

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.

Fold Change
Percentile
Similarity

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

For the Experiment IRBC HB3 (48 Hour scaled)

return protein coding Genes

that are up-regulated

with a Fold change $\geq$ 2

between each gene's average expression value

in the following 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

and its average expression value

in the following 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

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

Advanced Parameters

Get Answer

Example showing one gene that would meet search criteria

(Dots represent this gene's expression values for selected samples)

Up-regulated

Reference Samples Comparison Samples

Average Comparison

Average Reference

2 fold

A maximum of four samples are shown when more than four are selected. You are searching for genes that are up-regulated between at least two reference samples and at least two comparison samples.

For each gene, the search calculates:

$$\text{fold change} = \frac{\text{average expression value in comparison samples}}{\text{average expression value in reference samples}}$$

and returns genes when fold change $\geq$ 2. To narrow the window, use the maximum reference value, or minimum comparison value. To broaden the window, use the minimum reference value, or maximum comparison value.

See the [detailed help for this 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.

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 1)

☐ Exact

☒ Upstream: 1000 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: start -- 1000 bp

end at: start -- 1 bp

(684 Genes in Step 2)

☐ Exact

☒ Upstream: 1 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: start -- 1 bp

end at: start -- 1 bp

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

<u>Motif 1</u>	• 1.1e-072 • 72 sites		
<u>Motif 2</u>	• 2.9e-012 • 103 sites		
<u>Motif 3</u>	• 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.

- 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.)])

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

Comparison ?

sgSpz vs BS

Fold change >= ?

2

Direction ?

up-regulated

+ Give this search a weight

+ Give this search a name

Get Answer

- 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.

My Strategies: [New](#) [Opened \(4\)](#) [All \(69\)](#) [Basket](#) [Examples](#) [Help](#)

(Genes)

Py Expression*
Resume
Copy
Save As
Share
Delete

Py Expression
75 Genes
Step 1

Add Step

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

All Results	Ortholog Groups	<i>P. falciparum</i> 3D7	<i>P. falciparum</i> IT	<i>P. vivax</i>	<i>P. yoelii</i>	<i>P. berghei</i>	<i>P. chabaudi</i>	<i>P. knowlesi</i>
75	74	0	0	0	75	0	0	0

Py Expression - step 1 - 75 Genes

Add 75 Genes to Basket

Download 75 Genes

First 1 2 Next Last

Advanced Paging

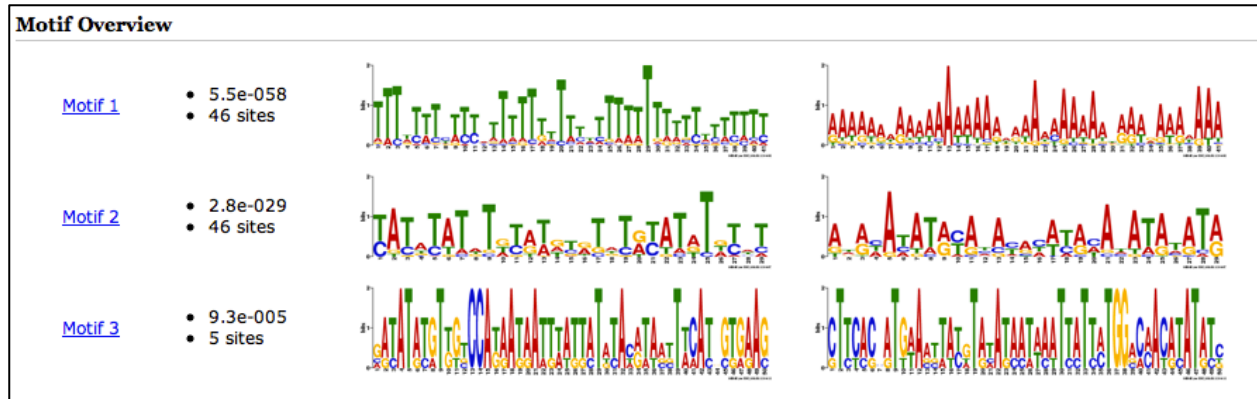
Select Columns

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
PY00456	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
PY01656	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)

- 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[TA]T[TA]T[CT][TA][TC][TC][ATC]TTTT[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[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[TT][GA][TC]AT[AG]T[GC][TC][AT]T

☐ Give this search a weight

☐ Give this search a name

[Get Answer](#)

- 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 collocation option when combining searches).

Genomic Collocation ?

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 whose **upstream region** overlaps the **exact region** of a Genomic Segment in Step 1 and is on

(7774 Genes in Step 2)

☐ Exact

☒ Upstream: 1000 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: bp

end at: bp

(1257 Genomic Segments in Step 1)

☒ Exact

☐ Upstream: 1000 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: bp

end at: 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

Run a new Search for

Transform by Orthology

Add contents of Basket

Add existing Strategy

Filter by assigned Weight

Add Step 4 : Transform by Orthology

Organism

- ☒ Plasmodium berghei
- ☒ Plasmodium chabaudi
- ☒ Plasmodium falciparum
- ☒ Plasmodium knowlesi
- ☒ Plasmodium vivax
- ☒ Plasmodium yoelii

Syntenic Orthologs Only?

Population Biology