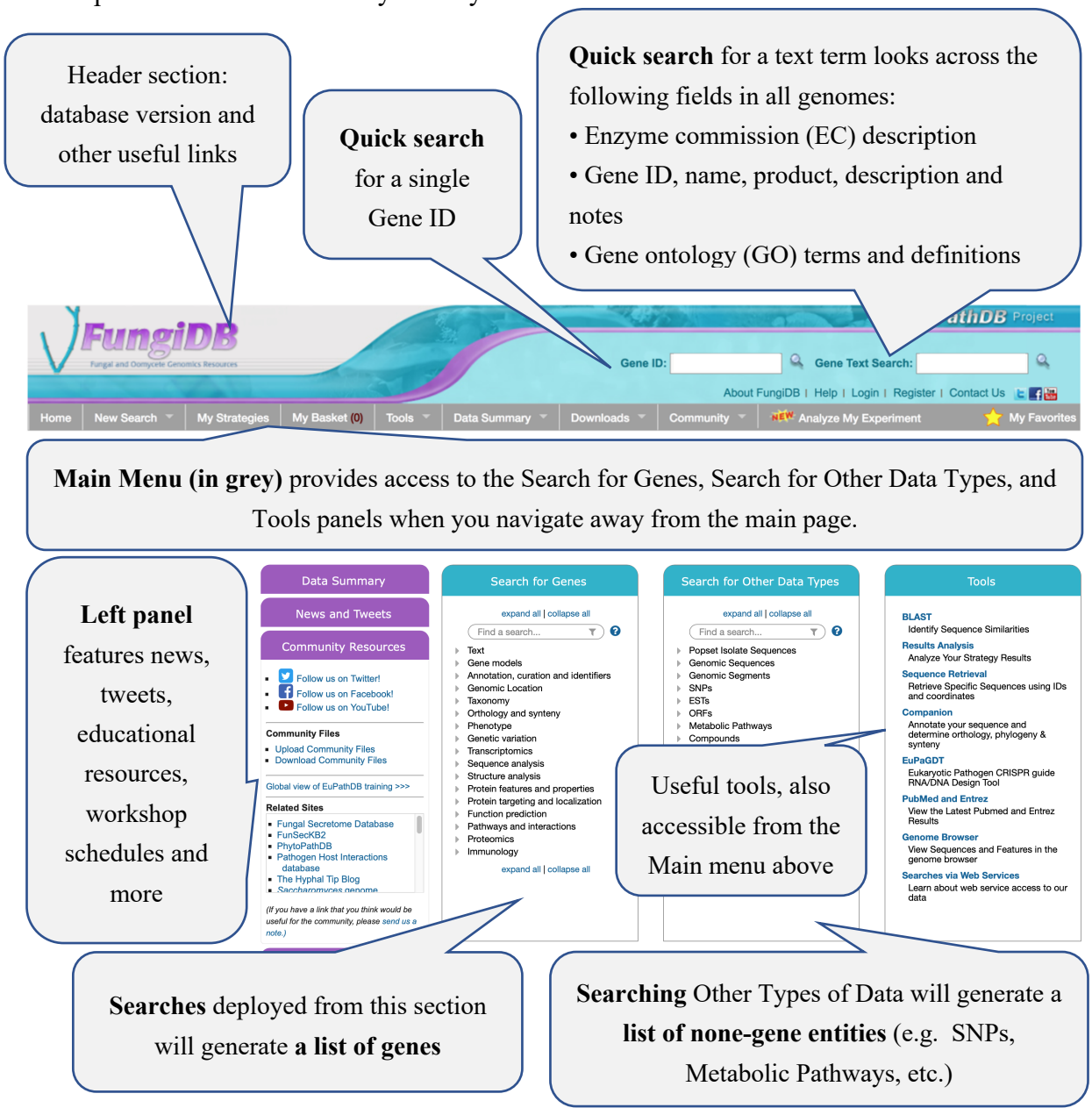
Anatomy of the main page.

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.

