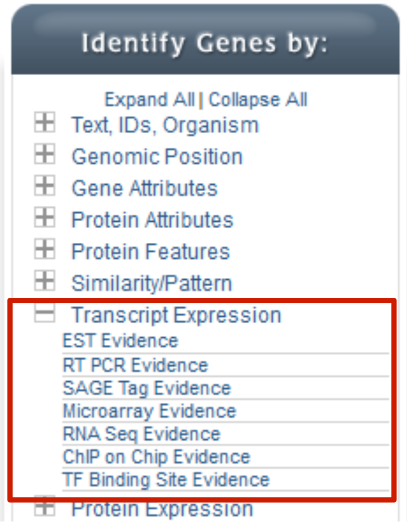
Exploring Transcriptomics Data
Exercise 13

13.1 Evidence of expression at the transcriptional level.
Note: For this exercise use http://www.eupathdb.org

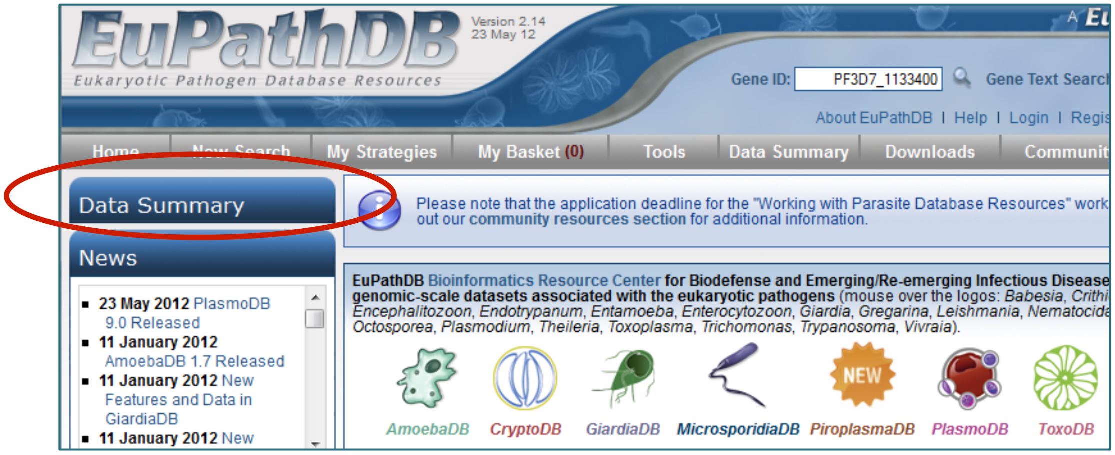
a. What kind of data types can be used to provide evidence of transcriptional activity? Hint: click on “Transcript Expression” to expand the list of possible searches.



b. Explore organisms that have microarray data. What organisms have expressed sequence tag (EST), RNA sequence, ChIP-chip or SAGE tag data?

c. What does RNA-seq data tell you that microarray data cannot? What does ChIP-chip data tell you about a gene?

d. Go to the Data Summary Section, can you find the same information there? Hint: data summary table in on the left side of the home page.



Please note that the application deadline for the "Working with Parasite Database Resources" work
out our community resources section for additional information.

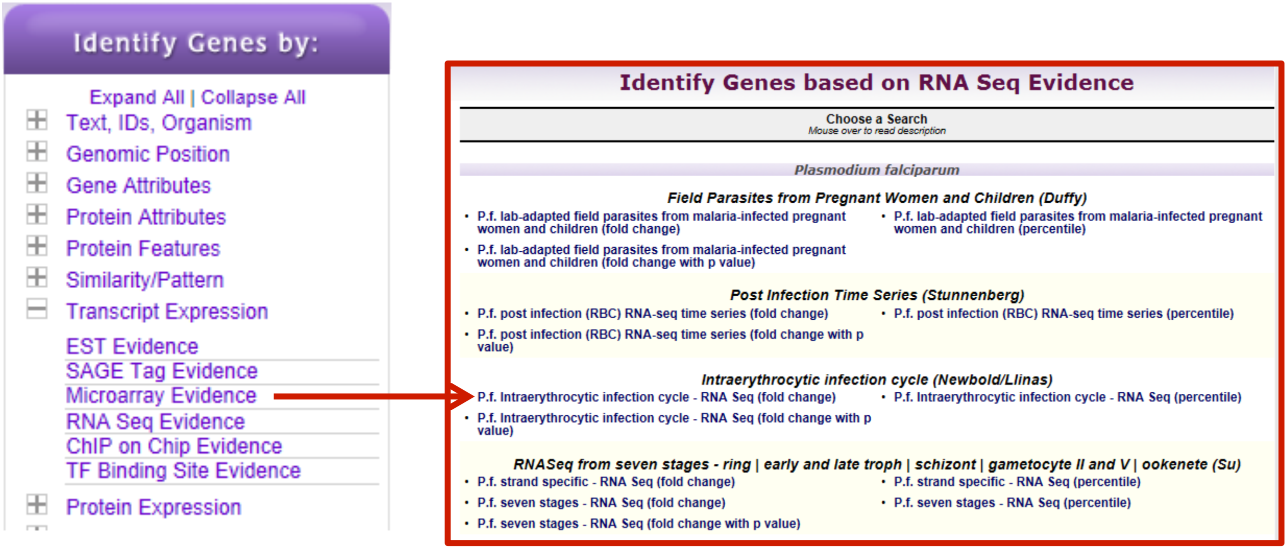
- **23 May 2012** PlasmoDB 9.0 Released
- **11 January 2012** AmoebaDB 1.7 Released
- **11 January 2012** New Features and Data in GiardiaDB
- **11 January 2012** New

EuPathDB Bioinformatics Resource Center **for Biodefense and Emerging/Re-emerging Infectious Disease**
genomic-scale datasets associated with the eukaryotic pathogens (mouse over the logos: *Babesia, Crithi*
Encephalitozoon, Endotrypanum, Entamoeba, Enterocytozoon, Giardia, Gregarina, Leishmania, Nematocida
Octosporea, Plasmodium, Theileria, Toxoplasma, Trichomonas, Trypanosoma, Vivraia).

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

a. Find all genes in *P. falciparum* that are upregulated based on RNA-seq data at late time points (30, 35 and 40-hours) compared to early time points in this experiment (1, 10, 15, 20, 25 hrs).

hint: for this exercise use “P.f. post infection (RBC) RNA-seq time series (fold change)”).



