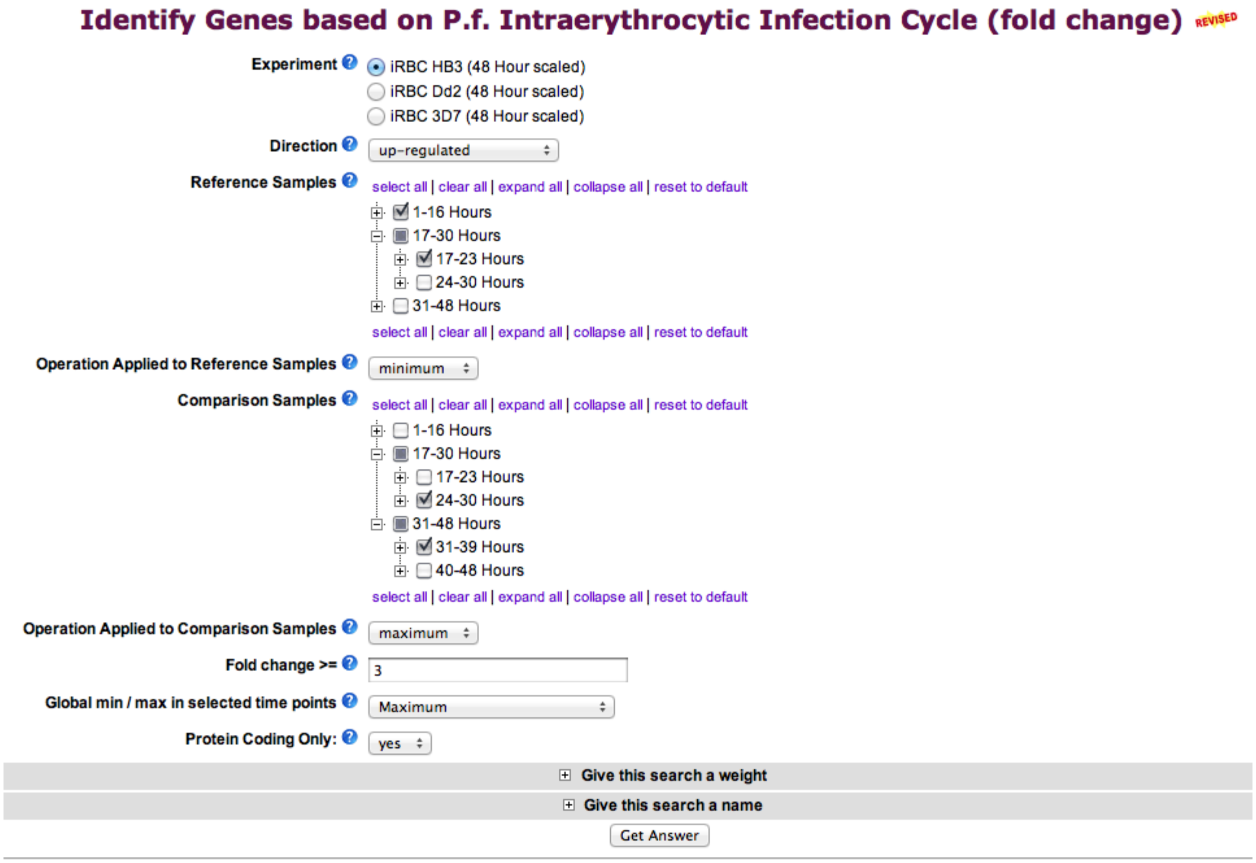
Complex strategies with Genomic Co-location
Exercise 13

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

Identify genes that are located within 1000 bp of each other and divergently transcribed, that are expressed maximally at day 30 of the iRBC cycle +- 8 hrs and show at least a 3-fold increase in expression.

- Hint: use the "Genes bases on Microarray Evidence" -> "***Intraerythrocytic Infection Cycle (DeRisi)***" -> "**P.f. Intraerythrocytic Infection Cycle (fold change)**" search.



