

Population data (SNPs)

Exercise 4

4.1 Identify SNPs within a group of Isolates

Note: For this exercise use <http://www.tritrypdb.org>

a. Go to the “Identify SNPs based on a Group of Isolates” search.

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

Identify Other Data Types:

- Expand All | Collapse All
- Isolates
- Genomic Sequences
- Genomic Segments (DNA Motif)
- SNPs
 - SNP ID(s)
 - A Group of Isolates **NEW**
 - Genomic Location **NEW**
 - Gene IDs **NEW**
 - Compare Two Groups of Isolates **NEW**
- ESTs
- ORFs
- SAGE Tags

Identify SNPs based on A Group of Isolates **NEW**

Organism:

Isolates: 16 selected | Host is human X

Select Isolates | View selection (16) | Collapse

Host	Count	Percentage
human	16	94.12%
unknown	1	5.88%

The red bar indicates the percentage of Isolates whose qualities you've already selected:

Read frequency threshold:

Minor allele frequency >=

Percent isolates with base call >=

Advanced Parameters

Get Answer

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 additional SNPs did you identify?
- Click on the Step 2 isolate group box to view the results of this specific search. What do you notice about the %minor alleles? (*many are quite low ... ie 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?


4.2 Find SNPs that distinguish *Toxoplasma gondii* strains isolated from chickens as compared to those isolated from cats.


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


Navigate to “Compare two groups of isolates”.


- Click select set A isolates and select hosts from the left column. Check the chicken box to select the 11 chicken isolates.
- Click select set B isolates and select hosts from the left column. Check the cat box to select the 12 cat isolates.
- Let’s run a very stringent search and change the “major allele frequency” parameters for both sets to 100. (*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..

Identify SNPs based on Compare Two Groups of Isolates NEW


Organism 


Toxoplasma gondii ME49 


Set A Isolates 

11 selected Host is Chicken 


Refine selection

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 


Refine selection

Set B read frequency threshold >= 


80% 

Set B major allele frequency >= 

100

Set B percent isolates with base call >= 

80

 Advanced Parameters

Get Answer

- How many SNPs did your search return? Are you surprised by this large number that distinguish these two fairly large groups of isolates?
- Optional (but highly encouraged 😊)*. 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?
 - 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](#)

- ☐ Elmeria
- ☐ 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 ?

[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

[Continue....](#)

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 whose **exact region** the **exact region** of a SNP in Step 1 and is on

(8322 Genes in Step)

Region

Gene

☒ Exact

☐ Upstream: 1000 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: bp

end at: bp

(10545 SNPs in Step)

Region

SNP

☒ Exact

☐ Upstream: 1000 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: bp

end at: bp

[Submit](#)

- 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 genes that is uncharacterized? 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>

4.3 Isolate comparison (SNP Chip)

Note: For this exercise use <http://www.plasmodb.org>

- Go to the “Identify SNPs based on Isolate Comparison (SNP chip)” search. *Hint: you can find this under “SNPs” in the “Identify Other Data Types” section.*

Identify SNPs based on Isolate Comparison (SNP chip)

Set A isolate identifiers

☒ Enter a list of IDs or text:

CP3.CF04.008_108, CP3.CF04.008_12G, CP3.CF04.008_1F, CP3.CF04.008_2G, CP3.CF04.008_7H, CP3.CF04.009_6D, CP3.CF04.010_108, CP3.CF08.008

☐ Upload a text file:

Choose File no file selected
Maximum size: 10MB. The file should contain the list of IDs.

☐ Copy from My Basket: 0 Isolates will be copied from your Basket.

☐ Copy from My Strategy: Choose a isolate strategy:

+

Minimum percentage of isolates in Set A with same allele >=

100

Set B isolate identifiers

☒ Enter a list of IDs or text:

CP3.1054, CP3.207-89, CP3.330-89, CP3.36-89, CP3.365_89, CP3.51, CP3.608, CP3.60888, CP3.7G8, CP3.9-411, CP3.ADA2, CP3.JST

☐ Upload a text file:

Choose File no file selected
Maximum size: 10MB. The file should contain the list of IDs.

☐ Copy from My Basket: 0 Isolates will be copied from your Basket.