Identify Genes based on RNA Seq Evidence

Choose a Search
Mouse over to read description

Plasmodium falciparum

Field Parasites from Pregnant Women and Children (Duffy)

- P.f. lab-adapted field parasites from malaria-infected pregnant women and children (fold change)
- P.f. lab-adapted field parasites from malaria-infected pregnant women and children (fold change with p value)
- P.f. lab-adapted field parasites from malaria-infected pregnant women and children (percentile)

Post Infection Time Series (Stunnenberg)

- P.f. post infection (RBC) RNA-seq time series (fold change)
- P.f. post infection (RBC) RNA-seq time series (fold change with p value)
- P.f. post infection (RBC) RNA-seq time series (percentile)

Intraerythrocytic infection cycle (Newbold/Llinas)

- P.f. Intraerythrocytic infection cycle - RNA Seq (fold change)
- P.f. Intraerythrocytic infection cycle - RNA Seq (fold change with p value)
- P.f. Intraerythrocytic infection cycle - RNA Seq (percentile)

RNASeq from seven stages - ring | early and late troph | schizont | gametocyte II and V | ookenete (Su)

- P.f. strand specific - RNA Seq (fold change)
- P.f. seven stages - RNA Seq (fold change)
- P.f. seven stages - RNA Seq (fold change with p value)
- P.f. strand specific - RNA Seq (percentile)
- P.f. seven stages - RNA Seq (percentile)

hint: there are several parameters to manipulate in this search:

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

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

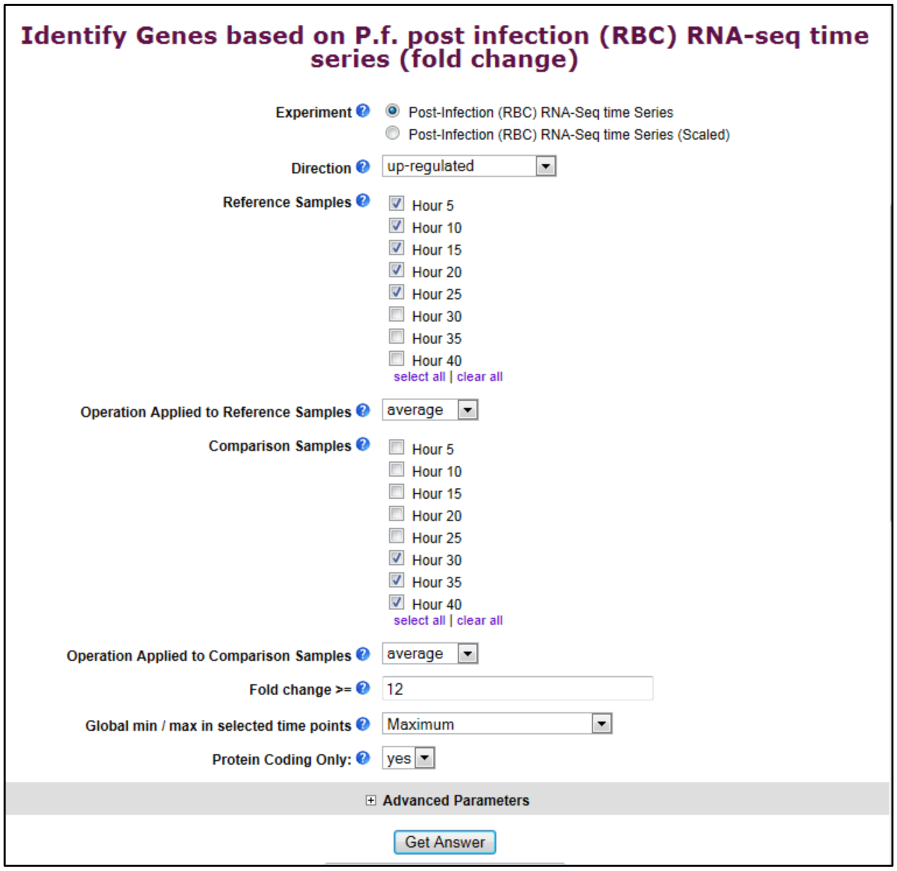
Operation Applied to Reference Samples: fold change is calculated as the ratio of two values (expression in reference)/(expression in comparison). When you choose multiple samples to serve as reference, we generate one number for the fold change calculation by using the minimum, maximum, or average. **Choose average**

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**

Operation Applied to Comparison Samples: see explanation above. **Choose average**

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.

Global min / max in selected time points: Choose whether the selected samples must contain the global minimum or maximum expression value of all samples. **Choose maximum** [Choosing minimum: if searching for up-regulated genes then the Reference values selected must be the global minimum. If searching for down-regulated genes, the comparison values selected must be the global minimum. Choosing maximum: if searching for up-regulated genes, then the Comparison values selected must be the global maximum. If searching for down-regulated genes, the reference values selected must be the global maximum.]



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

Experiment ● Post-Infection (RBC) RNA-Seq time Series
○ Post-Infection (RBC) RNA-Seq time Series (Scaled)

Direction: up-regulated

Reference Samples
☑ Hour 5
☑ Hour 10
☑ Hour 15
☑ Hour 20
☑ Hour 25
□ Hour 30
□ Hour 35
□ Hour 40
select all | clear all

Operation Applied to Reference Samples: average

Comparison Samples
□ Hour 5
□ Hour 10
□ Hour 15
□ Hour 20
□ Hour 25
☑ Hour 30
☑ Hour 35
☑ Hour 40
select all | clear all

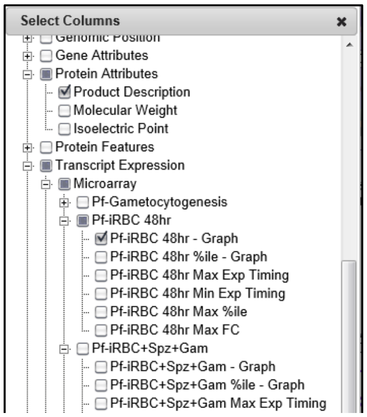
Operation Applied to Comparison Samples: average

Fold change >= 12

Global min / max in selected time points: Maximum

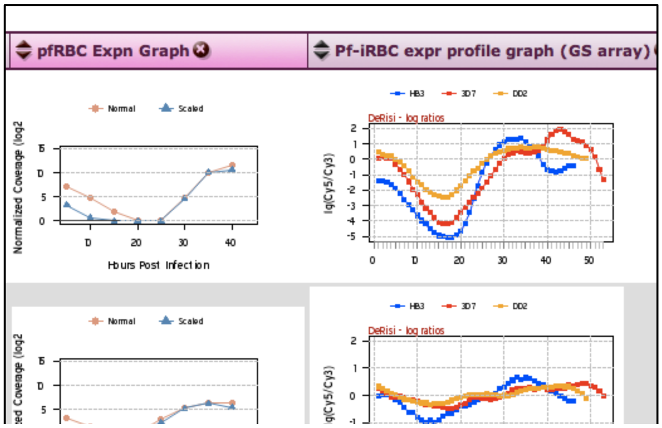
Protein Coding Only: yes

b. For the genes returned by the search, how does the RNA-sequence data compare to microarray data?
(hint: add the column called “Pf-iRBC expr profile graph (GS array)” and compare the RNA-seq to the microarray graphs).