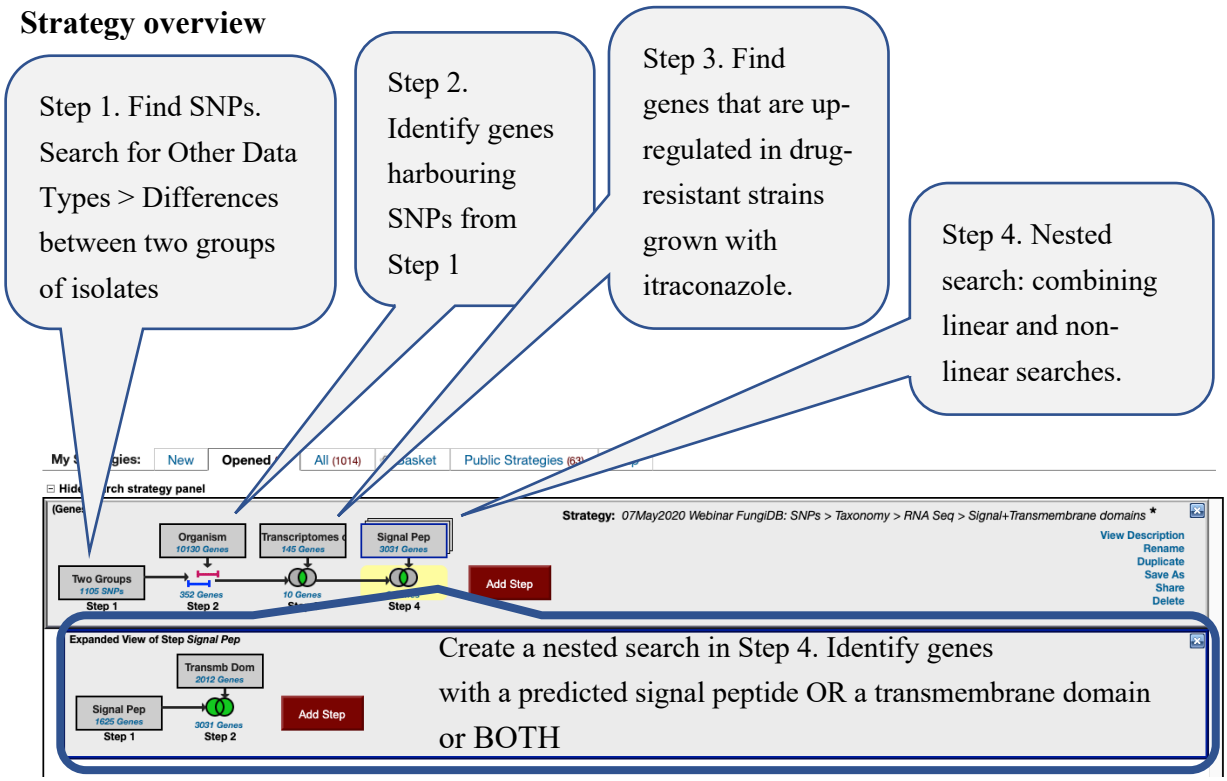


Building linear and nested search strategies.

