Genetic Exercises

SNPs and Population Genetics

Introduction to single nucleotide polymorphism searches in EuPathDB:

Each isolate's sequencing reads are aligned to the reference genome (Organism) and SNPs are recorded for each isolate based on the Read Frequency Threshold. Then, scanning SNP locations across isolates in Set A and B separately, SNPs within each set are recorded if they meet the major allele frequency and percent isolates with a base call. When two groups of isolates are compared to each other (set A and B), SNPs returned by the search are ones that differ between Isolates in Set A and B.

Organism: Choosing an Organism.

The Organism parameter defines the species of the parasite isolates and the genome in which the SNPs are determined. Choosing an Organism focuses the Isolates parameter to the list of isolates of that organism, thus changing the subset of isolates available when forming your group.

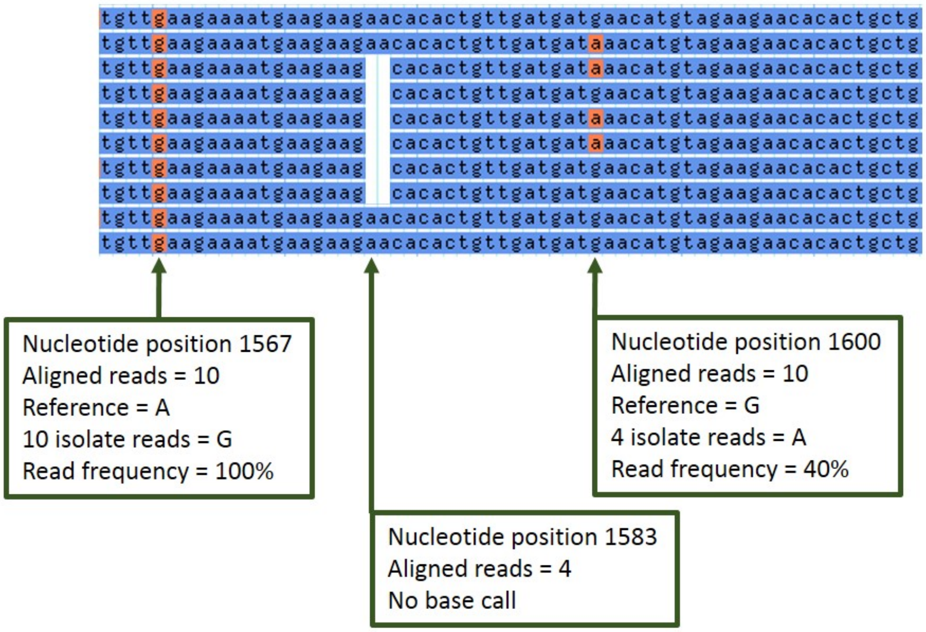
Isolates: Choosing isolate group.

Isolate sequences are accompanied by information about the sequencing data set - metadata such as the clinical phenotype of the host or the location collected. By default, the isolate group includes all isolates from the Organism you chose. You have the option to limit the group of isolates based on metadata characteristics. For example, you can find SNPs for isolates collected from a certain geographic location. To do this, click a metadata category (Country) on the left and choose your desired metadata characteristic (e.g. Peru) on the right. Multiple metadata categories and characteristics can be chosen when defining your group. Not all isolate data sets have the same metadata. A running total of your selections is displayed above the metadata filter.

Read Frequency Threshold: Calling SNPs for each isolate in a set.

Each isolate's sequencing reads are aligned to a reference genome (Organism) and then each nucleotide position with 5 or more aligned reads is examined. A base call is made if the aligned reads meet your Read Frequency Threshold. For example, Isolate X has 10 aligned reads at nucleotide position 1600. If 6 reads are G and 4 reads are A, the read frequency is 40%. Running this search with the Read Frequency Threshold set to 80% will prevent a base call and consequently exclude Isolate X when returning SNPs for nucleotide position 1600. The parameter lets you control the quality of the sequencing data and the confidence of the SNP calls. Read Frequency Threshold is a particularly important parameter when dealing with diploid (or aneuploid) organisms since a read frequency of ~50% is expected for heterozygous SNPs.

