FungiDB: SNPs and Population Genetics

Learning Objective:

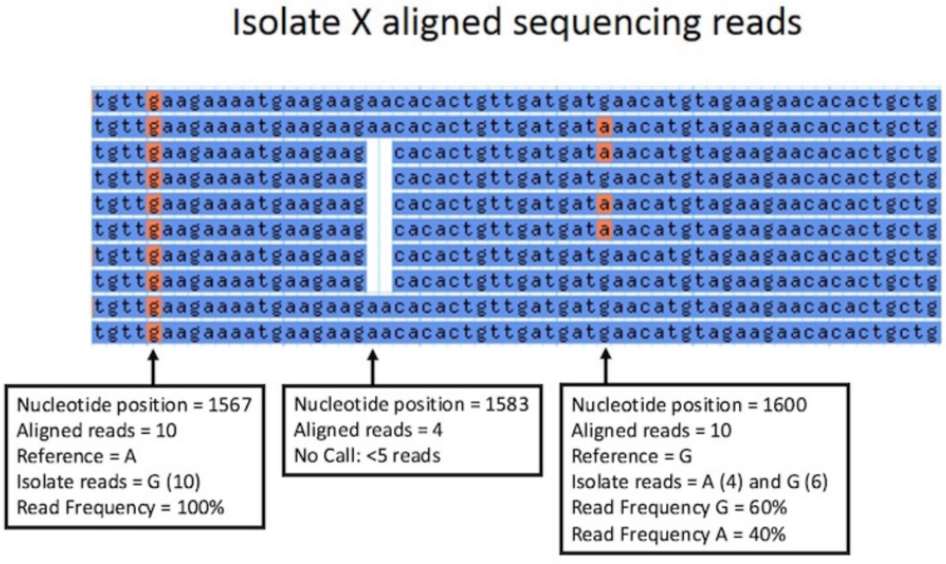
- Investigate SNP datasets using the following searches:
  - SNP characteristics,
  - SNPs between groups of isolates,
- Explore copy number variation records to identify aneuploidy cases.

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

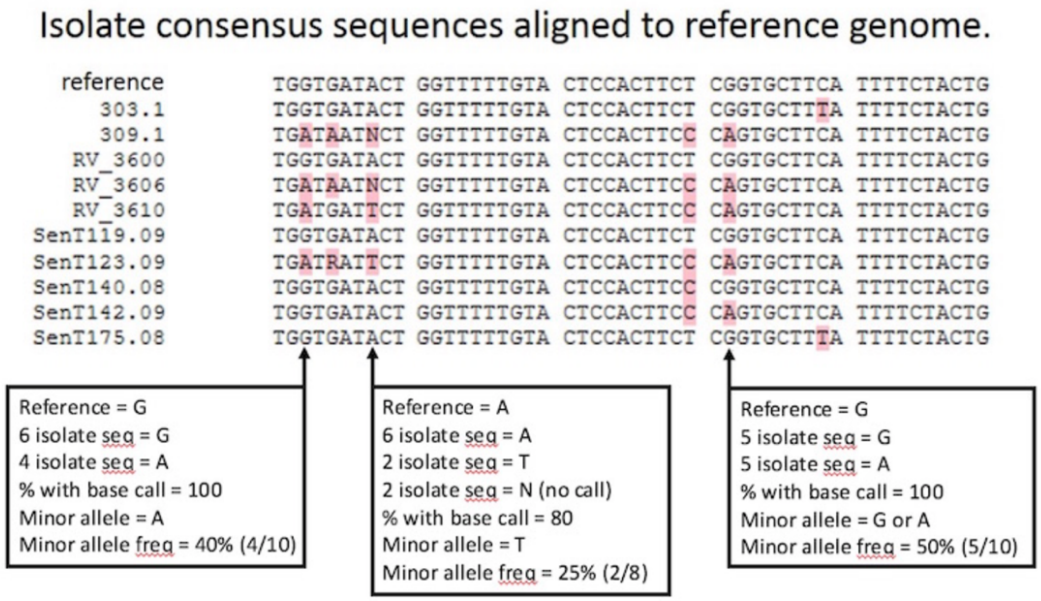
Read Frequency Threshold:

The parameter lets you control the quality of the sequencing data and the confidence of the SNP calls. Read Frequency Threshold is a particularly important parameter when dealing with diploid (or aneuploid) organisms since a read frequency of ~50% is expected for heterozygous SNPs.

Each isolate's sequencing reads are aligned to a reference genome and then each nucleotide position with 5 or more aligned reads is examined. A base call is made if the aligned reads meet your Read Frequency Threshold. For example, *Isolate X* has 10 aligned reads at nucleotide position 1600. If 6 reads are G and 4 reads are A, the read frequency is 60% for the G call and 40% for A. Running this search with the Read Frequency Threshold set to 80% will prevent a base call and consequently exclude *Isolate X* when returning SNPs for nucleotide position 1600. Running the search with the Read Frequency Threshold set to 60% will bring back a G for this isolate and a 40% threshold will return two calls (both G and A) at this position.



