

Anatomy of the main page.

The Eukaryotic Pathogen, Vector and Host Informatics Resources (<https://VEuPathDB.org>) are comprised of a family of bioinformatics resources including an integrated functional genomics database for fungi and oomycetes - FungiDB. FungiDB (<https://FungiDB.org>) is a free online resource for data-mining and functional genomics analysis. It provides an easy-to-use, interactive interface to explore genomes, annotation, functional data (transcriptomics or proteomics), metabolic pathways and results from numerous genome-wide analyses (ie. InterPro scan, signal peptide and transmembrane domain predictions, orthology, etc.). FungiDB contains an expanding number of genomes from species spanning the Oomycetes and Fungi groups including but not limited to plant, animal, and human pathogens.

The modules presented here are designed to introduce you to FungiDB resources and teach you how to construct basic and complex search strategies (*in silico* experiments). Navigate to <https://fungidb.org/> and examine the organisation of the main landing page. Try to set up a few searches to learn your way around the interface.

The image shows a screenshot of the FungiDB website interface with several callout boxes providing detailed information about its features:

- Header section:** database version and other useful links
- Quick search** for a single Gene ID
- Quick search** for a text term looks across the following fields in all genomes:
 - Enzyme commission (EC) description
 - Gene ID, name, product, description and notes
 - Gene ontology (GO) terms and definitions
- Main Menu (in grey)** provides access to the Search for Genes, Search for Other Data Types, and Tools panels when you navigate away from the main page.
- Left panel** features news, tweets, educational resources, workshop schedules and more
- Useful tools, also accessible from the Main menu above**
- Searches deployed from this section will generate a list of genes**
- Searching Other Types of Data will generate a list of none-gene entities (e.g. SNPs, Metabolic Pathways, etc.)**

The screenshot shows the FungiDB logo, navigation menu (Home, New Search, My Strategies, My Basket, Tools, Data Summary, Downloads, Community, Analyze My Experiment, My Favorites), search bars (Gene ID, Gene Text Search), and the main menu items: Data Summary, News and Tweets, Community Resources, Search for Genes, Search for Other Data Types, and Tools. The search panels show expand/collapse options and search input fields. The Tools panel lists various services like BLAST, Results Analysis, Sequence Retrieval, Companion, EUPaGDT, PubMed and Entrez, Genome Browser, and Searches via Web Services.

Building linear and nested search strategies.

Learning objectives:

1. Look for homozygous SNPs between two groups of *Aspergillus fumigatus* isolates (triazole sensitive and triazole resistant) to identify SNPs that may be important for drug resistance.
2. Map SNPs to *A. fumigatus* Af293 genes.
3. Identify genes up-regulated in drug-resistant strains A1160, HapB and 29.9 grown with itraconazole using the integrated RNA-Seq evidence.
4. Create a nested search for genes that have a predicted signal peptide OR a transmembrane domain OR both.

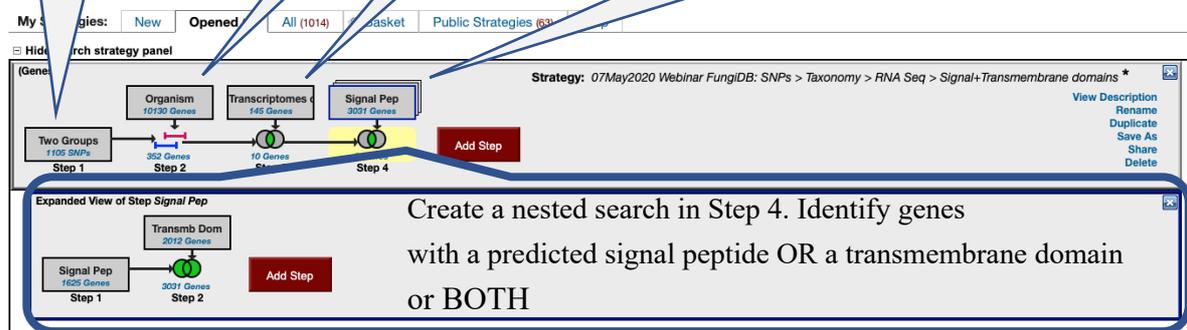
Strategy overview

Step 1. Find SNPs.
Search for Other Data
Types > Differences
between two groups
of isolates

Step 2.
Identify genes
harbouring
SNPs from
Step 1

Step 3. Find
genes that are up-
regulated in drug-
resistant strains
grown with
itraconazole.

Step 4. Nested
search: combining
linear and non-
linear searches.



