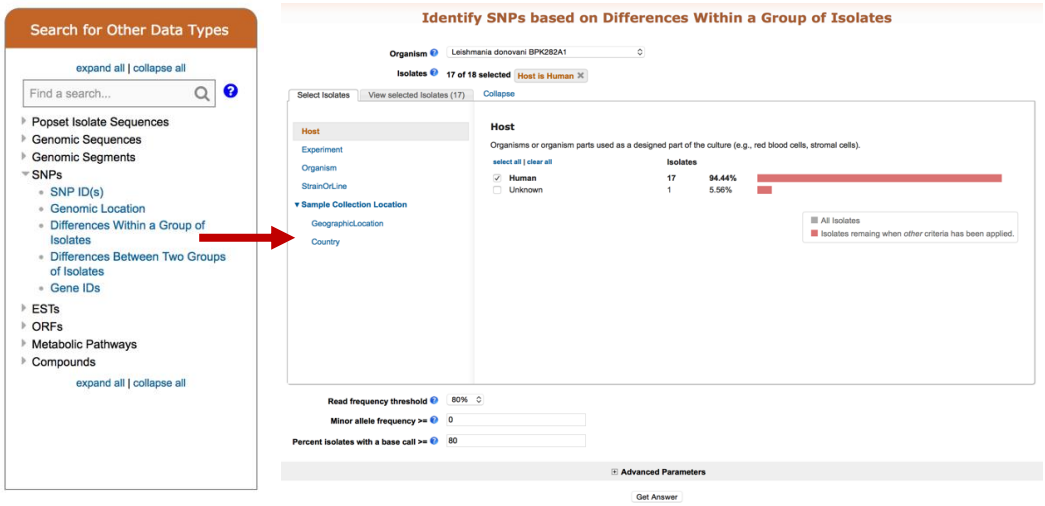
Genetic Exercises

SNPs and Population Genetics

Single Nucleotide Polymorphisms (SNPs) in EuPathDB can be used to characterize similarities and differences within a group of isolates or that distinguish between two groups of isolates. They can also be utilized to identify genes that may be under evolutionary pressure, either to stay the same (purifying selection) or to change (diversifying or balancing selection). Isolates are assayed for SNPs in EuPathDB by two basic methods; re-sequencing and then alignment of sequence reads to a reference genome or 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. 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.

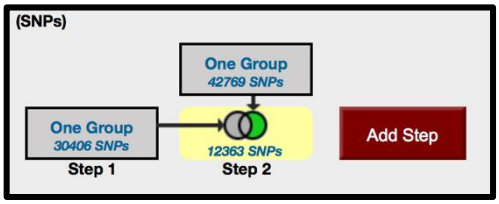


   Hint: you can find this under "SNPs" in the "Identify Other Data Types" section.

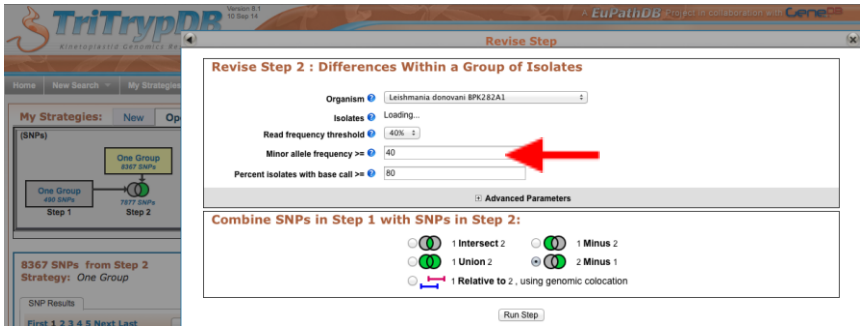
   b. **What does this search do?** Choose *Leishmania donovani* for the organism and select isolates from the human host. Use default parameters for the rest of the parameters. Run the query and look at your results.

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

- 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 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?

NOTE: Exercises 3 and 4 are similar. The first is in ToxoDB and explores the hypothesis that we can identify SNPs/genes involved in T. gondii host preference. The second is in PlasmoDB and identifies SNPs (regions of the genome) that could be involved in P. falciparum Artemesinin resistance. Note that the PlasmoDB exercise uses a SNP-chip assay rather than re-sequencing. Scan through both and choose the one that interests you the most to do first. Then if there is time at the end you can come back and do the other one.