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

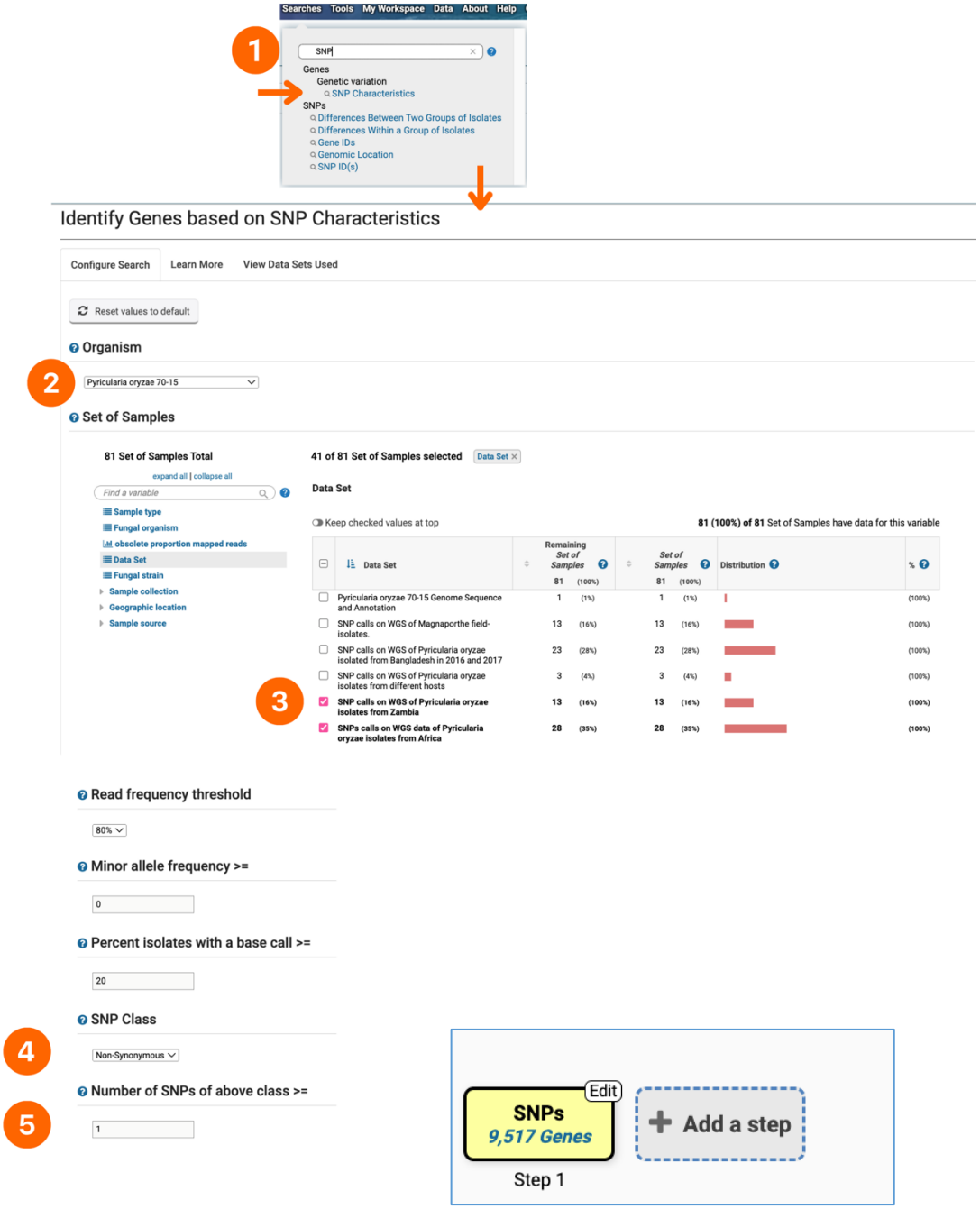


Percent isolates with a base call:
Sometimes an isolate does not have a base call at a certain nucleotide position because the Read Frequency Threshold was not met or because there were less than 5 aligned sequencing reads for that nucleotide position. In this case, a SNP can be returned by the search based on a subset of your isolate group. The 'Percent isolates with a base call' parameter defines the fraction of isolates that must have a base call before a SNP is returned for that nucleotide position. The default setting for this parameter is 80% or 8 out of 10 isolates in your group must have a base call for a SNP to be returned by the search. The higher this parameter, the more likely the SNP is to be high quality as regions difficult to align or difficult to sequence will tend to have a lower percentage of calls since the coverage and/or quality will be lower in that region.

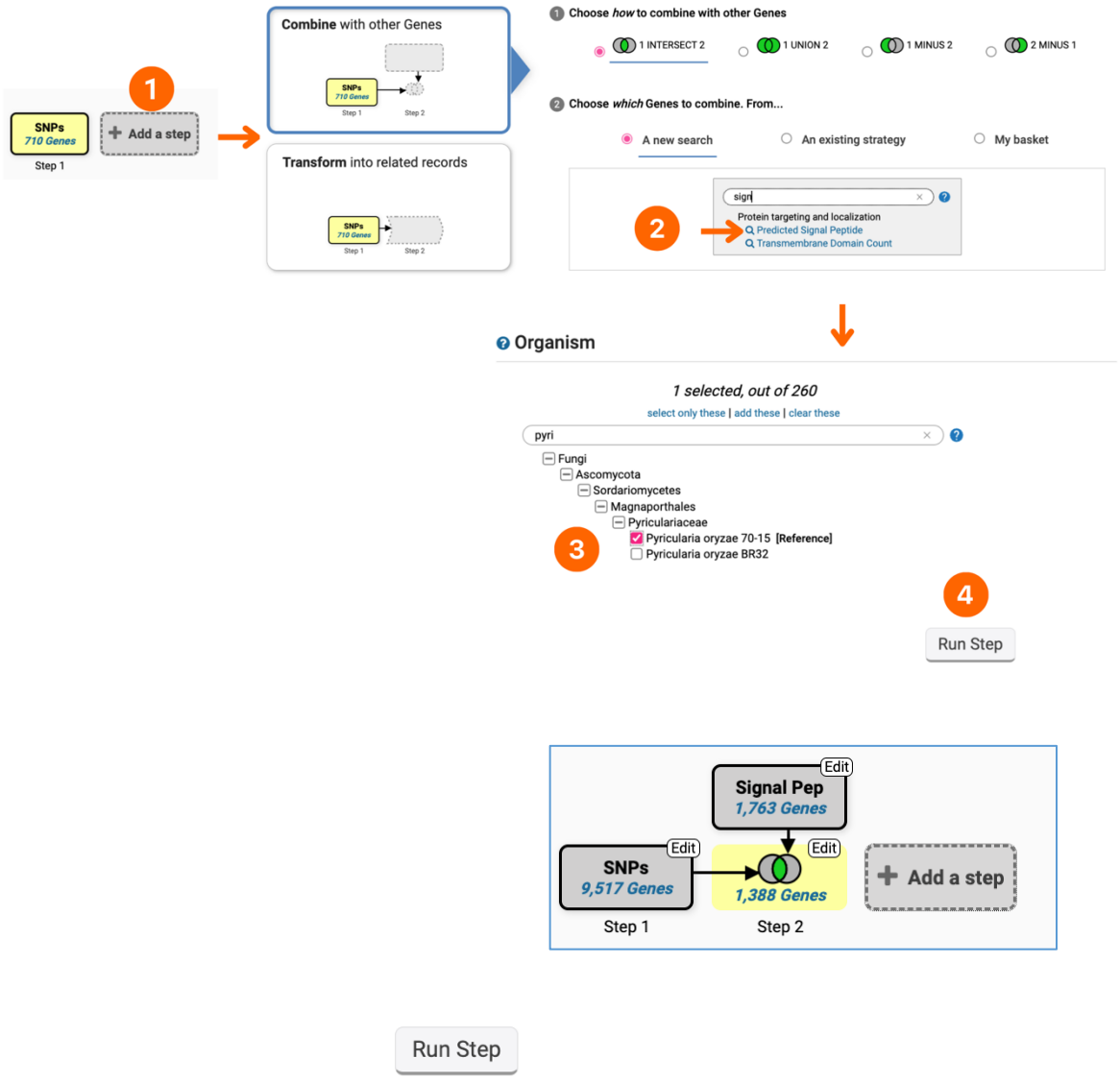
A. **Identify Genes based on SNP Characteristics search:** Identify putative nuclear effectors with at least 1 non-synonymous SNP in *Pyricularia oryzae*.

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

- **Identify genes with at least 1 non-synonymous SNP.**
    1. Deploy the “SNP characteristics’ search.
    2. Select *Pyricularia oryzae* 70-50 from the genome drop-down list.
    3. In the Data Set section, select the datasets where isoaltes were collected in Zambia and other African fields.
    4. Set the “SNP Class” parameter to “Non-Synonymous”.
    5. Choose to identify genes with at least 1 non-synonymous SNPs and click on the “Get Answer” button.