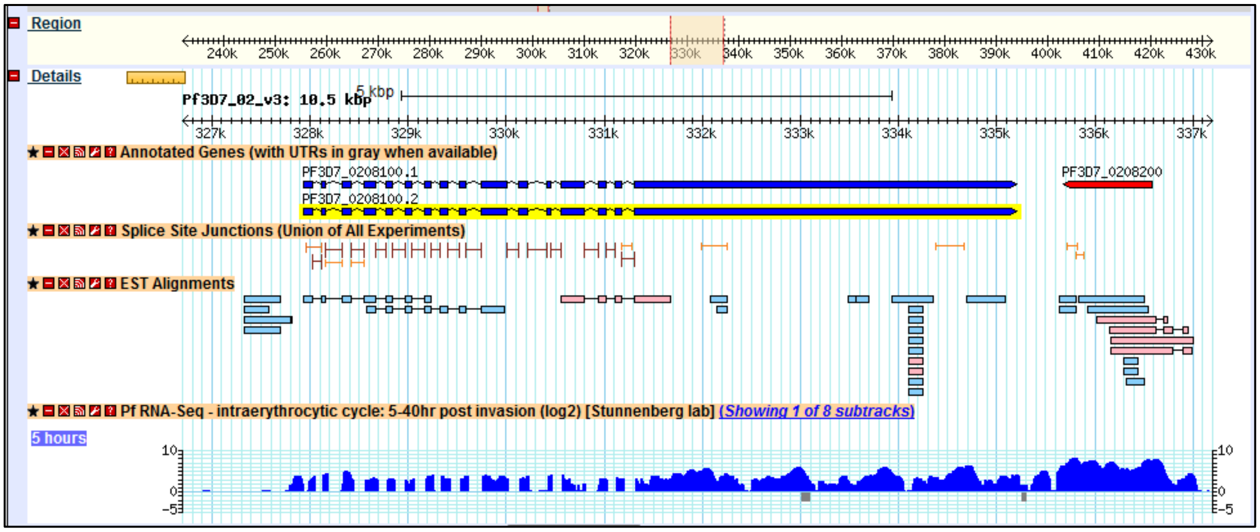


Select Columns ×

□ Genomic Position
⊞ □ Gene Attributes
⊟ ▪ Protein Attributes
☑ Product Description
□ Molecular Weight
□ Isoelectric Point
⊞ □ Protein Features
⊟ ▪ Transcript Expression
⊟ ▪ Microarray
⊞ □ Pf-Gametocytogenesis
⊟ ▪ Pf-iRBC 48hr
☑ Pf-iRBC 48hr - Graph
□ Pf-iRBC 48hr %ile - Graph
□ Pf-iRBC 48hr Max Exp Timing
□ Pf-iRBC 48hr Min Exp Timing
□ Pf-iRBC 48hr Max %ile
□ Pf-iRBC 48hr Max FC
⊟ □ Pf-iRBC+Spz+Gam
□ Pf-iRBC+Spz+Gam - Graph
□ Pf-iRBC+Spz+Gam %ile - Graph
□ Pf-iRBC+Spz+Gam Max Exp Timing



c. Which gene has 17 exons? (hint: use the columns)
d. Is this gene alternatively spliced? Look at the gene page. Take note of the Gene ID.
e. View this gene in the genome browser and load the RNA-seq tracks for this experiment “Pf RNA-Seq - intraerythrocytic cycle: 5-40hr post invasion (log2) [Stunnenberg lab]”. Do these tracks match the results you got above? (ie. is this gene differentially regulated between the early time points and the late ones?
f. Do you agree with the alternative splice call? Are there other possible splice variants? (hint: turn on the track called “Splice Site Junctions - Combined”).
g. 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).


Region

Details

Pf3D7_02_v3: 10.5 kbp

Annotated Genes (with UTRs in gray when available)

Splice Site Junctions (Union of All Experiments)

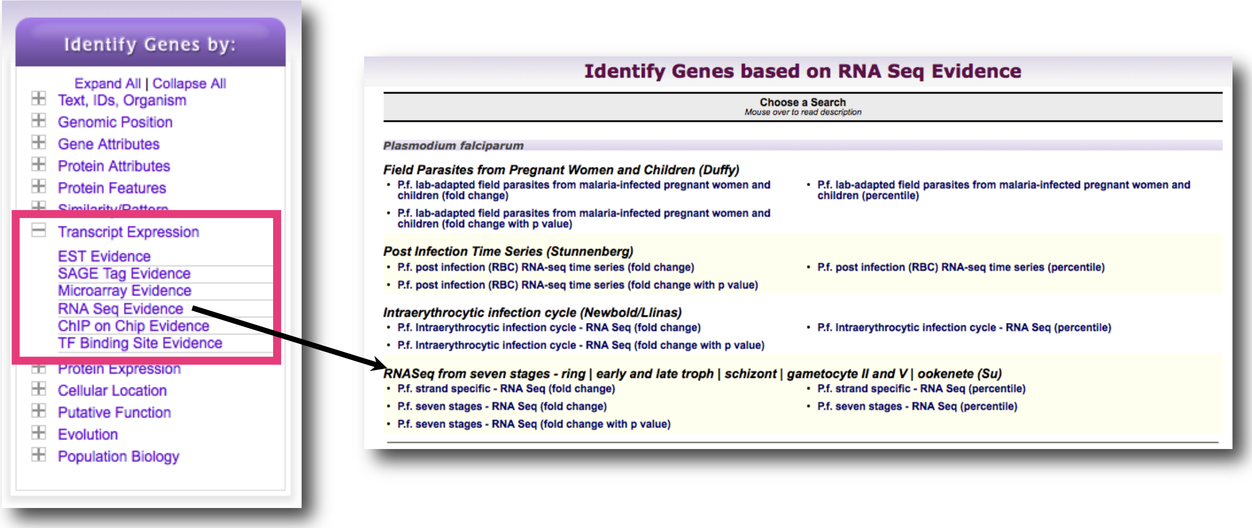
EST Alignments

Pf RNA-Seq - intraerythrocytic cycle: 5-40hr post invasion (log2) [Stunnenberg lab] (Showing 1 of 8 subtracks)

5 hours

13.3: Finding genes based on RNAseq evidence and inferring function of hypothetical genes. Note: Use http://plasmodb.org for this exercise.

1. Find all genes in *P. falciparum* that are upregulated at least 50-fold in ookinetes compared to other stages: P.f. seven stages - RNA Seq (fold change)



Identify Genes by:

Expand All | Collapse All

- ⊞ Text, IDs, Organism
- ⊞ Genomic Position
- ⊞ Gene Attributes
- ⊞ Protein Attributes
- ⊞ Protein Features
- ⊞ Similarity/Pattern
- ⊟ Transcript Expression
  - EST Evidence
  - SAGE Tag Evidence
  - Microarray Evidence
  - RNA Seq Evidence
  - ChIP on Chip Evidence
  - TF Binding Site Evidence
- ⊞ Protein Expression
- ⊞ Cellular Location
- ⊞ Putative Function
- ⊞ Evolution
- ⊞ Population Biology

Identify Genes based on RNA Seq Evidence

Choose a Search
Mouse over to read description

Plasmodium falciparum

Field Parasites from Pregnant Women and Children (Duffy)

- P.f. lab-adapted field parasites from malaria-infected pregnant women and children (fold change)
- P.f. lab-adapted field parasites from malaria-infected pregnant women and children (fold change with p value)
- P.f. lab-adapted field parasites from malaria-infected pregnant women and children (percentile)

Post Infection Time Series (Stunnenberg)

- P.f. post infection (RBC) RNA-seq time series (fold change)
- P.f. post infection (RBC) RNA-seq time series (fold change with p value)
- P.f. post infection (RBC) RNA-seq time series (percentile)

