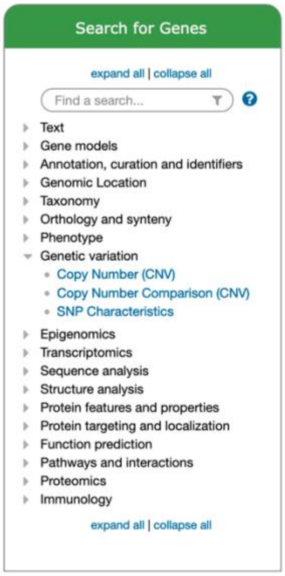
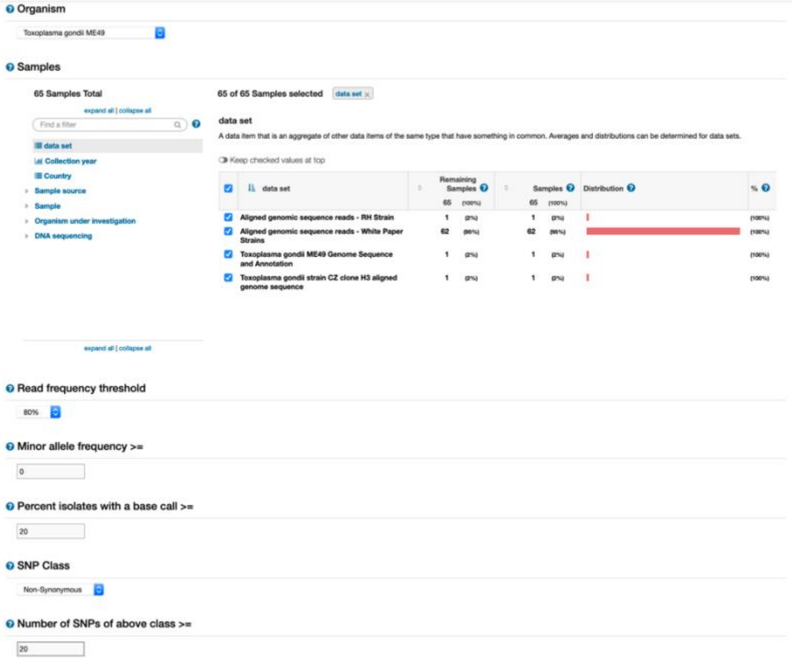
Genetic Exercises

SNPs and Population Genetics

Single Nucleotide Polymorphisms (SNPs) in EuPathDB 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 EuPathDB by two basic methods; re-sequencing and then alignment of sequence reads to a reference genome or DNA hybridization to a SNP-chip array. 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.
   - ***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
Keep checked values at top
expand all | collapse all