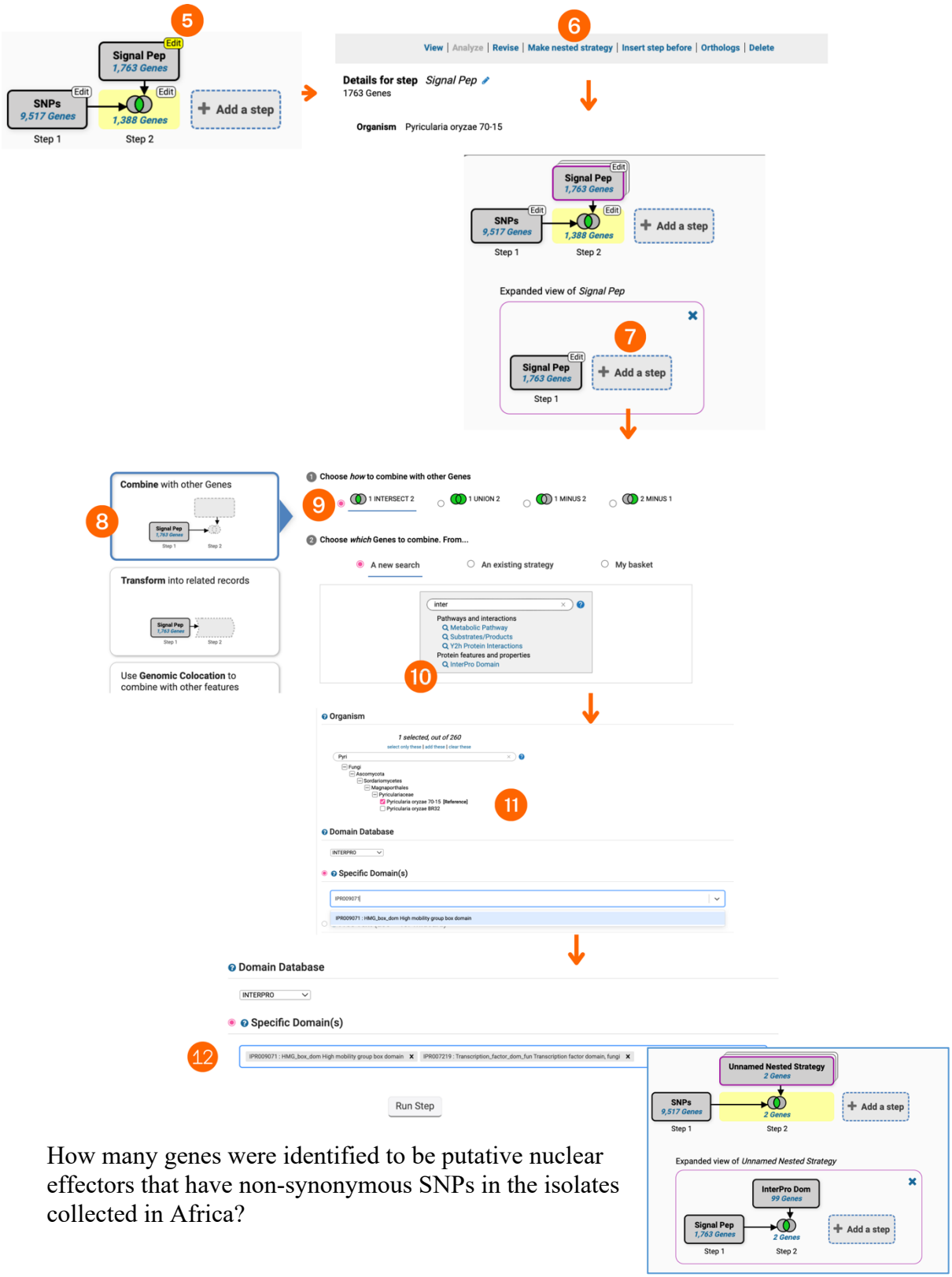
- **Identify putative nuclear effectors based on the presence of both a secretion signal and the DNA-binding domains IPR007219 or IPR009071 .**
    1. Click on the “Add a Step” button.
    2. Use the “Combine with Other Genes” option to deploy the “Predicted Signal Peptide” search.
    3. Set the genome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



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

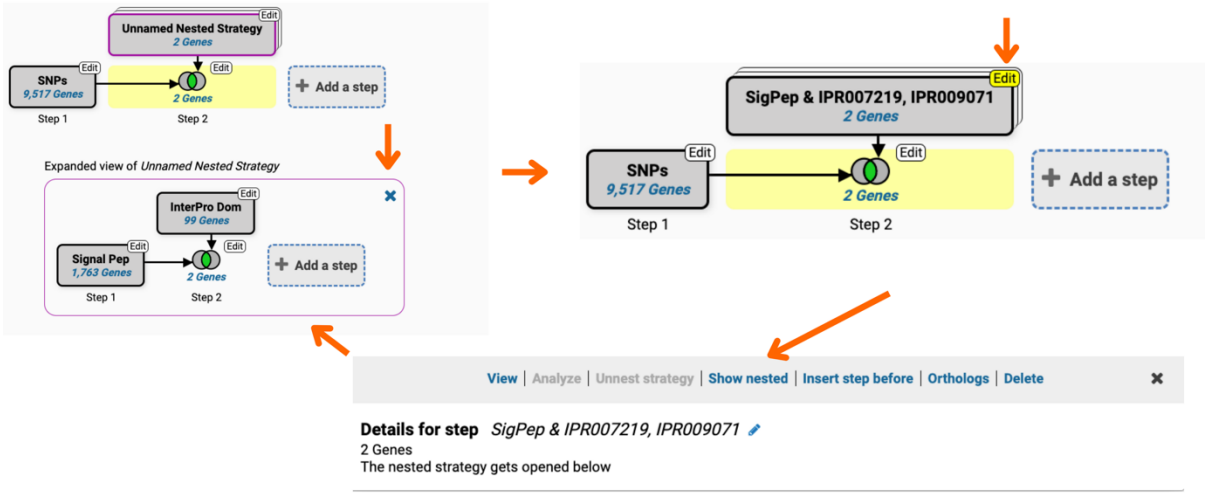
5. Hover over the “Signal Pep” search box and click on the “Edit” option.
6. Select the “Make nested strategy” option at the top.
7. Click on the “Add a Step” button within the “Expanded view of *Signal Pep”* (nested) strategy.
8. Select the “Combine with other Genes” search.

9. Set the Boolean operator to “1 intersect 2”.
10. Deploy the “InterPro Domain” search.
11. Set the genome to *Pyricularia oryzae* 70-50 and set the “Domain database” to InterPro and enter and select the following DNA binding domains from the dropdown menu: IPR007219, IPR009071.
12. Click on the “Run Step” once both domains are selected.



How many genes were identified to be putative nuclear effectors that have non-synonymous SNPs in the isolates collected in Africa?

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



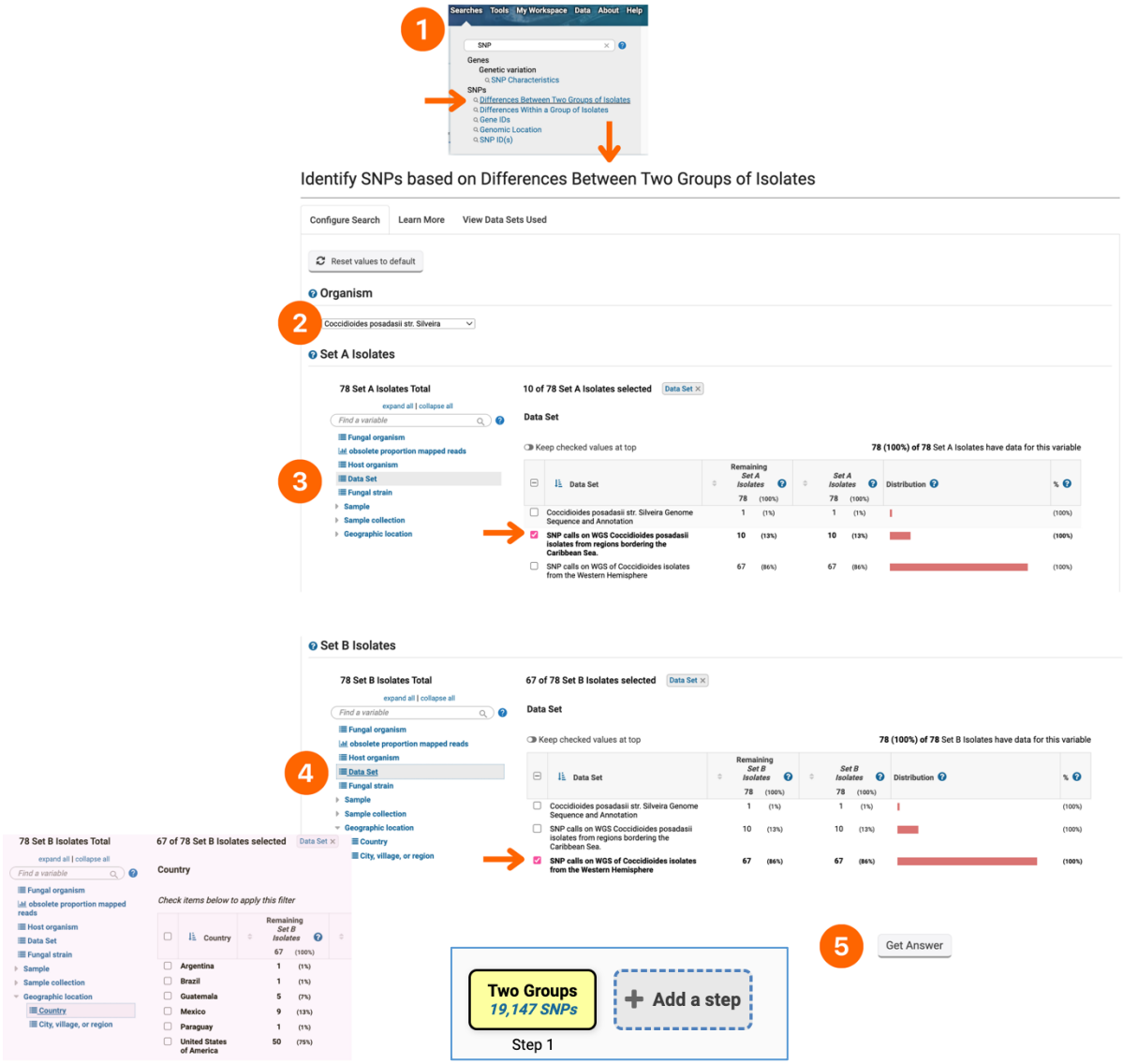
Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/bd657f5629cac5df

