

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 biological replicates. 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.	✓	

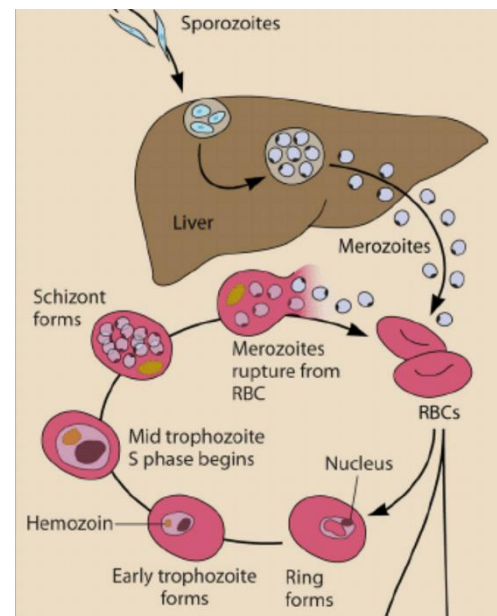
	This search identified genes whose 5' splice site location varies between samples.		
Metacycle	The MetaCycle package detects rhythmic signals from large scale time-series data, such as circadian rhythms within expression time courses, using either ARSER or JTK-Cycle. This search returns genes whose rhythmic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

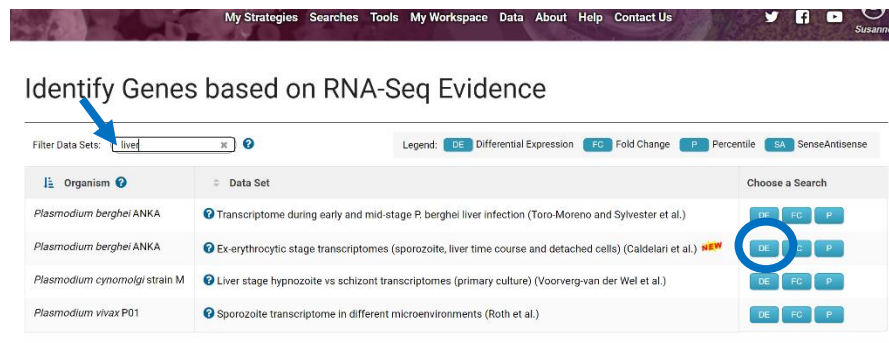
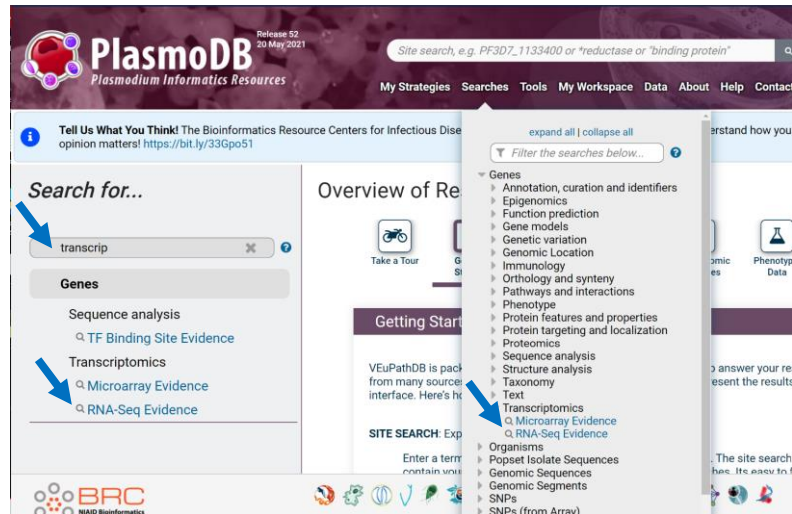
PlasmoDB contains RNA seq data from a study in the rodent model *Plasmodium berghei*, that includes a time course of liver infection as well as sporozoite and merozoite samples for comparison. ([Caledlari et al. 2019](#)) Seven samples were assayed in triplicate for RNA sequence:

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-value ≤ 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 with a p-value of $p < 0.001$.

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)

1,331 Genes

+ Add a step

Step 1

- 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. Click the Edit icon and choose revise from the options panel.

The screenshot shows a search strategy named "Unnamed Search Strategy" with a single step labeled "Pber ex-erythro RNAseq (de)" which has returned 1,331 genes. A blue circle highlights the "Edit" icon next to the search step. A details panel for this step is open, showing the following parameters:

Details for step: Pber ex-erythro RNAseq (de)	
Experiment	Ex-erythrocytic stage transcriptomes (sporozoite liver time course and detached cells) unstranded
Reference Sample	sporozoite
Comparator Sample	Liver 48h
Direction	up-regulated
fold difference >=	4
adjusted P value less than or equal to	0.001

A blue arrow points to the "Revise" button in the top navigation bar of the details panel. Another blue arrow points to the "0.001" value in the "adjusted P value" field, indicating where to click to change it.

