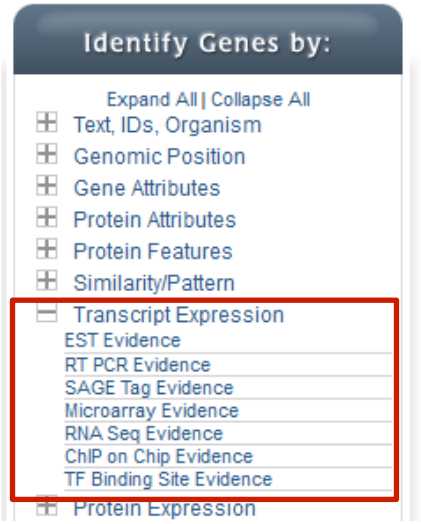
Exploring Transcriptomics Data
Exercise 12

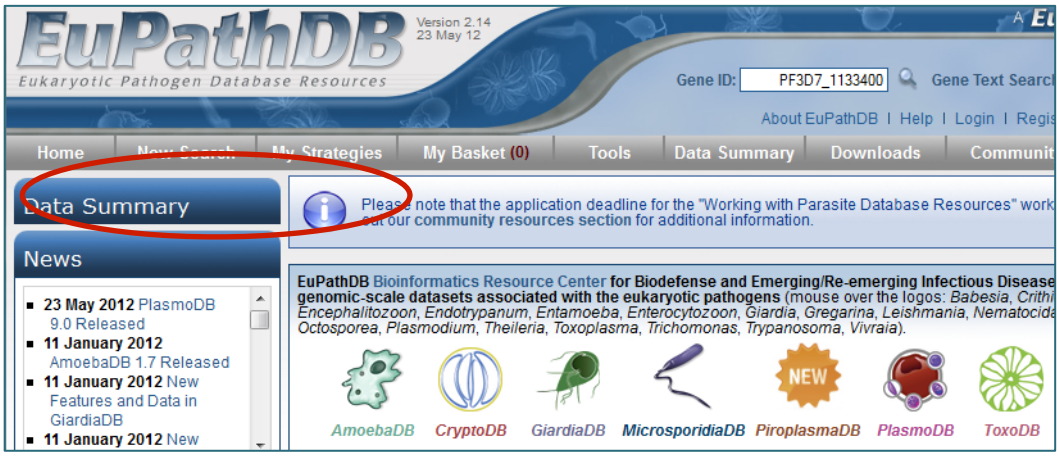
12.1 Evidence of expression at the transcriptional level.

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.



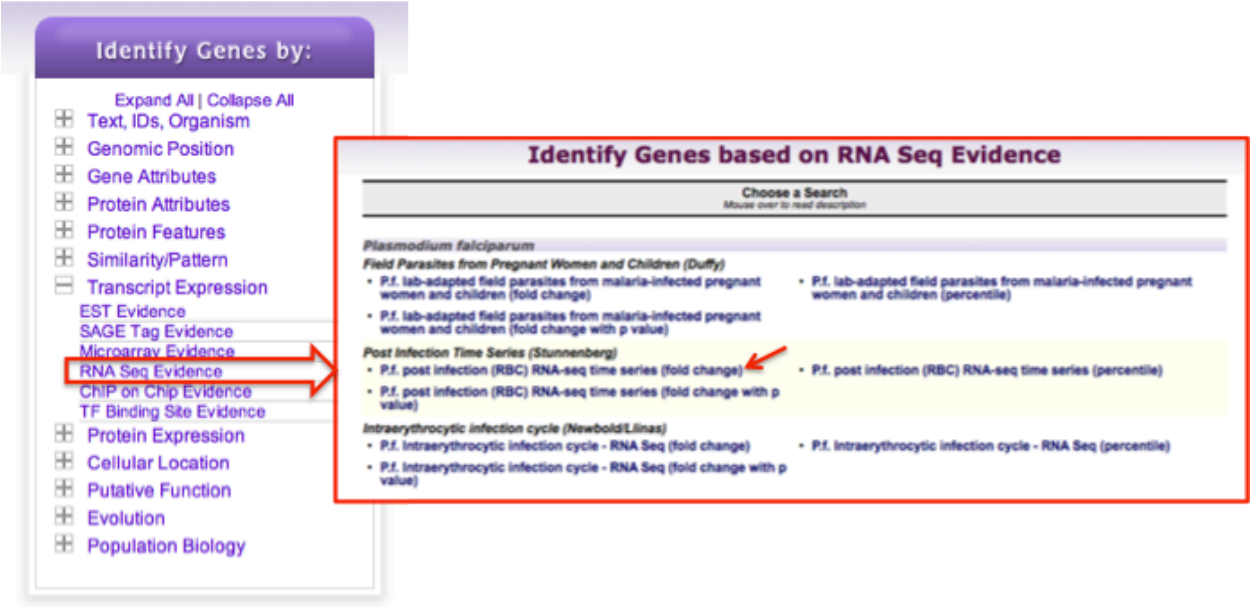
b. Explore organisms that have microarray data. What organisms have expressed sequence tag (EST), RNA sequence, ChIP-chip or SAGE tag data?
c. What does RNA-seq data tell you that microarray data cannot? What does ChIP-chip data tell you about a gene?
d. Go to the Data Summary Section, can you find the same information there?
Hint: data summary table in on the left side of the home page.



