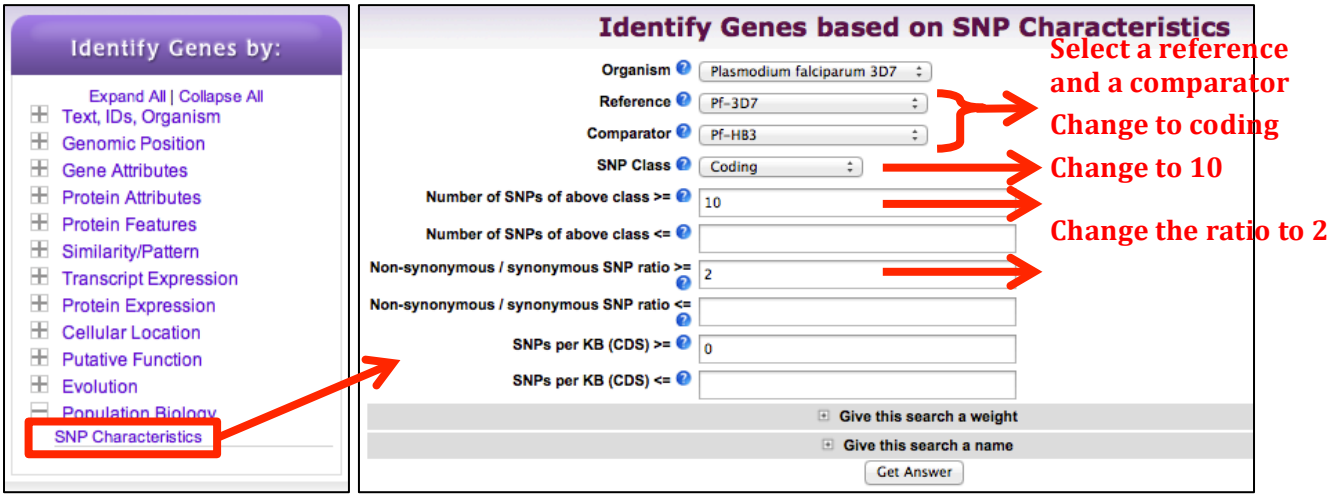
Population data, SNPs and alleles
Exercise 4

4.1 Diversifying or purifying selection

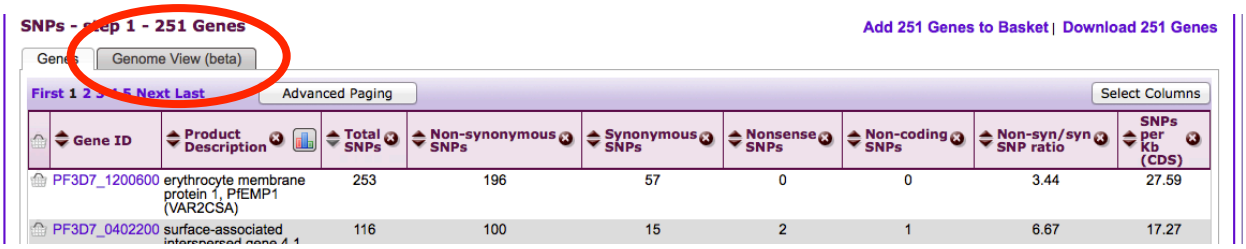
Note: For this exercise use http://www.plasmodb.org

a. Find the *P. falciparum* genes that are the most diverse among sequenced strains and thus appear to be under diversifying selection.

