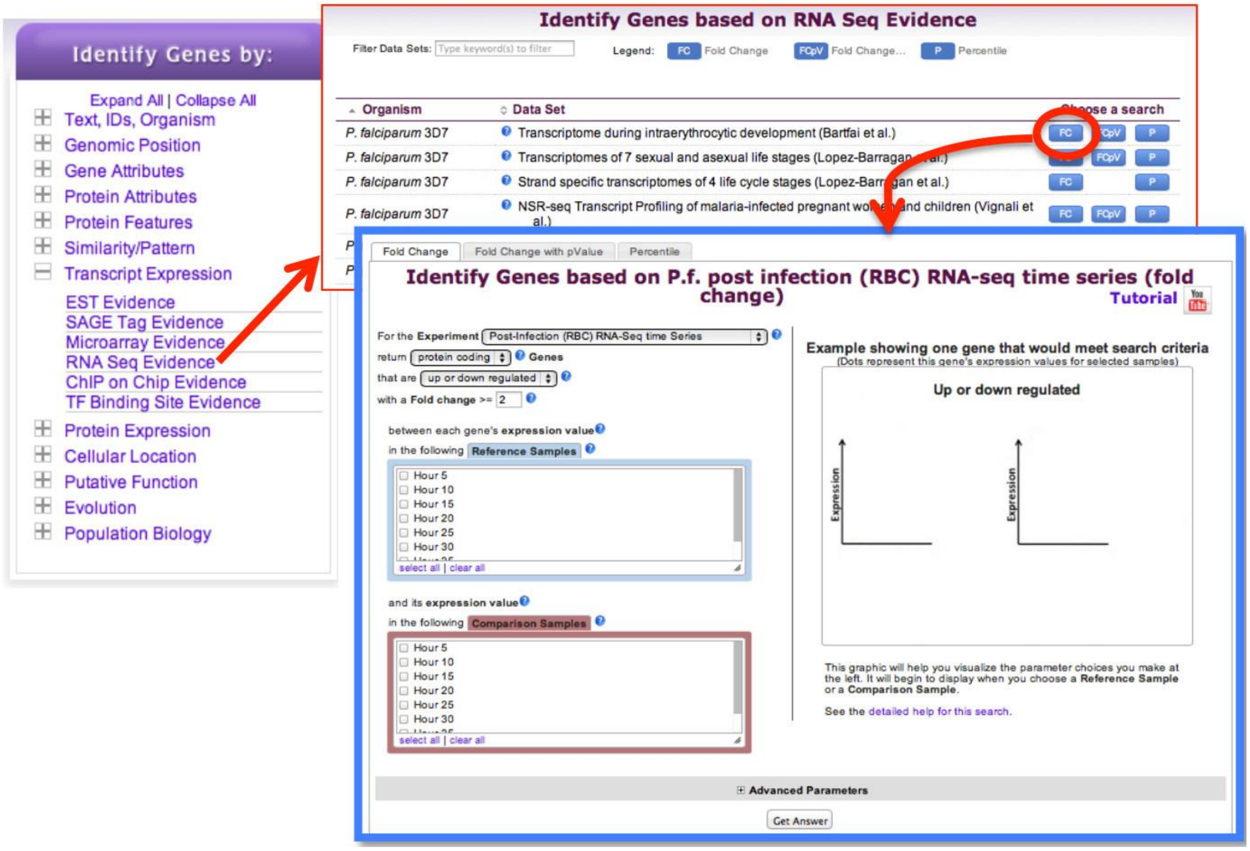
Functional Genomics
Transcriptomics and Proteomics

1. Exploring RNA sequence data in *Plasmodium falciparum.*
Note: For this exercise use http://www.plasmodb.org

1a. Find all genes in *P. falciparum* that are up-regulated during the later stages of the intraerythrocytic cycle.

- Hint: Use the fold change search for the data set "Transcriptome during intraerythrocytic development (Bartfai *et al.*)". For this data set, synchronized Pf3D7 parasites were assayed by RNA-seq at 8 time-points during the iRBC cycle. We want to find genes that are up-regulated in the later time points (30, 35, 40 hours) using the early time points (5, 10, 15, 20, 25 hours) as reference.



Identify Genes based on RNA Seq Evidence

Filter Data Sets: Type keyword(s) to filter

Legend: FC Fold Change FCpV Fold Change... P Percentile

Organism	Data Set	Choose a search
P. falciparum 3D7	Transcriptome during intraerythrocytic development (Bartfai et al.)	FC CpV P
P. falciparum 3D7	Transcriptomes of 7 sexual and asexual life stages (Lopez-Barragan al.)	FCpV P
P. falciparum 3D7	Strand specific transcriptomes of 4 life cycle stages (Lopez-Barr gan et al.)	FC P
P. falciparum 3D7	NSR-seq Transcript Profiling of malaria-infected pregnant wo and children (Vignali et al.)	FC FCpV P

Identify Genes based on P.f. post infection (RBC) RNA-seq time series (fold change)

For the Experiment: Post-Infection (RBC) RNA-Seq time Series

return: protein coding Genes

that are: up or down regulated

with a Fold change >= 2

between each gene's expression value

in the following Reference Samples

□ Hour 5
□ Hour 10
□ Hour 15
□ Hour 20
□ Hour 25
□ Hour 30

select all | clear all

and its expression value

in the following Comparison Samples

□ Hour 5
□ Hour 10
□ Hour 15
□ Hour 20
□ Hour 25
□ Hour 30

select all | clear all

Example showing one gene that would meet search criteria
(Dots represent this gene's expression values for selected samples)

This graphic will help you visualize the parameter choices you make at the left. It will begin to display when you choose a **Reference Sample** or a **Comparison Sample**.

See the detailed help for this search.

- Hint: there are a number of parameters to manipulate in this search. As you modify parameters on the left side note the dynamic help on the right side. See screenshots.

- **Direction**: the direction of change in expression. **Choose up-regulated**.