12.2 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)").



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

Select your ***reference time*** points – these are the time points you will compare to. In this case these are the early time points **(5,10,15,20,25)**.

Select your ***comparison time*** points – these are the time points you are interested in comparing to the reference. In these case you are interested in genes that are upregulated in later time points **(30,35,40)**.

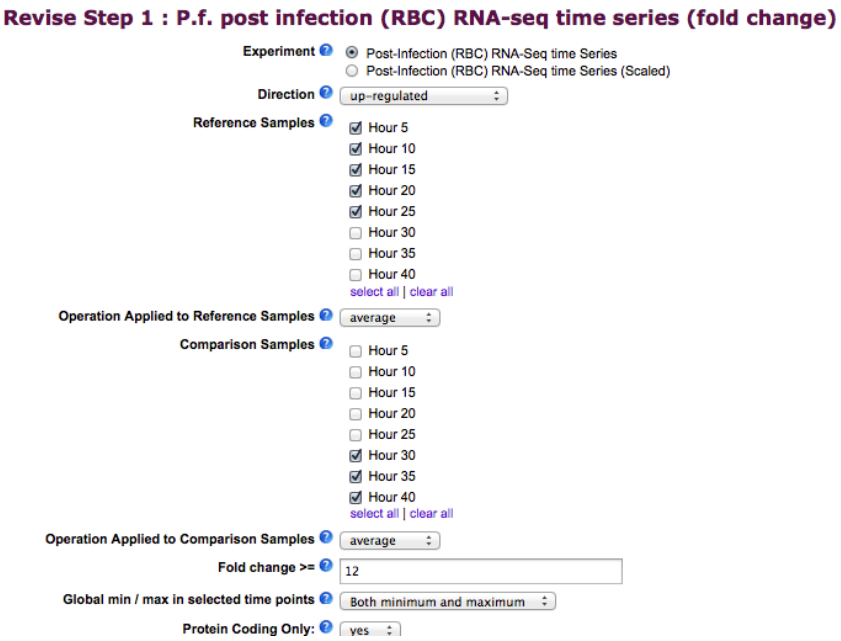
Fold induction: choose how induced do you want your comparison samples to be. For this exercise choose 12 but feel free to modify this.

Choose the ***direction*** of regulation – in this case **up-regulated**.