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

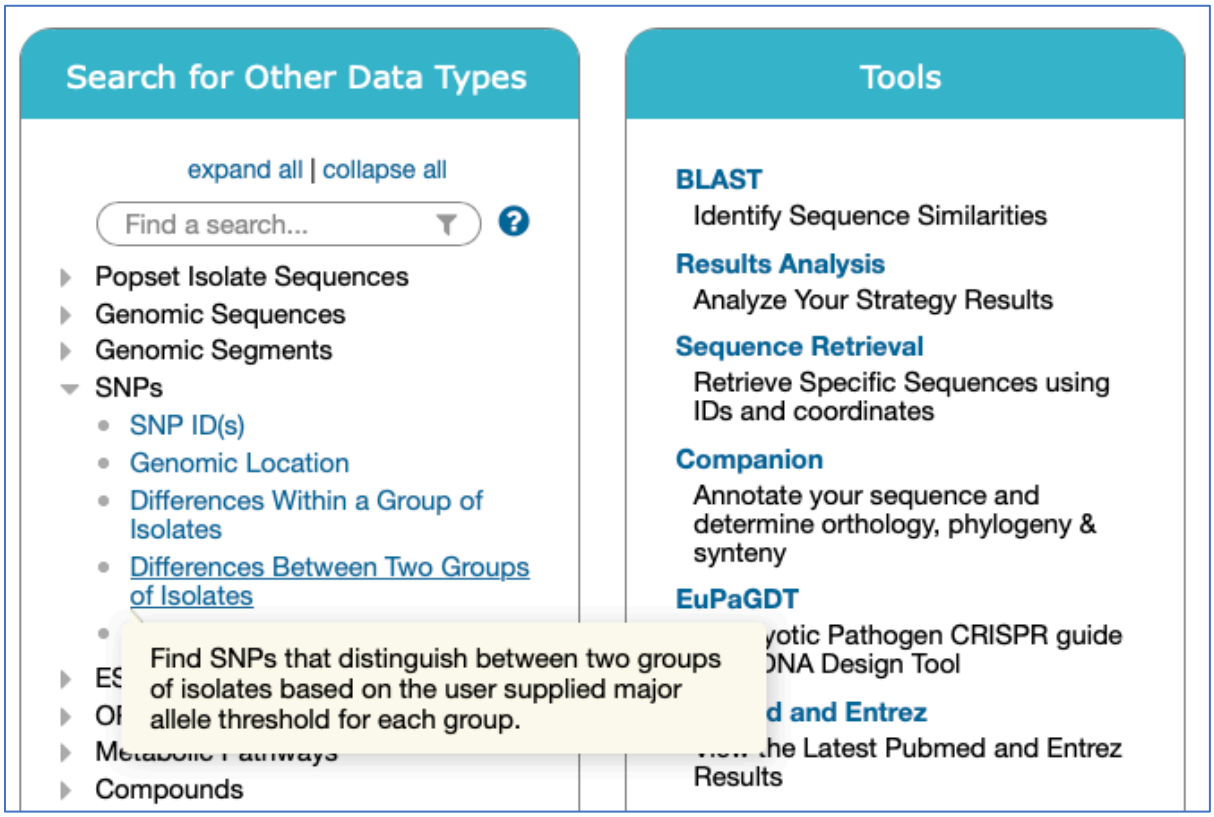


Genomic colocation operator. Maps SNPs from Step 1 to individual genes in *A. fumigatus* Af293 (Step 2). This operator will be the only search operator available when you create searches across different types of data (e.g. list of SNPs vs gene lists).

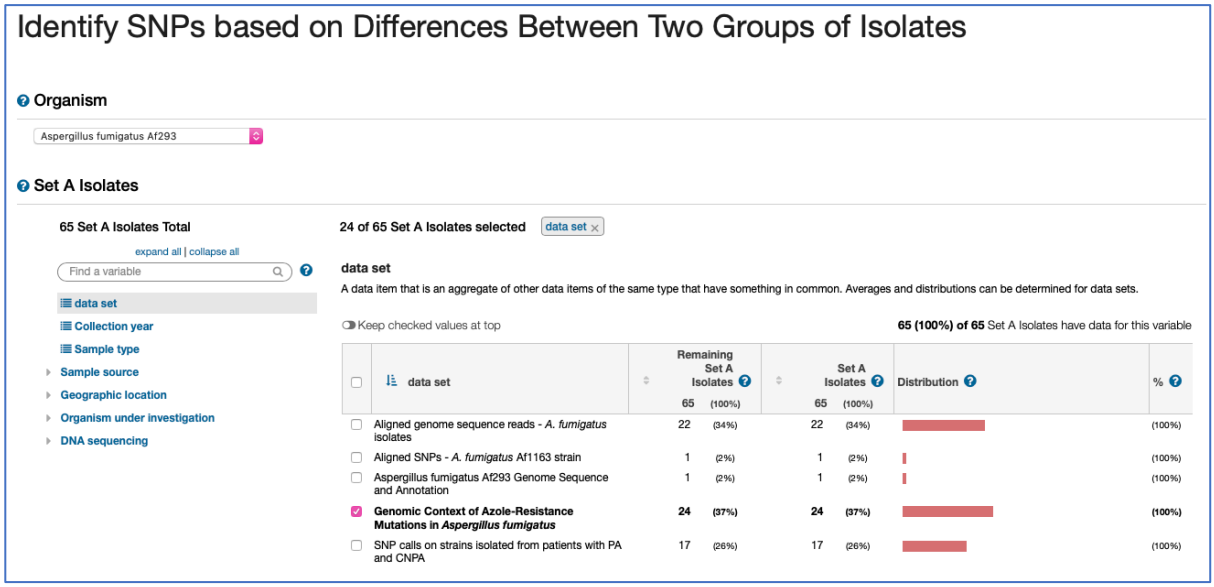
Intersect operator used in direct comparison (e.g. genes from Step2 that also fit search criteria in Step 3, etc.)