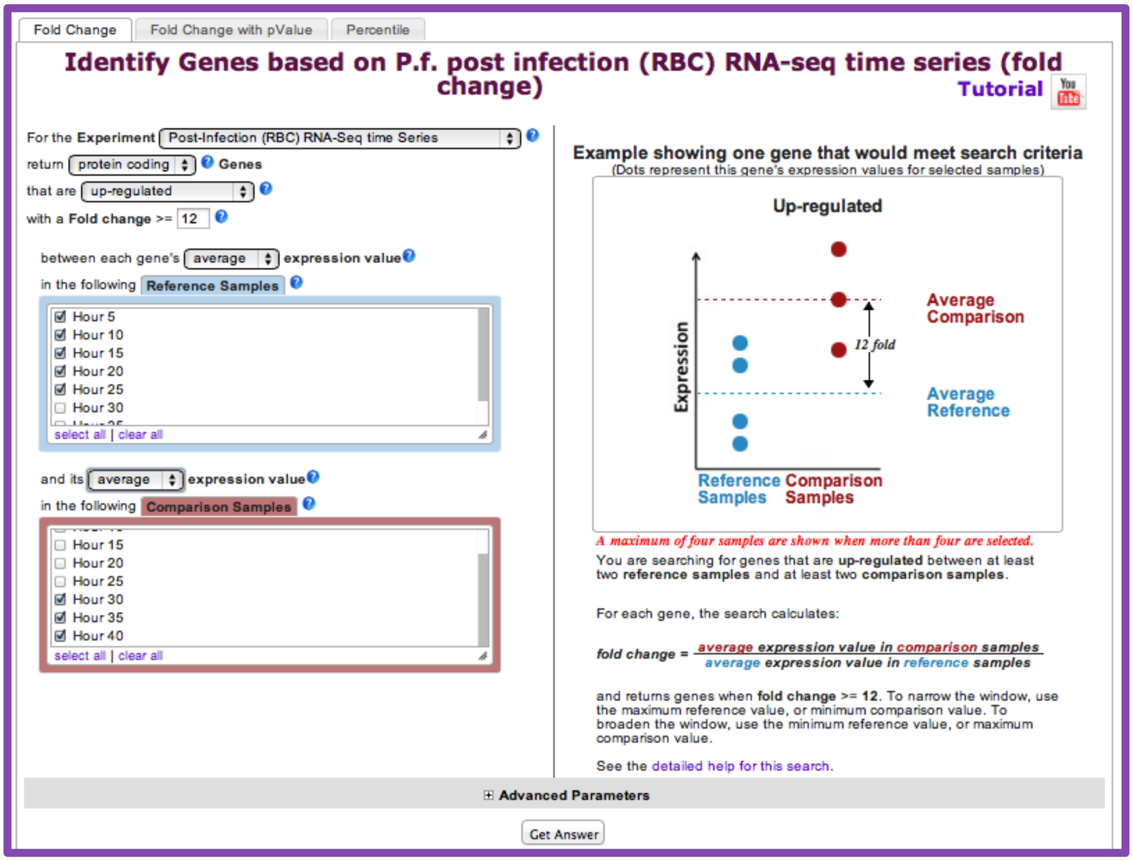
- **Fold Change>=:** the intensity of difference in expression needed before a gene is returned by the search. **Choose 12** but feel free to modify this.

- **Between each gene's AVERAGE expression value:** This parameter sets the operation applied to reference samples. Fold change is calculated as the ratio of two values (expression in reference)/(expression in comparison). When you choose multiple samples to serve as reference, we generate one number for the fold change calculation by using the minimum, maximum, or average. **Choose average**

- **Reference Sample**: the samples that will serve as the reference when comparing expression between samples. **choose 5, 10, 15, 20, 25**

- **And it's AVERAGE expression value:** This is the operation applied to comparison samples. see explanation above. **Choose average**

- **Comparison Sample**: the sample that you are comparing to the reference. In this case you are interested in genes that are up-regulated in later time points **choose 30, 35, 40**



Identify Genes based on P.f. post infection (RBC) RNA-seq time series (fold change)

For the **Experiment** Post-Infection (RBC) RNA-Seq time Series

return protein coding **Genes**

that are up-regulated

with a **Fold change** >= 12

between each gene's average **expression value**

in the following **Reference Samples**

☑ Hour 5
☑ Hour 10
☑ Hour 15
☑ Hour 20
☑ Hour 25
☐ Hour 30

select all | clear all

and its average **expression value**

in the following **Comparison Samples**

☐ Hour 15
☐ Hour 20
☐ Hour 25
☑ Hour 30
☑ Hour 35
☑ Hour 40

select all | clear all

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:

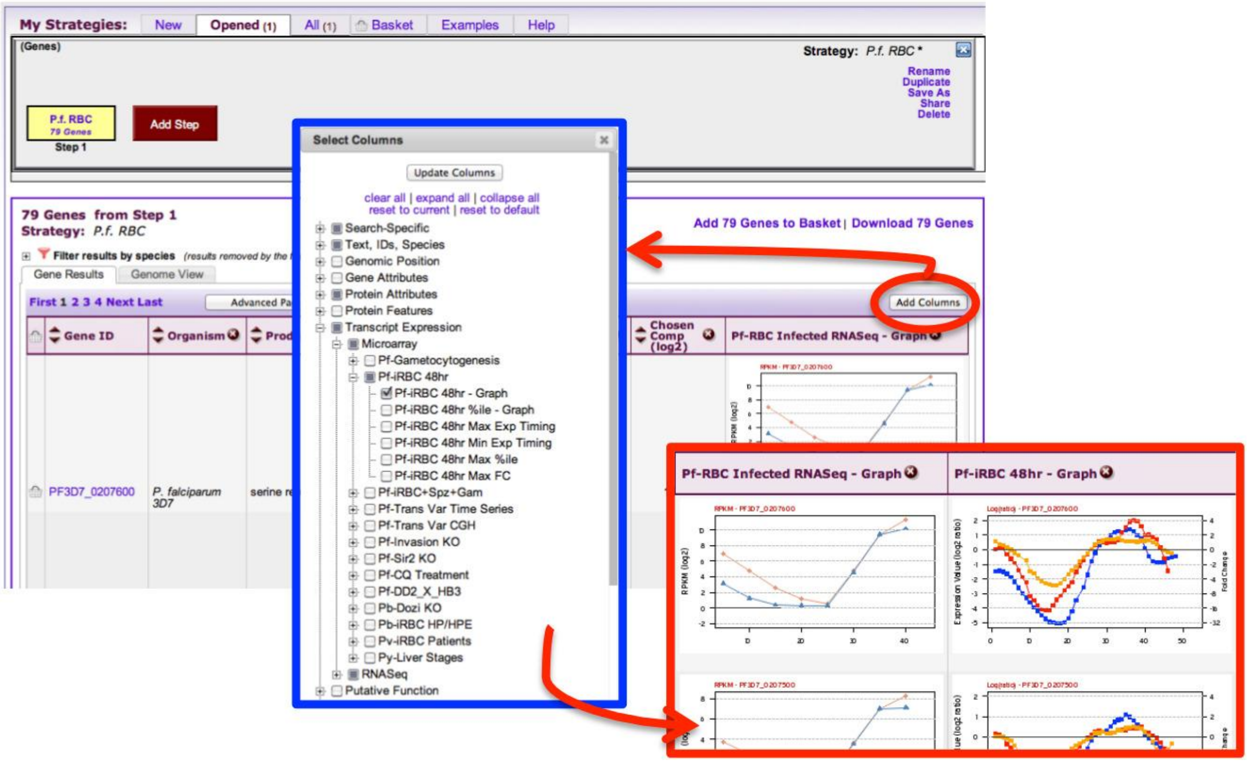
$$\textit{fold change} = \frac{\textit{average expression value in comparison samples}}{\textit{average expression value in reference samples}}$$

