Fungal Pathogen Genomics 9 – 13 May 2023 course timetable

Tuesday 9th May

Time (BST)	Content	Instructors/Speaker
10:00 – 10:30	Registration	
10:30 – 12:30	**Welcome, and Instructor/Database introductions** Wellcome Genome Science Ensembl Fungi SGD/CGD MycoCosm/JGI FungiDB	 Martin Aslett Aleena Mushtaq + Manuel Carbajo Jodi Lew-Smith Sara Calhoun D Roos & E Basenko
12:30 –	**1. Introduction to database queries** FungiDB site search & advanced search strategies - p.6 SGD YeastMine - p.16 Ensembl Fungi – BioMart - p.24 Ensembl Fungi Molecular Interactions - p.35	All instructors
13:00 – 14:00	Lunch	
– 15:30	Introduction to database queries, Cont/…	All instructors
15:30 –	**2. Transcriptomics & Proteomics** Ensembl Fungi Track Hubs – p. 46 SGD Expression tools - SPELL – p. 53 FungiDB Transcriptomic & Proteomic analysis– p. 57 Assessing (& editing) gene annotation (JBrowse/Apollo) – p.68	All instructors D Roos
16:00 – 16:30	Tea break	
– 18:30	Transcriptomics & Proteomics, cont	All instructors
18:30 – 19:00	**3. SNPs & Variants** SGD Variant viewer – p. 76 FungiDB SNP analysis & CNVs – p. 80 Exploring variant in Ensembl Fungi – p. 98	All instructors
19:00– 20:00	**Welcome reception & Dinner**	
20:00 – 21:00	Participant presentations (2 min flash talks)	All instructors

Wednesday 10th May 2023

Time (BST)	Content	Instructors/Speaker
– 09:00	Breakfast	
09:00 – 10:30	SNPs & Variants (Cont.)	All instructors
10:30 – 11:00	Tea break	
11:00 – 13:00	**4. Comparative Genomics & Orthology and Evolutionary analysis & cross-species inference** Ensembl Fungi – WGA – p. 111 MycoCosm CAZy enzymes– p. 118 MycoCosm Synteny– p. 127 Exploring protein domains and clusters across Ensembl & MycoCosm– p. 132 SGD predicting fungal biology– p. 141 Ensembl Fungi Evolutionary analysis (gene trees) – p. 149 FungiDB JBrowse synteny – p. 163 FungiDB &OrthoMCL: Orthology and Phyletic Patterns – p. 174	15 min intro: D Roos
13:00 – 14:00	Lunch	
14:00 –	Comparative Genomics & Orthology, Cont/…	All instructors
16:00– 16:30	Tea break	
– 17:00	Comparative Genomics & Orthology, Cont/…	All instructors
17:00 – 19:00	**5. NGS data analysis I** (note: RNA-Seq and Variant calling sessions are run in parallel) Background & Intro to VEuPathDB Galaxy Deploying workflows using pre-loaded data 5a. RNA-Seq – p. 186 5b. SNP Calling– p. 200	**Intro lecture** - Crouch RNA-Seq SNP calling Galaxy platform & user workspaces 5a: Crouch / Basenko 5b: Roos / Brown
19:00– 20:00	**Dinner**	
20:00 – 21:00	**Research Seminar (Michael Bromley, MRC)**	

Thursday 11th May 2023

Time (BST)	Content	Instructors/Speaker
– 09:00	Breakfast	
09:00 –	**6. Enrichment analysis** SGD GO Slim mapper – p. 207 CGD GO Term finder – p. 210 FungiDB GO enrichment – p. 214	10 min intro: S Brown
10:30 – 11:00	Tea break	
– 12:30	Enrichment analysis, Cont/…	All instructors
12:30 –	**7. NGS Data analysis II** (note: RNA-Seq and Variant calling sessions are run in parallel) 7a. RNA-Seq – p. 224 7b. SNP Calling – p. 236	Lecture: K. Crouch 7a. Crouch / Basenko 7b: Roos / Brown
13:00 – 14:00	Lunch	
14:00 – 15:00	**Sanger Tour**	Nishadi De Silva
15:00 –	**NGS Data analysis II**, Cont/…	All instructors 7a. Crouch / Basenko 7b: Roos / Brown
16:00– 16:30	Tea break	
– 18:30	**NGS Data analysis II**, Cont/… (note: RNA-Seq and Variant calling sessions are run in parallel)	All instructors 7a. Crouch / Basenko 7b: Roos / Brown
18:30 – 19:00	**Introduction to group projects**	Nishadi De Silva
19:00– 20:00	**Dinner**	
20:00 – 21:00	**Research Seminar (Ester Gaya, KEW)**	

Friday 12th May 2023

Time (BST)	Content	Instructors/Speaker
– 09:00	Breakfast	
09:00 –	**8. Functional analysis: Pathways & metabolites** MycoCosm KEGG Browser & Secondary metabolism clusters – p. 252 FungiDB pathways & metabolites – p. 263	All instructors
10:30 – 11:00	Tea break	
– 12:00	Functional analysis: Pathways & metabolites, Cont/…	All instructors
12:00 –	**Group Projects**	Nishadi De Silva & All instructors
13:00 – 14:00	Lunch	
14:00 –	Group Projects, Cont/…	
16:00– 16:30	Tea break	
– 19:00	Group Projects, Cont/…	All instructors
19:00– 20:00	**Dinner**	

Saturday 13th May 2023

Time (BST)	Content	Instructors/Speaker
– 09:00	Breakfast	
09:00 –	Group Projects, Cont/…	All instructors
10:30 – 11:00	Tea break	
11:00 – 13:00	**Group Projects Presentations (12 min/team max)**	All instructors
13:00 – 14:00	Lunch & Departure	

Site Search

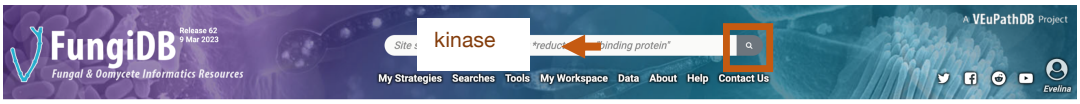
Learning objectives:

- Use keywords in site search.
- Explore site search results.
- Filter site search results by categories.
- Filter site search results by organisms.
- Filter site search results by category fields.
- Export results to a search strategy.
- Find a specific gene using its ID in site search.

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

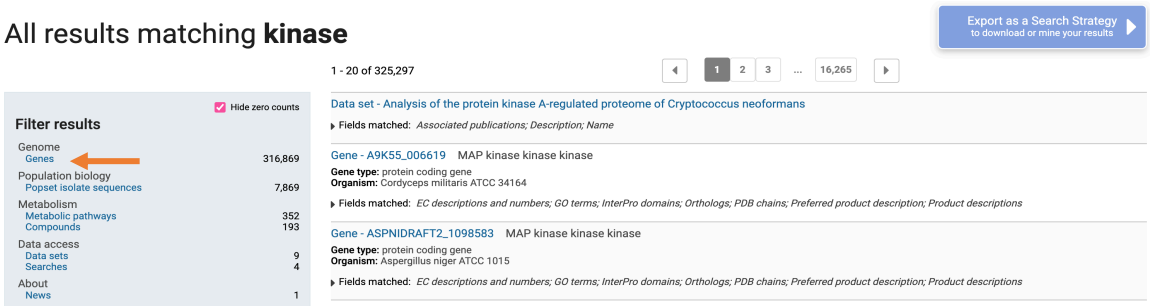
Site search: text, term or gene id.

1. 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.



2. How many results with the word kinase did you get? Are all of these records genes?

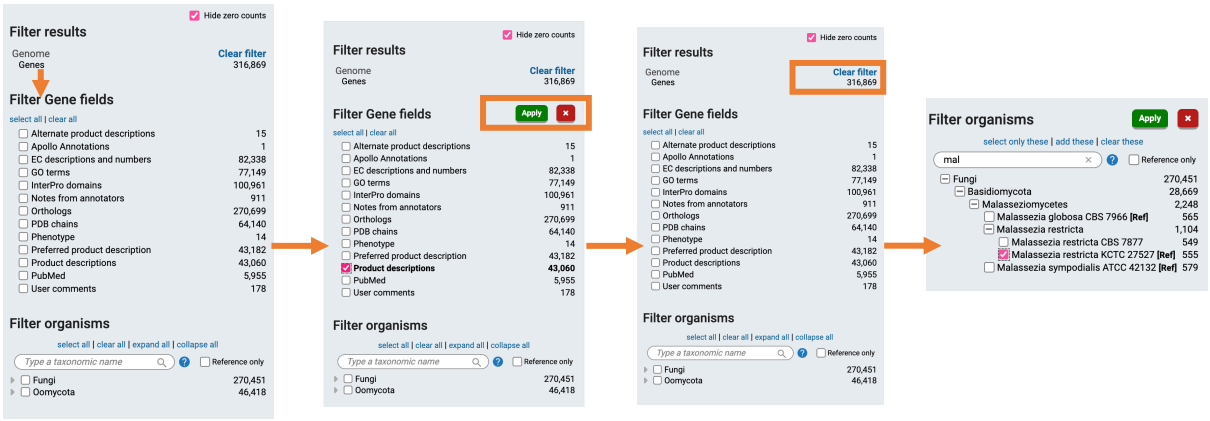
3. Explore the filter panel on the left side of the page. Filter the results to view gene results only (hint: click on the word *Genes* in the *Filter results* section):



Notice that clicking on the “Genes” category reveals additional filtering options.

4. Select and apply the *Product descriptions* filter.

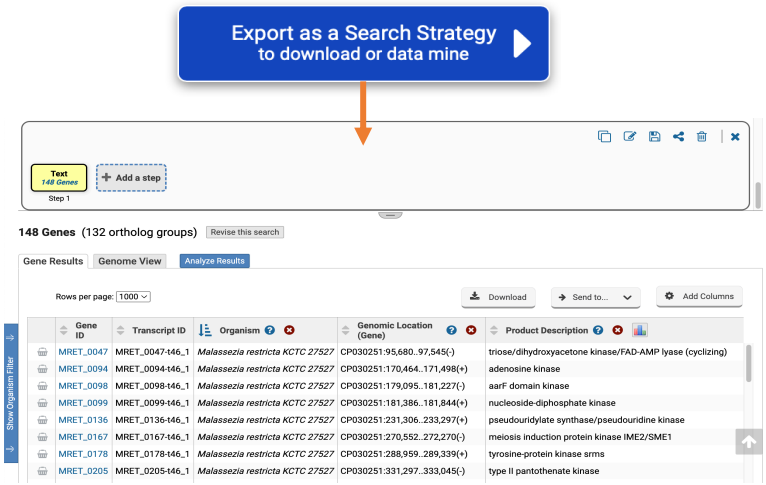
Note: The applied filter can be easily cleared by clicking on “Clear filter” option.



5. 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?

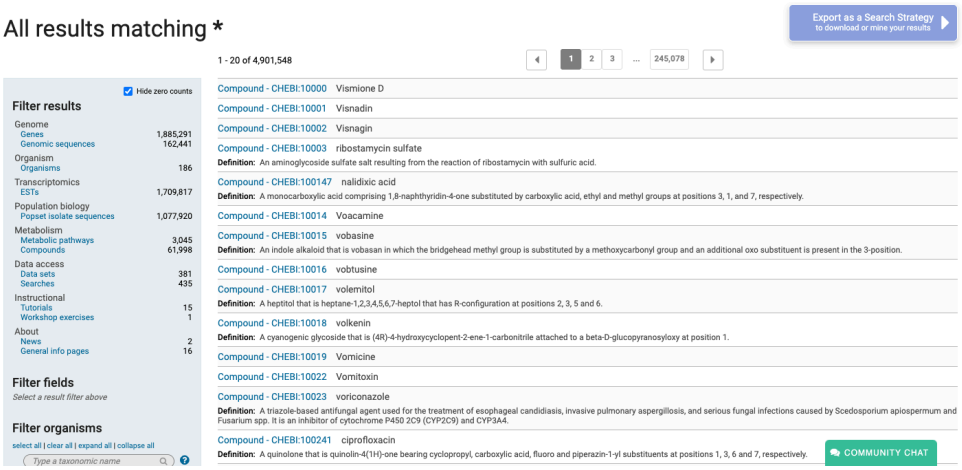
6. Export the results to a search strategy.

To achieve this, click on the blue button called “Export as a search strategy…” at the top right-hand side of the results page. Notice that before the Genes category was selected this button was inactive. This is because the search strategy can be deployed on a signle category only (e.g. Genes or Data sets, but not both).



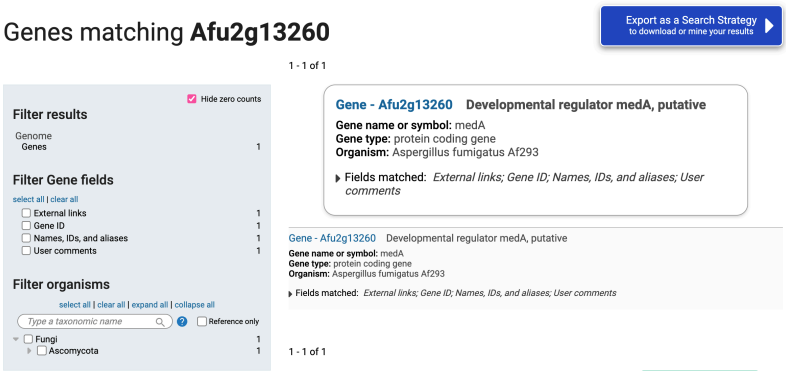
7. Try running the same search but this time use a wild card (*) (e.g., kinase*).

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



8. 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** (see screenshot below). Click on the gene link in blue to navigate to the gene record page for this gene.



Note: a single gene id can be also exported as a search strategy. This may be useful if you are interested in cross-referencing different types of data for one gene.

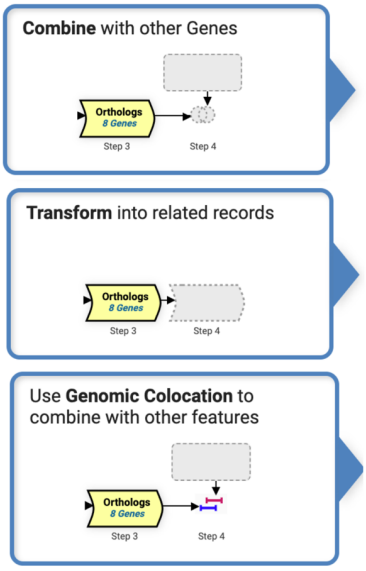
Advanced Search Strategies

Learning objectives:

- Use sites search and other types of searches to create a multi-step query across different types of records and genomes.

The strategy system offers over 100 structured searches that can be combined to produce multi-step strategies. Searches can be deployed from the site search, or 'Search For…' menu on the home page and from the 'Searches' dropdown menu in the header of every page. Searches listed under Genes will return a list of gene IDs, while searches listed under 'SNPs' or 'Metabolic Pathways' will return record IDs representing SNPs, or metabolic pathways, respectively, etc.

The searches can be combined via three major approaches:

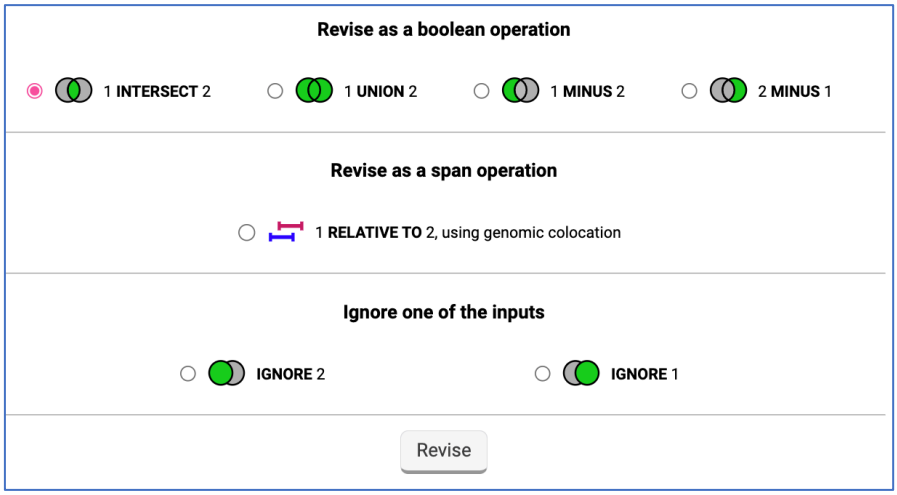


1. Combine with other Genes (compares results that are gene lists)

2. Transform into related records (transform results into orthologs, metabolic pathways or compounds)

3. Use Genomics Colocation to combine with other features (cross-references different types of data – e.g., gene lists with metabolic pathways)

Strategy steps are connected via the 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 can be masked off from the strategy with the help of "ignore step" Boolean operators.

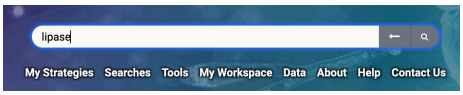


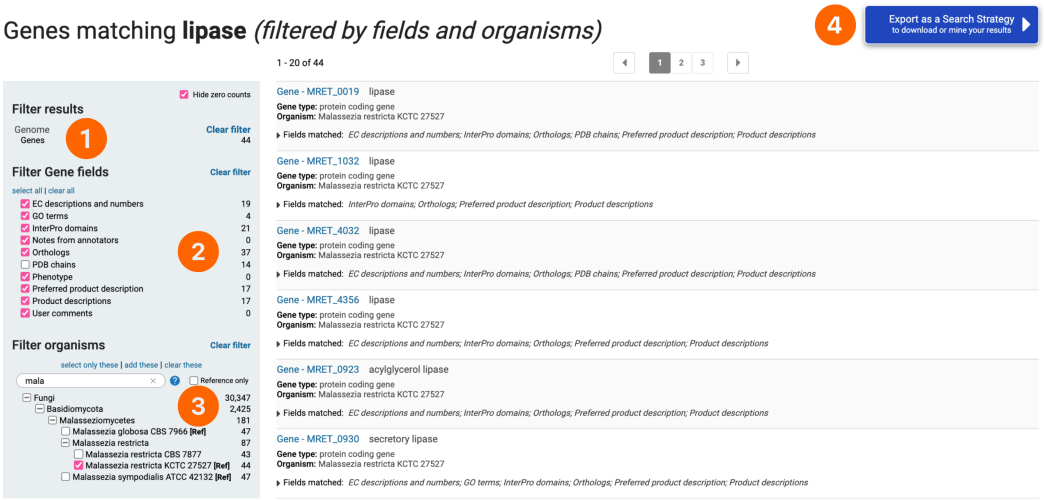
M. restricta can cause skin disorders and is one of the most common fungal species found on human skin. *Malassezia* cannot produce fatty acids and relies on fatty acid uptake from external sources. Secreted lipases are thought to contribute to *Malassezia* pathogenicity. In this strategy we will identify secreted lipases in *M. restricta* KCTC 27527, cross-reference annotation with InterPro domain annotations and find orthologs of *M. restricta* genes in another *Malassezia* strain and also *Candida albicans* (REF).

To build this strategy, use the following approach:

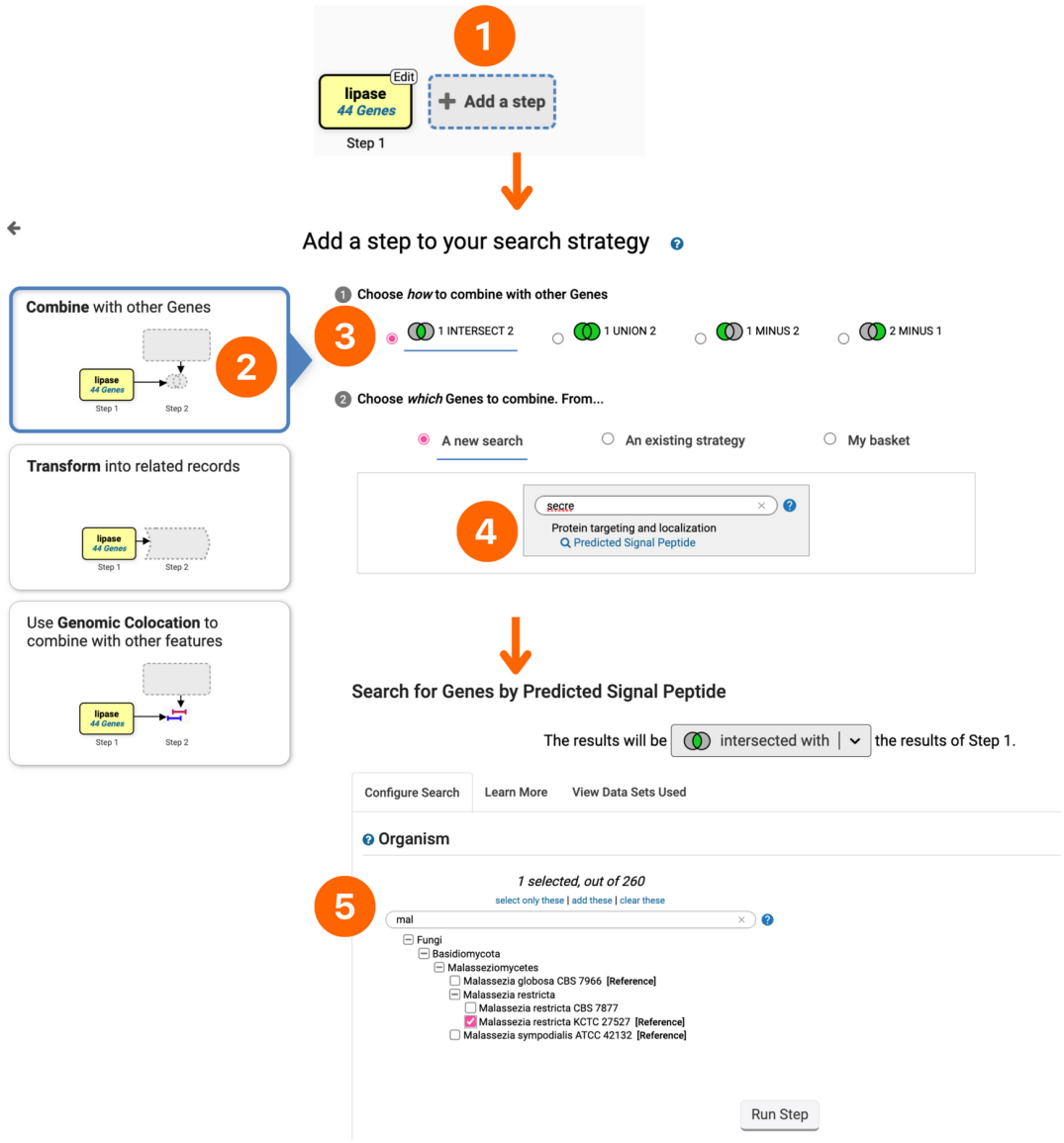
- **Use site search** to identify genes that have "lipase" annotation in *Malassezia restricta* KCTC 27527. This search identifies genes that have "lipase" annotation is several evidence fields.
- **Identify Genes by Signal peptide prediction**. This search returns genes predicted to have signal peptide.
- **Identify Genes based on InterPro domain**. This search identifies genes with specific domain signature – secreted lipase (LIP).
- **Transform by Orthology into another organism**. FungiDB integrates ortholog profiles from OrthoMCL. The OrthoMCL algorithm clusters proteins into ortholog groups based on BLAST similarity across at 150 genomes that span the tree of life. The transform we perform here will convert a list of genes in one organism to their orthologs in a different organism. In this case, we will transform a list of *M. restricta* KCTC 27527genes into their orthologs in *Malassezia restricta* CBS 7877 and *Candida albicans* SC5314.

- **Use site search** to identify genes that have “lipase” annotation in *Malassezia restricta* KCTC 27527
  1. Run site search for genes annotated with “lipase” and filter on Genes.
  2. Use Gene fields to filter your results as shown.
  3. Restrict your search to *M. restricta* KCTC 27527 genes.
  4. Export results as a search strategy.

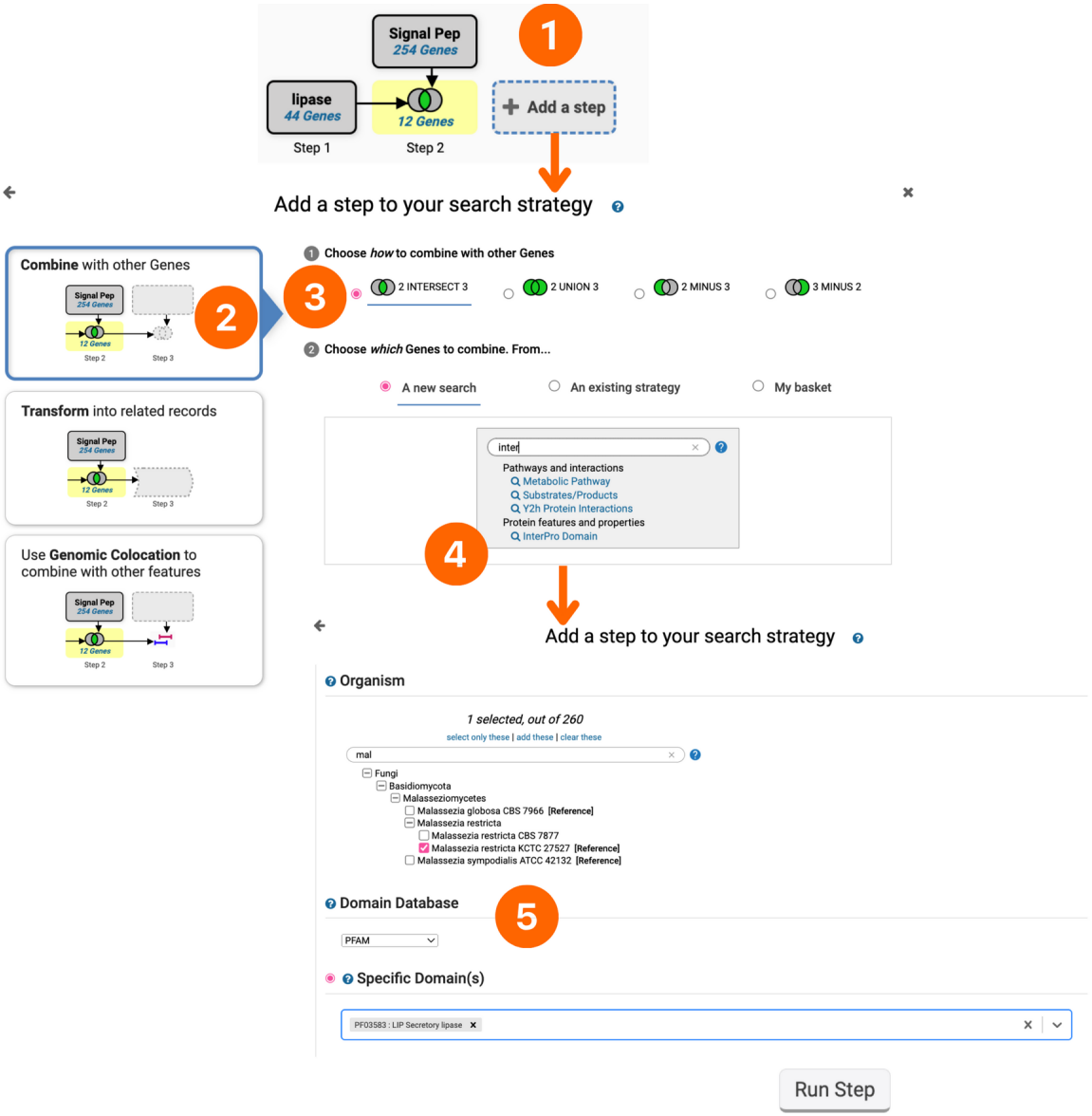




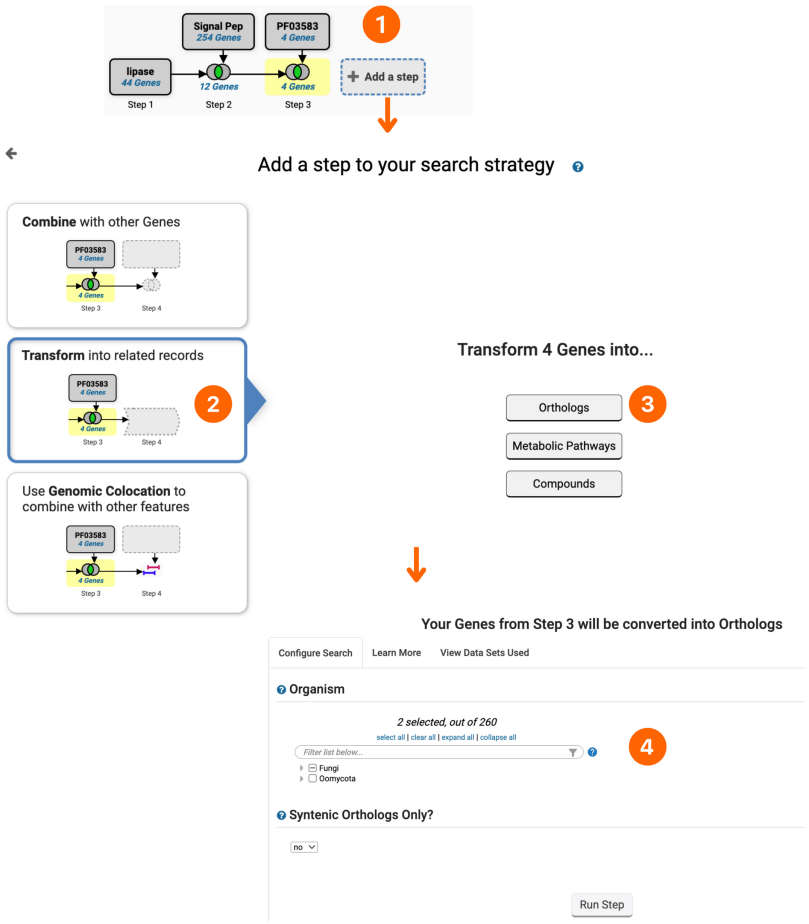
- **Identify Genes by Signal peptide prediction**. This step will identify lipases that may be secreted.
    1. Click on the “Add step” button.
    2. Choose “Combine with other genes” search.
    3. Choose to “intersect” your results with the previous step.
    4. Filter the available searches to deploy the “Predicated Signal Peptide” search.
    5. Restrict the search to *M. restricta* KCTC 27527 and click on the “Run Step” button.



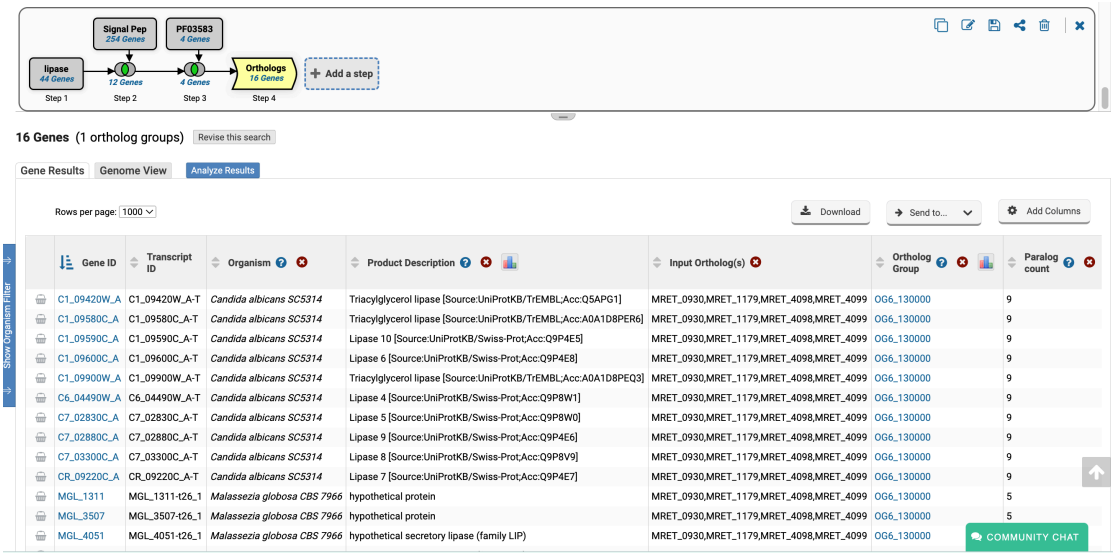
- **Identify Genes based on InterPro domain**. This search identifies genes with specific domain signature – secreted lipase (LIP).
    1. Click on the “Add step” button.
    2. Choose “Combine with other genes” search.
    3. Choose to “intersect” your results with the previous step.
    4. Filter the available searches to deploy the “InterPro domain” search.
    5. Restrict to *M. restricta* KCTC 27527, select “Secretory lipase” domain (PF03583 : LIP Secretory lipase), and click on the “Run Step” button.



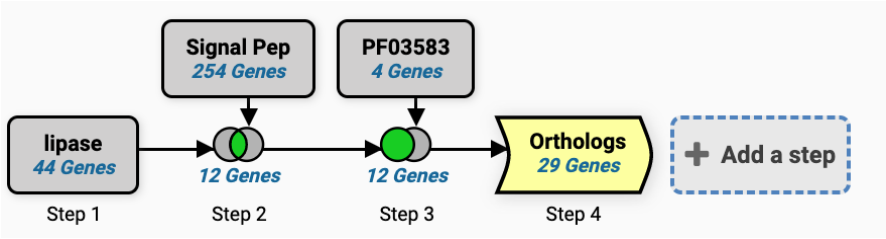
- **Transform by Orthology into another organism/s**. This search is particularly useful if you are working with a poorly annotated genome and want to take advantage of annotations from another, better annotated, genome. In this exercise, we will practice finding orthologs in *Malassezia globosa* CBS 7966 and *Candida albicans* SC5314.
    1. Click on the "Add step" button.
    2. Choose "Transform into related records" search.
    3. Choose to deploy the "Orthologs" search.
    4. Restrict the orthologs search to *M. globosa* and *Candida albicans* SC5314 and click on the "Run Step" button.



Examine your results. Do they make sense?



How can you lower the stringency of the search by removing the third step from the search without deleting it? (Hint: you will need to use a certain Boolean operator).



Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/d3a431b32ee7b32f

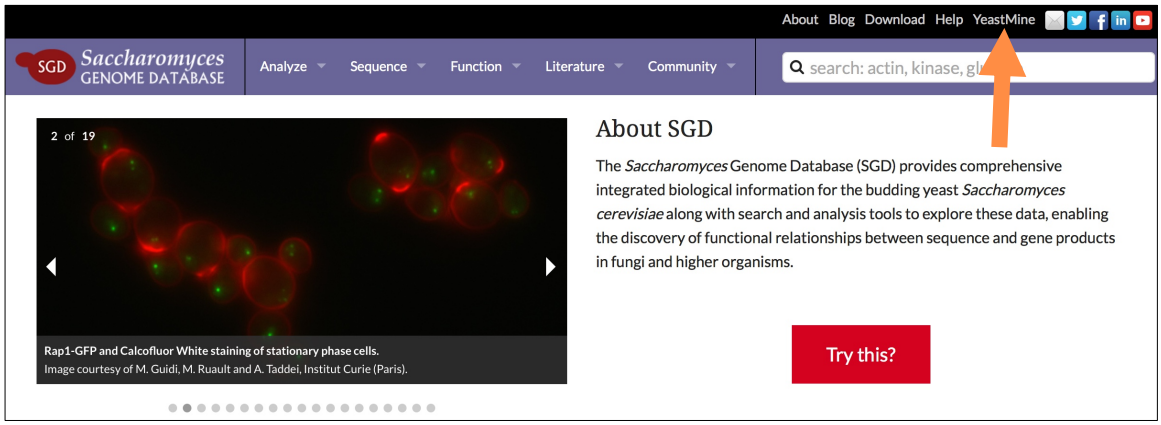
References:
Park et al. J. Microbiol. Biotechnol. 2021; 31(5): 637-644 doi:10.4014/jmb.2012.12048

Search Strategies in SGD

In addition to a faceted search tool, SGD provides **YeastMine** (https://yeastmine.yeastgenome.org/) as a means for users to conduct more advanced queries. YeastMine 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 YeastMine to search for as-yet undiscovered mitochondrial ribosomal proteins in yeast.

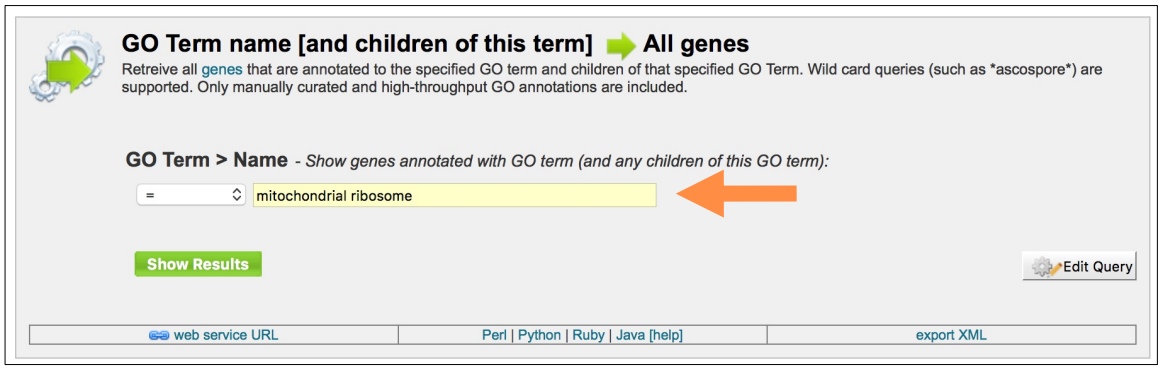
- Access YeastMine from SGD home page (http://www.yeastgenome.org); click on YeastMine in the upper right corner above the search box.



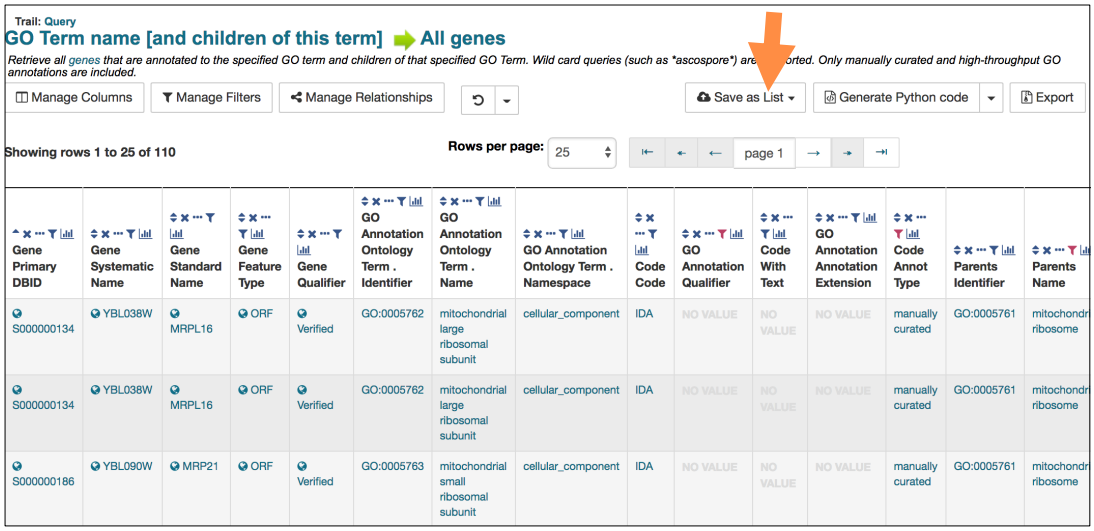
1. Create a list of proteins that are known subunits of the mitochondrial ribosome (MTR):

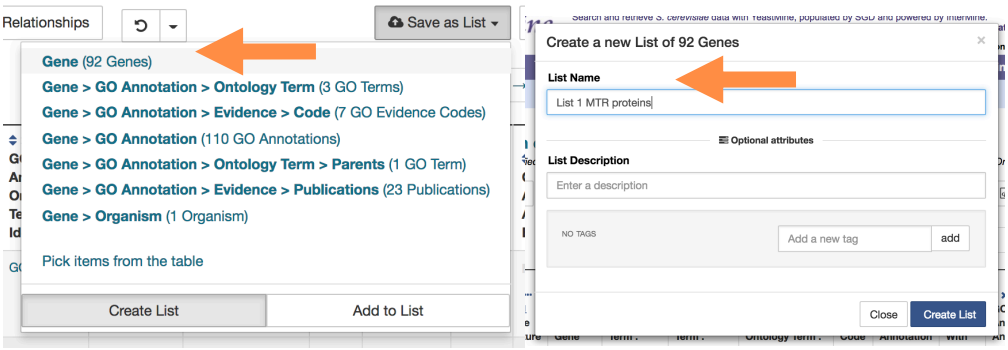
- Open **FUNCTION** tab and select **GO Term name [and children of this term] -> All genes**

- Enter **mitochondrial ribosome** into the query box; hit **Show Results**



- In the Results page, you should see a table with 110 rows. Click on **Save as List** and select the option **Gene (91 Genes).** Give your list a name, such as "**List 1 MTR proteins**" and hit **Create List** (you should see a green **Success** banner on top)

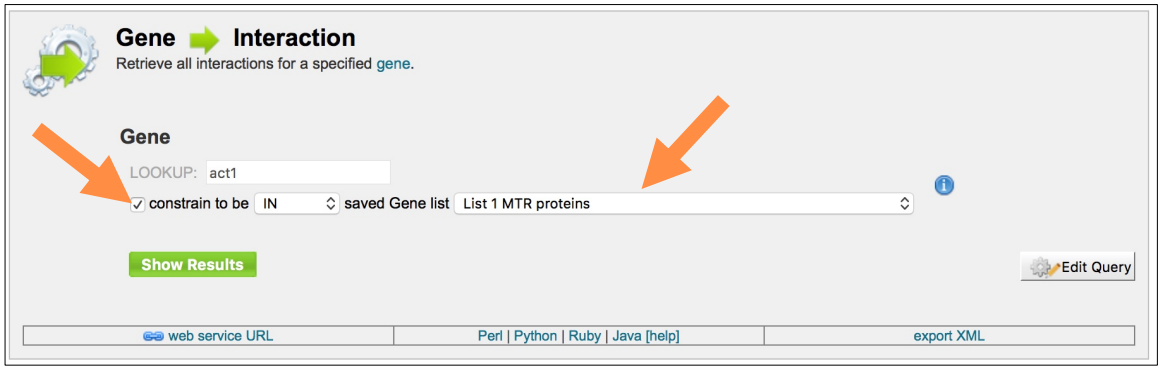




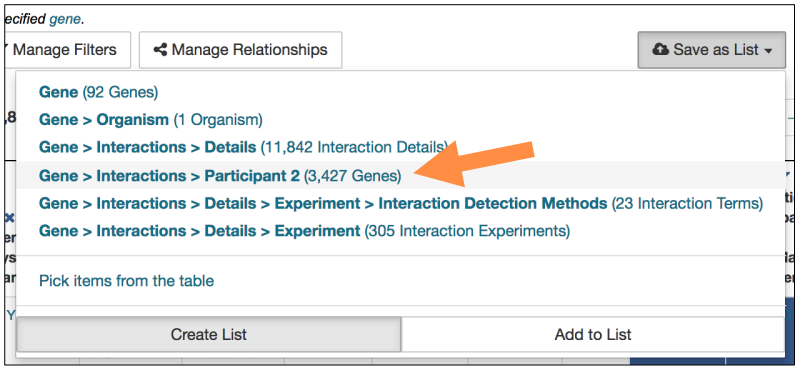
2. Find proteins that genetically interact with MTR proteins:

- Go back to YeastMine home page (click on **Home** in the purple banner on top). Open the **INTERACTIONS** tab and select **Gene -> Genetic Interactions**

- Check the box next to **constrain to be IN** and select your previously created list ("**List 1 MTR proteins**") from the menu; hit **Show Results**

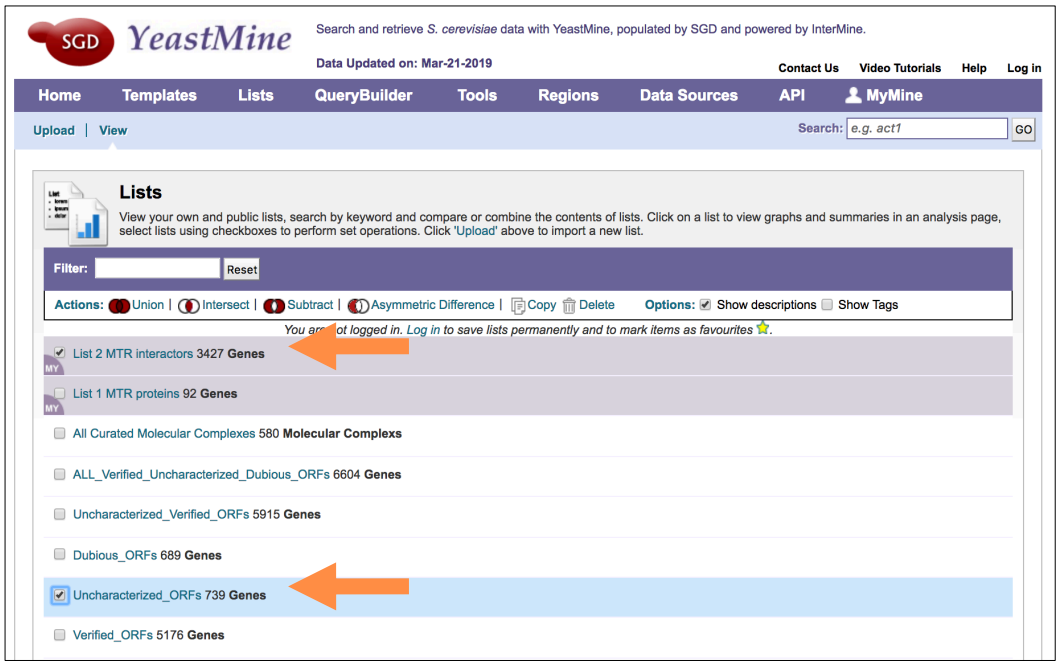


- The results page shows all genes/proteins with genetic or physical interactions with the MTR proteins from List 1. Save the MTR interactors by clicking on **Save as List** and selecting **Gene > Interactions > Participant 2**. Give your list a name ("**List 2 MTR interactors**") and hit **Create List.**

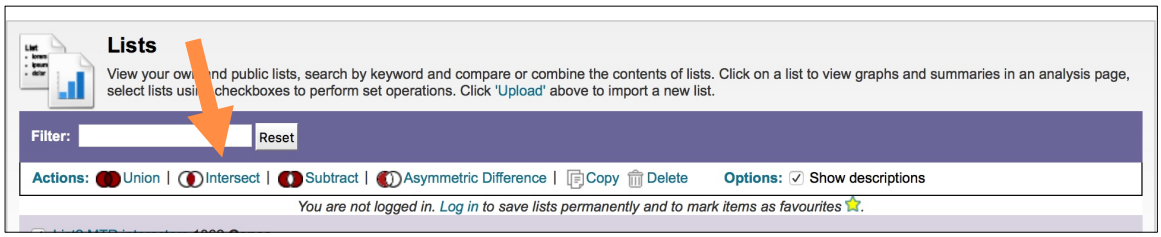


3. Find MTR interactors that are uncharacterized:

- Use a pre-made list of uncharacterized yeast genes: select **Lists** from the purple banner on top and click on **View** in the upper left corner. Scroll down the page to check the box next to **Uncharacterized_ORFs.** Also check your previously saved list ("**List 2 MTR interactors**") that should be on top, highlighted in purple.



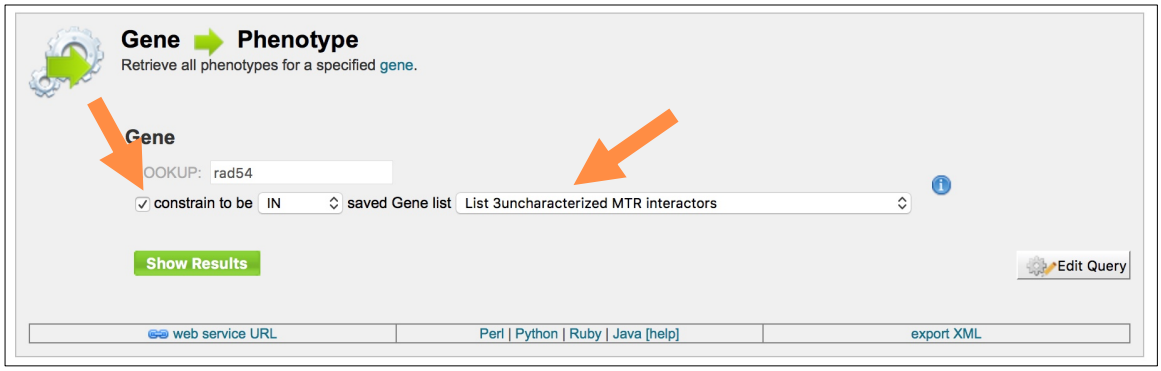
- From the **Actions**, click on **Intersect**, give your list a name ("**List 3 uncharacterized MTR interactors**") and click on **Save**; a green confirmation banner should appear on top.



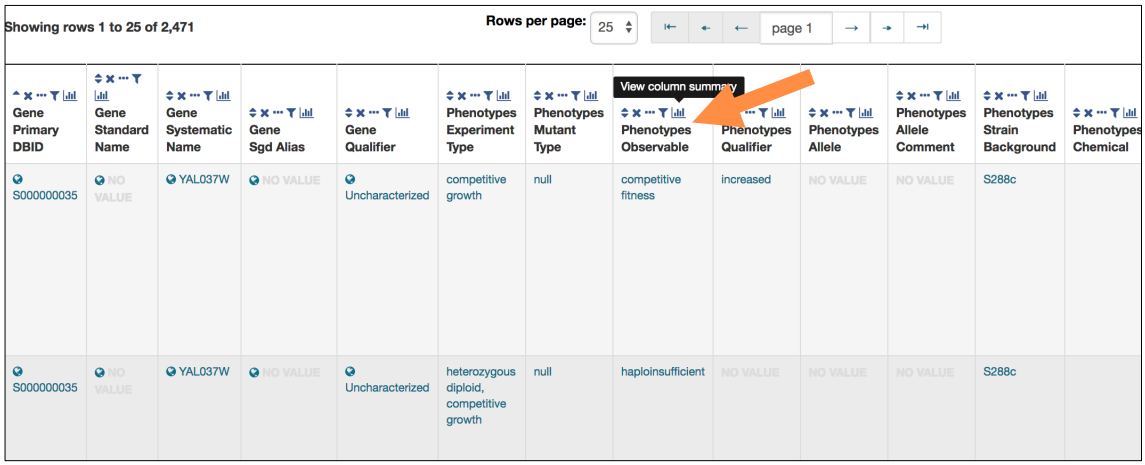
- Click on your list to see the results.
- Because we have over 200 genes in our results, it would be a good idea to narrow down our candidates even more. For example, because the MTR is a mitochondrial complex, we would expect that deleting uncharacterized (but bona fide) subunits of the MTR would disrupt aerobic respiration. Let's refine our list of predicted MTR subunits by seeing which genes disrupt respiratory growth when deleted.

- Return to YeastMine home page, open up **PHENOTYPES** tab and select the **Gene -> Phenotype** query

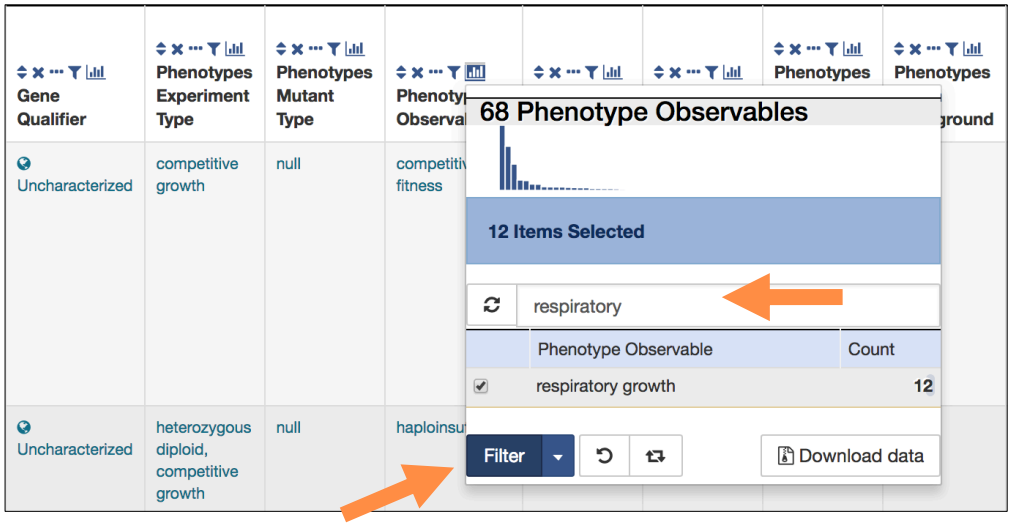
- Check the **constrain to be IN** checkbox and select your saved list ("**List 3 uncharacterized MTR interactors**"); click on **Show Results**



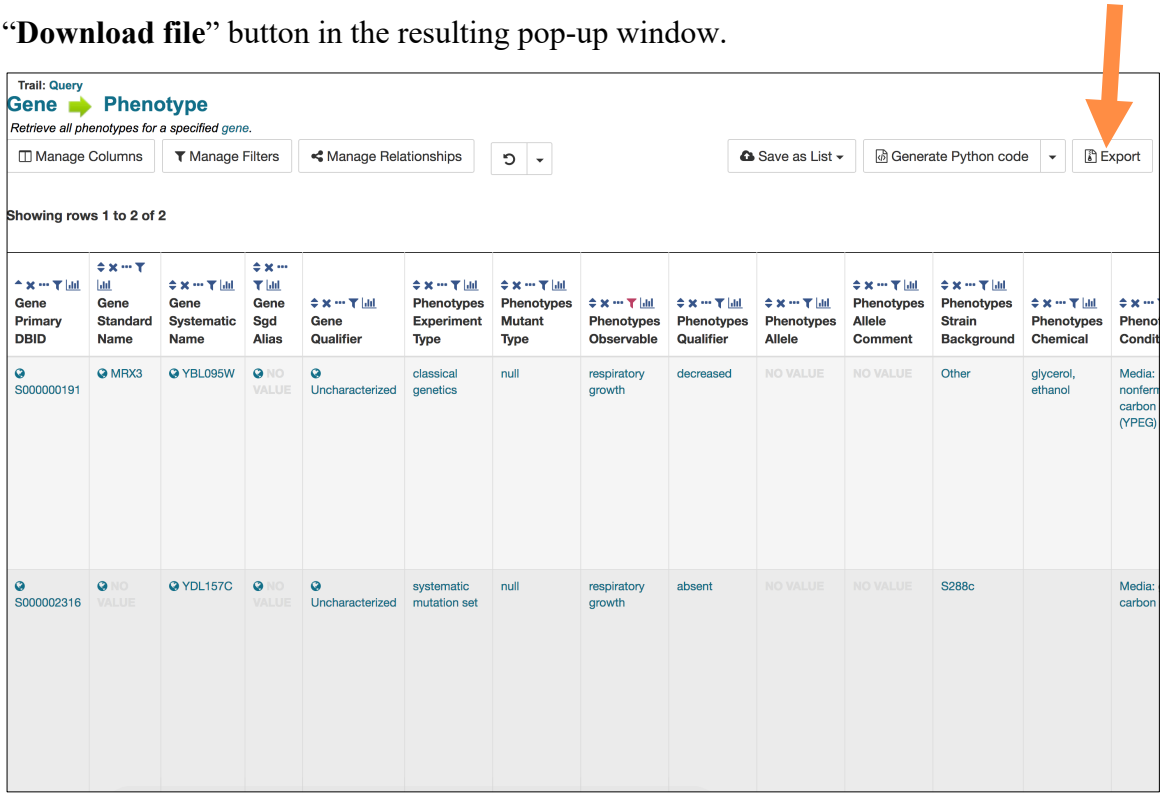
- In the Results table, find a column labeled **Phenotypes Observable**. Hover your mouse over the small icons above the column name and click on **View Column Summary** (the bar graph icon on the right).



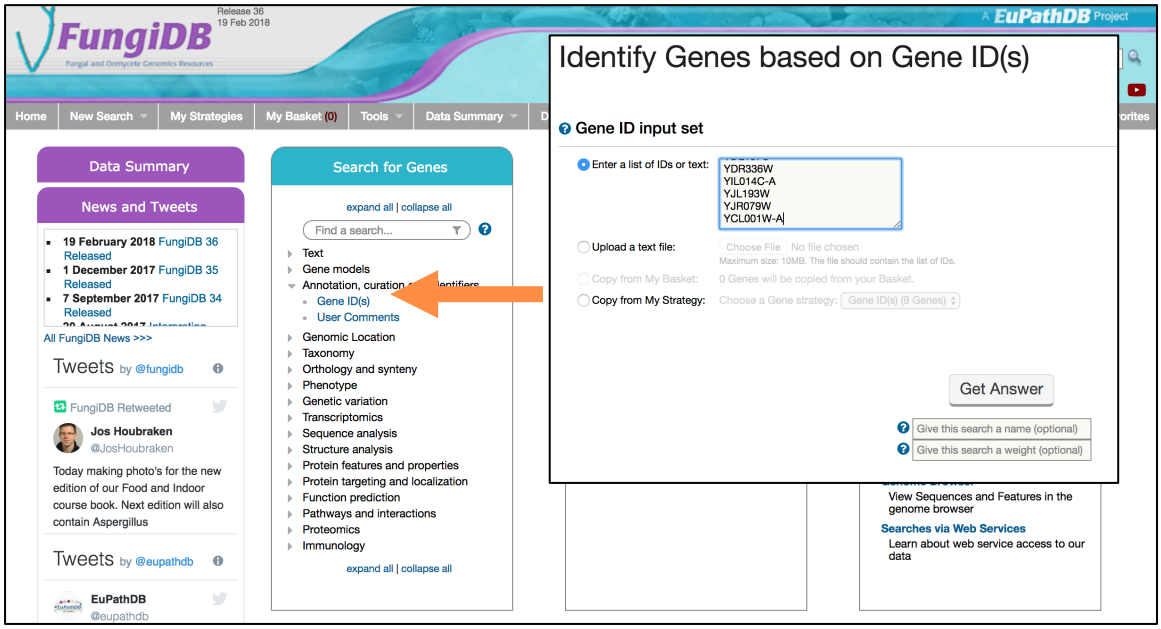
- In the **Filter values** box, enter **respiratory** and scroll down the list to check the box next to **Respiratory growth**; hit **Filter**.



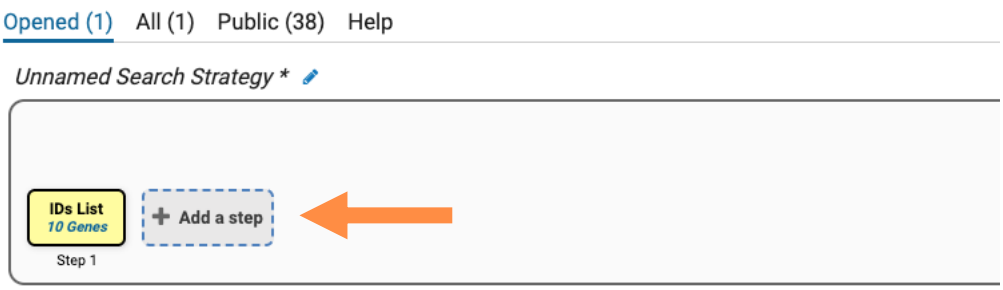
- To filter the phenotypes for those where respiratory growth is impeded, find the **Phenotype Qualifiers** column and open the **View Column Summary** menu. Select all items that refer to hindering respiratory growth: “decreased”, “decreased rate”, “absent”, etc. Then, hit Filter.
- You should now have a list of uncharacterized yeast genes whose products interact with mitochondrial ribosomes and mutations lead to respiratory growth defects. 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.



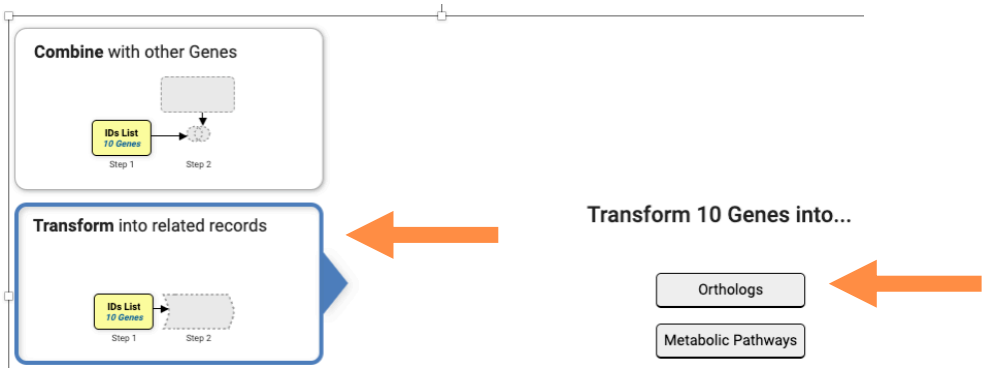
- The results of the above YeastMine analysis suggest 10 genes that potentially encode undiscovered subunits of the mitochondrial ribosome. Although these genes are uncharacterized, more data may exist on their orthologs in other organisms. Use FungiDB to survey the function of orthologs in Fungi and Oomycetes.
- Open the FungiDB homepage (http://fungidb.org/). In the "**Search for Genes**" box, open the "**Annotation, curation and identifiers**" section and click on "**Gene ID(s)**".



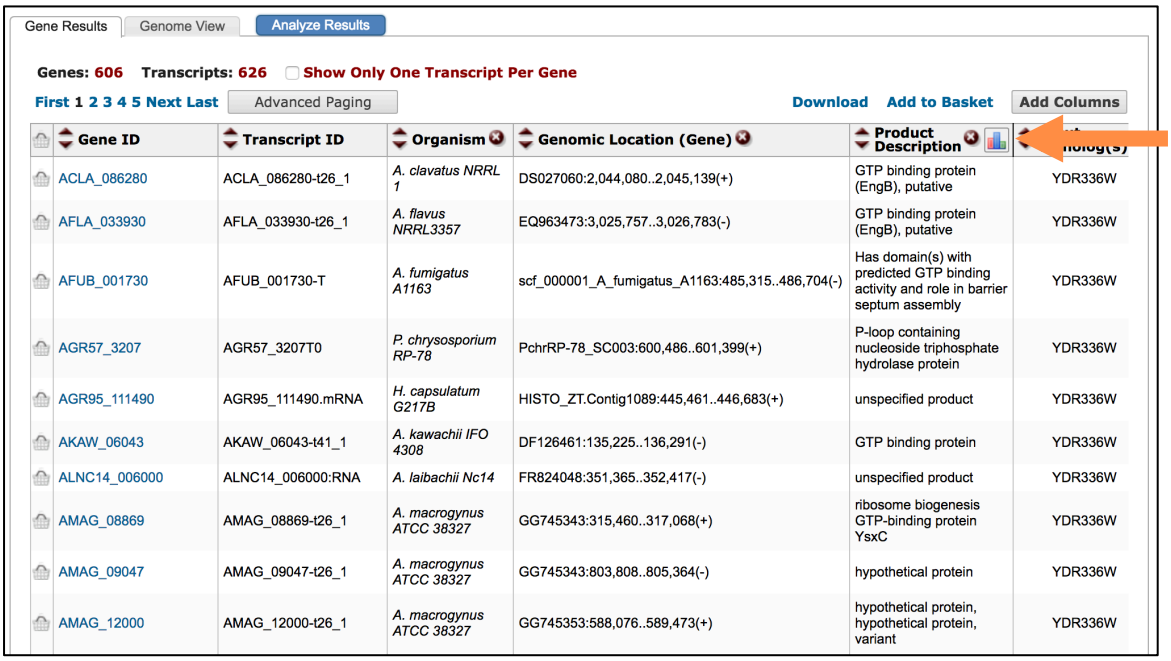
- Using your exported .tsv file from YeastMine, copy and paste the systematic names of your results into the box. Click on "**Get Answer**"
- Click on the "**Add a step**" button.



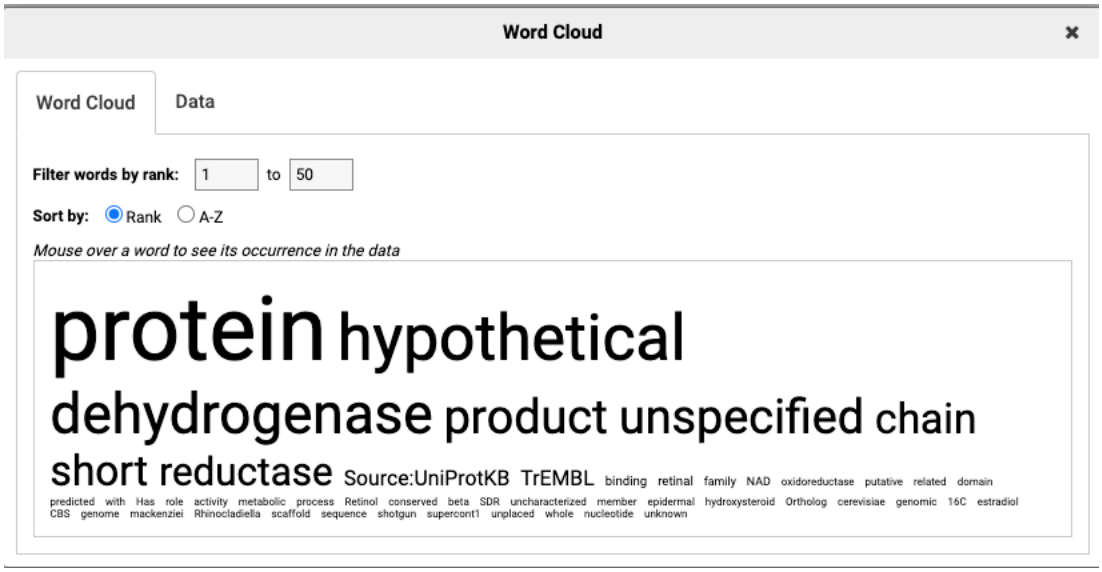
- In the resulting pop-up window, click on **Transform into Related Records**. Select **Orthologs** and then **Fungi** and **Oomycetes**, then click on **Run Step**.



- Orthologs from multiple species will be shown in the results table. Peruse the "Product Description" column. Do the descriptions of these orthologs support the prediction that the 8 yeast genes encode subunits of the mitochondrial ribosome? Click on the bar graph icon by the Product Description column to see a word cloud of entries in this column.



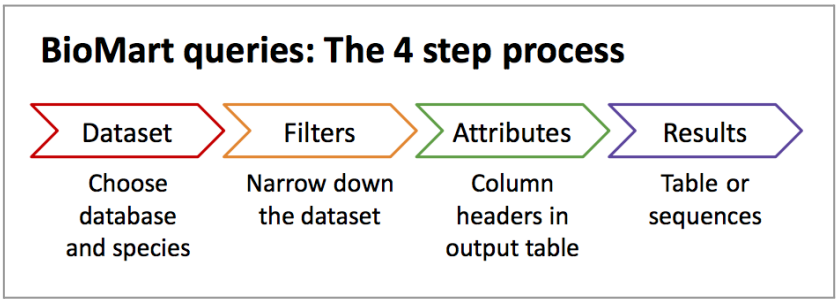
Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Ortholog(s)
ACLA_086280	ACLA_086280-t26_1	*A. clavatus NRRL 1*	DS027060:2,044,080..2,045,139(+)	GTP binding protein (EngB), putative	YDR336W
AFLA_033930	AFLA_033930-t26_1	*A. flavus NRRL3357*	EQ963473:3,025,757..3,026,783(-)	GTP binding protein (EngB), putative	YDR336W
AFUB_001730	AFUB_001730-T	*A. fumigatus A1163*	scf_000001_A_fumigatus_A1163:485,315..486,704(-)	Has domain(s) with predicted GTP binding activity and role in barrier septum assembly	YDR336W
AGR57_3207	AGR57_3207T0	*P. chrysosporium RP-78*	PchrRP-78_SC003:600,486..601,399(+)	P-loop containing nucleoside triphosphate hydrolase protein	YDR336W
AGR95_111490	AGR95_111490.mRNA	*H. capsulatum G217B*	HISTO_ZT.Contig1089:445,461..446,683(+)	unspecified product	YDR336W
AKAW_06043	AKAW_06043-t41_1	*A. kawachii IFO 4308*	DF126461:135,225..136,291(-)	GTP binding protein	YDR336W
ALNC14_006000	ALNC14_006000:RNA	*A. laibachii Nc14*	FR824048:351,365..352,417(-)	unspecified product	YDR336W
AMAG_08869	AMAG_08869-t26_1	*A. macrogynus ATCC 38327*	GG745343:315,460..317,068(+)	ribosome biogenesis GTP-binding protein YsxC	YDR336W
AMAG_09047	AMAG_09047-t26_1	*A. macrogynus ATCC 38327*	GG745343:803,808..805,364(-)	hypothetical protein	YDR336W
AMAG_12000	AMAG_12000-t26_1	*A. macrogynus ATCC 38327*	GG745353:588,076..589,473(+)	hypothetical protein, hypothetical protein, variant	YDR336W



Exercise: Ensembl Fungi BioMart

Follow these instructions to guide you through BioMart to answer the following query:

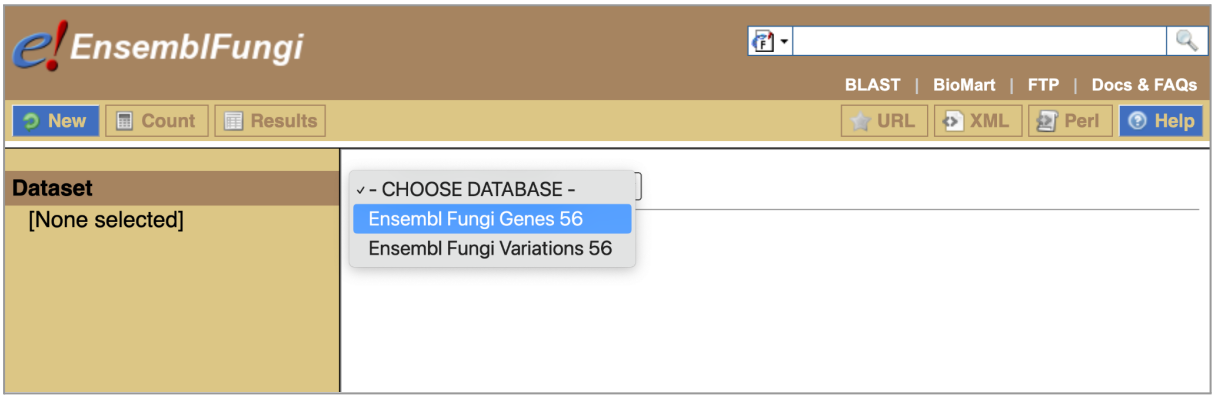
(a) How many genes within the 14:1128520-1142558 region are found in *Fusarium solani* that do not have an orthologue in *Fusarium verticillioides*?
(b) Export the gene name, locations and GO terms associated with these genes
(c) Export their cDNA sequences



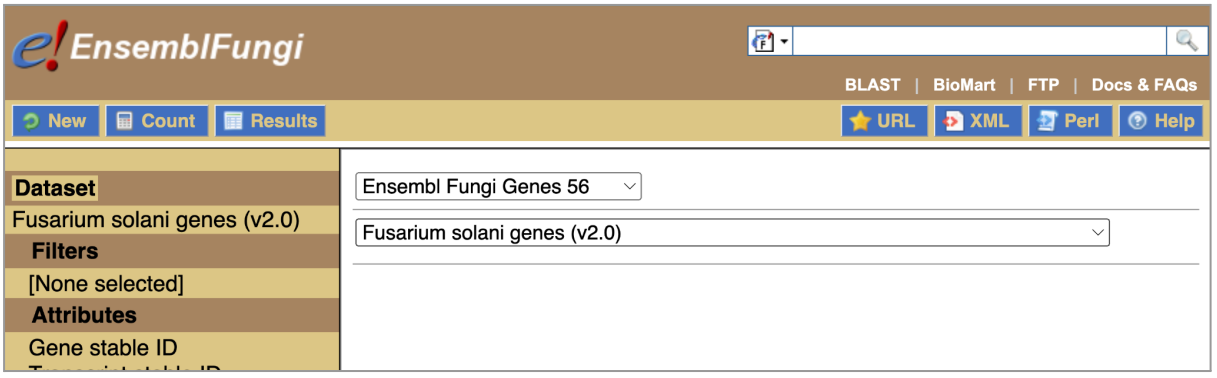
Click on BioMart in the top header of a fungi.ensembl.org page to go to: https://fungi.ensembl.org/biomart/martview/

NOTE: These answers were determined using BioMart Ensembl Fungi 56

Step 1a: Choose Ensembl Fungi Genes 56 as the database

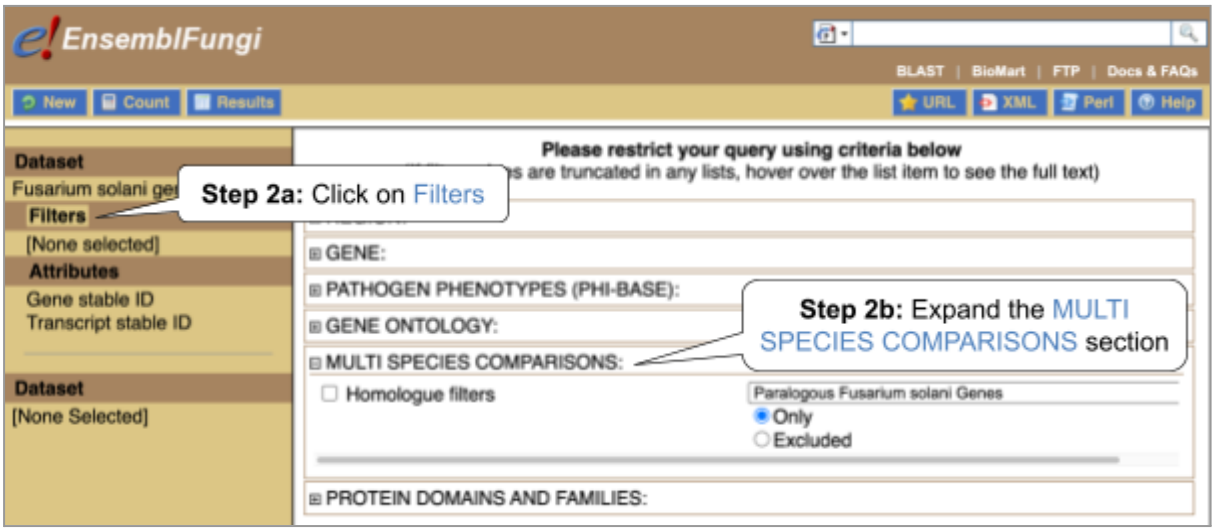


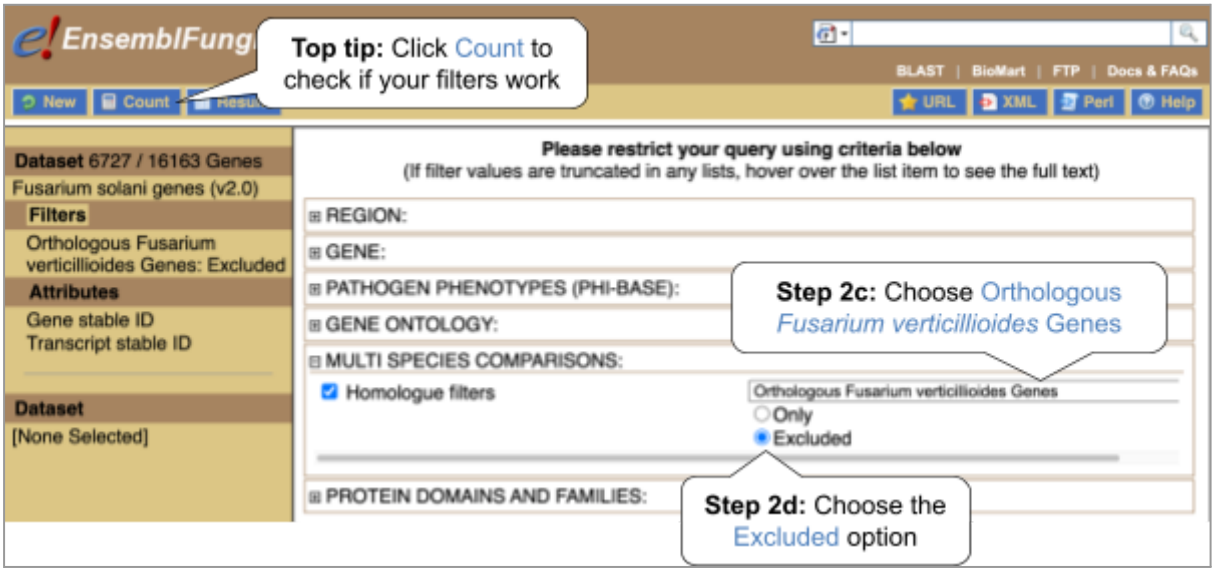
Step 1b: Choose *Fusarium solani* genes (v2.0) as the dataset

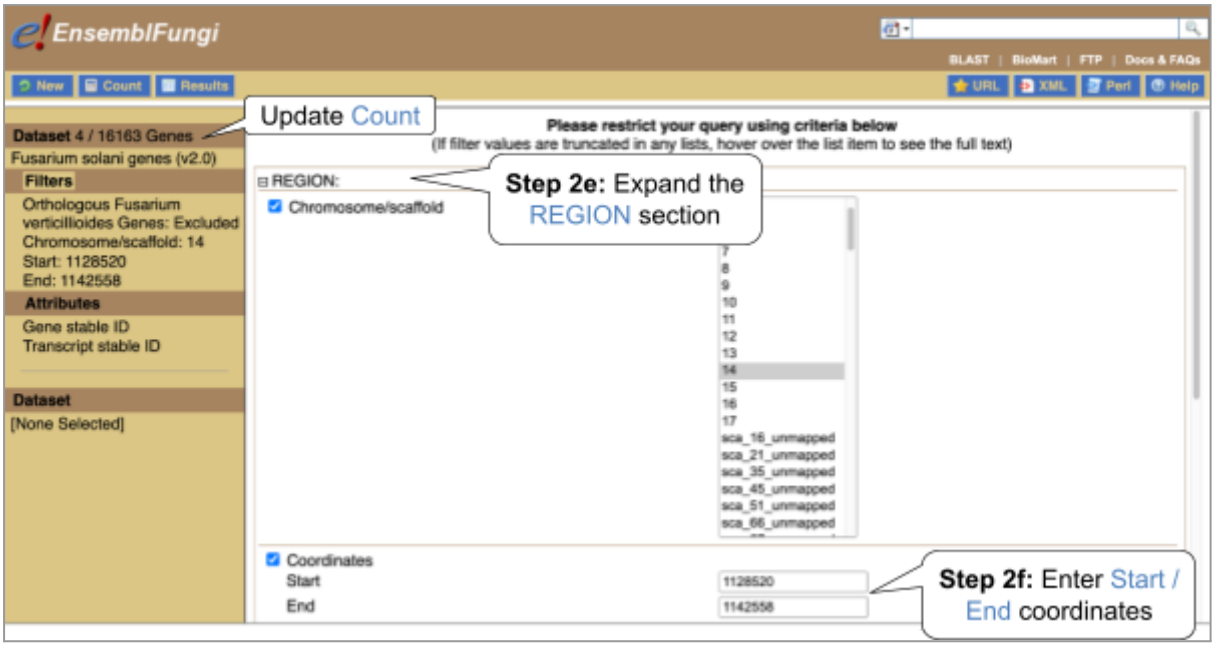


Step 2: Choose appropriate filters

We want to narrow down the dataset of all *F. solani* genes to a subset of genes matching our filters. We are interested in *F. solani* genes that **do not** have an orthologue with *F. verticillioides*. We need to filter the dataset to find these genes.







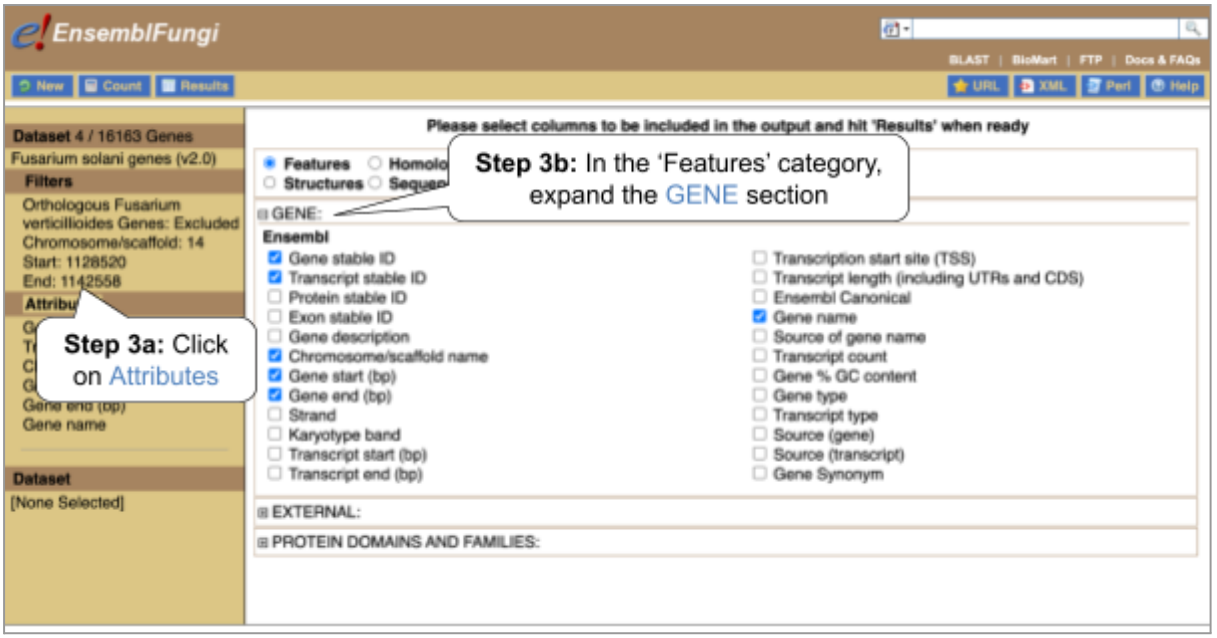
Using the count function we can see that there are 4 *F. solani* genes (out of a total of 16,163) in the 14:1128520-1142558 region that do not have an orthologue in *F. verticillioides*.

Step 3: Select Attributes

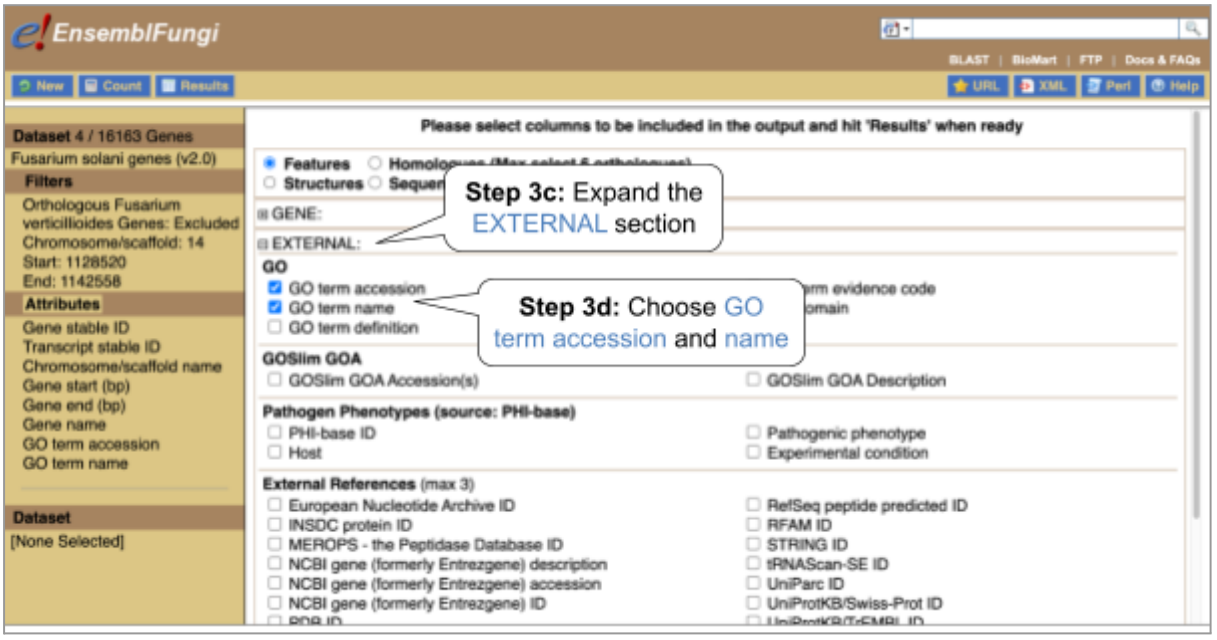
Attributes (our desired output) are defined by what we would like to learn about the data. We want to find out more information about these genes, including:

1. Gene name
2. Locations
3. Associated GO terms
4. cDNA sequences

There are four main attribute types: Features, Structures, Homologues and Sequences. BioMart allows querying only one type at a time. We can answer points 1-3 in a single query as they can all be found under Features, but we will need to build a second query to answer point 4 (Sequence type).



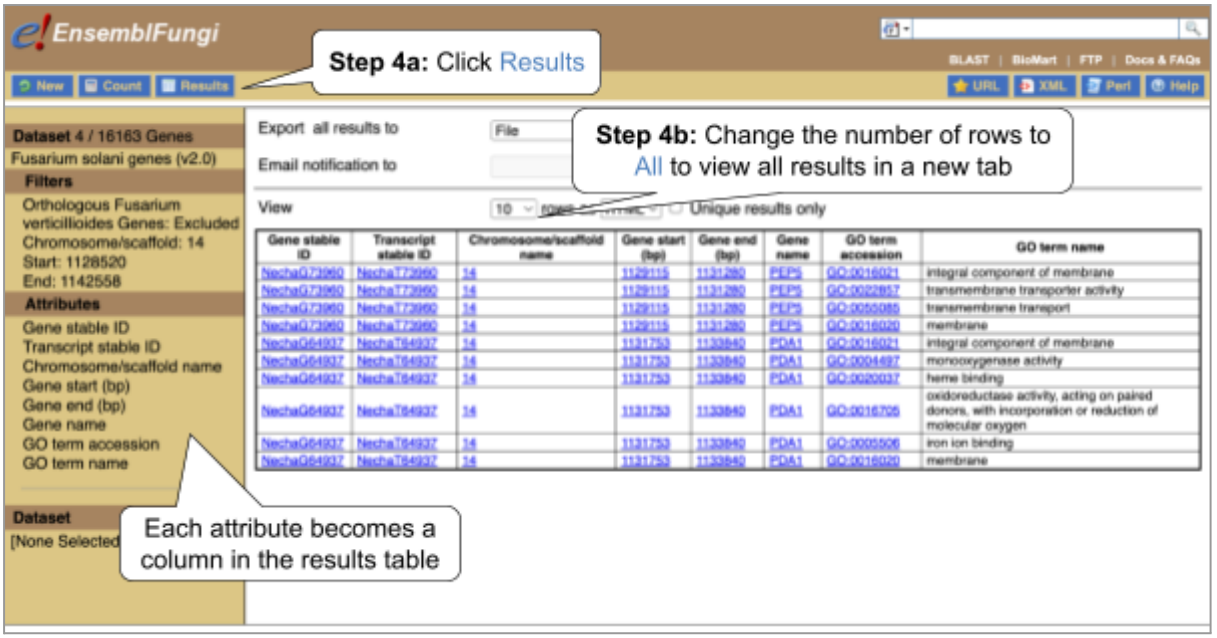
Make sure that Features is selected at the top of the page.
Expand the **GENE** section, and **select** Chromosome/scaffold name, Gene start and Gene end, and Gene name.



Expand the **EXTERNAL** section. This section contains lots of identifiers from databases outside of Ensembl. Select GO term accession and GO term name.

Step 4: Get results!

You can download the data if you'd like. The output table shows only 10 first rows by default.

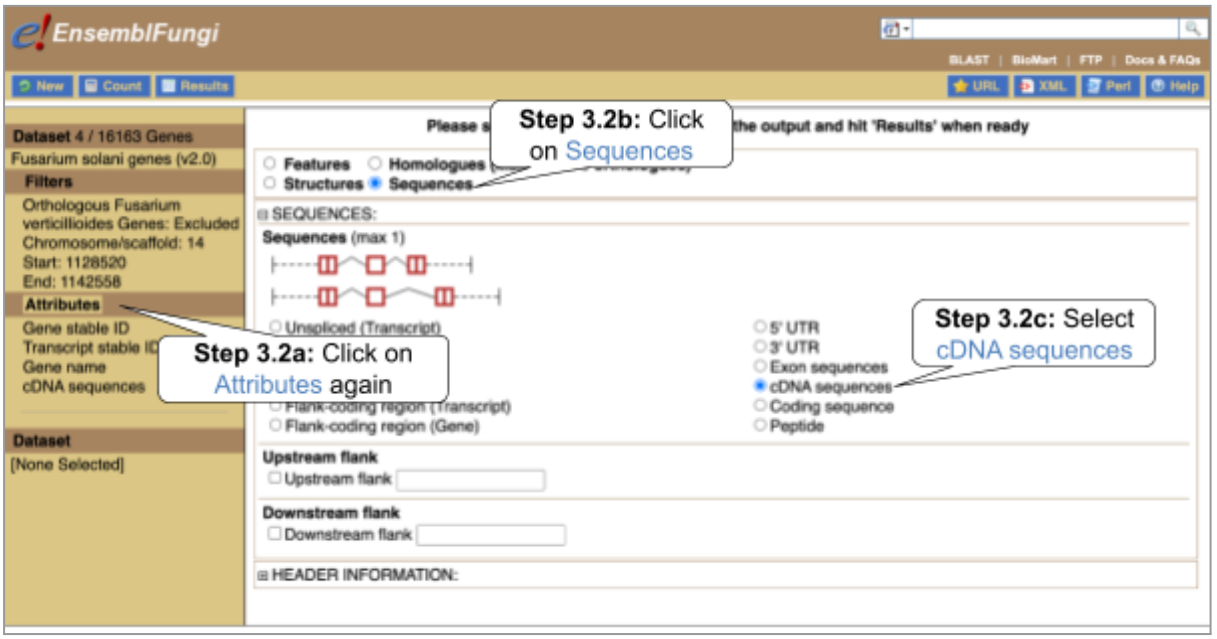


You can click on the location links and explore the synteny between the two species.

What about the last point? 'Export their cDNA sequences?'
In the Attributes section there are some 'radio buttons'. If you'd like to export Sequence data, you need to build a separate query.

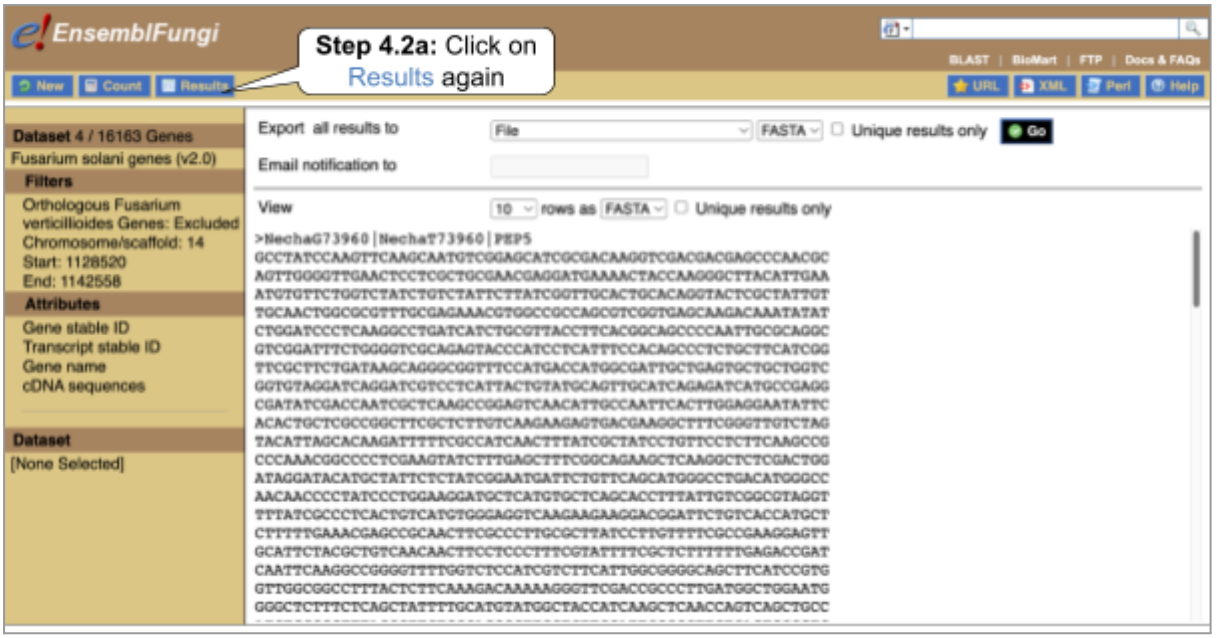
Step 3.2: Let's go back to step 3: Selecting attributes

From the results page, click back to Attributes in the left-hand navigation panel – there's no need to start from scratch.



Also expand the **HEADER INFORMATION** section and **select** Gene name.

Step 4.2: View results for the sequences



What did you learn about these genes in this exercise?
Could you learn these things from the Ensembl browser? Would it take longer?

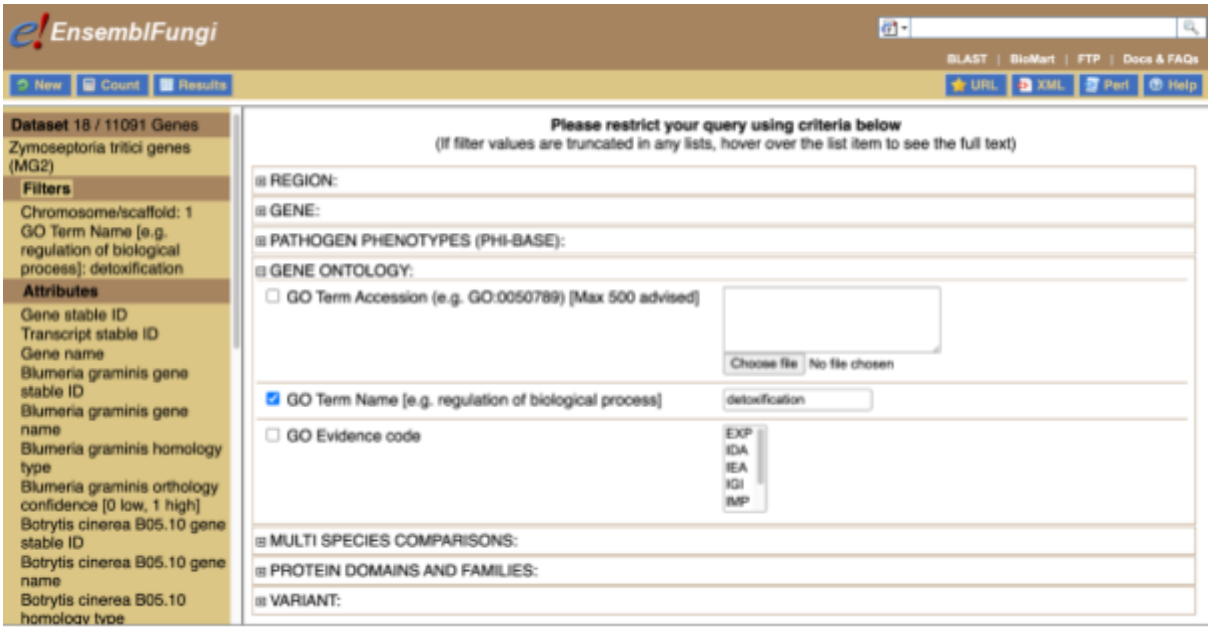
For more details on BioMart, have a look at this publication:

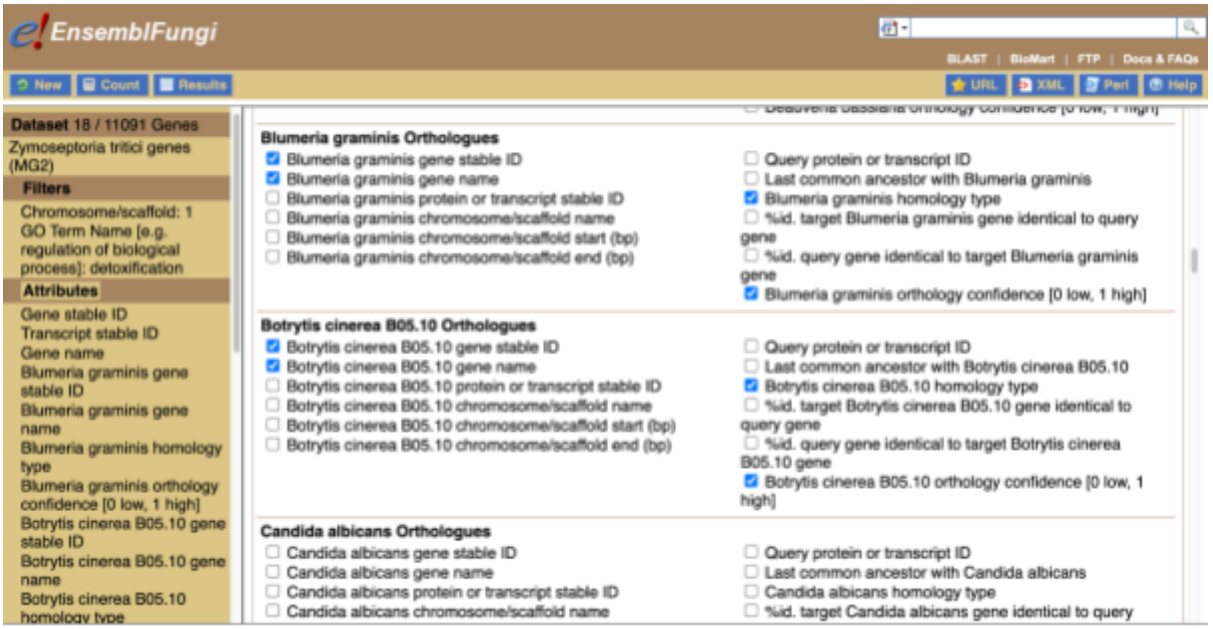
Kinsella RJ, Kähäri A, Haider S, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database : the Journal of Biological Databases and Curation. 2011;2011:bar030. DOI: 10.1093/database/bar030. PMID: 21785142; PMCID: PMC3170168.

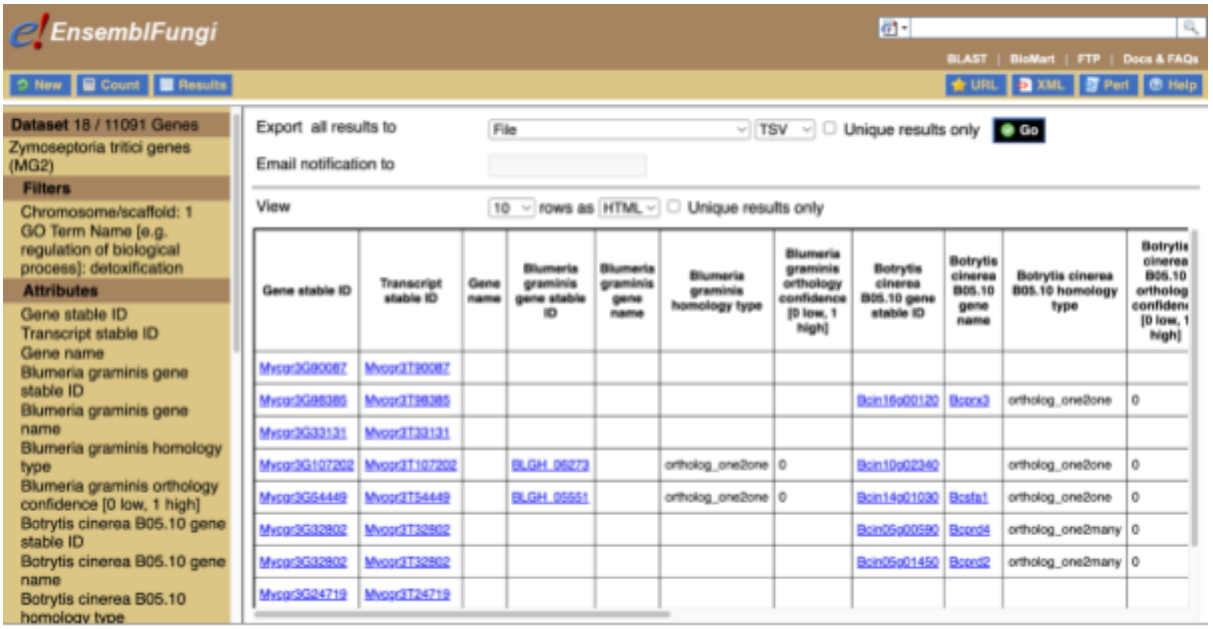
Additional BioMart Exercise 1 – Export orthologues

Use Ensembl Fungi BioMart to retrieve all *Zymoseptoria tritici* genes associated with the GO term detoxification located on chromosome 1. Export the gene IDs, names, homology type and confidence of their orthologues in *Blumeria graminis*, *Botrytis cinerea*, *Cryptococcus neoformans* and *Saccharomyces cerevisiae*.

(a) Do all of these *Z. tritici* genes have an orthologue in the other species? Which of these species are pathogenic? Do you see a correlation?
(b) Can you find an orthologue in *Cryptococcus neoformans* with high orthology confidence? What is the Gene ID? We will explore more about this orthologue in the exercise section for the Evolutionary Analysis module.

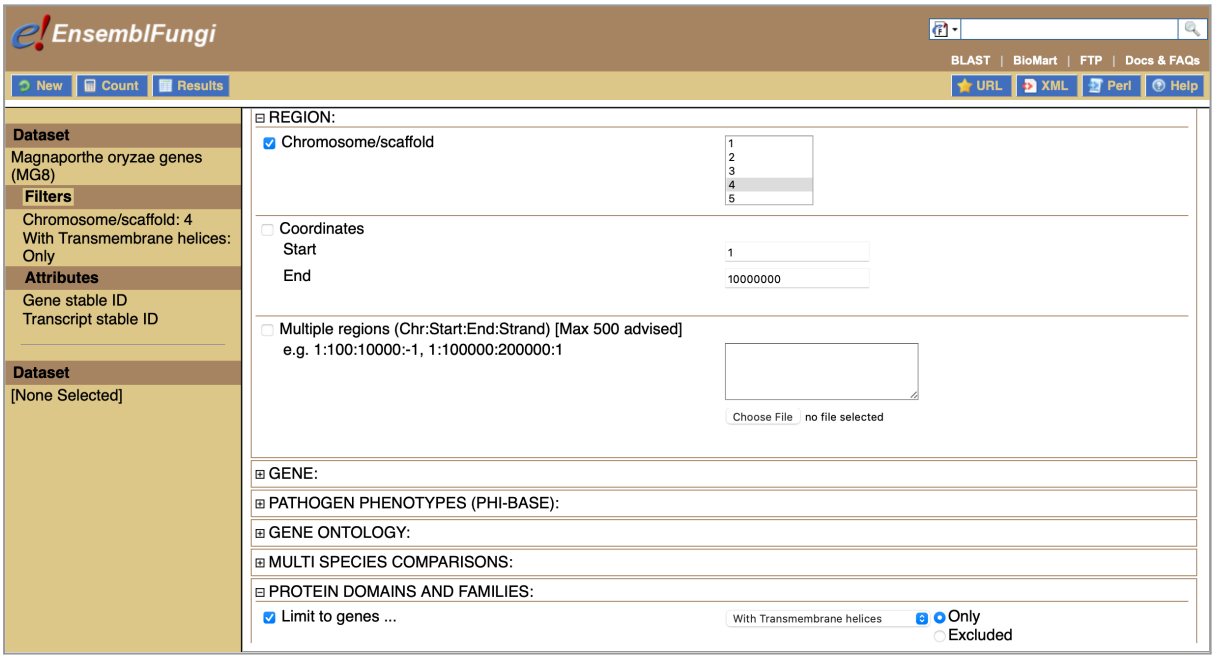


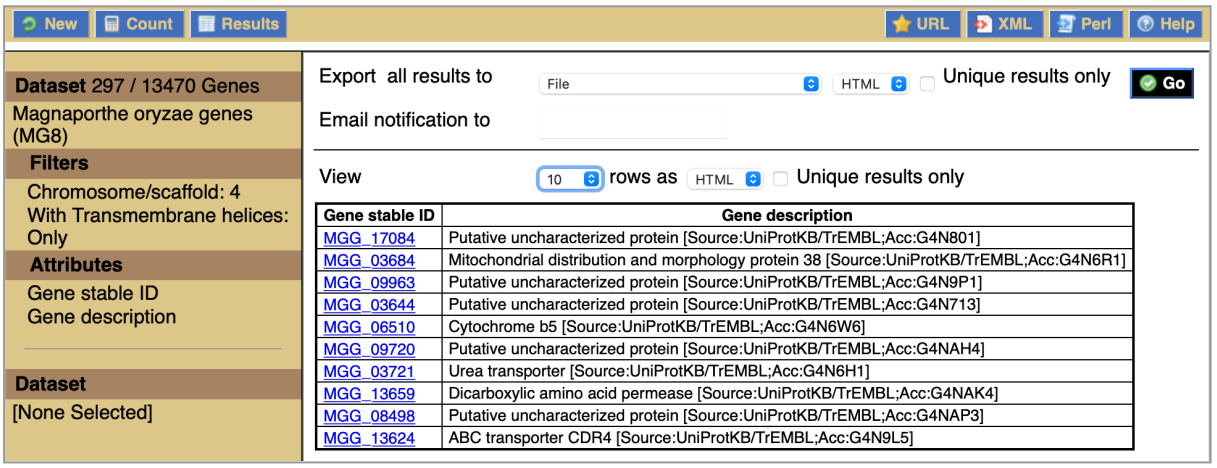



EnsemblFungi
BLAST | BioMart | FTP | Docs & FAQs
New
Count
Results
URL
XML
Perl
Help
Dataset 18 / 11091 Genes
Zymoseptoria tritici genes (MG2)
Filters
Chromosome/scaffold: 1
GO Term Name [e.g. regulation of biological process]: detoxification
Attributes
Gene stable ID
Transcript stable ID
Gene name
Blumeria graminis gene stable ID
Blumeria graminis gene name
Blumeria graminis homology type
Blumeria graminis orthology confidence [0 low, 1 high]
Botrytis cinerea B05.10 gene stable ID
Botrytis cinerea B05.10 gene name
Botrytis cinerea B05.10 homology type
Export all results to
File
TSV
Unique results only
Go
Email notification to
View
10
rows as
HTML
Unique results only
Gene stable ID
Transcript stable ID
Gene name
Blumeria graminis gene stable ID
Blumeria graminis gene name
Blumeria graminis homology type
Blumeria graminis orthology confidence [0 low, 1 high]
Botrytis cinerea B05.10 gene stable ID
Botrytis cinerea B05.10 gene name
Botrytis cinerea B05.10 homology type
Botrytis cinerea B05.10 orthology confidence [0 low, 1 high]
Mycgr3G90087
Mycgr3T90087
Mycgr3G98385
Mycgr3T98385
Bcin16g00120
Bcprx3
ortholog_one2one
0
Mycgr3G33131
Mycgr3T33131
Mycgr3G107202
Mycgr3T107202
BLGH_06273
ortholog_one2one
0
Bcin10g02340
ortholog_one2one
0
Mycgr3G54449
Mycgr3T54449
BLGH_05551
ortholog_one2one
0
Bcin14g01030
Bcsfa1
ortholog_one2one
0
Mycgr3G32802
Mycgr3T32802
Bcin05g00580
Bcprd4
ortholog_one2many
0
Mycgr3G32802
Mycgr3T32802
Bcin05g01450
Bcprd2
ortholog_one2many
0
Mycgr3G24719
Mycgr3T24719

Additional BioMart Exercise 2 – Finding genes by protein domain

Generate a list of all *Magnaporthe oryzae* genes on chromosome 4 that are annotated to contain Transmembrane domains/helices. Include the Ensembl Gene ID and description.