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

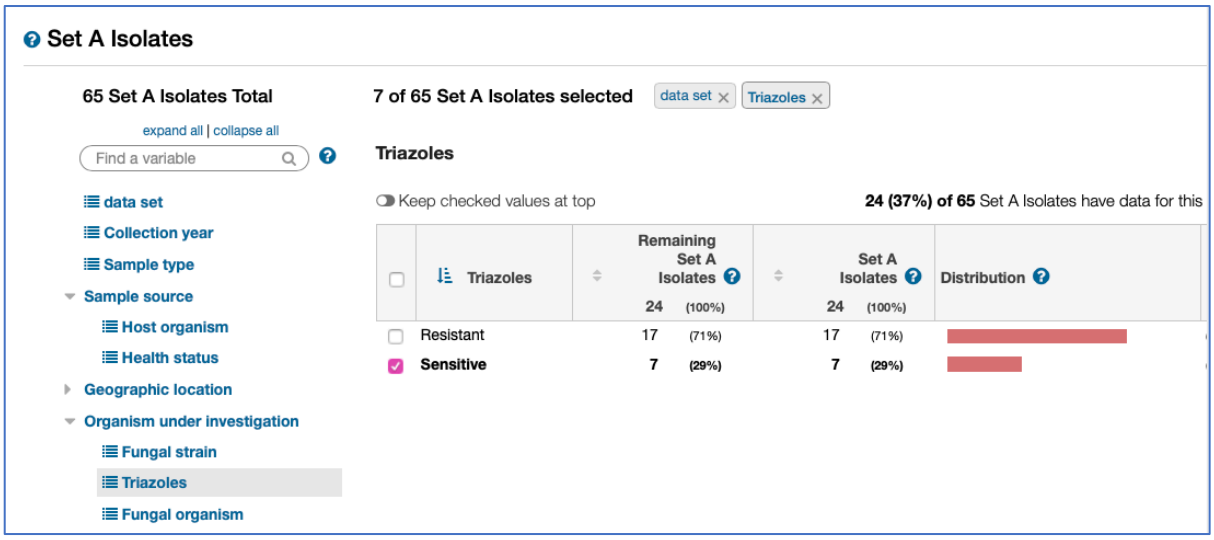
a. Click on the SNPs menu and then select the Differences Between Two Groups of Isolates link to deploy a search.



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*".



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



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%.

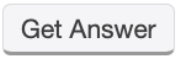
To find high quality haploid SNP in triazole sensitive and resistant isolate groups (Set A and Set B, respectively), set this parameter to 80% for both groups.

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.

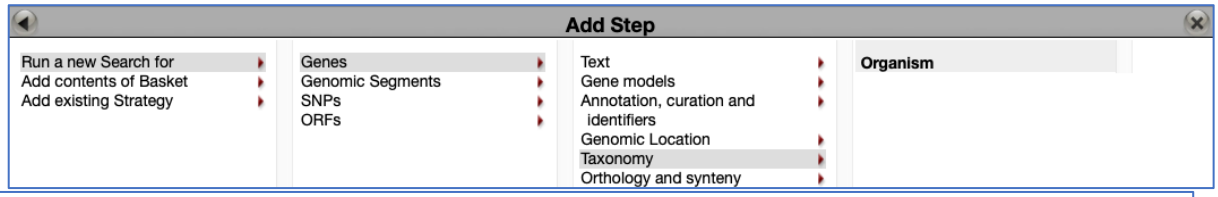
Percent isolates with base call >=
Percent isolates with a base call defines the fraction of the selected isolates that must have a base call before a SNP is returned for that nucleotide position based on the remaining isolates that do have data.
Set this parameter to 80% for Set A and Set B isolates.