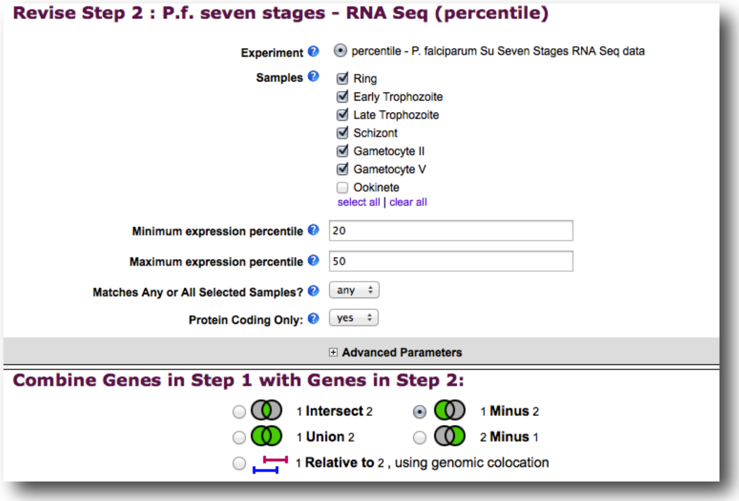
Intraerythrocytic infection cycle (Newbold/Llinas)

- P.f. Intraerythrocytic infection cycle - RNA Seq (fold change)
- P.f. Intraerythrocytic infection cycle - RNA Seq (fold change with p value)
- P.f. Intraerythrocytic infection cycle - RNA Seq (percentile)

RNASeq from seven stages - ring | early and late troph | schizont | gametocyte II and V | ookenete (Su)

- P.f. strand specific - RNA Seq (fold change)
- P.f. seven stages - RNA Seq (fold change)
- P.f. seven stages - RNA Seq (fold change with p value)
- P.f. strand specific - RNA Seq (percentile)
- P.f. seven stages - RNA Seq (percentile)

c. The above search will give you all genes that are upregulated by 50 fold in ookinetes compared to the other stages. However, this does not mean that these genes are not expressed in the other stages. How can you remove genes form the list that are likely not expressed in the other stages? (*hint: run a search for genes based on RNAseq evidence from the same experiment, but this time select the percentile search*): P.f. seven stages - RNA Seq (percentile)



Revise Step 2 : P.f. seven stages - RNA Seq (percentile)

Experiment: ● percentile - P. falciparum Su Seven Stages RNA Seq data

Samples:
- ☑ Ring
- ☑ Early Trophozoite
- ☑ Late Trophozoite
- ☑ Schizont
- ☑ Gametocyte II
- ☑ Gametocyte V
- □ Ookinete

select all | clear all

Minimum expression percentile: 20

Maximum expression percentile: 50

Matches Any or All Selected Samples?: any

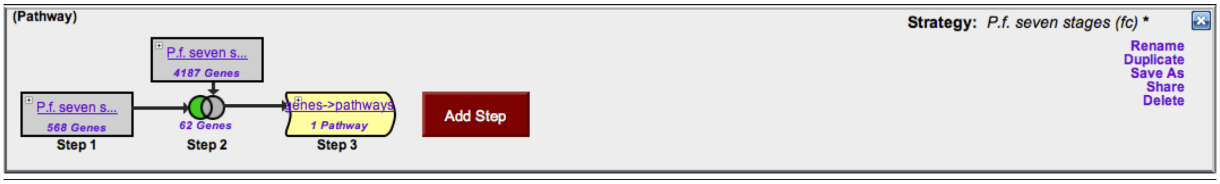
Protein Coding Only: yes

⊞ 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

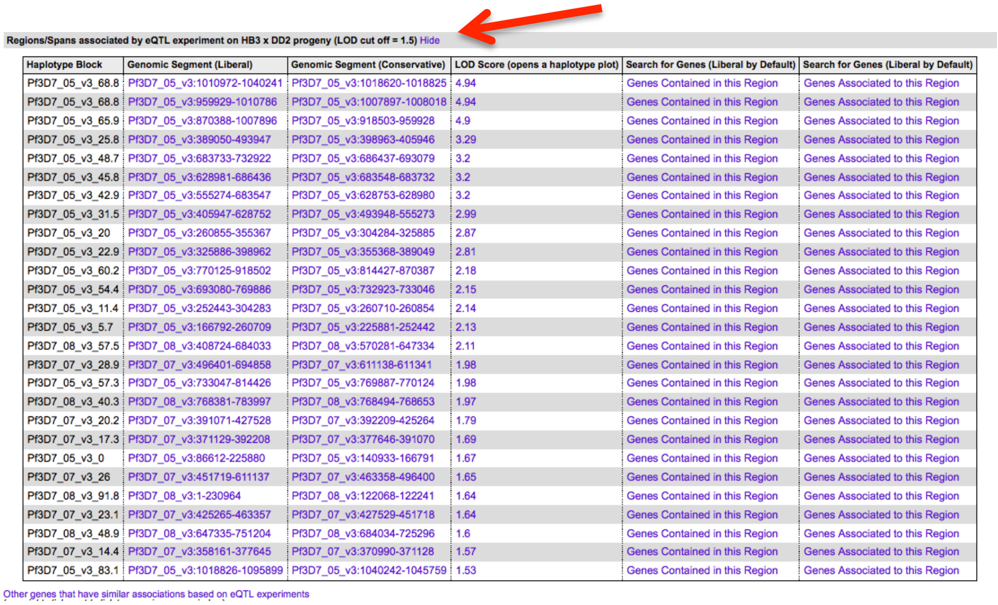
d. Which metabolic pathways are represented in this gene list? (*hint:* transform results to metabolic pathways.



13.3 Exploring Expression Quantitative Trait Locus (eQTL) data in PlasmoDB.

Genetic crosses were instrumental in implicating the PfCRT gene in chloroquine resistance. PlasmoDB contains expression quantitative trait locus data from Gonzales *et. al*. PLoS Biol 6(9): e238. The trait that was examined in this study was gene expression using microarray experiments.

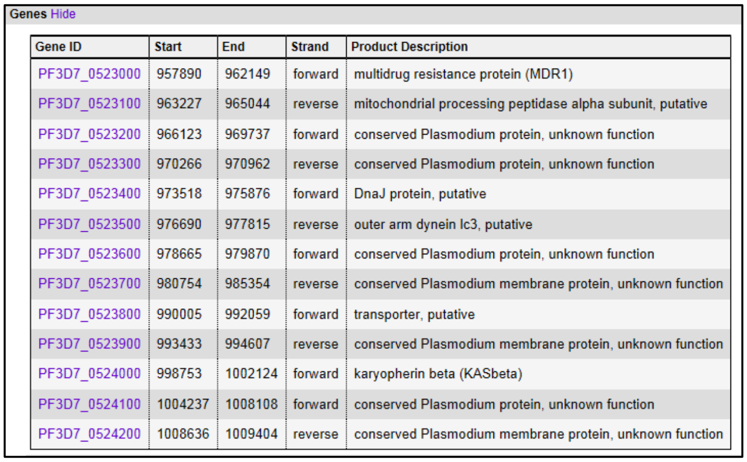
a. Go to the gene page for the gene with the ID PF3D7_0630200. Can you identify the genomic region (haplotype block) that is "most" associated with this gene, ie. has the highest LOD score? (Hint: examine the table called "Regions/Spans associated by eQTL experiment on HB3 x DD2 progeny" on the gene page.