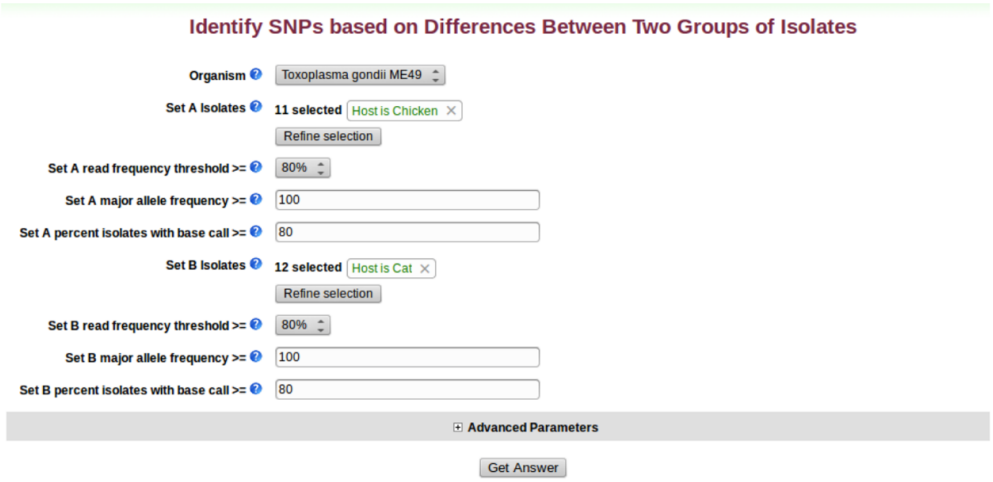
3. Find SNPs that distinguish *Toxoplasma gondii* strains isolated from chickens as compared to those isolated from cats.

   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 (*Gallus* gallus) box to select the 11 chicken isolates.

b. Click select set B isolates and select hosts from the left column. Check the cat (*Felis* catus) box to select the 12 cat isolates.

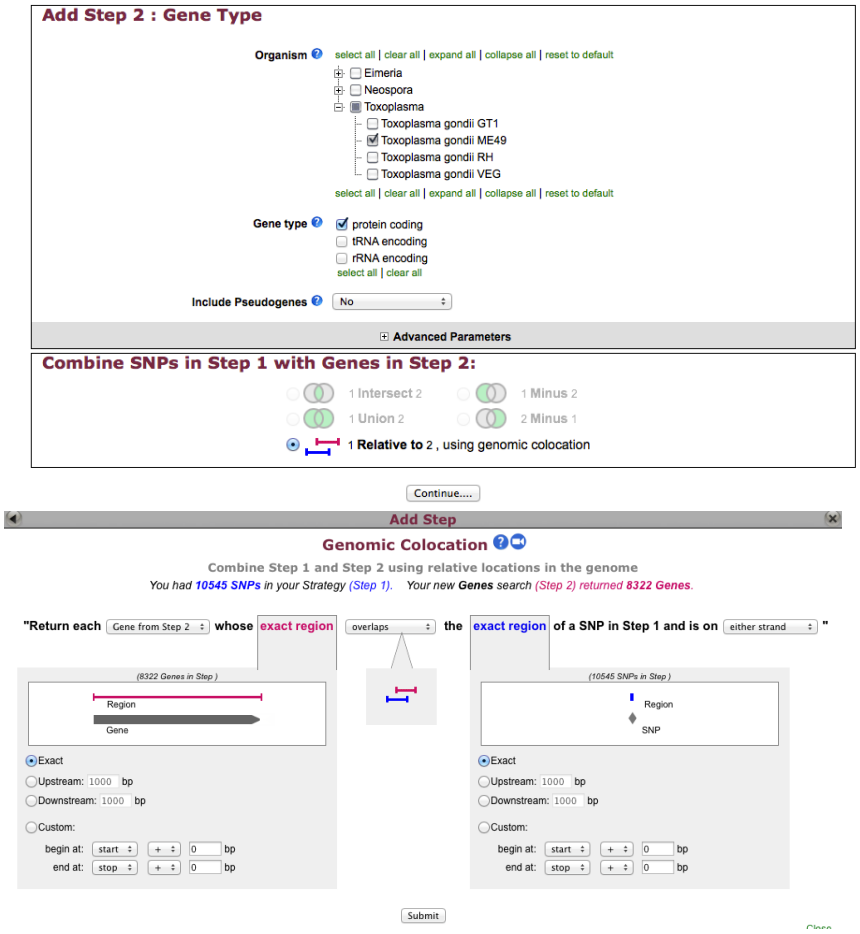


c. Let's run a very stringent search and change the "major allele frequency" parameters for both sets to 90. (*What does that mean?)*. We'll leave the other parameters at their default values, which are in themselves pretty stringent … but feel free to change them to see how this impacts your results.

