VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

Mining Transcriptomics Data

Genes up-regulated in liver stages of *Plasmodium* infection

Learning objectives

- Review the types of expression searches in VEuPathDB
- Use the differential expression and fold change searches to explore gene expression in liver stage *Plasmodium* infections

Transcriptomics data in VEuPathDB

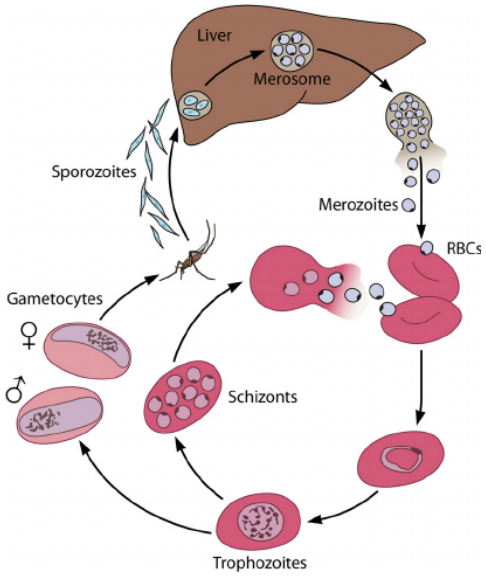
Transcript expression, or the abundance of an mRNA, can be determined in the laboratory with several different techniques including RNA sequencing, microarray, and RT-PCR. VEuPathDB supports these data types with several searches, as shown in the table below. For RNA sequence data, expression values are graphed on gene pages and mapped reads can be visualized in the genome browser. Using the search strategy system, it's easy to delve deep into a specific data set and to take advantage of several types of data when combining search results in the strategy system.

Search	Description	RNA- seq	Microarray
Differential Expression	**Statistical** analysis of studies whose experimental design includes **biological replicates**. Differential expression (DE) methods rely on replicates to estimate how much gene expression naturally varies within a condition. A DE search finds genes based on fold change difference between two samples with a user defined p-value cutoff. **Only pairwise comparisons** can be made with this search.	✓	
Fold Change	Expression differences between samples are calculated but **statistical analyses are not performed**. A fold change search finds genes whose expression value differs between samples without considering statistical parameters. This search offers a form of differential expression analysis when the experimental design **did not include replicates** and **allows for comparing groups of samples**, e.g. find genes whose expression is up-regulated in the liver time course (2, 24, 36, and 54 hours) vs the control (0 hours).	✓	✓
Percentile	For each sample in an experiment, each genes' **expression value is sorted from lowest to highest** and a **percentile rank** is determined. For example, a percentile search can find genes whose expression is in the highest 10% of expression values within a sample.	✓	✓

Sense/ Antisense	For strand-specific RNA sequence, expression values are determined in the sense and antisense direction. This search **finds genes that exhibit simultaneous changes in sense and antisense transcripts**. For example you can look for genes with increasing antisense transcripts and decreasing sense transcripts, as might occur when antisense transcription suppresses sense transcription.	✓	
Splice-site Location	This **trypanosome-specific search** takes advantage of the 'splice-leader' RNA-seq data which determines transcript abundance within the polycistronic mRNA using splice-leader specific primers. This search identified genes whose 5' splice site location varies between samples.	✓	
Metacycle	The MetaCycle package detects rhythmic signals from **large scale time-series data**, such as circadian rhythms within expression time courses, using either ARSER or JTK-Cycle. This search returns genes whose rhythmic signals match the conditions (period and amplitude range) you specify. The search will return the corresponding period, amplitude and p-value of genes that meet your search criteria.	✓	✓
Similarity	The similarity search returns genes whose **expression profile** within the experiment **follows a similar pattern** as the gene you specify.	✓	✓
Direct Comparison	**Microarray data** for two samples is often collected on the same glass slide. For these experiments, the direct comparison search returns genes whose expression varies between samples in **pairwise comparisons**.		✓
Coexpression	**Meta-analysis across multiple microarray experiments** defined a co-expression network. This search returns genes within the co-expression network of your gene(s) of interest.		✓

Introduction to the exercise

The life cycle of *Plasmodium* is split between the sexual mosquito stage and the asexual host phase. The host stage includes a 6-7 day asymptomatic liver stage which ends with the release of merozoites into the bloodstream where they infect erythrocytes. The erythrocytic stages are well studied compared to the liver stages.