and returns genes when **fold change >= 12**. 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.

1b. For the genes returned by the search, how does the RNA-sequence data compare to microarray data?

- Hint: PlasmoDB contains data from a similar experiment that was analyzed by microarray instead of RNA sequencing. This experiment is called: Erythrocytic expression time series (3D7, DD2, HB3) (Bozdech et al. and Linas et al.). To directly compare the data for genes returned by the RNA seq search that you just ran, add the column called “Pf-iRBC 48hr - Graph”.

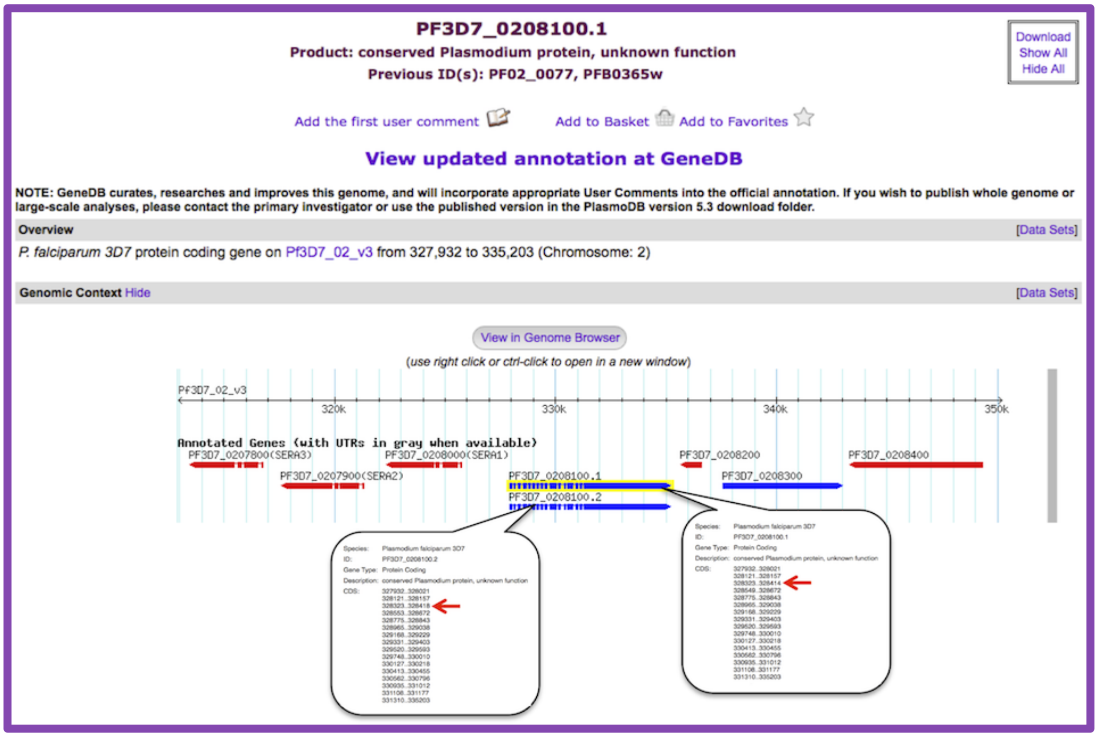


1c. How many genes in this result have 16 exons?

- Hint: add a column for number of exons. To help you find the genes with 16 exons, you can sort the columns using the arrows that precede the column heading. Also, clicking the histogram icon in the column heading will provides options for viewing the column data as a table or histogram.

- There are three gene IDs with 16 exons each. Two have similar gene IDs. What does this mean?

1d. Click on one of the two similar gene IDs from above. Look at the gene page. Take note of the Gene ID. Mouse over the gene models in the genomic context view and explore the popup. What information does it contain? Note that the CDS section includes exon coordinates. Compare the coordinates for the two alternative splice variants of this gene - can you identify the difference (it is very subtle)?



PF3D7_0208100.1

Product: conserved Plasmodium protein, unknown function

Previous ID(s): PF02_0077, PFB0365w

View updated annotation at GeneDB

NOTE: GeneDB curates, researches and improves this genome, and will incorporate appropriate User Comments into the official annotation. If you wish to publish whole genome or large-scale analyses, please contact the primary investigator or use the published version in the PlasmoDB version 5.3 download folder.

