

Population data (SNPs)

Exercise 4

4.1 Isolate comparison

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

a. Go to the “Identify SNPs based on Isolate Comparison” search.

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

The image shows two parts of the Plasmodb.org website. On the left is a sidebar titled "Identify Other Data Types:" with a list of categories. A red arrow points to "Isolate Comparison" under the "SNPs" category. On the right is the main search form titled "Identify SNPs based on Isolate Comparison".

Identify Other Data Types:

- Expand All | Collapse All
- Isolates
- Genomic Sequences
- Genomic Segments (DNA Motif)
- SNPs
 - SNP ID(s)
 - Gene ID
 - Genomic Location
 - Presence in isolate assay
 - Strain
 - Allele Frequency (HTS)
 - Isolate Comparison
- ESTs
- ORFs
- SAGE Tags
- Metabolic Pathways **BETA**
- Compounds **BETA**

Identify SNPs based on Isolate Comparison

Set A isolate identifiers Enter list: CP3.273609, CP3.273646, CP3.273647, CP3.273648, CP3.273649, CP3.273650, CP3.273651, CP3.273652

Copy Isolates from My Basket (0 Isolates)

Upload from a text file: Browse... Maximum size: 10MB.

Minimum percentage of isolates in Set A with same allele >= 100

Set B isolate identifiers Enter list: CP3.273597, CP3.273658, CP3.273660, CP3.273661, CP3.273663, CP3.273664, CP3.273665, CP3.273666, CP3.273667, CP3.273668, CP3.273670, CP3.273671

Copy Isolates from My Basket (0 Isolates)

Upload from a text file: Browse... Maximum size: 10MB.

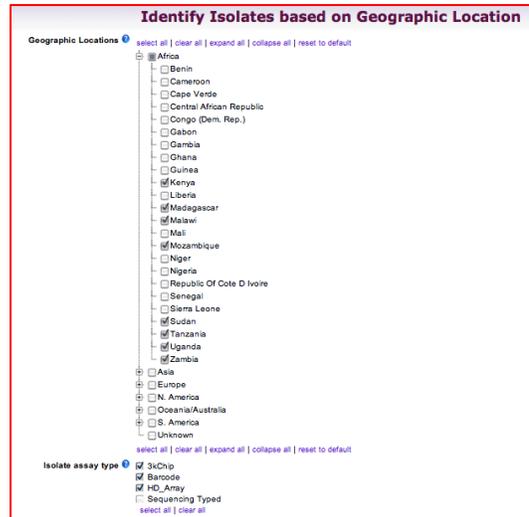
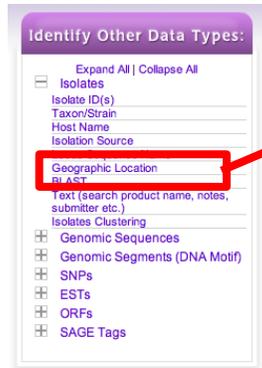
Minimum percentage of isolates in Set B with same allele >= 100

Advanced Parameters

b. **What does this search do?** What is in Set A and B? Run the query and look at your results. How many SNPs were identified between isolates from Brazil and Malawi? What could you use this information for?

c. **Find SNPs that differentiate isolates from East Africa and those from West Africa.**

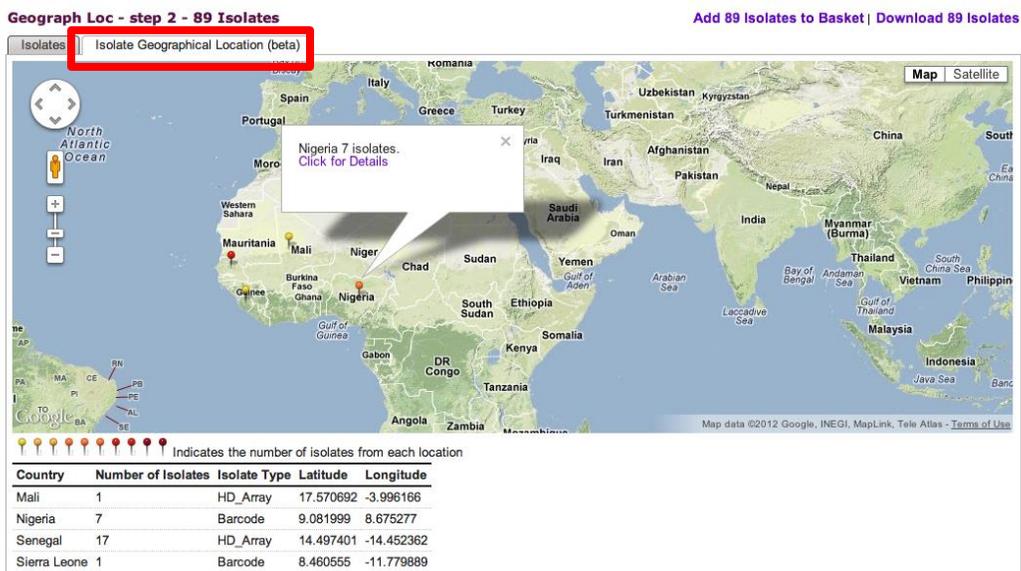
- For this exercise we are going to use the same ‘Identify SNPs by Isolate Comparison’ search as above. However, first we have to identify isolate IDs from West Africa and ones from East Africa. To do this use the ‘Identify Isolates by Geographic Location’ query under the isolates section (note that 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.