Organism: Toxoplasma gondii ME49

Samples

65 Samples Total

65 of 65 Samples selected

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%

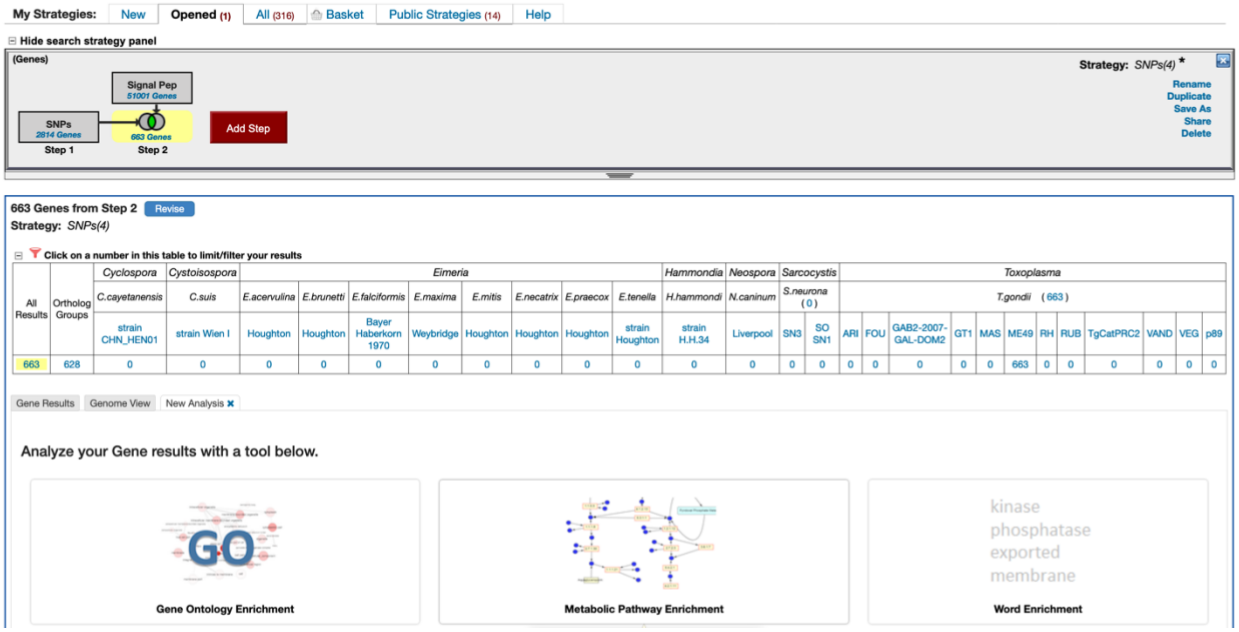
Minor allele frequency >= 0

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).
- **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.



663 Genes from Step 2 Revise

Strategy: *SNPs(4)*

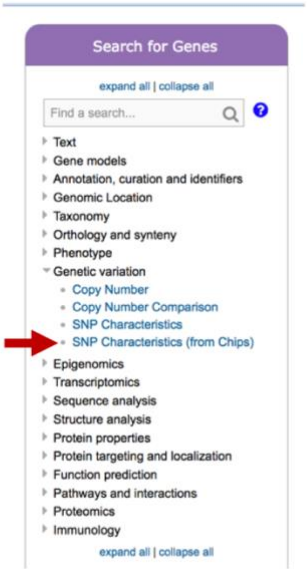
⊟ Click on a number in this table to limit/filter your results

All Results	Ortholog Groups	*Cyclospora*	*Cystoisospora*	*Eimeria*								*Hammondia*	*Neospora*	*Sarcocystis*		*Toxoplasma*											
		C.cayetanensis	*C.suis*	*E.acervulina*	*E.brunetti*	*E.falciformis*	*E.maxima*	*E.mitis*	*E.necatrix*	*E.praecox*	*E.tenella*	*H.hammondi*	*N.caninum*	*S.neurona* (0)		*T.gondii* (663)											
		strain CHN_HEN01	strain Wien I	Houghton	Houghton	Bayer Haberkorn 1970	Weybridge	Houghton	Houghton	Houghton	strain Houghton	strain H.H.34	Liverpool	SN3	SO SN1	ARI	FOU	GAB2-2007-GAL-DOM2	GT1	MAS	ME49	RH	RUB	TgCatPRC2	VAND	VEG	p89
663	628	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	663	0	0	0	0	0	0

Analyze your Gene results with a tool below.

