Fungal Pathogen Genomics 2 – 6 June 2025 course timetable

Monday, 2nd of June

Time (BST)	Content
13:00 – 13:30	Meet & greet
13:30 – 15:45	**1. Database queries** 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	**2. Transcriptomics & Proteomics** Ensembl Fungi Track Hubs – p. 32 FungiDB Transcriptomic & Proteomic analysis– p. 40 SGD Expression tools - SPELL – p. 53
18:00 – 18:30	Q & A session

Tuesday, 3rd of June

Time (BST)	Content
13:00 – 13:30	Q & A session
13:30 – 15:45	**3. Enrichment analysis** SGD GO Slim mapper – p. 57 CGD GO Term finder – p. 60 FungiDB GO enrichment – p. 65
15:45 – 16:00	Break
16:00 – 17:00	Research Seminar: Prof Jason Stajich
17:00 – 18:00	**4. SNPs & Variants** SGD Variant viewer – p. 75 FungiDB SNP analysis & CNVs – p. 79 Exploring variants in Ensembl Fungi – p. 97 VEP – p. 103
18:00 – 18:30	Q & A session

Wednesday, 4th of June

Time (BST)	Content
13:00 – 13:30	Q & A session
13:30 – 15:45	**5. Comparative Genomics & Orthology and Evolutionary analysis & cross-species inference** Ensembl Fungi – WGA – p. 112 MycoCosm CAZy enzymes– p. 121 MycoCosm Synteny– p. 130 SGD predicting fungal biology– p. 135 Ensembl Fungi Evolutionary analysis (gene trees) – p. 143 FungiDB &OrthoMCL: Orthology and Phyletic Patterns – p. 159
15:45 – 16:00	Break
16:00 – 17:55	**6. Functional analysis: Pathways & metabolites** MycoCosm KEGG Browser & Secondary metabolism clusters – p. 165 FungiDB pathways & metabolites – p. 170
17:55 – 18:00	**Introduction to Group Projects**
18:00 – 18:30	Q & A session

Thursday 5th of June

Time (BST)	Content
13:00 – 13:30	Q & A session
13:30 – 15:45	**7. My Data sets** My workspace in FungiDB – p. 173 Exploring gene models – p. 187
15:45 – 16:00	Break
16:00 – 18:00	**8. Group Projects**
18:00 – 18:30	Q & A session

Friday, 6th of June

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

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.

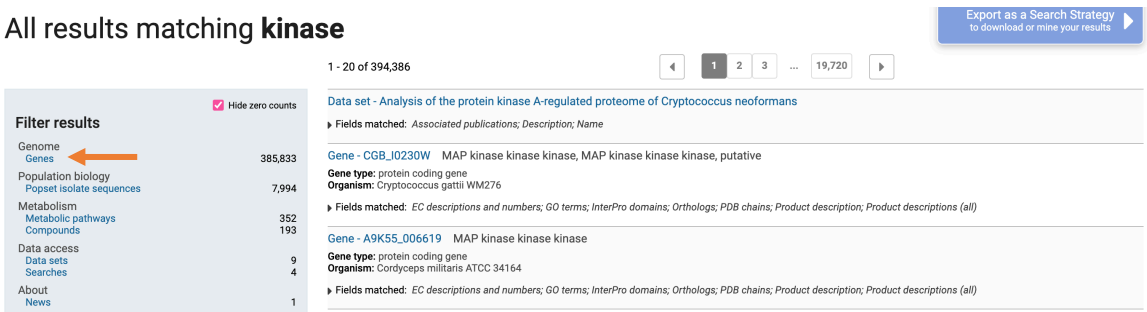
The site search can be accessed from the header of the site and is available from every page. The site search queries the database for a term (e.g., text) or a specific ID and returns a list of pages and documents that contain the query term.

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?

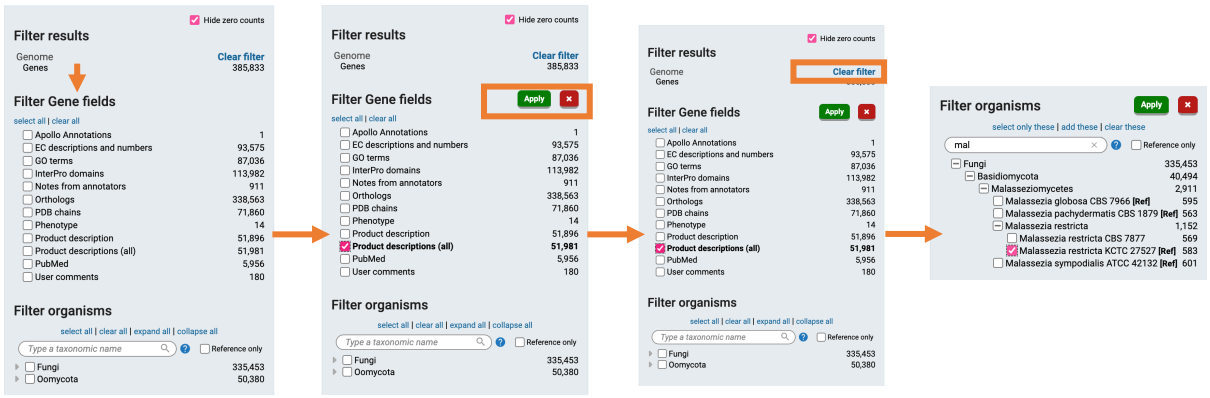
- Explore the filter panel on the left side of the page. Filter the results to view gene results only (hint: click on the word **Genes** in the "Filter results" section):



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.

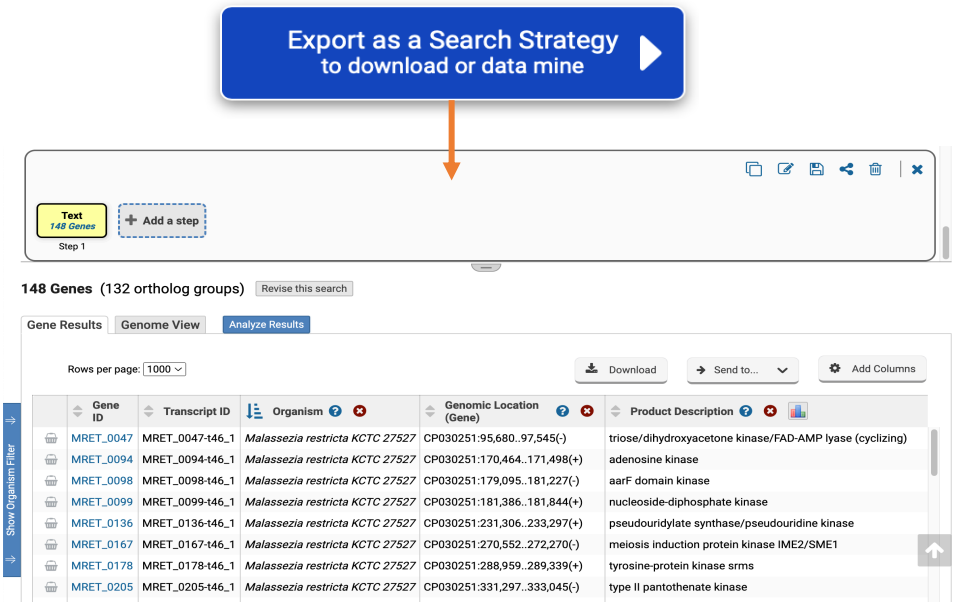
Note: The applied filter can be easily cleared by clicking on “Clear filter” option as shown in the screenshot below.



- 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?

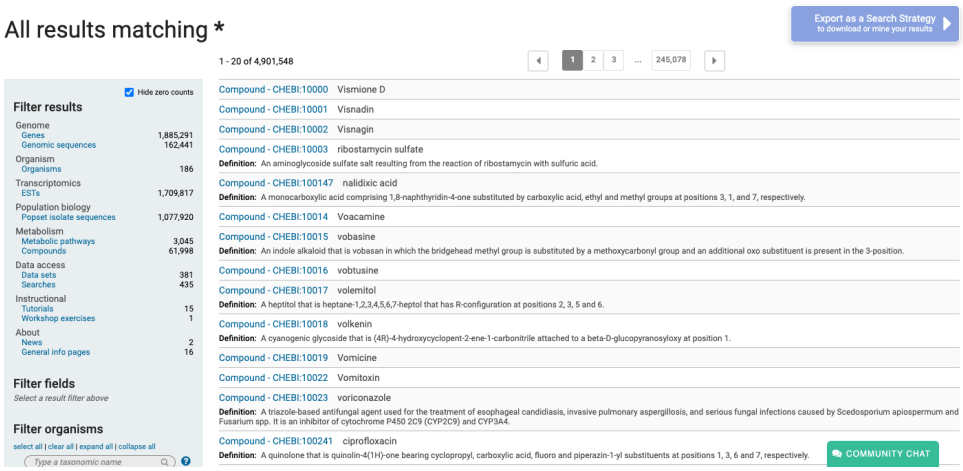
- Export the results to a search strategy.

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



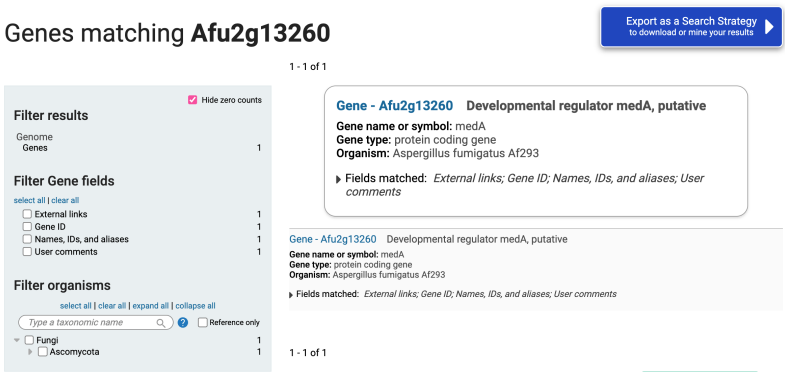
- Try running the same search but this time use a wild card (*) (e.g., kinase*).

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 - phosphofructo**kinase**). The wild card (*) can be used alone to retrieve all records available to the site search (see screenshot below).



- 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**.



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

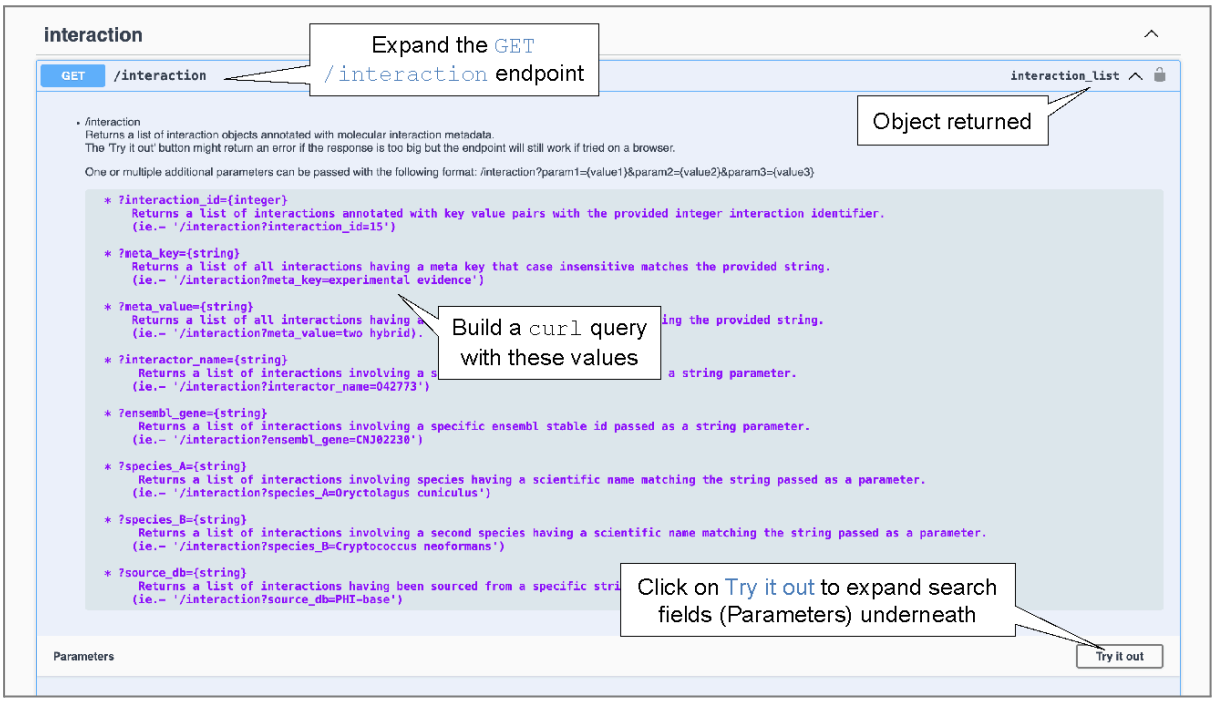
Links to be clicked shown in blue, text to be entered shown in red.

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.

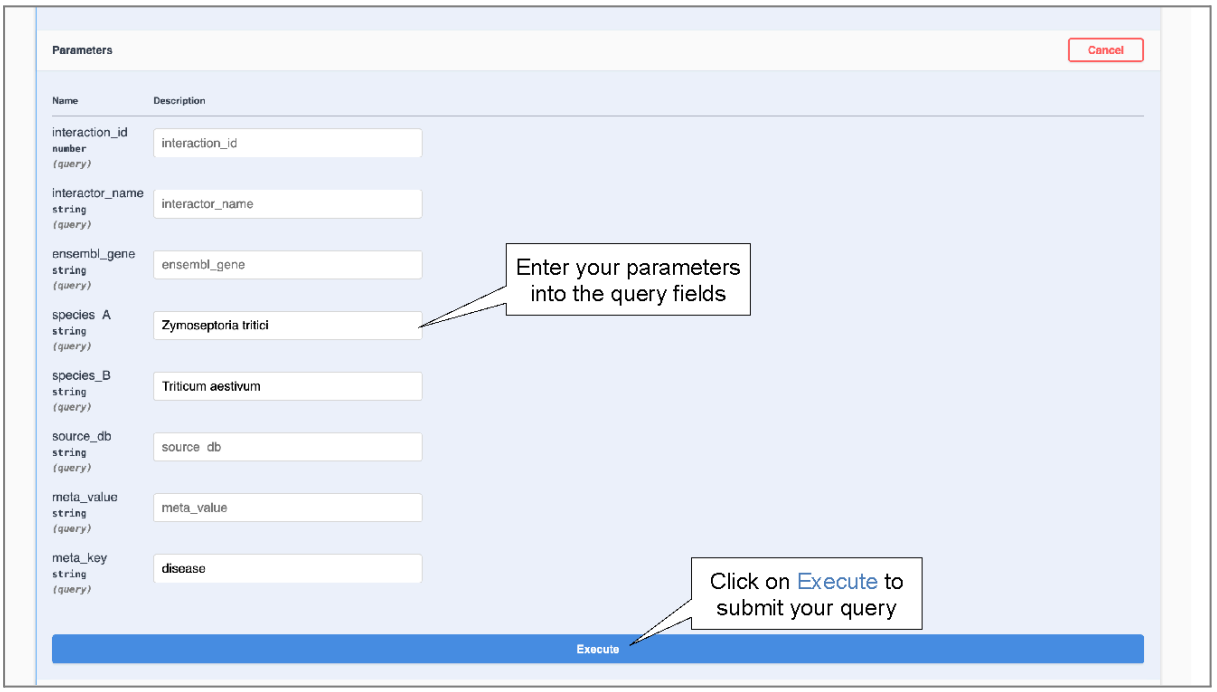
Step 1: Find all genes involved in molecular interactions for *Zymoseptoria tritici*.
From the Ensembl Interactions REST API page https://interactions.rest.ensembl.org, search for all *Zymoseptoria tritici* genes that have a pathogenic interaction with *Triticum aestivum* (wheat).

Enter https://interactions.rest.ensembl.org/interactions_by_prodname/ 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.

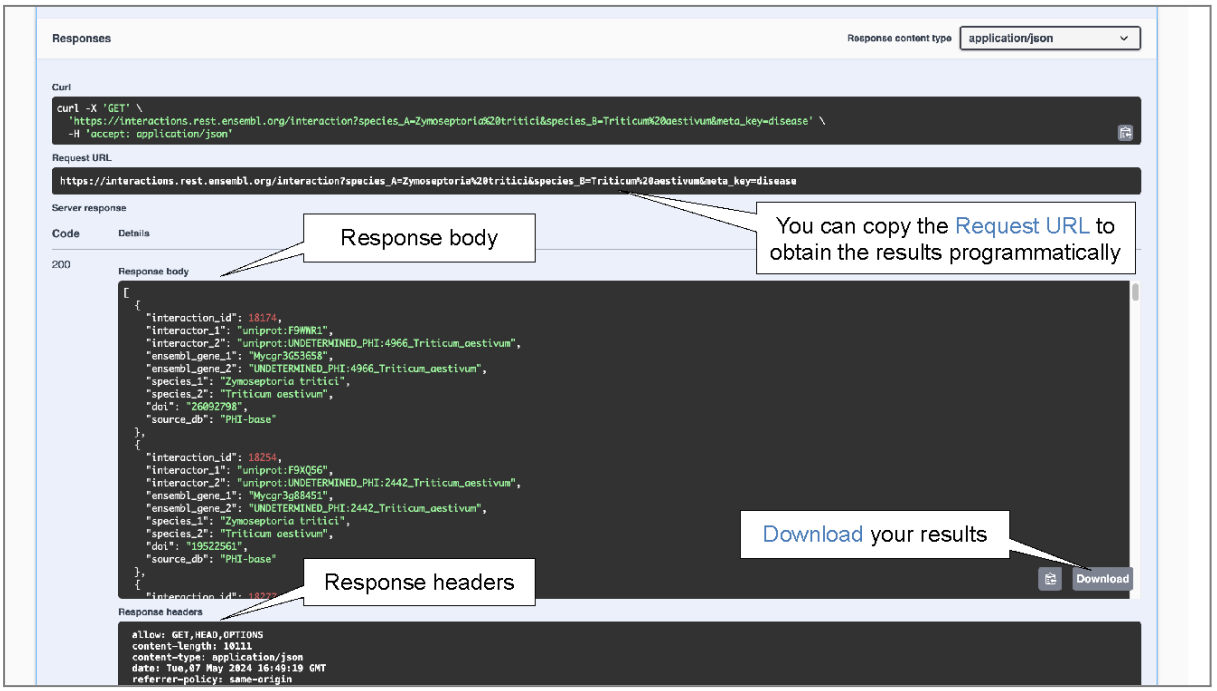


Scroll down to the 'Parameters' section and fill in the query fields as follows:
species_A: *Zymoseptoria tritici*
species_B: *Triticum aestivum*
meta_key: disease

Click on Execute to submit your request.



Scroll down to ‘Responses’ to view your output.



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

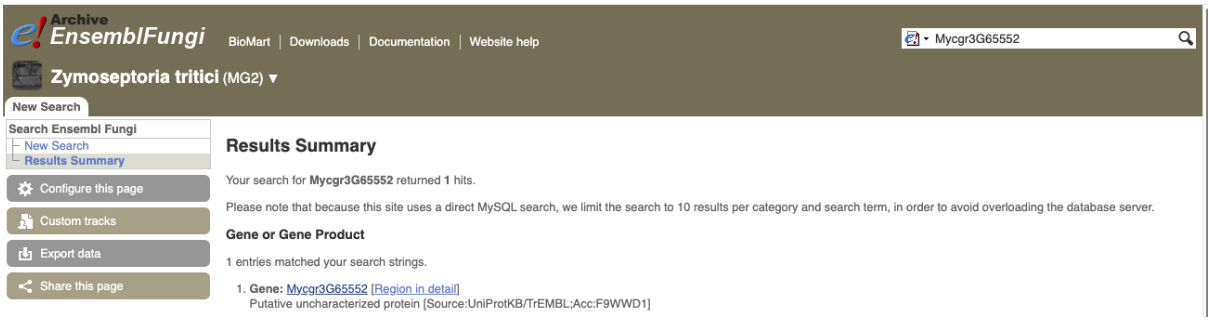
Under ‘Server response’, you should get the following output:

```
zymoseptoria_tritici": ["Mycgr3G53658", "Mycgr3g88451",
"Mycgr3G85040", "Mycgr3G40048", "Mycgr3G111221",
"Mycgr3G103264", "Mycgr3G89160", "Mycgr3G80707",
"Mycgr3G65552", "Mycgr3g105487", "Mycgr3G70181",
"Mycgr3G46840", "Mycgr3G93828", "Mycgr3G31676",
"Mycgr3G51018", "Mycgr3G36951", "Mycgr3G77528",
"Mycgr3G39611", "Mycgr3G96592", "Mycgr3G86705",
"Mycgr3G107320", "Mycgr3G74194", "Mycgr3G87000",
"Mycgr3G100355", "Mycgr3G92404", "Mycgr3G69942"]
```

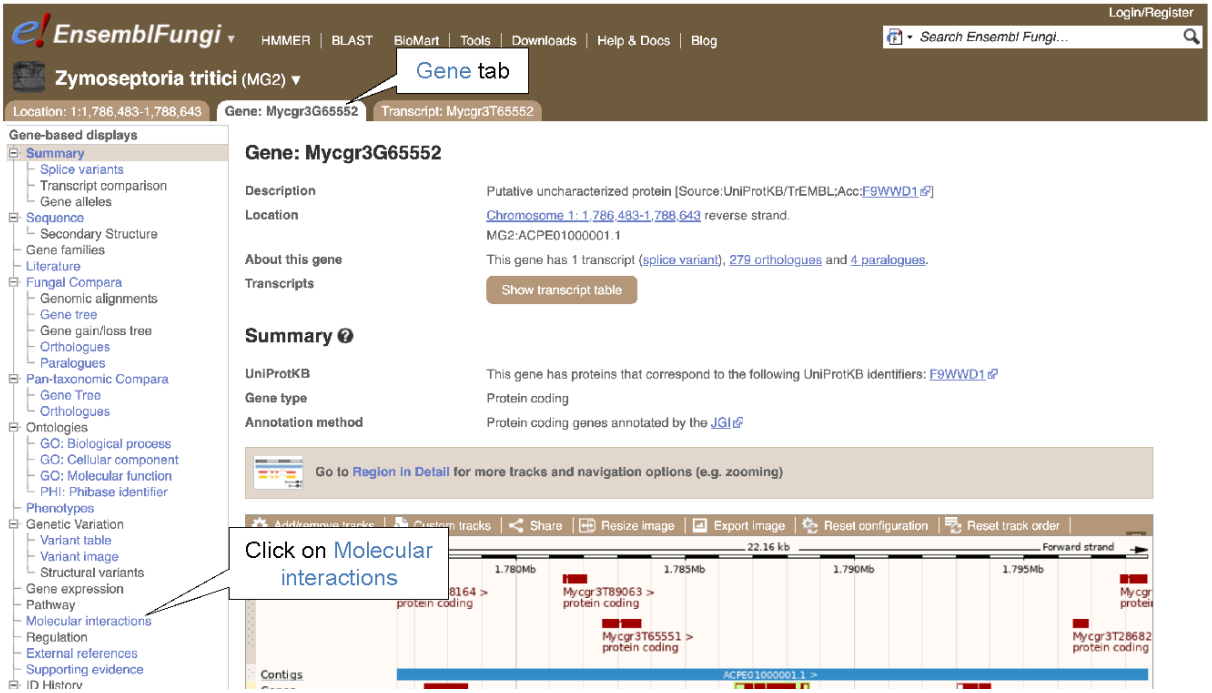

Step 2: Let's find out more about the gene Mycgr3G65552 in the Ensembl Fungi browser.

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

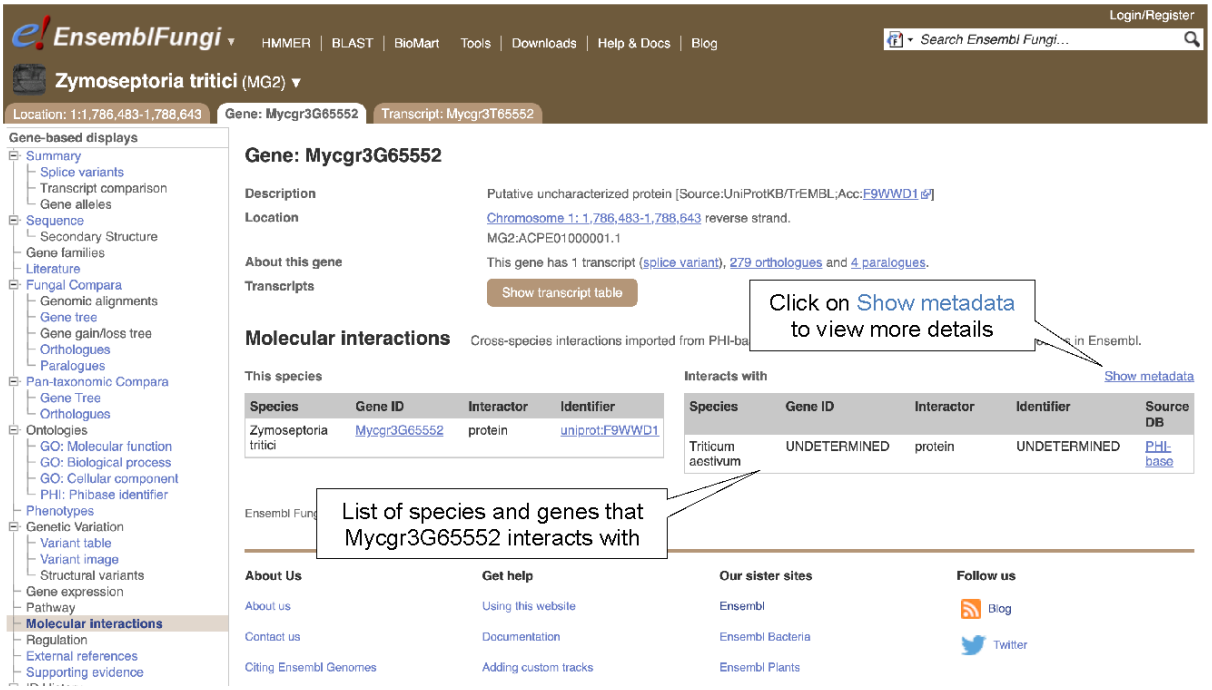
On the Ensembl Fungi 59 homepage, in the drop down menu "All genomes" select Zymoseptoria tritici. enter the gene ID Mycgr3G65552 in the top right-hand corner and hit Search. Click on the gene ID Mycgr3G65552 to open the 'Gene' tab.



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.



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



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.

Interacts with Show metadata

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

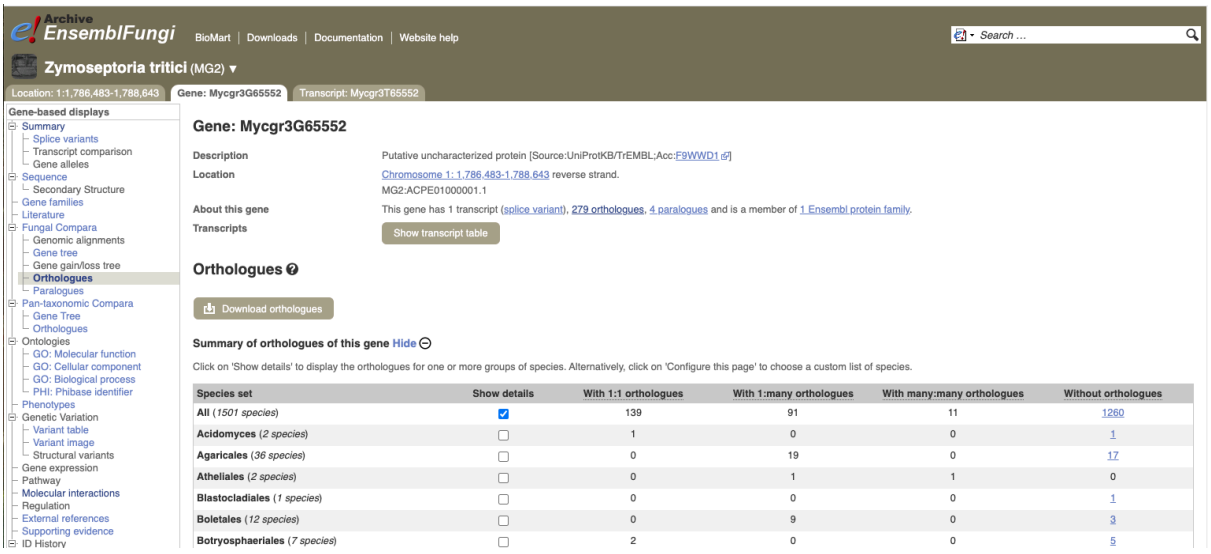
Can you find out what the phenotype for this interaction is? Click on Show metadata at the top right-hand corner of the 'Interacts with' table. Based on PHI-base, the interaction is associated with 'loss of pathogenicity'.

Interacts with Show metadata

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

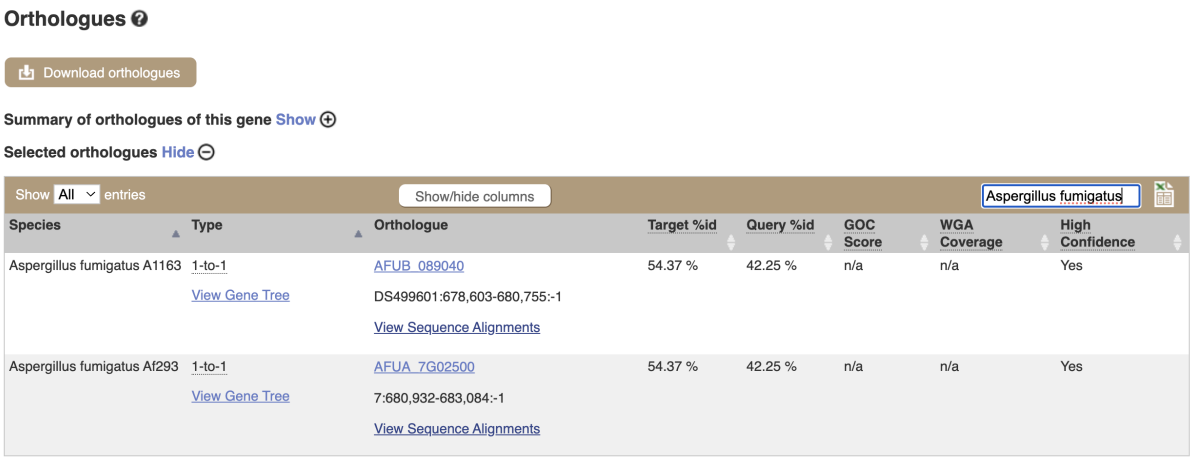
Experimental evidence	gene complementation
Interaction type	interspecies interaction
Interaction phenotype	PHIPO:0000010
Disease name	PHIDO:0000331
Pathogen protein modification	gene deletion: full
PHI-base high level term	Loss of pathogenicity
Pathogen experimental strain	IPO323
Host experimental strain	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.

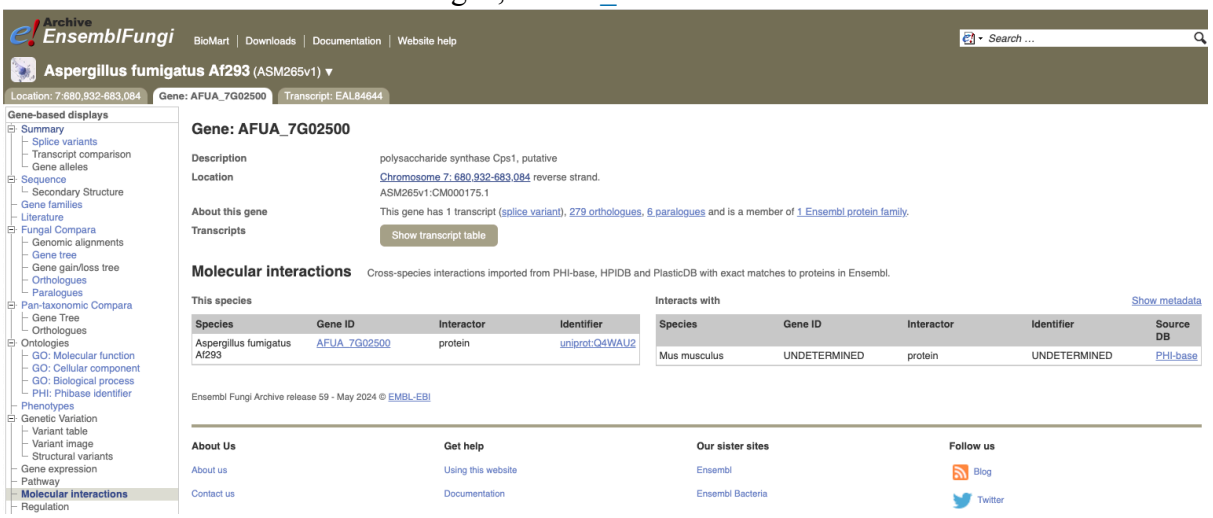


Can you find out if there are any orthologues in *Aspergillus fumigatus* with molecular interaction entries?

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.



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.



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

Interacts with Show metadata

Species	Gene ID	Interactor	Identifier	Source DB
Mus musculus	UNDETERMINED	protein	UNDETERMINED	PHI-base
Several experiments exist for this interaction. Please click here for more information				
Mus musculus	UNDETERMINED	protein	UNDETERMINED	PHI-base

Experimental evidence	gene complementation
Interaction type	interspecies interaction
Interaction phenotype	PHIPO:0000015
Disease name	PHIDO:0000020
Pathogen protein modification	gene deletion: full
PHI-base high level term	Reduced virulence
Pathogen experimental strain	Af293
Host experimental strain	C57BL/6

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

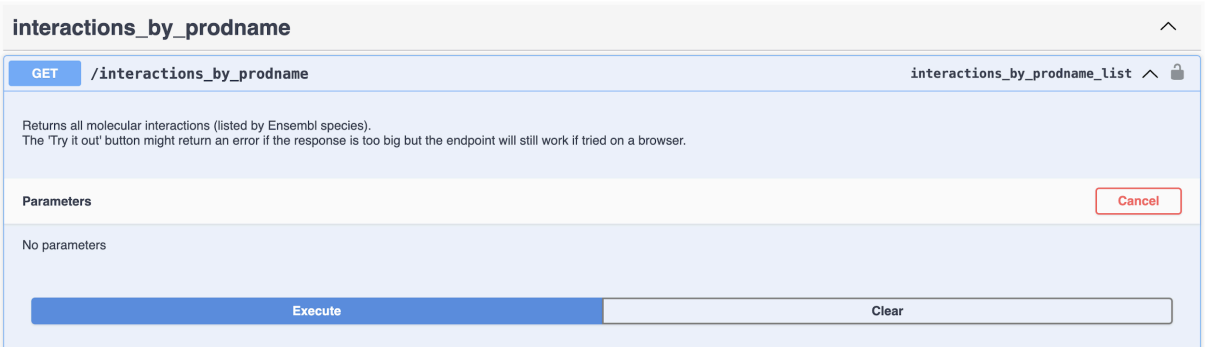
Additional host-pathogen exercise 1 – Exploring GO terms and phenotypes

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:

(a) Using the Ensembl Interactions REST API, can you retrieve all genes with molecular interaction information for *B. cinerea*?
(b) Open the 'Molecular interactions' page for the Bcin07g00720 gene in *B. cinerea*. What plant species does the gene interact with?
(c) Can you find the phenotype that is reported for each of the species the gene interacts with?
(d) Find all fungal orthologues. Is there any orthologue in *Magnaporthe oryzae* for Bcin07g00720? For which orthologue is molecular interaction information available?
(e) Which species does the *M. oryzae* orthologue interact with?
(f) 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:

(a) Go to the Ensembl Interactions REST API and expand the `GET /interactions_by_prodname` endpoint documentation. Click on Try it out and then Execute.

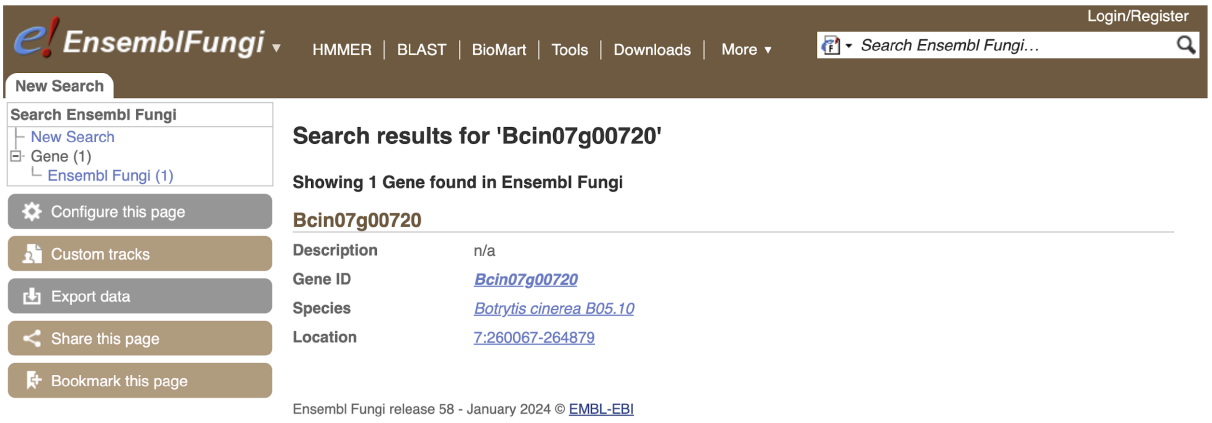


In the 'Response body', search for botrytis_cynerea. Alternatively, you can open the request URL https://interactions.rest.ensembl.org/interactions_by_prodname in your browser. You should get the following output:

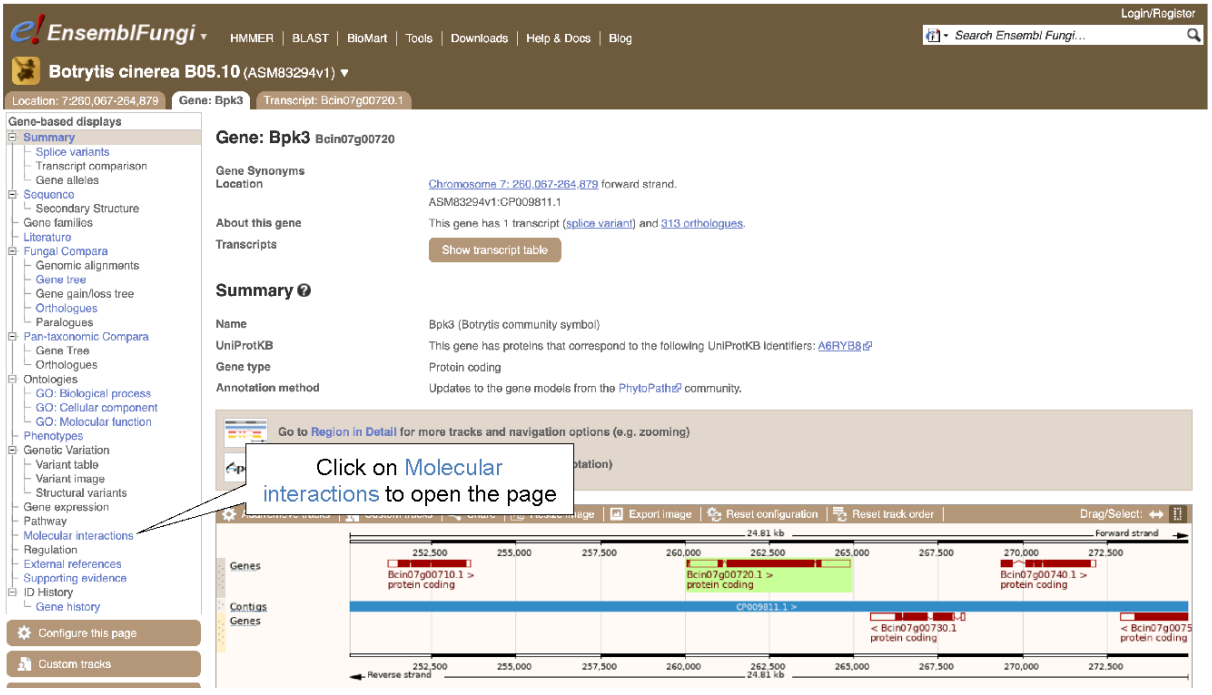
```
"botrytis_cinerea": ["Bcin07g00720", "Bcin02g02570",
"Bcin12g04900", "Bcin16g00630", "Bcin02g06770",
"Bcin03g07190", "Bcin09g02390", "Bcin09g01800",
"Bcin07g03050", "Bcin08g05150", "Bcin10g01250",
```

```
"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.



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



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

Species	Gene ID	Interactor	Identifier
Botrytis cinerea B05.10	Bcin07g00720	protein	uniprot:A6RYB8

Interacts with — Show metadata

Species	Gene ID	Interactor	Identifier	Source DB
Solanum lycopersicum	UNDETERMINED	protein	UNDETERMINED	PHI-base
Vitis vinifera	UNDETERMINED	protein	UNDETERMINED	PHI-base
Cucumis sativus	UNDETERMINED	protein	UNDETERMINED	PHI-base
Malus domestica	UNDETERMINED	protein	UNDETERMINED	PHI-base

(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.

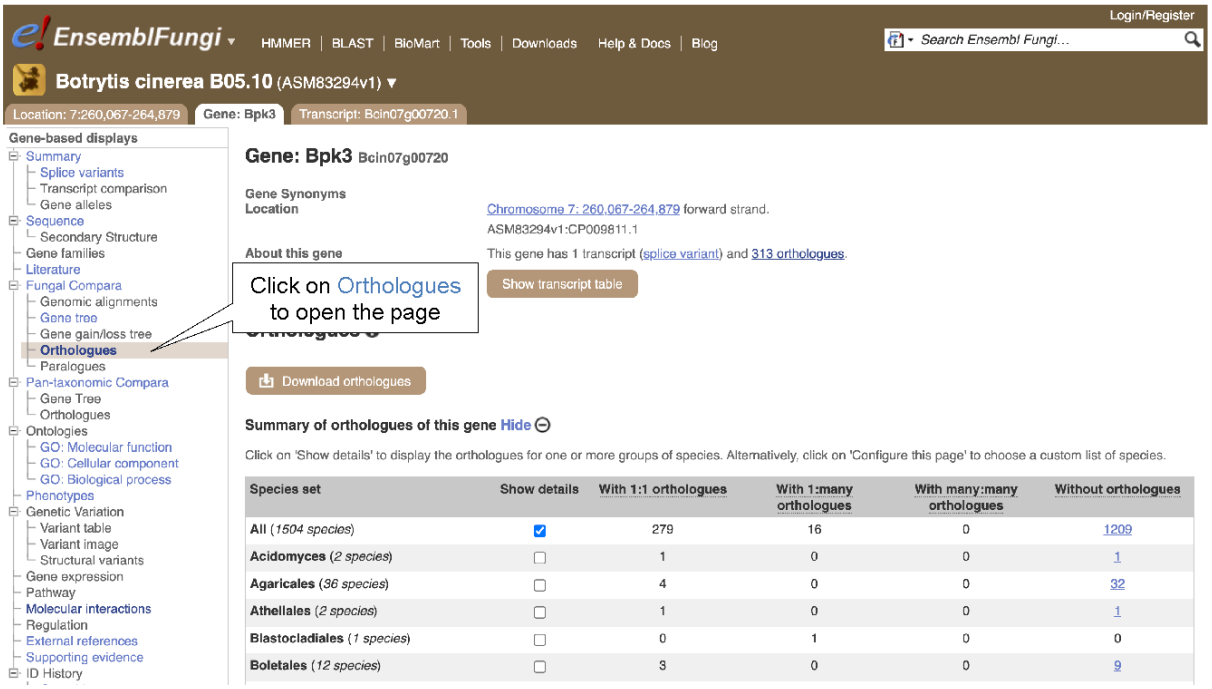
This species

Species	Gene ID	Interactor	Identifier
Botrytis cinerea B05.10	Bcin07g00720	protein	uniprot:A6RYB8

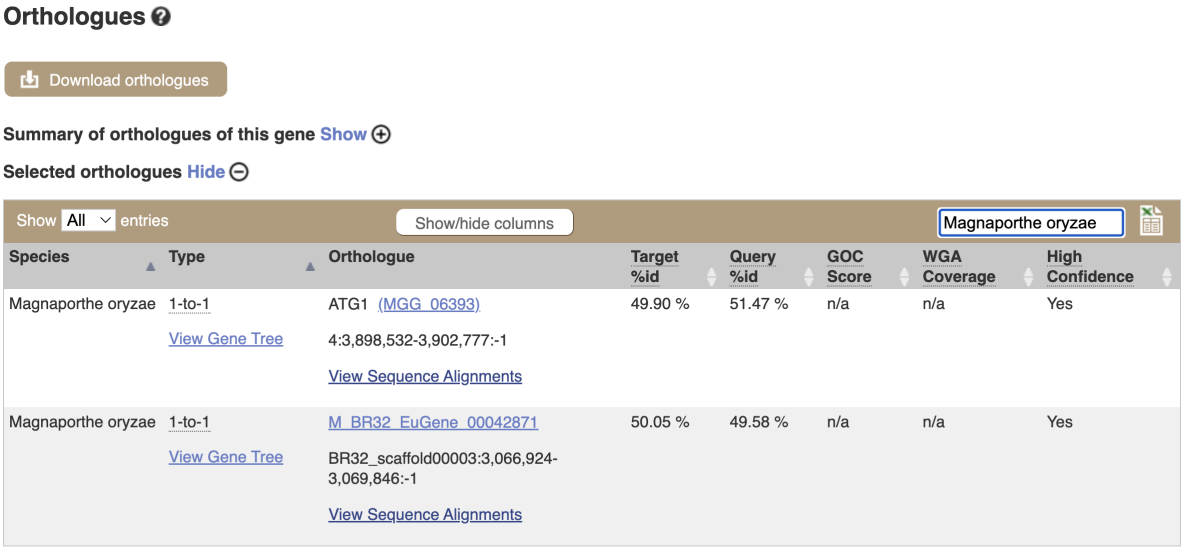
Interacts with — Show metadata

Species	Gene ID	Interactor	Identifier	Source DB
Solanum lycopersicum	UNDETERMINED	protein	UNDETERMINED	PHI-base
Interaction type	interspecies interaction			
Interaction phenotype	PHIPO:0000015			
Disease name	PHIDO:0000178			
Pathogen protein modification	gene mutation; gene complementation			
PHI-base high level term	Reduced virulence			
Pathogen experimental strain	B05.10			
Vitis vinifera	UNDETERMINED	protein	UNDETERMINED	PHI-base
Interaction type	interspecies interaction			
Interaction phenotype	PHIPO:0000015			
Disease name	PHIDO:0000178			
Pathogen protein modification	gene mutation; gene complementation			
PHI-base high level term	Reduced virulence			
Pathogen experimental strain	B05.10			
Cucumis sativus	UNDETERMINED	protein	UNDETERMINED	PHI-base
Interaction type	interspecies interaction			
Interaction phenotype	PHIPO:0000010			
Disease name	PHIDO:0000178			
Pathogen protein modification	gene mutation; gene complementation			
PHI-base high level term	Loss of pathogenicity			
Pathogen experimental strain	B05.10			
Malus domestica	UNDETERMINED	protein	UNDETERMINED	PHI-base
Interaction type	interspecies interaction			
Interaction phenotype	PHIPO:0000015			
Disease name	PHIDO:0000178			
Pathogen protein modification	gene mutation; gene complementation			
PHI-base high level term	Reduced virulence			
Pathogen experimental strain	B05.10			

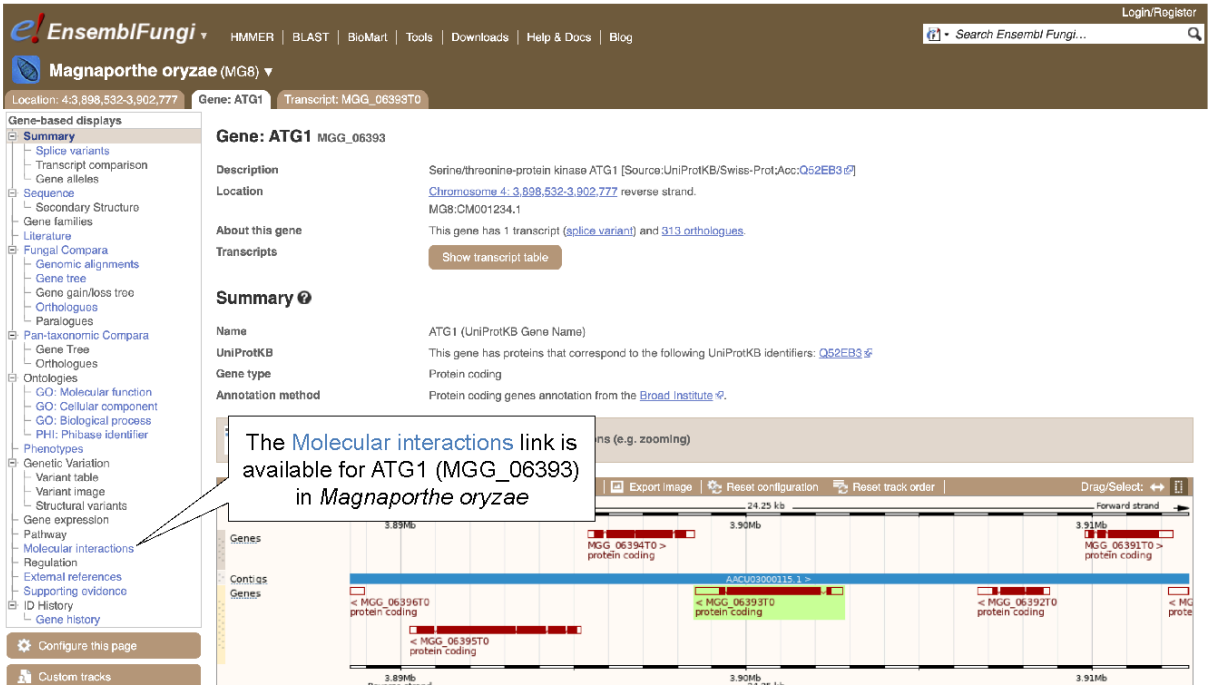
(d) To retrieve all fungal orthologues, go to Fungal Compara: Orthologues in the left-hand panel.



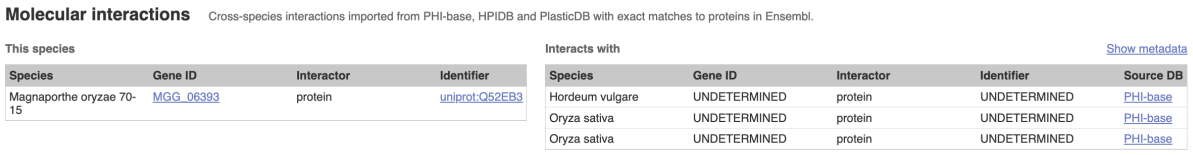
Scroll down to the Orthologues table and use the filter box in the top right-hand corner to search for *Magnaporthe oryzae*.



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).



(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

Species	Gene ID	Interactor	Identifier
Magnaporthe oryzae 70-15	MGG_06393	protein	uniprot:Q52EB3

Interacts with — Show metadata

Species	Gene ID	Interactor	Identifier	Source DB
Hordeum vulgare	UNDETERMINED	protein	UNDETERMINED	PHI-base
Oryza sativa	UNDETERMINED	protein	UNDETERMINED	PHI-base
Oryza sativa	UNDETERMINED	protein	UNDETERMINED	PHI-base

(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’

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

This species

Species	Gene ID	Interactor	Identifier
Botrytis cinerea B05.10	Bcin07g00720	protein	uniprot:A6RYB8

Interacts with — Show metadata

Species	Gene ID	Interactor	Identifier	Source DB
Solanum lycopersicum	UNDETERMINED	protein	UNDETERMINED	PHI-base
Interaction type	interspecies interaction			
Interaction phenotype	PHIPO:0000015			
Disease name	PHIDO:0000178			
Pathogen protein modification	gene mutation; gene complementation			
PHI-base high level term	Reduced virulence			
Pathogen experimental strain	B05.10			
Vitis vinifera	UNDETERMINED	protein	UNDETERMINED	PHI-base
Interaction type	interspecies interaction			
Interaction phenotype	PHIPO:0000015			
Disease name	PHIDO:0000178			
Pathogen protein modification	gene mutation; gene complementation			
PHI-base high level term	Reduced virulence			
Pathogen experimental strain	B05.10			
Cucumis sativus	UNDETERMINED	protein	UNDETERMINED	PHI-base
Interaction type	interspecies interaction			
Interaction phenotype	PHIPO:0000010			
Disease name	PHIDO:0000178			
Pathogen protein modification	gene mutation; gene complementation			
PHI-base high level term	Loss of pathogenicity			
Pathogen experimental strain	B05.10			
Malus domestica	UNDETERMINED	protein	UNDETERMINED	PHI-base
Interaction type	interspecies interaction			
Interaction phenotype	PHIPO:0000015			
Disease name	PHIDO:0000178			
Pathogen protein modification	gene mutation; gene complementation			
PHI-base high level term	Reduced virulence			
Pathogen experimental strain	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

Species	Gene ID	Interactor	Identifier
Magnaporthe oryzae 70-15	MGG_06393	protein	uniprot:Q52EB3

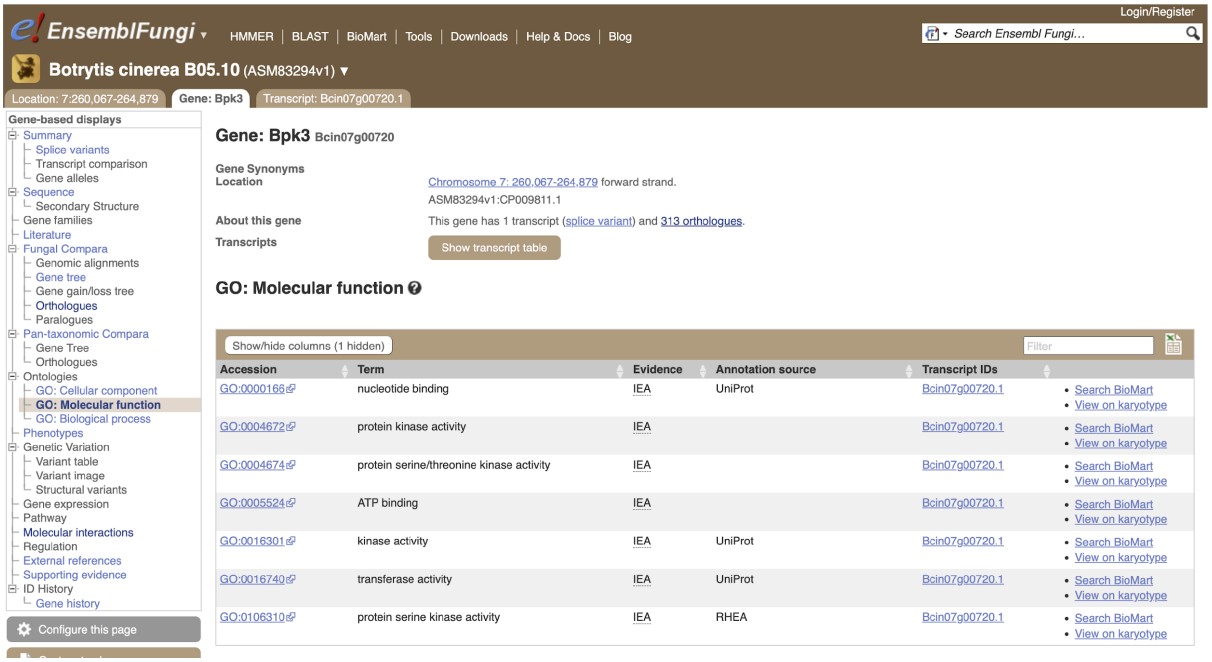
Interacts with — Show metadata

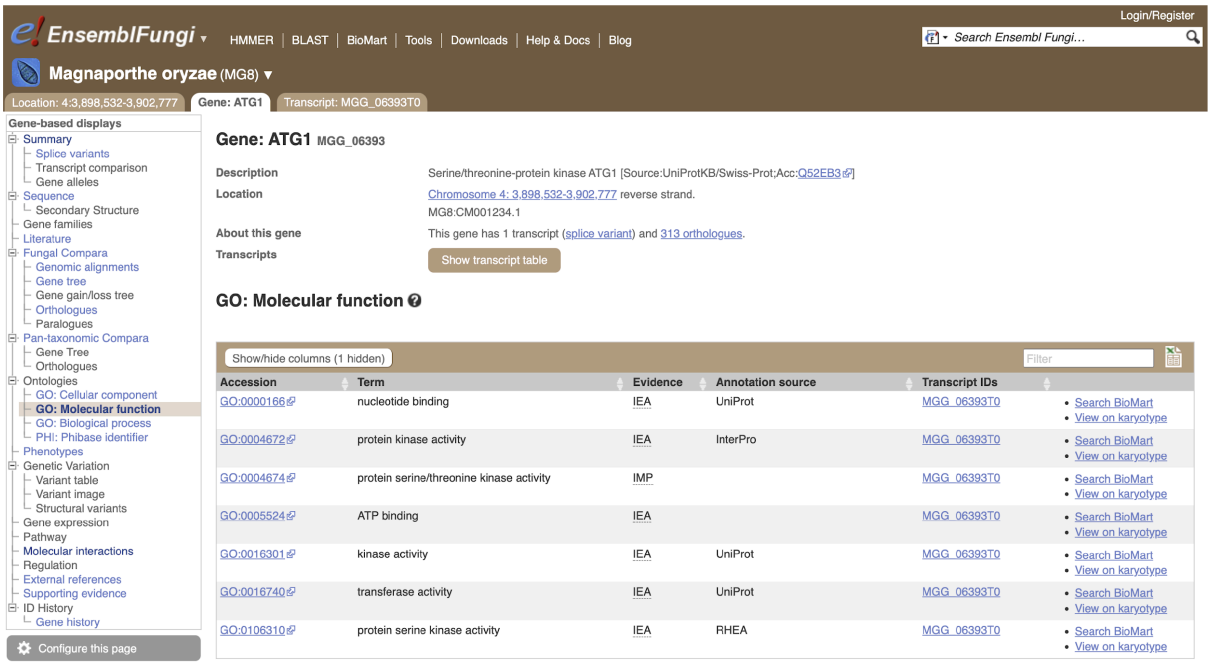
Species	Gene ID	Interactor	Identifier	Source DB
Hordeum vulgare	UNDETERMINED	protein	UNDETERMINED	PHI-base
Several experiments exist for this interaction. Please click here for more information				
Oryza sativa	UNDETERMINED	protein	UNDETERMINED	PHI-base
Several experiments exist for this interaction. Please click here for more information				
Oryza sativa	UNDETERMINED	protein	UNDETERMINED	PHI-base
Interaction type	interspecies interaction			
Interaction phenotype	PHIPO:0000010			
Disease name	PHIDO:0000315			
Pathogen protein modification	gene deletion: full			
PHI-base high level term	Loss of pathogenicity			
Pathogen experimental strain	Guy11			
Host experimental strain	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

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



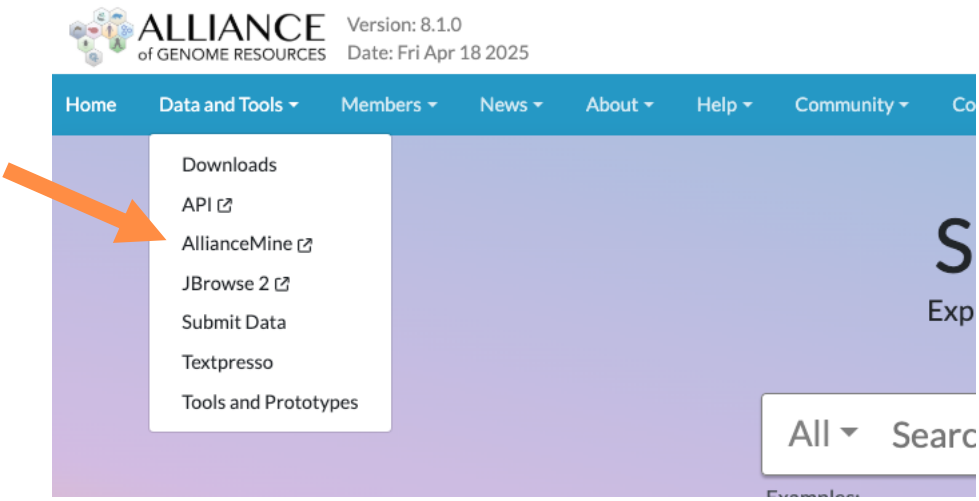


Search Strategies in SGD

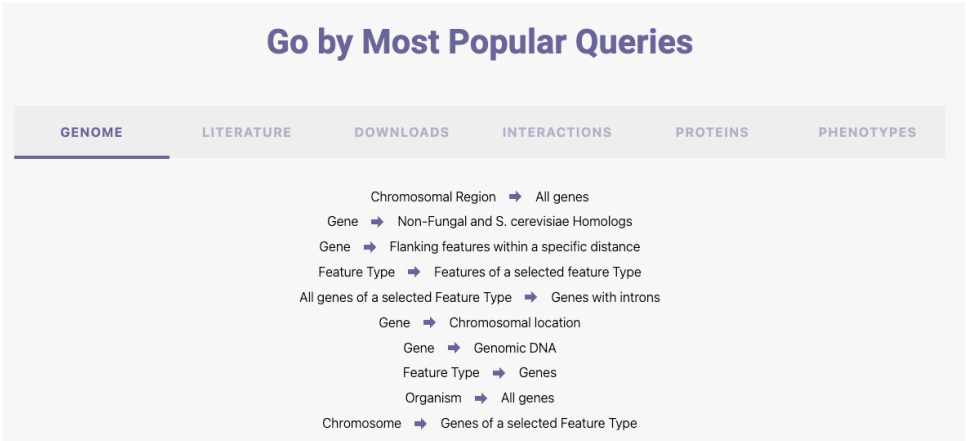
In addition to a faceted search tool, AllianceMine (within the Alliance of Genome Resources) is a means for SGD users to conduct more advanced queries. AllianceMine enables rapid retrieval and manipulation of curated biological data on *S. cerevisiae* genes and genomic features. By creating gene lists, 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.

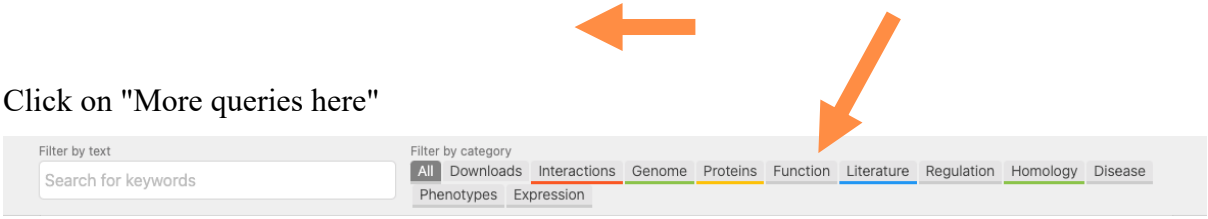
- Access AllianceMine from Alliance home page (https://www.alliancegenome.org); click on AllianceMine from the Data and Tools section at the top of the page.



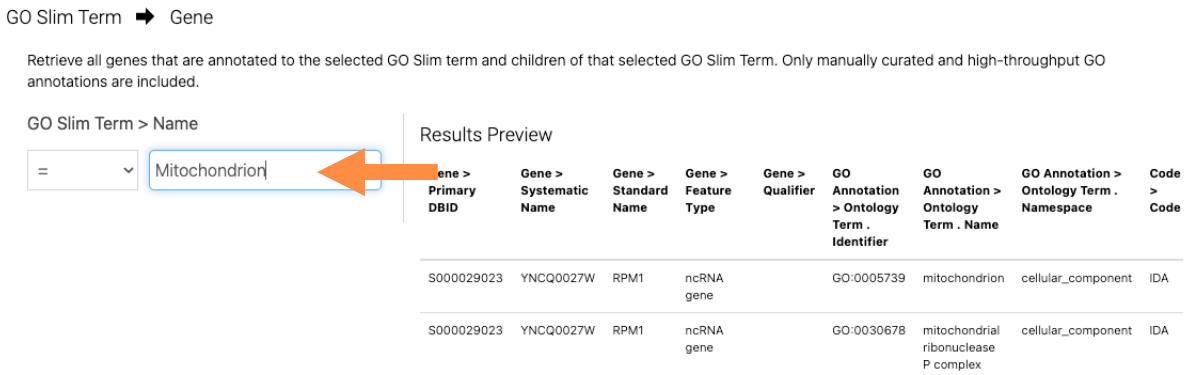
1. **Create a list of proteins that are known subunits of the mitochondrial ribosome (MTR):**



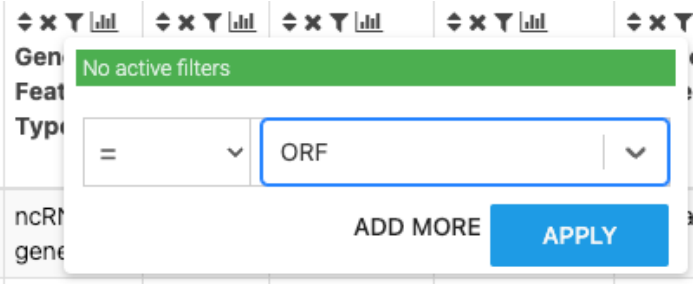
- Click on "More queries here"



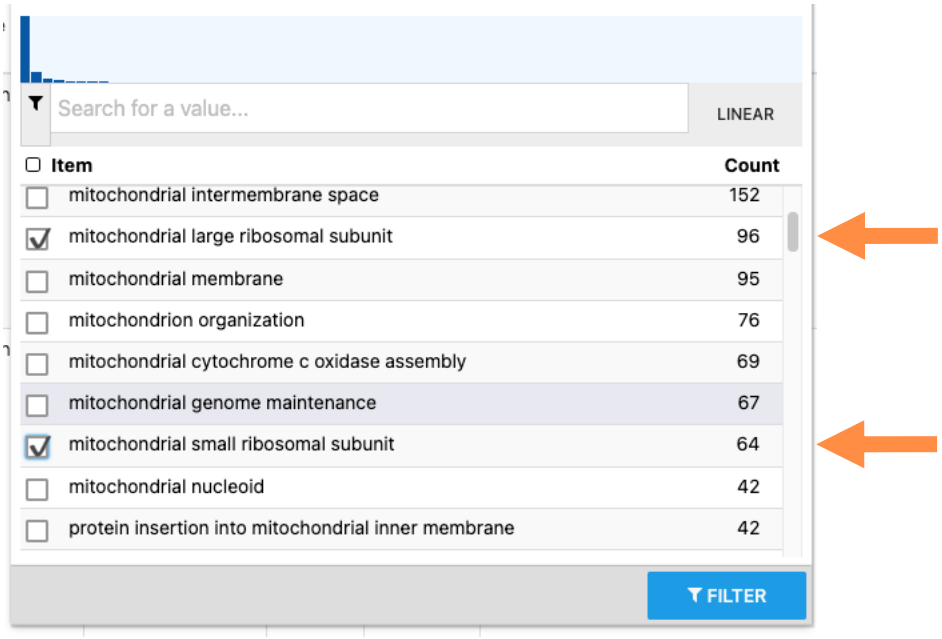
- And then select the **FUNCTION** tab and then **GO Slim Term => Gene**. Enter "mitochondrion" as your GO slim term. This will return many results. Go to the bottom of the results and click "**View ~39,000 rows**."
  [NOTE: if you get zero results, click the "Edit Query" button and use the trash can 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]



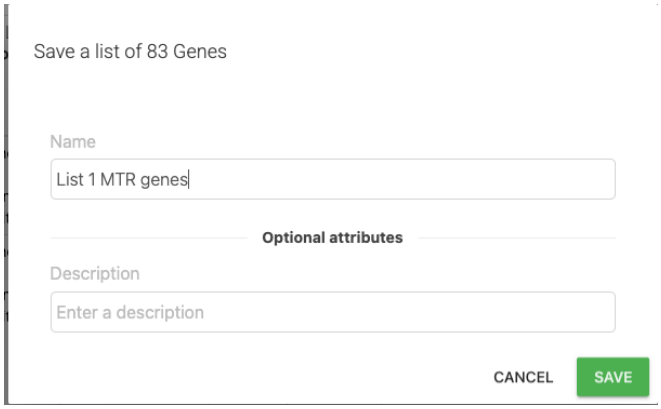
- In the Query Results, first go to the **Gene Feature Type** column, click the filter icon and then select "**ORF**" from the drop-down menu and "Apply."



- 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.

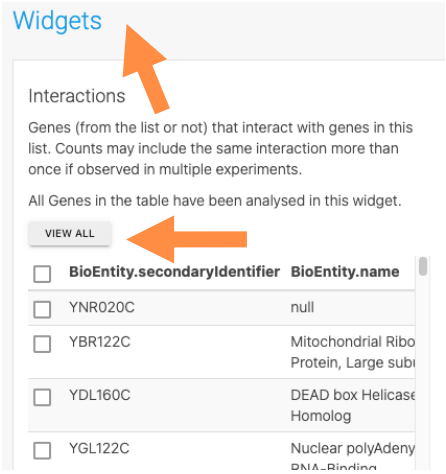


Save this list by clicking the **Save List** button on the upper right of the table and selecting **"Genes (83)"** at the top of the pull-down. Give it the name "List 1 MTR genes" and save.

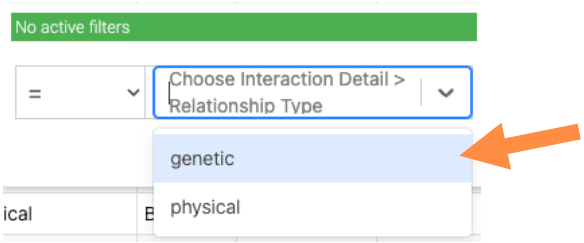


2. Find proteins that genetically interact with MTR proteins:

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

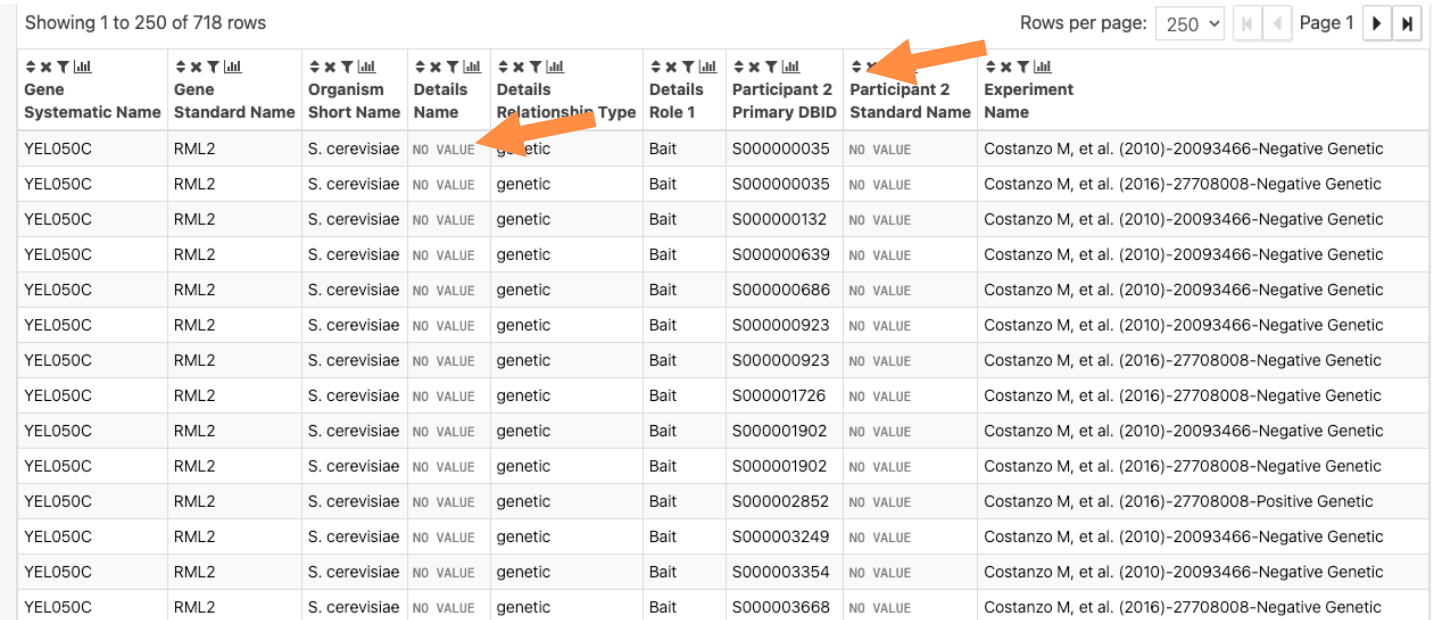


- The results table shows all genes/proteins with genetic or physical interactions with the MTR genes. There are over 17K of them.
- Go to the column for "**Details Relationship Type**" and filter for "**Genetic**." Hit Apply and you'll get a list of ~8500 rows.



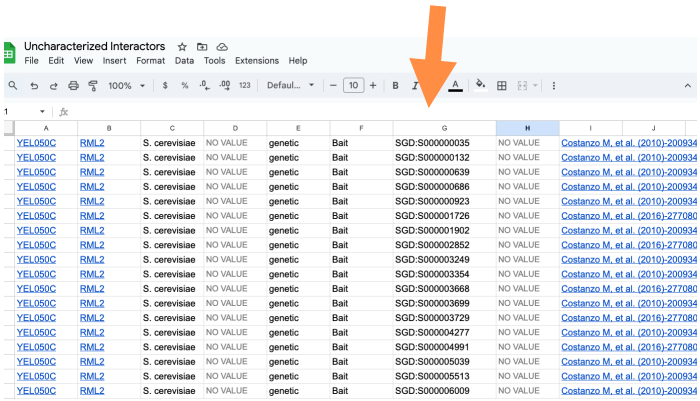
3. **Find MTR interactors that are uncharacterized:**

- 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.



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	S000000035	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000000035	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000000132	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000000639	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000000686	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000000923	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000000923	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

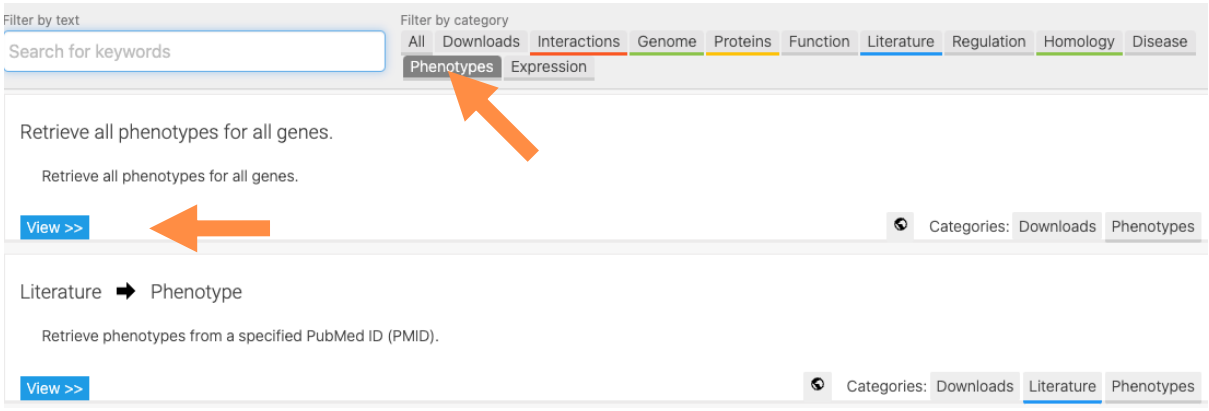
- Given the current state of bugginess in this tool (it's undergoing major revision), you'll do best to copy and paste the section of the list that includes just the rows that say "no value" in the Participant 2 column. Copy this section and paste it into a blank worksheet.
- In the worksheet, sort by **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)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000000132	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000000639	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000000686	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000000923	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000001726	NO VALUE	Costanzo M, et al. (2016)-27708	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000001902	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000002852	NO VALUE	Costanzo M, et al. (2016)-27708	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000003249	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000003354	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000003668	NO VALUE	Costanzo M, et al. (2016)-27708	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000003699	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000003729	NO VALUE	Costanzo M, et al. (2016)-27708	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000004277	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000004991	NO VALUE	Costanzo M, et al. (2016)-27708	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000005039	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000005513	NO VALUE	Costanzo M, et al. (2010)-20093	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000006009	NO VALUE	Costanzo M, et al. (2010)-20093	

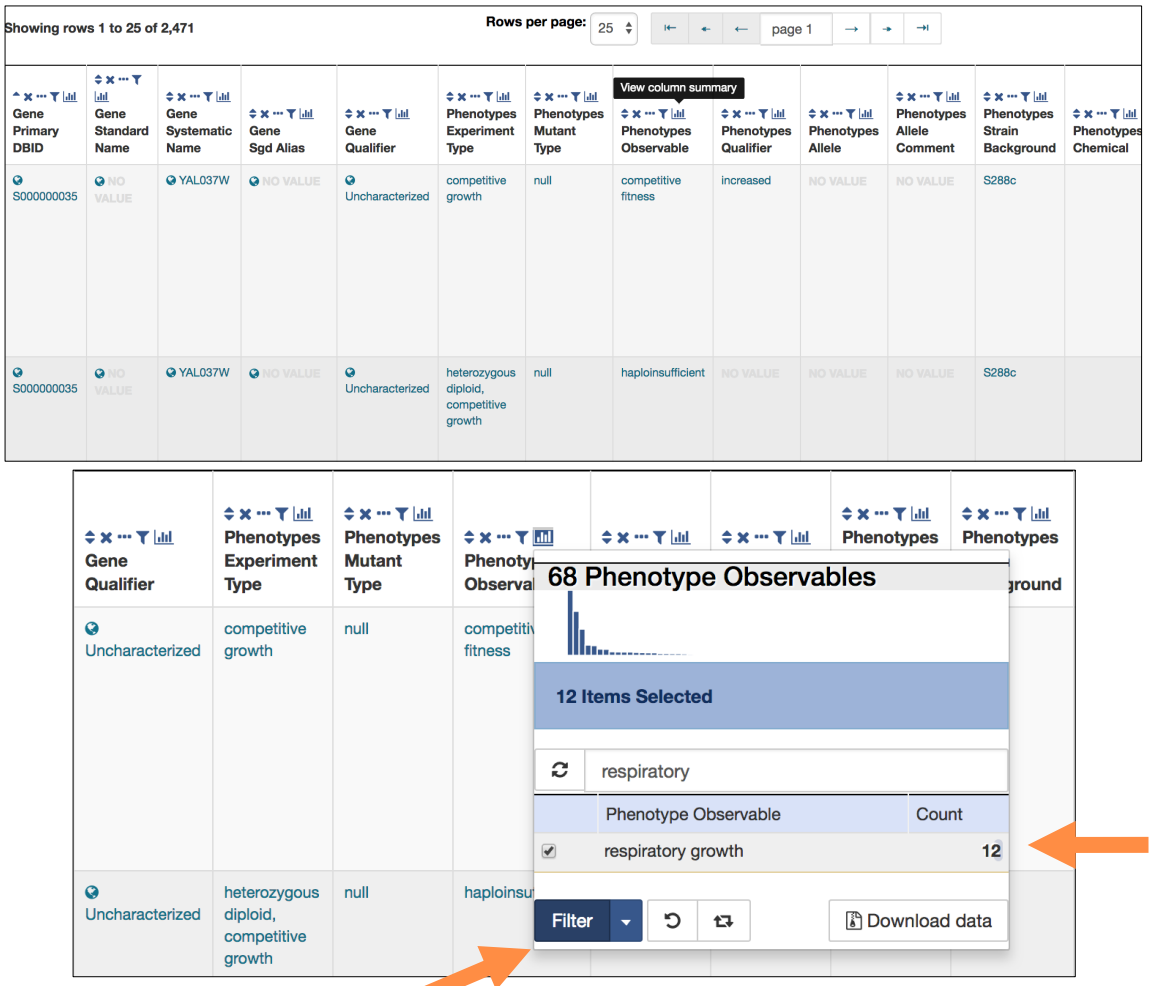
- It would be a good idea to narrow down our candidates even more. For example, because the MTR is a mitochondrial complex, we would expect that deleting uncharacterized (but bona fide) subunits of the MTR would disrupt aerobic respiration. Let’s refine our list of predicted MTR subunits by seeing which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

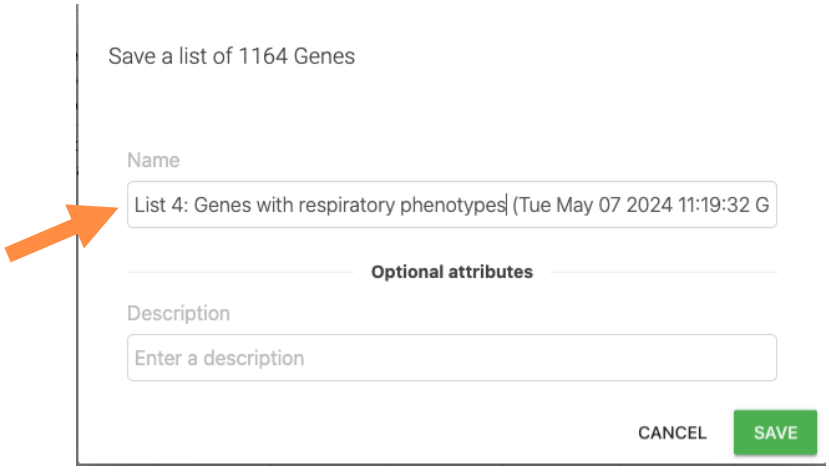


- You'll get a massive set of results but go ahead and open the table by clicking on the list of results at the bottom.
- In the Results table, find a column labeled **Phenotypes Observable**. Hover your mouse over the small icons above the column name and click on the bar graph icon.

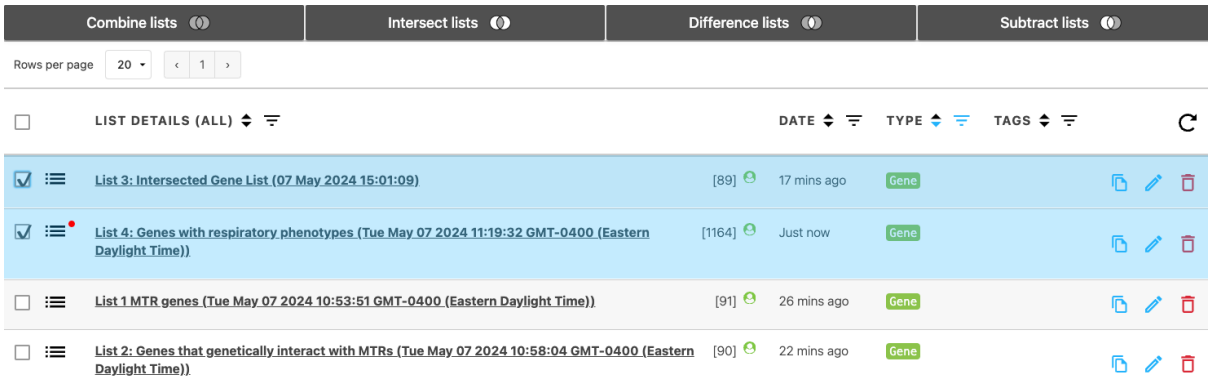
- In the **Filter values** box, enter **respiratory** and scroll down the list to check the box next to **Respiratory growth**; hit **Filter**.



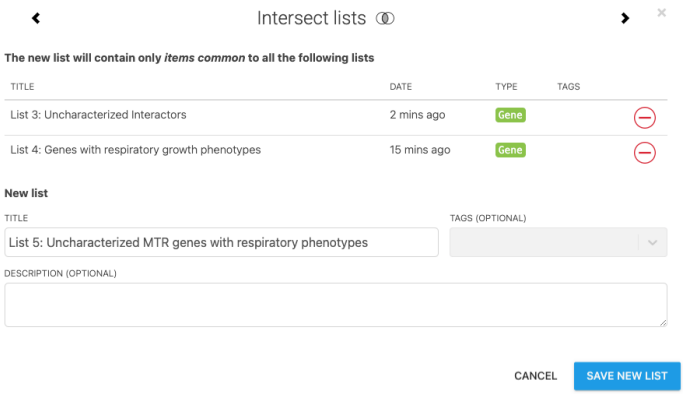
- To filter the phenotypes for those where respiratory growth is impeded, find the **Phenotype Qualifiers** column and click the bar graph icon again. Select all items that refer to hindering respiratory growth: “decreased”, “decreased rate”, “absent”, etc. Then, hit **Filter.**
- Now the save the gene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



- 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**."

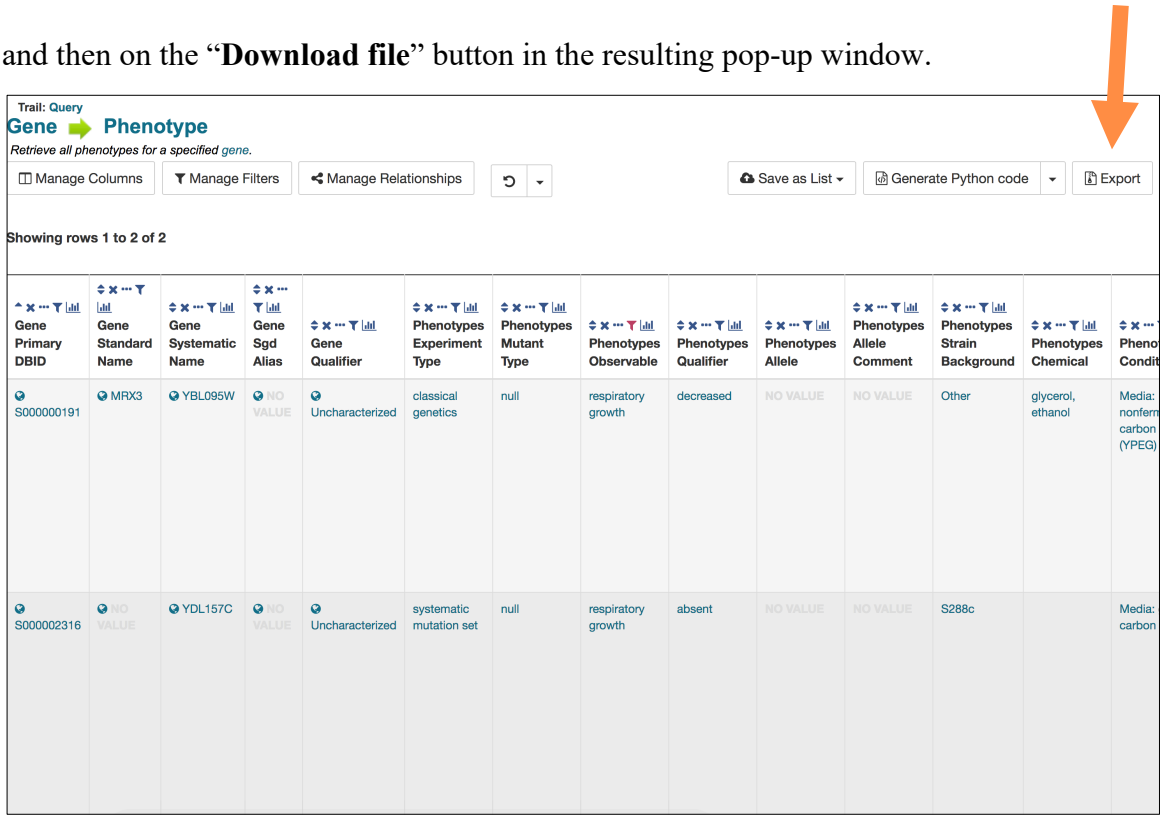


- Save your intersected list as " List 5: Uncharacterized MTR genes with respiratory phenotypes"



- 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.



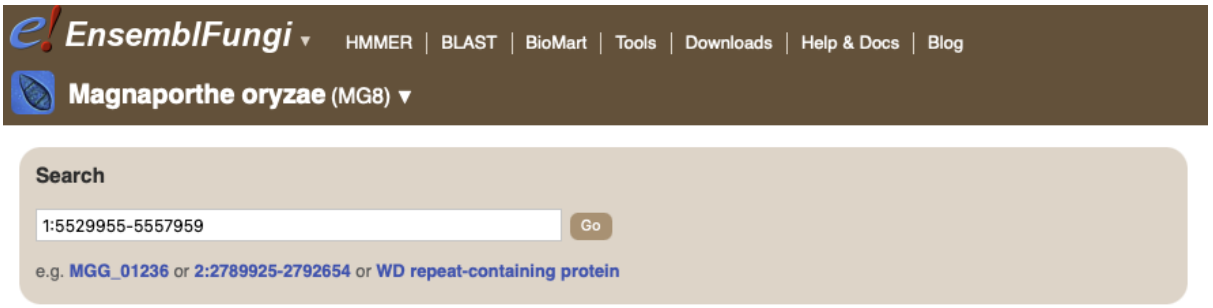
- The results of the above AllianceMine analysis suggest 2 genes that potentially encode undiscovered subunits of the mitochondrial ribosome. Although these genes are uncharacterized, more data may exist on their orthologs in other organisms. Using the exported Gene IDs, you can use FungiDB to survey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

Exercise: Attaching Track Hubs to Ensembl Fungi

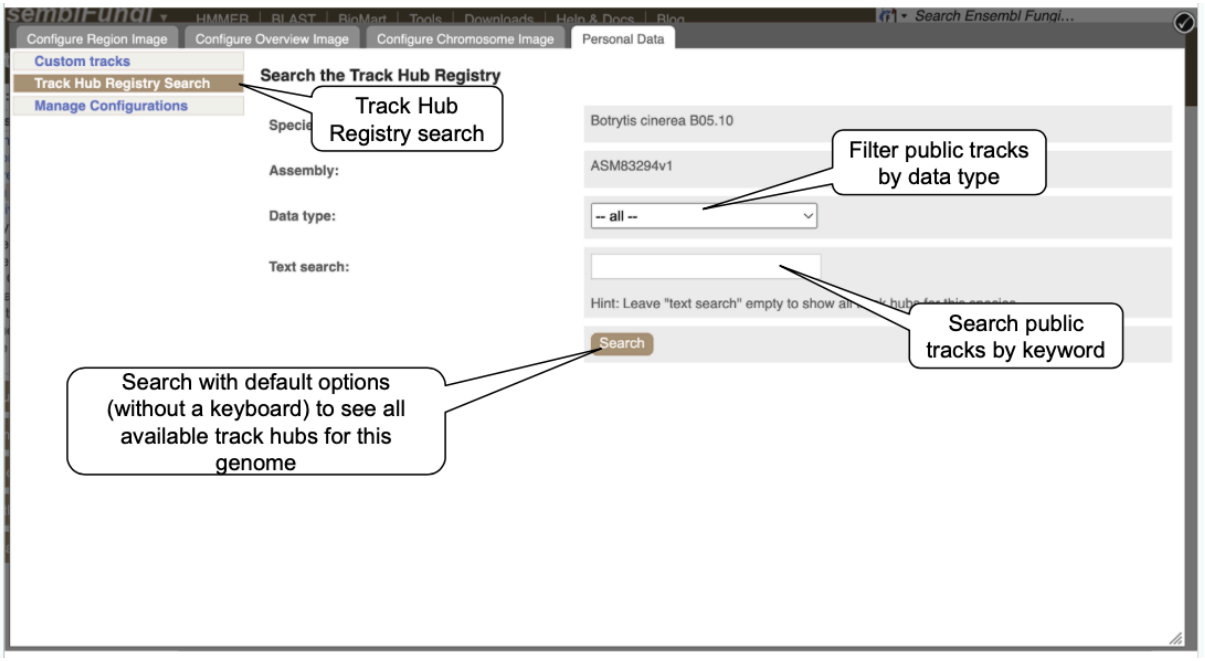
Links to be clicked shown in blue, text to be entered shown in red.

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.

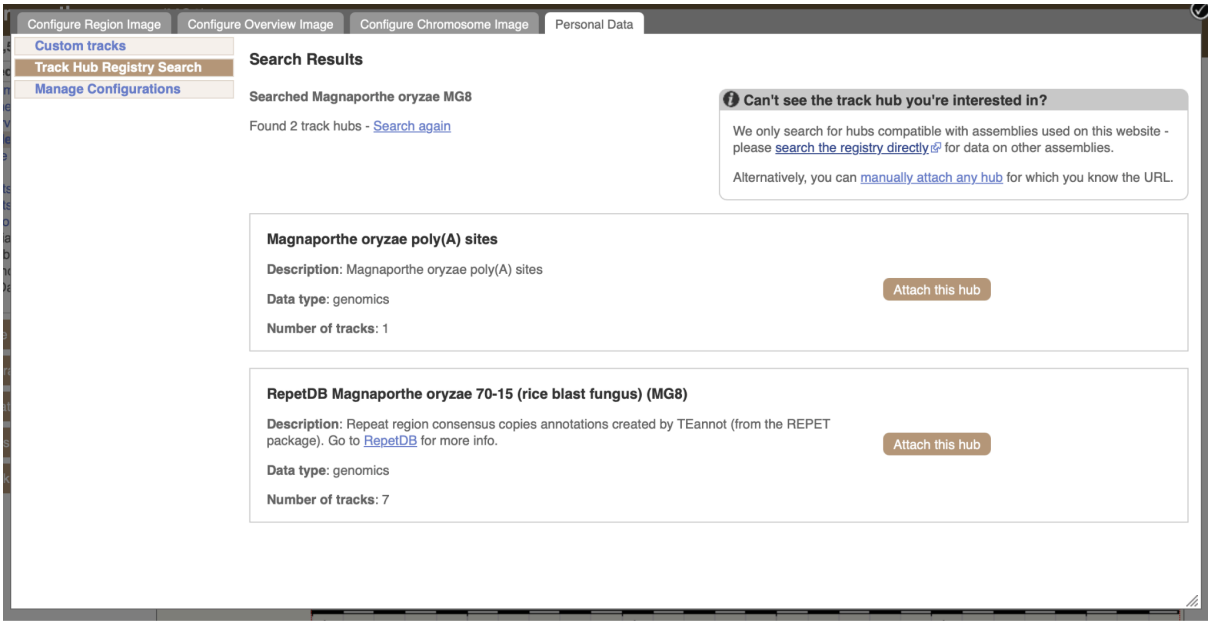
Go to fungi.ensembl.org on your browser and click the favourite species: *Magnaporthe oryzae (MG8)* then search for the region 1:5529955-5557959



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.



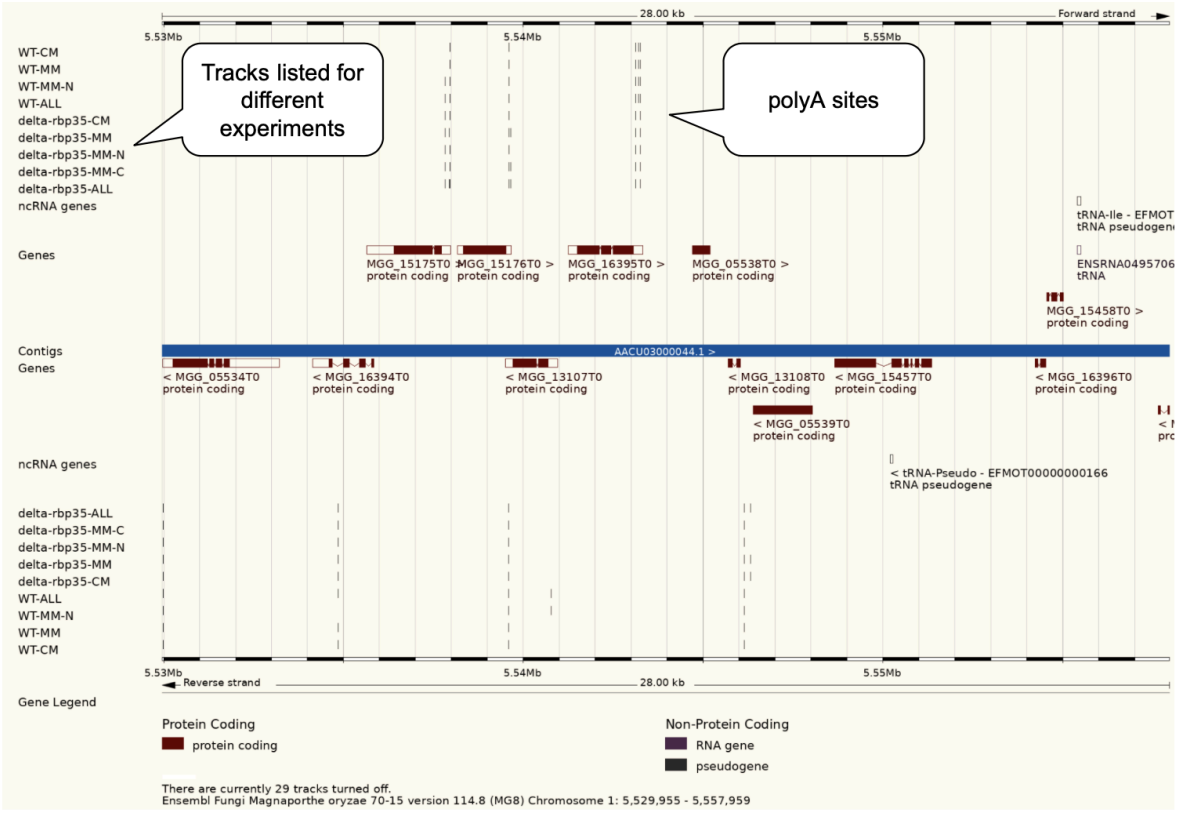
Just click Search with no options selected.



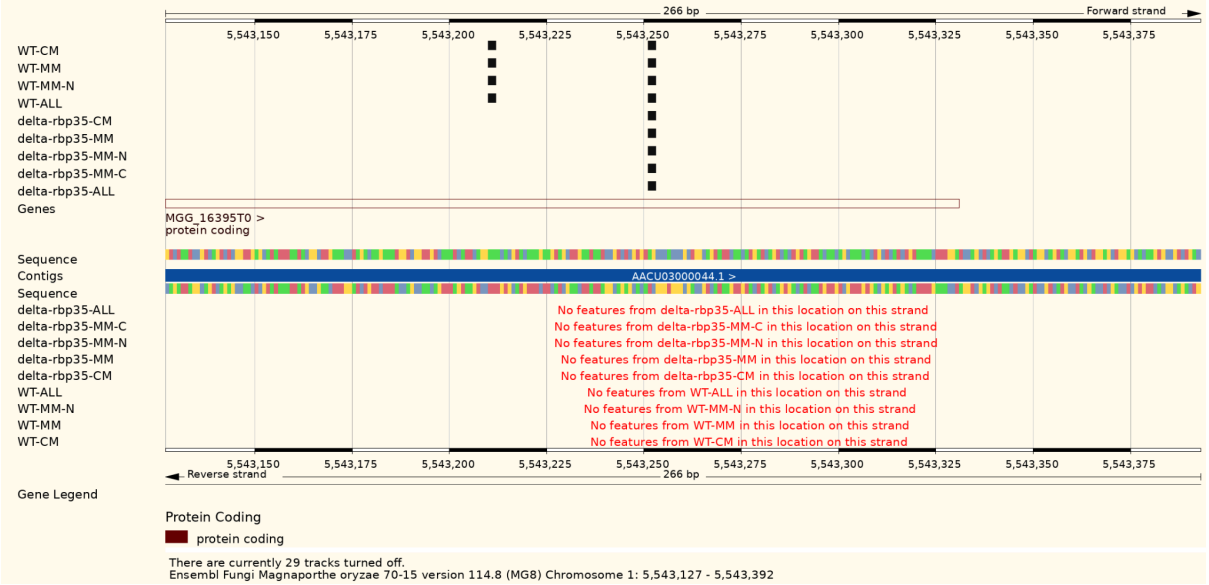
There are two available track hubs for this assembly, one of which is currently unavailable (e! genomes 61)

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

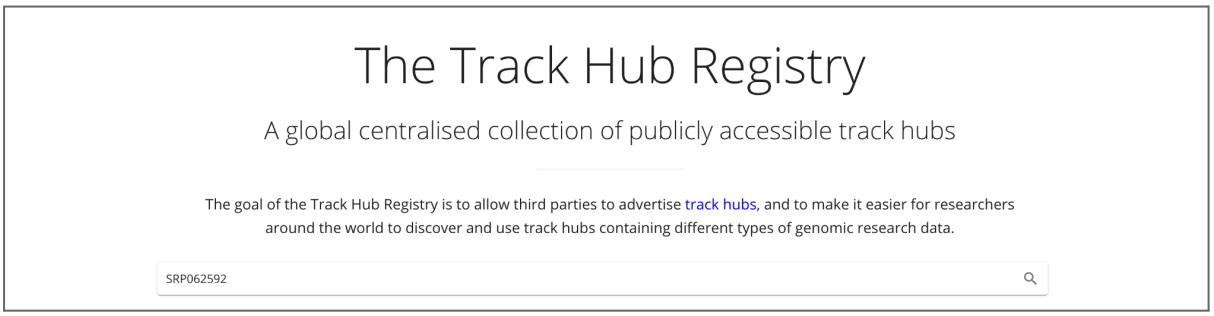
The track hub should now load and appear on the most-detailed image at the bottom of the ‘Region in detail’ page.

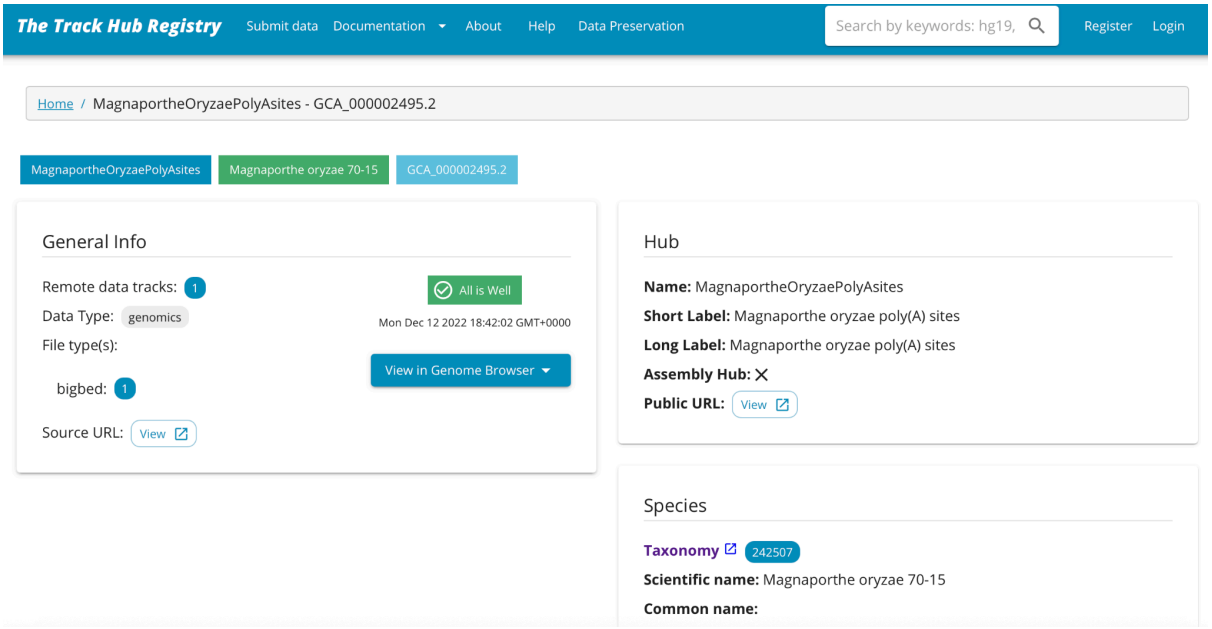


If you zoom in further, you can see a more detailed representation of the data:



(a) Go to 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)





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/

Let's take a look at that URL.

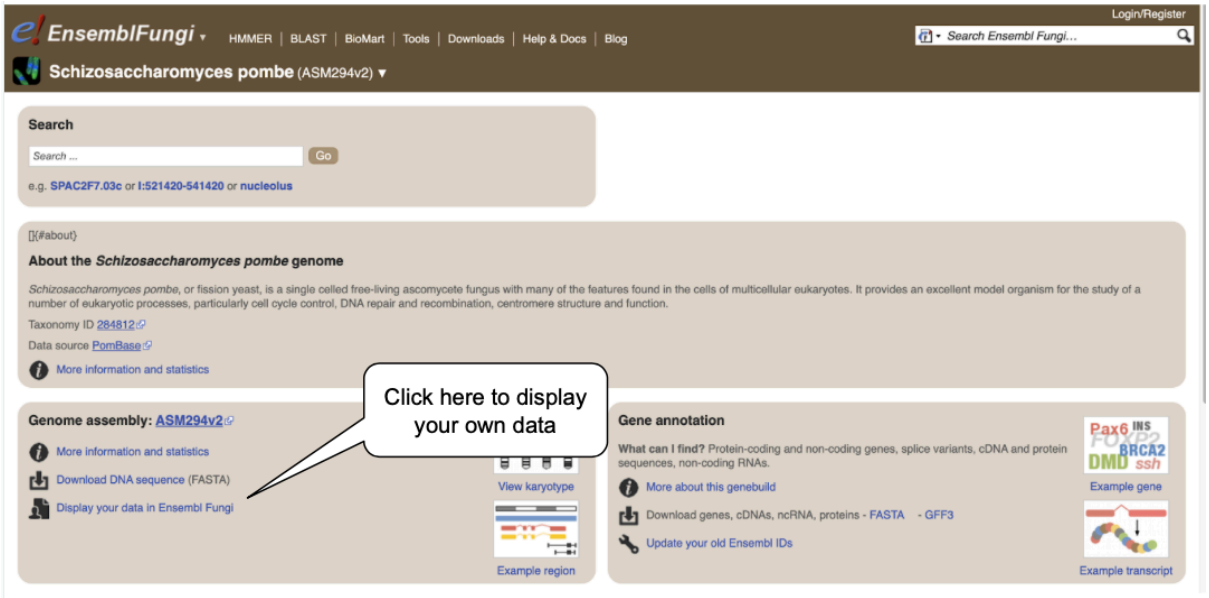
NOTE: Many internet browsers have recently dropped support for FTP, including the latest Firefox and Google Chrome versions. Firefox v87.0 still contains built-in FTP implementation. If you struggle to open the FTP site, try the HTTP version: https://ftp.ebi.ac.uk/ensemblgenomes/pub/misc_data/bam/fungi/Spom/

Index of /ensemblgenomes/pub/misc_data/bam/fungi/Spom

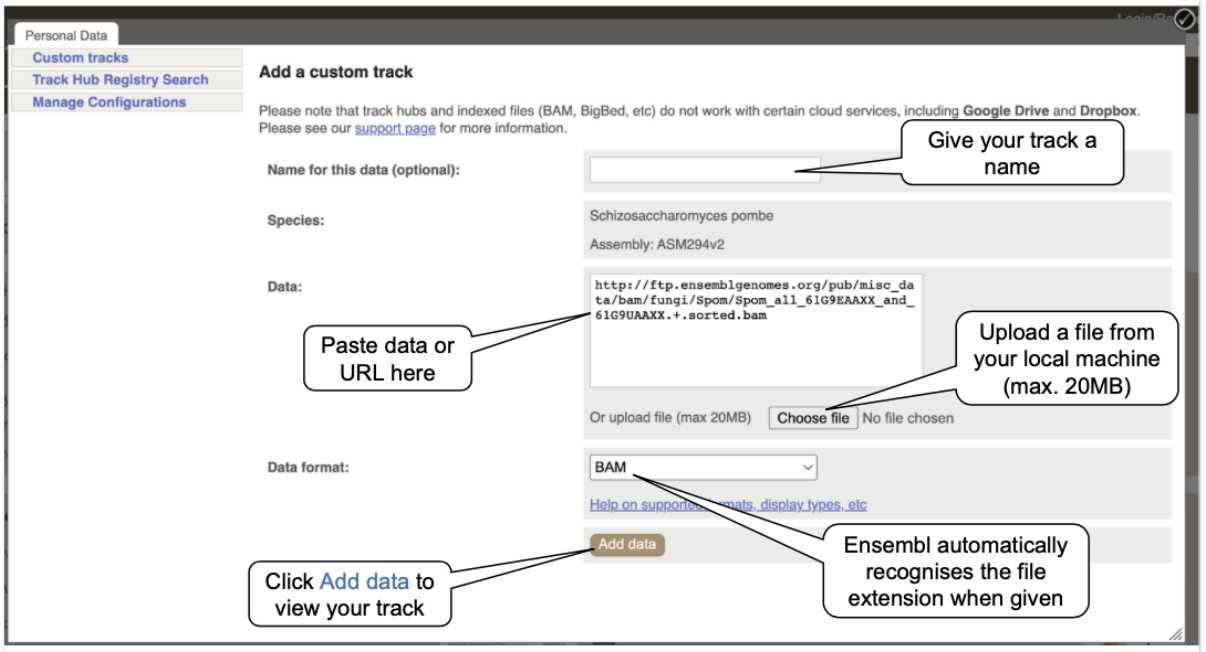
Name	Last modified	Size	Description
Parent Directory		-	
Spom_all_61G9EAAXX_and_61G9UAAXX.+.sorted.bam	2014-11-26 15:06	3.3G	
Spom_all_61G9EAAXX_and_61G9UAAXX.+.sorted.bam.bai	2014-11-26 15:06	36K	
Spom_all_61G9EAAXX_and_61G9UAAXX.-.sorted.bam	2014-11-26 15:04	3.8G	
Spom_all_61G9EAAXX_and_61G9UAAXX.-.sorted.bam.bai	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.



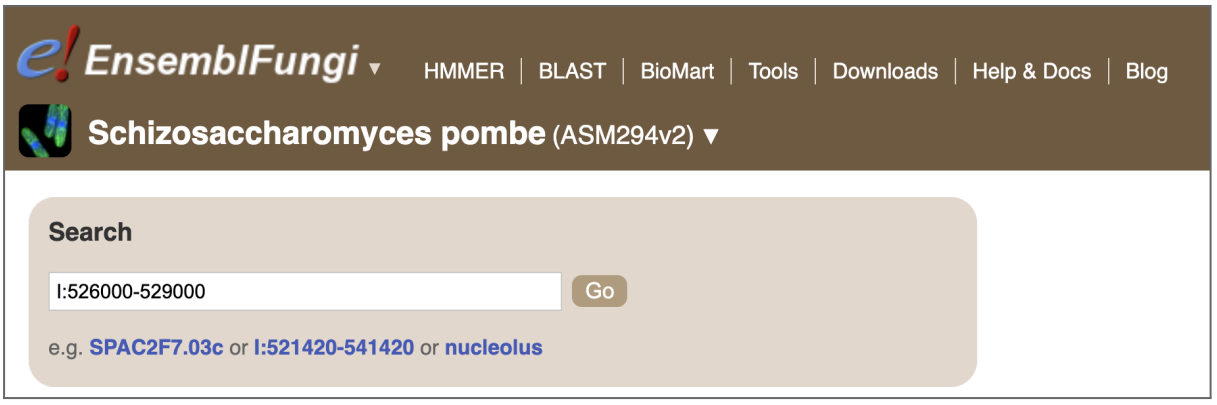
A menu will appear:

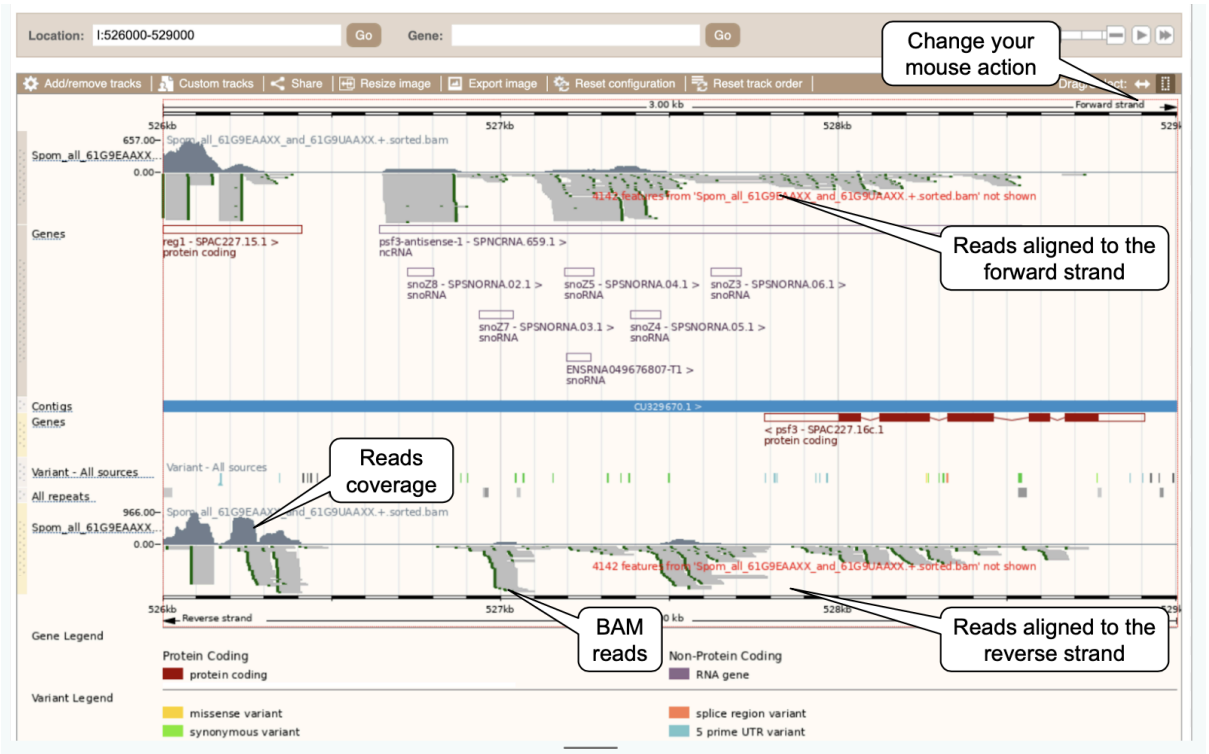


The interface detects file extensions if you upload or attach a file. If you want to upload a file, just click on Choose file, select the file from your local machine and it should automatically detect the file type you have submitted.

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

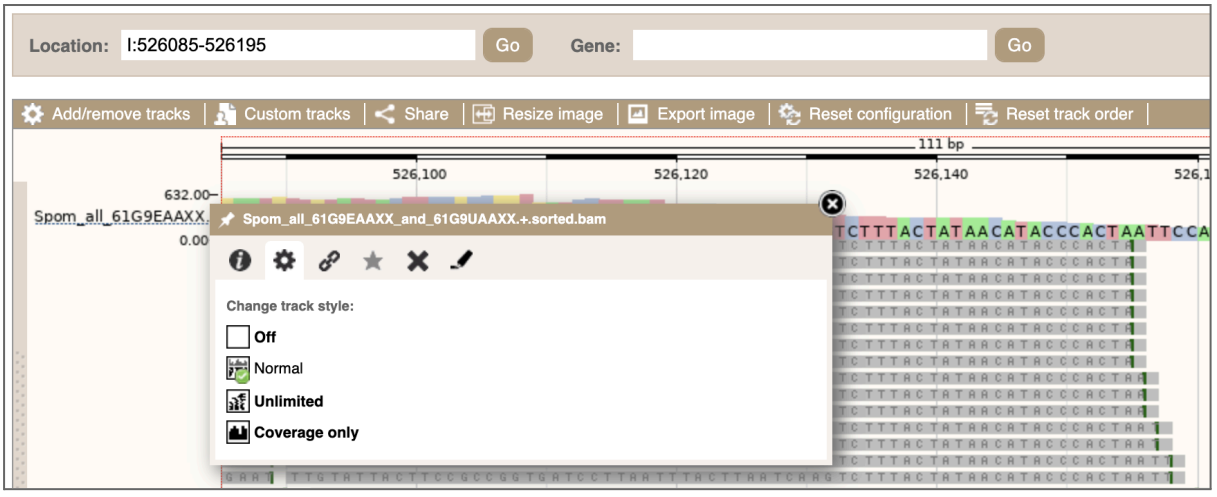
Since this is a file, the interface is able to detect the '.bam' file extension and automatically labels the format as BAM. Click on Add data and close the menu. It may take a while to load as there is a lot of data (Firefox tends to be fast). Once the data has been uploaded, you'll get a thank you message.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

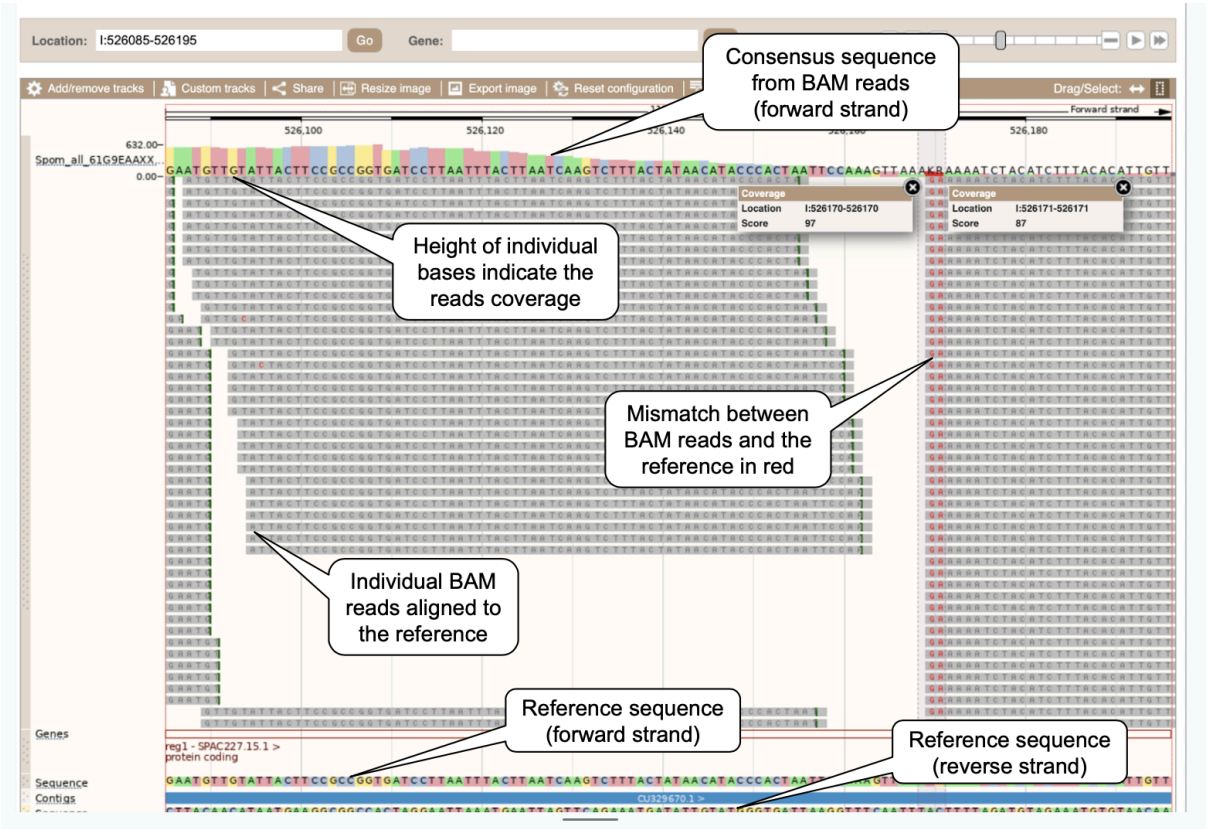




Newly added BAM file track split into forward and reverse stranded reads. You can zoom in to see the sequence itself. Drag out boxes in the view to zoom in, until you see a sequence of individual reads, or jump to a 110 bp region: I:526085-526195.

(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?




Location: I:526085-526195
Go
Gene:
Add/remove tracks
Custom tracks
Share
Resize image
Export image
Reset configuration
Drag/Select:
Forward strand
526,100
526,120
526,140
526,180
632.00
0.00
Spom_all_61G9EAAXX...
Coverage
Location I:526170-526170
Score 97
Coverage
Location I:526171-526171
Score 87
Consensus sequence from BAM reads (forward strand)
Height of individual bases indicate the reads coverage
Mismatch between BAM reads and the reference in red
Individual BAM reads aligned to the reference
Reference sequence (forward strand)
Reference sequence (reverse strand)
Genes
reg1 - SPAC227.15.1 >
protein coding
Sequence
Contigs

Advanced Search Strategies.

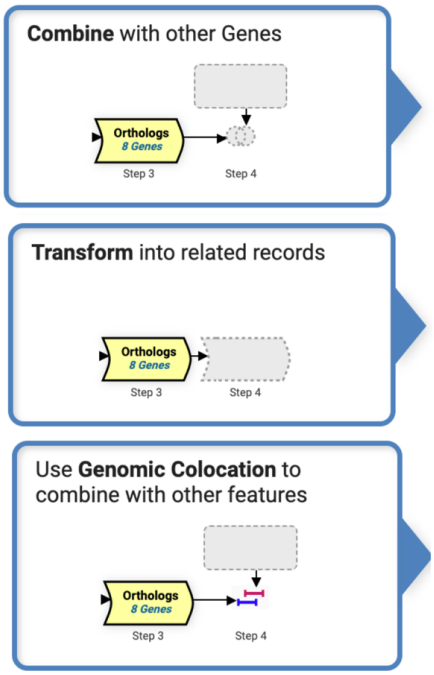
Learning objectives

- 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:

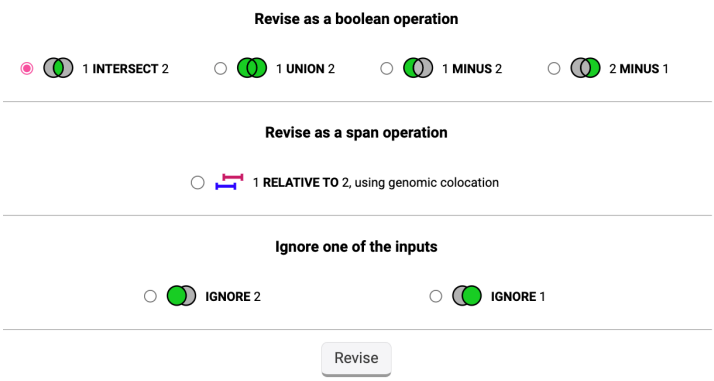


Combine with other Genes
Compare results that are gene lists

Transform into related records
Transform results into orthologs (e.g. Aspergillus > Candida), metabolic pathways or compounds.

Use Genomic Colocation
Combine different data types (e.g., cross-reference gene lists with metabolic pathway IDs).

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.

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:

- **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 in 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:

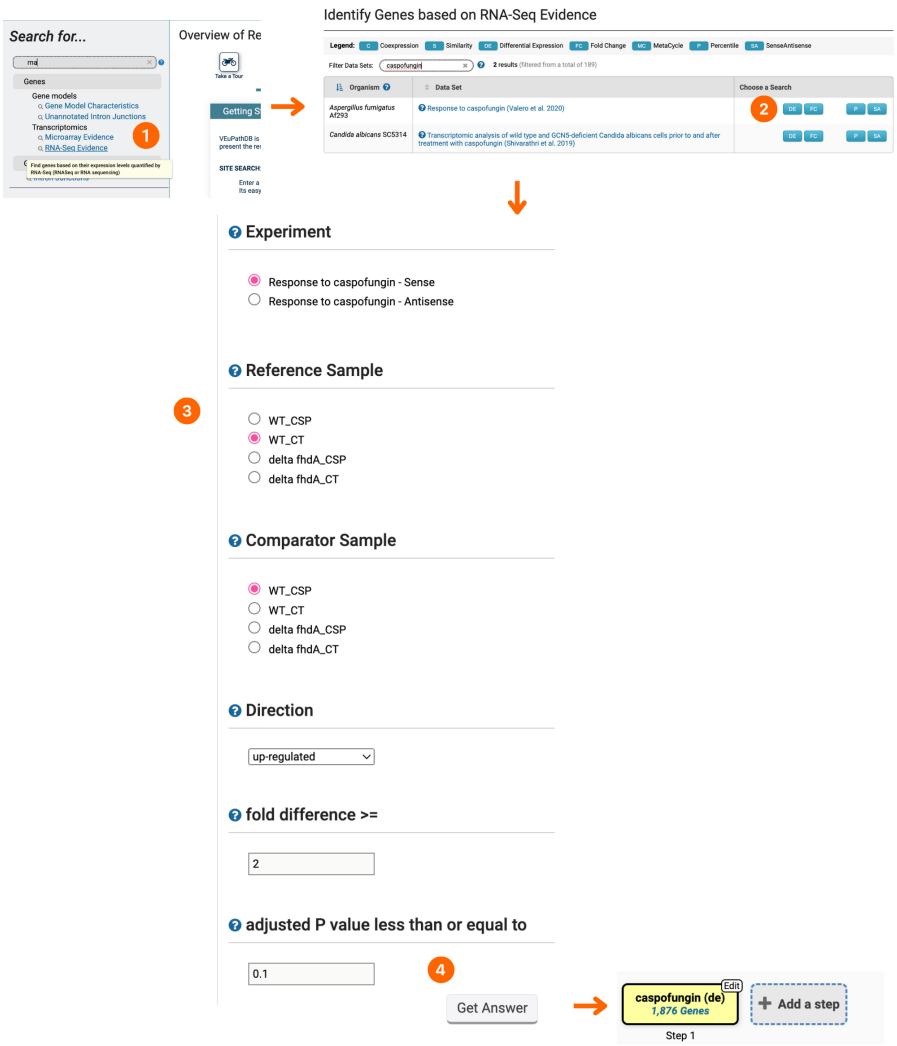
A. Are up-regulated when *Aspergillus* is exposed to human airway epithelial cells,

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.

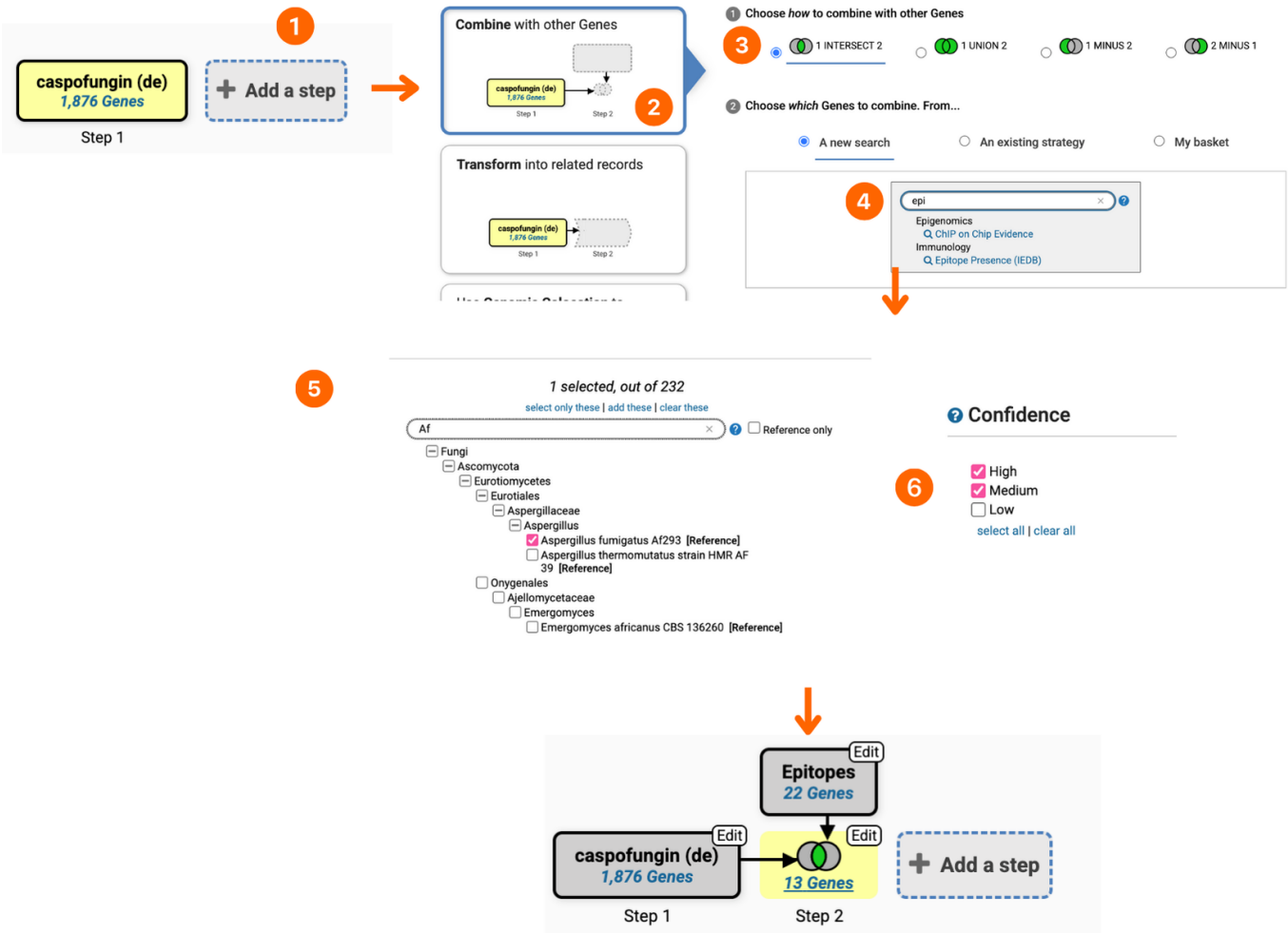
1. Select the search from the "Search for…" panel (shown below) or the "Searches" menu at the top of the page.
2. Identify the "Response to caspofungin (Valero et al. 2020)" dataset and click on the DE (Differential expression) button.
   Tip: Use the filter box to quickly find relevant search results.
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.



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

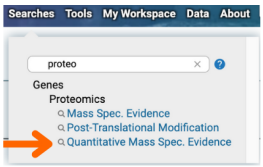
Advanced Search Strategies. Proteomics

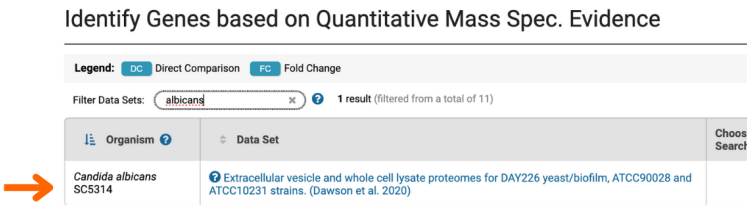
In this section, you will learn how to query quantitative mass spec data.

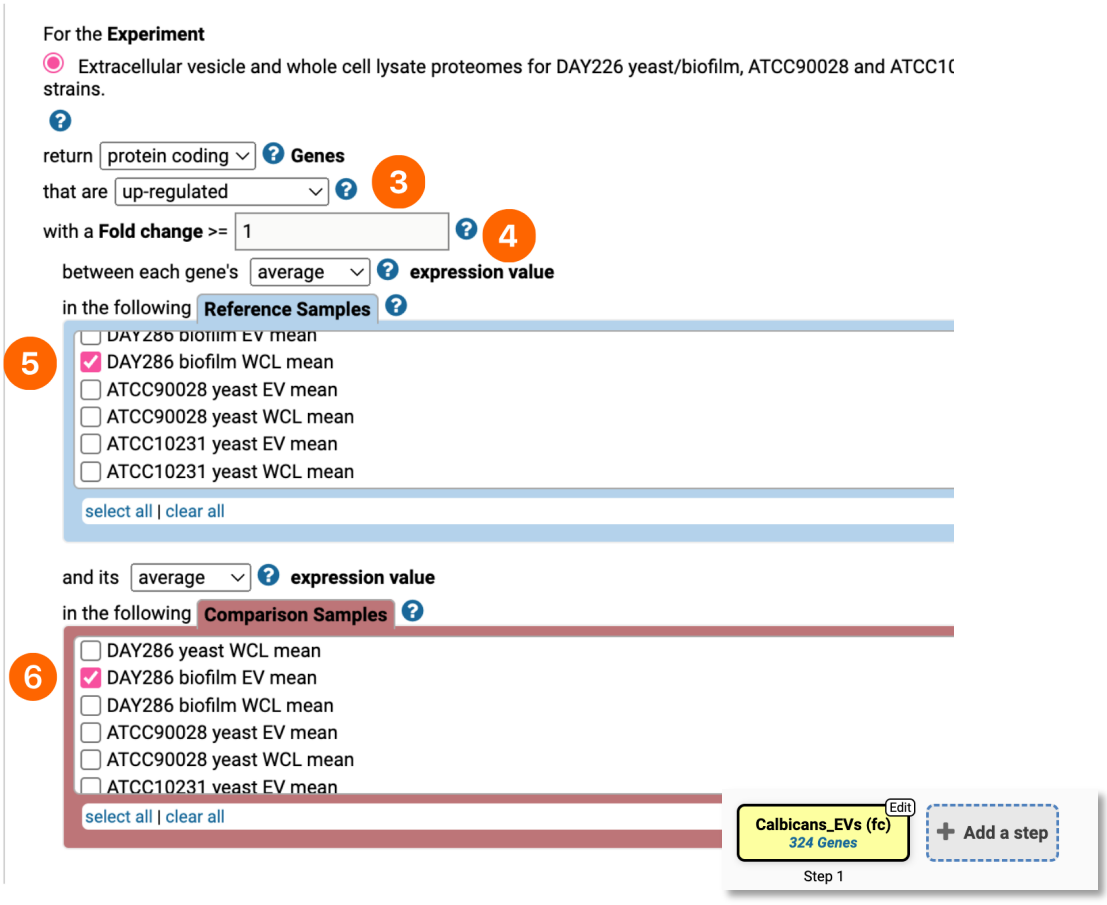
Fungal extracellular vesicles (EVs) have been implicated in host-pathogen and pathogen-pathogen communication. In this exercise, we will find proteins abundant in EVs when compared to whole cell lysate (WCL) in biofilm conditions.

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

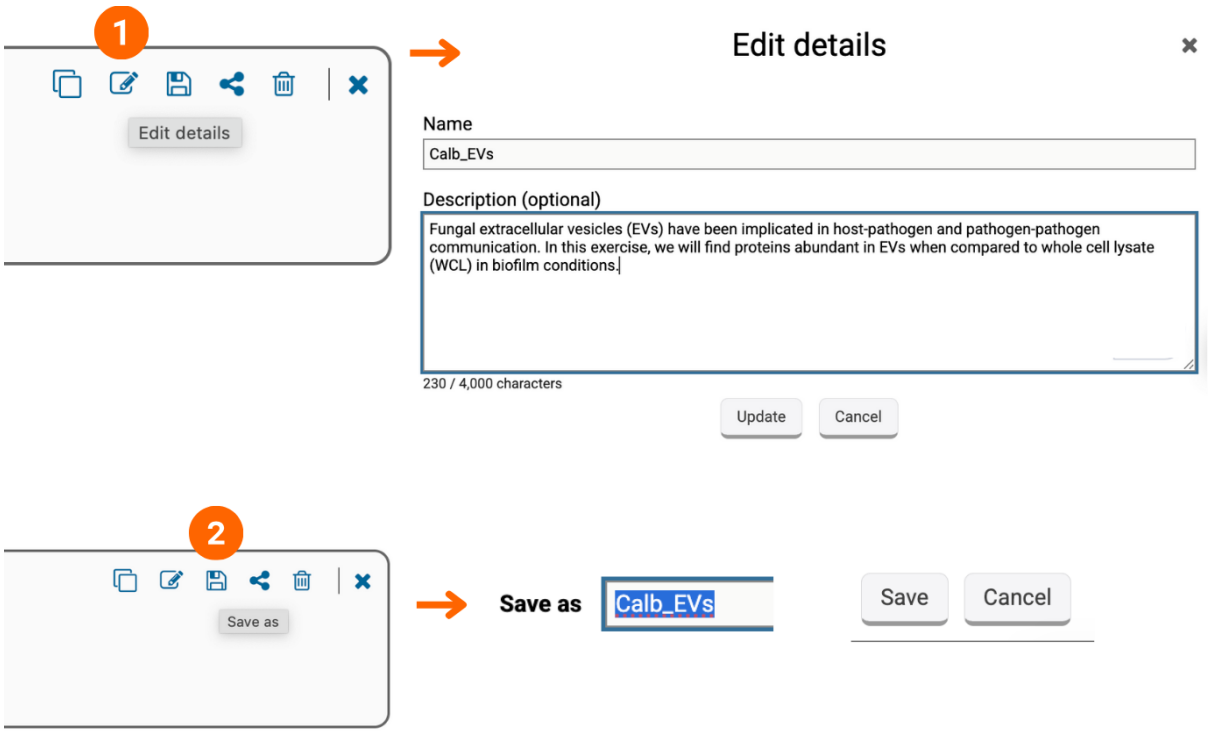
1. Navigate to the “Quantitative Mass Spec. Evidence” search.
2. Filter for “albicans” and click the “FC” button for the Dawson et al. 2020 dataset.
3. Look for up-regulated genes.
4. With a Fold change >= 1.
5. Set Reference strain to DAY286 biofilm WCL mean.
6. Set Comparison Sample to DAY286 biofilm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







- **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

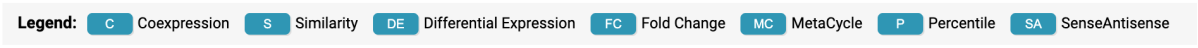
OPTIONAL
Exploring transcriptomics host-pathogen datasets in VEuPathDB

HostDB.org is a VEuPathDB knowledgebase providing pathogen host informatics resources.
Learning objectives:

- Query host-pathogen RNA-Seq data.
- Create a proteomics query and save this strategy to your account.

I. Transcriptomics.

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



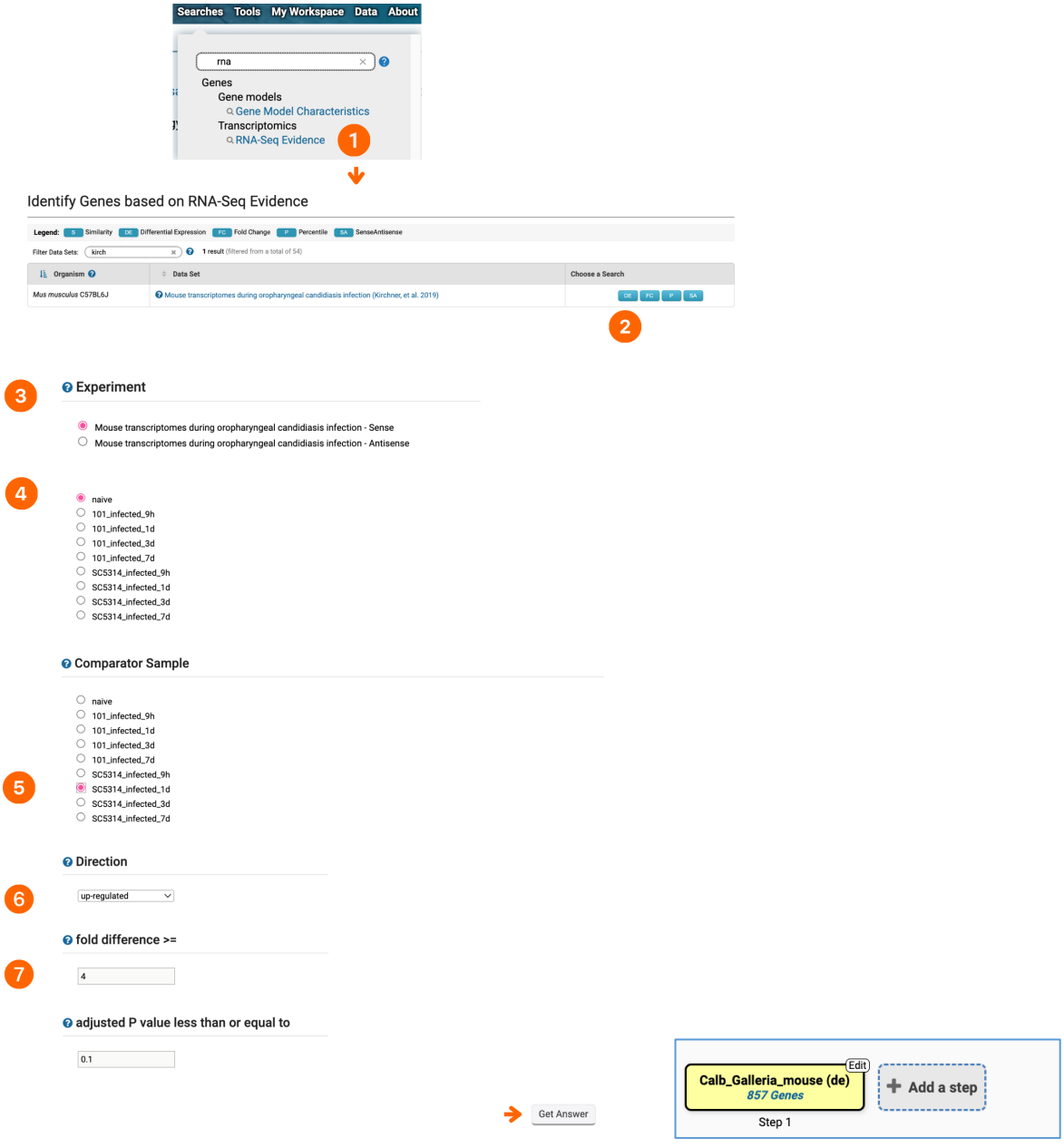
- **Coexpression.** Search for genes which have positive or negative correlations with a set of genes.
- **Similarity.** Search for genes which have a similar profile for an experiment.
- **Differential Expression (DE).** This search uses DESeq2 analysis results. You can choose the directionality and 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 in 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.

For this exercise, we will query the host (mouse) and pathogen (*Candida albicans*) RNA-Seq data produced 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.

Objective: Identify differentially expressed genes in mice (HostDB.org) during infection (1d).

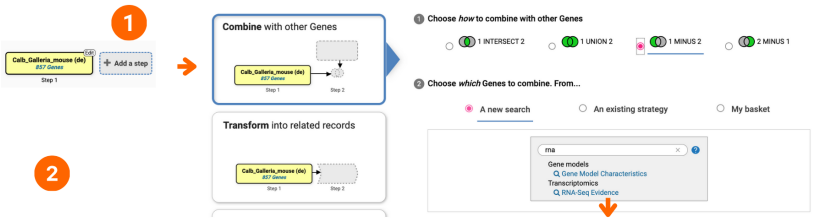
A. The next block of exercises will be carried out in HostDB.org

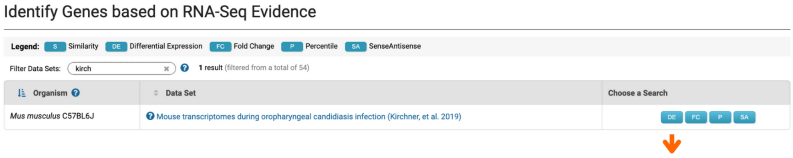
- **Identify genes up-regulated in mice infected with SC5314 at 1d**.
  1. Navigate to the RNA-Seq Evidence search and filter RNA-Seq datasets for "Kirch".
  2. Click on the "DE" button.
  3. Choose to examine the sense strand.
  4. Select reference sample: naïve.
  5. Select comparator sample: SC5314_infected_1d.
  6. Look for up-regulated genes.
  7. Select magnitude of upregulation: 4 fold.

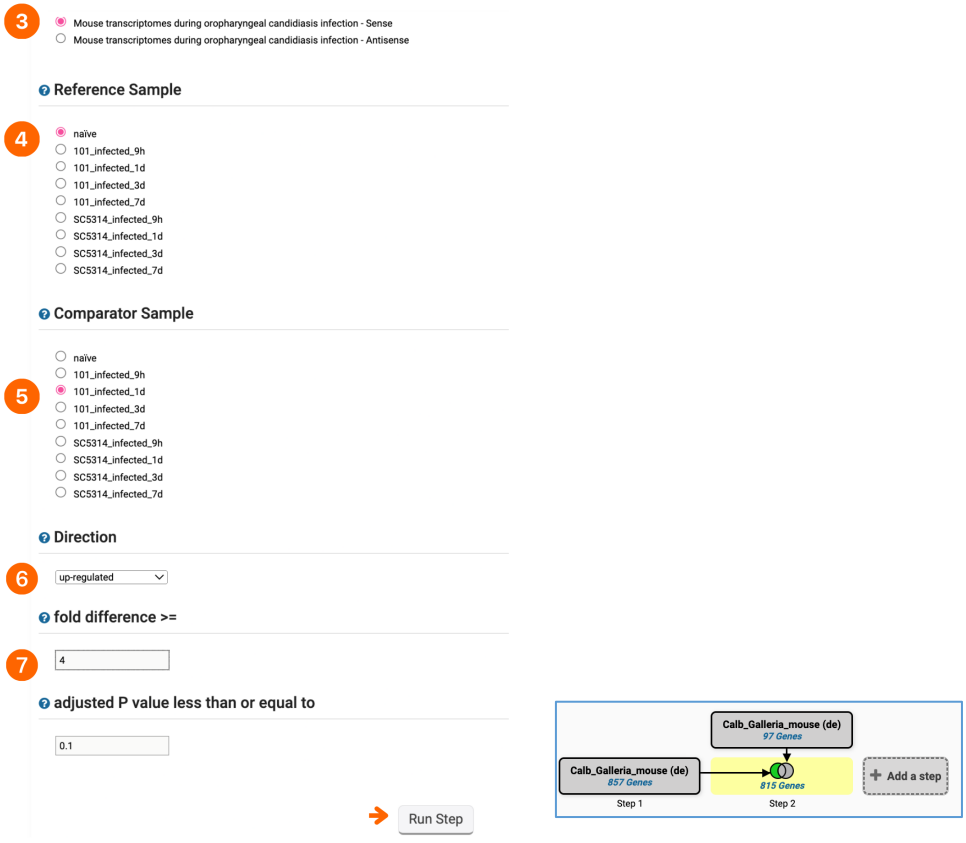


- **Identify host genes up-regulated by the SC5314 strain but not 101 at 1d of infection**.
    1. Click on the “Add Step” button.
    2. Navigate to the RNA-Seq Evidence search, select the “1 minus 2” Boolean operator, identify the RNA-Seq Evidence” search, filter for “Kirch” to quickly identify the dataset and click on the “DE” button.
    3. Choose to examine the sense strand.
    4. Select reference sample: naïve.
    5. Select comparator sample: 101_infected_1d.
    6. Look for up-regulated genes.
    7. Select magnitude of upregulation: 4 fold.

Note: The default Boolean operator is set to the “intersect” option. Make sure to select the correct Boolean operator for this search.

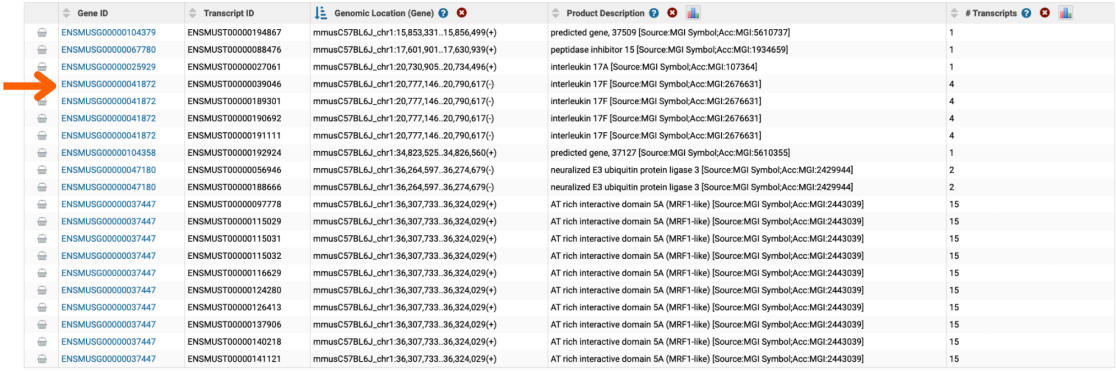






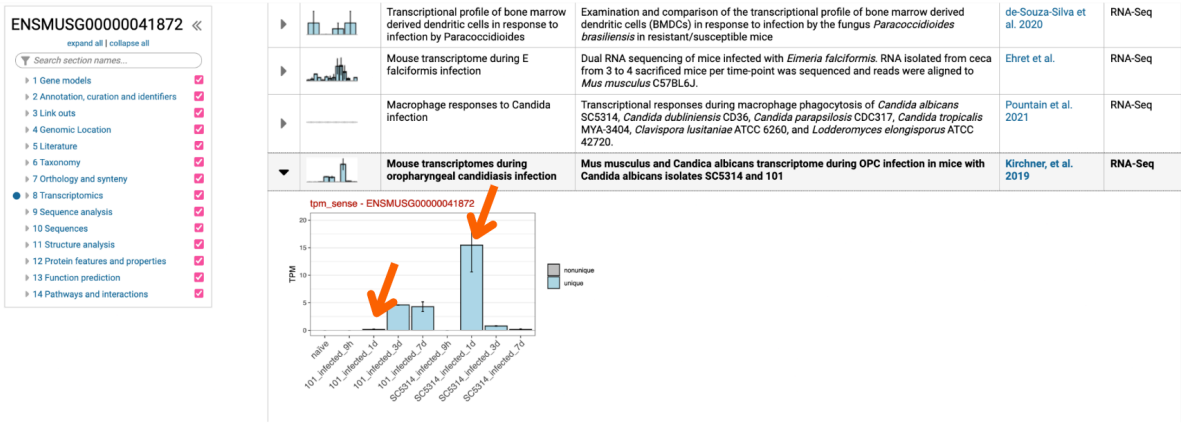
- **Examine the results in HostDB:**

    1. Click on the Gene ID link for “interleukin 17F” and navigate to the Transcript expression section within the gene record page.



Gene ID	Transcript ID	Genomic Location (Gene)	Product Description	# Transcripts
ENSMUSG00000104379	ENSMUST00000194867	mmusC57BL6J_chr1:15,853,331..15,856,499(+)	predicted gene, 37509 [Source:MGI Symbol;Acc:MGI:5610737]	1
ENSMUSG00000067780	ENSMUST00000088476	mmusC57BL6J_chr1:17,601,901..17,630,939(+)	peptidase inhibitor 15 [Source:MGI Symbol;Acc:MGI:1934659]	1
ENSMUSG00000025929	ENSMUST00000027061	mmusC57BL6J_chr1:20,730,905..20,734,496(+)	interleukin 17A [Source:MGI Symbol;Acc:MGI:107364]	1
ENSMUSG00000041872	ENSMUST00000039046	mmusC57BL6J_chr1:20,777,146..20,790,617(-)	interleukin 17F [Source:MGI Symbol;Acc:MGI:2676631]	4
ENSMUSG00000041872	ENSMUST00000189301	mmusC57BL6J_chr1:20,777,146..20,790,617(-)	interleukin 17F [Source:MGI Symbol;Acc:MGI:2676631]	4
ENSMUSG00000041872	ENSMUST00000190692	mmusC57BL6J_chr1:20,777,146..20,790,617(-)	interleukin 17F [Source:MGI Symbol;Acc:MGI:2676631]	4
ENSMUSG00000041872	ENSMUST00000191111	mmusC57BL6J_chr1:20,777,146..20,790,617(-)	interleukin 17F [Source:MGI Symbol;Acc:MGI:2676631]	4
ENSMUSG00000104358	ENSMUST00000192924	mmusC57BL6J_chr1:34,823,525..34,826,560(+)	predicted gene, 37127 [Source:MGI Symbol;Acc:MGI:5610355]	1
ENSMUSG00000047180	ENSMUST00000056946	mmusC57BL6J_chr1:36,264,597..36,274,679(-)	neuralized E3 ubiquitin protein ligase 3 [Source:MGI Symbol;Acc:MGI:2429944]	2
ENSMUSG00000047180	ENSMUST00000188666	mmusC57BL6J_chr1:36,264,597..36,274,679(-)	neuralized E3 ubiquitin protein ligase 3 [Source:MGI Symbol;Acc:MGI:2429944]	2
ENSMUSG00000037447	ENSMUST00000097778	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG00000037447	ENSMUST00000115029	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG00000037447	ENSMUST00000115031	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG00000037447	ENSMUST00000115032	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG00000037447	ENSMUST00000116629	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG00000037447	ENSMUST00000124280	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG00000037447	ENSMUST00000126413	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG00000037447	ENSMUST00000137906	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG00000037447	ENSMUST00000140218	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG00000037447	ENSMUST00000141121	mmusC57BL6J_chr1:36,307,733..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15

Notice that the interleukin 17F response is much stronger at 1d in response to SC5314 infection. This is consistent with the delayed mouse response to C. albicans strain 101 compared to strain SC5314. Now, you may want to look back at gene enrichment signatures in fungi to learn more about SC5314 and 101-driven responses.

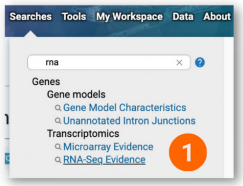


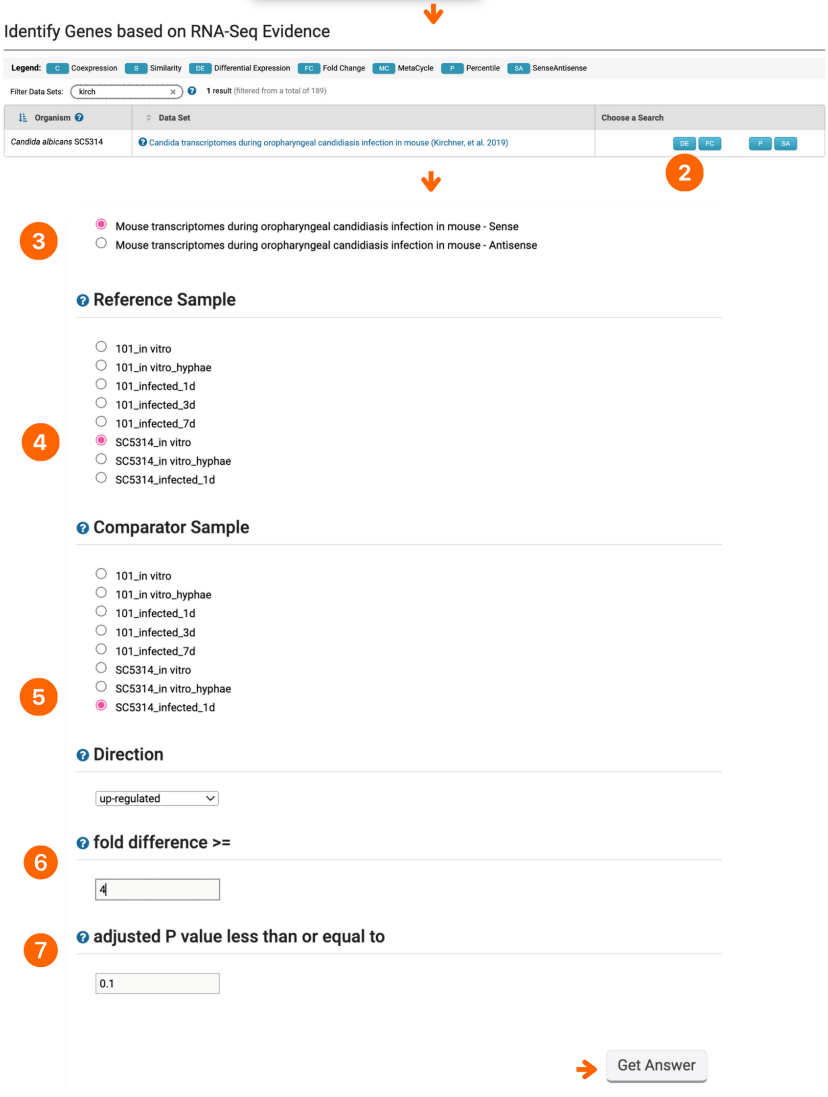
In summary, this strategy compared differentially expressed genes in mice in response to infection with SC5314 and 101 strains. It also identified genes up-regulated in response to SC5314 at 1d of infection while subtracting common genes upregulated in response to the exposure to the 101 strain.

Strategy URL: https://hostdb.org/hostdb/app/workspace/strategies/import/81476b9e123e341b

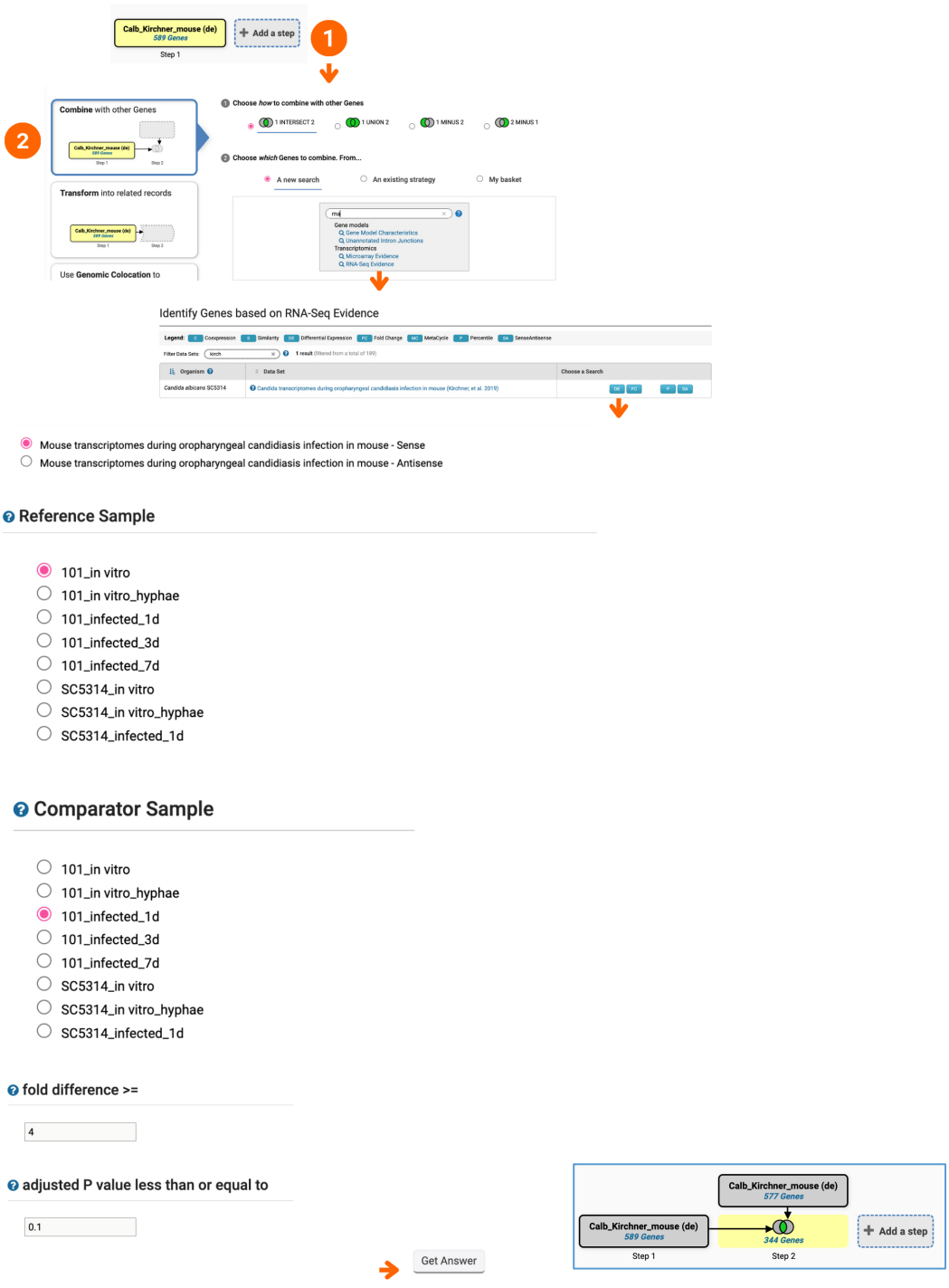
B. The next block of exercises will be carried out in FungiDB.org.

- **Identify genes up-regulated in SC5314 at 1d of infection**.
    1. Navigate to the RNA-Seq Evidence search and filter RNA-Seq datasets for "Kirch".
    2. Click on the "DE" button.
    3. Choose to examine the sense strand.
    4. Select reference sample: SC5314_in vitro.
    5. Select comparator sample: SC5314_infected_1d.
    6. Look for up-regulated genes.
    7. Select magnitude of upregulation: 4 fold.

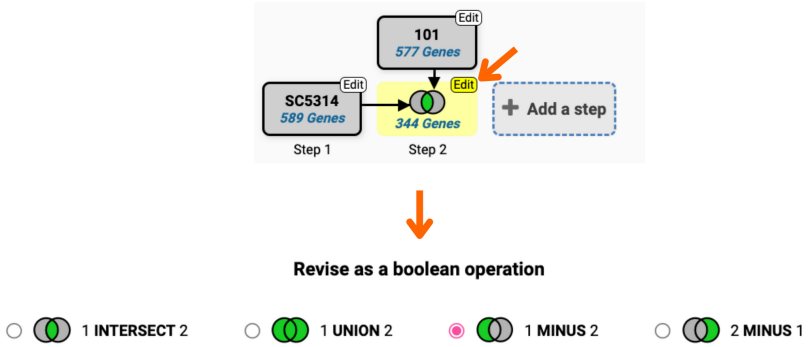




- **Identify genes up-regulated in SC5314 but not 101 strain at 1d of infection**.
    1. Click on the “Add Step” button.
    2. Navigate to the RNA-Seq Evidence search, filter for “Kirch” to quickly identify the dataset, and click the “DE” button.
    3. Choose to examine the sense strand.
    4. Select reference sample: 101_in vitro.
    5. Select comparator sample: 101_infected_1d.
    6. Look for up-regulated genes.
    7. Select magnitude of upregulation: 4 fold.

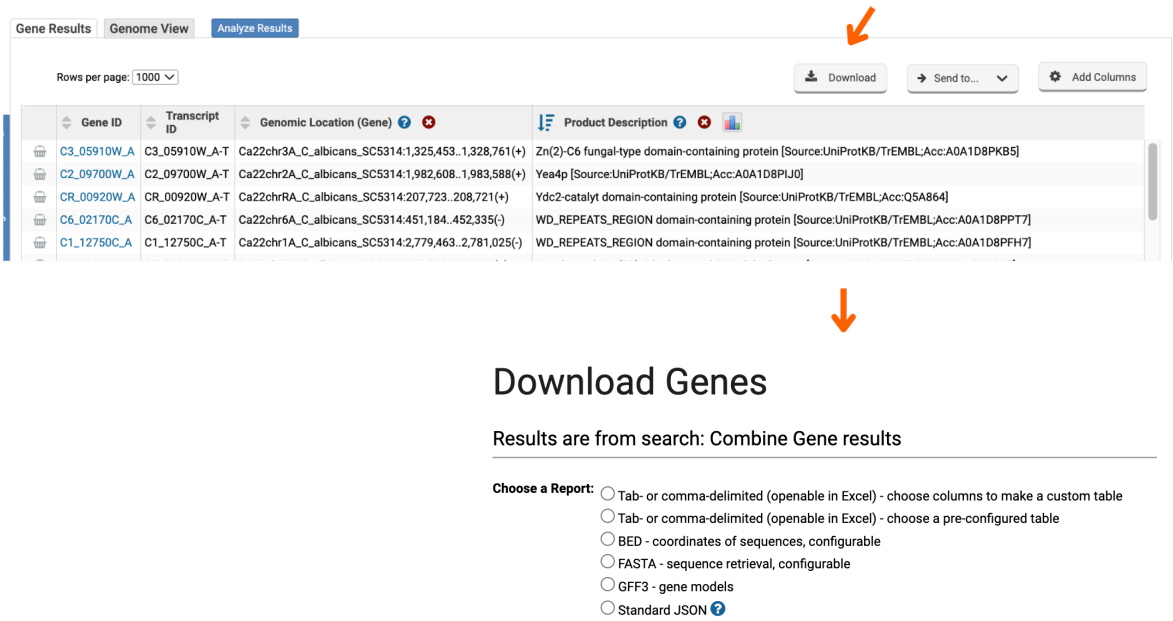


Note: You can always modify the Boolean operator by clicking on the Edit function as shown below:

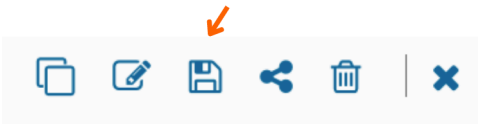


In summary, this strategy compared differentially expressed genes in SC5314 and 101 strains. It also identified genes up-regulated in SC5314 at 1d of infection while subtracting common upregulated genes in the 101 strain background.

Note: The results of this analysis can be exported. FungiDB offers several download options, including viewing them within the browser or exporting them locally to your computer.



You can save the strategy by clicking on the floppy disk icon on the right. We will return to this strategy in the module on GO Enrichment analysis.



Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/802d9f2b606fc1fa

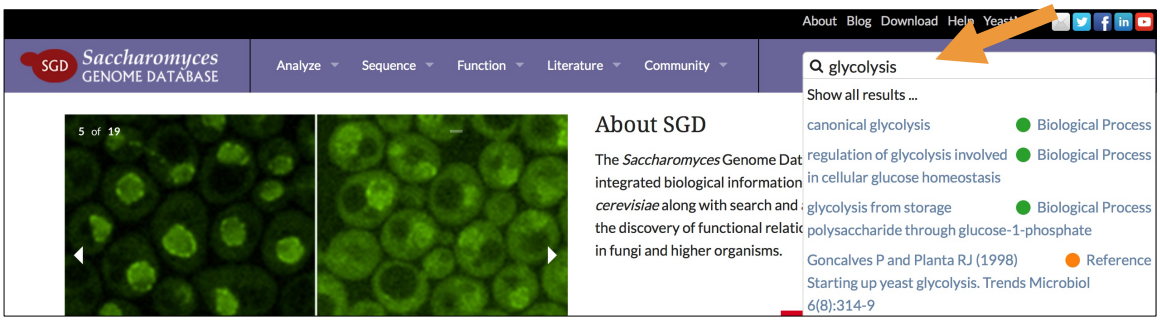
Using SPELL to Analyze Expression Datasets & Coexpressed Genes at SGD

SPELL (**S**erial **P**attern of **E**xpression **L**evels **L**ocator) 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.

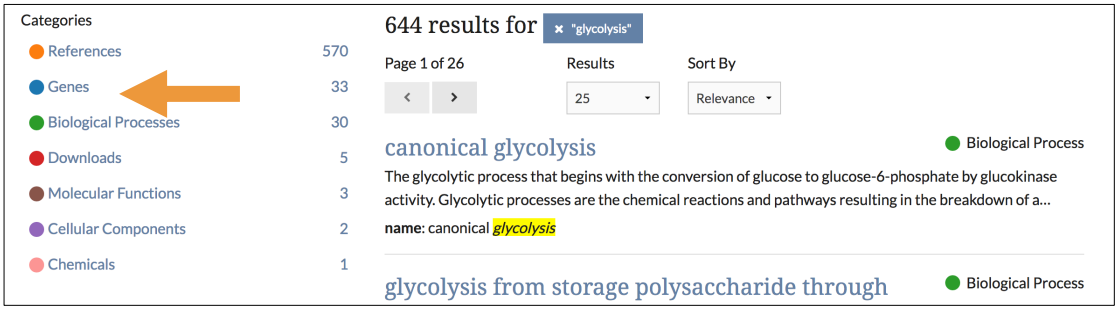
Use SPELL to find out which genes are coexpressed with genes involved in glycolysis.

Compile a list of genes involved in glycolysis.

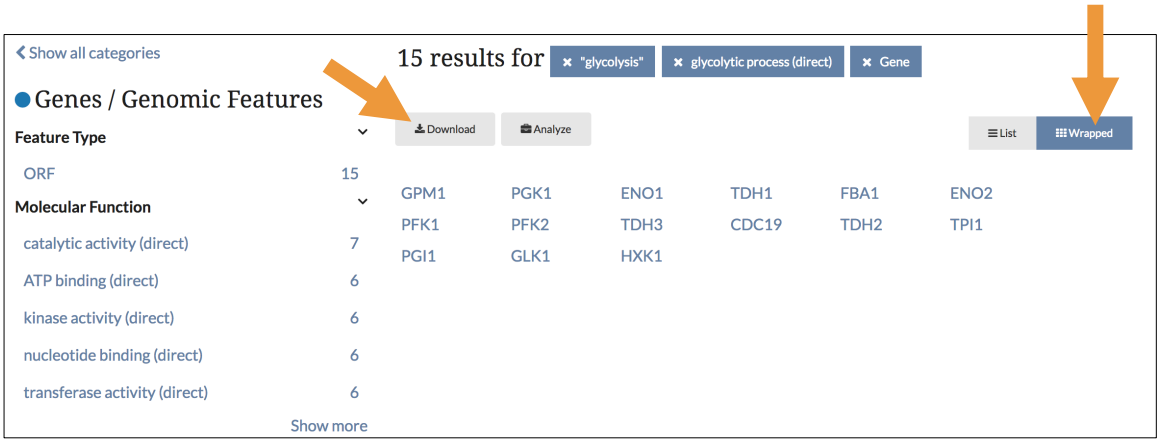
- On the SGD home page (www.yeastgenome.org), enter glycolysis into the search box and hit Enter.



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



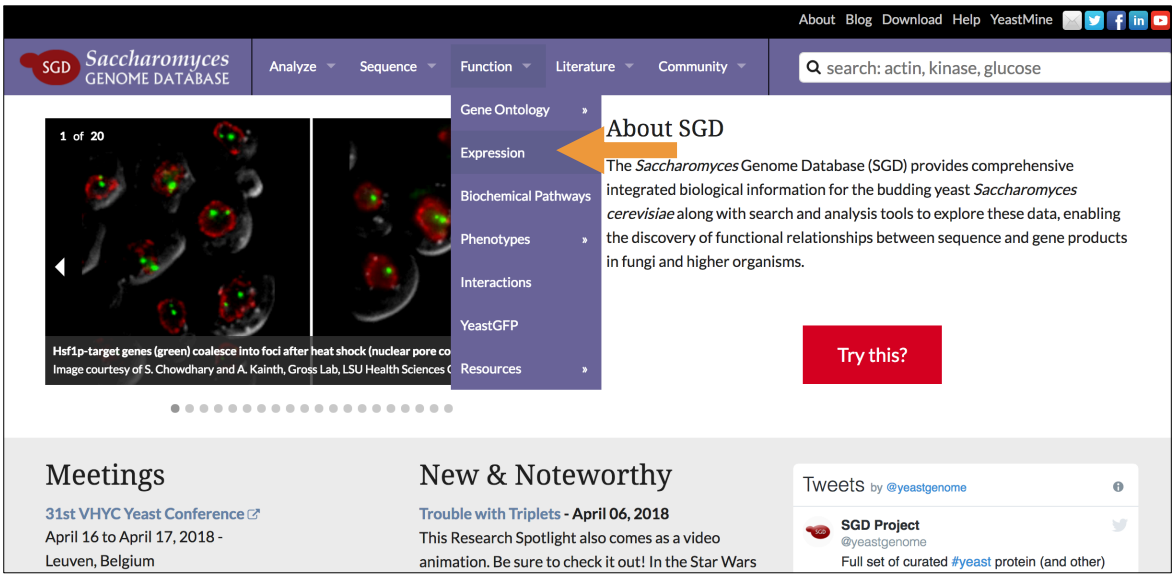
- Scroll down the page and find the **Biological Process** category on the left hand menu. Hit Show more and select **glycolytic process (direct).**
- To download the list of genes, click on **Wrapped** and then on **Download**.



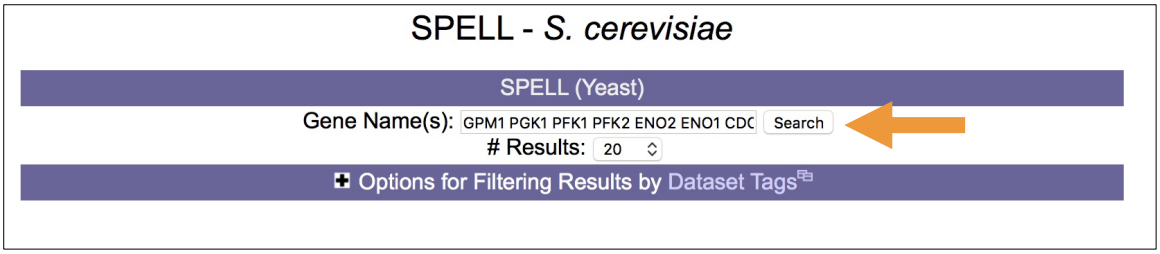
- The **Analyze** button, directly to the right of Download, enables you to import your search results directly into SPELL (among other tools at SGD). However, for the sake of demonstration, in this exercise we are instead going to enter our gene list into SPELL manually.

Import your gene list into SPELL and run a query:

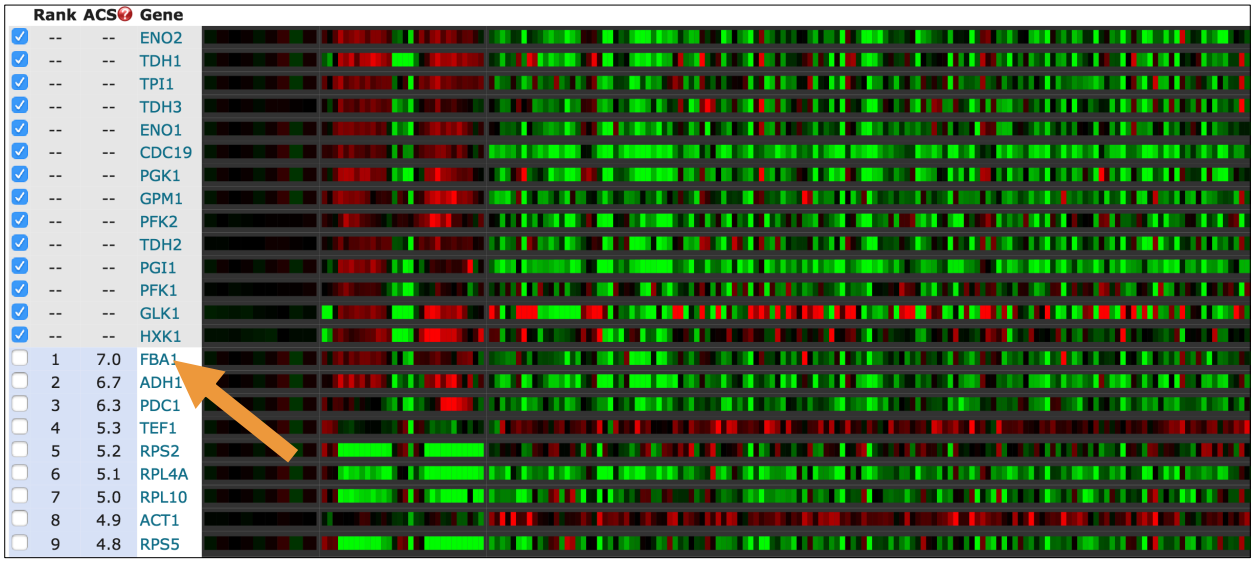
- To access SPELL, go to the SGD home page at www.yeastgenome.org, open the **Function** tab on top of the page and click on **Expression**. Or, if you are already on a Locus Summary page, open the Expression tab and click on the SPELL link under the histogram.



- 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.



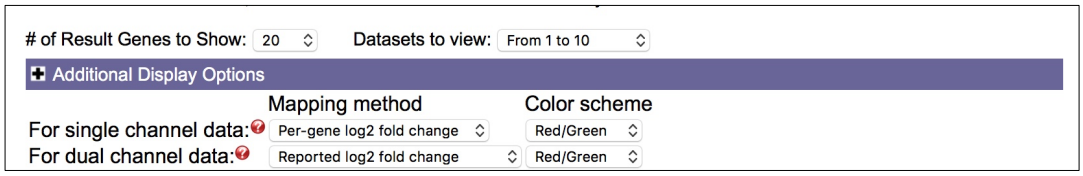
- Scroll down the list of genes on the left. Genes with checked boxes are from our query; the remaining genes are "hits", ordered from top to bottom according to their ranks. The rank reflects the correlation of expression of that gene with the query gene(s), given the relevance weight of that expression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



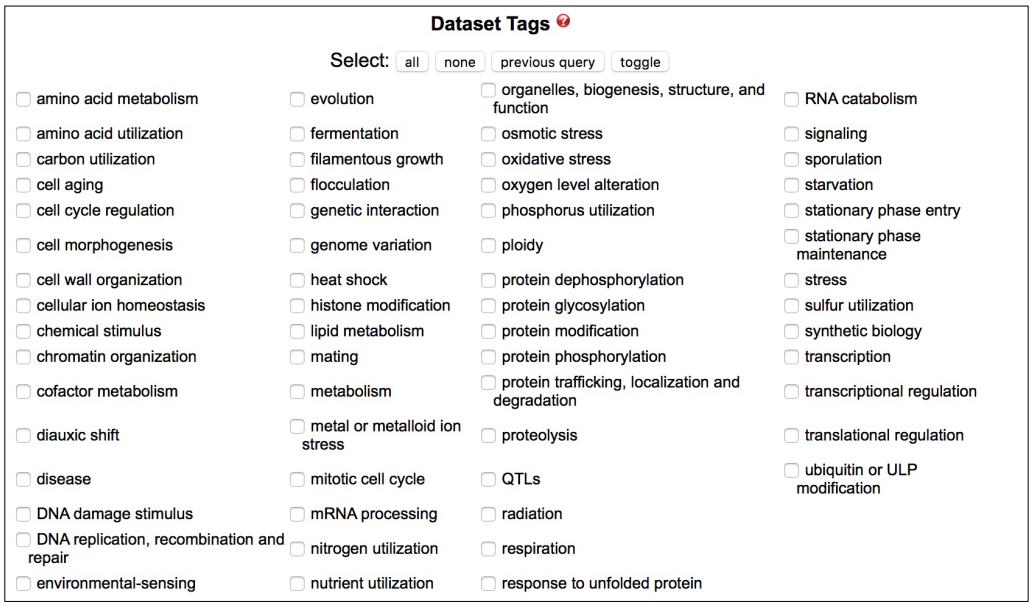
- Notice that the glycolysis gene we deleted earlier, FBA1, is indeed the highest-ranking gene!

- 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?

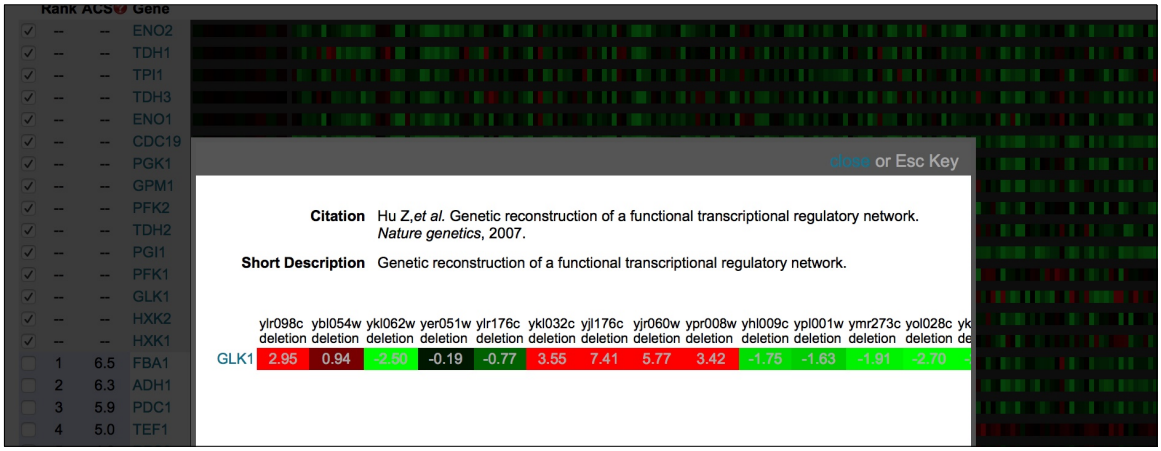
- Click on + **Additional Display Options** to change the default mapping method and color scheme to blue/yellow. Directly above this section are options to change the number of genes and datasets shown in your results.



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



- 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

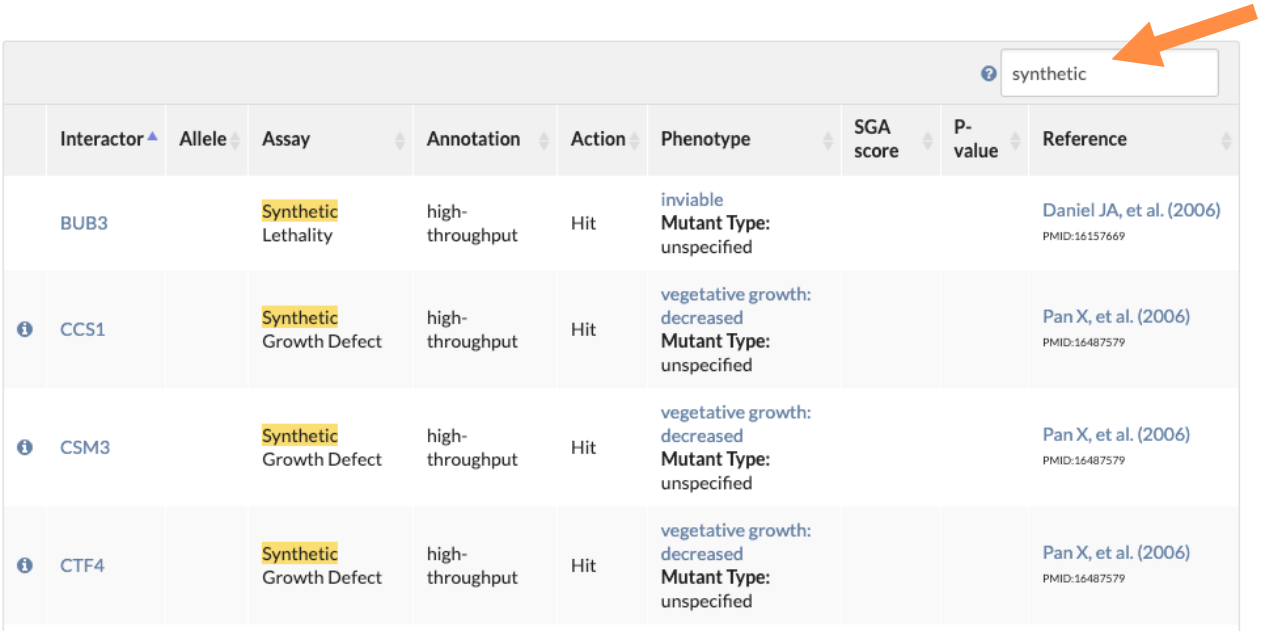
GOTerm	P-val	% query	% genome	Annotated Genes
glucose catabolic process (biological_process)	1.33e-29	19 of 35	52 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2
hexose catabolic process (biological_process)	2.39e-28	19 of 35	59 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

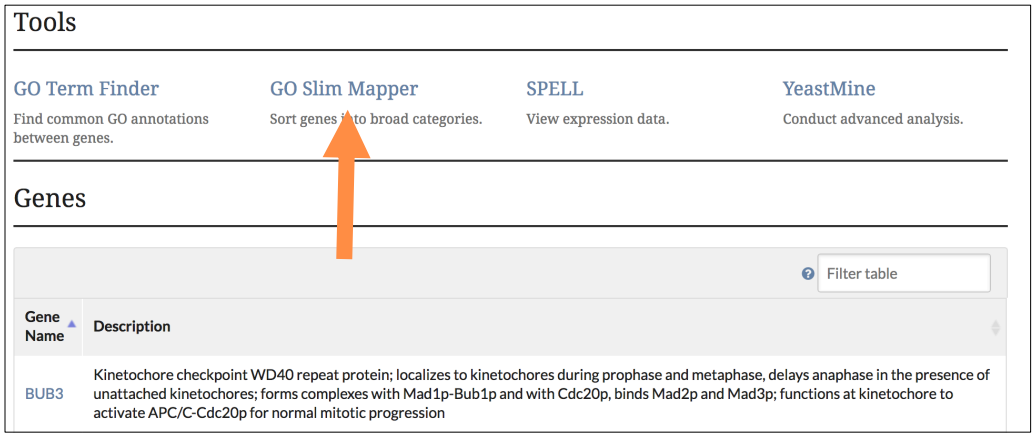
The Gene Ontology (GO) is structured in a hierarchy, such that granular terms ("perinuclear space") are connected and further down the hierarchy than their related broader terms ("nucleus"). However, for many purposes, such as reporting the upregulated cellular functions of a transcriptomics experiment, 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.

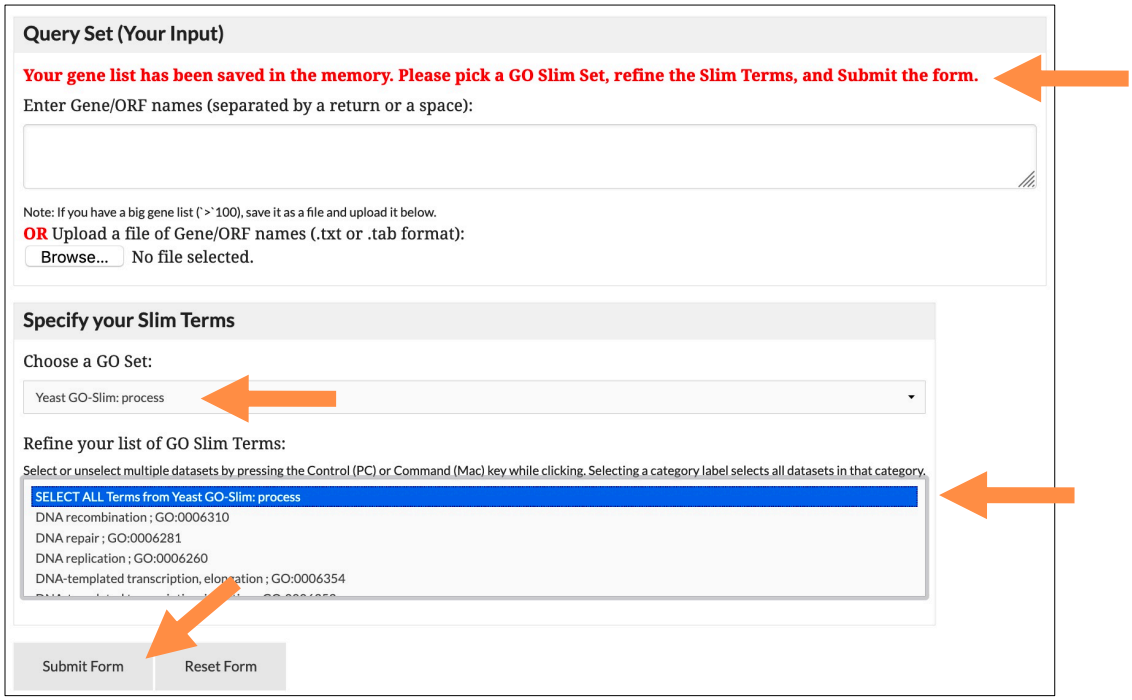
- From the SGD home page (www.yeastgenome.org), go to the Locus Summary page for the uncharacterized gene **YLR287C.**
- Select **Interactions** tab at the top and then scroll to the **Genetic Interactions** section on the page. Here, we are interested in finding genes that have a genetic interaction 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.



- 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**.

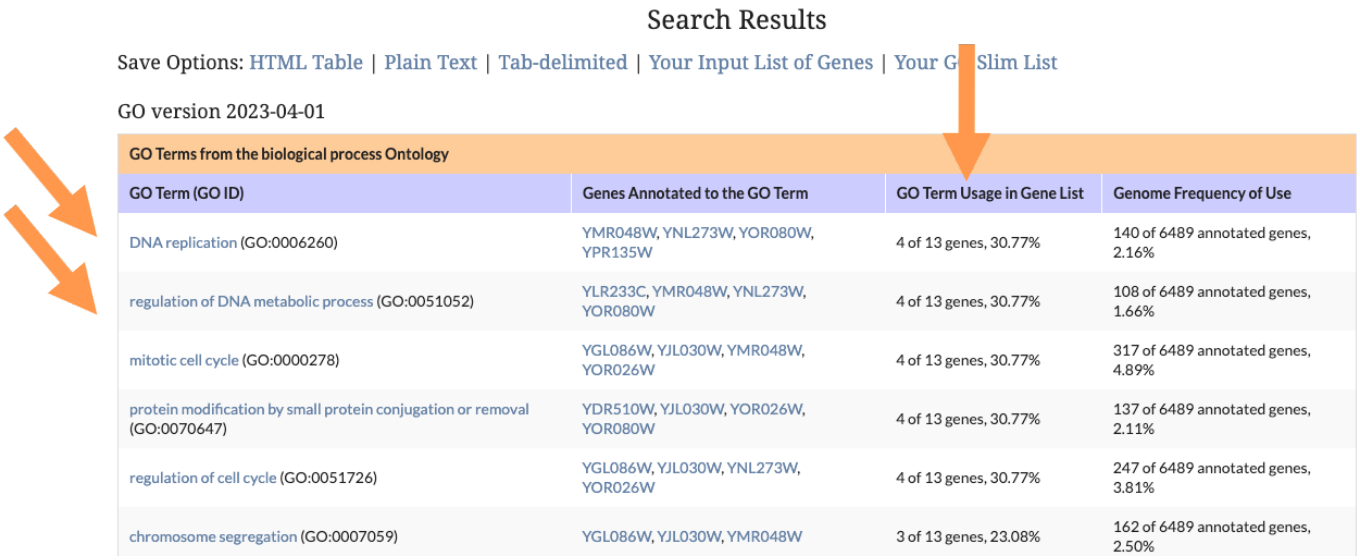


- The GO Slim Mapper has three steps (plus one optional step) in which you can specify your query. The Query Set (Your Input) box has been preloaded in memory with the list of genes you imported from the table.



- 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:
    a. GO Slim terms picked by GO Slim Mapper
    b. Genes from your list that are annotated to that term, hyperlinked to their Locus Summary pages.
    c. GO Term Usage in Gene List (cluster frequency), the number and percentage of genes in your list annotated to each term.
    d. 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 G 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?

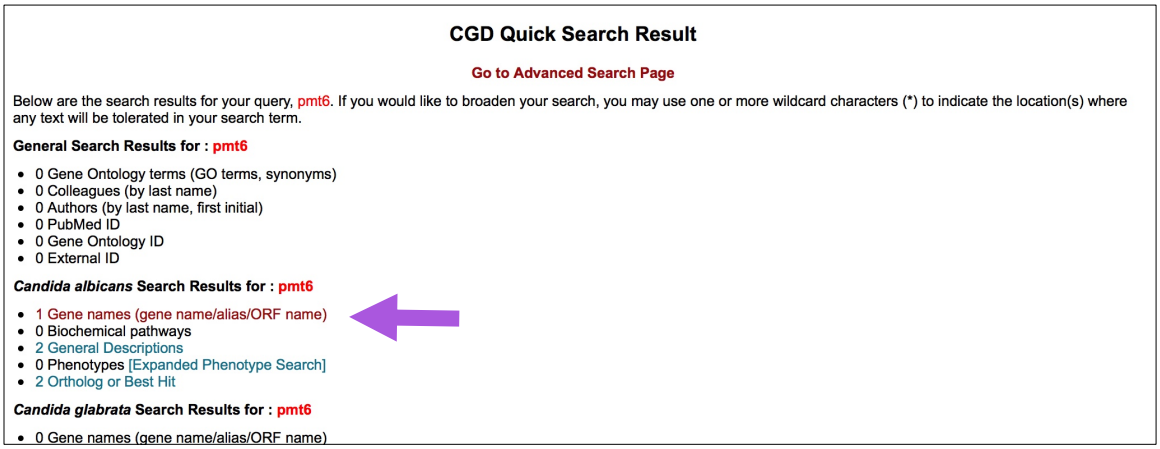
GO Enrichment, Phenotype Data at CGD

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.

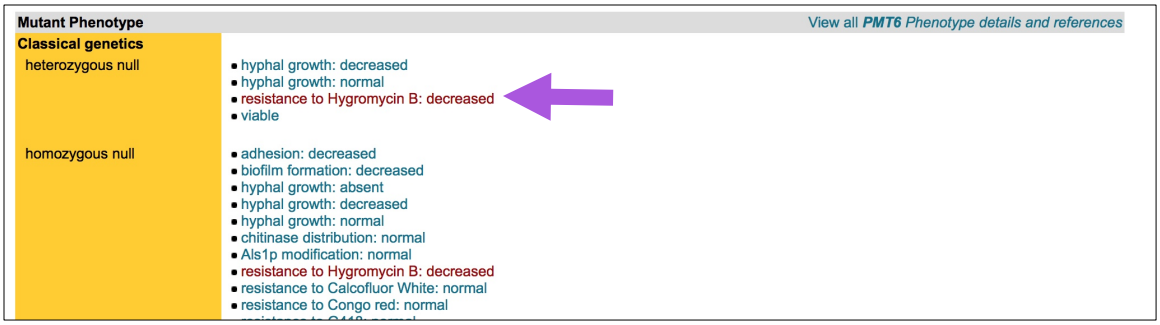
GO Term Finder takes a list of genes and identifies what GO terms are significant for the list. It is a powerful way to interpret the results of omics experiments or any situation where determining common functions and roles are important. For example, GO Term Finder can take a list of upregulated genes 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.

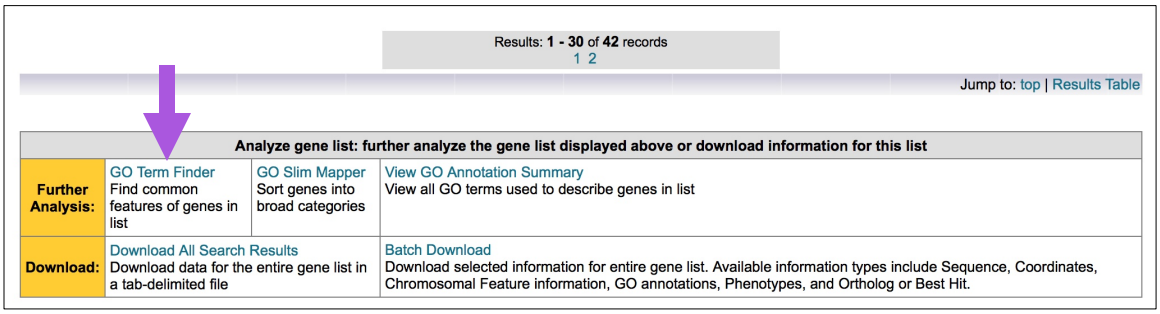
- From the CGD home page (www.candidagenome.org), go to the Locus Summary page for the hygromycin B-sensitivity gene PMT6. Enter **PMT6** into the **search our site** box and click **GO**. On the next page, under ***Candida albicans* Search Results**, click on hyperlinked **1 Gene names (gene name/alias/ORF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



- 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**

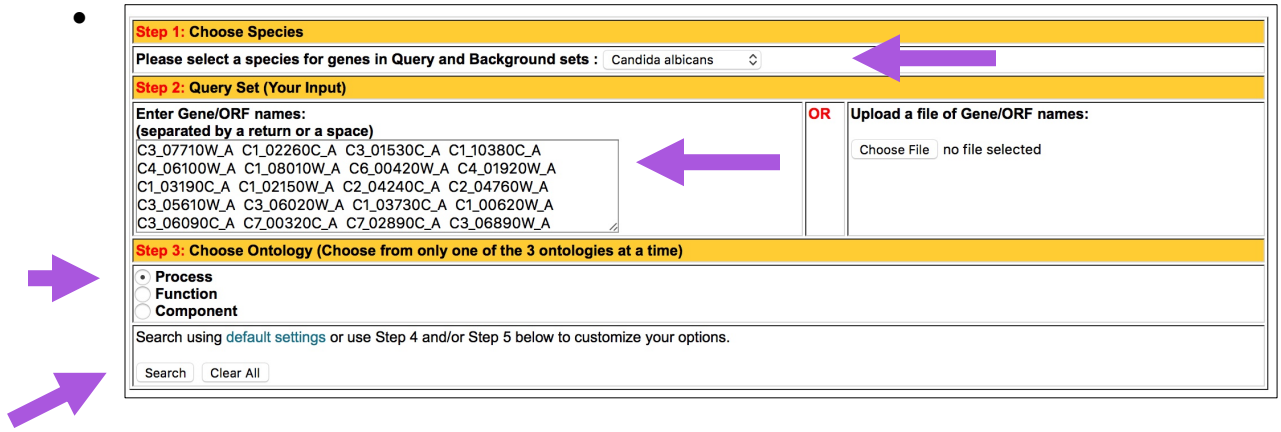


- On the **Phenotype Search Results** page, click on **Jump to: Analyze Gene List** above the table on the right (or simply scroll down to the bottom of the page). Click on **GO Term Finder** link.



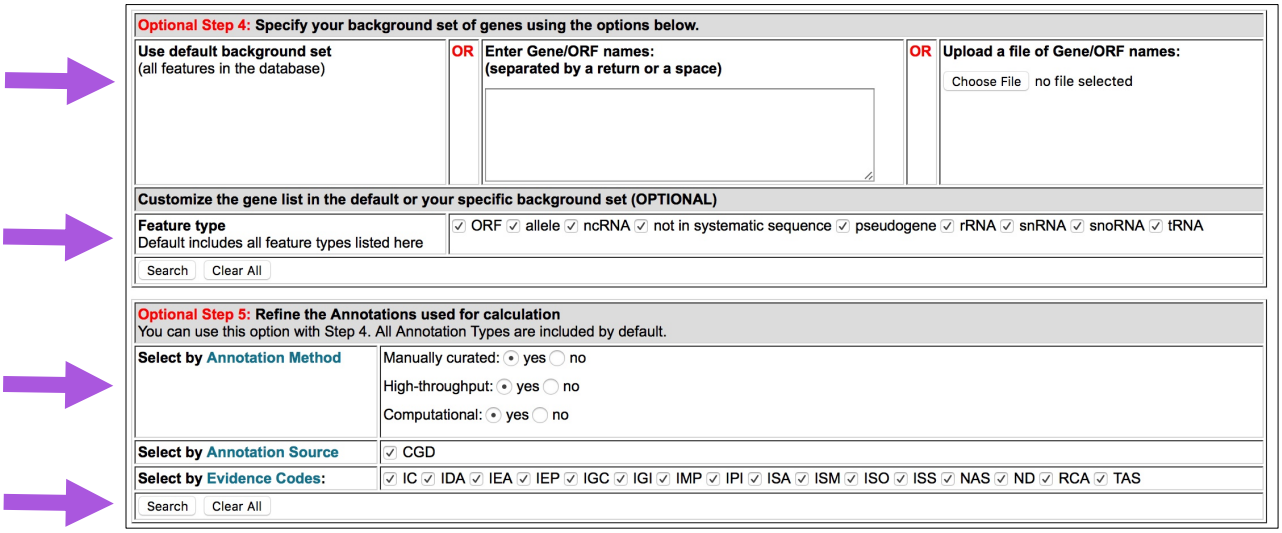
  - With your own list of genes, you can access GO Term Finder from any CGD page by opening **GO** menu in the banner on top and clicking on **GO Term Finder**. Or you use this URL: http://www.candidagenome.org/cgi-bin/GO/goTermFinder

- 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). Chose **Process** as the ontology. Click the **Search** button to use the default settings.
-

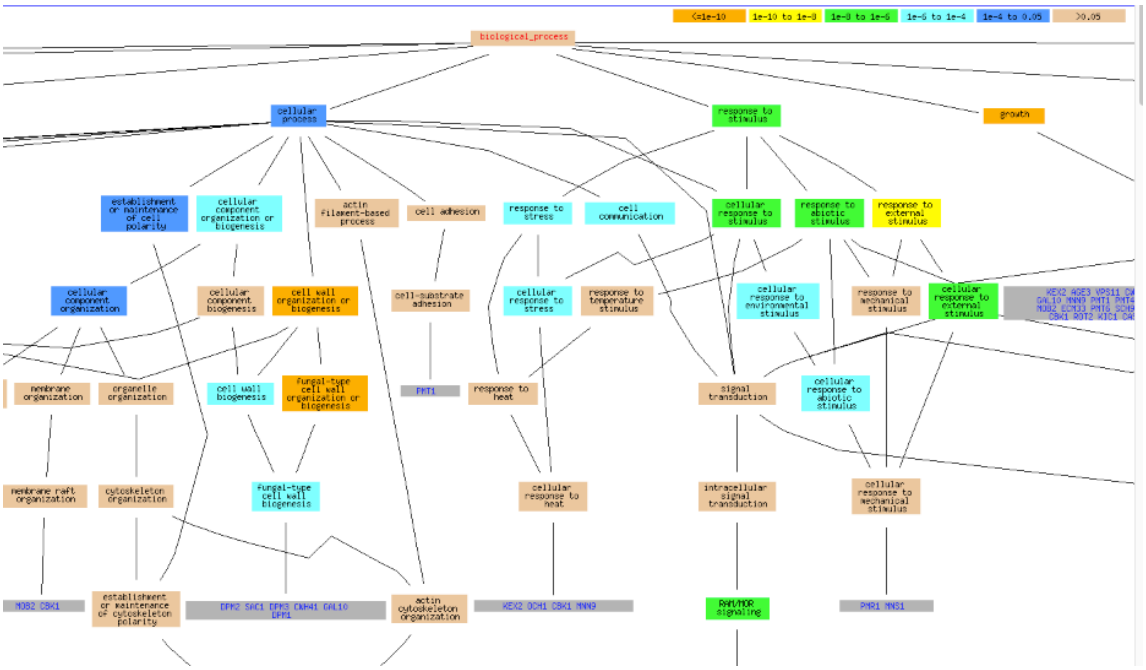


You can further customize your query in the next steps down the page:

- 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



- 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:

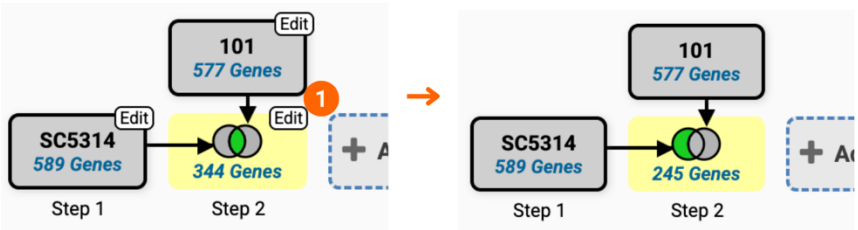
- Perform a GO enrichment analysis
- Create a complex search strategy using both FungiDB and SGD

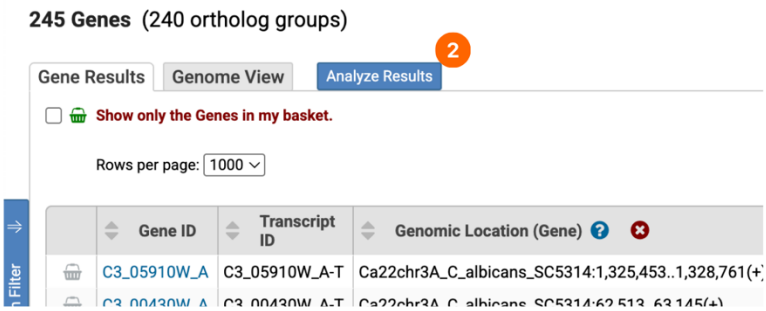
For this exercise, we will query Candida albicans transcriptomics by Kirchner et al. in 2019. The study focuses on the oropharyngeal candidiasis experimental model in mice, which was used to examine *C. albicans'* interaction with the host at mucosal surfaces in vivo. The study involved two strains of *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.**

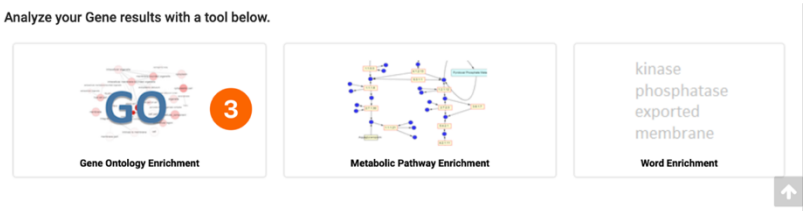
The following strategy was created to identify genes up-regulated in both SC5314 and 101 strains in response to mucosal surfaces:
https://fungidb.org/fungidb/app/workspace/strategies/import/25c2bed9444c4423

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.





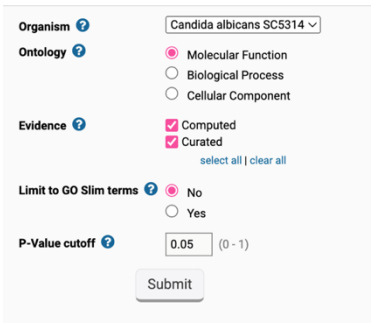
The enrichment analysis tools include Gene Ontology, Metabolic Pathway, and Word Enrichment tools.



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.

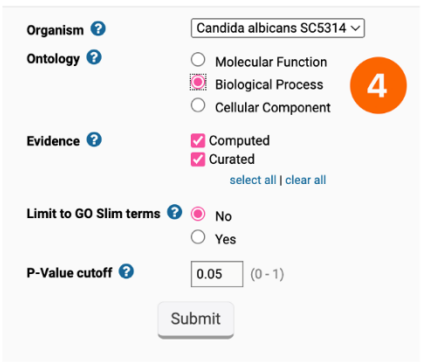
GO enrichment analysis can be performed on the following ontology groups:

- 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.



Analysis Results:

244 rows

Open in Revigo | Show Word Cloud | Download

GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0042273	ribosomal large subunit biogenesis	558	67	12.0	3.03	4.20	1.08e-17	1.68e-14	1.68e-14
GO:0000470	maturation of LSU-rRNA	440	55	12.5	3.16	4.20	3.31e-15	2.59e-12	5.17e-12
GO: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.

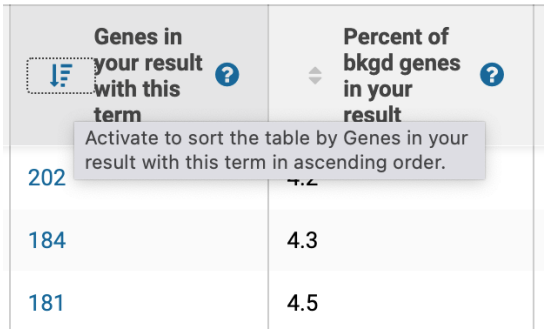
- **Odds ratio—Determines whether the odds of the GO term appearing in the list of interest are the same as those** for the background list.
- **P-value** - Assumptions under a null hypothesis, the probability of getting a result that is equal to or greater than what was observed.
- **Benjamini-Hochburg false discovery rate** - A method for controlling false discovery rates for type 1 errors.
- **Bonferroni adjusted P-values** - A method for correcting significance based on multiple comparisons.

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?

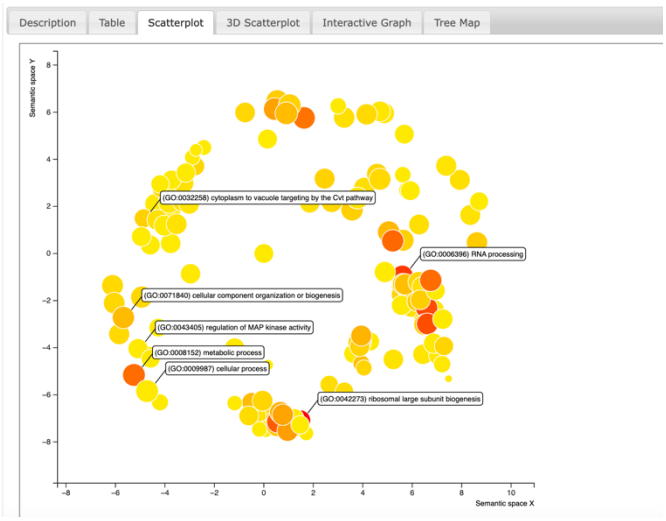
Note: you can sort genes in your results using the sort options within a column:



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

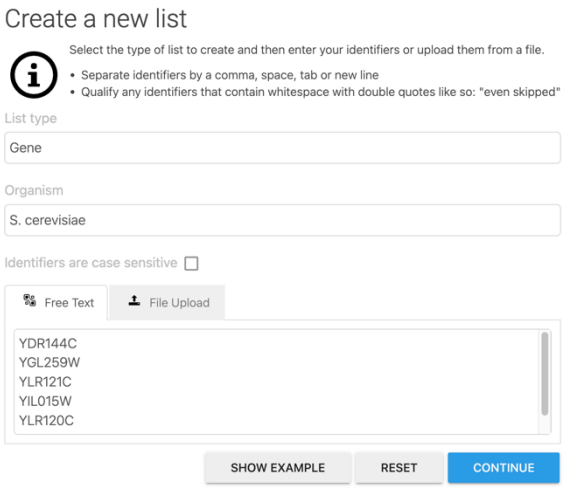
   - Run an InterPro search and a GPI anchor prediction tool. What did you learn about this protein?

   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 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.



- 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?

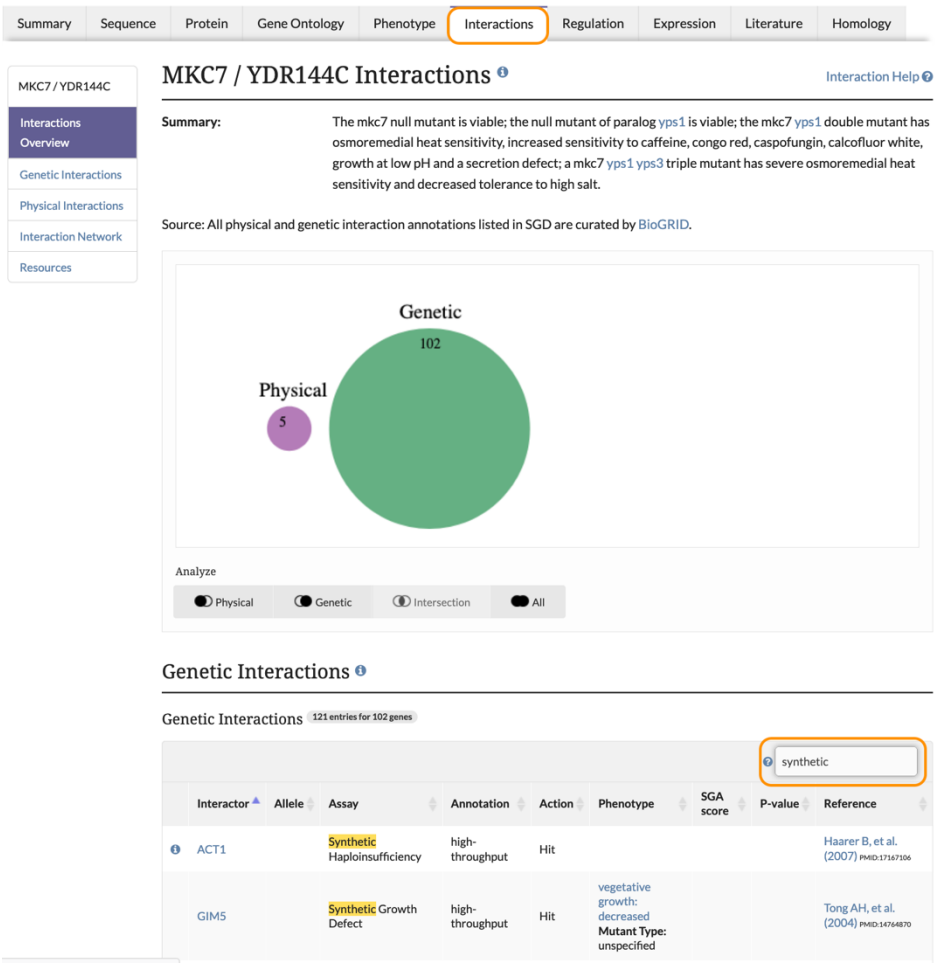
  Hint: see the **GO Annotation** section under the 'Data' on the locus page.

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.



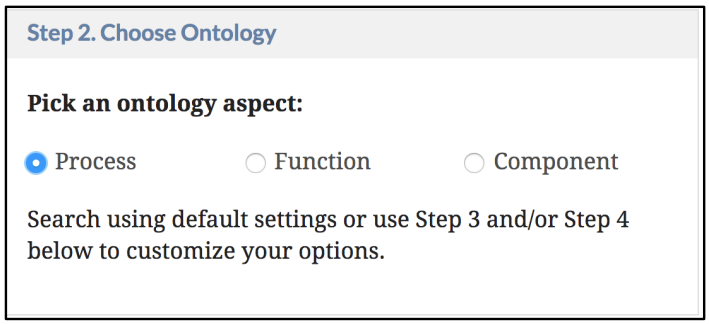
- Click on the **Download** button, which is located under the results table, and save this gene list. *Rename the file to* ***synthetic.txt***.

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.



- 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

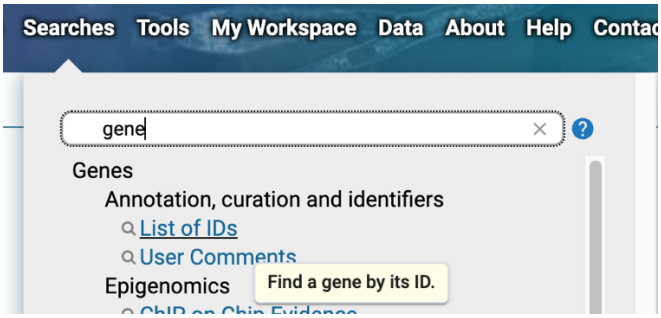
Now, let's go back to the file of MKC7 "synthetic" genetic interactors we downloaded earlier and find the orthologs of these genes in *Lomentospora prolificans*.

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

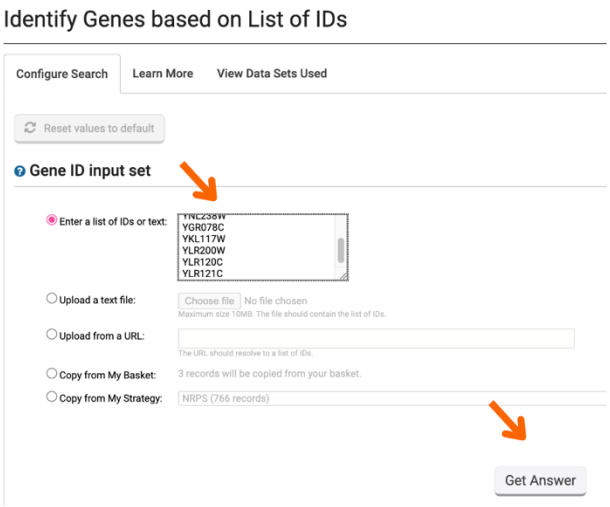
Interactor	Interactor Sy	Interactor	Interactor Systematic Name	Type	Assay	Annotation
MKC7	YDR144C	ACT1	YFL039C	Genetic	Synthetic Ha	high-through
MKC7	YDR144C	GIM5	YML094W	Genetic	Synthetic Gr	high-through
MKC7	YDR144C	IRE1	YHR079C	Genetic	Synthetic Gr	manually cur
MKC7	YDR144C	KEX2	YNL238W	Genetic	Synthetic 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

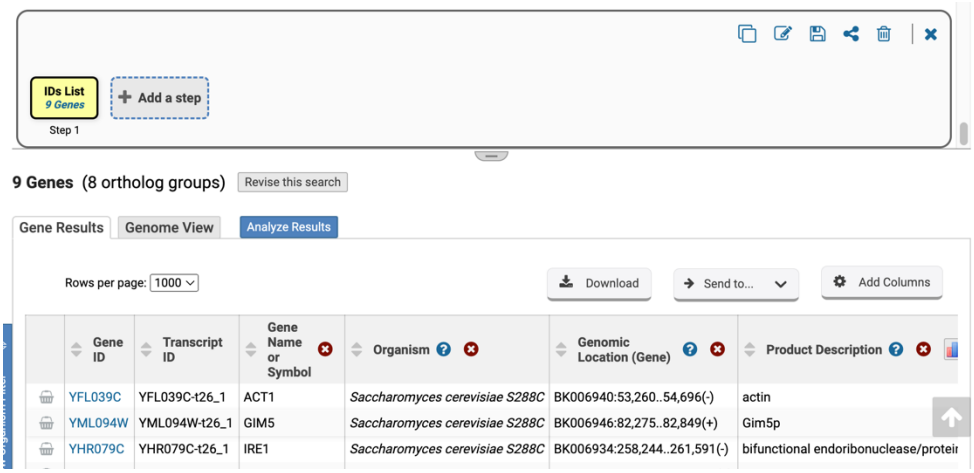
- Revisit FungiDB and initiate the List of IDs search query

The query can be deployed from the “Searches” menu at the top of the “Search for Genes” section on the main page.



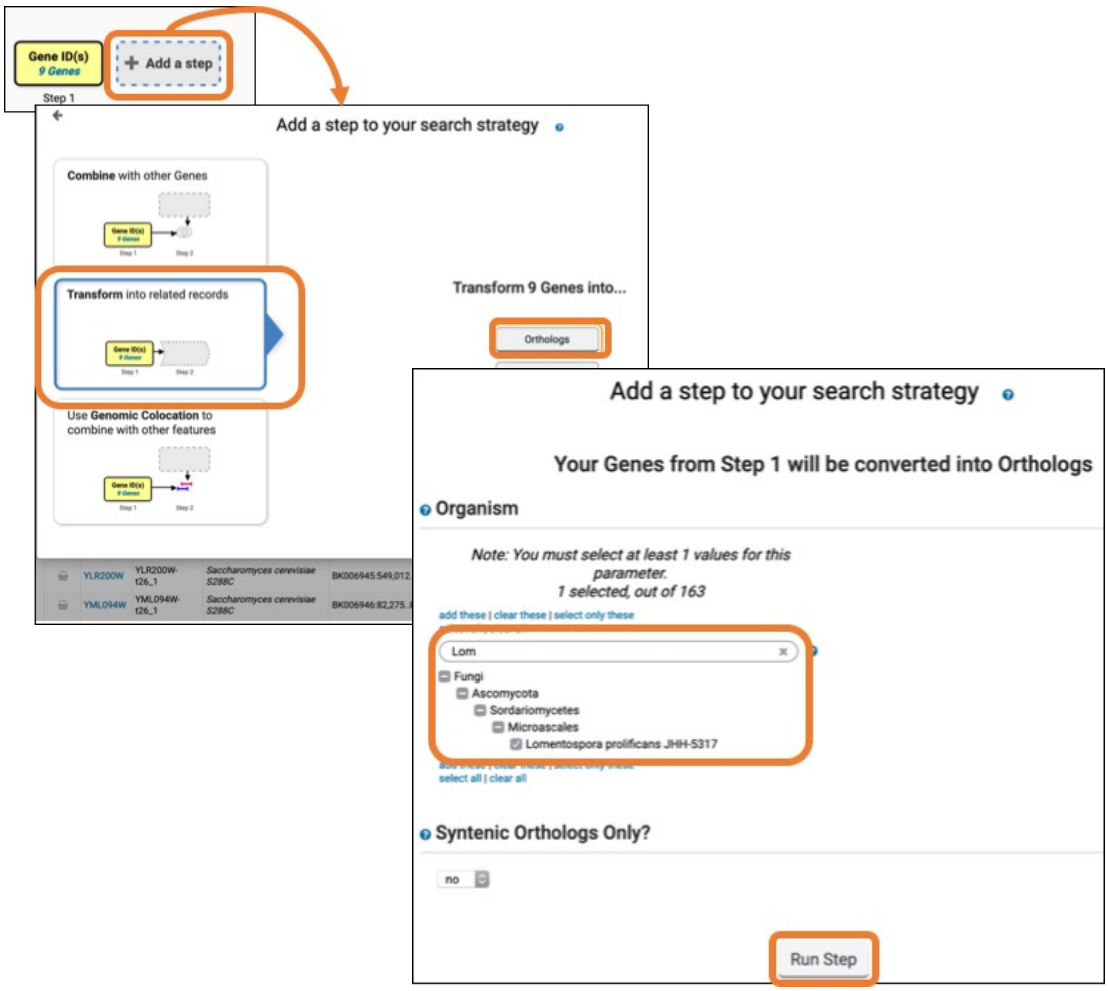
- Paste the list of Gene IDs with the “synthetic” genetic interactions with MKC7 into the FungiDB query and click the **Get Answer** button.

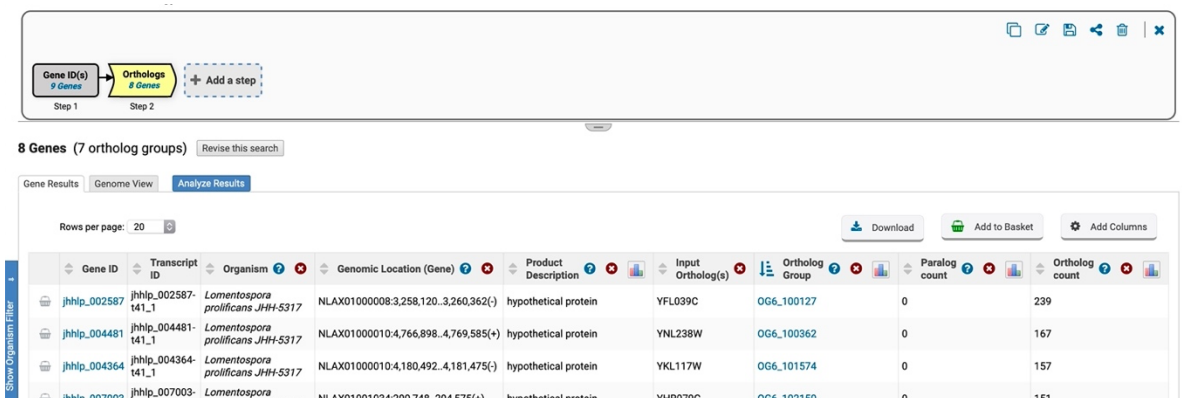




- Find orthologs in *Lomentospora prolificans*.

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**.





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

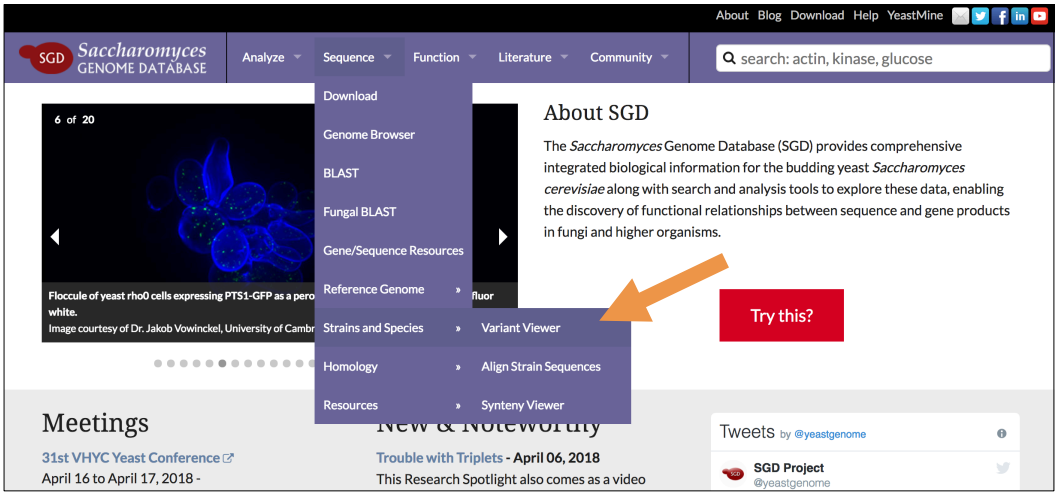
SGD's Variant Viewer (**https://yeastgenome.org/variant-viewer**) is an open-source web application that compares nucleotide and amino acid sequence differences between 12 common *S. cerevisiae* laboratory strains. For a given open reading frame, Variant Viewer breaks down the position and nature of 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.

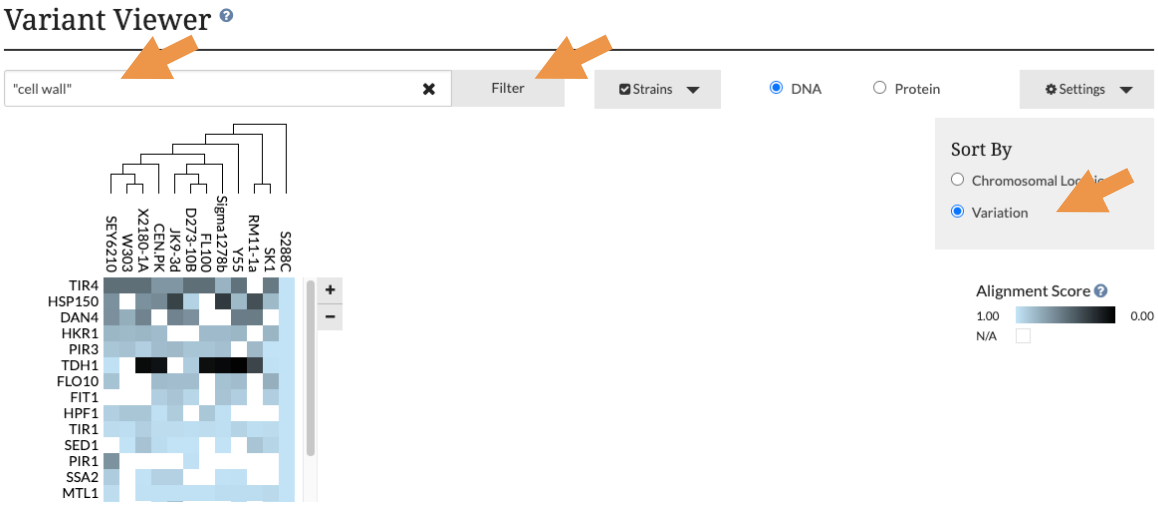
In this exercise, we will use Variant Viewer to find out what genetic differences between Sigma1278b and S288C explain why they differ in their ability to form pseudohyphae.

S288C vs. Sigma1278b: Cell Wall

- Open the SGD home page (www.yeastgenome.org), 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

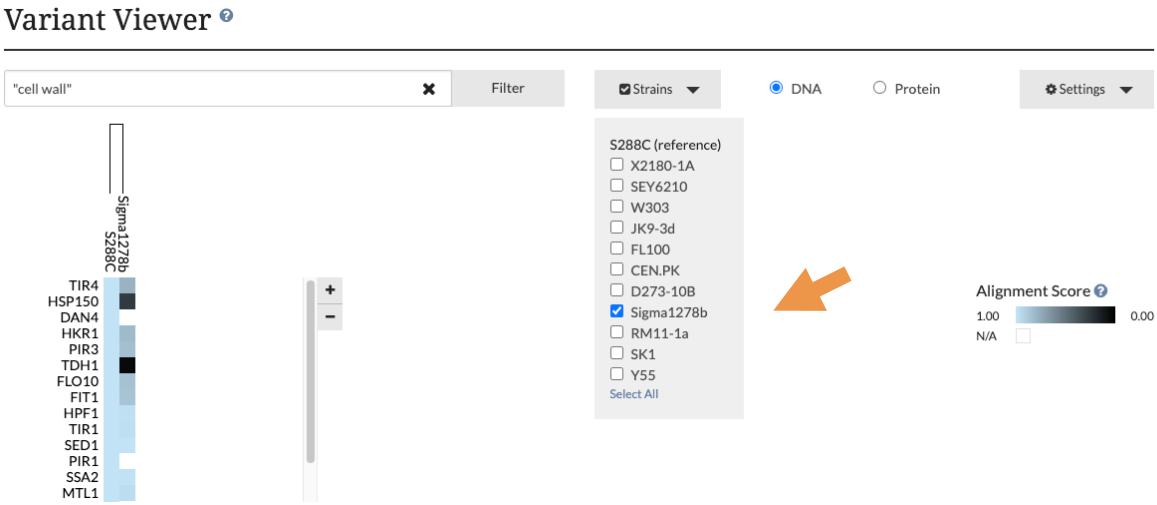


- The **Filter** box accepts one or more genes, as well as Gene Ontology (GO) terms. Because we are interested in genes involved in cell wall development, enter the GO term "**cell wall**" (don't forget the quotation marks) and select **variation** in the settings pull-down, then click **Filter**.

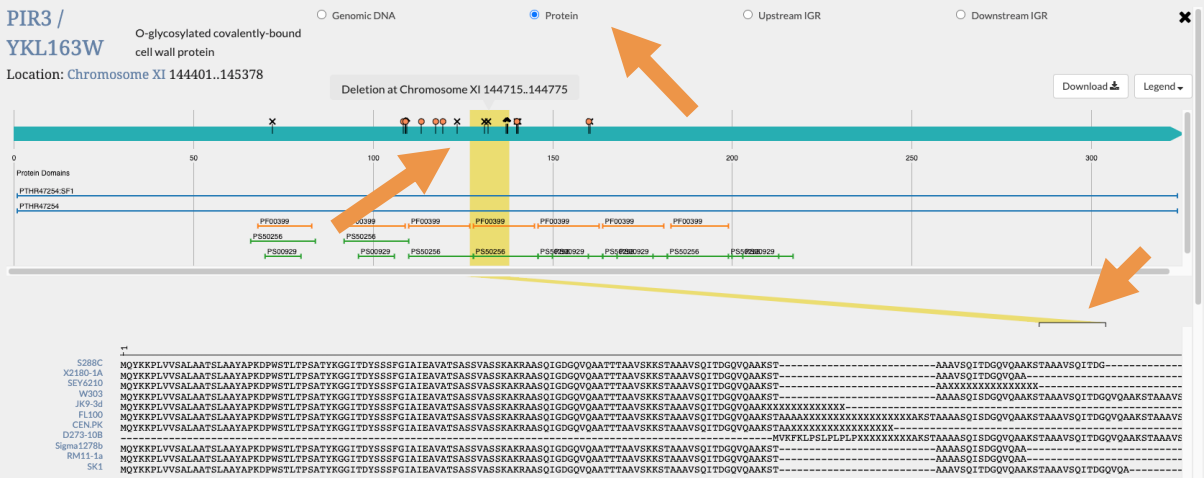


- The **matrix**, shown on the left, will have changed to only include the genes that localize to cell walls.
    - This matrix enables you to visualize high-level differences in multiple genes relative to strain S288C. Each square in the matrix corresponds to one of the twelve strains in Variant Viewer, 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 Sigma1287b). Open the **Strains** pull-down menu, press Deselect All, then re-select Sigma1287b.



- 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

1. 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
2. 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
3. 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 PubMed
4. Porter SE, et al. (2002) The yeast pafl-rNA 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
5. 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

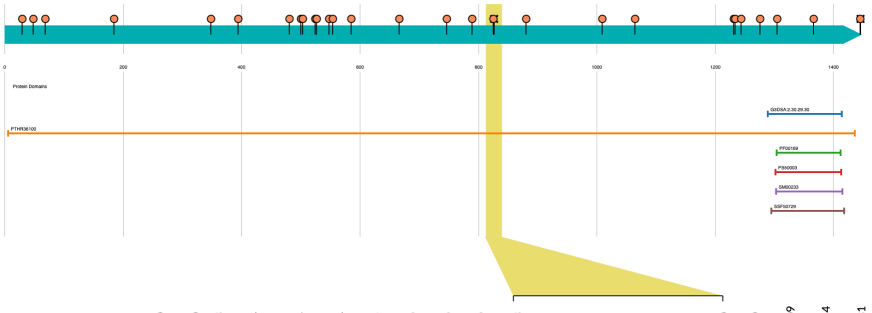
- Variant Viewer is also embedded in the Sequence tab of every gene page, with the data for the gene already pre-loaded from the results of the Variant Viewer search. This allows you to look at the variant information for a gene without starting from the tool's entry page.

Variants

Genomic DNA | Protein | Upstream IGR | Downstream IGR

Location: Chromosome X 598735..603078

Download | Legend



	480	500	503	524	547	554	585	666	746	789	813	840	880	1009	1064	1231
S288C	F	T	P	N P	S	R	P	S	I	L	GDITFNRGDLLSLSFDEELGQDFANFL		P	E	E	H
X2180-1A	F	T	P	N P	S	R	P	S	I	L	GDITFNRGDLLSLSFDEELGQDFANFL		P	E	E	H
SEY6210	F	T	P	N P	S	R	P	S	I	L	GDITFNRGDLLSLSFDEELGQDFANFL		P	E	E	H
W303	F	T	P	N P	S	R	P	S	I	L	---------------------------		P	E	E	H
JK9-3d	F	T	P	N P	S	R	P	S	I	L	GDITFNRGDLLSLSFDEELGQDFANFL		P	E	E	H
FL100	F	T	P	N P	S	R	P	S	I	L	GDITFNRGDLLSLSFDEELGQDFANFL		P	E	E	H
CEN.PK	F	T	P	N P	S	R	P	S	V	L	GDITFNRGDLLSLSFDEELGQDFANFL		L	E	K	H
D273-10B	F	T	P	N P	S	R	P	S	V	L	GDITFNRGDLLSLSFDEELGQDFANFL		L	G	E	P
Sigma1278b	F	T	P	N P	S	R	P	S	I	L	GDITFNRGDLLSLSFDEELGQDFANFL		P	E	E	H
RM11-1a	S	T	S	N P	S	S	S	S	V	P	GDITFNRGDLLSLSFDEELGQDFANFL		P	E	E	H
SK1	F	T	P	N P	S	R	P	S	I	L	GDITFNRGDLLSLSFDEELGQDFANFL		P	E	E	H
Y55	S	I	S	S L	T	R	P	Y	V	P	GDITFNRGDLLSSSFDEELGQDFANFL		P	E	E	H

FungiDB: SNPs and Population Genetics

Learning Objective:

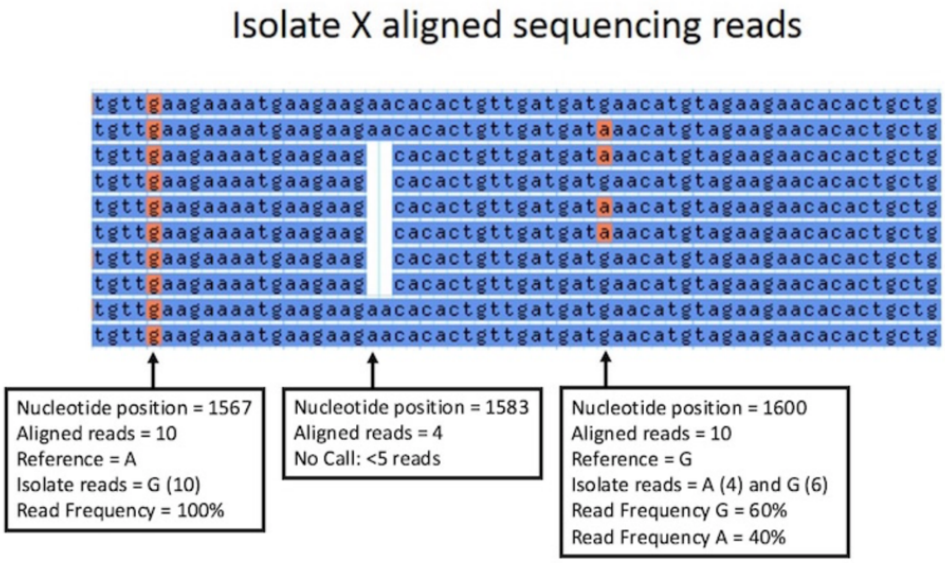
- Investigate SNP datasets using the following searches:
  - SNP characteristics,
  - SNPs between groups of isolates,
- Identify aneuploidy with the copy number variations search.

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).

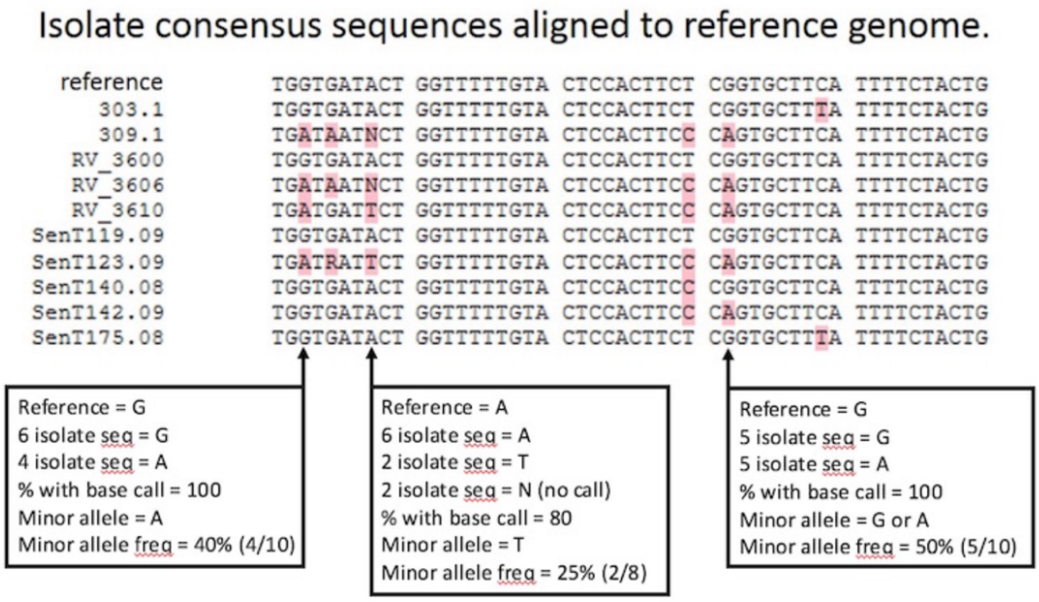
Read Frequency Threshold:

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.



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.



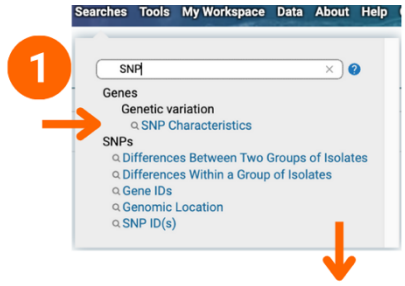
Percent isolates with a base call:
Sometimes an isolate does not have a base call at a certain nucleotide position because the Read Frequency Threshold was not met or because there were less than 5 aligned sequencing reads for that nucleotide position. In this case, an SNP can be returned by the search 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.

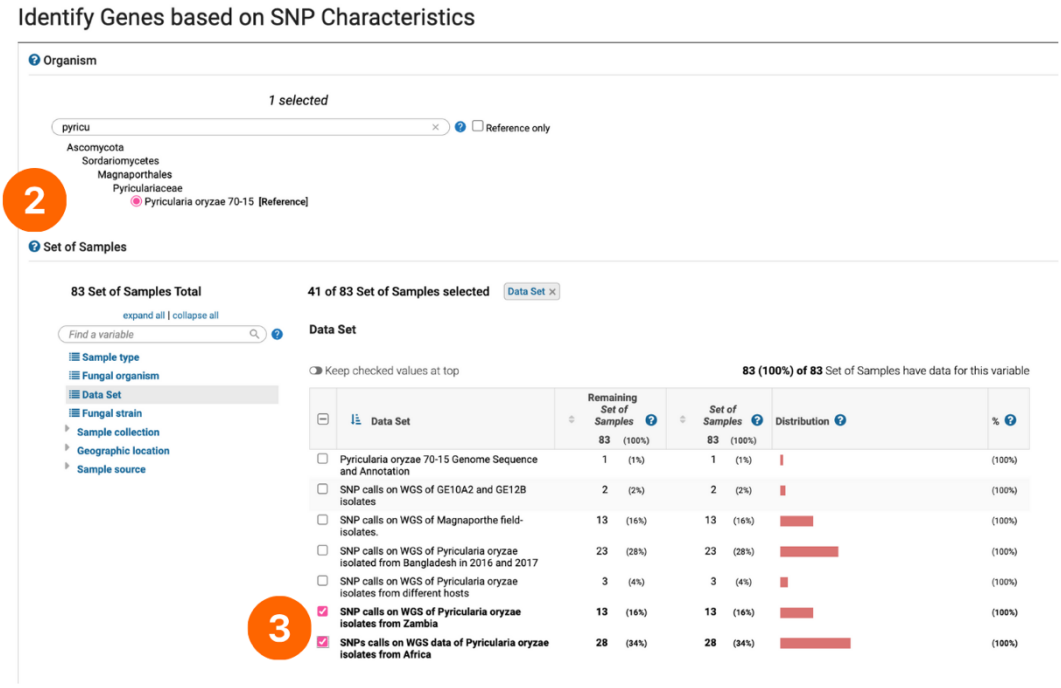
A. Identify putative nuclear effectors with at least 1 non-synonymous SNP in *Pyricularia oryzae*.

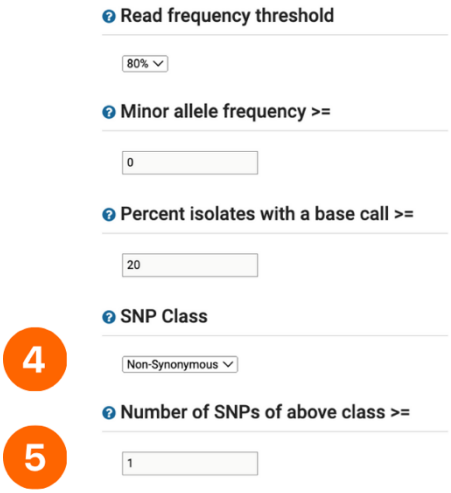
In this exercise, you will analyze a collection of *P. oryzae* isolates obtained from rice plants in different parts of Africa to identify genes with at least one non-synonymous SNP that also exhibit characteristics of nuclear effectors.

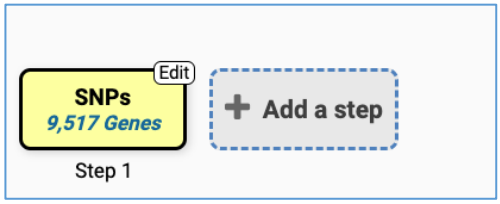
Pyricularia oryzae (also known as *Magnaporthe oryzae*) is a pathogen that affects rice crops, causing a severe disease known as rice blast. During the infection process, *P. oryzae* and other plant pathogens use various effectors to manipulate the plant's immune system. Nuclear effectors are a type 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.

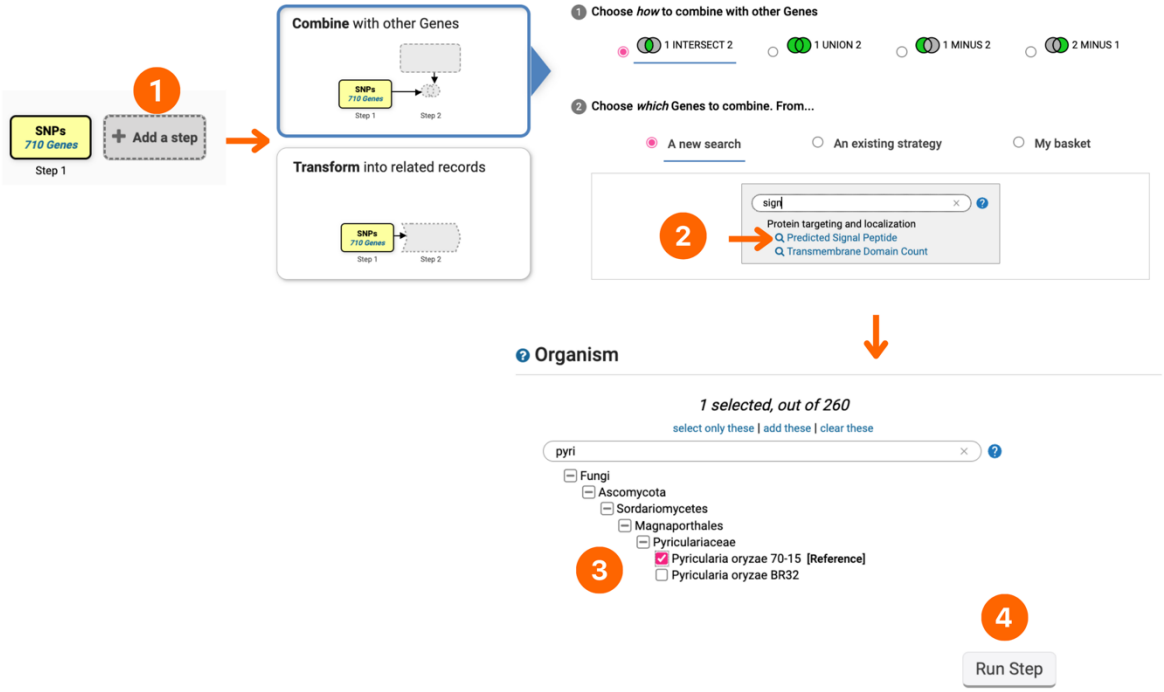


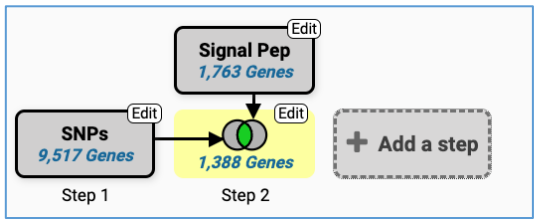






- **Identify putative nuclear effectors.**
    1. Click on the “Add a Step” button.
    2. Use the “Combine with Other Genes” option to deploy the “Predicted Signal Peptide” search.
    3. Set the genome to *Pyricularia oryzae* 70-50.
    4. Click on the “Run Step” button.

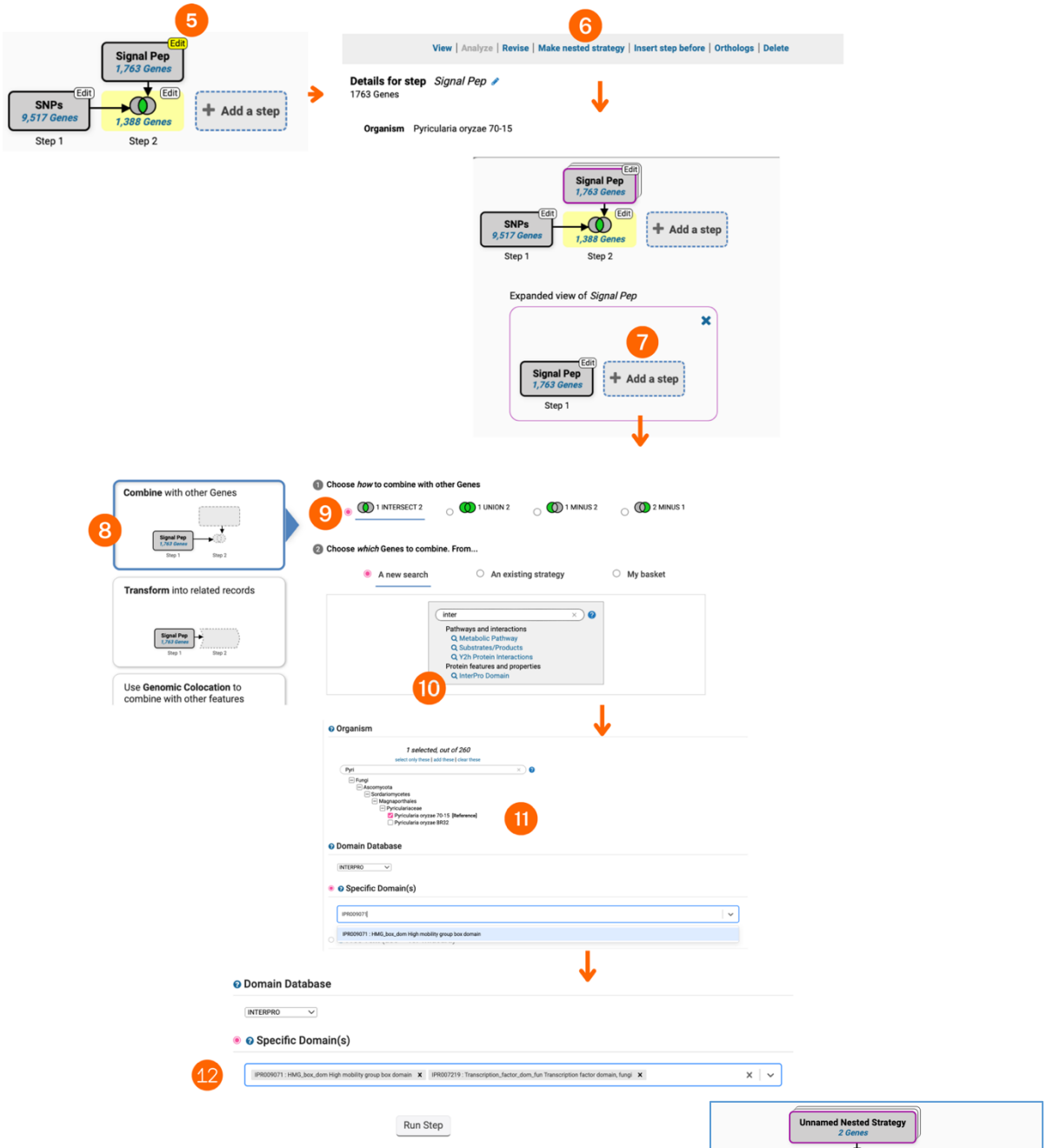




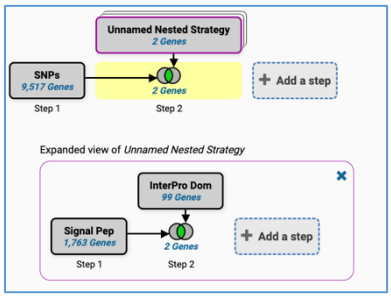
The strategy returns genes with at least 1 SNP and a predicted signal peptide domain. How can we identify genes with at least 1 SNP and a predicted signal peptide domain AND a DNA-binding domain (IPR007219 or IPR009071)? (Hint: you can do this with a nested strategy as described below).

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.

11. 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.
12. Click on the "Run Step" once both domains are selected.



How many genes were identified to be putative nuclear effectors that have non-synonymous SNPs in the isolates collected in Africa?

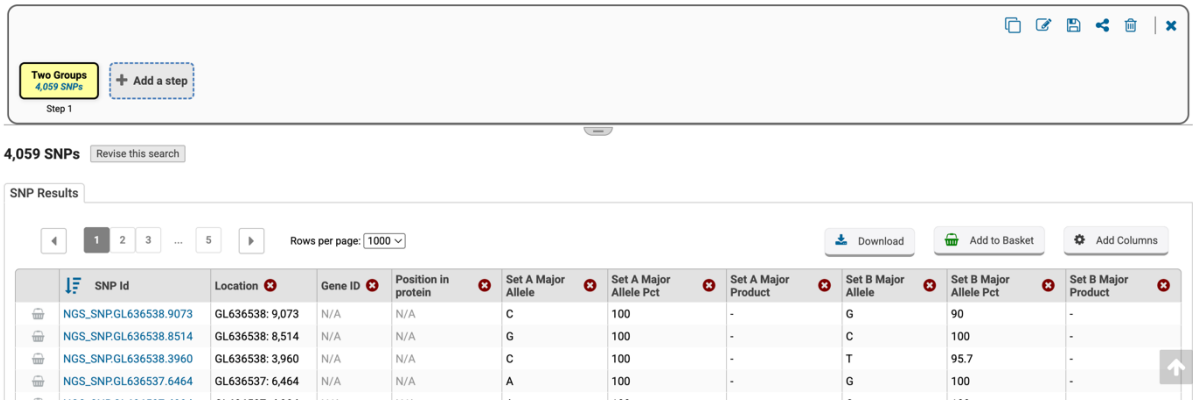


Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/bd657f5629cac5df
Reference: https://www.nature.com/articles/s41467-020-19624-w

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



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 Allele Pct	Set A Major Product	Set B Major Allele	Set B Major Allele 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	93.3	-
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	-

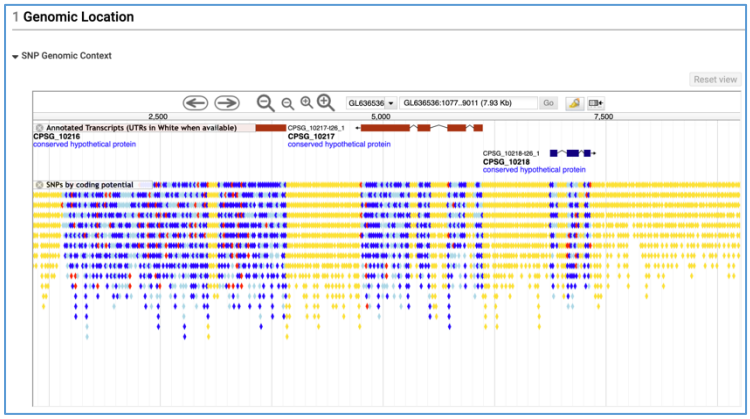
Each SNP is also linked to its own record page. Click on the following link to explore the SNP record page in more detail: NGS_SNP.GL636536.6075.

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.

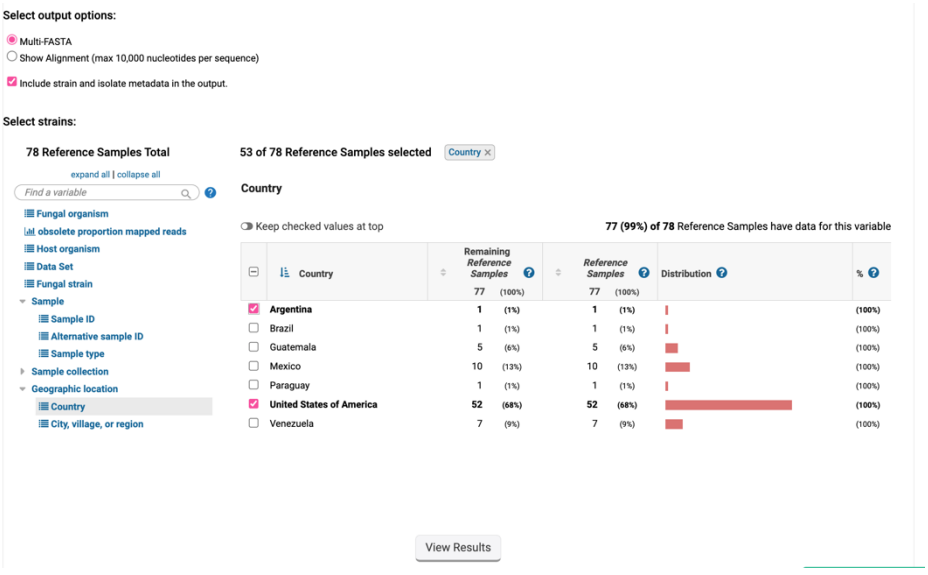
A summary of all SNPs detected in this gene across all datasets integrated into FungiDB is displayed in the SNP Genomic Context section:

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:



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

Country Summary Download Data Sets

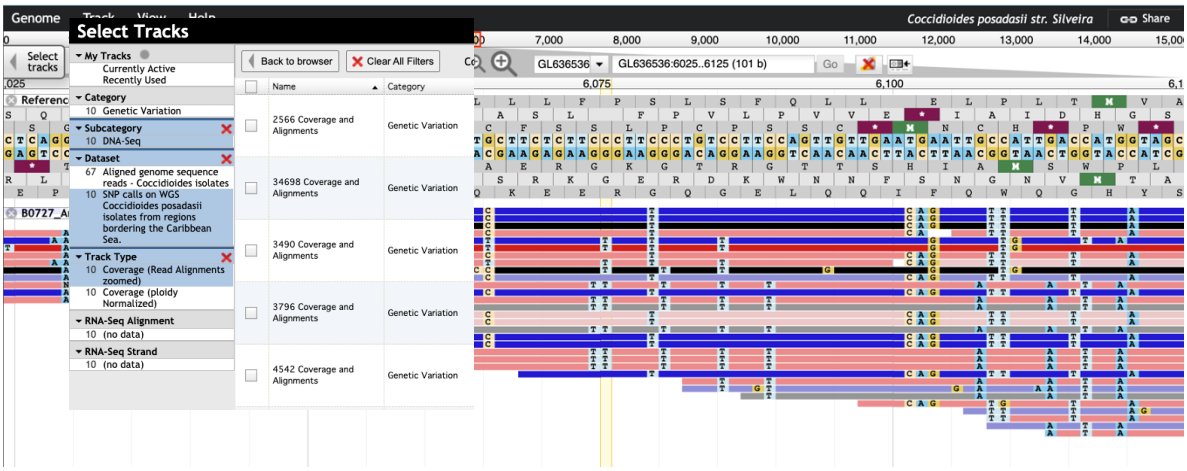
Search this table...

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

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

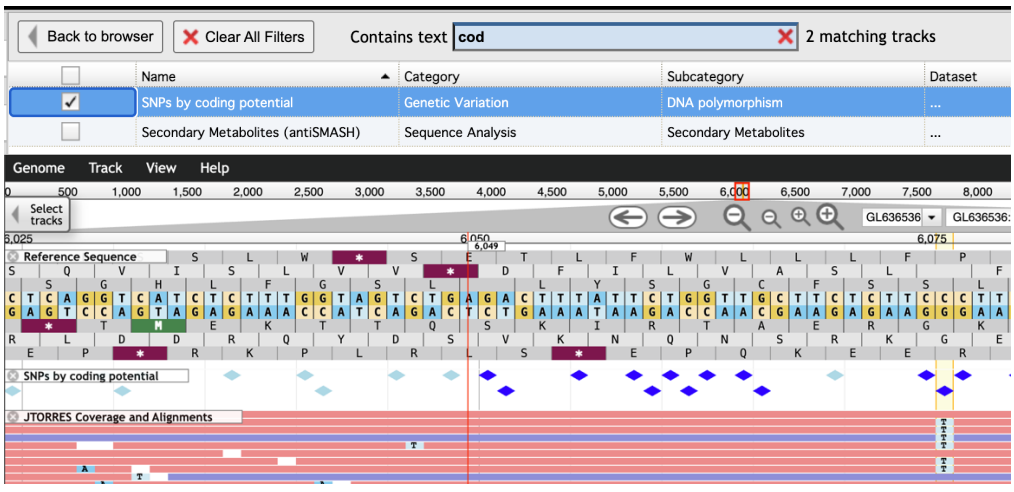
Venezuela	JTORRES	EUSMPL0102-1-7	C	G	C	75	100	view DNA-seq reads

- 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.



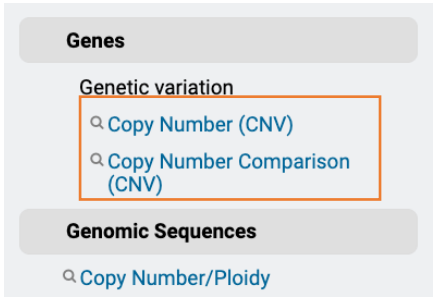
C. Lean how to use copy number variation (genes) search.

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.

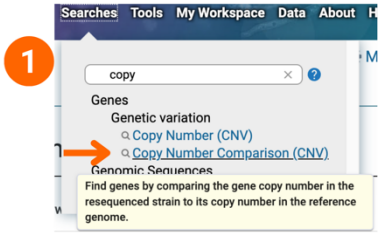
For example, a single-copy gene in a diploid organism has a haploid number of 1 and a gene dose of 2. You can choose to search for genes where at least one of your selected isolates meets your cutoff criteria for the chosen metric (By Strain/Sample), or where the median of the chosen metric across all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

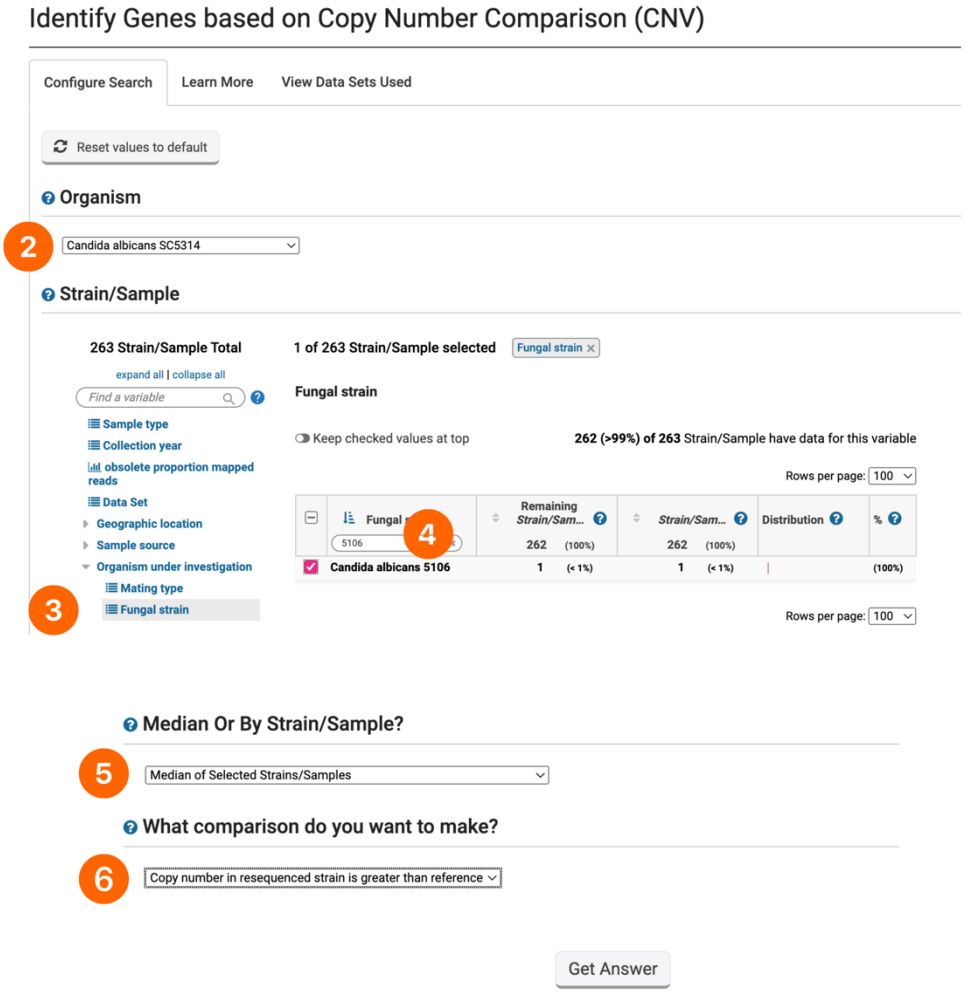
In the next exercise, we will discover aneuploidy in a clinical strain of *Candida albicans*.

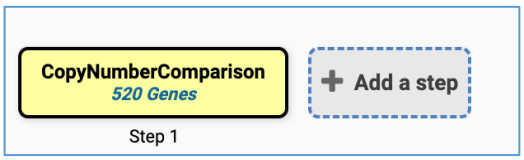
- **Discover regions of potential segmental aneuploidy in *Candida albicans* isolate 5106.**
    1. Deploy the “Copy Number Comparison (CNV)” search.
    2. Select the genome for “Candida albicans”.
    3. Navigate to the Fungal strain” metadata field.
    4. Filter isolates for “5106” and check the 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.

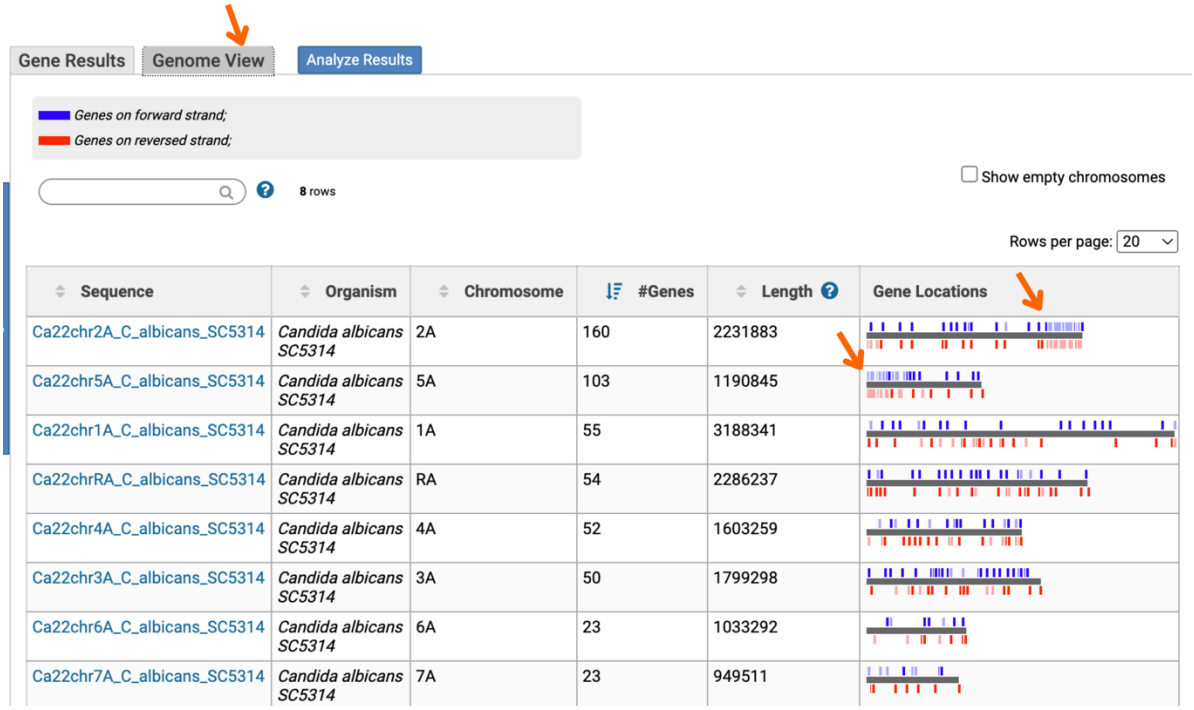
    6. From the drop-down menu select the “Copy number in resequenced strain is greater than reference” option.







Examine the results using the Genome View option.



Sequence	Organism	Chromosome	#Genes	Length	Gene Locations
Ca22chr2A_C_albicans_SC5314	*Candida albicans SC5314*	2A	160	2231883	
Ca22chr5A_C_albicans_SC5314	*Candida albicans SC5314*	5A	103	1190845	
Ca22chr1A_C_albicans_SC5314	*Candida albicans SC5314*	1A	55	3188341	
Ca22chrRA_C_albicans_SC5314	*Candida albicans SC5314*	RA	54	2286237	
Ca22chr4A_C_albicans_SC5314	*Candida albicans SC5314*	4A	52	1603259	
Ca22chr3A_C_albicans_SC5314	*Candida albicans SC5314*	3A	50	1799298	
Ca22chr6A_C_albicans_SC5314	*Candida albicans SC5314*	6A	23	1033292	
Ca22chr7A_C_albicans_SC5314	*Candida albicans SC5314*	7A	23	949511	

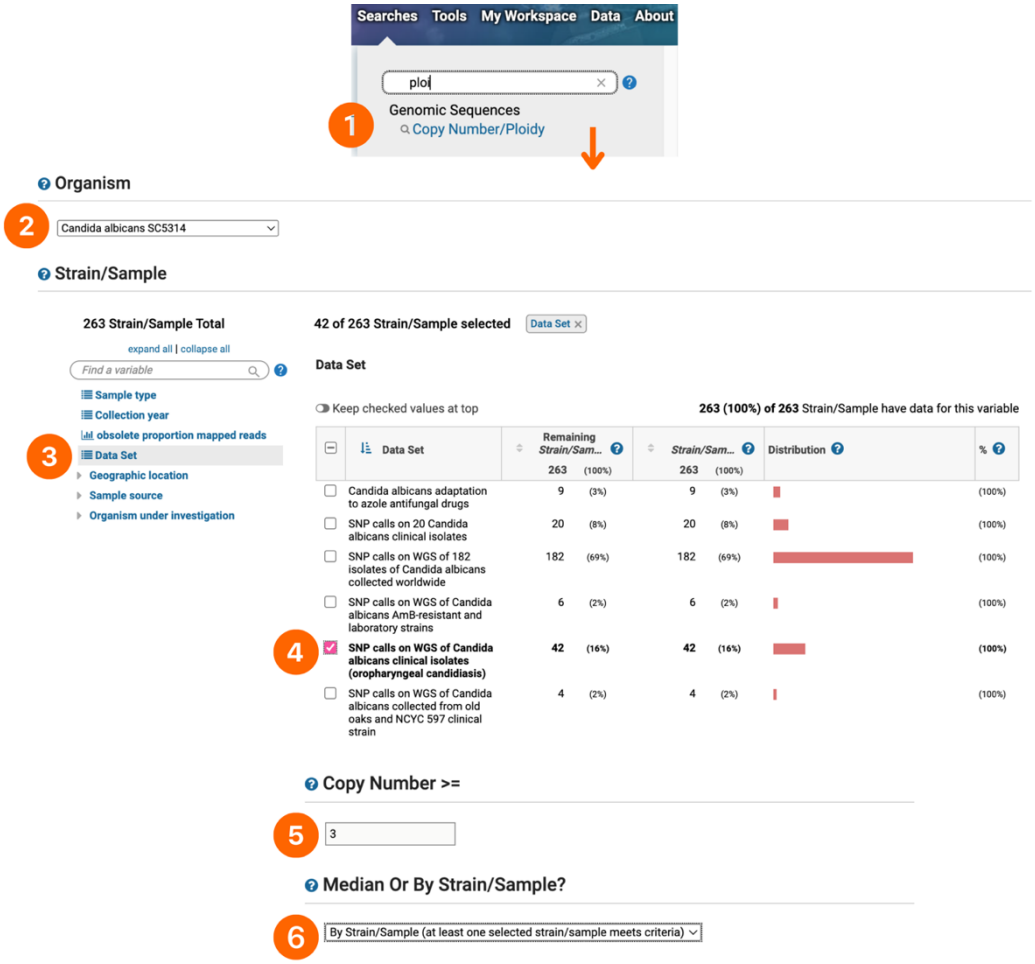
As you can see in the highlighted regions, large numbers of genes predicted to have increased copy numbers are clustered at the right-hand end of chromosome 2 and the left-hand end of chromosome 5, corresponding to the segmental aneuploidies shown in the JBrowse session above.

Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/07b439e0de5e9c6a

Optional

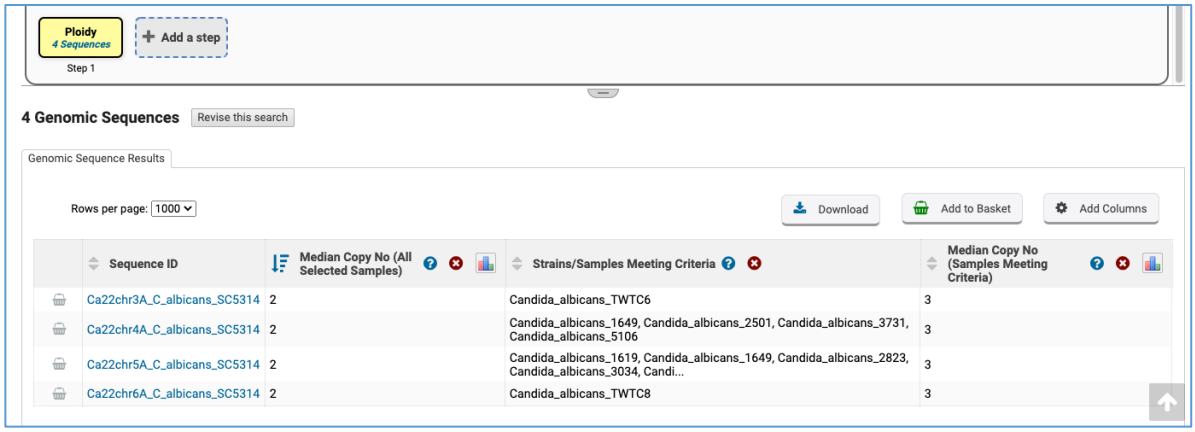
Copy Number/Ploidy search (Genomic Sequences)
Find genomic sequences or chromosomes based on their estimated copy number in resequenced strains. Genomic sequences returned by the search will either have a median estimated copy number greater than or equal to the value you entered for the Copy Number 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 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).

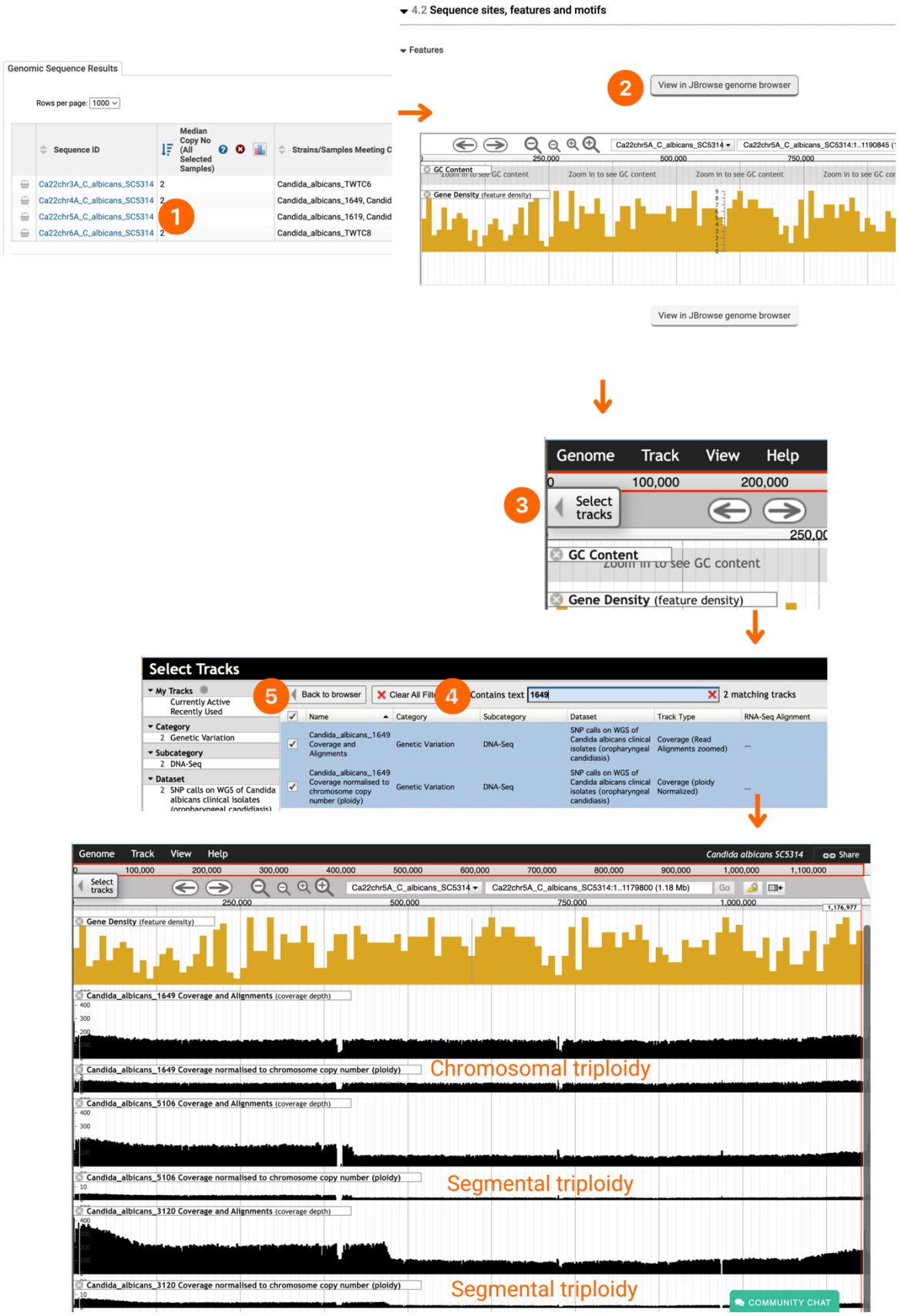


- **Explore segmental aneuploidy in JBrowse.**

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.



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/jbrowse/index.html?loc=Ca22chr5A_C_albicans_SC5314%3A1..1190845&data=%2Ffungidb%2Fservice%2Fjbrowse%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/

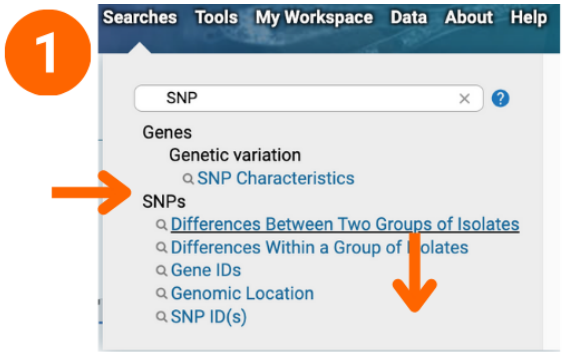
Supplement

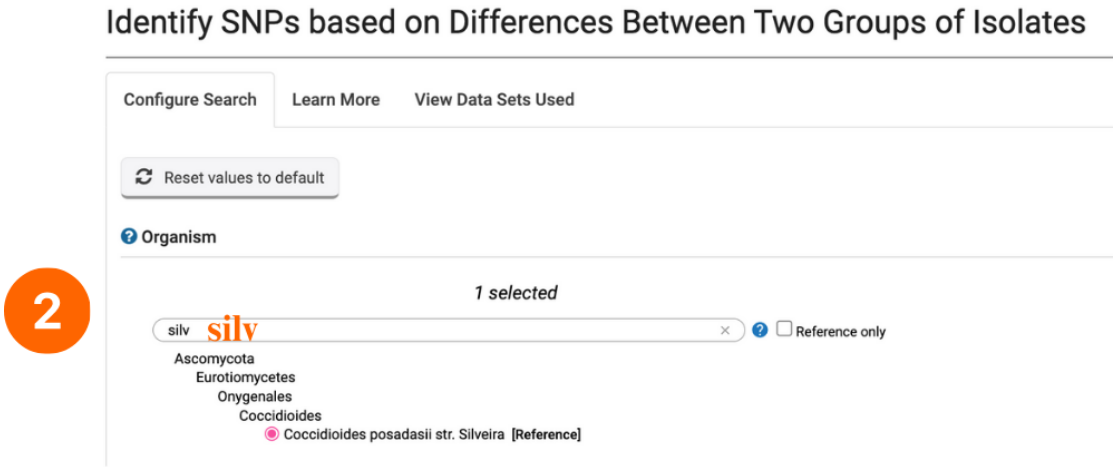
How to distinguish SNP between two groups of isolates.

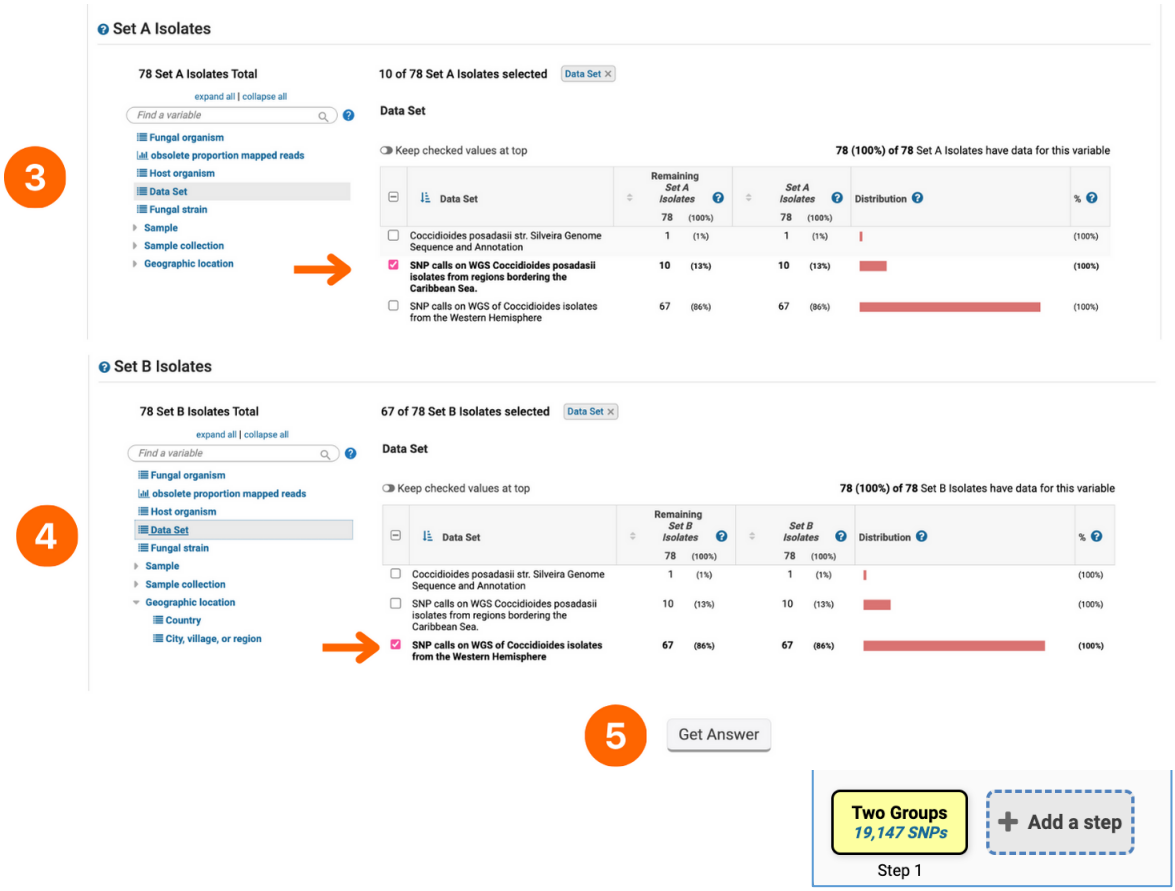
Create a strategy using the "Differences Between Two Groups of Isolates" search, which distinguishes SNPs between two groups of *C. posadasii* str Silveira isolates (A and B) collected in the Caribbean (group A) and western hemispheres (group B):

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

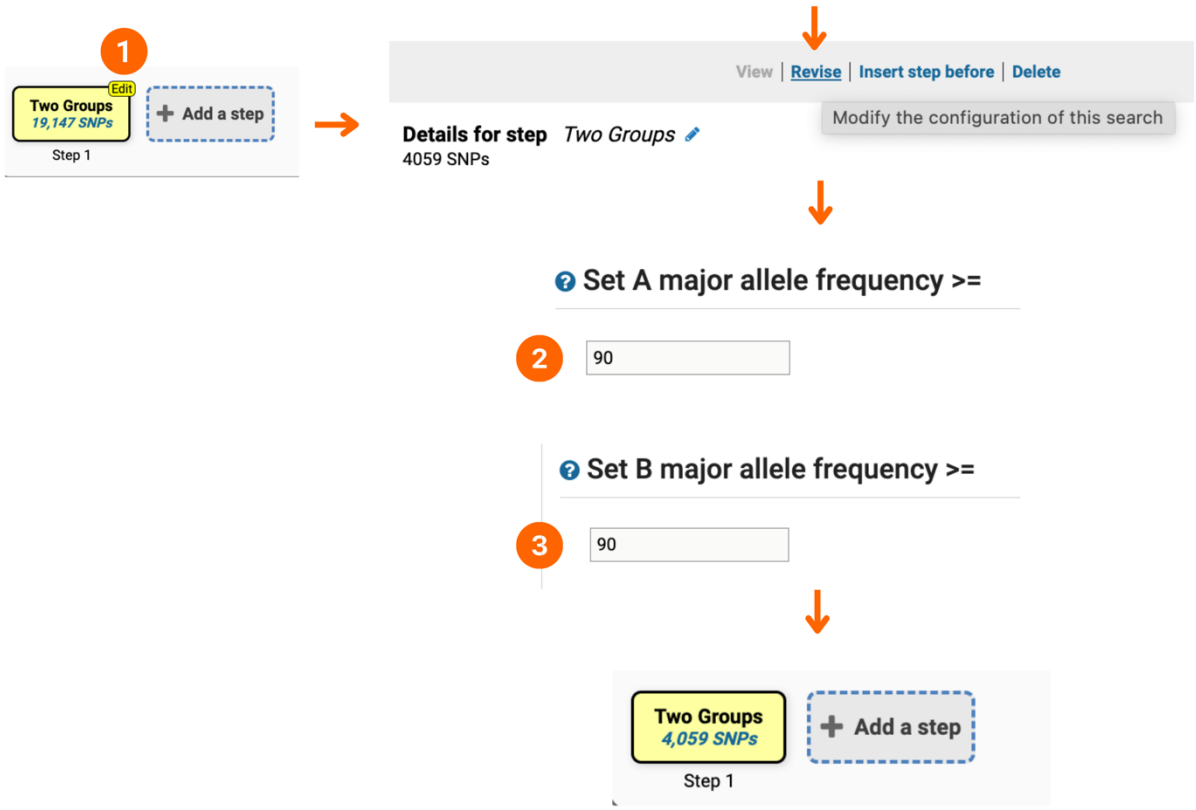
    1. Deploy the "Difference Between Two Groups of Isolates" search.
    2. Set the genome to *Coccidioides posadasii* strain Silveira.
    3. Select Set A isolates from the Data Set menu: Caribbean dataset.
    4. Select Set B isolates from the Data Set menu: Western hemisphere dataset.
    5. Click on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



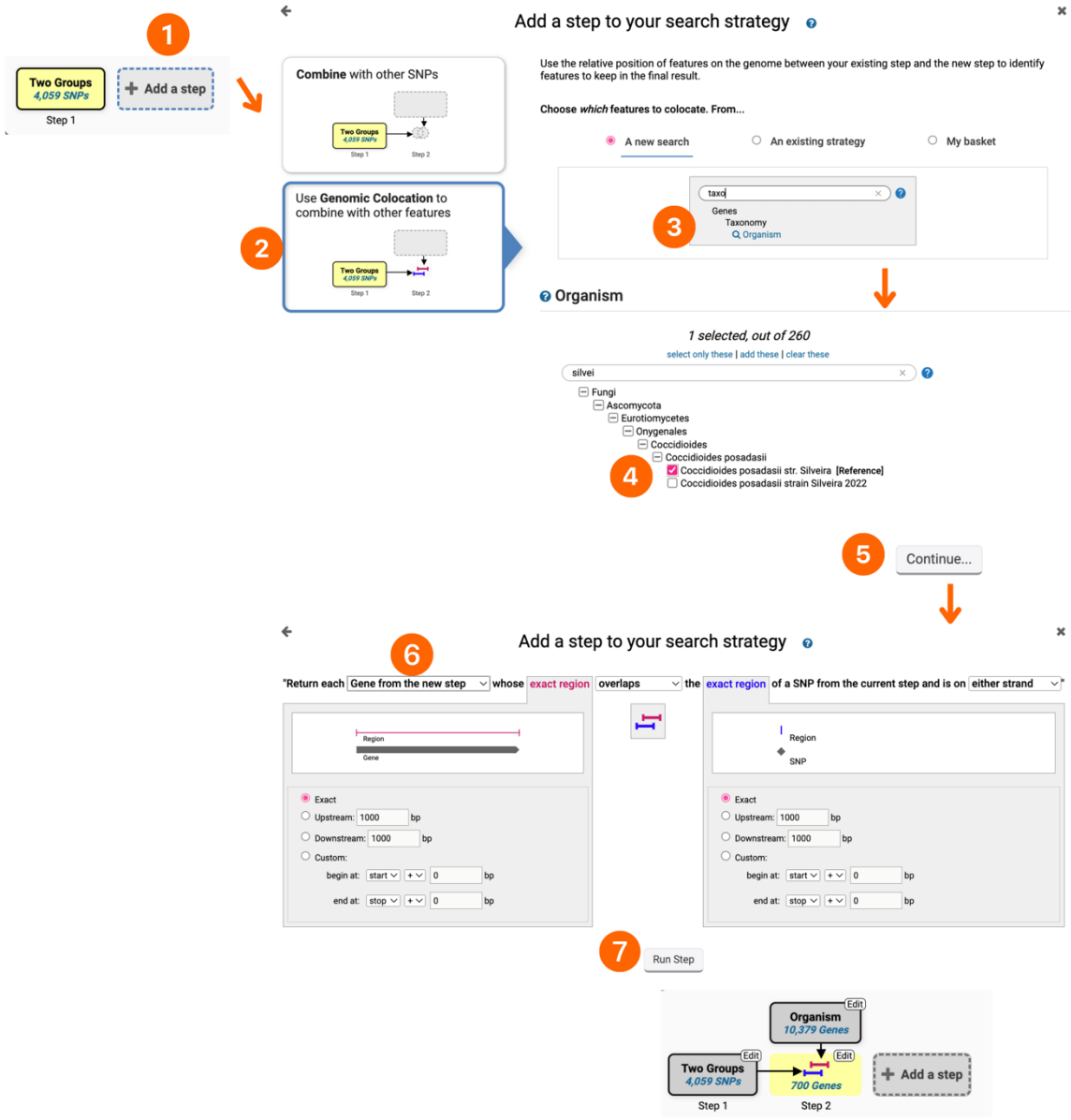




- **Change the stringency of your search to major allele frequency >= 90%**



- **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

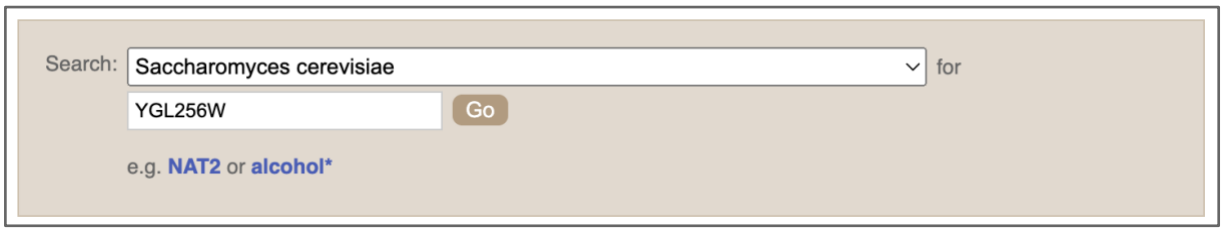
Exercise: Exploring variants in Ensembl Fungi

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

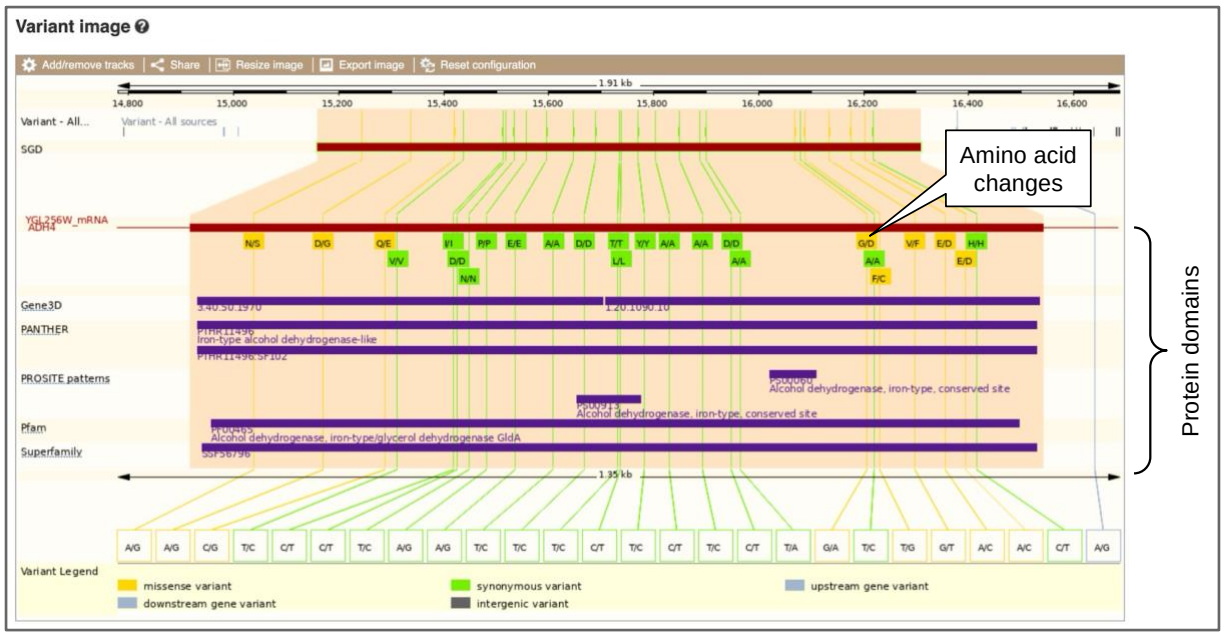
In any of the sequence views shown in the ‘Gene’ and ‘Transcript’ tabs, you can view variants on the sequence. You can do this by clicking on Configure this page 

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.



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.

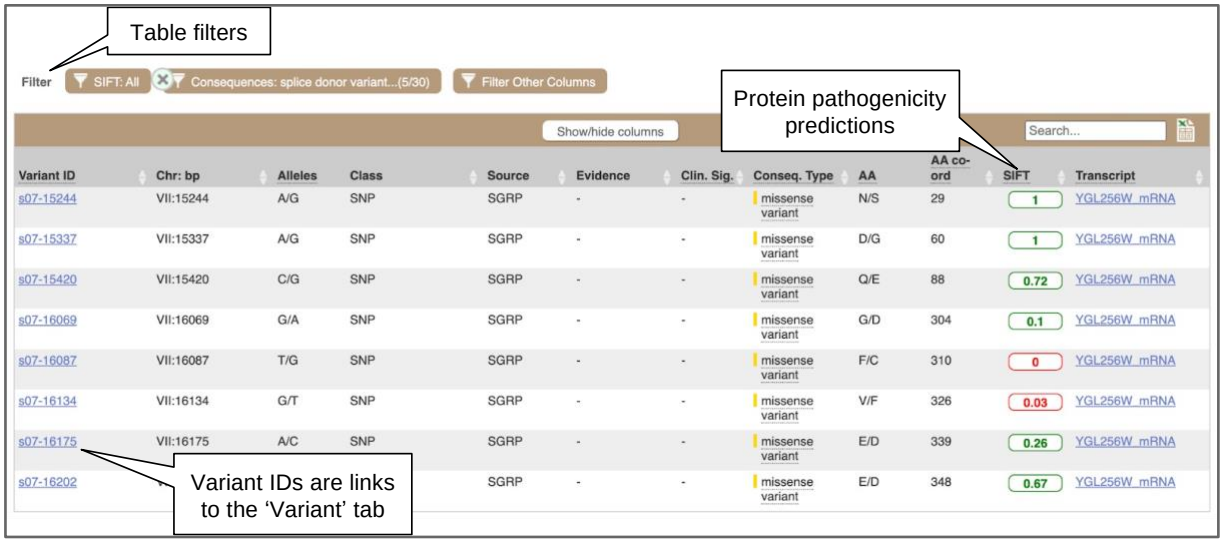
This table shows the variants in order of their occurrence throughout the genome, and they are reported on the forward strand. The gene *ADH4* is located on the forward strand, so we are first shown variants upstream of the gene (starting at the 5’ upstream region).

(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.

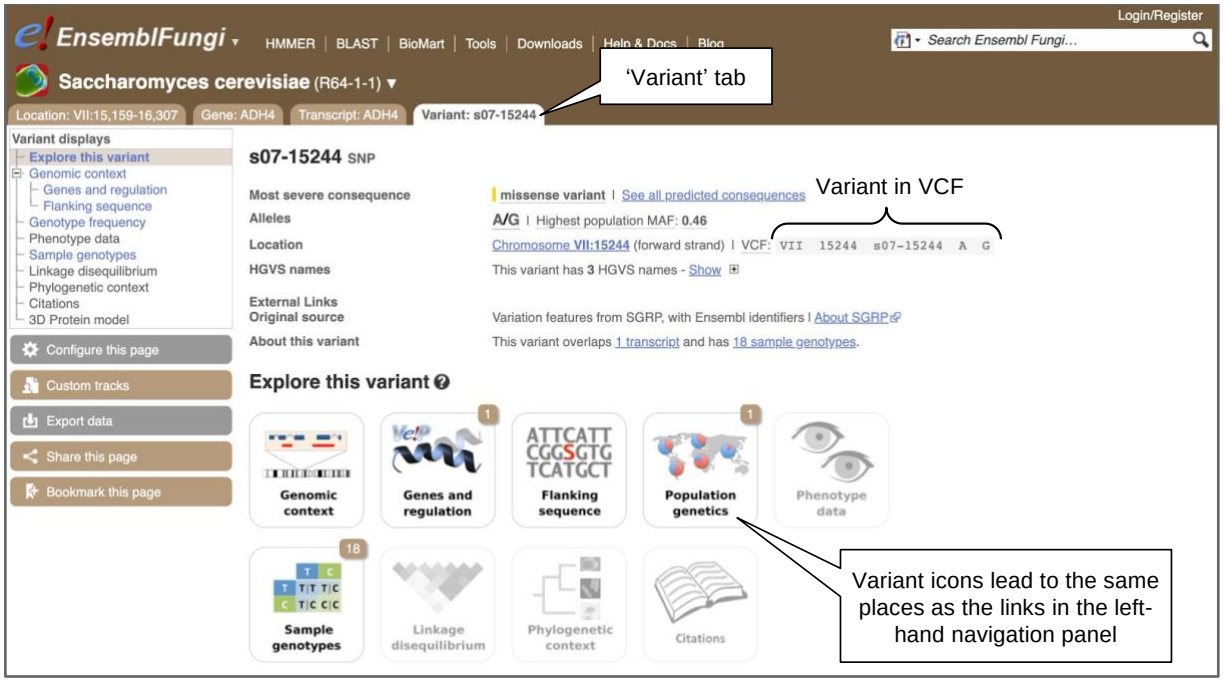
(b) Are there any known variants in this gene predicted to be deleterious?

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'.



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.

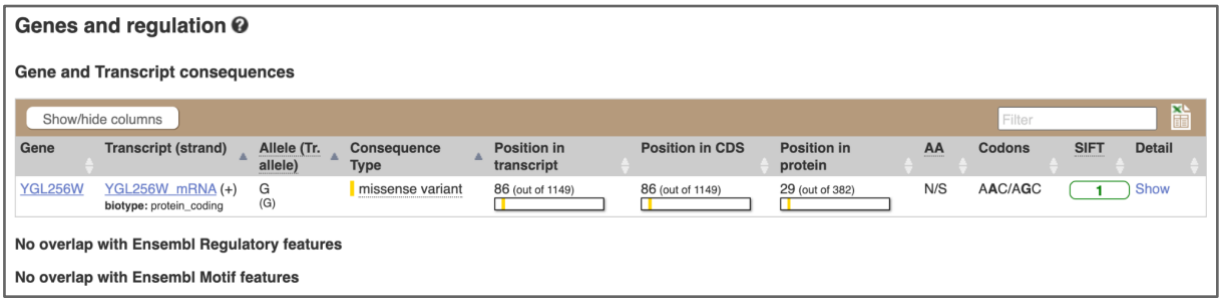
(c) What are the genomic coordinates of this variant?

(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?

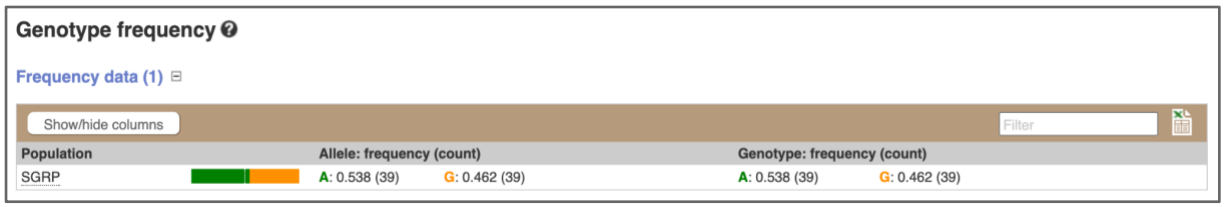
Click on the Genes and regulation icon, or follow the link in the left-hand panel.



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.

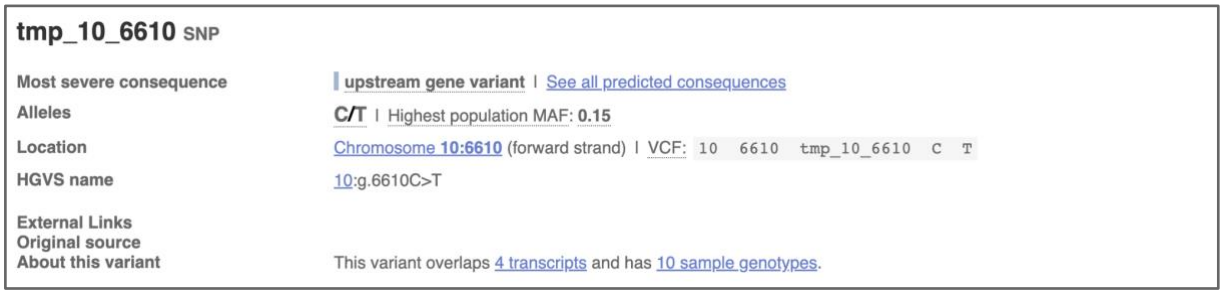


Additional Exercise – Variation data in *Fusarium oxysporum*

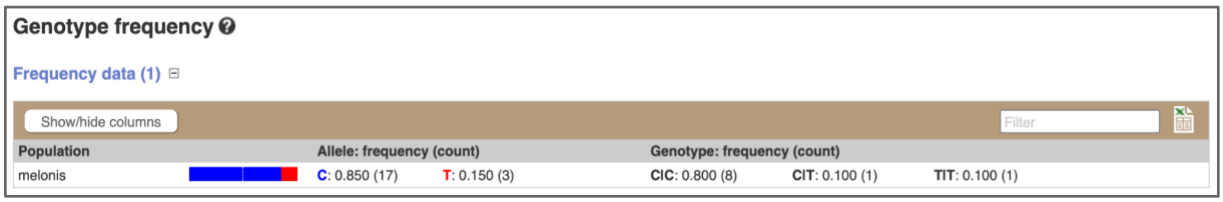
(a) Select the *Fusarium oxysporum* FO2 genome and search for FOXG_13574T0 gene. One of its upstream variants is SNP tmp_10_6610. What are the possible alleles for this polymorphic position? Which one is on the reference genome?
(b) What is the most frequent allele at this position? How many heterozygous individuals were observed in the melonis population?
(c) Which individuals have got genotypes C|T and T|T?

Answers

(a) 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.



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



(c) 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).

Sample genotypes

Search for a sample: Search (e.g. NA18507)

[back to top]

