Genetic Variation Exercises

SNPs and CNVs

Learning Objective:

- Run SNP searches in VEuPathDB
- Explore SNP search parameters and their effect on search results
- Use SNP searches to identify genes that are under diversifying or stabilizing selection
- Run CNV searches in VEuPathDB
- Explore CNV search parameters
- Use CNV searches to identify regions of a genome that exhibit duplications or deletions.

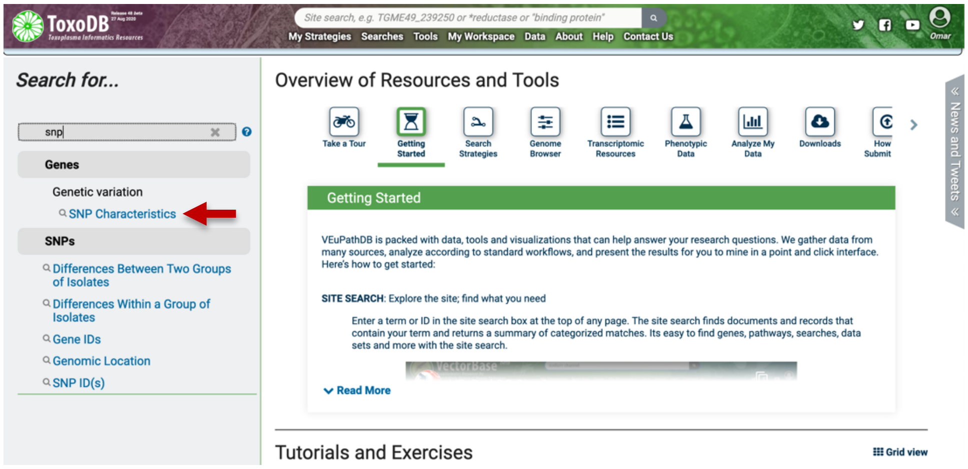
Single Nucleotide Polymorphisms (SNPs): single nucleotide changes between isolates or strains. SNPs have different functional effects with most having no consequential effect on gene function. SNPs may directly affect protein function when they are non-synonymous (results in a change in the amino acid; missense) or when they are cause a premature stop codon (nonsense). SNPs that do not fall within genes are non-coding (between genes or intronic). These types of SNPs may still affect splicing, mRNA stability, transcription, etc.
Copy number variation (CNV): variation in copy number of genes or regions of a genome. CNVs may be result of deletions or duplications.
See appendix for more information.

SNP Searches

In VEuPathDB SNPs can be used to characterize similarities and differences within a group of isolates or that distinguish between two groups of isolates. They can also be utilized to identify genes that may be under evolutionary pressure, either to stay the same (purifying selection) or to change (diversifying or balancing selection). Isolates are assayed for SNPs in VEuPathDB by two basic methods; re-sequencing and then alignment of sequence reads to a reference genome or DNA hybridization to a SNP-chip array (available in PlasmoDB only). In these exercises we'll explore both methods and ask a variety of questions to identify SNPs or genes of interest. If you do not understand the purpose of a parameter, please remember to mouse over the "?" icon and/or read the more detailed description at the bottom of the question page.

1. **Identify *T. gondii* genes that contain at least 20 nonsynonymous SNPs. For this exercise use http://toxodb.org**

   a. Start by running a search for genes based on SNP characteristics – this search can be found under the ‘Genetic Variation’ category.



VEuPathDB is packed with data, tools and visualizations that can help answer your research questions. We gather data from many sources, analyze according to standard workflows, and present the results for you to mine in a point and click interface. Here's how to get started:

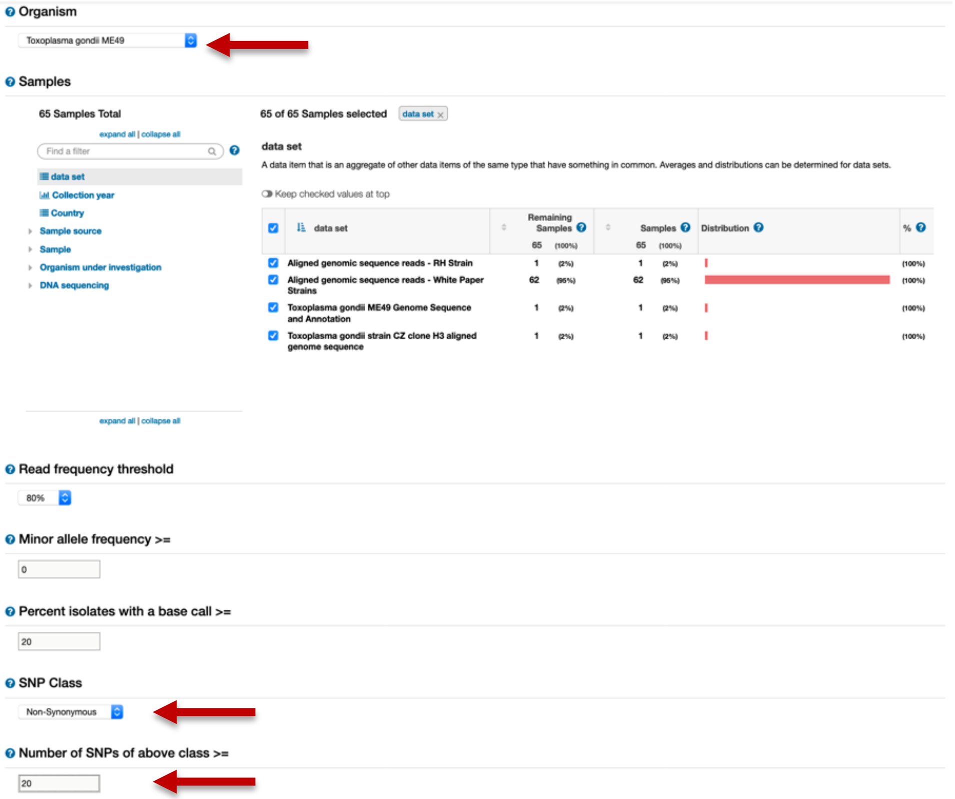
SITE SEARCH: Explore the site; find what you need

Enter a term or ID in the site search box at the top of any page. The site search finds documents and records that contain your term and returns a summary of categorized matches. Its easy to find genes, pathways, searches, data sets and more with the site search.

   b. Select Toxoplasma gondii ME49 from the drop-down list. Notice how the sample information changes when you change organism.

   c. In the sample section, select all available samples.

   d. Change the SNP class to Non-synonymous and the ‘number of SNPs of above class’ field to 20.


expand all | collapse all
Find a filter
data set
Collection year
Country
Sample source
Sample
Organism under investigation
DNA sequencing
expand all | collapse all
Keep checked values at top

Organism

Toxoplasma gondii ME49

Samples

65 Samples Total

65 of 65 Samples selected data set

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.

☑	data set	Remaining Samples 65 (100%)	Samples 65 (100%)	Distribution	%
☑	Aligned genomic sequence reads - RH Strain	1 (2%)	1 (2%)		(100%)
☑	Aligned genomic sequence reads - White Paper Strains	62 (95%)	62 (95%)		(100%)
☑	Toxoplasma gondii ME49 Genome Sequence and Annotation	1 (2%)	1 (2%)		(100%)
☑	Toxoplasma gondii strain CZ clone H3 aligned genome sequence	1 (2%)	1 (2%)		(100%)

Read frequency threshold

80%

Minor allele frequency >=

0

Percent isolates with a base call >=

20

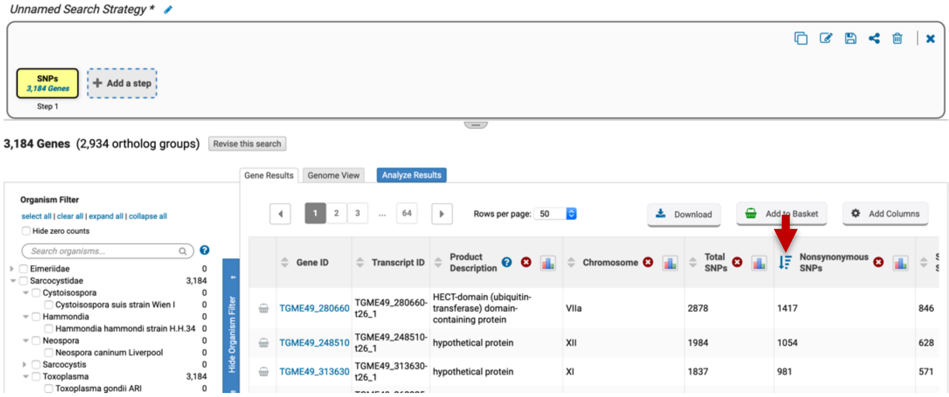
SNP Class

Non-Synonymous

Number of SNPs of above class >=

20

e. How many genes did you return? Which gene has the highest number of non-synonymous SNPs? (*hint*: sort the non-synonymous SNP columns).


*Unnamed Search Strategy **

3,184 Genes (2,934 ortholog groups) Revise this search

Organism Filter

select all | clear all | expand all | collapse all

□ Hide zero counts

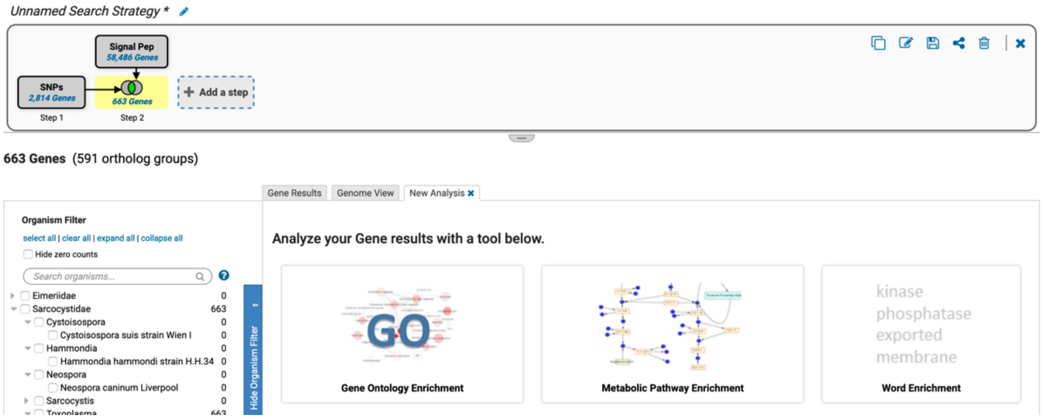
□ Eimeriidae	0
□ Sarcocystidae	3,184
□ Cystoisospora	0
□ Cystoisospora suis strain Wien I	0
□ Hammondia	0
□ Hammondia hammondi strain H.H.34	0
□ Neospora	0
□ Neospora caninum Liverpool	0
□ Sarcocystis	0
□ Toxoplasma	3,184
□ Toxoplasma gondii ARI	0

Gene ID	Transcript ID	Product Description	Chromosome	Total SNPs	Nonsynonymous SNPs	
TGME49_280660	TGME49_280660-t26_1	HECT-domain (ubiquitin-transferase) domain-containing protein	VIIa	2878	1417	846
TGME49_248510	TGME49_248510-t26_1	hypothetical protein	XII	1984	1054	628
TGME49_313630	TGME49_313630-t26_1	hypothetical protein	XI	1837	981	571

f. What happens if you revise this search and change the “Percent isolates with a base call >=” field to 100?

g. How many of these genes have a predicted secretory signal peptide? (*hint*: add a step that identifies all genes with a signal peptide).

h. What kinds of genes are in this result list? One way to determine if you have naything enriched in your results is to run an enrichment analysis. Click on the “Analyze Results” tab then compare the results you get from the GO enrichment and from the Word enrichment, we will disucss these results.


*Unnamed Search Strategy **

663 Genes (591 ortholog groups)

Organism Filter

select all | clear all | expand all | collapse all

□ Hide zero counts

□ Eimeriidae	0
□ Sarcocystidae	663
□ Cystoisospora	0
□ Cystoisospora suis strain Wien I	0
□ Hammondia	0
□ Hammondia hammondi strain H.H.34	0
□ Neospora	0
□ Neospora caninum Liverpool	0
□ Sarcocystis	0
□ Toxoplasma	663

Analyze your Gene results with a tool below.