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

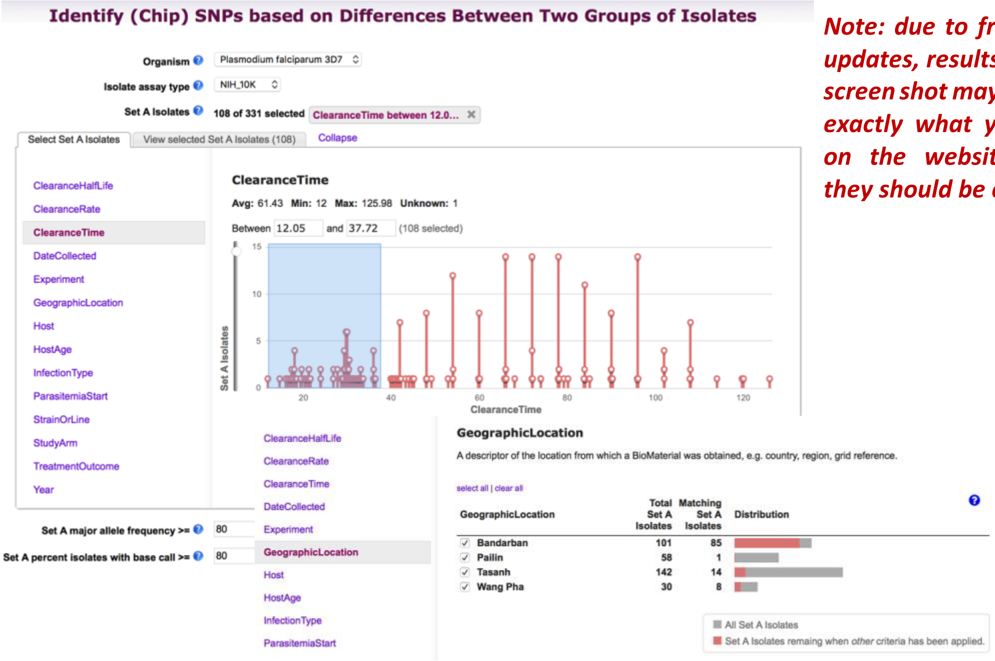


Navigate to the “Differences between two groups of isolates” search under “Search for SNPs (from Array).

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 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).*



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

Organism: Plasmodium falciparum 3D7

Isolate assay type: NIH_10K

Set A Isolates: 108 of 331 selected ClearanceTime between 12.0...

ClearanceTime

Avg: 61.43 Min: 12 Max: 125.98 Unknown: 1

Between 12.05 and 37.72 (108 selected)

Set A major allele frequency >= 80

Set A percent isolates with base call >= 80

GeographicLocation

A descriptor of the location from which a BioMaterial was obtained, e.g. country, region, grid reference.

GeographicLocation	Total Set A Isolates	Matching Set A Isolates	Distribution
✓ Bandarban	101	85	
✓ Pailin	58	1	
✓ Tasanh	142	14	
✓ Wang Pha	30	8	

▨ All Set A Isolates
🟥 Set A Isolates remaing when *other* criteria has been applied.

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.

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?

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.

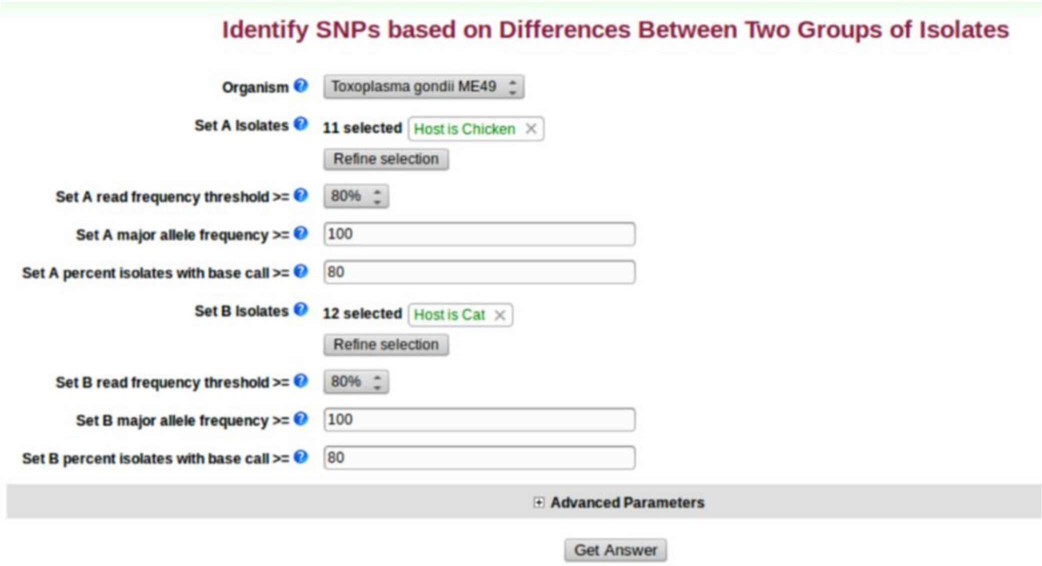
Hint: add a step to search for genes by text and search for kelch13. This will cause you to use the genomic co-location operation as outlined in exercise 3. Set it up the same way except choose custom and start – 10000, stop + 10000 to define the region.