Genomic collocation operator. Maps SNPs from Step 1 to individual genes in *A. fumigatus* Af293 (Step 2). This operator will be the only search operator available when you create searches across different types of data (e.g. list of SNPs vs gene lists).



Intersect operator used in direct comparison (e.g. genes from Step2 that also fit search criteria in Step 3, etc.)



Intersect operator used in direct comparison (e.g. genes that have one characteristic OR the other OR BOTH)

- a. Click on the SNPs menu and then select the Differences Between Two Groups of Isolates link to deploy a search.

Search for Other Data Types

expand all | collapse all

Find a search... ?

- ▶ Popset Isolate Sequences
- ▶ Genomic Sequences
- ▶ Genomic Segments
- ▼ SNPs
 - SNP ID(s)
 - Genomic Location
 - Differences Within a Group of Isolates
 - Differences Between Two Groups of Isolates
 - Find SNPs that distinguish between two groups of isolates based on the user supplied major allele threshold for each group.
- ▶ ES
- ▶ Of
- ▶ Metabolic Pathways
- ▶ Compounds

Tools

BLAST
Identify Sequence Similarities

Results Analysis
Analyze Your Strategy Results

Sequence Retrieval
Retrieve Specific Sequences using IDs and coordinates

Companion
Annotate your sequence and determine orthology, phylogeny & synteny

EuPaGDT
Prokaryotic Pathogen CRISPR guide
DNA Design Tool

Pubmed and Entrez
Fetch the Latest Pubmed and Entrez Results

- b. From the Organism drop down menu select *Aspergillus fumigatus* Af293 to bring up available datasets.
- c. Select the dataset titled “Genomic Context of Azole-Resistance Mutations in *Aspergillus fumigatus*”.

Identify SNPs based on Differences Between Two Groups of Isolates

Organism
Aspergillus fumigatus Af293

Set A Isolates
65 Set A Isolates Total
24 of 65 Set A Isolates selected [data set](#)

expand all | collapse all

Find a variable

- ▣ data set
- ▣ Collection year
- ▣ Sample type
- ▶ Sample source
- ▶ Geographic location
- ▶ Organism under investigation
- ▶ DNA sequencing

data set
A data item that is an aggregate of other data items of the same type that have something in common. Averages and distributions can be determined for data sets.

Keep checked values at top

65 (100%) of 65 Set A Isolates have data for this variable

<input type="checkbox"/> data set	Remaining Set A Isolates	Set A Isolates	Distribution	%
<input type="checkbox"/> Aligned genome sequence reads - <i>A. fumigatus</i> isolates	22 (34%)	22 (34%)	<div style="width: 34%;"></div>	(100%)
<input type="checkbox"/> Aligned SNPs - <i>A. fumigatus</i> Af1163 strain	1 (2%)	1 (2%)	<div style="width: 2%;"></div>	(100%)
<input type="checkbox"/> Aspergillus fumigatus Af293 Genome Sequence and Annotation	1 (2%)	1 (2%)	<div style="width: 2%;"></div>	(100%)
<input checked="" type="checkbox"/> Genomic Context of Azole-Resistance Mutations in <i>Aspergillus fumigatus</i>	24 (37%)	24 (37%)	<div style="width: 37%;"></div>	(100%)
<input type="checkbox"/> SNP calls on strains isolated from patients with PA and CNPA	17 (26%)	17 (26%)	<div style="width: 26%;"></div>	(100%)

- d. Define Set A isolates by expanding the “Organism under investigation” section on the left. Click on the “Triazoles” and select “Sensitive”.

Set A Isolates

65 Set A Isolates Total 7 of 65 Set A Isolates selected data set x Triazoles x

expand all | collapse all

Find a variable

Triazoles

Keep checked values at top **24 (37%) of 65 Set A Isolates have data for this**

	Remaining Set A Isolates	Set A Isolates	Distribution
<input type="checkbox"/> Resistant	24 (100%)	24 (100%)	
<input checked="" type="checkbox"/> Sensitive	17 (71%)	7 (29%)	

- e. Define Set B isolates, except this time choose the Resistant group.
 f. Define SNP search stringency by setting thresholds for the following parameters:

Read frequency threshold >=

This parameter defines a stringency for data supporting a SNP call between an isolate in a group (Set A or Set B) and the reference genome (*A. fumigatus* Af293). Each nucleotide position of each isolate is compared to the reference genome. A SNP call is made if the portion of the isolate's aligned reads that support the SNP is above the Read Frequency Threshold. Select 80% to find high quality haploid SNPs. For heterozygous diploid/aneuploid SNPs select 40%.

