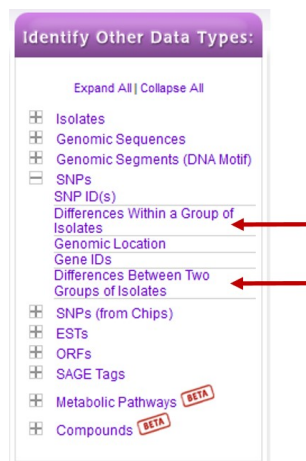


Metadata and Single Nucleotide Polymorphisms (SNPs)

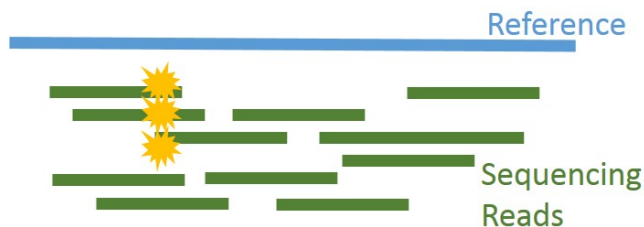
PlasmoDB contains several population biology data sets based on whole genome re-sequencing of isolates. Each isolate sequence is aligned to a reference genome and single nucleotide polymorphisms (SNPs) are recorded. Metadata associated with the isolates (*e.g.* country collected or drug sensitivity status) 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.

The screenshot shows a web-based interface for SNP identification. On the left, a sidebar titled 'Identify Other Data Types:' contains a tree view with categories like 'Isolates', 'Genomic Sequences', 'SNPs', and 'Differences Between Two Groups of Isolates'. The 'Differences Between Two Groups of Isolates' option is circled in red. A red arrow points from this circle to the main search area. The main area is titled 'Identify SNPs based on Differences Between Two Groups of Isolates' and is for the organism *Plasmodium falciparum* 3D7. It features two sections for 'Set A Isolates' and 'Set B Isolates', both with 204 isolates selected. Each section includes a list of quality filters (e.g., BioSampleID, BioSourceType, ClinicalPhenotype) and three parameter sliders: 'Set A read frequency threshold' (80%), 'Set A major allele frequency' (80), and 'Set A percent isolates with base call' (80). Similar parameters are present for Set B. An 'Advanced Parameters' section is visible at the bottom, and a 'Get Answer' button is located at the very bottom.

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 in 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 $\geq 80\%$
 Set A major allele frequency $\geq 70\%$
 Set A percent isolates with base call ≥ 50
 Set B Isolates = Thailand
 Set B read frequency threshold $\geq 80\%$
 Set B major allele frequency $\geq 70\%$
 Set B percent isolates with base call ≥ 50

Identify SNPs based on Differences Between Two Groups of Isolates

Organism

Set A Isolates Country is Mexico ✕

Set A read frequency threshold \geq

Set A major allele frequency \geq

Set A percent isolates with base call \geq

Set B Isolates Country is Thailand ✕

Set B read frequency threshold \geq

Set B major allele frequency \geq

Set B percent isolates with base call \geq

Select Set A Isolates View selected Set A Isolates (20) Collapse

BioSampleID BioSourceType DateCollected Host Experiment Organism StrainOrLine Year Sample Collection Location GeographicLocation Country	Country select all clear all <input type="checkbox"/> Brazil <input type="checkbox"/> China <input type="checkbox"/> Columbia <input type="checkbox"/> India <input type="checkbox"/> Mauritania <input checked="" type="checkbox"/> Mexico <input type="checkbox"/> Nicaragua <input type="checkbox"/> North Korea <input type="checkbox"/> Panama <input type="checkbox"/> Papua New Guinea <input type="checkbox"/> Peru <input type="checkbox"/> Thailand <input type="checkbox"/> Unknown <input type="checkbox"/> Vietnam
--	---

Select Set B Isolates View selected Set B Isolates (21) Collapse

BioSampleID BioSourceType DateCollected Host Experiment Organism StrainOrLine Year Sample Collection Location GeographicLocation Country	Country select all clear all <input type="checkbox"/> Brazil <input type="checkbox"/> China <input type="checkbox"/> Columbia <input type="checkbox"/> India <input type="checkbox"/> Mauritania <input type="checkbox"/> Mexico <input type="checkbox"/> Nicaragua <input type="checkbox"/> North Korea <input type="checkbox"/> Panama <input type="checkbox"/> Papua New Guinea <input type="checkbox"/> Peru <input checked="" type="checkbox"/> Thailand <input type="checkbox"/> Unknown <input type="checkbox"/> Vietnam
--	---

Look for parameter help here

- 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?

9656 SNPs from Step 1
Strategy: Two Groups(10)

SNP Results

SNP Id	Location	Gene ID	Position	Set A	Set A	Set A	Set A Mjr Prod Is Variable	Set B Major Allele	S M A P
NGS_SNPAAKM01000025.1679	AAKM01000025:1,679						no	G	
NGS_SNPAAKM01000025.1687	AAKM01000025:1,687						no	T	
NGS_SNPAAKM01000025.1692	AAKM01000025:1,692						no	C	
NGS_SNPAAKM01000025.1836	AAKM01000025:1,836						no	A	

Select Columns dialog box:

- Search-Specific
 - Set A Major Allele
 - Set A Major Allele Pct
 - Set A Is Triallelic
 - Set A Major Product
 - Set A Mjr Prod Is Variable
 - Set B Major Allele
 - Set B Major Allele Pct
 - Set B Is Triallelic
 - Set B Major Product
 - Set B Mjr Prod Is Variable
- Location
 - Left Flank
 - Reference Allele
 - Right Flank
- Gene ID
- Gene strand
- Coding
- Position in CDS
- Position in protein
 - Non-synonymous
 - Left Flank (gene strand)
 - Reference Allele (gene strand)
 - Right Flank (gene strand)
 - Major Allele Frequency
 - Minor Allele Frequency

- Download your SNPs to an Excel spreadsheet.

9656 SNPs from Step 1
Strategy: Two Groups(10)

SNP Results

SNP Id	Location	Gene ID	Position in protein	Set A Major Allele	Set A Major Allele Pct	Set A Is Triallelic	Set A Major Product	Set A Mjr Prod Is Variable	Set B Major Allele	S M A P
NGS_SNPAAKM01000025.1679	AAKM01000025:1,679		N/A	A	92.3	no	-	no	G	
NGS_SNPAAKM01000025.1687	AAKM01000025:1,687		N/A	G	92.8	no	-	no	T	
NGS_SNPAAKM01000025.1692	AAKM01000025:1,692		N/A						C	
NGS_SNPAAKM01000025.1836	AAKM01000025:1,836		N/A						A	

Please select a format from the dropdown:

- Select a format ---
- Select a format ---
- Tab delimited (Excel): choose from columns
- Text: choose from columns and/or tables
- XML: choose from columns and/or tables
- json: choose from columns and/or tables

PlasmidDB 24 14 Apr 15
©2015 The EuPathDB Project Team