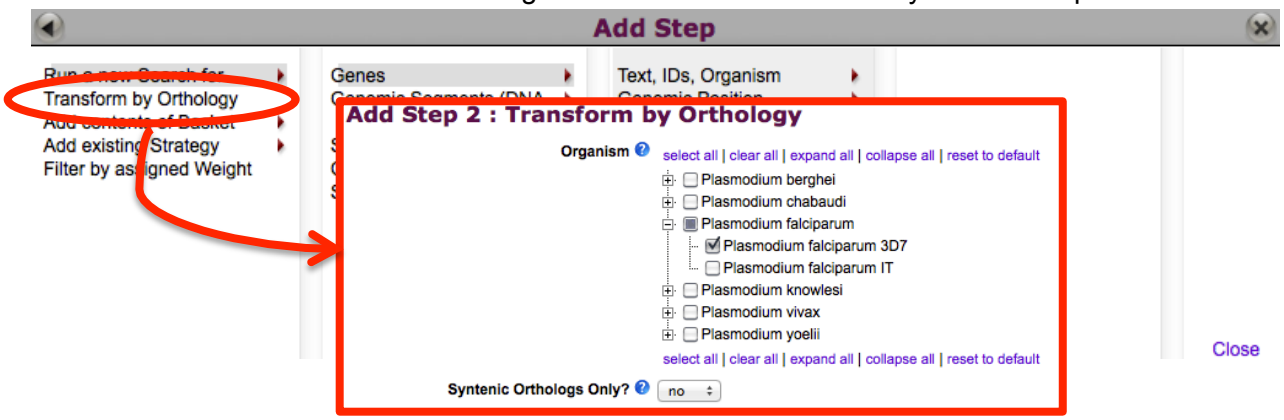
Hint: for this exercise you need to use the "Genes by SNP Characterisitics" search under "Identify Genes by:".

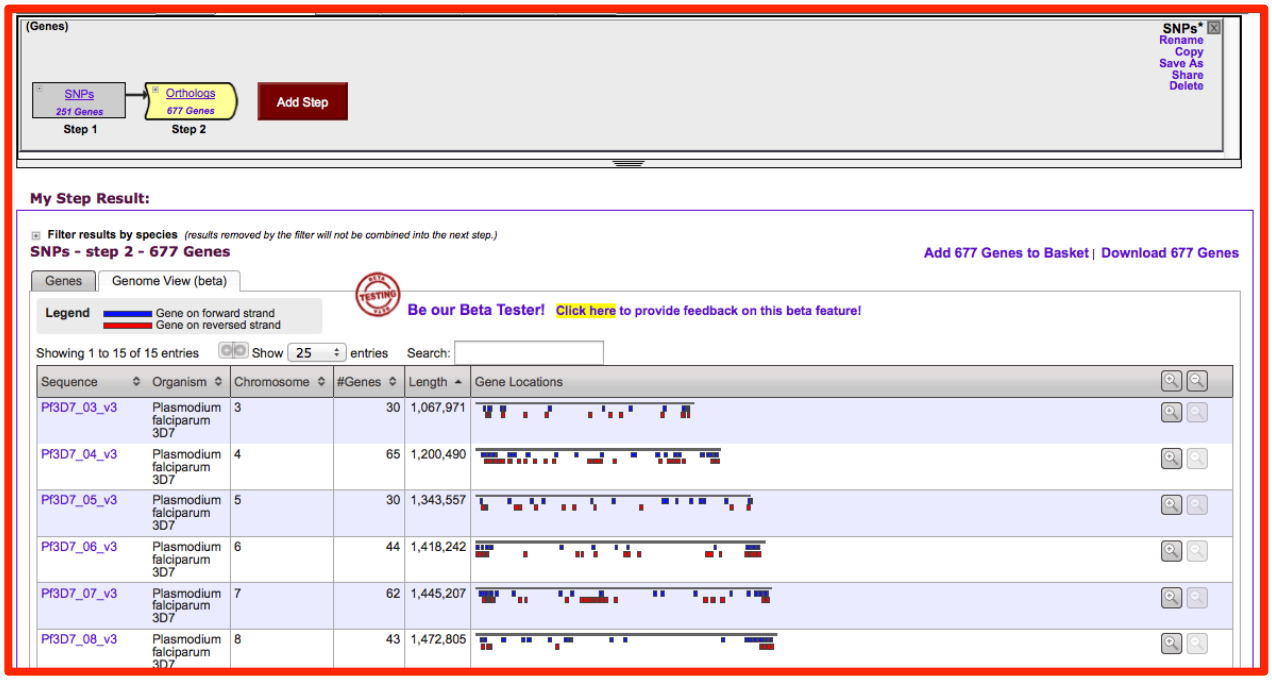


b. What strikes you about the known genes in this result set? What does the distribution of these genes on chromosomes look like? (hint: Click on the "Genome View" tab).