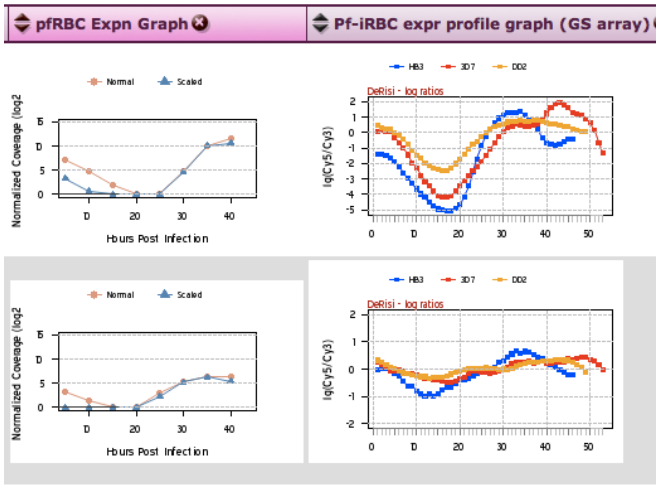
Choose the ***Operation when selecting multiple samples***: *Choosing minimum / maximum if doing up-regulated will use the minimum of the Reference and the maximum of the Comparison to calculate fold change (opposite for down-regulated). Choosing average will use the average of the samples selected in each group to calculate the fold change* – for this example use **average**.

Global min / max in selected time points: Choose whether the selected samples must be the global minimum or maximum or both. Choosing

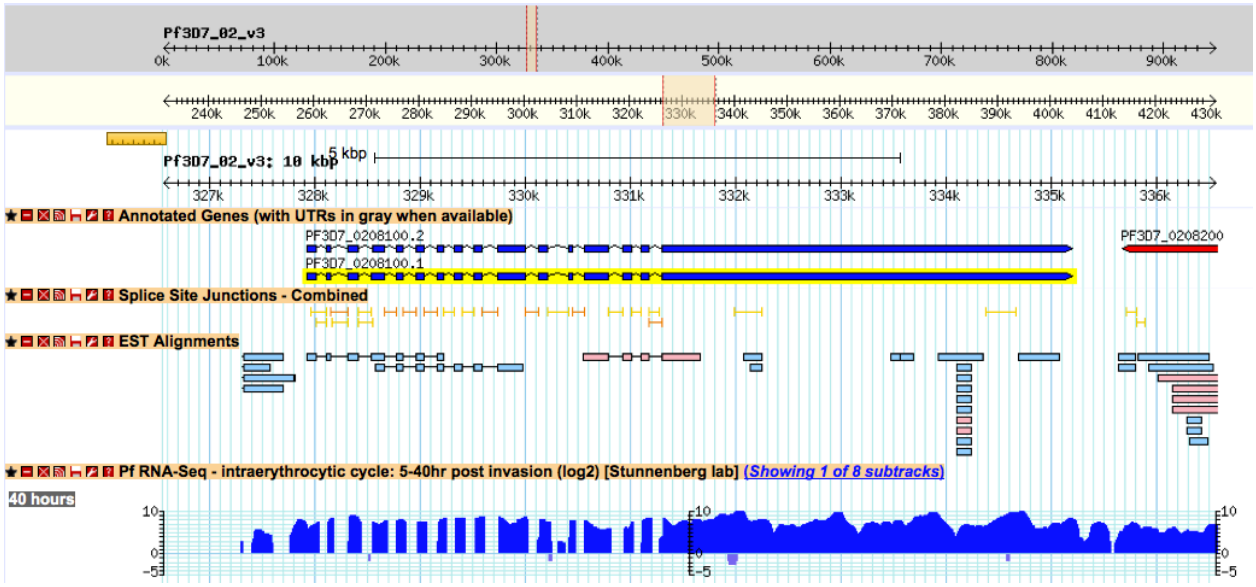
minimum: if doing up-regulated then the Reference values selected must be the global minimum, if doing down-regulated, the comparison values selected must be the global minimum. Choosing maximum: if doing up-regulated then the Comparison values selected must be the global maximum, if doing down-regulated, the reference values selected must be the global maximum – for this example use **Maximum**.



b. How do these RNA-sequence results compare to microarray results? (hint: add the column called "Pf-iRBC expr profile graph (GS array)" and compare the RNA-seq to the microarray graphs).