Regions/Spans associated by eQTL experiment on HB3 x DD2 progeny (LOD cut off = 1.5) Hide

Haplotype Block	Genomic Segment (Liberal)	Genomic Segment (Conservative)	LOD Score (opens a haplotype plot)	Search for Genes (Liberal by Default)	Search for Genes (Liberal by Default)
Pf3D7_05_v3_68.8	Pf3D7_05_v3:1010972-1040241	Pf3D7_05_v3:1018620-1018825	4.94	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_68.8	Pf3D7_05_v3:959929-1010786	Pf3D7_05_v3:1007897-1008018	4.94	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_65.9	Pf3D7_05_v3:870388-1007896	Pf3D7_05_v3:918503-959928	4.9	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_25.8	Pf3D7_05_v3:389050-493947	Pf3D7_05_v3:398963-405946	3.29	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_48.7	Pf3D7_05_v3:683733-732922	Pf3D7_05_v3:686437-693079	3.2	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_45.8	Pf3D7_05_v3:628981-686436	Pf3D7_05_v3:683548-683732	3.2	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_42.9	Pf3D7_05_v3:555274-683547	Pf3D7_05_v3:628753-628980	3.2	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_31.5	Pf3D7_05_v3:405947-628752	Pf3D7_05_v3:493948-555273	2.99	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_20	Pf3D7_05_v3:260855-355367	Pf3D7_05_v3:304284-325885	2.87	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_22.9	Pf3D7_05_v3:325886-398962	Pf3D7_05_v3:355368-389049	2.81	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_60.2	Pf3D7_05_v3:770125-918502	Pf3D7_05_v3:814427-870387	2.18	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_54.4	Pf3D7_05_v3:693080-769886	Pf3D7_05_v3:732923-733046	2.15	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_11.4	Pf3D7_05_v3:252443-304283	Pf3D7_05_v3:260710-260854	2.14	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_5.7	Pf3D7_05_v3:166792-260709	Pf3D7_05_v3:225881-252442	2.13	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_08_v3_57.5	Pf3D7_08_v3:408724-684033	Pf3D7_08_v3:570281-647334	2.11	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_28.9	Pf3D7_07_v3:496401-694858	Pf3D7_07_v3:611138-611341	1.98	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_57.3	Pf3D7_05_v3:733047-814426	Pf3D7_05_v3:769887-770124	1.98	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_08_v3_40.3	Pf3D7_08_v3:768381-783997	Pf3D7_08_v3:768494-768653	1.97	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_20.2	Pf3D7_07_v3:391071-427528	Pf3D7_07_v3:392209-425264	1.79	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_17.3	Pf3D7_07_v3:371129-392208	Pf3D7_07_v3:377646-391070	1.69	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_0	Pf3D7_05_v3:86612-225880	Pf3D7_05_v3:140933-166791	1.67	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_26	Pf3D7_07_v3:451719-611137	Pf3D7_07_v3:463358-496400	1.65	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_08_v3_91.8	Pf3D7_08_v3:1-230964	Pf3D7_08_v3:122068-122241	1.64	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_23.1	Pf3D7_07_v3:425265-463357	Pf3D7_07_v3:427529-451718	1.64	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_08_v3_48.9	Pf3D7_08_v3:647335-751204	Pf3D7_08_v3:684034-725296	1.6	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_14.4	Pf3D7_07_v3:358161-377645	Pf3D7_07_v3:370990-371128	1.57	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_83.1	Pf3D7_05_v3:1018826-1095899	Pf3D7_05_v3:1040242-1045759	1.53	Genes Contained in this Region	Genes Associated to this Region

Other genes that have similar associations based on eQTL experiments

b. What kinds of genes do you find in this region? Click on the first link in the column "Genomic segment (liberal)". Now examine the gene table on the genomic segment page.