Identify Genes based on P.f. Intraerythrocytic Infection Cycle (fold change) REVISED

Experiment ● iRBC HB3 (48 Hour scaled)
○ iRBC Dd2 (48 Hour scaled)
○ iRBC 3D7 (48 Hour scaled)

Direction: up-regulated

Reference Samples: select all | clear all | expand all | collapse all | reset to default
☑ 1-16 Hours
■ 17-30 Hours
☑ 17-23 Hours
□ 24-30 Hours
□ 31-48 Hours
select all | clear all | expand all | collapse all | reset to default

Operation Applied to Reference Samples: minimum

Comparison Samples: select all | clear all | expand all | collapse all | reset to default
□ 1-16 Hours
■ 17-30 Hours
□ 17-23 Hours
☑ 24-30 Hours
■ 31-48 Hours
☑ 31-39 Hours
□ 40-48 Hours
select all | clear all | expand all | collapse all | reset to default

Operation Applied to Comparison Samples: maximum

Fold change >=: 3

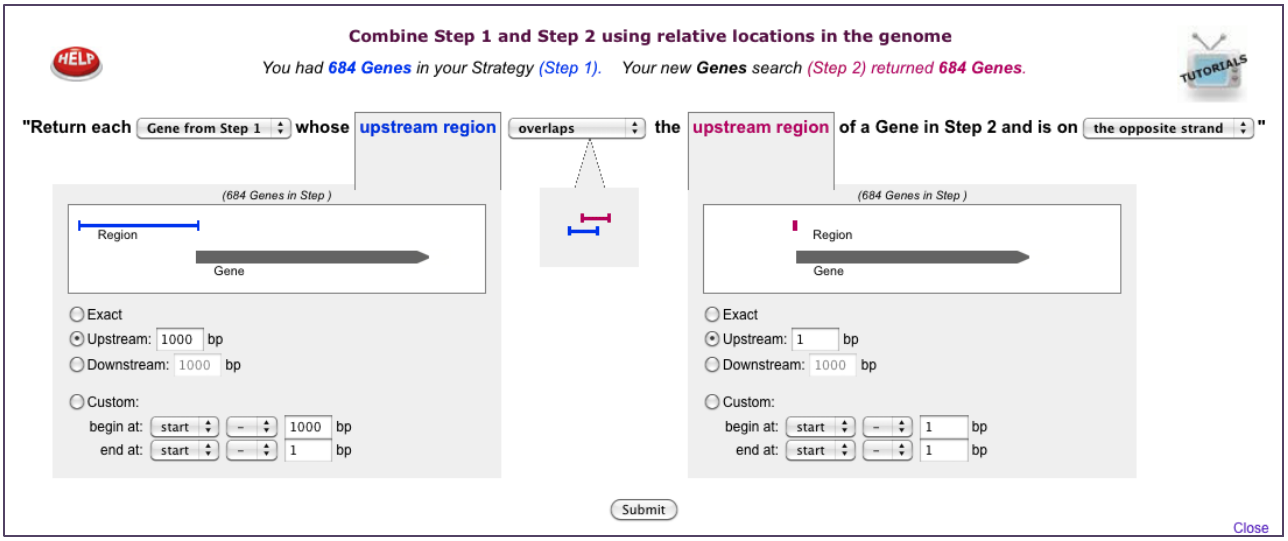
Global min / max in selected time points: Maximum

Protein Coding Only: yes

- Add a step that is the same as the first step. Note that you could copy the first step and then add an existing strategy to avoid setting the parameters again.
- Select the genomic colocation (Relative to … using relative genomic locations) 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.

- If you are having difficulty setting this up, you can see the strategy at:
- http://plasmodb.org/plasmo/im.do?s=6b8094bdb6738e05 Cut and paste the link into your browser if the hyperlink does not
- Turn on the “Pf-iRBC expr profile graph (GS array)” 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 next page 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.