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.

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.



Identify SNPs based on Differences Between Two Groups of Isolates

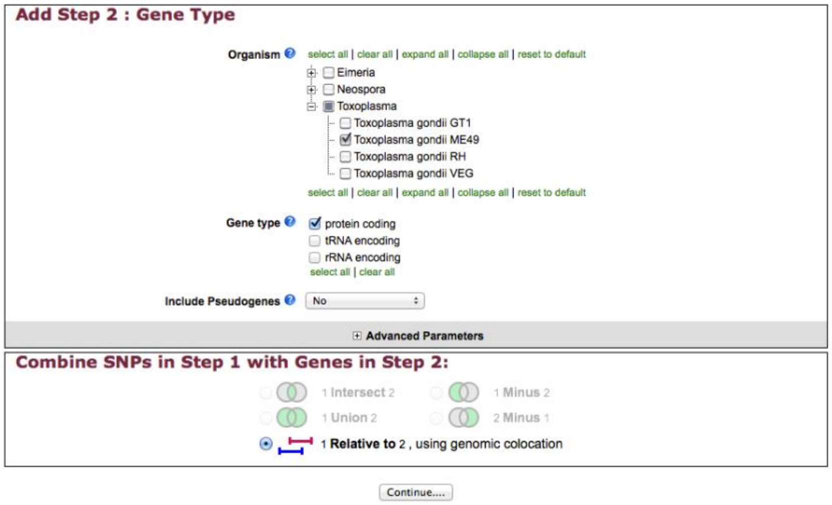
Organism	Toxoplasma gondii ME49
Set A Isolates	11 selected Host is Chicken ×
Set A read frequency threshold >=	80%
Set A major allele frequency >=	100
Set A percent isolates with base call >=	80
Set B Isolates	12 selected Host is Cat ×
Set B read frequency threshold >=	80%
Set B major allele frequency >=	100
Set B percent isolates with base call >=	80

c. Let’s run a very stringent search and change the “major allele frequency” parameters for both sets to 90. (*What does that mean?)*. We’ll leave the other parameters at their default values, which are in themselves pretty stringent … but feel free to change them to see how this impacts your results.

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

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*. What is the only operator that is available to you when you add this step? Why is this? 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 Step 2 : Gene Type

Organism: select all | clear all | expand all | collapse all | reset to default

- ☐ Eimeria
- ☐ Neospora
- ▨ Toxoplasma
  - ☐ Toxoplasma gondii GT1
  - ☑ Toxoplasma gondii ME49
  - ☐ Toxoplasma gondii RH
  - ☐ Toxoplasma gondii VEG

select all | clear all | expand all | collapse all | reset to default

Gene type:
- ☑ protein coding
- ☐ tRNA encoding
- ☐ rRNA encoding

select all | clear all

Include Pseudogenes: No

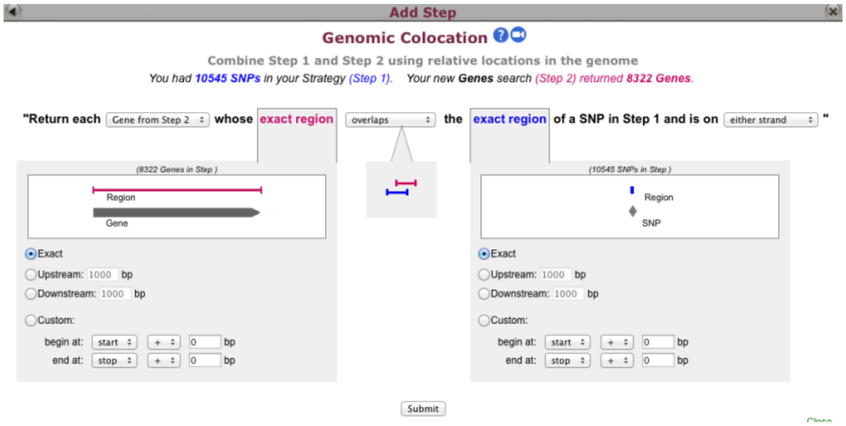
⊞ Advanced Parameters

Combine SNPs in Step 1 with Genes in Step 2:

○ 1 Intersect 2 ○ 1 Minus 2

○ 1 Union 2 ○ 2 Minus 1

● 1 **Relative to** 2 , using genomic colocation



Add Step

Genomic Colocation

Combine Step 1 and Step 2 using relative locations in the genome

*You had **10545 SNPs** in your Strategy (Step 1). Your new **Genes** search (Step 2) returned **8322 Genes**.*

"Return each [Gene from Step 2] whose **exact region** [overlaps] the **exact region** of a SNP in Step 1 and is on [either strand] "

(8322 Genes in Step)

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

(10545 SNPs in Step)

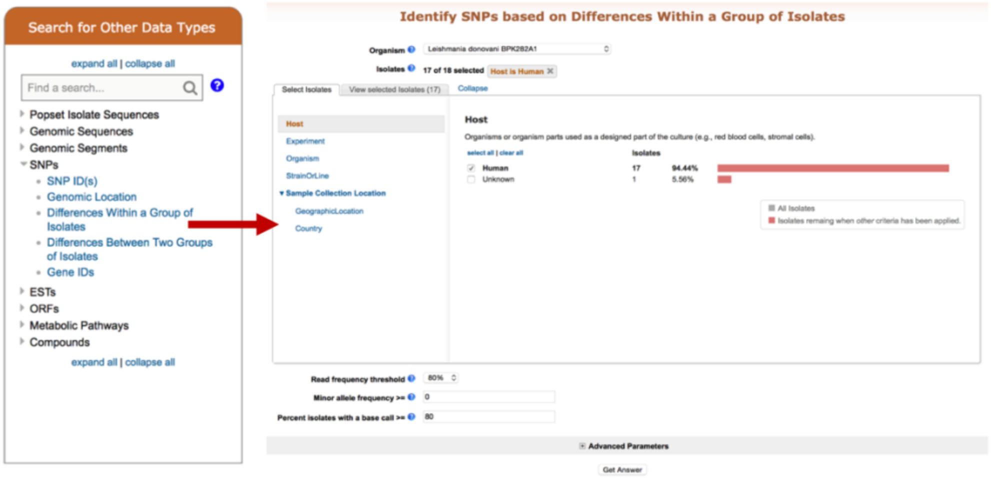
● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start + 0 bp
end at: stop + 0 bp

- 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 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.



Identify SNPs based on Differences Within a Group of Isolates

Organism: Leishmania donovani BPK282A1

Isolates: 17 of 18 selected Host is Human

Host

Organisms or organism parts used as a designed part of the culture (e.g., red blood cells, stromal cells).

	Isolates	
☑ Human	17	94.44%
□ Unknown	1	5.56%

Read frequency threshold: 80%

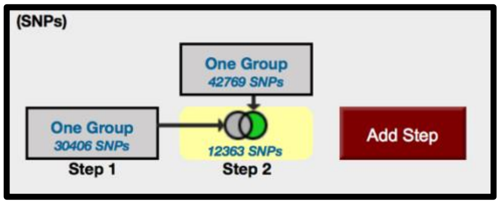
Minor allele frequency >=: 0

Percent isolates with a base call >=: 80

   Hint: you can find this under "SNPs" in the "Identify Other Data Types" section.

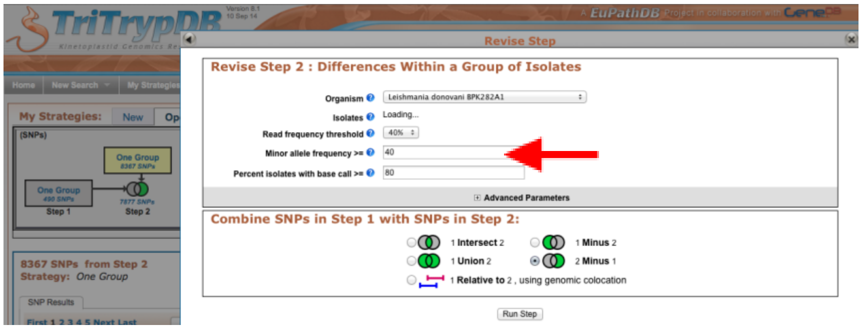
   b. What does this search do? Choose *Leishmania donovani* for the organism and select isolates from the human host. Use default parameters for the rest of the parameters.
   Run the query and look at your results.

