

FungiDB: SNPs and Population Genetics

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 aneuploidy

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 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.

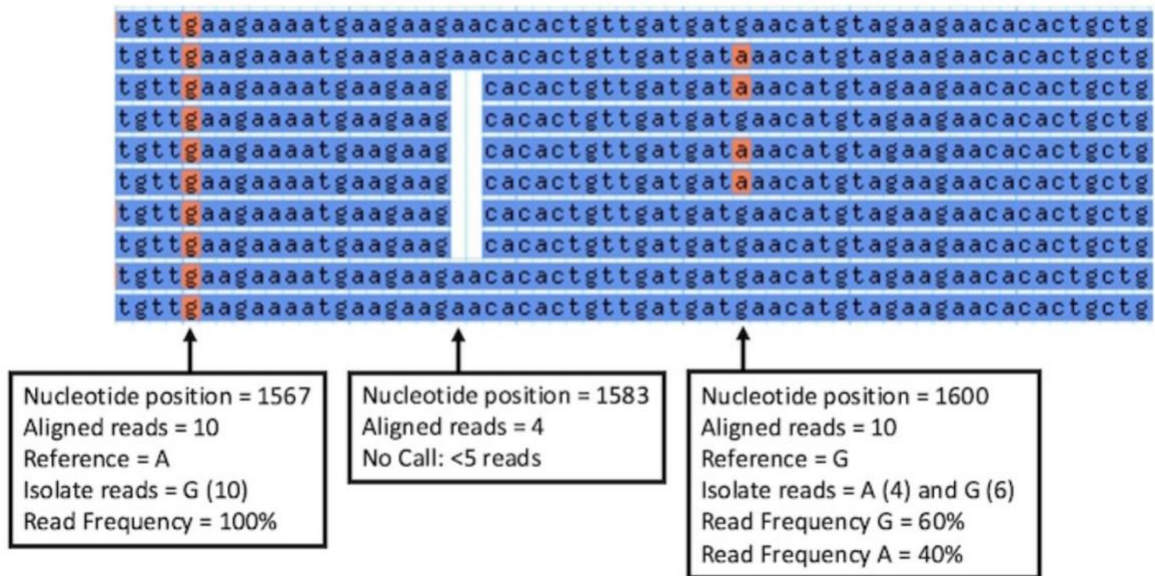
Single Nucleotide Polymorphisms (SNPs) can be used to characterize similarities and differences within a group of isolates or between two groups of isolates. They can also be used to identify genes that may be under evolutionary pressure, either to stay the same (purifying selection) or to change (diversifying or balancing selection).

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.

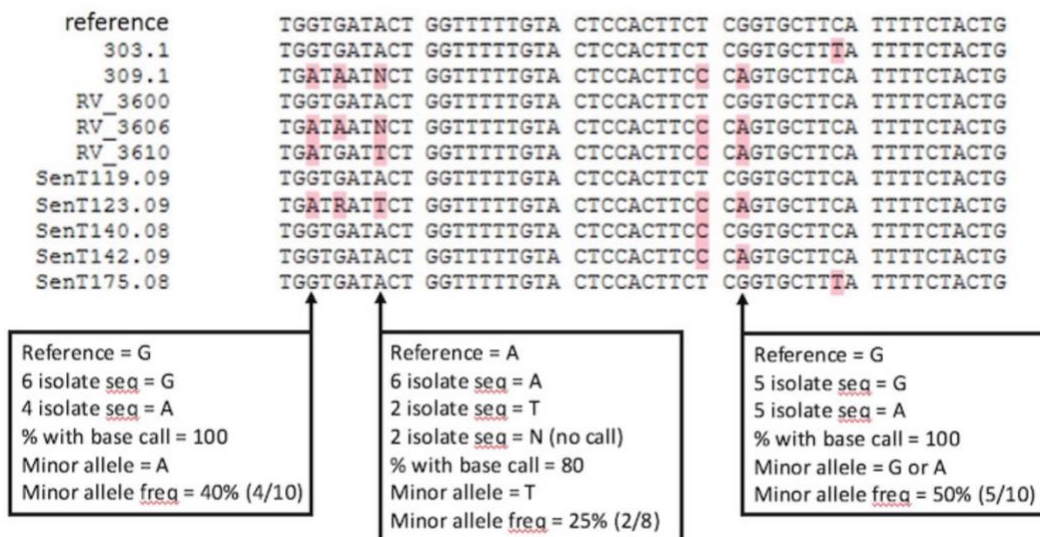
Isolate X aligned sequencing reads



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.