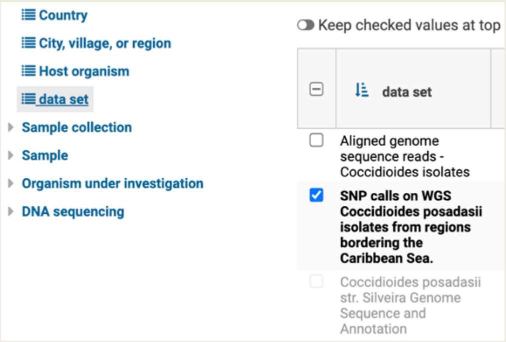
2. Identifying SNPs between fungal isolates collected in distinct geographical areas

For this exercise use https://fungidb.org

The example described below identifies SNPs in *Coccidioides posadasii* (*C. posadasii*) str. Silveira 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.

a) Identify SNPs based on Differences Between Two Groups of Isolates

- From the *Search for…*, navigate to the *Identify SNPs based on Differences Between Two Groups of Isolates*.
- From the drop-down menu select Coccidioides posadasii str. Silveira
- From the *Data set* check the box to select the data set titled "SNP calls on WGS Coccidioides posadasii isolates from regions bordering the Caribbean Sea". More information about the dataset: https://fungidb.org/fungidb/app/record/dataset/DS_d27c9dd420

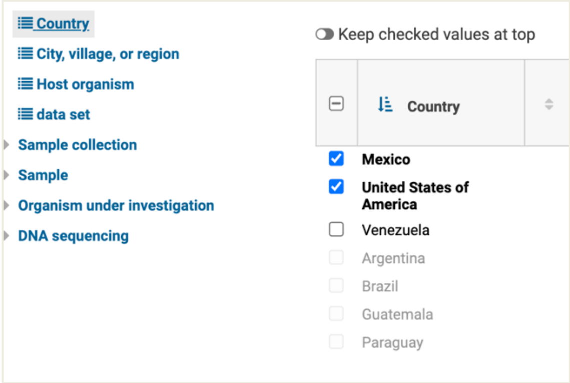


☐ Aligned genome sequence reads - Coccidioides isolates

☑ SNP calls on WGS Coccidioides posadasii isolates from regions bordering the Caribbean Sea.

☐ Coccidioides posadasii str. Silveira Genome Sequence and Annotation

- Next, click on the *Country* option and for the first group select Mexico and United States of America
  - For the second group (Set B isolates), use the same dataset and set the country parameter to Venezuela.
  - Set your search stringency: Major allele frequency = 90 and Percent of isolates with base call = 70 for both groups. Feel free to come back to this step and choose different settings to see how it affects your search.



☑ Mexico

☑ United States of America

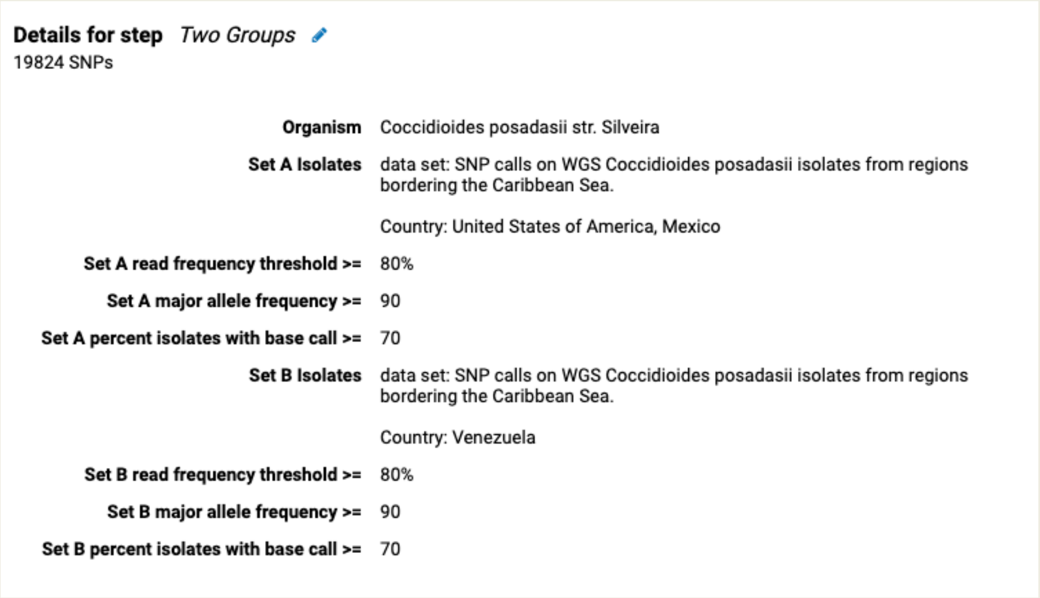
☐ Venezuela

☐ Argentina

☐ Brazil

☐ Guatemala

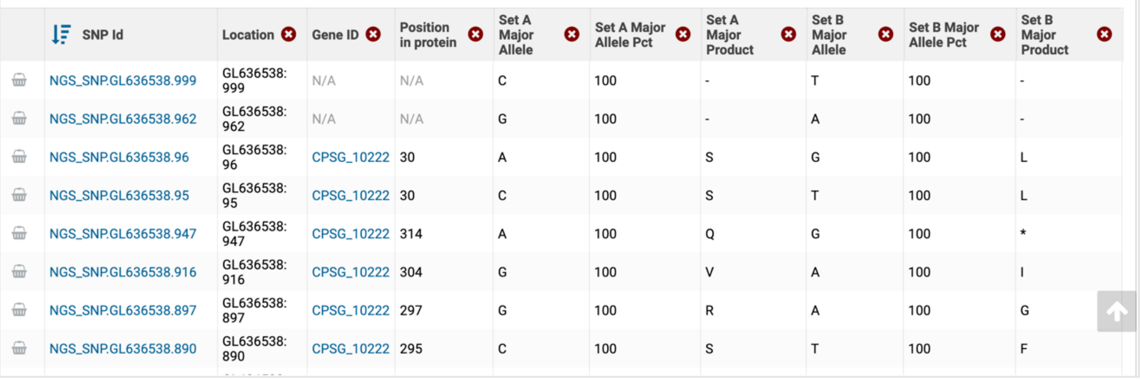
☐ Paraguay



Details for step *Two Groups*
19824 SNPs

Organism	Coccidioides posadasii str. Silveira
Set A Isolates	data set: SNP calls on WGS Coccidioides posadasii isolates from regions bordering the Caribbean Sea. Country: United States of America, Mexico
Set A read frequency threshold >=	80%
Set A major allele frequency >=	90
Set A percent isolates with base call >=	70
Set B Isolates	data set: SNP calls on WGS Coccidioides posadasii isolates from regions bordering the Caribbean Sea. Country: Venezuela
Set B read frequency threshold >=	80%
Set B major allele frequency >=	90
Set B percent isolates with base call >=	70

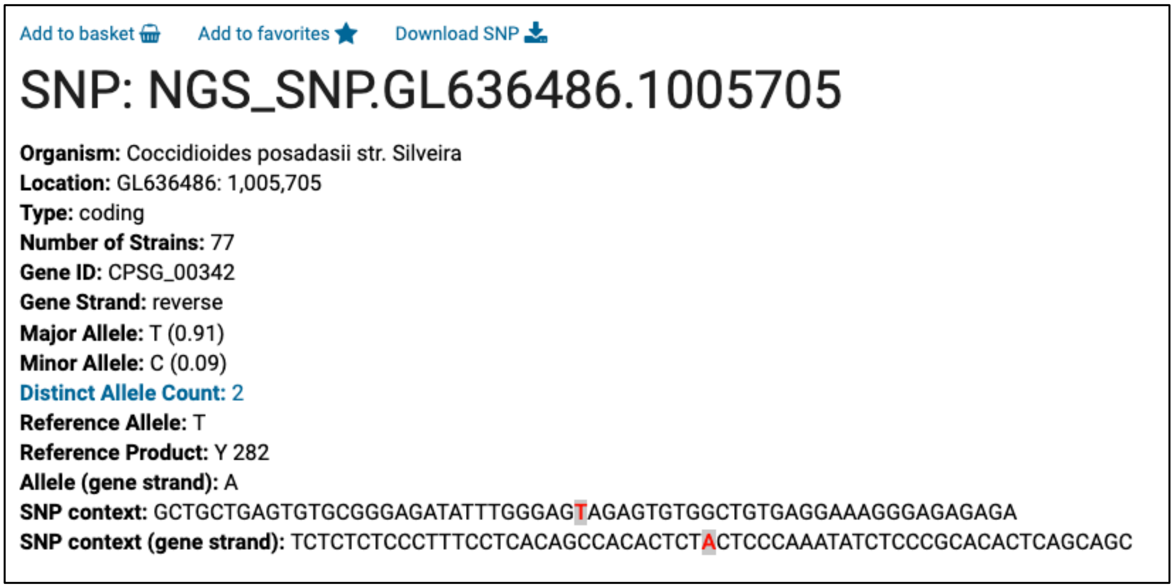
The search strategy returns SNPs rather than genes, which are classified by genomic location within the results table. When individual SNPs fall within a gene, its corresponding Gene ID is listed next to the SNP record.



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.GL636538.999	GL636538: 999	N/A	N/A	C	100	-	T	100	-
NGS_SNP.GL636538.962	GL636538: 962	N/A	N/A	G	100	-	A	100	-
NGS_SNP.GL636538.96	GL636538: 96	CPSG_10222	30	A	100	S	G	100	L
NGS_SNP.GL636538.95	GL636538: 95	CPSG_10222	30	C	100	S	T	100	L
NGS_SNP.GL636538.947	GL636538: 947	CPSG_10222	314	A	100	Q	G	100	*
NGS_SNP.GL636538.916	GL636538: 916	CPSG_10222	304	G	100	V	A	100	I
NGS_SNP.GL636538.897	GL636538: 897	CPSG_10222	297	G	100	R	A	100	G
NGS_SNP.GL636538.890	GL636538: 890	CPSG_10222	295	C	100	S	T	100	F

- To examine a SNP record page, click on the *NGS_SNP.xxxx* link.
- Let's take a look at the SNP record page for SNP: NGS_SNP.GL636486.1005705
  - If your results table looks somewhat different and you cannot easily locate the SNP mentioned above – can you think of other ways to locate this SNP within your results? *Hint: Click Add Step and look up the SNP by its ID.*

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.

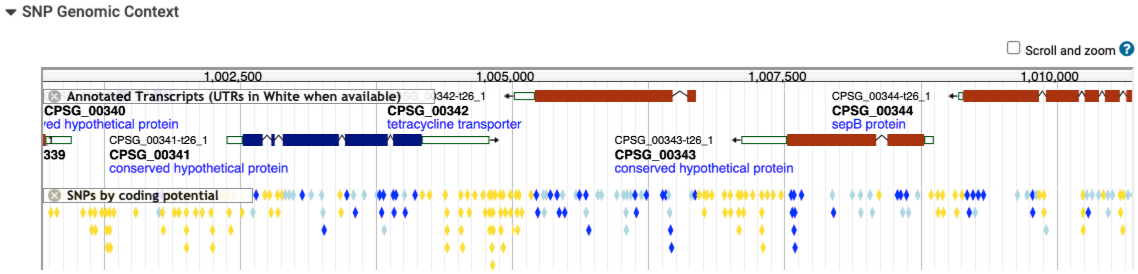


Add to basket Add to favorites Download SNP

SNP: NGS_SNP.GL636486.1005705

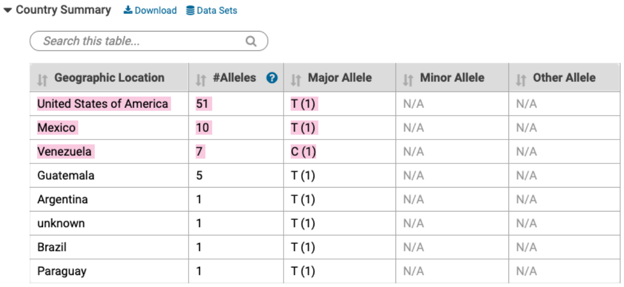
Organism: Coccidioides posadasii str. Silveira
Location: GL636486: 1,005,705
Type: coding
Number of Strains: 77
Gene ID: CPSG_00342
Gene Strand: reverse
Major Allele: T (0.91)
Minor Allele: C (0.09)
Distinct Allele Count: 2
Reference Allele: T
Reference Product: Y 282
Allele (gene strand): A
SNP context: GCTGCTGAGTGTGCGGGAGATATTTGGGAGTAGAGTGTGGCTGTGAGGAAAGGGAGAGAGA
SNP context (gene strand): TCTCTCTCCCTTTCCTCACAGCCACACTCTACTCCCAAATATCTCCCGCACACTCAGCAGC

Genomic location, SNP type and aligned reads are also displayed in JBrowse:



SNPs are denoted by diamonds that are colored based on the coding potential under DNA polymorphism in the Genetic variation section (see pre-workshop module for more information).

Examine SNP record page further. Note that in addition to the US, Mexico, and Venezuela isolates, the SNP records table also contains information for other isolates collected elsewhere.

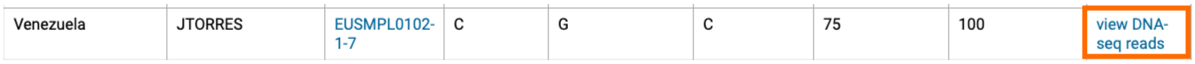


Country Summary Download Data Sets

Search this table...

