

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):

A screenshot of the search results page for 'kinase'. The title is 'All results matching kinase'. It shows 1 - 20 of 325,297 results. A blue button in the top right corner says 'Export as a Search Strategy'. On the left, a 'Filter results' sidebar has a red arrow pointing to the 'Genes' link under 'Population biology'. The main area lists several gene entries with their details and matching fields. A blue button at the bottom right says 'Export as a Search Strategy'.

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 single category only (e.g. Genes or Data sets, but not both).

Gene ID	Transcript ID	Organism	Genomic Location	Product Description
MRET_0047	MRET_0047-i46_1	Malassezia restricta KCTC 27527	CP030251:95,680..97,545(+)	triose/dihydroxyacetone kinase/FAD-AMP lyase (cyclizing)
MRET_0094	MRET_0094-i46_1	Malassezia restricta KCTC 27527	CP030251:170,464..171,498(+)	adenosine kinase
MRET_0098	MRET_0098-i46_1	Malassezia restricta KCTC 27527	CP030251:179,095..181,227(-)	aarF domain kinase
MRET_0099	MRET_0099-i46_1	Malassezia restricta KCTC 27527	CP030251:181,386..181,844(+)	nucleoside-diphosphate kinase
MRET_0136	MRET_0136-i46_1	Malassezia restricta KCTC 27527	CP030251:231,306..233,297(+)	pseudouridylate synthase/pseudouridine kinase
MRET_0167	MRET_0167-i46_1	Malassezia restricta KCTC 27527	CP030251:270,552..272,270(-)	meiosis induction protein kinase IME2/SME1
MRET_0178	MRET_0178-i46_1	Malassezia restricta KCTC 27527	CP030251:288,959..289,339(+)	tyrosine-protein kinase sms
MRET_0205	MRET_0205-i46_1	Malassezia restricta KCTC 27527	CP030251:331,297..333,045(-)	type II pantothenate kinase

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

All results matching *

1 - 20 of 4,901,548

Hide zero counts

Filter results

Genome	1,885,291
Genes	162,441
Genomic sequences	
Organism	186
Organisms	
Transcriptomics	
ESTs	1,709,817
Population biology	
Popset isolate sequences	1,077,920
Metabolism	
Metabolic pathways	3,045
Compounds	61,998
Data access	
Data sets	381
Searches	435
Instructional Tutorials	15
Workshop exercises	1
About	2
News	
General info pages	16

Filter fields
Select a result filter above

Filter organisms
select all | clear all | expand all | collapse all

Type a taxonomic name

Hide zero counts

Export as a Search Strategy to download or mine your results

Compound - CHEBI:10000 Vismone D
Compound - CHEBI:10001 Visanadin
Compound - CHEBI:10002 Visnagin
Compound - CHEBI:10003 ribostamycin sulfate
Definition: An aminoglycoside sulfate salt resulting from the reaction of ribostamycin with sulfuric acid.

Compound - CHEBI:10014 nalidixic acid
Definition: A monocarboxylic acid comprising 1,8-naphthyridin-4-one substituted by carboxylic acid, ethyl and methyl groups at positions 3, 1, and 7, respectively.

Compound - CHEBI:10015 visnagine
Definition: An indole alkaloid that is visanadin in which the bridgehead methyl group is substituted by a methoxycarbonyl group and an additional oxo substituent is present in the 3-position.

Compound - CHEBI:10016 volutusine
Compound - CHEBI:10017 volenitol
Definition: A heptol that is heptane-1,2,3,4,5,6,7-heptol that has R-configuration at positions 2, 3, 5 and 6.

Compound - CHEBI:10018 volkenin
Definition: A cyanogenic glycoside that is (4R)-4-hydroxycyclopent-2-ene-1-carbonitrile attached to a beta-D-glucopyranosyloxy at position 1.

Compound - CHEBI:10019 Vornicine
Compound - CHEBI:10022 Vomitoxin
Compound - CHEBI:10024 voriconazole
Definition: A triazole-based antifungal agent used for the treatment of esophageal candidiasis, invasive pulmonary aspergillosis, and serious fungal infections caused by *Scedosporium apiospermum* and *Fusarium* spp. It is an inhibitor of cytochrome P450 2C9 (CYP2C9) and CYP3A4.

Compound - CHEBI:100241 ciprofloxacin
Definition: A quinolone that is quinolin-4(1H)-one bearing cyclopropyl, carboxylic acid, fluoro and piperazin-1-yl substituents at positions 1, 3, 6 and 7, respectively.

COMMUNITY CHAT

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

The gene id search will return the gene record card for [Afu2g13260](#) (see screenshot below). Click on the gene link in blue to navigate to the gene record page for this gene.

Genes matching Afu2g13260

1 - 1 of 1

Hide zero counts

Filter results

Genome	1
Genes	1

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

Filter organisms
select all | clear all | expand all | collapse all

Type a taxonomic name Reference only

Fungi
 Ascomycota
1

Gene - Afu2g13260 Developmental regulator medA, putative
Gene name or symbol: medA
Gene type: protein coding gene
Organism: *Aspergillus fumigatus* Af293
▶ Fields matched: External links; Gene ID; Names, IDs, and aliases; User comments

Gene - Afu2g13260 Developmental regulator medA, putative
Gene name or symbol: medA
Gene type: protein coding gene
Organism: *Aspergillus fumigatus* Af293
▶ Fields matched: External links; Gene ID; Names, IDs, and aliases; User comments

1 - 1 of 1

Export as a Search Strategy to download or mine your results

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.

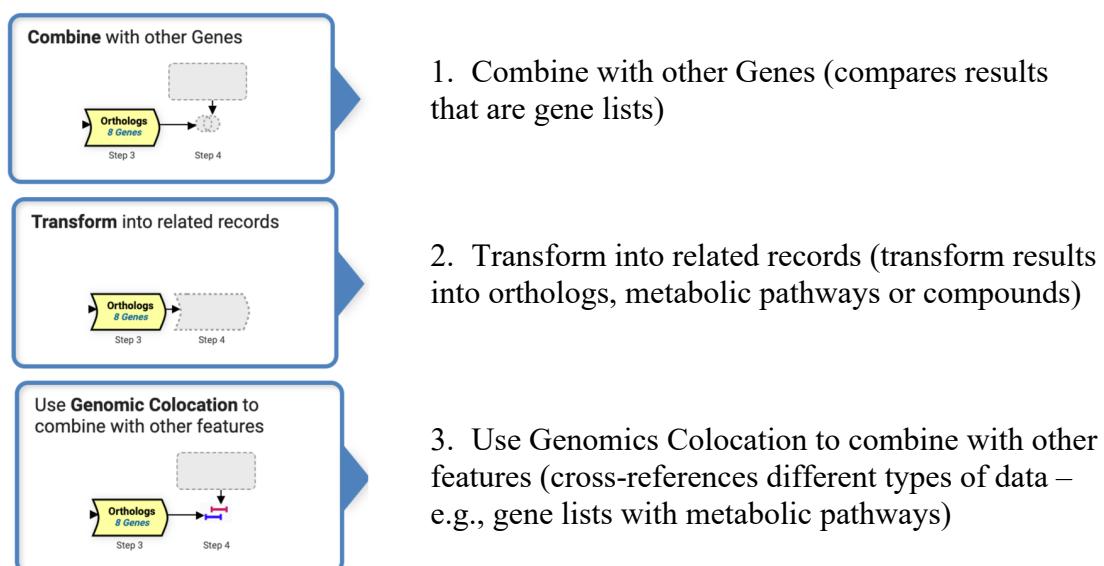
Advanced Search Strategies

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:



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.

Revise as a boolean operation	
<input checked="" type="radio"/> 1 INTERSECT 2	<input type="radio"/> 1 UNION 2
<input type="radio"/> 1 MINUS 2	<input type="radio"/> 2 MINUS 1
Revise as a span operation	
<input type="radio"/> 1 RELATIVE TO 2, using genomic colocation	
Ignore one of the inputs	
<input type="radio"/> IGNORE 2	<input type="radio"/> IGNORE 1
<input type="button" value="Revise"/>	

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 in 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 27527 genes into their orthologs in *Malassezia restricta* CBS 7877 and *Candida albicans* SC5314.

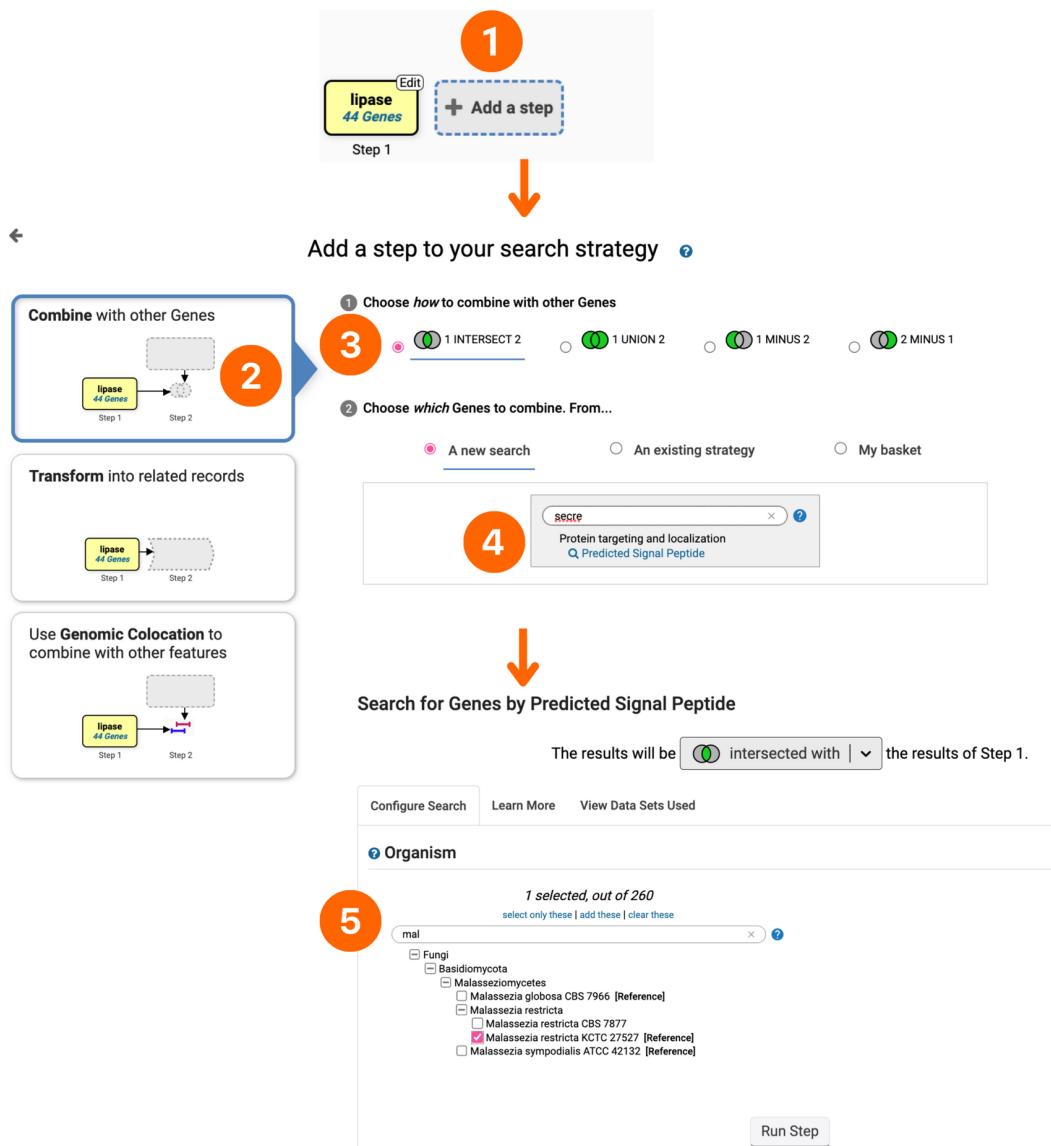
- Use site search to identify genes that have “lipase” annotation in *Malassezia restricta* KCTC 27527

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.

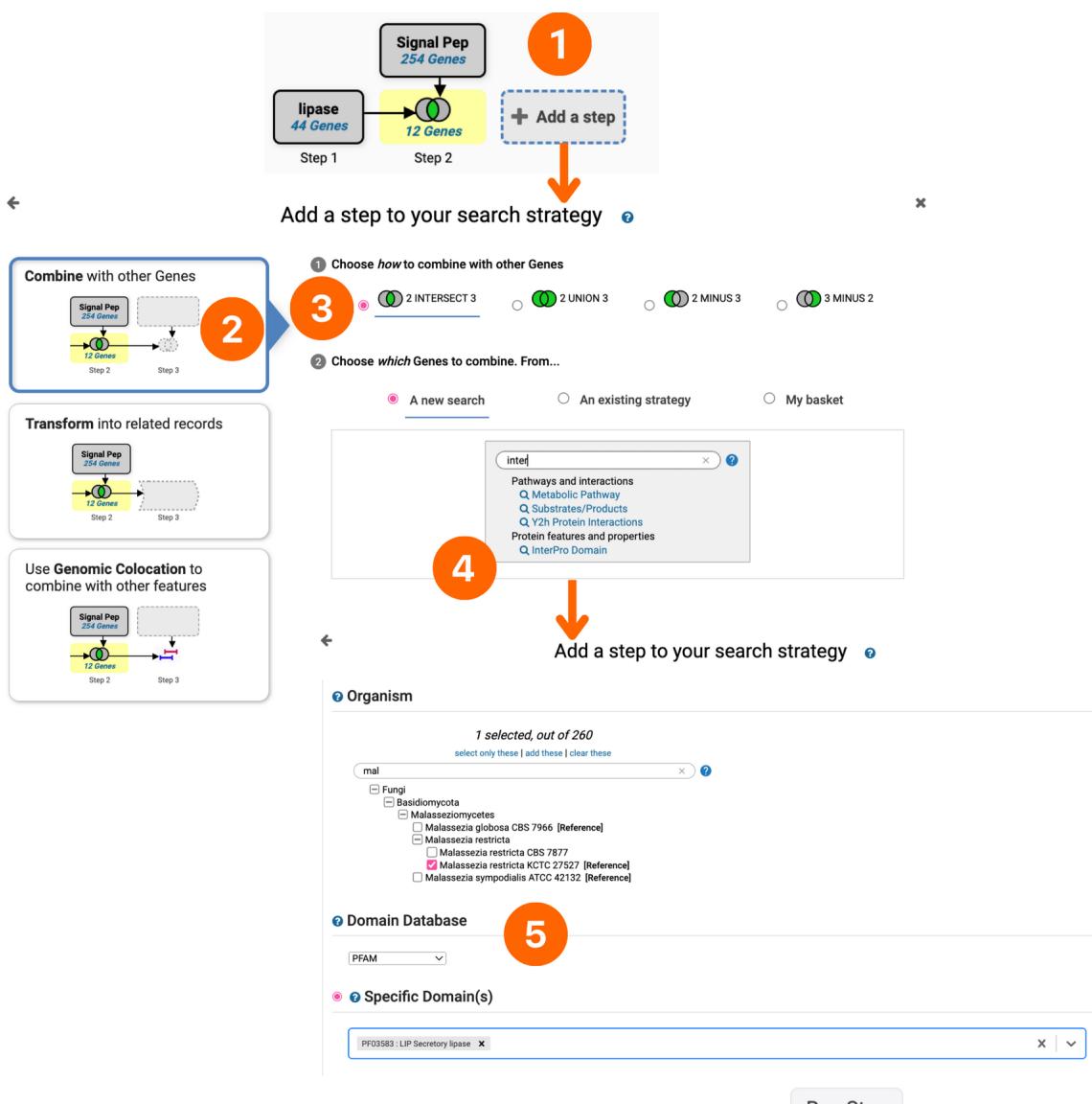
The screenshot shows a search interface with the following components:

- Search Bar:** The word "lipase" is typed into the search bar.
- Header:** My Strategies, Searches, Tools, My Workspace, Data, About, Help, Contact Us.
- Section 1 (Filter results):** A button labeled "1" is highlighted. It includes a checkbox for "Hide zero counts" and a "Clear filter" button with the number 44.
- Section 2 (Filter Gene fields):** A button labeled "2" is highlighted. It lists various gene field filters with their counts: EC descriptions and numbers (19), GO terms (4), InterPro domains (21), Notes from annotators (0), Orthologs (37), Product descriptions (14), Phenotype (0), Preferred product description (17), Product descriptions (17), and User comments (0). There is also a "Clear filter" button.
- Section 3 (Filter organisms):** A button labeled "3" is highlighted. It shows a tree view of organisms under "malassezia": Fungi > Basidiomycota > Malasseziomycetes > Malassezia. Specific entries include: Malassezia globosa CBS 7966 [Ref] (181), Malassezia restricta (87), Malassezia restricta CBS 7877 (43), Malassezia restricta KCTC 27527 [Ref] (44), and Malassezia sympodialis ATCC 42132 [Ref] (47). There is a "select only these" dropdown and a "Clear these" button. A "Reference only" checkbox is also present.
- Results List:** The main area displays a list of genes matching the filters. Each entry includes the gene name, type, organism, and a detailed description of the fields matched. For example:
 - Gene - MRET_0019 lipase: Gene type: protein coding gene; Organism: Malassezia restricta KCTC 27527
 - Gene - MRET_1032 lipase: Gene type: protein coding gene; Organism: Malassezia restricta KCTC 27527
 - Gene - MRET_4032 lipase: Gene type: protein coding gene; Organism: Malassezia restricta KCTC 27527
 - Gene - MRET_4356 lipase: Gene type: protein coding gene; Organism: Malassezia restricta KCTC 27527
 - Gene - MRET_0923 acylglycerol lipase: Gene type: protein coding gene; Organism: Malassezia restricta KCTC 27527
 - Gene - MRET_0930 secretary lipase: Gene type: protein coding gene; Organism: Malassezia restricta KCTC 27527
- Export Button:** A blue button labeled "4" and "Export as a Search Strategy" with a download icon.

- **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.

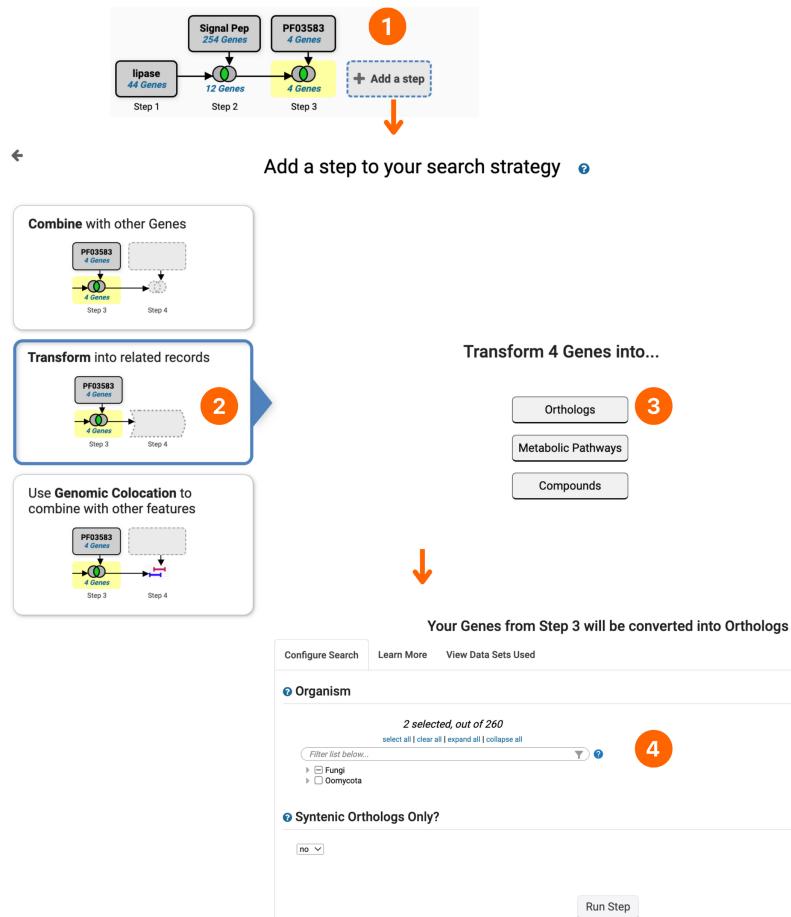


- **Identify Genes based on InterPro domain.** This search identifies genes with specific domain signature – secreted lipase (LIP).
 1. Click on the “Add step” button.
 2. Choose “Combine with other genes” search.
 3. Choose to “intersect” your results with the previous step.
 4. Filter the available 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.



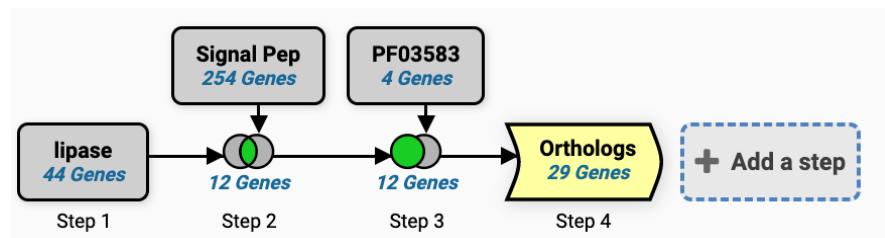
Examine your results. Do they make sense?

The screenshot shows the FungiDB ortholog search interface. At the top, there is a search history panel with four steps:

- Step 1:** lipase (44 Genes)
- Step 2:** Signal Pep (254 Genes) - resulting in 12 Genes
- Step 3:** PF03583 (4 Genes) - resulting in 4 Genes
- Step 4:** Orthologs (16 Genes)

Below this is a results table titled "16 Genes (1 ortholog groups)". The table has columns for Gene ID, Transcript ID, Organism, Product Description, Input Ortholog(s), Ortholog Group, and Paralog count. The results list various genes from *Candida albicans* and *Malassezia globosa*, along with their product descriptions and ortholog information.

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.

The SGD home page features a navigation bar with links for About, Blog, Download, Help, YeastMine, and social media icons. A search bar contains the query "search: actin, kinase, gli". An orange arrow points to the "YeastMine" link in the top right corner. Below the navigation bar is a banner image of yeast cells stained with Rap1-GFP and Calcofluor White. To the right of the banner is a "About SGD" section with a brief description of the database's purpose. At the bottom right is a red "Try this?" button.

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**

The SGD FUNCTION tab is active, indicated by a grey background. A red arrow points to the "FUNCTION" tab label. Another red arrow points to the "GO Term name [and children of this term] → All genes" option in the list of search queries. The SGD logo and "popular templates" watermark are visible in the bottom right corner.

FUNCTION

Read more

Query for function:

- GO Term → All genes
- Gene → GO Terms.
- GO Term name [and children of this term] → All genes
- Gene → Pathways
- Pathway → Genes
- GO ID → Genes.
- GO Term name → GO Term Identifier.
- Literature → GO annotations

» More queries

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

GO Term name [and children of this term] ➔ All genes
Retrieve all genes that are annotated to the specified GO term and children of that specified GO Term. Wild card queries (such as *ascospore*) are supported. Only manually curated and high-throughput GO annotations are included.

GO Term > Name - Show genes annotated with GO term (and any children of this GO term):
mitochondrial ribosome

Show Results | Edit Query

web service URL | Perl | Python | Ruby | Java [help] | export XML

- 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)

Showing rows 1 to 25 of 110 Rows per page: 25

Gene Primary DBID	Gene Systematic Name	Gene Standard Name	Gene Feature Type	Gene Qualifier	GO Annotation Ontology Term . Identifier	GO Annotation Ontology Term . Name	GO Annotation Ontology Term . Namespace	Code With Text	Code Qualifier	GO Annotation Extension	Code Annot Type	Parents Identifier	Parents Name
S000000134	YBL038W	MRPL16	ORF	Verified	GO:0005762	mitochondrial large ribosomal subunit	cellular_component	IDA	NO VALUE	NO VALUE	manually curated	GO:0005761	mitochondrial ribosome
S000000134	YBL038W	MRPL16	ORF	Verified	GO:0005762	mitochondrial large ribosomal subunit	cellular_component	IDA	NO VALUE	NO VALUE	manually curated	GO:0005761	mitochondrial ribosome
S000000186	YBL090W	MRP21	ORF	Verified	GO:0005763	mitochondrial small ribosomal subunit	cellular_component	IDA	NO VALUE	NO VALUE	manually curated	GO:0005761	mitochondrial ribosome

Relationships

- Gene (92 Genes)
- Gene > GO Annotation > Ontology Term (3 GO Terms)
- Gene > GO Annotation > Evidence > Code (7 GO Evidence Codes)
- Gene > GO Annotation (110 GO Annotations)
- Gene > GO Annotation > Ontology Term > Parents (1 GO Term)
- Gene > GO Annotation > Evidence > Publications (23 Publications)
- Gene > Organism (1 Organism)

Pick items from the table

Create List | Add to List

Create a new List of 92 Genes

List Name: List 1 MTR proteins

List Description: Enter a description

No TAGS | Add a new tag | add

Close | Create List

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**

GENOME PROTEINS FUNCTION PHENOTYPES INTERACTIONS REGULATION HOMOLOGY DISEASE LITERATURE

[Read more](#)

Query for interactions:

- Gene → Complex + Details
- Gene → Genetic Interactions
- Gene → Physical Interactions
- Literature → Interaction
- Complex → Details + Participants

» [More queries](#)

popular templates

- 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**

Gene Interaction
Retrieve all interactions for a specified gene.

Gene

LOOKUP: act1

constrain to be IN List 1 MTR proteins

[Show Results](#) [Edit Query](#)

[web service URL](#) [Perl | Python | Ruby | Java \[help\]](#) [export XML](#)

- 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**.

specified gene.

Manage Filters Manage Relationships Save as List ▾

Gene (92 Genes)
Gene > Organism (1 Organism)
Gene > Interactions > Details (11,842 Interaction Details)
Gene > Interactions > Participant 2 (3,427 Genes)

Gene > Interactions > Details > Experiment > **Interaction Detection Methods** (23 Interaction Terms)
Gene > Interactions > Details > Experiment (305 Interaction Experiments)

Pick items from the table

[Create List](#) [Add to 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.

The screenshot shows the YeastMine 'Lists' page. At the top, there's a navigation bar with SGD, YeastMine logo, search bar ('Search and retrieve S. cerevisiae data with YeastMine, populated by SGD and powered by InterMine.'), and various links like Home, Templates, Lists, QueryBuilder, Tools, Regions, Data Sources, API, and MyMine. Below the navigation is a search bar ('Search: e.g. act1'). The main content area is titled 'Lists' and contains a table of gene sets. The table has columns for 'List Name', 'Genes', and 'Actions'. The first row, 'List 2 MTR interactors 3427 Genes', is highlighted in purple and has a checked checkbox. The last row, 'Uncharacterized_ORFs 739 Genes', also has a checked checkbox. Arrows point to both of these rows.

List	Genes	Actions
<input checked="" type="checkbox"/> List 2 MTR interactors	3427 Genes	<input type="radio"/> Union <input type="radio"/> Intersect <input type="radio"/> Subtract <input type="radio"/> Asymmetric Difference <input type="checkbox"/> Copy <input type="checkbox"/> Delete <input checked="" type="checkbox"/> Options: Show descriptions <input type="checkbox"/> Show Tags
<input type="checkbox"/> List 1 MTR proteins	92 Genes	
<input type="checkbox"/> All Curated Molecular Complexes	580 Molecular Complexes	
<input type="checkbox"/> ALL_Verified_Uncharacterized_Dubious_ORFs	6604 Genes	
<input type="checkbox"/> Uncharacterized_Verified_ORFs	5915 Genes	
<input type="checkbox"/> Dubious_ORFs	689 Genes	
<input checked="" type="checkbox"/> Uncharacterized_ORFs	739 Genes	<input type="radio"/> Union <input type="radio"/> Intersect <input type="radio"/> Subtract <input type="radio"/> Asymmetric Difference <input type="checkbox"/> Copy <input type="checkbox"/> Delete <input checked="" type="checkbox"/> Options: Show descriptions <input type="checkbox"/> Show Tags
<input type="checkbox"/> Verified_ORFs	5176 Genes	

- 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.

The screenshot shows the same 'Lists' page as before, but now the 'List 3 uncharacterized MTR interactors' entry is highlighted in green, indicating it has been successfully created. The 'Actions' bar at the top is visible, and the overall layout is consistent with the previous screenshot.

- 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

GENOME PROTEINS FUNCTION PHENOTYPES INTERACTIONS REGULATION HOMOLOGY EXPRESSION LITERATURE

[Read more](#)

Query for phenotypes:

- Phenotype ➔ Genes
- Gene ➔ Phenotype
- Literature ➔ Phenotype

» [More queries](#)

popular templates

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

Gene ➔ Phenotype
Retrieve all phenotypes for a specified gene.

Gene
LOOKUP: rad54

constrain to be IN List 3uncharacterized MTR interactors

Show Results **Edit Query**

[web service URL](#) | Perl | Python | Ruby | Java [help] | export XML

- 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).

Showing rows 1 to 25 of 2,471												Rows per page:	25	←	←	←	page 1	→	→	→
Gene Primary DBID	Gene Standard Name	Gene Systematic Name	Gene Sgd Alias	Gene Qualifier	Phenotypes Experiment type	Phenotypes Mutant Type	Phenotypes Observable	Phenotypes Qualifier	Phenotypes Allele	Phenotypes Comment	Phenotypes Strain Background	Phenotypes Chemical								
S000000035	NO VALUE	YAL037W	NO VALUE	Uncharacterized	competitive growth	null	competitive fitness	increased	NO VALUE	NO VALUE	S288c									
S000000035	NO VALUE	YAL037W	NO VALUE	Uncharacterized	heterozygous diploid, competitive growth	null	haploinsufficient	NO VALUE	NO VALUE	NO VALUE	S288c									

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

68 Phenotype Observables

12 Items Selected

respiratory

Phenotype Observable	Count
respiratory growth	12

Filter Download data

- 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.

Showing rows 1 to 2 of 2

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

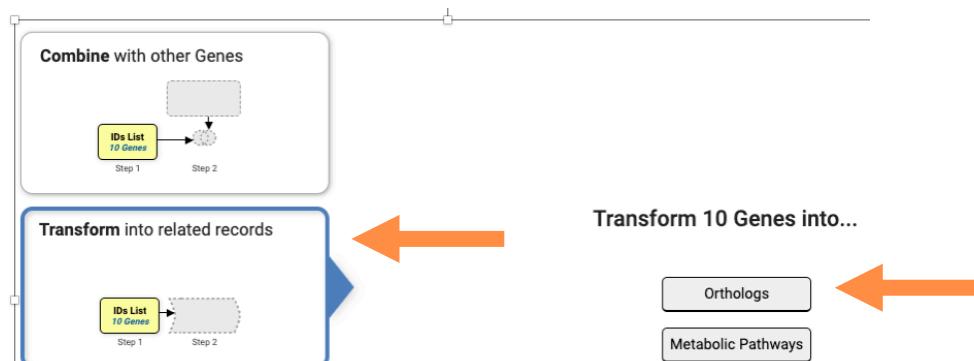
- The results of the above YeastMine analysis suggest 10 genes that potentially encode undiscovered subunits of the mitochondrial ribosome. Although these genes are uncharacterized, more data may exist on their orthologs in other organisms. Use FungiDB to survey the function of orthologs in Fungi and 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)”.

The screenshot shows the FungiDB homepage with the search interface open. The 'Gene ID input set' section contains a text input field with several gene IDs listed: YDR336W, YIL014C-A, YJL193W, YJR079W, and YCL001W-A. Below the input field are three options: 'Upload a text file', 'Copy from My Basket', and 'Copy from My Strategy'. At the bottom right of the search form are two optional fields: 'Give this search a name (optional)' and 'Give this search a weight (optional)'. To the right of the search form, there are links for 'View Sequences and Features in the genome browser' and 'Searches via Web Services'.

- 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.

The screenshot shows the 'Add a step' dialog box. It contains a yellow button labeled 'IDs List 10 Genes' and a blue dashed button labeled '+ Add a step'. The '+ Add a step' button is highlighted with an orange arrow.

- 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.

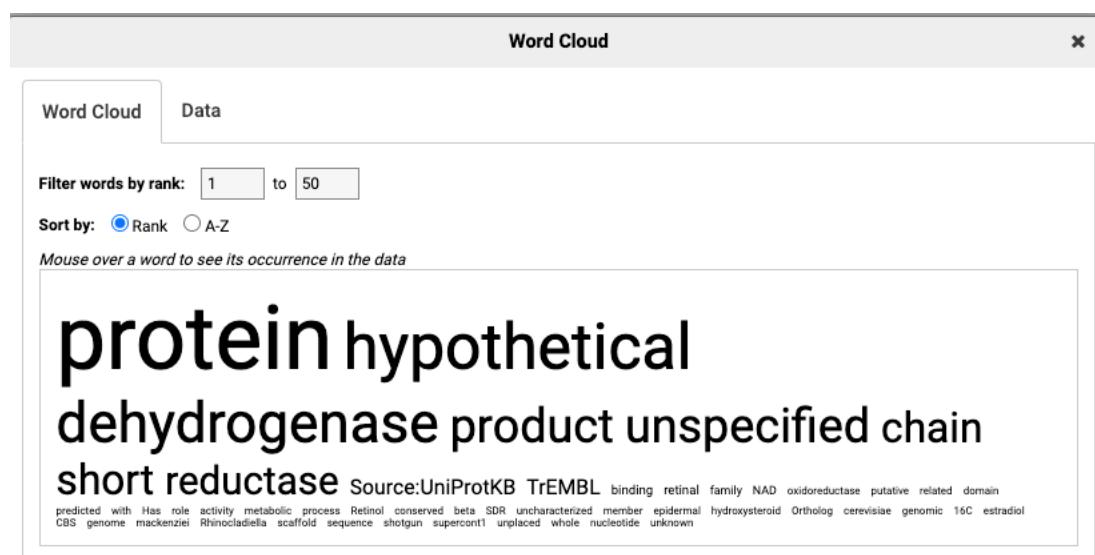
Gene Results Genome View Analyze Results

Genes: 606 Transcripts: 626 Show Only One Transcript Per Gene

First 1 2 3 4 5 Next Last Advanced Paging

Download Add to Basket Add Columns

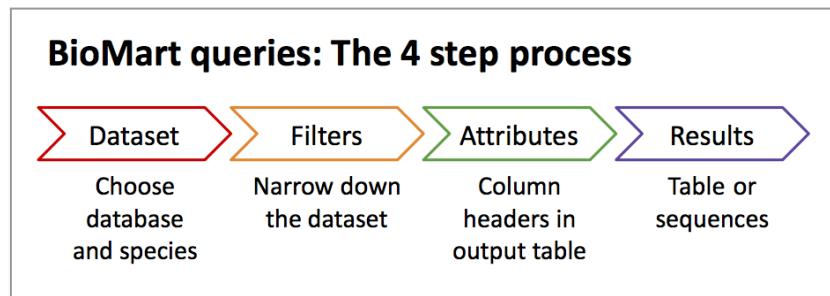
Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Ortholog(s)
ACLA_086280	ACLA_086280-t26_1	<i>A. clavatus</i> NRRL 1	DS027060:2,044,080..2,045,139(+)	GTP binding protein (EngB), putative	YDR336W
AFLA_033930	AFLA_033930-t26_1	<i>A. flavus</i> NRRL3357	EQ963473:3,025,757..3,026,783(-)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

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

The screenshot shows the Ensembl Fungi BioMart homepage. The top navigation bar includes links for BLAST, BioMart, FTP, Docs & FAQs, URL, XML, Perl, and Help. The main interface has tabs for New, Count, and Results. On the left, there's a sidebar with sections for Dataset (set to [None selected]), Filters (set to [None selected]), and Attributes (set to Gene stable ID). The central area shows a dropdown menu for 'Dataset' with options: Ensembl Fungi Genes 56 (selected), Ensembl Fungi Variations 56, and Ensembl Fungi Genes 55.

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

The screenshot shows the Ensembl Fungi BioMart interface after selecting 'Ensembl Fungi Genes 56' in Step 1a. The 'Dataset' section in the sidebar now shows 'Fusarium solani genes (v2.0)'. The central area displays a dropdown menu for 'Dataset' with the same options: Ensembl Fungi Genes 56 (selected), Ensembl Fungi Variations 56, and Ensembl Fungi Genes 55. The other sections in the sidebar remain the same: Filters (None selected) and Attributes (Gene stable ID).

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.

The screenshot shows the EnsemblFungi search interface. On the left, there's a sidebar with 'Dataset' set to 'Fusarium solani genes (v2.0)' and 'Filters' selected. Below that are 'Attributes' for Gene stable ID and Transcript stable ID. The main area has a header 'Please restrict your query using criteria below' and a note '(If filter values are truncated in any lists, hover over the list item to see the full text)'. It contains sections for 'GENE', 'PATHOGEN PHENOTYPES (PHI-BASE)', 'GENE ONTOLOGY', and 'MULTI SPECIES COMPARISONS'. Under 'MULTI SPECIES COMPARISONS', there's a checkbox for 'Homologue filters' and a dropdown menu for 'Paralogous Fusarium solani Genes' with options 'Only' (radio button selected) and 'Excluded'. A callout bubble 'Step 2a: Click on Filters' points to the 'Filters' link in the sidebar. Another callout bubble 'Step 2b: Expand the MULTI SPECIES COMPARISONS section' points to the 'Homologue filters' checkbox.

The screenshot shows the EnsemblFungi search interface after applying filters. The sidebar now shows 'Dataset 6727 / 16163 Genes' and 'Fusarium solani genes (v2.0)'. The 'Filters' section is expanded, showing 'Orthologous Fusarium verticillioides Genes: Excluded' under 'Attributes'. The main area shows the same filter sections as before, but the 'Homologue filters' checkbox is checked. The dropdown menu for 'Paralogous Fusarium solani Genes' now shows 'Orthologous Fusarium verticillioides Genes' with options 'Only' and 'Excluded'. A callout bubble 'Top tip: Click Count to check if your filters work' points to the 'Count' button in the top navigation bar. Another callout bubble 'Step 2c: Choose Orthologous Fusarium verticillioides Genes' points to the 'Excluded' option in the dropdown. A final callout bubble 'Step 2d: Choose the Excluded option' points to the 'Excluded' radio button in the dropdown menu.

The screenshot shows the EnsemblFungi BioMart interface. On the left, there's a sidebar with 'Dataset 4 / 16163 Genes' and 'Fusarium solani genes (v2.0)'. Under 'Filters', it says 'Orthologous Fusarium verticillioides Genes: Excluded Chromosome/scaffold: 14'. Under 'Attributes', it lists 'Gene stable ID' and 'Transcript stable ID'. The main area has a heading 'Update Count' with a note: 'Please restrict your query using criteria below (If filter values are truncated in any lists, hover over the list item to see the full text)'. A dropdown menu 'REGION:' is open, showing 'Chromosome/scaffold' selected. A callout box labeled 'Step 2e: Expand the REGION section' points to this dropdown. Below it, a scrollable list shows chromosomes 1 through 17, followed by several 'sca_xxx_unmapped' entries. At the bottom, there are 'Coordinates' checkboxes for 'Start' and 'End', with the values '1128520' and '1142558' entered into their respective fields. A callout box labeled 'Step 2f: Enter Start / End coordinates' points to these input fields.

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).

Dataset 4 / 16163 Genes
Fusarium solani genes (v2.0)

Filters
Orthologous Fusarium verticillioides Genes: Excluded Chromosome/scaffold: 14
Start: 1128520
End: 1142558

Attributes
Gene start (bp)
Gene end (bp)
Gene name

Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

Step 3b: In the 'Features' category, expand the GENE section

Ensembl

- Gene stable ID
- Transcript stable ID
- Protein stable ID
- Exon stable ID
- Gene description
- Chromosome/scaffold name
- Gene start (bp)
- Gene end (bp)
- Strand
- Karyotype band
- Transcript start (bp)
- Transcript end (bp)

Transcription start site (TSS)

Transcript length (including UTRs and CDS)

Ensembl Canonical

Gene name

Source of gene name

Transcript count

Gene % GC content

Gene type

Transcript type

Source (gene)

Source (transcript)

Gene Synonym

EXTERNAL:

PROTEIN DOMAINS AND FAMILIES:

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.**

Dataset 4 / 16163 Genes
Fusarium solani genes (v2.0)

Filters
Orthologous Fusarium verticillioides Genes: Excluded Chromosome/scaffold: 14
Start: 1128520
End: 1142558

Attributes
Gene stable ID
Transcript stable ID
Chromosome/scaffold name
Gene start (bp)
Gene end (bp)
Gene name
GO term accession
GO term name

Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

Step 3c: Expand the EXTERNAL section

GO

- GO term accession
- GO term name
- GO term definition

GOSlim GOA

- GOSlim GOA Accession(s)

Pathogen Phenotypes (source: PHI-base)

- PHI-base ID
- Host

External References (max 3)

- European Nucleotide Archive ID
- INSDC protein ID
- MEROPS - the Peptidase Database ID
- NCBI gene (formerly Entrezgene) description
- NCBI gene (formerly Entrezgene) accession
- NCBI gene (formerly Entrezgene) ID
- RNA ID
- RefSeq peptide predicted ID
- RFAM ID
- STRING ID
- tRNAscan-SE ID
- UniParc ID
- UniProtKB/Swiss-Prot ID
- UniPrintKB/T-CellML ID

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.

Step 4a: Click Results

Export all results to File

Email notification to

View Unique results only

Gene stable ID	Transcript stable ID	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)	Gene name	GO term accession	GO term name
NechaG73960	NechaT73960	14	1129115	1131280	PEPS	GO:0016021	integral component of membrane
NechaG73960	NechaT73960	14	1129115	1131280	PEPS	GO:0022857	transmembrane transporter activity
NechaG73960	NechaT73960	14	1129115	1131280	PEPS	GO:0050985	transmembrane transport
NechaG73960	NechaT73960	14	1129115	1131280	PEPS	GO:0016026	membrane
NechaG64937	NechaT64937	14	1131753	1133840	PDA1	GO:0016021	integral component of membrane
NechaG64937	NechaT64937	14	1131753	1133840	PDA1	GO:0004467	monoxygenase activity
NechaG64937	NechaT64937	14	1131753	1133840	PDA1	GO:0020037	heme binding
NechaG64937	NechaT64937	14	1131753	1133840	PDA1	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
NechaG64937	NechaT64937	14	1131753	1133840	PDA1	GO:0005506	iron ion binding
NechaG64937	NechaT64937	14	1131753	1133840	PDA1	GO:0016020	membrane

Step 4b: Change the number of rows to All to view all results in a new tab

Each attribute becomes a column in the results table

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.

Step 3.2a: Click on Attributes again

Please select the output and hit 'Results' when ready

Step 3.2b: Click on Sequences

Step 3.2c: Select cDNA sequences

Features Homologues
 Structures Sequences

SEQUENCES:
Sequences (max 1)

 Unspliced (Transcript)

5' UTR
 3' UTR
 Exon sequences
 cDNA sequences
 Coding sequence
 Peptide

Upstream flank
 Upstream flank

Downstream flank
 Downstream flank

HEADER INFORMATION:

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

Step 4.2: View results for the sequences

The screenshot shows the EnsemblFungi interface. In the top left, there's a logo and a navigation bar with 'New', 'Count', and 'Results' buttons. The 'Results' button is highlighted with a yellow box and has a callout bubble pointing to it with the text 'Step 4.2a: Click on Results again'. The main content area displays a dataset summary for 'Dataset 4 / 16163 Genes' and 'Fusarium solani genes (v2.0)'. It includes sections for 'Filters' (listing 'Orthologous Fusarium verticillioides Genes: Excluded Chromosome/scaffold: 14' with 'Start: 1128520' and 'End: 1142558') and 'Attributes' (listing 'Gene stable ID', 'Transcript stable ID', 'Gene name', and 'cDNA sequences'). Below this is a 'Dataset' section with '[None Selected]'. The right side of the screen shows a large block of FASTA sequence data for the gene NechaG73960, starting with the header '>NechaG73960 | NechaT73960 | PEP5'. The sequence itself is a long string of nucleotide bases.

*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

New Count Results

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]
- Botryotinia cinerea B05.10 gene stable ID
- Botryotinia cinerea B05.10 gene name
- Botryotinia cinerea B05.10 homology type

Please restrict your query using criteria below
(If filter values are truncated in any lists, hover over the list item to see the full text)

REGION:

GENE:

PATHOGEN PHENOTYPES (PHI-BASE):

GENE ONTOLOGY:

GO Term Accession (e.g. GO:0050789) [Max 500 advised]

GO Term Name (e.g. regulation of biological process)

GO Evidence code

MULTI SPECIES COMPARISONS:

PROTEIN DOMAINS AND FAMILIES:

VARIANT:

New | **Count** | **Results**

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]

Blumeria graminis Orthologues

- Blumeria graminis gene stable ID
- Blumeria graminis gene name
- Blumeria graminis protein or transcript stable ID
- Blumeria graminis chromosome/scaffold name
- Blumeria graminis chromosome/scaffold start (bp)
- Blumeria graminis chromosome/scaffold end (bp)
- Query protein or transcript ID
- Last common ancestor with Blumeria graminis
- Blumeria graminis homology type
- %id. target Blumeria graminis gene identical to query gene
- %id. query gene identical to target Blumeria graminis gene
- Blumeria graminis orthology confidence [0 low, 1 high]

Botrytis cinerea B05.10 Orthologues

- Botrytis cinerea B05.10 gene stable ID
- Botrytis cinerea B05.10 gene name
- Botrytis cinerea B05.10 protein or transcript stable ID
- Botrytis cinerea B05.10 chromosome/scaffold name
- Botrytis cinerea B05.10 chromosome/scaffold start (bp)
- Botrytis cinerea B05.10 chromosome/scaffold end (bp)
- Query protein or transcript ID
- Last common ancestor with Botrytis cinerea B05.10
- Botrytis cinerea B05.10 homology type
- %id. target Botrytis cinerea B05.10 gene identical to query gene
- %id. query gene identical to target Botrytis cinerea B05.10 gene
- Botrytis cinerea B05.10 orthology confidence [0 low, 1 high]

Candida albicans Orthologues

- Candida albicans gene stable ID
- Candida albicans gene name
- Candida albicans protein or transcript stable ID
- Candida albicans chromosome/scaffold name
- Query protein or transcript ID
- Last common ancestor with Candida albicans
- Candida albicans homology type
- %id. target Candida albicans gene identical to query gene

EnsemblFungi

New Count Results

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 Botrytis cinerea B05.10 ortholog confidence [0 low, 1 high]

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 ortholog confidence [0 low, 1 high]	
Mycg3G80087	Mycg3T390087										
Mycg3G88385	Mycg3T398385						Bcn16g00120	Bcox3	ortholog_one2one	0	
Mycg3G33131	Mycg3T33131										
Mycg3G107202	Mycg3T107202	BLGH_06273		ortholog_one2one	0		Bcn10g02340		ortholog_one2one	0	
Mycg3G54449	Mycg3T54449	BLGH_05551		ortholog_one2one	0		Bcn14g01092	Bcox1	ortholog_one2one	0	
Mycg3G32802	Mycg3T32802						Bcn05g02059	Bcox4	ortholog_one2many	0	
Mycg3G32802	Mycg3T32802						Bcn05g01450	Bcox2	ortholog_one2many	0	
Mycg3G44719	Mycg3T24719										

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.

EnsemblFungi

New Count Results BLAST BioMart FTP Docs & FAQs URL XML Perl Help

Dataset
Magnaporthe oryzae genes (MG8)

Filters
Chromosome/scaffold: 4
With Transmembrane helices: Only

Attributes
Gene stable ID
Transcript stable ID

Dataset
[None Selected]

REGION:
 Chromosome/scaffold
1
2
3
4
5

Coordinates
Start 1
End 10000000

Multiple regions (Chr:Start:End:Strand) [Max 500 advised]
e.g. 1:100:10000:-1, 1:100000:200000:1

GENE:

PATHOGEN PHENOTYPES (PHI-BASE):

GENE ONTOLOGY:

MULTI SPECIES COMPARISONS:

PROTEIN DOMAINS AND FAMILIES:
 Limit to genes ...
With Transmembrane helices Only Excluded

New Count Results URL XML Perl Help

Dataset 297 / 13470 Genes
Magnaporthe oryzae genes (MG8)

Filters
Chromosome/scaffold: 4
With Transmembrane helices: Only

Attributes
Gene stable ID
Gene description

Dataset
[None Selected]

Export all results to File HTML Unique results only Go
Email notification to

View 10 rows as HTML Unique results only

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

Please restrict your query using criteria below
(If filter values are truncated in any lists, hover over the list item to see the full text)

REGION:

GENE:
 Limit to genes (external references)... With ChEMBL ID(s) Only Excluded
 Input external references ID list [Max 500 advised] UniProtKB/Swiss-Prot ID(s) [e.g. A0ZWU1]
P31411
O13742
O60159
O94287
 Transcript count >= Choose File no file selected
 Transcript count <= Gene type
ncRNA
protein_coding
pseudogene
RNase_MRP_RNA
RNase_P_RNA
 Transcript type
ncRNA
protein_coding
pseudogene
RNase_MRP_RNA
RNase_P_RNA

New Count Results
 ★ URL XML Perl Help

Dataset 36 / 7268 Genes <i>Schizosaccharomyces pombe</i> genes (ASM294v2)	Export all results to <input type="text"/> File <input type="button" value="HTML"/> Unique results only <input type="checkbox"/> Go Email notification to <input type="text"/> View <input type="button" value="50"/> rows as <input type="button" value="HTML"/> Unique results only
Filters UniProtKB/Swiss-Prot ID(s) [e.g. A0ZWU1]: [ID-list specified]	
Attributes Gene stable ID Gene name Gene description PomBase ID	
Dataset [None Selected]	

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
SPAC16E.07c	vph1	V-type ATPase V0 subunit a (predicted) [Source:PomBase;Acc:SPAC16E.07c]	SPAC16E.07c_1
SPAC29B12.02c	sel2	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		CIC 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		CIC chloride channel (predicted) [Source:PomBase;Acc:SPBC887.02]	SPBC887.02_1
SPBC16D4.15	gpi16	pig-T, Gpi16 (predicted) [Source:PomBase;Acc:SPBC16D4.15]	SPBC16D4.15_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	gp13	pig-A, phosphatidylinositol N-acetylglucosaminyltransferase subunit Gp13 (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/Cog8 [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	glc2	glucosidase II alpha subunit Glc2 [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	gpb8	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%20tritici

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.

EnsemblFungi ▾ HMMER | BLAST | More ▾ Mycgr3G65552

New Search | Jobs ▾

Search Ensembl Fungi

- New Search
- Gene (1)
 - Ensembl Fungi (1)

Search results for 'Mycgr3G65552'

Showing 1 Gene found in Ensembl Fungi

Mycgr3G65552

Description	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:F9WWD1]
Gene ID	Mycgr3G65552
Species	Zymoseptoria tritici
Location	1:1786483-1788643
Gene trees	EGGT00050000025158 (Pan-taxonomic Compara) PTHR10587 (Fungi Compara)

Ensembl Fungi release 56 - Feb 2023 ©

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

Search Ensembl Fungi...

Zymoseptoria tritici (MG2) ▾

Location: 1:1,786,483-1,788,643 Gene: Mycgr3G65552 Transcript: Mycgr3T65552 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

Gene: Mycgr3G65552

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

Location Chromosome 1: 1,786,483-1,788,643 reverse strand.
MG2:ACPE0100001.1

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

Transcripts Hide transcript table

Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq	Flags
-	Mycgr3T65552	1868	471aa	Protein coding	F9WWD1	-	Ensembl Canonical

Summary

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

Gene type Protein coding

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

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

Add/remove tracks | Custom tracks | Share | Resize image | Export image | Reset configuration |

Genes

1.780Mb 1.785Mb 1.790Mb 1.795Mb Forward strand

Mycgr3T18164 > protein coding
Mycgr3T89063 > protein coding
Mycgr3T65551 > protein coding
ACPE0100001.1 >
< Mycgr3T23438 protein coding
< Mycgr3T65552 protein coding
< Mycgr3T107003 protein coding

Contigs

Genes

Configure this page

Custom tracks

Export data

Share this page

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.

[Login/Register](#)

Zymoseptoria tritici (MG2) ▾

Location: 1:1,786,483-1,788,643 Gene: Mycgr3G65552 Transcript: Mycgr3T65552 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

Gene: Mycgr3G65552

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

Location Chromosome 1: 1,786,483-1,788,643 reverse strand.
MG2:ACPE01000001.1

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

Transcripts [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq	Flags
-	Mycgr3T65552	1868	471aa	Protein coding	F9WWD1	-	Ensembl Canonical

Molecular interactions

This species

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

Interacts with [Show metadata](#)

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

Ensembl Fungi release 56 - Feb 2023 © EMBL-EBI

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](#).

[Login/Register](#)

Zymoseptoria tritici (MG2) ▾

Location: 1:1,786,483-1,788,643 Gene: Mycgr3G65552 Transcript: Mycgr3T65552 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: Molecular function
 - GO: Cellular component
 - GO: Biological process
 - PHI: Phibase identifier
- Phenotypes
- Genetic Variation
 - Variant table
 - Variant image
 - Structural variants
 - Gene expression
 - Pathway
- Molecular interactions**
- Regulation
- External references
- Supporting evidence
- ID
- Gene history

Gene: Mycgr3G65552

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

Location Chromosome 1: 1,786,483-1,788,643 reverse strand.
MG2:ACPE01000001.1

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

Transcripts [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq	Flags
-	Mycgr3T65552	1868	471aa	Protein coding	F9WWD1	-	Ensembl Canonical

Molecular interactions

This species

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

Interacts with [Show metadata](#)

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

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

Ensembl Fungi release 56 - Feb 2023 © EMBL-EBI

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.

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

Location: 1..1,786,485..1..1,788,643 Gene: Mycgr3G65552 Transcript: Mycgr3T65552 Jobs

Gene: Mycgr3G65552

Description Putative uncharacterized protein [Source-UniProtKB/EMBL_Acc:F99WW014]

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

About this gene MG2.ACPE01000001.1

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

Show/Hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq	Flags
-	Mycgr3T65552	1868	471aa	Protein coding	F99WW014	-	Ensembl Canonical

Orthologues

Download orthologues

Summary of orthologues of this gene Hide

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

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (370 species)	<input type="checkbox"/>	139	91	11	128
Aldomoryces (1 species)	<input type="checkbox"/>	1	0	0	0
Agaricales (21 species)	<input type="checkbox"/>	0	19	0	2
Athelioidales (2 species)	<input type="checkbox"/>	0	1	1	0
Blastocladiidae (1 species)	<input type="checkbox"/>	0	0	0	1
Boletales (9 species)	<input type="checkbox"/>	0	9	0	0
Botryosphaeriales (2 species)	<input type="checkbox"/>	2	0	0	0
Cantharellales (2 species)	<input type="checkbox"/>	0	3	0	0
Coprinodales (11 species)	<input type="checkbox"/>	9	2	0	0
Chlorobasidiomycetidae (8 species)	<input type="checkbox"/>	7	0	0	1
Chytridiomycota (6 species)	<input type="checkbox"/>	0	1	0	5
Corticales (7 species)	<input type="checkbox"/>	0	0	0	1
Cryptomycota (2 species)	<input type="checkbox"/>	0	0	0	2
Dacrymycetales (3 species)	<input type="checkbox"/>	0	3	0	0
Dothideales (1 species)	<input type="checkbox"/>	1	0	0	0
Dothideomycetes (3 species)	<input type="checkbox"/>	3	0	0	0
Eryiphylales (3 species)	<input type="checkbox"/>	3	0	0	0
Eurotiiales (17 species)	<input checked="" type="checkbox"/>	17	0	0	0
Fungi (16 species)	<input type="checkbox"/>	3	8	0	5

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

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

Selected orthologues Hide

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Aspergillus fumigatus A1293	1-to-1	AFUA_7G02500	54.37 %	42.25 %	n/a	n/a	Yes
		7_680_932_683_084..1					
		View Gene Tree					
		View Sequence Alignments					
Aspergillus fumigatus A1163	1-to-1	AFUB_080940	54.37 %	42.25 %	n/a	n/a	Yes
		DS499601_678_603_680_755..1					
		View Gene Tree					
		View Sequence Alignments					

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

 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

This species

Species	Gene ID	Interactor	Identifier	Source DB
Aspergillus fumigatus Af293	AFUA_7G02500	protein	uniprot:Q4WAU2	

Interacts with

Species	Gene ID	Interactor	Identifier	Source DB
Mus musculus	UNDETERMINED	protein	UNDETERMINED	PHI-base
Mus musculus	UNDETERMINED	protein	UNDETERMINED	PHI-base

Show metadata

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_prodid/. Search for *botrytis_cinerea*, 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.

[Login/Register](#)

e! EnsemblFungi ▾ HMMER | BLAST | BioMart | More ▾ [Bcin07g00720](#)

New Search Jobs ▾

Search Ensembl Fungi
 New Search
 Gene (1)
 Ensembl Fungi (1)

Configure this page

Custom tracks Export data Share this page Bookmark this page

Search results for 'Bcin07g00720'

Showing 1 Gene found in Ensembl Fungi

Bcin07g00720

Description	n/a
Gene ID	Bcin07g00720
Species	Botrytis cinerea B05.10
Location	7:260067-264879

Ensembl Fungi release 56 - Feb 2023 © EMBL-EBI

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

[Login/Register](#)

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

Zymoseptoria tritici (MG2) ▾

Location: 1:1,786,483-1,788,643 Gene: Mycgr3G65552 Transcript: Mycgr3T65552 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

Gene: Mycgr3G65552

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

Location Chromosome 1: 1,786,483-1,788,643 reverse strand.
MG2:ACPE01000001.1

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

Transcripts [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq	Flags
-	Mycgr3T65552	1868	471aa	Protein coding	F9WWD1	-	Ensembl Canonical

Summary

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

Gene type Protein coding

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

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

Add/remove tracks | Custom tracks | Share | Resize image | Export image | Reset configuration | Forward strand

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).

The screenshot shows the Ensembl Fungi interface for the gene **Bpk3** in *Botrytis cinerea* B05.10. The left sidebar contains a tree navigation for various species and databases. The main content area displays the gene's location (Chromosome 7: 260,067–264,879 forward strand), its protein coding sequence (AOA384JLJ46), and orthologs. A table of molecular interactions is shown, with a "Show metadata" link in the top right corner of the table header.

(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”.

The screenshot shows the Ensembl Fungi interface for the gene **Mycgr3G65552** in *Zymoseptoria tritici* (MG2). Similar to the previous screen, it includes a sidebar with navigation links. The main content area shows the gene's location (Chromosome 1: 1,788,483–1,788,643 reverse strand) and its orthologs. A detailed table of molecular interactions is displayed, with a "Show metadata" link in the top right corner of the "Interacts with" section.

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

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

Sordariidae 800.10 (ASM3294v1) ▾ Gene: Bpk3 (locusTag707270) [genes]

Gene-based displays

- Summary
- Splice variants
- Fungal comparison
- Gene alleles
- Sequence
- Secondary Structure
- Gene families
- Literature
- Fungal Comparisons
- Genomic alignments
- Gene gain/loss tree
- Gene gain/loss
- Orthologues
- Pan-taxonomic Comparisons
- Gene Tree
- Orthologues
- Ontologies
- GO: Biological process
- GO: Molecular function
- Phenotypes
- Gene expression
- Variant table
- Variant image
- Structural variants
- Gene expression
- Molecular interactions
- Regulation
- External references
- Supporting evidence
- Gene history

Orthologues

Orthologues

Show transcript table

Name: Transcript ID: bpk3_4631 Protein: Bpk3 Protein coding: AQA3H4L4B# (MGI384+)

Uniprot RefSeq Ensembl Canonical

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)		279	16	0	78
Aldomyces (1 species)		1	0	0	0
Ascomycota (1 species)		4	0	0	17
Basidiomycota (3 species)		1	0	0	1
Blastocladiellales (1 species)		0	1	0	0
Boletales (9 species)		3	0	0	6
Botryosphaeriales (2 species)		2	0	0	0
Candidales (1 species)		3	0	0	0
Cephalotrichomyces (8 species)		9	2	0	1
Chlorothyridomyces (6 species)		8	0	0	0
Chytridiomycota (4 species)		3	3	0	0
Cordyceps (1 species)		0	0	0	1
Curvulariaceae (2 species)		2	0	0	0
Dermatectomycetidae (3 species)		3	0	0	0
Dothideomycetes (1 species)		1	0	0	0
Erysiphales (3 species)		3	0	0	0
Endomycetidae (3 species)		17	0	0	0
Fungi (16 species)		13	0	0	3
Gastrales (1 species)		1	0	0	0
Geoglossales (1 species)		1	0	0	0
Glorenophytaceae (2 species)		0	1	0	1
Gymnosporangiales (1 species)		6	1	0	0
Heliotrichales (9 species)		8	1	0	0
Hymenochaetales (2 species)		0	0	0	2
Hypocreales (30 species)		30	0	0	0
Jaspliales (1 species)		0	0	0	1
Magnaporthe (1 species)		4	0	0	0
Microascales (4 species)		4	0	0	0

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

Selected orthologues Hide □

Species	Type	Show/hide columns	Orthologue	Target %id	Query %id	QGC Score	WGA Coverage	High Confidence
Magnaporthe poae	1-to-1		MAGP_0292T	51.17%	n/a	n/a	n/a	Yes
			supercontig_7.832.004-836.478-1					
			View Sequence Alignments					
Magnaporthe oryzae	1-to-1		M_8032_Eugene_00042871	50.05 %	49.58 %	n/a	n/a	Yes
			BR32_scfa0000003.3,056,924-3,069,846-1					
			View Sequence Alignments					
Geummannomyces tritici R3-111a-1	1-to-1		GGTG_00196	49.49 %	50.63 %	n/a	n/a	Yes
			supercontig_4.2,261,817-2,266,643-1					
			View Sequence Alignments					
Magnaporthe oryzae	1-to-1		ATG1 (MGG_06393)	49.90 %	51.47 %	n/a	n/a	Yes
			4.3,898,532-3,902,777-1					
			View Sequence Alignments					

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).

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

Magnaporthe oryzae (MG8) ▾ Location: 4:3,898,532-3,902,777 Gene: ATG1 Transcript: MGG_06393T0

Gene-based displays

Summary

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

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.

About this gene MG8:CM001234.1

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

Transcripts Show transcript table

Summary

Name ATG1 (UniProtKB Gene Name)

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

Gene type Protein coding

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

Go to Region In Detail for more tracks and navigation options (e.g. zooming)

Add/remove tracks Custom tracks Share Resize image Export image Reset configuration Reset track order

24.25 kb Forward strand

Genes 3.89Mb MGG_06394T0 > protein coding 3.90Mb MGG_06394T0 > protein coding 3.91Mb MGG_06391T0 > protein coding

Contigs Genes AACU03000115.1 > MGG_06394T0 MGG_06392T0 MGG_06390T0

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

The screenshot shows the Ensembl Fungi interface for the gene **ATG1** (MGG_063930) in *Magnaporthe oryzae* (MG8). The left sidebar has a 'Molecular interactions' section. The main content area displays the gene details and a table of interactions:

Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq	Flags
- MGG_063930	3957	982aa	Protein coding	Q52EB3	-		Ensembl Canonical

Molecular interactions

This species				Interacts with				Show metadata	
Species	Gene ID	Interactor	Identifier	Species	Gene ID	Interactor	Identifier	Source DB	
Magnaporthe oryzae 70-15	MGG_063930	protein	uniprot:Q52EB3	Hordeum vulgare	UNDETERMINED	protein	UNDETERMINED	PHI-base	
				Oryza sativa	UNDETERMINED	protein	UNDETERMINED	PHI-base	
				Oryza sativa	UNDETERMINED	protein	UNDETERMINED	PHI-base	

Ensembl Fungi release 56 - Feb 2023 © EMBL-EBI

(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/threonine kinase activity”, “ATP binding”, “kinase activity”, “transferase activity” and “protein serine kinase activity”.

The screenshot shows the Ensembl Fungi interface for the gene **Bpk3** (Bcln07g00720.1) in *Botrytis cinerea* B05.10. The left sidebar has a 'GO: Molecular function' section. The main content area displays the gene details and a table of GO annotations:

Accession	Term	Evidence	Annotation source	Transcript IDs
GO:0000156	nucleotide binding	IEA	UniProt	Bcln07g00720.1
GO:0004672	protein kinase activity	IEA		Bcln07g00720.1
GO:0004674	protein serine/threonine kinase activity	IEA		Bcln07g00720.1
GO:0005524	ATP binding	IEA		Bcln07g00720.1
GO:0016301	kinase activity	IEA	UniProt	Bcln07g00720.1
GO:0016740	transferase activity	IEA	UniProt	Bcln07g00720.1
GO:0106310	protein serine kinase activity	IEA	RHEA	Bcln07g00720.1

eEnsemblFungi ▾ HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog

Magnaporthe oryzae (MG8) ▾

Location: 3,898,532-3,902,777 Gene: ATG1 Transcript: MGQ_0639370 Jobs ▾

Gene-based display

- Summary
- Sanger variants
- Transcript comparison
- Gene alleles
- Secondary Structure
- Gene families
- Gene tree
- Fungal Comparisons
- Genomic alignments
- Gene expression
- Gene/gain/loss tree
- Orthologues
- Paralogues
- Pan-taxonomic Comparisons
- Gene Tree
- Orthologues
- Ontologies
- GO: Cellular component
- GO: Biological process
- **GO: Molecular function**
- P-Hit: Pfam domain identifier
- Phenotype
- Genetic Variation
- Variant table
- Linkage disequilibrium
- Structural variants
- Gene expression
- Gene body
- Molecular interactions
- Regulation
- Supporting references
- Supporting evidence
- ID History
- Gene history

Gene: ATG1 MGQ_0639370

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 ([pedice variant](#)) and 313 orthologues.