PlasmoDB contains RNA seq data from a study in the rodent model *Plasmodium berghei* that includes a time course of liver infection as well as sporozoite and merozoite samples for comparison (Caledlari et al. 2019).

Seven samples were assayed in triplicate for RNA sequence data:

- Sporozoites
- 6 hr liver infection

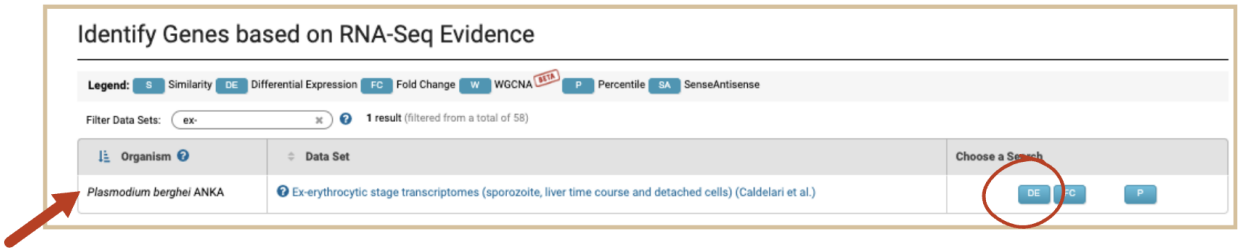
- 24 hr liver infection
- 48 hr liver infection
- 54 hr liver infection
- 60 hr liver infection
- Merozoites (detached cells)

The purpose of this two-part exercise is to use data from this study to explore gene up-regulation in liver stages of *Plasmodium* infection.

Part 1: Determine what genes are up-regulated at least 4 fold (p-value <= 0.001) at 48 hr post infection vs. the sporozoite stage

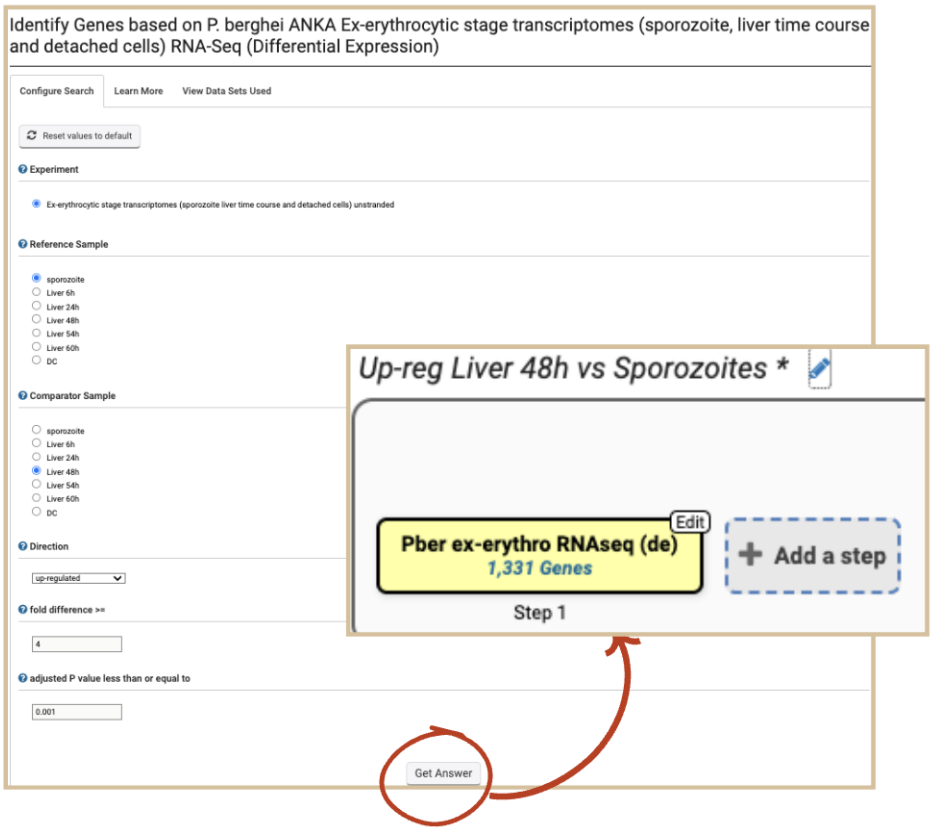
1. Navigate to the appropriate search for this question
    a. Go to PlasmoDB.org
    b. From the sidebar or header, search for “RNA” and click on **RNA-Seq Evidence**