Genotypes for melonis

Show/hide columns Filter

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:

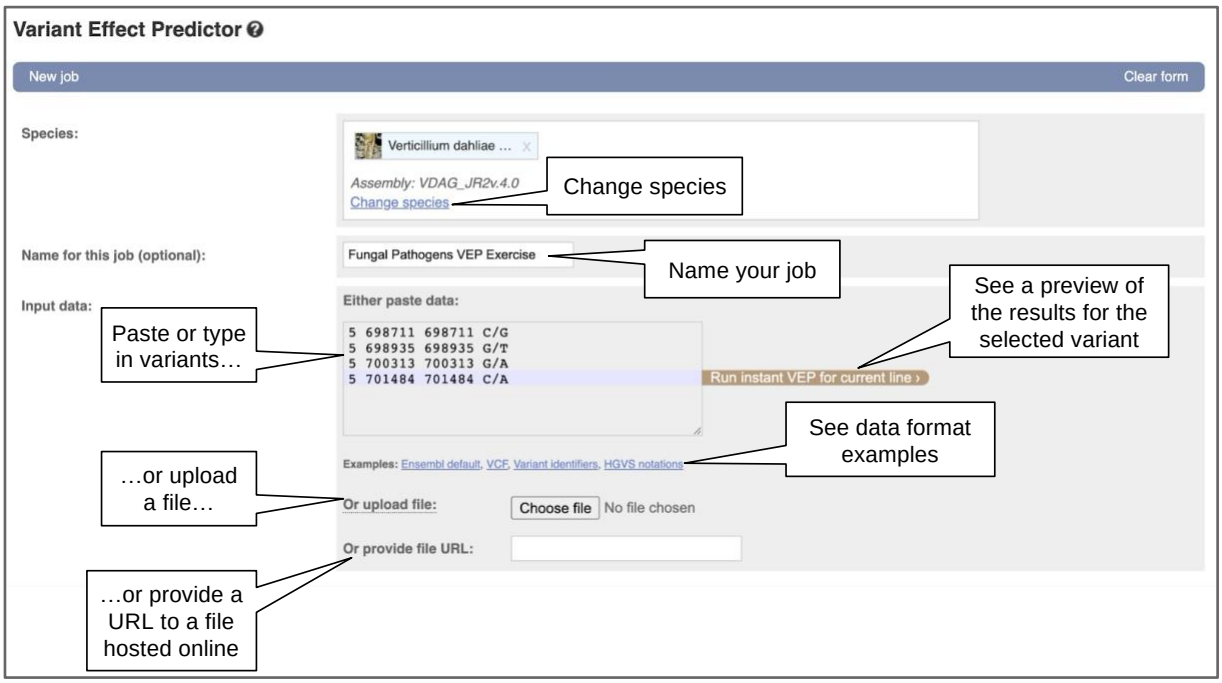
(a) Are your variants novel or have they already been annotated in Ensembl?
(b) What genes are affected by your variants?
(c) Do any of your variants affect gene regulation?

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.

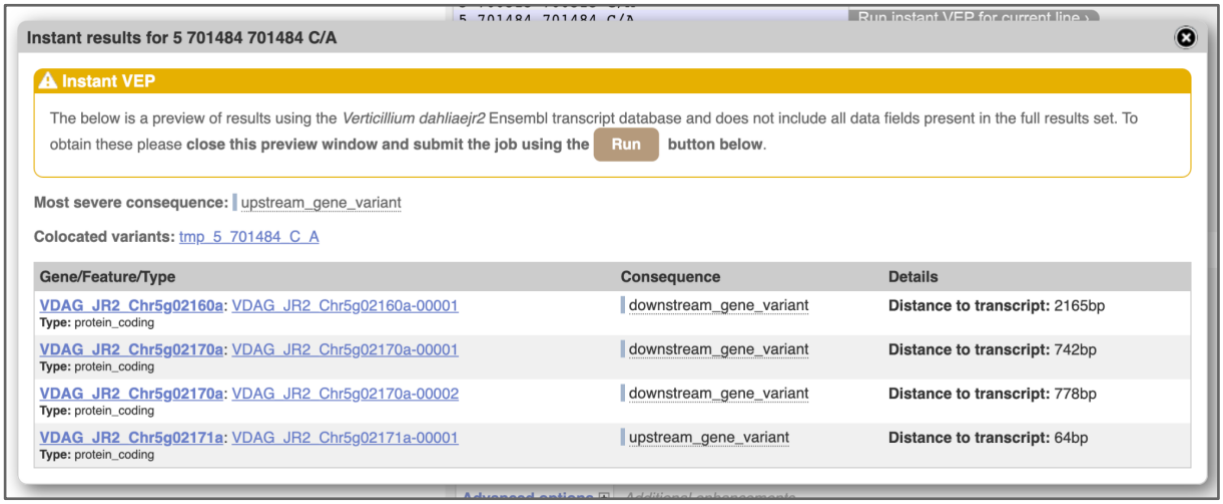
The VEP recognises a number of input formats including the Ensembl default format, variant call format (VCF), variant identifiers and HGVS notations. The HGVS nomenclature is a globally recognised standard for describing variants. You can read more about this here: https://hgvs-nomenclature.org/stable/.

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:

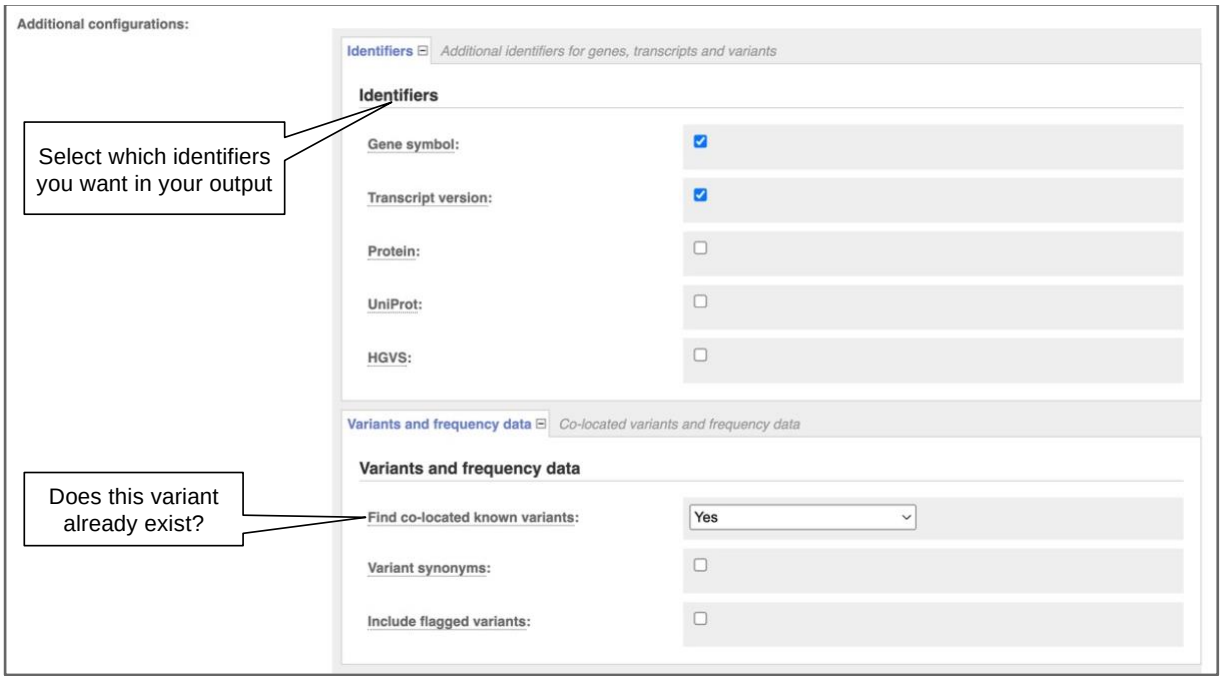
5 698711 698711 C/G
5 698935 698935 G/T
5 700313 700313 G/A
5 701484 701484 C/A

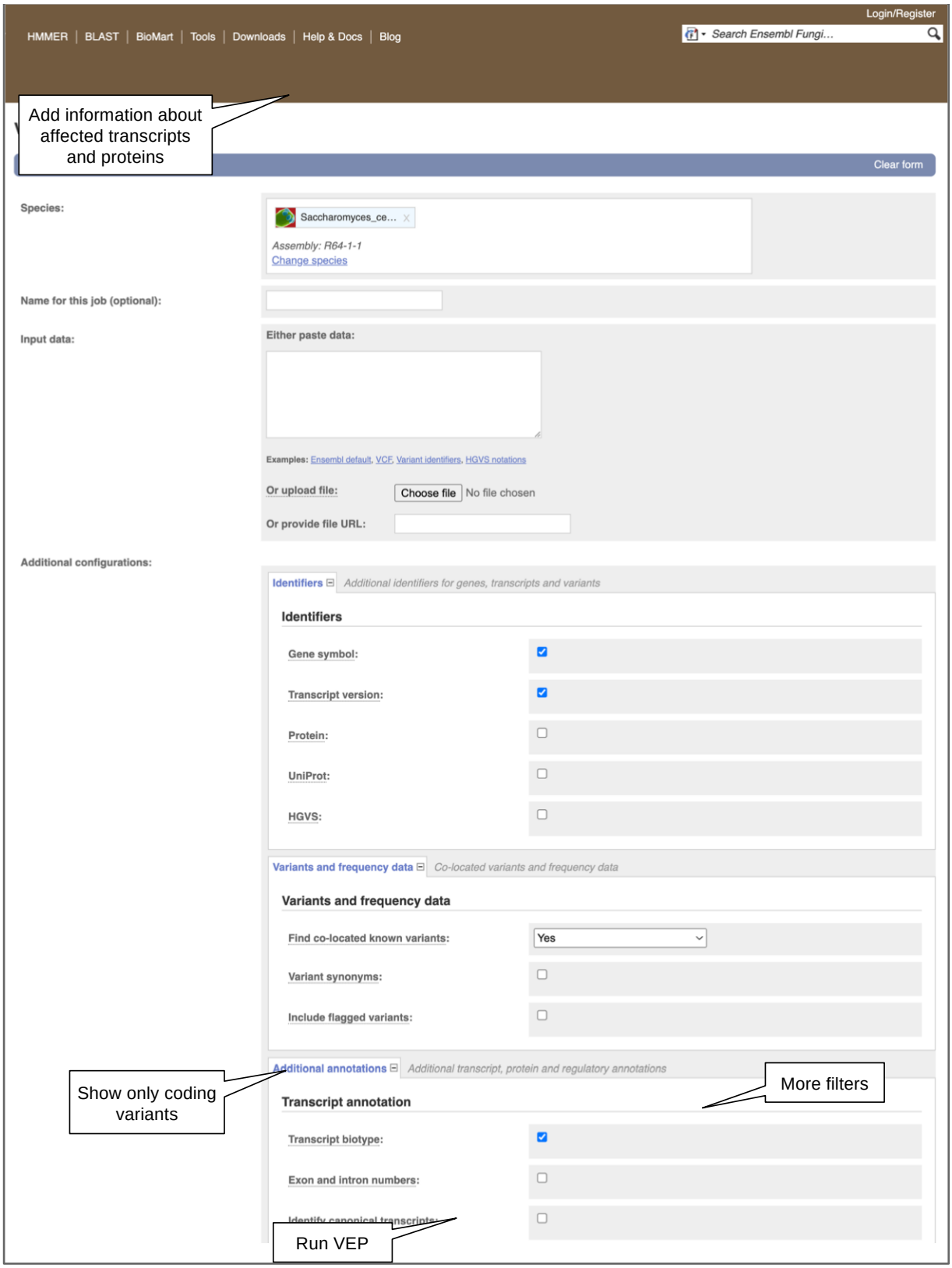


The VEP will automatically detect that the data is in Ensembl default format. Clicking on the Run instant VEP for current line option will generate a pop-up with summarised results for that individual variant.

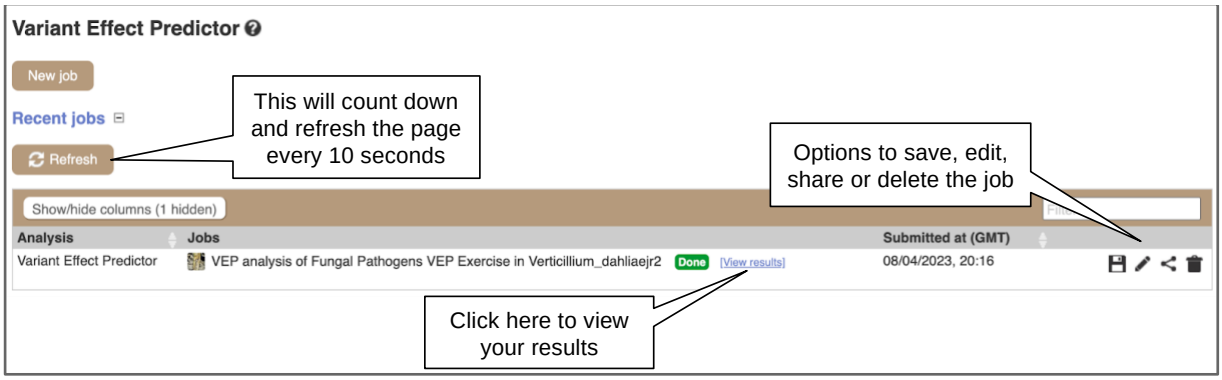


There are further options that you can choose for your output. These are categorised as Identifiers, Variants and frequency data, Additional annotations, Predictions, Filtering options and Advanced options. Let's open all the menus and take a look.





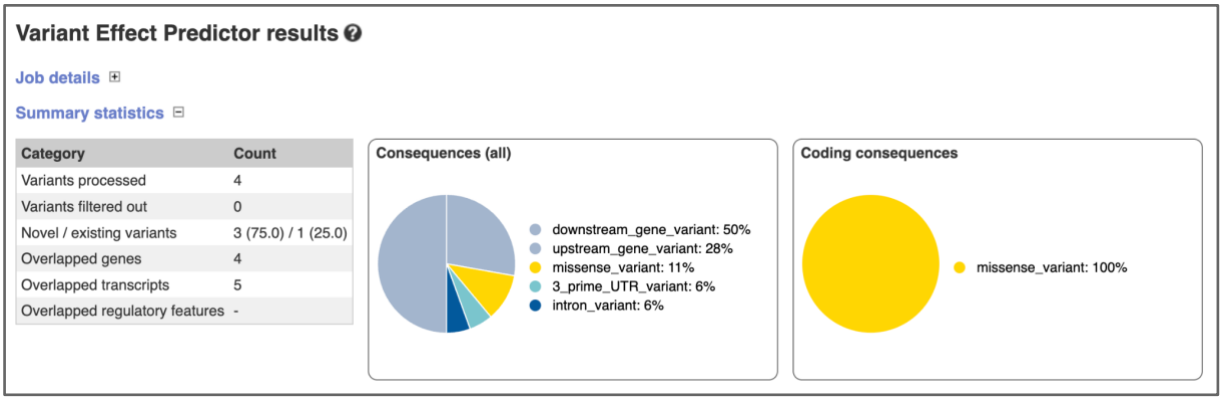
Hover over the options to see definitions. When you've selected everything you need, scroll to the bottom of the page and click Run.



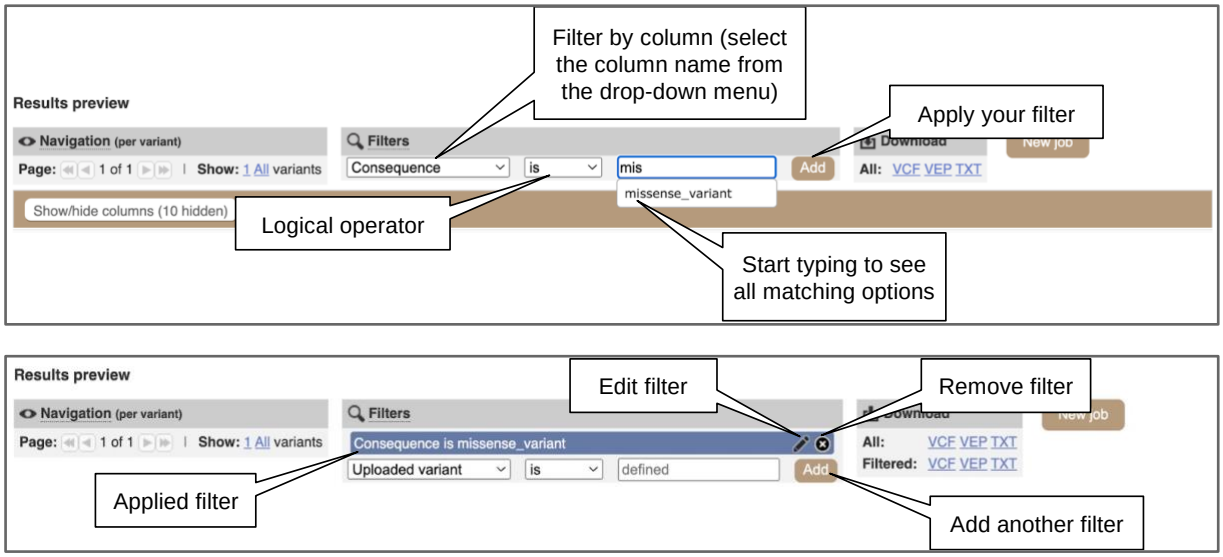
A table display will show you the status of your job. It will say Queued, then automatically switch to Done when the job is done, you do not need to refresh the page. You can edit or discard your job at this time. If you have submitted multiple jobs, they will all appear here.

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.

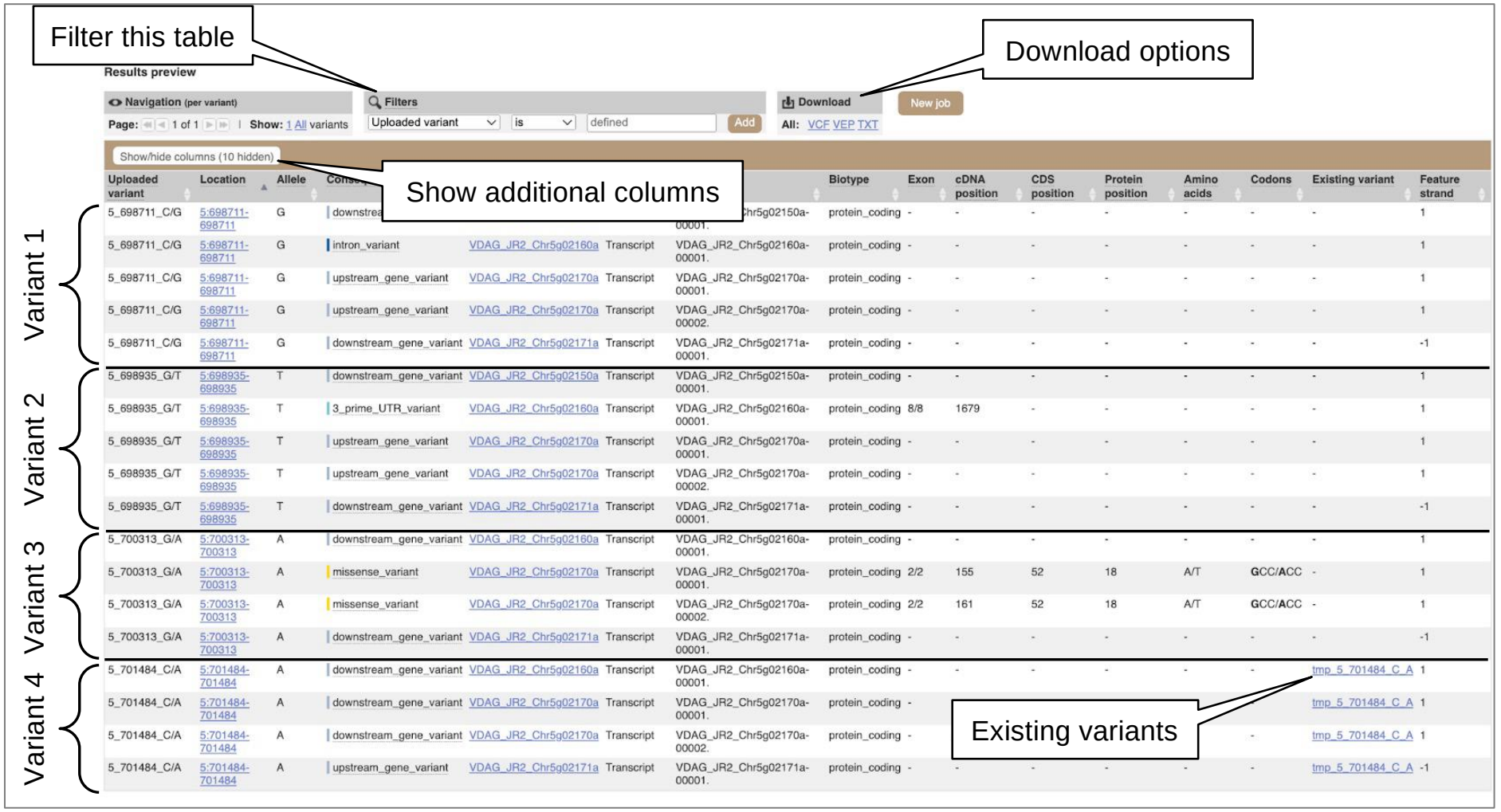
Let's come back to our questions:
(a) Are your variants novel or have they already been annotated in Ensembl?
(b) What genes are affected by your variants?
(c) Do any of your variants affect gene regulation?



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.



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.



Results preview

Navigation (per variant) | Page: 1 of 1 | Show: 1 All variants

Filters: Uploaded variant | is | defined | Add

Download | New job | All: VCF VEP TXT

Show/hide columns (10 hidden)

Uploaded variant	Location	Allele	Consequence			Biotype	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand
5_698711_C/G	5:698711-698711	G	downstrea		hr5g02150a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5:698711-698711	G	intron_variant	VDAG_JR2_Chr5g02160a	Transcript VDAG_JR2_Chr5g02160a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5:698711-698711	G	upstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript VDAG_JR2_Chr5g02170a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5:698711-698711	G	upstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript VDAG_JR2_Chr5g02170a-00002.	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5:698711-698711	G	downstream_gene_variant	VDAG_JR2_Chr5g02171a	Transcript VDAG_JR2_Chr5g02171a-00001.	protein_coding	-	-	-	-	-	-	-	-1
5_698935_G/T	5:698935-698935	T	downstream_gene_variant	VDAG_JR2_Chr5g02150a	Transcript VDAG_JR2_Chr5g02150a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_698935_G/T	5:698935-698935	T	3_prime_UTR_variant	VDAG_JR2_Chr5g02160a	Transcript VDAG_JR2_Chr5g02160a-00001.	protein_coding	8/8	1679	-	-	-	-	-	1
5_698935_G/T	5:698935-698935	T	upstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript VDAG_JR2_Chr5g02170a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_698935_G/T	5:698935-698935	T	upstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript VDAG_JR2_Chr5g02170a-00002.	protein_coding	-	-	-	-	-	-	-	1
5_698935_G/T	5:698935-698935	T	downstream_gene_variant	VDAG_JR2_Chr5g02171a	Transcript VDAG_JR2_Chr5g02171a-00001.	protein_coding	-	-	-	-	-	-	-	-1
5_700313_G/A	5:700313-700313	A	downstream_gene_variant	VDAG_JR2_Chr5g02160a	Transcript VDAG_JR2_Chr5g02160a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_700313_G/A	5:700313-700313	A	missense_variant	VDAG_JR2_Chr5g02170a	Transcript VDAG_JR2_Chr5g02170a-00001.	protein_coding	2/2	155	52	18	A/T	GCC/ACC	-	1
5_700313_G/A	5:700313-700313	A	missense_variant	VDAG_JR2_Chr5g02170a	Transcript VDAG_JR2_Chr5g02170a-00002.	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_Chr5g02171a	Transcript VDAG_JR2_Chr5g02171a-00001.	protein_coding	-	-	-	-	-	-	-	-1
5_701484_C/A	5:701484-701484	A	downstream_gene_variant	VDAG_JR2_Chr5g02160a	Transcript VDAG_JR2_Chr5g02160a-00001.	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A	1
5_701484_C/A	5:701484-701484	A	downstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript VDAG_JR2_Chr5g02170a-00001.	protein_coding	-						tmp_5_701484_C_A	1
5_701484_C/A	5:701484-701484	A	downstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript VDAG_JR2_Chr5g02170a-00002.	protein_coding	-					-	tmp_5_701484_C_A	1
5_701484_C/A	5:701484-701484	A	upstream_gene_variant	VDAG_JR2_Chr5g02171a	Transcript VDAG_JR2_Chr5g02171a-00001.	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A	-1

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

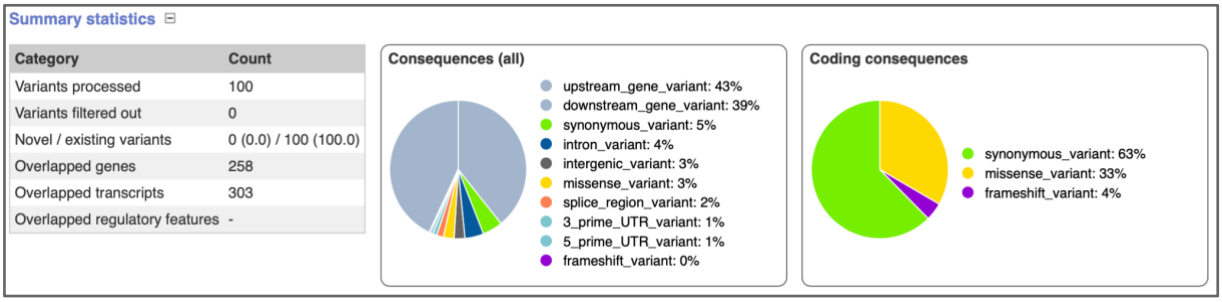
On the course file page, you will find a VCF file labelled VEP_exercise.vcf. This is a small subset of the outcome of *Puccinia graminis* (Ug99) whole genome sequencing and variant calling experiment. This file can also be found on our FTP site under the following link: http://ftp.ebi.ac.uk/pub/databases/ensembl/training/2021/FungalPathogens/VEP_exercise.vcf

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 HGVSp 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

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

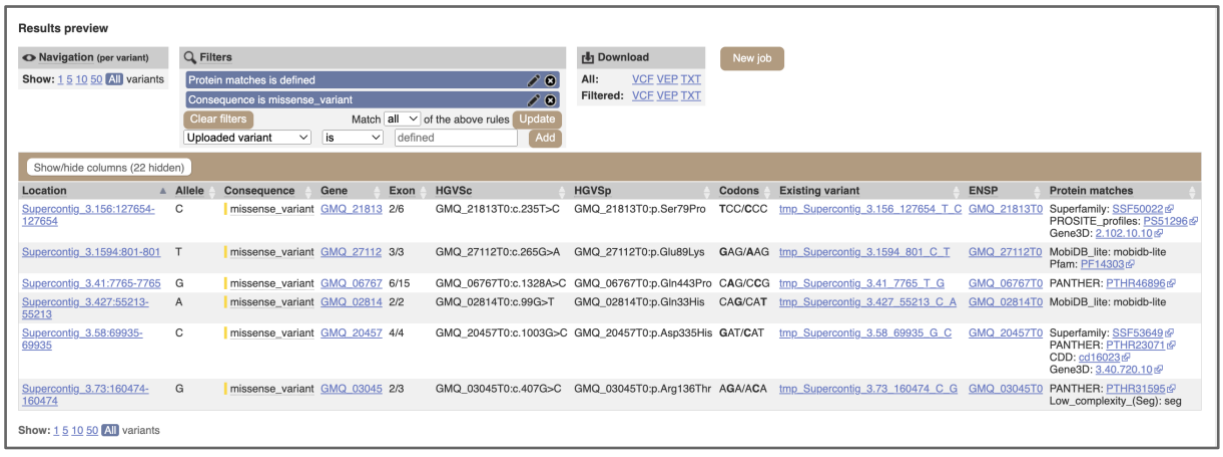
Identifiers: Protein (to include protein position information), HGVS (to include the HGVS notations)
Additional annotations: Protein matches (to include any overlapping protein domains)

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:

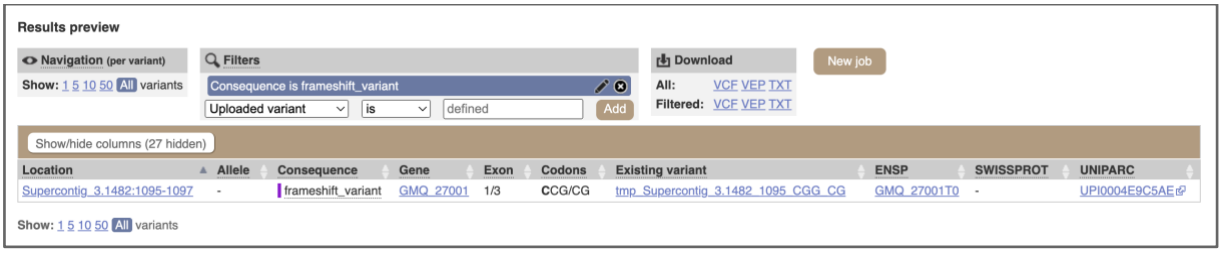
GMQ_21813T0:p.Ser79Pro
GMQ_27112T0:p.Glu89Lys
GMQ_06767T0:p.Gln443Pro
GMQ_02814T0:p.Gln33His
GMQ_20457T0:p.Asp335His
GMQ_03045T0:p.Arg136Thr



(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.

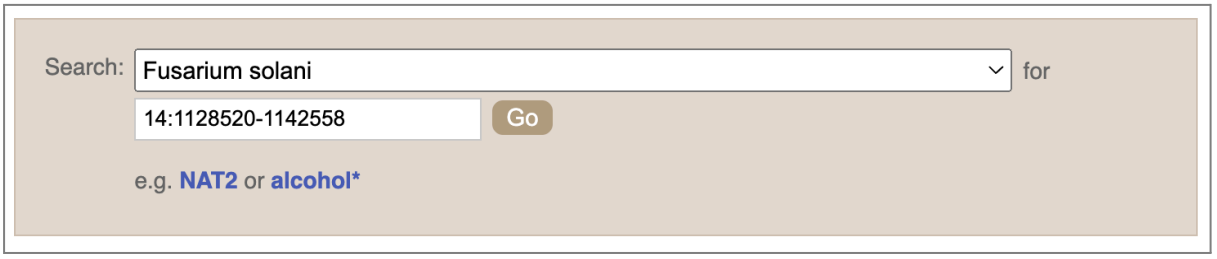


Exercise: Ensembl Fungi whole-genome alignments

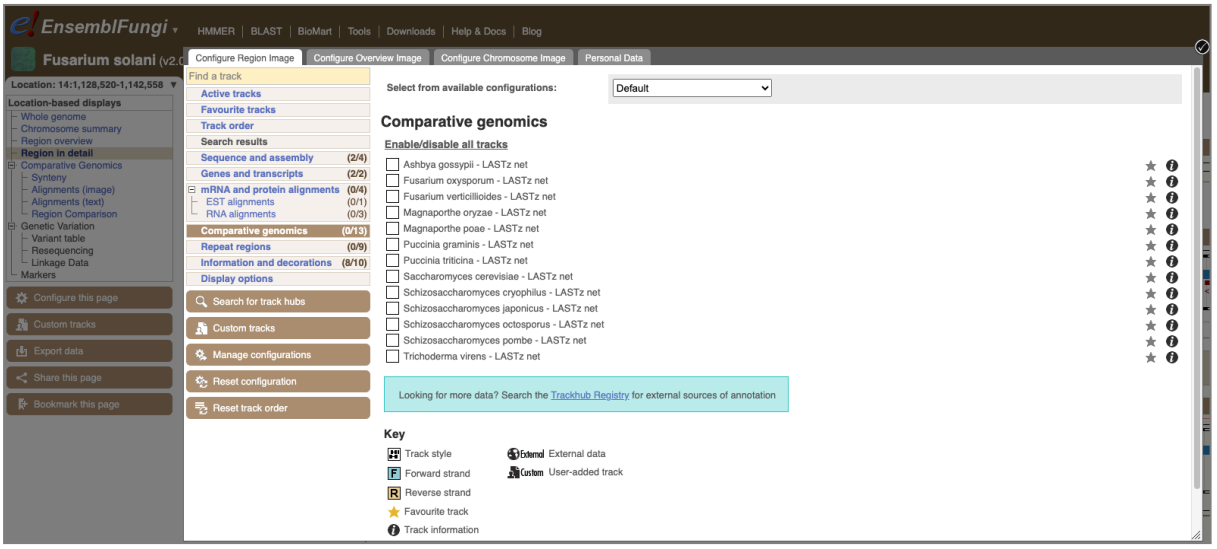
Links to be clicked shown in blue, text to be entered shown in red.

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.

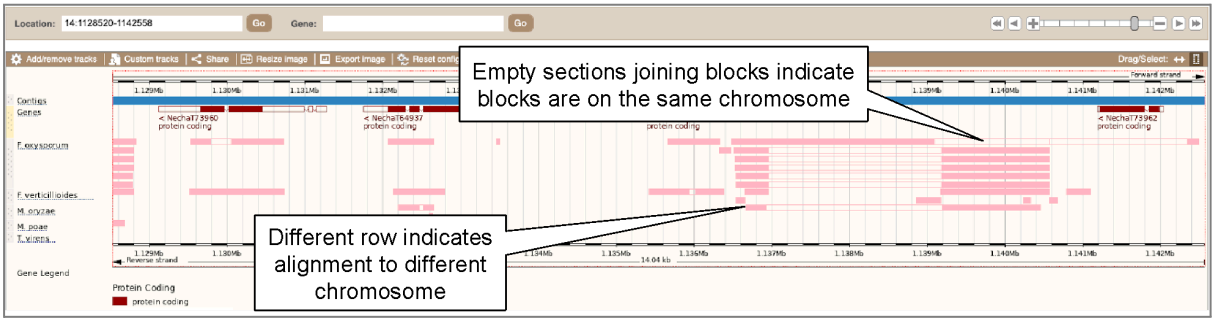
(a) Find the region 14:1128520-1142558 in *Fusarium solani* and go to the Region in detail page. This region includes four genes we identified from our first BioMart query: *PEP5*, *PDA1*, *ESP3* and *PEP5*.



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.

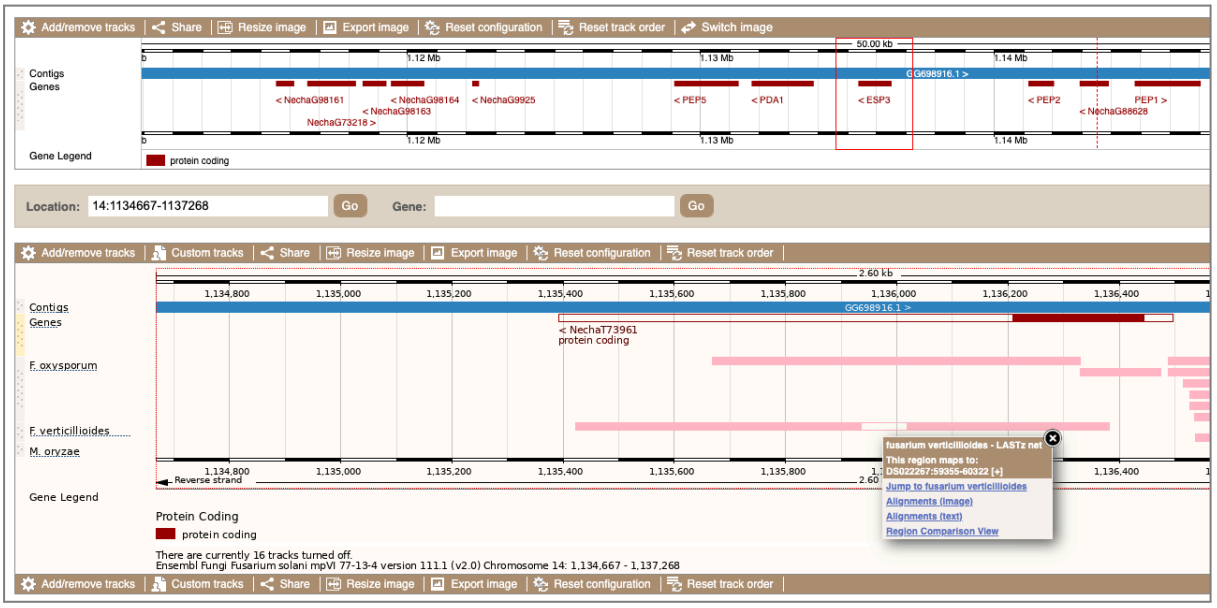


We can now see some pink alignments shown on the display. Alignments to the same chromosome are presented in a single row, and gaps in the alignment are shown by linking blocks. If there are alignments to multiple chromosomes in the aligned species these are represented on different rows.



(b) Looking at the pink alignment blocks, does this region in F. solani s align to multiple different chromosomes in the other species?

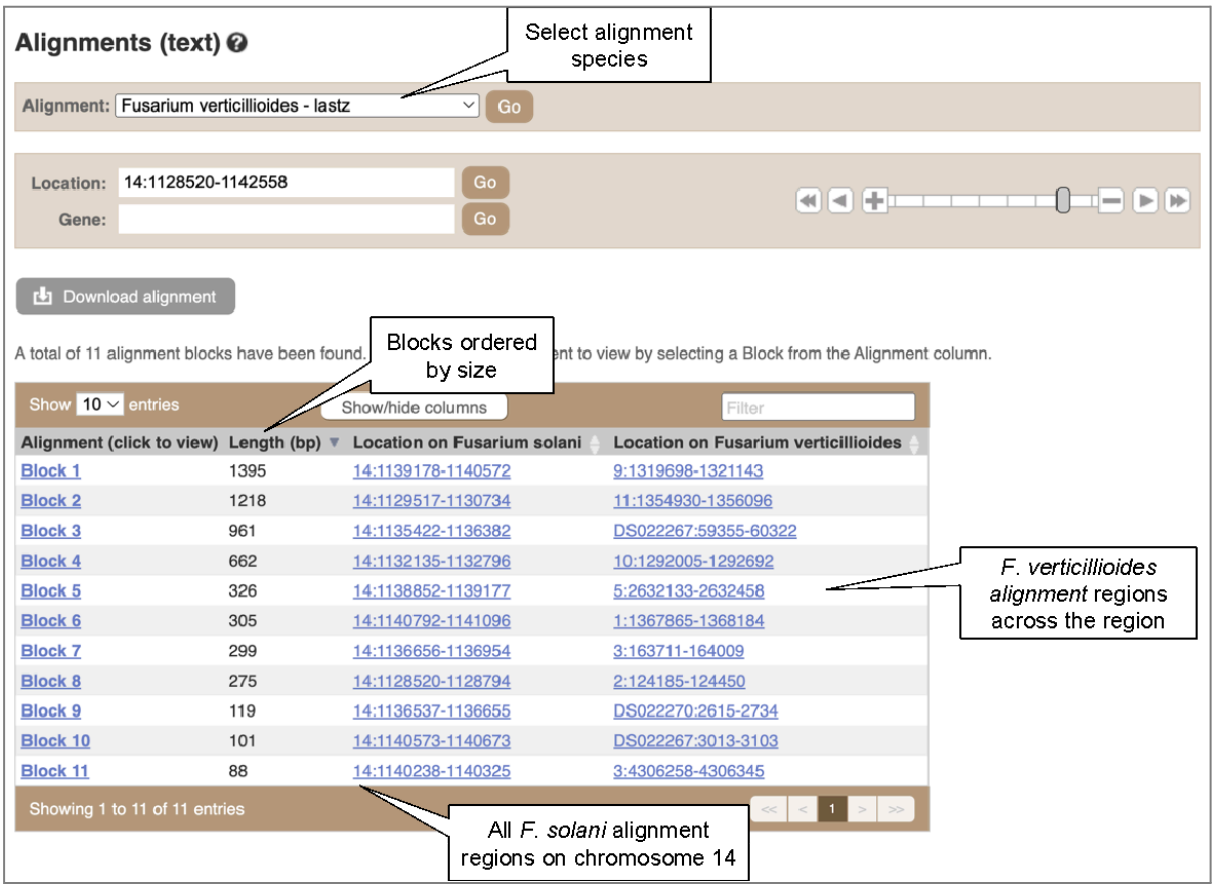
(c) Which chromosome(s) does the *F. solani ESP3* gene align to in *F. verticillioides*?



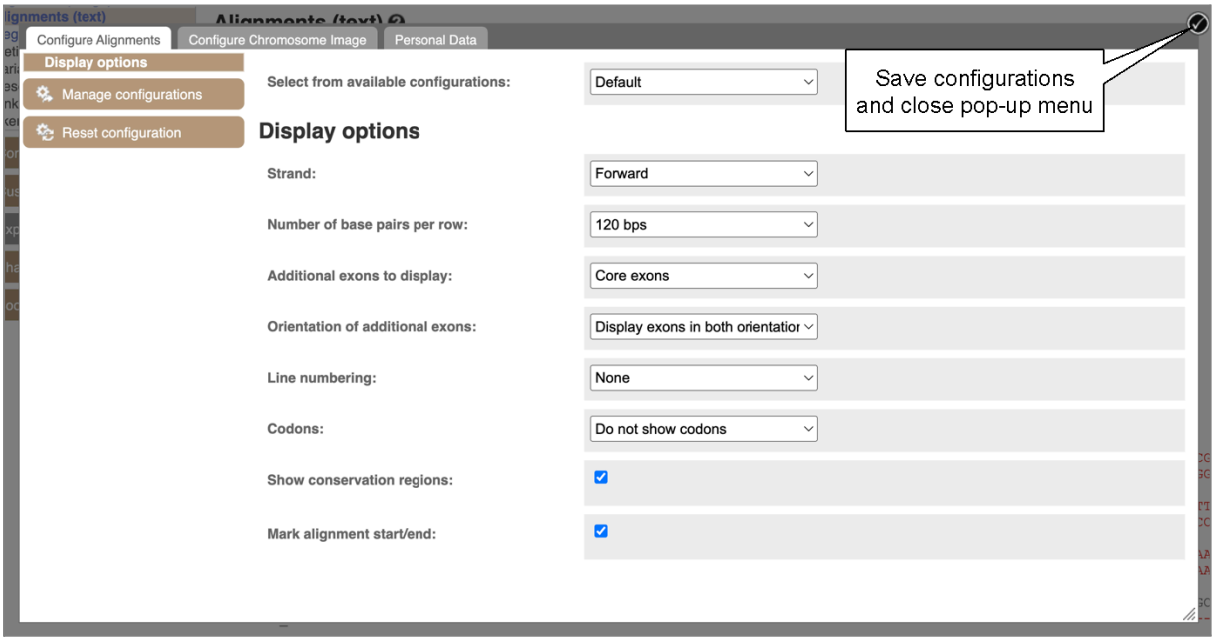
We can see that alignments in this region are quite poor for these species, with alignments spanning different chromosomes. This supports the lack of orthologues between these species.

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.

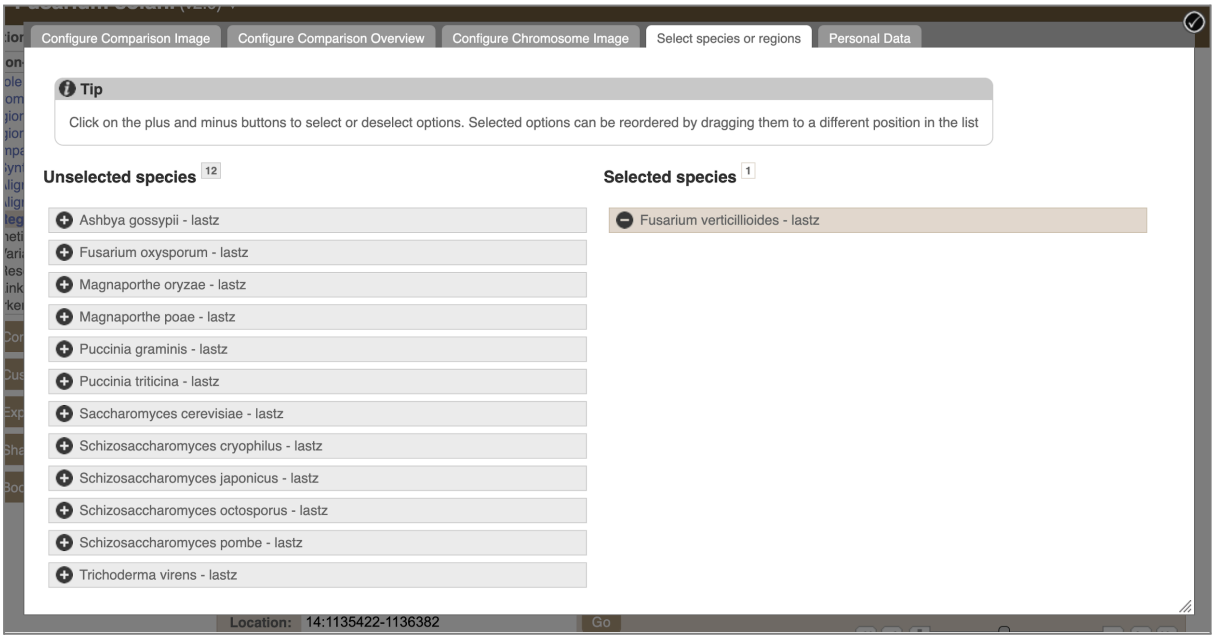


Let's click on Block 3. This takes you to a new page with a sample of the aligned sequence. Then click the button to Display full alignment. You will see a list of the regions aligned, followed by the sequence alignment. Exons are shown in red. Click on Configure this page, you can turn on the options 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.

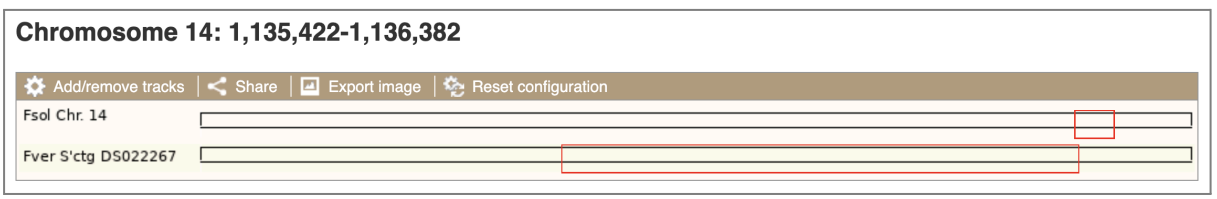


To view an image of the alignments, click on Region comparison in the left-hand navigation panel. This view is like the 'Region in detail' page as it shows three images of the genome at different scales. You can add multiple species to this view.

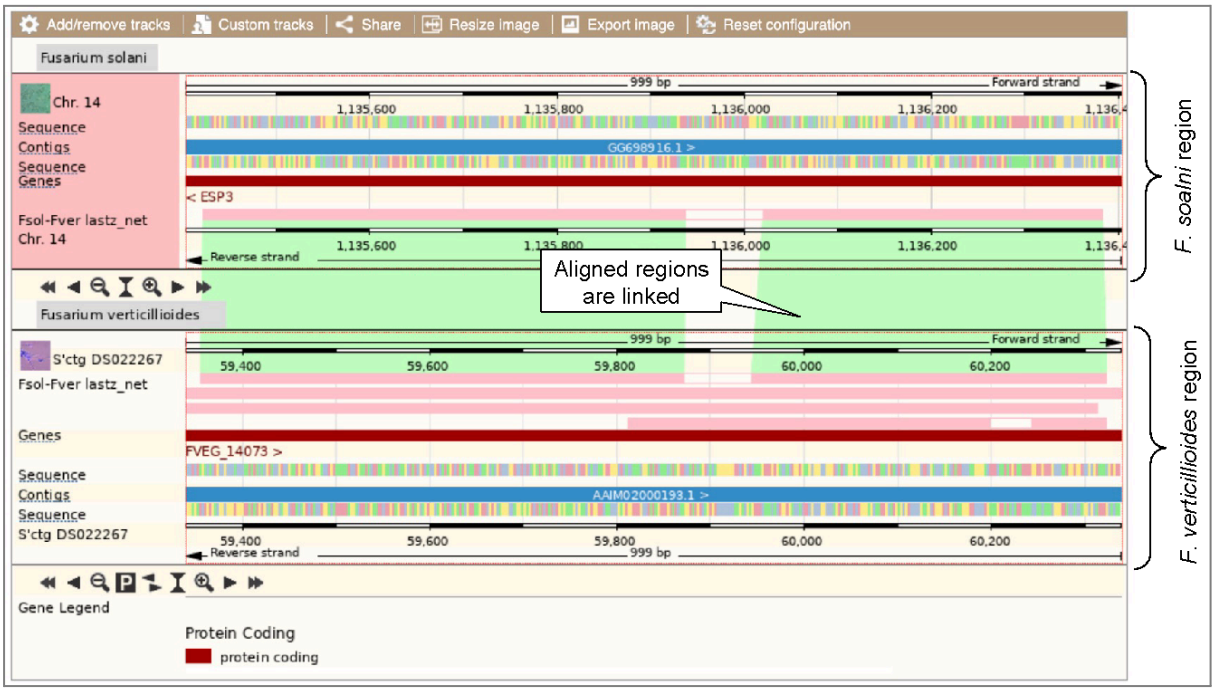
Click on the brown Select species or regions button. In the pop-up menu, select *Fusarium verticillioides* from the list. Close the window.



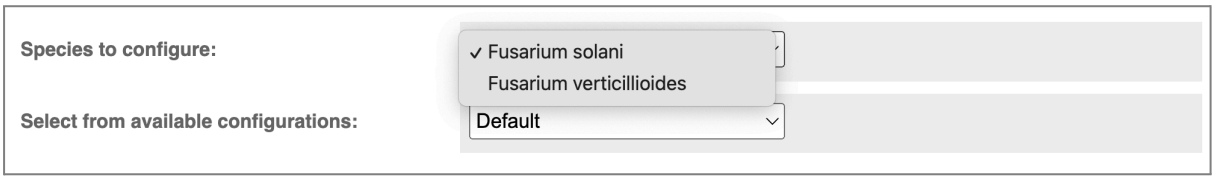
This page, similar to the region in detail page, shows the chromosome positions first. We can see the location of this alignment on the scaffold in *F. verticillioides*.



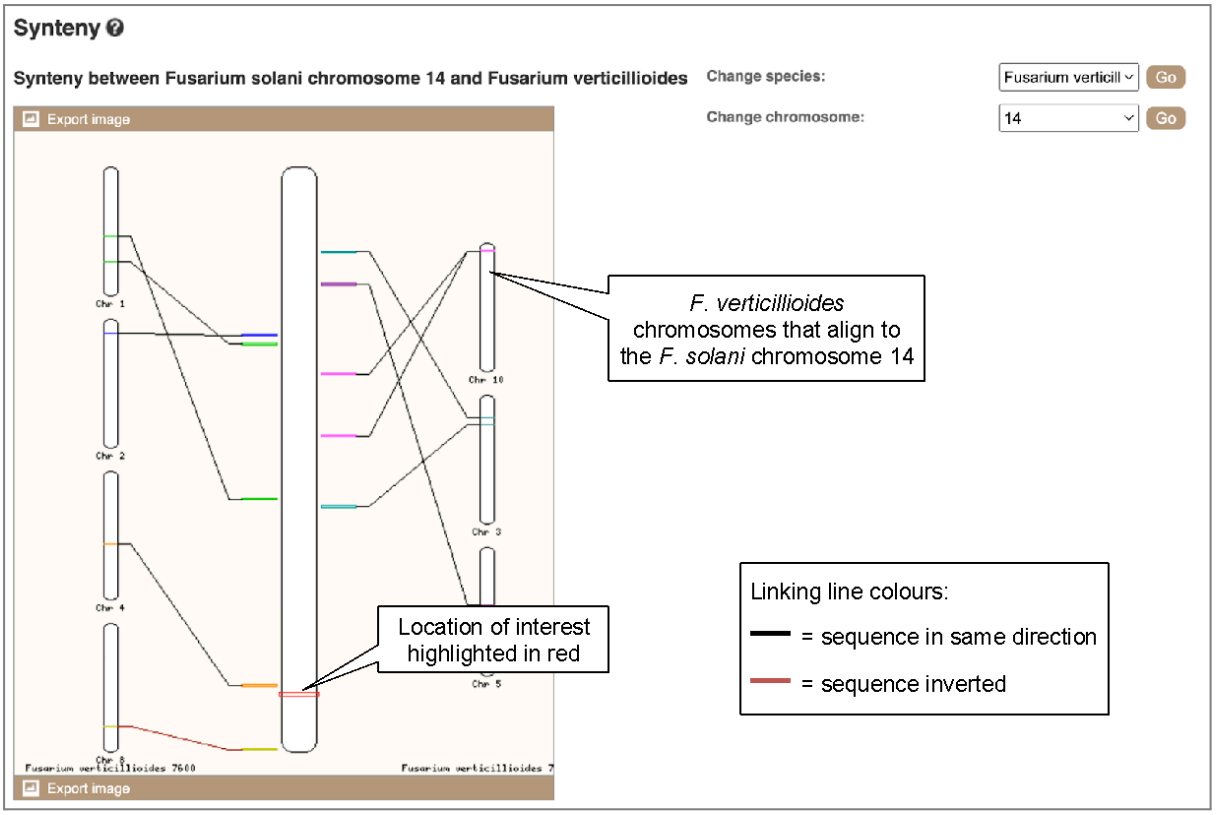
Scroll down to the most detailed image. An example image (of another alignment block) is below, and you should see something similar on your browser.



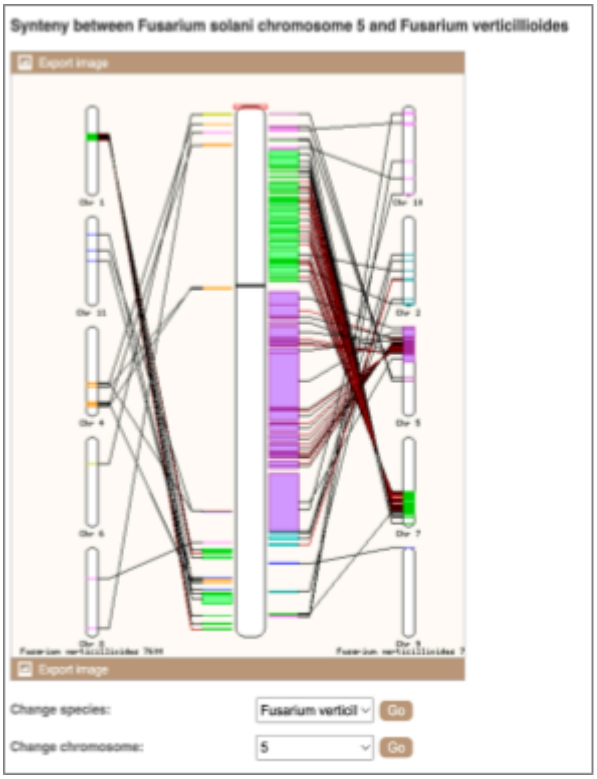
You can add data to both of these views with the same options you had in the 'Region in detail' page. Click on Configure this page and look at the top of the menu.

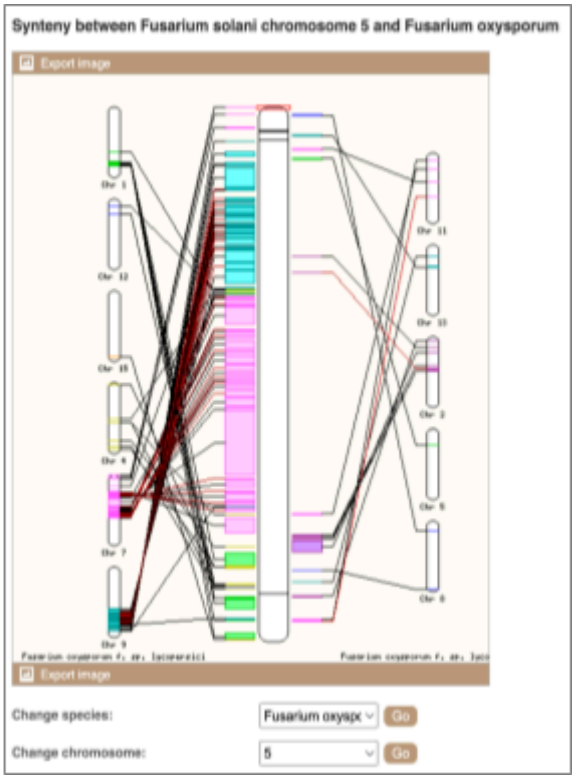


We can view chromosomal rearrangements in the 'Synteny' view. Click on Synteny in the left-hand navigation panel. (*note* - there is a bug causing this view to fail to load in later Ensembl versions, you can access the view in this archive - https://feb2023-fungi.ensembl.org/Fusarium_solani/Location/Synteny?db=core;otherspecies=Fusarium_verticillioides;r=14:1128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



(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*?





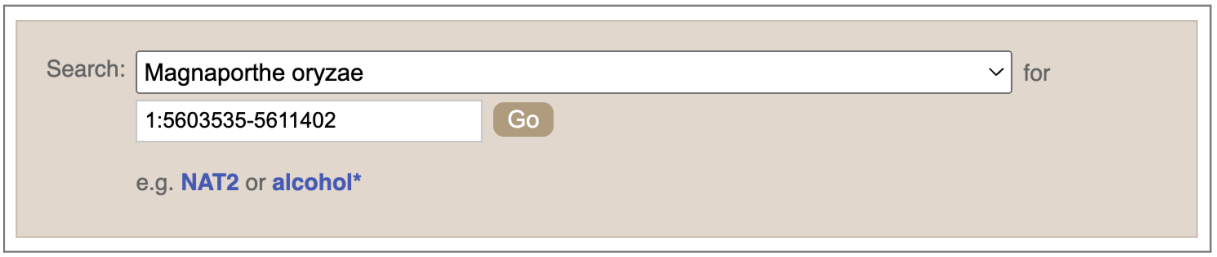
Additional Exercise - Rearrangements in *Magnaporthe* species

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*.

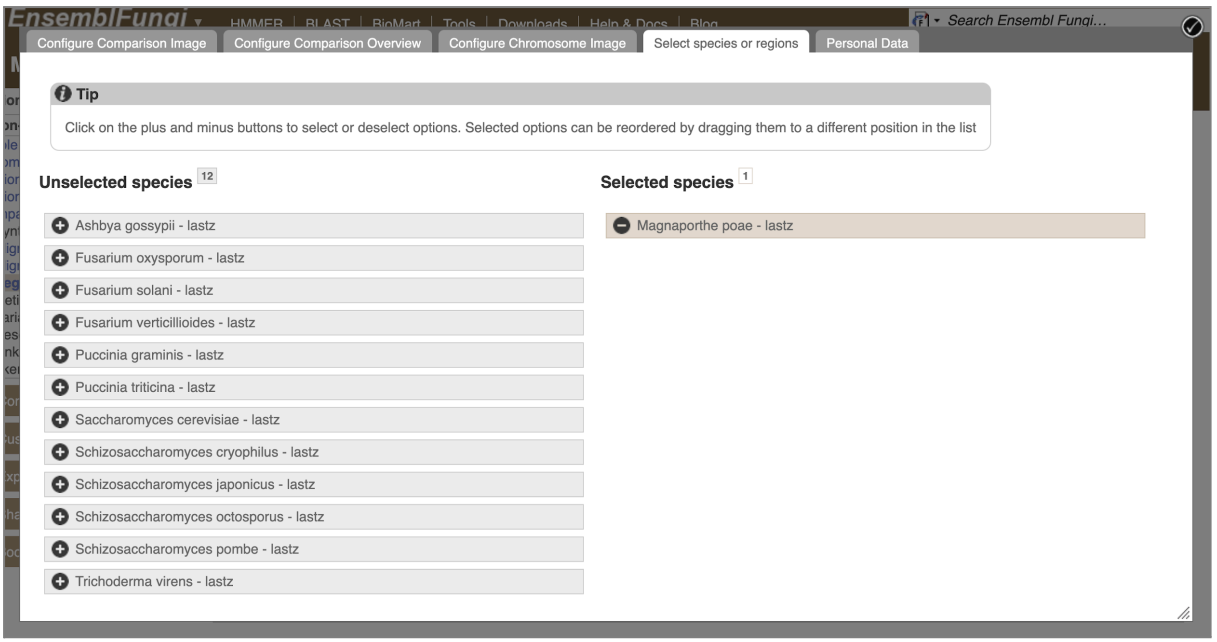
(a) Search for the region 1:5603535-5611402 in *Magnaporthe oryzae*.
(b) Click on Region comparison and choose *Magnaporthe poae* from the Select species or regions pop-up to display an alignment.
(c) Scroll down to the most detailed image. To what region (chromosome/scaffold/contig) does this region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

Answer - Rearrangements in *Magnaporthe* species

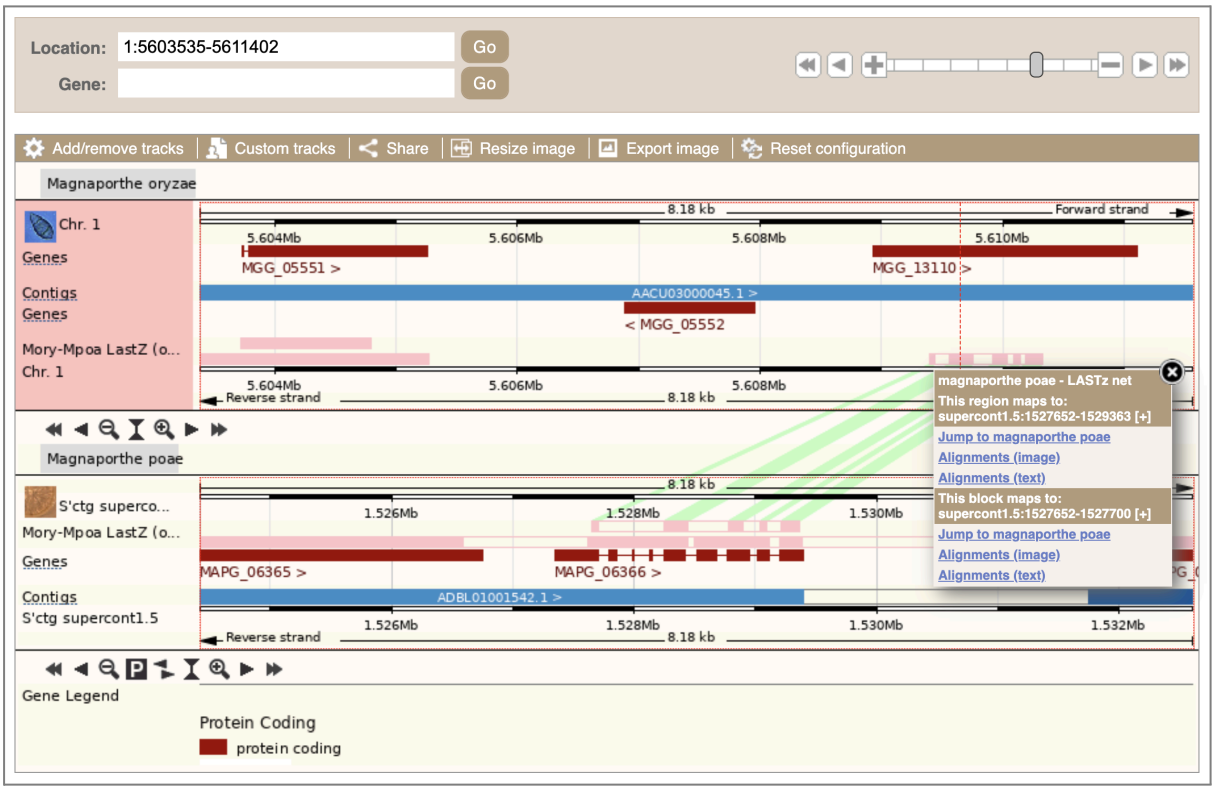
(a) Go to fungi.ensembl.org in your browser and search for the region 1:5603535-5611402 in *Magnaporthe oryzae*.



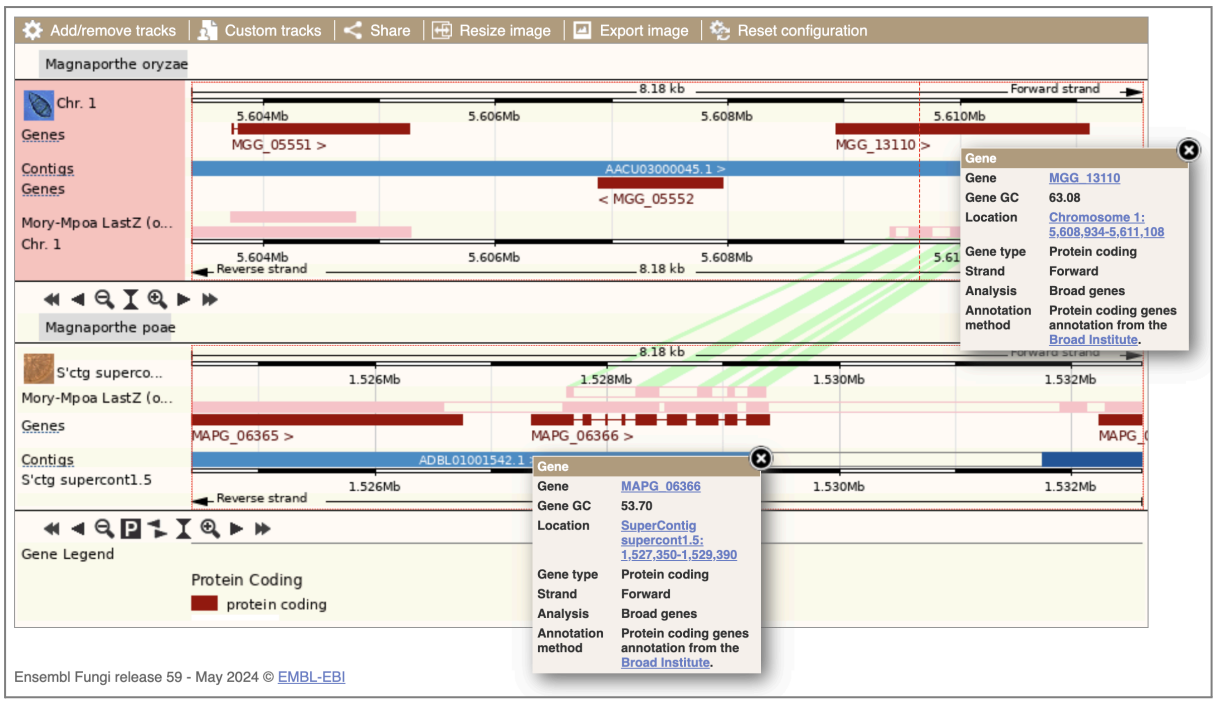
(b) Click on Region Comparison in the left-hand panel. Click on Select species or regions and select *Magnaporthe poae* - lastz in the pop-up menu to display the alignment.



(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.


Add/remove tracks
Custom tracks
Share
Resize image
Export image
Reset configuration
Magnaporthe oryzae
Chr. 1
Genes
Contigs
Genes
Mory-Mpoa LastZ (o...
Chr. 1
8.18 kb
Forward strand
5.604Mb
5.606Mb
5.608Mb
5.610Mb
MGG_05551 >
MGG_13110 >
AACU03000045.1 >
< MGG_05552
Reverse strand
Gene
Gene MGG_13110
Gene GC 63.08
Location Chromosome 1: 5,608,934-5,611,108
Gene type Protein coding
Strand Forward
Analysis Broad genes
Annotation method Protein coding genes annotation from the Broad Institute.
Magnaporthe poae
S'ctg superco...
Mory-Mpoa LastZ (o...
Genes
Contigs
S'ctg supercont1.5
1.526Mb
1.528Mb
1.530Mb
1.532Mb
MAPG_06365 >
MAPG_06366 >
ADBL01001542.1 >
Gene
Gene MAPG_06366
Gene GC 53.70
Location SuperContig supercont1.5: 1,527,350-1,529,390
Gene type Protein coding
Strand Forward
Analysis Broad genes
Annotation method Protein coding genes annotation from the Broad Institute.
Gene Legend
Protein Coding
protein coding
Ensembl Fungi release 59 - May 2024 © EMBL-EBI

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/

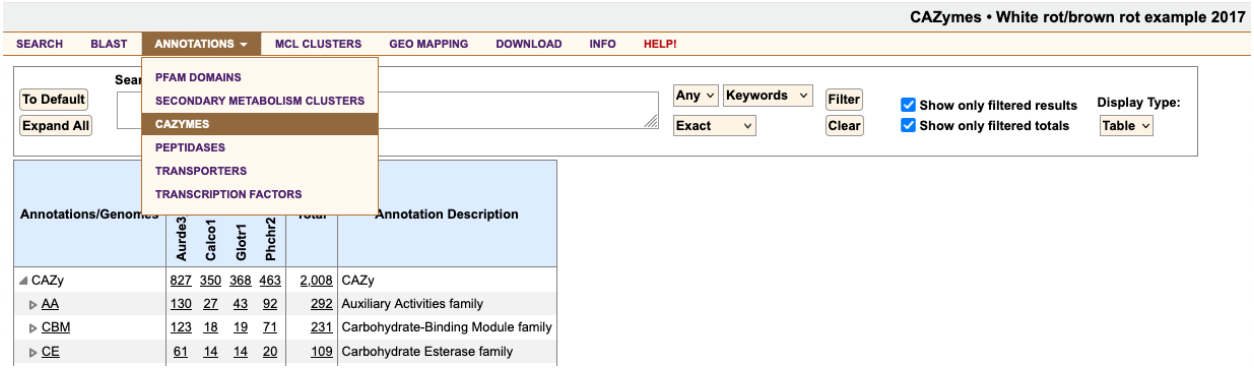
##	Name	Assembly Length	# Genes	Published
1	Auricularia subglabra v2.0	76,853,599	25,459	Floudas D et al., 2012
2	Calocera cornea v1.0	33,244,933	13,177	Nagy LG et al., 2016
3	Gloeophyllum trabeum v1.0	37,181,821	11,846	Floudas D et al., 2012
4	Phanerochaete chrysosporium RP-78 v2.2	35,149,519	13,602	Ohm RA et al., 2014

CAZy browser

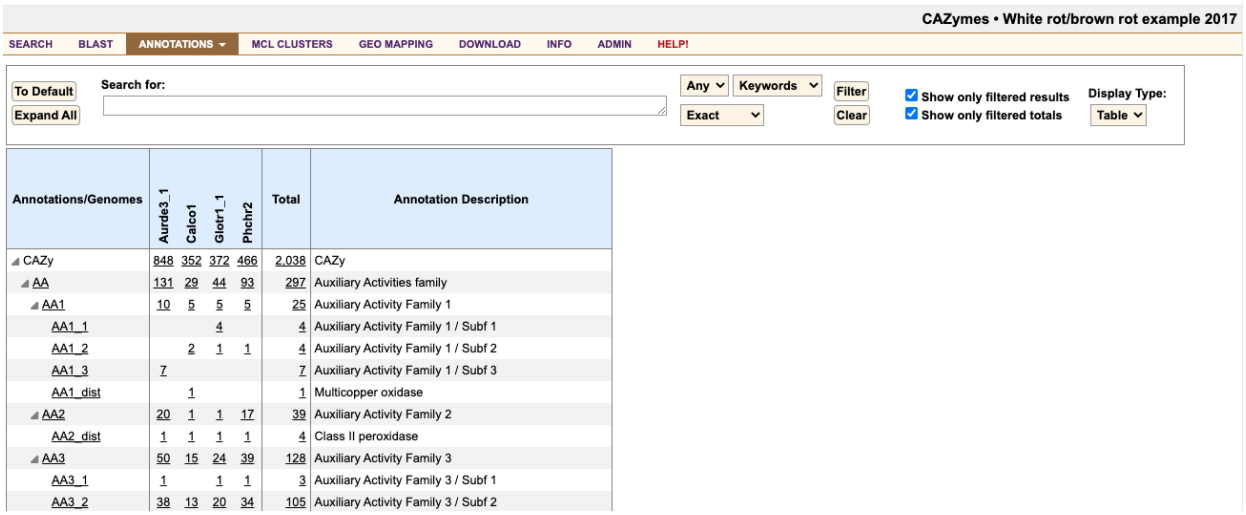
CAZymes (Carbohydrate-Active Enzymes) are enzymes that degrade, modify, and/or create glycosidic bonds (Levasseur et al. 2013). They can be classified into families of structurally-

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.

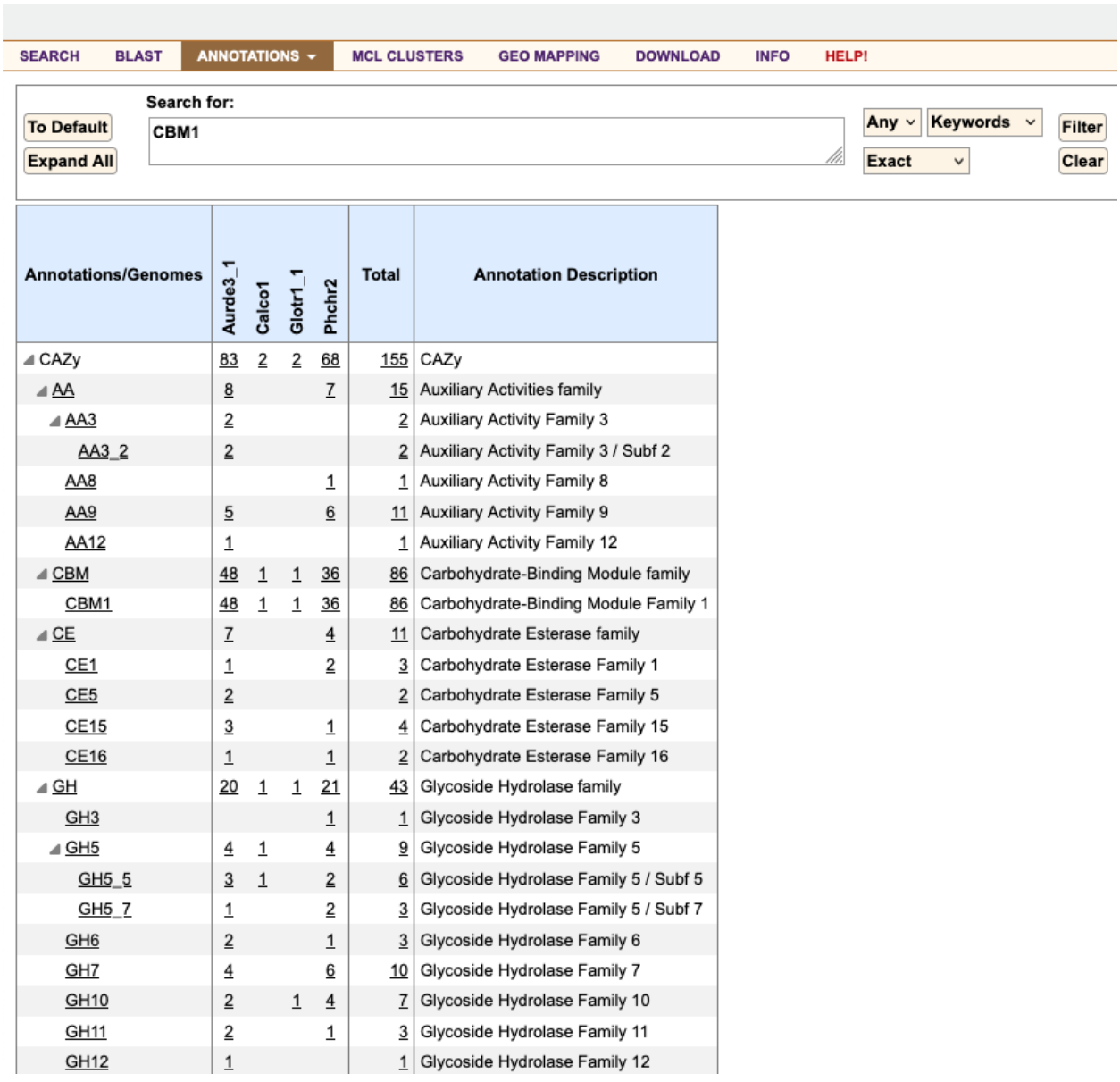


Here you will see a table representation of the predicted CAZymes in each species. The organisms are labeled along the top by genome portal identifier ("portal ID"). The CAZymes are organized hierarchically by family and labeled along the sides: CAZy family identifier on the left, and family description 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.



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!

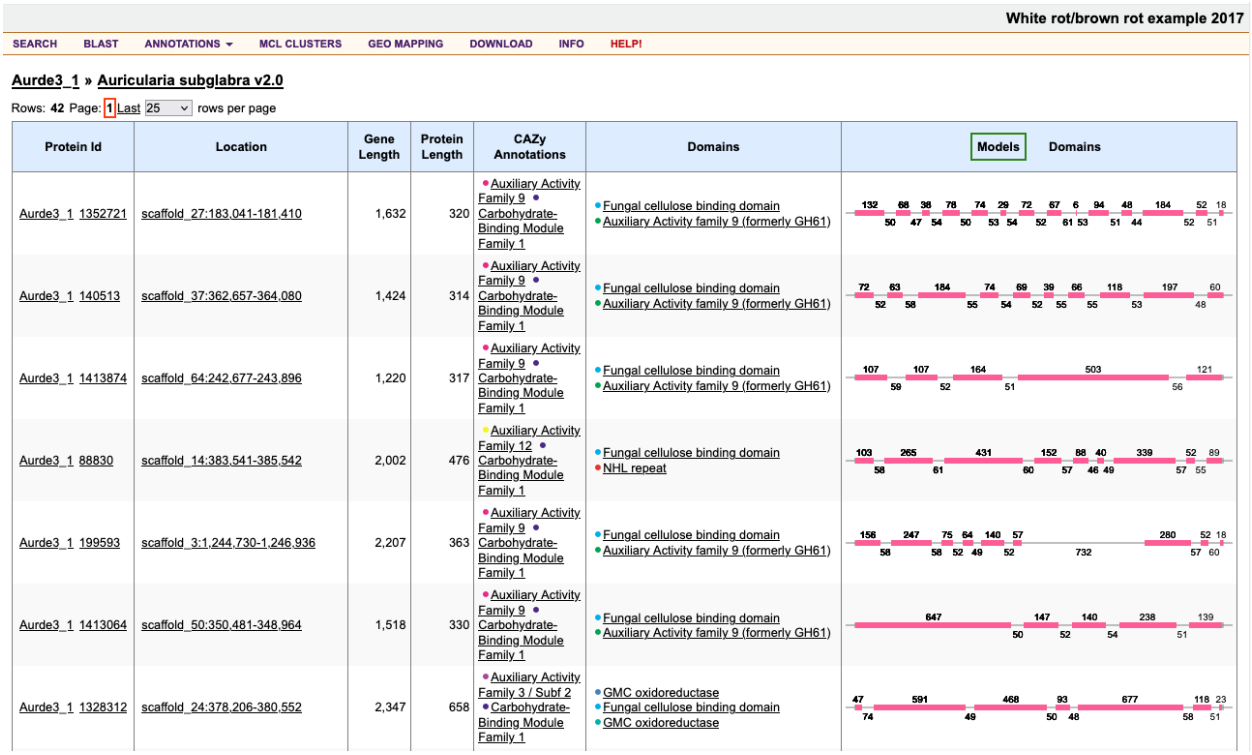
What about the carbohydrate binding motifs, CBM1? Let's say we don't want to scroll through the entire list of CAZymes. Type "CBM1" into the "CAZY terms" search box and click "Filter". This will limit the view to only those CAZymes that have a CBM1. Why do so many CAZymes besides CBM1 show up? Because 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.



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-occuring 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.



White rot/brown rot example 2017

SEARCH BLAST ANNOTATIONS MCL CLUSTERS GEO MAPPING DOWNLOAD INFO HELP!

Aurde3_1 » Auricularia subglabra v2.0

Rows: 42 Page: 1 Last 25 rows per page

Protein Id	Location	Gene Length	Protein Length	CAZy Annotations	Domains	Models Domains
Aurde3_1 1352721	scaffold_27:183,041-181,410	1,632	320	Auxiliary Activity Family 9 • Carbohydrate-Binding Module Family 1	Fungal cellulose binding domain • Auxiliary Activity family 9 (formerly GH61)	132 68 38 78 74 29 72 67 6 94 48 184 52 18 / 50 47 54 50 53 54 52 61 53 51 44 52 51
Aurde3_1 140513	scaffold_37:362,657-364,080	1,424	314	Auxiliary Activity Family 9 • Carbohydrate-Binding Module Family 1	Fungal cellulose binding domain • Auxiliary Activity family 9 (formerly GH61)	72 63 184 74 69 39 66 118 197 60 / 52 58 55 54 52 55 55 53 48
Aurde3_1 1413874	scaffold_64:242,677-243,896	1,220	317	Auxiliary Activity Family 9 • Carbohydrate-Binding Module Family 1	Fungal cellulose binding domain • Auxiliary Activity family 9 (formerly GH61)	107 107 164 503 121 / 59 52 51 56
Aurde3_1 88830	scaffold_14:383,541-385,542	2,002	476	Auxiliary Activity Family 12 • Carbohydrate-Binding Module Family 1	Fungal cellulose binding domain • NHL repeat	103 265 431 152 88 40 339 52 89 / 58 61 60 57 46 49 57 55
Aurde3_1 199593	scaffold_3:1,244,730-1,246,936	2,207	363	Auxiliary Activity Family 9 • Carbohydrate-Binding Module Family 1	Fungal cellulose binding domain • Auxiliary Activity family 9 (formerly GH61)	158 247 75 64 140 57 280 52 18 / 58 58 52 49 52 732 57 60
Aurde3_1 1413064	scaffold_50:350,481-348,964	1,518	330	Auxiliary Activity Family 9 • Carbohydrate-Binding Module Family 1	Fungal cellulose binding domain • Auxiliary Activity family 9 (formerly GH61)	647 147 140 238 139 / 50 52 54 51
Aurde3_1 1328312	scaffold_24:378,206-380,552	2,347	658	Auxiliary Activity Family 3 / Subf 2 • Carbohydrate-Binding Module Family 1	GMC oxidoreductase • Fungal cellulose binding domain • GMC oxidoreductase	47 591 468 93 677 118 23 / 74 49 50 48 58 51

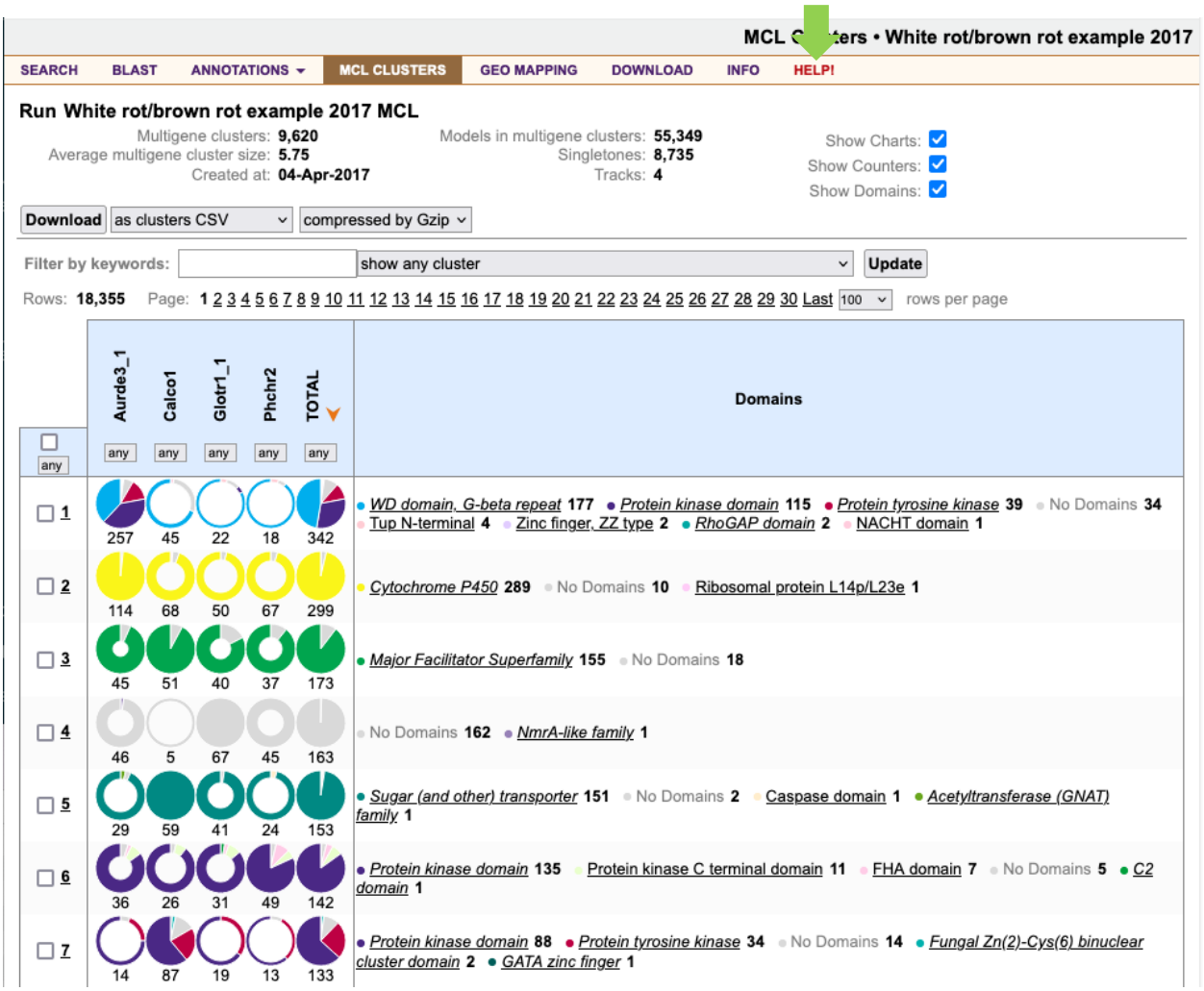
As an exercise, repeat the same search with GH6, GH7, and also the AA9 family of lytic polysaccharide monoxygenases, which may oxidatively act on lignin (Levasseur et al. 2013). Do the presence/absence patterns of these genes indicate the same conclusions about these fungi's mode of decay as we found 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.)

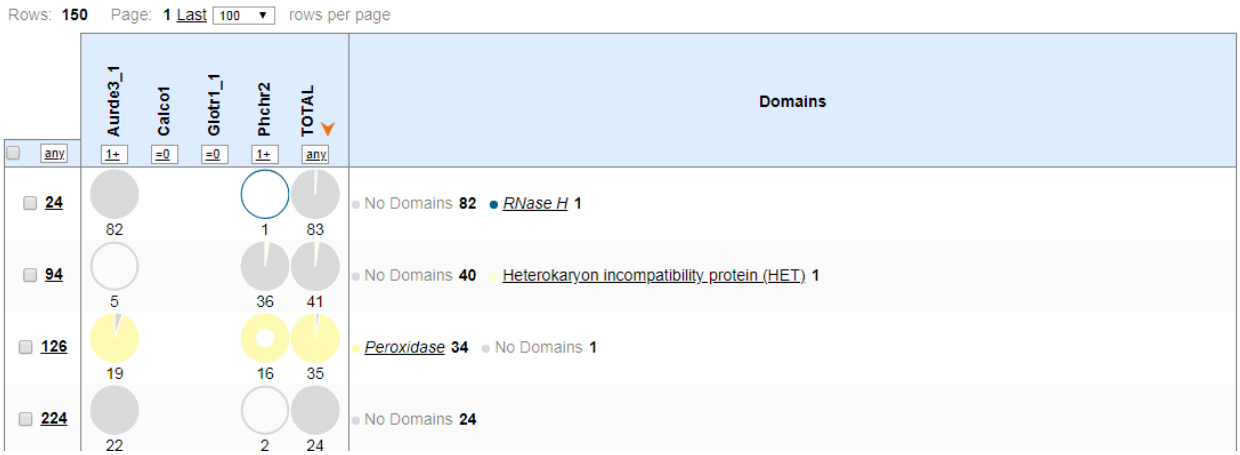
Cluster page

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.



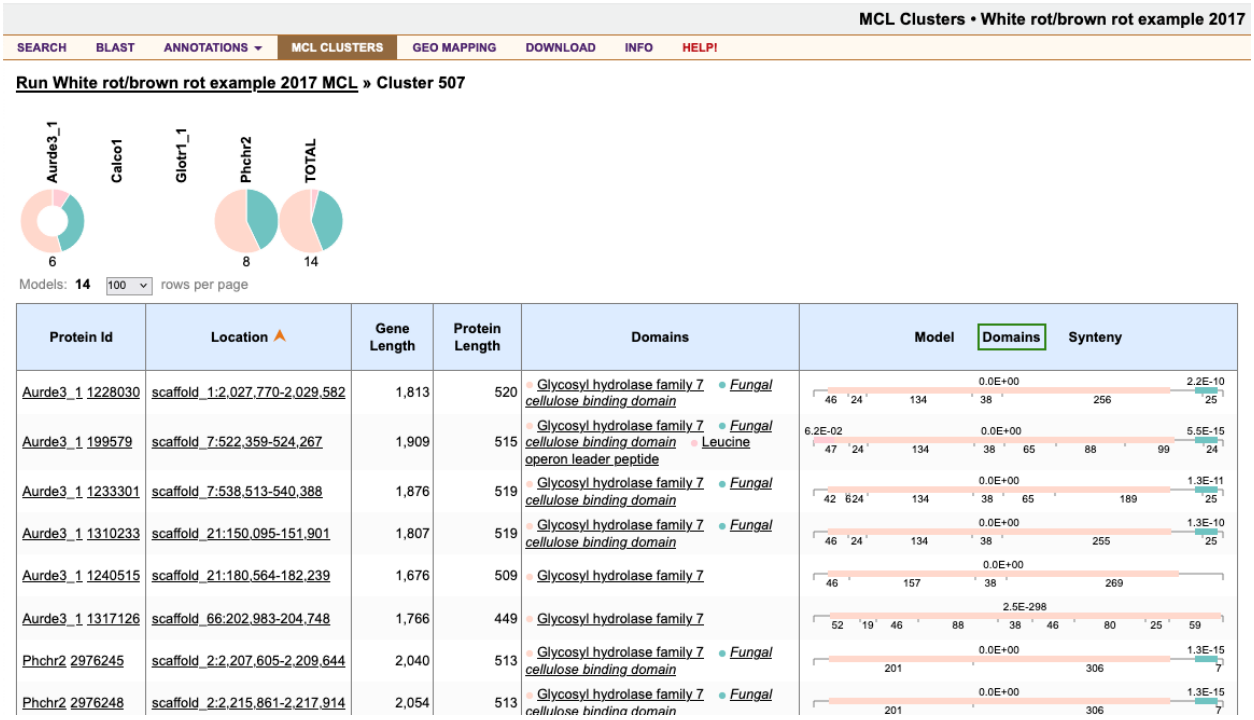
Notice that under each organism label is a button “any” that can be used to filter clusters by the number of proteins that organism contributes to a cluster, and thus limit which clusters are shown. As an experiment, set the white rot fungi (Aurde3_1 and Phchr2) to “1+” and the brown rot fungi (Calco1 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.



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.

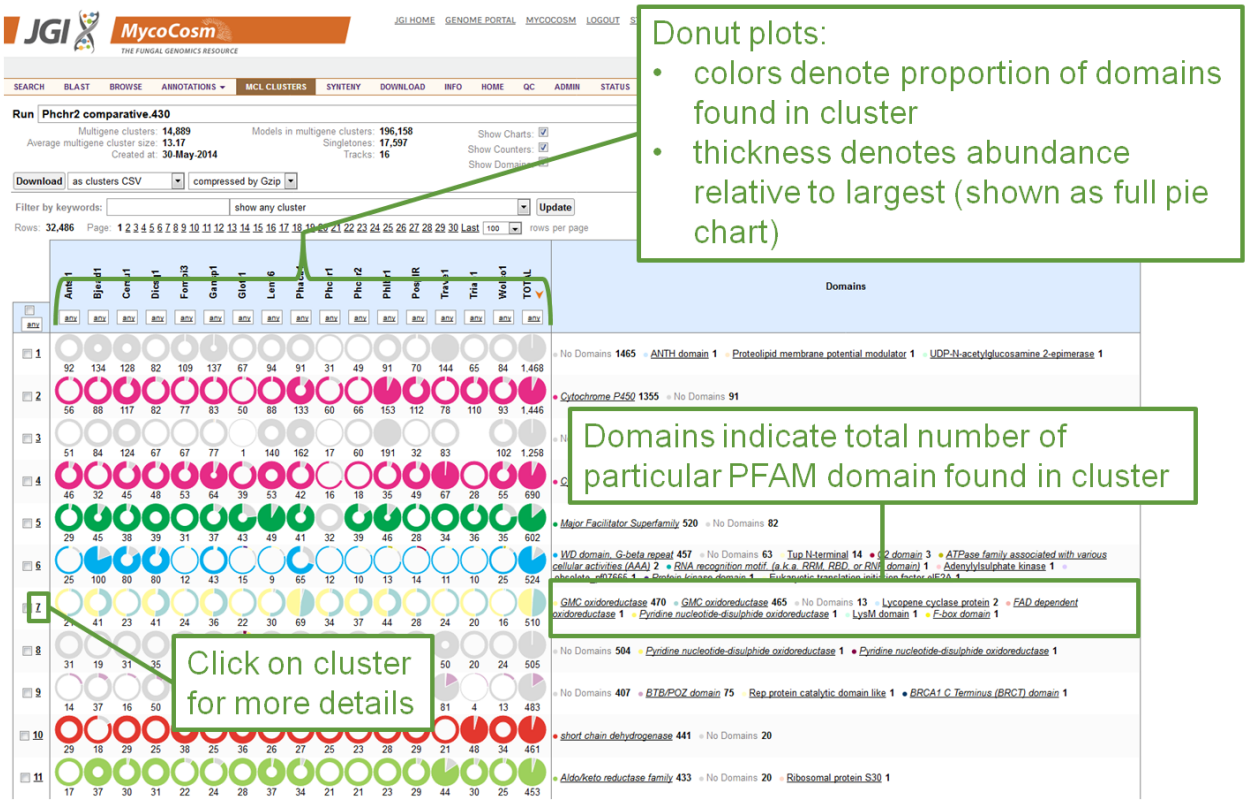
Scroll through the rest of the 150 clusters and you will see domains such as “Glycosyl hydrolase family 7” and “Fungal cellulose binding domain” in cluster 507, which roughly overlap with the GH7 and CBM1 families from the CAZy exercise. Click the “507” to explore that cluster in more detail. On the 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.



