

Fungal Pathogen Genomics

2-6 June 2024



course information

WELLCOME CONNECTING SCIENCE

Fungal Pathogen Genomics

2-6 June 2024

**A course held at the Wellcome Sanger
Institute, Hinxton**

Course Information Handbook



Table of Contents

COURSE TIMETABLE	8
PROGRAMME OVERVIEW	1
SUMMARY	1
TARGET AUDIENCE	2
PROGRAMME	2
LEARNING OUTCOMES:	3
COURSE TIMETABLE WITH CONTENT	4
FUNGIDB SITE SEARCH	10
<i>Site search: text, term or gene id</i>	10
ADVANCED SEARCH STRATEGIES	13
<i>Combine with other Genes</i>	13
<i>Transform into related records</i>	13
<i>Use Genomic Colocation</i>	13
<i>Creating advanced search strategies in FungiDB</i>	14
A. <i>Deploy the “RNA-Seq Evidence” search to identify genes that are up-regulated when Aspergillus is exposed to human airway epithelial cells.</i>	14
B. <i>Deploy SNP Characteristic search to identify genes with non-synonymous SNPs using WGS data of clinical isolates.</i>	15
C. <i>Identify genes with immune reactivity</i>	16
SEARCH STRATEGIES IN SGD	17
1. <i>CREATE A LIST OF PROTEINS THAT ARE KNOWN SUBUNITS OF THE MITOCHONDRIAL RIBOSOME (MTR)</i> :.....	17
2. <i>FIND PROTEINS THAT GENETICALLY INTERACT WITH MTR PROTEINS:</i>	19
3. <i>FIND MTR INTERACTORS THAT ARE UNCHARACTERIZED</i> :.....	20
4. <i>ANALYZE THE RESULTS IN FUNGIDB</i>	21
EXERCISE: ENSEMBL FUNGI BIOMART	23
ADDITIONAL BIOMART EXERCISE 1 – EXPORT ORTHOLOGUES	29
<i>Exercise 1 answers</i>	29
ADDITIONAL BIOMART EXERCISE 2 – FINDING GENES BY PROTEIN DOMAIN	32
<i>Exercise 2 answers</i>	32
ADDITIONAL BIOMART EXERCISE 3 – CONVERT IDS	35
<i>Exercise 3 answers</i>	35
EXERCISE: EXPLORING HOST-PATHOGEN INTERACTIONS IN ENSEMBL FUNGI	38
ADDITIONAL HOST-PATHOGEN EXERCISE 1 – EXPLORING GO TERMS AND PHENOTYPES	45
<i>Exercise 1 answers</i> :	45
EXERCISE: ATTACHING TRACK HUBS TO ENSEMBL FUNGI	52
COMPILE A LIST OF GENES INVOLVED IN GLYCOLYSIS.	60
IMPORT YOUR GENE LIST INTO SPELL AND RUN A QUERY:	61

EXPLORING TRANSCRIPTOMICS & PROTEOMICS DATASETS IN FUNGIDB	64
TRANSCRIPTOMICS	64
A. The next block of exercises will be carried out in HostDB.org	65
• Identify host genes up-regulated by the SC5314 strain but not 101 at 1d of infection.	66
• Examine the results in HostDB:.....	67
B. The next block of exercises will be carried out in FungiDB.org.....	68
• Identify genes up-regulated in SC5314 but not 101 strain at 1d of infection.....	69
PROTEOMICS	71
• Identify proteins more abundant in EVs than whole cell lysate (WCL).....	71
• Identify genes that are upregulated in SC5314 during infection and present in the EVs samples when <i>Candida</i> is grown in biofilm condition.	72
References	72
EXPLORING GENE MODELS IN JBROWSE.....	73
EXERCISE: ENSEMBL FUNGI WHOLE-GENOME ALIGNMENTS	76
MYCOCOSM: COMPARATIVE ANALYSIS OF GENE FAMILIES	85
CAZy browser.....	85
Cluster page.....	88
References:	93
MYCOCOSM: SYNTENY TUTORIAL	94
Exercises:	98
Reference:.....	98
EXPLORING PROTEIN DOMAINS AND CLUSTERS ACROSS SPECIES IN ENSEMBL AND MYCOCOSM.....	99
ANSWERS	100
1. OBTAIN A LIST OF ALL GENES ENCODED IN THE MITOCHONDRIAL GENOME OF <i>C. GLABRATA</i> :.....	108
5. IMPORT THE <i>S. CEREVISIAE</i> SYNTHETIC LETHAL INTERACTION GENES INTO FUNGIDB FOR FURTHER ANALYSIS:.....	114
EXERCISE: ENSEMBL FUNGI GENE TREES AND HOMOLOGUES	115
ADDITIONAL EXERCISE 2: MUSHROOM GENES	128
ADDITIONAL EXERCISE 2 ANSWER: MUSHROOM GENES	128
FUNGiDB & ORTHOLMCL: ORTHOLOGY AND PHYLETIC PATTERNS	131
About OrthoMCL	131
EXAMINING ORTHOMCL OUTPUT ON GENE RECORD PAGES IN FUNGIDB	132
• Do all <i>Cryptococcus</i> species currently integrated in FungiDB contain this protein?	133
USING THE PHYLETIC PATTERN SEARCH IN ORTHOMCL	135
• Run the default search for EUKA>=5T AND hsap>=10.	135
• Using the “Phyletic pattern” search, identify how many eukaryotic protein groups do not contain orthologs from bacteria and archaea.	136
• Find all groups that contain orthologs from at least one species of Ascomycota fungi (1T) but not from bacteria, archaea or metazoan (0T).	137
• Revise your search to find groups that:	137
Useful information:.....	138
Combining searches in OrthoMCL.....	139
Exploring a specific OrthoMCL group - examining the cluster graph.....	141

What is Galaxy?.....	142
RNA-SEQ ANALYSIS, PART I.....	143
Group assignments.....	146
Guide to RNA-Seq histories and file organisation.....	147
• Organize samples with replicates into collections:	148
Running a workflow in Galaxy.....	149
• Deploy a pre-configured workflow.....	149
• Configure an RNA-Seq workflow.....	150
How to work with Galaxy editor (optional)	153
VARIANT CALLING ANALYSIS, PART I.....	156
USING SGD GO SLIM MAPPER AND INTERACTION DATA TO PREDICT GENE FUNCTION.....	160
• Highlight SELECT ALL Terms from Yeast GO-Slim: Process.....	161
GO ENRICHMENT, PHENOTYPE DATA AT CGD.....	163
FungiDB: Performing GO Enrichment analysis.....	169
- Odds ratio—Determines whether the odds of the GO term appearing in the list of interest are the same as those for the background list.....	170
Optional exercise. Creating queries across FungiDB and SGD.....	172
1. Navigate to jhhlp_004726 in FungiDB and examine available records.....	172
2. Export orthologs of this gene and carry over <i>S. cerevisiae</i> gene IDs into SGD.....	172
3. Find known genetic interactions for MKC7.....	173
Additional resources:.....	178
RNA SEQUENCE DATA ANALYSIS VIA GALAXY, PART 2	179
Learning objectives:.....	179
• Sharing workflow histories with others.	179
• Importing workflow histories and output files into your own Galaxy workspace.....	180
• Explore the FastQC results.	182
Explore the differential expression results.....	183
B. DESeq2 results file—This table contains the actual differential expression results. While these can be viewed within Galaxy, it will be more beneficial to download this table and open it in Excel so you can sort the results.	183
• Download DESeq2 results (tabular format) by clicking the floppy disk save icon.....	183
• Explore the results in Excel.....	184
EXPORTING DATA TO VEuPATHDB/FUNGI DB	186
• Create a Dataset List with “htseq-count on data” files.....	186
2. Search for htseq-count files.	186
3. Select “htseq-count on data” files.....	187
4. “Build dataset list”.....	187
5. Rename each htseq-count sample, give the collection a name and create a dataset list.....	188
• Create a Dataset List with “BAM to BigWig on data” files.	188
• Use the HTSeqCountToTPM tool to convert counts to TPM.....	189
• Export TPM counts and BigWig data to VEuPathDB/FungiDB workspace.....	190
VARIANT CALLING ANALYSIS, PART 2: ANALYZING RESULTS (GROUP EXERCISE)	192
• Share workflow histories with others.....	192
• Importing workflow histories and output files into your own Galaxy workspace.....	192
• Examine your results.....	195
• Examine sequence quality based on FastQC quality scores.	196

• Examine <i>SnpEff</i> summaries (HTML).....	196
Summary statistics for variant types	197
Statistics for the variant effects and impacts:	198
Examining SNP information.....	199
Filtering VCF file data.	200
• Extract the filtered VCF file (<i>SnpSift</i> output) and convert it into an Excel document.....	200
• Manipulate Excel file to display SNP info in columns.....	202
• Analyze your data in Venny.	203
Viewing the VCF file results in the JBrowse genome browser.	205
• Run the newly created workflow to generate a compressed vcf and index files.....	207
• Download compressed vcf (vcf_bgzip) and index (tabix) files and view them in JBrowse.....	207
SGD VARIANT VIEWER	209
S288C vs. SIGMA1278B: CELL WALL	209
VARIANT VIEWER: SEQUENCE TAB.....	211
FUNGIDB: SNPs AND POPULATION GENETICS	213
Read Frequency Threshold:.....	213
Minor allele frequency:.....	214
Percent isolates with a base call:.....	214
• Identify genes with at least 1 non-synonymous SNP.	215
• identify putative nuclear effectors based on the presence of both a secretion signal and the DNA-binding domains IPR007219 or IPR009071.....	216
B. IDENTIFY SNPs BASED ON DIFFERENCES BETWEEN TWO GROUPS OF ISOLATES.....	219
• Identify SNPs between two groups of <i>C. posadasii</i> str. <i>Silveira</i> isolates.....	219
• Change the stringency of your search to major allele frequency $\geq 90\%$	220
C. Copy number variation & ploidy searches.....	224
C.1. Copy Number/Ploidy search (Genomic Sequences)	224
• Explore segmental aneuploidy in JBrowse.	225
C.2. Copy Number search (Genes) Using Gene Searches	227
EXERCISE: EXPLORING VARIANTS IN ENSEMBL FUNGI	231
Answers	235
EXERCISE: THE ENSEMBL FUNGI VARIANT EFFECT PREDICTOR (VEP)	237
ADDITIONAL EXERCISE: THE ENSEMBL FUNGI VARIANT EFFECT PREDICTOR (VEP).....	236
ANSWER.....	236
MYCOCOSM: KEGG BROWSER	239
Reference:.....	243
MYCOCOSM: SECONDARY METABOLISM CLUSTERS BROWSER	244
References:	248
FUNGIDB: SECONDARY METABOLITES AND CLUSTERS.....	249
• FINDING SECONDARY METABOLITES AND GENE CLUSTERS.....	249

COURSE ORGANISATION AND DEVELOPMENT STAFF

Wellcome Connecting Science Team

Dr Michelle Bishop	Associate Director
Dr Alice Matimba	Head of Training and Global Capacity
Mr Martin Aslett	Informatics Manager
Ms Vaishnavi Vikas Gangadhar	Informatics Technical Officer
Mrs Karon Chappell	Events Organiser

URL: <https://coursesandconferences.wellcomeconnectingscience.org/our-events/courses/>

Email: globaltraining@wellcomeconnectingscience.org

Course Organises & Support

Evelina Bassenko	FungiDB/VEuPathDB (University of Liverpool)
David Roos	FungiDB/VEuPathDB (University of Pennsylvania)
Nishadi De Silva	Ensembl Fungi (EBI)

Course Instructors

Jodi Lew-Smith	CGD/SGD Candida Genome Database, Stanford University
Steven Ahrendt	JGI/Mycocosm
Manuel Carbajo	Ensembl
Kathryn Crouch	FungiDB/VEuPathDB (University of Glasgow)
Stuart Brown	VEuPathDB
Louisee Mirabueno	Ensembl

Speakers

Jane Usher	Exeter University
Jason Rudd	Rothamsted Research

COPYRIGHT LICENSING

This manual is licensed under **CC BY-NC-SA**



You are free to:

- **Share** — copy and redistribute the material in any medium or format
- **Adapt** — remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial** — You may not use the material for [commercial purposes](#).
- **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.
- **No additional restrictions** — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

COURSE TIMETABLE

		Fungal Pathogen Genomics (2-7 June 2024)				
Time (BST)	Sunday 2 June	Monday 3 June	Tuesday 4 June	Wednesday 5 June	Thursday 6 June	Friday 7 June
07:30		BREAKFAST	BREAKFAST	BREAKFAST	BREAKFAST	BREAKFAST
09:00		Evaluating gene model evidence	Enrichment analysis	Galaxy: SNP analysis (lecture)	Functional analysis:	
09:30				SNP analysis	Pathways and metabolites	Group projects (cont.)
10:00						
10:30 - 11:00		TEA BREAK	TEA BREAK	TEA BREAK	TEA BREAK	TEA BREAK
11:00	REGISTRATION		Enrichment analysis, cont.			
11:30	WCS Orientation	Comparative Genomics & Orthology			Manual curation in Apollo	Group presentations
12:00	Instructor & Database introductions				Group projects	(12 min per team)
12:30			Galaxy: RNA-Seq Result analysis (lecture)	Introduction to group projects		
13:00-14:00	LUNCH	LUNCH	LUNCH	LUNCH	LUNCH	LUNCH & DEPARTURE
14:00						
14:30	Introduction to Database Queries	Comparative Genomics & Orthology (cont.)	RNA-Seq Analysis Part 2	Course photo & the Sanger Tour	Group projects (cont.)	
15:00						
15:30 - 16:00	TEA BREAK	TEA BREAK	TEA BREAK	TEA BREAK	TEA BREAK	
16:00	Introduction to Database Queries, cont.	Galaxy NGS data analysis				
16:30		Introductory lecture				
17:00			RNA-Seq Analysis Part 2	SNPs & variants		
17:30	Transcriptomics & Proteomics				Group projects (cont.)	
18:00		Deploy RNA-Seq & SNP workflows	Research Seminar	Research Seminar		
18:30			Dr Jane Usher	Dr Jason Rudd		
19:00-20:00	WELCOME RECEPTION & DINNER	DINNER	DINNER	DINNER	DINNER	
20:00-21:00	Flash participant presentations	Adjourn (bar open)	Adjourn (bar open)	Adjourn (bar open)	Adjourn (bar open)	
	Adjourn (bar open)					

Programme Overview

Summary

The kingdom of Fungi encompasses a diverse range of organisms adapted to various environmental niches, playing crucial roles in ecosystems, human/animal/plant health, and global food security. Species like *Fusarium*, *Pyricularia*, *Ustilago*, *Puccinia*, and *Zymoseptoria*, etc. threaten agricultural ecosystems and food security worldwide. Fungal pathogens such as *Aspergillus*, *Candida*, *Mucor*, *Cryptococcus*, *Histoplasma*, *Coccidioides*, *Batrachochytrium*, and others are of great concern for medical and veterinary professionals due to their potential to cause allergies, illnesses, and life-threatening infections. Furthermore, fungi serve as essential model systems in basic and applied research and play pivotal roles in biotechnology, food production, biomedical and pharmaceutical research and the biofuel industry.

High-throughput ‘omics’ data technologies empower scientists to conduct extensive analyses of the genomes, transcriptomes, proteomes, genetic variance data of a wide array of fungal and oomycete organisms. These analyses are essential for inquiries regarding pathogenicity, host-pathogen interactions, and the discovery of novel drug targets. To enhance accessibility and re-use of data, and facilitate analysis of different types of data, several web-based bioinformatic resources have been developed.

This week-long course represents a collaborative teaching effort involving the following resources dedicated to supporting research on fungal and oomycete species:

- FungiDB/VEuPathDB
- Ensembl Fungi/EBI
- SGD/CGD
- MycoCosm/JGI

The Fungal Pathogen Genomics course offers hands-on training on how to use unique, web-based tools provided by individual resources and apply them to both public and private datasets. Participants will learn to formulate testable hypotheses and explore genomes, functional omics datasets (transcriptomics, SNP, proteomics, etc.) and other data types across multiple databases.

Target audience

This course is aimed at graduate students, postdocs, clinical scientists, clinicians/healthcare professionals or lab heads working on fungal pathogens.

Programme

Daily activities will encompass both individual and group training exercises, complemented by supplementary lectures focusing on bioinformatics techniques and tools utilized by various databases. For example, the learning materials will address the following topics:

- Navigating gene record pages and genome browsers.
- Identification of orthologs and orthology-based inference.
- Comparative genomics, gene trees, whole-genome alignment.
- Creating in-silico experiments in FungiDB by employing search queries and public omics datasets and performing enrichment analysis on search results.
- Omics data analysis and visualization in VEuPathDB Galaxy, including RNA-Seq and Variant calling analysis (with Ensembl Variant Effect Predictor tool).
- Identifying genetic interactions, virulence genes, secondary metabolites.
- Introduction to manual genome curation using Apollo, a web-based platform for structural and functional genome annotation.

The programme also features talks by globally renowned guest speakers and provides occasions for engaging in scientific discussions and networking within a friendly environment.

Learning outcomes:

After attending this course, participants will be able to:

- Navigate each database and answer research questions by creating custom queries across multiple bioinformatics resources.
- Utilize built-in web-based bioinformatics tools to mine omics data, including:
 - finding genes
 - identifying orthologs in other species via orthologous transformations,
 - determining trends through GO and metabolic pathway enrichment analysis,
 - visualizing genomics, proteomics, and transcriptomics data,
- Perform RNA-Seq and Variant calling analysis in VEuPathDB Galaxy, exporting and sharing analysis results, and more.
- Gain an understanding of the practical advantages and limitations of the tools used in bioinformatics, ensuring that the skills acquired are pertinent to their research and future endeavors.
- Develop skills and knowledge necessary to actively contribute to community genome curation and annotation efforts.
- Apply their newfound skills in web-based bioinformatic resources in their research projects.

Course timetable with content

Sunday, 2nd of June

Time (BST)	Content
10:00 – 10:30	Registration
10:30 – 12:30	Welcome, and Instructor/Database introductions Wellcome Genome Science MycoCosm/JGI FungiDB SGD/CD Ensembl Fungi
12:30 –	1. Introduction to database queries <ul style="list-style-type: none"> • FungiDB site search & advanced search strategies • SGD YeastMine • Ensembl Fungi – BioMart • Ensembl Fungi Molecular Interactions
13:00 – 14:00	Lunch
– 15:30	Introduction to database queries, Cont/...
15:30 –	2. Transcriptomics & Proteomics <ul style="list-style-type: none"> • Ensembl Fungi Track Hubs • SGD Expression tools - SPELL • FungiDB Transcriptomic & Proteomic analysis
16:00 – 16:30	Tea break
– 19:00	Transcriptomics & Proteomics, Cont/...
19:00– 20:00	Welcome reception & Dinner
20:00 – 21:00	Participant presentations (2 min flash talks)

Monday, 3rd of June

Time (BST)	Content
09:00	Breakfast
09:00 – 10:30	3. Evaluating gene model evidence
10:30 – 11:00	Tea break
11:00 – 13:00	4. Comparative Genomics & Orthology and Evolutionary analysis & cross-species inference <ul style="list-style-type: none"> • Ensembl Fungi – WGA • MycoCosm CAZy enzymes • MycoCosm Synteny • Exploring protein domains and clusters across Ensembl & MycoCosm • SGD predicting fungal biology • Ensembl Fungi Evolutionary analysis (gene trees) • FungiDB & OrthoMCL: Orthology and Phylogenetic Patterns
13:00 – 14:00	Lunch
14:00 –	Comparative Genomics & Orthology, Cont/...
15:30– 16:00	Tea break
16:00 – 19:00	5. NGS data analysis I <p>Background & Intro to VEuPathDB Galaxy</p> <ul style="list-style-type: none"> • Deploying RNA-Seq • Deploying SNP workflows
19:00– 20:00	Dinner

Tuesday, 4th of June

Time (BST)	Content
– 09:00	Breakfast
09:00 –	<p>6. Enrichment analysis</p> <ul style="list-style-type: none"> • SGD GO Slim mapper • CGD GO Term finder • FungiDB GO enrichment
10:30 – 11:00	Tea break
– 12:30	Enrichment analysis, Cont/...
12:30 –	<p>7. NGS Data analysis II - RNA-Seq analysis Part 2</p> <p>Exporting and analysing data from Galaxy</p>
13:00 – 14:00	Lunch
15:00 –	NGS Data analysis, Cont/...
15:30– 16:00	Tea break
– 18:00	NGS Data analysis, Cont/...
18:00 – 19:00	Research Seminar
19:00– 20:00	Dinner

Wednesday 5th of June

Time (BST)	Content
– 09:00	Breakfast
09:00 –	8. NGS Data analysis II - SNP analysis Part 2 Exporting and analysing data from Galaxy
10:30 – 11:00	Tea break
– 12:00	NGS Data analysis, Cont/...
12:15 – 12:30	Introduction to group projects.
13:00 – 14:00	Lunch
14:00 – 15:00	Sanger Tour
15:30– 16:00	Tea break
16:00 – 18:00	3. SNPs & Variants <ul style="list-style-type: none"> • SGD Variant viewer • FungiDB SNP analysis & CNVs • Exploring variant in Ensembl Fungi
18:00 – 19:00	Research Seminar
19:00– 20:00	Dinner

Thursday 6th of June

Time (BST)	Content
– 09:00	Breakfast
09:00 –	8. Functional analysis: Pathways & metabolites <ul style="list-style-type: none"> • MycoCosm KEGG Browser & Secondary metabolism clusters • FungiDB pathways & metabolites
10:30 – 11:00	Tea break
11:30 – 12:00	Manual curation in Apollo
12:00 –	Group Projects
13:00 – 14:00	Lunch
14:00 – 15:00	Group Projects, Cont/...
15:30– 16:00	Tea break
– 19:00	Group Projects, Cont/...
19:00– 20:00	Dinner

Friday 7th of June

Time (BST)	Content
– 09:00	Breakfast
09:00 –	Group Projects, Cont/...
10:30 – 11:00	Tea break
11:00 – 13:00	Group Projects Presentations (12 min per team max)
13:00 – 14:00	Lunch & Departure



FungiDB Site Search

Learning objectives:

- Use keywords in site search.
- Filter site search results by categories, organisms, and other categories.
- Export results to a search strategy.
- Find genes with gene IDs.

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

Site search: text, term or gene id.

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



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

All results matching kinase

1 - 20 of 394,396

Export as a Search Strategy

Filter results

- Genes **Genes** (highlighted)
- Population biology
- Protein-protein interactions
- Metabolism
- Mutations pathways
- Components
- Data sources
- Data sets
- GeneSets
- AltMut News

Data set 1: Analysis of the protein kinase A-regulated proteome of *Cryptococcus neoformans*

Gene - **CGL_00200W** MAP kinase kinase kinase, MAP kinase kinase kinase, positive

Gene type: protein-encoding gene

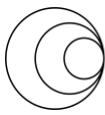
Organism: *Cryptococcus neoformans* ATCC 30542

Gene - **ANKE3_00041W** MAP kinase kinase kinase

Gene type: protein-encoding gene

Organism: *Cryptococcus neoformans* ATCC 30542

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



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

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

The screenshots illustrate the step-by-step process of applying filters. The first two steps focus on filtering gene fields, while the last two steps focus on filtering organisms. The 'Save' button is highlighted in the second step, indicating where to apply the selected filter.

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

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

The screenshot shows the final search results page with a clear call-to-action button for exporting the data.

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

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

All results matching *

1 - 10 of 4201 total

Compound	CHEBI19889: Vinblastine
Compound	CHEBI19881: Vinorelbine
Compound	CHEBI19882: Vinagrel
Compound	CHEBI19883: Fluorouracil sulfate
Definition	an antineoplastic cytostatic drug resulting from the reaction of fluorouracil with sulfuric acid
Compound	CHEBI198147: Inhibition; acidic
Definition	a monosubstituted and/or comprising 1,3-disubstituted 1,3-disubstituted by anisole or anisidine and methyl groups at positions 1, 3, and 7 respectively
Compound	CHEBI198144: Inhibition
Compound	CHEBI198119: Inhibition
Definition	an inhibitory pharmacological effect which the target molecule group is substituted by a methanesulfonate group and an additional non-sulfonate group is present in the molecule
Compound	CHEBI198122: Inhibition
Compound	CHEBI198117: Inhibition
Definition	a inhibitor that is capable of causing 1,2,4,5,6,7-hexamerine has 6 configuration at positions 1, 2, 3, 5 and 6
Compound	CHEBI198118: Inhibition
Definition	a conjugate product that is 260-methoxysubstitution is case 1,2-disubstituted structure for a family of glucosaminoglycans in position 1
Compound	CHEBI198119: Inhibition
Compound	CHEBI198123: Inhibition
Compound	CHEBI198122: Inhibition
Definition	a kinase-based biological agent used for the treatment of neoplastic conditions, known chemically as genistein, with various biologic influences toward the stimulation of assessments and function due to its inhibition of tyrosine kinase activity on tyrosine phosphorylation
Compound	CHEBI198141: Inhibition
Definition	a derivative that is a member of 100-methoxycongeners, selective and tissue specific lig of cyclin-dependent kinases 1, 5 and 7 respectively

Export as a Search Strategy
Download or view your results

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

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

Genes matching Afu2g13260

1 - 1 of 1

Filter results	Gene name or symbol: Afu2g13260
Filter Gene Fields	External links: 1; Name ID: 1; Names, IDs, and aliases: 1; Local comments: 1
Filter organisms	Fungi: Aspergillus fumigatus ATCC 46

Afu2g13260: Developmental regulator medA, putative

Gene name or symbol: medA
 Gene type: protein coding gene
 Organism: Aspergillus fumigatus ATCC 46

Fields matched: External links; Gene ID; Name; ID; and aliases; User comments

Afu2g13260: Developmental regulator medA, putative

Gene name or symbol: medA
 Gene type: protein coding gene
 Organism: Aspergillus fumigatus ATCC 46

Fields matched: External links; Gene ID; Name; ID; and aliases; User comments

Export as a Search Strategy
Download or view your results

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

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

Search strategy links:

kinase - <https://fungidb.org/fungidb/app/workspace/strategies/import/9c47e36cfaf7790f6>

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

Afu2g13260 -

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

Advanced Search Strategies

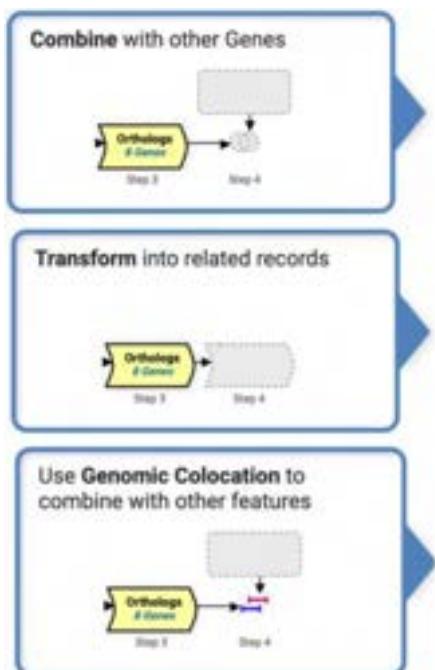
Learning objectives

- Deploy search for different types of data and create an advanced search strategy in FungiDB.

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

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

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



Combine with other Genes

Compare results that are gene lists

Transform into related records

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

Use Genomic Colocation

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

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

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



The screenshot shows the revision interface for a search strategy:

- Revise as a boolean operation:**
 - 1 INTERSECT 2
 - 1 UNION 2
 - 1 MINUS 2
 - 2 MINUS 1
- Revise as a span operation:**
 - 1 RELATIVE TO 2, using genomic collocation
- Ignore one of the inputs:**
 - IGNORE 2
 - IGNORE 1

A **Revise** button is located at the bottom right.



Creating advanced search strategies in FungiDB.

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

- A. Are up-regulated when *Aspergillus* is exposed to human airway epithelial cells,
- B. Have non-synonymous mutations identified by whole genome sequencing (WGS) of clinical isolates,
- C. Are known to be immune-reactive.

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

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

1. Select the search from the “Search for...” panel (shown below) or the “Searches” menu at the top of the page.
 Tip: Utilize the filter box to promptly retrieve relevant search results.
2. Identify the ‘Response to caspofungin (Valero et al. 2020)’ dataset and click on the DE (Differential expression) button.
3. Set up search parameters:
 - i. Reference sample: WT_CT
 - ii. Comparator sample: WT_CSP
 - iii. Direction: up-regulated
 - iv. Fold: 2Leave other parameters at default.
4. Click on the “Get Answer” button.

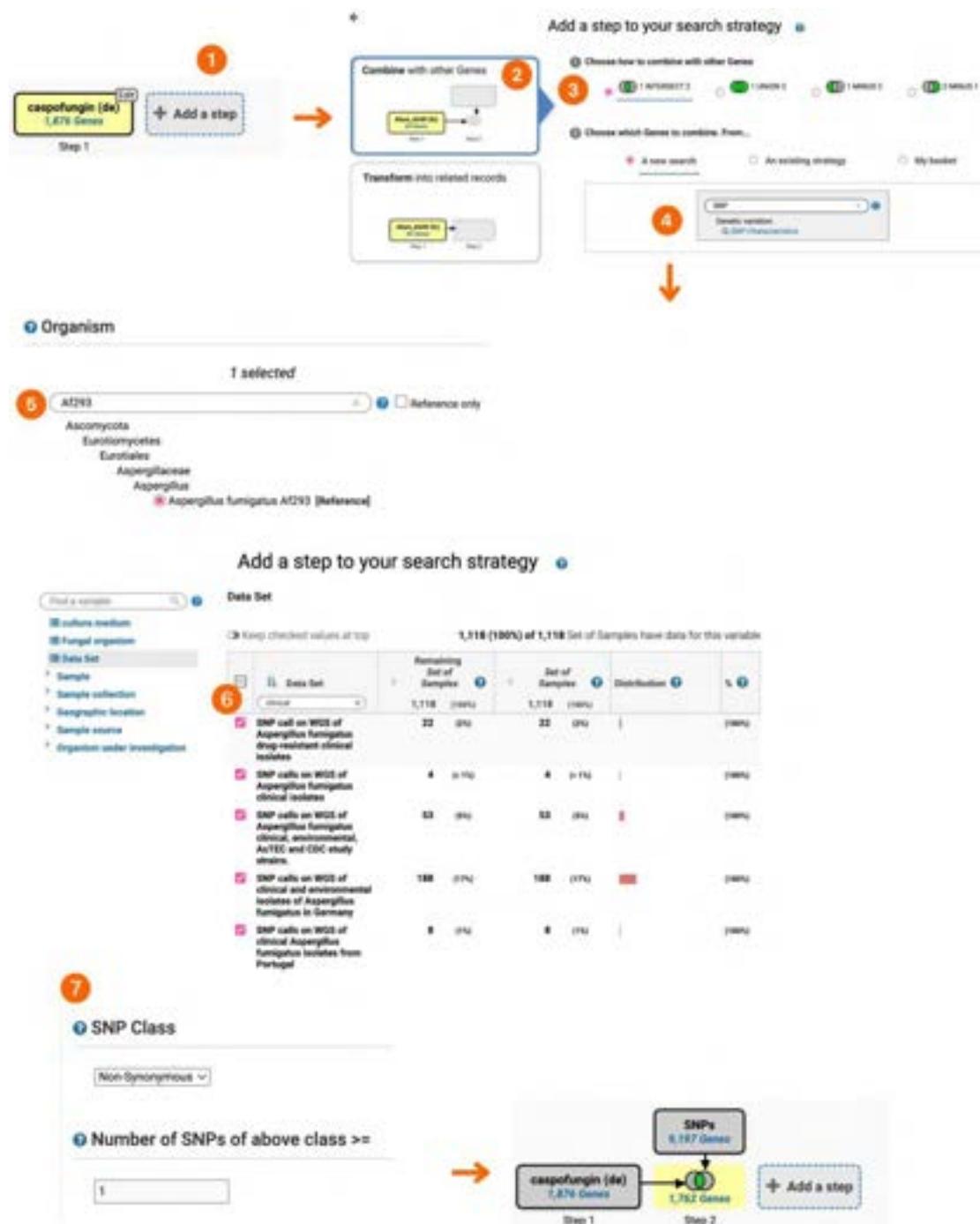
The screenshot shows the FungiDB search interface with the following steps highlighted:

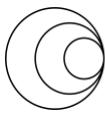
1. In the "Search for..." panel, the "Data Sources" section is expanded, showing options like "Gene Model Annotations", "Protein Model Annotations", "Transcriptome", "Metabolic Evidence", and "RNA-seq Evidence". The "RNA-seq Evidence" option is selected.
2. In the "Identify Genes based on RNA-Seq Evidence" search interface, the "Experiment" section shows "Response to caspofungin - Sensor" selected. A red arrow points from the "Search for..." panel to this section.
3. In the "Reference Sample" section, "WT_CT" is selected. A red circle with the number 3 is placed next to this selection.
4. In the "Comparator Sample" section, "WT_CSP" is selected. A red circle with the number 4 is placed next to this selection.
5. In the "Direction" section, "up-regulated" is selected. A red circle with the number 5 is placed next to this selection.
6. In the "fold difference >=" section, the value "2" is entered. A red circle with the number 6 is placed next to this input field.
7. In the "adjusted P value less than or equal to" section, the value "0.01" is entered. A red circle with the number 7 is placed next to this input field.
8. At the bottom right, the "Get Answer" button is highlighted with a red circle and the number 8. An orange arrow points from the "adjusted P value" section to this button. To the right of the button is a yellow box containing the text "caspofungin (de) 1,619 genes". Below the "Get Answer" button is a "Step 1" indicator.
9. To the right of the yellow box, there is a "+ Add a step" button.



B. Deploy SNP Characteristic search to identify genes with non-synonymous SNPs using WGS data of clinical isolates.

1. Within the search strategy, click on the “Add a step” button.
2. Make sure to use the “Combine with other Genes” option.
3. Select the “1 INTERSECT 2” Boolean operator (if not selected by default already)
4. Use filter to identify the new search to deploy – “SNP Characteristics”. Click on the “SNP Characteristics” link in blue to deploy the search.
5. Filter for ‘Af293’ and select “Aspergillus fumigatus Af293”
6. Filter datasets for “clinical” and select all 5 datasets
7. Specify specific SNP characteristics (SNP Class = Non-Synonymous; Number of SNPs of above class >=1) and run the analysis.

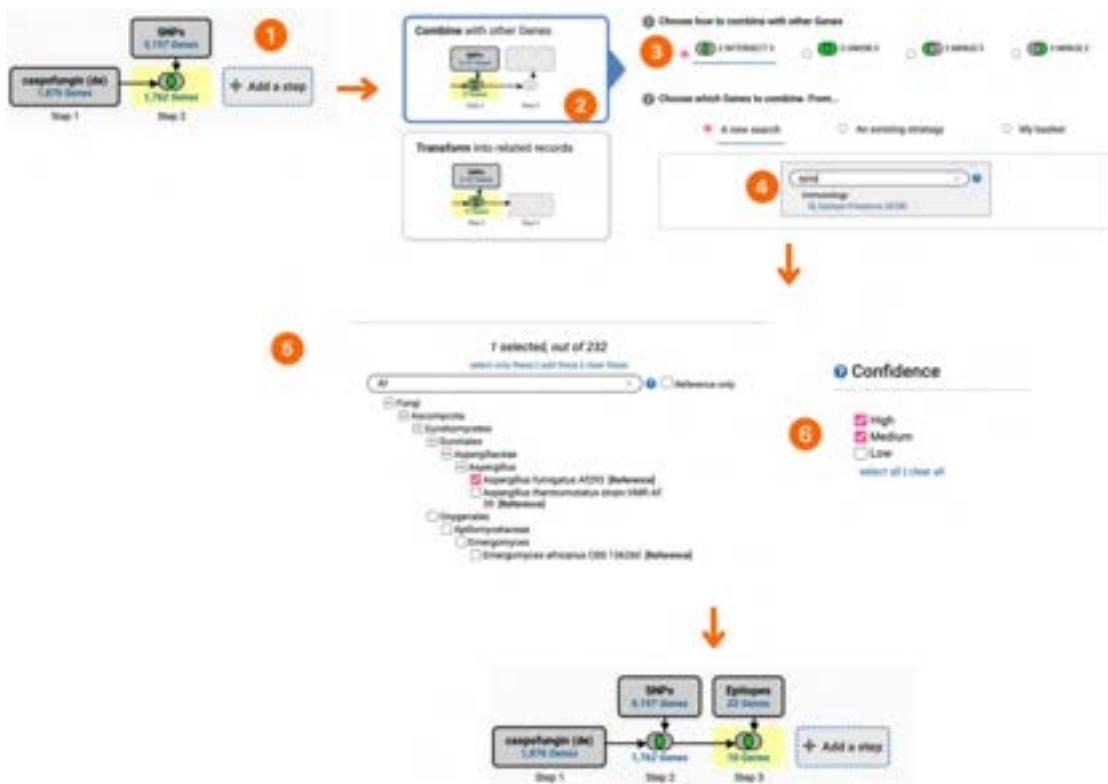




C. Identify genes with immune reactivity.

Epitopes are recognized by the immune system and can be used for vaccine development. Identify genes that have annotated epitope records.

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



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

Search strategy link:

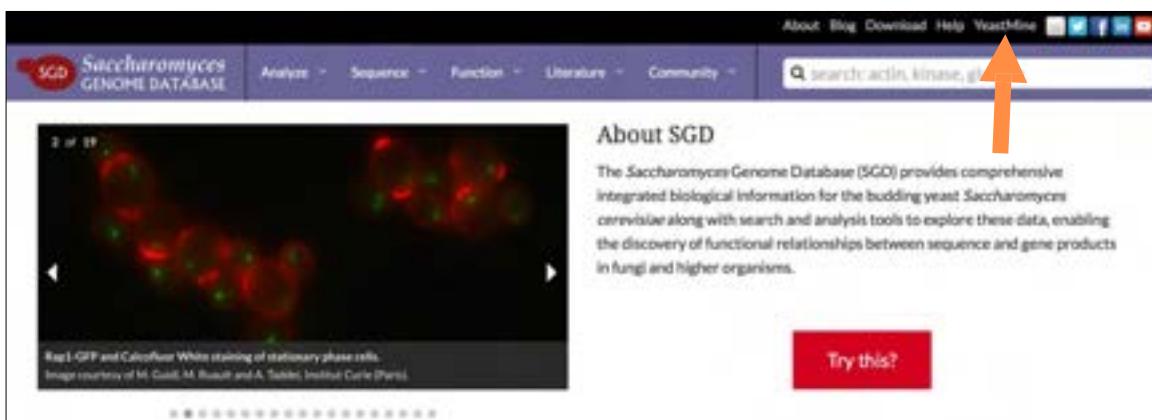
<https://fungidb.org/fungidb/app/workspace/strategies/import/0e675d26ca9287cb>

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?”

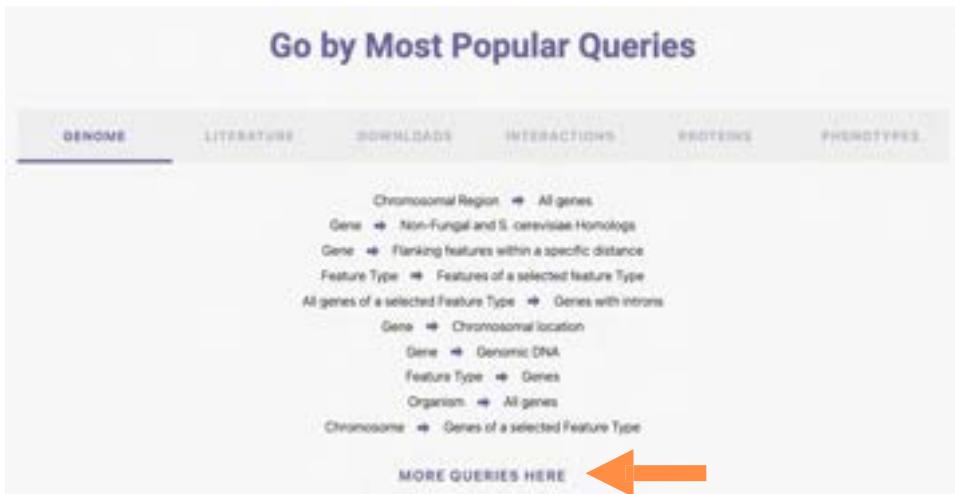
In this exercise, we will use YeastMine to search for as-yet undiscovered mitochondrial ribosomal proteins in yeast.

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



The screenshot shows the SGD home page with a navigation bar at the top. The "YeastMine" link is highlighted with an orange arrow pointing to it. Below the navigation bar is a search bar containing the query "search:actin, kinase, g".

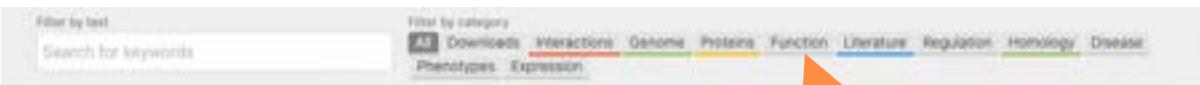
- Create a list of proteins that are known subunits of the mitochondrial ribosome (MTR):



The screenshot shows the "Go by Most Popular Queries" section of YeastMine. It displays a list of popular search queries, each with a corresponding icon and a brief description. An orange arrow points to the "MORE QUERIES HERE" button at the bottom of the list.

Query	Description
Chromosomal Region	All genes
Gene	Non-Fungal and <i>S. cerevisiae</i> Homologs
Gene	Ranking features within a specific distance
Feature Type	Features of a selected feature type
All genes of a selected Feature Type	Genes with introns
Gene	Chromosomal location
Gene	Genomic DNA
Feature Type	Genes
Organism	All genes
Chromosome	Genes of a selected Feature Type

- Click on "More queries here"



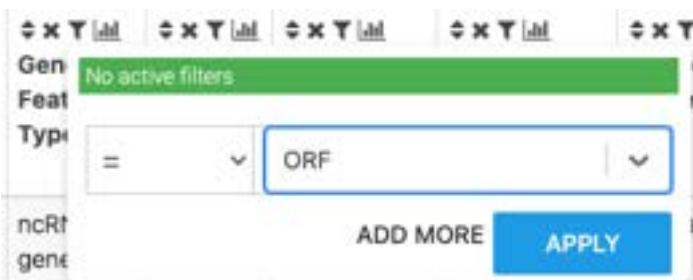
- And then select the **FUNCTION** tab and then **GO Slim Term => Gene**.

Enter "mitochondrion" as your GO slim term. This will return many bottom of the results and click "**View 3494 rows.**"

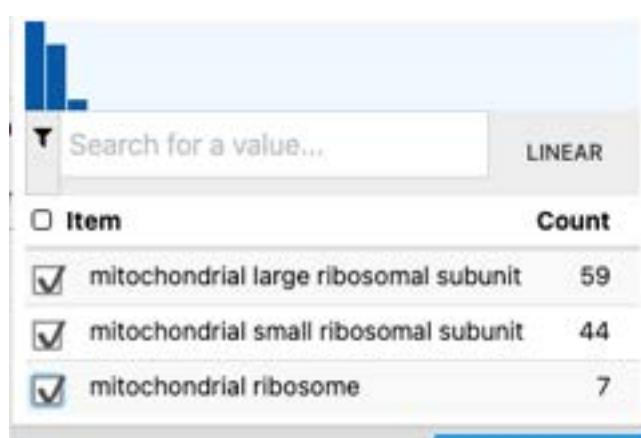


Gene > Primary DBID	Gene > Systematic Name	Gene > Standard Name	Gene > Feature Type	Gene > Qualifier	GO Annotation > Ontology Term . Identifier	GO Annotation > Ontology Term . Name	GO Annotation > Ontology Term . Namespace	Code > Code
9000029023	YNCQ0027W	8PMJ	ncRNA gene		GO:0005739	mitochondrion	cellular_component	C4
9000029023	YNCQ0027W	8PMJ	ncRNA gene		GO:0030678	mitochondrial ribonucleolar P complex	cellular_component	C4

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



- Next go to the "Ontology Term Name" column, hit the graph icon, and select the boxes for "**mitochondrial large ribosomal subunit**," "**mitochondrial small ribosomal subunit**," and "**mitochondrial ribosome**." Hit FILTER and you'll get 108 results.



- Save this list by clicking the **Save List** button on the upper right of the table and selecting "**Genes (89)**" at the top of the pull-down. Give it the name "List 1 MTR genes" and save.

Save a list of 91 Genes

Name:

Optional attributes

Description:

CANCEL **SAVE**

2. Find proteins that genetically interact with MTR proteins:

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

Widgets

Interactions

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

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

VIEW ALL

BioEntity.secondaryIdentifier	BioEntity.name
YNR020C	null
YBR122C	Mitochondrial Ribo Protein, Large subunit
YDL160C	DEAD box Helicase Homolog
YGL122C	Nuclear polyAenylyl Ribonucleoprotein

- The results table shows all genes/proteins with genetic or physical interactions with the MTR genes. There are over 17K of them.
- Go to the column for "**Details Relationship Type**" and filter for "**Genetic**." Hit Apply and you'll get a list of 8970 rows.

No active filters

Choose Interaction Detail > Relationship Type

genetic
physical

3. Find MTR interactors that are uncharacterized:

- Go to the "Gene Standard Name" column and click the filter icon. For the purposes of this exercise, filter to include ONLY the gene **RML2**, which yields 718 rows.
- Go to the column "Participant 2 Standard Name" (these are the genes that interact with RML2) and SORT this column by descending order, which puts the "no value" participants at the top. These are the potential

Showing 1 to 250 of 718 rows

Rows per page: 250 | Page 1 |

# × Trial Gene Systematic Name	# × Trial Gene Standard Name	# × Trial Organism Details Name	# × Trial Details Relationship Type	# × Trial Details Role 1	# × Trial Participant 2 Primary DBID	# × Trial Participant 2 Standard Name	# × Trial Experiment Name	
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500000035	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500000035	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500000132	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500000639	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500000686	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500000923	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500000923	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500001726	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500001902	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500001902	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500002852	NO VALUE	Costanzo M, et al. (2016)-27708008-Positive Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500003249	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500003354	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500003668	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500003699	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500003699	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500003729	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500004277	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500004277	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500004991	NO VALUE	Costanzo M, et al. (2016)-27708008-Positive Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500005039	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500005513	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YELO5OC	RML2	S. cerevisiae	NO VALUE	genetic	Bait	500006009	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic

- Given the current state of bugginess in this tool (it's undergoing major revision), you'll do best to copy and paste the section of the list that includes just the rows that say "no value" in the Participant 2 column. Copy this section and paste it into a blank worksheet.
- In the worksheet, sort by **column G** (the "Participant 2 Primary DBID") and then do a quick de-dupe of the list to leave 17 potential uncharacterized interactors. Save this worksheet as "**List 2: Uncharacterized MTR interactors**"
- Copy this list of 17 database IDs.



4. Analyze the results in FungiDB

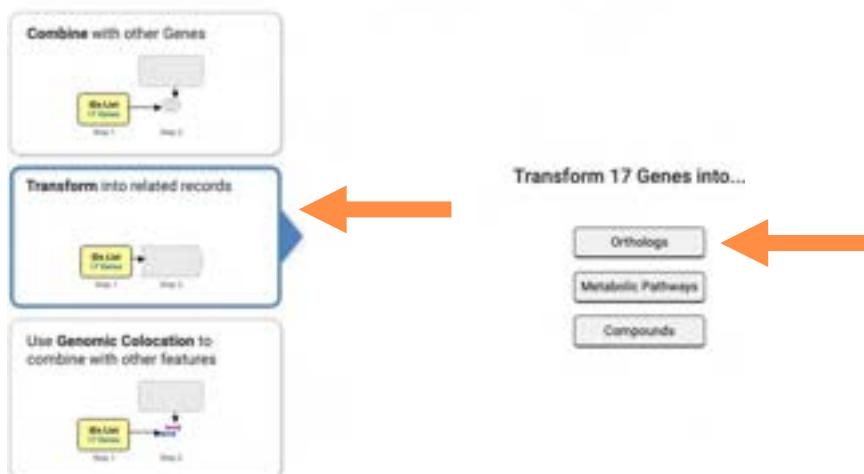
- The results of the above YeastMine analysis suggest 17 genes that potentially encode undiscovered subunits of the mitochondrial ribosome. Although these genes are uncharacterized, more data may exist on their orthologs in other organisms. Use FungiDB to survey the function of orthologs in other Fungi.
- Open the FungiDB homepage (<http://fungidb.org/>). In the “Search for...” section on the left, click "Genes" and then open the “Annotation, curation and identifiers” section and click on “List of ID(s)”.

The screenshot shows the FungiDB homepage. On the left, there is a search bar labeled "Search for..." with a dropdown menu showing "Genes" and "Annotation, curation and identifiers". Below this is a link to "List of IDs". On the right, there is a navigation bar with links like "Overview of Resources and Tools", "Take a Tour", "Getting Started", "Search Strategies", "Genome Browser", "Transcriptional Resources", "Phenotypic Data", and "Analyse My Data". A large orange arrow points from the "Getting Started" link towards the "List of IDs" link in the search dropdown.

- Using your list of DBIDs from YeastMine, copy and paste the systematic names of your results into the "Enter a list of IDs" box. Click on "Get Answer"
- Click on the "Add a step" button.

The screenshot shows the "My Search Strategies" page. At the top, there is a header with "Opened (1)", "All (2)", "Public (50)", and "Help". Below this is a section for "Unnamed Search Strategy *". It contains a yellow box labeled "ID's List 17 Genes" and a blue box labeled "+ Add a step Step 1". A large orange arrow points from the "Add a step" button towards the "ID's List" box. At the bottom, there is a footer with "17 Genes (17 ortholog groups)" and "Review this search".

- In the resulting pop-up window, click on **Transform into Related Records**. Select **Orthologs** and then **Fungi** and 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 17 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 Analyse Results

Genes: 3,345 Transcripts: 3,410 Show Only One Transcript Per Gene

1 2 3 ... 171 Rows per page: 25

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description
A105_06149	A105_06149_11	Cladophialophora psammophila CBS 119553	AMGX01000006.1;410,907..411,905(+)	YggS family pyridoxal phosphate enzyme
A107_01911	A107_01911_11	Cladophialophora yegrenii CBS 114405	AMGW01000001.5;292,758..5,293,729(-)	YggS family pyridoxal phosphate enzyme
A109_01176	A109_01176_11	Exophiala aquamarae CBS 17991#	AMGV01000001.3;248,178..3,265,017(-)	YggS family pyridoxal phosphate enzyme
A1Q1_02452	A1Q1_02452_11	Trichosporon asahii var. asahii CBS 2479	JH977584.125,762..126,580(+)	Pyridoxal phosphate homeostasis protein [Source:UniProtKB/TrEMBL]
ARK55_004020	ARK55_004020_11	Cordyceps militaris ATCC 34754	CP023325.955,409..956,339(+)	Alanine racemase Family (SRD)
AAP_04498	AAP_04498_11	Ascosporella apis ARSEF 740#	AZG291000022.15,983..16,801(+)	Pyridoxal phosphate homeostasis protein [Source:UniProtKB/TrEMBL]
A8675_3514	A8675_3514_11	Phialophora attinorum CBS 131958	LFJN01000014.675,275..676,243(-)	XMT_018143571.1

Word Cloud

Word Cloud Data

Filter words by rank: 1 to 50

Sort by: Rank A-Z

Mouse over a word to see its occurrence in the data

protein hypothetical Source:UniProtKB domain

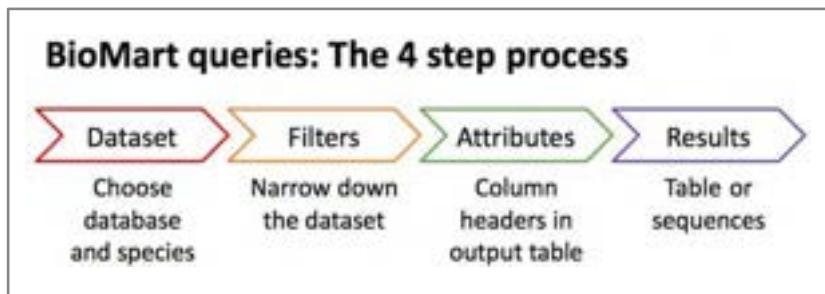
TrEMBL product unspecified containing reductase dehydrogenase alcohol oxidase Aldo-keto phosphatase Prostaglandin aryl-kinase oxidoreductase family NADP pyridoxal dependent Aldo-ketoreductase MFS transporter Ortholog dipeptidase enzyme membrane DUF966 YggS lipoate homotetrakis binding acid Pyridoxal kinase 2 monothioic conserved IML2 predicted DUF576 with AAO14 oxidase Pyridoxamine reductase role activity

Exercise: Ensembl Fungi BioMart

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

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

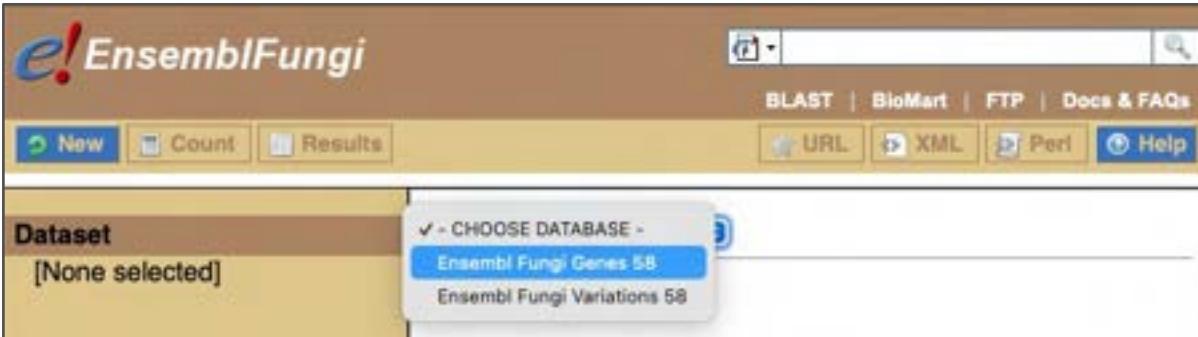
- How many genes within the 14:1128520-1142558 region are found in *Fusarium solani* that do not have an orthologue in *Fusarium verticillioides*?
- Export the gene name, locations and GO terms associated with these genes
- Export their cDNA sequences



Click on **BioMart** in the top header of any fungi.ensembl.org page or enter <https://fungi.ensembl.org/biomart/martview/> into your browser.

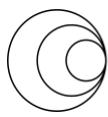
NOTE: These answers were determined using BioMart Ensembl Fungi 58

Step 1a: Choose [Ensembl Fungi Genes 58](#) as the database



The screenshot shows the Ensembl Fungi BioMart homepage. At the top, there's a navigation bar with links for BLAST, BioMart, FTP, and Docs & FAQs. Below the navigation is a toolbar with buttons for New, Count, and Results. On the left, there's a sidebar with a 'Dataset' section containing the text '[None selected]'. To the right of the sidebar, a dropdown menu is open under the heading '✓ - CHOOSE DATABASE -'. The menu lists two options: 'Ensembl Fungi Genes 58' (which is highlighted in blue) and 'Ensembl Fungi Variations 58'.

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



The screenshot shows the EnsemblFungi search interface. On the left, there's a sidebar with 'Dataset' set to 'Fusarium solani genes (v2.0)', 'Filters' set to '[None selected]', and 'Attributes' set to 'Gene stable ID'. The main area shows the selected dataset: 'Ensembl Fungi Genes 58' and 'Fusarium solani genes (v2.0)'.

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 'Filters' section expanded. A callout 'Step 2a: Click on Filters' points to the 'Filters' link in the sidebar. Another callout 'Step 2b: Expand the MULTI SPECIES COMPARISONS section' points to the 'MULTI SPECIES COMPARISONS' section in the main panel, which is currently expanded. In this section, there's a checkbox for 'Homologue filters' and two radio button options: 'Only' (selected) and 'Excluded'.

The screenshot shows the results after applying the filter. A callout 'Top tip: Click Count to check if your filter works' points to the 'Count' button in the top navigation bar. Another callout 'Step 2c: Choose Orthologous Fusarium verticillioides Genes' points to the 'Orthologous Fusarium verticillioides Genes' section in the main panel, where the 'Excluded' option is selected. A final callout 'Step 2d: Select the Excluded option' points to the 'Excluded' radio button in the same section.



The screenshot shows the EnsemblFungi BioMart interface. The left sidebar displays dataset information: Dataset 1 / 16163 Genes, Fusarium solani genes (v2.0), Filters: Orthologous Fusarium verticillioides Genes: Excluded, Chromosome/scaffold: 14, Start: 1128520, End: 1142558. The Attributes section includes Gene stable ID and Transcript stable ID. The Dataset section shows (None Selected). The main query form has 'Update Count' and 'Please restrict your query using criteria below' buttons. Step 2e: Expand the REGION section is highlighted with a callout pointing to the 'Chromosome/scaffold' dropdown menu, which lists chromosomes 4 through 18 and scaffolds sca_16_unmapped to sca_66_unmapped, with '14' selected. Step 2f: Enter Start/End coordinates is highlighted with a callout pointing to the 'Coordinates' section, which contains 'Start' (1128520) and 'End' (1142558) 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).



The screenshot shows the EnsemblFungi BioMart search interface. On the left, there's a sidebar with a 'Dataset' section containing '1 / 16163 Genes' and 'Fusarium solani genes (v2.0)'. Below that is a 'Filters' section for orthologous genes. Under 'Attributes', several options are listed: Gene stable ID, Transcript stable ID, Chromosome/scaffold name, Gene start (bp), Gene end (bp), and Gene name. A 'Dataset' section at the bottom says '(None Selected)'. The main area has a heading 'Please select columns to be included in the output and hit 'Results' when ready'. It includes sections for 'Step 3a: Click on Attributes' (with 'Features' selected), 'Step 3b: In the Features category, expand the GENE section', and 'Step 3c: Select GO term accession and name'. There are checkboxes for various gene and transcript details like start sites, lengths, and GC content.

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

This screenshot continues the EnsemblFungi BioMart search interface. The 'Attributes' section remains the same. The main area now shows the 'EXTERNAL' section, which is expanded. A callout 'Step 3d: Select GO term accession and name' points to the 'GO' section where 'GO term accession' and 'GO term name' checkboxes are checked. Other sections visible include 'GOSlim GOA' (with 'GOSlim GOA Accession(s)' checked), 'Pathogen Phenotypes (source: PHI-base)', and 'External References (max 3)'.

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 will retrieve your BioMart results in tabular format. Notice the order of the columns - these are in the same order in which you selected your **Attributes**.

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

EnsemblFungi

Step 4a: Click on Results

Dataset: 1716163 Genes
Fusarium solani genes (v2.0)
Filters
Orthologous Fusarium verticillioides Genes: Excluded
Chromosomes/sccaffold: 14
Start: 1128520
End: 1142558
Attributes
Gene stable ID
Transcript stable ID
Chromosome/sccaffold name
Gene start (bp)
Gene end (bp)
Gene name
GO term accession
GO term name

Export all results to: File
Email notification to: results only

Step 4b: Select All to view all results in a new tab

View: 10 rows as: HTML Unique results only

Gene stable ID	Transcript stable ID	Chromosome/sccaffold name	Gene start (bp)	Gene end (bp)	Gene name	GO term accession	GO term name
NeochG73360	NeochT73360	54	1129113	1121296	PEP5	GO-0016621	integral component of membrane
NeochG73360	NeochT73360	54	1129113	1121296	PEP5	GO-0022857	transmembrane transporter activity
NeochG73360	NeochT73360	54	1129115	1121296	PEP5	GO-0055085	transmembrane transport
NeochG73360	NeochT73360	54	1129175	1121296	PEP5	GO-0016622	membrane
NeochG64107	NeochT64107	54	1121263	1123846	POA1	GO-0016621	integral component of membrane
NeochG64107	NeochT64107	54	1121263	1123846	POA1	GO-0004497	monooxygenase activity
NeochG64107	NeochT64107	54	1121263	1123846	POA1	GO-0020937	heme binding
NeochG64107	NeochT64107	54	1121263	1123846	POA1	GO-0016703	zedoxinase activity, acting on peptidic donors, with incorporation or reduction of molecular oxygen
NeochG64107	NeochT64107	54	1121263	1123846	POA1	GO-0003302	iron ion binding
NeochG64107	NeochT64107	54	1121263	1123846	POA1	GO-0016622	membrane

Each attribute becomes a column in the results table

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

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.

EnsemblFungi

Step 3.2a: Click on Attributes again

Dataset: 1716163 Genes
Fusarium solani genes (v2.0)
Filters
Orthologous Fusarium verticillioides Genes: Excluded
Chromosomes/sccaffold: 14
Start: 1128520
End: 1142558
Attributes
Gene stable ID
Transcript stable ID
Gene name
cDNA sequences

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

Step 3.2b: Click on Sequences

Features Homologues (Max size)
 Structures Sequences

SEQUENCES:
Sequences (max 1)

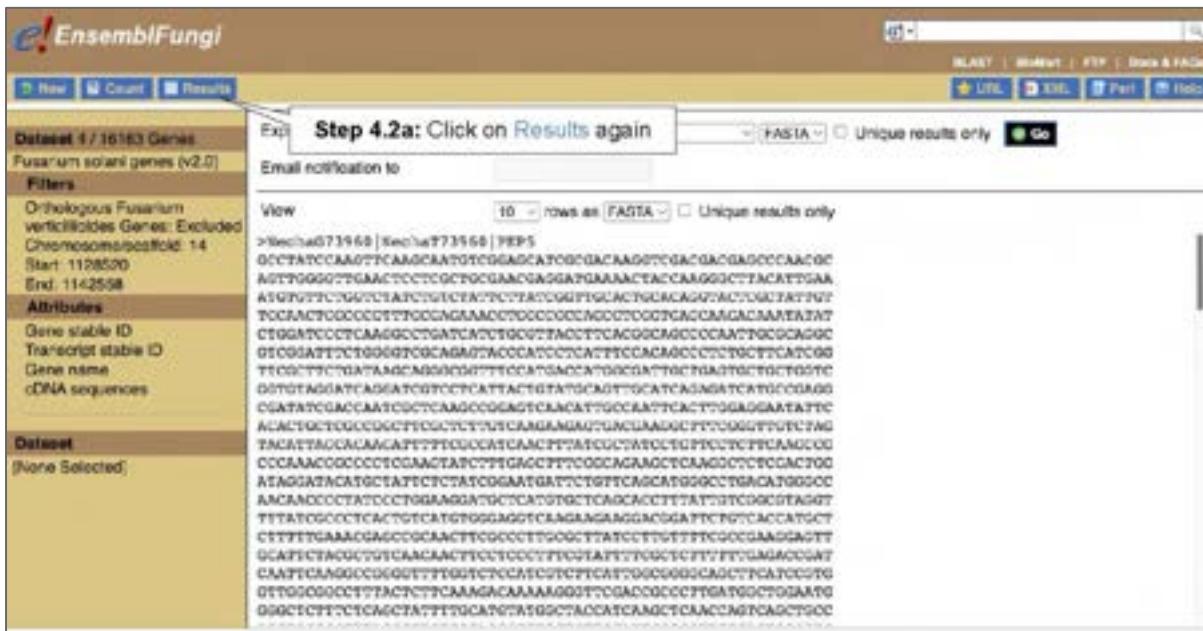
Step 3.2c: Select cDNA sequences

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

Upstream 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. The top navigation bar includes links for BLAST, BioMart, FTP, Data & Tools, URL, XML, Perl, and Help. The main search bar contains the query 'Nechtae#73968'. Below the search bar, there are tabs for 'How', 'Count', and 'Results'. The 'Results' tab is selected, displaying the following information:

- Dataset 1 / 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, Gene name, cDNA sequences
- Dataset**: (None Selected)

The main content area displays the sequence results for the gene Nechtae#73968. A prominent message in the center says 'Step 4.2a: Click on Results again'. The sequence is presented in FASTA format, starting with the header 'Nechtae#73968 | Nechtae#73968 | PEPS'. 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.

Exercise 1 answers

You can open BioMart by clicking [BioMart](#) in the navigation bar at the top of any Ensembl Fungi page, or by entering the URL <https://fungi.ensembl.org/biomart/martview/> in your browser. Using the four-step process described above, we need to select the following in BioMart options in the left-hand panel:

1. Dataset: Click on [CHOOSE DATABASE](#) and select [Ensembl Fungi Genes 58](#) from the drop-down menu. Click on [CHOOSE DATASET](#) and select [Zymoseptoria tritici genes \(MG2\)](#) from the drop-down menu.



The screenshot shows the Ensembl Fungi BioMart interface. At the top, there's a navigation bar with links for BLAST, BioMart, FTP, and Docs & FAQ. Below the navigation bar, there are buttons for New, Count, and Results. The main area has two dropdown menus under 'Dataset'. The first dropdown is set to 'Ensembl Fungi Genes 58' and the second is set to 'Zymoseptoria tritici genes (MG2)'. On the left side, there's a sidebar with sections for 'Dataset' (set to 'Zymoseptoria tritici genes (MG2)'), 'Filters' (set to '[None selected]'), and 'Attributes' (listing 'Gene stable ID' and 'Transcript stable ID'). At the bottom of the interface, it says 'Ensembl Genomes release 58 - January 2024 (v1)'.

2. Filters: Open the [REGION](#) tab and select [1](#) under [Chromosome/scaffold](#). Open the [GENE ONTOLOGY](#) tab and enter [detoxification](#) under [GO Term Name](#). Click on the [Count](#) button in the top left-hand corner. Your filter should apply to 19/11,091 genes.

The screenshot shows the EnsemblFungi search interface. On the left, a sidebar displays the dataset as "Zymoseptoria tritici genes (MG2)" and lists filters for "Chromosome/scaffold: 1" and "GO Term Name [e.g. regulation of biological process]: detoxification". Below that, attributes like "Gene stable ID" and "Transcript stable ID" are listed. A "Dataset" section shows "[None Selected]". The main panel contains several search fields and dropdown menus. Under "PATHOGEN PHENOTYPES (PHI-BASE)", there's a dropdown for "GO Evidence code" with options EXP, IDA, IFA, IGI, and IMP. Under "MULTI SPECIES COMPARISONS", there are sections for "PROTEIN DOMAINS AND FAMILIES" and "VARIANT". At the bottom, a status bar indicates "Ensembl Genomes release 58 - January 2024 (x) []".

3. Attributes: Select **Homologues** from the options on the top. Open the **GENE** tab and select **Gene name**. Open the **ORTHOLOGUES [A-E]** tab and select the following options:

- *Blumeria graminis* gene stable ID
- *Blumeria graminis* gene name
- *Blumeria graminis* homology type
- *Blumeria graminis* orthology confidence [0 low, 1 high]
- *Botrytis cinerea* B05.10 gene stable ID
- *Botrytis cinerea* B05.10 gene name
- *Botrytis cinerea* B05.10 homology type
- *Botrytis cinerea* B05.10 orthology confidence [0 low, 1 high]
- *Cryptococcus neoformans* var. *neoformans* JEC21 gene stable ID
- *Cryptococcus neoformans* var. *neoformans* JEC21 gene name
- *Cryptococcus neoformans* var. *neoformans* JEC21 homology type
- *Cryptococcus neoformans* var. *neoformans* JEC21 orthology confidence [0 low, 1 high]

Open the **ORTHOLOGUES [P-T]** tab and select the following options:

- *Saccharomyces cerevisiae* gene stable ID
- *Saccharomyces cerevisiae* gene name
- *Saccharomyces cerevisiae* homology type
- *Saccharomyces cerevisiae* orthology confidence [0 low, 1 high]

EnsemblFungi

[New](#) [Count](#) [Results](#)

[BLAST](#) | [BioMart](#) | [FTP](#) | [Docs & FAQ](#)

[★ URL](#) [XML](#) [Perl](#) [Help](#)

Dataset Zymoseptoria tritici genes (MG2)	<input type="checkbox"/> Pyrenopora tritici-repentis P1-1C-BFP chromosome/scaffold end (bp)	<input type="checkbox"/> Pyrenopora tritici-repentis P1-1C-BFP orthology confidence [0 low, 1 high]
Filters	<input type="checkbox"/> Saccharomyces cerevisiae gene stable ID <input checked="" type="checkbox"/> Saccharomyces cerevisiae gene name <input type="checkbox"/> Saccharomyces cerevisiae protein or transcript stable ID <input type="checkbox"/> Saccharomyces cerevisiae chromosome/scaffold name <input type="checkbox"/> Saccharomyces cerevisiae chromosome/scaffold start (bp) <input type="checkbox"/> Saccharomyces cerevisiae chromosome/scaffold end (bp)	
Attributes	<input type="checkbox"/> Query protein or transcript ID <input type="checkbox"/> Last common ancestor with Saccharomyces cerevisiae <input checked="" type="checkbox"/> Saccharomyces cerevisiae homology type <input type="checkbox"/> %id. target Saccharomyces cerevisiae gene identical to query gene <input type="checkbox"/> %id. query gene identical to target Saccharomyces cerevisiae gene <input checked="" type="checkbox"/> Saccharomyces cerevisiae orthology confidence [0 low, 1 high]	
	Saccharomyces cerevisiae Orthologues	
	Schizosaccharomyces cryophilus Orthologues	
	<input type="checkbox"/> Schizosaccharomyces cryophilus gene stable ID <input type="checkbox"/> Schizosaccharomyces cryophilus gene name <input type="checkbox"/> Schizosaccharomyces cryophilus protein or transcript stable ID <input type="checkbox"/> Schizosaccharomyces cryophilus chromosome/scaffold name	
	<input type="checkbox"/> Query protein or transcript ID <input type="checkbox"/> Last common ancestor with Schizosaccharomyces cryophilus <input type="checkbox"/> Schizosaccharomyces cryophilus homology type <input type="checkbox"/> %id. target Schizosaccharomyces	

Ensembl Genomes release 58 - January 2024 ([x](#)) [\[?\]](#)

- Click on the **Results** button in the top left-hand corner to view your output table. Select **All** from the drop-down menu to open the full table in a new tab.

EnsemblFungi

[New](#) [Count](#) [Results](#)

[BLAST](#) | [BioMart](#) | [FTP](#) | [Docs & FAQ](#)

[★ URL](#) [XML](#) [Perl](#) [Help](#)

Dataset Zymoseptoria tritici genes (MG2)	Export all results to <input type="text"/> File <input type="radio"/> TSV <input type="checkbox"/> Unique results only Go							
Filters	Email notification to <input type="text"/>							
Attributes	View <input type="radio"/> 10 <input type="radio"/> 20 <input type="radio"/> 50 <input type="radio"/> 100 <input type="radio"/> 150 <input type="radio"/> 200 <input type="radio"/> All rows as <input type="radio"/> HTML <input type="checkbox"/> Unique results only							
Gene stable ID	Transcript stable ID	ne me	Blumeria graminis gene stable ID	Blumeria graminis gene name	Blumeria graminis homology type	Blumeria graminis orthology confidence [0 low, 1 high]	Botrytis cinerea B05.1 gene stable ID	Botrytis cinerea B05.1 gene name
Mycg3G900087	Mycg3T900087				ortholog_oneZone	1	Bon03g01480	B0906
Mycg3G102589	Mycg3T102589		BLGH_05232		ortholog_oneZone	1	Bon03g01480	B0906
Mycg3G102589	Mycg3T102589		BLGH_05232		ortholog_oneZone	1	Bon03g01480	B0906
Mycg3G102589	Mycg3T102589		BLGH_05232		ortholog_oneZone	1	Bon03g01480	B0906
Mycg3G102589	Mycg3T102589		BLGH_05232		ortholog_oneZone	1	Bon03g01480	B0906
Mycg3G983865	Mycg3T983865						Bon16g00120	B0906
Mycg3G33131	Mycg3T33131							
Mycg3G1027202	Mycg3T1027202		BLGH_06273		ortholog_oneZone	0	Bon10g02340	
Mycg3G54449	Mycg3T54449		BLGH_06551		ortholog_oneZone	0	Bon14g01030	B07a1

Ensembl Genomes release 58 - January 2024 ([x](#)) [\[?\]](#)

- No, not all *Z. tritici* genes located on chromosome 1 with the associated GO term 'detoxification' have an orthologue in the other species. *B. graminis* causes powdery mildew on grasses (e.g. cereals), *B. cinerea* is known to cause botrytis bunch rot in grape and *C. neoformans* is the causative agent of cryptococcosis and cryptococcal meningitis. Do you see a correlation?

- (b) CNM01690 in *C. neoformans* has high orthology confidence.

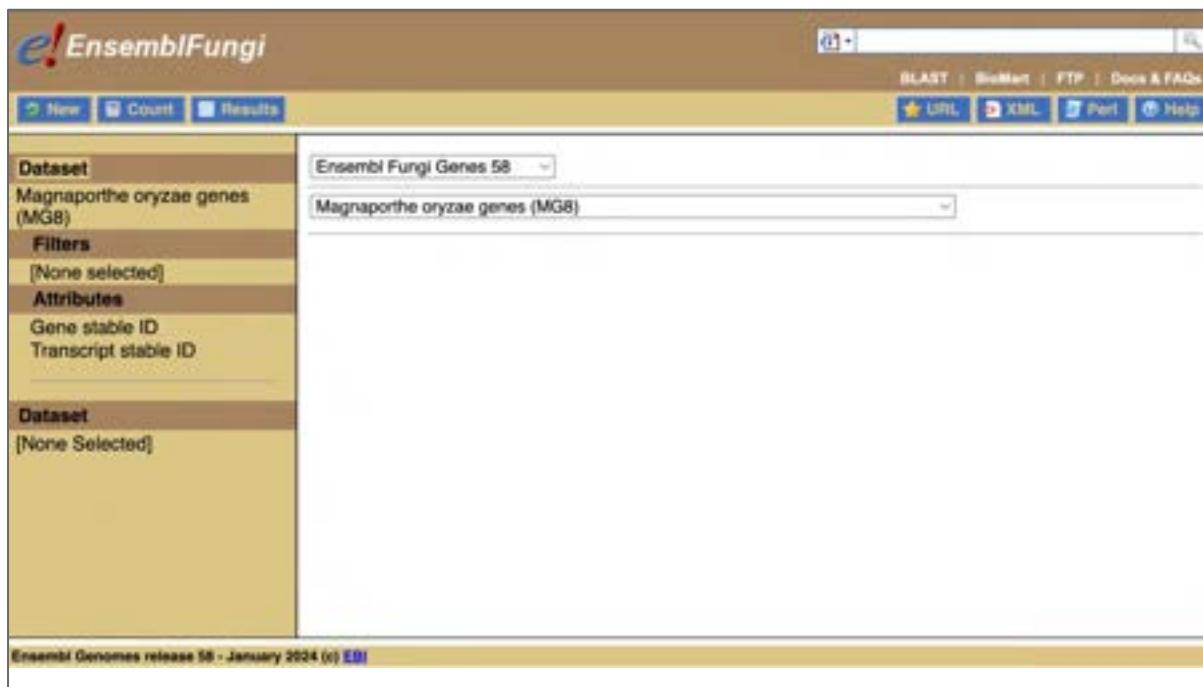
Additional BioMart Exercise 2 – Finding genes by protein domain

Generate a list of all *Magnaporthe oryzae* (MG8) genes on chromosome 4 that are annotated to contain Transmembrane domains/helices. Include the Ensembl gene stable ID and description.

Exercise 2 answers

Click on the **New** button in the top left-hand corner to start a new BioMart query. Using the four-step process described above, we need to select the following in BioMart options in the left-hand panel:

1. Dataset: Click on **CHOOSE DATABASE** and select **Ensembl Fungi Genes 58** from the drop-down menu. Click on **CHOOSE DATASET** and select ***Magnaporthe oryzae* genes (MG8)** from the drop-down menu.