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.

1. Identify *Pyricularia oryzae* genes that contain at least 10 non-synonymous SNPs.

- Start by running a search for genes based on SNP characteristics - this search can be found under the 'Genetic Variation' category.
- Select *Pyricularia oryzae* 70-50 from the drop-down list.
- In the sample section, select the dataset titled “SNP calls on WGS of *Pyricularia oryzae* isolated from Bangladesh in 2016 and 2017”.

Identify Genes based on SNP Characteristics

24 Set of Samples Total expand all | collapse all

Find a variable

Sample type

data set

Organism under investigation

Fungal strain

Fungal organism

DNA sequencing

23 of 24 Set of Samples selected Fungal organism x data set x

data set

☐ Keep checked values at top

24 (100%) of 24 Set of Samples have data for this variable

<input checked="" type="checkbox"/>	data set	Remaining Set of Samples	Set of Samples	Distribution	%
<input checked="" type="checkbox"/>	SNP calls on WGS of <i>Pyricularia oryzae</i> isolated from Bangladesh in 2016 and 2017	23 (100%)	24 (100%)	23 (96%)	(100%)

- Change the SNP class to Non-synonymous and the ‘number of SNPs of above class’ field to 10 and click on the Get Answer button.

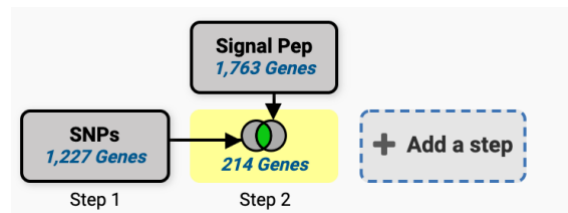
SNP Class

Non-Synonymous ▼

Number of SNPs of above class >=

10

- How many genes did you return? Which gene has the highest number of non-synonymous SNPs? (hint: sort the non-synonymous SNP columns).
- What happens if you revise this search and change the “Percent isolates with a base call >=” field to 100?
- How many of these genes have a predicted secretory signal peptide? (hint: add a step that identifies all genes with a signal peptide).



- What kinds of genes are in this result list? One way to determine if you have anything 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.
- Is there a difference between your results? Why do you think this is? Hint: The word enrichment tool uses product description to find words enriched in your genes results while GO Enrichment analysis uses Gene ontology terms assigned to your genes.

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

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

Country
City, village, or region
Host organism
data set
Sample collection
Sample
Organism under investigation
DNA sequencing

Keep checked values at top

data set

☐ 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.

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.

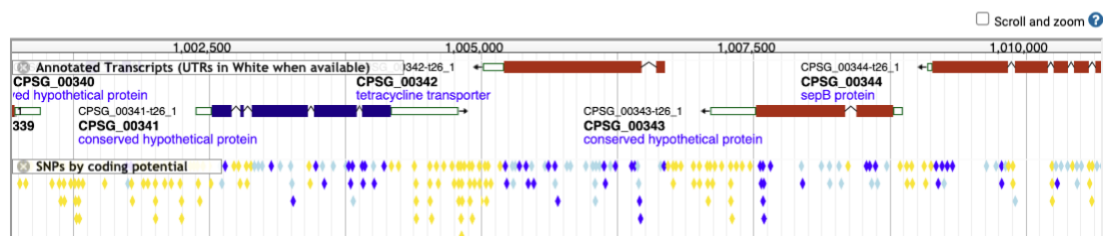
[Add to basket](#)
[Add to favorites](#)
[Download SNP](#)