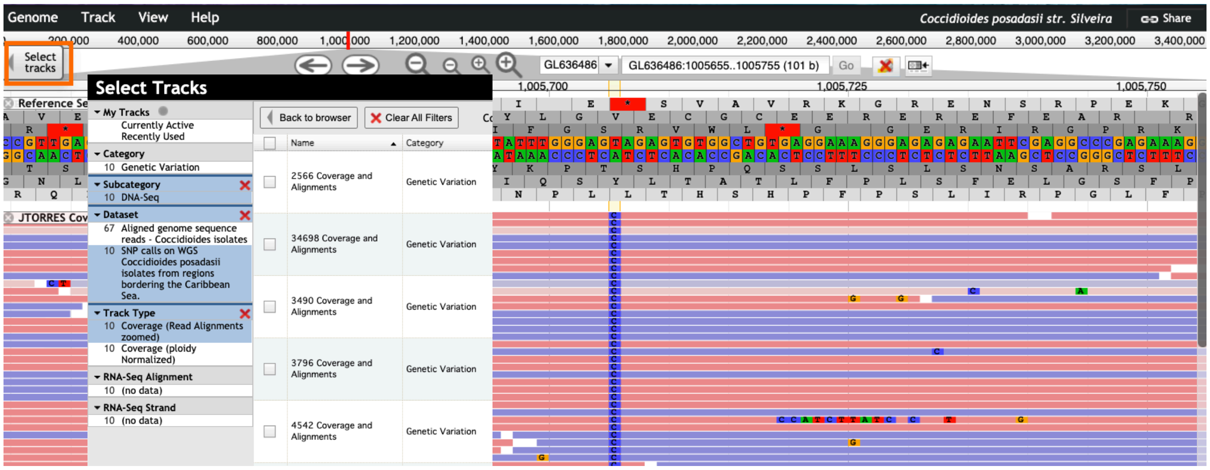
Geographic Location	#Alleles	Major Allele	Minor Allele	Other Allele
United States of America	51	T (1)	N/A	N/A
Mexico	10	T (1)	N/A	N/A
Venezuela	7	C (1)	N/A	N/A
Guatemala	5	T (1)	N/A	N/A
Argentina	1	T (1)	N/A	N/A
unknown	1	T (1)	N/A	N/A
Brazil	1	T (1)	N/A	N/A
Paraguay	1	T (1)	N/A	N/A

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

This action will re-direct you to a JBrowse session where you can select even more isolate tracks by clicking on the Select Tracks tab on the left.



Select Tracks

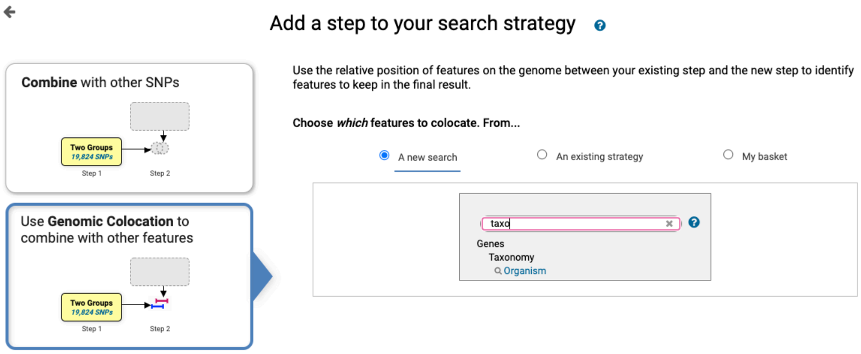
- My Tracks
  - Currently Active
  - Recently Used
- Category
  - 10 Genetic Variation
- Subcategory
  - 10 DNA-Seq
- Dataset
  - 67 Aligned genome sequence reads - Coccidioides isolates
  - 10 SNP calls on WGS Coccidioides posadasii isolates from regions bordering the Caribbean Sea.
- Track Type
  - 10 Coverage (Read Alignments zoomed)
  - 10 Coverage (ploidy Normalized)
- RNA-Seq Alignment
  - 10 (no data)
- RNA-Seq Strand
  - 10 (no data)

Back to browser | Clear All Filters

	Name	Category
□	2566 Coverage and Alignments	Genetic Variation
□	34698 Coverage and Alignments	Genetic Variation
□	3490 Coverage and Alignments	Genetic Variation
□	3796 Coverage and Alignments	Genetic Variation
□	4542 Coverage and Alignments	Genetic Variation

b) Determine genes that map to the SNPs identified in Step 1.

- Add Step and use Genomic Colocation search to combine the results in Step 1 with organism search in Step 2:



Add a step to your search strategy

Combine with other SNPs

Use **Genomic Colocation** to combine with other features

Use the relative position of features on the genome between your existing step and the new step to identify features to keep in the final result.

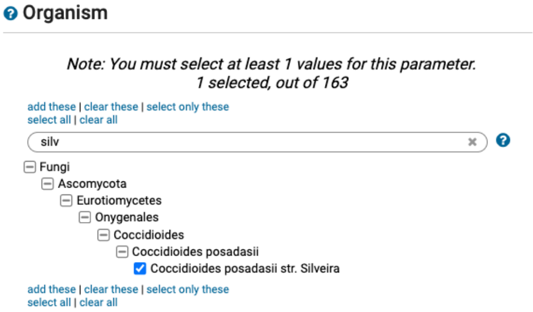
Choose *which* features to colocate. From...

● A new search ○ An existing strategy ○ My basket

taxo

Genes
Taxonomy
Organism

- Next window will bring up an organism selection window, choose Silveira strain.



Organism

Note: You must select at least 1 values for this parameter.
1 selected, out of 163

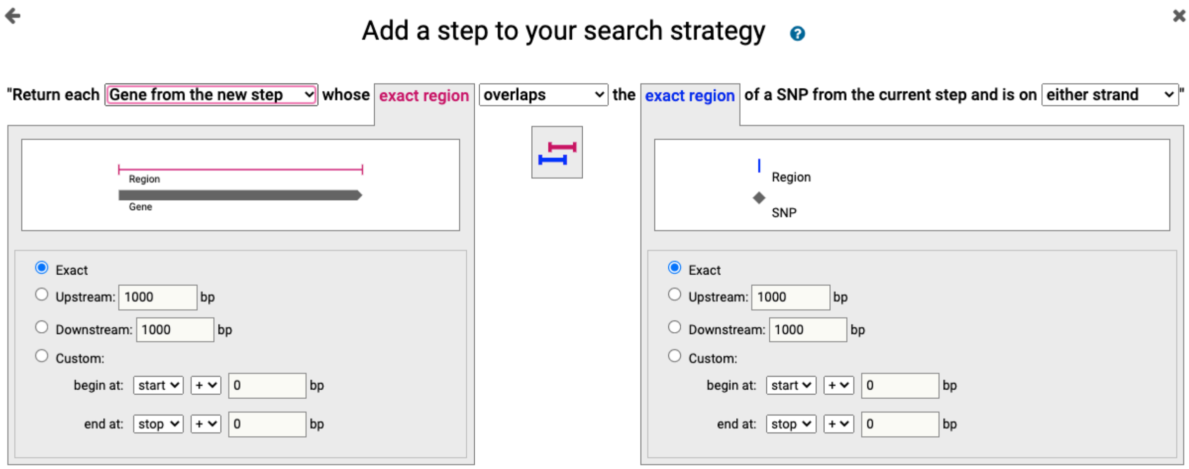
add these | clear these | select only these
select all | clear all

silv

- Fungi
  - Ascomycota
    - Eurotiomycetes
      - Onygenales
        - Coccidioides
          - Coccidioides posadasii
            - ☑ Coccidioides posadasii str. Silveira

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

- Next, set up your colocation parameters and Choose to Return each *Gene from the new Step* whose <u>exact region</u> *overlaps* the <u>exact region</u> of a SNP in Step 1 and is on *either strand*



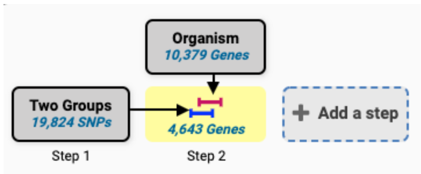
Add a step to your search strategy

"Return each Gene from the new step whose exact region overlaps the exact region of a SNP from the current step and is on either strand"

● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start + 0 bp
end at: stop + 0 bp

● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start + 0 bp
end at: stop + 0 bp

- Examine your results. How many gene were identified in your search?

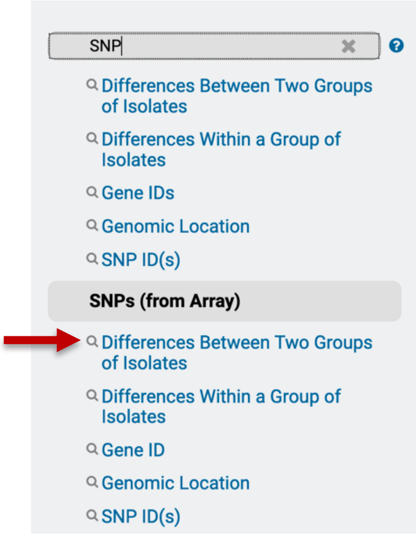


- How can you analyze this data further?

Hint: you can extract genes that have *hypothetical* in the product description via the *Text* search. *You can also perform GO enrichment or identify orthologs in other species, or map to metabolic pathways etc., or you can use other resources as shown previously to cross reference the integrated data. In addition, you may also run a SNP search* <u>*within*</u> *a group of isolates to identify heterozygous or homozygous SNPs…*

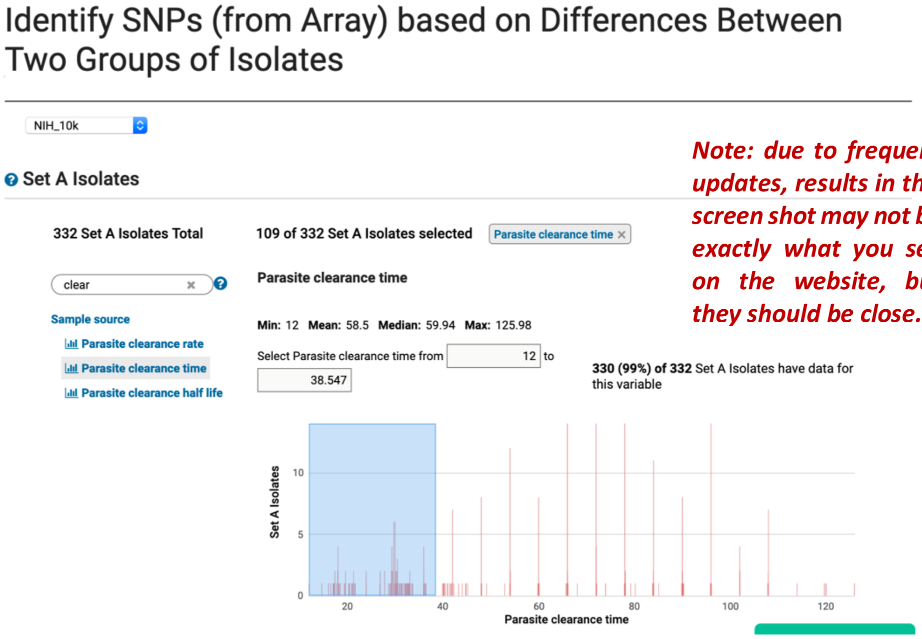
3. **Identify SNPs that distinguish parasites with rapid clearance times following treatment with the anti-malarial drug Artesunate vs. those that have delayed clearance times.** We have a published study in PlasmoDB (Takala-Harrison et. al.) with sufficient meta-data about the samples to ask this interesting question.
   For this exercise use http://PlasmoDB.org

Navigate to the "Differences between two groups of isolates" search under "Search for SNPs (from Array).