The screenshot shows the Ensembl Fungi BioMart interface. The top navigation bar includes links for BLAST, BioMart, FTP, Docs & FAQs, URL, XML, Perl, and Help. The main interface has three tabs at the top: New, Count, and Results. The 'New' tab is selected. On the left, there's a sidebar with sections for Dataset (set to Magnaporthe oryzae genes (MG8)), Filters (None selected), and Attributes (Gene stable ID, Transcript stable ID). The main panel shows the chosen dataset: Ensembl Fungi Genes 58 and Magnaporthe oryzae genes (MG8). At the bottom, it says Ensembl Genomes release 58 - January 2024 (x) EBI.

2. Filters: Open the REGION tab and select 4 under Chromosome/scaffold. Open the PROTEIN DOMAINS AND FAMILIES tab and select With transmembrane helices - Only under Limit to genes.... Click on the Count button in the top left-hand corner. Your filter should apply to 297/13,470 genes.

e!EnsemblFungi

New Count Results

Dataset 297 / 13470 Genes
Magnaporthe oryzae genes (MG8)

Filters
Chromosome/scaffold: 4
With Transmembrane helices: Only

Attributes
Gene stable ID
Transcript stable ID

Dataset
[None Selected]

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:

MULTI SPECIES COMPARISONS:

PROTEIN DOMAINS AND FAMILIES:
 Limit to genes ... With Transmembrane helices Only Excluded
 Limit to genes with these family or domain IDs [Max 500 advised] Interpro ID(s) (e.g. IPR000008)
 Choose file No file chosen

Ensembl Genomes release 58 - January 2024 (x) EBI

3. Attributes: Select Features from the options on the top. Open the GENE tab, unselect Transcript stable ID and select Gene description.

e!EnsemblFungi

New Count Results

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]

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

Features Homologues (Max select 6 orthologues)
 Structures Sequences

GENE:

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 name
 Source of transcript name
 Transcript count
 Gene % GC content
 Gene type
 Transcript type
 Source (gene)
 Source (transcript)
 Gene Synonym

Ensembl Genomes release 58 - January 2024 (x) EBI

4. Results: Click on the Results button in the top left-hand corner to view your output table. Select All from the drop-down menu to open the full table in a new tab.

e! EnsemblFungi

New Count Results

BLAST | BiMart | FTP | Docs & FAQs

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 Unique results only

Email notification to

View rows as Unique results only

Gene stable ID	Gene description
MGG_17084	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4N801]
MGG_03684	Mitochondrial distribution and morphology protein 38 [Source:UniProtKB/TrEMBL;Acc:Q4N6R1]
MGG_09963	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4N9P1]
MGG_02644	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4N713]
MGG_06610	Cytochrome b6 [Source:UniProtKB/TrEMBL;Acc:Q4N6W8]
MGG_09720	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4NAH4]
MGG_03721	Urea transporter [Source:UniProtKB/TrEMBL;Acc:Q4N6H1]
MGG_13659	Dicarboxylic amino acid permease [Source:UniProtKB/TrEMBL;Acc:Q4NAK4]
MGG_05498	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4NAP3]
MGG_13624	ABC transporter CDI4 [Source:UniProtKB/TrEMBL;Acc:Q4N6L5]

Ensembl Genomes release 58 - January 2024 (x) EBI

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.

- (a) Do these 36 protein IDs correspond to 36 genes?

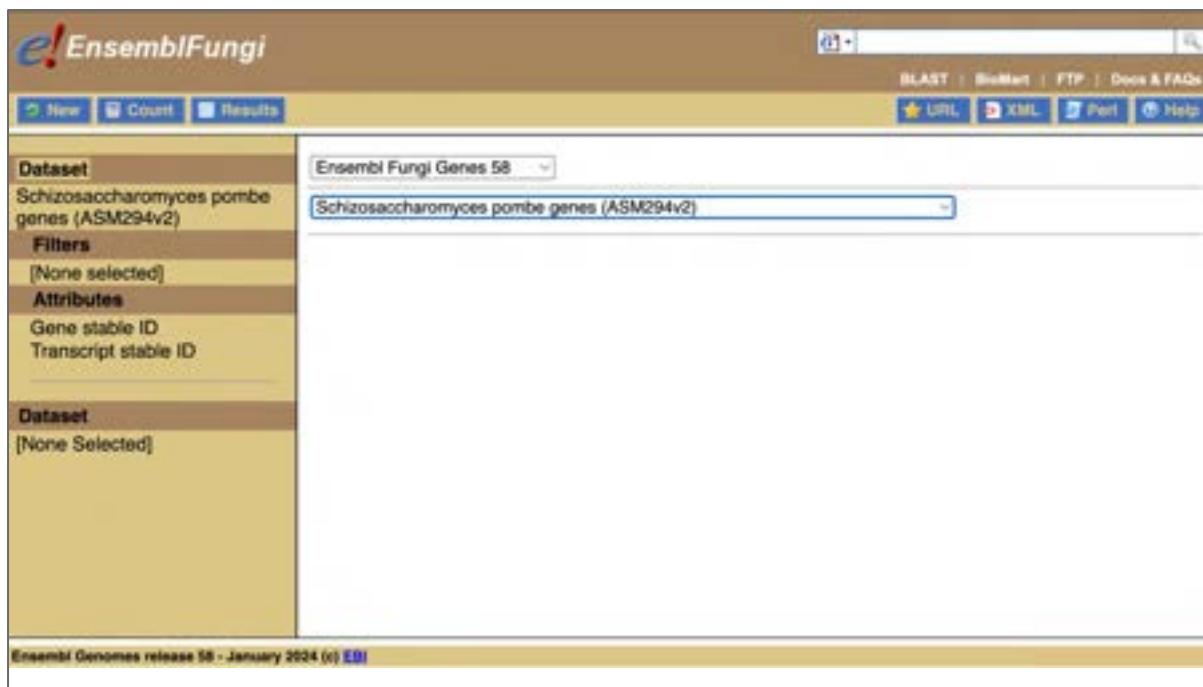
Input list of IDs:

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	O9428

Exercise 3 answers

Click on the **New** button in the top left-hand corner to start a new BioMart query. Using the four-step process described above, we need to select the following in BioMart options in the left-hand panel:

- Dataset: Click on **CHOOSE DATABASE** and select **Ensembl Fungi Genes 58** from the drop-down menu. Click on **CHOOSE DATASET** and select ***Schizosaccharomyces pombe* genes (ASM294v2)** from the drop-down menu.



The screenshot shows the Ensembl Fungi BioMart interface. The top navigation bar includes links for BLAST, BioMart, FTP, and Docs & FAQs, along with buttons for URL, XML, Perl, and Help. The main interface has three tabs: New, Count, and Results, with New selected. On the left, there's a sidebar with sections for Dataset, Filters, and Attributes. The Dataset section shows 'Ensembl Fungi Genes 58' selected and 'Schizosaccharomyces pombe genes (ASM294v2)' chosen under it. The Filters section says '[None selected]'. The Attributes section lists 'Gene stable ID' and 'Transcript stable ID'. Below the sidebar, another Dataset section shows '[None Selected]'. At the bottom, a footer bar indicates 'Ensembl Genomes release 58 - January 2024 (c) EBI'.



2. Filters: Open the **GENE** tab and paste your list of IDs into the text box under **Input external references ID list**. Select **UniProtKB/Swiss-Prot ID(s)** [e.g. A0ZWU1] from the drop-down menu above to specify the type of IDs you are giving. Click on the **Count** button in the top left-hand corner. Your filter should apply to 36/7,268 genes.

The screenshot shows the Ensembl Fungi interface for Dataset 36 / 7268 Genes. On the left, there's a sidebar with 'Filters' set to 'UniProtKB/Swiss-Prot ID(s) [e.g. A0ZWU1]' and 'Attributes' set to 'Gene stable ID' and 'Transcript stable ID'. The main area has a heading 'Please restrict your query using criteria below' with a note '(If filter values are truncated in any lists, hover over the list item to see the full text)'. It includes sections for 'REGION', 'GENE' (with options for 'Limit to genes (external references)...' and 'Input external references ID list [Max 500 advised]'), and 'Dataset' (set to '[None Selected]'). At the bottom, it says 'Ensembl Genomes release 58 - January 2024 (x) EBI'.

3. Attributes: Select **Features** from the options on the top. Open the **GENE** tab, unselect **Transcript stable ID** and select **Gene name** and **Gene description**. Open the **EXTERNAL** tab, scroll down to External References and select **PomBase ID**.

The screenshot shows the Ensembl Fungi interface for Dataset 36 / 7268 Genes. The sidebar has 'Filters' set to 'UniProtKB/Swiss-Prot ID(s) [e.g. A0ZWU1]', 'Attributes' set to 'Gene stable ID', 'Gene name', 'Gene description', and 'PomBase ID'. The main area shows the 'External References (max 3)' section with a large list of checkboxes. The 'PomBase ID' checkbox is selected. Other options include ChEMBL ID, Enzyme EC Number ID, European Nucleotide Archive ID, Fission Yeast Phenotype Ontology ID, INSDC protein ID, KEGG ID, MEROPE - the Peptidase Database ID, NCBI gene (formerly Entrezgene) description, NCBI gene (formerly Entrezgene) accession, NCBI gene (formerly Entrezgene) ID, Orthologous Gene ID, PDB ID, RefSeq mRNA ID, RefSeq mRNA predicted ID, RefSeq peptide ID, RefSeq peptide predicted ID, RFAM ID, Sequence Ontology ID, Sequence Publications ID, SPD ID, STRING ID, tRNAscan-SE ID, UniParc ID, UniProtKB/SpliceVariant ID, UniProtKB/Swiss-Prot ID, UniProtKB/TrEMBL ID, WikiGene description, WikiGene name, and WikiGene ID. At the bottom, it says 'Ensembl Genomes release 58 - January 2024 (x) EBI'.

4. Results: Click on the **Results** button in the top left-hand corner to view your output table. Select **All** from the drop-down menu to open the full table in a new tab.

Gene stable ID	Gene name	Gene description	PomBase ID
SPBC29A3.14c	trt1	telomerase reverse transcriptase 1 protein Trt1 [Source:PomBase;Acc:SPBC29A3.14c]	SPBC29A3.14c.1
SPAC15A10.08	ain1	alpha-actinin [Source:PomBase;Acc:SPAC15A10.08]	SPAC15A10.08.1
SPAC16E8.07c	vph1	V-type ATPase V0 subunit a (predicted) [Source:PomBase;Acc:SPAC16E8.07c]	SPAC16E8.07c.1
SPAC29B12.02c	set2	histone lysine methyltransferase Set2 [Source:PomBase;Acc:SPAC29B12.02c]	SPAC29B12.02c.1
SPAC2C4.07c	dis32	3'-5'-exoribonuclease activity Dis3L2 [Source:PomBase;Acc:SPAC2C4.07c]	SPAC2C4.07c.1
SPACUNK4.10		glyoxylate reductase (predicted) [Source:PomBase;Acc:SPACUNK4.10]	SPACUNK4.10.1
SPBC16E9.11c	mub3	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	atll	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
SPBC88T.02		CIC chloride channel (predicted) [Source:PomBase;Acc:SPBC88T.02]	SPBC88T.02.1
SPBC1604.15	gpi16	pig-T, Gpi16 (predicted) [Source:PomBase;Acc:SPBC1604.15]	SPBC1604.15.1
SPCC1620.11	nug97	nucleoporin Nie96 homolog [Source:PomBase;Acc:SPCC1620.11]	SPCC1620.11.1
SPBC609.02	ptnl	phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase Ptn1 [Source:PomBase;Acc:SPBC609.02]	SPBC609.02.1
SPCC18.18c	fan1	fumurate 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	ash1	bis(5'-nucleosidyl)-tetraphosphatase [Source:PomBase;Acc:SPCC4G3.02]	SPCC4G3.02.1
SPCC290.03c	nap186	nucleoporin Nap186 [Source:PomBase;Acc:SPCC290.03c]	SPCC290.03c.1
SPBC3D6.07	gpi3	pig-A, phosphatidylinositol N-acetylglucosaminyltransferase subunit Gpi3 (predicted) [Source:PomBase;Acc:SPBC3D6.07]	SPBC3D6.07.1
SPCC18B5.11c	cds1	replication checkpoint kinase Cds1 [Source:PomBase;Acc:SPCC18B5.11c]	SPCC18B5.11c.1
SPBC428.01c	nup107	nucleoporin Nap107 [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	psd4	proline specific plasma membrane permease Psd4 (predicted) [Source:PomBase;Acc:SPAC869.10c]	SPAC869.10c.1
SPAC1002.03c	gln2	glucosidase II alpha subunit Gln2 [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	snf8	pig-K [Source:PomBase;Acc:SPCC11E10.02c]	SPCC11E10.02c.1
SPAC1805.15c	mub2	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

(a) Yes, the 36 UniProt IDs correspond to 36 genes. However, not all of them have a gene name assigned to them (e.g. SPACUNK4.10).

Exercise: Exploring host-pathogen interactions in Ensembl Fungi

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

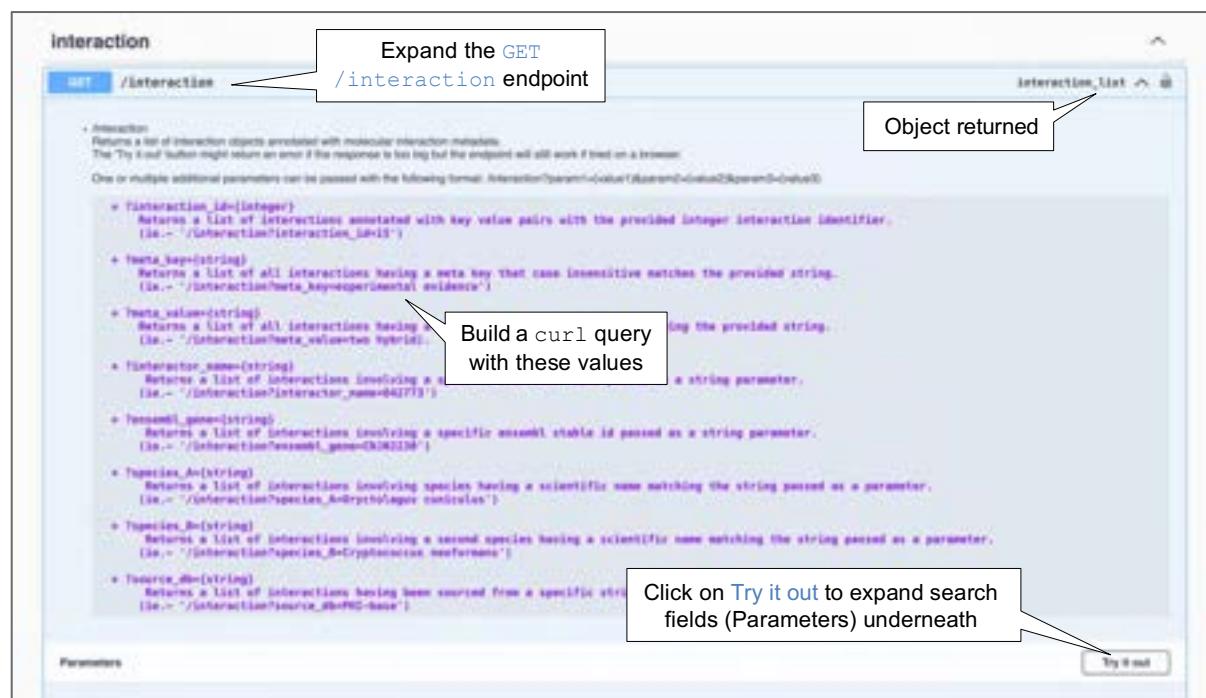
Zymoseptoria tritici (also known as *Septoria tritici* and *Mycosphaerella graminicola*) is a fungal pathogen that causes septoria leaf blotch disease in *Triticum aestivum* (wheat). This fungus is considered a major threat to wheat production worldwide, and its ability to rapidly adapt to fungicides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

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

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

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

Enter https://interactions.rest.ensembl.org/interactions_by_prodid/ into your browser and expand the GET /interaction documentation by clicking on `interaction_list`. This opens a description and all available parameters for the endpoint. Click on `Try it out` to start your REST API request.



interaction

Expand the `GET /interaction` endpoint

Object returned

Build a curl query with these values

Click on `Try it out` to expand search fields (Parameters) underneath

Try it out

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

`species_A: Zymoseptoria tritici`

`species_B: Triticum aestivum`

`meta_key: disease`

Click on `Execute` to submit your request.

Parameters

Name	Description
interaction_id	interaction_id
interactor	interactor
interactor_name	interactor_name
ensembl_gene	ensembl_gene
species_A	Zymoseptoria tritici
species_B	Triticum aestivum
source_db	source_db
meta_value	meta_value
meta_key	disease

Cancel

Enter your parameters into the query fields

Execute

Click on [Execute](#) to submit your query

Scroll down to ‘Responses’ to view your output.

Responses

Response content type: Application/json

Curl

```
curl -X GET "https://interactions.react.ensembl.org/InteractionSpecies_A-Zymoseptoria_tritici/InteractionSpecies_B-Triticum_aestivum?meta_value=disease"
```

Request URL

https://interactions.react.ensembl.org/InteractionSpecies_A-Zymoseptoria_tritici/InteractionSpecies_B-Triticum_aestivum?meta_value=disease

Server response

Code: 200

Response body

```
[{"interaction_id": "Mycgr3G53658", "interactor": "Zymoseptoria tritici", "interactor_name": "Zymoseptoria tritici", "ensembl_gene": "ENSG00000244222", "species": "Zymoseptoria tritici", "species_name": "Zymoseptoria tritici", "source_db": "Reactome", "meta_value": "disease"}, {"interaction_id": "Mycgr3G88451", "interactor": "Triticum aestivum", "interactor_name": "Triticum aestivum", "ensembl_gene": "ENSG00000244224", "species": "Triticum aestivum", "species_name": "Triticum aestivum", "source_db": "Reactome", "meta_value": "disease"}, {"interaction_id": "Mycgr3G85040", "interactor": "Triticum aestivum", "interactor_name": "Triticum aestivum", "ensembl_gene": "ENSG00000244225", "species": "Triticum aestivum", "species_name": "Triticum aestivum", "source_db": "Reactome", "meta_value": "disease"}, {"interaction_id": "Mycgr3G40048", "interactor": "Triticum aestivum", "interactor_name": "Triticum aestivum", "ensembl_gene": "ENSG00000244226", "species": "Triticum aestivum", "species_name": "Triticum aestivum", "source_db": "Reactome", "meta_value": "disease"}, {"interaction_id": "Mycgr3G111221", "interactor": "Triticum aestivum", "interactor_name": "Triticum aestivum", "ensembl_gene": "ENSG00000244227", "species": "Triticum aestivum", "species_name": "Triticum aestivum", "source_db": "Reactome", "meta_value": "disease"}, {"interaction_id": "Mycgr3G103264", "interactor": "Triticum aestivum", "interactor_name": "Triticum aestivum", "ensembl_gene": "ENSG00000244228", "species": "Triticum aestivum", "species_name": "Triticum aestivum", "source_db": "Reactome", "meta_value": "disease"}, {"interaction_id": "Mycgr3G89160", "interactor": "Triticum aestivum", "interactor_name": "Triticum aestivum", "ensembl_gene": "ENSG00000244229", "species": "Triticum aestivum", "species_name": "Triticum aestivum", "source_db": "Reactome", "meta_value": "disease"}, {"interaction_id": "Mycgr3G80707", "interactor": "Triticum aestivum", "interactor_name": "Triticum aestivum", "ensembl_gene": "ENSG00000244230", "species": "Triticum aestivum", "species_name": "Triticum aestivum", "source_db": "Reactome", "meta_value": "disease"}]
```

You can copy the Request URL to obtain the results programmatically

Download your results

Response headers

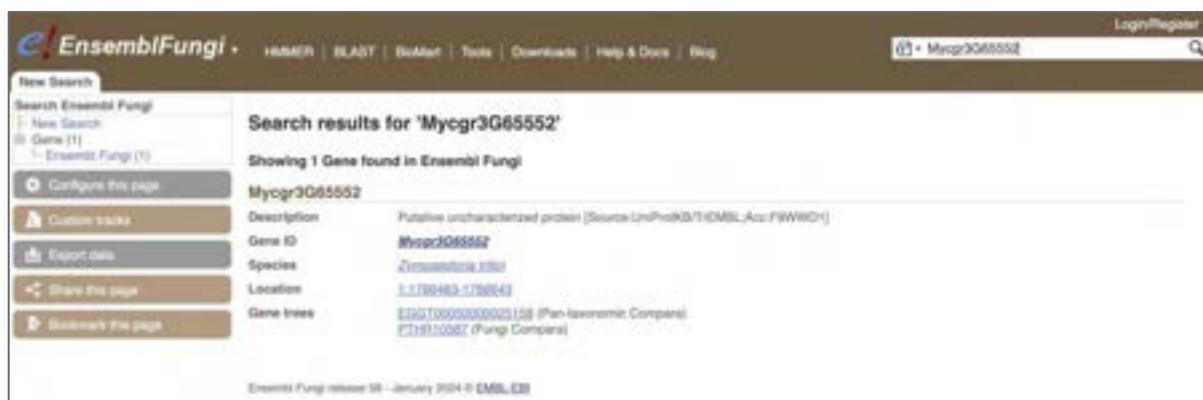
```
HTTP/1.1 200 OK
Content-Type: application/json
Content-Length: 38232
Date: Tue, 07 May 2024 16:48:29 UTC
Referer-Policy: same-origin
```

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

Under ‘Server response’, you should get the following output: zymoseptoria_tritici”:
 [“Mycgr3G53658”, “Mycgr3g88451”, “Mycgr3G85040”, “Mycgr3G40048”,
 “Mycgr3G111221”, “Mycgr3G103264”, “Mycgr3G89160”, “Mycgr3G80707”,

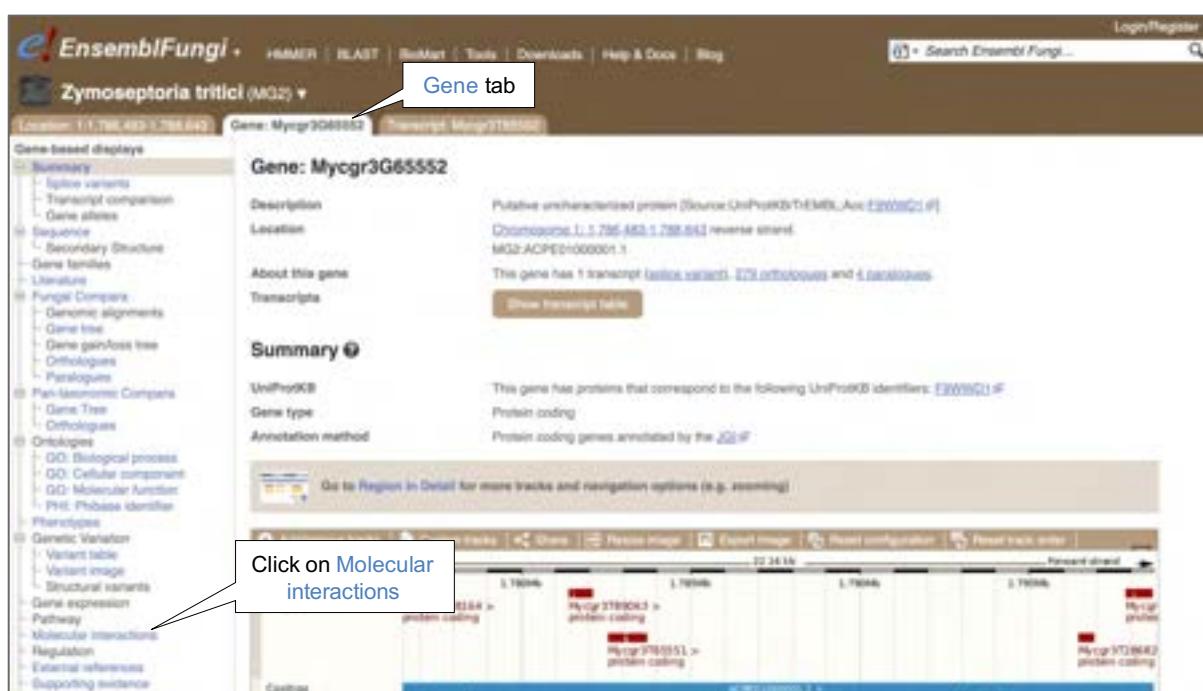
"Mycgr3G65552", "Mycgr3g105487", "Mycgr3G70181", "Mycgr3G46840", "Mycgr3G93828", "Mycgr3G31676", "Mycgr3G51018", "Mycgr3G36951", "Mycgr3G77528", "Mycgr3G39611", "Mycgr3G96592", "Mycgr3G86705", "Mycgr3G107320", "Mycgr3G74194", "Mycgr3G87000", "Mycgr3G100355", "Mycgr3G92404", "Mycgr3G69942"]

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



Search results for 'Mycgr3G65552'	
Showing 1 Gene found in Ensembl Fungi	
Mycgr3G65552	Description Putative uncharacterized protein [Source:UniProtKB/Trembl;Acc:FJWWDH]
	Gene ID Mycgr3G65552
	Species Zymoseptoria tritici
	Location 1:1789483-1789643 (Pan-taxonomic Composite)
	Gene Alias ESGT000000000002158 (Pan-taxonomic Composite)
	PFT1815087 (Fungi Composite)

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.



Gene tab

Gene: Mycgr3G65552

Molecular interactions

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

The screenshot shows the Ensembl Fungi interface for the gene Mycgr3G65552. The left sidebar contains navigation links for Gene-based displays, Molecular interactions, and Molecular Interactions (highlighted). The main content area shows the gene details, including its UniProt ID (EWEW02), a putative uncharacterized protein, and cross-species interactions with *Triticum aestivum*. A callout box points to the 'Show metadata' button with the text 'Click on Show metadata to view more details'. Another callout box points to the interaction table with the text 'List of species and genes that Mycgr3G65552 interacts with'.

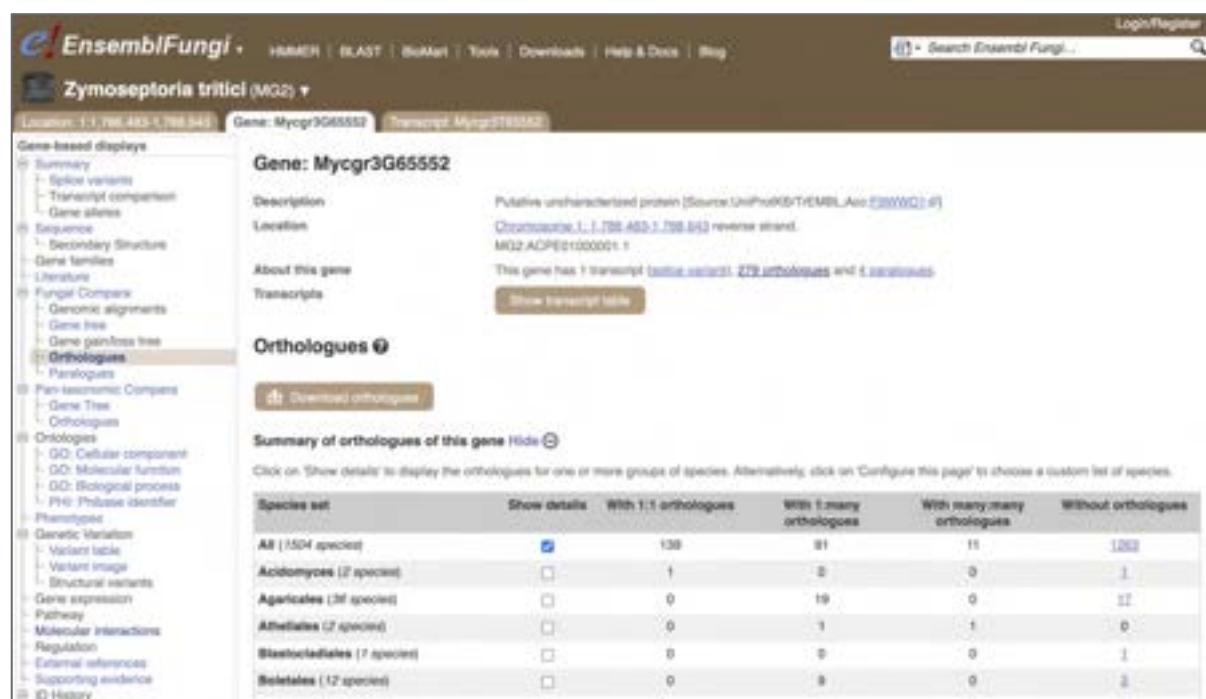
Can you find the wheat gene ID that Mycgr3G65552 interacts with? Look at the [Interacts with](#) table. The gene ID is 'UNDETERMINED'. This means a molecular interaction has been experimentally verified between Mycgr3G65552 and wheat, but the former gene has not been identified yet.

Interacts with					Show metadata
Species	Gene ID	Interactor	Identifier	Source DB	
Triticum aestivum	UNDETERMINED	protein	UNDETERMINED	PHI-base	

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

Interacts with					Show metadata
Species	Gene ID	Interactor	Identifier	Source DB	
Triticum aestivum	UNDETERMINED	protein	UNDETERMINED		PHI-base
Experimental evidence		gene complementation			
Interaction type		interspecies interaction			
Interaction phenotype		PHIPO:0000010			
Disease name		PHIDO:0000331			
Pathogen protein modification		gene deletion: full			
PHI-base high level term		Loss of pathogenicity			
Pathogen experimental strain		IPO323			
Host experimental strain		cv. Riband			

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



The screenshot shows the Ensembl Fungi interface for the gene `Mycgr3G65552` in the species `Zymoseptoria tritici (M02)`. The left sidebar contains a tree view of orthologous groups across various species. The main content area displays the gene details for `Mycgr3G65552`, including its description as a putative uncharacterized protein, its location on chromosome 1, and its transcript information. Below this, the 'Orthologues' section is shown, featuring a table of orthologous genes across different species groups. The table includes columns for the species set, the number of orthologues, and the type of orthology relationship (1:1, many-to-many, many-to-many). A 'Download orthologues' button is also present.

Species set	Show details	With 1:1 orthologues	With many orthologues	With many-many orthologues	Without orthologues
AB (1504 species)	<input checked="" type="checkbox"/>	139	81	11	1263
Ascomycota (27 species)	<input type="checkbox"/>	1	0	0	1
Agaricales (36 species)	<input type="checkbox"/>	0	19	0	17
Atheliales (27 species)	<input type="checkbox"/>	0	1	1	0
Blastocladiales (7 species)	<input type="checkbox"/>	0	0	0	1
Boletales (12 species)	<input type="checkbox"/>	0	8	0	3

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

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

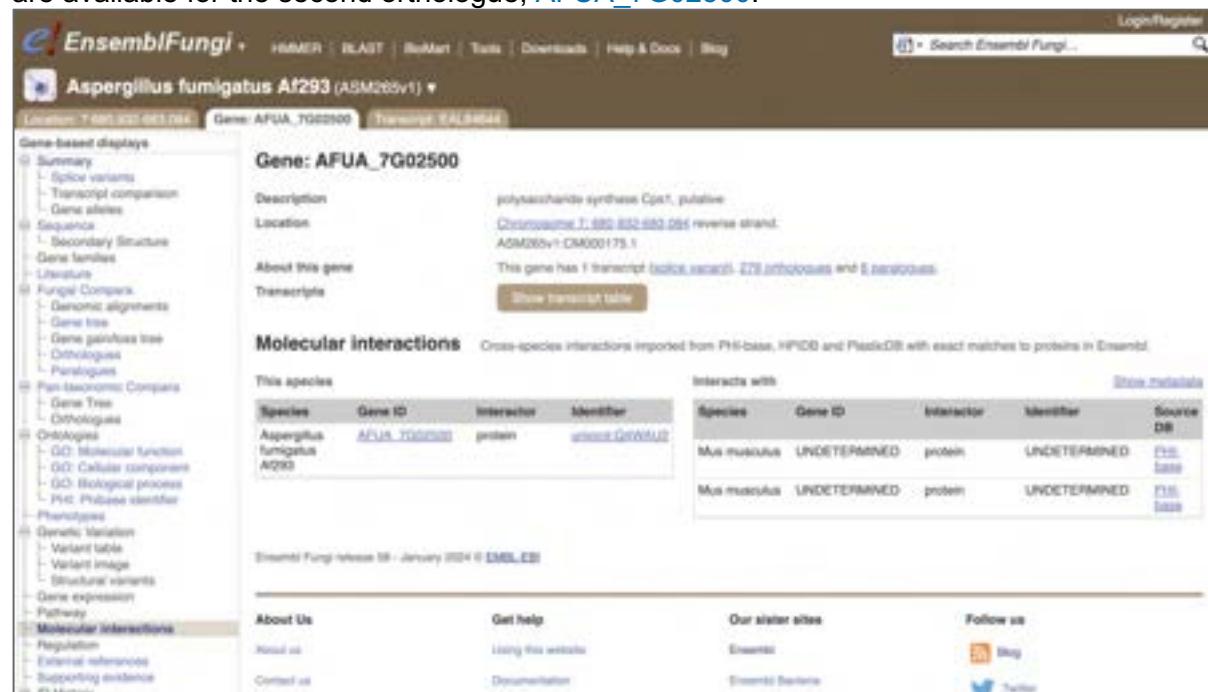
Orthologues 
 Download orthologues

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

Selected orthologues [Hide](#) 

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Aspergillus fumigatus Af160	1-to-1	AFUA_7G02500	54.37 %	42.25 %	n/a	n/a	Yes
		View Gene Tree		DS499601:878,603-680,755-1			
				View Sequence Alignments			
Aspergillus fumigatus Af293	1-to-1	AFUA_7G02500	54.37 %	42.25 %	n/a	n/a	Yes
		View Gene Tree		7,680,932-683,084-1			
				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](#).



The screenshot shows the Ensembl Fungi interface for the gene AFUA_7G02500. The left sidebar contains a tree-based navigation menu. The main content area displays the gene's details, including its description as a polysaccharide synthase CpxA1, putative, its location on chromosome 7, and its transcript information. Below this, the 'Molecular interactions' section is shown, featuring a table of interactions with mice. The table includes columns for Species, Gene ID, Interactor, and Identifier. Two entries are listed: one for Mus musculus with an unetermined identifier, and another for Mus musculus with an unassigned identifier. The bottom of the page includes links for 'About Us', 'Get help', 'Our sister sites', and social media links for 'Follow us'.

Species	Gene ID	Interactor	Identifier	Source DB
Aspergillus fumigatus Af293	AFUA_7G02500	protein	unassigned	FMS
Mus musculus	UNDETERMINED	protein	UNDETERMINED	FMS
Mus musculus	UNDETERMINED	protein	UNDETERMINED	FMS

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

Interacts with					Show metadata
Species	Gene ID	Interactor	Identifier	Source DB	
Mus musculus	UNDETERMINED	protein	UNDETERMINED	PHI-base	
Several experiments exist for this interaction. Please click here for more information					
Mus musculus	UNDETERMINED	protein	UNDETERMINED	PHI-base	
Experimental evidence	gene complementation				
Interaction type	interspecies interaction				
Interaction phenotype	PHIPO:0000015				
Disease name	PHIDO:0000020				
Pathogen protein modification	gene deletion: full				
PHI-base high level term	Reduced virulence				
Pathogen experimental strain	A/293				
Host experimental strain	C57BL/6				

The phenotype for the orthologue in mice is 'reduced virulence'.

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

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

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

Exercise 1 answers:

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



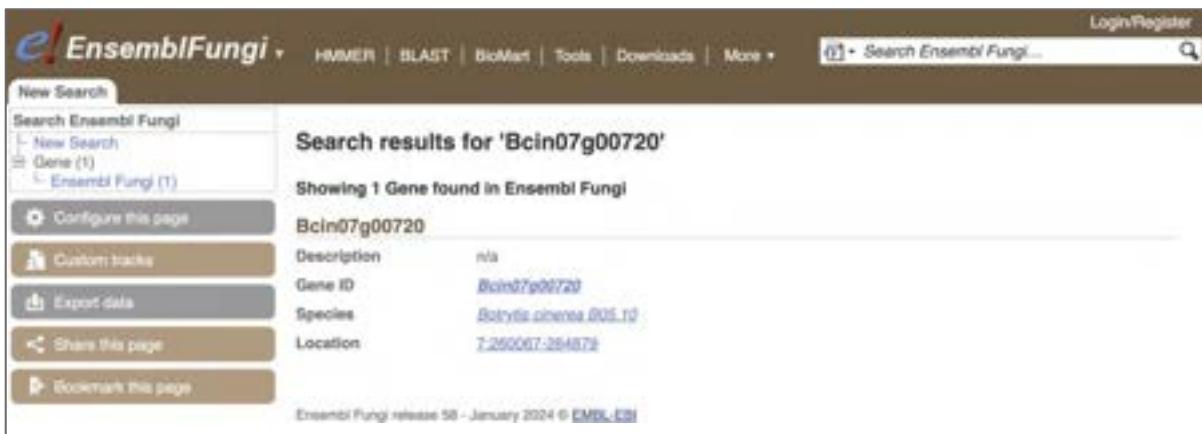
The screenshot shows the 'interactions_by_prodbname' endpoint documentation. At the top, there is a 'Try it out' button. Below it, the URL is shown as `GET /interactions_by_prodbname`. A note below the URL states: '(Returns all molecular interactions (lated by Ensembl species). The "Try it out" button might return an error if the response is too big but the endpoint will still work if tried on a browser.)'. Under 'Parameters', it says 'No-parameters'. At the bottom, there are 'Execute' and 'Clear' buttons.

In the ‘Response body’, search for `botrytis_cinerea`. Alternatively, you can open the request URL https://interactions.rest.ensembl.org/interactions_by_prodbname in your browser. You should get the following output:

```
"botrytis_cinerea": ["Bcin07g00720", "Bcin02g02570",
"Bcin12g04900", "Bcin16g00630", "Bcin02g06770", "Bcin03g07190",
"Bcin09g02390", "Bcin09g01800", "Bcin07g03050", "Bcin08g05150",
"Bcin10g01250", "Bcin14g01870", "Bcin06g04870", "Bcin06g00240",
"Bcin06g03440", "Bcin03g07900", "Bcin03g06840",
```

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

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

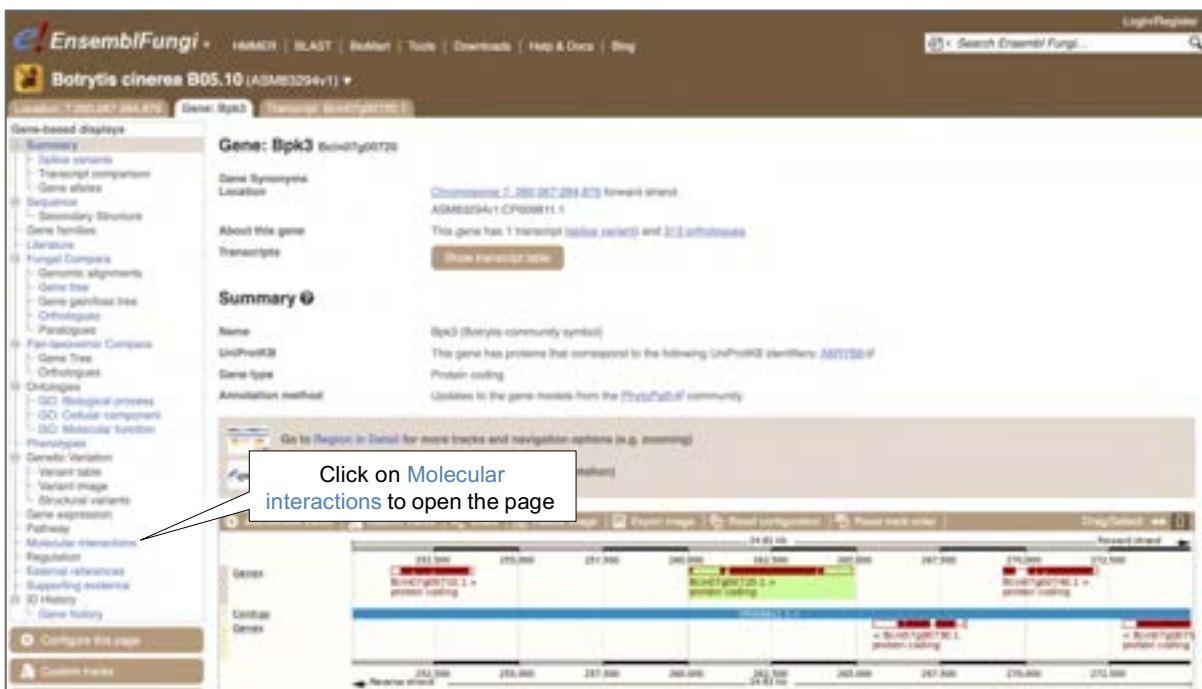


The screenshot shows the Ensembl Fungi homepage with a search bar at the top. Below it, a sidebar on the left contains buttons for 'New Search', 'Search Ensembl Fungi' (with options for 'New Search', 'Gene (1)', and 'Ensembl Fungi (1)'), 'Configure this page', 'Custom tracks', 'Export data', 'Share this page', and 'Bookmark this page'. The main content area displays search results for 'Bcin07g00720', showing 'Showing 1 Gene found in Ensembl Fungi'. The gene details are as follows:

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

At the bottom of the page, it says 'Ensembl Fungi release 58 - January 2024 © EMBL-EBI'.

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



The screenshot shows the Ensembl Fungi Gene page for **Bpk3** (Bcin07g00720). The left sidebar includes links for 'Summary', 'Transcript variants', 'Transcript comparisons', 'Gene alleles', 'Alleles', 'Secondary structures', 'Gene families', 'Libraries', 'Fungal Comparisons', 'Genomic alignments', 'Gene tree', 'Gene paralogs', 'Orthologs', 'Protein families', 'Protein-protein contacts', 'GO Cellular component', 'GO Molecular function', 'Phenotypes', 'Gene-based displays', 'Gene-based variants', 'Variant tables', 'Variant image', 'Structural variants', 'Gene expression', 'Pathway', 'Molecular interactions', 'Regulation', 'External references', 'Supporting evidence', 'History', and 'Gene history'. A callout box points to the 'Molecular interactions' link in the sidebar with the text 'Click on Molecular interactions to open the page'.

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

Molecular interactions Cross-species interactions imported from PHI-base, PHOdb and PlasmoDB with exact matches to proteins in Ensembl					Show metadata			
This species				Interacts with		Show metadata		
Species	Gene ID	Interaction	Identifier	Species	Gene ID	Interaction	Identifier	Source DB
Botryotinia cinerea 905.10	Bac07400720	protein	UNDETERMINED	Solanum lycopersicum	UNDETERMINED	protein	UNDETERMINED	PHI-base
				Vitis vinifera	UNDETERMINED	protein	UNDETERMINED	PHI-base
				Cucumis sativus	UNDETERMINED	protein	UNDETERMINED	PHI-base
				Matus domesticus	UNDETERMINED	protein	UNDETERMINED	PHI-base

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

Molecular interactions Cross-species interactions imported from PHI-base, PHOdb and PlasmoDB with exact matches to proteins in Ensembl					Show metadata			
This species				Interacts with		Show metadata		
Species	Gene ID	Interaction	Identifier	Species	Gene ID	Interaction	Identifier	Source DB
Botryotinia cinerea 905.10	Bac07400720	protein	UNDETERMINED	Solanum lycopersicum	UNDETERMINED	protein	UNDETERMINED	PHI-base
				Interaction type	Interspecies interaction			
				Interaction phenotype	PHOGO 0000015			
				Disease name	PHOGO 0000178			
				Pathogen protein modification	gene mutation, gene complementation			
				PHI-base high level term	Reduced virulence			
				Pathogen experimental strain	905.10			
				Vitis vinifera	UNDETERMINED	protein	UNDETERMINED	PHI-base
				Interaction type	Interspecies interaction			
				Interaction phenotype	PHOGO 0000015			
				Disease name	PHOGO 0000178			
				Pathogen protein modification	gene mutation, gene complementation			
				PHI-base high level term	Reduced virulence			
				Pathogen experimental strain	905.10			
				Cucumis sativus	UNDETERMINED	protein	UNDETERMINED	PHI-base
				Interaction type	Interspecies interaction			
				Interaction phenotype	PHOGO 0000010			
				Disease name	PHOGO 0000178			
				Pathogen protein modification	gene mutation, gene complementation			
				PHI-base high level term	Loss of pathogenicity			
				Pathogen experimental strain	905.10			
				Matus domesticus	UNDETERMINED	protein	UNDETERMINED	PHI-base
				Interaction type	Interspecies interaction			
				Interaction phenotype	PHOGO 0000015			
				Disease name	PHOGO 0000178			
				Pathogen protein modification	gene mutation, gene complementation			
				PHI-base high level term	Reduced virulence			
				Pathogen experimental strain	905.10			

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

EnsemblFungi · HAMMER · BLAST · BioMart · Tools · Downloads · Help & Docs · Blog · Login/Register · Search Ensembl Fungi ...

Botrytis cinerea B05.10 (ASM3294v1) · Gene: Bpk3 · Transcript: Bpk3g00795.1 · Location: 1:260,947,944-879

Gene-based displays

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

Gene: Bpk3 Bpk3g00795.1

Gene Synonyms

Location

Chromosome 1: 260,947,944-879 forward strand; ASM3294v1.CP008811.1

About this gene

This gene has 1 transcript (basic extent) and 213 orthologues.

Show transcript table

Click on Orthologues to open the page

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 (1154 species)	<input checked="" type="checkbox"/>	279	18	0	875
Achlya (7 species)	<input type="checkbox"/>	1	0	0	6
Agaricales (38 species)	<input type="checkbox"/>	4	0	6	32
Athelioidales (2 species)	<input type="checkbox"/>	1	0	0	1
Blastocladiales (7 species)	<input type="checkbox"/>	0	1	0	6
Botryotinia (12 species)	<input type="checkbox"/>	3	0	0	9

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

Orthologues

Download orthologues

Summary of orthologues of this gene Show

Selected orthologues Hide

Show	All	Others	Hide	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	Magnaporthe oryzae	High Confidence
Magnaporthe oryzae	1-to-1			ATG1 (MGG_06393)	49.90 %	51.47 %	n/a	n/a		Yes
				View Gene Tree	43,898,532-3,902,777>1					
				View Sequence Alignments						
Magnaporthe oryzae	1-to-1			M_BR02_EuGene_00042871	50.05 %	49.58 %	n/a	n/a		Yes
				View Gene Tree	BR02_scaffold00003.3,066,924-3,069,846>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 interaction information is available for the orthologue ATG1 ([MGG_06393](#)).

The Molecular interactions link is available for ATG1 (MGG_06393) in *Magnaporthe oryzae*

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

Molecular interactions					Cross-species interactions imported from Phrbase, HPODB and PintoDB with exact matches to proteins in Ensembl.				
This species				Interacts with				More metadata	
Species	Gene ID	Interactor	Identifier	Species	Gene ID	Interactor	Identifier	Source DB	
Magnaporthe oryzae 7-15	MGI_0000002	protein	protein	Hordeum vulgare	UNDETERMINED	protein	UNDETERMINED	DfBase	
				Dryja saliva	UNDETERMINED	protein	UNDETERMINED	DfBase	
				Dryja saliva	UNDETERMINED	protein	UNDETERMINED	DfBase	

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

Molecular interactions <small>Cross-species interactions imported from PPI-base, HPODB and PhenoDB with exact matches to proteins in Ensembl</small>								
This species		Interacts with			Show metadata			
Species	Gene ID	Interactor	Identifier	Species	Gene ID	Interactor	Identifier	Source DB
Botryotinia cinerea (BOS_10)	Bos07600288	protein	460011607128	Solanum lycopersicum	UNDETERMINED	protein	UNDETERMINED	PPI-base
				Interaction type	Interspecies interaction			
				Interaction phenotype	PPIPO_0000215			
				Disease name	PHOIO_00002178			
				Pathogen protein modification	gene mutation; gene complementation			
				PPI-base high-level term	Reduced virulence			
				Pathogen experimental strain	BOS_10			
				Vitis vinifera	UNDETERMINED	protein	UNDETERMINED	PPI-base
				Interaction type	Interspecies interaction			
				Interaction phenotype	PPIPO_0000215			
				Disease name	PHOIO_00002178			
				Pathogen protein modification	gene mutation; gene complementation			
				PPI-base high-level term	Reduced virulence			
				Pathogen experimental strain	BOS_10			
				Cucumis sativus	UNDETERMINED	protein	UNDETERMINED	PPI-base
				Interaction type	Interspecies interaction			
				Interaction phenotype	PPIPO_0000210			
				Disease name	PHOIO_00002178			
				Pathogen protein modification	gene mutation; gene complementation			
				PPI-base high-level term	Loss of pathogenicity			
				Pathogen experimental strain	BOS_10			
				Matus domesticus	UNDETERMINED	protein	UNDETERMINED	PPI-base
				Interaction type	Interspecies interaction			
				Interaction phenotype	PPIPO_0000215			
				Disease name	PHOIO_00002178			
				Pathogen protein modification	gene mutation; gene complementation			
				PPI-base high-level term	Reduced virulence			
				Pathogen experimental strain	BOS_10			

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

Molecular interactions <small>Cross-species interactions imported from PPI-base, HPODB and PhenoDB with exact matches to proteins in Ensembl</small>								
This species		Interacts with			Show metadata			
Species	Gene ID	Interactor	Identifier	Species	Gene ID	Interactor	Identifier	Source DB
Magnaporthe oryzae T0	MOGL_06000115	protein	4700011210002	Hordium vulgare	UNDETERMINED	protein	UNDETERMINED	PPI-base
				Several experiments exist for this interaction. Please click here for more information				
				Oryza sativa	UNDETERMINED	protein	UNDETERMINED	PPI-base
				Several experiments exist for this interaction. Please click here for more information				
				Oryza sativa	UNDETERMINED	protein	UNDETERMINED	PPI-base
				Interaction type	Interspecies interaction			
				Interaction phenotype	PPIPO_0000215			
				Disease name	PHOIO_00002178			
				Pathogen protein modification	gene deletion; full			
				PPI-base high-level term	Loss of pathogenicity			
				Pathogen experimental strain	Guy11			
				Host experimental strain	cv. CO-99			

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

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

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

EnsemblFungi • HOMER | BLAST | PMAKER | Tools | Downloads | Help & Docs | Log in

Botrytis cinerea B05.10 (ASMB3294v1) • Gene: Bpk3 [Transcript: Bpk3g01729]

Gene-based displays

- Summary
- Slice variants
- Transcript comparison
- Gene effects
- Sequence
- Secondary Structure
- Gene families
- Literature
- Fungal Comparisons
- Genomic alignments
- Gene tree
- Gene paralogs tree
- Orthologues
- Paralogues
- Pan-eukaryotic Comparisons
- Gene Tree
- Orthologues
- Ontologies
- GO: Cellular component
- GO: Molecular function
- GO: Biological process
- Phenotypes
- Genetic Variation
- Variant table
- Variant image
- Structural variants
- Gene expression
- Pathway
- Molecular interactions
- Regulation
- External references
- Supporting evidence
- GI History
- Gene history

Configure this page

Gene: Bpk3 Bpk3g01729

Gene Synonyms:

Location: Chromosome 2, 260,367-264,825 forward strand.
ATAGC294v1.CP008811.1

About this gene: This gene has 1 transcript (splice variant) and 23.3 kilobases.

Transcripts: [Show transcript table](#)

GO: Molecular function

Show/Hide columns (11 hidden)				
Annotation	Term	Evidence	Annotation source	Transcript ID
GO:0001180	nucleotide binding	EA	UniProt	Bpk3g01729L1
GO:0004672	protein kinase activity	EA	UniProt	Bpk3g01729L1
GO:0004673	protein serine/threonine kinase activity	EA	UniProt	Bpk3g01729L1
GO:0005524	ATP binding	EA	UniProt	Bpk3g01729L1
GO:0018811	kinase activity	EA	UniProt	Bpk3g01729L1
GO:0032057	transferase activity	EA	UniProt	Bpk3g01729L1
GO:0038020	protein serine kinase activity	EA	PhEA	Bpk3g01729L1

EnsemblFungi • HOMER | BLAST | PMAKER | Tools | Downloads | Help & Docs | Log in

Magnaporthe oryzae (MO) • Gene: ATG1 MO_08038

Gene-based displays

- Summary
- Slice variants
- Transcript comparison
- Gene effects
- Sequence
- Secondary Structure
- Gene families
- Literature
- Fungal Comparisons
- Genomic alignments
- Gene tree
- Gene paralogs tree
- Orthologues
- Paralogues
- Pan-eukaryotic Comparisons
- Gene Tree
- Orthologues
- Ontologies
- GO: Cellular component
- GO: Molecular function
- GO: Biological process
- Phenotypes
- Genetic Variation
- Variant table
- Variant image
- Structural variants
- Gene expression
- Pathway
- Molecular interactions
- Regulation
- External references
- Supporting evidence
- GI History
- Gene history

Configure this page

Gene: ATG1 MO_08038

Description: Beta-tubulin-like protein kinase ATG1 (Source: UniProtKB/Swiss-Prot; ID: Q51233v1)

Location: Chromosome 6, 3,898,532-3,902,277 reverse strand.
MO_08038

About this gene: This gene has 1 transcript (splice variant) and 21.5 kilobases.

Transcripts: [Show transcript table](#)

GO: Molecular function

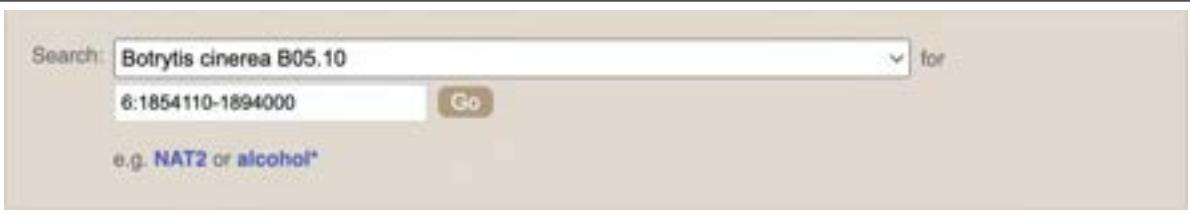
Show/Hide columns (11 hidden)				
Annotation	Term	Evidence	Annotation source	Transcript ID
GO:0001180	nucleotide binding	EA	UniProt	MO_08038T1
GO:0004672	protein kinase activity	EA	UniProt	MO_08038T1
GO:0004673	protein serine/threonine kinase activity	MP	UniProt	MO_08038T1
GO:0005524	ATP binding	EA	UniProt	MO_08038T1
GO:0018811	kinase activity	EA	UniProt	MO_08038T1
GO:0032057	transferase activity	EA	UniProt	MO_08038T1
GO:0038020	protein serine kinase activity	EA	PhEA	MO_08038T1

Exercise: Attaching Track Hubs to Ensembl Fungi

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

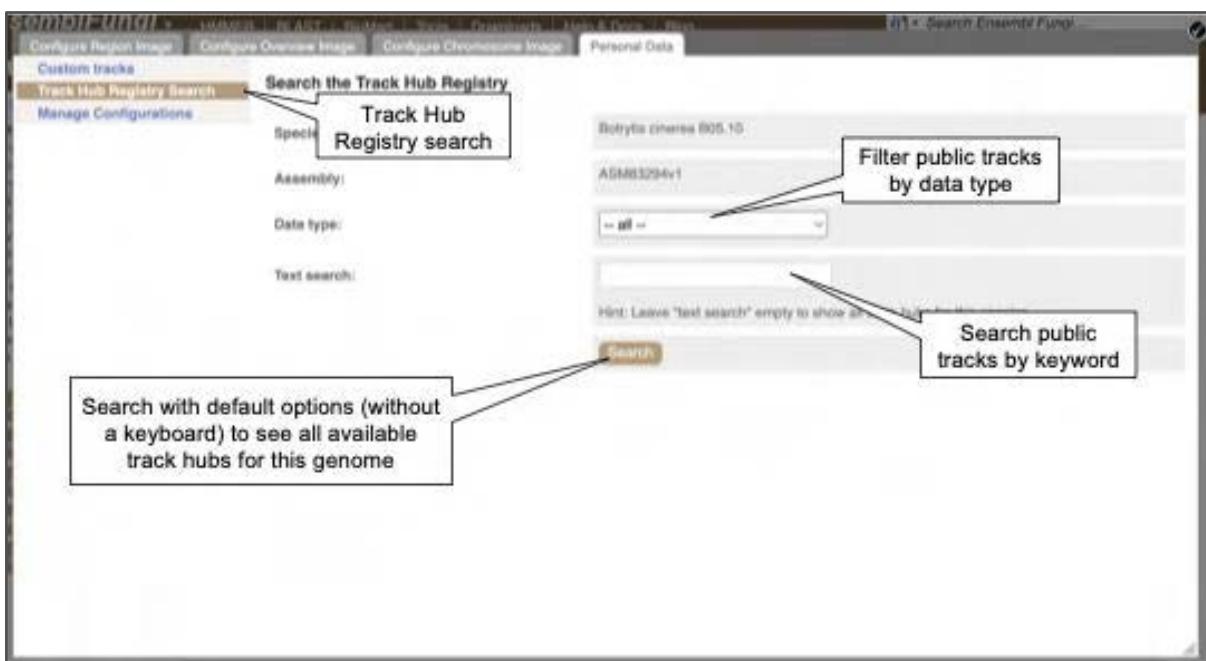
There are a number of publicly available datasets that are available to add to views in Ensembl. You can find full lists of these at <https://trackhubregistry.org/>. We're going to search and add these files from within Ensembl.

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

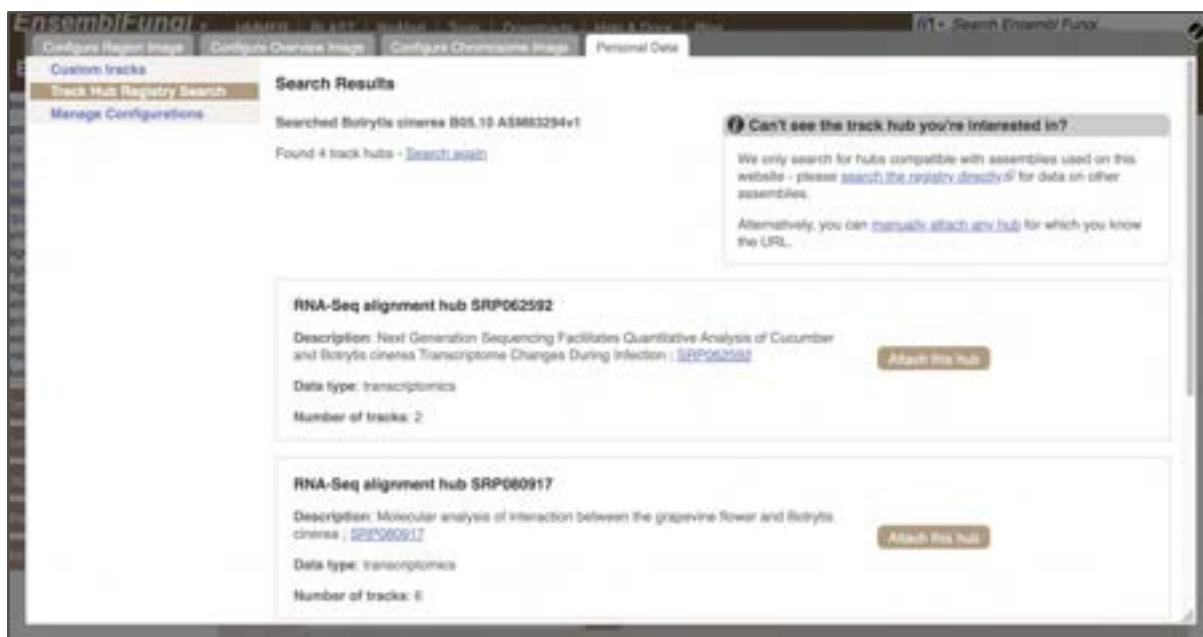


Search: Botrytis cinerea B05.10
 6:1854110-1894000 Go
 e.g. NAT2 or alcohol*

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. In the pop-up menu, click on **Track Hub Registry Search** on the left-hand navigation panel.



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



Ensembl/Final

Search Results

Searched Botrytis cinerea B05.10 ASMR03294v1

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

Can't see the track hub you're interested in?

We only search for hubs compatible with assemblies used on this website - please [search the registry directly](#) for data on other assemblies.

Alternatively, you can [manually attach any hub](#) for which you know the URL..

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 SRP060917

Description: Molecular analysis of interaction between the grapevine flower and *Botrytis cinerea* : [SRP060917](#)

Data type: transcriptomics

Number of tracks: 0

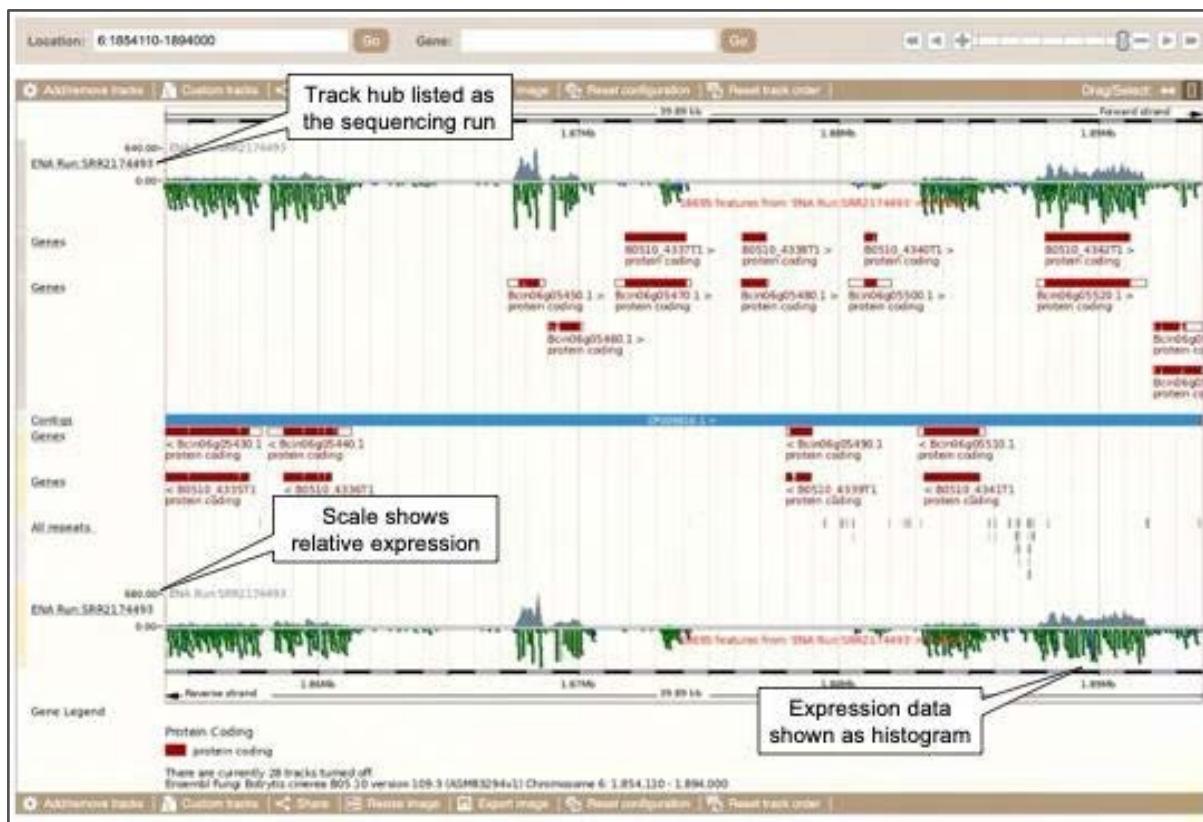
Attach this hub

Attach this hub

There are four available track hubs for this assembly.

Choose the 'RNA-Seq alignment hub SRP062592' by clicking on the [Attach this hub](#) button on the right. It is a next-generation sequencing (NGS) quantitative analysis of cucumber and *B. cinerea* transcriptome changes during infection. Close the pop-up window.

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



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



- (a) Go to www.trackhubregistry.org on your browser and search for [SRP062592](#). Can you jump to Ensembl Fungi directly from the Track Hub Registry page?

The Track Hub Registry

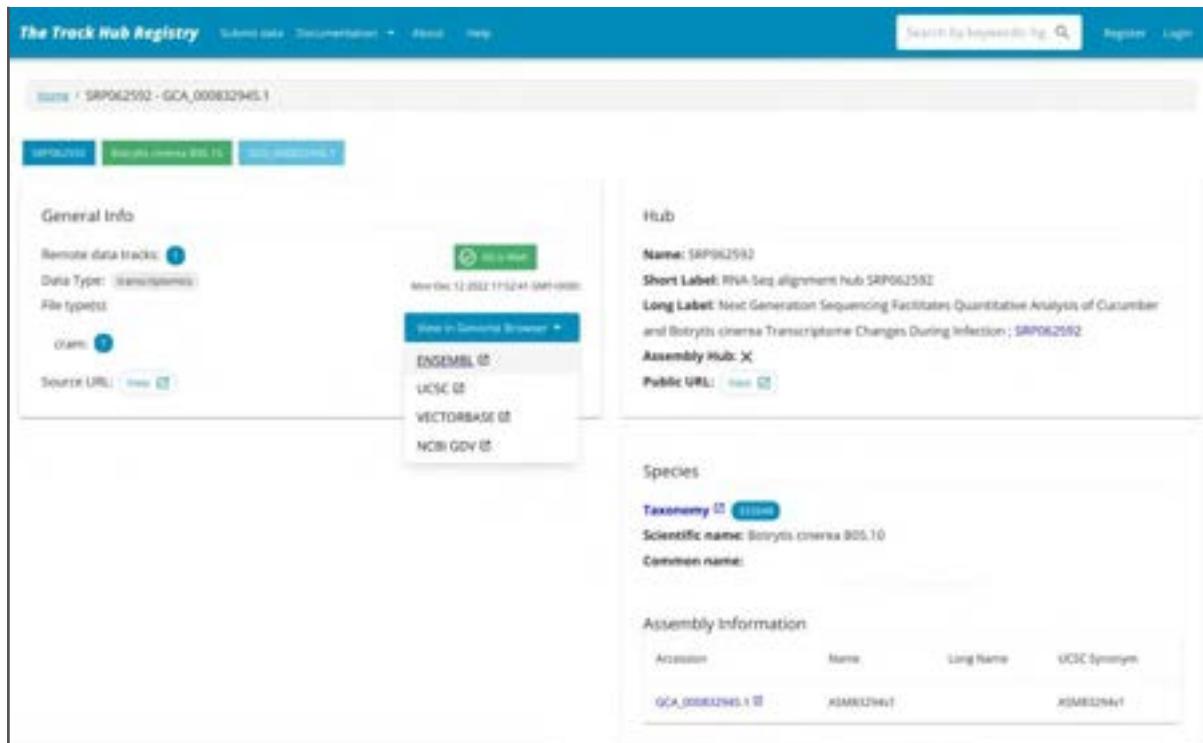
A global centralised collection of publicly accessible track hubs

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



Register | Log in



The screenshot shows the SRP062592 hub details page. At the top, there's a navigation bar with links for 'Submit data', 'Documentation', 'About', and 'Help'. A search bar is also present. Below the header, the URL 'https://www.trackhubregistry.org/submit/SRP062592' is shown. The main content area is divided into sections: 'General Info', 'Hub', 'Species', and 'Assembly Information'. The 'General Info' section includes fields for 'Remote data tracks' (with a count of 1), 'Data Type' (BAM/BAM/FASTQ), 'File type(s)' (bam), and 'Source URL' (with a link to 'View in Genome Browser'). The 'Hub' section provides detailed information: 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), Assembly Hub (X), and Public URL (with a link). The 'Species' section lists Taxonomy (Botrytis cinerea 805.10), Scientific name (Botrytis cinerea 805.10), and Common name. The 'Assembly Information' section shows a table with columns for Accession, Name, Long Name, and UCSC Synonym, listing GCA_000832945.1 and X3M8L2H4v1.

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

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

Let's take a look at that URL.

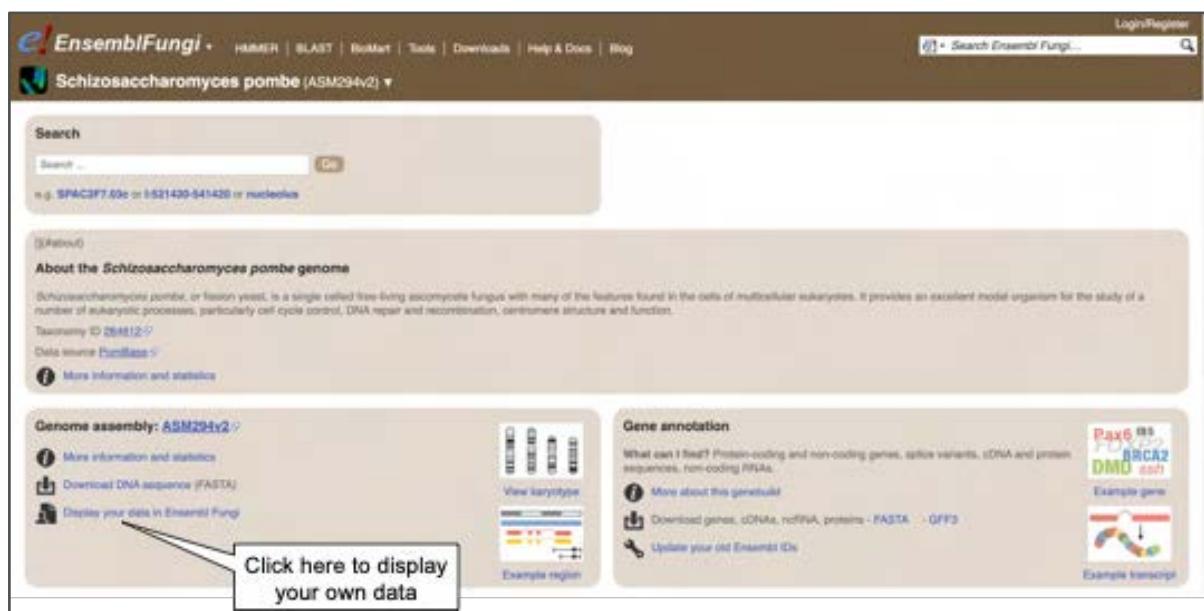
NOTE: Many internet browsers have recently dropped support for FTP, including the latest Firefox and Google Chrome versions. Firefox v87.0 still contains built-in FTP implementation. If you struggle to open the FTP site, try the HTTP version:
https://ftp.ebi.ac.uk/ensemblgenomes/pub/misc_data/bam/fungi/Spom/

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

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

Here you can see two BAM files (file names ending in '.bam') with corresponding index files (file names ending in '.bam.bai'). We're interested in the files [Spom_all_61G9EAAXX_and_61G9UAAXX.+sorted.bam](#) and [Spom_all_61G9EAAXX_and_61G9UAAXX.+sorted.bam.bai](#). These files are the BAM file and the index file respectively. When attaching a BAM file to Ensembl, there must be an index file in the same folder.

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



The screenshot shows the Ensembl Fungi homepage for *Schizosaccharomyces pombe* (ASM294v2). The top navigation bar includes links for HOMER, BLAST, BioMart, Tools, Downloads, Help & Docs, and Blog. The search bar contains the query "SPAC3F7.60c or I521429-541420 or nucleolus". Below the search bar, a "About the Schizosaccharomyces pombe genome" section provides a brief overview of the organism. The "Genome assembly: ASM294v2" section includes links for "More information and statistics", "Download DNA sequence (FASTA)", and "Display your data in Ensembl Fungi". A callout box points to the "Display your data in Ensembl Fungi" link. The "Gene annotation" section features a diagram showing genes (Pax6, POF3, BBC1, DMD) and their mRNA and protein products. Other links in this section include "More about this genome!", "Download genes, cDNAs, ncRNAs, proteins - FASTA", and "Update your old Ensembl IDs".

A menu will appear:

Personal Data Custom tracks Track Hub Registry Search Manage Configurations

Add a custom track

Please note that track hubs and indexed files (BAM, BigBed, etc) do not work with certain cloud services, including Google Drive and Dropbox. Please see our [support page](#) for more information.

Name for this data (optional):

Species: **Schizosaccharomyces pombe**
Assembly: ASM294v2

Data:

Or upload file (max 20MB): No file chosen

Data format: **BAM**

[Help on supported formats, display types & file formats](#)

Click [Add data](#) to view your track 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](#), select the file from your local machine and it should automatically detect the file type you have submitted.

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

(http://ftp.ensemblgenomes.org/pub/misc_data/bam/fungi/Spom/Spom_all_61G9EAAXX_and_61G9UAAXX.+sorted.bam).