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): TCTCTCTCCCTTTCCTCACAGCCACACTCTACTCCCAAATATCTCCGCACACTCAGCAGC

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

▼ SNP Genomic Context



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](#)

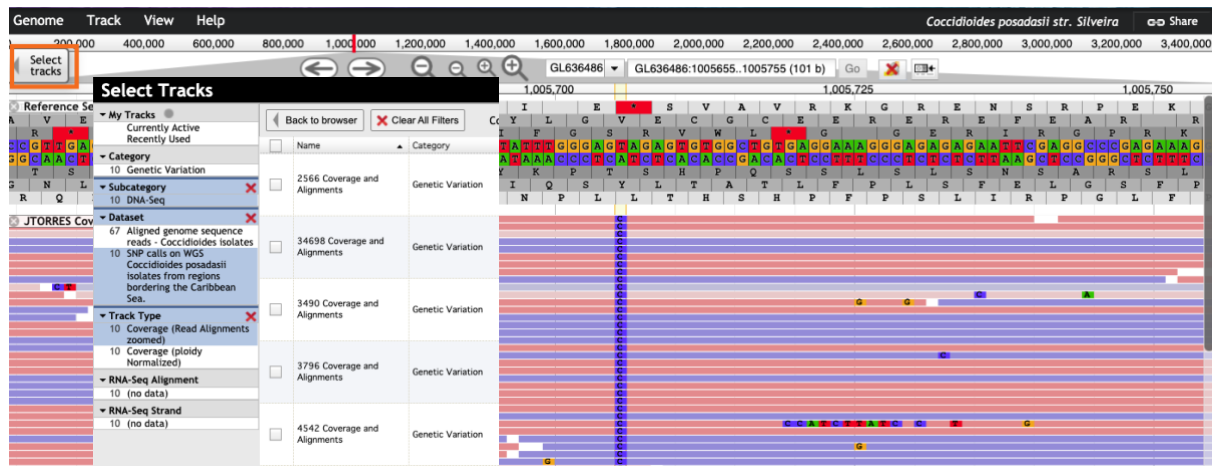
Search this table...

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.



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 ?

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

taxid

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

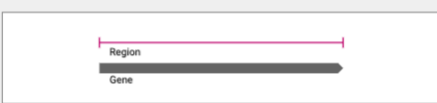
- ☐ Fungi
 - ☐ Ascomycota
 - ☐ Eurotiomycetes
 - ☐ Onygenales
 - ☐ Coccidioides
 - ☐ Coccidioides posadasii
 - ☒ Coccidioides posadasii str. Silveira

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 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"



Region
Gene

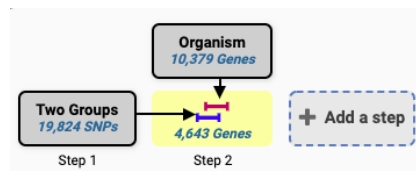
☒ Exact
☐ Upstream: 1000 bp
☐ Downstream: 1000 bp
☐ Custom:
begin at: start + 0 bp
end at: stop + 0 bp



Region
SNP

☒ 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 within* a group of isolates to identify *heterozygous* or *homozygous* SNPs...*

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

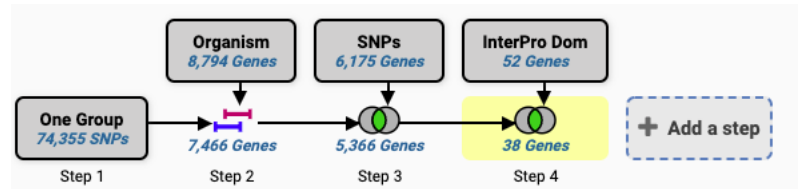
- Deploy the SNP search called “Differences Within a Group of Isolates”
- Look for homozygous SNPs in *Batrachochytrium dendrobatidis* WGS (Hammersmith). For example, here is one way to set your search:

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.

4. 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 .