Overview [Data Sets]

P. falciparum 3D7 protein coding gene on Pf3D7_02_v3 from 327,932 to 335,203 (Chromosome: 2)

Genomic Context Hide [Data Sets]

(*use right click or ctrl-click to open in a new window*)

Species:	Plasmodium falciparum 3D7
ID:	PF3D7_0208100.2
Gene Type:	Protein Coding
Description:	conserved Plasmodium protein, unknown function
CDS:	327932..328021 328121..328157 328323..328418 328553..328672 328775..328843 328965..329038 329168..329229 329331..329403 329520..329593 329748..330010 330127..330218 330413..330455 330562..330796 330935..331012 331108..331177 331310..335203

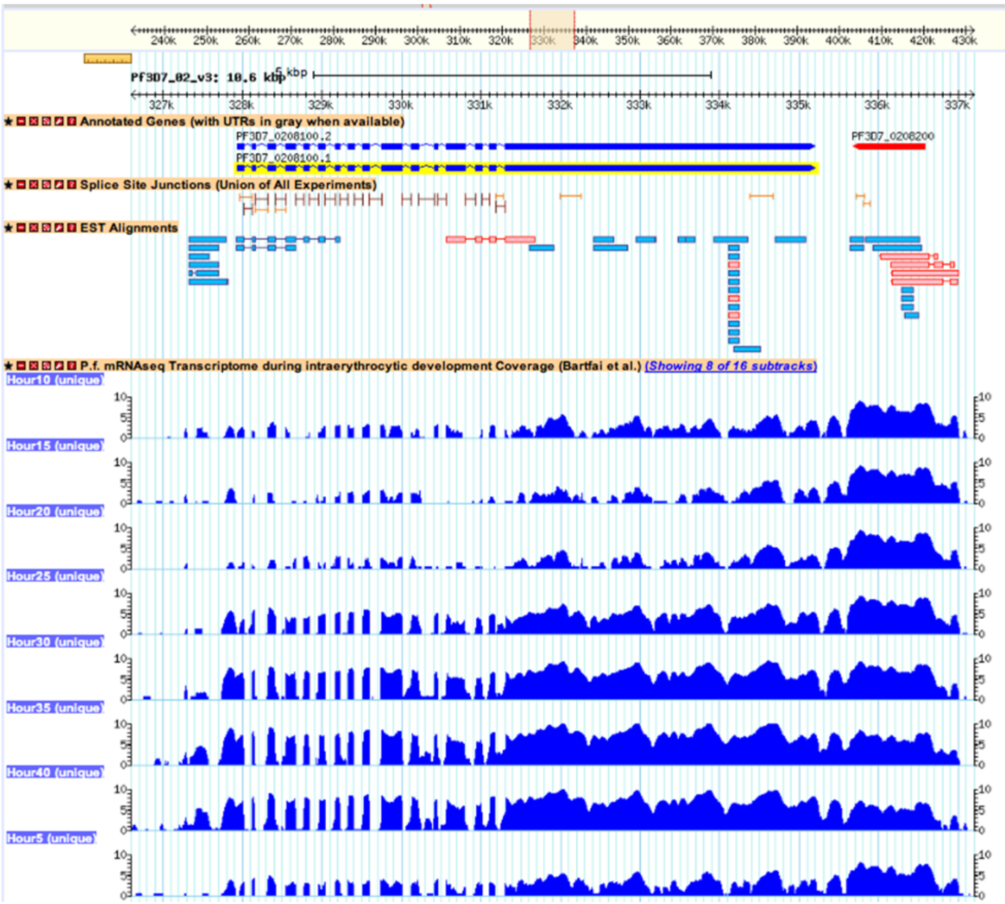
Species:	Plasmodium falciparum 3D7
ID:	PF3D7_0208100.1
Gene Type:	Protein Coding
Description:	conserved Plasmodium protein, unknown function
CDS:	327932..328021 328121..328157 328323..328414 328549..328672 328775..328843 328965..329038 329168..329229 329331..329403 329520..329593 329748..330010 330127..330218 330413..330455 330562..330796 330935..331012 331108..331177 331310..335203

1e. View this gene in the genome browser and load the RNA-seq tracks for this experiment. The track is named: “Transcriptome during intraerythrocytic development mRNAseq Coverage aligned to P falciparum 3D7 (Bartfai et al.) (log plot).

- Do these tracks match the differential expression results you got above? Is this gene differentially regulated between the early time points and the late ones?

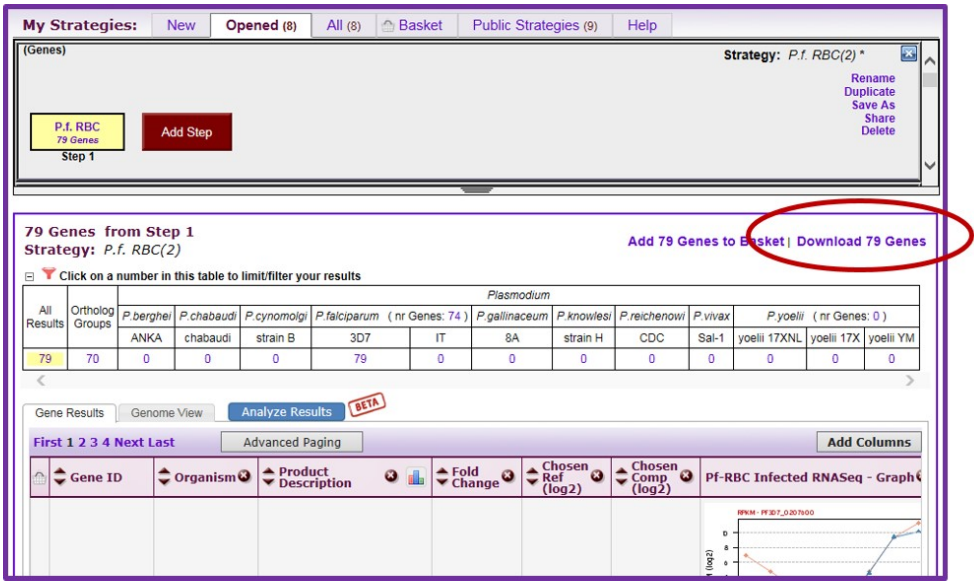
- Do you agree with the alternative splice call? Are there other possible splice variants? (*Hint*: turn on the track called “Splice Site Junctions (Union of All Experiments)”).

1f. What other data type can you load to help in looking at gene structure? (*Hint*: Look in the transcript expression section of the gbrowse tracks... how about ESTs?).


