

Transcriptomics: RNA sequence and microarray data searches

Learning Objectives

- Review the types of expression searches in VEuPathDB
- Use the differential expression, fold change and percentile search to explore gene expression in liver stage *plasmodium* infections
- Compare the expression searches to reveal advantages and disadvantages of each search
- Run a co-expression search.

Transcript expression or the abundance of an mRNA, can be determined in the laboratory with several different techniques including RNA-sequence, microarray, and RT-PCR. VEuPathDB supports these data types with several searches (see table below) and for RNA seq, expression is graphed on gene pages and 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	Micro-array
Differential Expression	Statistical analysis of studies whose experimental design includes replicate samples. A differential expression 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.	✓	

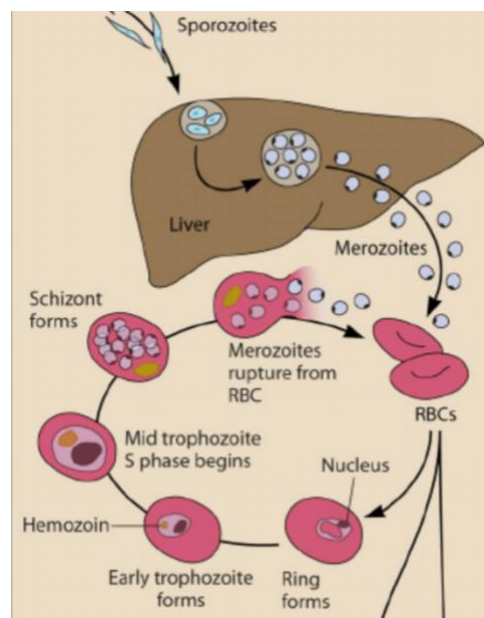
	This search identified genes whose 5' splice site location varies between samples.		
Metacycle	The MetaCycle package detects rhythmic signals from large scale time-series data, such as circadian rhythms within expression time courses, using either ARSER or JTK-Cycle. This search returns genes whose rhythmic 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 follow 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.		✓

1. Find genes that are up-regulated in the later liver stages of *Plasmodium* infection. [PlasmoDB.org](https://plasmodb.org/)

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.

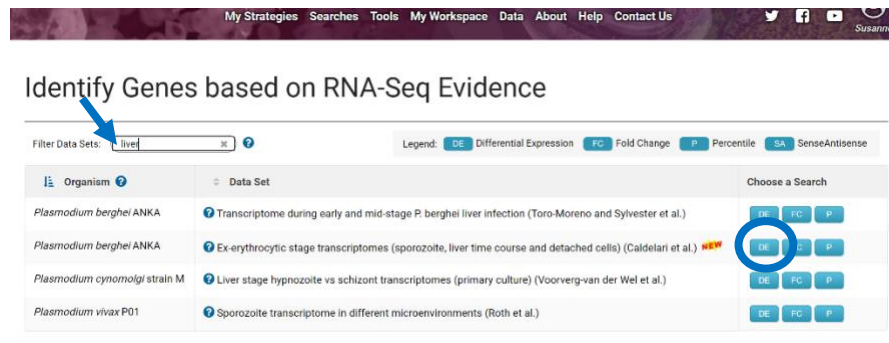
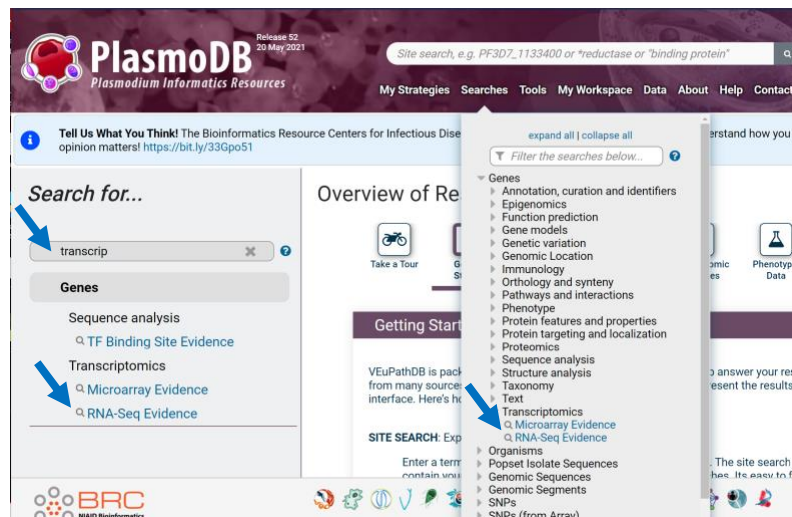
PlasmoDB contains RNA seq data from a study in the rodent model *Plasmodium berghei*, that includes a time course of liver infection as well as sporozoite and merozoite samples for comparison. ([Caledlari et al. 2019](#)) Seven samples were assayed in triplicate for RNA sequence:

1. Sporozoites
2. 6 hr liver infection
3. 24 hr liver infection
4. 48 hr liver infection
5. 54 hr liver infection
6. 60 hr liver infection
7. Merozoites (detached cells).



Use this data set to determine what genes are upregulated at least 4 fold ($p\text{-value} \leq 0.001$) at 48 hr post infection vs the sporozoite stage.

- a. Navigate to the RNA seq search page and find the data set called Ex-erythrocytic stage transcriptomes (sporozoite, liver time course and detached cells) (Caldelari et al.). Searches are available from the Search For... menu on the left side of the home page, as well as the Searches drop down menu in the header.



- b. Arrange the differential expression search to return genes that are at least 4 fold up-regulated in the 48-hour liver infection compared to sporozoites.

Differential Expression Fold Change Percentile

Identify Genes based on P. berghei ANKA Ex-erythrocytic stage transcriptomes (sporozoite, liver time course and detached cells) RNA-Seq (Differential Expression)

[Reset values](#)

Experiment

- Ex-erythrocytic stage transcriptomes (sporozoite liver time course and detached cells) unstranded

Reference Sample

- sporozoite
- Liver 6h
- Liver 24h
- Liver 48h
- Liver 54h
- Liver 60h
- DC

Comparator Sample

- sporozoite
- Liver 6h
- Liver 24h
- Liver 48h
- Liver 54h
- Liver 60h
- DC

Direction

up-regulated

fold difference >=

4

adjusted P value less than or equal to

0.001

[Get Answer](#)

[Pber ex-erythro RNAseq \(de\)](#) 7,331 Genes
 [Add a step](#)

Step 1

1,331 Genes (1,291 ortholog groups)

[Revise this search](#)

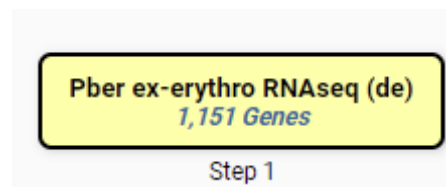
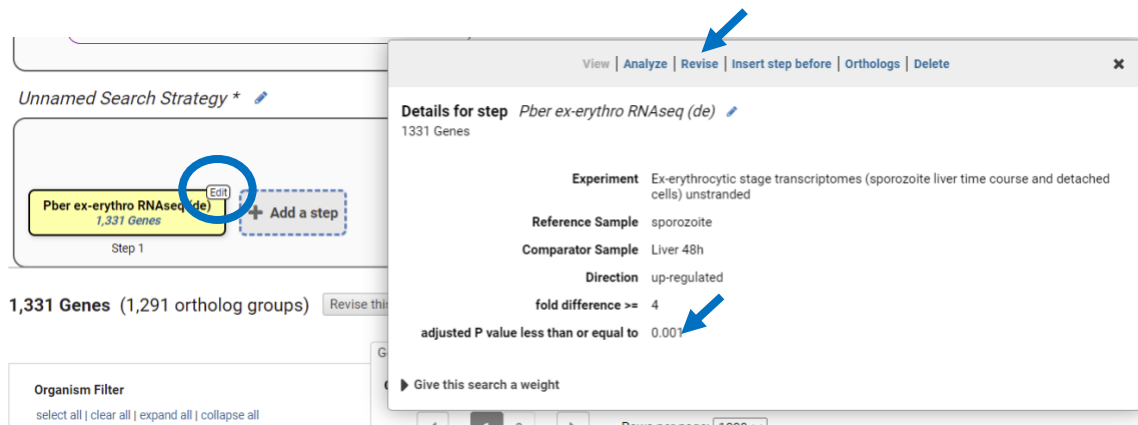
Gene Results Genome View **Analyze Results**