   - 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).*



Revise Step 2 : Differences Within a Group of Isolates

Organism: Leishmania donovani BPK282A1
Isolates: Loading...
Read frequency threshold: 40%
Minor allele frequency >=: 40
Percent isolates with base call >=: 80

⊞ Advanced Parameters

Combine SNPs in Step 1 with SNPs in Step 2:

○ 1 Intersect 2 ○ 1 Minus 2
○ 1 Union 2 ● 2 Minus 1
○ 1 Relative to 2 , using genomic colocation

Run Step

- 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?

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

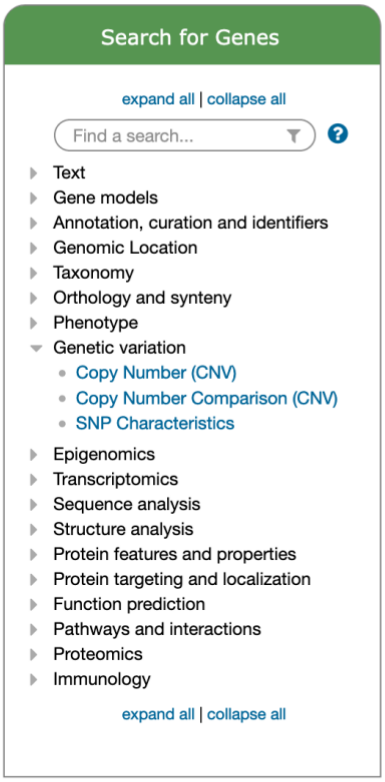
In addition to being useful for variant calling, high throughput sequencing data can be used for determining regions of copy number variation (CNV). All reads in ToxoDB are mapped to the same reference strain ME49, as a result we can estimate a gene's copy number in each of the aligned strains.

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.

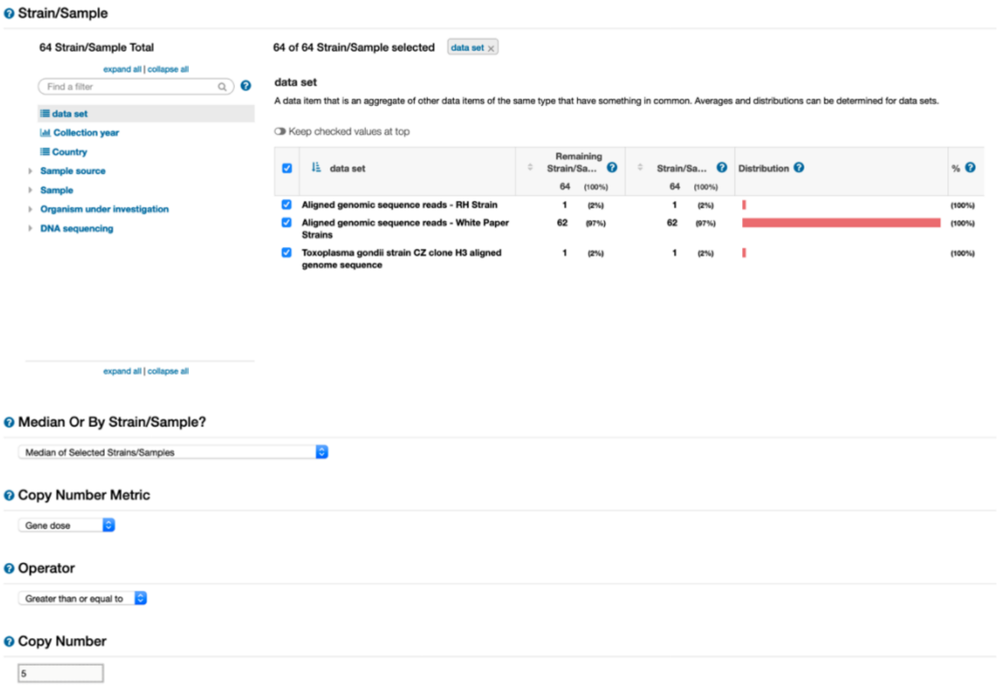
Copy number comparison: This search compares the estimated copy number of a gene in the re-sequenced strain with the copy number in the reference annotation. The copy number in the reference annotation is calculated as the number of genes that are in the same ortholog group as the gene of interest. We 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 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.