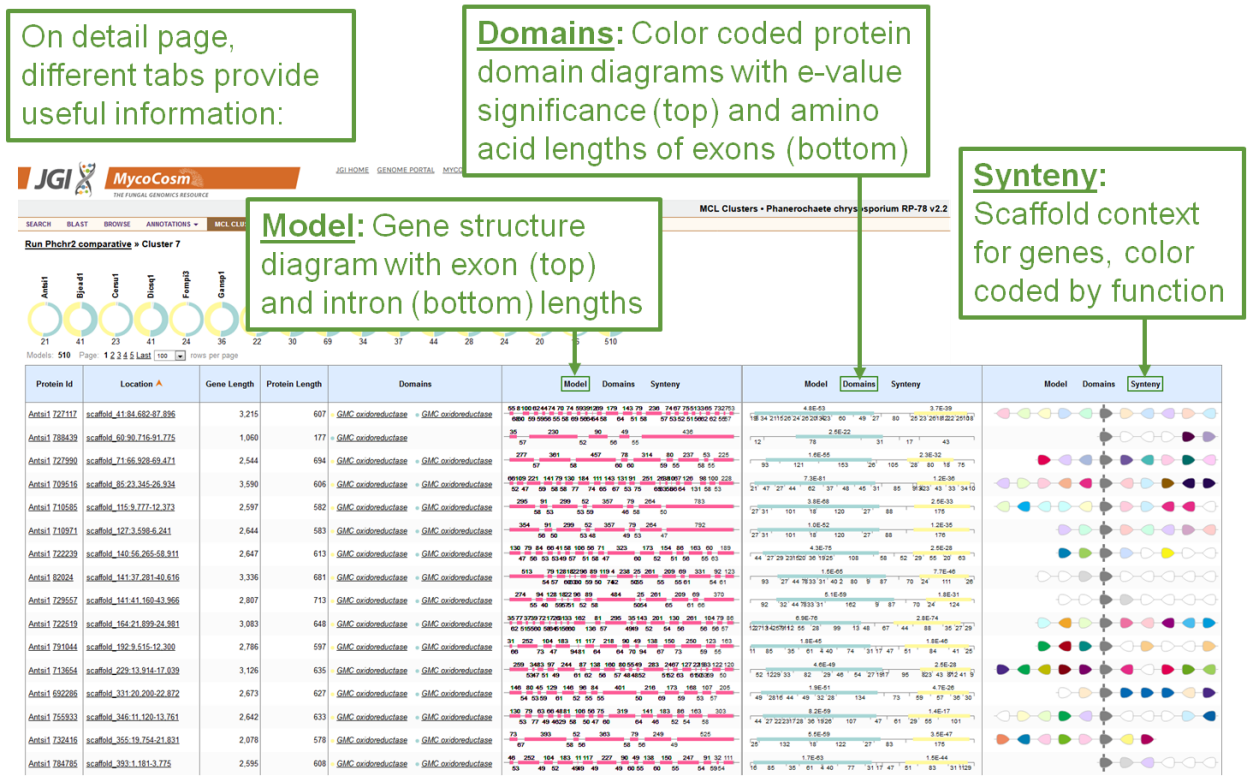
Let’s look at what other proteins have the CBM1 carbohydrate-binding motifs in them. Returning to the cluster run page (click the “MCL CLUSTERS” tab). Enter the phrase “fungal cellulose binding domain” (be sure to include the quotes) into the “filter by keywords” field and select “Update”. This returns 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 monoxygenases (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.

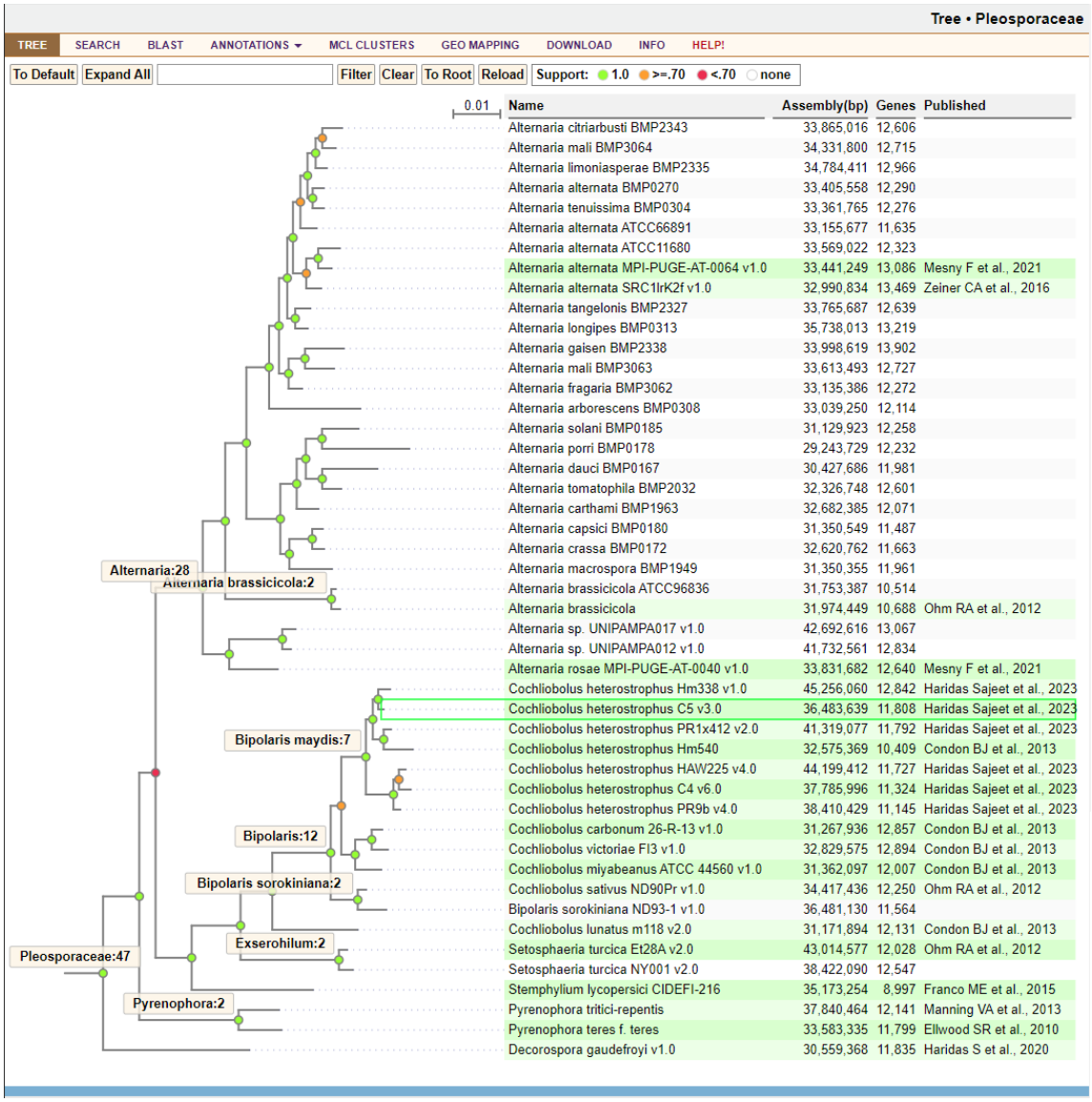
MycoCosm: Synteny Tutorial

<u>Objective</u>: 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.



Name	Assembly(bp)	Genes	Published
Alternaria citriarbusti BMP2343	33,865,016	12,606	
Alternaria mali BMP3064	34,331,800	12,715	
Alternaria limoniasperae BMP2335	34,784,411	12,966	
Alternaria alternata BMP0270	33,405,558	12,290	
Alternaria tenuissima BMP0304	33,361,765	12,276	
Alternaria alternata ATCC66891	33,155,677	11,635	
Alternaria alternata ATCC11680	33,569,022	12,323	
Alternaria alternata MPI-PUGE-AT-0064 v1.0	33,441,249	13,086	Mesny F et al., 2021
Alternaria alternata SRC1lrK2f v1.0	32,990,834	13,469	Zeiner CA et al., 2016
Alternaria tangelonis BMP2327	33,765,687	12,639	
Alternaria longipes BMP0313	35,738,013	13,219	
Alternaria gaisen BMP2338	33,998,619	13,902	
Alternaria mali BMP3063	33,613,493	12,727	
Alternaria fragaria BMP3062	33,135,386	12,272	
Alternaria arborescens BMP0308	33,039,250	12,114	
Alternaria solani BMP0185	31,129,923	12,258	
Alternaria porri BMP0178	29,243,729	12,232	
Alternaria dauci BMP0167	30,427,686	11,981	
Alternaria tomatophila BMP2032	32,326,748	12,601	
Alternaria carthami BMP1963	32,682,385	12,071	
Alternaria capsici BMP0180	31,350,549	11,487	
Alternaria crassa BMP0172	32,620,762	11,663	
Alternaria macrospora BMP1949	31,350,355	11,961	
Alternaria brassicicola ATCC96836	31,753,387	10,514	
Alternaria brassicicola	31,974,449	10,688	Ohm RA et al., 2012
Alternaria sp. UNIPAMPA017 v1.0	42,692,616	13,067	
Alternaria sp. UNIPAMPA012 v1.0	41,732,561	12,834	
Alternaria rosae MPI-PUGE-AT-0040 v1.0	33,831,682	12,640	Mesny F et al., 2021
Cochliobolus heterostrophus Hm338 v1.0	45,256,060	12,842	Haridas Sajeet et al., 2023
Cochliobolus heterostrophus C5 v3.0	36,483,639	11,808	Haridas Sajeet et al., 2023
Cochliobolus heterostrophus PR1x412 v2.0	41,319,077	11,792	Haridas Sajeet et al., 2023
Cochliobolus heterostrophus Hm540	32,575,369	10,409	Condon BJ et al., 2013
Cochliobolus heterostrophus HAW225 v4.0	44,199,412	11,727	Haridas Sajeet et al., 2023
Cochliobolus heterostrophus C4 v6.0	37,785,996	11,324	Haridas Sajeet et al., 2023
Cochliobolus heterostrophus PR9b v4.0	38,410,429	11,145	Haridas Sajeet et al., 2023
Cochliobolus carbonum 26-R-13 v1.0	31,267,936	12,857	Condon BJ et al., 2013
Cochliobolus victoriae FI3 v1.0	32,829,575	12,894	Condon BJ et al., 2013
Cochliobolus miyabeanus ATCC 44560 v1.0	31,362,097	12,007	Condon BJ et al., 2013
Cochliobolus sativus ND90Pr v1.0	34,417,436	12,250	Ohm RA et al., 2012
Bipolaris sorokiniana ND93-1 v1.0	36,481,130	11,564	
Cochliobolus lunatus m118 v2.0	31,171,894	12,131	Condon BJ et al., 2013
Setosphaeria turcica Et28A v2.0	43,014,577	12,028	Ohm RA et al., 2012
Setosphaeria turcica NY001 v2.0	38,422,090	12,547	
Stemphylium lycopersici CIDEFI-216	35,173,254	8,997	Franco ME et al., 2015
Pyrenophora tritici-repentis	37,840,464	12,141	Manning VA et al., 2013
Pyrenophora teres f. teres	33,583,335	11,799	Ellwood SR et al., 2010
Decorospora gaudefroyi v1.0	30,559,368	11,835	Haridas S et al., 2020

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.

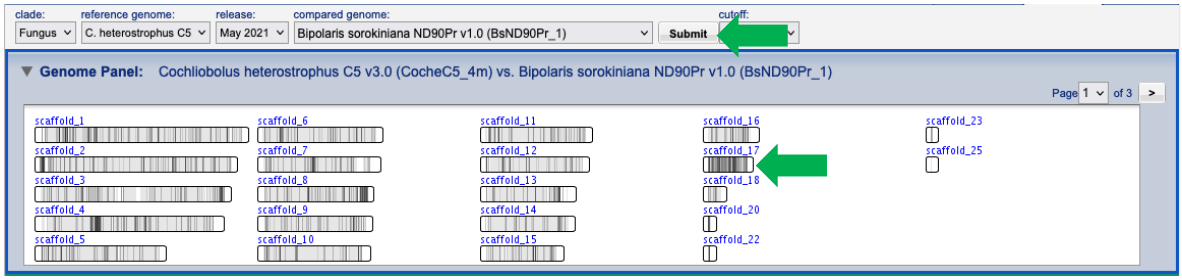
Click on the SYNTENY tab in the organism portal (*Cochliobolus heterostrophus* C5).

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

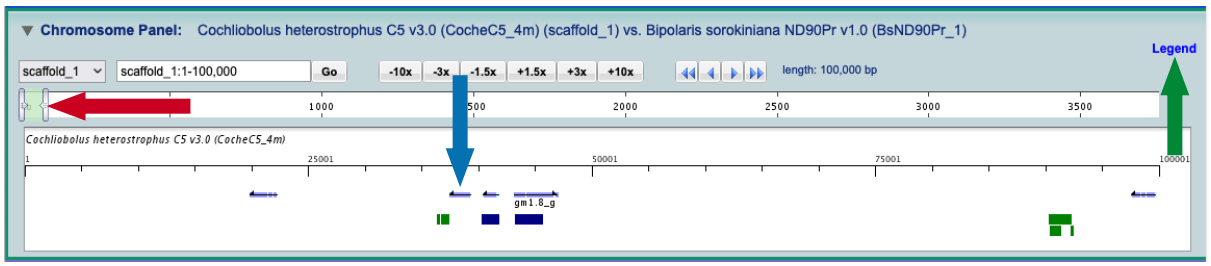
A. the Genome Panel
B. the Chromosome Panel
C. the Comparison Panel.

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

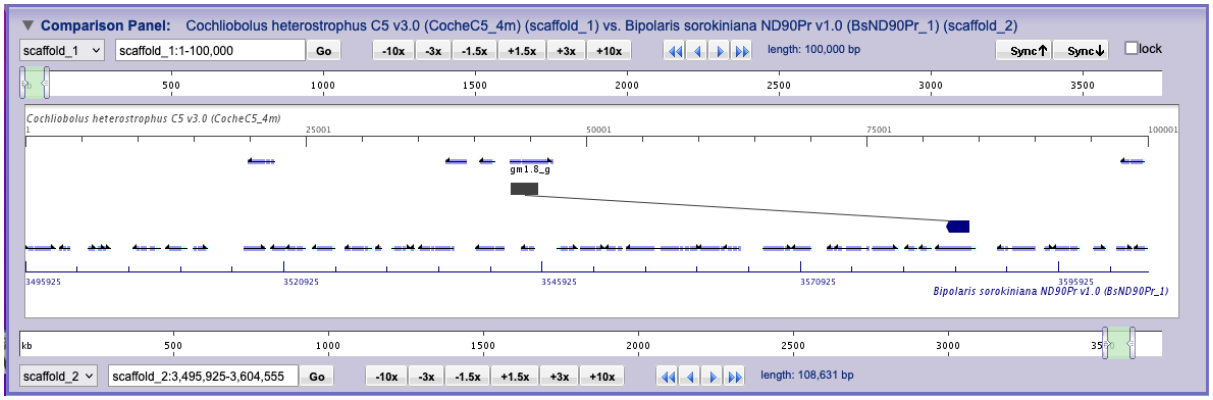
A. The Genome Panel depicts alignment density for all scaffolds in the reference genome against all chromosomes in the compared genome. Here, alignment density is defined for a region in the reference genome as the number of syntenic regions in the compared genome. Darker regions in the image have higher 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.



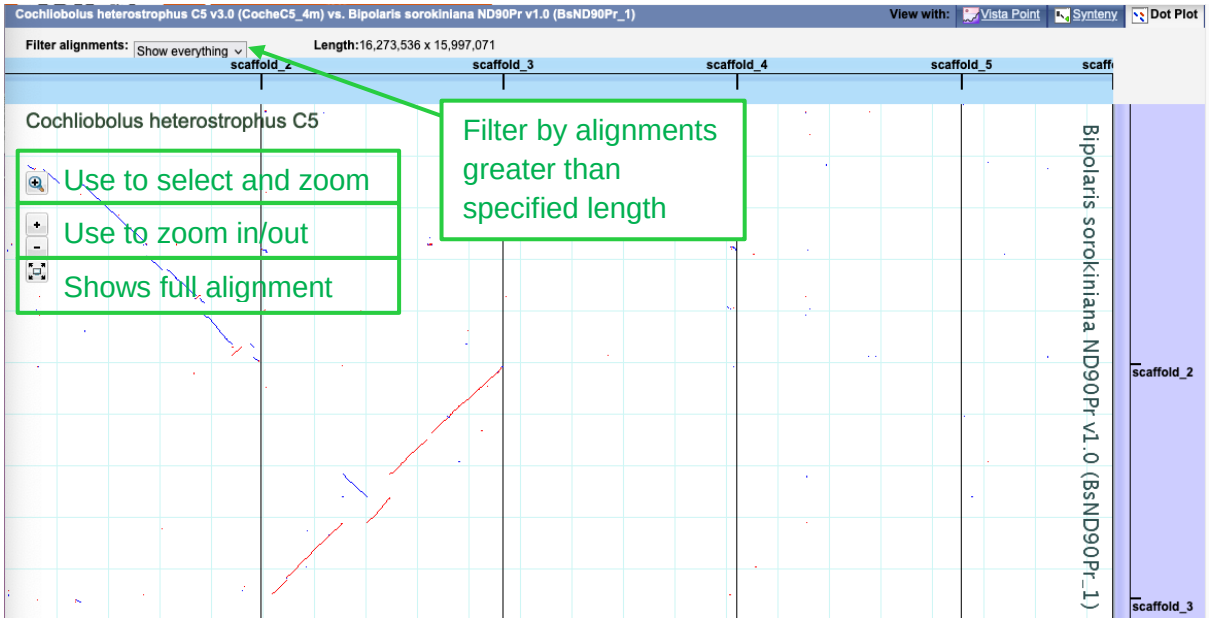
Syntenic blocks and gene models are both interactive, as described above for the Chromosome Panel. Navigation controls allow the user to switch chromosomes, zoom and pan independently over the reference and compared genomes.

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.

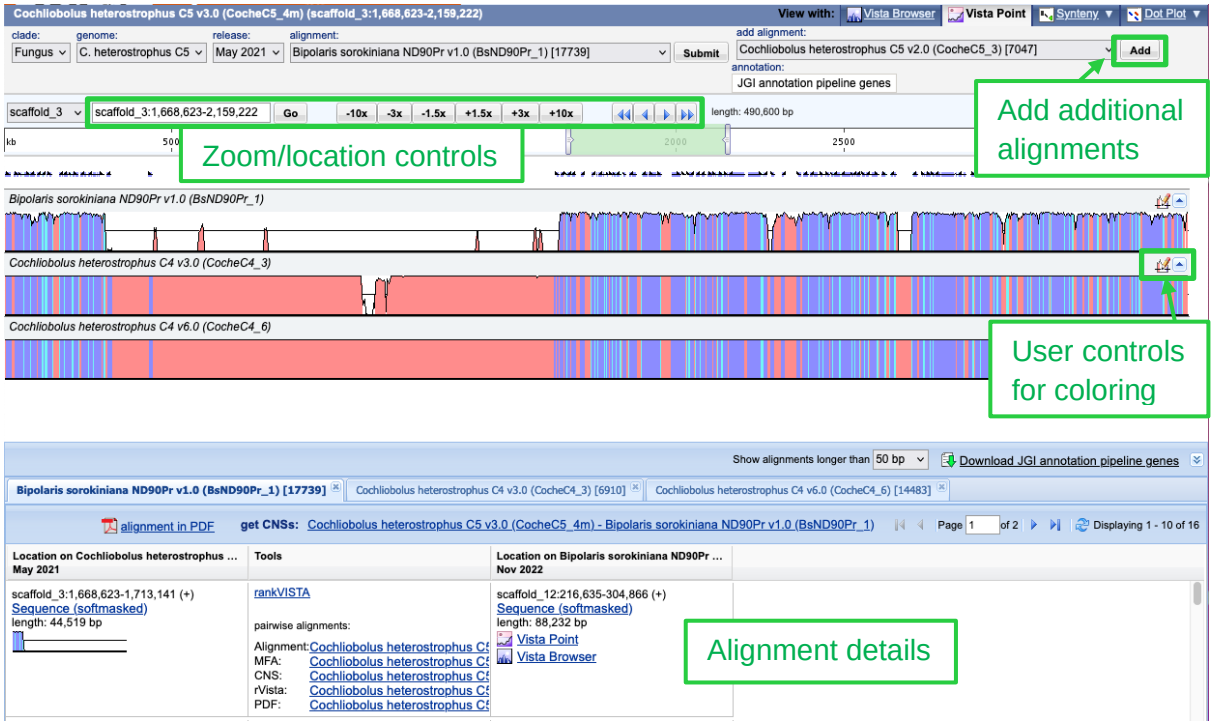
In the main view window, DNA coordinates of the reference genome are presented on the X axis, and DNA coordinates of the compared genome are presented on the Y axis. All chromosomes or scaffolds are concatenated together, usually in a descending order by size. The diagonal lines in the image display 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 syntenic 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:

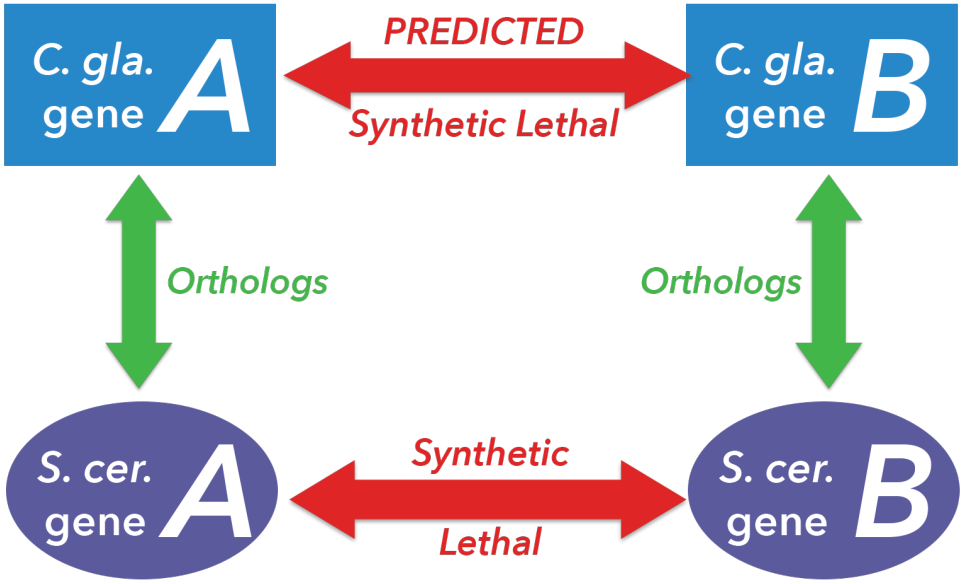
- Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, et al. (2012) Diverse Lifestyles and Strategies of Plant Pathogenesis Encoded in the Genomes of Eighteen Dothideomycetes Fungi. PLOS Pathogens 8(12): e1003037.

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.

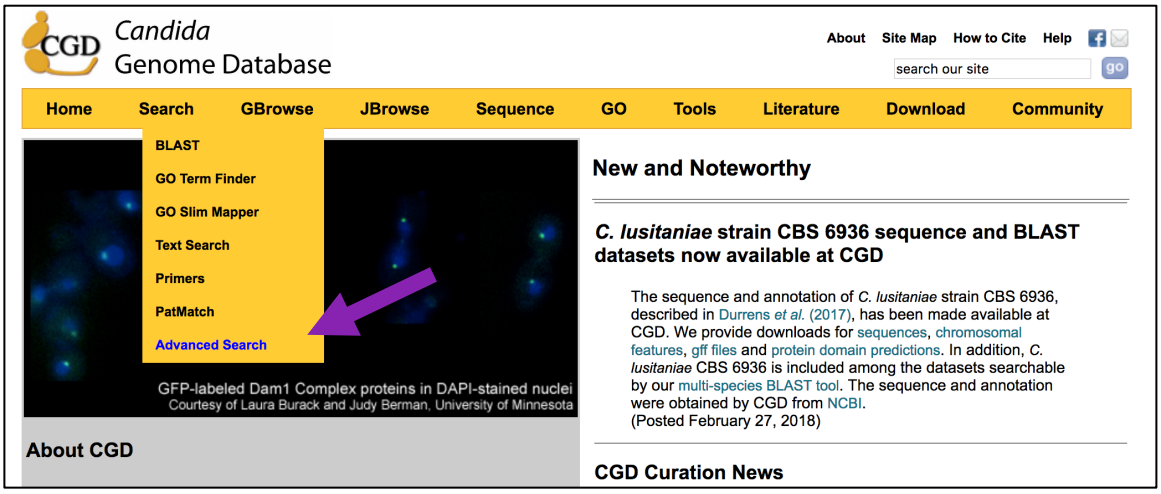
Genetic interactions such as synthetic lethality are richly documented for the budding yeast *S. cerevisiae*, but not as much for many other fungal species. By examining known genetic interactions in *S. cerevisiae*, we can predict synthetic lethal relationships in *C. glabrata* and other fungal pathogens.



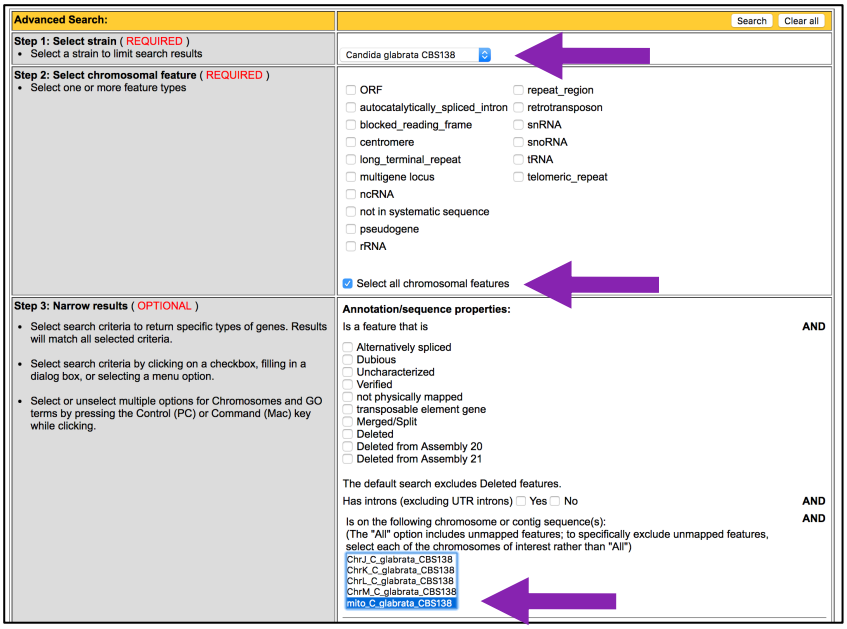
If conserved, these synthetic lethal interactions may reveal future antifungal targets for use against azole-resistant strains in the clinic. Using known synthetic lethal interactions in the *S. cerevisiae* genome, predict potentially conserved synthetic lethal interactions for mitochondrial genes in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

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

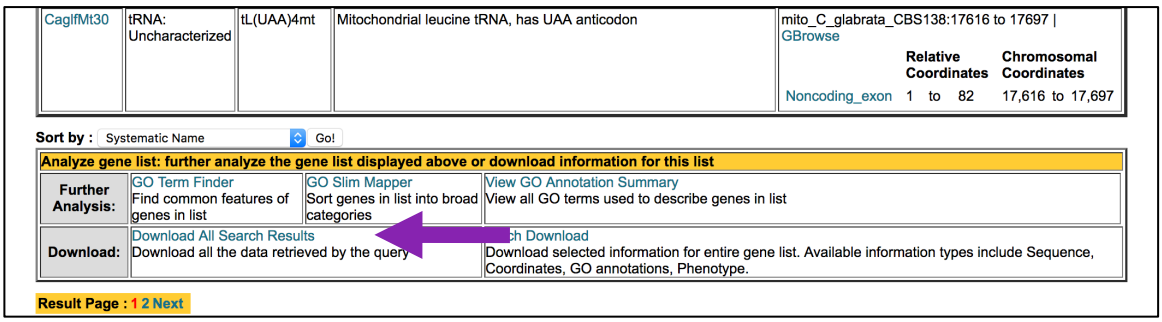
- On the CGD homepage (**http://www.candidagenome.org**), open the Search tab in the yellow toolbar and select **Advanced Search**.



- 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.

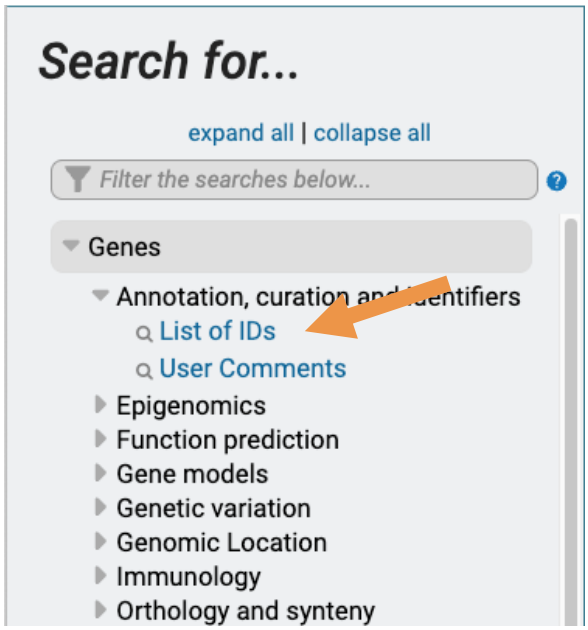


- Scroll to the bottom of the page and click on the "**Download All Search Results**" link. The results will download in an Excel sheet.

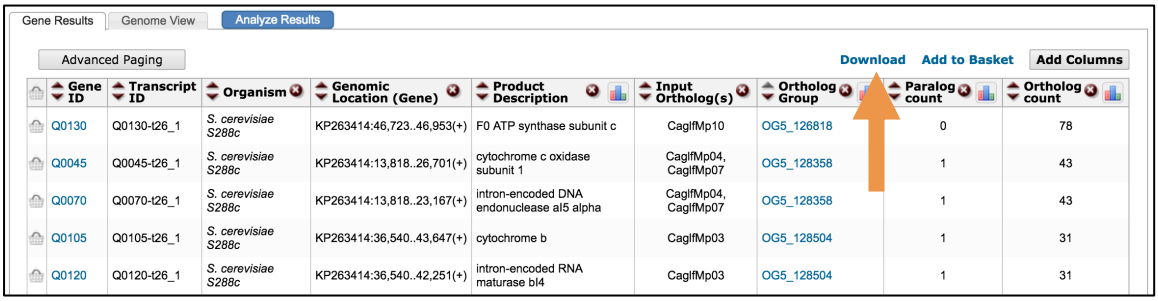


2. Use FungiDB to find *S. cerevisiae* orthologs of *C. glabrata* mitochondrial genes:

- 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**".



- 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.

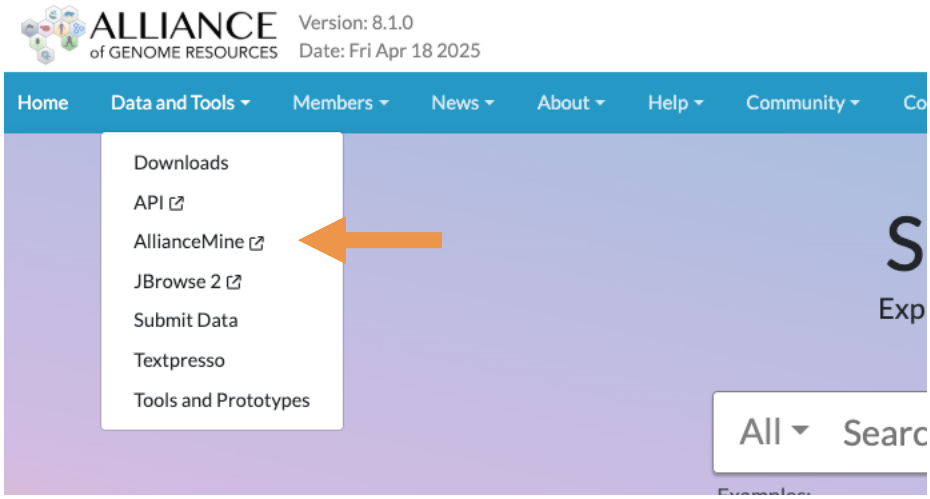


Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Input Ortholog(s)	Ortholog Group	Paralog count	Ortholog count
Q0130	Q0130-t26_1	*S. cerevisiae S288c*	KP263414:46,723..46,953(+)	F0 ATP synthase subunit c	CaglfMp10	OG5_126818	0	78
Q0045	Q0045-t26_1	*S. cerevisiae S288c*	KP263414:13,818..26,701(+)	cytochrome c oxidase subunit 1	CaglfMp04, CaglfMp07	OG5_128358	1	43
Q0070	Q0070-t26_1	*S. cerevisiae S288c*	KP263414:13,818..23,167(+)	intron-encoded DNA endonuclease aI5 alpha	CaglfMp04, CaglfMp07	OG5_128358	1	43
Q0105	Q0105-t26_1	*S. cerevisiae S288c*	KP263414:36,540..43,647(+)	cytochrome b	CaglfMp03	OG5_128504	1	31
Q0120	Q0120-t26_1	*S. cerevisiae S288c*	KP263414:36,540..42,251(+)	intron-encoded RNA maturase bI4	CaglfMp03	OG5_128504	1	31

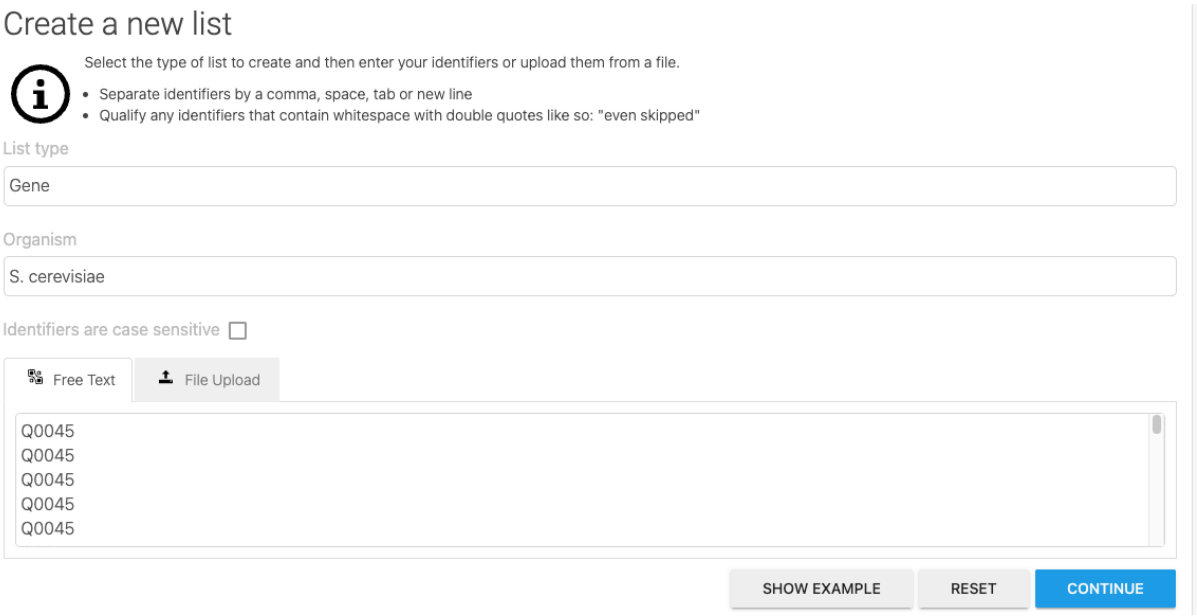
- In the download options menu, select "**Tab delimited (openable in Excel) – choose a pre-configured table**". Set the Download Type as **Comma-delimited (.csv) file***, then hit **Get**.

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

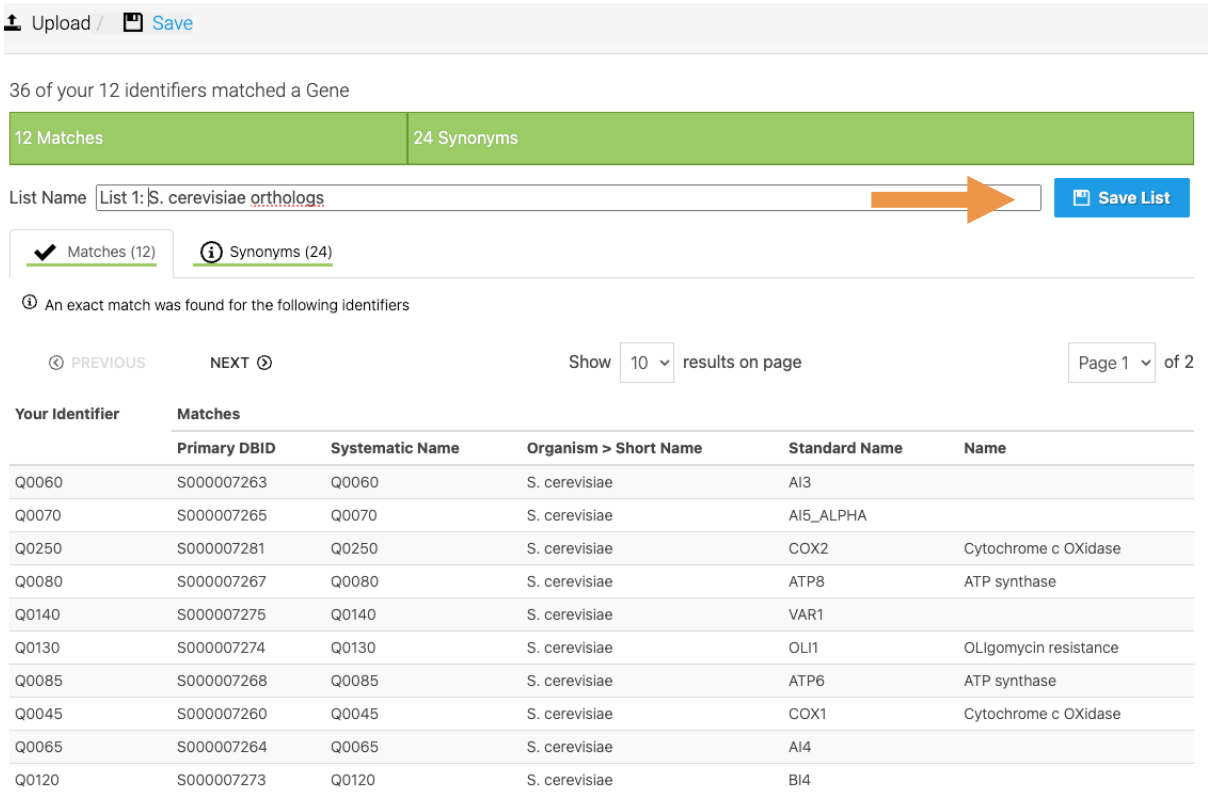
- Access AllianceMine from Alliance home page (https://www.alliancegenome.org); click on AllianceMine from the Data and Tools section at the top of the page.



- 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.

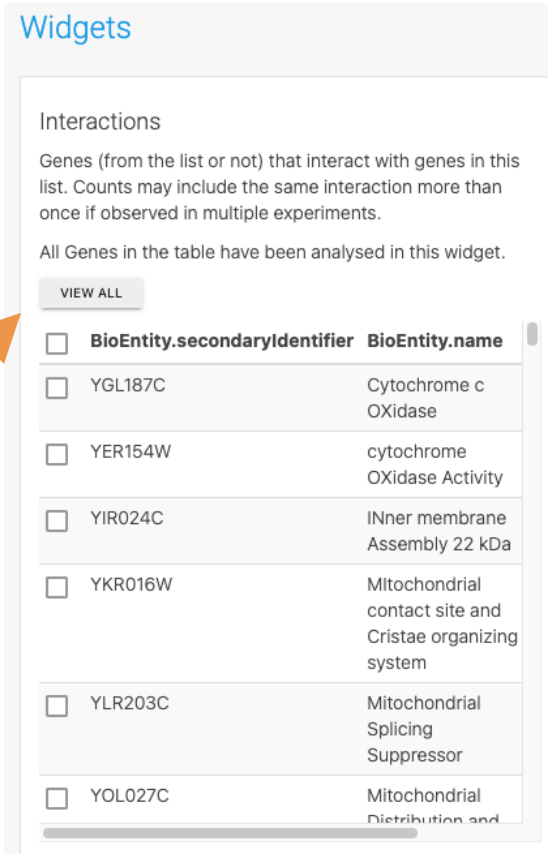


- You can save this list of genes as "**List 1: S. cerevisiae orthologs**" (type it in). Click on the blue "**Save List**" button.

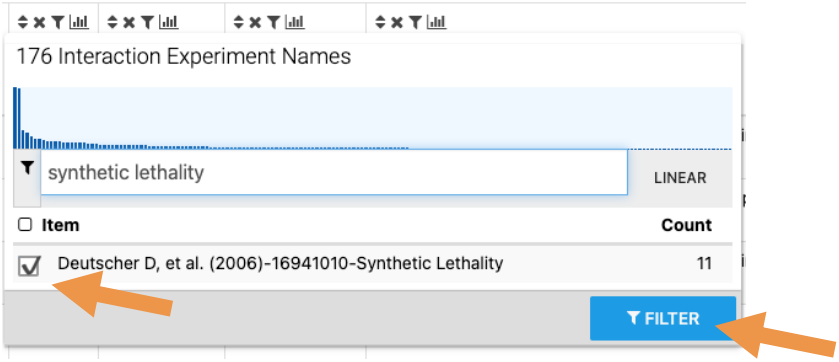


Your Identifier	Matches				
	Primary DBID	Systematic Name	Organism > Short Name	Standard Name	Name
Q0060	S000007263	Q0060	S. cerevisiae	AI3	
Q0070	S000007265	Q0070	S. cerevisiae	AI5_ALPHA	
Q0250	S000007281	Q0250	S. cerevisiae	COX2	Cytochrome c OXidase
Q0080	S000007267	Q0080	S. cerevisiae	ATP8	ATP synthase
Q0140	S000007275	Q0140	S. cerevisiae	VAR1	
Q0130	S000007274	Q0130	S. cerevisiae	OLI1	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	

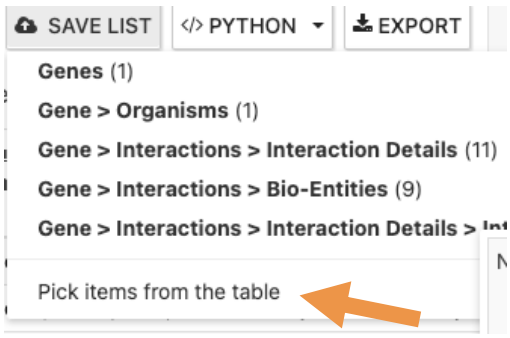
4. After you save the list, you'll get query results with options for running searches. In the **Widgets** section below the table, click "view all" in the **Interactions** section. You'll get roughly 1000 results that you can see by clicking the **forward button** (opposite the back button at the top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



- 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**.

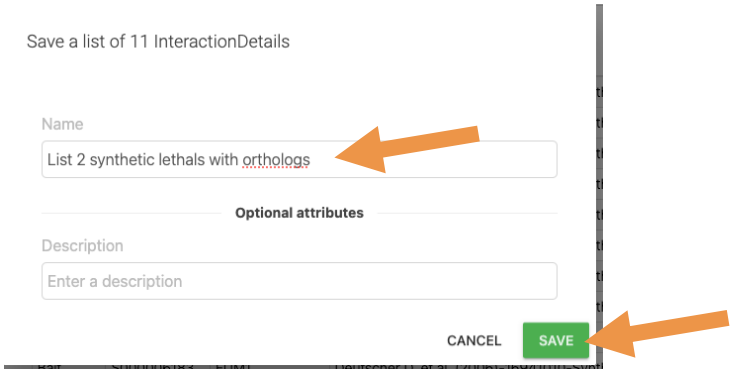


- You'll get a list of eleven genes that are synthetically lethal with a member of your original ortholog list. (COX1)
- Click "**Save List**" at the top right and then choose "**Pick items from the table**." Radio buttons will appear in the table and you want to check all the items in the "**Participant 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.

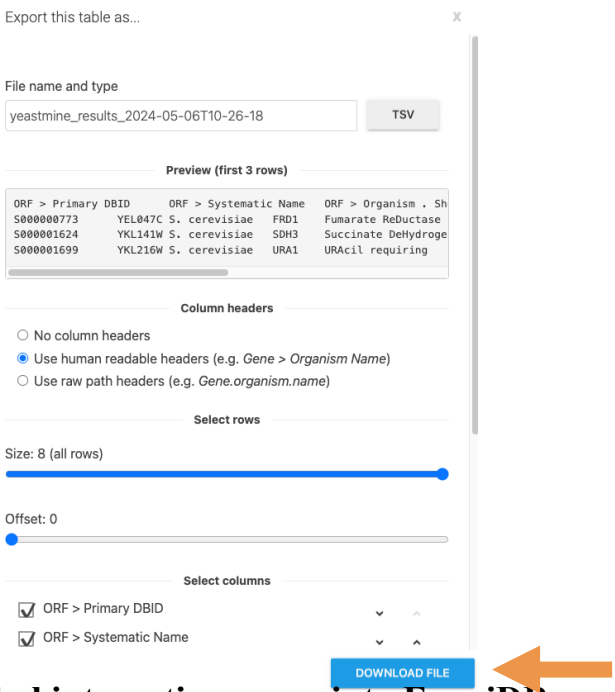


Participant 2 Primary DBID	Participant 2 Standard Name	Experiment Name
S000000773	FRD1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
S000001624	SDH3	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
S000001699	URA1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
S000001699	URA1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

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



- 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.



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

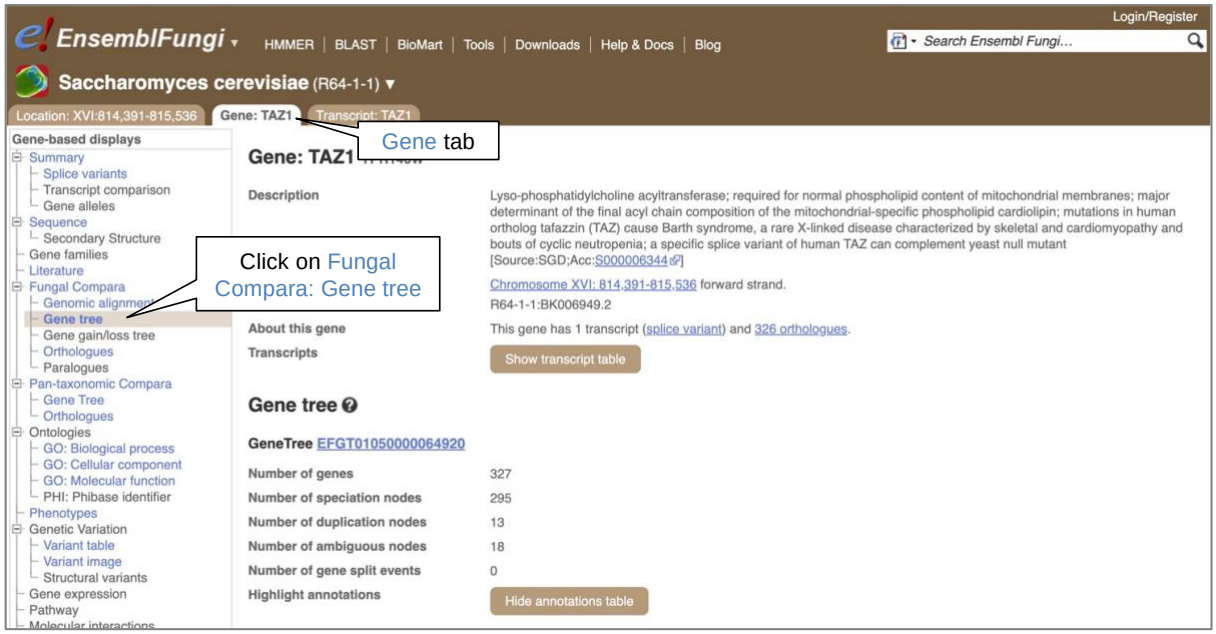
- Open the FungiDB homepage (**http://fungidb.org/**). In the "**Search for…"** section on the left, select "**Genes**," then "**Annotation, curation and identifiers**" section and click on "**List of IDs**".
- Using your exported file from AllianceMine, copy and paste the ORF names of the *S. cerevisiae* 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?

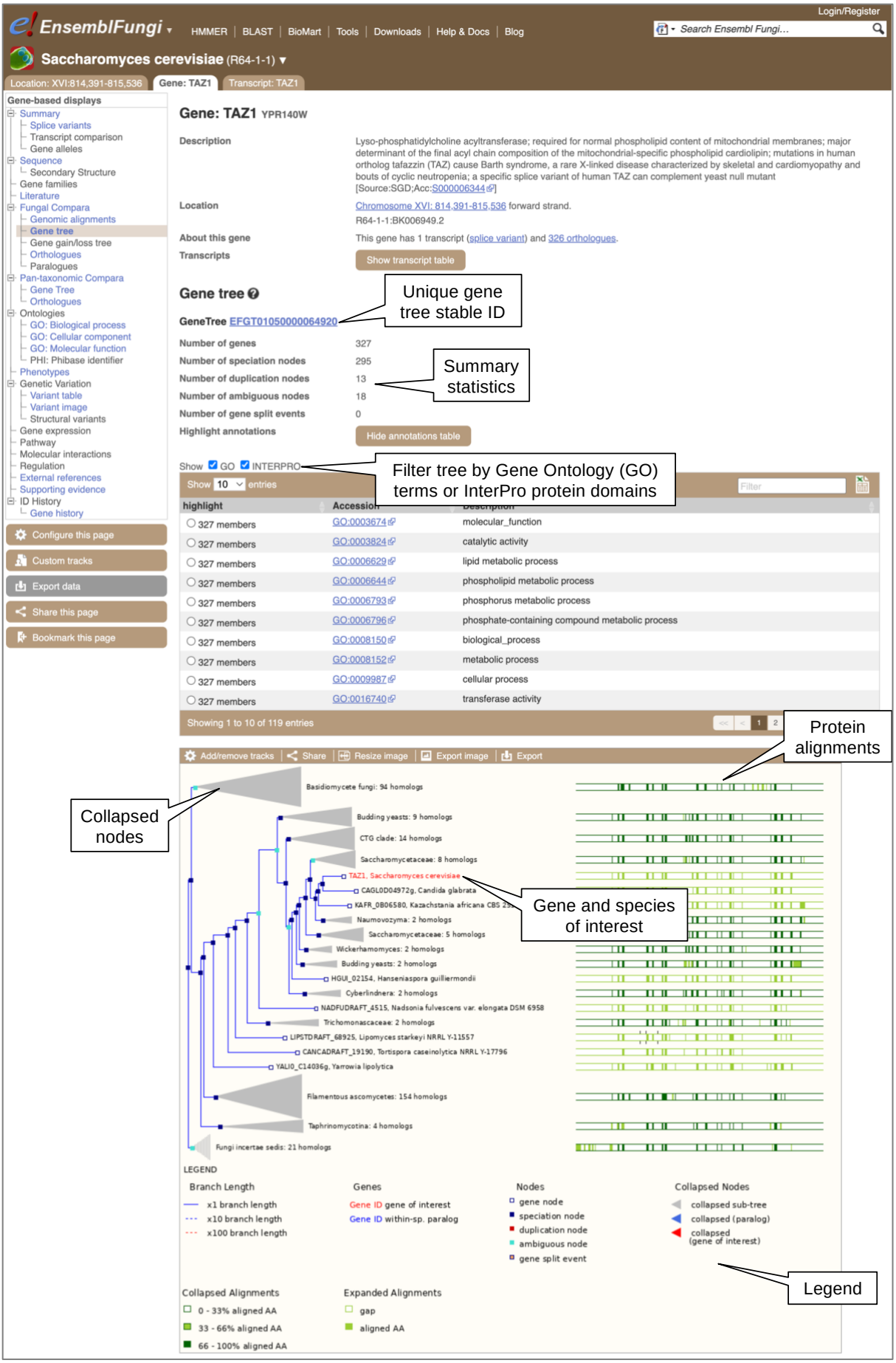
Exercise: Ensembl Fungi Gene Trees and Homologues

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.

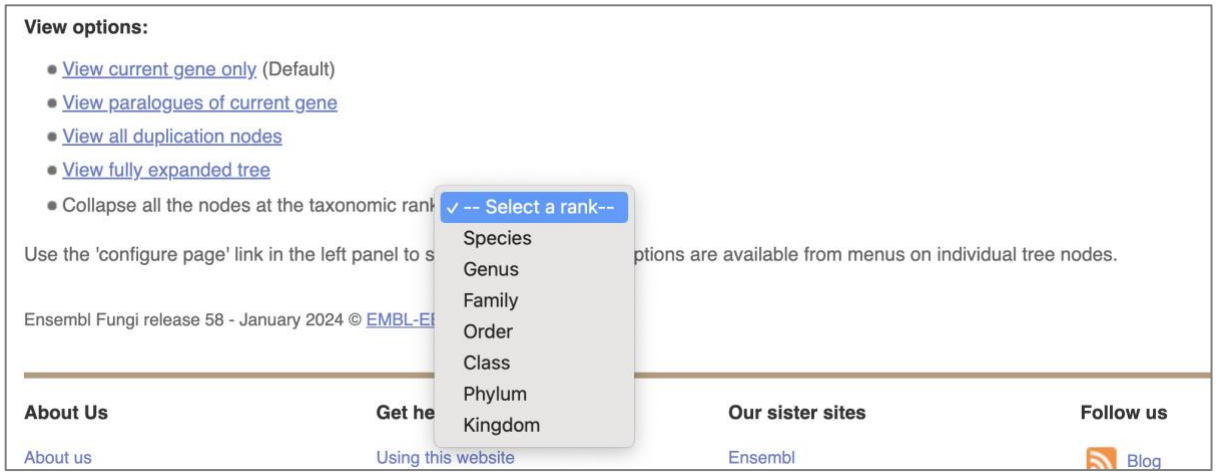


Click on Fungal Compara: Gene tree on the left-hand menu, which will display the current gene (in the context of a phylogenetic tree) used to determine orthologues and paralogues


EnsemblFungi
HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog
Login/Register
Search Ensembl Fungi…
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 Compara
Genomic alignments
Gene tree
Gene gain/loss tree
Orthologues
Paralogues
Pan-taxonomic Compara
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
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
Summary statistics
Highlight annotations
Hide annotations table
Show GO INTERPRO
Filter tree by Gene Ontology (GO) terms or InterPro protein domains
Show 10 entries
Filter
highlight
Accession
Description
327 members GO:0003674 molecular_function
327 members GO:0003824 catalytic activity
327 members GO:0006629 lipid metabolic process
327 members GO:0006644 phospholipid metabolic process
327 members GO:0006793 phosphorus metabolic process
327 members GO:0006796 phosphate-containing compound metabolic process
327 members GO:0008150 biological_process
327 members GO:0008152 metabolic process
327 members GO:0009987 cellular process
327 members GO:0016740 transferase activity
Showing 1 to 10 of 119 entries
Protein alignments
Add/remove tracks | Share | Resize image | Export image | Export
Collapsed nodes
Basidiomycete fungi: 94 homologs
Budding yeasts: 9 homologs
CTG clade: 14 homologs
Saccharomycetaceae: 8 homologs
TAZ1, Saccharomyces cerevisiae
CAGL0D04972g, Candida glabrata
KAFR_0B06580, Kazachstania africana CBS 2
Naumovozyma: 2 homologs
Saccharomycetaceae: 5 homologs
Wickerhamomyces: 2 homologs
Budding yeasts: 2 homologs
HGUI_02154, Hanseniaspora guilliermondii
Cyberlindnera: 2 homologs
NADFUDRAFT_4515, Nadsonia fulvescens var. elongata DSM 6958
Trichomonascaceae: 2 homologs
LIPSTDRAFT_68925, Lipomyces starkeyi NRRL Y-11557
CANCADRAFT_19190, Tortispora caseinolytica NRRL Y-17796
YALI0_C14036g, Yarrowia lipolytica
Filamentous ascomycetes: 154 homologs
Taphrinomycotina: 4 homologs
Fungi incertae sedis: 21 homologs
Gene and species of interest
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

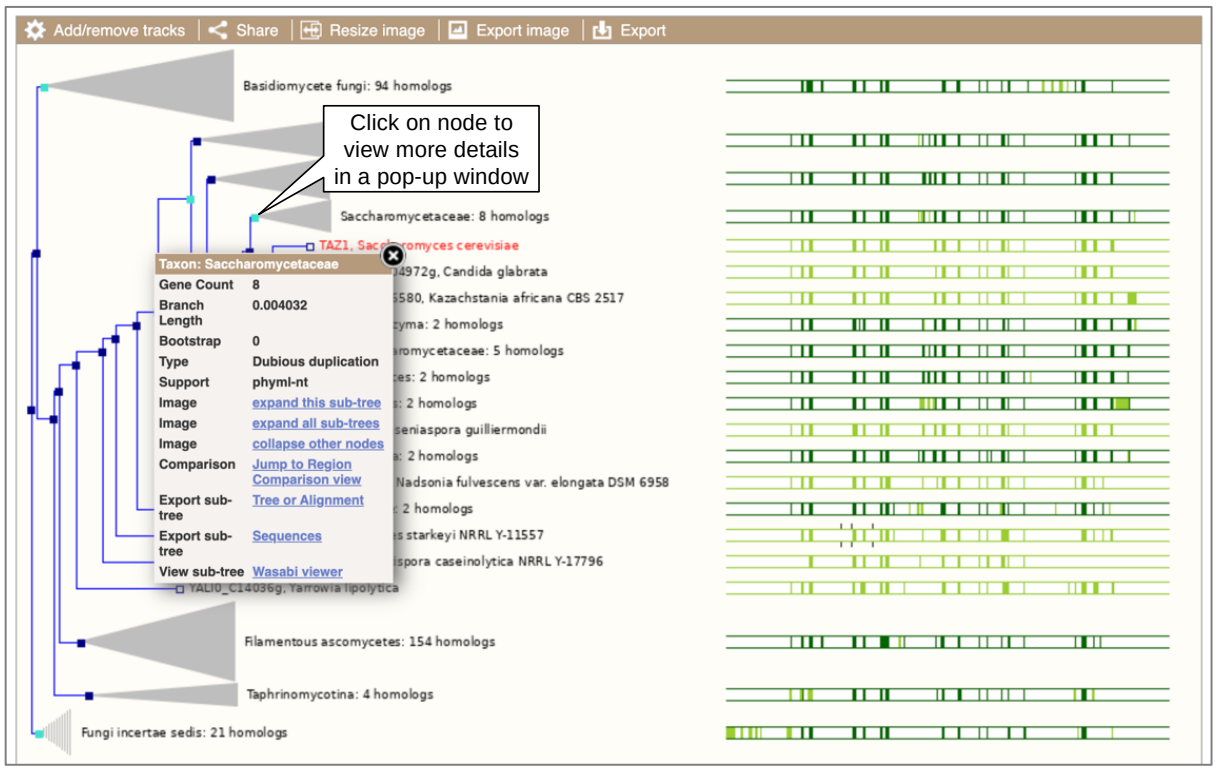
(a) How many duplication events are there in this tree?

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.



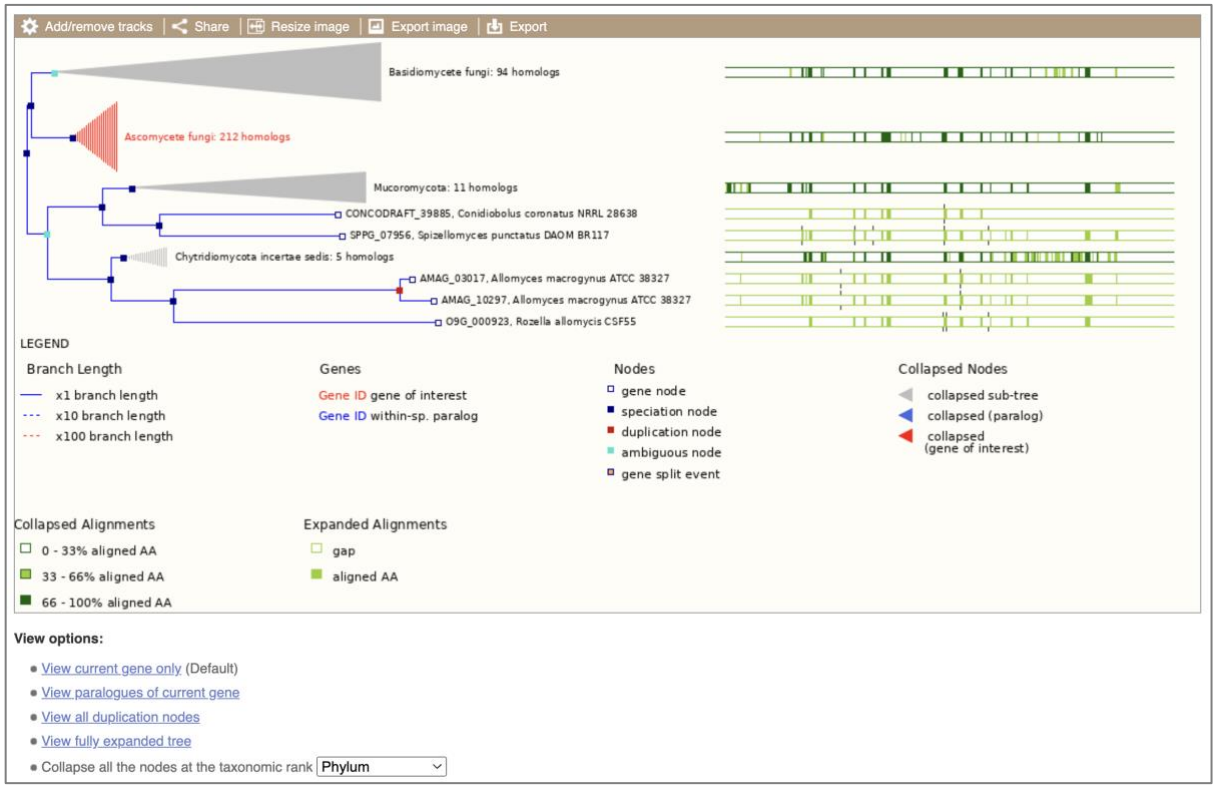
Click on View all duplication nodes. This will expand the tree so that all duplication nodes are visible. Count the number of red nodes. There are 13 duplication events in the tree.

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.

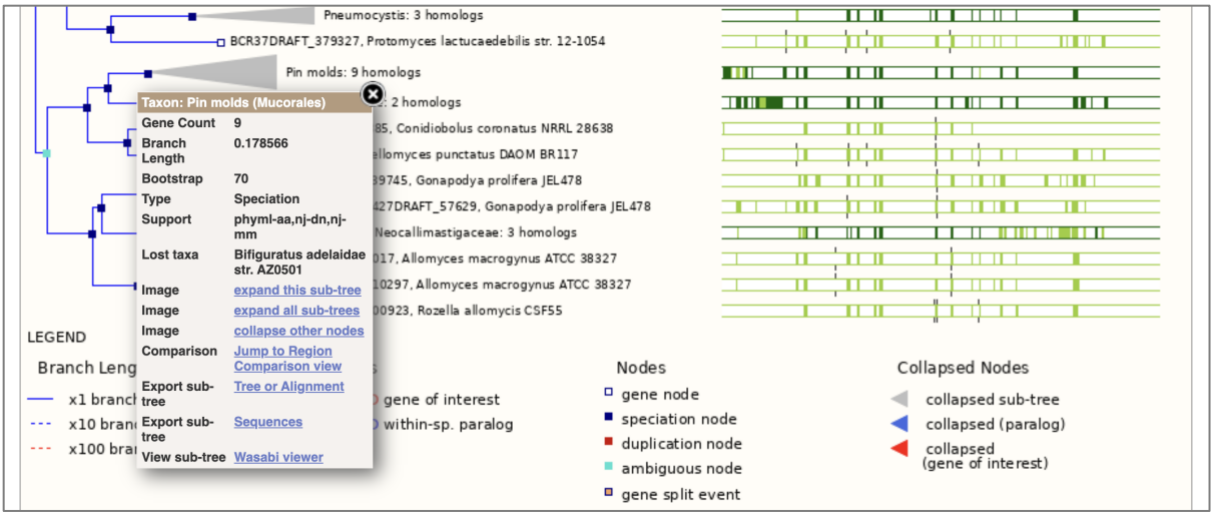


The phylum with the highest number of *TAZ1* homologues is Ascomycete fungi.

(c) What is the bootstrap support of the pin moulds (*Mucorales*) class in this tree?

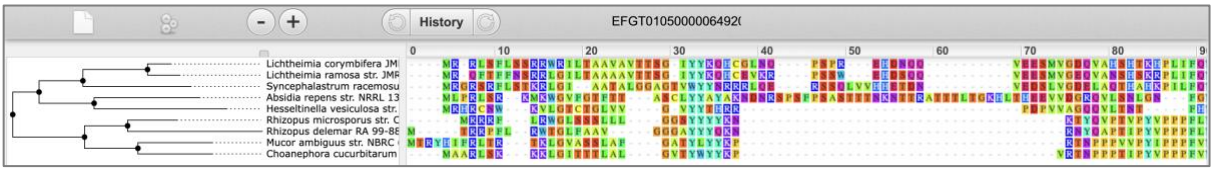
Bootstrap values in a phylogenetic tree indicate that out of 100, how many times the same branch is observed when repeating the generation of a phylogenetic tree on a resampled set of data. Bootstrap values in Ensembl gene trees are calculated using a tool called TreeBeST, and the final consensus trees 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.

Click on the Pin molds node to view more details. In the pop-up window, you will find the bootstrap value to be 70.

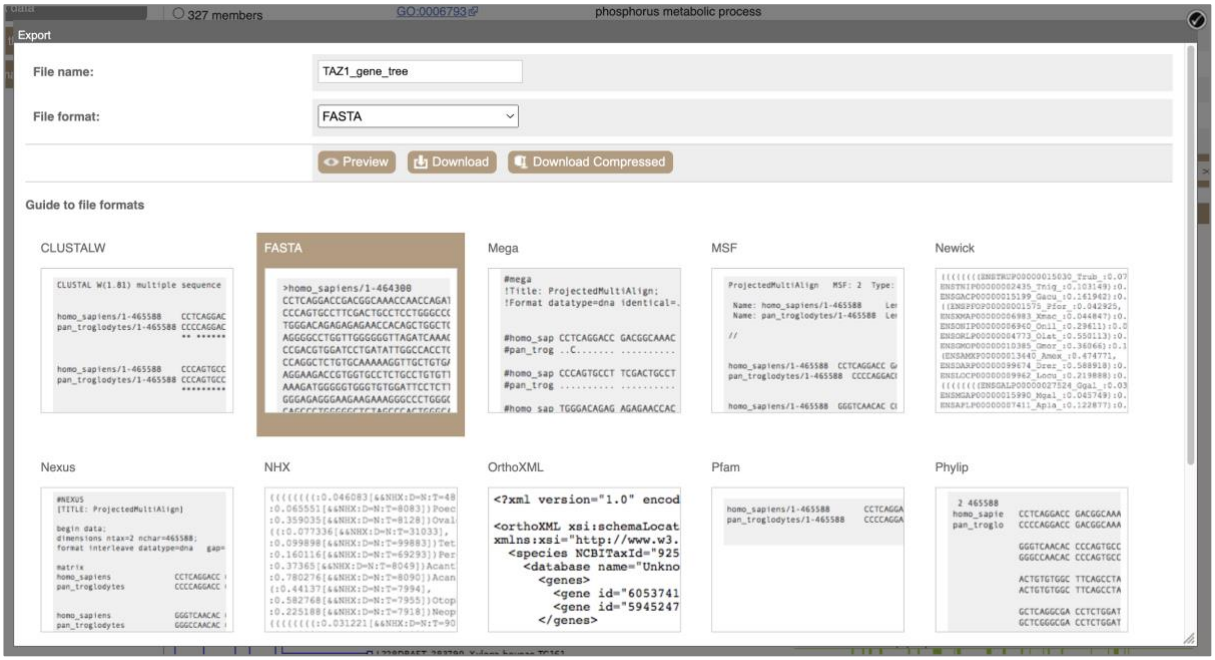


(d) Can you display the sequence alignment of all the homologues in this Class (*Hint: Use the Wasabi viewer*)?

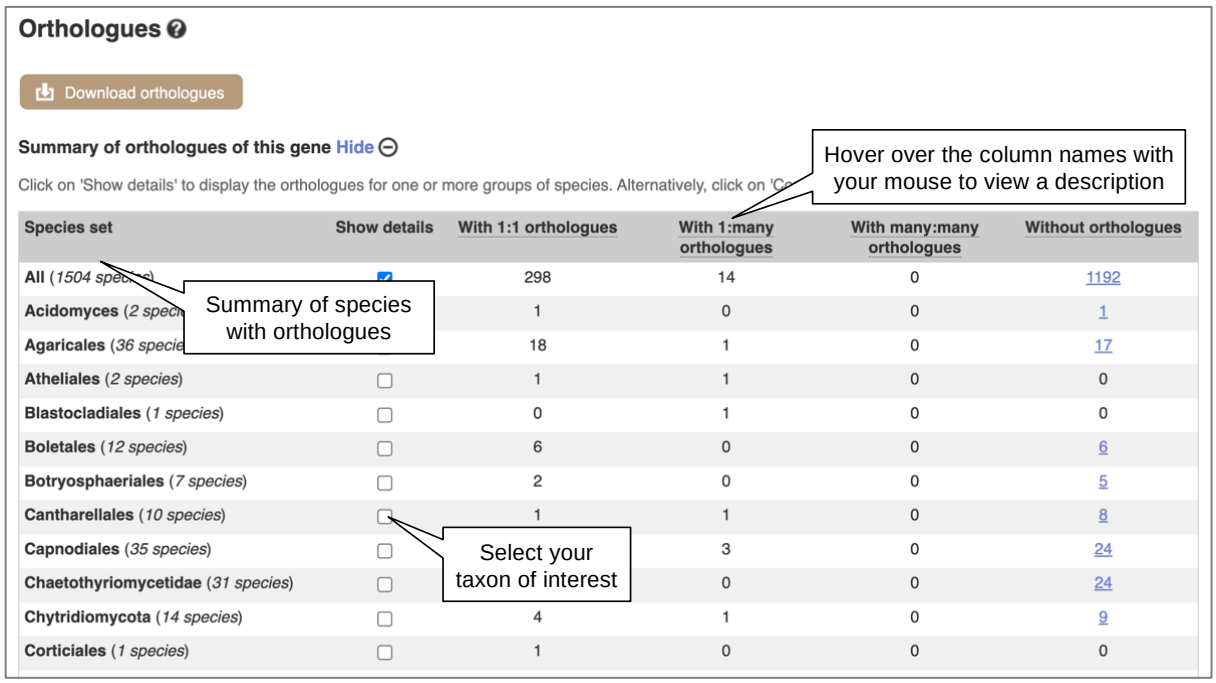
Wasabi is an open-source, web-based environment for visualising sequence data alongside phylogenetic trees. You can read more about the platform in this publication: https://europepmc.org/article/MED/26635364.



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.

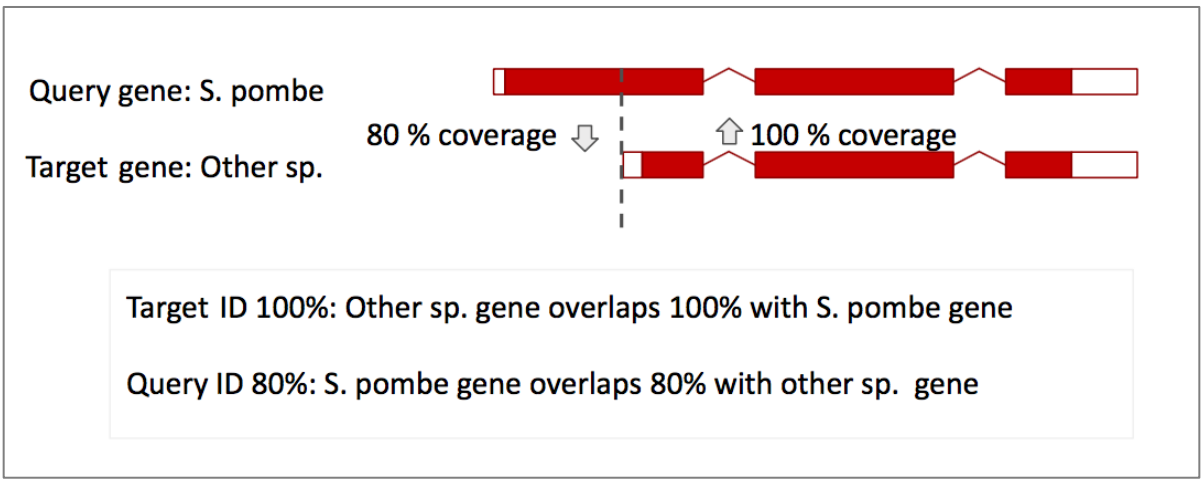


We can look at homologues in the Orthologues and Paralogues pages, which can be accessed from the left-hand menu. If there are no orthologues or paralogues, then the link(s) will be greyed out. Click on Orthologues to see the orthologues available.

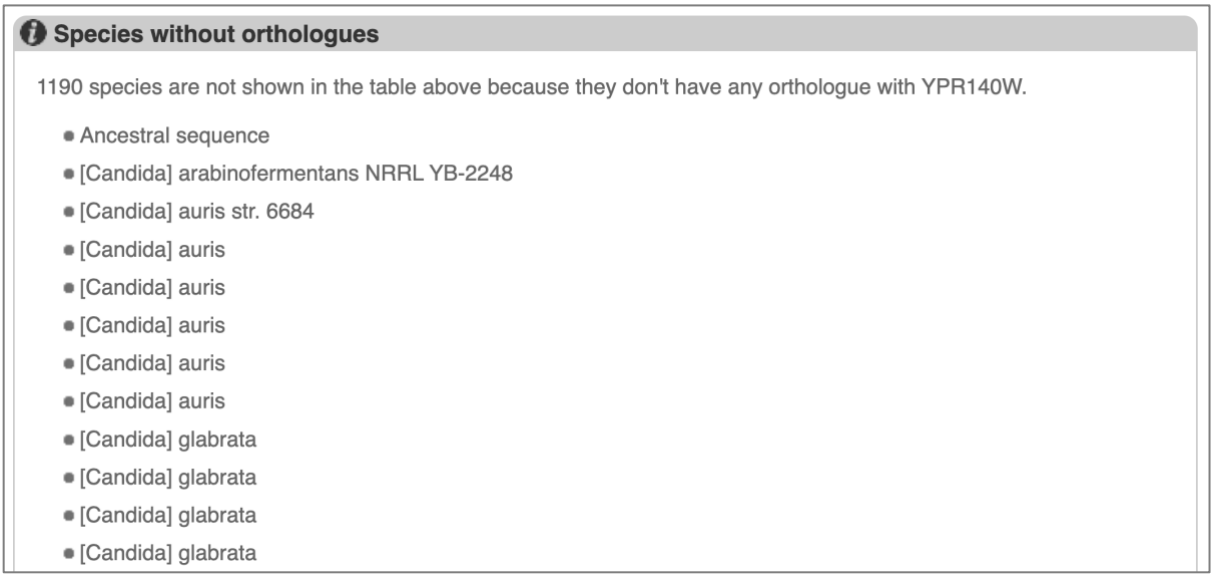


(e) What is the difference between Target %id and Query %id? (*Hint: Mouse over*)

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!



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.

Click on Pan-taxonomic Compara: Orthologues.

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 '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*)	☑	4	0	0	1500
Acidomyces (*2 species*)	☐	0	0	0	2
Agaricales (*36 species*)	☐	0	0	0	36
Atheliales (*2 species*)	☐	0	0	0	2
Blastocladiales (*1 species*)	☐	0	0	0	1
Boletales (*12 species*)	☐	0	0	0	12
Botryosphaeriales (*7 species*)	☐	0	0	0	7
Cantharellales (*10 species*)	☐	0	0	0	10

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 View Gene Tree	AAEL001564 2:21,496,991-21,541,309:-1 View Sequence Alignments	23.81 %	18.37 %	n/a	n/a	No
Amborella trichopoda	1-to-1 View Gene Tree	AMTR_s00022p00068080 AmTr_v1.0_scaffold00022:710,032-717,504:-1 View Sequence Alignments	23.43 %	17.59 %	n/a	n/a	No
Amphimedon queenslandica (Demosponge)	1-to-1 View Gene Tree	LOC100632622 GL345242.1:108,662-110,163:-1 View Sequence Alignments	24.73 %	18.11 %	n/a	n/a	No
Anopheles gambiae (African malaria mosquito, PEST)	1-to-1 View Gene Tree	AGAP007599 2L:48,133,715-48,137,634:-1 View Sequence Alignments	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*)	☑	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*)	☑	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?

Selected orthologues Hide

Show All entries | Show/hide columns | human

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Human	1-to-1	TAFAZZIN (ENSG00000102125)	24.66 %	18.90 %	n/a	n/a	No
	View Gene Tree	X:154,411,524-154,421,726:1					
		View Sequence Alignments					
Pediculus humanus	1-to-1	PHUM309640		18.90 %	n/a	n/a	No
	View Gene Tree	DS235308:45,836-47,144:-1					
		View Sequence Alignments					

Orthologue Aligment
View Protein Alignment
View cDNA Alignment

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	ENSP00000469981	292 aa	24 %	85 %	X:154411524-154421726

