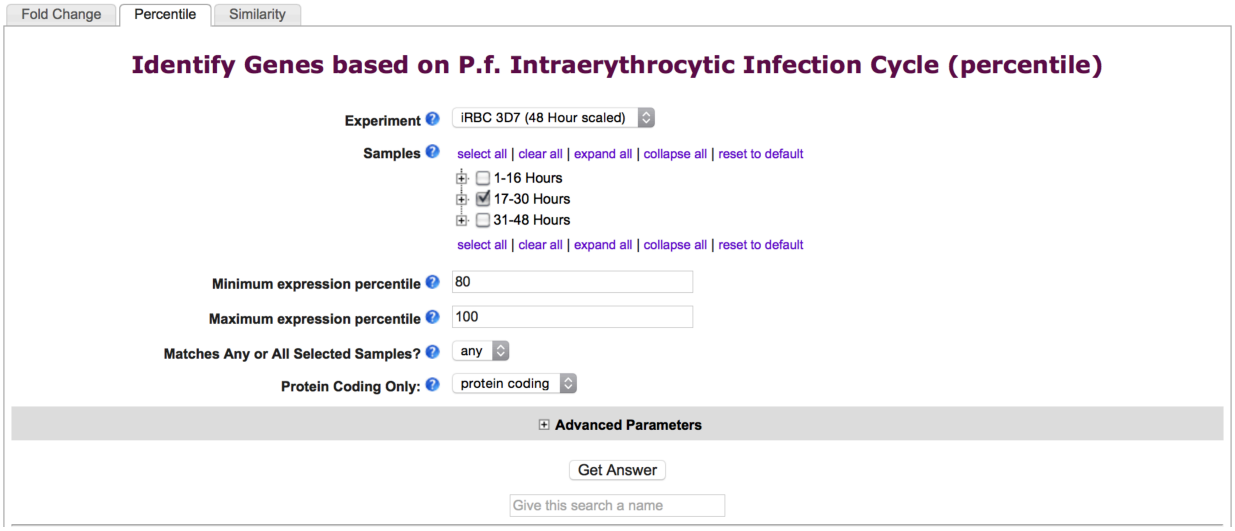
Data Integration Exercises - I
Complex strategies with Genomic Colocation

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

Identify genes that meet these four criteria:

1. are highly expressed (80-90th percentile) in *P.* falciparum 3D7 parasites at 24-30 hours of the iRBC cycle and,
2. are located within 1000 bp of each other,
3. are divergently transcribed,

Hint: first use the "Genes based on Microarray Evidence" -> "Erythrocytic expression time series (3D7,DD2, HB3) (Bozdech et al. and Linas et al.)" -> "**P"** search (percentile).