☐ Copy from My Strategy: Choose a isolate strategy:

+

Minimum percentage of isolates in Set B with same allele >=

100

Advanced Parameters

[Get Answer](#)

- c. What does this search do?** Run the default query and look at your results. How many SNPs were identified between isolates from Brazil and Malawi? What could you use this information for? *Hint: If I were to provide an allele call for each of these 35 SNPs from an isolate could you tell me whether it came from Brazil or Malawi (or neither)?.*

4.3a Optional Exercise: Find SNPs that differentiate isolates from East Africa and those from West Africa. **Note,** this exercise has been moved to the end of this set and should not be done until after section 4.4. We will be replacing the current mechanism for selecting sets of isolates with the advanced parameter you have been using so while the exercise is well worth doing, it will be much simpler in a couple of months.

4.4 Identify genes that appear to be under diversifying selection based on isolates from Senegal. For this exercise use <http://www.plasmodb.org>

- Go to the “Identify Genes based on SNP Characteristics” search. *Hint: you can find this under “Identify Genes” in the “Population Biology” section.*
 - Choose strains from organism P. falciparum that are from the geographic region of Thies, Senegal.
 - Set the number of coding SNPs to be ≥ 30 and the non-synonymous / synonymous SNP ratio to be ≥ 3 . (see image below for help configuring the search if you have problems).
 - How many genes did you find? What types of genes do you see in your list? Does this make sense as genes that might advantageous to the parasite to be under diversifying selection (ie, the protein sequence is changing)?
 - What is the gene with the highest non-synonymous / synonymous ratio? *Hint: sort by this column.*
 - What gene has the most total SNPs?
 - Save this strategy as we will use it as a starting point for some comparisons and it will be quicker for you to reopen the saved strategy than to re-run the search.

Identify Genes based on SNP Characteristics NEW

Organism Plasmodium falciparum 3D7

Isolates 69 selected GeographicLocation is Thies,S...

Select Isolates View selection (69)

Year

Host

StrainOrLine

GeographicLocation

☐ Gambia

☒ Thies,Senegal

☐ unknown

64

69

11

44.44%

47.92%

7.64%

The red bar indicates the percentage of Isolates whose qualities you've already selected:

Minor allele frequency \geq

Percent isolates with base call \geq

Read frequency threshold

SNP Class

Number of SNPs of above class \geq

Number of SNPs of above class \leq

Non-synonymous / synonymous SNP ratio \geq

Non-synonymous / synonymous SNP ratio \leq

SNPs per KB (CDS) \geq

SNPs per KB (CDS) \leq

Advanced Parameters

Get Answer

- Add a step to this result to compare this list of genes with genes that may be under diversifying selection based on isolates from Gambia (an African country essentially contained within Senegal).
 - *Hint: click add step -> Genes -> population biology -> SNP Characteristics. Configure as above except choose isolates from Gambia.*
 - How many genes are in common between these two regions? **NOTE:** save this strategy as we'll use it again later in this exercise.
 - Is AMA1 still the gene with the largest NS/S ratio? *Hint: Add a column for HTS NS/S ratio.* Why is the ratio lower than for either of the specific results (Senegal or Gambia)? *Hint: This ratio is based on a read frequency threshold of 20% which is very low for haploid organisms so likely contains sequencing errors.*
 - How would you identify genes under selection in Senegal but not Gambia (and vice versa)? *Hint: revise the operator to use 1 not 2 or 2 not 1 operator.* Play with relaxing the parameters a bit of the result being subtracted to increase the likelihood that your result is specific. For example, set the number of coding SNPs to 20 and/or set the NS/S ratio to 2.5.
- **Comparing your results with a published list:** You just read the recent paper by Tetteh *et.al.* (<http://www.ncbi.nlm.nih.gov/pubmed/19440377>) where they perform an analysis of SNPs on a set of *P. falciparum* genes. Their conclusion is that these genes are under “balancing” selection – under diversifying selection due to their exposure to the host’s immune pressure. You decide you would like to analyze their list of genes in PlasmoDB.