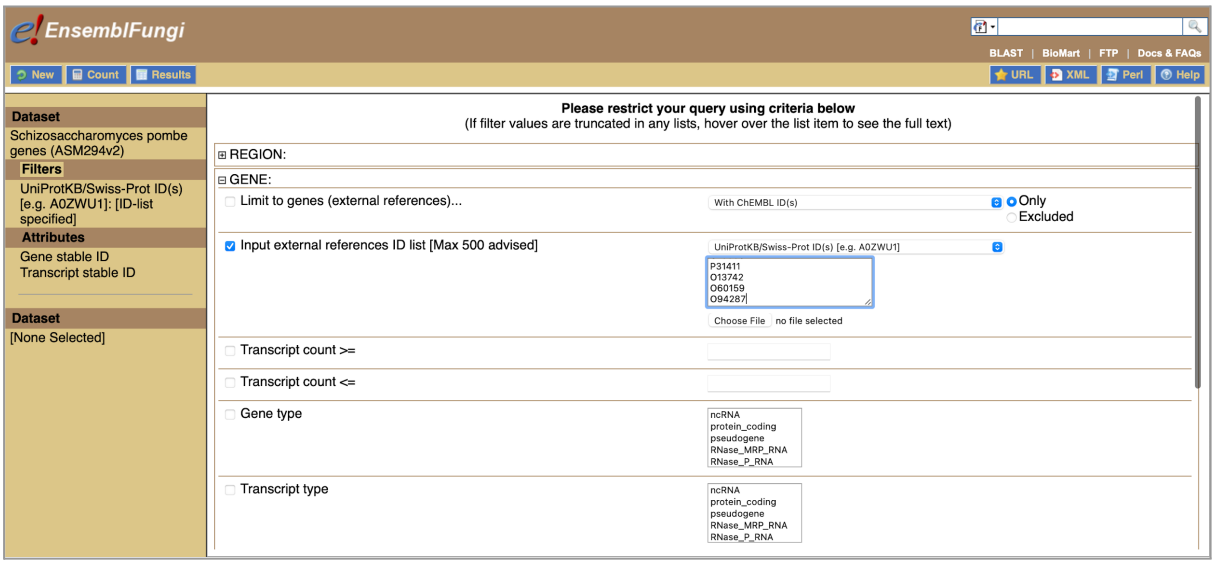
Gene stable ID	Gene description
MGG_17084	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:G4N801]
MGG_03684	Mitochondrial distribution and morphology protein 38 [Source:UniProtKB/TrEMBL;Acc:G4N6R1]
MGG_09963	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:G4N9P1]
MGG_03644	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:G4N713]
MGG_06510	Cytochrome b5 [Source:UniProtKB/TrEMBL;Acc:G4N6W6]
MGG_09720	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:G4NAH4]
MGG_03721	Urea transporter [Source:UniProtKB/TrEMBL;Acc:G4N6H1]
MGG_13659	Dicarboxylic amino acid permease [Source:UniProtKB/TrEMBL;Acc:G4NAK4]
MGG_08498	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:G4NAP3]
MGG_13624	ABC transporter CDR4 [Source:UniProtKB/TrEMBL;Acc:G4N9L5]

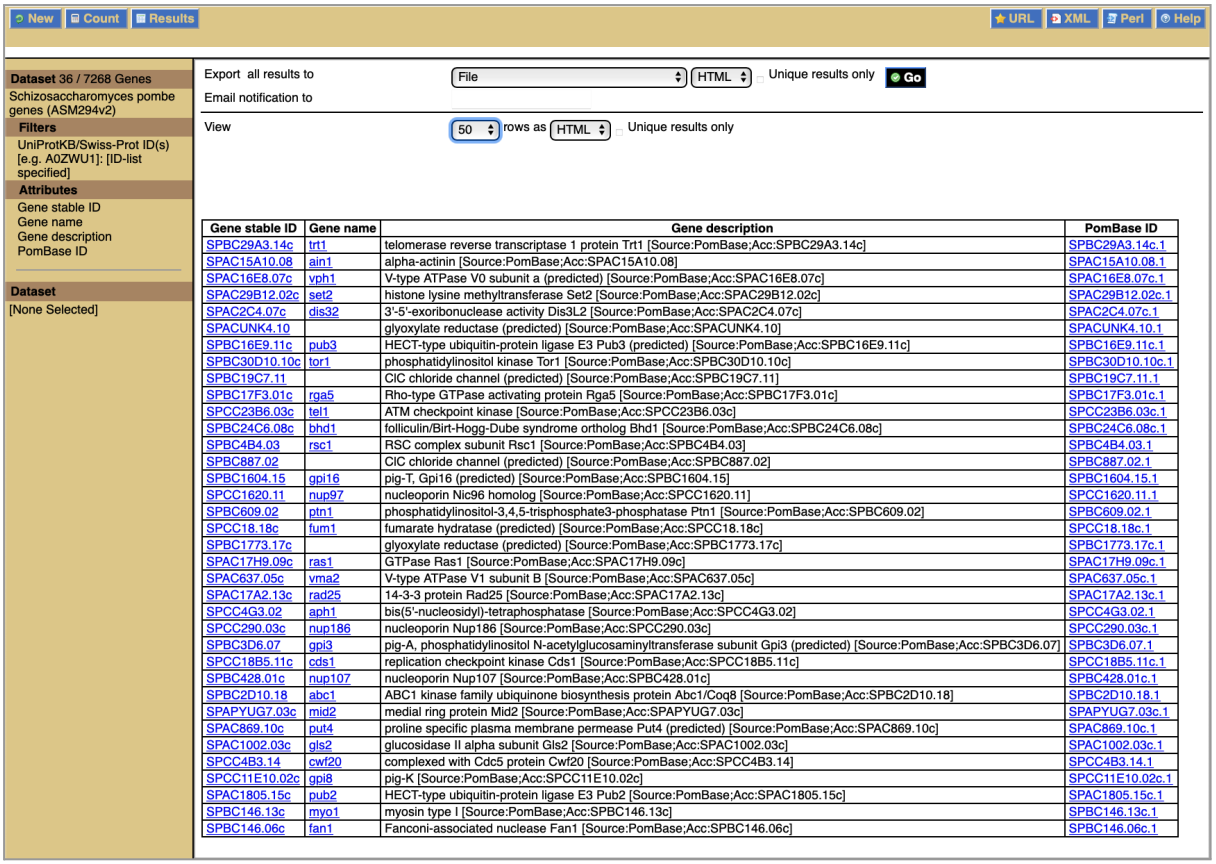
Additional BioMart Exercise 3 – Convert IDs

For a list of *Schizosaccharomyces pombe* UniProt (UniProtKB/Swiss-Prot) IDs, export the Gene name and description, as well as the PomBase IDs. Do these 36 protein IDs correspond to 36 genes?

Input list of IDs:

Q92338	Q9US55	P78847	O74964
O13728	O14075	O94418	O14026
P49776	O94574	O94526	O74630
O74769	O94380	Q9UTG2	O14356
Q09170	P87172	O14326	O13339
Q9USK4	Q9USP5	Q9URZ3	P31411
O14040	Q9P7Y8	P42657	O13742
Q9Y804	Q9Y7Z8	P08647	O60159
O94552	Q10331	O74335	O94287





Gene stable ID	Gene name	Gene description	PomBase ID
SPBC29A3.14c	trt1	telomerase reverse transcriptase 1 protein Trt1 [Source:PomBase;Acc:SPBC29A3.14c]	SPBC29A3.14c.1
SPAC15A10.08	ain1	alpha-actinin [Source:PomBase;Acc:SPAC15A10.08]	SPAC15A10.08.1
SPAC16E8.07c	vph1	V-type ATPase V0 subunit a (predicted) [Source:PomBase;Acc:SPAC16E8.07c]	SPAC16E8.07c.1
SPAC29B12.02c	set2	histone lysine methyltransferase Set2 [Source:PomBase;Acc:SPAC29B12.02c]	SPAC29B12.02c.1
SPAC2C4.07c	dis32	3'-5'-exoribonuclease activity Dis3L2 [Source:PomBase;Acc:SPAC2C4.07c]	SPAC2C4.07c.1
SPACUNK4.10		glyoxylate reductase (predicted) [Source:PomBase;Acc:SPACUNK4.10]	SPACUNK4.10.1
SPBC16E9.11c	pub3	HECT-type ubiquitin-protein ligase E3 Pub3 (predicted) [Source:PomBase;Acc:SPBC16E9.11c]	SPBC16E9.11c.1
SPBC30D10.10c	tor1	phosphatidylinositol kinase Tor1 [Source:PomBase;Acc:SPBC30D10.10c]	SPBC30D10.10c.1
SPBC19C7.11		ClC chloride channel (predicted) [Source:PomBase;Acc:SPBC19C7.11]	SPBC19C7.11.1
SPBC17F3.01c	rga5	Rho-type GTPase activating protein Rga5 [Source:PomBase;Acc:SPBC17F3.01c]	SPBC17F3.01c.1
SPCC23B6.03c	tel1	ATM checkpoint kinase [Source:PomBase;Acc:SPCC23B6.03c]	SPCC23B6.03c.1
SPBC24C6.08c	bhd1	folliculin/Birt-Hogg-Dube syndrome ortholog Bhd1 [Source:PomBase;Acc:SPBC24C6.08c]	SPBC24C6.08c.1
SPBC4B4.03	rsc1	RSC complex subunit Rsc1 [Source:PomBase;Acc:SPBC4B4.03]	SPBC4B4.03.1
SPBC887.02		ClC chloride channel (predicted) [Source:PomBase;Acc:SPBC887.02]	SPBC887.02.1
SPBC1604.15	gpi16	pig-T, Gpi16 (predicted) [Source:PomBase;Acc:SPBC1604.15]	SPBC1604.15.1
SPCC1620.11	nup97	nucleoporin Nic96 homolog [Source:PomBase;Acc:SPCC1620.11]	SPCC1620.11.1
SPBC609.02	ptn1	phosphatidylinositol-3,4,5-trisphosphate3-phosphatase Ptn1 [Source:PomBase;Acc:SPBC609.02]	SPBC609.02.1
SPCC18.18c	fum1	fumarate hydratase (predicted) [Source:PomBase;Acc:SPCC18.18c]	SPCC18.18c.1
SPBC1773.17c		glyoxylate reductase (predicted) [Source:PomBase;Acc:SPBC1773.17c]	SPBC1773.17c.1
SPAC17H9.09c	ras1	GTPase Ras1 [Source:PomBase;Acc:SPAC17H9.09c]	SPAC17H9.09c.1
SPAC637.05c	vma2	V-type ATPase V1 subunit B [Source:PomBase;Acc:SPAC637.05c]	SPAC637.05c.1
SPAC17A2.13c	rad25	14-3-3 protein Rad25 [Source:PomBase;Acc:SPAC17A2.13c]	SPAC17A2.13c.1
SPCC4G3.02	aph1	bis(5'-nucleosidyl)-tetraphosphatase [Source:PomBase;Acc:SPCC4G3.02]	SPCC4G3.02.1
SPCC290.03c	nup186	nucleoporin Nup186 [Source:PomBase;Acc:SPCC290.03c]	SPCC290.03c.1
SPBC3D6.07	gpi3	pig-A, phosphatidylinositol N-acetylglucosaminyltransferase subunit Gpi3 (predicted) [Source:PomBase;Acc:SPBC3D6.07]	SPBC3D6.07.1
SPCC18B5.11c	cds1	replication checkpoint kinase Cds1 [Source:PomBase;Acc:SPCC18B5.11c]	SPCC18B5.11c.1
SPBC428.01c	nup107	nucleoporin Nup107 [Source:PomBase;Acc:SPBC428.01c]	SPBC428.01c.1
SPBC2D10.18	abc1	ABC1 kinase family ubiquinone biosynthesis protein Abc1/Coq8 [Source:PomBase;Acc:SPBC2D10.18]	SPBC2D10.18.1
SPAPYUG7.03c	mid2	medial ring protein Mid2 [Source:PomBase;Acc:SPAPYUG7.03c]	SPAPYUG7.03c.1
SPAC869.10c	put4	proline specific plasma membrane permease Put4 (predicted) [Source:PomBase;Acc:SPAC869.10c]	SPAC869.10c.1
SPAC1002.03c	gls2	glucosidase II alpha subunit Gls2 [Source:PomBase;Acc:SPAC1002.03c]	SPAC1002.03c.1
SPCC4B3.14	cwf20	complexed with Cdc5 protein Cwf20 [Source:PomBase;Acc:SPCC4B3.14]	SPCC4B3.14.1
SPCC11E10.02c	gpi8	pig-K [Source:PomBase;Acc:SPCC11E10.02c]	SPCC11E10.02c.1
SPAC1805.15c	pub2	HECT-type ubiquitin-protein ligase E3 Pub2 [Source:PomBase;Acc:SPAC1805.15c]	SPAC1805.15c.1
SPBC146.13c	myo1	myosin type I [Source:PomBase;Acc:SPBC146.13c]	SPBC146.13c.1
SPBC146.06c	fan1	Fanconi-associated nuclease Fan1 [Source:PomBase;Acc:SPBC146.06c]	SPBC146.06c.1

Exercise: Exploring host-pathogen interactions in Ensembl Fungi

Zymoseptoria tritici, also known as *Septoria tritici* and *Mycosphaerella graminicola*, is a fungal pathogen that causes septoria leaf blotch disease in wheat. This fungus is considered a major threat to wheat production world-wide, 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.

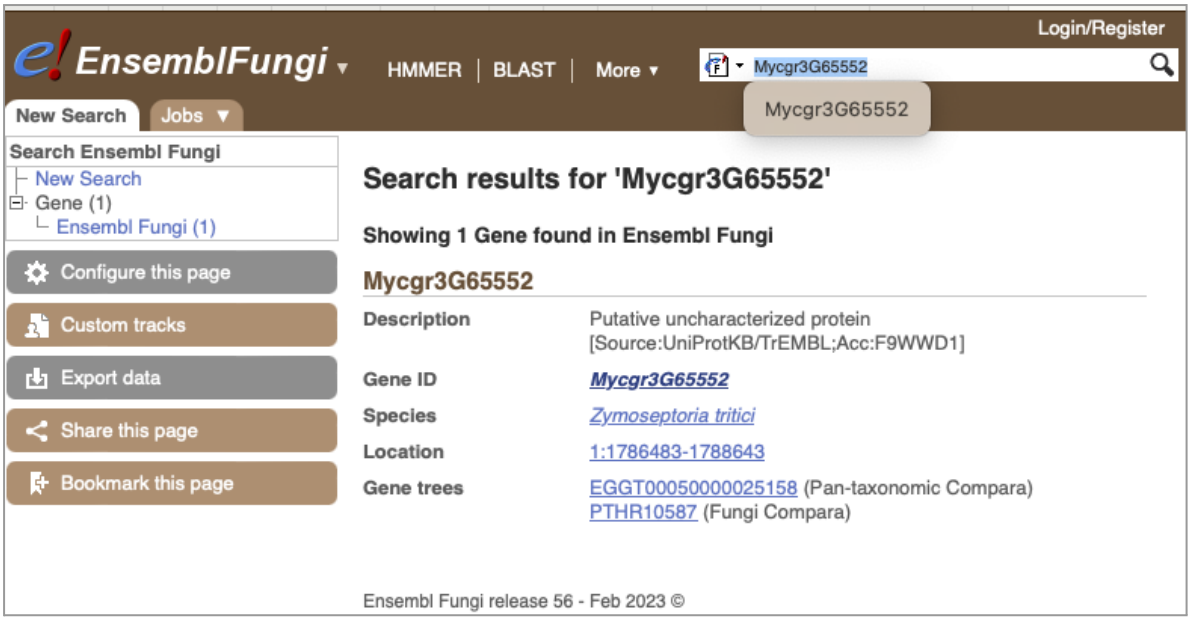
Step 1: Find all genes with molecular interactions for *Zymoseptoria tritici*.
From the endpoint API doc page https://interactions.rest.ensembl.org, search for all zymoseptoria_tritici genes

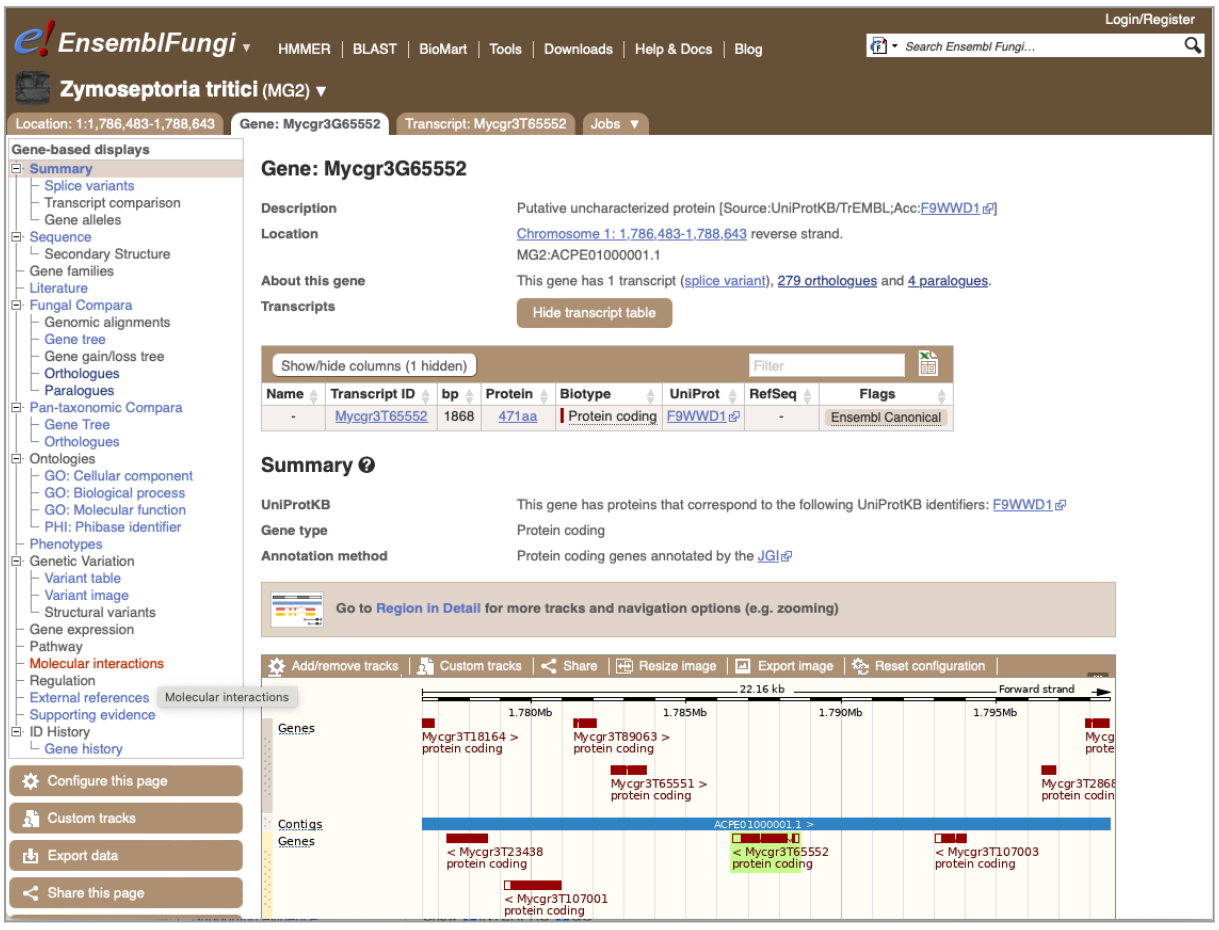
https://interactions.rest.ensembl.org/ensembl_gene?scientific_name=zymoseptoria tritici

You should get the following output:

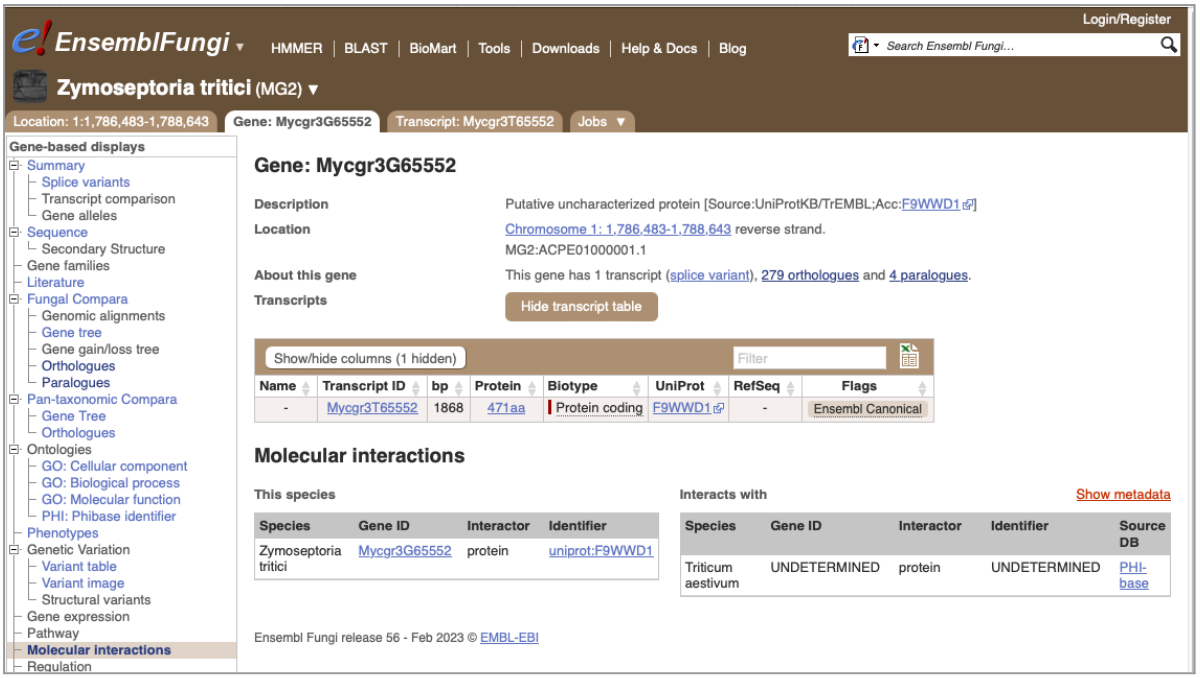
```
zymoseptoria_tritici": ["Mycgr3G53658", "Mycgr3g88451",
"Mycgr3G85040", "Mycgr3G40048", "Mycgr3G111221", "Mycgr3G103264",
"Mycgr3G89160", "Mycgr3G80707", "Mycgr3G65552", "Mycgr3g105487",
"Mycgr3G70181", "Mycgr3G46840", "Mycgr3G93828", "Mycgr3G31676",
"Mycgr3G51018", "Mycgr3G36951", "Mycgr3G77528",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

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

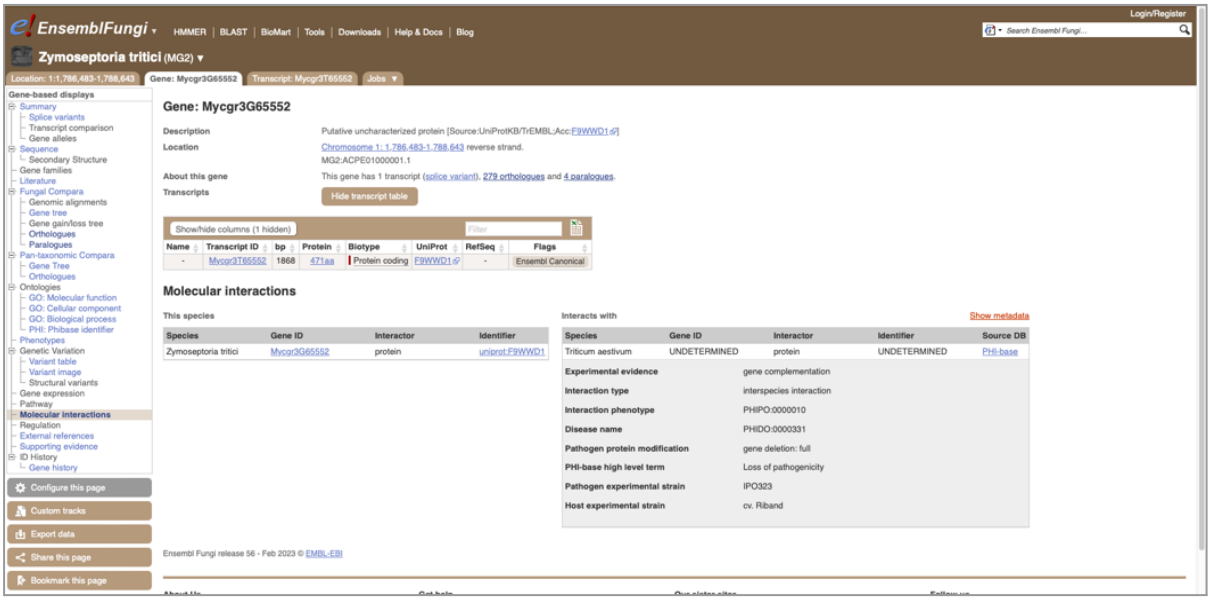




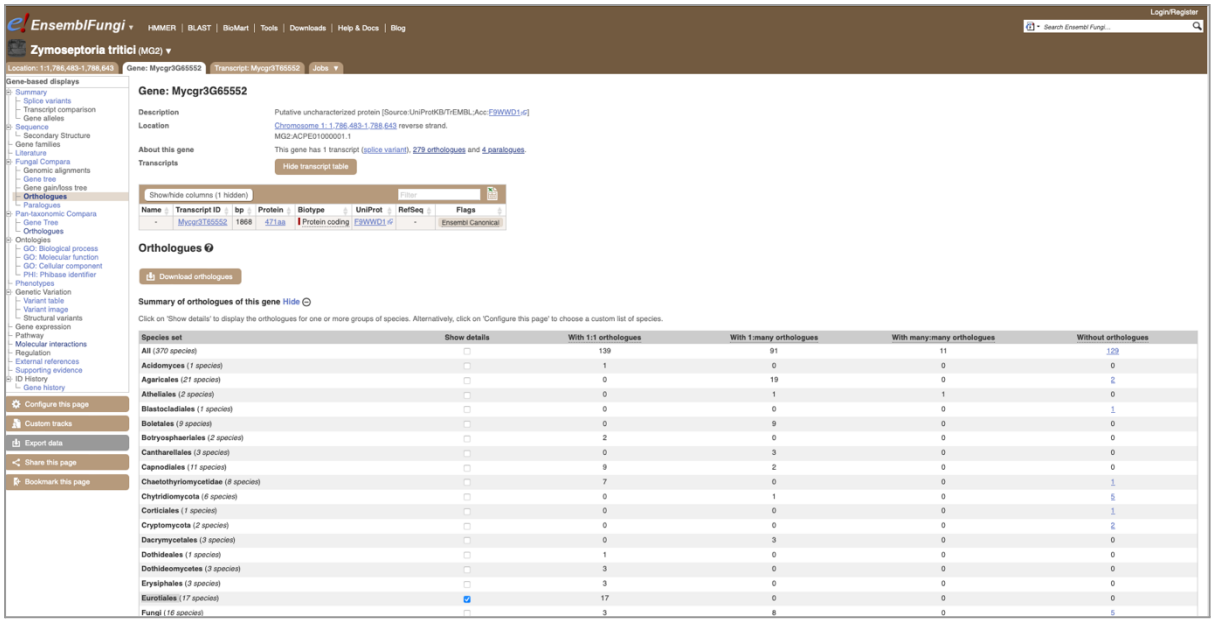
To find a list of species with which this particular *Z. tritici* gene has molecular interactions with, click on Molecular interactions in the left-hand panel. From this page, we can see that *Z. tritici* is known to interact with *Triticum aestivum* (wheat). Can you find the wheat Gene ID that Mycgr3G65552 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 hasn't been identified yet.



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.

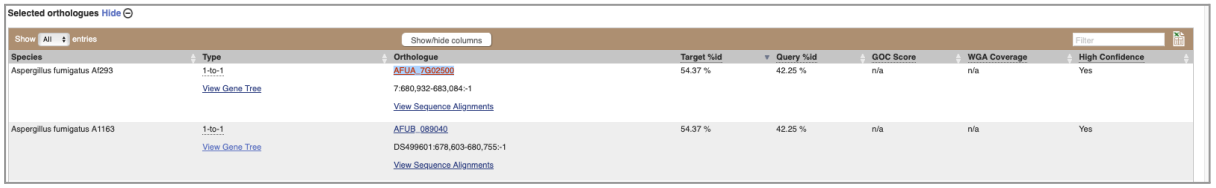


Step 3: Next, let's find all fungal orthologues. There are several ways of doing this. One way is to go to Fungal Compara: Orthologues in the left-hand panel.

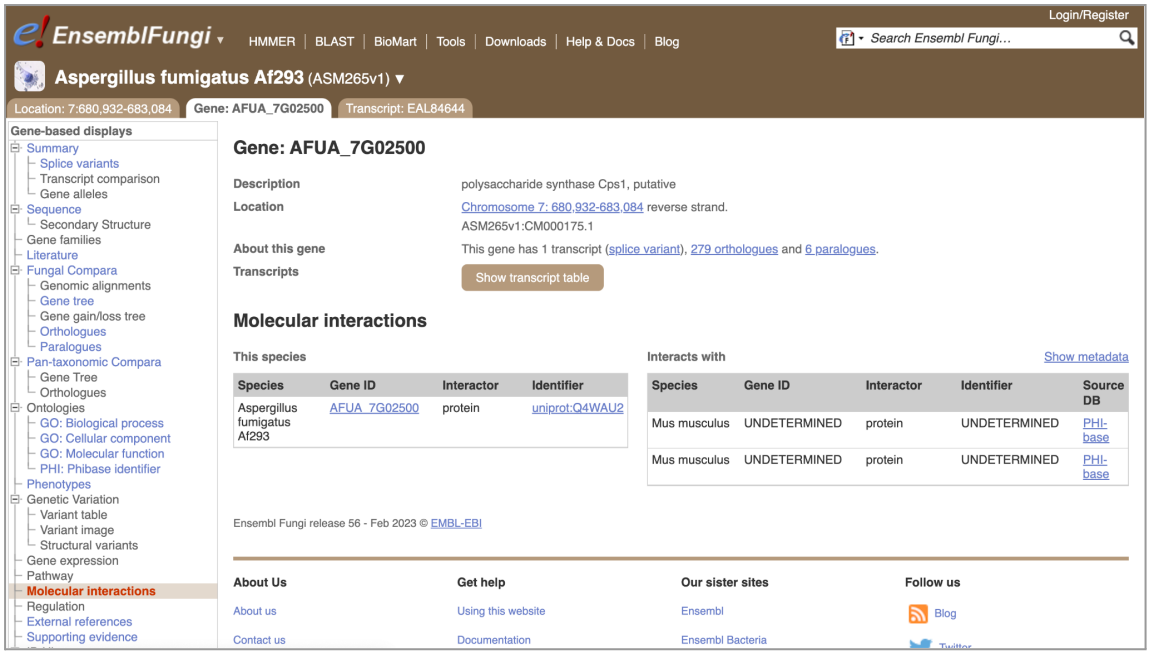


Can you find out if there are any orthologues in *Aspergilus fumigatus* with molecular interactions entries?

Step 4: You can hide the Summary of orthologues of this gene table by clicking the Hide button. Enter *Aspergilus 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?


Login/Register
EnsemblFungi
HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog
Search Ensembl Fungi…
Aspergillus fumigatus Af293 (ASM265v1)
Location: 7:680,932-683,084
Gene: AFUA_7G02500
Transcript: EAL84644
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
Gene: AFUA_7G02500
Description
polysaccharide synthase Cps1, putative
Location
Chromosome 7: 680,932-683,084 reverse strand.
ASM265v1:CM000175.1
About this gene
This gene has 1 transcript (splice variant), 279 orthologues and 6 paralogues.
Transcripts
Show transcript table
Molecular interactions
This species
Species
Gene ID
Interactor
Identifier
Aspergillus fumigatus Af293
AFUA_7G02500
protein
uniprot:Q4WAU2
Interacts with
Show metadata
Species
Gene ID
Interactor
Identifier
Source DB
Mus musculus
UNDETERMINED
protein
UNDETERMINED
PHI-base
Mus musculus
UNDETERMINED
protein
UNDETERMINED
PHI-base
Ensembl Fungi release 56 - Feb 2023 © EMBL-EBI
About Us
About us
Contact us
Get help
Using this website
Documentation
Our sister sites
Ensembl
Ensembl Bacteria
Follow us
Blog
Twitter

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. Use Ensembl Fungi to find out more information about molecular interactions in the species and answer the following questions:

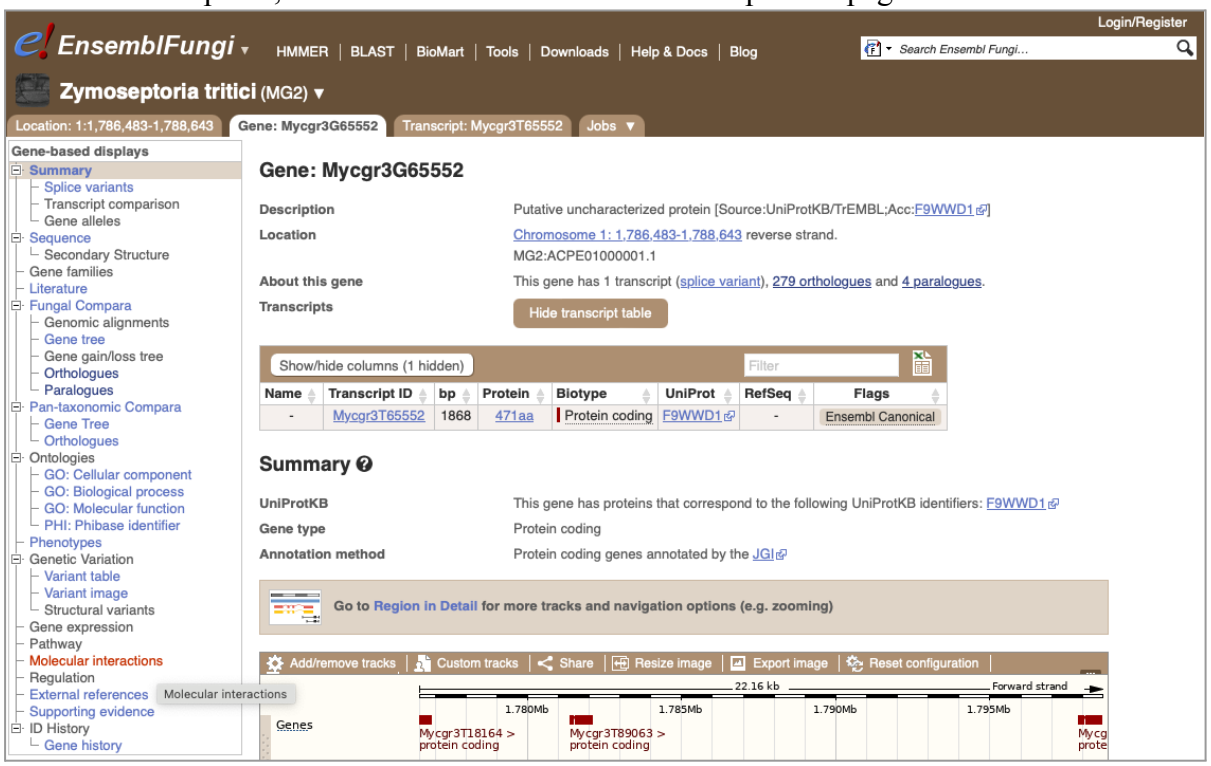
(a) Using the Ensembl REST API, can you retrieve all genes with molecular interactions information for *B. cinerea*?
(b) Open the Molecular interactions page for the Bcin07g00720 gene in *B. cinerea*. What plant species does the gene interact with?
(c) Can you find the phenotype that is reported for 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?

Answers:
(a) From the endpoint https://interactions.rest.ensembl.org/interactions_by_prodname/. Search for botrytis_cynerea, which will give you 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"]
```

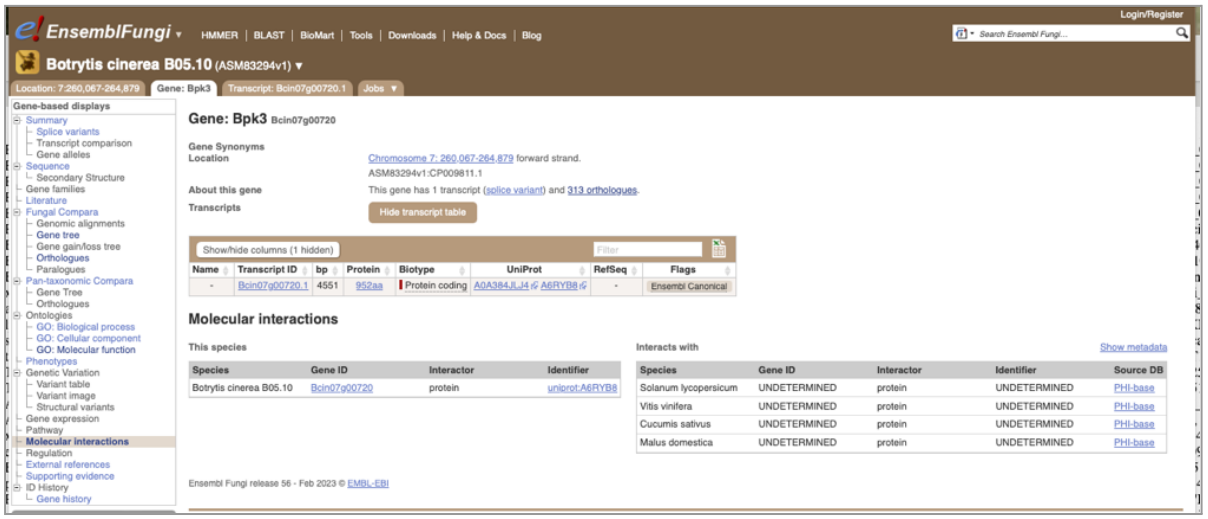

(b) Go to the Ensembl Fungi homepage and search for Bcin07g00720. In the results, click on the Gene ID to open the Gene tab.

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

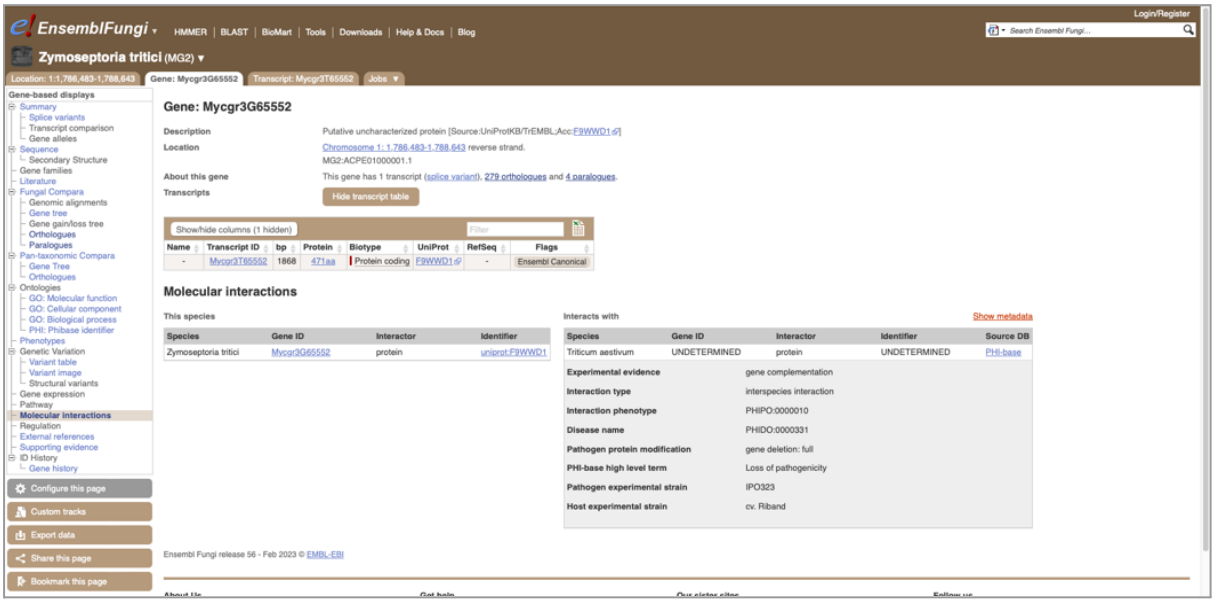


In the Molecular interactions page, you can find all species the gene interacts with in the right-hand table. These include *Solanum lycopersicum* (tomato), *Vitis vinifera* (grape),

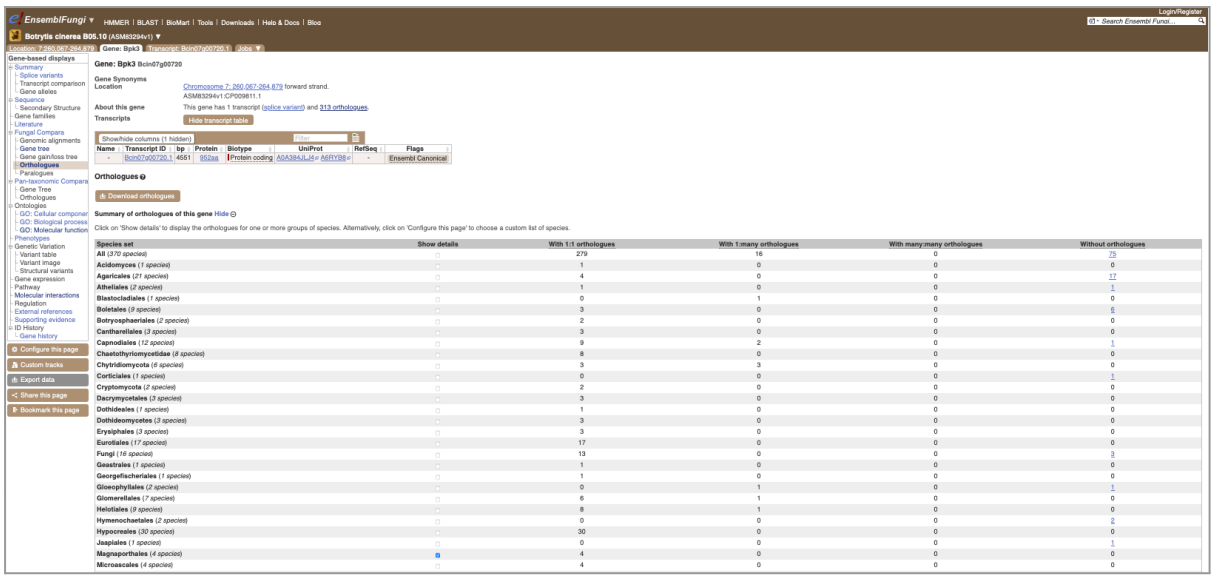
Cucumis sativus (cucumber) and *Malus domestica* (apple).



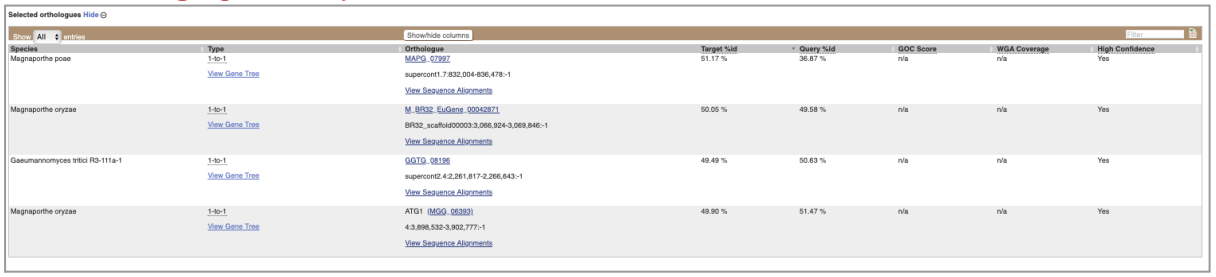
(c) Click on Show metadata in the right-hand corner of the Interacts with table. You can find associated phenotypes under PHI-base high level term. The gene is associated with “Reduced virulence” and / or “Loss of pathogenicity”.



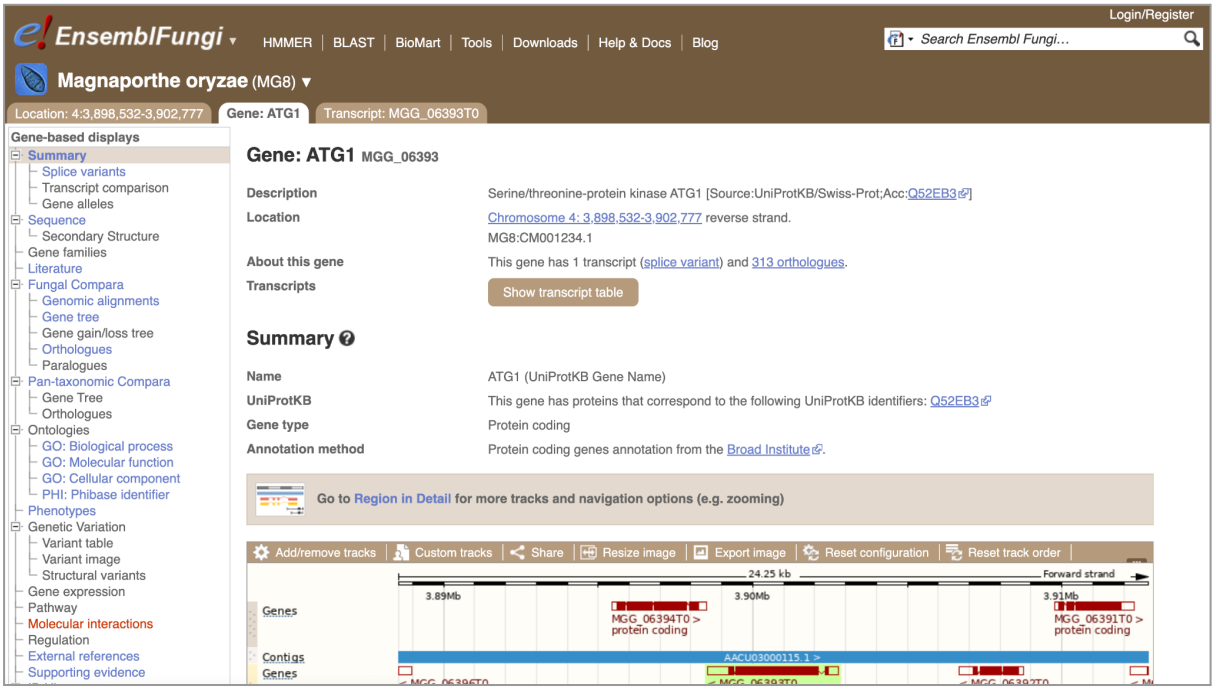
(d) To retrieve all fungal orthologues, go to Fungal Compara: Orthologues in the left-hand panel.



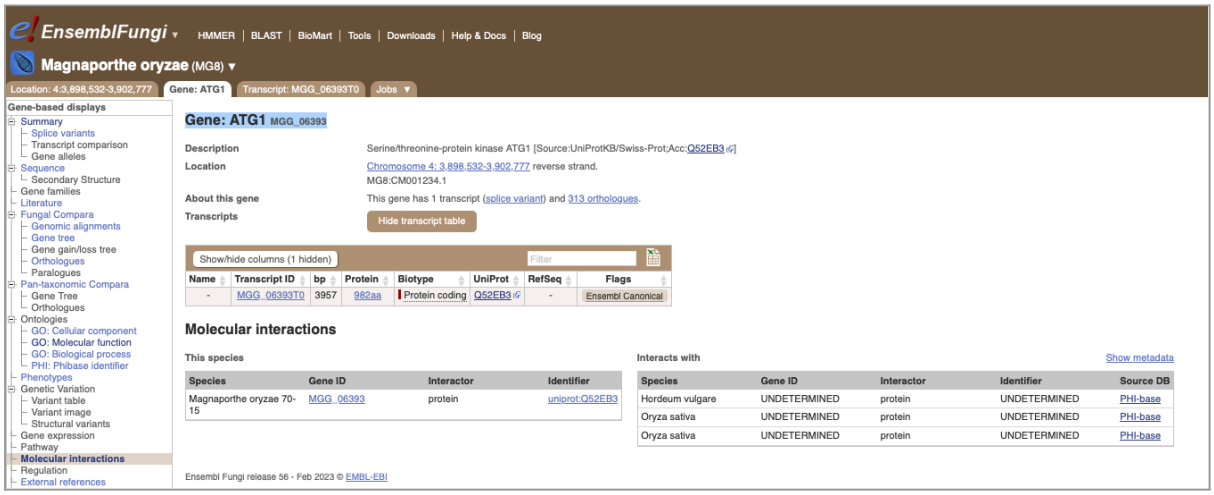
Scroll down to the Orthologues table and use the filter box in the top right-hand corner to search for *Magnaporthe oryzae*.



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 interactions information is available for the orthologue ATG1 (MGG_06393).

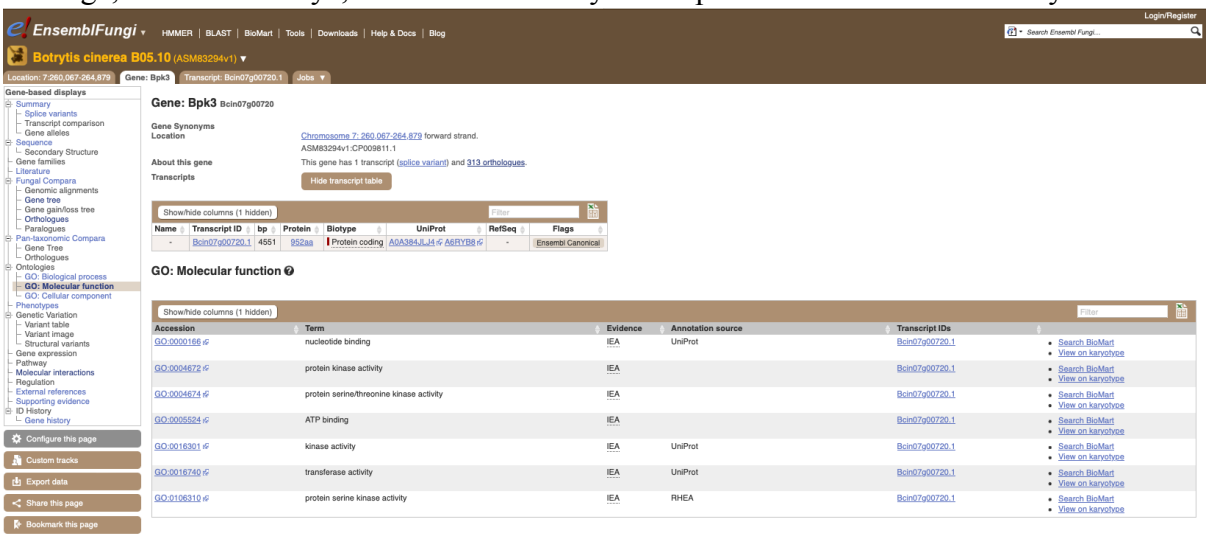


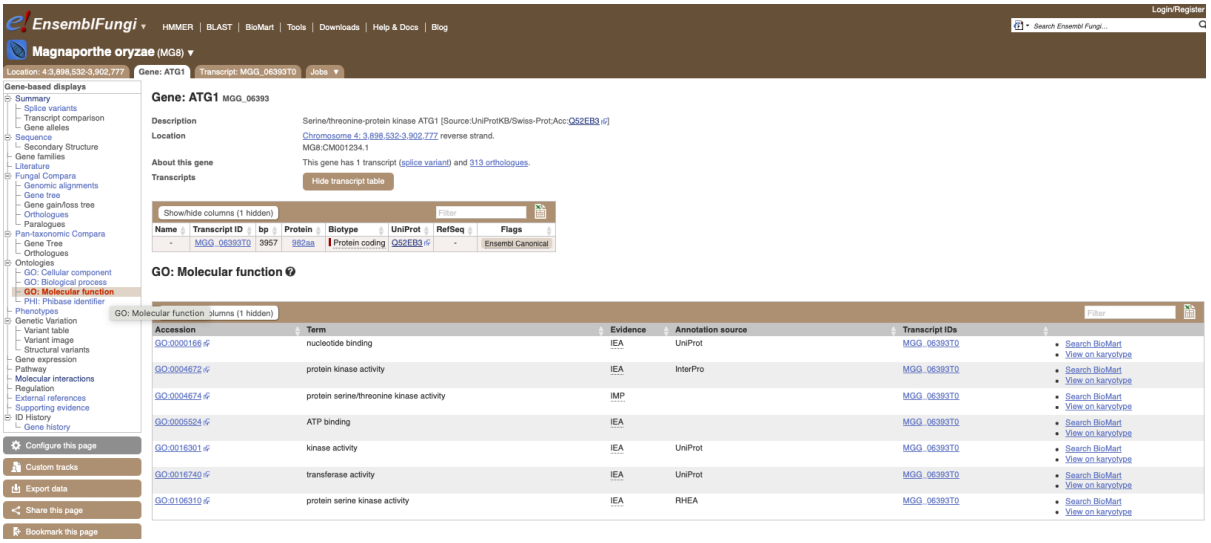
(e) Click on Molecular interactions in the left-hand panel. The ATG1 protein interacts with *Hordeum vulgare* (barley) and *Oryza sativa* (rice).



(f) Click on Show metadata to view the phenotypes associated with the molecular interactions. In *B. cinerea*, the phenotype is "Loss of pathogenicity" and in *M. oryzae* the phenotype is "Loss of pathogenicity" and "Reduced virulence".

(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/theonine kinase activity","ATP binding", "kinase activity","transferase activity" and "protein serine kinase activity".

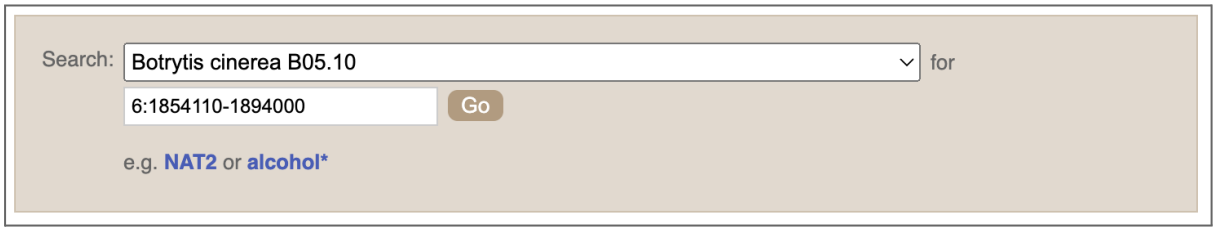



EnsemblFungi
HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog
Login/Register
Search Ensembl Fungi...
Magnaporthe oryzae (MG8)
Location: 4:3,898,532-3,902,777
Gene: ATG1
Transcript: MGG_06393T0
Jobs
Gene-based displays
Summary
Splice variants
Transcript comparison
Gene alleles
Sequence
Secondary Structure
Gene families
Literature
Fungal Compara
Genomic alignments
Gene tree
Gene gain/loss tree
Orthologues
Paralogues
Pan-taxonomic Compara
Gene Tree
Orthologues
Ontologies
GO: Cellular component
GO: Biological process
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: ATG1 MGG_06393
Description
Serine/threonine-protein kinase ATG1 [Source:UniProtKB/Swiss-Prot;Acc:Q52EB3]
Location
Chromosome 4: 3,898,532-3,902,777 reverse strand.
MG8:CM001234.1
About this gene
This gene has 1 transcript (splice variant) and 313 orthologues.
Transcripts
Hide transcript table
Show/hide columns (1 hidden)
Filter
Name
Transcript ID
bp
Protein
Biotype
UniProt
RefSeq
Flags
MGG_06393T0
3957
982aa
Protein coding
Q52EB3
Ensembl Canonical
GO: Molecular function
GO: Molecular function olumns (1 hidden)
Filter
Accession
Term
Evidence
Annotation source
Transcript IDs
GO:0000166
nucleotide binding
IEA
UniProt
MGG_06393T0
Search BioMart
View on karyotype
GO:0004672
protein kinase activity
IEA
InterPro
MGG_06393T0
Search BioMart
View on karyotype
GO:0004674
protein serine/threonine kinase activity
IMP
MGG_06393T0
Search BioMart
View on karyotype
GO:0005524
ATP binding
IEA
MGG_06393T0
Search BioMart
View on karyotype
GO:0016301
kinase activity
IEA
UniProt
MGG_06393T0
Search BioMart
View on karyotype
GO:0016740
transferase activity
IEA
UniProt
MGG_06393T0
Search BioMart
View on karyotype
GO:0106310
protein serine kinase activity
IEA
RHEA
MGG_06393T0
Search BioMart
View on karyotype

Exercise: Attaching Track Hubs

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

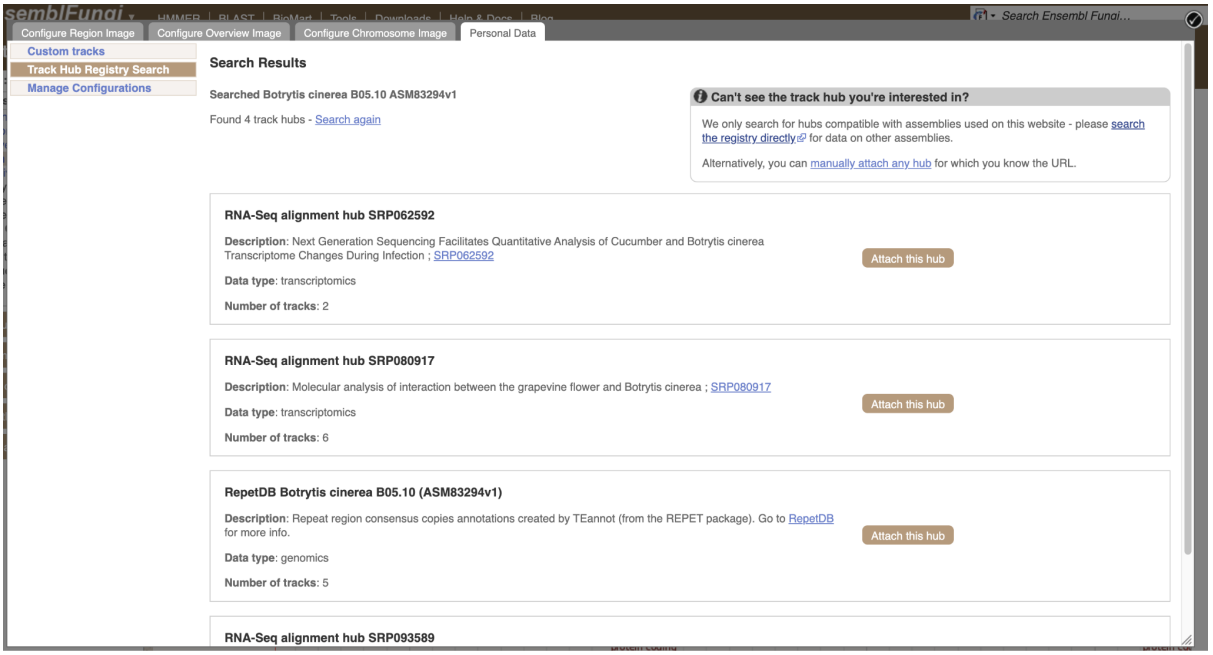
Go to fungi.ensembl.org and search for the region 6:1854110-1894000 in the species *Botrytis cinerea* B05.10.



This will take you directly to the Region in Detail page in the location tab. Click on the

button, found just below the Configure this page button on the left. A pop-up will appear, click on Track Hub Registry Search on the left-hand navigation panel.

Just click Search with no options selected.



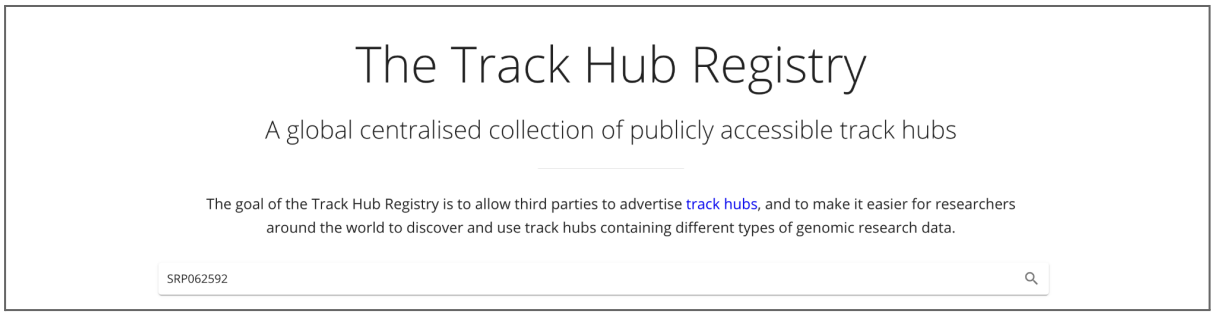
There are four available TrackHubs for this assembly.

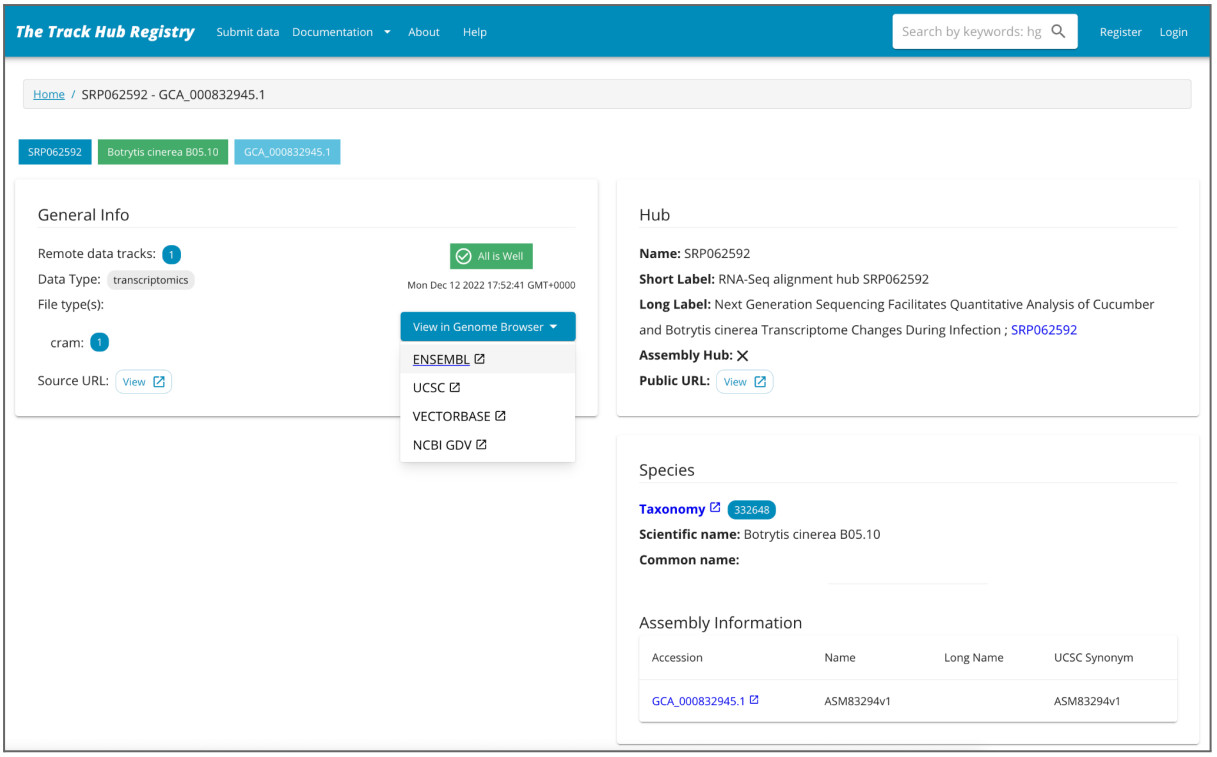
Choose the **RNA-Seq alignment hub SRP062592** by clicking on the 'Attach this hub' button. It is a next generation sequencing quantitative analysis of cucumber and *Botrytis cinerea* transcriptome changes during Infection. Close the pop-up window.

The TrackHub 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 and search for SRP062592. Can you jump to Ensembl directly from the Track Hub Registry page?

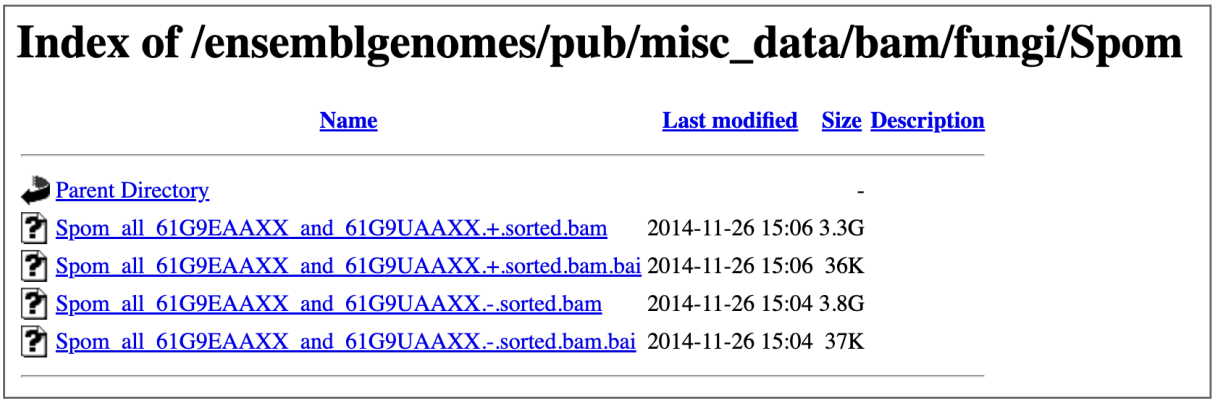




If you have your own files, or know a file you want to attach that is not present on the TrackHub registry, you can also attach these. There are two ways to do this, either by URL or by file upload.

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/

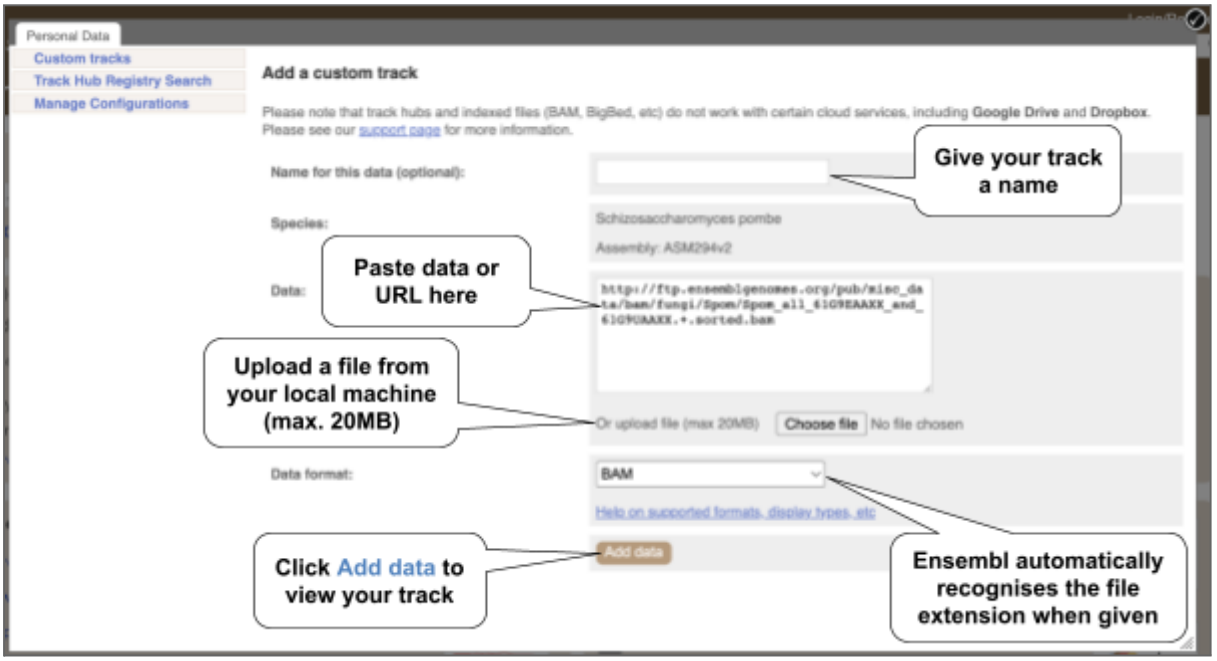
Let's take a look at that URL.
NOTE: Many internet browsers have recently dropped support for FTP, including the latest Firefox and Google Chrome versions. Firefox v87.0 still contains built-in FTP implementation. If you struggle to open the FTP site, try the HTTP version: https://ftp.ebi.ac.uk/ensemblgenomes/pub/misc_data/bam/fungi/Spom/



Here you can see two BAM files (.bam) with corresponding index files (.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 Genomes, there must be an index file in the same folder.

From the Ensembl Fungi homepage, click on *Schizosaccharomyces pombe*, then on Display your data in Ensembl Fungi.

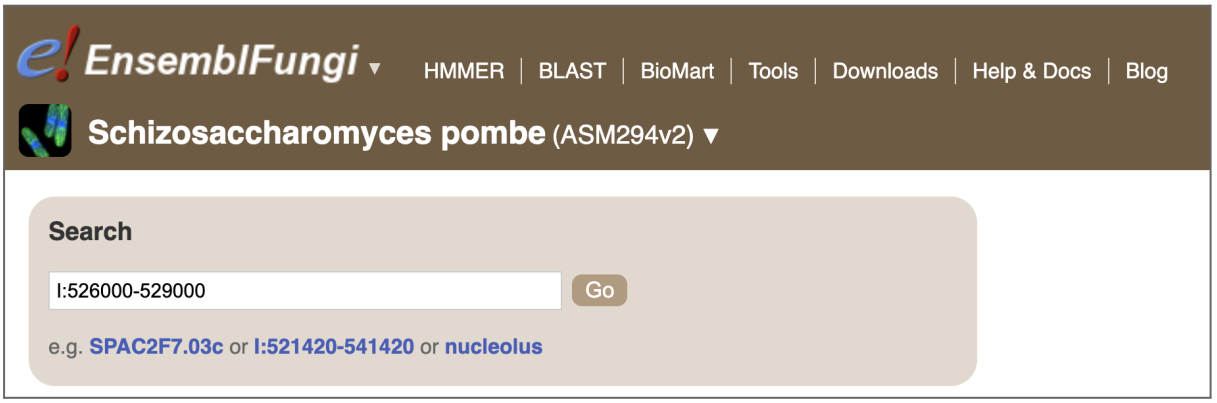
A menu will appear:

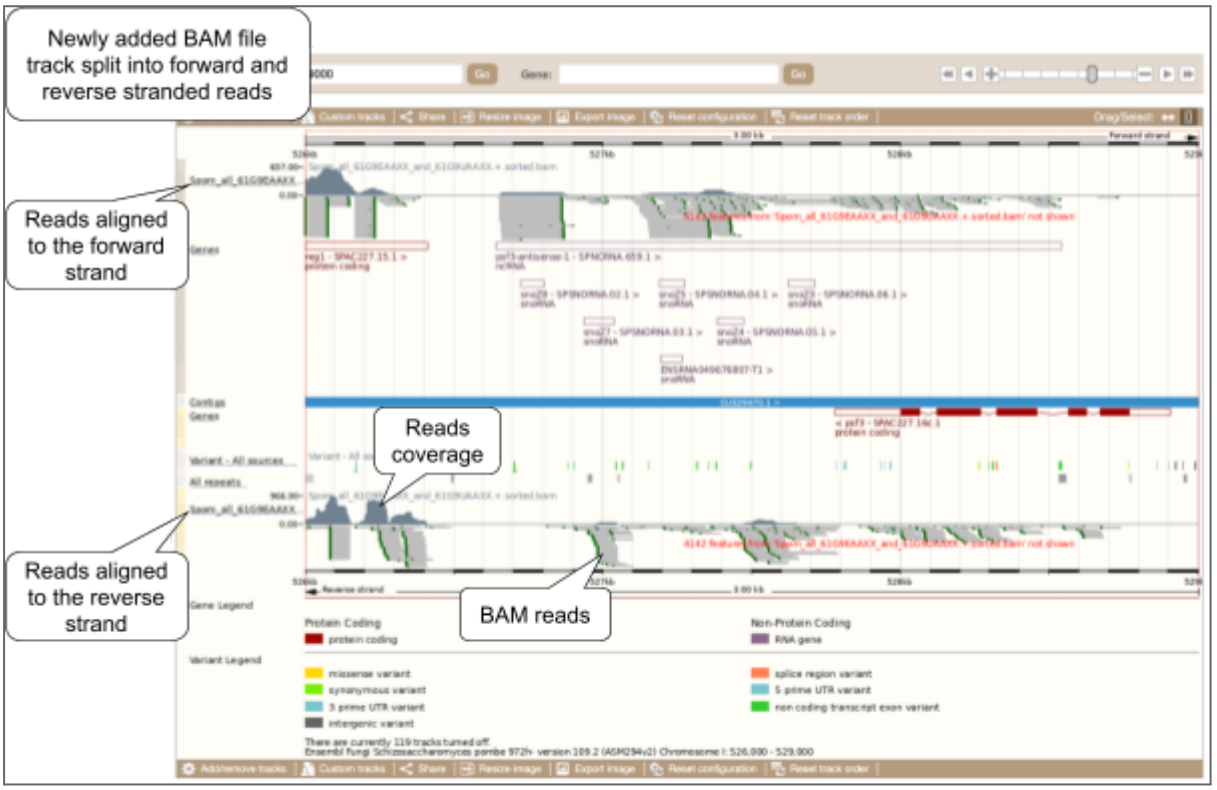


The interface detects file extensions if you upload or attach a file. If you want to upload a file just click on Choose file, choose the file and it should automatically detect the file type you have submitted.

If you have a URL, like the one we located earlier, paste in the URL of the BAM file itself (ftp://ftp.ensemblgenomes.org/pub/misc_data/bam/fungi/Spom/Spom_all_61G9EAAXX_and_61G9UAAXX.+.sorted.bam).

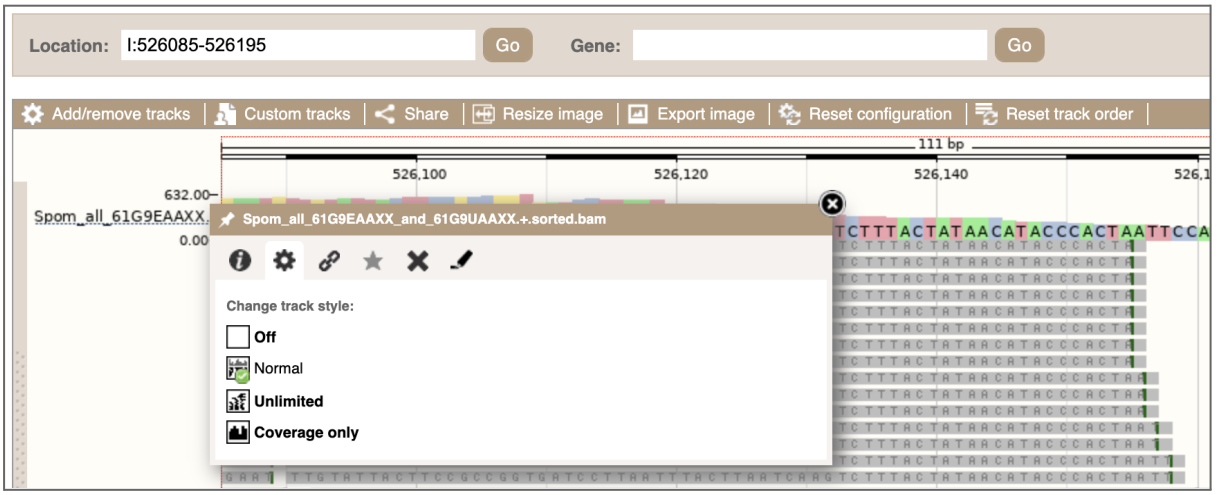
Since this is a file, the interface is able to detect the ".BAM" file extension and automatically labels the format as BAM. Click on Add data and close the menu. It may take a while to load as there is a lot of data (Firefox tends to be fast). Once the data has been uploaded, you'll get a thank you message.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

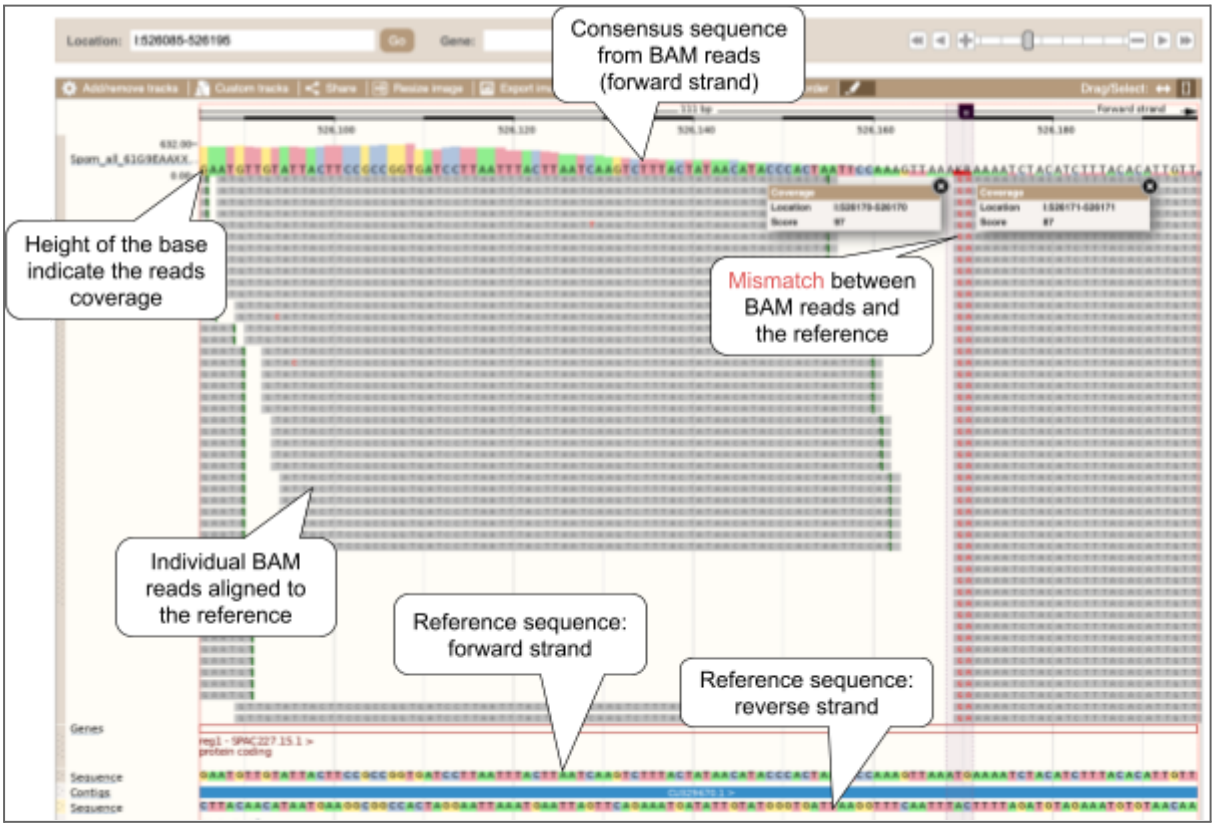




You can zoom in to see the sequence itself. Drag out boxes in the view to zoom in, until you see a sequence of individual reads, or jump to a 110 bp region: I:526085-526195.

(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?





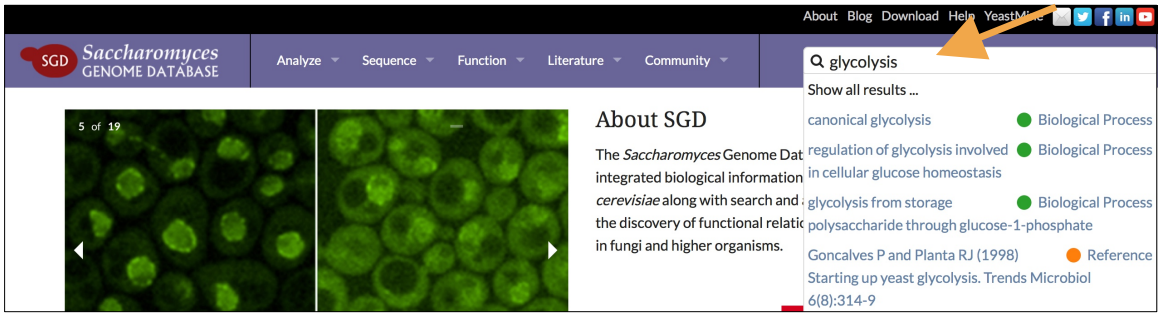
Using SPELL to Analyze Expression Datasets & Coexpressed Genes at SGD

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.

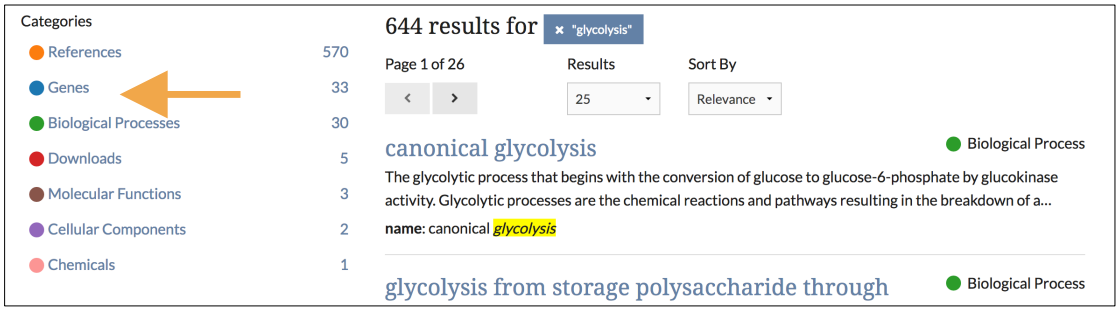
Use SPELL to find out which genes are coexpressed with genes involved in glycolysis.

Compile a list of genes involved in glycolysis.

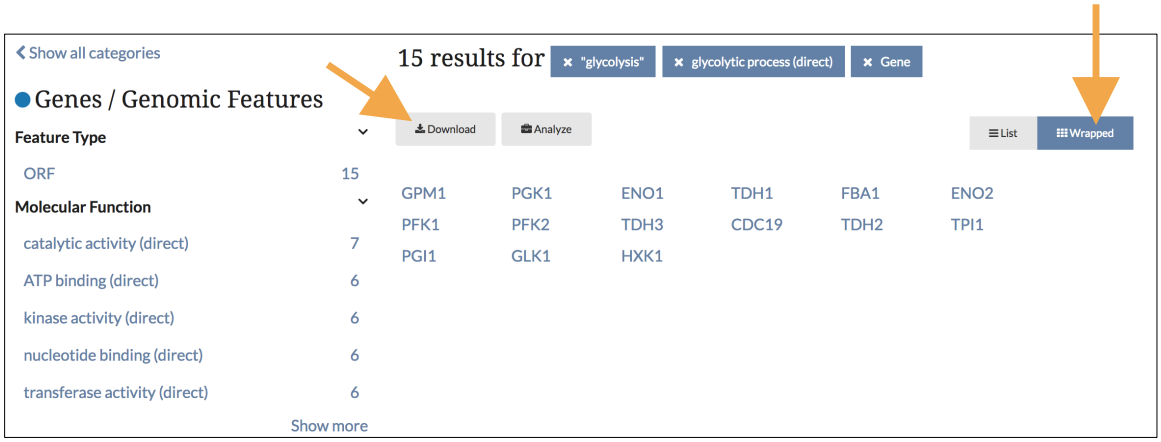
- On the SGD home page (www.yeastgenome.org), enter glycolysis into the search box and hit Enter.



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



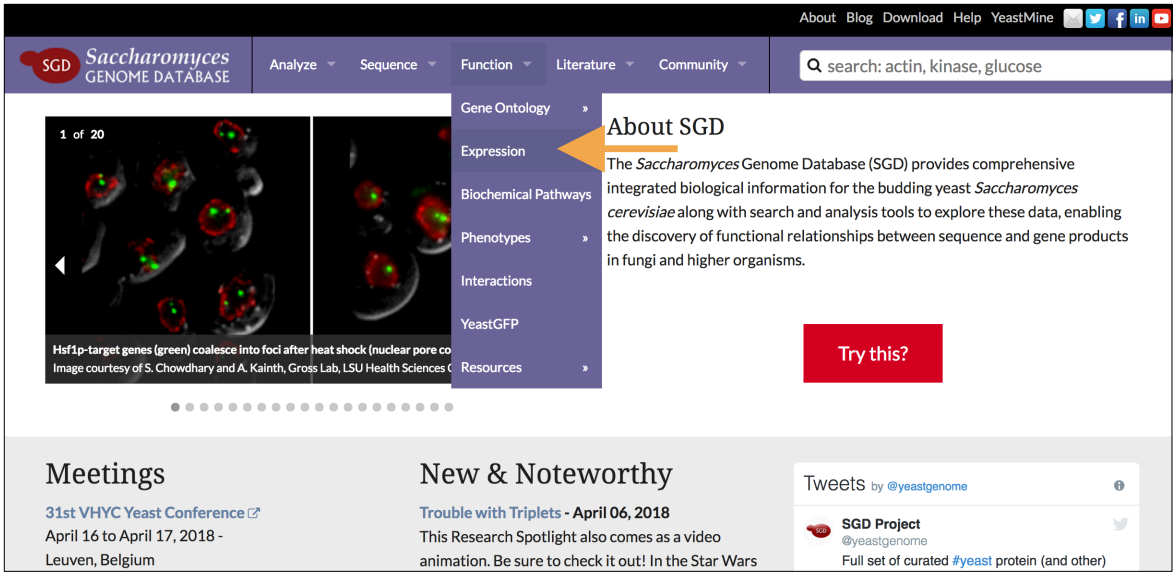
- Scroll down the page and find the **Biological Process** category on the left hand menu. Hit Show more and select **glycolytic process (direct).**
- To download the list of genes, click on **Wrapped** and then on **Download**.



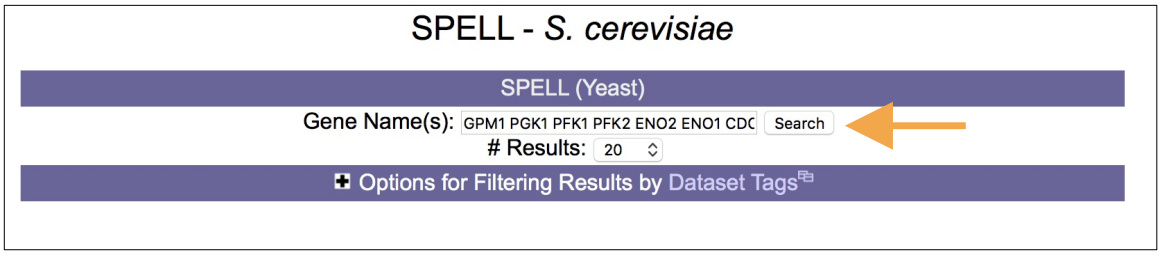
- The **Analyze** button, directly to the right of Download, enables you to import your search results directly into SPELL (among other tools at SGD). However, for the sake of demonstration, in this exercise we are instead going to enter our gene list into SPELL manually.

Import your gene list into SPELL and run a query:

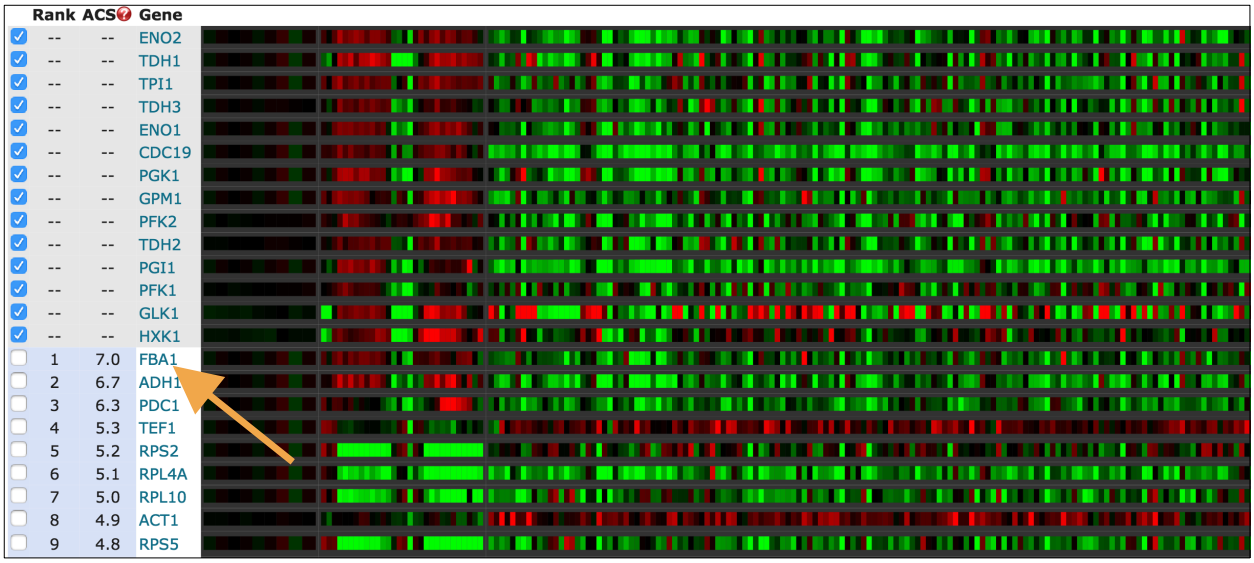
- To access SPELL, go to the SGD home page at www.yeastgenome.org, open the **Function** tab on top of the page and click on **Expression**. Or, if you are already on a Locus Summary page, open the Expression tab and click on the SPELL link under the histogram.



- 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.



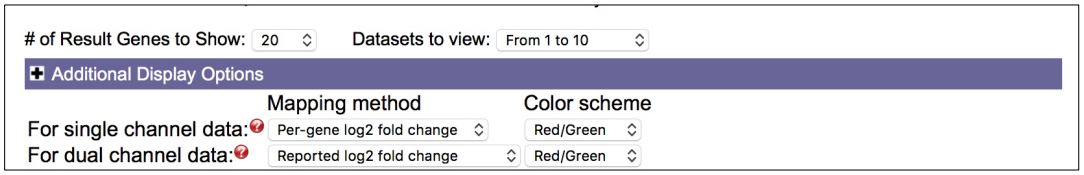
- Scroll down the list of genes on the left. Genes with checked boxes are from our query; the remaining genes are "hits", ordered from top to bottom according to their ranks. The rank reflects the correlation of expression of that gene with the query gene(s), given the relevance weight of that expression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



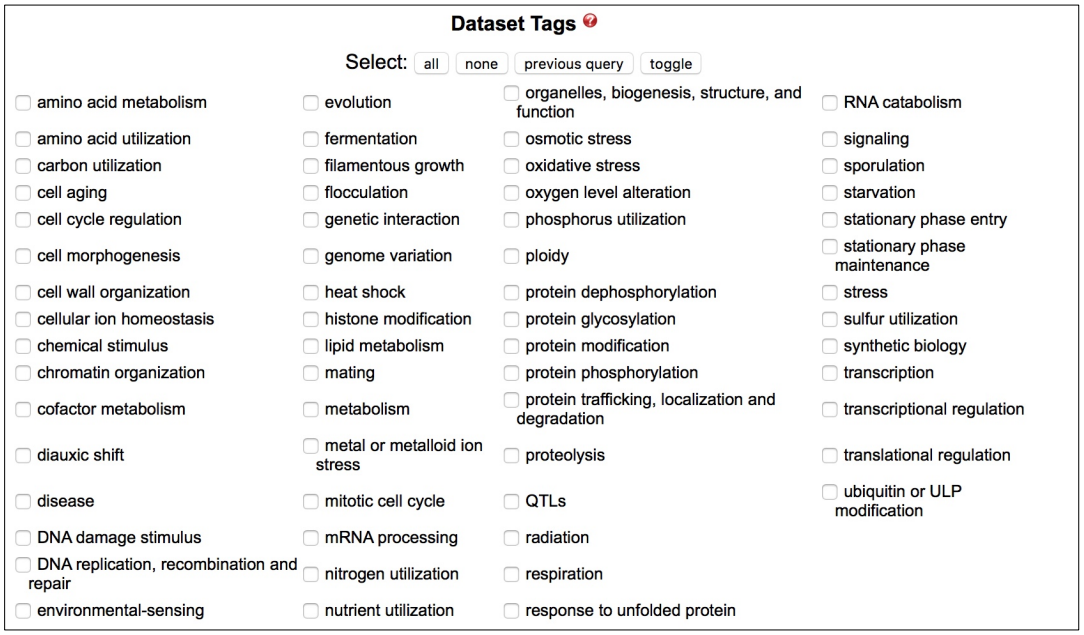
- Notice that the glycolysis gene we deleted earlier, FBA1, is indeed the highest-ranking gene!

- 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?

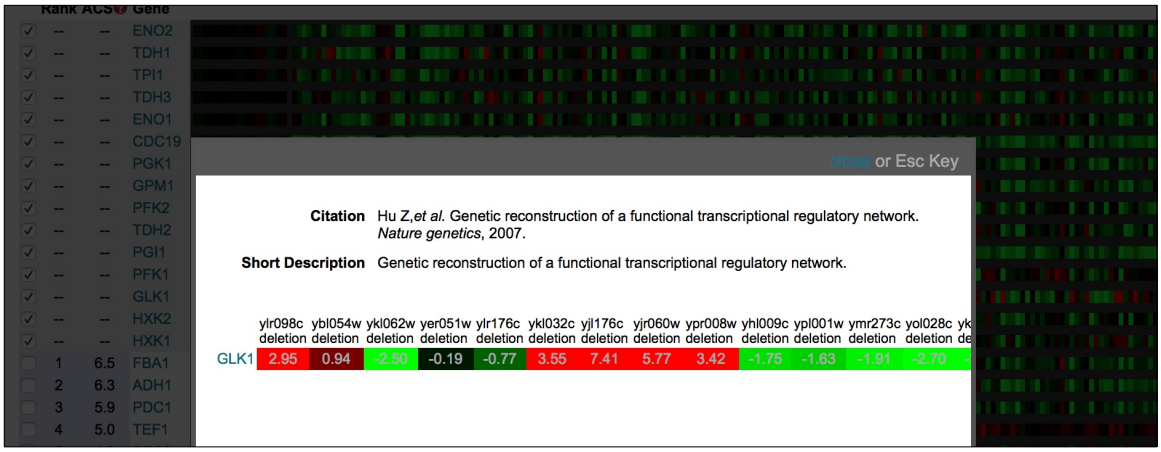
- Click on + **Additional Display Options** to change the default mapping method and color scheme to blue/yellow. Directly above this section are options to change the number of genes and datasets shown in your results.



- 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

Exploring transcriptomics & proteomics datasets in FungiDB

Transcriptomics

Learning objectives:

- Query host-pathogen RNA-Seq data in HostDB and FungiDB, respectively.
- Create a proteomics query and save this strategy to your account.

Transcriptomics datasets can be analyzed using Fold Change (FC), Differential Expression (DE), Percentile (P), and Sense/Antisense searches (SA).

Percentile (P). For each Experiment and Sample, genes are ranked by expression level (e.g., search for low/high levels of gene expression).

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 regulates, fold difference of 2, etc.)

Differential Expression (DE). This search uses DESeq2 analysis results. You can choose the directionality and the 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 cut off 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.

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 chosen Comparison samples and the chosen Reference samples.

MetaCycle. This search is applied to circadian datasets. For each study/experiment, you can choose either ARSER or JTK_Cycle method for detecting rhythmic signals. The search will return the corresponding period, amplitude and p-value.

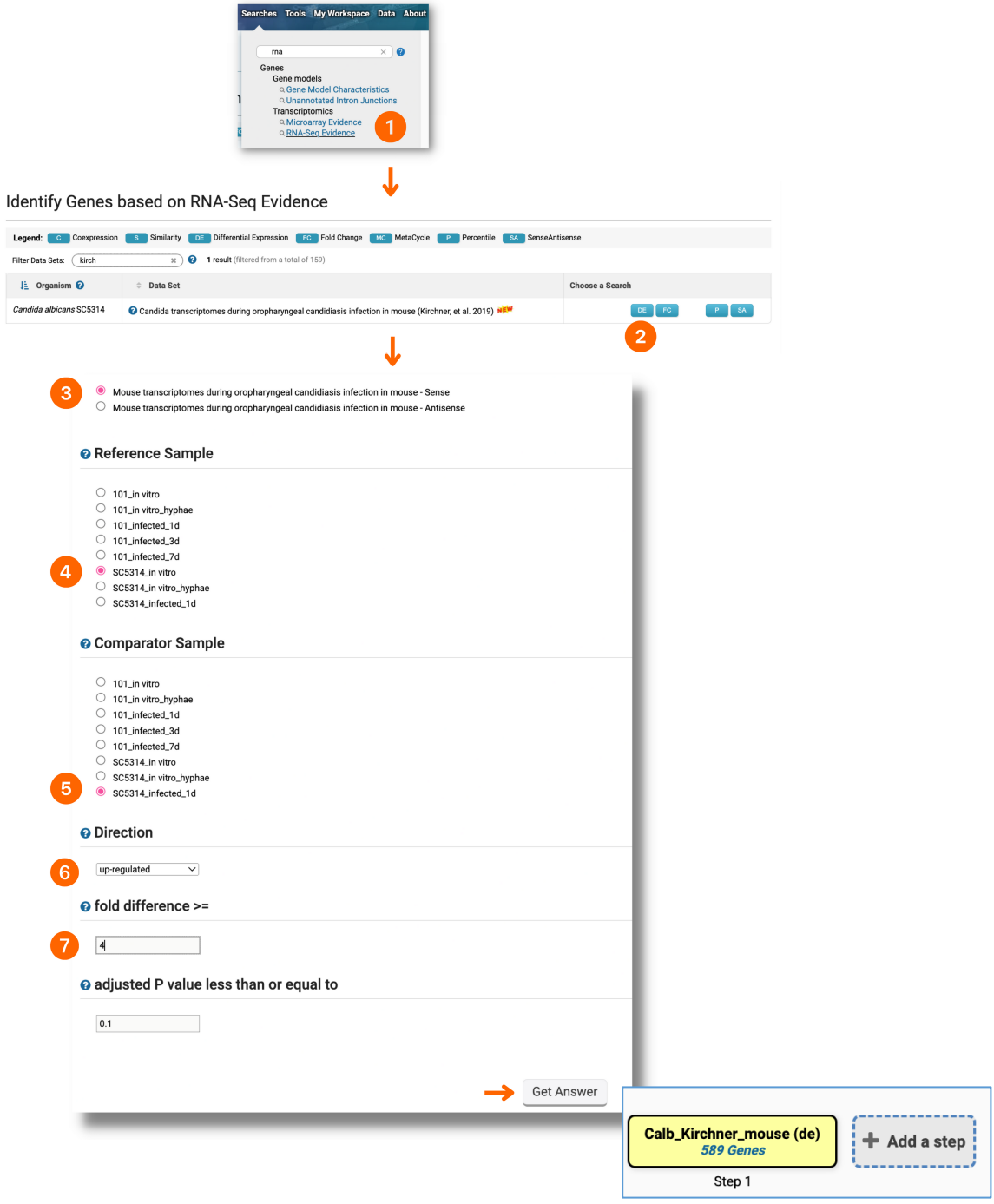
In this exercise we will query host (mouse) and pathogen (*Candida albicans*) RNA-Seq data generated by Kirchner et al. 2019. The authors used the experimental model of oropharyngeal candidiasis in mice to understand the interaction of *C. albicans* with the host at mucosal surfaces *in vivo*. Two *C. albicans* strains were used in this study – SC5314 (virulent lab strain) and the persistent strain 101.

Objectives:

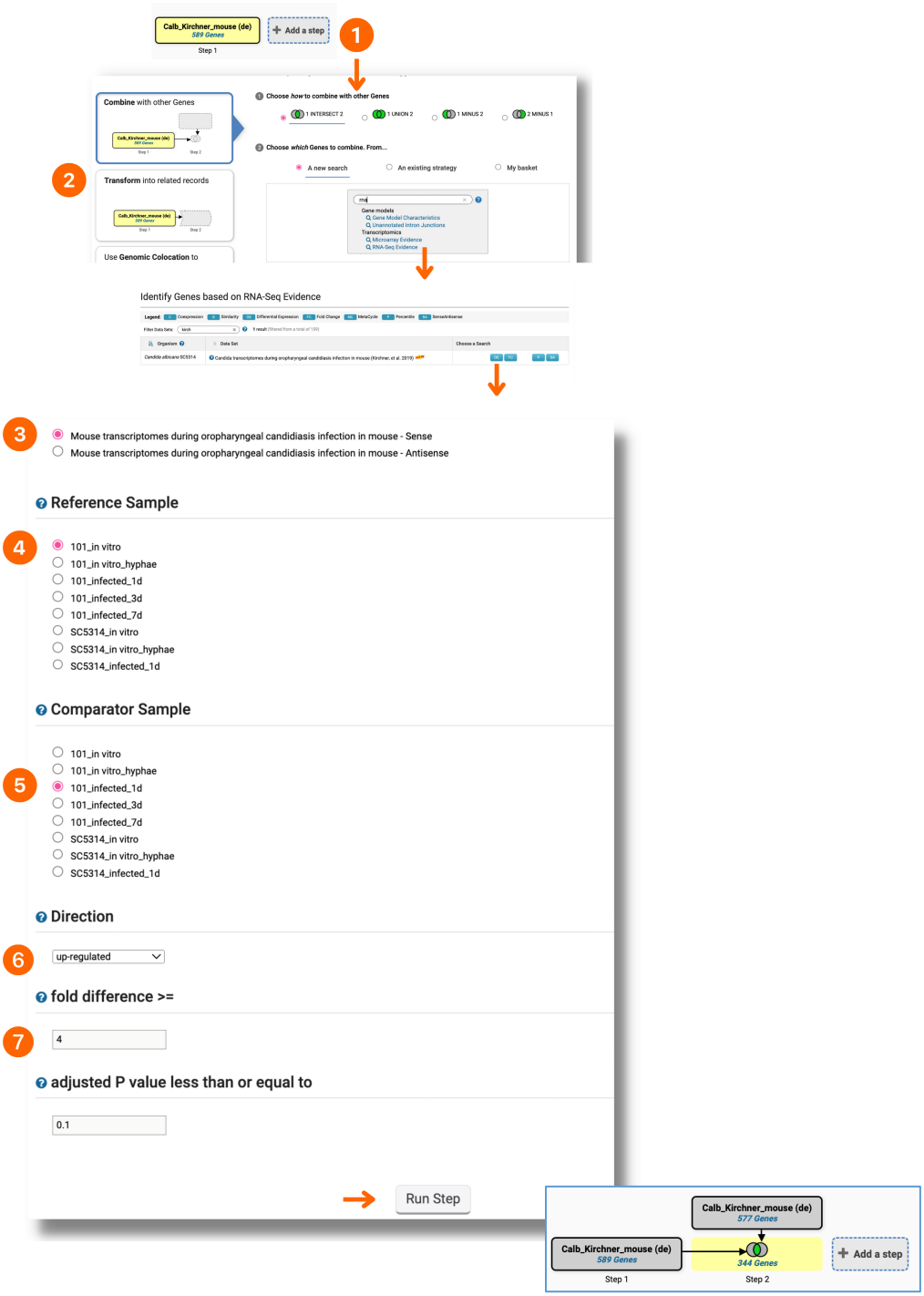
1. Identify differentially expressed genes in the virulent SC5314 strain compared to strain 101 using FungiDB.
2. Identify genes upregulated in mouse in response to the infection with SC5314 but not strain 101.

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

- **Identify genes that are 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.

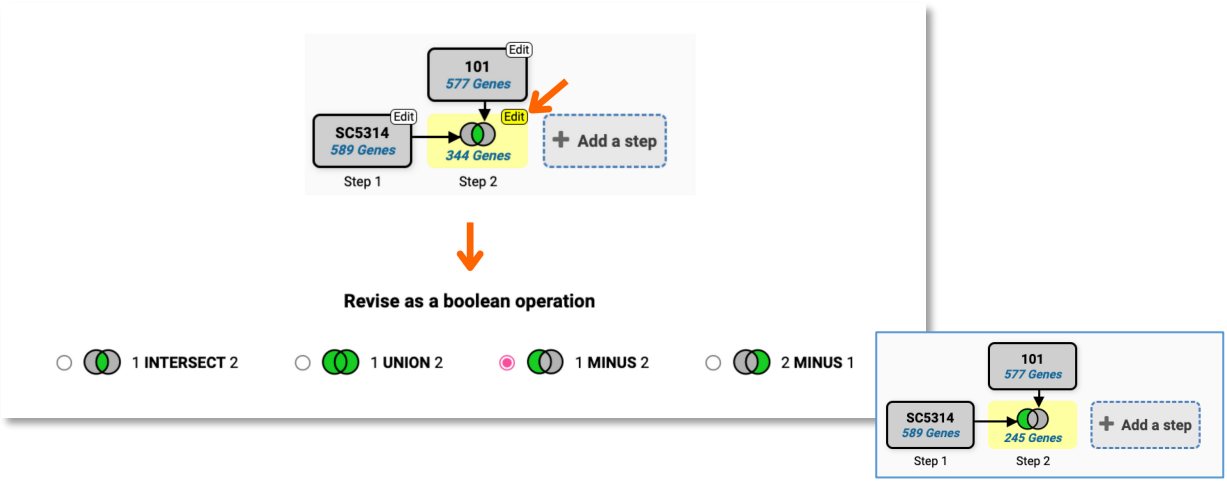


- **Identify genes that are up-regulated in SC5314 but not 101 persistent strain at 1d of infection**.
    1. Click on the “Add Step” button
    2. Navigate to the RNA-Seq Evidence search, filter for “Kirch” to quickly identify the dataset and click on the “DE” button.
    3. Choose to examine the sense 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.

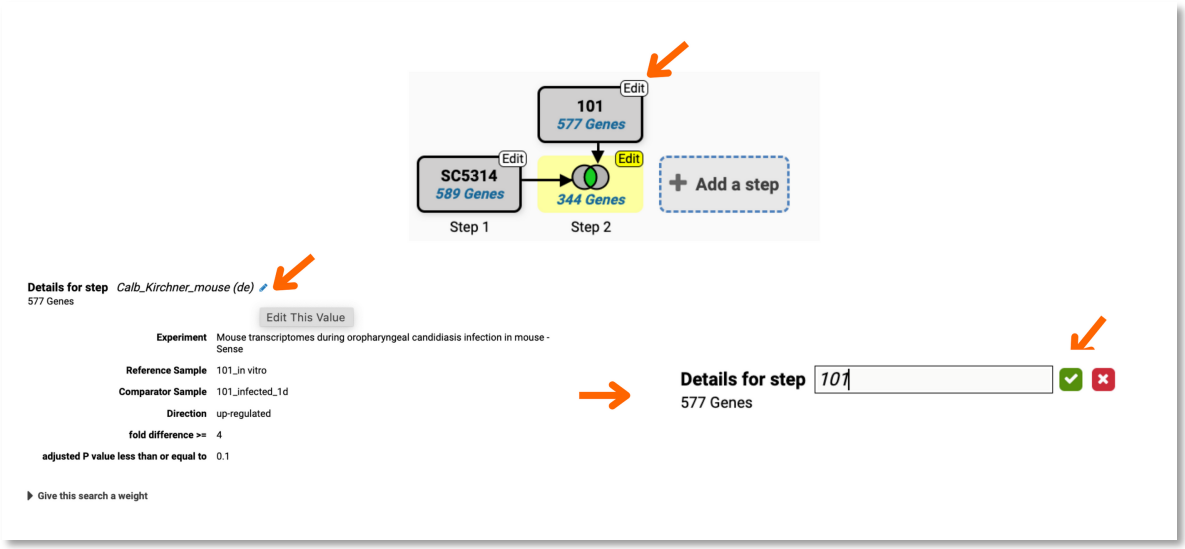


The default setting of the Boolean operators was set to the “intersect” option, which returns genes that are up-regulated by 4 fold in both strains.

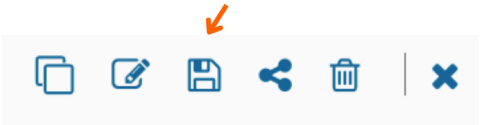
- **Change the search criteria to display genes upregulated in SC5314 only.**



Note: you can rename steps to keep track of the datasets/search results:



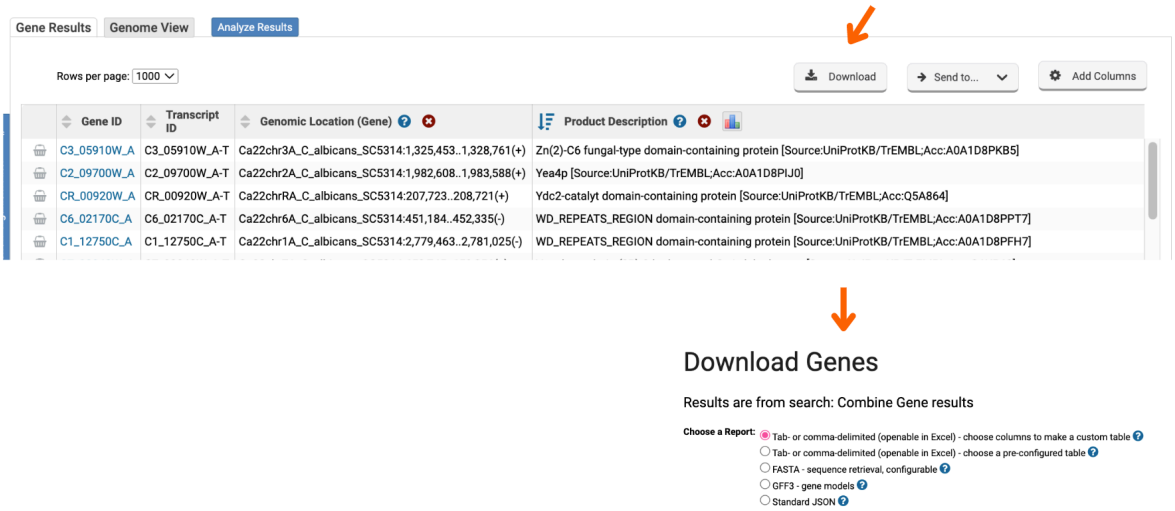
Save the strategy by clicking on the floppy disk icon on the right.



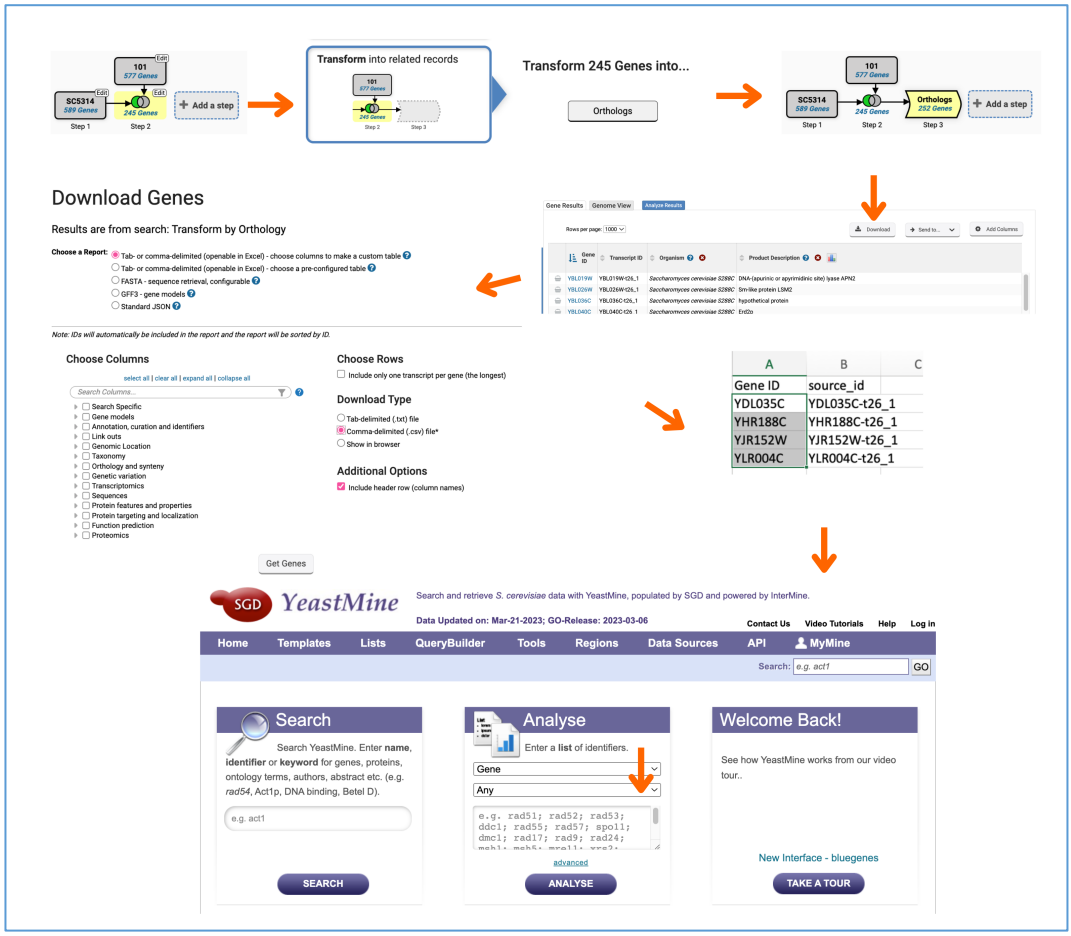
In summary, this strategy identified genes up-regulated in SC5314 when infecting mice at 1d while subtracting any genes that are also up-regulated in strain 101.

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

Note: this data can be exported and FungiDB offers several download options that can be access by clicking on the Download button located about the results table.