Returned SNPs by searches in EuPathDB are found by aligning the isolates (with SNPs recorded as described above) and applying 2 parameters across the group of isolates: Percent isolates with a base call and Minor allele frequency. When comparing two groups of isolates, the SNPs returned are found by comparing SNPs from Set A Isolates with those from Set B. SNPs are recorded for each set by aligning the isolates in each set (with isolate specific SNPs recorded as described above for read frequency threshold) and applying 2 parameters across the isolate sets: Percent isolates with a base call and Major allele frequency. SNPs returned by the search have different base call between Set A isolates and Set B isolates.


Isolate X aligned sequencing reads

tgttgaagaaaatgaagaagaacacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaagaacacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgataaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaag cacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaagaacacactgttgatgatgaacatgtagaagaacacactgctg
tgttgaagaaaatgaagaagaacacactgttgatgatgaacatgtagaagaacacactgctg
Nucleotide position 1567
Aligned reads = 10
Reference = A
10 isolate reads = G
Read frequency = 100%
Nucleotide position 1600
Aligned reads = 10
Reference = G
4 isolate reads = A
Read frequency = 40%
Nucleotide position 1583
Aligned reads = 4
No base call

Percent isolates with a base call: Parameter 1 for calling SNPs across your isolate group

Sometimes an isolate does not have a base call at a certain nucleotide position because the Read Frequency Threshold was not met or because there were less than 5 aligned sequencing reads for that nucleotide position. In this case, a SNP can be returned by the search based on a subset of your isolate group. The 'Percent isolates with a base call' parameter defines the fraction of isolates that must have a base call before a SNP is returned for that nucleotide position. The default setting for this parameter is 80% or 8 out of 10 isolates in your group must have a base call for a SNP to be returned by the search.

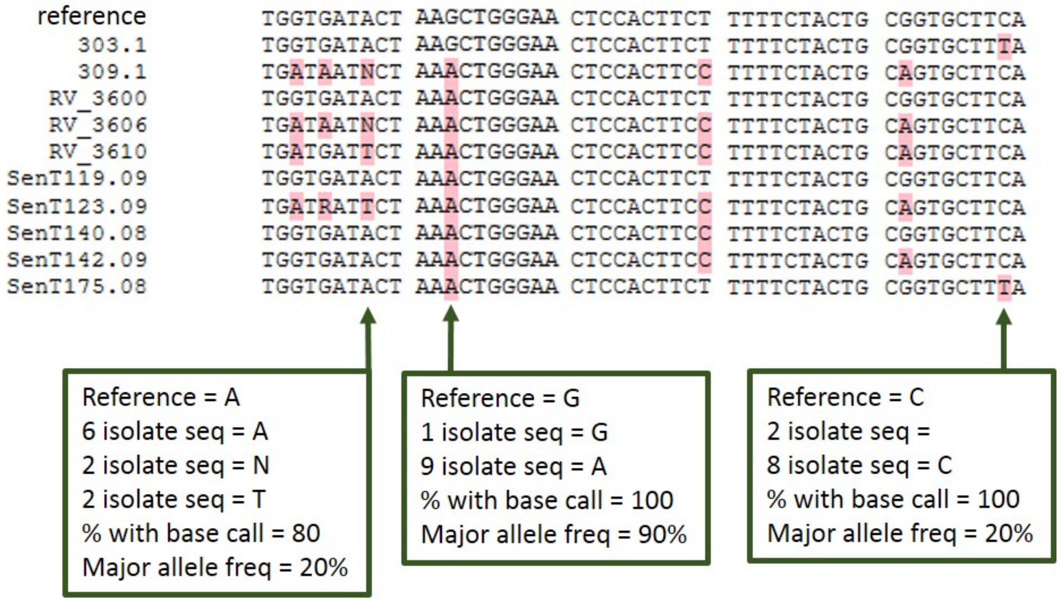
Minor allele frequency: Parameter 2 for calling SNPs across your isolate group.

