

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.

Search for Genes

expand all | collapse all

Find a search...

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Orthology and synteny
- Phenotype
- Genetic variation
 - Copy Number (CNV)
 - Copy Number Comparison (CNV)
 - SNP Characteristics
- Epigenomics
- Transcriptomics
- Sequence analysis
- Structure analysis
- Protein features and properties
- Protein targeting and localization
- Function prediction
- Pathways and interactions
- Proteomics
- Immunology

expand all | collapse all

Organism

Toxoplasma gondii ME49

Samples

65 Samples Total 65 of 65 Samples selected

Find a filter

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.

Keep checked values at top

data set	Remaining Samples	Samples	Distribution	%
Aligned genomic sequence reads - RH Strain	1 (2%)	1 (2%)		(100%)
Aligned genomic sequence reads - White Paper Strains	62 (96%)	62 (96%)		(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

- 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, we will discuss these results.

My Strategies: [New](#) [Opened \(1\)](#) [All \(316\)](#) [Basket](#) [Public Strategies \(14\)](#) [Help](#)

Hide search strategy panel

(Genes) Strategy: SNPs(4) * [Rename](#) [Duplicate](#) [Save As](#) [Share](#) [Delete](#)

SNPs 2814 Genes Step 1 → Signal Pep 51071 Genes Step 2 → Add Step

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.cayatanensis		C.suis		E.acervulina	E.brunetti	E.falciformis	E.maxima	E.mits	E.necatrix	E.praecox	E.tenella	H.hammondi	N.caninum	S.neurona (0)	T.gondii (663)															
		strain CHN_HEND1	strain Wien I	Houghton	Houghton	Bayer Haberkorn 1970	Weybridge	Houghton	Houghton	Houghton	Houghton	strain Houghton	strain H.H.34	Liverpool	SN3	SO SN1	ARI	FOU	GAB2-2007-GAL-DM2	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	0	0	0	0	0	0	0	0	0	0	0	0

Gene Results | [Genome View](#) | [New Analysis](#)

Analyze your Gene results with a tool below.



Gene Ontology Enrichment



Metabolic Pathway Enrichment

kinase
phosphatase
exported
membrane

Word Enrichment

- 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”.

- 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

Set A Isolates

Set A read frequency threshold >=

Set A major allele frequency >=

Set A percent isolates with base call >=

Set B Isolates

Set B read frequency threshold >=

Set B major allele frequency >=

Set B percent isolates with base call >=

- 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 distinguishes 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

Add Step 2 : Gene Type

Organism

Eimeria

Neospora

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 whose overlaps the of a SNP in Step 1 and is on

(8322 Genes in Step)

Gene

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: bp

end at: bp

(10545 SNPs in Step)

SNP

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: bp

end at: bp

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”

- 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: <http://toxodb.org/toxo/im.do?s=f6cdf8edcda494b>

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

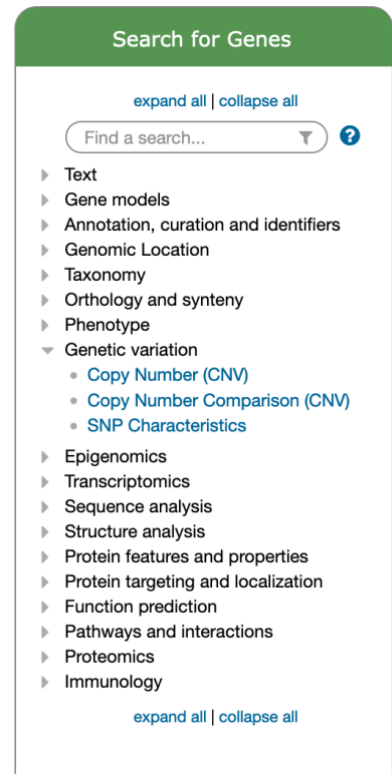
In addition to being useful for variant calling, high throughput sequencing data can be used for determining regions of copy number variation (CNV). All reads in ToxoDB are mapped to the same reference strain ME49, as a result we can estimate a gene's copy number in each of the aligned strains.

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 resequenced 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 both on the same chromosome and 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 resequenced 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 greater 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 expand all | collapse all

64 of 64 Strain/Sample selected data set x

Find a filter

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.

Keep checked values at top

<input checked="" type="checkbox"/>	data set	Remaining Strain/Sa...	Strain/Sa...	Distribution	%
		64 (100%)	64 (100%)		
<input checked="" type="checkbox"/>	Aligned genomic sequence reads - RH Strain	1 (2%)	1 (2%)		(100%)
<input checked="" type="checkbox"/>	Aligned genomic sequence reads - White Paper Strains	62 (97%)	62 (97%)		(100%)
<input checked="" type="checkbox"/>	Toxoplasma gondii strain CZ clone H3 aligned genome sequence	1 (2%)	1 (2%)		(100%)

expand all | collapse all

Median Or By Strain/Sample?
Median of Selected Strains/Samples

Copy Number Metric
Gene dose

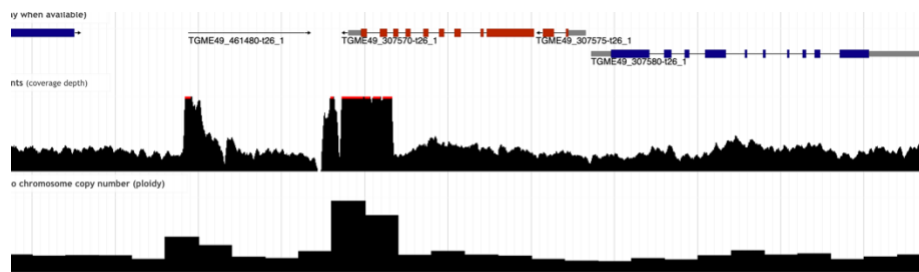
Operator
Greater than or equal to

Copy Number

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).

The screenshot shows a genomic analysis tool interface. At the top, there's a search panel with "CopyNumber" selected. Below it, a table lists 92 genes from Step 1, categorized by organism groups like Cyclospora, Eimeria, and Toxoplasma. A red arrow points to a specific entry in the table. Below the table, there's a "Genome View" tab showing a genomic map with red and blue lines representing genes. A detailed view of a region on TGME49_chrVII is shown, listing genes like TGME49_240310, TGME49_240325, TGME49_240330, TGME49_240340, TGME49_240350, TGME49_240360, and TGME49_240370. A tooltip for TGME49_240370 shows its start and end coordinates and a link to view the record.

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 gene has the highest CNV? (*hint*: sort the median gene dose column from highest to lowest) – do you believe this? Take a look at the coverage data of CZ strain aligned to ME49. What do you see? <http://tinyurl.com/y44yyxkb>



You can zoom out to see if there is any bias in read alignment.