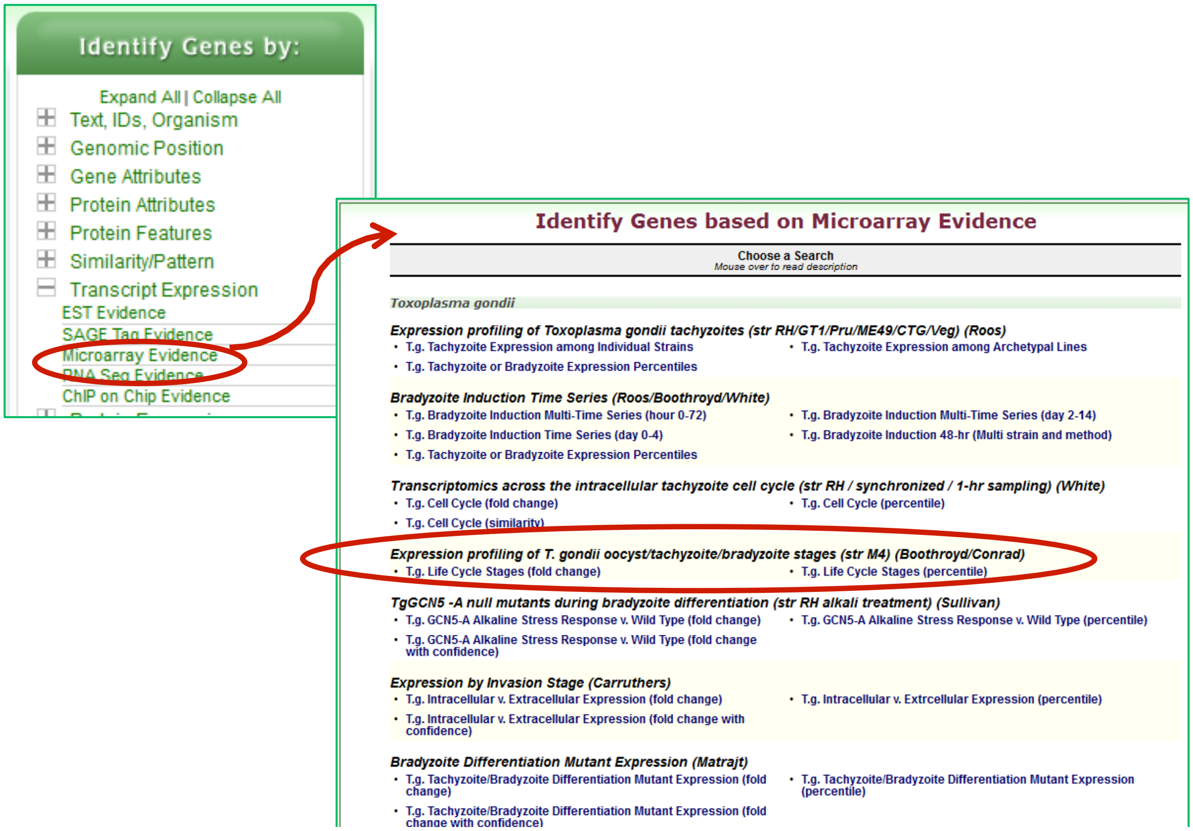
Genes Hide

Gene ID	Start	End	Strand	Product Description
PF3D7_0523000	957890	962149	forward	multidrug resistance protein (MDR1)
PF3D7_0523100	963227	965044	reverse	mitochondrial processing peptidase alpha subunit, putative
PF3D7_0523200	966123	969737	forward	conserved Plasmodium protein, unknown function
PF3D7_0523300	970266	970962	reverse	conserved Plasmodium protein, unknown function
PF3D7_0523400	973518	975876	forward	DnaJ protein, putative
PF3D7_0523500	976690	977815	reverse	outer arm dynein lc3, putative
PF3D7_0523600	978665	979870	forward	conserved Plasmodium protein, unknown function
PF3D7_0523700	980754	985354	reverse	conserved Plasmodium membrane protein, unknown function
PF3D7_0523800	990005	992059	forward	transporter, putative
PF3D7_0523900	993433	994607	reverse	conserved Plasmodium membrane protein, unknown function
PF3D7_0524000	998753	1002124	forward	karyopherin beta (KASbeta)
PF3D7_0524100	1004237	1008108	forward	conserved Plasmodium protein, unknown function
PF3D7_0524200	1008636	1009404	reverse	conserved Plasmodium membrane protein, unknown function

c. What other genes are associated with this block?
(Hint: go back to the gene page eQTL table, and click the "genes associated with this region" link. Run the search on the next page and examine the list of genes. It might be useful to sort this list based on the LOD scores.)

13.4 Finding oocyst expressed genes in *T. gondii* based on microarray evidence.
Note: For this exercise use http://toxodb.org

a. Find genes that are expressed at 10 fold higher levels in one of the oocyst stages than in any other stage in the Expression Profiling of T. gondii ***Oocyst/Tachyzoite/Bradyzoite stages (Boothroyd/Conrad)*** microarray experiment.



Identify Genes based on Microarray Evidence

Choose a Search
Mouse over to read description

Toxoplasma gondii

Expression profiling of Toxoplasma gondii tachyzoites (str RH/GT1/Pru/ME49/CTG/Veg) (Roos)
- T.g. Tachyzoite Expression among Individual Strains
- T.g. Tachyzoite Expression among Archetypal Lines
- T.g. Tachyzoite or Bradyzoite Expression Percentiles

Bradyzoite Induction Time Series (Roos/Boothroyd/White)
- T.g. Bradyzoite Induction Multi-Time Series (hour 0-72)
- T.g. Bradyzoite Induction Multi-Time Series (day 2-14)
- T.g. Bradyzoite Induction Time Series (day 0-4)
- T.g. Bradyzoite Induction 48-hr (Multi strain and method)
- T.g. Tachyzoite or Bradyzoite Expression Percentiles

Transcriptomics across the intracellular tachyzoite cell cycle (str RH / synchronized / 1-hr sampling) (White)
- T.g. Cell Cycle (fold change)
- T.g. Cell Cycle (percentile)
- T.g. Cell Cycle (similarity)

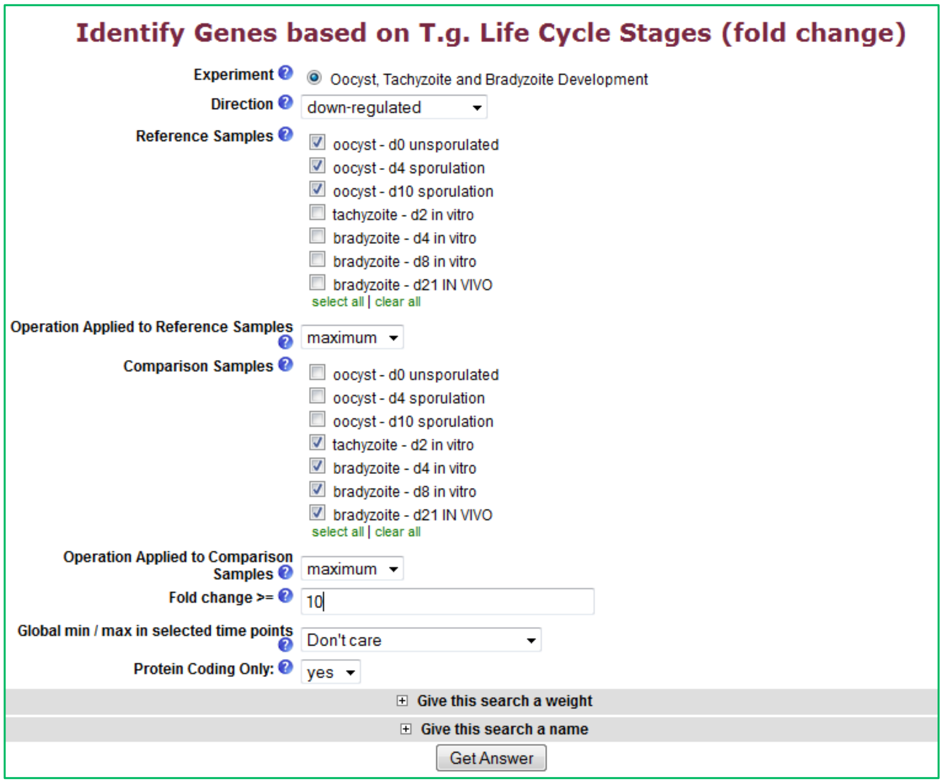
Expression profiling of T. gondii oocyst/tachyzoite/bradyzoite stages (str M4) (Boothroyd/Conrad)
- T.g. Life Cycle Stages (fold change)
- T.g. Life Cycle Stages (percentile)

TgGCN5 -A null mutants during bradyzoite differentiation (str RH alkali treatment) (Sullivan)
- T.g. GCN5-A Alkaline Stress Response v. Wild Type (fold change)
- T.g. GCN5-A Alkaline Stress Response v. Wild Type (percentile)
- T.g. GCN5-A Alkaline Stress Response v. Wild Type (fold change with confidence)

Expression by Invasion Stage (Carruthers)
- T.g. Intracellular v. Extracellular Expression (fold change)
- T.g. Intracellular v. Extrcellular Expression (percentile)
- T.g. Intracellular v. Extracellular Expression (fold change with confidence)

Bradyzoite Differentiation Mutant Expression (Matrajt)
- T.g. Tachyzoite/Bradyzoite Differentiation Mutant Expression (fold change)
- T.g. Tachyzoite/Bradyzoite Differentiation Mutant Expression (percentile)
- T.g. Tachyzoite/Bradyzoite Differentiation Mutant Expression (fold change with confidence)

- There are multiple parameters that need to be set.
- **Experiment**: choose **Oocyst, Tachyzoite and Bradyzoite Development**.
- **Direction**: choose **down-regulated** since we want to find things more highly expressed in oocysts than in other stages.
- Notice setting the Direction to down-regulated automatically changes the ***Operations Applied to Reference Samples*** from average to **maximum** and minimum for the comparator samples. This would enable you to find the genes with the maximum difference between these two sets of samples. Let's leave the reference set to maximum.
- **Reference Samples**: choose the **three oocyst samples: (unsporulated, 4 days sporulated and 10 days sporulated**.
- **Comparison Samples**: choose the **4 non-oocyst samples: 2 days, 4 days, 8 days *in vitro*, and 21 days *in vivo***. (ie, tachyzoite and three bradyzoite samples)
- **Operation Applied to Comparison Samples:** choose maximum since the goal is to find genes with 10-fold higher expression in at least one of the oocyst samples compared to any of the non-oocyst samples.
- **Fold Change >=** 10.
- **Global min/max in selected time points:** choose "**don't care**". Since we have selected all the samples between the reference and comparator time points, the global max and the global min will have to be within the selected time points. If we had not selected all the time points, then changing this parameter would make a difference as the global min or max could be in a time point that we didn't select.
- **Protein coding only** as **yes**. We want to only look at polyadenylated transcripts.