The minor allele frequency refers to the least common base call for a single nucleotide position across all isolates. The default setting for this parameter is 0% and returns all SNPs - instances where at least one isolate has a base call that differs from reference. Increase the Minor allele frequency to ensure that SNPs returned by the search are shared by two or more isolates in your group.

Major Allele Frequency: Parameter 2 for recording SNPs in isolate sets

The major allele frequency is the frequency of the most common SNP across the isolates in a Set. The default setting for this parameter is 80%. SNPs returned by this search have a different SNP call between Set A and Set B. See image below.

Set A aligned isolate sequences.



```
  reference   TGGTGATACT AAGCTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTCA
      303.1   TGGTGATACT AAGCTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTTA
      309.1   TGATAATNCT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
    RV_3600   TGGTGATACT AAACTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTCA
    RV_3606   TGATAATNCT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
    RV_3610   TGATGATTCT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
 SenT119.09   TGGTGATACT AAACTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTCA
 SenT123.09   TGATRATTCT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
 SenT140.08   TGGTGATACT AAACTGGGAA CTCCACTTCC TTTTCTACTG CGGTGCTTCA
 SenT142.09   TGGTGATACT AAACTGGGAA CTCCACTTCC TTTTCTACTG CAGTGCTTCA
 SenT175.08   TGGTGATACT AAACTGGGAA CTCCACTTCT TTTTCTACTG CGGTGCTTTA
```

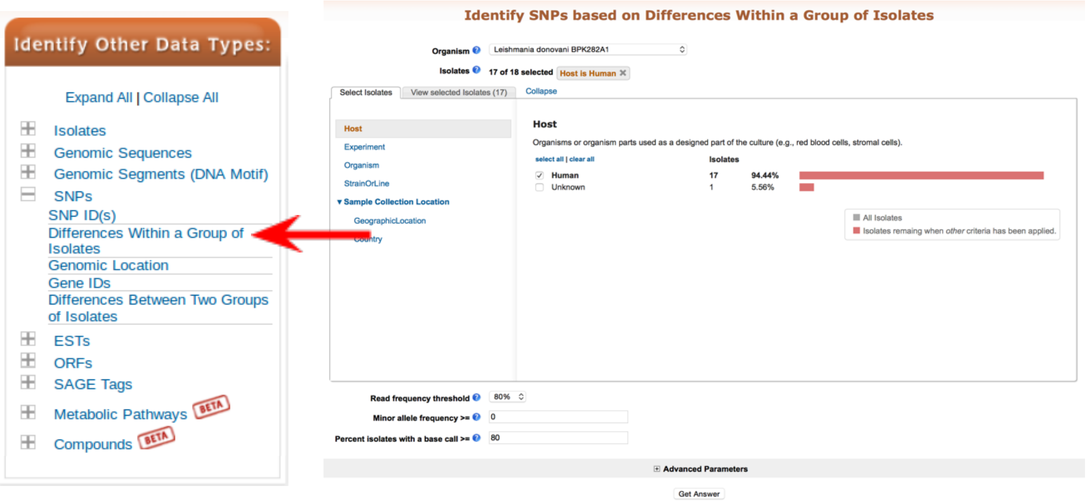
Reference = A
6 isolate seq = A
2 isolate seq = N
2 isolate seq = T
% with base call = 80
Major allele freq = 20%

Reference = G
1 isolate seq = G
9 isolate seq = A
% with base call = 100
Major allele freq = 90%

Reference = C
2 isolate seq =
8 isolate seq = C
% with base call = 100
Major allele freq = 20%