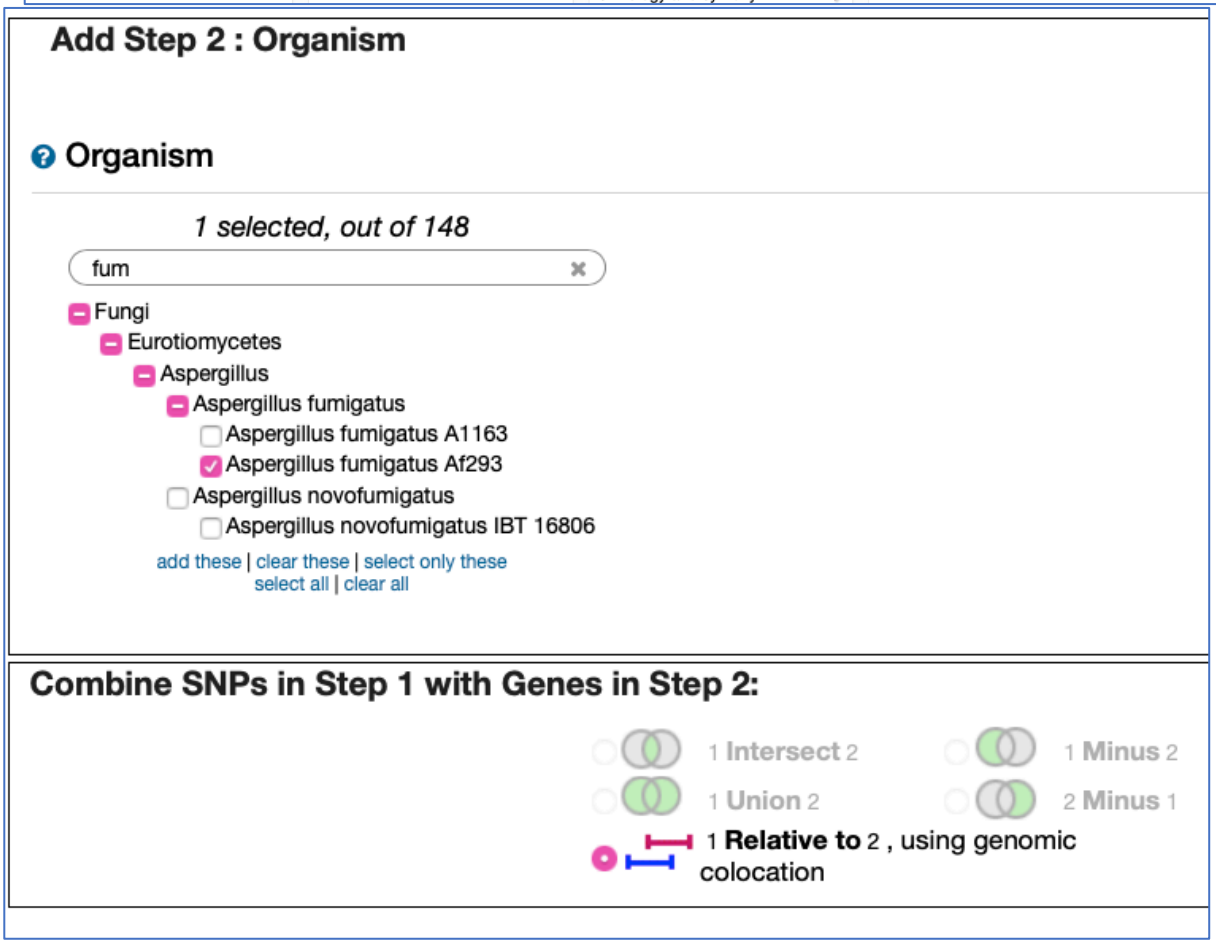
g. Deploy the search by clicking on the "Get Answer" button



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

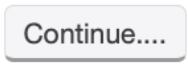
a. Click on Add Step and then Run a new Search for Genes. Select Taxonomy and then organism for identify genes in *Aspergillus fumigatus* Af293