Genes: 1,331 Transcripts: 1,333 ☐ Show Only One Transcript Per Gene

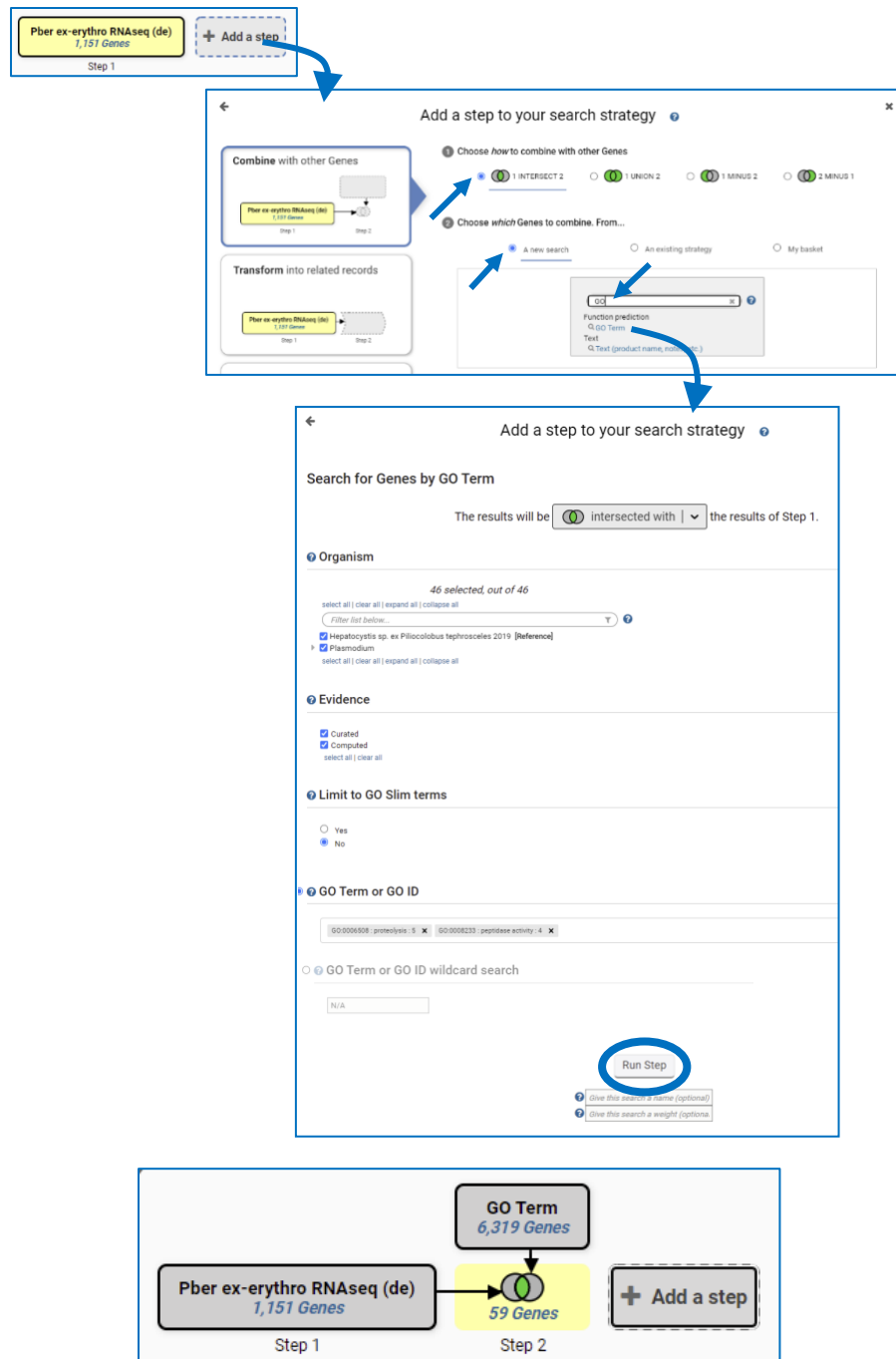
Download Add to Basket Add Columns

Gene ID	Transcript ID	Organism	Product Description	Fold Change	Adjusted P value
PBANKA_1003000	PBANKA_1003000.1	Plasmodium berghei ANKA	liver specific protein 2	1325.8	1.022E-90
PBANKA_1237900	PBANKA_1237900.1	Plasmodium berghei ANKA	mitochondrial-processing peptidase subunit alpha, putative	722.63	2.433E-10
PBANKA_1358600	PBANKA_1358600.1	Plasmodium berghei ANKA	isocitrate dehydrogenase (NADP), mitochondrial, putative	607.2	9.103E-10
PBANKA_0518900	PBANKA_0518900.1	Plasmodium berghei ANKA	conserved Plasmodium membrane protein, unknown function	592.5	8.881E-19
PBANKA_1127800	PBANKA_1127800.1	Plasmodium berghei ANKA	DnaJ protein, putative	486.84	3.223E-09
PBANKA_1338700	PBANKA_1338700.1	Plasmodium berghei ANKA	plasmepsin V, putative	382.22	4.480E-08
PBANKA_1358700	PBANKA_1358700.1	Plasmodium berghei ANKA	conserved Plasmodium protein, unknown function	355.4	1.272E-08
PBANKA_1305100	PBANKA_1305100.1	Plasmodium berghei ANKA	60S ribosomal protein L1, putative	354.09	3.799E-09

- c. How many genes were returned by the search? Do you believe these results? To convince yourself, you could browse the product description column. Are there clues that these genes are liver-specific
- d. Increase the statistical stringency of the search from $p \leq 0.001$ to $p < 0.0001$. How many genes are returned by the search now? 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.