Transcripts [Hide transcript table](#)

Name	Transcript ID	bp	Protein	BioType	UniProt	RefSeq	Flags	Annotations
-	MGQ_0639370	3957	982aa	Protein coding	Q52EB3	-	-	Ensembl Canonical

GO: Molecular function ▾

GO: Molecular function (columns: 1 hidden) [Filter](#)

Accession	Term	Evidence	Annotation source	Transcript IDs
GO_0000168	nucleotide binding	IEA	UniProt	MGQ_0639370
GO_0004672	protein kinase activity	IEA	InterPro	MGQ_0639370
GO_0004674	protein serine/threonine kinase activity	IMP		MGQ_0639370
GO_0005524	ATP binding	IEA		MGQ_0639370
GO_0016301	kinase activity	IEA	UniProt	MGQ_0639370
GO_0016740	transferase activity	IEA	UniProt	MGQ_0639370
GO_0106310	protein serine kinase activity	IEA	RHEA	MGQ_0639370

Exercise: Attaching Track Hubs

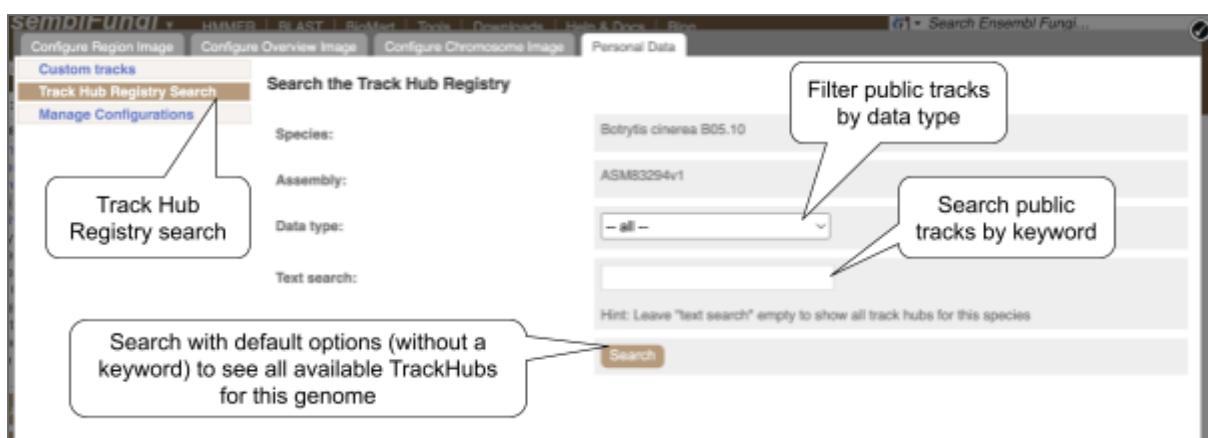
There are a number of publicly available datasets that are available to add onto views in Ensembl. You can find full lists of these at www.trackhubregistry.org. We're going to search and add these files from within Ensembl.

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

The screenshot shows a search interface. At the top, there is a search bar containing "Botrytis cinerea B05.10" and a dropdown arrow. Below it is another input field with "6:1854110-1894000". To the right of these fields is a "Go" button. Below the input fields, there is a note: "e.g. NAT2 or alcohol*".

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

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



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

ensemblFunai | HMMER | BLAST | BioMart | Tools | Downloads | Help & Done | Play

Configure Region Image | Configure Overview Image | Configure Chromosome Image | Personal Data

Search Results

Searched *Botrytis cinerea* B05.10 ASM83294v1

Found 4 track hubs - [Search again](#)

RNA-Seq alignment hub SRP062592

Description: Next Generation Sequencing Facilitates Quantitative Analysis of Cucumber and *Botrytis cinerea* Transcriptome Changes During Infection ; [SRP062592](#)

Data type: transcriptomics

Number of tracks: 2

RNA-Seq alignment hub SRP080917

Description: Molecular analysis of interaction between the grapevine flower and *Botrytis cinerea* ; [SRP080917](#)

Data type: transcriptomics

Number of tracks: 6

RepetDB Botrytis cinerea B05.10 (ASM83294v1)

Description: Repeat region consensus copies annotations created by TEannot (from the REPET package). Go to [RepetDB](#) for more info.

Data type: genomics

Number of tracks: 5

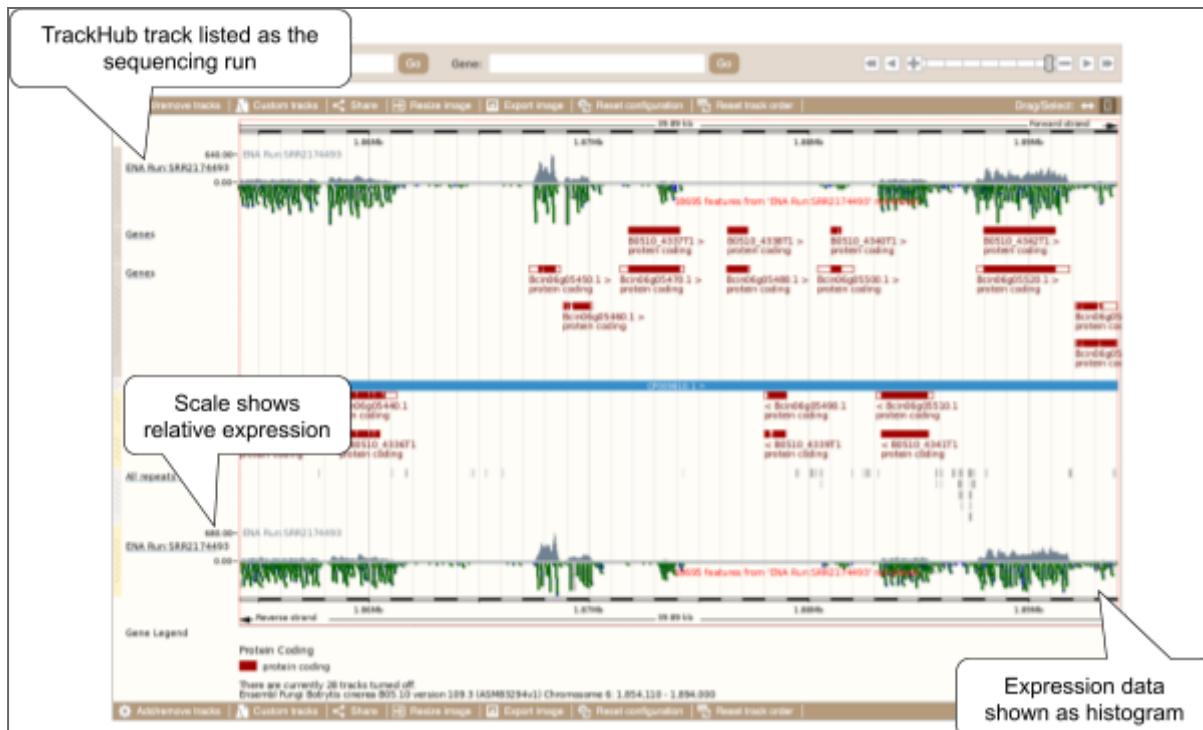
RNA-Seq alignment hub SRP093589

PROCESSING

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?

The Track Hub Registry

A global centralised collection of publicly accessible track hubs

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

SRP062592

The screenshot shows the 'The Track Hub Registry' interface. At the top, there are links for 'Submit data', 'Documentation', 'About', and 'Help'. A search bar with placeholder 'Search by keywords: hg' and a magnifying glass icon is on the right. Below the header, the URL 'Home / SRP062592 - GCA_000832945.1' is shown. The main content area has tabs for 'SRP062592', 'Botrytis cinerea B05.10', and 'GCA_000832945.1'. The 'GCA_000832945.1' tab is active. Under 'General Info', it lists 'Remote data tracks: 1', 'Data Type: transcriptomics', 'File type(s): cram: 1', and 'Source URL: [View](#)'. A dropdown menu for 'View in Genome Browser' offers options like ENSEMBL, UCSC, VECTORBASE, and NCBI GDV. To the right, under 'Hub', are fields for 'Name: SRP062592', 'Short Label: RNA-Seq alignment hub SRP062592', 'Long Label: Next Generation Sequencing Facilitates Quantitative Analysis of Cucumber and Botrytis cinerea Transcriptome Changes During Infection; [SRP062592](#)', 'Assembly Hub: X', and 'Public URL: [View](#)'. Under 'Species', it shows 'Taxonomy: 332648', 'Scientific name: Botrytis cinerea B05.10', and 'Common name:'. Under 'Assembly Information', a table shows 'Accession: GCA_000832945.1', 'Name: ASM83294v1', 'Long Name: ', and 'UCSC Synonym: ASM83294v1'.

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

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

Let's take a look at that URL.

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

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

	Name	Last modified	Size	Description
	Parent Directory			
	Spom_all_61G9EAAXX_and_61G9UAAXX.+.sorted.bam	2014-11-26 15:06	3.3G	
	Spom_all_61G9EAAXX_and_61G9UAAXX.+.sorted.bam.bai	2014-11-26 15:06	36K	
	Spom_all_61G9EAAXX_and_61G9UAAXX.-.sorted.bam	2014-11-26 15:04	3.8G	
	Spom_all_61G9EAAXX_and_61G9UAAXX.-.sorted.bam.bai	2014-11-26 15:04	37K	

Here you can see two BAM files (.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](#).

The screenshot shows the Ensembl Fungi homepage for the *Schizosaccharomyces pombe* genome (ASM294v2). The main content area includes a search bar, a brief overview of the genome, and links to genome assembly (ASM294v2), gene annotation, and other resources. A callout bubble points to the "Display your data in Ensembl Fungi" link.

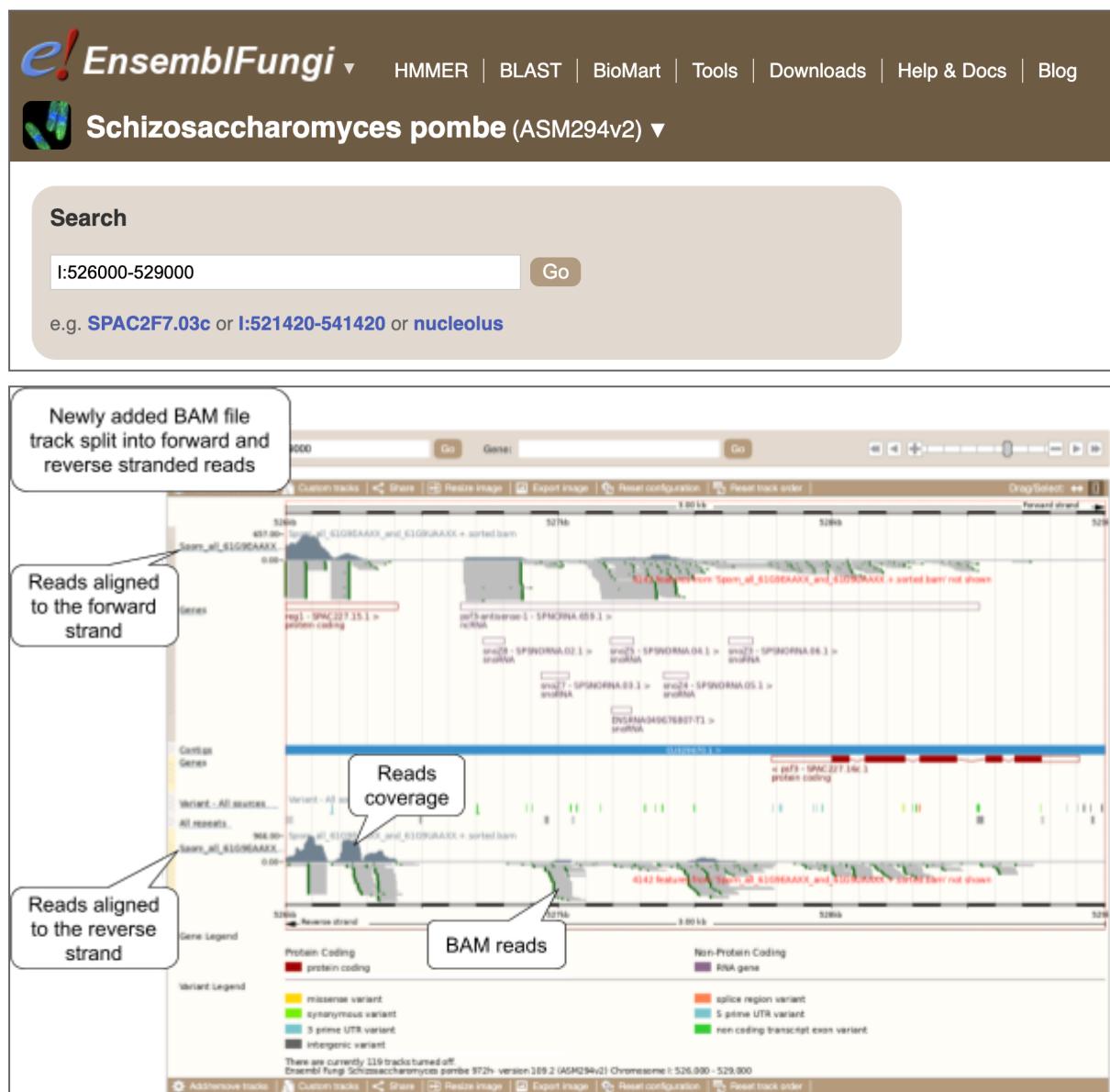
A menu will appear:

The screenshot shows the "Add a custom track" dialog box. It includes fields for "Name for this data (optional)", "Species" (set to *Schizosaccharomyces pombe*), "Data" (with a URL input field containing <http://ftp.ensemblgenomes.org/pub/micr...>), "Data format" (set to "BAM"), and an "Add data" button. Callout bubbles point to each of these elements: "Give your track a name", "Paste data or URL here", "Upload a file from your local machine (max. 20MB)", "Click Add data to view your track", and "Ensembl automatically recognises the file extension when given".

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.

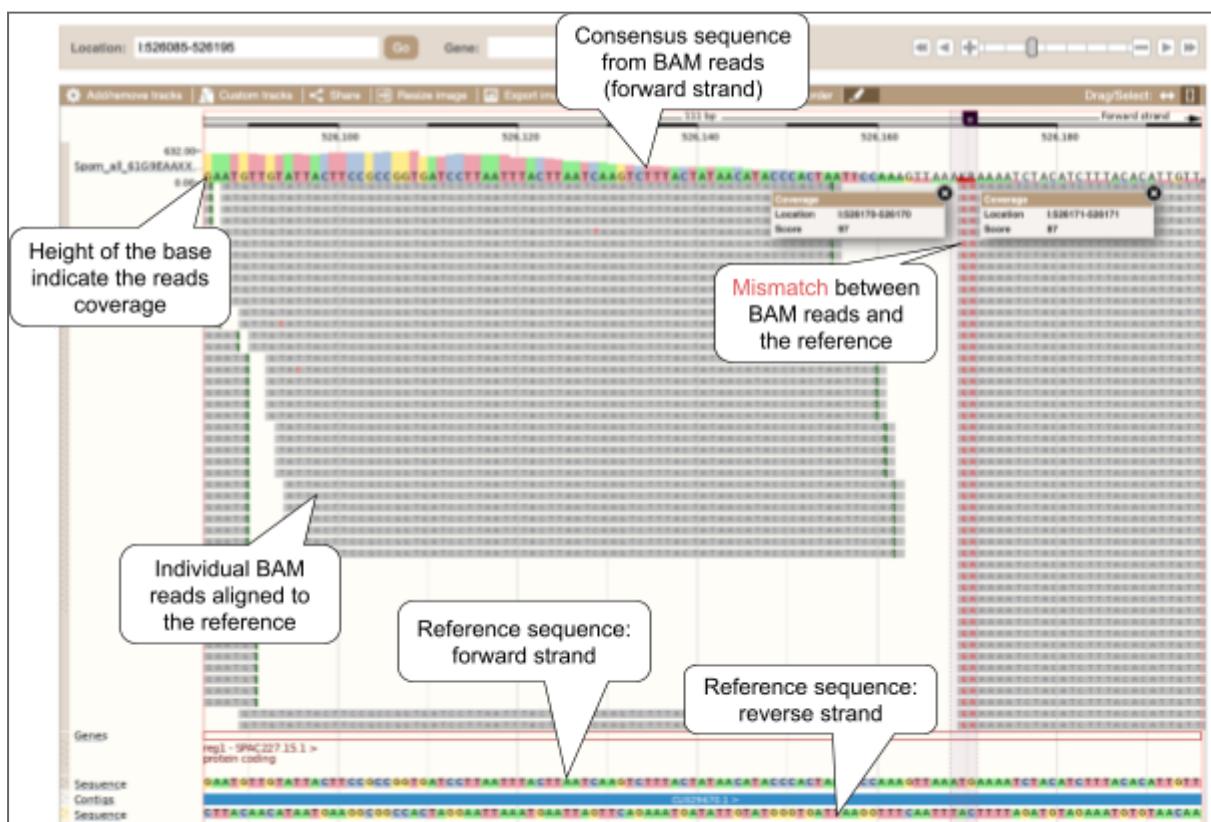
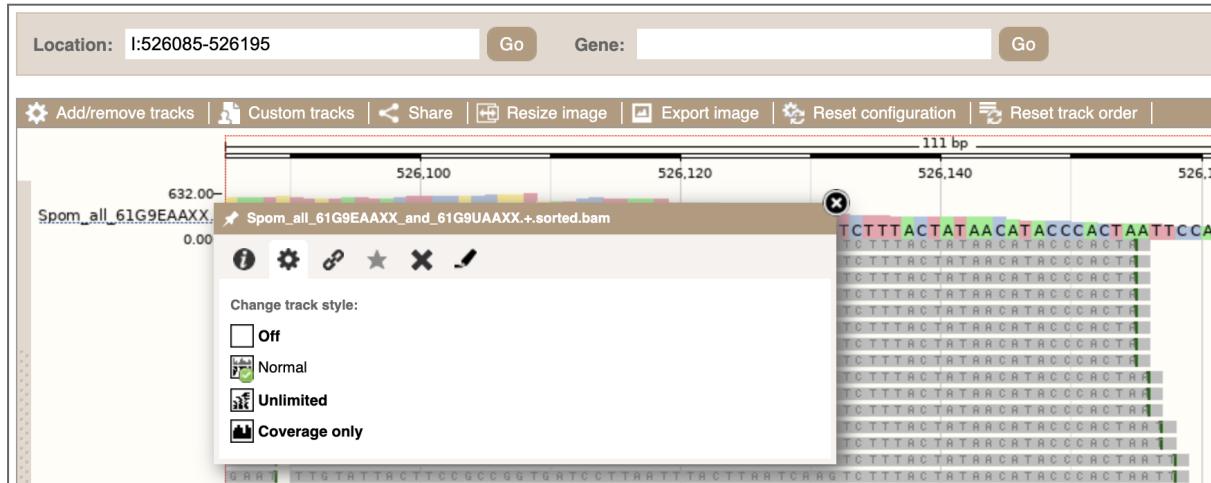
If you have a URL, like the one we located earlier, paste in the URL of the BAM file itself (ftp://ftp.ensemblgenomes.org/pub/misc_data/bam/fungi/Spom/Spom_all_61G9EAAXX_and_61G9UAAXX.+sorted.bam).

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

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

Compile a list of genes involved in glycolysis.

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

The screenshot shows the SGD (Saccharomyces Genome Database) homepage. At the top, there is a navigation bar with links for About, Blog, Download, Help, YeastMine, and social media icons. Below the navigation bar is a search bar containing the query "glycolysis". To the right of the search bar, there is a summary box titled "About SGD" with text about the database and some search results. An orange arrow points from the text "On the SGD home page" in the instructions above to the search bar on the screen.

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

The screenshot shows the SGD search results page for the query "glycolysis". The results are categorized by type: References (570), Genes (33), Biological Processes (30), Downloads (5), Molecular Functions (3), Cellular Components (2), and Chemicals (1). A left sidebar lists these categories with corresponding icons. An orange arrow points from the text "On the Results page" in the instructions above to the "Genes" category in the sidebar. The main results area shows the first few entries under the "Genes" category, including "canonical glycolysis" and "glycolysis from storage polysaccharide through".

- 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 screenshot shows the SGD search results page for the query "glycolysis" with additional filters applied: "glycolysis", "glycolytic process (direct)", and "Gene". The results are categorized by Feature Type: ORF (15), Molecular Function (catalytic activity (direct) 7, ATP binding (direct) 6, kinase activity (direct) 6, nucleotide binding (direct) 6, transferase activity (direct) 6). A left sidebar shows the "Genes / Genomic Features" category selected. An orange arrow points from the text "Find the Biological Process category" in the instructions above to the "Genes / Genomic Features" category in the sidebar. Another orange arrow points from the text "To download the list of genes" in the instructions above to the "Download" button in the top right corner of the results table. The results table lists genes such as GPM1, PGK1, ENO1, TDH1, FBA1, ENO2, PFK1, PFK2, TDH3, CDC19, TDH2, TPI1, PGI1, GLK1, and HXK1.

- 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:

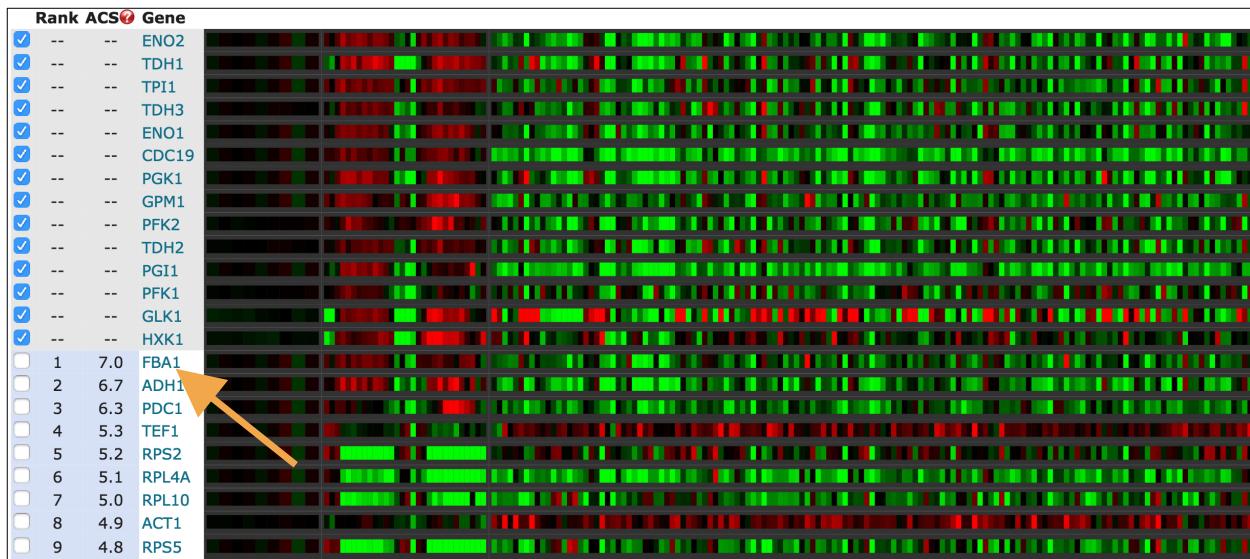
- To access SPELL, go to the SGD home page at www.yeastgenome.org, open the **Function** tab on top of the page and click on **Expression**. Or, if you are already on a Locus Summary page, open the Expression tab and click on the SPELL link under the histogram.

The screenshot shows the SGD home page. At the top, there is a navigation bar with links for About, Blog, Download, Help, YeastMine, and social media icons. A search bar contains the query "search: actin, kinase, glucose". Below the search bar, there is a main content area featuring a microscopy image of yeast cells with green and red fluorescence. To the right of the image is a sidebar titled "About SGD" which provides a brief overview of the database. The "Expression" menu item in the top navigation bar is highlighted with an orange arrow. At the bottom of the page, there are sections for Meetings (31st VHYC Yeast Conference) and New & Noteworthy (Trouble with Triplets - April 06, 2018), along with a Twitter feed from the SGD Project (@yeastgenome).

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

The screenshot shows the SPELL search interface for *S. cerevisiae*. The title "SPELL - *S. cerevisiae*" is at the top. Below it is a purple header bar with the text "SPELL (Yeast)". The main search form has a "Gene Name(s)" input field containing "GPM1 PGK1 PFK1 PFK2 ENO2 ENO1 CDC" and a "Search" button. An orange arrow points to the "Search" button. Below the search form is another purple bar with the text "+ Options for Filtering Results by Dataset Tags".

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



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

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

+ Additional Display Options

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

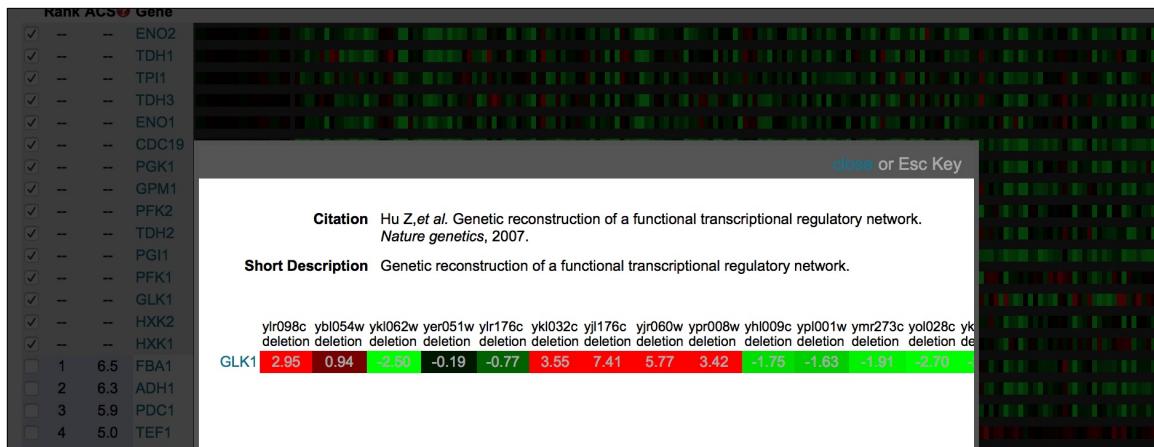
- To select only datasets with particular tags, click on + **Options for Filtering Results**.

Dataset Tags ⓘ

Select: all none previous query toggle

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

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



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

GO Term Enrichment	
GOTerm	P-val
glucose catabolic process (biological_process)	1.33e-29
hexose catabolic process (biological_process)	2.39e-28
monosaccharide catabolic process (biological_process)	2.91e-27
glycolysis (biological_process)	4.79e-27
glucose metabolic process (biological_process)	1.66e-23
single-organism carbohydrate catabolic process (biological_process)	3.62e-22
hexose metabolic process (biological_process)	4.32e-22
monosaccharide metabolic process (biological_process)	1.42e-21
carbohydrate catabolic process (biological_process)	1.97e-21
generation of precursor metabolites and energy (biological_process)	7.97e-18
single-organism carbohydrate metabolic process (biological_process)	1.60e-13
gluconeogenesis (biological_process)	3.72e-13
hexose biosynthetic process (biological_process)	5.25e-13
monosaccharide biosynthetic process (biological_process)	7.33e-13

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 is not available (e.g. no replicates). After selecting samples, you have the option to take the average, minimum, or maximum expression value within each group. If choosing only one sample from a group, the selected 'operation' will not affect your results. Time-series experiments will offer an extra parameter called "Global min/max" which allows you to filter your results further. Finally, you can choose the directionality and the magnitude of the difference (e.g., up/down 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.

The screenshot shows the FungiDB.org search interface with the following steps highlighted:

- Step 1:** The search bar contains "rma". The sidebar shows "Genes" selected, with "RNA-Seq Evidence" highlighted by a red circle.
- Step 2:** The main search results page for "kirch" shows one result: "Candida transcriptomes during oropharyngeal candidiasis infection in mouse (Kirchner, et al. 2019)". The "DE" button is highlighted by a red circle.
- Step 3:** The search parameters for the transcriptome dataset are shown. The "Reference Sample" dropdown is set to "SC5314_in vitro" (highlighted by a red circle).
- Step 4:** The "Comparator Sample" dropdown is set to "SC5314_infected_1d" (highlighted by a red circle).
- Step 5:** The "Direction" dropdown is set to "up-regulated" (highlighted by a red circle).
- Step 6:** The "fold difference >=" input field is set to "4" (highlighted by a red circle).
- Step 7:** The "adjusted P value less than or equal to" input field is set to "0.1" (highlighted by a red circle).

At the bottom right, there is a yellow box labeled "Calb_Kirchner_mouse (de) 589 Genes" and a blue dashed box labeled "+ Add a step".

- Identify genes that are up-regulated in SC5314 but not 101 persistent strain at 1d of infection.

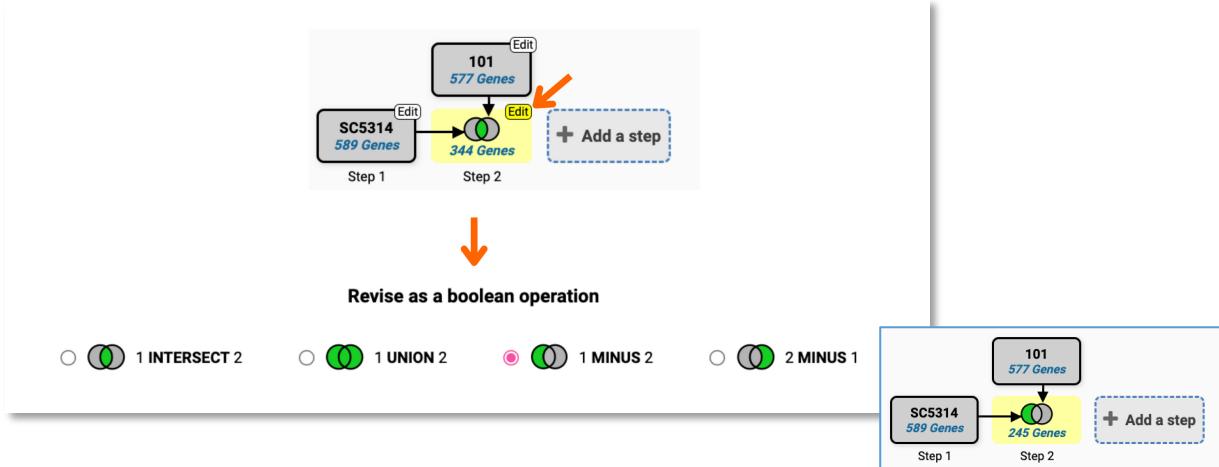
1. Click on the “Add Step” button
2. Navigate to the RNA-Seq Evidence search, filter for “Kirch” to quickly identify the dataset and click on the “DE” button.
3. Choose to examine the sense strand.
4. Select reference sample: 101 _ in vitro.
5. Select comparator sample: 101_infected_1d.
6. Look for up-regulated genes.
7. Select magnitude of upregulation: 4 fold.

The screenshot shows the IPA software interface with the following steps highlighted:

- Step 1:** A yellow box labeled "Calb_Kirchner_mouse (de) 589 Genes". An orange circle labeled "1" points to the "+ Add a step" button.
- Step 2:** A blue box titled "Combine with other Genes". It shows a flowchart where "Calb_Kirchner_mouse (de) 589 Genes" is combined with "Calb_Kirchner_mouse (de) 589 Genes". An orange circle labeled "2" points to the "Transform into related records" section.
- Step 3:** A grey box titled "Identify Genes based on RNA-Seq Evidence". It shows a search bar with "Calb_Kirchner SC5314" and a dropdown menu with "Gene Model Characteristics", "Unannotated Interactions", "Transcriptomics", "Protein Evidence", and "RNA-Seq Evidence". An orange circle labeled "3" points to the "RNA-Seq Evidence" option.
- Step 4:** A grey box titled "Reference Sample". It lists various samples with radio buttons. An orange circle labeled "4" points to the "101_in vitro" option.
- Step 5:** A grey box titled "Comparator Sample". It lists various samples with radio buttons. An orange circle labeled "5" points to the "101_infected_1d" option.
- Step 6:** A grey box titled "Direction". A dropdown menu is set to "up-regulated". An orange circle labeled "6" points to this field.
- Step 7:** A grey box titled "fold difference >= ". A text input field contains the value "4". An orange circle labeled "7" points to this field.
- Step 8:** A grey box titled "adjusted P value less than or equal to". A text input field contains the value "0.1".
- Run Step:** A grey button with an orange arrow pointing right.
- Result:** A blue box titled "Calb_Kirchner_mouse (de) 589 Genes". It shows a green circle with a minus sign and the text "344 Genes". An orange circle labeled "8" points to this result.

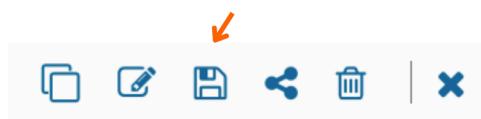
The default setting of the Boolean operators was set to the “intersect” option, which returns genes that are up-regulated by 4 fold in both strains.

- 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 accessed by clicking on the Download button located about the results table.

Gene ID	Transcript ID	Genomic Location (Gene)	Product Description
C2_05910W_A	C2_05910W_A-T	Ca22chr3A_C._albicans_SC5314:1,325,453..1,328,761(+)	Zn(2)-C6 fungal-type domain-containing protein [Source:UniProtKB/TrEMBL;Acc:A0A1D8PKB5]
C2_09700W_A	C2_09700W_A-T	Ca22chr2A_C._albicans_SC5314:1,982,608..1,983,588(+)	Yea4p [Source:UniProtKB/TrEMBL;Acc:A0A1D8PJJ0]
CR_00920W_A	CR_00920W_A-T	Ca22chrRA_C._albicans_SC5314:207,723..208,721(+)	Ydc2-catalyt domain-containing protein [Source:UniProtKB/TrEMBL;Acc:Q5A864]
C6_02170C_A	C6_02170C_A-T	Ca22chr6A_C._albicans_SC5314:451,184..452,335(-)	WD_REPEATS_REGION domain-containing protein [Source:UniProtKB/TrEMBL;Acc:A0A1D8PPT7]
C1_12750C_A	C1_12750C_A-T	Ca22chr1A_C._albicans_SC5314:2,779,463..2,781,025(-)	WD_REPEATS_REGION domain-containing protein [Source:UniProtKB/TrEMBL;Acc:A0A1D8PFH7]

Download Genes

Results are from search: Combine Gene results

- Choose a Report:
- Tab- or comma-delimited (openable in Excel) - choose columns to make a custom table [?](#)
 - Tab- or comma-delimited (openable in Excel) - choose a pre-configured table [?](#)
 - FASTA - sequence retrieval, configurable [?](#)
 - GFF3 - gene models [?](#)
 - Standard JSON [?](#)

You can also export yeast orthologs of *C. albicans* genes into Yeastmine. 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:

The screenshot shows the YeastMine interface with the following components:

- Workflow Steps:** A series of boxes connected by arrows, showing the transformation of 101 genes into 245 orthologs, and finally into 252 orthologs.
- Download Genes Section:** Shows results from a search: Transform by Orthology. It includes a 'Choose a Report' dropdown and a table of orthologous genes.
- Results Table:** A table showing orthologous genes for SCS314, grouped by source_id. The table includes columns for Gene ID, source_id, and a preview of the data.
- Bottom Navigation:** Includes links for Home, Templates, Lists, QueryBuilder, Tools, Regions, Data Sources, API, Contact Us, Video Tutorials, Help, Log In, and a search bar.

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.

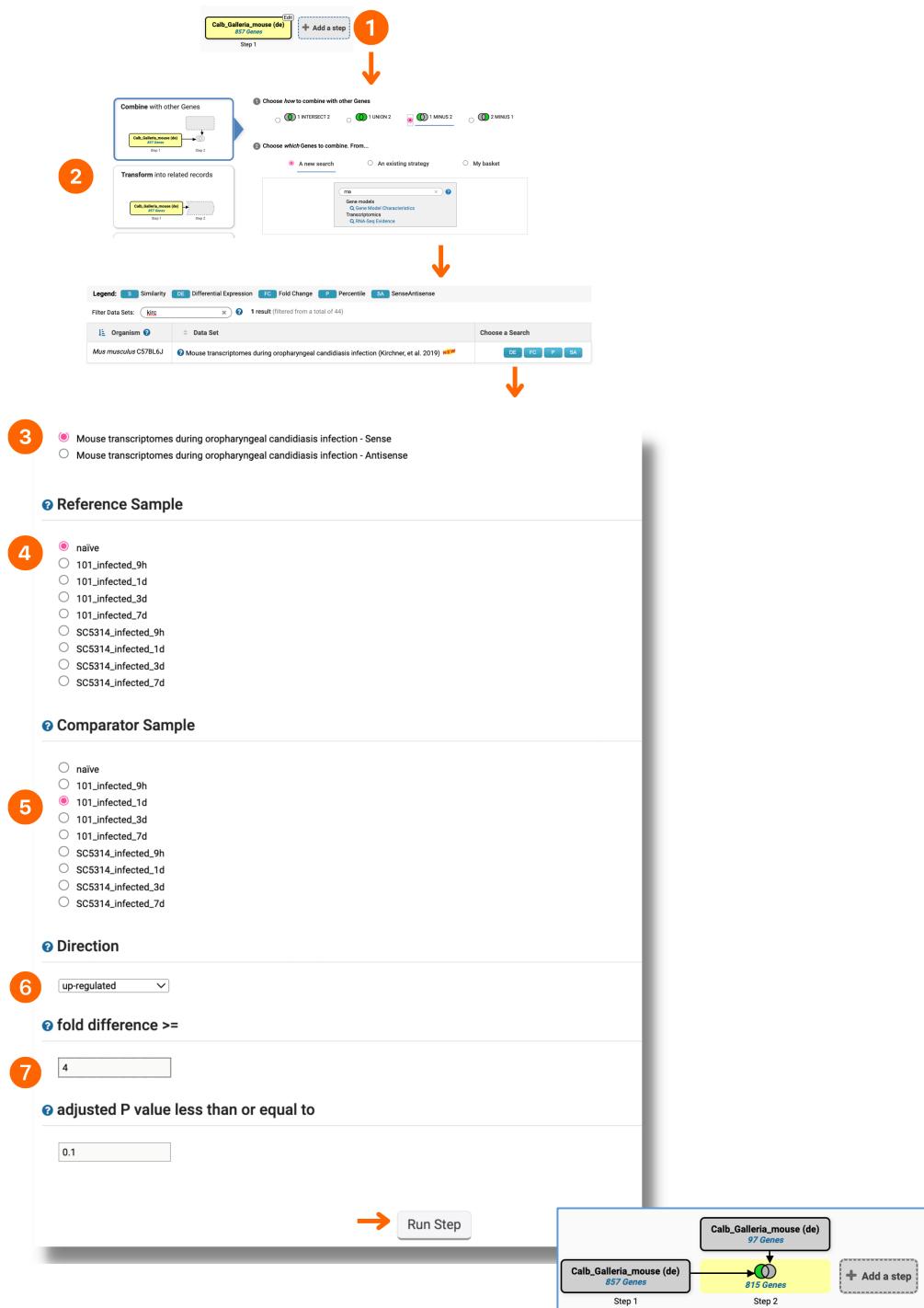
The screenshot shows the HostDB.org search interface with the following steps highlighted:

- Step 1:** The search bar contains "rna". The "RNA-Seq Evidence" button is circled in orange.
- Step 2:** The search results for "kirch" show one result: "Mouse transcriptomes during oropharyngeal candidiasis infection (Kirchner, et al. 2019)". The "DE" button is circled in orange.
- Step 3:** The search results page shows the selected dataset. The "Reference Sample" dropdown is open, showing options like "naive" and various infected time points. The "naive" option is circled in orange.
- Step 4:** The "Reference Sample" dropdown is shown again, with "naive" selected. Other options include 101_infected_9h, 101_infected_1d, 101_infected_3d, 101_infected_7d, SC5314_infected_9h, SC5314_infected_1d, SC5314_infected_3d, and SC5314_infected_7d. The "naive" option is circled in orange.
- Step 5:** The "Comparator Sample" dropdown is open, showing options like "naive" and various infected time points. The "SC5314_infected_1d" option is circled in orange.
- Step 6:** The "Direction" dropdown is set to "up-regulated". The "fold difference >=" dropdown is set to "4".
- Step 7:** The "adjusted P value less than or equal to" input field is set to "0.1".

At the bottom right, there is a box labeled "Calb_Galleria_mouse (de) 857 Genes" with an "Edit" button, and a "Revise" button above it. A "Step 1" label is also present.

- 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.

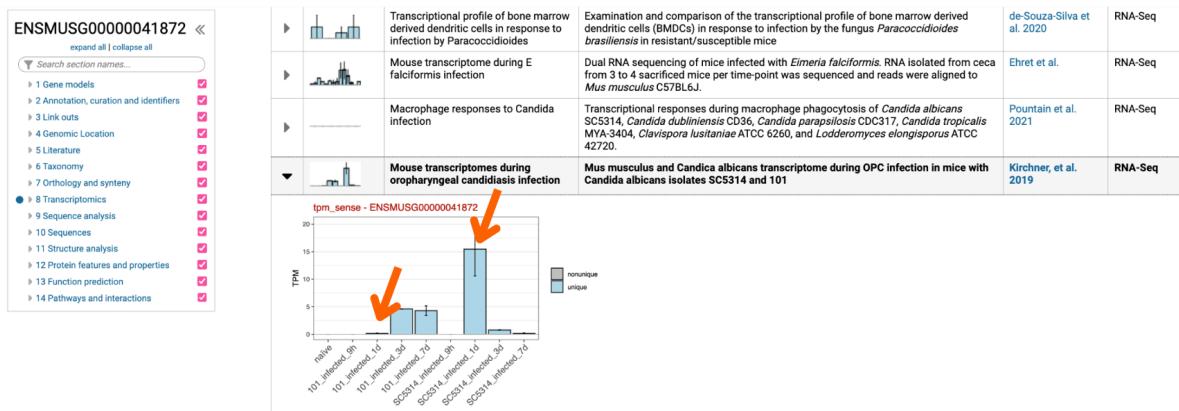


- Examine the results in HostDB:

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
ENSMUSG000000041879	ENSMUST00000194867	mmusC57BL6J_chr1:15,853,331..15,856,499(+)	predicted gene, 37500 [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	ENSMUST0000027061	mmusC57BL6J_chr1:20,730,905..20,734,496(+)	interleukin 17A [Source:MGI Symbol;Acc:MGI:107364]	1
ENSMUSG000000041872	ENSMUST0000039046	mmusC57BL6J_chr1:20,777,146..20,790,617(-)	interleukin 17F [Source:MGI Symbol;Acc:MGI:2676631]	4
ENSMUSG000000041872	ENSMUST00000189301	mmusC57BL6J_chr1:20,777,146..20,790,617(-)	interleukin 17F [Source:MGI Symbol;Acc:MGI:2676631]	4
ENSMUSG000000041872	ENSMUST00000190692	mmusC57BL6J_chr1:20,777,146..20,790,617(-)	interleukin 17F [Source:MGI Symbol;Acc:MGI:2676631]	4
ENSMUSG000000041872	ENSMUST00000191111	mmusC57BL6J_chr1:20,777,146..20,790,617(-)	interleukin 17F [Source:MGI Symbol;Acc:MGI:2676631]	4
ENSMUSG0000000104358	ENSMUST00000192924	mmusC57BL6J_chr1:34,823,525..34,826,560(+)	predicted gene, 37127 [Source:MGI Symbol;Acc:MGI:5610355]	1
ENSMUSG000000047180	ENSMUST00000056946	mmusC57BL6J_chr1:36,264,597..36,274,679(-)	neuronal E3 ubiquitin protein ligase 3 [Source:MGI Symbol;Acc:MGI:2429944]	2
ENSMUSG000000047180	ENSMUST00000188666	mmusC57BL6J_chr1:36,264,597..36,274,679(-)	neuronal E3 ubiquitin protein ligase 3 [Source:MGI Symbol;Acc:MGI:2429944]	2
ENSMUSG000000037447	ENSMUST00000099778	mmusC57BL6J_chr1:36,307,731..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG000000037447	ENSMUST00000115029	mmusC57BL6J_chr1:36,307,731..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG000000037447	ENSMUST00000115031	mmusC57BL6J_chr1:36,307,731..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG000000037447	ENSMUST00000115032	mmusC57BL6J_chr1:36,307,731..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG000000037447	ENSMUST00000116629	mmusC57BL6J_chr1:36,307,731..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG000000037447	ENSMUST00000124280	mmusC57BL6J_chr1:36,307,731..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG000000037447	ENSMUST00000126413	mmusC57BL6J_chr1:36,307,731..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG000000037447	ENSMUST00000137906	mmusC57BL6J_chr1:36,307,731..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG000000037447	ENSMUST00000140218	mmusC57BL6J_chr1:36,307,731..36,324,029(+)	AT rich interactive domain 5A (MRF1-like) [Source:MGI Symbol;Acc:MGI:2443039]	15
ENSMUSG000000037447	ENSMUST00000141121	mmusC57BL6J_chr1:36,307,731..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.

1

2

3

4

5

6

Circadian proteomic analysis - delta csp-1
Circadian proteomic analysis - Wild Type

Direction

Comparison

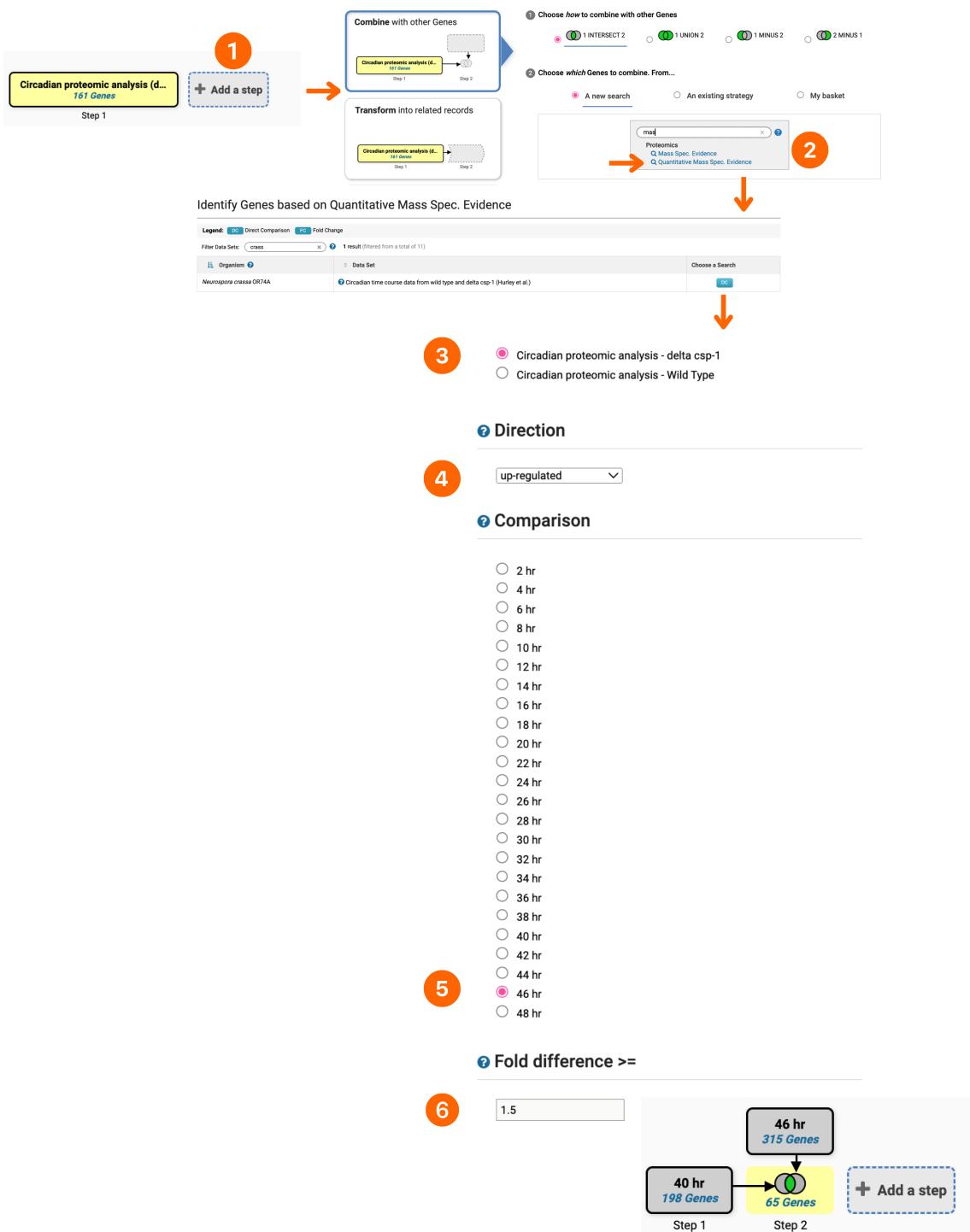
Fold difference >=

40 hr
198 Genes

Add a step

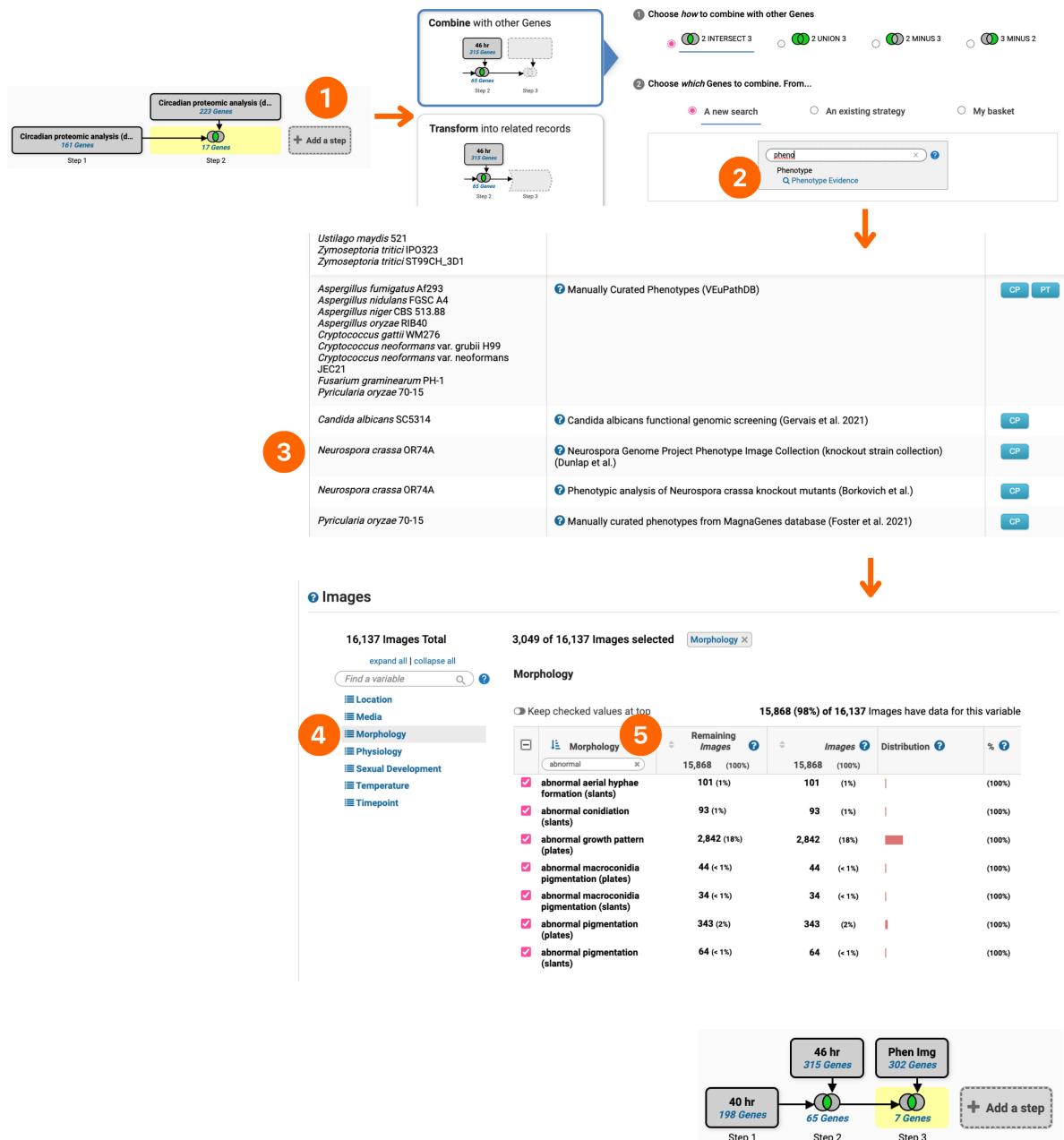
- Identify proteins expressed in culture at 46 hr.

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

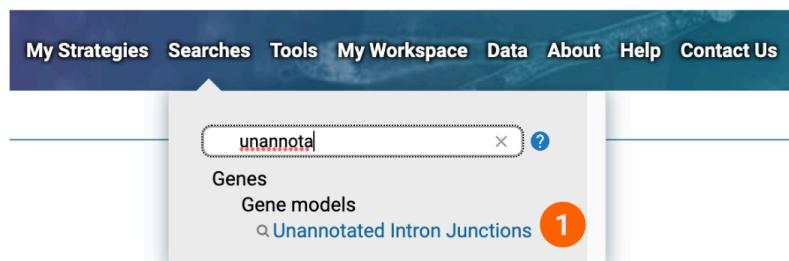
Assessing (& editing) gene annotation (JBrowse/Apollo) (optional)

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.

Identify Genes based on Unannotated Intron Junctions

Organism
Note that this search is only available for genomes to which RNA sequencing reads have been mapped.
1 selected, out of 30

muco
 Fungi
 Mucormycota
 Mucor lusitanicus CBS 277.49 [Reference]

Minimum number of unique reads >=

Percent of Most Abundant Intron (MAI)
 to

Consider 5' & 3' Flanking Sequence up to (bp)

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.

Gene ID	Transcript ID	Product Description	# Unannotated Junctions	Max % Mai	Max Unique Reads	Max ISRPM
QYA_157425	QYA_157425T0	1-Acyl dihydroacetone phosphate reductase and related dehydrogenases	1	62.1	1906	801.98
QYA_148011	QYA_148011T0	2-oxoisovalerate dehydrogenase subunit alpha [Source:UniProtKB/TrEMBL;Acc:ADA168 9R9]	1	22.5	118	68.54
QYA_116745	QYA_116745T0	3-hydroxy-3-methylglutaryl coenzyme A reductase [Source:UniProtKB/TrEMBL;Acc:ADA168 MWS]	1	42	170	82.25
QYA_72770	QYA_72770T0	3-hydroxyacyl-[acyl-carrier-protein] dehydratase [Source:UniProtKB/TrEMBL;Acc:A0A162M011]	1	47.2	50	16.58
QYA_153891	QYA_153891T0	3-hydroxyacyl-CoA dehydrogenase	1		351	181.81

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 multi-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.

The screenshot shows a results table with various columns. A red circle highlights the 'Add Columns' button at the top right of the table. An orange arrow points from this button to a 'Select Columns' dialog box. The dialog box title is 'Select Columns' and it displays '6 columns selected, out of 80 columns allowed'. It includes a checkbox for 'JBrowse' under the 'Gene models' section, which is checked. A red arrow points to the 'Update Columns' button at the bottom right of the dialog box.

	Gene ID	Transcript ID	Product Description	# Unannotated Junctions	% MAI	Max Unique Reads	Max ISRPM	JBrowse
	QYA_157425	QYA_157425T0	1-Acyl dihydroacetone phosphate reductase and related dehydrogenases	1	100	1906	801.98	
	QYA_148011	QYA_148011T0	2-oxoisovalerate dehydrogenase subunit alpha [Source:UniProtKB/TREMBL;Acc:A0A168I9R9]	1	62.1	118	68.54	
	QYA_116745	QYA_116745T0	3-hydroxy-3-methylglutaryl coenzyme A reductase [Source:UniProtKB/TREMBL;Acc:A0A168MWS8]	2	22.5	170	82.25	
	QYA_72770	QYA_72770T0	3-hydroxyacyl[acyl-carrier protein] dehydratase [Source:UniProtKB/TREMBL;Acc:A0A162MQ1]	1	42	50	16.58	
	QYA_153891	QYA_153891T0	3-hydroxyacyl-CoA dehydrogenase	1	47.2	351	181.81	
	QYA_151919	QYA_151919T0	3-hydroxyisobutyryl-CoA hydrolase, mitochondrial [Source:UniProtKB/TREMBL;Acc:A0A168ND88]	1	23.5	28	13.63	
	QYA_157539	QYA_157539T0	40S ribosomal protein S1 [Source:UniProtKB/TREMBL;Acc:A0A168MX07]	1	100	2155	1162.14	

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 bring up the pop-up window):

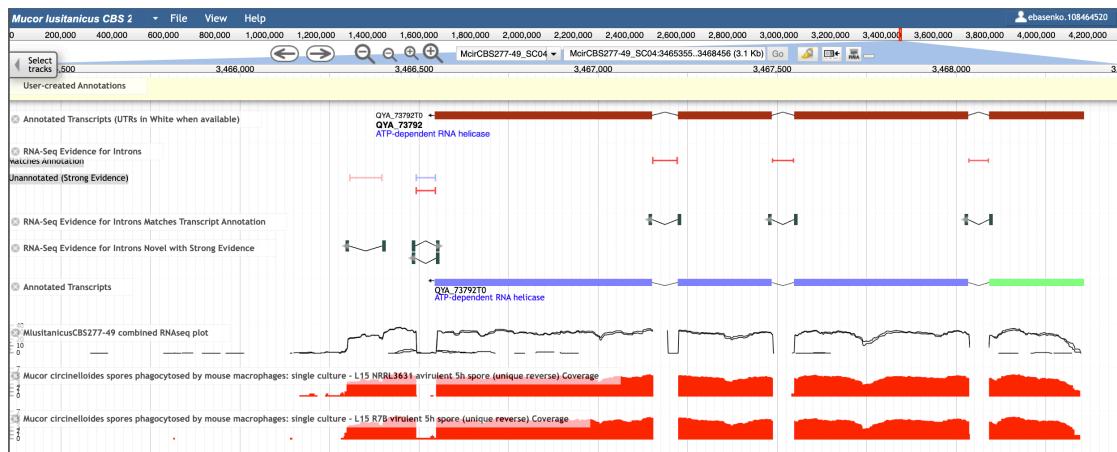
Once in Apollo, use the right panel to select the “Tracks” tab to bring up the following tracks:

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.

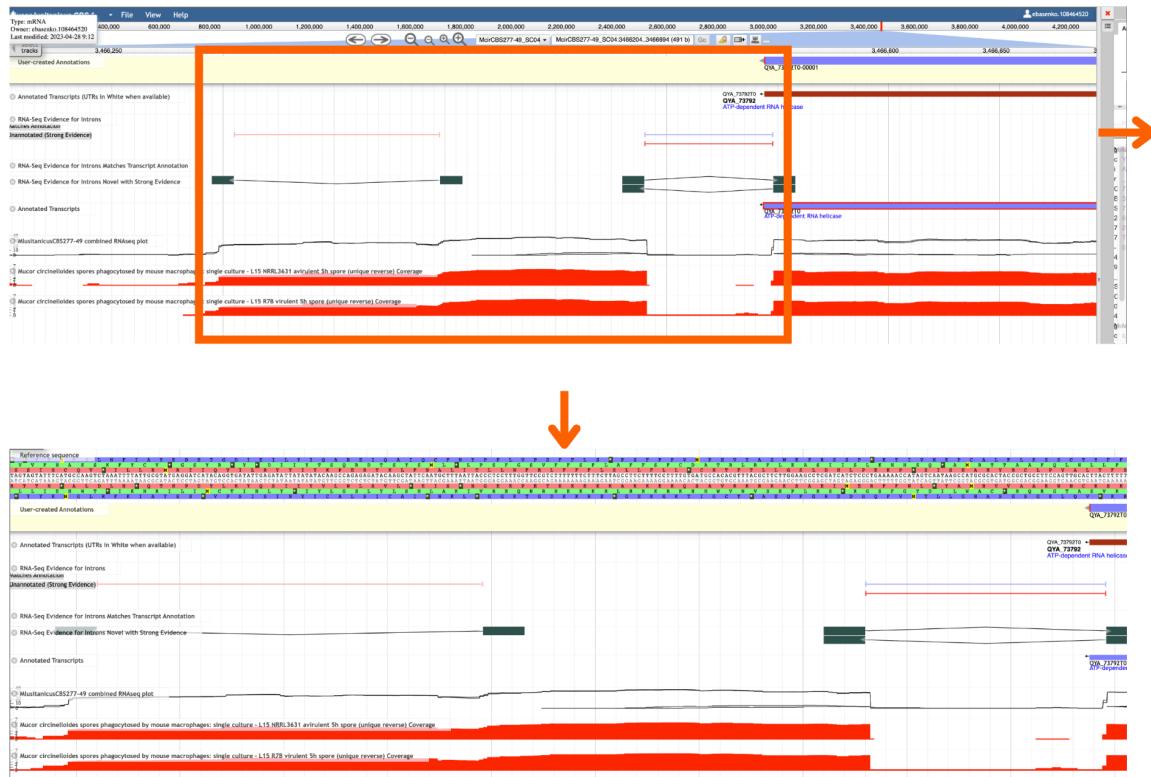
To do this, double click on the gene in the “Annotated Transcripts” tracks which was selected from the “Draggable Annotation” section on the right. Double-clicking will highlight the whole gene rather than an individual component. Note: The “Annotated Transcripts (UTRs in White when available)” track 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:



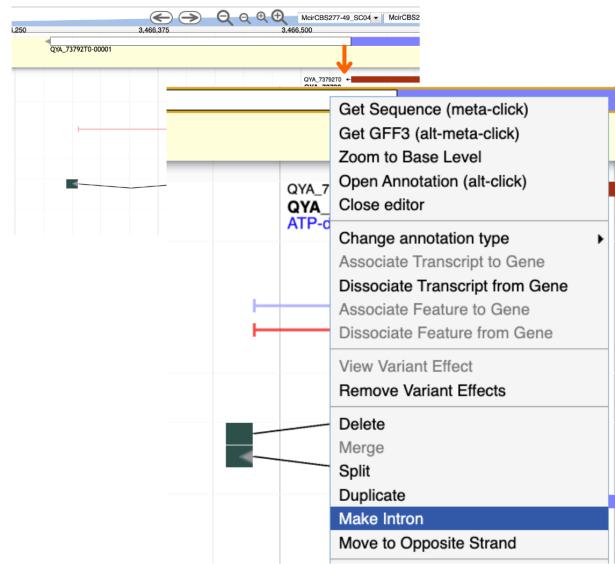
3. Extend gene model using evidence tracks for guidance.

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:



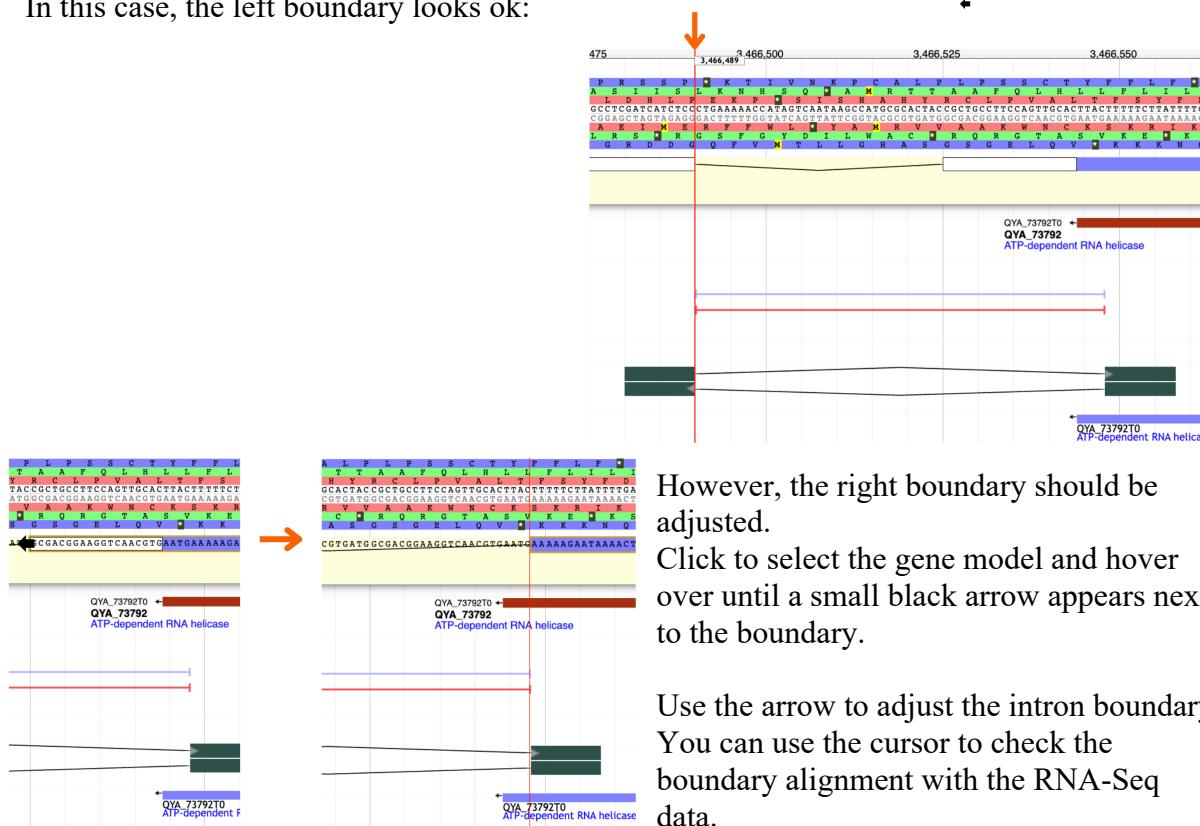
4. Create intron.

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:

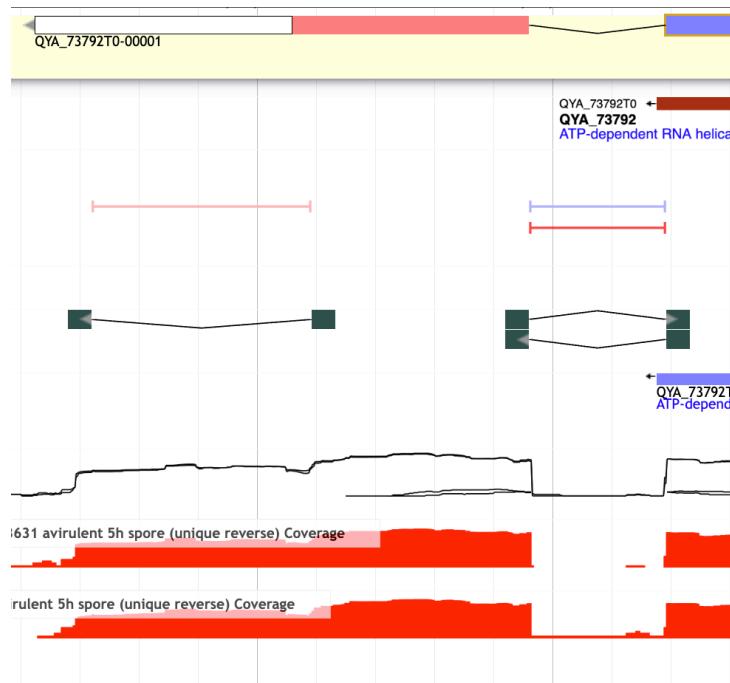


However, the right boundary should be adjusted.

