Genetic Exercises

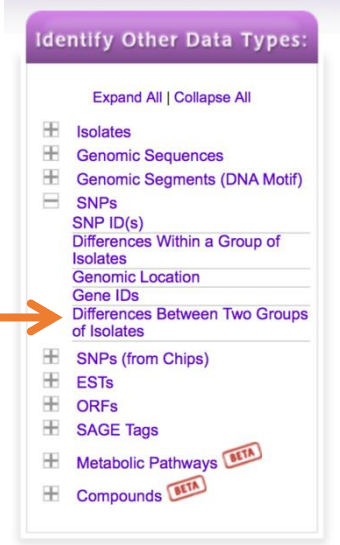
SNPs and Population Genetics

Single Nucleotide Polymorphisms (SNPs) in EuPathDB can be used to characterize a group of isolates or to distinguish between two groups of isolates. They can also be used to identify genes that may be under evolutionary pressure, either to stay the same (purifying selection) or to change (diversifying or balancing selection). There are two types of isolate SNP data in EuPathDB – predetermined SNPs that come to us as SNP-chip array data and SNPs that EuPathDB calls from whole genome re-sequencing data. EuPathDB analyzes whole genome resequencing data by aligning each isolate's sequencing reads to a reference genome and comparing the sequence base-by-base to call SNPs against the reference genome. SNP-chip array data is analyzed much like microarray data with SNP calls based on 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. **Find SNPs that differentiate between isolates from Gambia and isolates from Senegal.**
   For this exercise use http://plasmodb.org

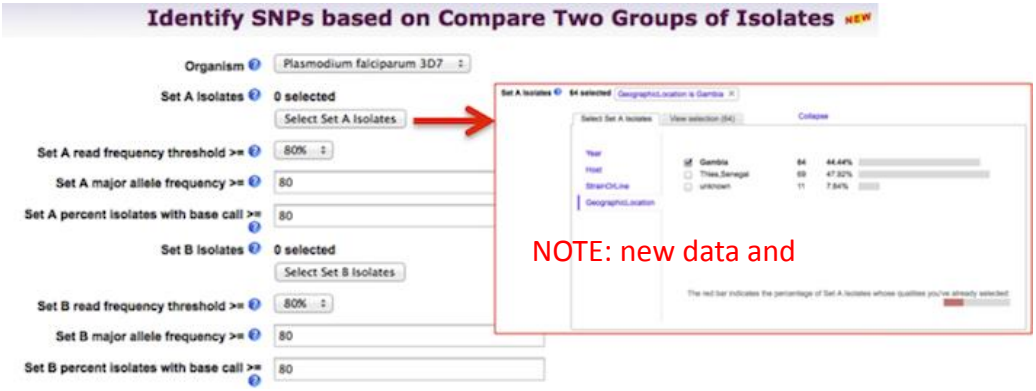
   Comparing SNPs that differentiate between two groups of isolates can be useful to distinguish drug sensitive and resistant parasites or to track changes between isolates from different geographic regions. EuPathDB integrates the isolate characteristics such as collection location and age of host and offers a way to group isolates based on these characteristics. 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.

- Open the Set A Isolates parameter. Click Geographic Location on the left and then choose Gambia.
- Define the SNP characteristics for Set A Isolates to be as follows:
  - Read frequency threshold >= 80%
  - Major allele frequency >= 70
  - Percent isolates with base call >= 50
  - What do these parameters mean? You can see definitions by mousing over the "?" icon by each parameter or by reading the more detailed description of the search at the bottom of the search page.
- Repeat this process for Set B Isolates choosing Senegal for the genomic location.



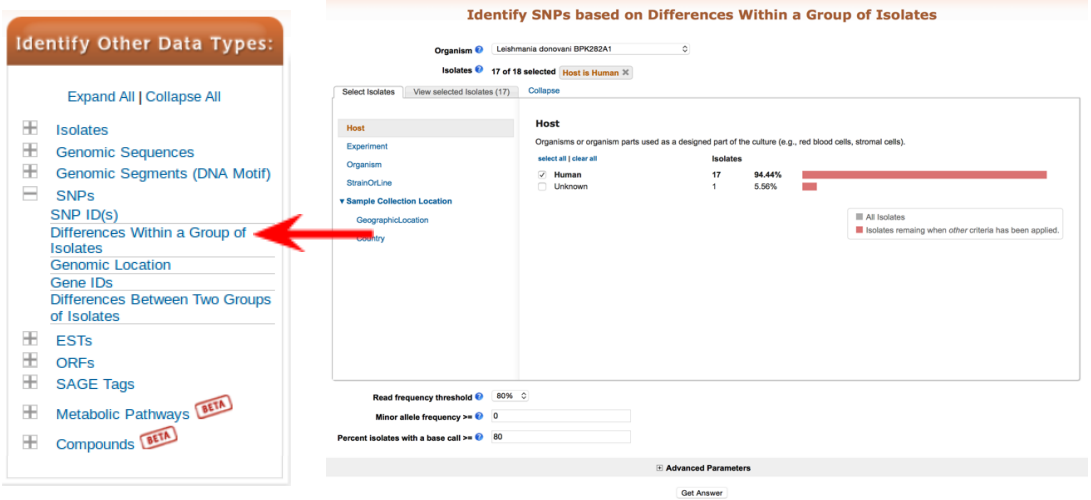
c. How many results did you get? Run this search again but compare SNPs from French Guiana with SNPs from Mali. Leave the other parameters at default values which is more stringent. How many SNPs did you get? Why would you expect more SNPs in this comparison than in the previous search?