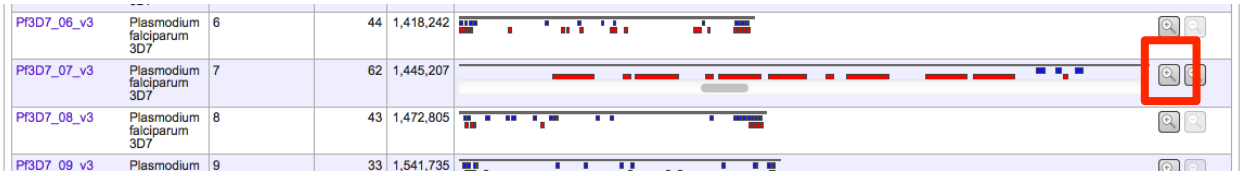


c. Find all paralogs of these genes in *P. falciparum*. (hint: add an orthology transform step). Now look at the distribution of the genes on chromosomes – do you notice a pattern?

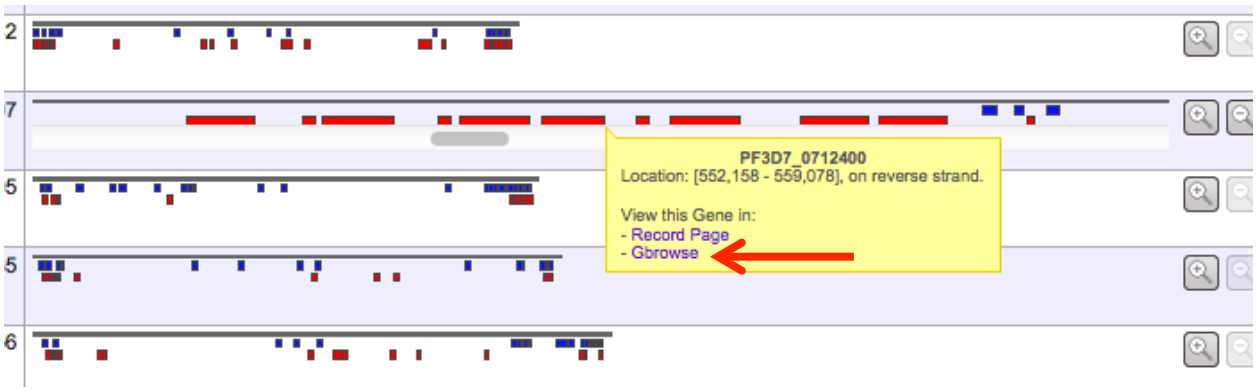




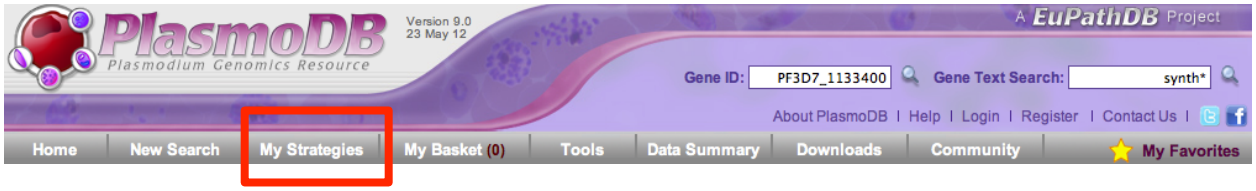
d. How many genes are in the middle cluster on chromosome 7? It might help to zoom in on this are (hint: use the magnifying glass icon on the left hand side).



