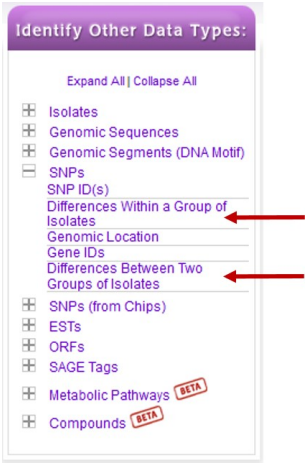
Metadata and Single Nucleotide Polymorphisms (SNPs)

PlasmoDB contains several population biology data sets based on whole genome re-sequencing of isolates. Each isolate sequence is aligned to a reference genome and single nucleotide polymorphisms (SNPs) are recorded. Metadata associated with the isolates (*e.g*. country collected or drug sensitivity status) are integrated in a way that allows for grouping isolates by shared characteristics despite originating from different data sets. Two searches are available to query these data. One search returns SNPs that are shared by one set of isolates. The other compares two sets of isolates to find SNPs that differ between the two sets.

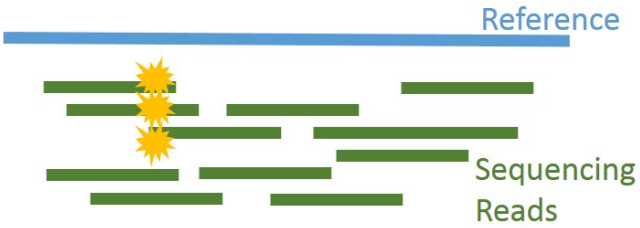


1. **Find SNPs that differentiate between *P. vivax* isolates collected in Mexico from isolates collected in Thailand.**

a. Navigate to the search – "Identify SNPs based on Differences between Two Groups of Isolates".

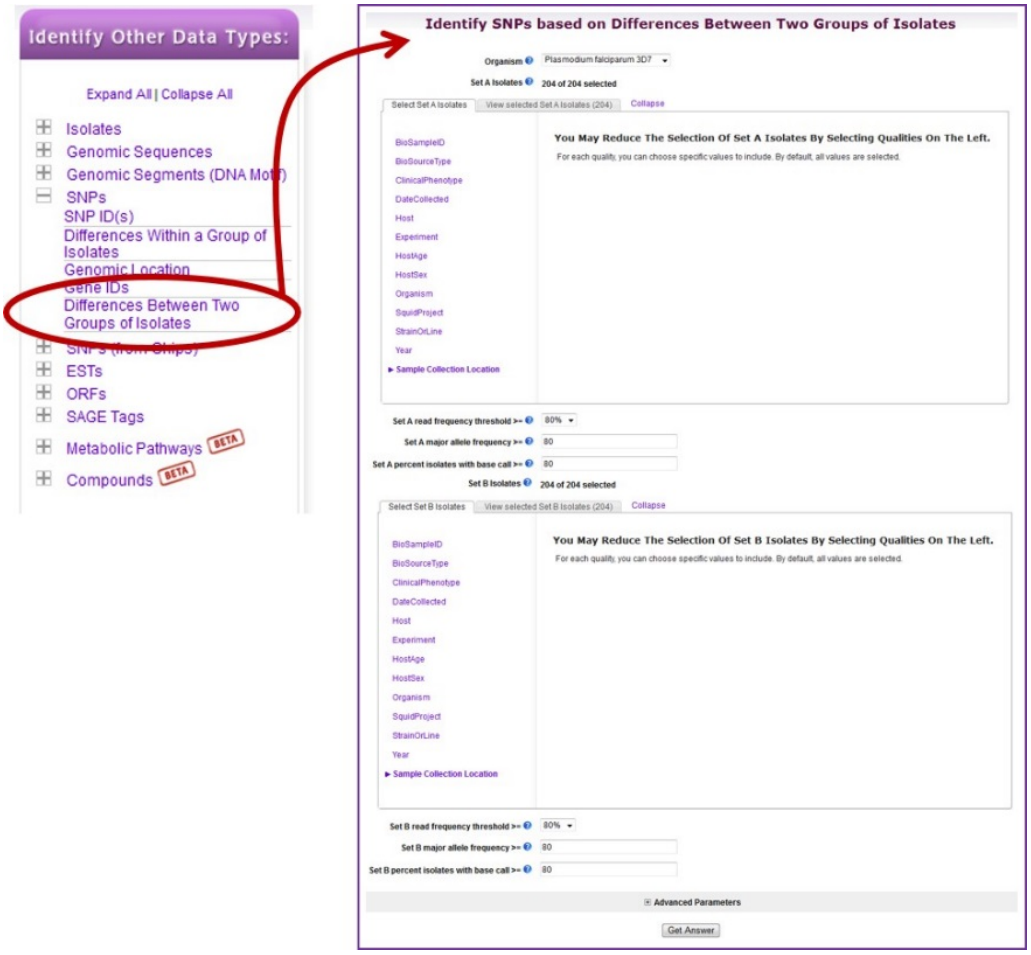
The search does three things:

- First each isolate's sequencing reads are aligned to a reference genome (*P. vivax* Sal1) and nucleotide differences are recorded as SNPs. You control whether a nucleotide difference is called a SNP with the **read frequency threshold** parameter. The portion of an isolates aligned reads that support a SNP call is the read frequency threshold.



3/5, frequency = 0.6

- The search then aligns isolate sequences (separately for Set A and Set B) and applies your SNP characteristic parameters, returning SNPs that meet your parameter values.
- The search compares Set A and Set B SNPs and returns those that differ.



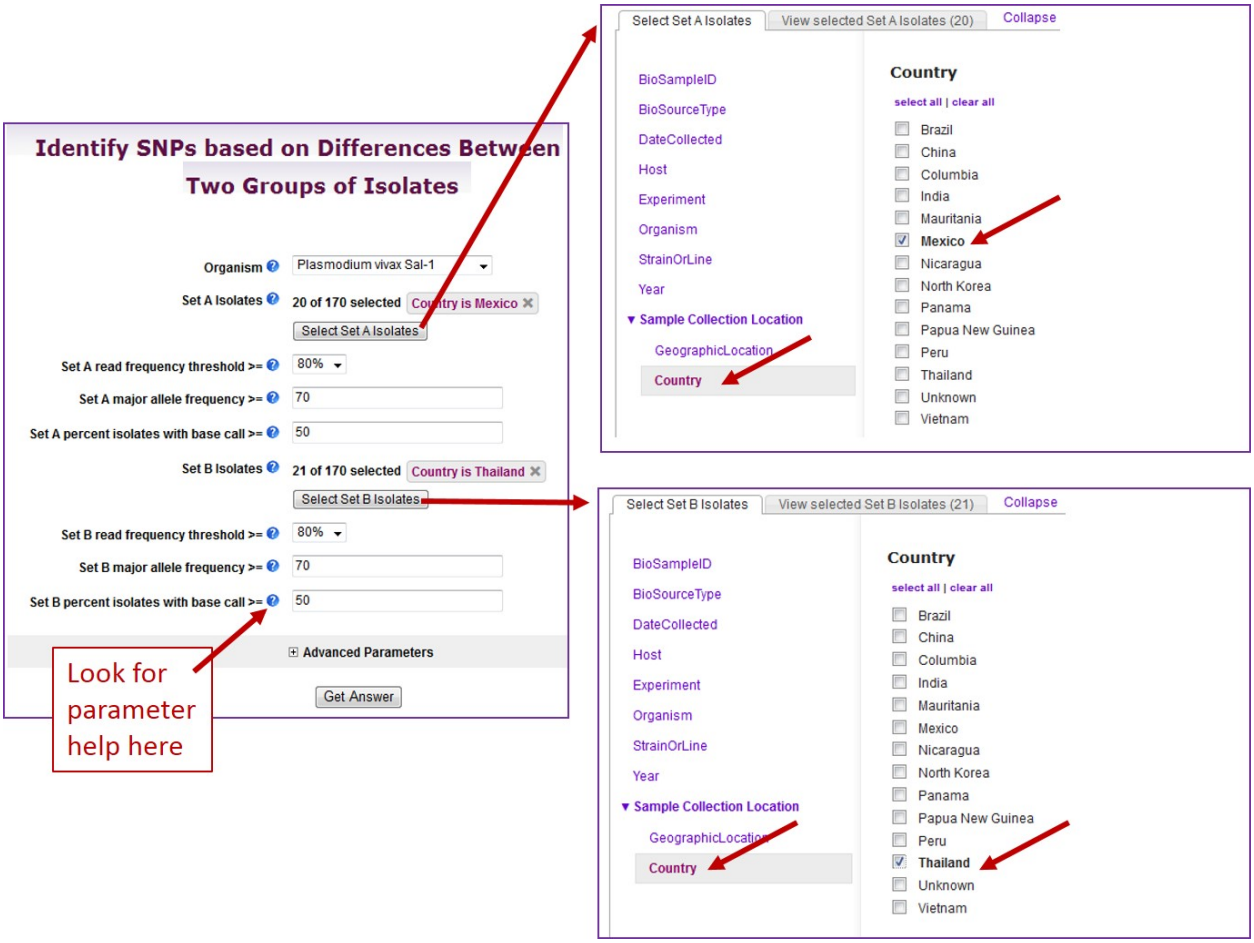
There are two main things you need to do:

- Define the two sets of isolates based on available metadata.
- Define the characteristics of the SNPs in each set of isolates using parameters Read Frequency Threshold, Major Allele Frequency, and Percent Isolates with a Base Call.

Here are suggested parameter values. Please hover over the help icons next to the parameter name for more information. There is also help information is the search description below the Get Answer button.

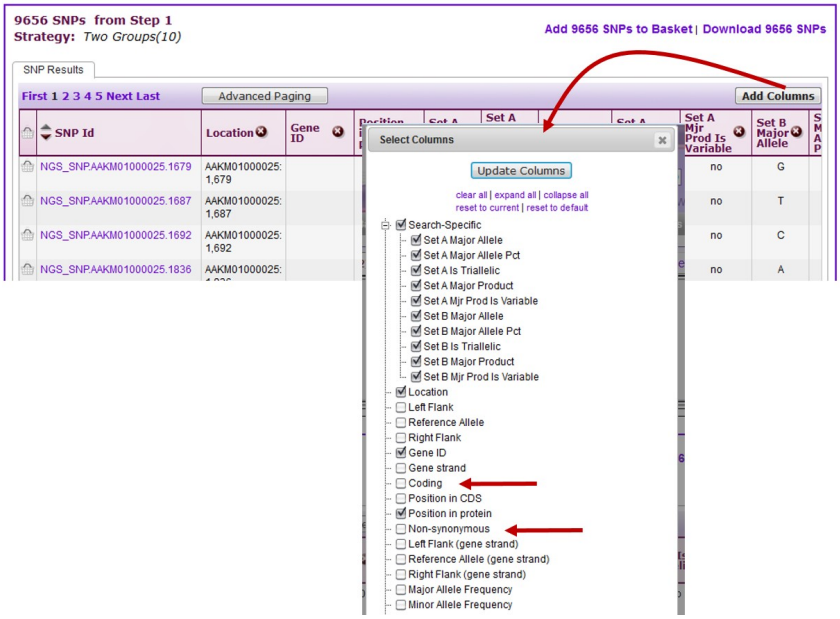
Important for this workshop, please use the following parameters.

Organism = *Plasmodium vivax* Sal1
Set A Isolates = Mexico
Set A read frequency threshold >= 80%
Set A major allele frequency >= 70%
Set A percent isolates with base call >= 50
Set B Isolates = Thailand
Set B read frequency threshold >= 80%
Set B major allele frequency >= 70%
Set B percent isolates with base call >= 50



- How many SNPs did you get?

- How many of these SNPs are in coding sequence? Add columns to your result table called Coding and Non-synonymous. There are many more columns of data to add. What other information about the SNPs can you learn?



- Download your SNPs to an Excel spreadsheet.