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. Once the status is changed to “Finished” the new gene model will become visible for other users.

Annotations Tracks Ref Sequence Search Organism Users Groups Admin

Show All Show Visible Only

Annotation Name ID All Types GO GP Prov

Reference Sequence All Users All Statuses

Rows 25 1-50 of 1,334

Name	Seq	Type	Length	Updated
QYA_73792T0	MciCBS277-49_SC04	gene	2,076	Apr 28, 2023
MciCBS277-49_SC01aaaaag	MciCBS277-49_SC01	gene	2,435	Jan 03, 2023
QYA_105303T0	MciCBS277-49_SC01	gene	523	Sep 08, 2022
QYA_157431T0	MciCBS277-49_SC10	gene	1,629	Sep 08, 2022
QYA_149982T0	MciCBS277-49_SC10	gene	873	Sep 08, 2022
QYA_149802T0	MciCBS277-49_SC10	gene	3,209	Sep 08, 2022
QYA_14515T0	MciCBS277-49_SC06	gene	4,234	Sep 08, 2022
MciCBS277-49_SC06c	MciCBS277-49_SC06	gene	2,193	Sep 08, 2022
MciCBS277-49_SC06a	MciCBS277-49_SC06	gene	1,879	Sep 08, 2022
QYA_75161T0	MciCBS277-49_SC06	gene	1,011	Sep 08, 2022
MciCBS277-49_SC03aad	MciCBS277-49_SC03	gene	1,502	Sep 07, 2022
QYA_152667T0	MciCBS277-49_SC03	gene	2,578	Sep 07, 2022
MciCBS277-49_SC03aac	MciCBS277-49_SC03	gene	745	Sep 07, 2022
QYA_140074T0	MciCBS277-49_SC03	gene	809	Sep 07, 2022
QYA_93110T0	MciCBS277-49_SC09	gene	5,230	Sep 07, 2022
QYA_75293T0	MciCBS277-49_SC09	gene	2,461	Sep 07, 2022
QYA_15569T0-00001	MciCBS277-49_SC04	gene	1,338	Sep 07, 2022
QYA_15679RT0	MmrRS277-49_SC06	gene	1,912	Sep 07, 2022

Link to annotation Close(x)

Details GO Gene Product Provenance DoXref Comment Attributes

Go ID Sync name with transcript Obsolete Annotations Delete

Type: gene Status: Finished

Name: QYA_73792T0

Symbol:

Aliases (";" separated):

Description:

Location: 3466281 - 3468356 strand()

Ref Sequence: MciCBS277-49_SC04

Owner: ebasenko.108464520

Created: Apr 28, 2023 09:14 AM

Updated: Apr 28, 2023 09:14 AM

SGD Variant Viewer

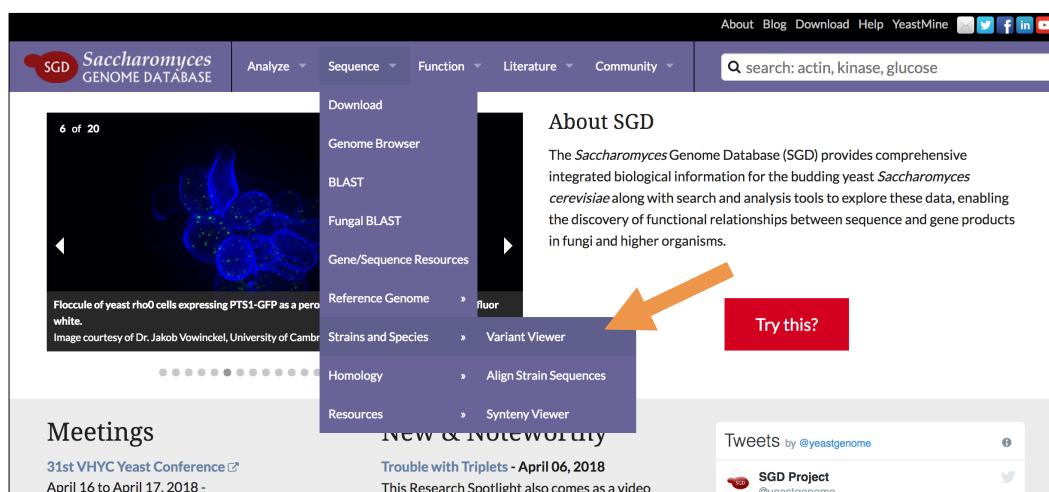
SGD's Variant Viewer (<https://yeastgenome.org/variant-viewer>) is an open-source web application that compares nucleotide and amino acid sequence differences between 12 common *S. cerevisiae* laboratory strains. For a given open reading frame, Variant Viewer breaks down the position and nature of any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

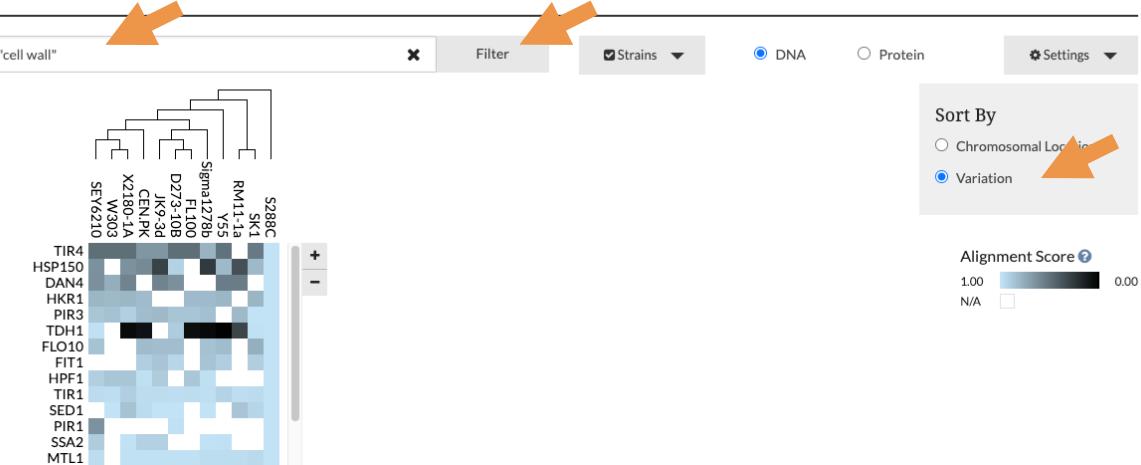
S288C vs. Sigma1278b: Cell Wall

- 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



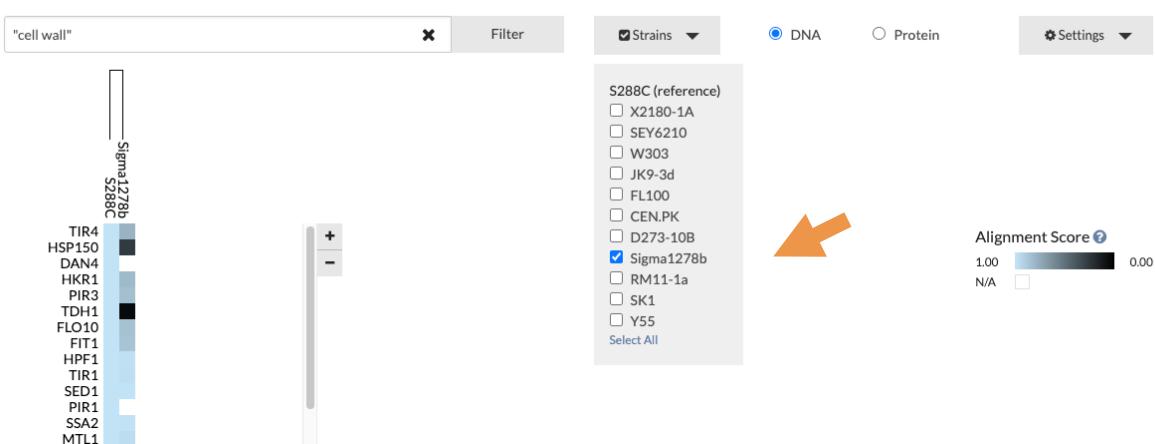
- The **Filter** box accepts one or more genes, as well as Gene Ontology (GO) terms. Because we are interested in genes involved in cell wall development, search for the GO term “**cell wall**,” sort by variation in the settings pull-down, and then click Filter.

Variant Viewer

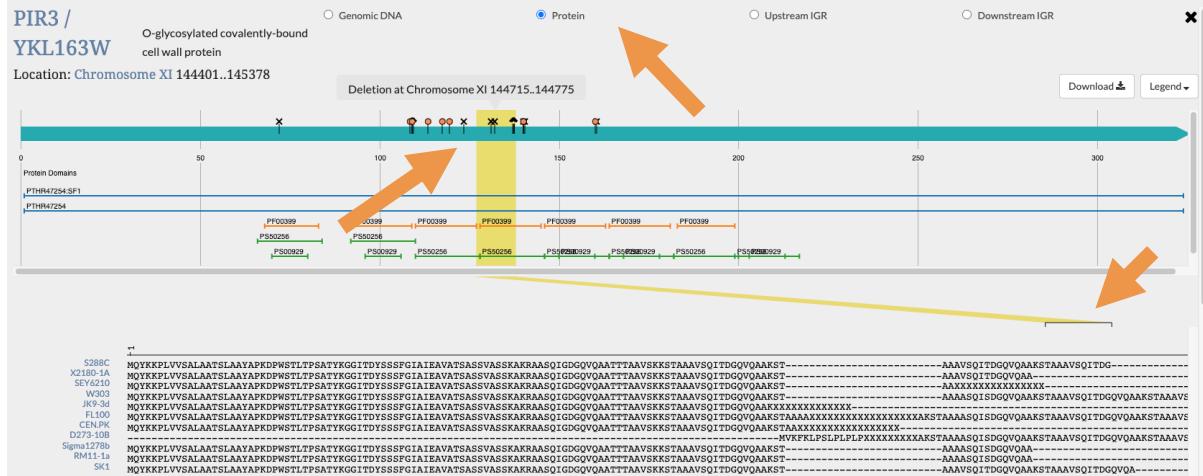


- The **matrix**, shown on the left, will have changed to only include the genes that localize to cell walls.
 - This matrix enables you to visualize high-level differences in multiple genes relative to strain S288C. Each square in the matrix corresponds to one of the twelve strains in Variant Viewer,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

Variant Viewer



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



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

References i 9

1. Toh-e A, et al. (1993) Three yeast genes, PIR1, PIR2 and PIR3, containing internal tandem repeats, are related to each other, and PIR1 and PIR2 are required for tolerance to heat shock. *Yeast* 9(5):481-94 PMID: 8322511
[SGD Paper](#) [DOI full text](#) [PubMed](#)

2. Yun DJ, et al. (1997) Stress proteins on the yeast cell surface determine resistance to osmotin, a plant antifungal protein. *Proc Natl Acad Sci U S A* 94(13):7082-7 PMID: 9192695
[SGD Paper](#) [DOI full text](#) [PMC full text](#) [PubMed](#)

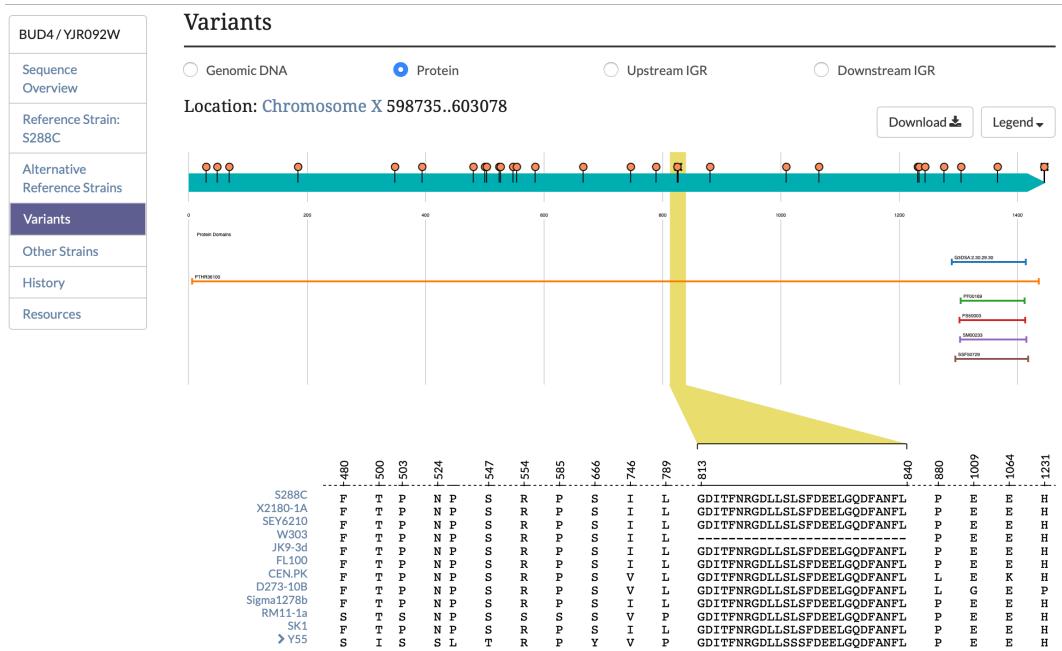
3. Doolin MT, et al. (2001) Overlapping and distinct roles of the duplicated yeast transcription factors Ace2p and Swi5p. *Mol Microbiol* 40(2):422-32 PMID: 11309124
[SGD Paper](#) [DOI full text](#) [PubMed](#)

4. Porter SE, et al. (2002) The yeast pafl-rNA polymerase II complex is required for full expression of a subset of cell cycle-regulated genes. *Eukaryot Cell* 1(5):830-42 PMID: 12455700
[SGD Paper](#) [DOI full text](#) [PMC full text](#) [PubMed](#)

5. Jung US and Levin DE (1999) Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway. *Mol Microbiol* 34(5):1049-57 PMID: 10594829
[SGD Paper](#) [DOI full text](#) [PubMed](#)

Variant Viewer: Sequence Tab

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



FungiDB: SNPs and Population Genetics

Learning Objective:

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

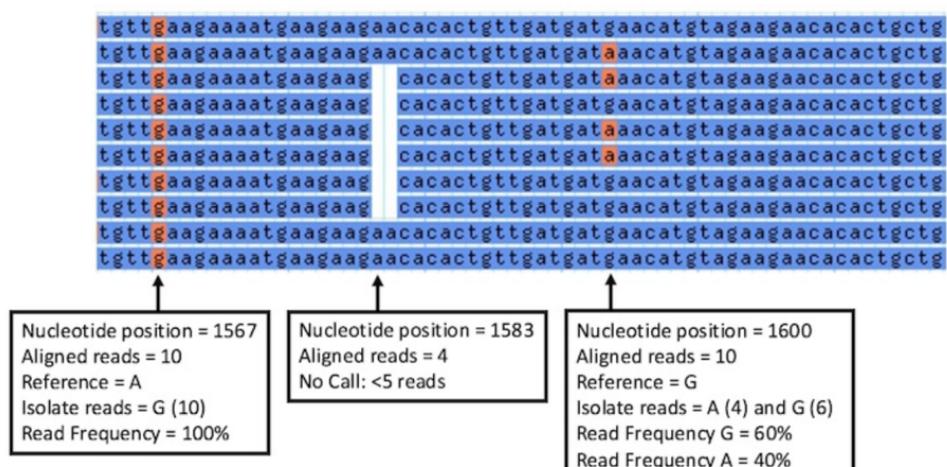
SNPs have different functional effects with most having no consequential effect on gene function. SNPs may directly affect protein function when they are non-synonymous (results in a change in the amino acid; missense) or when they cause a premature stop codon (nonsense). SNPs that do not fall within 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.

Isolate X aligned sequencing reads



Minor allele frequency:

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

Isolate consensus sequences aligned to reference genome.

reference	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
303.1	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTT A TTTTCTACTG
309.1	TGATAA T NCT GGTTTTGTA CTCCACTTC C CAGTGCTTCA TTTTCTACTG
RV_3600	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
RV_3606	TGATAA T NCT GGTTTTGTA CTCCACTTC C CAGTGCTTCA TTTTCTACTG
RV_3610	TGATGATT C GGTTTTGTA CTCCACTTC C CAGTGCTTCA TTTTCTACTG
SenT119.09	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
SenT123.09	TGATRAT T CT GGTTTTGTA CTCCACTTC C CAGTGCTTCA TTTTCTACTG
SenT140.08	TGGTGATACT GGTTTTGTA CTCCACTTC C CAGTGCTTCA TTTTCTACTG
SenT142.09	TGGTGATACT GGTTTTGTA CTCCACTTC C CAGTGCTTCA TTTTCTACTG
SenT175.08	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTT A TTTTCTACTG

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

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

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

Percent isolates with a base call:

Sometimes an isolate does not have a base call at a certain nucleotide position because the Read Frequency Threshold was not met or because there were less than 5 aligned sequencing reads for that nucleotide position. In this case, 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 Genes based on SNP Characteristics

Configure Search Learn More View Data Sets Used

Reset values to default

Organism: Pyricularia oryzae 70-5

Set of Samples

41 of 81 Set of Samples selected (Data Set)

Data Set	Remaining Set of Samples	Set of Samples	Distribution	%
Pyricularia oryzae 70-15 Genome Sequence and Annotation	1 (1%)	1 (1%)	1	(100%)
SNP calls on WGS of Magnaporthe field-isolates.	13 (16%)	13 (16%)	13	(100%)
SNP calls on WGS of Pyricularia oryzae isolated from Bangladesh in 2016 and 2017	23 (28%)	23 (28%)	23	(100%)
SNP calls on WGS of Pyricularia oryzae isolates from different hosts	3 (4%)	3 (4%)	3	(100%)
SNP calls on WGS of Pyricularia oryzae isolates from Zambia	13 (16%)	13 (16%)	13	(100%)
SNPs calls on WGS data of Pyricularia oryzae isolates from Africa	28 (35%)	28 (35%)	28	(100%)

Read frequency threshold: 80%

Minor allele frequency >= 0

Percent isolates with a base call >= 20

SNP Class: Non-Synonymous

Number of SNPs of above class >= 1

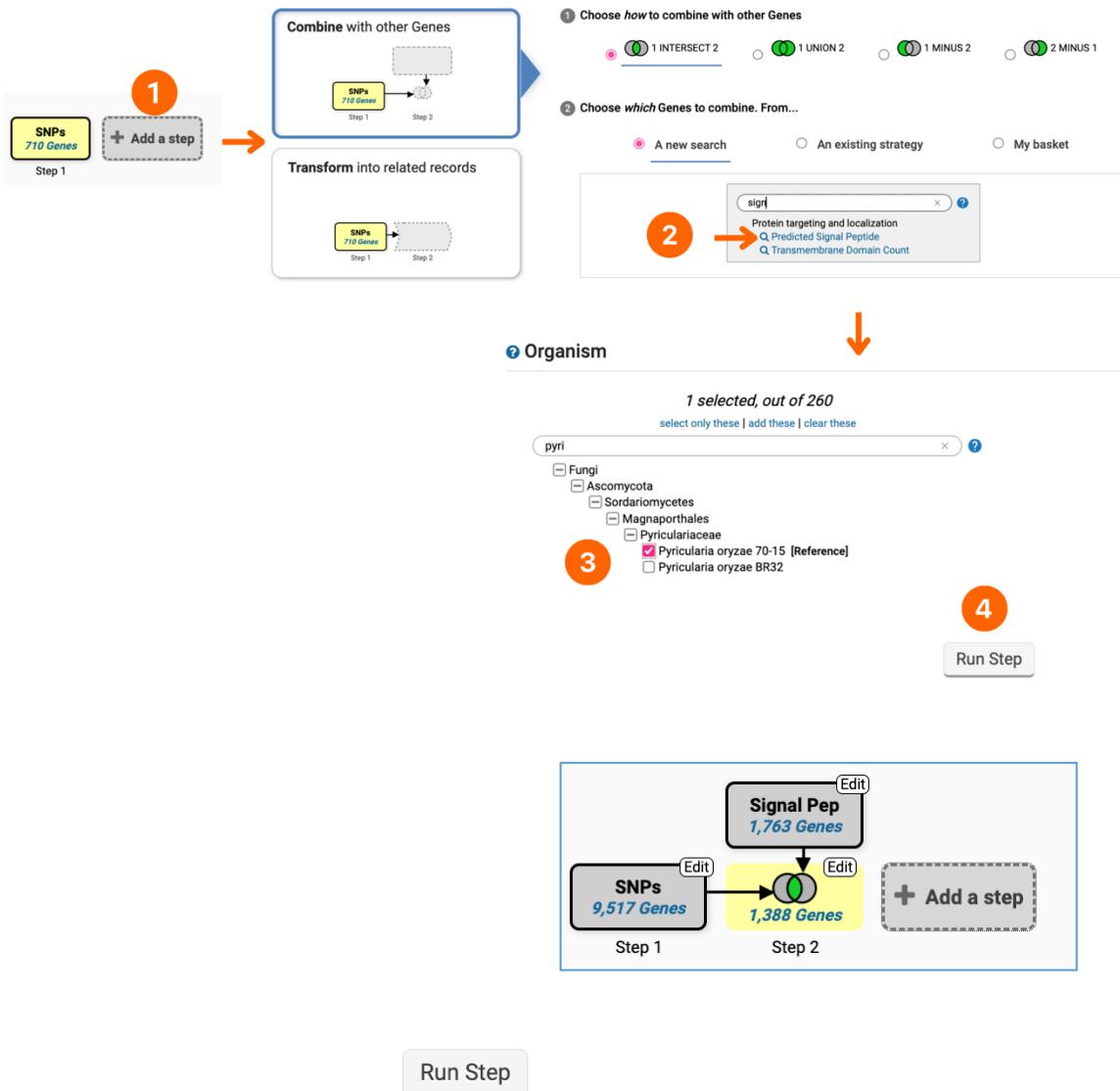
Step 1

Edit

Add a step

- Identify putative nuclear effectors based on the presence of both a secretion signal and the DNA-binding domains IPR007219 or IPR009071 .

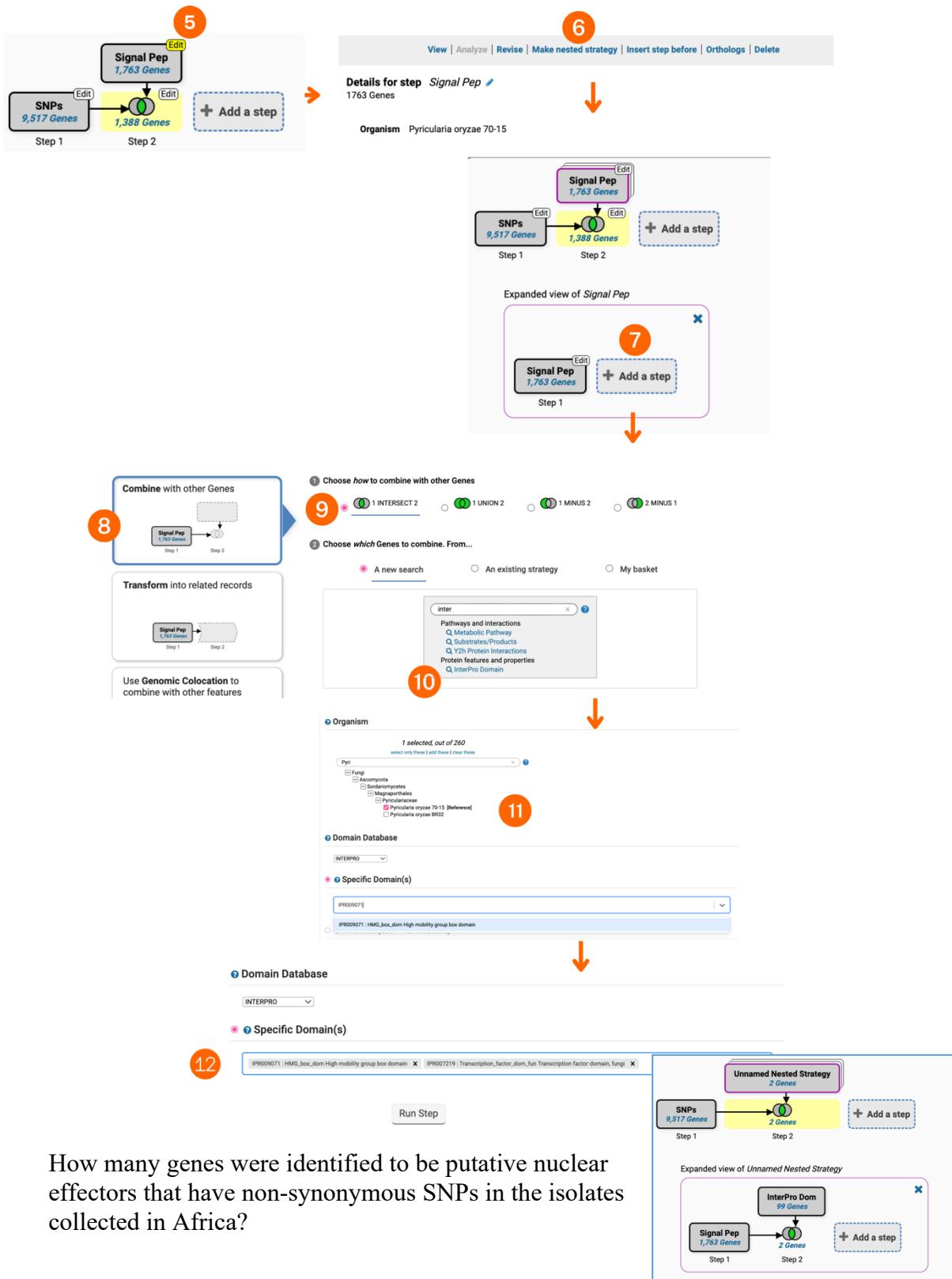
1. Click on the “Add a Step” button.
2. Use the “Combine with Other Genes” option to deploy the “Predicted Signal Peptide” search.
3. Set the genome to *Pyricularia oryzae* 70-50.
4. Click on the “Run Step” button.



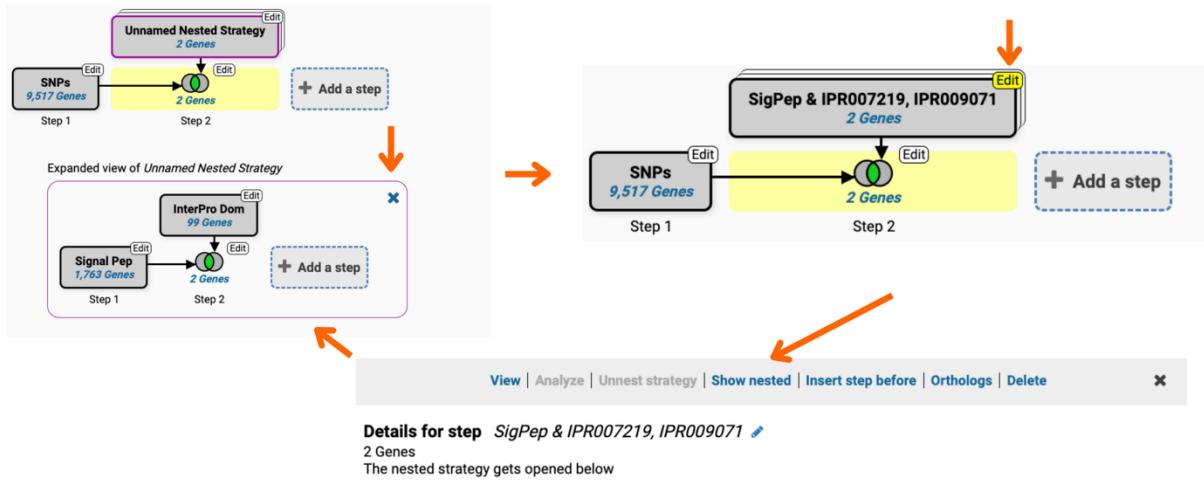
Note that currently, our strategy returns genes that have at least 1 SNP and also a predicted signal peptide domain. How can we identify that that have at least 1 SNP and ALSO a predicted signal peptide domain AND a DNA-binding domain? (Hint: create a nested strategy as described below).

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.



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



Strategy URL:

<https://fungidb.org/fungidb/app/workspace/strategies/import/bd657f5629cac5df>

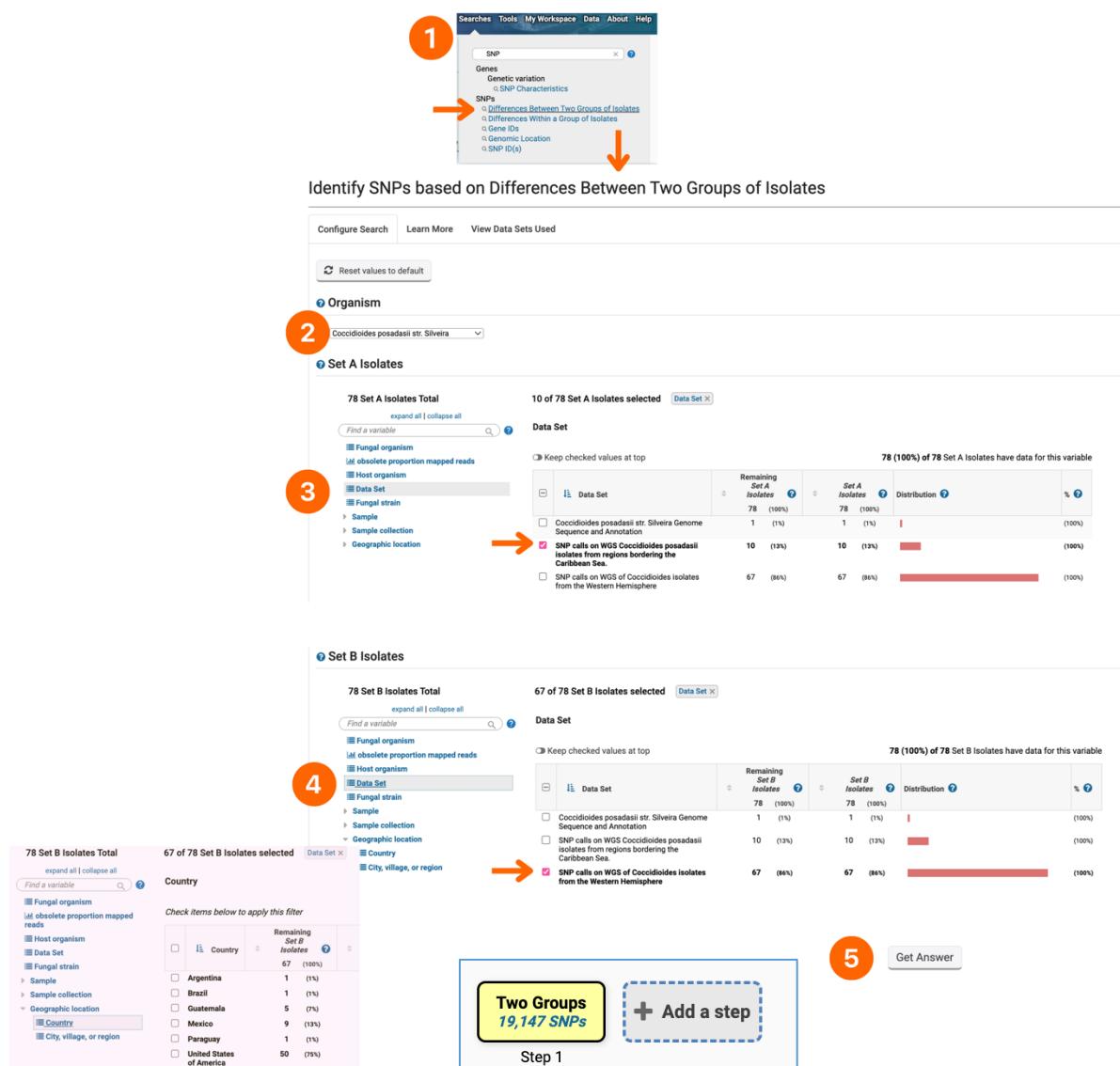
References: <https://www.nature.com/articles/s41467-020-19624-w>

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

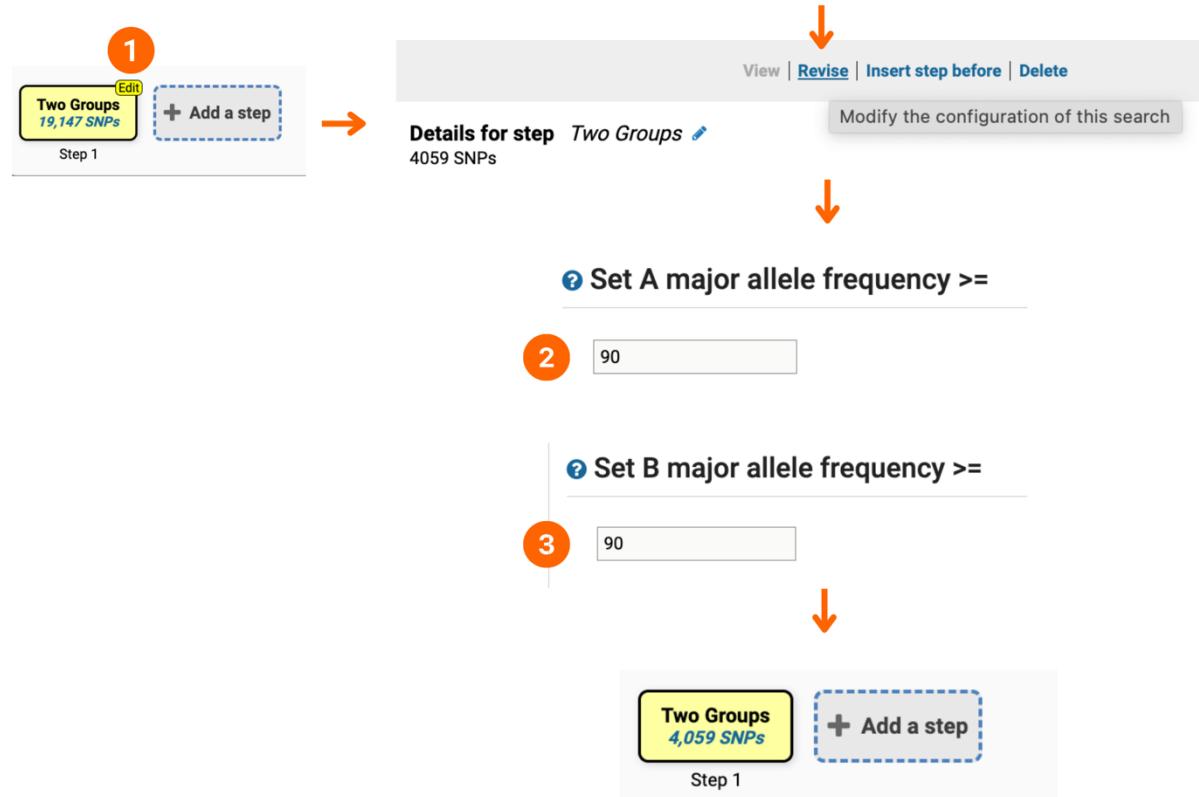
Coccidioidomycosis, also known as Valley fever, is caused by two closely related species – *C. immitis* and *C. posadasii*. The disease is associated with high morbidity and mortality rates that affects tens of thousands of people each year. The two fungal species are endemic to several regions in the 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 metadata (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 $\geq 90\%$



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_10216	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	CPSP_10216	738	T	100	T	C	93.3	A
NGS_SNP:GL636536.1541	GL636536: 1,541	CPSP_10216	753	G	100	A	A	95.8	V
NGS_SNP:GL636536.13558	GL636536: 13,558	CPSP_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: TCTGAGACTTTATTCTGGTTGCTTCCTTC

CCTTCCCTGTCCCTCCAGTTGTTGAATGAAT

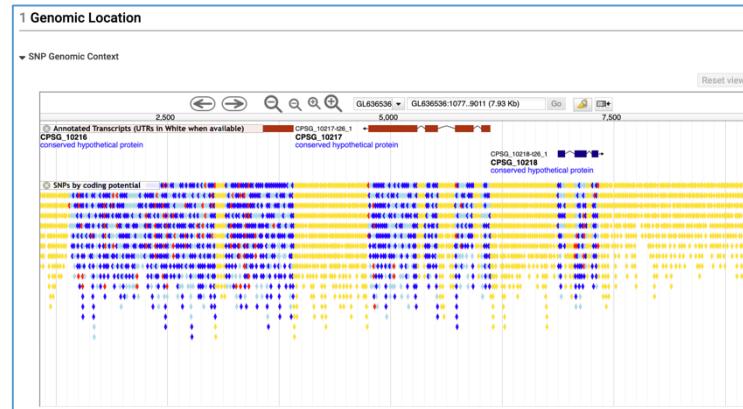
SNP context (gene strand): ATTCAATCACAACTGGAAGCAGGGAAG

GAAAGAGAAGCAACCAGAACATAAGTCTCAGA

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

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

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



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

Select output options:

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

Select strains:

78 Reference Samples Total 53 of 78 Reference Samples selected Country

expand all | collapse all Find a variable

Country		Remaining Reference Samples		Distribution		%
	Count	Count	(%)	Count	(%)	(%)
<input checked="" type="checkbox"/> Argentina	1 (1%)	1 (1%)		1 (1%)		(100%)
<input type="checkbox"/> Brazil	1 (1%)	1 (1%)		1 (1%)		(100%)
<input type="checkbox"/> Guatemala	5 (6%)	5 (6%)		5 (6%)		(100%)
<input type="checkbox"/> Mexico	10 (13%)	10 (13%)		10 (13%)		(100%)
<input type="checkbox"/> Paraguay	1 (1%)	1 (1%)		1 (1%)		(100%)
<input checked="" type="checkbox"/> United States of America	52 (68%)	52 (68%)		52 (68%)		(100%)
<input type="checkbox"/> Venezuela	7 (9%)	7 (9%)		7 (9%)		(100%)

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

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

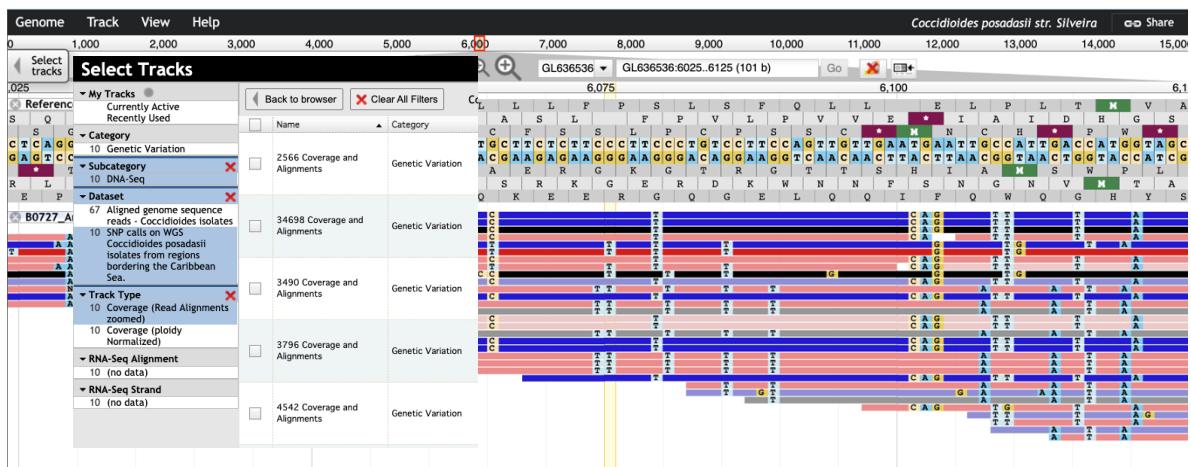
Search this table... [?](#)

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.

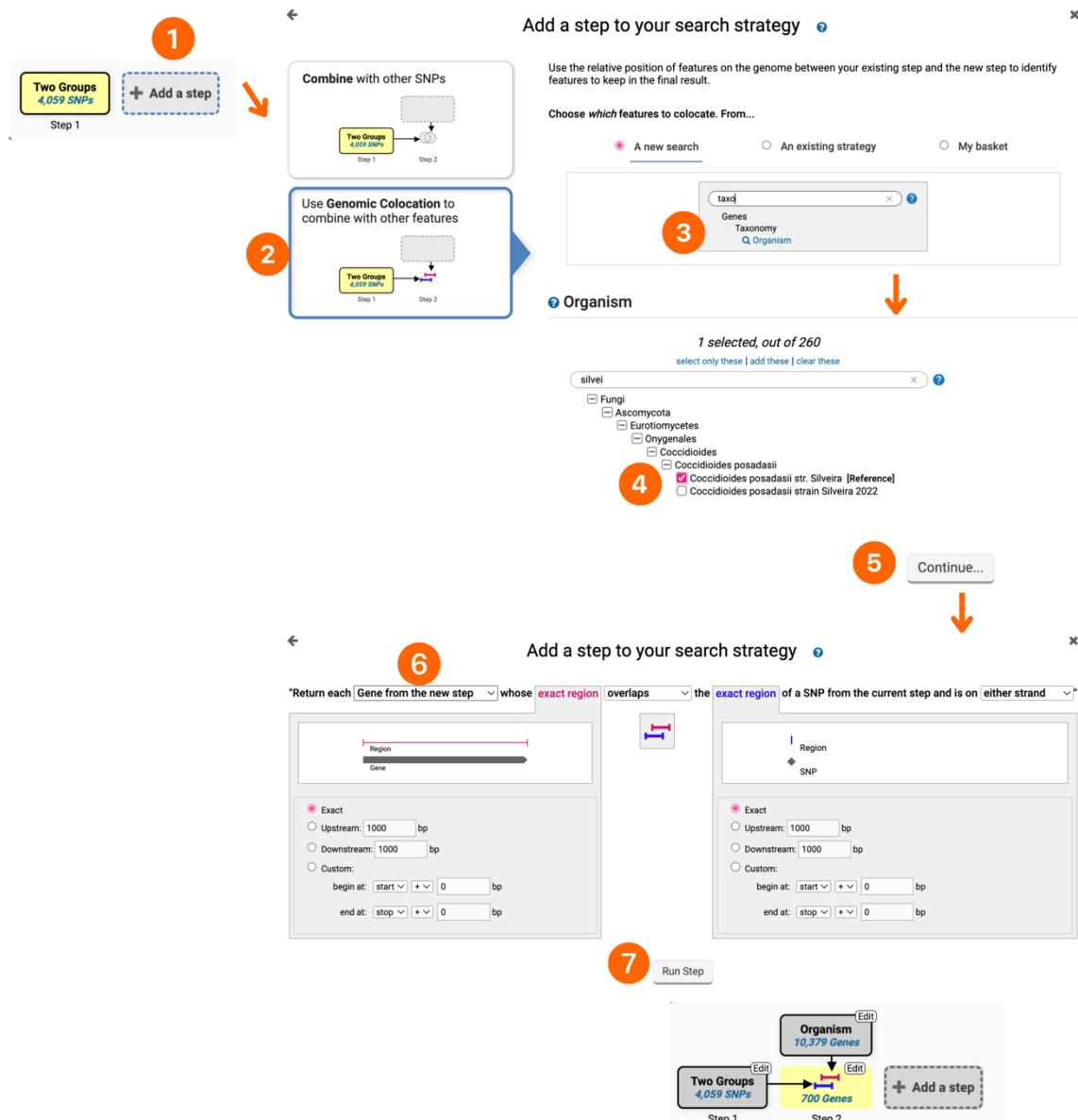
Venezuela	JTORRES	EUSMPL0102-1-7	C	G	C	75	100	view DNA-seq reads
-----------	---------	----------------	---	---	---	----	-----	------------------------------------

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:

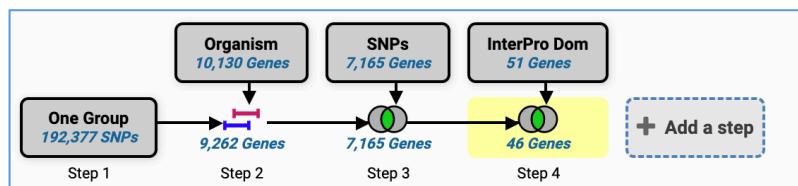
<https://fungidb.org/fungidb/app/workspace/strategies/import/d9d0fff2dbda229d>

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 screenshot shows the BioNumerics software interface with the following steps highlighted:

- Search Bar:** The search bar contains "plo". A circled "1" is next to it.
- Organism Selection:** The dropdown menu shows "Candida albicans SC5314". A circled "2" is next to it.
- Data Set Selection:** The "Data Set" section is expanded, showing various datasets. A circled "3" is next to the "Data Set" heading.
- Dataset Selection:** The "SNP calls on WGS of Candida albicans clinical isolates (oropharyngeal candidiasis)" dataset is selected, indicated by a checked checkbox. A circled "4" is next to it.
- Copy Number Input:** The "Copy Number >=" input field contains the value "3". A circled "5" is next to it.
- Output Selection:** The "Median Or By Strain/Sample?" dropdown is set to "By Strain/Sample (at least one selected strain/sample meets criteria)". A circled "6" is next to it.

Get Answer

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).

Genomic Sequence Results			
Genomic Sequence Results Rows per page: 1000 <input type="button" value="▼"/> Download Add to Basket Add Columns			
	Sequence ID	Median Copy No (All Selected Samples)	Strains/Samples Meeting Criteria
	Ca22chr3A_C_albicans_SC5314	2	Candida_albicans_TWTC6
	Ca22chr4A_C_albicans_SC5314	2	Candida_albicans_1649, Candida_albicans_2501, Candida_albicans_3731, Candida_albicans_5106
	Ca22chr5A_C_albicans_SC5314	2	Candida_albicans_1619, Candida_albicans_1649, Candida_albicans_2823, Candida_albicans_3034, Candi...
	Ca22chr6A_C_albicans_SC5314	2	Candida_albicans_TWTC8

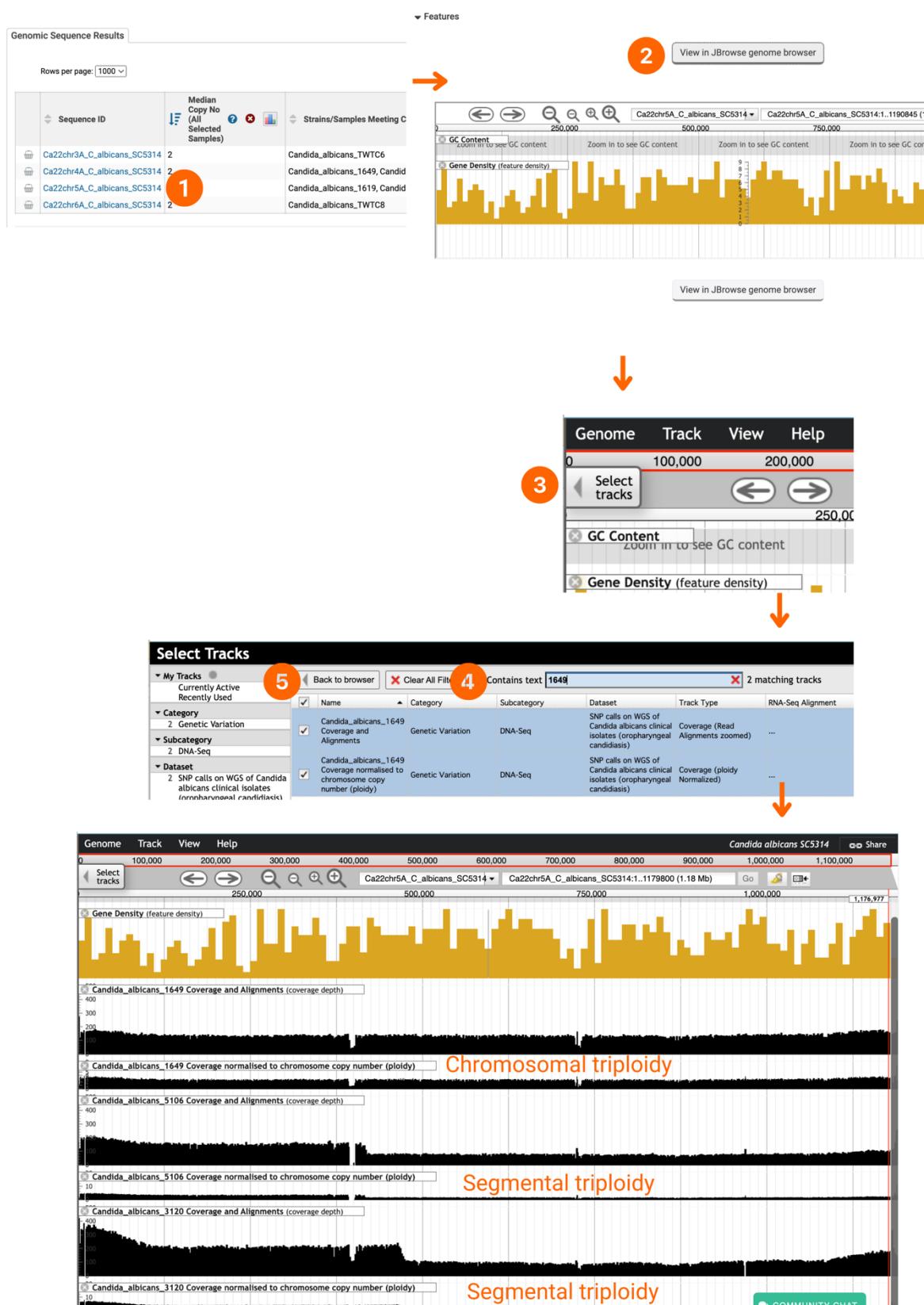
- **Explore segmental aneuploidy in JBrowse.**

JBrowse has two coverage tracks:

- Raw coverage from the alignment (available for every isolate where we have whole genome sequencing, whether we ran the copy number pipeline or not)
- Normalized coverage in bins (only available for isolates where we have run the copy number pipeline)

1. Click on one of the Sequence ID Ca22chr5A_C_albicans_SC5314 (in blue).
2. Navigate to JBrowse by clicking on the “View in JBrowse genome browser” button.
3. When in JBrowse, click on the Select tracks tab to customize your view.
4. Use the “Contains text” filter to identify and select tracks for the following isolates: 1649, 5106, and 3120.
5. Click on the “Back to browse” tab to return to JBrowse view with selected tracks.

▼ 4.2 Sequence sites, features and motifs



Notice examples of chromosomal (1649) and segmental triploidy (5106 and 3120). 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=](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>

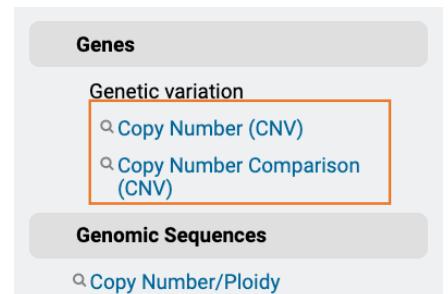
References: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383195/>

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 “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.



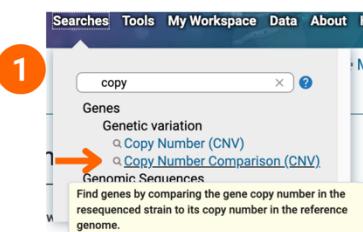
Different metrics for defining copy number:

- **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.



Identify Genes based on Copy Number Comparison (CNV)

Configure Search Learn More View Data Sets Used

Reset values to default

Organism

2

Strain/Sample

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

expand all | collapse all

	Fungal	Remaining Strain/Sam...	Strain/Sam...	Distribution	%
5106	4	262 (100%)	262 (100%)		(100%)
<input checked="" type="checkbox"/> Candida albicans 5106	1 (< 1%)	1 (< 1%)			

Rows per page: 100

3

Median Or By Strain/Sample?

5

What comparison do you want to make?

6

Get Answer

CopyNumberComparison
520 Genes

+ Add a step

Step 1

Examine the results using the Genome View option.

The screenshot shows a user interface for examining genomic data. At the top, there are three tabs: "Gene Results", "Genome View" (which is highlighted with a red arrow), and "Analyze Results". Below the tabs is a legend indicating "Genes on forward strand" (blue) and "Genes on reversed strand" (red). A search bar and a "Rows per page" dropdown set to 20 are also present. The main area displays a table of gene results:

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

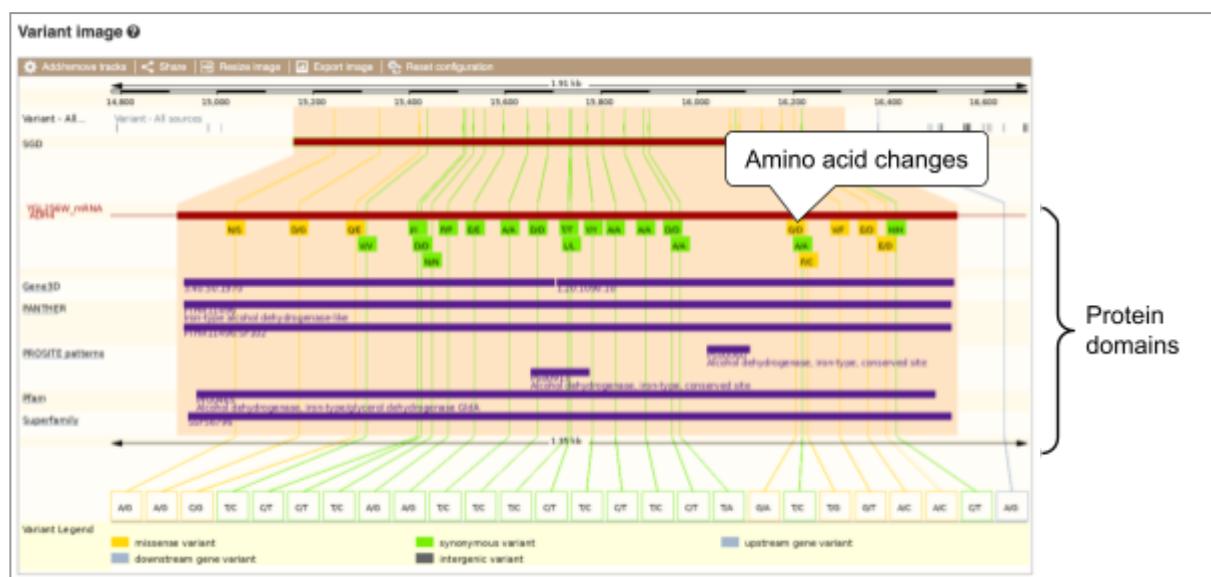
Exercise: Exploring variants in Ensembl Fungi

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 at the 5' upstream region).

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

You can filter the table to view variants that alter the protein sequence. Click on the [Consequences: All](#) button above the table. Click the option '[PTV and Missense](#)' in the pop

up, then [Apply](#). You can also filter by other columns such as variant [Class](#).

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

The SIFT scores 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.

The screenshot shows a dialog box titled "Consequences: All" with a "Filter Other Columns" button. It lists 30 variant types, each with a color-coded SIFT score (Off, On, or Off). The variants include transcript ablation, splice donor variant, splice acceptor variant, splice donor 5th base variant, stop gained, frameshift variant, stop lost, start lost, transcript amplification, inframe insertion, inframe deletion, protein altering variant, missense variant, splice region variant, splice polypyrimidine tract variant, incomplete terminal codon variant, splice donor region variant, synonymous variant, start retained variant, and stop retained variant. Buttons at the bottom include "Turn All Off", "PTV", "PTV & Missense", "Only Exonic", "Turn All On", "Apply", and "Cancel". Below the dialog are genomic coordinates VII:15518, C/T, SNP, SGRP, and a synonymous D.

Variant Type	SIFT Score
transcript ablation	Off
splice donor variant	On
splice acceptor variant	On
splice donor 5th base variant	Off
stop gained	On
frameshift variant	On
stop lost	Off
start lost	Off
transcript amplification	Off
inframe insertion	Off
inframe deletion	Off
protein altering variant	Off
missense variant	On
splice region variant	Off
splice polypyrimidine tract variant	Off
incomplete terminal codon variant	Off
splice donor region variant	Off
synonymous variant	Off
start retained variant	Off
stop retained variant	Off
coding sequence variant	Off
mature miRNA variant	Off
5 prime UTR variant	Off
3 prime UTR variant	Off
non coding transcript exon variant	Off
intron variant	Off
NMD transcript variant	Off
non coding transcript variant	Off
upstream gene variant	Off
downstream gene variant	Off

The screenshot shows a table with "Table filters" and "Protein pathogenicity predictions" applied. The table includes columns for Variant ID, Chr:bp, Alleles, Class, Source, Evidence, Clin. Sig., Conseq_Type, AA, AA coord, SIFT, and Transcript. A callout points to the "Variant IDs are links to variant tab" and another points to the "Protein pathogenicity predictions" column. The SIFT column shows values like 1, 0.72, 0.1, 0, 0.03, 0.26, and 0.67.

Variant ID	Chr:bp	Alleles	Class	Source	Evidence	Clin. Sig.	Conseq_Type	AA	AA coord	SIFT	Transcript
s07-15244	VII:15244	A/G	SNP	SGRP	-	-	missense variant	N/S	29	1	YGL256W_mRNA
s07-15337	VII:15337	A/G	SNP	SGRP	-	-	missense variant	D/G	60	1	YGL256W_mRNA
s07-15420	VII:15420	C/G	SNP	SGRP	-	-	missense variant	G/E	88	0.72	YGL256W_mRNA
s07-16069	VII:16069	G/A	SNP	SGRP	-	-	missense variant	G/D	304	0.1	YGL256W_mRNA
s07-16087	VII:16087	T/G	SNP	SGRP	-	-	missense variant	F/C	310	0	YGL256W_mRNA
s07-16134	VII:16134	A/T	SNP	SGRP	-	-	missense variant	V/F	326	0.03	YGL256W_mRNA
s07-16175	VII:16175	A/T	SNP	SGRP	-	-	missense variant	E/D	339	0.26	YGL256W_mRNA
s07-16202	VII:16202	A/C	SNP	SGRP	-	-	missense variant	E/D	348	0.67	YGL256W_mRNA

Let's have a look at a specific variant. Click on the top result in the filtered table, or search for [s07-15244](#). This will open up the variation tab.

Variant displays

- Explore this variant
- Genomic context
- Flanking sequence
- Genotype frequency
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- 3D Protein model

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

s07-15244 SNP

Most severe consequence: missense variant | See all predicted consequences

Alleles: A/G | Highest population MAF: 0.46

Location: Chromosome VII:15244 (forward strand) | VCF: VII: 15244 s07-15244 A G

HGVS names: This variant has 3 HGVS names - Show

External Links: Variation features from SGRP, with Ensembl identifiers | About SGRP

Original source: This variant overlaps 1 transcript and has 18 sample phenotypes.

About this variant

Explore this variant

Genomic context, Genes and regulation, Flanking sequence, Population genetics, Phenotype data, Sample genotypes, Linkage disequilibrium, Phylogenetic context, Citations

Variation icons (these go to the same places as the links in the left-hand navigation panel)

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.

Gene	Transcript (strand)	Allele (Tr. allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	AA	Codons	SIFT	Detail
YGL256W	YGL256W_mRNA (+)	G (G)	missense variant	86 (out of 1148)	86 (out of 1148)	29 (out of 382)	N/S	AAC/AGC	1	Show

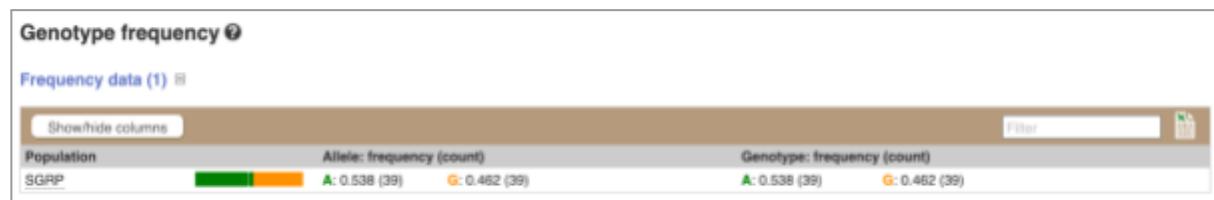
No overlap with Ensembl Regulatory features

No overlap with Ensembl Motif features

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

(f) Which allele is major in the 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?

tmp_10_6610 SNP

Most severe consequence: upstream gene variant | See all predicted consequences

Alleles: C/T | Highest population MAF: 0.15

Location: Chromosome 10:6610 (forward strand) | VCF: 10 6610 tmp_10_6610 C T

HGVS name: 10:g.6610C>T

External Links

Original source

About this variant

This variant overlaps 4 transcripts and has 10 sample genotypes.

- (b) What is the most frequent allele at this position? How many heterozygous individuals were observed in the melonis population?

Genotype frequency

Frequency data (1) ▾

Show/hide columns Filter

Population	Allele: frequency (count)	Genotype: frequency (count)
melonis	C: 0.850 (17) T: 0.150 (3)	CC: 0.800 (8) CT: 0.100 (1) TT: 0.100 (1)

- (c) Which individuals have got genotypes C|T and T|T?

Sample genotypes

Search for a sample: Search (e.g. NA18507) [back to top]

Genotypes for melonis ▾

Show/hide columns Filter

Sample (Male/Female/Unknown)	Genotype (forward strand)	Population(s)	Father	Mother
86939 (U)	CC	melonis	-	-
86944 (U)	CC	melonis	-	-
86945 (U)	CC	melonis	-	-
86946 (U)	CC	melonis	-	-
86947 (U)	CC	melonis	-	-
86948 (U)	CC	melonis	-	-
86949 (U)	CC	melonis	-	-
909453 (U)	CC	melonis	-	-
909454 (U)	CT	melonis	-	-
909455 (U)	TT	melons	-	-

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

Variant Effect Predictor ⓘ

New job Clear form

Species: *Verticillium dahliae* JR2 ⓘ Assembly: VDAg_JR2v4.0 Change species

Name for this job (optional): Name your job

Input data: Paste or type in data... Either paste data: See preview of the results for a selected line

Examples: Ensembl default, VCF Variant identifiers, HGVS notations Run instant VEP for current line ⓘ

...or upload a file... Or upload file: Choose file No file chosen See data format examples

...or provide a URL to a file hosted online Or provide file URL:

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.

The screenshot shows a browser window titled "Instant results for 5 701484 701484 C/A". A yellow header bar says "Instant VEP". Below it, a message states: "The below is a preview of results using the *Verticillium dahliae* Jr2 Ensembl transcript database and does not include all data fields present in the full results set. To obtain these please close this preview window and submit the job using the Run button below." The main content area displays a table of variants:

Gene/Feature/Type	Consequence	Details
VDAG_JR2_Chr5g02160a: VDAG_JR2_Chr5g02160a-00001 Type: protein_coding	downstream_gene_variant	Distance to transcript: 2165bp
VDAG_JR2_Chr5g02170a: VDAG_JR2_Chr5g02170a-00001 Type: protein_coding	downstream_gene_variant	Distance to transcript: 742bp
VDAG_JR2_Chr5g02170a: VDAG_JR2_Chr5g02170a-00002 Type: protein_coding	downstream_gene_variant	Distance to transcript: 778bp
VDAG_JR2_Chr5g02171a: VDAG_JR2_Chr5g02171a-00001 Type: protein_coding	upstream_gene_variant	Distance to transcript: 64bp

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

The screenshot shows the "Identifiers" and "Variants and frequency data" sections of the VEP configuration interface. A callout bubble points to the "Find co-located known variants" dropdown, which is set to "Yes". Another callout bubble points to the "Variant synonyms" checkbox.

Identifiers

- Gene symbol:
- Transcript version: Which identifiers do you want in the output?
- Protein:
- UniProt:
- HOVS:

Variants and frequency data

- Find co-located known variants: Yes
- Variant synonyms: Does this variant already exist?
- Include flagged variants:

Additional annotations Additional transcript, protein and regulatory annotations

Transcript annotation

Transcript biotype: Add information about affected transcripts and proteins

Exon and intron numbers:

Identify canonical transcripts:

Upstream/Downstream distance (bp): 5000

mRNA structure:

NMD:

UTRAnnotator:

Protein annotation

Protein matches:

IntAct: Disabled
Enabled

Predictions Variant predictions, e.g. SIFT, PolyPhen

Splicing predictions

dbSCSNV:

MaxEntScan:

SpliceAI: Disabled
Enabled

Conservation

BLOSUM62:

Ancestral allele:

Phenotype data and citations

Phenotypes:

Gene Ontology:

DiseaseNET:

Mastermind:

Filtering options Pre-filter results by frequency or consequence type

Filters

Return results for variants in coding regions only: Show only coding variants

Restrict results: Show all results ▾ More filter

NB: Restricting results may exclude biologically important variants!

Advanced options Additional enhancements

Run: Run VEP

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

This will count down and refresh the page every 10 seconds

Click here to view results

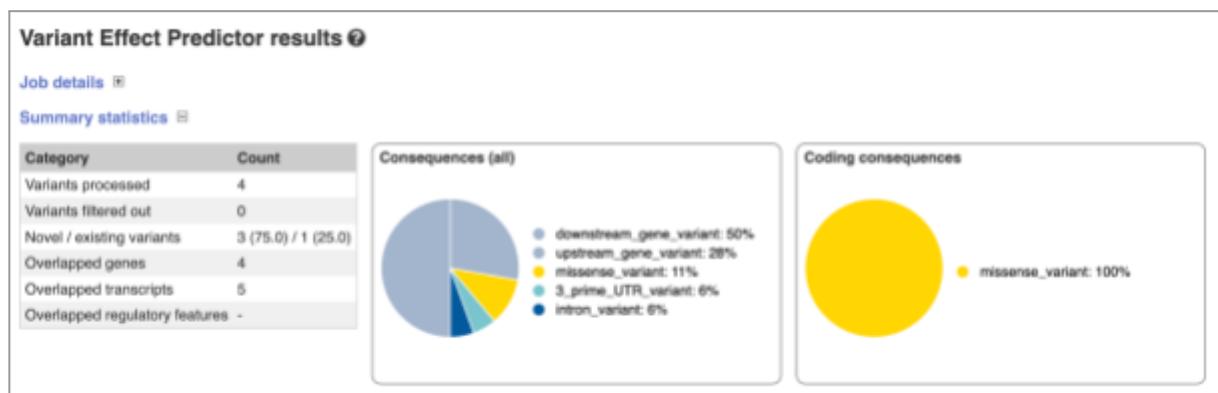
Options to save, edit, share or delete the job

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

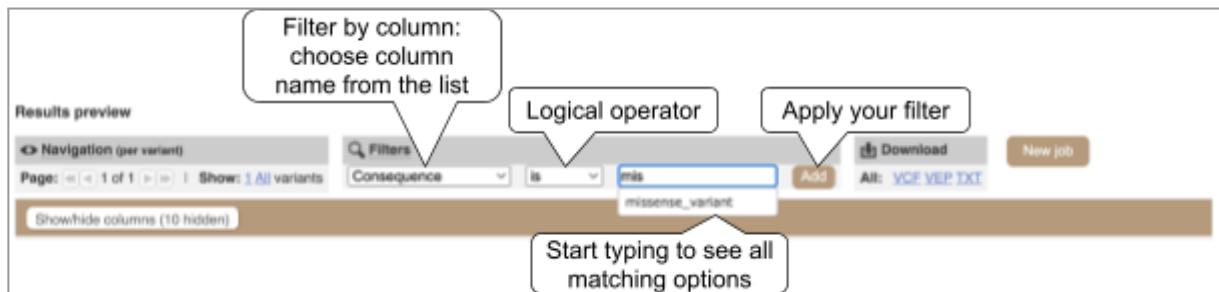
Click **View results** once your job is done. In your results you will see a graphical summary of your data, as well as a table of your results.