- 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.
- If the above steps are taking too long, feel free to copy the IDs from the following link: <http://goo.gl/rhRdO>
- 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 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 >= ?

Give this search a weight
 Give this search a name

- 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).

4.2 Analyzing SNPs on a defined list of genes.

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

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

- Create a strategy step with the gene IDs from the paper.

Hint: enter the above list into the 'Identify Genes based on Gene ID(s)' search option.

The image shows a screenshot of the PlasmODB web interface. On the left, a purple sidebar titled 'Identify Genes by:' contains a list of search criteria. The 'Gene ID(s)' option is circled in red. A red arrow points from this option to the main search form on the right. The search form is titled 'Identify Genes based on Gene ID(s)' and features a 'Gene ID input set' section with three radio buttons: 'Enter list' (selected), 'Copy Genes from My Basket (0 Genes)', and 'Upload from a text file:'. The 'Enter list' option is active, and a text area contains the gene IDs: Pf13_0191, PF13_0192, PF13_0194, PFL1385c, PFB0340c, MAL7P1.208, PF13_0348, PF10_0144, PF14_0102, PFE0080c, PFE0075c, PFD0955w. Below the input area are 'Advanced Parameters' and a 'Get Answer' button. At the bottom, there is a text box labeled 'Give this search a name'.

- Add a step to your strategy to identify how many of these genes are under diversifying selection.

Hint: the "Identify Genes based on SNP Characteristics" is found under the population biology menu (see figure on the next page).

- What parameters would you choose?
- Would you expect genes under balancing selection to have a high or low non-synonymous/synonymous SNP ratio?
- How many genes were returned by your search? Of these, how many intersect with the set of genes from the paper?
- Click on the result for your ID search in the first step (26 genes) and add columns for SNP characteristics (under population biology). Do all these genes appear to be under balancing selection? Is this consistent with the results of your strategy?

(Genes)

Gene ID(s)
26 Genes

Step 1

Add Step

Add Step

- Run a new Search for
- Transform by Orthology
- Add contents of Basket
- Add existing Strategy
- Filter by assigned Weight

- Genes
- Genomic Segments (DNA Motif)
- SNPs
- ORFs
- SAGE Tags

- Text, IDs, Species
- Genomic Position
- Gene Attributes
- Protein Attributes
- Protein Features
- Similarity/Pattern
- Transcript Expression
- Protein Expression
- Cellular Location
- Putative Function
- Evolution
- Population Biology

SNP Characteristics

Close

Add Step 2 : SNP Characteristics

Organism: Plasmodium falciparum 3D7

Reference: PF-3D7

Comparator: Pf-GHANA1

SNP Class: All SNPs

Number of SNPs of above class >= 1

Number of SNPs of above class <= 1000

Non-synonymous / synonymous SNP ratio >= 1

Non-synonymous / synonymous SNP ratio <= 10

SNPs per KB (CDS) >= 0

SNPs per KB (CDS) <= 10

Give this search a weight

Give this search a name

Combine Genes in Step 1 with Genes in Step 2:

- 1 Intersect 2
- 1 Union 2
- 1 Relative to 2, using genomic colocation
- 1 Minus 2
- 2 Minus 1

Run Step

SNPs
1275 Genes

Gene ID(s)
26 Genes
Step 1

8 Genes
Step 2

Add Step

4.3 Find genes with at least 30 non-synonymous SNPs that appear to be under diversifying selection when comparing all human isolates to a bear isolate.

For this exercise use <http://toxodb.org/toxo/>

- a. Navigate to the search 'Identify Genes based on HTS SNP Characteristics'
 - Use this search to find genes that contain SNPs that were identified by comparing high throughput sequencing data of parasite isolates.

