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).

SNPs - step 1 - S Genes Genom	251 Genes ne View (beta) xt Last Advan	ced Paging				Add 251 Genes	to Basket Downlo	ad 251 Genes
🕼 🗢 Gene ID	Product Description	Total SNPs	♣ Non-synonymous SNPs	♣ Synonymous SNPs	♦ Nonsense SNPs	Non-coding SNPs	♣ Non-syn/syn ③ SNP ratio	SNPs per Kb (CDS)
PF3D7_1200600	erythrocyte membrane protein 1, PfEMP1 (VAR2CSA)	253	196	57	0	0	3.44	27.59
PF3D7_0402200	surface-associated	116	100	15	2	1	6.67	17.27

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?

	Add Step	×
Bun e new Search for Transform by Orthology Add contents of Decker Add existing Strategy Filter by as igned Weight	Genes Text, IDs, Organism Add Step 2 : Transform by Orthology Organism Select all clear all expand all collapse all reset to default Plasmodium falciparum Plasmodium falciparum 0D7 Plasmodium falciparum 1T Plasmodium falciparum 1T Plasmodium falciparum 1T Plasmodium tralciparum 1T Plasmodium tralci	Close

(Genes)	Crthologs					SNPs* X Rename Copy Save As Share Delete
251 Genes Step 1	677 Genes Step 2	Add Step				
My Step Resul	t:					
Filter results by s SNPs - step 2	species (results re - 677 Genes	ernoved by the filter will	not be combine	d into the next	step.)	Add 677 Genes to Basket Download 677 Genes
Genes Gene	ome View (beta)		TESTING			
Legend	Gene on forwa	ard strand rsed strand	TH	Be our B	leta Tester! Click here to provide feedback on this beta feature!	
Showing 1 to 15 of	f 15 entries	Show 25	+ entries	Search:		
Sequence 0	Organism \$	Chromosome \$	#Genes \$	Length 🔺	Gene Locations	
Pf3D7_03_v3	Plasmodium falciparum 3D7	3	30	1,067,971	**	
Pf3D7_04_v3	Plasmodium falciparum 3D7	4	65	1,200,490		
Pf3D7_05_v3	Plasmodium falciparum 3D7	5	30	1,343,557	A CONTRACT OF A CONTRACT OF	
Pf3D7_06_v3	Plasmodium falciparum 3D7	6	44	1,418,242	<u></u>	
Pf3D7_07_v3	Plasmodium falciparum 3D7	7	62	1,445,207	The second s	
Pf3D7_08_v3	Plasmodium falciparum 3D7	8	43	1,472,805	<u> </u>	

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).

Pf3D7_06_v3	Plasmodium falciparum 3D7	6	44	1,418,242	
Pf3D7_07_v3	Plasmodium falciparum 3D7	7	62	1,445,207	
Pf3D7_08_v3	Plasmodium falciparum 3D7	8	43	1,472,805	
Pf3D7 09 v3	Plasmodium	9	33	1,541,735	

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:

2			
7			••
5		Location: [552,158 - 559,078], on reverse strand. View this Gene in: - Record Page	0
5	w	- Gbrowse	
6	No. 1996 and 1997	F	

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.

	noDR	Version 9.0 23 May 12	the states		A EuP	athDB Project
Plasmodium Gen	omics Resource	33	Gene ID	PF3D7_1133400	Gene Text Search:	synth*
100 201 11	82.8 ·····			About PlasmoDB 1	Help I Login I Register	I Contact Us I 🕒 🚮
Home New Search	My Strategies	My Basket (0)	Tools Data Summa	ry Downloads	Community	🔶 My Favorites

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.