- e. 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. 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 have these activities?



2. **Find genes that are upregulated 4 fold in any liver stage compared to sporozoites.** Hint: use the Fold change search to compare the 6, 24, 48, 54 and 60-hour time points to sporozoites.
 - a. Navigate to the RNA Seq search page and choose the Fold Change search for the Ex-erythrocytic (Caldelari et al 2019) data set as in 1a above.

- b. Arrange the fold change search to return genes that are up-regulated in the average expression across the liver stages compared to the sporozoites.

For the Experiment:

- Ex-erythrocytic stage transcriptomes (sporozoite, liver time course and detached cells) unstranded

return protein coding Genes that are up-regulated

with a Fold change ≥ 4

between each gene's minimum expression value (or a Floor of 10 reads) in the following Reference Samples:

- ☒ sporozoite
- ☐ Liver 6h
- ☐ Liver 24h
- ☐ Liver 48h
- ☐ Liver 54h
- ☐ Liver 60h
- ☐ DC

select all | clear all

and its average expression value (or the Floor selected above) in the following Comparison Samples:

- ☐ sporozoite
- ☒ Liver 6h
- ☒ Liver 24h
- ☒ Liver 48h
- ☒ Liver 54h
- ☒ Liver 60h
- ☐ DC

select all | clear all

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

Up-regulated

A maximum of four samples are shown when more than four are selected.

For each gene, the search calculates:

$$\text{fold change} = \frac{\text{average expression value in comparison}}{\text{reference expression value}}$$
 and returns genes when fold change ≥ 4 .