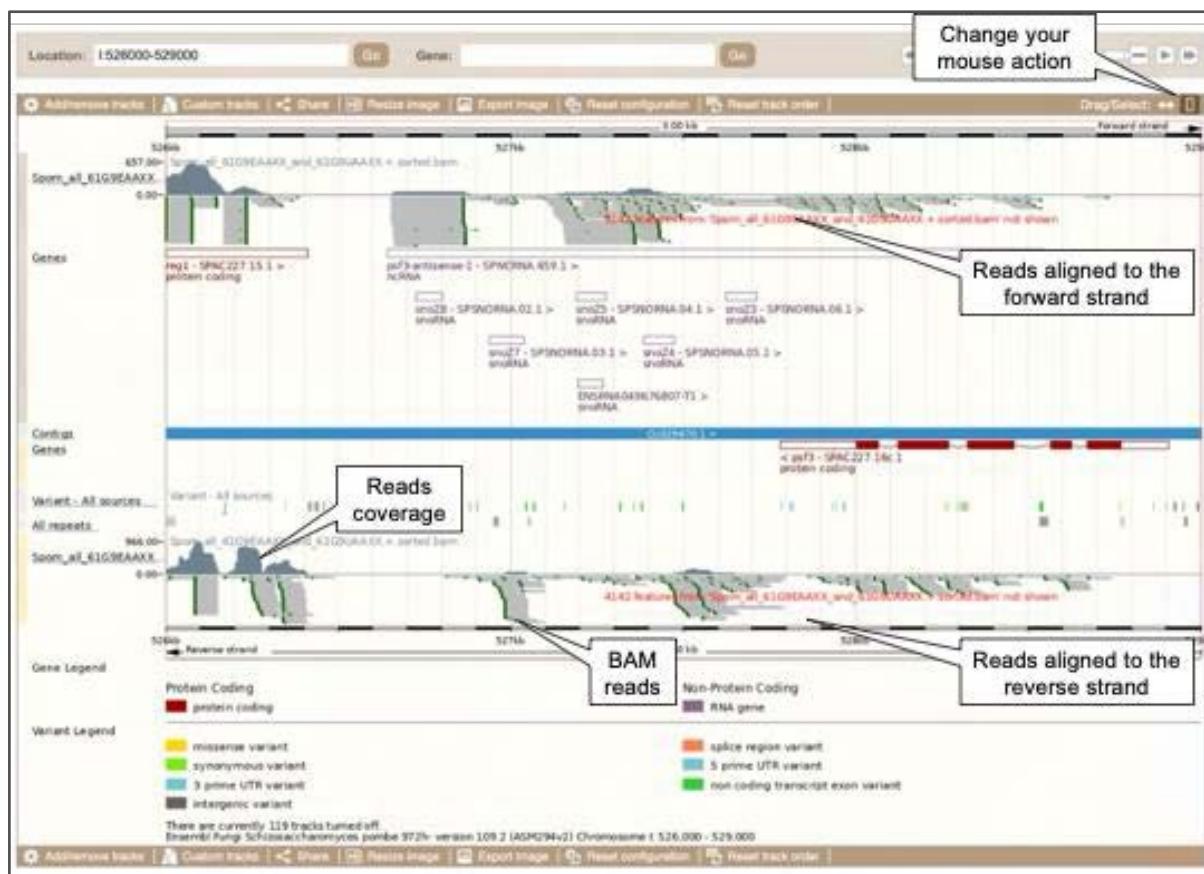
Since this is a file, the interface is able to detect the '.bam' file extension and automatically labels the format as **BAM**. Click on [Add data](#) and close the menu. It may take a while to load as there is a lot of data (Firefox tends to be fast). Once the data has been uploaded, you'll get a thank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

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

Schizosaccharomyces pombe (ASM294v2) ▾

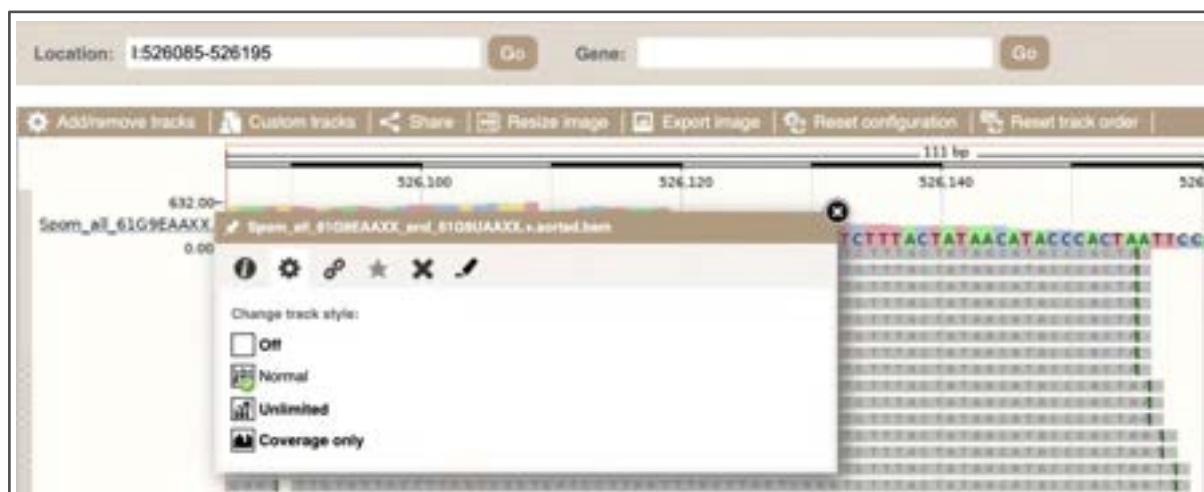
Search

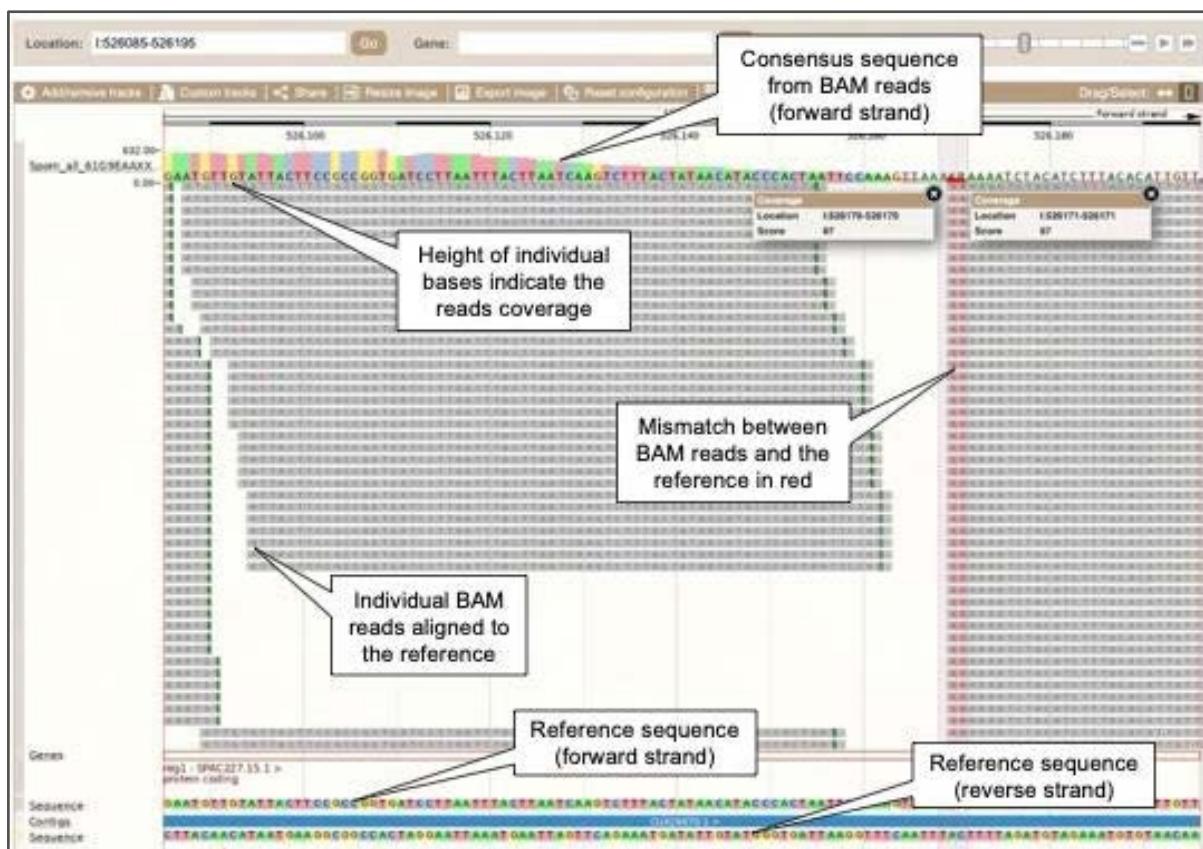
e.g. [SPAC2F7.03c](#) or [I:521420-541420](#) or [nucleolus](#)



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

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







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's a navigation bar with links for About, Blog, Download, Help, YeastDB, and social media icons. Below the navigation is a search bar with the query "Q: glycolysis". To the right of the search bar, there's a "Show all results..." link. On the left, there's a thumbnail image of yeast cells. In the center, there's a section titled "About SGD" with a brief description of the database. On the right, there's a sidebar with a list of biological processes related to glycolysis, each accompanied by a green circular icon. At the bottom of the sidebar, there are two references: "Goncalves P and Planta RJ (1998)" and "Starting up yeast glycolysis, Trends Microbiol 6(8):314-9".

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

The screenshot shows the SGD results page for the query "glycolysis". On the left, there's a sidebar with categories: References (orange arrow), Genes (blue arrow), Biological Processes, Downloads, Molecular Functions, Cellular Components, and Chemicals. The "Genes" category is selected. The main content area shows "644 results for 'glycolysis'" and "Page 1 of 26". There are buttons for Results (25), Sort By (Relevance), and a "Show more" button. Below this, there are two sections: "canonical glycolysis" (5 results) and "glycolysis from storage polysaccharide through" (2 results). Each section has a brief description and a green circular icon indicating it's a biological process.

- 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 results page for the query "glycolysis" with the "Genes / Genomic Features" category selected (blue arrow). The main content area shows "15 results for 'glycolysis'". There are buttons for Download (grey) and Analyse (grey). To the right, there are two blue buttons: "Wrap" and "Download" (orange arrow). Below this, there's a table of 15 genes: GPM1, PGK1, ENO1, TDH1, FBA1, ENO2, PFK1, PFK2, TDH3, CDC19, TDH2, TP1, PGI1, GLK1, HOK1. At the bottom, there's a "Show more" link.



- 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:

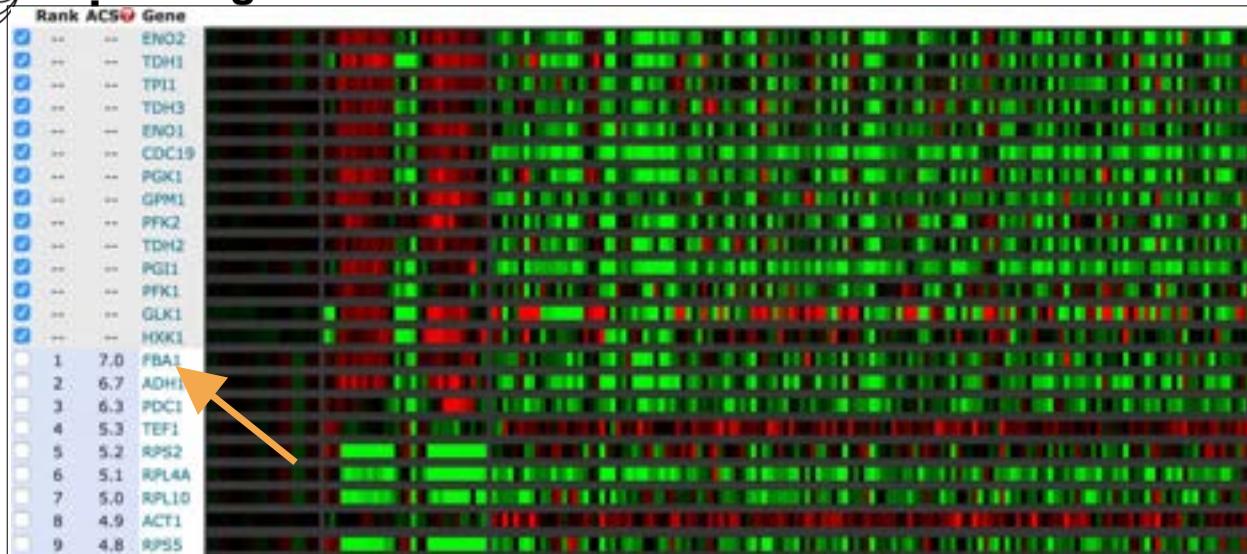
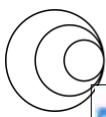
- To access SPELL, go to the SGD home page at www.yeastgenome.org, open the **Function** tab on top of the page and click on **Expression**. Or, if you are already on a Locus Summary page, open the Expression tab and click on the SPELL link under the histogram.

The screenshot shows the SGD home page. At the top, there's 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's a banner image of yeast cells with green and red fluorescence. To the right of the banner is a dropdown menu titled "Expression" which is currently expanded. Other items in the menu include Gene Ontology, Biochemical Pathways, Phenotypes, Interaction, YeastGDP, and Resources. A yellow arrow points to the "Expression" menu item. To the right of the menu, a modal window titled "About SGD" is displayed, containing text about the database and a "Try this?" button. Below the banner and menu, there are sections for Meetings (listing the 31st Vrije Universiteit Brussel Yeast Conference) and New & Noteworthy (listing "Trouble with Tripletts" from April 06, 2018).

- 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. The title is "SPELL - *S. cerevisiae*". Below it is a purple header bar with the text "SPELL (Yeast)". Underneath is a search form with the "Gene Name(s)" field containing "GPM1 PGK1 PFK1 PFK2 ENO2 ENO1 CDC". To the right of the field is a "Search" button and a "# Results: 20" indicator. 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	% query	% genome
glucose catabolic process (biological_process)	1.33e-29	19 of 35	52 of 6381
hexose catabolic process (biological_process)	2.39e-29	19 of 35	59 of 6381
monosaccharide catabolic process (biological_process)	2.91e-27	19 of 35	66 of 6381
glycolysis (biological_process)	4.79e-27	16 of 35	32 of 6381
glucose metabolic process (biological_process)	1.86e-23	19 of 35	99 of 6381
single-organism carbohydrate catabolic process (biological_process)	3.62e-22	19 of 35	115 of 6381
hexose metabolic process (biological_process)	4.32e-22	19 of 35	116 of 6381
monosaccharide metabolic process (biological_process)	1.42e-21	19 of 35	123 of 6381
carbohydrate catabolic process (biological_process)	1.97e-21	19 of 35	125 of 6381
generation of precursor metabolites and energy (biological_process)	7.80e-18	19 of 35	190 of 6381
single-organism carbohydrate metabolic process (biological_process)	1.60e-13	19 of 35	319 of 6381
gluconeogenesis (biological_process)	3.72e-13	10 of 35	33 of 6381
hexose biosynthetic process (biological_process)	5.25e-13	10 of 35	34 of 6381
monosaccharide biosynthetic process (biological_process)	7.33e-13	10 of 35	35 of 6381
Annotated Genes			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, HKK1, HKK2, PKK2, GLK1, GPM1, PFK1, TP1, TDH1, PGK1, ENO2, PGK1, ACH1, PGK2, CDC19, PGK3, TDK2			
FBA1, TDH3, ENO1, GPM1, TP1, TDH1, PGK1, ENO2, PGK1, TDK2			

Exploring transcriptomics & proteomics datasets in FungiDB

Learning objectives:

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

Transcriptomics.

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

Legend:  Coexpression  Similarity  Differential Expression  Fold Change  MetaCycle  Percentile  SenseAntisense

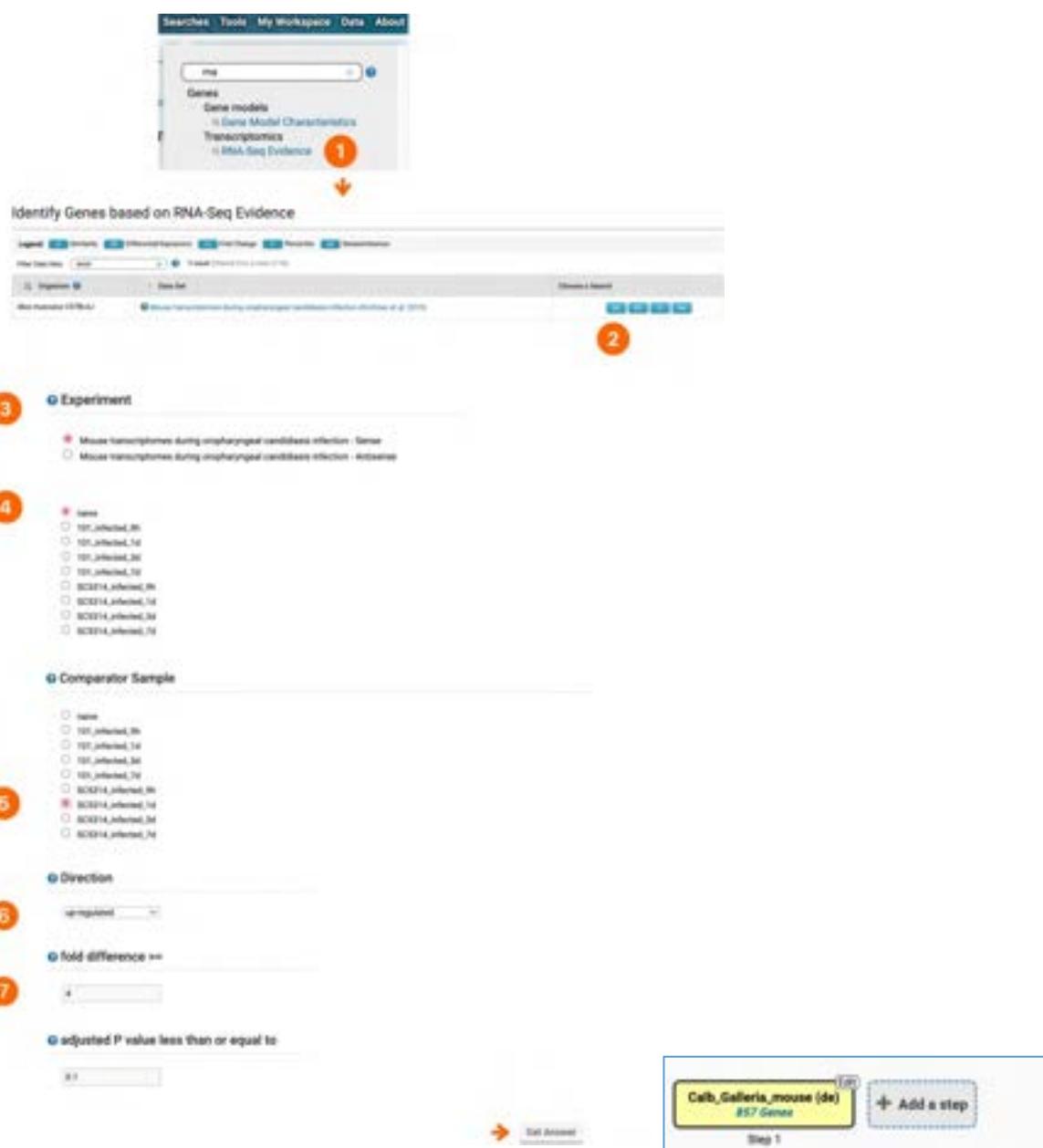
- **Coexpression.** Search for genes which have positive or negative correlations with a set of genes.
- **Similarity.** Search for genes which have a similar profile for an experiment.
- **Differential Expression (DE).** This search uses DESeq2 analysis results. You can choose the directionality and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

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

Objective: Identify differentially expressed genes in mice (HostDB.org) and Candida albicans SC5314 (FungiDB.org) during infection (1d).

A. The next block of exercises will be carried out in [HostDB.org](#)

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



The screenshot shows the HostDB.org interface for identifying differentially expressed genes based on RNA-Seq Evidence. The steps are numbered as follows:

1. In the main navigation bar, click on "RNA-Seq Evidence".
2. In the search interface, click on the "DE" button.
3. Under "Experiment", select "Mouse transcriptomes during oropharyngeal candida infection - Sense".
4. Under "Reference Sample", select "Naive".
5. Under "Comparator Sample", select "SC5314_infected_1d".
6. Under "Direction", select "up-regulated".
7. Under "fold difference >=", enter "4".
8. Under "adjusted P value less than or equal to:", enter ".05".
9. Click the "Search" button.
10. The results show "Calb, Galleria, mouse (de) #57 Genes".



- Identify host genes up-regulated by the SC5314 strain but not 101 at 1d of infection.

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

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

The screenshot shows the QIIME pipeline interface with the following steps:

- Step 1:** A box labeled "Combine with other Genes" with a "1 minus 2" button highlighted.
- Step 2:** A box labeled "Identify Genes based on RNA-Seq Evidence". It includes:
 - A legend: "Legend: DE (downregulated) Inferred (inferred) DE (upregulated) Inferred (inferred)"
 - "This Data Set: 101" and "Reads: 101 reads from 101 samples"
 - "Reference Sample": "naive" selected, with other options like "101_infected_0h", "101_infected_1d", etc.
 - "Comparator Sample": "101_infected_1d" selected, with other options like "naive", "101_infected_0h", etc.
 - "Direction": "upregulated" selected.
 - "fold difference >= 4" selected.
 - "adjusted P value less than or equal to 0.05" selected.
- Step 3:** A box showing the results of the search, with a legend:
 - Red dot: "Mouse transcriptomes during respiratory candidiasis infection - Sense"
 - Green dot: "Mouse transcriptomes during respiratory candidiasis infection - Antisense"
- Step 4:** A box showing the "Combine & Run" step, which has been completed successfully.

- Examine the results in HostDB:

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

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



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

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



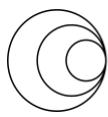
B. The next block of exercises will be carried out in [FungiDB.org](#).

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

The screenshot shows the FungiDB RNA-Seq Evidence search interface. The steps are numbered 1 through 7:

1. A dropdown menu is open, showing options like "Series", "Gene models", "Gene models", "Transcriptome", and "Transcripts Evidence".
2. The "DE" button is highlighted.
3. The "Sense" radio button is selected.
4. The "Reference Sample" section is shown, with "SC5314_in vitro" selected.
5. The "Comparator Sample" section is shown, with "SC5314_infected_1d" selected.
6. The "Direction" dropdown is set to "up-regulated".
7. The "fold difference >=" input field contains the value "4".

At the bottom right, there is a yellow box labeled "Calb_Kirchner_mouse (de) SFP Genes" and a blue dashed box labeled "+ Add a step Step 1".



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

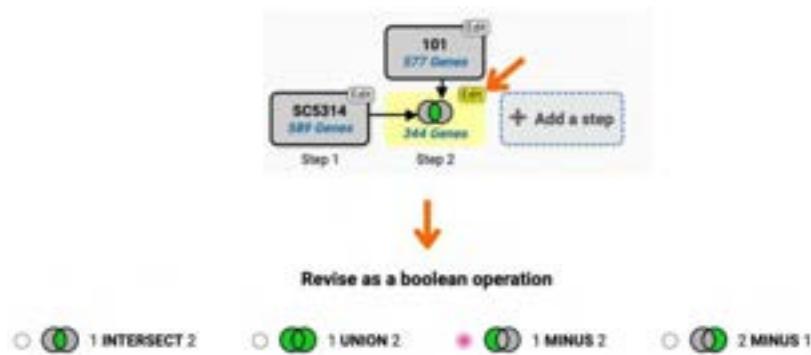
1. Click on the “Add Step” button.
2. Navigate to the RNA-Seq Evidence search, filter for “Kirch” to quickly identify the dataset, and click the “DE” button.
3. Choose to examine the sense strand.
4. Select reference sample: 101_in vitro.
5. Select comparator sample: 101_infected_1d.
6. Look for up-regulated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

The screenshot shows a step-by-step process for identifying genes up-regulated in SC5314 but not 101 strain at 1d of infection. The steps are numbered 1 through 7.

- Step 1:** A yellow box labeled "Cell_Kirchmar_mouse (96) 377 dataset" has an "Add a step" button next to it. A red circle with the number 1 is above the button.
- Step 2:** The user is in the "Compare with other Genes" section. They have selected "Mouse transcriptions during oropharyngeal candidiasis infection in mouse - Sense" as the reference and "Mouse transcriptions during oropharyngeal candidiasis infection in mouse - Antisense" as the comparator. A red circle with the number 2 is on the left side of the screen.
- Step 3:** The user is in the "Identify Genes based on RNA-Seq Evidence" section. A red circle with the number 3 is on the left side of the screen.
- Step 4:** Under "Reference Sample", the "101_in vitro" option is selected. A red circle with the number 4 is on the left side of the screen.
- Step 5:** Under "Comparator Sample", the "101_infected_1d" option is selected. A red circle with the number 5 is on the left side of the screen.
- Step 6:** Under "fold difference >= ", a value of "4" is entered. A red circle with the number 6 is on the left side of the screen.
- Step 7:** Under "adjusted P value less than or equal to", a value of "0.1" is entered. A red circle with the number 7 is on the left side of the screen.

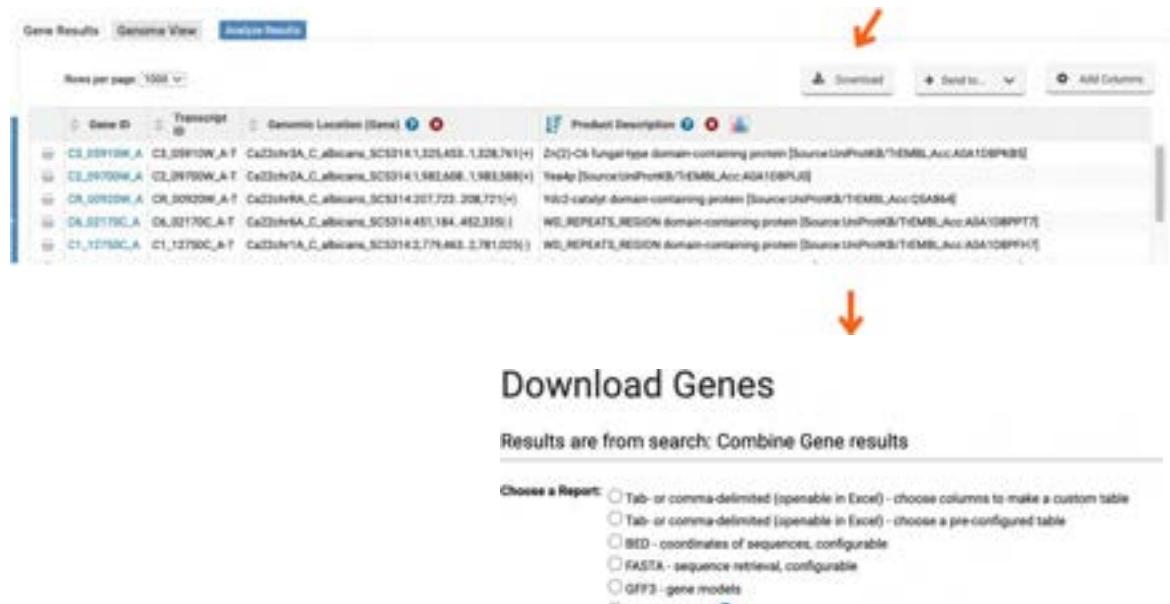
At the bottom right, there is a "Get Answer" button and a summary diagram showing the flow from "Cell_Kirchmar_mouse (96) 377 dataset" to "Cell_Kirchmar_mouse (96) 377 dataset" with a green circle labeled "446 genes".

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



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

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



Gene Results | Genome View | Analyze Results

Rows per page: 1000 ✓

Gene ID	Transcript ID	Genomic Location (Sanger)	Product Description	
CE_00910W_A	CE_00910W_A-T	Ca23chr3A_C_allelicane_SC5314;1,325,453..1,328,761(+)	Zn(II)-Ca_fungal-type domain-containing protein [Source-UniProtKB/ThEMBL_Acc:Q8A1D8PF03]	
CE_00910W_A	CE_00910W_A-T	Ca23chr3A_C_allelicane_SC5314;1,983,468..1,983,589(+)	YsdAp [Source-UniProtKB/ThEMBL_Acc:Q8A1D8PF03]	
CR_00920W_A	CR_00920W_A-T	Ca23chr3A_C_allelicane_SC5314;207,723..208,721(+)	Wtcl-carboxyl domain-containing protein [Source-UniProtKB/ThEMBL_Acc:Q8A1D8PF03]	
DA_00778CA	DA_00778CA-T	Ca23chr3A_C_allelicane_SC5314;481,184..482,335(+)	WIG_REPEATS_REGION domain-containing protein [Source-UniProtKB/ThEMBL_Acc:Q8A1D8PF03]	
CI_12750C_A	CI_12750C_A-T	Ca23chr3A_C_allelicane_SC5314;2,779,463..2,781,025(+)	WIG_REPEATS_REGION domain-containing protein [Source-UniProtKB/ThEMBL_Acc:Q8A1D8PF03]	

Download

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
- BED - coordinates of sequences, configurable
- FASTA - sequence retrieval, configurable
- GFF3 - gene models
- Standard JSON

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

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



Proteomics

Objective: Query proteomics data for *Candida albicans*.

Fungal extracellular vesicles (EVs) have been implicated in host-pathogen and pathogen-pathogen communication. In this exercise, we will query quantitative mass spec data, compare protein abundance in EVs vs. whole cell lysate (WCL) in biofilm conditions, and cross-reference these results with the RNA-Seq evidence search created above.

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

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

Identify Genes based on Quantitative Mass Spec. Evidence

Legend: Direct Comparison Fold Change

Filter Data Sets: albicans 1 result (based on a total of 11)

Organism: Data Set: Choose Search

Calbicans albicans (Extracellular vesicle and whole cell lysate proteomes for DAY226 yeast/biofilm, ATCC90028 and ATCC10231 strains. (Dawson et al. 2020))

For the Experiment

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

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

DAY286 biofilm EV mean (5) DAY286 biofilm WCL mean ATCC90028 yeast EV mean ATCC90028 yeast WCL mean ATCC10231 yeast EV mean ATCC10231 yeast WCL mean

select all | clear all

and its average expression value in the following Comparison Samples (6)

DAY286 yeast WCL mean DAY286 biofilm EV mean DAY286 biofilm WCL mean ATCC90028 yeast EV mean ATCC90028 yeast WCL mean ATCC10231 yeast EV mean

select all | clear all

Calbicans_EVs (fc) 324 Genes

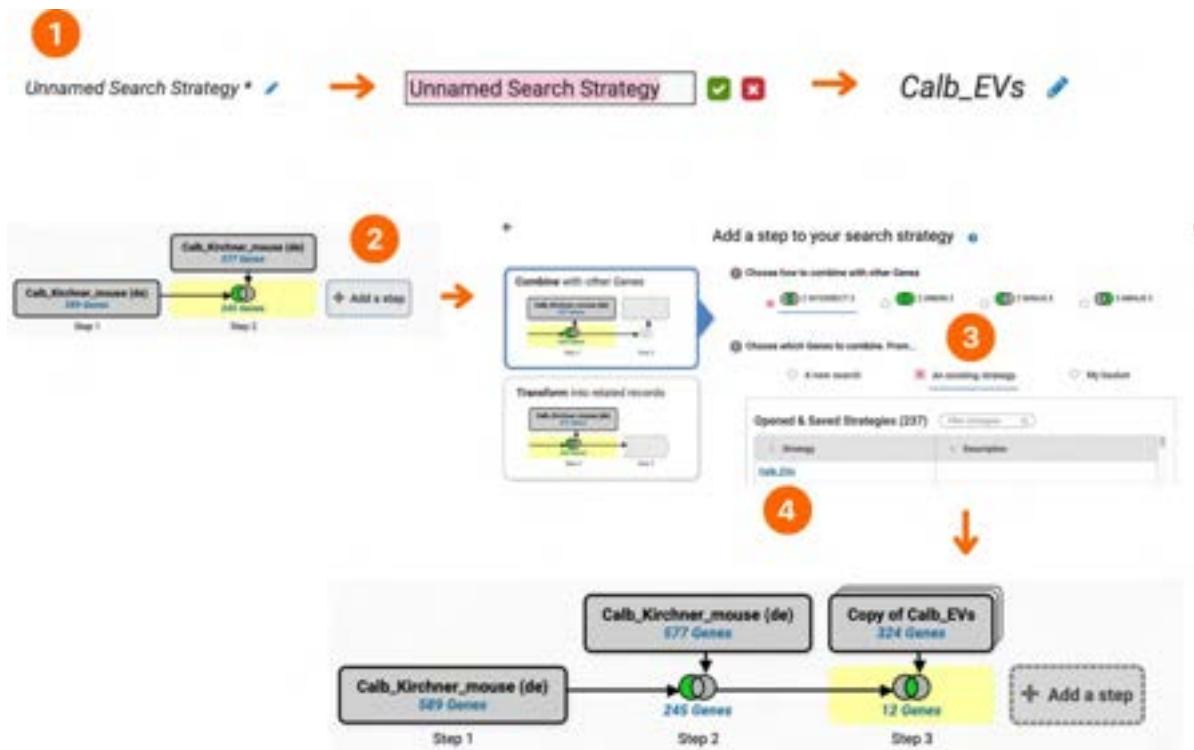
+ Add a step

Step 1



- Identify genes that are upregulated in SC5314 during infection and present in the EVs samples when Candida is grown in biofilm condition.

1. Give your proteomics search a name (e.g. Calb_EVs)
2. Click on the previous RNA_Seq search to activate it and “Add a step”
3. Select to add an existing strategy
4. Click on the Calb_EVs proteomics strategy to import the results.



Strategy URL:

Calb_EVs: <https://fungidb.org/fungidb/app/workspace/strategies/import/c971467cff5062fa>

Transcriptomics & Proteomics query:

<https://fungidb.org/fungidb/app/workspace/strategies/import/18b90faee40fe0db>

References.

1. Hughes ME, Hogenesch JB, Kornacker K. 2010. JTK_CYCLE: an efficient non-parametric algorithm for detecting rhythmic components in genome-scale datasets. *J Biol Rhythms*. 25(5):372–380. doi:10.1177/0748730410379711.
2. Yang R, Su Z. 2010. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*. 26(12):i168–i174. doi:10.1093/bioinformatics/btq189.

Exploring Gene Models in JBrowse

Learning objectives:

- Leverage omics data (e.g., RNA-Seq) to evaluate gene models.
- Determine if a gene model is accurate or if alternate models are possible

In previous exercises, you spent some time learning about gene pages and examining genes in the context of the JBrowse genome browser. It is important to recognize that gene models (structural annotation) are often open to interpretation, especially with respect to:

- transcript initiation and termination sites (5' and 3' untranslated regions, or UTRs)
- alternative processing events ... if you sequence deep enough, virtually all genes (in organisms that process transcripts) display alternative splicing, even for single exon genes.
- the potential significance of non-coding RNAs

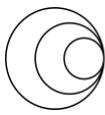
Even heavily curated genomes do not fully reflect all available knowledge about stage-specific splicing, as new information is emerging all the time! In addition, many gene models were computationally derived using methods that may have not relied on experimental evidence supporting intron/exon boundaries (e.g. RNA-seq data).

In this exercise, we will explore genome browser track configuration options in greater detail, focusing on the interpretation of RNA-seq datasets and using this information to examine the genes in *Aspergillus fumigatus* Af293. The genes will be provided in class. Examine gene models and the underlying evidence as a group.

How to create your JBrowse view:

1. Navigate to JBrowse.
2. Select *Aspergillus fumigatus* Af293 from the dropdown menu.
3. Within JBrowse, click the 'Select tracks' tab and choose the 'Transcriptomics' category.
4. Select the dataset called 'Response to caspofungin'.
5. Choose to visualize 'unique' tracks from the 'RNA-Seq Alignment' category.
6. You may also want to activate the 'Syntenic sequences and genes' track to visualize gene conservation in other species.

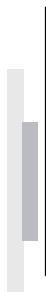
During your discussion, evaluate gene models for missing or incorrectly annotated introns/exons, UTRs, merged genes or unannotated genes in the vicinity. Be prepared to discuss one gene per group in the classroom.



- Apollo
- BLAST (multi-query capable)
- Companion
- CRISPR guide design tool
- Galaxy
- [Genome browser](#)

1

- Aspergillus fischeri NRRL 181
- Aspergillus flavus NRRL3357
- Aspergillus flavus NRRL3357 2020
- Aspergillus fumigatus A1163
- Aspergillus fumigatus Af293**



Select Tracks

• My Tracks

- Currently Active
- Recently Used

.. Category X

- 10 Transcriptomics

• Subcategory X

- 10 RNA-Seq

• Dataset X

- 6 Adaptation to oxygen limitation
- 7 Comparative transcriptomics of dormant and germinating conidia
- 4 Determining Aspergillus fumigatus transcription factor expression and function during invasion of the mammalian lung
- 22 Gene expression in WT, hrmA deletion, hrmA OE, hrmA_REV, EVOL under hypoxia and normoxia conditions
- 6 Gene expression under oxidative and iron stresses
- 6 Mycelial gene expression in response to treatment with 5,8-diHODE
- 10 Response to caspofungin
- 38 Sensitivity of transcription factor mutants of Aspergillus fumigatus to Congo Red
- 6 Transcriptome analysis of conidium germination of Aspergillus fumigatus in different growth conditions
- 10 Transcriptome of wild-type vs veA and mtfA deletion mutants
- 6 Transcriptome under normoxia and hypoxia conditions
- 8 Transcriptomes of WT, nctA, and nctB mutants in response to itraconazole.
- 8 Transcriptomes of

itraconazole-resistant strains
10 Transcriptomics of Aspergillus fumigatus upon exposure to human airway epithelial cells

• Track Type

- 8 Coverage
- 1 Multi X:f plot
- 1 Multi-Density

- RNA-Seq Alignment
- 8 non-unique
- 10 unique

e

Back to bro

X Clear All Filters

Contains text L

[71]

Name	
Response to caspofungin - 001.1 - WT_CSP (unique forward) Coverage	Tr
Response to caspofungin - 001.3 - WT_CSP (unique reverse) Coverage	Tr
Response to caspofungin - 002.1 - WLCT (unique forward) Coverage	Tr
Response to caspofungin - 002.3 - WT_CT (unique reverse) Coverage	Tr
Response to caspofungin - 003.1 - delta fhdA_CSP (unique forward) Coverage	Tr
Response to caspofungin - 003.3 - delta fhdA_CSP (unique reverse) Coverage	Tr
Response to caspofungin - 004.1 - delta fhdA_CT (unique forward) Coverage	Tr
Response to caspofungin - 004.3 - delta fhdA_CT (unique reverse) Coverage	Tr
Response to caspofungin Density - Unique Only	Tr
Response to caspofungin XYPlot - Unique Only	Tr

e

Select Tracks

• My Tracks

- Currently Active
- Recently Used

• category X

2 unique and non-uniq

- 40 (no data)
- 1 Comparative Genomics
- 21 Epigenomics
- 3 Gene Models

Contains text [

category

Exercise: Ensembl Fungi whole-genome alignments

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

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

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



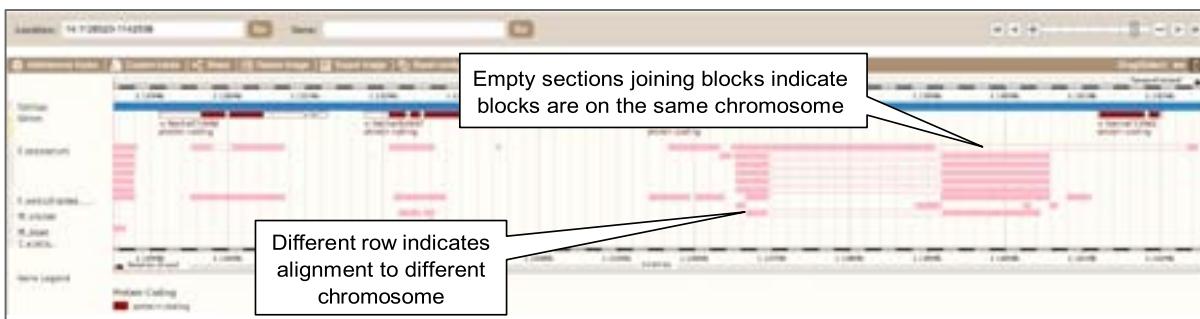
The screenshot shows the Ensembl Fungi search interface. In the search bar, 'Fusarium solani' is entered. Below the search bar, the region '14:1128520-1142558' is specified. A 'Go' button is visible. Below the search bar, there is a placeholder text 'e.g. NAT2 or alcohol*'. The background is light grey, and the overall interface is clean and modern.

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.

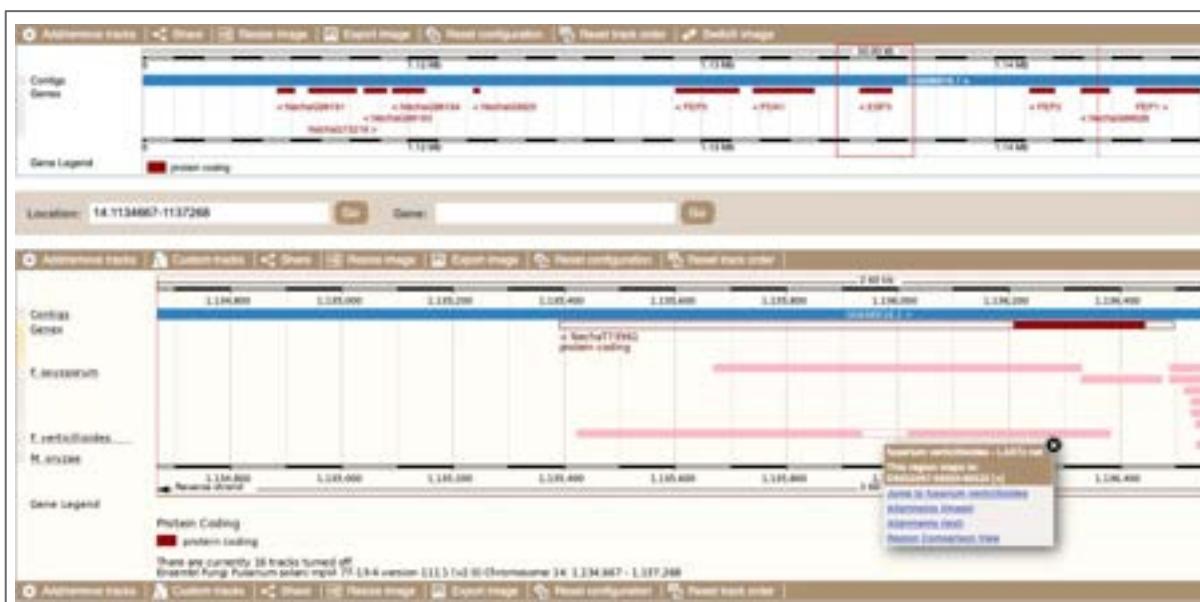


The screenshot shows the 'Configure this page' interface for the 'Comparative genomics' section. On the left, there is a sidebar with various options like 'Active tracks', 'Comparative tracks', 'Track order', etc. The main area is titled 'Comparative genomics' and shows a list of species pairs with checkboxes next to them. Most checkboxes are unchecked, except for a few which are pink, indicating they are currently selected. A legend at the bottom defines symbols for 'Track style', 'Forward strand', 'Reverse strand', 'Resource strand', 'Feature track', and 'Track information'. A note at the bottom says 'Using this page? Learn the Ensembl Glossary for essential terms in annotation'.

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



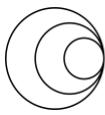
- (b) Looking at the pink alignment blocks, does this region in *F. solani* s align to multiple different chromosomes in the other species?
- (c) Which chromosome(s) does the *F. solani* *ESP3* gene align to in *F. verticillioides*?



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

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

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



The screenshot shows a search interface for alignments. At the top, there's a dropdown for 'Alignment' set to 'Fusarium verticillioides - lastz' with a 'Go' button. Below it is a 'Location' field with '14:1128520-1142558' and a 'Go' button. A 'Gene:' field and a 'Go' button are also present. To the right is a genomic track with arrows. Below these are download buttons for 'Download alignment' and 'Blocks ordered by size'. A message says 'A total of 11 alignment blocks have been found' and 'Sort by clicking on the column headers or by selecting a Block from the Alignment column.' A 'Show 10 entries' dropdown and a 'Filter' input field are at the top of the table. The table has columns: 'Alignment (click to view)', 'Length (bp)', 'Location on *Fusarium solani*', and 'Location on *Fusarium verticillioides*'. The data is as follows:

Alignment (click to view)	Length (bp)	Location on <i>Fusarium solani</i>	Location on <i>Fusarium verticillioides</i>
Block_1	1395	14:1139178-1140572	9:1318698-1321143
Block_2	1218	14:1128517-1130734	11:1354930-1356096
Block_3	961	14:1135422-1136382	Q502267:38355-60322
Block_4	662	14:1132135-1132790	10:1322005-1322692
Block_5	326	14:1138852-1139177	5:2632133-2632458
Block_6	305	14:1140792-1141096	1:1367865-1368184
Block_7	299	14:1136656-1136954	2:183711-164009
Block_8	275	14:1128520-1128794	2:124185-124460
Block_9	119	14:1136637-1136655	Q502270:2615-2738
Block_10	101	14:1140573-1140673	Q502267:3013-3103
Block_11	88	14:1140238-1140325	3:4306258-4306345

At the bottom, it says 'Showing 1 to 11 of 11 entries'.

Annotations in the screenshot:

- A callout points to the 'Select alignment species' dropdown with the text 'Select alignment species'.
- A callout points to the 'Blocks ordered by size' link with the text 'Blocks ordered by size'.
- A callout points to the 'F. verticillioides alignment regions across the region' text with the text 'F. verticillioides alignment regions across the region'.
- A callout points to the 'All *F. solani* alignment regions on chromosome 14' text with the text 'All *F. solani* alignment regions on chromosome 14'.

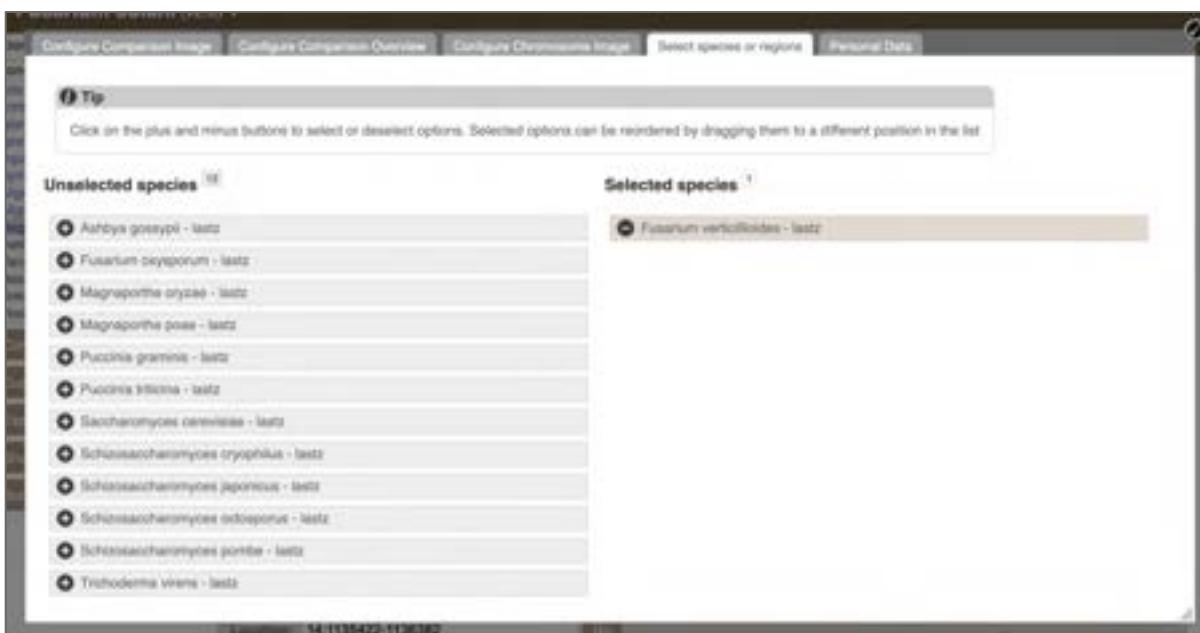
Let's click on **Block 3**. This takes you to a new page with a sample of the aligned sequence. Then click the button to **Display full alignment**. You will see a list of the regions aligned, followed by the sequence alignment. Exons are shown in red. Click on **Configure this page**, you can turn on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

The screenshot shows a configuration menu with tabs: 'Configure Aligned', 'Alignments (text)', 'Configure Display', and 'Normal text'. The 'Configure Display' tab is active. It contains several configuration options:

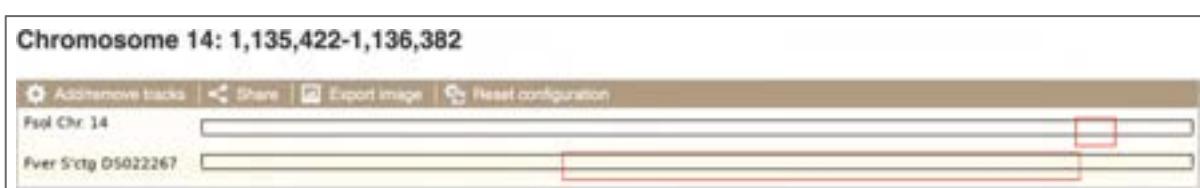
- 'Display options':
 - 'Select from available configurations': dropdown set to 'Default'.
 - 'Save configurations and close pop-up menu' button.
- 'Display options' settings:
 - 'Strand': dropdown set to 'Forward'.
 - 'Number of base pairs per row': dropdown set to '120 bps'.
 - 'Additional exons to display': dropdown set to 'Core exons'.
 - 'Orientation of additional exons': dropdown set to 'Display exons in both orientation'.
 - 'Line numbering': dropdown set to 'None'.
 - 'Codons': dropdown set to 'Do not show codons'.
 - 'Show conservation regions': checked checkbox.
 - 'Mark alignment start/end': checked checkbox.

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

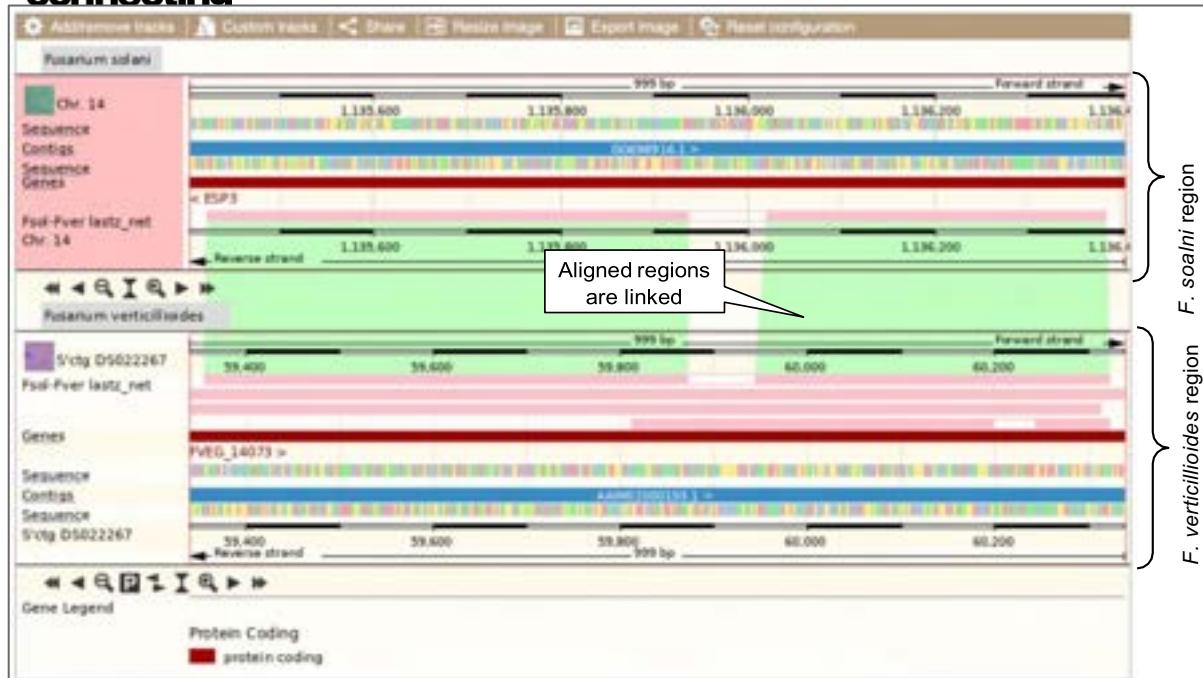
Click on the brown [Select species or regions](#) button. In the pop-up menu, select *Fusarium verticillioides* from the list. Close the window.



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



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



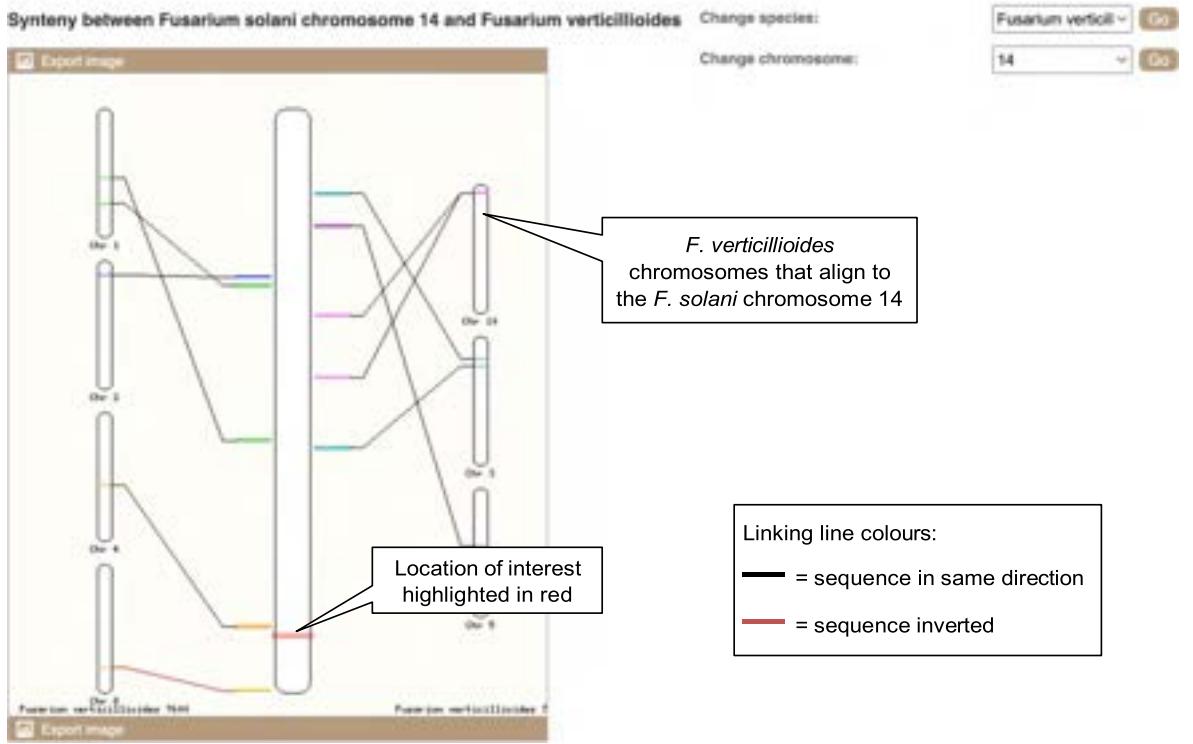
You can add data to both of these views with the same options you had in the 'Region in detail' page. Click on [Configure this page](#) and look at the top of the menu.

This figure shows a configuration menu. It has two sections: 'Species to configure:' and 'Select from available configurations:'. In the 'Species to configure:' section, there are checkboxes for 'Fusarium solani' (which is checked) and 'Fusarium verticillioides'. In the 'Select from available configurations:' section, there is a dropdown menu set to 'Default'.

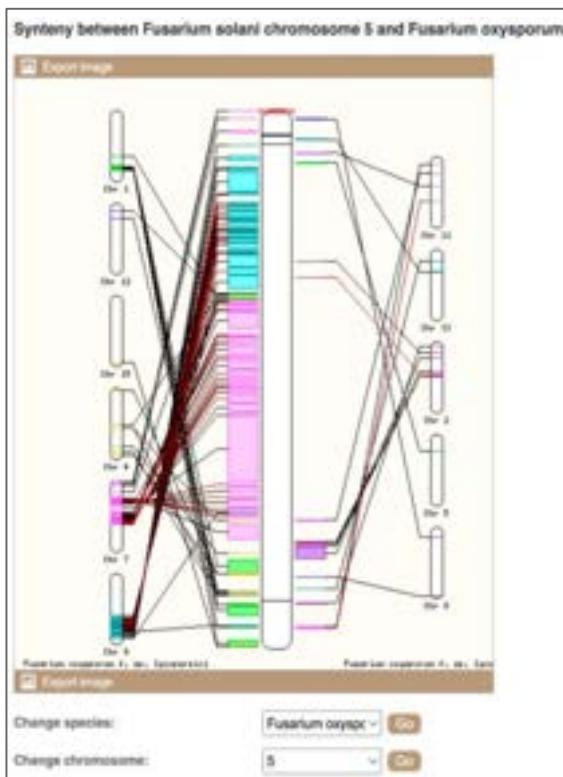
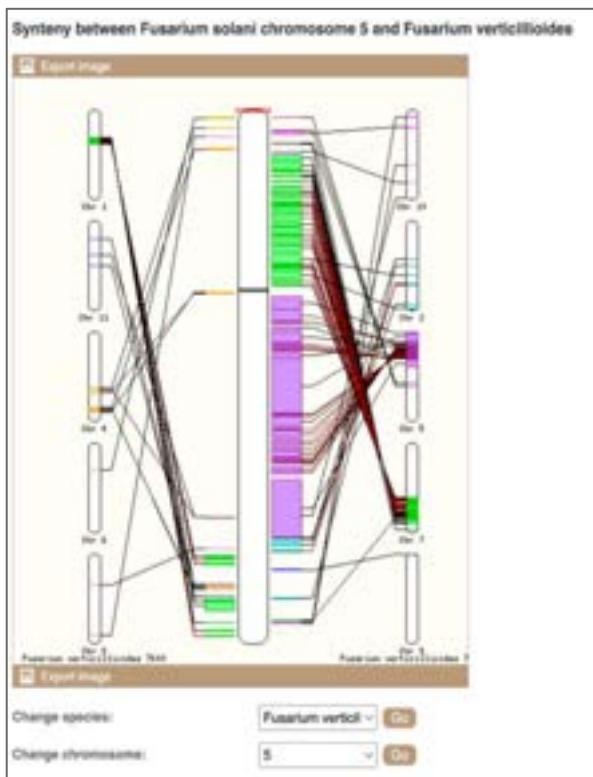
We can view chromosomal rearrangements in the 'Synteny' view. Click on [Synteny](#) in the left-hand navigation panel.



Synteny



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



Additional Exercise - Rearrangements in *Magnaporthe* species

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

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

Answer - Rearrangements in *Magnaporthe* species

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



The screenshot shows a search interface for the fungi.ensembl.org database. The search bar at the top contains the text 'Magnaporthe oryzae'. Below the search bar, there is a text input field containing the genomic coordinates '1:5603535-5611402'. To the right of this input field is a brown 'Go' button. Below the input field, there is a small note 'e.g. NAT2 or alcohol*'.

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



Ensembl Fungi - Search - BLAST - BLAT - Tools - Downloads - Help & Docs - Home

Search Ensembl Fungi

Configure Ensembl Image | Configure Detailed Overview | Configure Chromosome Images | Select species or regions | Previous Data

Tip
Click on the plus and minus buttons to select or deselect options. Selected options can be reordered by dragging them to a different position in the list.

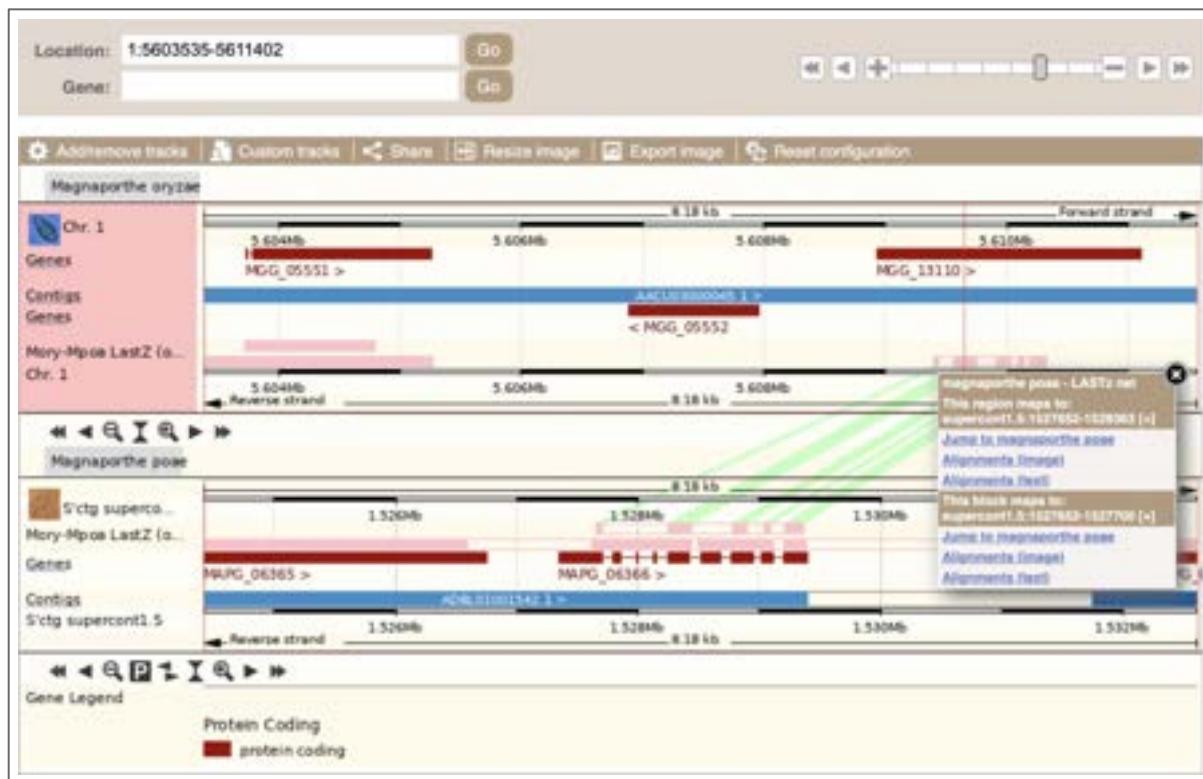
Unselected species (12)

- Achlya galactogae - lastz
- Fusarium coeruleum - lastz
- Fusarium solani - lastz
- Fusarium verticillioides - lastz
- Puccinia graminis - lastz
- Puccinia tritici - lastz
- Saccharomyces cerevisiae - lastz
- Schizosaccharomyces cryophilus - lastz
- Schizosaccharomyces japonicus - lastz
- Schizosaccharomyces octosporus - lastz
- Schizosaccharomyces pombe - lastz
- Tilchoderma viride - lastz

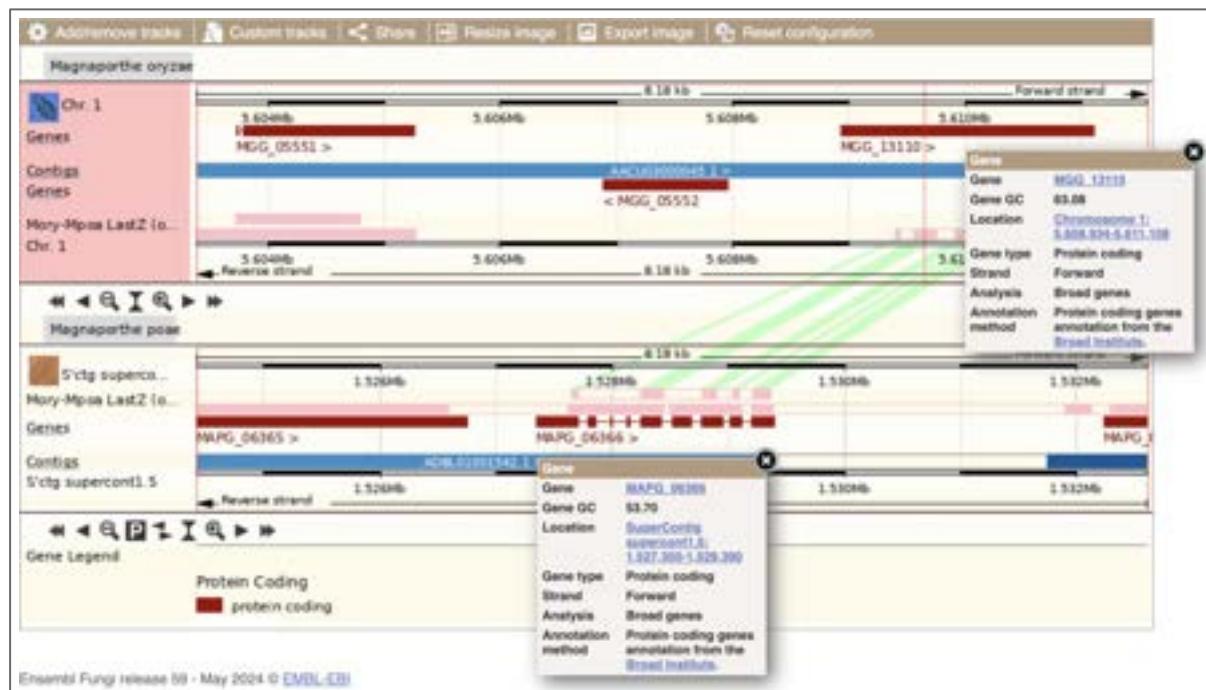
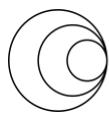
Selected species (1)

- Magnaporthe poae - lastz

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



- (d) In the ‘Genes’ track, find out which features overlap the alignment regions. Click on the feature to find out more information. In *M. oryzae*, the gene MGG_13110 is present. In *M. poae*, the gene MAPG_06366 is present. Both genes are protein coding.



MycoCosm: Comparative Analysis of Gene Families

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

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

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

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

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

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

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

CAZy browser

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

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

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

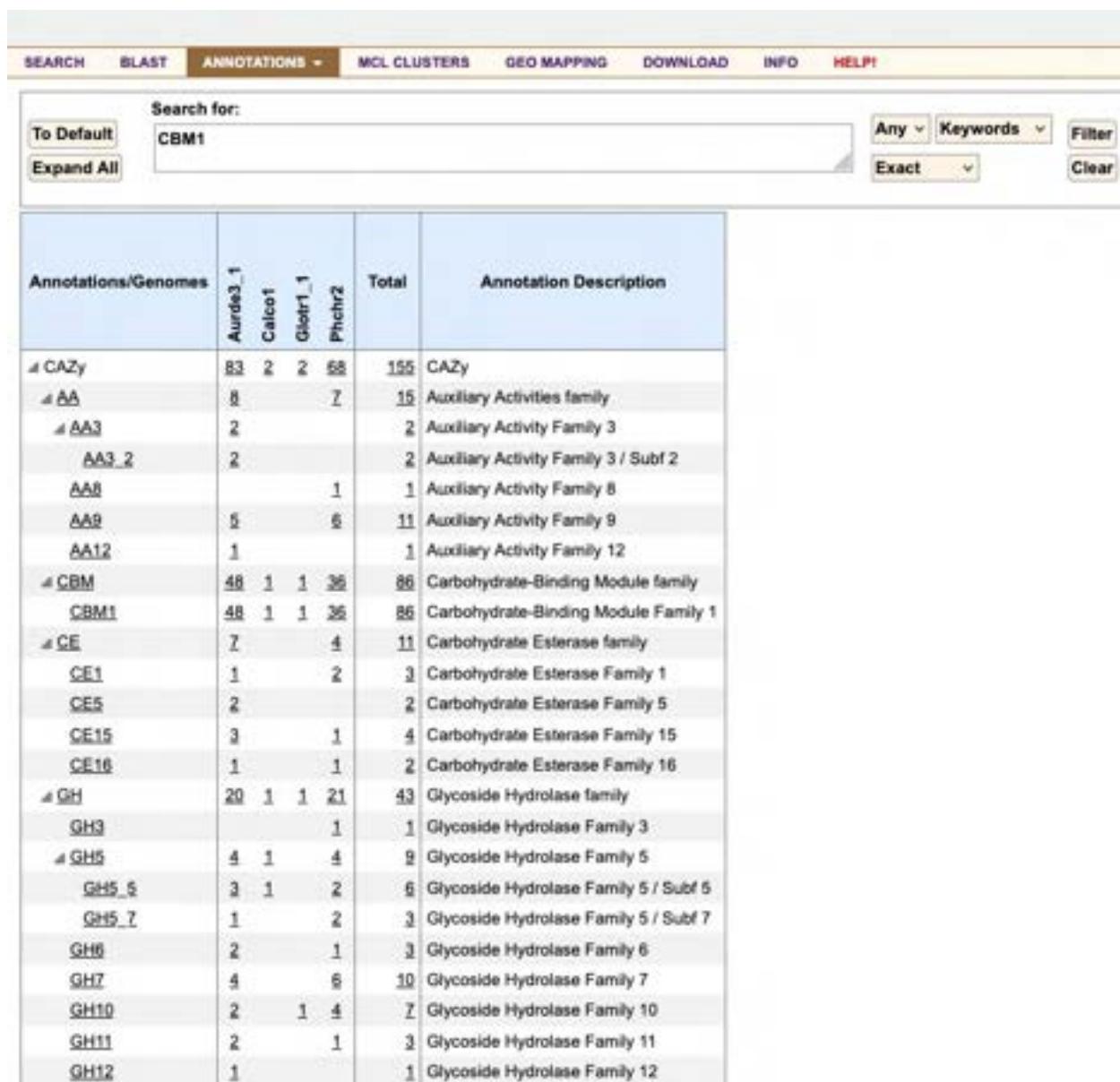
Annotations/Genomes	Annotation Description			
	AA1	AA2	AA3	AA4
✓ CAZy	627	350	368	653
✓ AA	132	27	43	92
✓ BBM	122	38	33	21
✓ CE	91	26	34	29
				2,008 CAZy
				292 Auxiliary Activities family
				221 Carbohydrate-Binding Module family
				152 Carbohydrate Esterase family

Here you will see a table representation of the predicted CAZymes in each species. The organisms are labeled along the top by genome portal identifier (“portal ID”). The CAZymes are organized hierarchically by family and labeled along the sides: CAZy family identifier on the left, and family description on the right. The numbers in the table represent how many proteins from each organism’s gene catalog were annotated with a given CAZyme, with a total provided for each row. Notice that the CAZymes are hierarchically organized: you can see the total number of genes assigned to the top level enzyme category (e.g. “AA”). To view family (e.g. “AA1”, “AA2”) and subfamily (e.g. “AA1_1”, “AA1_2”) designations, click on the small arrow to the left of each category, or use the “Expand All” button at the top of the page.

Annotations/Genomes	Annotation Description			
	Total	AA1	AA2	AA3
✓ CAZy	2,038	886	252	372
✓ AA	297	321	29	44
✓ AA1	25	32	—	—
✓ AA1_1	5	—	—	—
✓ AA1_2	8	—	—	—
✓ AA1_3	7	—	—	—
✓ AA1_4	1	—	—	—
✓ AA2	20	1	3	17
✓ AA2_1	6	1	1	1
✓ AA2_2	14	1	1	1
✓ AA2_3	1	1	1	1
✓ AA3	528	92	38	28
✓ AA3_1	3	1	1	1
✓ AA3_2	525	38	33	26
				1 Multicopper oxidase
				2 Auxiliary Activity Family 2
				3 Class II peroxidase
				4 Auxiliary Activity Family 3
				5 Auxiliary Activity Family 3.1 Subf 1
				6 Auxiliary Activity Family 3.1 Subf 2

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

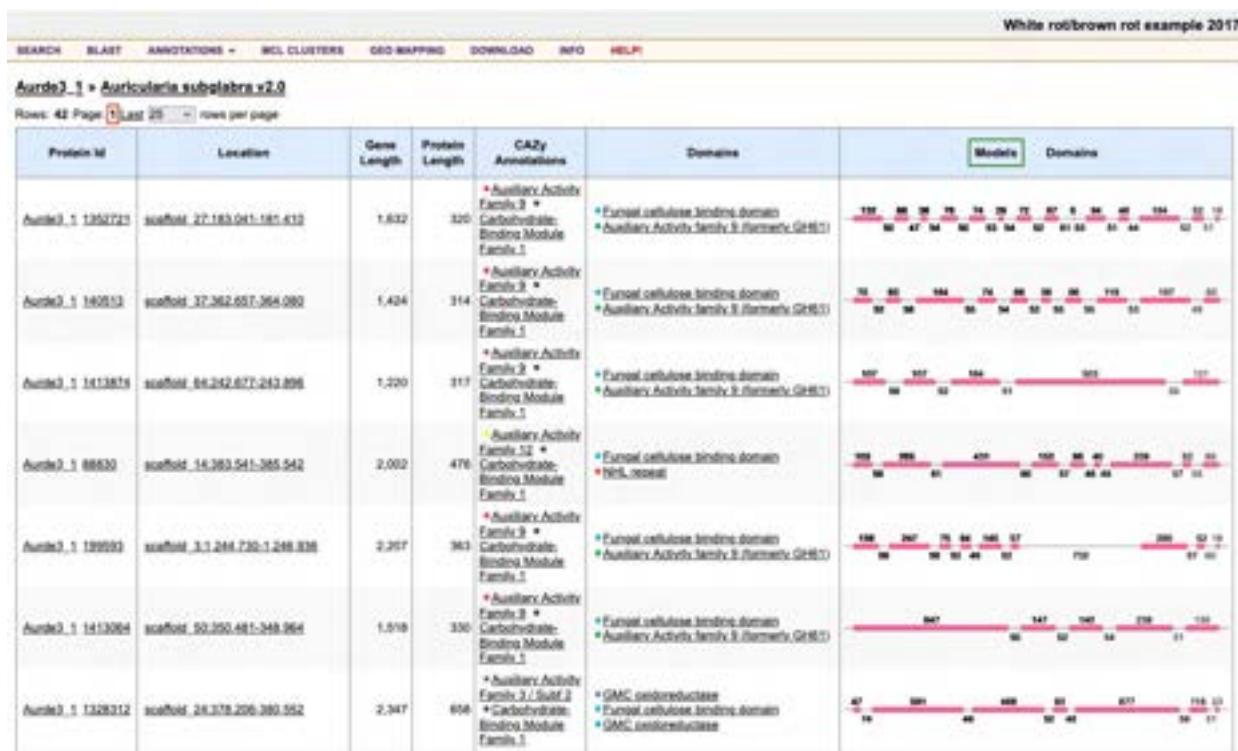
What about the carbohydrate binding motifs, CBM1? Let's say we don't want to scroll through the entire list of CAZymes. Type "CBM1" into the "CAZY terms" search box and click "Filter". This will limit the view to only those CAZymes that have a CBM1. Why do so many CAZymes besides CBM1 show up? Because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



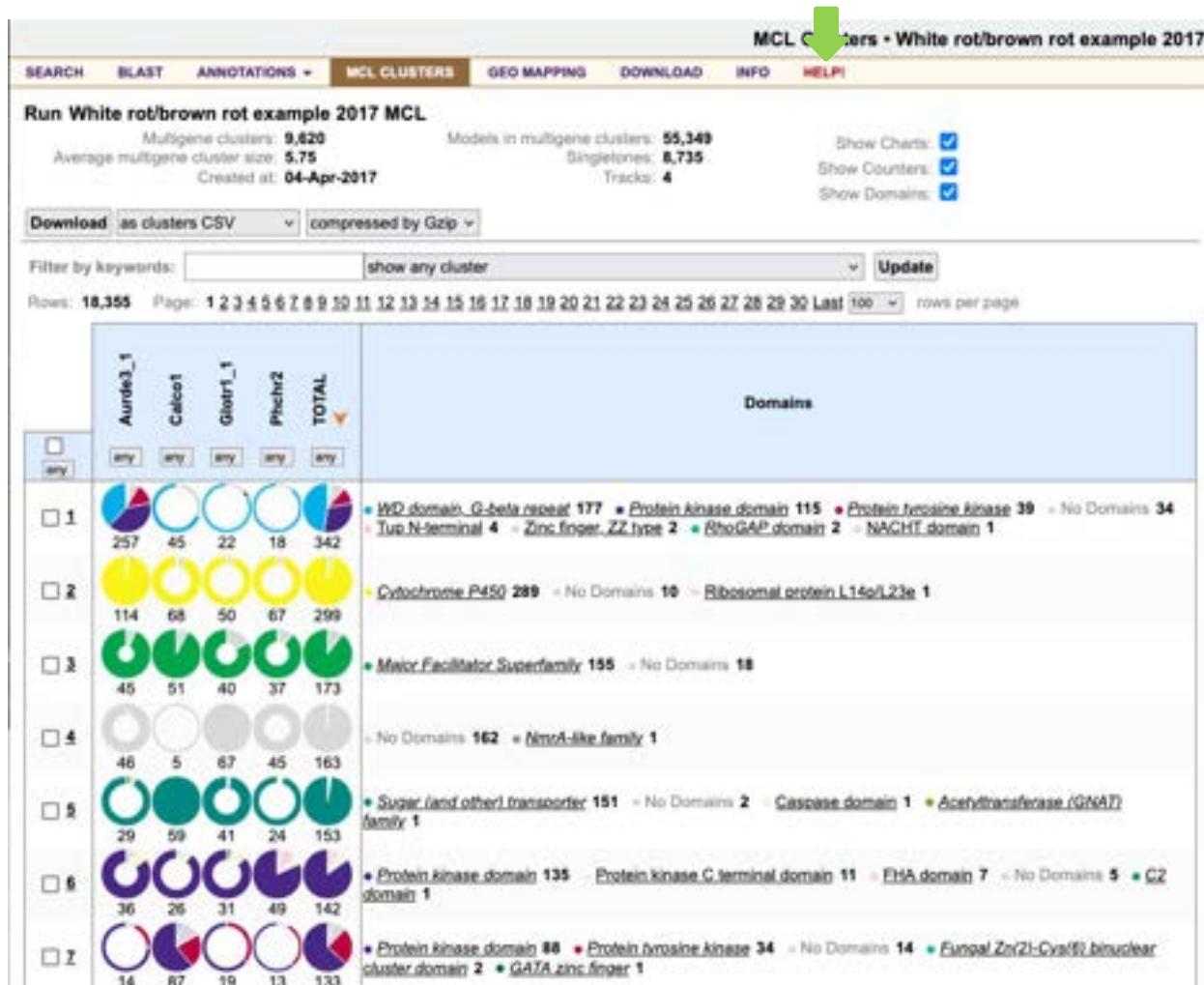
Annotations/Genomes	Aurde3_1	Calcof	Glietr1_1	Phchr2	Total	Annotation Description
CAZY	83	2	2	58	155	CAZY
AA	8			7	15	Auxiliary Activities family
AA3	2				2	Auxiliary Activity Family 3
AA3_2	2				2	Auxiliary Activity Family 3 / Subf 2
AA8				1	1	Auxiliary Activity Family 8
AA9	5			6	11	Auxiliary Activity 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	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.