SNP
Differences Between Two Groups of Isolates
Differences Within a Group of Isolates
Gene IDs
Genomic Location
SNP ID(s)
SNPs (from Array)
Differences Between Two Groups of Isolates
Differences Within a Group of Isolates
Gene ID
Genomic Location
SNP ID(s)

a. Unlike re-sequencing experiments that can identify any SNPs in the sequence, SNP-Chips have a pre-determined set of SNPs that are assayed and there are multiple different Chips on which these assays can be run. For this study, the authors used the NIH_10K Chip, an array with approximately 10,000 SNPs of which ~8000 can be assayed. Choose this in the Isolate assay type parameter.

b. Once this is done, an interesting set of characteristics are seen in the parameters to choose isolates. In addition to geographic location, there are clinical parameters like Clearance Time, Parasitemia levels, etc. In this exercise we want to identify SNPs that distinguish parasites with rapid clearance times from those with delayed clearance times but you could try other



Identify SNPs (from Array) based on Differences Between Two Groups of Isolates

NIH_10k

Set A Isolates

332 Set A Isolates Total

109 of 332 Set A Isolates selected Parasite clearance time ×

clear

Sample source
- Parasite clearance rate
- Parasite clearance time
- Parasite clearance half life

Parasite clearance time

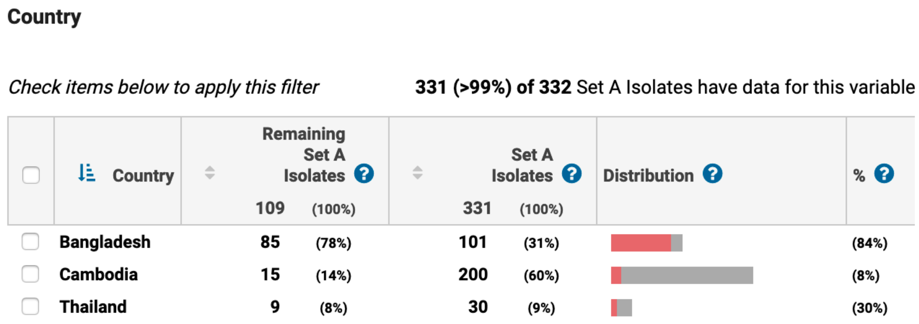
Min: 12 **Mean:** 58.5 **Median:** 59.94 **Max:** 125.98

Select Parasite clearance time from 12 to 38.547

330 (99%) of 332 Set A Isolates have data for this variable

Note: due to frequent updates, results in this screen shot may not be exactly what you see on the website, but they should be close.

possibilities once you are finished. In Set A Isolates, click on some of the characteristics to explore the data. Then choose Clearance Time and select 0 – 38 or 39 minutes. Do these rapid clearance samples appear to be evenly distributed geographically? *Hint: click on Geographic Location to view the distribution of these selected samples (pink section of histogram).*



Country

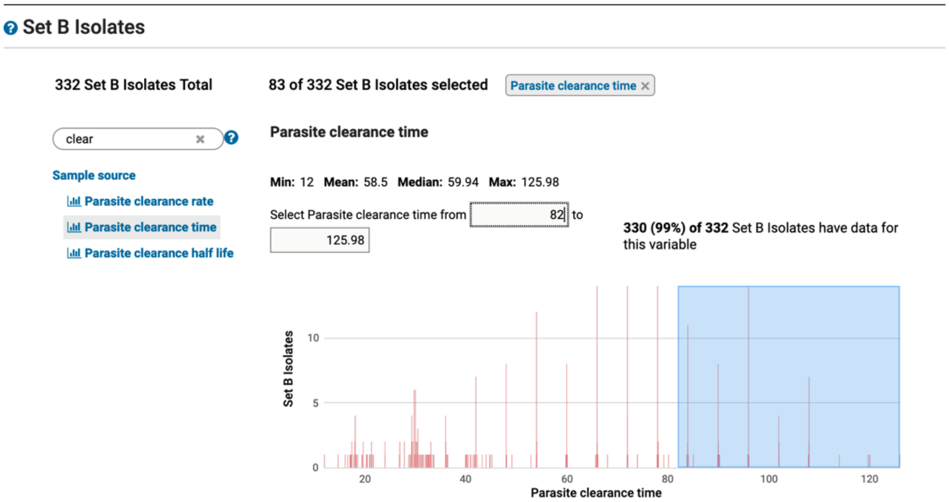
Check items below to apply this filter

331 (>99%) of 332 Set A Isolates have data for this variable

□	Country	Remaining Set A Isolates		Set A Isolates		Distribution	%
		109	(100%)	331	(100%)		
□	Bangladesh	85	(78%)	101	(31%)		(84%)
□	Cambodia	15	(14%)	200	(60%)		(8%)
□	Thailand	9	(8%)	30	(9%)		(30%)

c. We'll keep the defaults of 80 for both Major Allele Frequency and Percent Isolates with Call for this exercise.

d. Now select Clearance times of 82 – end for Set B Isolates. Are these isolates geographically biased?



Set B Isolates

332 Set B Isolates Total

83 of 332 Set B Isolates selected

Parasite clearance time ×

clear ×

Parasite clearance time

Min: 12 **Mean:** 58.5 **Median:** 59.94 **Max:** 125.98

Select Parasite clearance time from 82 to 125.98

330 (99%) of 332 Set B Isolates have data for this variable

e. Keep defaults for Major Allele and Percent with call and run the search. How many SNPs did you find?

A gene (Kelch13) has been identified that is involved in Artemesinin resistance in South East Asia. Is one or more of your SNPs in the region (+/- 10 KB) of the kelch13 gene? Note that we are not expecting that the SNP would be within the gene as this is a Chip experiment where the SNPs were pre-determined and there may not be a SNP on the array within a particular gene that we care about. However, if there is a haplotype that is being selected for in the presence of artemesinin, any SNPs within that haplotype (region of the genome) should likewise be selected.

Hint: add a step to search for genes by text and search for kelch13. This will require you to use the genomic co-location operation as outlined in exercise 3. Set it up the same way except choose custom and start – 10000, stop + 10000 to define the region.

4. **Using resequencing data to identify regions of copy number variation (CNV) For this exercise use https://toxodb.org**

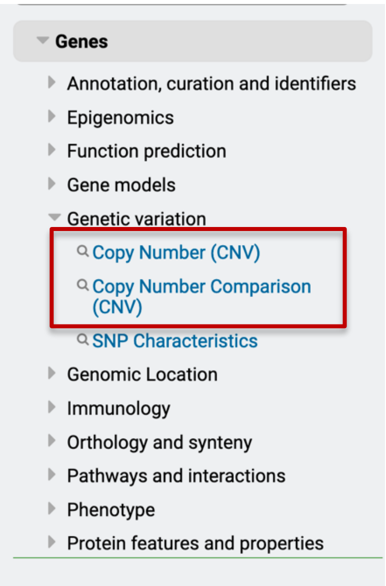
In addition to being useful for variant calling, high throughput sequencing data can be used for determining regions of copy number variation (CNV). All reads in ToxoDB are mapped to the same reference strain ME49, as a result we can estimate a gene's copy number in each of the aligned strains.

The goal of this exercise is to identify

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.

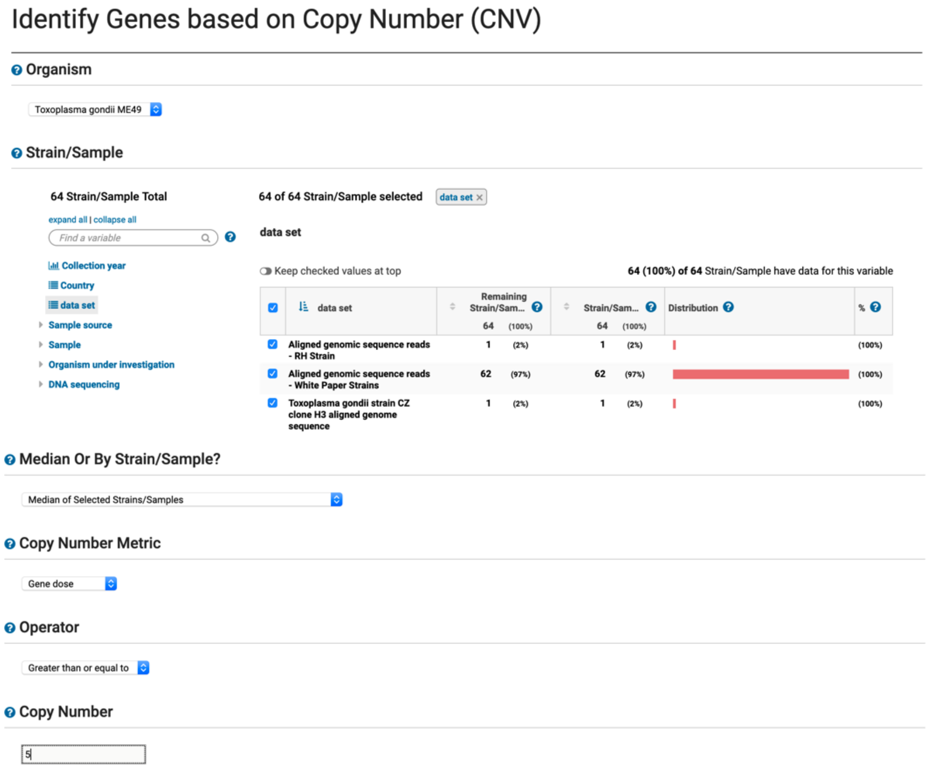
Copy number comparison: This search compares the estimated copy number of a gene in the re-sequenced strain with the copy number in the reference annotation. The copy number in the reference annotation is calculated as the number of genes that are in the same ortholog group as the gene of interest. We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



You have the choice between two different metrics for defining copy number: ***haploid number or gene dose***. 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)

Begin by choosing an Organism (reference genome) and one or more re-sequenced isolates. Choose whether you want to apply your search criteria to individual samples or to the median of your chosen samples. Then choose your Metric, Operator and Copy Number, and initiate the search by clicking the GET ANSWER button. Genes returned by the search will have a copy number based on your chosen metric within the range that you specified. For example, searching with the haploid number equal to 4 will return genes with 4 copies on a chromosome.

a. Use the copy number search to identify genes that are present at a copy number great than 5. Set up the copy number search to include all available isolates/strains, select the median of selected strains/samples, use Gene Dose for copy number metric and set the copy number to 5.



Identify Genes based on Copy Number (CNV)

Organism

Toxoplasma gondii ME49

Strain/Sample

64 Strain/Sample Total

64 of 64 Strain/Sample selected data set ×

data set

Keep checked values at top

64 (100%) of 64 Strain/Sample have data for this variable

☑	data set	Remaining Strain/Sam... 64 (100%)	Strain/Sam... 64 (100%)	Distribution	%
☑	Aligned genomic sequence reads - RH Strain	1 (2%)	1 (2%)		(100%)
☑	Aligned genomic sequence reads - White Paper Strains	62 (97%)	62 (97%)		(100%)
☑	Toxoplasma gondii strain CZ clone H3 aligned genome sequence	1 (2%)	1 (2%)		(100%)

Median Or By Strain/Sample?

Median of Selected Strains/Samples

Copy Number Metric

Gene dose

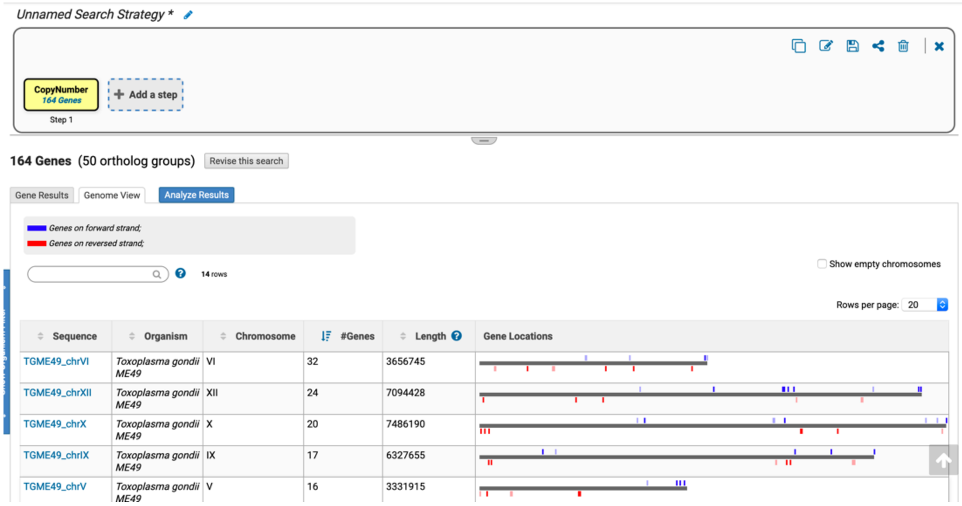
Operator

Greater than or equal to

Copy Number

5

How many genes did you get? Are any of these genes clustered in the same location? (*hint*: click on the “Genome view” tab and examine the red and blue lines in the gene location column – wider lines indicate more than one gene in that location, click on the line to view what is there).