You are searching for genes that are up-regulated between one reference sample and at least two comparison samples.

To narrow the window, use the minimum comparison value. To broaden the window, use the maximum comparison value.

Get Answer

Pber ex-erythro RNAseq (fc)
2,098 Genes

+ Add a step

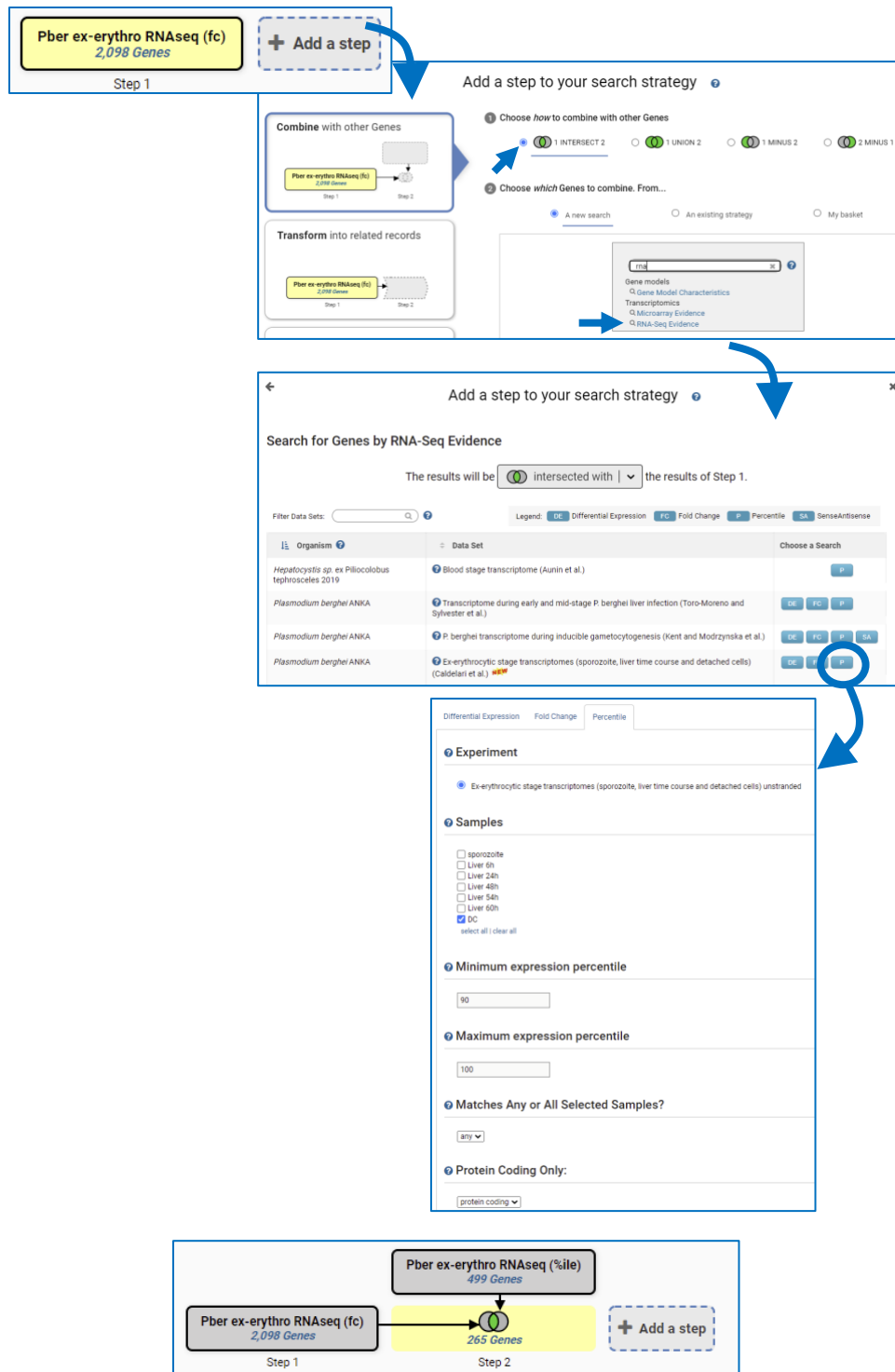
Step 1

Gene Results: 2,098 (2,098) — Show Only One Transcript Per Gene

Download Add to Bookmarks Add Columns

Gene ID	Transcript ID	Organism	Product Description	Fold Change	Chosen Ref (Base)	Chosen Comp (Base)	Pber ex-erythro RNAseq - log2 (log)
PBANKA_1203800	PBANKA_1203800.1	Plasmodium berghei AK24	liver specific protein 2	1201.7	0.33 (7.05)	1140.01	
PBANKA_0318900	PBANKA_0318900.1	Plasmodium berghei AK24	conserved Plasmodium membrane protein, unknown function	500.2	0.88 (5.06)	4491.2	