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



Notice that under each organism label is a button “any” that can be used to filter clusters by the number of proteins that organism contributes to a cluster, and thus limit which clusters are shown. As an experiment, set the white rot fungi (Aurde3_1 and Phchr2) to “1+” and the brown rot fungi (Calco1 and Glotr1_1) to “=0”. Doing so will return only those clusters which are present in Aurde3_1/Phchr2 and absent in Calco1/Glotr1_1.



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

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

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

MCL Clusters - White rot/brown rot example 2017

SEARCH BLAST ANNOTATIONS MCL CLUSTERS GEO MAPPING DOWNLOAD INFO HELP

Run White rot/brown rot example 2017 MCL > Cluster 507

Aurde3_1 Calco1 Glotr1_1 Phchr2 TOTAL

Models: 14 rows per page

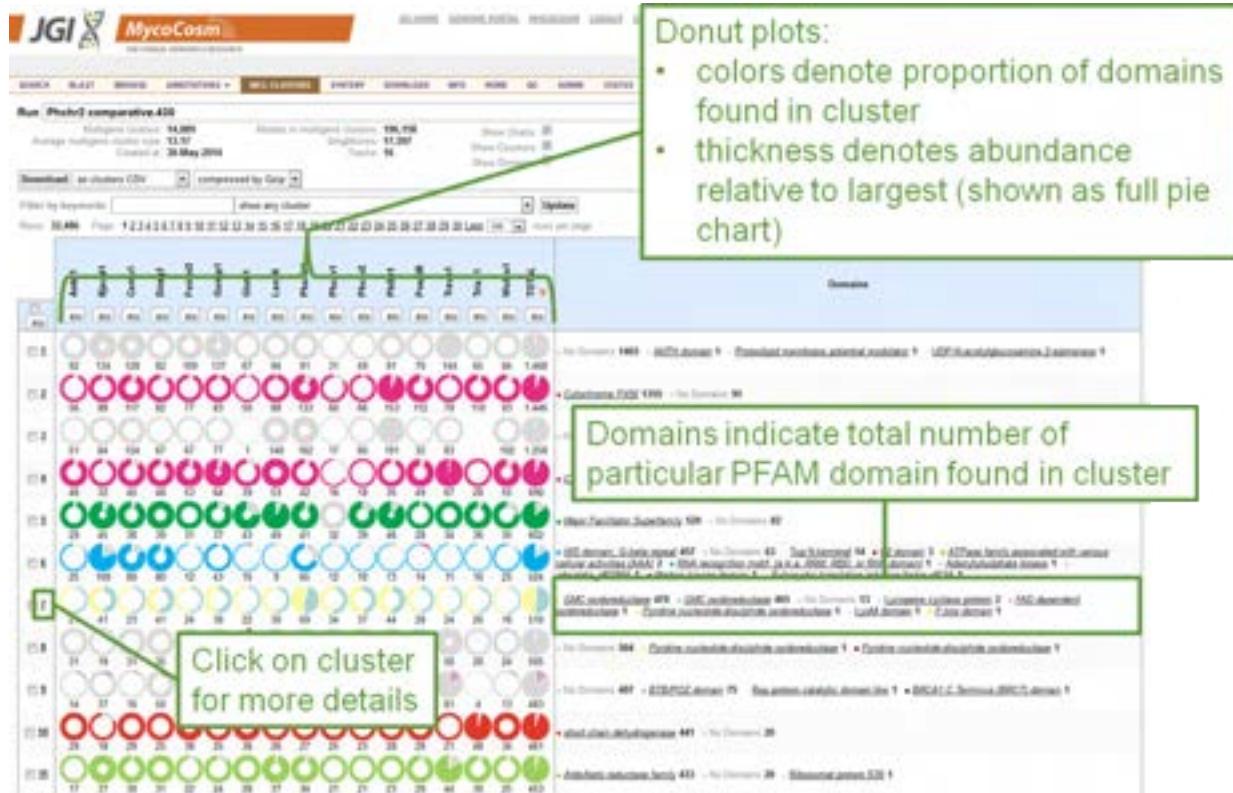
Protein Id	Location	Gene Length	Protein Length	Domains	Model	Domains	Synteny	
Aurde3_1_5228038	scaffold_1:2,027,770-2,029,582	1,813	520	Glycosyl hydrolase family 7 cellulose binding domain	45-14	134	38 216	0.0E+00 0.0E+10
Aurde3_1_1095579	scaffold_7:532,359-534,267	1,909	515	Glycosyl hydrolase family 7 cellulose binding domain peptid leader peptide	47-24	134	38 83 89 80	0.0E+00 0.0E+10
Aurde3_1_5233361	scaffold_7:534,512-540,368	1,876	519	Glycosyl hydrolase family 7 cellulose binding domain	42-824	134	38 83	0.0E+00 0.0E+10
Aurde3_1_1310232	scaffold_21:150,095-151,901	1,807	519	Glycosyl hydrolase family 7 cellulose binding domain	46-14	134	38 205	0.0E+00 0.0E+10
Aurde3_1_1240519	scaffold_21:180,564-182,239	1,676	509	Glycosyl hydrolase family 7	46	137	38 269	0.0E+00 0.0E+10
Aurde3_1_1317128	scaffold_66,202,983-204,748	1,786	449	Glycosyl hydrolase family 7	52-19	46	98 188 46 85 23 39	0.0E+00 0.0E+10
Phchr2_2976245	scaffold_2:2,207,605-2,209,644	2,040	513	Glycosyl hydrolase family 7 cellulose binding domain	201		306	0.0E+00 0.0E+10
Phchr2_2976248	scaffold_2:2,215,861-2,217,914	2,054	513	Glycosyl hydrolase family 7 cellulose binding domain	201		310	0.0E+00 0.0E+10

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

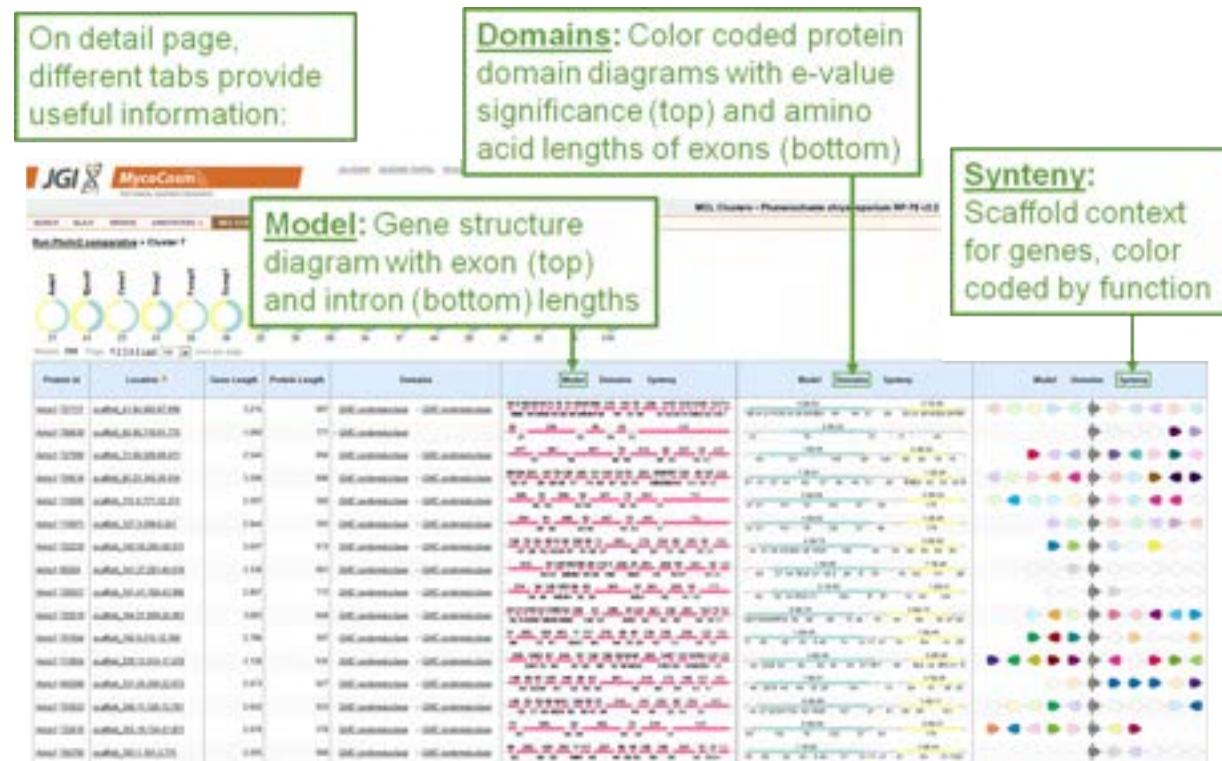
Returning to the cluster run page (click the "MCL CLUSTERS" tab). Enter the phrase "fungal cellulose binding domain" (be sure to include the quotes) into the "filter by keywords" field and select "Update". This returns some 26 clusters, all of which have the Pfam domain CBM_1 (PF00734). We see that CBM1 motifs occur in a wide array of domain combinations: often with GMC oxidoreductases, AA9 lytic polysaccharide monooxygenases (formerly Glycosyl hydrolase family 61), and many hydrolytic enzymes such as GH5, GH6, and GH7. Notice that while these proteins typically are found in expanded copy number in the white rot fungi (Aurde3_1 and Phchr2) they are sometimes found, albeit in lower copy number, in the brown rot fungi (Calco1 and Glotr1_1).

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

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



Clicking in Cluster number provides additional tools as shown below:



References:

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

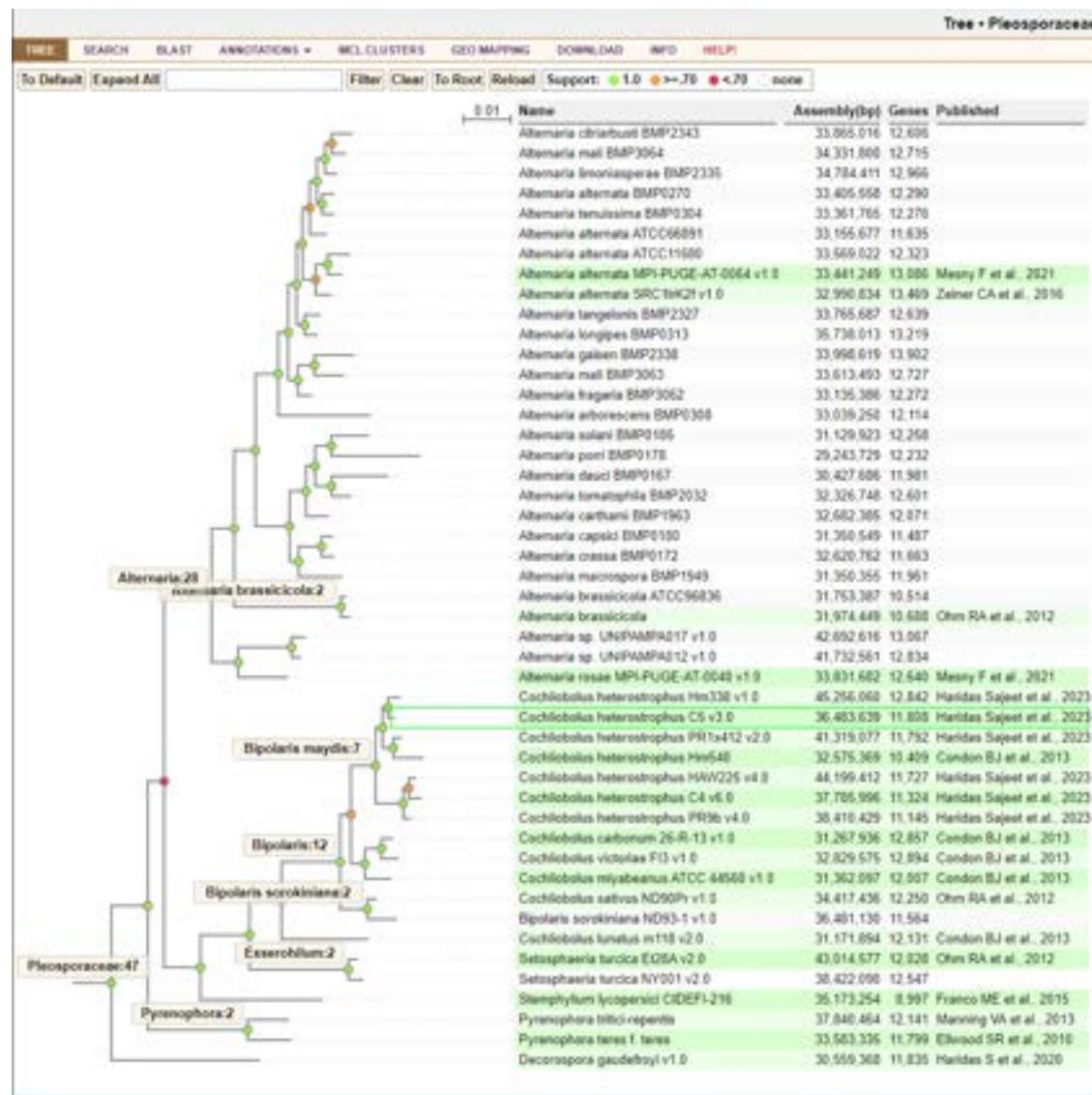
MycoCosm: Synteny Tutorial

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

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

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

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



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

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

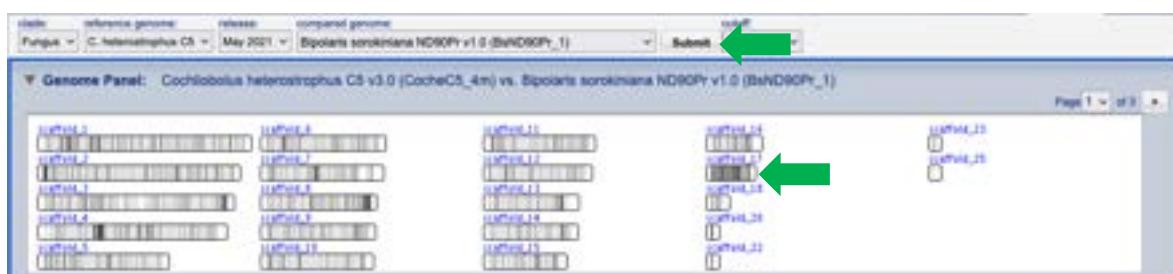


Genomic synteny is displayed in three collapsible panels in the Synteny Browser:

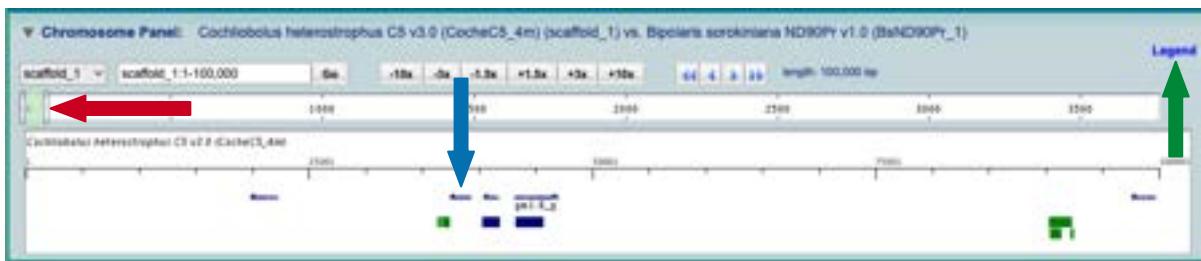
- A. the Genome Panel
- B. the Chromosome Panel
- C. the Comparison Panel.

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

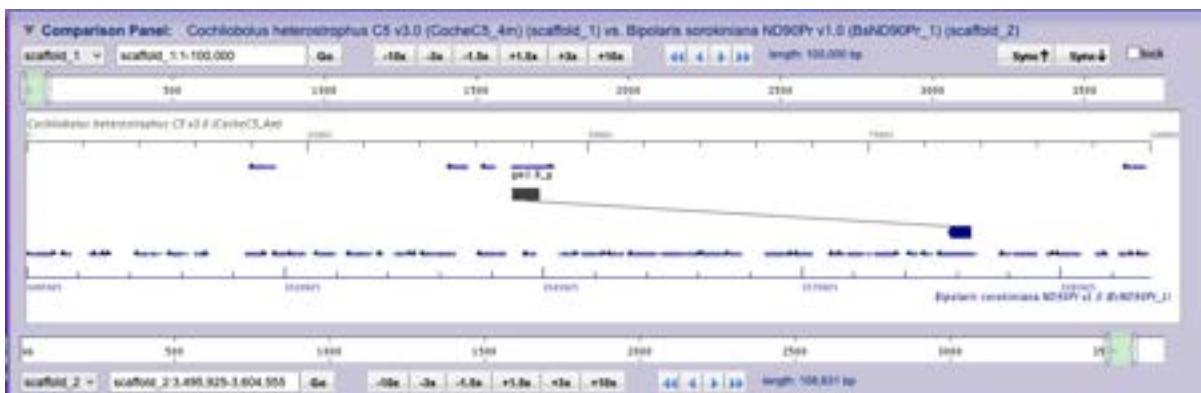
- A. The Genome Panel depicts alignment density for all scaffolds in the reference genome against all chromosomes in the compared genome. Here, alignment density is defined for a region in the reference genome as the number of syntenic regions in the compared genome. Darker regions in the image have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



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



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



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

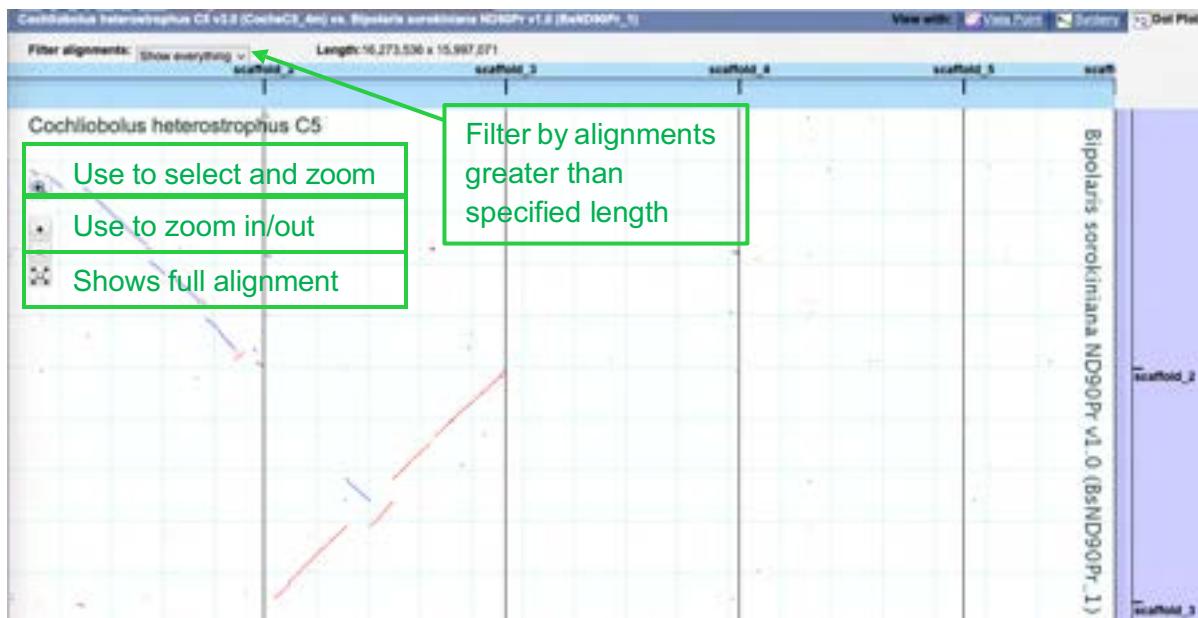
The SYNTENY page also allows whole genome pairwise comparison and comparison of one-to-many using the 'Dot Plot' and 'Vista Point' views respectively.

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



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

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



'Dot Plot' hides the genome portal navigation bar. You can click the "Synteny" view to restore it.

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



Exercises:

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

Reference:

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

Exploring protein domains and clusters across species in Ensembl and MycoCosm

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

We're going to use the HMMER tool, 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/phmmер> 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 exercise, 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 NечаG73962.

- a) Search *Fusarium solani* for NечаG73962 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 on your browser and search for *Fusarium solani*. Select *Fusarium solani* FSSC 5 v1.0, then click on the **MCL clusters** option at the top of the page. Search for the protein domain we identified, *SnoaL_4*.
 - I. For the first cluster, 4,213, which species is missing any hits?

- II. Click on the link for the domain (4213) to explore in more detail. We can see from the donut plots that only 4 species contain this SnoaL-like domain.
 - III. Which species have the most similar protein lengths, and contain the SnoaL-like domain?
- f) Click on Synteny in the final column.
- I. Are the clusters around this conserved between these species? Click on them to highlight the occurrence in the image.
 - II. Which other species has the most similar synteny to *Fusarium solani* (Fusso1) protein 495533? Why do you think this is?

Answers

We're going to use the HMMER tool, 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.

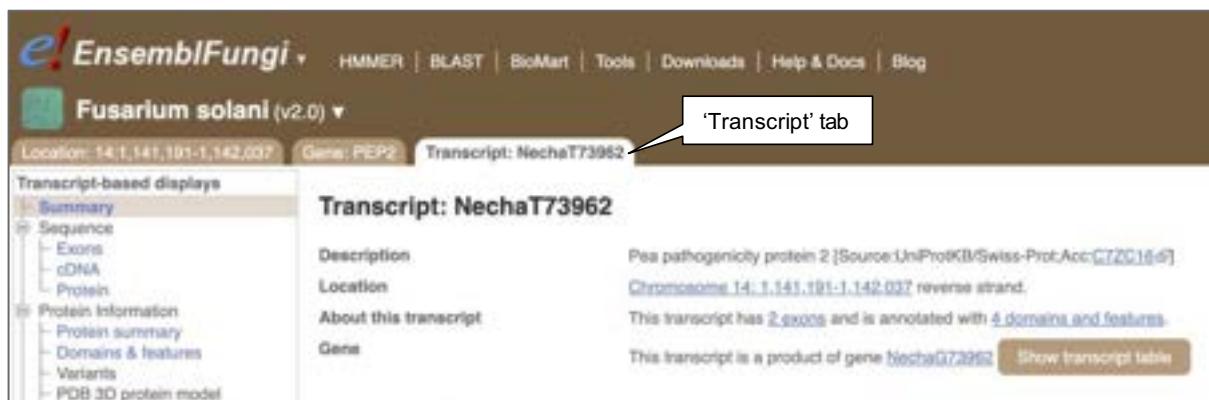
In the BioMart demo, we found genes that were associated with reduced virulence in *Fusarium solani*, and which did not have an orthologue in *F. oxysporum*. We're going to explore one of these genes in more detail: *PEP2* or NechaG73962.

- 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.

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

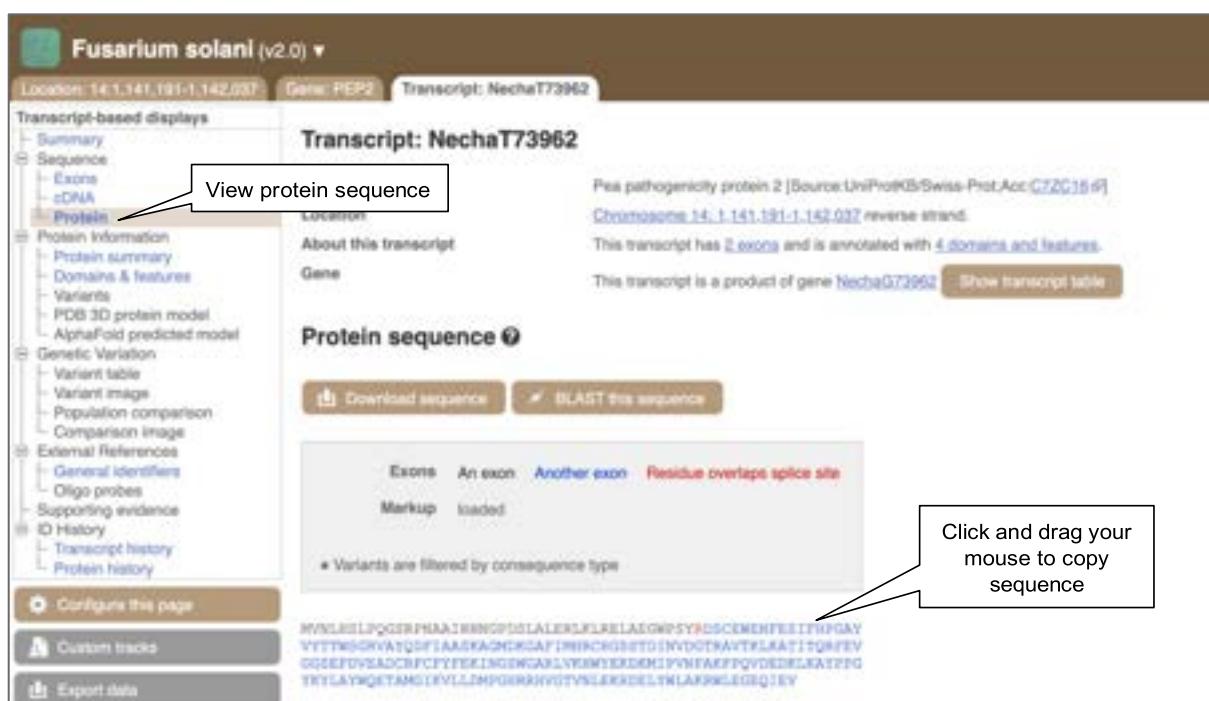


Click on the gene name hyperlink on the results page, this will take you to the gene tab. Click on the transcript tab [Transcript: NechaT73962](#) to go to the transcript tab.



The screenshot shows the Ensembl Fungi interface for the *Fusarium solani* genome. The top navigation bar includes links for HMMER, BLAST, BioMart, Tools, Downloads, Help & Docs, and Blog. Below the navigation is a search bar with the query "Fusarium solani (v2.0)". The main content area displays transcript information for NechaT73962, including its description as "Pea pathogenicity protein 2 [Source:UniProtKB/Swiss-Prot;Acc:Q7ZG14]", location on Chromosome 14: 1,141,191-1,142,037 reverse strand, and details about its domains and features. A callout box points to the "Transcript" tab in the header.

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



The screenshot shows the Ensembl Fungi interface for the *Fusarium solani* genome. The left navigation panel has "Protein" highlighted under the "Sequence" header. A callout box points to this link. The main content area displays the protein sequence for NechaT73962, which is described as "Pea pathogenicity protein 2 [Source:UniProtKB/Swiss-Prot;Acc:Q7ZG14]" and located on Chromosome 14: 1,141,191-1,142,037 reverse strand. The protein sequence itself is shown as a long string of amino acids: MVALIEELPQGERPHAAIZHNGPDALALERLFLRELAEGWPSYDSCMEMHFEETITPGAYVITTMISGVATQGIFTAMSKAKCHKCAPFIRFGCHGHTDINVDGTRAVTKLAKATITQHPEVQQGEFTDVIADCRFCFYFEKINGWGAFLYKEMYERDKHIPVNFAKEPPQVDEDEKLAKTPPGTTTLYATIQETANGLLEVLLIMPUEKHOVUTYSLERKDELTHLAKPLKLEQTYV. A callout box points to the sequence with the instruction "Click and drag your mouse to copy sequence".

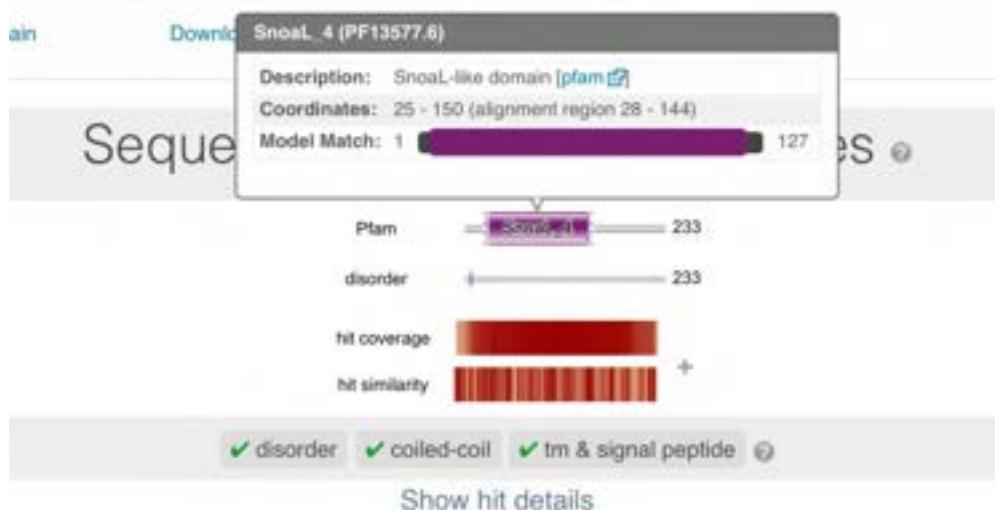
Using the link in the Ensembl Fungi header navigate to the HMMER tool. Paste in your sequence and click **Submit**.



The screenshot shows the Ensembl Fungi HMMER search interface. At the top, there's a navigation bar with links for HMMER, BLAST, BioMart, Tools, Downloads, Help & Docs, and Log in. A search bar is also at the top. Below the navigation, there's a large 'HMMER' logo with a stylized red and blue figure. The main area is titled 'phmmmer' and 'protein sequence vs protein sequence database'. It has a text input field containing a sequence of amino acids: 'MNGLSELDQGGRPHAAATRNGPFLCALKLRLSRLRLADQDHPYRAGCCTWETATPFGKATVYPTTQHQAAYQFLAARVAGHCGAFTPRNGKHSRSLIIVDQHATVYKAKETDQHSEVDEEIPDVAEADCPCTTERDQHEDQHQLYVWEEEDRDKLDPHSHHHPVYDQHGRQHATPPSYVTEGLAQNQETGAACTVLLDQDGRHMMRHTVLSGQELTVLAKENLGEQEV'. Below the sequence is a 'Clear' button.

- III. What is the PFAM domain identified in this sequence?
- IV. 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.



- b) 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) 	 	2.6e-163	

Customise Results

Select Visible Columns  <input type="checkbox"/> Row Count <input type="checkbox"/> Known Structure <input type="checkbox"/> Secondary Accessions and Ids <input type="checkbox"/> Identical Seqs <input checked="" type="checkbox"/> Description <input type="checkbox"/> Number of Hits <input checked="" type="checkbox"/> Species <input type="checkbox"/> Number of Significant Hits <input checked="" type="checkbox"/> Cross-references <input type="checkbox"/> Bit Score <input type="checkbox"/> Kingdom <input type="checkbox"/> Hit Positions <input checked="" type="checkbox"/> Phylum	Rows Per Page  <input checked="" type="radio"/> 50 <input type="radio"/> 100 <input type="radio"/> 250 <input type="radio"/> 1000 <input type="radio"/> 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'

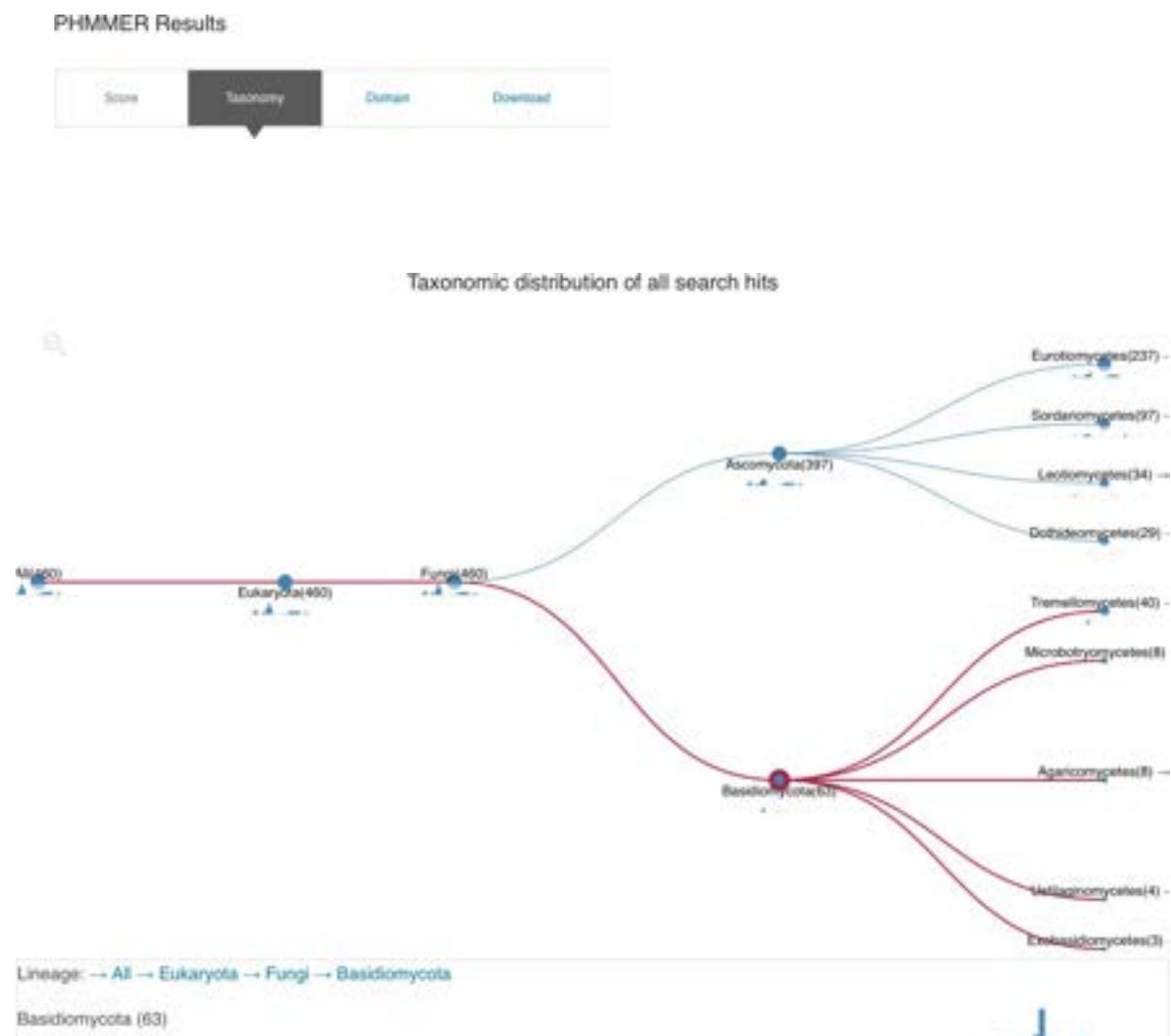
II. Which species is reported as the top hit? Is this the same as *Fusarium solani* (think about the reproductive cycle...)?

Answer: The sexual form (teleomorph) of *Fusarium solani* (the anamorph) is *Nectria haematococca*. Note that *Fusarium vanettenii* is another name for *Nectria haematococca*.

Significant Query Matches (460) in <i>peasubgenomes</i> (v.4)							Customise
Target	Description	Phylum	Species	Cross-references	E-value		
> NечаG73962#	Pea pathogenicity protein 2 [Source:UniProtKB/Swiss-Prot;Acc:C7ZC16]	Ascomycota	Nectria haematococca (strain 77-13-4 / ATCC MYA-4622 / FGSC 9596 / MPVI) 	 	2.6e-163		
> LW93_4799#	Uncharacterized protein	Ascomycota	Gibberella fujikuroi #	 	1.6e-137		
> FFB14_04603#	Pea pathogenicity protein 2	Ascomycota	<i>Fusarium fujikuroi</i> (GCA_900096505) #	 	2.1e-137		
> AU210_001920#	Hypothetical protein	Ascomycota	<i>Fusarium oxysporum</i> f. sp. <i>radicis-cucumerinum</i> #	 	6.2e-137		
> FDWG_10080#	pea pathogenicity protein 2	Ascomycota	<i>Fusarium oxysporum</i> f. sp. <i>lycopersici</i> MN25 (GCA_000259975) #	 	6.2e-137		

- c) 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?

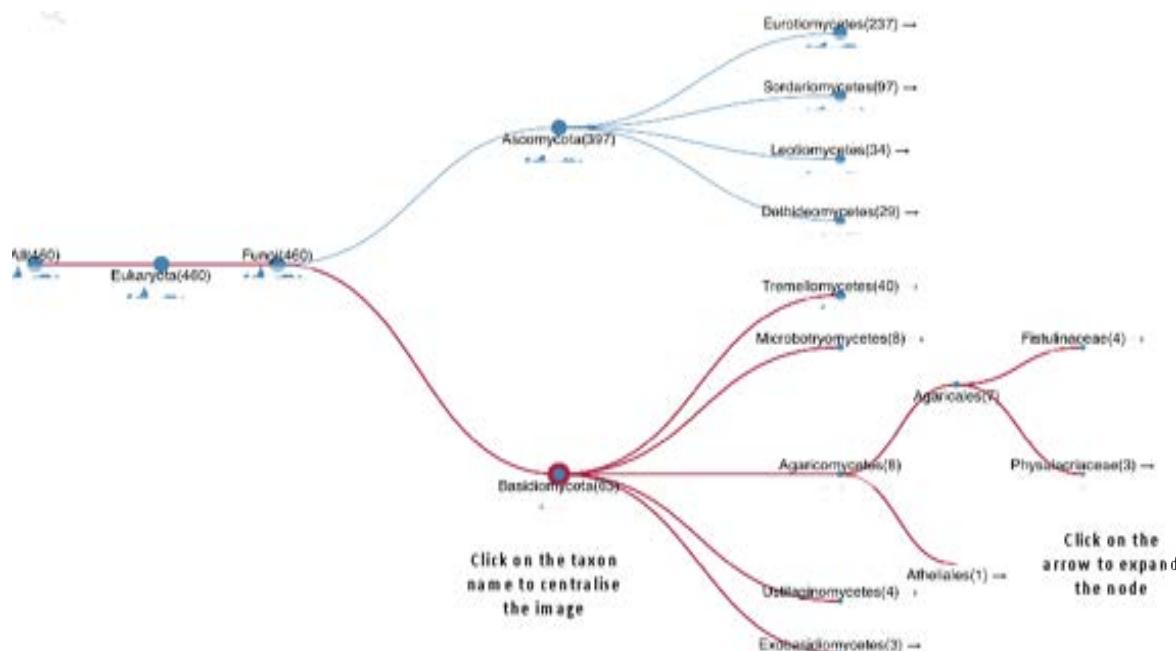
Answer: We can see from the number in the parentheses that there are 63 hits.



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

Answer: Fistulinaceace and Physalaciaceae 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.



- d) 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 Show Charts:
 Average multigene cluster size: 10.09 Singletons: 6,116 Show Counters:
 Created at: 30-Mar-2018 Tracks: 9 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.

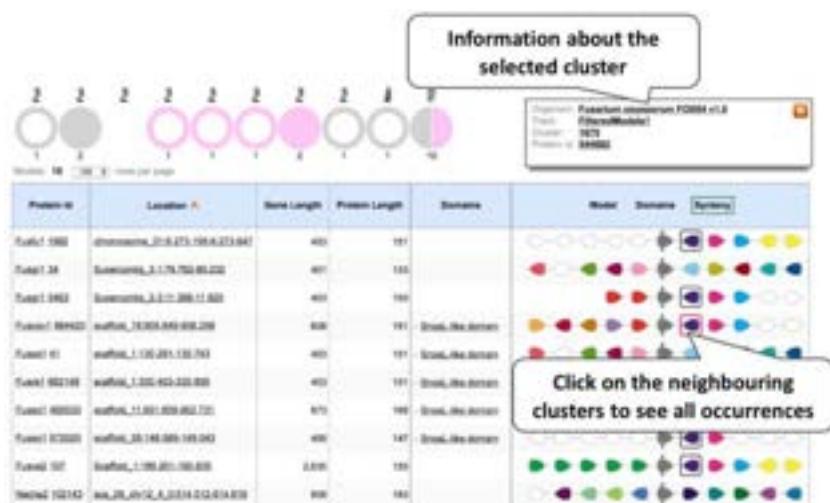
- III. Which species have the most similar protein lengths, and contain the Snoal-like domain?

Protein Id	Location	Gene Length	Protein Length	Domains	Model	Domains	Synteny
Fusel1_3882	chromosome_21.6-273.185-6-273.647	410	181				
Fusel1_38	chromosome_3.1-79.162-89.232	411	133				
Fusel1_5463	chromosome_3.3-11.368-11.820	410	150				
Fusel1_884020	scaffold_18.800-649-906-298	606	181	Snoal-like domain			
Fusel1_47	scaffold_1.130-291-130.213	412	181	Snoal-like domain			
Fusel1_582149	scaffold_1.330-403-330-859	403	181	Snoal-like domain			
Fusel1_495533	scaffold_11.951-859-950.731	873	188	Snoal-like domain			
Fusel1_572000	scaffold_28.148-589-149.043	455	147	Snoal-like domain			
Fusel2_102143	scaffold_1.188-201-180.830	2,635	199				
Neoch1_392143	WA_28_chv12_4.0.014.013-014.019	808	183				

These three have the same protein length

- e) Click on Synteny in the final column.

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



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

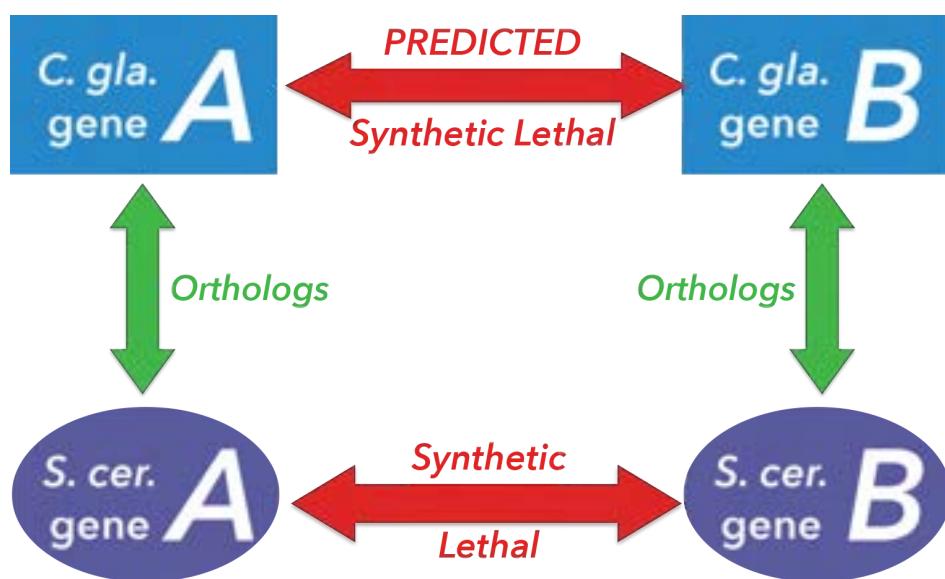
Answer: *Nectria haematococca* v2.0 FilteredModels1. We know this to be the sexual form of *F. solani* so this is expected.

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

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

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

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



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

1. Obtain a list of all genes encoded in the mitochondrial genome of *C. glabrata*:
 - On the CGD homepage (<http://www.candidagenome.org>), open the Search tab in the yellow toolbar and select Advanced Search.



The screenshot shows the CGD homepage. In the sidebar, under the 'Search' section, there is a link labeled 'Advanced Search'. A purple arrow points to this link. The main content area features a 'New and Noteworthy' section about the availability of *C. lusitaniae* strain CBS 6936 sequence and BLAST datasets. Below this, there is a photograph of GFP-labeled Dam1 Complex proteins in DAPI-stained nuclei, with a caption crediting Laura Burack and Judy Berman from the University of Minnesota.

- In Step 1 of the Advanced Search, select ***Candida glabrata CBS138*** as your strain.
- In Step 2, check the “**Select all chromosomal features**” checkbox.
- In Step 3, specify that that you are looking for mitochondrial genes by selecting “**mito_C_glabrata_CBS138**” as the chromosome.

The screenshot shows the 'Advanced Search' interface. On the left, there are three steps: Step 1 (Select strain), Step 2 (Select chromosomal feature), and Step 3 (Narrow results). Step 1 has a dropdown menu with 'Candida glabrata CBS138' selected. Step 2 has a checkbox for 'Select all chromosomal features' which is checked. Step 3 has a dropdown menu for 'Chromosome' with 'mito_C_glabrata_CBS138' selected. There are also sections for 'Annotation/sequence properties' and 'The default search excludes Deleted features.'

- Click on “Search” (bottom left). A results page will follow, listing out 37 features in the *C. glabrata* mitochondrial genome.
- Scroll to the bottom of the page and click on the “**Download All Search Results**” link. The results will download in an Excel sheet.



The screenshot shows a search result for CGD ID 30. The gene is described as a mitochondrial leucine tRNA with UAA anticodon. It spans from position 17,616 to 17,697. A purple arrow points to the "Download" link at the bottom of the page.

Cg30t30	SRNA: Uncharacterized	SL(UAA)mt	Mitochondrial leucine tRNA, has UAA anticodon.	mito_C_glabrata_CG3138:17616 to 17697 Browse	Relative Coordinates	Chromosomal Coordinates
				Noncoding_exon	1 to 82	17,616 to 17,697

Sort by : Systematic Name Get

Analyze gene list: further analyse the gene list displayed above or download information for this list

Further Analysis: [GO Term Finder](#) [ID Slim Mapper](#) [New GO Annotation Summary](#)
[Find common features of genes in list](#) [Sort genes in list into broad categories](#) [View all GO terms used to describe genes in list](#)

Download: [Download All Search Results](#) [Batch Download](#)
[Download all the data retrieved by the query](#) [Download selected information for entire gene list. Available information types include Sequence, Coordinates, GO annotations, Phenotype.](#)

Result Page : 1 2 Next

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

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

The screenshot shows the FungiDB search interface. An orange arrow points to the “List of IDs” link under the “Annotation, curation and identifiers” section.

Search for...

expand all | collapse all

Filter the searches below... [?](#)

Genes

Annotation, curation and identifiers [List of IDs](#) [User Comments](#)

Epigenomics

Function prediction

Gene models

Genetic variation

Genomic Location

Immunology

Orthology and synteny

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

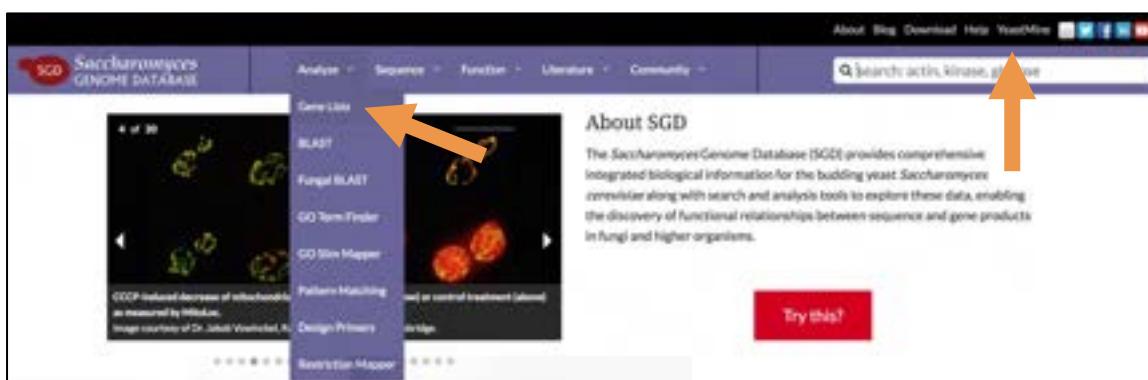


Gene Results									
Gene ID		Transcript ID		Organism		Genomic Location (Gene)		Product Description	
Q0136	Q0136-026_1	S. cerevisiae	S288c	KP263414:46,723..46,952(+)		ND ATP synthase subunit ii	CagMap13	OGI_128019	0 78
Q0045	Q0045-026_1	S. cerevisiae	S288c	KP263414:13,818..26,707(+)		Cytochrome c oxidase subunit 1	CagMap04, CagMap07	OGI_128058	1 43
Q0075	Q0075-026_1	S. cerevisiae	S288c	KP263414:13,818..23,187(+)		Intron-encoded DNA endonuclease att alpha	CagMap04, CagMap07	OGI_128058	1 43
Q0105	Q0105-026_1	S. cerevisiae	S288c	KP263414:36,540..43,847(+)		cytochrome b	CagMap03	OGI_128054	1 31
Q0120	Q0120-026_1	S. cerevisiae	S288c	KP263414:36,540..42,291(+)		Intron-encoded RNA matruse 04	CagMap03	OGI_128054	1 31

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

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 Excel file of *S. cerevisiae* orthologs that you downloaded earlier. To import these orthologs into YeastMine, go to the Upload tab and then create

Create a new list



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

i List type
 Gene
 Organism
 Strain
 Allele

Organism: **S. cerevisiae**

Identifiers are case sensitive

Identifiers:

Free Text:

SHOW EXAMPLE



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

Upload / Save

36 of your 12 identifiers matched a Gene

12 Matches 24 Synonyms

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

Matches (12) Synonyms (24)

① An exact match was found for the following identifiers

② PREVIOUS NEXT ③ Show 10 results on page Page 1 of 2

Your Identifier	Matches	Primary DBID	Systematic Name	Organism > Short Name	Standard Name	Name
Q0060	5000007263	Q0060	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	A13	
Q0070	5000007265	Q0070	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	A15_ALPHA	
Q0250	5000007281	Q0250	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	COX2	Cytochrome c OXidase
Q0080	5000007267	Q0080	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	ATPB	ATP synthase
Q0140	5000007275	Q0140	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	VART1	
Q0130	5000007274	Q0130	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	OUT1	OLigomycin resistance
Q0085	5000007268	Q0085	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	ATPE	ATP synthase
Q0045	5000007260	Q0045	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	COX1	Cytochrome c OXidase
Q0065	5000007264	Q0065	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	A14	
Q0120	5000007273	Q0120	<i>S. cerevisiae</i>	<i>S. cerevisiae</i>	BM	

4. After you save the list, you'll get query results with options for running searches. In the **Widgets** section below the table, click "view all." You'll get roughly 750 results.

Widgets

Interactions

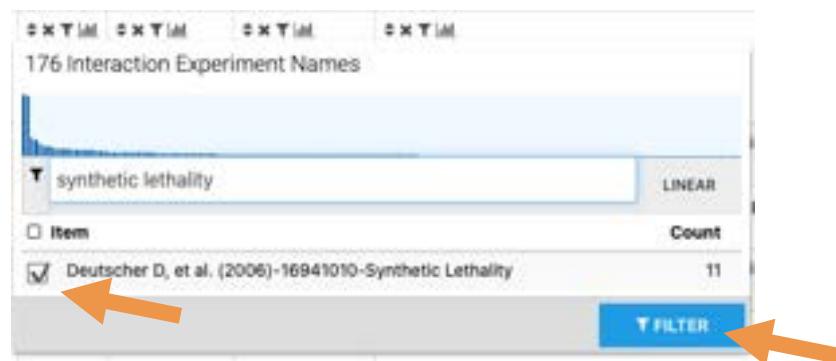
Genes (from the list or not) that interact with genes in this list. Counts may include the same interaction more than once if observed in multiple experiments.

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

VIEW ALL 

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

- In the column for "Interaction Experiment Names," search for "synthetic lethality," click the box next to any results, and then click **Filter**.



176 Interaction Experiment Names

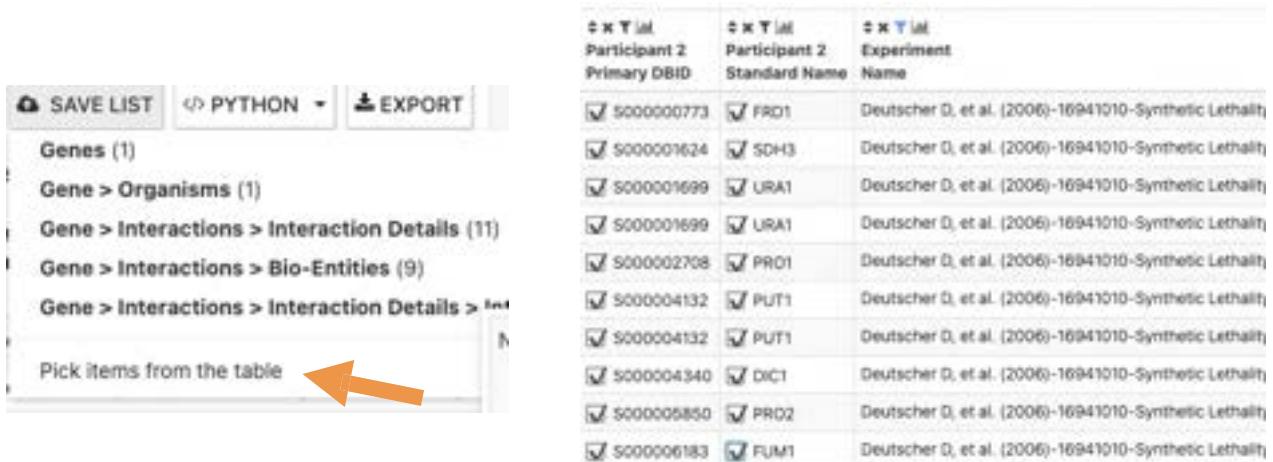
synthetic lethality

Item Count

Deutscher D, et al. (2006)-16941010-Synthetic Lethality 11

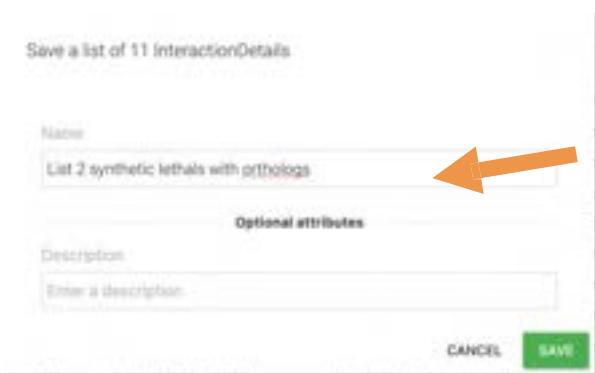
FILTER

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



Participant 2 Primary DBID	Participant 2 Standard Name	Experiment Name
5000000773	FRD1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
5000001624	SDH3	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
5000001699	URA1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
5000001699	URA1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
5000002708	PRO1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
5000004132	PUT1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
5000004132	PUT1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
5000004340	DIC1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
5000005850	PRO2	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
5000006183	FUM1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality

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



Save a list of 11 InteractionDetails

Name

List 2 synthetic lethals with orthologs

Optional attributes

Description

Enter a description

CANCEL SAVE



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

The screenshot shows a data export interface. At the top, there's a 'File name and type' input field containing 'yeastmine_results_2024-05-06T10-26-18' and a 'TSV' button. Below that is a 'Preview (first 3 rows)' section showing a table with columns: ORF > Primary ID, ORF > Systematic Name, ORF > Organism > Strain, and Function. The preview rows include entries like YEL047C, YKL141W, etc. Underneath the preview is a 'Column headers' section with three radio button options: 'No column headers' (unchecked), 'Use human readable headers (e.g. Gene > Organism Name)' (checked), and 'Use raw path headers (e.g. Gene.organism.name)' (unchecked). A 'Select rows' section follows, with a 'Size: 8 (all rows)' dropdown set to 8 and a horizontal scroll bar. Below it is an 'Offset: 0' slider. To the right is a 'Select columns' section with checkboxes for 'ORF > Primary ID' and 'ORF > Systematic Name', and a 'Download File' button. An orange arrow points to the 'Download File' button.

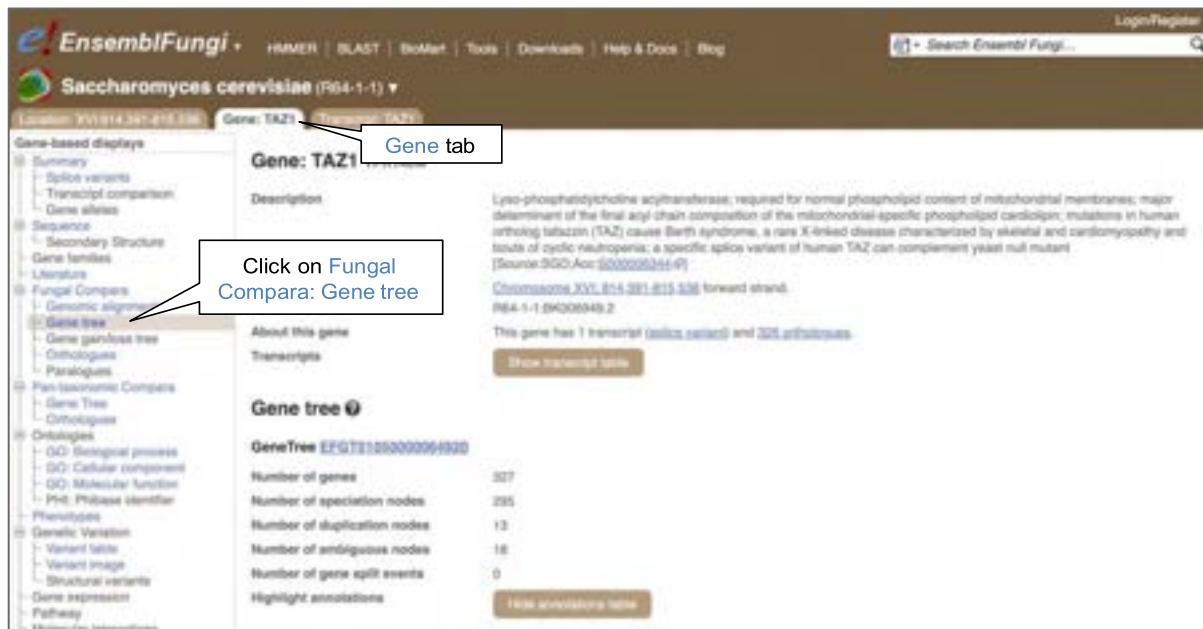
5. Import the *S. cerevisiae* synthetic lethal interaction genes into FungiDB for further analysis:

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

Exercise: Ensembl Fungi Gene Trees and Homologues

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

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



The screenshot shows the Ensembl Fungi interface for the *Saccharomyces cerevisiae* R64-1-1 genome. The URL is <https://www.ensembl.org/Fungi/Gene?g=YPR140W>. The 'Gene' tab is active. On the left, a navigation menu is open, showing the 'Fungal Compara' section expanded, with 'Gene tree' highlighted. A callout box points to this 'Gene tree' link. On the right, the main content area is titled 'Gene: TAZ1'. It includes a 'Description' section with a detailed paragraph about the gene's function and clinical significance, and a 'Gene tree' section with a table of statistics:

Number of genes	327
Number of speciation nodes	295
Number of duplication nodes	13
Number of ambiguous nodes	16
Number of gene split events	0

A 'View transcript table' button is located below the table. A 'Download' button is also visible.

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

EnsemblFungi · HAMMER · BLAST · BioMart · Tools · Downloads · Help & Docs · Blog · Login/Register · Search Ensembl Fungi...

Saccharomyces cerevisiae (R64-1-1) · Gene: TAZ1 · Transcript: TAZ1

Gene-based displays

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

Configure this page

Custom tools

Export data

Share this page

Document this page

Gene: TAZ1 YPR142W

Description

Lysophosphatidylcholine acyltransferase, required for normal phospholipid content of mitochondrial membranes; major determinant of the first acyl chain composition of the mitochondrial-specific phospholipid cardiolipin; mutations in human ortholog TAZ1 cause Barth syndrome, a rare X-linked disease characterized by skeletal and cardiomyopathy and bouts of cyclic neutropenia; a specific splice variant of human TAZ can complement yeast null mutant [Source SGD Acc 20020003644].

Location

Chromosome XVI, 814,391-815,536 forward strand.
R64-1-1:8K30004032

About this gene

This gene has 1 transcript (YPR142W) and 329 orthologues.

Transcripts

Show transcript table

Gene tree

GeneTree [EGGT21290000004930](#)

Unique gene tree stable ID

Summary statistics

Number of genes	327
Number of speciation nodes	296
Number of duplication nodes	13
Number of ambiguous nodes	18
Number of gene split events	0
Highlight annotations	Hide annotation table

Show GO InterPro

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

Accession

327 members	GO:0030749	molecular_function
327 members	GO:0030234	activity
327 members	GO:0008209	lipid metabolic_process
327 members	GO:0008584	phospholipid metabolic_process
327 members	GO:0006730	phosphorus metabolic_process
327 members	GO:0006796	phosphate-containing compound metabolic_process
327 members	GO:0008132	biological_process
327 members	GO:0008332	metabolic_process
327 members	GO:0008827	cellular_process
327 members	GO:0018210	transferase_activity

Showing 1 to 10 of 110 entries

Protein alignments

Collapsed nodes

Gene and species of interest

Legend

Branch Length

- x1 branch length
- x10 branch length
- x100 branch length

Genes

- Gene** gene of interest
- Gene** without sp. paralog

Nodes

- gene node
- speciation node
- duplication node
- ambiguous node
- gene split event

Collapsed Nodes

- collapsed sub-tree
- collapsed (paralog)
- collapsed (gene of interest)

Collapsed Alignments

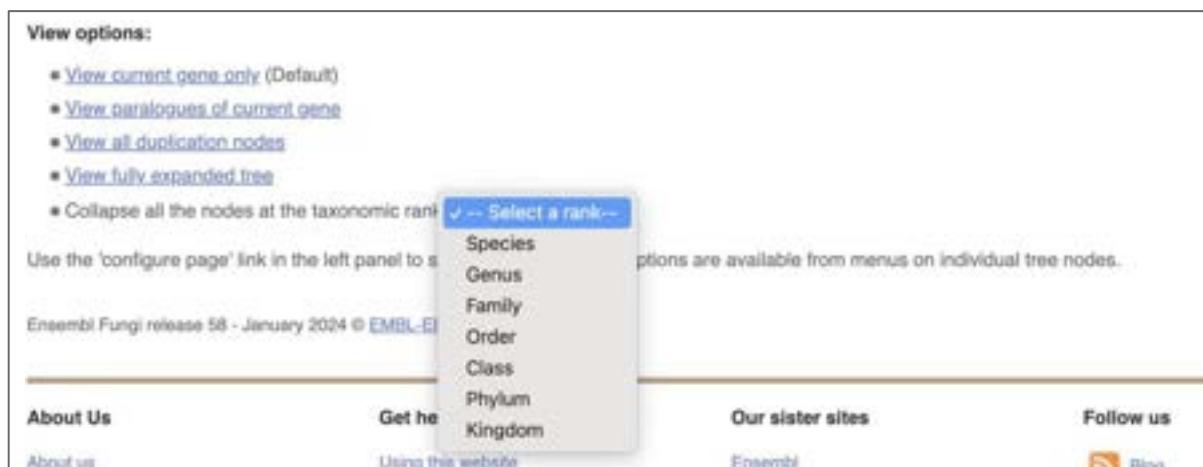
- 0 - 33% aligned AA
- 33 - 66% aligned AA
- 66 - 100% aligned AA

Expanded Alignments

- gap
- aligned AA

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

Scroll to 'View options' at the bottom of the page. Here, you can find some quick filtering options. You can view paralogues and quickly expand or collapse nodes based on class, phylum etc.



The screenshot shows the 'View options:' dropdown menu open. The menu items are:

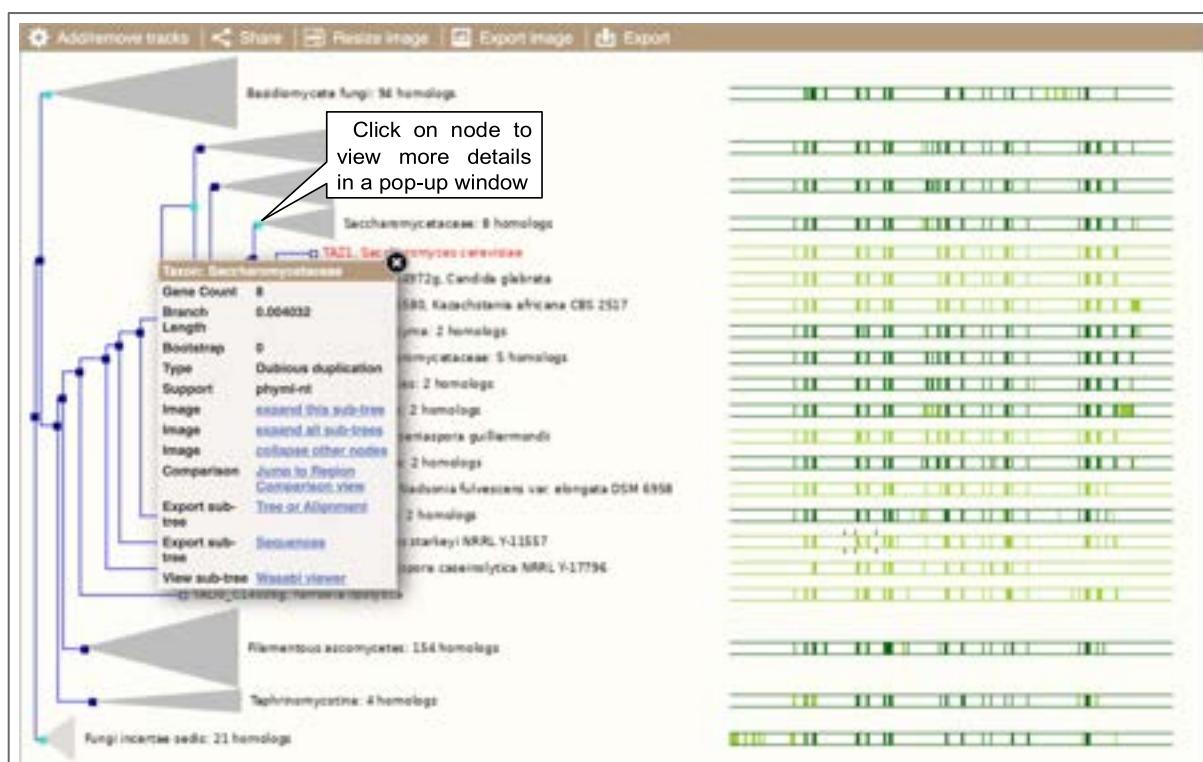
- View current gene only (Default)
- View paralogues of current gene
- View all duplication nodes
- View fully expanded tree
- Collapse all the nodes at the taxonomic rank

A tooltip for 'Select a rank...' indicates: 'Species, Genus, Family, Order, Class, Phylum, Kingdom'. Below the menu, a note says: 'Use the 'configure page' link in the left panel to see more options available from menus on individual tree nodes.'

At the bottom of the page, there are links for 'About Us', 'Get help', 'Our sister sites', 'Follow us', and 'Blog'.

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

Funnels indicate collapsed nodes. Click on a node (coloured square) to open a pop-up window, which tells you what type of node it is, some statistics and options to expand or export the sub-tree:



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

Under 'View options', collapse all nodes at the taxonomic rank **Phylum**.

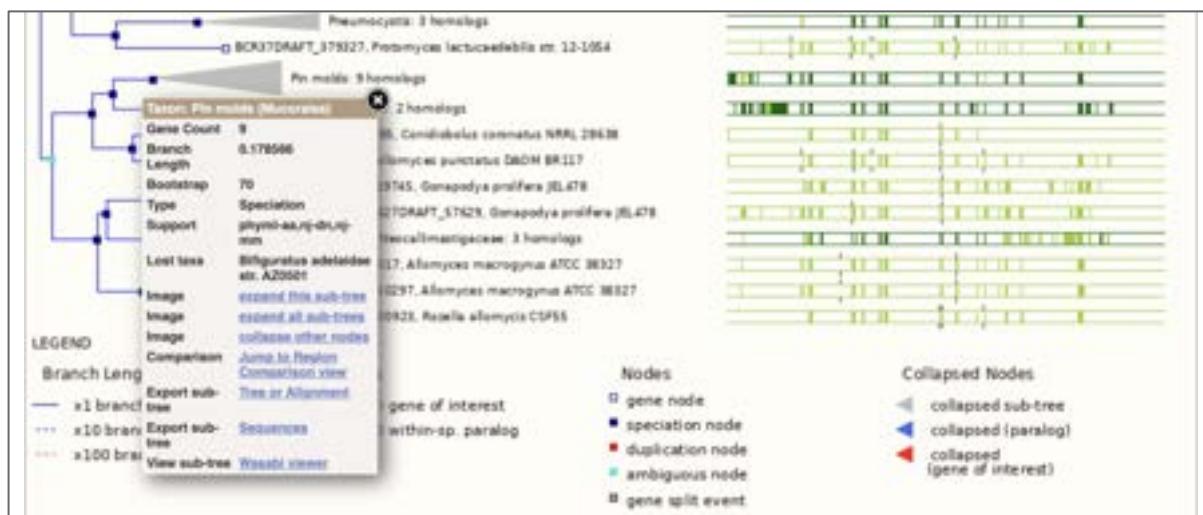


The phylum with the highest number of *TAZ1* homologues is Ascomycete fungi.

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

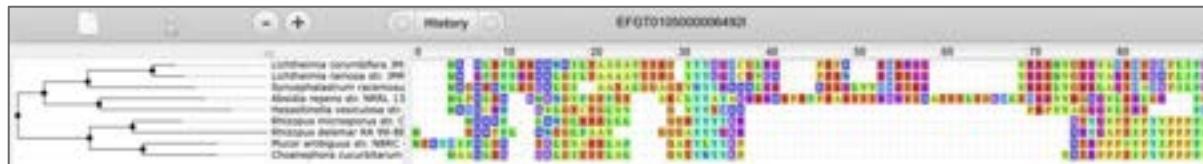
Bootstrap values in a phylogenetic tree indicate that out of 100, how many times the same branch is observed when repeating the generation of a phylogenetic tree on a resampled set of data. Bootstrap values in Ensembl gene trees are calculated using a tool called TreeBeST, and the final consensus trees consist of clades chosen to minimise the number of duplications, losses inferred and have the highest bootstrap support. More on this process is available at https://www.ensembl.org/info/genome/compara/homology_method.html.

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



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

[Wasabi](#) is an open-source, web-based environment for visualising sequence data alongside phylogenetic trees. You can read more about the platform in this publication: <https://europemc.org/article/MED/26635364>.



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





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

Orthologues

Download orthologues

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

Click on 'Show details' to display the orthologues for one or more groups of species. Alternatively, click on 'C' to

Hover over the column names with your mouse to view a description

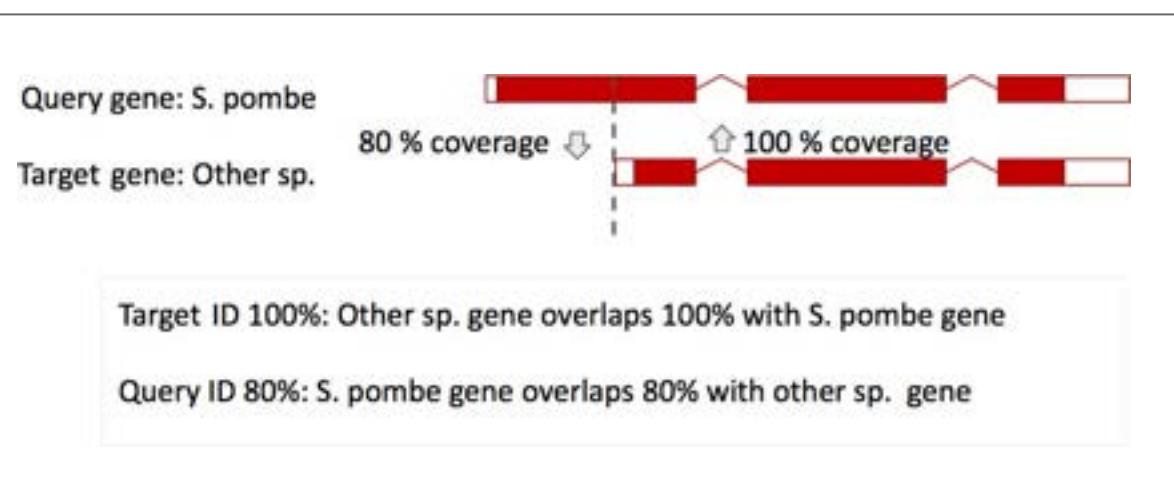
Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many-many orthologues	Without orthologues
All (1504 species)		298	14	0	1192
Acidomyces (2 species)		1	0	0	1
Agaricales (36 species)		18	1	0	12
Atheliales (2 species)		1	1	0	0
Blastocladiales (7 species)		0	1	0	0
Boletales (12 species)		6	0	0	6
Botryosphaerales (7 species)		2	0	0	5
Cantharellales (10 species)		1	1	0	9
Capnodiales (35 species)		3	0	0	24
Chaetothyriomycetidae (31 species)		0	0	0	24
Chytridiomycota (14 species)		4	1	0	9
Corticiales (7 species)		1	0	0	0

Selected orthologues Hide					
Species	Type	Orthologue	Link to orthologue gene tab	Filter table	Download table
Absidia repens str. NPIFL 1336	1-to-1	BCR42DRAFT_405738	View Gene Tree	Compare Regions (BCR42Draft)	Similarity metrics
Acaromyces str. MCA 4196	1-to-1	BCR42DRAFT_381454	View Gene Tree	Compare Regions (K2819639-712,383-713,790-1)	View Sequence Alignments
Acidomyces richmondensis BFW	1-to-1	M433DRAFT_132335	View Gene Tree	Compare Regions (scaffold_55-22,999-24,304-1)	View Sequence Alignments
Acremonium chrysogenum ATCC 11650	1-to-1	ACRE_D50350	View Gene Tree	Compare Regions (scaffold53:63,537-64,720-1)	View Sequence Alignments
Agaricus bisporus var. burretii JB137-58	1-to-1	AGABUDRAFT_914626	View Gene Tree	Compare Regions (JH971389:1,858,411-1,859,869-1)	View protein or cDNA sequence alignment

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



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



Click on [Hide](#) above the table or scroll to the bottom of the page to see a list of the species that do not have any orthologues with *TAZ1* in *S. cerevisiae*... there are a lot!

Species without orthologues	
1190 species are not shown in the table above because they don't have any orthologue with YPR140W.	
• Ancestral sequence	
• [Candida] arabinofermentans NRRL YB-2248	
• [Candida] auris str. 6684	
• [Candida] auris	
• [Candida] glabrata	

S. cerevisiae is part of Ensembl's pan-taxonomic-compara (often shortened to pan-compara), which compares a subset of fungal species with representative species from other taxa, such as plants, protists, bacteria and vertebrates. This offers a broad view of homologous relationships from across the taxonomy. Go to [Pan-taxonomic Compara: Gene Tree](#). Let's look at the pan-taxonomic tree with nodes collapsed at the [Kingdom](#) rank.