Let's come back to our questions:

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



The 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.

Results preview

Navigation (per variant) | Filters | Show: 1 All variants | Uploaded variant | is | defined | Add | Download | New job

Show/hide columns (10 hidden)

Variant 1

Uploaded variant	Location	Allele	Type	Description	Gene	Protein	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand
5_698711_C/G	5_698711-698711	G	downstream_gene_variant	VDAG_JR2_Chromosome 5: 2150a Transcript	VDAG_JR2_Chromosome 5: 2150a	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5_698711-698711	G	intron_variant	VDAG_JR2_Chromosome 5: 2160a Transcript	VDAG_JR2_Chromosome 5: 2160a	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5_698711-698711	G	upstream_gene_variant	VDAG_JR2_Chromosome 5: 2170a Transcript	VDAG_JR2_Chromosome 5: 2170a	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5_698711-698711	G	upstream_gene_variant	VDAG_JR2_Chromosome 5: 2170a Transcript	VDAG_JR2_Chromosome 5: 2170a	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5_698711-698711	G	downstream_gene_variant	VDAG_JR2_Chromosome 5: 2171a Transcript	VDAG_JR2_Chromosome 5: 2171a	protein_coding	-	-	-	-	-	-	-	-1

Variant 2

Uploaded variant	Location	Allele	Type	Description	Gene	Protein	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand
5_698835_G/T	5_698835-698835	T	downstream_gene_variant	VDAG_JR2_Chromosome 5: 2150a Transcript	VDAG_JR2_Chromosome 5: 2150a	protein_coding	-	-	-	-	-	-	-	1
5_698835_G/T	5_698835-698835	T	3_prime_UTR_variant	VDAG_JR2_Chromosome 5: 2160a Transcript	VDAG_JR2_Chromosome 5: 2160a	protein_coding	8/8	1679	-	-	-	-	-	1
5_698835_G/T	5_698835-698835	T	upstream_gene_variant	VDAG_JR2_Chromosome 5: 2170a Transcript	VDAG_JR2_Chromosome 5: 2170a	protein_coding	-	-	-	-	-	-	-	1
5_698835_G/T	5_698835-698835	T	upstream_gene_variant	VDAG_JR2_Chromosome 5: 2170a Transcript	VDAG_JR2_Chromosome 5: 2170a	protein_coding	-	-	-	-	-	-	-	1
5_698835_G/T	5_698835-698835	T	downstream_gene_variant	VDAG_JR2_Chromosome 5: 2171a Transcript	VDAG_JR2_Chromosome 5: 2171a	protein_coding	-	-	-	-	-	-	-	-1

Variant 3

Uploaded variant	Location	Allele	Type	Description	Gene	Protein	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand
5_700313_G/A	5_700313-700313	A	downstream_gene_variant	VDAG_JR2_Chromosome 5: 2150a Transcript	VDAG_JR2_Chromosome 5: 2160a	protein_coding	-	-	-	-	-	-	-	1
5_700313_G/A	5_700313-700313	A	missense_variant	VDAG_JR2_Chromosome 5: 2170a Transcript	VDAG_JR2_Chromosome 5: 2170a	protein_coding	2/2	155	52	18	A/T	GCC/ACC	-	1
5_700313_G/A	5_700313-700313	A	missense_variant	VDAG_JR2_Chromosome 5: 2170a Transcript	VDAG_JR2_Chromosome 5: 2170a	protein_coding	2/2	161	52	18	A/T	GCC/ACC	-	1
5_700313_G/A	5_700313-700313	A	downstream_gene_variant	VDAG_JR2_Chromosome 5: 2171a Transcript	VDAG_JR2_Chromosome 5: 2171a	protein_coding	-	-	-	-	-	-	-	-1

Variant 4

Uploaded variant	Location	Allele	Type	Description	Gene	Protein	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand
5_701484_C/A	5_701484-701484	A	downstream_gene_variant	VDAG_JR2_Chromosome 5: 2150a Transcript	VDAG_JR2_Chromosome 5: 2160a	protein_coding	-	-	-	-	-	-	tmo_5_701484_C_A 1	1
5_701484_C/A	5_701484-701484	A	downstream_gene_variant	VDAG_JR2_Chromosome 5: 2170a Transcript	VDAG_JR2_Chromosome 5: 2170a	protein_coding	-	-	-	-	-	-	tmo_5_701484_C_A 1	1
5_701484_C/A	5_701484-701484	A	downstream_gene_variant	VDAG_JR2_Chromosome 5: 2170a Transcript	VDAG_JR2_Chromosome 5: 2170a	protein_coding	-	-	-	-	-	-	tmo_5_701484_C_A 1	1
5_701484_C/A	5_701484-701484	A	upstream_gene_variant	VDAG_JR2_Chromosome 5: 2171a Transcript	VDAG_JR2_Chromosome 5: 2171a	protein_coding	-	-	-	-	-	-	tmo_5_701484_C_A -1	

Show additional columns

Download options

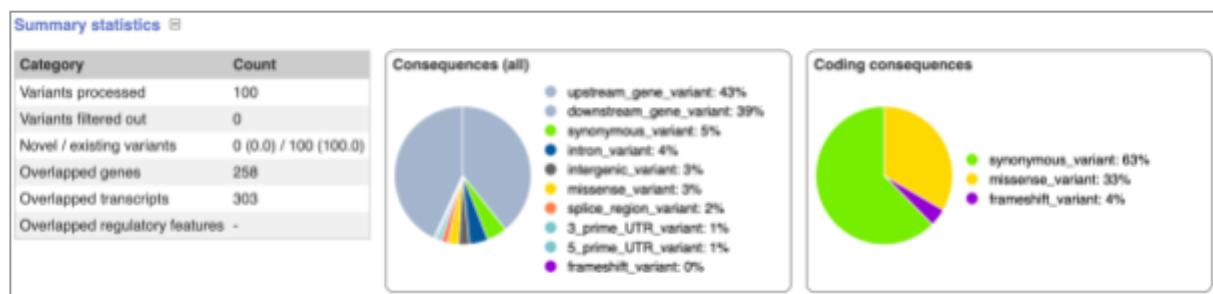
Existing variants

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).

- How many variants have been processed?
- How many genes and transcripts are overlapped by variants in this file?



- 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.*)

Locus	Allele	Consequence	Gene	Exon	HGVSg	HGVSg	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Domains
Supergene_3.1904801-001	T	missense_variant	GMQ_27112	39	GMQ_27112T0:c.289G>A	GMQ_27112T0:p.Gly80Asp	289	290	89	EK	GAGGAG	In :Supergene_3.1904801_C_T	Plas-PP14303 PANTHER:PTHR4125 PANTHER:PTHR4125-SF3 MotifDB: No motifs fits
Supergene_3.190427064-00004	C	missense_variant	GMQ_21813	29	GMQ_21813T0:c.209T>C	GMQ_21813T0:p.Ser70Pro	209	210	79	SP	YQHEDD	In :Supergene_3.190427064_T_G	PROSITE profiles PS51290-P Superfamily 5975692
Supergene_3.190480030-00006	C	missense_variant	GMQ_20460	414	GMQ_20460T0:c.1090G>C	GMQ_20460T0:p.Arg369His	1090	1090	335	DH	GATIAD	In :Supergene_3.190480030_D_C	Genid# 34079910 PANTHER:PTHR423871 Superfamily 59753840 CDD:e11639
Supergene_3.48118082-110082	T	missense_variant	GMQ_00311	114	GMQ_00311T0:c.79G>A	GMQ_00311T0:p.Glu27Lys	79	79	26	EK	GAIAAKA	In :Supergene_3.48118082_C_T	-
Supergene_3.41.2780-7780	G	missense_variant	GMQ_08787	618	GMQ_08787T0:c.1228A>C	GMQ_08787T0:p.DamIDPro_1328	1228	1228	443	GP	CAGICDD	In :Supergene_3.41.2780_T_B	PANTHER:PTHR48896 PANTHER:PTHR48896-SF3
Supergene_3.191713801-173301	T	missense_variant	GMQ_04080	102	GMQ_04080T0:c.287G>A	GMQ_04080T0:p.Gly98Glu	287	287	96	QE	GGACGAA	In :Supergene_3.191713801_Q_T	-
Supergene_3.73.180474-160474	G	missense_variant	GMQ_03045	293	GMQ_03045T0:c.467G>C	GMQ_03045T0:p.Arg157Thr	467	467	136	R/T	AGAINDA	In :Supergene_3.73.180474_C_R	Low complexity (Seq&seq) PANTHER:PTHR31595 PANTHER:PTHR31595-SF1
Supergene_3.427758213-56413	A	missense_variant	GMQ_028114	212	GMQ_028114T0:c.198G>F	GMQ_028114T0:p.Gly62Pro	198	199	33	QH	CAGCAT	In :Supergene_3.427758213_C_A	PANTHER:PTHR31361 PANTHER:PTHR21361-SF15 MotifDB: No motifs fits

- What are the HGVS notations of missense variants falling in known protein domains?

Results preview

Navigation (per variant) **Show: 1 5 10 50 All variants** **Filters** **Download** **New job**

Consequence is defined
Consequence is missense_variant
Clear filters Match all of the above rules **Update**
Uploaded variant **1** **10** **all** **Advanced**

Show/hide columns (22 hidden)

Locales	Allele	Consequence	Gene	Exon	HGVSc	HGVSp	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Domains
Supercodon_3_1584_801_801	T	missense_variant	GMQ_27112	95	GMQ_27112T9.c.285G>A	GMQ_27112T9.p.Glu85Asp	285	285	85	EK	GAG>GAQ	me_Supercodon_3_1584_801_C	Plas_Pf14083 PANTHER_PTHN4126 PANTHER_PTHN4125_SF3 MeIOB_RemoDB-life
Supercodon_3_158_127984_127984	C	missense_variant	GMQ_21813	26	GMQ_21813T0.c.238T>C	GMQ_21813T0.p.Ser79Pro	238	238	79	SF	TGCG>CCG	me_Supercodon_3_158_127984_T	Prosite_profiles_P031296-P Superfamily_SF05832
Supercodon_3_58_89930_89930	C	missense_variant	GMQ_20457	414	GMQ_20457T0.c.1980G>C	GMQ_20457T0.p.Arg639Pro	1980	1980	395	DH	GAAT>CAT	me_Supercodon_3_58_89930_D	GeneID:346720,10,10 PANTHER_PTHN20371 Superfamily_SF05848 CCD_sfH903
Supercodon_3_41_7785_7785	G	missense_variant	GMQ_08797	615	GMQ_08797T0.c.1328A>C	GMQ_08797T0.p.Gly443Pro	1328	1328	443	GP	CAG>CGG	me_Supercodon_3_41_7785_T	PANTHER_PTHN48956 PANTHER_PTHN48956_SF3
Supercodon_3_73_169474_169474	G	missense_variant	GMQ_08045	20	GMQ_08045T0.c.467G>C	GMQ_08045T0.p.Arg136Thr	467	467	136	RT	AGA>ACA	me_Supercodon_3_73_169474_C	Low_complexity_(Seg_asg) PANTHER_PTHN21395_SF1 PANTHER_PTHN21395_SF1
Supercodon_3_427_56213_56213	A	missense_variant	GMQ_08814	20	GMQ_08814T0.c.98G>T	GMQ_08814T0.p.Gln29Asp	98	98	33	QH	CAG>CAT	me_Supercodon_3_427_56213_C	PANTHER_PTHN21361 PANTHER_PTHN21361_SF15 MeIOB_RemoDB-life

(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?

Results preview

Navigation (per variant) **Show: 1 5 10 50 All variants** **Filters** **Download** **New job**

Consequence is frameshift
Uploaded variant **1** **10** **all** **Advanced**

Show/hide columns (22 hidden)

Locales	Allele	Consequence	Gene	Exon	HGVSc	HGVSp	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	UNIPARC
Supercodon_3_1482_1086_1086	G	frameshift_variant	GMQ_27081	10	GMQ_27081T0.c.201del	GMQ_27081T0.p.Phe74ArgfsTer27	200-201	200-201	74	PX	GGG>GGG	me_Supercodon_3_1482_1086_CGG_CG	UP000000000

Exercise: Ensembl Fungi whole-genome alignments

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

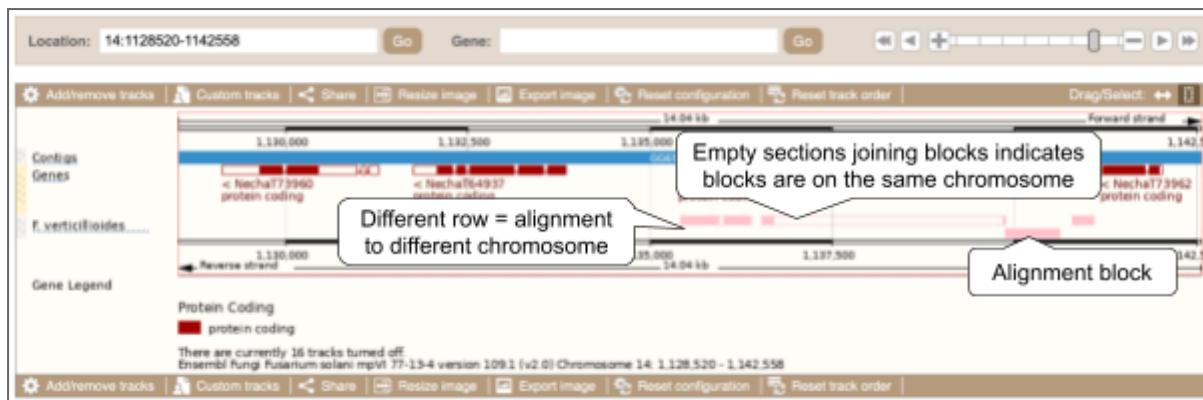
- (a) Find the region **14:1128520-1142558** in *Fusarium solani* and go to the [Region in detail](#) page. This region includes four genes we identified from our first BioMart query: *PEP5*, *PDA1*, *ESP3* and *PEP5*.

The screenshot shows the Ensembl search interface. In the search bar, 'Fusarium solani' is entered. Below it, the specific genomic location '14:1128520-1142558' is typed into a field. A brown 'Go' button is to the right of the location field. Below the search bar, there is a placeholder text 'e.g. NAT2 or alcohol*'.

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

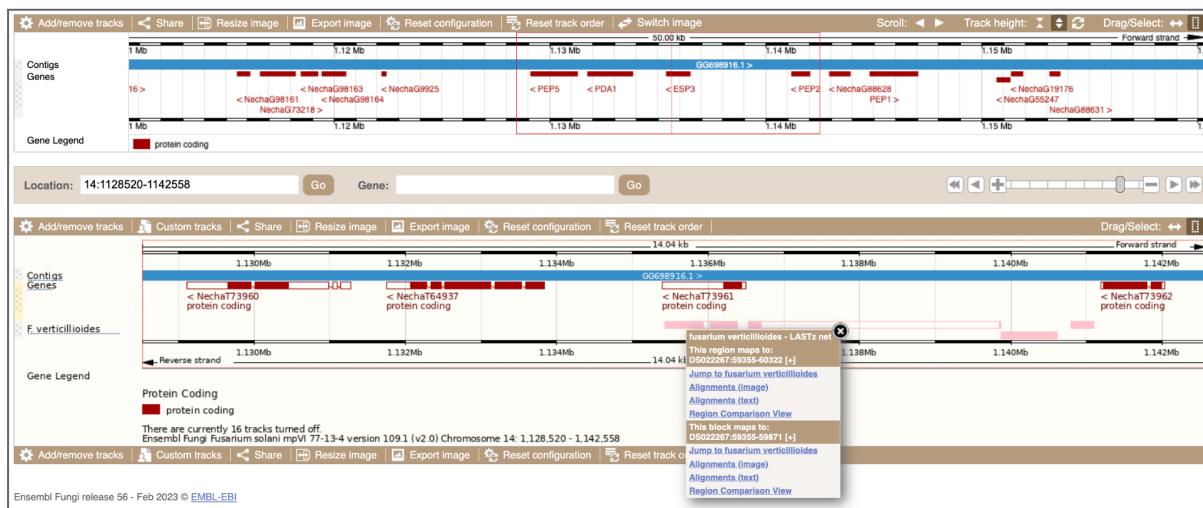
The screenshot shows the 'Comparative genomics' configuration page. On the left, a sidebar lists various track categories: Active tracks, Favourite tracks, Track order, Search results, Sequence and assembly (2/4), Genes and transcripts (2/2), mRNA and protein alignments (6/6), EST alignments (0/1), RNA alignments (0/3), Comparative genomics (1/1), Repeat regions (0/9), Information and decorations (8/10), and Display options. Under 'Comparative genomics', there is a single track listed: 'Fusarium verticillioides - LASTz net'. A note below the track says 'Looking for more data? Search the [TrackHub Registry](#) for external sources of annotation'. On the right, a key defines symbols: a grey square for 'Track style', a grey square with a dot for 'External data', a blue square for 'Forward strand', a red square for 'Reverse strand', a yellow star for 'Favourite track', and a blue info icon for 'Track information'. At the bottom, notes state: 'Please note that the content of external tracks is not the responsibility of the Ensembl project.' and 'URL-based tracks may either slow down your ensembl browsing experience OR may be unavailable as these are served and stored from other servers elsewhere on the Internet.'

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.

A total of 4 alignment blocks have been found. Please select an alignment to view by selecting a Block from the Alignment column.

Show/hide columns	Filter		
Alignment (click to view)	Length (bp)	Location on <i>Fusarium solani</i>	Location on <i>Fusarium verticillioides</i>
Block 1	3335	14:1136537-1139871	DS022270:2542-2734
Block 2	961	14:1135422-1136382	DS022267:59355-60322
Block 3	746	14:1139872-1140617	5:2630678-2631429
Block 4	305	14:1140792-1141096	1:1367865-1368184

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.

Display options	
Strand:	Forward <input type="button" value="▼"/>
Number of base pairs per row:	120 bps <input type="button" value="▼"/>
Additional exons to display:	Core exons <input type="button" value="▼"/>
Orientation of additional exons:	Display exons in both orientations <input type="button" value="▼"/>
Line numbering:	None <input type="button" value="▼"/>
Codons:	Do not show codons <input type="button" value="▼"/>
Show conservation regions:	<input checked="" type="checkbox"/>
Mark alignment start/end:	<input checked="" type="checkbox"/>

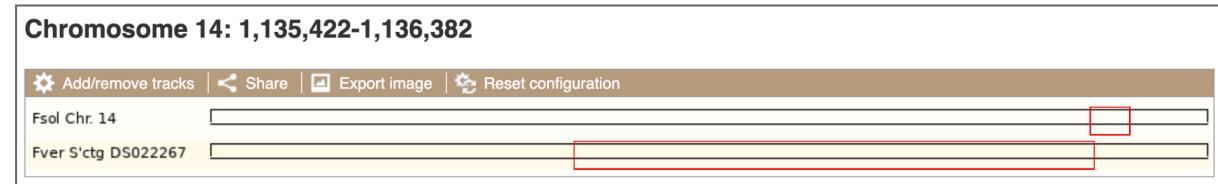
Fusarium_solani	ACACAATCATGGCCATGGCGGACTCGCCATCCCACAGGTTTCCAGCTGGACCCGCCGCTCAAGCAGCAGAACATTCTTGAGCACCCATCTTACCGCCGC
Fusarium_verticillioides	ACCCAATCACAGCATCTGTGCTCTCACCTTCGAAGCACCTCCATCACTGGAGATCAGTCCACCGGCTTGAGCCGGCAGGTCACCGTGGACAGTGTGATTTCACCGGCAA
Fusarium_solani	CCTCTGGACTCTCTGAGGCTGG-----AATCTGTGGCCAGTCGAGTGTATCTTGAAAGCAGATTGAGGCGAGATCAGAGAAGAGGCAAAGACATGGCAAGATTCT
Fusarium_verticillioides	CTTCTGGACTCTCTGAGGCTGGAGATTAAGAATAGCG-CATTCTGAGG-----ATTCTTCAC--GGAGATTCTGGAGGAGAACTCACCACCTTCACCAAAACCGTGTGATTCTTCT
Fusarium_solani	CGCGCTTCTTATGGATCACAGTGTGCTCTGGTAGATTCGCATCAACTGGAAACACTGGGACCCAGGGAGGCGAANCCCTG-CTTGTGTTA-GGAGCCGCCAGAGTGTCAAGTCTGGA
Fusarium_verticillioides	CGGGCTCGAA---GCCACATGTGGCGACTTAAGCGTCTGGCGATAAATTCGGCCCTACACCTTCTCCGGGACTTCAGCGACAGCCGATTCGCACTTCAG
Fusarium_solani	GGCAGACCAAGTCAGGATAGAACATGATGGACACTACCATCTTGTACTGGGCAACTACGGCTAC-GCGAACGAATCAACCTGTAGTTCTGGCCATGCAACACAECAAACATCA
Fusarium_verticillioides	TCCAAAACGAACTGGATGGAACTGGTACACATTCATCTCTTGTCTCTGGCGTTATGGCTTCCAGCTGGCTAAACAGCTGCTGGAGTCGGCCCTTGACAACACTCCAGAATCA
Fusarium_solani	CTCAGGGATCTTCCCACAGATCGCAAGGCAAGATCTGGATTTTCTCTGTGACTTAAATCCCGCCCGATCAGGGTTGGCTCTGTGCAATTAGACGCTTCGCTCTCTGCA
Fusarium_verticillioides	TCCAGGGATCTTACATCGCAACAGCAAGATGGTACTTCTGCTGATCTTCTGCTGAGTGGCTTCTGCTGAGTGGCTCTGCTGAGTGGCTCTGCTGAGTGGGCTCTGCTGCA
Fusarium_solani	GTGACAGGTGATCAACTATAGTGAACCTCAACCTTGTCCAAAGAACCCATCGCATCAGCATCAAACCGAGCGC-GGCCAGGCCAGTCGAGAAAGTGTAGTCAGTCTAGGAC
Fusarium_verticillioides[REDACTED].....CACTGACTTTTATTCTGGCCATGTCCTACGCTGAGCACTGGCCAGGCTGAGTAGCTGAGCCTAGGGCTGCTGAGTCTCAGATAGGGAC
Fusarium_solani	ATGGCACTCGGATATGGCGATTTCTGAGCATCTGTCAACTGCTGTGGTGGG-----TGCCCTTTCTACCAAGCGCTGTCATTCAAGGAGCAC-TGGGCCCTCGGTAC
Fusarium_verticillioides	ATGGCACTCGGCGATGGCGTTCTGGAGACACTGGCTTACGCAATGCTGTTTACCGTATGGTGTATGGCTCTGGCTCCCTTACCTGCTATCCTCAGGAGACATCGTGGAGTTTCTGTC
Fusarium_solani	TGCCAACAGAGGAGGACACAGACGGCTTATGGCTTGGAGAAGGCATTTGGTCTCTCTGAGTGTATGGCCCGTATAAGGCTGTAAGTGTAAAGGCTCTCGCGTGGCTG
Fusarium_verticillioides	CGACAGAACAGAGGGCGAACAGGGCGCTTGTGGG-TACCATTCAGGTTGGCTCCAGAGATGATGTCATAGGTGTTATAAGACTGCTCTGGACTCTAGGAAACATATGCTGGCTGGT
Fusarium_solani	GCGAGCGCTACTGGCCCTGGT-TGGAGGAA
Fusarium_verticillioides	ATGGTGTCTCTGGCATGGTAAAGAGAAA

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.

Unselected species	0	Selected species	1
		 Fusarium verticillioides - lastz	

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.

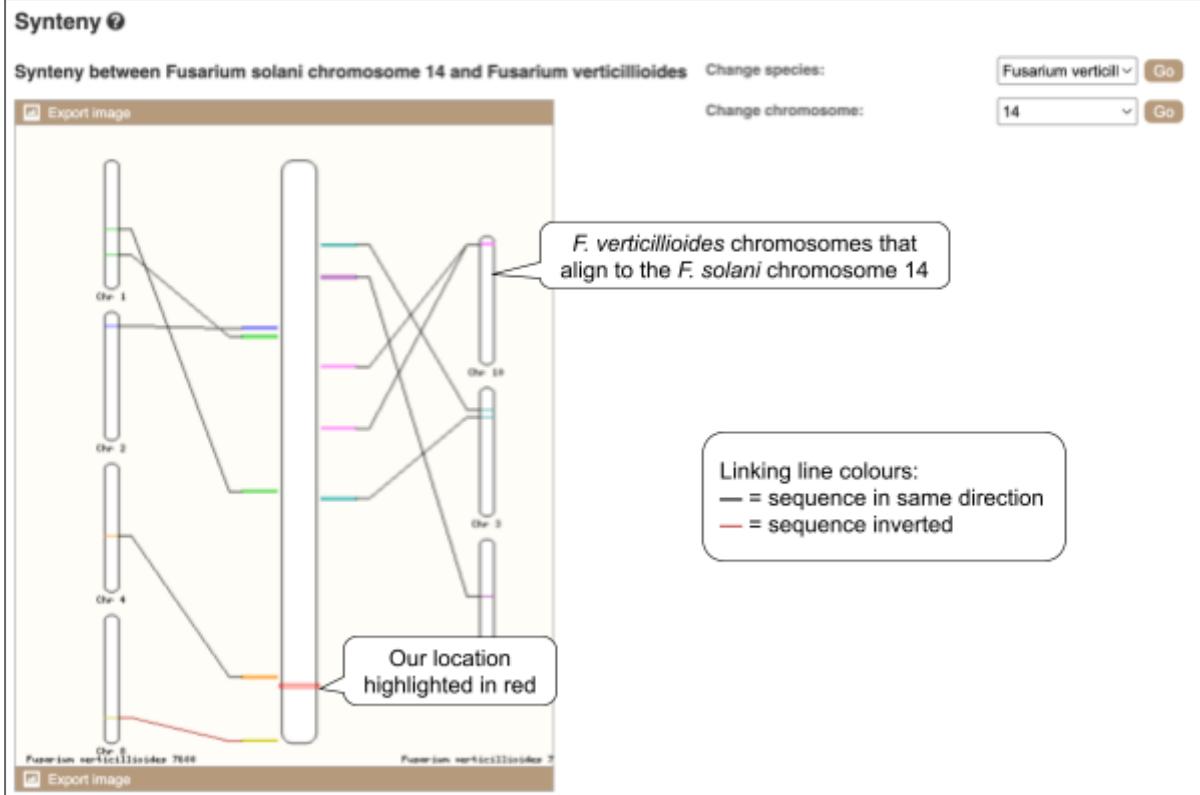


You can add data to both of these views with the same options you had in the Region in Detail page. Click on [Configure this page](#) and look at the top of the menu.

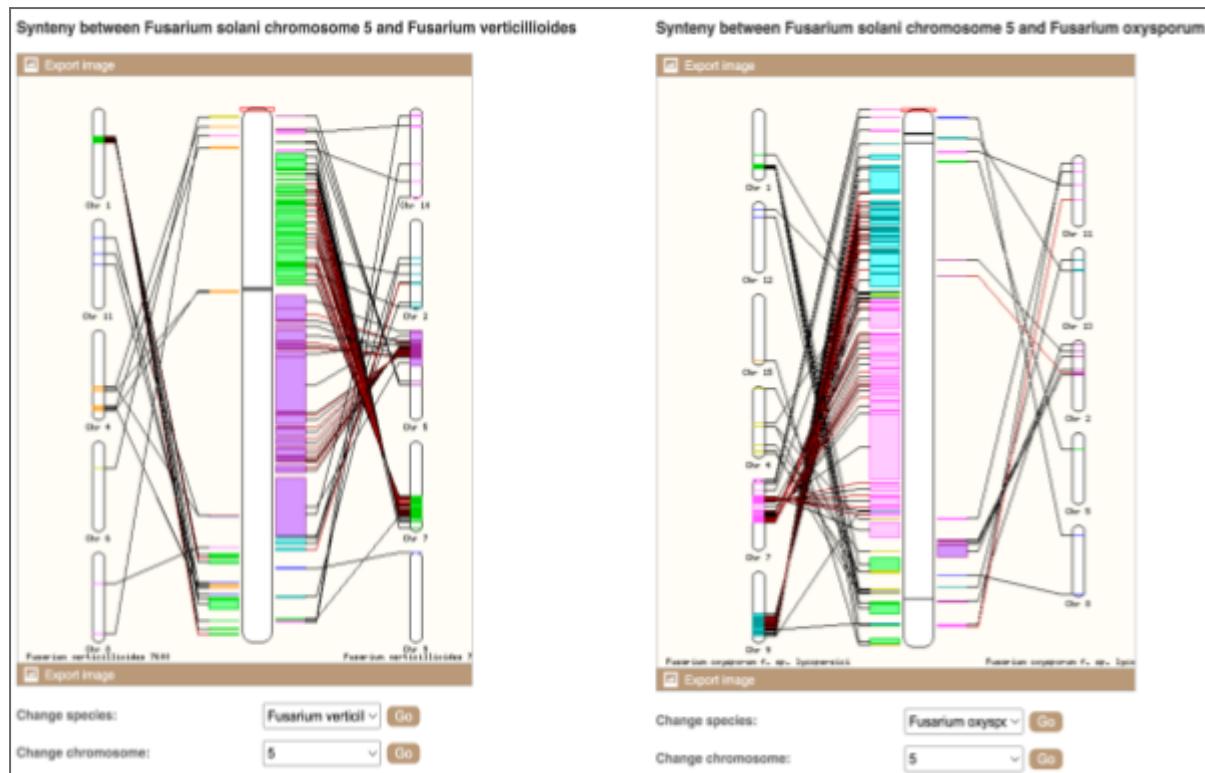
Species to configure:

Select from available configurations:

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*.

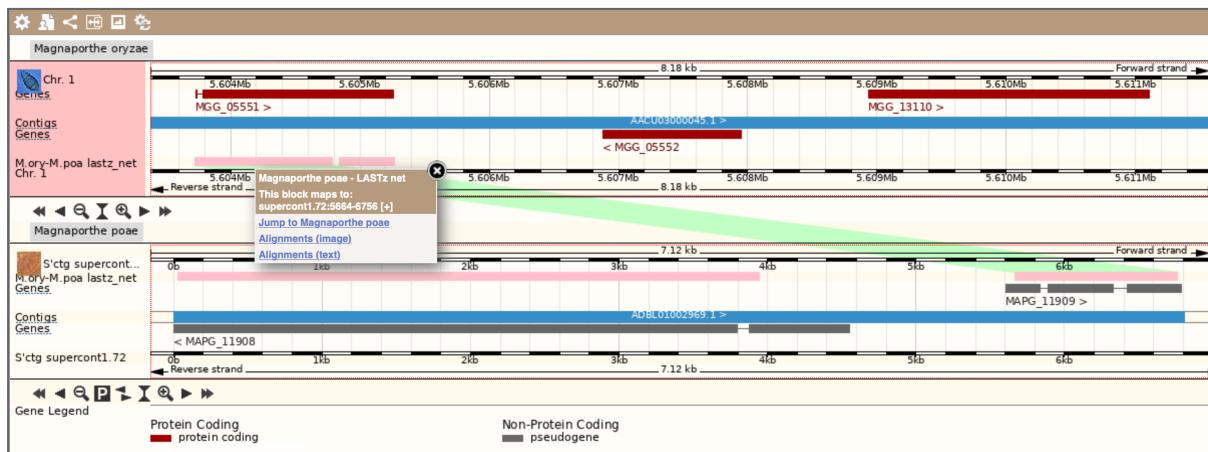
Search: Magnaporthe oryzae for

1:5603535-5611402 Go

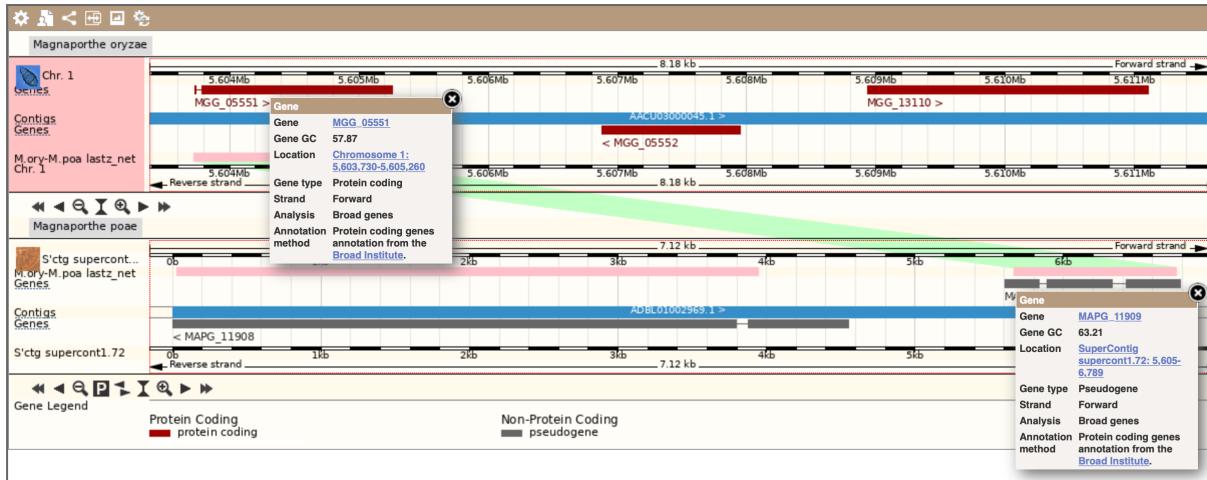
e.g. NAT2 or alcohol*

(b) Click on **Region comparison** and choose *Magnaporthe poae* from the **Select species or regions** pop-up to display an alignment.

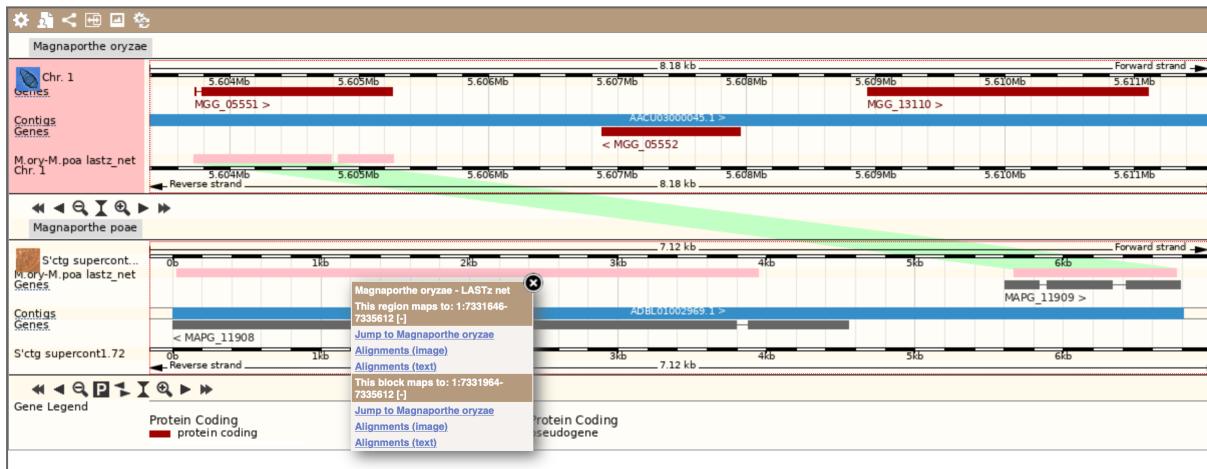
(c) Scroll down to the most detailed image. To what region (chromosome/scaffold/contig) does this region align to on the *M. poae* assembly?



(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/

Info • White rot/brown rot example 2017						
SEARCH	BLAST	ANNOTATIONS	MCL CLUSTERS	GEO MAPPING	DOWNLOAD	INFO
## 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.

The screenshot shows the CAZy browser interface with the title "CAZymes • White rot/brown rot example 2017". The top navigation bar includes links for SEARCH, BLAST, ANNOTATIONS (selected), MCL CLUSTERS, GEO MAPPING, DOWNLOAD, INFO, and HELP!. The ANNOTATIONS dropdown menu is open, showing categories like PFAM DOMAINS, SECONDARY METABOLISM CLUSTERS, CAZYMES (selected), PEPTIDASES, TRANSPORTERS, and TRANSCRIPTION FACTORS. On the right, there are search filters for "Keywords" (set to "Exact") and checkboxes for "Show only filtered results" and "Show only filtered totals" (both checked). A "Display Type" dropdown is set to "Table". The main content area displays a table with columns for Annotations/Genomes (Aurde3, Calco1, Gior1, Phchr2) and Total, followed by an Annotation Description column. The data shows counts for various CAZy families across different organisms.

Annotations/Genomes	Aurde3	Calco1	Gior1	Phchr2	Total	Annotation Description
CAZy	827	350	368	463	2,008	CAZy
AA	130	27	43	92	292	Auxiliary Activities family
CBM	123	18	19	71	231	Carbohydrate-Binding Module family
CE	61	14	14	20	109	Carbohydrate Esterase family

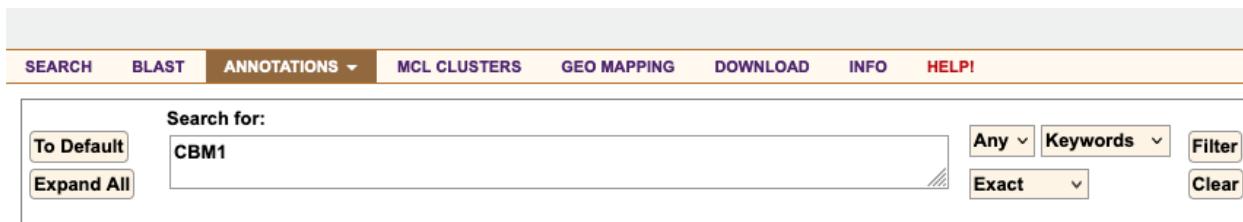
Here you will see a table representation of the predicted CAZymes (Levasseur et al. 2013). The organisms are labeled along the top. The CAZymes are organized by family and labeled along the sides. The numbers in the table tell you how many proteins from each organism's gene catalog were annotated with 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.

This screenshot shows a more detailed view of the CAZy browser interface. The top navigation bar and search/filter section are identical to the previous screenshot. The main content area displays a table with columns for Annotations/Genomes (Aurde3_1, Calco1, Gior1_1, Phchr2) and Total, followed by an Annotation Description column. The data is highly detailed, showing subfamilies and specific enzyme names for each category. For example, under the AA category, it lists AA1, AA1_1, AA1_2, AA1_3, AA1_dist, AA2, AA2_dist, AA3, AA3_1, and AA3_2, with further breakdowns for each.

Annotations/Genomes	Aurde3_1	Calco1	Gior1_1	Phchr2	Total	Annotation Description
CAZy	848	352	372	466	2,038	CAZy
AA	131	29	44	93	297	Auxiliary Activities family
AA1	10	5	5	5	25	Auxiliary Activity Family 1
AA1_1				4	4	Auxiliary Activity Family 1 / Subf 1
AA1_2		2	1	1	4	Auxiliary Activity Family 1 / Subf 2
AA1_3		1			1	Auxiliary Activity Family 1 / Subf 3
AA1_dist		1			1	Multicopper oxidase
AA2	20	1	1	17	39	Auxiliary Activity Family 2
AA2_dist	1	1	1	1	4	Class II peroxidase
AA3	50	15	24	39	128	Auxiliary Activity Family 3
AA3_1	1	1	1		3	Auxiliary Activity Family 3 / Subf 1
AA3_2	38	13	20	34	105	Auxiliary Activity Family 3 / Subf 2

If we read Levasseur et al. 2013 we know that the AA2 family consists of peroxidases that may degrade lignin. Browsing the table, we see that 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.

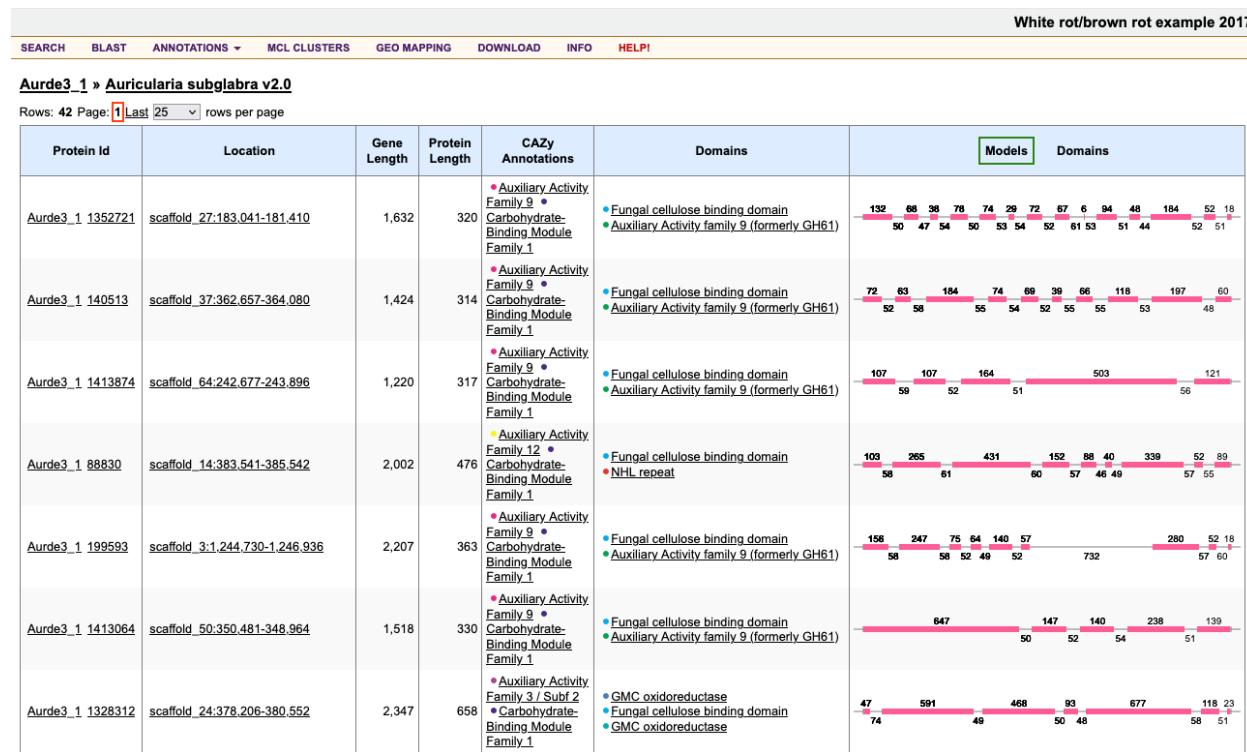


The screenshot shows the CAZy database search interface. The search term 'CBM1' has been entered into the 'Search for:' field. The 'Annotations' dropdown is selected. The search results table includes columns for Annotations/Genomes, Aurde3_1, Calco1, Glotr1_1, Phchr2, Total, and Annotation Description. The table lists various CAZy families and their counts, with 'CBM1' appearing in multiple rows across different families.

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

Notice the abundance of CBM1-encoding genes in *P. chrysosporium* and *A. subglabra*, while *G. trabeum* and *C. cornea* have only a single CBM1-encoding gene each (co-occurring with GH5_5 and GH10 proteins). All of this indicates that we might be looking at two white-rot and two brown-rot fungi.

Click on the number (e.g., 48 for Aurde3_1) to see the CBM1-containing proteins of *A. subglabra* in more detail. Notice a variety of CAZymes co-occur with CBM1, including GH5 (various subfamilies), GH6, and many others.



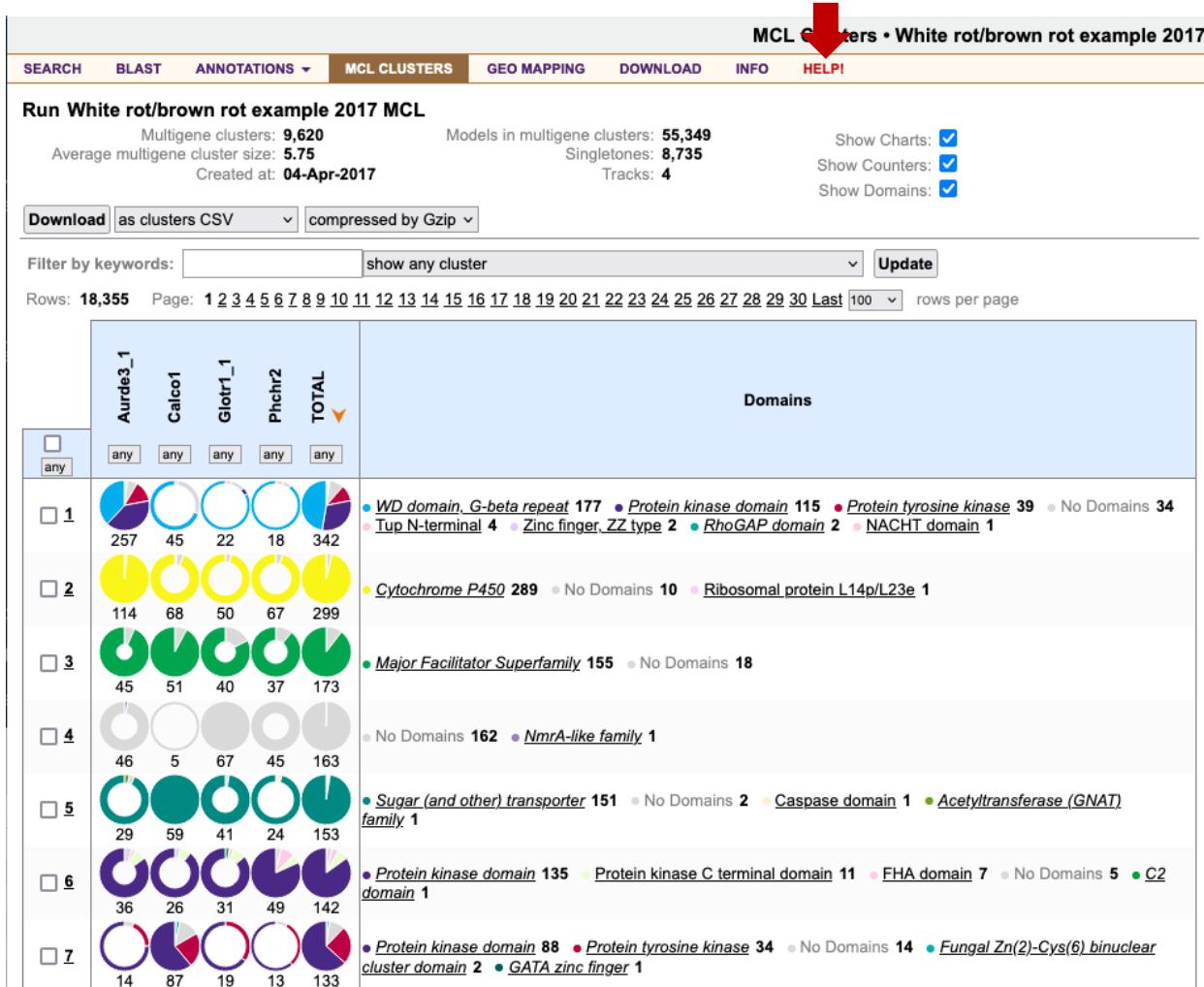
As an exercise, repeat the same search with GH6, GH7, and also the AA9 family of lytic polysaccharide monooxygenases, which may oxidatively act on lignin (Levasseur et al. 2013). Do the presence/absence patterns of these genes indicate the same conclusions about these fungi's mode of decay as we found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

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

Cluster page

Now that we have an idea which fungus uses which decay mode, let's ask the reverse question: what are the genes present in one lifestyle, and absent in the other? To do this, click the 'MCL CLUSTERS' item of the Main menu. Here you will see the results of protein sequence clustering by the 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.

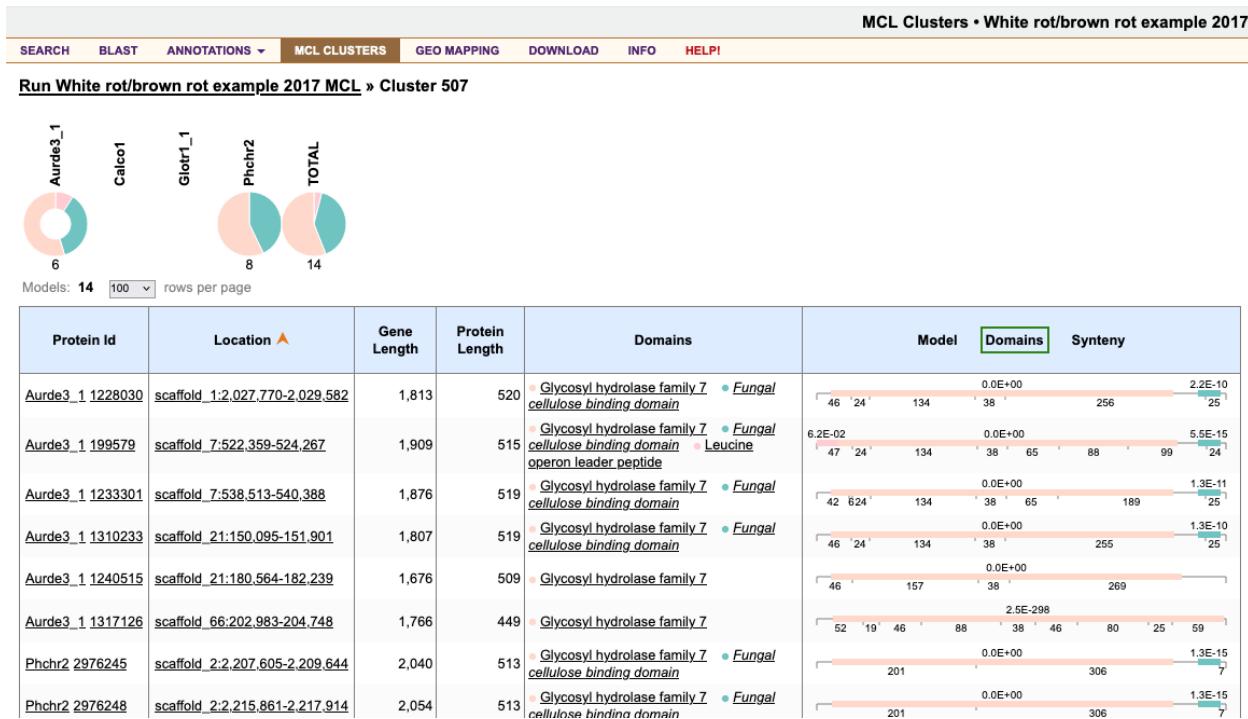
Rows: 150 Page: 1 Last 100 rows per page



150 clusters fit these criteria. These clusters might include genes important to the white rot decay mode, because they are present in white rot fungi and absent in brown rot fungi. 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.

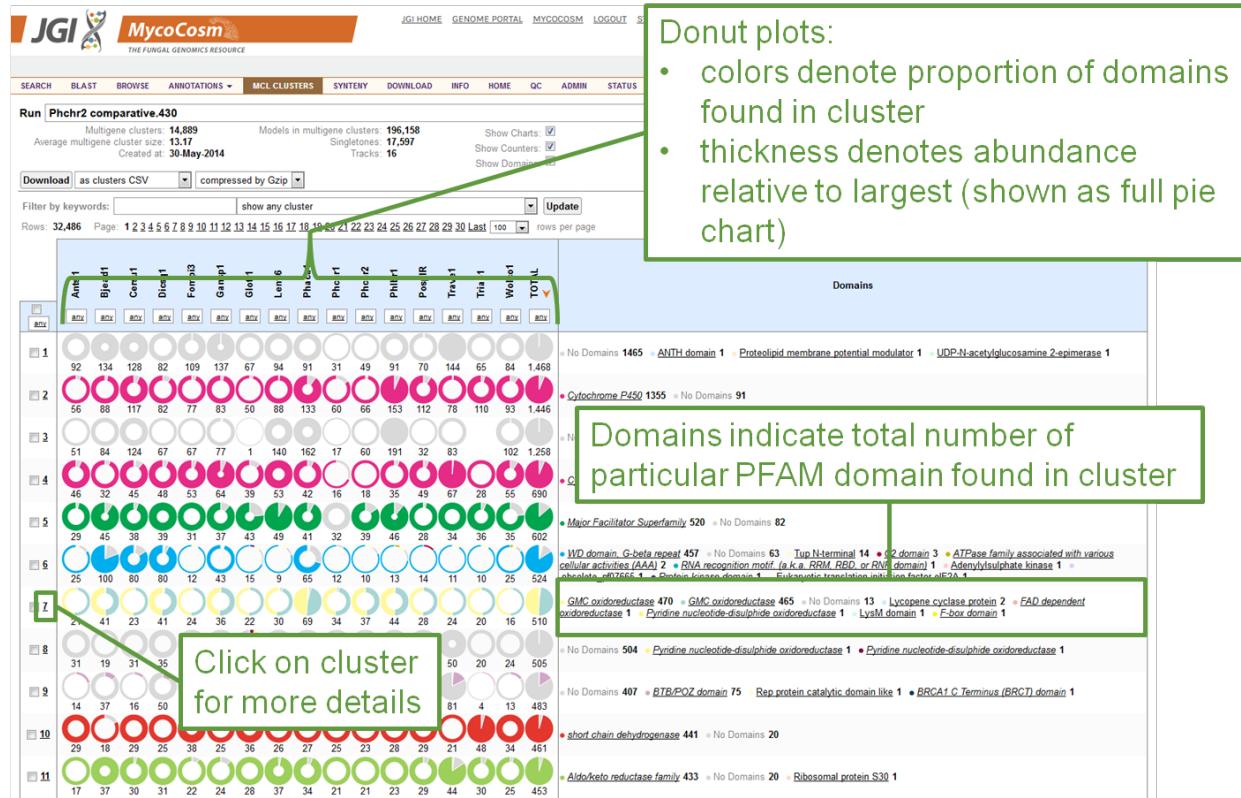


Let's look at what other proteins have the CBM1 carbohydrate-binding motifs in them.

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 monooxygenases (formerly Glycosyl hydrolase family 61), and many hydrolytic enzymes such as GH5, GH6, and GH7. Notice that while these proteins typically are found in expanded copy number in the white rot fungi (Aurde3_1 and Phchr2) they are sometimes found, albeit in lower copy number, in the brown rot fungi (Calco1 and Glotr1_1).

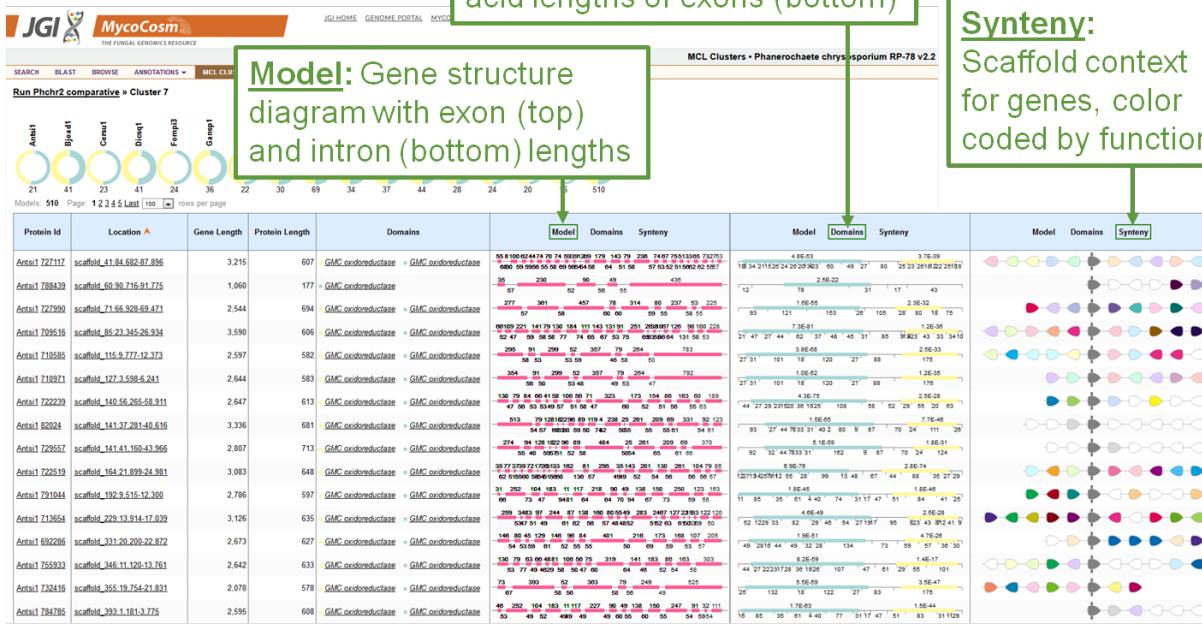
As additional exercises you can (a) search for gene families absent in both white rot fungi; (b) find gene families absent in white rot but present in both brown rot fungi and look at functional domains associated with these families; (c) check if any of these domains are present only in brown rot fungi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

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



Clicking in Cluster number provides additional tools as shown below.

On detail page, different tabs provide useful information:



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., Patyshakulyeva, 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.

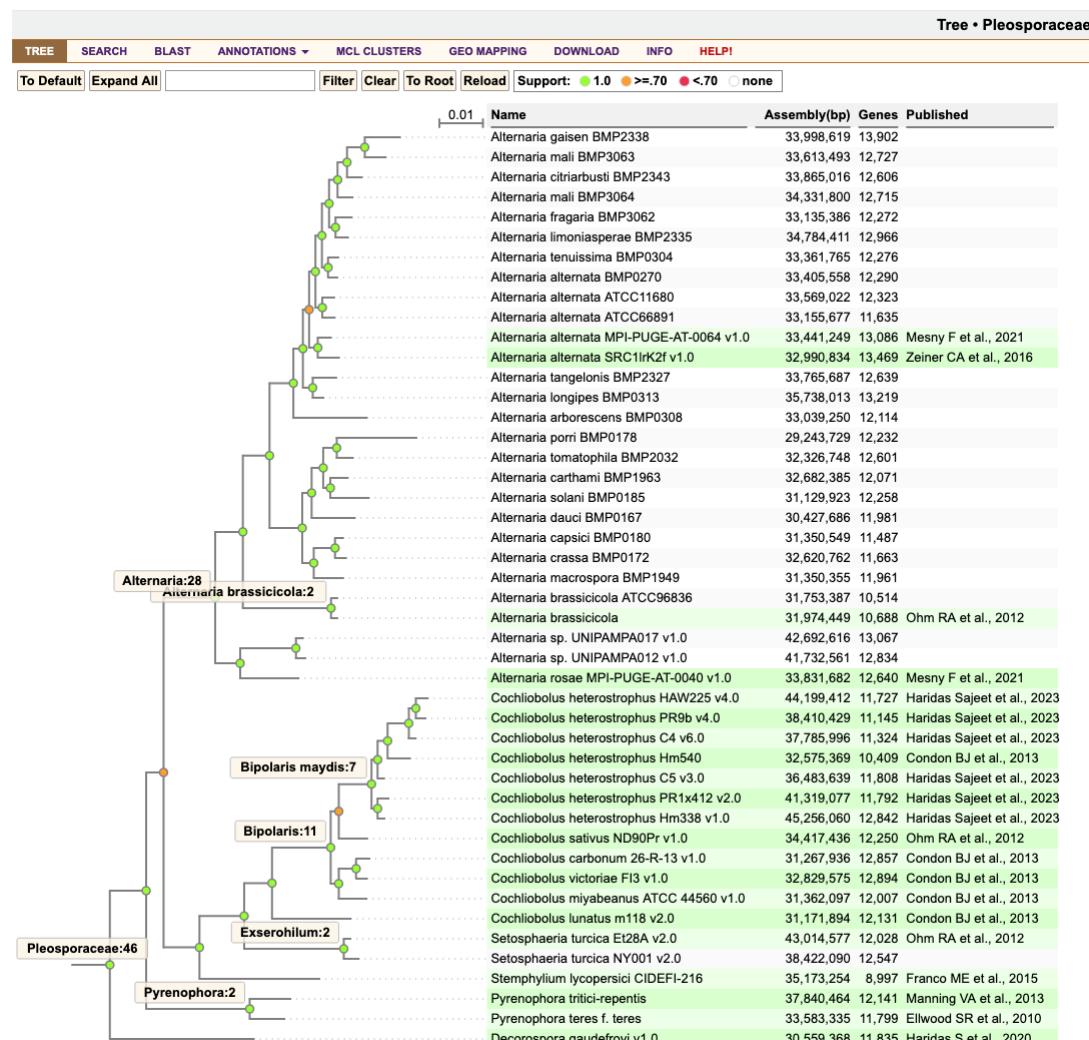
MycoCosm: Synteny Tutorial

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.



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.

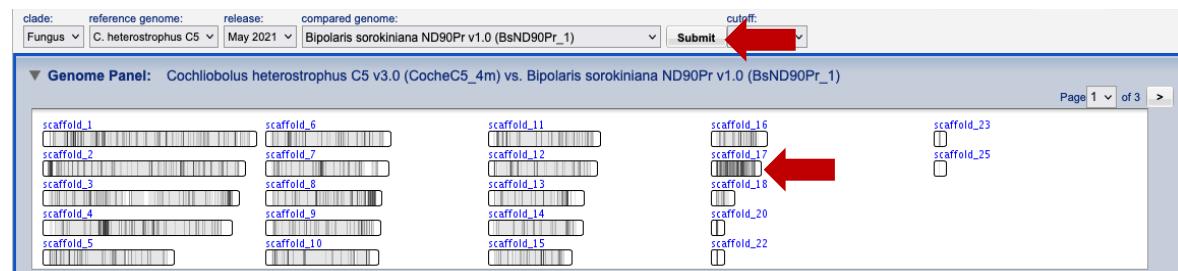
Click on “*Cochliobolus heterostrophus* C5” to go to the organism genome portal. Ideally, you should do this in another tab or window so that you can follow the exercises below keeping the phylogenetic placement of this organism in mind.

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

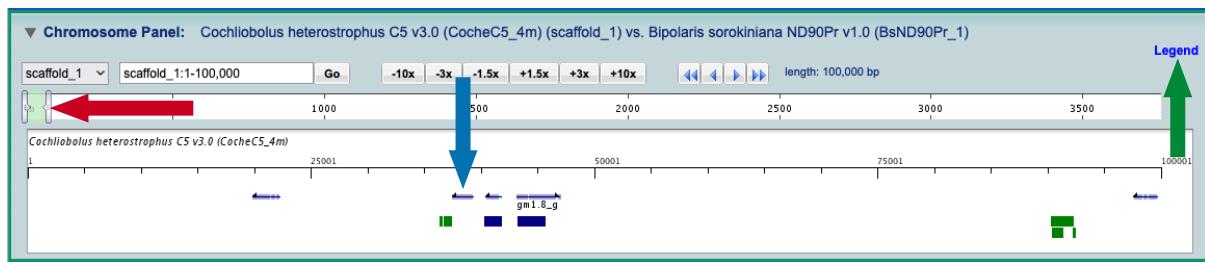


Genomic synteny is displayed in three collapsible panels in the Synteny Browser: the Genome Panel, the Chromosome Panel and the Comparison Panel.

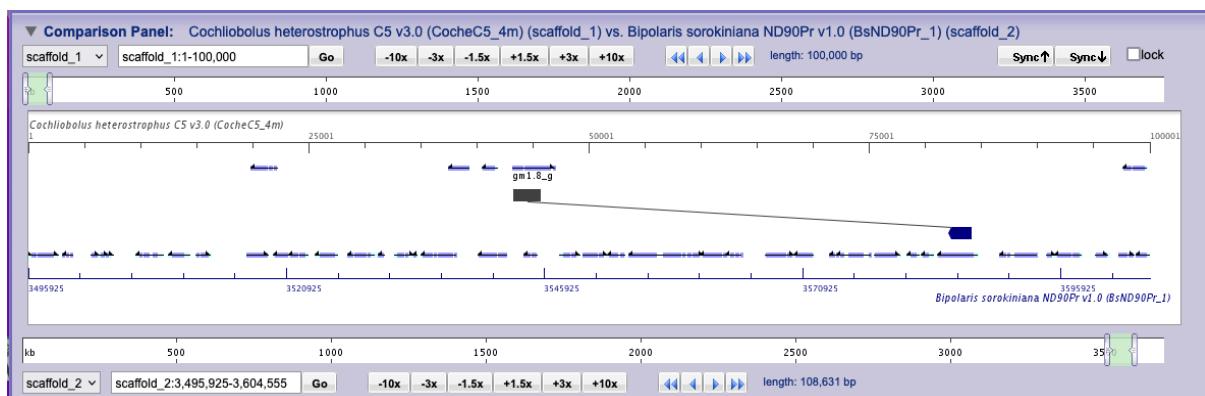
The compared genome can be changed from the dropdown menu and clicking “Submit”. The Genome Panel depicts alignment density for all scaffolds in the reference genome against all chromosomes in the compared genome. Here, alignment density is defined for a region in the reference genome as the number of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



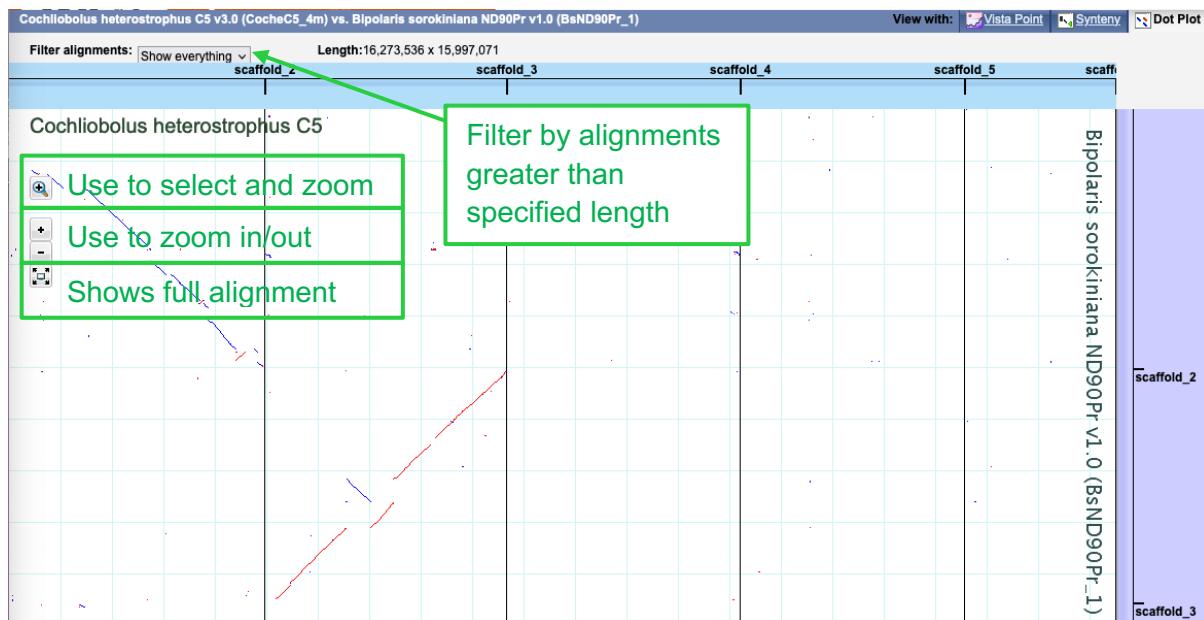
Syntenic blocks and gene models are both interactive, as described above for the Chromosome Panel. Navigation controls allow the user to switch chromosomes, zoom and pan independently over the reference and compared genomes. The SYNTENY page also allows whole genome pairwise comparison and comparison 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.