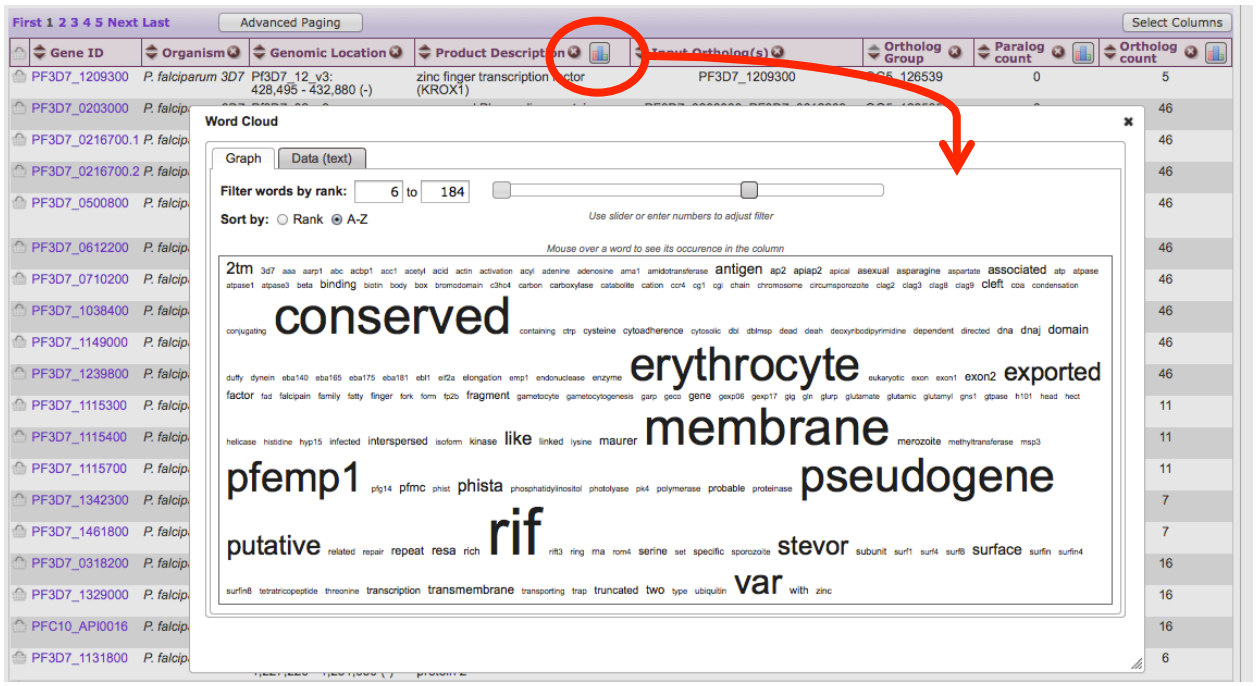
e. What are these genes? Mouse over one of them then click on Gbrowse in the popup. When you get to the gbrowse view you can mouse over the genes in the region to see what they are:



f. Go back to your search strategy page (hint: you can easily get back to your strategies by clicking on the “My Strategies” link in the grey tool bar.



g. How can you quickly get an idea about which genes are represented in your results? (hint: click on the column analysis icon next to the product description). You can customize the filter and sorting of the resulting word cloud. You can also click on the “Data” tab in the popup to get a table of the results.

