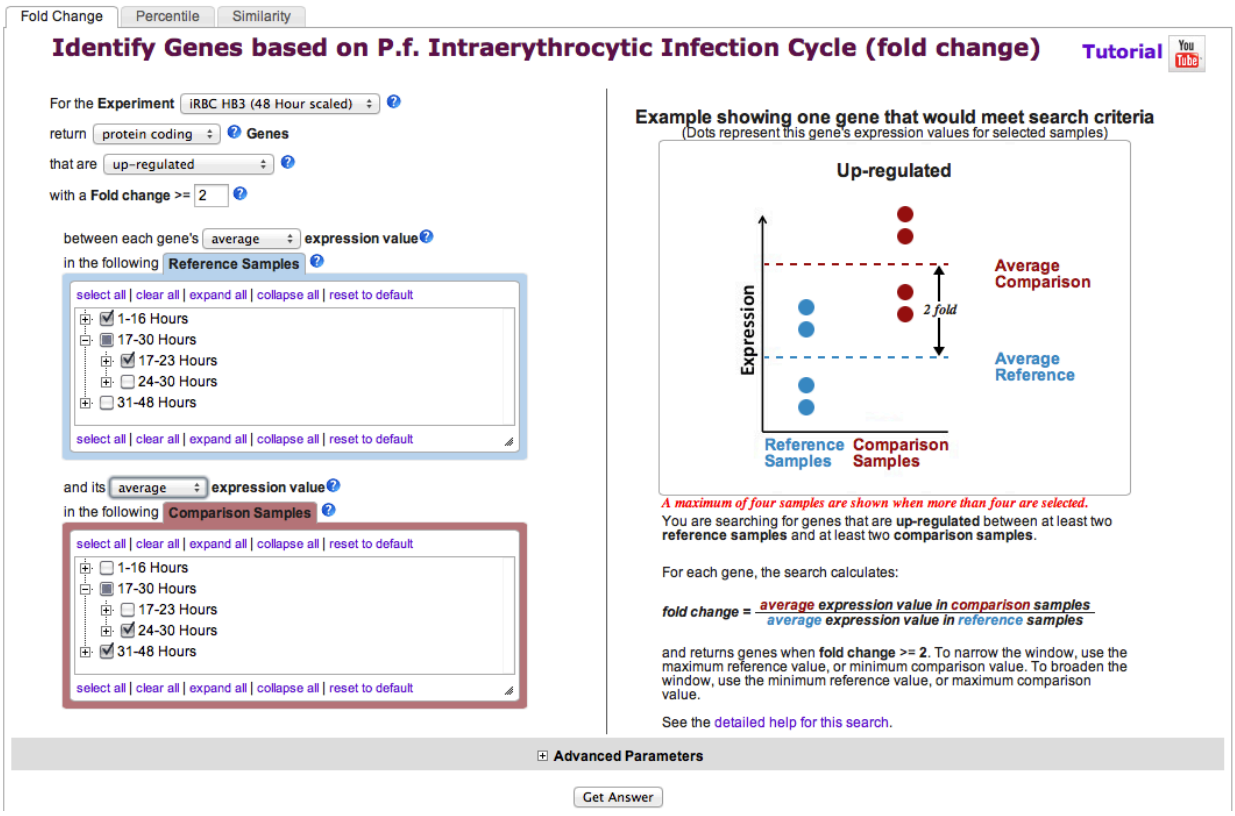
Complex strategies with Genomic Colocation
Exercise 10

1. **Divergent genes with similar expression profiles.**
   Note: for this exercise use http://plasmodb.org.