240k 250k 260k 270k 280k 290k 300k 310k 320k 330k 340k 350k 360k 370k 380k 390k 400k 410k 420k 430k
Pf3D7_02_v3: 10.6 kbp
5 kbp
327k 328k 329k 330k 331k 332k 333k 334k 335k 336k 337k
Annotated Genes (with UTRs in gray when available)
PF3D7_0208100.2
PF3D7_0208100.1
PF3D7_0208200
Splice Site Junctions (Union of All Experiments)
EST Alignments
P.f. mRNAseq Transcriptome during intraerythrocytic development Coverage (Bartfai et al.) (Showing 8 of 16 subtracks)
Hour10 (unique)
Hour15 (unique)
Hour20 (unique)
Hour25 (unique)
Hour30 (unique)
Hour35 (unique)
Hour40 (unique)
Hour5 (unique)
10
5
0

1g. You decide that you would like to present this data at your lab meeting and possibly use the data as supporting evidence in a grant proposal. The data you want to present is much like the graphs on individual gene pages but you would like to make a composite graph showing the profiles of several genes from your list.

- Download to your computer the list of genes returned by the search. Include the following information in your downloaded file: Gene ID, Product Description, Genomic Location, Annotated 5' UTR length, Annotated 3' UTR length.



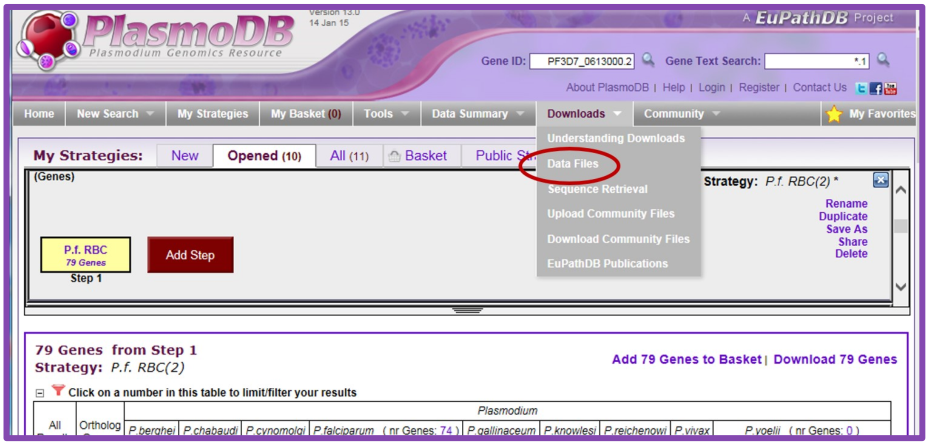
79 Genes from Step 1
Strategy: P.f. RBC(2)

Add 79 Genes to Basket | Download 79 Genes

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

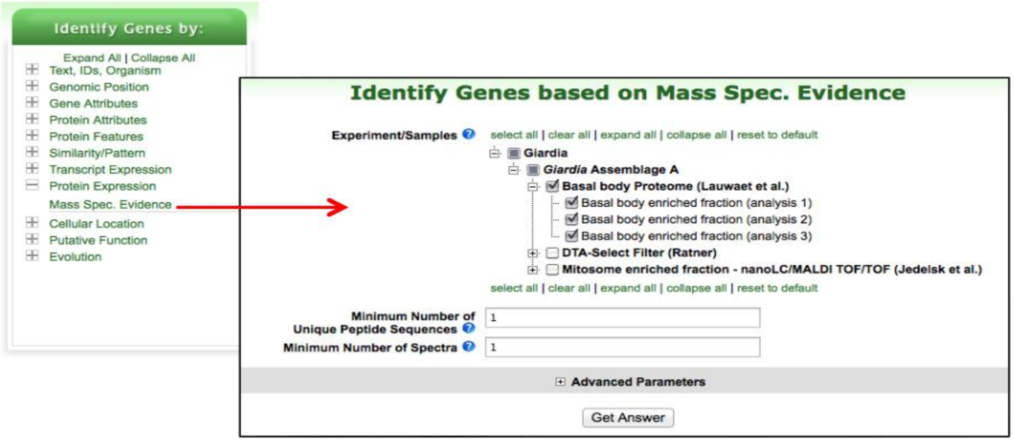
All Results	Ortholog Groups	Plasmodium												
		P.berghei	P.chabaudi	P.cynomolgi	P.falciparum (nr Genes: 74)		P.gallinaceum	P.knowlesi	P.reichenowi	P.vivax	P.yoelii (nr Genes: 0)		
		ANKA	chabaudi	strain B	3D7	IT	8A	strain H	CDC	Sal-1	yoelii 17XNL	yoelii 17X	yoelii YM
79	70	0	0	0	79	0	0	0	0	0	0	0	0

- Download to your computer the data file containing expression values (FPKM) for the RNA the sense strand uniquely aligned reads.



2. **Find all Giardia genes with evidence of basal body expression based on mass spec/proteomics data.**
   For this exercise use http://giardiadb.org

   The proteomics search is available under the heading “Protein Expression” in the “Identify gene by” section.



Identify Genes by:

Expand All | Collapse All

- ⊞ Text, IDs, Organism
- ⊞ Genomic Position
- ⊞ Gene Attributes
- ⊞ Protein Attributes
- ⊞ Protein Features
- ⊞ Similarity/Pattern
- ⊞ Transcript Expression
- ⊟ Protein Expression
  - Mass Spec. Evidence
- ⊞ Cellular Location
- ⊞ Putative Function
- ⊞ Evolution

Identify Genes based on Mass Spec. Evidence

Experiment/Samples select all | clear all | expand all | collapse all | reset to default

- ⊟ ▣ Giardia
  - ⊟ ▣ *Giardia* Assemblage A
    - ⊟ ☑ Basal body Proteome (Lauwaet et al.)
      - ☑ Basal body enriched fraction (analysis 1)
      - ☑ Basal body enriched fraction (analysis 2)
      - ☑ Basal body enriched fraction (analysis 3)
    - ⊞ ☐ DTA-Select Filter (Ratner)
    - ⊞ ☐ Mitosome enriched fraction - nanoLC/MALDI TOF/TOF (Jedelsk et al.)