Click on Pan-taxonomic Compara: Orthologues.

Orthologues

[Download orthologues](#)

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

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

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (1504 species)	<input checked="" type="checkbox"/>	4	0	0	1500
Ascomycota (2 species)	<input type="checkbox"/>	0	0	0	2
Agaricales (36 species)	<input type="checkbox"/>	0	0	0	36
Asteriales (2 species)	<input type="checkbox"/>	0	0	0	2
Blastocladiales (1 species)	<input type="checkbox"/>	0	0	0	1
Boletales (12 species)	<input type="checkbox"/>	0	0	0	12
Botryosphaerales (7 species)	<input type="checkbox"/>	0	0	0	7
Cantharellales (10 species)	<input type="checkbox"/>	0	0	0	10

Species	Type	Orthologue	Target %Id	Query %Id	GOC Score	WGA Coverage	High Confidence
Aedes aegypti (Yellow fever mosquito, LVP_AGWG)	1-to-1	AAEL001564 -P	23.81 %	18.37 %	n/a	n/a	No
		View Gene Tree	221,496,991-21,541,309>1				
			View Sequence Alignments				
Ambonella trichopoda	1-to-1	AMTR_460322c00068580 -P	23.43 %	17.59 %	n/a	n/a	No
		View Gene Tree	AmTr_v1.0_scaffold00022-710,032-717,504>1				
			View Sequence Alignments				
Amphimedon queenslandica (Demosponge)	1-to-1	LOC100632622 -P	24.73 %	18.11 %	n/a	n/a	No
		View Gene Tree	GL345242.1:108,662-110,163>1				
			View Sequence Alignments				
Anopheles gambiae (African malaria mosquito, PEST)	1-to-1	AGAP007599 -P	24.57 %	18.64 %	n/a	n/a	No
		View Gene Tree	2L>48,133,715-48,137,634>1				
			View Sequence Alignments				

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

Fungal Compara:

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (370 species)	<input checked="" type="checkbox"/>	298	14	0	58

Pan-taxonomic Compara:

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (370 species)	<input checked="" type="checkbox"/>	4	0	0	366

- (g) How many animal orthologues are there? Does this number agree with the Pan-taxonomic tree above? Hint: Click the [Show details](#) box for [Vertebrates](#) and [Metazoa](#), and count the number of orthologues in the table below).
- (h) Filter the second table to view the human orthologue. How much sequence identity does the human protein have to the *S. cerevisiae* one? Is it a high-confidence homologue? Click on the [View Sequence Alignment](#) link in the 'Orthologue' column to [View Protein Alignment](#) in ClustalW format. Does it support your conclusions?

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	ENSP00000460981	292 aa	24 %	85 %	X:154411524-154421726

CLUSTAL W (1.81) multiple sequence alignment

YFR147W/1-381 MEF2DYL-----ERGIDEFLEAYPRRS-----FLMLRFLSTTSLLTGFQWSKILLPFTCYVNA
E3HSP0000463981/1-282 MEF2DYL-----ERGIDEFLEAYPRRS-----FLMLRFLSTTSLLTGFQWSKILLPFTCYVNA

TPR142W/1-381 KLNQFKEKLETALREKKRNNEGLMTVWNNRNGHIVDPLNWA71PFKLTPELINTRMELGAA
EMEP03Q00669981/1-292 TVHSKREVLYELIEK-RGPAFLITVVEHHQSCDHQCLWHLWELLELRLRWWLKLMLRWTFAAAS

YPR14C/M/1-381 ICPQKRIIFLANTFTSLOQVLSTER ***** - PIVVUPPQEE
EHEP02005469981/1-292 ICPTKRLIIRHFFSFLGKCVFVCRGAEEFPDAMREBKGKVILYCHGIMPAGDKRERIGNGNYQK

Additional Exercise 1: *Zymoseptoria* Orthologues

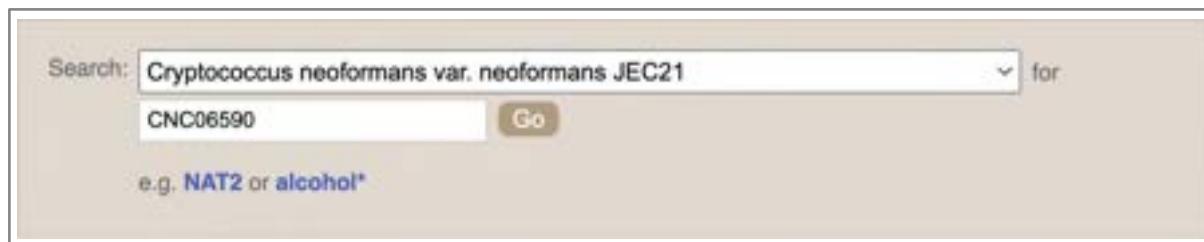
We will now explore an orthologue that we identified using BioMart (additional exercise 1 in the BioMart module). We identified 18 genes associated with the GO term detoxification in *Zymoseptoria tritici*. We then found a single high-confidence orthologue in *Cryptococcus neoformans* (CNM01690) which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

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

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

Additional Exercise 1 Answer: *Zymoseptoria* Orthologues

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



Search: for

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

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



External references 	
This gene corresponds to the following database identifiers:	
External database	Database identifier
NCBI gene (formerly Entrezgene)	3240240  [View all locations]
UniGene	Ene_7413  [View all locations]
UniProtKB-Gene Ontology Annotation	Q5K7P9  [View all locations]

- Go to [Fungal Compara: Orthologues](#) in the left-hand panel. Click on the [Hide !\[\]\(ac0a0ffb36288fed1ab84e1d6f252138_img.jpg\)](#) button for the 'Summary of orthologues of this gene' table. In the 'Selected orthologues' table, use the search bar in the top right-hand corner to search for *Zymoseptoria tritici*. Click on [View Sequence Alignments](#) and in the pop-up menu select [View Protein Alignment](#).

- (c) You can find a description of the different symbols by clicking the question mark icon  next to 'Orthologue alignment'. This opens the corresponding help page in a new tab.

Orthologue alignment ?						
Download homology Click on ? to open the corresponding help page in a new browser tab						
Type: 1-to-1 orthologues						
Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Cryptococcus neoformans var. neoformans JEC21	CNM01690	AAW46801	383 aa	71 %	98 %	13.510531-512507
Zymoseptoria tritici	Myco3254449	Myco3254449	385 aa	71 %	98 %	1_5638024-5639654

In the help page, we can find a description of the conservation codes:

For protein alignments, the conservation codes are:

- * when amino acids are identical
- : when amino acids are different but the function is conserved
- when amino acids are different but the function is semi-conserved.
- space when amino acids are different and there is no conservation of function.

Dashes in the sequence (for both nucleotides and amino acids) indicate gaps in the alignment.

Looking at the ClustalW alignment and referring to the conservation codes, we can see that the N-terminus is more highly conserved than the C-terminus as there is a gap in the alignment in the C-terminus:



CLUSTAL W (1.81) multiple sequence alignment

AAM46801/1-383 Mycgr3P54449/1-385	MSTEQQVITCKAAIAWAEAKPLSIETVEVAPPKQGEVRINKILYTLGLCHTDAYTLSGNDPE MSTEQQVITCKAAVANKEAGKDLVIEDVEVLFFRAMEEVRIKVAYTOVCHTDAYMLSGKDPE *****,******;***** * * * *; * * *; * * * * * * * *;
AAM46801/1-383 Mycgr3P54449/1-385	GAPPVILGHEGGGIVESVGEGVDNVKVGDNVVFLYTAECRECKFCRSRKTNLCGKVRRTQ GAPPVIAGREGAGIYESIGEGVDNVKGDTVVAVLYTPECKECKFCRSRKTNLCGKIRATQ ***** * * *;***** * * * * * * * *;*****;*****;
AAM46801/1-383 Mycgr3P54449/1-385	GKGVMPDGTTIRFKNCQDQDILHEFMCGSTFAQYTVVSKFSVVAINPKAPLKTSCILLOCITT GKGVMPDGSSRFNCKGDLLIEFMCGSTFRQYTVVADISVVAVTDKAPEMDRTCLLOCITT *****;*****;*****;*****;*****;*****;*****;*****;
AAM46801/1-383 Mycgr3P54449/1-385	GYGAATKSP---GI-EGSNVAIPEGVGCVGLSVLQOAKAKOCKRIFAIIDTNPKKKKEAWKF GYGAATITAGIONGVENGDONVAVFGAGCVGLSVIQGAAASRIMAGKIVVVDVNDSKKKEWASKF ***** 1. * * 1 * * * 1 * * * 1 * * * 1 * * * 1 * * * 1 * * * 1 * * * *;
AAM46801/1-383 Mycgr3P54449/1-385	GATDFINP-KOLPEGKRTIVDYLIKEETDGGLDFTFDATGNVGVMRNALEACHKGWGVCTII GATDFVNPTKDLKEGEKIQDRLVEMTDGGCDYTFDCTGNVHVVMRSALEACHKGWGESIII *****;** ***;** * * * * * * * *;*****;*****;*****;*****;
AAM46801/1-383 Mycgr3P54449/1-385	GVAPAGAEISTRPFQIVTGRVWNGSAFGGVKGRTTELPGIVEODYLAGKLWVNEFVTHINGEL GVAAAGQEIASTRPFQIVTGRVWNGCAFGGVKGRSQMGGLIDDYMQGKLMVDEFITHRQNL *****;***;*****;*****;*****;*****;*****;*****;*****;*****;
AAM46801/1-383 Mycgr3P54449/1-385	EGINKGFDDMEAGDCIRCVVDMOF-NEAP GGINDAFAHDMMEAGDCIRCVVDMQKL----- *****;*****;

You can read more about Clustal alignments in the '[The Clustal Omega Multiple Alignment Package](#)' publication by Sievers and Higgins (2020).

Additional Exercise 2: Mushroom Genes

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

From the 'Gene' tab, click to view the [Gene tree](#). At the bottom of the image click to collapse all the nodes at the taxonomic rank of [Class](#).

- (a) What do you notice about the types of fungi shown in the gene tree?
- (b) Does this match with what you would expect from the gene description?
(*Hint: Agaricomycetes class belongs to the Basidiomycota phylum*)
- (c) Based on the protein alignment shown at the right, can you predict which end of the gene/protein is most conserved?
- (d) Click to view the [Orthologues](#) page. In the Selected orthologues table, find the entry for the species *Amanita thiersii* and click to view a protein alignment. Does this support your conclusion about the conserved region of the gene/protein?

Additional Exercise 2 Answer: Mushroom Genes

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



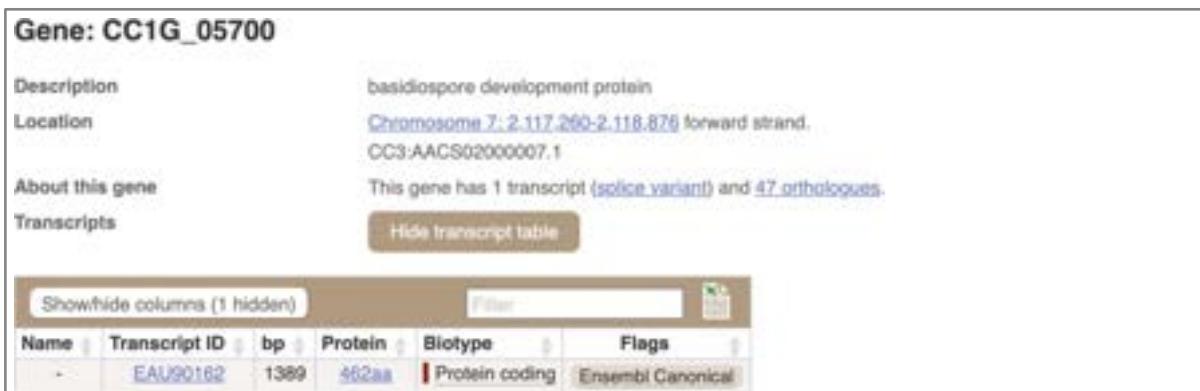
Search: for
CC1G_05700
e.g. [NAT2](#) or [alcohol](#)*

Click on the gene stable ID [CC1G_05700](#) to open the 'Gene' tab. In the left-hand panel, click on [Fungal Compara: Gene tree](#). Scroll to the bottom of the gene tree and collapse all the nodes at the taxonomic rank [Class](#) under 'View options'.

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

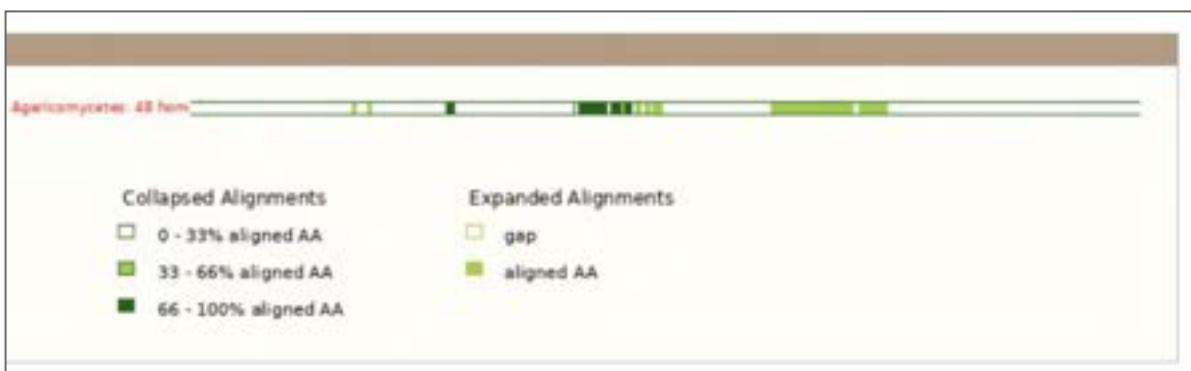


(b) The gene description is as follows:



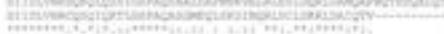
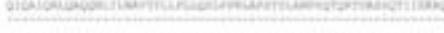
The class Agaricomycetes belongs to the phylum Basidiomycota, therefore we would expect the gene encoding the basidiospore development protein to be conserved across Agaricomycetes species.

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



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

Selected orthologues Hide							
Show		All	Show/hide columns		Amanita thiersii		
Species	Type	Orthologue	Target Id	Query Id	GOC Score	WGA Coverage	High Confidence
Amanita thiersii	1-to-1	AMANTHOSAFT_122148	42.73 %	10.17 %	n/a	n/a	No
Skyo4041		K2301993:102,546-102,928:-1					
		View Gene Tree					
		View Sequence Alignments					

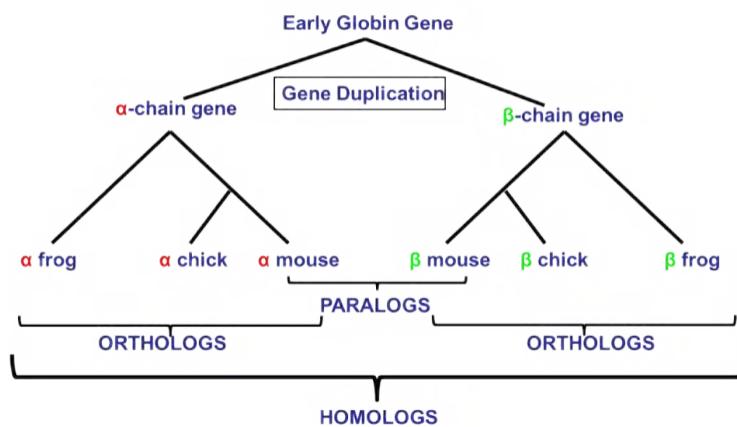
Type: 1-to-1 orthologues						
Species	Gene ID	Peptide ID	Peptide length	% Identity (Protein)	% coverage	Genome location
Coprinopsis cinerea kuyama7W130	CC1G_86706	EAU96162	462 aa	10 %	21 %	T2111720-2119876
Amanita thiersii Stay4041	AMATHI0RAFT_120148	PTT96162	110 aa	42 %	60 %	K2D118A_1022446-1023028
COPROTIN N (1-41) multiple sequence alignment:						
RAE96162/1-410						
PTT96162/1-110						
RAE96162/1-410						
PTT96162/1-110						
RAE96162/1-410						
PTT96162/1-110						
RAE96162/1-410						
PTT96162/1-110						
RAE96162/1-410						
PTT96162/1-110						
RAE96162/1-410						
PTT96162/1-110						
RAE96162/1-410						
PTT96162/1-110						
RAE96162/1-410						
PTT96162/1-110						

FungiDB & OrthoMCL: Orthology and Phylogenetic Patterns

Learning objectives:

- Run searches in OrthoMCL.
- Run phylogenetic pattern searches using checkboxes 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 database that groups orthologous protein sequences across the tree of life. An orthogroup contains genes descended from a common ancestor by a process of duplication and speciation (see figure above), so a single orthogroup may contain both genes across different species with similar function and paralogs within a single species.

Each protein in every OrthoMCL species is assigned to precisely one ortholog group (e.g. [OG6_162879](#)). Importantly, proteins within a single OrthoMCL group have been shown to display a high degree of functional conservation (e.g., a group's proteins have consistent EC numbers) ([Li et al. 2003](#)). Orthology is important in predicting the function of the rapidly increasing number of newly identified proteins produced by genome sequencing and the automated discovery of protein sequences ([Glover et al. 2019](#)). Within VEuPathDB, orthology can be used to transform a list of genes from one species into their closest equivalents in another species.

OrthoMCL contains two sets of genomes. A **Core** set of 150 genomes have been chosen as well annotated reference organisms that broadly represent the major branches of the tree of life. The OrthoMCL algorithm uses BLAST to calculate pairwise distances among all proteins in the 150 core genomes, normalizes the scores for sequence length and evolutionary distance, then uses MCL clustering ([Dongen 2000](#); [www.micans.org/mcl](#)) to create orthogroups of similar proteins. All of the non-core VEuPathDB species (pathogens, hosts, and vectors) have been added as **Peripheral** organisms, in some cases including multiple strains and genome assemblies for the same species. All proteins from the Peripheral organisms are assigned to the most similar Core cluster by best BLAST score, but proteins that do not match any Core protein with an e-value better than 10^{-5} are set aside as **Residuals**.

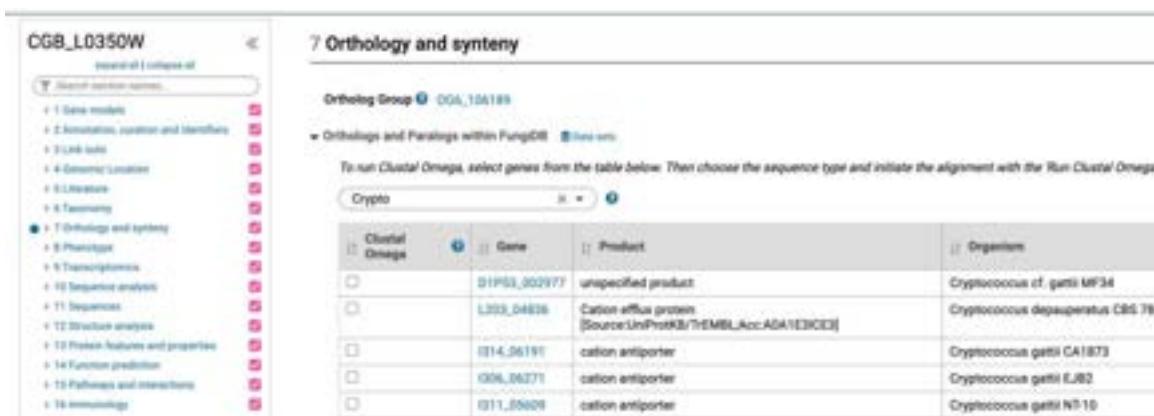
Pairwise BLAST distances among all Residual proteins are computed and used for a second round of MCL clustering to create Residual groups (e.g. [OG6r20_100305](#))

The OrthoMCL website offers the ability to explore orthogroups by taxonomy, number of proteins or species, sequence similarity, EC numbers, PFam domains, and text search of gene descriptions. Users can use the Ortholog Group or Protein queries in the grey Search box to the left or the Searches menu in the header bar, or just type a search term in the 'Site search' box above which will result in a list of proteins and groups to explore. In addition, users can use a VEuPathDB Galaxy workflow to map their own set of proteins (e.g. protein sequences derived from a genome sequence of an organism) to OrthoMCL groups. See the [Assign Proteins to Groups](#) page.

For more information, see the [About OrthoMCL](#) and [OrthoMCL FAQ](#) pages.

Examining OrthoMCL output on gene record pages in FungiDB

- Go to the FungiDB gene record page for [CGB_L0350W](#), a hypothetical protein in *Cryptococcus gattii*.
 - 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?

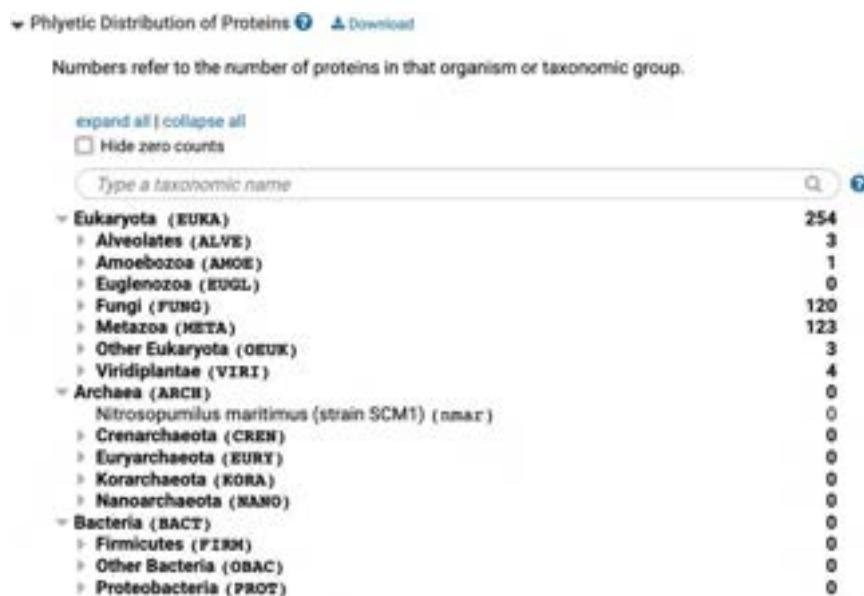


Ortholog Group	Gene	Product	Organism
OG6_106189	B1P03_002977	unspecified product	<i>Cryptococcus cf. gattii</i> MF34
	L203_046836	Cation efflux protein [Source:UniProt&THEMBL_Acc:ADA1E3C03]	<i>Cryptococcus depauperatus</i> CBS764
	BT14_061911	cation antiporter	<i>Cryptococcus gattii</i> CA1873
	OG6_06271	cation antiporter	<i>Cryptococcus gattii</i> EJB2
	BT11_056026	cation antiporter	<i>Cryptococcus gattii</i> NJ16

- 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
- Is this gene found in both Ascomycetes and Basidiomycetes?

- Does this protein have orthologs in Archaea and Bacteria (Hint: uncheck the box for "Hide zero counts")?

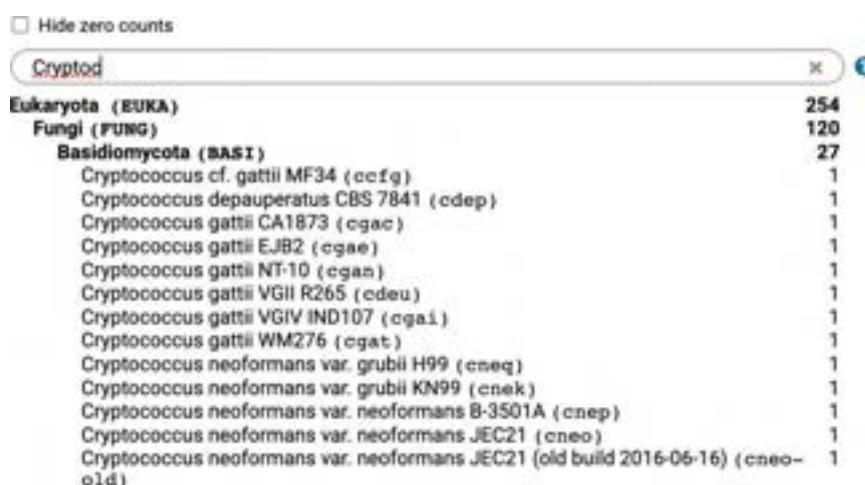
Phylogenetic distribution: Numbers refer to the number of proteins in that organism or taxonomic group. In order to see organisms and taxonomic groups without proteins in this ortholog group, uncheck 'Hide zero counts.'



Group summary breaks down summary by protein types: A core protein is from one of the 150 core species that were initially used to form 'core' groups. A peripheral protein is from a peripheral species whose entire proteome was mapped into the 'core' groups.

Peripheral proteins that do not map into a 'core' group are placed into residuals groups.

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





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

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 create protein alignments for *Cryptococcus* genes?

(Hint: Open List of All Proteins" section and use the "Search this table" filter to limit the alignment to "Cryptococcus", check all boxes, then hit "Run Clustal Omega for selected proteins" button at the bottom of this section).

To align sequences, select proteins from the table below. Then choose the 'Output format' and click the 'Run Clustal genes' button.

cne| X 6 rows (filtered from a total of 288)

Clustal Omega	Accession	Description	Organism	Taxon
<input checked="" type="checkbox"/>	cnev LQV05_001641	unknown	<i>Cryptococcus neoformans</i> strain:VNII	Fungi
<input checked="" type="checkbox"/>	cnek CKF44_05394	unknown	<i>Cryptococcus neoformans</i> var. grubii KN99	Fungi
<input checked="" type="checkbox"/>	cneq CNAG_05394	Cation:cation antiporter [Source:UniProtKB/TrEMBL;Acc:J9VZE1]	<i>Cryptococcus neoformans</i> var. grubii H99	Fungi
<input checked="" type="checkbox"/>	cneo-old CNH00620	cation:cation antiporter, putative	<i>Cryptococcus neoformans</i> var. neoformans JEC21 (old build 2016-06-16)	Fungi
<input checked="" type="checkbox"/>	cneo CNH00620	Cation:cation antiporter, putative [Source:UniProtKB/TrEMBL;Acc:Q5KCD4]	<i>Cryptococcus neoformans</i> var. neoformans JEC21	Fungi
<input checked="" type="checkbox"/>	cneq CNBL0590	unknown	<i>Cryptococcus neoformans</i> var. neoformans B-3501A	Fungi

Check All Uncheck All

Please note: selecting a large number of proteins will take several minutes to align.

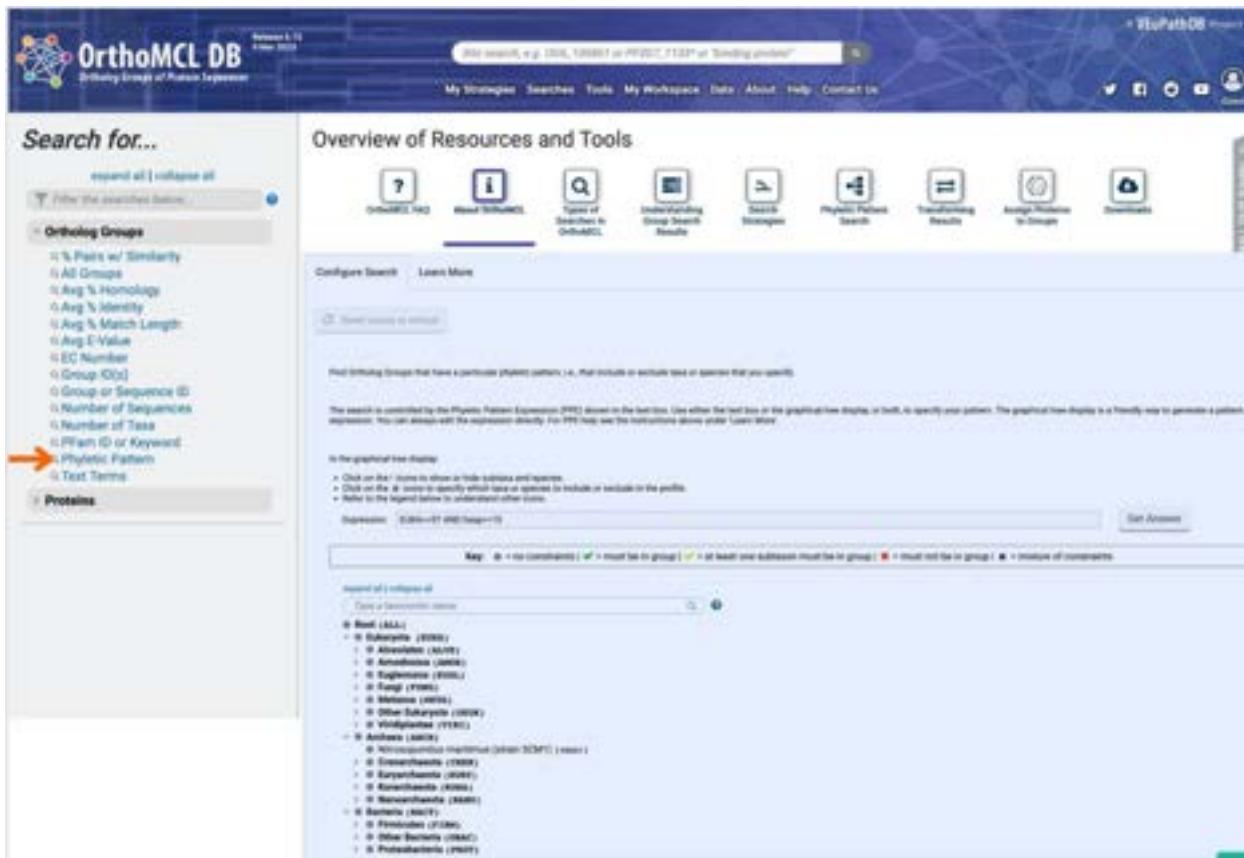
Output format: Mismatches highlighted

Run Clustal Omega for selected proteins

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 search interface. On the left, there is a sidebar titled "Search for..." with a dropdown menu set to "Ortholog Groups". Under "Ortholog Groups", several search criteria are listed, including "% Pairs w/ Similarity", "All Groups", "Avg % Homology", "Avg % Identity", "Avg % Match Length", "Avg E-Value", "EG Number", "Group ID", "Group or Sequence ID", "Number of Sequences", "Number of Taxa", "Pfam ID or Keyword", "Phyletic Pattern", "Text Terms", and "Proteins". An orange arrow points to the "Phyletic Pattern" option. The main content area is titled "Overview of Resources and Tools" and includes a row of icons for "OrthoMCL FAQ", "About OrthoMCL", "Types of Searches in OrthoMCL", "Understanding Group Search Results", "Search Strategies", "Phyletic Pattern Search", "Transforming Results", "Assign Proteins to Groups", and "Downloads". Below this is a "Configure Search" section with a "Learn More" link and a "Key" for search operators. The search bar contains the expression "EUKA>=5T AND hsap>=10". A "Get Answer" button is located to the right of the search bar.

There are two ways to specify a phyletic pattern:

1. Using the expression box.

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



The screenshot shows the search results for the expression "EUKA>=5T AND hsap>=10". The results are grouped under "Class of taxonomic source". The first group is "Bivalve (584)", which includes: *Bivalvia* (208), *Bivalvia* (50), *Ampullaria* (100), *Bivalvia* (100), *Ramp* (100), *Mollusca* (100), *Other Bivalviate* (100), *Vibriginea* (100). The second group is "Annelida (340)", which includes: *Annelida* (340), *Clitellata* (340), *Clitellata* (100), *Eurypharetida* (100), *Clitellata* (100), *Monopeltischaeta* (100), *Monopeltischaeta* (100), *Buridae* (100), *Prostomidae* (100), *Other Annelids* (100), *Prostomatoidea* (100). The search bar at the bottom contains the same expression "EUKA>=5T AND hsap>=10". A "Get Answer" button is located to the right of the search bar.

- 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. 1, 2, 5).

Examples

These expressions find ortholog groups in which...

hsap>=5

there are five or more human sequences

hsap+ecol=2

both human and E. coli are present.

hsap+ecol=1

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)
 - > 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)

- **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	Amoebozoa	Euglenozoans	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>

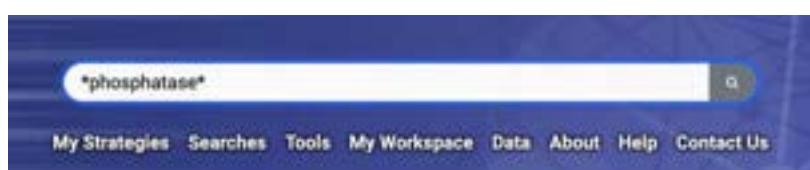
Useful information:

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

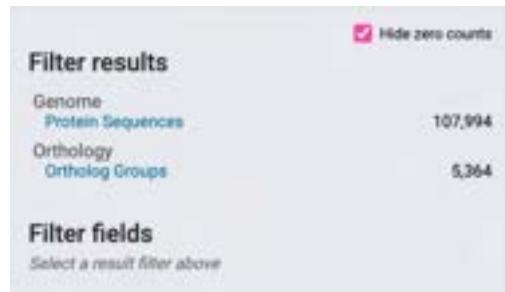
- Find all fungal proteins that are likely to be phosphatases and that do not have orthologs outside of fungal kingdom.
 - a. Use the site search to look for *phosphatase* (use asterisks to find any combination of the word “phosphatase”).



The screenshot shows the OrthoMCL search interface. In the top search bar, the query "*phosphatase*" is entered. Below the search bar, there is a navigation menu with links: My Strategies, Searches, Tools, My Workspace, Data, About, Help, and Contact Us. The main area displays a table titled "Filter results" with two rows:

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

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



The screenshot shows the OrthoMCL search results. At the top right is a checkbox labeled "Hide zero counts". Below it is a table titled "Filter results" with two rows:

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

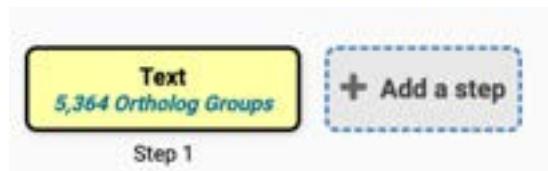
Below the table is a section titled "Filter fields" with the sub-instruction "Select a result filter above".

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



The screenshot shows the OrthoMCL search results. An orange arrow points to the "Ortholog Groups" row in the "Filter results" table. To the right of the table is a blue button labeled "Export as a Search Strategy" with the sub-instruction "to download or mine your results".

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

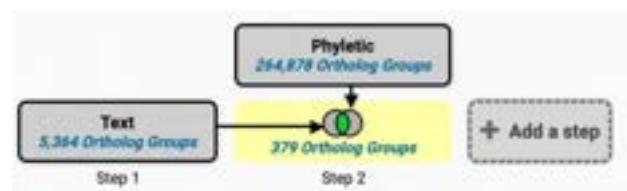


The screenshot shows the OrthoMCL search results. A yellow box highlights the "Text" field, which contains "5,364 Ortholog Groups". To the right of the "Text" field is a dashed box with a plus sign and the text "Add a step". Below the "Text" field is the label "Step 1".

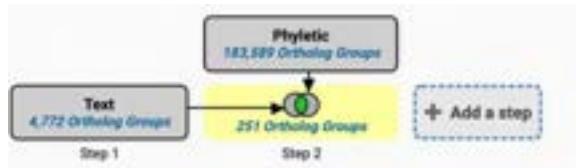
- c. Add a step and run a phyletic pattern search for groups that contain any fungi proteins but do not contain any other organism outside fungi. (Hint: make sure everything has a red X on it except for fungi, which should be a grey circle (no constraints)). How would this search be different if you used a green check instead of the grey circle for Fungi?

- * Root (ALL)
 - * Eukaryota (EUKA)
 - > ✘ Alveolates (ALVE)
 - > ✘ Amoebozoa (AMOE)
 - > ✘ Euglenozoia (EUGL)
 - > ⚡ Fungi (FUNG)
 - > ✘ Metazoa (META)
 - > ✘ Other Eukaryota (OEUKE)
 - > ✘ 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?

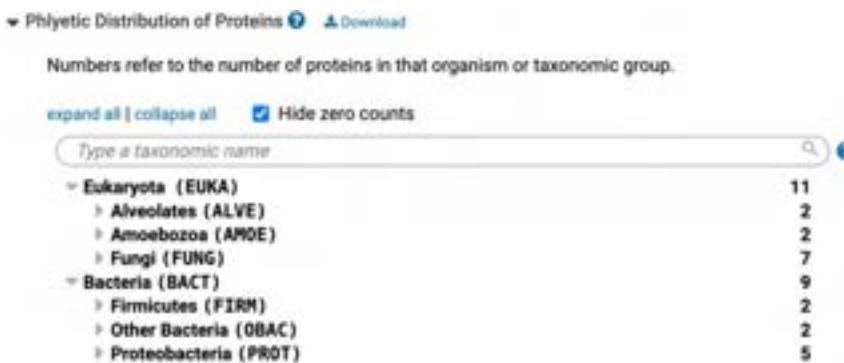


Strategy URL: <https://orthomcl.org/orthomcl/app/workspace/strategies/326574223/430551843>



Exploring a specific OrthoMCL group - examining the cluster graph.

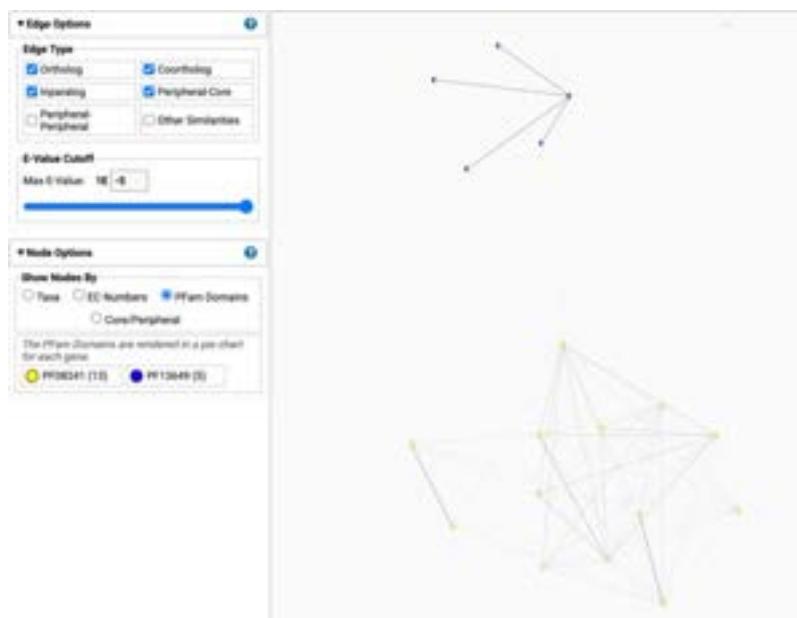
- Visit the OrthoMCL record page for the group **OG6_129371**
- Examine the phyletic distribution tree. What taxa does this group contain?



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

You can interact with the cluster graph. For example, move the slider to the left to increase to remove less significant edges connections between proteins. 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 PFam domains. Is there a pattern to the subclusters?



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-source, web-based platform for data analysis. It adheres to the FAIR principles of data sharing and re-use and eliminates the need for command-line scripting, enabling users to conduct, replicate, and disseminate comprehensive large-scale data analyses. In collaboration with Globus [1], the VEuPathDB project has established its own instance of Galaxy.

VEuPathDB Galaxy, accessible at <https://fungidb.org/fungidb/app/galaxy-orientation>, provides users with pre-loaded genomes, pre-configured workflows, and a suite of tools for private data analysis and visualization. Additionally, a tailored selection of tools facilitates the export of Galaxy results to private workspaces within VEuPathDB sites, accessible via the "My Workspace > My data sets" section. These datasets within the workspace can be explored using the FungiDB interface and tools, seamlessly integrating with public data housed in FungiDB. Accessing VEuPathDB Galaxy necessitates an account with FungiDB/VEuPathDB, which is freely available and applicable across all VEuPathDB genomics sites.

It's important to note that the VEuPathDB Galaxy instance isn't designed for long-term data storage, with datasets automatically purged after 60 days. To retain data, users are advised to download their analysis results locally and subsequently delete and purge files to create space for future analyses.

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.
- All Galaxy exercises will be conducted using the workshop instance of Galaxy (link provided below).
- This a group exercise module. Group activities within Galaxy will be organized into teams of four individuals. Only one person per group should deploy workflows within the workshop Galaxy environment. After the workflow is completed, everyone will get a copy of the workflow.

RNA-Seq analysis, Part I

Learning objectives:

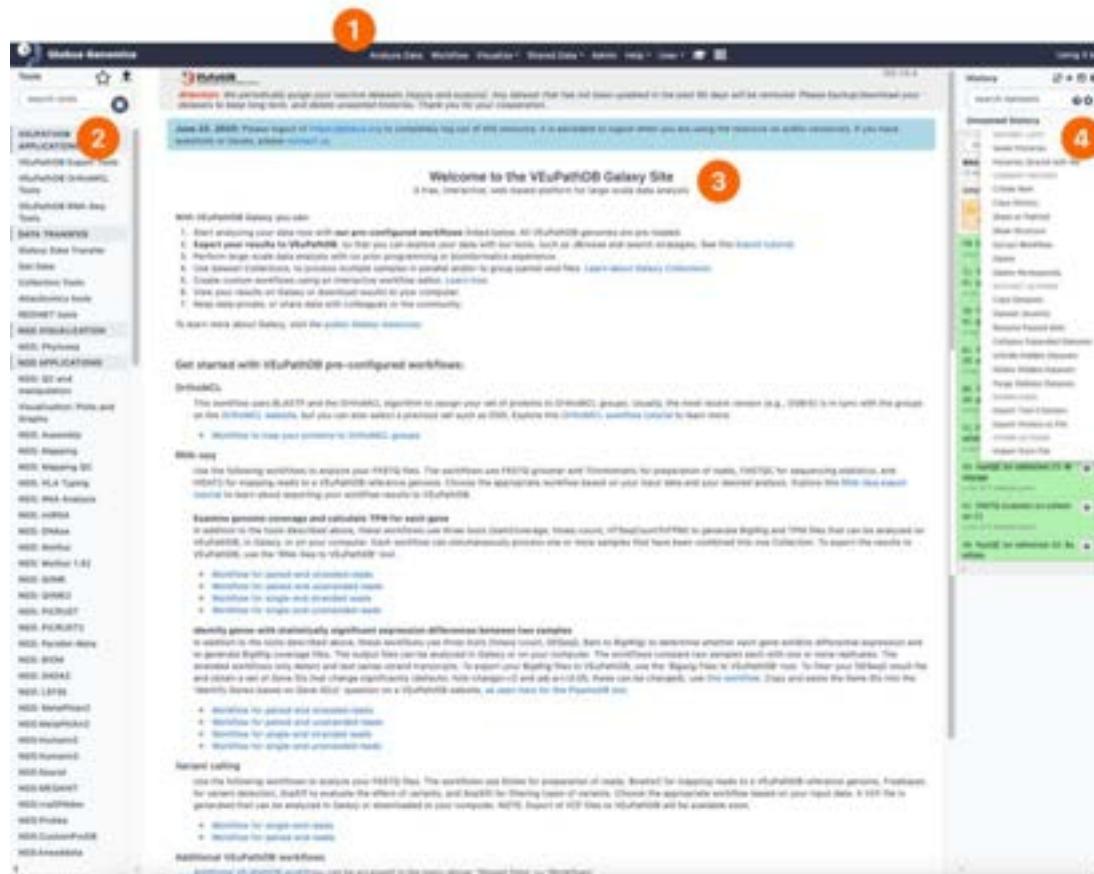
- Become familiar with the VEuPathDB Galaxy workspace.
- Upload raw data into Galaxy workspace and run a pre-configured SNP workflow

During this exercise, we will obtain raw sequence files from the "shared history" section in the workshop instance of VEuPathDB Galaxy. Subsequently, we will process these files using a pre- configured RNA-Seq workflow for paired or single-end reads. This workflow will entail aligning the data to a reference genome, computing gene expression, and interpreting the data.

The anatomy of the VEuPathDB Galaxy landing page.

The workspace comprises four major components:

1. The top menu, which governs the main interface, offers access to the landing page, shared data, public and private workflows, and additional features.
2. The left panel contains a list of available tools, with VEuPathDB export tools featured at the top.
3. The main welcome (landing) page serves as an interactive interface and houses pre- configured workflows, workflow editors, and more.
4. The right panel provides access to histories, and options to delete and purge datasets.



Note: Don't see Galaxy tools needed for your research? – Let us know by sending an email to help@fungidb.org

Important:

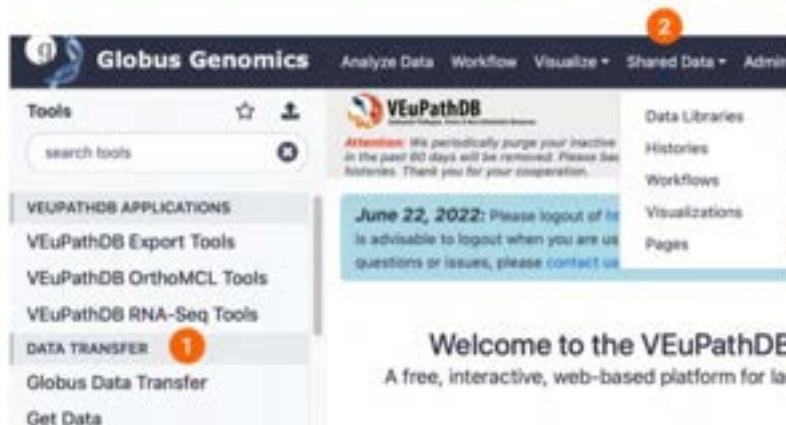
- If you do not have an account with VEuPathDB/FungiDB, please create one now.
 - **Access the workshop instance of VEuPathDB Galaxy.**
1. Click on the following URL to begin: <https://veupathdb1.globusgenomics.org/>
 2. On the next page, choose the “VEuPathDB” option and click on the ‘Continue’ button.
 3. If you are not already logged into VEuPathDB, you will be prompted to do so.
 4. Click ‘Continue’ on the next page (no need to link an existing account).
 5. Select “non-profit” and agree to the Terms of Service. Click ‘Continue’.
 6. When asked for grant permissions to use this Galaxy instance. Click ‘Allow’.



The diagram illustrates the workflow for logging into the VEuPathDB Galaxy instance. It consists of six numbered steps connected by arrows:

- Step 1:** The URL <https://veupathdb1.globusgenomics.org/> is entered in a browser. An orange circle labeled "1" is placed over the URL.
- Step 2:** The user is prompted to log in using their organizational credentials. A dropdown menu shows "VEuPathDB" selected. An orange circle labeled "2" is placed over the "Continue" button.
- Step 3:** The VEuPathDB login page is shown, featuring fields for "Username or Email" and "Password". An orange circle labeled "3" is placed over the "Continue" button.
- Step 4:** The user has successfully logged in, as indicated by the message "Welcome – You've Successfully Logged In". An orange circle labeled "4" is placed over the "Continue" button.
- Step 5:** The user is prompted to "Complete Your Sign Up For veupathdb1@veupathdb.org". They are asked to provide their "Name" and "Email" (both redacted here). They must also select the "Account will be used for" options ("non-profit research or educational purposes" and "commercial purposes") and agree to the "I have read and agree to the Globus Terms of Service and Privacy Policy". An orange circle labeled "5" is placed over the "Continue" button.
- Step 6:** The user is asked if they want to "View your identity", "Manage data using Globus Transfer", "View your email address", or "View identity details". These options are preceded by orange circles labeled "6". Below this, a note states: "By clicking 'Allow', you allow veupathdb1 (this client) to use the above listed information and services. You can rescind this and other consents at any time." At the bottom are "Allow" and "Deny" buttons.

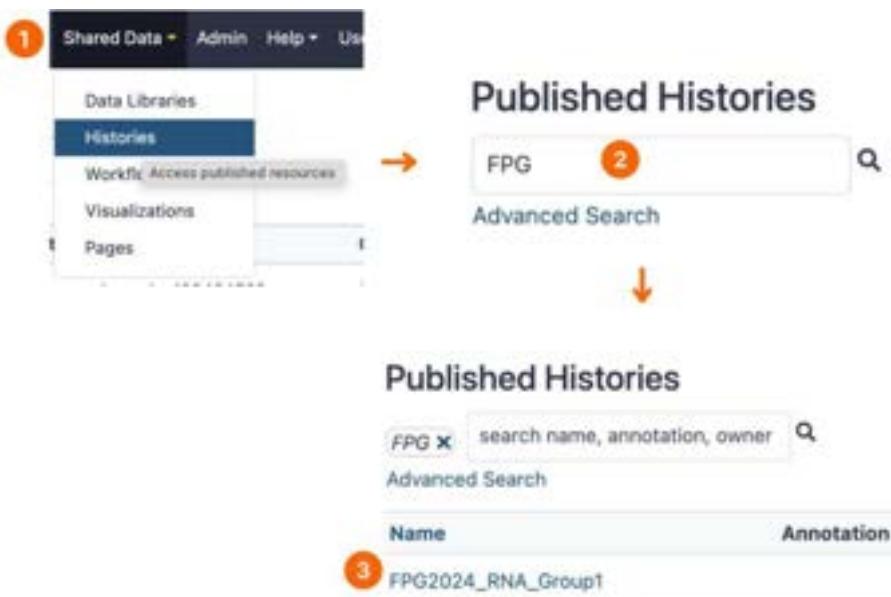
There are multiple ways to import data into your Galaxy workspace. For example, you can transfer data via tools located under the “Data Transfer” section in menu on the left (1). You can also transfer data from the “Shared Data” section in the main menu (2).



The screenshot shows the Globus Genomics interface. The top navigation bar includes "Analyze Data", "Workflow", "Visualize", "Shared Data", and "Admin". A sub-menu for "Shared Data" is open, showing options like "Data Libraries", "Histories", "Workflows", "Visualizations", and "Pages". A notification bar at the top right says "Attention: We periodically purge your inactive histories. If no activity occurs in the past 60 days, your histories will be removed. Please save histories. Thank you for your cooperation." Below the navigation, there's a "VEuPathDB" section with a message about logging out. The main content area is titled "Welcome to the VEuPathDB" and describes it as a free, interactive, web-based platform for large-scale bioinformatics analysis.

Important:

- In this exercise, we will use the ‘Shared data’ menu to access pre-loaded raw data.
- Only one person per each group should import data files and deploy an SNP workflow. Note: Everyone will get a chance to practice data analysis in the 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 the ‘Shared Data > Histories’ option.
 2. Filter all public workflows on “FPG”.
 3. Click on the history link that corresponds to your group number (e.g., FPG2024_RNA_Group1) to import the data into the Galaxy workspace.



The image consists of two screenshots. The top screenshot shows the "Published Histories" search interface. The search bar contains "FPG" and has a magnifying glass icon. Below the search bar is an "Advanced Search" button. The results table has columns for "Name" and "Annotation". One row is highlighted with a red circle labeled 3, which corresponds to the history "FPG2024_RNA_Group1". The bottom screenshot shows the details of the selected history. It includes a "Name" field containing "FPG2024_RNA_Group1", an "Annotation" field, and a "Description" field that reads "This history was created by the FPG group for the RNA-seq analysis of Group 1 samples. It contains raw RNA-seq data and a workflow for processing it using the VEuPathDB OrthoMCL tool." There is also a "View" button.



Group assignments.

Group 1 *Candida auris*. Analyze transcriptomes from cells grown under high concentrations of tunicamycin. Control: no drug. Single-read data.

Comparison	No drug vs Tunicamycin
History name for download (in Galaxy)	FPG2024_RNA_Group1
Ref genome (in Galaxy)	FungiDB-37_CaurisB8441_Genome

Reference: PMID: n/a BioProject: PRJEB60034

Group 2 *Rhizopus delemar*. Analyze transcriptomes from germinated spores. Control: 0 h. Paired-end data.

Comparison	0 h vs 6 h
History name for download (in Galaxy)	FPG2024_RNA_Group2
Ref genome (in Galaxy)	FungiDB-29_RdelemarRA99-880_Genome

Reference: PMID: PRJNA472797 BioProject: PRJNA472797

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

Comparison	Planktonic vs Biofilm
History name for download (in Galaxy)	FPG2024_RNA_Group3
Ref genome (in Galaxy)	FungiDB-42_CparapsilosisCDC317_Genome

Reference: PMID: 25233198 BioProject: PRJNA246482

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

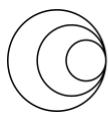
Comparison	Mycelia vs Spherules
History name for download (in Galaxy)	FPG2024_RNA_Group4
Ref genome (in Galaxy)	FungiDB-61_CposadasiiSilveira2022_Genome

Reference: PMID: 22911737 BioProject: PRJNA169242

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

Comparison	Spores vs Mycelia
History name for download (in Galaxy)	FPG2024_RNA_Group5
Ref genome (in Galaxy)	FungiDB-31_FgraminearumPH-1_Genome

Reference: PMID: 24625133 BioProject: PRJNA239711

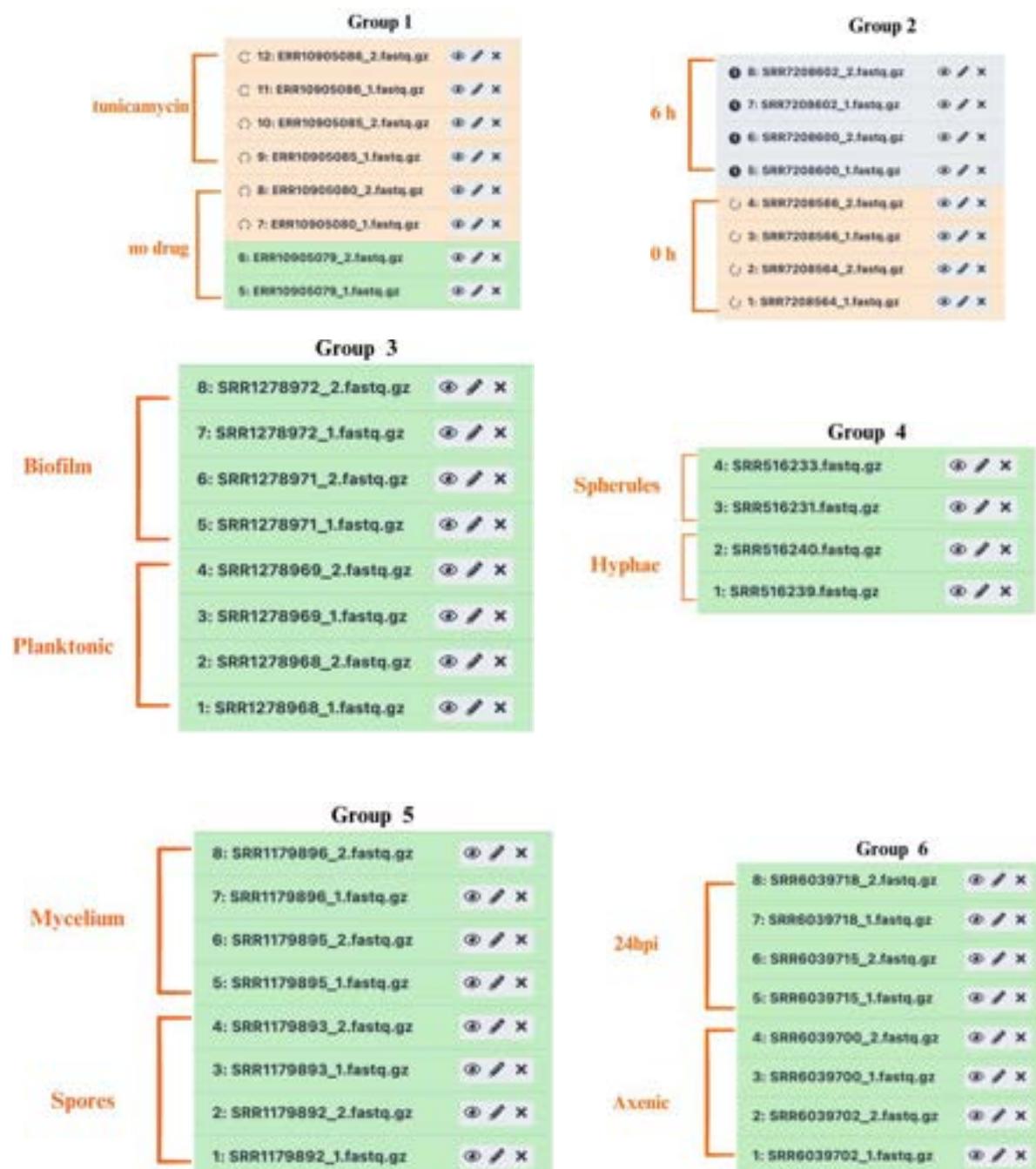


Group 6 *Ustilago maydis*. Analyze transcriptomes from plant-associated development samples (axenic culture vs 24 hours post infection (hpi)). Paired-end data.

Comparison	axenic vs 24hpi
History name for download (in Galaxy)	FPG2024_RNA_Group6
Ref genome (in Galaxy)	FungiDB-51_Umaydis521_Genome

Reference: PMID: 33653886 BioProject: PRJNA407369

Guide to 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 of Galaxy workflows.

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

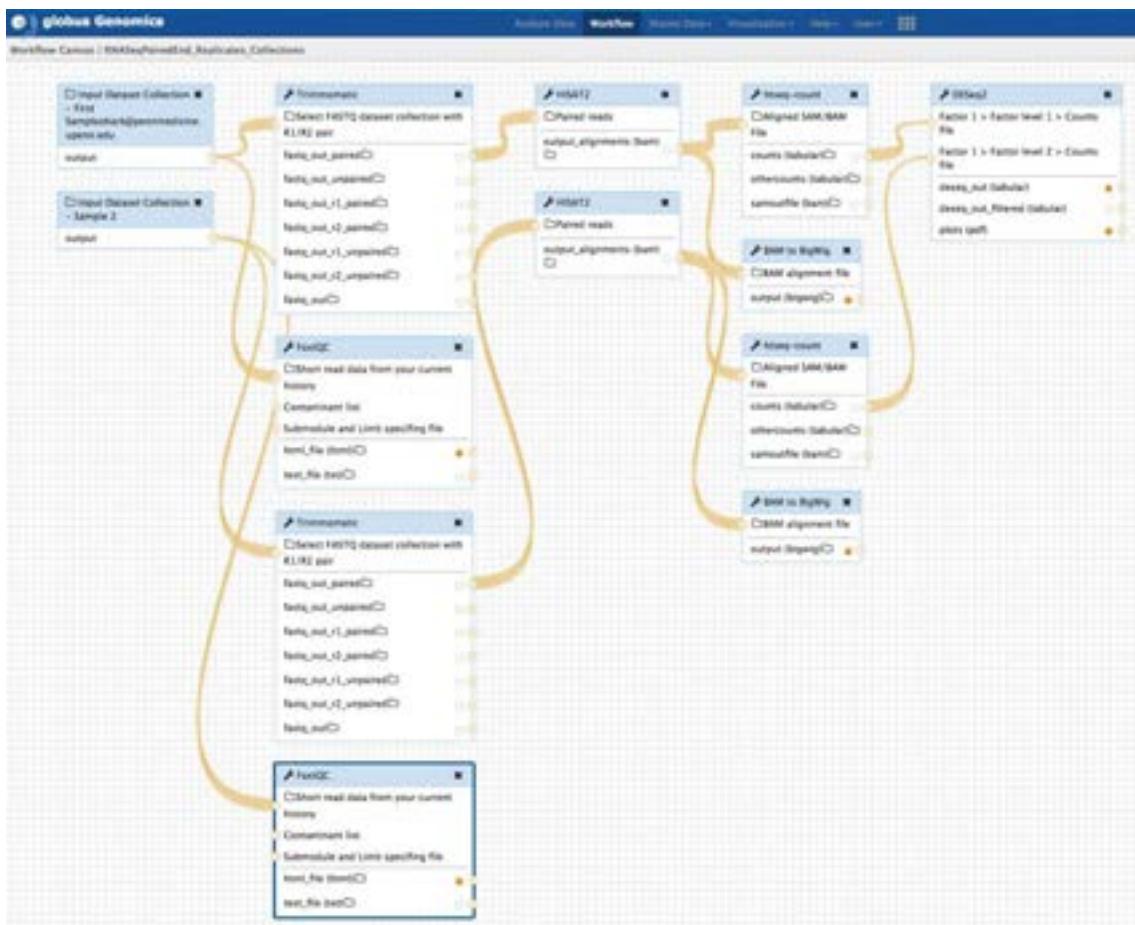
1. Click on the checkbox function “operation on multiple datasets”.
2. Select samples that belong to the same condition (control samples will appear at the bottom, see file mapping notes for each group below).
3. Click on “For all selected” and choose “Build List of Dataset Pairs”.
4. Name the sample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

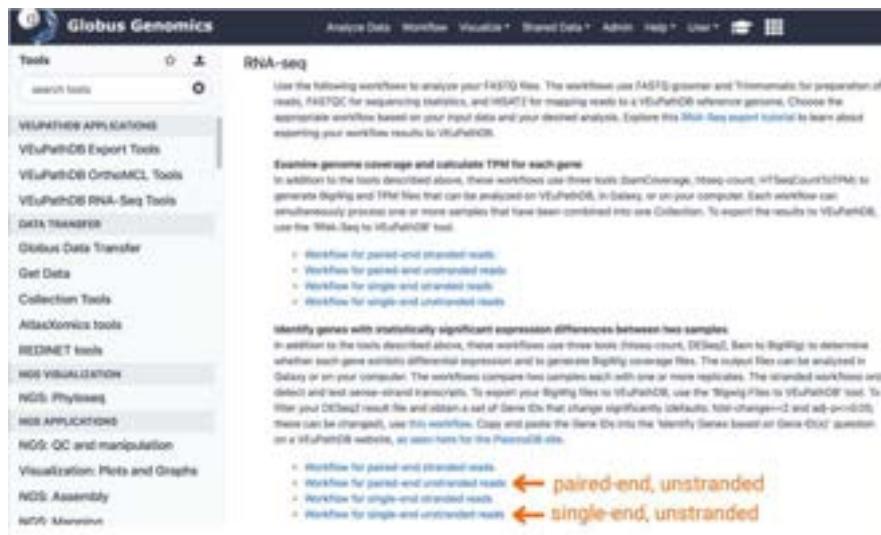
1. Input: raw data, dataset collections.
2. FASTQC: analyse for quality, generate read quality reports.
3. Trimmomatic: trims the reads based on their quality scores and adaptor sequences.
4. HISAT2: align reads to a reference and generate coverage plots.
5. HTSeq: estimate abundance (read counts per 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”.



Use the following workflow to analyze your FASTQ files. This workflow uses FASTQ-gz compressor 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 expert tutorial to learn about submitting 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 (htseq-count, htSeqCountToTPM) to generate BigWig and TDF files that can be visualized 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 'Workflow 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 (DESeq2, DESeq), 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 commandline. 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' files to VEuPathDB tool. To filter your DESeq2 result file and obtain a set of genes that change significantly (defaults: fold-change=2 and adj-pval<0.05), these can be changed, use this workflow. Clean and pack the gene IDs into file. Identify genes based on GeneID/Chromosome on a VEuPathDB website, as user name for file (User@VDB).

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

← paired-end, unstranded
← single-end, unstranded

- **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

1. 1. Input Dataset Collection - Sample 1
13 samples

2. 2. Input Dataset Collection - Sample 2
19 samples

3. HISAT2_Genome (Galaxy Version 2.0.0)

4. FastQC (Galaxy Version: FastQC_0.11.0)

5. HISAT2_Genome (Galaxy Version 2.0.0)

6. FastQC (Galaxy Version: FastQC_0.11.0)

7. Trimmomatic (Galaxy Version: 0.36.0)

8. Trimmomatic (Galaxy Version: 0.36.0)

9. HISAT2_Genome Version: 2.0.0

10. HISAT2_Genome Version: 2.0.0

11. BAM to BED (Galaxy Version: 0.2.0)

12. htseq-count - You can use exon or CDS as feature type. You must use exon_id as ID Attribute. Galaxy Version: HTSEQCOUNT_0.1.0 SAMTOOLS_1.3 PICARD_1.134

13. htseq-count - You can use exon or CDS as feature type. You must use exon_id as ID Attribute. Galaxy Version: HTSEQCOUNT_0.1.0 SAMTOOLS_1.3 PICARD_1.134

14. BAM to BED (Galaxy Version: 0.2.0)

15. DESeq2 (Galaxy Version: 2.11.40.0) (Galaxy Version: 2.11.40.0)

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 2.0.0)

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

10. HISAT2 (Galaxy Version 2.0.0)

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

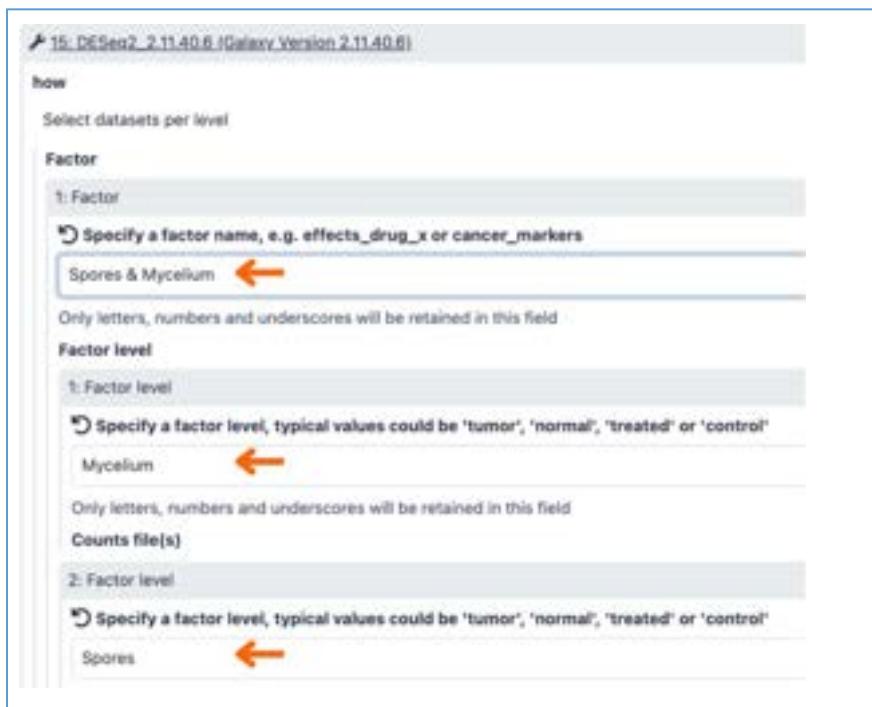
13. htseq-count - You can use exon or CDS as feature type.

Aligned SAM/BAM File
 Is this library mate-paired?
paired-end

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

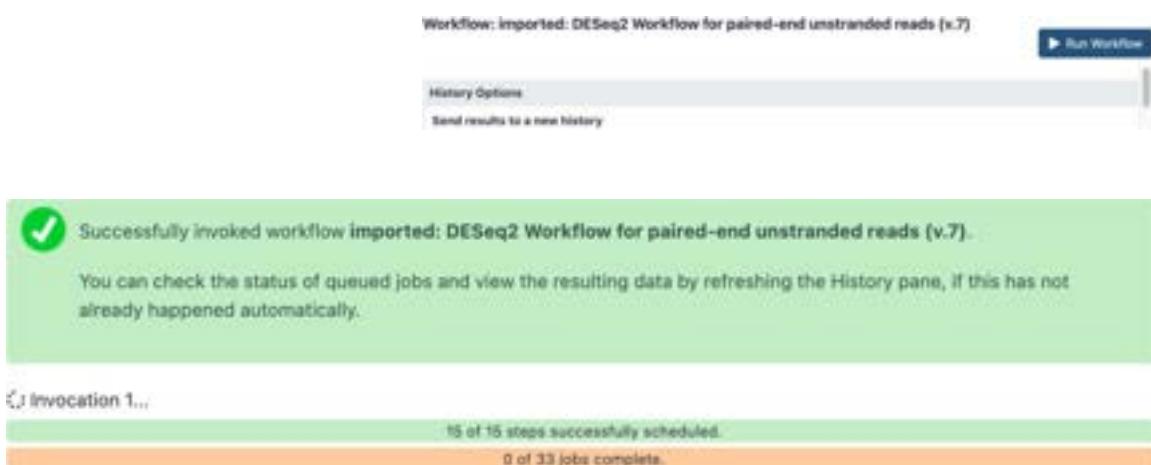
Select a genome annotation
FungiDB-31_FgraminearumPh-1_Genome

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



The screenshot shows the configuration of a DESeq2 workflow. It includes sections for 'Factor' and 'Factor level'. The 'Factor' section has a field labeled '1: Factor' containing 'Specify a factor name, e.g. effects_drug_x or cancer_markers' with the value 'Spores & Mycelium' highlighted by a red arrow. The 'Factor level' section contains two entries: '1: Factor level' with 'Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'' and the value 'Mycelium' highlighted by a red arrow, and '2: Factor level' with the same specification and the value 'Spores' highlighted by a red arrow.

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



The screenshot shows the Galaxy interface after running a workflow. It displays a success message: "Successfully invoked workflow imported: DESeq2 Workflow for paired-end unstranded reads (v.7)." Below this, it says "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." At the bottom, there is a progress bar for "Invocation 1..." showing "15 of 15 steps successfully scheduled" and "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.

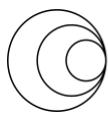


How to work with Galaxy editor (optional)

You can create your own workflows. An interactive workflow editor allows you to add and configure tools.

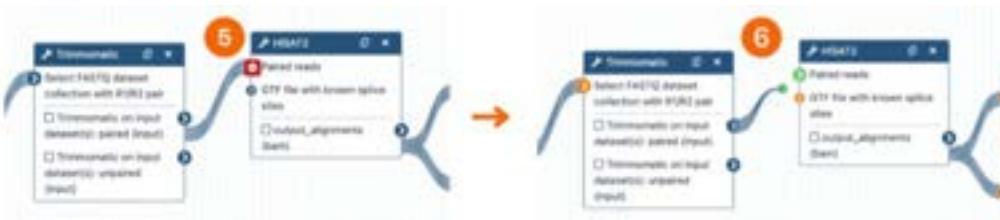
1. Navigate to the “Shared Data” menu.
2. Click on “Workflows”.
3. Left-click on the “FPG2023 workflow editor practice” work to “import”
4. Once the workflow is imported in your workspace, left-click and select “edit”.



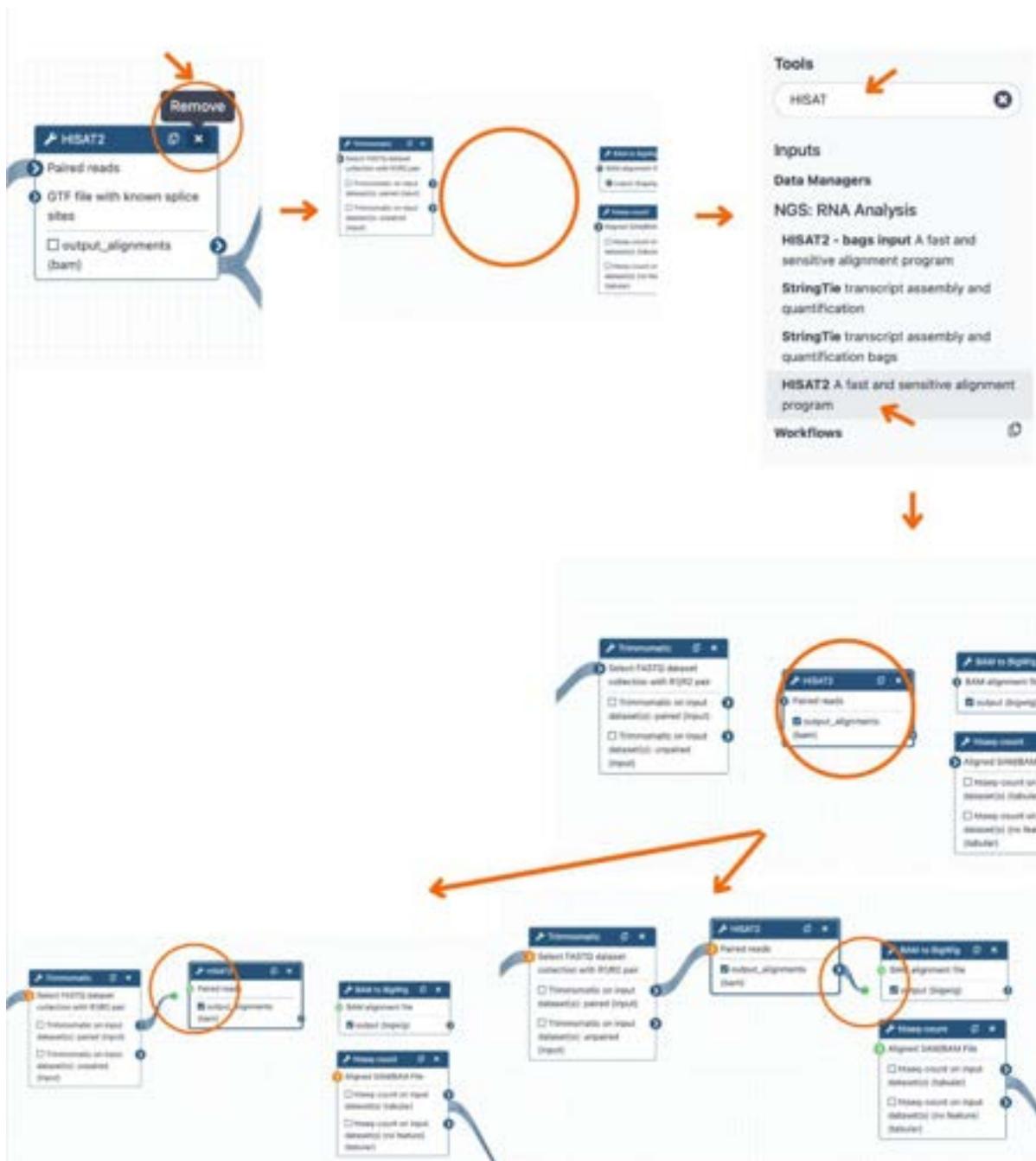


Once you are in the workflow editor:

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 a connection. When this happens, take a look at the tool documentation notes in the right panel, and 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 principles 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.

Variant Calling analysis, Part I.

Learning objectives:

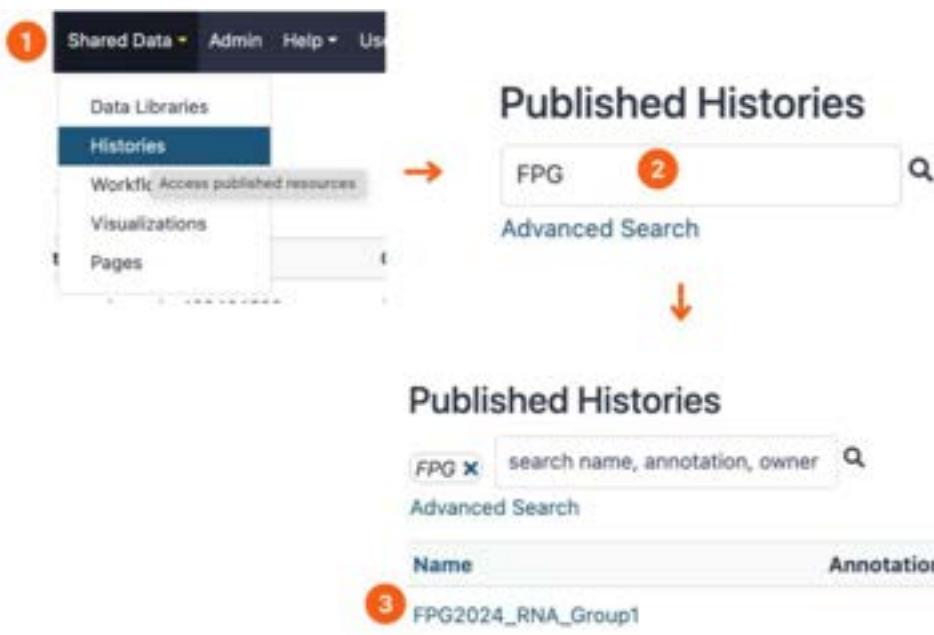
- Become familiar with the VEuPathDB Galaxy workspace.
- Upload raw data into Galaxy workspace and run a pre-configured SNP workflow

Important:

- In this exercise, we will use the ‘Shared data’ menu to access pre-loaded raw data.
- Only one person per each group should import data files and deploy a SNP workflow. Note: Everyone will get a chance to practice data analysis in the 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 “FPG2024” ..
3. Click on the history link that correspond to your group number (e.g., FPG2024_SNP_Group1) to import the data into the Galaxy workspace.