Genes

Genetic variation

- Copy Number (CNV)
- Copy Number Comparison (CNV)

Genomic Sequences

- Copy Number/Ploidy ➔

Identify Genomic Sequences based on Copy Number/Ploidy

68 Strain/Sample Total 42 of 68 Strain/Sample selected data set

expand all | collapse all

Find a variable

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

☒ Keep checked values at top

	68 (100%)	42 (62%)	68 (100%)	Distribution	%
<input checked="" type="checkbox"/> Aligned genome sequence reads - <i>Candida albicans</i> clinical isolates	42	42	68	<div style="width: 62%;"></div>	(62%)
<input type="checkbox"/> Aligned genome sequence reads - <i>Candida albicans</i> evolution of resistance to amphotericin B	6	6	6	<div style="width: 10%;"></div>	(10%)
<input type="checkbox"/> SNP calls on 20 <i>Candida albicans</i> clinical isolates	20	20	20	<div style="width: 29%;"></div>	(29%)

- Set search criteria:

Copy Number \geq

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).

Ploidy Add a step

Step 1

4 Genomic Sequences Revise this search

Genomic Sequence Results

Rows per page: 1000

Download
 Add to Basket
 Add Columns

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

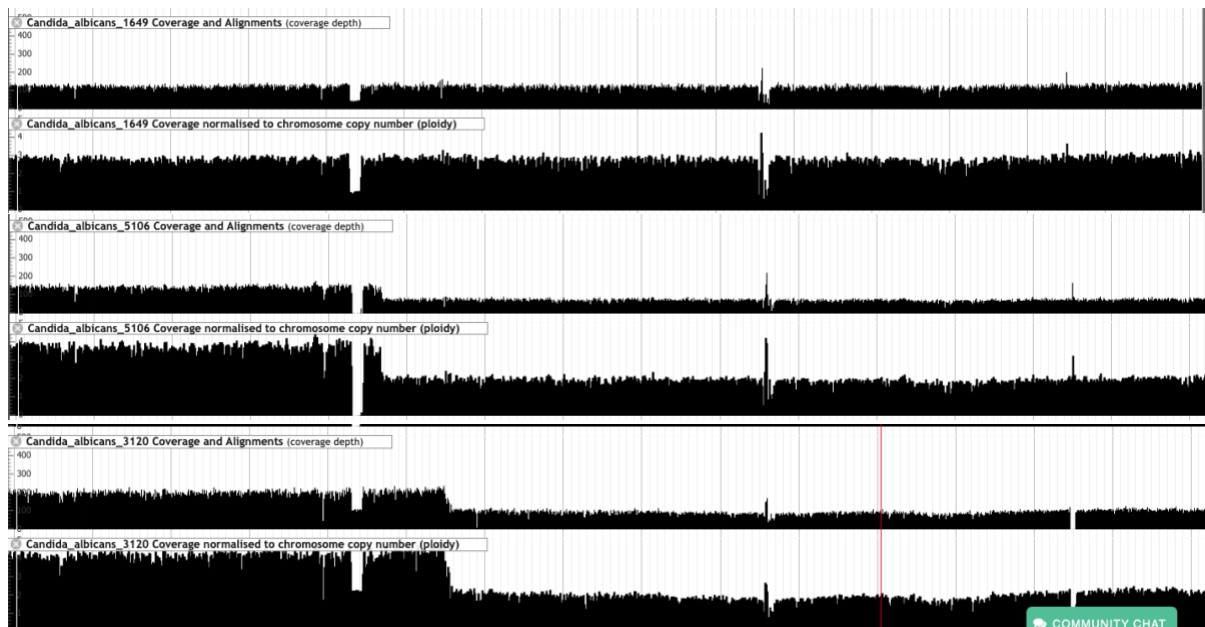
In JBrowse we have two coverage tracks:

- 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)