```
CLUSTAL W (1.81) multiple sequence alignment


YPR140W/1-381          MSFRDVL-------ERGDEFLEAYPRRS----PLWRFLSYSTSLLTFGVSKLLLFTCYNV
ENSP00000469981/1-292  -------MPLHVKW--------PFP---AVPPLTWTLASSVVMGLVGTYSCFWTKYMNHL
                                 .:*         * : *  .  *.   * :      ::


YPR140W/1-381          KLNGFEKLETALERSKRENRGLMTVMNHMSMVDDPLVWATLPYKLFTSLDNIRWSLGAHN
ENSP00000469981/1-292  TVHNREVLYELIEK-RGPATPLITVSNHQSCMDDPHLWGILKLRHIWNLKLMRWTPAAAD
                       .::. * *   :*: :     *:** ** * :*** :*. *  : : .*. :**: .* :


YPR140W/1-381          ICFQNKFLANFFSLGQVLSTER----------------------------FGVGPFQGS
ENSP00000469981/1-292  ICFTKELHSHFFSLGKCVPVCRGAEFFQAENEGKGVLDTGRHMPGAGKRREKGDGVYQKG
                       *** ::: ::*****: :.. *                              * * :* .
```

Additional Exercise 1: *Zymoseptoria* Orthologues

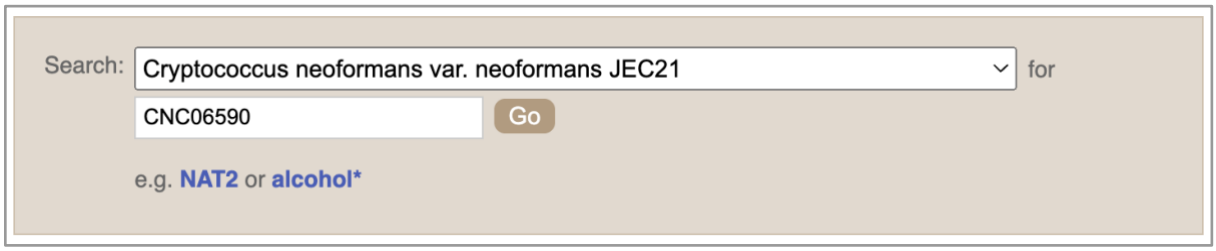
We will now explore an orthologue that we identified using BioMart (additional exercise 1 in the BioMart module). We identified 18 genes associated with the GO term detoxification in *Zymoseptoria tritici*. We then found a single high-confidence orthologue in *Cryptococcus neoformans* (CNM01690) which we will now explore further.

Search for CNM01690 in *Cryptococcus neoformans* var. *neoformans* JEC21 and go to the gene page.

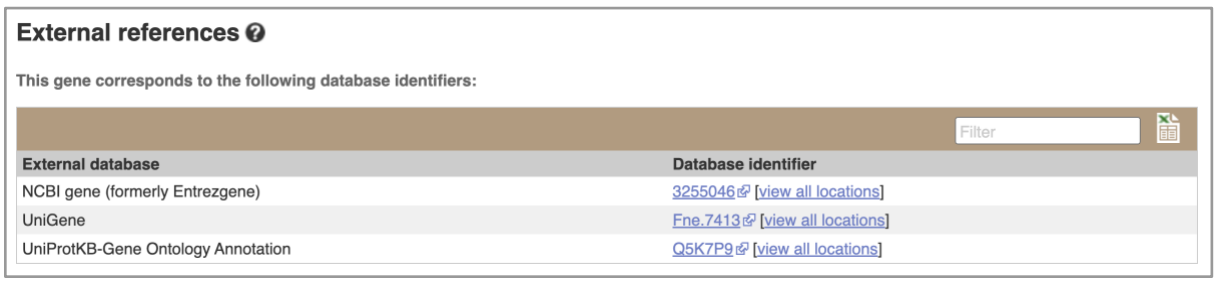
(a) Does this gene in *C. neoformans* have a UniProtKB-Gene Ontology annotation?
(b) Find the *Z. tritici* orthologue in the Orthologues page and view a protein alignment.
(c) At which end of the protein (N- or C-terminus) does the alignment between these two genes become worse?

Additional Exercise 1 Answer: *Zymoseptoria* Orthologues

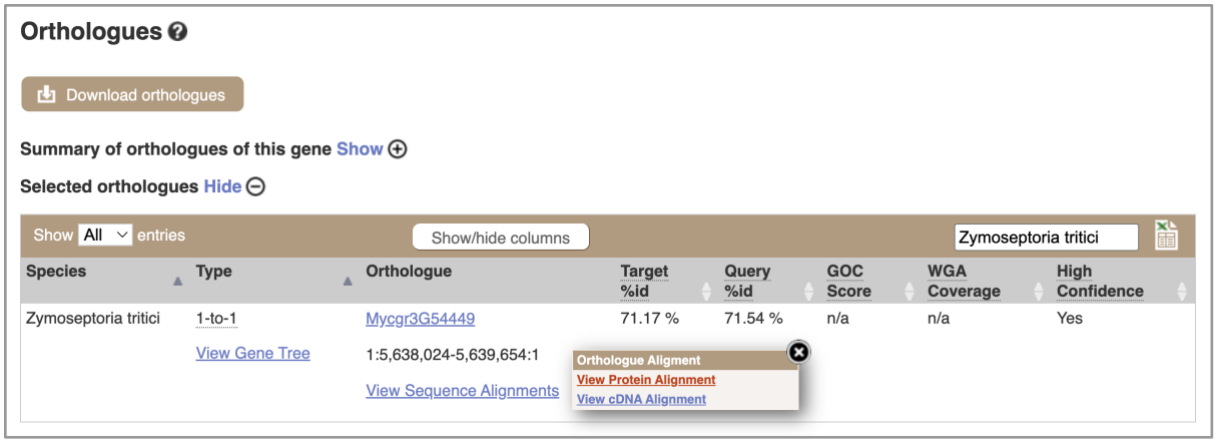
Go to fungi.ensembl.org in your browser. In the species-specific search box, select *Cryptococcus neoformans* var. *neoformans* JEC21 from the drop-down list and enter CNM01690. In the results page, click on the gene stable ID CNM01690 to navigate to the 'Gene' tab.



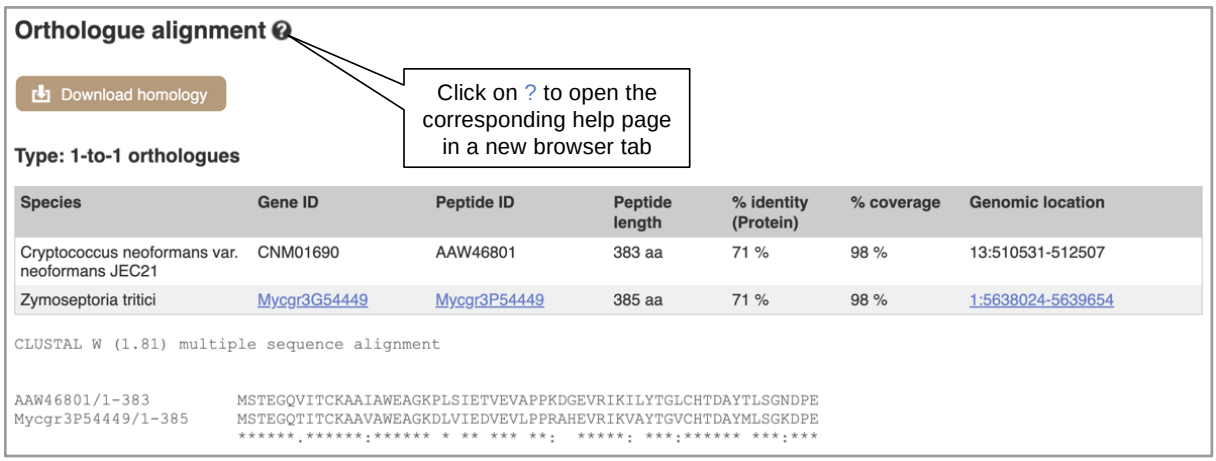
(a) In the left-hand panel under 'Gene-based displays', click on External references. Yes, the this gene has a UniProtKB-Gene Ontology annotation. The database ID for the UniProtKB-Gene Ontology annotation is Q5K7P9.



(b) 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.



(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.



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:

```
CLUSTAL W (1.81) multiple sequence alignment


AAW46801/1-383          MSTEGQVITCKAAIAWEAGKPLSIETVEVAPPKDGEVRIKILYTGLCHTDAYTLSGNDPE
Mycgr3P54449/1-385      MSTEGQTITCKAAVAWEAGKDLVIEDVEVLPPRAHEVRIKVAYTGVCHTDAYMLSGKDPE
                        ******.******:****** * ** *** **:  *****: ***:****** ***:***


AAW46801/1-383          GAFPVILGHEGGGIVESVGEGVDNVKVGDHVVPLYTAECRECKFCKSGKTNLCGRVRTTQ
Mycgr3P54449/1-385      GAFPVIAGHEGAGIVESIGEGVTNVKVGDTVVALYTPECKECKFCKSGKTNLCGKIRATQ
                        ****** ****.*****:**** ****** **.***.**:***************::*:**


AAW46801/1-383          GKGVMPDGTTRFKCKGQDILHFMGCSTFAQYTVVSKFSVVAINPKAPLKTSCLLGCGITT
Mycgr3P54449/1-385      GKGVMPDGSSRFRCKGKDLLHFMGCSTFSQYTVVADISVVAVTDKAPMDRTCLLGCGITT
                        ********::**:***:*:*********:*****:.:****:. ***:. :*********


AAW46801/1-383          GYGAATKSP---GI-EGSNVAIFGVGCVGLSVLQGAKAKGCKRIFAIDTNPKKKEWAVKF
Mycgr3P54449/1-385      GYGAATITAGKNGVEKGDNVAVFGAGCVGLSVIQGAASRNAGKIIVVDVNDSKKEWASKF
                        ****** :.   *: :*.***:**.*******:*** ::.. :*:.:*.* .***** **


AAW46801/1-383          GATDFINP-KDLPEGKTIVDYLIEETDGGLDFTFDATGNVGVMRNALEACHKGWGVCTII
Mycgr3P54449/1-385      GATDFVNPTKDLKEGEKIQDRLVEMTDGGCDYTFDCTGNVHVMRSALEACHKGWGESIII
                        *****:** *** **:.* * *:* **** *:***.**** ***.********** . **


AAW46801/1-383          GVAPAGAEISTRPFQLVTGRVWKGSAFGGVKGRTELPGIVEDYLAGKLWVNEFVTHNETL
Mycgr3P54449/1-385      GVAAAGQEIATRPFQLVTGRVWKGCAFGGVKGRSQMGGLIDDYMGGKLKVDEFITHRQNL
                        ***.** **:***************.********::: *:::**:.*** *:**:**.:.*


AAW46801/1-383          EGINKGFDDMHAGDCIRCVVDMGF-NEAP
Mycgr3P54449/1-385      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

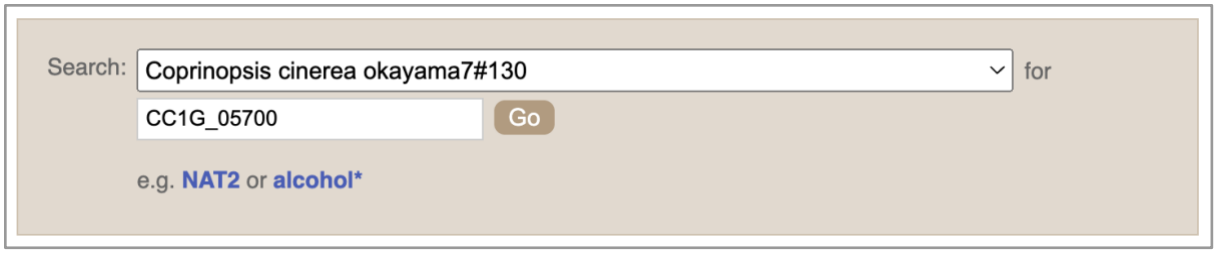
We're going to take a look at the gene CC1G_05700 in *Coprinopsis cinerea* okayama7#130.

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.

(a) What do you notice about the types of fungi shown in the gene tree?
(b) Does this match with what you would expect from the gene description? (*Hint: Agaricomycetes class belongs to the Basidiomycota phylum*)
(c) Based on the protein alignment shown at the right, can you predict which end of the gene/protein is most conserved?
(d) 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

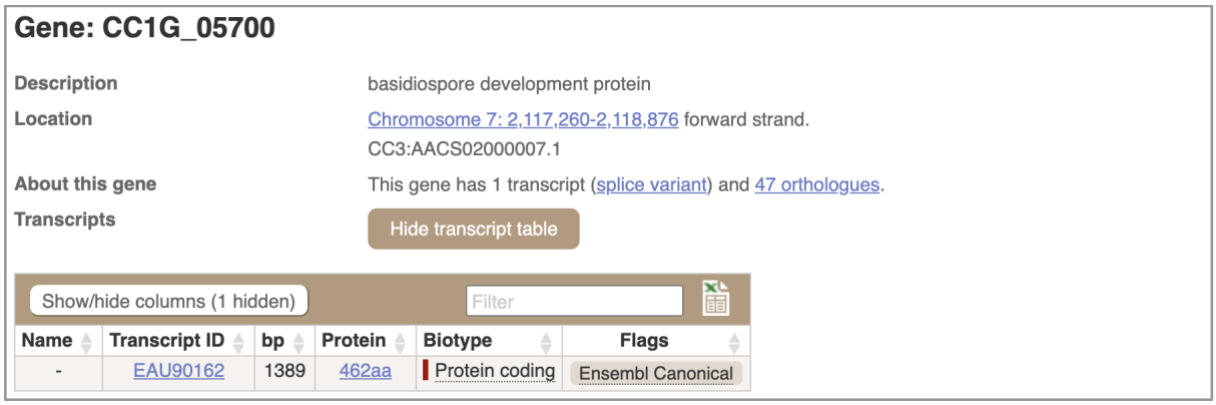
Go to fungi.ensembl.org in your browser. In the species-specific search box, select *Coprinopsis cinerea* okayama7#130 from the drop-down list and enter CC1G_05700.



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'.

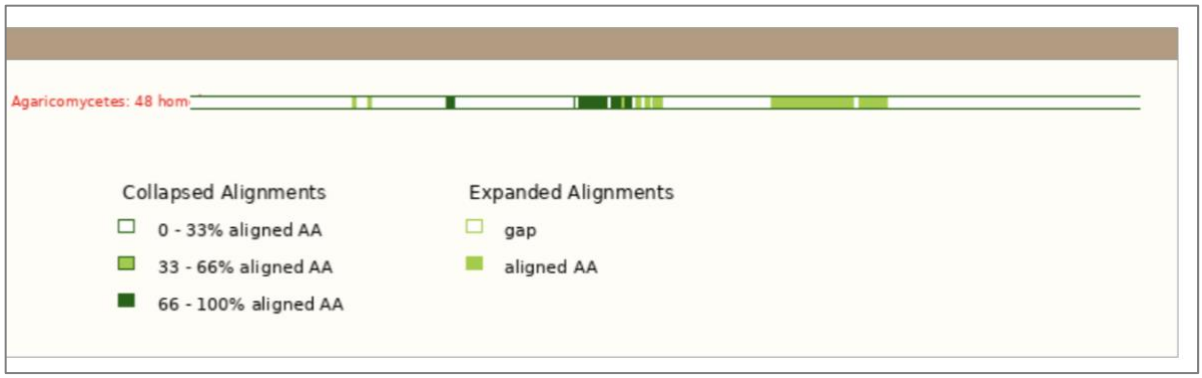
(a) All fungi shown in the gene tree are Agaricomycetes:

(b) The gene description is as follows:

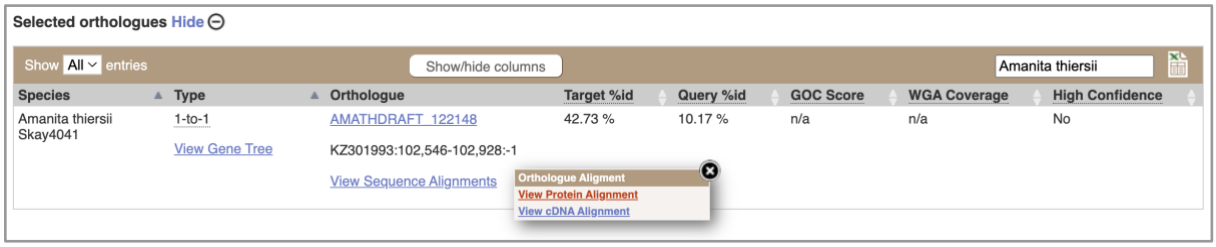


The class Agaricomycetes belongs to the phylum Basidiomycota, therefore we would expect the gene encoding the basidiospore development protein to be conserved across Agaricomycetes species.

(c) Dark green regions in the alignment indicate highly conserved sequences (see 'Collapsed Alignments' legend):



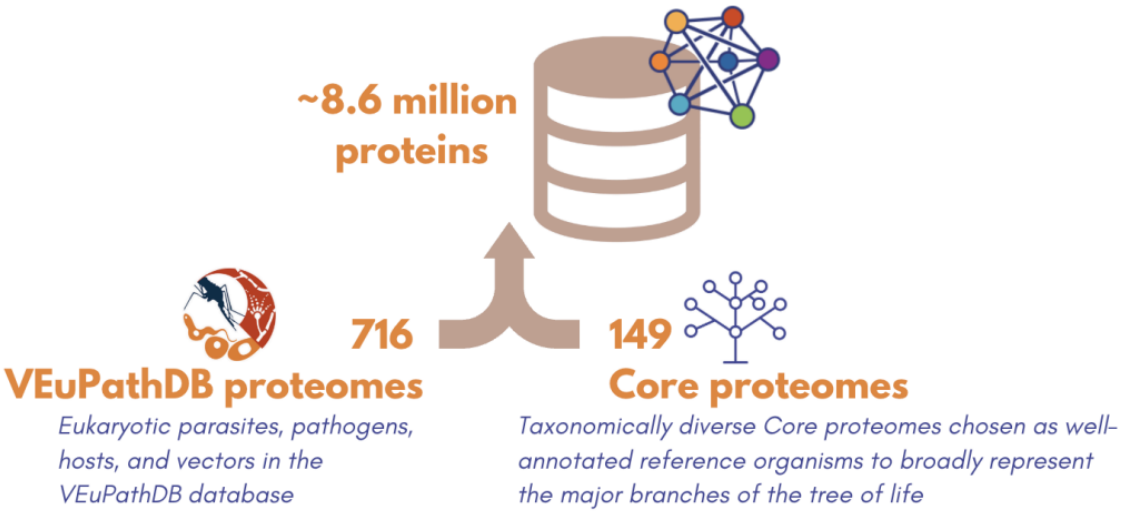
(d) Go to Fungal Compara: Orthologues in the left-hand panel. Click on the Hide ⊖ button for the 'Summary of orthologues of this gene' table. In the 'Selected orthologues' table, use the search bar in the top right-hand corner to search for *Amanita thiersii*. Click on View Sequence Alignments and in the pop-up menu select View Protein Alignment.



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.

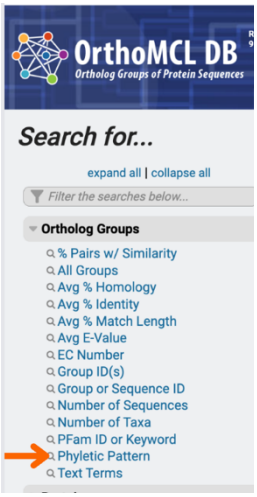


OrthoMCL 7 is a genome-scale database and website (orthomcl.org) that uses protein sequence similarity and phylogenetic relationships among proteins to create groups of orthologous protein sequences called orthogroups. Orthogroups are clustered with OrthoFinder, which uses DIAMOND for much faster BLAST 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.



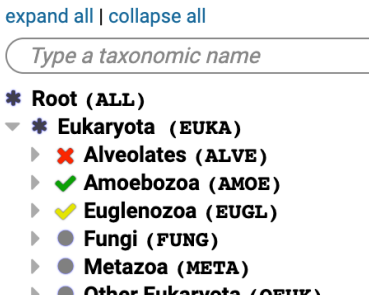
There are two ways to specify a phyletic pattern:

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).

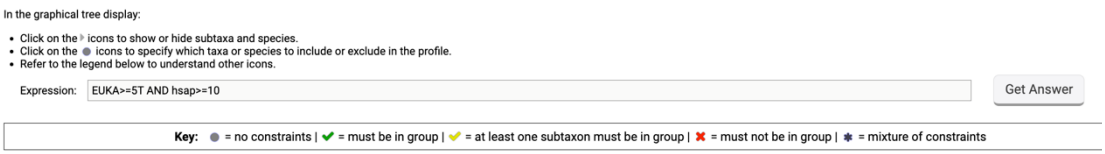
Note: A grey asterisk (*) is shown when multiple constraints are selected for sub-groups within a clade.



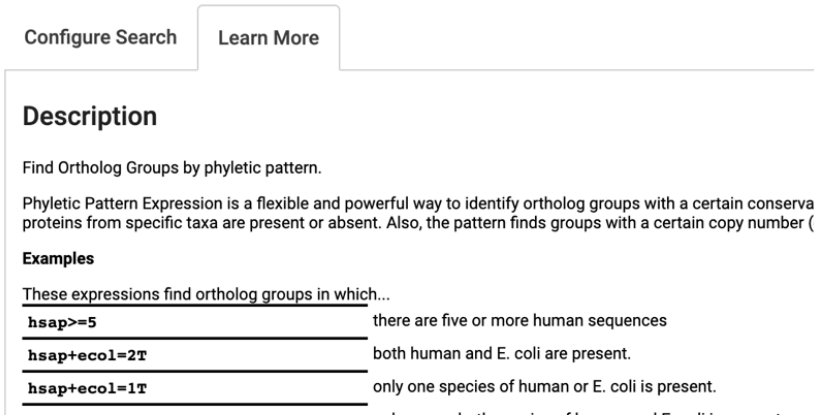
2. Using the expression box

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**.



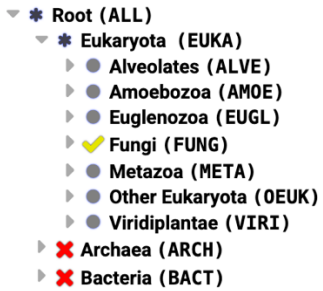
The phyletic expression syntax is explained on the **Learn More** tab of the search page.



Let's practice running a few searches:

A. Using the selectable tree, identify how many Fungal (FUNG) orthogroups <u>do not</u> contain any proteins from bacteria and archaea.

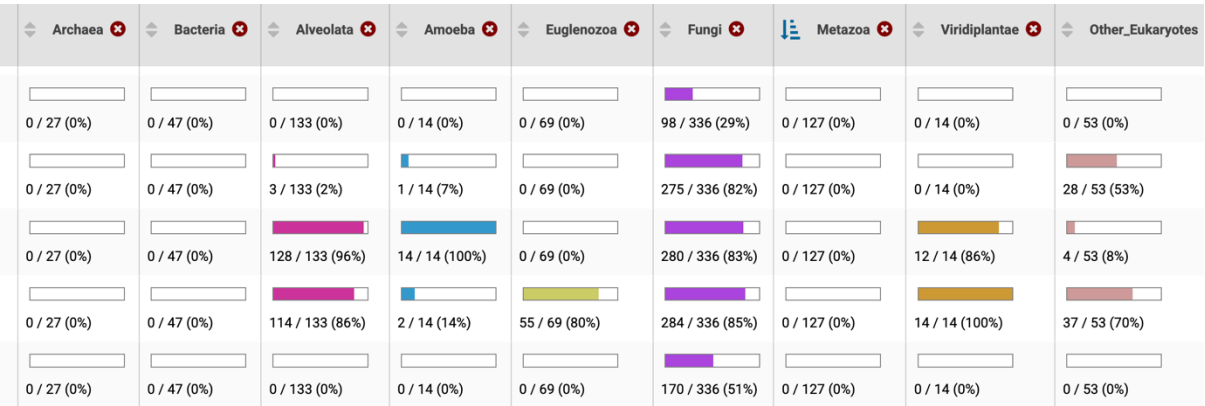
The yellow check box allows orthogroups that contain a protein in <u>any</u> taxa that is within Fungi. How many orthogroups are found?



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).

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]



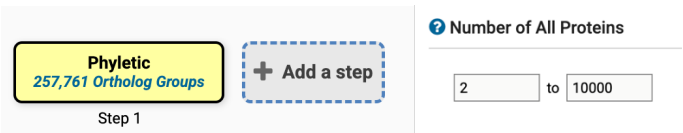
Archaea	Bacteria	Alveolata	Amoeba	Euglenozoa	Fungi	Metazoa	Viridiplantae	Other_Eukaryotes
0 / 27 (0%)	0 / 47 (0%)	0 / 133 (0%)	0 / 14 (0%)	0 / 69 (0%)	98 / 336 (29%)	0 / 127 (0%)	0 / 14 (0%)	0 / 53 (0%)
0 / 27 (0%)	0 / 47 (0%)	3 / 133 (2%)	1 / 14 (7%)	0 / 69 (0%)	275 / 336 (82%)	0 / 127 (0%)	0 / 14 (0%)	28 / 53 (53%)
0 / 27 (0%)	0 / 47 (0%)	128 / 133 (96%)	14 / 14 (100%)	0 / 69 (0%)	280 / 336 (83%)	0 / 127 (0%)	12 / 14 (86%)	4 / 53 (8%)
0 / 27 (0%)	0 / 47 (0%)	114 / 133 (86%)	2 / 14 (14%)	55 / 69 (80%)	284 / 336 (85%)	0 / 127 (0%)	14 / 14 (100%)	37 / 53 (70%)
0 / 27 (0%)	0 / 47 (0%)	0 / 133 (0%)	0 / 14 (0%)	0 / 69 (0%)	170 / 336 (51%)	0 / 127 (0%)	0 / 14 (0%)	0 / 53 (0%)

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.