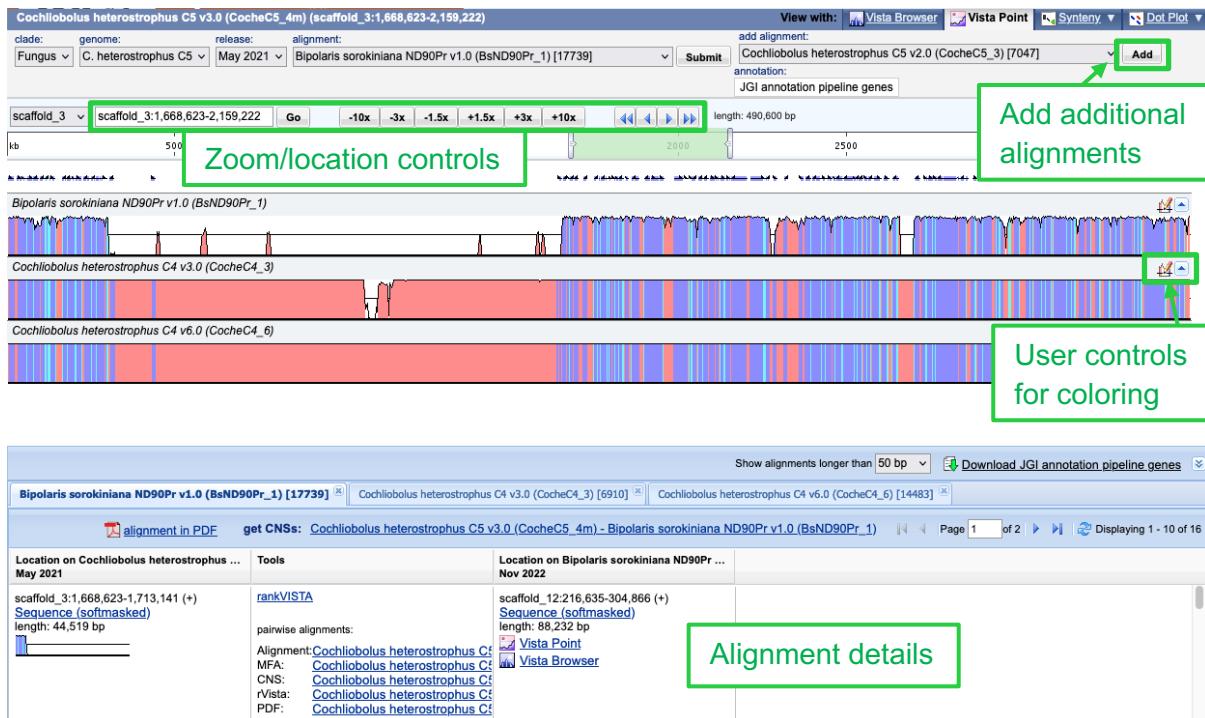


In the main view window, DNA coordinates of the reference genome are presented on the X axis, and DNA coordinates of the compared genome are presented on the Y axis. All chromosomes or scaffolds are concatenated together, usually in a descending order by size. The diagonal lines in the image display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



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



Exercises:

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

Reference:

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

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/phmmr> 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 Basidiomycota?
 - 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.
- Are the clusters around this conserved between these species? Click on them to highlight the occurrence in the image.
 - 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/hmmmer/search/phmmmer> 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.

The screenshot shows a search interface for the fungi.ensembl.org website. At the top, there is a search bar with two input fields. The first field contains the text "Fusarium solani" and the second field contains "NechaG73962". To the right of the search bar is a dropdown menu with the word "for" next to it. Below the search bar is a brown rectangular button with the word "Go" in white. At the bottom left of the interface, there is a note in blue text that says "e.g. NAT2 or alcohol*".

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.

Fusarium solani (v2.0) ▾

Location: 14:1,141,191-1,142,037 Gene: PEP2 Transcript: NechaT73962

Transcript tab

Transcript: NechaT73962

Description Pea pathogenicity protein 2 [Source:UniProtKB/Swiss-Prot;Acc:[C7ZC16](#)]

Location Chromosome 14: 1,141,191-1,142,037 reverse strand.

About this transcript This transcript has [2 exons](#) and is annotated with [4 domains and features](#).

Gene This transcript is a product of gene [NechaG73962](#) [Show transcript table](#)

Transcript-based displays

- Summary
- Sequence
 - Exons
 - cDNA
 - Protein
- Protein Information
 - Protein summary
 - Domains & features
 - Variants
 - PDB 3D protein model

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

Fusarium solani (v2.0) ▾

Location: 14:1,141,191-1,142,037 Gene: PEP2 Transcript: NechaT73962

Transcript: NechaT73962

Pea pathogenicity protein 2 [Source:UniProtKB/Swiss-Prot;Acc:[C7ZC16](#)]

Chromosome 14: 1,141,191-1,142,037 reverse strand.

About this transcript This transcript has [2 exons](#) and is annotated with [4 domains and features](#).

Gene This transcript is a product of gene [NechaG73962](#) [Show transcript table](#)

Transcript-based displays

- Summary
- Sequence
 - Exons
 - cDNA
 - Protein**
- Protein Information
 - Protein summary
 - Domains & features
 - Variants
 - PDB 3D protein model
 - AlphaFold predicted model
- Genetic Variation
 - Variant table
 - Variant image
 - Population comparison
 - Comparison Image
- External References
 - General identifiers
 - Oligo probes
 - Supporting evidence
- ID History
 - Transcript history
 - Protein history

[View protein sequence](#)

Protein sequence ?

[Download sequence](#) [BLAST this sequence](#)

Exons An exon Another exon Residue overlaps splice site

Markup loaded

* Variants are filtered by consequence type

MVNLLHSLPQGSRPNAAIRNNGPDSLALERLKLRELJAEGWPSYRDSCCEWFESIFHPGAY
VYTITWSGRVAYQDFIAASKAGMDKGAFIMIRCHGSSTDINVDGTRAVTKLKATITQRPEV
GGSEFDVVAEADCRFCFYFEKINGSWGARLVRKHWYEKDRMIPVNPAAKFPQVQDSDLKLKAYPPG
YKYLAYWQETANGIKVLLDMPGHRRRVGTVNLKDELYWLAKRMILEGEQIEV

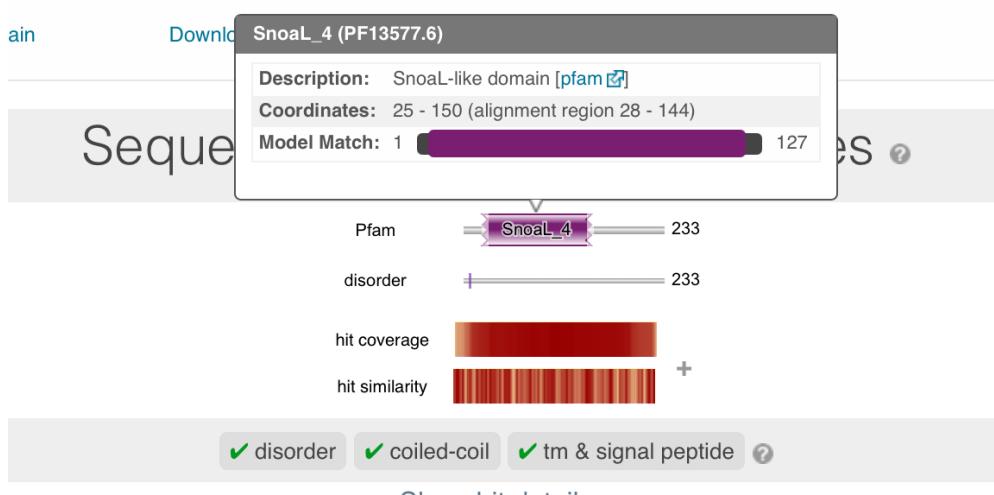
Copy protein sequence

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

The screenshot shows the Ensembl Fungi website with the HMMER search tool. The main navigation bar includes links for HMMER, BLAST, BioMart, Tools, Downloads, Help & Docs, and Blog. A 'Login/Register' link is in the top right. A search bar at the top right contains the placeholder 'Search Ensembl Fungi...'. Below the search bar is a large 'HMMER' logo with a yellow circle above it. To the right of the logo is the text 'phmmmer' and 'protein sequence vs protein sequence database'. A text input field is labeled 'Paste in your sequence or use the example' with a small help icon. The input field contains a sample sequence: 'MVLLESLFOGSRPHAAITRNGPDSALEREKLERELAEQWPTDSCENENFESIPHPGAY VYTWTSGRVAYQDFIAAESRAGHDGAFIIRHRCNGSSTDINVEGTRAVKLKATITQRFFEV GGSREFPSVADCRNCFITFERINGWGAFLVKHNYEKDHITFVSPAKFPQVNEDELKAYFPG TKTLAYMHQETAMSLKVLLOMPQHHRHRYGTVNLEFHDELYWLAKHNLGEQLEV'. Below the input field are 'Submit' and 'Clear' buttons.

- 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.

			Customise
Species	Cross-references	E-value	
Nectria haematococca (strain 77-13-4 / ATCC MYA-4622 / FGSC 9596 / MPVI) ↗	XXX Crosses File	2.6e-163	

Customise Results ↗

Select Visible Columns ↗

Row Count Known Structure
 Secondary Accessions and Ids Identical Seqs
 Description Number of Hits
 Species Number of Significant Hits
 Cross-references Bit Score
 Kingdom Hit Positions
 Phylum

Rows Per Page ↗

50 100 250 1000 2500

Update Restore Defaults

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...)?

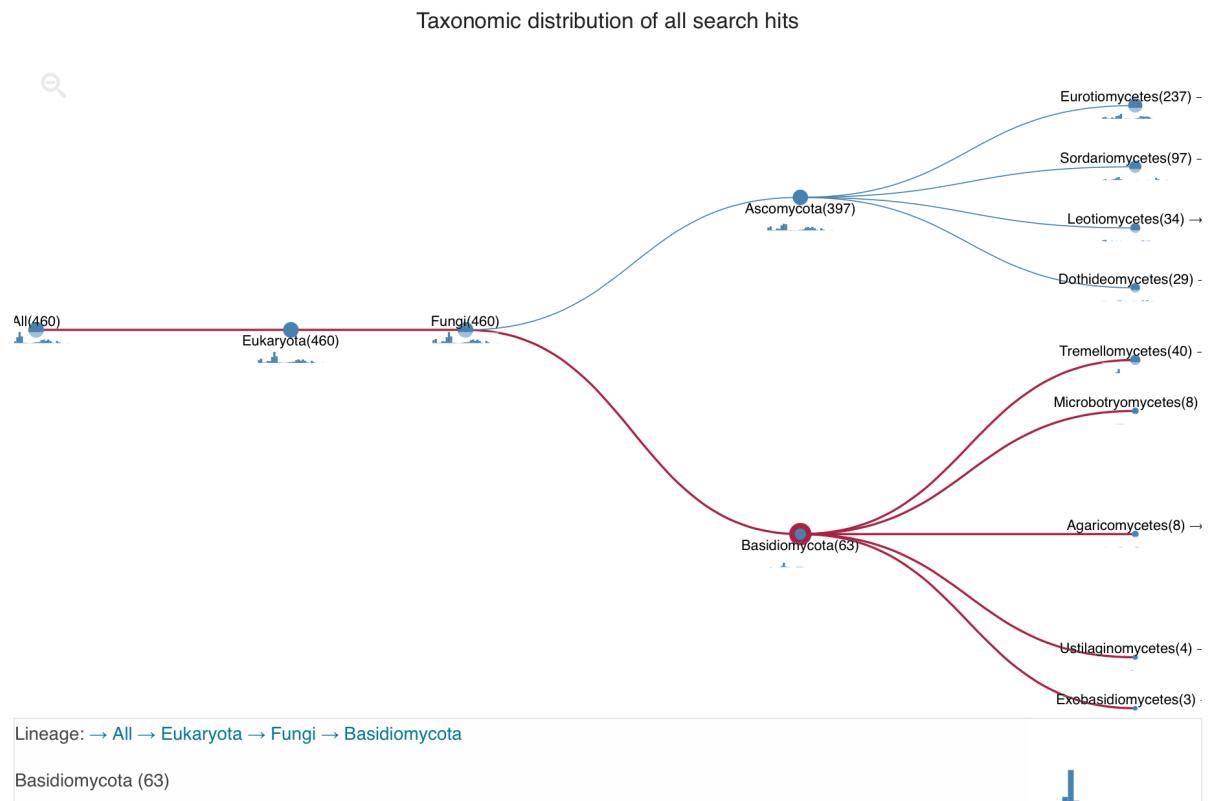
Answer: The sexual form (teleomorph) of Fusarium solani (the anamorph) is Nectria haematococca.

Significant Query Matches (460) in <i>ensemblgenomes</i> (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) ↗	XXX Crosses File	2.6e-163
>	LW93_4799 ↗	Uncharacterized protein	Ascomycota	Gibberella fujikuroi ↗	Crosses File	1.6e-137
>	FFB14_04603 ↗	Pea pathogenicity protein 2	Ascomycota	Fusarium fujikuroi (GCA_900096505) ↗	Crosses File	2.1e-137
>	AU210_001920 ↗	hypothetical protein	Ascomycota	Fusarium oxysporum f. sp. radicis-cucumerinum ↗	Crosses File	6.2e-137
>	FOWG_10080 ↗	pea pathogenicity protein 2	Ascomycota	Fusarium oxysporum f. sp. lycopersici MN25 (GCA_000259975) ↗	Crosses File	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 Basidiomycota?

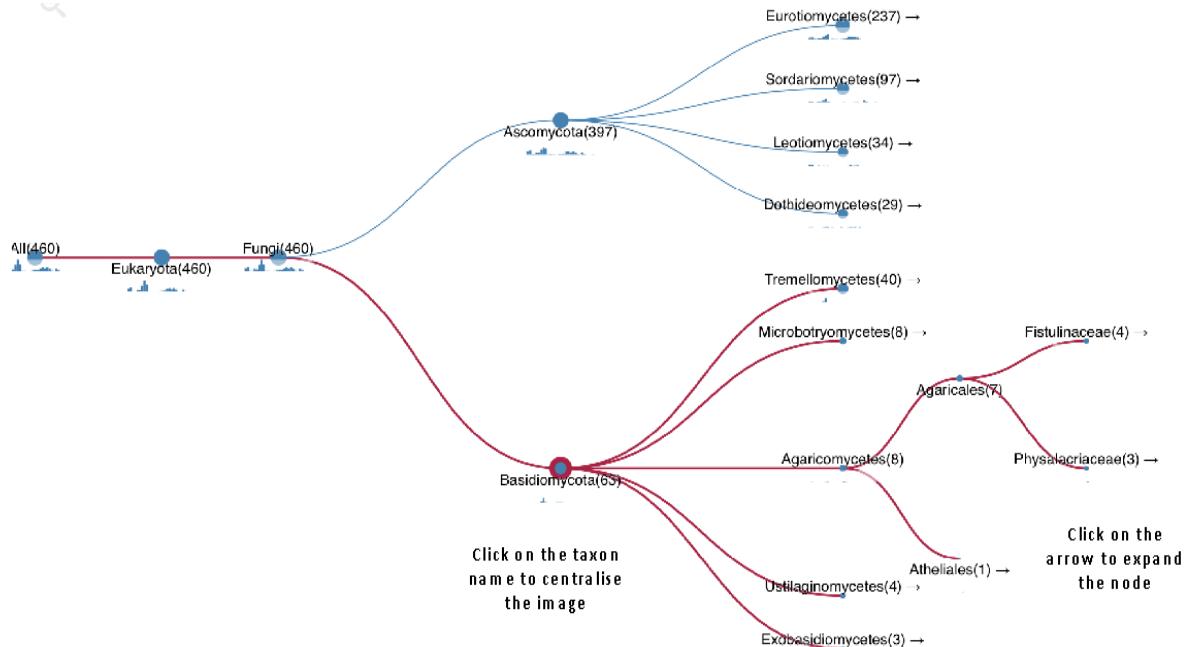
Answer: We can see from the number in the parentheses that there are 63 hits.



- 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.

JGI MycoCosm
THE FUNGAL GENOMICS RESOURCE

JGI HOME GENOME PORTAL MYCOCOSM LOGIN

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

Run **Fusso1 comparative clustering.2371**

Multigene clusters: **15,016** Models in multigene clusters: **150,173**
Average multigene cluster size: **10.00** Singletones: **6,116**
Created at: **30-Mar-2018** Tracks: **9**

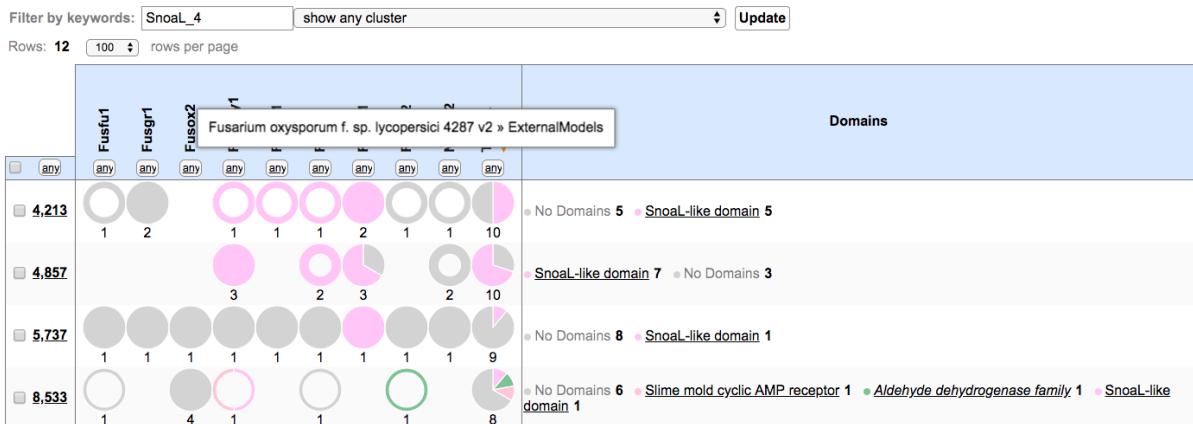
Show Charts: Show Counters: Show Domains:

Download as clusters CSV compressed by Gzip

Filter by keywords: **SnoaL_4** show any cluster Update

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?

Protein Id	Location ▲	Gene Length	Protein Length	Domains	Model	Domains	Synteny
Fusfu1_1982	chromosome_016.273.195-6.273.647	453	151			453	
Fusgr1_34	Supercontig_3.1:79.782-80.232	451	133			92	310
Fusgr1_5463	Supercontig_3.3:11.368-11.820	453	150			453	
Fusoxy1_684420	scaffold_18:905.649-906.256	608	151	Snoal-like domain		18	530
Fuspa1_41	scaffold_1:130.291-130.743	453	151	Snoal-like domain		60	453
Fusre1_682149	scaffold_1:330.403-330.855	453	151	Snoal-like domain		453	453
Fusso1_495533	scaffold_11:951.859-952.731	873	168	Snoal-like domain		873	
Fusso1_572020	scaffold_28:148.589-149.043	455	147	Snoal-like domain		455	
Fusve2_107	Scaffold_1:188.201-190.835	2,635	150			98	2,474
Necha2_102143	sca_29_chr12_4_0-514.012-514.819	808	183			63	
						566	66
							174

These three have the same protein length

- 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.



Protein Id	Location ▲	Gene Length	Protein Length	Domains	Model	Domains	Synteny
Fusfu1_1982	chromosome_016.273.195-6.273.647	453	151				
Fusgr1_34	Supercontig_3.1:79.782-80.232	451	133				
Fusgr1_5463	Supercontig_3.3:11.368-11.820	453	150				
Fusoxy1_684420	scaffold_18:905.649-906.256	608	151	Snoal-like domain			
Fuspa1_41	scaffold_1:130.291-130.743	453	151	Snoal-like domain			
Fusre1_682149	scaffold_1:330.403-330.855	453	151	Snoal-like domain			
Fusso1_495533	scaffold_11:951.859-952.731	873	168	Snoal-like domain			
Fusso1_572020	scaffold_28:148.589-149.043	455	147	Snoal-like domain			
Fusve2_107	Scaffold_1:188.201-190.835	2,635	150				
Necha2_102143	sca_29_chr12_4_0-514.012-514.819	808	183				

Click on the neighbouring clusters to see all occurrences

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

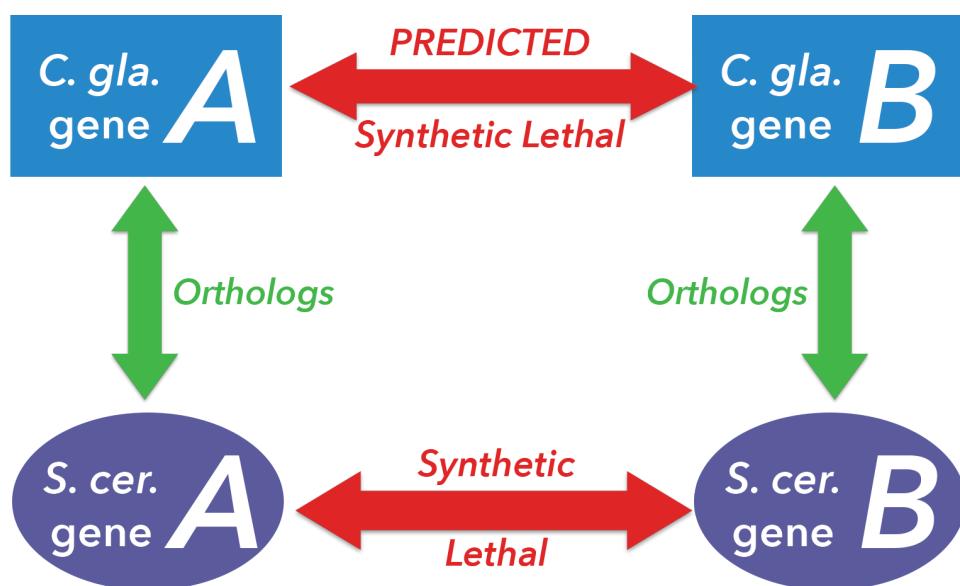
Answer: *Nectria haematococca* v2.0 FilteredModels1. We know this to be the sexual form of *F. solani* so this is expected.

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 toazole 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.

Candida Genome Database

Home Search GBrowse JBrowse Sequence GO Tools Literature Download Community

BLAST
GO Term Finder
GO Slim Mapper
Text Search
Primers
PatMatch
Advanced Search

GFP-labeled Dam1 Complex proteins in DAPI-stained nuclei
Courtesy of Laura Burack and Judy Berman, University of Minnesota

New and Noteworthy

***C. lusitaniae* strain CBS 6936 sequence and BLAST datasets now available at CGD**

The sequence and annotation of *C. lusitaniae* strain CBS 6936, described in Durrens et al. (2017), has been made available at CGD. We provide downloads for sequences, chromosomal features, gff files and protein domain predictions. In addition, *C. lusitaniae* CBS 6936 is included among the datasets searchable by our multi-species BLAST tool. The sequence and annotation were obtained by CGD from NCBI.
(Posted February 27, 2018)

About CGD

CGD Curation News

- 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.

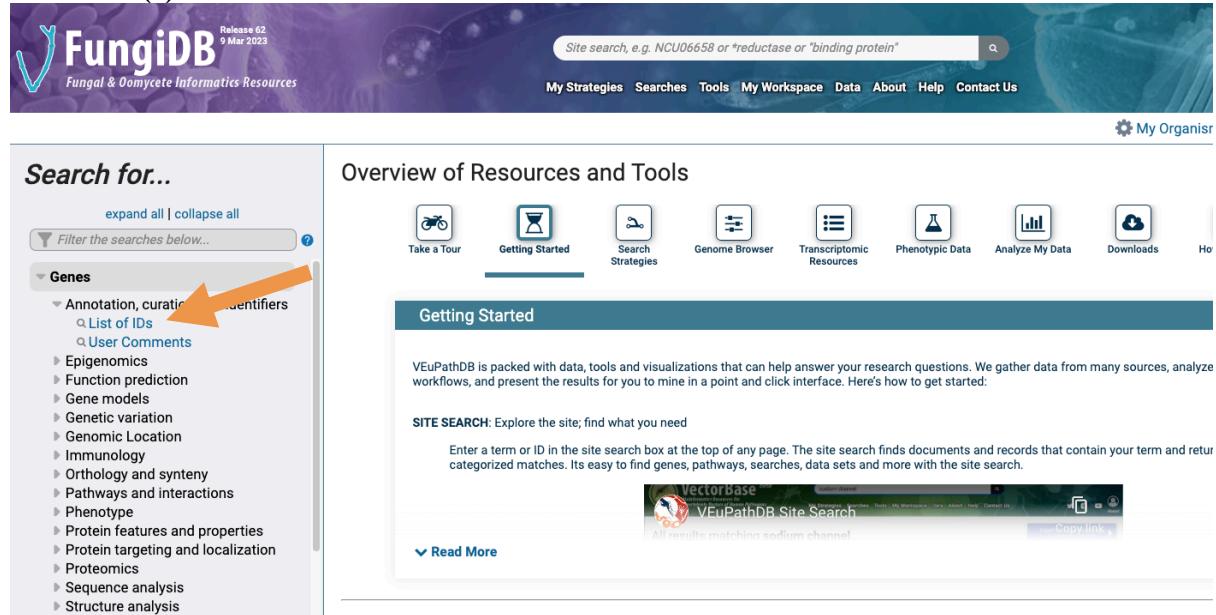
Advanced Search:	
Step 1: Select strain (REQUIRED) • Select a strain to limit search results Candida glabrata CBS138  	
Step 2: Select chromosomal feature (REQUIRED) • Select one or more feature types <input type="checkbox"/> ORF <input type="checkbox"/> repeat_region <input type="checkbox"/> autocatalytically_spliced_intron <input type="checkbox"/> retrotransposon <input type="checkbox"/> blocked_reading_frame <input type="checkbox"/> snRNA <input type="checkbox"/> centromere <input type="checkbox"/> snoRNA <input type="checkbox"/> long_terminal_repeat <input type="checkbox"/> tRNA <input type="checkbox"/> multigene_locus <input type="checkbox"/> telomeric_repeat <input type="checkbox"/> ncRNA <input type="checkbox"/> not_in_systematic_sequence <input type="checkbox"/> pseudogene <input type="checkbox"/> rRNA <input checked="" type="checkbox"/> Select all chromosomal features 	
Step 3: Narrow results (OPTIONAL) • Select search criteria to return specific types of genes. Results will match all selected criteria. • Select search criteria by clicking on a checkbox, filling in a dialog box, or selecting a menu option. • Select or unselect multiple options for Chromosomes and GO terms by pressing the Control (PC) or Command (Mac) key while clicking. Annotation/sequence properties: Is a feature that is AND <input type="checkbox"/> Alternatively_spliced <input type="checkbox"/> Dubious <input type="checkbox"/> Uncharacterized <input type="checkbox"/> Verified <input type="checkbox"/> not_physically_mapped <input type="checkbox"/> transposable_element_gene <input type="checkbox"/> Merged/Split <input type="checkbox"/> Deleted <input type="checkbox"/> Deleted_from_Assembly_20 <input type="checkbox"/> Deleted_from_Assembly_21 The default search excludes Deleted features. Has introns (excluding UTR introns) <input type="checkbox"/> Yes <input type="checkbox"/> No AND Is on the following chromosome or contig sequence(s): AND (The "All" option includes unmapped features; to specifically exclude unmapped features, select each of the chromosomes of interest rather than "All") ChrJ_C_glabrata_CBS138 ChrK_C_glabrata_CBS138 ChrL_C_glabrata_CBS138 ChrM_C_glabrata_CBS138 mito_C_glabrata_CBS138 	

- 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.

CaglfMt30	tRNA: Uncharacterized	tL(UAA)4mt	Mitochondrial leucine tRNA, has UAA anticodon	mito_C_glabrata_CBS138:17616 to 17697 GBrowse	Relative Coordinates	Chromosomal Coordinates
				Noncoding_exon 1 to 82	17,616 to 17,697	
Sort by : Systematic Name <input style="float: right;" type="button" value="Go!"/>						
Analyze gene list: further analyze the gene list displayed above or download information for this list						
Further Analysis:	GO Term Finder Find common features of genes in list	GO Slim Mapper Sort genes in list into broad categories	View GO Annotation Summary View all GO terms used to describe genes in list			
Download:	<input type="button" value="Download All Search Results"/> Download all the data retrieved by query	<input type="button" value="Batch Download"/> Download selected information for entire gene list. Available information types include Sequence, Coordinates, GO annotations, Phenotype.				
Result Page : 1 2 Next						

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

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



The screenshot shows the FungiDB homepage. At the top, there's a banner with the text "Release 62 9 Mar 2023". Below the banner is a search bar with placeholder text "Site search, e.g. NCU06658 or *reductase or *binding protein". To the right of the search bar are links for "My Strategies", "Searches", "Tools", "My Workspace", "Data", "About", "Help", and "Contact Us". On the far right, there's a "My Organism" link. The main content area is divided into two columns. The left column is titled "Search for..." and contains a sidebar with sections like "Genes", "Epigenomics", "Function prediction", etc., and a "List of IDs" option under "Annotation, curation and identifiers". An orange arrow points to the "List of IDs" link. The right column is titled "Overview of Resources and Tools" and includes sections for "Getting Started", "Search Strategies", "Genome Browser", "Transcriptomic Resources", "Phenotypic Data", "Analyze My Data", and "Downloads". There's also a "Take a Tour" button.

- 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”.

Gene ID input set

Enter a list of IDs or text:

Upload a text file: No file chosen
Maximum size 10MB. The file should contain the list of

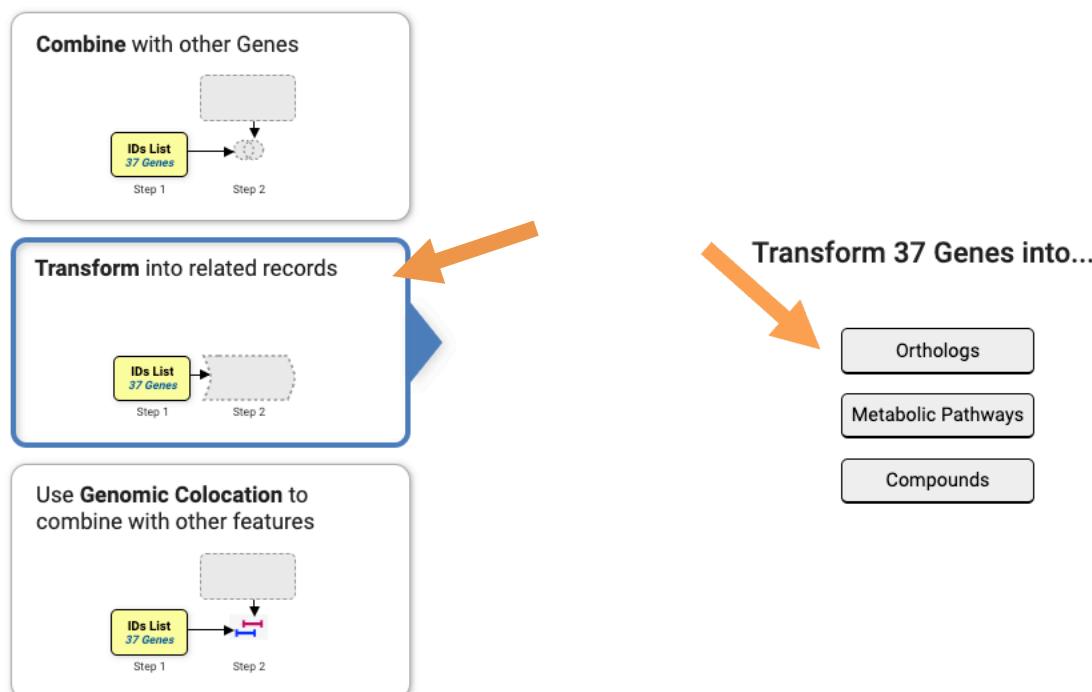
Upload from a URL:

The URL should resolve to a list of IDs

- 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.”



Add a step to your search strategy [?](#)



- 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 Results		Genome View		Analyze Results						
Advanced Paging										
Gene	ID	Transcript	ID	Organism	Genomic Location (Gene)	Product Description	Input Ortholog(s)	Ortholog Group	Paralog count	Ortholog count
Q0130	Q0130-t26_1	<i>S. cerevisiae</i> S288c		KP263414:46,723..46,953(+)	F0 ATP synthase subunit c	CaglMp10	OG5_126818		0	78
Q0045	Q0045-t26_1	<i>S. cerevisiae</i> S288c		KP263414:13,818..26,701(+)	cytochrome c oxidase subunit 1	CaglMp04, CaglMp07	OG5_128358		1	43
Q0070	Q0070-t26_1	<i>S. cerevisiae</i> S288c		KP263414:13,818..23,167(+)	intron-encoded DNA endonuclease aif5 alpha	CaglMp04, CaglMp07	OG5_128358		1	43
Q0105	Q0105-t26_1	<i>S. cerevisiae</i> S288c		KP263414:36,540..43,647(+)	cytochrome b	CaglMp03	OG5_128504		1	31
Q0120	Q0120-t26_1	<i>S. cerevisiae</i> S288c		KP263414:36,540..42,251(+)	intron-encoded RNA maturase bl4	CaglMp03	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.

Identifier you provided	Match	symbol	organism short name	name	length	secondary identifier	primary identifier	class
Q0060		A13	<i>S. cerevisiae</i>		6179	Q0060	S000007263	ORF
Q0070		A15_ALPHA	<i>S. cerevisiae</i>		9350	Q0070	S000007265	ORF

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**.

[Read more](#)

Query for interactions:

- Gene → Complex + Details
- Gene → Genetic Interactions
- Gene → Physical Interactions
- Literature → Interaction
- Complex → Details + Participants

[» More queries](#)


popular templates

- 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.

Gene → **Interaction**
Retrieve all interactions for a specified gene.

Gene

LOOKUP: act1

constrain to be IN ↗ saved Gene list List 1: S. cerevisiae orthologs

Show Results

Edit Query

web service URL | Perl | Python | Ruby | Java [help] | export XML



- 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.

Trail: Query
Gene → **Genetic Interactions**
Retrieve all genetic interactions for a specified gene.

Manage Columns | Manage Filters | Manage Relationships

Showing 1 to 25 of 70 rows

Gene Primary DBID	Gene Standard Name	Gene Systematic Name	Gene Sgd Alias	Gene Name	Interaction Detection Methods Identifier
S000007260	COX1	Q0045	cytochrome c oxidase subunit 1 OXI3	Cytochrome c Oxidase	Synthetic lethality
S000007260	COX1	Q0045	cytochrome c oxidase subunit 1 OXI3	Cytochrome c Oxidase	Synthetic lethality
S000007260	COX1	Q0045	cytochrome c oxidase subunit 1 OXI3	Cytochrome c Oxidase	Synthetic lethality
S000007260	COX1	Q0045	cytochrome c oxidase subunit 1 OXI3	Cytochrome c Oxidase	Synthetic lethality

7 Interaction Term Identifiers

Interaction Term Identifier	Count
Dosage Rescue	26
Synthetic Rescue	23
Synthetic Lethality	11
Phenotypic Suppression	5
Dosage Lethality	2
Phenotypic Enhancement	2
Synthetic Growth Defect	1

Filter | Download data

- 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***”.

Showing 1 to 9 of 9 rows

Gene Primary DBID	Gene Standard Name	Gene Systematic Name	Gene Sgd Alias	Gene (1 Gene)	Gene > Organism (1 Organism)	Gene > Interactions > Details (9 Interaction Details)	Gene > Interactions > Participant 2 (9 Genes)	Gene > Interactions > Details > Experiment > Interactant Detection Methods (1 Interaction Term)	Gene > Interactions > Details > Experiment (1 Interaction Experiment)	Experiment Name	Details interactionType
S000007260	COX1	Q0045		cytochrome c oxidase subunit 1 OXI3						Deutscher D, et al. (2006)-16941010-Synthetic Lethality	genetic
S000007260	COX1	Q0045		cytochrome c oxidase subunit 1 OXI3	Cytochrome c Oxidase	<i>S. cerevisiae</i>	genetic interactions	inversible	Bait		
										Synthetic Lethality	
										Deutscher D, et al. (2006)-16941010-Synthetic Lethality	genetic

- 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).

SGD YeastMine Search and retrieve *S. cerevisiae* data with YeastMine, populated by SGD and powered by InterMine. Last Updated on: Apr-14-2018

Contact Us Video Tutorials Help Log in

Home Templates Lists QueryBuilder Tools Regions Data Sources API MyMine

Upload | View Search: e.g. act1 GO

Lists

View your own and public lists, search by keyword and compare or combine the contents of lists. Click on a list to view graphs and summaries in an analysis page, select lists using checkboxes to perform set operations. Click 'Upload' above to import a new list.

Filter: Reset

Actions: Union | Intersect | Subtract | Asymmetric Difference | Copy Delete Options: Show descriptions Show Tags

You are not logged in. [Log in](#) to save lists permanently and to mark items as favourites ★.

List 2: Synthetic lethal interactors, *S. cerevisiae* 9 Genes MY

List 1: *S. cerevisiae* orthologs 11 Genes MY

All Curated Macromolecular Complexes 594 Molecular Complexes

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

List Analysis for List 2: Synthetic lethal interactors, *S. cerevisiae* (9 Genes)

Showing 1 to 9 of 9 rows

Gene Primary DBID	Gene Systematic Name	Gene Organism . Short Name	Gene Standard Name	Gene Name
S00000773	YEL047C	<i>S. cerevisiae</i>	FRD1	Fumarate ReDuctase

A. fischeri NRRL 181 (13) A. flavus NRRL3357 (15) A. fumigatus Af293 (13) A. gambiae (6) A. nidulans FGSC A4 (16) A. niger ATCC 1015 (18) C. albicans SC5314 (7) C. albicans WO-1 (7) C. dubliniensis CD36 (7) C. elegans (21) C. gattii VGII R265 (10) C. gattii WM276 (12) C. krusei ATCC 6259 (10) C. neoformans var. neofomans JEC21 (10) C. parapsilosis CDC317 (10) C. neofomans var. neofomans JEC21 (10) C. parapsilosis CDC317 (10) C. posadasii C735 delta SW0gp (11) D. melanogaster (30) D. rerio (21) I. galbrait CBS 138 (10) H. capsulatum G166AR (12) H. capsulatum NAM1 (12) H. sapiens (20) M. musculus (20) M. oryzae 70-15 (12) N. crassa OR74A (12) R. norvegicus (22) S. cerevisiae (6) S. pombe (8) T. marneffei ATCC 18224 (16) U. maydis 521 (12)

View homologues in other Mines:

FlyMine D. melanogaster x

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.

The screenshot shows the FungiDB Gene Results interface. At the top, there are tabs for 'Gene Results', 'Genome View', and 'New Analysis'. Below these, a message says 'Analyze your Gene results with a tool below.' A box contains a 'Gene Ontology Enrichment' tool, which features a large blue 'GO' logo surrounded by small red circles representing biological processes. An orange arrow points to the 'GO' logo. On the left side of the page, there is a vertical bar with the text 'Hide organism filter'.

- 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

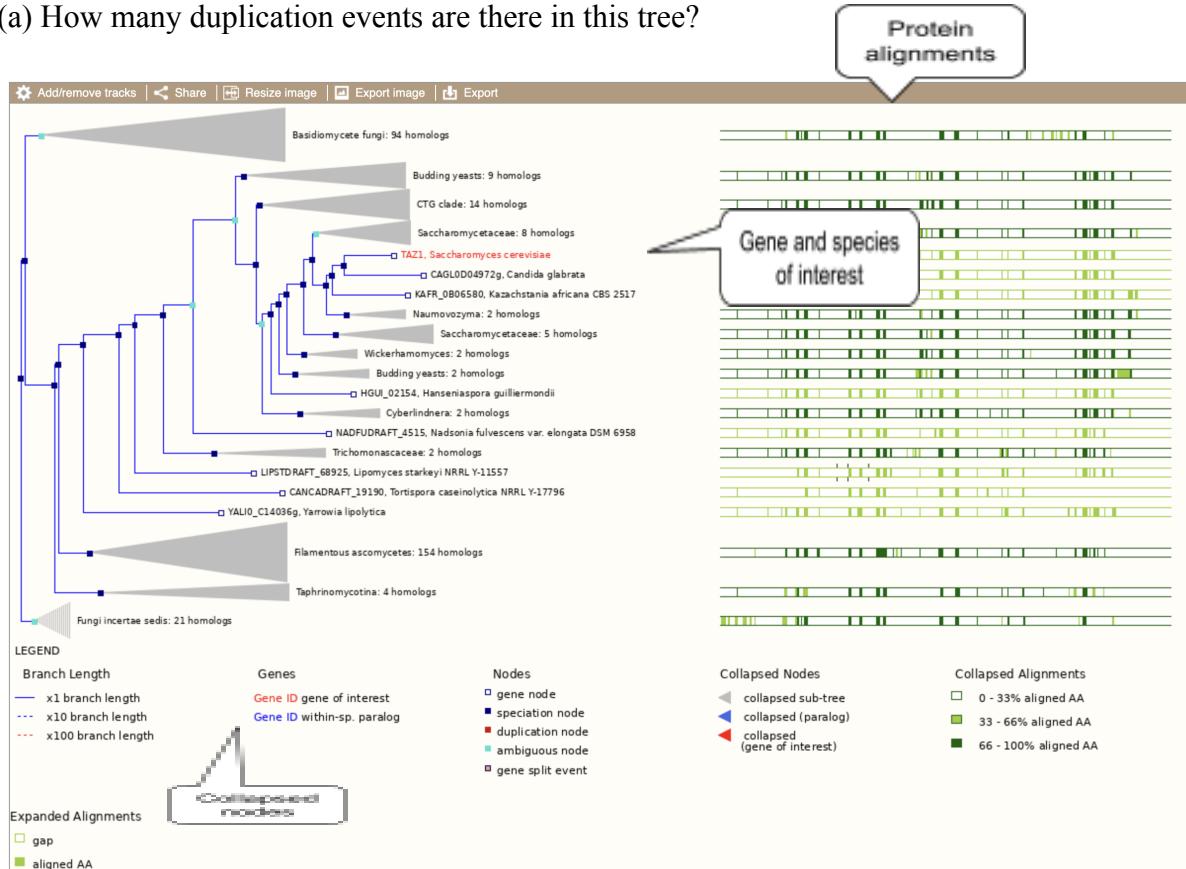
Let's look at the homologues of *Saccharomyces cerevisiae* TAZ1 (Gene stable ID: YPR140W). This gene is involved in stress response and conserved across different taxonomic domains. Search for the gene and go to the Gene tab.

The screenshot shows the Ensembl Fungi interface for *Saccharomyces cerevisiae* (R64-1-1). The main panel displays the gene information for TAZ1 (YPR140W). On the left, there is a sidebar titled "Gene-based displays" with several sections: 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), and Pan-taxonomic Compara (Gene Tree, Orthologues). The "Gene tree" section is currently selected. The main panel shows the gene details: Gene: TAZ1 YPR140W, Description, Location, About this gene, Transcripts, and a large "Gene tree" button.

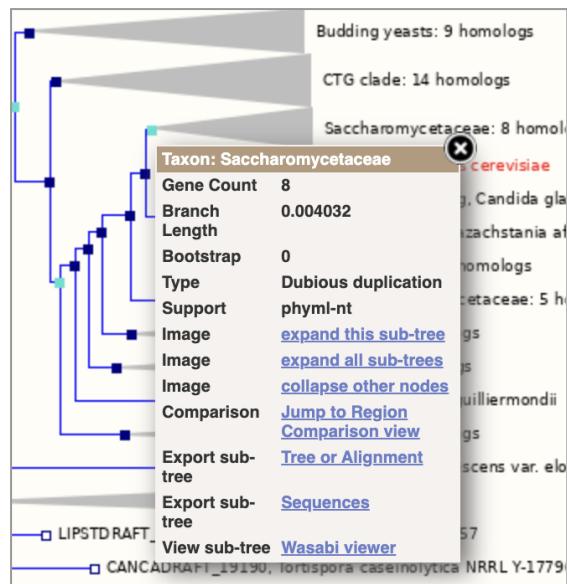
Click on **Fungal Compara: Gene tree**, which will display the current gene in the context of a phylogenetic tree used to determine orthologues and paralogues.

The screenshot shows the "Gene tree" page for TAZ1. At the top, it displays the GeneTree ID EFGT01050000064920 and two summary statistics: Number of genes (327) and Number of speciation nodes (295). Below this is a table of annotations, with a callout pointing to the "Show 10 entries" button and the "Filter tree by Gene Ontology (GO) terms or InterPro protein domains" input field. The main table lists 121 entries of highlight annotations, each with an accession number and a description. The first few entries include molecular_function, catalytic activity, lipid metabolic process, phospholipid metabolic process, phosphorus metabolic process, phosphate-containing compound metabolic process, biological_process, metabolic process, cellular process, and transferase activity.

(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:

View options:

- [View current gene only](#) (Default)
- [View paralogues of current gene](#)
- [View all duplication nodes](#)
- [View fully expanded tree](#)
- Collapse all the nodes at the taxonomic rank

Use the 'configure page' link in the left panel to see more options available from menus on individual tree nodes.

Ensembl Fungi release 56 - Feb 2023 © EMBL-EBI

About Us

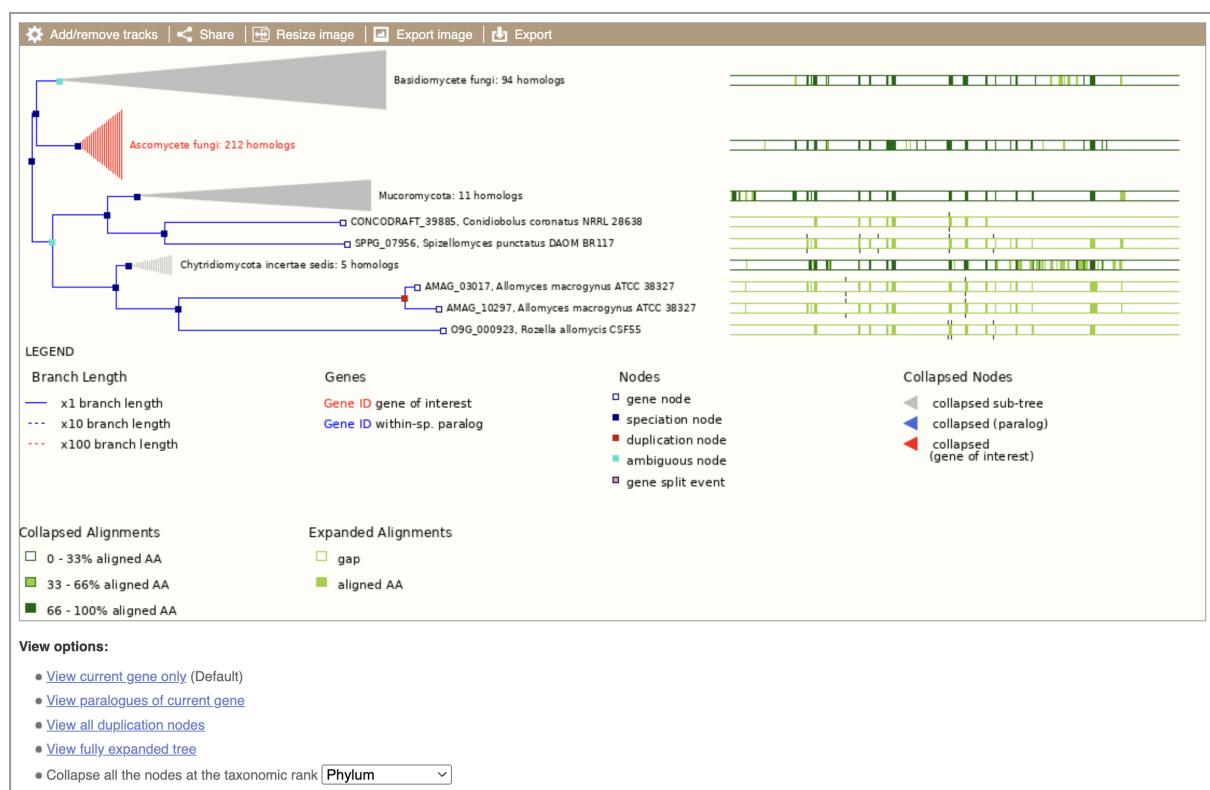
[About us](#)

[Using this website](#)

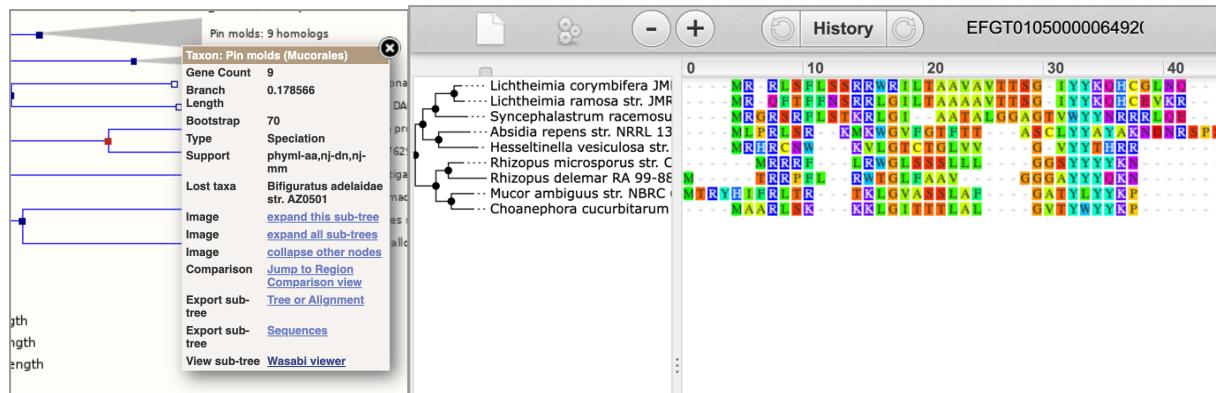
Our sister sites

[Ensembl](#)

(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)	<input checked="" type="checkbox"/>	298	14	0	58
Acidomyces (1 species)	<input type="checkbox"/>	1	0	0	0
Agaricales (21 species)	<input type="checkbox"/>	18	1	0	2
Atheliales (2 species)	<input type="checkbox"/>	1	1	0	0
Blastocladiales (1 species)	<input type="checkbox"/>	0	1	0	0
Boletales (9 species)	<input type="checkbox"/>	6	0	0	3
Botryosphaerales (2 species)	<input type="checkbox"/>	2	0	0	0
Cantharellales (3 species)	<input type="checkbox"/>	1	1	0	1

Select taxon of interest

Download table

Similarity metrics

Filter table

Orthologue details by species

Selected orthologues [Hide](#)

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Absidia repens str. NRRL 1336	1-to-1	BCR42DRAFT_405738 View Gene Tree View Sequence Alignments	23.74 %	17.32 %	n/a	n/a	No
Acaromyces ingoldii str. MCA 4198	1-to-1	FA10DRAFT_281454 View Gene Tree View Sequence Alignments	24.93 %	n/a	n/a	n/a	No
Acidomyces richmondensis BFW	1-to-1	M433DRAFT_12235 View Gene Tree View Sequence Alignments	27.41 %	20.25 %	n/a	n/a	Yes
Acromonium chrysogenum ATCC 11550	1-to-1	ACRE_050350 View Gene Tree View Sequence Alignments	32.27 %	n/a	n/a	n/a	Yes
Agaricus bisporus var. burrettii JB137-S8	1-to-1	AGABI1DRAFT_91626 View Gene Tree View Sequence Alignments	29.32 %	23.62 %	n/a	n/a	No

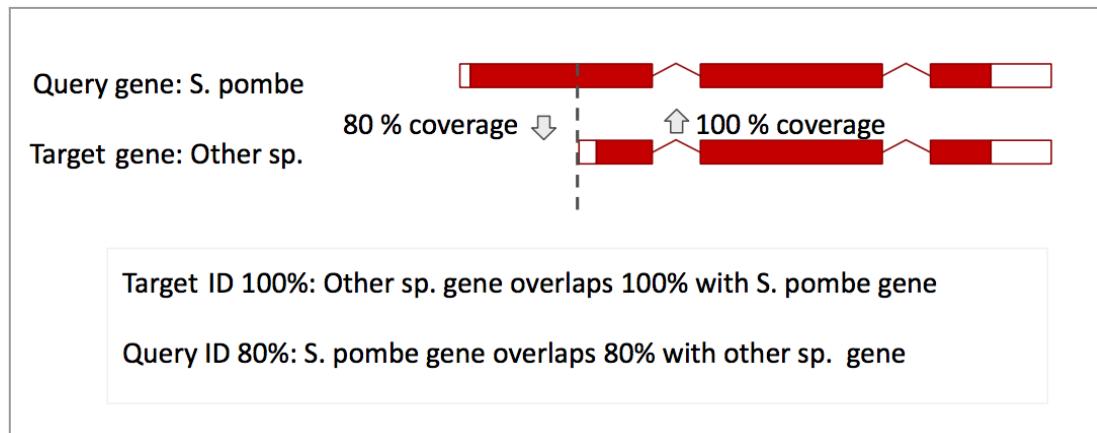
Link to orthologue gene tab

View region comparison of orthologues

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!

Species without orthologues

58 species are not shown in the table above because they don't have any orthologue with YPR140W.

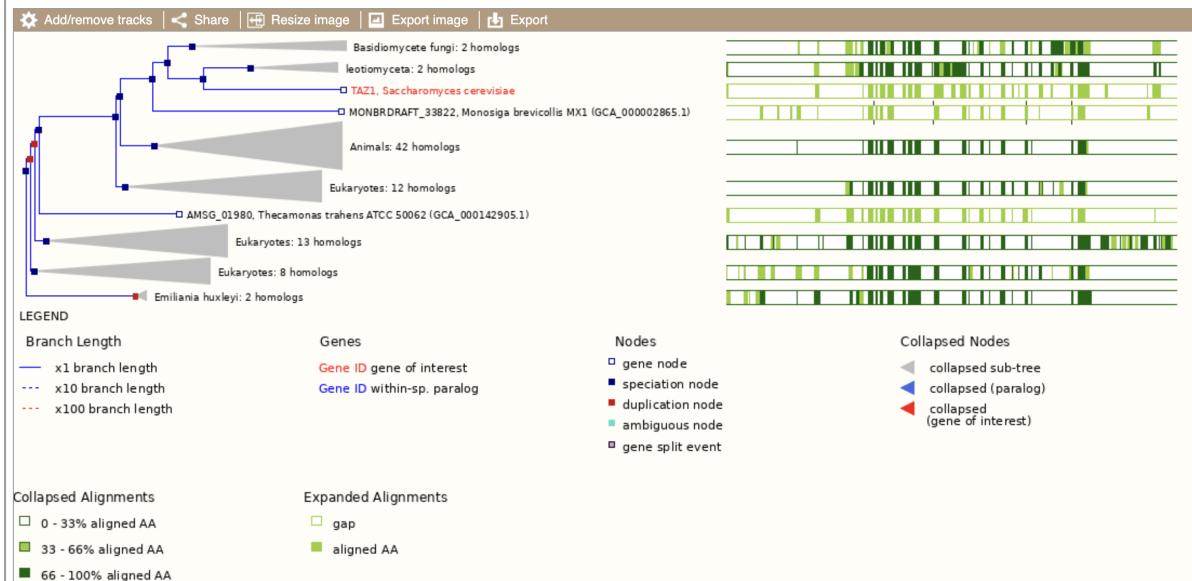
- *Amphiamblus* sp. WSBS2006
- *Anncalilia algerae* PRA339
- *Aspergillus flavus* NRRL3357
- *Aspergillus nidulans*
- *Aspergillus terreus* NIH2624
- *Batrachochytrium dendrobatidis* JEL423
- *Bifiguratus adelaideae* str. AZ0501

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 a look at the Pan-taxonomic tree with nodes collapsed at the Kingdom rank.

Gene Tree

GeneTree EGGT0005000021121

Number of genes	84
Number of speciation nodes	64
Number of duplication nodes	16
Number of ambiguous nodes	3
Number of gene split events	0



Click on Pan-taxonomic Compara > Orthologues now.

Orthologues

Download orthologues

Summary of orthologues of this gene [Hide](#)

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

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (370 species)	<input checked="" type="checkbox"/>	4	0	0	366
Acidomyces (1 species)	<input type="checkbox"/>	0	0	0	1
Agaricales (21 species)	<input type="checkbox"/>	0	0	0	21
Atheliales (2 species)	<input type="checkbox"/>	0	0	0	2
Blastocladiales (1 species)	<input type="checkbox"/>	0	0	0	1
Boletales (9 species)	<input type="checkbox"/>	0	0	0	9

Selected orthologues [Hide](#)

Show All entries		Show/hide columns				Filter	
Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Aedes aegypti (LVP_AGWG)	1-to-1	AAEL001564	25.85 %	19.95 %	n/a	n/a	No
	View Gene Tree	2:21,496,991-21,541,309:-1					
	View Sequence Alignments						
Amborella trichopoda	1-to-1	AMTR_s00022p00068080	23.08 %	17.32 %	n/a	n/a	No
	View Gene Tree	AmTr_v1.0_scaffold00022:710,032-717,504:-1					
	View Sequence Alignments						

(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)	<input checked="" type="checkbox"/>	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)	<input checked="" type="checkbox"/>	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?