2. Choose appropriate experiments/ samples & parameters within the search
    a. Find the data set called “Ex-erythrocytic stage transcriptomes (sporozoite, liver time course and detached cells) (Caldelari et al.)”
    b. Select the DE (differential expression) search

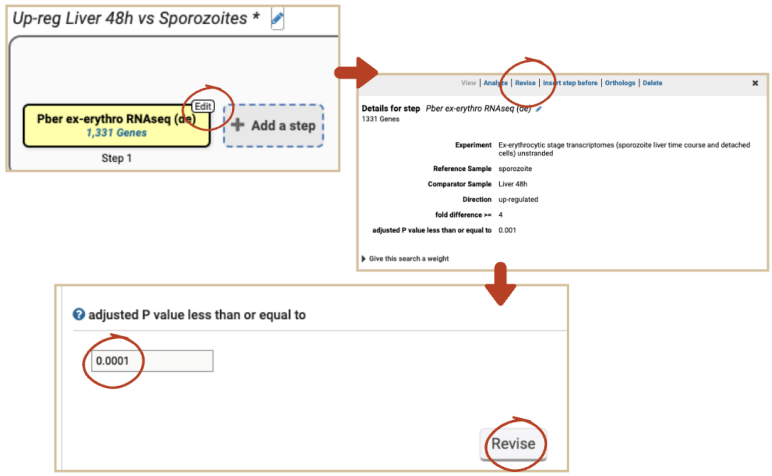


3. Configure the differential expression search to return genes that are at least 4 fold up-regulated in the 48-hour liver infection compared to sporozoites with a p-value of p<0.001.

   How many genes did you get? Do you believe these results? To convince yourself, you could browse the product description column of the results table. Are there clues that these genes are liver-specific?

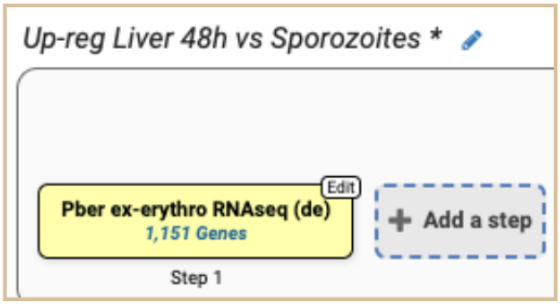


4. Increase the statistical stringency of the search from p <= 0.001 to p<0.0001.



Hint: revise the search and change the p-value. Hover over the yellow search box until the Edit icon appears. Click the Edit icon and choose revise from the options panel.

How many genes are returned by the search now?



5. What other properties would you expect of a late liver stage gene/protein? Since the next step is to emerge from the hepatocyte, these genes may have proteolytic activity. How could you integrate functional annotation into your search strategy?