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

Minimum Number of Unique Peptide Sequences 1

Minimum Number of Spectra 1

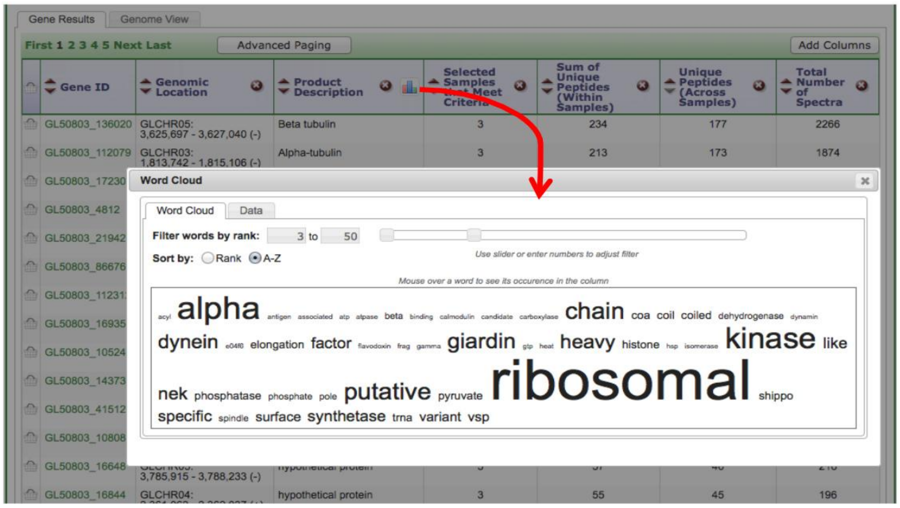
⊞ **Advanced Parameters**

Get Answer

2a. How many genes did you identify?

2b. What kinds of genes are in your result list?

Hint 1: analyze the Product Description column. The **word cloud tool** counts the number of times a word appears in the column and then draws a word cloud in which the size of the word reflects how many times the word appears in the product description column. Click on the little graphic icon next to the column called “Product description”.



Gene ID	Genomic Location	Product Description	Selected Samples that Meet Criteria	Sum of Unique Peptides (Within Samples)	Unique Peptides (Across Samples)	Total Number of Spectra
GL50803_136020	GLCHR05: 3,625,697 - 3,627,040 (-)	Beta tubulin	3	234	177	2266
GL50803_112079	GLCHR03: 1,813,742 - 1,815,106 (-)	Alpha-tubulin	3	213	173	1874
GL50803_17230						
GL50803_4812						
GL50803_21942						
GL50803_86676						
GL50803_11231						
GL50803_16935						
GL50803_10524						
GL50803_14373						
GL50803_41512						
GL50803_10808						
GL50803_16648	3,785,915 - 3,788,233 (-)	hypothetical protein				
GL50803_16844	GLCHR04:	hypothetical protein	3	55	45	196

Word Cloud

Word Cloud | Data

Filter words by rank: 3 to 50

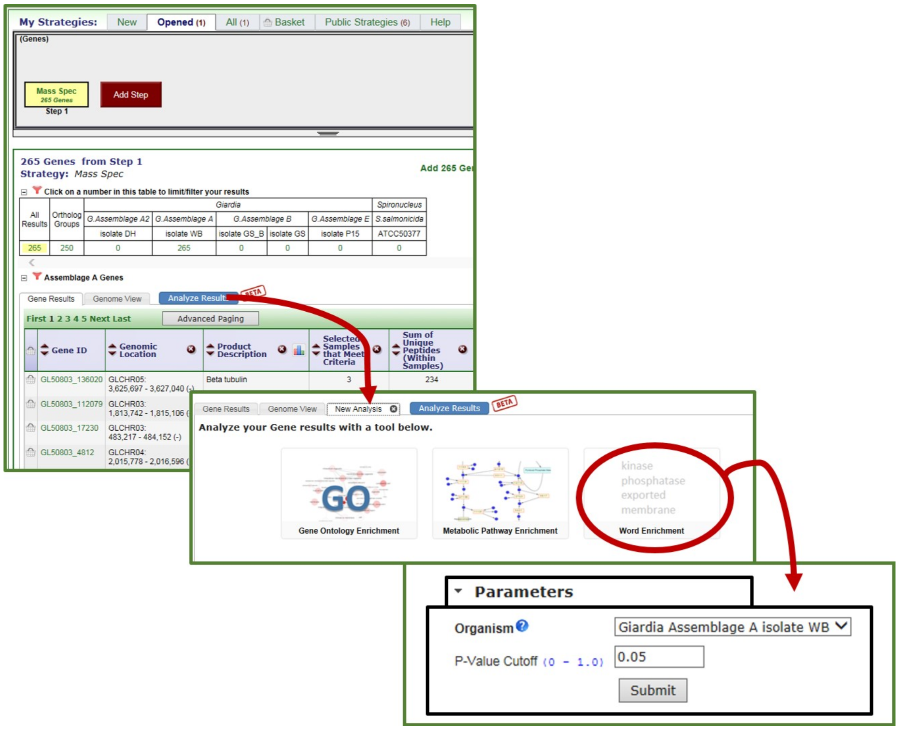
Use slider or enter numbers to adjust filter

Sort by: ○ Rank ● A-Z

Mouse over a word to see its occurence in the column

acyl alpha antigen associated atp atpase beta binding calmodulin candidate carboxylase chain coa coil coiled dehydrogenase dynamin dynein e04f6 elongation factor flavodoxin frag gamma giardin gtp heat heavy histone hsp isomerase kinase like nek phosphatase phosphate pole putative pyruvate ribosomal shippo specific spindle surface synthetase trna variant vsp

Hint 2: apply a statistical analysis to the Product Description column. The Word Enrichment Analysis Tool (Click the Blue Analyze Results button) considers the words in the Product Description column and applies the Fischer's Exact test to compare your gene result to the product descriptions of the entire genome.