Selected orthologues [Hide](#)

Show All entries		Show/hide columns		human				
Species	Type	Orthologue		Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Human	1-to-1	TAFazzin (ENSG00000102125)	View Sequence Alignments	24.66 %	18.90 %	n/a	n/a	No
		X:154,411,524-154,421,726:1						
			Orthologue Alignment					
Pediculus humanus	1-to-1	PHUM309640	View Protein Alignment	18.90 %	n/a	n/a	n/a	No
		DS235308:45,836-47,144:-1	View cDNA Alignment					
			View Sequence Alignments					

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----PLWRFLSYSTSLLTGVSKLLLFTCYNV
ENSP00000469981/1-292 -----MPLHVKW-----PFP---AVPPLTWTLASSVVMGLVGTYSFCFWTKYMNHL
. : *           * : * . * . * :   :
```

```
YPR140W/1-381      KLNQFEKLETALERSKRENRGLMTVMNHMSMVDDPLVVATLPYKLFTSLDNIRWSLGAHN
ENSP00000469981/1-292 TVHNREVLYELIEK-RGPATPLITVSNHQSCMDDPHLGILKLRHIWNLKLMRWTPAAAD
. ::. * * ;*: : *;** ** * ;*** :*. * : : .*. ;**: .* :
```

```
YPR140W/1-381      ICFQNKFPLANFFSLGQLSTER-----FGVGPQGS
ENSP00000469981/1-292 ICFTKELHSHFFSLGKCVPCRGAEFFQAENEKGKVLDTGRHMPGAGKRREKGDGVYQKG
*** :: :;*****: :.. *
```

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.

The screenshot shows the BioMart search interface. The search bar contains "Cryptococcus neoformans var. neoformans JEC21" and the search term "CNC06590". Below the search bar is a "Go" button and a note "e.g. NAT2 or alcohol*".

(a) Does this gene in *C. neoformans* have a UniProtKB-Gene Ontology annotation?

The screenshot shows the Ensembl gene page for CNM01690. The page header includes the location (13:510,531-512,507), gene ID (CNM01690), and transcript ID (AAW46801). The left sidebar has a tree menu with "External references" selected. The main content area displays the gene's description, location (Chromosome 13: 510,531-512,507 reverse strand, ASM9104v1 AE017353.1), and external references. It lists database identifiers for NCBI gene, UniGene, and UniProtKB-Gene Ontology Annotation, along with a table of transcripts and their corresponding database identifiers.

(b) Find the *Z. tritici* orthologue in the [Orthologues](#) page and view a protein alignment.

The screenshot shows the Orthologues page for Zymoseptoria tritici. A table lists orthologues for the species Zymoseptoria tritici, with one entry for MGHOG1 (Mycgr3G76502). The table columns include Species, Type, Orthologue, Target %id, Query %id, GOC Score, WGA Coverage, and High Confidence. The "Orthologue Alignment" link is highlighted in a tooltip. The "Zymoseptoria tritici" column header has a dropdown arrow icon.

(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 Mycgr3P76502/1-357		MADEFVKLSIFGTVFEVTTTRYVDLQPVGMGAFLGLVCSAKDQLSCTSVAIKKIMKPFPSTPVLM MAEFVRAQIFGTTFEITTSRYTDLQPVGMGAFLGLVCSAKDQLTGQAVAVKKIMKPFPSTPVLM ***:*** .*****.*;*:***.*****:*****:*****:*****:*****:*****:*****:*****				
AAW42642/1-365 Mycgr3P76502/1-357		SKRTYRELKLLKHILRHENIISLSDIFISPLEDIYFVTTELLGTDLHRLLTSPLEKQFIQY SKRTYRELKLLKHILHENVIISLSDIFISPLEDIYFVTTELLGTDLHRLLTSPLEKQFIQY *****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****				
AAW42642/1-365 Mycgr3P76502/1-357		FLYQILRGLKYVHSAGVVHRLDKPSNILVNENCDLKICDFGLARIQDPQMTGYVSTRYYR FLYQILRGLKYVHSAGVVHRLDKPSNILVNENCDLKICDFGLARIQDPQMTGYVSTRYYR *****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****				
AAW42642/1-365 Mycgr3P76502/1-357		APEIMLTWQKYDVAVDIWSTGCIFAEMLEGKPLFPKGDKHVNQFSIITELLGTPPPDDVIQT APEIMLTWQKYDVEDIWSACCIPAEMLEGKPLFPKGDKHVNQFSIITDLLGTPPPDDVIST *****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****				
AAW42642/1-365 Mycgr3P76502/1-357		IASENTLRFVQSLPKREKVPFSTKFPNADPVSLDLLEKMLVFDPRTRISAAEGLAHEYLA ICSENTLRFVOSLPKRERQPLKNKFKNADPQAIELLERMLVFDPRKRVAGEALADPYLS *.*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****				
AAW42642/1-365 Mycgr3P76502/1-357		PYHDPTDEPVAAEVFDWSFNDADLPVDTWKVMMYSEILDFHNLGDISQNE--AEGPVVTGE PYHDPTDEPEAEERKDWSFNDADLPVDTWKIMMYSEILDYHNVDS-ANNGEGQE--NGG *****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****				
AAW42642/1-365 Mycgr3P76502/1-357		VPAAPAS A-----				
		.				

Additional Exercise 2 - Mushroom genes

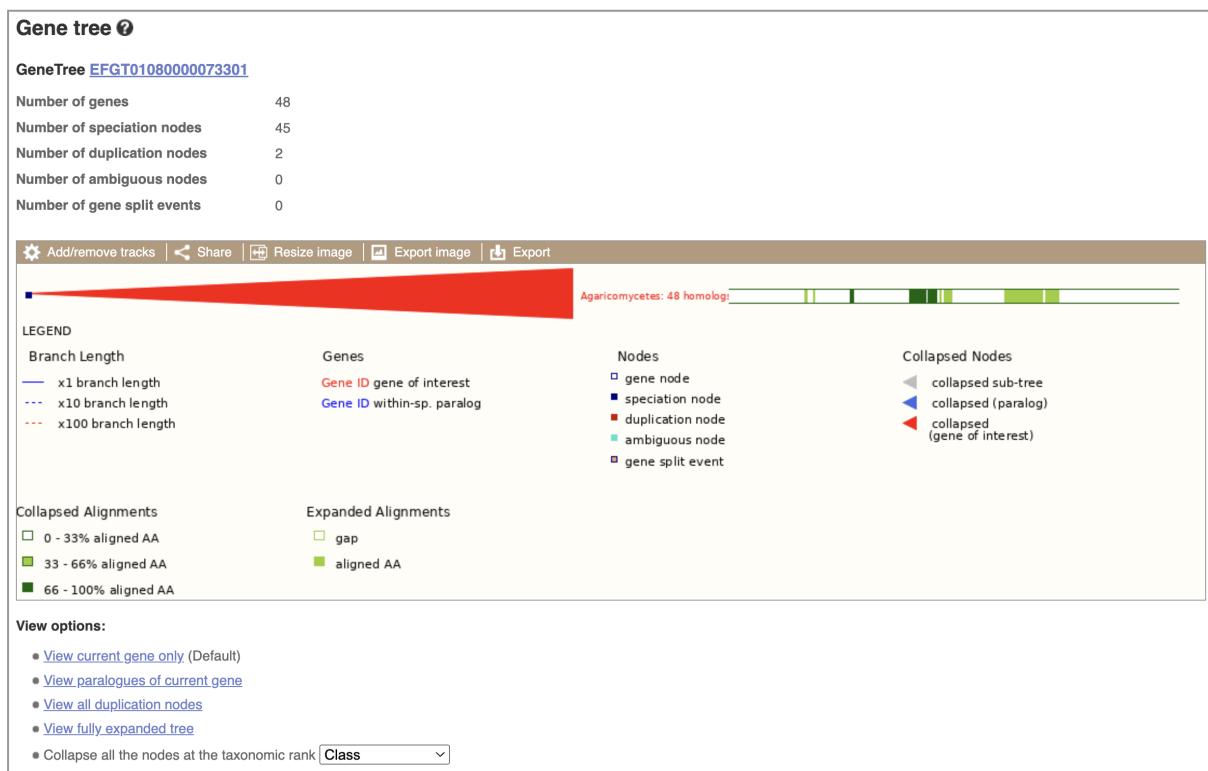
We're going to take a look at the gene [CC1G_05700](#) in *Coprinopsis cinerea* okayama7#130.

Search: for

e.g. [NAT2](#) or [alcohol](#)*

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*)

Gene: CC1G_05700

Description	basidiospore development protein
Location	Chromosome 7: 2,117,260-2,118,876 forward strand. CC3:AACS02000007.1
About this gene	This gene has 1 transcript (splice variant) and 47 orthologues .
Transcripts	Hide transcript table

Show/hide columns (1 hidden)		Filter	Export		
Name	Transcript ID	bp	Protein	Biotype	Flags
-	EAU90162	1389	462aa	Protein coding	Ensembl Canonical

(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?

Selected orthologues Hide								
Show All entries Amanita thiersii 								
Show/hide columns								
Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence	
Amanita thiersii	1-to-1	AMATHDRAFT_122148	42.73 %	10.17 %	n/a	n/a	No	
Skay4041		KZ301993:102,546-102,928:-1						
		View Sequence Alignments	Orthologue Alignment View Protein Alignment View cDNA Alignment					

Orthologue alignment ②

 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      -----MRVLLHDHQMNLEKFSGHVEALISNVKETSQELRKTSSTFEQEHDKLLG
PFH51030/1-110      PLTFLDKNATSMRVLHLDTQANFEKFSTRVNDFNLGAETKESEINLVKSLSLFERGOETLTN
                     ***** *:**** :*:: :...: **..*: ..* **: ::.* .

EAU90162/1-462      DIIDLVNRCSQIQLGSPPAQASGMQLSKDINQLDCLDKRLDAIQTV-----
PFH51030/1-110      *****: .*: .:*****:;:: : .::: **:.*:****:;.

EAU90162/1-462      QIQAIQNLLQQQNLLINAVTPLLLQLPQLPRLAPSTSLANFNSQTQRTDASSQTIEKRO
PFH51030/1-110      ----

EAU90162/1-462      PSYHQETLRRQRVRVDSDIQEISPCKPLPGSAQKKRRIESPRSVQKPSLELTQRLFPSSSP
PFH51030/1-110      ----

EAU90162/1-462      DLIKYSTDSEGPKTPQVNERSAPIVTPRRPLQDLFPFFPGSNQRSVSKRPMPPSSTRLV
PFH51030/1-110      ----

EAU90162/1-462      GPGKSATPGPSRVGAESRAALARPLIKPLAIAPLAFSSTSKTPVHISNFTPKPVTPSL
PFH51030/1-110      ----

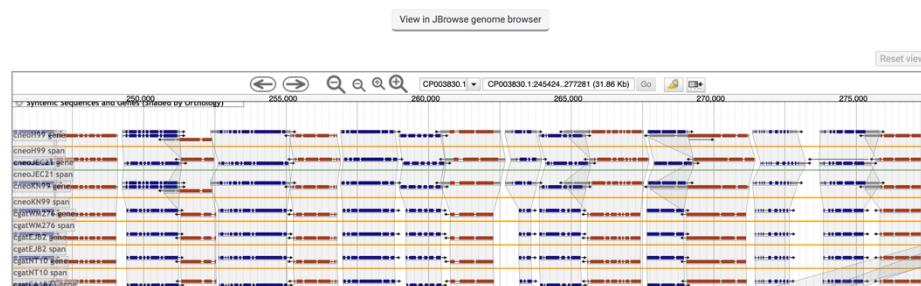
EAU90162/1-462      RNAVAGEGRALKIAQTPQVLKNERMTSQAAKNNTMPPAGMVSLRSSTTTATAKPTS
PFH51030/1-110      ----

EAU90162/1-462      NTPRFGPEANKPPLLRAPTNNGPRLQERMKEPVREGRRFIPLVTDDEDDSD
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?

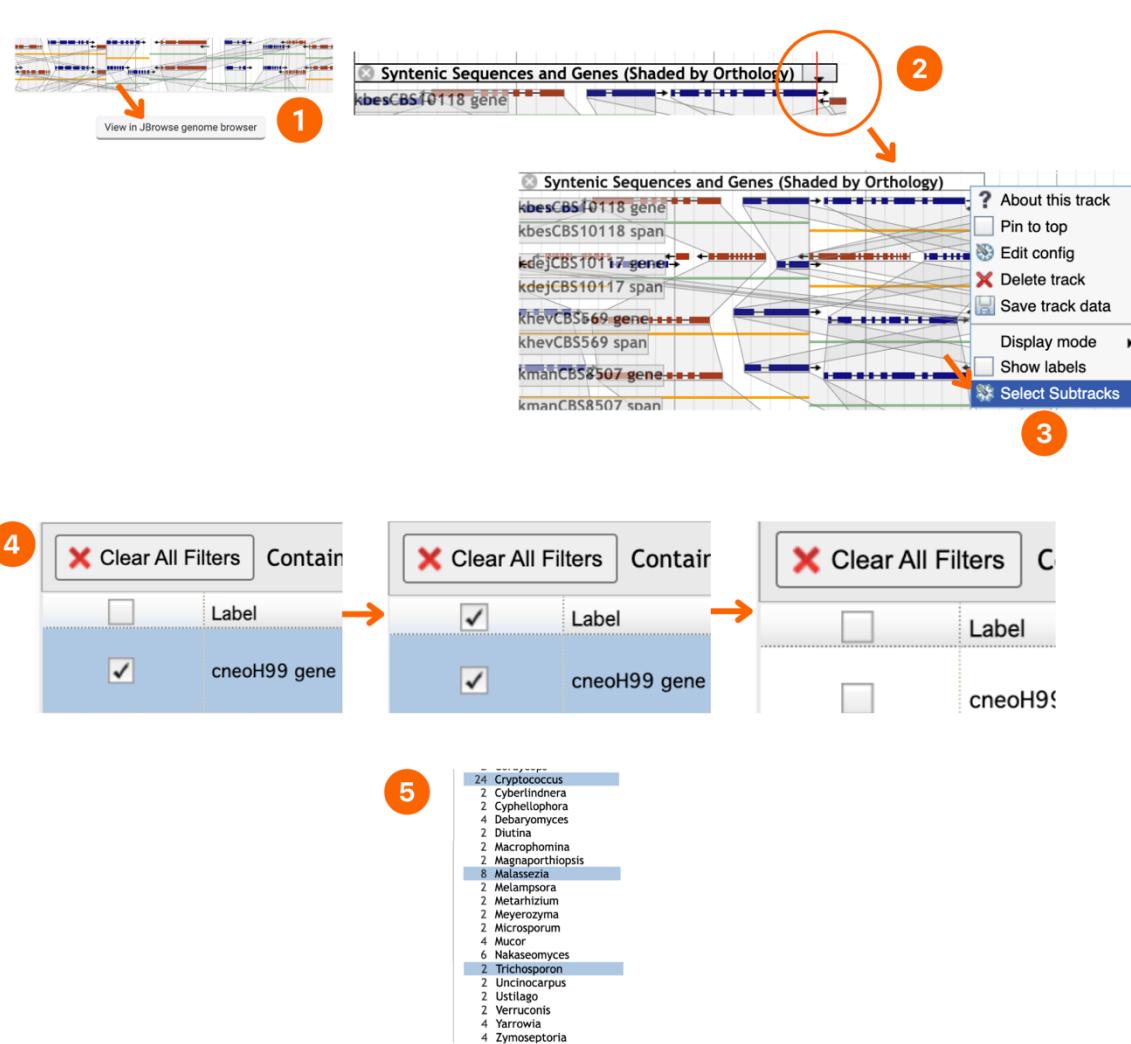
Ortholog Group OG6_531912

Orthologs and Paralogs within FungiDB Data sets

To run Clustal Omega, select genes from the table below. Then choose the sequence type and initiate the analysis.

Clustal Omega	Gene	Product
<input type="checkbox"/>	D1P53_003607	unspecified product
<input type="checkbox"/>	L203_05099	GATA-type domain-containing protein [Source:UniProtKB/Trembl;Acc:A0A1E3I9A1]
<input type="checkbox"/>	I314_00715	hypothetical protein
<input type="checkbox"/>	I306_00695	hypothetical protein

- 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.

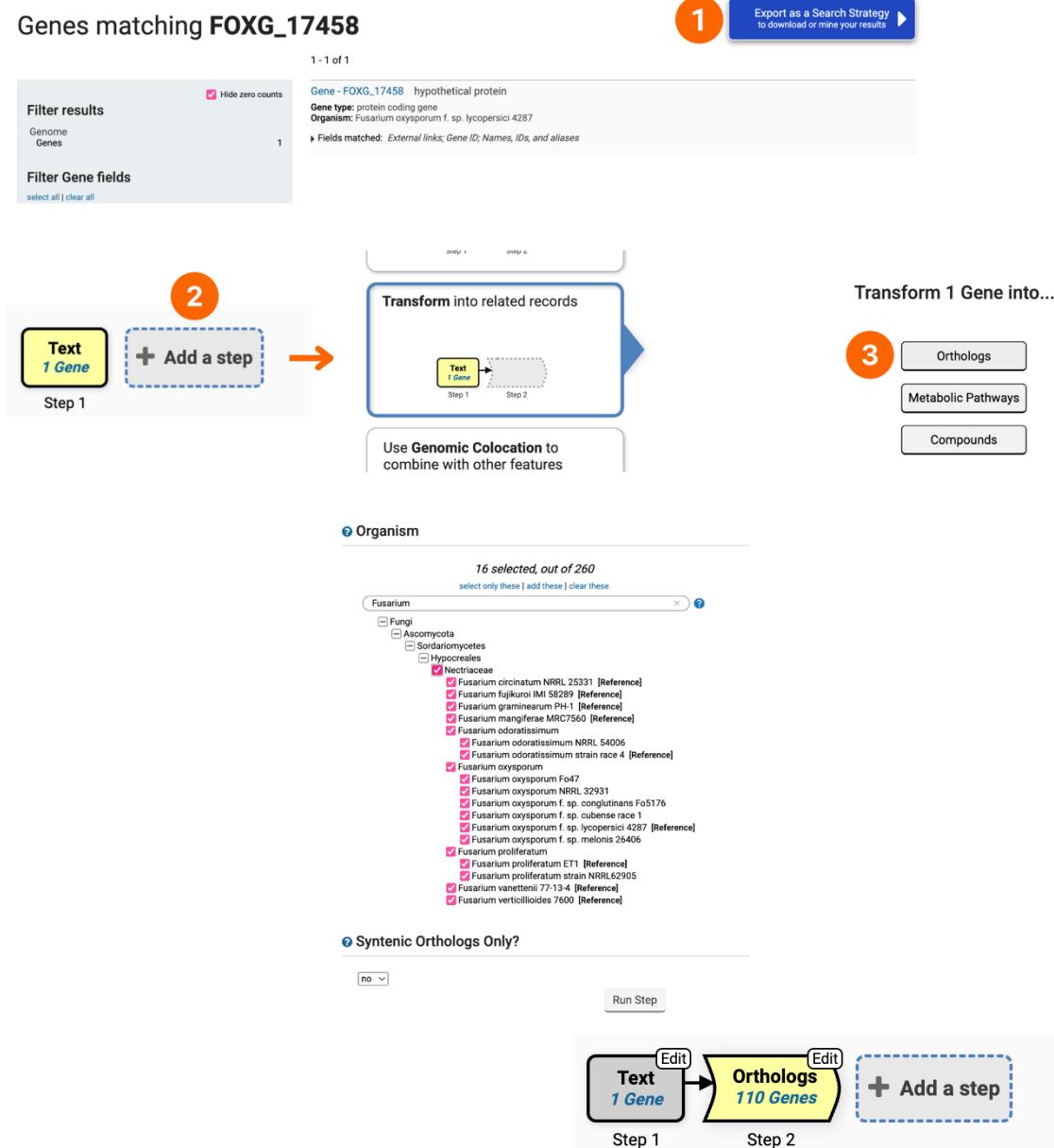


Notice that this hypothetical protein in *Fusarium oxysporum* f. sp. *lycopercisi* 4287 has no syntenic orthologs. Why do you think this is?

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.

The screenshot shows the FungiDB interface. At the top, there's a navigation bar with links for 'My Strategies', 'Searches', 'Tools', 'My Workspace', 'Data', 'About', 'Help', and 'Contact Us'. A user profile for 'Velina' is shown on the right. Below the header, a search bar contains 'FOGX_17458'. The main content area displays 'Genes matching FOXG_17458' with a single result: 'Gene - FOXG_17458 hypothetical protein'. The result details its type as a 'protein coding gene' from 'Fusarium oxysporum f. sp. lycopersici 4287'. There are also links for 'External links', 'Gene ID', 'Names', 'IDs', and 'aliases'. A button labeled 'Export as a Search Strategy' is available.



- Examine phyletic distribution by clicking on the “Ortholog Group” link within the results table.

This screenshot shows a results table with columns for 'Input Ortholog(s)' and 'Ortholog Group'. The 'Ortholog Group' column contains links to ortholog groups, with one link highlighted by an orange arrow. The table rows are as follows:

	Input Ortholog(s)	Ortholog Group
	FOGX_17458	OG6..115926

▼ Phylogenetic Distribution of Proteins [?](#) [Download](#)

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

Hide zero counts

fusarium

		x	?
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 (fpzo)	8		
Fusarium proliferatum strain NRRL62905 (fpzn)	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!



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 (Michelise 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., Micheli,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 "FOGX_17458T0" (Transcript 0 of FOXG_17458). You will find that the proteinID of FOXG_17458 in Fusox2 is 23236.

FOXG_17458T0	<input type="button" value="Search"/>	
Search By:	Across: Terms:	
Keywords	Default exact - fast	
<input type="button" value="Download"/> as CSV	compressed by Gzip	
Total genes found: 1 25		
Gene	Gene Ontology	Annotations
Portal: Fusox2 Portal Name: Fusarium oxysporum f. sp. lycopersici 4287 v2 Protein Id: 23236 Transcript Id: 23236 Location: Scaffold_51:76759-80046 (+) Model Name: FOXG_17458T0 Track: ExternalModels	GO:0003677 • DNA binding GO:0003700 • DNA-binding transcription factor activity GO:0005634 • nucleus GO:0006351 • DNA-templated transcription GO:0006355 • regulation of DNA-templated transcription GO:0008270 • zinc ion binding	PF04082 • Fungal specific transcription factor domain PF00172 • Fungal Zn(2)-Cys(6) binuclear cluster domain IPR007219 • IPR001138 • IPR002409 • missing_ipr002409

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

MCL Clusters • Fusarium oxysporum f. sp. lycopersici 4287 v2

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

Run: **Fusox2 comparative clustering.829**

Filter: **Fusarium-orthomcl 1.5.2900**

Rows: any all

Clusters: 76,126 Singletons: 5,743 Tracks: 5

Show Charts: Show Counters: Show Domains:

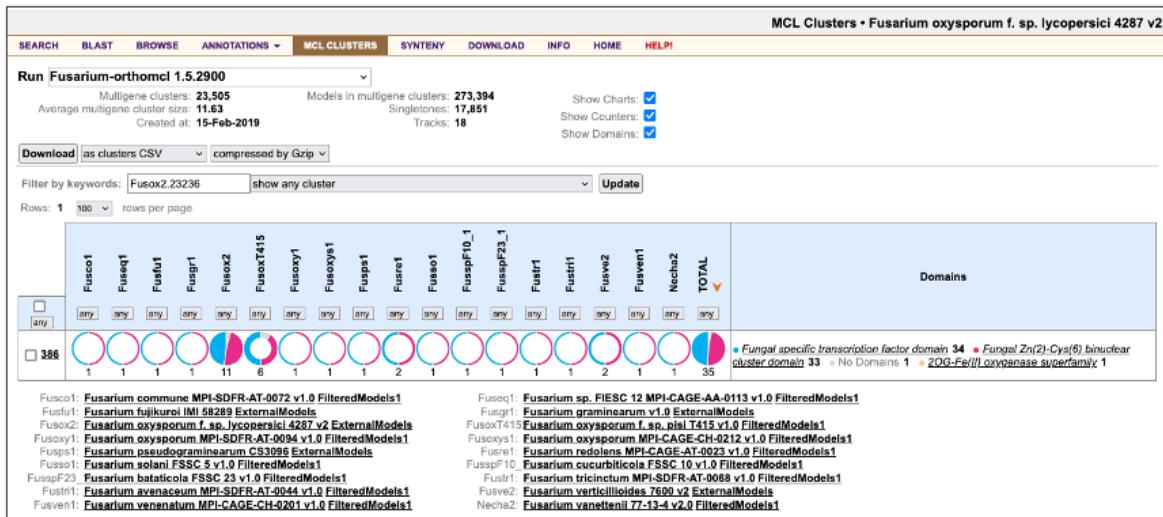
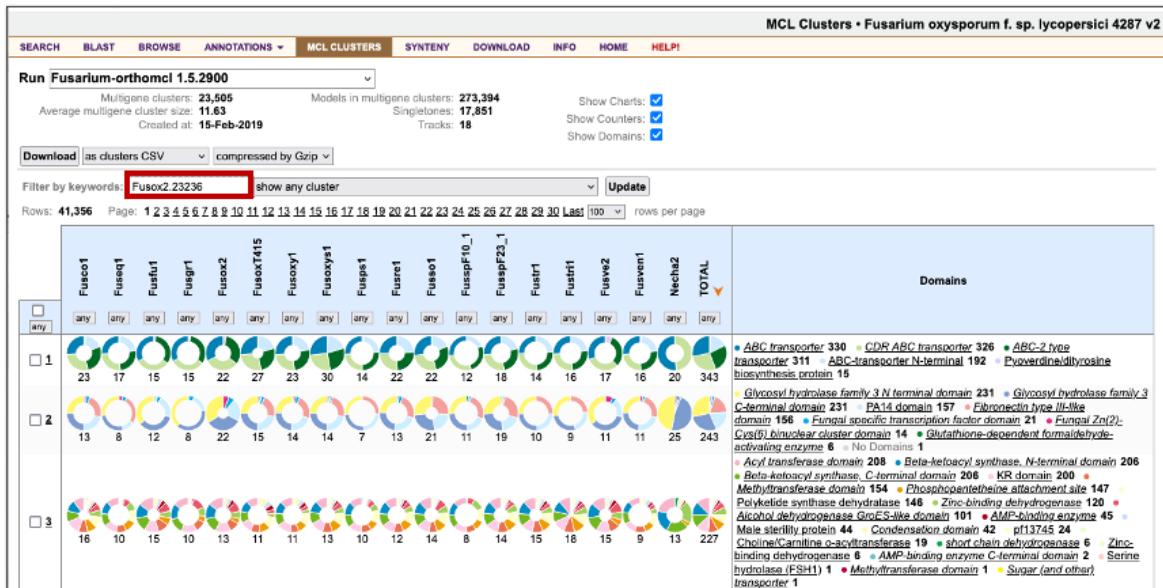
21 22 23 24 25 26 27 28 29 30 Last 100 rows per page

Domains

- Enzyme 363 • No Domains 6 • Zinc-binding dehydrogenase 3 • Shikimate / isoleucine 1
- Transporter family 262 • No Domains 95 • Sugar (and other) transporter 4 • ABC transporter 1
- Hydrolase 182 • No Domains 132 • Methyltransferase domain 41 • ABC transporter 1

<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 Fusoxyl1.

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:

Fusarium oxysporum (FO2) ▾

Location: 14:1,108,371-1,111,923 Gene: FOXG_17458 Transcript: FOXG_17458T0 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: Biological process
- GO: Molecular function
- GO: Cellular component
- PHI: Phibase identifier
- Phenotypes
- Genetic Variation
- Variant table
- Variant image
- Structural variants
- Gene expression
- Pathway
- Regulation
- External references

Gene: FOXG_17458

Description: conserved hypothetical protein [Source:BROAD_F_oxysporum;Acc:FOGX_17458]

Location: Chromosome 14: 1,108,371-1,111,923 forward strand. FO2:CM000602.1

About this gene: This gene has 1 transcript (splice variant), 318 orthologues, 15 paralogues and is a member of 2 Ensembl protein families.

Transcripts: Hide transcript table

Name	Transcript ID	bp	Protein	Biotype	UniProt	Flags
Novel	FOGX_17458T0	2832	943aa	Protein coding	J9NQH9	rf

Gene tree ?

GeneTree ENSGT0093000001158

Number of genes: 345
 Number of speciation nodes: 270
 Number of duplication nodes: 47
 Number of ambiguous nodes: 26
 Number of gene split events: 1
 Highlight annotations: Hide annotations table

Legend:

- Branch Length: $\times 1$ branch length, $\times 10$ branch length, $\times 100$ branch length
- Nodes:
 - Gene node (blue square)
 - Speciation node (black square)
 - Duplication node (red square)
 - Ambiguous node (green square)
 - Gene split event (grey square)
- Collapsed Nodes:
 - Collapsed sub-tree (grey triangle)
 - Collapsed (paralog) (blue triangle)
 - Collapsed (gene of interest) (red triangle)
- Collapsed Alignments:
 - 0 - 33% aligned AA (light green)
 - 33 - 66% aligned AA (medium green)
 - 66 - 100% aligned AA (dark green)

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:

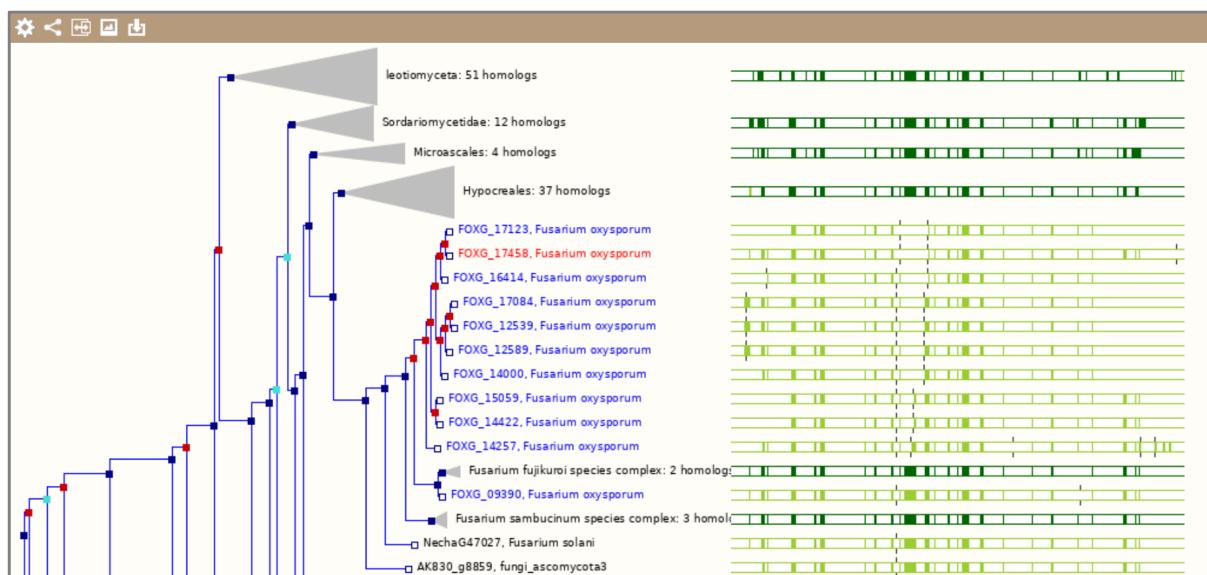
Expanded Alignments
gap
aligned AA
Add/remove tracks | Share | Resize image | Export image | Export

View options:

- View current gene only (Default)
- View paralogues of current gene
- View all duplication nodes
- View fully expanded tree
- Collapse all the nodes at the taxonomic rank -- Select a rank --

Use the 'configure page' link in the left panel to set the default. Further options are available from menus on individual tree nodes.

Ensembl Fungi release 56 - Feb 2023 © EMBL-EBI



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

Download data from this image

Ascomycetes: 5 homologs

leotiomyceta: 51 homologs

File name:	FOGX_17458_gene_tree
File format:	FASTA
Guide to file formats	
Preview Download Download Compressed	
CLUSTALW	<pre>>homo_sapiens 1-46588 CCTCAGGACCCGACGGCAAACCAACCGAA CCCCATGCTTGCAGACTGCCCTCTGGGCCC TGGGACAGAGAGAACCAACAGCTGGCTC AGGGGGCTTGGGTTAGATAACAAK CCAGAGCTGGATCTGGCTATTTGGGCCACCTC CCAGGCTCTGTGCAAAGAGGGTGGCTGTGTA AGGAAGACCTGGTGGCTCTGGCTGTGTT AAAGATGGGGTGGTTGGTGGATTCTCTT GGGAGGGAGGAGAGAAAAGGGCCCTGGGG CAGGGCTCTGGGCTCTGGCTCTGGCTCTGGG</pre>
FASTA	<pre>>homo_sapiens 1-46588 CCTCAGGACCCGACGGCAAACCAACCGAA CCCCATGCTTGCAGACTGCCCTCTGGGCCC TGGGACAGAGAGAACCAACAGCTGGCTC AGGGGGCTTGGGTTAGATAACAAK CCAGAGCTGGATCTGGCTATTTGGGCCACCTC CCAGGCTCTGTGCAAAGAGGGTGGCTGTGTA AGGAAGACCTGGTGGCTCTGGCTGTGTT AAAGATGGGGTGGTTGGTGGATTCTCTT GGGAGGGAGGAGAGAAAAGGGCCCTGGGG CAGGGCTCTGGGCTCTGGCTCTGGCTCTGGG</pre>
Mega	<pre>#mega !Title: ProjectedMultiAlign !Format datatype=dna identical=.</pre> <pre>#homo_sap CCTCAGGACC GACGGAAAC #pan_trog</pre> <pre>#homo_sap CCCAGTGCCT TCGACTGCCT #pan_trog</pre> <pre>#homo_sap TGGGACAGAG AGAGAACAC</pre>
MSF	<pre>ProjectedMultiAlign MSF: 2 Type: Name: homo_sapiens 1-46588 Name: pan_troglodytes 1-46588 Ler //</pre> <pre>homosapiens 1-46588 CCTCAGGACC GACGGAAAC &pan_trog</pre> <pre>homosapiens 1-46588 CCTCAGGACC GACGGAAAC &pan_trog</pre> <pre>homosapiens 1-46588 GGTCAACAC C</pre>
Newick	<pre>(((((((BNGTWP00000015030_Trib :0.07 ENSGTP00000002435_Trisig :0.10349) :0.0 ENSGCP00000015199_Gacu :0.161942) :0. (ENSGFP000000010157_Pfor :0.042925, ENSGCP000000010158_Pfor :0.042925, ENSGOP000000084694_Onl1 :0.29811) :0.0 ENSGORLP00000004773_Olat :0.55011) :0.0 ENSGMOP00000010385_Gmor :0.36066) :0.1 (ENSGTWP00000002435_Trisig :0.10349) :0.0 ENSGDMP00000089674_Dzer :0.588918) :0.0 ENSLCOP000000059962_Locu :0.219888) :0. ((((((ENSGGLP00000027524_Gopal :0.0275 ENSGNAP00000015990_Mgal :0.045769) : ENSAIFP000000020741_Apm1 :0.122877) :0.</pre>
Nexus	<pre>##NEXUS [TITLE: ProjectedMultiAlign]</pre> <pre>begin data; dimen:taxa nchar=46588; format interleave datatype=dna gap=.</pre> <pre>matrix homo_sapiens CCTCAGGACC pan_troglodytes CCCAGGACC</pre> <pre>homosapiens 1-46588 GGTCACACAC &pan_troglo</pre>
NHX	<pre>(((((((0.046083 64NHIX:D=N:T=48 0.065551 64NHIX:D=N:T=8083) :Poec :0.359035 64NHIX:D=N:T=8128)) :oval ((0.077336 64NHIX:D=N:T=31033), 0.099898 64NHIX:D=N:T=99883)) :Tet :0.160116 64NHIX:D=N:T=69293)) :Per :0.078027 64NHIX:D=N:T=8090)) :Acan :0.44137 64NHIX:D=N:T=7994 :0.582768 64NHIX:D=N:T=7955)) :stop :0.225188 64NHIX:D=N:T=7918)) :Neop ((((((((0.031221 64NHIX:D=N:T=90</pre>
OrthoXML	<pre><?xml version="1.0" encoding="UTF-8"?> <orthoXML xsi:schemaLocation="http://www.w3.org/2005/11/orthoxml/ns/orthoxml.xsd" xsi:type="orthoxml"></pre> <pre><species NCBITaxId="925 <database name="Unkno <genes> <gene id="6053741 <gene id="5945247 </genes></pre>
Pfam	<pre>homosapiens 1-46588 CCTCAGGACC &pan_troglo :0.045769)</pre>
Phylip	<pre>2 46588 homosapien CCTCAGGACC GACGGAAAC pan_troglodyt CCCAGGACC GACGGAAAC CCCCAGGA GGGCAACAC CCACTGGC GGGCAACAC CCACTGGC ACTGTGTCGGC TTACGCTTA ACTGTGTCGGC TTACGCTTA GCTCAGGCA GCTCTGGAT GCTCAGGCA GCTCTGGAT</pre>
PhyloXML	<pre><?xml version="1.0" encoding="UTF-8"?> <phyloxml xsi:schemaLocation="http://www.phylogeny.roots="true" type="gene" /></pre>
PSI	<pre>homosapiens CCTCAGGACC GACGGAAAC &pan_troglo :0.045769)</pre>
Stockholm	<pre># STOCKHOLM 1.0</pre>
Text	<pre>(B=0 T=Euteleostomi 10335 ---(B=67 T=Neopterygii 1; ---(B=2 T=Clupeocean)</pre>

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

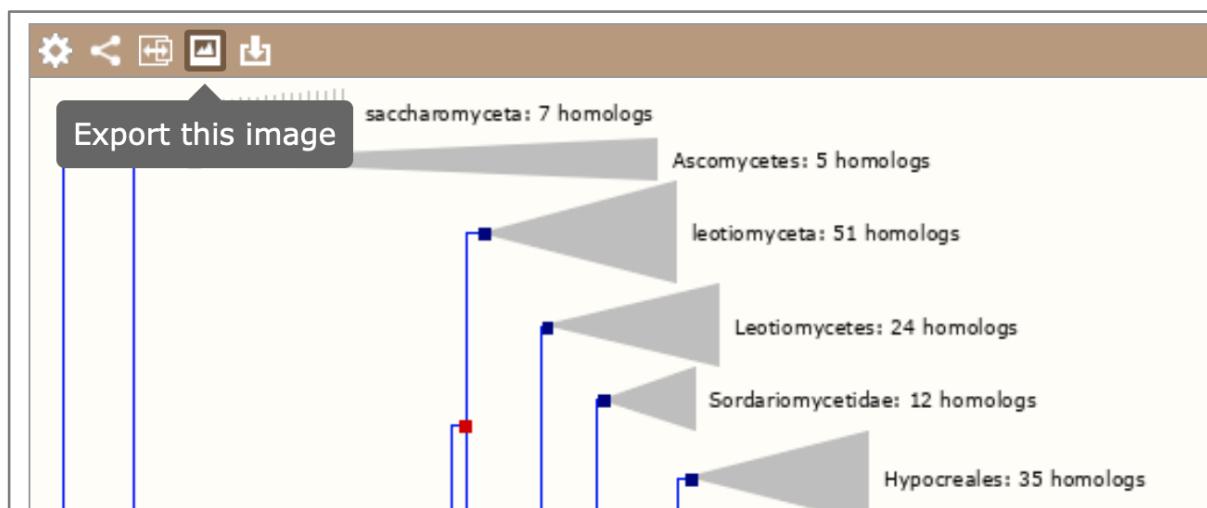


Image download

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

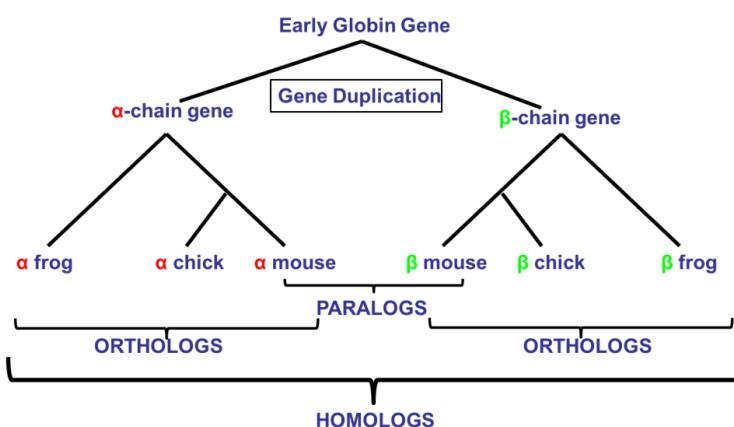
[Download](#)

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.

Homology



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?

CGB_L0350W
«

expand all | collapse all

Search section names...

▶	1 Gene models	<input checked="" type="checkbox"/>
▶	2 Annotation, curation and identifiers	<input checked="" type="checkbox"/>
▶	3 Link outs	<input checked="" type="checkbox"/>
▶	4 Genomic Location	<input checked="" type="checkbox"/>
▶	5 Literature	<input checked="" type="checkbox"/>
▶	6 Taxonomy	<input checked="" type="checkbox"/>
●	7 Orthology and synteny	<input checked="" type="checkbox"/>
▶	8 Phenotype	<input checked="" type="checkbox"/>
▶	9 Transcriptomics	<input checked="" type="checkbox"/>
▶	10 Sequence analysis	<input checked="" type="checkbox"/>
▶	11 Sequences	<input checked="" type="checkbox"/>
▶	12 Structure analysis	<input checked="" type="checkbox"/>
▶	13 Protein features and properties	<input checked="" type="checkbox"/>
▶	14 Function prediction	<input checked="" type="checkbox"/>
▶	15 Pathways and interactions	<input checked="" type="checkbox"/>
▶	16 Immunology	<input checked="" type="checkbox"/>

7 Orthology and synteny

Ortholog Group OG6_106189

Orthologs and Paralogs within FungiDB Data sets

To run Clustal Omega, select genes from the table below. Then choose the sequence type and initiate the alignment with the 'Run Clustal Omega' button.

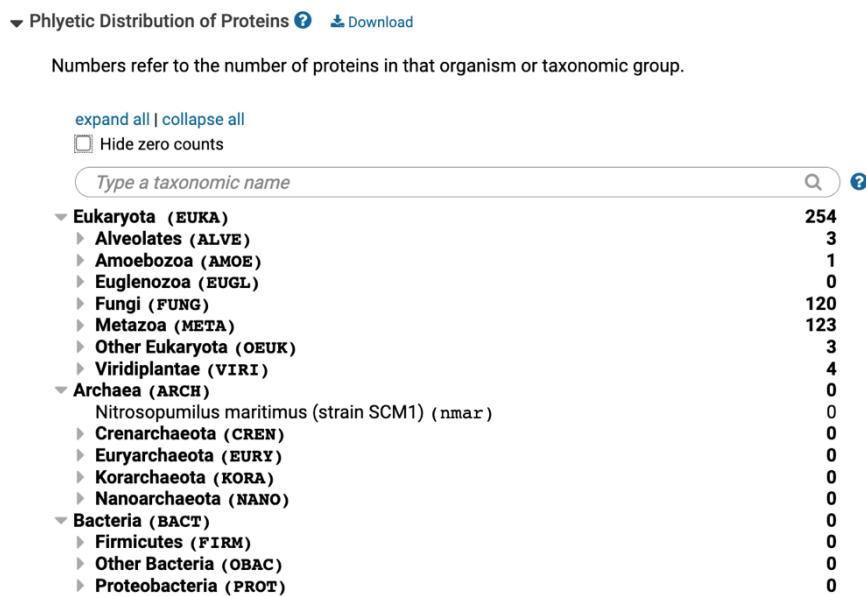
Crypto X ?

Clustal Omega	Gene	Product	Organism
<input type="checkbox"/>	D1P53_002977	unspecified product	Cryptococcus cf. gattii MF34
<input type="checkbox"/>	L203_04836	Cation efflux protein [Source:UniProtKB/TrEMBL;Acc:A0A1E3ICE3]	Cryptococcus depauperatus CBS 784
<input type="checkbox"/>	I314_06191	cation antiporter	Cryptococcus gattii CA1873
<input type="checkbox"/>	I306_06271	cation antiporter	Cryptococcus gattii EJB2
<input type="checkbox"/>	I311_05609	cation antiporter	Cryptococcus gattii NT-10

- 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 orthologs in Archaea 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?

Hide zero counts

Cryptococcus

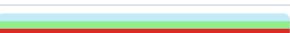
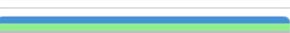
Eukaryota (EUKA)	254
Fungi (FUNG)	120
Basidiomycota (BASI)	27
Cryptococcus cf. gattii MF34 (ccfg)	1
Cryptococcus depauperatus CBS 7841 (cdep)	1
Cryptococcus gattii CA1873 (cgac)	1
Cryptococcus gattii EJB2 (cgae)	1
Cryptococcus gattii NT-10 (cgan)	1
Cryptococcus gattii VGII R265 (cdeu)	1
Cryptococcus gattii VGIV IND107 (cgai)	1
Cryptococcus gattii WM276 (cgat)	1
Cryptococcus neoformans var. grubii H99 (cneq)	1
Cryptococcus neoformans var. grubii KN99 (cnek)	1
Cryptococcus neoformans var. neoformans B-3501A (cnep)	1
Cryptococcus neoformans var. neoformans JEC21 (cneo)	1
Cryptococcus neoformans var. neoformans JEC21 (old build 2016-06-16) (cneo-old)	1

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

4 PFam domains

▼ PFam Legend [Download](#)

Search this table... ?

Accession	Symbol	Description	Count	Legend
PF01545	Cation_efflux	Cation efflux family	251	
PF03645	Tctex-1	Tctex-1 family	2	
PF03102	NeuB	NeuB family	1	
PF01423	LSM	LSM domain	1	

- How can you look up protein alignments for *Cryptococcus*?
 Hint: run ClustalOmega 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.

The screenshot shows the OrthoMCL DB homepage. On the left, there is a sidebar titled "Search for..." with sections for "Ortholog Groups" and "Proteins". A red arrow points to the "Phyletic Pattern" option under "Ortholog Groups". The main content area is titled "Overview of Resources and Tools" and includes a navigation bar with links like "OrthoMCL FAQ", "About OrthoMCL", "Types of Searches in OrthoMCL", "Understanding Group Search Results", "Search Strategies", "Phyletic Pattern Search" (which is highlighted), "Transforming Results", "Assign Proteins to Groups", and "Downloads". Below the navigation bar, there is a "Configure Search" tab and a "Learn More" link. A text input field contains the expression "EUKA>=5T AND hsap>=10". A "Get Answer" button is located to the right of the expression. At the bottom of the page, there is a taxonomic tree with nodes labeled with scientific names like "Root (ALL)", "Eukaryota (EUKA)", "Archaea (ARCA)", and "Bacteria (BACT)".

There are two ways to specify a phyletic pattern:

1. Using the expression box.

- Run the default search for EUKA>=5T AND hsap>=10.

In the graphical tree display:

- Click on the icons to show or hide subtaxa and species.
- Click on the icons to specify which taxa or species to include or exclude in the profile.
- Refer to the legend below to understand other icons.

Expression: EUKA>=5T AND hsap>=10

Get Answer

Key: = no constraints | = must be in group | = at least one subtaxon must be in group | = must not be in group | = mixture of constraints

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

[Configure Search](#)[Learn More](#)

Description

Find Ortholog Groups by phyletic pattern.

Phyletic Pattern Expression is a flexible and powerful way to identify ortholog groups with a certain conservation pattern. Proteins from specific taxa are present or absent. Also, the pattern finds groups with a certain copy number (e.g., both human and E. coli are present).

Examples

These expressions find ortholog groups in which...

hsap>=5 there are five or more human sequences

hsap+ecol=2T both human and E. coli are present.

hsap+ecol=1T only one species of human or E. coli is present.

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.

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

Type a taxonomic name



- * Root (ALL)
- * Eukaryota (EUKA)
 - ▶ **Alveolates (ALVE)**
 - ▶ **Amoebozoa (AMOE)**
 - ▶ **Euglenozoa (EUGL)**
 - ▶ **Fungi (FUNG)**
 - ▶ **Metazoa (META)**
 - ▶ **Other Eukaryota (OEUK)**
 - ▶ **Viriplantae (VIRI)**
- Archaea (ARCH)
 - Nitrosopumilus maritimus (strain SCM1) (nmarr)
 - Crenarchaeota (CREN)
 - Euryarchaeota (EURY)
 - Korarchaeota (KORA)
 - Nanoarchaeota (NANO)
- Bacteria (BACT)
 - Firmicutes (FIRM)
 - Other Bacteria (OBAC)
 - Proteobacteria (PROT)

- 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.

Phyletic
882,708 Ortholog Groups

+ Add a step

Step 1

Strategy URL: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/eebc49abcf1d99f>

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

Phyletic
120,871 Ortholog Groups

+ Add a step

Step 1

- 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.

	Archaea	Bacteria	Alveolata	Amoeba	Euglenozoa	Fungi	Metazoa	Viridiplantae
0 / 27 (0%)	0 / 47 (0%)	0 / 137 (0%)	0 / 14 (0%)	0 / 73 (0%)	7 / 309 (2%)	0 / 124 (0%)	0 / 14 (0%)	
0 / 27 (0%)	0 / 47 (0%)	0 / 137 (0%)	0 / 14 (0%)	0 / 73 (0%)	1 / 309 (0%)	0 / 124 (0%)	0 / 14 (0%)	
0 / 27 (0%)	0 / 47 (0%)	109 / 137 (80%)	4 / 14 (29%)	27 / 73 (37%)	59 / 309 (19%)	0 / 124 (0%)	1 / 14 (7%)	
0 / ALVEOLATA Ciliates: 0 / 2 Apicomplexa: Haemosporida: 60 / 60 Coccidia: 48 / 51 Piroplasmida: 17 / 17 Other apicomplexa: 4 / 4 Other alveolata: 3 / 3	132 / 137 (96%)	14 / 14 (100%)	72 / 73 (99%)	1 / 309 (0%)	0 / 124 (0%)	1 / 14 (7%)		
0 / 27 (0%)	0 / 47 (0%)	0 / 137 (0%)	0 / 14 (0%)	0 / 73 (0%)	1 / 309 (0%)	0 / 124 (0%)	0 / 14 (0%)	

Strategy URL: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/555af9c529d4927>

- Revise your search to find groups that:
 - do not contain orthologs from Alveolates, Amoebozoa, archaea, bacteria and Ascomycetes.
 - contain at least one ortholog group 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.

Phyletic
1,631 Ortholog Groups

+ Add a step

Step 1

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.

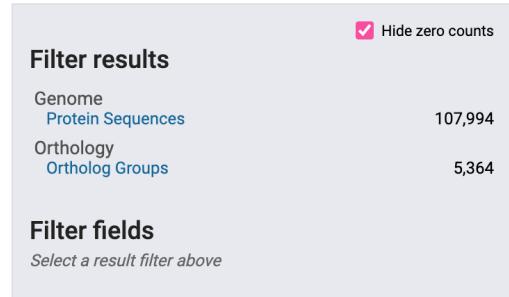


Combining searches in OrthoMCL

- 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 results as a search strategy.

Hide zero counts

Filter results

Genome Protein Sequences	107,994
Orthology Ortholog Groups	5,364

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

Step 1

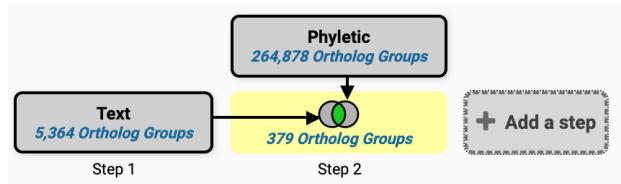
Text
5,364 Ortholog Groups

+ Add a step

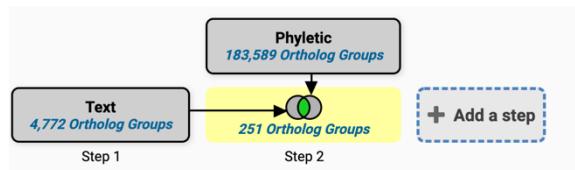
- c. Add a step and run a phyletic pattern search for groups that contain any fungi proteins but do not contain any other organism outside fungi. (hint: make sure everything has a red x on it except for fungi, which should be a grey circle (no constraints)).

- * Root (ALL)
- * Eukaryota (EUKA)
 - Alveolates (ALVE)
 - Amoebozoa (AMOE)
 - Euglenozoa (EUGL)
 - Fungi (FUNG)
 - Metazoa (META)
 - Other Eukaryota (OEUK)
 - Viridiplantae (VIRI)
- Archaea (ARCH)
 - Nitrosopumilus maritimus (strain SCM1) (nmar)
 - Crenarchaeota (CREN)
 - Euryarchaeota (EURY)
 - Korarchaeota (KORA)
 - Nanoarchaeota (NANO)
- Bacteria (BACT)
 - Firmicutes (FIRM)
 - Other Bacteria (OBAC)
 - Proteobacteria (PROT)

How many groups did the search return?

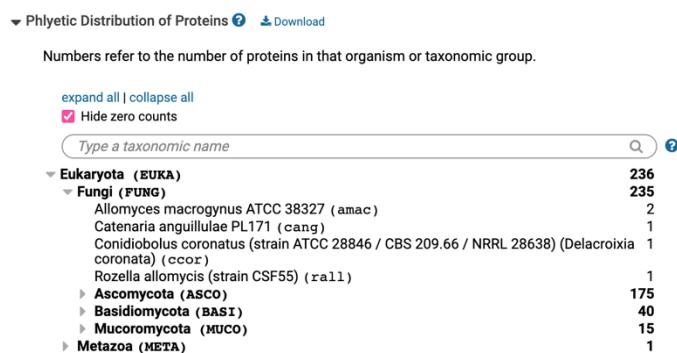


Strategy URL: <https://orthomcl.org/orthomcl/app/workspace/strategies/326574223/430551843>



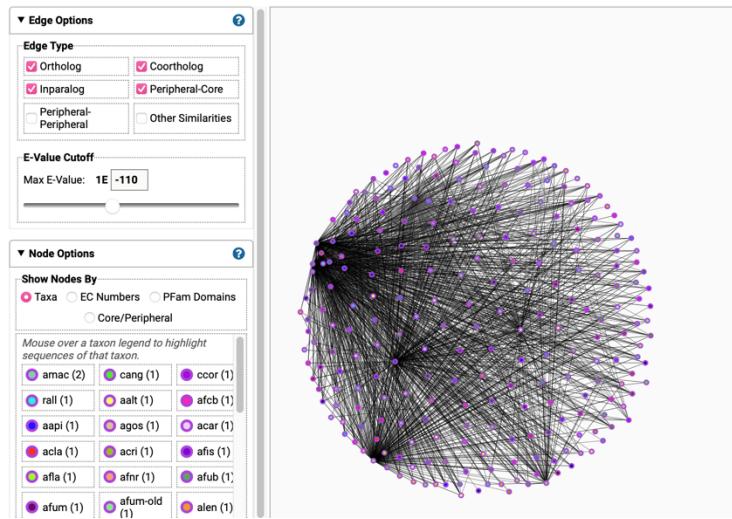
Exploring a specific OrthoMCL group - examining the cluster graph.

- 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)

Cluster Graph: OG6_115064 (245 proteins) [?](#)



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.

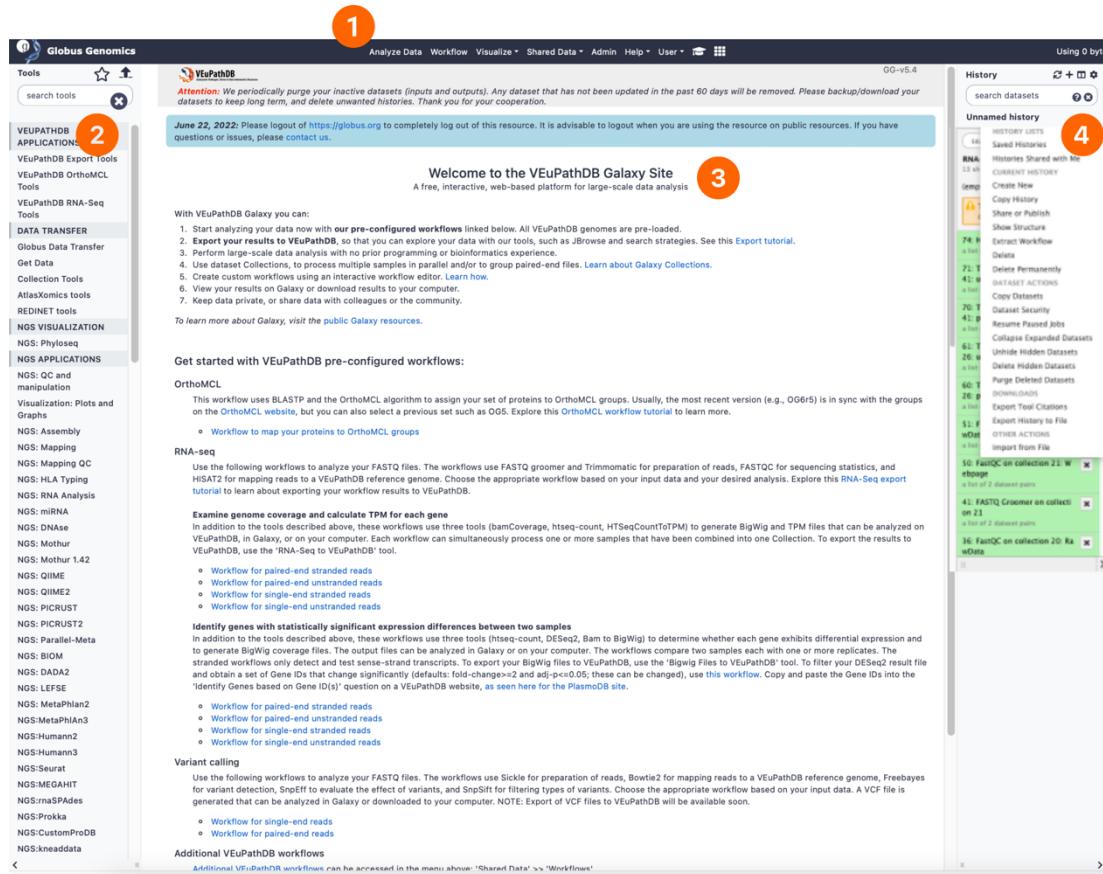
1. Click on the following URL to begin: <https://veupathdb1.globusgenomics.org/>
2. On the next page, you will be asked to define your organization. Choose the “VEuPathDB” option and click on the “Continue” button.
3. If you are not already logged into VEuPathDB, you will be prompted to do so.
4. Click 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 import 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 provides access to pre-loaded raw data, publicly shared workflows, or workflow results (histories), etc.

The screenshot shows the Globus Genomics interface. At the top, there's a navigation bar with links for Analyze Data, Workflow, Visualize, Shared Data, and Admin. Below the navigation bar, there's a sidebar with sections for Tools, VEUPATHDB APPLICATIONS, and DATA TRANSFER. The DATA TRANSFER section is highlighted with a red circle labeled '1'. To the right of the sidebar, there's a main content area for the VEuPathDB application. The VEuPathDB logo and name are at the top. Below that is a message about purging inactive histories. A blue box contains a message about logging out on June 22, 2022. The main content area has a heading "Welcome to the VEuPathDB" and a sub-heading "A free, interactive, web-based platform for larg...". On the far right, there's a sidebar with links for Data Libraries, Histories, Workflows, Visualizations, and Pages. A red circle labeled '2' is on the "Shared Data" link in this sidebar.

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.

The screenshot shows the Published Histories page. At the top, there's a header with "Published Histories" and a search bar containing "FPG2023". Below the header, there's a "Advanced Search" button. The main content area shows a table of published histories. The first row in the table is highlighted with a red circle labeled '3', which corresponds to the history entry for "FPG2023 Group 2 RNA-Seq raw files". The table has columns for Name, Annotation, Owner, Community Rating, and Community Tags. There are also edit and delete icons for each history entry.

Name	Annotation	Owner	Community Rating	Community Tags
FPG2023				
Group 2				
RNA-Seq				
raw files				
FPG2023		ebasenko.108464520	★★★★★	
Group 1				
RNA-Seq				
raw files				

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 parapsilosis*. 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

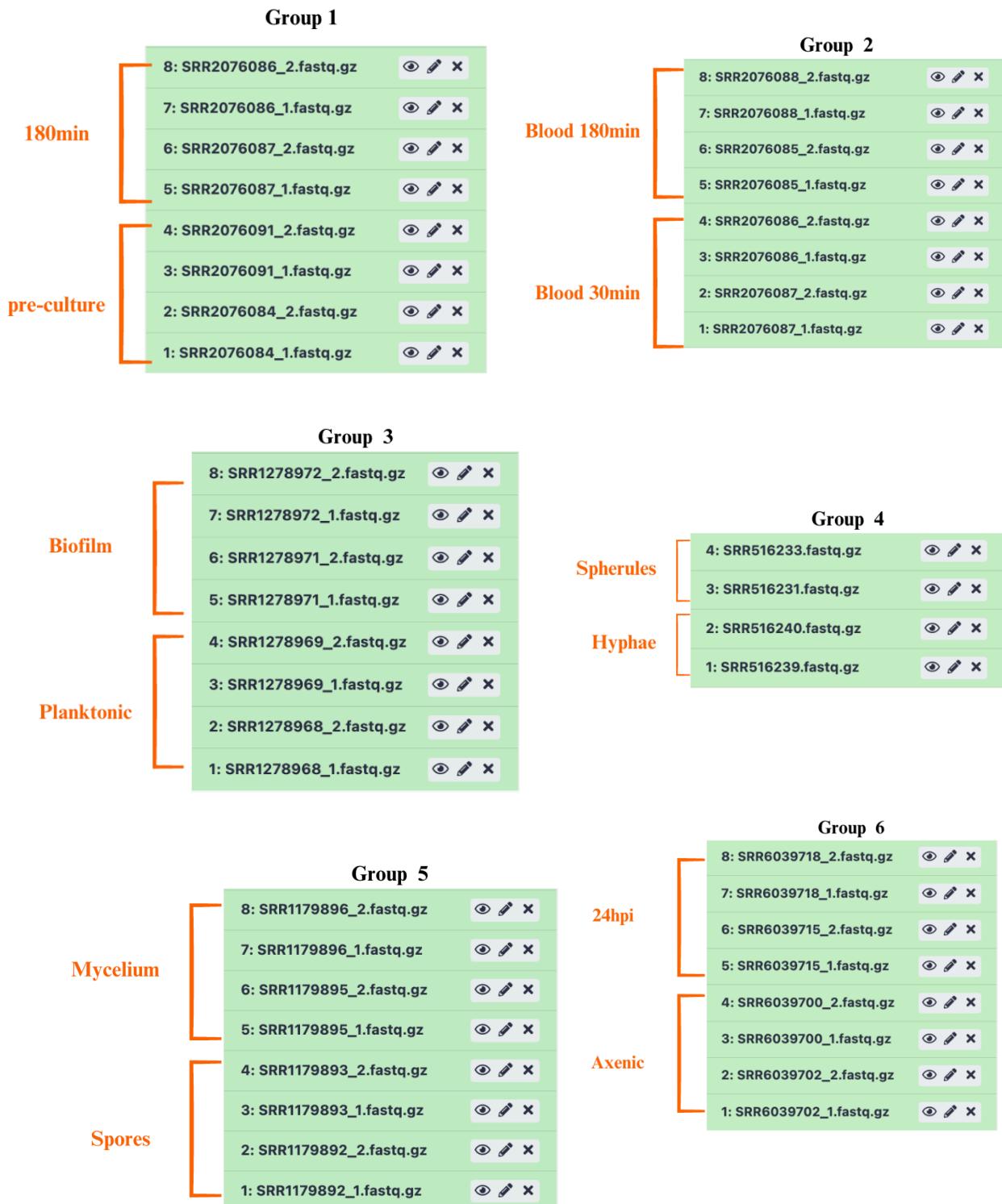
Reference: PMID: 24625133 BioProject: PRJNA239711

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:**

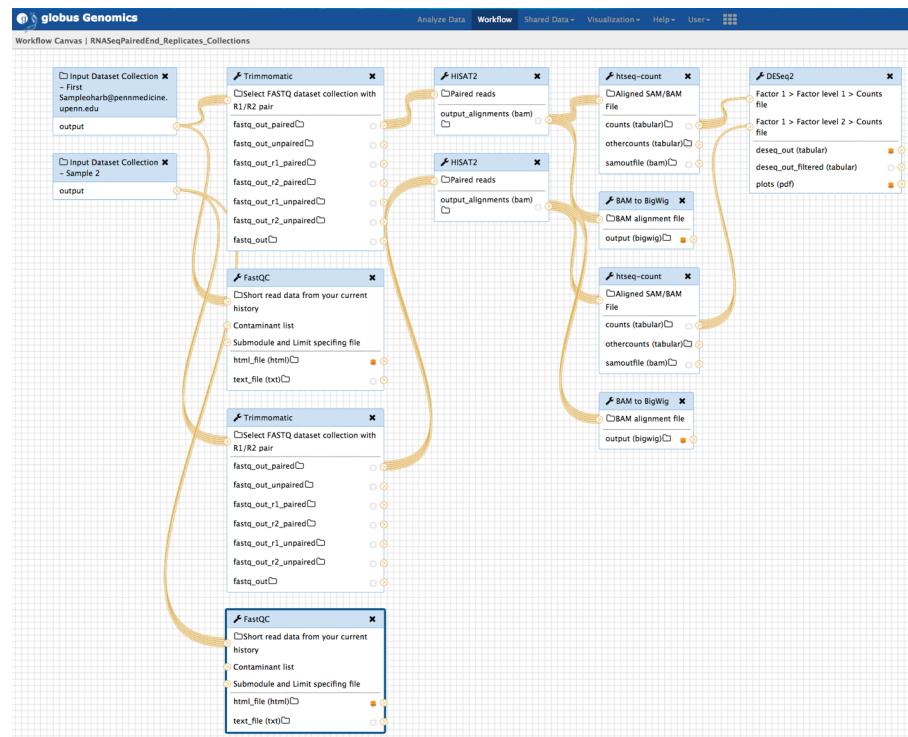
1. Click on the checkbox function “operation on multiple datasets”.
 2. Select samples that belong to the same condition (control samples will appear at the bottom, see file mapping notes for each group below).
 3. Click on “For all selected” and choose “Build List of Dataset Pairs”.
- Note: for single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

RNA-seq

Use the following workflows to analyze your FASTQ files. The workflows use FASTQ groomer and Trimmomatic for preparation of reads, FASTQC for sequencing statistics, and HISAT2 for mapping reads to a VEuPathDB reference genome. Choose the appropriate workflow based on your input data and your desired analysis. Explore this [RNA-Seq export tutorial](#) to learn about exporting your workflow results to VEuPathDB.

Examine genome coverage and calculate TPM for each gene

In addition to the tools described above, these workflows use three tools (bamCoverage, htseq-count, HTSeqCountToTPM) to generate BigWig and TPM files that can be analyzed on VEuPathDB, in Galaxy, or on your computer. Each workflow can simultaneously process one or more samples that have been combined into one Collection. To export the results to VEuPathDB, use the 'RNA-Seq to VEuPathDB' tool.

- Workflow for paired-end stranded reads
- Workflow for paired-end unstranded reads
- Workflow for single-end stranded reads
- Workflow for single-end unstranded reads

Identify genes with statistically significant expression differences between two samples

In addition to the tools described above, these workflows use three tools (htseq-count, DESeq2, Bam to BigWig) to determine whether each gene exhibits differential expression and to generate BigWig coverage files. The output files can be analyzed in Galaxy or on your computer. The workflows compare two samples each with one or more replicates. The stranded workflows only detect and test sense-strand transcripts. To export your BigWig files to VEuPathDB, use the 'Bigwig to VEuPathDB' tool. To filter your DESeq2 result file and obtain a set of Gene IDs that change significantly (defaults: fold-change>=2 and adj-p<=0.05; these can be changed), use [this workflow](#). Copy and paste the Gene IDs into the 'Identify Genes based on Gene ID(s)' question on a VEuPathDB website, as seen here for the PlasmodDB site.

- Workflow for paired-end stranded reads
- Workflow for paired-end unstranded reads
- Workflow for single-end stranded reads
- 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).

History Options

Send results to a new history

Yes No

13: Input Dataset Collection - Sample 1
13: spores

13: Input Dataset Collection - Sample 2
18: mycelium

3: FASTQ Groomer (Galaxy Version 1.0.4)

4: FastQC (Galaxy Version FASTQC: 0.11.3)

5: FASTQ Groomer (Galaxy Version 1.0.4)

6: FastQC (Galaxy Version FASTQC: 0.11.3)

7: Trimmomatic (Galaxy Version 0.36.5)

8: Trimmomatic (Galaxy Version 0.36.5)

9: HISAT2 (Galaxy Version 2.0.5)

10: HISAT2 (Galaxy Version 2.0.5)

11: BAM to BigWig (Galaxy Version 0.2.0)

12: htseq-count - You can use exon or CDS as feature type. You must use gene_id as ID Attribute. (Galaxy Version HTSEQ: default; SAMTOOLS: 1.2; PICARD: 1.134)

13: htseq-count - You can use exon or CDS as feature type. You must use gene_id as ID Attribute. (Galaxy Version HTSEQ: default; SAMTOOLS: 1.2; PICARD: 1.134)

14: BAM to BigWig (Galaxy Version 0.2.0)

15: DESeq2_2.11.40.6 (Galaxy Version 2.11.40.6)

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:

9: HISAT2 (Galaxy Version 2.0.5)

10: HISAT2 (Galaxy Version 2.0.5)

Input data format

FASTQ

Single end or paired reads?

Collection of paired reads

Paired reads

Paired-end options

Specify paired-end parameters

Disable alignments of individual mates

false

Disable discordant alignments

false

Skip reference strand of reference

false

Source for the reference genome to align against

Use a built-in genome

Select a reference genome

FungiDB-31_FgraminearumPH-1_Genome

12: htseq-count - You can use exon or CDS as feature type.

13: htseq-count - You can use exon or CDS as feature type.

Aligned SAM/BAM File

Is this library mate-paired?

paired-end

Will you select an annotation file from your history or use a built-in annotation

Use a built-in annotation

Select a genome annotation

FungiDB-31_FgraminearumPH-1_Genome

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

The screenshot shows the configuration of a DESeq2 workflow. It includes fields for specifying a factor name (Spores & Mycelium), a factor level (Mycelium), and another factor level (Spores). Orange arrows point to each of these three entries.

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

Workflow: imported: DESeq2 Workflow for paired-end unstranded reads (v.7)

Run Workflow

History Options

Send results to a new history



Successfully invoked workflow imported: DESeq2 Workflow for paired-end unstranded reads (v.7).

You can check the status of queued jobs and view the resulting data by refreshing the History pane, if this has not already happened automatically.

Invocation 1...

15 of 15 steps successfully scheduled.

0 of 33 jobs complete.

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)

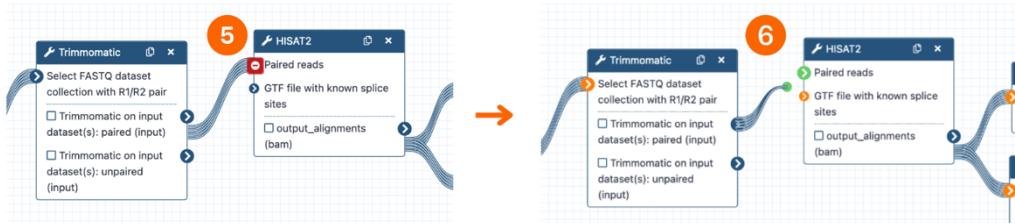
You can create your own workflows. The tools can all be added and configured in an interactive workflow editor.

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 imported into your workspace, left-click and select “edit.”

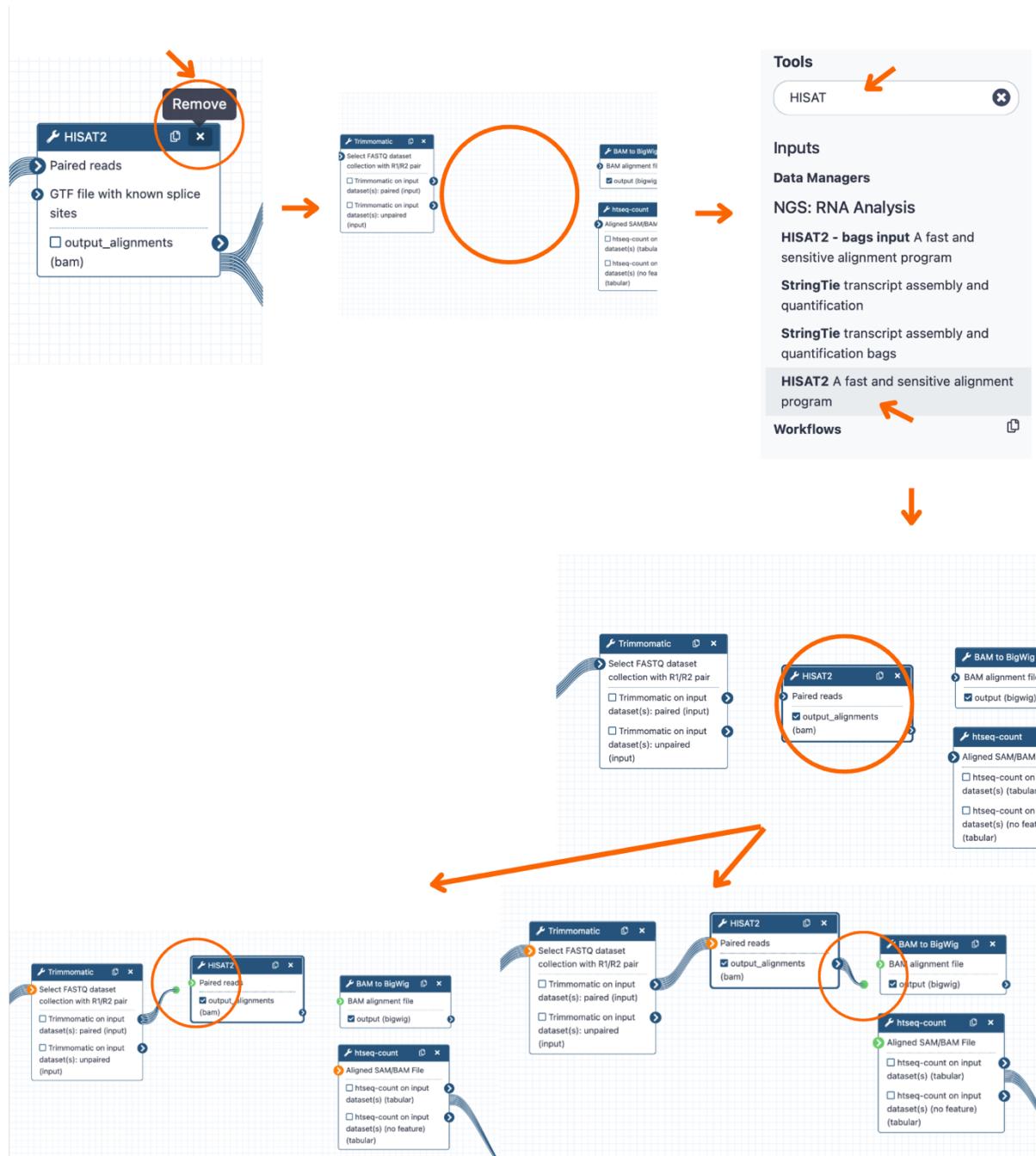


Once you are in the workflow editor:

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 your 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 exiting 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”

1 <https://veupathdb1.globusgenomics.org/>

2

Log in to use veupathdb1

Use your existing organizational login
e.g., university, national lab, facility, project

VEuPathDB

By selecting Continue, you agree to Globus [terms of service](#) and [privacy policy](#).

Continue

OR

[Sign in with Google](#) [Sign in with ORCID iD](#)

Didn't find your organization? Then use [Globus ID](#) to sign in. (What's this?)

3

Please log in

Username or Email: _____

Password: _____

[Forgot Password?](#) [Register/Subsribe](#)

Visit our partner Bioinformatics Resource Center, [B2-BRC](#)

4

Welcome – You've Successfully Logged In

This is the first time you are accessing Globus with your **EuPathDB** login.
If you have previously used Globus with another login you can link it to your **EuPathDB** login. When linked, both logins will be able to access the same Globus account permissions and history.

Continue Link to an existing account Why should I link accounts?

5

Complete Your Sign Up For
[REDACTED]@eupathdb.org

Name [REDACTED]

Email [REDACTED]

Organization test account*

Account will be used for

non-profit research or educational purposes
 commercial purposes
 I have read and agree to the Globus [Terms of Service](#) and [Privacy Policy](#).

Continue

* This field is specified by the identity provider, and cannot be modified by Globus. If you change it with your identity provider, it will propagate to Globus the next time you log in.

veupathdb1 would like to:

6

View your identity [\(i\)](#)

Manage data using Globus Transfer [\(i\)](#)

View your email address [\(i\)](#)

View identity details [\(i\)](#)

To work, the above will need to: [▼](#)

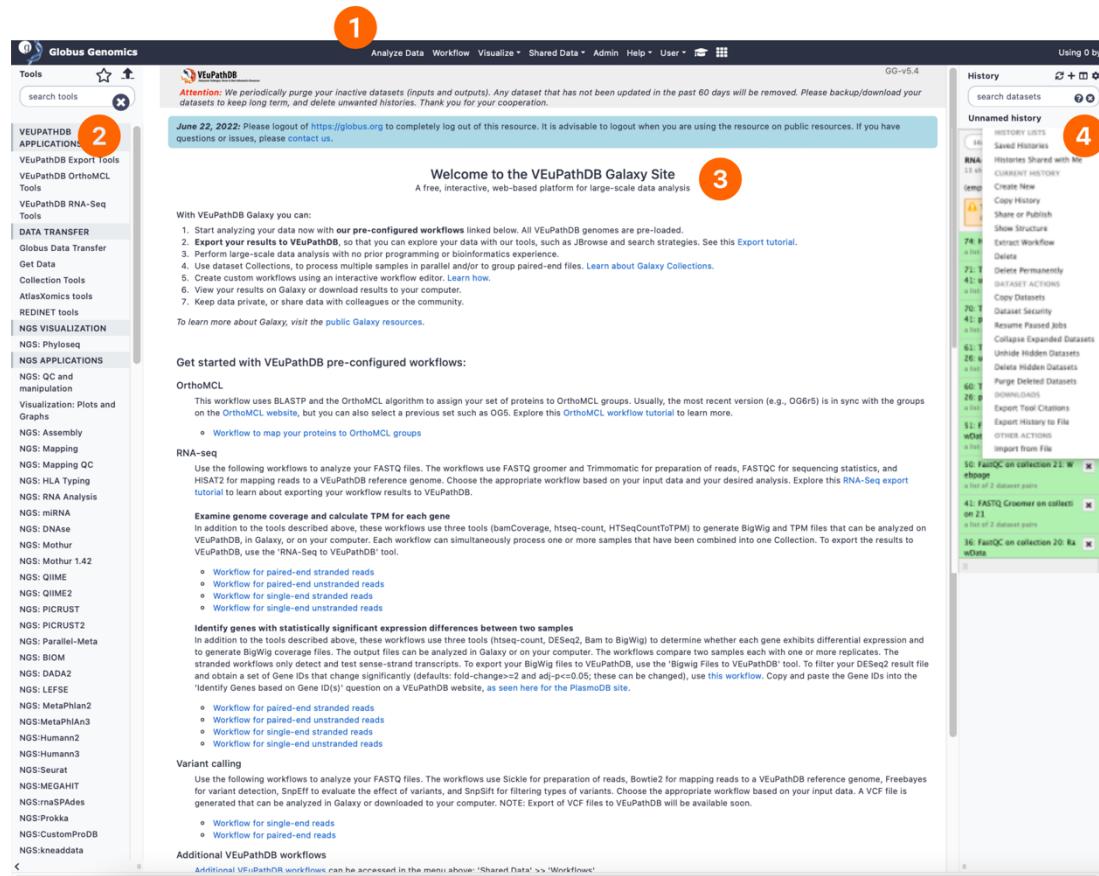
By clicking 'Allow', you allow **veupathdb1** (this client has not provided terms of service or a privacy policy to Globus) to use the above listed information and services. You can rescind this and other consents [\(i\)](#) at any time.

Allow Deny

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
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 import 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 provides access to pre-loaded raw data, publicly shared workflows, or workflow results (histories), etc.

The screenshot shows the VEuPathDB homepage. At the top, there is a navigation bar with links for Analyze Data, Workflow, Visualize, Shared Data, and Admin. Below the navigation bar, there is a search bar labeled "search tools". On the left side, there is a sidebar with sections for Tools, VEuPathDB APPLICATIONS, and DATA TRANSFER. The "DATA TRANSFER" section is highlighted with a red circle containing the number 1. Under "DATA TRANSFER", there are links for Globus Data Transfer and Get Data. To the right of the sidebar, there is a message box with the following text:
Attention: We periodically purge your inactive histories. Thank you for your cooperation.
June 22, 2022: Please logout of the system. It is advisable to logout when you are using the system. If you have any questions or issues, please contact us.
Below the message box, there is a welcome message: Welcome to the VEuPathDB. A free, interactive, web-based platform for large-scale bioinformatics analysis.

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 corresponds to your group number (e.g., FPG2023 SNP Group1) to import the data into your Galaxy workspace.

The screenshot shows the "Published Histories" search interface. At the top, there is a search bar with the text "FPG2023" and a magnifying glass icon. Below the search bar, there is a link for "Advanced Search". On the left, there is a sidebar with links for Data Libraries, Histories, Workflows, Visualizations, and Pages. The "Histories" link is highlighted with a red circle containing the number 1. An arrow points from the "Histories" link to the search bar. Below the search bar, there is a table with columns for Name and Annotation. The first row in the table is highlighted with a red circle containing the number 3, corresponding to the "FPG2023 SNP Group1" entry.

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).

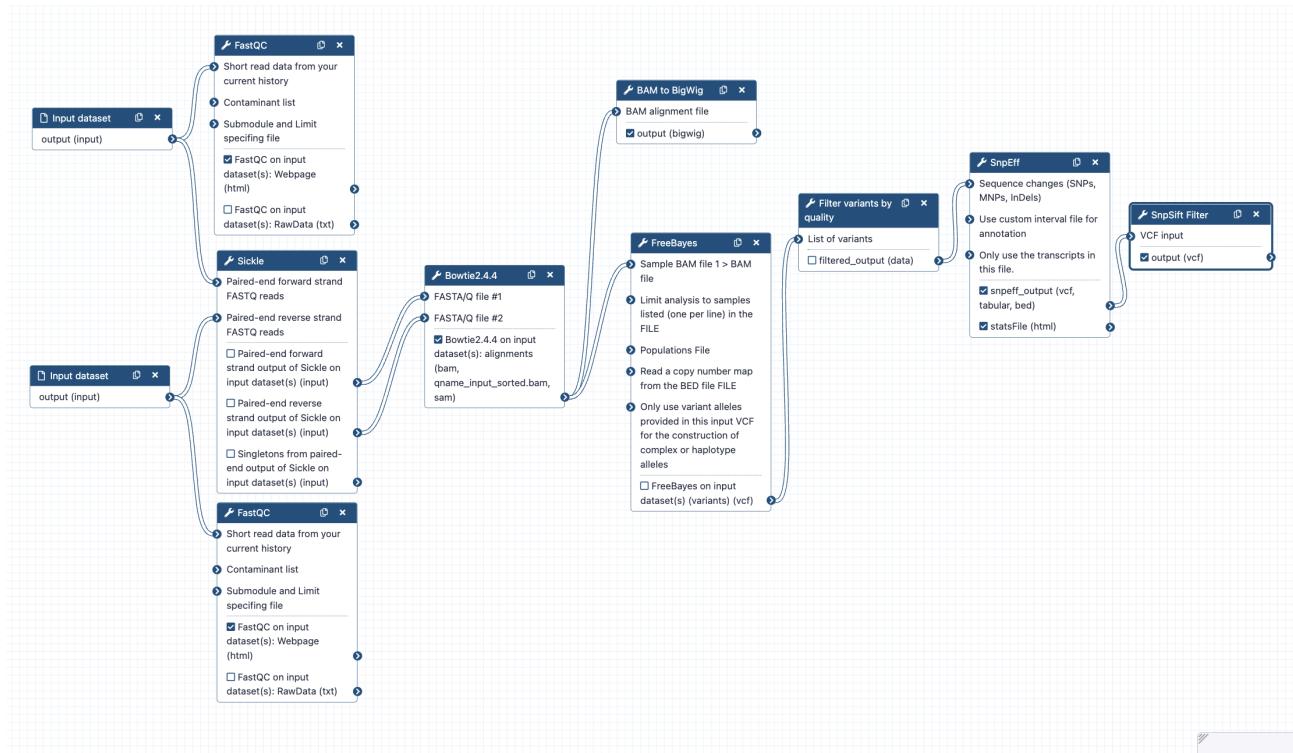
Variant calling

Use the following workflows to analyze your FASTQC detection, SnpEff to evaluate the effect of variants analyzed in Galaxy or downloaded to your computer.

- Workflow for single-end reads
- Workflow for paired-end reads

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.

Workflow: imported: Variant Calling Workflow for paired-end reads (v.7)

5  Run Workflow

History Options
Send results to a new history
Yes No

1: Input dataset - 1
1: SRR11785185_1.fastq.gz 

2: Input dataset - 8
2: SRR11785185_2.fastq.gz 

3: FastQC - 2 (Galaxy Version FASTQC: 0.11.3)

4: Sickle (Galaxy Version 1.33.2)

5: FastQC - 9 (Galaxy Version FASTQC: 0.11.3)

6: Bowtie2_4.4 (Galaxy Version 2.4.4+galaxy0) 

7: BAM to BigWig - 11 (Galaxy Version 0.2.0)

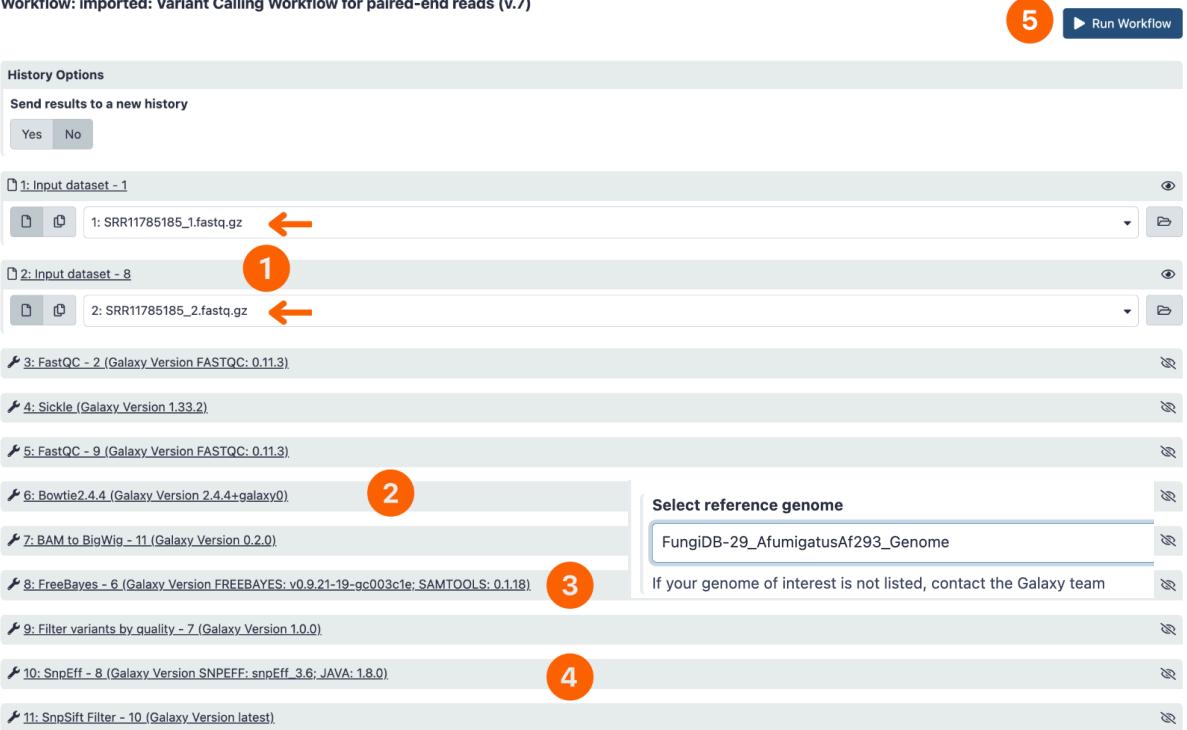
8: FreeBayes - 6 (Galaxy Version FREEBAYES: v0.9.21-19-gc003c1e; SAMTOOLS: 0.1.18) 

9: Filter variants by quality - 7 (Galaxy Version 1.0.0)

10: SnpEff - 8 (Galaxy Version SNPEFF:.snpEff_3.6; JAVA: 1.8.0) 

11: Snpsift Filter - 10 (Galaxy Version latest)

Select reference genome
FungiDB-29_AfumigatusAf293_Genome
If your genome of interest is not listed, contact the Galaxy team

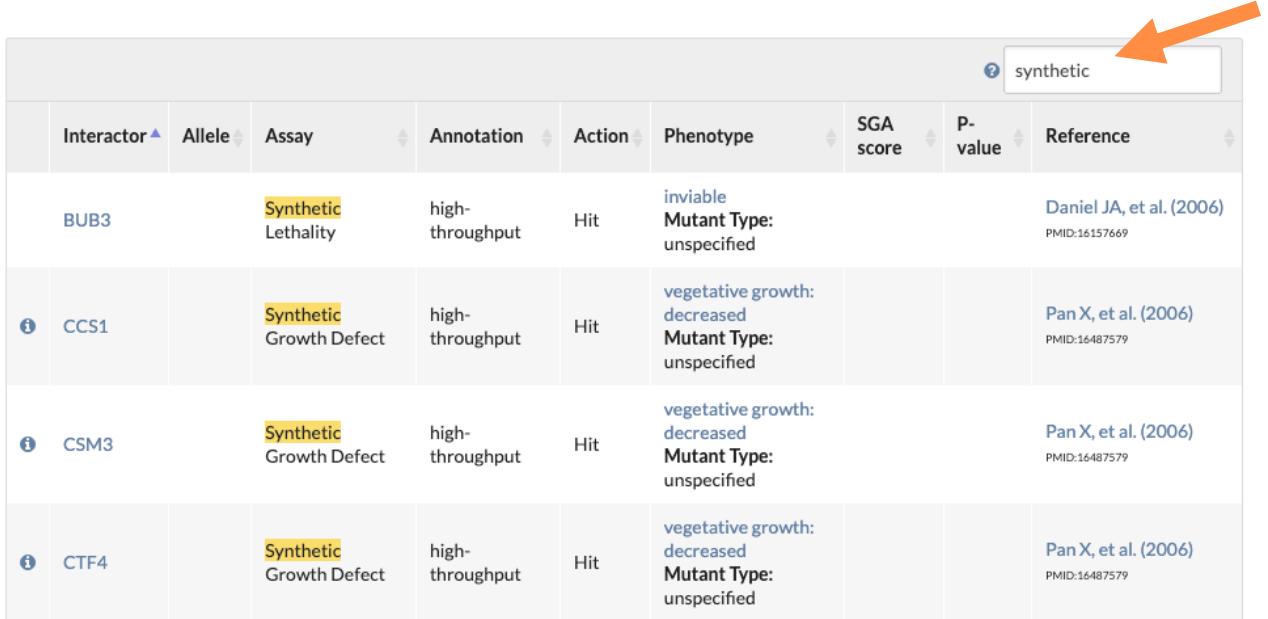


Using SGD GO Slim Mapper and Interaction Data to Predict Gene Function

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.



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

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

Tools

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

Genes

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

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

Query Set (Your Input)

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

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

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

Specify your Slim Terms

Choose a GO Set:

 Yeast GO-Slim: process

Refine your list of GO Slim Terms:

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

SELECT ALL Terms from Yeast GO-Slim: process

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

 Submit Form  Reset Form

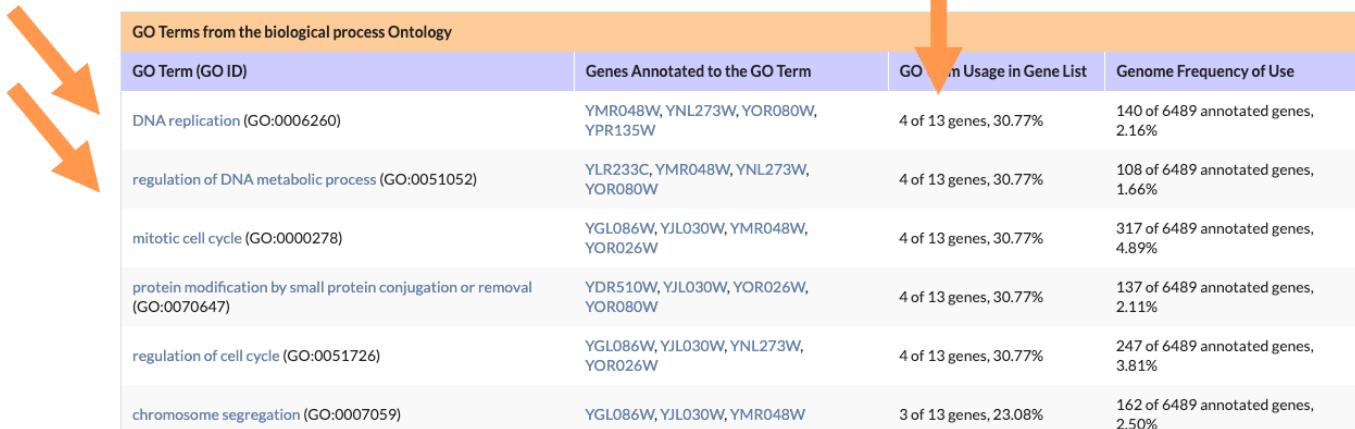
- Choose a **GO Set** by selecting **Yeast GO-Slim: Process** from the pull-down.
- Highlight **SELECT ALL Terms from Yeast GO-Slim: Process**.
- Click the **Submit Form** button to use the default settings or go further down to customize your query.

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

Search Results

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

GO version 2023-04-01



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

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

GO Enrichment, Phenotype Data at CGD

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

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

CGD Quick Search Result

[Go to Advanced Search Page](#)

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

General Search Results for : pmt6

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

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

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

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

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

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

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

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

Results: 1 - 30 of 42 records
1 2

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

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

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

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

Step 2: Query Set (Your Input)
 Enter Gene/ORF names:
 (separated by a return or a space)
 C3_07710W_A C1_02260C_A C3_01530C_A C1_10380C_A
 C4_06100W_A C1_08010W_A C6_00420W_A C4_01920W_A
 C1_03190C_A C1_02150W_A C2_04240C_A C2_04760W_A
 C3_05610W_A C3_06020W_A C1_03730C_A C1_00620W_A
 C3_06090C_A C7_00320C_A C7_02890C_A C3_06890W_A

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

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

Process
 Function
 Component

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

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

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

Optional Step 4: Specify your background set of genes using the options below.

Use default background set (all features in the database)	OR	Enter Gene/ORF names: (separated by a return or a space)	OR	Upload a file of Gene/ORF names: <input type="button" value="Choose File"/> no file selected
--	----	---	----	---

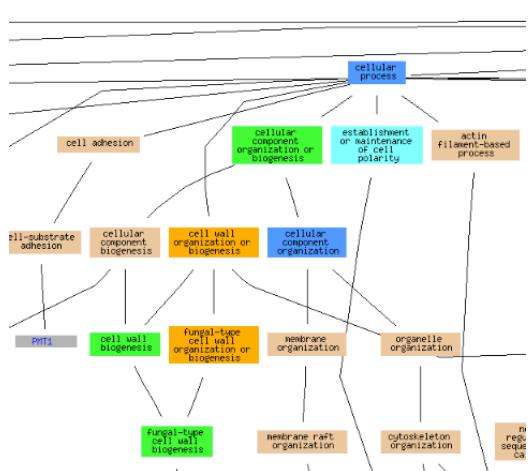
Customize the gene list in the default or your specific background set (OPTIONAL)

Feature type Default includes all feature types listed here	<input checked="" type="checkbox"/> ORF <input checked="" type="checkbox"/> allele <input checked="" type="checkbox"/> ncRNA <input checked="" type="checkbox"/> not in systematic sequence <input checked="" type="checkbox"/> pseudogene <input checked="" type="checkbox"/> rRNA <input checked="" type="checkbox"/> snRNA <input checked="" type="checkbox"/> snoRNA <input checked="" type="checkbox"/> tRNA
<input type="button" value="Search"/> <input type="button" value="Clear All"/>	

Optional Step 5: Refine the Annotations used for calculation
You can use this option with Step 4. All Annotation Types are included by default.

Select by Annotation Method	Manually curated: <input checked="" type="radio"/> yes <input type="radio"/> no High-throughput: <input checked="" type="radio"/> yes <input type="radio"/> no Computational: <input checked="" type="radio"/> yes <input type="radio"/> no
Select by Annotation Source	<input checked="" type="checkbox"/> CGD
Select by Evidence Codes:	<input checked="" type="checkbox"/> IC <input checked="" type="checkbox"/> IDA <input checked="" type="checkbox"/> IEA <input checked="" type="checkbox"/> IEP <input checked="" type="checkbox"/> IGC <input checked="" type="checkbox"/> IGI <input checked="" type="checkbox"/> IMP <input checked="" type="checkbox"/> IPI <input checked="" type="checkbox"/> ISA <input checked="" type="checkbox"/> ISM <input checked="" type="checkbox"/> ISO <input checked="" type="checkbox"/> ISS <input checked="" type="checkbox"/> NAS <input checked="" type="checkbox"/> ND <input checked="" type="checkbox"/> RCA <input checked="" type="checkbox"/> TAS
<input type="button" value="Search"/> <input type="button" value="Clear All"/>	

- Click **Search**. The input is checked and any genes that are not recognized as valid for the selected *Candida* species are rejected; click on **Proceed** in the following window.
- The results page displays the significant shared GO terms (or their parents) in both graphic and table form, within the set of genes associated with hygromycin B sensitivity entered on the previous page:



The graph shows the GO tree that includes terms used directly or indirectly in annotations for the genes in your list. The terms are color-coded to indicate their statistical significance (p-value score). 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

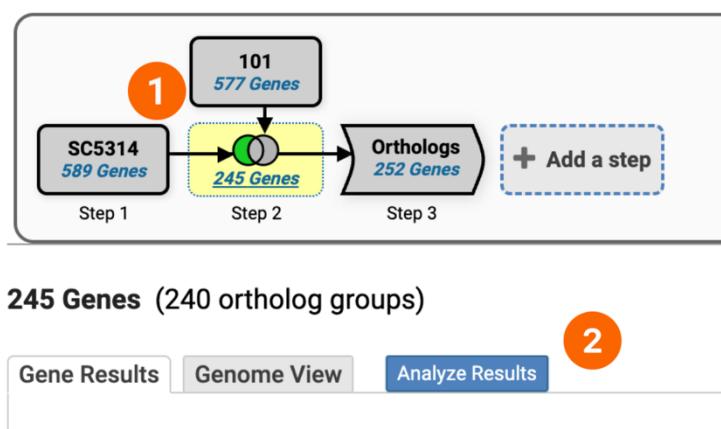
Learning objectives:

- 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>

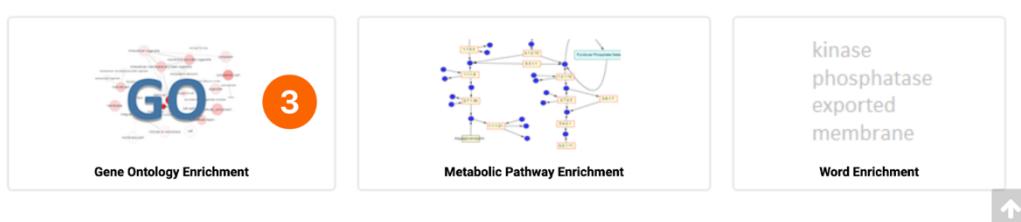
1. Click on the Step 2 to identify upregulated gene in *C. albicans* SC5314 only.
2. Click on the “Analyze Results” tab to bring up enrichment analysis options.



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.

Analyze your Gene results with a tool below.



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.

Organism: Candida albicans SC5314

Ontology:

- Molecular Function
- Biological Process
- Cellular Component

Evidence:

- Computed
- Curated

select all | clear all

Limit to GO Slim terms: No

P-Value cutoff: 0.05 (0 - 1)

Submit

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.

Organism: Candida albicans SC5314

Ontology:

- Molecular Function
- Biological Process
- Cellular Component

Evidence:

- Computed
- Curated

select all | clear all

Limit to GO Slim terms: No

P-Value cutoff: 0.05 (0 - 1)

Submit

Analysis Results:

244 rows

Open in Revigo | Show Word Cloud | Download

GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0042273	ribosomal large subunit biogenesis	558	67	12.0	3.03	4.20	1.08e-17	1.68e-14	1.68e-14
GO:0000470	maturition of LSU-rRNA	440	55	12.5	3.16	4.20	3.31e-15	2.59e-12	5.17e-12
GO:0000463	maturition of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	432	53	12.3	3.10	4.07	2.62e-14	1.37e-11	4.10e-11

The results table includes several additional statistical measurements:

- **Fold enrichment** - The ratio of the proportion of genes in the list of interest with a specific GO term over the proportion of genes in the background with that term.
- **Odds ratio** - Determines 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-Hochberg false discovery rate** - A method for controlling false discovery rates for type 1 errors.
- **Bonferroni adjusted P-values** - A method for correcting significance based on multiple comparisons.

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

Notice that the table contains columns with GO IDs and GO terms along with the number of genes in the background and those specific to the RNA-Seq analysis results presented (linked in blue).

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

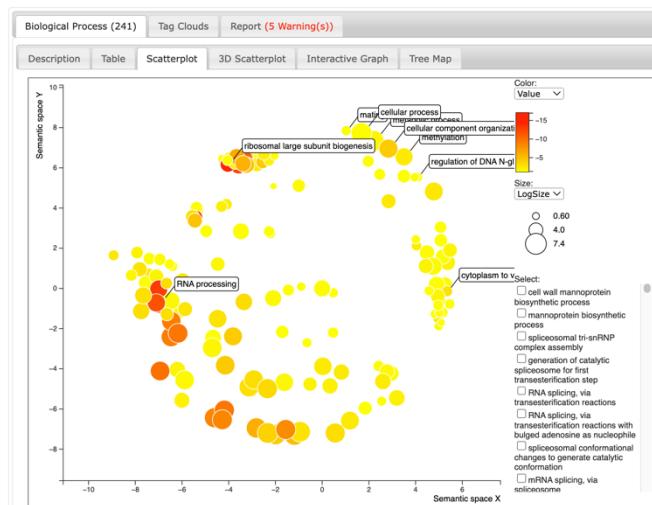
Note: you can sort genes in your results using the sort options within a column:

Genes in your result with this term	Percent of bkgd genes in your result
Activate to sort the table by Genes in your result with this term in ascending order.	
202	7.2
184	4.3
181	4.5

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

The screenshot shows the FungiDB gene record page for jhhlp_004726, a hypothetical protein. At the top, there are links for 'Add to basket' and 'Add to favorites'. Below that is the gene ID 'jhhlp_004726 hypothetical protein'. A horizontal line separates this from the main content area. In the main area, there's a 'Download Gene' link followed by 'jhhlp_004726'. Below this, there are two sections: 'Choose a Report' (radio buttons for 'Text - choose from columns and/or tables' and 'FASTA - sequence retrieval, configurable') and 'Choose Attributes' (checkboxes for various gene models like Gene models, Annotation, curation and identifiers, Genomic Location, Orthology and synteny, etc.). To the right is a 'Choose Tables' section with a similar list of checkboxes. An orange arrow points from the 'Download Gene' link down to the 'Choose Tables' section, specifically highlighting the checkbox for 'Orthologs and Paralogs within VEuPathDB' which is checked.

- 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.

SUMMARY

Sequence Protein Gene Ontology Phenotype Interactions Regulation Expression Literature Homology

MKC7 / YDR144C Interactions

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

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

Analyze

Physical Genetic Intersection All

Genetic Interactions

Genetic Interactions 121 entries for 102 genes

Interactor	Allele	Assay	Annotation	Action	Phenotype	SGA score	P-value	Reference
ACT1		Synthetic Haploinsufficiency	high-throughput	Hit				Haarer B, et al. (2007) PMID:17167106
GIM5		Synthetic Growth Defect	high-throughput	Hit	vegetative growth: decreased Mutant Type: unspecified			Tong AH, et al. (2004) PMID:15474670

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

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

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

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

Step 2. Choose Ontology

Pick an ontology aspect:

Process Function Component

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

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

Result Table

Terms from the Process Ontology of gene_association.sgd with p-value <= 0.01

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

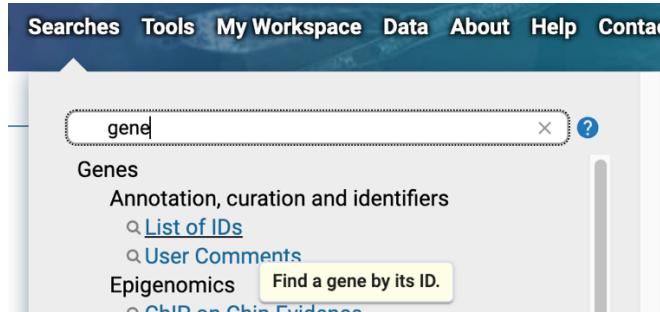
Now, let's go back to the file of MKC7 "synthetic" genetic interactors we downloaded earlier and find the orthologs of these genes in *Lomentospora prolificans*.

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

Interactor	Interactor Systematic Name	Interactor	Interactor Systematic Name	Type	Assay	Annotation
MKC7	YDR144C	ACT1	YFL039C	Genetic	Synthetic	Ha high-through
MKC7	YDR144C	GIMS	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.

Identify Genes based on List of IDs

Configure Search Learn More View Data Sets Used

Gene ID input set

Enter a list of IDs or text:

Upload a text file: No file chosen
Maximum size 10MB. The file should contain the list of IDs.

Upload from a URL:
The URL should resolve to a list of IDs.

Copy from My Basket: 3 records will be copied from your basket.

Copy from My Strategy: NRPS (766 records)

The screenshot shows the BioPax search interface. At the top left, there is a yellow button labeled "IDs List 9 Genes". To its right is a blue button labeled "+ Add a step". Below these buttons is a section labeled "Step 1". At the top right of the interface are several icons for file operations: copy, paste, save, print, etc.

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

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

Rows per page: 1000 [▼](#)

[Download](#)

[Send to...](#) [▼](#)

[Add Columns](#)

	Gene ID	Transcript ID	Gene Name or Symbol	Organism	Genomic Location (Gene)	Product Description
YFL039C	YFL039C-126_1	ACT1	<i>Saccharomyces cerevisiae</i> S288C	BK006940:53,260..54,696(-)	actin	
YML094W	YML094W-126_1	GIM5	<i>Saccharomyces cerevisiae</i> S288C	BK006946:82,275..82,849(+)	Gim5p	
YHR079C	YHR079C-126_1	IRE1	<i>Saccharomyces cerevisiae</i> S288C	BK006934:258,244..261,591(-)	bifunctional endoribonuclease/protein	

- 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**.

The screenshot shows the "Add a step to your search strategy" interface. At the top left, there is a yellow button labeled "Gene ID(s) 9 Genes" and a blue button labeled "+ Add a step". A red arrow points from the "+ Add a step" button to the main content area.

The main content area has three sections:

- Combine with other Genes**: Shows a flowchart from "Gene ID(s)" to "Step 1" and "Step 2".
- Transform into related records**: Shows a flowchart from "Gene ID(s)" to "Step 1" and "Step 2". This section is highlighted with a blue box and a red arrow pointing to it from the "+ Add a step" button.
- Use Genomic Colocation to combine with other features**: Shows a flowchart from "Gene ID(s)" to "Step 1" and "Step 2".

To the right of these sections, there is a box titled "Transform 9 Genes into..." with a blue button labeled "Orthologs".

Below the sections, there is another "Add a step to your search strategy" box. It contains the following steps:

- Your Genes from Step 1 will be converted into Orthologs**
- Organism**: A note says "Note: You must select at least 1 values for this parameter." A dropdown menu shows "Lom" selected, with "Fungi", "Ascomycota", "Sordariomycetes", "Microscales", and "Lomentospora prolificans JHH-5317" listed below it. A red box highlights the dropdown menu.
- Syntenic Orthologs Only?**: A radio button is set to "no".

A red box highlights the "Run Step" button at the bottom right of the second box.

The screenshot shows a search interface for orthologs. Step 1: Gene ID(s) (9 Genes). Step 2: Orthologs (8 Genes). A dashed box labeled '+ Add a step' is present. Below the steps is a search bar: '8 Genes (7 ortholog groups)' and 'Revise this search'. Underneath are tabs: 'Gene Results' (selected), 'Genome View', and 'Analyze Results'. A 'Rows per page' dropdown set to 20. At the top right are icons for download, add to basket, and add columns. The main area displays a table with 8 rows of gene information:

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Input Ortholog(s)	Ortholog Group	Paralog count	Ortholog count
jhhlp_002587	jhhlp_002587-141_1	Lomentospora prolificans JHH-5317	NLAX01000008:3,258,120..3,260,362(-)	hypothetical protein	YFL039C	OG6_100127	0	239
jhhlp_004481	jhhlp_004481-141_1	Lomentospora prolificans JHH-5317	NLAX01000010:4,766,898..4,769,585(+)	hypothetical protein	YNL238W	OG6_100362	0	167
jhhlp_004364	jhhlp_004364-141_1	Lomentospora prolificans JHH-5317	NLAX01000010:4,180,492..4,181,475(-)	hypothetical protein	YKL117W	OG6_101574	0	157
					VLM070P	OG6_100362	0	167
					VLM070P	OG6_100362	0	167
					VLM070P	OG6_100362	0	167
					VLM070P	OG6_100362	0	167
					VLM070P	OG6_100362	0	167

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.

The screenshot shows the Galaxy History Actions menu. On the left, there's a 'History' button with a red circle containing the number '1'. To its right is a search bar with placeholder text 'search datasets' and two small icons. Below the search bar is the history title 'Mycelium vs Spore' and some statistics: '15 shown, 19 deleted, 148 hidden' and '49.74 GB'. To the right of the history title are three small icons. An orange arrow points from the 'History' button to the 'Share or Publish' button in the dropdown menu. The dropdown menu itself has four items: 'Copy', 'Share or Publish' (which is highlighted in blue), and 'Show Structure'.

Make History Accessible via Link and Publish It

This history is currently restricted so that only you and the users listed below can access it. You can:

Also make all objects within the History accessible.

Generates a web link that you can share with other people so that they can view and import the history.

2

Also make all objects within the History accessible.

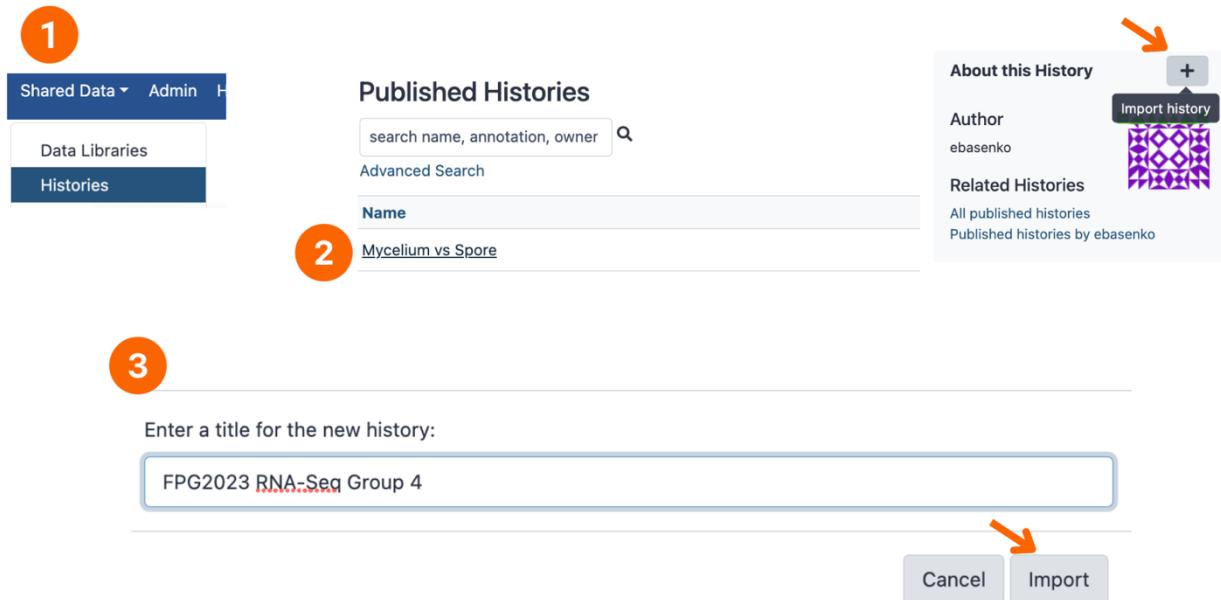
Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, v

Share History with Individual Users

You have not shared this history with any users.

- 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.

Many more output files are available to explore →

Mycelium vs Spore
16 shown, 18 deleted 148 hidden
49.74 GB

Differential expression data on the two collection →

94: DESeq2 plots on data 88, data 86, and others

93: DESeq2 result file on data 88, data 86, and others

90: BAM to BigWig on collection 72
a list with 2 items

75: BAM to BigWig on collection 69
a list with 2 items

39: FastQC on collection 18: Webpage
a list of pairs with 2 items

24: FastQC on collection 13: Webpage
a list of pairs with 2 items

18: mycelium
a list of pairs with 2 items

13: spores
a list of pairs with 2 items

8: SRR1179896_2.fastq.gz

7: SRR1179896_1.fastq.gz

6: SRR1179895_2.fastq.gz

Coverage data in BigWig format →

FastQC results (one per each file submitted) →

- 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 (eye) on either forward or reverse. Note that each FastQ file will have its own FastQC results.

24: FastQC on collection 13: X

Webpage ↗
a list of pairs with 2 items

SRR1179892.fastq a pair of datasets	SRR1179893.fastq a pair of datasets
	forward ↗
	reverse ↗

Summary																	
✓ Basic Statistics ✗ Per base sequence quality ✓ Per tile sequence quality ✓ Per sequence quality scores ! Per base sequence content ! Per sequence GC content ✓ Per base N content ✓ Sequence Length Distribution ✓ Sequence Duplication Levels ✗ Overrepresented sequences ✓ Adapter Content ✗ Kmer Content	✓ Basic Statistics <table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th style="background-color: #000080; color: white;">Measure</th> <th style="background-color: #000080; color: white;">Value</th> </tr> </thead> <tbody> <tr><td>Filename</td><td>SRR11785185_2.fastq.gz</td></tr> <tr><td>File type</td><td>Conventional base calls</td></tr> <tr><td>Encoding</td><td>Sanger / Illumina 1.9</td></tr> <tr><td>Total Sequences</td><td>7649791</td></tr> <tr><td>Sequences flagged as poor quality</td><td>0</td></tr> <tr><td>Sequence length</td><td>251</td></tr> <tr><td>%GC</td><td>50</td></tr> </tbody> </table> ✗ Per base sequence quality <p style="text-align: center;">Quality scores across all bases (Sanger / Illumina 1.9 encoding)</p>	Measure	Value	Filename	SRR11785185_2.fastq.gz	File type	Conventional base calls	Encoding	Sanger / Illumina 1.9	Total Sequences	7649791	Sequences flagged as poor quality	0	Sequence length	251	%GC	50
Measure	Value																
Filename	SRR11785185_2.fastq.gz																
File type	Conventional base calls																
Encoding	Sanger / Illumina 1.9																
Total Sequences	7649791																
Sequences flagged as poor quality	0																
Sequence length	251																
%GC	50																

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.

The screenshot shows the Galaxy interface with a green header bar. The header bar contains the text "94: DESeq2 plots on d ata 88, data 86, and ot hers" and "93: DESeq2 result file on data 88, data 86, and others". Below the header, there is a section titled "DESeq2 run information" which lists sample table details: myceliumXvsXspores, SRR1179895.fastq mycelium, SRR1179896.fastq mycelium, SRR1179892.fastq spores, and SRR1179893.fa. At the bottom of the screenshot, there is a table with three columns: GeneID, Base mean, and Log2 fold change. The first row shows FGRAMPH1_01G25635 with values 82750.1380783884 and 13.7. The second row shows FGRAMPH1_01G08385 with values 24133.4229278897 and -12.1. The third row shows FGRAMPH1_01G15589 with values 25128.2417004296 and 12.6.

1. GeneID	2. Base mean	3. Log2 fold change
FGRAMPH1_01G25635	82750.1380783884	13.7
FGRAMPH1_01G08385	24133.4229278897	-12.1
FGRAMPH1_01G15589	25128.2417004296	12.6

- **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_0029	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
7 D8B26_0066	342.517653	3.85530043	0.20041015	12.7787265	3.08E-40	3.71E-38

Identify Genes based on List of IDs

Configure Search Learn More View Data Sets Used

[Reset values to default](#)

Gene ID input set

Enter a list of IDs or text:
D8B26_000344
D8B26_007030
D8B26_003055
D8B26_006187
D8B26_005929
D8B26_003310

Upload a text file: Choose file | No file chosen
Maximum size 10MB. The file should contain the list of IDs.

Upload from a URL:
The URL should resolve to a list of IDs.

Copy from My Basket: 3 records will be copied from your basket.

Copy from My Strategy: ID list search (? records)

[Get Answer](#)

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

Exporting data to VEuPathDB

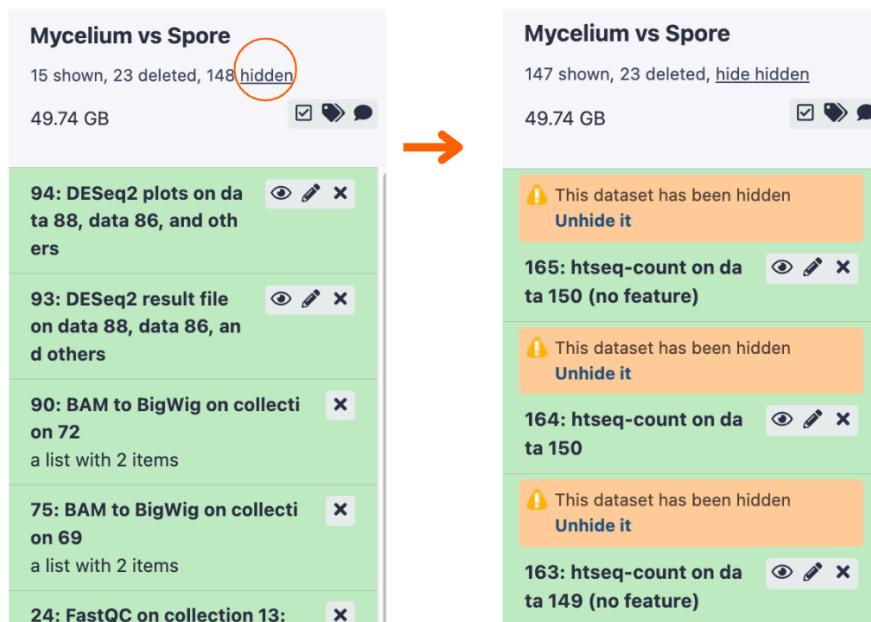
The VEuPathDB RNAseq export tool provides a mechanism to export your RNAseq results (TPM values) and BigWig RNAseq coverage files. The advantage of doing this is that it allows you to search the TPM data using the RNAseq search in VEuPathDB and view the BigWig files in the genome browser.

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.**

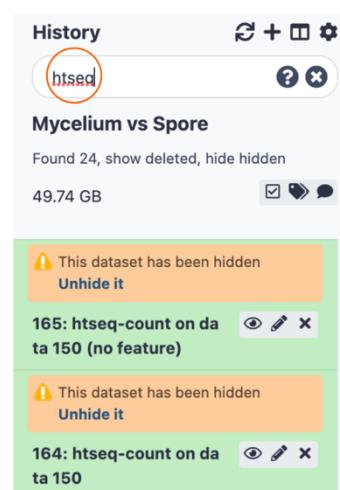
1. **Reveal hidden files.**

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.

Mycelium vs Spore

Found 24, show deleted, hide hidden

49.74 GB



All None

For all selected... ▾

⚠ This dataset has been hidden
[Unhide it](#)

165: htseq-count on data 150
(no feature)

⚠ This dataset has been hidden
[Unhide it](#)

164: htseq-count on data 150

⚠ This dataset has been hidden
[Unhide it](#)

163: htseq-count on data 149
(no feature)

4. “Build dataset list”.

Click on the “For all selected” button and choose the “Build dataset list” option.

Mycelium vs Spore

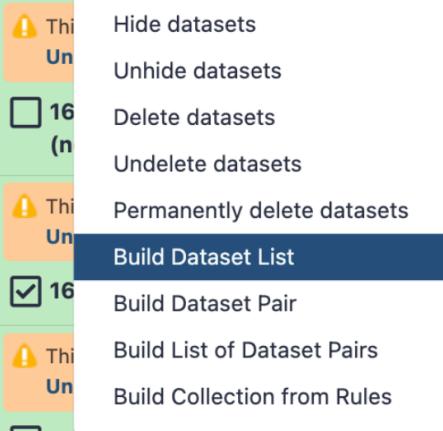
Found 24, show deleted, hide hidden

49.74 GB



All None

For all selected... ▾



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.

The screenshot shows a user interface for creating a dataset collection. At the top, there's a list of four htseq-count samples: "htseq-count on data 74", "htseq-count on data 73", "htseq-count on data 71", and "htseq-count on data 70". An orange arrow points to the "htseq-count on data 73" item. A modal window titled "htseq-count on data 74" is open, prompting the user to "Click to rename". Below the original element, there's a text input field with the placeholder "Enter a name for your new collection" and a value "mycelium 2". There are "Cancel" and "OK" buttons at the bottom of the modal. After clicking "OK", the modal closes, and the renamed element "mycelium 2" appears in the list. The original element "htseq-count on data 73" has been removed. The list now contains "mycelium 2", "mycelium 1", "spore 2", and "spore 1". A second orange arrow points to the "Create list" button at the bottom right of the interface. This button is circled in orange.

- **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.

The screenshot shows the Galaxy web interface with the following steps highlighted:

- 1**: Points to the "HTSeqCountToTPM" tool in the left sidebar under "VEUPATHDB APPLICATIONS".
- 2**: Points to the "gene counts of sense-strand aligned RNA-Seq reads" input field, which contains "175: Mycelium vs spores". A note below says: "This is a batch mode input field. Separate jobs will be triggered for each dataset selection."
- 3**: Points to the "Select a genome annotation" dropdown, which is set to "FungiDB-31_FgraminearumPH-1_Genome".
- 4**: Points to the "Execute" button at the bottom of the form.

- **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.

Tools

VEUPathDB APPLICATIONS

VEuPathDB Export Tools

Gene List to VEuPathDB Export a gene list to VEuPathDB

Bigwig File to VEuPathDB Export one or more bigwig files to VEuPathDB where they can be viewed as tracks in the Genome Browser.

RNA-Seq to VEuPathDB Export an RNA-Seq result to VEuPathDB

VEuPathDB OrthoMCL Tools

VEuPathDB RNA-Seq Tools

HTSeqCountToFPKM compute FPKM from per-gene read counts and reference genome

HTSeqCountToTPM compute TPM from per-gene read counts and reference genome

DATA TRANSFER

Globus Data Transfer

Get Data

Collection Tools

AtlasXomics tools

REDINET tools

agat conversion tools

NGS VISUALIZATION

RNA-Seq to VEuPathDB Export an RNA-Seq result to VEuPathDB (Galaxy Version 1.0.0)

My Data Set name:

hyphae vs spherules

specify a name for the new dataset

Are you exporting sense and antisense TPM/FPKM datasets?

No

Select yes if your experiment is strand-specific and you are including sense and antisense datasets in this export.

BigWig collection:

70: h vs s

Select the BigWig collection to include in the new VEuPathDB My Data Set. The BigWig collection you select here must be mapped to the reference genome that you select below.

TPM or FPKM collection:

72: HTSeqCountToTPM on collection 65: gene expression

Select the TPM or FPKM collection. For an unstranded dataset, its name should include the phrase 'gene expression'.

My Data Set summary:

pathogenic and non-pathogenic stages

My Data Set description:

https://veupathdbprod.globusgenomics.org/

Email notification

Yes No

Send an email notification when the job completes.

Execute

- **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.

My Data Set: *Afumigatus* pre-blood vs 180min

Status: This data set is installed and ready for use in FungiDB.

Owner: Me

Description: Afumigatus

ID: 4032963

Data type: RNA-Seq (RnaSeq 1.0)

Summary: pre blood - 180

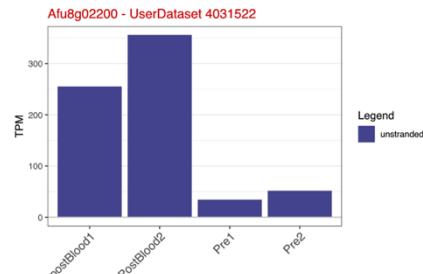
Created: 2 years ago

Data set size: 271.05 M

Quota usage: 2.84% of 10.00 G

Available searches: • RNA-Seq user dataset (fold change)

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 VCF 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.

1 History

Mycelium vs Spore

15 shown, 19 deleted, 148 hidden

49.74 GB

History Actions

Copy

Share or Publish

Show Structure

Make History Accessible via Link and Publish It

This history is currently restricted so that only you and the users listed below can access it. You can:

Also make all objects within the History accessible.

Generates a web link that you can share with other people so that they can view and import the history.

Also make all objects within the History accessible.

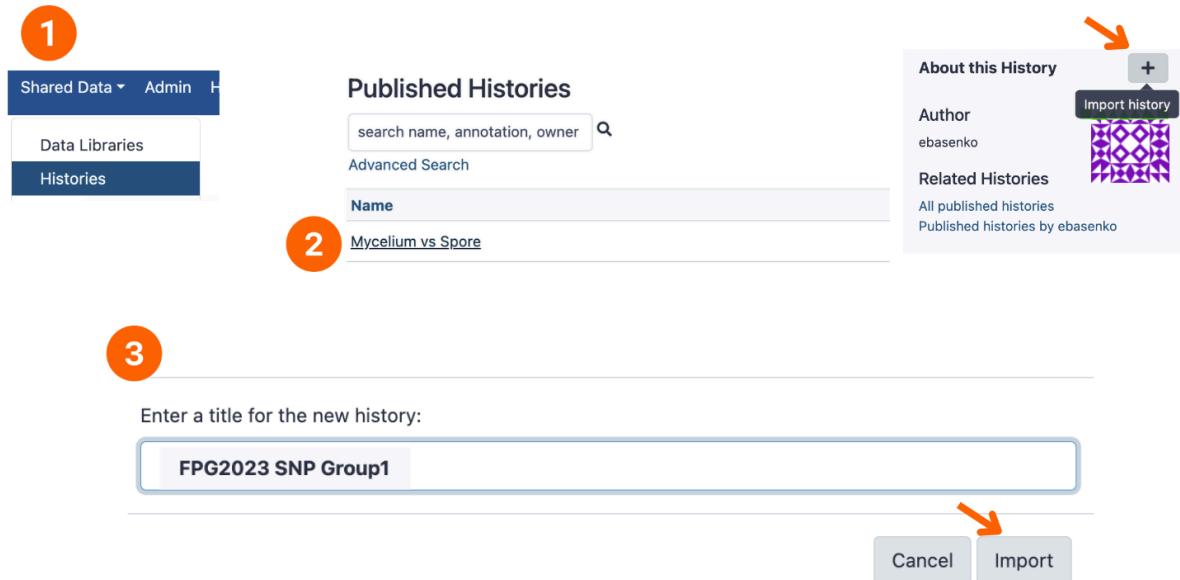
Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, v

Share History with Individual Users

You have not shared this history with any users.

• 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.

FPG2023 SNP GROUP5	
9 shown, 2 deleted, 7 hidden	(orange circle)
11.86 GB	
filter VCF files using arbitrary expressions	→
SnpEff: Analyze and annotate of variants, and calculation of the effects	→
Bowtie: Align reads to a reference genome	→
18: SnpSift Filter on data 16	edit, delete
17: SnpEff on data 15	edit, delete
16: SnpEff on data 15	edit, delete
13: BAM to BigWig on data 12	edit, delete
12: Bowtie2.4.4 on data 8 and data 7: alignments	edit, delete
10: FastQC on data 4: Webpage	edit, delete
5: FastQC on data 3: Webpage	edit, delete
4: SRR10728586_2.fastq.gz	edit, delete
3: SRR10728586_1.fastq.gz	edit, delete

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>

FPG2023 SNP GROUP5
9 shown, 2 deleted, 7 hidden
11.86 GB

18: SnpSift Filter on data 16
17: SnpEff on data 15
16: SnpEff on data 15
13: BAM to BigWig on data 12
12: Bowtie2.4.4 on data 8 and data 7: alignments
10: FastQC on data 4: Webpage
5: FastQC on data 3: Webpage
4: SRR10728586_2.fastq.gz
3: SRR10728586_1.fastq.gz

18: SnpSift Filter on data 16
17: SnpEff on data 15
16: SnpEff on data 15
15: FreeBayes on data 12 (variants) filtered by quality
14: FreeBayes on data 12 (variants)
13: BAM to BigWig on data 12
12: Bowtie2.4.4 on data 8 and data 7: alignments
11: FastQC on data 4: RawData
10: FastQC on data 4: Webpage
9: Singletons from paired-end output of Sickle on data 4 and data 3
8: Sickle on data 4 and data 3

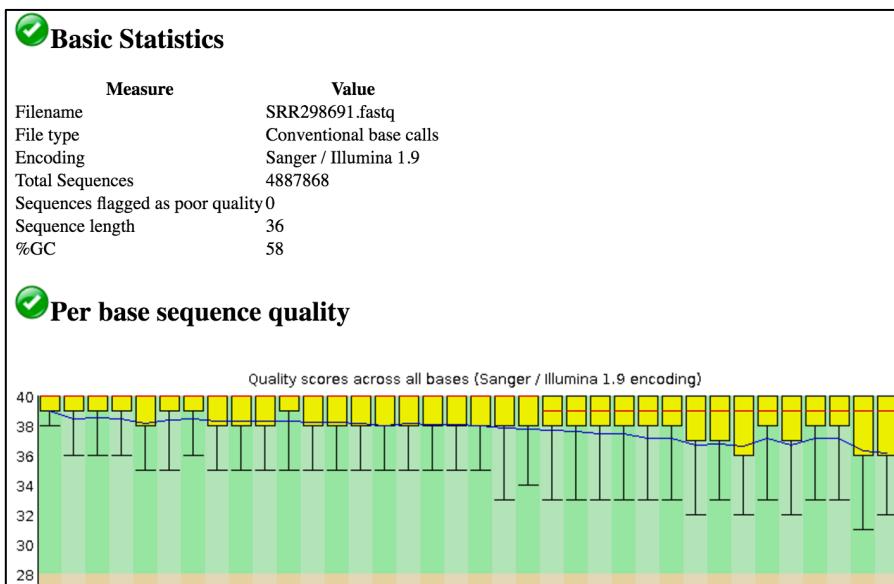
15: FreeBayes on data 12 (variants) filtered by quality
~300,000 lines
format: vcf, database: FungiDB-34_ZtriticilPO323_Genome
Traceback (most recent call last):
File "metadata/set.py", line 1, in <module>
from galaxy_ext.metadata.set_metadata import set_metadata;
set_metadata()
File "/opt/galaxy/lib/galaxy_ext/metadata/set_metadata.py", line 20, in <module>
from gal

display with IGV local

1. Chrom
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth a
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth pe
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of al
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of al
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele fre

- Examine sequence quality based on FastQC quality scores.

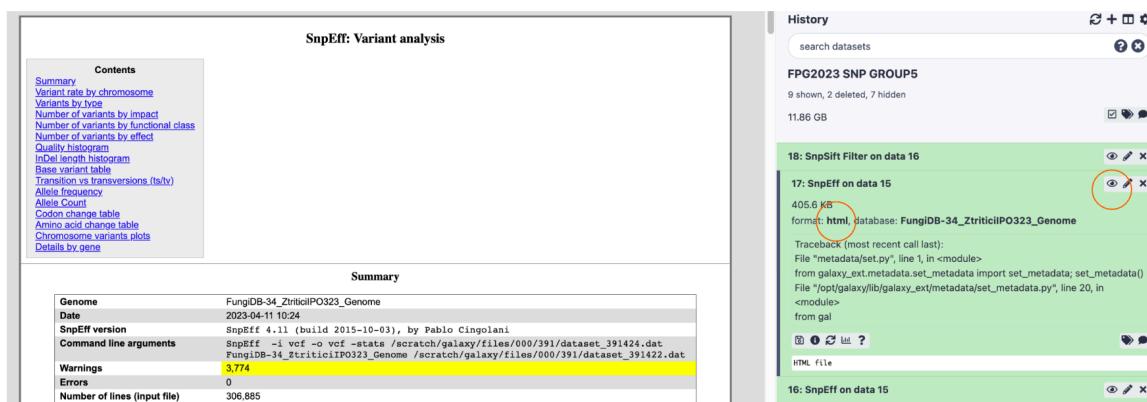
FastQC provides an easy-to-navigate visual representation sequencing data quality and distribution of nucleotides per read position. What does the report tell you about the quality ?



- 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_ZtriticiciIPO323_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_ZtriticiciIPO323_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:

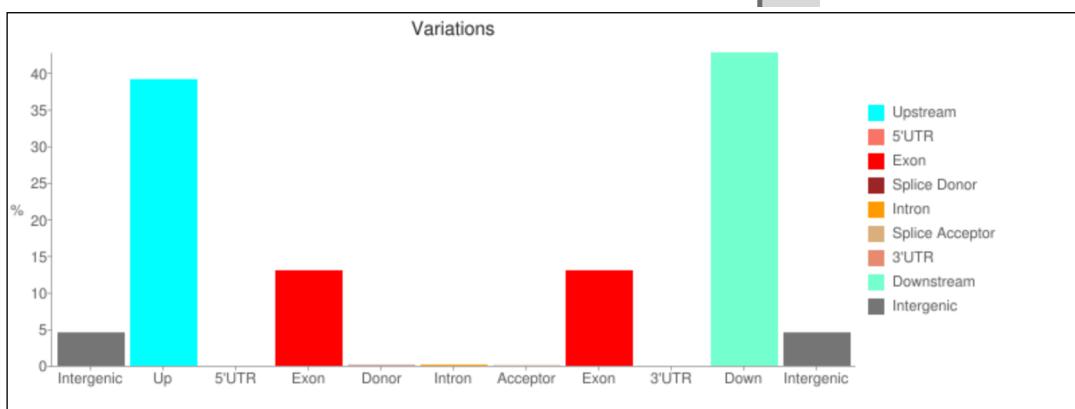
Number of effects by impact

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.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
Ztri_chr_1	133	.	CC	GT	59.2437	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	195	.	CATA	CATG	169.043	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	1565	.	A	G	68.5388	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	1603	.	C	T	140.924	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	1651	.	C	T	114.529	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	1927	.	G	A	113.199	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	1985	.	C	T	250.268	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	2168	.	G	A	100.41	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	2272	.	CAATG	TAATG	191.809	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	2293	.	G	A	206.133	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	2367	.	G	A	54.2829	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	2630	.	C	T	112.111	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	2975	.	C	T	62.699	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	3119	.	GAATG	CAATG	58.621	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	3180	.	C	T	80.1965	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	3723	.	G	A	125.847	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	3812	.	T	C	50.3	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	4453	.	G	A	74.9798	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	4465	.	G	A	109.005	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	4479	.	GC	CT	129.602	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	4495	.	T	C	63.6211	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	5145	.	T	C	132.17	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	5265	.	TA	CG	298.39	.	AB=0;ABP=0;AC=2;AF=1;
Ztri_chr_1	5325	.	G	A	321.168	.	AB=0;ABP=0;AC=2;AF=1;

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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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:QF	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 will 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.