The image shows a screenshot of the ToxoDB search interface. On the left, there is a green header 'Identify Genes by:' followed by a list of search criteria with expand/collapse icons. The criteria include: Text, IDs, Organism; Genomic Position; Gene Attributes; Protein Attributes; Protein Features; Similarity/Pattern; Transcript Expression; Protein Expression; Cellular Location; Putative Function; Evolution; Population Biology; and SNP Characteristics. A red arrow points from the 'SNP Characteristics' option in the left menu to a search box on the right. The search box is titled 'Identify Genes based on SNP Characteristics' and contains a dropdown menu with two options: 'SNP Characteristics' and 'HTS SNP Characteristics'. Below the dropdown is a button labeled 'Choose a Search' with the text 'Mouse over to read description'.

- b. Group isolates by host.
 - The datasets used in this search are high throughput sequencing data obtained on isolates. Meta data associated with these isolates include year collected, host, haplogroup, geographic location, ATCC#, strain or line name. In order to compare a bear isolate with all human isolates, we need to group the samples according to their host.
- c. Choose reference and comparison samples.
 - We want to use a bear isolate as reference and all human as comparator. Expanding the comparator Bear category shows that the B41 and B73 isolates are from a bear, so we want to choose either of these for our reference. Expand the comparator Human category shows about 20 samples sequenced from humans. Choose all samples by clicking the box next to Human.
- d. Set other parameters
 - SNP class = non-synonymous
 - Number of SNPs of above class ≥ 30
 - Non-synonymous / synonymous SNP ratio > 1

Identify Genes based on HTS SNP Characteristics

Organism

Group Comparator Samples by Meta Data

Reference

Comparator [select all](#) | [clear all](#) | [expand all](#) | [collapse all](#) | [reset to default](#)

- Bear
- Cat
- Chicken
- Cougar
- Dog
- Goat
- Human
- Jaguar
- Pig
- Sheep
- Toucan
- unknown

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#) | [reset to default](#)

Number of aligned reads >=

Allele frequency >=

P value <=

SNP Class

Number of SNPs of above class >=

Number of SNPs of above class <=

Non-synonymous / synonymous SNP ratio >=

Non-synonymous / synonymous SNP ratio <=

SNPs per KB (CDS) >=

SNPs per KB (CDS) <=

Advanced Parameters

e. How many genes did you get?

f. How can you quickly get an idea of the function of the genes in your result set?

- Hint: use the analysis function on the Product Description column. Click the graph icon to analyse the data in the column. Since Product Description is text, the analysis performed is a word cloud. You can exclude common words using the Filter Words By Rank

My Strategies: New Opened (1) All (115) Basket Exam

(Genes)

HTS SNPs
216 Genes
Step 1

Add Step

216 Genes from Step 1
Strategy: HTS SNPs(4)

Filter by organism or strain (results removed by the filter will not be combined into the next step)
Filter by strains (advanced) (results removed by the filter will not be combined into the next step)

Gene Results Genome View

First 1 2 3 Next Last Advanced Paging

Gene ID	Gene Group (representative gene)	Product Description	total HTS SNPs
TGME49_275710	TGGT1_000620	hypothetical protein	40
TGME49_253400	TGGT1_002480	hypothetical protein	124
TGME49_254720	TGGT1_004270	dense granule protein	57
TGME49_254820	TGGT1_004380		
TGME49_281980	TGGT1_005490		

Word Cloud

Graph Data (text)

Filter words by rank: 7 to 50
Use slider or enter numbers to adjust filter

Sort by: Rank A-Z

Mouse over a word to see its occurrence in the column

sag sequence rhoptry dense factor
granule ap2 finger kinase putative aaa atpase gondii phosphatase toxoplasma
transcription type brp1 catalytic gra15 gra4 gra6 gra7 gra8 hmc3 incomplete kruf mic17a ron4 rop17 rop2a rop39 rop40 rop35 ers22c ers25c ers35a ers36a ers36b
ers33a ers43 ers49c ers57 tracd

- g. How can you quickly get an idea of the function of the genes in your result set?
- The genome view tab displays your gene results mapped onto the chromosome (or corresponding genomic sequence).
 - Hover over the glyphs to see information about the gene and to link to GBrowse at the location of the gene.

216 Genes from Step 1
Strategy: HTS SNPs(4)

Filter by organism or strain (results removed by the filter will not be combined into the next step.)
 Filter by strains (advanced) (results removed by the filter will not be combined into the next step.)

Gene Results **Genome View**

First 1 2 3 Next Last Advanced Paging

Gene ID	Gene Group (representative gene)	Product Description	Total HTS SNPs
TGME49_275710	TGGT1_000620	hypothetical protein	40
TGME49_253400	TGGT1_002480	hypothetical protein	40
TGME49_254720	TGGT1_004270	derivative of GR	40

[Add 216 Genes to Basket](#) | [Download 216 Genes](#)

Filter by organism or strain (results removed by the filter will not be combined into the next step.)
 Filter by strains (advanced) (results removed by the filter will not be combined into the next step.)