*Unnamed Search Strategy **

CopyNumber
164 Genes
Step 1

164 Genes (50 ortholog groups) Revise this search

🟦 *Genes on forward strand;*
🟥 *Genes on reversed strand;*

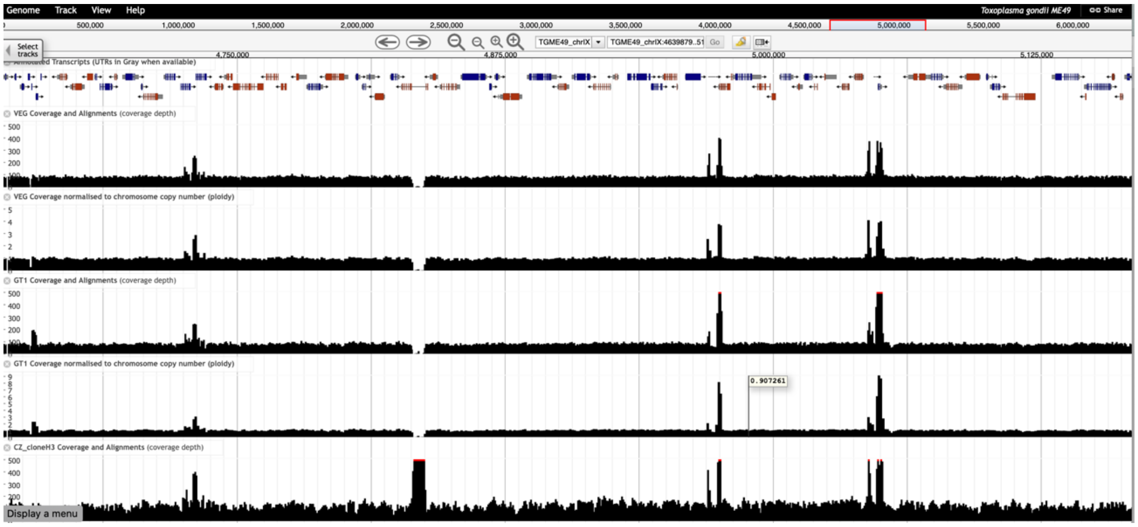
14 rows

□ Show empty chromosomes

Rows per page: 20

Sequence	Organism	Chromosome	#Genes	Length	Gene Locations
TGME49_chrVI	*Toxoplasma gondii ME49*	VI	32	3656745	
TGME49_chrXII	*Toxoplasma gondii ME49*	XII	24	7094428	
TGME49_chrX	*Toxoplasma gondii ME49*	X	20	7486190	
TGME49_chrIX	*Toxoplasma gondii ME49*	IX	17	6327655	
TGME49_chrV	*Toxoplasma gondii ME49*	V	16	3331915	

What happens if you edit this step and change the “Median Or By Strain/Sample” parameter to “By Strain/Sample (at least one selected strain/sample meets criteria)”? Do you get more or less genes? Which genes have the highest CNV? (*hint*: sort the median gene dose column from highest to lowest). Is this what you expected? Does the coverage of reads from resequenced strains aligned to the reference support this conclusion? Here is a link to a JBrowse view with some of the reseqeunced strain coverage data turned on: https://tinyurl.com/2yweuthr

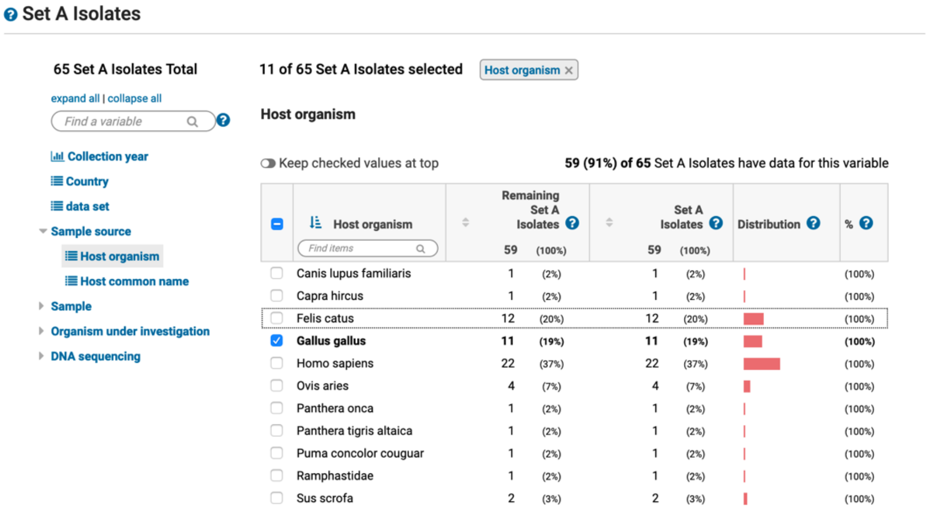


Additional optional exedrcises

5. **Find SNPs that distinguish *Toxoplasma gondii* strains isolated from chickens as compared to those isolated from cats. *NOTE: This exercise in ToxoDB explores the hypothesis that we can identify SNPs/genes involved in T. gondii host preference.***

Navigate to "Identify SNPs based on Differences Between Two Groups of Isolates".

b. Click select set A isolates and select hosts from the left column. Check the chicken (*Gallus gallus*) box to select the 11 chicken isolates.



Set A Isolates

65 Set A Isolates Total

expand all | collapse all

Find a variable

- Collection year
- Country
- data set
- Sample source
  - Host organism
  - Host common name
- Sample
- Organism under investigation
- DNA sequencing

11 of 65 Set A Isolates selected   Host organism ×

Host organism

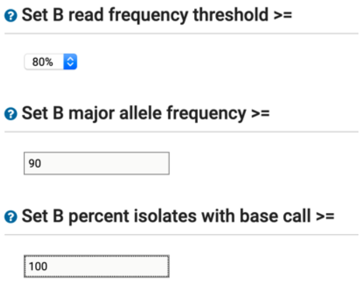
Keep checked values at top   59 (91%) of 65 Set A Isolates have data for this variable

▬	Host organism	Remaining Set A Isolates		Set A Isolates		Distribution	%
	Find items	59	(100%)	59	(100%)		
□	Canis lupus familiaris	1	(2%)	1	(2%)	▬	(100%)
□	Capra hircus	1	(2%)	1	(2%)	▬	(100%)
□	Felis catus	12	(20%)	12	(20%)	▬	(100%)
☑	**Gallus gallus**	**11**	**(19%)**	**11**	**(19%)**	▬	**(100%)**
□	Homo sapiens	22	(37%)	22	(37%)	▬	(100%)
□	Ovis aries	4	(7%)	4	(7%)	▬	(100%)
□	Panthera onca	1	(2%)	1	(2%)	▬	(100%)
□	Panthera tigris altaica	1	(2%)	1	(2%)	▬	(100%)
□	Puma concolor couguar	1	(2%)	1	(2%)	▬	(100%)
□	Ramphastidae	1	(2%)	1	(2%)	▬	(100%)
□	Sus scrofa	2	(3%)	2	(3%)	▬	(100%)

c. Click select set B isolates and select hosts from the left column. Check the cat (*Felis catus*) box to select the 12 cat isolates.

d. Let's run a very stringent search and change the "major allele frequency" parameters for both sets to 90. (*What does that mean?)*. Also, set the isolates with base call parameter to 100 for both sets A and B.

- How many SNPs did your search return? Does this large number that distinguish these two fairly large groups of isolates surprise you?



Set B read frequency threshold >=

80%

Set B major allele frequency >=

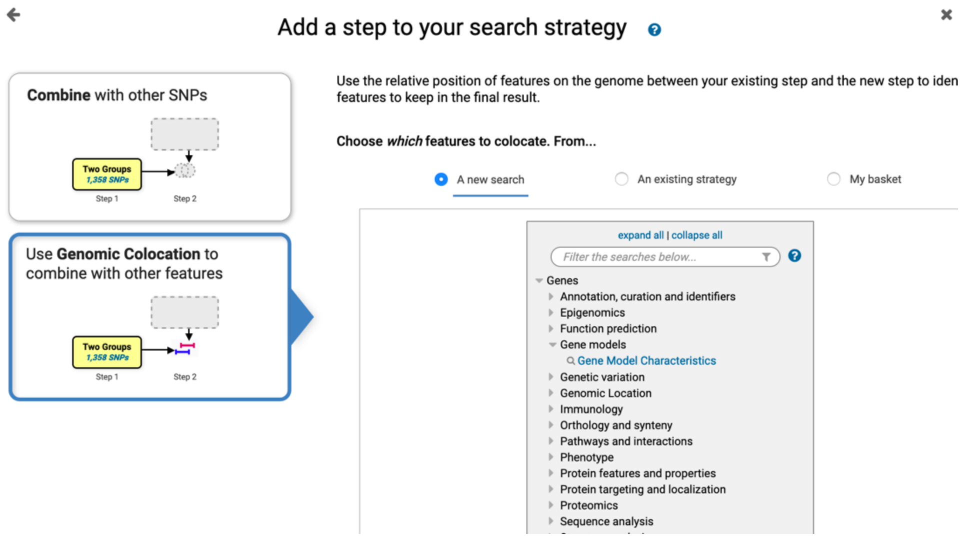
90

Set B percent isolates with base call >=

100

You want to identify genes that could potentially be involved in host preference in *Toxoplasma gondii* and you expect that the SNPs from this search you just ran may be in protein coding regions of genes involved in this preference. How might you identify genes containing these SNPs?

e. Add a step to identify *protein-coding genes* in *Toxoplasma gondii ME49*. Select the “Use Genomic Colocation…” option. Then select the gene search called “Gene Model Characteristics”.



Add a step to your search strategy

Combine with other SNPs

Use **Genomic Colocation** to combine with other features

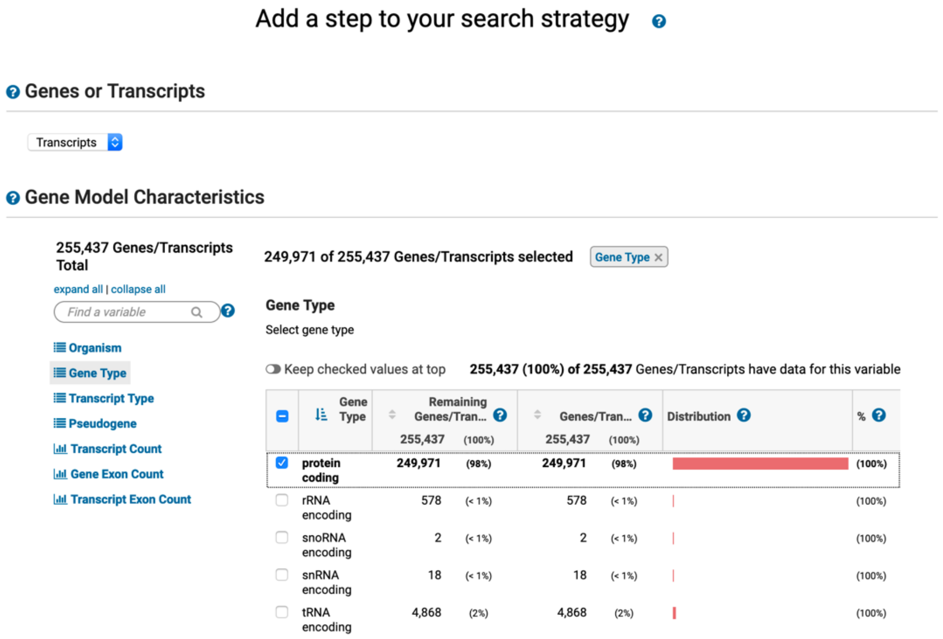
Use the relative position of features on the genome between your existing step and the new step to iden features to keep in the final result.

Choose *which* features to colocate. From...

● A new search ○ An existing strategy ○ My basket

expand all | collapse all
Filter the searches below...
Genes
Annotation, curation and identifiers
Epigenomics
Function prediction
Gene models
Gene Model Characteristics
Genetic variation
Genomic Location
Immunology
Orthology and synteny
Pathways and interactions
Phenotype
Protein features and properties
Protein targeting and localization
Proteomics
Sequence analysis

f. Configure the gene model characteristics search to find protein coding genes



Add a step to your search strategy

Genes or Transcripts

Transcripts

Gene Model Characteristics

255,437 Genes/Transcripts Total

249,971 of 255,437 Genes/Transcripts selected Gene Type ×

Gene Type

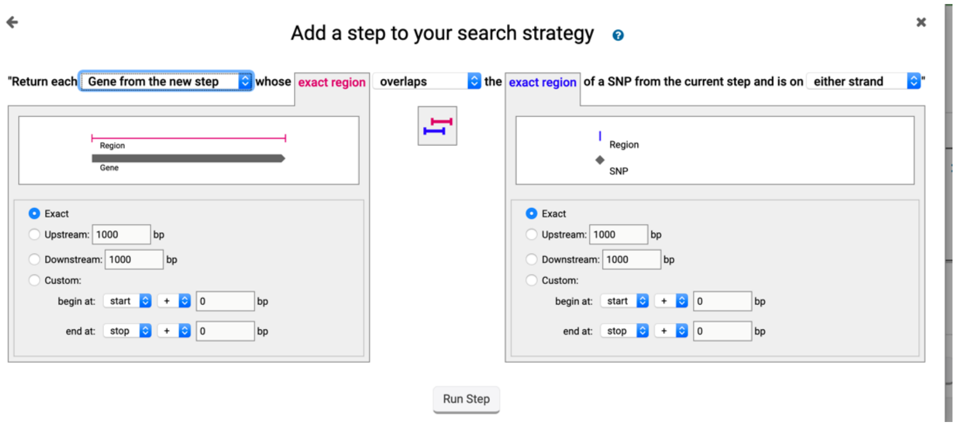
Select gene type

Keep checked values at top **255,437 (100%) of 255,437** Genes/Transcripts have data for this variable