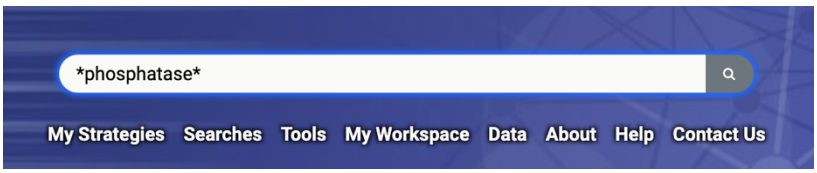
Step 1. Use the selection tree to choose only Fungi.
Step 2. +Add a step and use the **Number of Sequences** search for 2 to 10000 proteins.

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.

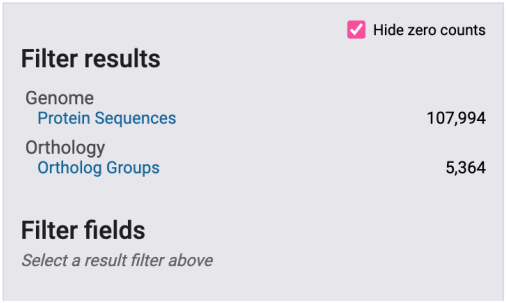


E. Find all fungal proteins that are likely to be phosphatases and that do not have orthologs outside of fungal kingdom.

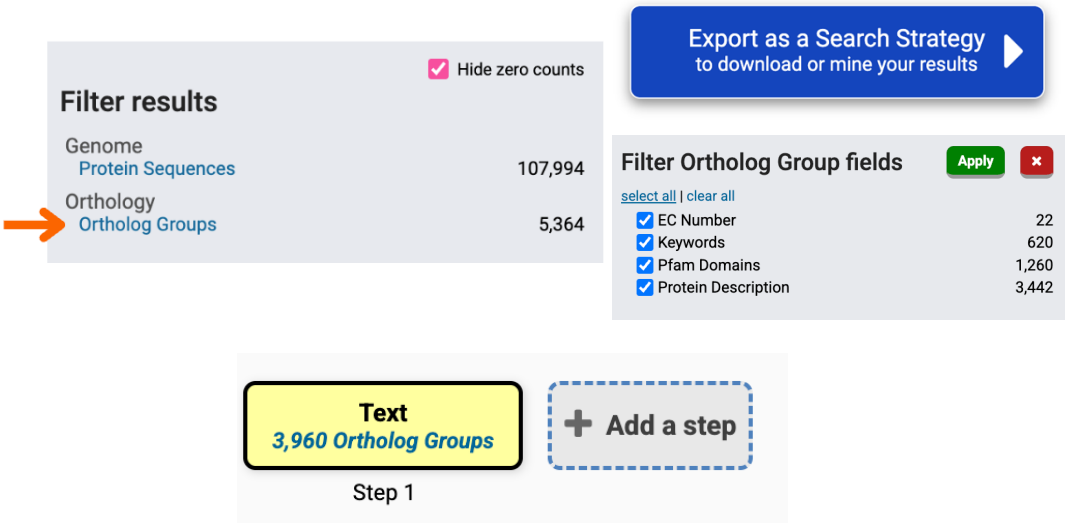
a. Use the site search to look for *phosphatase* (use asterisks to find any combination of the word "phosphatase").



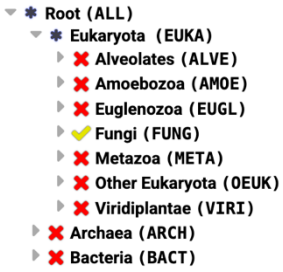
How many protein sequences were identified? How many ortholog groups did you identify?



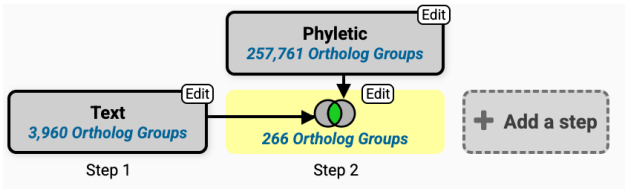
b. Display the ortholog groups containing the word **phosphatase** in **all ortho group fields** and export the results as a search strategy.



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).



How many groups did the search return?



Strategy URL:

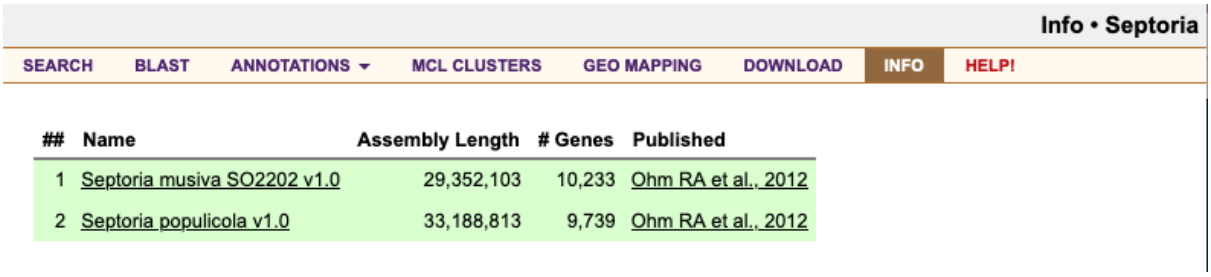
Strategies:
A: https://orthomcl.org/orthomcl/app/workspace/strategies/import/4764a59520b7e517
B: https://orthomcl.org/orthomcl/app/workspace/strategies/import/77a327acd20b5a2d
C: https://orthomcl.org/orthomcl/app/workspace/strategies/import/2a7a12997022153b
D: https://orthomcl.org/orthomcl/app/workspace/strategies/import/2a7a12997022153b
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 populicola*, 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:



##	Name	Assembly Length	# Genes	Published
1	Septoria musiva SO2202 v1.0	29,352,103	10,233	Ohm RA et al., 2012
2	Septoria populicola v1.0	33,188,813	9,739	Ohm RA et al., 2012

2) Click on '***Septoria musiva* SO2202 v1.0**' to go to its genome portal:

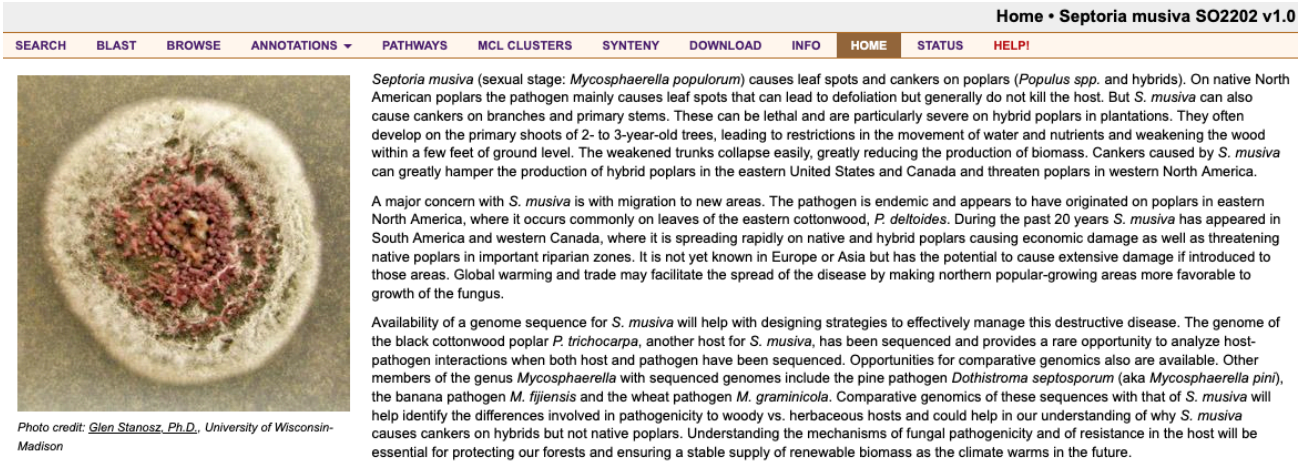


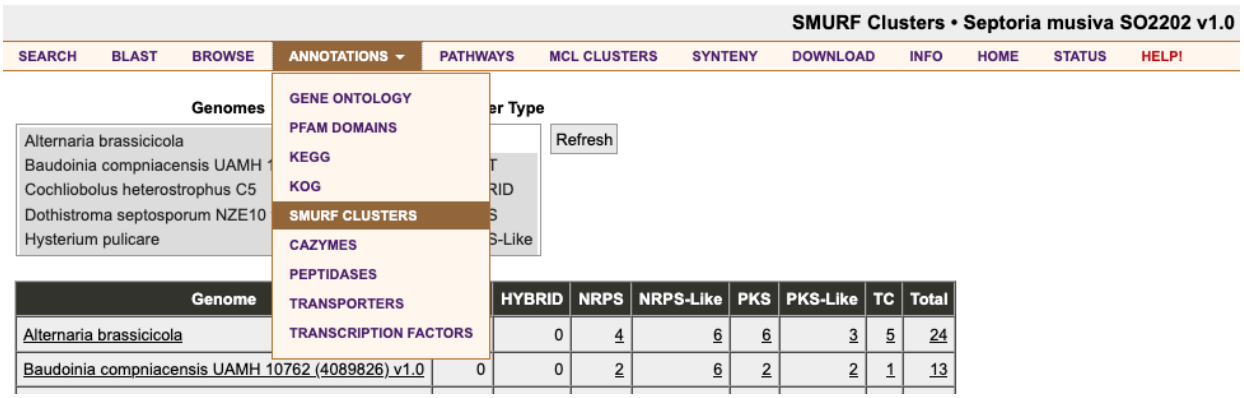
Photo credit: Glen Stanosz, Ph.D., University of Wisconsin-Madison

Septoria musiva (sexual stage: *Mycosphaerella populorum*) 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.

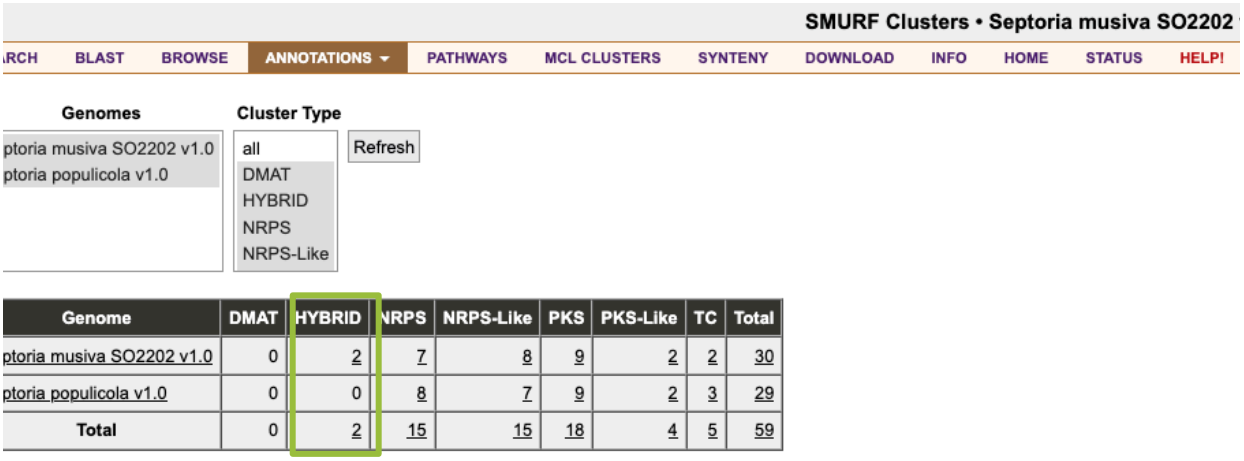
A major concern with *S. musiva* is with migration to new areas. The pathogen is endemic and appears to have originated on poplars in eastern North America, where it occurs commonly on leaves of the eastern cottonwood, *P. deltoides*. During the past 20 years *S. musiva* has appeared in South America 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.

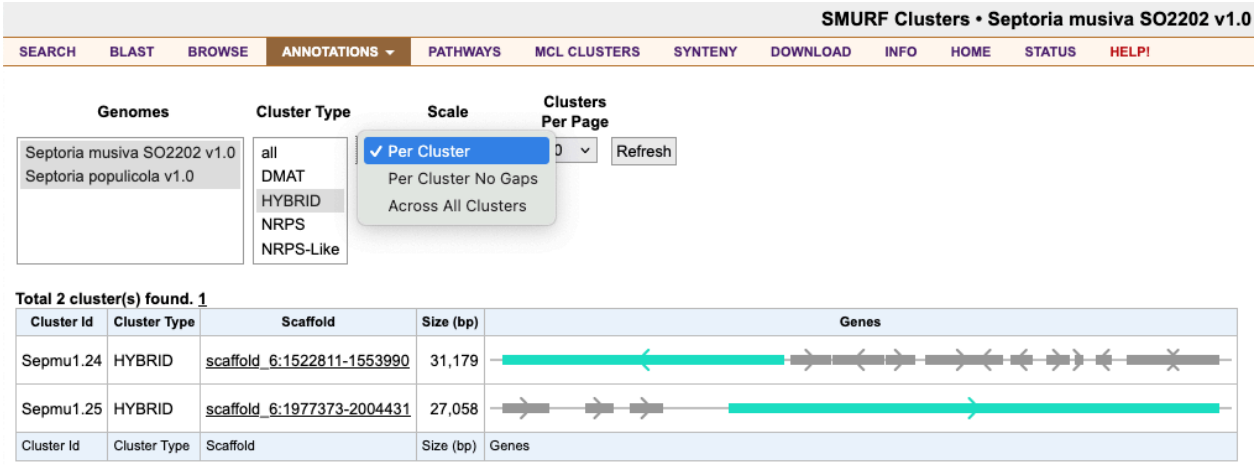
3) Click on "**ANNOTATIONS => SMURF CLUSTERS**" to go to the portal's SM clusters browser:



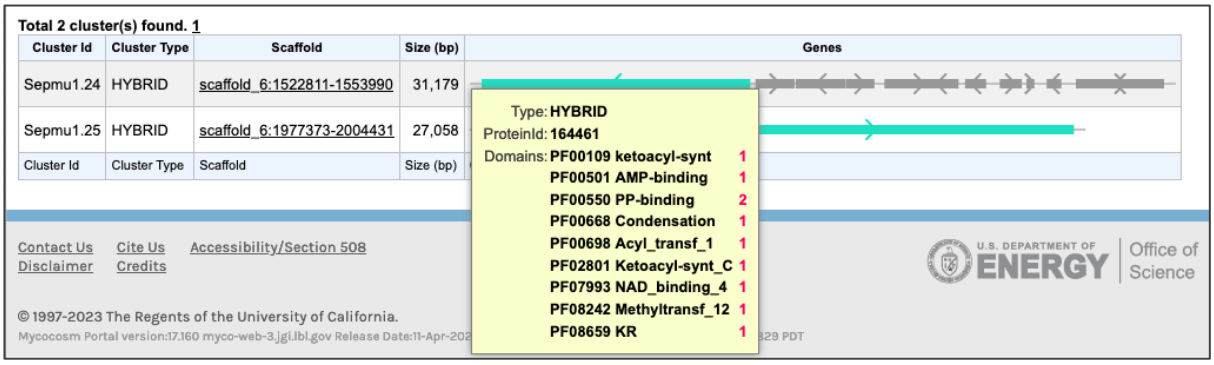
4) Scroll through the 'Genomes' list box and select both '*Septoria musiva*' and '*Septoria populica*', and only those 2 species. Click the '**Refresh**' button. Only the SM cluster core gene counts of the 2 *Septoria* spp. are shown, and may be directly compared. *S. musiva* has 2 hybrid core genes (PKS-NRPS genes) while *S. populicola* has none:



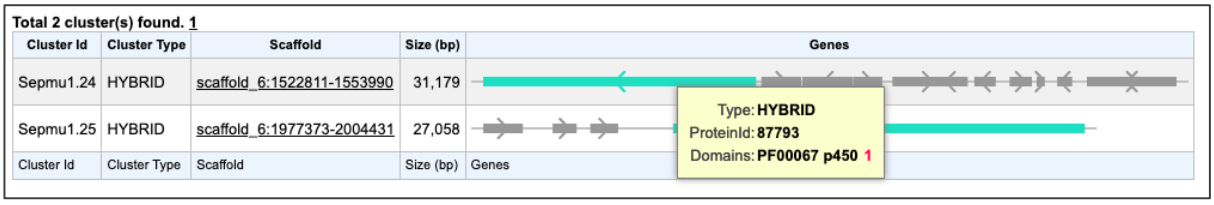
5) There are a total of 2 genes in the Hybrid column. Click on the number to show a graphical representation of the 2 gene clusters in *S. musiva*. The 'Size' column displays each cluster's length, and the 'Genes' column displays each cluster's core PKS-NRPS gene (in color) and its accessory, decorator, 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:



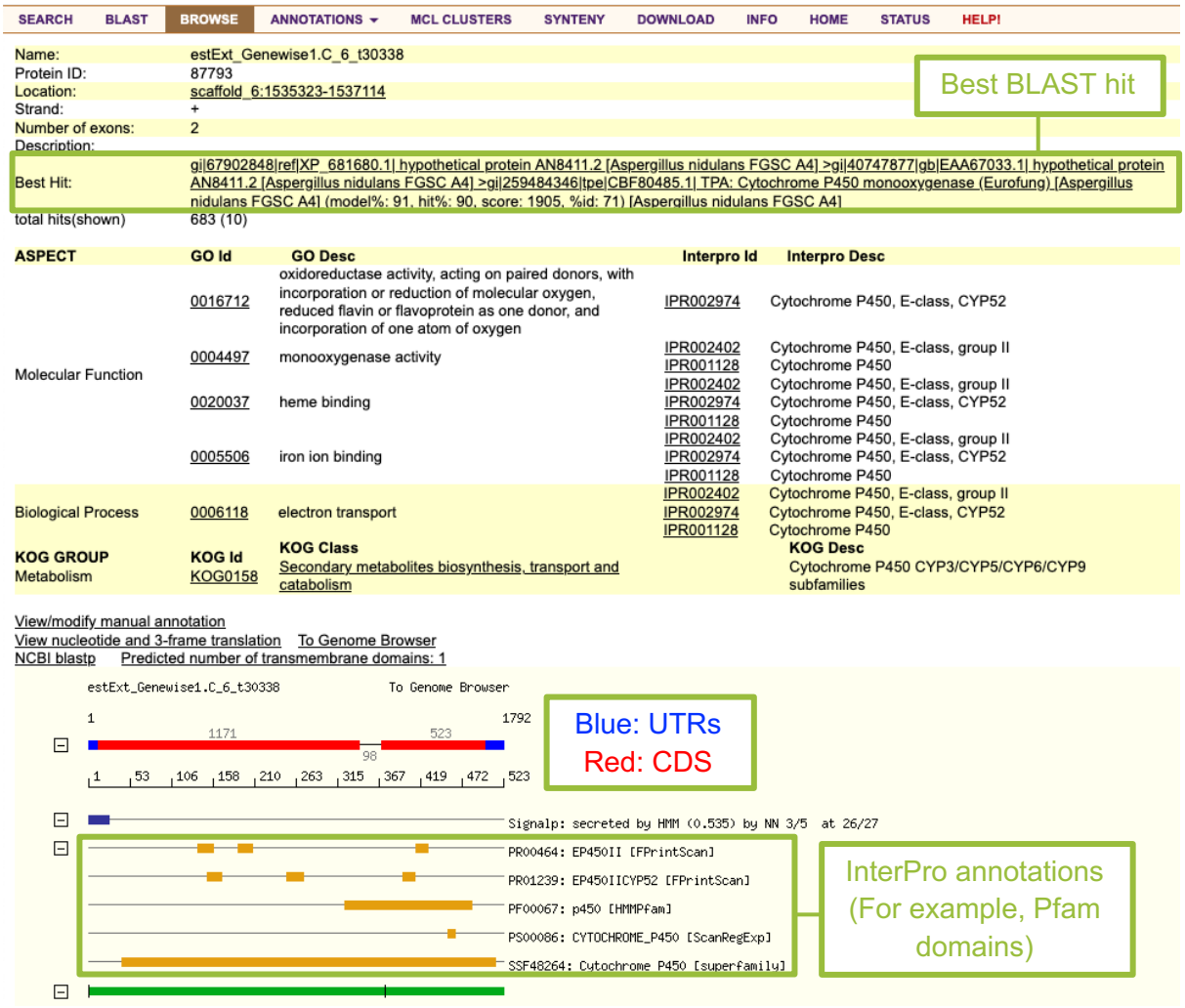
6) Each gene in the clusters is represented by an arrow with a single pair of fletching that indicates the gene's 5' to 3' direction. Mouse-over the top cluster's core gene to get more information about the PKS-NRPS hybrid. The listed domains are typical of a hybrid enzyme:



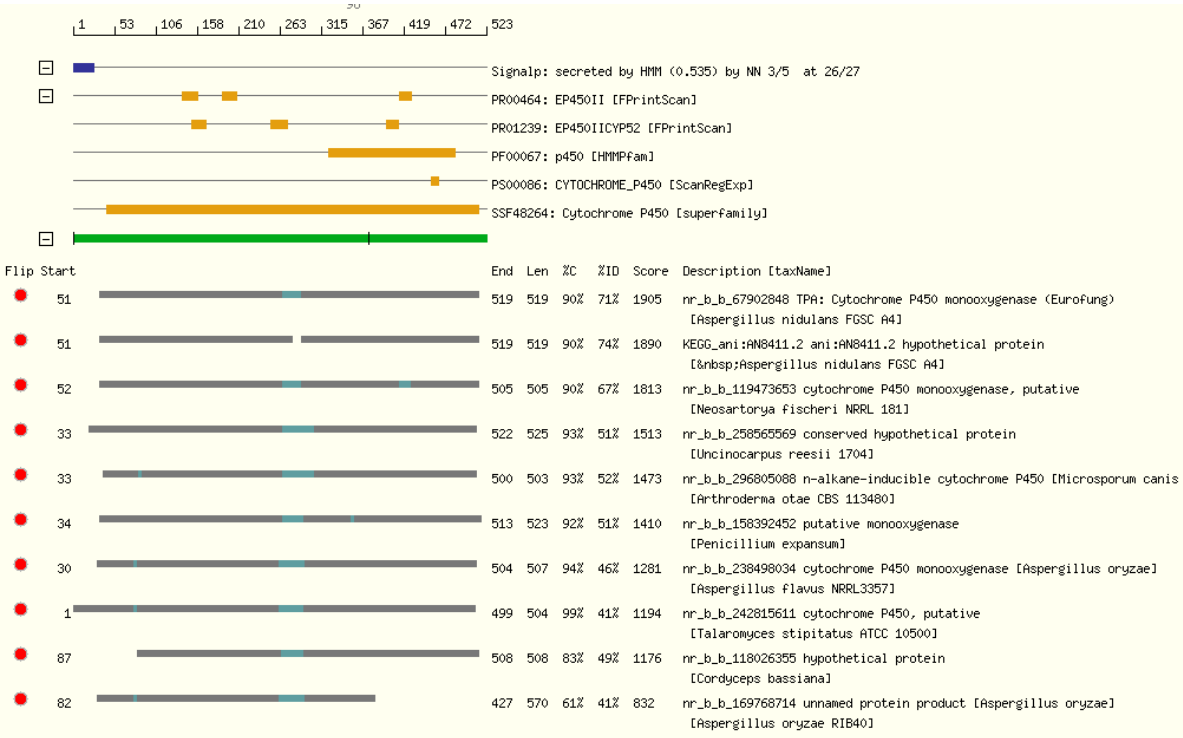
7) To get domain information about the other genes in the SM cluster, mouse-over them too. The next gene 3' to the core gene has a p450 domain:



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:
    a) The gene is expressed. The blue bars represent UTRs, which can be inferred only from transcriptomic data.
    b) The protein has p450 Pfam and other annotations indicative of a cytochrome p450 monooxygenase.
    c) 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).



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

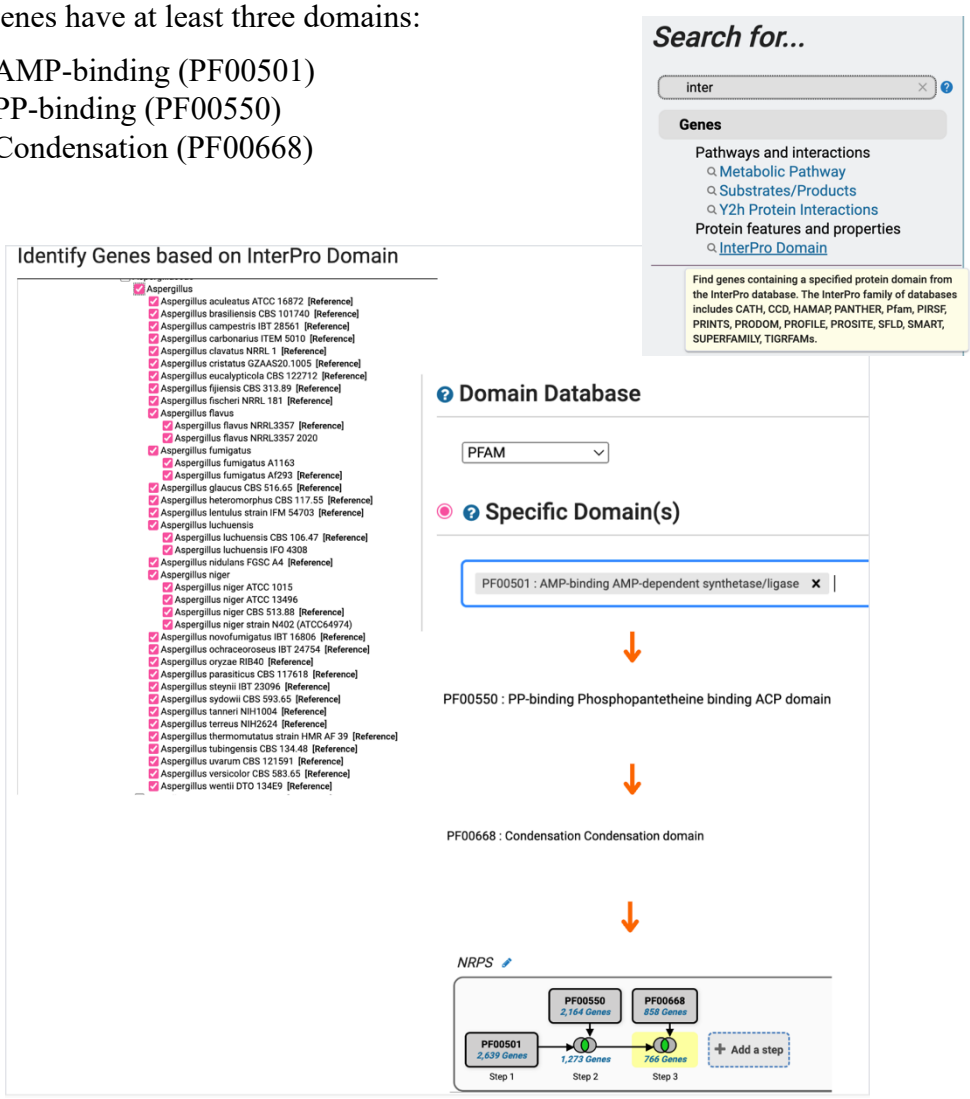
Fungi generate a wide variety of secondary metabolites. These metabolites can be categorised based on the initial stage of their biosynthesis, particularly the essential "key enzymes" involved: Non-ribosomal peptide synthetases (NRPSs), NRPS-like enzymes, Polyketide synthases (PKSs), PKS-like enzymes, 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:

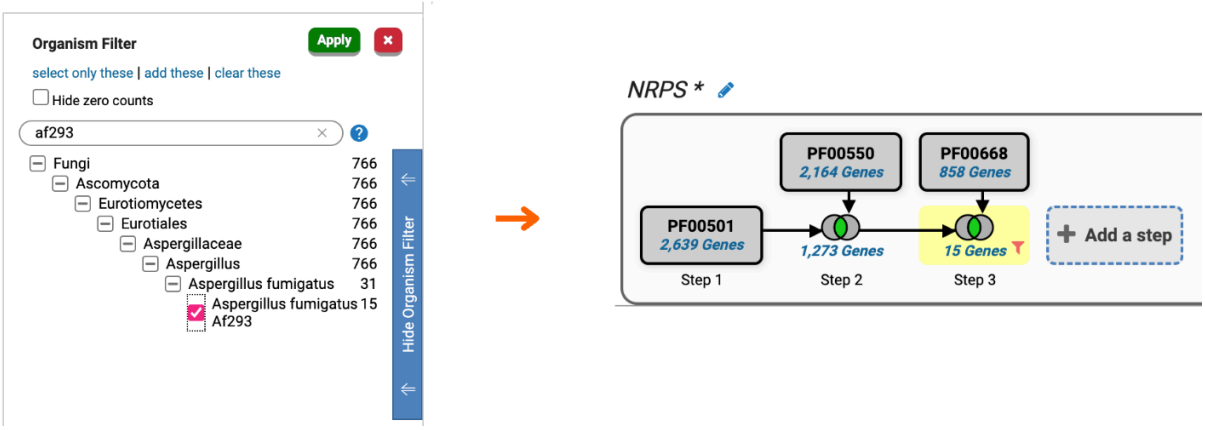
- AMP-binding (PF00501)
- PP-binding (PF00550)
- Condensation (PF00668)



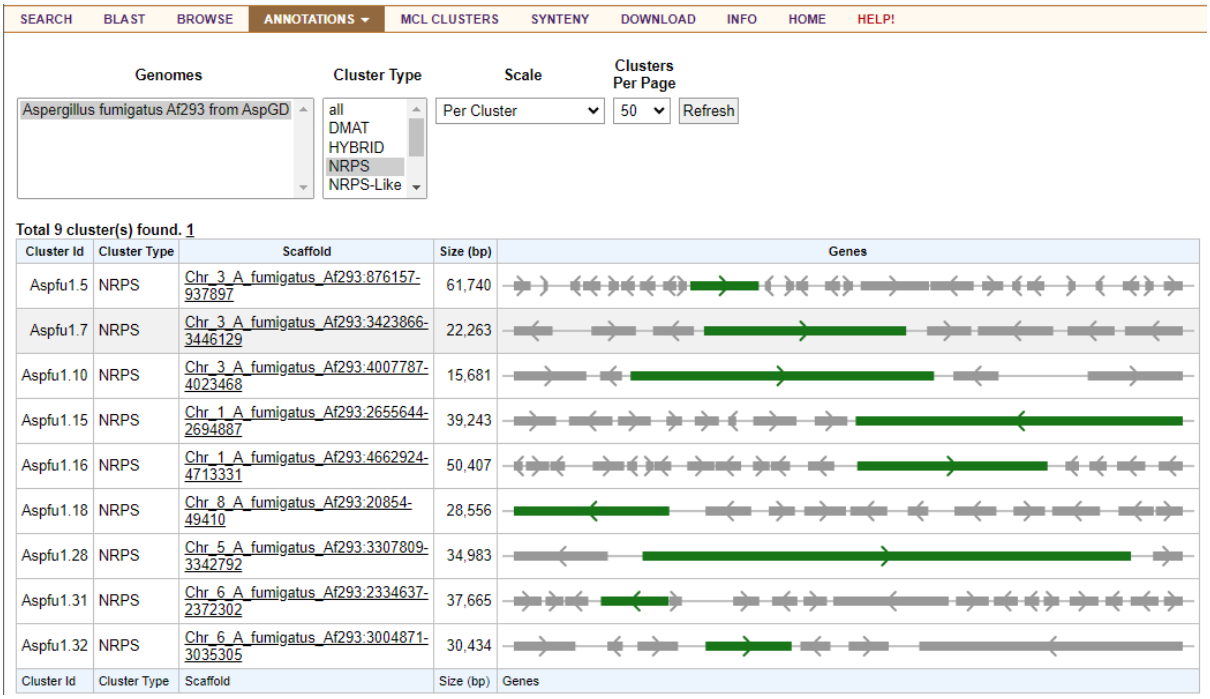
Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/85a1e3a5a603efc6

- 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?



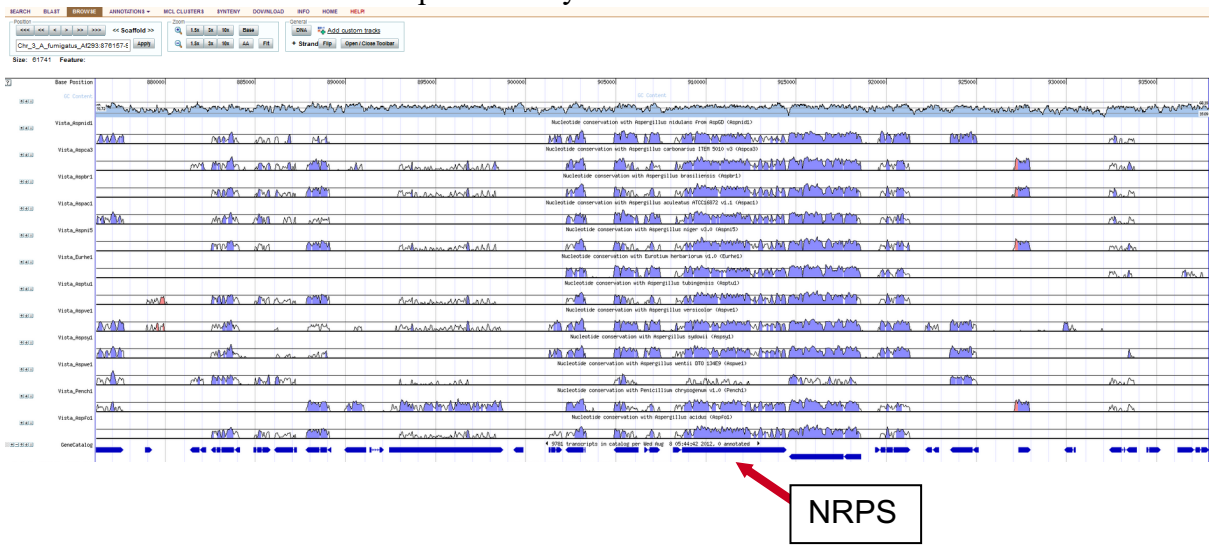
- What do you think may be causing the difference in the predicted gene number?

This view on MycoCosm allows you to analyze backbone and auxiliary proteins across the entire predicted secondary metabolism cluster.

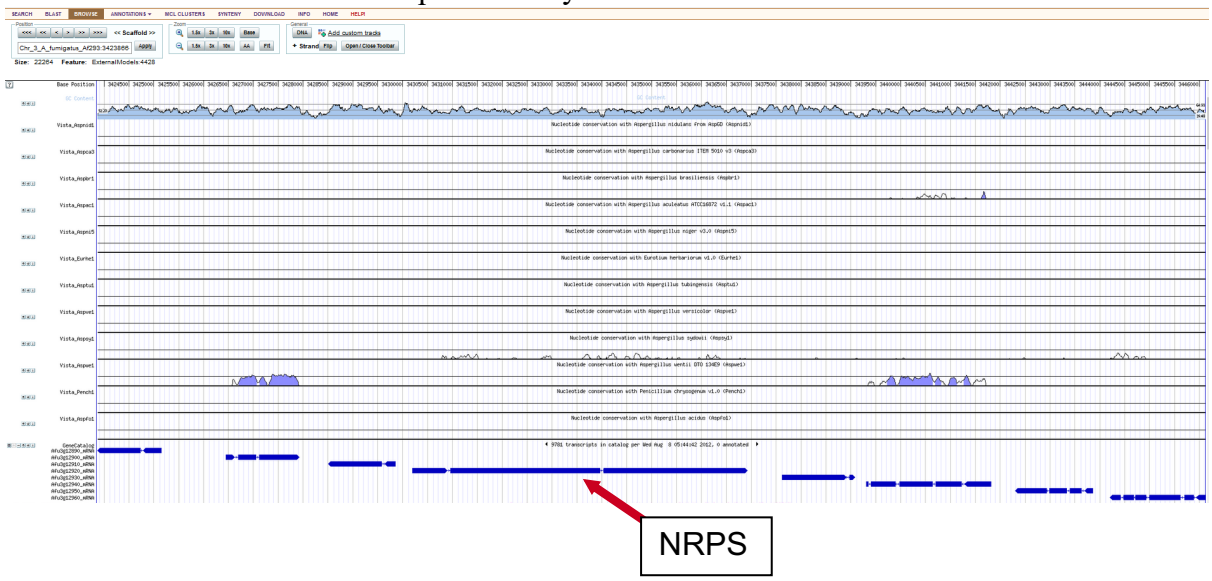
- How conserved are these secondary metabolite clusters across related Aspergilli?
- Click on the scaffold coordinates for Aspfu1.5 and analyze the Vista curve tracks in the genome browser.
- How many related Aspergilli show some synteny with this region?
- Optional: Repeat this exercise for the next 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:



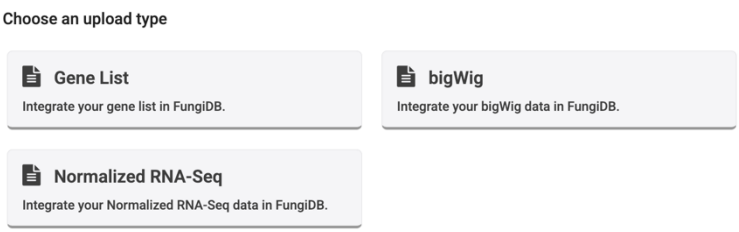
Working with user datasets in FungiDB

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),
2. **bigWig** files (as a .bw file),
3. **Normalized RNA-Seq data** (as a zip file containing RNA-seq data with or without the supporting bigwig files).



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.

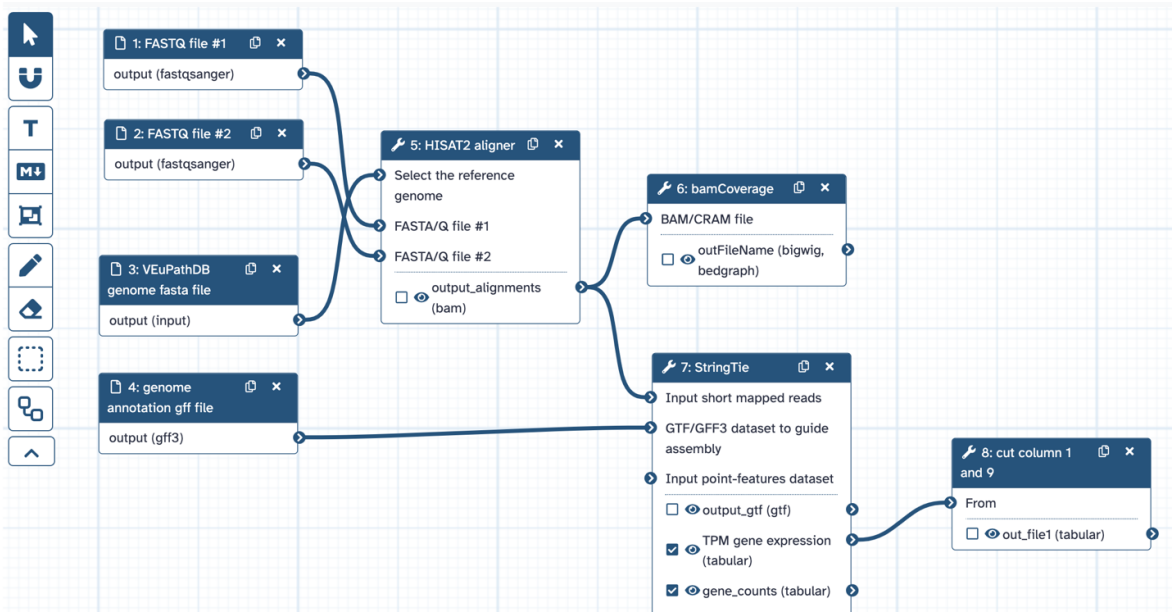
Data analysis in UseGalaxy.org

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.

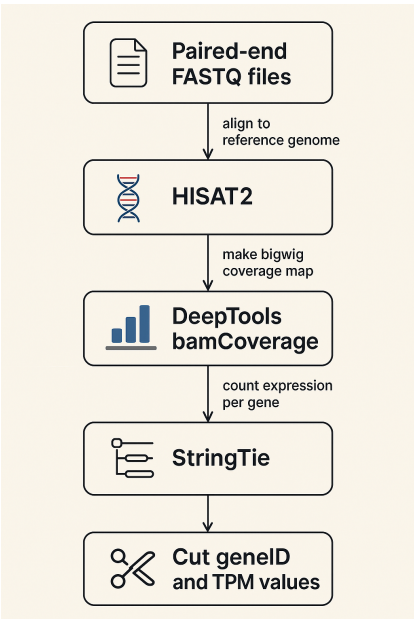
A Galaxy workflow resembles a recipe or a sequential method for data analysis. This link directs you to a publicly shared workflow that utilizes a VEuPathDB reference genome to analyze RNA-Seq data and yield results compatible with FungiDB tools: https://usegalaxy.org/u/stuart/w/rnaseq-with-hisat2-stringtie--tpm.

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.

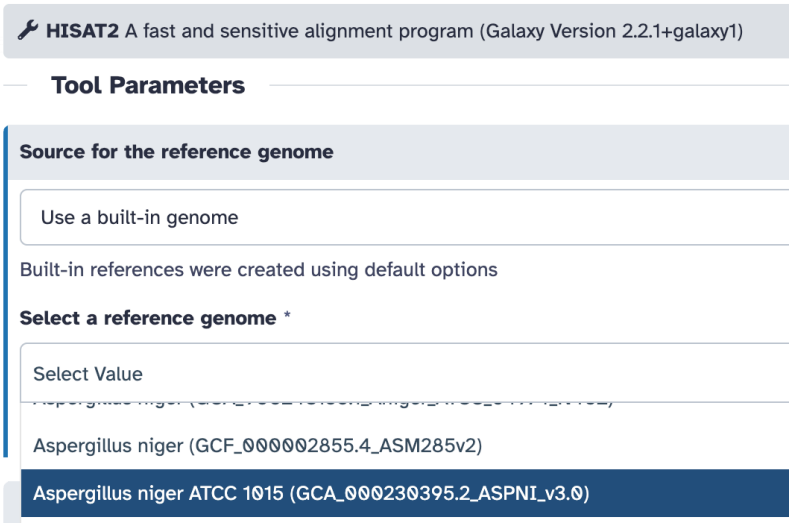
bamCoverage is part of a toolkit called DeepTools. It turns the alignment data (from HISAT2) into a BigWig file, which is a visual summary showing where in the genome there is high or low RNA expression.This is useful for creating graphs and seeing patterns in gene activity.

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.

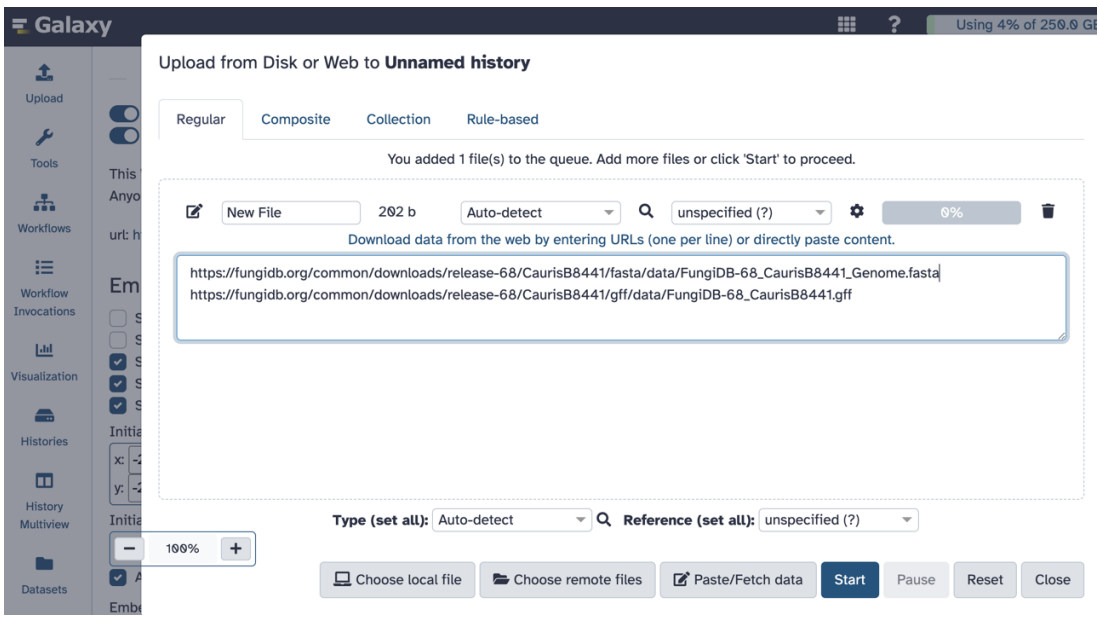
TPM is a way of standardizing gene expression data so you can compare across genes and samples more easily.

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.

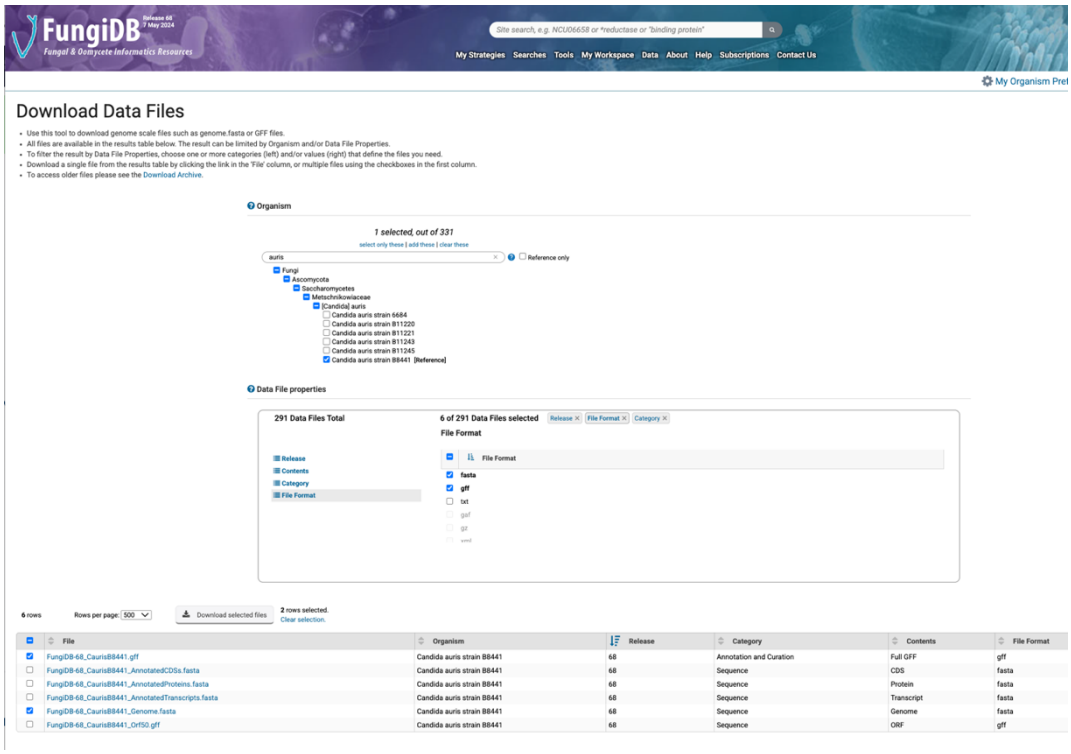


2. Providing a direct URL pointing to the fasta and gff files in VEuPathDB (the easiest way to make sure you are using the latest FungiDB). To use the URL, select the Paste/Fetch data from the Tools section in Galaxy.

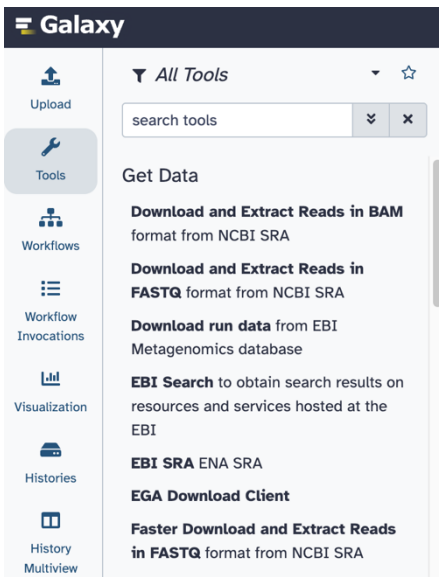


3. Manually uploading both files.

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.



Getting raw sequence data into Galaxy.

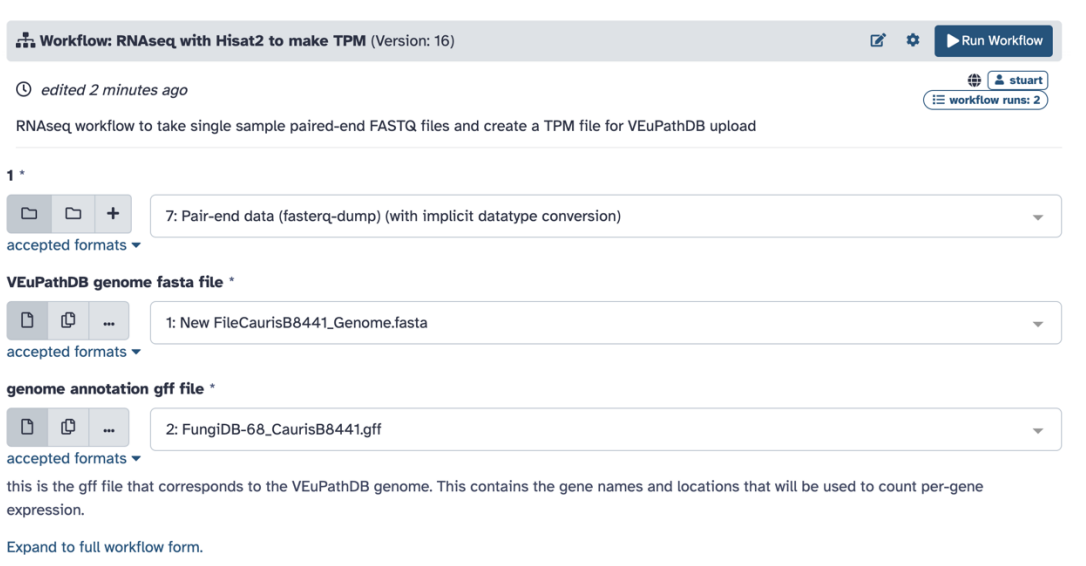


Above are just three methods to input your data into Galaxy, though many more options are available.

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.



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.

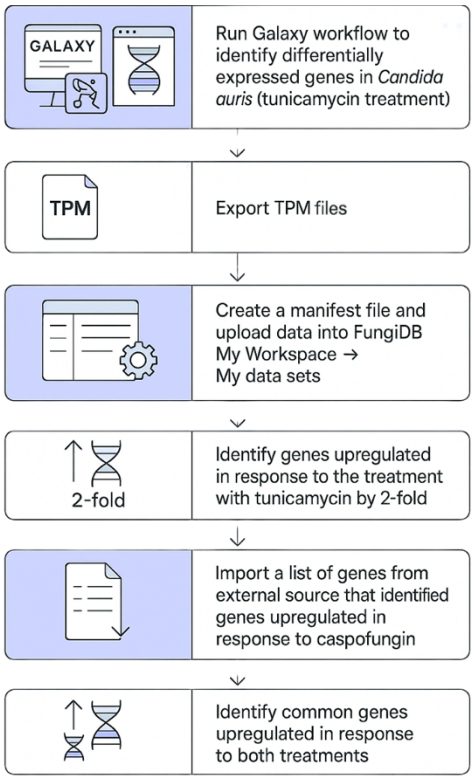
We will work with two data types for *Candida auris*:

- **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.)

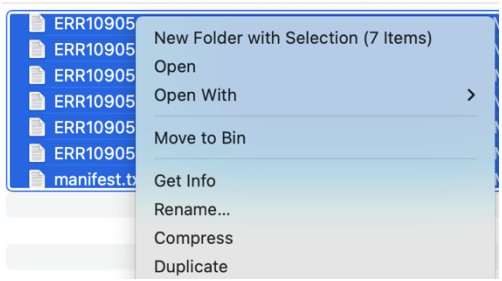
manifest.txt

```
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
```

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”.

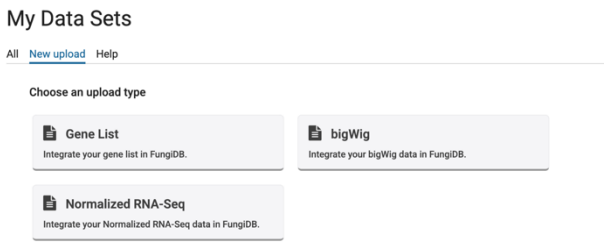


Once you have compressed all files, you are ready to upload the data for the “Normalized RNA-Seq” data tool.

- **Upload your dataset to FungiDB.**
- Navigate to the My Workspace > My data sets https://fungidb.org/fungidb/app/workspace/datasets/new/rnaseq

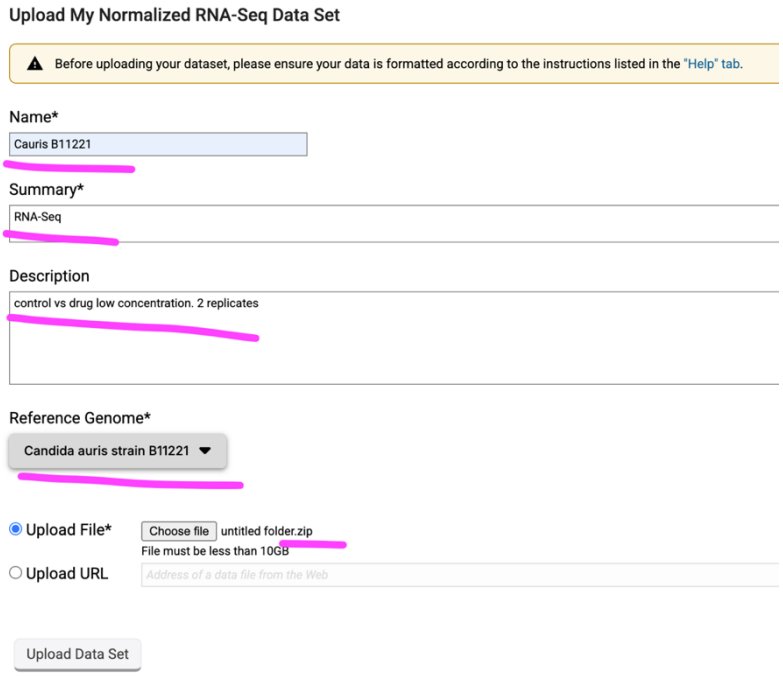
- Click on the “New upload” tab,

- 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.



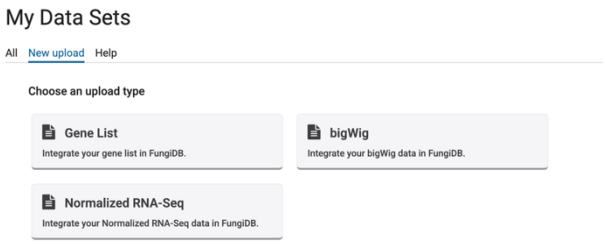
Additional instructions are here: https://fungidb.org/fungidb/app/workspace/datasets/help.

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,
- Click on the "Gene List" button,

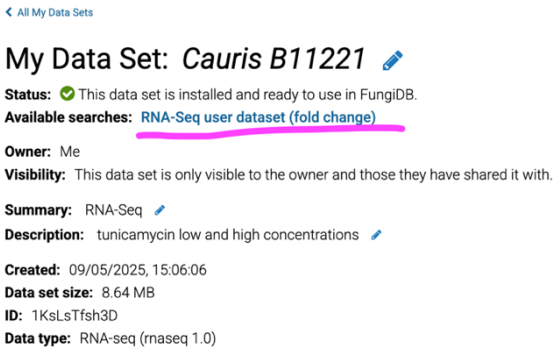


- Fill in information about the dataset
  Note: Set the reference genome to *C. auris* B8441
- Attach the "Study2_list_B8441.txt" file you downloaded from LMS.
- Click on the "Upload Data Set" button.
- Check that the dataset was installed successfully

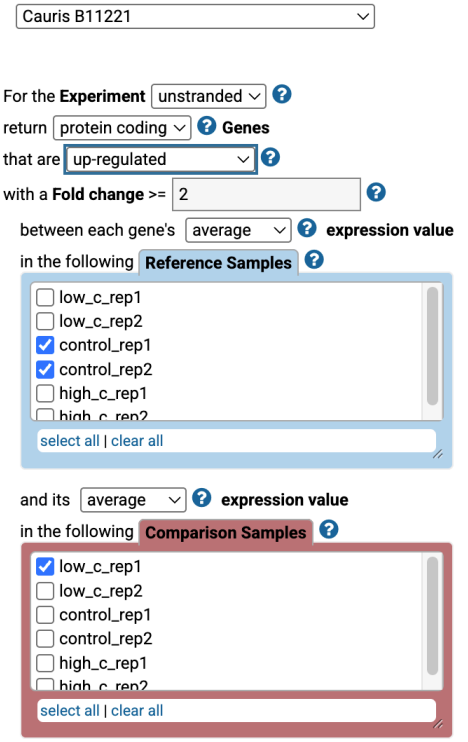
C. This following section will be completed in class during Zoom breakout sessions.

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.



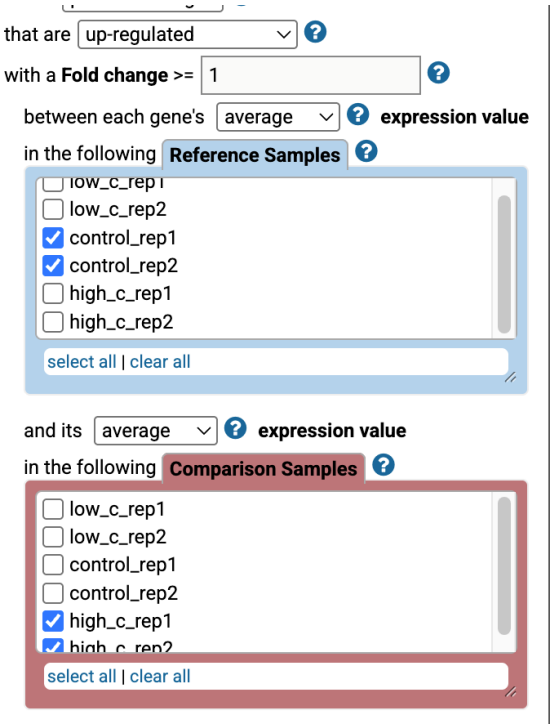
d. Find genes upregulated **1 or more fold** during the treatment with tunamycin (low concentration) when compared to control (reference sample). This search should be familiar to you as you have used to explore the integrated dataset in FungiDB before.



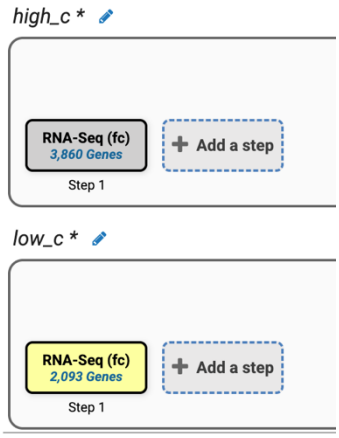
e. Name you strategy as “low_c”

f. Find genes that are **uniquely upregulated** in cells treated with high doses of tunicamycin

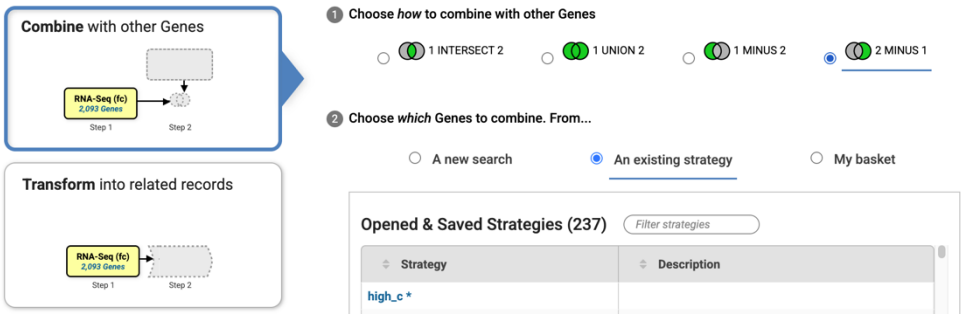
   i. Duplicate your original strategy first and modify it to use the high_c data,



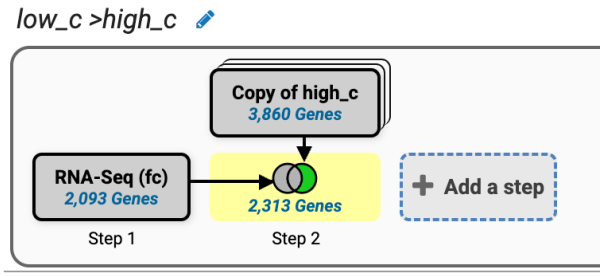
ii. Name your new strategy as "high_c",



iii. Return to the "low_c" strategy, click on the "Add Step" and select the "high_c" strategy and appropriate Boolean operator.

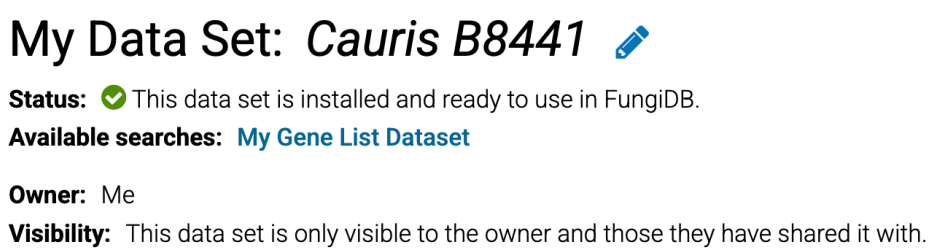


g. How many genes did the search return?

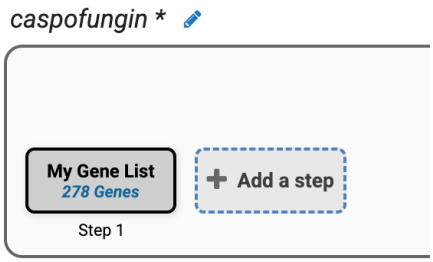


Next, incorporate Study 2 gene list (caspofungin) into the current search strategy to look for upregulated genes in response to both antifungal drugs.

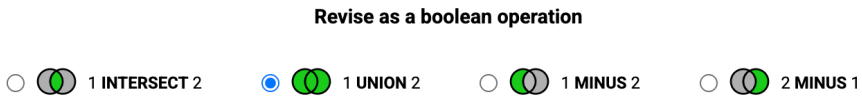
- **Deploy the Gene List search.**
  - Navigate to the My Data Sets
  - Click on the gene list dataset from Study 2
  - Deploy the Gene List Dataset search



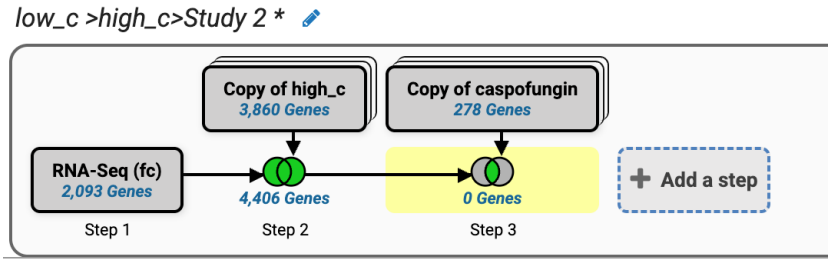
  - Save the search strategy (e.g. caspofungin).



  - 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)



  - Click on the "Add a step" and cross-reference your results with caspofungin gene list.



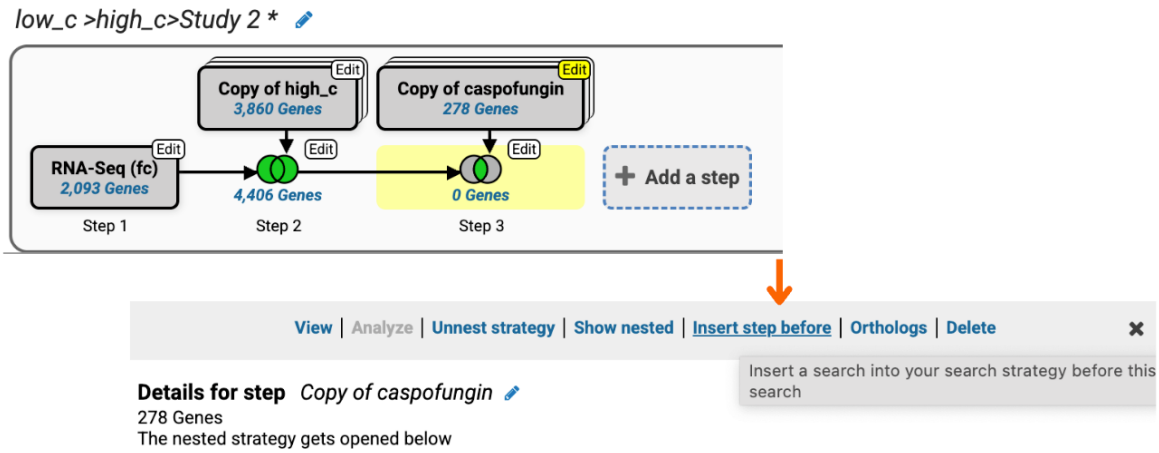
Do you think the results make sense?

Hint: Steps 1 & 2 are using gene IDs for the Ca B11221 genome, while Step 3 uses IDs for the Ca B8441.

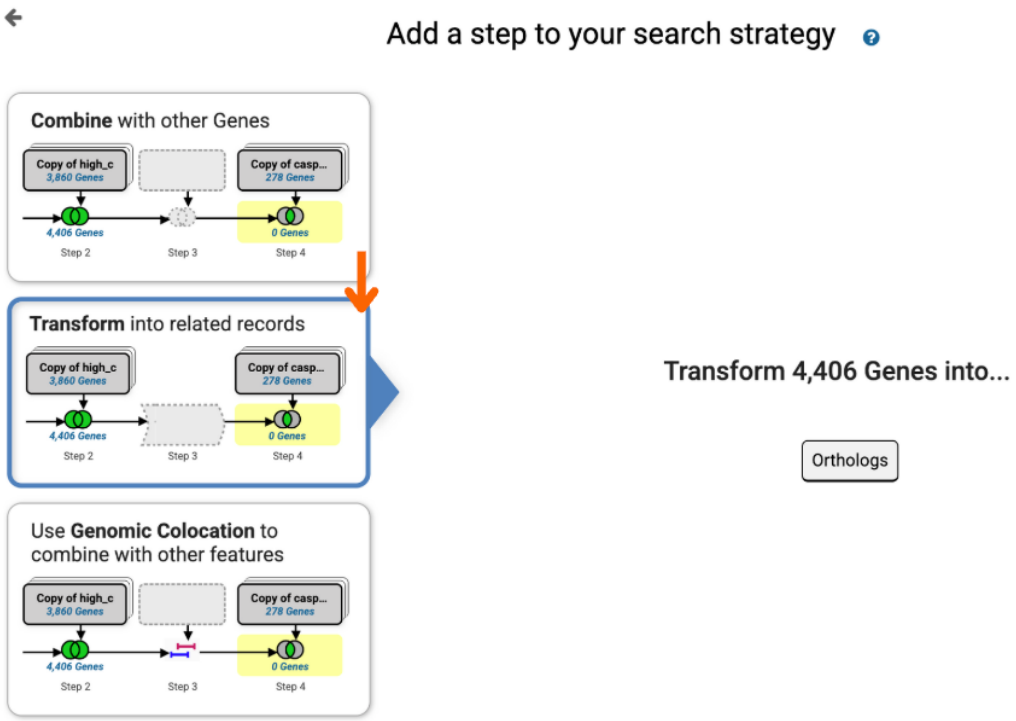
How can you ensure that all steps are using the same reference genome?

Hint: convert the results in Step 2 into B8441 Gene IDs to cross-reference both studies properly.

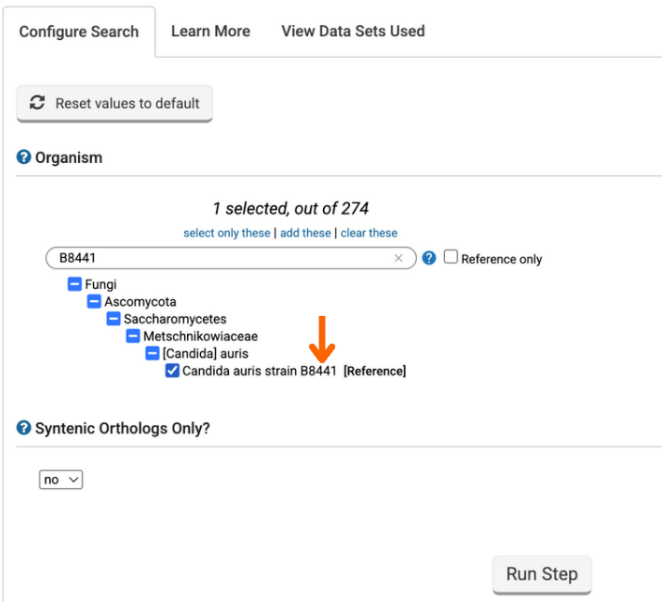
- Hover over the last step and click on the “Edit” option and click on the “Insert a step before”.

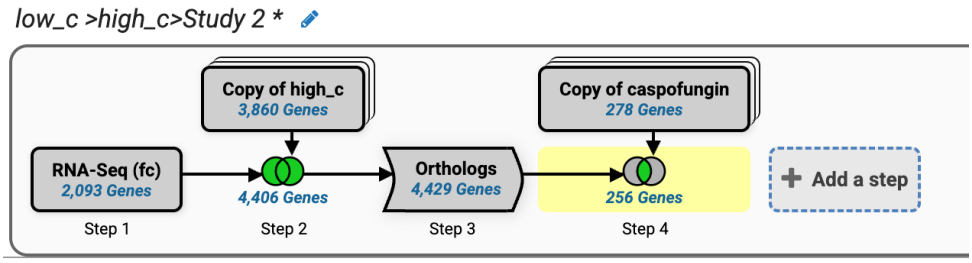


- Select the “Transform into related records” > “Orthologs” option.



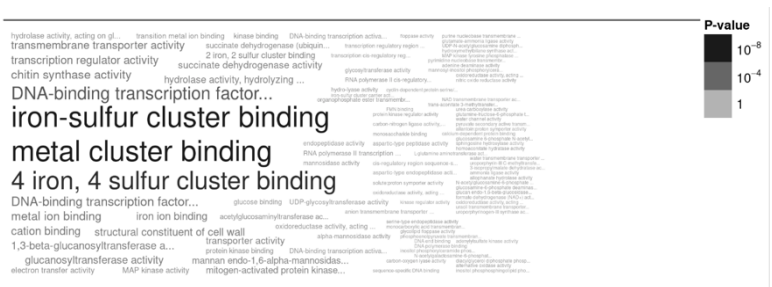
- Choose the reference genome – B8441 and click the “Run Step” button.





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