13.2 Finding possible oocyst expressed genes based on DNA motifs.



Combine Step 1 and Step 2 using relative locations in the genome

*You had **684 Genes** in your Strategy (Step 1). Your new **Genes** search (Step 2) returned **684 Genes**.*

"Return each [Gene from Step 1] whose **upstream region** [overlaps] the **upstream region** of a Gene in Step 2 and is on [the opposite strand]"

(684 Genes in Step)

○ Exact
● Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start - 1000 bp
end at: start - 1 bp

(684 Genes in Step)

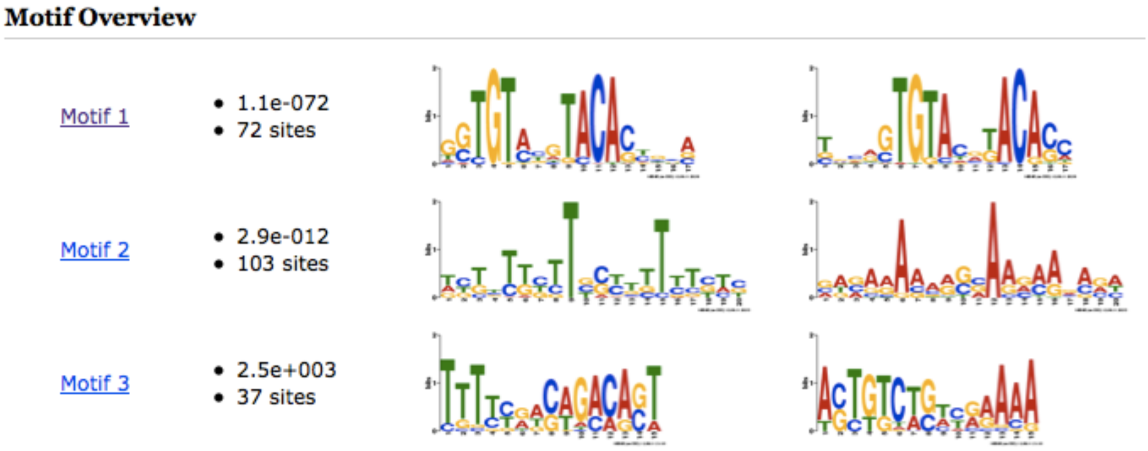
○ Exact
● Upstream: 1 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start - 1 bp
end at: start - 1 bp

Note: for this exercise use http://toxodb.org

In exercise 12.4 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

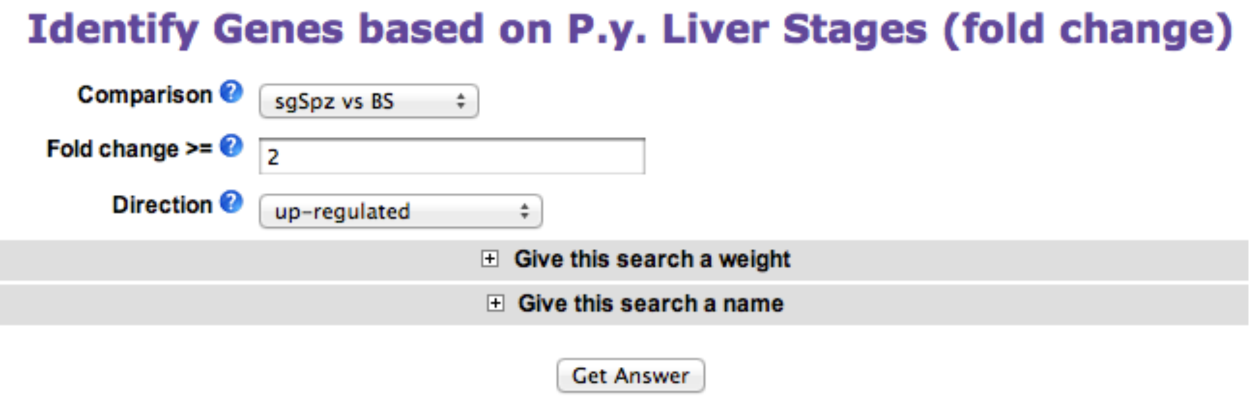
Motif 2
- 2.9e-012
- 103 sites

Motif 3
- 2.5e+003
- 37 sites

13.3. 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. For the purpose of this exercise, the goal is to identify such elements in genes upregulated in salivary gland sporozoites.

a. Identify genes that are upregulated in malaria sporozoites compared to blood stage parasites. Examine the microarray section of PlasmoDB. Can you identify an experiment that would give you this answer? (hint: look at other *Plasmodium* species, ie. *P. yoelii* [Parasite Liver Stages Survey (Kappe) ---> P.y. Liver Stages (fold change)]