- 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"

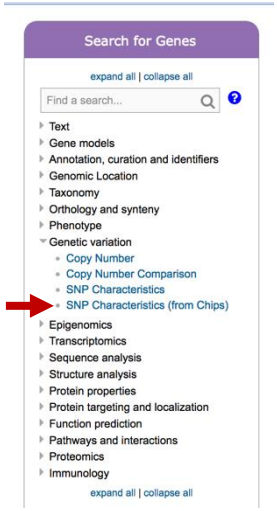


- How many genes are returned?
- What is the gene that contains the most SNPs on your list? *Hint: sort the list high to low by match count.*
- Does this gene have orthologs in other species from ToxoDB? *Hint: go to the gene page and look at the genomic context and orthologs/paralogs in ToxoDB table.*
- Does it have orthology in any other species? *Hint: click on the link under the orthologs table and look at in OrthoMCL.*

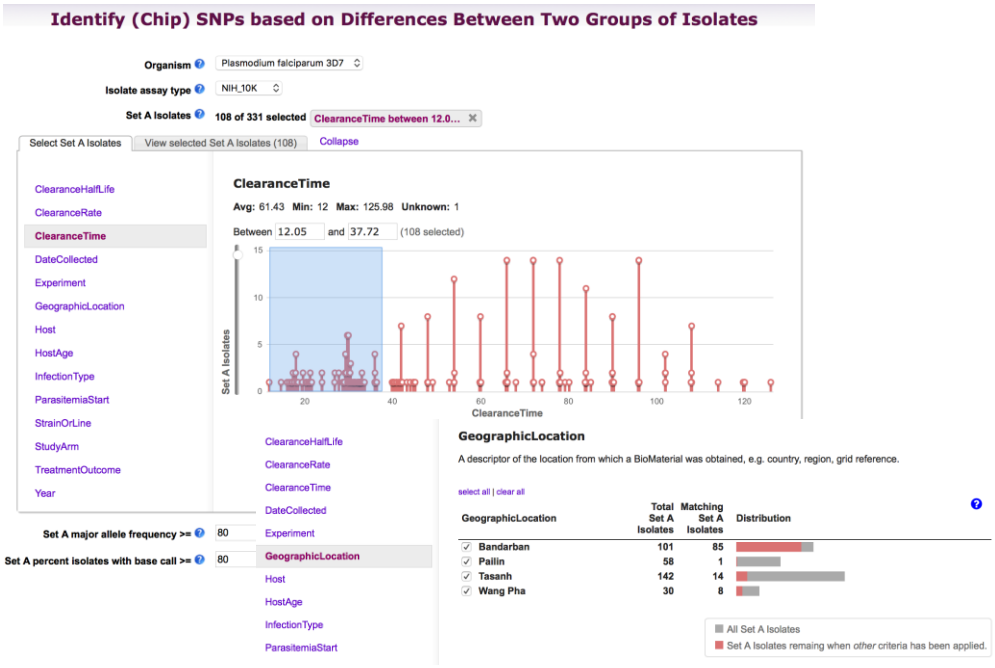
- What does this say about this gene? How can you follow up on what what role this gene may be playing for the organism? *Hint: you are a biologist and will need to look at the data on the gene record page and interpret it based on your experience and intuition.*
- Do these genes appear to be randomly 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 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).*