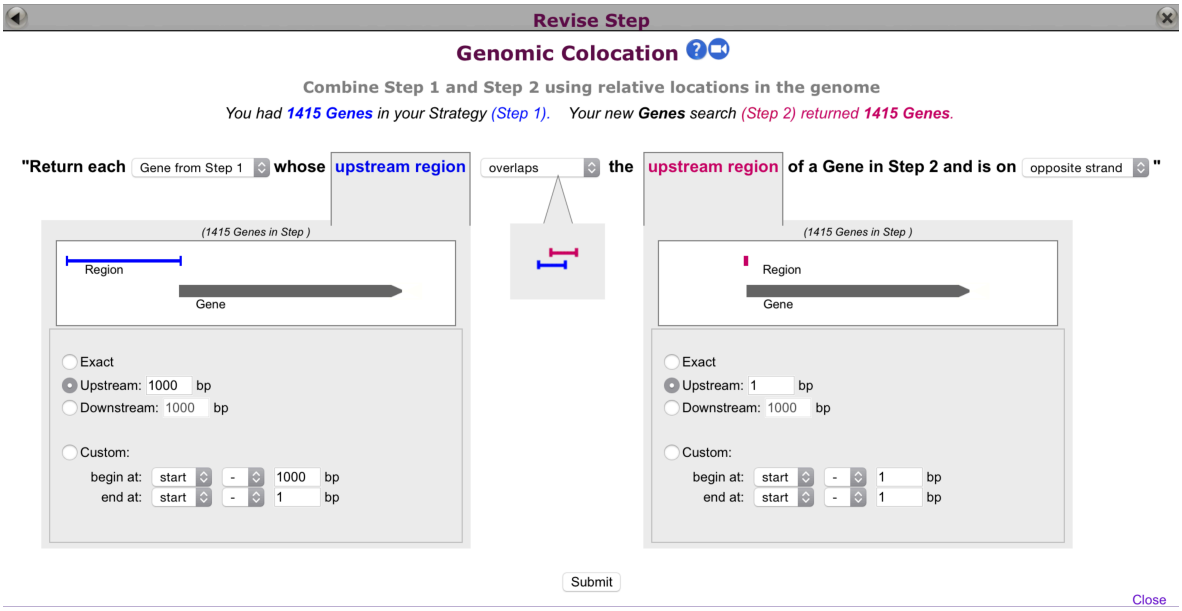


- Add a step that is the same as the first step and select the genomic colocation (1 relative to 2) operation.
- Set up the form to identify those genes that are transcribed on the opposite strand that have their starts located within 1000 bp of another gene's start.
- If you are having difficulty setting this up, you can see the strategy at:

  http://plasmodb.org/plasmo/im.do?s=97840366c30611ef

  Cut and paste the link into your browser if the hyperlink does not work.

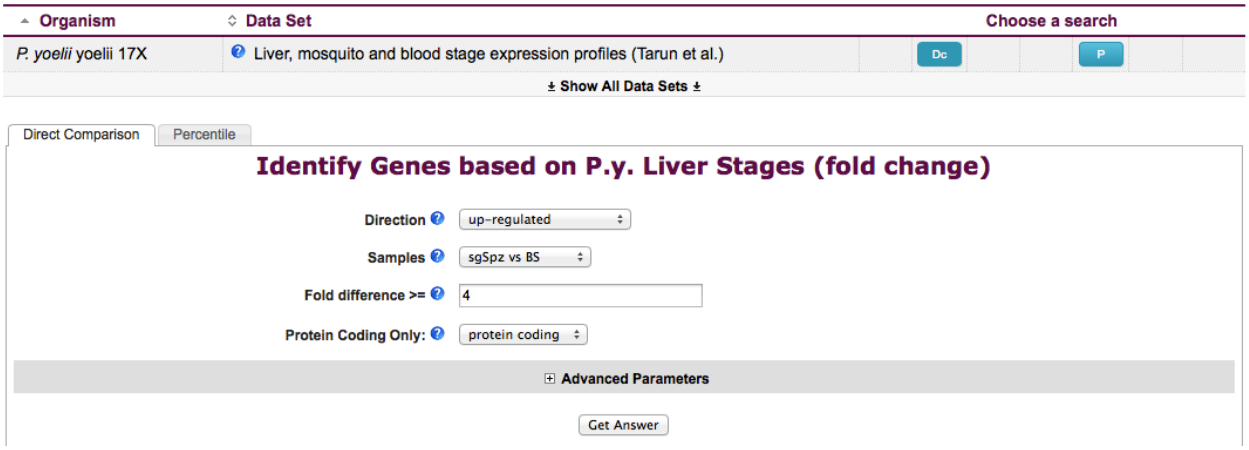
- Turn on the “Pf-iRBC 48hr - Graph” column to assess how well the pairs of genes compare in terms of expression. The pairs of genes are located one above the other in the result table if sorted by location.
- Note that you could do similar types of experiments to look at potential co-regulation / shared 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. Identifying conserved DNA elements upstream of genes