Strain/Sample

64 Strain/Sample Total

64 of 64 Strain/Sample selected

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 Strain/Sa...		Strain/Sa...		Distribution	%
		64	(100%)	64	(100%)		
☑	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

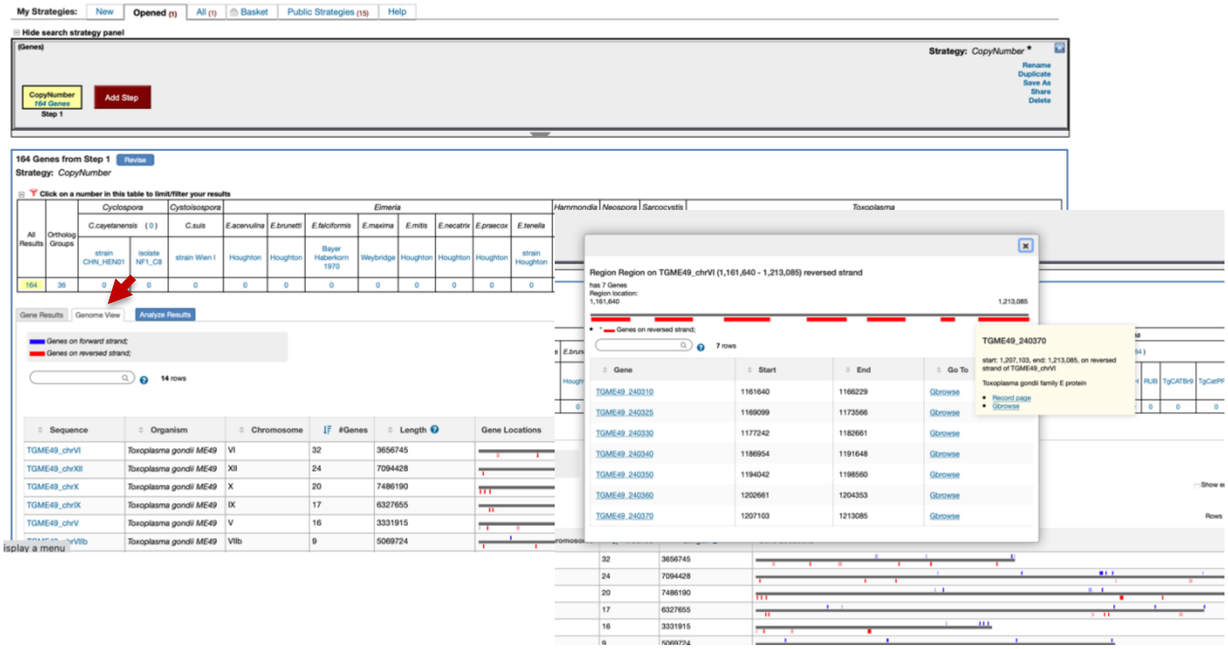
Greater than or equal to

Copy Number

5

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

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). 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/y3mc53zm**



My Strategies: New | Opened (1) | All (1) | Basket | Public Strategies (15) | Help

⊟ Hide search strategy panel

(Genes)

CopyNumber 164 Genes — Step 1 | Add Step

Strategy: CopyNumber * — Rename, Duplicate, Save As, Share, Delete

164 Genes from Step 1 Revise

Strategy: CopyNumber

⊟ Click on a number in this table to limit/filter your results

All Results	Ortholog Groups	Cyclospora		Cystoisospora	Eimeria								Hammondia	Neospora	Sarcocystis	Toxoplasma
		C.cayetanensis (0)		C.suis	E.acervulina	E.brunetti	E.falciformis	E.maxima	E.mitis	E.necatrix	E.praecox	E.tenella				
		strain CHN_HEN01	isolate NF1_C8	strain Wien I	Houghton	Houghton	Bayer Haberkorn 1970	Weybridge	Houghton	Houghton	Houghton	strain Houghton				
164	36	0	0	0	0	0	0	0	0	0	0	0				