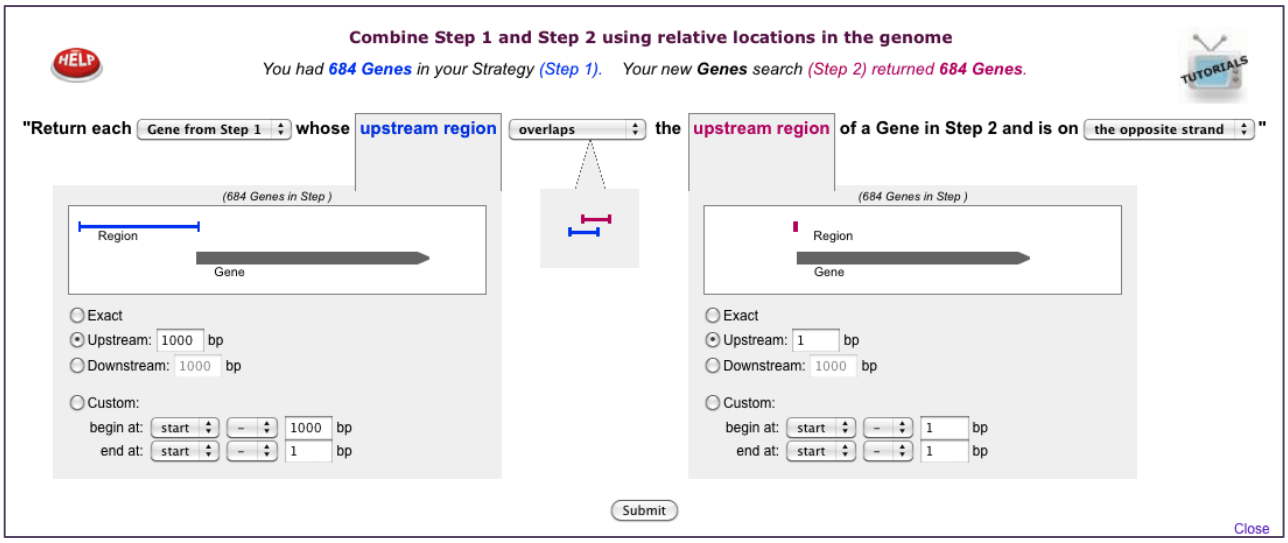
   Identify genes that meet these four criteria:
   1. are located within 1000 bp of each other.
   2. are divergently transcribed.
   3. are upregulated by at least 2-fold between between 24-48 hrs compared to 1-23 hrs.

- Hint: first use the "Genes bases on Microarray Evidence" -> "***Erythrocytic expression time series (3D7,DD2, & HB3) (Bozdech et al. and Linas et al.)***" and run a fold change search.



- Add a step that is the same as the first step and select the genomic colocation (1 relative to 2) operation.
- Set up the form to identify those genes that are transcribed on the opposite strand that have their starts located within 1000 bp of another genes start.

- Turn on the “Pf-iRBC 48hr - Graph” column to assess how well the pairs of genes compare in terms of expression. The pairs of genes are located one above the other in the result table if sorted by location.
- Note that you could do similar types of experiments to look at potential co-regulation / shared enhancers / divergent promoters with other sorts of data such as:
    - Genes by ChiP-chip peaks in ToxoDB.
    - DNA motifs for transcription factor binding sites.
    - Of course other expression queries.
    - Etc …
- The screenshot below shows one way (there are MANY) to configure the genome colocation form to identify genes that are divergently transcribed located with their start within 1000 bp of each other.



2. Finding possible oocyst expressed genes based on DNA motifs.
Note: for this exercise use http://toxodb.org

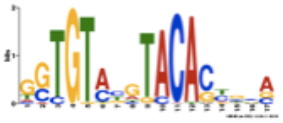
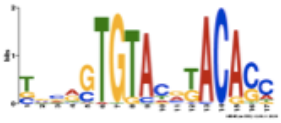
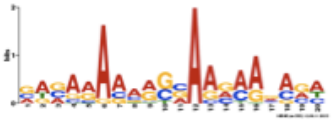
In an earlier exercise you defined a number of *T. gondii* genes that are preferentially expressed in the oocyst stages. How can you use this information to expand the number of possible oocyst regulated genes? One possibility is to try and define common elements in promoter or 5’UTR regions (ie. 5’ to the start of the genes). For this you will have to be able to retrieve 5’ sequence from all of the genes in the oocyst list. How would you do this? (*Hint*: click on download genes then select FASTA format from the drop down menu). The amount of upstream sequence you retrieve is up to you.