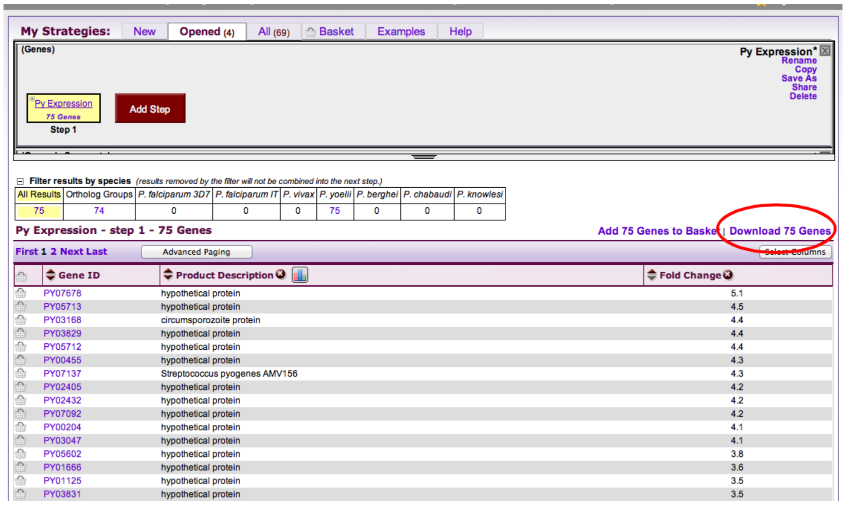
Identify Genes based on P.y. Liver Stages (fold change)

Comparison: sgSpz vs BS

Fold change >=: 2

Direction: up-regulated

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.



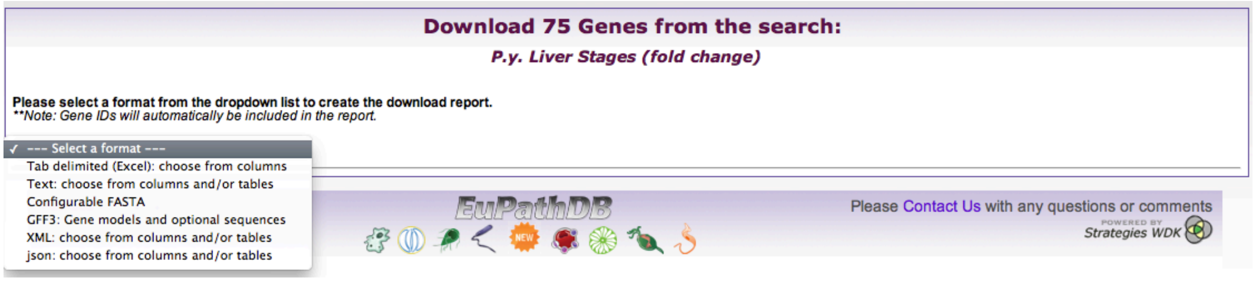
⊟ Filter results by species *(results removed by the filter will not be combined into the next step.)*

All Results	Ortholog Groups	*P. falciparum 3D7*	*P. falciparum IT*	*P. vivax*	*P. yoelii*	*P. berghei*	*P. chabaudi*	*P. knowlesi*
75	74	0	0	0	75	0	0	0

Py Expression - step 1 - 75 Genes

Gene ID	Product Description	Fold Change
PY07678	hypothetical protein	5.1
PY05713	hypothetical protein	4.5
PY03168	circumsporozoite protein	4.4
PY03829	hypothetical protein	4.4
PY05712	hypothetical protein	4.4
PY00455	hypothetical protein	4.3
PY07137	Streptococcus pyogenes AMV156	4.3
PY02405	hypothetical protein	4.2
PY02432	hypothetical protein	4.2
PY07092	hypothetical protein	4.2
PY00204	hypothetical protein	4.1
PY03047	hypothetical protein	4.1
PY05602	hypothetical protein	3.8
PY01666	hypothetical protein	3.6
PY01125	hypothetical protein	3.5
PY03831	hypothetical protein	3.5

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)



Download 75 Genes from the search:

P.y. Liver Stages (fold change)