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

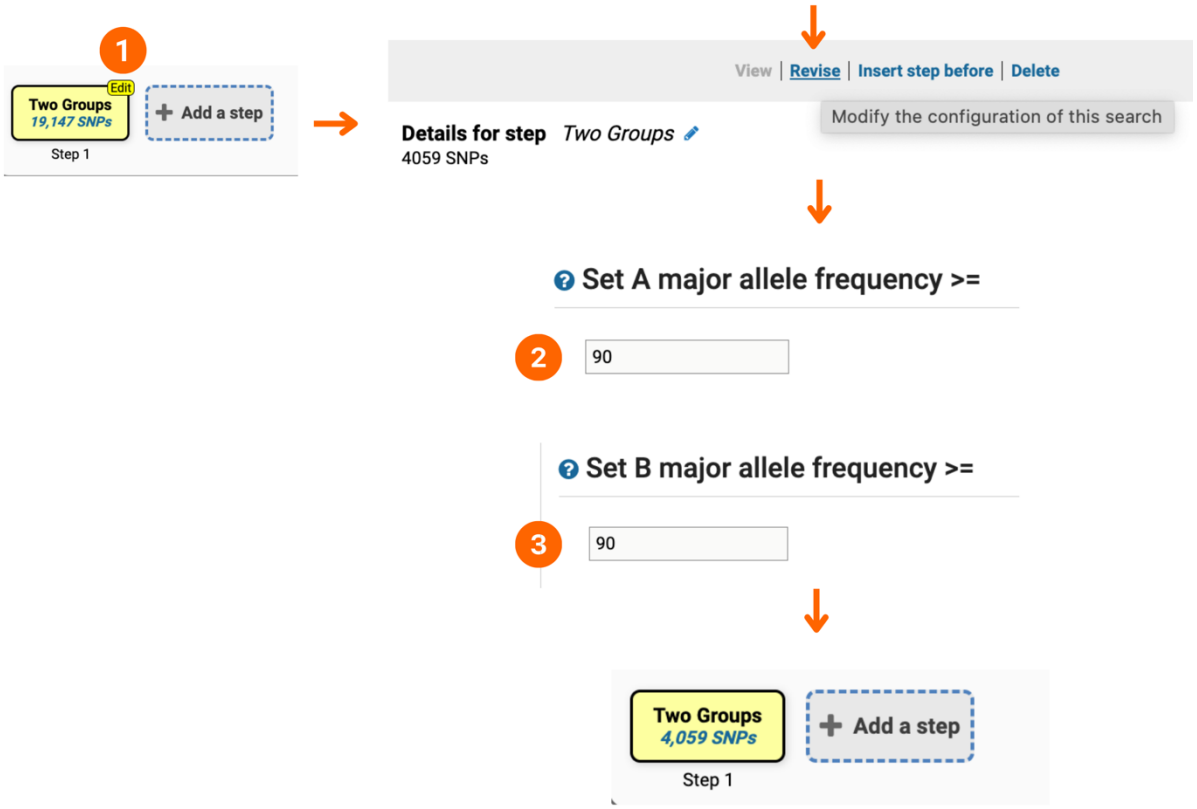
B. Identify SNPs based on Differences Between Two Groups of Isolates

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

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



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



The search strategy returns SNPs rather than genes, which are classified by genomic location within the results table. When individual SNPs fall within a gene, its corresponding Gene ID is listed next to the SNP record (Gene ID column).

SNP Id	Location	Gene ID	Position in protein	Set A Major Allele	Set A Major Allele Pct	Set A Major Product	Set B Major Allele	Set B Major Allele Pct	Set B Major Product
NGS_SNP.GL636538.9073	GL636538: 9,073	N/A	N/A	C	100	-	G	90	-
NGS_SNP.GL636538.8514	GL636538: 8,514	N/A	N/A	G	100	-	C	100	-
NGS_SNP.GL636538.3960	GL636538: 3,960	N/A	N/A	C	100	-	T	95.7	-
NGS_SNP.GL636537.6464	GL636537: 6,464	N/A	N/A	A	100	-	G	100	-
NGS_SNP.GL636537.4384	GL636537: 4,384	N/A	N/A	A	100	-	G	100	-
NGS_SNP.GL636537.1402	GL636537: 1,402	N/A	N/A	A	100	-	G	93.3	-
NGS_SNP.GL636536.8746	GL636536: 8,746	N/A	N/A	A	100	-	G	100	-
NGS_SNP.GL636536.6075	GL636536: 6,075	CPSG_10217	15	T	100	E	C	92.3	G
NGS_SNP.GL636536.532	GL636536: 532	N/A	N/A	T	100	-	A	100	-
NGS_SNP.GL636536.4473	GL636536: 4,473	N/A	N/A	T	100	-	C	92.3	-
NGS_SNP.GL636536.1587	GL636536: 1,587	CPSG_10216	738	T	100	T	C	93.3	A
NGS_SNP.GL636536.1541	GL636536: 1,541	CPSG_10216	753	G	100	A	A	95.8	V
NGS_SNP.GL636536.13558	GL636536: 13,558	CPSG_10220	295	A	100	F	G	90	F
NGS_SNP.GL636536.12038	GL636536: 12,038	N/A	N/A	G	100	-	A	91.4	-
NGS_SNP.GL636536.11250	GL636536: 11,250	N/A	N/A	T	100	-	C	91.3	-

- Each SNP is linked to its own record page. Click on the NGS_SNP.GL636536.6075.

SNP location, allele summary, associated GeneID, major and minor allele records can be found at the top of the page, followed by DNA polymorphism summary and SNP records table that is searchable by isolate IDs.