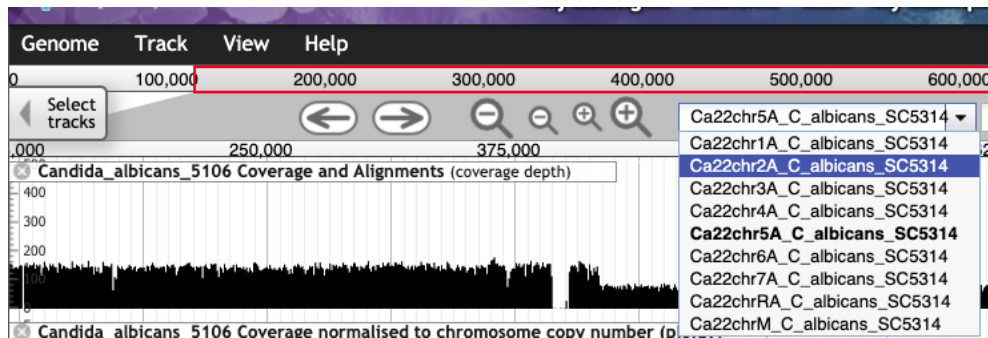
A	B	C	D
Sequence ID	Median Copy No (All Selected Samples)	Strains/Samples Meeting Criteria	Median Copy No (Samples Meeting Criteria)
Ca22chr3A_C_albicans_SC5314	2	Candida_albicans_TWTC6	3
Ca22chr4A_C_albicans_SC5314	2	Candida_albicans_1649, Candida_albicans_2501, Candida_albicans_3731, Candida_albicans_5106	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



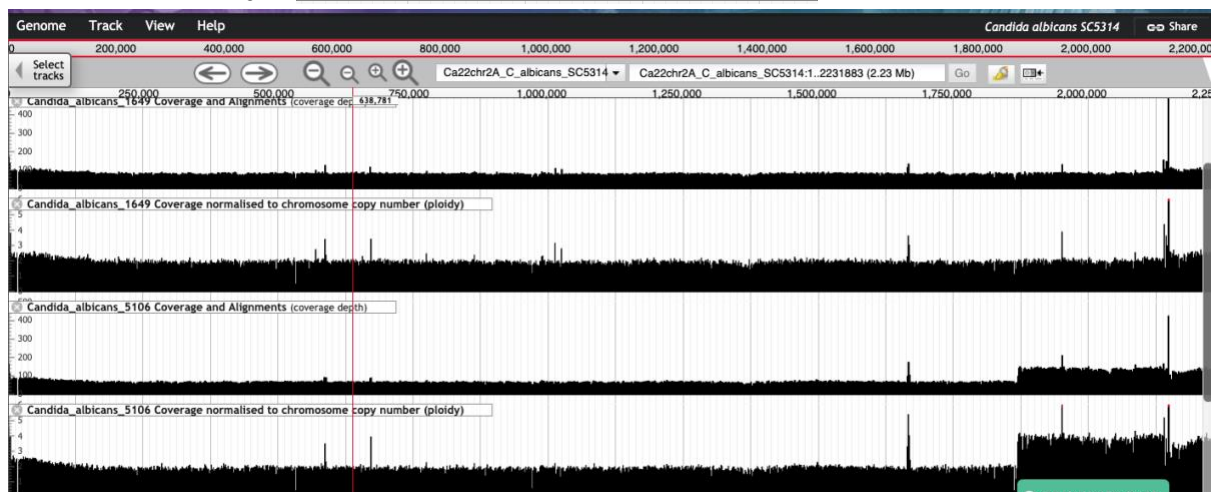
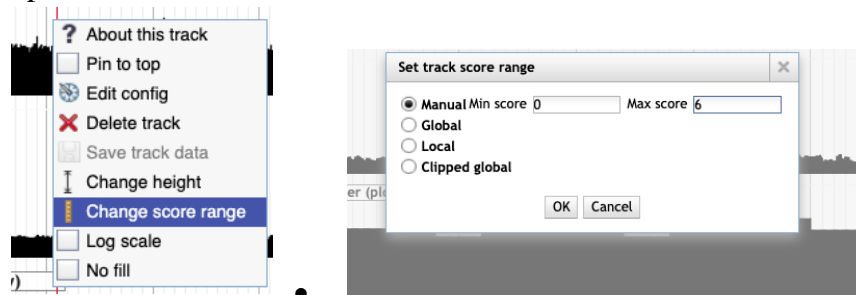
Notice examples of chromosomal (1649) and segmental triploidy (5106,3120). Note that the whole chromosome is shown in both screenshots, and both tracks are shown for each sample. We are not currently normalizing for telomere proximity.