After you have your sequences you will need to run them through a DNA pattern finder like MEME (http://meme.sdsc.edu/meme/intro.html). Results from a submission to MEME could take up to several hours so for your convenience 300

nucleotides upstream of all the oocyst results were analyzed using MEME – results can be visualized here:

Can you take one of the generated motifs and find additional genes in *T. gondii* that contain this motif in their upstream regions? What do your results look like? Did you get too many or too few results? How would you modify the motif to change your results?

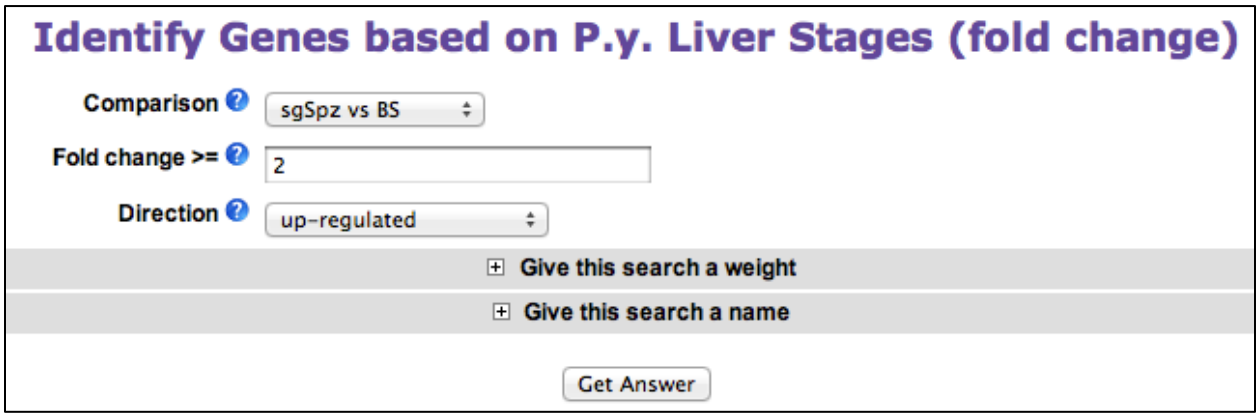
Motif Overview

Motif 1	• 1.1e-072 • 72 sites		
Motif 2	• 2.9e-012 • 103 sites		
Motif 3	• 2.5e+003 • 37 sites		

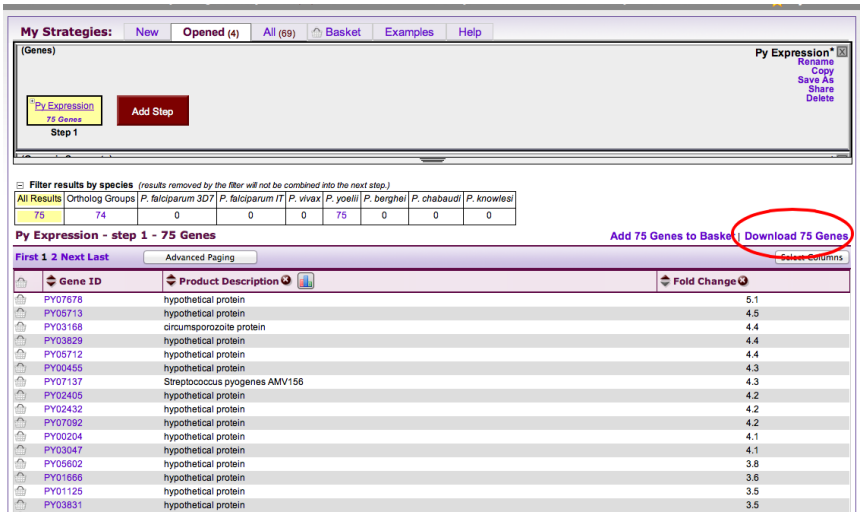
3. Identifying conserved DNA elements upstream of genes

The goal of this exercise is to identify a DNA element in the upstream region of similarly regulated genes.