Add to basket Add to favorites Download SNP

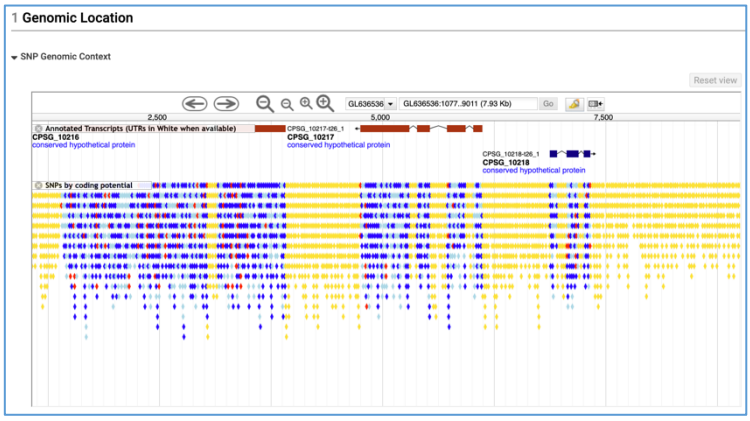
SNP: NGS_SNP.GL636536.6075

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

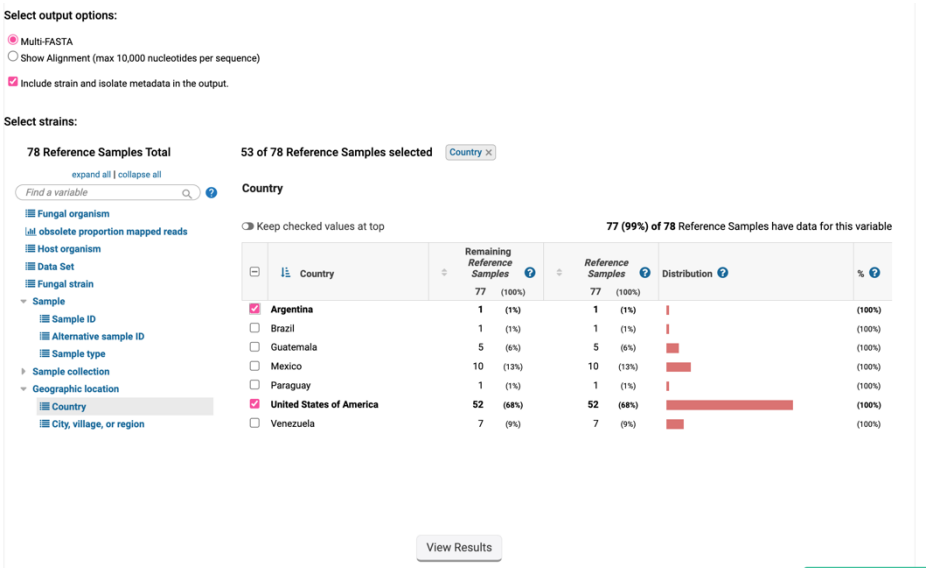
A summary of all SNPs detected in this gene across all datasets integrated into FungiDB is displayed in the SNP Genomic Context section:

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

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



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



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

Country Summary Download Data Sets

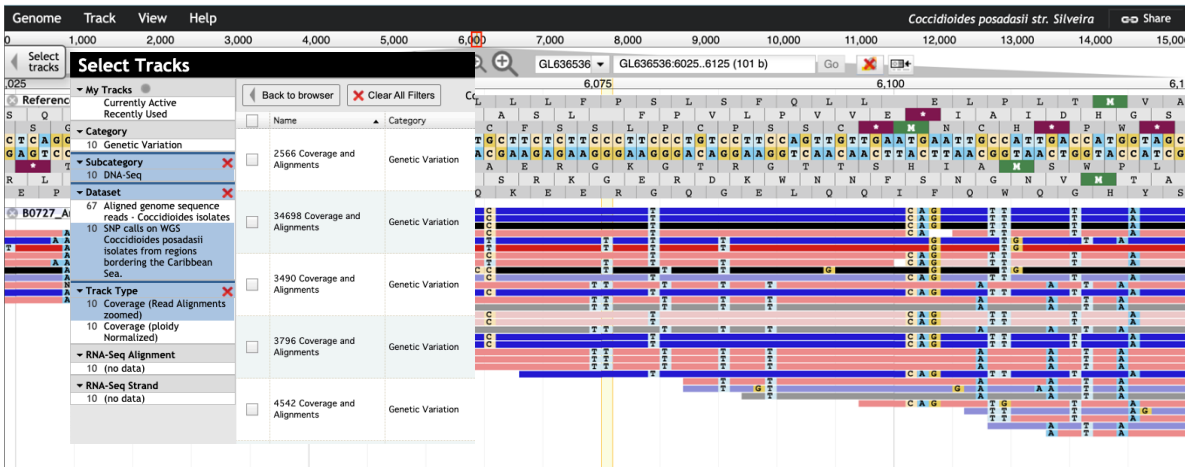
Search this table...

Geographic Location	#Alleles	Major Allele	Minor Allele	Other Allele
United States of America	65	C (.62)	T (.38)	N/A
Mexico	15	C (.53)	T (.47)	N/A
Venezuela	10	T (.7)	C (.3)	N/A
Guatemala	6	C (.83)	T (.17)	N/A
Argentina	2	C (.5)	T (.5)	N/A
Brazil	2	C (.5)	T (.5)	N/A
Paraguay	2	C (.5)	T (.5)	N/A
unknown	1	C (1)	N/A	N/A

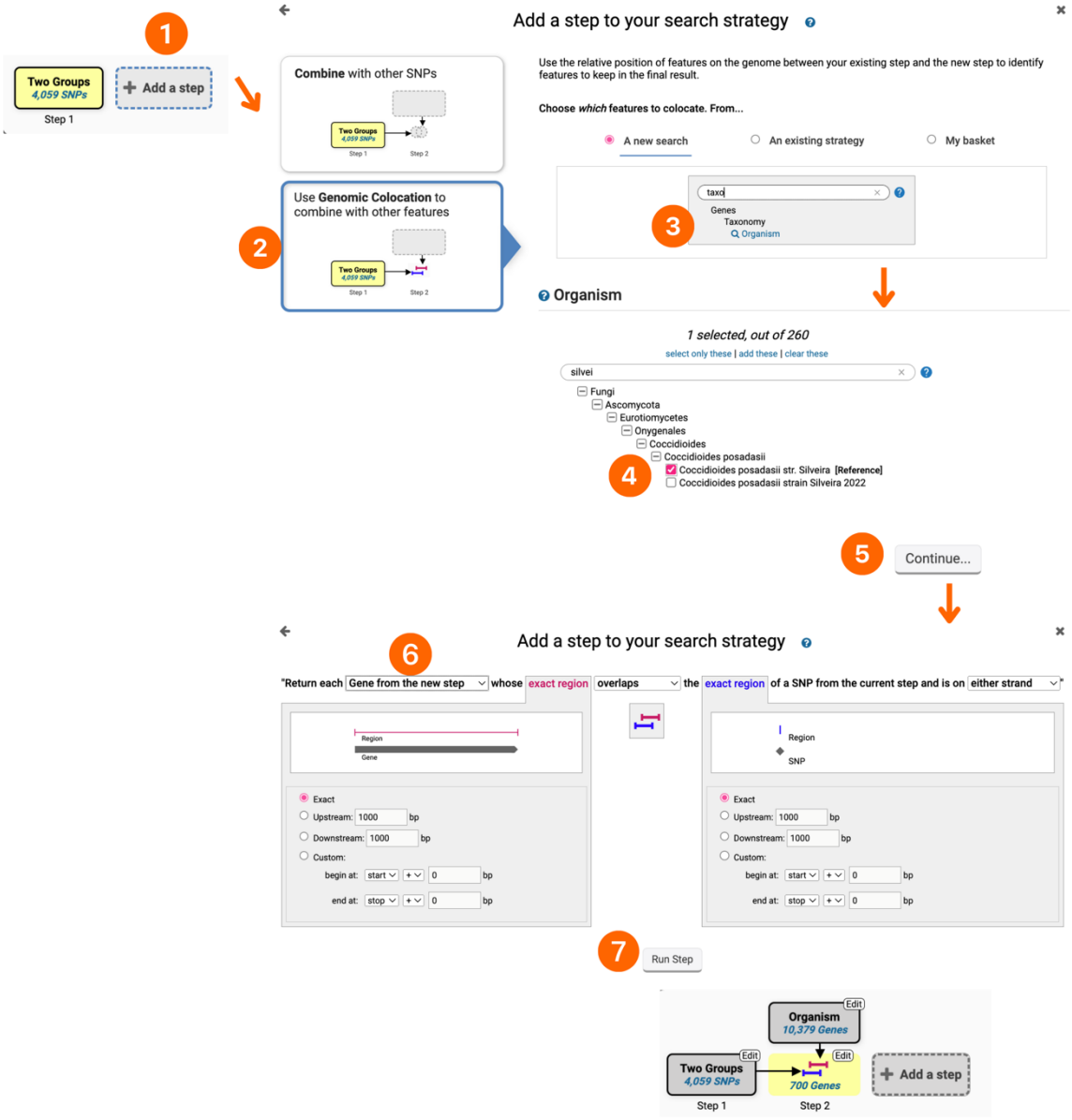
DNA-seq reads can be viewed by clicking on the *view DNA-seq reads* link from within the table.