To find high quality haploid SNP in triazole sensitive and resistant isolate groups (Set A and Set B, respectively), set this parameter to 80% for both groups.

Major allele frequency >=

The major allele frequency is the frequency of the most common SNP across the isolates in a Set. The default setting for this parameter is 80%. SNPs returned by the two groups search will have a different major allele call between Set A and Set B. NOTE: 100% is permissible and the most stringent since we are first identifying a SNPs from this set and then comparing it with the allele SNP in set B.

In this exercise, leave the parameter selection at default (80%) for Set A and Set B isolates.

Percent isolates with base call >=

Percent isolates with a base call defines the fraction of the selected isolates that must have a base call before a SNP is returned for that nucleotide position based on the remaining isolates that do have data.

Set this parameter to 80% for Set A and Set B isolates.

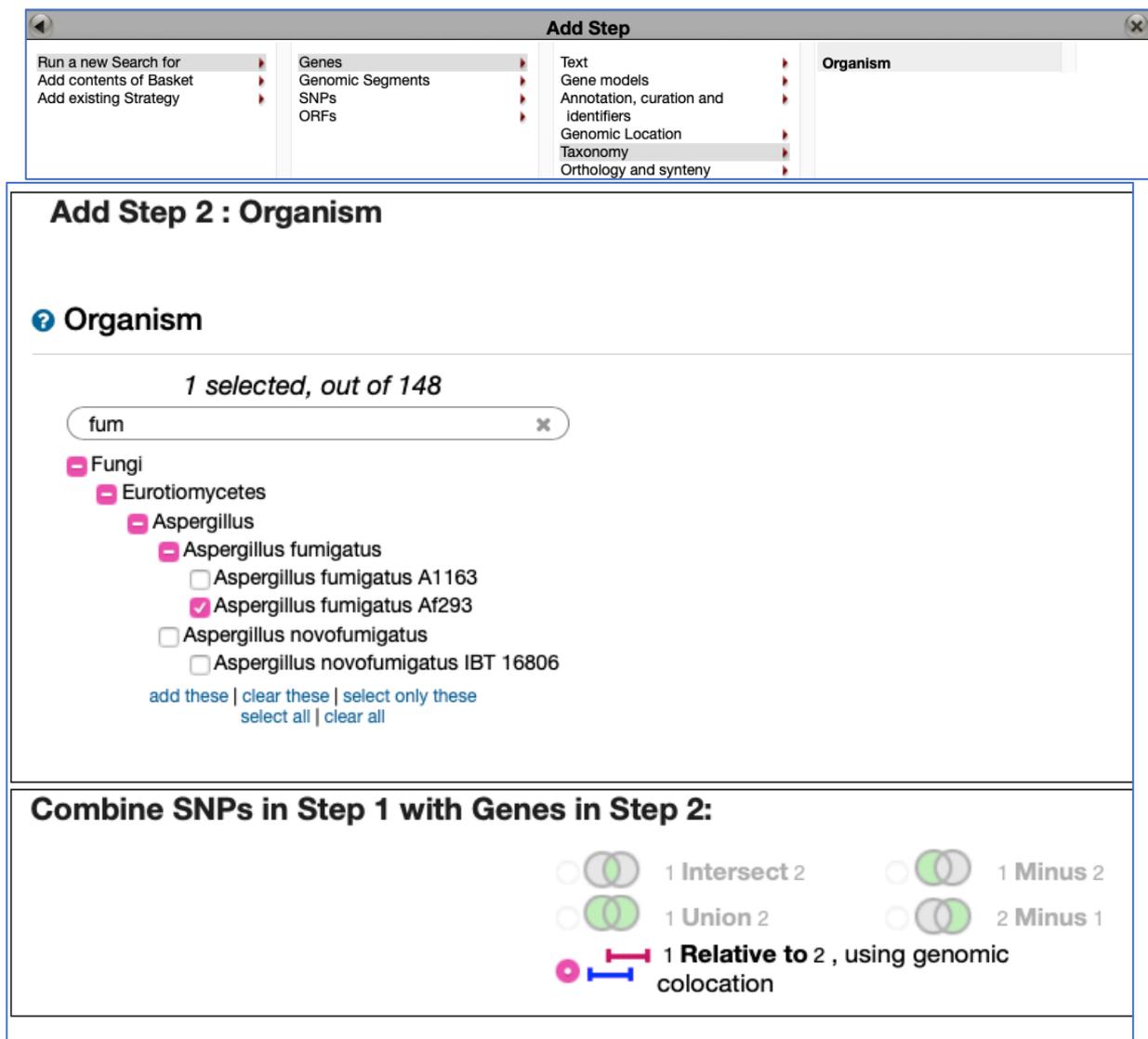
g. Deploy the search by clicking on the “Get Answer” button