```

9;SRF=6;SRP=26.4622;SRR=23;TYPE=snp;ANN=A|missense_variant|MODERATE|Afu1g00140|Afu1g00140|transcript|Afu1g00140-T|Coding|1;SRP=29.6108;SRR=14;TYPE=snp;ANN=A|missense_variant&splice_region_variant|MODERATE|Afu1g00140|Afu1g00140|transcript|Afu1g00140-T|Coding|1;SRP=5.18177;SRR=0;TYPE=snp;ANN=C|missense_variant|MODERATE|Afu1g00230|Afu1g00230|transcript|Afu1g00230-T|Coding|1;SRP=0;SRR=0;TYPE=complex;ANN=G|GC|missense_variant|MODERATE|Afu1g00230|Afu1g00230|transcript|Afu1g00230-T|Coding|1;SRP=0;SRR=0;TYPE=complex;ANN=G|GC|missense_variant|MODERATE|Afu1g00230|Afu1g00230|transcript|Afu1g00230-T|Coding|1;SRP=5.18177;SRR=0;TYPE=snp;ANN=T|missense_variant|MODERATE|Afu1g00230|Afu1g00230|transcript|Afu1g00230-T|Coding|1;SRP=0;SRR=5.18177;SRR=0;TYPE=snp;ANN=A|missense_variant|MODERATE|Afu1g00230|Afu1g00230|transcript|Afu1g00230-T|Coding|1;SRP=5.18177;SRR=0;TYPE=snp;ANN=G|missense_variant|MODERATE|Afu1g00230|Afu1g00230|transcript|Afu1g00230-T|Coding|1;SRP=5.18177;SRR=0;TYPE=snp;ANN=G|missense_variant|MODERATE|Afu1g00230|Afu1g00230|transcript|Afu1g00230-T|Coding|1;SRP=0;SRR=0;TYPE=complex;ANN=GATCGA|missense_variant|MODERATE|Afu1g00230|Afu1g00230|transcript|Afu1g00230-T|Coding|1;SRP=5.18177;SRR=0;TYPE=snp;ANN=G|missense_variant|MODERATE|Afu1g00300|Afu1g00300|transcript|Afu1g00300-T|Coding|1;SRP=5.18177;SRR=0;TYPE=snp;ANN=C|splice_acceptor_variant&intron_variant|HIGH|Afu1g00300|Afu1g00300|transcript|Afu1g00300-T|Coding|1;SRP=5.18177;SRR=0;TYPE=snp;ANN=C|missense_variant|MODERATE|Afu1g00300|Afu1g00300|transcript|Afu1g00300-T|Coding|1;SRP=5.18177;SRR=0;TYPE=snp;ANN=G|missense_variant|MODERATE|Afu1g00300|Afu1g00300|transcript|Afu1g00300-T|Coding|3/5;SRP=5.18177;SRR=1;TYPE=mnp;ANN=AGT|missense_variant|MODERATE|Afu1g00300|Afu1g00300|transcript|Afu1g00300-T|Coding|1;SRP=5.18177;SRR=0;TYPE=snp;ANN=A|stop_gained|HIGH|Afu1g00550|Afu1g00550|transcript|Afu1g00550-T|Coding|2/c.575G>A;SRP=0;SRR=5.18177;SRR=0;TYPE=snp;ANN=T|missense_variant|MODERATE|Afu1g00550|Afu1g00550|transcript|Afu1g00550-T|Coding|1;SRP=0;SRR=5.18177;SRR=0;TYPE=snp;ANN=T|missense_variant|MODERATE|Afu1g00550|Afu1g00550|transcript|Afu1g00550-T|Coding|1;SRP=0;SRR=5.18177;SRR=0;TYPE=snp;ANN=A|missense_variant|MODERATE|Afu1g00550|Afu1g00550|transcript|Afu1g00550-T|Coding|2/c.697G>A;TYPE=mnp;ANN=TT|missense_variant|MODERATE|Afu1g00570|Afu1g00570|transcript|Afu1g00570-T|Coding|1;c.910_911delGTinsA;SRP=0;SRR=5.18177;SRR=0;TYPE=snp;ANN=T|missense_variant|MODERATE|Afu1g00570|Afu1g00570|transcript|Afu1g00570-T|Coding|1;SRP=0;SRR=0;TYPE=complex;ANN=TATT|stop_gained|HIGH|Afu1g00570|Afu1g00570|transcript|Afu1g00570-T|Coding||c.892_896delGAAT