The screenshot illustrates the workflow for importing data:

1. The top navigation bar shows 'Shared Data' selected, highlighted with a red circle.
2. The search bar contains the filter 'FPG', also highlighted with a red circle.
3. The search results show a single entry: 'FPG2024_RNA_Group1', which is also highlighted with a red circle.

Group assignments.

Group 1 *Aspergillus fumigatus*. AFIS2503 clinical isolate from pleural fluid of a patient. Paired-end data.

History name	FPG2024_SNP_Group1
Ref genome (in Galaxy)	FungiDB-29_AfumigatusAf293_Genome

Group 2 *Aspergillus fumigatus*. AFIS1415 clinical isolate from pleural fluid of a patient. Paired-end data.

History name	FPG2024_SNP_Group2
Ref genome (in Galaxy)	FungiDB-29_AfumigatusAf293_Genome

Group 3 *Zymoseptoria tritici*. ST16CH_1A27 isolate collected from common wheat (*Triticum aestivum*) in Eschikon, Switzerland.

Paired-end data.

History name	FPG2024_SNP_Group3
Ref genome (in Galaxy)	FungiDB-34_ZtriticiPO323_Genome

Group 4 *Zymoseptoria tritici*. ORE15_Mad_G1isolate collected from common wheat (*Triticum aestivum*) in Oregon, USA.

Paired-end data.

History name	FPG2024_SNP_Group4
Ref genome (in Galaxy)	FungiDB-34_ZtriticiPO323_Genome

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

History name	FPG2024_SNP_Group5
Ref genome (in Galaxy)	FungiDB-37_CaurisB8441_Genome

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

History name	FPG2024_SNP_Group6
Ref genome (in Galaxy)	FungiDB-37_CaurisB8441_Genome

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

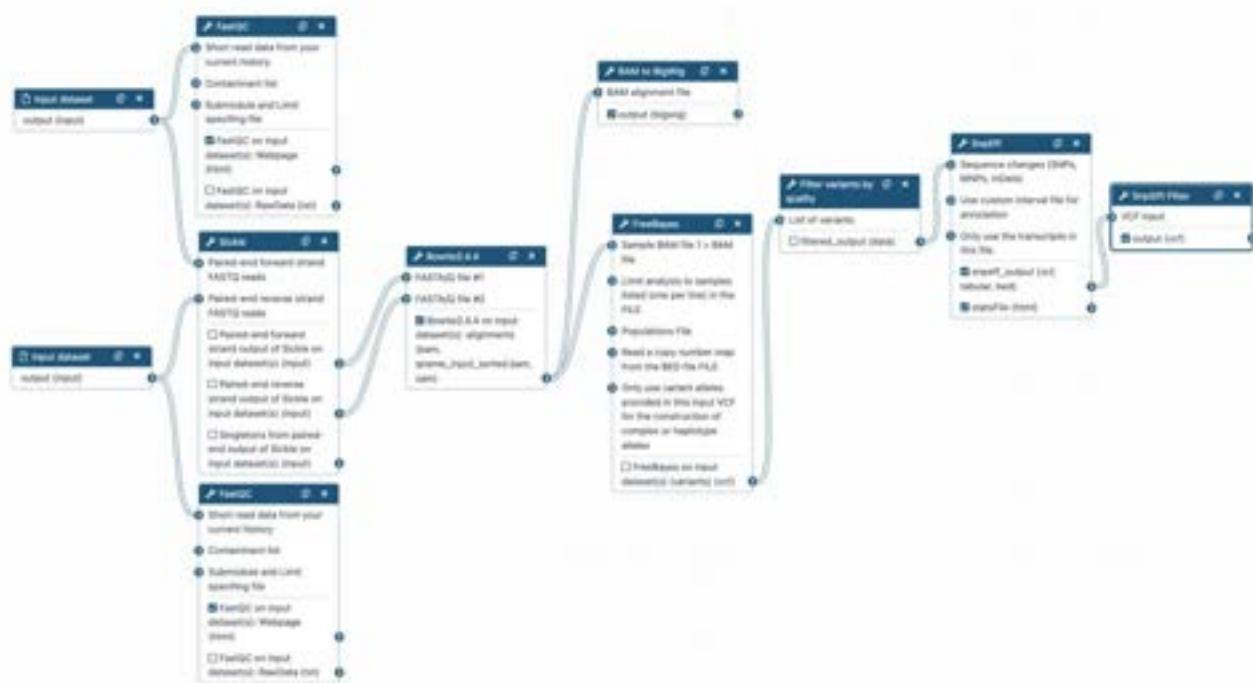
Variant calling

Use the following workflows to analyze your FASTQ 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](#)

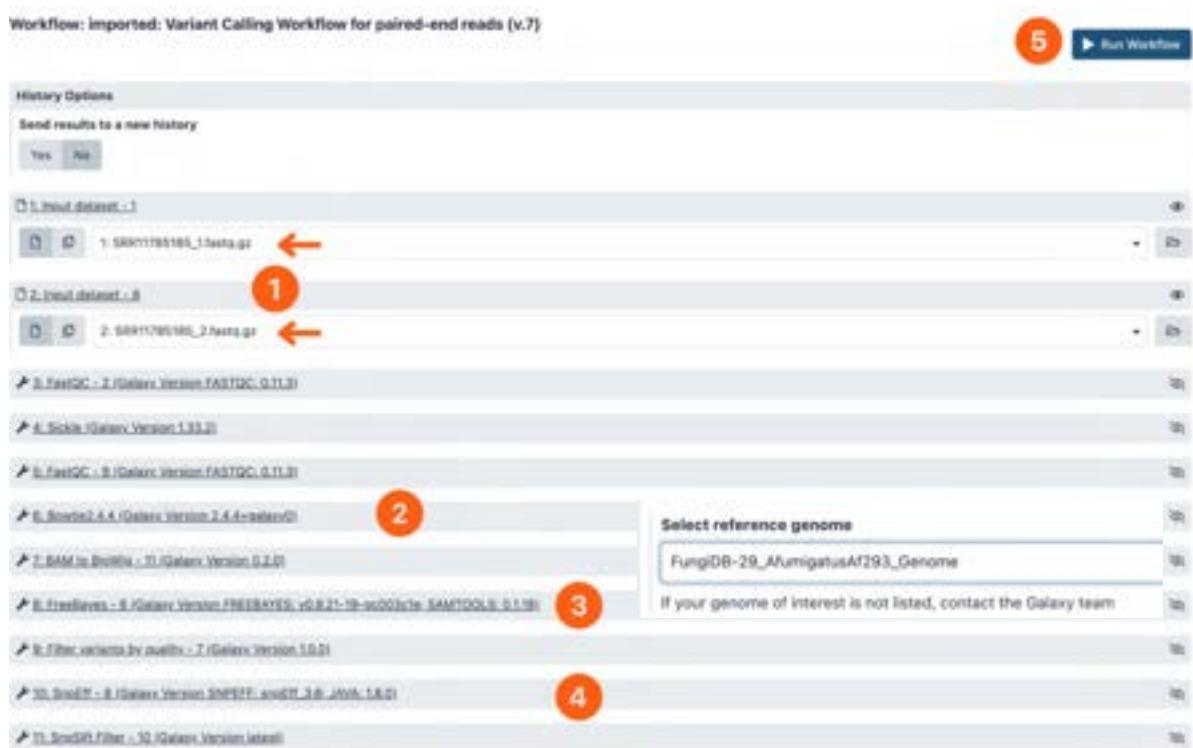
The pre-configured workflows follow these steps:

- 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.



- Define 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 (Default will have the same file selected for both input files). Hint: for single read data, you will have only one file.
2. Select reference genome for Bowtie2
Hint: or reference genome information, see group assignment table above).
3. Repeat genome selection for FreeBayes.
4. Select the same genome for SnpEff.
5. Click Run Workflow.



References

1. Foster I. 2011. Globus online: accelerating and democratizing science through cloud-based services. IEEE Internet Comput. 15(3):70–73.
doi:10.1109/MIC.2011.64.

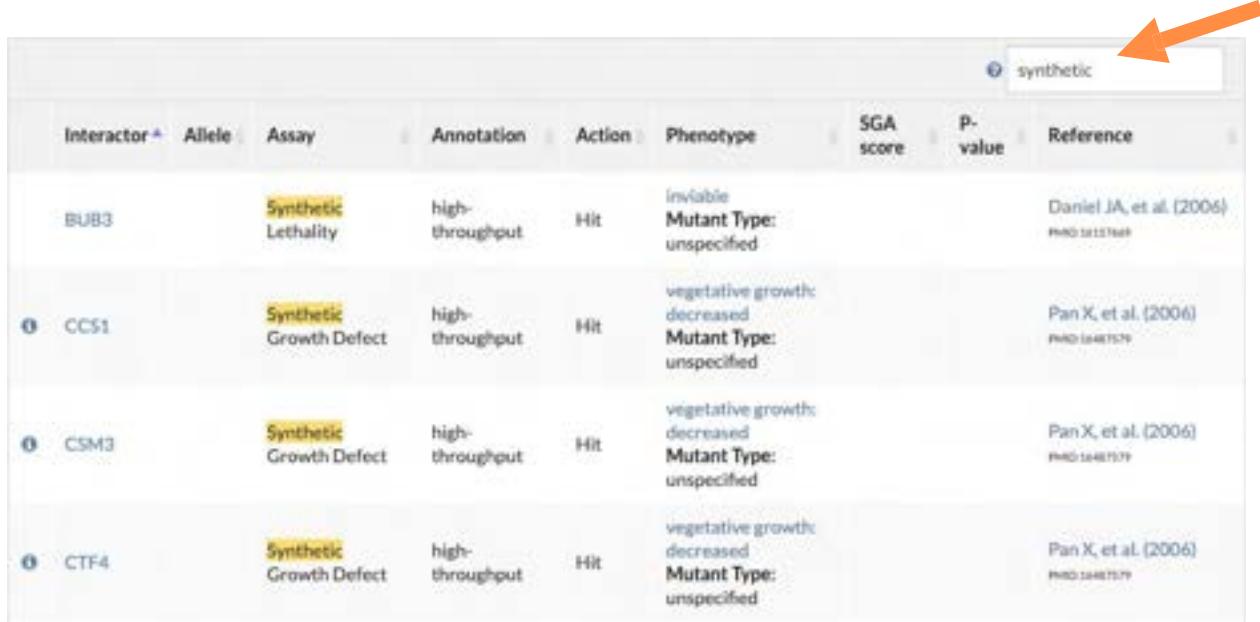
Using SGD GO Slim Mapper and Interaction Data to Predict Gene Function

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, it 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.



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:16518848
CCS1		Synthetic Growth Defect	high-throughput	Hit	vegetative growth: decreased Mutant Type: unspecified			Pan X, et al. (2006) PMID:16481579
CSM3		Synthetic Growth Defect	high-throughput	Hit	vegetative growth: decreased Mutant Type: unspecified			Pan X, et al. (2006) PMID:16481579
CTF4		Synthetic Growth Defect	high-throughput	Hit	vegetative growth: decreased Mutant Type: unspecified			Pan X, et al. (2006) PMID:16481579

- 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

Filter table

- 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 (> 300), 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 statements in that category.

SELECT ALL Terms from Yeast GO-Slim: process 

DNA recombination, GO:0006310
DNA repair, GO:0006281
DNA replication, GO:0006280
DNA-templated transcription, elongation ; GO:0006254

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

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

Search Results

Save Options: HTML Table | Plain Text | Tab-delimited | Your Input List of Genes | Your GO Slim List

GO version 2023-04-01

GO Terms from the biological process Ontology			
GO Term (GO ID)	Genes Annotated to the GO Term	GO Term Usage in Gene List	Genome Frequency of Use
DNA replication (GO:0004260)	YMR048W, YNL273W, YDR080W, YPR135W	4 of 13 genes, 30.77%	140 of 6489 annotated genes, 2.16%
regulation of DNA metabolic process (GO:0051052)	YLR233C, YMR048W, YNL273W, YCR080W	4 of 13 genes, 30.77%	106 of 6489 annotated genes, 1.66%
mitotic cell cycle (GO:0000278)	YGL084W, YJL030W, YMR048W, YOR022W	4 of 13 genes, 30.77%	317 of 6489 annotated genes, 4.89%
protein modification by small protein conjugation or removal (GO:0070447)	YDR510W, YJL030W, YCR021W, YOR080W	4 of 13 genes, 30.77%	137 of 6489 annotated genes, 2.11%
regulation of cell cycle (GO:0051726)	YGL084W, YJL030W, YNL273W, YOR022W	4 of 13 genes, 30.77%	247 of 6489 annotated genes, 3.81%
chromosome segregation (GO:0007059)	YGL084W, YJL030W, YMR048W	3 of 13 genes, 23.08%	162 of 6489 annotated genes, 2.50%

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

GO Enrichment, Phenotype Data at CGD

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

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

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

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

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 PM78 Phenotype details and references

Classical genetics	<ul style="list-style-type: none"> ▪ hyphal growth: decreased ▪ hyphal growth: normal ▪ resistance to Hygromycin B: decreased ▪ viable
heterozygous null	
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 Calcifluor 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

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 : **Candida albicans**

Step 2: Query Set (Your Input)

Enter Gene/ORF names:
(separated by a return or a space)
C3_07710W_A C1_02360C_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:
Choose File: no file selected

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

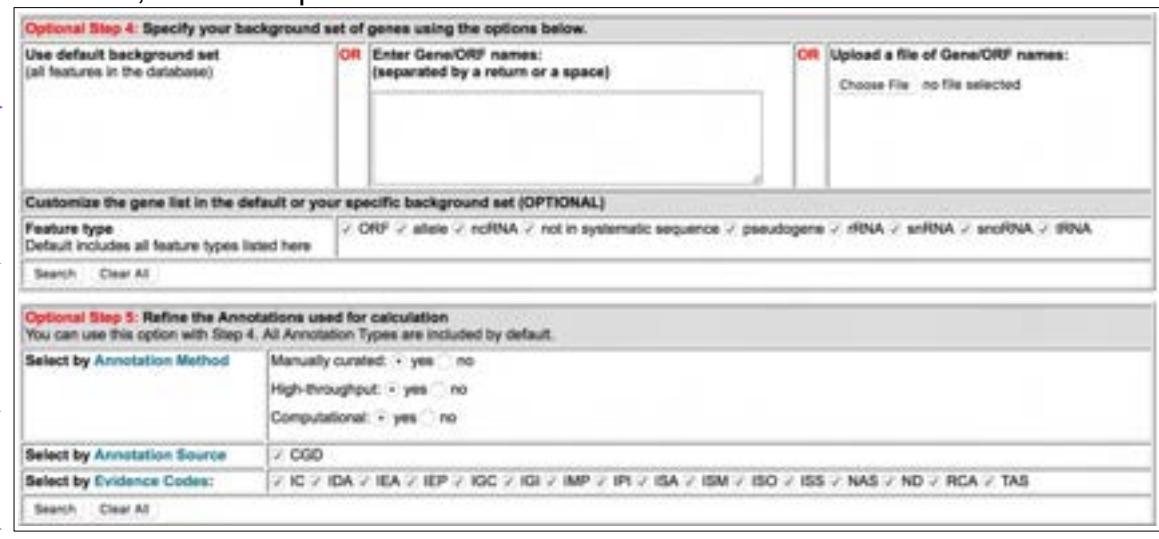
- Process
- Function
- Component

Search using [default settings](#) or use Step 4 and/or Step 5 below to customize your options.

Search Clear All

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: 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

✓ ORF ✓ allele ✓ ncRNA ✓ not in systematic sequence ✓ pseudogene ✓ rRNA ✓ snRNA ✓ snoRNA ✓ tRNA

Search 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.

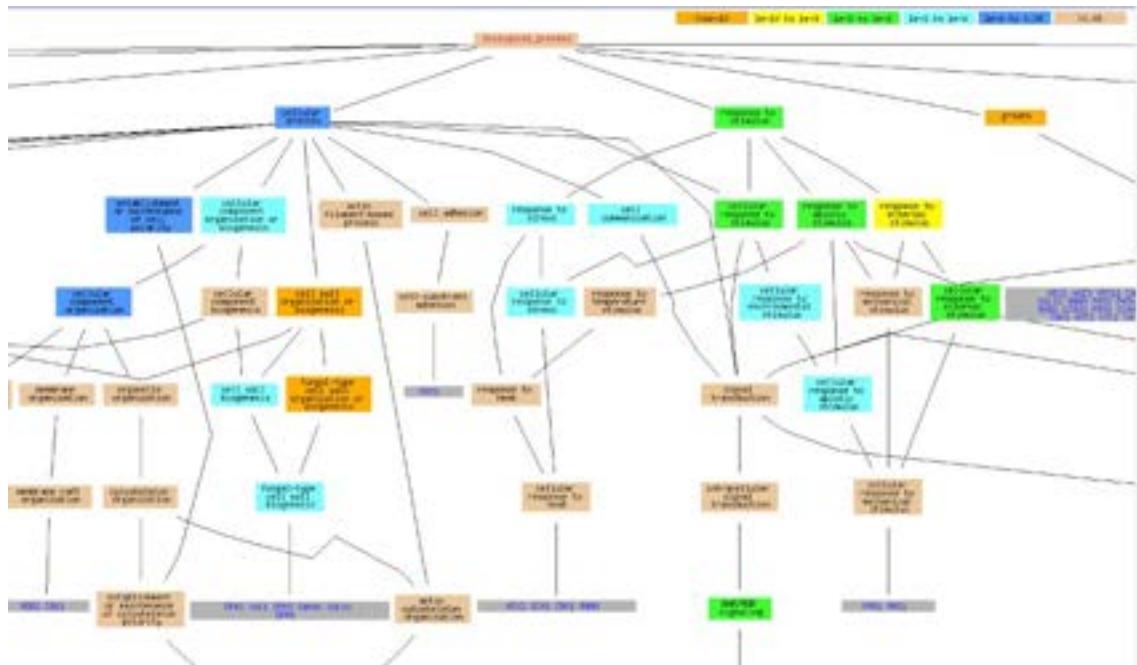
Select by Annotation Method	Manually curated: <input checked="" type="checkbox"/> yes <input type="checkbox"/> no
	High-throughput: <input checked="" type="checkbox"/> yes <input type="checkbox"/> no
	Computational: <input checked="" type="checkbox"/> yes <input type="checkbox"/> no

Select by Annotation Source: ✓ CGD

Select by Evidence Codes: ✓ IC ✓ IDA ✓ IEA ✓ IEP ✓ IGC ✓ IGI ✓ IMP ✓ IPI ✓ ISA ✓ ISM ✓ ISO ✓ ISS ✓ NAS ✓ ND ✓ RGA ✓ TAS

Search Clear All

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



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

Terms from the Process Ontology

Gene Ontology term	Cluster frequency	Background frequency	Corrected P-value	False discovery rate	Genes annotated to the term
cell wall organization or biogenesis [AmiGO]	27 out of 41 genes, 65.3%	242 out of 6473 background genes, 3.7%	6.92e-27	0.00%	CASA, CBK1, CWH1, DPM1, DPM2, DPM3, ECM33, GAL10, HBD1, HYM1, KIC1, MNW9, MNST1, MOB2, OCH1, PMR1, PMT1, PMT4, RHBL1, ROT2, SAC1, SAP10, SAP9, SEC23, S002, VPC1, VPC2, VPC3
fungi-type cell wall organization or biogenesis [AmiGO]	25 out of 41 genes, 61.0%	213 out of 6473 background genes, 3.3%	6.17e-25	0.00%	CASA, CBK1, CWH1, DPM1, DPM2, DPM3, ECM33, GAL10, HBD1, HYM1, KIC1, MNW9, MNST1, MOB2, PMR1, PMT1, PMT4, RHBL1, ROT2, SAC1, SAP10, SAP9, SEC23, S002, VPC1
glycoprotein metabolic process [AmiGO]	16 out of 41 genes, 43.9%	130 out of 6473 background genes, 2.0%	5.31e-18	0.00%	CWH1, DPM1, DPM2, DPM3, GAL10, MNH14, MNH9, MNST1, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHBL1, ROT2, SAC1, VRG4
macromolecule glycosylation [AmiGO]	16 out of 41 genes, 39.0%	117 out of 6473 background genes, 1.8%	1.48e-15	0.00%	CWH1, DPM1, DPM2, DPM3, GAL10, MNH14, MNW9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHBL1, SAC1, VRG4
protein glycosylation [AmiGO]	16 out of 41 genes, 39.0%	117 out of 6473 background genes, 1.8%	1.48e-15	0.00%	CWH1, DPM1, DPM2, DPM3, GAL10, MNH14, MNW9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHBL1, SAC1, VRG4
glycosylation [AmiGO]	16 out of 41 genes, 38.5%	118 out of 6473 background genes, 1.8%	1.69e-15	0.00%	CWH1, DPM1, DPM2, DPM3, GAL10, MNH14, MNW9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHBL1, SAC1, VRG4
glycoprotein biosynthetic process [AmiGO]	16 out of 41 genes, 39.0%	121 out of 6473 background genes, 1.9%	2.57e-15	0.00%	CWH1, DPM1, DPM2, DPM3, GAL10, MNH14, MNW9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHBL1, SAC1, VRG4
fungi-type cell wall organization [AmiGO]	17 out of 41 genes, 41.5%	155 out of 6473 background genes, 2.4%	4.88e-15	0.00%	CASA, CBK1, ECM33, HBD1, HYM1, KIC1, MNW9, MOB2, PMR1, PMT1, PMT4, RHBL1, SAP10, SAP9, SEC23, S002, VPC1
external encapsulating structure organization [AmiGO]	17 out of 41 genes, 41.5%	156 out of 6473 background genes, 2.4%	5.45e-15	0.00%	CASA, CBK1, ECM33, HBD1, HYM1, KIC1, MNW9, MOB2, PMR1, PMT1, PMT4, RHBL1, SAP10, SAP9, SEC23, S002, VPC1
cell wall organization [AmiGO]	17 out of 41 genes, 41.5%	156 out of 6473 background genes, 2.4%	5.45e-15	0.00%	CASA, CBK1, ECM33, HBD1, HYM1, KIC1, MNW9, MOB2, PMR1, PMT1, PMT4, RHBL1, SAP10, SAP9, SEC23, S002, VPC1
filamentous growth [AmiGO]	26 out of 41 genes, 63.4%	626 out of 6473 background genes, 9.7%	1.84e-14	0.00%	AGE3, CASA, CBK1, CWH1, ECM33, GAL10, HYM1, KEX2, KIC1, MNW9, MNST1, MOB2, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHBL1, ROT2, SAC1, SAP10, SAP9, SEC23, S002, VPC1, VRG4
growth [AmiGO]	26 out of 41 genes, 63.4%	653 out of 6473 background genes, 9.8%	2.43e-14	0.00%	AGE3, CASA, CBK1, CWH1, ECM33, GAL10, HYM1, KEX2, KIC1, MNW9, MNST1, MOB2, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHBL1, ROT2, SAC1, SAP10, SAP9, SEC23, S002, VPC1, VRG4

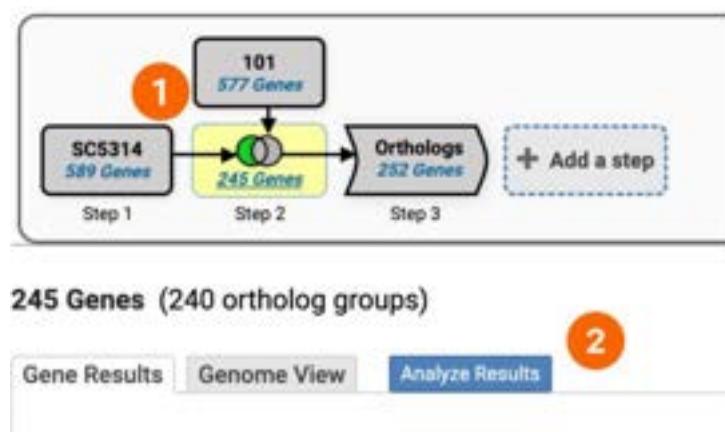
- Additional columns list the p-value, the false discovery rate (FDR), and a list of all the genes annotated, either directly or indirectly, to the term. FDR is an estimate of the percent chance that a particular GO term might actually be a false positive. It represents the fraction of the nodes with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

FungiDB: Performing GO Enrichment analysis

Learning objectives:

- Perform a GO enrichment analysis
- Create a complex search strategy using both FungiDB and SGD
- **Perform enrichment analysis on *C. albicans* SC5314 gene upregulated when the pathogen is exposed to mucosal surfaces.**
 - Use a search strategy created in the 'Transcriptomics & Proteomics' Strategy URL:
<https://fungidb.org/fungidb/app/workspace/strategies/import/802d9f2b606fc1fa>

1. Click on Step 2 to highlight upregulated genes in *C. albicans* SC5314 only.
2. Click on the “Analyze Results” tab for enrichment analysis options.



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

3. Deploy GO enrichment analysis by clicking the “Gene Ontology Enrichment” button.

Analyze your Gene results with a tool below.



The screenshot shows three analysis tools available for gene results: Gene Ontology Enrichment, Metabolic Pathway Enrichment, and Word Enrichment. The 'Gene Ontology Enrichment' button is highlighted with a blue border and a '3' icon, indicating it is the selected tool.

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

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

Other parameters limit users' analysis to either "Curated" or "Computed" annotations or both. Those with a GO evidence code inferred from electronic annotation (IEA) are denoted "Computed," while all others have some curation. The default P-value is set to 0.05 but can be adjusted manually.

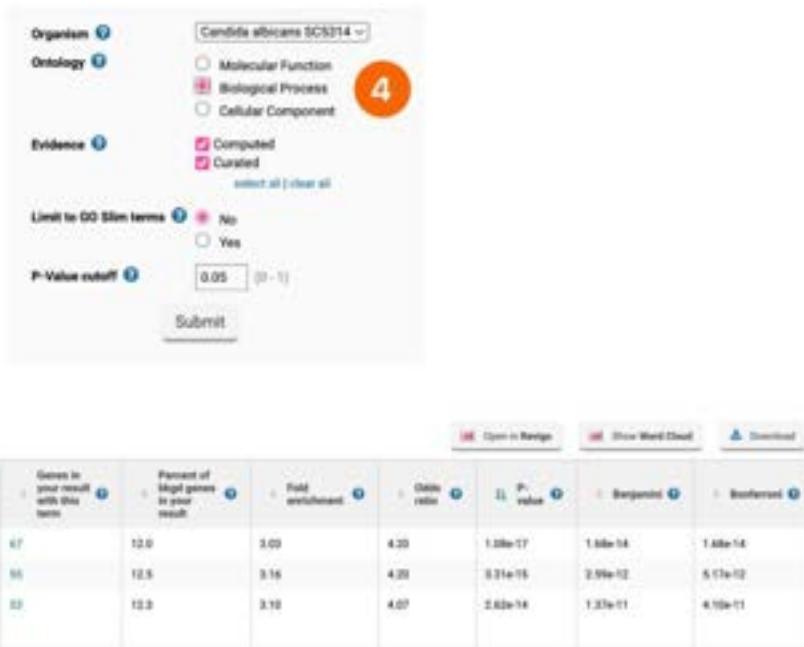


The screenshot shows the Wellcome GO enrichment analysis interface. The search parameters are as follows:

- Organism: Candida albicans SC5314
- Ontology: Molecular Function, Biological Process, Cellular Component
- Evidence: Computed, Curated
- Limit to GO Slim terms: No
- P-Value cutoff: 0.05

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

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



The screenshot shows the results table for the enrichment analysis. The table includes the following columns:

GO ID	GO Term	Genes in the list with this term	Genes in your result with this term	Percent of total genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0042273	ribosomal large subunit biogenesis	338	67	12.0	3.03	4.20	1.88e-17	1.88e-14	1.88e-14
GO:0000470	maturational of (L3U) rRNA	440	56	12.5	3.16	4.39	9.31e-15	2.59e-12	5.17e-12
GO:0000463	maturational of (L3U) rRNA from precursor rRNA transcript (5S(r)-rRNA, 5.8S rRNA, L3U rRNA)	432	52	12.3	3.18	4.07	2.62e-14	1.33e-11	4.15e-11

Analysis Results: 344 items

Open in Revigo | Show Word Cloud | Download

The results table includes several additional statistical measurements:

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

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

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

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

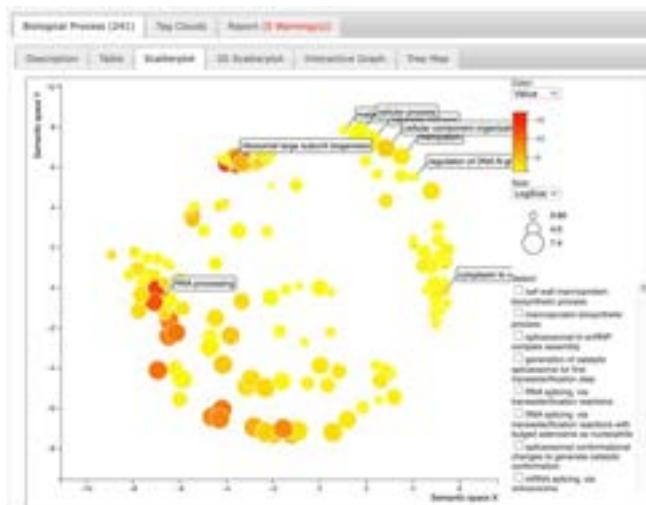
Note: you can sort genes in your results using the sort options within a column:

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

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

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

Bubble size represents the frequency of the GO term in the underlying database.



The table tab provides a detailed overview of the GO terms, P-values and also parent GO terms used to describe a group of related GO terms (<http://geneontology.org/docs/ontology-relations/>)

Optional exercise. Creating queries across FungiDB and SGD.

Use case: During a genetic screen in *Lomentospora prolificans*, you identified several exciting genes, including jhhlp_004726, a hypothetical protein. Use FungiDB and SGD records to learn more about this gene.

1. Navigate to jhhlp_004726 in FungiDB and examine available records.

https://fungidb.org/fungidb/app/record/gene/jhhlp_004726

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

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

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

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



The screenshot shows the FungiDB gene record page for jhhlp_004726. At the top, there are three buttons: 'Add to basket', 'Add to favorites', and 'Download Gene'. Below the title 'jhhlp_004726 hypothetical protein' is a large button labeled 'Download Gene: jhhlp_004726'. Underneath are sections for 'Choose a Report' (with 'Text' selected), 'Choose Attributes' (with 'Gene models' checked), 'Choose Tables' (with 'Orthologs and syntenic' checked), and 'Download Type' ('Text File' selected). An orange arrow points from the 'Download Gene' button to the 'Choose Tables' section.

The exported text file can be opened with Excel.

- Sort genes on the [Organism].
- Copy GenIDs for *S. cerevisiae* (e.g., YDR144C).
- Navigate to the SGD gene lists search to create a new upload.
- Paste *S. cerevisiae* orthologs for jhhlp_004726 in the form: <https://www.yeastgenome.org/locus/YDR144C>.



Create a new list

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

List type: Gene

Organism: *S. cerevisiae*

Identifiers are case sensitive

File Type: Free Text File uploaded

CONTINUE

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

3. Find known genetic interactions for MKC7.

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

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



Summary Sequence Protein Gene Ontology Phenotype Interactions Regulation Expression Literature Homology

MKC7 / YDR144C Interactions

Interactions Overview Genetic Interactions Physical Interactions Interaction Network Resources

Summary: The mkc7 null mutant is viable; the null mutant of paralog ypr1 is viable; the mkc7 ypr1 double mutant has concomitant heat sensitivity, increased sensitivity to caffeine, congo-red, caspofungin, calcifluor white, growth at low pH and a secretion defect; a mkc7 ypr1 ypr2 triple mutant has severe concomitant heat sensitivity and decreased tolerance to high salt.

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

Genetic

Physical

Analyze Physical Genetic Interaction All

Genetic Interactions

Genetic Interactions 13 interactions

Interactor Allele Assay Annotation Action Phenotype SGA score P-value Reference

ACT1 Synthetic Haploinsufficiency High-throughput HI Haarer B. et al. (2007) [\[PubMed\]](#)

CDC2 Synthetic Growth Defect High-throughput HI vegetative growth decreased Mutant Type unspecified Tong AH. et al. (2004) [\[PubMed\]](#)

- 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%	50 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
fungi-type cell wall organization	4 of 9 genes, 44.4%	205 of 7166 genes, 2.9%	0.00678	0.40%	0.02	YHR079C; YLR120C; YLR121C; YFL039C

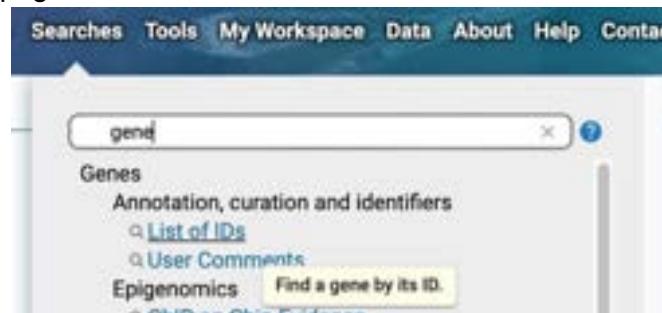
Now, let's go back to the file of MKC7 "synthetic" genetic interactors we downloaded earlier and find the orthologs of these genes in *Lomentospora prolificans*.

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

Interactor	Interactor/Sy Interactor	Interactor Systematic Name	Type	Assay	Annotation
MKC7	YDR344C	ACT1	YFL039C	Genetic	Synthetic Hz high-through
MKC7	YDR344C	GIM1	YML094W	Genetic	Synthetic Gm high-through
MKC7	YDR344C	IKE1	YHR027C	Genetic	Synthetic Gm manually cur
MKC7	YDR344C	IKX2	YNL238W	Genetic	Synthetic Let manually cur
MKC7	YDR344C	PAC10	YGR078C	Genetic	Synthetic Let high-through
MKC7	YDR344C	SBAS1	YKL117W	Genetic	Synthetic Let high-through
MKC7	YDR344C	YKE2	YLR200W	Genetic	Synthetic Gm high-through
MKC7	YDR344C	YPS1	YLR120C	Genetic	Synthetic Let manually cur
MKC7	YDR344C	YPS1	YLR120C	Genetic	Synthetic Let manually cur
MKC7	YDR344C	YPS1	YLR120C	Genetic	Synthetic Gm manually cur
MKC7	YDR344C	YPS1	YLR121C	Genetic	Synthetic Let manually cur

- Revisit FungiDB and initiate the List of IDs search query

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



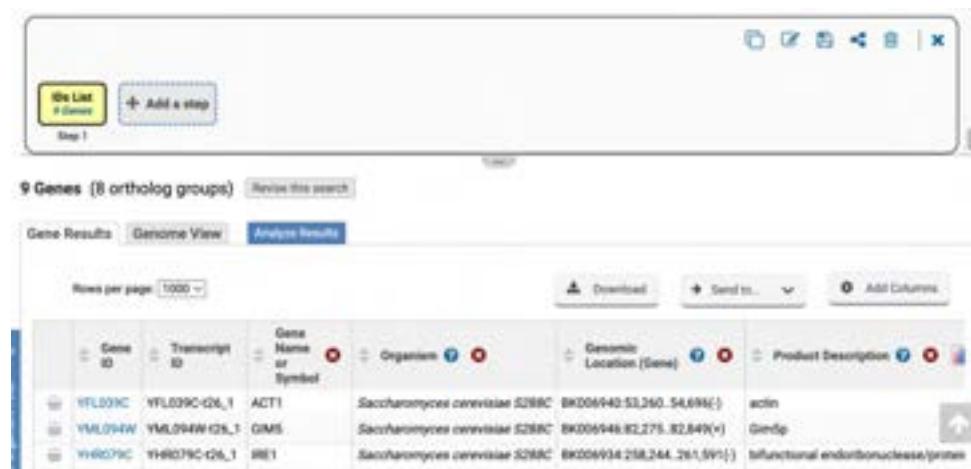
- Paste the list of Gene IDs with the “synthetic” genetic interactions with MKC7 into the FungiDB query and click the **Get Answer** button.



A screenshot of the "Identify Genes based on List of IDs" search form. At the top, there are three buttons: Configure Search, Learn More, and View Data Sets Used. Below these are several input options:

- Gene ID input set:** A text input field containing the gene IDs: YNL039W, YMR027C, YBL117W, YLR040W, YLR209C, and YLR175C. An orange arrow points to this input field.
- Upload a text file:** A file input field labeled "Choose file" with the placeholder "Select file. This file will be automatically converted to a list of IDs".
- Upload from a URL:** A text input field labeled "Enter URL" with the placeholder "The URL should point to a file or list of IDs".
- Copy from My Basket:** A checkbox with the placeholder "Checklist with the selected items copied to basket".
- Copy from My Strategy:** A checkbox with the placeholder "Checklist with the selected items copied to strategy".

At the bottom right of the form is a large orange "Get Answer" button with an orange arrow pointing to it.



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

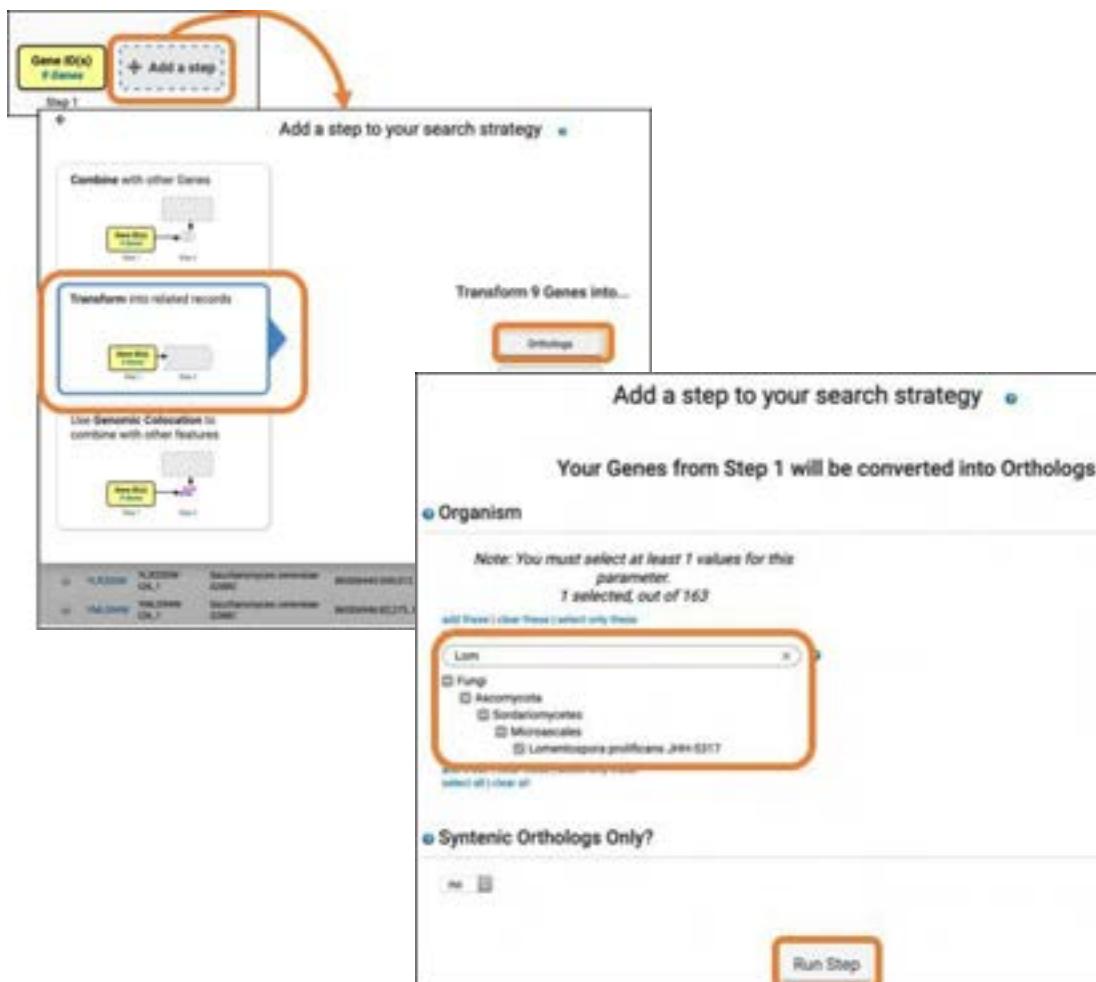
Gene Results [Genome View](#) [Analyse Results](#)

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-026_1	ACT1	Saccharomyces cerevisiae S288C	BR0008942-53,260..54,694(+)	actin
YML094W	YML094W-026_1	GMS1	Saccharomyces cerevisiae S288C	BR0008948-82,275..82,849(+)	Gms1p
YHR078C	YHR078C-026_1	IRF1	Saccharomyces cerevisiae S288C	BR0008934-218,244..219,591(+)	transcriptional endoribonuclease/protein

- Find orthologs in *Lomentospora prolificans*.

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



Gene ID(x)
9 Genes [Add a step](#)

Add a step to your search strategy

Transform 9 Genes Into... [Orthologs](#)

Add a step to your search strategy

Your Genes from Step 1 will be converted into Orthologs

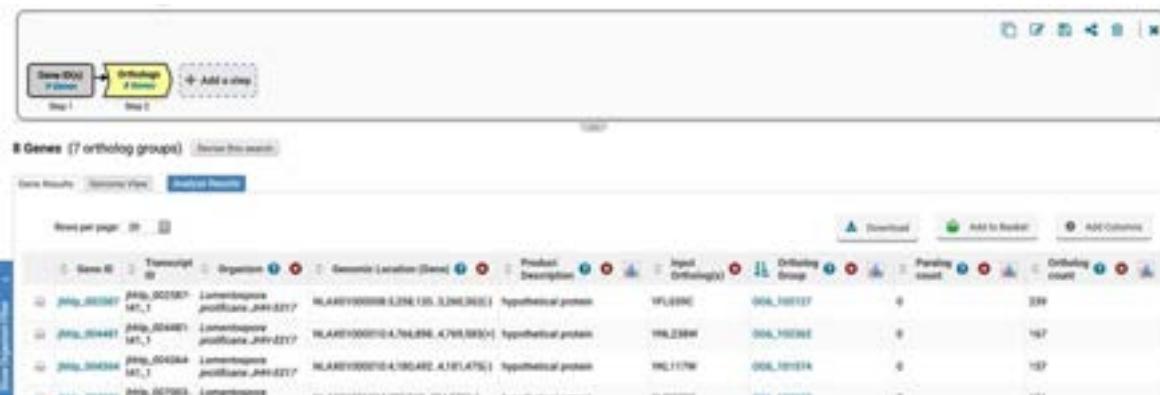
Organism

Note: You must select at least 1 values for this parameter.
7 selected, out of 163
[Add these](#) | [Clear these](#) | [Select only these](#)

Lom
 Fungi
 Ascomycota
 Saccharomycetes
 Mucorales
 Lomentospora prolificans JPH-5317

Syntenic Orthologs Only?

[Run Step](#)



The screenshot shows a search interface for orthologs. Step 1 shows a gene ID (YGL00100W) and Step 2 shows orthologs for Lomentospora prolificans. The results table has columns for Gene ID, Transcript ID, Organism, GeneLoc, Product Description, Input Ortholog, Orthology Group, Pending count, and Conflicting count. Three rows are shown:

Gene ID	Transcript ID	Organism	GeneLoc	Product Description	Input Ortholog	Orthology Group	Pending count	Conflicting count
JHhp_000347	JHhp_000347	Lomentospora prolificans	JAH-2271	YIL039C	YIL039C	YIL039C	0	339
JHhp_004471	JHhp_004471	Lomentospora prolificans	JAH-2271	YIL039C	YIL039C	YIL039C	0	167
JHhp_004474	JHhp_004474	Lomentospora prolificans	JAH-2271	YIL039C	YIL039C	YIL039C	0	157

How many interacting *S. cerevisiae* genes have a hypothetical protein ortholog in *Lomentospora prolificans*? Can you find jhhlp_004726 amongst these genes?

Strategy URL:

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

Glycosylphosphatidylinositol (GPI)-anchored proteins are involved in cell wall integrity and cell-cell interactions, and perturbations in GPI biosynthesis lead to hypersensitivity to host defences. Given the accumulated biological information we uncovered at SGD and FungiDB, summarize your predictions about the hypothetical *L. prolificans* protein jhhlp_004726.

- What is the likely jhhlp_004726 ortholog in *S. cerevisiae*?
 - Is this gene a GPI protein in yeast?
- Do you have sufficient information to think the hypothetical gene in *L. prolificans* may be a putative GPI-anchor protein?
- How many “synthetic” genetic interactors exist in SGD for MKC7 in yeast?
 - What GO terms were enriched in biological processes associated with MKC7 interactors in *S. cerevisiae*?
 - How many orthologs of these genes are found in *L. prolificans*?
 - Why do you think the number of genes varies between *S. cerevisiae* and *L. prolificans*?

Additional resources:

More info on Fischer's exact test:

<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/>

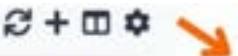
RNA sequence data analysis via Galaxy, Part 2

Learning objectives:

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

• Sharing workflow histories with others.

1. Ensure your history has a useful name (e.g., Mycelium vs Spore, RNA Group3, etc.) and click the history action menu icon.
2. Select the “Make History Accessible and Publish” option and check to ensure all objects within History are accessible.

1 History  

search datasets 

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.

2 

Also make all objects within the History accessible.

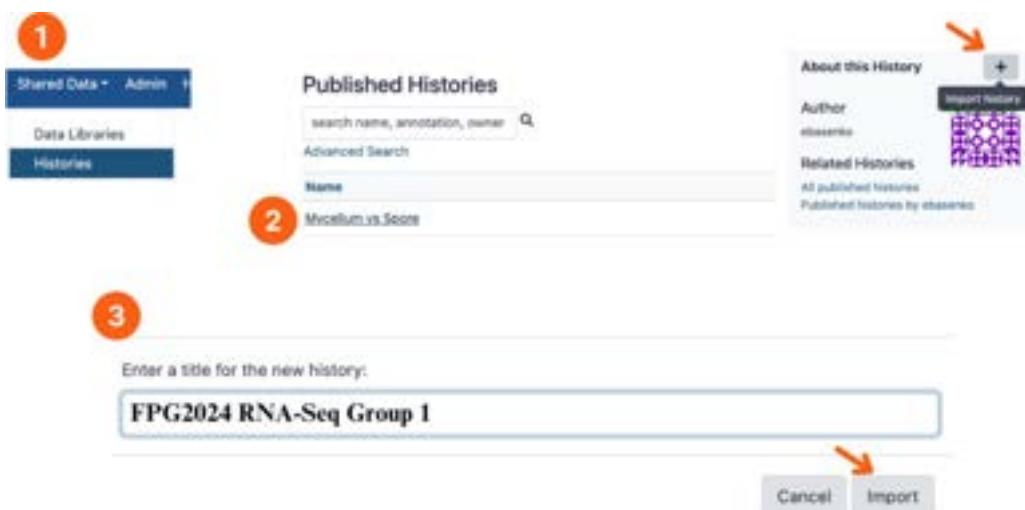
Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, so that other users can view and import it.

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 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, but not all are visible. You can explore all the hidden files by clicking on the word “hidden” (orange circle)—this will reveal all hidden files.

Many more output files are available to explore →

Differential expression data on the two collection →

Coverage data in BigWig format →

FastQC results (one per each file submitted) →

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



- Explore the FastQC results.

To do this, find the step called “FastQC on collection ##: Webpage.” Click on the name. This will open the FastQ pairs. Click on one of them, then click on the 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:
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
- Oversimplified sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	SRR1179893_2.fastq.gz
File type	Conventional base calling
Encoding	Sanger / Illumina 1.3
Total Sequences	1645791
Sequences flagged as poor quality	0
Sequence length	251
KQC	50

Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.3 encoding)

Explore the differential expression results.

We will explore two output files:

- A. **DESeq2 Plots** – you can view these directly in Galaxy by clicking 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 table contains the actual differential expression results. While these can be viewed within Galaxy, it will be more beneficial to download this table and open it in Excel so you can sort the results.**

The tabular file contains seven 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 the floppy disk save icon.

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



- **Explore the results in Excel.**
 1. Sort them based on the log2 fold change – column 3.

2. Pick a list of upregulated gene IDs from column 3 with a good corrected P value (column 7) and load them into FungiDB using the “List of IDs” search.

A	B	C	D	E	F	G
D8826_0012	5492.946886	4.14837844	0.25276937	19.4973844	1.16E-84	7.25E-82
D8826_0041	3459.15095	4.12535507	0.23288538	19.3773525	1.20E-83	6.95E-81
D8826_0047	549.884174	4.11535522	0.34755985	11.8407366	2.40E-32	1.99E-30
D8826_0029	12514.2357	4.09248482	0.17888878	22.8775688	7.77E-156	7.88E-113
D8826_0065	297.307163	4.03803435	0.2780354	14.50955963	1.05E-47	1.64E-45
D8826_0031	5682.61669	4.03468031	0.22941832	17.5865813	3.12E-69	1.33E-66
D8826_0065	242.253832	4.01587422	0.29254924	13.726489	7.05E-43	9.53E-41
D8826_0024	1129.38482	3.97988586	0.26221324	15.1780507	4.94E-52	1.00E-49
D8826_0079	401.277324	3.9573969	0.27562766	14.3599407	9.23E-47	1.39E-44
D8826_003118						

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:

Upload from a URL:

Copy From My Basket:

Copy From My Strategy:

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

Exporting data to VEuPathDB/FungiDB

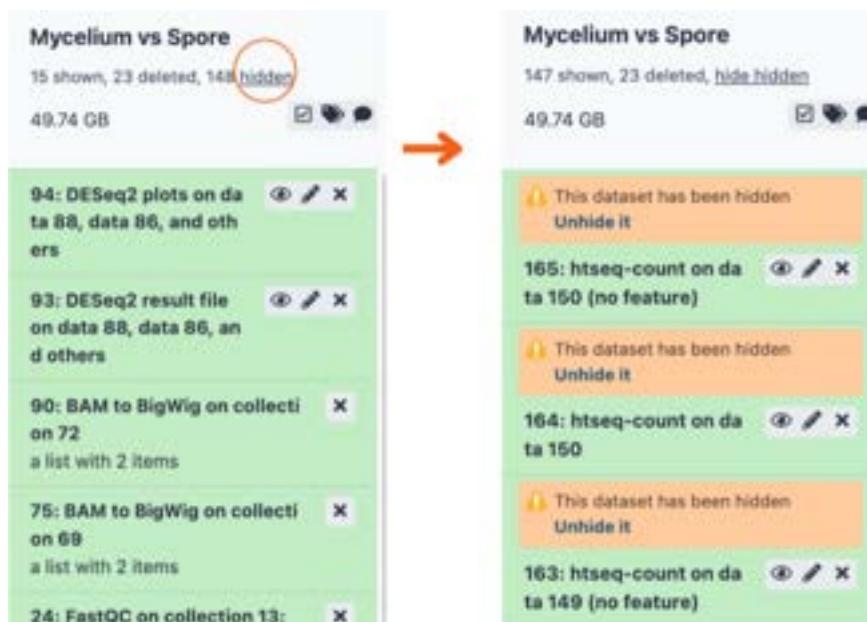
The VEuPathDB RNA-Seq export tool provides a mechanism to export your RNASeq results (TPM values) and BigWig RNASeq coverage files. The advantage of doing this is that it allows you to search the TPM data using the RNASeq search in VEuPathDB and view the BigWig files in the genome browser.

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.**

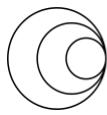
Click on the link at the top of your history that says “## hidden”. This will show all hidden files.



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 word “the-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, you should have selected four 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 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... ▾

- This dataset has been hidden Hide datasets
- This dataset has been unhidden Unhide datasets
- 165: htseq-count on data 150 Delete datasets
- 164: htseq-count on data 150 Undelete datasets
- This dataset has been hidden Permanently delete datasets
- 163: htseq-count on data 149 Build Dataset List
- 162: htseq-count on data 148 Build Dataset Pair
- This dataset has been hidden Build List of Dataset Pairs
- This dataset has been hidden Build Collection from Rules



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

Note: the htseq-count files will be in the same order as the raw files loaded into the history. For more info, use the “Guide to FPG2023 RNA-Seq histories and file organization” in Part 1.

Create a collection from a list of datasets

Collections of datasets are permanent, ordered lists of datasets that can be passed to tools and workflows in order to have analyses done on each member of the entire group. This interface allows you to crea... More help

Start over

htseq-count on data 74

htseq-count on data 73

htseq-count on data 71

htseq-count on data 70

Cancel

htseq-count on data 74

Click to rename

htseq-count on data 71

htseq-count on data 70

veupathdbprod.globusgenomics.org says

Enter a new name for the element:

mycelium 2

Cancel OK

Create a collection from a list of datasets

Collections of datasets are permanent, ordered lists of datasets that can be passed to tools and workflows in order to have analyses done on each member of the entire group. This interface allows you to crea... More help

Start over

mycelium 2

mycelium 1

score 2

score 1

Hide original elements?

Name: Mycelium vs score

Create list

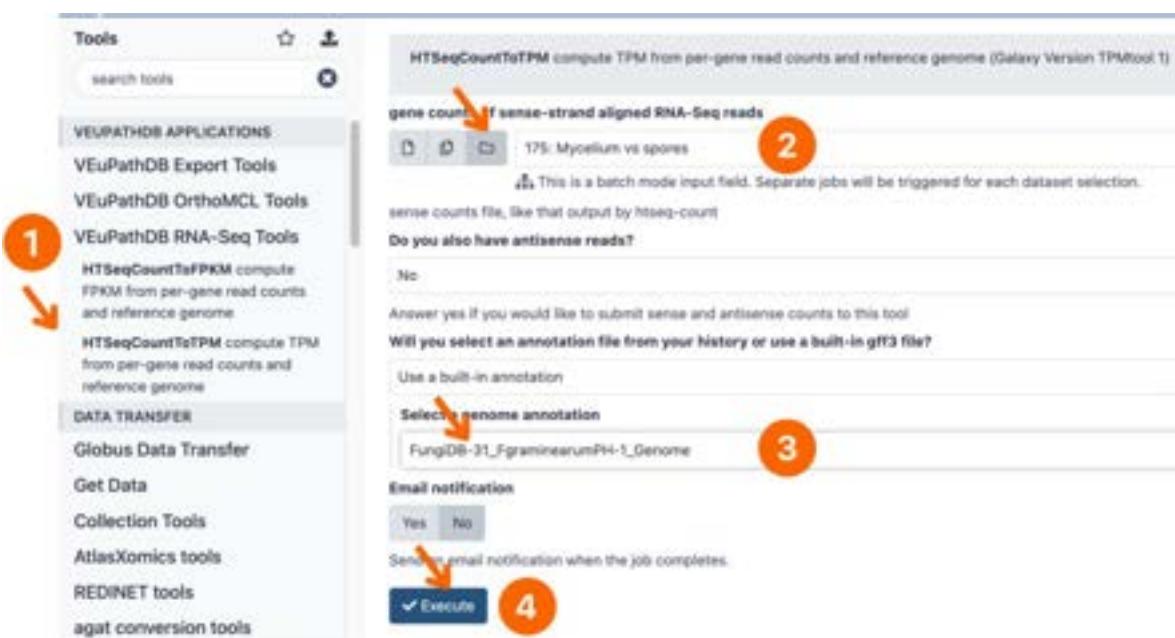
• Create a Dataset List with “BAM to BigWig on data” files.

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.**

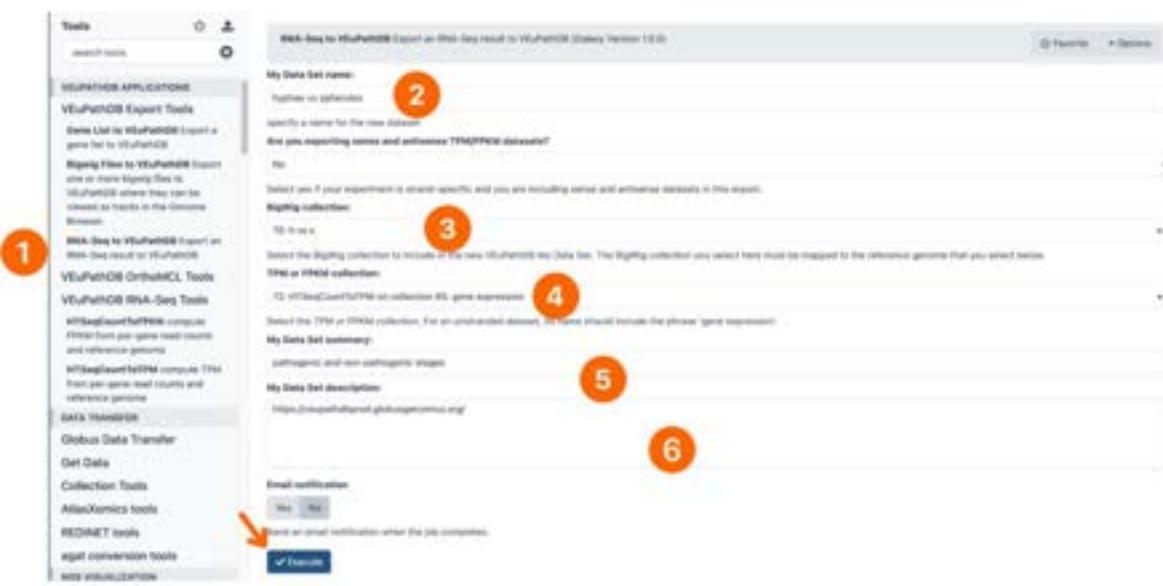
1. Select the HTSeqCountToTPM tool (under the VEupathDB RNAseq tools in the left menu).
2. Make sure the list of count files is selected.
3. Select the reference organism.
4. Click on the “Execute” button.



The screenshot shows the HTSeqCountToTPM tool interface in Galaxy. On the left, there's a sidebar with a 'Tools' section and a 'VEUPATHDB APPLICATIONS' section containing several tools. A red circle with the number '1' has an arrow pointing to the 'HTSeqCountToTPM' tool in the applications list. The main panel shows the 'HTSeqCountToTPM' tool configuration. It has a header: 'HTSeqCountToTPM compute TPM from per-gene read counts and reference genome (Galaxy Version TPMtool 1)'. Below the header, there's a 'gene counts' input field with three dropdown menus, one of which is highlighted with a red circle and the number '2'. A note below says 'This is a batch mode input field. Separate jobs will be triggered for each dataset selection.' There's also a question 'Do you also have antisense reads?' with 'No' selected. A note below asks 'Will you select an annotation file from your history or use a built-in gff3 file?' followed by 'Use a built-in annotation'. Under 'Select a genome annotation', a dropdown menu shows 'FungiDB-31_FgraminearumPH-1_Genome', highlighted with a red circle and the number '3'. At the bottom, there's an 'Email notification' section with 'Yes' selected and a note 'Send an email notification when the job completes.', followed by a 'Execute' button highlighted with a red circle and the number '4'.

- 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.



1

2

3

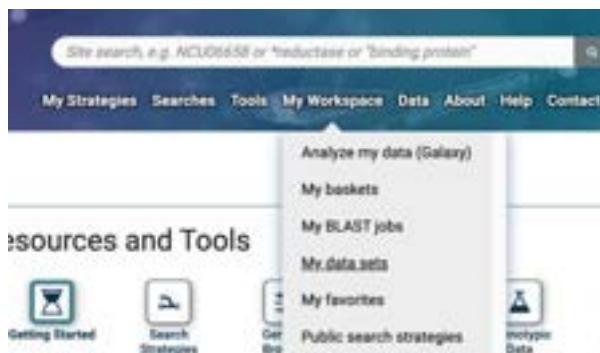
4

5

6

- 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

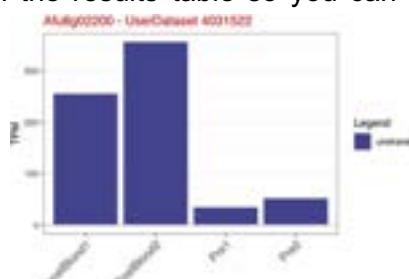
Data set size: 271.05 M

Quota usage: 2.84% of 10.00 G

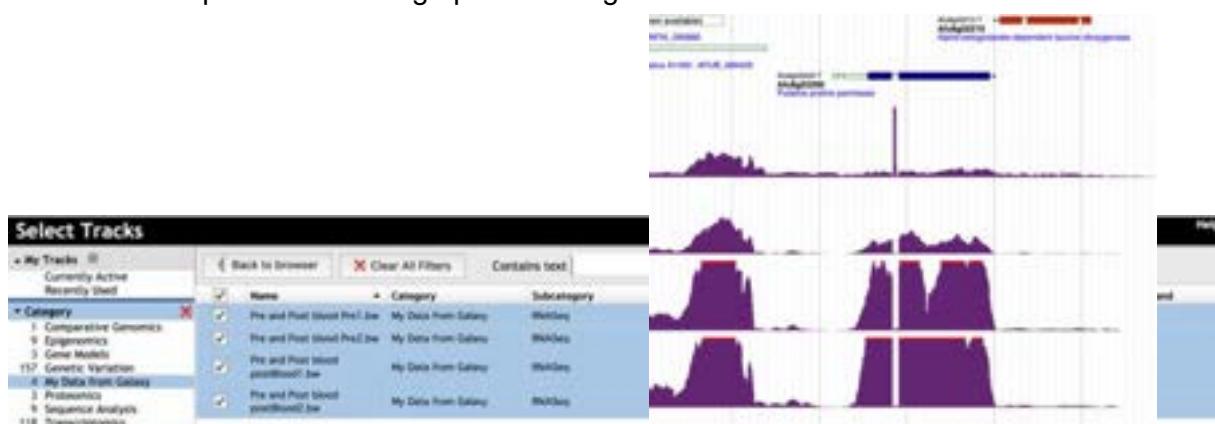
Available searches: • RNA-Seq user dataset (fold change) ↗



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.



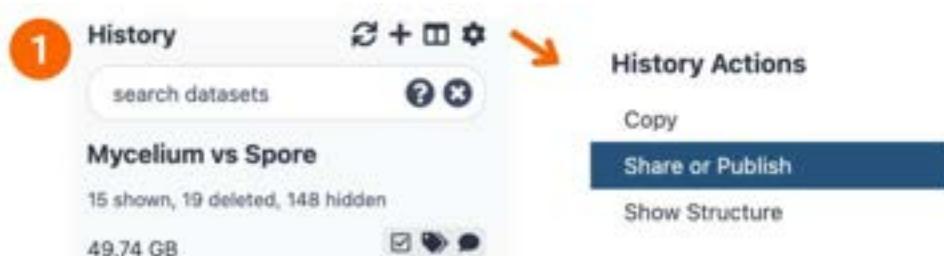
Variant Calling analysis, Part 2: Analyzing results (Group Exercise)

Learning objectives:

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

• Share workflow histories with others.

1. Make sure your history has a useful name (e.g., Group3 SNPs, etc.) and click on the history action menu icon.
2. Select the “Make History Accessible and Publish” option and check to ensure all objects within History are accessible.



The screenshot shows the Galaxy History interface. A red circle labeled '1' highlights the 'History' tab. An orange arrow points from the 'Actions' icon (three vertical dots) to a dropdown menu labeled 'History Actions'. The 'Share or Publish' option in the menu is highlighted with a red rectangle. Below the history list, there are buttons for 'Copy' 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, ↗

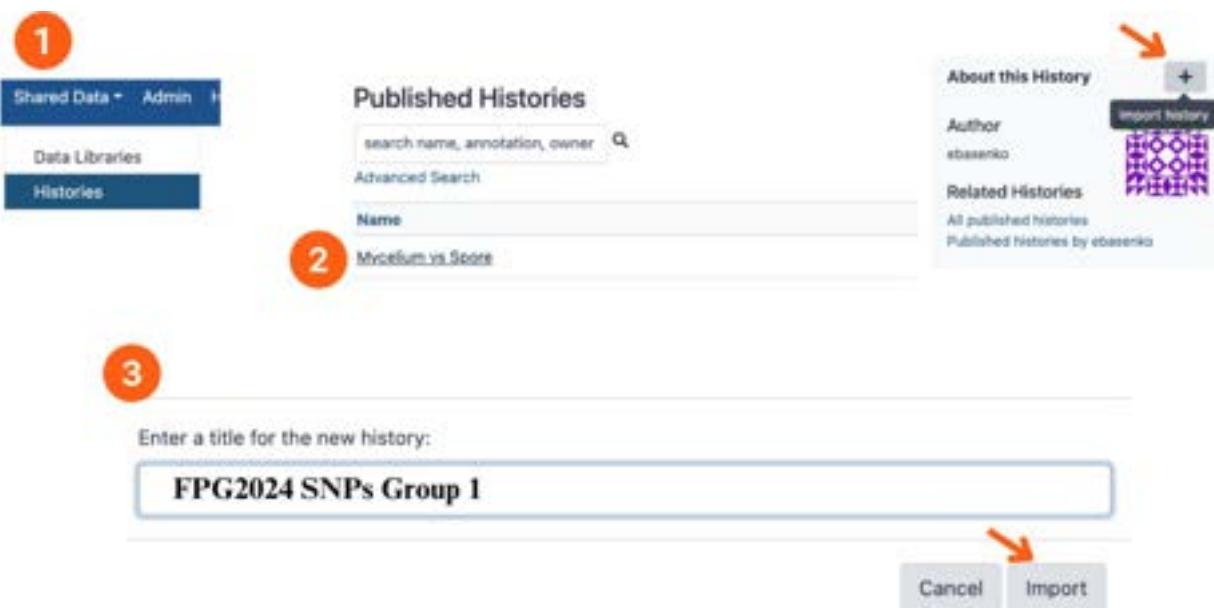
Share History with Individual Users

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 or leave it as is.



Published Histories

search name, annotation, owner Advanced Search

Name: Mycelium vs Spore

About this History
Author ebaserenko
Related Histories All published histories Published Histories by ebaserenko

Import history

Enter a title for the new history:
FPG2024 SNPs Group 1

Cancel Import

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 them are visible. You can explore all the hidden files by clicking on the word “hidden” (orange circle)—this will reveal all hidden files.

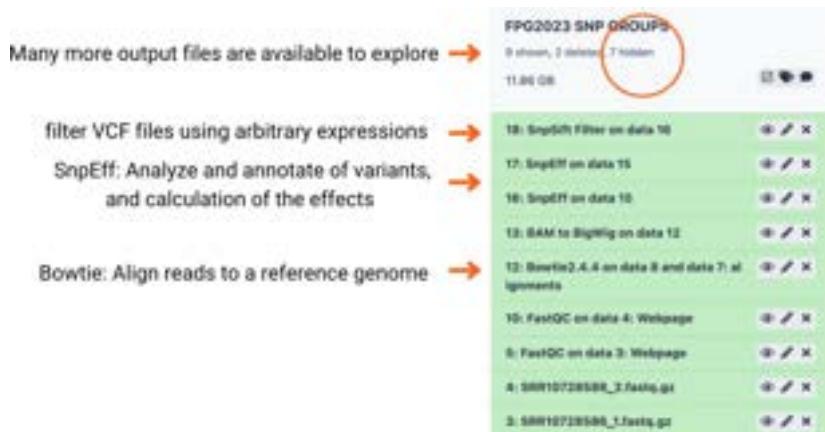
The Variant calling workflow has three major components: (1) mapping of raw reads to the reference genome, (2) calling variants, and (3) annotating variants. This workflow can be used to call single nucleotide polymorphisms, insertions and deletions (also defined as indels), and multiple nucleotide polymorphisms.

We used Bowtie2 to align and map sequences to a reference genome in this workflow. 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 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 we are using is called Sort and belongs to the SAMtools suite. The sorted file is an input for downstream FreeBayes variants.

Analysis and annotation of the genomic variants are carried out by the SnpEff tool. SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes). It uses reference genomes to annotate genomic variants based on their location and 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 impact of the amino acid



Many more output files are available to explore →

FPG2023 SNP GROUPS
9 drivers, 2 indels, 7 hidden
11.86 GB

- 18: SnpEff Filter on data 10
- 17: SnpEff on data 15
- 16: SnpEff on data 10
- 13: BAM to BigWig on data 12
- 12: Bowtie2-0.4.4 on data 8 and data 7 alignments
- 10: FastQC on data 4: Webpage
- 9: FastQC on data 3: Webpage
- 8: SRR1072385SR_2.fastq.gz
- 9: SRR1072385SR_1.fastq.gz

change. They are classified into synonymous and non-synonymous, gain or loss of start codons, gain or loss of stop codons, and frameshifts.

The SnpSift tool annotates, filters, and manipulates genomic annotated variants. Once you have 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.

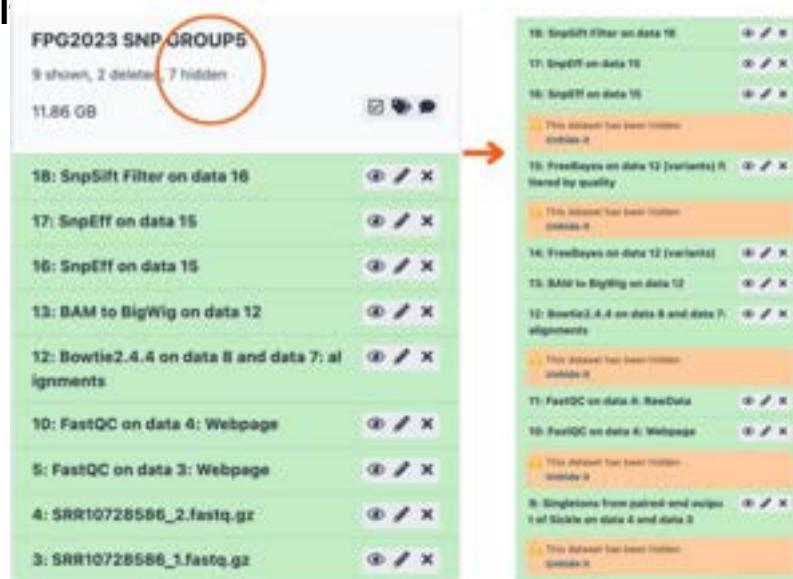
Bowtie generates a file called a BAM file. You will likely hear of file formats called SAM or

BAM when dealing with sequence alignment files. SAM stands for Sequence Alignment/Map format, and BAM is the binary version of a SAM file.

Many 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 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

9: FastQC on data 3: Webpage

8: SRR10728586_2.fastq.gz

3: SRR10728586_1.fastq.gz

18: SnpSift Filter on data 16
This dataset has been filtered: remove it

16: SnpEff on data 15
This dataset has been filtered: remove it

13: BAM to BigWig on data 12 (variants) R
Filtered by quality

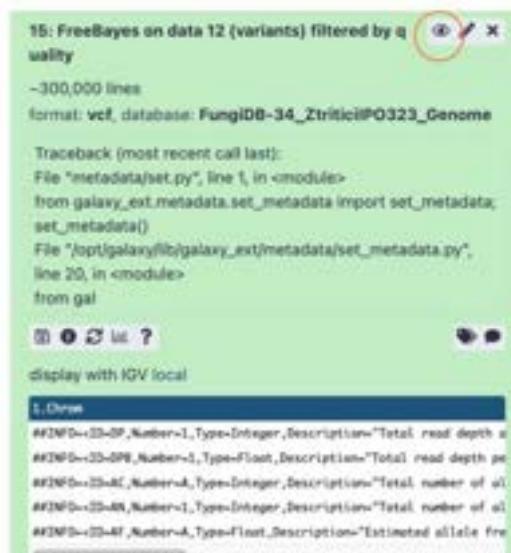
12: Bowtie2.4.4 on data 8 and data 7: alignments
This dataset has been filtered: remove it

10: FastQC on data 4: RawData
This dataset has been filtered: remove it

9: FastQC on data 3: Webpage
This dataset has been filtered: remove it

8: SRR10728586_2.fastq.gz
This dataset has been filtered: remove it

3: SRR10728586_1.fastq.gz
This dataset has been filtered: remove it



18: FreeBayes on data 12 (variants) filtered by quality

-300,000 lines:
format: vcf, database: FungiDB-34_ZtriticumPO323_Genome

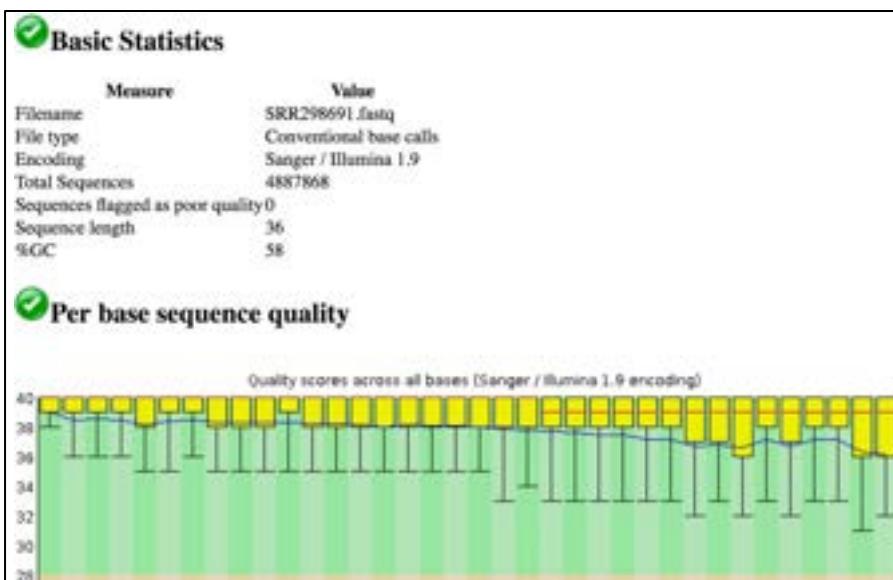
Traceback (most recent call last):
File "metadata/set.py", line 1, in <module>
from galaxy_ext.metadata.set_metadata import set_metadata;
set_metadata()
File "/opt/galaxy/lib/galaxy_ext/metadata/set_metadata.py",
line 20, in <module>
from gal

display with IGV local

```
#INFO<--ID=DP,Number=1,Type=Integer,Description="Total read depth"
#INFO<--ID=DPW,Number=1,Type=Float,Description="Total read depth per
#INFO<--ID=AC,Number=1,Type=Integer,Description="Total number of alt
#INFO<--ID=AN,Number=1,Type=Integer,Description="Total number of alt
#INFO<--ID=AF,Number=1,Type=Float,Description="Estimated allele fre
```

- Examine sequence quality based on FastQC quality scores.

FastQC provides an easy-to-navigate visual representation of 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 with the HTML format.