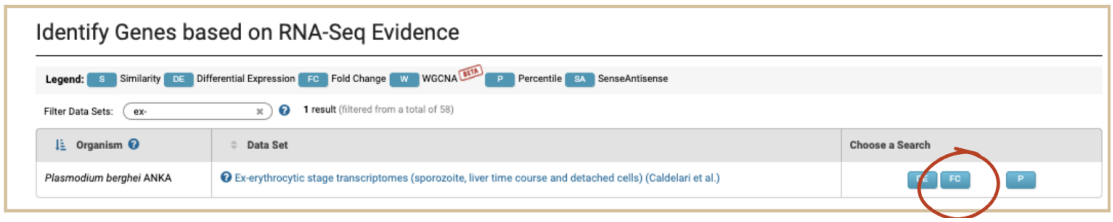
6. Add a step to your search to add functional annotation. Intersect your RNA seq search with a GO term search to see if any of your genes are annotated with proteolytic or peptidase activity. (GO:0008233 peptidase activity GO:0006508 proteolysis).

   How many genes are up-regulated in late liver stages compared to sporozoites AND have been annotated with proteolytic or peptidase activity?

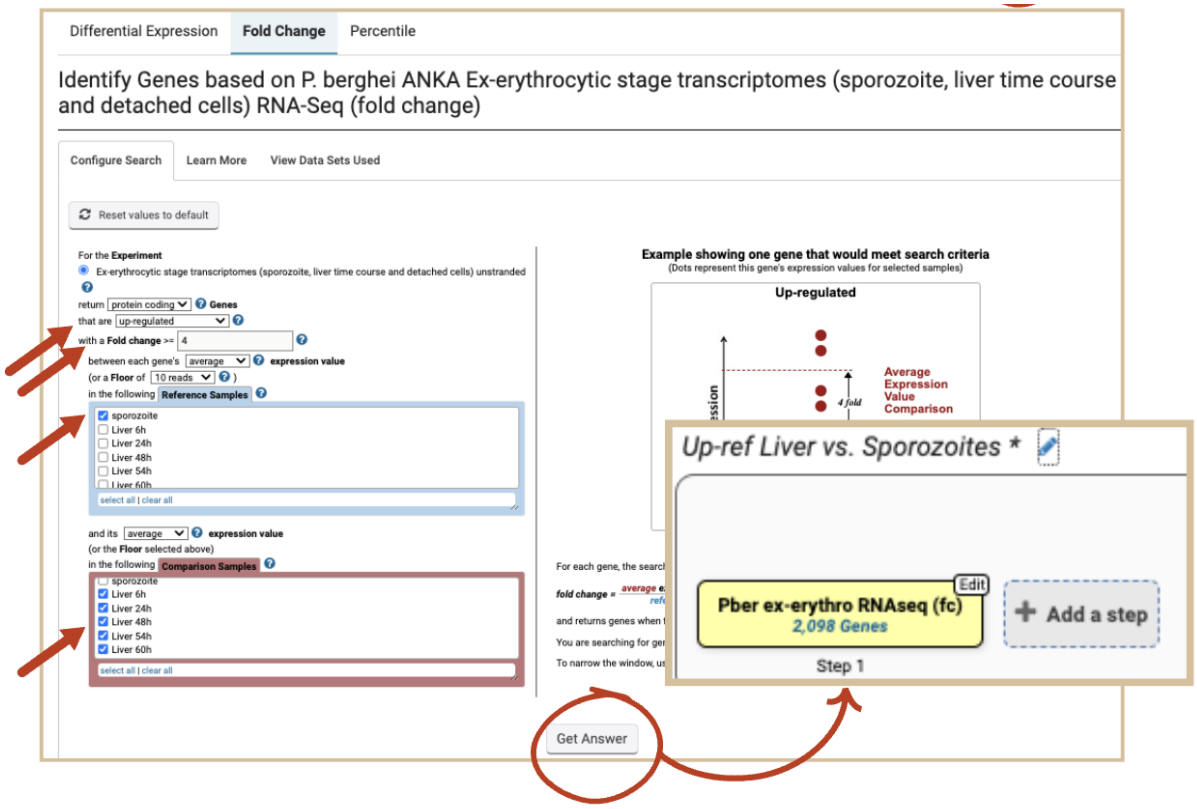
Part 2: Determine what genes are up-regulated 4 fold in any liver stage compared to sporozoites

1. Navigate to the same dataset that was used in the first part of this exercise- “Ex-erythrocytic stage transcriptomes (sporozoite, liver time course and detached cells) (Caldelari et al.)”