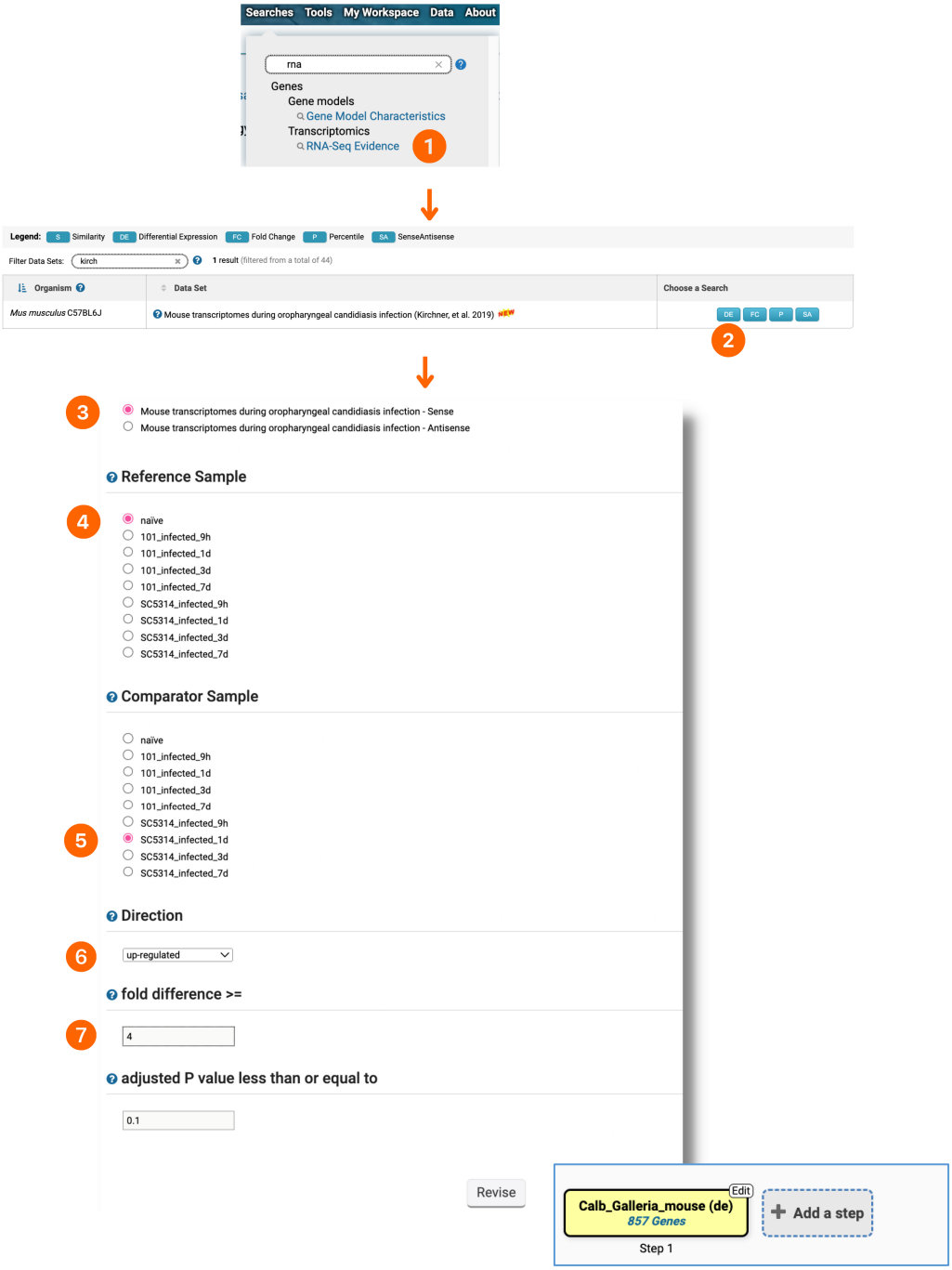
You can also export yeast orthologs of *C. albicans* genes into Yeast mine. YeastMine enables rapid retrieval and manipulation of curated biological data on yeast, which you can use to make predictions about orthologs in fungal pathogens. Here is an outline of the workflow extracting Gene IDs compatible with SGD searches:



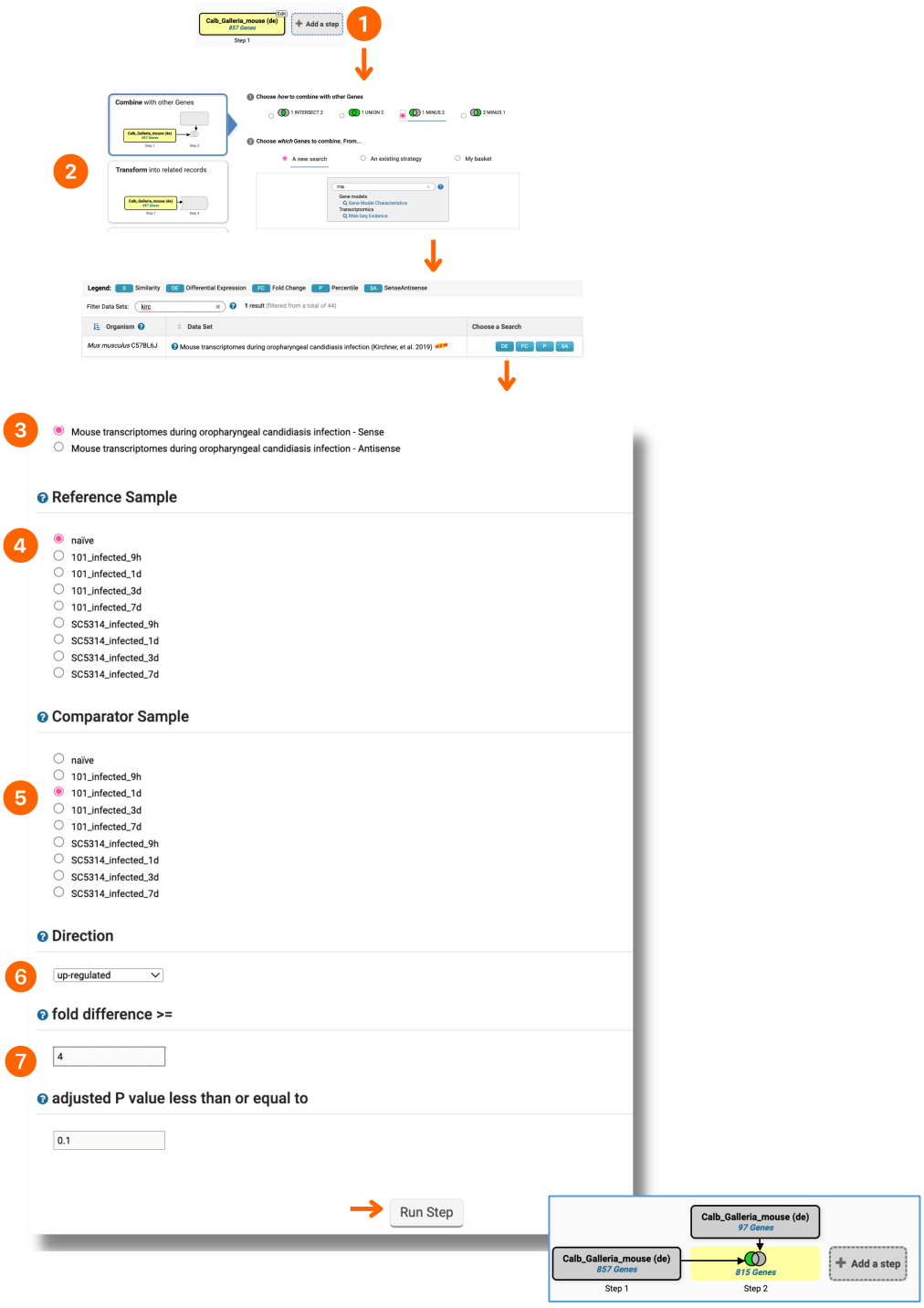
Next, we will identify gene up-regulated in mice when infected with SC5314 and 101 and select for SC5314-specific responses.

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

- **Identify genes that are 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.

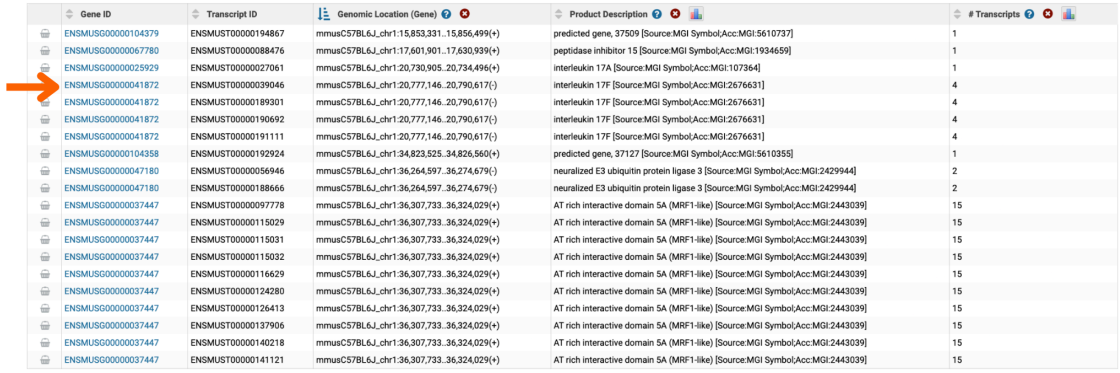


- **Identify genes that are up-regulated in SC5314 but not 101 persistent strain at 1d of infection**.
    1. Click on the “Add Step” button.
    2. Navigate to the RNA-Seq Evidence search, select “1 minus 2” Boolean operator, filter for “Kirch” to quickly identify the dataset and click on the “DE” button.
    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.



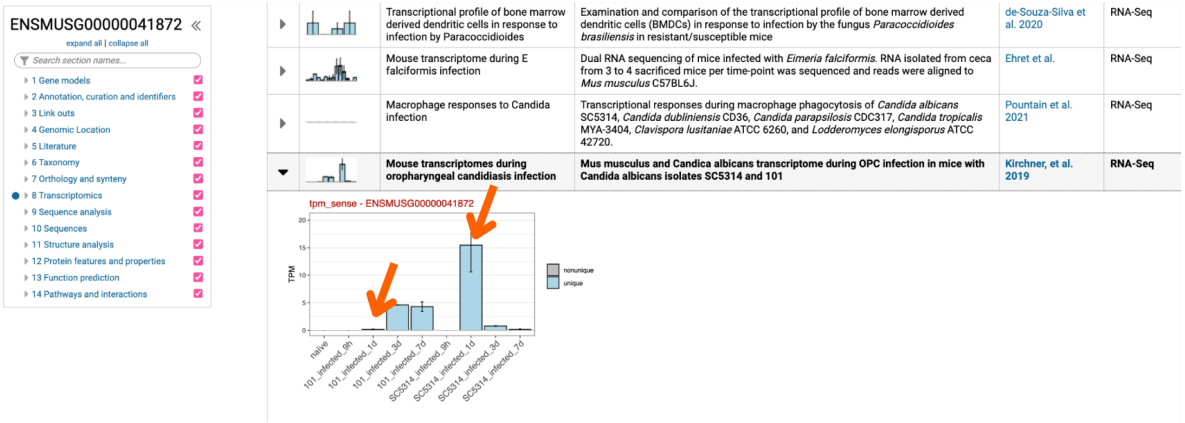
- **Examine the results in HostDB:**

    1. Click on the Gene ID link for “interleukin 17F” and navigate to the transcriptomics expression section.



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

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



Strategy URL: https://hostdb.org/hostdb/app/workspace/strategies/import/de6763c0b7f9916c

Dataset reference: Kirchner et al. 2019 DOI: 10.3389/fimmu.2019.00330

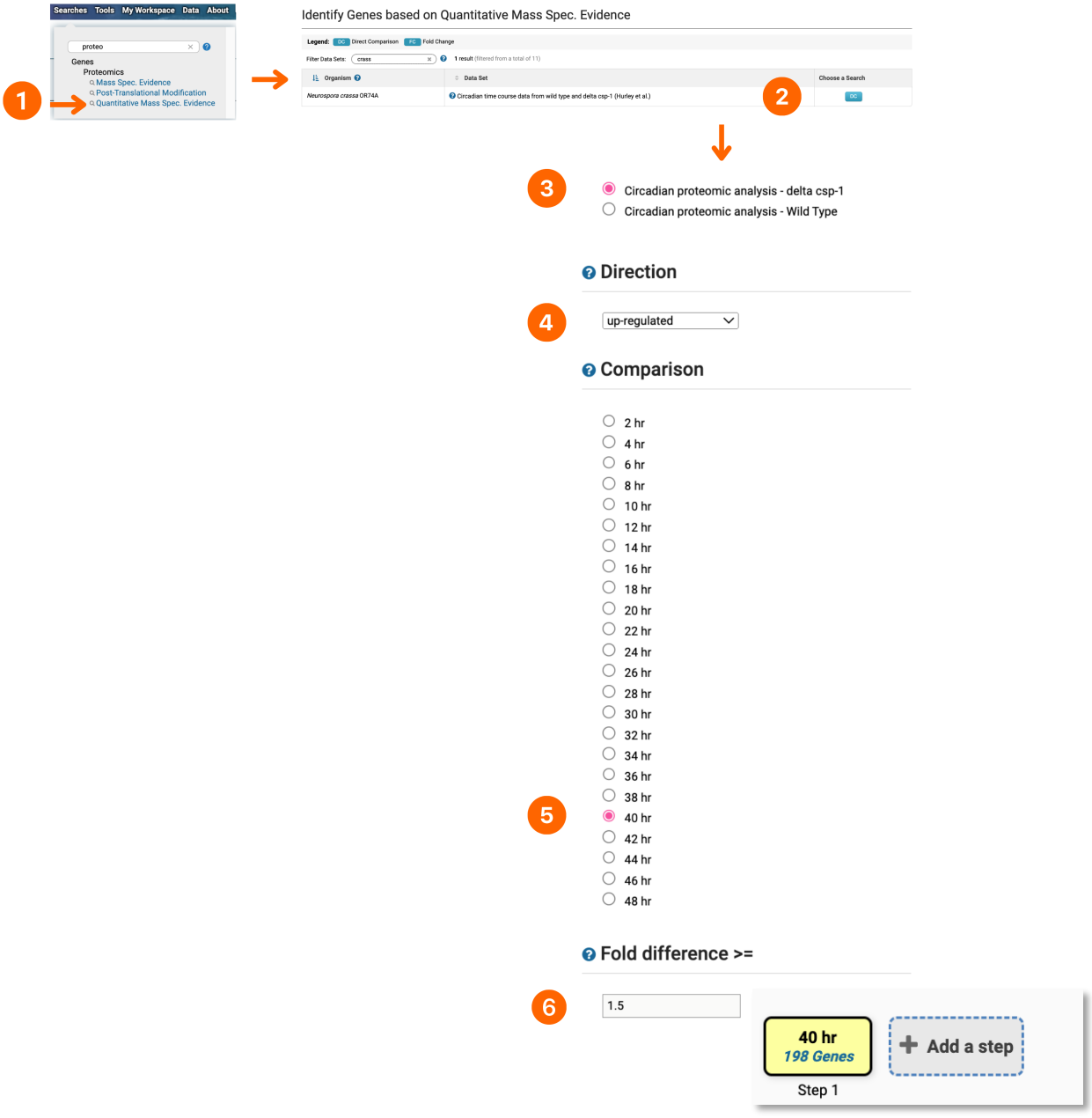
Proteomics

Learning objectives:

- Query proteomics data for *N. crassa* (e.g., genes upregulated between 40 and 46hr of incubation) and map results to *N. crassa* knockout phenotypes.

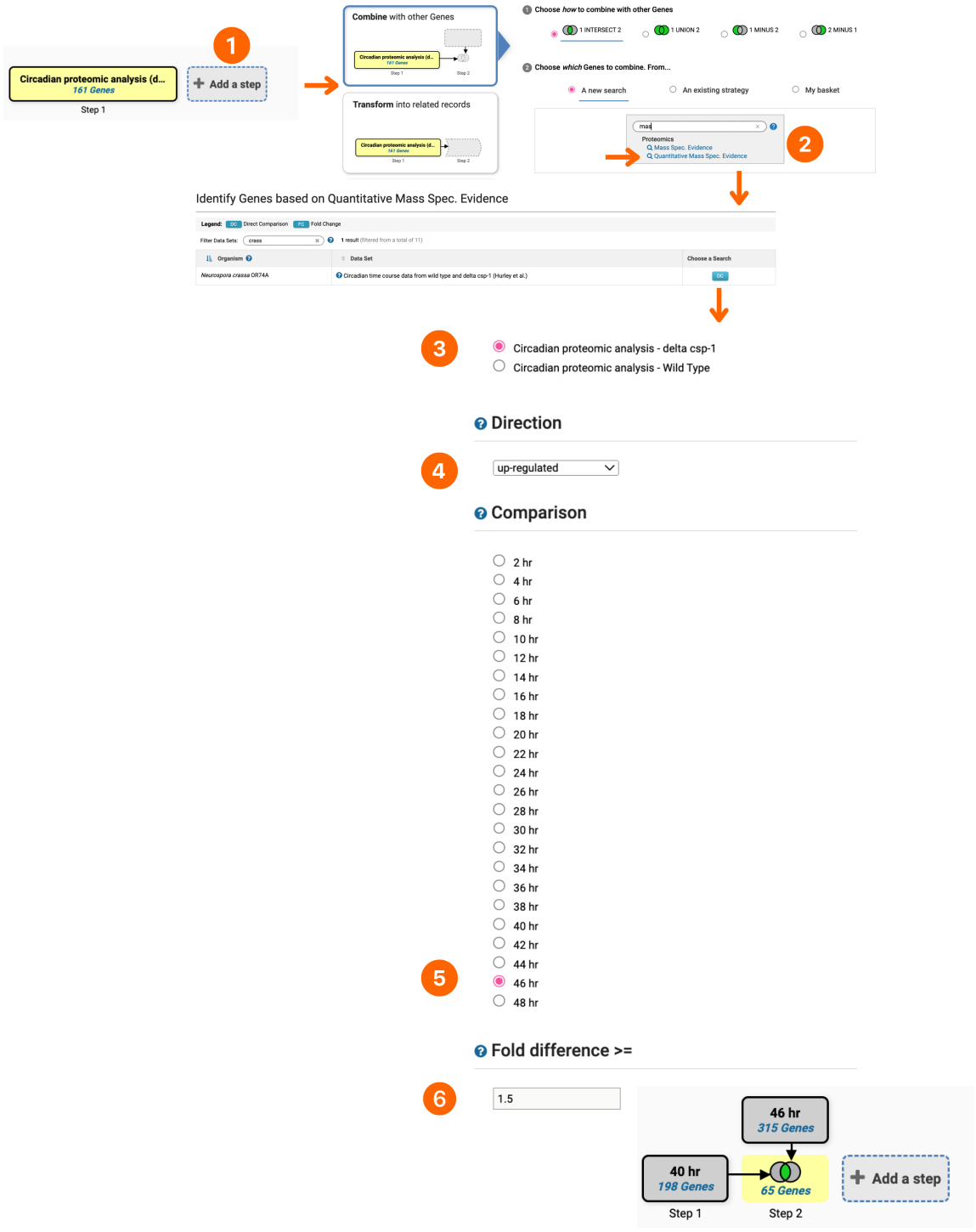
- **Identify proteins expressed in culture at 40 hr.**

1. Navigate to the “Quantitative Mass Spec. Evidence” search.
2. Click on the “DC” button for Hurley et al. 2019 dataset.
3. Select delta-csp1 mutant.
4. Choose to look for up-regulated genes.
5. Set Comparison to 40hr.
6. Leave the fold difference parameter at default.



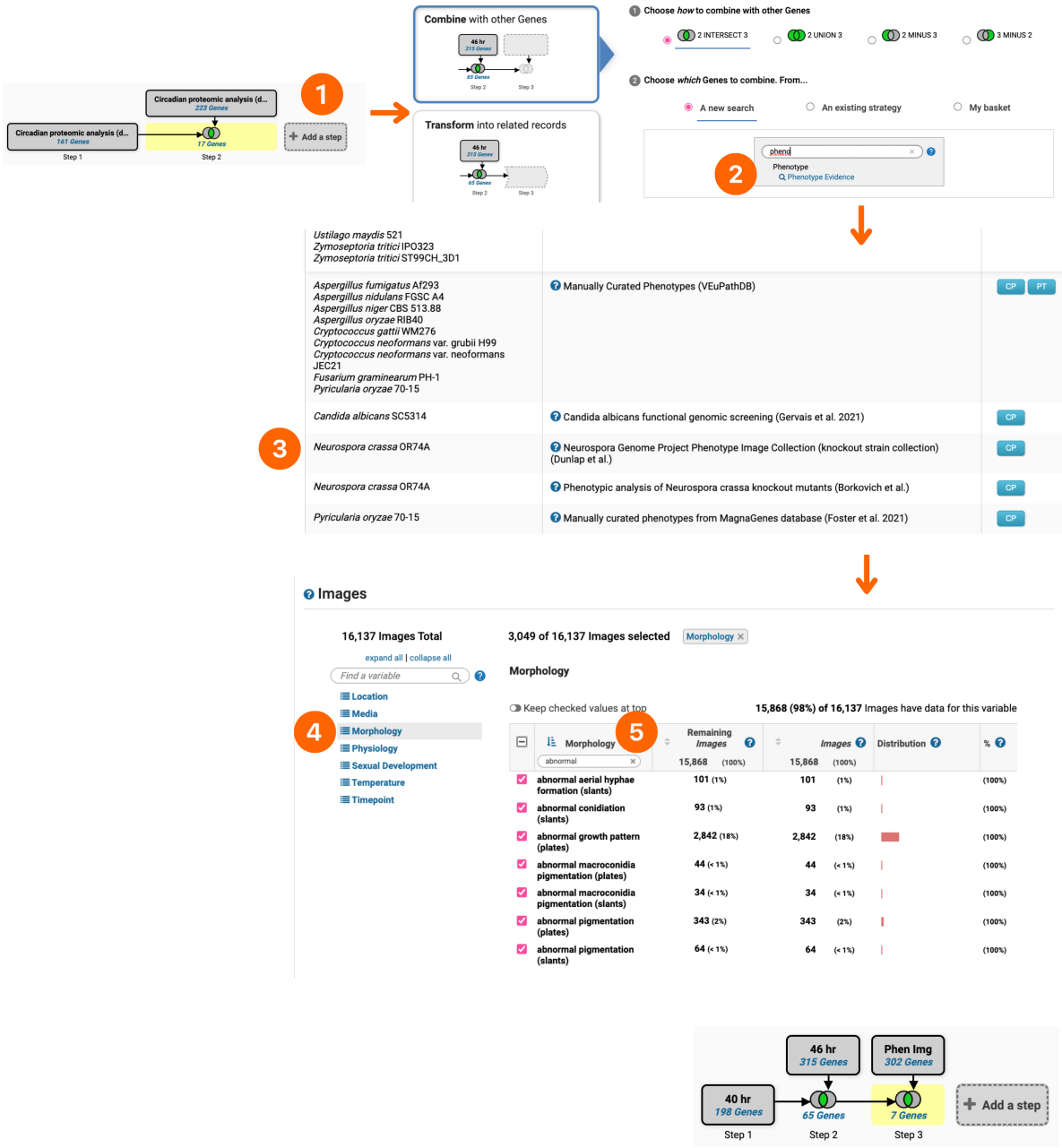
- **Identify proteins expressed in culture at 46 hr.**

1. Click on the “Add step”.
2. Select the “Combine with other Genes” search and navigate to the Quantitative Mass. Spec Evidence search.
3. Click on the “DC” button for Hurley et al. 2019 dataset.
4. Select WT sample.
5. Choose to look for up-regulated genes.
6. Set Comparison to 46hr.



- **Identify genes required for normal growth morphology in *N. crassa*.**

    1. Click on the “Add step”.
    2. Select the “Combine with other Genes” search and navigate to the Phenotype Evidence search.
    3. Click on the curated phenotypes (CP) button to investigate records from Neurospora Genome Project Phenotype Image Collection (Dunlap et al.).
    4. Navigate to the “Morphology” section.
    5. Filter on “abnormal” and select all annotated abnormal phenotypes.



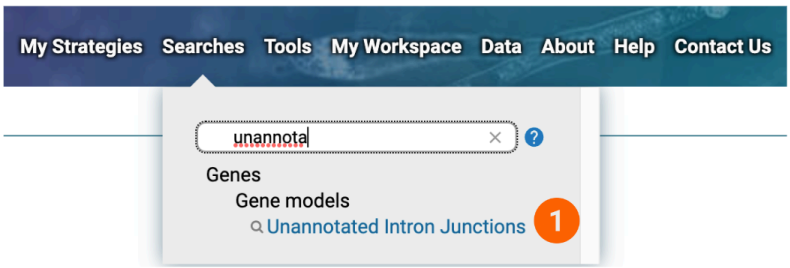
Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/0cae335cef282483
Reference: Hurley et al. 2018 DOI: 10.1016/j.cels.2018.10.014

Assessing (& editing) gene annotation (JBrowse/Apollo) (optional)

In this tutorial, we will show you how to identify possible incorrect gene structures and correct them in Apollo.

The "Unannotated Intron Junctions" search enables users to identify genes that contain, or are flanked by, unannotated high confidence intron junction-spanning reads from RNA-seq data. These genes may be incompletely or inaccurately annotated due to missing introns/exons and/or alternative splice variants. Once you've identified the genes with unannotated introns you can explore them in JBrowse and correct gene structures in Apollo, an open-source software enabling users to inspect, refine and add gene models to the current genome annotations.
Note: This search is only available for genomes with mapped RNA-Seq datasets.

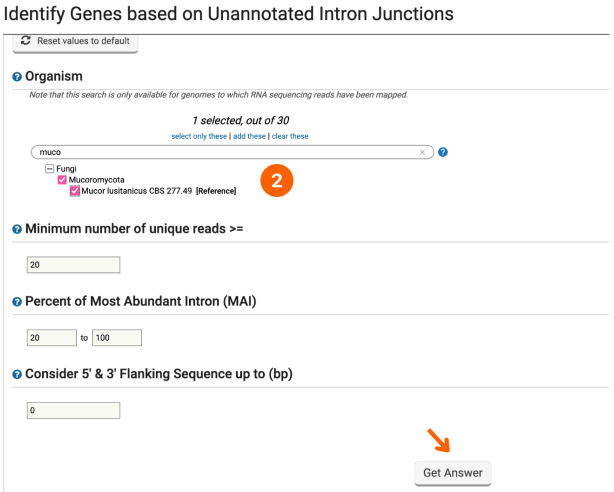
- **Identifying possible incorrect gene structures via the "Unannotated Intron Junctions "search.**
    1. **Deploy the "Unannotated Intron Junctions" search.**



    2. **Set search parameters.**

Organism: *Mucor lusitanicus* CBS 277.49
Minimum number of unique reads: keep at default.
Percent of most abundant intron (MAI): keep at default.

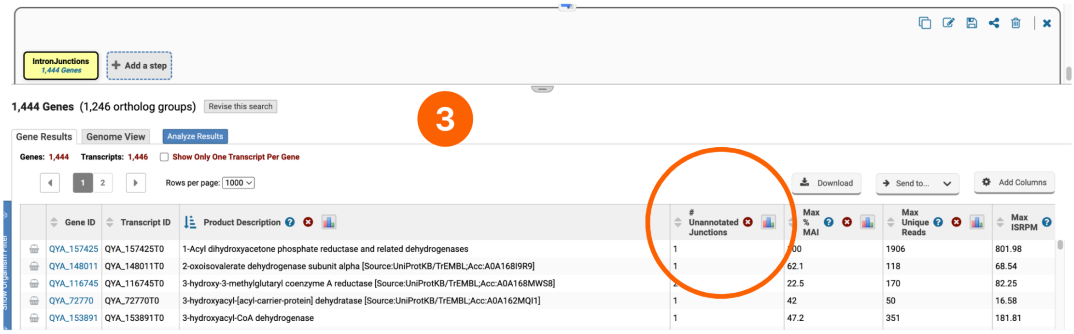
Note: The most abundant intron (MAI; supported by the highest number of intron-spanning reads; ISRs) provides context for the expected observation frequency: introns mistakenly omitted from the gene model are likely to be as abundant as correctly annotated introns.
- Consider 5' and 3' Flanking sequence up to (bp): keep at default.
Note: Here you can enter the maximum number of nucleotides flanking the annotated gene model to explore when looking for unannotated introns. Search automatically includes the annotated gene model.



Using the default parameters on this search you will get a first impression on the number of genes with unannotated introns. If you think this number is too high to explore the data, change the search parameters, the minimum number of unique reads or percent of most abundant intron.

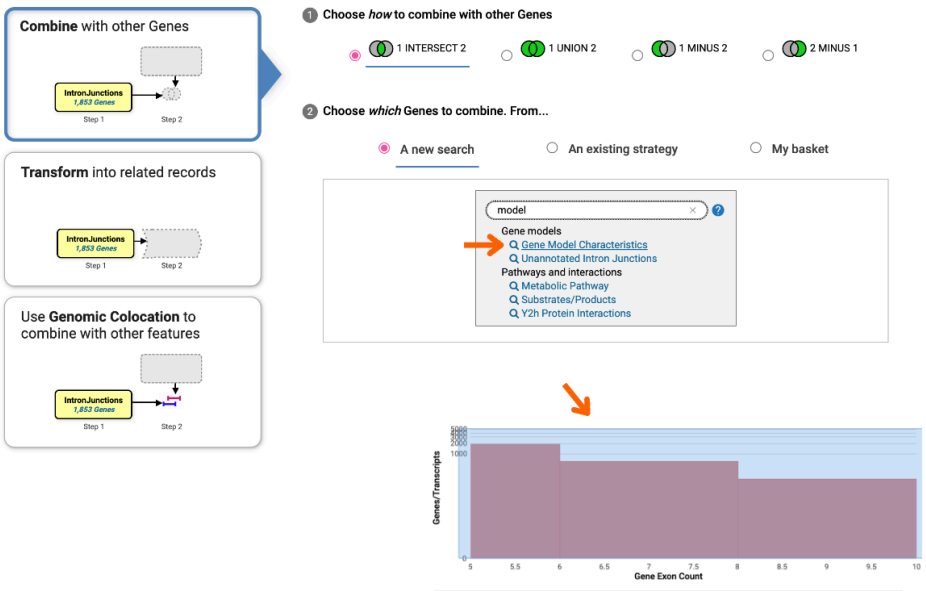
3. Explore the results table.

Note: Search results can be ordered by using the “Novel junctions” filter.

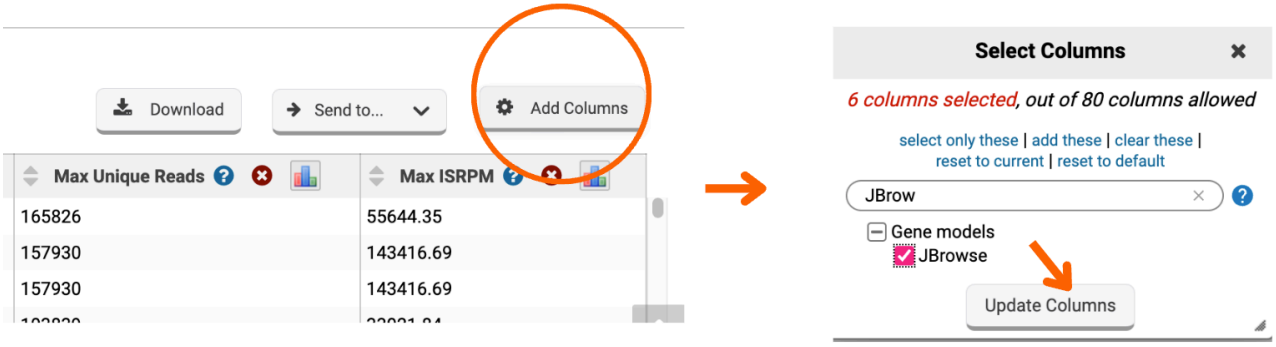


Number of Novel Splice Junctions. In case the number is 1, this means your gene has 1 possible unannotated intron. If this number is quite high, i.e. over 50 there is a possibility that your gene of interest is a rRNA, located in a repetitive region or it is part of a gene family. Therefore, it is important to explore the results in JBrowse/Apollo with additional evidence.
Max % Mai: Maximum percentage of intron with the maximum total unique reads in this gene for the novel introns that met search criteria.
Max Unique Reads: Maximum total unique reads for the novel introns that met search criteria.
Max ISRPM: Maximum total ISRPM (Intron Spanning Reads Per Million) for the novel introns.

This search can be combined with the “Gene Model Characteristics” search to limit the results on the number of exons in the gene. This may be useful if you want to look possible structural annotation errors in muti-exon genes only:



Exploring evidence in JBrowse.
Clicking on the Gene ID linked in blue will re-direct you to the gene record page where you can click on JBrowse button. However, you can also modify the results table to include direct JBrowse links for easy navigation. To do this, click on the “Add Columns” button and select JBrowse from the menu.

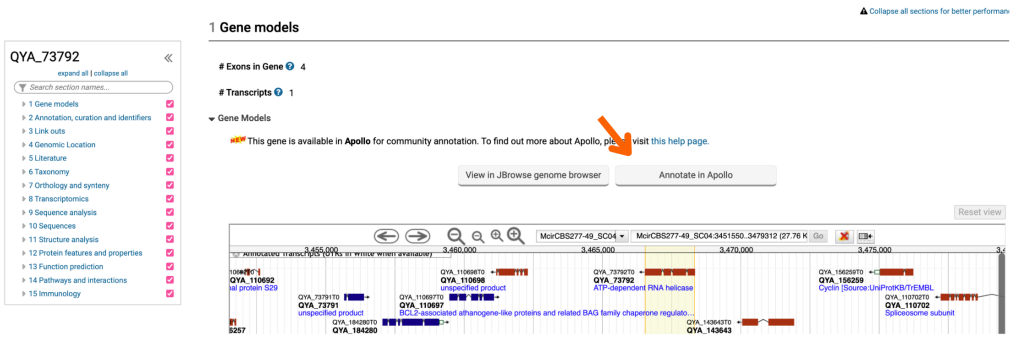


Gene ID	Transcript ID	Product Description	# Unannotated Junctions	Max % MAI	Max Unique Reads	Max ISRPM	JBrowse
QYA_157425	QYA_157425T0	1-Acyl dihydroxyacetone phosphate reductase and related dehydrogenases	1	100	1906	801.98	JBrowse
QYA_148011	QYA_148011T0	2-oxoisovalerate dehydrogenase subunit alpha [Source:UniProtKB/TrEMBL;Acc:A0A168I9R9]	1	62.1	118	68.54	JBrowse
QYA_116745	QYA_116745T0	3-hydroxy-3-methylglutaryl coenzyme A reductase [Source:UniProtKB/TrEMBL;Acc:A0A168MWS8]	2	22.5	170	82.25	JBrowse
QYA_72770	QYA_72770T0	3-hydroxyacyl-[acyl-carrier-protein] dehydratase [Source:UniProtKB/TrEMBL;Acc:A0A162MQI1]	1	42	50	16.58	JBrowse
QYA_153891	QYA_153891T0	3-hydroxyacyl-CoA dehydrogenase	1	47.2	351	181.81	JBrowse
QYA_151919	QYA_151919T0	3-hydroxyisobutyryl-CoA hydrolase, mitochondrial [Source:UniProtKB/TrEMBL;Acc:A0A168ND88]	1	23.5	28	13.63	JBrowse
QYA_157539	QYA_157539T0	40S ribosomal protein S1 [Source:UniProtKB/TrEMBL;Acc:A0A168MX07]	1	100	2155	1162.14	JBrowse

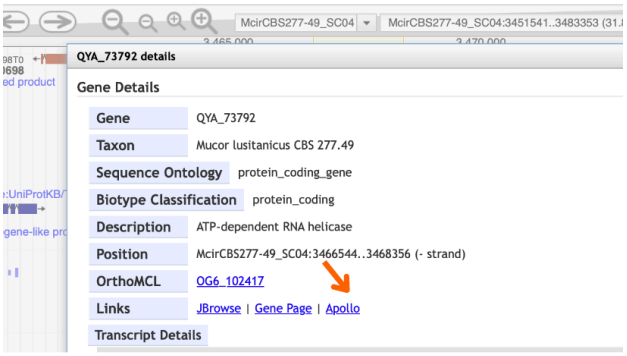
Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/d5cddb62413777fa

Correcting gene structure in Apollo.

Apollo can be accessed from gene record pages:



and also in JBrowse (left click on the gene to brin up the pop-up window):

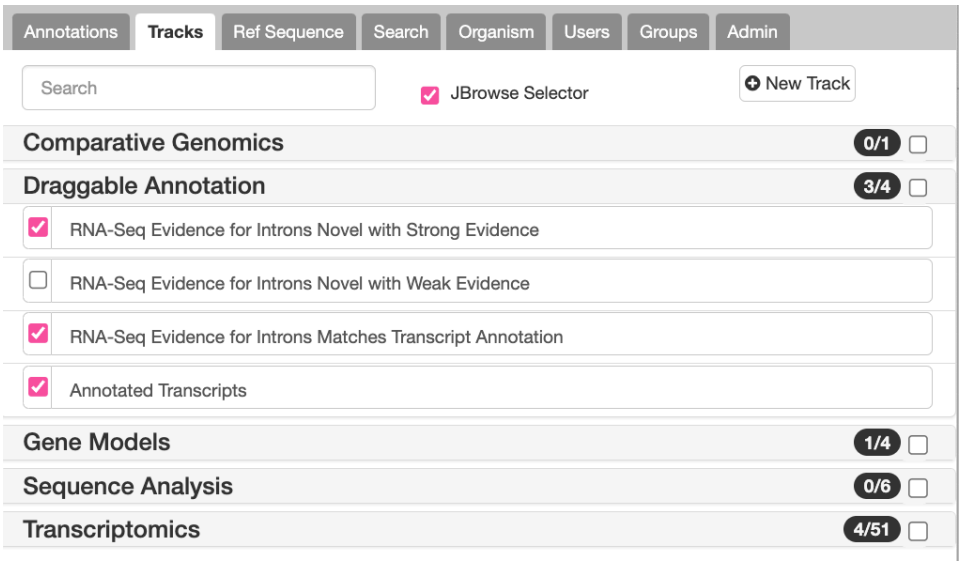


Once in Apollo, use the right panel to select the "Tracks" tab to bring up the following tracks:

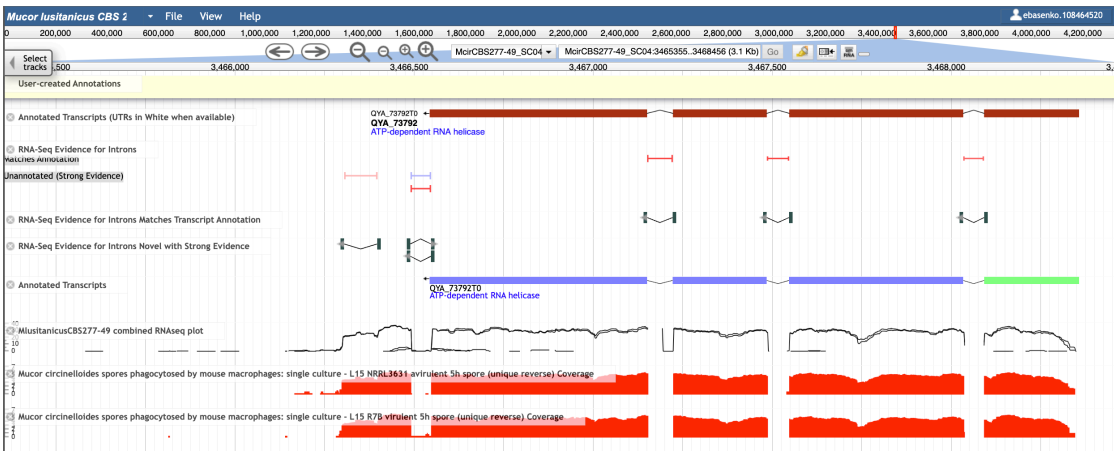
Draggable Annotation

Check the box to select the following tracks:

- RNA-Seq Evidence for Introns Novel with Strong Evidence
- RNA-Seq Evidence for Introns Matches Transcript Annotation
- Annotated transcripts



Note: You can also deploy several unique reverse RNA-Seq tracks as a guide when making changes to the structural gene annotation. The RNA-Seq tracks are available under the “transcriptomics” menu.



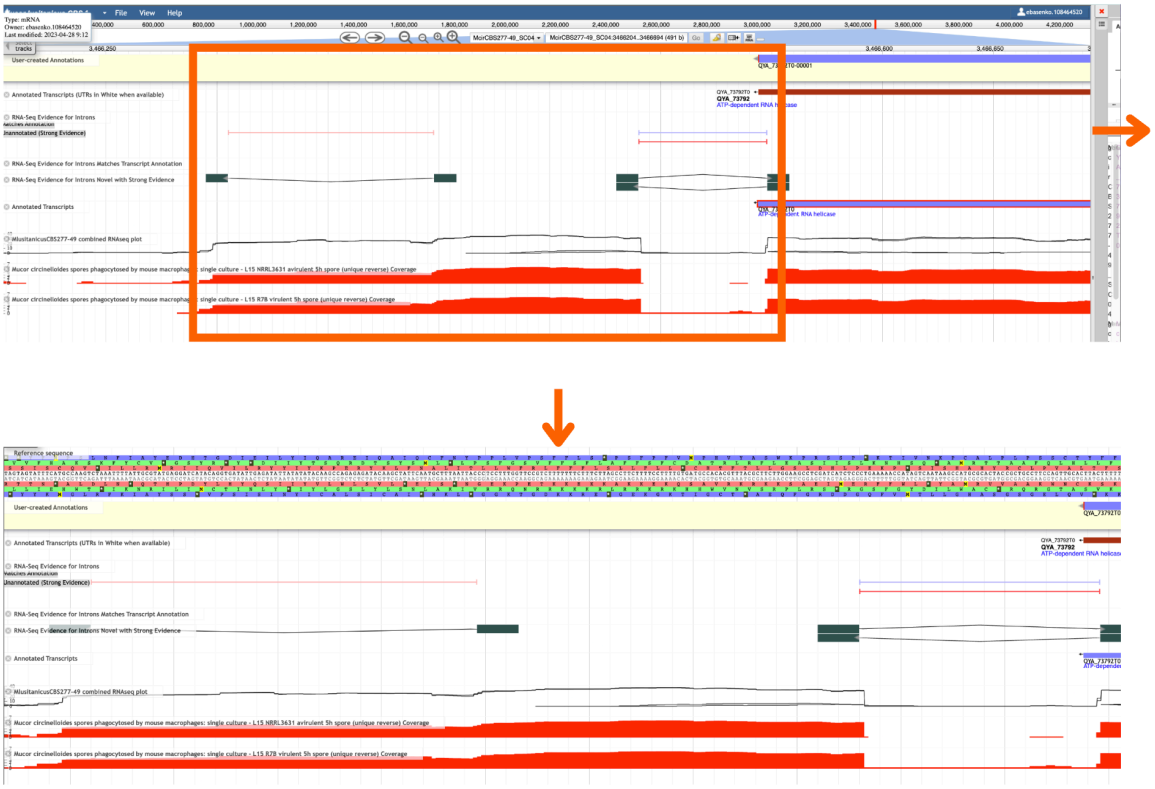
1. **Drag a gene model into the User-created Annotation workspace.**

To do this, double click on the gene in the “Annotated Transcripts” tracks which was selected from the “Draggable Annotation” section on the right. Double-clicking will highlight the whole gene rather an individual component. Note: The “Annotated Transcripts (UTRs in White when available” track cannot be used for this purpose.

2. Activate the “Reference sequence” track to guide gene correction.

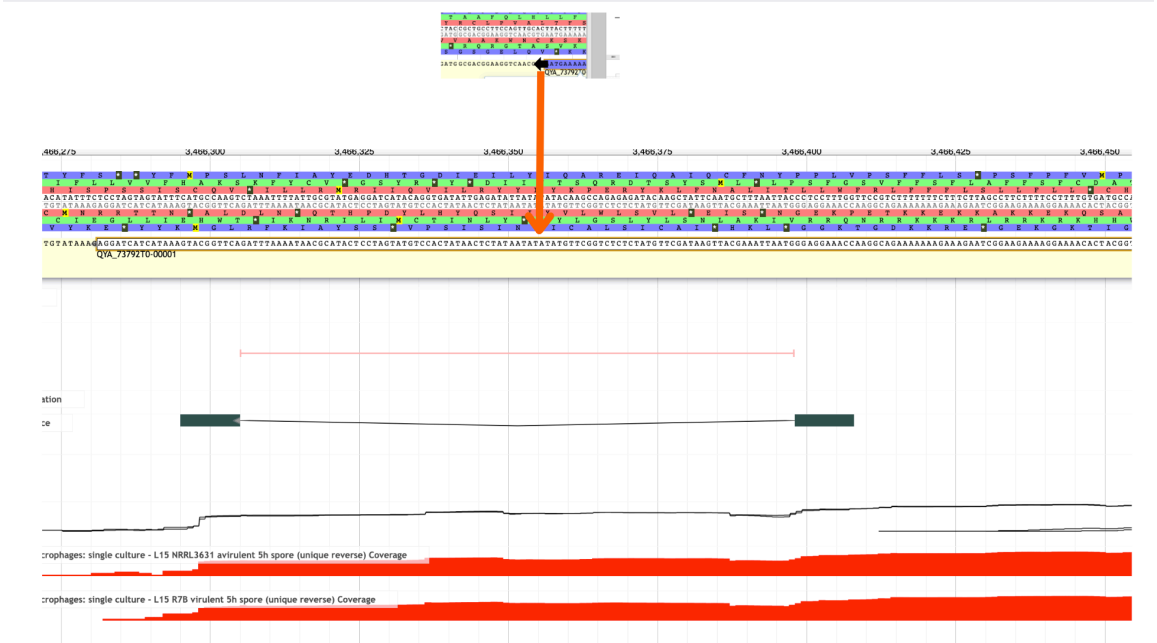
Activate the “Reference sequence” track from the Track menu on the left. To be able to view the Reference sequence track, you must be zoomed in a good bit.

You may want to drag the right window to the right to create more working space within the Apollo editor and then use the cursor to zoom in to the section highlighted by the orange box:



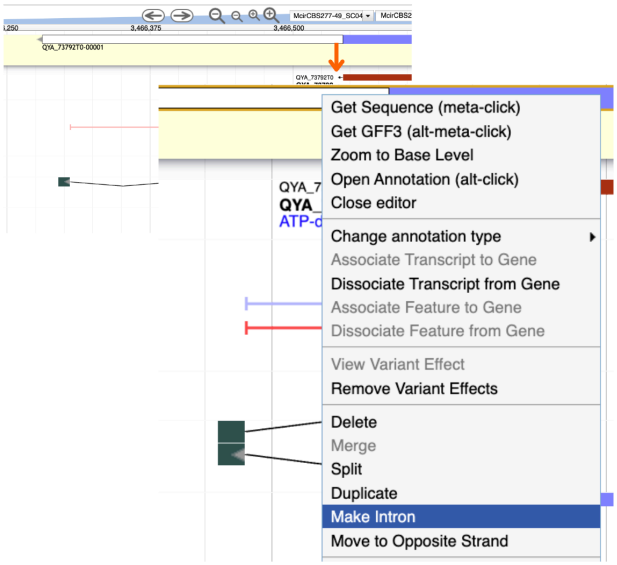
3. Extend gene model using evidence tracks for guidance.

Hover over the track in the “User-created Annotation” until a small black arrow appears at the end of the track. Left click on the arrow and extend gene model:



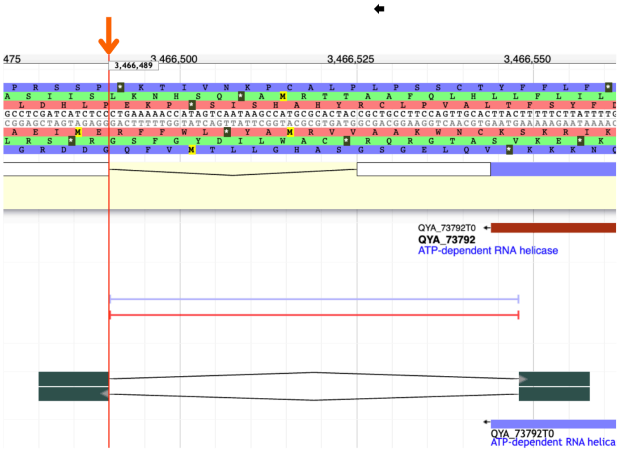
4. Create intron.

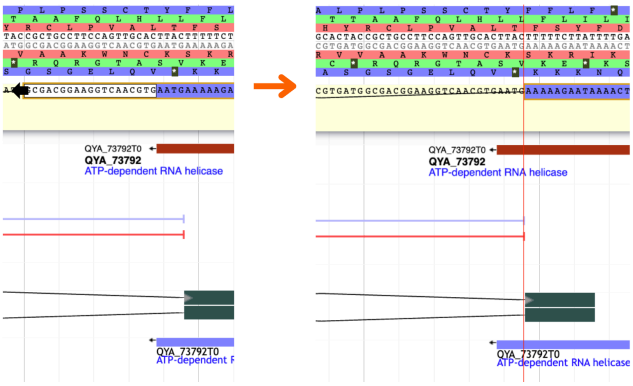
Zoom out, right-click on the white box gene feature created by Apollo as a result of the gene model extension, and select "Make intron" option.



5. Modify intron boundaries.

Apollo will automatically create an intron feature. Now, zoom in to adjust the boundaries and use the "RNA-Seq Evidence for introns Novel with Strong Evidence" track for guidance. In this case, the left boundary looks ok:



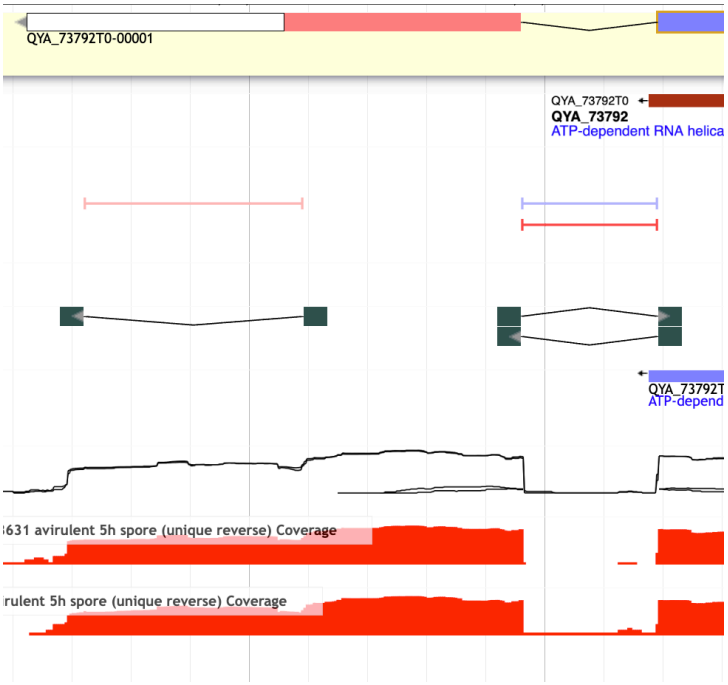


However, the right boundary should be adjusted.

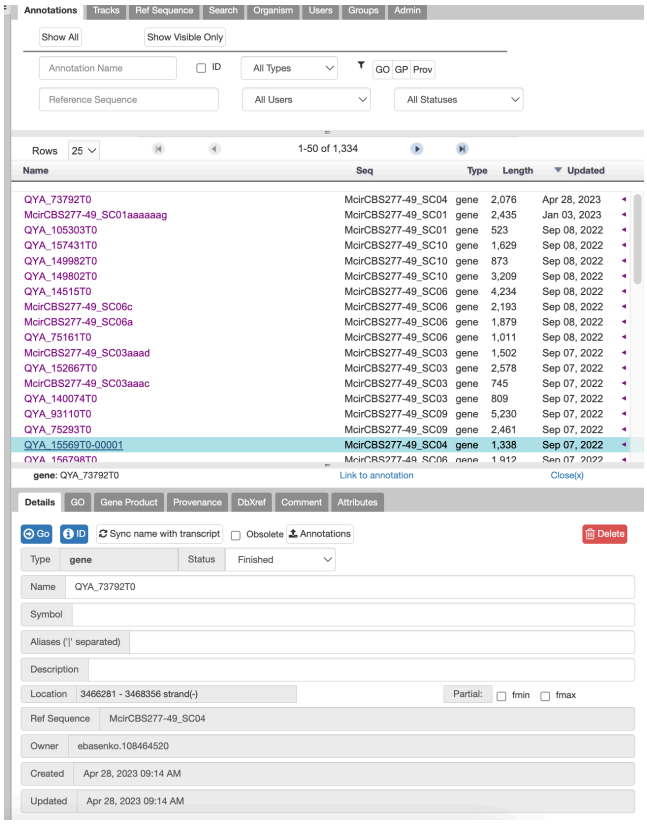
Click to select the gene model and hover over until a small black arrow appears next to the boundary.

Use the arrow to adjust the intron boundary. You can use the cursor to check the boundary alignment with the RNA-Seq data.

Notice that this adjustment automatically annotated an extra exon. Apollo also automatically predicted a UTR.



Once the new gene model is complete, navigate to the Annotations > Details, etc. tabs to provide evidence and comments. One the status is changed to “Finished” the new gene model will become visible for other users.



SGD Variant Viewer

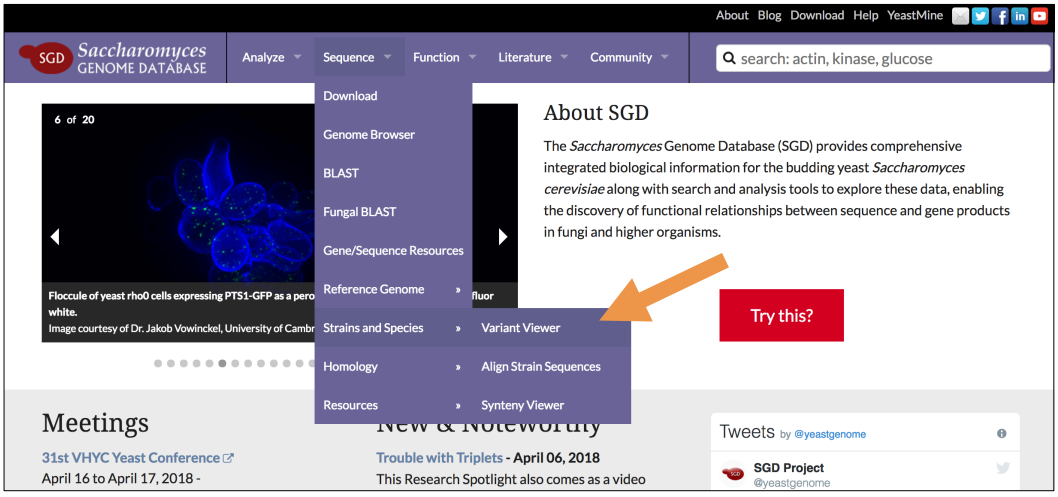
SGD's Variant Viewer (**https://yeastgenome.org/variant-viewer**) is an open-source web application that compares nucleotide and amino acid sequence differences between 12 common *S. cerevisiae* laboratory strains. For a given open reading frame, Variant Viewer breaks down the position and nature of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

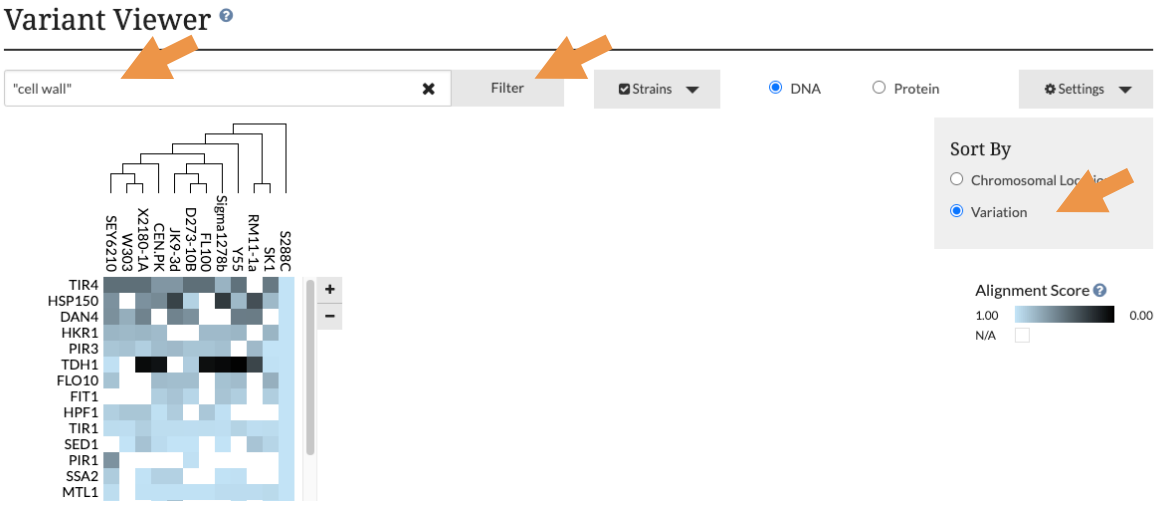
In this exercise, we will use Variant Viewer to find out what genetic differences between Sigma1278b and S288C explain why they differ in their ability to form pseudohyphae.

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

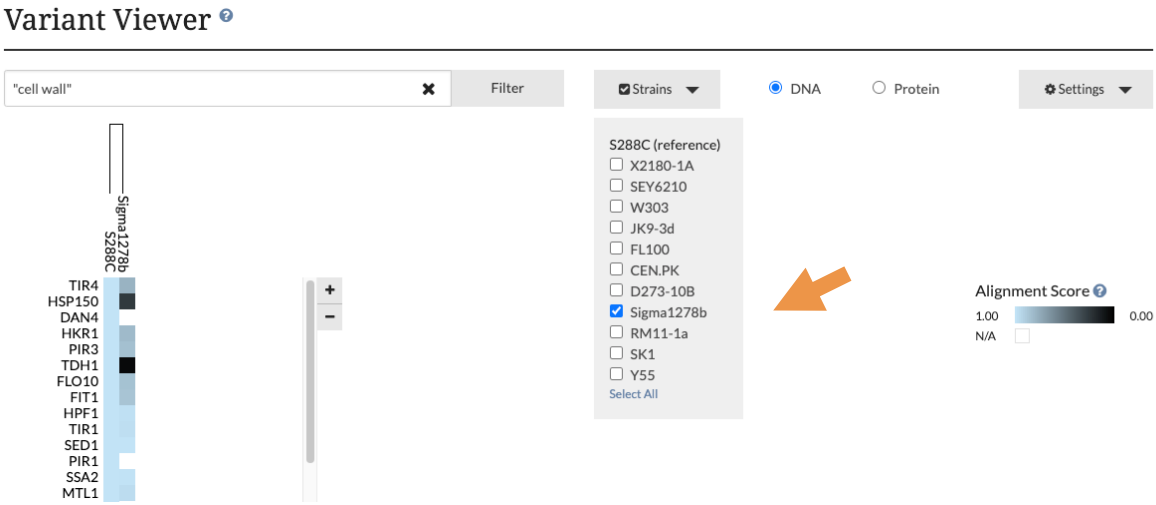


- The **Filter** box accepts one or more genes, as well as Gene Ontology (GO) terms. Because we are interested in genes involved in cell wall development, search for the GO term "**cell wall,**" sort by variation in the settings pull-down, and then click Filter.



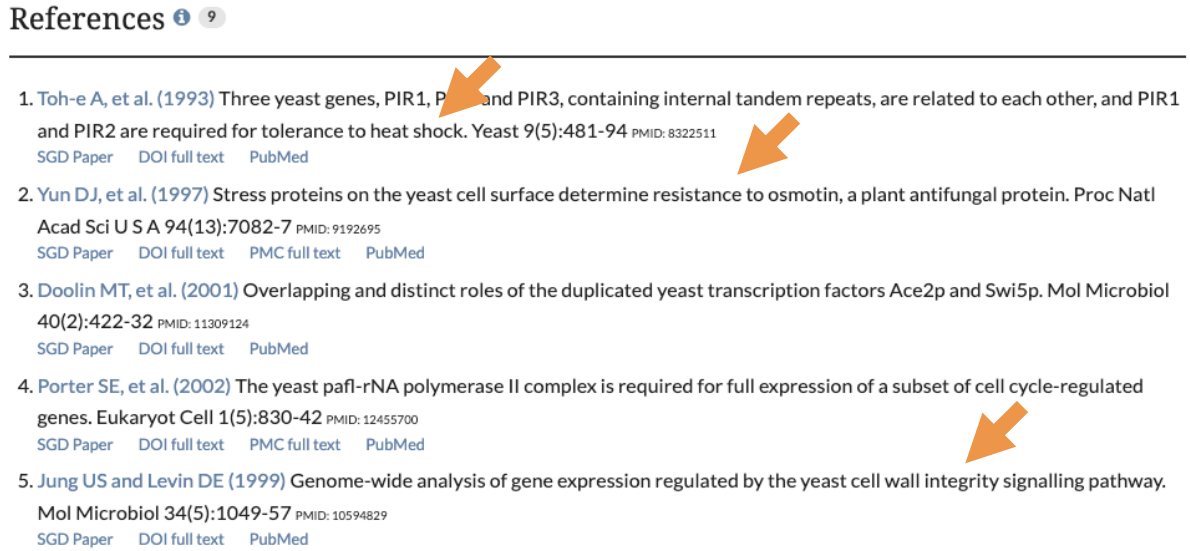
- The **matrix**, shown on the left, will have changed to only include the genes that localize to cell walls.
    - This matrix enables you to visualize high-level differences in multiple genes relative to strain S288C. Each square in the matrix corresponds to one of the twelve strains in Variant Viewer, 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 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



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.

BUD4 / YJR092W

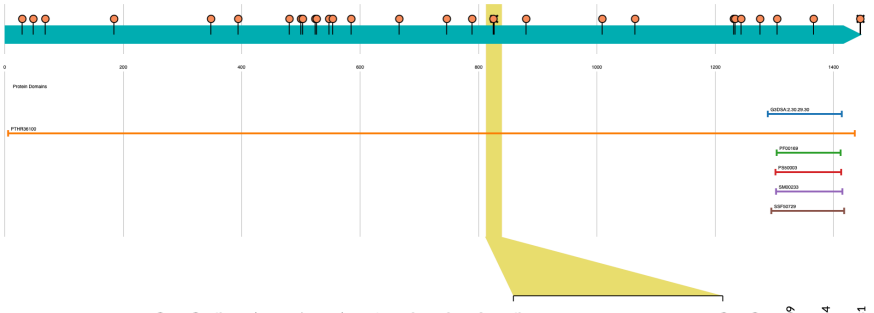
- Sequence Overview
- Reference Strain: S288C
- Alternative Reference Strains
- Variants
- Other Strains
- History
- Resources

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:

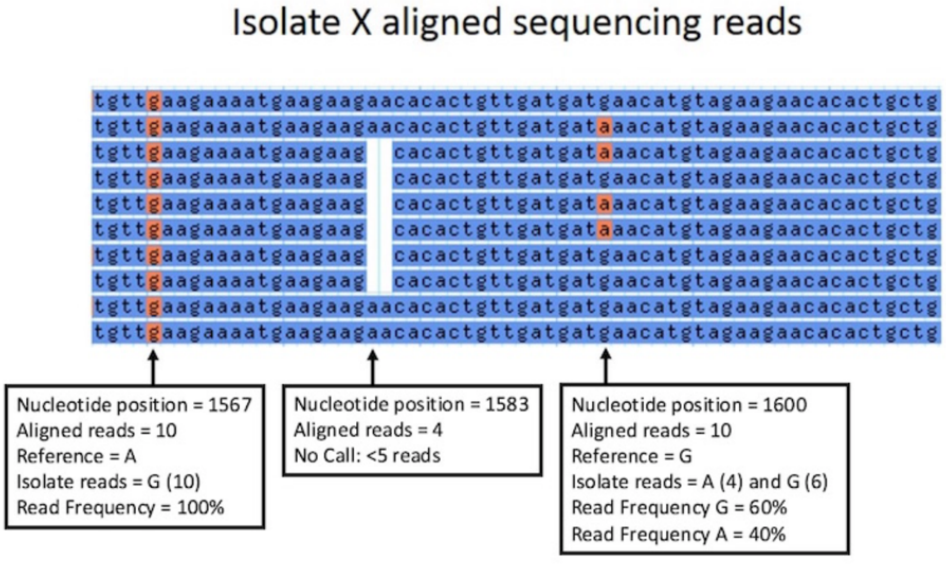
- Investigate SNP datasets using the following searches:
  - SNP characteristics,
  - SNPs between groups of isolates,
- Explore copy number variation records to identify aneuploidy cases.

SNPs have different functional effects with most having no consequential effect on gene function. SNPs may directly affect protein function when they are non-synonymous (results in a change in the amino acid; missense) or when they cause a premature stop codon (nonsense). SNPs that do not fall within genes are non-coding, but they may still affect splicing, mRNA stability, transcription, etc. SNPs can be used to characterize similarities and differences within a group of isolates or between two groups of isolates. They can also be used to identify genes that may be under evolutionary pressure, either to stay the same (purifying selection) or to change (diversifying or balancing selection).

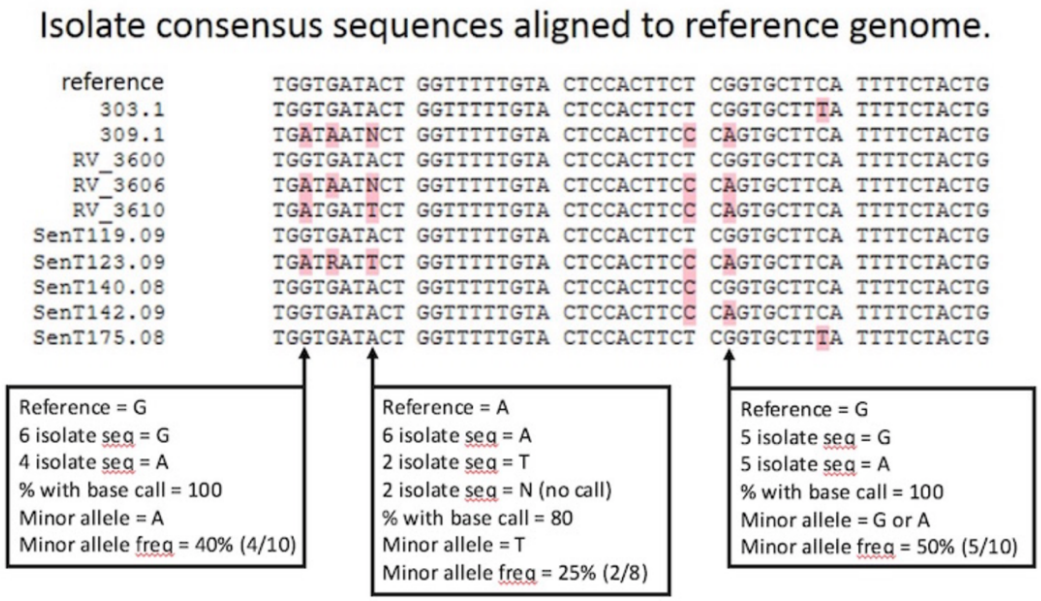
Read Frequency Threshold:

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.

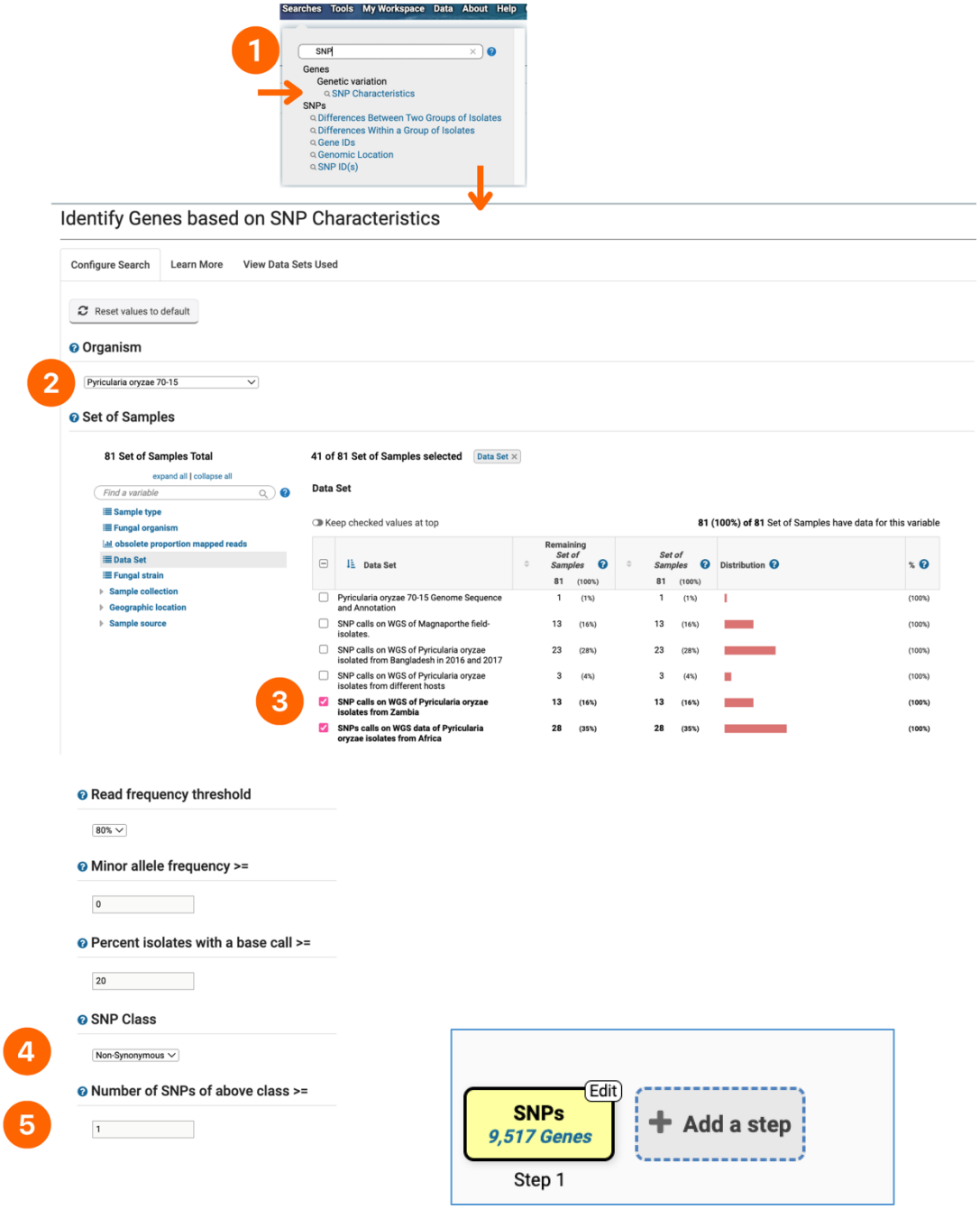


Percent isolates with a base call:
Sometimes an isolate does not have a base call at a certain nucleotide position because the Read Frequency Threshold was not met or because there were less than 5 aligned sequencing reads for that nucleotide position. In this case, a SNP can be returned by the search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

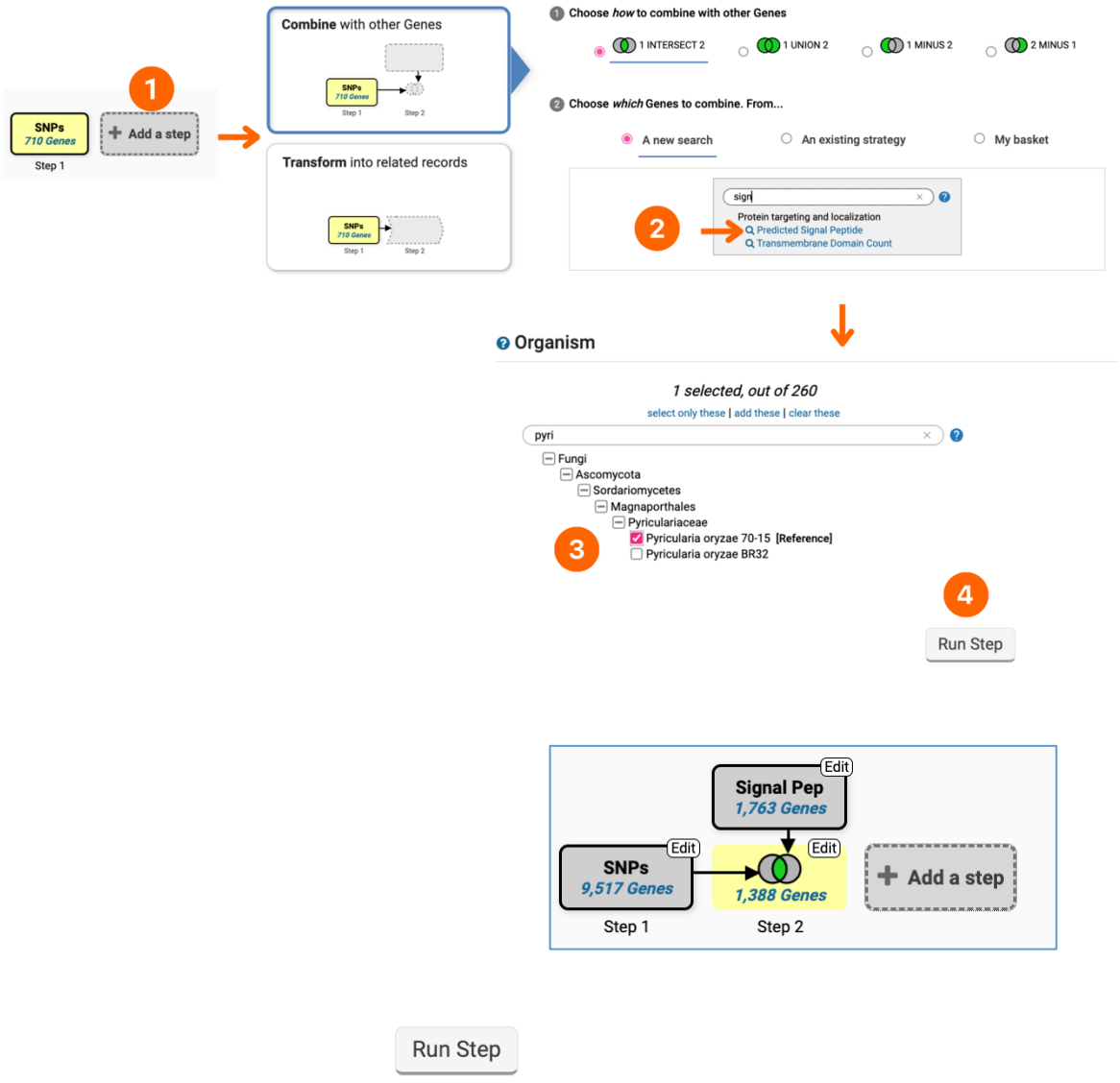
A. **Identify Genes based on SNP Characteristics search:** Identify putative nuclear effectors with at least 1 non-synonymous SNP in *Pyricularia oryzae*.

   P. oryzae is a plant pathogen that causes a devastating rice blast disease. *P. oryzae* and other plant pathogens use different types of effectors to modulate plant immunity during infection. Nuclear effectors have both a secretion signal and a DNA-binding domain. In the next exercise, we will examine *P. oryzae* isolates collected from infected rice plants in different locations in Africa and identify genes with at least one non-synonymous SNP that also carry signatures of nuclear effectors.

- **Identify genes with at least 1 non-synonymous SNP.**
    1. Deploy the “SNP characteristics’ search.
    2. Select *Pyricularia oryzae* 70-50 from the genome drop-down list.
    3. In the Data Set section, select the datasets where isoaltes 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.



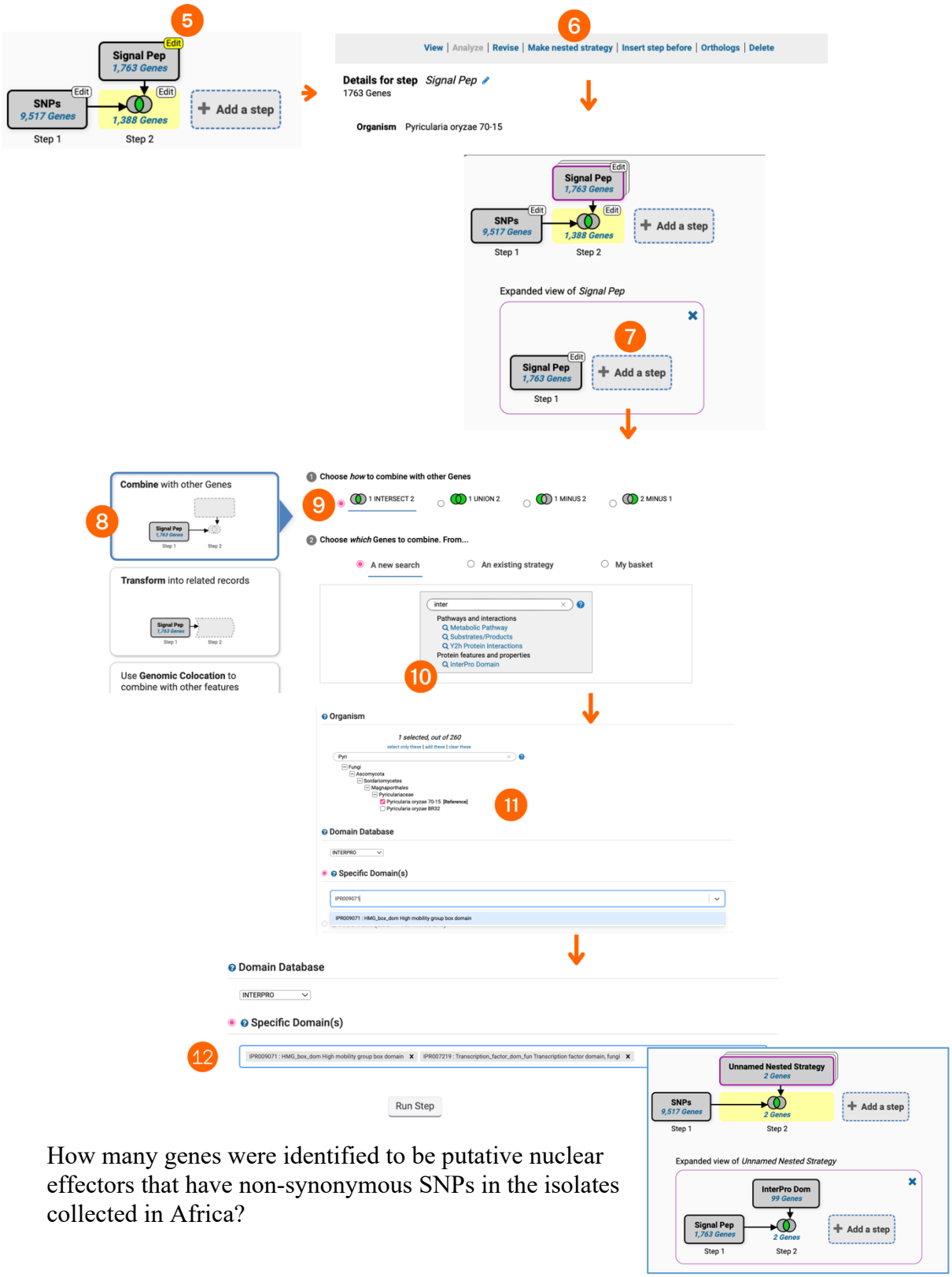
- **Identify putative nuclear effectors based on the presence of both a secretion signal and the DNA-binding domains IPR007219 or IPR009071 .**
    1. Click on the “Add a Step” button.
    2. Use the “Combine with Other Genes” option to deploy the “Predicted Signal Peptide” search.
    3. Set the genome to *Pyricularia oryzae* 70-50.
    4. Click on the “Run Step” button.



Note that currently, our strategy returns genes that have at least 1 SNP and also a predicted signal peptide domain. How can we identify that that have at least 1 SNP and ALSO a predicted signal peptide domain AND a DNA-binding domain? (Hint: create a nested strategy as described below).

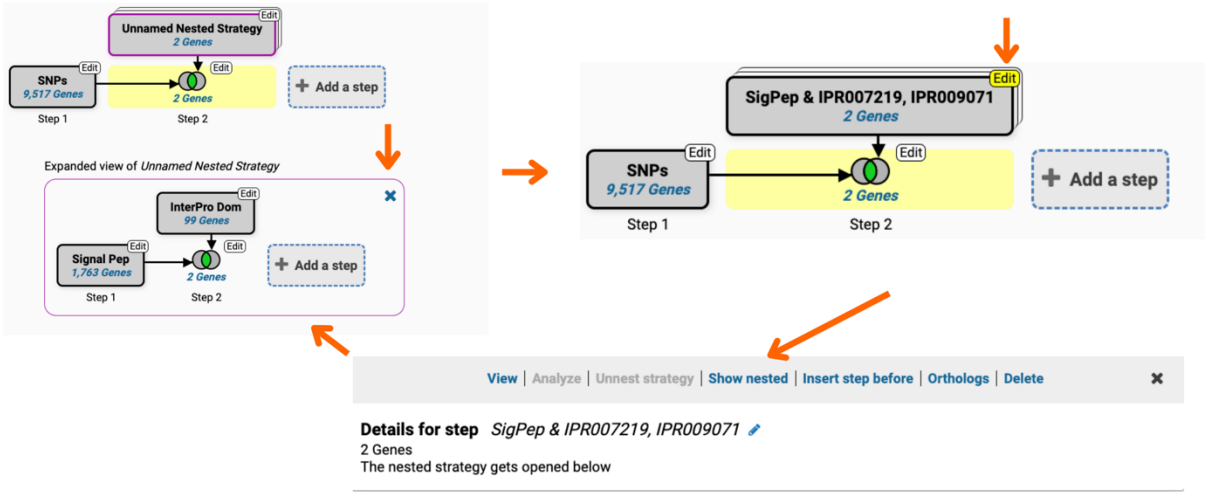
5. Hover over the “Signal Pep” search box and click on the “Edit” option.
6. Select the “Make nested strategy” option at the top.
7. Click on the “Add a Step” button within the “Expanded view of *Signal Pep”* (nested) strategy.
8. Select the “Combine with other Genes” search.

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?

Note: Nested strategy can be collapsed and expanded later as needed:



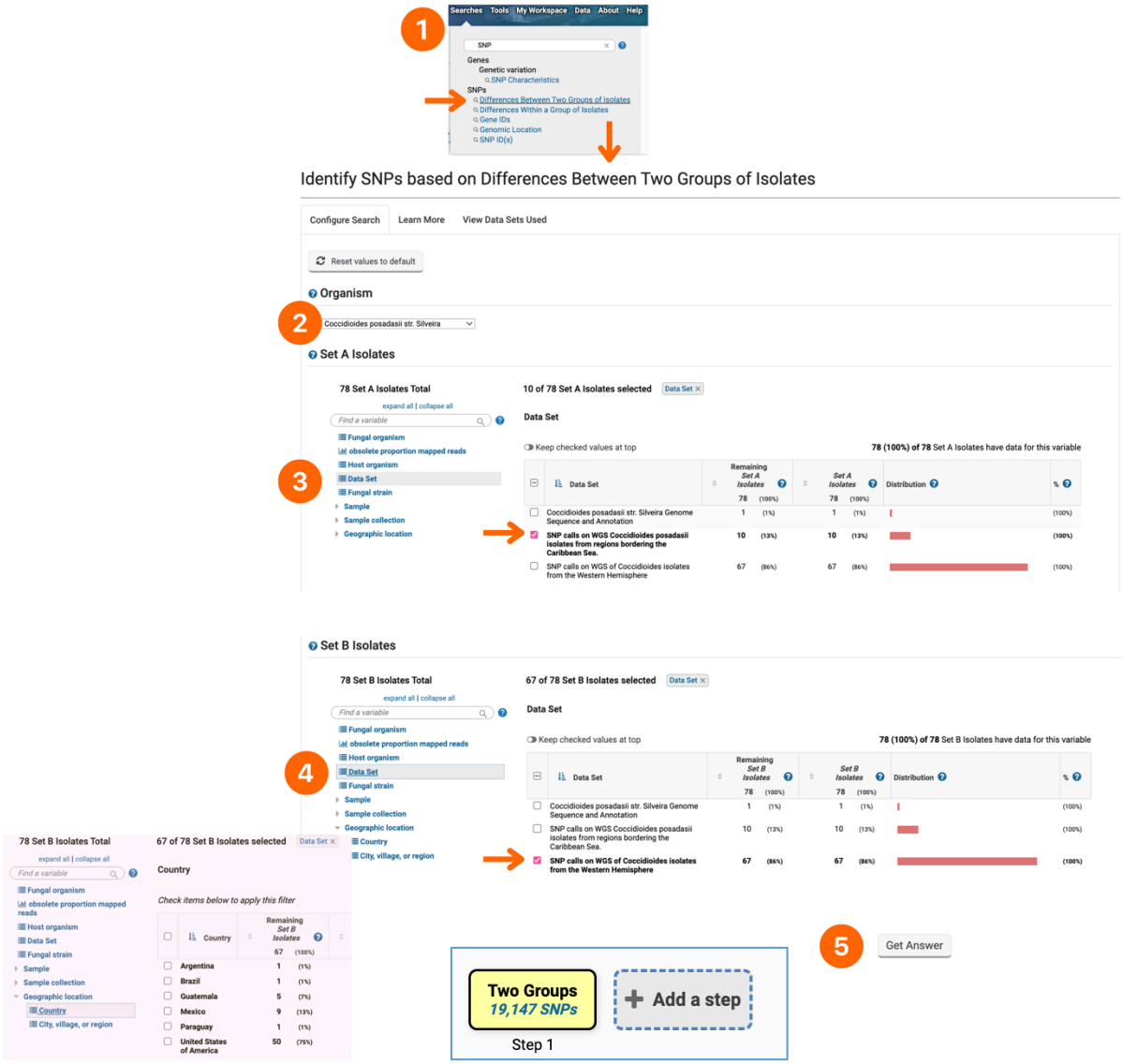
Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/bd657f5629cac5df

References: https://www.nature.com/articles/s41467-020-19624-w

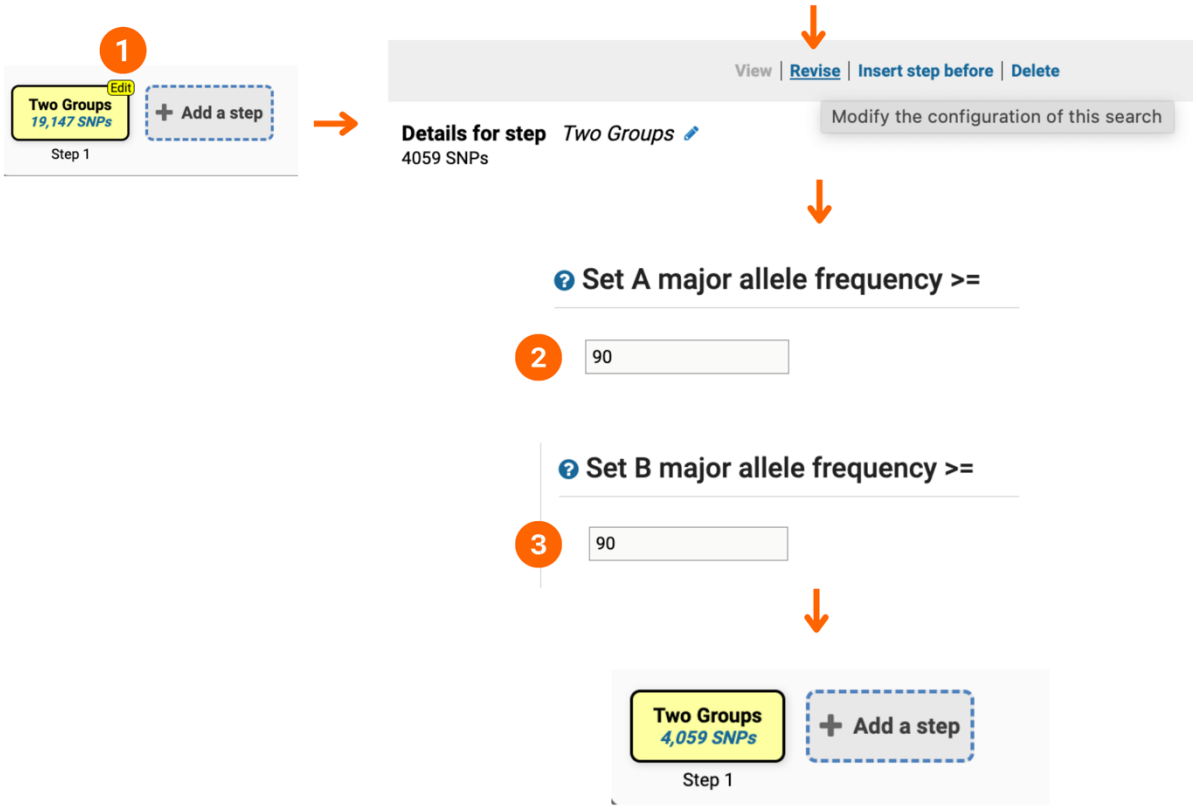
B. Identify SNPs based on Differences Between Two Groups of Isolates

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

- **Identify SNPs between two groups of *C. posadasii* str. Silveira isolates** (collected in Caribbean and Western hemisphere).
    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 Data Set menu: Caribbean dataset.
    4. Select Set B isolates from Data Set menu: Western hemisphere dataset. (Note: you can always examine other isolate medata (e.g., countries) as shown in the offset screenshot below).
    5. Click on the "Get Answer" button to get the results.



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



The search strategy returns SNPs rather than genes, which are classified by genomic location within the results table. When individual SNPs fall within a gene, 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 linked to its own record page. Click on the NGS_SNP.GL636536.6075.

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.

Add to basket Add to favorites Download SNP

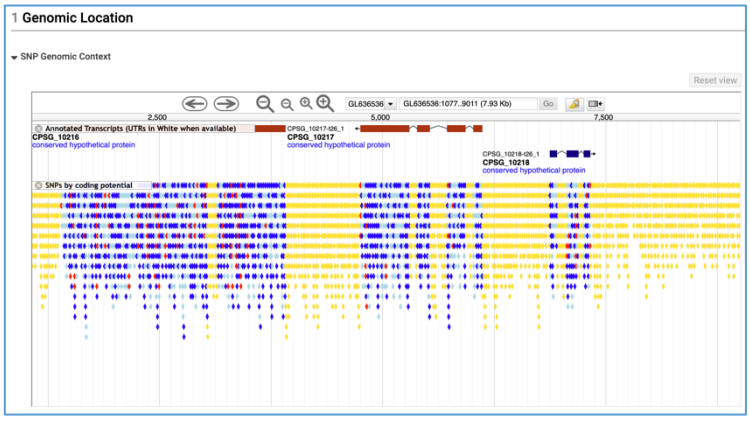
SNP: NGS_SNP.GL636536.6075

Organism: Coccidioides posadasii str. Silveira
Location: GL636536: 6,075
Type: coding
Number of Strains: 66
Gene ID: CPSG_10217
Gene Strand: reverse
Major Allele: C (0.58)
Minor Allele: T (0.42)
Distinct Allele Count: 2
Reference Allele: C
Reference Product: G 15
Allele (gene strand): G
SNP context: TCTGAGACTTTATTCTGGTTGCTTCTCTTCCCTTCCCTGTCCTTCCAGTTGTTGAATGAAT
SNP context (gene strand): ATTCATTCAACAACTGGAAGGACAGGGAAGGGAAGAGAAGCAACCAGAATAAAGTCTCAGA

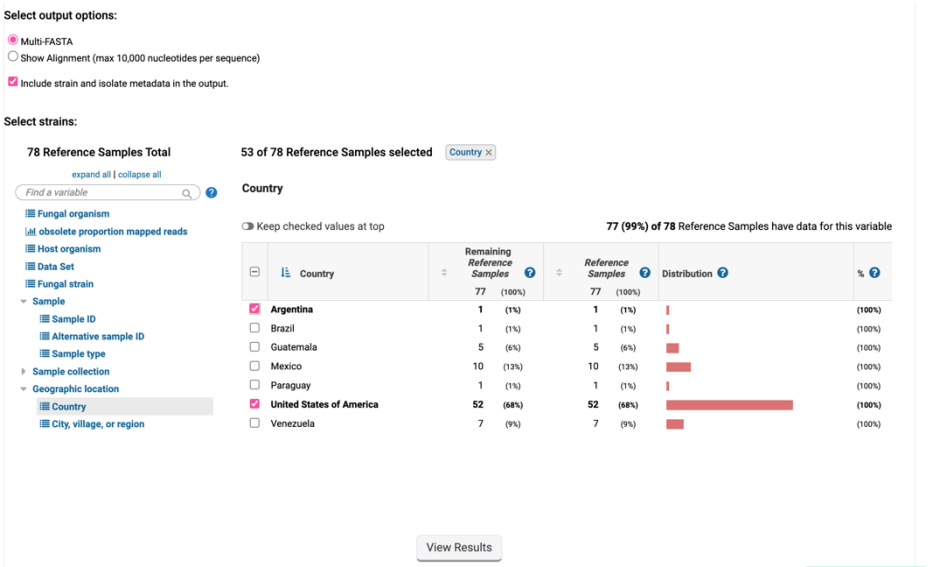
A summary of all SNPs detected in this gene across all datasets integrated into FungiDB is displayed in the SNP Genomic Context section:

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

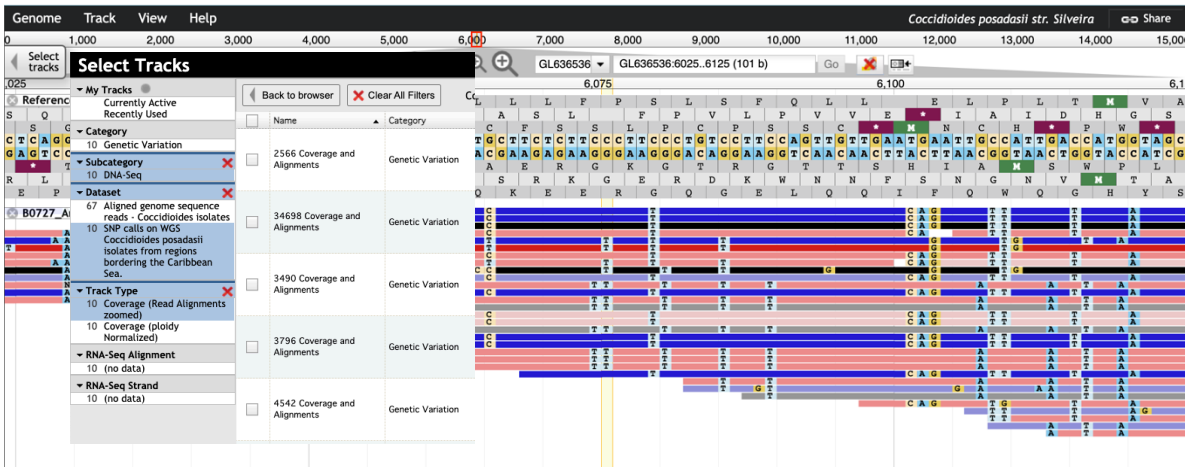
Search this table...

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

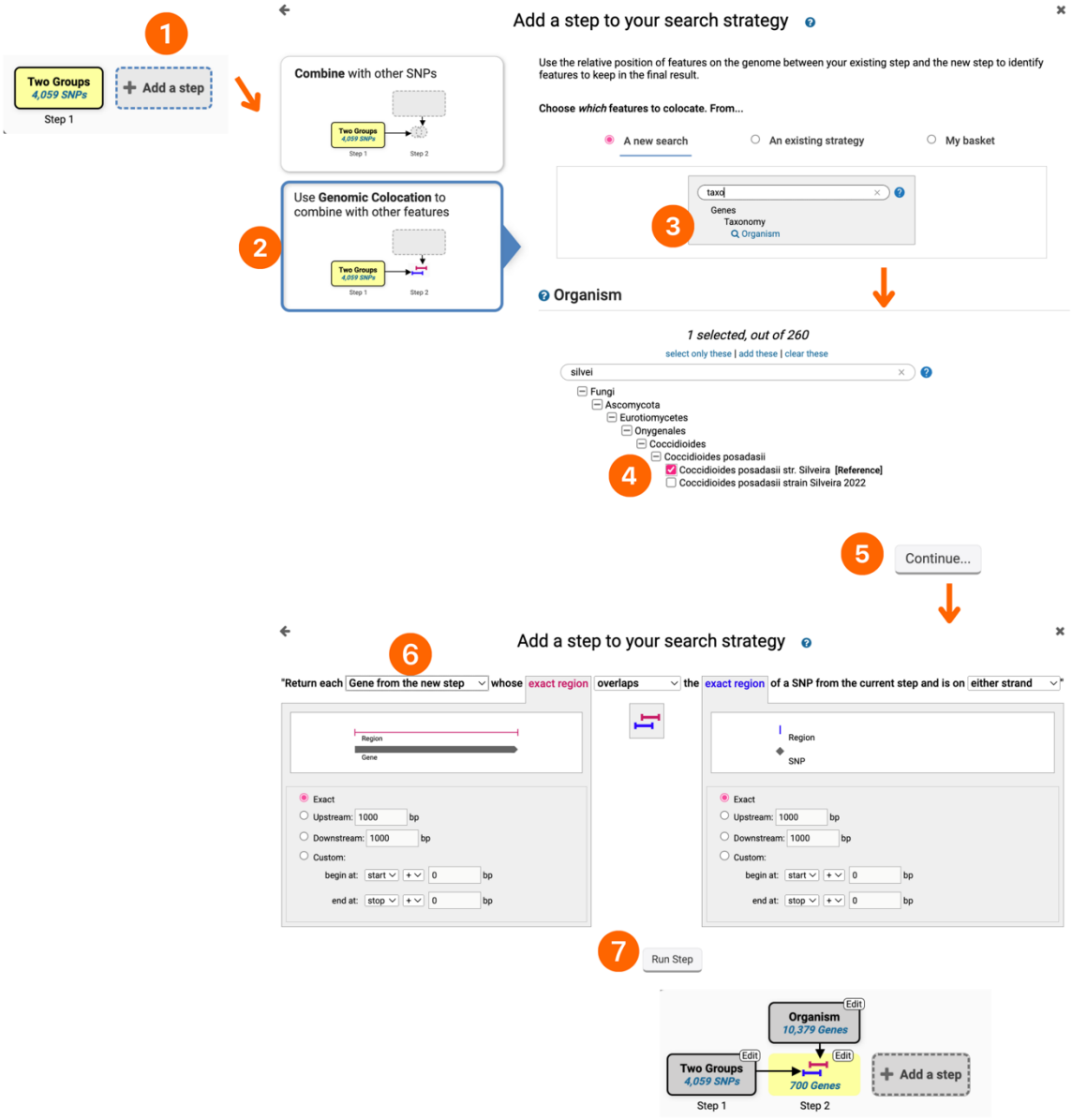
DNA-seq reads can be viewed by clicking on the *view DNA-seq reads* link from within the table.

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

Clicking on the "view DNA-seq reads" link will re-direct you to a JBrowse highlighting SNPs detected. You can select more tracks to examine by clicking on the Select Tracks tab on the left.



- **Identify *C. posadasii* str. Silveira genes that harbor geographic-specific SNPs.**
    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 genes by choosing the "Gene from the new step" from the drop-down menu while leaving other selections at default.
    7. Click on the "Run Step" button for results.



In this strategy we identified 700 genes that incurred different SNPs in different geographical locations. For those genes that are not well characterized (e.g., conserved hypothetical proteins) you can use other searches and tool to understand their function. You may also run a SNP search within a group of isolates to identify heterozygous (e.g., read frequency threshold 60%) or homozygous (e.g., read frequency threshold 80%) SNPs…

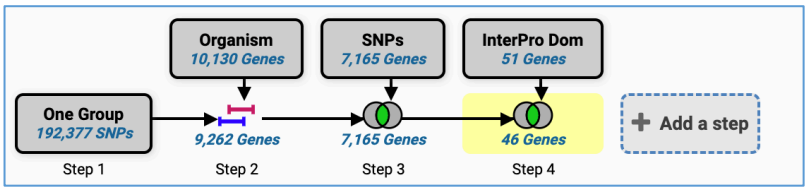
Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/d9d0fff2dbda229d

C. Identify SNPs within a group of isolates (optional)

- Deploy the SNP search called "Differences Within a Group of Isolates".
- Look for homozygous SNPs in *Aspergillus fumigatus* Af293 WGS (azole-resistance dataset). For example, here is one way to set your search:

Organism	Aspergillus fumigatus Af293
Samples	Data Set: Genomic Context of Azole-Resistance Mutations in Aspergillus fumigatus
Read frequency threshold	80%
Minor allele frequency >=	0
Percent isolates with a base call >=	20

Next, combine cross-reference homozygous mutations with *A. fumigatus* genes (Step 2) and identify genes that carry non-synonymous mutations only (Step 3; Hint: requires SNP Characteristics search), and look for ABC-transporters (Step 4; Hint: Requires InterPro Domain search; this example uses PF00005).



Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/ee44a65f5b67697a

Note: To identify heterozygous SNPs, set the read frequency threshold parameter to 40% and increase the minor allele frequency threshold (try 20 or 40).

Read frequency threshold *applies to the sequencing reads of individual isolates and defines a stringency for data supporting a SNP call between an isolate and the reference genome (Organism). Each nucleotide position of each isolate is compared to the reference genome and a SNP call is made if the portion of the isolate's aligned reads that support the SNP is above the Read Frequency Threshold (RFT). Find high quality haploid SNPs with 80% RFT or heterozygous diploid/aneuploid SNPs with 40%.*

Minor Allele Frequency *parameter applies to your group of isolates. A SNP can occur in any number of isolates in your group and the least frequent SNP call across all isolates is the Minor Allele Frequency. A SNP will be returned by the search if the frequency of the minor allele is equal to or greater than your Minor Allele Frequency.*

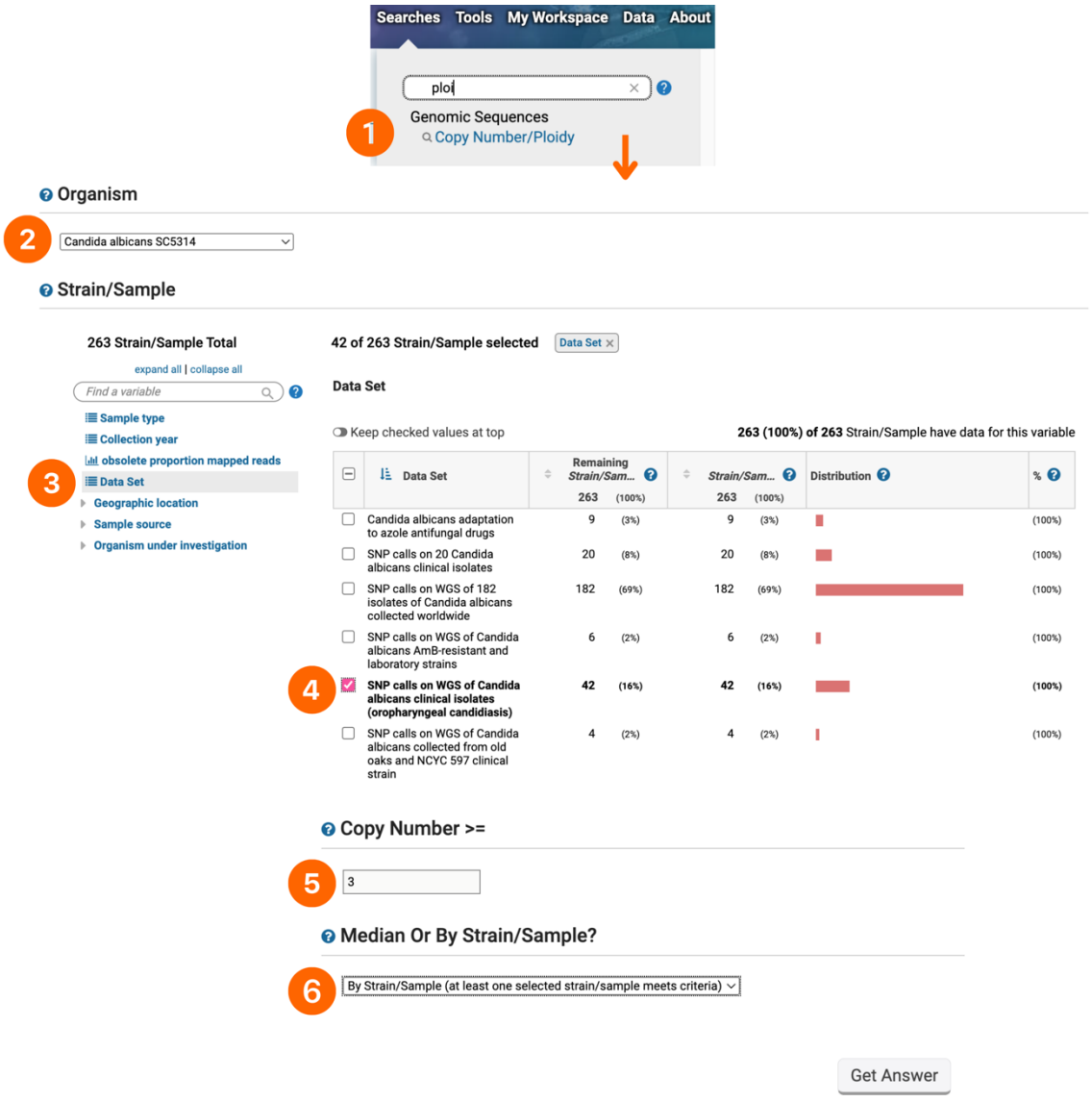
D. Copy number variation & ploidy searches.

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.

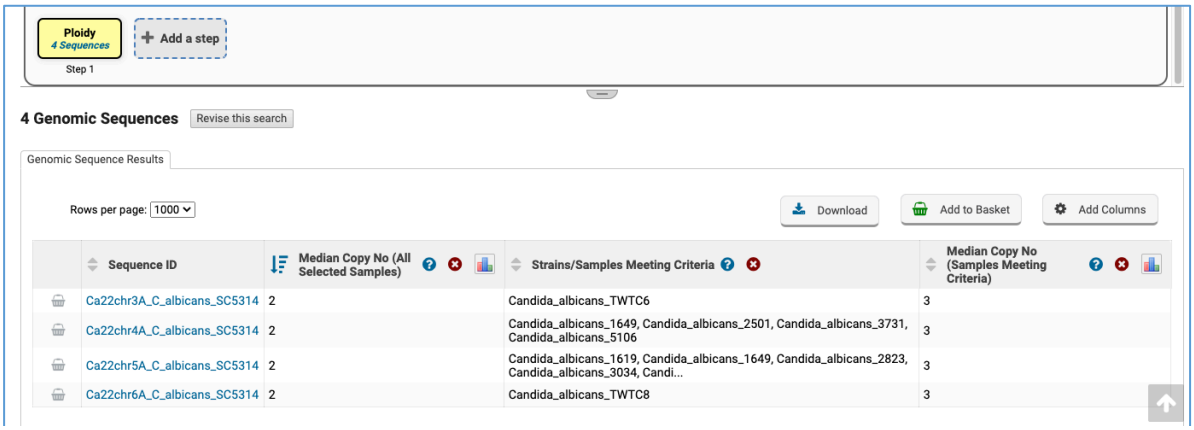
D.1. Copy Number/Ploidy search (Genomic Sequences)

Find genomic sequences or chromosomes based on their estimated copy number in resequenced strains. Genomic sequences returned by the search will have 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 has to 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).

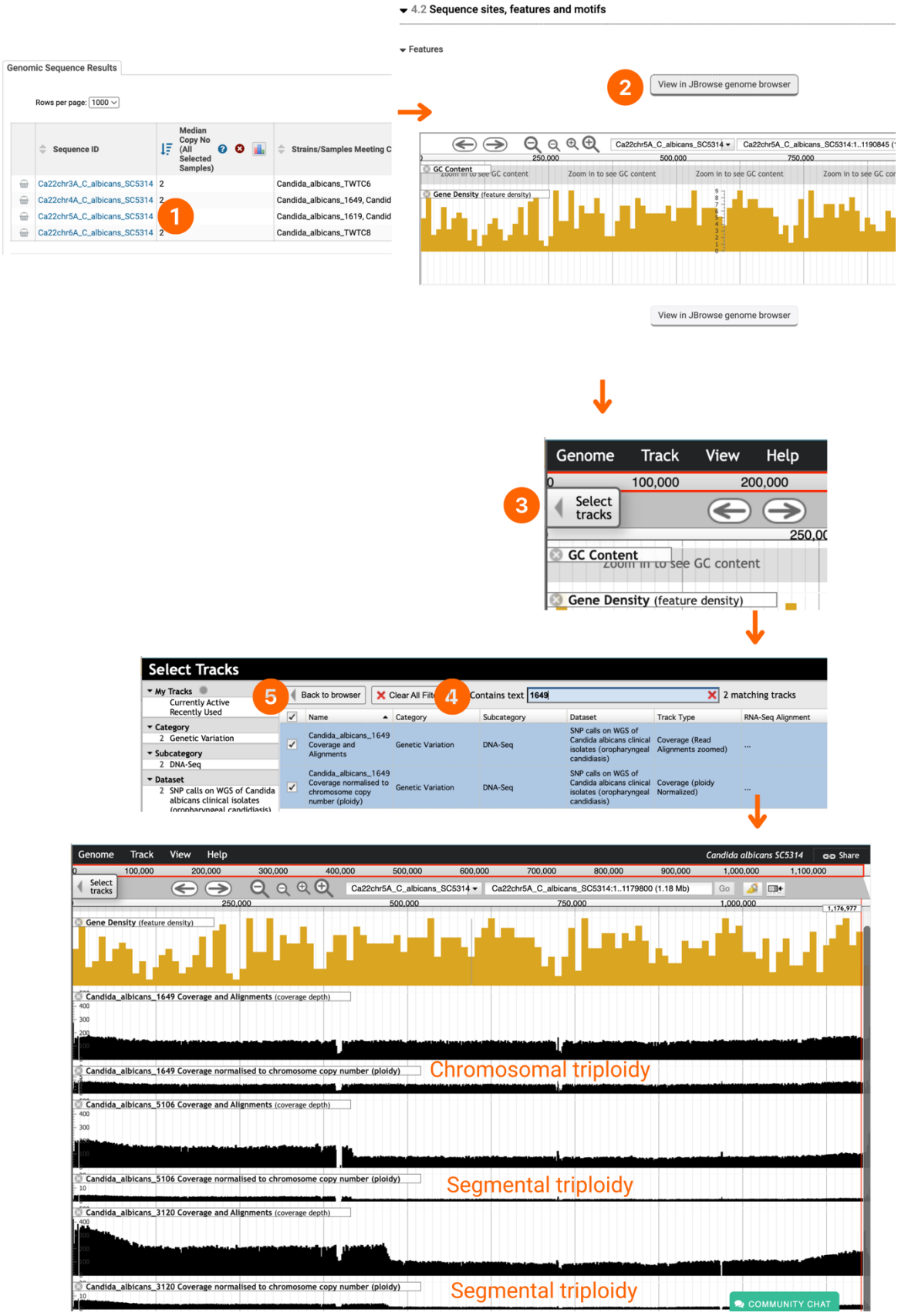


- **Explore segmental aneuploidy in JBrowse.**

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). Note that the whole chromosome is shown in both screenshots, and both tracks are shown for each sample. Note: VEuPathDB is not currently normalizing for telomere proximity.