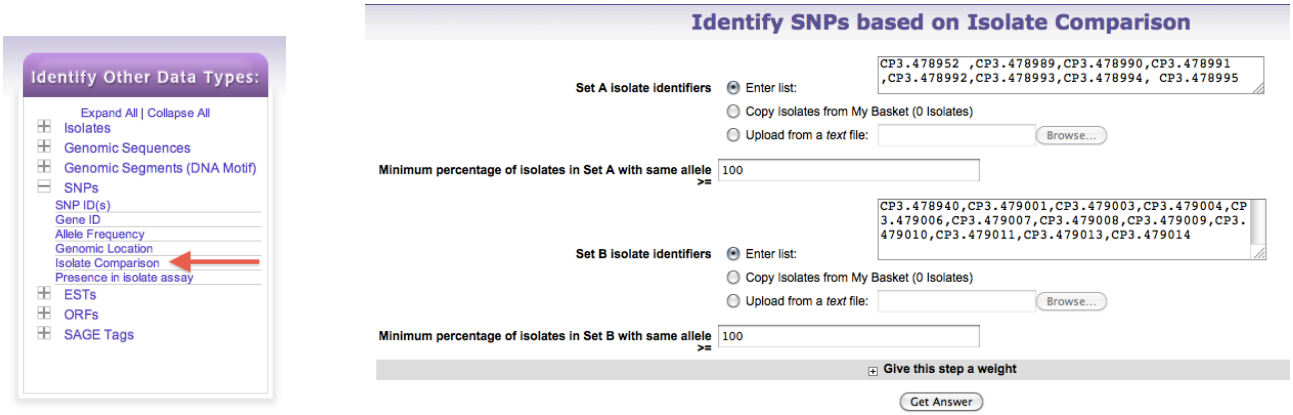


4.2 Isolate comparison

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

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

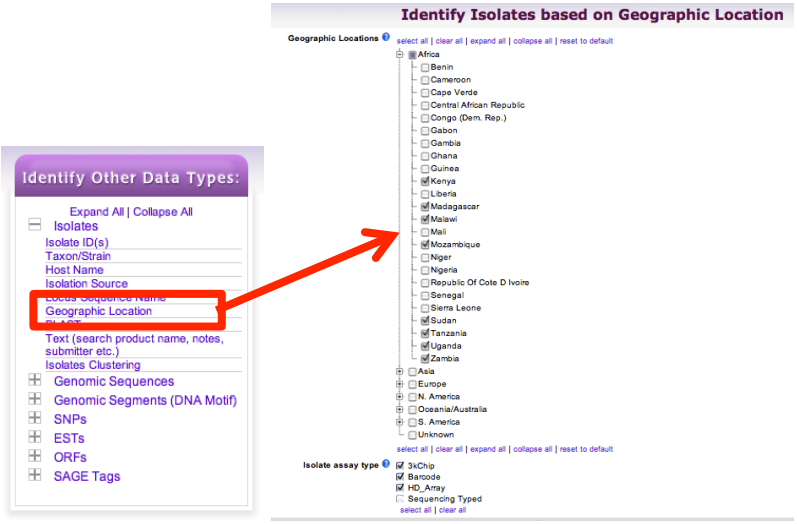
Hint: you can find this under "SNPs" in the "Identify other data types" section.



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.

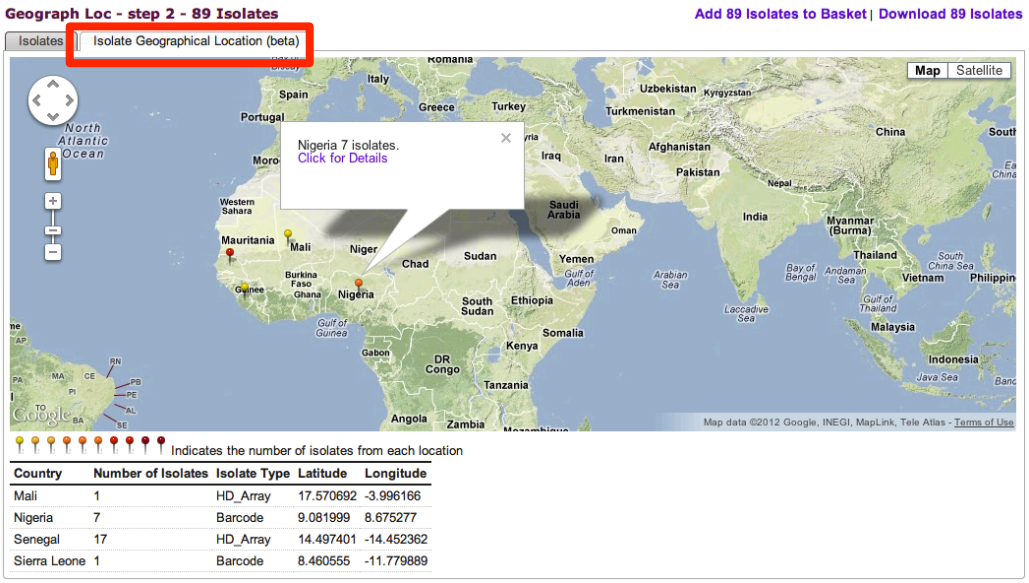
- For this exercise we are going to use the same SNPs by isolate comparison search as above. However, first we have to identify isolate IDs from West Africa and ones from East Africa. To do this use the Geographic location query under the isolates section (note that you will need to run this query twice, once for each set of countries):



Some East African countries: Kenya, Madagascar, Malawi, Mozambique, Tanzania, Sudan, Uganda, Zambia

