Exploring Transcriptomics Data
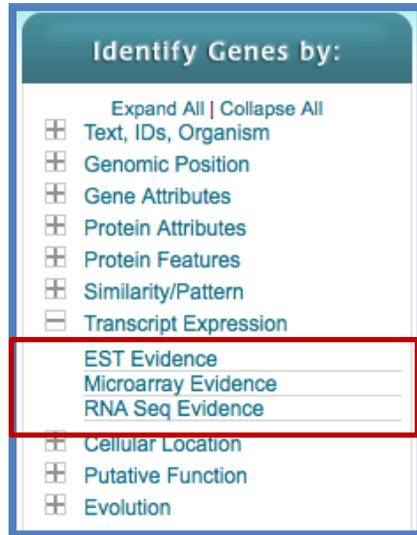## Exercise 13

**13.1 Evidence of expression at the transcriptional level.
  Note: For this exercise use http://www.fungidb.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), or RNA sequence?
c. What does RNA-seq data tell you that microarray data cannot?
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.
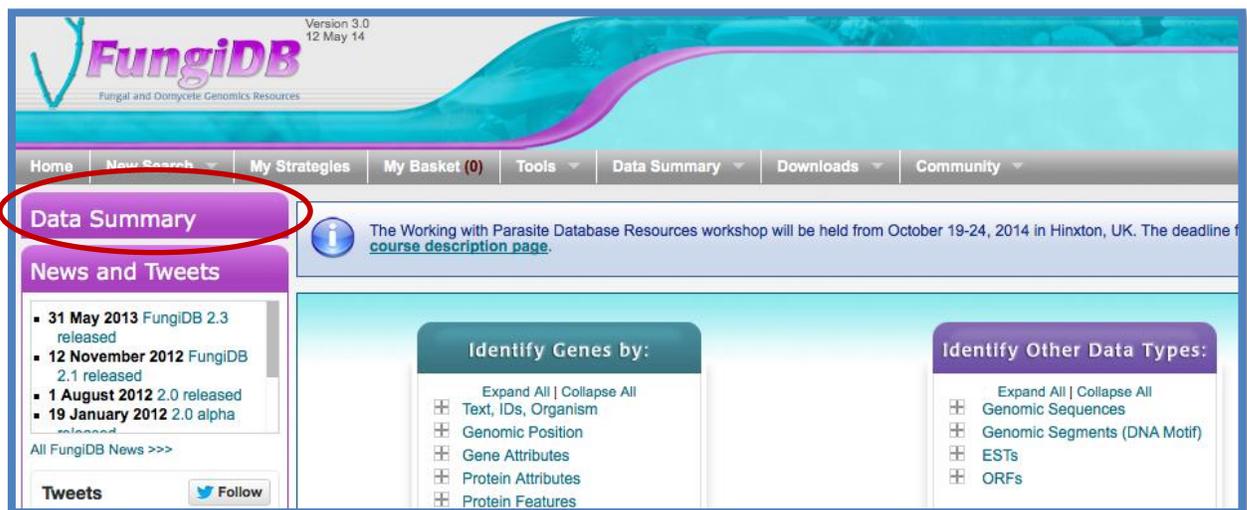
13.2 Exploring RNA sequence data in *FungiDB.*
   Note: For this exercise use http://www.fungidb.org

a. Find all genes in *C. posadasii C735 delta SOWgp* that are upregulated based on RNA-seq data at Parasitic spherule phase compared to Saprobic hyphae.

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

Experiment: Choose the experiment of interest, in this case the only option available: Saprobic vs Parasitic Growth

Genes: gene format of the results. Choose protein coding.

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

Fold Change>=: fold change is calculated as the ratio of two values (expression in reference)/(expression in comparison). The intensity of difference in expression needed before a gene is returned by the search. Choose 2 but feel free to modify this.

Reference Sample: the samples that will serve as the reference when comparing expression between samples. Choose Saprobic Hyphae

Comparison Sample: the sample that you are comparing to the reference. In this case you are interested in genes that are up-regulated in Parasitic Spherules phase

Identify Genes based on C posadasii C735 delta SOWgp Saprobic vs Parasitic Growth RNASeq (fold change)

Tutorial

For the **Experiment** [Saprobic vs Parasitic Growth ⬍] ❓

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** ❓

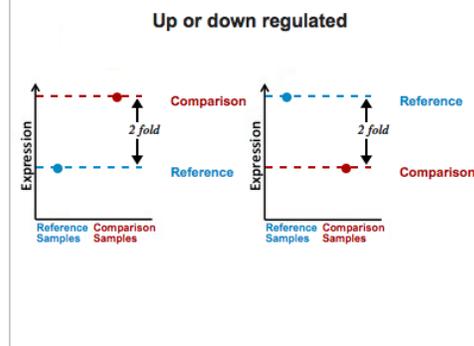☑ Saprobic Hyphae
☐ Parasitic Spherules

and its **expression value** ❓
in the following **Comparison Samples** ❓

☐ Saprobic Hyphae
☑ Parasitic Spherules

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

Up or down regulated

You are searching for genes that are **up or down regulated** between one **reference sample** and one **comparison sample**.