Here is the list of gene IDs from their paper:

PFF0615c, Pf13_0338, PFE0395c, PF14_0201, PFF0995c, PF10_0346, PF10_0347, PF10_0348, PF10_0352, PF13_0197, PF13_0196, MAL13P1.174, PF13_0193, MAL13P1.173, Pf13_0191, PF13_0192, PF13_0194, PFL1385c, PFB0340c, MAL7P1.208, PF13_0348, PF10_0144, PF14_0102, PFE0080c, PFE0075c, PFD0955w

- Add a step to your strategy to see if any of these genes are present in your list of genes with high NS/S ratios. *Hint: click add step -> genes -> Test,IDs,organism -> Gene IDs and paste in the list above.*
 - How many genes are shared? *Hint: The above strategy is very stringent, try decreasing stringency of SNP searches to 10 coding SNPs and NS/S ratio >= 1.5.*
- **Comparing your results with genes targeted by the host immune system.** You want to explore whether your list of stringent diversifying genes appear to be targets of the host immune response. Re-open your saved strategy and

compare the results to genes that appear to be targeted by the host immune response.

- *Hint: click add step -> Genes -> Host Response -> Serum Antibody levels.*
- We'll try to identify genes that are more responsive in adults before the malaria season (May timepoint) as compared to children ages 1-5 at the same timepoint.
- Select reference samples: Sampling time point (May), age (select 1-5)
- Select comparison samples: Sampling time point (May), age (select adults)
- Leave other parameter values at defaults, select intersect and submit.

Add Step 4 : Serum Antibody Levels

Reference Samples 95 selected Sampling Time Point is May Age is between 1.36 and 5.52

Refine selection

Comparison Samples 51 selected Sampling Time Point is May Age is between 17.03 and 25.64

Select Comparison Samples View selection (51) Collapse

IPCR Result

Age

Sampling Time Point

Time To Onset

Disease State

Sex

The time period elapsed since an identifiable point in the life cycle of an organism. If a development... [read more](#)

Avg: 8.79 Min: 2 Max: 25

Select Age between and (88 items selected)

The red bar indicates the percentage of Comparison Samples whose qualities you've already selected:

Direction increased immunogenicity

P value less than or equal to

Metadata category to color graph by Age

Advanced Parameters

Combine Genes in Step 3 with Genes in Step 4:

☒ 3 Intersect 4

☐ 3 Union 4

☐ 3 Minus 4

☐ 4 Minus 3

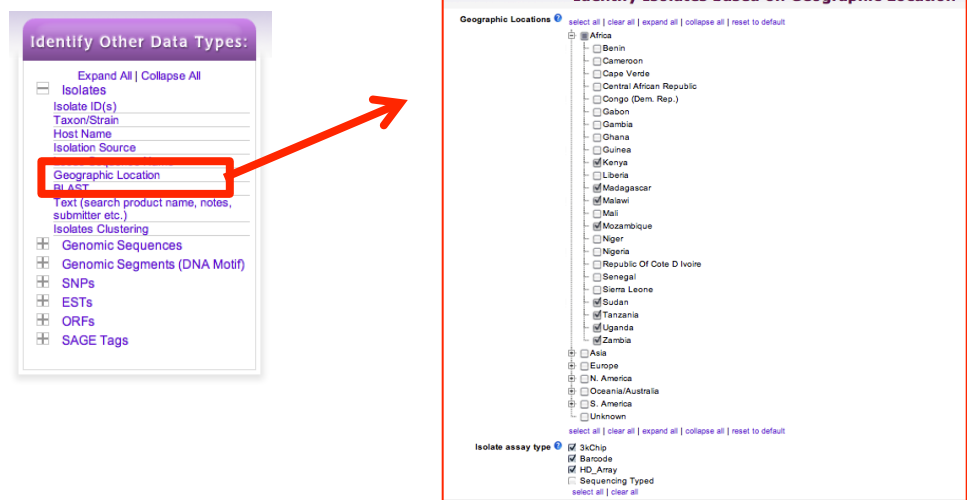
☐ 3 Relative to 4, using genomic colocation

Run Step

- How many genes are returned by this specific step?
- Of these, how many are present in your stringent list earlier?
- Now relax the parameters as you did in the ID list comparison to see how many genes are shared. *Hint: one quick way to do this is to delete the ID step from your previous strategy and add the host response step rather than changing the parameters for two steps.*

4.3a Optional Exercise: Find SNPs that differentiate isolates from East Africa and those from West Africa.