Identify Genes based on T.g. Life Cycle Stages (fold change)

Experiment: ● Oocyst, Tachyzoite and Bradyzoite Development

Direction: down-regulated

Reference Samples:
- ☑ oocyst - d0 unsporulated
- ☑ oocyst - d4 sporulation
- ☑ oocyst - d10 sporulation
- □ tachyzoite - d2 in vitro
- □ bradyzoite - d4 in vitro
- □ bradyzoite - d8 in vitro
- □ bradyzoite - d21 IN VIVO

select all | clear all

Operation Applied to Reference Samples: maximum

Comparison Samples:
- □ oocyst - d0 unsporulated
- □ oocyst - d4 sporulation
- □ oocyst - d10 sporulation
- ☑ tachyzoite - d2 in vitro
- ☑ bradyzoite - d4 in vitro
- ☑ bradyzoite - d8 in vitro
- ☑ bradyzoite - d21 IN VIVO

select all | clear all

Operation Applied to Comparison Samples: maximum

Fold change >=: 10

Global min / max in selected time points: Don't care

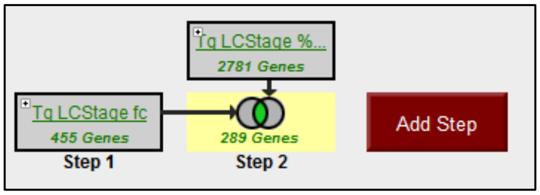
Protein Coding Only: yes

⊞ Give this search a weight

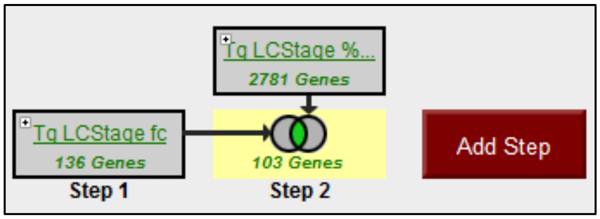
⊞ Give this search a name

Get Answer

b. Add a step to limit this set of genes to only those for which all the non-oocyst stages are expressed below 50th percentile … ie likely not expressed at those stages.
- Hint: use the ***Expression Profiling of T. gondii Oocyst/Tachyzoite/Bradyzoite stages (str M4) (Boothroyd/Conrad)*** -> T.g. Life Cycle Stages (percentile) search.
- Select the 4 non-oocyst samples .
- We want all to have less than 50th percentile so set ***minimum percentile*** **to 0** and ***maximum percentile*** **to 50**.
- Since we want all of them to be in this range, choose **ALL** in the "***Matches Any or All Selected Samples***".
- Set ***Protein Coding Only*** to **YES**.
- Note: you can turn on the column for "M4 Life Cycle Stages – graph" to see the graphs in the final result table.



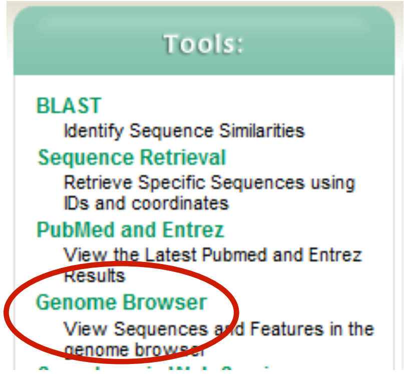
c. Revise the first step of this strategy to find genes where all oocyst stages (d0, 4, 10) are 10 fold higher than any of the non-oocyst stages.
   - Hint, change the “***Operation applied to reference samples***” to **minimum**.
   - Does this result in cleaner, more convincing looking graphs? Why?
   - Would you consider these genes to be oocyst specific?
   - **Save this strategy as we’ll use this strategy for an exercise we are doing this afternoon**.



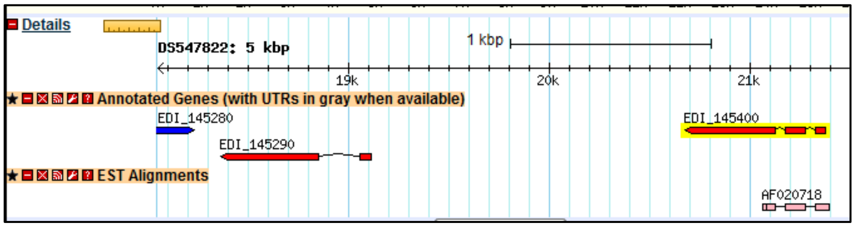
d. Revise the first step of this strategy to find genes that are 3 fold higher in d4 oocysts than any other life cycle stage in this experiment.
   - Do all these genes have d4 oocysts as the global maximum time point?
   - Note that we still have the step to limit the percentile of non-oocyst samples to <= 50th percentile. What happens if you revise this step to also include the d0 and d10 oocyst samples in this percentile range? Do you get more of fewer results back? Why?

13.5 Exploring EST evidence in *Entamoeba.*
Note: For this exercise use http://amoebadb.org

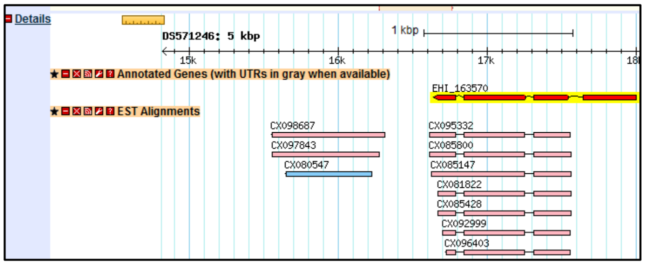
a. Find all *Entamoeba* genes that have EST evidence.
b. Which gene has the highest number of ESTs?
c. Go to the following *E. dispar* gene in the genome browser: **EDI_145400**
   Hint: from the home page click on “Genome Browser” under the tools section, enter the ID in the “Landmark or Region” box.


Tools:
BLAST
Identify Sequence Similarities
Sequence Retrieval
Retrieve Specific Sequences using IDs and coordinates
PubMed and Entrez
View the Latest Pubmed and Entrez Results
Genome Browser
View Sequences and Features in the genome browser

d. Look at the EST evidence for this gene. Does it support the gene model? Does the EST data tell you something else about the gene?
Hint: it would be easier to view this if you only have the gene model and EST tracks on, also you may have to zoom in or out to get a good view of the gene.



e. Now, go to the following *E. histolytica* gene in the genome browser: **EHI_163570** (just like you did in step 'c' above). What does the EST evidence look like? Are there any ESTs that support an alternative gene model? Look to the left of this gene. What do these ESTs mean?