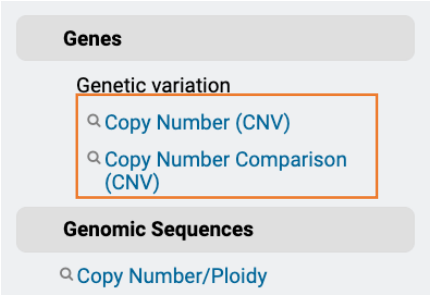
URL:
https://fungidb.org/fungidb/jbrowse/index.html?loc=Ca22chr5A_C_albicans_SC5314%3A1..1190845&data=%2Ffungidb%2Fservice%2Fjbrowse%2Ftracks%2FcalbSC5314&tracks=gcContent%2CgeneDensity%2CCandida_albicans_1649%20Coverage%20and%20Alignments%2CCandida_albicans_1649%20Coverage%20normalised%20to%20chromosome%20copy%20number%20(ploidy)%2CCandida_albicans_5106%20Coverage%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/6dc86b214d14a5f3
References: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383195/

D.2. Copy Number search (Genes)

E. 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 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:

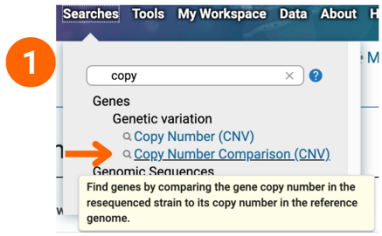
- **Haploid number** is the number of genes on an individual chromosome.
- **Gene dose** is the total number of genes in an organism, accounting for copy number of the chromosome.

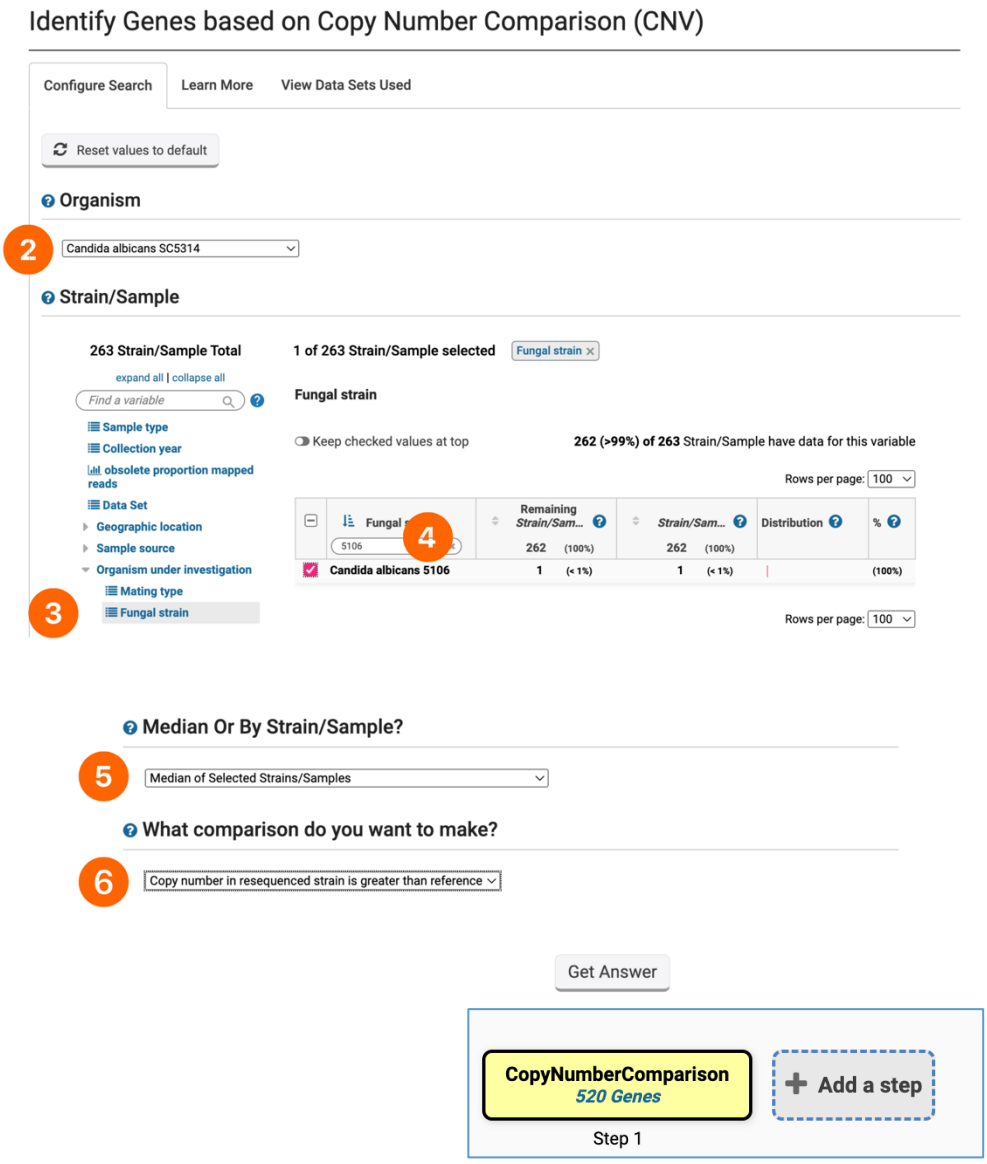
For example, a single-copy gene in a diploid organism has a haploid number of 1 and a gene dose of 2. You can choose to search for genes where at least one of your selected isolates meets your cutoff criteria for the chosen metric (By Strain/Sample), or where the median of the chosen metric across all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

- **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.

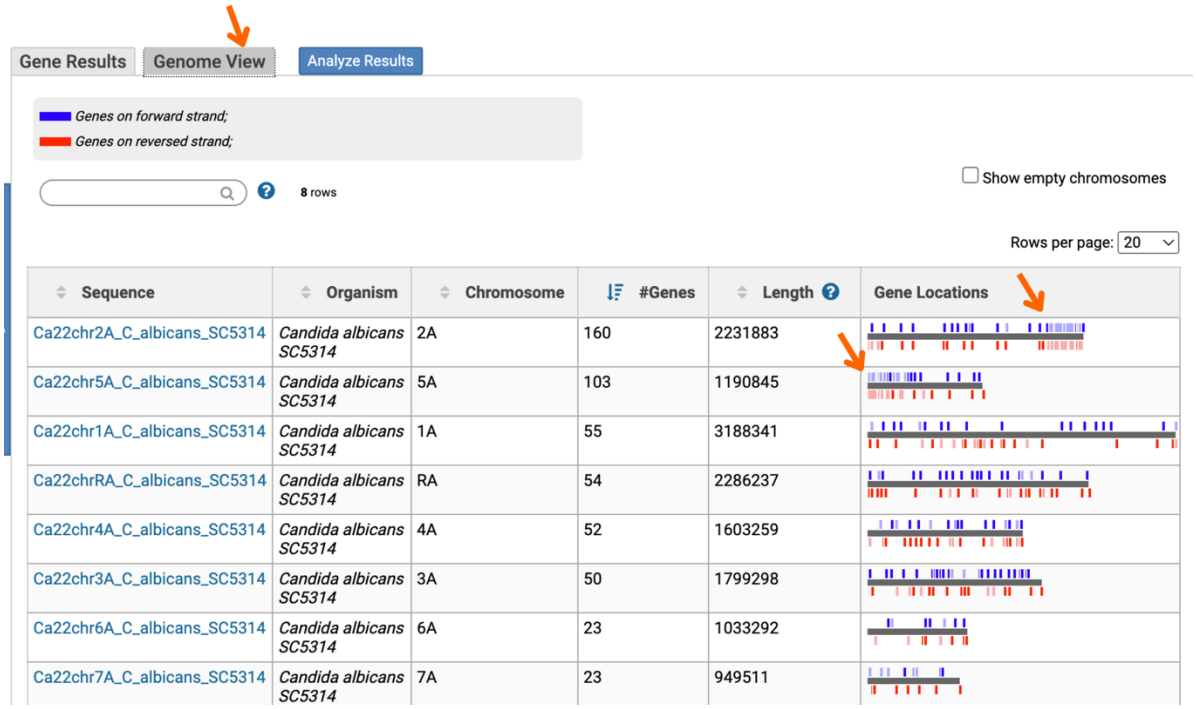
    Note: Choosing Median or By Strain/Sample will only make a difference if you have multiple strains.

    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.



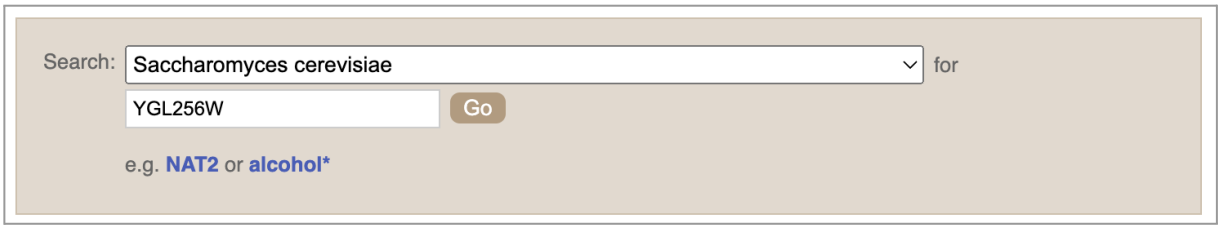
Sequence	Organism	Chromosome	#Genes	Length	Gene Locations
Ca22chr2A_C_albicans_SC5314	*Candida albicans SC5314*	2A	160	2231883	
Ca22chr5A_C_albicans_SC5314	*Candida albicans SC5314*	5A	103	1190845	
Ca22chr1A_C_albicans_SC5314	*Candida albicans SC5314*	1A	55	3188341	
Ca22chrRA_C_albicans_SC5314	*Candida 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 that are 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.

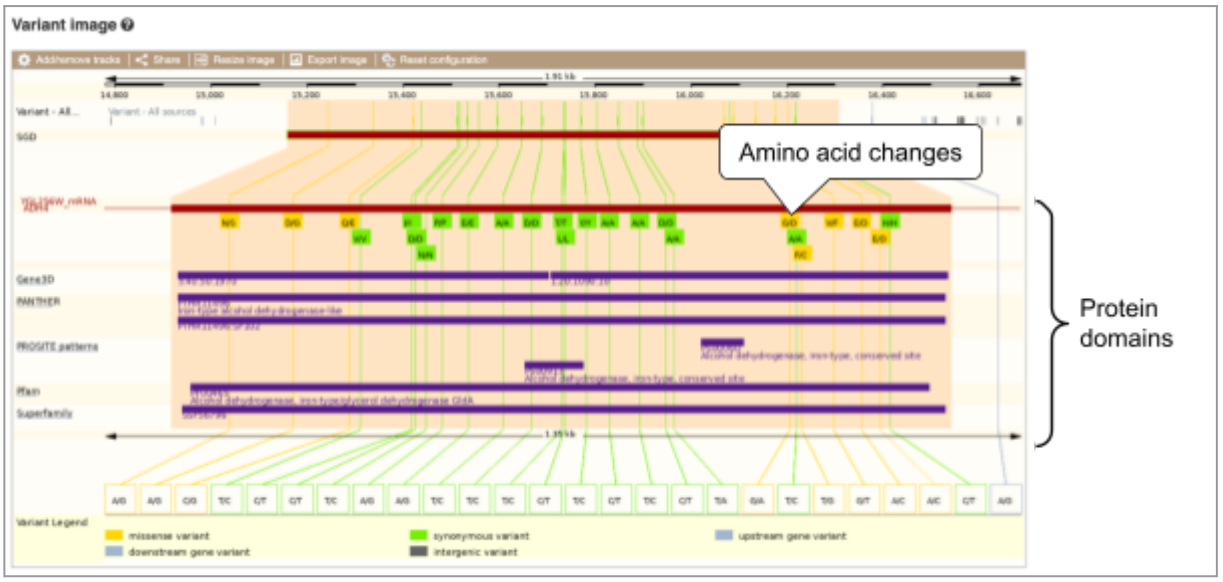
Exercise: Exploring variants in Ensembl Fungi

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 ones we are interested in using the variant table. Click on the Variant table link.

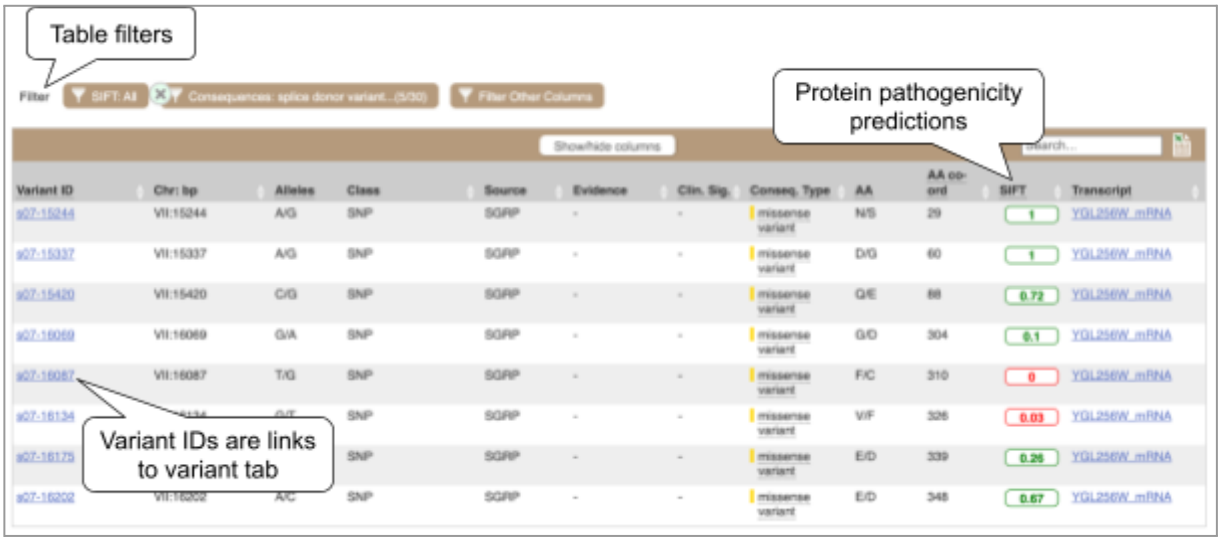
This table shows the variants in order of their occurrence through the genome, and they are reported on the forward strand. The gene *ADH4* is located on the forward strand, so we are first shown variants upstream of the gene (starting a the 5' upstream region).

(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 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 up the variation tab.

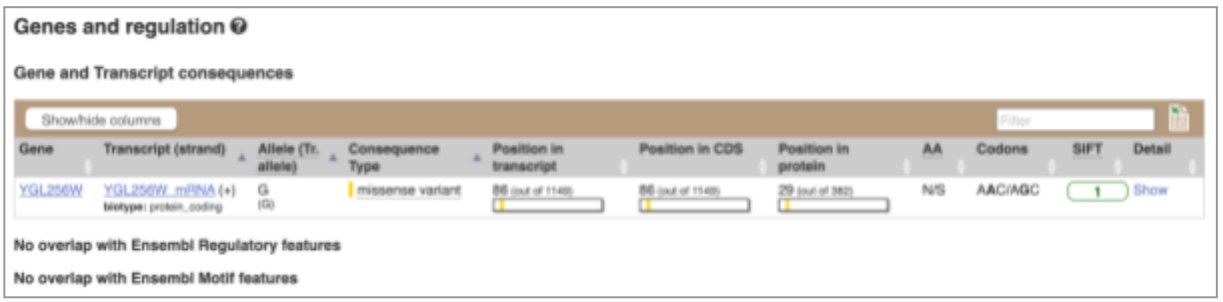
The icons show you what information is available for this variant.

(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.*)

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

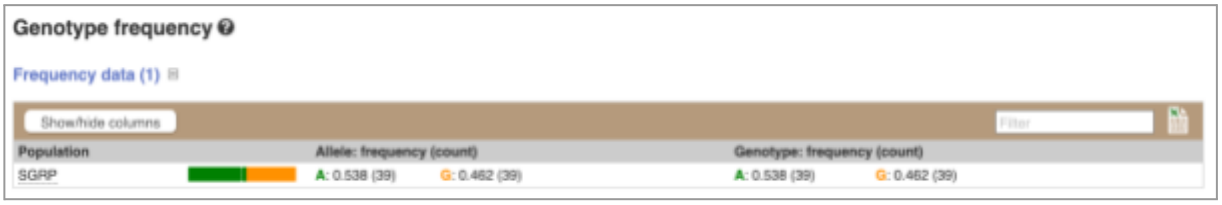
Click on Genes and regulation, or follow the link at the left.



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 SGRP study?

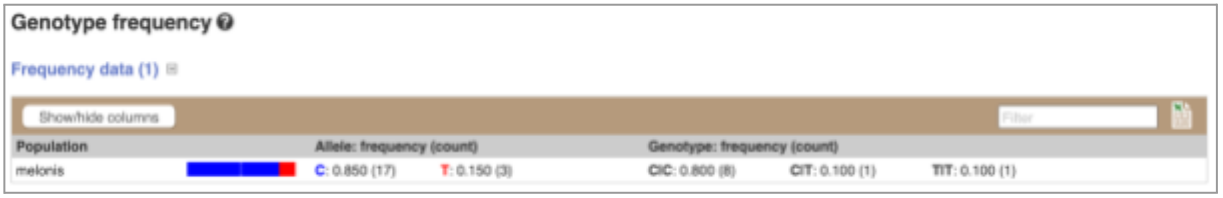
Click on Genotype frequency in the left-hand menu. Note that the reference allele is more frequent than alternative allele in this case.



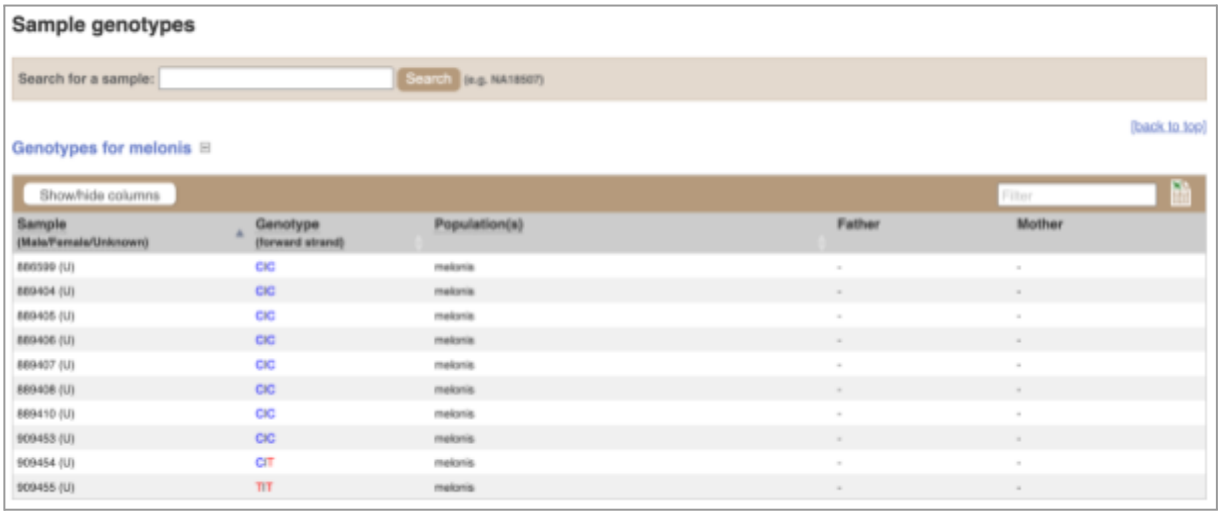
Additional Exercise – Variation data in *Fusarium oxysporum*

(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?



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.

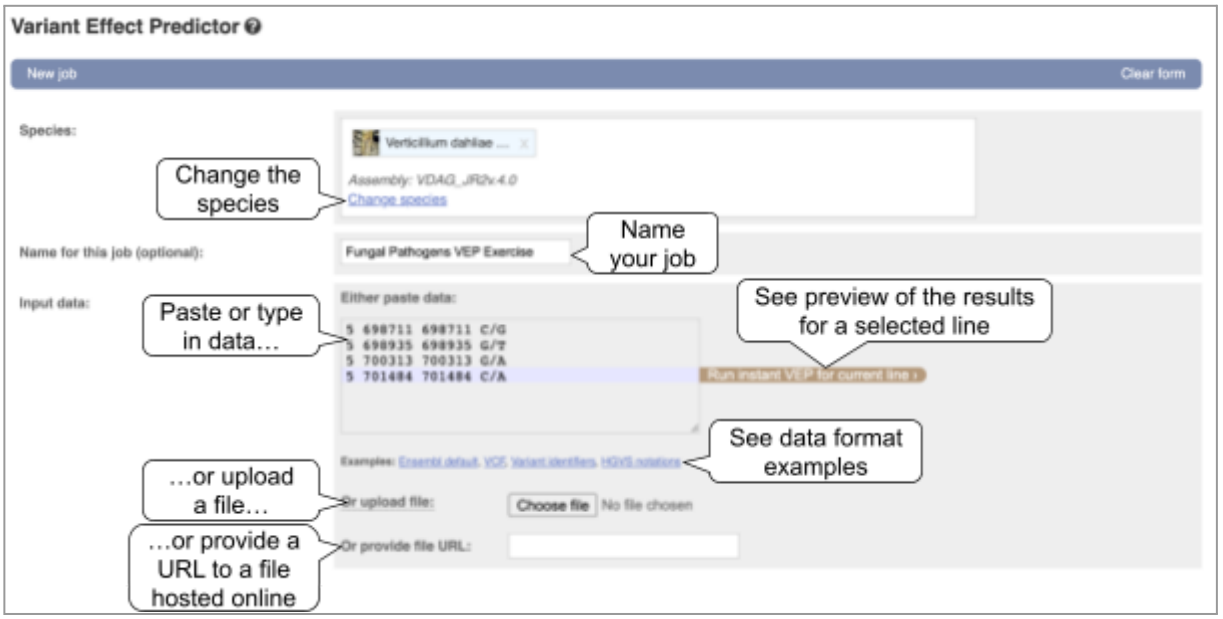
Use the Ensembl VEP to determine:
(a) Are your variants novel or have they already been annotated in Ensembl? (b) What genes are affected by your variants?
(c) Do any of your variants affect gene regulation?

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.

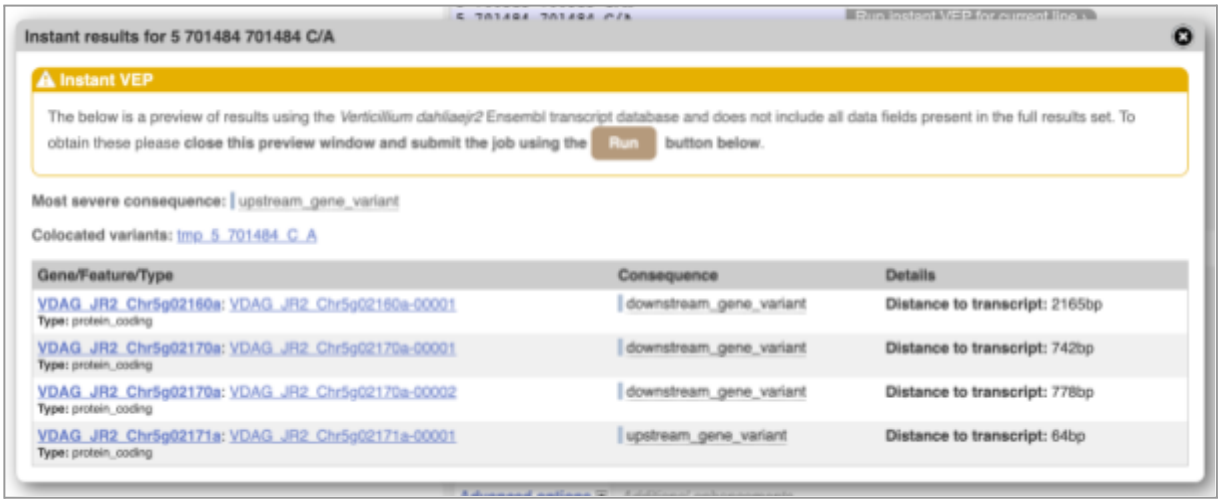
The VEP recognises a number of input formats including the Ensembl default format, VCF, Variant identifiers and HGVS notations.

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:

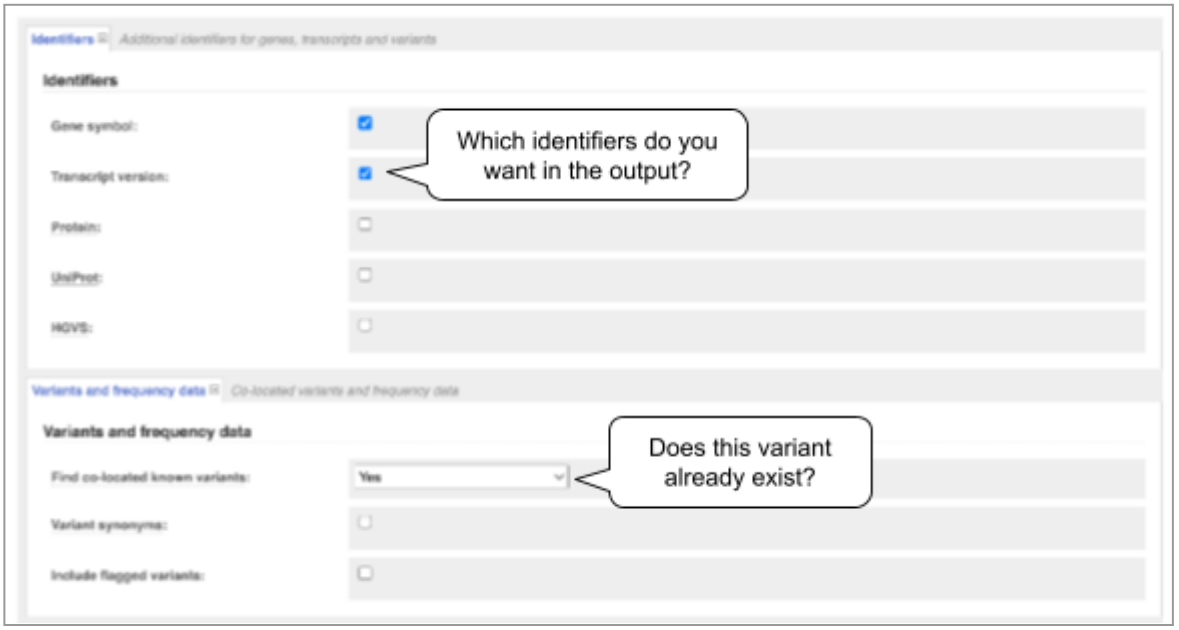
5 698711 698711 C/G
5 698935 698935 G/T
5 700313 700313 G/A
5 701484 701484 C/A

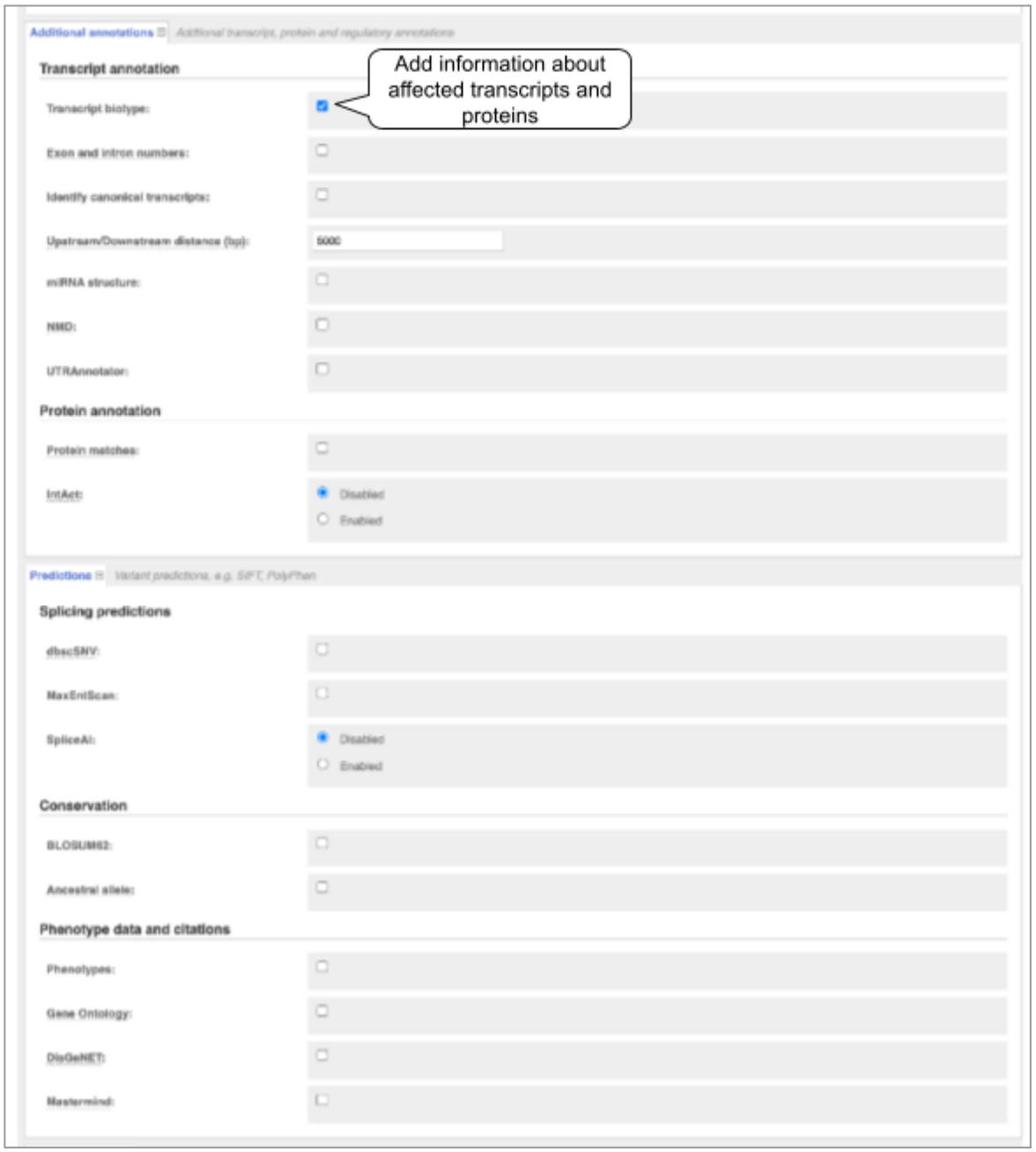


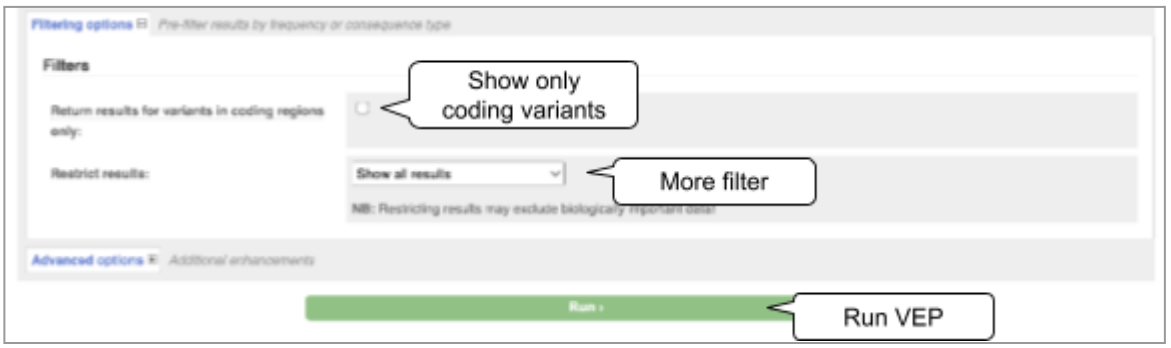
The VEP will automatically detect that the data is in Ensembl default format.
Clicking on the ‘Run instant VEP for current line’ will generate a pop-up with summarised results for that individual variant.



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.





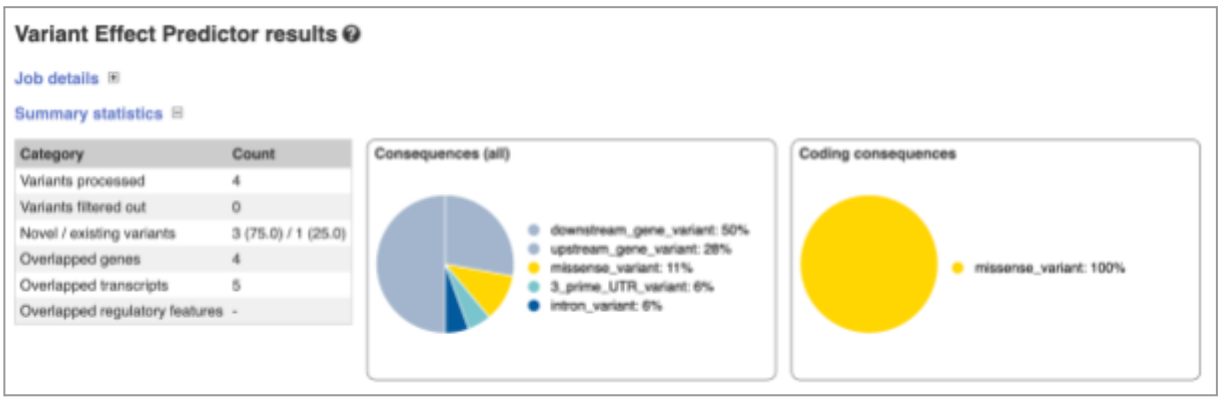


Hover over the options to see definitions. When you've selected everything you need, scroll right to the bottom and click Run.

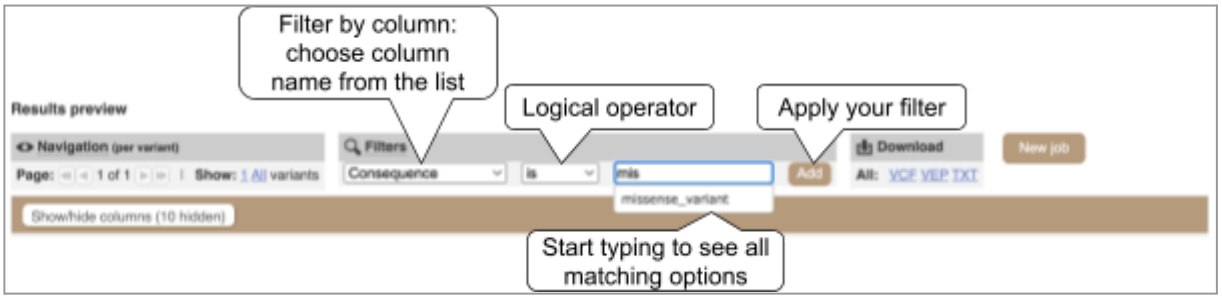
A table display will show you the status of your job. It will say Queued, then automatically switch to Done when the job is done, you do not need to refresh the page. You can edit or discard your job at this time. If you have submitted multiple jobs, they will all appear here.

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.

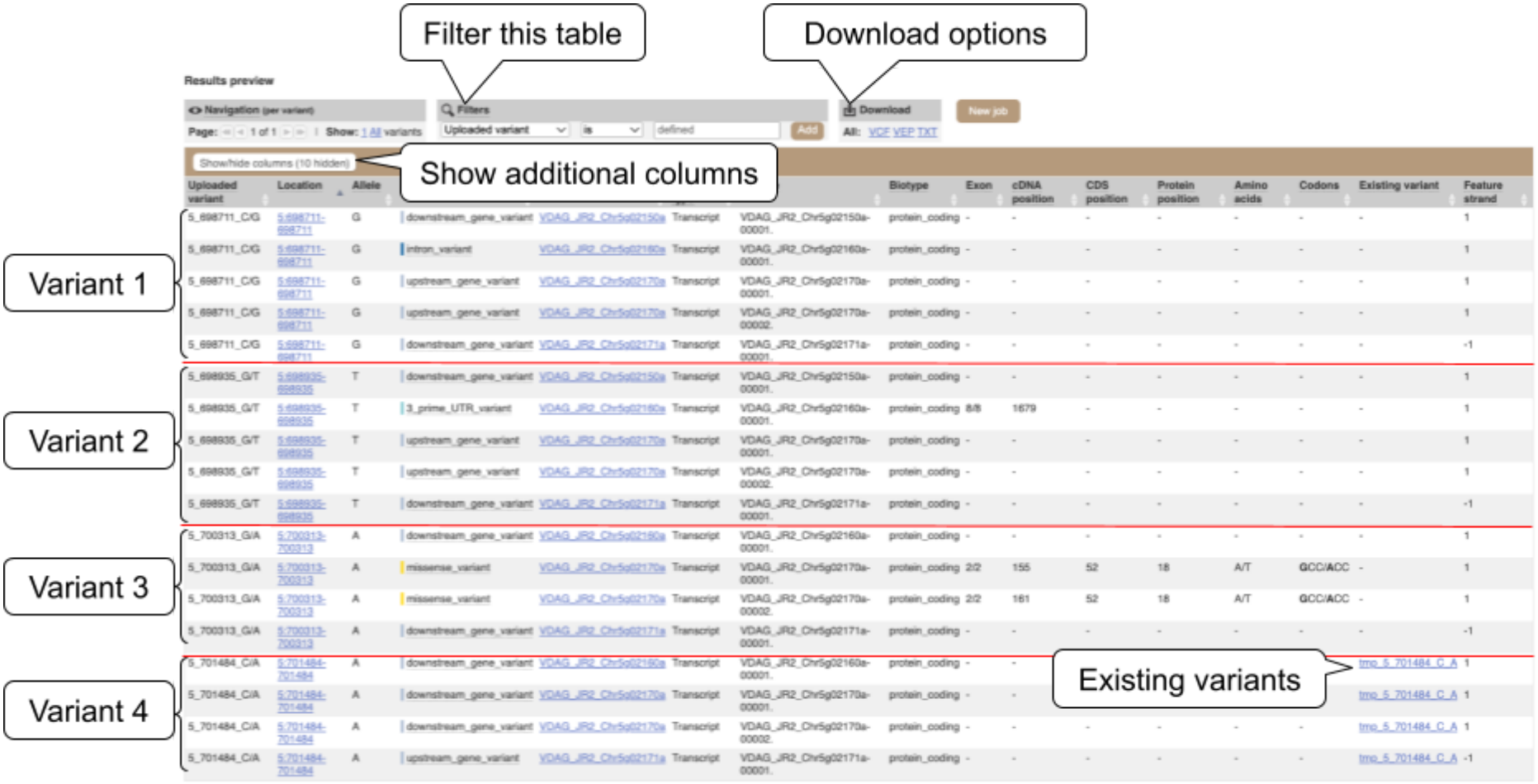
Let's come back to our questions:
(a) Are your variants novel or have they already been annotated in Ensembl?
(b) What genes are affected by your variants?
(c) Do any of your variants affect gene regulation?



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 ‘define’ in the search box, just leave it as it is.

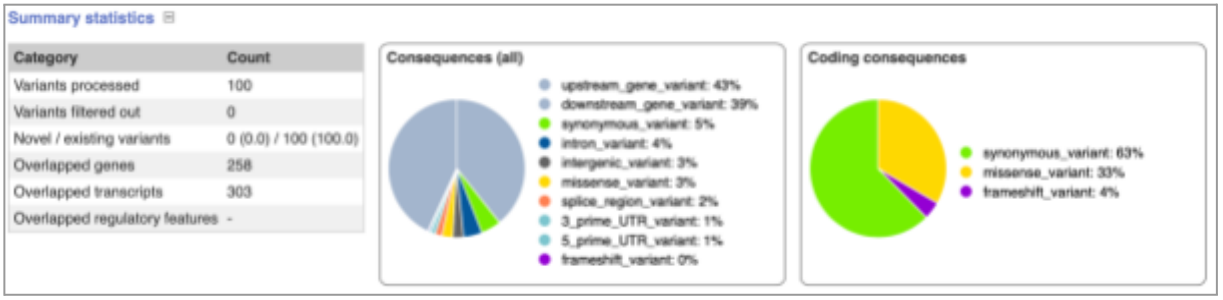

Filter this table
Download options
Results preview
Navigation (per variant)
Filters
Download
New job
Page: 1 of 1
Show: 1 All variants
Uploaded variant
is
defined
Add
All: VCF VEP TXT
Show/hide columns (10 hidden)
Show additional columns
Uploaded variant
Location
Allele
Biotype
Exon
cDNA position
CDS position
Protein position
Amino acids
Codons
Existing variant
Feature strand
Variant 1
5_698711_C/G
Variant 2
5_698935_G/T
Variant 3
5_700313_G/A
Variant 4
5_701484_C/A
Existing variants

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

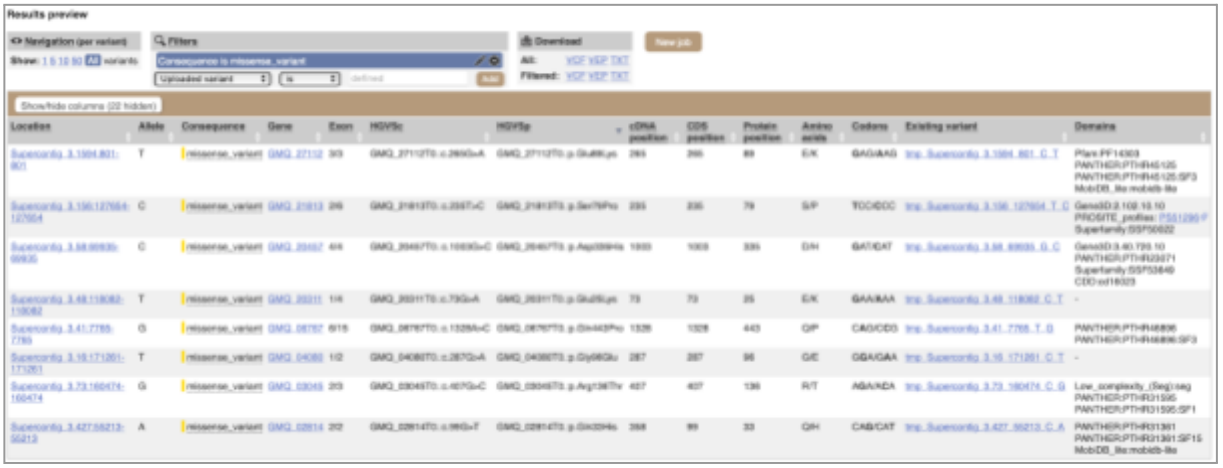
On the course file page, you will find a VCF file labelled VEP_exercise.vcf. This is a small subset of the outcome of *Puccinia graminis Ug99* whole genome sequencing and variant calling experiment. This file can also be found on our FTP site under the following link: http://ftp.ebi.ac.uk/pub/databases/ensembl/training/2021/FungalPathogens/VEP_exercise.vcf

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

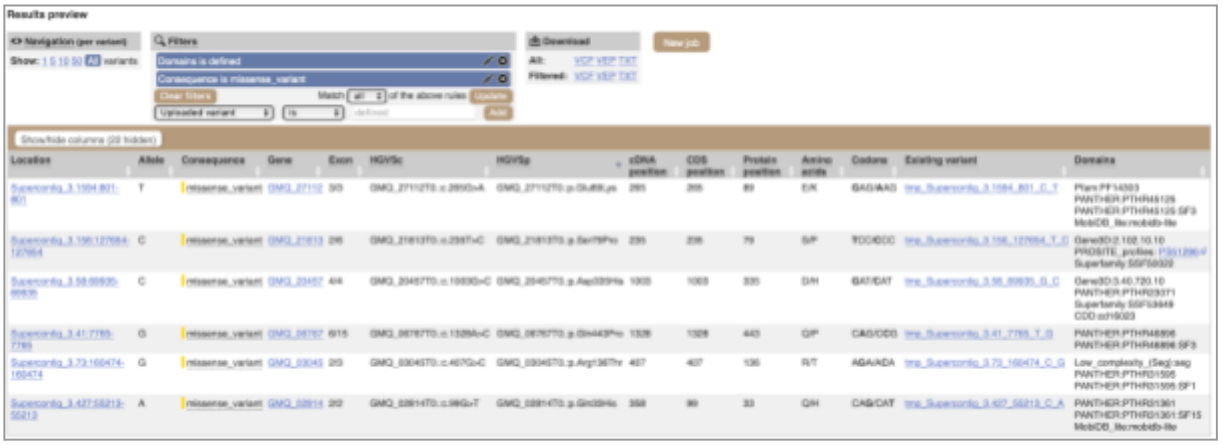
(a) How many variants have been processed?
(b) How many genes and transcripts are overlapped by variants in this file?



(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?

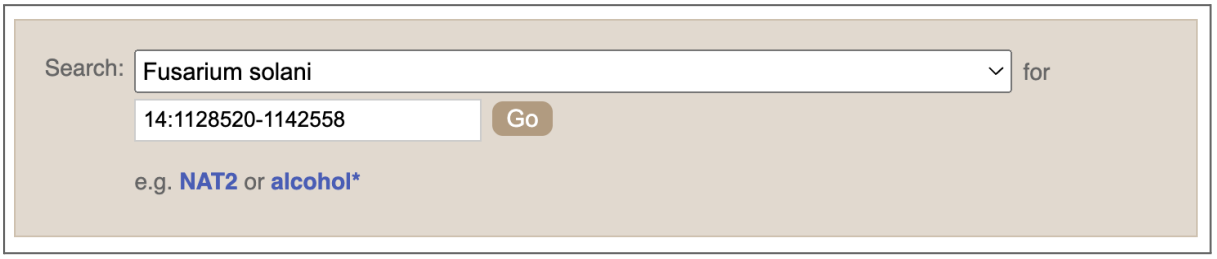


(e) How many variants are frameshift? Which gene(s) do they fall in and which exons? Can you find a UniParc ID of protein(s) affected by this variant?

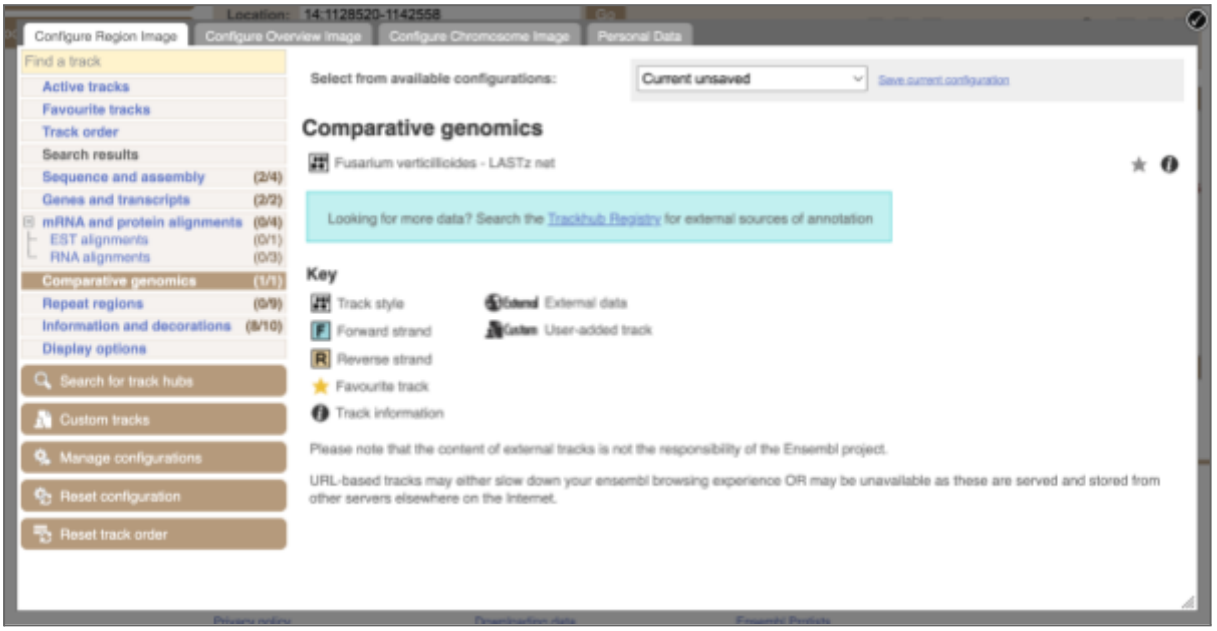
Exercise: Ensembl Fungi whole-genome alignments

Let's look at some of the comparative genomics views in the Location tab.

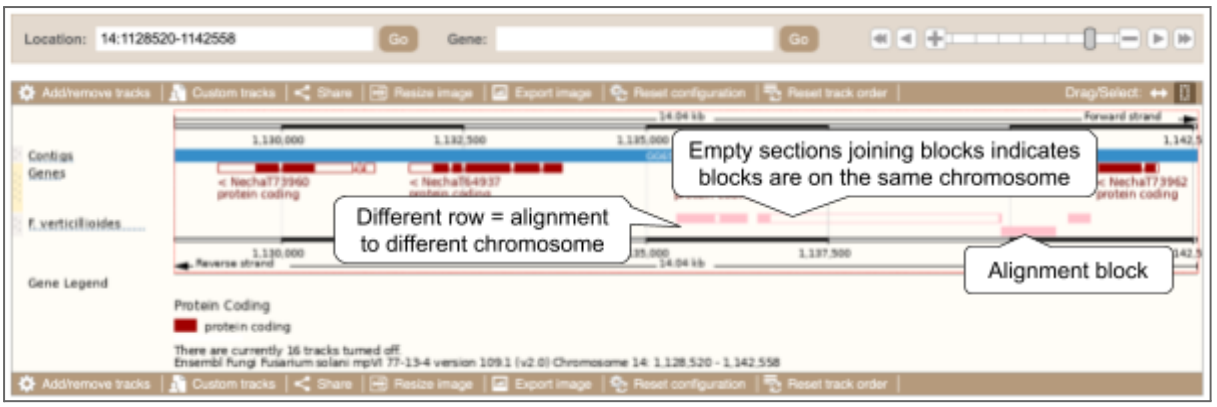
(a) Find the region 14:1128520-1142558 in *Fusarium solani* and go to the Region in detail page. This region includes four genes we identified from our first BioMart query: *PEP5*, *PDA1*, *ESP3* and *PEP5*.



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.

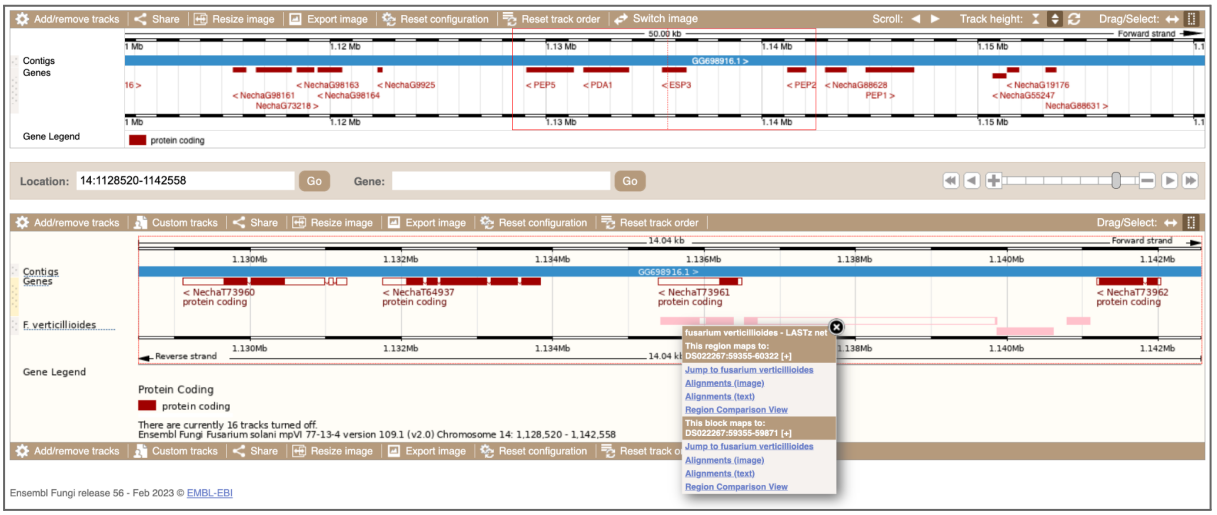


We can now see some pink alignments shown on the display. Alignments to the same chromosome are presented in a single row, and gaps in the alignment are shown by linking blocks. If there are alignments to multiple chromosomes in the aligned species these are represented on different rows.



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

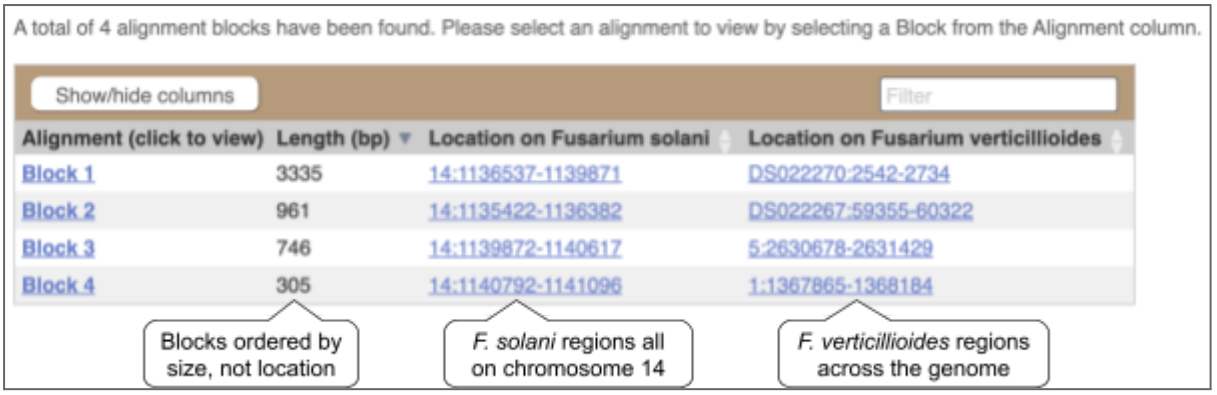
(c) Which chromosome(s) does the *F. solani ESP3* gene align to in *F. verticillioides*?



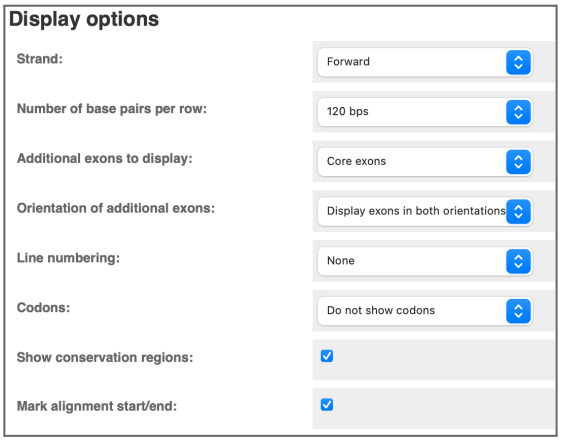
We can see that alignments in this region are quite poor for these species, with alignments spanning different chromosomes. This supports the lack of orthologues between these species.

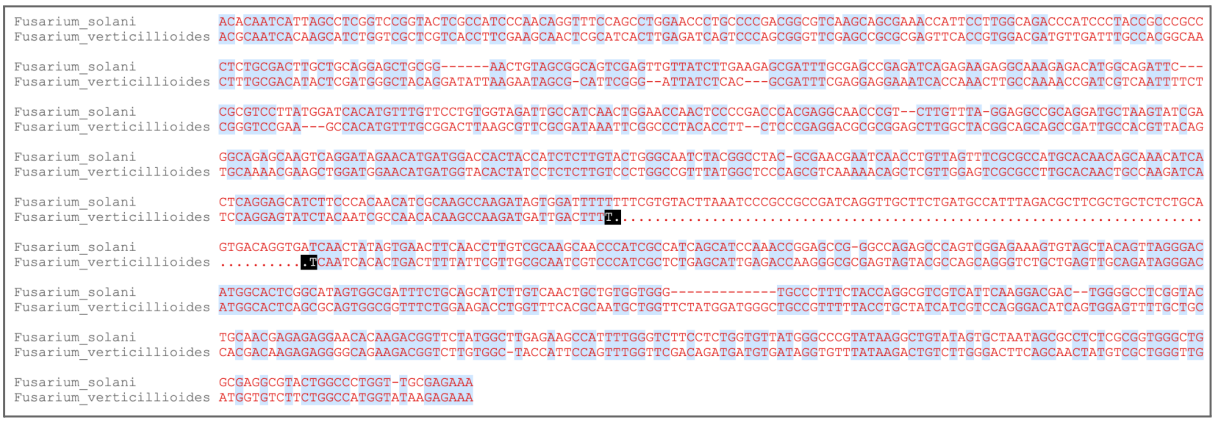
We can view more detailed alignments in the alignment's text / image and region comparison views. Let's first view a text alignment in this region. Click on Alignments (text) on the left and choose *Fusarium verticillioides* from the drop-down menu.

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.



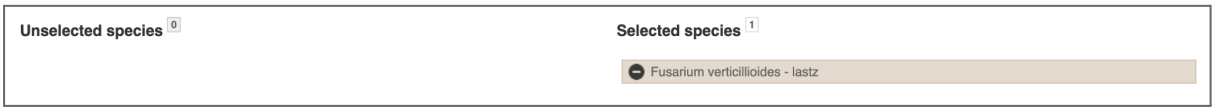
Let's click on Block 2. This takes you to a new page with a sample of the aligned sequence. Then click the button to Display full alignment. You will see a list of the regions aligned, followed by the sequence alignment. Exons are shown in red. Click on Configure this page, you can turn on the options to view Show conservation regions and Mark alignment start/end. This will add highlights where the sequence matches.



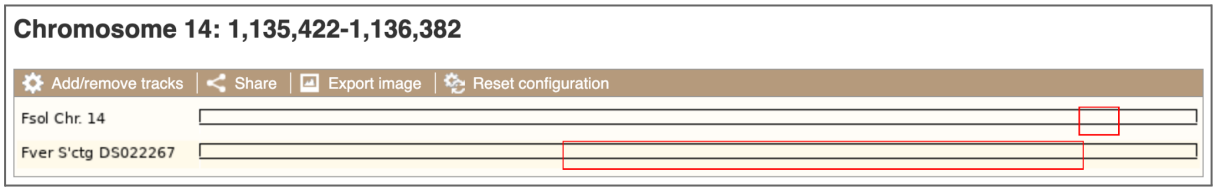


To view an image of the alignments, click on Region comparison in the left-hand navigation panel. This view is like the Region in Detail page as it shows three images of the genome at different scales. You can add multiple species to this view.

Click on the brown Select species or regions button. Choose *Fusarium verticillioides* species by clicking on the name. Close the window.

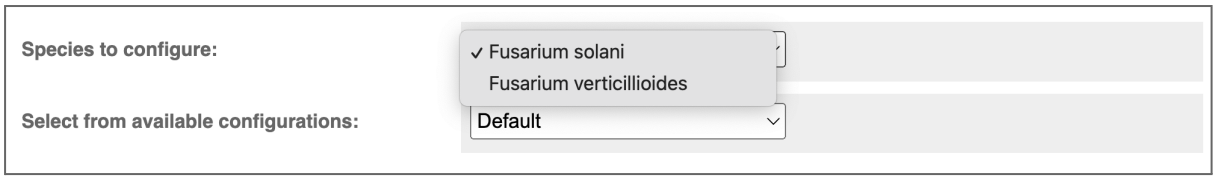


This page, similar to the region in detail page, shows the chromosome positions first. We can see the location of this alignment on the scaffold in *F. verticillioides*.

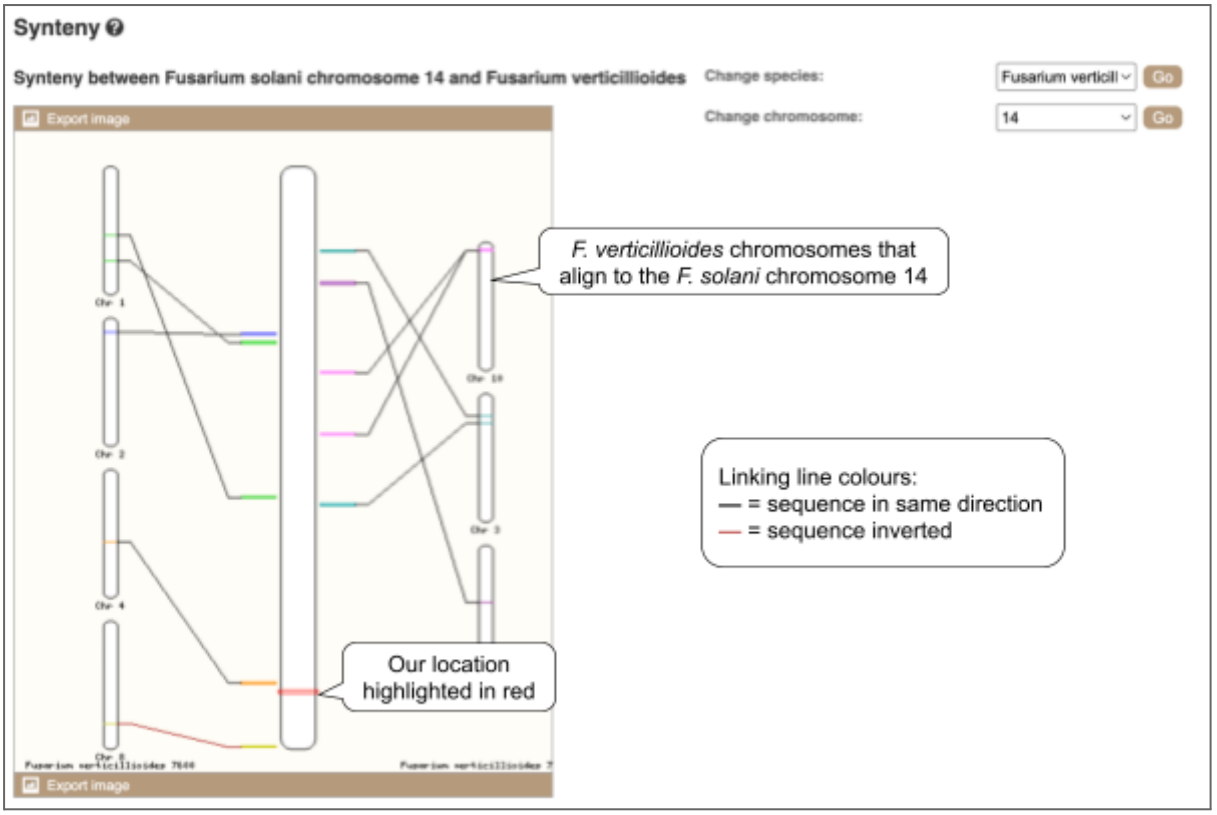


Scroll down to the most detailed image.

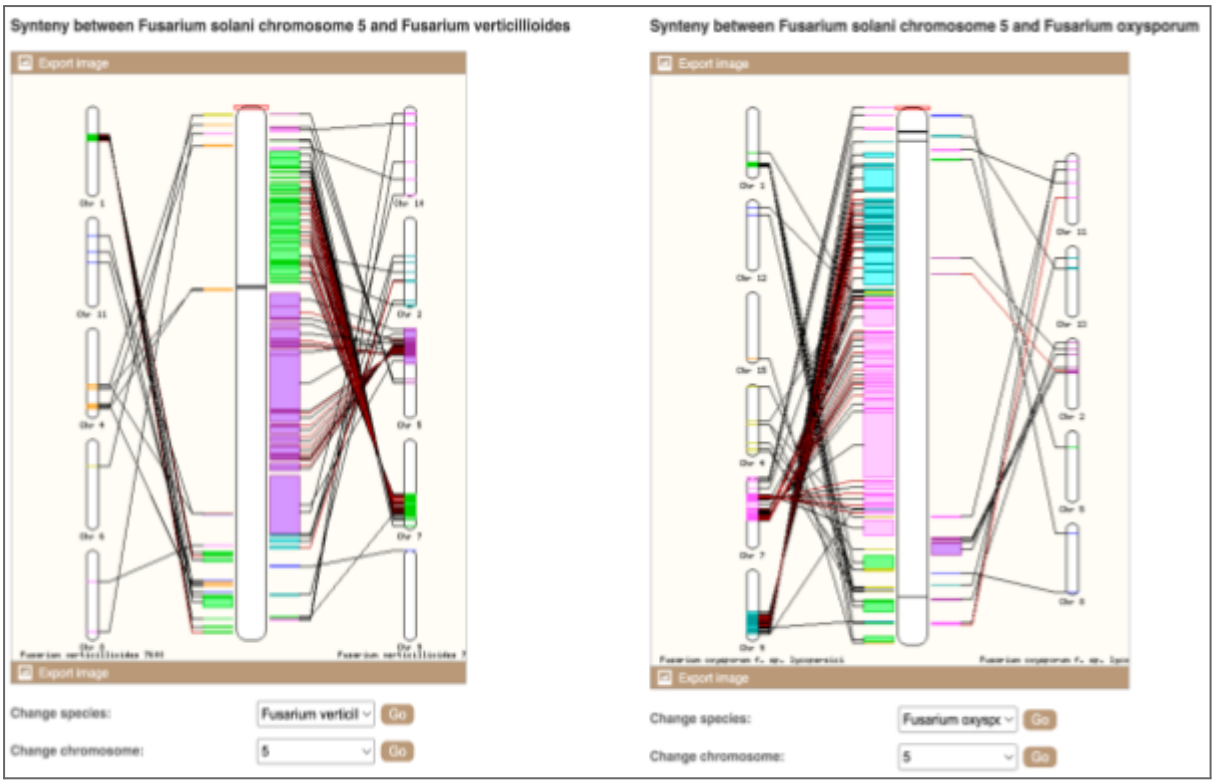
You can add data to both of these views with the same options you had in the Region in Detail page. Click on Configure this page and look at the top of the menu.



We can view chromosomal rearrangements in the Synteny view. Click on Synteny in the left-hand navigation panel.



(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*?



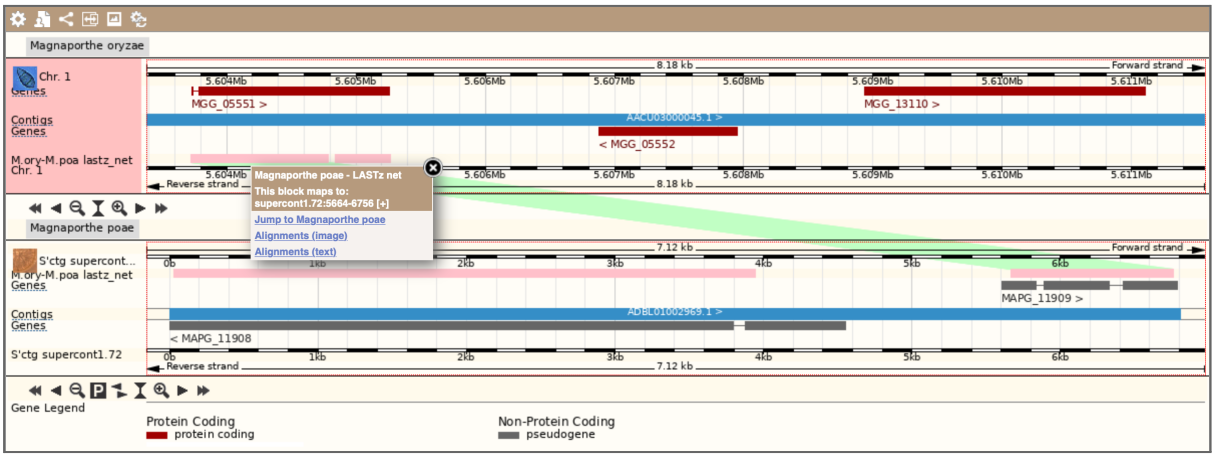
Additional Exercise - Rearrangements in *Magnaporthe* species

A recent paper Bao et al (2017) 'PacBio sequencing reveals transposable elements as a key contributor to genomic plasticity and virulence variation in *Magnaporthe oryzae*' 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?

(e) There is another alignment block in the *M. poae* display. Where does this region map to in *M. oryzae*?

MycoCosm: Comparative Analysis of Gene Families

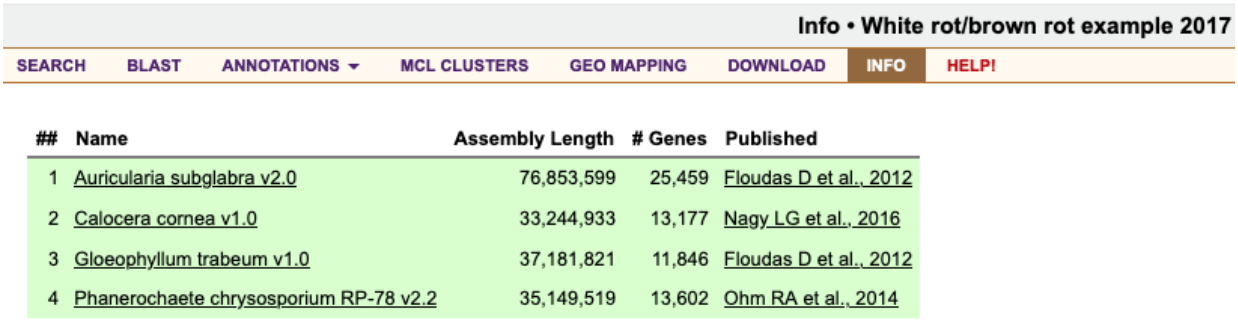
Objective: Compare genomes of wood decay fungi to identify gene families which can be used to distinguish white rot and brown rot fungi

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 the white rot fungi's genomes encode class II peroxidase enzymes (CAZy: AA2) to break down lignin; carbohydrate-binding motifs (CAZy: CBM1) to bind cellulose; and glycoside hydrolases of families 6 and 7 (CAZY: GH6 and GH7) to break down cellulose. The genome of a brown-rot fungus tends to lack genes encoding these enzymes, or have them in reduced numbers compared to white rot fungi.

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):

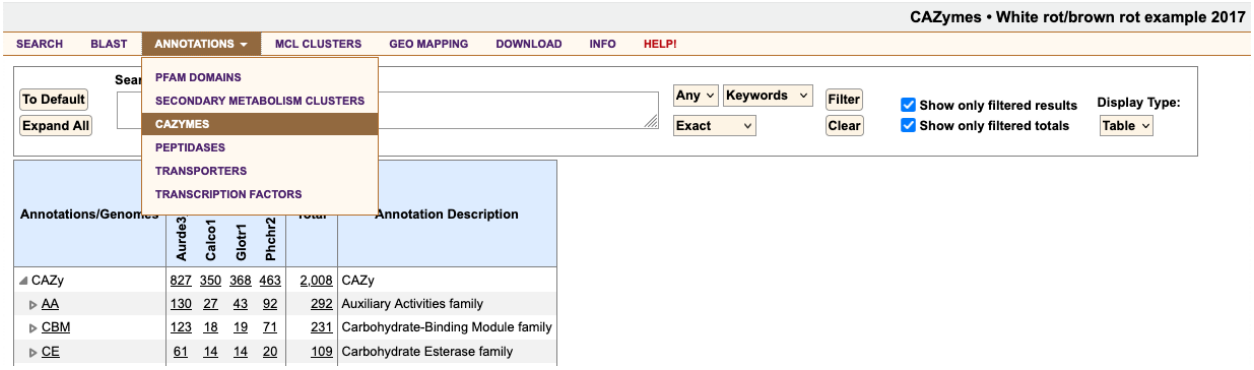
https://mycocosm.jgi.doe.gov/WR_BR_example_2017/



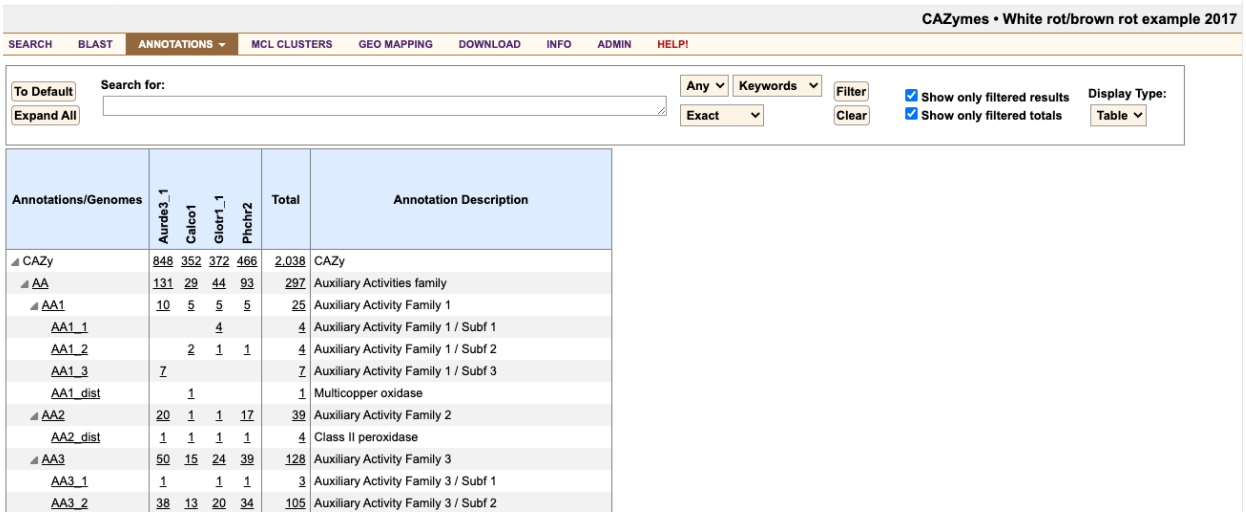
##	Name	Assembly Length	# Genes	Published
1	Auricularia subglabra v2.0	76,853,599	25,459	Floudas D et al., 2012
2	Calocera cornea v1.0	33,244,933	13,177	Nagy LG et al., 2016
3	Gloeophyllum trabeum v1.0	37,181,821	11,846	Floudas D et al., 2012
4	Phanerochaete chrysosporium RP-78 v2.2	35,149,519	13,602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

CAZy browser

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

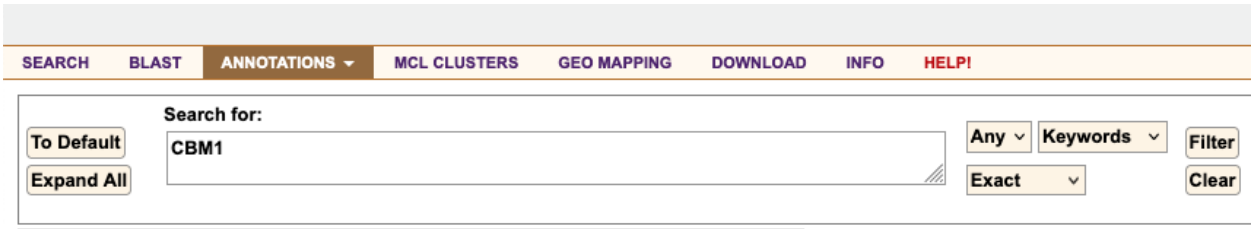


Here you will see a table representation of the predicted CAZymes (Levasseur et al. 2013). The organisms are labeled along the top. The CAZymes are organized by family and labeled along the sides. The numbers in the table tell you how many proteins from each organism's gene catalog were annotated with a given CAZyme. There is also a totals column. Notice that the CAZymes are hierarchically organized: you can see the total number of genes assigned to the general enzyme category (e.g. 'AA'). To expand top level assignment, click on the small arrow left of the category, or use the "Expand All" button at the top. Family designations ('AA1', 'AA2', etc.), and to subfamilies ('AA1_1', 'AA1_2', etc.) will then show up.



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 for AA2, *P. chrysosporium* and *A. subglabra* possess 20 and 17 copies of AA2, whereas *G. trabeum* and *C. cornea* possess only one AA2 copy each. This might suggest that the former two are white rot fungi and the latter two brown rot fungi!

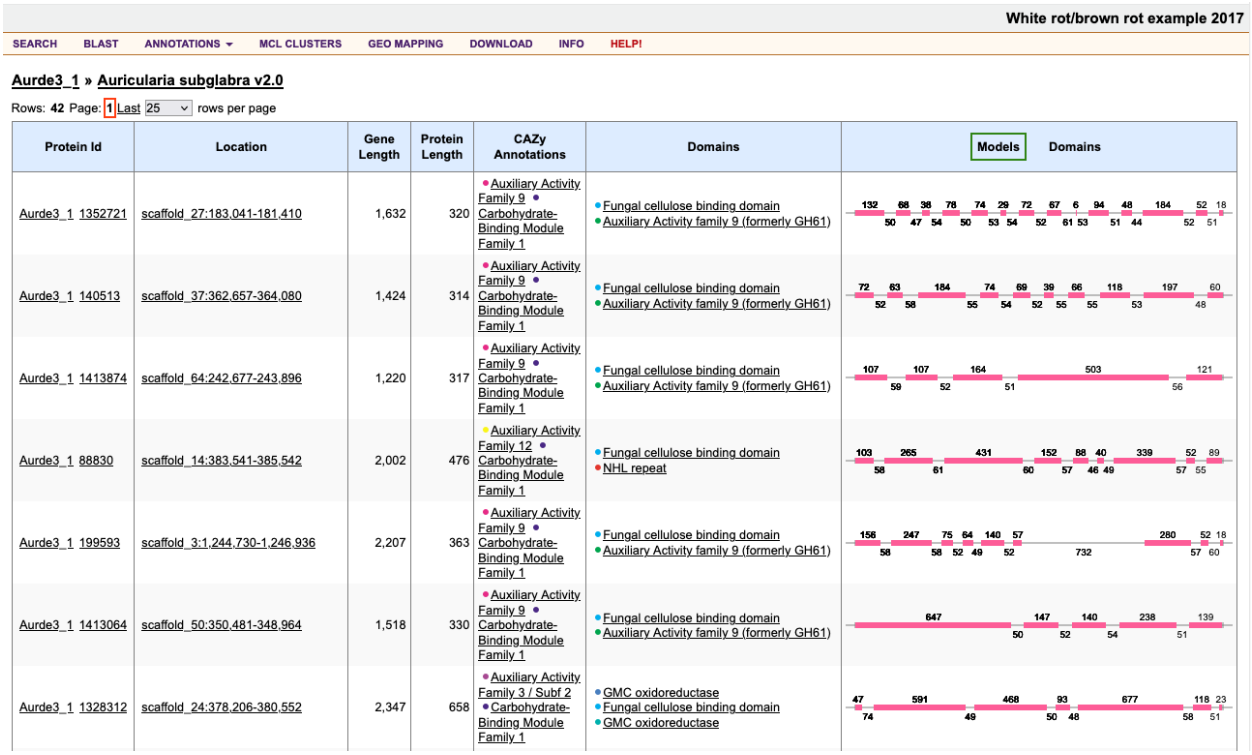
What about the carbohydrate binding motifs, CBM1? Let's say we don't want to scroll through the entire list of CAZymes. Type 'CBM1' into the 'CAZY terms' search box and select "Filter". This will limit the view to only those CAZymes that have a CBM1. Why do so many CAZymes besides CBM1 show up? Because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



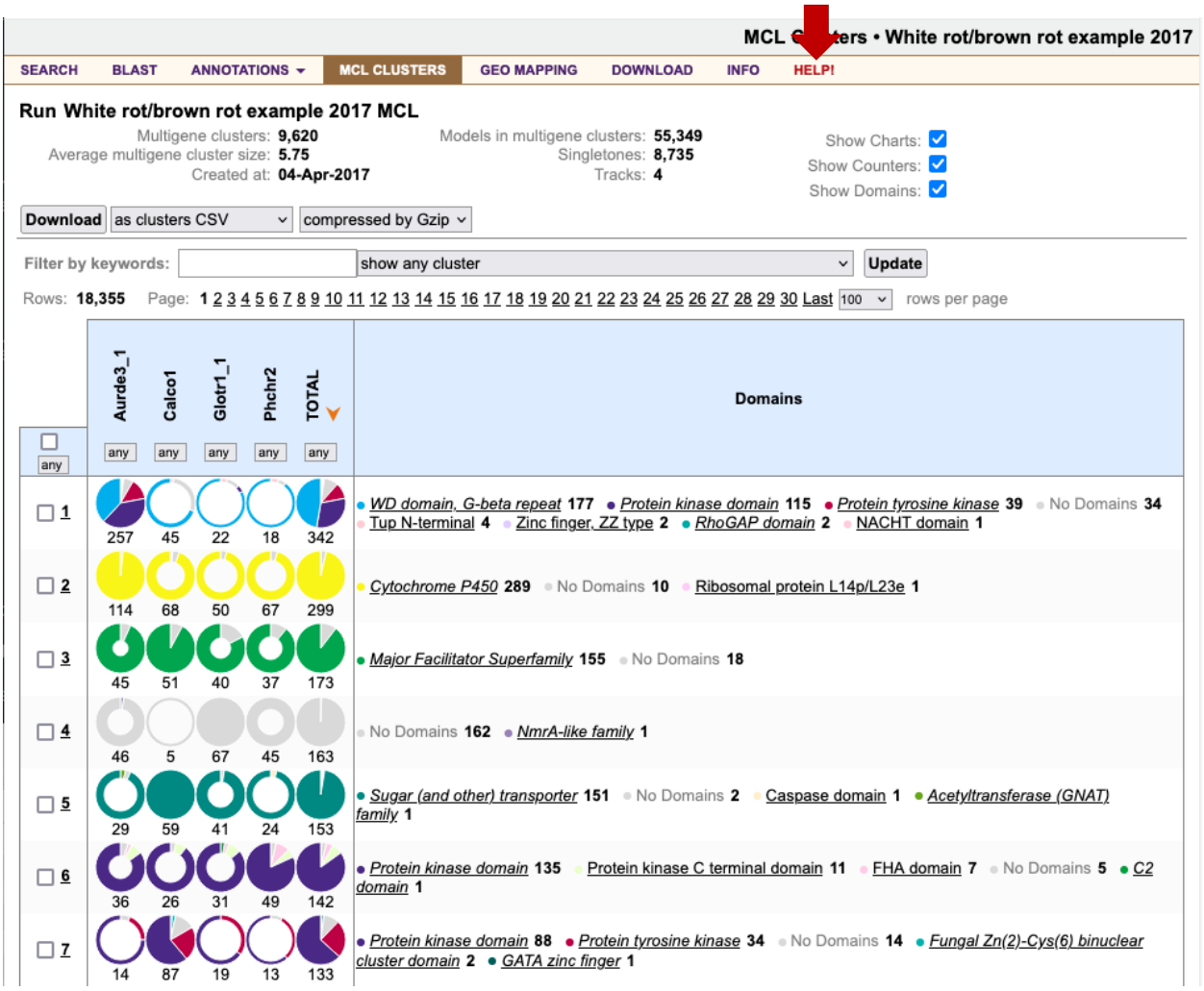
As an exercise, repeat the same search with GH6, GH7, and also the AA9 family of lytic polysaccharide monoxygenases, which may oxidatively act on lignin (Levasseur et al. 2013). Do the presence/absence patterns of these genes indicate the same conclusions about these fungi's mode of decay as we found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

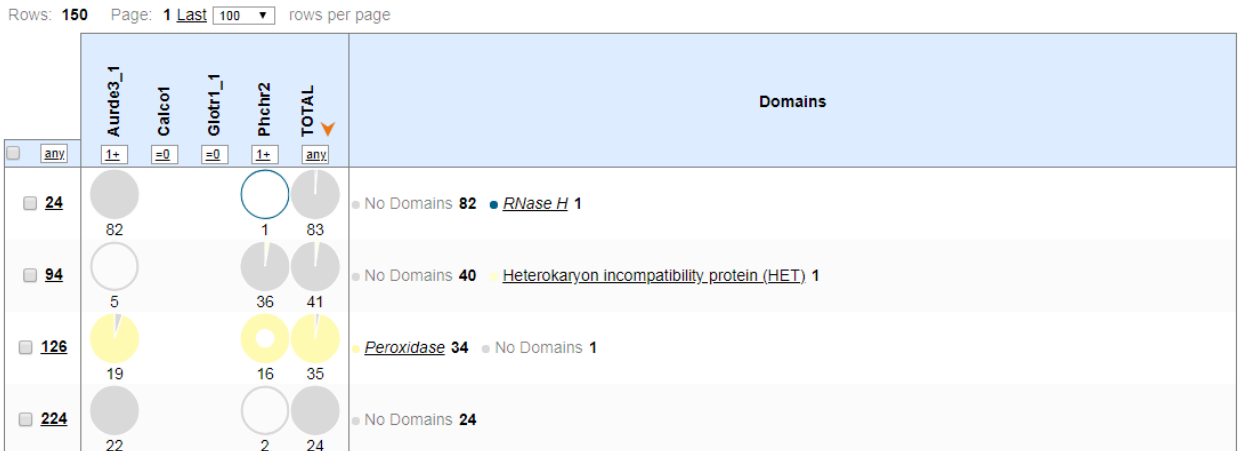
Cluster page

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 MCL algorithm (Enright et al. 2002). You can think of clusters as protein families. As with the CAZy browser, the columns indicate organisms. The rows indicate a

protein cluster, one cluster per row, with the number of proteins each organism contributes to a cluster. See the HELP Menu for a full explanation of the cluster page.



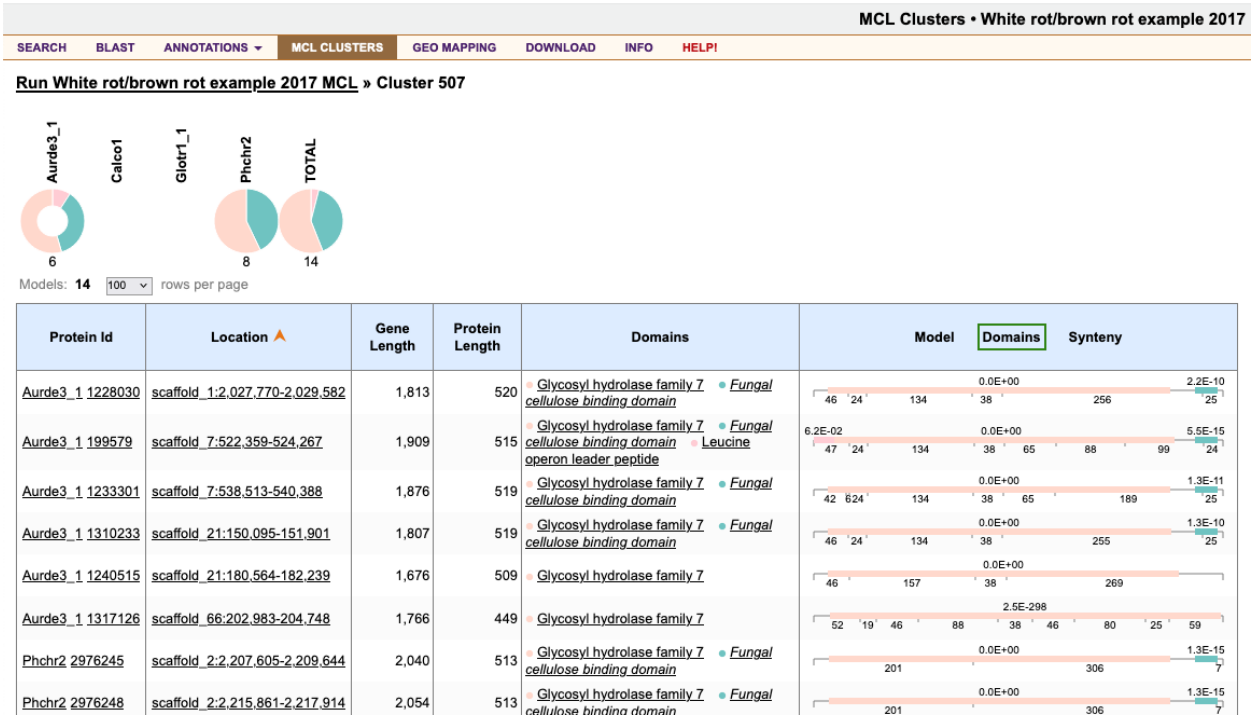
Notice that under each organism label is a button ‘any’ that can be used to filter clusters by the number of proteins that organism contributes to a cluster, and thus limit which clusters are shown. As an experiment, set the white rot fungi (Aurde3_1 and Phchr2) to “1+” and the brown rot fungi (Calco1 and Glotr1_1) to “=0”. Doing this returns 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. But 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 (http://pfam.xfam.org) 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.

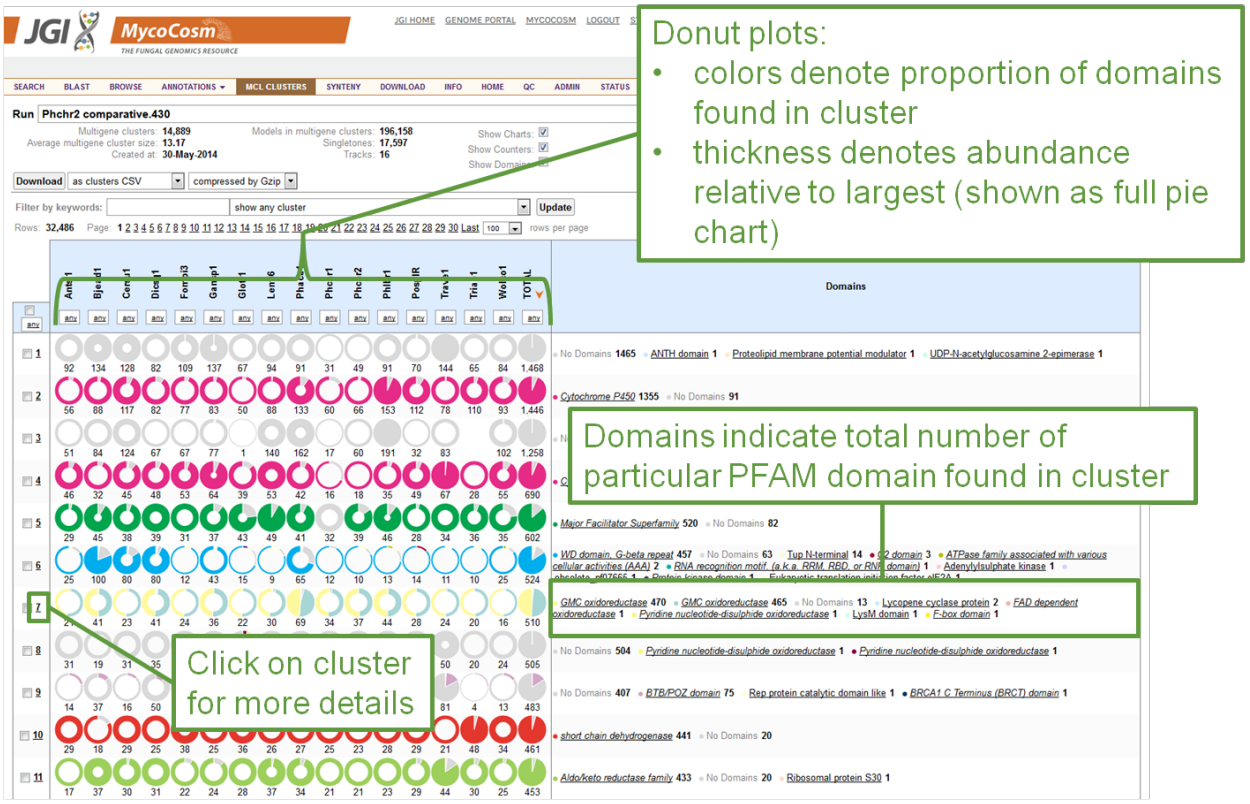
Scroll through the rest of the 150 clusters and you will see domains such as Glycosyl hydrolase family 7 and Fungal cellulose binding domain in cluster 507, which roughly overlap with the CAZy GH7 and CBM1 families. Click the '507' to explore that cluster in more detail. On the cluster detail page, a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



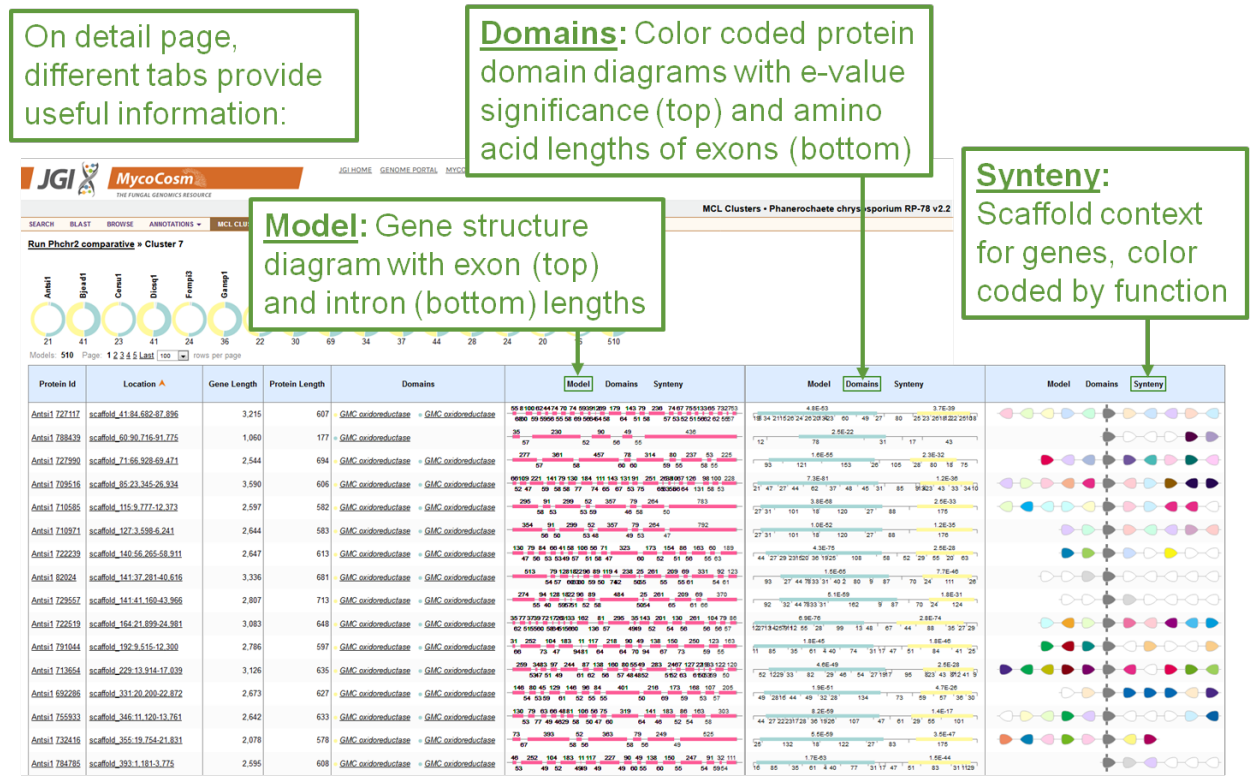
Let's look at what other proteins have the CBM1 carbohydrate-binding motifs in them. Returning to the cluster run page (click the "MCL CLUSTERS" tab). Enter the phrase "fungal cellulose binding domain" (be sure to include the quotes) into the "filter by keywords" field and select "Update". This returns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

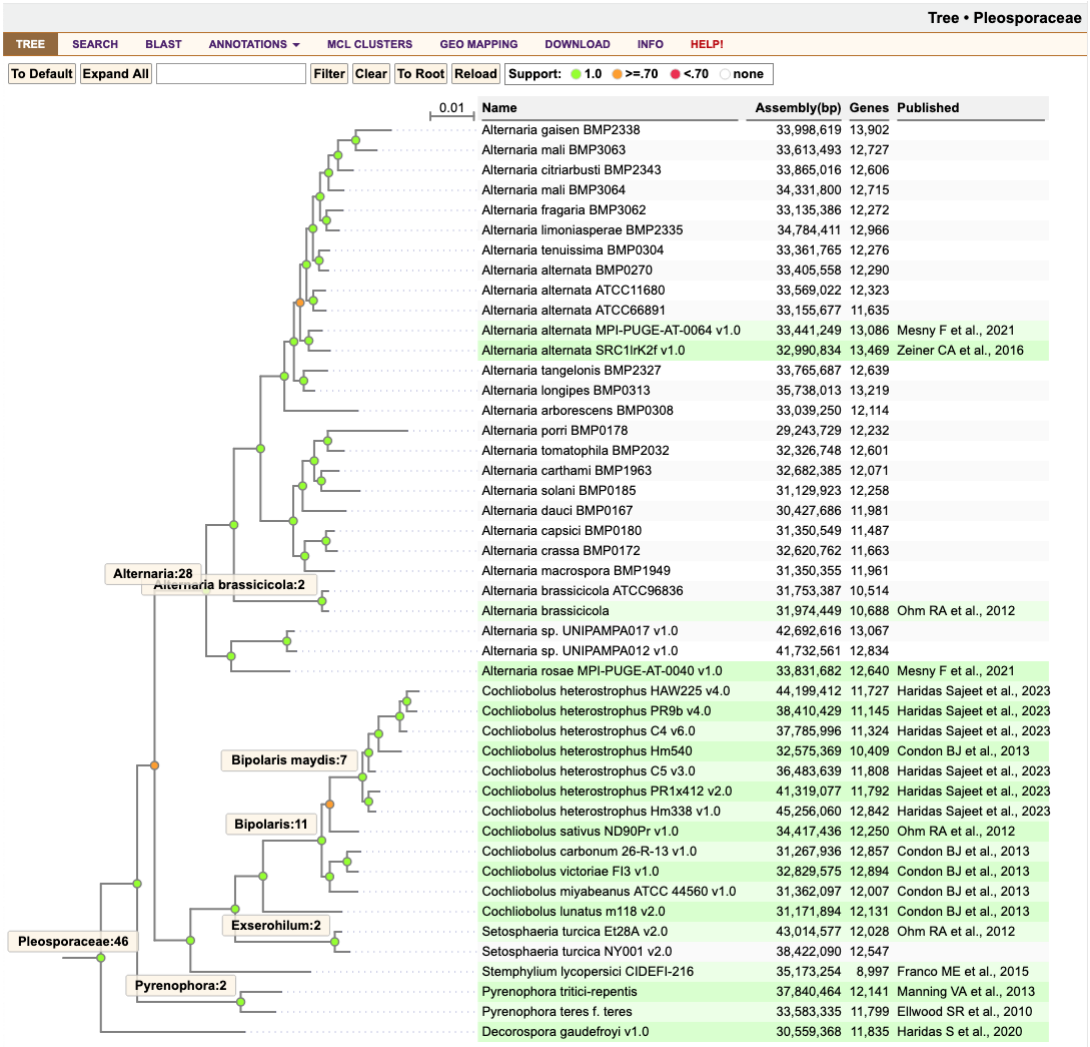
MycoCosm: Synteny Tutorial

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

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

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 gaisen BMP2338	33,998,619	13,902	
Alternaria mali BMP3063	33,613,493	12,727	
Alternaria citriarbusti BMP2343	33,865,016	12,606	
Alternaria mali BMP3064	34,331,800	12,715	
Alternaria fragaria BMP3062	33,135,386	12,272	
Alternaria limoniasperae BMP2335	34,784,411	12,966	
Alternaria tenuissima BMP0304	33,361,765	12,276	
Alternaria alternata BMP0270	33,405,558	12,290	
Alternaria alternata ATCC11680	33,569,022	12,323	
Alternaria alternata ATCC66891	33,155,677	11,635	
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 arborescens BMP0308	33,039,250	12,114	
Alternaria porri BMP0178	29,243,729	12,232	
Alternaria tomatophila BMP2032	32,326,748	12,601	
Alternaria carthami BMP1963	32,682,385	12,071	
Alternaria solani BMP0185	31,129,923	12,258	
Alternaria dauci BMP0167	30,427,686	11,981	
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 HAW225 v4.0	44,199,412	11,727	Haridas Sajeet et al., 2023
Cochliobolus heterostrophus PR9b v4.0	38,410,429	11,145	Haridas Sajeet et al., 2023
Cochliobolus heterostrophus C4 v6.0	37,785,996	11,324	Haridas Sajeet et al., 2023
Cochliobolus heterostrophus Hm540	32,575,369	10,409	Condon BJ et al., 2013
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 Hm338 v1.0	45,256,060	12,842	Haridas Sajeet et al., 2023
Cochliobolus sativus ND90Pr v1.0	34,417,436	12,250	Ohm RA et al., 2012
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 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

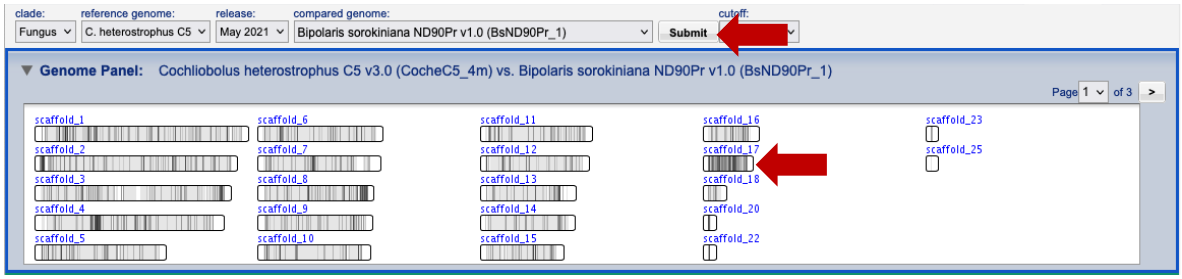
Note the green selection box while mousing over the tree. A left-click collapses and expands the selection box. You can also use shift+click to zoom into the selection. The browser back button does not work on the tree page. Click the TREE tab again to restore the default view.

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.

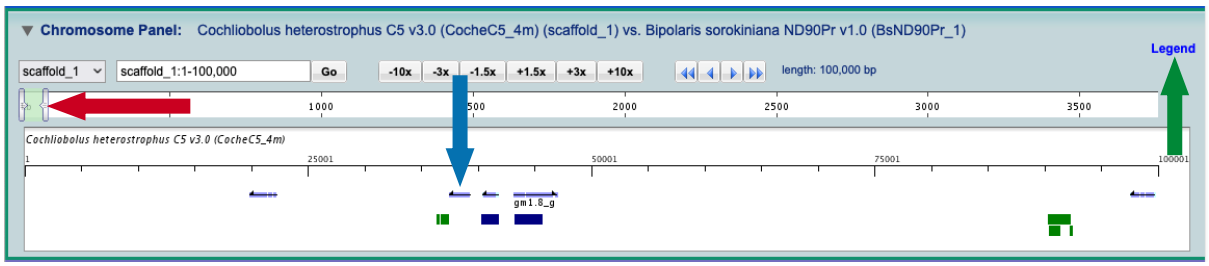
Click on the SYNTENY tab in the organism portal (*Cochliobolus heterostrophus* C5).

Genomic synteny is displayed in three collapsible panels in the Synteny Browser: the Genome Panel, the Chromosome Panel and the Comparison Panel.

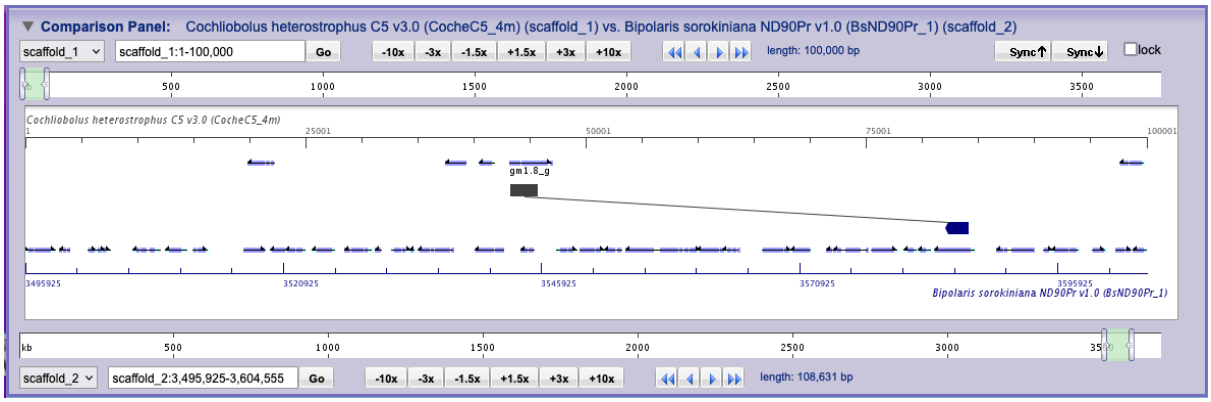
The compared genome can be changed from the dropdown menu and clicking "Submit". The Genome Panel depicts alignment density for all scaffolds in the reference genome against all chromosomes in the compared genome. Here, alignment density is defined for a region in the reference genome as the number of 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.



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 below) 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 below) 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.



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.

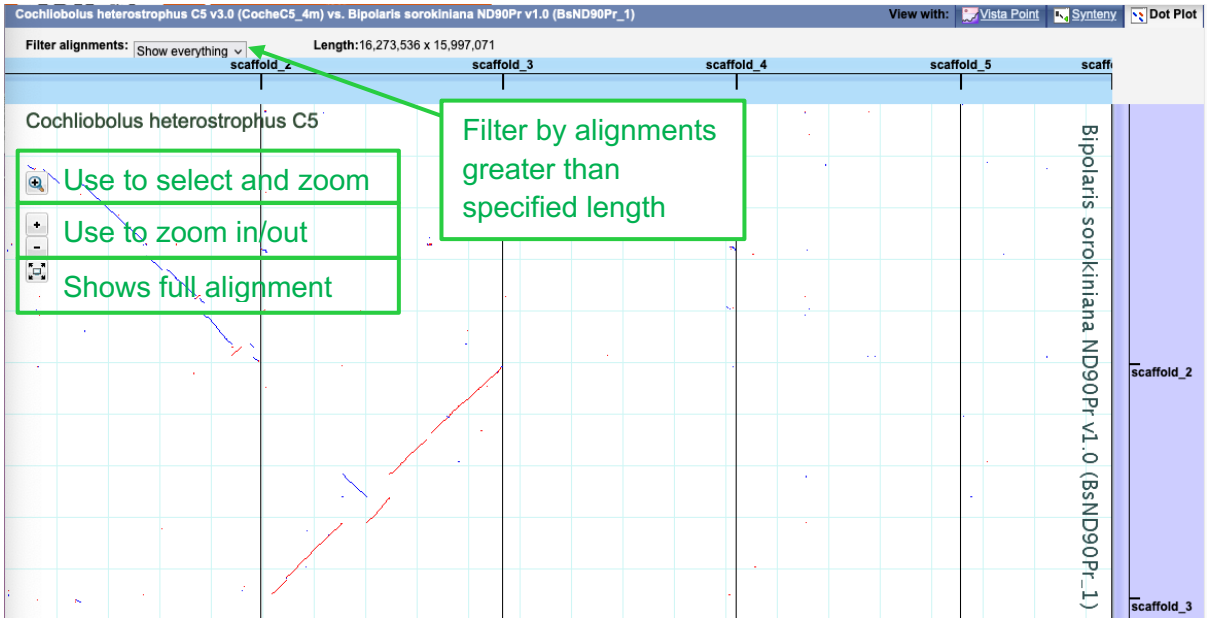


Syntenic blocks and gene models are both interactive, as described above for the Chromosome Panel. Navigation controls allow the user to switch chromosomes, zoom and pan independently over the reference and compared genomes. The SYNTENY page also allows whole genome pairwise comparison and comparison of one-to-many using the ‘Dot Plot’ and ‘Vista Point’ views respectively.

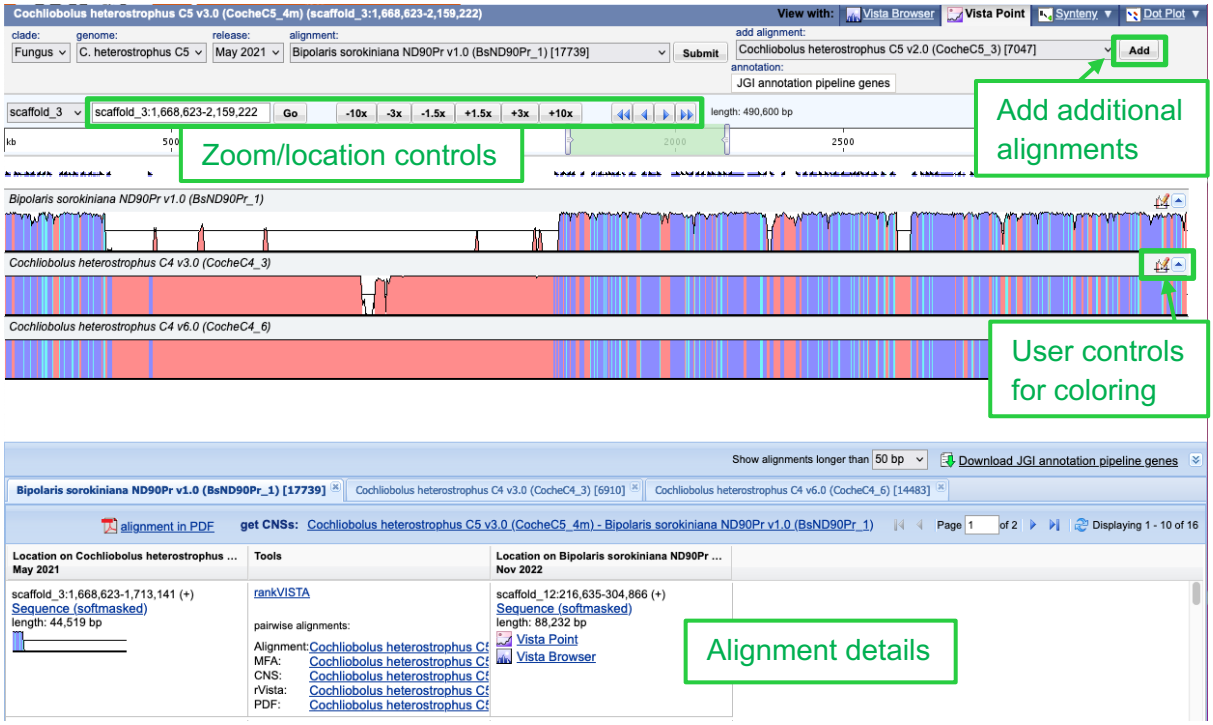
‘Dot Plot’ (VistaDot) is an interactive tool that enables users to look at the DNA conservation between two genome assemblies at different levels of resolution and across multiple chromosomes/scaffolds.

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.

Exploring protein domains and clusters across species in Ensembl and MycoCosm

We're going to use the HMMER tool, which is embedded in Ensembl Fungi with the Ensembl Fungi species as reference proteomes and standard search parameters. You can use the full tool here: https://www.ebi.ac.uk/Tools/hmmer/search/phmmer if you want to work with other species and set your own thresholds. This tool uses protein sequences as input and finds domains within them and maps against all species, like a more powerful, domain oriented BLAST. We'll then explore this at a finer scale looking at a handful of closely related species in the MycoCosm Cluster tool.

In the BioMart demo, we found genes that were associated with reduced virulence in Fusarium solani, and which did not have an orthologue in F. oxysporum. We're going to explore one of these genes in more detail: *PEP2* or NechaG73962.

a) Search *Fusarium solani* for NechaG73962 at fungi.ensembl.org. Navigate to the Transcript tab and either export the protein sequence in FASTA format or highlight and copy it.

b) Using the link in the Ensembl Fungi header navigate to the HMMER tool. Paste in your sequence and click Submit.
   I. What is the PFAM domain identified in this sequence?
   II. Hover over the domain image, what is the length of the aligned region with our submitted protein sequence?

c) In the Significant Query Matches table at the bottom of the page, click on the black Customise button and add 'Phylum' to the table.
   I. To which Phylum do the top hits belong to?
   II. Which species is reported as the top hit? Is this the same as Fusarium solani (think about the reproductive cycle...)?

d) We can explore the taxonomy more broadly elsewhere. Click on the Taxonomy tab just above the domain image.
   I. How many hits were there in the Basdiomycota?
   II. Click to expand the Agaricomycetes node by clicking on the arrow, and then the Agaricales. Which families are represented?