PBANKA_1203800	PBANKA_1203800.1	Plasmodium berghei AK24	DnaJ protein, putative	500.5	51.26	24963.16	

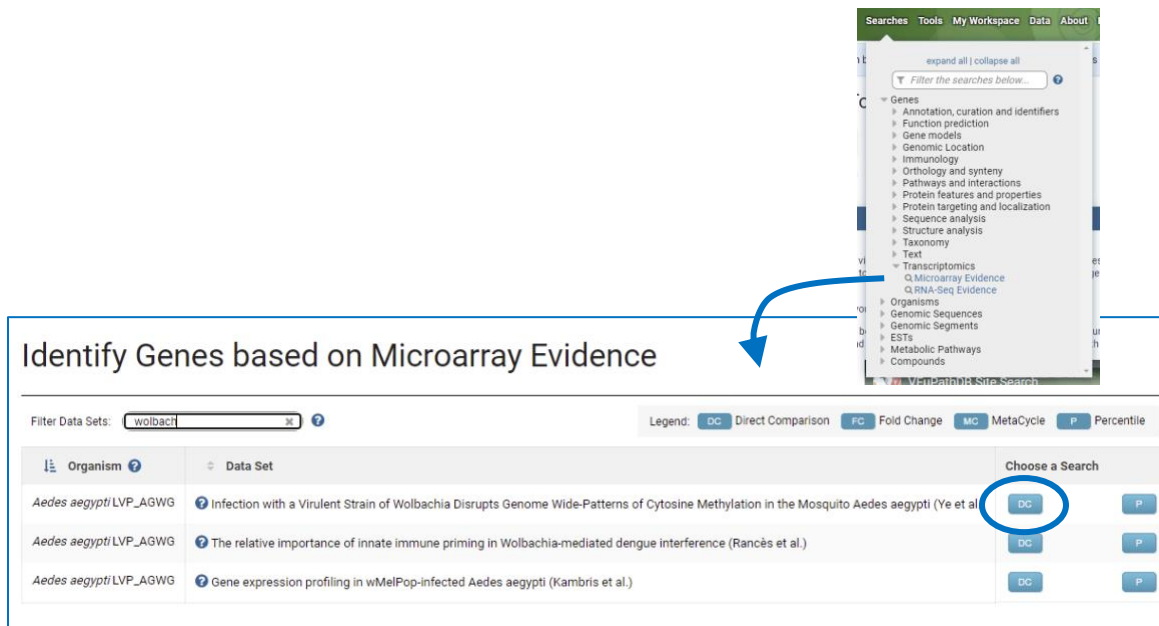
- c. Explore your results. Did the search return more genes or fewer genes than the differential expression search?
- d. Use the Add Columns to turn on the TPM graph for the 'Ex-erythrocytic stages' data set. Notice the error bars for the DNAJ protein PBANKA_1203800. Would this gene be returned by the Differential Expression search that applies statistics before returning genes?
- e. Use the Percentile search to determine what genes in this result list are also expressed in the top 10% of genes in the merozoite (detached cells) sample? Hint: Add a step to the strategy that intersects your current result with search that returns the 90-100th percentile genes of the merozoite sample.