The screenshot shows the updated search strategy "Pber ex-erythro RNAseq (de)" which now returns 1,151 genes. The step is labeled "Step 1".

- 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 both of these activities?

Pber ex-erythro RNAseq (de)
1,151 Genes
Step 1

+ Add a step

Add a step to your search strategy

Combine with other Genes

Transform into related records

1 Choose how to combine with other Genes

2 Choose which Genes to combine. From...

GO

Add a step to your search strategy

Search for Genes by GO Term

The results will be ☐ intersected with ☐ the results of Step 1.

Organism

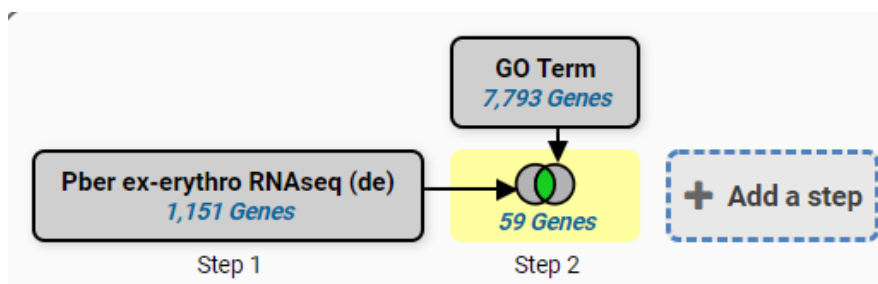
Evidence

Limit to GO Slim terms

GO Term or GO ID

GO Term or GO ID wildcard search

Run Step



2. Find genes that are upregulated 4 fold in any liver stage compared to sporozoites. Hint: Search the same data set but 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 Genes

that are

with a Fold change \geq

between each gene's expression value
(or a Floor of)

in the following Reference Samples

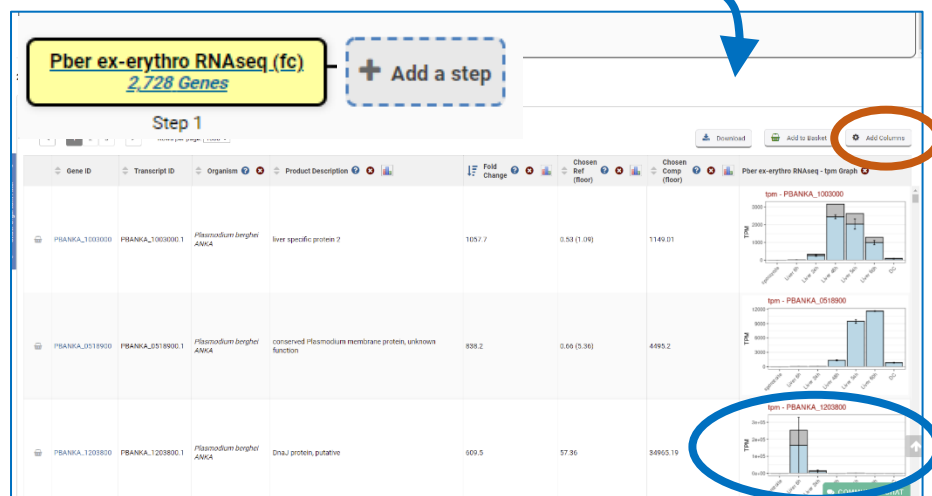
☒ sporozoite
☐ Liver 6h
☐ Liver 24h
☐ Liver 48h
☐ Liver 54h
☐ Liver 60h

and its expression value
(or the Floor selected above)