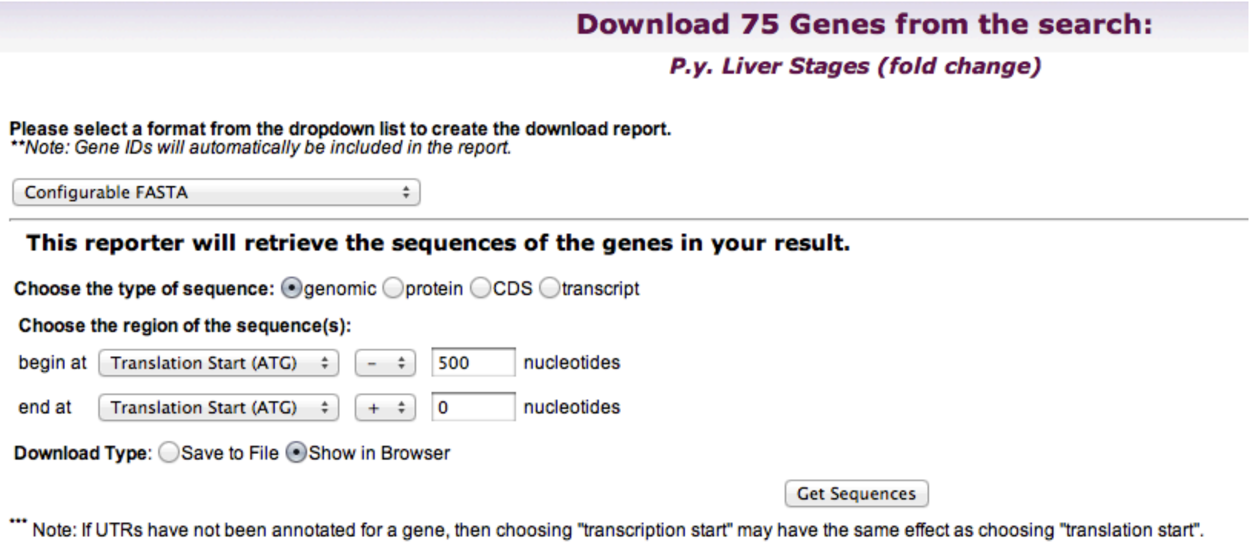
Please select a format from the dropdown list to create the download report.
***Note: Gene IDs will automatically be included in the report.*

✓ --- Select a format ---
Tab delimited (Excel): choose from columns
Text: choose from columns and/or tables
Configurable FASTA
GFF3: Gene models and optional sequences
XML: choose from columns and/or tables
json: choose from columns and/or tables
EuPathDB
NEW

Please Contact Us with any questions or comments

POWERED BY
Strategies WDK

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



Download 75 Genes from the search:

P.y. Liver Stages (fold change)

Please select a format from the dropdown list to create the download report.
***Note: Gene IDs will automatically be included in the report.*

Configurable FASTA

This reporter will retrieve the sequences of the genes in your result.

Choose the type of sequence: ● genomic ○ protein ○ CDS ○ transcript

Choose the region of the sequence(s):

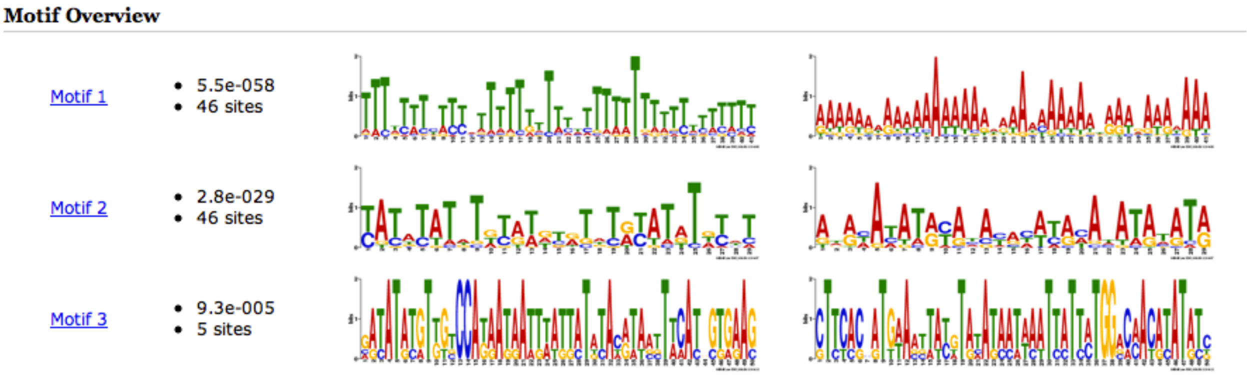
begin at Translation Start (ATG) – 500 nucleotides