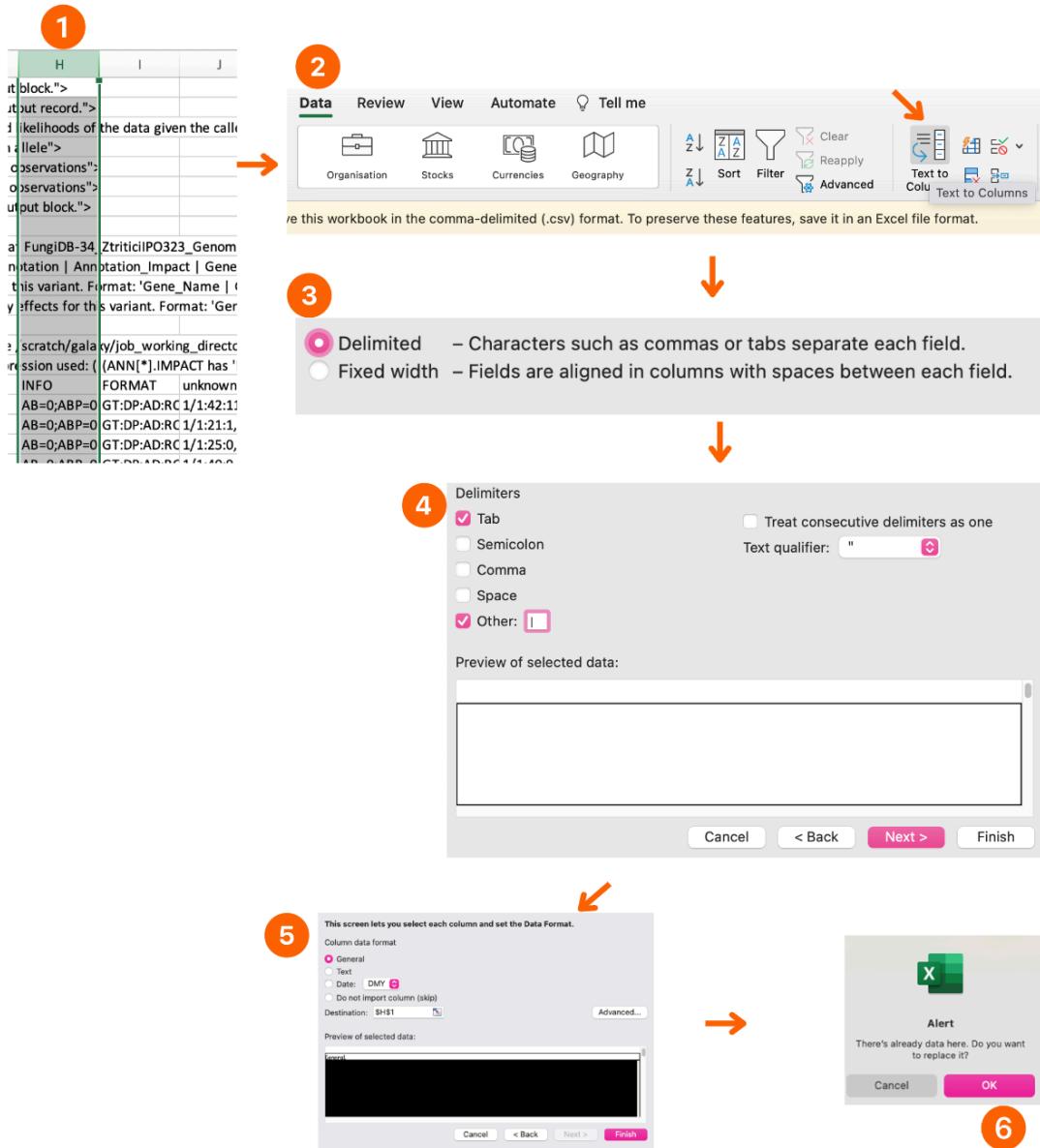
```

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.

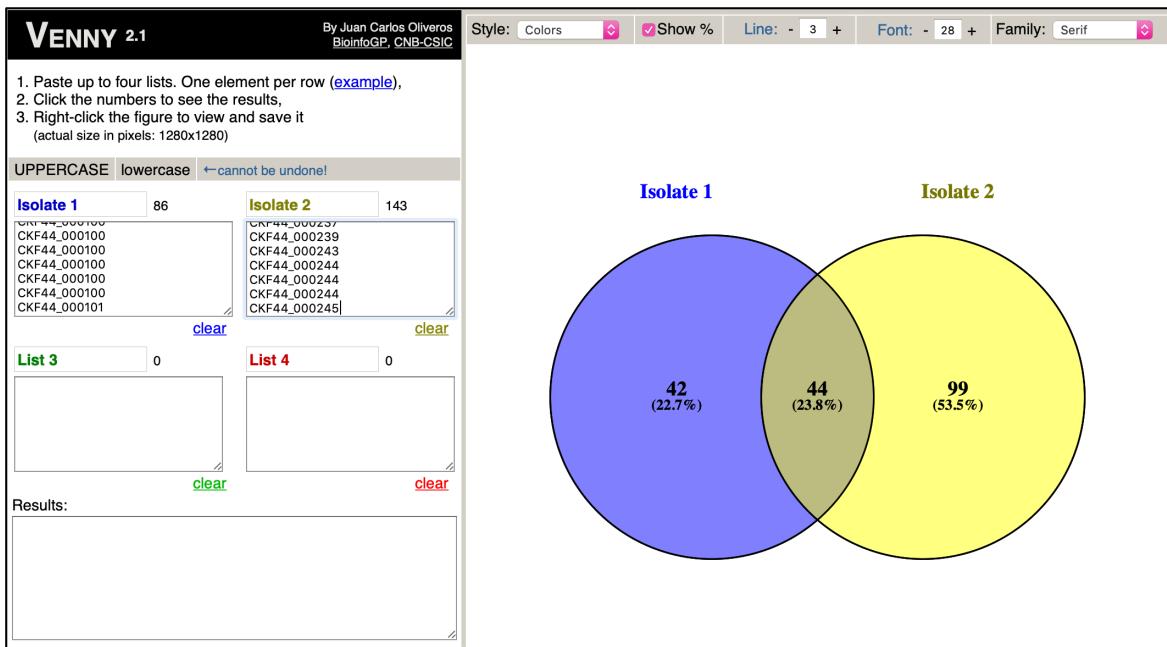
e_ID	Feature_Typ	Feature_ID	Transcript_E_Rank	HGVS.c	HGVS.p	cDNA.pos / tCDN.pos / Cl_AA.pos / AA	Distance	ERRORS / WARNINGS / INFO!
49	Genotype Quality, the Phred-scaled marginal (or unconditional) probability of the called genotype">							
50	Genotype Likelihood, log10-scaled likelihoods of the data given the called genotype for each possible genotype generated from the reference and alternate alleles given the sample ploidy">							
51	"Read Depth">							
52	"Reference allele observation count">							
53	"Sum of quality of the reference observations">							
54	"Alternate allele observation count">							
55	"Sum of quality of the alternate observations">							
56	"an">							
57	les/008/dataset_8077.dat PlasmoDB-29_Pfalciparum3D7_Genome /scratch/galaxy/files/008/dataset_8075.dat "							
58	e_ID	Feature_Typ	Feature_ID	Transcript_E_Rank	HGVS.c	HGVS.p	cDNA.pos / tCDN.pos / Cl_AA.pos / AA	Distance
59	59	scriptsAffected">						
60	59	scriptsAffected">						
61	61	o Cingolani"						
62	62	008/dataset_8076.dat -e /scratch/galaxy/job_working_directory/004/4170/tmpAQDb8H"						
63								
64	QUAL	FILTER	INFO	FORMAT	unknown			
65	163.615...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	c.1729_1730 p.Asp577Pro	t229/6492	1729/6492 577/2163
66	59.2743...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	2-Jan c.1773A>T p.Lys591Asn	t173/6492	173/6492 591/2163
67	112.419...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	c.4420_4421 p.Thr1474Gln	t420/6492	4420/6492 1474/2163
68	123.945...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	2-Jan c.4432C>G p.Gln1478Gln	t432/6492	4432/6492 1478/2163
69	70.7189...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	2-Jan c.4466C>A p.Thr1489Ile	t4466/6492	4466/6492 1489/2163
70	203.132...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	2-Jan c.4655T>G p.Leu1552Asn	t4655/6492	4655/6492 1552/2163
71	149.708...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	c.4733_4734 p.Asp1578Ala	t4733/6492	4733/6492 1578/2163
72	101.922...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	2-Jan c.4741C>A p.Gln1581Ile	t4741/6492	4741/6492 1581/2163
73	106.751...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	2-Feb c.5647A>G p.Asn1883Asn	t5647/6492	5647/6492 1883/2163
74	68.702...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	2-Feb c.5873C>G p.Thr1958Ser	t5873/6492	5873/6492 1958/2163
75	599.479...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	c.6472_6474 p.Ala2158Ser	t6472/6492	6472/6492 2158/2163
76	6.44607...	A B 0 AB =0;missense_va MODERATE	Pf3D7_0100100	Pf3D7_0100 transcript	Pf3D7_0100 Coding	c.6490_6492 p.Ile2160Leu	t6490/6492	6490/6492 2160/2163

- 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.

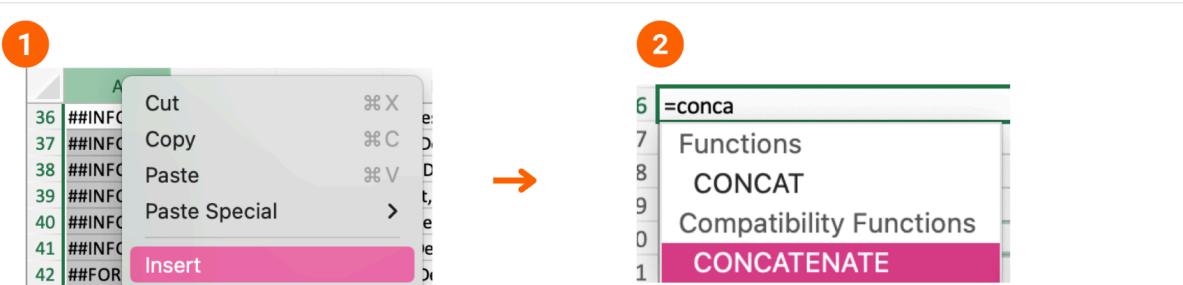
If you are comparing two or more strains, you may want to extract gene IDs from all VCF files and identify common signatures across isolates or strains. For this type of analysis, you can use <http://bioinfogp.cnb.csic.es/tools/venny/> to generate a Venn diagram:



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



3

SUM	A	B	C	D	E	F	G	H	I	J	K	L	M	N
50		#INFO<=ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this variant. Format: 'Gene_Name Gene_ID Number_of_transcripts_in_gene Percent_of_transcripts_in_gene Description'"												
51		#INFO<=ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for this variant. Format: 'Gene_Name Gene_ID Number_of_transcripts_in_gene Percent_of_transcripts_in_gene Description'"												
52		#SnpSiftVersion="SnpSift 4.1 (build 2015-10-03), by Pablo Cingolani"												
53		#SnpSiftCmd="SnpSift filter -f /scratch/galaxy/files/000/391/dataset_391071.dat -e /scratch/galaxy/job_working_directory/000/260/260223/configs/tmpufmf1sa3"												
54		#FILTER=<ID=SnpSift,Description="SnpSift 4.1 (build 2015-10-03), by Pablo Cingolani, Expression used: (((ANN*)>IMPACT has 'HIGH') & ((ANN*)>MODERATE) & ((na FILTER)"												
55		#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO					
56	=CONCATENATE(B56,".",C56,".",M56)	Chr1_A_fumigatus_Af293	1314781		AATA	GATG	85.5736	.	AB=0:ABP=0:/ missense Va MODERATE	Afu1g00410	Afu1g00410	transcript	/	
57		Chr1_A_fumigatus_Af293	131514.		T	C	72.9308	.	AB=0:ABP=0:/ missense Va MODERATE	Afu1g00410	Afu1g00410	transcript	/	
58		Chr1_A_fumigatus_Af293	143640.		T	C	97.7793	.	AB=0:ABP=0:/ missense Va MODERATE	Afu1g00450	Afu1g00450	transcript	/	
59		Chr1_A_fumigatus_Af293	144396.		G	A	135.073	.	AB=0:ABP=0:/ missense Va MODERATE	Afu1g00450	Afu1g00450	transcript	/	

You should get unique SNP IDs that look like this (for example):

CP022321.1:15259:CKF44_000003. Copy this function for other entries:

Chr1_A_fumigatus_Af293:185468:Afu1g00580	Chr1_A_fumigatus_Af293	185468	.	TTC
Chr1_A_fumigatus_Af293:185521:Afu1g00580	Chr1_A_fumigatus_Af293	185521	.	A
Chr1_A_fumigatus_Af293:401061:Afu1g01110	Chr1_A_fumigatus_Af293	401061	.	G
Chr1_A_fumigatus_Af293:402973:Afu1g01120	Chr1_A_fumigatus_Af293	402973	.	GG
Chr1_A_fumigatus_Af293:403260:Afu1g01120	Chr1_A_fumigatus_Af293	403260	.	A
Chr1_A_fumigatus_Af293:405284:Afu1g01130	Chr1_A_fumigatus_Af293	405284	.	T
Chr1_A_fumigatus_Af293:405434:Afu1g01130	Chr1_A_fumigatus_Af293	405434	.	A
Chr1_A_fumigatus_Af293:406035:Afu1g01140	Chr1_A_fumigatus_Af293	406035	.	G
Chr1_A_fumigatus_Af293:406481:Afu1g01140	Chr1_A_fumigatus_Af293	406481	.	G
Chr1_A_fumigatus_Af293:407398:Afu1g01160	Chr1_A_fumigatus_Af293	407398	.	A
	Chr1_A_fumigatus_Af293	407406	.	A
	Chr1_A_fumigatus_Af293	410505	.	C

4. Copy these newly generated unique IDs into List 1 and List 2 on Venny <http://bioinfogp.cnb.csic.es/tools/venny/> and examine the data.

4

Chr1_A_fumigatus_Af293:145783:Afu1g00460
Chr1_A_fumigatus_Af293:148888:Afu1g00470
Chr1_A_fumigatus_Af293:148933:Afu1g00470
Chr1_A_fumigatus_Af293:148945:Afu1g00470
Chr1_A_fumigatus_Af293:185087:Afu1g00580
Chr1_A_fumigatus_Af293:185100:Afu1g00580
Chr1_A_fumigatus_Af293:185439:Afu1g00580
Chr1_A_fumigatus_Af293:185468:Afu1g00580
Chr1_A_fumigatus_Af293:185521:Afu1g00580
Chr1_A_fumigatus_Af293:401061:Afu1g01110
Chr1_A_fumigatus_Af293:402973:Afu1g01120
Chr1_A_fumigatus_Af293:403260:Afu1g01120
Chr1_A_fumigatus_Af293:405284:Afu1g01130
Chr1_A_fumigatus_Af293:405434:Afu1g01130
Chr1_A_fumigatus_Af293:406035:Afu1g01140
Chr1_A_fumigatus_Af293:406481:Afu1g01140

VENNY 2.1 By Juan Carlos Olvera (bioinfogp.cnb.csic.es)

1. Paste up to four lists. One element per row ([example](#)).
2. Click the numbers to see the results.
3. Right-click the figure to view and save it
(actual size in pixels: 1280x1280)

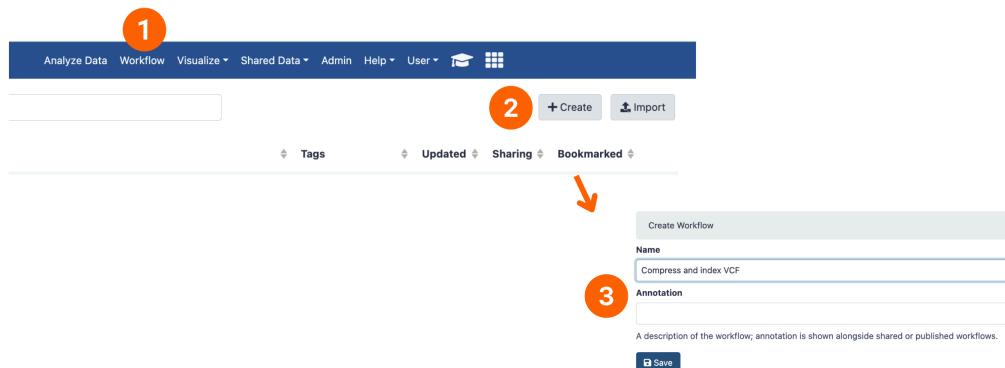
UPPERCASE	lowercase	← cannot be undone!	
List 1	12	List 2	12
Chr1_A_fumigatus_Af293:145783:Afu1g00460	Chr1_A_fumigatus_Af293:185439:Afu1g00580	Chr1_A_fumigatus_Af293:185439:Afu1g00580	Chr1_A_fumigatus_Af293:185468:Afu1g00580
Chr1_A_fumigatus_Af293:148888:Afu1g00470	Chr1_A_fumigatus_Af293:185468:Afu1g00580	Chr1_A_fumigatus_Af293:185468:Afu1g00580	Chr1_A_fumigatus_Af293:185521:Afu1g00580
Chr1_A_fumigatus_Af293:148933:Afu1g00470	Chr1_A_fumigatus_Af293:185521:Afu1g00580	Chr1_A_fumigatus_Af293:185521:Afu1g00580	Chr1_A_fumigatus_Af293:401061:Afu1g01110
Chr1_A_fumigatus_Af293:148945:Afu1g00470	Chr1_A_fumigatus_Af293:401061:Afu1g01110	Chr1_A_fumigatus_Af293:401061:Afu1g01110	Chr1_A_fumigatus_Af293:402973:Afu1g01120
Chr1_A_fumigatus_Af293:185087:Afu1g00580	Chr1_A_fumigatus_Af293:402973:Afu1g01120	Chr1_A_fumigatus_Af293:402973:Afu1g01120	
Chr1_A_fumigatus_Af293:185100:Afu1g00580			
Chr1_A_fumigatus_Af293:185439:Afu1g00580			
Chr1_A_fumigatus_Af293:185468:Afu1g00580			
Chr1_A_fumigatus_Af293:185521:Afu1g00580			
Chr1_A_fumigatus_Af293:401061:Afu1g01110			
Chr1_A_fumigatus_Af293:402973:Afu1g01120			
Chr1_A_fumigatus_Af293:403260:Afu1g01120			
Chr1_A_fumigatus_Af293:405284:Afu1g01130			
Chr1_A_fumigatus_Af293:405434:Afu1g01130			
Chr1_A_fumigatus_Af293:406035:Afu1g01140			
Chr1_A_fumigatus_Af293:406481:Afu1g01140			

Viewing VCF file results in the JBrowse genome browser.

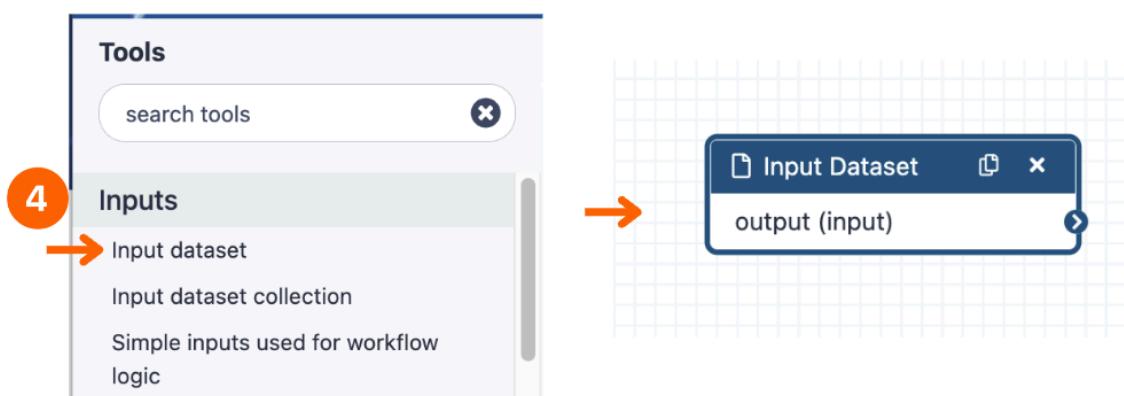
- **Create a workflow to generate a compressed vcf and index files for viewing your data in JBrowse.**

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:

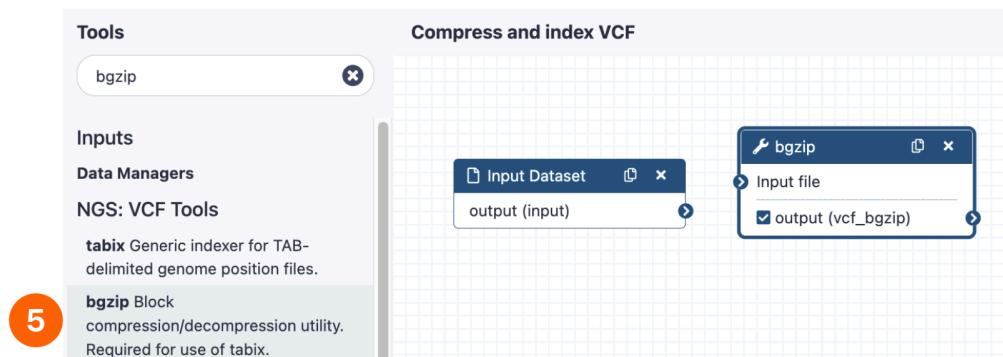
1. Click on the “Workflow” menu.
2. Click on the “Create” button to start a new workflow.
3. Give the workflow a name (e.g. Compress and index VCF) and click on the save button. This will open a workflow canvas.



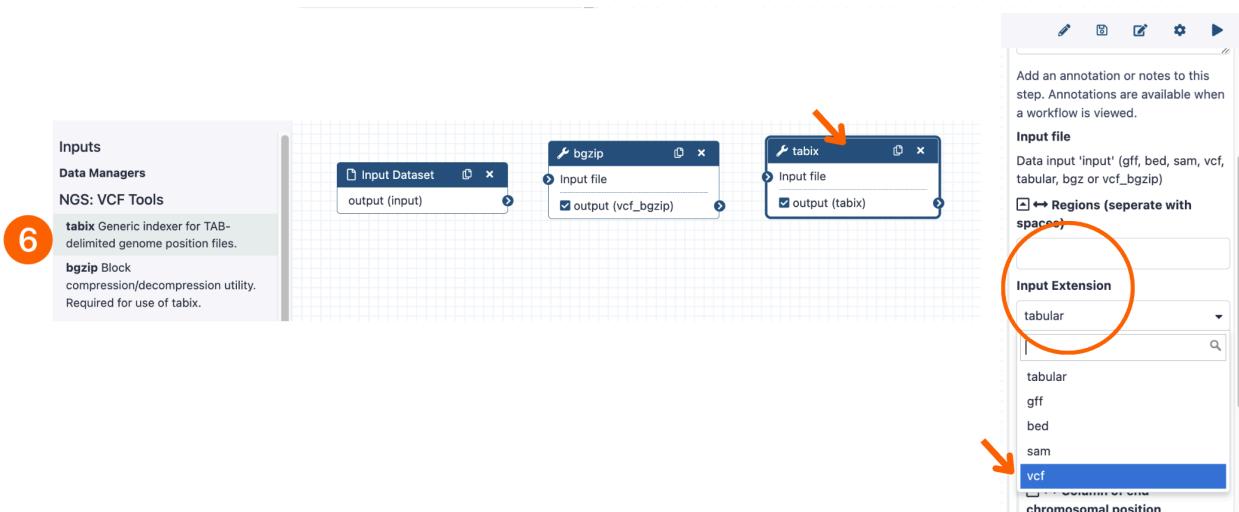
4. All workflows must start with an input file so add the “Input Dataset” step to the workflow using the menu on the left (you must click on the tool for it to appear in the workflow editor canvas).



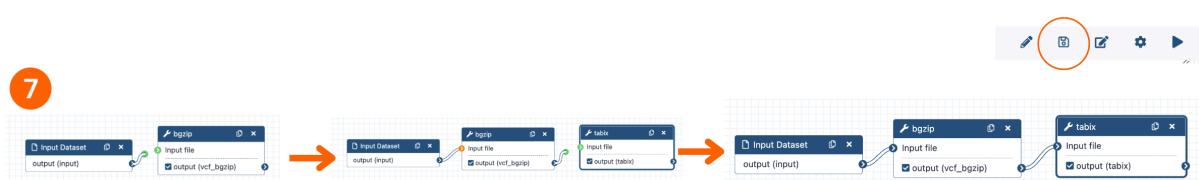
5. Using the menu on the left, search for and add the “bgzip” tool.



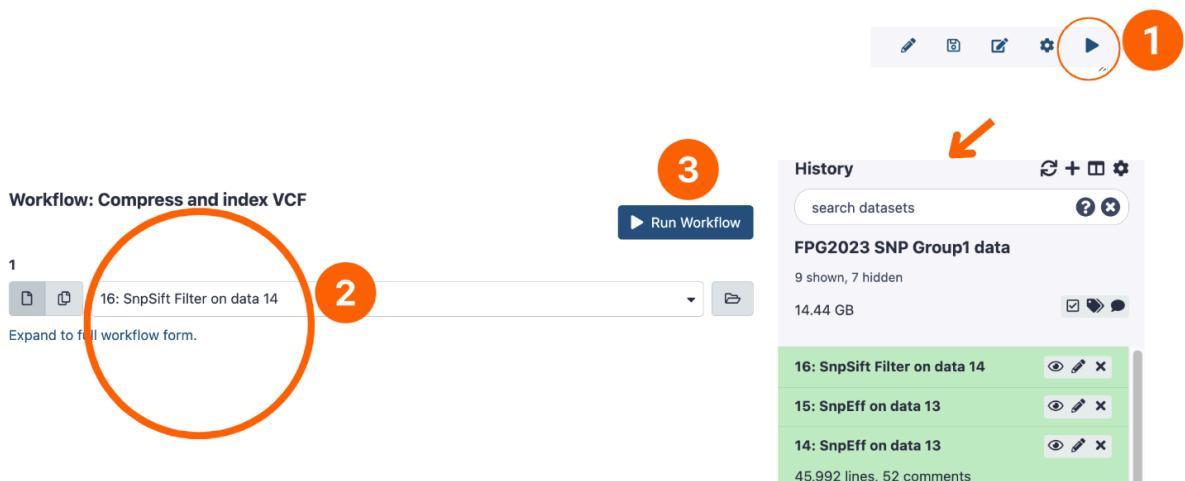
6. Using the menu on the left, search for and add the “tabix” tool. Left-click on the “tabix” icon and select “vcf” under “input selection” on the right (tool option section)



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.
 - Click on the “Play” button to start your workflow.
 - Select the VCF file you want to process.
 - 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).

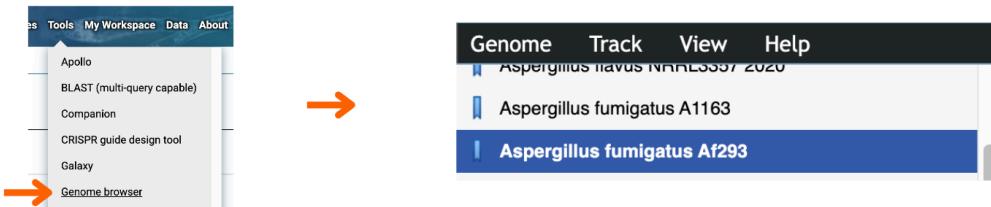
The screenshot shows the Galaxy interface after the workflow has run. On the left, a message box indicates the workflow was successfully invoked. Below it are buttons for "View Report 1" and "Download BioCompute Object". To the right, the "History" panel shows two new entries: "20: tabix on data 19" and "19: bgzip on data 14". Both entries have their download icons circled in red.

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

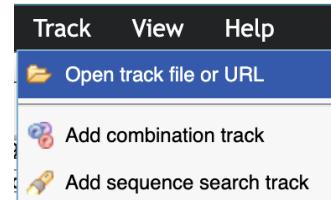
The screenshot shows the Galaxy interface displaying the details of the downloaded files. "19: bgzip on data 14" is listed as 7.2 MB, format vcf_bgzip, and "20: tabix on data 19" is listed as 14.2 KB, format tabix. Both are categorized as "binary data". Their download icons are circled in red.

- After the files are downloaded, rename them as follows:
 - The **vcf_bgzip** file to “**group#.vcf.gz**” (i.e. **group1.vcf.gz**)
 - 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.



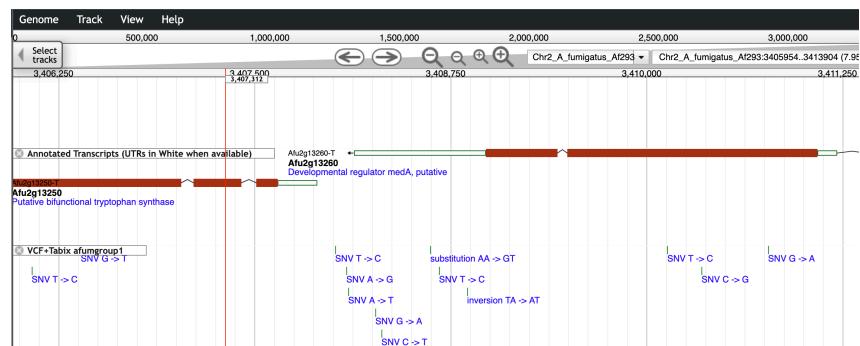
4. Click on the Track menu, select "Open track file or URL".



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

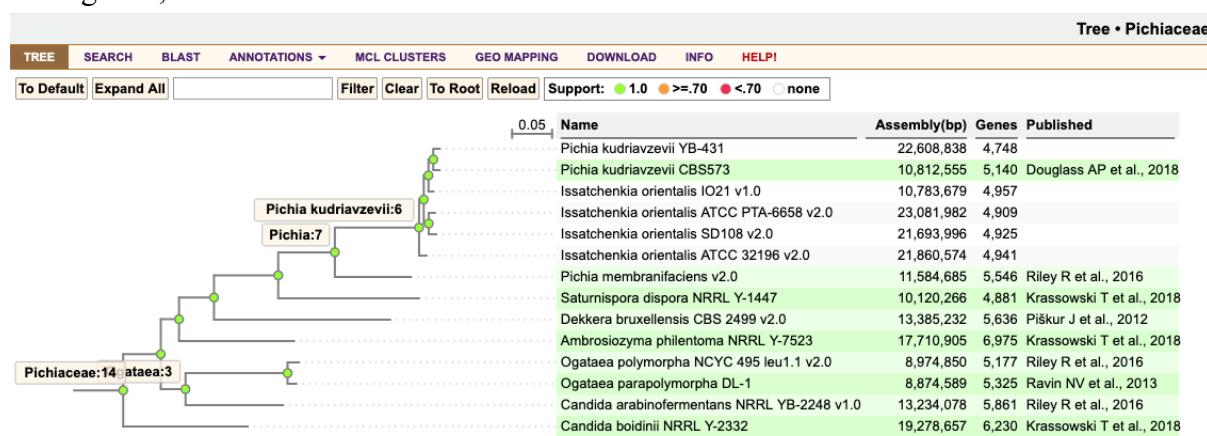
KEGG stands for Kyoto Encyclopedia of Genes and Genomes at <http://www.genome.jp/kegg/>, which maintains a curated set of EC-annotated enzymes and their pathways. Each portal's KEGG Browser facilitates display and discovery of MycoCosm's KEGG-annotated genes. Using the KEGG browser, one can search or 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.

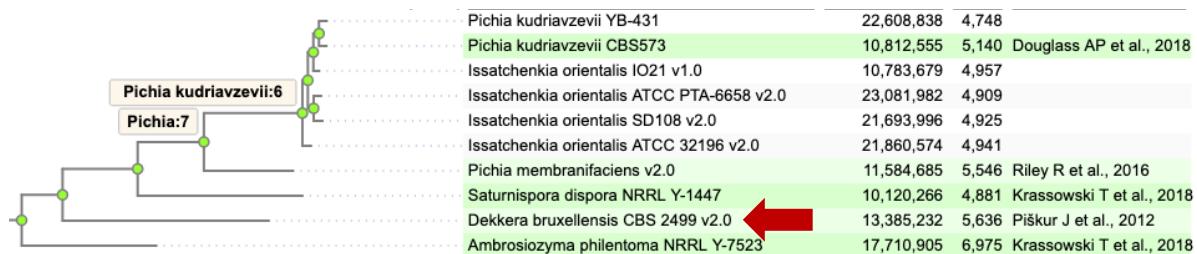
- 1) Go to the MycoCosm Pichiaceae PhyloGroup at mycocosm.jgi.doe.gov/Pichiaceae:

Info • Pichiaceae								
TREE	SEARCH	BLAST	ANNOTATIONS ▾	MCL CLUSTERS	GEO MAPPING	DOWNLOAD	INFO	HELP!
## 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:

The screenshot shows the KEGG browser interface for Dekkera bruxellensis CBS 2499 v2.0. The navigation bar includes SEARCH, BLAST, BROWSE, ANNOTATIONS (selected), MCL CLUSTERS, SYNTENY, DOWNLOAD, INFO, HOME, STATUS, and HELP. The ANNOTATIONS dropdown menu is open, showing options like GENE ONTOLOGY, PFAM DOMAINS, KEGG (selected), KOG, and SECONDARY METABOLISM CLUSTERS. Below the menu, a sidebar displays “models in Dekkera bruxellensis CBS 2499 v2.0 FilteredModels1 (ver 1)”, listing 206 models in total. The main content area shows the KEGG Metabolic Pathway for Amino Acid Metabolism, with links to Alanine, aspartate and glutamate metabolism (27 models) and Arginine and proline metabolism (45 models).

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

Carbohydrate Metabolism	332
Amino sugar and nucleotide sugar metabolism	<u>68</u>
Ascorbate and aldarate metabolism	<u>21</u>
Butanoate metabolism	<u>34</u>
C5-Branched dibasic acid metabolism	<u>2</u>
Citrate cycle (TCA cycle)	<u>28</u>
Fructose and mannose metabolism	<u>46</u>
Galactose metabolism	<u>24</u>
Glycolysis / Gluconeogenesis	<u>47</u>
Glyoxylate and dicarboxylate metabolism	<u>10</u>
Inositol phosphate metabolism	<u>27</u>

- 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*:

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](#)**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):

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

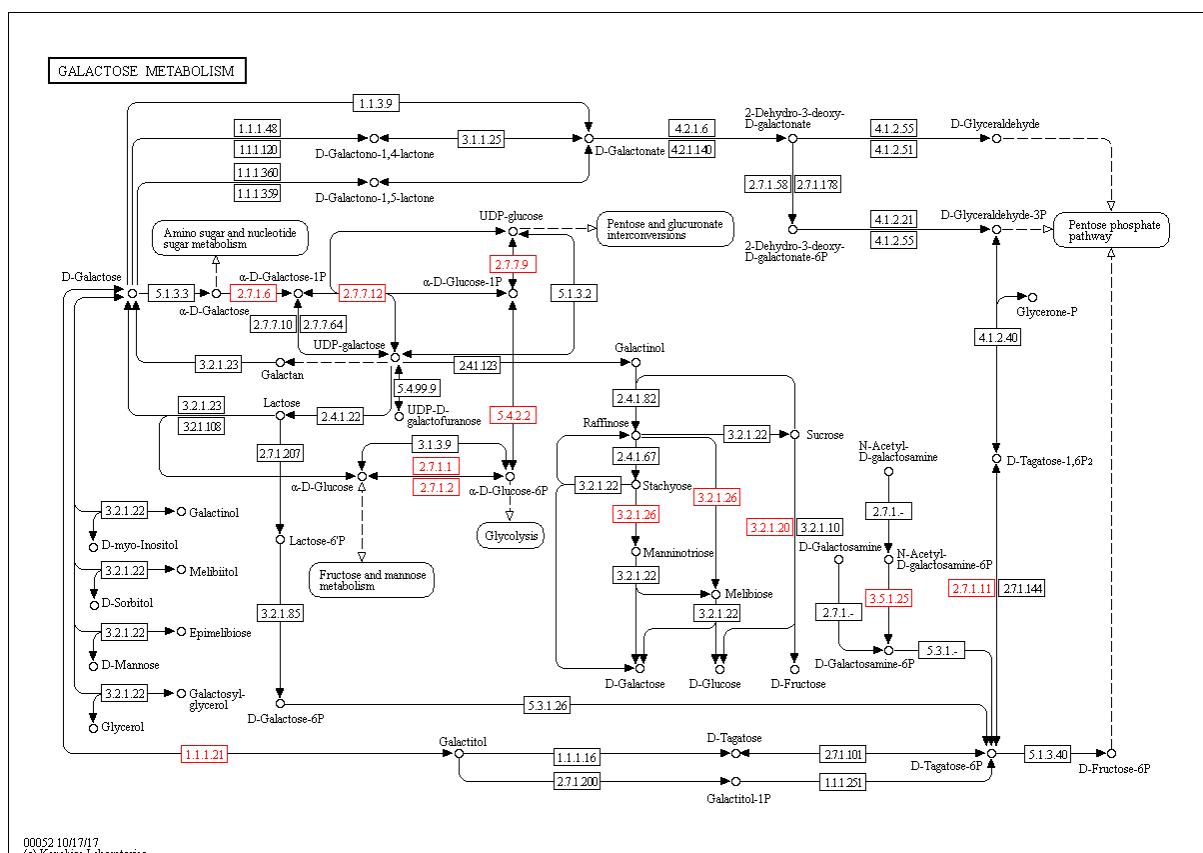
MAP00052: Galactose metabolism

[Summary View | Model View | View KEGG Map]

[Click here to open KEGG Map](#)

EC Number Description

models in
19 v2.0 Pichia membranifaciens v2.0 models in
all selected
FilteredModels1
FilteredModels1



- 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.

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)

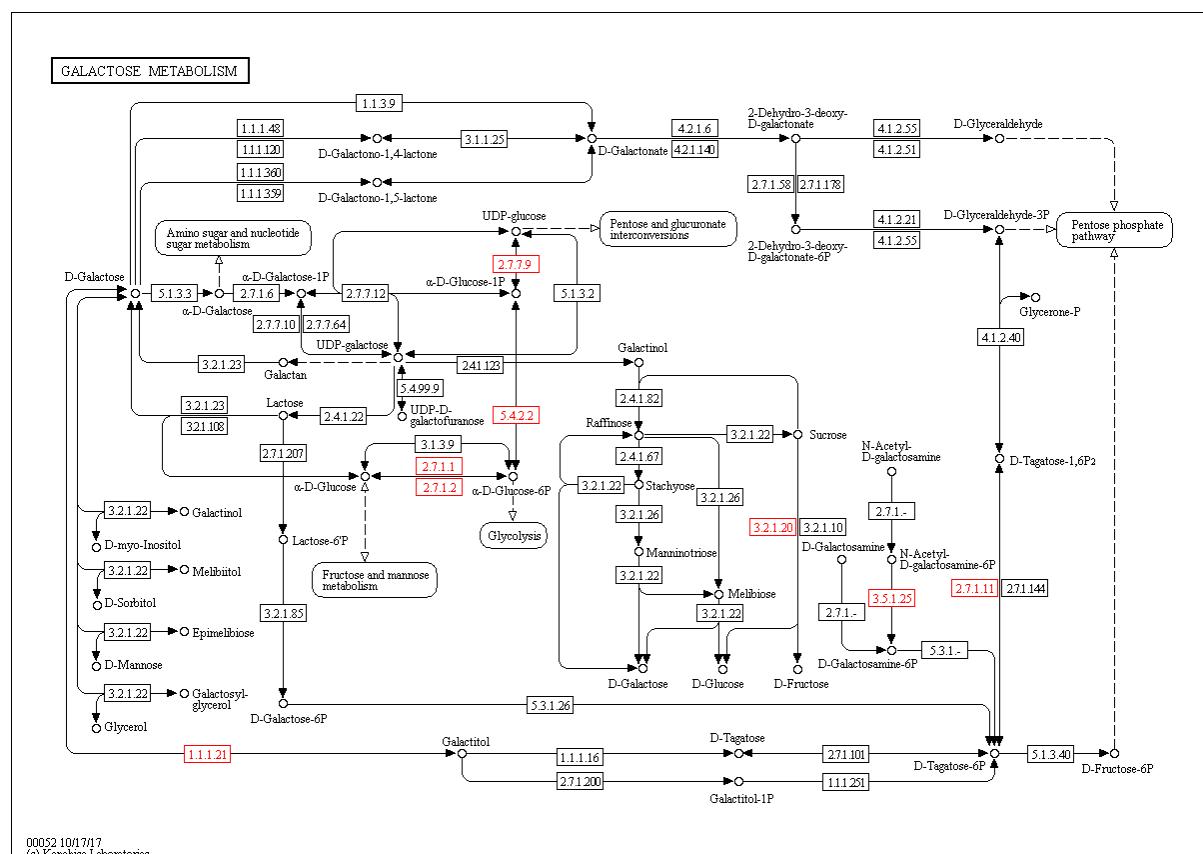
Other Functions

[View KEGG Metabolic Pathways](#)
[View KEGG Regulatory Pathways](#)
[Search KEGG](#)

MAP00052: Galactose metabolism

[Summary View | Model View | View KEGG Map]

- 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.



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 populincola*, which infects a different species of poplar and does not induce cankers. The SM's structure suggests that its biosynthetic pathway may have as its core enzyme a hybrid PKS-NRPS (polyketide synthase-nonribosomal peptide synthetase). Use MycoCosm to find candidate gene clusters for this pathway.

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

Info • Septoria

SEARCH	BLAST	ANNOTATIONS ▾	MCL CLUSTERS	GEO MAPPING	DOWNLOAD	INFO	HELP!
## Name Assembly Length # Genes Published							
1 Septoria musiva SO2202 v1.0 29,352,103 10,233 Ohm RA et al., 2012							
2 Septoria populincola v1.0 33,188,813 9,739 Ohm RA et al., 2012							

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

Home • **Septoria musiva SO2202 v1.0**

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



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

Septoria musiva (sexual stage: *Mycosphaerella populinorum*) causes leaf spots and cankers on poplars (*Populus spp.* and hybrids). On native North American poplars the pathogen mainly causes leaf spots that can lead to defoliation but generally do not kill the host. But *S. musiva* can also cause cankers on branches and primary stems. These can be lethal and are particularly severe on hybrid poplars in plantations. They often develop on the primary shoots of 2- to 3-year-old trees, leading to restrictions in the movement of water and nutrients and weakening the wood within a few feet of ground level. The weakened trunks collapse easily, greatly reducing the production of biomass. Cankers caused by *S. musiva* can greatly hamper the production of hybrid poplars in the eastern United States and Canada and threaten poplars in western North America.

A major concern with *S. musiva* is with migration to new areas. The pathogen is endemic and appears to have originated on poplars in eastern North America, where it occurs commonly on leaves of the eastern cottonwood, *P. deltoides*. During the past 20 years *S. musiva* has appeared in South America and western Canada, where it is spreading rapidly on native and hybrid poplars causing economic damage as well as threatening native poplars in important riparian zones. It is not yet known in Europe or Asia but has the potential to cause extensive damage if introduced to those areas. Global warming and trade may facilitate the spread of the disease by making northern popular-growing areas more favorable to growth of the fungus.

Availability of a genome sequence for *S. musiva* will help with designing strategies to

- 3) Click on “ANNOTATIONS => SECONDARY METABOLISM CLUSTERS” to go to the portal’s SM clusters browser:

Secondary Metabolism Clusters • Septoria musiva SO2202 v1.0

SEARCH	BLAST	BROWSE	ANNOTATIONS ▾	MCL CLUSTERS	SYNTENY	DOWNLOAD	INFO	HOME	STATUS	HELP!												
			GENE ONTOLOGY PFAM DOMAINS KEGG KOG SECONDARY METABOLISM CLUSTERS CAZYMES PEPTIDASES TRANSPORTERS TRANSCRIPTION FACTORS																			
			<input type="button" value="Refresh"/>																			
			Genomes <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Alternaria brassicicola</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Baudoinia compniacensis UAMH 1</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Cochliobolus heterostrophus C5</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Dothistroma septosporum NZE10</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Hysterium pulicare</div> Genome <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Alternaria brassicicola</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Baudoinia compniacensis UAMH 10762 (4089826) v1.0</div>	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>NRPS</th> <th>NRPS-Like</th> <th>PKS</th> <th>PKS-Like</th> <th>TC</th> <th>Total</th> </tr> </thead> <tbody> <tr> <td>4</td> <td>6</td> <td>6</td> <td>3</td> <td>5</td> <td>24</td> </tr> <tr> <td>2</td> <td>6</td> <td>2</td> <td>2</td> <td>1</td> <td>13</td> </tr> </tbody> </table>	NRPS	NRPS-Like	PKS	PKS-Like	TC	Total	4	6	6	3	5	24	2	6	2	2	1	13
NRPS	NRPS-Like	PKS	PKS-Like	TC	Total																	
4	6	6	3	5	24																	
2	6	2	2	1	13																	

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

Secondary Metabolism Clusters • Septoria musiva SO2202 v1.0

SEARCH	BLAST	BROWSE	ANNOTATIONS ▾	MCL CLUSTERS	SYNTENY	DOWNLOAD	INFO	HOME	STATUS	HELP!																													
			Genomes <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Septoria musiva SO2202 v1.0</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Septoria populincola v1.0</div> Cluster Type <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">all</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">DMAT</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">HYBRID</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">NRPS</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">NRPS-Like</div>																																				
			<input type="button" value="Refresh"/>																																				
			<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Genome</th> <th>DMAT</th> <th>HYBRID</th> <th>NRPS</th> <th>NRPS-Like</th> <th>PKS</th> <th>PKS-Like</th> <th>TC</th> <th>Total</th> </tr> </thead> <tbody> <tr> <td>Septoria musiva SO2202 v1.0</td> <td>0</td> <td>2</td> <td>7</td> <td>8</td> <td>9</td> <td>2</td> <td>2</td> <td>30</td> </tr> <tr> <td>Septoria populincola v1.0</td> <td>0</td> <td>0</td> <td>8</td> <td>7</td> <td>9</td> <td>2</td> <td>3</td> <td>29</td> </tr> <tr> <td>Total</td> <td>0</td> <td>2</td> <td>15</td> <td>15</td> <td>18</td> <td>4</td> <td>5</td> <td>59</td> </tr> </tbody> </table>	Genome	DMAT	HYBRID	NRPS	NRPS-Like	PKS	PKS-Like	TC	Total	Septoria musiva SO2202 v1.0	0	2	7	8	9	2	2	30	Septoria populincola v1.0	0	0	8	7	9	2	3	29	Total	0	2	15	15	18	4	5	59
Genome	DMAT	HYBRID	NRPS	NRPS-Like	PKS	PKS-Like	TC	Total																															
Septoria musiva SO2202 v1.0	0	2	7	8	9	2	2	30																															
Septoria populincola v1.0	0	0	8	7	9	2	3	29																															
Total	0	2	15	15	18	4	5	59																															

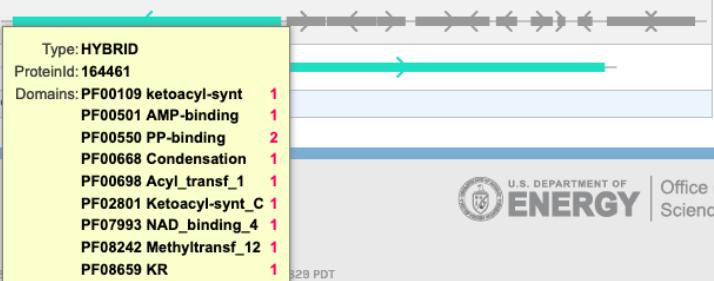
- 5) There is a total of 2 genes in the Hybrid column. Click on the number to show a graphical representation of the 2 *S. musiva* gene clusters. The ‘Size’ column displays each cluster’s length, and the ‘Genes’ column displays each cluster’s core PKS-NRPS gene (in color) and its accessory, decorator,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

Secondary Metabolism Clusters • Septoria musiva SO2202 v1.0

SEARCH	BLAST	BROWSE	ANNOTATIONS ▾	MCL CLUSTERS	SYNTENY	DOWNLOAD	INFO	HOME	STATUS	HELP!													
			Genomes <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Septoria musiva SO2202 v1.0</div> Cluster Type <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">all</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">DMAT</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">HYBRID</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">NRPS</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">NRPS-Like</div> Scale <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">✓ Per Cluster</div> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Per Cluster No Gaps</div> <div style="border: 1px solid #ccc; padding: 2px; background-color: #0070C0; color: white; margin-bottom: 5px;">Across All Clusters</div>																				
			<input type="button" value="Refresh"/>																				
			<p>Total 2 cluster(s) found. 1</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Cluster Id</th> <th>Cluster Type</th> <th>Scaffold</th> <th>Size (bp)</th> <th>Genes</th> </tr> </thead> <tbody> <tr> <td>Sepmu1.24</td> <td>HYBRID</td> <td>scaffold_6:1522811-1553990</td> <td>31,179</td> <td></td> </tr> <tr> <td>Sepmu1.25</td> <td>HYBRID</td> <td>scaffold_6:1977373-2004431</td> <td>27,058</td> <td></td> </tr> <tr> <td>Cluster Id</td> <td>Cluster Type</td> <td>Scaffold</td> <td>Size (bp)</td> <td>Genes</td> </tr> </tbody> </table>	Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes	Sepmu1.24	HYBRID	scaffold_6:1522811-1553990	31,179		Sepmu1.25	HYBRID	scaffold_6:1977373-2004431	27,058		Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes
Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes																			
Sepmu1.24	HYBRID	scaffold_6:1522811-1553990	31,179																				
Sepmu1.25	HYBRID	scaffold_6:1977373-2004431	27,058																				
Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes																			

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

Total 2 cluster(s) found. 1

Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes
Sepmu1.24	HYBRID	scaffold_6:1522811-1553990	31,179	
Sepmu1.25	HYBRID	scaffold_6:1977373-2004431	27,058	
Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes

Contact Us Cite Us Accessibility/Section 508
[Disclaimer](#) [Credits](#)

© 1997-2023 The Regents of the University of California.
Mycocosm Portal version:17.160 myco-web-3.jgi.lbl.gov Release Date:11-Apr-2023 129 PDT

U.S. DEPARTMENT OF ENERGY Office of Science

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

Total 2 cluster(s) found. 1

Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes
Sepmu1.24	HYBRID	scaffold_6:1522811-1553990	31,179	
Sepmu1.25	HYBRID	scaffold_6:1977373-2004431	27,058	
Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes

- 8) To get more detailed information about a gene, click on it directly. Click on the gene with the p450 domain to see its ‘protein page’. Examination of the protein page reveals that:
- The gene is expressed. The blue bars represent UTRs, which can be inferred only from transcriptomic data.
 - The protein has p450 Pfam and other annotations indicative of a cytochrome p450 monooxygenase.
 - The best Blast hit in nr is a cytochrome p450 monooxygenase from *Aspergillus nidulans*, which belongs to a different class of fungi (Eurotiomycetes) from *Septoria* (Dothideomycetes).

Best BLAST hit

SEARCH	BLAST	BROWSE	ANNOTATIONS	MCL CLUSTERS	SYNTENY	DOWNLOAD	INFO	HOME	STATUS	HELP!
Name:	estExt_Genewise1.C_6_t30338									
Protein ID:	87793									
Location:	scaffold_6:1535323-1537114									
Strand:	+									
Number of exons:	2									
Description:	gi 67902848 ref XP_681680.1 hypothetical protein AN8411.2 [Aspergillus nidulans FGSC A4]>gi 40747877 gb EAA67033.1 hypothetical protein AN8411.2 [Aspergillus nidulans FGSC A4]>gi 259484346 tp CBF80485.1 TPA: Cytochrome P450 monooxygenase (Eurofung) [Aspergillus nidulans FGSC A4] (model%: 91, hit%: 90, score: 1905, %id: 71) [Aspergillus nidulans FGSC A4]									
Best Hit:	total hits(shown)	683 (10)								

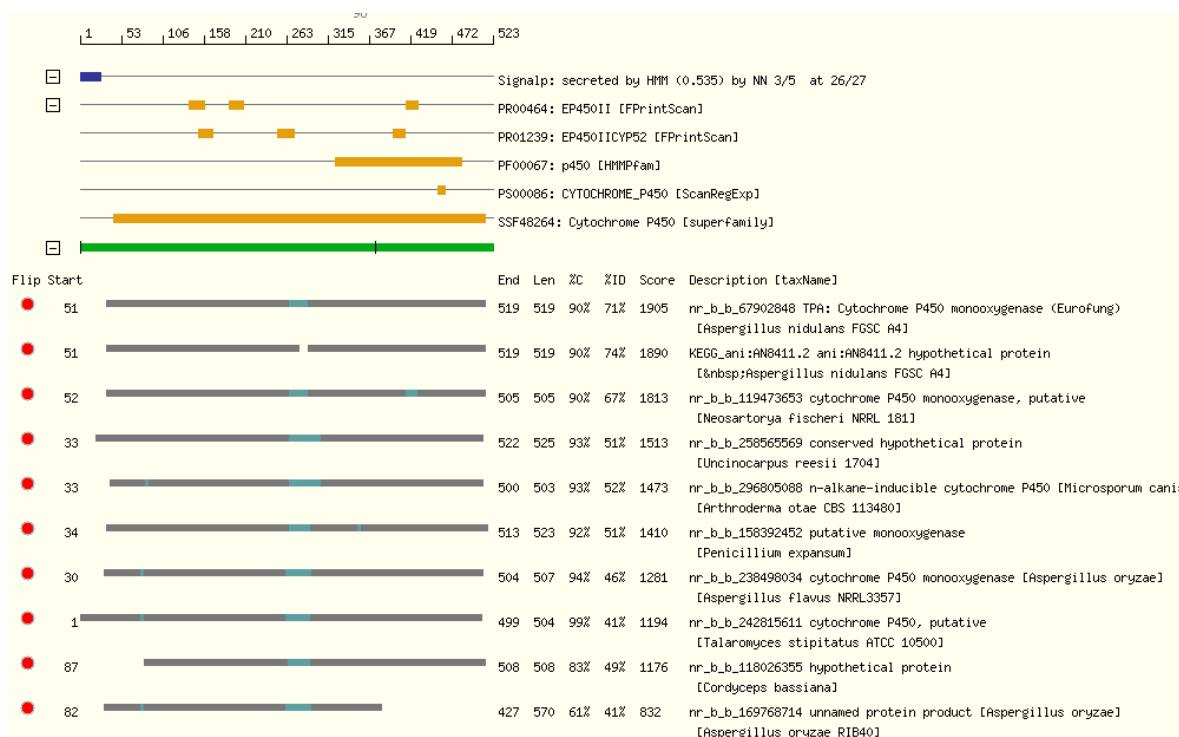
ASPECT	GO Id	GO Desc	Interpro Id	Interpro Desc
Molecular Function	<u>0016712</u>	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	IPR002974	Cytochrome P450, E-class, CYP52
	<u>0004497</u>	monooxygenase activity	IPR002402	Cytochrome P450, E-class, group II
	<u>0020037</u>	heme binding	IPR001128	Cytochrome P450
	<u>0005506</u>	iron ion binding	IPR002402	Cytochrome P450, E-class, group II
Biological Process	<u>0006118</u>	electron transport	IPR002974	Cytochrome P450, E-class, CYP52
			IPR001128	Cytochrome P450
KOG GROUP	KOG Id	KOG Class	KOG Desc	
Metabolism	KOG0158	Secondary metabolites biosynthesis, transport and catabolism	Cytochrome P450 CYP3/CYP5/CYP6/CYP9 subfamilies	

[View/modify manual annotation](#)
[View nucleotide and 3-frame translation](#) [To Genome Browser](#)
[NCBI blast](#) [Predicted number of transmembrane domains: 1](#)

Blue: UTRs
Red: CDS

InterPro annotations (For example, Pfam domains)

- 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

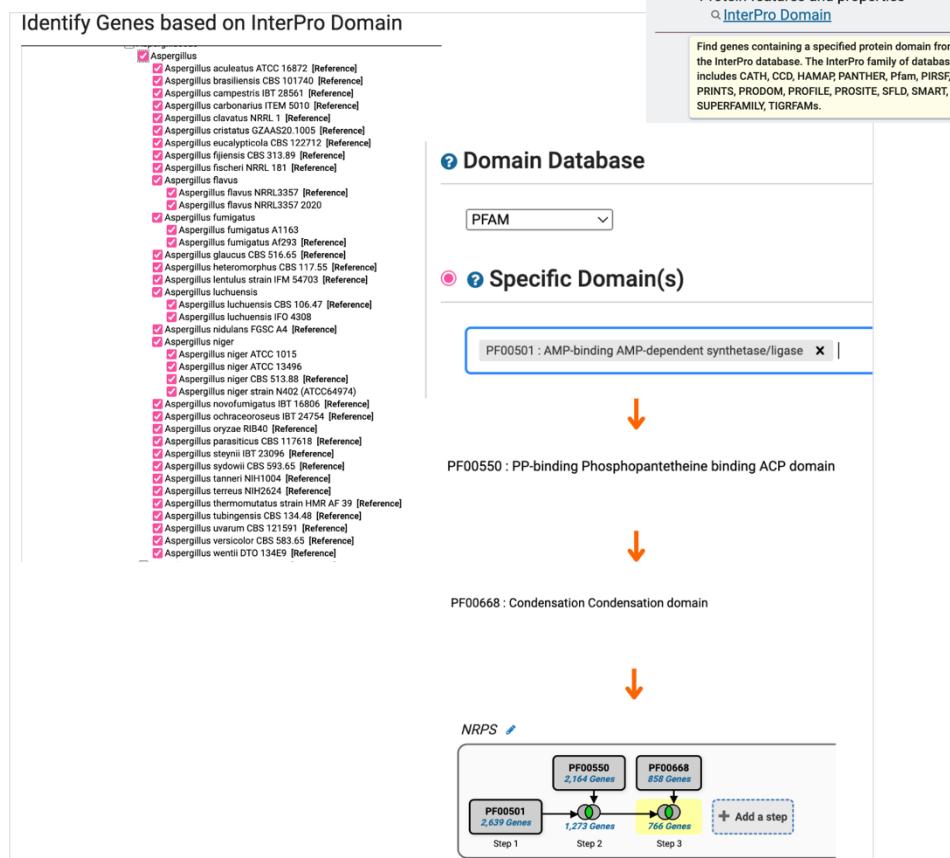
• Finding secondary metabolites and gene clusters

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)

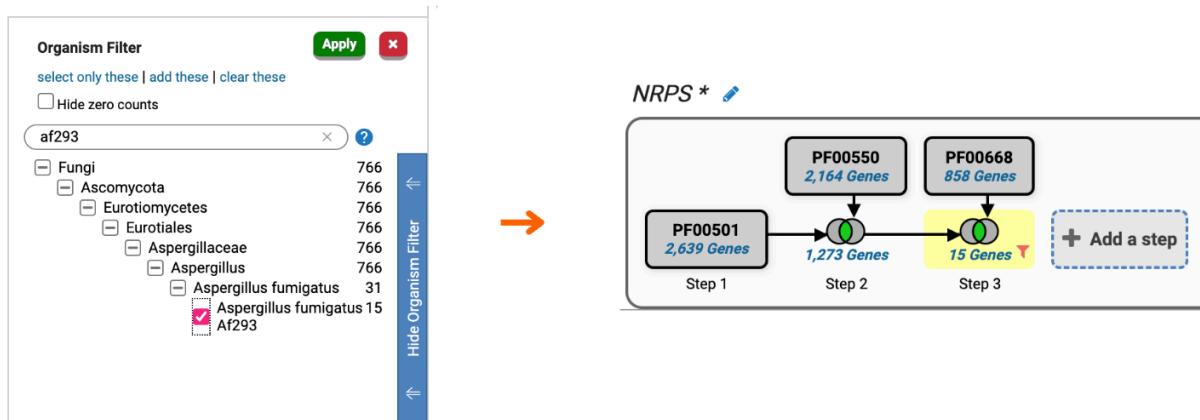


Strategy URL:

<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?

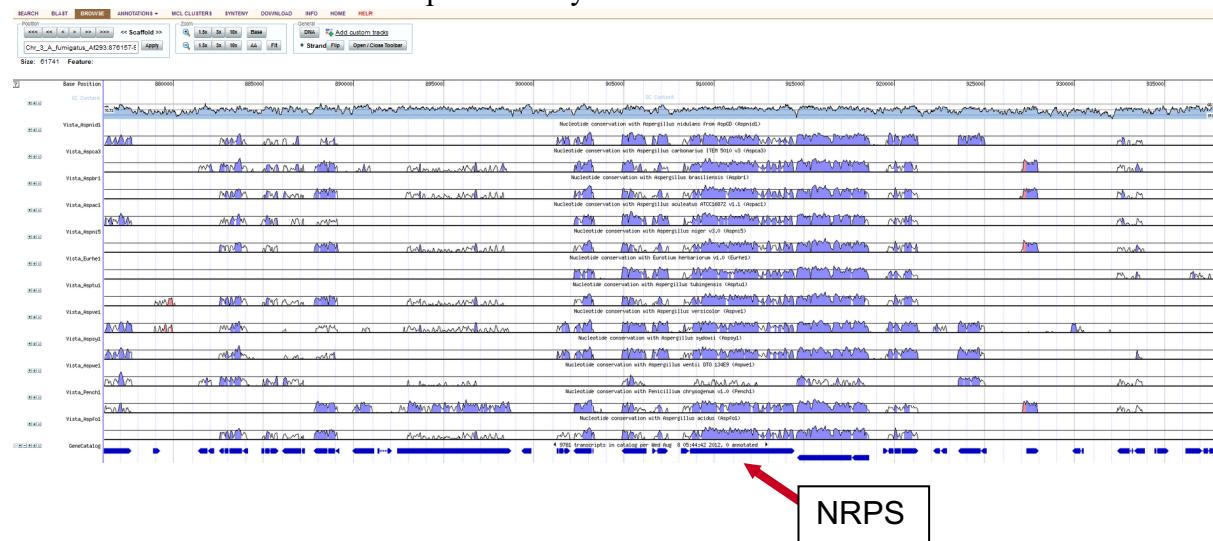
This screenshot shows the 'Annotations' tab in the MycoCosm interface, specifically the 'Secondary Metabolism Clusters' section for *Aspergillus fumigatus* Af293. The top navigation bar includes links for SEARCH, BLAST, BROWSE, ANNOTATIONS (selected), MCL CLUSTERS, SYNTENY, DOWNLOAD, INFO, HOME, QC, ADMIN, and HELP!. The main content area has tabs for Genomes, Cluster Type (set to NRPS), Scale (Per Cluster), and Clusters Per Page (set to 50). Below this, a table lists 9 clusters found:

Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes
Aspfu1.5	NRPS	Chr_3_A_fumigatus_Af293:876157-937897	61,740	Diagram showing gene orientation and size.
Aspfu1.7	NRPS	Chr_3_A_fumigatus_Af293:3423866-3446129	22,263	Diagram showing gene orientation and size.
Aspfu1.10	NRPS	Chr_3_A_fumigatus_Af293:4007787-4023468	15,681	Diagram showing gene orientation and size.
Aspfu1.15	NRPS	Chr_1_A_fumigatus_Af293:2655644-2694887	39,243	Diagram showing gene orientation and size.
Aspfu1.16	NRPS	Chr_1_A_fumigatus_Af293:4662924-4713331	50,407	Diagram showing gene orientation and size.
Aspfu1.18	NRPS	Chr_8_A_fumigatus_Af293:20854-49410	28,556	Diagram showing gene orientation and size.
Aspfu1.28	NRPS	Chr_5_A_fumigatus_Af293:3307809-3342792	34,983	Diagram showing gene orientation and size.
Aspfu1.31	NRPS	Chr_6_A_fumigatus_Af293:2334637-2372302	37,665	Diagram showing gene orientation and size.
Aspfu1.32	NRPS	Chr_6_A_fumigatus_Af293:3004871-3035305	30,434	Diagram showing gene orientation and size.

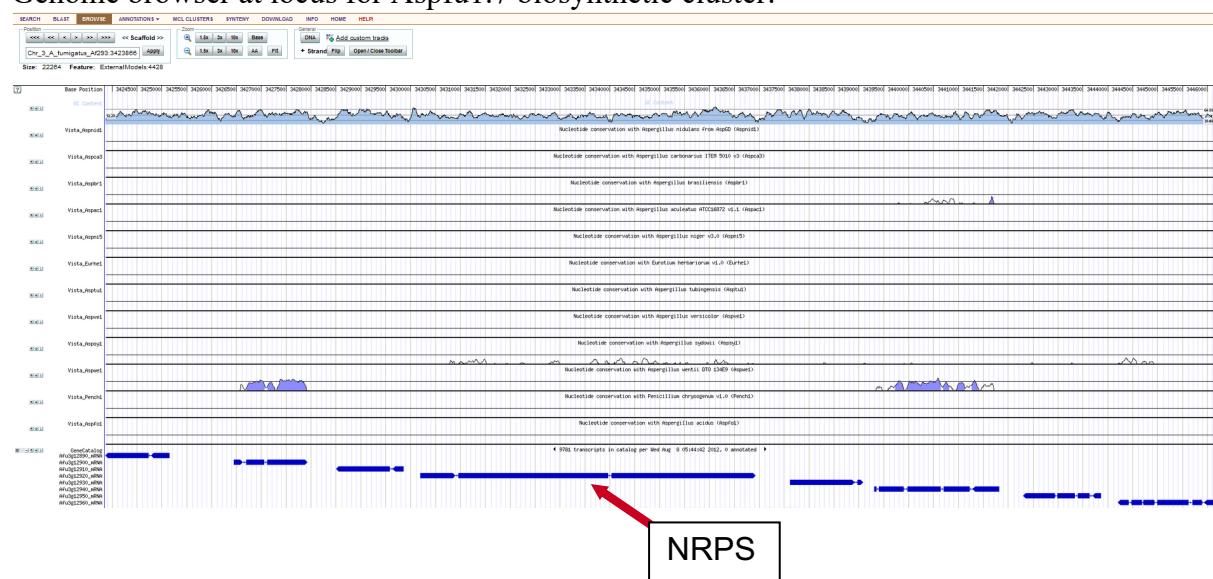
- What do you think may be causing the difference in the predicted gene number?
- This view on MycoCosm allows you to analyze backbone and auxiliary proteins across the entire predicted secondary metabolism cluster. How conserved are these secondary metabolite clusters across related Aspergilli? Click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