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



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

The Summary includes warnings about possible genomic annotation errors or inconsistencies identified in the reference genome. If there are many, use caution when 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 lines (input file) - number of lines in the vcf file
- Number of non-variants: 0—some packages report non-variant observations for nt positions between the reference genome and the vcf file generated.
- 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

Summary	
Genome	FungiDB-34_Zm稟iPO323_Genome
Date	2023-04-11 19:24
SnpEff version	snpEff 4.11 (build 2015-10-03), by Pablo Cingolani
Command line arguments	java -jar /opt/snpEff/snpEff.jar -v vcf -o vcf -stats /scratch/galaxy/tiles/000/291/dataset_391424.vcf FungiDB-34_Zm稟iPO323_Genome /scratch/galaxy/tiles/000/381/dataset_391422.fasta
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	1,280,619
Genome total length	39,730,198
Genome effective length	39,730,198
Variant rate	1 variant every 129 bases

Variants rate details	
Chromosome	Length
Zm1_MiscC scaffold	43,947
Zm chr_1	6,088,777
Zm chr_10	1,682,575
Zm chr_11	1,924,292
Zm chr_12	1,462,624
Zm chr_13	1,185,774
Zm chr_14	773,098
Zm chr_15	639,501
Zm chr_16	807,044
	Variants
	2,441
	137
	111
	110
	114
	110
	374
	81
	118
	Variants rate

mice and human projects) any recognised variants will be listed here

- 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 variants by type

Type	Total
SNP	154,254
MNP	12,864
INS	6,907
DEL	7,304
MIXED	2,180
INTERVAL	0
Total	143,289

Type	What it 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** commonly refers to frameshift or new stop codon detections, as those changes will profoundly affect 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:

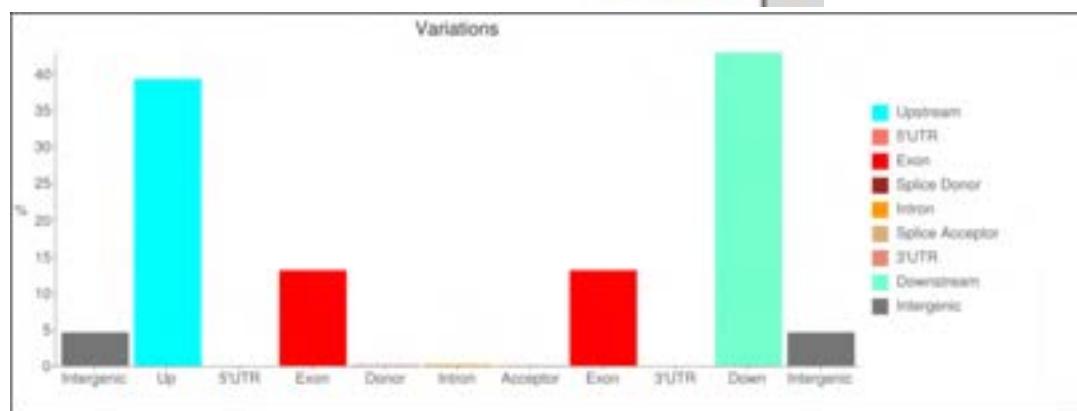
Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	1,857	0.145%
LOW	87,874	6.861%
MODERATE	41,970	3.277%
MODIFIER	1,148,118	89.777%

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%



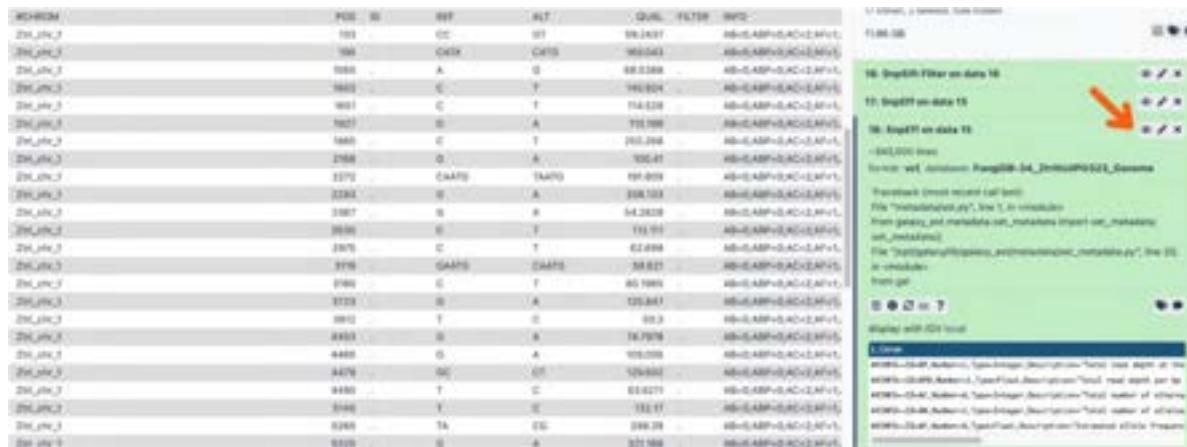
Additionally, you may see several SNPs being reported in several classes: missense variant + splice region variant. This means that some SNPs found within certain splice sites also contain a missense variant. SNPs in the splice sequences may affect intron splicing and lead to readthrough.

- The 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 usually represented by a bar graph where count = number of SNPs and X axis is quality score (a higher score means better p-values and high confidence in 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 specific alleles (a high ratio suggests that genes may be under selective pressure).
- Allele frequency statistics report the frequency of alleles and help identify potential sequencing artefacts due to the PCR enrichment step (generation of heterozygous counts in a haploid organism).

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

Examining SNP information.

You can view the SNP information by clicking the “eye” icon within the SnpEff vcf file.



The screenshot shows the Artemis genome browser interface with a VCF file loaded. On the left, the genomic sequence is displayed with various SNPs highlighted. On the right, there are several panels providing detailed information about the SNPs. One panel, titled "SnpSift Filter on data 16", lists 11 SNPs with their effects and filters applied. Another panel shows a histogram of allele frequencies. The bottom right corner displays the command used to generate the VCF file, which includes SnpSift filtering.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
Zm1_snp_1	101		C	T	29.2437	AB	ABP>0, AC>0, NV>0
Zm1_snp_2	100		CATG	CATG	90.0443	AB	ABP>0, AC>0, NV>0
Zm1_snp_3	1000		A	G	89.0388	AB	ABP>0, AC>0, NV>0
Zm1_snp_4	1001		T	TG	191.8294	AB	ABP>0, AC>0, NV>0
Zm1_snp_5	1001		C	T	114.1238	AB	ABP>0, AC>0, NV>0
Zm1_snp_6	1002		A	A	105.1088	AB	ABP>0, AC>0, NV>0
Zm1_snp_7	1000		T	TG	201.2648	AB	ABP>0, AC>0, NV>0
Zm1_snp_8	2168		G	A	105.81	AB	ABP>0, AC>0, NV>0
Zm1_snp_9	2170		CATTG	CATTG	191.009	AB	ABP>0, AC>0, NV>0
Zm1_snp_10	2284		G	A	318.103	AB	ABP>0, AC>0, NV>0
Zm1_snp_11	2307		G	A	14.2818	AB	ABP>0, AC>0, NV>0
Zm1_snp_12	3806		T	TG	111.01	AB	ABP>0, AC>0, NV>0
Zm1_snp_13	3870		C	T	62.0998	AB	ABP>0, AC>0, NV>0
Zm1_snp_14	3919		CATTG	CATTG	308.821	AB	ABP>0, AC>0, NV>0
Zm1_snp_15	3980		C	T	85.7989	AB	ABP>0, AC>0, NV>0
Zm1_snp_16	3988		G	A	125.844	AB	ABP>0, AC>0, NV>0
Zm1_snp_17	4012		T	C	99.3	AB	ABP>0, AC>0, NV>0
Zm1_snp_18	4053		G	A	14.7978	AB	ABP>0, AC>0, NV>0
Zm1_snp_19	4481		G	A	105.009	AB	ABP>0, AC>0, NV>0
Zm1_snp_20	4479		AC	CT	124.000	AB	ABP>0, AC>0, NV>0
Zm1_snp_21	4480		T	C	131.021	AB	ABP>0, AC>0, NV>0
Zm1_snp_22	5146		T	C	131.17	AB	ABP>0, AC>0, NV>0
Zm1_snp_23	5281		TG	TC	388.39	AB	ABP>0, AC>0, NV>0
Zm1_snp_24	5329		G	A	321.998	AB	ABP>0, AC>0, NV>0

The vcf file generated by SnpEff contains information about SNPs and their 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;AF=0;GT;DP;RO;O;1/1;143;0;0;143;5341;207.887;-43.0473;0		
CM001231	483825		G	A	64.8756	.	AB=0;AF=0;GT;DP;RO;O;1/1;4;0;0;4;146;-10.0999;-1.20412;0		
CM001231	518226		G	C	51.7908	.	AB=0;AF=0;GT;DP;RO;O;1/1;8;0;0;7.227;-11.5007;2.10721;0		
CM001231	574021		C	G	237.265	.	AB=0;AF=0;GT;DP;RO;O;1/1;17;0;0;17.583;-39.079;5.11751;0		
CM001231	609879		GAA	CAG	55.2785	.	AB=0;AF=0;GT;DP;RO;O;1/1;32;8;277.22;463;-18.1711;-0.694735;0		
CM001231	1090073		G	T	79.4156	.	AB=0;AF=0;GT;DP;RO;O;1/1;8;2;75.6;238;-11.5539;-1.36362;0		
CM001231	1090104		A	T	70.961	.	AB=0;AF=0;GT;DP;RO;O;1/1;8;0;0;6;235;-12.5346;-1.80618;0		
CM001231	1153611		CCTC	CTTG	111.123	.	AB=0;AF=0;GT;DP;RO;O;1/1;8;5;188.97;-9.30616;-6.1461;0		
CM001231	1159150		CT	GC	126.126	.	AB=0;AF=0;GT;DP;RO;O;1/1;31;0;0;19.741;-29.7713;-5.71957;0		
CM001231	1159438		C	G	82.3312	.	AB=0;AF=0;GT;DP;RO;O;1/1;47.30;1992;17.640;0;-9.53002;-3.56795		
CM001231	1159465		G	C	249.656	.	AB=0;AF=0;GT;DP;RO;O;1/1;126.47;1770;79.3651;-53.8644;-25.2134;0		
CM001231	1159499		T	C	124.95	.	AB=0;AF=0;GT;DP;RO;O;1/1;143;3;1567;111;4248;-76.1575;-33.4865;0		
CM001231	1181576		CC	TG	191.675	.	AB=0;AF=0;GT;DP;RO;O;1/1;27;0;0;25.924;-41.7448;-7.52575;0		
CM001231	1293309		C	G	51.22	.	AB=0;AF=0;GT;DP;RO;O;1/1;2;0;0;2.78;-6.92763;-0.60206;0		
CM001231	1323058		TT	GC	71.3001	.	AB=0;AF=0;GT;DP;RO;O;1/1;6;0;0;6;223;-12.5485;-1.80618;0		
CM001231	1485397		A	G	3558.42	.	AB=0;AF=0;GT;DP;RO;O;1/1;499;0;0;497;18671;-804.678;-149.612;0		
CM001231	1485429		G	A	3783.33	.	AB=0;AF=0;GT;DP;RO;O;1/1;517;1;38.516;2660;0;-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 helpful 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 the filtered VCF file (SnpSift output) and convert it into an Excel document.

For this exercise, two groups will share data SnpSift outputs: group 1 and 2, group 3 and 4, and group 5 and 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, meaning you can extract them programmatically or using a program like Excel.



The screenshot shows a Galaxy tool interface with a green header bar. The bar contains the title '29_AlumigatusAt293_Genome.vcf', a 'View data' button, and other tool-related icons. Below the header, there's a text area with a large amount of VCF file content. At the bottom of the interface, there's a command line section with the following text:

```
History
1: SnpSift Filter on data[1] View data
10,309 lines, 64 comments
format vcf database: /tmp/4f8_4
29_AlumigatusAt293_Genome
Command to execute: java -Xms4g -Xmx8g
java -jar /mnt/galaxyTools/tools/snpSift/snpSift.jar
filter Filter -F
scratch/galaxy/files/09ff/dataset_9634
-e
scratch/galaxy/job_working_directory/
display with IGV local
```

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

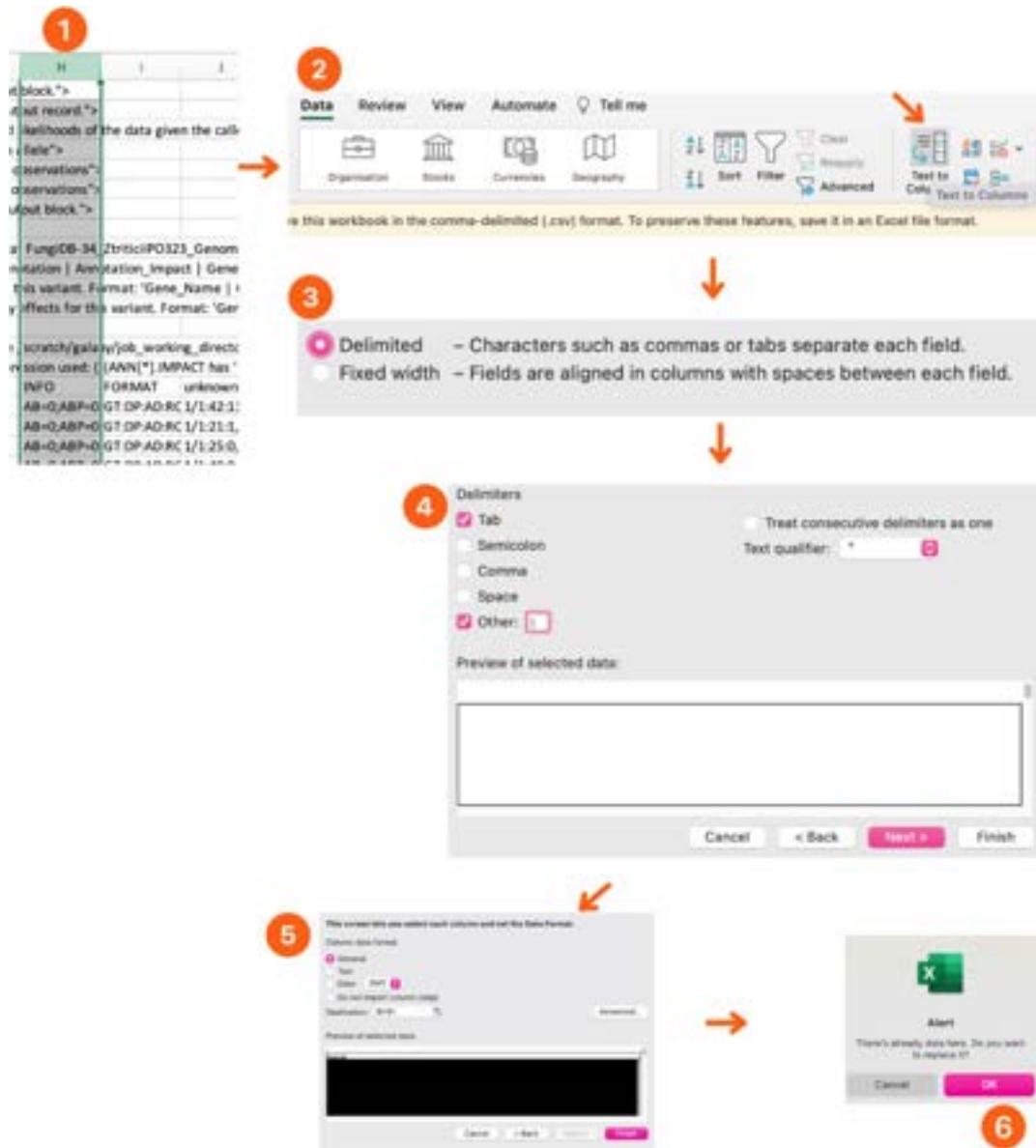
The screenshot shows an Excel spreadsheet with the following details:

- Title:** Galaxy29-SnpSift_Filter_on_data_21.xls
- Columns:** The columns are labeled as follows: ID, Feature_id, Feature_type, Feature_name, Transcript_id, Rank, Ref, Alt, Effect, AA_change, Distance, and SCORE, WARNING, ERROR.
- Data:** The data consists of approximately 100 rows of genomic variants. Each row contains information such as the variant ID, transcript ID, rank, reference allele, alternate allele, effect type (e.g., missense, splice), amino acid change, distance from start, and scores for errors and warnings.



- Manipulate Excel file to display SNP info in columns.

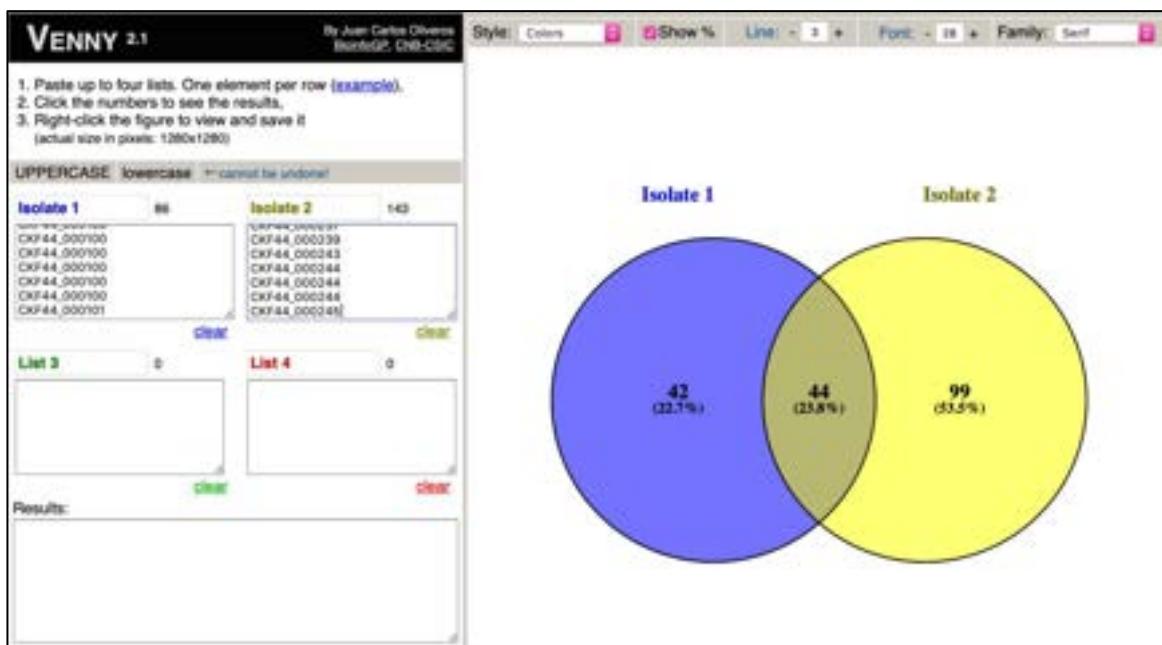
- Select the “INFO” column.
- Navigate to the “Data” tab in Excel and choose “Text to Columns”.
- Use the “Delimited” option.
- Set delimiters to the “Tab” and “|” in the “Other” and click “Next”
- Leave other criteria at default and click on the “Finish” button.
- Click “OK” on the Alert pop-up.



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 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 shows a comparison between lists of GenIDs. 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 an SNP in the WT and an SNP in the mutant that will be at 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:genelD

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 GenelD



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	=CONCATENATE(B34,"/",C34,"/",D34)																								
36	#INFC	Cut	Ctrl X																						
37	#INFC	Copy	Ctrl C																						
38	#INFC	Paste	Ctrl V																						
39	#INFC	Paste Special	>																						
40	#INFC																								
41	#FOR	Insert																							
42																									

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

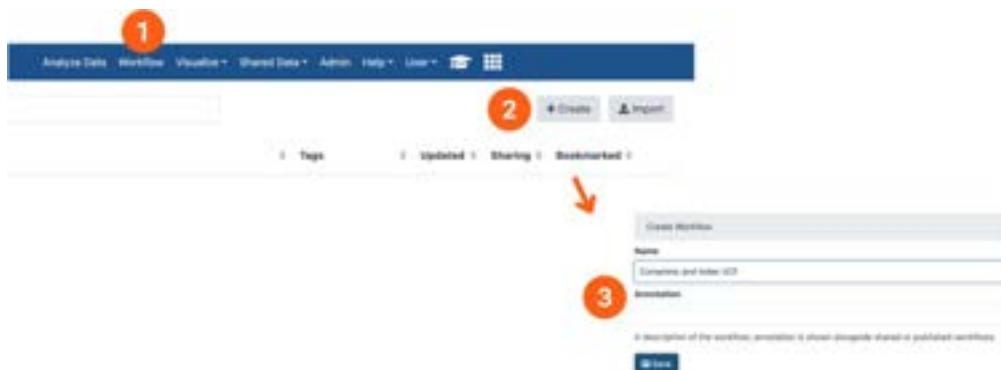
4. Copy these newly generated unique IDs into List 1 and List2 on Venny <http://bioinfogp.cnb.csic.es/tools/venny/> and examine the data.

Viewing the VCF file results in the JBrowse genome browser.

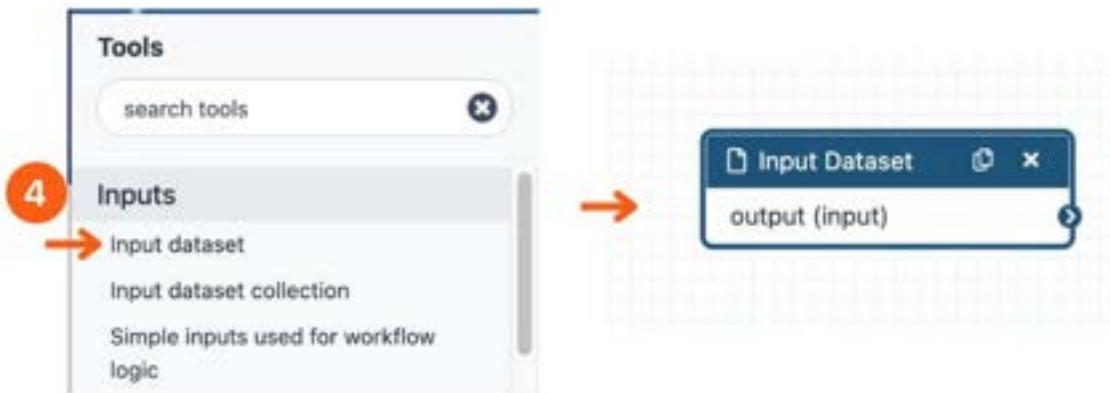
- **Create a workflow to generate compressed VCF and index files to view your data in JBrowse.**

To view a VCF file in JBrowse, it must first 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. Name the workflow (e.g., Compress and Index VCF) and click 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.

Tools

bgzip

Inputs

Data Managers

NGS: VCF Tools

tabix: Generic indexer for TAB-delimited genome position files.

bgzip Block
compression/decompression utility.
Required for use of tabix.

5

Compress and index VCF

Input Dataset
output (input)

bgzip
Input file
output (vcf_bgzip)



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)

Inputs

Data Managers

NGS: VCF Tools

tabix: Generic indexer for TAB-delimited genome position files.

bgzip Block
compression/decompression utility.
Required for use of tabix.

6

tabix
Input file
output (vcf_bgzip)

tabix
Input file
output (vcf)

Add an annotation or notes to this step. Annotations are available when a workflow is viewed.

Input file

Data input: 'input' (gtf, bed, sam, vcf, tabular, bgz or vcf_bgzip)
Regions (separate with whitespace)

Input Extension

tabular

gtf

bed

sam

vcf



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.

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

- Click on the “Run Workflow” button.

The screenshot shows the Galaxy web interface. On the left, there's a workflow titled "Workflow: Compress and index VCF". It contains a step labeled "16: SnpSift Filter on data 14". A red circle labeled "2" encircles the "Run Workflow" button. In the top right, there's a "Run" button with a red arrow labeled "1" pointing to it. On the far right, there's a "History" tab with a red circle labeled "3" around it. The history list shows three completed steps: "16: SnpSift Filter on data 14", "15: SnpEff on data 13", and "14: SnpEff on data 13". Each step has a green checkmark and a delete icon.

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

This screenshot shows the Galaxy history panel. It lists two completed workflow steps: "16: bgzip on data 14" and "17: tabix on data 19". Both files are located in the "PfPG2023 SNP Group1 data" dataset. The "bgzip" file is 1.2 MB and the "tabix" file is 14.2 KB. Each entry has a green checkmark and a delete icon.

- Download compressed vcf (vcf_bgzip) and index (tabix) files and view them in JBrowse.

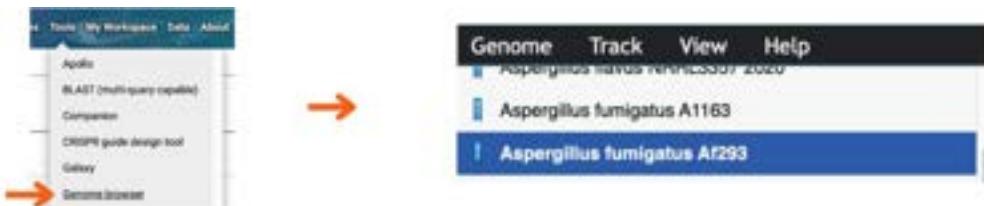
- Download both files by clicking on the download icon. You will need both files.

This screenshot shows the Galaxy history panel again. It lists the same two completed workflow steps: "16: bgzip on data 14" and "17: tabix on data 19". The "bgzip" file is highlighted with a red circle. Both files are located in the "PfPG2023 SNP Group1 data" dataset. The "bgzip" file is 1.2 MB and the "tabix" file is 14.2 KB. Each entry has a green checkmark and a delete icon.

- 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 and select "Open track file or URL".



5. Drag and drop your files into the window that appears. The file formats are autodetected. Then click on the “Open” button at the bottom of the pop-up.



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



SGD Variant Viewer

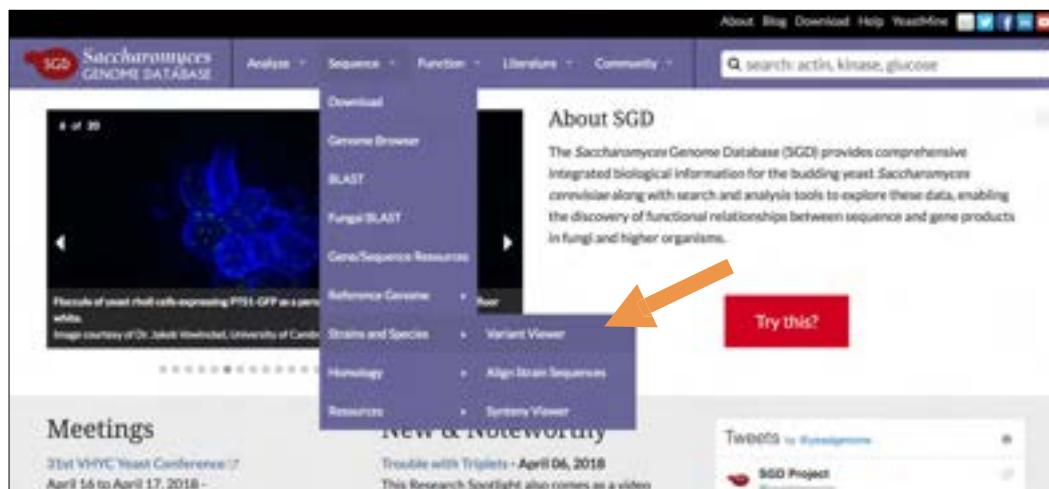
SGD's Variant Viewer (<https://yeastgenome.org/variant-viewer>) is an open-source web application that compares nucleotide and amino acid sequence differences between 12 common *S. cerevisiae* laboratory strains. For a given open reading frame, Variant Viewer breaks down the position and nature of any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

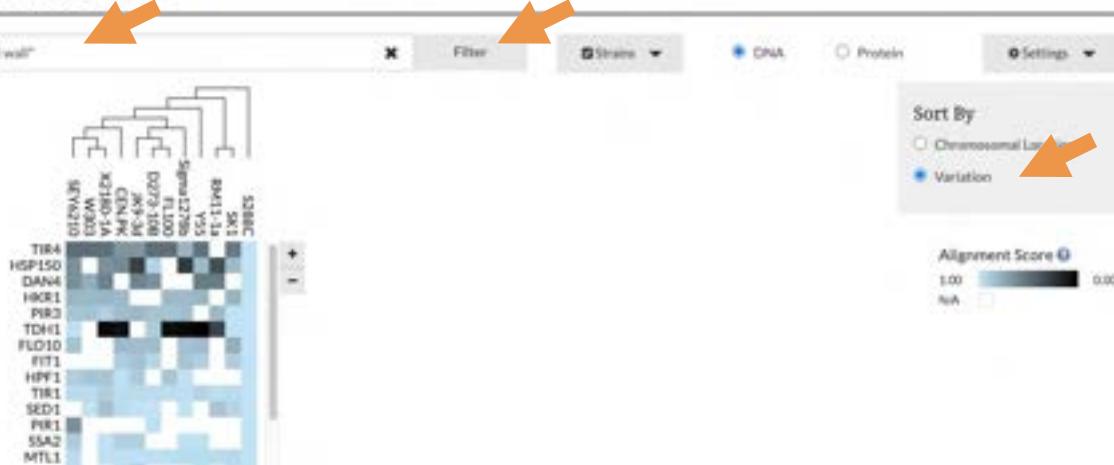
- Open the SGD home page (www.yeastgenome.org), open the Sequence tab on top of the page, then select Strains and Species followed by Variant Viewer from the pull-down menus. Or just type in the URL: yeastgenome.org/variant-viewer



- 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.

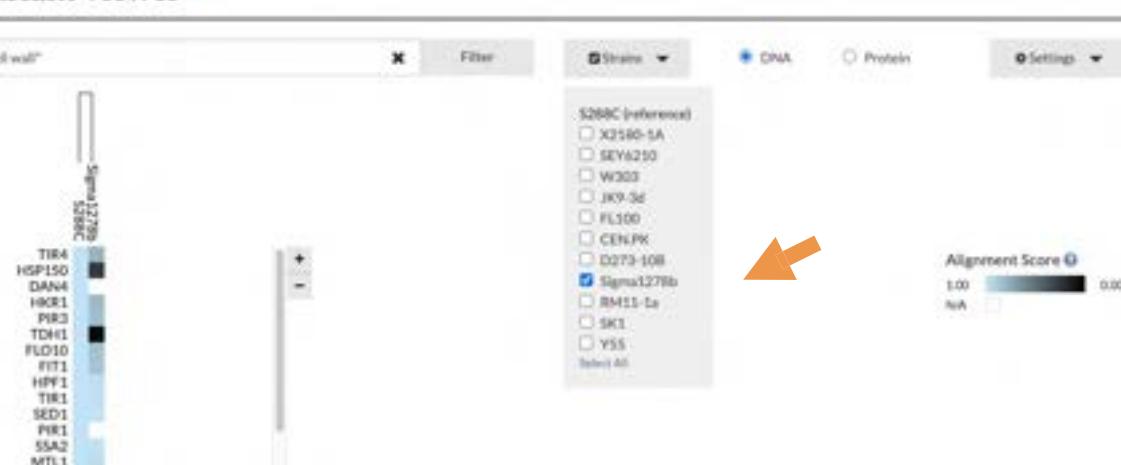


Variant Viewer ®

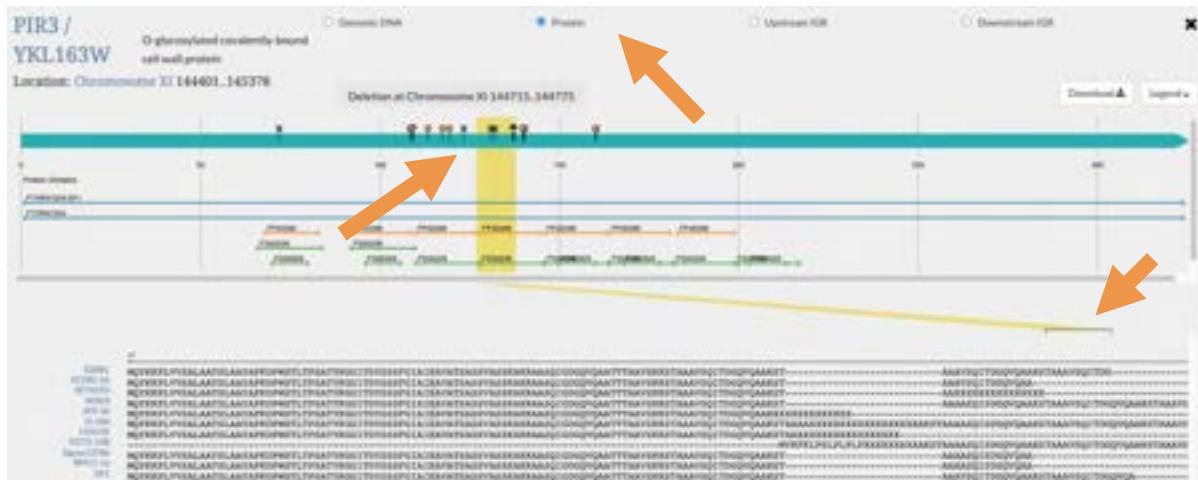


- The **matrix**, shown on the left, will have changed to only include the genes that localize to cell walls.
 - This matrix enables you to visualize high-level differences in multiple genes relative to strain S288C. Each square in the matrix corresponds to one of the twelve strains in Variant Viewer, 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 in the upper left of the page to go to the locus summary page. From the PIR3 Locus Summary page, you can see in the Description that this protein is known to vary between strains.
- In the list of references below, you'll find papers referring to the role of this cell wall protein (and its relations) in heat shock, response to toxins, and cell wall integrity. The differences in this protein between strains might contribute to variations in behavior, such as differences in pseudohyphal growth for Sigma1278b relative to S288C

References

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: 8322111
[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: 9194585
[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: 11369124
[SGD Paper](#) [DOI full text](#) [PubMed](#)
4. Porter SE, et al. (2002) The yeast paf1-rRNA polymerase II complex is required for full expression of a subset of cell cycle-regulated genes. *Eukaryot Cell* 1(5):830-42 PMID: 12465700
[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: 10394627
[SGD Paper](#) [DOI full text](#) [PubMed](#)

Variant Viewer: Sequence Tab

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

FungiDB: SNPs and Population Genetics

Learning Objective:

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

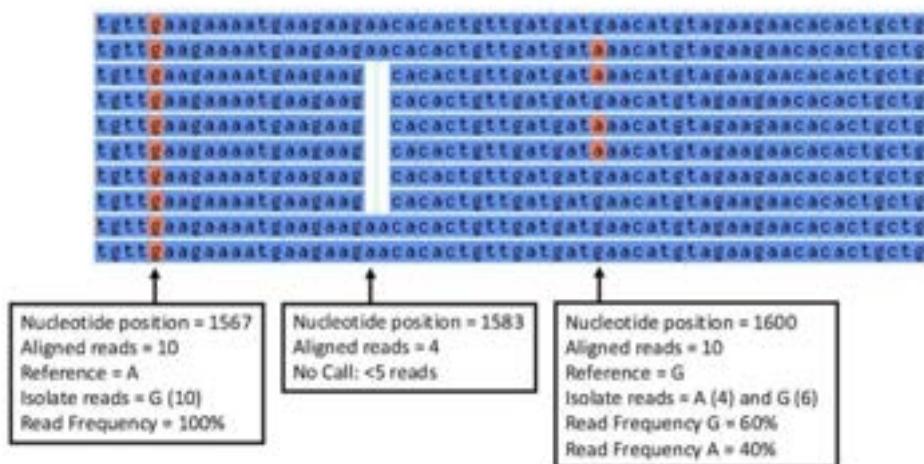
Single nucleotide polymorphisms (SNPs) are genetic variations that may or may not have an impact on the function of a gene. Most SNPs do not affect gene function. However, some SNPs that lead to a change in the amino acid or a premature stop codon (nonsense) can directly affect protein function. SNPs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

Read Frequency Threshold:

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

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

Isolate X aligned sequencing reads





Minor allele frequency:

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

Isolate consensus sequences aligned to reference genome.

reference	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
303.1	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
309.1	TGATGATCT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
RV_3600	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
RV_3606	TGATGATCT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
RV_3610	TGATGATCT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
SenT119.09	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
SenT123.09	TGATGATCT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
SenT140.08	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
SenT142.09	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
SenT175.08	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA 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, an SNP can be returned by the search based on a subset of your isolate group. The 'Percent isolates with a base call' parameter defines the fraction of isolates that must have a base call before an SNP is returned for that nucleotide position. The default setting for this parameter is 80%, or 8 out of 10 isolates in your group must have a base call for an SNP to be returned by the search. The higher this parameter, the more likely the SNP is to be high quality as regions difficult to align or difficult to sequence will tend to have a lower percentage of calls since the coverage and/or quality will be lower in that region.

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

Pyricularia oryzae (also known as *Magnaporthe oryzae*) is a pathogen that affects rice crops, causing a severe disease known as rice blast. During the infection process,

P. oryzae and other plant pathogens make use of various types of effectors to manipulate the plant's immune system. Nuclear effectors are a type of effector that contain both a secretion signal and a DNA-binding domain. In the upcoming exercise, we will be analyzing a collection of *P. oryzae* isolates obtained from rice plants in different parts of Africa. Our goal is to identify genes with at least one non-synonymous SNP, which also exhibit characteristics 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 organism tree.
3. In the Data Set section, select the datasets where isolates were collected in Zambia and other African fields.
4. Set the “SNP Class” parameter to “Non-Synonymous”.
5. Choose to identify genes with at least 1 non-synonymous SNPs and click on the “Get Answer” button.

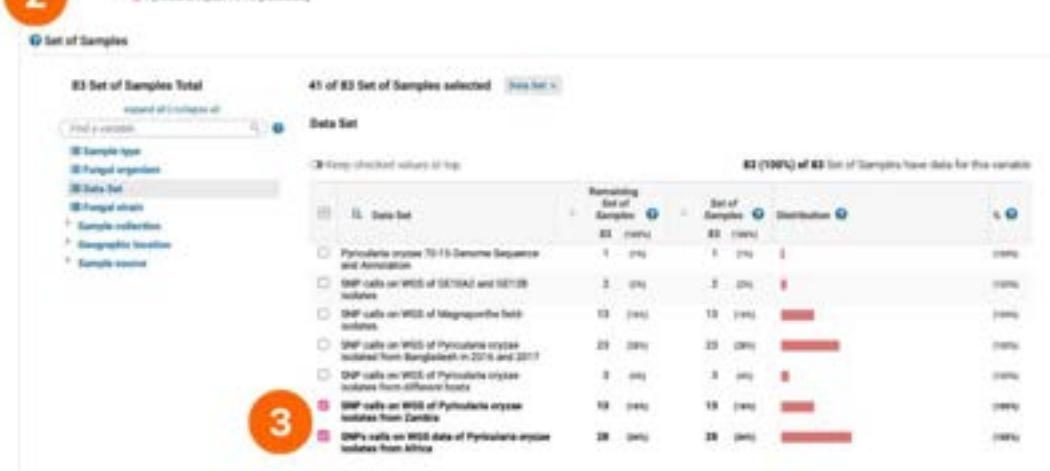
1



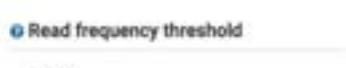
2



3



4



5



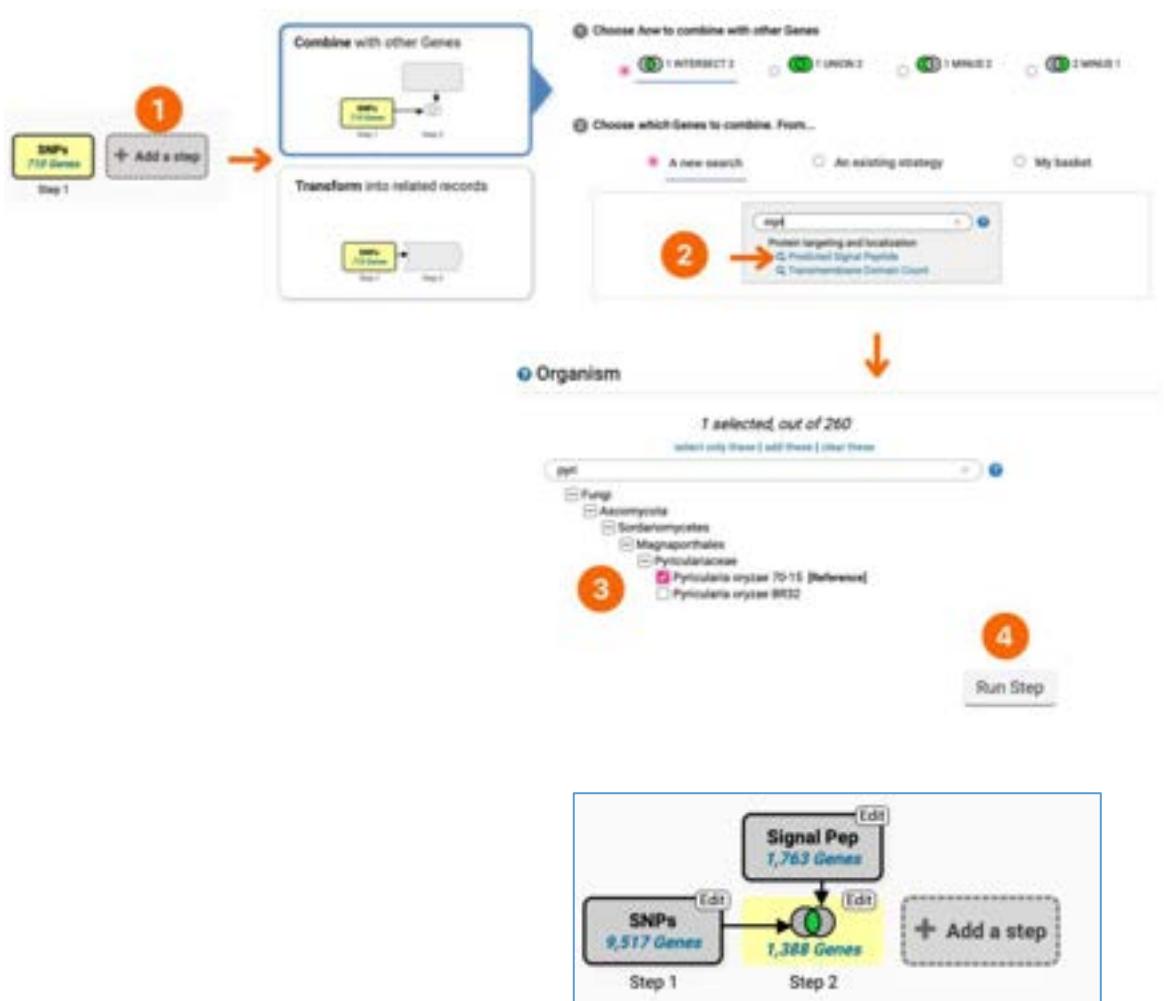
SNPs
9,517 Genes

+ Add a step

Step 1

- 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.

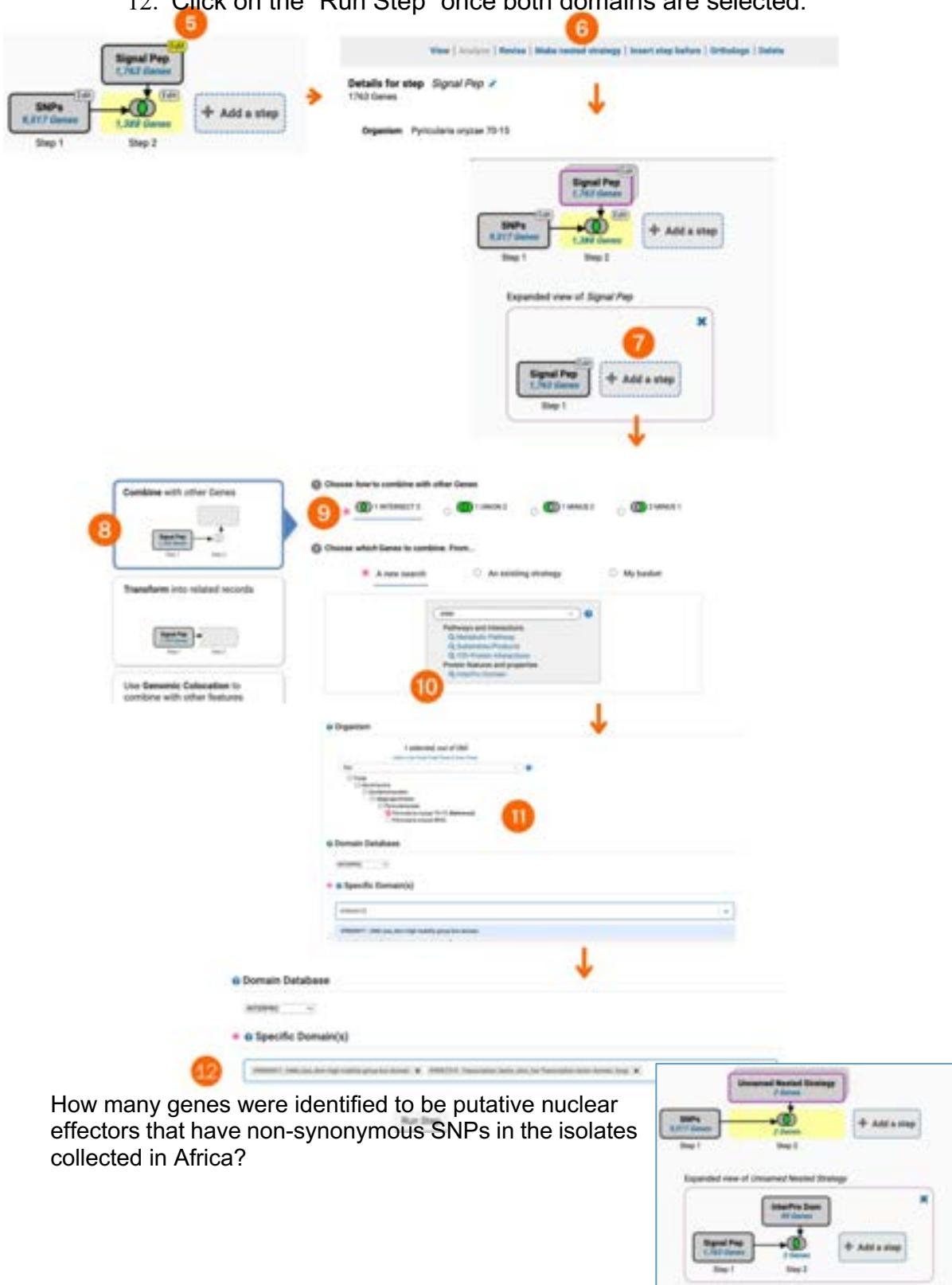


Note that currently, our strategy returns genes with at least 1 SNP and a predicted signal peptide domain. How can we identify genes with at least 1 SNP and a predicted signal peptide domain AND a DNA-binding domain? (Hint: you can do this with a nested strategy as described below).

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

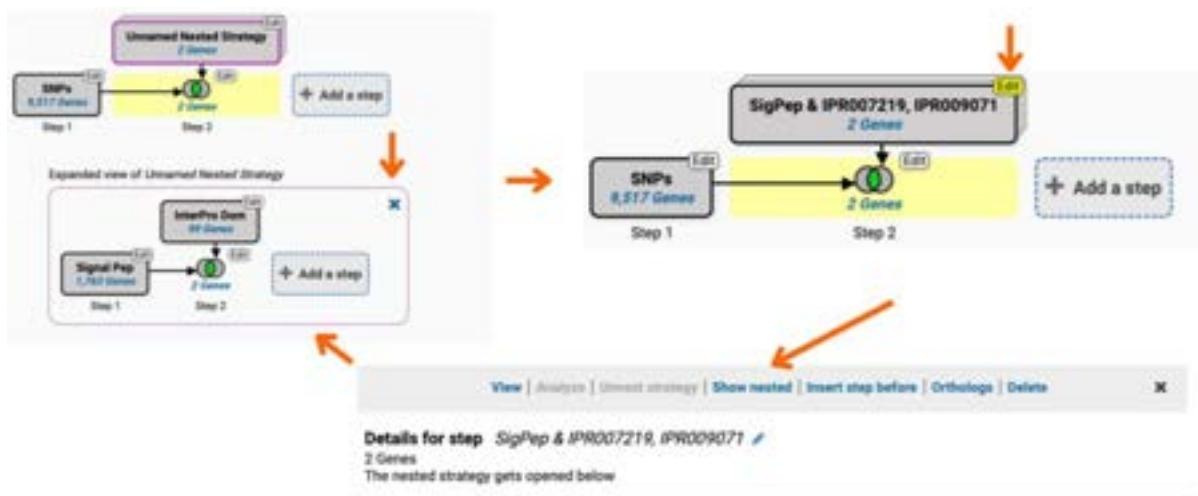


11. Set the genome to *Pyricularia oryzae* 70-50 and set the “Domain database” to InterPro and enter and select the following DNA binding domains from the dropdown menu: IPR007219, IPR009071.
12. Click on the “Run Step” once both domains are selected.





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 a disease caused by two closely related species of fungi – *Coccidioides immitis* (*C. immitis*) and *Coccidioides posadasii* (*C. posadasii*). The disease is associated with high morbidity and mortality rates that affect tens of thousands of people every year. These two fungal species are found in several regions in the Western Hemisphere, but recent studies suggest that their geographic range is expanding. The following example describes the identification of SNPs (single nucleotide polymorphisms) in *C. posadasii* str. Silveira isolates that were collected from different geographic locations.

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

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



Identify SNPs based on Differences Between Two Groups of Isolates

Configure Search Learn More View Data Sets Used

Reset values to default

Organism

1 selected

sly sly

Ascomycota
Eurotiomycetes
Oryzales
Coccidioides
Coccidioides posadasii str. Silveira [Reference]



3

Data Set	Remaining Set A isolates	Set A isolates	Distribution
Coccolomyces present in Shire-Gomes Sequence and Annotation	79 (100%)	1 (1%)	1 (1%)
SNP calls on WGS of Coccolomyces present across from regions bordering the Caribbean Sea	79 (100%)	10 (13%)	10 (13%)
SNP calls on WGS of Coccolomyces isolates from the Western Hemisphere	79 (100%)	47 (60%)	47 (60%)

4

Data Set	Remaining Set B isolates	Set B isolates	Distribution
Coccolomyces present in Shire-Gomes Sequence and Annotation	79 (100%)	1 (1%)	1 (1%)
SNP calls on WGS of Coccolomyces present across from regions bordering the Caribbean Sea	79 (100%)	10 (13%)	10 (13%)
SNP calls on WGS of Coccolomyces isolates from the Western Hemisphere	79 (100%)	47 (60%)	47 (60%)

5

Get Answer

Two Groups
19,147 SNPs

+ Add a step

Step 1

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

1

2

Details for step Two Groups ↗
4059 SNPs

3

View | **Revise** | Insert step before | Delete
Modify the configuration of this search.

↓

⑦ Set A major allele frequency \geq

2 90

↓

⑦ Set B major allele frequency \geq

3 90

↓

Two Groups
4,059 SNPs

+ Add a step

Step 1

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_SNPGL636536.6075	GL636536.6:075		10-10	C	100		G	90	
NGS_SNPGL636536.6114	GL636536.6:114		10-10	G	100		C	100	
NGS_SNPGL636536.5945	GL636536.5:945		10-10	C	100		T	95.7	
NGS_SNPGL636536.5454	GL636536.5:454		10-10	A	100		G	100	
NGS_SNPGL636536.4354	GL636536.4:354		10-10	A	100		G	100	
NGS_SNPGL636536.1402	GL636536.1:402		10-10	A	100		G	95.3	
NGS_SNPGL636536.8746	GL636536.8:746		10-10	A	100		G	100	
NGS_SNPGL636536.6075	GL636536.6:075	CPGS_10217	15	T	100		C	92.3	
NGS_SNPGL636536.5332	GL636536.5:332		10-10	T	100		A	100	
NGS_SNPGL636536.4479	GL636536.4:479		10-10	T	100		G	92.3	
NGS_SNPGL636536.1587	GL636536.1:587	CPGS_10216	739	T	100		G	95.3	
NGS_SNPGL636536.1541	GL636536.1:541	CPGS_10216	752	G	100		A	95.8	
NGS_SNPGL636536.1338	GL636536.1:338	CPGS_10220	295	A	100		G	90	
NGS_SNPGL636536.1220	GL636536.1:220		10-10	G	100		A	91.6	
NGS_SNPGL636536.1120	GL636536.1:120		10-10	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 GenoID, 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: CPGS_10217

Gene Strand: reverse

Major Allele: C (0.58)

Minor Allele: T (0.42)

Distinct Allele Count: 2

Reference Allele: C

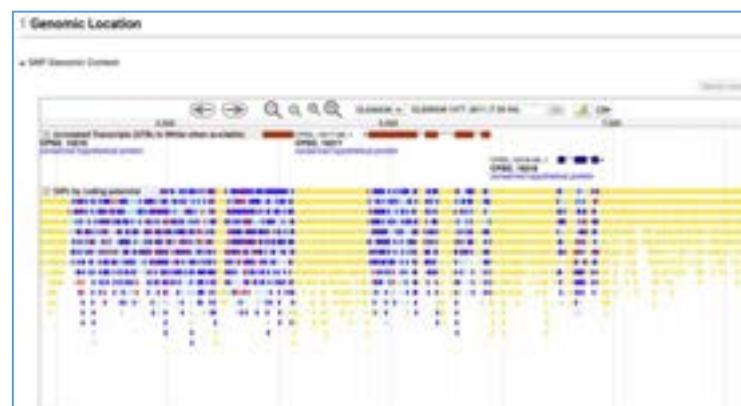
Reference Product: G 15

Allele (gene strand): G

SNP context: TCTGAGACTTATTCTGGTTGCTTCCTTC**C**CTTCCCTGTCCTTCCAGTTGTTGAATGAAT

SNP context (gene strand): ATTCAATTCAACA**T**GGAAAGGACAGGGAAAG**G**GAAGAGAAGCAACCAGAATAAAGTCTCAGA

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 coloured 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 species metadata in the output

Select strains:

78 Reference Samples Total 63 of 78 Reference Samples selected Country

Filter a column

	EU - Country	Remaining Reference Samples	Reference Samples	Distribution	%
<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	3 (3%)	3 (3%)	3 (3%)	100%
<input type="checkbox"/>	Mexico	10 (10%)	10 (10%)	10 (10%)	100%
<input type="checkbox"/>	Paraguay	1 (1%)	1 (1%)	1 (1%)	100%
<input checked="" type="checkbox"/>	United States of America	62 (80%)	62 (80%)	62 (80%)	100%
<input type="checkbox"/>	Unknown	1 (1%)	1 (1%)	1 (1%)	100%

27 (9%) of 78 Reference Samples have data for this variable

[View Results](#)

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

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

Search this table...

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

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

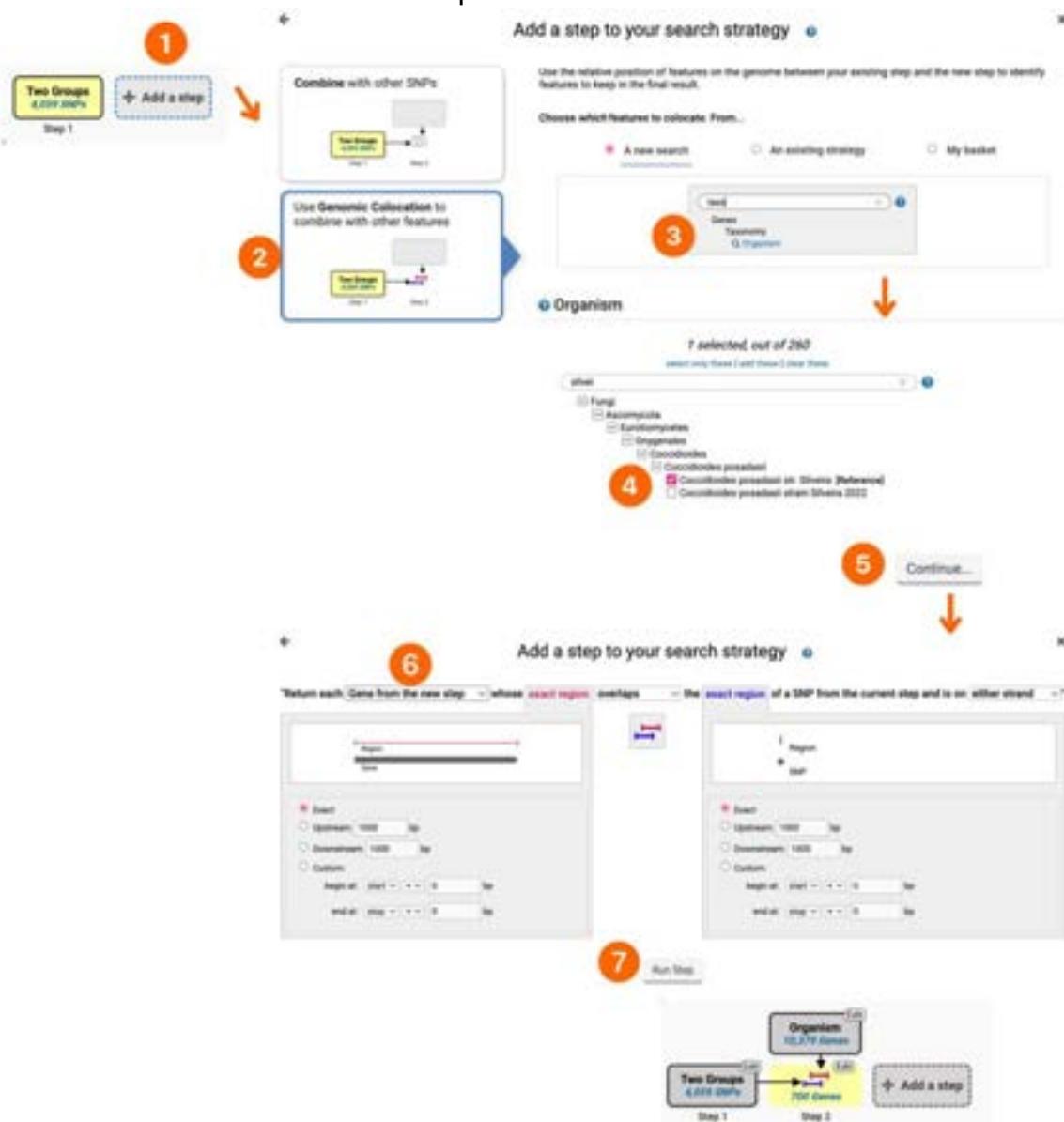
Clicking on the “view DNA-seq reads” link will re-direct you to a JBrowse highlighting SNPs detected. You can select more tracks to examine by clicking on the Select Tracks tab on the left.





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

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



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

Strategy URL:

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



C. 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.

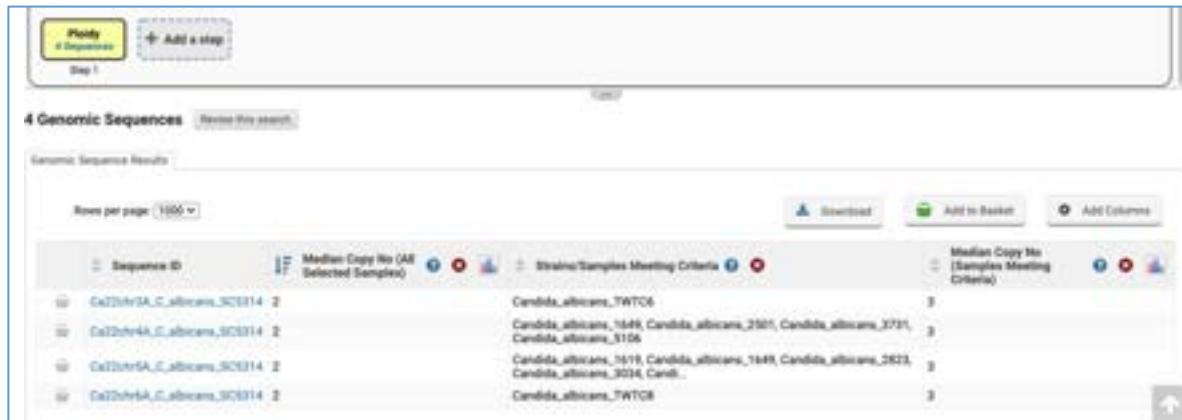
C.1. Copy Number/Ploidy search (Genomic Sequences)

Find genomic sequences or chromosomes based on their estimated copy number in resequenced strains. Genomic sequences returned by the search will 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 FungiDB search interface. Step 1 highlights the search bar with 'ploid' and the 'Copy Number/Ploidy' dropdown. Step 2 highlights the 'Organism' dropdown set to 'Candida albicans SC5314'. Step 3 highlights the 'Data Set' section where 'SNP calls on WGS of Candida albicans clinical isolates (oropharyngeal candidiasis)' is selected. Step 4 highlights the 'Copy Number >=' input field set to '3'. Step 5 highlights the 'Median Or By Strain/Sample?' dropdown set to 'By Strain/Sample'. Step 6 highlights the 'By Strain/Sample (at least one selected strain/sample must be chosen)' dropdown.

The search by strain/sample (i.e., at one or more of the selected strains must match the criteria rather than the median of the selected strains matching) is intended to find chromosomes where the whole chromosome is duplicated. It may find chromosomes where partial aneuploidy involves most of the chromosome but is unlikely to find chromosomes where partial aneuploidy only covers a small region. Also, because this search currently relies on coverage alone, it will not find instances of global genome duplication (e.g., all chromosomes became triploid).



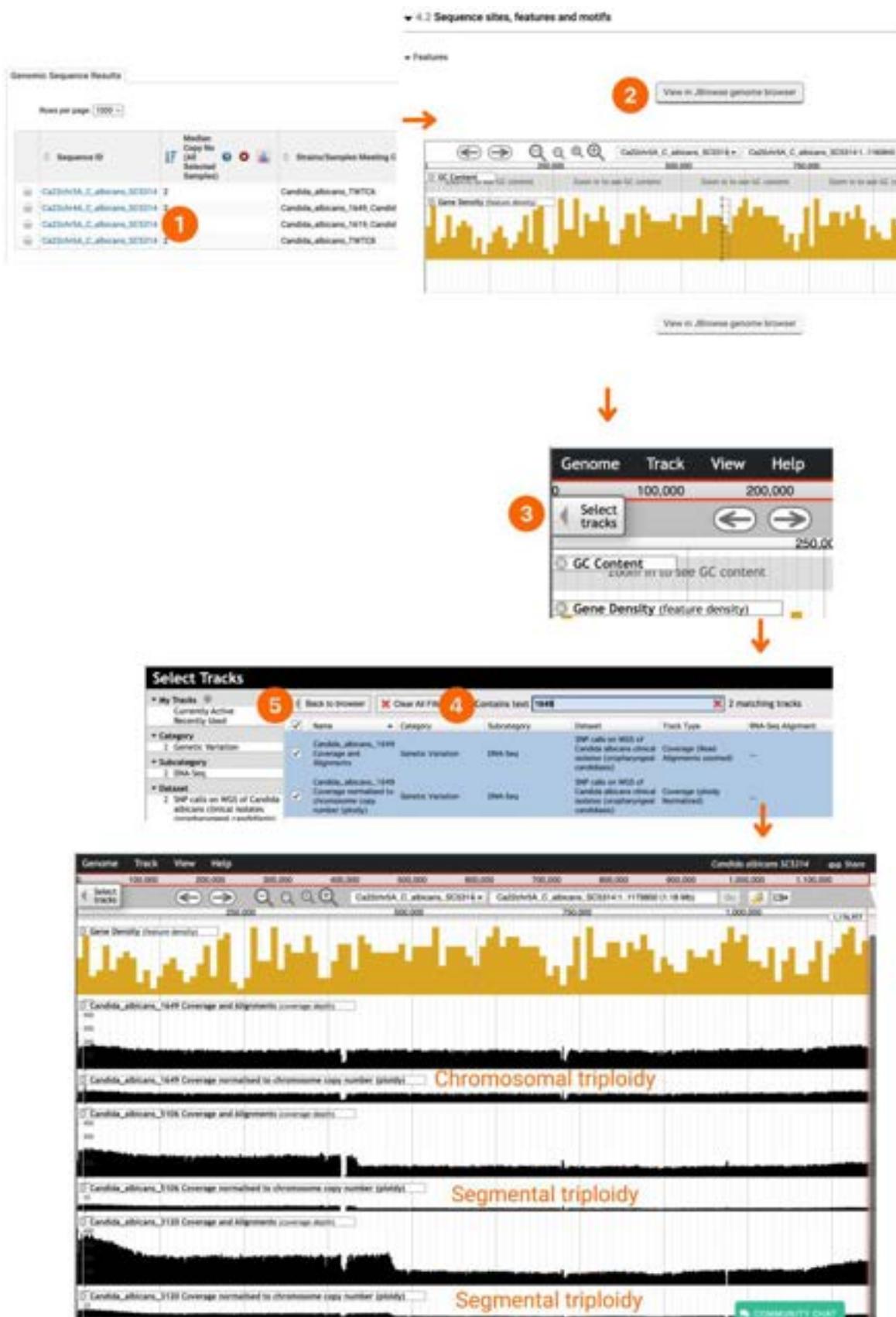
Sequence ID	Median Copy No (All Selected Samples)	Strains/Samples Meeting Criteria	Median Copy No (Samples Meeting Criteria)
Ca22chr5A_C_albicans_SC5314	2	Candida_albicans_TWTC6	3
Ca22chr4A_C_albicans_SC5314	2	Candida_albicans_1649, Candida_albicans_2501, Candida_albicans_3731, Candida_albicans_3106	3
Ca22chr5A_C_albicans_SC5314	2	Candida_albicans_1649, Candida_albicans_1649, Candida_albicans_2501, Candida_albicans_3024, Candida_albicans_TWTC8	3
Ca22chr5A_C_albicans_SC5314	2	Candida_albicans_TWTC8	3

- **Explore segmental aneuploidy in JBrowse.**

JBrowse has two coverage tracks:

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

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



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

URL:

[https://fungidb.org/fungidb/browse/index.html?loc=Ca22chr5A_C_albicans_SC5314%3A1..1190845&data=%2Ffungidb%2Fservice%2Fjb%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%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/browse/index.html?loc=Ca22chr5A_C_albicans_SC5314%3A1..1190845&data=%2Ffungidb%2Fservice%2Fjb%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%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/>

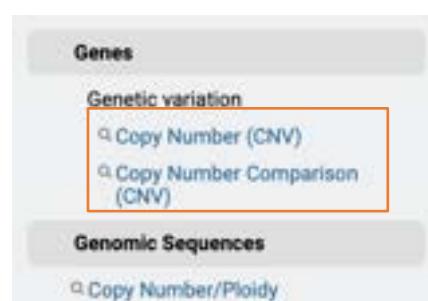
C.2. Copy Number search

(Genes) Using Gene

Searches

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

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



Different metrics for defining copy number:

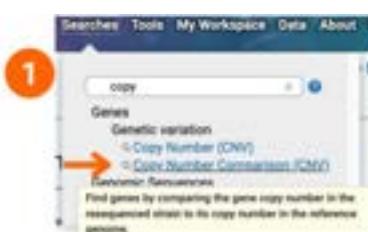
- **Haploid number** is the number of genes on an individual chromosome.
- **Gene dose** is the total number of genes in an organism, accounting for copy number of the chromosome.

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



- 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

1 *Candida albicans* SC5314

Strain/Sample

2 263 Strain/Sample Total 1 of 263 Strain/Sample selected

3 **Fungal strain**

4 *Candida albicans* 5106

5 Median Or By Strain/Sample?

6 What comparison do you want to make?

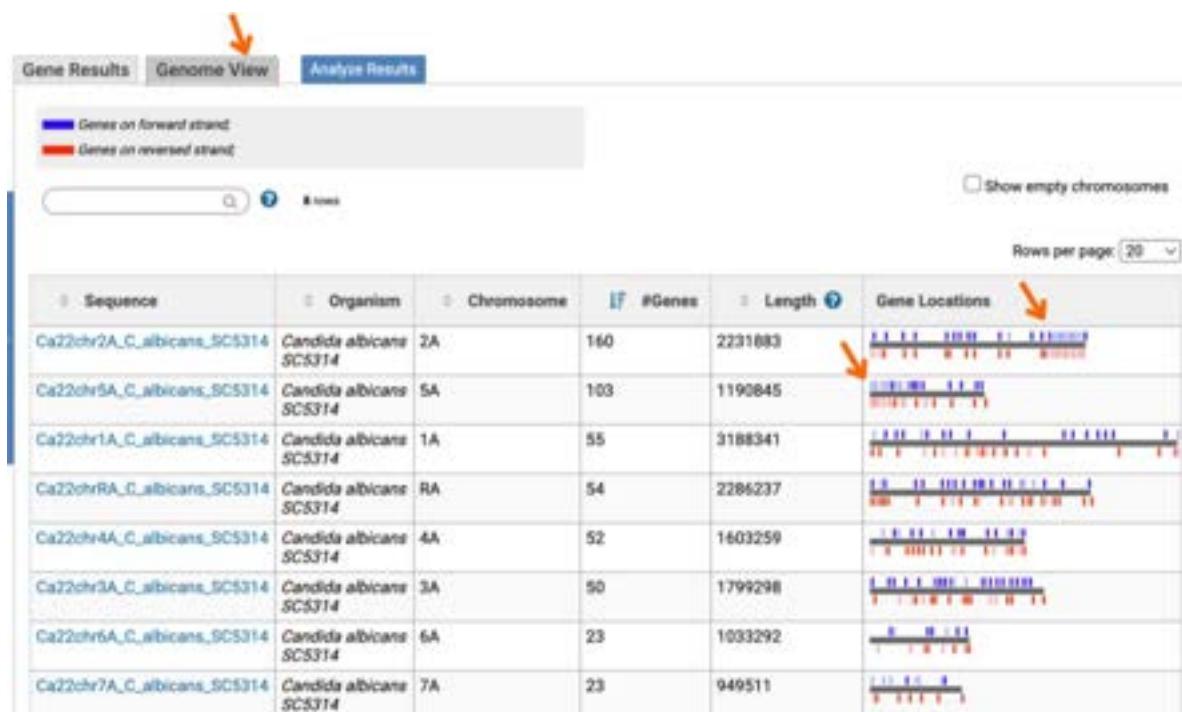
Get Answer

CopyNumberComparison
520 Genes

+ Add a step

Step 1

Examine the results using the Genome View option.



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

Strategy URL:

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

Exercise: Exploring variants in Ensembl Fungi

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

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

 from any of these views.

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



The screenshot shows the Ensembl search interface. In the search bar, 'Saccharomyces cerevisiae' is selected, and 'YGL256W' is entered. Below the search bar, there is a note: 'e.g. NAT2 or alcohol*'. A 'Go' button is visible.

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

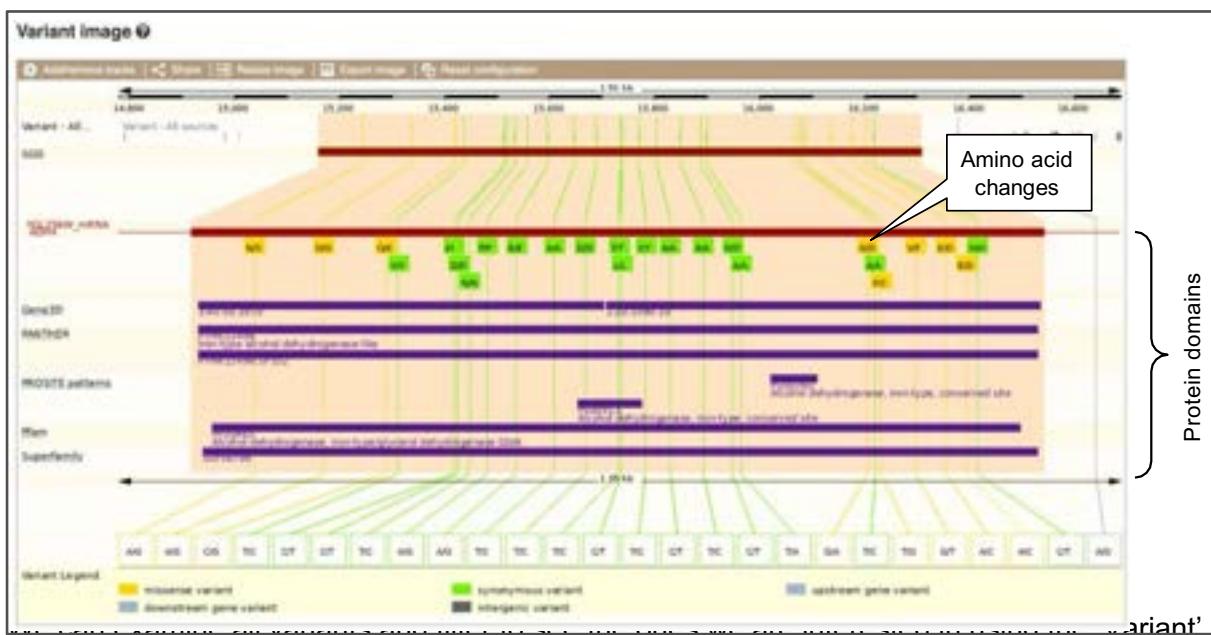
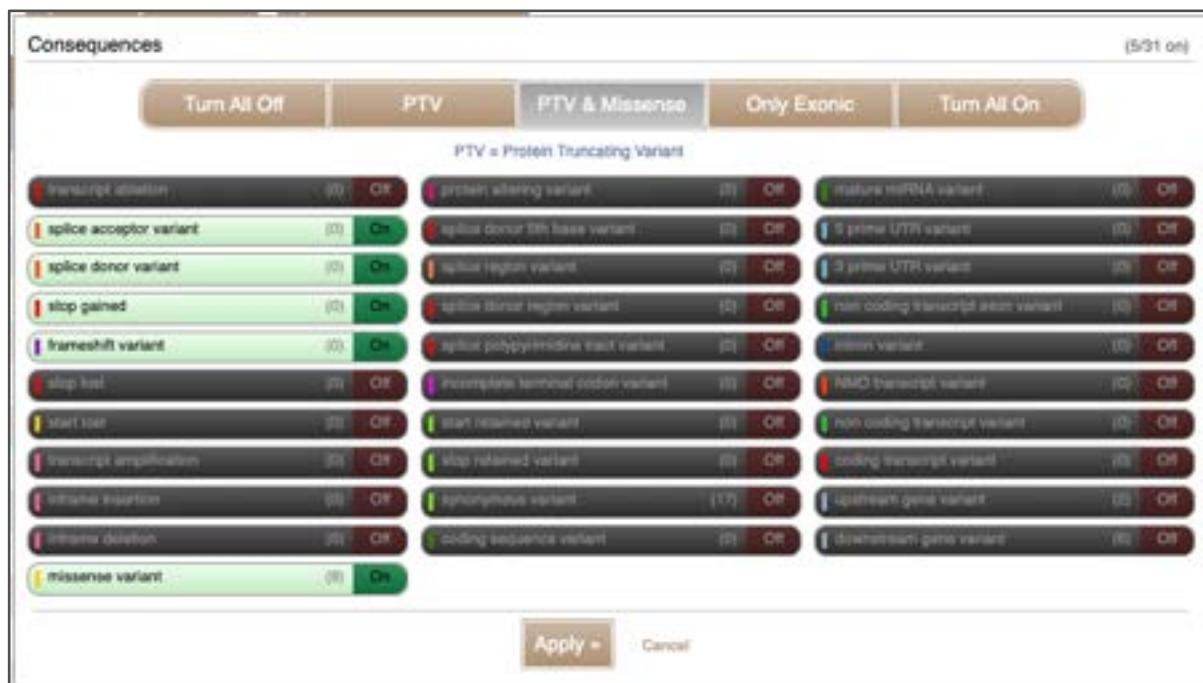


table. Click on the [Variant table](#) link on the left-hand menu.