- Find *Aedes aegypti* genes that are upregulated in both head and muscle during infection with *Wolbachia*. The *Wolbachia* strain wMelPop, which reduces longevity in *Drosophila melanogaster*, has been introduced into the Dengue virus mosquito vector, *Aedes aegypti* as a strategy to reduce

disease transmission. VectorBase has a microarray data set that compared *Wolbachia* infected and uninfected mosquito head and muscle. This exercise uses VectorBase.org.

- a. Navigate to the microarray search and choose the Direct Comparison search for the dataset titled 'Infection with a Virulent Strain of Wolbachia Disrupts Genome Wide-Patterns of Cytosine Methylation in the Mosquito *Aedes aegypti* (Ye et al.)'



Identify Genes based on Microarray Evidence

Filter Data Sets:

Legend: **DC** Direct Comparison **FC** Fold Change **MC** MetaCycle **P** Percentile

Organism	Data Set	Choose a Search
<i>Aedes aegypti</i> LVP_AGWG	Infection with a Virulent Strain of Wolbachia Disrupts Genome Wide-Patterns of Cytosine Methylation in the Mosquito <i>Aedes aegypti</i> (Ye et al.)	DC P
<i>Aedes aegypti</i> LVP_AGWG	The relative importance of innate immune priming in Wolbachia-mediated dengue interference (Rancès et al.)	DC P
<i>Aedes aegypti</i> LVP_AGWG	Gene expression profiling in wMelPop-infected <i>Aedes aegypti</i> (Kambris et al.)	DC P

- b. Initiate a search that returns genes that are upregulated 2 fold in infected head vs uninfected.

Identify Genes based on *A. aegypti* LVP_AGWG Infection with a Virulent Strain of Wolbachia Disrupts Genome Wide-Patterns of Cytosine Methylation in the Mosquito *Aedes aegypti* Microarray (direct comparison)

Experiment

☒ Infection with a Virulent Strain of Wolbachia Disrupts Genome Wide-Patterns of Cytosine Methylation in the Mosquito *Aedes aegypti*

Direction

Comparison

☒ head infected v head uninfected

Fold difference >=

Protein Coding Only:

Wolbachia infection in head an...
695 Genes

Step 1

- c. Intersect your search result with another search that returns genes upregulated 2 fold in muscle vs uninfected. Your combined result will be genes that are upregulated in head and muscle in response to *Wolbachia* infection.

← Add a step to your search strategy ⓘ

Combine with other Genes

Wolbachia infection in head an...
857 Genes

Step 1

Step 2

Transform into related records

Wolbachia infection in head an...
857 Genes

Step 1

Step 2

Use Genomic Colocation to

1 Choose how to combine with other Genes

☒ 1 INTERSECT 2 ☐ 1 UNION 2 ☐ 1 MINUS 2 ☐ 2 MINUS 1

2 Choose which Genes to combine. From...

☒ A new search ☐ An existing strategy ☐ My basket

RNA

Gene models
Q Gene Model Characteristics
Transcriptomics
Q Microarray Evidence
Q RNA-Seq Evidence

Run Step

- d. Determine enriched Molecular Function GO terms for the upregulated genes. Make sure you are viewing the combined result (the Step 2 result will be highlighted yellow) and click Analyze Result to open the Enrichment Tool. What gene functions are shared by the combined result? What biological role can you envision for these mosquito genes during the *wolbachia* infection?

Wolbachia infection in head an...
857 Genes

Wolbachia infection in head an...
857 Genes

Step 1

Step 2

394 Genes (355 ortholog groups)

Organism Filter

select all | clear all | expand all | collapse all

Hide zero counts

Search organisms...