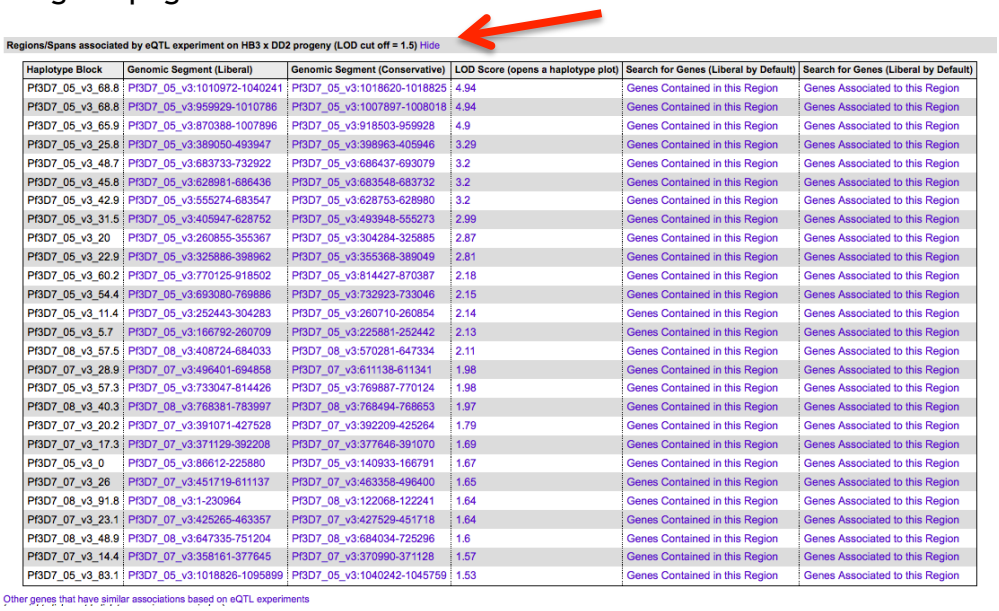
c. Which gene has the most number of exons? (hint: use the columns).
d. Is this gene alternatively spliced? Look at the gene page. Take note of the Gene ID.
e. View this gene in the genome browser and load the RNA-seq tracks for this experiment “5 - 40 hours post infection (Stunnenberg lab)”. Do these 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).



12.3 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.

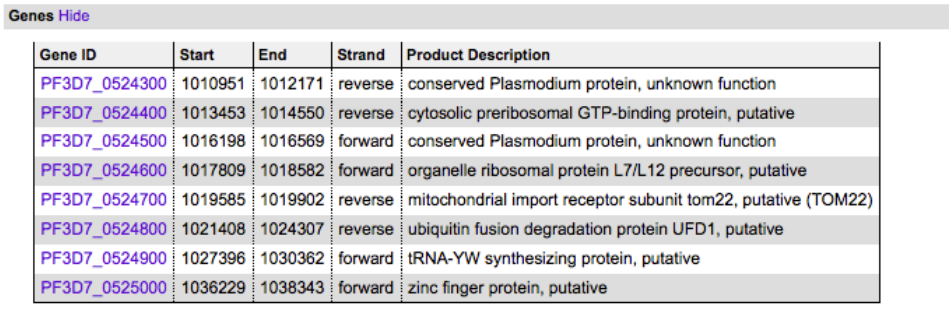


Regions/Spans associated by eQTL experiment on HB3 x DD2 progeny (LOD cut off = 1.5) Hide

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 next page.