in the following Comparison Samples

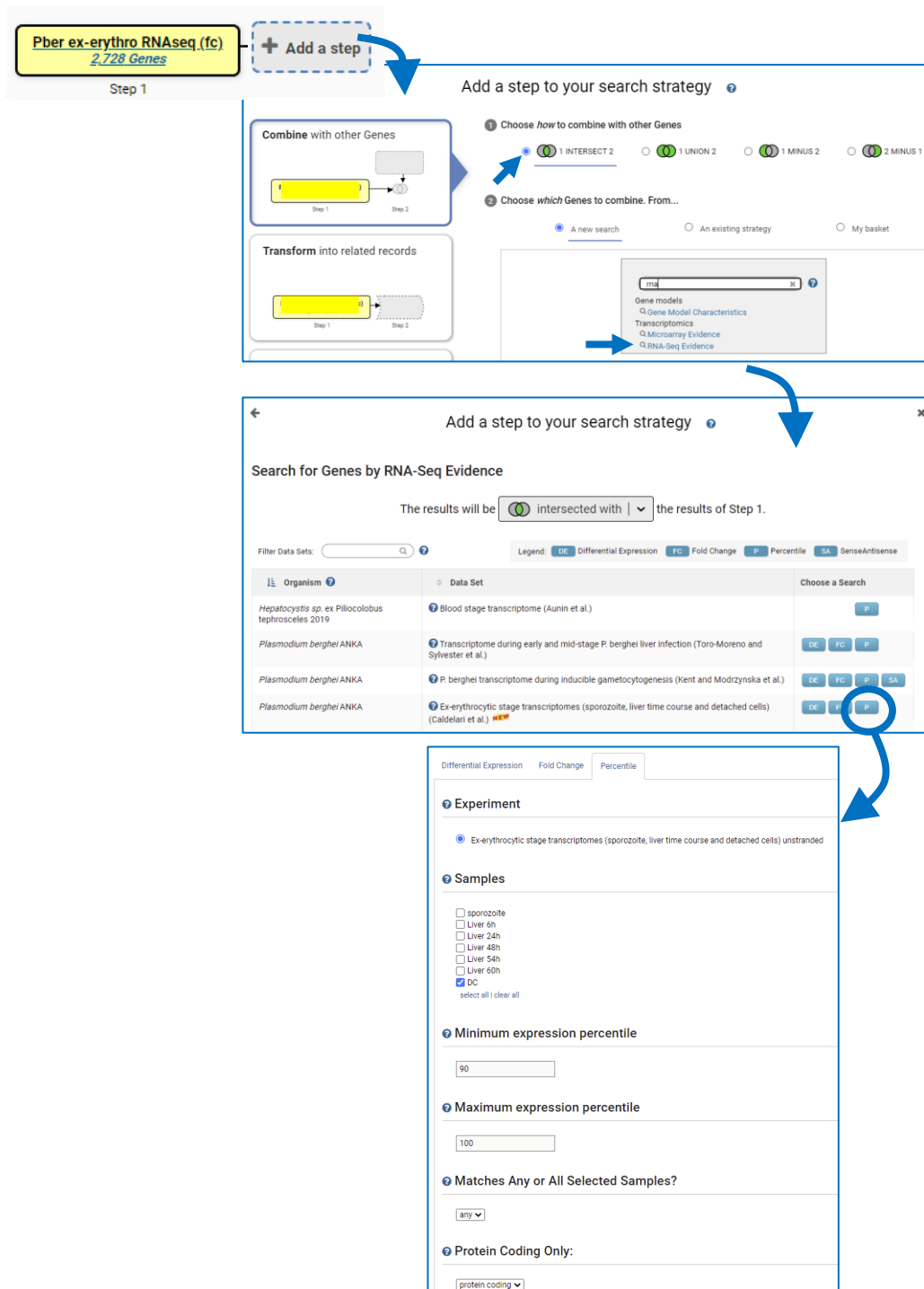
☐ sporozoite
☒ Liver 6h
☒ Liver 24h
☒ Liver 48h
☒ Liver 54h
☒ Liver 60h

Get Answer



- 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 are also expressed in the top 10% of genes in the merozoite (detached cells, DC) 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 (DC) sample.



Step 1
Pber ex-erythro RNAseq (fc)
2,728 Genes

Add a step

Add a step to your search strategy

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

Search for Genes by RNA-Seq Evidence

The results will be ☒ intersected with the results of Step 1.

Filter Data Sets:

Legend: ☒ DE Differential Expression ☐ FC Fold Change ☐ P Percentile ☐ SA SenseAntisense

Organism	Data Set	Choose a Search
Hepatoctyitis sp. ex Piliocolobus tephrosceles 2019	Blood stage transcriptome (Aunin et al.)	<input type="button" value="P"/>
Plasmodium berghei ANKA	Transcriptome during early and mid-stage P. berghei liver infection (Toro-Moreno and Sylvester et al.)	<input type="button" value="DE"/> <input type="button" value="FC"/> <input type="button" value="P"/>
Plasmodium berghei ANKA	P. berghei transcriptome during inducible gametocytogenesis (Kent and Modrzynska et al.)	<input type="button" value="DE"/> <input type="button" value="FC"/> <input type="button" value="P"/> <input type="button" value="SA"/>
Plasmodium berghei ANKA	Ex-erythrocytic stage transcriptomes (sporozoite, liver time course and detached cells) (Caldelari et al.)	<input type="button" value="DE"/> <input type="button" value="FC"/> <input type="button" value="P"/>

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

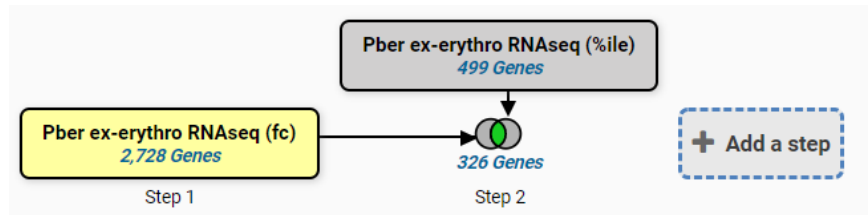
Samples
☐ sporozoite
☐ Liver 6h
☐ Liver 24h
☐ Liver 48h
☐ Liver 54h
☐ Liver 60h
☒ DC
select all | clear all

Minimum expression percentile

Maximum expression percentile

Matches Any or All Selected Samples?