2. For this question, select the FC (fold change) search. A fold change search finds genes whose expression value differs between samples without considering statistical parameters. It allows for comparing groups of samples.



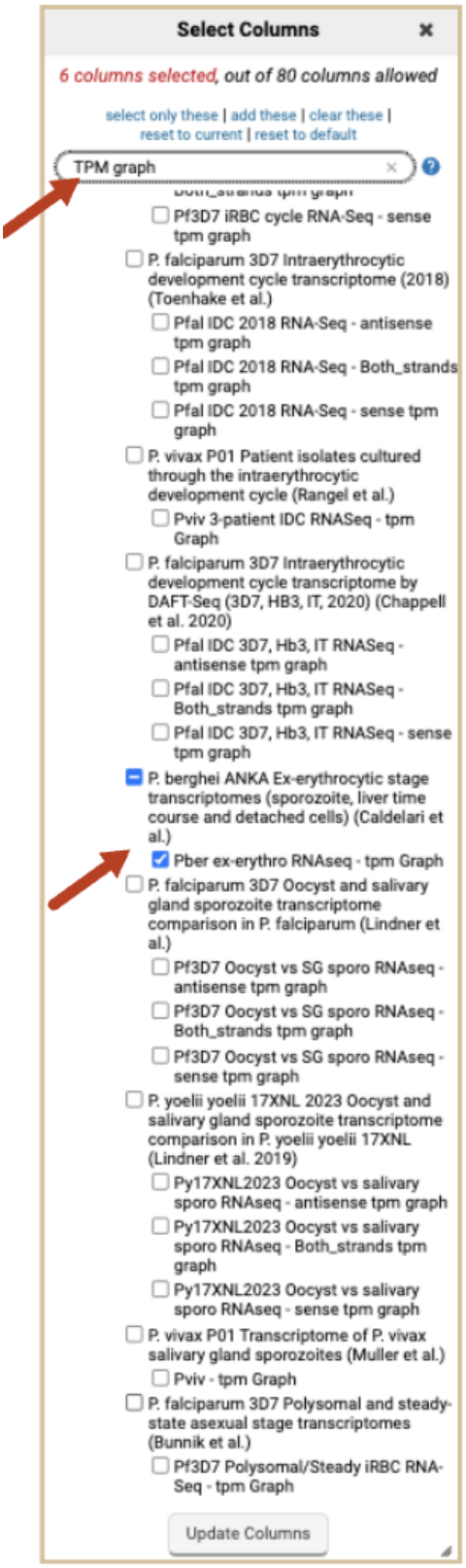
3. Configure the fold change search to return genes that are 4 fold up-regulated in the average expression across all the liver stages compared to the sporozoites.



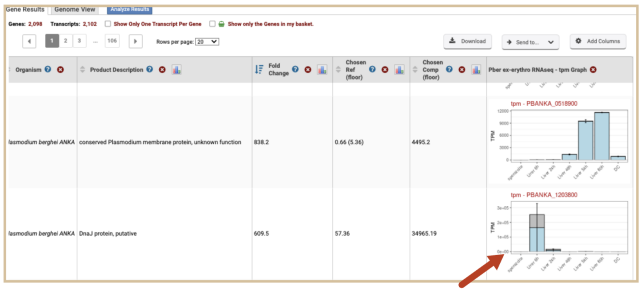
How many genes did you get?

Did the search return more genes or fewer genes than the differential expression search?

4. Explore search results. Use the Add Columns to turn on the TPM graph for the ‘Ex-erythrocytic stages’ data set. Notice the error bars for the DNAJ protein PBANKA_1203800.



Would this gene be returned by the Differential Expression search that applies statistics before returning genes?



5. Add a step to the search strategy to determine what genes in this result are also represented in the top 10% of genes in the merozoite (aka detached cells or DC) sample.
   How many genes are up-regulated in liver stages compared to sporozoites and also represented in the top 10% of genes expressed in detached cells (merozoites)?