   NOTE: You may need to click on the node name (e.g. Agaricales), to reposition the image.

e) Let's explore this protein domain further using the JGI MycoCosm cluster tool. This will enable us to compare the presence genes containing this domain across a closely related group. Navigate to mycocosm.jgi.doe.gov and search for *Fusarium solani*. Select Fusarium solani FSSC 5 v1.0, then click on the MCL clusters option at the top of the page. Search for the protein domain we identified, SnoaL_4.
   I. For the first cluster, 4,213, which species is missing any hits?
   II. Click on the link for the domain (4213) to explore in more detail. We can see from the donut plots that only 4 species contain this SnoaL-like domain.
   III. Which species have the most similar protein lengths, and contain the SnoaL-like domain?

f) Click on Synteny in the final column.
   I. Are the clusters around this conserved between these species? Click on them to highlight the occurrence in the image.
   II. Which other species has the most similar synteny to *Fusarium solani* (Fusso1) protein 495533? Why do you think this is?

Answers

Exploring protein domains and clusters across species in Ensembl and MycoCosm

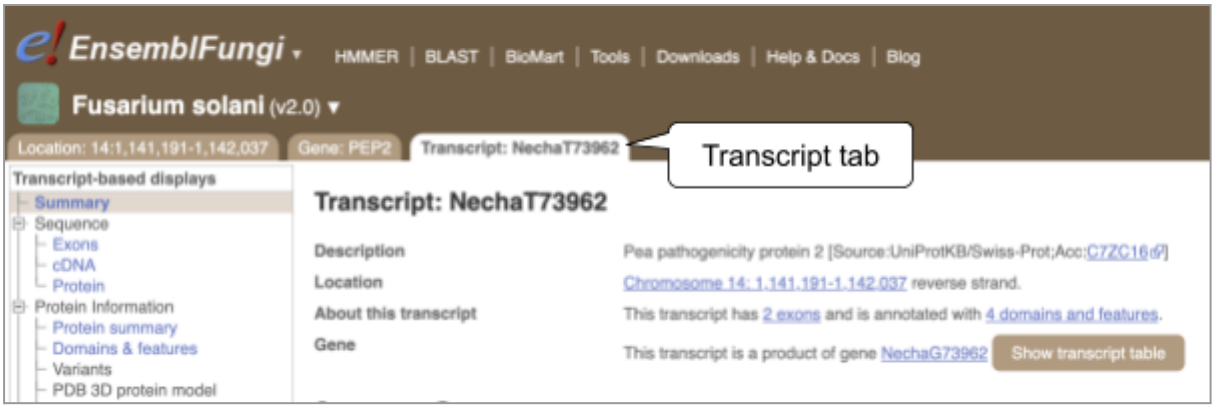
We're going to use the HMMER tool, which is embedded in Ensembl Fungi with the Ensembl Fungi species as reference proteomes and standard search parameters. You can use the full tool here: https://www.ebi.ac.uk/Tools/hmmer/search/phmmer if you want to work with other species and set your own thresholds. This tool uses protein sequences as input and finds domains within them and maps against all species, like a more powerful, domain oriented BLAST. We'll then explore this at a finer scale looking at a handful of closely related species in the MycoCosm Cluster tool.

In the BioMart demo, we found genes that were associated with reduced virulence in Fusarium solani, and which did not have an orthologue in F. oxysporum. We're going to explore one of these genes in more detail: *PEP2* or NechaG73962.

g) Search *Fusarium solani* for NechaG73962 at fungi.ensembl.org. Navigate to the Transcript tab and either export the protein sequence in FASTA format, or highlight and copy it.

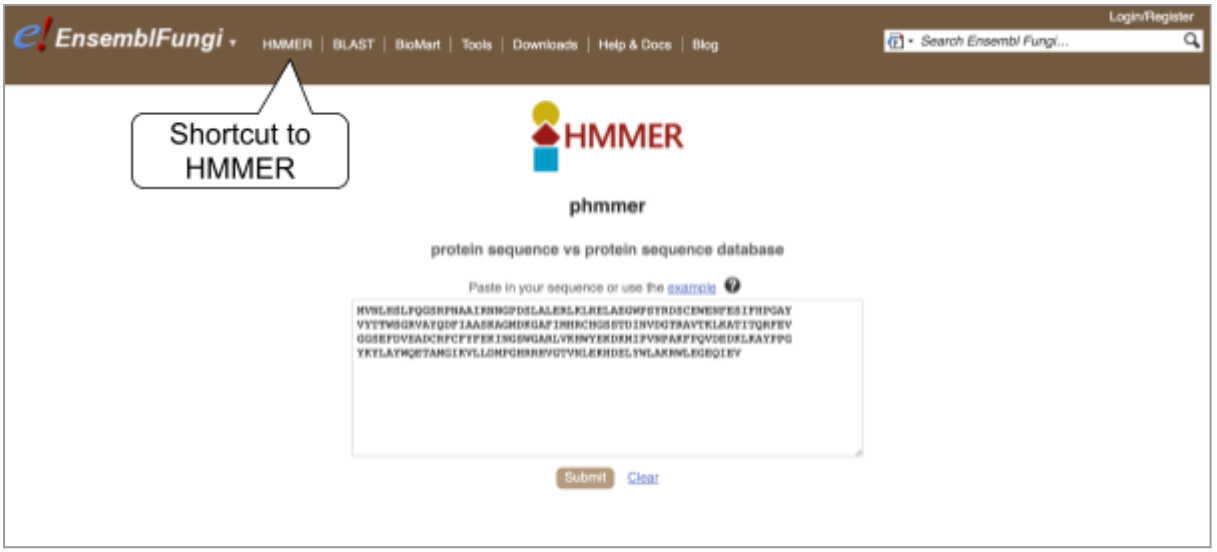
Answer: Go to fungi.ensembl.org. From the homepage select *Fusarium solani* from the drop-down list and type in NechaG73962. Hit Go.

Click on the gene name hyperlink on the results page, this will take you to the gene tab. Click on the transcript tab Transcript: NechaT73962 to go to the transcript tab.



On the left-hand navigation panel there is a link for Protein under the Sequence header. Highlight the protein sequence and copy it.

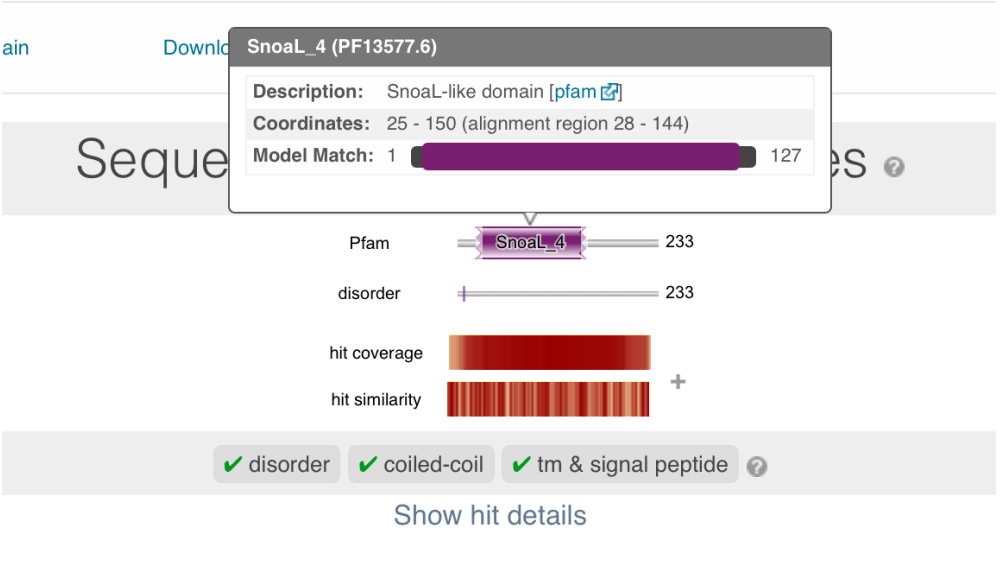
Using the link in the Ensembl Fungi header navigate to the HMMER tool. Paste in your sequence and click Submit.



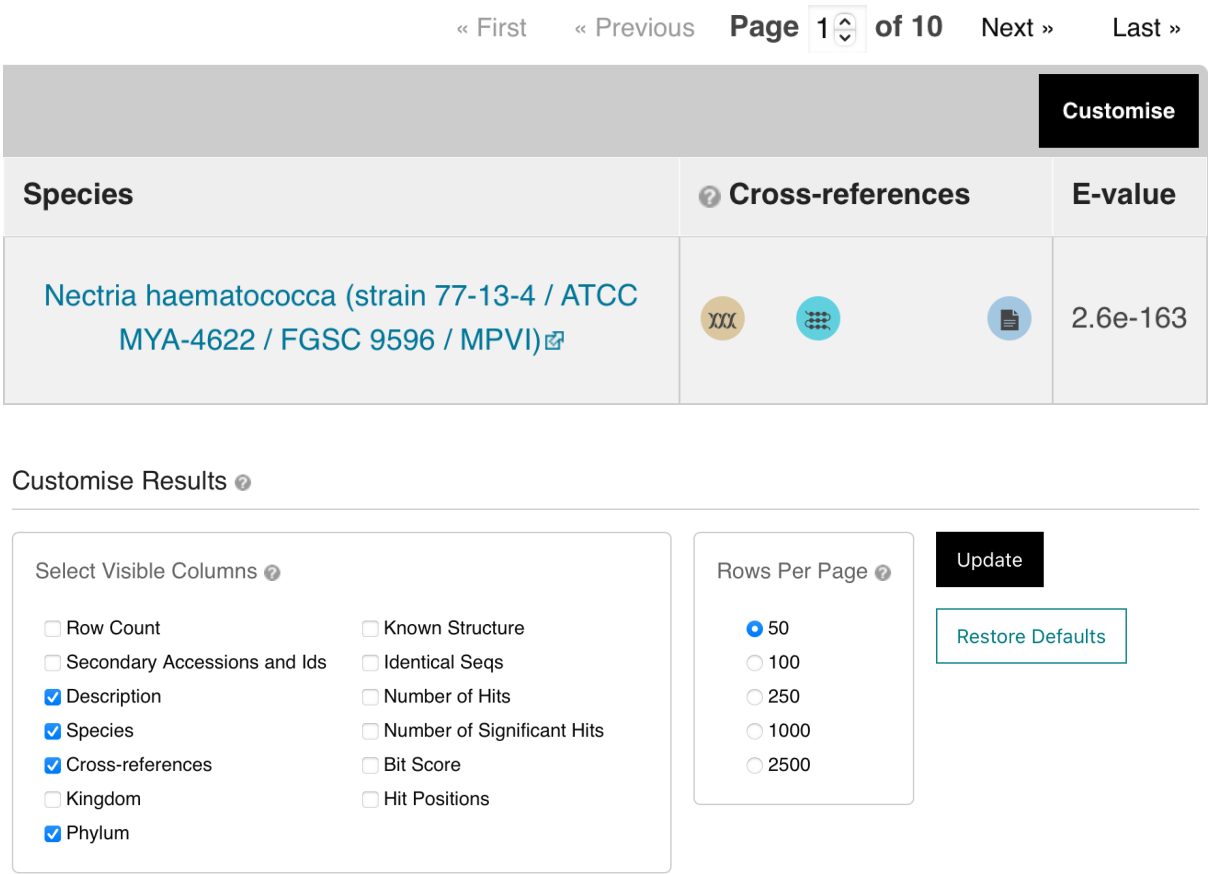
I. What is the PFAM domain identified in this sequence?

II. Hover over the domain image, what is the length of the aligned region with our submitted protein sequence?

Answers: The image shown in the centre middle of the page shows the domain (or domains) matched in your sequence. Hovering over the domain will give you some summary information, including the length of the overlapping sequence.



h) In the Significant Query Matches table at the bottom of the page, click on the black Customise button and add ‘Phylum’ to the table.



I. To which Phylum do the top hits belong to?

Answer: We can see that the column of the first hits are all listed as 'Ascomycota'

II. Which species is reported as the top hit? Is this the same as Fusarium solani (think about the reproductive cycle...)?

Answer: The sexual form (teleomorph) of Fusarium solani (the anamorph) is Nectria haematococca.

« First « Previous Page 1 of 10 Next » Last »

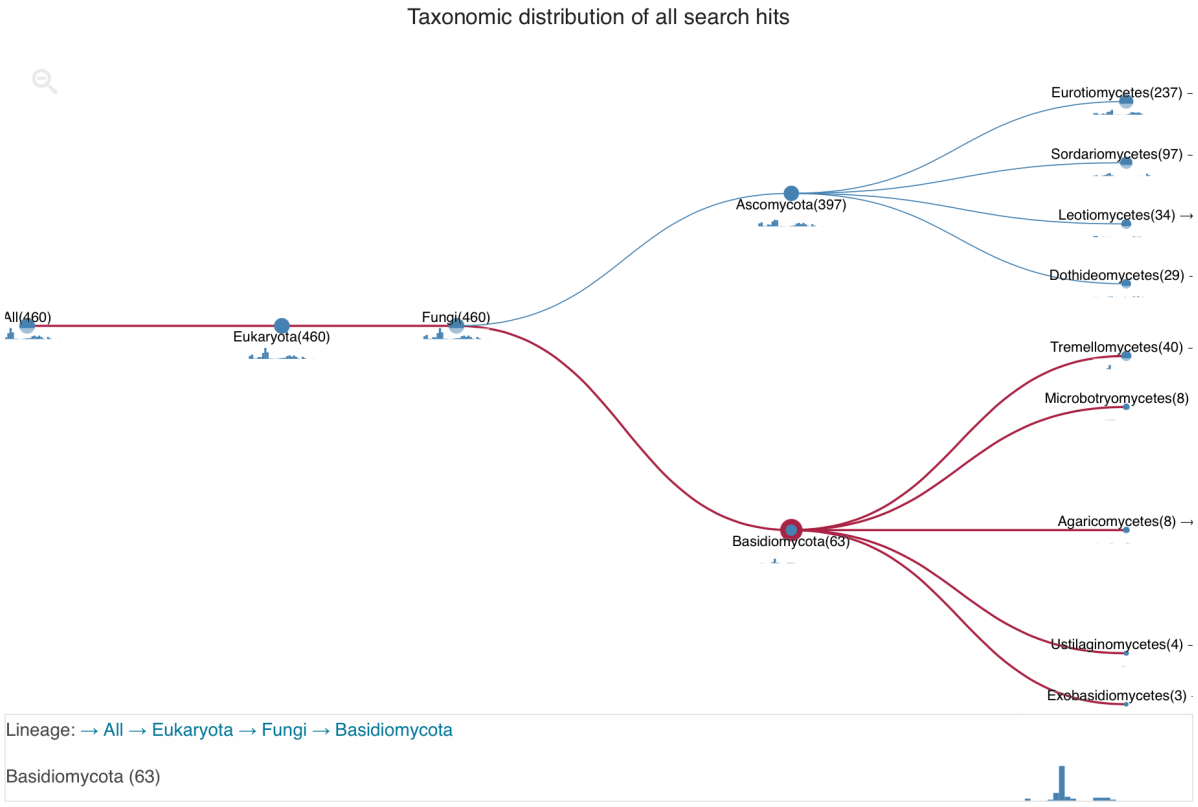
Significant Query Matches (460) in *ensemblgenomes* (v.44) Customise

	Target	Description	Phylum	Species	Cross-references	E-value
>	NechaG73962	Pea pathogenicity protein 2 [Source:UniProtKB/Swiss-Prot;Acc:C7ZC16]	Ascomycota	Nectria haematococca (strain 77-13-4 / ATCC MYA-4622 / FGSC 9596 / MPVI)		2.6e-163
>	LW93_4799	Uncharacterized protein	Ascomycota	Gibberella fujikuroi		1.6e-137
>	FFB14_04603	Pea pathogenicity protein 2	Ascomycota	Fusarium fujikuroi (GCA_900096505)		2.1e-137
>	AU210_001920	hypothetical protein	Ascomycota	Fusarium oxysporum f. sp. radicis-cucumerinum		6.2e-137
>	FOWG_10080	pea pathogenicity protein 2	Ascomycota	Fusarium oxysporum f. sp. lycopersici MN25 (GCA_000259975)		6.2e-137

i) We can explore the taxonomy more broadly elsewhere. Click on the Taxonomy tab just above the domain image.

I. How many hits were there in the Basdiomycota?

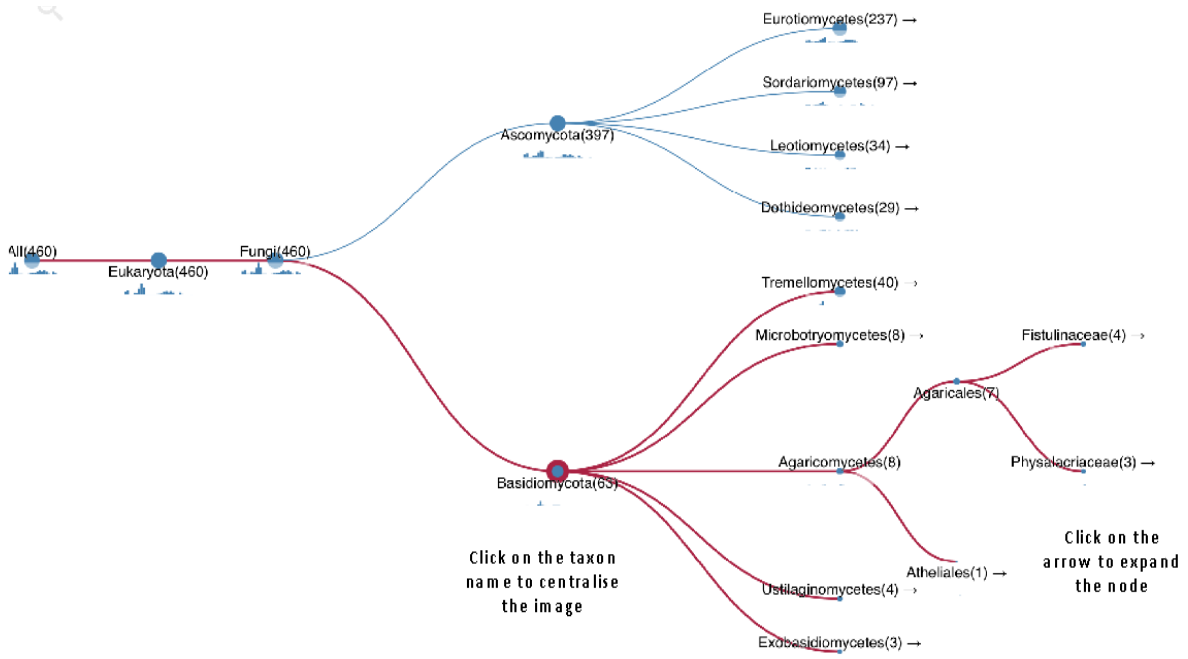
Answer: We can see from the number in the parentheses that there are 63 hits.



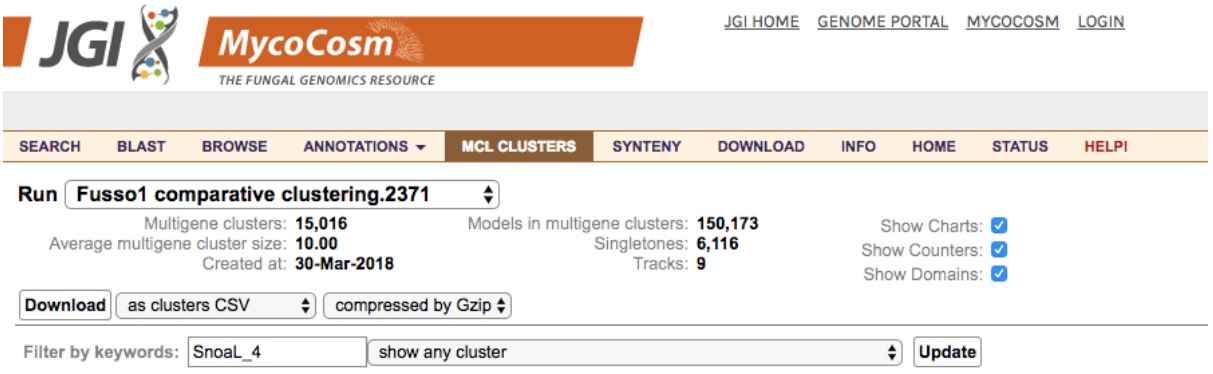
II. Click to expand the Agaricomycetes node by clicking on the arrow, and then the Agaricales. Which families are represented?

Answer: Fistulinaceae and Physalacriaceae families are shown here with 4 and 3 members respectively.

NOTE: You may need to click on the node name (e.g. Agaricales), to reposition the image.

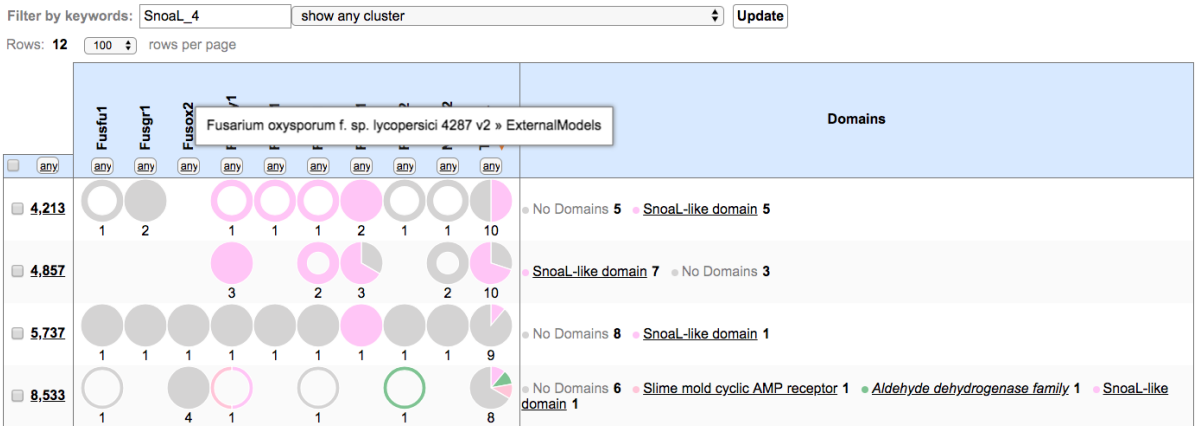


j) Let's explore this protein domain further using the JGI MycoCosm cluster tool. This will enable us to compare the presence genes containing this domain across a closely related group. Navigate to mycocosm.jgi.doe.gov and search for *Fusarium solani.* Select Fusarium solani FSSC 5 v1.0, then click on the MCL clusters option at the top of the page. Search for the protein domain we identified, SnoaL_4.



I. For the first cluster, 4,213, which species is missing any hits?

Answer: There is no 'donut' in the first row for the species Fusox2. Hover over the name or look at the list below the table to see what this species/assembly full name is, it is *Fusarium oxysporum* f. sp. lycopersici 4287 v2 ExternalModels.



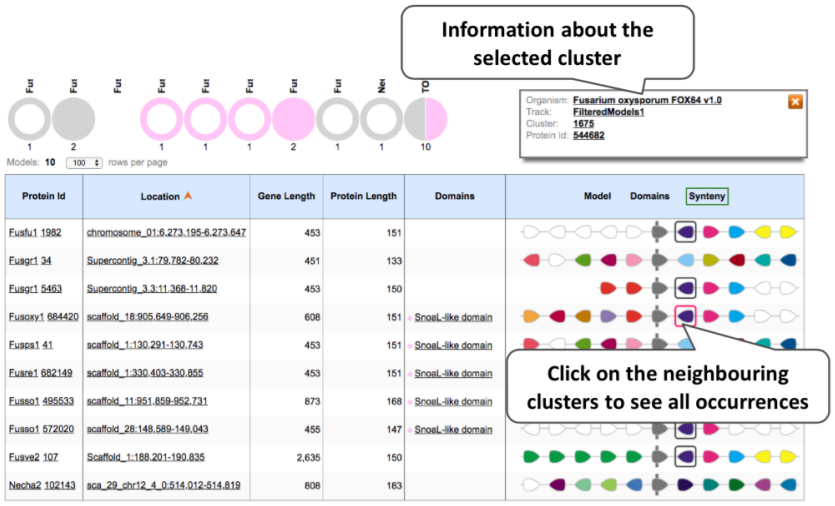
II. Click on the link for the domain (4213) to explore in more detail. We can see from the donut plots that only 4 species contain this SnoaL-like domain.

Answer: The pink colour corresponds to the SnoaL-like domain.

III. Which species have the most similar protein lengths, and contain the SnoaL-like domain?

k) Click on Synteny in the final column.

I. Are the clusters around this conserved between these species? Click on them to highlight the occurrence in the image.



II. Which other species has the most similar synteny to *Fusarium solani* (Fusso1) protein 495533? Why do you think this is?

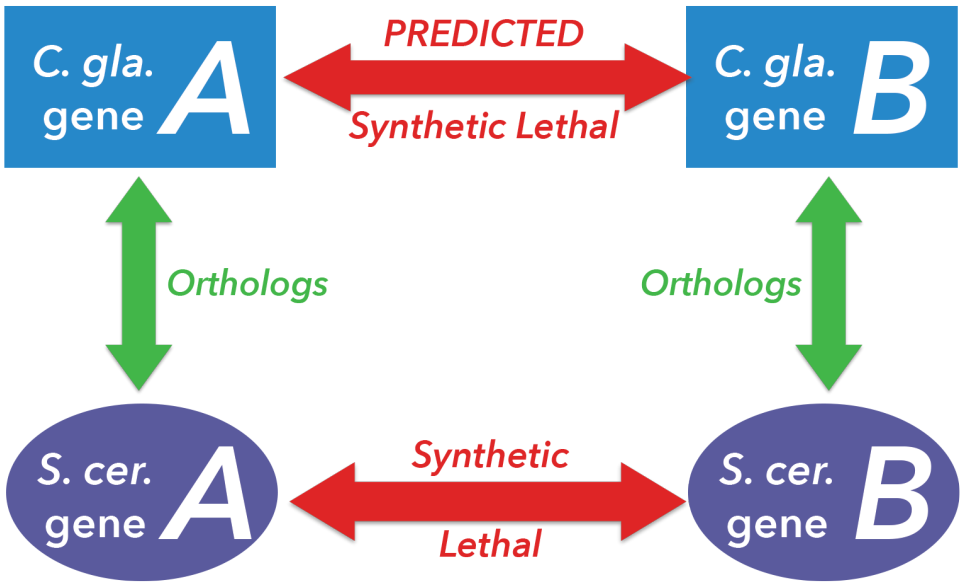
Answer: *Nectria haematococca* v2.0 FilteredModels1. We know this to be the sexual form of *F. solani* so this is expected.

Using *S. cerevisiae* Orthologs of *Candida glabrata* Genes 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.

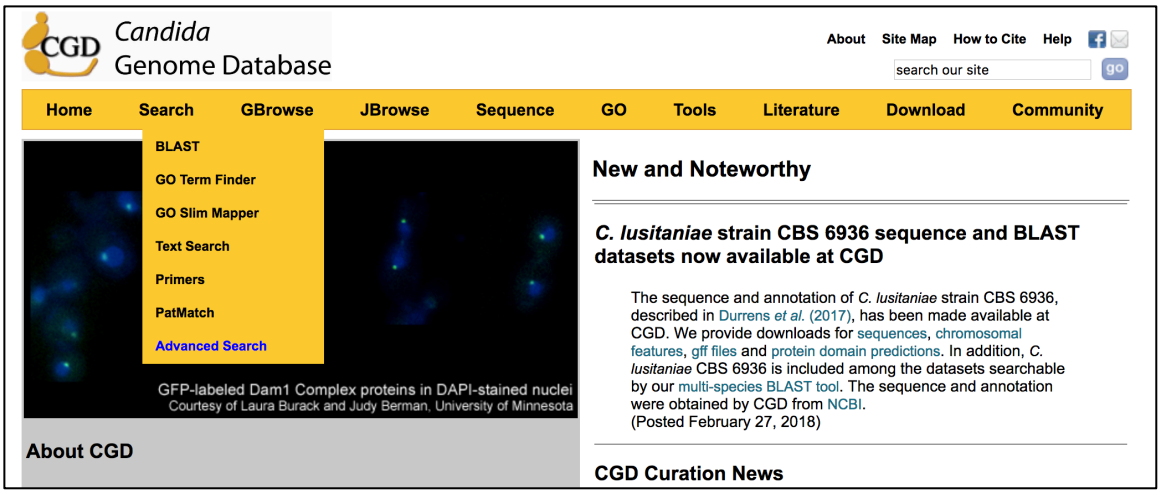
Genetic interactions such as synthetic lethality are richly documented for the budding yeast *S. cerevisiae*, but not as much for many other fungal species. By examining known genetic interactions in *S. cerevisiae*, we can predict synthetic lethal relationships in *C. glabrata* and other fungal pathogens.



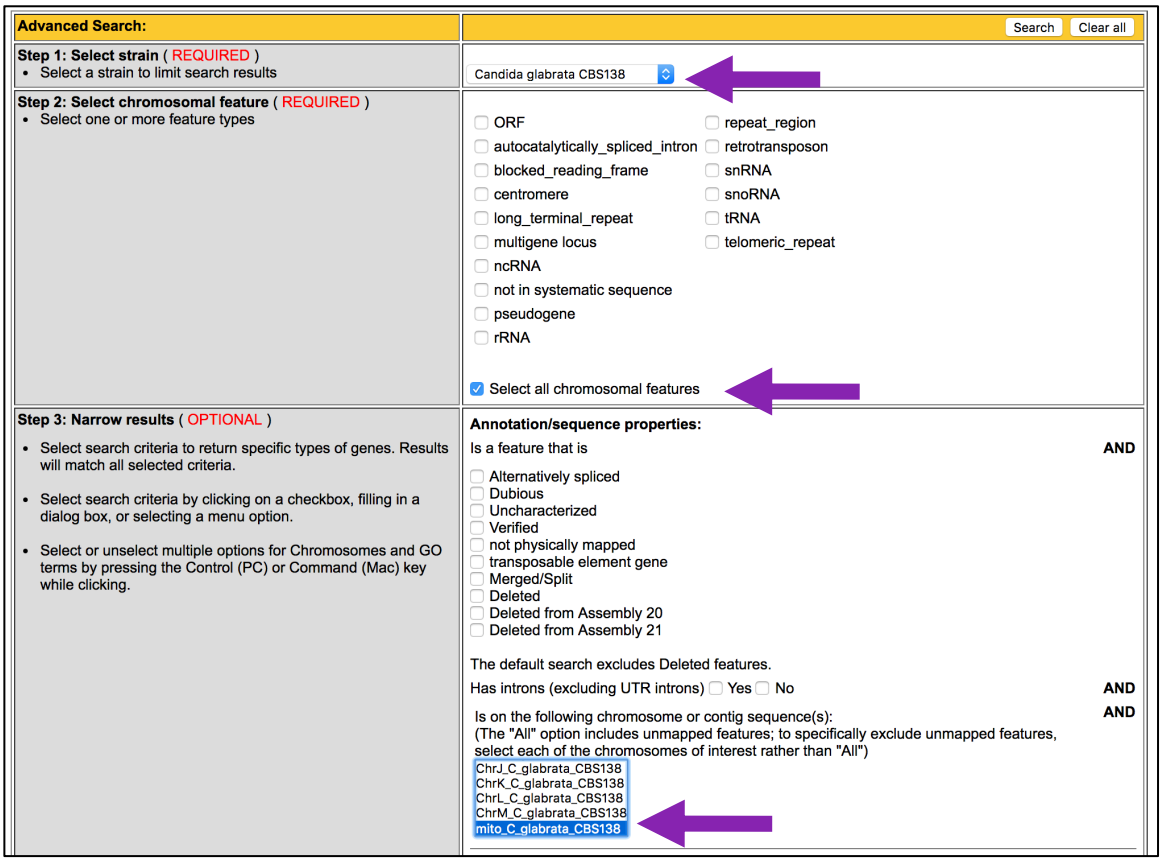
If conserved, these synthetic lethal interactions may reveal future antifungal targets for use against azole-resistant strains in the clinic. Using known synthetic lethal interactions in the *S. cerevisiae* genome allows prediction of potentially conserved synthetic lethal interactions for mitochondrial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

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

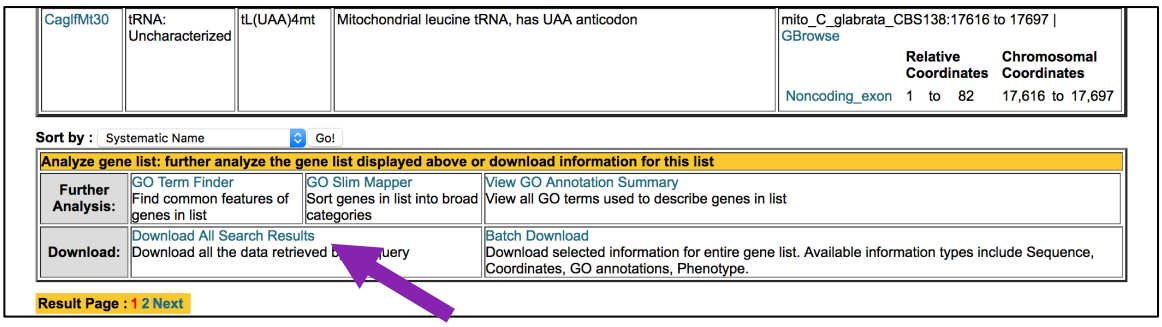
- On the CGD homepage (**http://www.candidagenome.org**), open the Search tab in the yellow toolbar and select **Advanced Search**.



- 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.

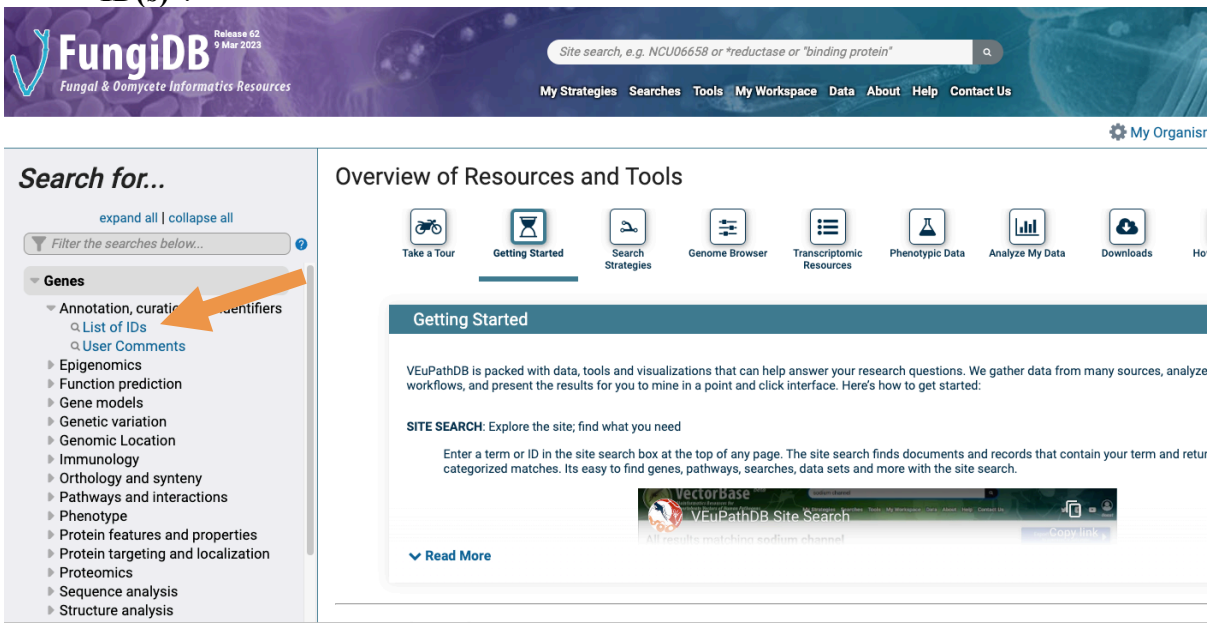


- Click on "Search". A results page will follow, listing out 37 features in the *C. glabrata* mitochondrial genome.
- Scroll to the bottom of the page and click on the "**Download All Search Results**" link.

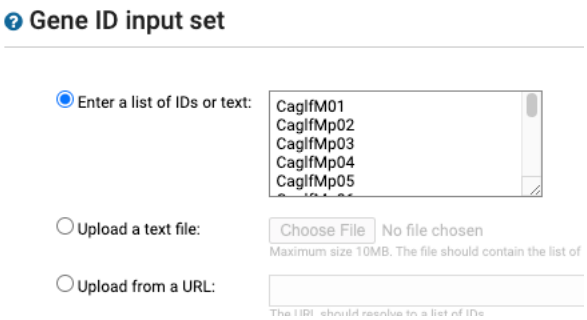


2. Use FungiDB to find *S. cerevisiae* orthologs of *C. glabrata* mitochondrial genes:

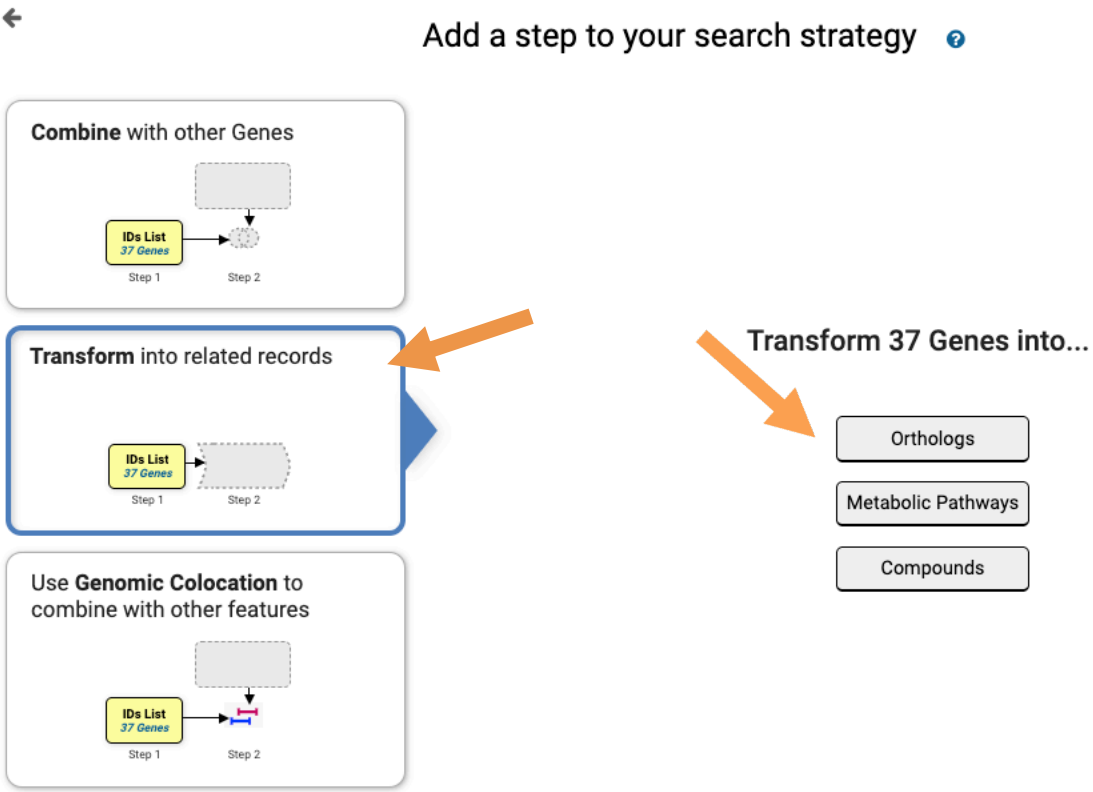
- Open the FungiDB homepage (**http://fungidb.org/**). In the "**Search for Genes**" box, open the "**Annotation, curation and identifiers**" section and click on "**List of ID(s)**".



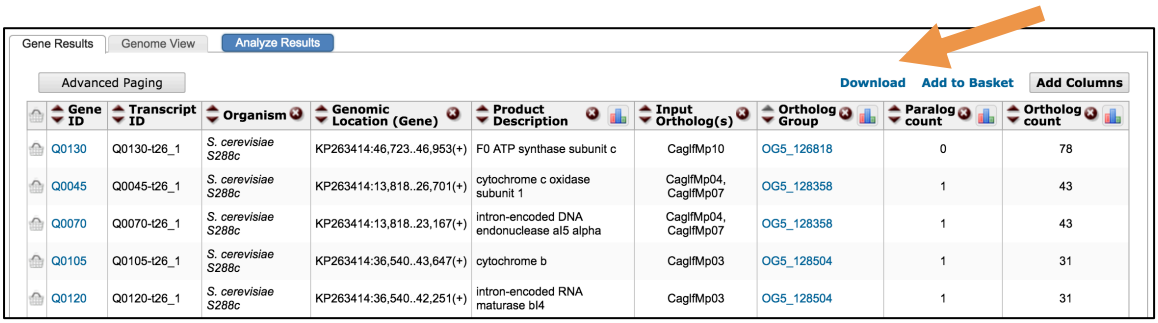
- Using your exported file from CGD, copy and paste the ORF names of the *C. glabrata* mitochondrial genes into the box. Click on "**Get Answer**".



- 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 on the right select "**Orthologs**."



- In the "**Organism**" list, type in "cerevisiae" to search and then select "Saccharomyces cerevisiae S288C", 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.

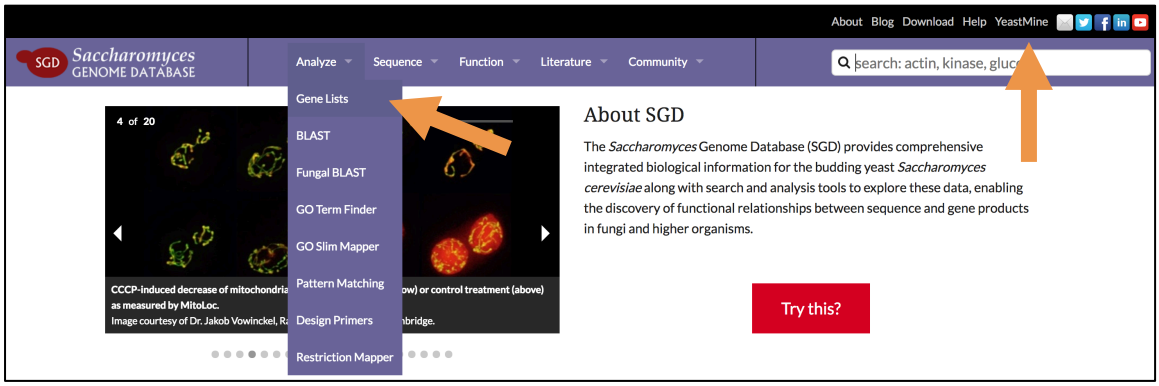


Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Input Ortholog(s)	Ortholog Group	Paralog count	Ortholog count
Q0130	Q0130-t26_1	*S. cerevisiae S288c*	KP263414:46,723..46,953(+)	F0 ATP synthase subunit c	CaglfMp10	OG5_126818	0	78
Q0045	Q0045-t26_1	*S. cerevisiae S288c*	KP263414:13,818..26,701(+)	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

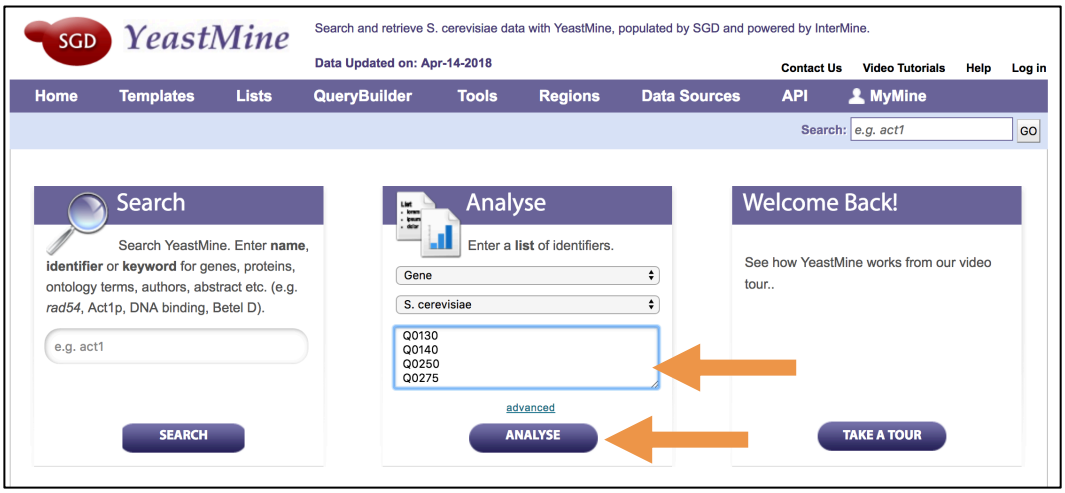
- In the download options menu, select "**Tab- or comma-delimited (openable in Excel) – choose a pre-configured table**". Set the Download Type as **Tab-delimited (.txt) file**, then hit **Get**. Open the text file, copy and paste all the data into an Excel sheet and go to Data/Text-to-column. Using the tool, select your delimiter as "tab" and hit finish. Now you will have a first column of gene IDs for pasting into YeastMine.

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

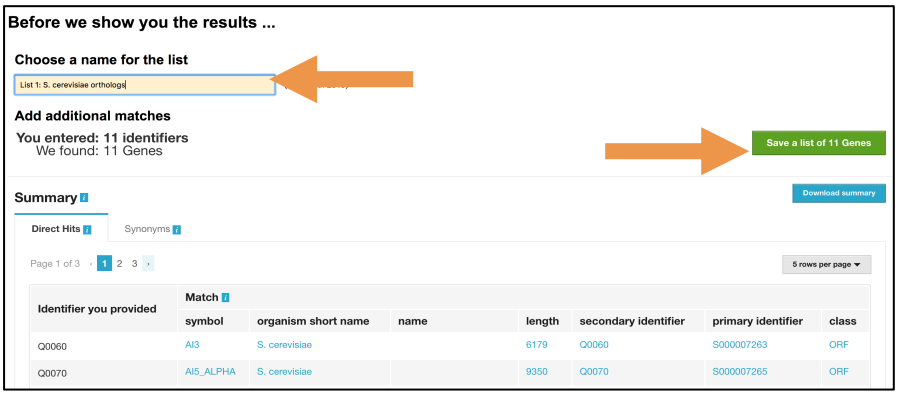
- Open the YeastMine homepage. You can access YeastMine from SGD by opening the Analyze tab and selecting **Gene Lists**, clicking the YeastMine link in the upper right corner of the homepage, or by entering in the URL:https://yeastmine.yeastgenome.org



- Open the file of *S. cerevisiae* orthologs that you downloaded earlier. To import these orthologs into YeastMine, copy and paste all entries in the **Gene ID** column of the text file into the "**Analyse**" box. Then, click on the purple "**ANALYSE**" button.

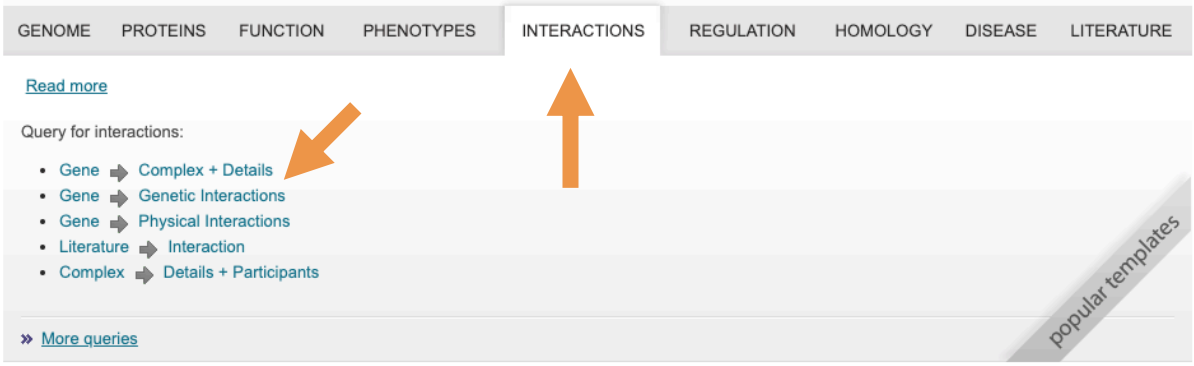


- A disambiguation page will be shown confirming your matches. 12 results should be shown. Name your gene list something descriptive, such as: "**List 1: S. cerevisiae orthologs**". Click on the green "**Save a list of Genes**" button.

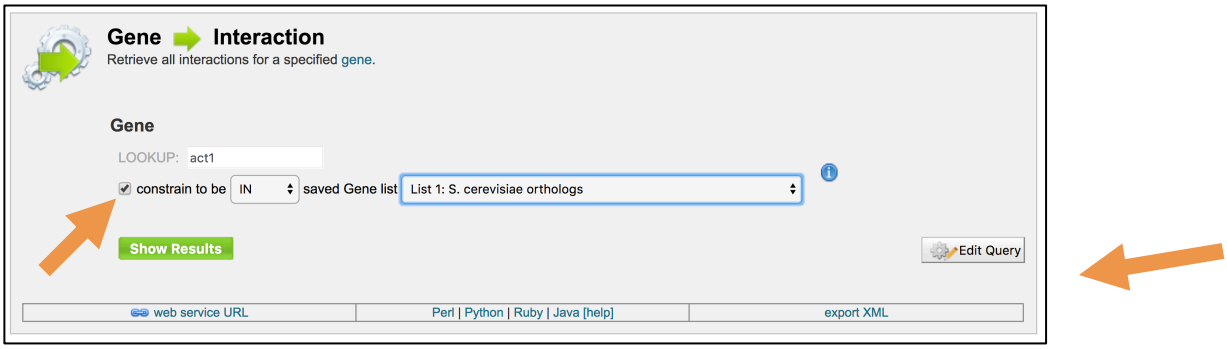


4. In YeastMine, find all synthetic lethal interactions for the *S. cerevisiae* orthologs by using the Gene → Interaction query:

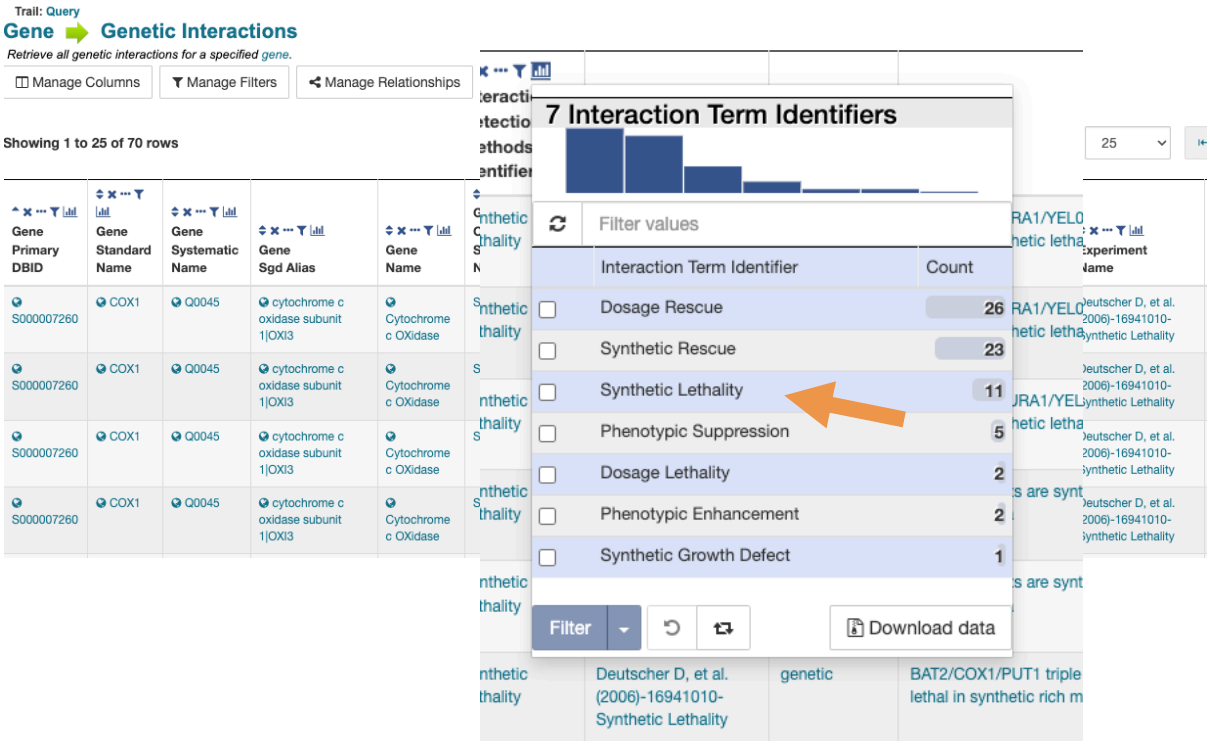
- Return to the YeastMine homepage: https://yeastmine.yeastgenome.org
- In the "popular templates" toolbar in the middle of the page, open the **INTERACTIONS** tab and select the query **Gene → Genetic Interactions**.



- Check the "**constrain to be IN**" checkbox. This allows you to input a list of genes. From the dropdown menu, select the list of *S. cerevisiae* orthologs you saved earlier in part 3. Click on the green **Show Results** button.

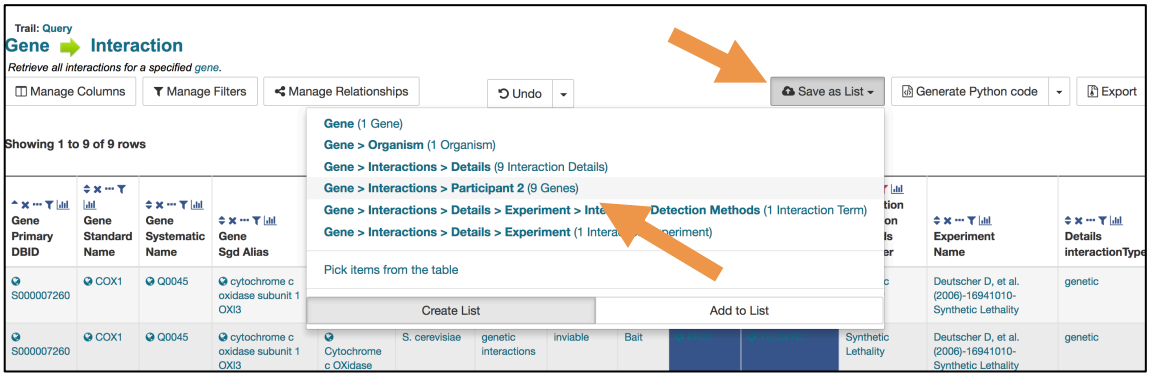


- The results table contains all genetic interactions for the list of *S. cerevisiae* orthologs you inputted. To filter for only **synthetic lethal** interactions, find the **Interaction Detection Methods Identifier** column. At the top of this column is a set of small blue icons. Click on the rightmost **View Column Summary** icon, which looks like a bar graph.

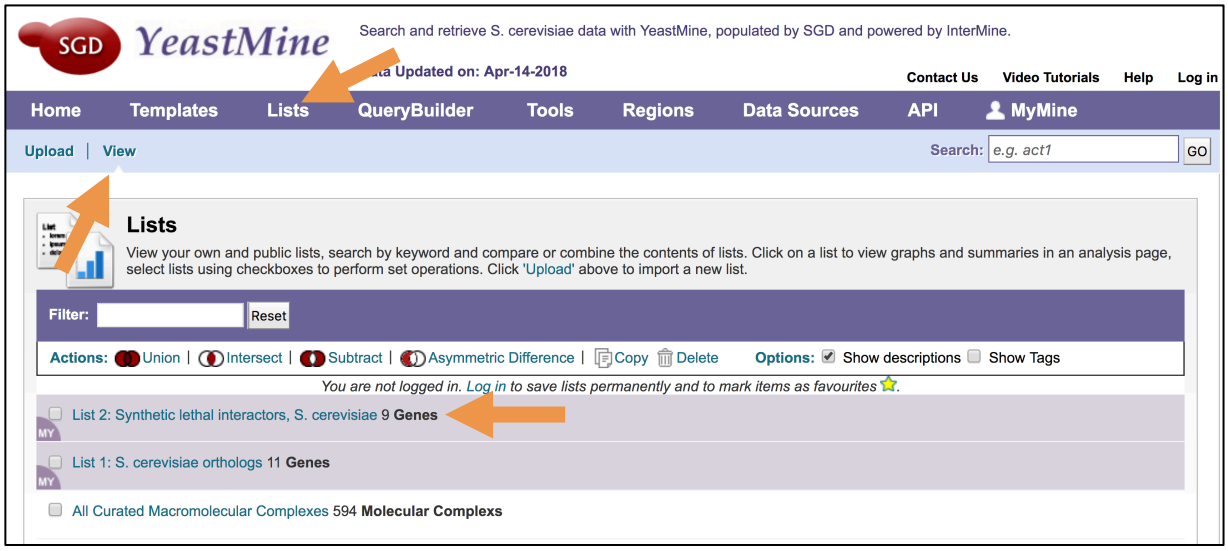


- A window summarizing all entries for this column will open. Check the entry for **Synthetic Lethality** and hit Filter.

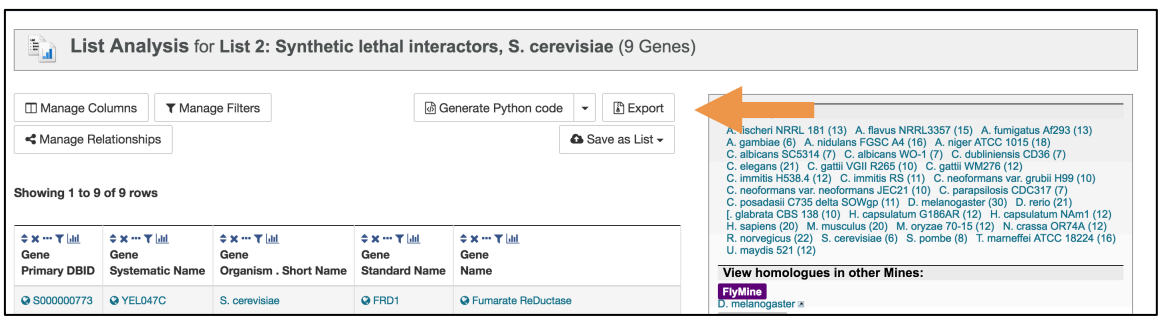
- The table now contains only synthetic lethal interactions. To save the interactors into a gene list, click on the **Save as List** button and select the entry **Gene > Interactions > Participant 2**. Give your list a descriptive name such as "**List 2: Synthetic lethal interactors, S. cerevisiae**".



- Access your new gene list by clicking on the **Lists** link in the top purple toolbar. Make sure that the **View** tab is open (see arrows).

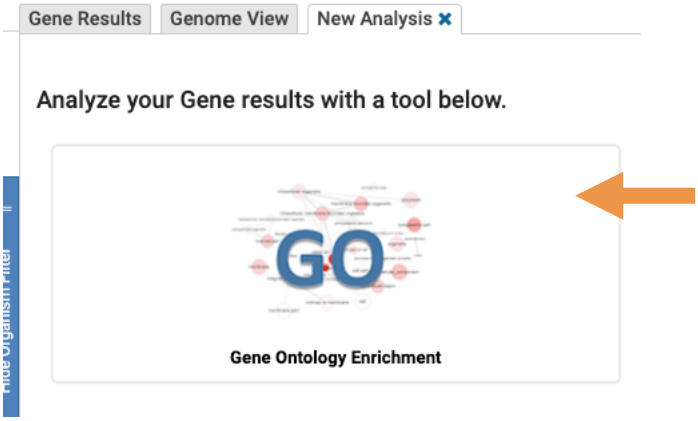


- Export the list of synthetic lethal interactors by clicking on the **Export** button, and then on the **Download file** button.



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

- Open the FungiDB homepage (http://fungidb.org/). Similar to part 2 of this exercise, in the **Search for Genes** box, open the **Annotation, curation and identifiers** section and click on Gene ID(s).
- Copy and paste all of the systematic *S. cerevisiae* gene names (YEL047C, YKL141W, etc.) from the downloaded list obtained in part 4 of this exercise. Hit **Get Answer.**
- To the right of the Gene Results table, click on the **Analyze Results** button. Select **Gene Ontology Enrichment** and run an enrichment for Biological Process.

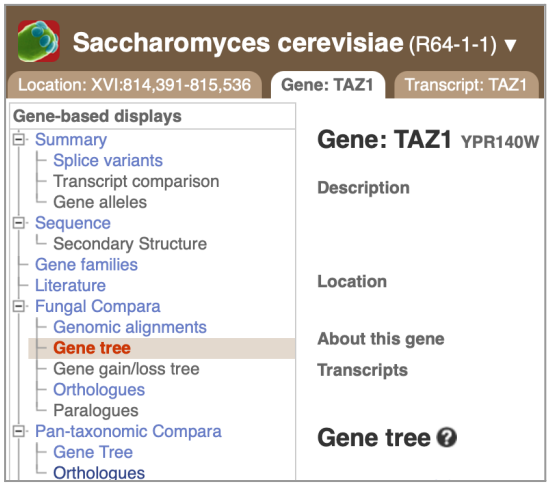


- Are the results surprising? Remember that these *S. cerevisiae* genes have synthetic lethal interactions with mitochondrial genes. Do the results suggest any biological processes that, if disrupted, might possibly inhibit mitochondria-defective *C. glabrata* clinical isolates?

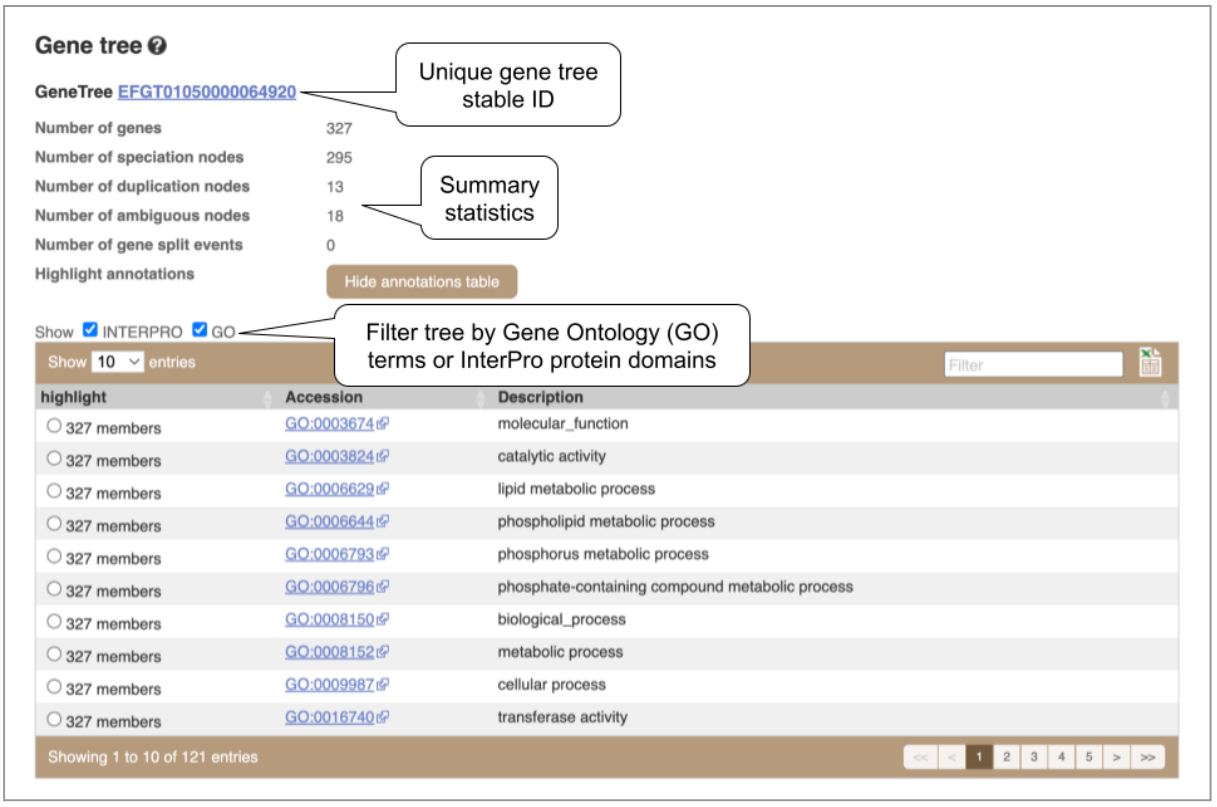
- Use the "Transform by Orthology" function to convert the *S. cerevisiae* genes into *C. glabrata* orthologs. These *C. glabrata* genes are predicted to have synthetic lethal interactions with *C. glabrata* mitochondrial genes.

Exercise: Ensembl Fungi gene trees and homologues

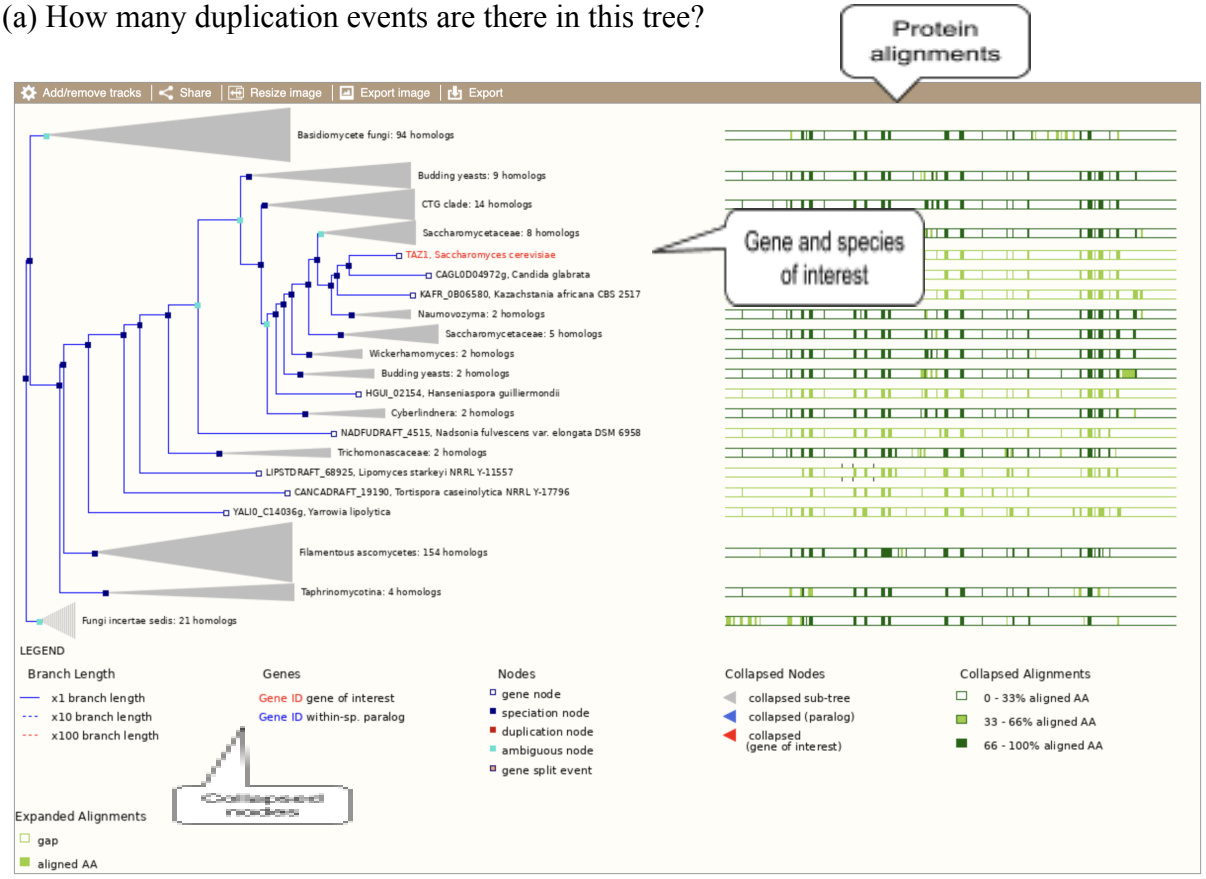
Let's look at the homologues of *Saccharomyces cerevisiae TAZ1* (Gene stable ID: YPR140W). This gene is involved in stress response and conserved across different taxonomic domains. Search for the gene and go to the Gene tab.



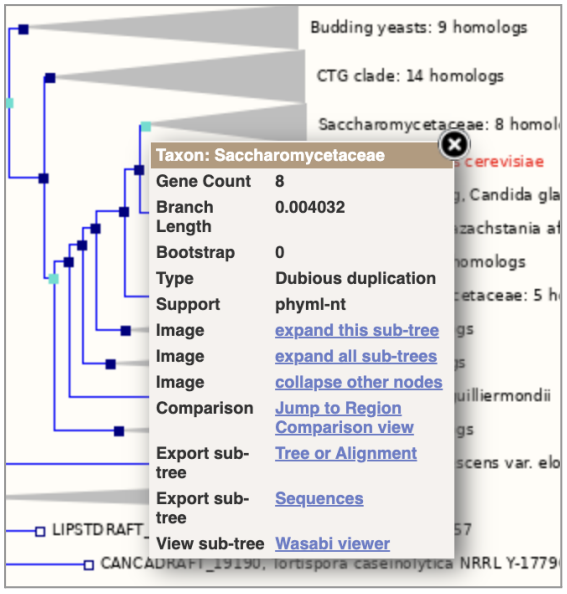
Click on Fungal Compara: Gene tree, which will display the current gene in the context of a phylogenetic tree used to determine orthologues and paralogues.



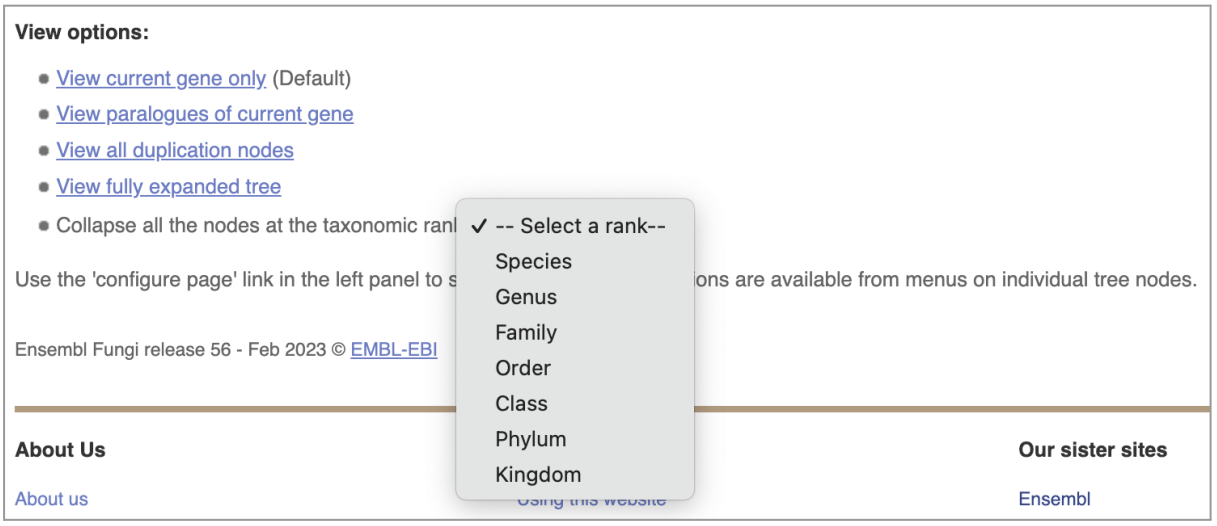
(a) How many duplication events are there in this tree?



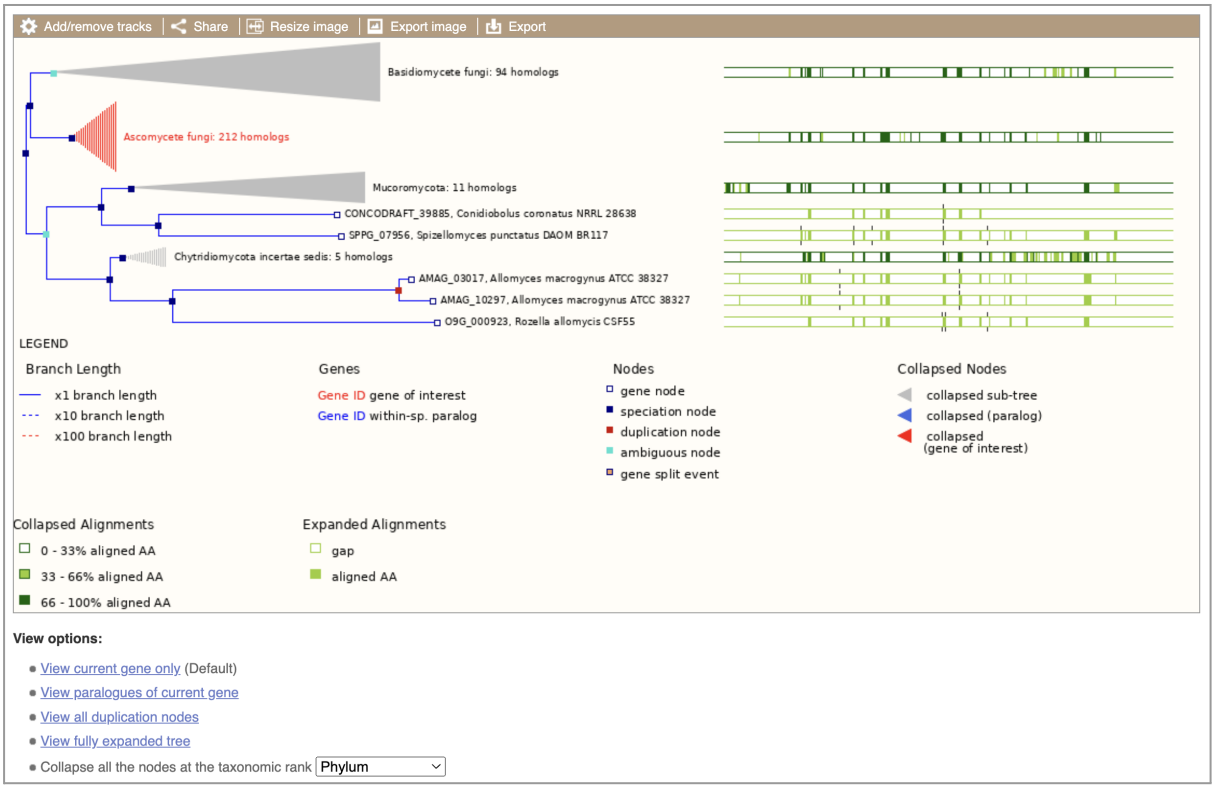
Funnels indicate collapsed nodes. Click on a node (coloured square) to get a pop-up.
We can then see what type of node this is, some statistics and options to expand or export the sub-tree:



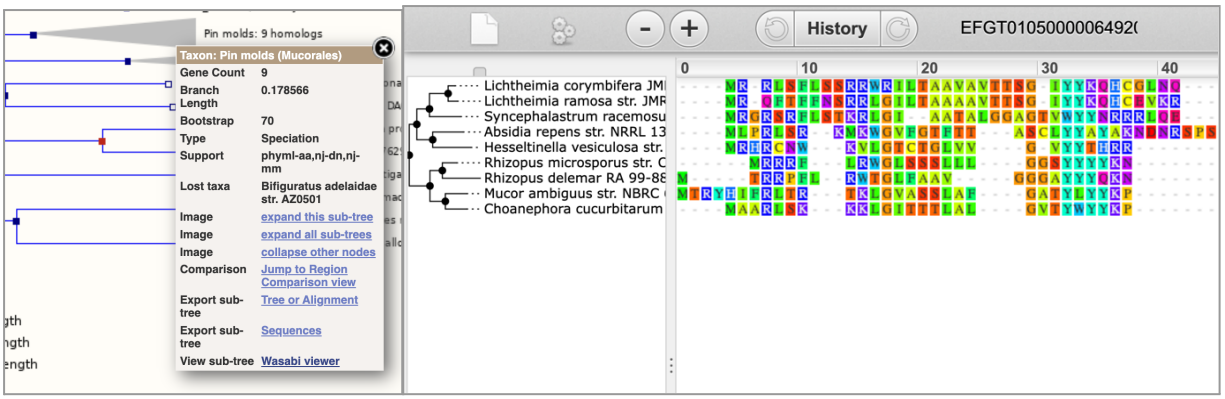
There are some quick filtering options below the image, where you can add paralogues, and quickly expand or collapse nodes:



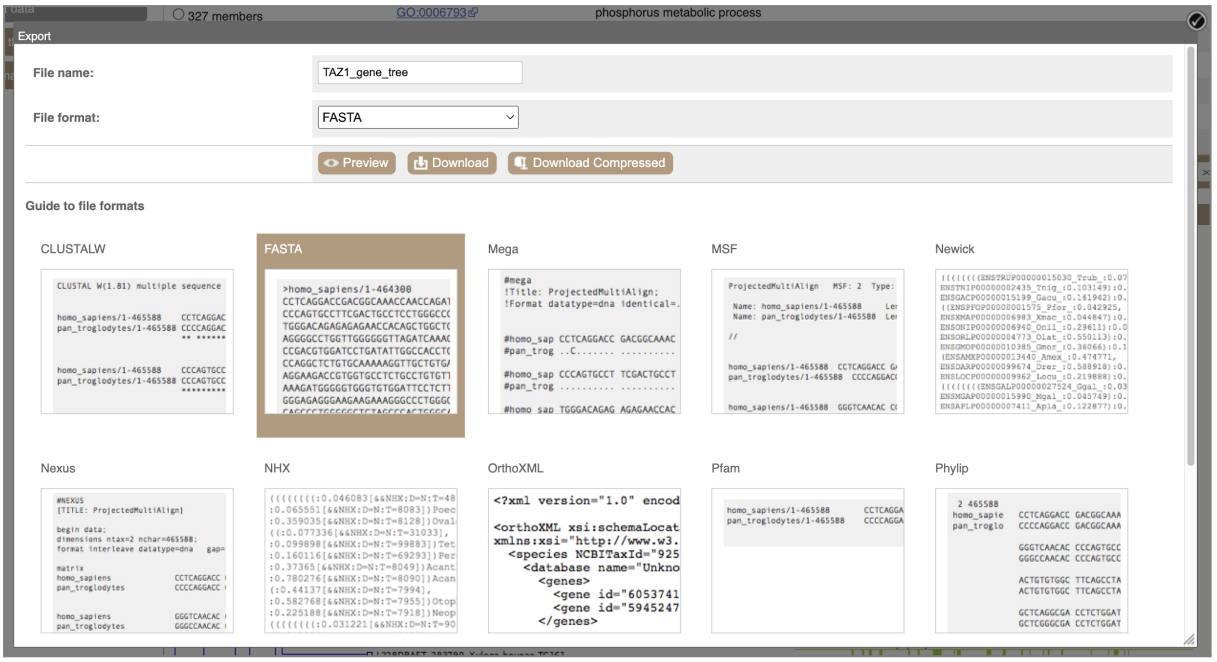
(b) What is the Phylum with the highest number of *TAZ1* homologues?



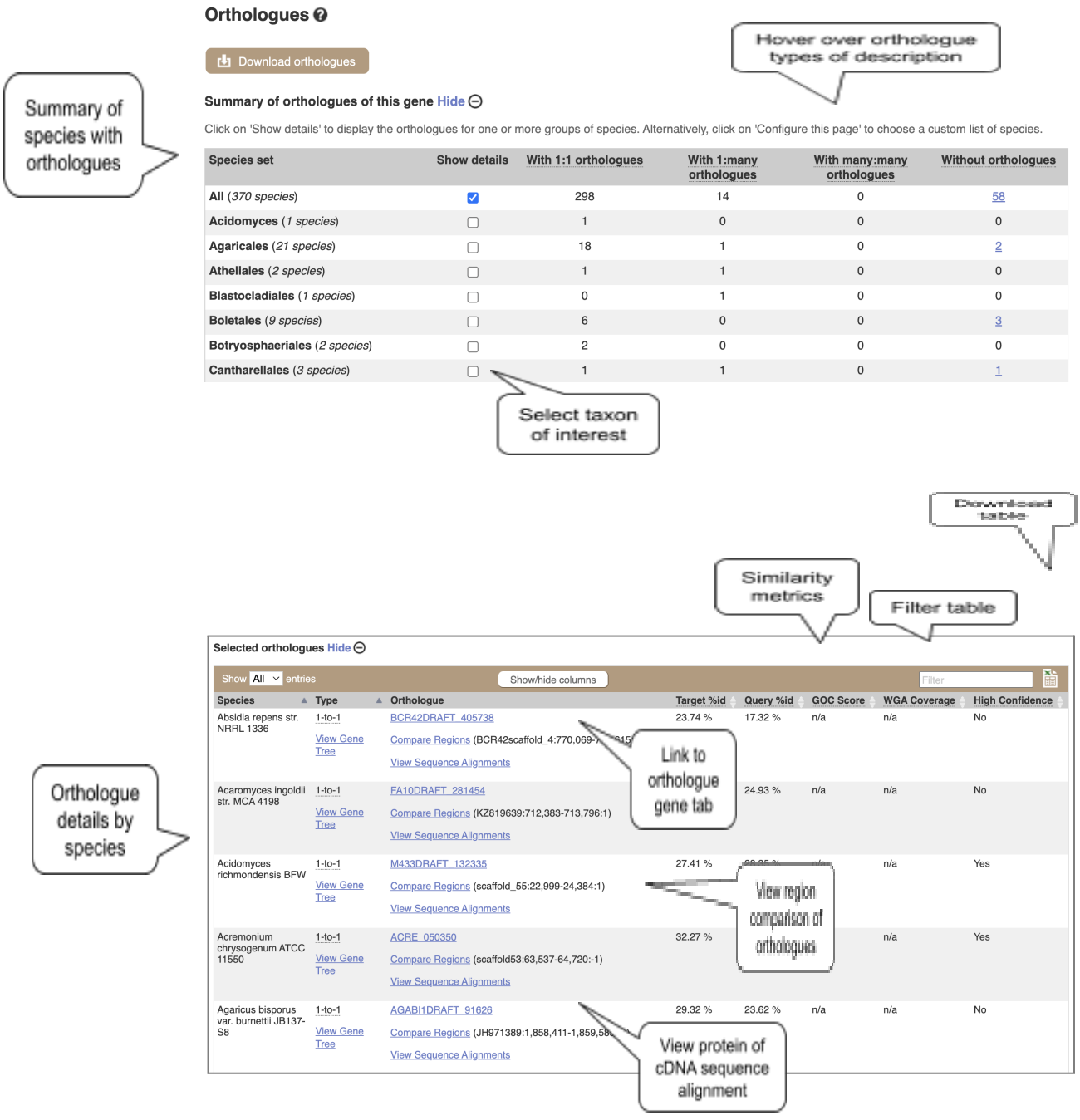
(c) What is the bootstrap support of the pin moulds (*Mucorales*) Class? Can you display the sequence alignment of all the homologues in this Class (*Hint: Use the Wasabi viewer*)?



You can download the tree in a variety of formats. Click on the Export icon  in the bar at the top of the image to get a pop-up where you can choose your format. You can preview this file before you download.

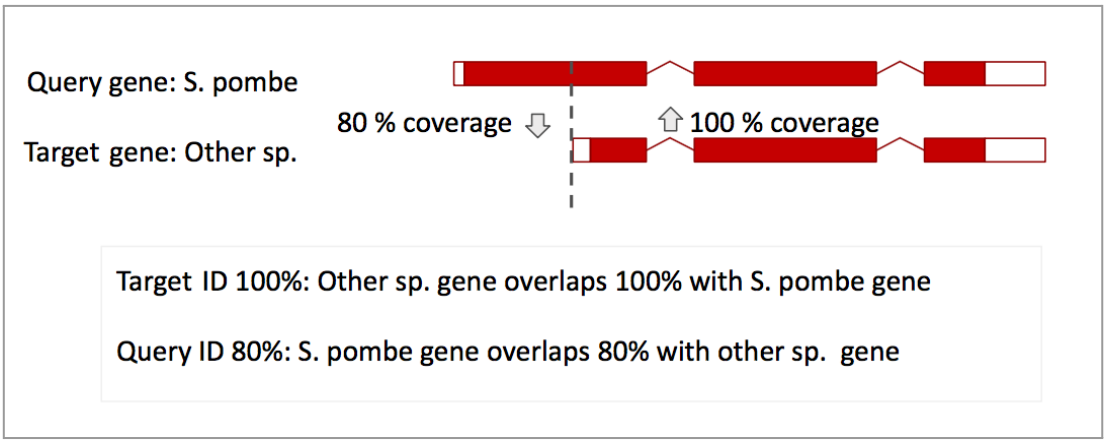


We can look at homologues in the Orthologues and Paralogues pages, which can be accessed from the left-hand menu. If there are no orthologues or paralogues, then the name will be greyed out. Click on Orthologues to see the orthologues available.

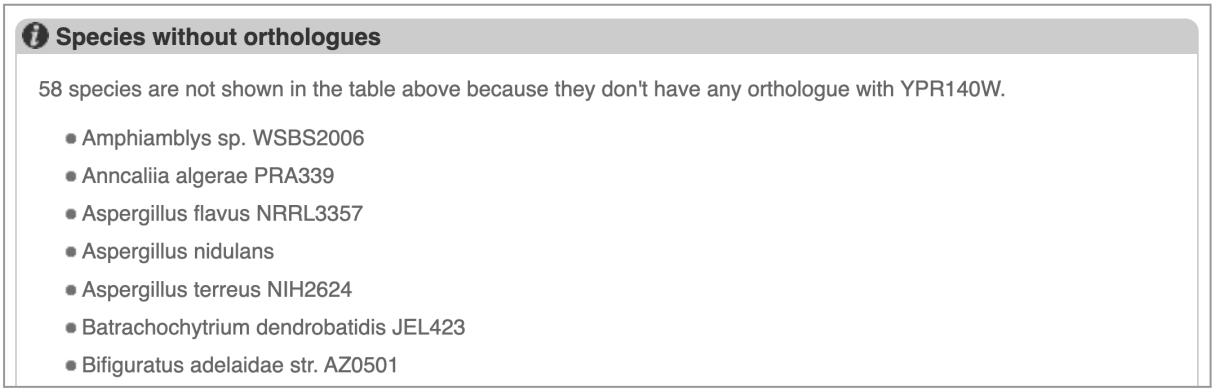

Orthologues
Download orthologues
Hover over orthologue types of description
Summary of species with orthologues
Summary of orthologues of this gene Hide
Click on 'Show details' to display the orthologues for one or more groups of species. Alternatively, click on 'Configure this page' to choose a custom list of species.
Species set | Show details | With 1:1 orthologues | With 1:many orthologues | With many:many orthologues | Without orthologues
All (370 species) | 298 | 14 | 0 | 58
Acidomyces (1 species) | 1 | 0 | 0 | 0
Agaricales (21 species) | 18 | 1 | 0 | 2
Atheliales (2 species) | 1 | 1 | 0 | 0
Blastocladiales (1 species) | 0 | 1 | 0 | 0
Boletales (9 species) | 6 | 0 | 0 | 3
Botryosphaeriales (2 species) | 2 | 0 | 0 | 0
Cantharellales (3 species) | 1 | 1 | 0 | 1
Select taxon of interest
Download table
Similarity metrics
Filter table
Selected orthologues Hide
Show All entries
Show/hide columns
Filter
Species | Type | Orthologue | Target %id | Query %id | GOC Score | WGA Coverage | High Confidence
Absidia repens str. NRRL 1336 | 1-to-1 | BCR42DRAFT_405738 | 23.74 % | 17.32 % | n/a | n/a | No
View Gene Tree | Compare Regions (BCR42scaffold_4:770,069-...615) | View Sequence Alignments
Link to orthologue gene tab
Orthologue details by species
Acaromyces ingoldii str. MCA 4198 | 1-to-1 | FA10DRAFT_281454 | 24.93 % | n/a | n/a | No
View Gene Tree | Compare Regions (KZ819639:712,383-713,796:1) | View Sequence Alignments
Acidomyces richmondensis BFW | 1-to-1 | M433DRAFT_132335 | 27.41 % | n/a | n/a | Yes
View Gene Tree | Compare Regions (scaffold_55:22,999-24,384:1) | View Sequence Alignments
View region comparison of orthologues
Acremonium chrysogenum ATCC 11550 | 1-to-1 | ACRE_050350 | 32.27 % | n/a | n/a | Yes
View Gene Tree | Compare Regions (scaffold53:63,537-64,720:-1) | View Sequence Alignments
Agaricus bisporus var. burnettii JB137-S8 | 1-to-1 | AGABI1DRAFT_91626 | 29.32 % | 23.62 % | n/a | n/a | No
View Gene Tree | Compare Regions (JH971389:1,858,411-1,859,58...) | View Sequence Alignments
View protein of cDNA sequence alignment

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

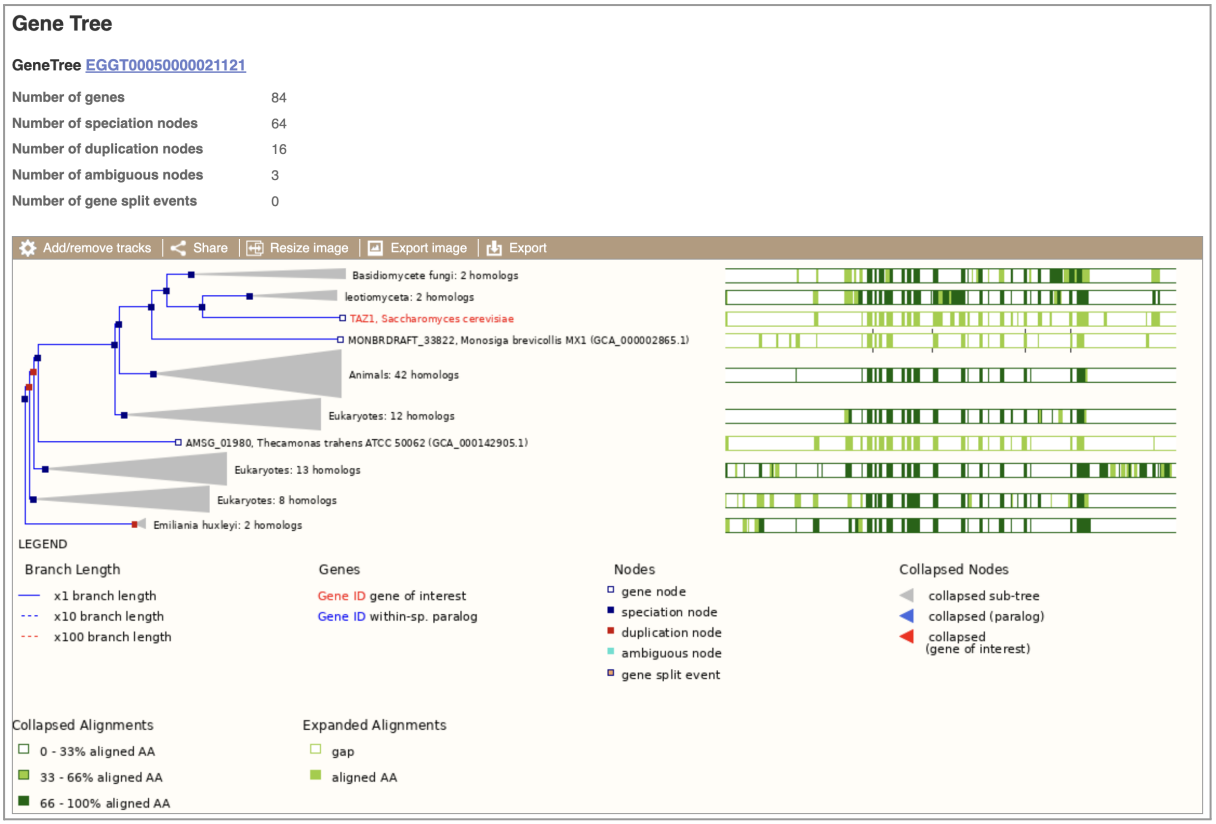
The sequence identity is reported in two ways, Target %id is how much of the orthologue or 'target gene' overlaps with the query gene, or our *S. cerevisiae* gene. The Query %id is the inverse of this. For example:



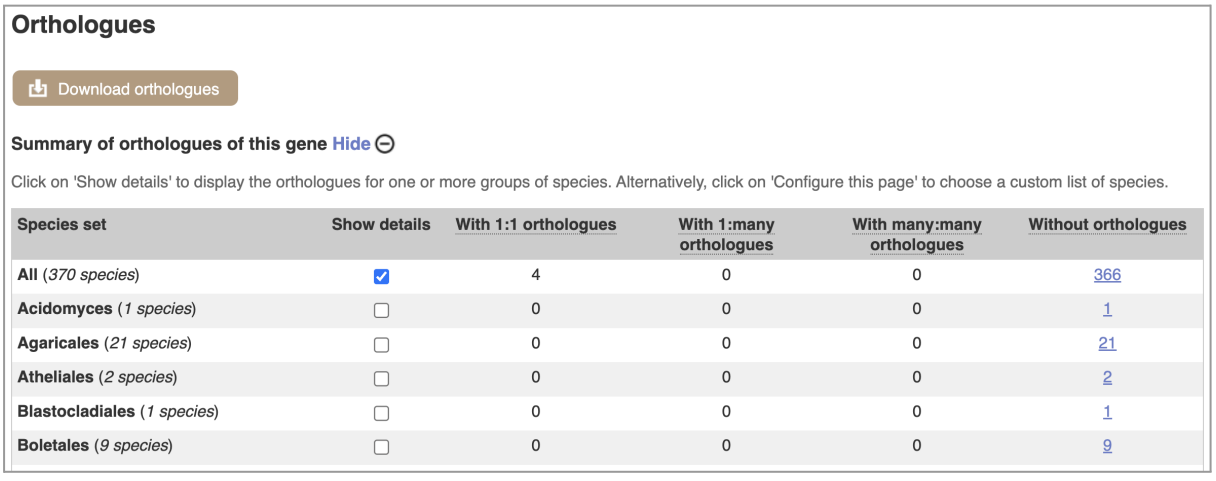
Scroll to the bottom of the page to see a list of the species that do not have any orthologues with *TAZ1* in *Saccharomyces cerevisiae*… there's a lot!

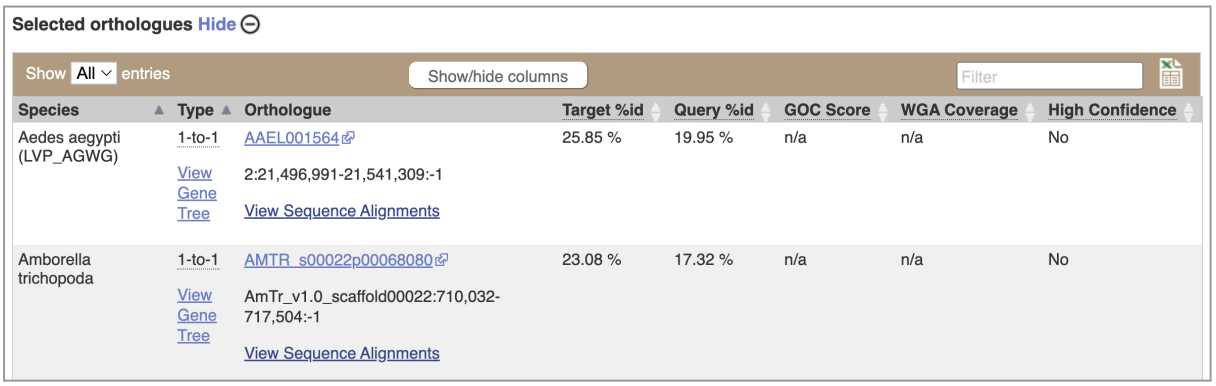


Saccharomyces cerevisiae is part of Pan-compara, which compares a subset of fungal species with species from other taxa, such as plants, bacteria and vertebrates. Go to Pan-taxonomic Compara > Gene Tree. Let's have look at the Pan-taxonomic tree with nodes collapsed at the Kingdom rank.



Click on Pan-taxonomic Compara > Orthologues now.





(d) How many species with predicted orthologues for this gene are there in Fungal Compara? What about in Pan-compara?

Fungal Compara:

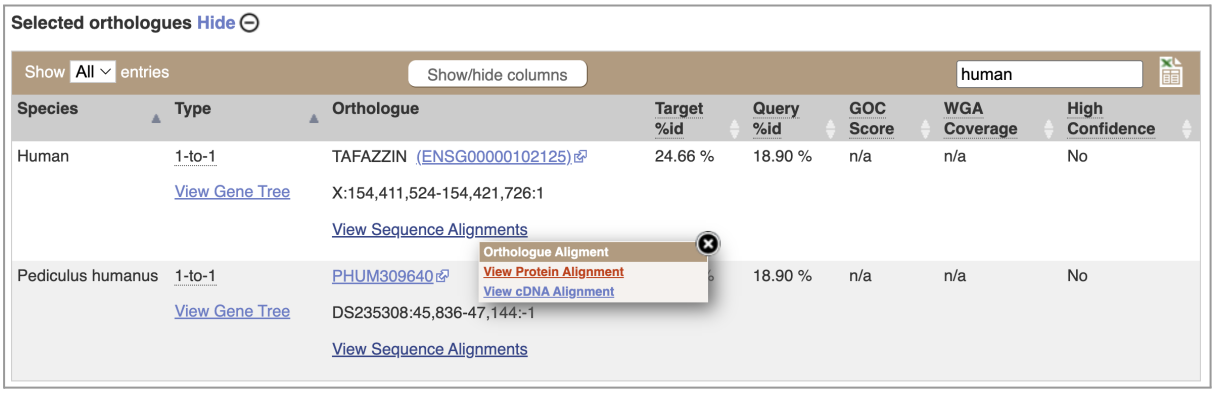
Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (*370 species*)	☑	298	14	0	58

Pan 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

(e) 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*).

(f) Filter the second table to view the human orthologue. How much sequence identity does the human protein have to the *Saccharomyces cerevisiae* one? Is it a high confidence homology? Click on the View Sequence Alignment link in the Orthologue column to View Protein Alignment in Clustal W format. Does it support your conclusions?



Orthologue Alignment

Download homology

Type: 1-to-1 orthologues

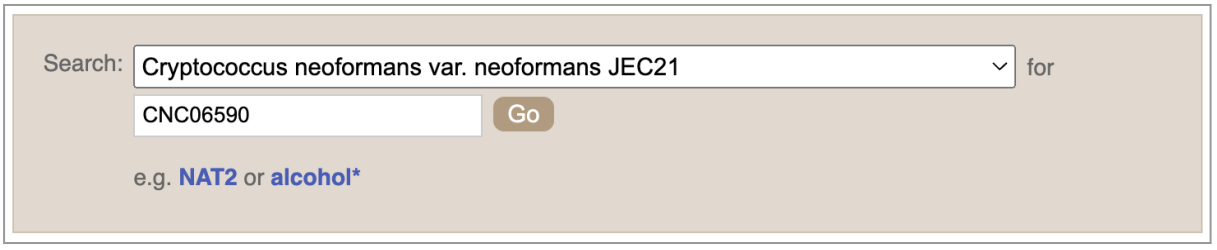
Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Saccharomyces cerevisiae	YPR140W	YPR140W	381 aa	18 %	65 %	XVI:814391-815536
Human	ENSG00000102125	ENSP00000469981	292 aa	24 %	85 %	X:154411524-154421726

```
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

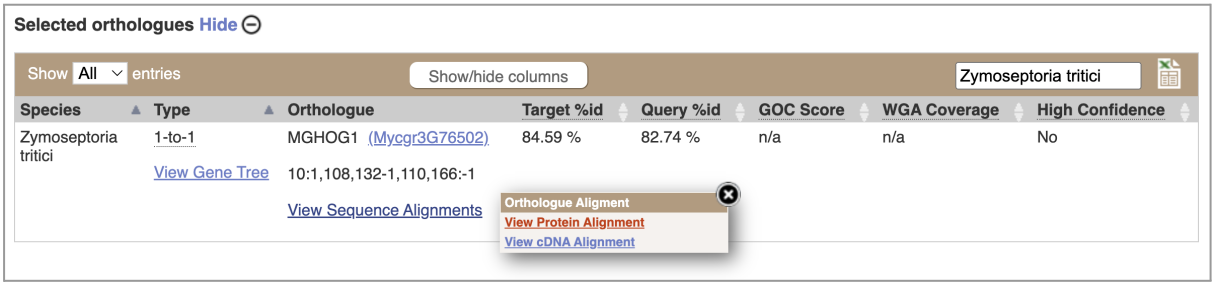
Exploring an orthologue that we identified using BioMart (additional exercise 1). We identified 18 genes associated with the GO term detoxification in *Zymoseptoria tritici*. We then found a single high confidence orthologue in *Cryptococcus neoformans* which we will now explore further.

Search for CNC06590 in *Cryptococcus neoformans* var. neoformans JEC21 to go to the gene page. Click on the gene ID CNM01690 to 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?

Orthologue alignment

Download homology

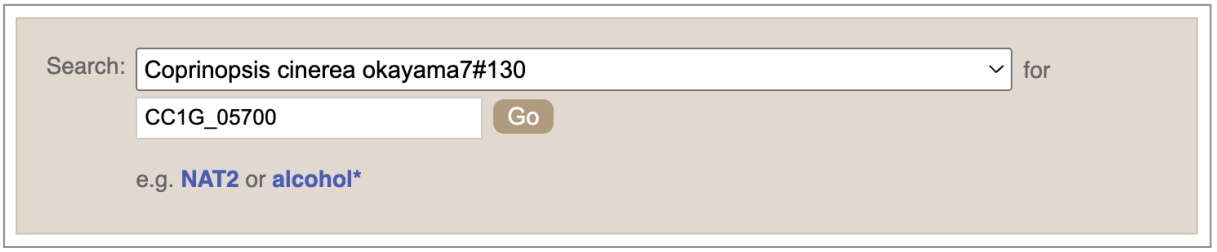
Type: 1-to-1 orthologues

Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Cryptococcus neoformans var. neoformans JEC21	CNC06590	AAW42642	365 aa	82 %	97 %	3:1927422-1929917
Zymoseptoria tritici	Mycgr3G76502	Mycgr3P76502	357 aa	84 %	99 %	10:1108132-1110166

```
CLUSTAL W (1.81) multiple sequence alignment


AAW42642/1-365          MADFVKLSIFGTVFEVTTRYVDLQPVGMGAFGLVCSAKDQLSGTSVAIKKIMKPFSTPVL
Mycgr3P76502/1-357      MAEFVRAQIFGTTFEITSRYTDLQPVGMGAFGLVCSAKDQLTGQAVAVKKIMKPFSTPVL
                        **:**: .****.**:*:**.********************:* :**:************


AAW42642/1-365          SKRTYRELKLLKHLRHENIISLSDIFISPLEDIYFVTELLGTDLHRLLTSRPLEKQFIQY
Mycgr3P76502/1-357      SKRTYRELKLLKHLKHENVISLSDIFISPLEDIYFVTELLGTDLHRLLTSRPLEKQFIQY
                        **************:***:****************************************


AAW42642/1-365          FLYQILRGLKYVHSAGVVHRDLKPSNILVNENCDLKICDFGLARIQDPQMTGYVSTRYYR
Mycgr3P76502/1-357      FLYQILRGLKYVHSAGVVHRDLKPSNILVNENCDLKICDFGLARIQDPQMTGYVSTRYYR
                        ************************************************************


AAW42642/1-365          APEIMLTWQKYDVAVDIWSTGCIFAEMLEGKPLFPGKDHVNQFSIITELLGTPPDDVIQT
Mycgr3P76502/1-357      APEIMLTWQKYDVEVDIWSAGCIFAEMLEGKPLFPGKDHVNQFSIITDLLGTPPDDVIST
                        ************ ***** :*************************:**********.*


AAW42642/1-365          IASENTLRFVQSLPKREKVPFSTKFPNADPVSLDLLEKMLVFDPRTRISAAEGLAHEYLA
Mycgr3P76502/1-357      ICSENTLRFVQSLPKRERQPLKNKFKNADPQAIELLERMLVFDPRKRVKAGEALADPYLS
                        *.***************: *:..** **** :::***:*******.*:.*.*.**. **:


AAW42642/1-365          PYHDPTDEPVAAEVFDWSFNDADLPVDTWKVMMYSEILDFHNLGDISQNE--AEGPVTGE
Mycgr3P76502/1-357      PYHDPTDEPEAEEKFDWSFNDADLPVDTWKIMMYSEILDYHNVDS-ANNGEGQE---NGG
                        ******** * * ****************:********:**:.. ::*    *   .*


AAW42642/1-365          VPAAPAS
Mycgr3P76502/1-357      A------
                        .
```

Additional Exercise 2 - Mushroom genes

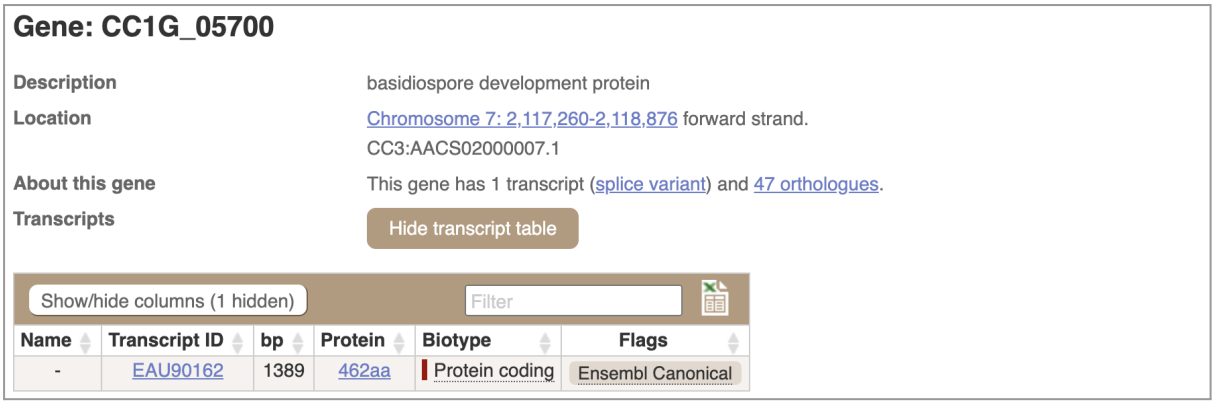
We're going to take a look at the gene CC1G_05700 in *Coprinopsis cinerea* okayama7#130.



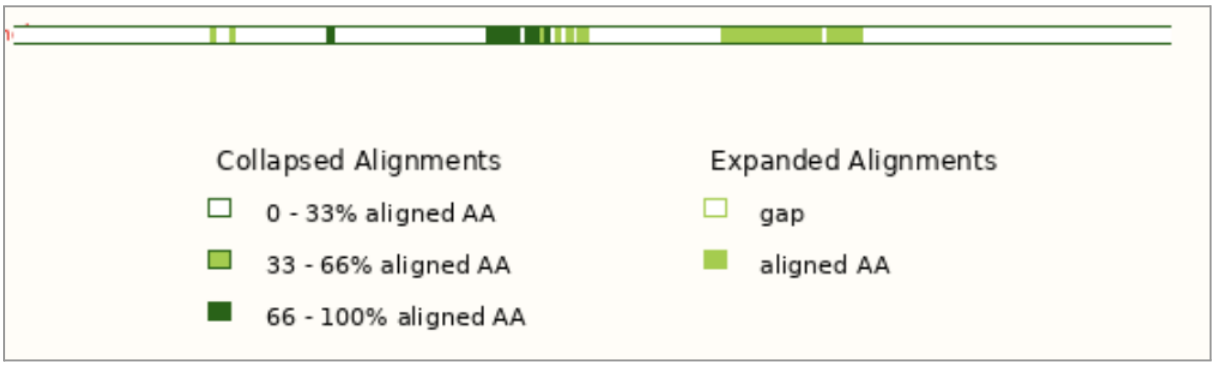
From the gene tab, click to view the Gene tree. At the bottom of the image click to collapse all the nodes at the taxonomic rank of Class.

(a) What do you notice about the types of fungi shown in the gene tree?

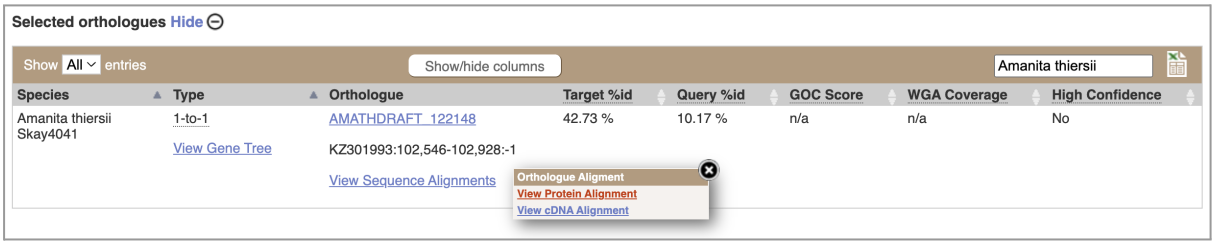
(b) Does this match with what you would expect from the gene description? (*Hint: Agaricomycetes class belongs to the Basidiomycota phylum*)



(c) Based on the protein alignment shown at the right, can you predict which end of the gene/protein is most conserved?



(e) 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?



Orthologue alignment

Download homology

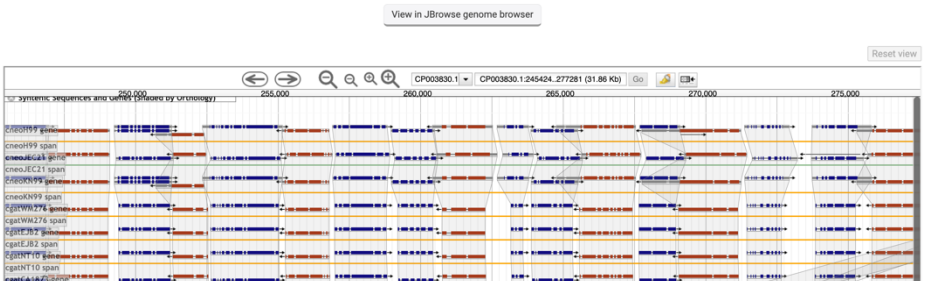
Type: 1-to-1 orthologues

Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Coprinopsis cinerea okayama7#130	CC1G_05700	EAU90162	462 aa	10 %	21 %	7:2117260-2118876
Amanita thiersii Skay4041	AMATHDRAFT_122148	PFH51030	110 aa	42 %	90 %	KZ301993:102546-102928

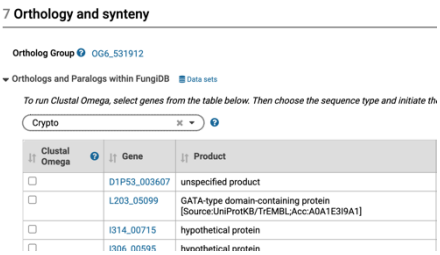
```
CLUSTAL W (1.81) multiple sequence alignment


EAU90162/1-462          ----------MRVLLHDTQMNLEKFSGHVEALISNVKETSQELRKTSSTFEQEHDKLLG
PFH51030/1-110          PLTPLDKNATSMRVLLHDTQANFEKFSTRVDNFLNGLAETKSEINLVKSLFERGQETLTN
                                 ********* *:**** :*: ::..: **..*:. ..* **: ::.* .