For each gene, the search calculates:

$$fold\ change_{up} = \frac{comparison\ expression\ value}{reference\ expression\ value}$$

$$fold\ change_{down} = \frac{reference\ expression\ value}{comparison\ expression\ value}$$

and returns genes when **fold change$_{up}$ >= 2 or fold change$_{down}$ >= 2.**

See the detailed help for this search.

Select Columns

reset to current | reset to default

- ☐ Search-Specific
  - ☐ Fold Change (Avg)
  - ☑ Fold Change
  - ☐ Avg Comp (log2)
  - ☐ Avg Ref (log2)
  - ☐ Min Comp (log2)
  - ☐ Min Ref (log2)
  - ☐ Max Comp (log2)
  - ☐ Max Ref (log2)
  - ☑ Chosen Comp (log2)
  - ☑ Chosen Ref (log2)
  - ☐ Time of Min Expression
  - ☐ Time of Max Expression
- ☐ Text, IDs, Species
  - ☐ Genomic Sequence ID
  - ☑ Organism
  - ☐ UniProt ID
  - ☐ Gene Name or Symbol
  - ☐ Previous ID(s)
  - ☐ Entrez Gene ID
- ☐ Genomic Position
- ☐ Gene Attributes
- ☐ Protein Attributes
- ☐ Protein Features
  - ☐ Protein Length
  - ☐ # TM Domains
  - ☐ SignalP Scores
  - ☑ SignalP Peptide
- ☐ Transcript Expression
- ☐ Putatitve Function
  - ☐ EC Numbers
  - ☑ Annotated GO Function
  - ☑ Annotated GO Process
  - ☐ Annotated GO Component
  - ☐ Predicted GO Function
  - ☐ Predicted GO Process
  - ☐ Predicted GO Component
- ☐ Evolution
- ☐ Population Biology

b.   For the genes returned by the search, what are the top 15 upregulated genes in the parasitic phase compared to the saprobic phase?

c.   Can you find more information for the "hypothetical genes"? hint: add columns from the putative function option.

d.   Are some of these upregulated genes secreted? Choose the SignalP Peptide box under the Protein Feature option.

e.   Are these genes unique to C. posadasii? Can the ortholog data help us find this information?

f.   What does the paralog count tell us about these top upregulated genes?

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_05_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. (fold change)
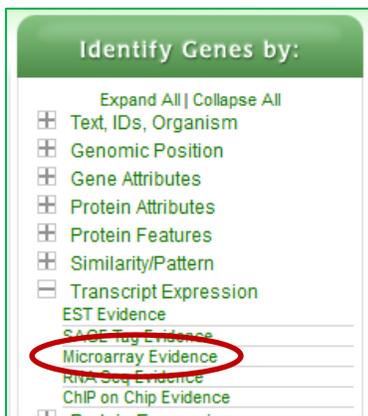
- There are multiple parameters that need to be set.
- **Experiment**: choose Oocyst, Tachyzoite and Bradyzoite Development.
- **Direction**: choose down-regulated since we want to find things more highly expressed in oocysts than in other stages.
- Notice setting the Direction to down-regulated automatically changes the **expression value** for reference sample 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)
- choose maximum **expression value** for comparison sample 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.
- **Select Protein coding genes**. 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)*** -> **T.g. Life Cycle Stages (percentile)** search.
- Select the 4 *in-vitro* 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***".
- Select ***Protein Coding genes***.
- Note: you can turn on the column for "M4 Life Cycle Stages – graph" to see the graphs in the final result table. (add column; transcript expression; microarray; tg-life cycle; tg m4 life cycles stages graph)
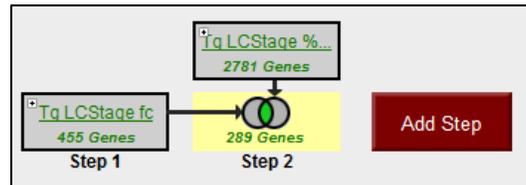
c. Revise the first step of this strategy to find genes where all oocyst stages (d0, 4, 10) are 10 fold higher than any of the non-oocyst stages.
- Hint, change the "***expression values*** to reference samples to ==minimum==.
- Does this result in cleaner, more convincing looking graphs? Why?
- Would you consider these genes to be oocyst specific?
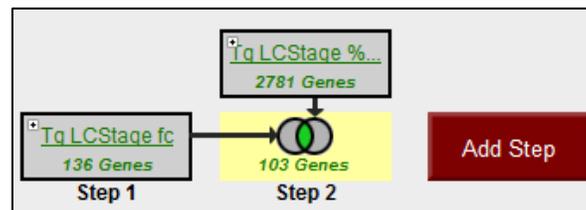
13.5   Exploring EST evidence in *Phytophthora infestans.*

a. Find all genes that have EST evidence.
b. Which gene has the highest number of ESTs?
c. Can you find some gene models that do not match their ESTs?
   Check out the Genome Browser linked off the gene page. Go to 'select tracks' and make sure ESTs are shown.

   Try sorting by the number of ESTs to find those with just a few alignments. Those with just 1 EST aren't very interesting but maybe those in the 5-20 range would be better. You can revise your search to return genes with greater than 5, or 10, etc EST hits and then sort by ESTs.