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

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



- **Identify *C. posadasii* str. Silveira genes that harbor geographic-specific SNPs.**
    1. Click on the "Add a step" button.
    2. Select the "Use Genomic Colocation to combine with other features" tool.
    3. Filter searches on "taxonomy" to identify the "Organism" search.
    4. Select *C. posadasii* strain Silveira genome.
    5. Click on the "Continue…" button to specify colocation search parameters.
    6. Select to return genes by choosing the "Gene from the new step" from the drop-down menu while leaving other selections at default.
    7. Click on the "Run Step" button for results.



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

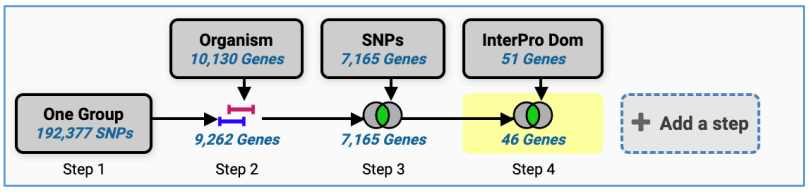
Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/d9d0fff2dbda229d

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

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

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

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



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

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

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

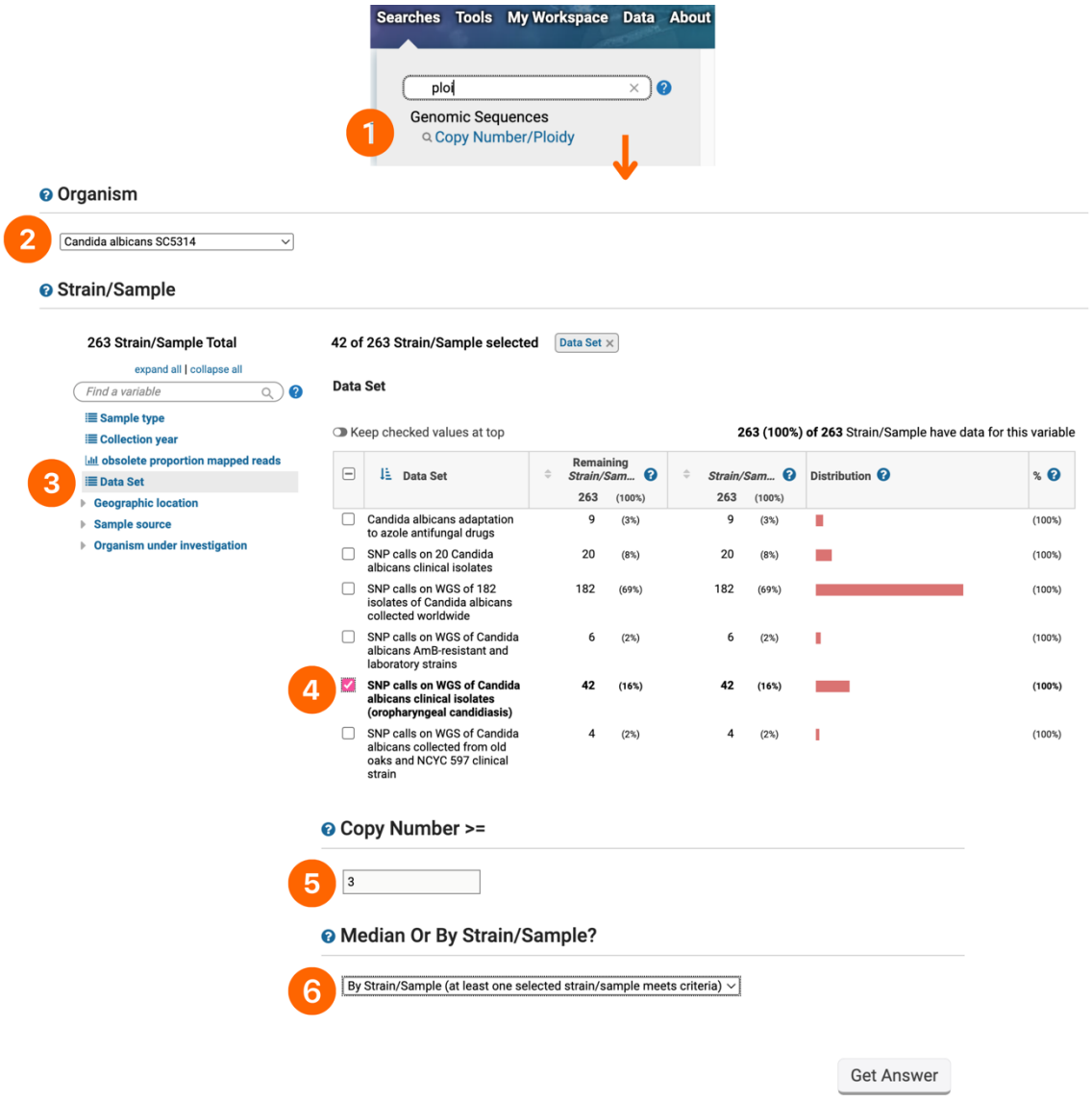
Minor Allele Frequency *parameter applies to your group of isolates. A SNP can occur in any number of isolates in your group and the least frequent SNP call across all isolates is the Minor Allele Frequency. A SNP will be returned by the search if the frequency of the minor allele is equal to or greater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

D. Copy number variation & ploidy searches.

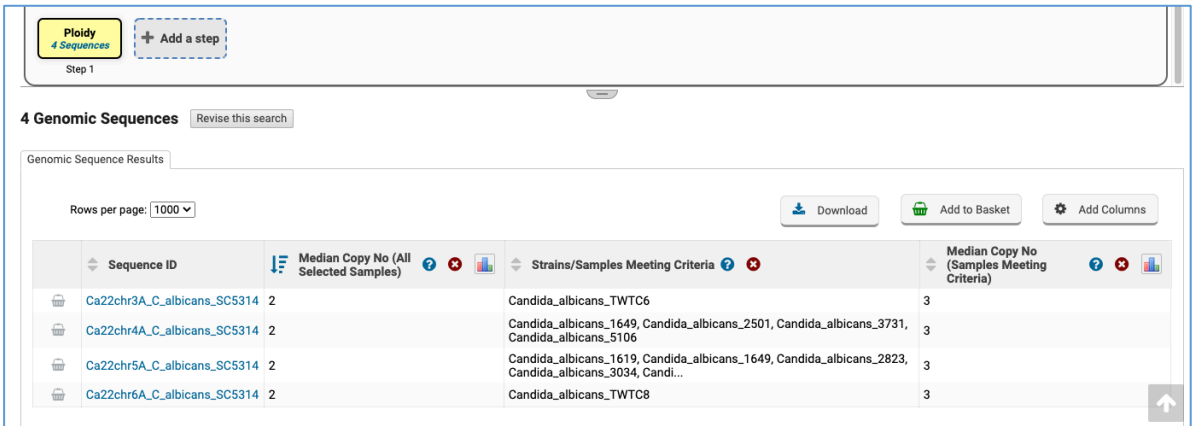
Gene copy number variation can be caused by deletions or duplications. In addition to being useful for variant calling, high throughput sequencing data can be used to determine regions with copy number variation (CNV). All reads in FungiDB are mapped to the same reference strain as SNP datasets and, as a result, we can estimate a gene's copy number in each of the aligned strains.