1. Identify 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 “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)
Differences Within a Group of Isolates
Genomic Location
Gene IDs
Differences Between Two Groups of Isolates
ESTs
ORFs
SAGE Tags
Metabolic Pathways
BETA
Compounds
BETA
Select Isolates
View selected Isolates (17)
Collapse
Host
Experiment
Organism
StrainOrLine
▾ Sample Collection Location
GeographicLocation
Country
select all | clear all
■ All Isolates
■ Isolates remaing when *other* criteria has been applied.
⊞ Advanced Parameters
Get Answer

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%

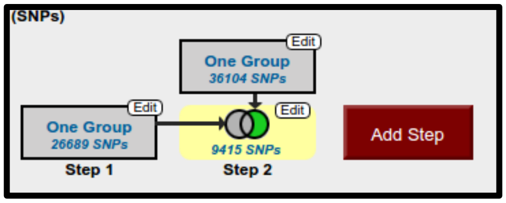
Read frequency threshold: 80%

Minor allele frequency >=: 0

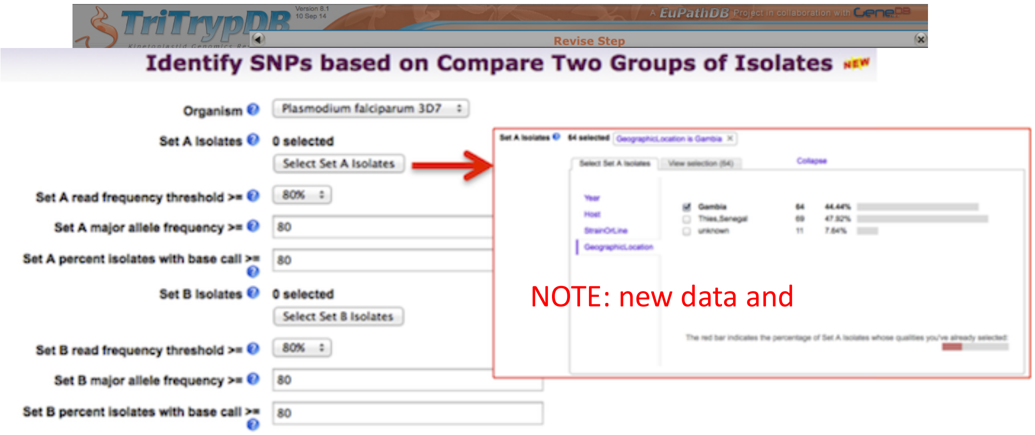
Percent isolates with a base call >=: 80

   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 second step results to view them. 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).*



Identify SNPs based on Compare Two Groups of Isolates NEW

Organism: Plasmodium falciparum 3D7

Set A Isolates: 0 selected

Select Set A Isolates

Set A read frequency threshold >= 80%

Set A major allele frequency >= 80

Set A percent isolates with base call >= 80

Set B Isolates: 0 selected

Select Set B Isolates

Set B read frequency threshold >= 80%

Set B major allele frequency >= 80

Set B percent isolates with base call >= 80

Set A Isolates 64 selected GeographicLocation is Gambia

Select Set A Isolates | View selection (64) | Collapse

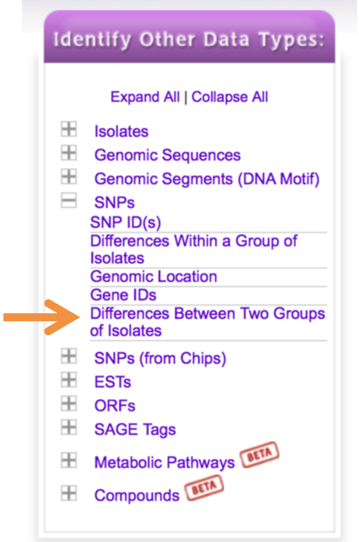
Year			
Host	☑ Gambia	64	44.44%
StrainOrLine	□ Thies,Senegal	69	47.92%
GeographicLocation	□ unknown	11	7.64%

The red bar indicates the percentage of Set A isolates whose qualities you've already selected

NOTE: new data and

- 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?

