

Genetic Variation Exercises

SNPs and CNVs

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.

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.

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

The screenshot shows the ToxoDB website. On the left, under 'Search for...', the 'Genetic variation' category is selected, and 'SNP Characteristics' is highlighted with a red arrow. The 'Overview of Resources and Tools' section is visible on the right, showing various tools like 'Take a Tour', 'Getting Started', 'Search Strategies', etc.

- 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.

The screenshot shows the ToxoDB search results page. The 'Organism' dropdown is set to 'Toxoplasma gondii ME49' (highlighted with a red arrow). The 'Samples' section shows 65 samples selected. The 'SNP Class' dropdown is set to 'Non-Synonymous' (highlighted with a red arrow), and the 'Number of SNPs of above class >=' field is set to 20 (highlighted with a red arrow).

data set	Remaining Samples	Samples	Distribution	%
Aligned genomic sequence reads - RH Strain	65 (100%)	65 (100%)		(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%)

- 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

SNPs 3,184 Genes
Step 1

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

Organism Filter
select all | clear all | expand all | collapse all
Hide zero counts
Search organisms...

Gene Results | Genome View | **Analyze Results**

Rows per page: 50

Download Add to Basket Add Columns

Gene ID	Transcript ID	Product Description	Chromosome	Total SNPs	Nonsynonymous SNPs	
TCME49_280660	TCME49_280660-t26_1	HECT-domain (ubiquitin-transferase) domain-containing protein	VIIa	2878	1417	846
TCME49_248510	TCME49_248510-t26_1	hypothetical protein	XII	1984	1054	628
TCME49_313630	TCME49_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 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, we will discuss these results.

Unnamed Search Strategy

SNPs 2,814 Genes
Step 1

Signal Pep 38,485 Genes
Add Genes
Step 2

663 Genes (591 ortholog groups)

Gene Results | Genome View | **New Analysis**

Analyze your Gene results with a tool below.

Gene Ontology Enrichment

Metabolic Pathway Enrichment

Word Enrichment

kinase
phosphatase
exported
membrane

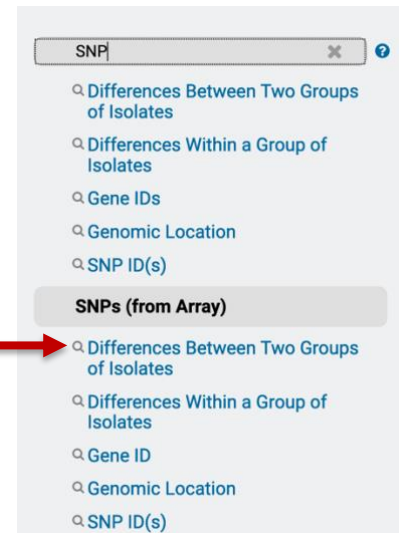
2. **Identify SNPs that distinguish parasites with rapid clearance times following treatment with the anti-malarial drug Artesunate vs. those that have delayed clearance times.**

We have a published study in PlasmoDB (Takala-Harrison et. al.) with sufficient meta-data about the samples to ask this interesting question.

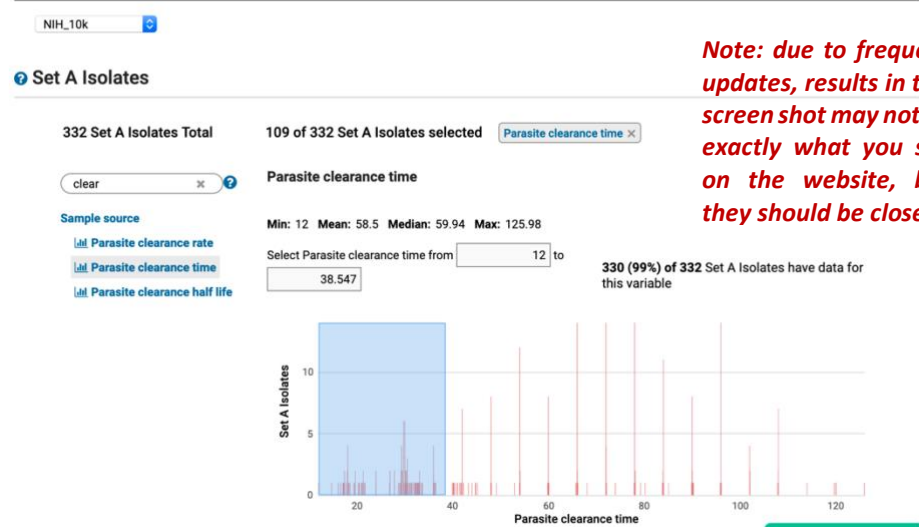
For this exercise use <http://PlasmoDB.org>

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

- 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.
- 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 possibilities once you are finished. In Set A Isolates, click on some of the characteristics to explore the data. Then choose



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






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.