Genes Hide

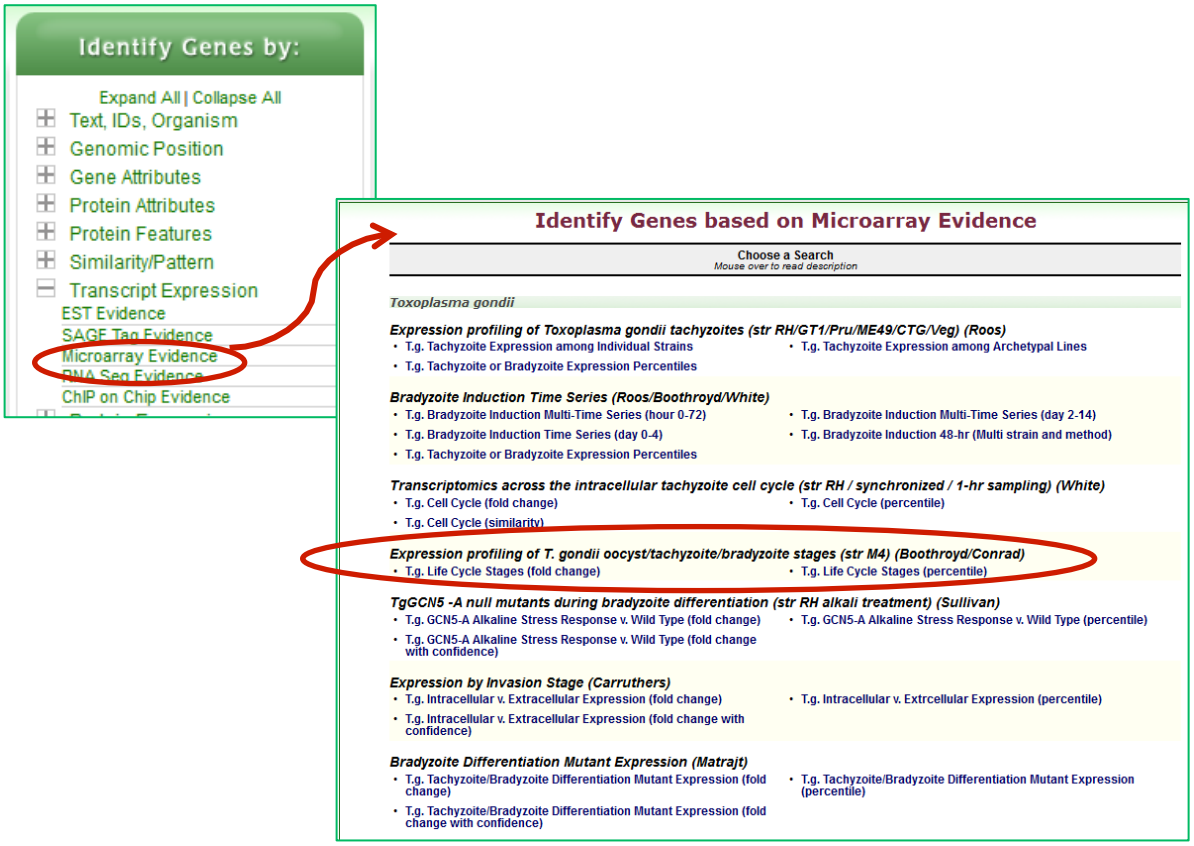
Gene ID	Start	End	Strand	Product Description
PF3D7_0524300	1010951	1012171	reverse	conserved Plasmodium protein, unknown function
PF3D7_0524400	1013453	1014550	reverse	cytosolic preribosomal GTP-binding protein, putative
PF3D7_0524500	1016198	1016569	forward	conserved Plasmodium protein, unknown function
PF3D7_0524600	1017809	1018582	forward	organelle ribosomal protein L7/L12 precursor, putative
PF3D7_0524700	1019585	1019902	reverse	mitochondrial import receptor subunit tom22, putative (TOM22)
PF3D7_0524800	1021408	1024307	reverse	ubiquitin fusion degradation protein UFD1, putative
PF3D7_0524900	1027396	1030362	forward	tRNA-YW synthesizing protein, putative
PF3D7_0525000	1036229	1038343	forward	zinc finger protein, putative

c. What other genes are associated with this block? (hint: click on 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.