2. Find SNPs that differentiate between groups of isolates. Drug sensitive vs. drug resistant.
For this exercise use http://plasmodb.org


Identify Other Data Types:
Expand All | Collapse All
Isolates
Genomic Sequences
Genomic Segments (DNA Motif)
SNPs
SNP ID(s)
Differences Within a Group of Isolates
Genomic Location
Gene IDs
Differences Between Two Groups of Isolates
SNPs (from Chips)
ESTs
ORFs
SAGE Tags
Metabolic Pathways BETA
Compounds BETA

Why would you want to compare between groups of isolates? One possibility is to compare between drug sensitive and resistant parasites, another is to compare strains between different geographic regions. Grouping isolates requires some knowledge about isolate characteristics (metadata). You can identify SNPs between two groups of isolates using the "Compare Two Groups of Isolates" query found under the SNPs heading in the "Identify other Data Types" section.

To set this query up, there are two main things you need to do:

a. Define the two sets of isolates (set A and B) based on available metadata or based on your own knowledge of individual isolate/strain characteristics.
b. Define the SNP characteristics in each set of isolates.
    - For this exercise find all SNPs that differentiate isolates from Gambia compared to those from Senegal.

    - Define the SNP characteristics to be as follows:

- Read frequency threshold >= 80%
- Major allele frequency >= 70
- Percent isolates with base call >= 50

- What do these parameters mean? Here are some definitions (you can get these by mousing over the blue question mark icons next to the parameter names).
- *Frequency Threshold*: The percent of aligned reads at this SNP location with this allele. In a perfect world with a haploid organism 100% of reads should support all SNP calls. For diploid or polyploid organisms, the read frequency threshold to identify heterozygous SNPs would be 50% or less.
- *Major Allele Frequency*: The percent of major alleles at this location (for the specified isolates at the specified read frequency). The major allele in a haploid organism is defined as the allele that is found in the majority of isolates/strains in a defined set. So, if you choose a major allele frequency of 70% in a group of 10 isolates, than 7 of those isolates should have the same allele.
- *Percent Isolates with Base Call*: The percent of isolates with the SNP call. For example, if you choose 20 isolates and a 'Min percent of isolates with base calls' of 75%, then the SNP will be ignored if there are less than 15 isolates that have a qualifying allele.

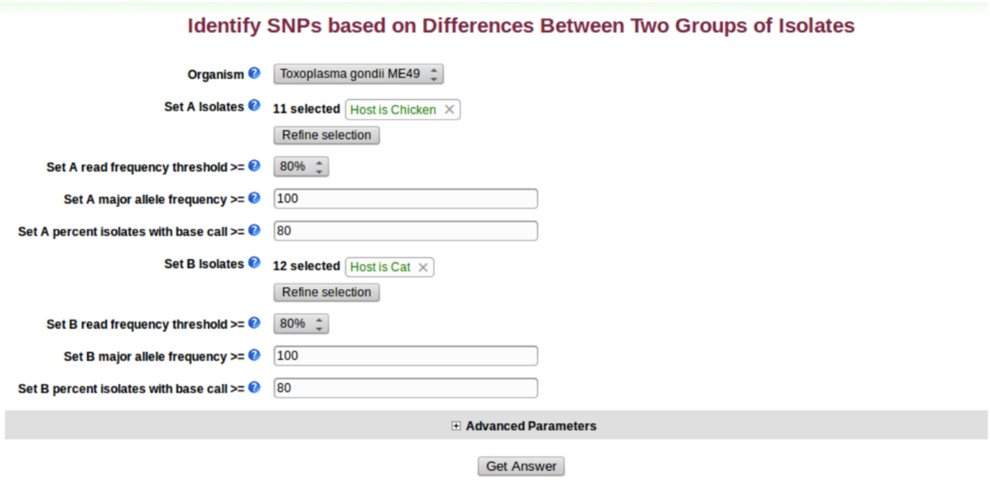
c. How many results did you get? Run this search again but compare SNPs from French Guiana with SNPs from Mali. Leave the other parameters at default values which is more stringent. How many SNPs did you get? Why would you expect more SNPs in this comparison than in the previous search?
   - You can define sets of isolates by other criteria. For example, you may wish to compare known chloroquine resistant and sensitive strains. Use the compare two groups query to compare these isolates:
     - Set A Isolates: 7G8, Dd2-1, Dd2-2, GB4 (resistant)
     - Set B Isolates: 3D7, CS2, HB3, IT (sensitive)
   - Define the SNP characteristics. For example, read frequency threshold >=80%, major allele frequency >= 100, percent isolates with base call >= 100. What happens if you change these numbers?