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

<input type="checkbox"/>	Country	Remaining Set A Isolates	Set A Isolates	Distribution	%
		109 (100%)	331 (100%)		
<input type="checkbox"/>	Bangladesh	85 (78%)	101 (31%)		(84%)
<input type="checkbox"/>	Cambodia	15 (14%)	200 (60%)		(8%)
<input type="checkbox"/>	Thailand	9 (8%)	30 (9%)		(30%)

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

Sample source

Parasite clearance rate

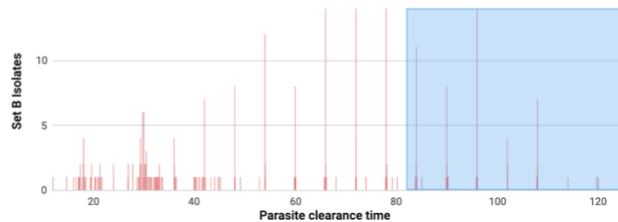
Parasite clearance time

Parasite clearance half life

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

Select Parasite clearance time from 125.98 to 82

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



- 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 Artemisinin 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 artemisinin, 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.

3. 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”.

- a. 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 11 of 65 Set A Isolates selected Host organism X

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

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	%
<input type="checkbox"/> Canis lupus familiaris	1 (2%)	1 (2%)		(100%)
<input type="checkbox"/> Capra hircus	1 (2%)	1 (2%)		(100%)
<input type="checkbox"/> Felis catus	12 (20%)	12 (20%)		(100%)
<input checked="" type="checkbox"/> Gallus gallus	11 (19%)	11 (19%)		(100%)
<input type="checkbox"/> Homo sapiens	22 (37%)	22 (37%)		(100%)
<input type="checkbox"/> Ovis aries	4 (7%)	4 (7%)		(100%)
<input type="checkbox"/> Panthera onca	1 (2%)	1 (2%)		(100%)
<input type="checkbox"/> Panthera tigris altaica	1 (2%)	1 (2%)		(100%)
<input type="checkbox"/> Puma concolor cougar	1 (2%)	1 (2%)		(100%)
<input type="checkbox"/> Ramphastidae	1 (2%)	1 (2%)		(100%)
<input type="checkbox"/> Sus scrofa	2 (3%)	2 (3%)		(100%)

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

- c. 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?

- d. 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

Two Groups
1,358 SNPs
Step 1

Step 2

Use Genomic Colocation to combine with other features

Two Groups
1,358 SNPs
Step 1

Step 2

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

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

- e. Configure the gene model characteristics search to find protein coding genes only.

Add a step to your search strategy ?

Genes or Transcripts

Transcripts

Gene Model Characteristics

255,437 Genes/Transcripts
Total

expand all | collapse all

Find a variable

Organism

Gene Type

Transcript Type

Pseudogene

Transcript Count

Gene Exon Count

Transcript Exon Count

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... 255,437 (100%)	Genes/Tran... 255,437 (100%)	Distribution	%
<input checked="" type="checkbox"/>	protein coding	249,971 (98%)	249,971 (98%)		(100%)
<input type="checkbox"/>	rRNA encoding	578 (< 1%)	578 (< 1%)		(100%)
<input type="checkbox"/>	snoRNA encoding	2 (< 1%)	2 (< 1%)		(100%)
<input type="checkbox"/>	snRNA encoding	18 (< 1%)	18 (< 1%)		(100%)
<input type="checkbox"/>	tRNA encoding	4,868 (2%)	4,868 (2%)		(100%)

- f. 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 *

Gene from the new step

Region
Gene

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

SNP from the current step

Region
SNP

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

Run Step

- 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>

4. Identifying SNPs within a group of isolates

For this exercise use <http://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