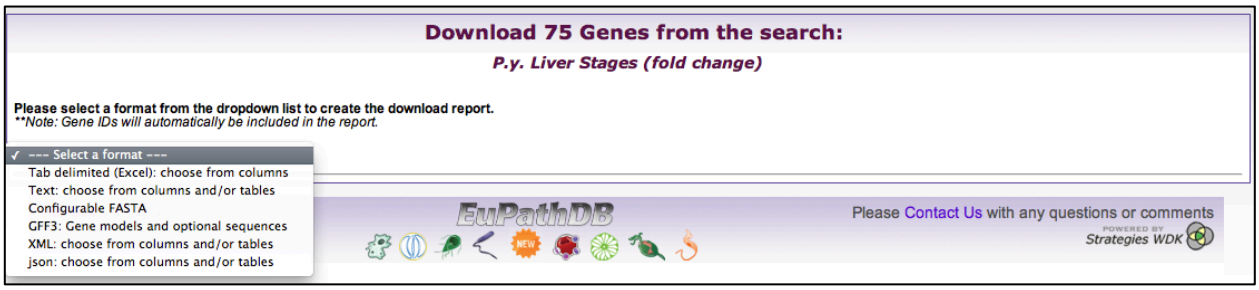
a. Identify genes that are up-regulated in malaria sporozoites compared to blood stage parasites. Examine the list of searchable experiments on the PlasmoDB microarray search page: Identify Genes based on Microarray Evidence. Can you identify an experiment that would give you this answer? (*Hint*: look at *Plasmodium* species other than *P. falciparum*, ie. *P. yoelii* [Liver, mosquito and blood stage expression profiles (Tarun et al.)]



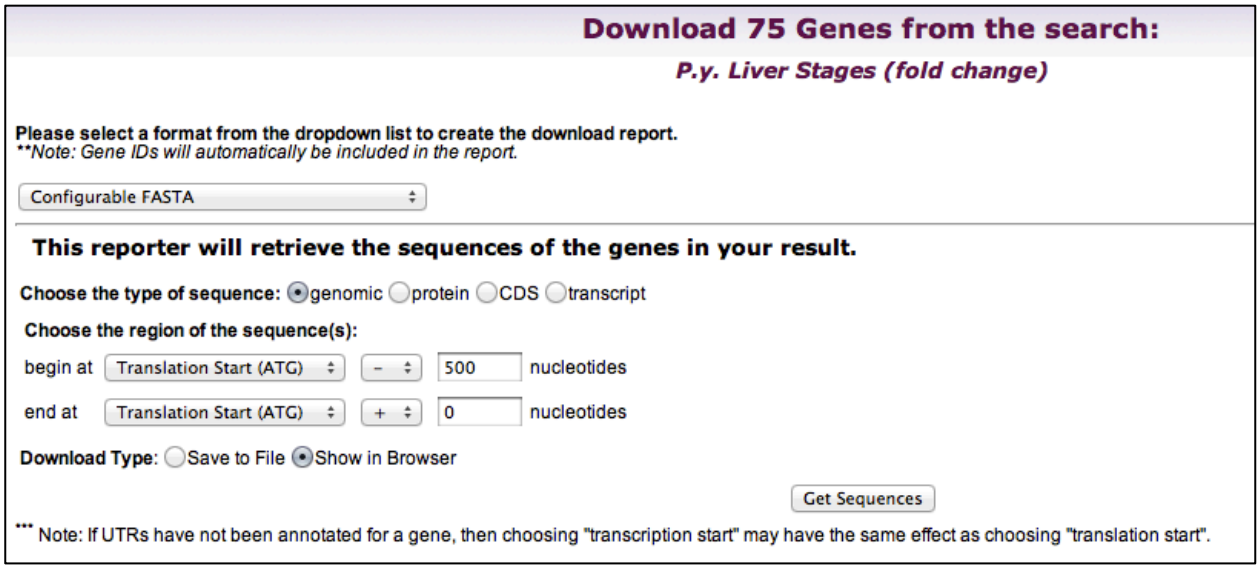
b. How many genes did you find? What you are interested in is looking at the nucleotide sequence upstream of the start sites of these genes. How can you do this in bulk? PlasmoDB has a sequence retrieval tool that allows you to download results of your searches in bulk. This includes a tool that allows you to specify the sequence you want.