265 Genes from Step 1
Strategy: *Mass Spec*

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

All Results	Ortholog Groups	*Giardia*					*Spironucleus*
		G.Assemblage A2	*G.Assemblage A*	*G.Assemblage B*		*G.Assemblage E*	*S.salmonicida*
		isolate DH	isolate WB	isolate GS_B	isolate GS	isolate P15	ATCC50377
265	250	0	265	0	0	0	0

Assemblage A Genes

Gene ID	Genomic Location	Product Description	Selected Samples that Meet Criteria	Sum of Unique Peptides (Within Samples)
GL50803_136020	GLCHR05: 3,625,697 - 3,627,040 (-)	Beta tubulin	3	234
GL50803_112079	GLCHR03: 1,813,742 - 1,815,106 (			
GL50803_17230	GLCHR03: 483,217 - 484,152 (-)			
GL50803_4812	GLCHR04: 2,015,778 - 2,016,596 (			

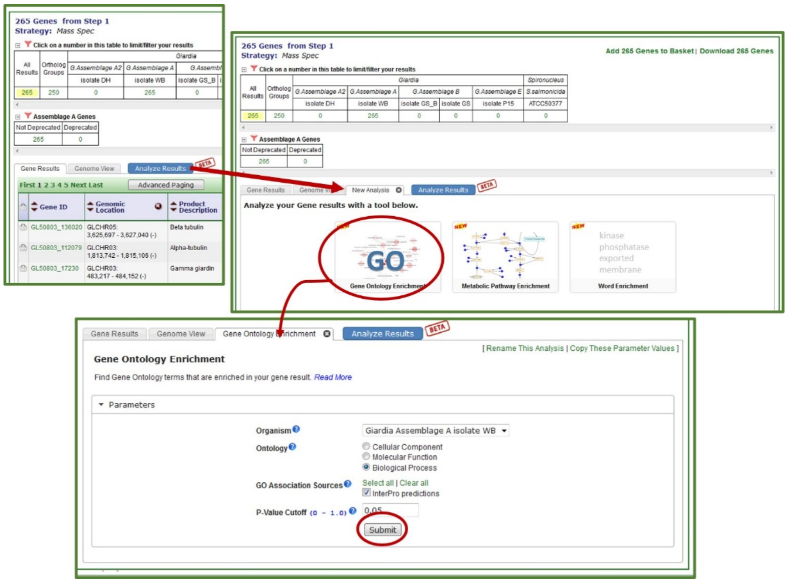
Analyze your Gene results with a tool below.

Parameters

Organism: Giardia Assemblage A isolate WB

P-Value Cutoff (0 - 1.0): 0.05

Hint 3: analyze the Gene Ontology terms assigned to the genes in your result list. **The Result Analysis/Enrichment tool** applies the Fischer's Exact test to compare your gene result to the entire genome. Use the Gene Ontology Enrichment to find Biological Process ontology terms that are enriched in your gene result.


265 Genes from Step 1
Strategy: *Mass Spec*

Add 265 Genes to Basket | Download 265 Genes

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

All Results	Ortholog Groups	*Giardia*					*Spironucleus*
		G.Assemblage A2	*G.Assemblage A*	*G.Assemblage B*		*G.Assemblage E*	*S.salmonicida*
		isolate DH	isolate WB	isolate GS_B	isolate GS	isolate P15	ATCC50377
265	250	0	265	0	0	0	0

Assemblage A Genes

Not Deprecated	Deprecated
265	0

Gene Results
Genome View
Analyze Results
BETA
First 1 2 3 4 5 Next Last
Advanced Paging

Gene ID	Genomic Location	Product Description
GL50803_136020	GLCHR05: 3,625,697 - 3,627,040 (-)	Beta tubulin
GL50803_112079	GLCHR03: 1,813,742 - 1,815,106 (-)	Alpha-tubulin
GL50803_17230	GLCHR03: 483,217 - 484,152 (-)	Gamma giardin

Analyze your Gene results with a tool below.

Gene Ontology Enrichment

Find Gene Ontology terms that are enriched in your gene result. *Read More*

Parameters

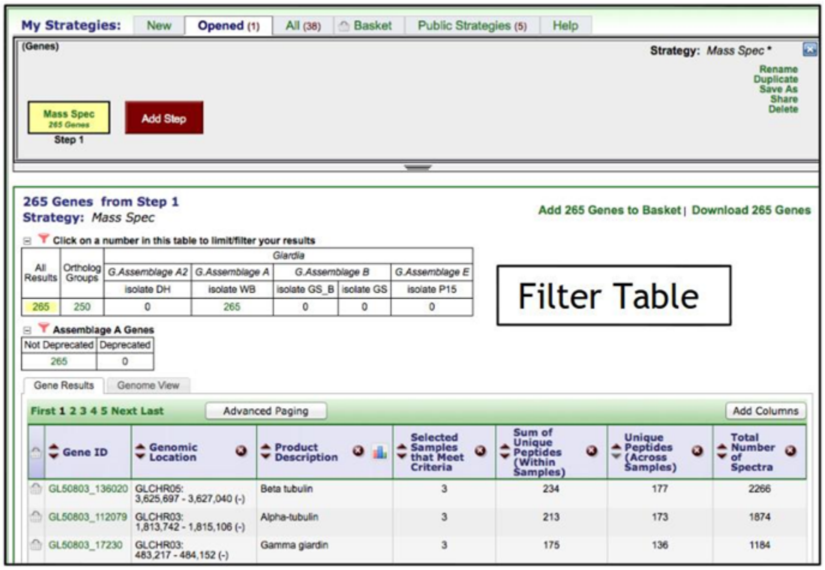
Organism: Giardia Assemblage A isolate WB

Ontology:
○ Cellular Component
○ Molecular Function
● Biological Process

GO Association Sources: Select all | Clear all
☑ InterPro predictions

P-Value Cutoff (0 - 1.0): 0.05

2c. Examine the filter table below the strategy section. Why is the number in the ortholog groups column less than the number in the Assemblage A column? Why do the other cells have a zero?