end at Translation Start (ATG) + 0 nucleotides

Download Type: ○ Save to File ● Show in Browser

*** Note: If UTRs have not been annotated for a gene, then choosing "transcription start" may have the same effect as choosing "translation start".

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:



Motif Overview

Motif 1
- 5.5e-058
- 46 sites

Motif 2
- 2.8e-029
- 46 sites

Motif 3
- 9.3e-005
- 5 sites

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]

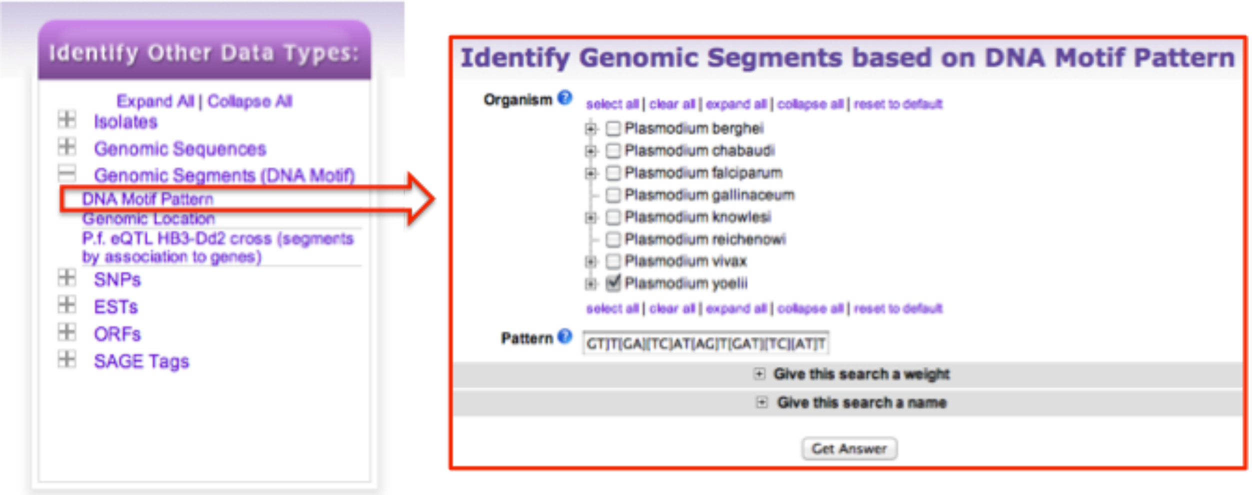
Motif 2:

[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)


Identify Other Data Types:
Expand All | Collapse All
Isolates
Genomic Sequences
Genomic Segments (DNA Motif)
DNA Motif Pattern
Genomic Location
P.f. eQTL HB3-Dd2 cross (segments by association to genes)
SNPs
ESTs
ORFs
SAGE Tags