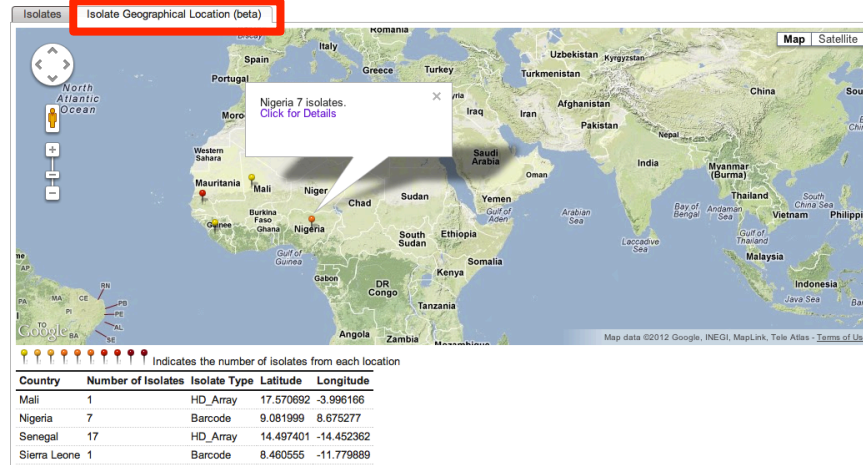
- For this exercise we are going to use the same 'Identify SNPs by Isolate Comparison' search as above. However, first we have to identify isolate IDs from West Africa and ones from East Africa. To do this use the 'Identify Isolates by Geographic Location' query under the isolates section (note that you will need to run this query twice, once for each set of countries):



Some East African countries: Kenya, Madagascar, Malawi, Mozambique, Tanzania, Sudan, Uganda, Zambia

Some West African Countries: Cameroon, Gabon, Liberia, Mali, Nigeria, Senegal, Sierra Leone

- For isolate assay type select HD_Array since this array has the most SNPs. You could also try the 3K_chip or even Barcode but shouldn't mix the assay types in one analysis.
- Confirm the distribution of the isolates you get by clicking on the "Geographic location" tab of the result page:



- Once you have isolates based on geographic location you will need to copy the IDs and paste them into the SNPs by isolate comparison query (make sure you put isolates from one set of countries into the input box for set A and the other set in the input box for Set B). You might find it useful to use the NotePad on your PC or open the query in another window or tab.
 - o To do this easily, click on “Download Results”, select “Tab delimited (Excel);” then unselect all the columns and click on “Get Report”. Now copy the list of IDs.
 - o If the above steps are taking too long, feel free to copy the IDs from the following link: [Need to add link here](#)
- Once you have the isolate IDs pasted in the isolate comparison query, run it and examine your results. Did you get any results? Revise the query and change the minimum percentage parameters to 70 for both set A and B:

◀
✕
Revise Step

Revise Step 1 : Isolate Comparison

Set A isolate identifiers ?

BC.458086; BC.458090; BC.458091;
BC.458092; BC.458093; BC.458101;
BC.458105; BC.458120; BC.458124;
BC.458125; BC.458126; BC.458127;
BC.458128; BC.458129; BC.458130;

☒ Enter list:
☐ Copy Isolates from My Basket (0 Isolates)

Minimum percentage of isolates in Set A with same allele >= ?

Set B isolate identifiers ?

BC.458098; BC.458110; BC.458111;
BC.458112; BC.458113; BC.458114;
BC.458115; BC.458116; BC.458117;
BC.458118; BC.458119; BC.458150;
BC.458168; BC.458169; CP3.273609;

☒ Enter list:
☐ Copy Isolates from My Basket (0 Isolates)

Minimum percentage of isolates in Set B with same allele >= ?

- What do your results look like now?
 - Which SNP differentiates more isolates (hint: look at the numbers in the columns for Set A and Set B)?
 - Do you think these SNPs are synonymous or non-synonymous? (hint: click on “select columns” and add the column called “non-synonymous”).
 - What are the genes that include these SNPs? (hint: click on the gene IDs in the “Gene ID” column).