Gene Results | Genome View | Analyze Results

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

14 rows

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	
TGME49_chrVIIb	*Toxoplasma gondii ME49*	VIIb	9	5069724	

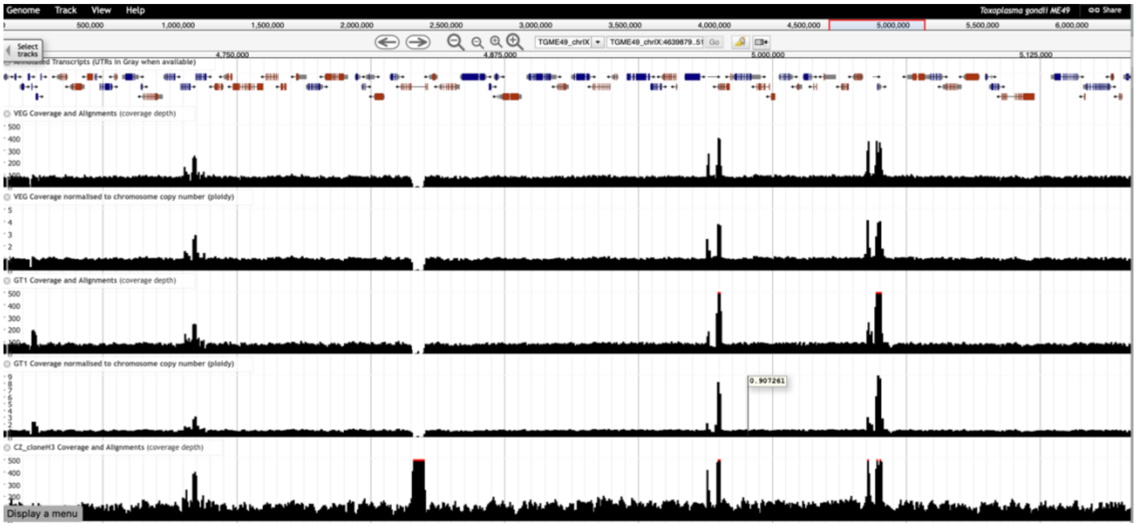
Region Region on TGME49_chrVI (1,161,640 - 1,213,085) reversed strand
has 7 Genes
Region location:
1,161,640 1,213,085

- 🟥 Genes on reversed strand;

7 rows

Gene	Start	End	Go To
TGME49_240310	1161640	1166229	Gbrowse
TGME49_240325	1169099	1173566	Gbrowse
TGME49_240330	1177242	1182661	Gbrowse
TGME49_240340	1186954	1191648	Gbrowse
TGME49_240350	1194042	1198560	Gbrowse
TGME49_240360	1202661	1204353	Gbrowse
TGME49_240370	1207103	1213085	Gbrowse

TGME49_240370
start: 1,207,103, end: 1,213,085, on reversed strand of TGME49_chrVI
Toxoplasma gondii family E protein
- Record page
- Gbrowse


Genome
Track
View
Help
Toxoplasma gondii ME49
Share
500,000
1,000,000
1,500,000
2,000,000
2,500,000
3,000,000
3,500,000
4,000,000
4,500,000
5,000,000
5,500,000
6,000,000
Select tracks
TGME49_chrIX
TGME49_chrIX:4639879..51
Go
4,750,000
4,875,000
5,000,000
5,125,000
Annotated Transcripts (UTRs in Gray when available)
VEG Coverage and Alignments (coverage depth)
VEG Coverage normalised to chromosome copy number (ploidy)
GT1 Coverage and Alignments (coverage depth)
GT1 Coverage normalised to chromosome copy number (ploidy)
0.907261
CZ_cloneH3 Coverage and Alignments (coverage depth)
500
400
300
200
100
Display a menu