Gene Results **Genome View**

Genes on forward strand:
 Genes on reverse strand:

Showing 1 to 20 of 20 entries Show 25 entries Search:

Sequence	Organism	Chromosome	#Genes	Length	Gene Locations
TGME49_chrX	Toxoplasma gondii ME49	X	29	7486190	
TGME49_chrXII	Toxoplasma gondii ME49	XII	29	7094428	

TGME49_227115
 start: 1,080,826, end: 1,086,233, on forward strand of TGME49_chrX
 hypothetical protein
[Record page](#)
[Browse](#)

4.5 Optional Exercise: Find genes that contain SNPs that distinguish *Toxoplasma gondii* strains isolated from chickens as compared to those isolated from cats.

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

Note that this exercise combines three different data types (Isolated, SNPs and Genes). Lets start by doing a SNP search to identify SNPs that distinguish these two populations. Navigate to “Identify SNPs based on Isolate Comparison (HTS)”.

- Choose Host from the “Group Comparator Samples by Meta Data” parameter. This will generate the reference and comparator menus using the meta data associated with these samples. In this case the isolates are organized the host from which they were isolated. These isolates have a fairly rich set of meta data so there are other options such as haplogroup or geographic location.
- Choose all Cat samples from the reference list. Note that you should “clear all” or check off the initial samples that are checked.

- c. Choose all Chicken samples from the comparator list.
- d. We let the percentages stay at 80 but you can play with these parameters to get more or fewer SNPs back. For this exercise 80 works well.

Identify SNPs based on Isolate Comparison (HTS)

Organism ?

Group Comparator Samples by Meta Data ?

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

- Bear
- Cat
- Chicken
- Cougar
- Dog
- Goat
- Human
- Jaguar
- Pig
- Sheep
- Toucan
- unknown

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

Minimum percentage of isolates in Set A with same allele \geq ?

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

- Bear
- Cat
- Chicken
- Cougar
- Dog
- Goat
- Human
- Jaguar
- Pig
- Sheep
- Toucan
- unknown

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

Minimum percentage of isolates in Set B with same allele \geq ?

+ Advanced Parameters

How many SNPs were returned? Were you surprised at this fairly large number? Why might there be so many? It turns out that the *T. gondii* genome is quite divergent ... there are on the order of 2.3 million SNPs between all the strains that have been re-sequenced in the 65 MB genome.

If the Gene ID column is not present please add it. How many of the SNPs seem to be in genes (estimate, please don't count). Are some of the genes enriched for SNPs? I.e., does it seem that some of the genes may be involved in the differential host preference?

How might you ask this question directly? What we really want is to identify the genes that contain these SNPs that we've found. We can do this directly ... can you figure out how to do it without looking further at the help?

- e. Click the “Add Step” button and add a search for “Genes by gene type” and set the parameters to return all *T. gondii* ME49 protein coding genes. Notice that the ONLY option available to Combine the steps is the last one Why is this??
- f. In the Genomic Colocation window “Return each gene from step 2 whose exact region overlaps with the exact region of a SNP in step 1 on either strand”. NOTE: we will cover genomic colocation in more detail later so don’t worry if you are confused!!

Add Step 2 : Gene Type

Organism ? [select all](#) | [clear all](#) | [expand all](#) | [collapse all](#) | [reset to default](#)

- Eimeria
 - Eimeria tenella strain Houghton
- Neospora
 - Neospora caninum Liverpool
- 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...](#)

Genomic Colocation

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

You had 1261 SNPs in your Strategy (Step 1). Your new Genes search (Step 2) returned 8320 Genes.

"Return each whose **exact region** overlaps the **exact region** of a SNP in Step 1 and is on

(8320 Genes in Step)



Region

Gene

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: bp

end at: bp

(1261 SNPs in Step)



Region

SNP

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: bp

end at: bp

[Close](#)

How many genes were returned? Note that some of the genes have many matched regions ... ie SNPs from step 1 that are contained within (overlap since SNPs are 1 bp long) the gene. In order to make this easier to look at, remove the "region" and "matched regions" columns. What do you think it means that there are so many SNPs in some of these genes? What kinds of genes are these (look at the product description column). Might these genes be involved in host preference?

Is there anything interesting about how these genes are distributed in the genome? *Hint: click the Genome View tab.* What do you think this means?

Are there other searches that you might want to add to this strategy in order to better understand this result and help refine your hypotheses?