- Switch the JBrowse view to the chromosome 2



- 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

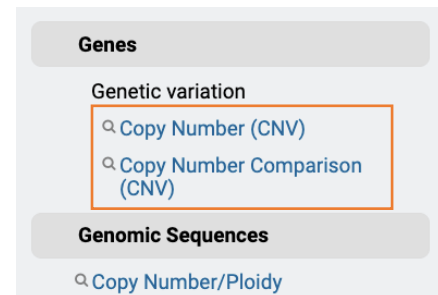


5. Using Gene Searches

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 "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.

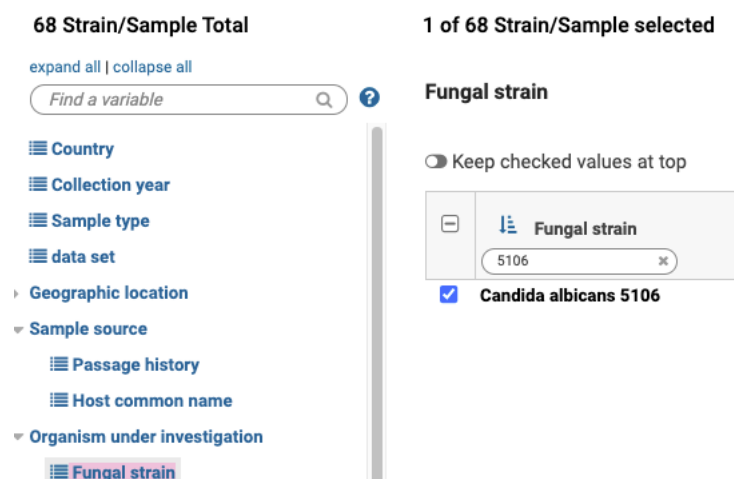


You have the choice between two different metrics for defining copy number: haploid number or gene dose:

- **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.



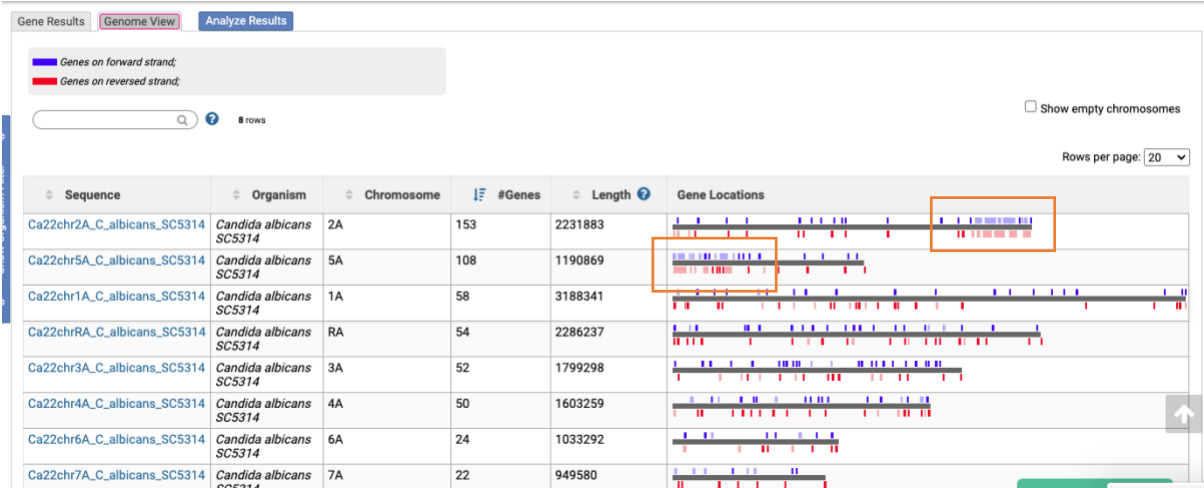
CopyNumberComparison
521 Genes

+ Add a step

Step 1

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:



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.