Get Answer



2. Determine SNPs that map to *A. fumigatus* Af293 genes.

a. Click on Add Step and then Run a new Search for Genes. Select Taxonomy and then organism for identify genes in *Aspergillus fumigatus* Af293



Add Step

Run a new Search for
Add contents of Basket
Add existing Strategy

Genes
Genomic Segments
SNPs
ORFs

Text
Gene models
Annotation, curation and identifiers
Genomic Location
Taxonomy
Orthology and synteny

Organism

Add Step 2 : Organism

Organism

1 selected, out of 148

fum

Fungi
Eurotiomycetes
Aspergillus
Aspergillus fumigatus
Aspergillus fumigatus A1163
 Aspergillus fumigatus Af293
Aspergillus novofumigatus
Aspergillus novofumigatus IBT 16806

[add these](#) | [clear these](#) | [select only these](#)
[select all](#) | [clear all](#)

Combine SNPs in Step 1 with Genes in Step 2:

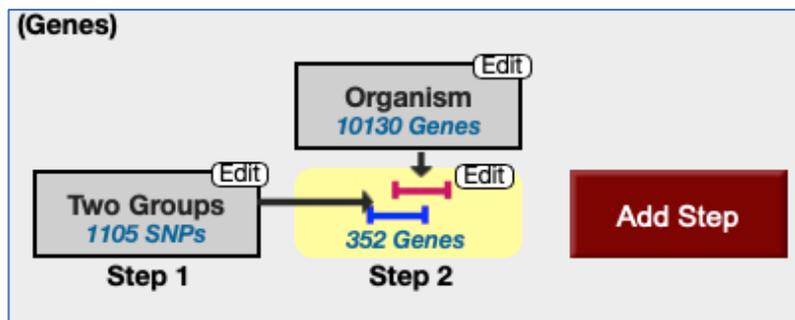
1 Intersect 2 1 Minus 2
 1 Union 2 2 Minus 1
 1 Relative to 2 , using genomic colocation

b. Notice that only one search operator is available. This is because you are comparing SNPs, which is a non-gene list, to a list of *A. fumigatus* Af293 genes.

- c. Click on the Continue.... button to proceed to a search parameter selection window.
- d. Define parameters to return each Gene from Step 2 that overlaps with a SNP or multiple SNPs identified in Step 1.

Continue....

"Return each **Gene from Step 2** whose **exact region** overlaps the **exact region** of a SNP in Step 1 and is on **either strand**"



The search you deployed mapped the location of 1105 SNPs to the genome of *A. fumigatus* and identified genes that harbor SNPs within ORFs.

3. Identify genes up-regulated in drug-resistant isolates using RNA-Seq evidence.

- a. Click Add Step to Run a new Search for Genes using RNA Seq Evidence

- b. Filter datasets for fumigatus ('fum') and select the dataset titled "Transcriptomes of itraconazole-resistant strains (Bowyer 2016)"

- c. Select search parameters, reference and comparison samples.

Look for up-regulated genes (at least 3 fold).

Reference Samples: Three strains grown in the absence of the itraconazole (-Itra).

Comparison Samples: The same strains grown with itraconazole (+Itra)

Add Step 3 : A. fumigatus Af293 Transcriptomes of itraconazole-resistant strains RNASeq (fold change)

For the Experiment **Transcriptomes of Itraconazole-resistant strains unstranded**

return **protein coding** Genes
 that are **up-regulated**
 with a **Fold change >= 3**

between each gene's **average** expression value
 (or a **Floor of 10 reads (13 FPKM)**)

In the following **Reference Samples**

- A1160-itra
- A1160+itra
- HapB-itra
- HapB+itra
- 29.9-itra

and its **average** expression value
 (or the **Floor** selected above)

In the following **Comparison Samples**

- A1160-itra
- HapB-itra
- HapB+itra
- 29.9-itra
- 29.9+itra

Example showing one gene that would meet search criteria
 (Dots represent this gene's expression values for selected samples)

Up-regulated

You are searching for genes that are **up-regulated** between at least two reference samples and at least two comparison samples.

For each gene, the search calculates:

$$\text{fold change} = \frac{\text{average expression level in comparison}}{\text{average expression level in reference}}$$

and returns genes when **fold change >= 3**.