Samples

252 Samples Total

208 of 252 Samples selected

Host organism

Keep checked values at top

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

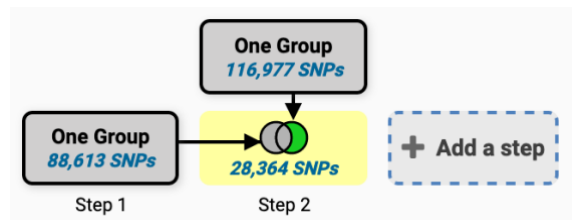
b. What does this search do? Choose *Leishmania donovani* for the organism and select isolates from the human host. Use default parameters for the rest of the parameters.

Run the query and look at your results.

- 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).*

Read frequency threshold

40%

Minor allele frequency >=

40

Percent isolates with a base call >=

20

Revise

- 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?

CNV Searches

5. Using resequencing data to identify regions of copy number variation (CNV)

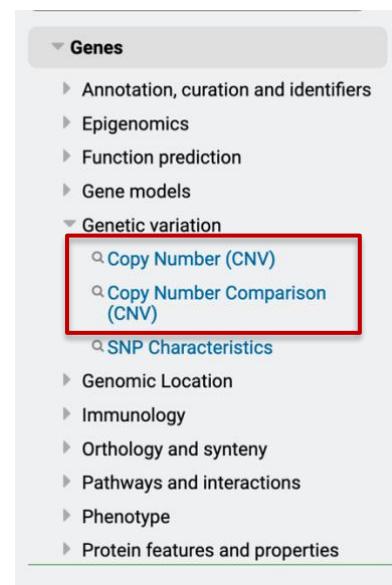
In addition to being useful for variant calling, high throughput sequencing data can be used for determining regions of copy number variation (CNV). All reads in ToxoDB are mapped to the same reference strain ME49, as a result we can estimate a gene’s copy number in each of the aligned strains.

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 in ToxoDB 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

Find a variable

Collection year

Country

data set

Sample source

Sample

Organism under investigation

DNA sequencing

Keep checked values at top

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

data set	Remaining Strain/Sam...	Strain/Sam...	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 275 Genes Add a step Step 1

275 Genes (29 ortholog groups) Revise this search

Gene Results Genome View Analyze Results

Organism Filter select all | clear all | expand all | collapse all Hide zero counts Search organisms...

Organism Filter

- ☐ Elmeridae 0
 - ☐ Cyclospora 0
 - ☐ Cyclospora cayentensis isolate NF1_C8 0
 - ☐ Cyclospora cayentensis strain CHN_HEN01 0
 - ☐ Elmeria 0
 - ☐ Elmeria acerulina Houghton 0
 - ☐ Elmeria brunetti Houghton 0
 - ☐ Elmeria falciformis Bayer Haberkorn 0
 - ☐ Elmeria maxima Weybridge 0
 - ☐ Elmeria mitis Houghton 0
 - ☐ Elmeria necatrix Houghton 0
 - ☐ Elmeria praecox Houghton 0
 - ☐ Elmeria tenella 0
 - ☐ Elmeria tenella Houghton 2021 0
 - ☐ Elmeria tenella strain Houghton 0
 - ☐ Sarcocystidae 275
 - ☐ Besnoitia besnoiti strain Bb-Ger1 0
 - ☐ Cyclospora suis strain Wien I 0
 - ☐ Hammondia hammondi strain H.H.34 0

Genes on forward strand: Genes on reversed strand: Show empty chromosomes Rows per page: 20

Sequence	Organism	Chromosome	#Genes	Length	Gene Locations
TGME49_chrVI	Toxoplasma gondii ME49	VI	40	3656745	
TGME49_chrXII	Toxoplasma gondii ME49	XII	39	7094428	
TGME49_chrX	Toxoplasma gondii ME49	X	32	7486190	
TGME49_chrXI	Toxoplasma gondii ME49	XI	27	6623461	
TGME49_chrIX	Toxoplasma gondii ME49	IX	23	6327655	
TGME49_chrIV	Toxoplasma gondii ME49	IV	20	2686605	
TGME49_chrV	Toxoplasma gondii ME49	V	20	3331915	

COMMUNITY CHAT

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 resequenced strain coverage data turned on:

<https://tinyurl.com/2k7kps54>