Protein Coding Only:



3. **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**.

Experiment

● Infection with a Virulent Strain of *Wolbachia* Disrupts Genome Wide-Patterns of Cytosine Methylation

Direction

up-regulated

Comparison

● head infected v head uninfected
○ muscle infected v muscle uninfected

Fold difference >=

2.0

Protein Coding Only:

protein coding

Get Answer

Wolbachia infection in head an...
695 Genes

+ Add a step

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

1 Choose *how* to combine with other Genes

● 1 INTERSECT 2 ○ 1 UNION 2 ○ 1 MINUS 2

2 Choose *which* Genes to combine. From...

● A new search ○ An existing strategy

microd

Transcriptomics
Q, Microarray Evidence

Experiment

● Infection with a Virulent Strain of *Wolbachia* Disrupts Genome Wide-Patterns of Cytosine Methylation

Direction

up-regulated

Comparison

○ head infected v head uninfected
● muscle infected v muscle uninfected

Fold difference >=

2.0

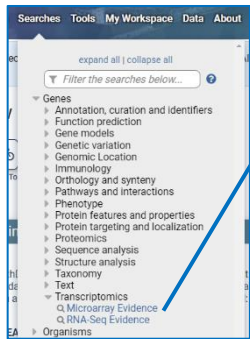
Protein Coding Only:

protein coding

- 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?

The screenshot displays the Gene Ontology Enrichment tool interface. At the top, a workflow diagram shows two steps: 'Step 1' with 'Wolbachia infection in head an...' (695 Genes) and 'Step 2' with 'Wolbachia infection in head an...' (827 Genes). The combined result for Step 2 is highlighted in yellow and labeled '394 Genes'. Below the workflow, a box shows '394 Genes (355 ortholog groups)' and tabs for 'Gene Results', 'Genome View', and 'Gene Ontology Enrichment*'. A blue circle highlights the 'Analyze Results' button. To the right, a box titled 'Analyze your Gene results with a tool' shows a 'GO' logo and 'Gene Ontology Enrichment'. Below this, the 'Gene Ontology Enrichment' section is shown with parameters: Organism (Aedes aegypti LVP_AGWG), Ontology (Molecular Function selected), Evidence (Computed and Curated selected), Limit to GO Slim terms (No selected), and P-Value cutoff (0.05). A 'Submit' button is at the bottom.

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 and default values for the other parameters.



Identify Genes based on Microarray Evidence

Filter Data Sets: Legend: C Coexpression DC Direct Comparison FC Fold Change P Percentile

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

Identify Genes based on Coexpression

Reset values

Gene ID input set

☒ Enter a list of IDs or text: An04g07430

☐ Upload a text file: Choose File No file chosen
Maximum size 10MB. The file should contain the list of IDs.

☐ Copy from My Basket: 0 records will be copied from your basket.

☐ Copy from My Strategy: Pyrimidine metabolism (ec00240) (KEGG) (60 records)

Correlation

Positive Correlation

Spearman coefficient (greater or equal to)

Get Answer

Coexpression
107 Genes

+ Add a step

Step 1

- 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?

107 Genes (102 ortholog groups) [Revise this search](#)

Gene Results **Genome View** **Analyze Results**

1 2 3 ... 6 Rows per page: 20 [Download](#) [Add to Basket](#) [Add Columns](#)

Gene ID	Transcript ID	Organism	Product Description	Input ID	Minimum coefficient	Ma	cor
Ortholog of A. nidulans FGSC...							

Word Cloud

Word Cloud Data

Filter words by rank: 25 to 50

Sort by: ☒ Rank ☐ A-Z

Mouse over a word to see its occurrence in the data

transport protein reduction
sydowii formation NRRL CBS acting
metabolic mRNA regulation ATCC carrier DNA
donors oxidation transporter versicolor component
electron function heme integral iron paired transcription

An01g14000	An01g14000-T	Aspergillus niger CBS 513.88	transmembr... 3-oxoacyl[acyl-carrier-protein] reductase	An04g07430	0.8	0.8
------------	--------------	------------------------------	---	------------	-----	-----