D.1. Copy Number/Ploidy search (Genomic Sequences)
Find genomic sequences or chromosomes based on their estimated copy number in resequenced strains. Genomic sequences returned by the search will have either have a median estimated copy number greater than or equal to the value you entered for the Copy Number across the selected strains/samples or will have an estimated copy number greater than or equal to the value you entered for the Copy Number in at least one of the selected strains/samples.

- **Identify trisomic chromosomes in clinical isolates of *Candida albicans*.**
    1. Deploy the "Copy Number/Ploidy" search.
    2. Set the genome to Candida albicans SC5314.
    3. Navigate to the Data Set section.
    4. Select the dataset called "SNP calls on WGS of Candida albicans clinical isolates (oropharyngeal candidiasis)".
    5. Set the Copy Number to "3".
    6. Select to identify ploidy "By strain/sample" and click on the "Get Answer" button.



The search by strain/sample (i.e., at one or more of the selected strains has to match the criteria rather than the median of the selected strains matching) is intended to find chromosomes where the whole chromosome is duplicated. It may find chromosomes where partial aneuploidy involves most of the chromosome but is unlikely to find chromosomes where partial aneuploidy only covers a small region. Also, because this search currently relies on coverage alone, it will not find instances of global genome duplication (e.g., all chromosomes became triploid).

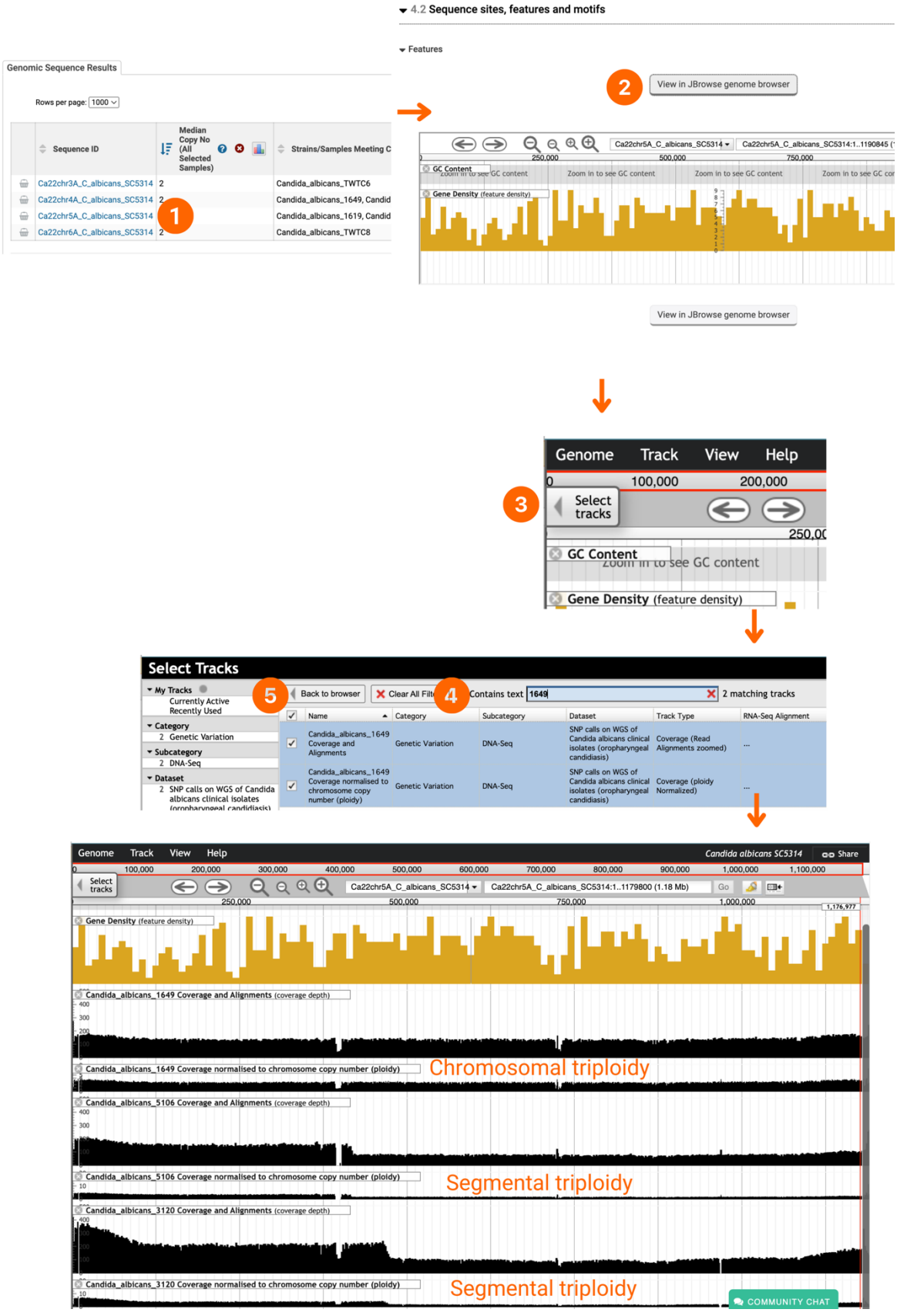


- **Explore segmental aneuploidy in JBrowse.**

JBrowse has two coverage tracks:

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

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



Notice examples of chromosomal (1649) and segmental triploidy (5106 and 3120). Note that the whole chromosome is shown in both screenshots, and both tracks are shown for each sample. Note: VEuPathDB is not currently normalizing for telomere proximity.

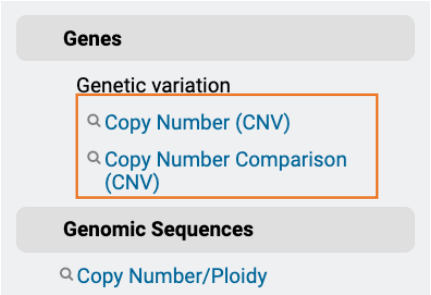
URL:
https://fungidb.org/fungidb/jbrowse/index.html?loc=Ca22chr5A_C_albicans_SC5314%3A1..1190845&data=%2Ffungidb%2Fservice%2Fjbrowse%2Ftracks%2FcalbSC5314&tracks=gcContent%2CgeneDensity%2CCandida_albicans_1649%20Coverage%20and%20Alignments%2CCandida_albicans_1649%20Coverage%20normalised%20to%20chromosome%20copy%20number%20(ploidy)%2CCandida_albicans_5106%20Coverage%20and%20Alignments%2CCandida_albicans_5106%20Coverage%20normalised%20to%20chromosome%20copy%20number%20(ploidy)%2CCandida_albicans_3120%20Coverage%20and%20Alignments%2CCandida_albicans_3120%20Coverage%20normalised%20to%20chromosome%20copy%20number%20(ploidy)&highlight=

Strategy URL:
https://fungidb.org/fungidb/app/workspace/strategies/import/6dc86b214d14a5f3
References: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383195/

D.2. Copy Number search (Genes)

E. Using Gene Searches

One way to discover regions of potential segmental aneuploidy is to use the searches for genes by copy number. We have two searches: Gene searches taking advantage of sequence alignment data can be found under the under the "Genetic Variation" category. Two available searches that define regions of CNV are:



- **Copy number:** This search returns genes that are present at copy numbers (haploid number or gene dose) within a range that you specify.

- **Copy number comparison:** This search compares the estimated copy number of a gene in the re-sequenced strain with the copy number in the reference annotation. The copy number in the reference annotation is calculated as the number of genes that are in the same ortholog group as the gene of interest. We infer that these genes have arisen as a result of tandem duplication of a common ancestor.