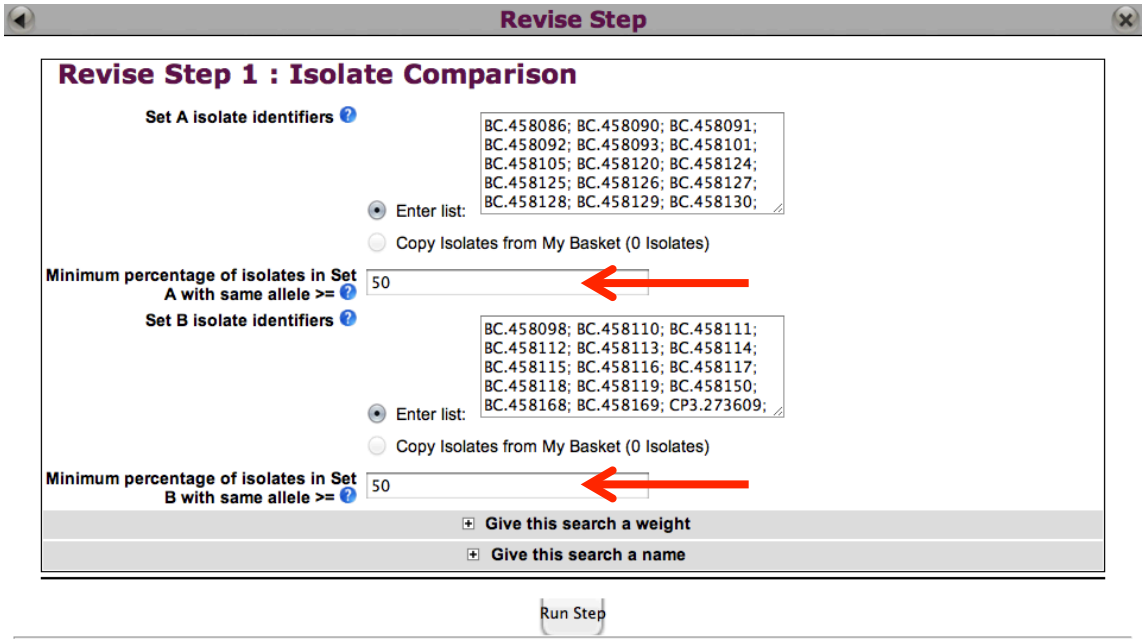
Some West African Countries: Cameroon, Gabon, Liberia, Mali, Nigeria, Senegal, Siera Leone

- 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:



Country	Number of Isolates	Isolate Type	Latitude	Longitude
Mali	1	HD_Array	17.570692	-3.996166
Nigeria	7	Barcode	9.081999	8.675277
Senegal	17	HD_Array	14.497401	-14.452362
Sierra Leone	1	Barcode	8.460555	-11.779889

- 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:



- What do your results look like now?
  - Which SNP differentiates more isolates (hint: look at the numbers in the columns for SetA and SetB)?
  - 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).

4.3 Analyzing SNPs on a defined list of genes.

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

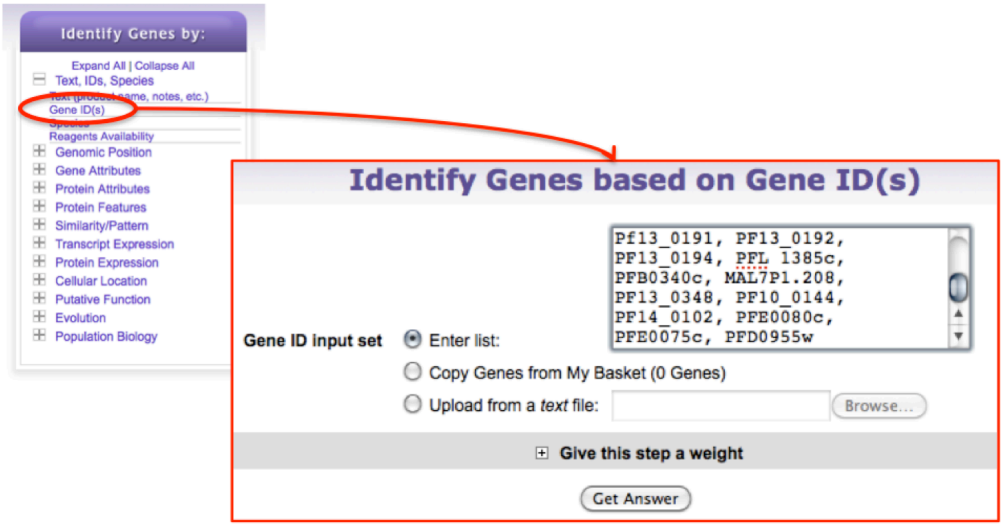
You just read the recent paper by Tetteh *et.al.* (http://www.ncbi.nlm.nih.gov/pubmed/19440377) where they perform an analysis of SNPs on a set of *P. falciparum* genes. Their conclusion is that these genes are under "balancing" selection – under diversifying selection due to their exposure to the host's 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

- **Put the above list into a step.**

  Hint: use the Identify Genes based on Gene ID(s) search option.