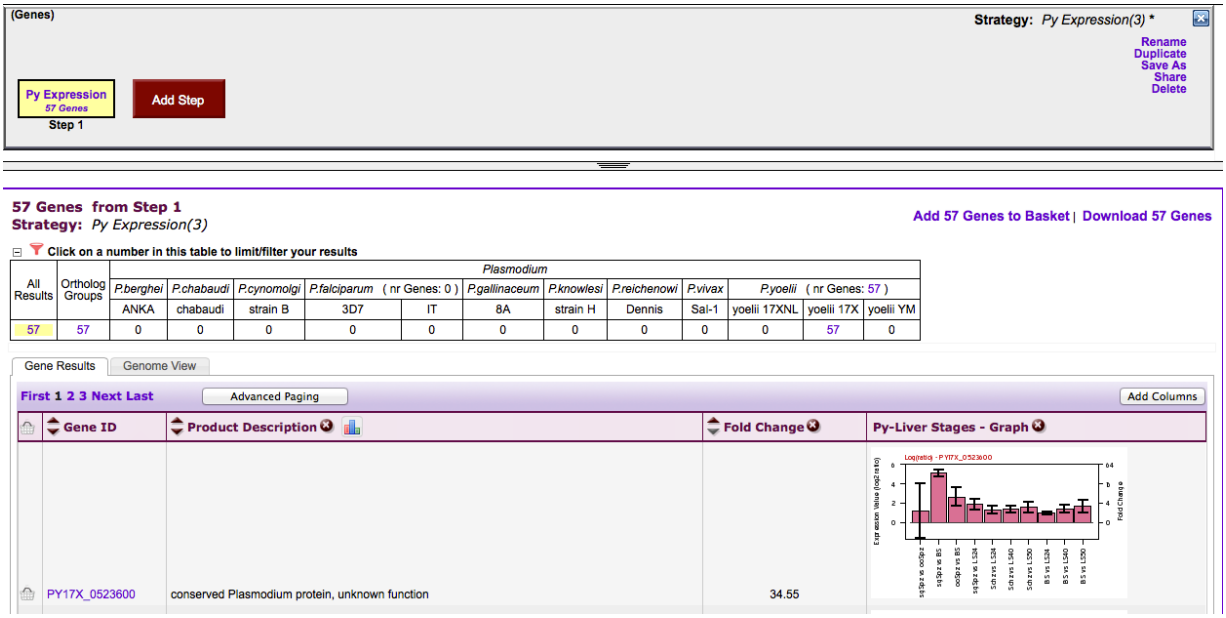
The goal of this exercise is to identify a DNA element in the upstream region of similarly regulated genes. You can use the same logic in this exercise with any life-cycle stage or organism of interest with available data .

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

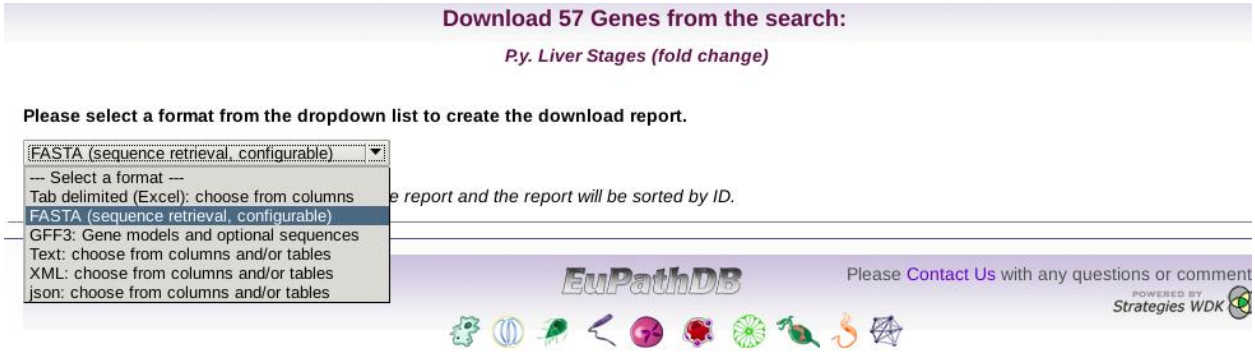


blood stage expression profiles (Tarun et al.) (direct comparison)]