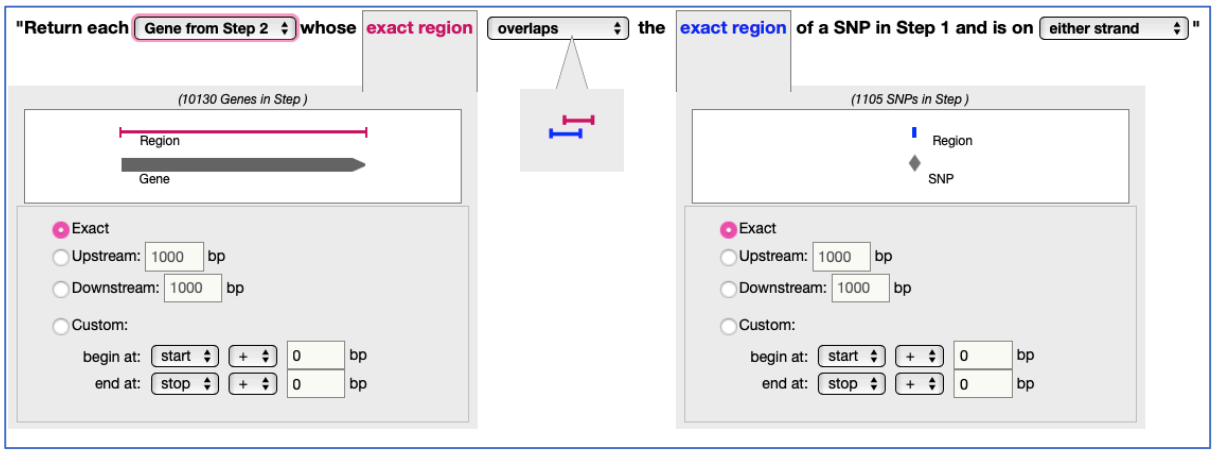


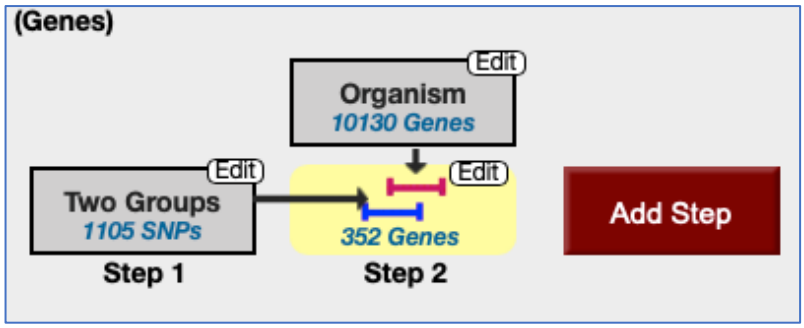
b. Notice that only one search operator is available. This is because you are comparing SNPs, which is a non-gene list, to a list of *A. fumigatus* Af293 genes.

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.

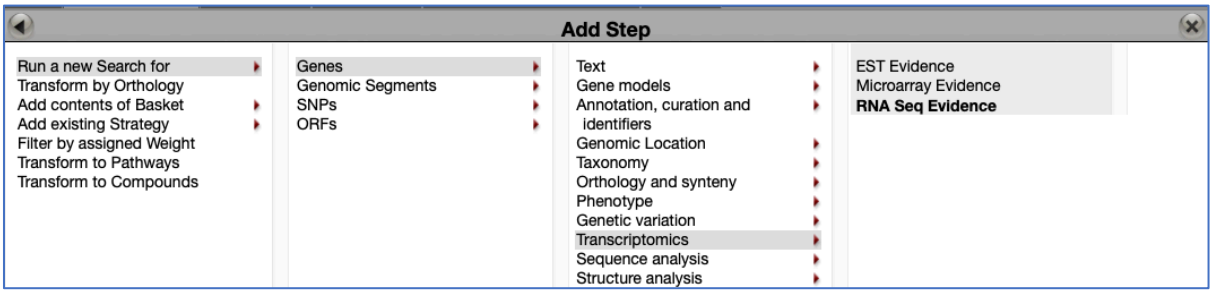




The search you deployed mapped the location of 1105 SNPs to the genome of *A. fumigatus* and identified genes that harbor SNPs within ORFs.