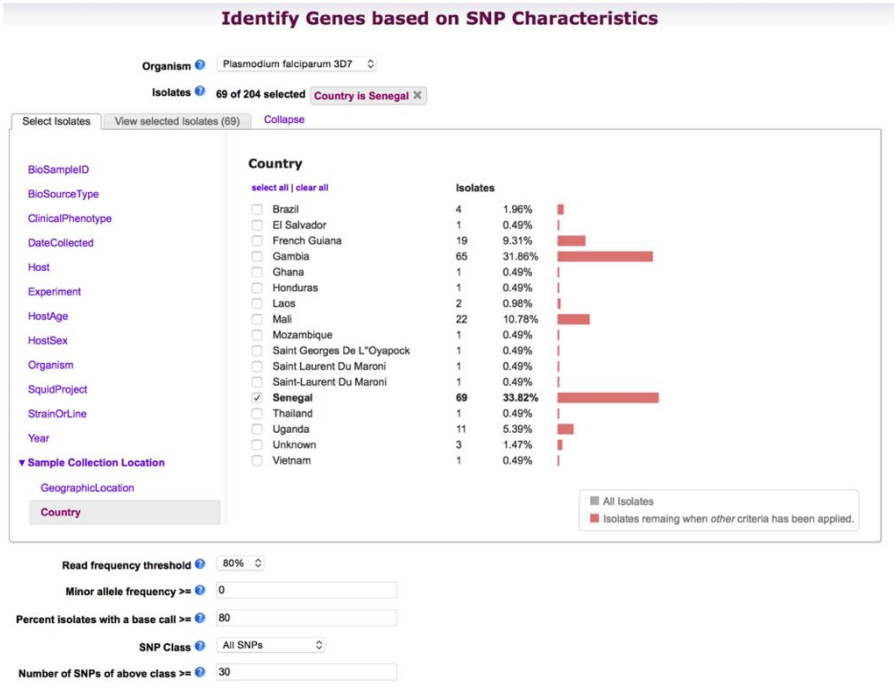
c. We'll keep the defaults of 80 for both Major Allele Frequency and Percent Isolates with Call for this exercise.
d. Now select Clearance times of 82 – end for Set B Isolates. Are these isolates geographically biased?
e. Keep defaults for Major Allele and Percent with call and run the search. How many 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.

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.*



- 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.

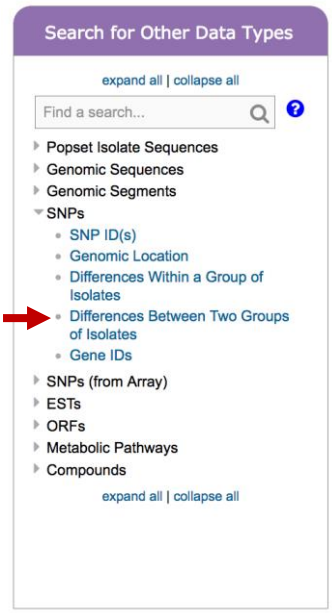
5. Comparing your results with a published list: You just read the recent paper by Tetteh *et.al.* (http://www.ncbi.nlm.nih.gov/pubmed/19440377) where they perform an analysis of SNPs on a set of *P. falciparum* genes. Their conclusion is that these genes are under "balancing" selection – under diversifying 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.*

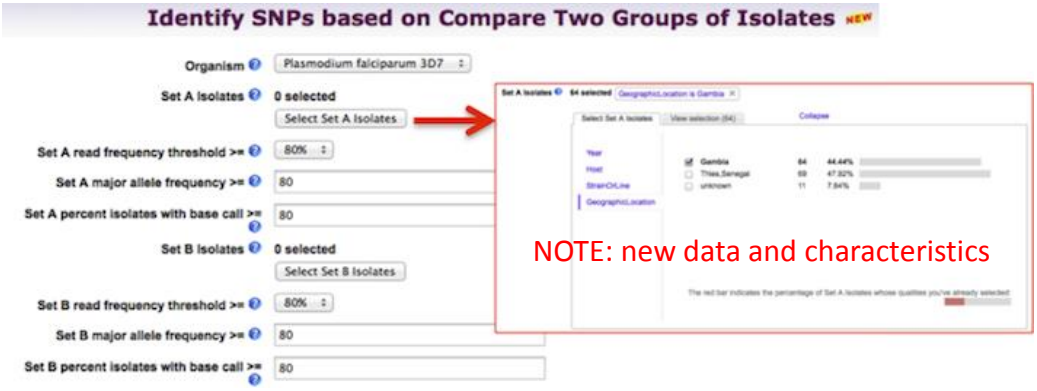
2. OPTIONAL EXERCISE: Find SNPs that differentiate between groups of isolates. For example, those from Gambia vs. Senegal.
For this exercise use http://plasmodb.org



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? 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.

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?