EAU90162/1-462          DIIDLVNRSQKQLQSSIGSPAQSAALDAFMNKVELRLESLDQRLDAMQAFNQTHSQALQT
PFH51030/1-110          DIIDLVNRCQSQIQKTLGSPAQASGMEQLSKDINQRLDCLDKRLDAIQTV----------
                        ********.*.*:*.::*****::.:: : :.:: **:.**:****:*:.


EAU90162/1-462          QIQAIQNLQAQQNLILNAVTPLLPLLQSLPPRLAPSTSLANPSQTQRTDASSQTIIEKRQ
PFH51030/1-110          ------------------------------------------------------------


EAU90162/1-462          PSYHQETLRKRQRVDSDIQEISPPKPLPGSAQKKRRIESPRSVQKPSLELTQRLFPSSSP
PFH51030/1-110          ------------------------------------------------------------


EAU90162/1-462          DLIKYSTDSEGPSTKPQVNERSAPLVTPRRPLQDLFPFPPGSNQRSVSKRPMPPSSTRLV
PFH51030/1-110          ------------------------------------------------------------


EAU90162/1-462          GPGKSATPGPSRVGAEESRAALARRPLIKPLAIAPLAFSSTSKTPVHISNFTPKPVTPSL
PFH51030/1-110          ------------------------------------------------------------


EAU90162/1-462          RNAVAGEGRALKIAQTPQVLKNERMTSQAAKNTTMPPPAGMVSLSRSSTITTATATKPTS
PFH51030/1-110          ------------------------------------------------------------