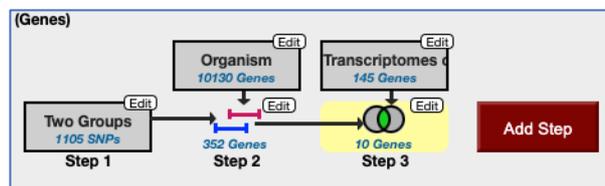
To narrow the window, use the maximum reference value, or minimum comparison value. To broaden the window, use the minimum reference value, or maximum comparison value.

See the detailed help for this search.

* or FPKM Floor, whichever is greater

Combine Genes in Step 2 with Genes in Step 3:

- 2 Intersect 3
- 2 Union 3
- 2 Minus 3
- 3 Minus 2
- 2 Relative to 3, using genomic colocation



4. Create a nested search for genes that have a predicted signal peptide OR a transmembrane domain OR both.

Up until now we have been creating linear, non-nested searches where a single operator (colocation or intersect) combined data types from two search steps (e.g. Step 2 and Step 1).

In the next search, we will use a nested strategy approach to:

- access two data types (signal peptide and transmembrane domain predictions)
- find genes that (1) have a predicted signal peptide or (2) have a transmembrane domain, or (3) have both
- intersect signal peptide and transmembrane domain predictions results with Step 3.

- Click Add Step to Run a new Search for Genes in Protein targeting and localization, Predicted Signal peptide data

Add Step

Run a new Search for	Genes	Text	Predicted Signal Peptide
Transform by Orthology	Genomic Segments	Gene models	Transmembrane Domain Count
Add contents of Basket	SNPs	Annotation, curation and identifiers	
Add existing Strategy	ORFs	Genomic Location	
Filter by assigned Weight		Taxonomy	
Transform to Pathways		Orthology and synteny	

Add Step 4 : Predicted Signal Peptide

Organism

1 selected, out of 148

fum

- Fungi
 - Eurotiomycetes
 - Aspergillus
 - Aspergillus fumigatus
 - Aspergillus fumigatus A1163
 - Aspergillus fumigatus Af293
 - Aspergillus novofumigatus
 - Aspergillus novofumigatus IBT 16806

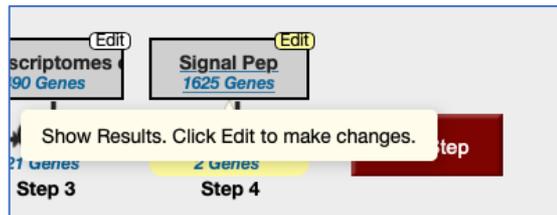
[add these](#) | [clear these](#) | [select only these](#)
[select all](#) | [clear all](#)

► **Advanced Parameters**

Combine Genes in Step 3 with Genes in Step 4:

 3 **Intersect** 4
  3 **Minus** 4
  3 **Union** 4
  4 **Minus** 3
  3 **Relative to** 4, using genomic colocation

b. Hover over the step, click Edit, and select Make Nested Strategy



Rename | View | Analyze | Revise | **Make Nested Strategy** | Insert Step Before | Orthologs | Delete

STEP 4 : Sign Expand this step in a new panel to add nested steps. (Use this to build a non-linear strategy)

Organism : Aspergillus fumigatus Af293

Minimum SignalP-NN Conclusion Score : 3

Minimum SignalP-NN D-Score : 0.5

Minimum SignalP-HMM Signal Probability : 0.5

Matches any or all advanced parameters : any

Results: 1625 Genes

Give this search a weight

A new search window, highlighted in different color, will appear underneath of your current strategy (Expanded View of Step Signal Pep).

- c. Use the nested strategy window to create a non-linear search - click Add Step within the new window. Run a search for transmembrane domains and choose union operator to look for all combinations:

Add Step 2 : Transmembrane Domain Count

Organism

1 selected, out of 148

fum

- Fungi
- Eurotiomycetes
- Aspergillus
 - Aspergillus fumigatus
 - Aspergillus fumigatus A1163
 - Aspergillus fumigatus A1293
 - Aspergillus novofumigatus
 - Aspergillus novofumigatus IBT 16806

add these | clear these | select only these
select all | clear all

Minimum Number of Transmembrane Domains

Maximum Number of Transmembrane Domains

Combine Genes in Step 1 with Genes in Step 2:

1 Intersect 2
 1 Minus 2
 1 Union 2
 2 Minus 1
 1 Relative to 2, using genomic colocation



To close the expanded view, click on the X on the right. You can always review/modify this step by clicking on the Edit button, which is located at the top corner of each search step.