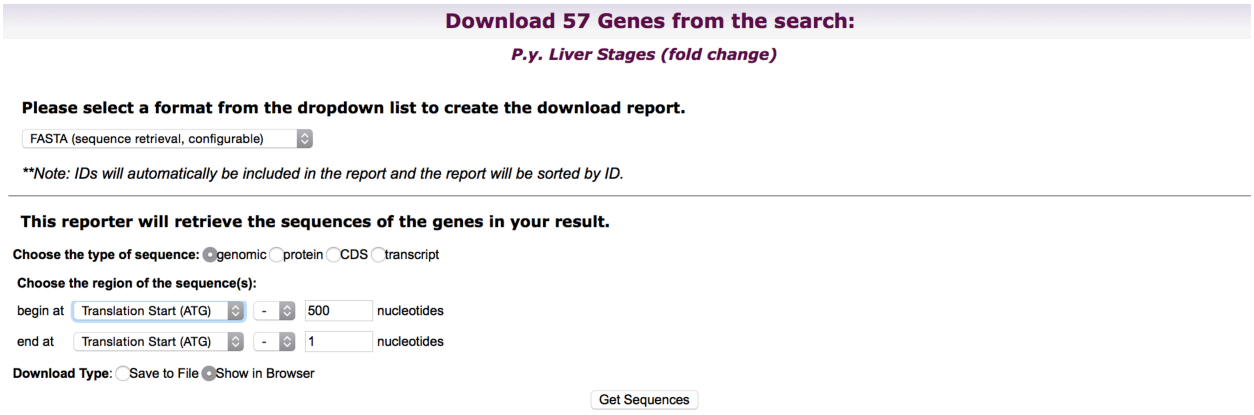
b. How many genes did you find? What you are interested in is looking at the nucleotide sequence upstream of the start sites of these genes. How can you do this in bulk? PlasmoDB has a sequence retrieval tool that allows you to download results of your searches in bulk. This includes a tool that allows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



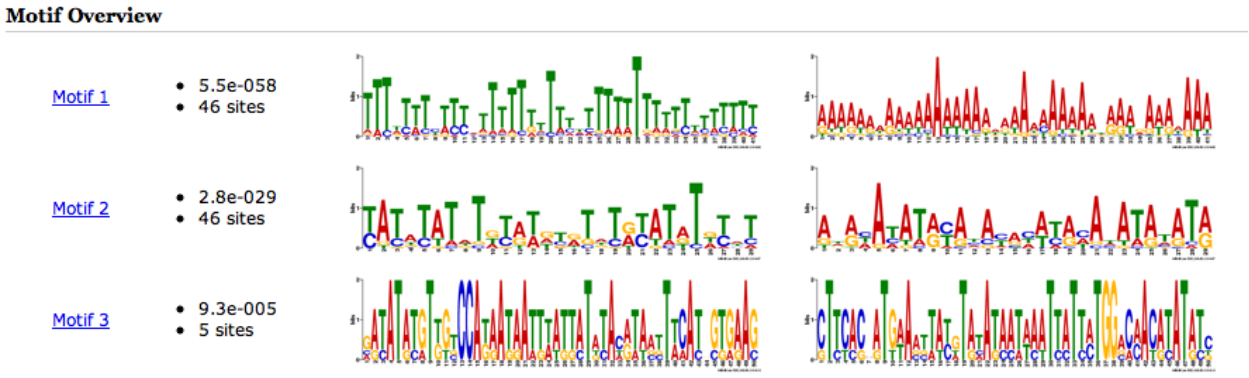
c. After you click on “Download ### Genes”, you are offered a drop down menu of options. Explore these; which one will allow you to specify the sequence to download. (*Hint*: Configurable FASTA)



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



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]