	Gene Type	Remaining Genes/Tran...		Genes/Tran...		Distribution	%
		255,437	(100%)	255,437	(100%)		
☑	**protein coding**	**249,971**	**(98%)**	**249,971**	**(98%)**		**(100%)**
□	rRNA encoding	578	(< 1%)	578	(< 1%)		(100%)
□	snoRNA encoding	2	(< 1%)	2	(< 1%)		(100%)
□	snRNA encoding	18	(< 1%)	18	(< 1%)		(100%)
□	tRNA encoding	4,868	(2%)	4,868	(2%)		(100%)

only.

g. Configure the genome colocation page to return “Gene from Step 2 whose exact region overlaps the exact region of a SNP in Step 1 and is on either strand”



Add a step to your search strategy

"Return each Gene from the new step whose exact region overlaps the exact region of a SNP from the current step and is on either strand"

● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start + 0 bp
end at: stop + 0 bp

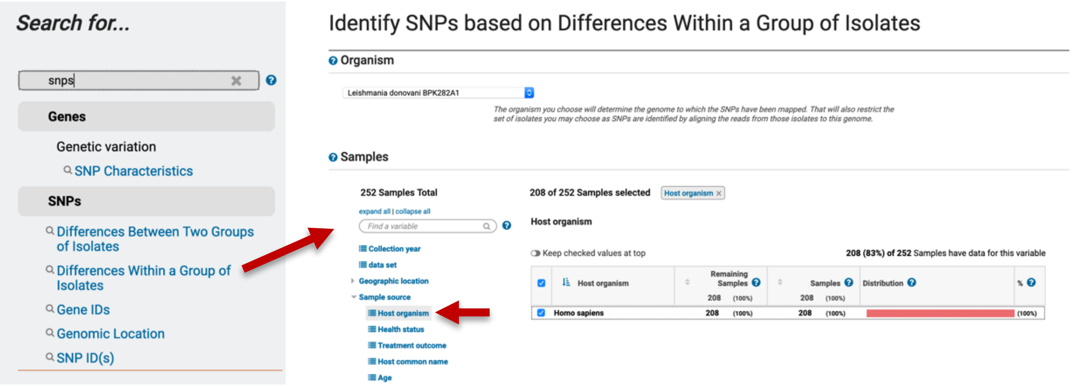
● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start + 0 bp
end at: stop + 0 bp

- How many genes are returned?
- What is the gene that contains the most SNPs on your list? *Hint: sort the list high to low by match count.*
- Does this gene have orthologs in other species from ToxoDB? *Hint: go to the gene page and look at the genomic context and orthologs/paralogs in ToxoDB table.*
- Does it have orthology in any other species? *Hint: click on the link under the orthologs table and look at in OrthoMCL.*
- What does this say about this gene? How can you follow up on what role this gene may be playing for the organism? *Hint: you are a biologist and will need to look at the data on the gene record page and interpret it based on your experience and intuition.*
- Do these genes appear to be randomly distributed along the genome? *Hint: click the “Genome View” tab to view the distribution.* If you are a *Toxoplasma* biologist, do you have any hypotheses why the distribution may be skewed?
  As a last resort: https://toxodb.org/toxo/im.do?s=4fe2f7409d4ba4d6

6. Identifying SNPs within a group of isolates

For this exercise use https://tritrypdb.org

a. Go to the “Differences Within a Group of Isolates” search.
Hint: you can find this under the “SNPs” category (remember you can filter the searches by typing a key word like “snps” in the filter box.



Search for...

snps

Genes

Genetic variation

SNP Characteristics

SNPs

Differences Between Two Groups of Isolates

Differences Within a Group of Isolates

Gene IDs

Genomic Location

SNP ID(s)

Identify SNPs based on Differences Within a Group of Isolates

Organism

Leishmania donovani BPK282A1

The organism you choose will determine the genome to which the SNPs have been mapped. That will also restrict the set of isolates you may choose as SNPs are identified by aligning the reads from those isolates to this genome.

Samples

252 Samples Total

expand all | collapse all

Find a variable

Collection year
data set
Geographic location
Sample source
Host organism
Health status
Treatment outcome
Host common name
Age

208 of 252 Samples selected Host organism ×

Host organism

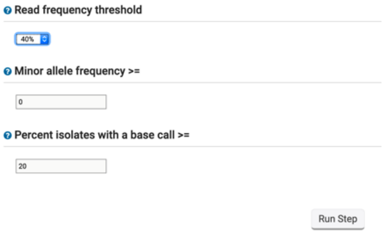
Keep checked values at top

208 (83%) of 252 Samples have data for this variable

☑	Host organism	Remaining Samples 208 (100%)	Samples 208 (100%)	Distribution	%
☑	Homo sapiens	208 (100%)	208 (100%)	🟥	(100%)

b. What does this search do? Choose *Leishmania donovani* for the organism and select isolates from the human host. Use default parameters for the rest of the parameters.

Run the query and look at your results.



Read frequency threshold

40%

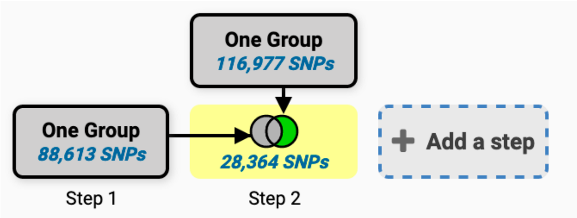
Minor allele frequency >=

0

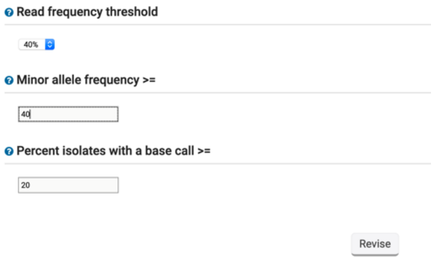
Percent isolates with a base call >=

20

- How many SNPs were returned?
- Are any of these heterozygous SNPs?
- How would you identify heterozygous SNPs? Add a step to your strategy to identify SNPs from these isolates that may be heterozygous. *Hint: choose a read frequency threshold of 40% and select the 2 minus 1 operation.*
- How many SNPs did you identify?

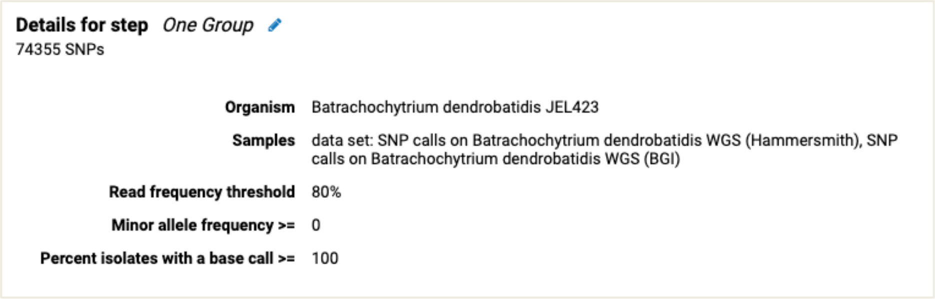


- Click on the second step results to view them. What do you notice about the %minor alleles? (*many are quite low … i.e. in one or two of the isolates*). How can you remove these from your search results? *Hint: revise this search and increase the minor allele frequency threshold (try 20 and 40 and compare results).*
- Why might you want to increase the minor allele threshold when you run SNP searches?
- Try increasing / decreasing the "Percent isolates with base call". How does this impact your results? Why might you want to change this parameter?
- Go to a record page for a SNP with a high minor allele frequency. What do you see in the Strains table? Why are many of the strains repeated?


Read frequency threshold
40%
Minor allele frequency >=
40
Percent isolates with a base call >=
20

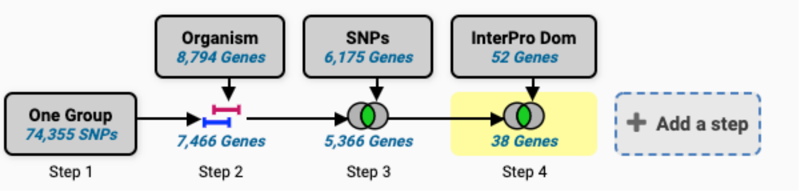
7. Identify SNPs within a group of isolates

- Deploy the SNP search called "Differences Within a Group of Isolates"
- Look for homozygous SNPs in Batrachochytrium dendrobatidis WGS (Hammersmith). For example, here is one way to set your search:


Details for step *One Group*
74355 SNPs

Organism	Batrachochytrium dendrobatidis JEL423
Samples	data set: SNP calls on Batrachochytrium dendrobatidis WGS (Hammersmith), SNP calls on Batrachochytrium dendrobatidis WGS (BGI)
Read frequency threshold	80%
Minor allele frequency >=	0
Percent isolates with a base call >=	100

Batrachochytrium dendrobatidis (Bd) causes chytridiomycosis in amphibians. Next combine your search of homozygous mutation that arose across all isolates in this study to map SNPs to *Bd* genes (Step 2; Hint: colocation tool), identify genes that carry non-synonymous mutations (Step 3; Hint: requires SNP Characteristics search), and look for ABC-transporters (Step 4; Hint: Requires InterPro Domain search; this example uses PF00005)



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.*

8. Use resequencing data to identify regions of copy number variation (CNV)

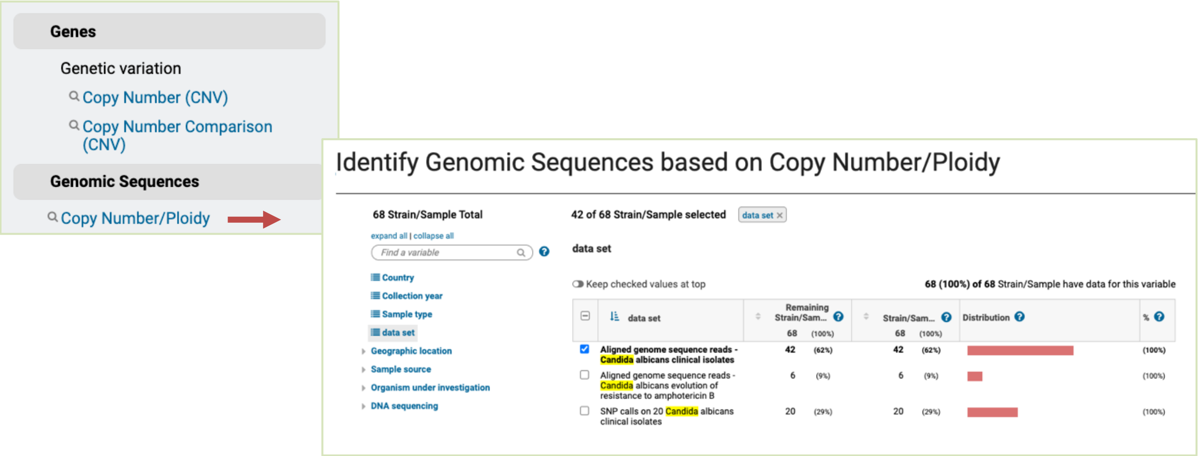
In addition to being useful for variant calling, high throughput sequencing data can be used for determining regions of copy number variation (CNV). All reads in FungiDB are mapped to the same reference strain as SNP datasets and, as a result, we can estimate a gene's copy number in each of the aligned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

One of the datasets we have loaded is isolates from *Candida albicans* clinical isolates described in this paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383195/ The data on aneuploidy is shown in figure 4: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383195/figure/fig4/

a. Find trisomic chromosomes.

- Use the Genomic Sequences by Copy Number/Ploidy search, select Candida albicans, and choose the dataset titled "Aligned genome sequence reads – Candida albicans clinical isolates".

Copy Number/Ploidy: 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. For example, to find supernumerary chromosomes in a diploid organism, search for genomic sequences where the Copy Number is >= 3.