12.4 Finding oocyst expressed genes in *T. gondii* based on microarray evidence.

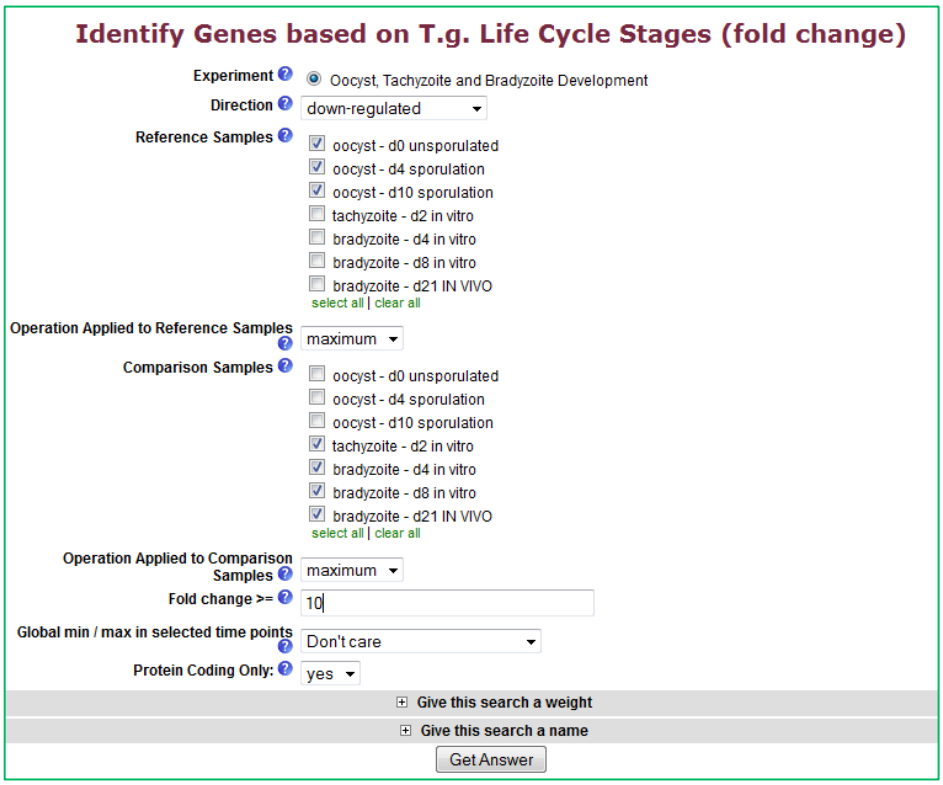
Note: For this exercise use http://toxodb.org

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.



- There are multiple parameters that need to be set for this query.
- For ***Experiment*** choose **Oocyst, Tachyzoite and Bradyzoite Development**.
- For the ***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.
- For the ***Reference Samples*** choose the **three oocyst samples**.

- For the ***Comparison Samples*** choose the **4 non-oocyst samples** (ie, tachyzoite and three bradyzoite samples)
- We want to change the ***Operation Applied to Comparison Samples*** to 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.
- Set the ***Fold Change >=*** **10**.
- Can leave ***global min/max in selected time points*** as "**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.
- Select ***protein coding only*** as **yes**. We want to only look at polyadenylated transcripts.



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.

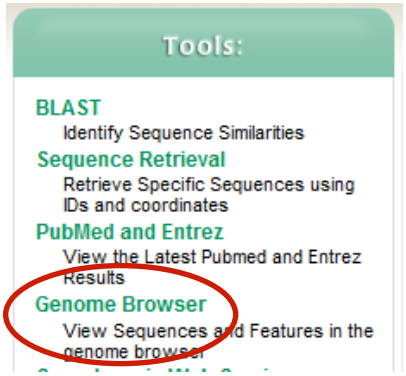
- We want all to have less than 50th percentile so set ***minimum percentile*** to 0 and ***maximum percentile*** to 50.
- Since we want all of them to be in this range, choose **ALL** in the "***Matches Any or All Selected Samples***".
- Set ***Protein Coding Only*** to **YES**.
- Note: you can turn on the column for "M4 Life Cycle Stages – graph" to see the graphs in the final result.

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?

12.5 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.



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, are what do these ESTs mean?

12.6 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 "Extend of Gene Overlap" search under the heading "Identify Other Data Types".
b. How many ESTs did you get? Hint: make sure you change the ***base overlap to '0'*** and select "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".

12.7 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 scaffold "CH991769". 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?

12.8 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?

12.9 Exploring microarray data in TriTyrpDB.

Note: For this exercise use http://www.tritrypdb.org

a. Find all genes in *T. brucei* that are upregulated 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.