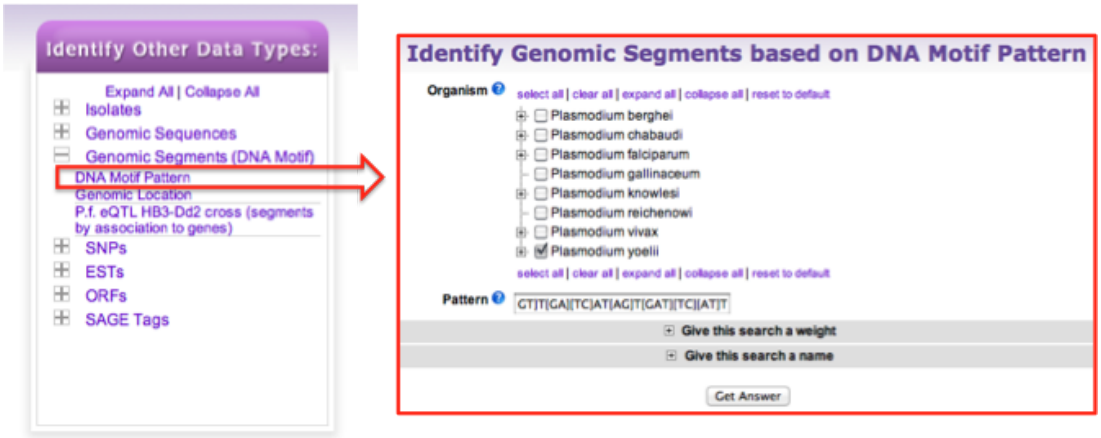
Motif 2:

[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

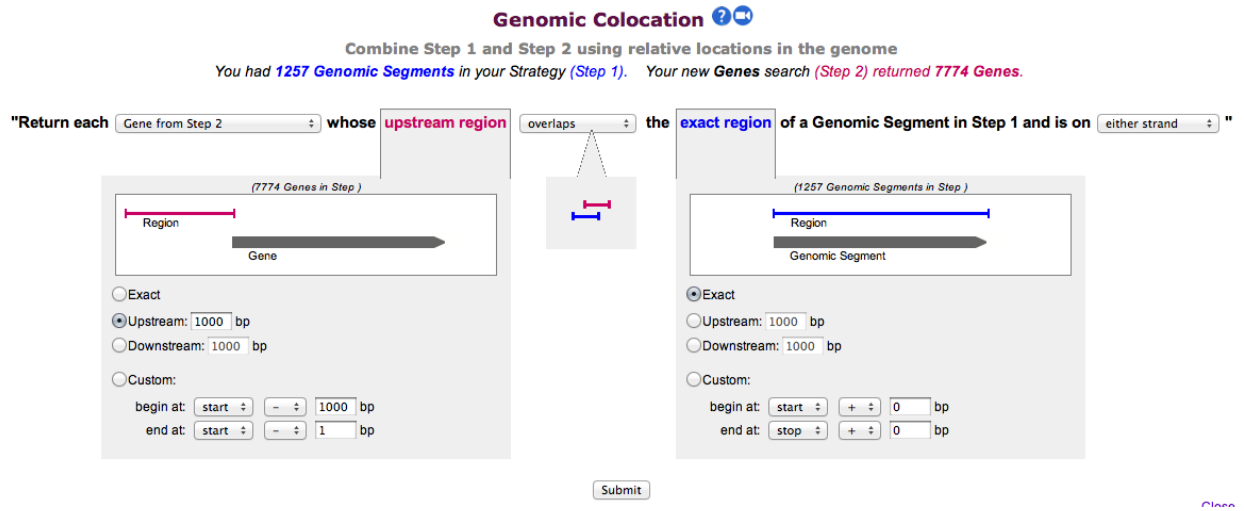
Motif 3:

[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]

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



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).



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

Optional: add a step and do the motif search on these orthologs to find out how many of them also contain the motif.