Identify Genomic Sequences based on Copy Number/Ploidy

68 Strain/Sample Total

42 of 68 Strain/Sample selected data set ×

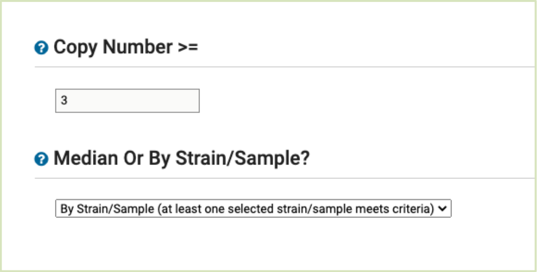
data set

Keep checked values at top

68 (100%) of 68 Strain/Sample have data for this variable

	data set	Remaining Strain/Sam... 68 (100%)		Strain/Sam... 68 (100%)		Distribution	%
☑	Aligned genome sequence reads - Candida albicans clinical isolates	42	(62%)	42	(62%)		(100%)
□	Aligned genome sequence reads - Candida albicans evolution of resistance to amphotericin B	6	(9%)	6	(9%)		(100%)
□	SNP calls on 20 Candida albicans clinical isolates	20	(29%)	20	(29%)		(100%)

- Set search criteria:



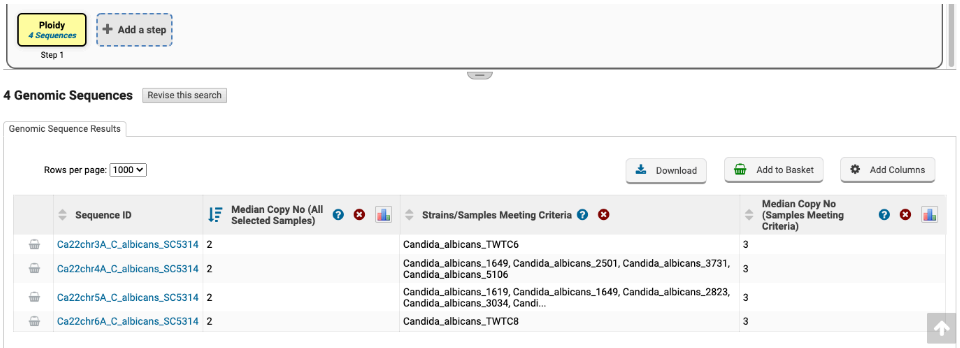
Copy Number >=

3

Median Or By Strain/Sample?

By Strain/Sample (at least one selected strain/sample meets criteria)

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).



4 Genomic Sequences Revise this search

Genomic Sequence Results

Rows per page: 1000

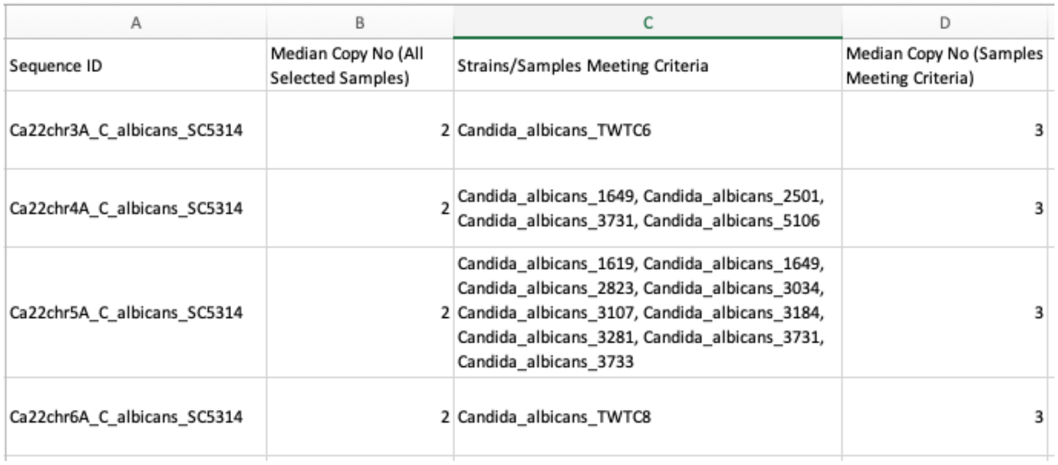
Sequence ID	Median Copy No (All Selected Samples)	Strains/Samples Meeting Criteria	Median Copy No (Samples Meeting Criteria)
Ca22chr3A_C_albicans_SC5314	2	Candida_albicans_TWTC6	3
Ca22chr4A_C_albicans_SC5314	2	Candida_albicans_1649, Candida_albicans_2501, Candida_albicans_3731, Candida_albicans_5106	3
Ca22chr5A_C_albicans_SC5314	2	Candida_albicans_1619, Candida_albicans_1649, Candida_albicans_2823, Candida_albicans_3034, Candi...	3
Ca22chr6A_C_albicans_SC5314	2	Candida_albicans_TWTC8	3

b. Explore segmental aneuploidy in JBrowse

In JBrowse we have two coverage tracks:

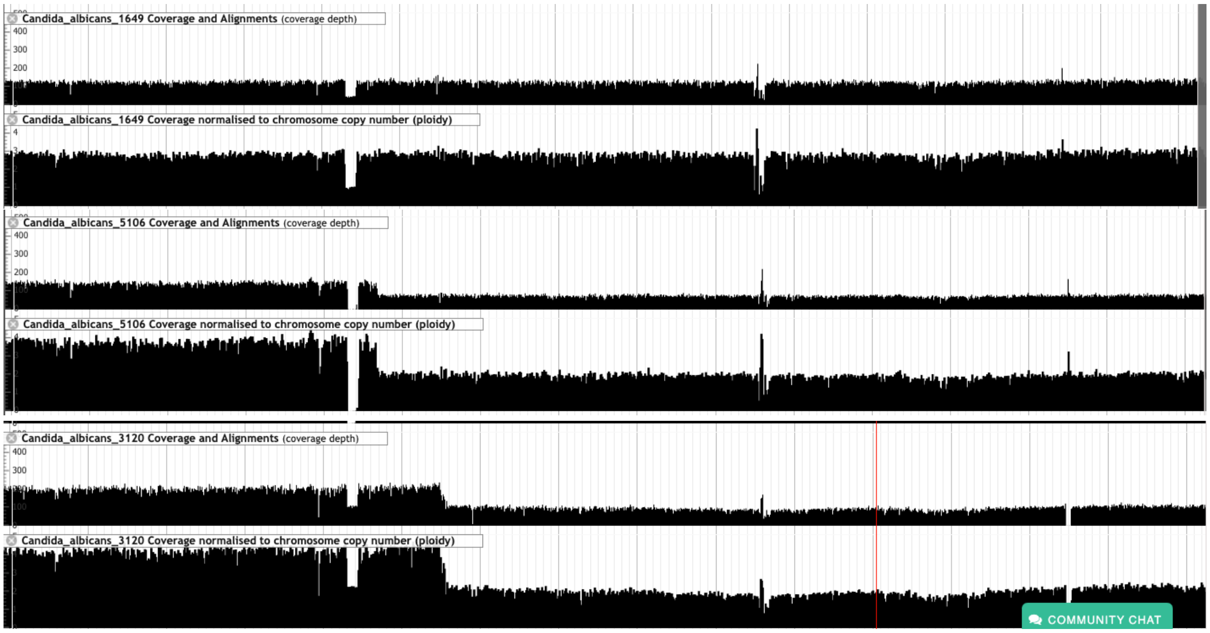
- Raw coverage from the alignment (available for every isolate where we have whole genome sequencing, whether we ran the copy number pipeline or not)
- Normalised coverage in bins (only available for isolates where we have run the copy number pipeline)

  Note: You can download the results as a .tsv file and then open it in Excel to view all results (Hint: Click on the Download button located above the results table and select the first export option from the top)



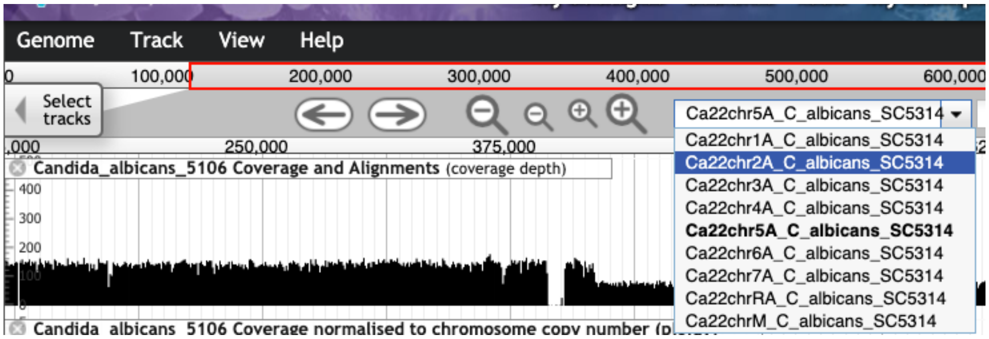
A	B	C	D
Sequence ID	Median Copy No (All Selected Samples)	Strains/Samples Meeting Criteria	Median Copy No (Samples Meeting Criteria)
Ca22chr3A_C_albicans_SC5314	2	Candida_albicans_TWTC6	3
Ca22chr4A_C_albicans_SC5314	2	Candida_albicans_1649, Candida_albicans_2501, Candida_albicans_3731, Candida_albicans_5106	3
Ca22chr5A_C_albicans_SC5314	2	Candida_albicans_1619, Candida_albicans_1649, Candida_albicans_2823, Candida_albicans_3034, Candida_albicans_3107, Candida_albicans_3184, Candida_albicans_3281, Candida_albicans_3731, Candida_albicans_3733	3
Ca22chr6A_C_albicans_SC5314	2	Candida_albicans_TWTC8	3

- Click on one of the Sequence ID Ca22chr5A_C_albicans_SC5314 (in blue) and then click on the View in JBrowse genome browser button.
- When in JBrowse, click on the Select tracks tab to customize your view:
- Select tracks for isolates 1649, 5106, and 3120


Candida_albicans_1649 Coverage and Alignments (coverage depth)
400
300
200
Candida_albicans_1649 Coverage normalised to chromosome copy number (ploidy)
4
Candida_albicans_5106 Coverage and Alignments (coverage depth)
400
300
200
Candida_albicans_5106 Coverage normalised to chromosome copy number (ploidy)
Candida_albicans_3120 Coverage and Alignments (coverage depth)
400
300
Candida_albicans_3120 Coverage normalised to chromosome copy number (ploidy)
COMMUNITY CHAT

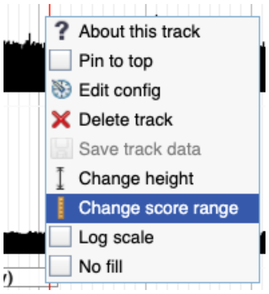
Notice examples of chromosomal (1649) and segmental triploidy (5106,3120). Note that the whole chromosome is shown in both screenshots, and both tracks are shown for each sample. We are not currently normalizing for telomere proximity.

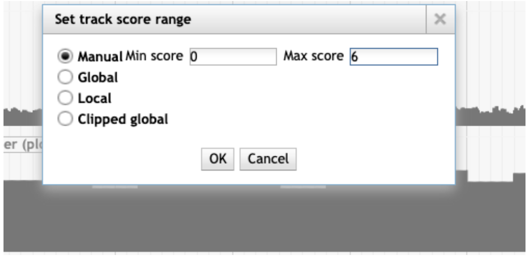
- Switch the JBrowse view to the chromosome 2


Genome
Track
View
Help
100,000
200,000
300,000
400,000
500,000
600,000
Select tracks
Ca22chr5A_C_albicans_SC5314
Ca22chr1A_C_albicans_SC5314
Ca22chr2A_C_albicans_SC5314
Ca22chr3A_C_albicans_SC5314
Ca22chr4A_C_albicans_SC5314
Ca22chr5A_C_albicans_SC5314
Ca22chr6A_C_albicans_SC5314
Ca22chr7A_C_albicans_SC5314
Ca22chrRA_C_albicans_SC5314
Ca22chrM_C_albicans_SC5314
Candida_albicans_5106 Coverage and Alignments (coverage depth)
400
300
200
100
Candida_albicans_5106 Coverage normalised to chromosome copy number

- Notice segmental aneuploidy in the chromosome 2 right arm.

