

Complex strategies with Genomic Colocation

Exercise 14

14.1 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 expressed maximally at day 30 of the iRBC cycle +/- 8 hrs and,
4. show at least a 3-fold increase in expression.

- Hint: first use the “Genes bases on Microarray Evidence” -> “Erythrocytic expression time series (3D7,DD2, & HB3) (Bozdech et al. and Linas et al.)” -> “Fc” 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 3

between each gene's minimum 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 maximum 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
- ☒ 31-39 Hours

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

Example showing one gene that would meet search criteria

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

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:

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

and returns genes when **fold change** \geq 3. This calculation creates the **broadest** window of expression values in which to look for genes that meet your fold change cutoff. To narrow the window, use the average or maximum reference value, or average or minimum comparison value.

[See the detailed help for this search.](#)

Advanced Parameters

Weight 10

Global min / max in selected time points Maximum

[Get Answer](#)

- 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.
- If you are having difficulty setting this up, you can see the strategy at: <http://plasmodb.org/plasmo/im.do?s=0ebfe58b1c9b42cc>. 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.

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)

Region

Gene

☐ Exact

☒ Upstream: 1000 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: start -- 1000 bp

end at: start -- 1 bp

(684 Genes in Step 2)

Region

Gene

☐ Exact

☒ Upstream: 1 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: start -- 1 bp

end at: start -- 1 bp

Submit

Close

14.2 Finding possible oocyst expressed genes based on DNA motifs.

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

In exercise 13.4 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.

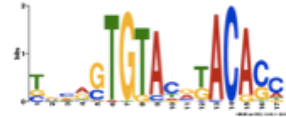
After you have your sequences you will need to run them through a DNA pattern finder like MEME (<http://meme.sdsc.edu/meme/intro.html>). Results from a submission to MEME could take up to several hours so for your convenience 300 nucleotides upstream of all the oocyst results were analyzed using MEME – 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

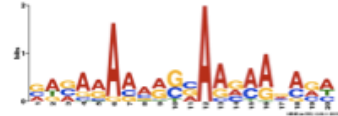
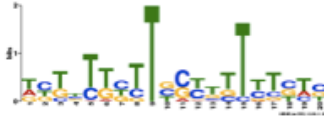
[Motif 1](#)

- 1.1e-072
- 72 sites



[Motif 2](#)

- 2.9e-012
- 103 sites



[Motif 3](#)

- 2.5e+003
- 37 sites



14.3. 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.) (direct comparison)])

Organism

Data Set

Choose a search

P. yoelii yoelii 17X

Liver, mosquito and blood stage expression profiles (Tarun et al.)

Do

P

Show All Data Sets

Direct Comparison

Percentile

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

Direction

up-regulated

Samples

sgSpz vs BS

Fold difference >=

4

Protein Coding Only:

protein coding

Advanced Parameters

Get Answer

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.

(Genes)

Strategy: Py Expression(3) *

Rename
Duplicate
Save As
Share
Delete

Py Expression
57 Genes
Step 1

Add Step

57 Genes from Step 1
Strategy: Py Expression(3)

Add 57 Genes to Basket | Download 57 Genes

Click on a number in this table to limit/filter your results

All Results	Ortholog Groups	Plasmodium										
		P.berghiei	P.chabaudi	P.cynomolgi	P.falciparum	(nr Genes: 0)	P.gallinaceum	P.knowlesi	P.reichenowi	P.vivax	P.yoelii	(nr Genes: 57)
		ANKA	chabaudi	strain B	3D7	IT	8A	strain H	Dennis	Sal-1	yoelii 17XNL	yoelii 17X
57	57	0	0	0	0	0	0	0	0	0	0	57

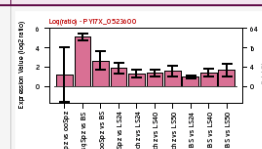
Gene Results

Genome View

First 1 2 3 Next Last

Advanced Paging

Add Columns

Gene ID	Product Description	Fold Change	Py-Liver Stages - Graph
PY17X_0523600	conserved Plasmodium protein, unknown function	34.55	

- 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
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 up-stream 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

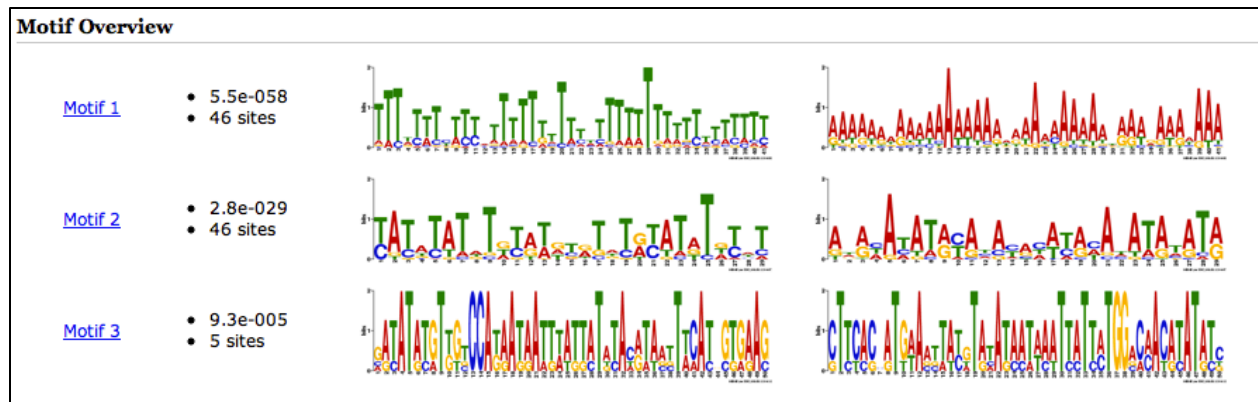
end at Translation Start (ATG) + 0 nucleotides

Download Type: ☐ Save to File ☒ Show in Browser

[Get Sequences](#)

*** 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:



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[T C][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][T A][AT][TG][TG][AC]T[AGT][TC]A[CAT][AG][TA][AT][ACG][TCG]T[TA][CA]A[TC][GA CTA][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

Pattern GTT[GA][TC]AT[AG]T[GAT][TC][AT]T

☐ Give this search a weight

☐ Give this search a name

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).

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 2)

☐ Exact

☒ Upstream: bp

☐ Downstream: bp

☐ Custom:

begin at: - bp

end at: - bp

(1257 Genomic Segments in Step 1)

☒ Exact

☐ Upstream: bp

☐ Downstream: 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 Genes

Transform by Orthology

Add contents of Basket Genom

Add existing Strategy Motif)

Filter by assigned Weight SNPs

ORFs

SAGE T

Add Step 4 : Transform by Orthology

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

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

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

Syntenic Orthologs Only? ☐ no

Give this search a name

Population Biology [Close](#)

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