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

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

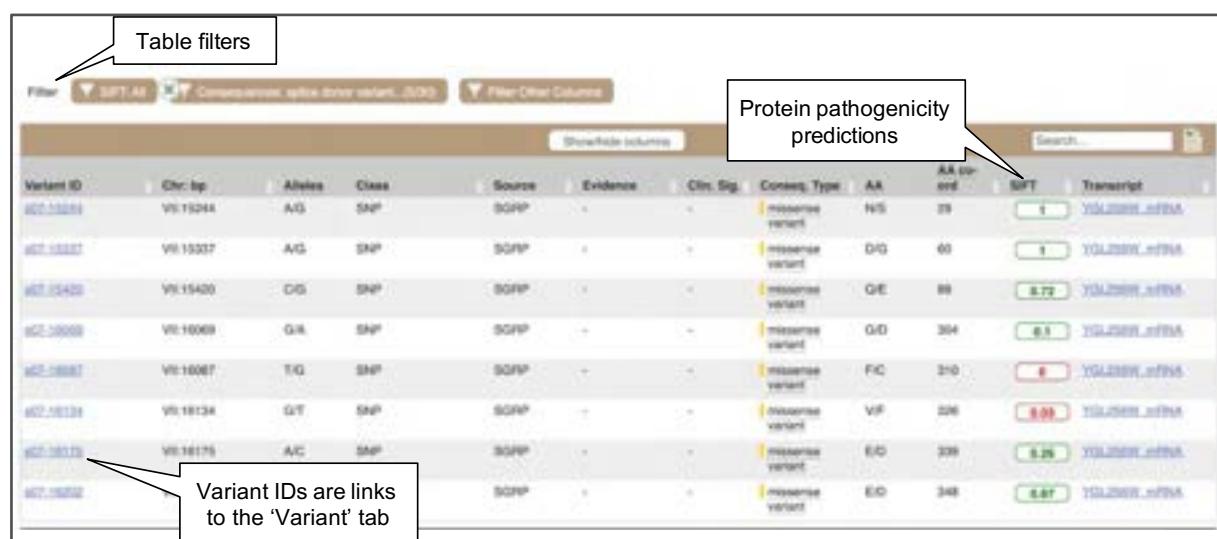
You can filter the table to view variants that alter the protein sequence. Click on the **Consequences: All** button above the table. Click the option '**PTV and Missense**' in the pop-up, then **Apply**. You can also filter by other columns such as variant **Class**.



Consequence Type	Description	Status
transcript start	OK	OK
splice acceptor variant	ON	protein altering variant
splice donor variant	ON	splice donor 5' flanking variant
stop gained	ON	splice region variant
frameshift variant	ON	splice donor 3' flanking variant
stop lost	OFF	splice polyypyrimidine tract variant
start lost	OFF	incomplete terminal codon variant
transcript amplification	OFF	start retained variant
intronic insertion	OFF	stop retained variant
intronic deletion	OFF	synonymous variant
missense variant	ON	coding sequence variant
protein altering variant	OFF	mature mRNA variant
5 prime UTR variant	OFF	5 prime UTR variant
3 prime UTR variant	OFF	non coding transcript exon variant
non coding transcript variant	OFF	non variant
NMD transcript variant	OFF	NMD transcript variant
non coding transcript variant	OFF	non coding transcript variant
coding transcript variant	OFF	coding transcript variant
upstream gene variant	OFF	upstream gene variant
downstream gene variant	OFF	downstream gene variant

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

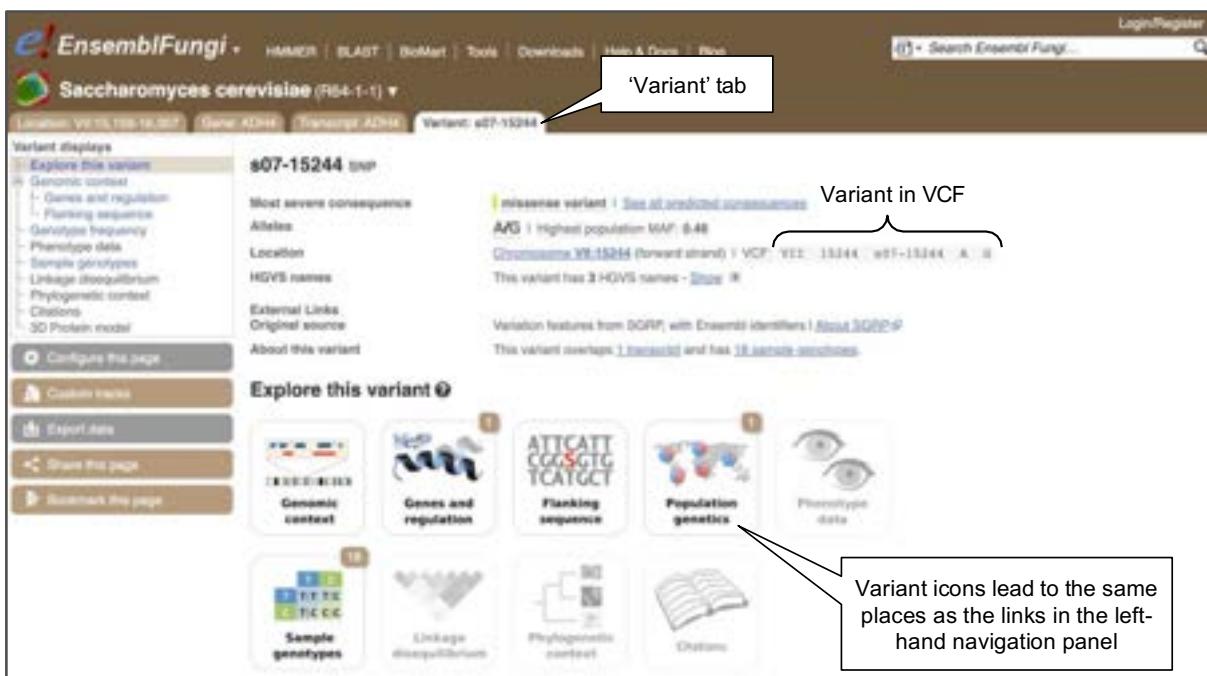
The SIFT scores (<https://doi.org/10.1093/nar/gkg509>) predict the consequence of the variant on the function of the protein taking into account chemical changes and conservation of amino acids. Scores <0.05 and coloured red are 'deleterious' while scores >0.05 and coloured green are 'tolerated'.



Showable columns										Search...
Variant ID	Chr:bp	Alleles	Class	Source	Evidence	Clin. Sig.	Conseq. Type	AA change	SIFT	Transcript
rs223261	VI:15244	A/G	SNP	SGRP	-	-	missense variant	N.S.	0.99	YOL200W_mRNA
rs223267	VI:15337	A/G	SNP	SGRP	-	-	missense variant	D.G.	0.01	YOL200W_mRNA
rs223269	VI:15420	C/G	SNP	SGRP	-	-	missense variant	G.E.	0.99	YOL200W_mRNA
rs223270	VI:16069	G/A	SNP	SGRP	+	-	missense variant	G.D.	0.01	YOL200W_mRNA
rs223271	VI:16067	T/G	SNP	SGRP	+	-	missense variant	F.I.C.	0.01	YOL200W_mRNA
rs223274	VI:16134	G/T	SNP	SGRP	-	-	missense variant	V.I.F.	0.01	YOL200W_mRNA
rs223275	VI:16175	A/C	SNP	SGRP	+	-	missense variant	E.I.O.	0.01	YOL200W_mRNA
rs223272				SGRP	-	-	missense variant	E.I.O.	0.01	YOL200W_mRNA

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

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



The screenshot shows the Ensembl Fungi interface for *Saccharomyces cerevisiae*. The 'Variant' tab is active. In the top right, there's a search bar and a 'Login/Register' button. The main content area displays variant details: ID s07-15244, consequence missense variant, allele AAG, population frequency 8.4%, and chromosomal location VII:15244. Below this, a 'Variant in VCF' section is shown. To the right, a box contains the text: 'Variant icons lead to the same places as the links in the left-hand navigation panel'. The bottom right of the main content area has a box containing the text: 'This variant overlaps 1 transcript and has 18 transcript consequences.'

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

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

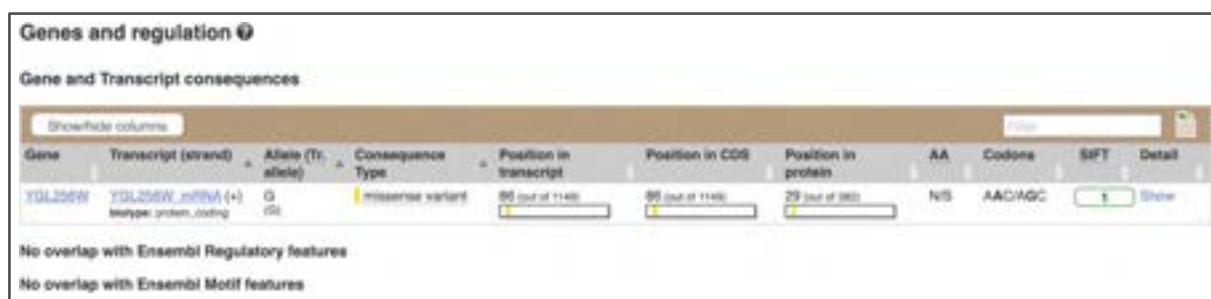
Location	Chromosome VII:15244 (forward strand) VCF: VII 15244 s07-15244 A G
----------	--

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

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

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

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

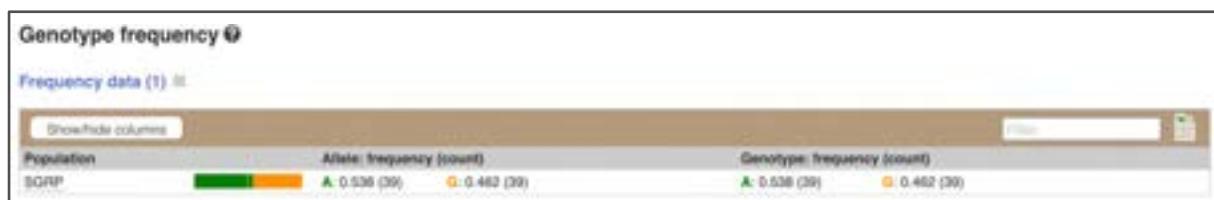


The screenshot shows the 'Genes and regulation' section for the variant s07-15244. It includes a table of gene and transcript consequences. The table has columns for Gene, Transcript (strand), Allele (Tr. allele), Consequence Type, Position in transcript, Position in CDS, Position in protein, AA, Codons, SIFT, and Detail. One row is shown for YGL256W, which is annotated as a missense variant at position 86 of 1140, 86 of 1140, and 29 of 982, with AA N15 and Codons AAC/AGC. The SIFT score is 1. Below the table, two small text boxes state: 'No overlap with Ensembl Regulatory features' and 'No overlap with Ensembl Motif features'.

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

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

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



Additional Exercise – Variation data in *Fusarium oxysporum*

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

Answers

- You can find the alleles in the summary information at the top of the ‘Variant’ tab. The reference allele for tmp_10_6610 is C and the alternative allele is T.

tmp_10_6610 SNP

Most severe consequence	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.

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

Genotype frequency

Frequency data (1)	
Population	Allele: frequency (count)
melonis	C: 0.850 (17) T: 0.150 (3)
	Genotype: frequency (count)
	C/C: 0.800 (8) C/T: 0.100 (1) T/T: 0.100 (1)

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

Sample genotypes

Search for a sample:		Search (e.g. W419507)	Back to list	
Genotypes for melonis				
Sample (Male/Female/Unknown)	Genotype (forward strand)	Population(s)	Father	Mother
W419509 (U)	CC	melonis	-	-
W419504 (U)	CG	melonis	-	-
W419505 (U)	CC	melonis	-	-
W419506 (U)	CC	melonis	-	-
W419507 (U)	CC	melonis	-	-
W419508 (U)	CC	melonis	-	-
W419510 (U)	CC	melonis	-	-
W419509 (U)	CC	melonis	-	-
W419504 (U)	CT	melonis	-	-
W419505 (U)	TT	melonis	-	-

Exercise: The Ensembl Fungi Variant Effect Predictor (VEP)

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

- 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?

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

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

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

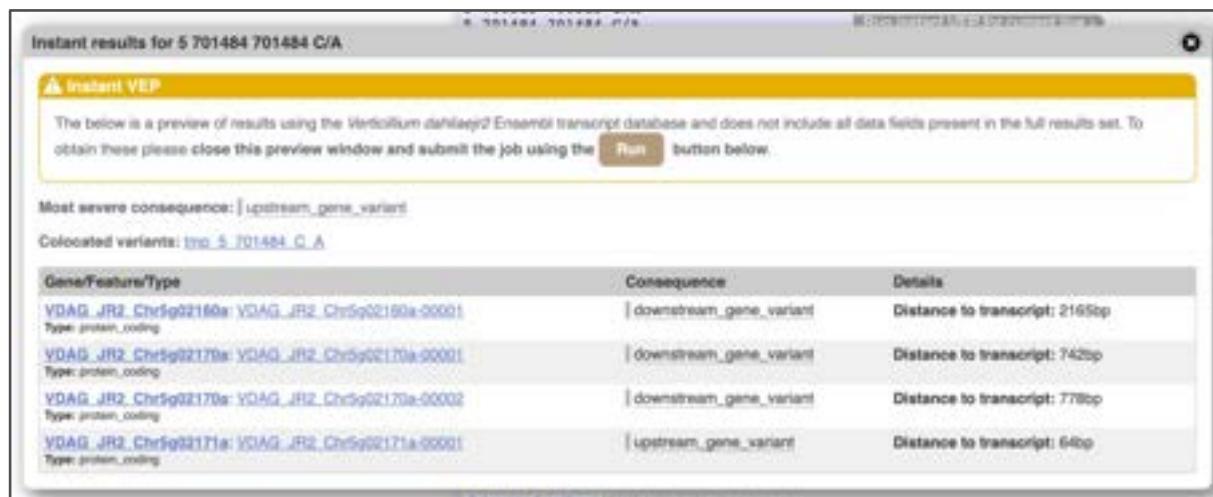
5 698711 698711 C/G
 5 698935 698935 G/T
 5 700313 700313 G/A
 5 701484 701484 C/A



The screenshot shows the 'Variant Effect Predictor' interface. Key elements include:

- Species:** Set to *Verticillium dahliae* JR2 v6.0. A 'Change species' button is available.
- Name for this job (optional):** A field labeled 'Name your job'.
- Input data:** A text area containing the variants: 5 698711 698711 C/G; 5 698935 698935 G/T; 5 700313 700313 G/A; 5 701484 701484 C/A. Below it are buttons for 'Paste or type in variants...', '...or upload a file...', and '...or provide a URL to a file hosted online'.
- Preview section:** Shows the variants and their effects. It includes a 'Run instant VEP' button and a link to 'See a preview of the results for the selected variant'.
- Help section:** A 'See data format examples' link.

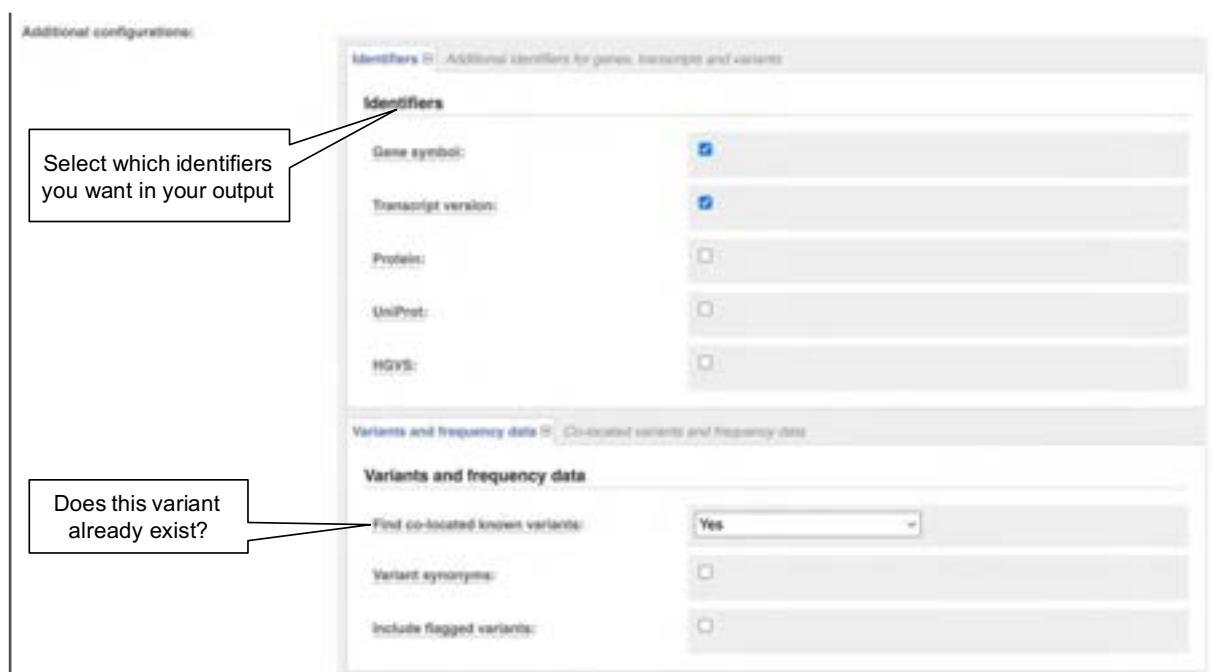
The VEP will automatically detect that the data is in Ensembl default format. Clicking on the [Run instant VEP for current line](#) option will generate a pop-up with summarised results for that individual variant.



The screenshot shows the 'Instant results' window for variant 5_701484_701484_C/A. It displays a preview of results using the *Verticillium dahliae* Ensembl transcript database. The most severe consequence is listed as 'upstream_gene_variant'. Below this, a table lists co-located variants for the gene **VDAG_JR2**. The variants are categorized by consequence: downstream_gene_variant (distance to transcript: 216bp), downstream_gene_variant (distance to transcript: 742bp), downstream_gene_variant (distance to transcript: 778bp), and upstream_gene_variant (distance to transcript: 64bp). All variants are of type protein_coding.

Gene/Feature/Type	Consequence	Details
VDAG_JR2_Chromosome:5g02160a:VDAG_JR2_Chromosome:5g02160a-00001 Type: protein_coding	downstream_gene_variant	Distance to transcript: 216bp
VDAG_JR2_Chromosome:5g02170a:VDAG_JR2_Chromosome:5g02170a-00001 Type: protein_coding	downstream_gene_variant	Distance to transcript: 742bp
VDAG_JR2_Chromosome:5g02170a:VDAG_JR2_Chromosome:5g02170a-00002 Type: protein_coding	downstream_gene_variant	Distance to transcript: 778bp
VDAG_JR2_Chromosome:5g02171a:VDAG_JR2_Chromosome:5g02171a-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 'Additional configurations' menu. Two specific sections are highlighted with callouts:

- Select which identifiers you want in your output**: Points to the 'Identifiers' section, which includes fields for Gene symbol, Transcript version, Protein, UniProt, and HGVS.
- Does this variant already exist?**: Points to the 'Variants and frequency data' section, which includes fields for Find co-located known variants (set to Yes), Variant synonymy, and Include flagged variants.

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

Login/Register

Search Ensembl Fungi...

Add information about affected transcripts and proteins

Species: **Saccharomyces_cerevisiae** Assembly: R54-1-1 Chaperones

Name for this job (optional):

Input data: Either paste data:
Example: Ensembl_id:1, ID:1, UniProt IDs: 1Q83,jobname
Or upload file: Choose file: No file chosen
Or provide file URL:

Additional configurations:

Identifiers: Additional identifiers for genes, transcripts and variants

Identifiers

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

Variants and frequency data: Co-localized variants and frequency data

Variants and frequency data

- Find co-localized known variants: Yes
- Variant synonyms:
- Include flagged variants:

Additional annotations: Additional transcript, protein and regulatory annotations

Transcript annotation

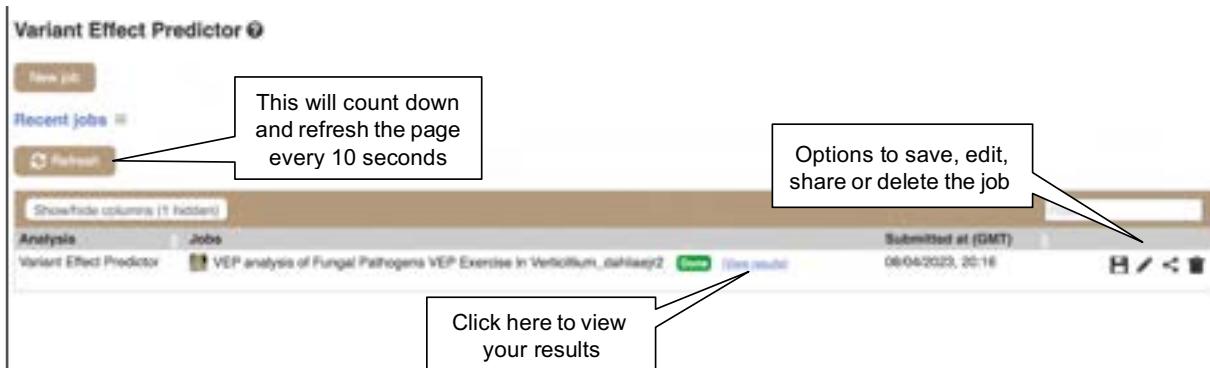
- Transcript biotype:
- Exon and Intron numbers:
- Intronic/exon/intron junctions

Show only coding variants

More filters

Run VEP

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



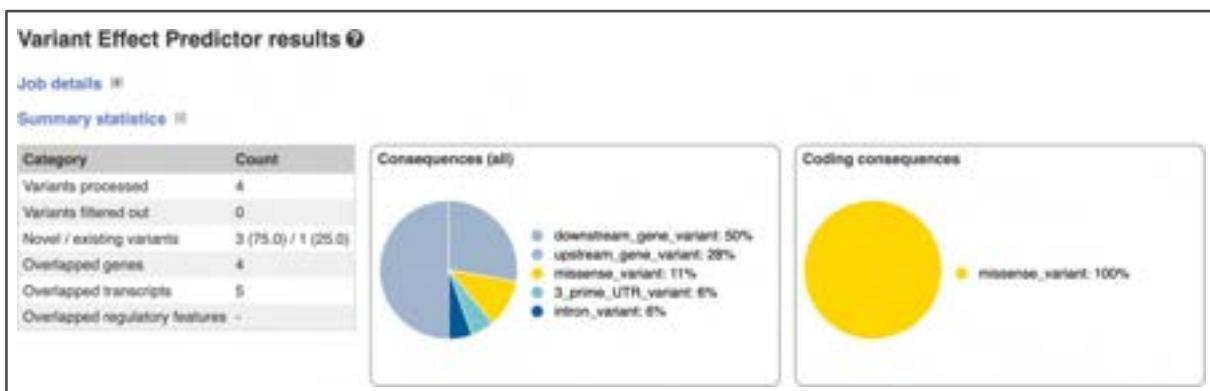
This screenshot shows the Variant Effect Predictor interface. At the top left, there are buttons for 'New job', 'Recent jobs' (with a dropdown menu), and 'Refresh'. A callout box points to the 'Refresh' button with the text: 'This will count down and refresh the page every 10 seconds'. At the top right, there are options to 'Save', 'Edit', 'Share', and 'Delete' the job. Another callout box points to these options with the text: 'Options to save, edit, share or delete the job'. Below the header, the analysis type is listed as 'Variant Effect Predictor' and the job title is 'VEP analysis of Fungal Pathogens VEP Exercise In Verticillium_candidum'. The status is shown as 'Done result'. The date and time of submission are 'Submitted at (GMT) 08/06/2023, 20:18'. A large button in the center says 'Click here to view your results'.

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

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

Let's come back to our questions:

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



This screenshot shows the 'Variant Effect Predictor results' interface. On the left, 'Job details' and 'Summary statistics' are listed. The 'Summary statistics' table includes:

Category	Count
Variants processed	4
Variants filtered out	0
Novel / existing variants	3 (75.0) / 1 (25.0)
Overlapped genes	4
Overlapped transcripts	5
Overlapped regulatory features	-

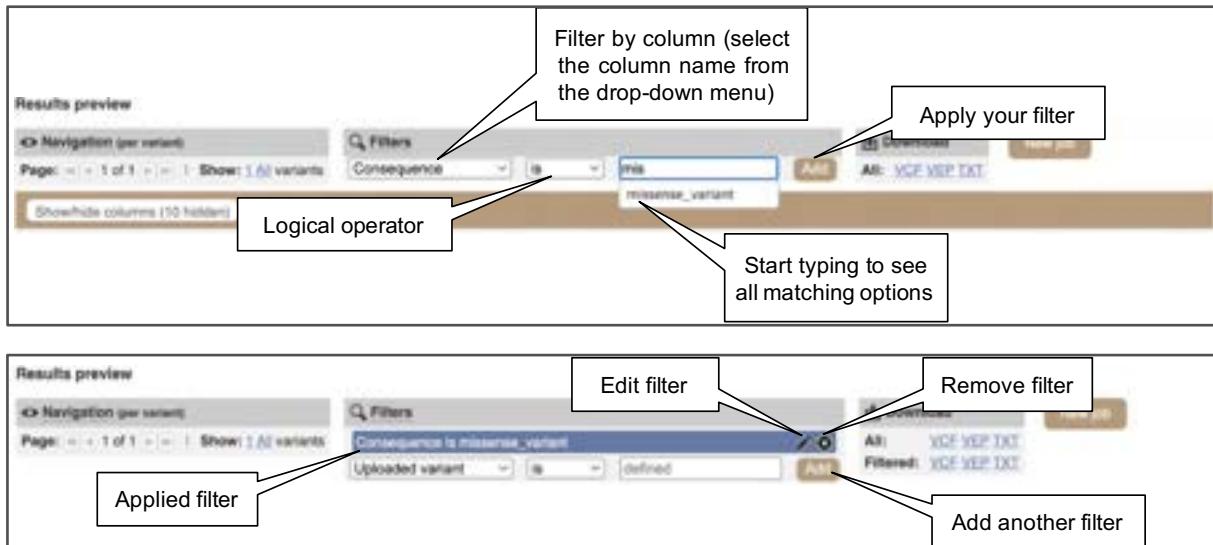
In the center, there are two charts: 'Consequences (all)' and 'Coding consequences'. The 'Consequences (all)' chart is a pie chart showing the distribution of variant consequences:

Consequence	Percentage
downstream_gene_variant	50%
upstream_gene_variant	25%
missense_variant	11%
3_prime_UTR_variant	8%
intron_variant	6%

The 'Coding consequences' chart is a pie chart showing the distribution of coding variant consequences:

Consequence	Percentage
missense_variant	100%

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.



Results preview

Navigation (per variant)

Page: 1 of 1 | Show: 1,61 variants

Filters

Consequence: miss

Show/hide columns (10 hidden)

Logical operator

Filter by column (select the column name from the drop-down menu)

Start typing to see all matching options

Apply your filter

Results preview

Navigation (per variant)

Page: 1 of 1 | Show: 1,61 variants

Filters

Consequence is missense_variant

Uploaded variant: defined

All: XCF-VEP-TCI
Filtered: XCF-VEP-TCI

Applied filter

Edit filter

Remove filter

Add another filter

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

Filter this table
Download options

Results preview
Navigation (per variant)
Filters
Downloaded
New job

Show/hide columns (10 hidden)
Show additional columns

Variant 1
Variant 2
Variant 3
Variant 4

Uploaded variant	Location	Allele	Context		Biotype	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand
5_698711_C/G	5_698711_698711	G	downstream		VDAG_JR2_Chromosome	1							
5_698711_C/G	5_698711_698711	G	intron_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_698711_C/G	5_698711_698711	G	upstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_698711_C/G	5_698711_698711	G	upstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_698711_C/G	5_698711_698711	G	downstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	-1						
5_698935_G/T	5_698935_698935	T	downstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_698935_G/T	5_698935_698935	T	3_prime_UTR_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_698935_G/T	5_698935_698935	T	upstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_698935_G/T	5_698935_698935	T	upstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_698935_G/T	5_698935_698935	T	downstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	-1						
5_700313_G/A	5_700313_700313	A	downstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_700313_G/A	5_700313_700313	A	missense_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_700313_G/A	5_700313_700313	A	missense_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_700313_G/A	5_700313_700313	A	downstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	-1						
5_701484_C/A	5_701484_701484	A	downstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_701484_C/A	5_701484_701484	A	downstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_701484_C/A	5_701484_701484	A	downstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	1						
5_701484_C/A	5_701484_701484	A	upstream_gene_variant	VDAG_JR2_Chromosome	Transcript	VDAG_JR2_Chromosome	-1						

Existing variants

235

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

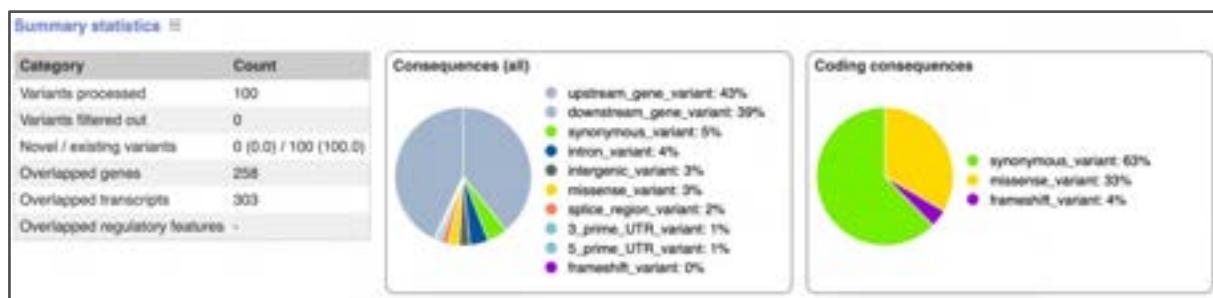
On the course file page, you will find a VCF file labelled VEP_exercise.vcf. This is a small subset of the outcome of *Puccinia graminis* (Ug99) whole genome sequencing and variant calling experiment. This file can also be found on our FTP site under the following link:
http://ftp.ebi.ac.uk/pub/databases/ensembl/training/2021/FungalPathogens/VEP_exercise.vcf

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

- 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.*)
- What are the HGVSp notations of missense variants falling in known protein domains?
- How many variants are frameshift? Which gene(s) do they fall in and which exons? Can you find a UniParc ID of protein(s) affected by this variant?

Answer

- 100 variants have been processed.
- The variants overlap 258 genes and 303 transcripts.



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

GMQ_21813
 GMQ_27112
 GMQ_04080
 GMQ_06767
 GMQ_02814
 GMQ_20311
 GMQ_20457
 GMQ_03045



Results preview													
All Navigation per variant		Filters		Download		More info							
Show: 1,000 variants		Protein matches		All		Filtered							
Downloaded variants (0) Filtered variants (0)													
Location	Amino	Consequence	Date	Effect	HGVSp	HGVSp	Protein position	Protein domain	Protein name	Protein variant	Protein domains		
Supernova_2.158.127961-001	T	missense_variant	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	200	200	EF	GMQ_27112T0:p.Glu89Lys	Phen:PT14079 PROTEIN_DOMAINS:130 PROTEIN_DOMAINS:131-139 SwissProt:640402960		
Supernova_2.158.127961-002	C	missense_variant	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	200	200	EF	GMQ_27112T0:p.Glu89Lys	GeneID:316750 16-16 PROTEIN_profiles:771028P Superfamily:SF070002		
Supernova_2.158.127961-003	G	missense_variant	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	1000	1000	EF	GMQ_27112T0:p.Glu89Lys	GeneID:316750 16-16 ProteoDB:PT14079 Superfamily:SF070002 CDD:0000639		
Supernova_2.158.127961-004	T	missense_variant	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	70	70	EF	GMQ_27112T0:p.Glu89Lys	—		
Supernova_2.158.127961-005	G	missense_variant	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	1000	1000	EF	GMQ_27112T0:p.Glu89Lys	PANTHER:PT14079 PANTHER:PT14079-SP1		
Supernova_2.158.127961-006	T	missense_variant	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	300	300	EF	GMQ_27112T0:p.Glu89Lys	GeneID:316750 16-16 PROTEIN_profiles:771028P Superfamily:SF070002		
Supernova_2.158.127961-007	G	missense_variant	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	407	407	EF	GMQ_27112T0:p.Glu89Lys	GeneID:316750 16-16 PROTEIN_profiles:771028P Superfamily:SF070002		
Supernova_2.158.127961-008	T	missense_variant	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	407	407	EF	GMQ_27112T0:p.Glu89Lys	GeneID:316750 16-16 PROTEIN_profiles:771028P Superfamily:SF070002		
Supernova_2.158.127961-009	A	missense_variant	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	300	300	EF	GMQ_27112T0:p.Glu89Lys	PANTHER:PT14079 PANTHER:PT14079-SP1 SwissProt:640402960		

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

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

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

In the VEP results table, apply the following filters:

Consequence is missense_variant
Protein matches is [leave text box empty]

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

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

Results preview													
All Navigation per variant		Filters		Download		More info							
Show: 1,000 variants		Protein matches is defined		All		Filtered							
Downloaded variants (0) Filtered variants (0)													
Location	Amino	Consequence	Date	Effect	HGVSp	HGVSp	Protein position	Protein domain	Protein name	Protein variant	Protein domains		
Supernova_2.158.127961-001	C	missense_variant	GMQ_27112T0	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	900 000	900 000	EF	GMQ_27112T0:p.Glu89Lys	Superfamily:SF070002P PROTEIN_profiles:771028P GeneID:2158127961 16-16 ProteoDB:PT14079		
Supernova_2.158.127961-002	T	missense_variant	GMQ_27112T0	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	900 000	900 000	EF	GMQ_27112T0:p.Glu89Lys	ProteoDB:PT14079		
Supernova_2.158.127961-003	G	missense_variant	GMQ_27112T0	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	900 000	900 000	EF	GMQ_27112T0:p.Glu89Lys	PANTHER:PT14079 PANTHER:PT14079-SP1 SwissProt:640402960		
Supernova_2.158.127961-004	T	missense_variant	GMQ_27112T0	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	900 000	900 000	EF	GMQ_27112T0:p.Glu89Lys	GeneID:2158127961 16-16 PROTEIN_profiles:771028P Superfamily:SF070002		
Supernova_2.158.127961-005	G	missense_variant	GMQ_27112T0	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	900 000	900 000	EF	GMQ_27112T0:p.Glu89Lys	GeneID:2158127961 16-16 PROTEIN_profiles:771028P Superfamily:SF070002		
Supernova_2.158.127961-006	A	missense_variant	GMQ_27112T0	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	900 000	900 000	EF	GMQ_27112T0:p.Glu89Lys	ProteoDB:PT14079		
Supernova_2.158.127961-007	C	missense_variant	GMQ_27112T0	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	900 000	900 000	EF	GMQ_27112T0:p.Glu89Lys	GeneID:2158127961 16-16 PROTEIN_profiles:771028P Superfamily:SF070002		
Supernova_2.158.127961-008	G	missense_variant	GMQ_27112T0	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	900 000	900 000	EF	GMQ_27112T0:p.Glu89Lys	GeneID:2158127961 16-16 PROTEIN_profiles:771028P Superfamily:SF070002		
Supernova_2.158.127961-009	T	missense_variant	GMQ_27112T0	2003-02-12T00:00:00	GMQ_27112T0:p.Glu89Lys	GMQ_27112T0:p.Glu89Lys	900 000	900 000	EF	GMQ_27112T0:p.Glu89Lys	GeneID:2158127961 16-16 PROTEIN_profiles:771028P Superfamily:SF070002		

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

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

Apply the [Consequence is frameshift_variant](#) under 'Filters'. There is 1 frameshift variant which falls in the GMQ_27001 gene on exon 1 (out of 3). The UniParc ID is UPI0004E9C5AE.

Results preview

Navigation per variant

Show: 1 5 10 25 All variants

Filters: Consequence is frameshift_variant

Download: All: VCF VEP TSV New job

Uploaded variant: 14 defined

Show/Hide column (LT) hidden

Location	Allele	Consequence	Gene	Exon	Codons	Existing variant	ENSP	SWISSPROT	UNIPARC
Swinecyste 3.1482_1086-1087	-	frameshift_variant	GMQ_27001	1/3	GGG/GCG	iso_Swinecyste 3.1482_1086_GGG_GCG	GMQ_2700179	-	UPI0004E9C5AE

Show: 1 5 10 25 All variants

MycoCosm: KEGG Browser

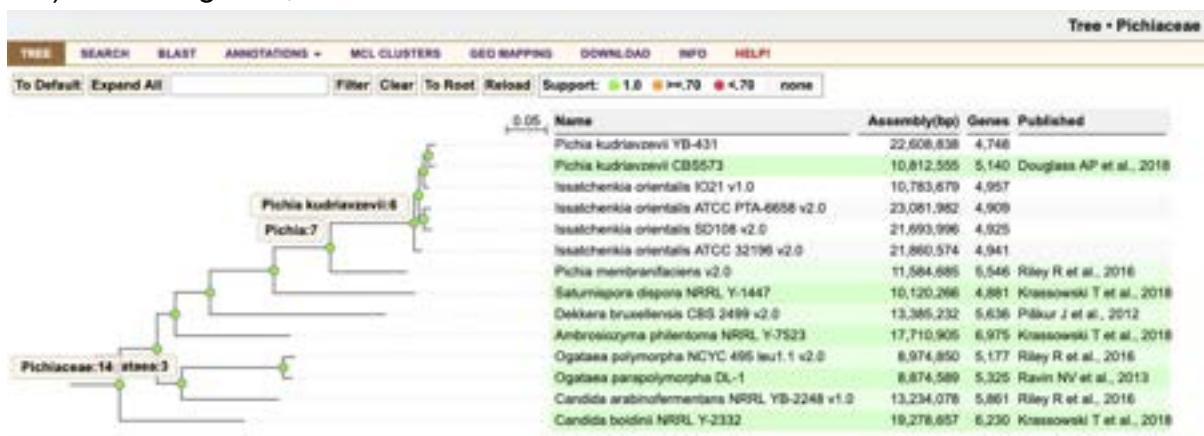
KEGG (Kyoto Encyclopedia of Genes and Genomes; <http://www.genome.jp/kegg/>), is a resource which maintains a curated set of annotated enzymes and their associated metabolic pathways. Each portal's KEGG Browser facilitates display and discovery of MycoCosm's KEGG-annotated genes. Using the KEGG browser, 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.

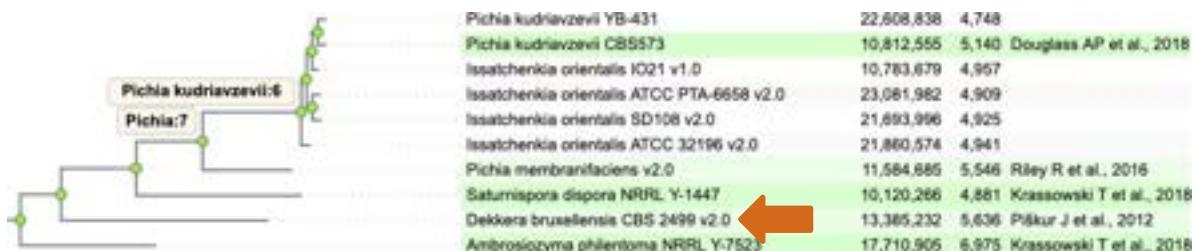
- 1) Go to the MycoCosm Pichiaceae PhyloGroup at mycocosm.jgi.doe.gov/Pichiaceae:

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 boidini NRRL Y-2332	19,278,657	6,230	Kressowski T et al., 2018					
3 Dekkera bruxellensis CBS 2499 v2.0	13,385,232	5,636	Pilkur 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 parapolymerophora 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 kudriavzevi CBS5573	10,812,555	5,140	Douglass AP et al., 2018					
11 Pichia kudriavzevi YB-431	22,608,838	4,748						
12 Pichia membranifaciens v2.0	11,584,685	5,546	Riley R et al., 2016					
13 Saturnispora dispersa NRRL Y-1447	10,120,266	4,881	Kressowski T et al., 2018					

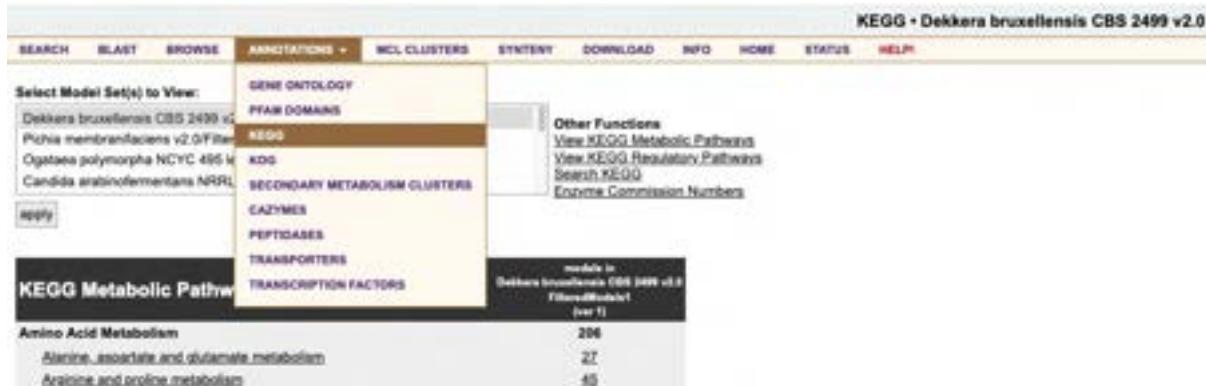
- 2) To verify that *Dekkera* (which grows on galactose) and *Pichia* (which does not) are sibling taxa, click on "TREE":



- 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. The top navigation bar includes SEARCH, BLAST, BROWSE, ANNOTATIONS (which is selected), NCCL CLUSTERS, STRENGTH, DOWNLOAD, INFO, HOME, STATUS, and HELP. A dropdown menu under ANNOTATIONS shows options like GENE CATALOGUE, PFAM DOMAINS, KEGG (selected), KOG, and SECONDARY METABOLISM CLUSTERS. Below the search bar, a list of model sets is shown with 'Dekkera bruxellensis CBS 2499 v2.0' selected. The main content area displays KEGG Metabolic Pathways for Amino Acid Metabolism, including Alanine, aspartate and glutamate metabolism (22 genes) and Arginine and proline metabolism (45 genes). A sidebar on the right provides links to KEGG Metabolic Pathways, Regulatory Pathways, and Enzyme Commission Numbers.

- 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	68
Ascorbate and aldarate metabolism	21
Butanoate metabolism	34
C5-Branched dibasic acid metabolism	2
Citrate cycle (TCA cycle)	28
Fructose and mannose metabolism	46
Galactose metabolism	24
Glycolysis / Gluconeogenesis	47
Glyoxylate and dicarboxylate metabolism	10
Inositol phosphate metabolism	27

- 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, 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
1.1.1.18 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. Here, the red boxes indicate enzymes present in *Dekkera*. These include both 2.7.1.6 (Galactokinase) and 2.7.7.12 (UDP-glucose--hexose-1-phosphate uridylyltransferase):

KEGG • Dekkera bruxellensis CBS 2499 v2.0

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

Select Model Set(s) to View:

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

Other Functions

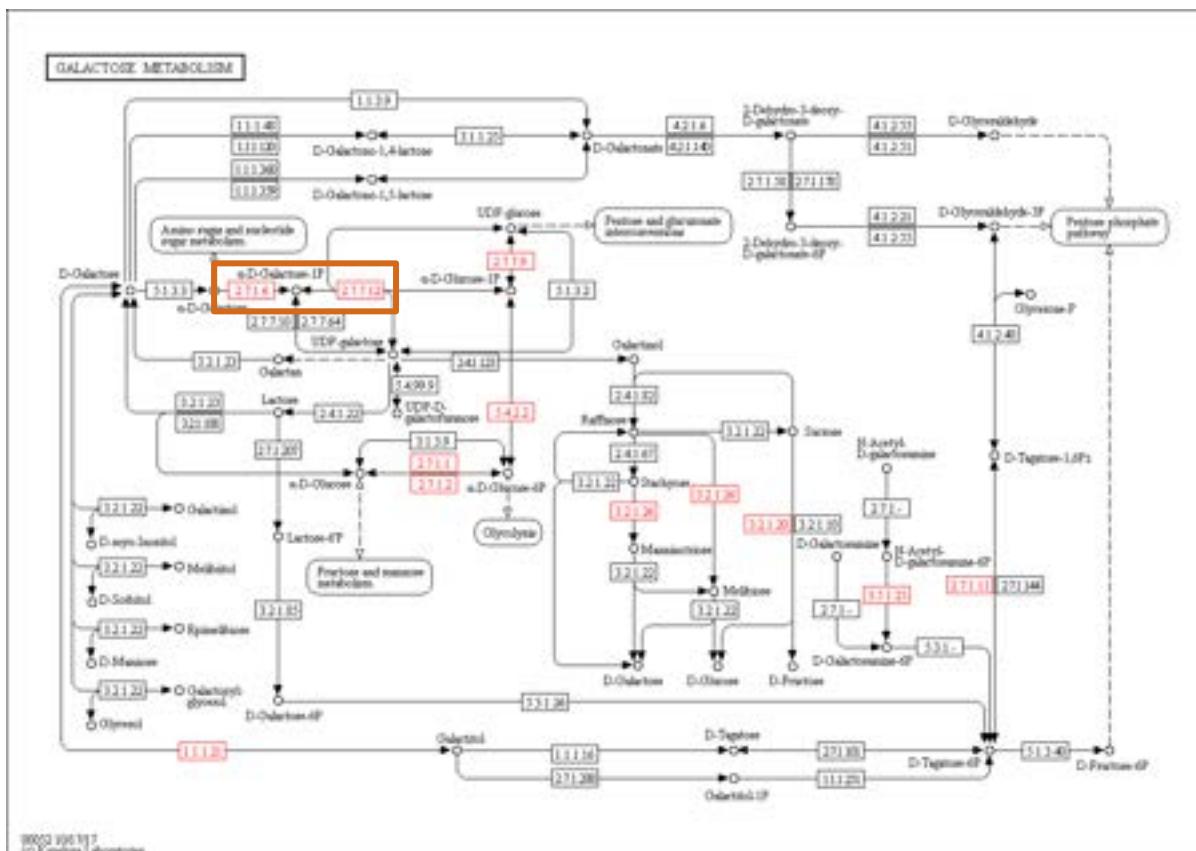
[View KEGG Metabolic Pathways](#)
[View KEGG Regulatory Pathways](#)
[Search KEGG](#)

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
-----------------------	---	--	---

Click here to open KEGG Map



- 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.

KEGG • *Dekkera bruxellensis CBS 2499 v2.0*

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

Select Model Set(s) to View:

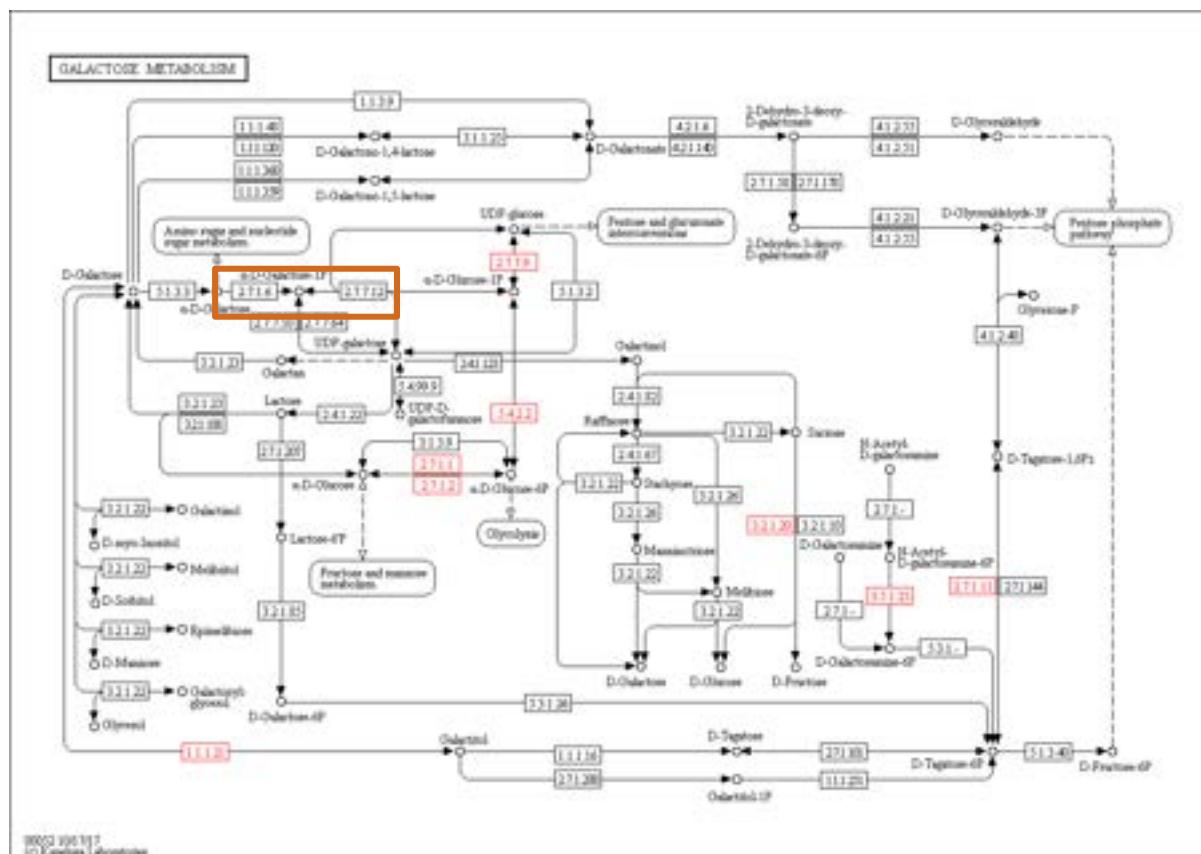
Dekkera bruxellensis CBS 2499 v2.0/FilteredModels1 (ver 1)
Pichia membranifaciens v2.0/FilteredModels1 (ver 1)
Ogataea polymorpha NCYC 495 leu1.1 v2.0/FilteredModels2 (ver 1)
Candida arabinofermentans NRRL YB-2248 v1.0/FilteredModels1 (ver 1)

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

apply

MAP00052: Galactose metabolism
[Summary View | Model View | View KEGG Map]

- 11) Click ‘View KEGG Map’ again. This time, note that neither 2.7.1.6 nor 2.7.7.12 are colored red. No wonder *Pichia* cannot grow on galactose – it is missing the genes coding for key enzymes in the galactose utilization pathway.



Exercise:

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

Reference:

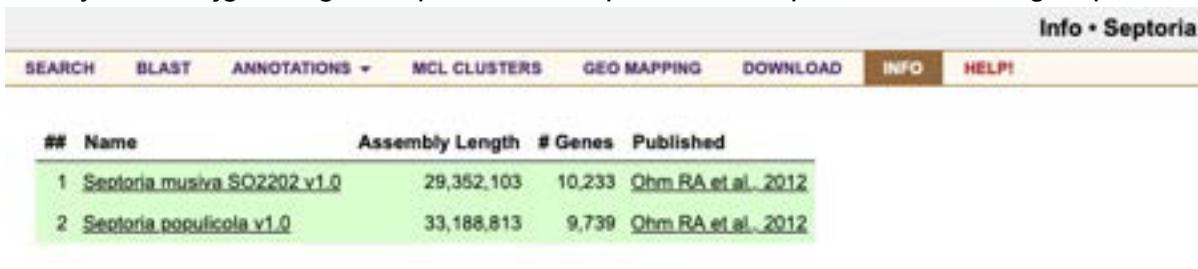
- Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Göker M, Salamov AA, Wisecaver JH, Long TM, Calvey CH, Aerts AL, Barry KW, Choi C, Clum A, Coughlan AY, Deshpande S, Douglass AP, Hanson SJ, Klenk HP, LaButti KM, Lapidus A, Lindquist EA, Lipzen AM, Meier-Kolthoff JP, Ohm RA, Otillar RP, Pangilinan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

MycoCosm: Secondary Metabolism Clusters Browser

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

Scenario: You have identified a toxic SM produced by *Septoria musiva*, a pathogenic fungus that induces cankers in the poplar tree, but not produced by *Septoria 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:



##	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:



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 its 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!	
Genomes			GENE ONTOLOGY PFAM DOMAINS KEGG KOG								
Alternaria brassicicola Baudoinia compniacensis UAMH 1 Cochliobolus heterostrophus C5 Dothistroma septosporum NZE10 Hysterium pulicare			SECONDARY METABOLISM CLUSTERS Refresh								
Genome			CAZYMES PEPTIDASES TRANSPORTERS TRANSCRIPTION FACTORS	NRPS	NRPS-Like	PKS	PKS-Like	TC	Total		
Alternaria brassicicola				4	6	5	3	5	26		
Baudoinia compniacensis UAMH 10762 (4089826) v1.0			?	?	2	5	2	1	13		

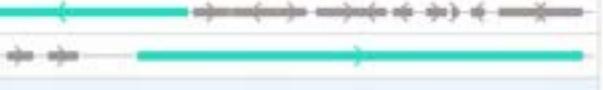
- 4) Scroll through the 'Genomes' list box and select both '*Septoria musiva*' and '*Septoria populincola*', and only those 2 species. Click the 'Refresh' button. Only the SM cluster core gene counts of the 2 *Septoria* spp. are shown, and may be directly compared. *S. musiva* has 2 hybrid core genes (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			Cluster Type								
Septoria musiva SO2202 v1.0 Septoria populincola v1.0			all DMAT HYBRID NRPS NRPS-Like	Refresh							
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	38	4	5	59	

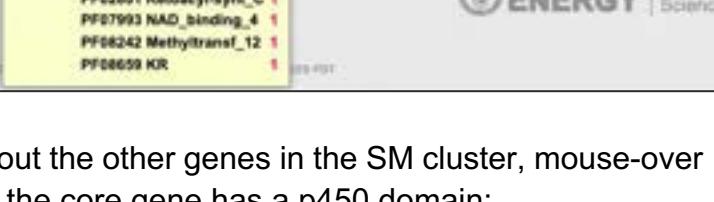
- 5) There are a total of 2 genes in the Hybrid column. Click on the number to show a graphical representation of the 2 gene clusters in *S. musiva*. The 'Size' column displays each cluster's length, and the 'Genes' column displays each cluster's core PKS-NRPS gene (in color) and its accessory, decorator, and other genes (in gray). A core hybrid gene is typically very large, but the total cluster size can be highly variable. To resize the 2 clusters to scale to each other, go to the 'Scale' pull-down menu, select 'Across All Clusters', and click on the 'Refresh' button:

Secondary Metabolism Clusters • Septoria musiva SO2202 v1.0

SEARCH	BLAST	BROWSE	ANNOTATIONS	MCL CLUSTERS	SYNTENY	DOWNLOAD	INFO	HOME	STATUS	HELP!	
Genomes			Cluster Type	Scale	Clusters Per Page						
Septoria musiva SO2202 v1.0			all DMAT HYBRID NRPS NRPS-Like	✓ Per Cluster Per Cluster No Gaps Across All Clusters	Refresh						
Total 2 cluster(s) found. 1											
Cluster ID	Cluster Type	Scaffold	Size (bp)	Genes							
Septmu1.24	HYBRID	scaffold_6_1522811-1553890	31,179								
Septmu1.25	HYBRID	scaffold_6_1977373-2004431	27,058								
Cluster N	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
Sepmut_24	HYBRID	scaffold_6:1522811-1553990	31,179	
Sepmut_25	HYBRID	scaffold_6:1977373-2004431	27,058	
Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes

Contact Us | GtoUs | Accessibility | Selection 508
 Disclaimer | Credits
 © 1997-2023 The Regents of the University of California.
 Biowulf-Portal version 170 (https://biowulf.tugrul.org). Release Date: 11-Aug-2023



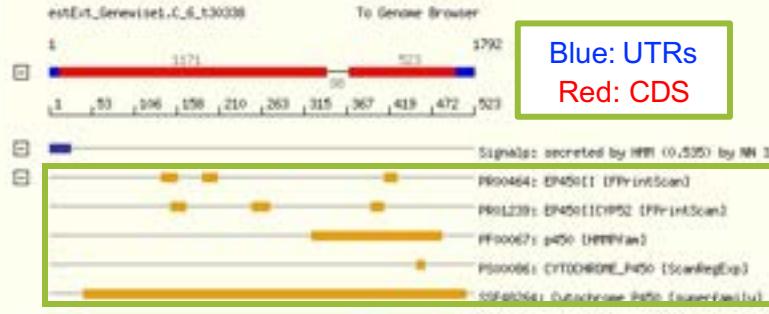
- 7) To get domain information about the other genes in the SM cluster, mouse-over them too. The next gene 3' to the core gene has a p450 domain:

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

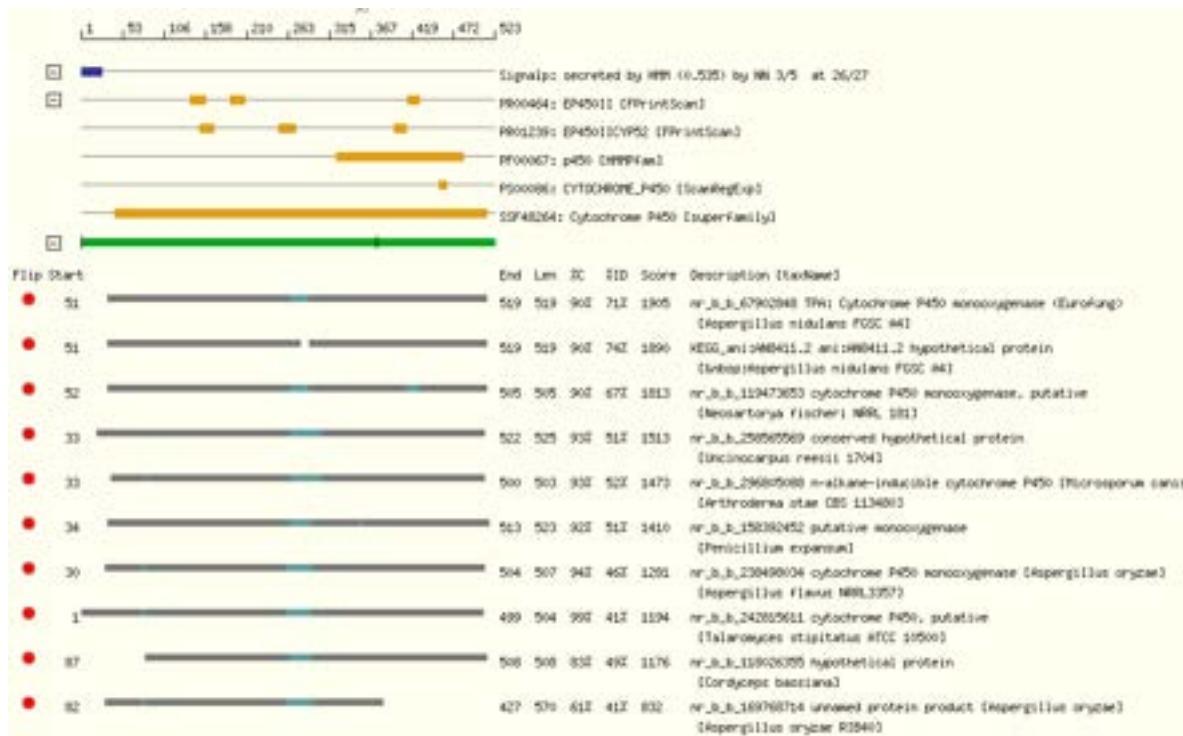
8) To get more detailed information about a gene, click on it directly. Click on the gene with the p450 domain to see its 'protein page'. Examination of the protein page reveals that:

- 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).

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:											
Best Hit:	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 pep CBP60485.1 TPA: Cytochrome P450 monooxygenase (Eurofung) [Aspergillus nidulans FGSC A4] (model%: 91, hit%: 90, score: 1905, %id: 71) [Aspergillus nidulans FGSC A4]										
total hits(shown)	683 (10)										
ASPECT	GO Id	GO Desc			Interpro Id	Interpro Desc					
Molecular Function	0016712	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					
	0004497	monooxygenase activity			IPR002402	Cytochrome P450, E-class, group II					
	0020037	heme binding			IPR001128	Cytochrome P450					
	0005536	iron ion binding			IPR002402	Cytochrome P450, E-class, group II					
Biological Process	0006118	electron transport			IPR002974	Cytochrome P450, E-class, CYP52					
					IPR002402	Cytochrome P450, E-class, group II					
					IPR001128	Cytochrome P450					
KOG GROUP Metabolism	KOG Id KOG0158	KOG Class Secondary metabolites biosynthesis, transport and catabolism			KOG Desc 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											
<p>estExt_Genewise1.C_6.t30338 To Genome Browser</p>  <p>Blue: UTRs Red: CDS</p>											
<p>Signal peptide: secreted by HMM (0.535) by MN 1/5 - 46 26/27</p> <p>InterPro annotations (For example, Pfam domains)</p>											



- 9) Based on the annotations and top hits, it seems that this gene is indeed a cytochrome p450 monooxygenase, a class of enzymes that often modify core structures of SM biosynthetic pathways. Similar perusal of the other genes of the cluster says that this cluster is an excellent candidate for synthesis of your SM.



- 10) One explanation for *S. musiva* having this cluster and the congeneric *S. populica* not is that the former acquired the cluster by horizontal gene transfer from a phylogenetically distant source. The ‘best Blast hit’ of the cytochrome p450 enzyme supports this hypothesis. To see if the core enzyme can shed some light, click the web browser back button to go back to the SM CLUSTERS graphic, and click on the same PKS-NRPS core gene we moused over earlier. The protein page is rich in details, including domains and the top 10 hits. All of the hits are high quality and are from Eurotiomycetes. This cluster is an excellent candidate for horizontal gene transfer from the Eurotiomycetes!

References:

- Dhillon B, Feau N, Aerts AL, Beauseigle S, Bernier L, Copeland A, Foster A, Gill N, Henrissat B, Herath P, LaButti KM, Levasseur A, Lindquist EA, Majoor E, Ohm RA, Pangilinan JL, Pribowo A, Saddler JN, Sakalidis ML, de Vries RP, Grigoriev IV, Goodwin SB, Tanguay P, Hamelin RC. Horizontal gene transfer and gene dosage drives adaptation to wood colonization in a tree pathogen. Proc Natl Acad Sci U S A. 2015 Mar 17;112(11):3451-6. doi: 10.1073/pnas.1424293112. Epub 2015 Mar 2. PubMed PMID: 25733908
- Schümann J, Hertweck C. Molecular basis of cytochalasan biosynthesis in fungi: gene cluster analysis and evidence for the involvement of a PKS-NRPS hybrid synthase by RNA silencing. J Am Chem Soc. 2007 Aug 8;129(31):9564-5. Epub 2007 Jul 18. PubMed PMID: 17636916.

FungiDB: Secondary Metabolites and clusters

Learning objectives:

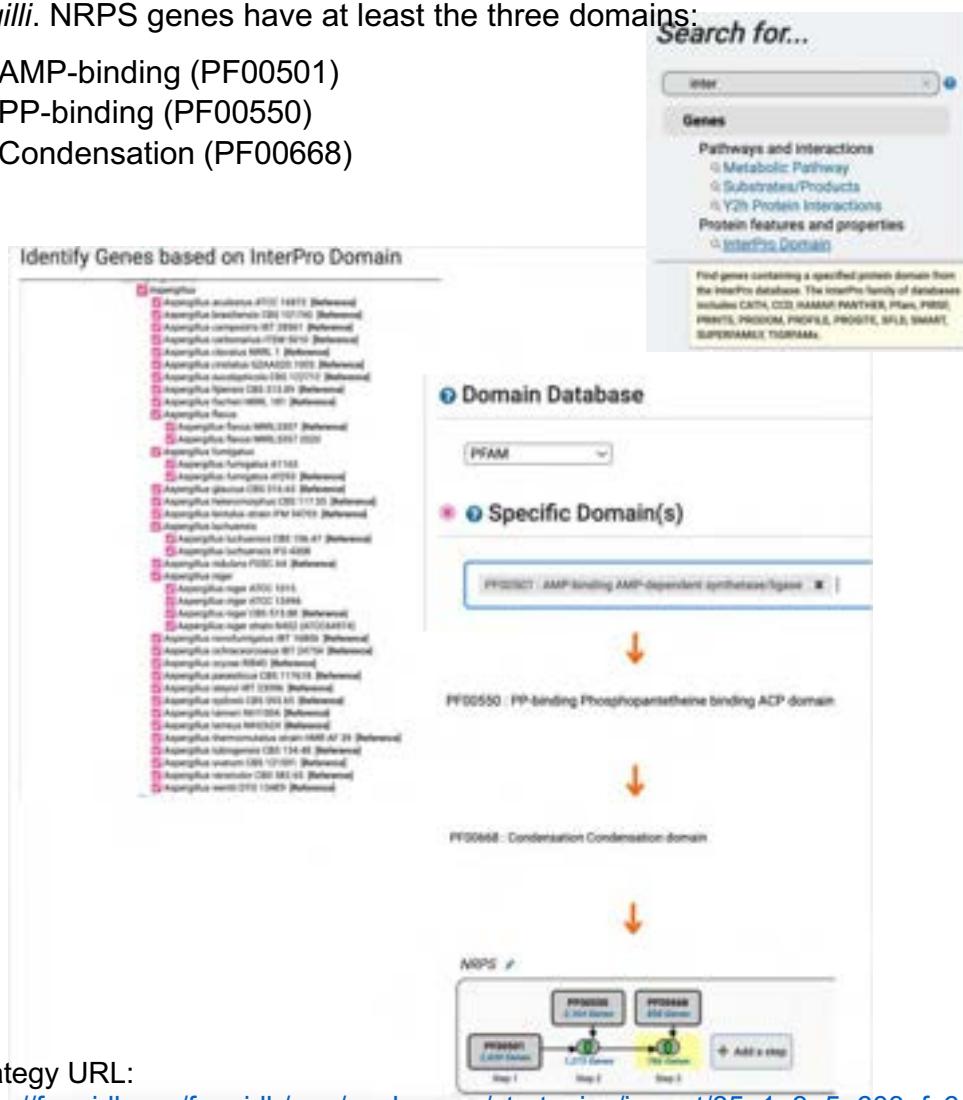
- Explore InterPro search in FungiDB
- Cross-reference the results with MycoCosm data
- **Finding secondary metabolites and gene clusters**

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)



The screenshot shows the FungiDB search interface for InterPro domains. In the top right, a search bar is set to 'Inter' and 'Genes'. Below it, a sidebar lists various databases: Pathways and Interactions, Metabolic Pathway, Substrates/Products, Y2h Protein Interactions, Protein features and properties, and InterPro Domains (which is selected). The main area is titled 'Identify Genes based on InterPro Domain' and shows a list of Aspergillus species. A specific domain search for 'PF00501 - AMP-binding AMP-dependent synthetase/ligase' is highlighted. Three arrows point down from this search result to three separate boxes: 'PF00550 - PP-binding Phosphopantetheine binding ACP domain', 'PF00668 - Condensation Condensation domain', and a diagram of an NRPS pathway.

Search for...

Inter
Genes

Pathways and Interactions

Metabolic Pathway

Substrates/Products

Y2h Protein Interactions

Protein features and properties

InterPro Domains

Identify Genes based on InterPro Domain

PF00501 - AMP-binding AMP-dependent synthetase/ligase

PF00550 - PP-binding Phosphopantetheine binding ACP domain

PF00668 - Condensation Condensation domain

NRPS #

Step 1: PF00501 (AMP-binding AMP-dependent synthetase/ligase)
Step 2: PF00550 (PP-binding Phosphopantetheine binding ACP domain)
Step 3: PF00668 (Condensation Condensation domain)

Add a step

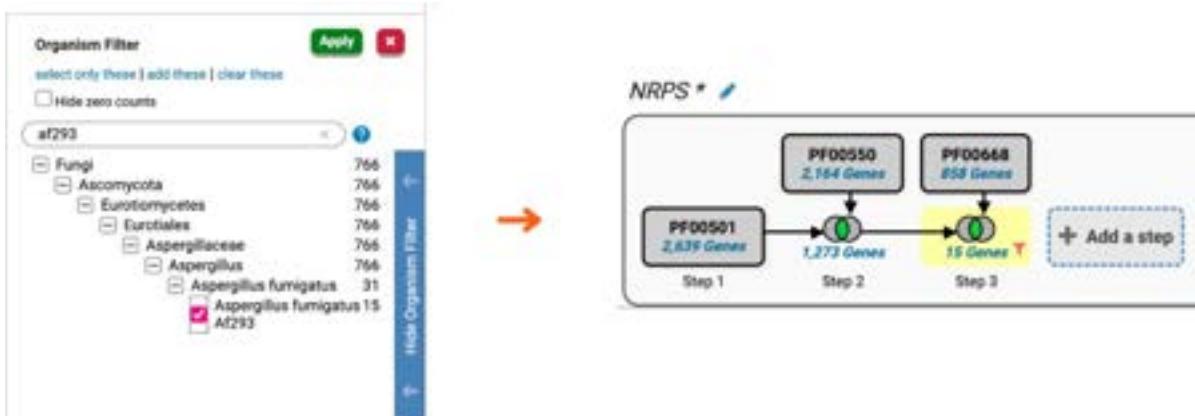
Strategy URL:

<https://fungidb.org/fungidb/app/workspace/strategies/import/85a1e3a5a603efc6>

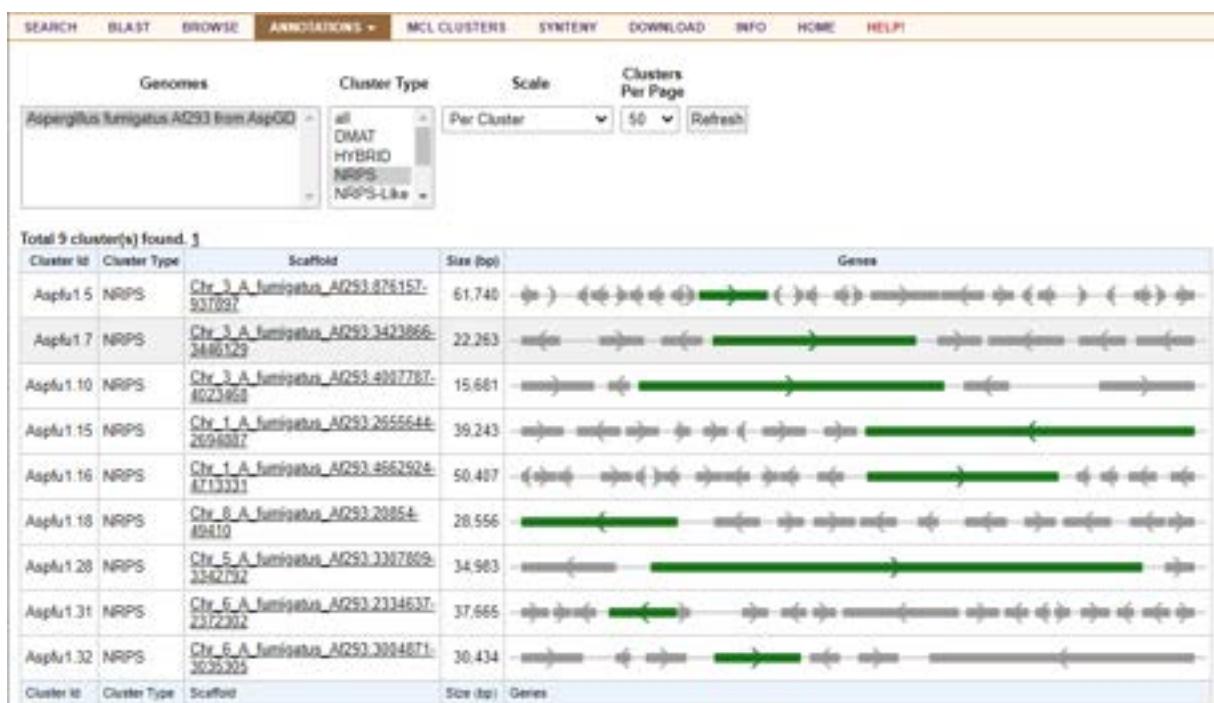


- How many genes were identified in *Aspergillus fumigatus* Af293?

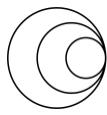
Hint: use the organism filter on the left to limit your search results to Af293 genes only.



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

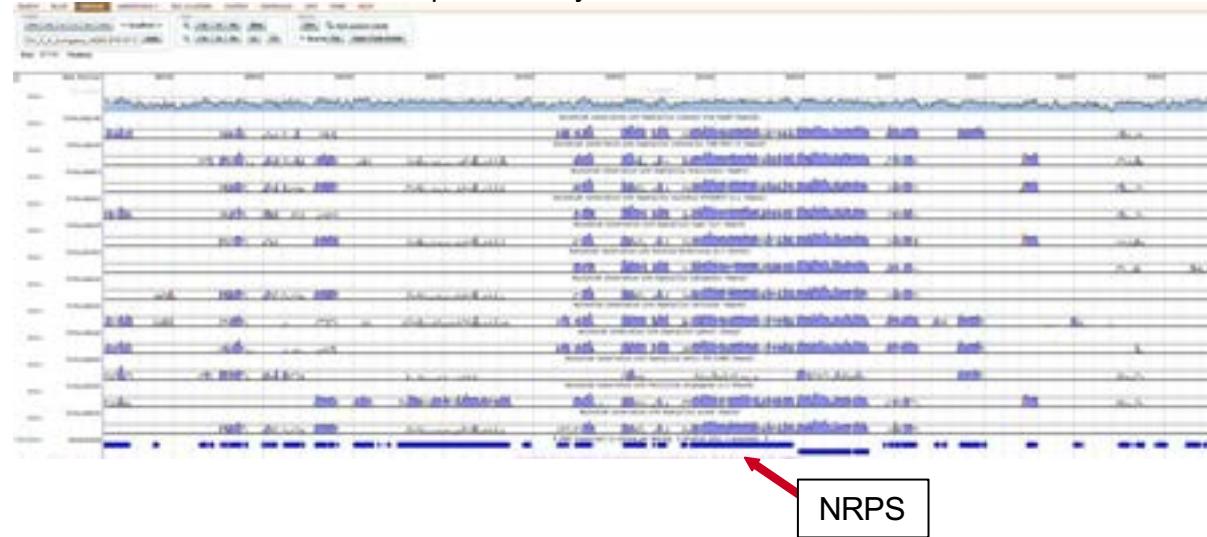


- What do you think may be causing the difference in the predicted gene number?
- This view on MycoCosm allows you to analyze backbone and auxiliary proteins across the entire predicted secondary metabolism cluster. How conserved are these secondary metabolite clusters across related Aspergilli? Click on the scaffold coordinates for Aspfu1.5 and analyze the Vista curve tracks in the genome browser. How many related Aspergilli show some synteny with this region? Repeat this exercise for the next cluster, Aspfu1.7.
 - Answer: Synteny is observed across most Aspergilli for Aspfu1.5, raising the possibility that this SM cluster is widespread across the genus. However,



Aspfu1.7 shows no synteny except for at a couple auxiliary genes in *Aspergillus wentii*, suggesting that it is possibly lineage specific.

Genome browser at locus for Aspfu1.5 biosynthetic cluster:



Genome browser at locus for Aspfu1.7 biosynthetic