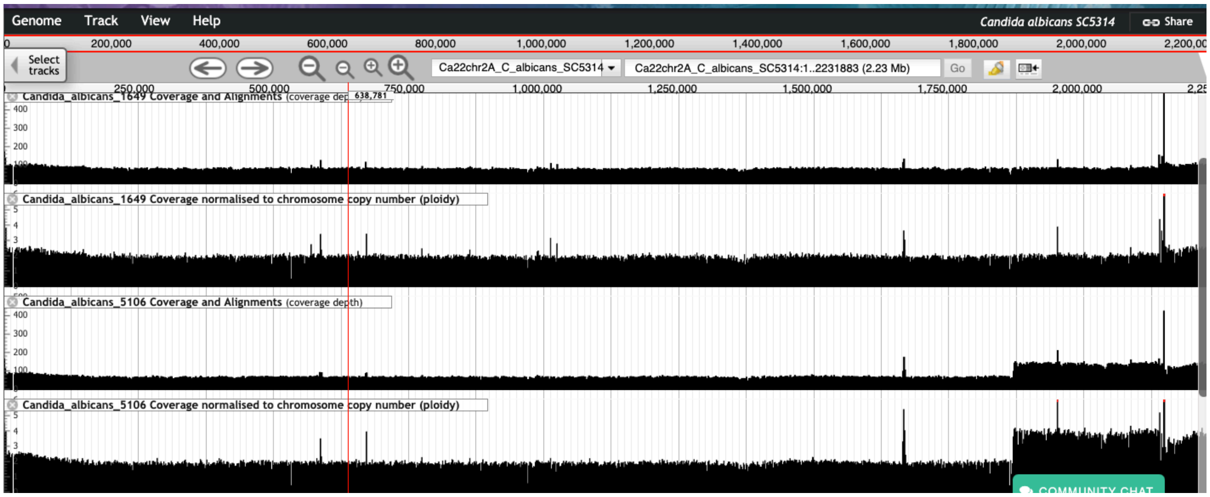
Note: you may need to zoom out and/or adjust settings in the Change Score range track option





● Manual Min score 0 Max score 6

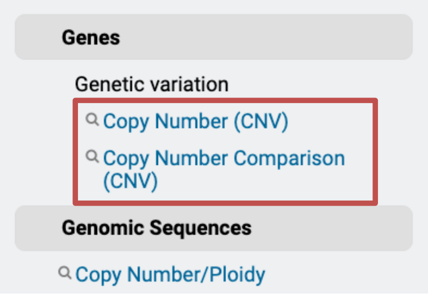
-


Genome
Track
View
Help
Candida albicans SC5314
Share
Select tracks
Ca22chr2A_C_albicans_SC5314
Ca22chr2A_C_albicans_SC5314:1..2231883 (2.23 Mb)
Go
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
1,600,000
1,800,000
2,000,000
250,000
500,000
750,000
1,250,000
1,500,000
1,750,000
Candida_albicans_1649 Coverage and Alignments (coverage depth)
638,781
Candida_albicans_1649 Coverage normalised to chromosome copy number (ploidy)
Candida_albicans_5106 Coverage and Alignments (coverage depth)
Candida_albicans_5106 Coverage normalised to chromosome copy number (ploidy)
400
300
200
5
4
3
COMMUNITY CHAT

Using Gene Searches

Looking through JBrowse is fine if you know what you are looking for, but it can be difficult for data mining. 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.

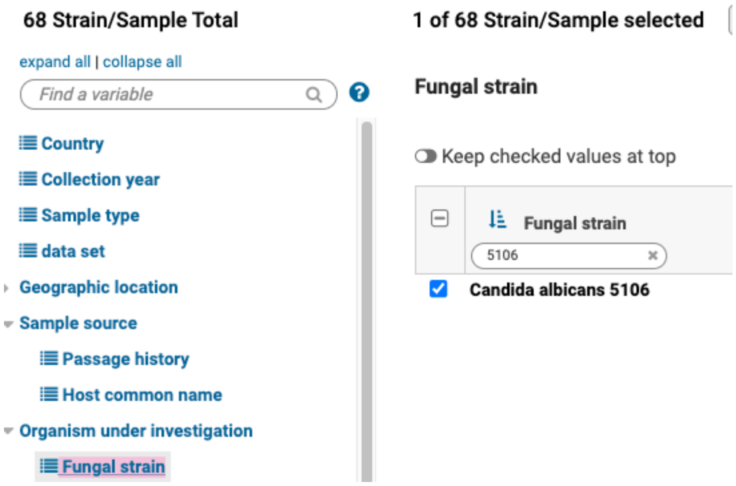
- **Copy number comparison:** This search compares the estimated copy number of a gene in the re-sequenced strain with the copy number in the reference annotation. The copy number in the reference annotation is calculated as the number of genes that are in the same ortholog group as the gene of interest.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

You have the choice between two different metrics for defining copy number: haploid number or gene dose:

- **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)

- To discover regions of potential segmental aneuploidy, use the *Genes by Copy Number Comparison* search to look for genes where the predicted haploid number is *greater than the number of copies in the reference annotation*. For clarity, restrict your search to isolate 5106.



68 Strain/Sample Total

expand all | collapse all

Find a variable

- Country
- Collection year
- Sample type
- data set
- Geographic location
- Sample source
  - Passage history
  - Host common name
- Organism under investigation
  - Fungal strain

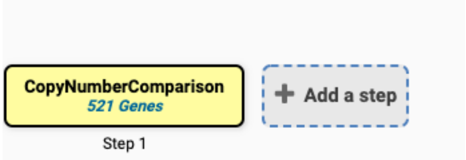
1 of 68 Strain/Sample selected

Fungal strain

Keep checked values at top

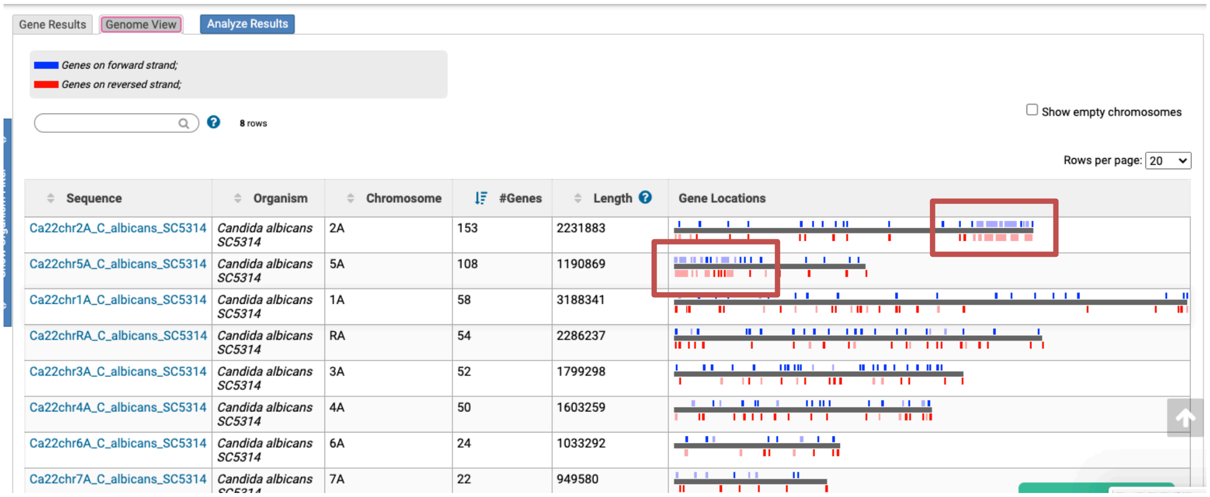
Fungal strain: 5106

☑ Candida albicans 5106



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

- You can export the list of genes and also visualize them in the Genome View, which highlights the locations of hits:



🟦 Genes on forward strand;
🟥 Genes on reversed strand;

8 rows

□ Show empty chromosomes

Rows per page: 20

Sequence	Organism	Chromosome	#Genes	Length	Gene Locations
Ca22chr2A_C_albicans_SC5314	*Candida albicans SC5314*	2A	153	2231883	
Ca22chr5A_C_albicans_SC5314	*Candida albicans SC5314*	5A	108	1190869	
Ca22chr1A_C_albicans_SC5314	*Candida albicans SC5314*	1A	58	3188341	
Ca22chrRA_C_albicans_SC5314	*Candida albicans SC5314*	RA	54	2286237	
Ca22chr3A_C_albicans_SC5314	*Candida albicans SC5314*	3A	52	1799298	
Ca22chr4A_C_albicans_SC5314	*Candida albicans SC5314*	4A	50	1603259	
Ca22chr6A_C_albicans_SC5314	*Candida albicans SC5314*	6A	24	1033292	
Ca22chr7A_C_albicans_SC5314	*Candida albicans SC5314*	7A	22	949580	

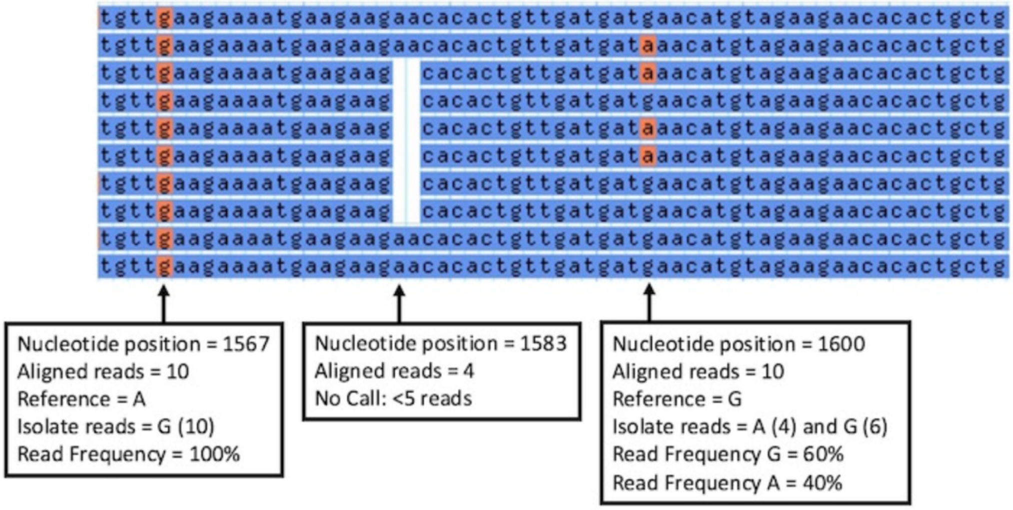
As you can see in the highlighted regions, large numbers of genes that are predicted to have increased copy numbers are clustered at the right-hand end of chromosome 2 and the left-hand end of chromosome 5, corresponding to the segmental aneuploidies shown in the JBrowse session above.

Appendix

Isolates are assayed for SNPs in VEuPathDB by two basic methods: re-sequencing and the alignment of sequence reads to a reference genome or DNA hybridization to a SNP-chip array.

Read Frequency Threshold: Calling SNPs for each isolate in your group. Each isolate's sequencing reads are aligned to a reference genome (Organism) 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. 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.

Isolate X aligned sequencing reads


tgttgaagaaaatgaagaagaacacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaagaacacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaagaacacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaagaacacactgttgatgatgaacatgtagaagaacacactgctg

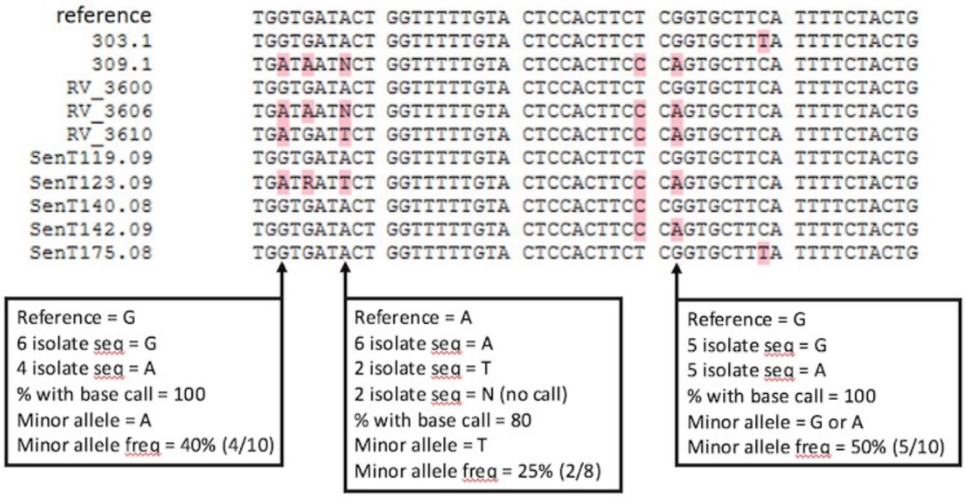
Nucleotide position = 1567
Aligned reads = 10
Reference = A
Isolate reads = G (10)
Read Frequency = 100%

Nucleotide position = 1583
Aligned reads = 4
No Call: <5 reads

Nucleotide position = 1600
Aligned reads = 10
Reference = G
Isolate reads = A (4) and G (6)
Read Frequency G = 60%
Read Frequency A = 40%

Minor allele frequency: Parameter for calling SNPs across your isolate group. 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 GGTTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
    303.1    TGGTGATACT GGTTTTTGTA CTCCACTTCT CGGTGCTTTA TTTTCTACTG
    309.1    TGATAATNCT GGTTTTTGTA CTCCACTTCC CAGTGCTTCA TTTTCTACTG
  RV_3600    TGGTGATACT GGTTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
  RV_3606    TGATAATNCT GGTTTTTGTA CTCCACTTCC CAGTGCTTCA TTTTCTACTG
  RV_3610    TGATGATTCT GGTTTTTGTA CTCCACTTCC CAGTGCTTCA TTTTCTACTG
SenT119.09   TGGTGATACT GGTTTTTGTA CTCCACTTCT CGGTGCTTCA TTTTCTACTG
SenT123.09   TGATRATTCT GGTTTTTGTA CTCCACTTCC CAGTGCTTCA TTTTCTACTG
SenT140.08   TGGTGATACT GGTTTTTGTA CTCCACTTCC CGGTGCTTCA TTTTCTACTG
SenT142.09   TGGTGATACT GGTTTTTGTA CTCCACTTCC CAGTGCTTCA TTTTCTACTG
SenT175.08   TGGTGATACT GGTTTTTGTA CTCCACTTCT CGGTGCTTTA 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: Parameter for calling SNPs across your isolate group

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.