Different metrics for defining copy number:

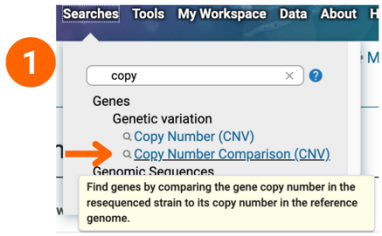
- **Haploid number** is the number of genes on an individual chromosome.
- **Gene dose** is the total number of genes in an organism, accounting for copy number of the chromosome.

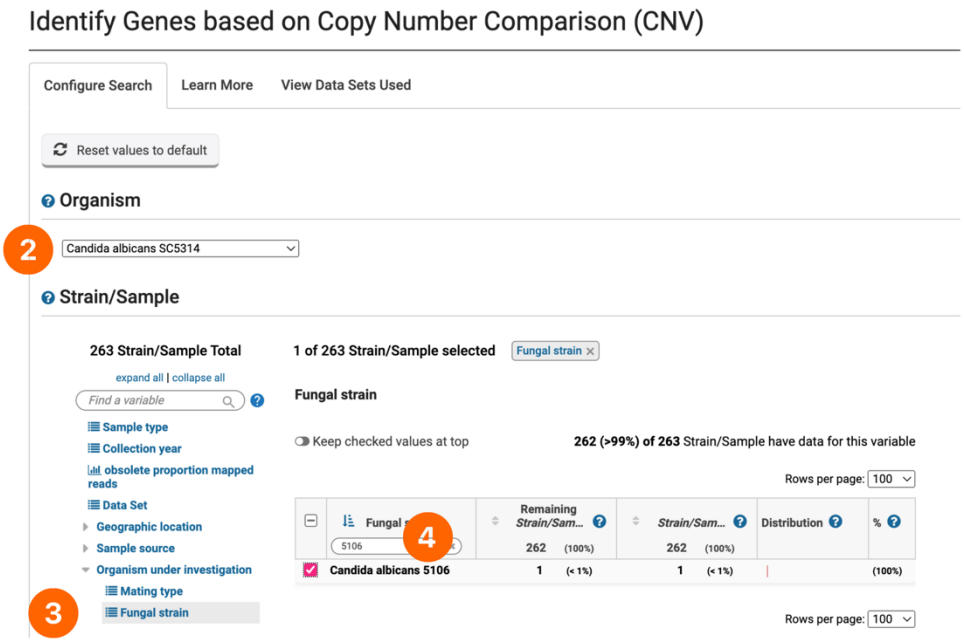
For example, a single-copy gene in a diploid organism has a haploid number of 1 and a gene dose of 2. You can choose to search for genes where at least one of your selected isolates meets your cutoff criteria for the chosen metric (By Strain/Sample), or where the median of the chosen metric across all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

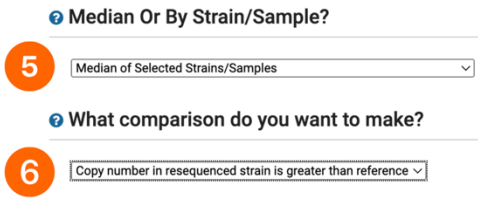
- **Discover regions of potential segmental aneuploidy in *Candida albicans* isolate 5106.**
    1. Deploy the “Copy Number Comparison (CNV)” search.
    2. Select the genome for “Candida albicans”.
    3. Navigate to the Fungal strain” metadata field.
    4. Filter isolates for “5106” and check the box to select this isolate.
    5. Leave the “Median or By Strain/Sample” parameter at default.

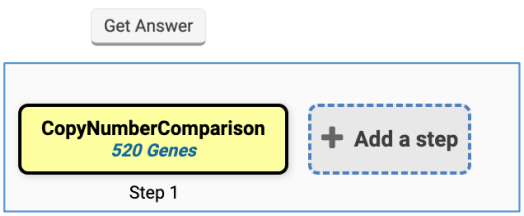
    Note: Choosing Median or By Strain/Sample will only make a difference if you have multiple strains.

    6. From the drop-down menu select the “Copy number in resequenced strain is greater than reference” option.

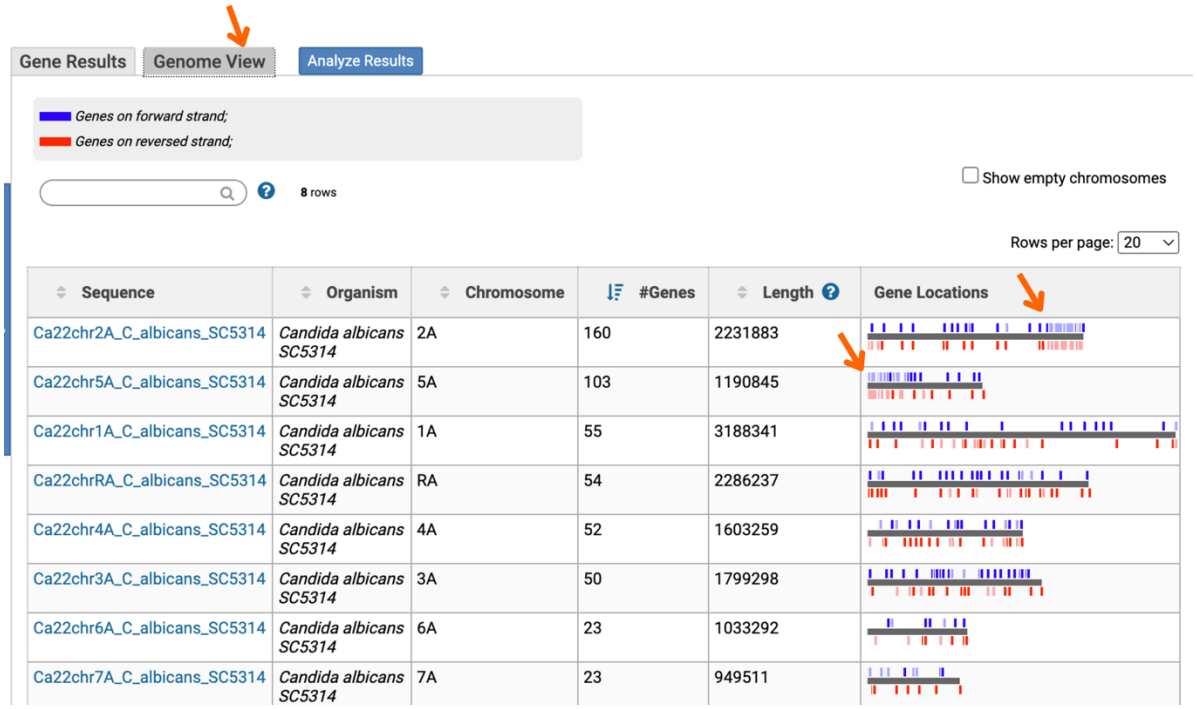








Examine the results using the Genome View option.



Sequence	Organism	Chromosome	#Genes	Length	Gene Locations
Ca22chr2A_C_albicans_SC5314	*Candida albicans SC5314*	2A	160	2231883	
Ca22chr5A_C_albicans_SC5314	*Candida albicans SC5314*	5A	103	1190845	
Ca22chr1A_C_albicans_SC5314	*Candida albicans SC5314*	1A	55	3188341	
Ca22chrRA_C_albicans_SC5314	*Candida albicans SC5314*	RA	54	2286237	
Ca22chr4A_C_albicans_SC5314	*Candida albicans SC5314*	4A	52	1603259	
Ca22chr3A_C_albicans_SC5314	*Candida albicans SC5314*	3A	50	1799298	
Ca22chr6A_C_albicans_SC5314	*Candida albicans SC5314*	6A	23	1033292	
Ca22chr7A_C_albicans_SC5314	*Candida albicans SC5314*	7A	23	949511	

As you can see in the highlighted regions, large numbers of genes that are predicted to have increased copy numbers are clustered at the right-hand end of chromosome 2 and the left hand end of chromosome 5, corresponding to the segmental aneuploidies shown in the JBrowse session above.