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

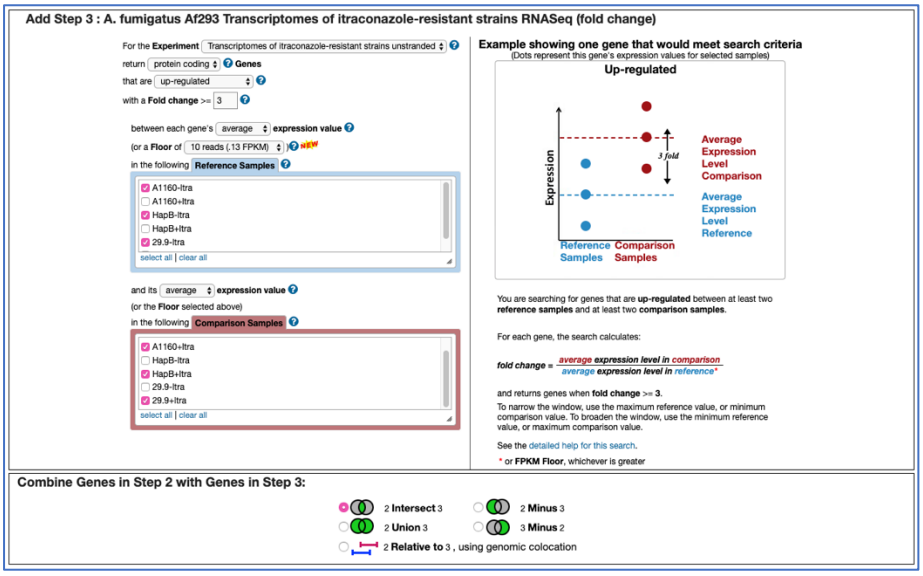
a. Click Add Step to Run a new Search for Genes using RNA Seq Evidence

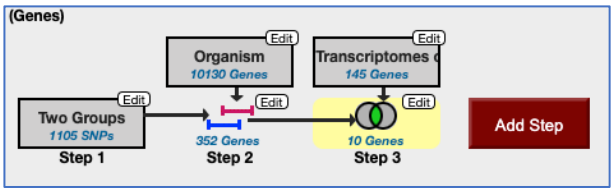


b. Filter datasets for fumigatus ('fum") and select the dataset titled "Transcriptomes of itraconazole-resistant strains (Bowyer 2016)"

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)



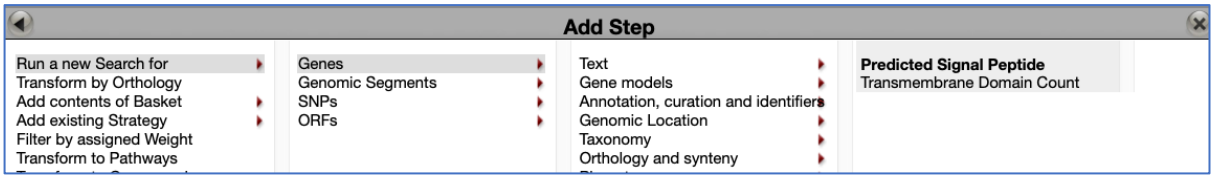


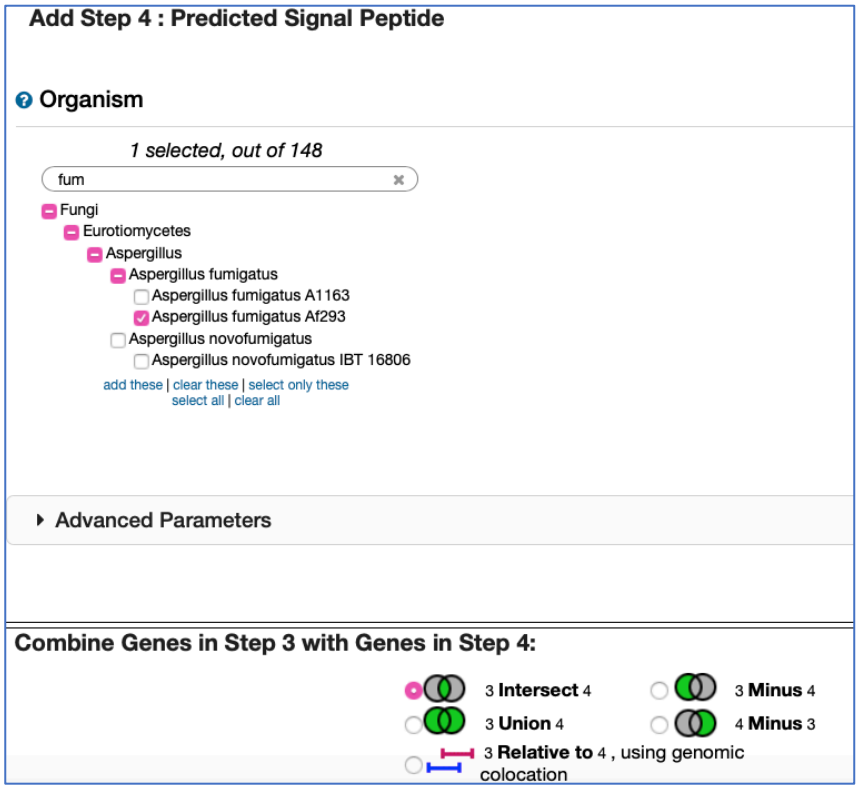
4. **Create a nested search for genes that have a predicted signal peptide OR a transmembrane domain OR both.**