3. **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 “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 box to select the 11 chicken isolates.

b. Click select set B isolates and select hosts from the left column. Check the cat 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 ✕

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

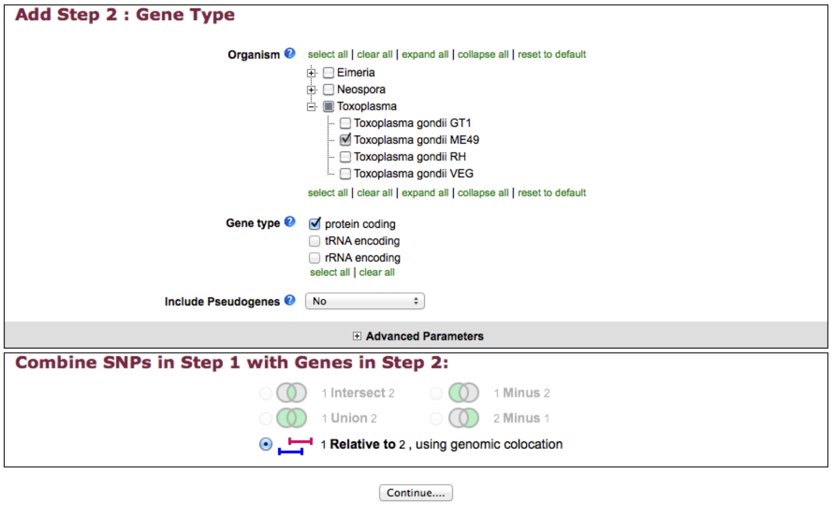
⊞ Advanced Parameters

Get Answer

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?
- *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?

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

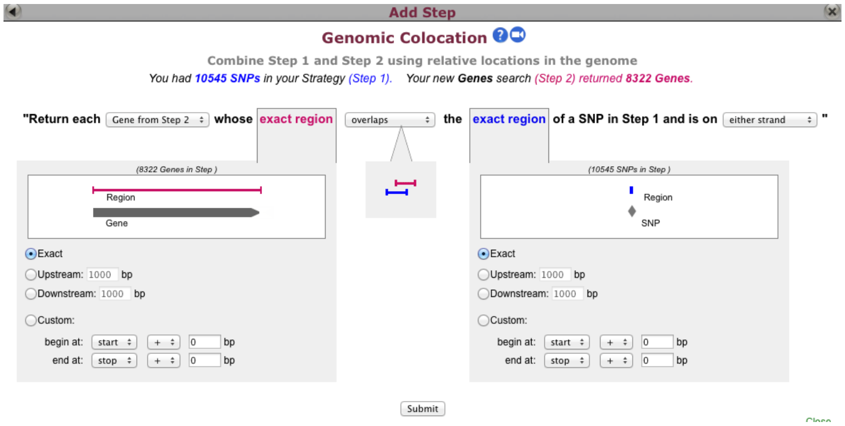
Include Pseudogenes: No

⊞ 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

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 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:

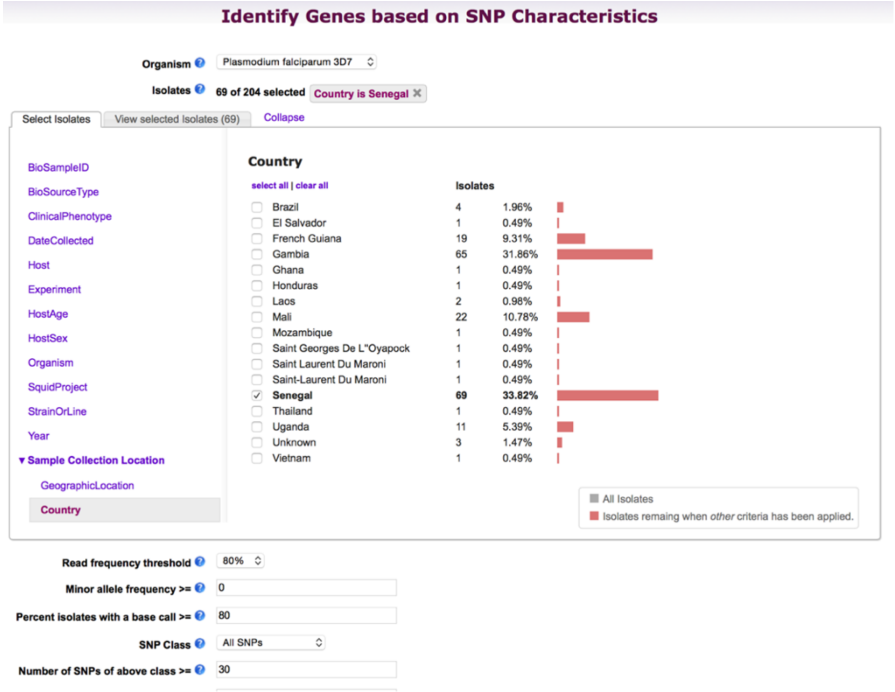
begin at: start + 0 bp

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: http://toxodb.org/toxo/im.do?s=f6cdff8edcda494b

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

a. Go to the "Identify Genes based on SNP Characteristics" search. *Hint: you can find this under "Identify Genes" in the "Population Biology" section.*



Identify Genes based on SNP Characteristics

Organism: Plasmodium falciparum 3D7

Isolates: 69 of 204 selected Country is Senegal ✕

Country

	Isolates	
□ Brazil	4	1.96%
□ El Salvador	1	0.49%
□ French Guiana	19	9.31%
□ Gambia	65	31.86%
□ Ghana	1	0.49%
□ Honduras	1	0.49%
□ Laos	2	0.98%
□ Mali	22	10.78%
□ Mozambique	1	0.49%
□ Saint Georges De L"Oyapock	1	0.49%
□ Saint Laurent Du Maroni	1	0.49%
□ Saint-Laurent Du Maroni	1	0.49%
☑ **Senegal**	**69**	**33.82%**
□ Thailand	1	0.49%
□ Uganda	11	5.39%
□ Unknown	3	1.47%
□ Vietnam	1	0.49%

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

Read frequency threshold: 80%

Minor allele frequency >=: 0

Percent isolates with a base call >=: 80

SNP Class: All SNPs

Number of SNPs of above class >=: 30

- Choose strains from *P. falciparum* (organism) that are from 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? (*Hint: use the Enrichment Analysis tool to get a quick overview*). Does this make sense as genes that might be 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.

b. 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 PF3D7_1475800 still the gene with the largest NS/S ratio? *Hint: Add a column for Population Biology 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.

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 revising it to decrease stringency of SNP searches to 10 coding SNPs and NS/S ratio >= 1.5.*