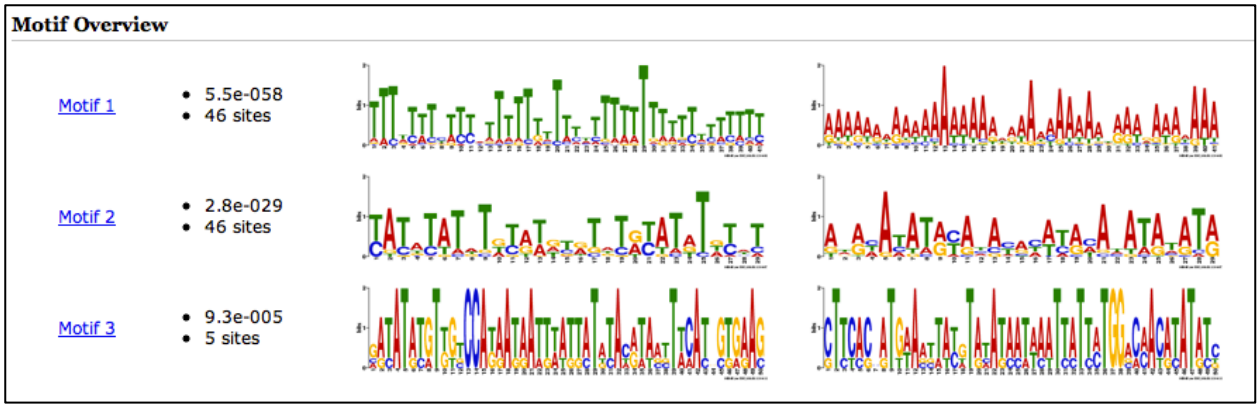
c. After you click on “Download ### Genes”, you are offered a drop down menu of options. Explore these; which one will allow you to specify the sequence to download. (hint: Configurable FASTA)



d. Define the sequence you want to retrieve. For this exercise retrieve 500 nucleotides up-stream of the start of translation.



e. The next step is to take this sequence and run it through a DNA motif finder such as MEME (http://meme.sdsc.edu/meme/intro.html). To speed up this process we have pre-run the motif finder and results are presented here:



The regular expression for each of these motifs is presented here:

Motif 1:

TTT[TAG]T[TA]T[CT][TA][TC][TC][ATC]TTTTT[TG]TTT[TC][TA]TTT[TA]TTTT[TA]T[TC][TA][TC][TA][TC]TT[TC]

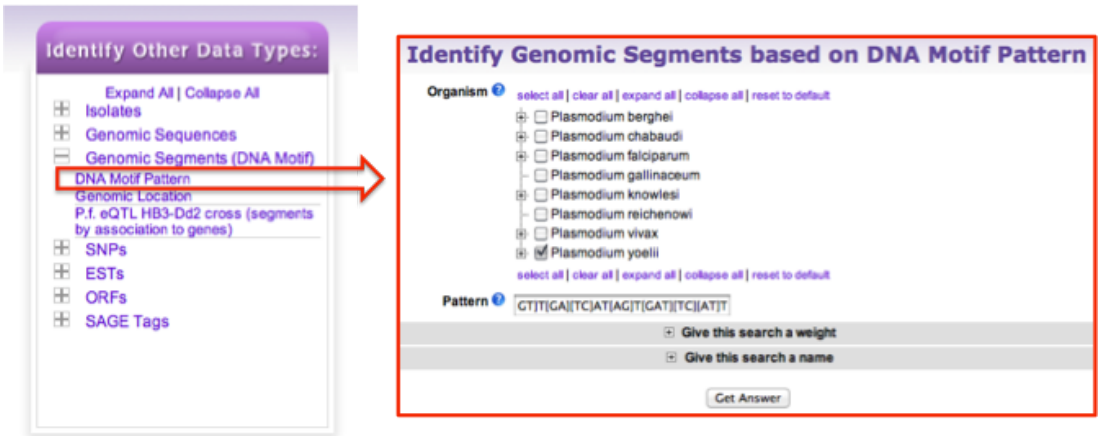
Motif 2:

[TC]A[TC][AT][TC]AT[ATG]T[GTA][TC][AG][TA][GAT][TC][GA]T[AGT]T[GA][TC]AT[AG]T[GAT][TC][AT]T

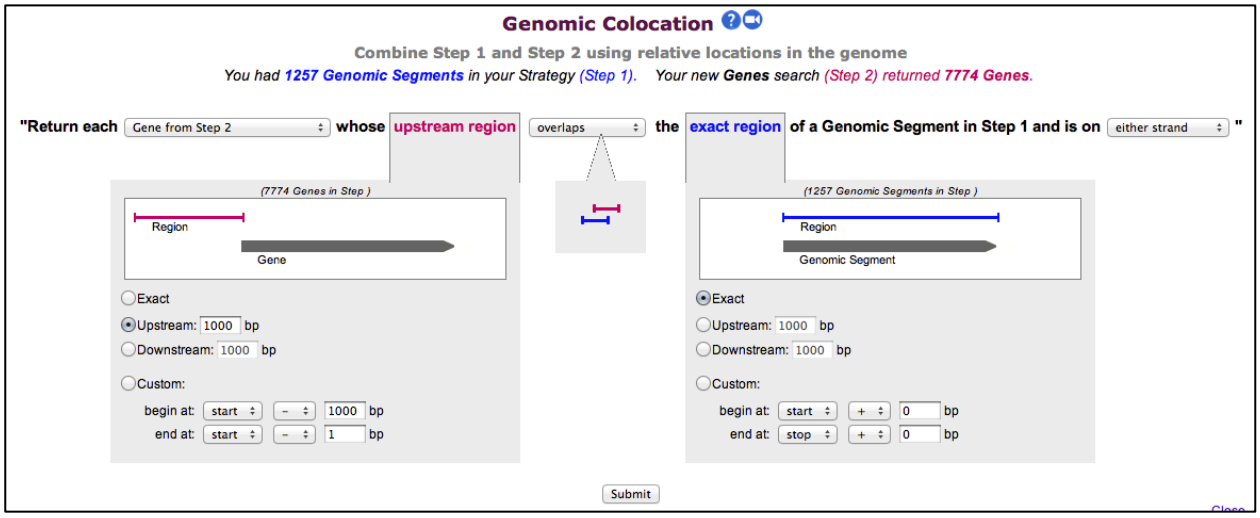
Motif 3:

[GAC][AG][TC]AT[AG][TC][GA]T[TG][GT][TCG]CCA[TG][AG]A[TG][AG]A[TA][TG][TA][AT][TG][TG][AC]T[AGT][TC]A[CAT][AG][TA][AT][ACG][TCG]T[TA][CA]A[TC][GACTA][GC][TG][GA][AG]A[GC]

f. Can you find any of these motifs in the *P. yoelii* genome? (*Hint*: use the DNA motif query)



g. How many times did this motif occur in the genome? How many of them are in the upstream region of genes? Can you find all *P. yoelii* genes that are within 1000 nucleotides downstream of the motif? (*Hint*: use the genomic colocation option when combining searches).



h. Do these genes have orthologs in other *Plasmodium* species? (*Hint*: add a step to your search strategy and transform the results to their orthologs).