Identify Genomic Segments based on DNA Motif Pattern

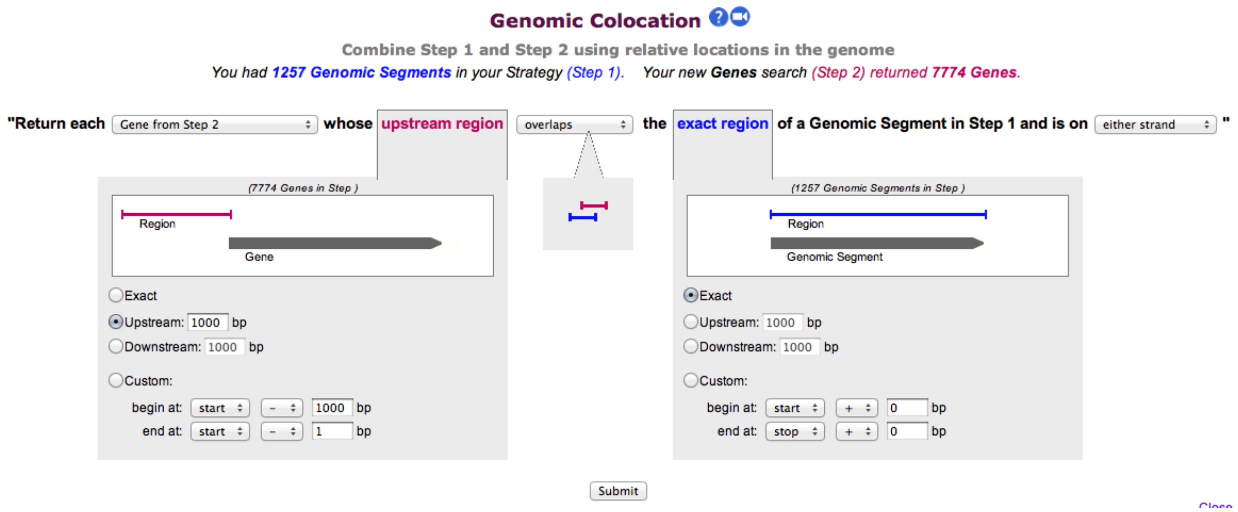
Organism: select all | clear all | expand all | collapse all | reset to default

- □ Plasmodium berghei
- □ Plasmodium chabaudi
- □ Plasmodium falciparum
- □ Plasmodium gallinaceum
- □ Plasmodium knowlesi
- □ Plasmodium reichenowi
- □ Plasmodium vivax
- ☑ Plasmodium yoelii

select all | clear all | expand all | collapse all | reset to default

Pattern: GT]T[GA][TC]AT[AG]T[GAT][TC][AT]T

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 genes in *P. yoelii* that are within 1000 nucleotides downstream of the motif? (hint: use the genomic colocation search).



Genomic Colocation

Combine Step 1 and Step 2 using relative locations in the genome

You had 1257 Genomic Segments in your Strategy (Step 1). Your new Genes search (Step 2) returned 7774 Genes.

"Return each Gene from Step 2 whose upstream region overlaps the exact region of a Genomic Segment in Step 1 and is on either strand "

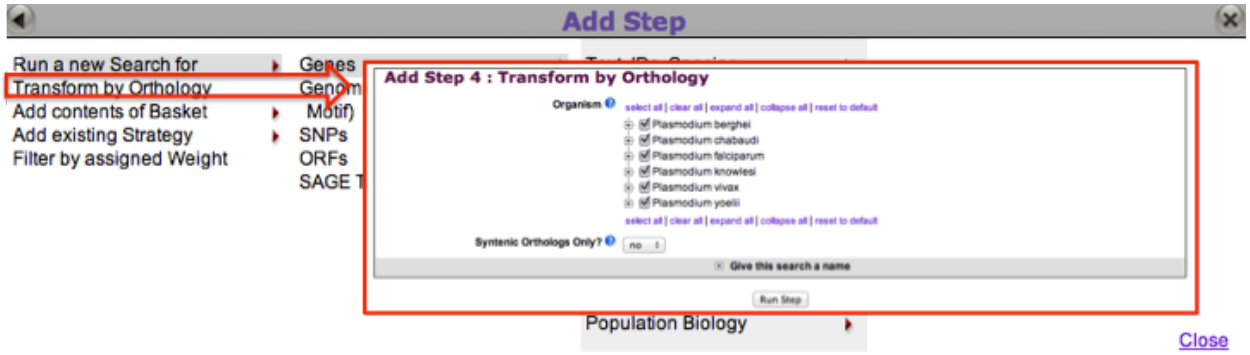
(7774 Genes in Step)

○ Exact
● Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start − 1000 bp
end at: start − 1 bp

(1257 Genomic Segments in Step)

● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start + 0 bp
end at: stop + 0 bp

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).



Add Step 4 : Transform by Orthology

Organism
☑ Plasmodium berghei
☑ Plasmodium chabaudi
☑ Plasmodium falciparum
☑ Plasmodium knowlesi
☑ Plasmodium vivax
☑ Plasmodium yoelii

Syntenic Orthologs Only? no