2. Identify SNPs within a group of Isolates

For this exercise use http://TriTrypdb.org

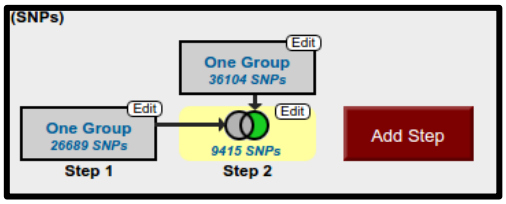
a. Go to the "Differences Within a Group of Isolates" search.

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

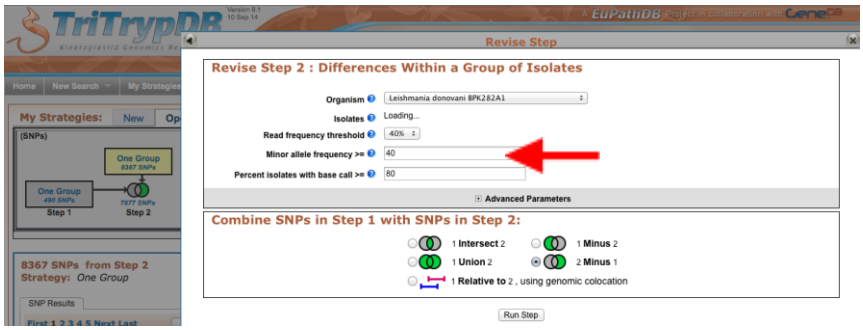


b. **What does this search do?** Choose *Leishmania donovani* for the organism and select isolates from the human host. Use default parameters for the rest of the parameters. Run the query and look at your results.

- How many SNPs were returned?
- Are any of these heterozygous SNPs? 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?



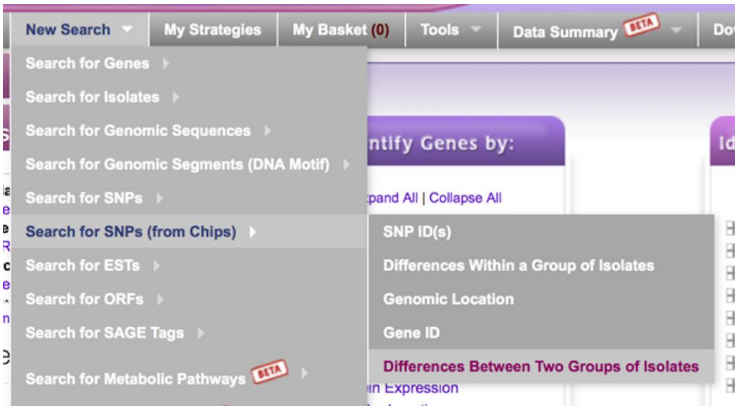
- Click on the second step results to view them. What do you notice about the %minor alleles? (*many are quite low … ie in one or two of the isolates*). How can you remove these from your search results? *Hint: revise this search and increase the minor allele frequency threshold (try 20 and 40 and compare results).*



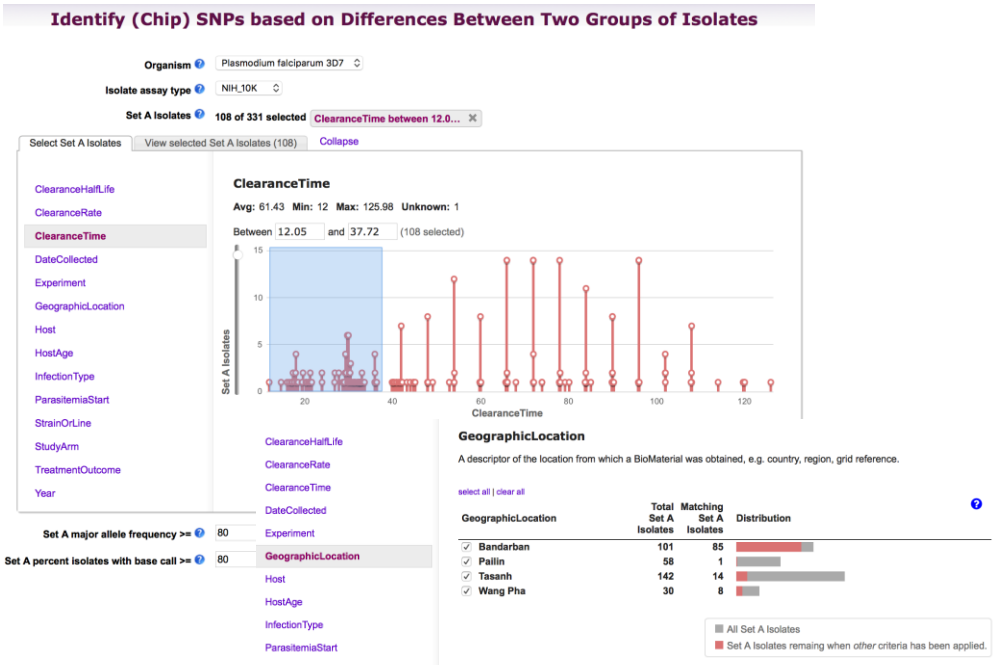
- Why might you want to increase the minor allele threshold when you run SNP searches?
- Try increasing / decreasing the "Percent isolates with base call". How does this impact your results? Why might you want to change this parameter?
- Go to a record page for a SNP with a high minor allele frequency. What do you see in the Strains table? Why are many of the strains repeated?
- OPTIONAL: Can you identify putative heterozygous SNPs in *Trypanosoma brucei* TREU927? How would you do this? How many do you find? Why is there a difference if you choose both TREU927 and TREU927_ resequence1 rather than just TREU927_resequence1? *Hint: see http://tinyurl.com/hetSNPs and look at the strains table on a record page noting the read frequency column for strain TREU927.*

3. **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 Chips).



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).*



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 + 1000 to define the region.*