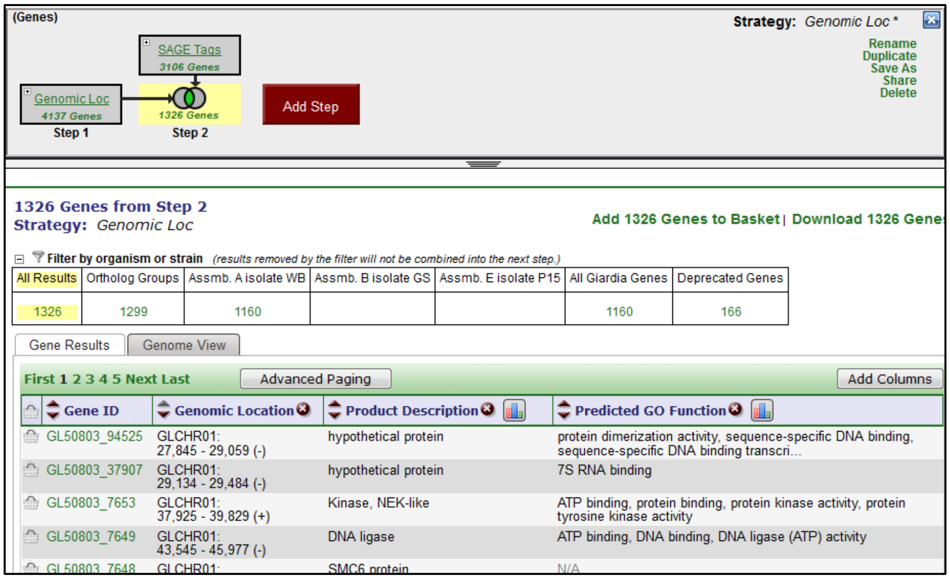


13.6 Finding all ESTs that do not coincide with a gene model in *Entamoeba.* Note: For this exercise use http://www.amoebadb.org

a. Find all ESTs that do not overlap with genes. Hint: Use the "Extent of Gene Overlap" search under the heading "Identify Other Data Types".
b. How many ESTs did you get? Hint: make sure you changed the ***base overlap to '0'*** and selected "does not overlap with a gene".
c. Visit one of the EST pages and explore it. Can you get to the genome browser from here to see this EST? Hint: Look at the "Alignments to genomic sequence" section, click on "view".

13.7 Exploring genes expressed during encystation in *Giardia* based on SAGE tag evidence. Note: For this exercise use http://www.giardiadb.org

a. Find all genes on chromosome " GLCHR05". Hint: search for genes based on genomic location, under the menu “Genomic Position” in the “Identify Genes by” section.
b. How many of those genes have SAGE tag evidence during encystation? Hint: add a SAGE tag evidence (under Identify Genes By Transcript Expression) step using the following parameters: allow the tag to align 20 bp from either end and align to only one place in the genome. Find only genes with a tag count >= 5.
c. Do any of these genes have nucleic acid binding activity? Can you tell this from the product names? What other information can you add that would help? Hint: add a “Predicted GO Function” column to the list of your results.
d. How would you identify other genes in the *Giardia* that have a nucleic acid binding function and are expressed during encystation?



1326 Genes from Step 2
Strategy: *Genomic Loc*

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

All Results	Ortholog Groups	Assmb. A isolate WB	Assmb. B isolate GS	Assmb. E isolate P15	All Giardia Genes	Deprecated Genes
1326	1299	1160			1160	166

Gene ID	Genomic Location	Product Description	Predicted GO Function
GL50803_94525	GLCHR01: 27,845 - 29,059 (-)	hypothetical protein	protein dimerization activity, sequence-specific DNA binding, sequence-specific DNA binding transcri...
GL50803_37907	GLCHR01: 29,134 - 29,484 (-)	hypothetical protein	7S RNA binding
GL50803_7653	GLCHR01: 37,925 - 39,829 (+)	Kinase, NEK-like	ATP binding, protein binding, protein kinase activity, protein tyrosine kinase activity
GL50803_7649	GLCHR01: 43,545 - 45,977 (-)	DNA ligase	ATP binding, DNA binding, DNA ligase (ATP) activity
GL50803_7648	GLCHR01:	SMC6 protein	N/A

13.8 Exploring ChIP-chip data in *Toxoplasma*.
Note: For this exercise use http://www.toxodb.org

a. Use ChIP-chip searches to identify all genes that are differentially expressed between Type I and Type II strains of *T. gondii*.
How many genes are likely transcriptionally active in Type I but not Type II strains?
b. Go to the gbrowse view of one of those genes. Hint: use the column to add the gbrowse link to your list of results.
c. Turn on the tracks for ChIP-chip data. Does the data agree with your search results?
d. Zoom out and explore neighboring genes.
e. Can you find centromeres? Turn off all tracks except the ChIP-chip Centromeres track. Hint: zoom out so that you have a view of the entire chromosome.
f. Once you locate the centromeric peaks, zoom in to explore what they look like.

g. Turn on one of the ChIP-chip graphs. What do you notice?

13.9 Exploring microarray data in TriTyrpDB.
Note: For this exercise use http://www.tritrypdb.org

a. Find all genes in *T. brucei* that are up-regulated 48 hours (as compared to the 0 time point) post induction of differentiation (look for at least 4-fold induction).
   Hint: notice that there are two experiments in the "**T.b. differentiation time series (fold change)**" search. You need to combine results from both experiments. Also, in one of the experiments the "blood form – high density" is equal to the 0 time point. How did you combine the above two experiments? Union or intersect?
b. How do these results compare to the RNA-seq data? Can you view RNA-seq data for all of these results? Hint: explore the columns in your result list. Add the column that corresponds to the RNA-seq data.
c. Start a new search and look for genes that are greater than 4-fold down regulated in amasitgotes compared to metacyclics. How many genes did you get? Which gene is the most repressed gene? Hint: sort the "fold change avg." column in the list of results.
d. Go to the gene page of the top gene from step 'c'. This gene is annotated as hypothetical – but can you get some hints to its possible function? Hint: look at the protein feature section.

13.10: Comparing RNA abundance and Protein abundance data. For this exercise use http://TriTrypDB.org.

In this exercise we want to compare the list of genes that show differential RNA abundance levels between procyclic and blood form stages in *T. brucei* with the list of genes that show differential protein abundance in these same stages.

a. Go to the genes by microarray expression -> T.b. Expression profiling of five life cycle stages (fold change). Configure the search to return protein coding genes that are down-regulated 2 fold in procyclic form (PCF) (I chose both log and Stat and averaged them) relative to the Blood Form reference sample.
b. Add a step to compare with protein expression. Genes by protein expression -> Quantitative Mass Spec Evidence. Configure this search to return genes that are downregulated in procyclic form relative to Blood form.
c. How many genes are in the intersection? Does this make sense … make certain that you set the directions correctly (*it can be quite confusing ☺).*
d. Try changing directions and compare up-regulated genes/proteins. (*hint, revise the existing strategy … you might want to duplicate it so you can keep both*). When you change one of the steps but not the other do you have any genes in the intersection? Why might this be??

e. Can you think of ways to provide more confidence (or cast a broader net) in the microarray step? *Hint: you could insert steps to restrict based on percentile or add a RNASequencing step that has the same samples.*