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).

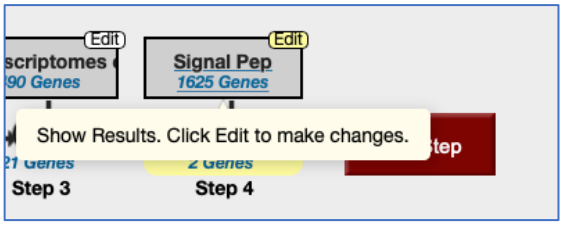
In the next search, we will use a nested strategy approach to:
- access two data types (signal peptide and transmembrane domain predictions)
- find genes that (1) have a predicted signal peptide or (2) have a transmembrane domain, or (3) have both
- intersect signal peptide and transmembrane domain predictions results with Step 3.

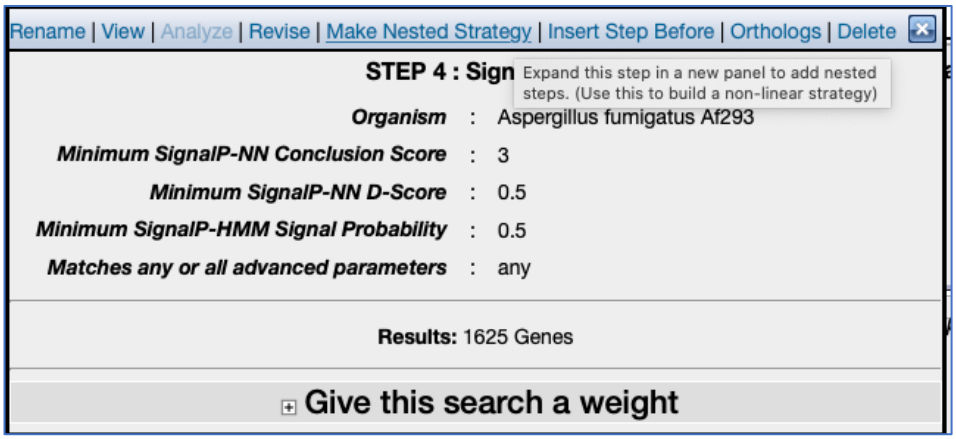
a. Click Add Step to Run a new Search for Genes in Protein targeting and localization, Predicted Signal peptide data





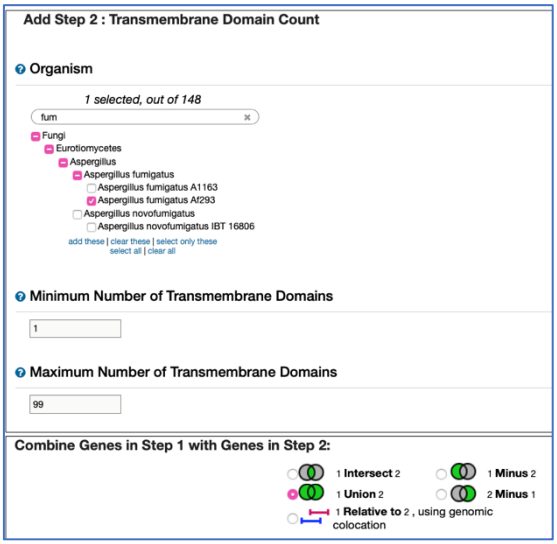
b. Hover over the step, click Edit, and select Make Nested Strategy





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:



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.