Gene Results

Genes: 394 Transcripts: 799

Rows per page: 1000

Transcript

Genomic Location

Analyze Results

Analyze your Gene results with a to

Gene Ontology Enrichment

Find Gene Ontology terms that are enriched in your gene result. [Read More](#)

Parameters

Organism:

Ontology: ☐ Biological Processes ☒ Molecular Function ☐ Cellular Component

Evidence: ☒ Computed ☐ Curated

Limit to GO Slim terms: ☒ No ☐ Yes

P-Value cutoff: (0-1)

Submit

Analysis Results:

GO ID GO Term Genes in the input with this term Genes in your result with this term Percent of input genes in your result Fold enrichment Odds ratio P-value Benjamin

GO:0003024	catalytic activity	3791	705	4.5	1.62	2.40	5.15e-14	1.22e-11
GO:0014787	hydrolase activity	1628	96	5.9	2.12	2.78	6.77e-14	1.22e-11
GO:0004175	endorphinase activity	472	40	9.5	3.40	4.18	1.90e-13	1.81e-11
GO:0004282	serine-type endopeptidase activity	346	38	11.0	3.98	4.83	2.21e-13	1.87e-11

4. **Find genes that are likely co-expressed with An04g07430, an *Aspergillus niger* protein coding gene with little functional annotation.** By finding genes that are expressed at the same time as An04g07430, we may find clues about its function and the biological processes that it participates in. This exercise uses [FungiDB](#).
 - a. Navigate to the microarray searches in FungiDB and choose the Coexpression search for the data set titled *Aspergillus niger* gene co-expression network (Vera Meyer). [Schape et al Nucleic Acids Research 2019](#). This data are the results of a meta-analysis of 155 publicly available transcriptomics analyses for *A. niger*, which were used to generate a genome-level co-expression network and sub-networks for >9,500 genes.
 - b. Run the search to find the co-expression network for An04g07430.

The screenshot shows the FungiDB interface. On the left, a sidebar lists search categories: Genes, Annotation, curation and identifiers, Function prediction, Gene models, Genetic variation, Genomic Location, Immunology, Orthology and synten, Pathways and interactions, Phenotype, Protein targeting and localization, Proteomics, Sequence analysis, Structure analysis, Taxonomy, Text, Transcriptomics, Microarray Evidence, RNA-Seq Evidence, and Organisms. The main search area is titled "Identify Genes based on Microarray Evidence". It includes a "Filter Data Sets" section with a search bar and a legend for Coexpression (C), Direct Comparison (DC), Fold Change (FC), and Percentile (P). Below this is a table of data sets:

Organism	Data Set	Choose a Search
<i>Aspergillus fumigatus</i> Af293	Response to hypoxia (Barker et al. 2012)	FC P
<i>Aspergillus niger</i> CBS 513.88	<i>Aspergillus niger</i> gene co-expression network (Vera Meyer)	C
<i>Candida albicans</i> SC5314	Antifungal Benzimidazole Derivative Response (Steffen Rupp)	DC P

Below the table is a section titled "Identify Genes based on Coexpression". It includes a "Reset values" button and a "Gene ID input set" section. The "Gene ID input set" section has four options: "Enter a list of IDs or text" (selected), "Upload a text file", "Copy from My Basket", and "Copy from My Strategy". The "Enter a list of IDs or text" option has a text input field containing "An04g07430". Below this is a "Correlation" section with a dropdown menu set to "Positive Correlation". Below that is a "Spearman coefficient (greater or equal to)" section with a text input field containing "0.75". At the bottom right of this section is a "Get Answer" button. Below the "Get Answer" button is a summary box showing "Coexpression 107 Genes" and a button to "Add a step".

- c. What genes share the co-expression profile of An04g07430? Several genes have a correlation coefficient of 0.85. What are these genes? Visit their gene pages to learn more.

- d. Scan the product description column for genes with known functions. Use the Column Histogram tool to view a word cloud of the product descriptions in the column. Set the rank range to 25-50. What words occur most often in the product descriptions of An04g07430 co-expressed genes?