EAU90162/1-462          NTPRFGPEANKPLLLRAPTNNGPRPLQERMKEPVREGRRFIPLVDTDDEDDSD
PFH51030/1-110          -----------------------------------------------------
```

FungiDB: Synteny in JBrowse

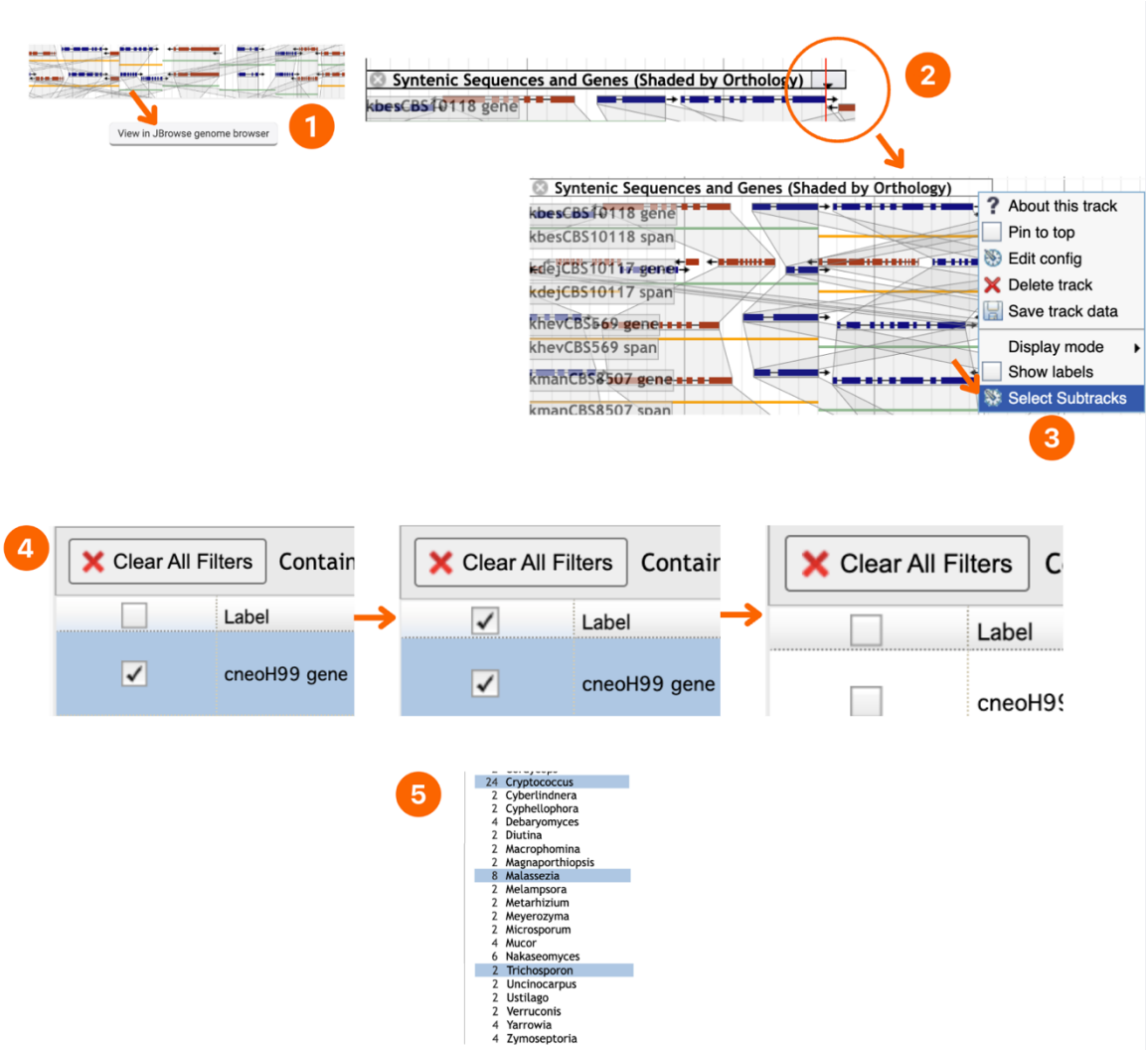
- **Navigate to the gene record page for Gat201 in *Cryptococcus neoformans* H99 and examine the evidence within the Orthology and Synteny section.**
    1. Use site search to locate the gene record page.
    2. Use the contents menu on the left to navigate to the Orthology and Synteny section.



    3. Filter for Cryptococcus species in the Orthology and Synteny table.
        - What can you tell about its conservation across *Cryptococcus* species?



- **Navigate to JBrowse and create a custom JBrowse view for this gene's synteny across *Cryptococcus*, *Malassezia*, and *Trichosporon*.**
    1. Click on the "View in JBrowse gene browser" button.
    2. When in JBrowse, left click at the end of the "Syntenic sequences and Genes (Shaded by Orthology)" tracks to bring up the pull-down menu.
    3. Click on the "Select subtracks" option.
    4. Use the main check box to clear all selections.
    5. Select tracks for *Cryptococcus*, *Malassezia*, and *Trichosporon (and don't forget save your choices by clicking on the "save" option at the bottom of the track).*



Gat201 is a positive regulator of titanization under specific conditions in *Cryptococcus*. Titan cell formation is a rare phenomenon in *C. neoformans*/*C. gattii* species complex. What can you conclude about the conservation of this gene across the selected fungal pathogens? Do the results make sense based on what you know about Gat201?

Examine neighboring genes in *Cryptococcus* species. Can you spot any genes that have undergone expansions, possible truncations or simply not present in all gene models?

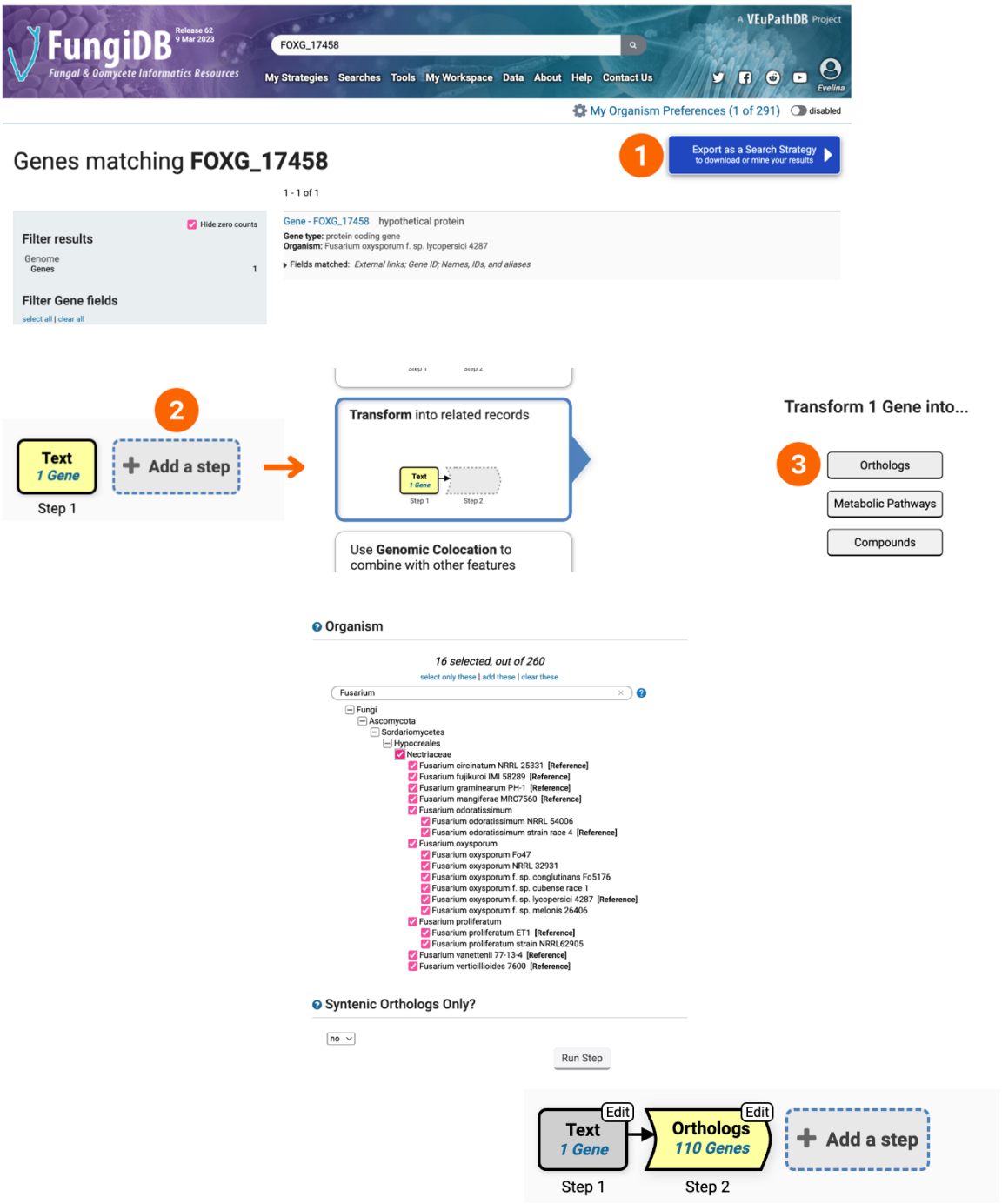
Mining synteny and orthology information for hypothetical genes.

- **Navigate to FOXG_17458 gene record page in FungiDB** and view the Orthology and Synteny section.

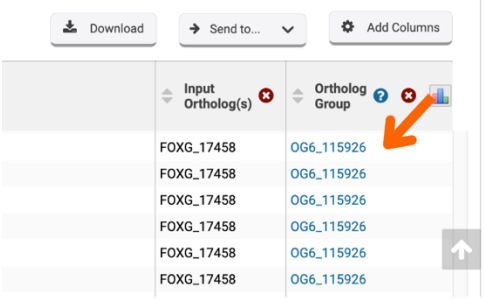
Notice that this hypothetical protein in *Fusarium oxysporum* f. sp. *lycopercisi* 4287 has no syntenic orthologs. Why do you think this is?

How can you test if there are orthologs of this gene in other *Fusarium* species?

- **Create a search strategy looking for orthologs of FOXG_17458 in *Fusarium* species**.
    1. Use site search and export FOXG_17458 as a strategy step.
    2. Click on the Add Step button.
    3. Deploy the “Transform into related record “ search and choose to transform genes into “Orthologs
    4. Select all Fusarium species and click on the “Run Step” button.



- **Examine phyletic distribution by clicking on the "Ortholog Group" link within the results table.**



Phlyetic Distribution of Proteins Download

Numbers refer to the number of proteins in that organism or taxonomic group.

Hide zero counts

fusarium

Eukaryota (EUKA)	**514**
Fungi (FUNG)	**514**
Ascomycota (ASCO)	**465**
Fusarium circinatum NRRL 25331 (fcir)	7
Fusarium fujikuroi IMI 58289 (ffuj)	7
Fusarium graminearum PH-1 (fgra)	2
Fusarium mangiferae MRC7560 (fman)	6
Fusarium odoratissimum NRRL 54006 (foxc)	9
Fusarium odoratissimum strain race 4 (foxt)	9
Fusarium oxysporum Fo47 (foxf)	3
Fusarium oxysporum NRRL 32931 (foxa)	4
Fusarium oxysporum f. sp. conglutinans Fo5176 (focf)	10
Fusarium oxysporum f. sp. cubense race 1 (foxr)	6
Fusarium oxysporum f. sp. lycopersici 4287 (foxy)	14
Fusarium oxysporum f. sp. melonis 26406 (foxm)	8
Fusarium proliferatum ET1 (fpro)	8
Fusarium proliferatum strain NRRL62905 (fprn)	6
Fusarium vanettenii 77-13-4 (fvan)	8
Fusarium verticillioides 7600 (fver)	4

Is there evidence of possible expansion across different Fusarium species?

- **Examine evidence for non-syntenic orthologs in MycoCosm.**

1. Navigate to Mycocosm main page and select a *Fusarium oxysporum* f. sp. *lycopersici* strain 4287 genome. [mycocosm.jgi.doe.gov/Fusox2]

Home • Fusarium oxysporum f. sp. lycopersici 4287 v2

SEARCH | BLAST | BROWSE | ANNOTATIONS | MCL CLUSTERS | SYNTENY | DOWNLOAD | INFO | HOME | HELP!

Fusarium wilt of tomato caused by Fusarium oxysporum f.sp. lycopersici.
Image Credit: David B. Langston, University of Georgia, Bugwood.org

The genome of *Fusarium oxysporum* f. sp. lycopersici strain 4287 (race 2, VCG 0030) was sequenced by the Broad Institute and the text below is copied from there. In order to allow comparative analyses with other fungi, a copy of this genome was imported into MycoCosm.

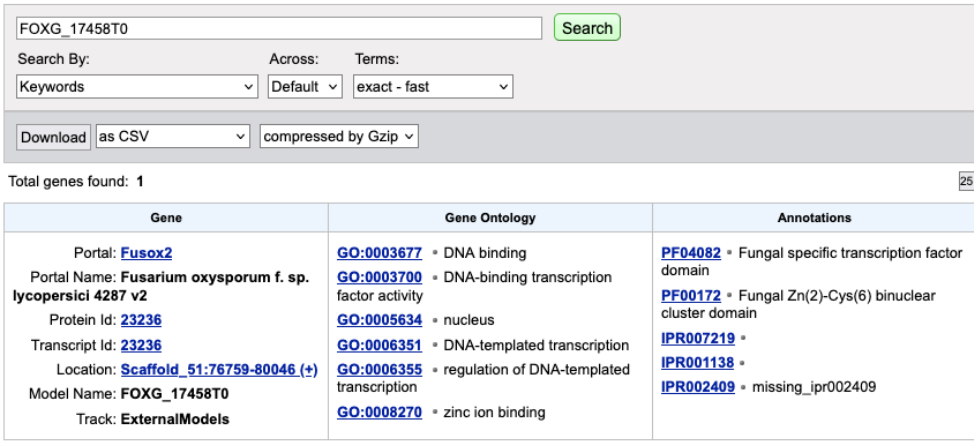
Fungi of the *Fusarium oxysporum* species complex (FOSC) are ubiquitous soil and plant inhabiting microbes. As plant pathogens, FOSC strains can cause wilt and root rot diseases on over 120 plant species (Michielse and Rep, 2009). Many FOSC strains can infect plant roots without apparent effect or can even protect plants from subsequent infection (Alabouvette et al., 2009). FOSC isolates also have been identified as human pathogens causing localized or disseminated infections that may become life-threatening in neutropenic individuals (O'Donnell et al., 2004).

The first genome made available in 2007 was from a tomato wilt strain FOL 4287 (NRRL 34936) which was used for comparative analysis with the genomes of *F. graminearum* and *F. verticillioides*. Results of this comparison led to the discovery of mobile supernumerary chromosomes in this strain of *F. oxysporum* f. sp. lycopersici (race 2 - VCG 0030) containing genes required for host specific infection and disease (Ma et al., 2010).

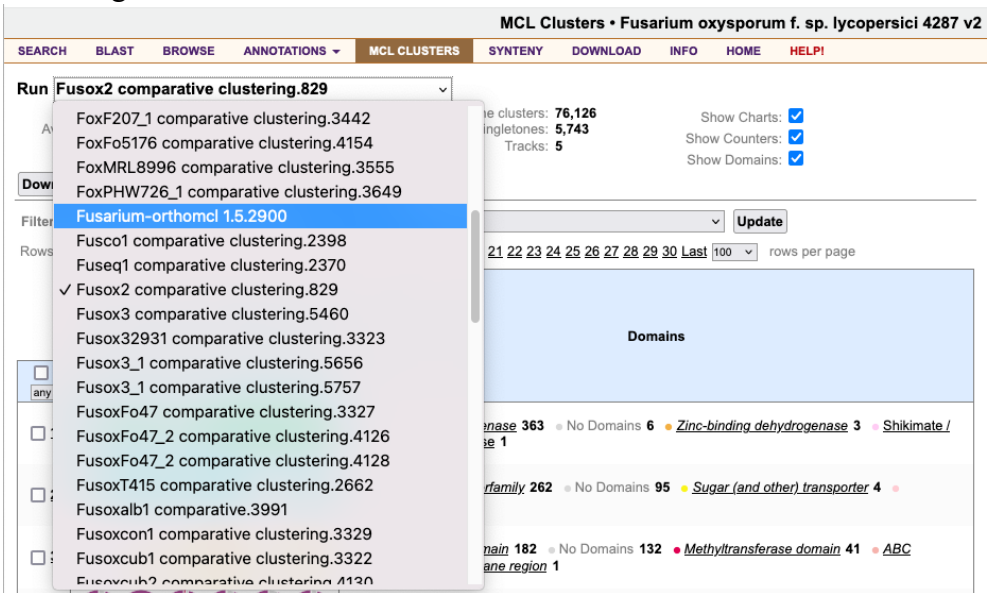
References :

- Alabouvette,C., Olivain,C., Migheli,Q., and Steinberg,C. (2009) Microbiological control of soil-borne phytopathogenic fungi with special emphasis on wilt-inducing Fusarium oxysporum. New Phytologist 184: 529-544.
- Ma,L.J., van der Does,H.C., Borkovich,K.A., Coleman,J.J., Daboussi,M.J., Di Pietro,A. et al. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 464: 367-373.
- O'Donnell,K., Sutton,D.A., Rinaldi,M.G., Magnon,K.C., Cox,P.A., Revankar,S.G. et al. (2004) Genetic diversity of human pathogenic members of the Fusarium oxysporum complex inferred from multilocus DNA sequence data and amplified fragment length polymorphism analyses: Evidence for the recent dispersion of a geographically widespread clonal lineage and nosocomial origin. Journal of Clinical Microbiology 42: 5109-5120.

2. Use the Fusox2 portal's search page to identify the proteinID of "FOXG_17458T0" (Transcript 0 of FOXG_17458). You will find that the proteinID of FOXG_17458 in Fusox2 is 23236.

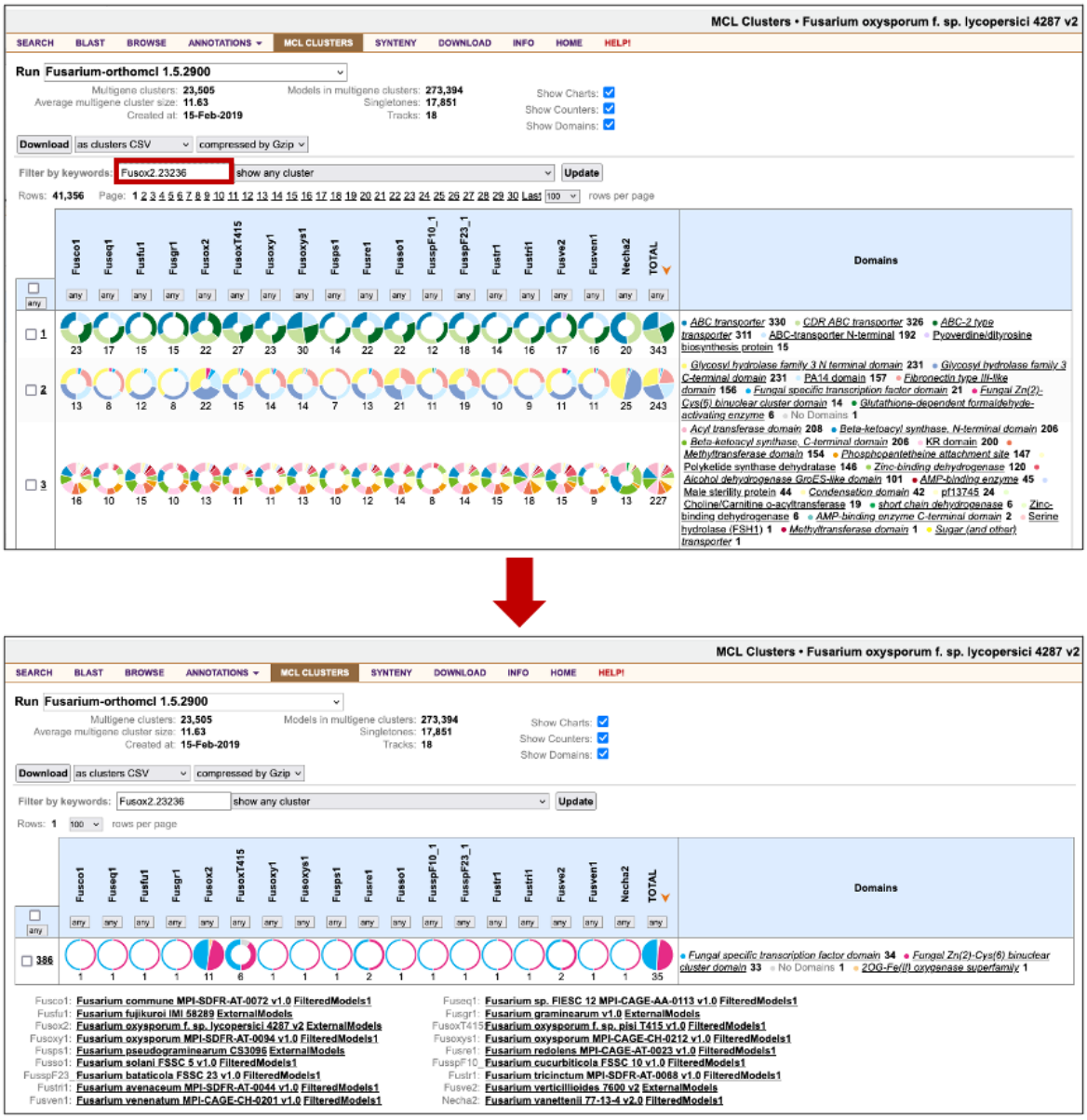


3. Click on MCL clusters tab and then use the pull down menu to select clustering run “Fusarium-orthomcl 1.5.2900”.



https://mycocosm.jgi.doe.gov/clm/run/Fusarium-orthomcl.2900?organism=Fusox2

4. Enter keyword Fusox2.23236 (databaseID.proteinID) and select “Update” to find clusters with that protein in it. Remember, for FOXG_17458 (FOXG_17458T0 protein ID in MycoCosm is 23236 and genome ID is Fusox2)

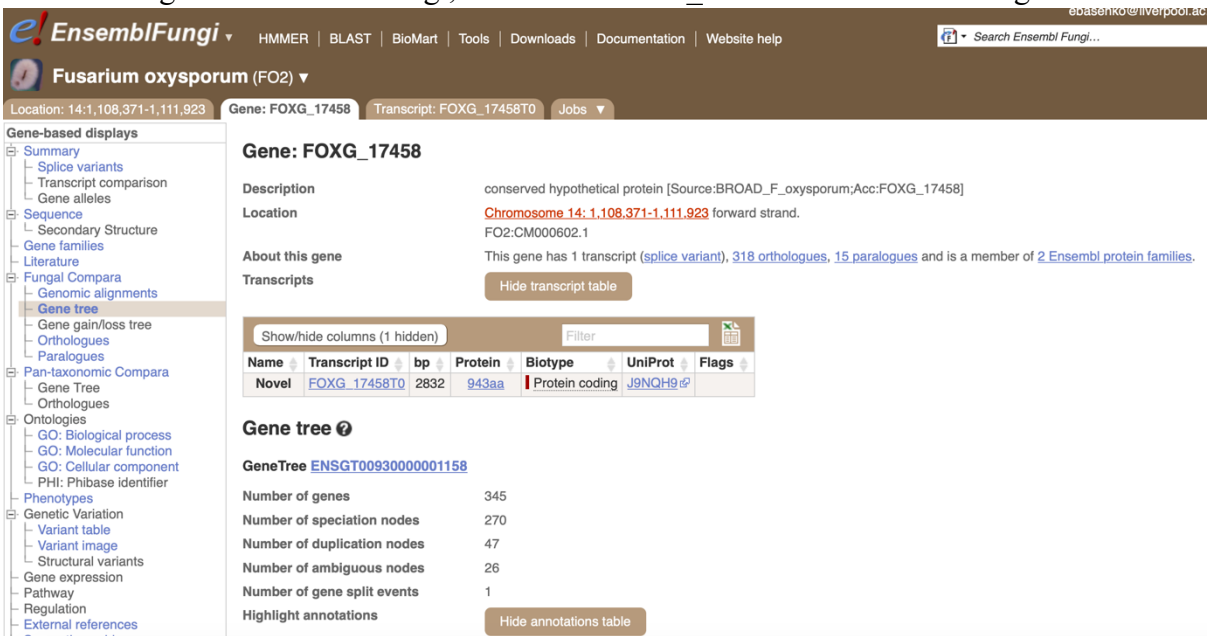


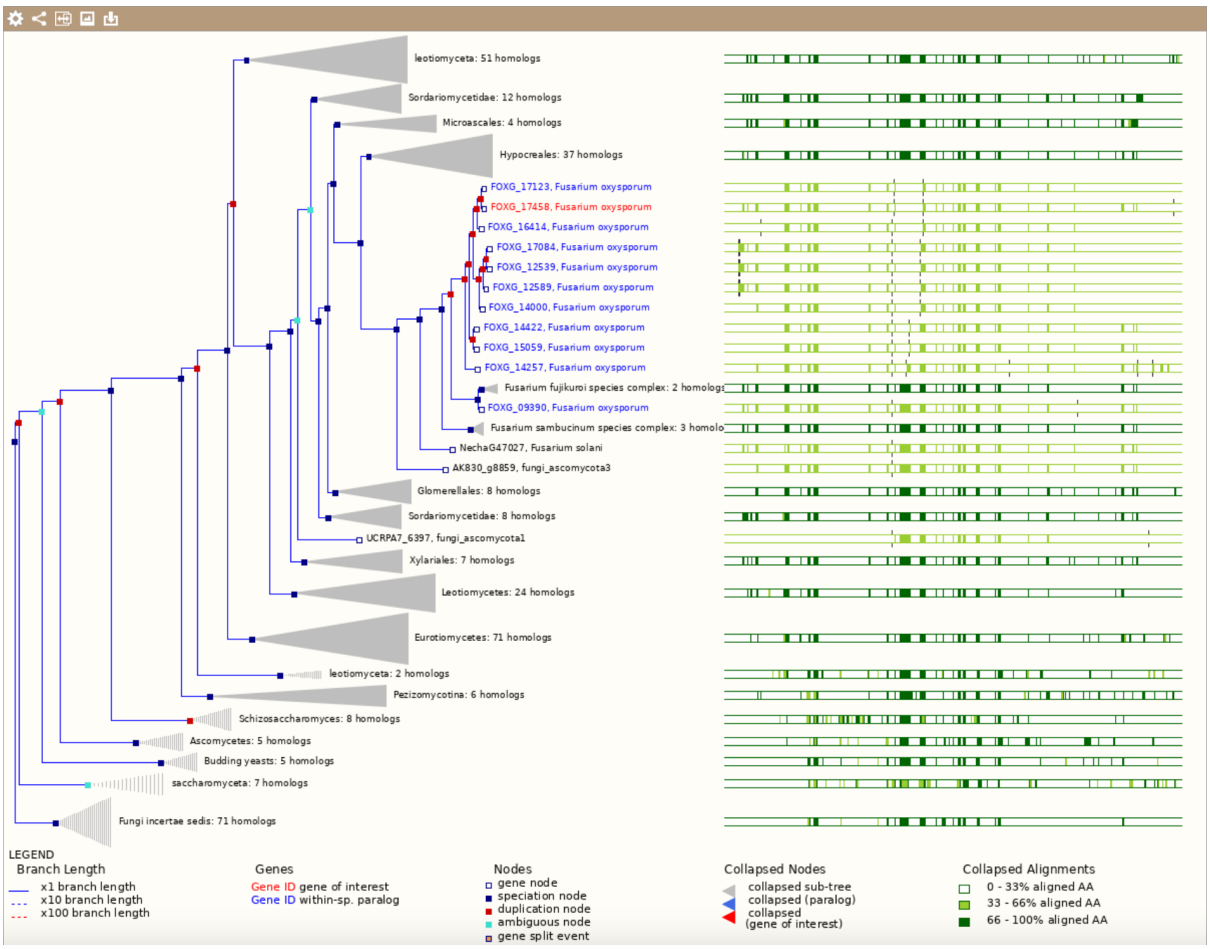
This will bring up cluster #386. Notice that this family is expanded only in the two known pathogens of the *Fusarium oxysporum* species complex with dispensable chromosomes (Fusox2 and FusoxT415), but not in other Fusarium species including endophytic *Fusarium oxysporum* like Fusoxys1 and Fusoxy1.

Now having this information at hand, you can either return to FungiDB and examine underlying transcriptomics, proteomics, *etc.* data or use other databases to enrich your analysis. For example:

- Navigate to Ensembl Fungi, search for FOXG_17458 and visualize the gene-tree:

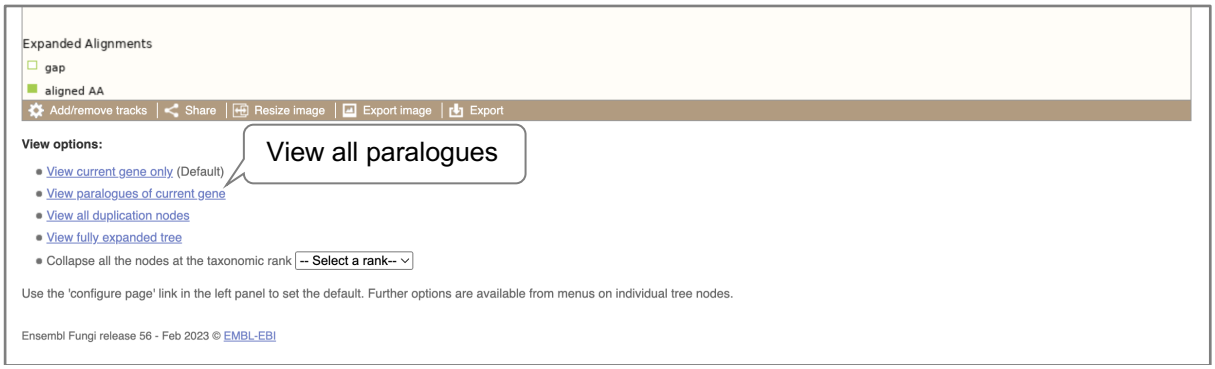
- Navigate to Ensembl Fungi, search for FOXG_17458 and visualize the gene-tree:

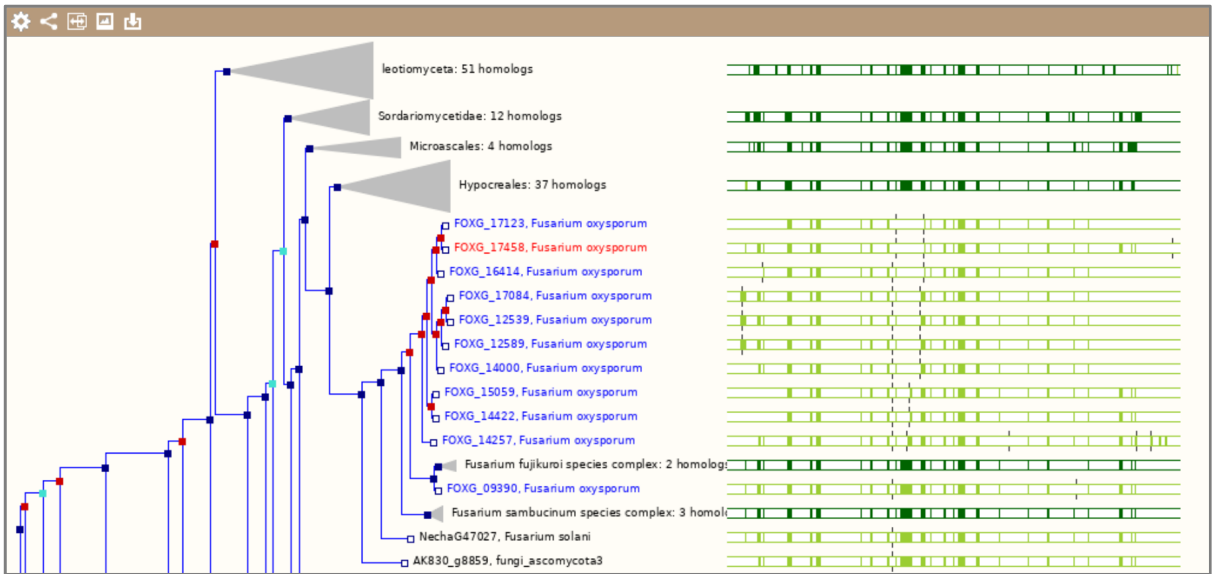




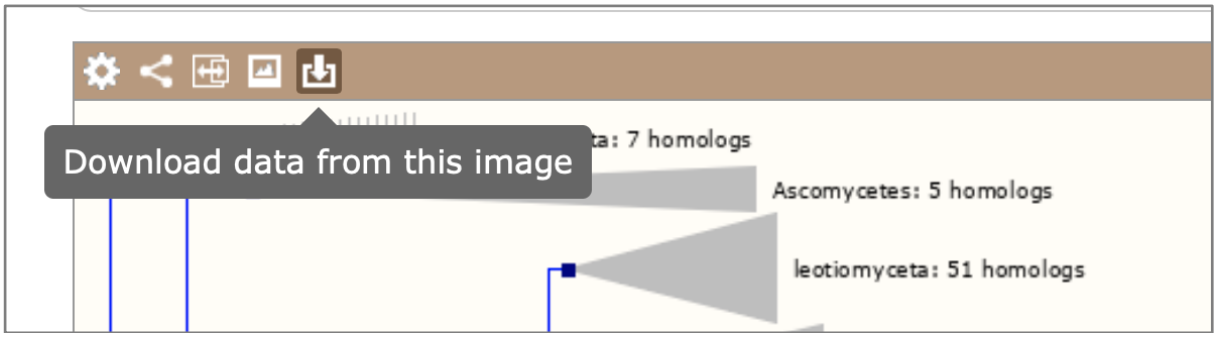
Take a look at the between species paralogues. Is your data consistent with observations in MycoCosm? (Hint: look for duplication nodes).

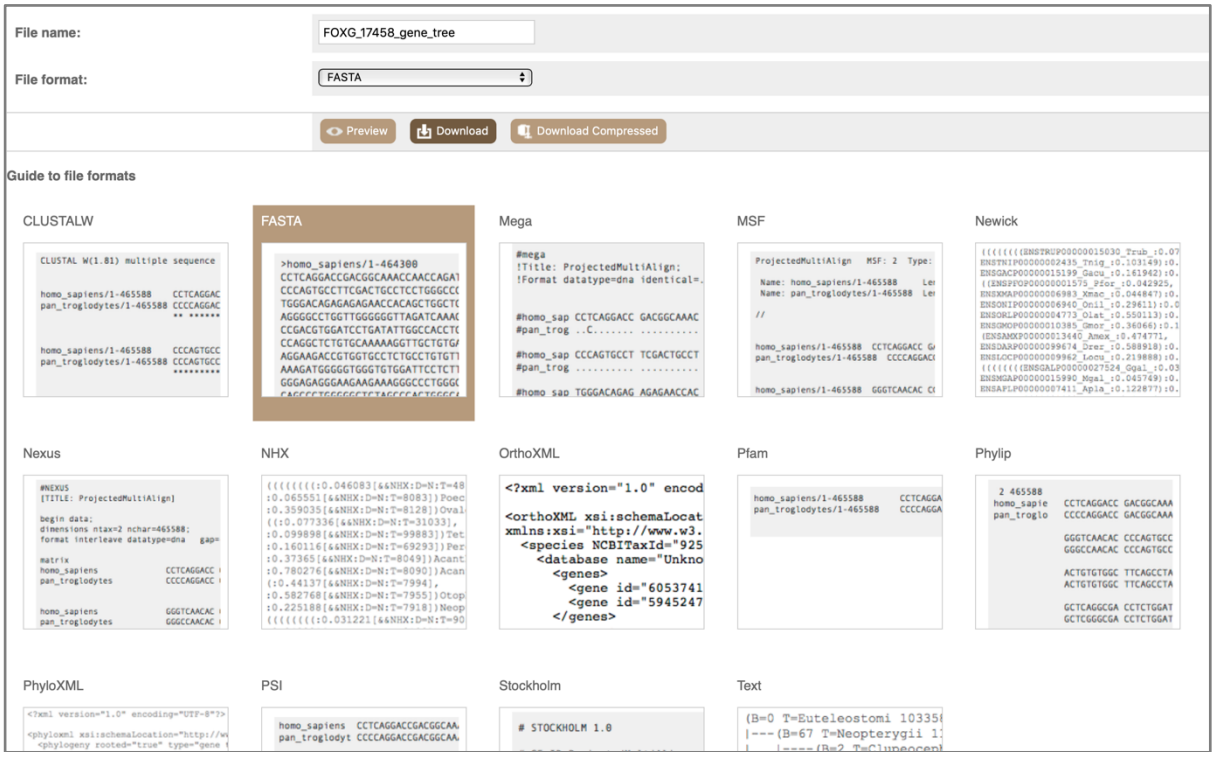
- Click on the link at the bottom of the gene tree image to view all paralogues on the tree:



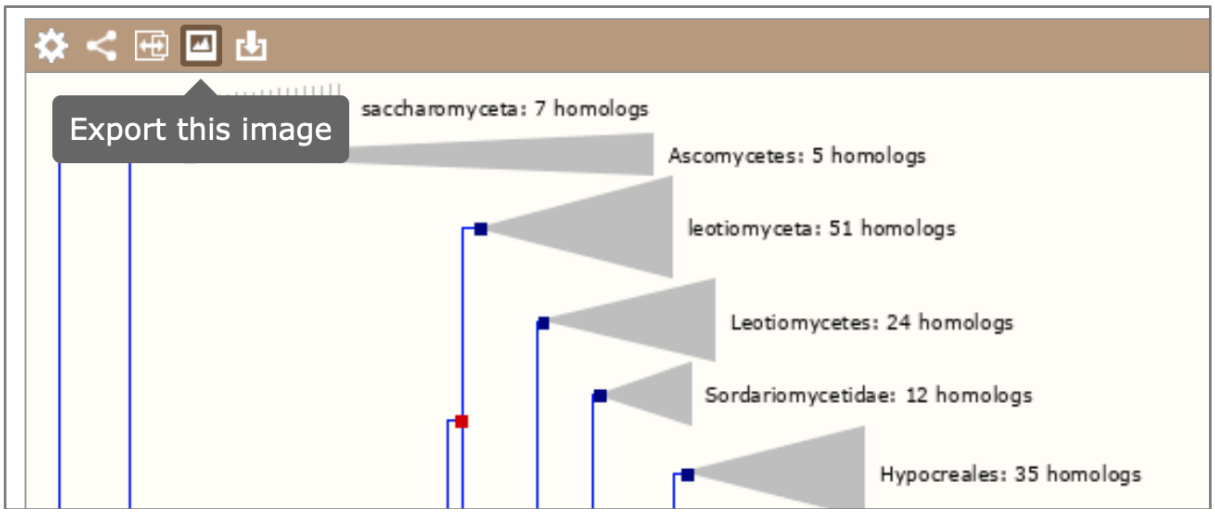


- To export this data you can click on the *Download data for this image* button and choose form multiple formats:





Or you can choose to download the image as shown by clicking on the Export this image button:



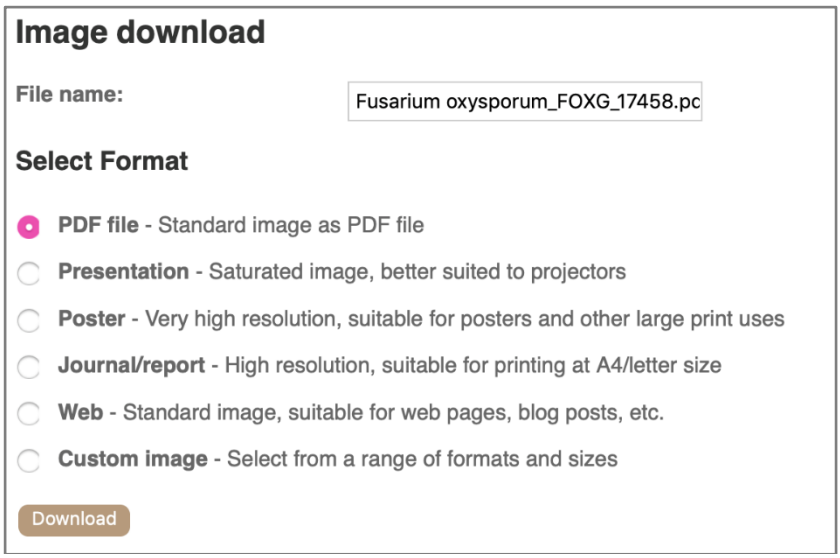
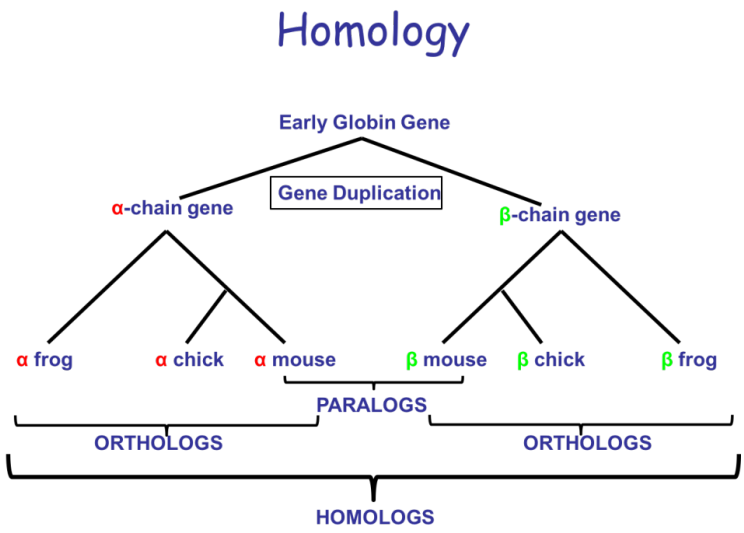


Image download
File name:
Fusarium oxysporum_FOXG_17458.pc
Select Format
PDF file - Standard image as PDF file
Presentation - Saturated image, better suited to projectors
Poster - Very high resolution, suitable for posters and other large print uses
Journal/report - High resolution, suitable for printing at A4/letter size
Web - Standard image, suitable for web pages, blog posts, etc.
Custom image - Select from a range of formats and sizes
Download

FungiDB & OrthoMCL: Orthology and Phyletic Patterns

Learning objectives:

- Run searches in OrthoMCL.
- Run phyletic pattern searches using check boxes or an expression.
- Combine searches using the strategy system.
- Explore individual ortholog group pages.
- Explore the group cluster graphs.



About OrthoMCL.

OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. It provides not only groups shared by two or more species/genomes, but also groups representing species-specific gene expansion families. Thus, it serves as an important utility for automated eukaryotic genome annotation. OrthoMCL starts with reciprocal best hits within each genome as potential in-paralog/recent paralog pairs and reciprocal best hits across any two genomes as potential ortholog pairs. Related proteins are interlinked in a similarity graph. Then, MCL (Markov Clustering algorithm; Dongen 2000; www.micans.org/mcl) is invoked to split mega-clusters. This process is analogous to the manual review in COG construction. MCL clustering is based on weights between each pair of proteins, so to correct for differences in evolutionary distance the weights are normalized before running MCL.

Background on Orthology and Prediction

Orthologs are homologs separated by speciation events. Paralogs are homologs separated by duplication events. Detection of orthologs is becoming much more important with the rapid progress in genome sequencing (Glover et al. 2019).

OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. It provides not only groups shared by two or more species/genomes, but also groups representing species-specific gene expansion families. Thus, it serves as an important utility for automated eukaryotic genome annotation. OrthoMCL starts with reciprocal best hits within each genome as potential in-paralog/recent paralog pairs and reciprocal best hits across any two genomes as potential ortholog pairs. Related proteins are interlinked in a similarity graph. Then, MCL (Markov Clustering algorithm; Dongen 2000;

www.micans.org/mcl) is invoked to split mega-clusters. This process is analogous to the manual review in COG construction. MCL clustering is based on weights between each pair of proteins, so to correct for differences in evolutionary distance the weights are normalized before running MCL.

OrthoMCL is similar to the INPARANOID algorithm (Remm et al. 2001) but is extended to cluster orthologs from multiple species. OrthoMCL clusters are coherent with groups identified by EGO (Lee et al. 2002), and an analysis using EC number suggests a high degree of reliability (Li et al. 2003).

We evaluated the performance of seven widely-used orthology detection algorithms that use three general prediction strategies: phylogeny-based, evolutionary distance-based and BLAST-based (Chen, et al. 2007). Specifically, we used Latent Class Analysis (LCA), a statistical technique appropriate for testing large data sets when no gold standard is available. Our results show an overall trade-off between sensitivity and specificity among these algorithms, with INPARANOID and OrthoMCL performing best with False Positive (FP) and False Negative (FN) error rates lower than 20%.

Method for Forming and Expanding Ortholog Groups in OrthoMCL.
Proteins are placed into Ortholog Groups by the following steps:

1. The OrthoMCL algorithm (see below) is employed on proteins from a set of 150 Core species to form Core ortholog groups. These species were carefully chosen based on proteome quality and widespread placement across the tree of life. Each Core protein is placed by the algorithm into a Core ortholog group consisting of one or more proteins. Core group names have the format OG6_xxxxxx (e.g., OG6_101327). OG6 refers to OrthoMCL release 6; for each sub-release (e.g., 6.1, 6.2, etc), the Core species and the Core ortholog group names will remain constant.

2. The proteins from hundreds of additional organisms, termed Peripheral organisms, are mapped into the Core groups. To do this, NCBI BLASTP is used to compare each Peripheral protein to each Core protein in the Core groups. (Note that Peripheral proteins that were previously added to the Core group are NOT used in the BLASTP.) Then, each Peripheral protein is assigned to the Core group containing the Core protein with the best BLAST score, but only if the E-Value is <1e-5 and the percent match length is >=50%.

3. All Peripheral proteins that fail to map to a Core group are collected and subjected to independent OrthoMCL analysis, forming Residual groups consisting of one or more proteins. Residual group names have the format OG6r1_xxxxxx (e.g., OG6r1_101327), where OG6 refers to release 6 and r1 refers to sub-release 1.

4. For each subsequent sub-release (which will occur every ~3 months along with other VEuPathDB sites), proteomes from additional Peripheral organisms will be processed as in steps 2 and 3 above. However, step 3 will differ slightly because the previous set of Residual groups will be disassembled, leaving the previous unmapped Peripheral proteins to be combined with the new unmapped Peripheral proteins. All of these proteins will be used to form new Residual groups (e.g., OG6r2_xxxxxx).

5. During a sub-release, the proteomes of some species will be updated to the latest version. This can be easily done for a Peripheral species: the old set of proteins are removed from ortholog groups and then the new set is mapped into groups as above. However, this is not possible for Core species because these proteins are used to define Core groups. Thus, the Core species with the older proteome remains on the site but is superficially retired by appending its abbreviation with -old (e.g., aaeg becomes aaeg-old). Then, the latest version of the proteome is mapped in as a peripheral species and obtains the original species abbreviation (e.g., aaeg is a peripheral with a more recent proteome than aaeg-old). These retired species will be eliminated fully when a new set of Core species is defined, as described in the next point.

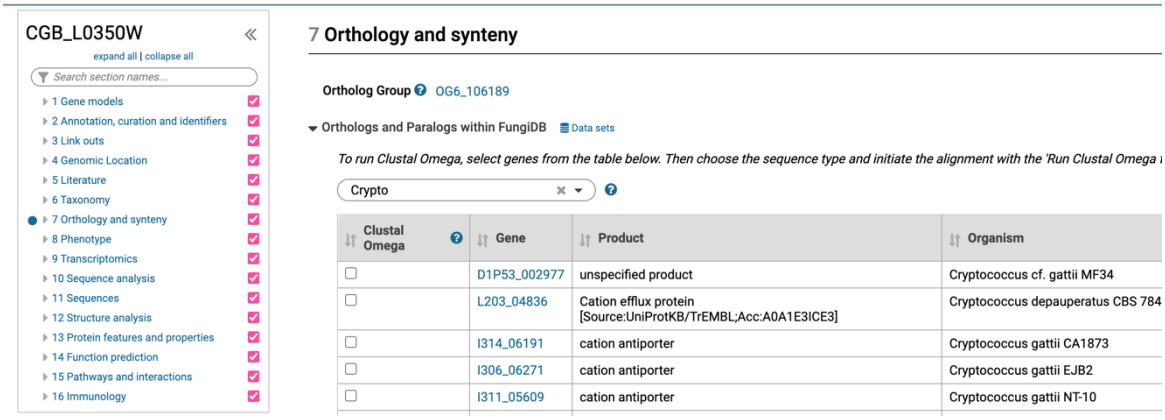
6. On occasion, the set of Core species will be re-defined, as more appropriate proteomes become available and/or when a large number of Core species are retired. In this case, new Core groups (e.g., OG7_xxxxxx) and Residual groups (e.g., OG7r1_xxxxxx) will be formed from the latest version of proteomes from a carefully-chosen set of core species.

This design allows for the addition of proteomes at every sub-release (e.g., 6.1, 6.2, etc). Note that Core groups (e.g., OG6_101327) will remain between sub-releases, though these groups will expand as Peripheral proteins are mapped in. In contrast, Residual groups will exist only for that sub-release; thus, Residual groups are useful in allowing the user to find proteins related to their protein(s) of interest, but are not stable groups.

Examining OrthoMCL output on gene record pages in FungiDB

- **Go to the gene record page for the CGB_L0350W, a hypothetical protein CNBL0590.**

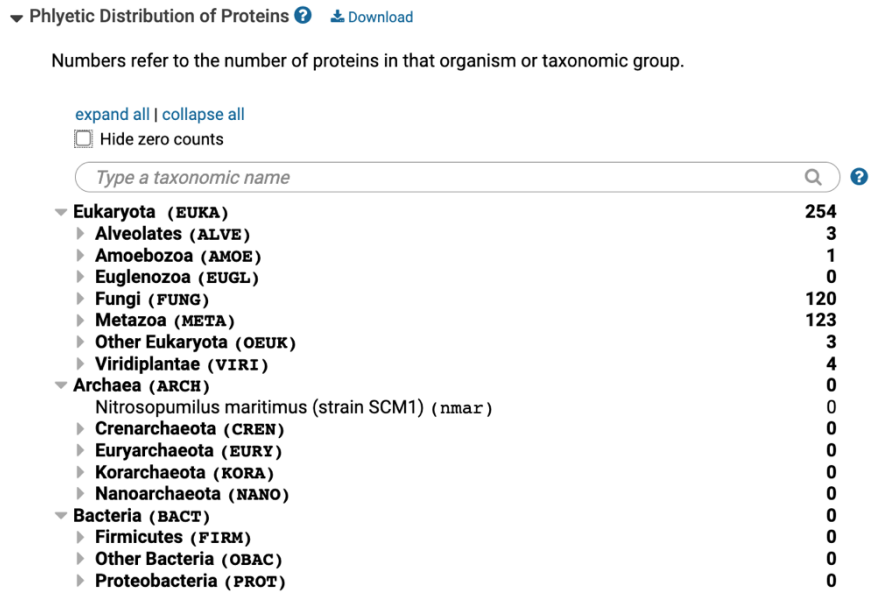
    a. What is the function of this gene? How can you infer its function?
        i. Click on the "Orthology and Synteny" link in the Contents menu on the left. Does this gene have orthologs in other *Cryptococcus* species?



b. Examine evidence in the “Function prediction” section.
c. What about other organisms outside fungi? (Hint: click on the Ortholog Group OG6_106189).
d. The OrthoMCL group page is divided into 5 sections:
    1. Phyletic distribution
    2. Group summary
    3. List of proteins
    4. PFam domains
    5. Cluster graph

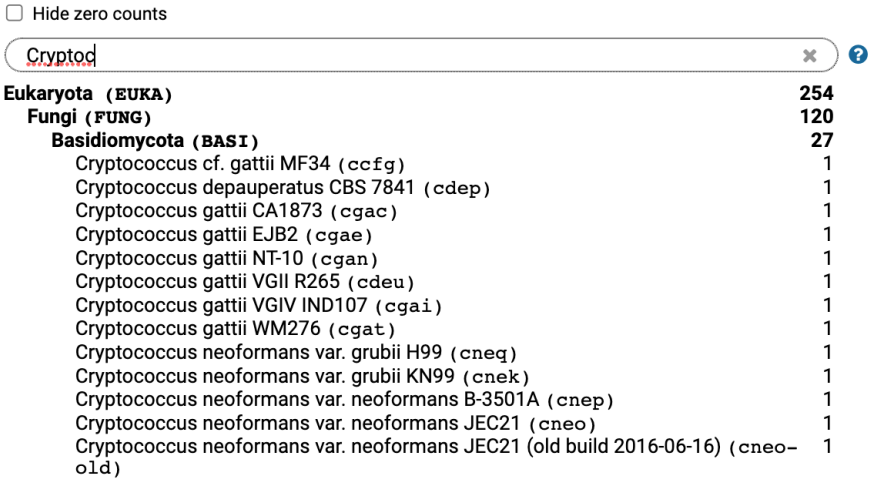
- **Does this protein have othologs in Archae and Bacteria?**

Phyletic distribution: Numbers refer to the number of proteins in that organism or taxonomic group. In order to see organisms and taxonomic groups without proteins in this ortholog group, uncheck 'Hide zero counts.'

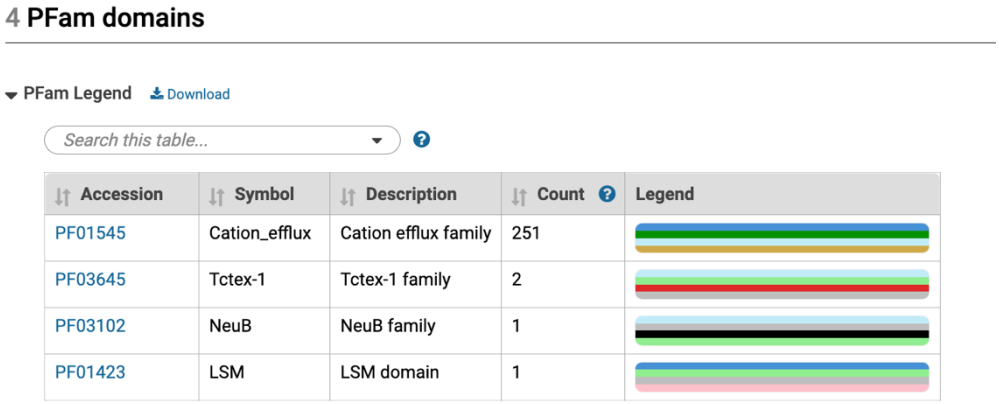


Group summary breaks down summary by protein types: A core protein is from one of the 150 core species that were initially used to form 'core' groups. A peripheral protein is from a peripheral species whose entire proteome was mapped into the 'core' groups. Peripheral proteins that do not map into a 'core' group are placed into residuals groups.

- **Do all *Cryptococcus* species currently integrated in FungiDB contain this protein?**



- **What is the most common PFAM domain associated with the proteins in this group?**

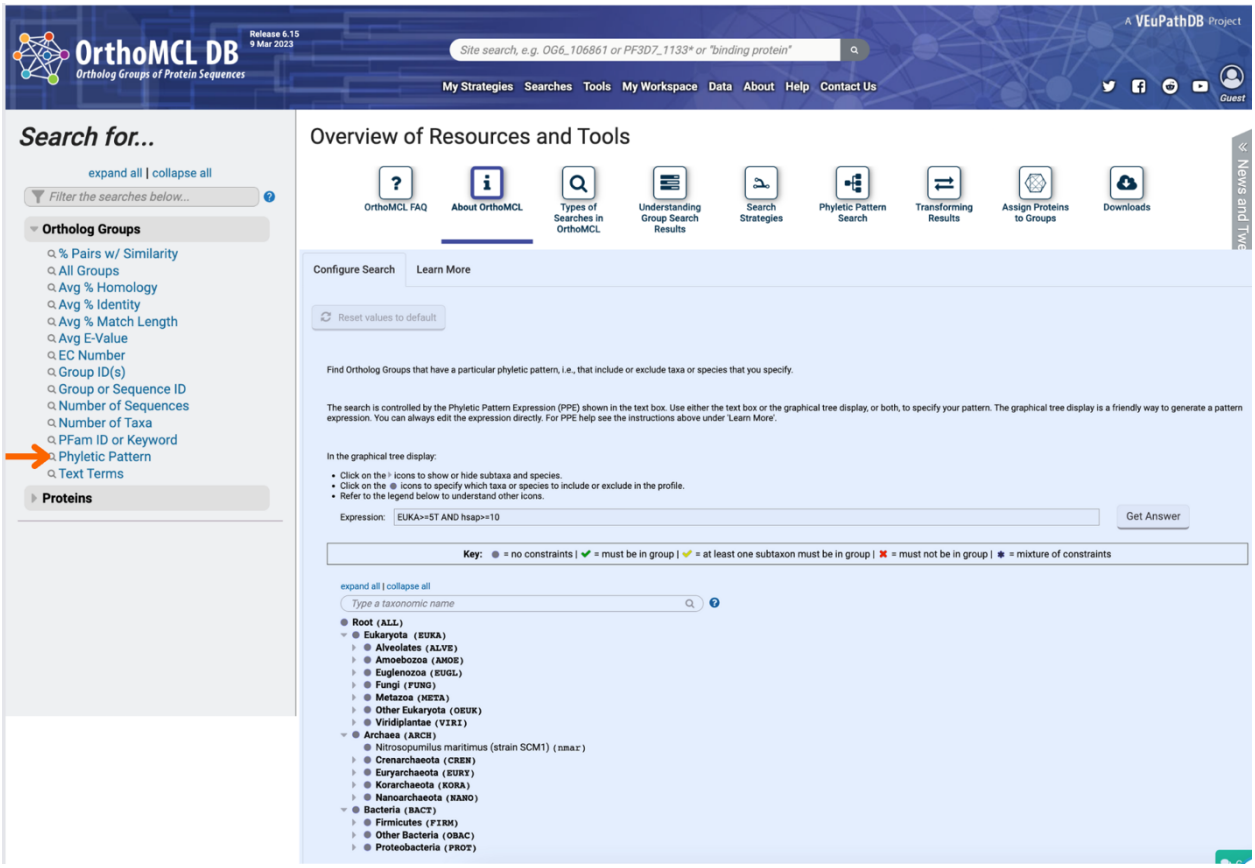


- **How can you look up protein alignments for *Cryptococcus*?**
(Hint: run ClustalOmerga tool and use the "Search this table" filter to limit the alignment to "Cryptococcus").

Using the Phyletic Pattern search in OrthoMCL

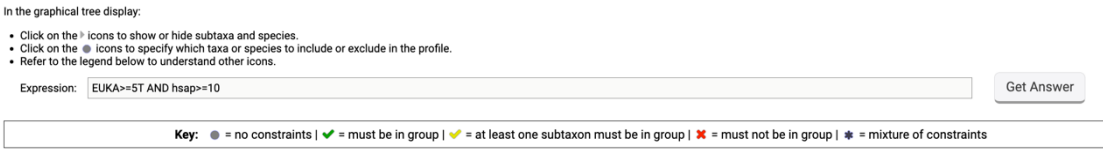
The “Phyletic Pattern” search is an ortholog group search – look under the ortholog groups category and explore the available searches.

- Find the “Phyletic Pattern” search.

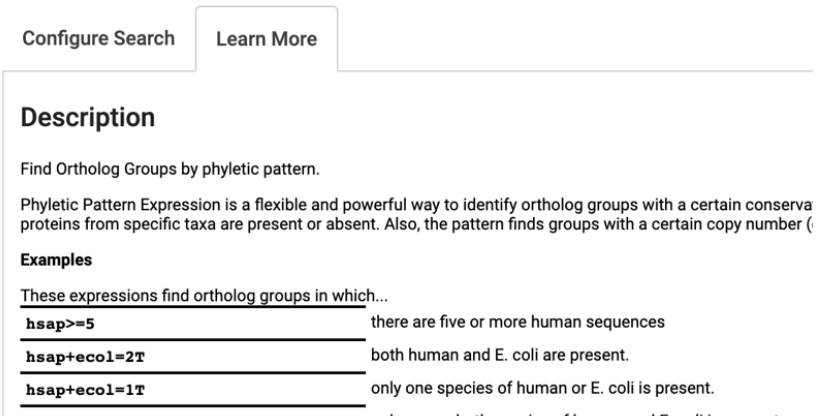


There are two ways to specify a phyletic pattern:

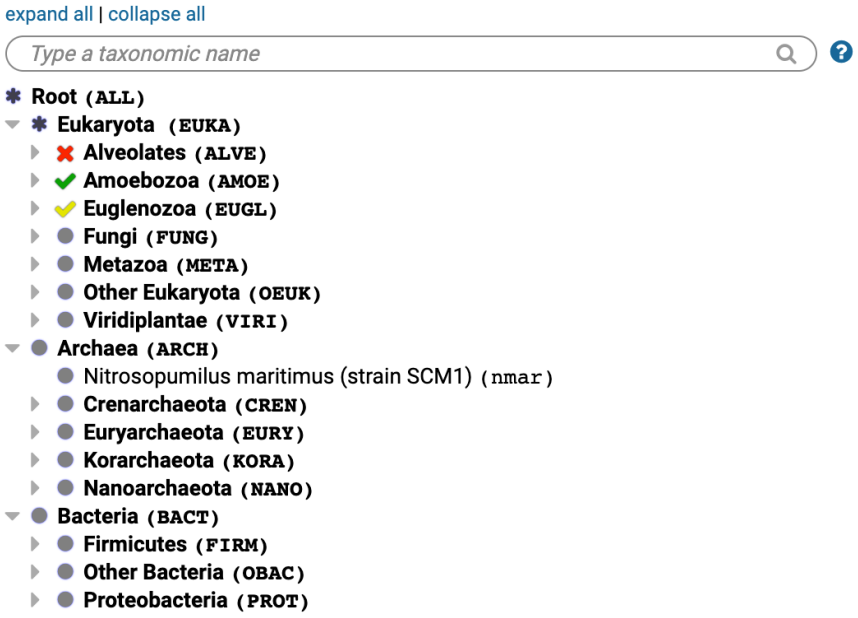
1. Using the expression box.
    - **Run the default search for EUKA>=5T AND hsap>=10.**



    - Use the “Learn More” tab to decipher the expression used above.

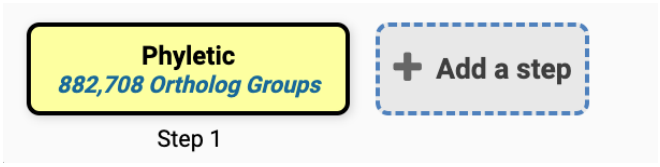


2. Using the selectable tree menu.
   You can click on the circle next to the taxon you want to include or exclude it from the search.



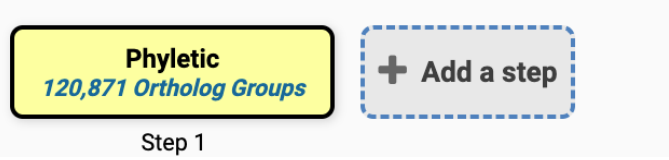
- **Using the "Phyletic pattern" search, identify how many eukaryotic protein groups do not contain orthologs from bacteria and archaea.**

  Hint: leave EUKA class with no constraints.

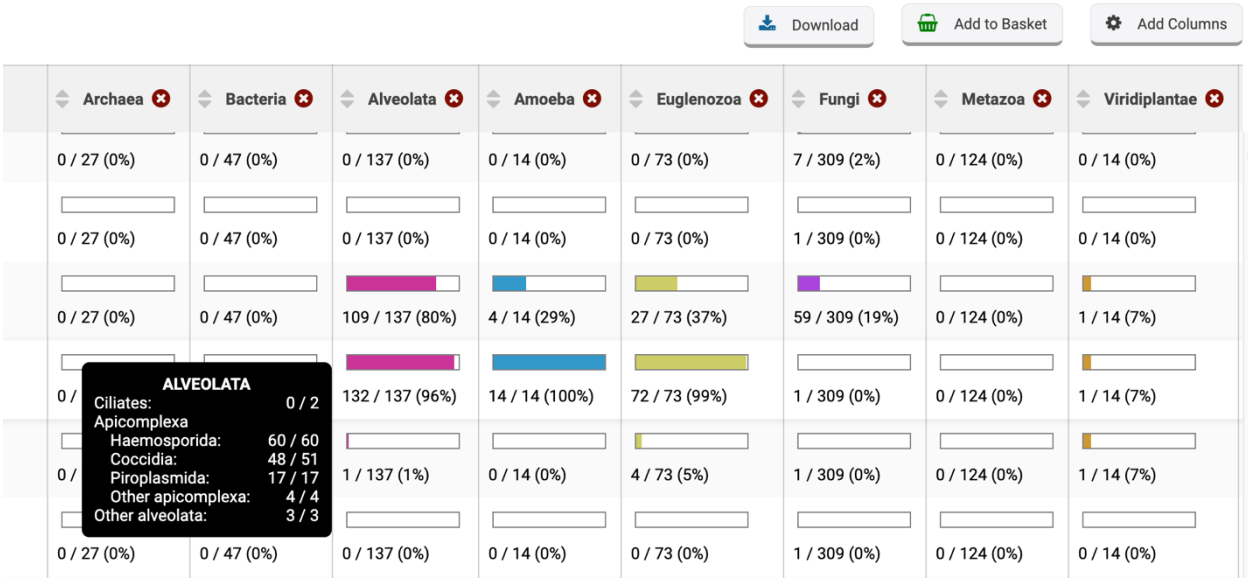


Strategy URL: https://orthomcl.org/orthomcl/app/workspace/strategies/import/eebc49abcff1d99f

- **Find all groups that contain orthologs from <u>at least one species</u> of *Ascomycota fungi* (1T) but <u>not from</u> bacteria, archaea or metazoan (0T).**



- **Examine your results and learn how to interpret the graphical representation for each group.**
  Scroll to the right of the results table examine graphical representation of the results. You can hover over each graph to learn more about phyletic distribution for each class.



Strategy URL: https://orthomcl.org/orthomcl/app/workspace/strategies/import/555afd9c529d4927

- **Revise your search to find groups that:**
  - <u>do not</u> contain orthologs from Alveolates, Amoebozoa, archaea, bacteria and Ascomycetes.
  - contain <u>at least one ortholog group</u> from *Mucor circinelloides* f. *circinelloides* 1006PhL (mcic) **AND** *Mucor circinelloides* f. *lusitanicus* CBS 277.49 (mcir).

Hint: You cannot answer this question by using the check boxes alone. For Mucor, use the expression field to finish the parameter set up manually.

If you are getting frustrated trying to figure this one out, you have a right to be! If your results look different, hover over the search step and click to revise the parameter search. The cool thing about OrthoMCL is that has an added feature that allows you to enter an expression to define the phyletic pattern. This option provides additional flexibility. Can you figure out what expression to use to answer this question? (hint: start by assigning the "do not contain" parameter (x) using check boxes to Alveolates, Amebozoa, archaea, bacteria and Ascomycetes. Next, use the expression window to add "AND" followed by specific criteria for Mucor spp. Use the learn more tab for more information.

Strategy URL: https://orthomcl.org/orthomcl/app/workspace/strategies/import/88e60b823cb2c959

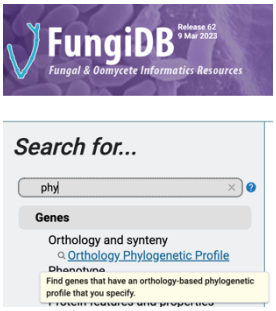
If you ran a search using just check boxes, the search will be configured to look for groups that:

- do not contain orthologs from Alveolates, Amoebozoa, archaea, bacteria and Ascomycetes.
- contain ortholog groups from both *Mucor circinelloides* f. *circinelloides* 1006PhL (mcic) **AND** *Mucor circinelloides* f. *lusitanicus* CBS 277.49 must be present

Strategy URL: https://orthomcl.org/orthomcl/app/workspace/strategies/326574153/430551723

Useful information:

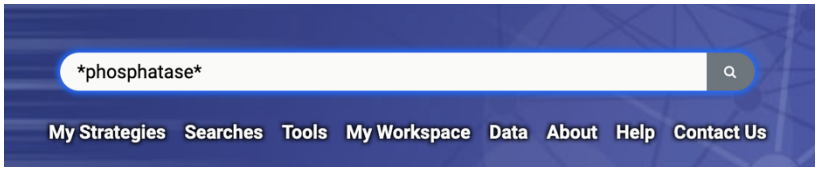
All VEuPathDB genomics sites (e.g., FungiDB) have an integrated phyletic pattern search that uses OrthoMCL to return lists of genes. For example, you use the "Orthology Phylogenetic Profile" search to identify genes in your organism of interest that are restricted in their profile. For example, you frequently want to identify genes that are conserved among organisms in your genus of interest but not present in the host as these genes may make good drug targets or vaccine candidates.



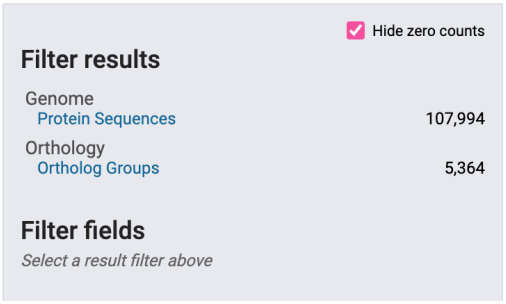
Combining searches in OrthoMCL

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

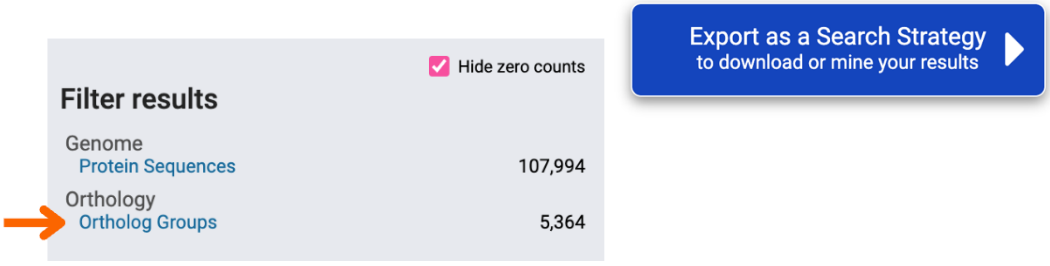
  a. Use the site search to look for *phosphatase* (use asterisks to find any combination of the word "phosphatase").

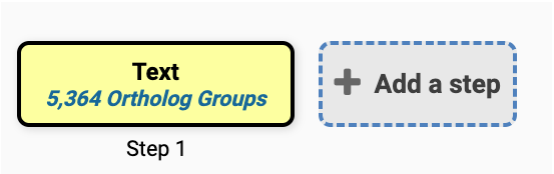


How many protein sequences were identified? How many ortholog groups did you identify?

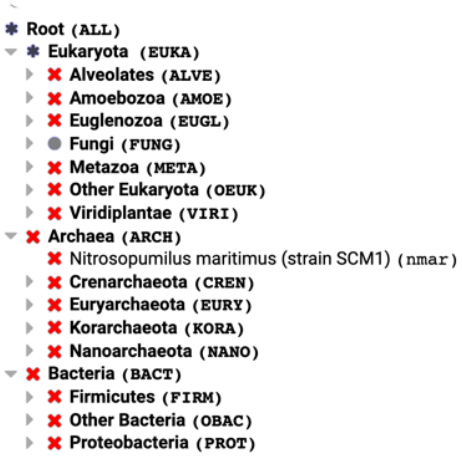


b. Display the ortholog groups containing the word phosphatase and export the resuts as a search strategy.

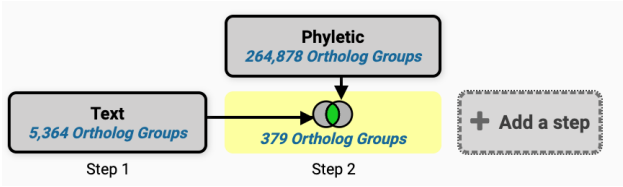




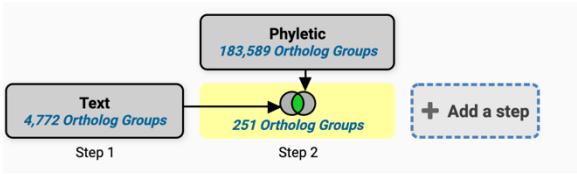
c. Add a step and run a phyletic pattern search for groups that contain any fungi proteins but do not contain any other organism outside fungi. (hint: make sure everything has a red x on it except for fungi, which should be a grey circle (no constrains)).



How many groups did the search return?



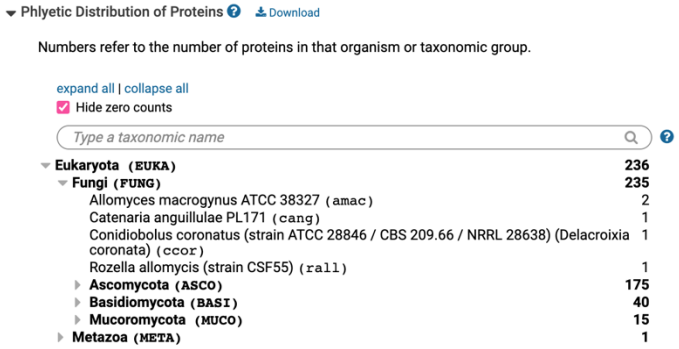
Strategy URL: https://orthomcl.org/orthomcl/app/workspace/strategies/326574223/430551843



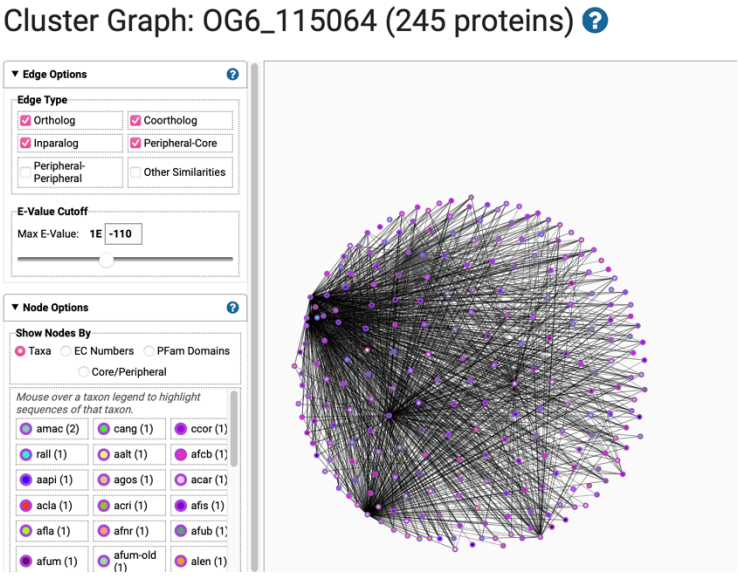
Exploring a specific OrthoMCL group - examining the cluster graph.

- **Visit the OrthoMCL record page for the group OG6_115064.**

- Examine the phyletic distribution tree. What taxa does this group contain?



- Examine the cluster graph for this group (it can be accessed at the bottom of the page)



You can interact with the cluster graph. For example, move the slide to increase the E-value cutoff stringency (e.g., to a more negative number). Can you identify subclusters? Click on the nodes in the graph – notice how the organism is updated on the right.

On the left of the page in the *Node Options* panel, click on PFam Domains to see which proteins have the various PFam domains.

In the *Node Options* panel, you can click on *Core/Peripheral* to observe which proteins were derived from Core species and which proteins were derived from Peripheral species. Proteins from Core species were used in the initial OrthoMCL algorithm to form Core ortholog groups. Proteins from Peripheral species were mapped into these Core groups by sequence similarity (determined by BLAST score).

What is Galaxy?

Galaxy is an open, web-based platform for data analysis under the FAIR principles of data sharing and re-use. Galaxy is an open-source platform that allows you to perform, reproduce, and share complete analyses without the use of command line scripting. The VEuPathDB project developed its own Galaxy instance in collaboration with Globus.

The VEuPathDB Galaxy offers pre-loaded genomes, pre-configured workflows and other tools for private data analysis and display. A custom-built set of tools also allows the ability to export Galaxy results into private workspaces within VEuPathDB sites (My Workspace > My data sets section). The datasets within the "My data sets" workspace can be explored using the FungiDB interface and tools and cross-referenced with the public data integrated in FungiDB.

VEuPathDB Galaxy access requires an account with FungiDB/VEuPathDB. The account is free and can be used to sign-in into any VEuPathDB genomics site.

The Galaxy instance is not meant for long term data storage. Datasets are automatically deleted after 60 days. To save your data, download your analysis results locally and then *delete and purge* files to free up space for your next analysis.

The Galaxy project offers extensive learning materials that can be accessed here: https://wiki.galaxyproject.org/Learn#Galaxy_101

Important: The Galaxy module consists of RNA-Seq and SNP analysis modules. These are concurrent sessions. This exercise will be carried out in groups of 4 people using the workshop Galaxy instance. Please do not use live FungiDB.org for this exercise. The detailed tutorials for both modules are available to all course participants.

RNA sequence data analysis via VEuPathDB Galaxy, Part I

Learning objectives:

- Become familiar with the VEuPathDB Galaxy workspace.
- Create collections of datasets from the pre-loaded data.
- Run a pre-configured RNA-Seq workflow.

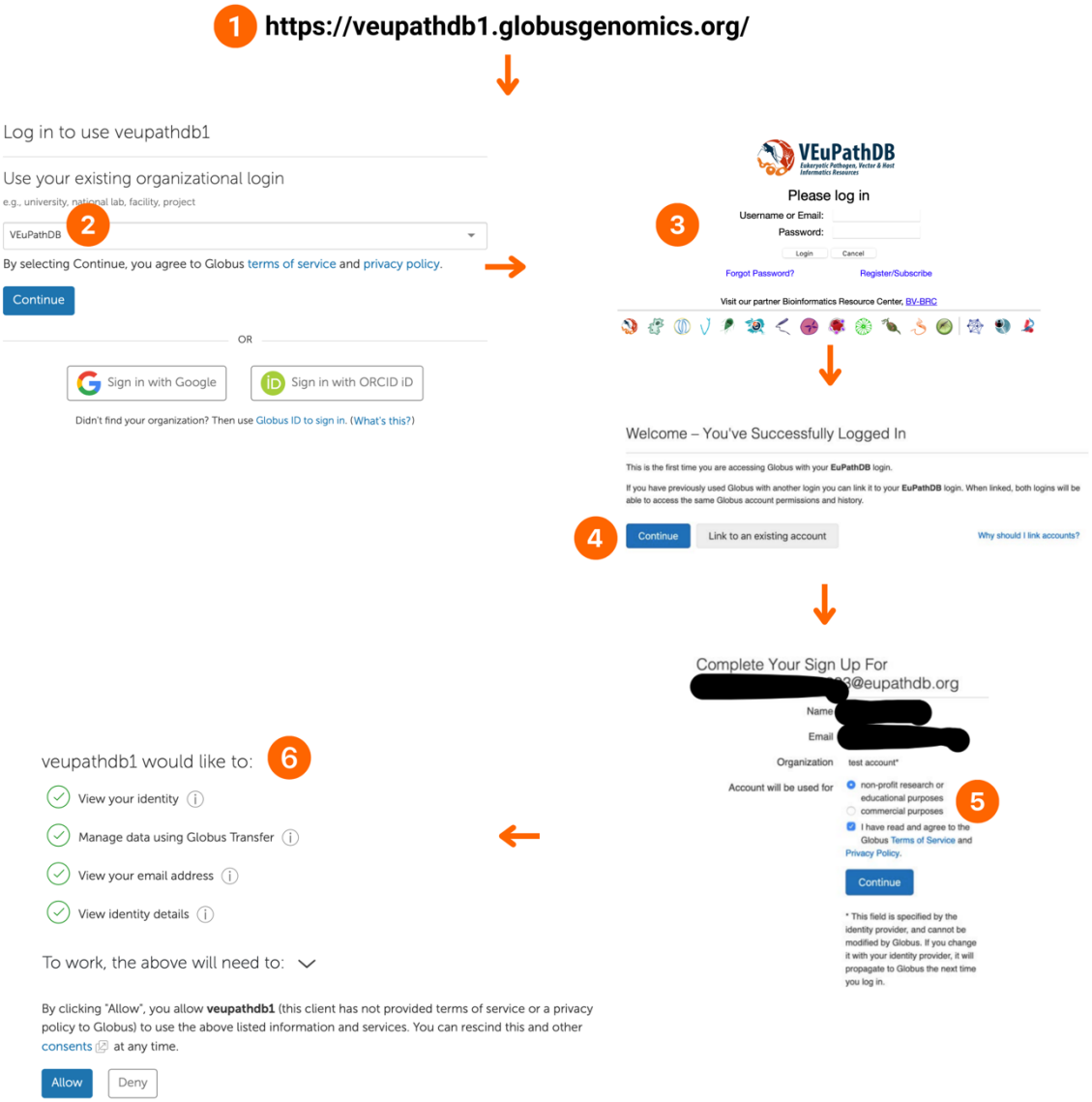
For this exercise, we will retrieve raw sequence files from the "shared history" section in VEuPathDB Galaxy and then run files through a pre-configured RNA-Seq workflow that will align the data to a reference genome, calculate expression values and determine differential expression.

Important: We will be working in groups of four people but only one person in each group should download data and deploy the pre-configured workflow. The other members' roles are to ensure that the correct datasets are used and that the correct workflow parameters are selected. In the Part 2 of this exercise, everyone will get a copy of the workflow output and practice how to perform data analysis.

- **Access the VEuPathDB Galaxy workshop instance.**

If you do not have an account with VEuPathDB/FungiDB, please create one now.

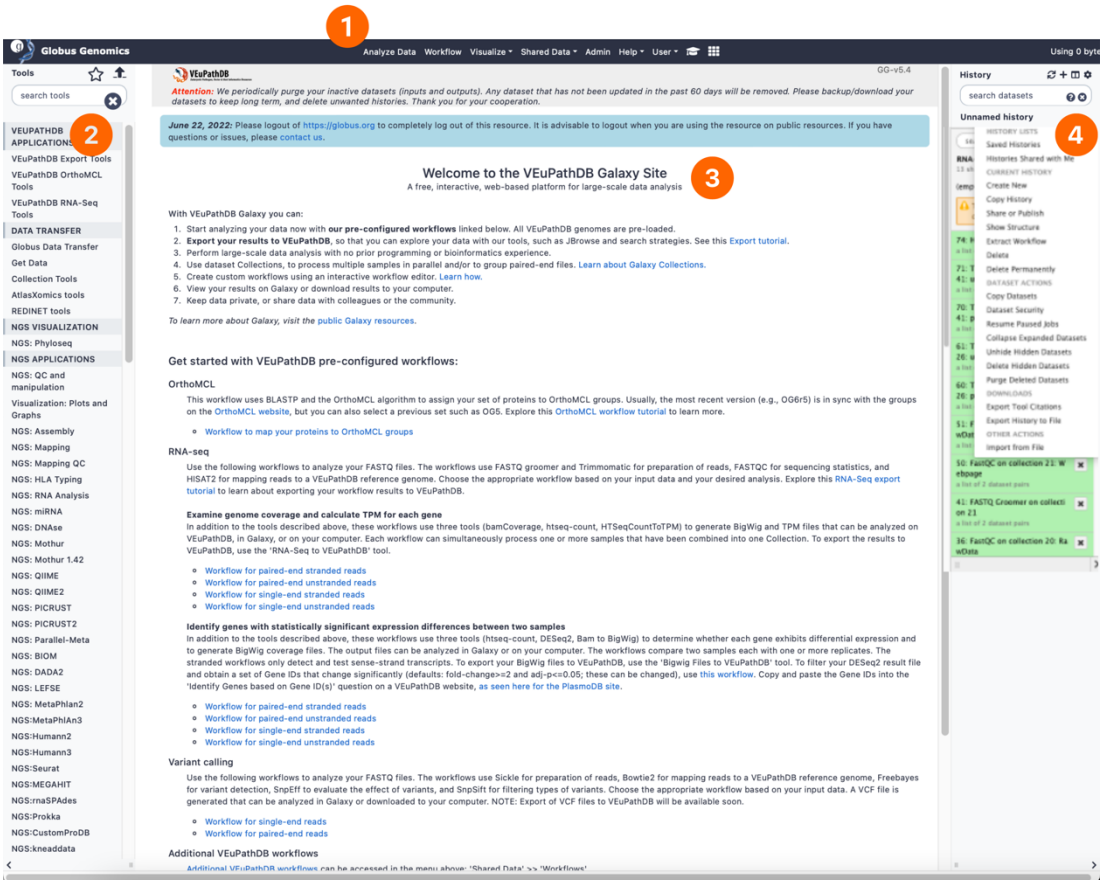
1. Click on the following URL to begin: https://veupathdb1.globusgenomics.org/
2. On the next page, you will be asked to define your organization. Choose the "VEuPathDB" option and click on the "Continue" button.
3. If you are not already logged into VEuPathDB, you will be prompted to do so.
4. Click on "Continue" on the next page (no need to link an existing account).
5. Select the "non-profit" option and agree to the Terms of Service. Click continue.
6. The next page will ask for permissions required to use this Galaxy instance. Click on "Allow"



The anatomy of the VEuPathDB Galaxy landing page.

The workspace has four major components:

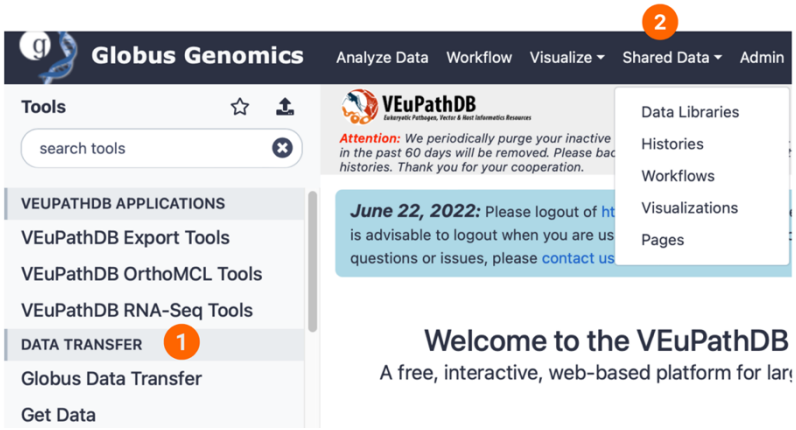
1. The top menu controls the main interface, provides access to the landing page, shared data, public and private workflows & more.
2. The left panel has a list of available tools where the VEuPathDB export tools are listed at the top.
3. The main welcome (landing) page is the interactive interface that houses pre-configured workflows, workflows editor, etc.
4. The panel on the right provides access to histories, deleted datasets, and other useful functions, including options to delete and purge datasets.



Don't see a tool you need for your research? – Let us know by sending an email to help@fungidb.org

Importing data for your workflow.

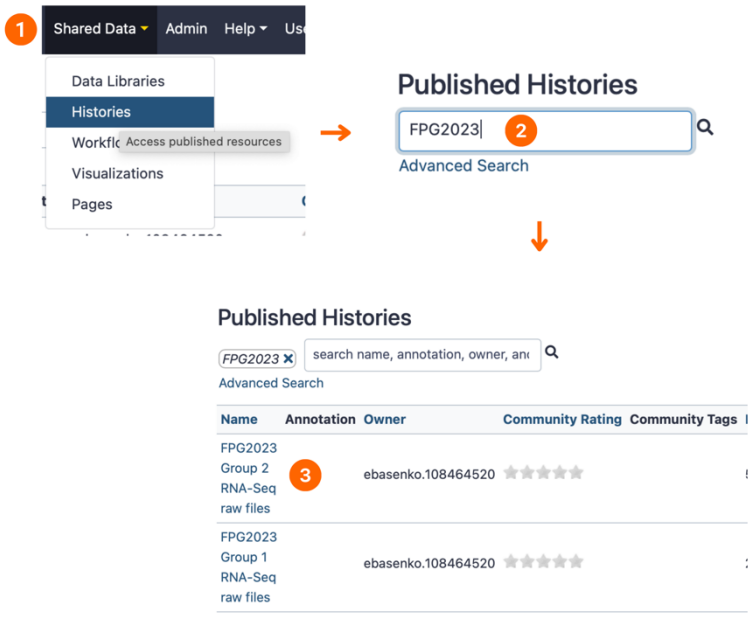
There are multiple ways to important data into your Galaxy workspace. You can transfer data via tools located under the "Data Transfer" section in menu on the left (1). You can also transfer data from the "Shared Data" section in the main menu (2). The latter prides access to pre-loaded raw data, publicly shared workflows, or workflow results (histories), etc.



For this exercise, pre-loaded raw files should be imported from the "Shared Data" > Histories.
Only one person per each group should import data files and deploy an RNA-Seq workflow. Everyone will practice data analysis in NGS Part 2 module. For group assignments, see below.

- **Import data for your RNA-Seq workflow via the Shared histories option.**

1. From the top menu, select "Shared Data > Histories" option.
2. Filter all public workflows on "FPG2023" .
3. Click on the history link that correspond to your group number to import the data into your Galaxy workspace.



Group assignments (see more information about the files below)

Groups 1 & 2 *Aspergillus fumigatus*. Paired-end data. Analyze transcriptomes from cells incubated in human blood (B) and defined minimal media (M) for 30 and 180 min.

Group Number	1	2
Comparison	M30 vs B30	B30 vs B180
History name for download (in Galaxy)	FPG2023 Group 1 RNA-Seq raw files	FPG2023 Group 2 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-29_AfumigatusAf293_Genome	

Reference: PMID: 26311470 BioProject: PRJNA287921

Group 3 *Candida parapsilopsis.* Paired-end data. Analyze transcriptomes from cells grown under planktonic and biofilm-inducing conditions. Control: planktonic.

Comparison	Planktonic vs Biofilm
History name for download (in Galaxy)	FPG2023 Group3 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-42_CparapsilosisCDC317_Genome

Reference: PMID: 25233198 BioProject: PRJNA246482

Group 4 *Coccidioides posadasii*. Single read data. Analyze transcriptomes from mycelia (non-pathogenic stage) and spherules (pathogenic stage).

Comparison	Mycelia vs Spherules
History name for download (in Galaxy)	FPG2023 Group 4 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-61_CposadasiiSilveira2022_Genome

Reference: PMID: 22911737 BioProject: PRJNA169242

Group 5 *Fusarium graminearum*. Paired-end data. Analyze spore and mycelial transcriptomes.

Comparison	Spores vs Mycelia
History name for download (in Galaxy)	FPG2023 Group 5 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-31_FgraminearumPH-1_Genome

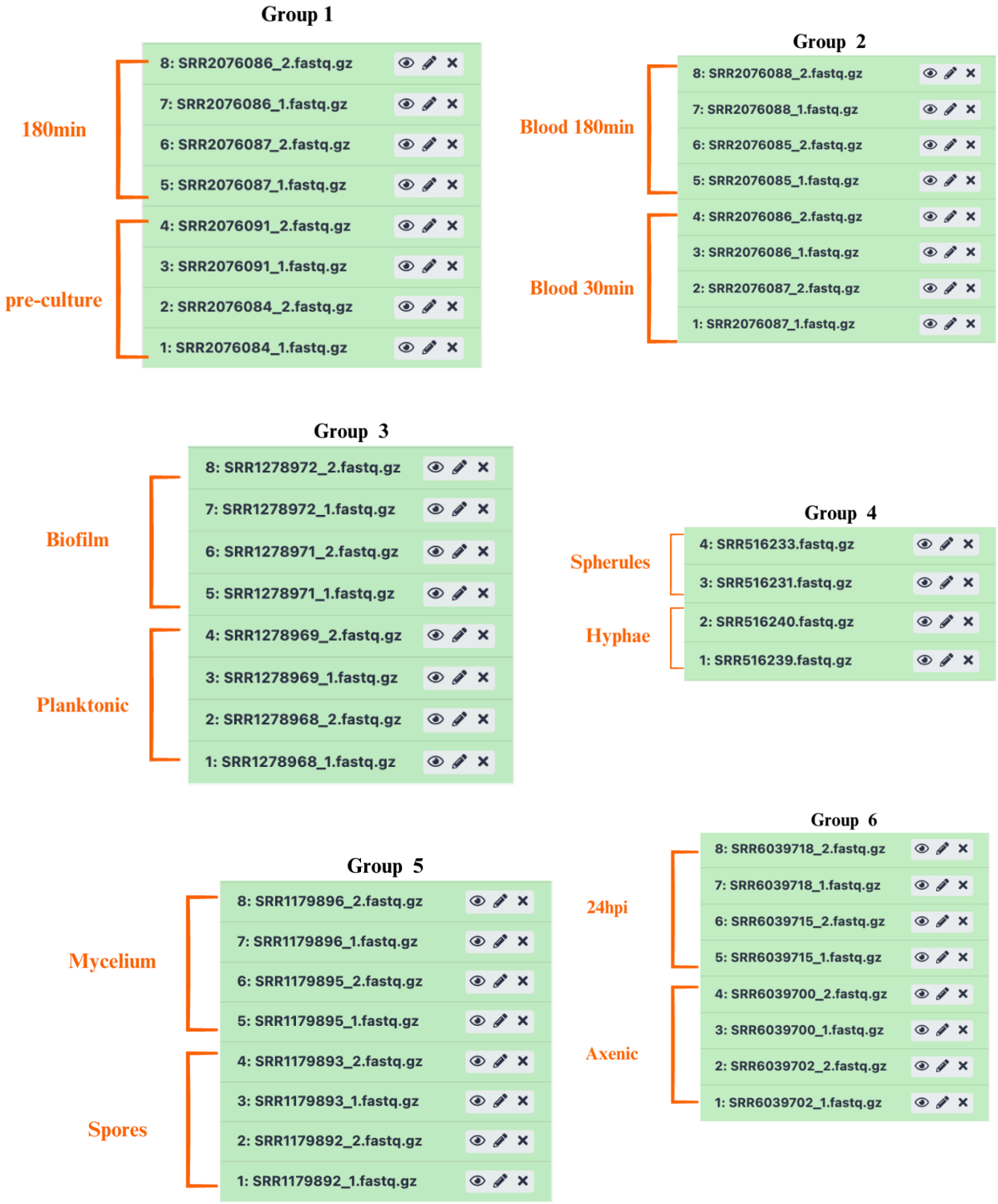
Reference: PMID: 24625133 BioProject: PRJNA239711

Group 6 *Ustilago maydis*. Paired-end data. Analyze transcriptomes from plant-associated development samples (axenic culture vs 12 days post infection (dpi)).

Comparison	0h vs 12 dpi
History name for download (in Galaxy)	FPG2023 Group 6 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-51_Umaydis521_Genome

Reference: PMID: 33653886 BioProject: PRJNA407369

Guide to FPG2023 RNA-Seq histories and file organisation.



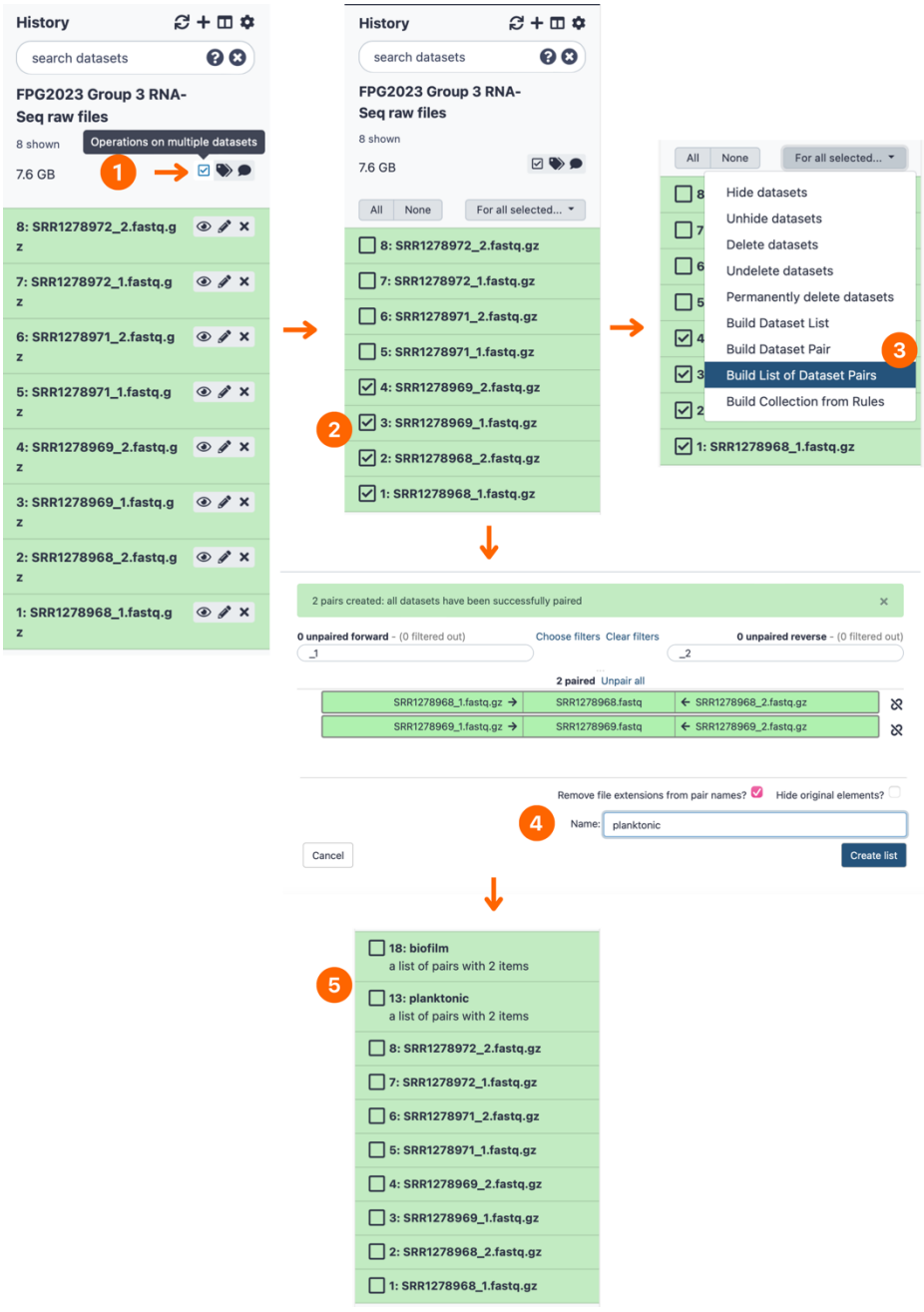
Each dataset contains two replicates. For datasets with multiple samples (e.g., containing biological replicates), it is useful to organize them into "Collections" (e.g., spore and mycelia). Organizing samples with replicates into collections also reduces the complexity Galaxy workflows.

- **Organize samples with replicates into collections:**

1. Click on the checkbox function “operation on multiple datasets”.
2. Select samples that belong to the same condition (control samples will appear at the bottom, see file mapping notes for each group below).
3. Click on “For all selected” and choose “Build List of Dataset Pairs”.

Note: for single read data, choose “Build Data List” option instead.

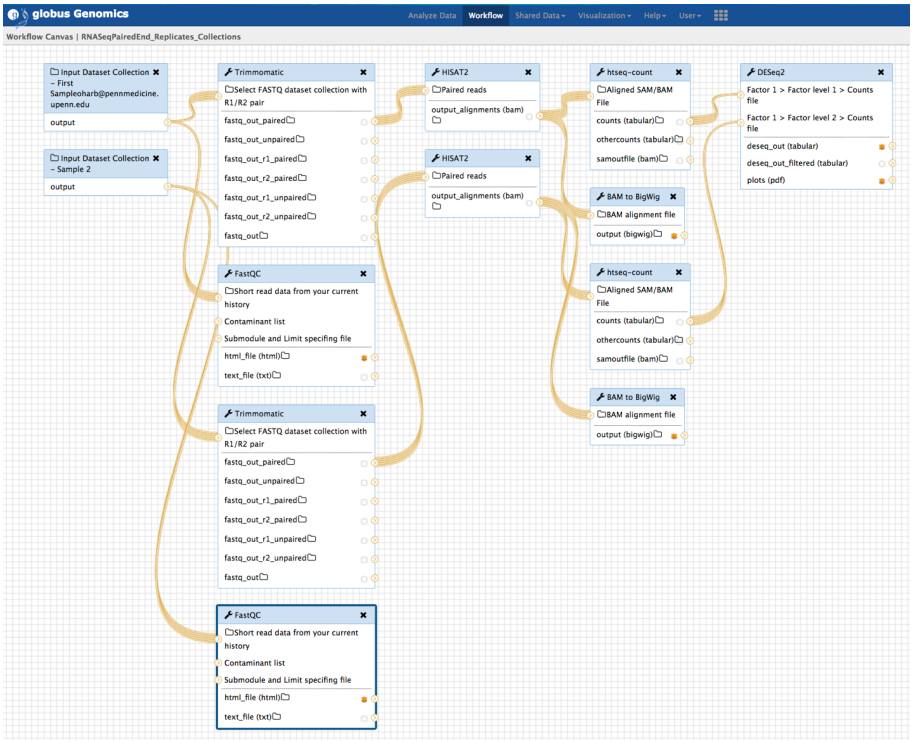
4. Name the sample (e.g. planktonic) and click “Create List”. Note: Usually the correct pairs are auto selected.
5. Repeat for the comparator sample. You should end up with 2 datasets (e.g., planktonic and biofilm).



Running a workflow in Galaxy

You can create your own workflows in galaxy using the tools from the menu on the left. For this exercise we will use a preconfigured workflow that consists of the following steps:

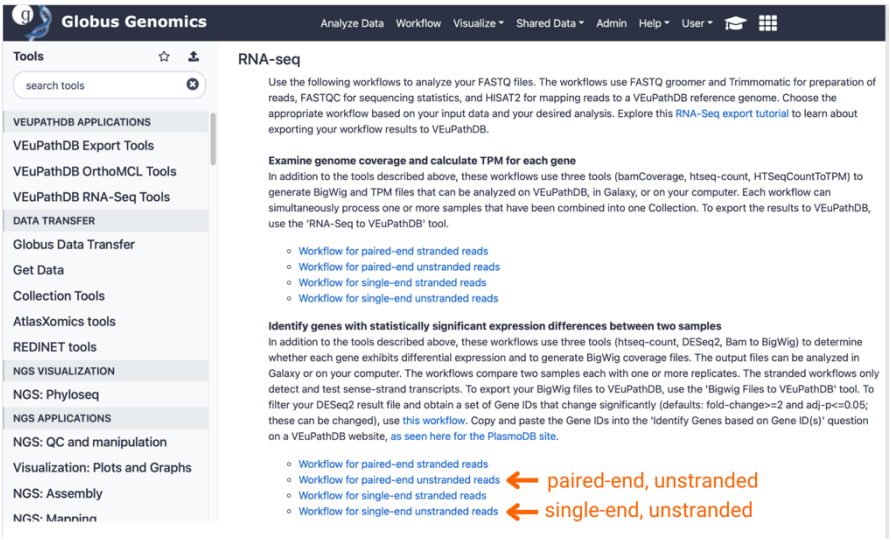
1. Input: raw data, dataset collections.
2. FASTQC: analyse for quality, generate read quality reports.
3. Trimmomatic: trims the reads based on their quality scores and adaptor sequences.
4. HISAT2: align reads to a reference and generate coverage plots.
5. HTSeq: estimate abundance (read counts per gene), generate coverage plots for JBrowse (BAM to BigWig).
6. DESeq2: differential expression of genes between samples.



- **Deploy a pre-configured workflow.**

To do this, navigate to the Galaxy home page and select the workflow appropriate for your dataset:

  - For paired-read datasets choose “Workflow for paired-end unstranded reads”.
  - For single read data, choose “Workflow for single-end unstranded reads”.

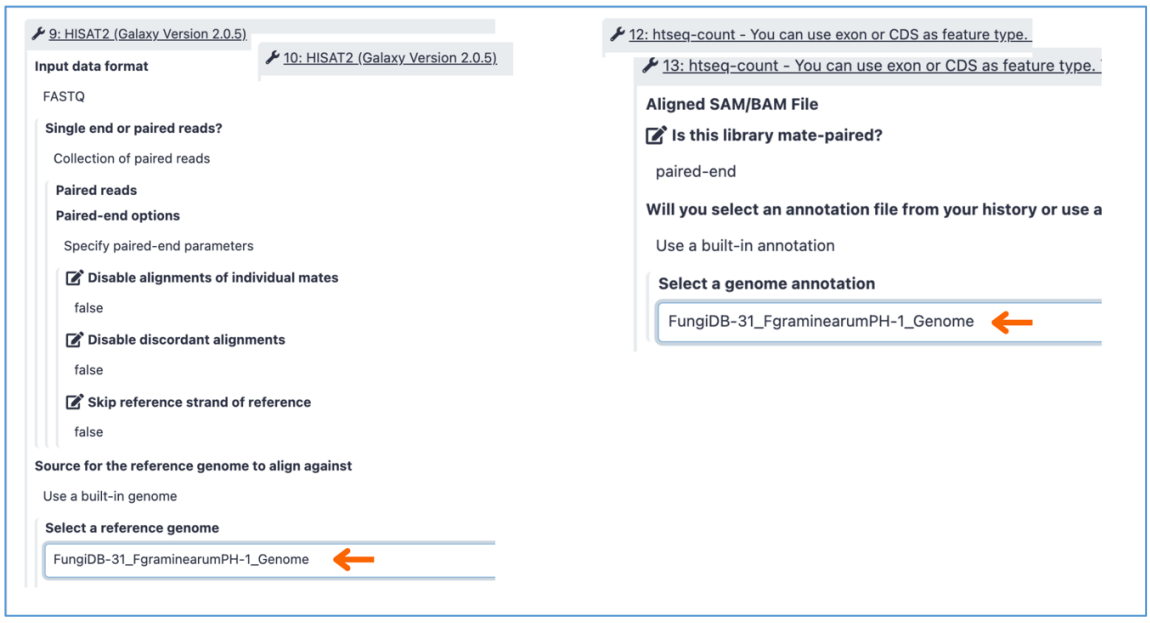


- **Configure an RNA-Seq workflow.**

There are multiple steps in the workflow, but you do not need to configure all of them. For this exercise, you will need to configure the following:

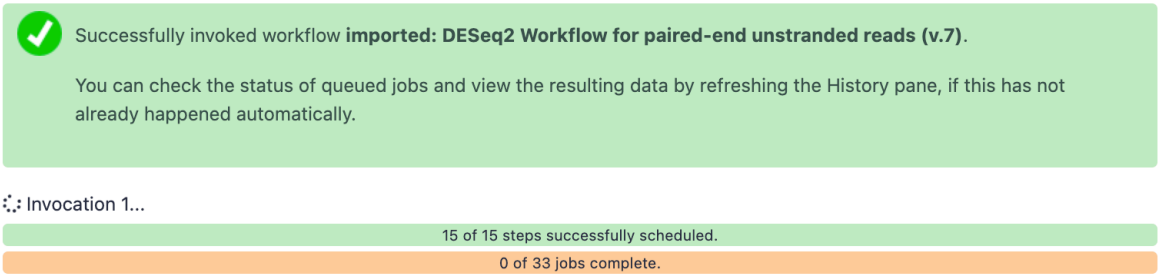
1. Input dataset collection 1 (e.g., planktonic).
2. Input dataset collection 2 (e.g., biofilm).
3. Both HISAT2 steps (requires reference genome – refer to the group assignments section above for this info).
4. Both htseq-count steps (requires reference genome – refer to the group assignments section above for this info).
5. DESeq2 (requires reference genome – refer to the group assignments section above for this info).

Make sure to set the correct reference genomes for HISAT2, htseq-count, and DESeq2 steps. It is critical that you select the correct genome that matches the experimental organism for your samples:



Name you factor levels. This helps keep everything organized and name properly in the workflow. Each factor level is typically the name of the condition, like “mycelia” or “spore”.

a. Once you are sure everything is configured correctly, click on “Run Workflow” at the top.



The steps will start running in the history section on the right. Grey means they are waiting to start. Yellow means they are running. Green means they have completed. Red means there was an error in the step.

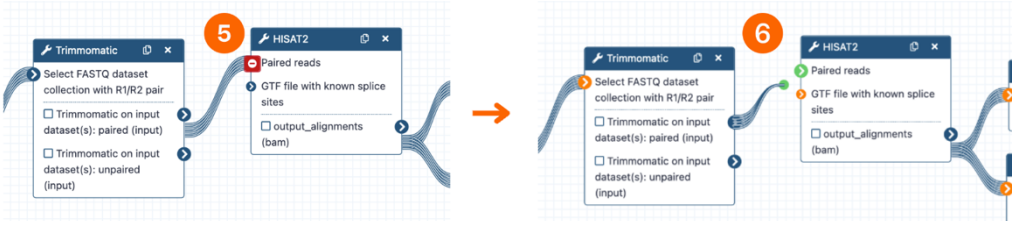
How to work with Galaxy editor (optional)

You can create your own workflows. The tools can all be added and configured in an interactive workflow editor.

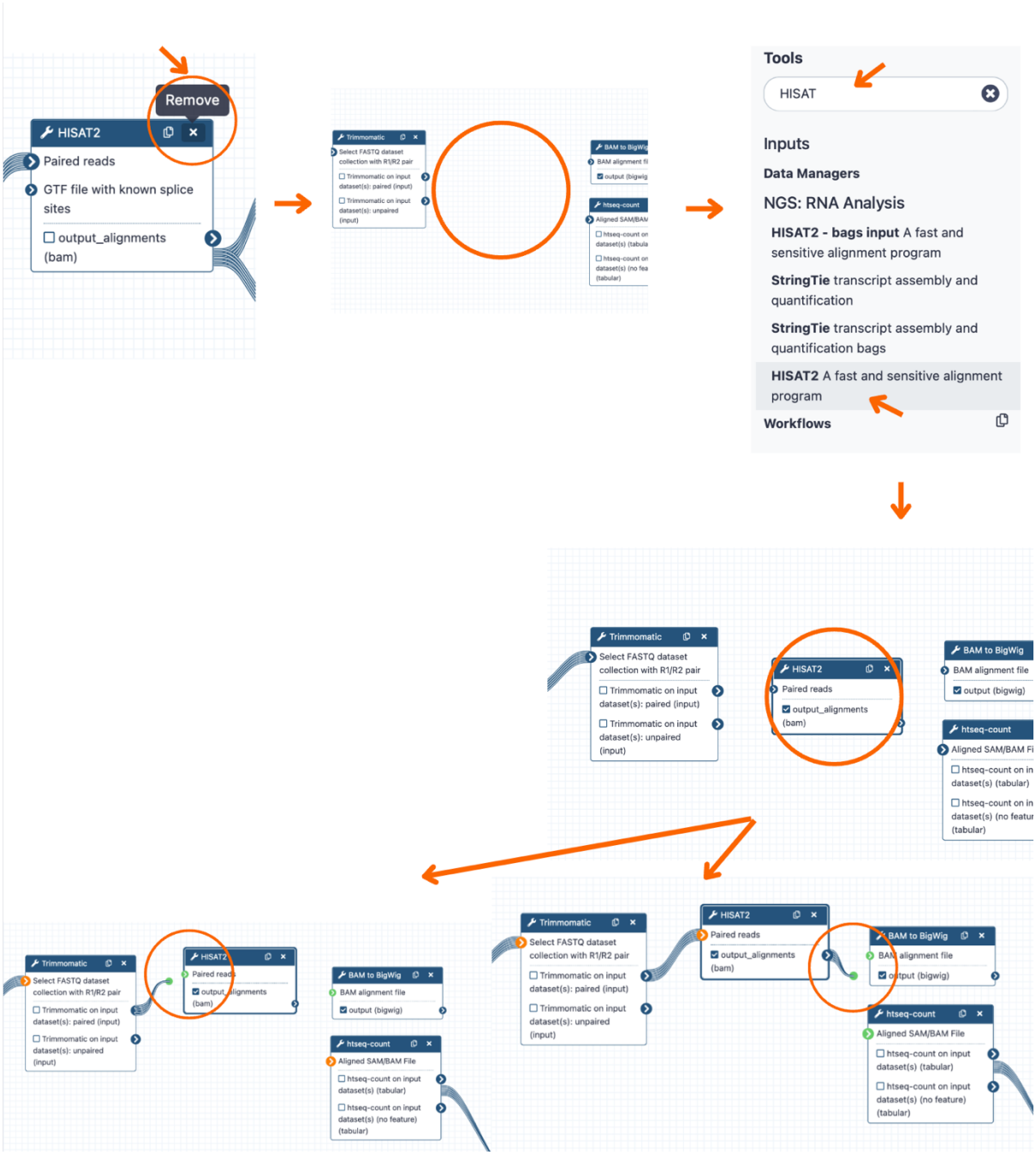
1. Navigate to the “Shared Data” menu.
2. Click on the “Workflows”.
3. Left-click on the “FPG2023 workflow editor practice” work to “import”
4. Once the workflow is important into your workspace, left-click and select “edit.

Once you are in the workflow editor:

5. Delete the Trimmomatic - HISAT2 connection.
6. Re-establish the connection by linking the “Trimmomatic on input dataset(s): paired (input) step to the “Paired reads” option in the HISTAS2.



7. Delete HISAT2 step completely by clicking on the “x” in the top right corner and use the tools menu on the left to insert it back.



Note: Sometimes you may be unable to re-establish connection. When this happens, take a look at the tool documentation notes in the right panel, check you r selection for single-read or paired-end setting in particular (paired-end setting must be selected if you are dealing with reverse and forward reads).

Now that you have learned the principals of workflow editing, you can either practice saving the workflow by clicking on the wheel at the far top corner or simply existing the workflow editor without saving.

What is Galaxy?

Galaxy is an open, web-based platform for data analysis under the FAIR principles of data sharing and re-use. Galaxy is an open-source platform that allows you to perform, reproduce, and share complete analyses without the use of command line scripting. The VEuPathDB project developed its own Galaxy instance in collaboration with Globus.

The VEuPathDB Galaxy offers pre-loaded genomes, pre-configured workflows and other tools for private data analysis and display. A custom-built set of tools also allows the ability to export Galaxy results into private workspaces within VEuPathDB sites (My Workspace > My data sets section). The datasets within the "My data sets" workspace can be explored using the FungiDB interface and tools and cross-referenced with the public data integrated in FungiDB.

VEuPathDB Galaxy access requires an account with FungiDB/VEuPathDB. The account is free and can be used to sign-in into any VEuPathDB genomics site.

The Galaxy instance is not meant for long term data storage. Datasets are automatically deleted after 60 days. To save your data, download your analysis results locally and then *delete and purge* files to free up space for your next analysis.

The Galaxy project offers extensive learning materials that can be accessed here: https://wiki.galaxyproject.org/Learn#Galaxy_101

Important: The Galaxy module consists of RNA-Seq and SNP analysis modules. These are concurrent sessions. This exercise will be carried out in groups of 4 people using the workshop Galaxy instance. Please do not use live FungiDB.org for this exercise. The detailed tutorials for both modules are available to all course participants.

Variant Calling analysis, Part I.

Learning objectives:

- Become familiar with the VEuPathDB Galaxy workspace.
- Upload raw data into Galaxy workspace and run a pre-configured SNP workflow

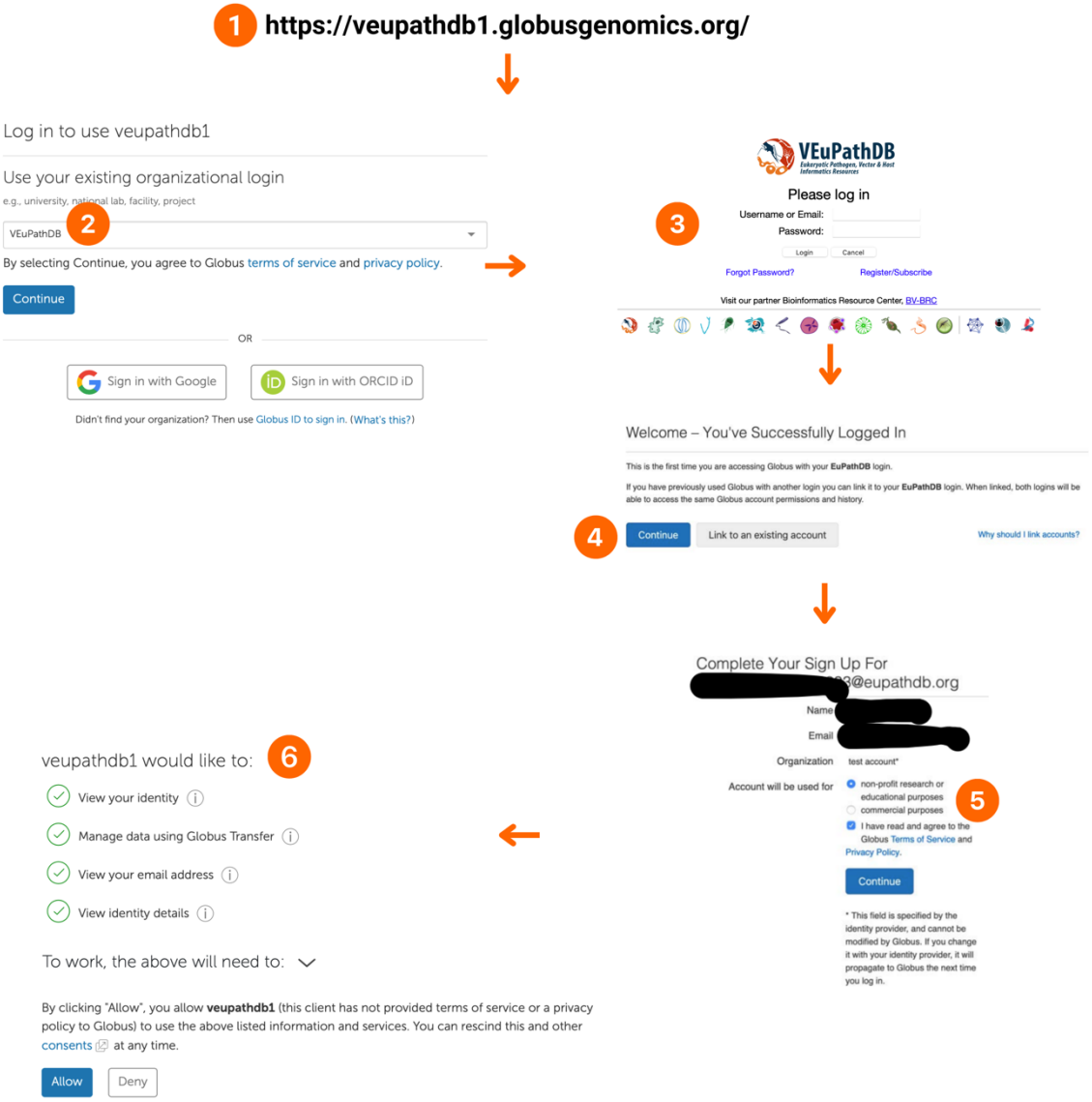
For this exercise, we will retrieve raw sequence files from the "shared history" section in VEuPathDB Galaxy and then run files through a pre-configured RNA-Seq workflow that will align the data to a reference genome, calculate expression values and determine differential expression.

Important: We will be working in groups of four people but only one person in each group should download data and deploy the pre-configured workflow. The other members' roles are to ensure that the correct datasets are used and that the correct workflow parameters are selected. In the Part 2 of this exercise, everyone will get a copy of the workflow output and practice how to perform data analysis.

- **Access the VEuPathDB Galaxy workshop instance.**

If you do not have an account with VEuPathDB/FungiDB, please create one now.

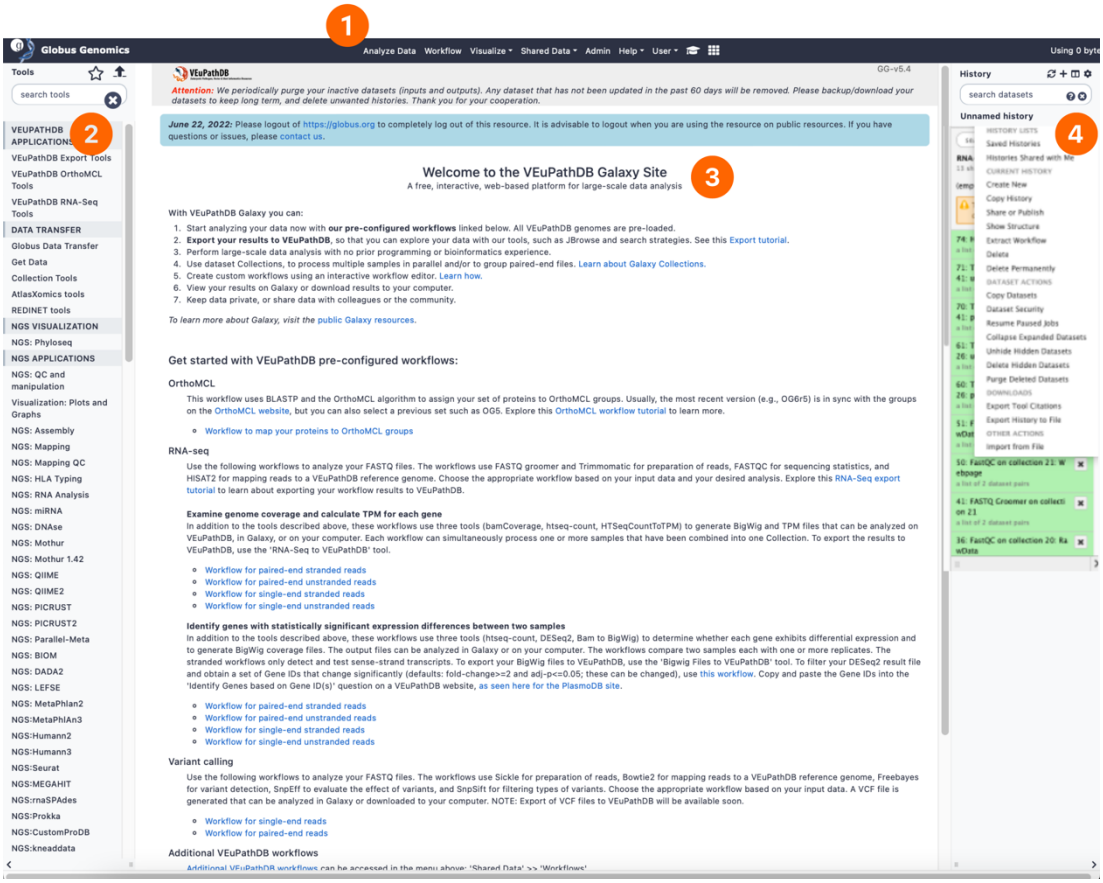
1. Click on the following URL to begin: https://veupathdb1.globusgenomics.org/
2. On the next page, you will be asked to define your organization. Choose the "VEuPathDB" option and click on the "Continue" button.
3. If you are not already logged into VEuPathDB, you will be prompted to do so.
4. Click on "Continue" on the next page (no need to link an existing account).
5. Select the "non-profit" option and agree to the Terms of Service. Click continue.
6. The next page will ask for permissions required to use this Galaxy instance. Click on "Allow"



The anatomy of the VEuPathDB Galaxy landing page.

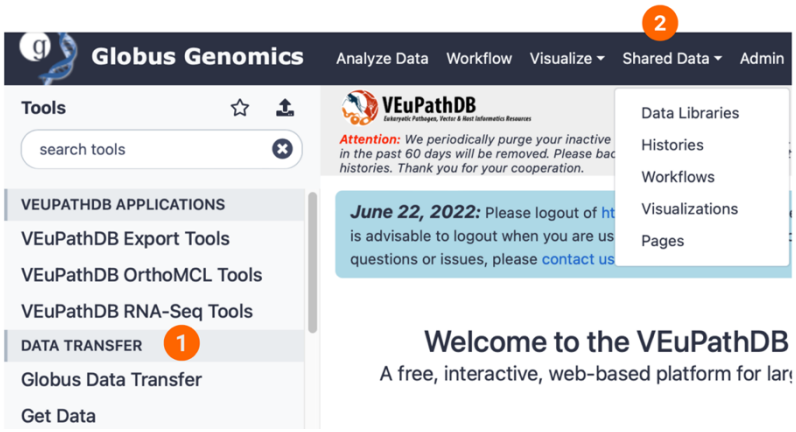
The workspace has four major components:

1. The top menu controls the main interface, provides access to the landing page, shared data, public and private workflows & more.
2. The left panel has a list of available tools where the VEuPathDB export tools are listed at the top.
3. The main welcome (landing) page is the interactive interface that houses pre-configured workflows, workflows editor, etc.
4. The panel on the right provides access to histories, deleted datasets, and other useful functions, including options to delete and purge datasets.



Don't see a tool you need for your research? – Let us know by sending an email to help@fungidb.org

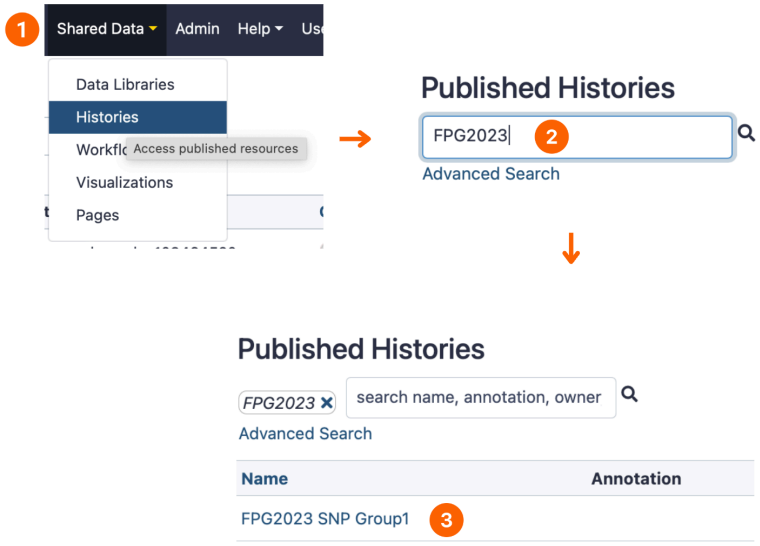
Importing data for your workflow.
There are multiple ways to important data into your Galaxy workspace. You can transfer data via tools located under the "Data Transfer" section in menu on the left (1). You can also transfer data from the "Shared Data" section in the main menu (2). The latter prides access to pre-loaded raw data, publicly shared workflows, or workflow results (histories), etc.



For this exercise, pre-loaded raw files should be imported from the "Shared Data" > Histories.
Only one person per each group should import data files and deploy an SNP workflow. Everyone will practice data analysis in NGS Part 2 module. For group assignments, see below.

- **Import data for your SNP workflow via the Shared histories option.**

1. From the top menu, select "Shared Data > Histories" option.
2. Filter all public workflows on "FPG2023" .
3. Click on the history link that correspond to your group number (e.g., FPG2023 SNP Group1) to import the data into your Galaxy workspace.



Group assignments

Groups 1 *Aspergillus fumigatus*. Paired-end data. A clinical isolate from pleural fluid of a patient. Isolate: AFIS2503.

History name for download (in Galaxy)	FPG2023 SNP Group1
Ref genome (in Galaxy)	FungiDB-29_AfumigatusAf293_Genome

Groups 2 *Aspergillus fumigatus*. Paired-end data. A clinical isolate from pleural fluid of a patient. Isolate: AFIS1415.

History name for download (in Galaxy)	FPG2023 SNP Group2
Ref genome (in Galaxy)	FungiDB-29_AfumigatusAf293_Genome

Group 3 *Zymoseptoria tritici*. Paired-end data. An isolate collected from common wheat (Triticum aestivum) in Switzerland: Eschikon. Isolate: ST16CH_1A27.

History name for download (in Galaxy)	FPG2023 SNP Group3
Ref genome (in Galaxy)	FungiDB-34_ZtriticilPO323_Genome

Group 4 *Zymoseptoria tritici*. Paired-end data. An isolate collected from common wheat (Triticum aestivum) in Oregon: USA. Isolate: ORE15_Mad_G1.

History name for download (in Galaxy)	FPG2023 SNP Group4
Ref genome (in Galaxy)	FungiDB-34_ZtriticilPO323_Genome

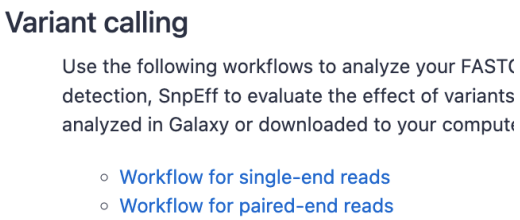
Group 5 *Candida auris*. Paired-end data. An isolated collected from an apple surface in India. Isolate: VPCI-F37-B-2021.

History name for download (in Galaxy)	FPG2023 SNP Group5
Ref genome (in Galaxy)	FungiDB-37_CaurisB8441_Genome

Group 6 *Candida auris*. Paired-end data. An isolated collected from an apple surface in India. Isolate: VPCI-F1-A-2020.

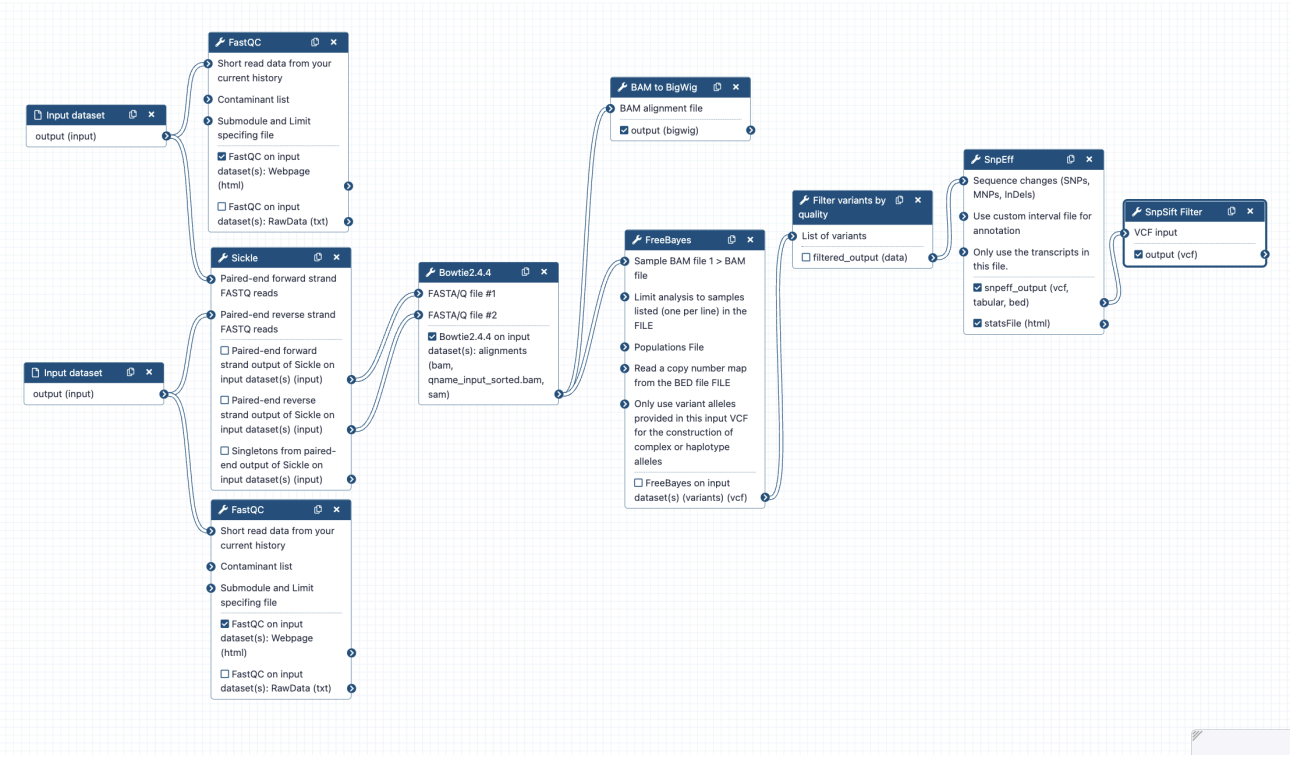
History name for download (in Galaxy)	FPG2023 SNP Group6
Ref genome (in Galaxy)	FungiDB-37_CaurisB8441_Genome

Once the data files have been transferred into your galaxy history you need to choose a workflow appropriate for your data (paired or single -read).



The pre-configured workflows follow these steps:

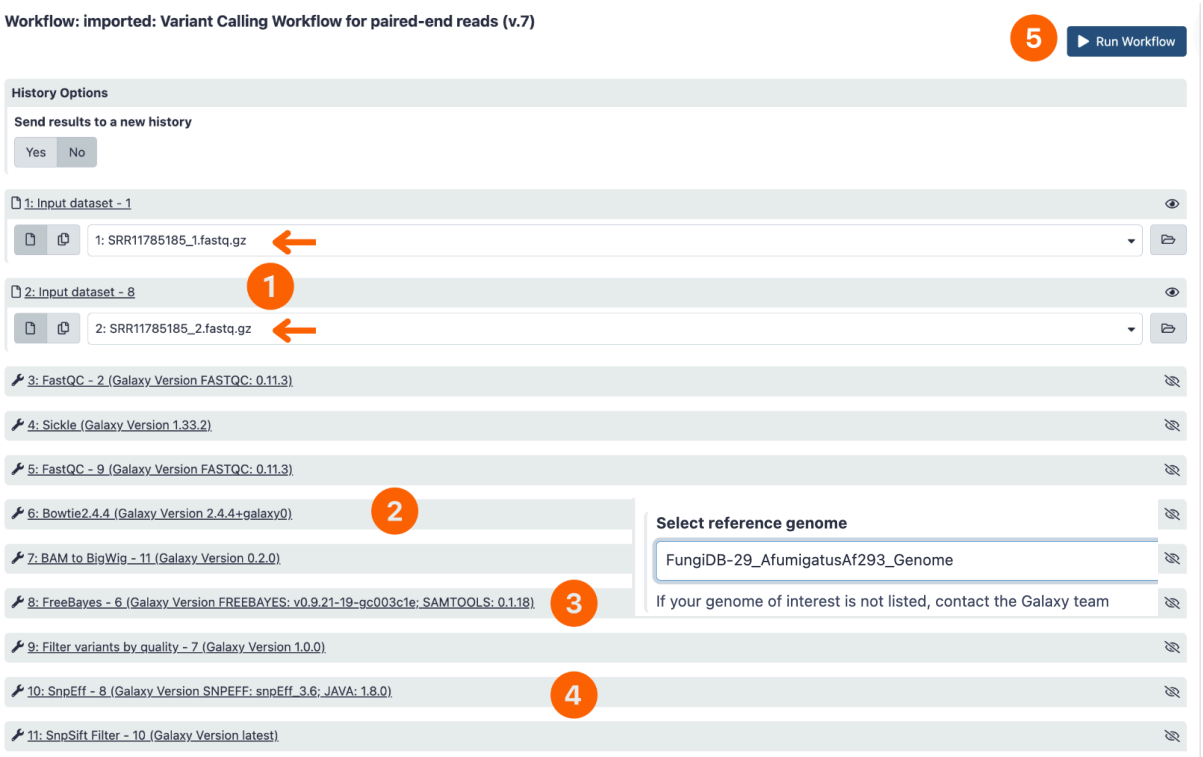
- Determine quality of the reads in your files and generates FASTQC reports.
- Trim reads based on their quality scores.
- Align reads to a reference genome using Bowtie2 and generating coverage plots.
- Sort alignments with respect to their chromosomal positions.
- Detect variants using FreeBayes.
- Filter SNP candidates.
- Analyze and annotate of variants, and calculation of the effects via SnpEff.



- **Set workflow parameters.**
    1. For paired-end data, make sure that the input steps are set to the *xxxx_1.fastq.gz* and *xxxx_2.fastq.gz* as by default both have the same one selected. Here is an example (disregards "7" and "8" and it simply refers to the ordered file number).

Note: for single read data, you will have only one file.

2. Select the correct reference genome for Bowtie2 (see group assignment above).
3. Select the correct reference genome for FreeBayes (see group assignment above).
4. Select the correct reference genome for SnpEff (see group assignment above).
5. Click Run Workflow.

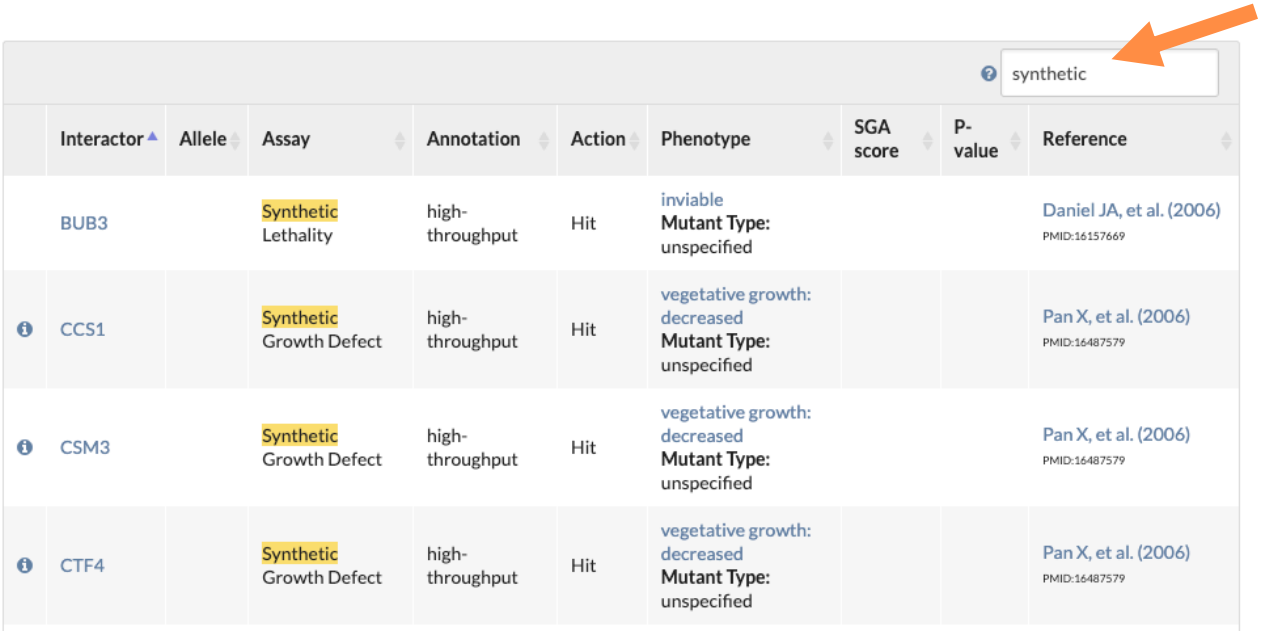


Using SGD GO Slim Mapper and Interaction Data to Predict Gene Function

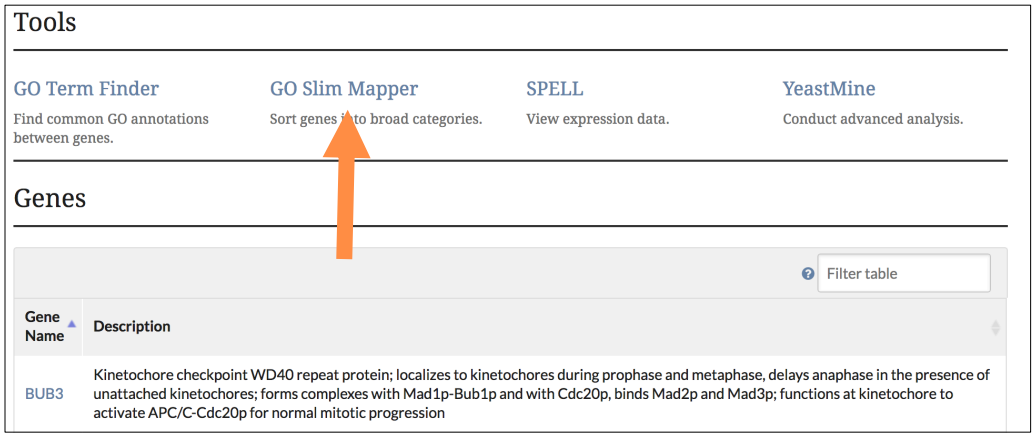
The Gene Ontology (GO) is structured in a hierarchy, such that granular terms ("perinuclear space") are connected and further down the hierarchy than their related broader terms ("nucleus"). However, for many purposes, such as reporting the upregulated cellular functions of a transcriptomics experiment,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

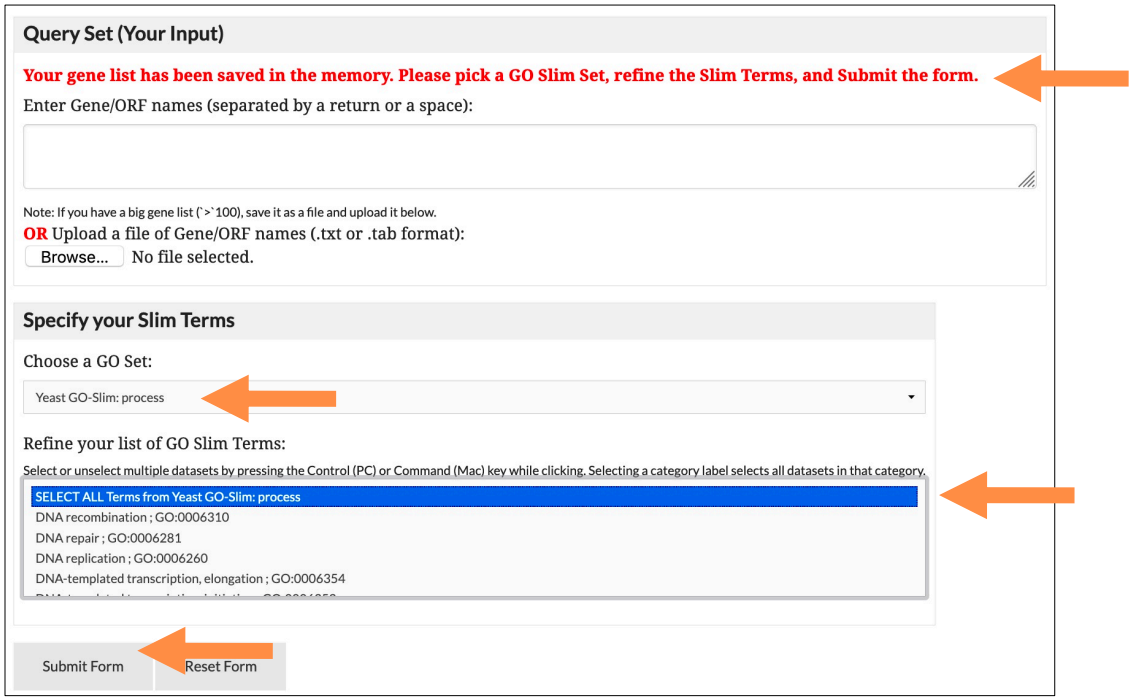
- From the SGD home page (www.yeastgenome.org), go to the Locus Summary page for the uncharacterized gene **YLR287C.**
- Select **Genetic Interactions** tab. Here, we are interested in finding genes that have a genetic interaction with YLR287C, as the function of these genes may provide hints about the function of YLR287C.
- Search for "synthetic" in 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**.

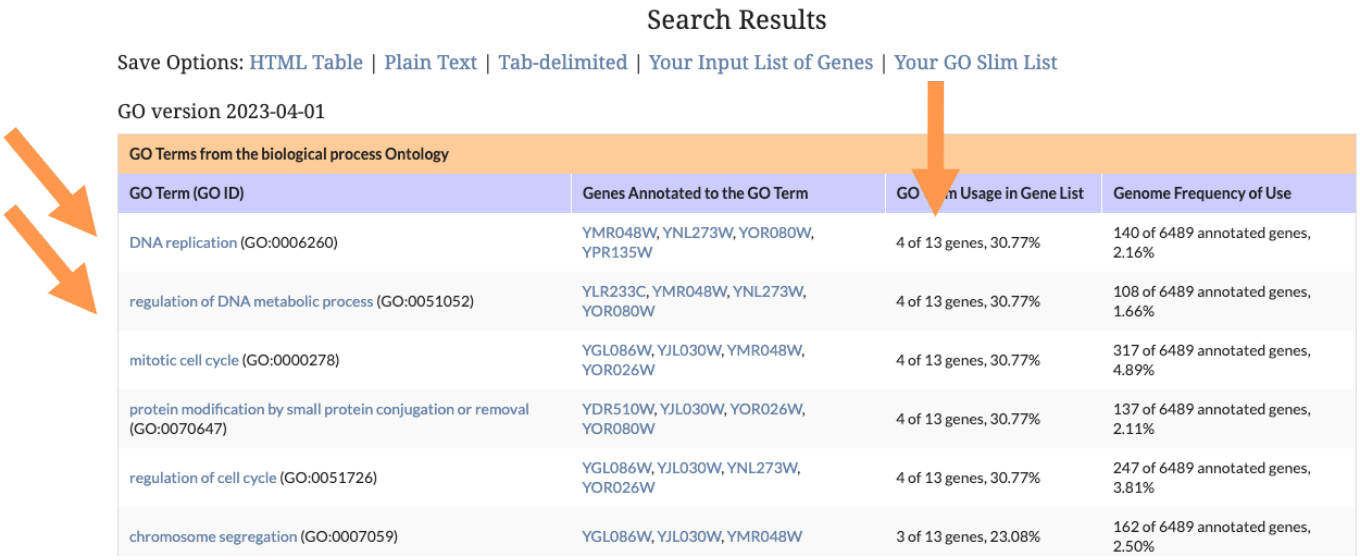


- The GO Slim Mapper has three steps (plus one optional step) in which you can specify your query. The Query Set (Your Input) box has been preloaded in memory with the list of genes you imported from the table.



- 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 GO Slim List

GO version 2023-04-01

GO Terms from the biological process Ontology			
GO Term (GO ID)	Genes Annotated to the GO Term	GO n 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?

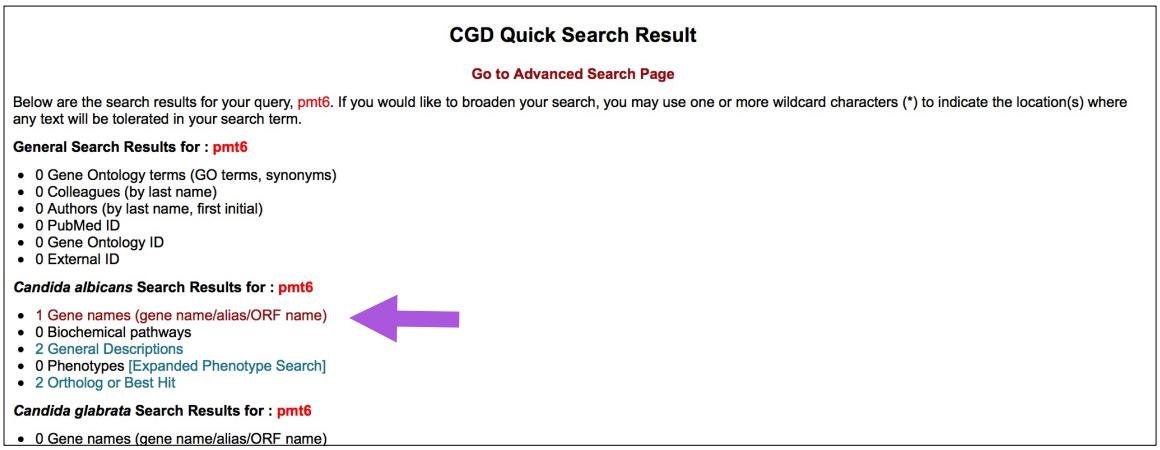
GO Enrichment, Phenotype Data at CGD

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.

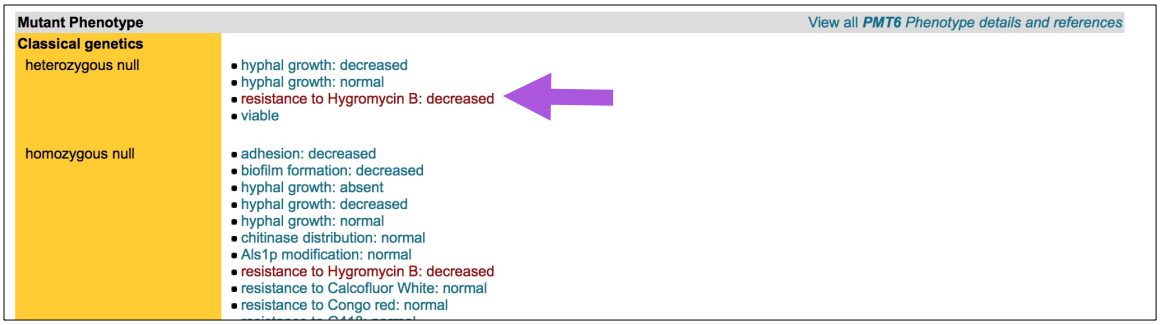
GO Term Finder takes a list of genes and identifies what GO terms are significant for the list. It is a powerful way to interpret the results of omics experiments or any situation where determining common functions and roles are important. For example, GO Term Finder can take a list of upregulated genes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

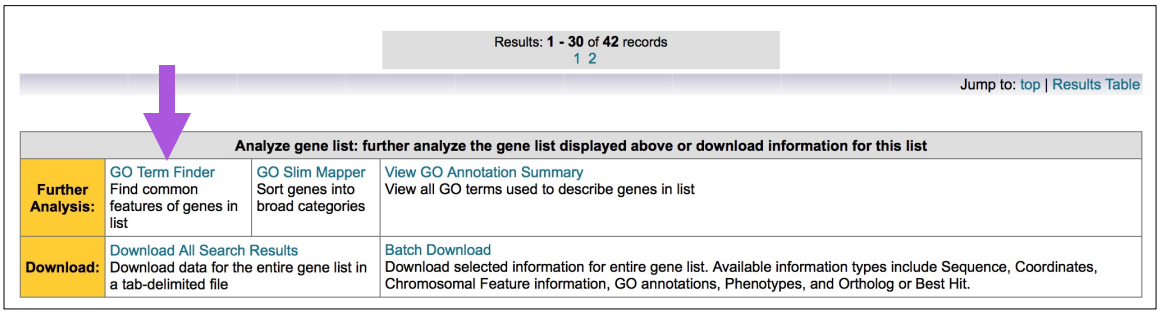
- From the CGD home page (www.candidagenome.org), go to the Locus Summary page for the hygromycin B-sensitivity gene PMT6. Enter **PMT6** into the **search our site** box and click **GO**. On the next page, under ***Candida albicans* Search Results**, click on hyperlinked **1 Gene names (gene name/alias/ORF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

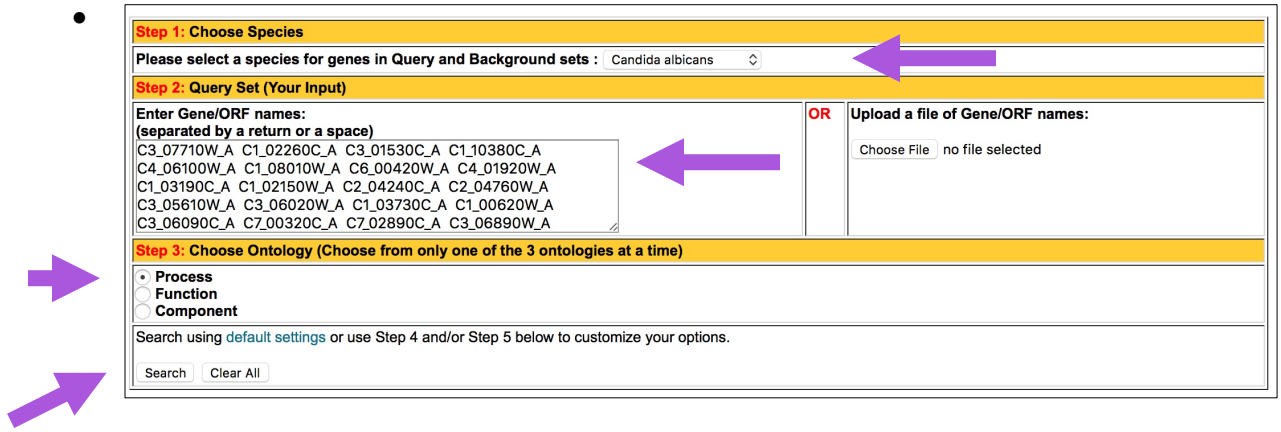


- On the **Phenotype Search Results** page, click on **Jump to: Analyze Gene List** above the table on the right (or simply scroll down to the bottom of the page). Click on **GO Term Finder** link.



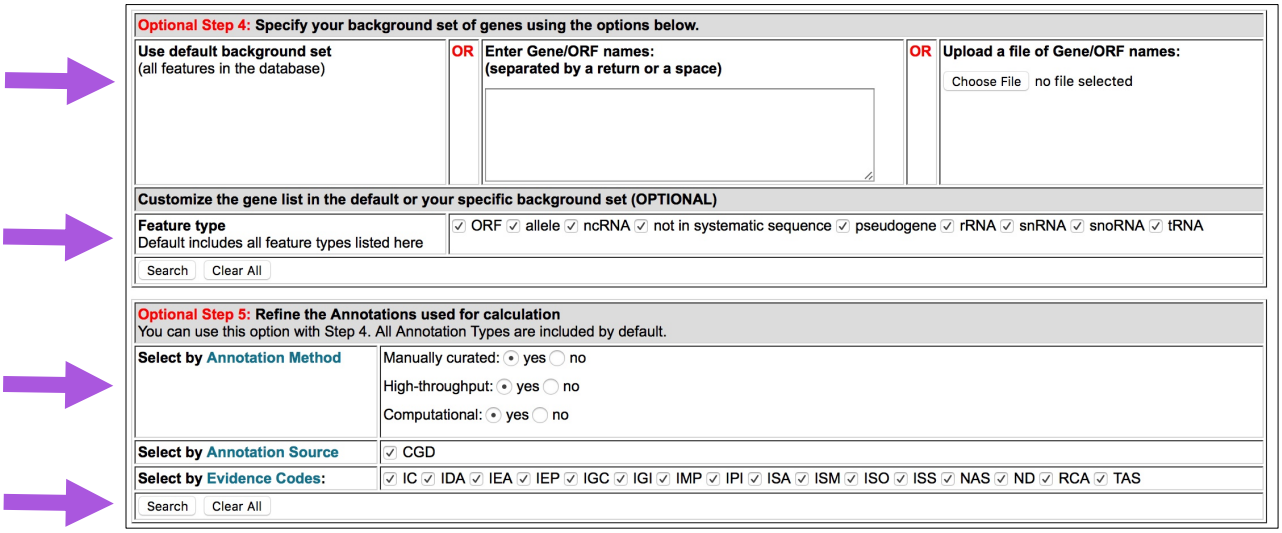
  - With your own list of genes, you can access GO Term Finder from any CGD page by opening **GO** menu in the banner on top and clicking on **GO Term Finder**. Or you use this URL: http://www.candidagenome.org/cgi-bin/GO/goTermFinder

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

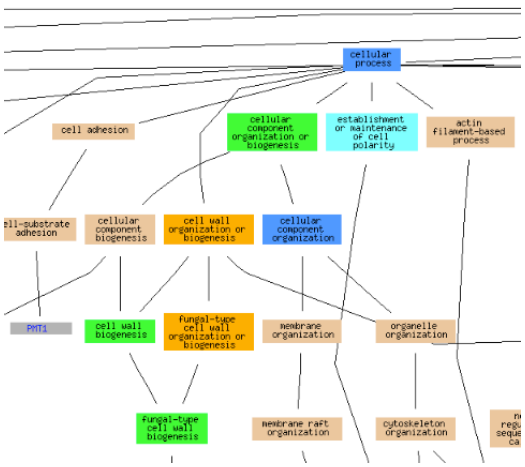


You can further customize your query in the next steps down the page:

- 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

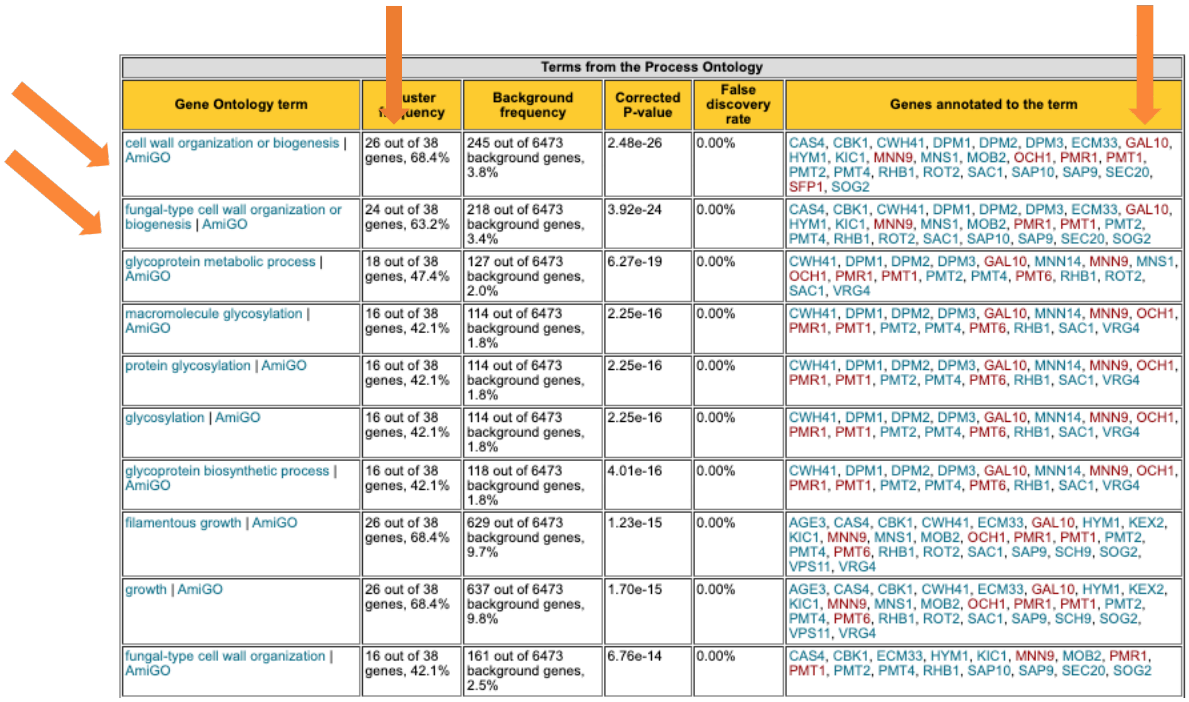


- Click **Search**. The input is checked and any genes that are not recognized as valid for the selected *Candida* species are rejected; click on **Proceed** in the following window.
- The results page displays the significant shared GO terms (or their parents) in both graphic and table form, within 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). 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	26 out of 38 genes, 68.4%	245 out of 6473 background genes, 3.8%	2.48e-26	0.00%	CAS4, CBK1, CWH41, DPM1, DPM2, DPM3, ECM33, GAL10, HYM1, KIC1, MNN9, MNS1, MOB2, OCH1, PMR1, PMT1, PMT2, PMT4, RHB1, ROT2, SAC1, SAP10, SAP9, SEC20, SFP1, SOG2