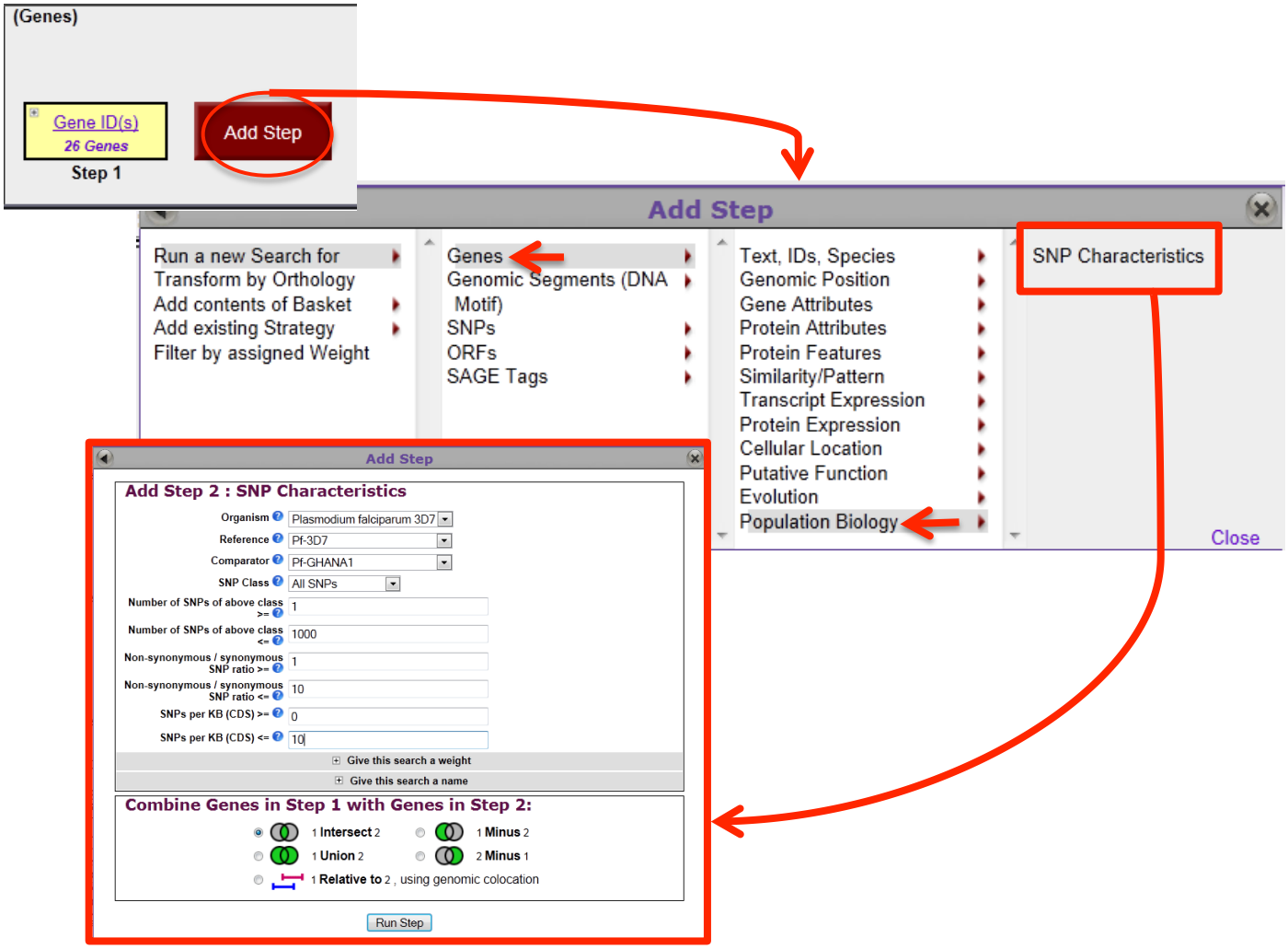


Add a step to your strategy to identify how many of these genes are under diversifying selection.

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 chose?
- Would you expect genes under balancing selection to have a high or low non-synonymous/synonymous SNP ration?
- How many genes were returned by your search? Of these, now many intersect with the set of genes from the paper?

Click on the result for your ID search in the first step (25 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
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 Minus 2
1 Union 2
2 Minus 1
1 Relative to 2 , using genomic colocation
Run Step