

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 (>80th 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).

Fold Change

Percentile

Similarity

Identify Genes based on P.f. Intraerythrocytic Infection Cycle (percentile)

Experiment ? iRBC 3D7 (48 Hour scaled)

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

1-16 Hours

17-30 Hours

31-48 Hours

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

Minimum expression percentile ? 80

Maximum expression percentile ? 100

Matches Any or All Selected Samples? ? any

Protein Coding Only: ? protein coding

+ Advanced Parameters

Get Answer

Give this search a name

- 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.

Revise Step

Genomic Colocation ?

Combine Step 1 and Step 2 using relative locations in the genome
 You had **1415 Genes** in your Strategy (Step 1). Your new **Genes** search (Step 2) returned **1415 Genes**.

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

(1415 Genes in Step 1)

☐ Exact
☒ Upstream: 1000 bp
☐ Downstream: 1000 bp

☐ Custom:
 begin at: start - 1000 bp
 end at: start - 1 bp

(1415 Genes in Step 2)

☐ Exact
☒ Upstream: 1 bp
☐ Downstream: 1000 bp

☐ Custom:
 begin at: start - 1 bp
 end at: start - 1 bp

Submit

[Close](#)

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

Organism

Data Set

P. yoelii yoelii 17XLiver, mosquito and blood stage expression profiles (Tarun et al.)

DoP

Show All Data Sets

Direct ComparisonPercentile

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

Directionup-regulated

SamplessgSpz vs BS

Fold difference >=4

Protein Coding Only:protein coding

Advanced Parameters

Get Answer

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? PlasmDB 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) *

Py Expression
#7 Genes
Step 1

Add Step

Rename
Duplicate
Save As
Share
Delete

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

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

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 57 Genes from the search:
P.y. Liver Stages (fold change)

Please select a format from the dropdown list to create the download report.

FASTA (sequence retrieval, configurable)

- Select a format ---
- Tab delimited (Excel): choose from columns
- FASTA (sequence retrieval, configurable)
- GFF3: Gene models and optional sequences
- Text: choose from columns and/or tables
- XML: choose from columns and/or tables
- json: choose from columns and/or tables

e report and the report will be sorted by ID.


Please [Contact Us](#) with any questions or comment
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 57 Genes from the search:
P.y. Liver Stages (fold change)

Please select a format from the dropdown list to create the download report.

FASTA (sequence retrieval, configurable)

****Note: IDs will automatically be included in the report and the report will be sorted by ID.**

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

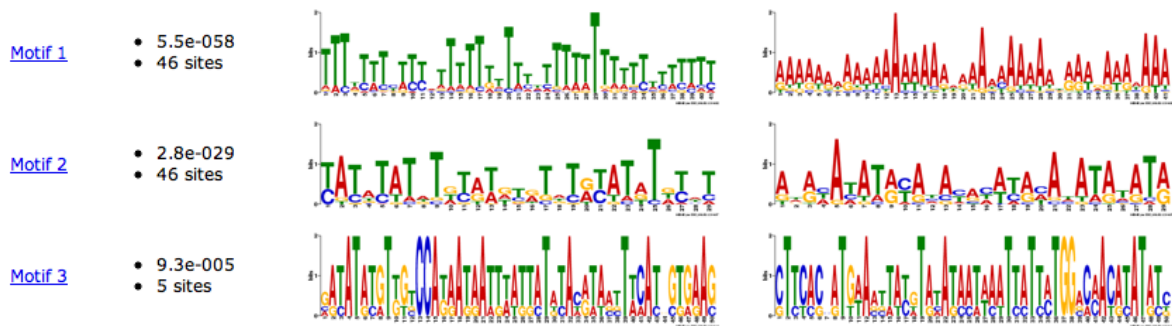
end at Translation Start (ATG) - 1 nucleotides

Download Type: ☐ Save to File ☒ Show in Browser

[Get Sequences](#)

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



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[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)

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

Get Answer

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)

Region

Gene

☐ Exact

☒ Upstream: 1000 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: start - 1000 bp

end at: start - 1 bp

(1257 Genomic Segments in Step)

Region

Genomic Segment

☒ Exact

☐ Upstream: 1000 bp

☐ Downstream: 1000 bp

☐ Custom:

begin at: start + 0 bp

end at: stop + 0 bp

Submit

[Close](#)

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 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

Run Step

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.

2. Find *Plasmodium falciparum* antigens that are immunogenic.

For this exercise use <http://plasmodb.org>

- a. You hope to identify antigens (genes) that antibodies to which may be protective to malaria infection. You can do this by identifying genes that exhibit an

increased immunogenicity in children (ages 0-12) with no disease (normal) compared to children with disease (infected). *Hint:* the “Serum Antibody Levels” search is available in the “Host Response” menu item in the “Identify Genes By” section of the home page.

Identify Genes by:

Expand All | Collapse All

- Text, IDs, Organism
- Genomic Position
- Gene Attributes
- Protein Attributes
- Protein Features
- Similarity/Pattern
- Transcript Expression
- Protein Expression
- Cellular Location
- Putative Function
- Evolution
- Population Biology
- Host Response
 - Serum Antibody Levels

Identify Genes based on Serum Antibody Levels

Reference Samples 263 of 421 selected Age between 0 and 12 X Disease State is Infected X

Select Reference Samples

Comparison Samples 70 of 421 selected Age between 0 and 12 X Disease State is Normal X

Select Comparison Samples View selected Comparison Samples (70) Collapse

Disease State

IPCR Result

Environmental history

General Information Of Study Subject

Age

Sex

Disease State

The name of the pathology diagnosed in the organism from which the biomaterial was derived. The disease state is normal if no disease has been diagnosed.

select all | clear all

☐ Infected

☒ Normal

Comparison Samples	
272	64.61%
149	35.39%

Legend: All Comparison Samples (grey bar), Comparison Samples remaining when other criteria has been applied (red bar)

Metadata category to color graph by DiseaseState

Direction increased immunogenicity

P value less than or equal to 0.05

Note that you will be comparing “comparison” samples to “reference” samples. So in this example, your comparison samples will be normal children and your reference samples will be infected children. You can configure the samples by making selection from left. What do your results look like? Could these represent potential protective antigens? Analyze your result (Analyze tab) to see if there are any biological processes that are enriched. Do these make sense based on your search?