fungal-type cell wall organization or biogenesis | AmiGO	24 out of 38 genes, 63.2%	218 out of 6473 background genes, 3.4%	3.92e-24	0.00%	CAS4, CBK1, CWH41, DPM1, DPM2, DPM3, ECM33, GAL10, HYM1, KIC1, MNN9, MNS1, MOB2, PMR1, PMT1, PMT2, PMT4, RHB1, ROT2, SAC1, SAP10, SAP9, SEC20, SOG2
glycoprotein metabolic process | AmiGO	18 out of 38 genes, 47.4%	127 out of 6473 background genes, 2.0%	6.27e-19	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 38 genes, 42.1%	114 out of 6473 background genes, 1.8%	2.25e-16	0.00%	CWH41, DPM1, DPM2, DPM3, GAL10, MNN14, MNN9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, SAC1, VRG4
protein glycosylation | AmiGO	16 out of 38 genes, 42.1%	114 out of 6473 background genes, 1.8%	2.25e-16	0.00%	CWH41, DPM1, DPM2, DPM3, GAL10, MNN14, MNN9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, SAC1, VRG4
glycosylation | AmiGO	16 out of 38 genes, 42.1%	114 out of 6473 background genes, 1.8%	2.25e-16	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 38 genes, 42.1%	118 out of 6473 background genes, 1.8%	4.01e-16	0.00%	CWH41, DPM1, DPM2, DPM3, GAL10, MNN14, MNN9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, SAC1, VRG4
filamentous growth | AmiGO	26 out of 38 genes, 68.4%	629 out of 6473 background genes, 9.7%	1.23e-15	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 38 genes, 68.4%	637 out of 6473 background genes, 9.8%	1.70e-15	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
fungal-type cell wall organization | AmiGO	16 out of 38 genes, 42.1%	161 out of 6473 background genes, 2.5%	6.76e-14	0.00%	CAS4, CBK1, ECM33, HYM1, KIC1, MNN9, MOB2, PMR1, PMT1, PMT2, PMT4, RHB1, SAP10, SAP9, SEC20, SOG2

- 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

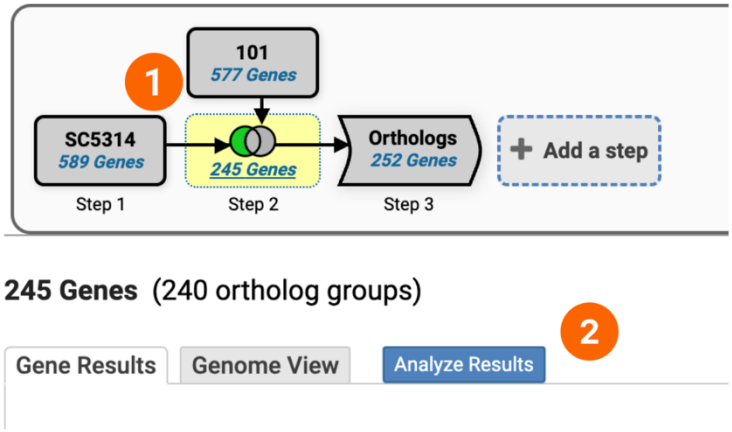
Learning objectives:

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

- **Use a previously created search strategy to perform Gene Ontology enrichment analysis on genes upregulated (identified by RNA-Seq) in *C. albicans* SC5314 only.**

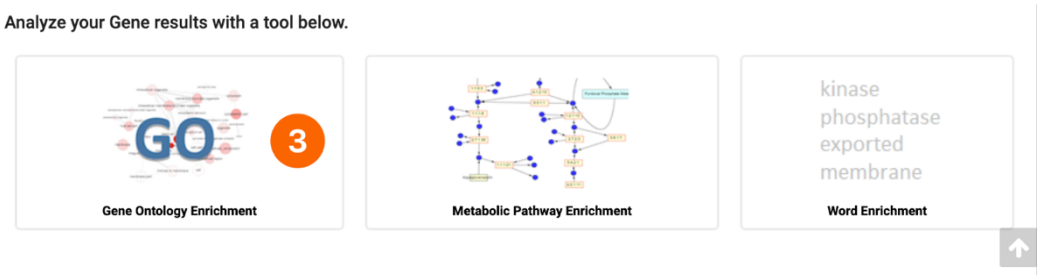
  Strategy URL:
  https://fungidb.org/fungidb/app/workspace/strategies/import/802d9f2b606fc1fa

1. Click on the Step 2 to identify upregulated gene in *C. albicans* SC5314 only.
2. Click on the "Analyze Results" tab to bring up enrichment analysis options.

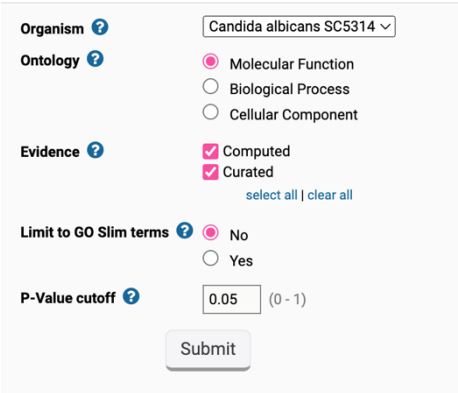


The enrichment analysis tools can be accessed under the blue Analyze Results tab and it includes Gene Ontology, Metabolic Pathway, and Word Enrichment tools. The three types of analysis apply Fisher's Exact test to evaluate ontology terms, over-represented pathways, and product description terms. Enrichment is carried out using a Fisher's Exact test with the background defined as all genes from the organism being queried. P-values corrected for multiple testing are provided using both the Benjamini-Hochberg false discovery rate method and the Bonferroni method.

3. Deploy GO enrichment analysis by clicking on the "Gene Ontology Enrichment" button.

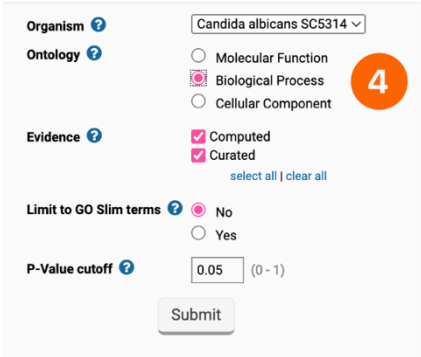


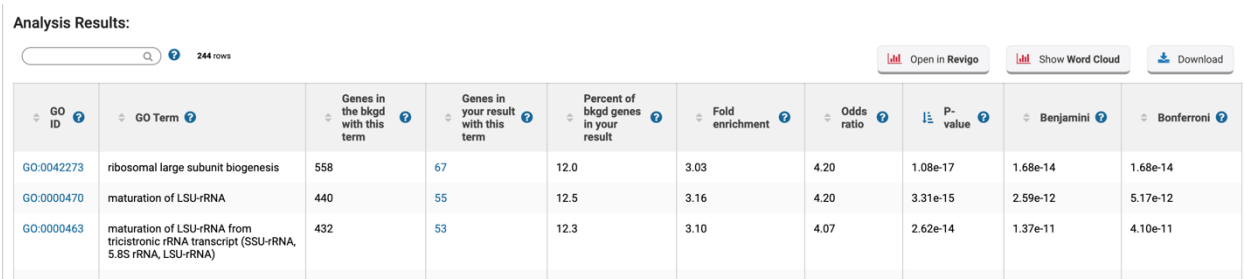
GO enrichment analysis can be performed on the following ontology groups: molecular function, cellular component, and biological processes. Also, other parameters allow users to limit their analysis on either “Curated” or “Computed” annotations, or both. Those with a GO evidence code inferred from electronic annotation (IEA) are denoted “Computed”, while all others have some degree of curation. The default P-value is set to 0.05 but can be adjusted manually.



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

4. Perform GO enrichment analysis (Biological Process) at default selection criteria.





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

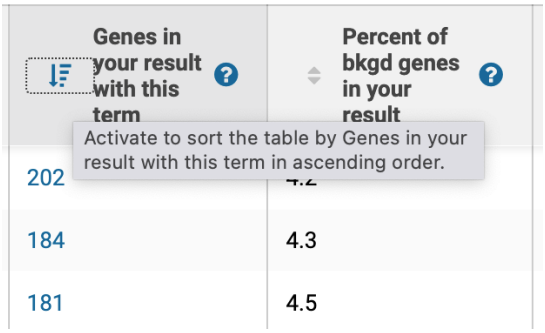
- **Benjamini-Hochburg false discovery rate** - A method for controlling false discovery rates for type 1 errors.
- **Bonferroni adjusted P-values** - A method for correcting significance based on multiple comparisons.

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?

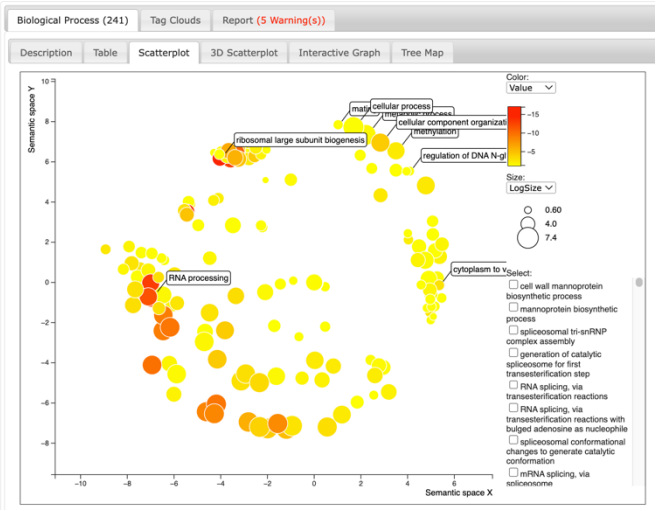
Note: you can sort genes in your results using the sort options within a column:



6. Visualize the results in Revigo by clicking on 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.

Bubble color corresponds to the user-provided p-value (see legend in upper right-hand corner)

Bubble size represent 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 (http://geneontology.org/docs/ontology-relations/)

Creating queries across FungiDB and SGD (optional exercise)

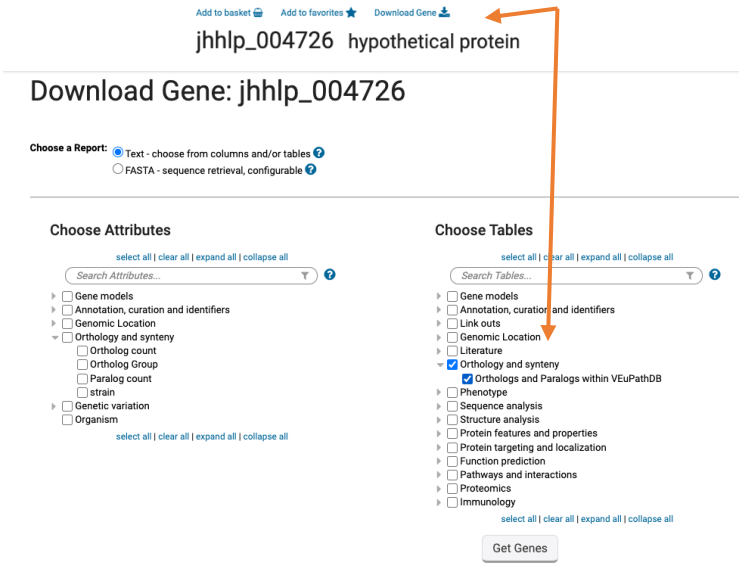
During a genetic screen in *Lomentospora prolificans*, you identified several interesting genes, including jhhlp_004726, which is a hypothetical protein. Take advantage of FungiDB and SGD records to learn more about this gene.

1. Navigate to jhhlp_004726 in FungiDB and examine available records.
https://fungidb.org/fungidb/app/record/gene/jhhlp_004726

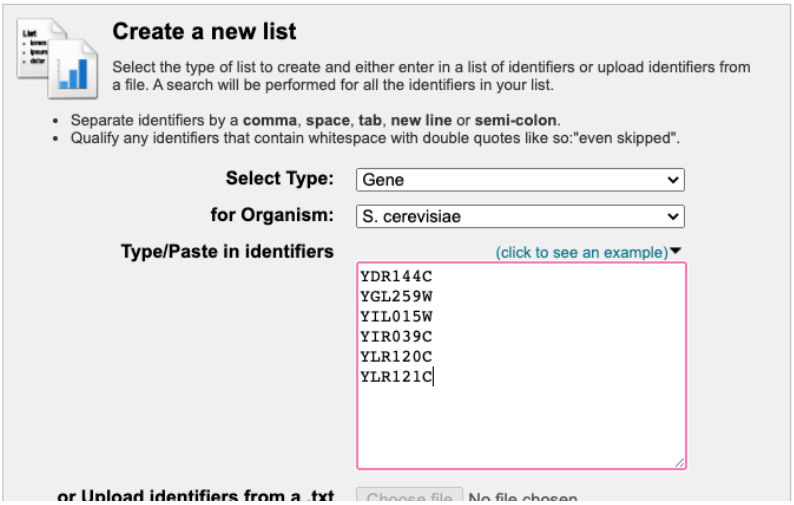
- Run an InterPro search and a GPI anchor prediction tool. What did you learn about this protein?

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

- Export orthologs of this gene.
Click on the Download gene link and select to export orthologs in VEuPathDB option



- Navigate to the SGD gene lists search and copy and paste *S. cerevisiae* orthologs for jhhlp_004726: https://www.yeastgenome.org/locus/YDR144C



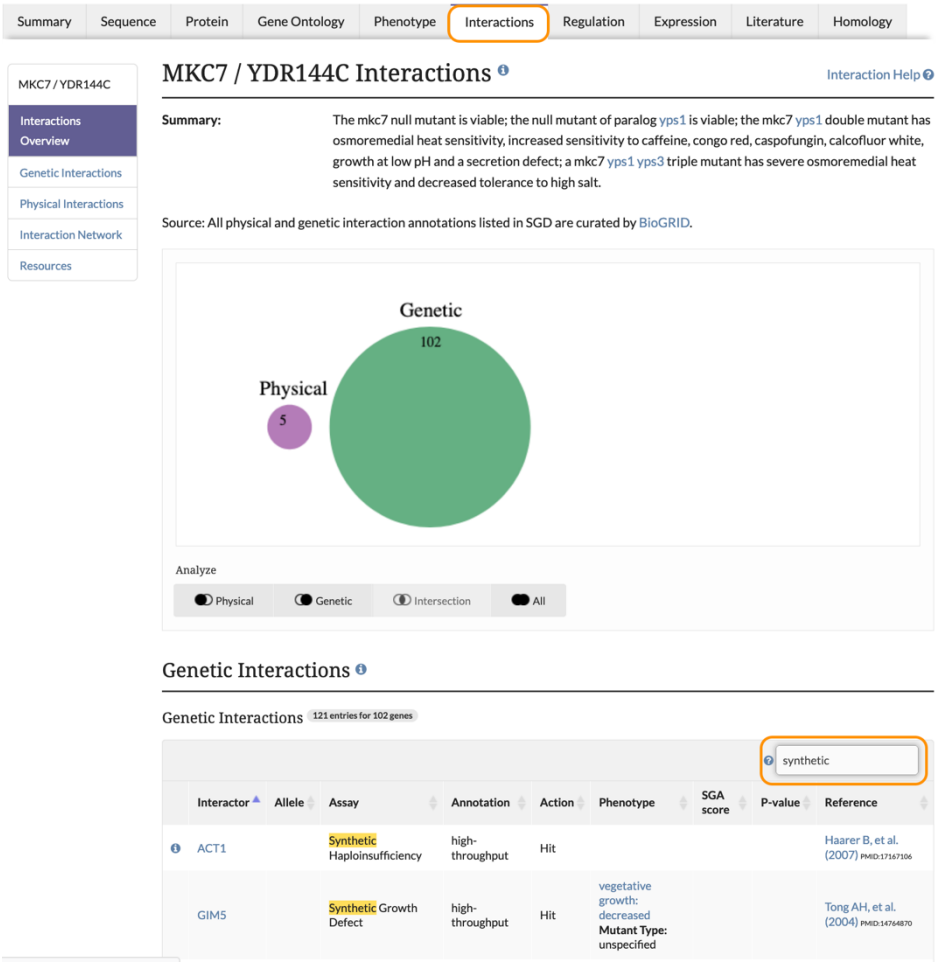
- Give your list a name such as 'Yeast orthologs 1'.
- Click on the GeneIDs to examine *S. cerevisiae* genes. What is the function of MKC7 (YDR144C) in *S. cerevisiae?* Does it encode a protein with enzymatic activity? Where in the cell does the protein execute its function? What biological process? Hint: see the **Gene Ontology** section on the locus page or click on the Gene Ontology tab at the top of the page.

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 a phenotype that is unexpected, given the phenotypes of the single mutants.

- **Find known genetic interactions for MKC7.**

    - In SGD, find the MKC7 locus page and navigate to the **Interactions** tab, which is listed in the Quick Links panel near the top. The interactions are divided into separate physical interactions and genetic interactions tables below the summary.

    - Filter the **Genetic Interactions** table on “synthetic”. This table will show only the genetic interactions where some sort of synthetic growth defect, haploinsufficiency, or lethality is produced.



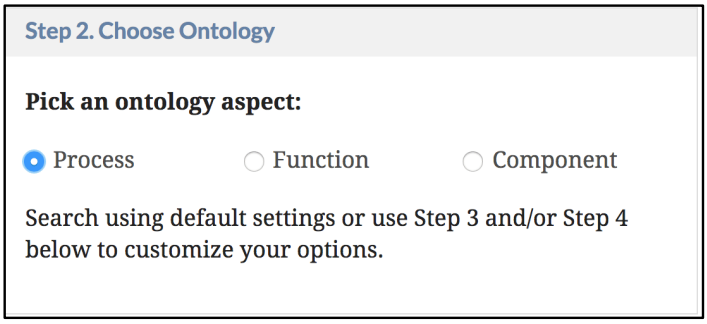
- Click on the **Download** button, which is located under the results table, and save this gene list. *Rename the file to* ***synthetic.txt***.

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

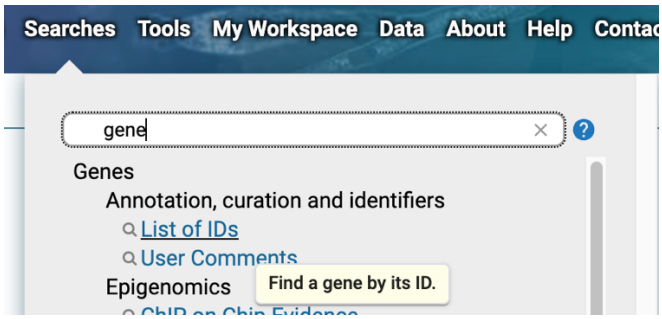
Now, let's go back to the file of MKC7 "synthetic" genetic interactors we downloaded earlier and find the orthologs of these genes in *Lomentospora prolificans*.

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

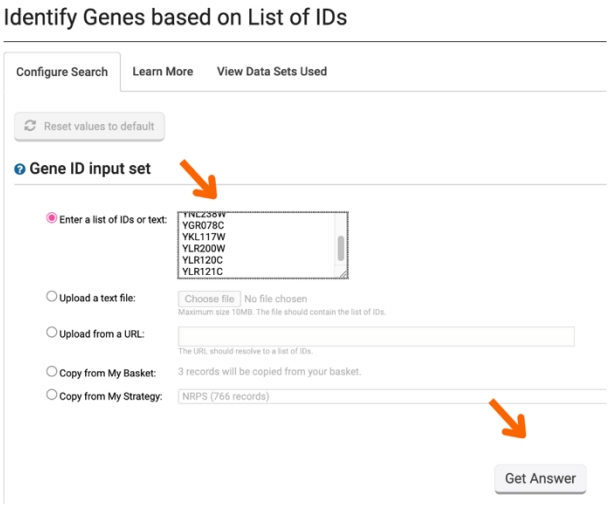
Interactor	Interactor Sy	Interactor	Interactor Systematic Name	Type	Assay	Annotation
MKC7	YDR144C	ACT1	YFL039C	Genetic	Synthetic Ha	high-through
MKC7	YDR144C	GIM5	YML094W	Genetic	Synthetic Gr	high-through
MKC7	YDR144C	IRE1	YHR079C	Genetic	Synthetic Gr	manually cur
MKC7	YDR144C	KEX2	YNL238W	Genetic	Synthetic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

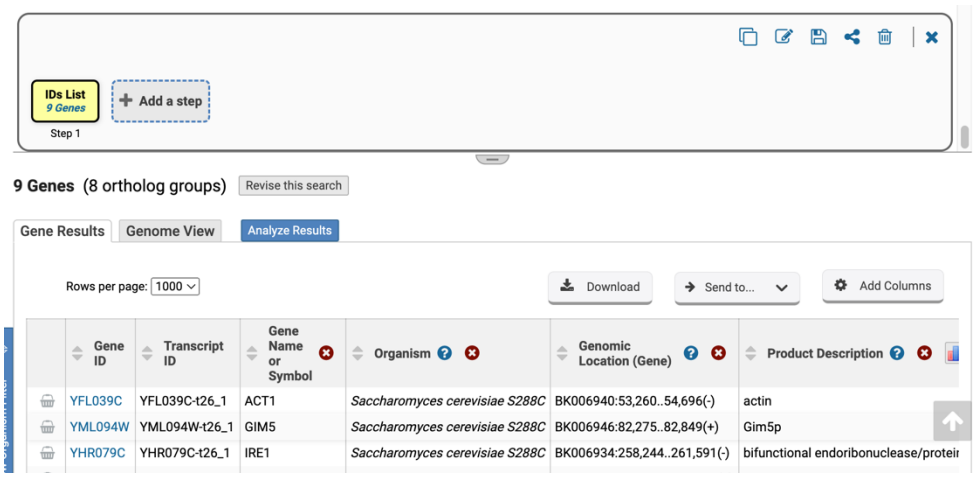
- Visit FungiDB again and initiate the List of IDs search query

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



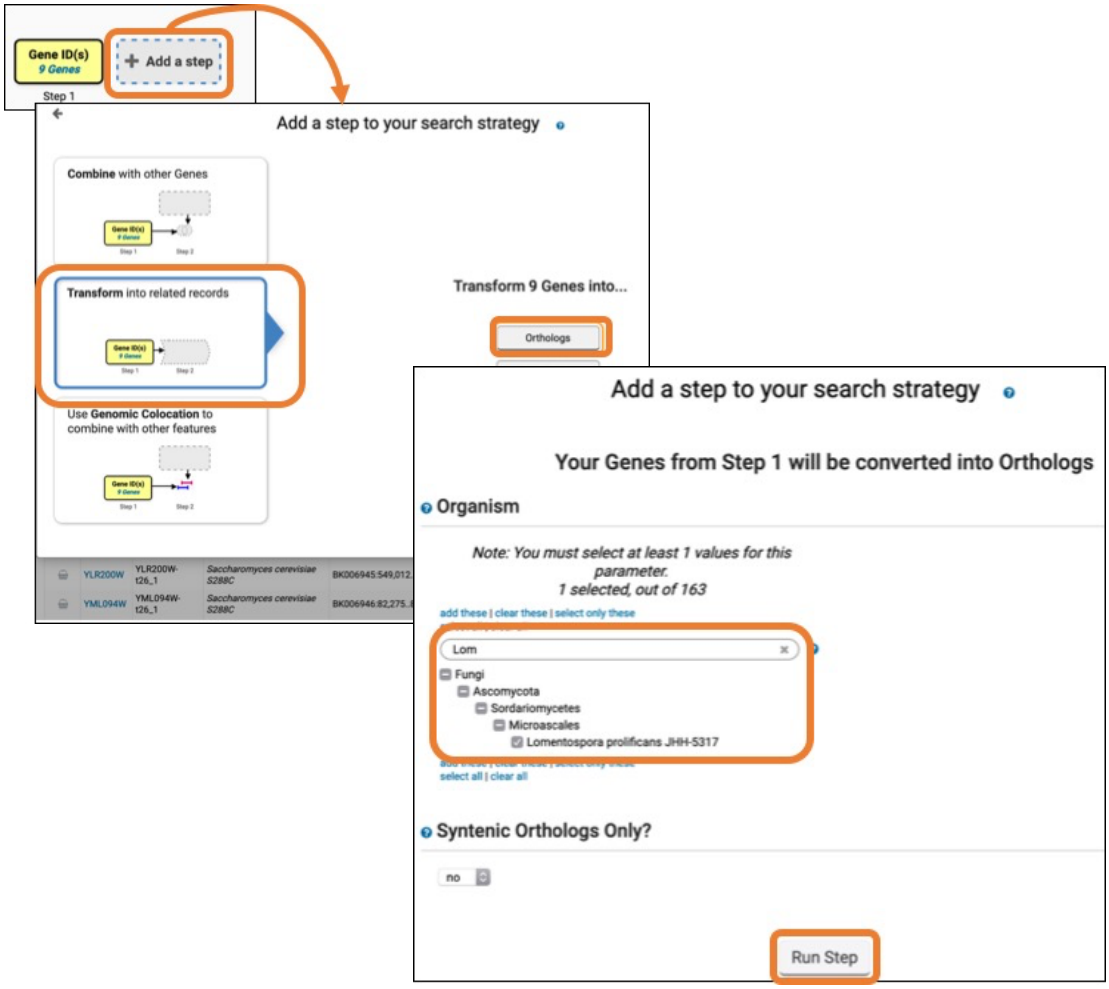
- Paste the list of Gene IDs that had the “synthetic” genetic interactions with MKC7 into FungiDB query and click on the **Get Answer** button.

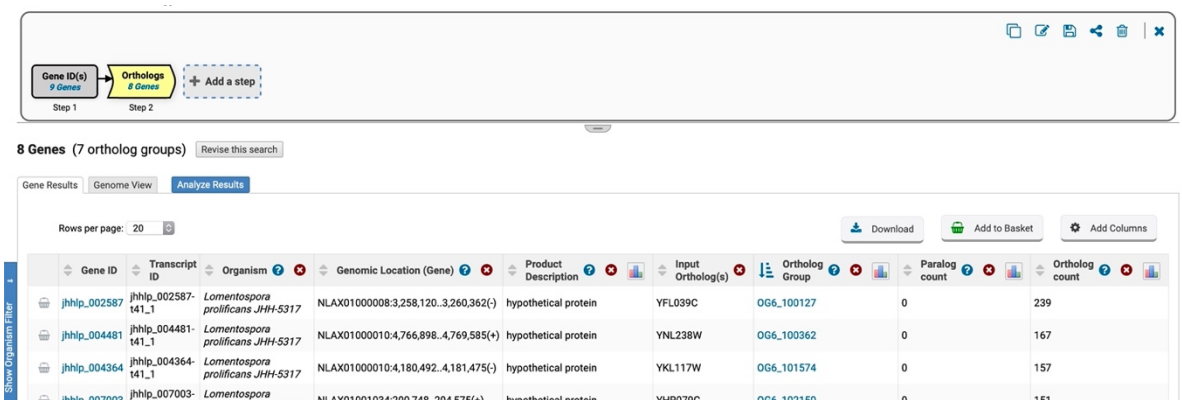




- Find orthologs in *Lomentospora prolificans*.

Click on 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 of the 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 defenses. 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 vary between *S. cerevisiae* and *L. prolificans*?

Additional resources:

More info on Fischer's exact test:
http://udel.edu/~mcdonald/statfishers.html

Some more info about Odds ratios:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/

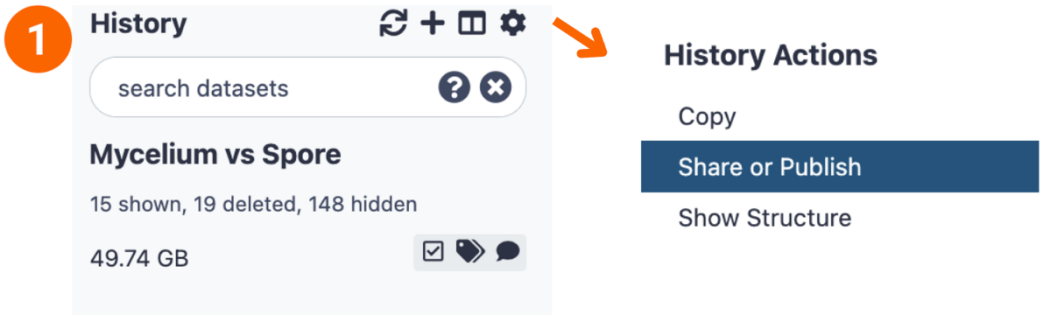
False discovery rates and P value correction:
http://brainder.org/2011/09/05/fdr-corrected-fdr-adjusted-p-values/

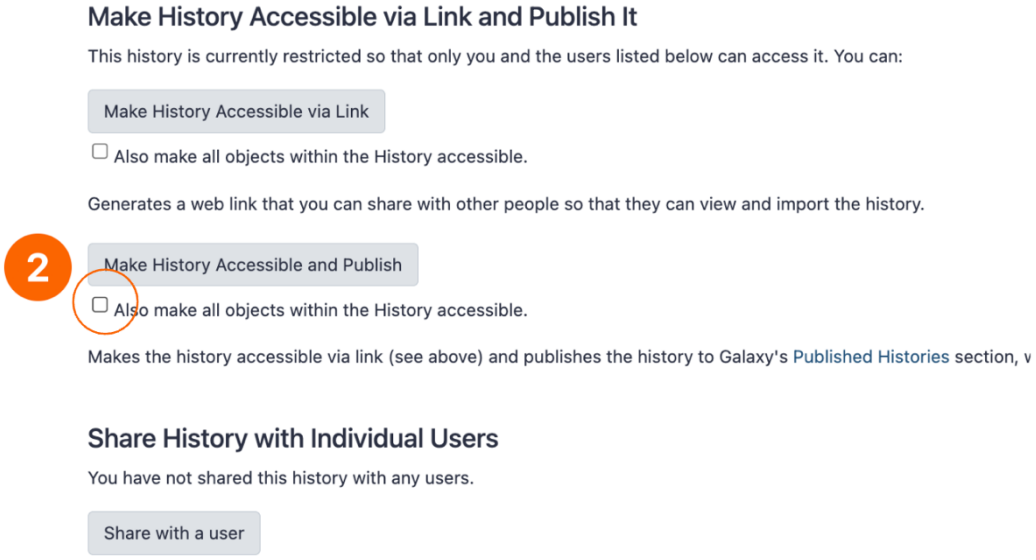
RNA sequence data analysis via Galaxy, Part 2

Learning objectives:

- Examine RNA-Seq analysis workflow and outputs.
- Import data from Galaxy to FungiDB My Workspace.
- Analyze the results using FungiDB interface and tools.

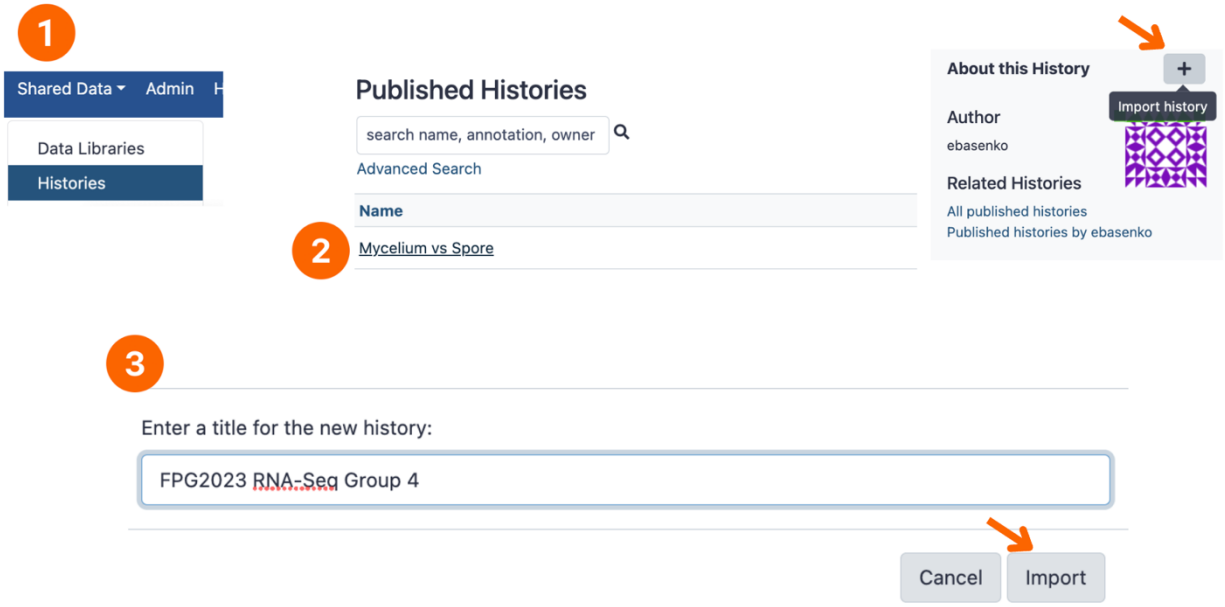
- **Sharing workflow histories with others.**
    1. Make sure your history has a useful name (e.g, Mycelium vs Spore, RNA Group3, etc.) and click on the history action menu icon.
    2. Select the "Make History Accessible and Publish" option and check to make sure that all objects within History are accessible.





- **Importing workflow histories and output files into your own Galaxy workspace.**

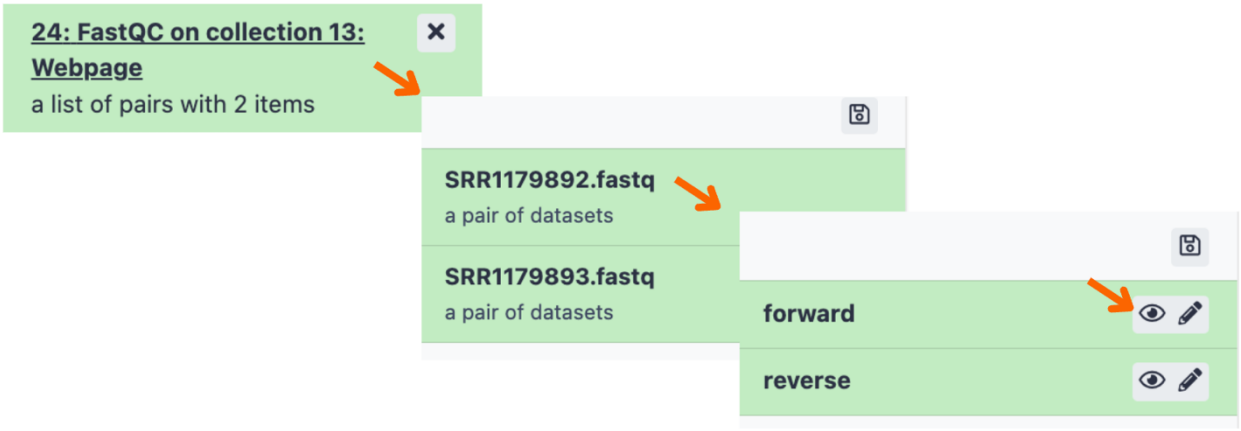
1. Click on “Shared Data” at the top and select “Histories”.
2. Click on the history shared by your colleague, click on the plus icon on the far right and choose to import the history.
3. You can give it a descriptive name if you prefer or leave it as is.

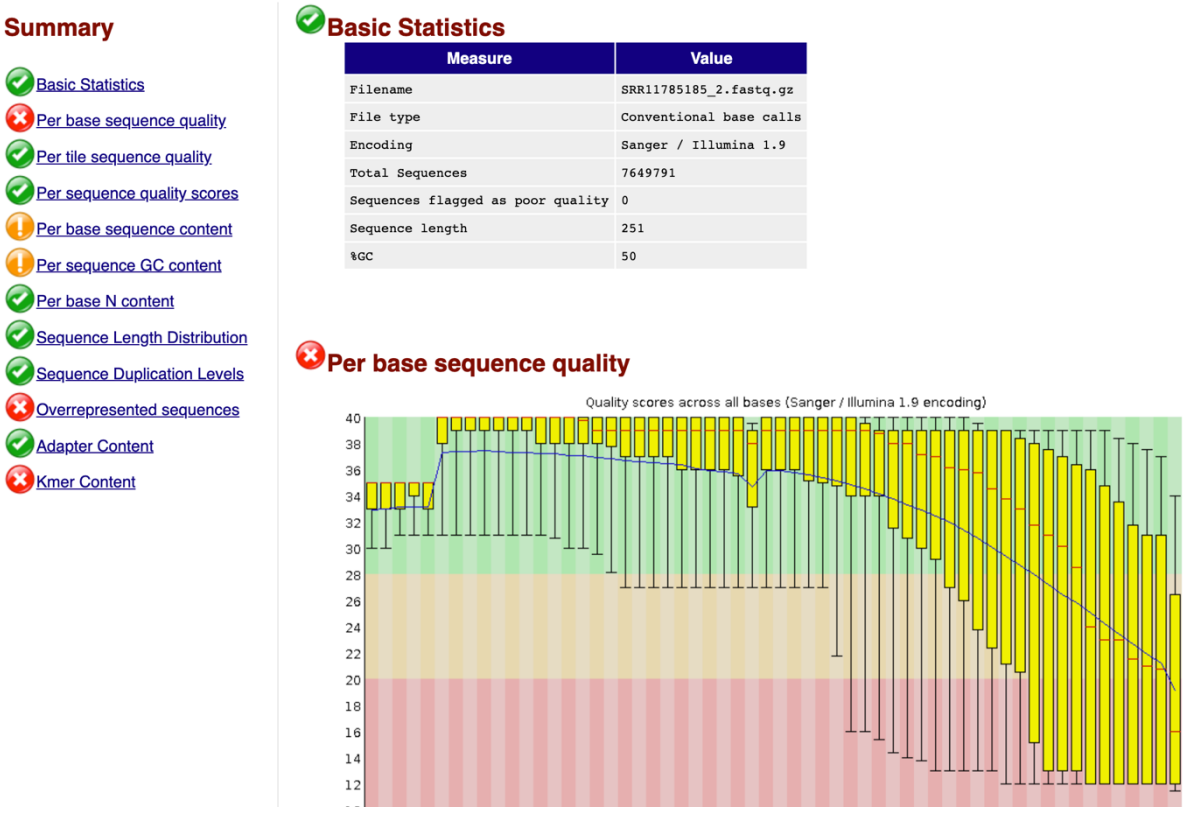


If everything worked out, you should see a list of completed workflow steps highlighted in green. The workflow generates many output files, however not all of the output files are visible. You can explore all the hidden files clicking on the word “hidden” (orange circle) – this will reveal all hidden files.

- **Explore the FastQC results.**

  To do this find the step called “FastQC on collection ##: Webpage”. Click on the name this will open up the FastQ pairs, click on one of them then click on view data icon ( ) on either forward or reverse. Note that each FastQ file will have its own FastQC results.





Explore the differential expression results.

We will explore two output files:

A. **DESeq2 Plots** – you can view these directly in galaxy by clicking on the view icon. These plots give you an idea about the quality of the experiment. The link above includes a detailed description of the graphs.

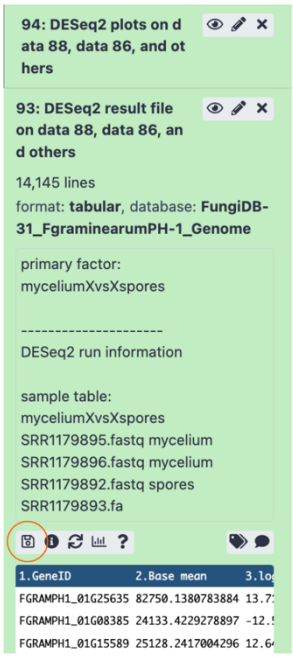
B. **DESeq2 results file** – this is a table which contains the actual differential expression results. These can be viewed within galaxy but it will be more useful to download this table and open in Excel so you can sort results and big genes of interest.

The tabular file contains 7 columns:

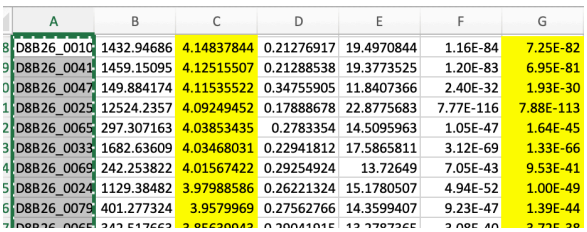
COLUMN	DESCRIPTION
1	Gene Identifiers
2	mean normalized counts, averaged over all samples from both conditions
3	the logarithm (to basis 2) of the fold change (See the note in inputs section)
4	standard error estimate for the log2 fold change estimate
5	Wald statistic
6	p value for the statistical significance of this change
7	p value adjusted for multiple testing with the Benjamini-Hochberg procedure which controls false discovery rate (FDR)

- **Download DESeq2 results (tabular format) by clicking on the floppy disk save icon.**

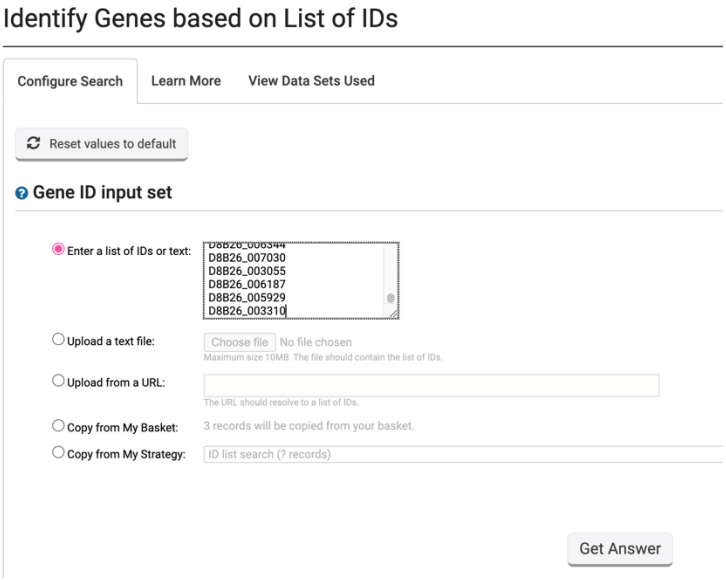
*** **Important**: the file name ends with the extension ". tabular " change this to .txt and then open the file in Excel.



- **Explore the results in Excel.**
  1. Sort them based on the log2 fold change – column 3.
  2. Pick a list of gene IDs from column 3 that are upregulated with a good corrected P value (column 7) and load then into FungiDB using the "List of IDs" search.



	A	B	C	D	E	F	G
8	D8B26_0010	1432.94686	4.14837844	0.21276917	19.4970844	1.16E-84	7.25E-82
9	D8B26_0041	1459.15095	4.12515507	0.21288538	19.3773525	1.20E-83	6.95E-81
0	D8B26_0047	149.884174	4.11535522	0.34755905	11.8407366	2.40E-32	1.93E-30
1	D8B26_0025	12524.2357	4.09249452	0.17888678	22.8775683	7.77E-116	7.88E-113
2	D8B26_0065	297.307163	4.03853435	0.2783354	14.5095963	1.05E-47	1.64E-45
3	D8B26_0033	1682.63609	4.03468031	0.22941812	17.5865811	3.12E-69	1.33E-66
4	D8B26_0069	242.253822	4.01567422	0.29254924	13.72649	7.05E-43	9.53E-41
5	D8B26_0024	1129.38482	3.97988586	0.26221324	15.1780507	4.94E-52	1.00E-49
6	D8B26_0079	401.277324	3.9579969	0.27562766	14.3599407	9.23E-47	1.39E-44



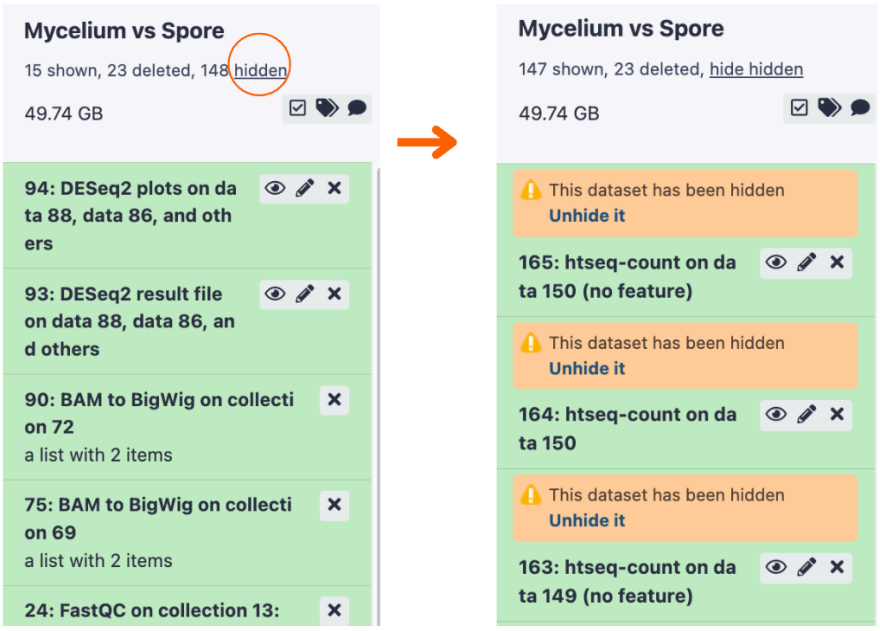
  3. Next, analyze results with GO or metabolic enrichment tools. Note: you can do the same for down-regulated genes.

Exporting data to VEuPathDB

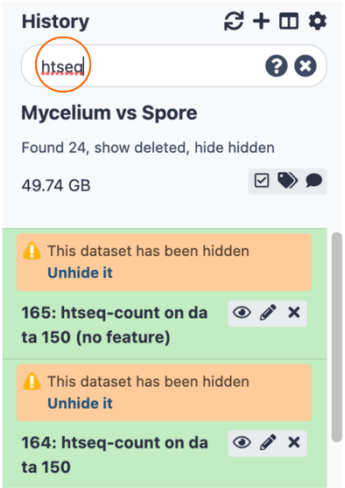
The VEuPathDB RNAseq export tool provides a mechanism to export your RNAseq results (TPM values) and BigWig RNAseq coverage files. The advantage of doing this is that it allows you to search the TPM data using the RNAseq search in VEuPathDB and view the BigWig files in the genome browser.

However, to use this feature you need to generate TPM values for genes in your datasets and organize your results into two collections, one for the TPMs and one for the BigWigs.

- **Create a Dataset List with "htseq-count on data" files.**
  1. **Reveal hidden files**.
     Click on the link at the top of your history that says "## hidden". This will show all hidden files.

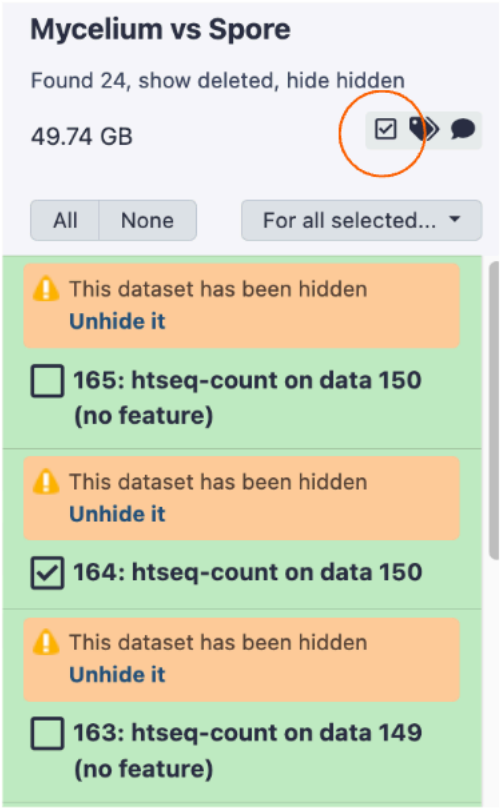


  2. **Search for htseq-count files.**
     Use the search datasets box at the top of your history to find any file in your history with the work "htseq-count". To do this, type "htseq" and click the "Enter" key on your keyboard.

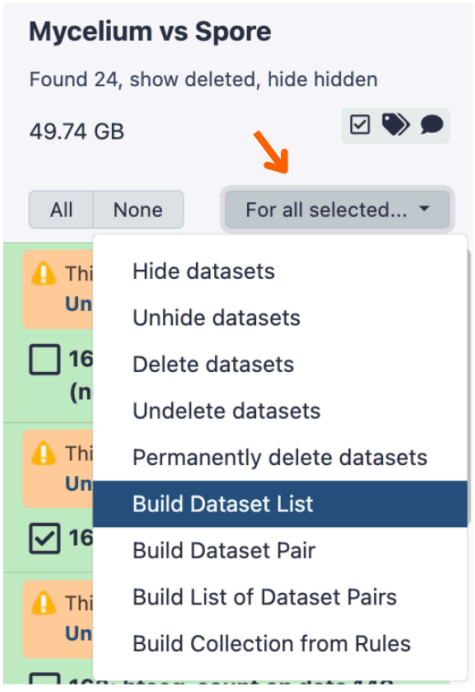


3. **Select "htseq-count on data" files.** Click on the "operation on multiple datasets" tool and select the individual htseq-count files. These should look something like this: **htseq-count on data** xx. Do not select "no feature" or "..on collection" files.

   Note: if you are comparing two conditions each done in duplicate then you should have selected 4 files.

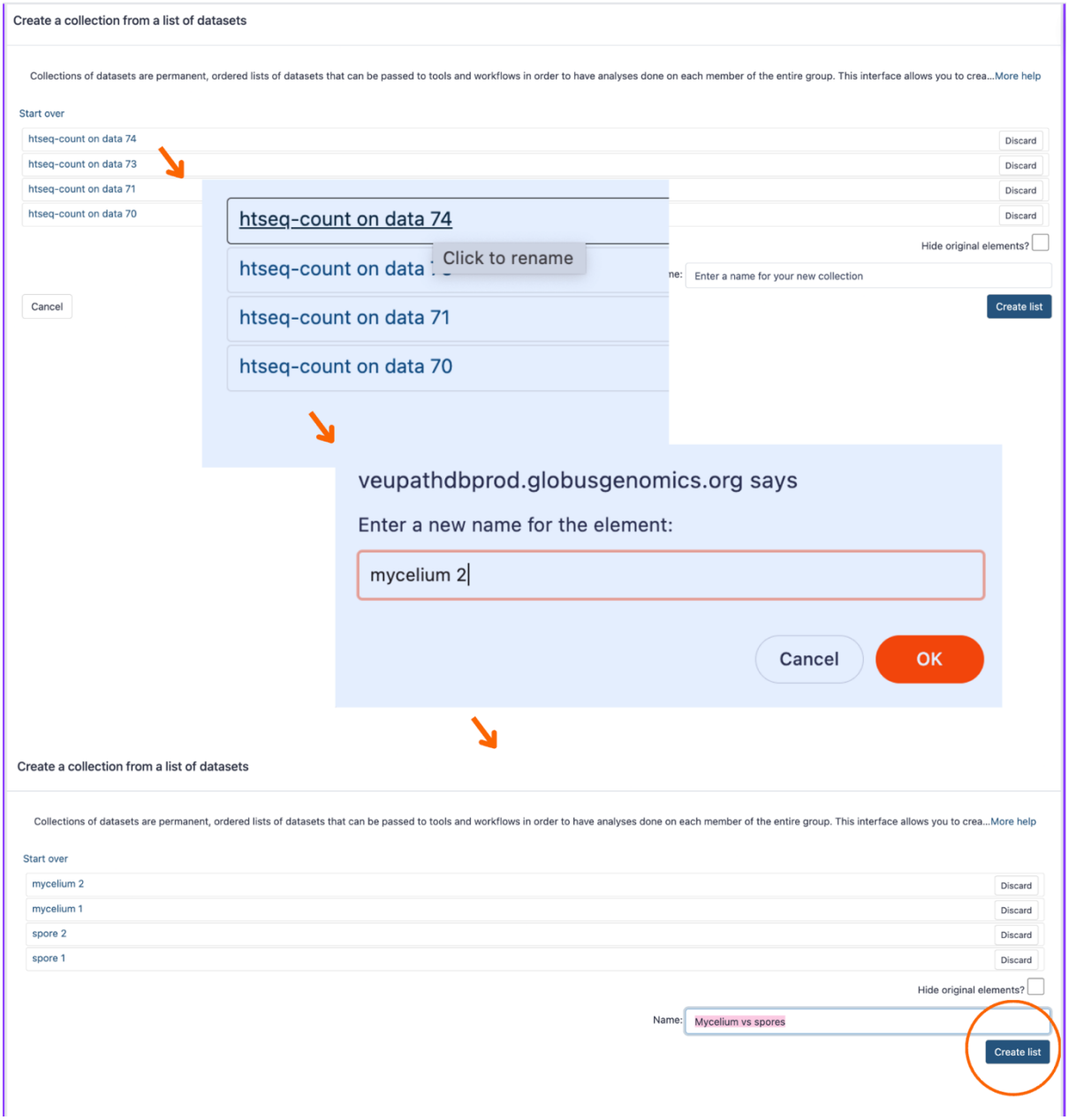


4. **"Build dataset list".** Click on the "For all selected" button and choose the "Build dataset list" option.



5. **Rename each htseq-count sample, give the collection a name and create a dataset list.**

Note: the htseq-count files will in the same order as the raw files loaded into the history. Use the "Guide to FPG2023 RNA-Seq histories and file organisation" in Part 1 for more info.



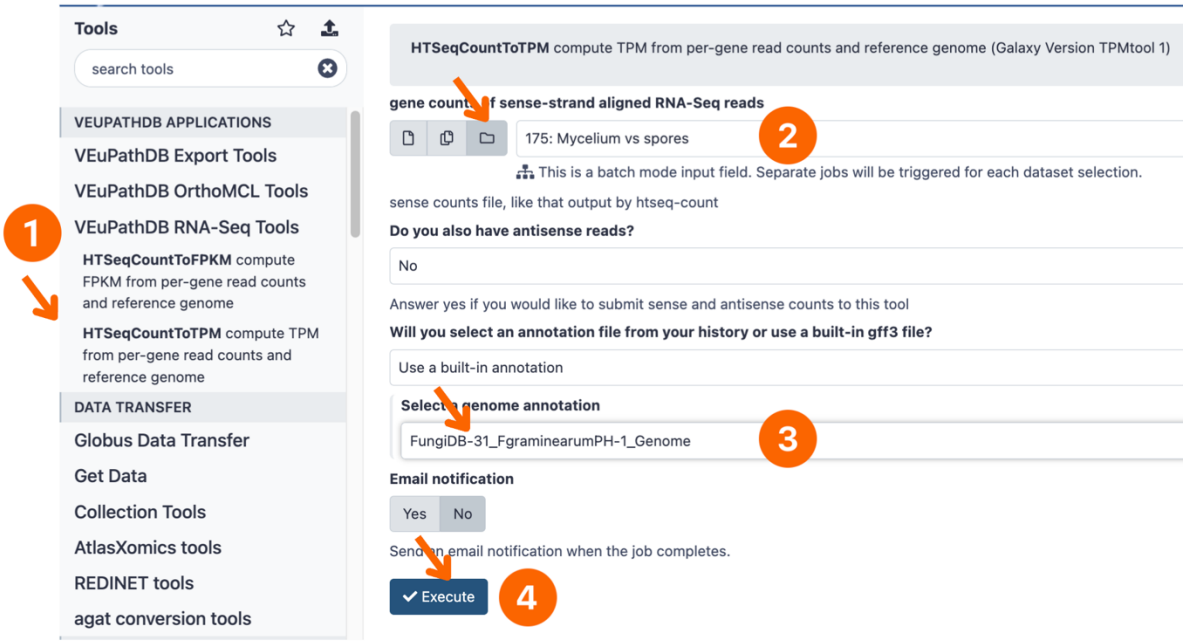
- **Create a Dataset List with "BAM to BigWig on data" files.**

Use the tutorial for htseq-count files to create a dataset list with BigWig files. Do not use "BAM to BigWig on collection" files.

Now that your count and bigwig files are nice and organized, the next step is to convert the counts into TPMs.

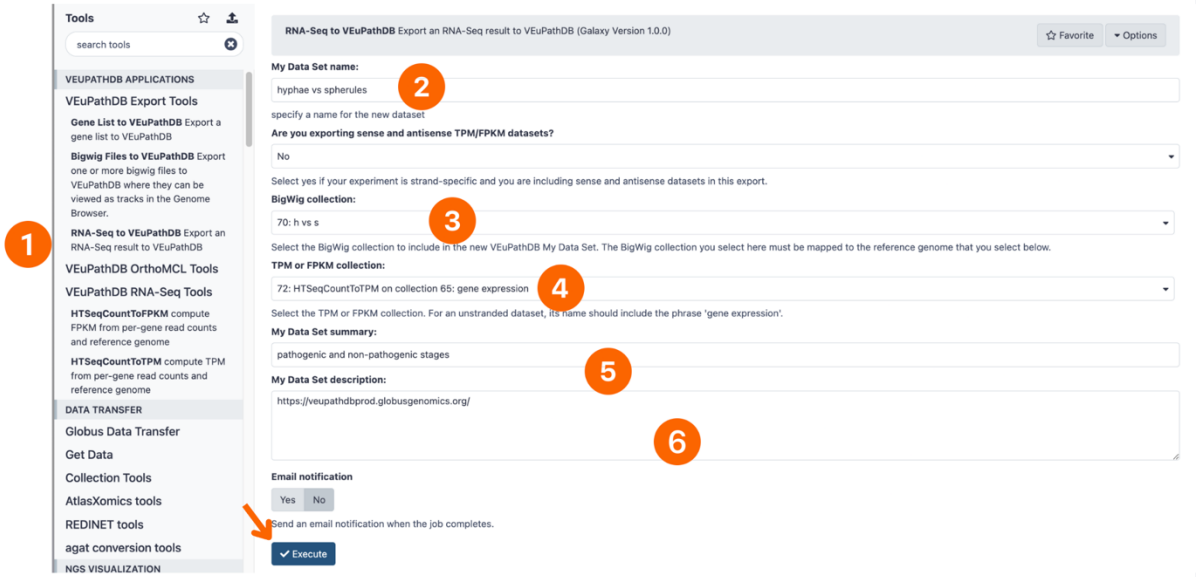
- **Use the HTSeqCountToTPM tool to convert counts to TPM**

    1. Select the HTSeqCountToTPM tool (under the VEupathDB RNAseq tools in the left menu).
    2. Make sure the list of count files is selected.
    3. Select the reference organism.
    4. Click on the “Execute” button.

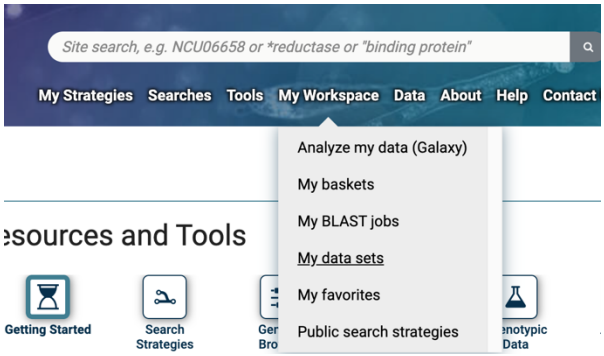


- **Export TPM counts and BigWig data to VEuPathDB/FungiDB workspace.**

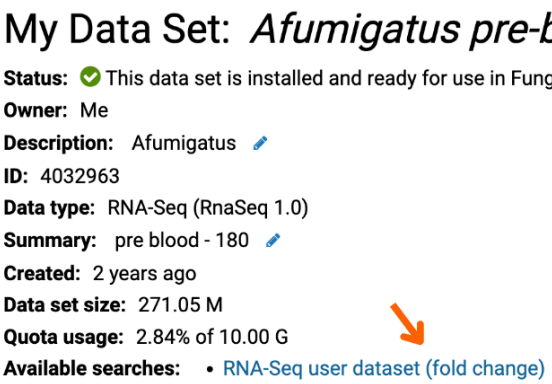
    1. Click on “VEuPathDB Export Tools” > “RNA-Seq to VEuPathDB”
    2. Enter a Data Set name.
    3. Choose, if not already selected, the correct BigWig collection.
    4. Choose, if not already selected, the correct TPM collection.
    5. Provide a data set summary.
    6. Provide a data set description and click on the “Execute” button.



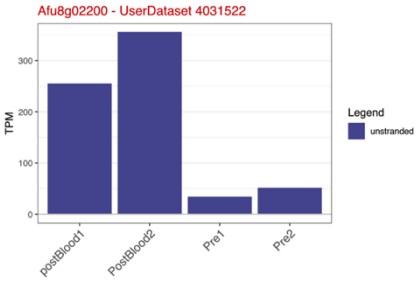
- **Explore your data in FungiDB**
  1. Click on the "My Workspace" link in the grey menu bar. Then select "My data sets" from the list.



  2. Explore the RNA-Seq dataset via the fold-change search in FungiDB.



Note that custom graphs are generated for your data in the results table so you can easily visualize the results for each gene.



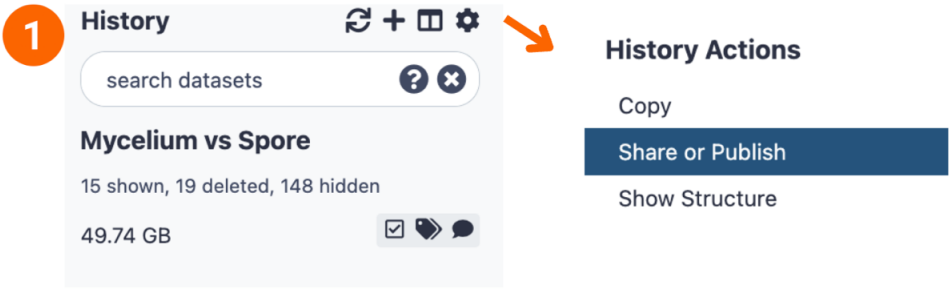
  3. Explore the coverage plots in the genome browser.

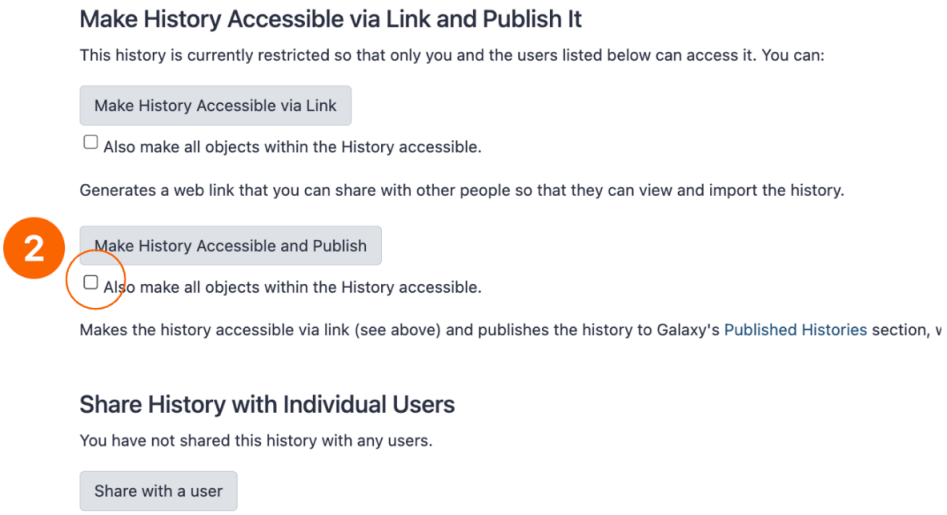
Variant Calling analysis, Part 2: Analyzing results (Group Exercise)

Learning objectives:

- Share and publish your workflow histories.
- Examine the outputs.
- View VCF files in JBrowse.
- Examine the filtered VFC file, extract Gene IDs, and create a Venny diagram.

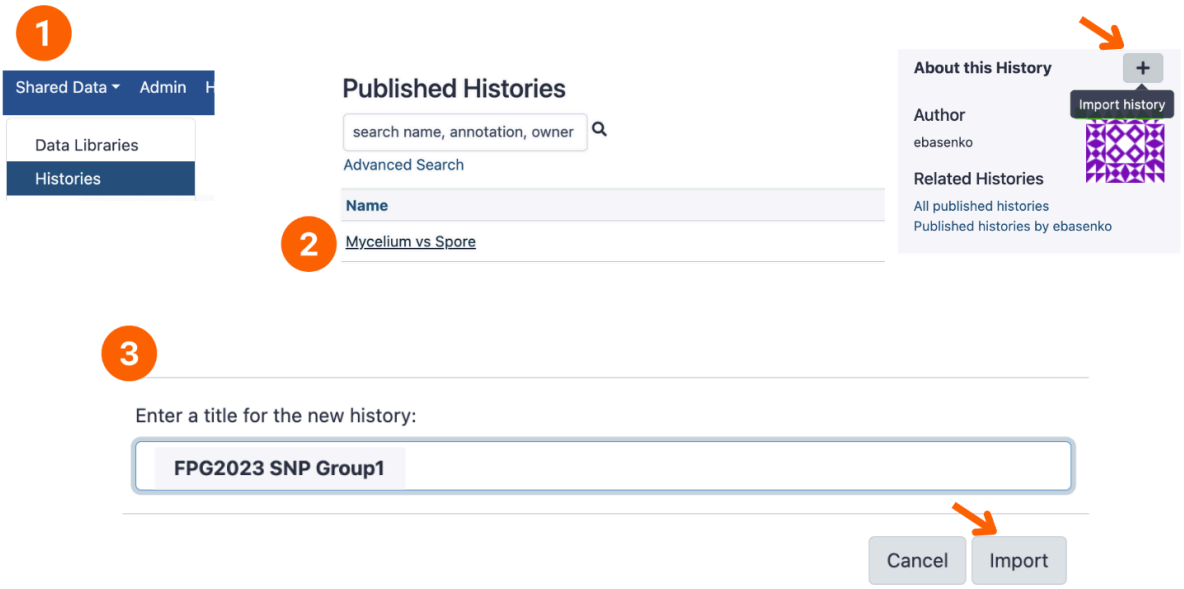
- **Share workflow histories with others.**
    1. Make sure your history has a useful name (e.g., Group3 SNPs, etc.) and click on the history action menu icon.
    2. Select the "Make History Accessible and Publish" option and check to make sure that all objects within History are accessible.





- **Importing workflow histories and output files into your own Galaxy workspace.**

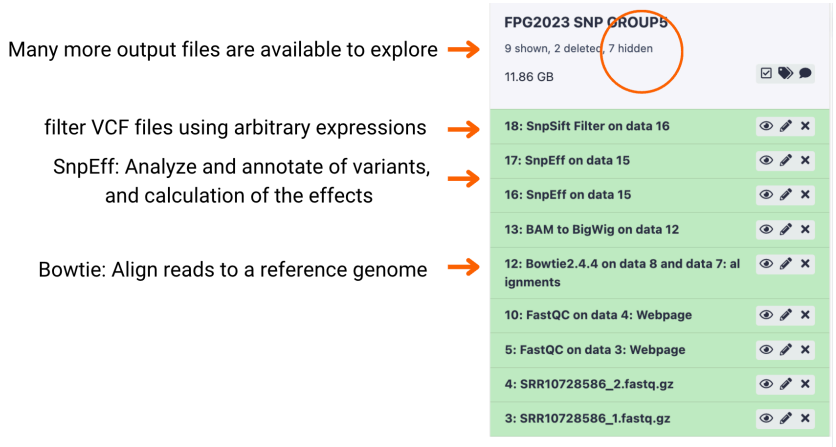
1. Click on "Shared Data" at the top and select "Histories".
2. Click on the history shared by your colleague, click on the plus icon on the far right and choose to import the history.
3. You can give it a descriptive name if you prefer or leave it as is.



If everything worked out, you should see a list of completed workflow steps highlighted in green. The workflow generates many output files, however not all of the output files are visible. You can explore all the hidden files clicking on the word "hidden" (orange circle) – this will reveal all hidden files.

The Variant calling workflow has three major components: (1) mapping of raw reads to the reference genome, (2) calling variants, and (3) annotating variants. This workflow can be used to call single nucleotide polymorphisms, insertions and deletions (also defined as indels), and multiple nucleotide polymorphisms.

In this workflow, we used Bowtie2 to align and map sequences to a reference genome. Once they are aligned it may be worth checking the quality of this process because misalignments lead to false SNP calls. SAM or BAM files provide sore this information and you can find these files to export in the hidden workflow steps.



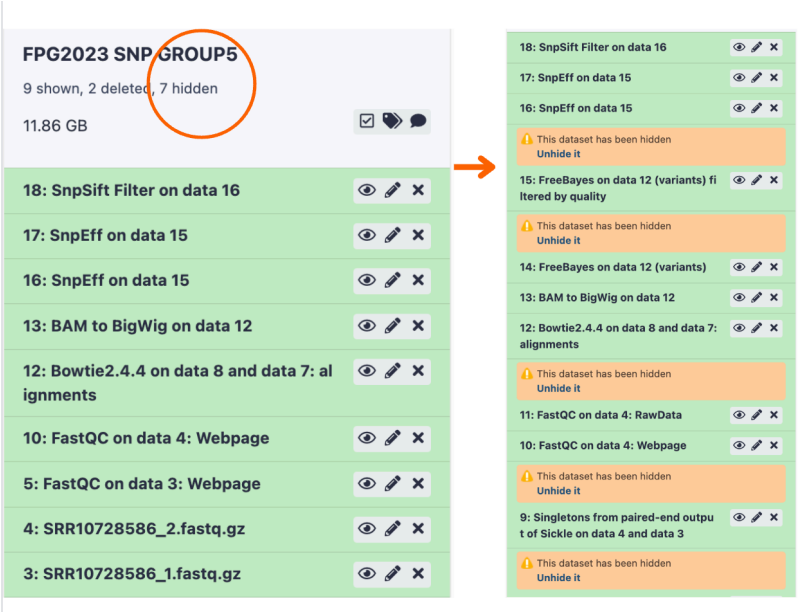
After reads have been aligned, they are sorted based on the chromosomal position. The tool that we are using is called Sort and it belongs to the suite of SAMtools. The sorted file is an input for downstream FreeBayes that calls SNPs and outputs into SnpEff that annotates variants.

Analysis and annotation of the genomic variants are carried out by the SnpEff tool. SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes). It uses reference genome to annotate genomic variants based on their genomic location and also predicts SNP coding effects. The genomic location features are intronic regions, 5' and 3' UTRs, and upstream, downstream, splice site and intergenic regions. SNP coding effects are categorized based on the effect of the amino

acid change and are classified into synonymous and non-synonymous, gain or loss of start codons, gain of loss of stop codon, and frame shifts.

The SnpSift tool annotates, filters, and manipulates genomic annotated variants. Once you annotated your files using SnpEff, you can use SnpSift to help you filter large genomic datasets (e.g. sort on high or moderate impact SNPs, etc.).

- **Examine your results.**

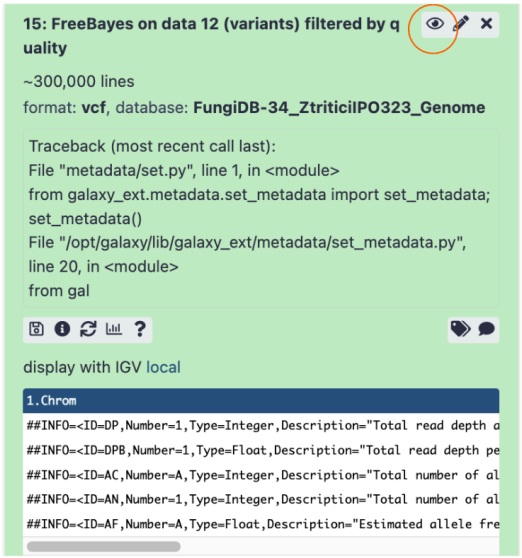


1. Click on the *hidden* files link in the history panel to reveal all workflow output files.
2. Examine the output files.
3. What does the tool FASTQC do?
4. What about Sickle?

The output of Sickle is used by a program called Bowtie2. Bowtie generates a file called a BAM file. Whenever dealing with sequence alignment files you will likely hear of file formats called SAM or BAM. SAM stands for Sequence Alignment/Map format, and BAM is the binary version of a SAM file.

Many of the downstream analysis programs that use BAM files require a sorted BAM file. This allows for more efficient analysis.

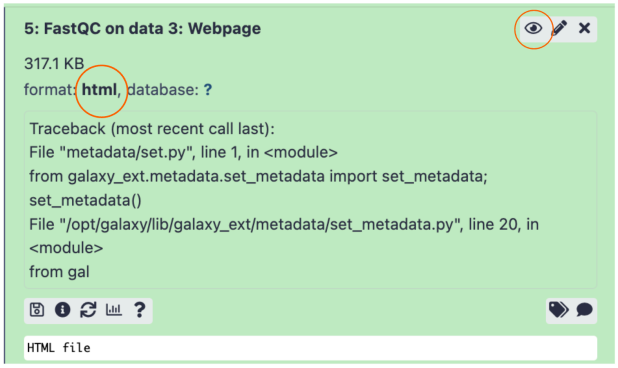
The sorted BAM file is the input for a program called FreeBayes. This program is a Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs (single-nucleotide polymorphisms), indels (insertions and deletions), MNPs (multi-nucleotide polymorphisms), and complex events (composite insertion and substitution events) smaller than the length of a short-read sequencing alignment. The output for many variant callers is a file called a VCF file. VCF stands for variant interchange format.

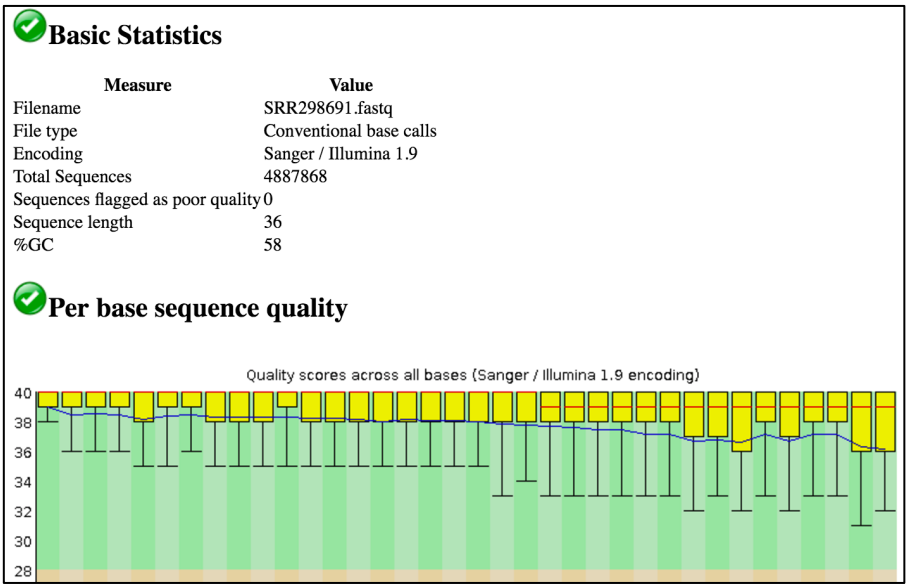


5. Examine the VCF file in your results (click on the *eye* icon to view its contents). Detailed information about VCF file content is available here: https://samtools.github.io/hts-specs/VCFv4.2.pdf

- **Examine sequence quality based on FastQC quality scores.**

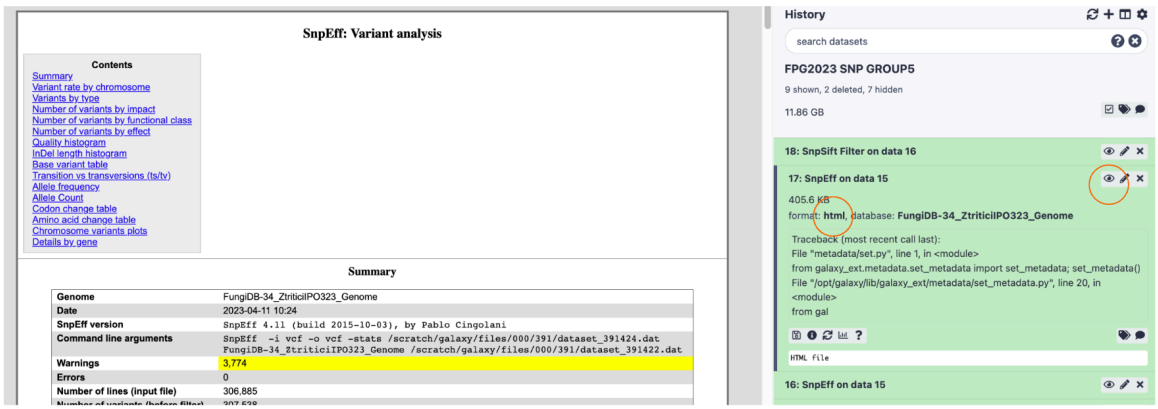
FastQC provides an easy-to-navigate visual representation sequencing data quality and distribution of nucleotides per read position. What does the report tell you about the quality ?





- **Examine SnpEff summaries (html)**
    - Click on the *View data icon* (eye) in the SnpEff output file that has the html format.

This will open the html file in Galaxy for your review.



The header contains a short summary and information about the run and it has several major components:

The Summary contains warnings about possible genomic annotation errors or inconsistencies identified in the reference genome. If there are many, use caution interpreting results and examine associated gff files for any issues (*e.g.* missing feature values in gff files, incomplete gene sequences, more than one stop codon per gene, *etc.*). Other components:

- Number of line (input file) - number of lines in vcf file
- Number of not variants: 0 - some packages report non-variant observations for nt positions between reference genome and vcf file generate.
- Number of known variants and multi-allelic VCF entries - if you work with a model organism where some variants were given an accession number (most commonly in mice and human projects) any recognised variants will be listed here

Summary

Genome	FungiDB-34_ZtriticilPO323_Genome
Date	2023-04-11 10:24
SnpEff version	`SnpEff 4.11 (build 2015-10-03), by Pablo Cingolani`
Command line arguments	`SnpEff -i vcf -o vcf -stats /scratch/galaxy/files/000/391/dataset_391424.dat FungiDB-34_ZtriticiIPO323_Genome /scratch/galaxy/files/000/391/dataset_391422.dat`
Warnings	3,774
Errors	0
Number of lines (input file)	306,885
Number of variants (before filter)	307,538
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	307,538
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	653
Number of effects	1,280,819
Genome total length	39,730,198
Genome effective length	39,730,198
Variant rate	1 variant every 129 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
Ztri_MitoScaffold	43,947	18	2,441
Ztri_chr_1	6,088,797	44,156	137
Ztri_chr_10	1,682,575	15,039	111
Ztri_chr_11	1,624,292	14,012	115
Ztri_chr_12	1,462,624	12,767	114
Ztri_chr_13	1,185,774	10,694	110
Ztri_chr_14	773,098	2,064	374
Ztri_chr_15	639,501	7,821	81
Ztri_chr_16	607,044	5,094	119

- Number of effects - SNP effects summary by type and regions
- Genome total length - number of bp in the reference genome
- Genome effective length - how many nucleotides can be mapped back to the genome
- Variant rate - higher frequency of variants before samples can indicate selective pressure

Summary statistics for variant types

Here is an example of variant calls and what they mean in terms of nucleotide changes:

Number variantss by type

Type	Total
SNP	114,034
MNP	12,864
INS	6,907
DEL	7,304
MIXED	2,180
INTERVAL	0
Total	143,289

Type	What is means	Example
SNP	Single-Nucleotide Polymorphism	Reference = 'A', Sample = 'C'
Ins	Insertion	Reference = 'A', Sample = 'AGT'
Del	Deletion	Reference = 'AC', Sample = 'C'
MNP	Multiple-nucleotide polymorphism	Reference = 'ATA', Sample = 'GTC'
MIXED	Multiple-nucleotide and an InDel	Reference = 'ATA', Sample = 'GTCAGT'

Statistics for the variant effects and impacts:

- **High impact** normally refers to frame shift or new stop codon detections as those changes will generate profound effects on gene function.

- **Modifier SNPs** can affect promoter function, while low and moderate SNPs are most commonly identified inside genes and are either non-coding or non-synonymous SNPs.

- Base changes summary. SnpEff html files provide a breakdown of SNPs across gene features:

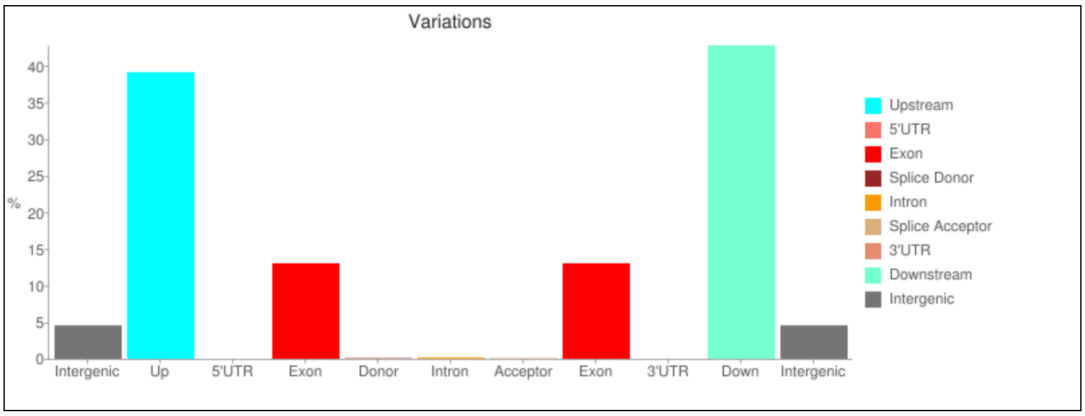
Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	1,857	0.145%
LOW	87,874	6.861%
MODERATE	41,970	3.277%
MODIFIER	1,149,118	89.717%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	29,331	28.472%
NONSENSE	370	0.359%
SILENT	73,317	71.169%

Type (alphabetical order)	Count	Percent
DOWNSTREAM	321,858	40.292%
EXON	67,505	8.451%
INTERGENIC	74,749	9.358%
INTRON	1,064	0.133%
NONE	1	0%
SPLICE_SITE_ACCEPTOR	5	0.001%
SPLICE_SITE_DONOR	4	0.001%
SPLICE_SITE_REGION	176	0.022%
TRANSCRIPT	12	0.002%
UPSTREAM	333,432	41.741%



Additionally, you may see several SNPs being reported in several classes: missense variant + splice region variant. This means that some SNPs that are found within certain splice sites

also contain a missense variant. SNPs in the splice sequences may affect intron splicing and lead to read through.

- Quality of reads is indicated in Phred's scale and is a good indicator of the quality of your datasets and results. Quality scores are normally represented by a bar graph where count = number of SNPs and X axis is quality score (higher score mean better p-values and high confidence of the results)

- Base changes: Reflects the frequency of base changes (purine-purine, purine-pyrimidine, pyrimidine-purine, pyrimidine-pyrimidine).

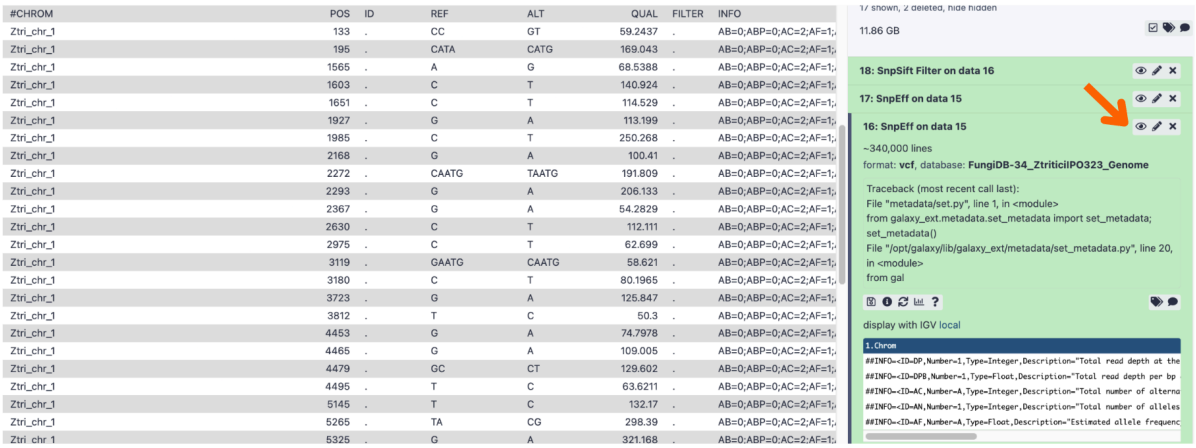
- Transition and transversion ratio help to identify if you may have a selective pressure on certain alleles (high ratio suggests that genes may be under selective pressure).

- Allele frequency statistics reports frequency of alleles and help to identify potential sequencing artifacts due to PCR enrichment step (generation of heterozygous counts in a haploid organism).

The vcf file generated by SnpEff contains information about SNPs and the genomic location. Post-processing of SNP data is normally required to make sense of thousands of SNPs and to decide which ones have biological and functional importance. Data processing can help you to extract SNP distribution and parse associated data including GeneIDs, protein-coding annotations, and effects in sequence ontology terms such as missense or synonymous variants, stop codon gain, etc. and also link changes to the genome model. SnpSift is among other programs that is often in SNP data post-processing. It can be installed and run locally to manipulate vcf files. Alternatively, you can also visualize vcf files in Artemis (additional steps are required to format the data).

Examining SNP information.

You can view the SNP information by clicking on the "eye" icon within the SnpEff vcf file.



The vcf file generated by SnpEff contains information about SNPs and the genomic location. Here is an example of a file opened in Excel:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	unknown
CM001231	189057	.	AG	CT	787.449	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:143:0:0:143:5341:-207.887,-43.0473,0
CM001231	483825	.	G	A	64.8756	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:4:0:0:4:146:-10.0999,-1.20412,0
CM001231	518226	.	G	C	51.7908	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:8:0:0:7:276:-11.5007,-2.10721,0
CM001231	574021	.	C	G	237.265	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:17:0:0:17:583:-39.079,-5.11751,0
CM001231	609879	.	GAA	CAG	55.2785	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:32:8:277:22:861:-18.1711,-0.694735,0
CM001231	1090073	.	G	T	79.4156	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:8:2:75:6:238:-11.5539,-1.36362,0
CM001231	1090104	.	A	T	70.961	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:6:0:0:6:220:-12.5146,-1.80618,0
CM001231	1153611	.	CCTC	GCTG	111.123	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:8:5:188:3:97:-9.30616,-6.1461,0
CM001231	1159150	.	CT	GC	126.126	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:31:0:0:19:741:-29.7713,-5.71957,0
CM001231	1159438	.	C	G	82.3312	.	AB=0;ABP=0;	GT:DP:RO:QI	0/0:47:30:1092:17:640:0,-9.53002,-3.50705
CM001231	1159465	.	G	C	249.656	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:126:47:1770:79:3013:-53.8644,-25.2134,0
CM001231	1159499	.	T	C	124.95	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:143:32:1167:111:4248:-76.1575,-33.4865,0
CM001231	1181576	.	CC	TG	191.675	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:27:0:0:25:924:-41.7448,-7.52575,0
CM001231	1293309	.	C	G	51.22	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:2:0:0:2:78:-6.92763,-0.60206,0
CM001231	1323058	.	TT	GC	71.3001	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:6:0:0:6:223:-12.5485,-1.80618,0
CM001231	1485397	.	A	G	3558.42	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:499:0:0:497:18671:-804.678,-149.612,0
CM001231	1485429	.	G	A	3783.33	.	AB=0;ABP=0;	GT:DP:RO:QI	1/1:517:1:38:516:20010:-843.425,-151.978,0

Filtering VCF file data.
VCF files contain a lot of data about variants and their positions. SnpEff generates various analyses/summaries of VCF files (including GeneIDs that overlap variant positions). However, it is often necessary to filter VCF files further to obtain useful information for your specific question. For example, you may want to filter out SNP positions that have an impact on the coding sequence.

One tool that can be used is called SnpSift Filter (look at the last step of the pipeline you just ran). This tool allows you to write complex expressions to filter a VCF file. Your workflow is set up to use an expression that filters VCF files on moderate and high impact SNPs (this setting can be adjusted manually in the workflow editor). Here is the exact expression used:

(((ANN[*].IMPACT has 'HIGH') | (ANN[*].IMPACT has 'MODERATE')) & ((na FILTER) | (FILTER = 'PASS')))

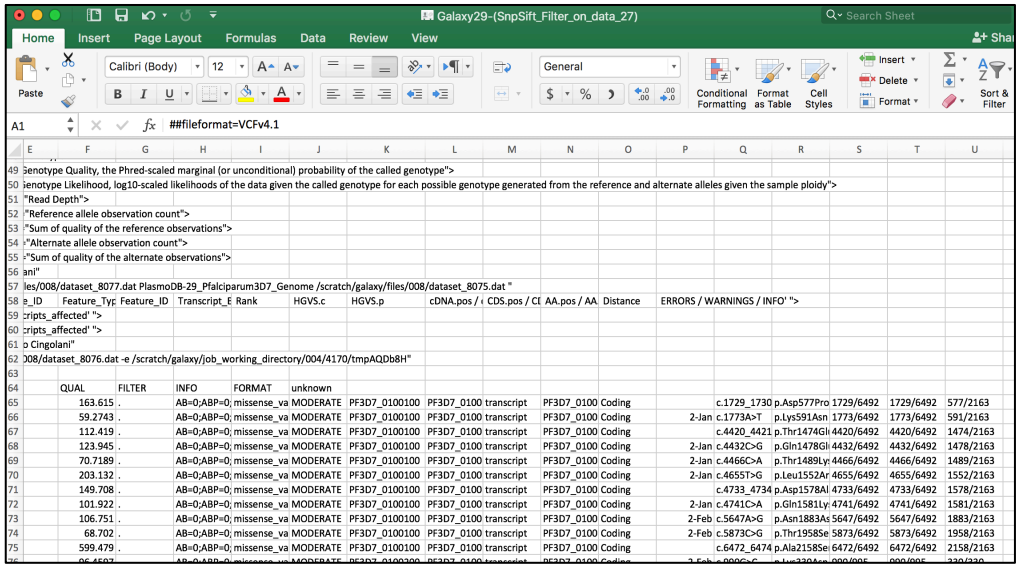
- **Extract filtered VCF file (SnpSift output) and convert into an Excel document.**

For this exercise, two groups with be sharing data SnpSift outputs: group 1 & 2, group 3 & 4, and group 5 & 6. File manipulations should be performed on both SnpSift vcf files.

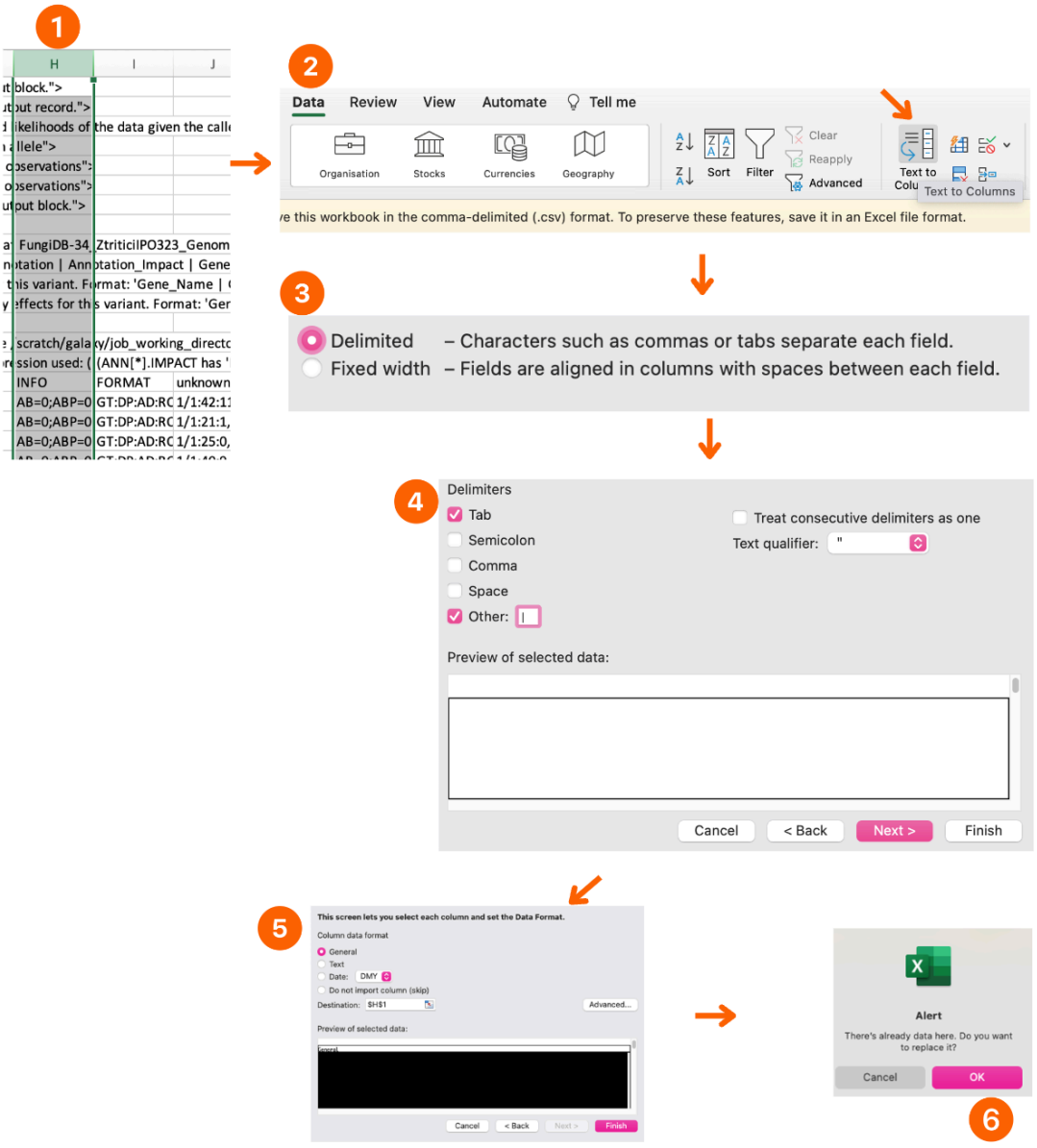
Look at the filtered vcf file in Galaxy. Notice that the Gene IDs are buried in the file, but the file has some structure which means you can extract them either programmatically or using a program like Excel.

Here are some steps you can take to extract Gene IDs from two VCF files then compare them to identify genes that are in common or that distinguish the two files.

1. Download the SnpSift Filter output by clicking on the save icon.
2. Right click and open this file with Excel.

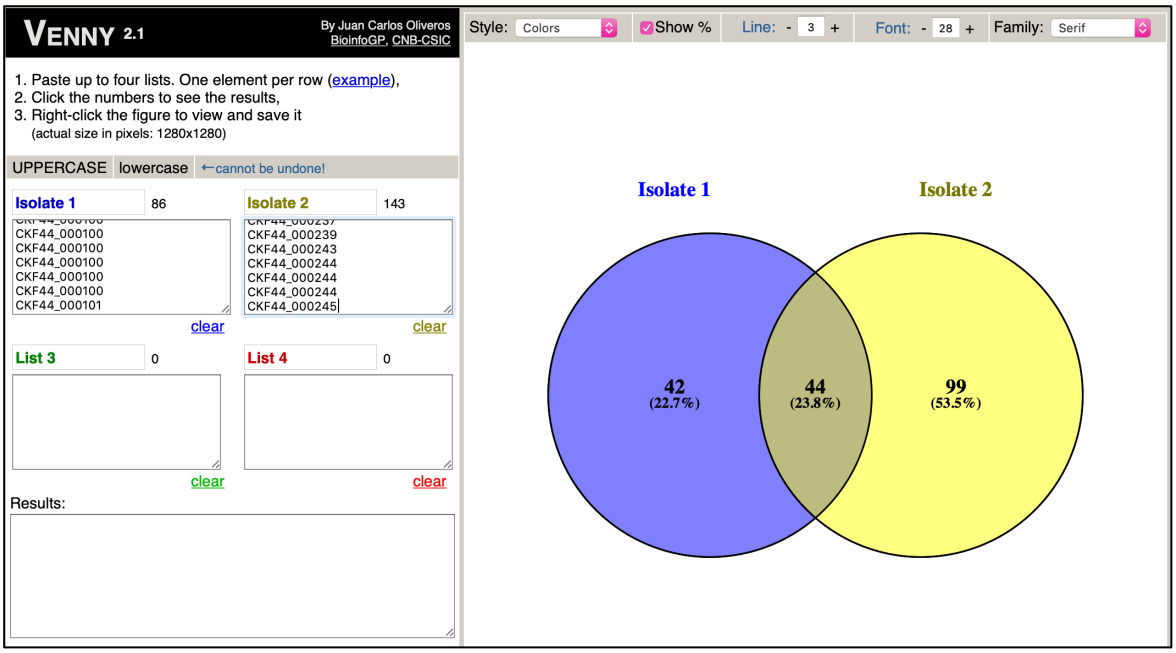


- **Manipulate Excel file to display SNP info in columns.**
  1. Select the “INFO” column.
  2. Navigate to the “Data” tab in Excel and choose “Text to Columns”.
  3. Use the “Delimited” option.
  4. Set delimiters to the “Tab” and “ | ” in the “Other” and click “Next”
  5. Leave other criteria at default and click on the “Finish” button.
  6. Click “OK” on the Alert pop-up.



Now you can look for Gene IDs of interest in the excel file. For example, if this is a known drug resistant line you can sort and examine SNPs based on their characteristics.

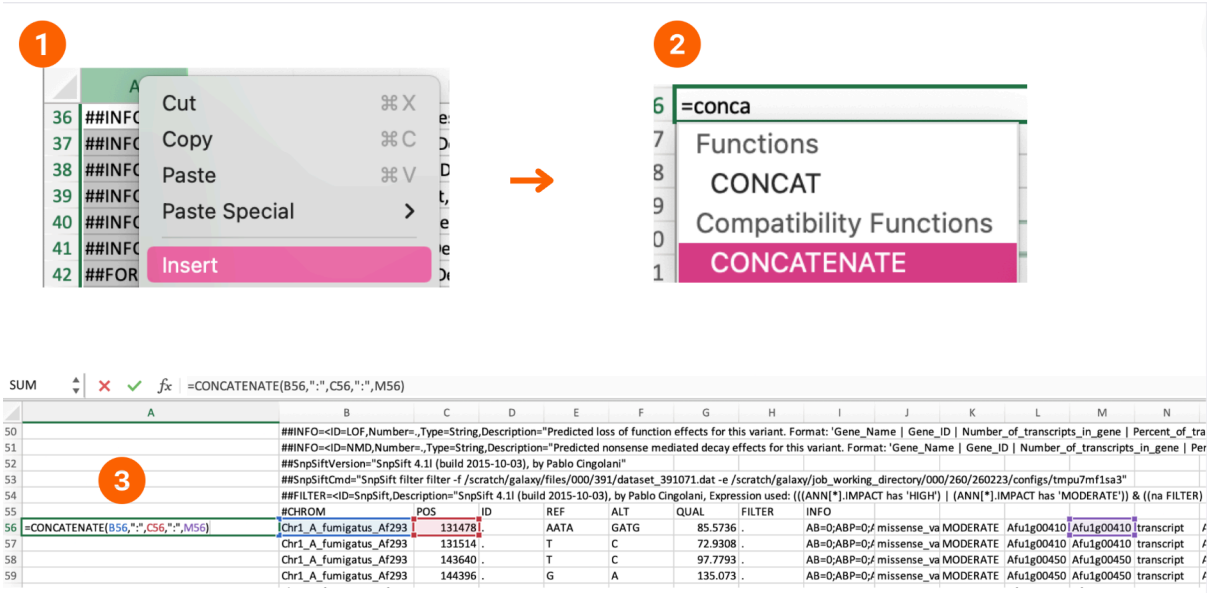
If you are comparing two or more strains, you may want to extract gene IDs from all VCF files and identify common signatures across isolates or strains. For this type of analysis, you can use http://bioinfogp.cnb.csic.es/tools/venny/ to generate a Venn diagram:



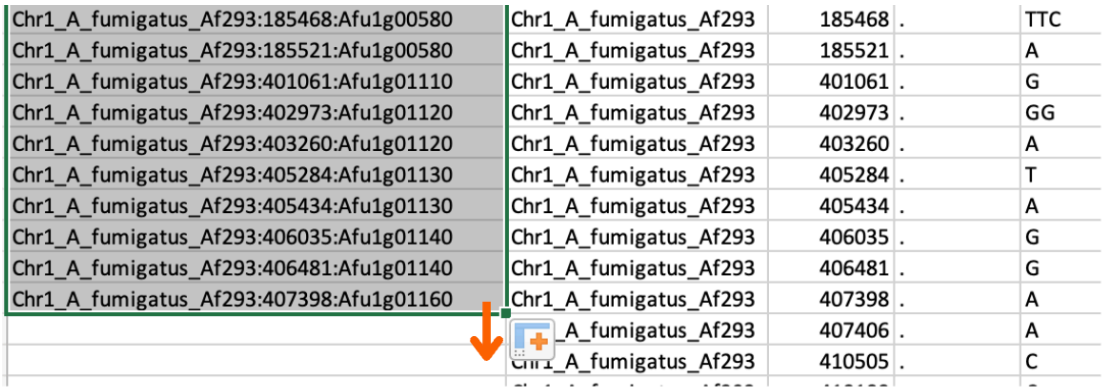
The screenshot above is showing comparison of between lists of GeneIDs. Is it possible to miss some important polymorphisms using this method? Of course, the answer is yes☺
For example, it is quite possible that a gene with a SNP in the WT and a SNP in the mutant that will be in the intersection of the two gene lists, contains different SNPs – you will miss this by doing the above steps. Below is a description of steps you can take to create a list of unique IDs for SNPs. This list of unique IDs can then be used in Venny.

- **Analyze your data in Venny.**

1. Start with the same excel files that you opened in the above section. Insert an empty column before the data.
2. Deploy the concatenate function in Excel.
3. Create a unique ID for SNPs by combining information from multiple columns to create something that looks like this: **chromosome:position:geneID**
   To do this you will use the concatenate function in Excel:
   =concatenate(cell#1,":",cell#2,":",cell#3)
   Cell#1 = cell with chromosome number
   Cell#2 = cell with position
   Cell#3 = cell with GeneID



You should get unique SNP IDs that look like this (for example): CP022321.1:15259:CKF44_000003. Copy this function for other entries:



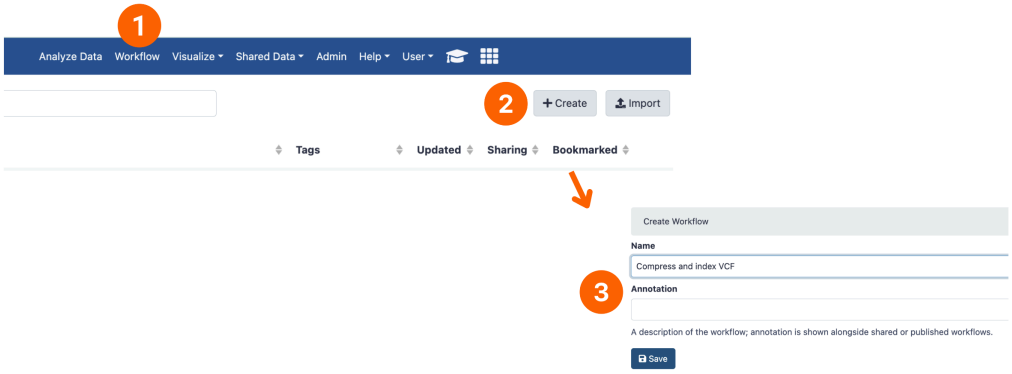
4. Copy these newly generated unique IDs into List 1 and Lis2 on Venny http://bioinfogp.cnb.csic.es/tools/venny/ and examine the data.

Viewing VCF file results in the JBrowse genome browser.

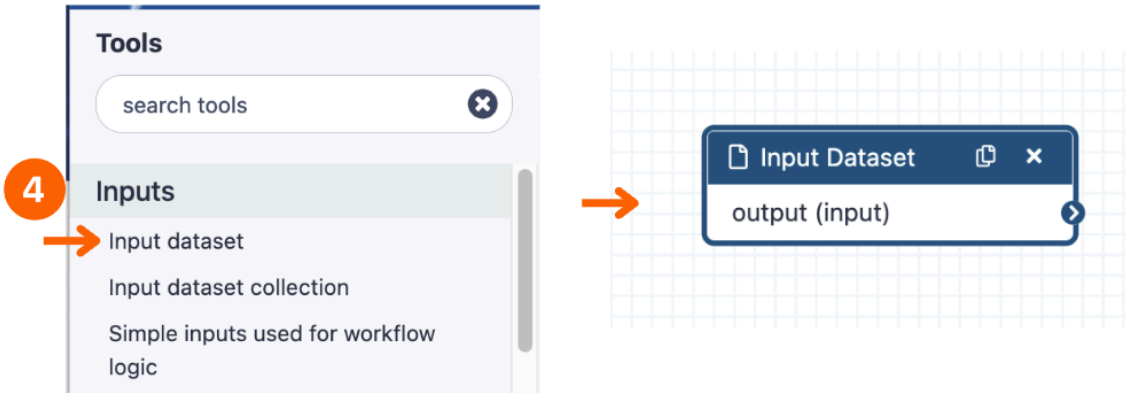
- **Create a workflow to generate a compressed vcf and index files for viewing your data in JBrowse.**

To view a VCF file in JBrowse, it first has to be indexed and compressed. This is done using two tools: bgzip and tabix, respectively. You can run these tools sequentially or you can set up a mini workflow and then run the workflow to generate the output files as follows:

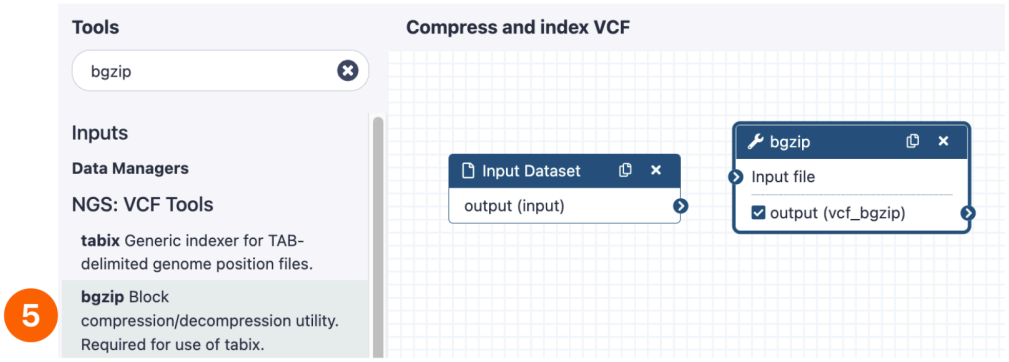
1. Click on the "Workflow" menu.
2. Click on the "Create" button to start a new workflow.
3. Give the workflow a name (e.g. Compress and index VCF) and click on the save button. This will open a workflow canvas.



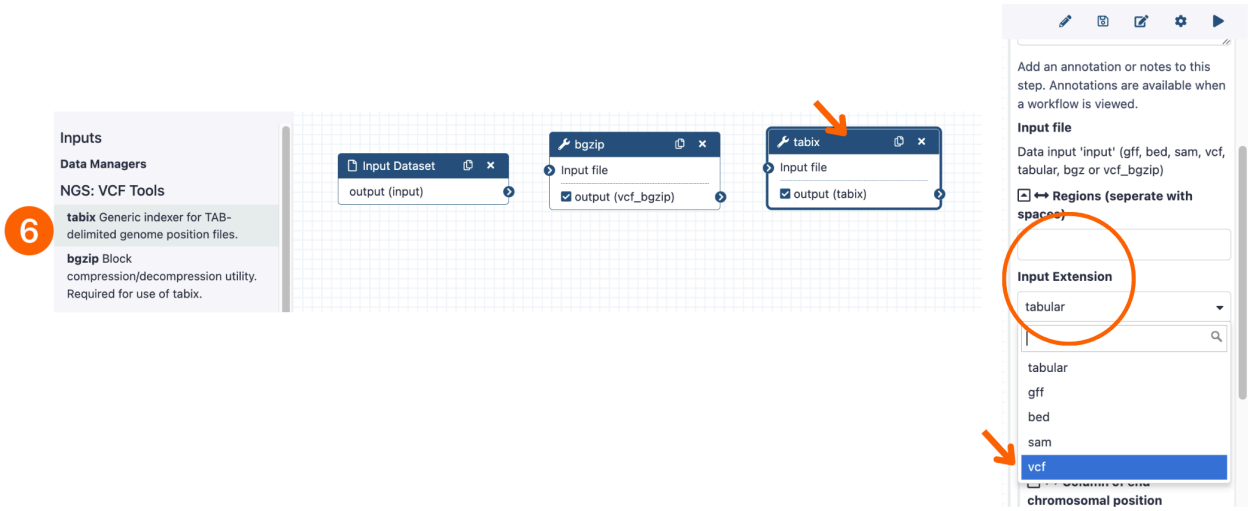
4. All workflows must start with an input file so add the "Input Dataset" step to the workflow using the menu on the left (you must click on the tool for it to appear in the workflow editor canvas).



5. Using the menu on the left, search for and add the "bgzip" tool.



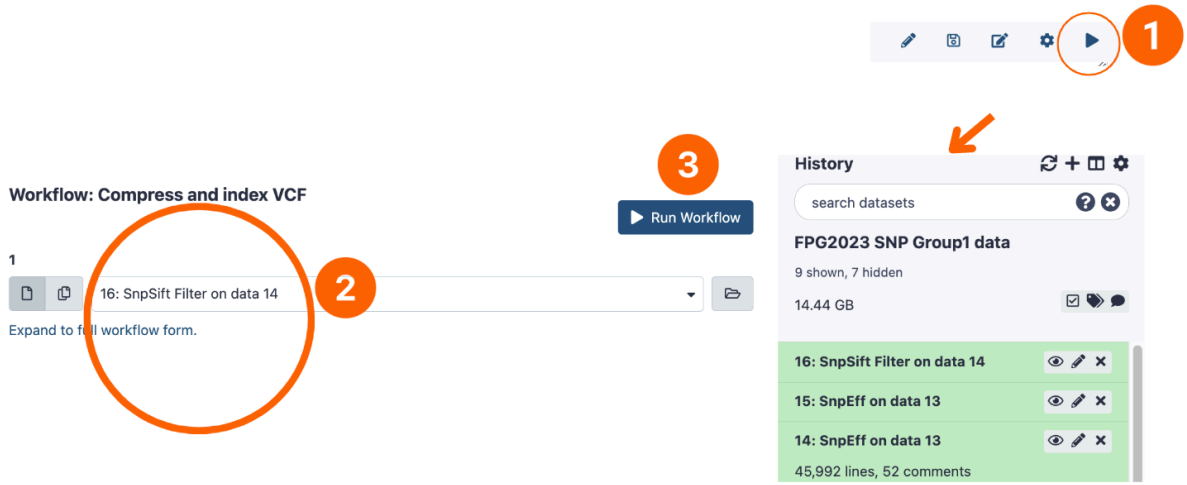
6. Using the menu on the left, search for and add the “tabix” tool. Left-click on the “tabix” icon and select “vcf” under “input selection” on the right (tool option section)



7. Connect each step/tool into a workflow and save it (the button is at the top of the screen)

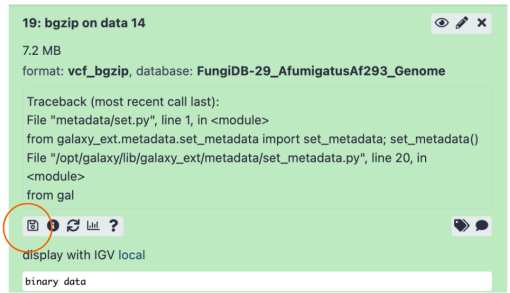
- **Run the newly created workflow to generate a compressed vcf and index files.**
    1. Click on the "Play" button to start your workflow.
    2. Select the VCF file you want to process.

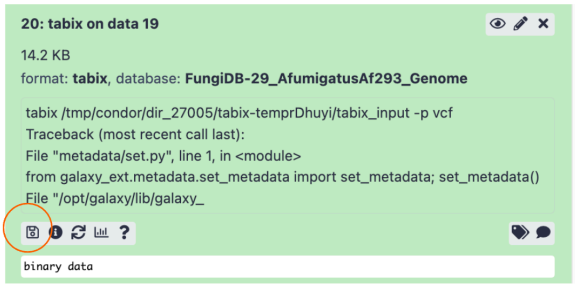
    Note that the workflow produced several vcf files - SnpSift, SnpEff.. In the screenshot below we will use a vcf filtered for high and moderate impact SNPs but if you are interested in all mutations, you may want to choose another file.
    3. Click on the "Run Workflow" button.



After the workflow completed running, you should have 2 new files in the history on the right (tabix and bgzip).

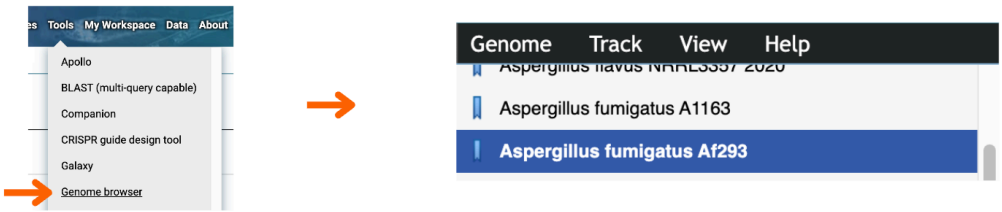
- **Download compressed vcf (vcf_bgzip) and index (tabix) files and view them in JBrowse.**
    1. Download both files by clicking on the download icon. You will need both files.



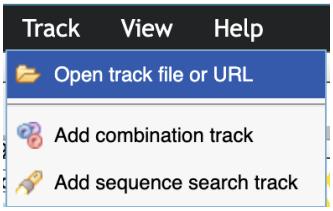


    2. After the files are downloaded, rename them as follows:
        a. The **vcf_bgzip** file to "**group#.vcf.gz"** (i.e. **group1.vcf.gz**)
        b. The **tabix** file to "**group#.vcf.gz.tbi**" (i.e. **group1.vcf.gz.tbi**)

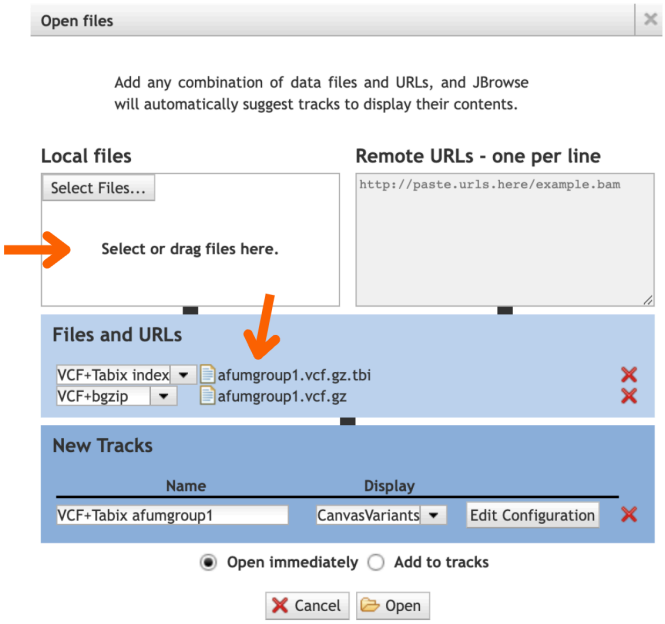
3. Navigate to JBrowse in FungiDB and select the correct genome from the Genome drop-down menu.



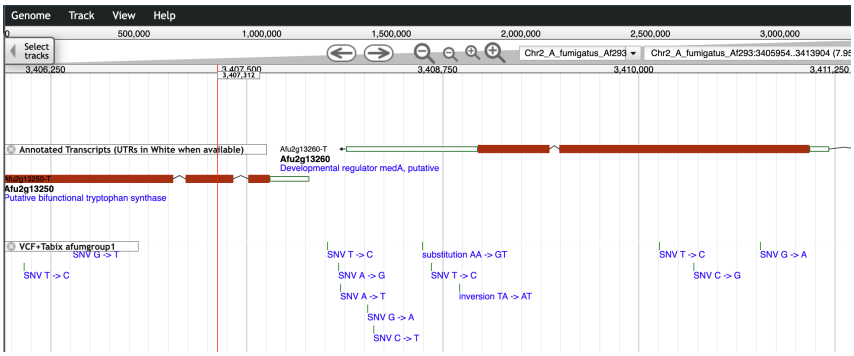
4. Click on the Track menu, select "Open track file or URL".



5. Drag and drop your files in the window that appears. Notice that the file formats are autodetected. Click on the "Open" button at the bottom of the pop-up.



You should now be able to view the SNPs in JBrowse.



MycoCosm: KEGG Browser

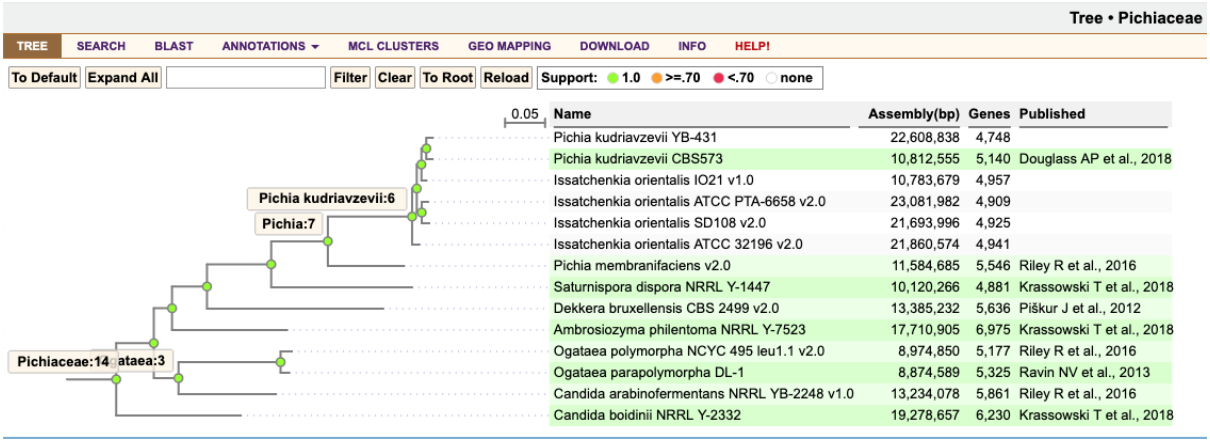
KEGG stands for Kyoto Encyclopedia of Genes and Genomes at http://www.genome.jp/kegg/, which maintains a curated set of EC-annotated enzymes and their pathways. Each portal's KEGG Browser facilitates display and discovery of MycoCosm's KEGG-annotated genes. Using the KEGG browser, one can search or browse through KEGG metabolic and regulatory pathways to retrieve information about the enzymes, pathways, and proteins associated with the KEGG annotations.

Scenario: You have plated a variety of yeasts on a variety of carbon sources, and discovered that some members of the Pichiaceae grow on galactose (e.g., *Dekkera bruxellensis*) and some do not (e.g., *Pichia membranifaciens*). Use MycoCosm to find genes that could explain this metabolic difference.

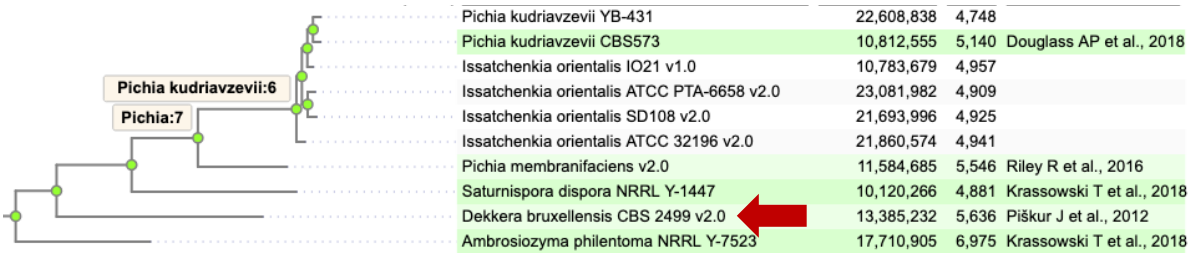
1) Go to the MycoCosm Pichiaceae PhyloGroup at mycocosm.jgi.doe.gov/Pichiaceae:

##	Name	Assembly Length	# Genes	Published
1	Candida arabinofermentans NRRL YB-2248 v1.0	13,234,078	5,861	Riley R et al., 2016
2	Candida boidinii NRRL Y-2332	19,278,657	6,230	Krassowski T et al., 2018
3	Dekkera bruxellensis CBS 2499 v2.0	13,385,232	5,636	Piškur J et al., 2012
4	Issatchenkia orientalis ATCC 32196 v2.0	21,860,574	4,941	
5	Issatchenkia orientalis ATCC PTA-6658 v2.0	23,081,982	4,909	
6	Issatchenkia orientalis IO21 v1.0	10,783,679	4,957	
7	Issatchenkia orientalis SD108 v2.0	21,693,996	4,925	
8	Ogataea parapolymorpha DL-1	8,874,589	5,325	Ravin NV et al., 2013
9	Ogataea polymorpha NCYC 495 leu1.1 v2.0	8,974,850	5,177	Riley R et al., 2016
10	Pichia kudriavzevii CBS573	10,812,555	5,140	Douglass AP et al., 2018
11	Pichia kudriavzevii YB-431	22,608,838	4,748	
12	Pichia membranifaciens v2.0	11,584,685	5,546	Riley R et al., 2016
13	Saturnispora dispora NRRL Y-1447	10,120,266	4,881	Krassowski T et al., 2018

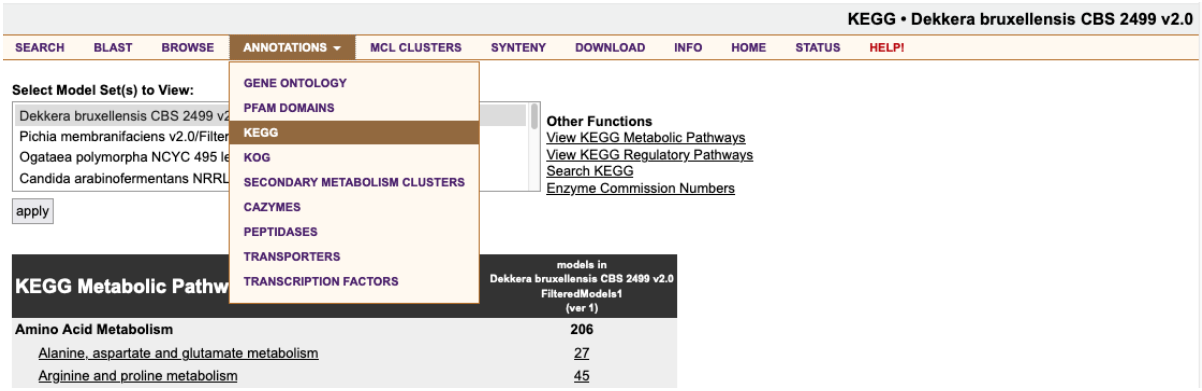
2) To verify that *Dekkera* (which grows on galactose) and *Pichia* (which does not) are sibling taxa, click on '**TREE**':



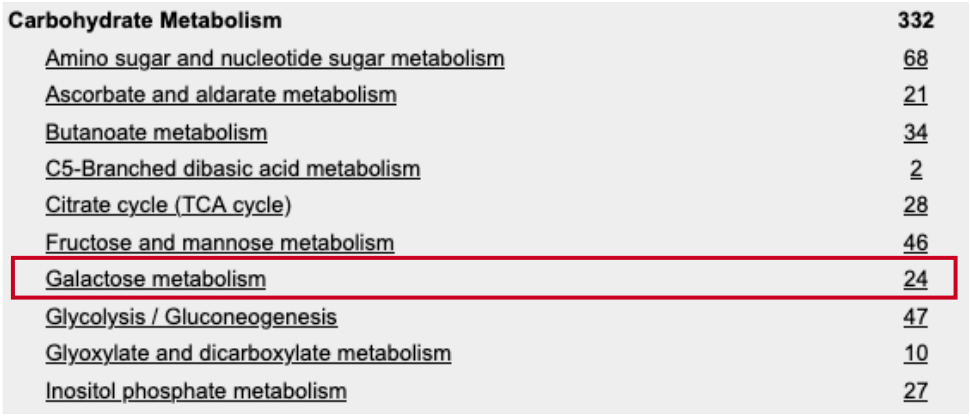
3) Click on '**Dekkera bruxellensis CBS 2499 v2.0**' to go to its genome portal:



4) Click on "**ANNOTATIONS** => **KEGG**" to go to the portal's KEGG browser:

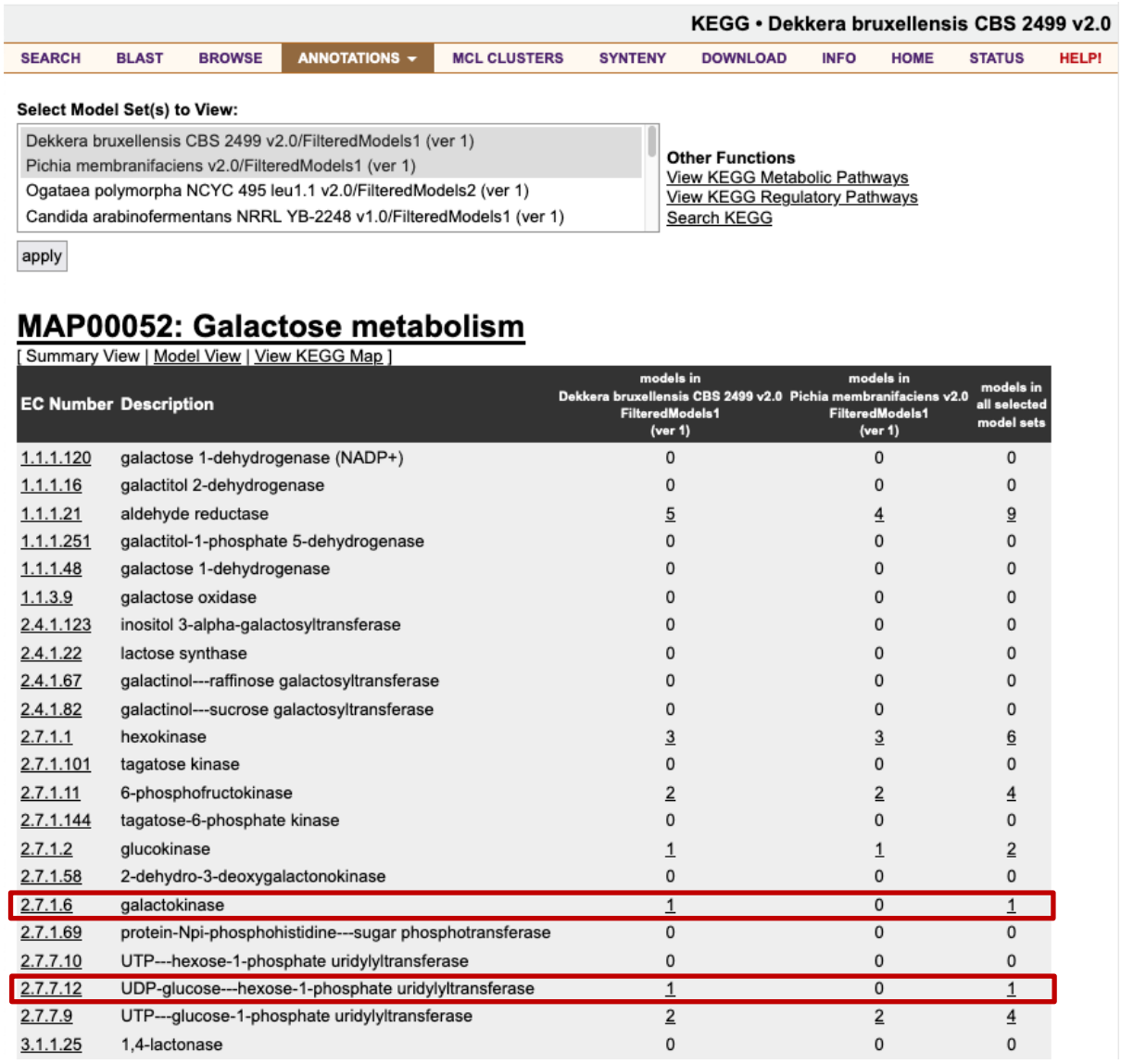


5) Scroll down to the '**Carbohydrate Metabolism**' section, and find the subsection '**Galactose metabolism**'. *Dekkera* has 24 genes annotated to this metabolic pathway:



6) Click on '**Galactose metabolism**' to drill down into the KEGG hierarchy and list the EC numbers associated with that pathway.

7) Go to the '**Select Model Set(s) to View**' list box and select *Dekkera bruxellensis* and *Pichia membranifaciens* and click the '**apply**' button. The *Dekkera* and *Pichia* galactose metabolism gene counts are side-by-side and may be directly compared. Galactokinase (EC = 2.7.1.6) and UDP-glucose--hexose-1-phosphate uridylyltransferase (2.7.7.12) are each present in *Dekkera* but not in *Pichia*:



KEGG • Dekkera bruxellensis CBS 2499 v2.0

SEARCH | BLAST | BROWSE | ANNOTATIONS | MCL CLUSTERS | SYNTENY | DOWNLOAD | INFO | HOME | STATUS | HELP!

Select Model Set(s) to View:

Dekkera bruxellensis CBS 2499 v2.0/FilteredModels1 (ver 1)
Pichia membranifaciens v2.0/FilteredModels1 (ver 1)
Ogataea polymorpha NCYC 495 leu1.1 v2.0/FilteredModels2 (ver 1)
Candida arabinofermentans NRRL YB-2248 v1.0/FilteredModels1 (ver 1)

Other Functions
View KEGG Metabolic Pathways
View KEGG Regulatory Pathways
Search KEGG

apply

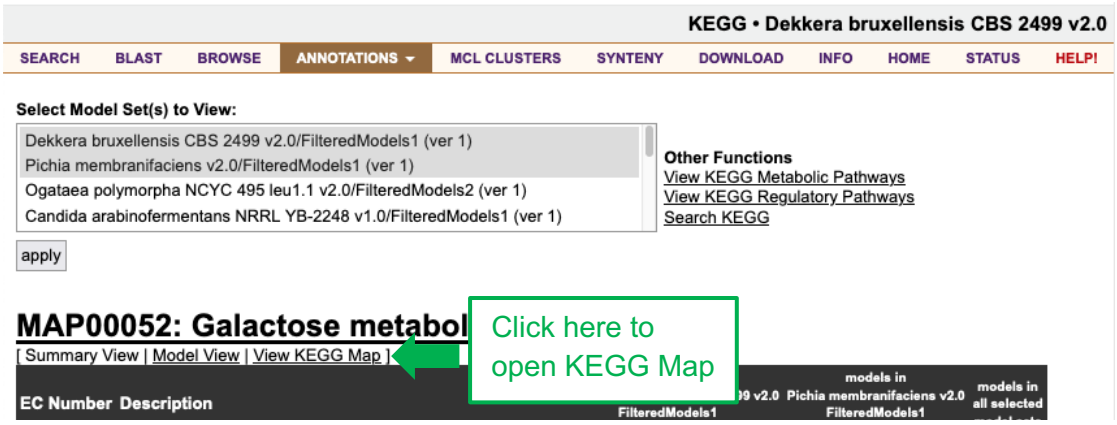
MAP00052: Galactose metabolism

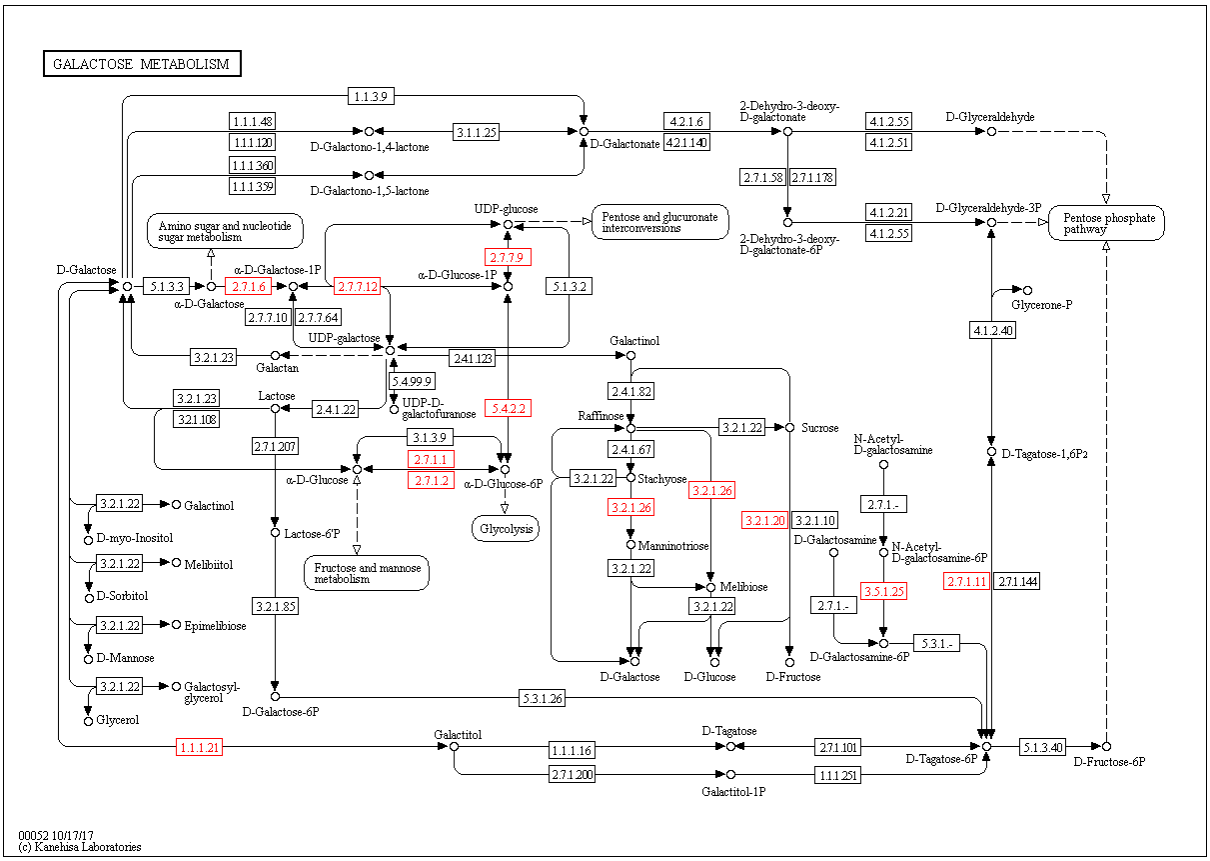
[Summary View | Model View | View KEGG Map]

EC Number	Description	models in Dekkera bruxellensis CBS 2499 v2.0 FilteredModels1 (ver 1)	models in Pichia membranifaciens v2.0 FilteredModels1 (ver 1)	models in all selected model sets
1.1.1.120	galactose 1-dehydrogenase (NADP+)	0	0	0
1.1.1.16	galactitol 2-dehydrogenase	0	0	0
1.1.1.21	aldehyde reductase	5	4	9
1.1.1.251	galactitol-1-phosphate 5-dehydrogenase	0	0	0
1.1.1.48	galactose 1-dehydrogenase	0	0	0
1.1.3.9	galactose oxidase	0	0	0
2.4.1.123	inositol 3-alpha-galactosyltransferase	0	0	0
2.4.1.22	lactose synthase	0	0	0
2.4.1.67	galactinol---raffinose galactosyltransferase	0	0	0
2.4.1.82	galactinol---sucrose galactosyltransferase	0	0	0
2.7.1.1	hexokinase	3	3	6
2.7.1.101	tagatose kinase	0	0	0
2.7.1.11	6-phosphofructokinase	2	2	4
2.7.1.144	tagatose-6-phosphate kinase	0	0	0
2.7.1.2	glucokinase	1	1	2
2.7.1.58	2-dehydro-3-deoxygalactonokinase	0	0	0
2.7.1.6	galactokinase	1	0	1
2.7.1.69	protein-Npi-phosphohistidine---sugar phosphotransferase	0	0	0
2.7.7.10	UTP---hexose-1-phosphate uridylyltransferase	0	0	0
2.7.7.12	UDP-glucose---hexose-1-phosphate uridylyltransferase	1	0	1
2.7.7.9	UTP---glucose-1-phosphate uridylyltransferase	2	2	4
3.1.1.25	1,4-lactonase	0	0	0

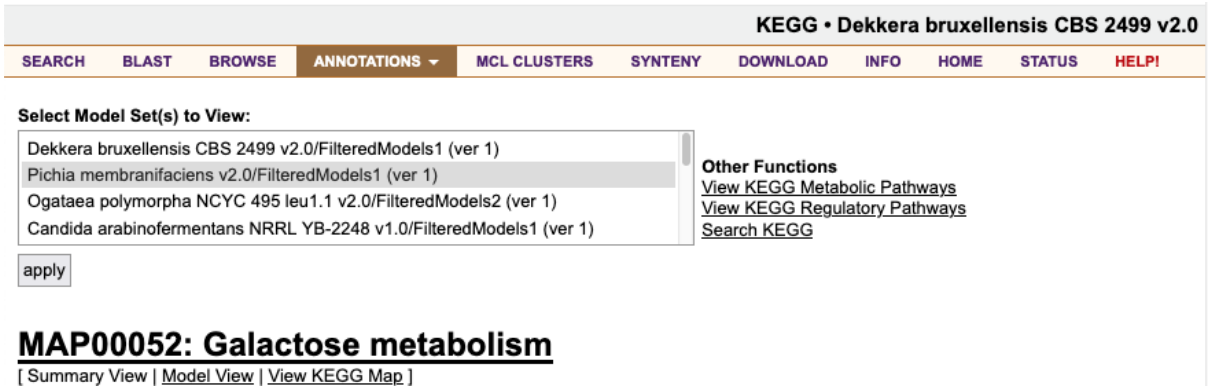
8) Scroll back up to the '**Select Model Set(s) to View**' list box and select *Dekkera bruxellensis* only. Click '**apply**' to show the *Dekkera* counts only.

9) Click '**View KEGG Map**' to see a graphical display of the pathway. Only those enzyme boxes colored red are annotated as such in *Dekkera*. These include both 2.7.1.6(Galactokinase) and 2.7.7.12 (UDP-glucose--hexose-1-phosphate uridylyltransferase):

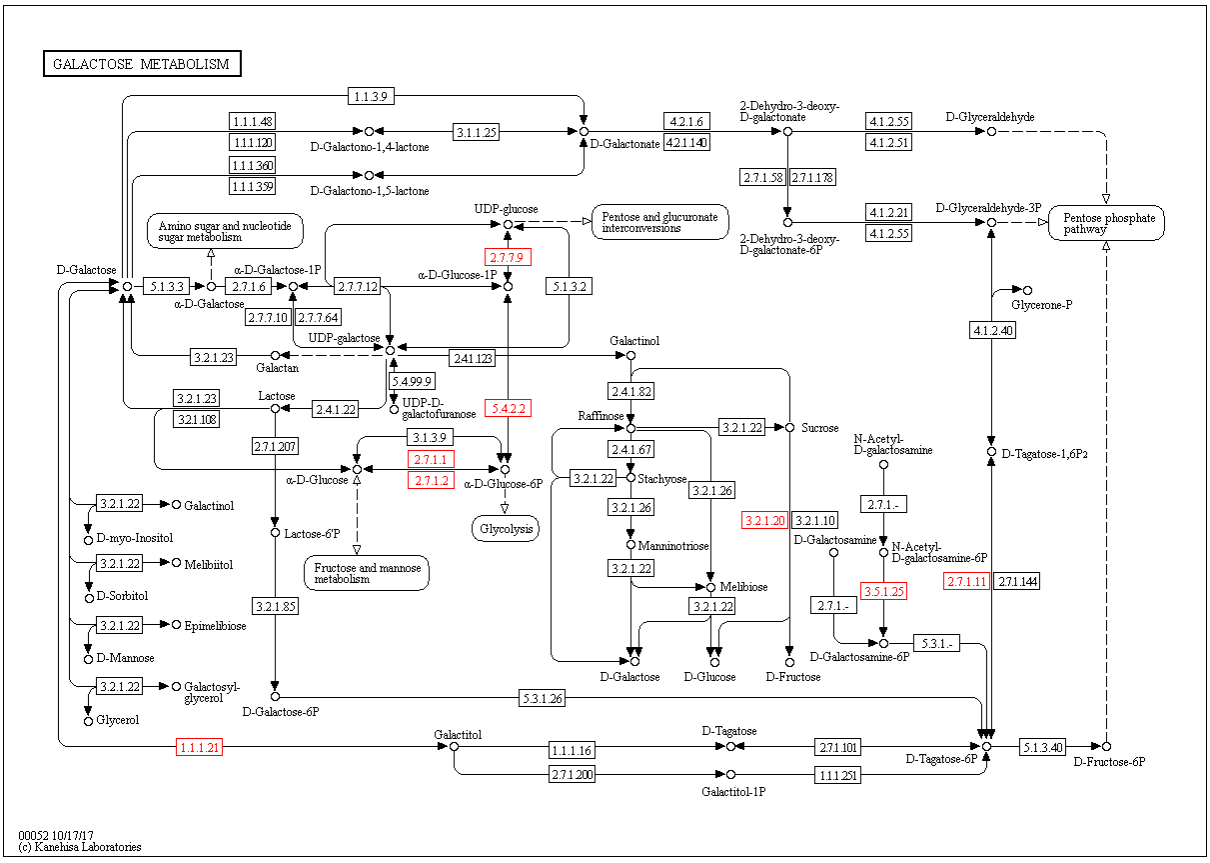




10) Use the web browser back button return to the *Dekkera* galactose metabolism page and select *Pichia* only. Click '**apply**' to show the *Pichia* counts only.



11) Click '**View KEGG Map**' again, and again only those enzyme boxes colored in red are annotated as such in *Pichia*. These include neither 2.7.1.6 nor 2.7.7.12. No wonder *Pichia* cannot grow on galactose – it is missing the genes coding for key enzymes in the galactose utilization pathway.



Exercise:

Based on the KEGG annotations, can you predict whether *Ogataea polymorpha*, *Saccharomyces cerevisiae*, and *Nadsonia fulvescens* can grow on galactose?

Reference:

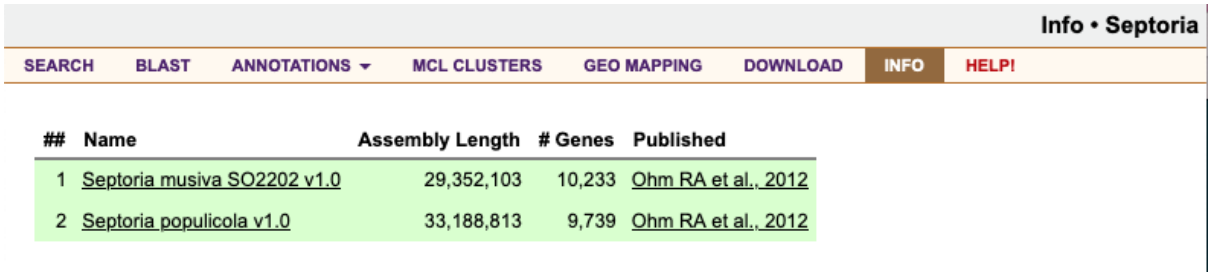
- Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Göker M, Salamov AA, Wisecaver JH, Long TM, Calvey CH, Aerts AL, Barry KW, Choi C, Clum A, Coughlan AY, Deshpande S, Douglass AP, Hanson SJ, Klenk HP, LaButti KM, Lapidus A, Lindquist EA, Lipzen AM, Meier-Kolthoff JP, Ohm RA, Otillar RP, Pangilinan JL, Peng Y, Rokas A, Rosa CA, Scheuner C, Sibirny AA, Slot JC, Stielow JB, Sun H, Kurtzman CP, Blackwell M, Grigoriev IV, Jeffries TW. Comparative genomics of biotechnologically important yeasts. Proc Natl Acad Sci U S A. 2016 Aug 30;113(35):9882-7. doi: 10.1073/pnas.1603941113. Epub 2016 Aug 17. PubMed PMID: 27535936; PubMed Central PMCID: PMC5024638.

MycoCosm: Secondary Metabolism Clusters Browser

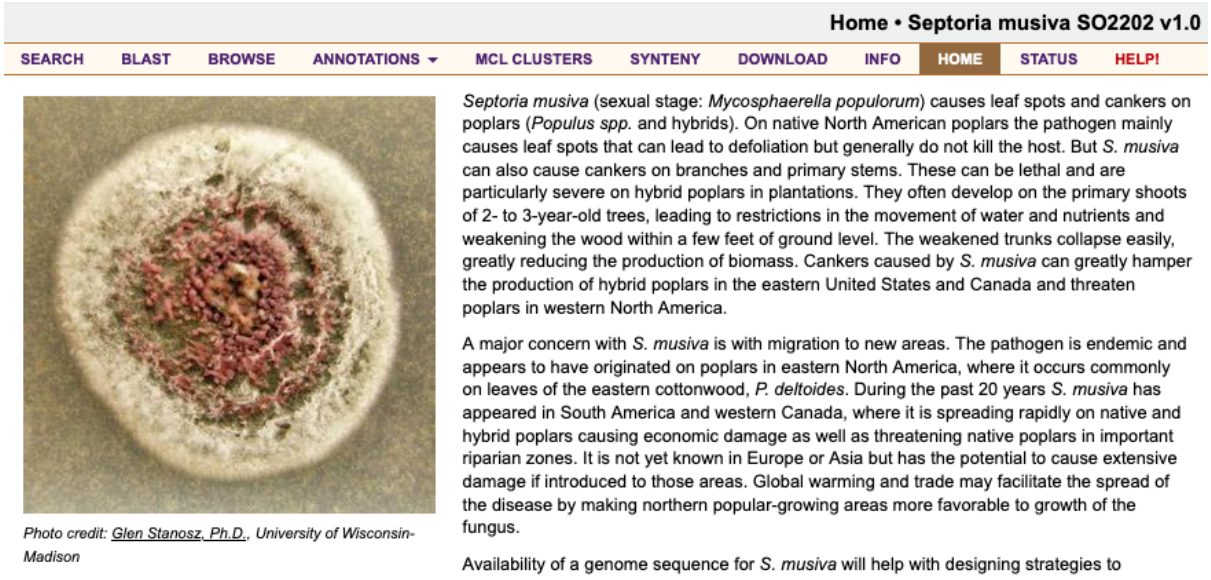
In fungi, secondary metabolite (SM) genes are often organized in chromosomal clusters dedicated to that metabolite's biosynthetic pathway. Each portal's SM Clusters Browser facilitates display and discovery of MycoCosm's SM-annotated genes.

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.

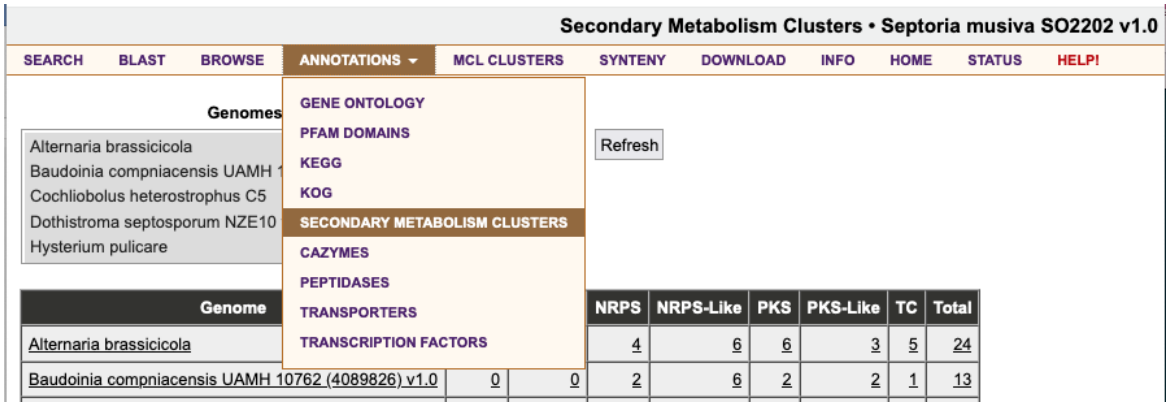
1) Go to the MycoCosm Septoria PhyloGroup at mycocosm.jgi.doe.gov/Septoria. Both species are represented in the group:



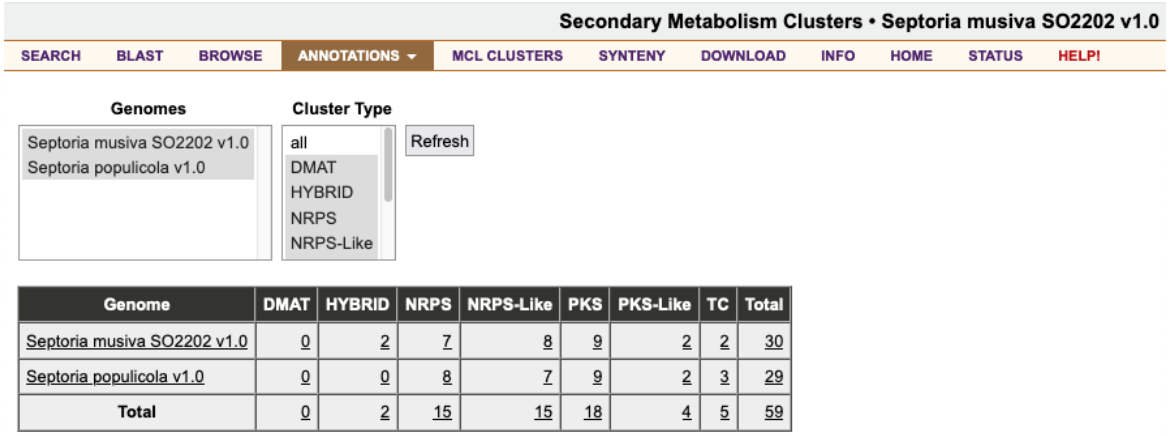
2) Click on '***Septoria musiva* SO2202 v1.0**' to go to its genome portal:



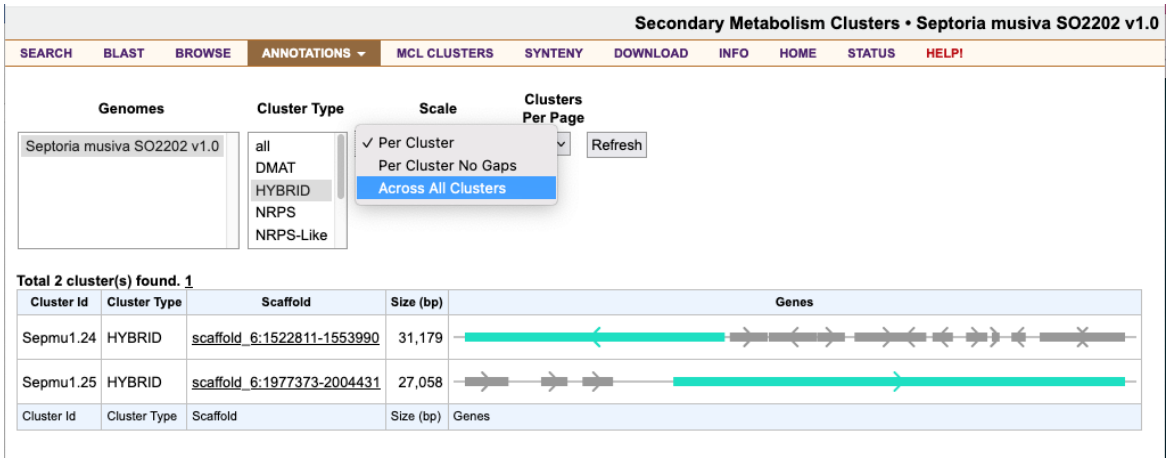
3) Click on "**ANNOTATIONS => SECONDARY METABOLISM CLUSTERS**" to go to the portal's SM clusters browser:



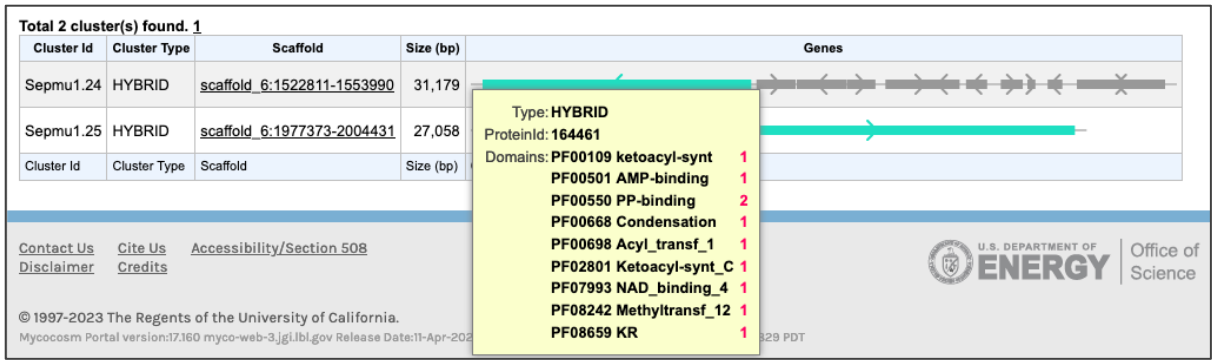
4) Scroll through the 'Genomes' list box and select both '*Septoria musiva*' and '*Septoria populica*', and only those 2 species. Click the '**Refresh**' button. Only the SM cluster core gene counts of the 2 *Septoria* sp. are shown, and may be directly compared. *S. musiva* has 2 hybrid core genes (PKS-NRPS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



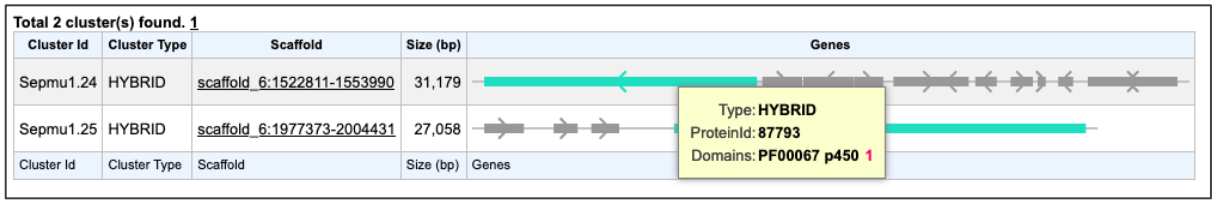
5) There is a total of 2 genes in the Hybrid column. Click on the number to show a graphical representation of the 2 *S. musiva* gene clusters. The 'Size' column displays each cluster's length, and the 'Genes' column displays each cluster's core PKS-NRPS gene (in color) and its accessory, decorator,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



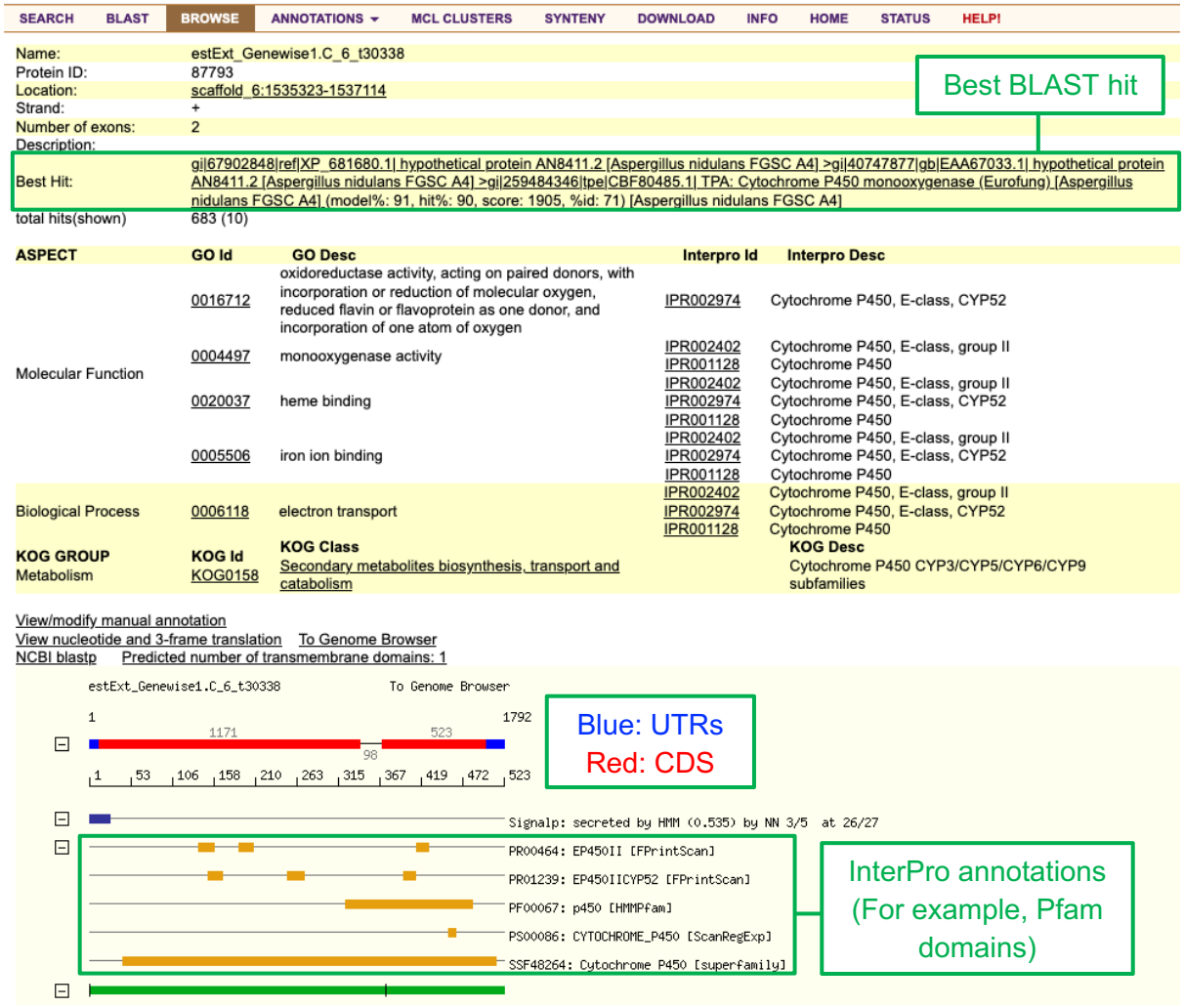
6) Each gene in the clusters is represented by an arrow with a single pair of fletching that indicates the gene's 5' to 3' direction. Mouse-over the top cluster's core gene to get more information about the PKS-NRPS hybrid. The listed domains are typical of a hybrid enzyme:



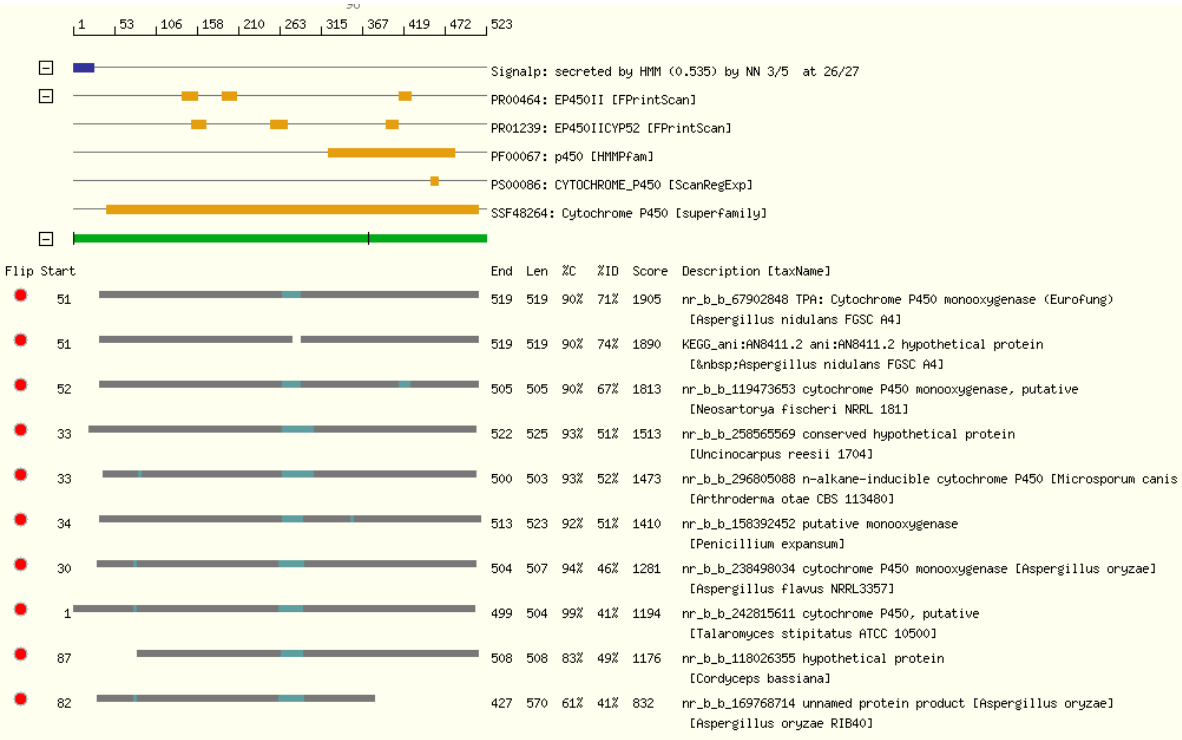
7) To get domain information about the other genes in the SM cluster, mouse-over them too. The next gene 3' to the core gene has a p450 domain:



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: Secondary Metabolites and clusters

Learning objectives:

- Explore InterPro search in FungiDB
- Cross-reference the results with MycoCosm data

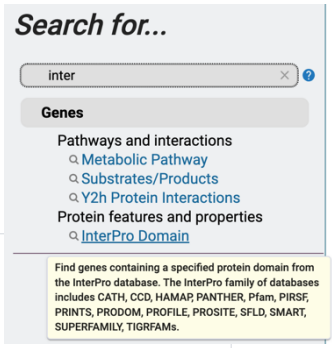
• Finding secondary metabolites and gene clusters

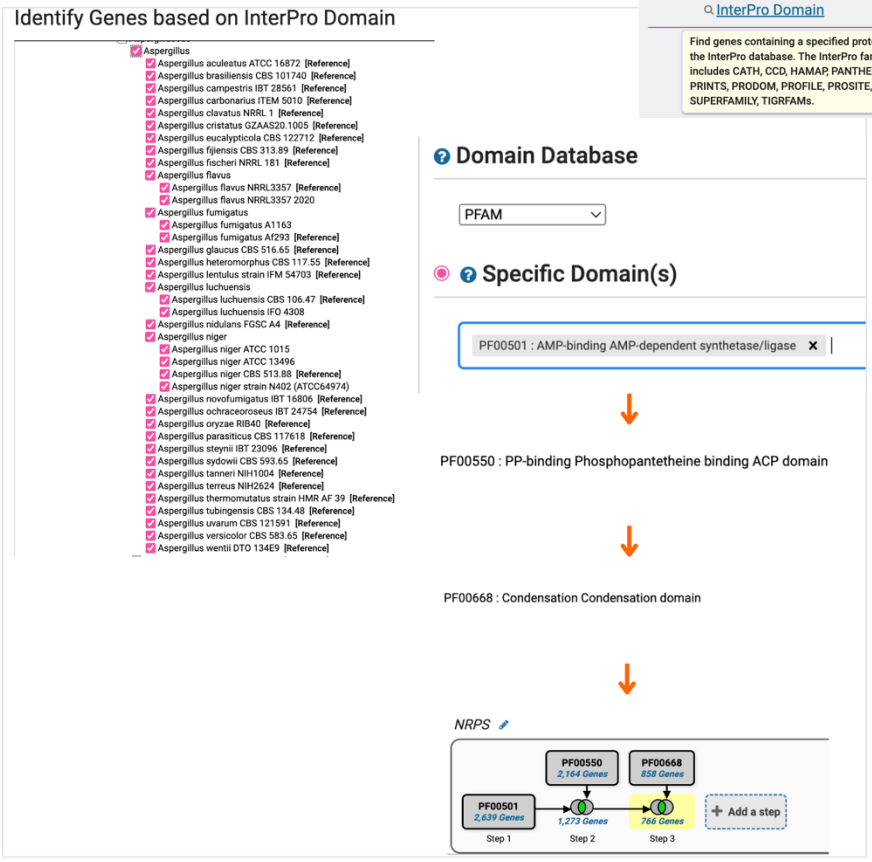
Fungi produce a plethora of secondary metabolites. The secondary metabolites can be segregated into groups based on the first step of their biosynthesis, more specifically, the "key enzymes" that are required: Non-ribosomal peptide synthetases (NRPSs), NRPS-like, Polyketide synthases (PKSs), PKS-like, Hybrid PKS – NRPS, Prenyltransferases (DMAT), Terpene cyclases/synthase (TC).

1. Use the InterPro search to identify NRPS genes in all *Aspergilli*.

NRPS genes have at least the three domains:

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

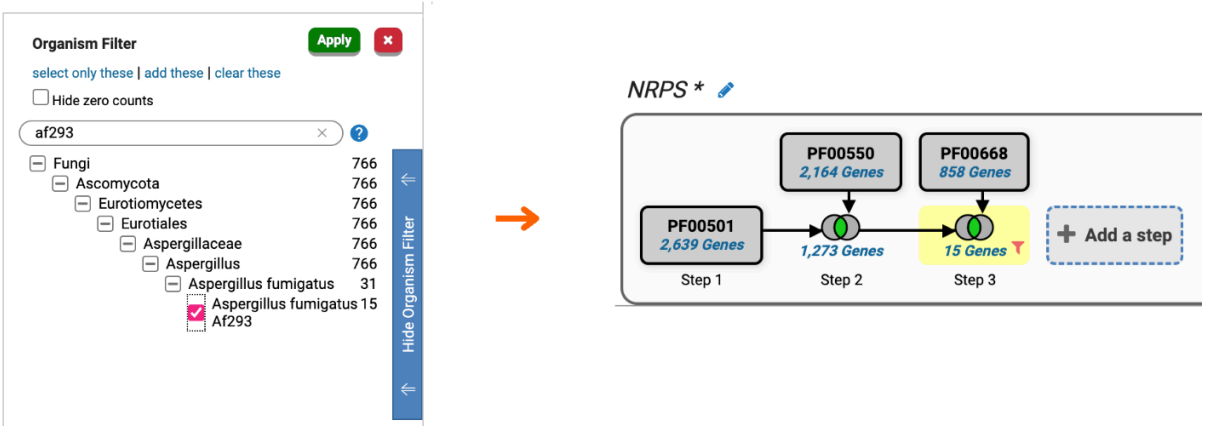




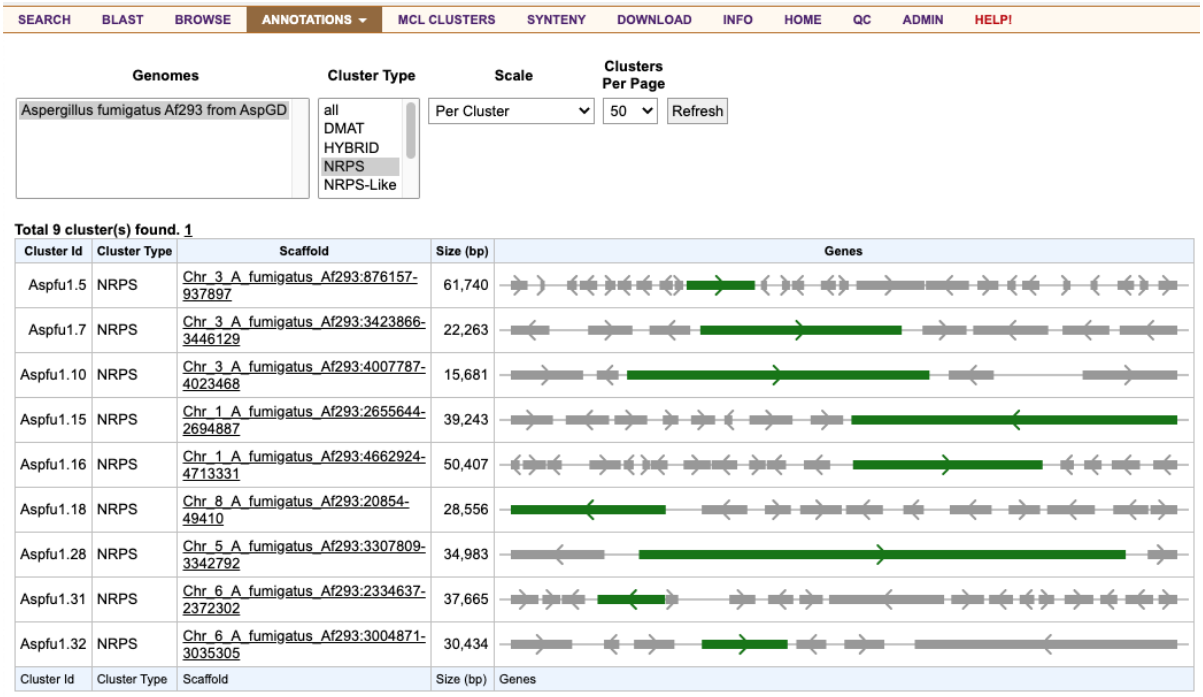
Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/85a1e3a5a603efc6

- 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.



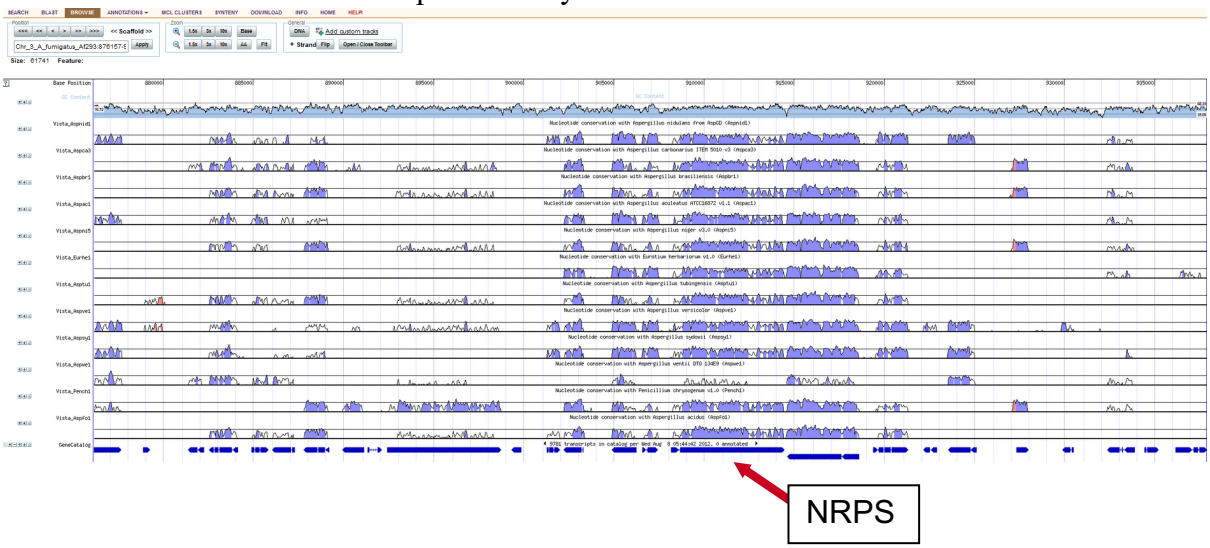
- Create a search for NRPS genes in MycoCosm. Access the *A. fumigatus* Af293 portal (https://mycocosm.jgi.doe.gov/Aspfu1) and navigate to the Secondary Metabolism Clusters page (under the 'Annotations' tab). How many genes did you get?



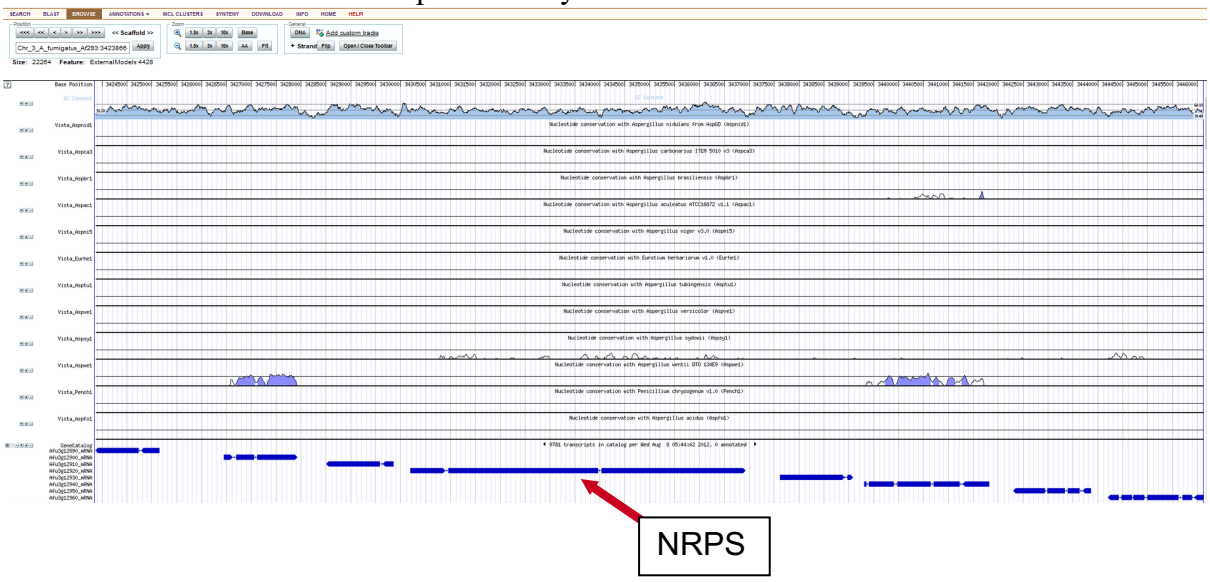
- What do you think may be causing the difference in the predicted gene number?
- This view on MycoCosm allows you to analyze backbone and auxiliary proteins across the entire predicted secondary metabolism cluster. How conserved are these secondary metabolite clusters across related Aspergilli? Click 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? 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 at a couple 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:



Reference: PMID:24692239