265 Genes from Step 1
Strategy: *Mass Spec*

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

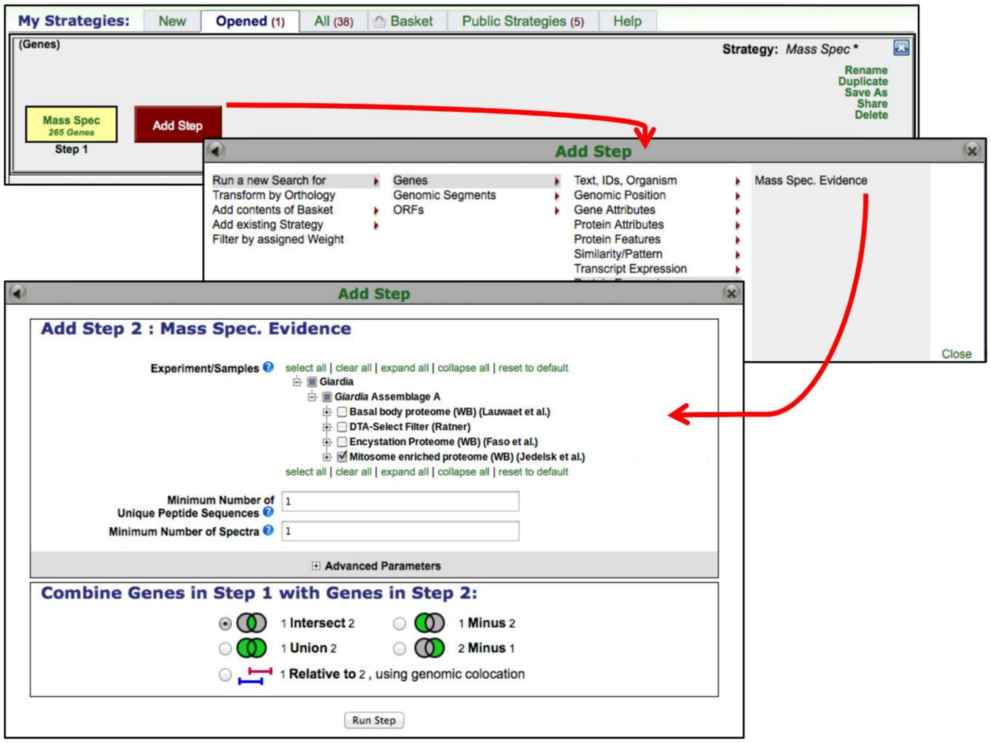
All Results	Ortholog Groups	*Giardia*				
		G.Assemblage A2	*G.Assemblage A*	*G.Assemblage B*		*G.Assemblage E*
		isolate DH	isolate WB	isolate GS_B	isolate GS	isolate P15
265	250	0	265	0	0	0

Assemblage A Genes

Not Deprecated	Deprecated
265	0

Gene ID	Genomic Location	Product Description	Selected Samples that Meet Criteria	Sum of Unique Peptides (Within Samples)	Unique Peptides (Across Samples)	Total Number of Spectra
GL50803_136020	GLCHR05: 3,625,697 - 3,627,040 (-)	Beta tubulin	3	234	177	2266
GL50803_112079	GLCHR03: 1,813,742 - 1,815,106 (-)	Alpha-tubulin	3	213	173	1874
GL50803_17230	GLCHR03: 483,217 - 484,152 (-)	Gamma giardin	3	175	136	1184

2d. How many genes with mass spec data from the basal proteome experiment also have mass spec evidence from the “Mitosome enriched proteome (WB) (Jedelsk et al.)”?



Add Step 2 : Mass Spec. Evidence

Experiment/Samples: select all | clear all | expand all | collapse all | reset to default

- Giardia
  - *Giardia* Assemblage A
    - ☐ Basal body proteome (WB) (Lauwaet et al.)
    - ☐ DTA-Select Filter (Ratner)
    - ☐ Encystation Proteome (WB) (Faso et al.)
    - ☑ Mitosome enriched proteome (WB) (Jedelsk et al.)

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

Minimum Number of Unique Peptide Sequences: 1

Minimum Number of Spectra: 1

Advanced Parameters

Combine Genes in Step 1 with Genes in Step 2:

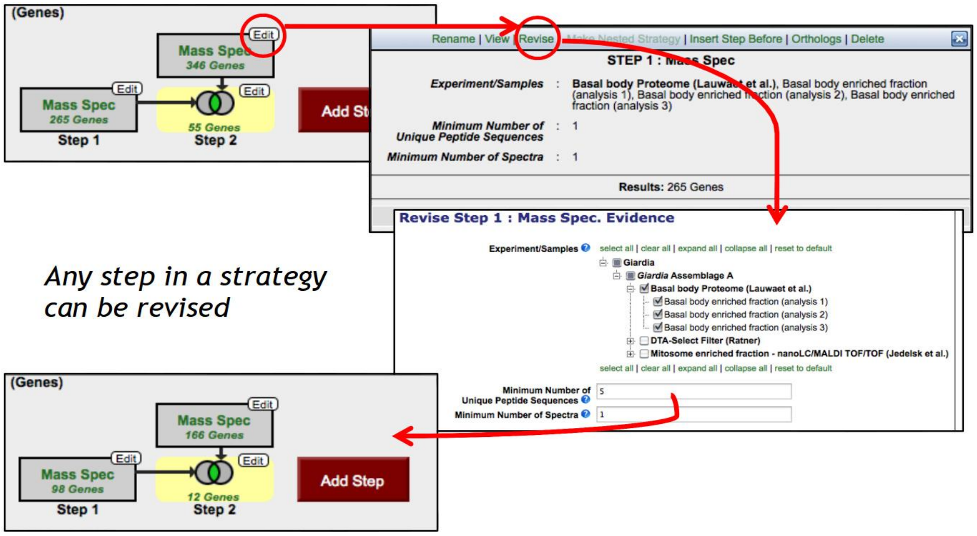
● 1 **Intersect** 2 ○ 1 **Minus** 2

○ 1 **Union** 2 ○ 2 **Minus** 1

○ 1 **Relative to** 2 , using genomic colocation

Run Step

2e. The default parameters of the mass spec search are set to identify any gene with at least 1 mapped peptide identified. How will your results change if you revise the two steps in your search strategy to only return genes with at least 5 peptides identified?



STEP 1 : Mass Spec

Experiment/Samples : **Basal body Proteome (Lauwaet et al.)**, Basal body enriched fraction (analysis 1), Basal body enriched fraction (analysis 2), Basal body enriched fraction (analysis 3)

Minimum Number of Unique Peptide Sequences : 1

Minimum Number of Spectra : 1

Results: 265 Genes

Revise Step 1 : Mass Spec. Evidence

Experiment/Samples select all | clear all | expand all | collapse all | reset to default

- ▣ Giardia
  - ▣ *Giardia* Assemblage A
    - ☑ Basal body Proteome (Lauwaet et al.)
      - ☑ Basal body enriched fraction (analysis 1)
      - ☑ Basal body enriched fraction (analysis 2)
      - ☑ Basal body enriched fraction (analysis 3)
    - □ DTA-Select Filter (Ratner)
    - □ Mitosome enriched fraction - nanoLC/MALDI TOF/TOF (Jedelsk et al.)

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

Minimum Number of Unique Peptide Sequences: 5

Minimum Number of Spectra: 1

Any step in a strategy can be revised