First 1 2 3 4 5 Next	Last	Advanced Paging	Select Columns
🙆 ≑ Gene ID	Organ	ism Q 🗢 Genomic Location Q 🗢 Product Descript on Q 👔 🗧 Teach Ortholog(s) Q 🔷 Ortholog Q 🖨 Paralog Q 👔 🗢	Ortholog 🥝 📊
PF3D7_1209300	P. falcipar.	m 3D7 Pf3D7_12_v3: zinc finger transcription_ctor PF3D7_1209300 05_126539 0 428,495 - 432,880 (-) (KROX1)	5
PF3D7_0203000	P. falcip	Word Cloud	4 6
PF3D7_0216700.1	P. falcip	Craph Data (layt)	46
PF3D7_0216700.2	P. falcip		46
PF3D7_0500800	P. falcip	Filter words by rank: 6 to 184 Sort by: O Rank @ A-Z Use alider or enter numbers to adjust filter	46
PF3D7_0612200	P. falcip	Mouse over a word to see its occurence in the column	46
@ PF3D7_0710200	P. falcip	2tm 367 as any 1 at adapt act adopt act adopt at advantant advantant advantant and advanta and advantanterse and antigen ap2 apiag2 apical assexual aspartagine aparta associated at apase atpaset apased beta binding both body box transformation and the carbon carbonylase catabolitie cation cere of og thain dromosome dromsportable cag2 cag3 cag3 cag3 cag3 cag3 cag3 cag3 cag3	46
PF3D7_1038400	P. falcip		46
PF3D7_1149000	P. falcip	or throat	46
PF3D7_1239800	P. falcip	dutty dynein eta 140 eta 165 eta 175 eta 151 eta 1 etz elongation empt endonuclease enzyme el yli il OCyle eukayotic euon exont exon2 exported	46
PF3D7_1115300	P. falcip	factor fad falcipain family fatty finger fork form foto fragment gametocyto gametocytogenesis garp geo gene geop06 geop07 pig gin plurp glutamate glutamic glutamyl grist gloase h101 head hect	11
PF3D7_1115400	P. falcip	helcase histore hyp15 infected interspersed isotom kinase like linked lysine maturer memory memory merozoite methyltransferase msp3	11
PF3D7_1115700	P. falcip	pfemp1 pseudogene	11
PF3D7_1342300	P. falcip		7
PF3D7_1461800	P. falcip		7
PF3D7_0318200	P. falcip		16
PF3D7_1329000	P. falcip	surfinë tetratricoppotide threenine transcription transmembrane transporting trap truncated two type ubiquitin Var with zinc	16
PFC10_API0016	P. falcip		16
PF3D7_1131800	P. falcip	יואטיזיע () איז און און איז איז און און איז איז אוואטער () איז	6

4.2 Isolate comparison Note: For this exercise use <u>http://www.plasmodb.org</u>

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.

	Id	entify SNPs b	ased on Isolat	e Comparison
Identify Other Data Types:	Set A isolate identifiers	Enter list: Convulsations from Mv	CP3.478952 ,CP3.4789 ,CP3.478992,CP3.4789	989,CP3.478990,CP3.478991 993,CP3.478994, CP3.478995
Isolates Genomic Sequences		 Upload from a text file: 	Dasket (0 Isolates)	Browse
Genomic Segments (DNA Motif) SNPs	Minimum percentage of isolates in Set A with same allele >=	100		
SNP ID(s) Gene ID Allele Frequency			CP3.478940,CP3.47900 3.479006,CP3.479007, 479010,CP3.479011,CP	D1,CP3.479003,CP3.479004,CP ,CP3.479008,CP3.479009,CP3. P3.479013,CP3.479014
Genomic Location Isolate Comparison	Set B isolate identifiers	Enter list:		1
Presence in isolate assay		Copy Isolates from My	Basket (0 Isolates)	
B ORFs		Upload from a text file:		Browse
SAGE Tags	Minimum percentage of isolates in Set B with same allele >=	100		
		[+ Give this step a weight	
			Get Answer	

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:



- 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: <u>http://goo.gl/rhRdO</u>
- 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:

	Revise Step	
Revise Step 1 : Isol	ate Comparison	
Set A isolate identifiers	 BC.458086; BC.458090; BC.458091; BC.458092; BC.458093; BC.458101; BC.458105; BC.458120; BC.458124; BC.458125; BC.458126; BC.458127; BC.458128; BC.458129; BC.458130; Copy Isolates from My Basket (0 Isolates) 	
Minimum percentage of isolates in S A with same allele >= (50	
Set B isolate identifiers	 BC.458098; BC.458110; BC.458111; BC.458112; BC.458113; BC.458114; BC.458115; BC.458116; BC.458117; BC.458115; BC.458116; BC.458150; BC.458168; BC.458169; CP3.273609; Copy Isolates from My Basket (0 Isolates) 	
Minimum percentage of isolates in S B with same allele >=	50	
	+ Give this search a weight	
	Give this search a name	

- 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 <u>http://www.plasmodb.org</u>

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.

Identify Genes by: Expand All Collapse All				
 Text, IDs, Species Text, μrouget same, notes, etc.) Gene ID(s) 	_			
Reagents Availability Genomic Position			N	
Gene Attributes Protein Attributes Brotein Ecotures	Ide	entify Genes b	based on Gene ID(s)	
SimilarityPattern SimilarityPattern Transcript Expression Protein Expression Cellular Location Putative Function Evolution Evolution Population Biology	Gene ID input set	Enter list: Copy Genes from My B Unload from a text file:	Pf13_0191, PF13_0192, PF13_0194, PFL_1385c, PFB0340c, MAL7P1.208, PF13_0348, PF10_0144, PF14_0102, PFE0080c, PFE0075c, PFD0955w asket (0 Genes)	
		Give t	his step a weight	
		G	et Answer	

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?

