

Strategies 2

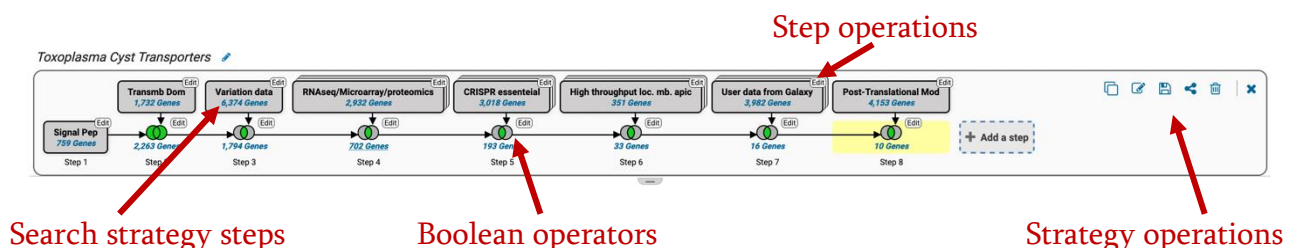
Data Integration Through Search Strategies

This exercise illustrates how to combine search results from different data types and how to effectively explore the results. **Specific objectives include:**

1. Understanding search strategy functions including adding/revising/deleting steps, copying search strategies, and saving and sharing strategies.
2. Interacting with gene results and adding columns
3. Navigating transcriptomic searches
4. Exploring proteomics data
5. Exploring subcellular localization data
6. Exploring genome wide CRISPR data
7. Exploring variation data
8. Leveraging orthology searches
9. Running enrichment analyses

Search strategies

Search strategies in VEuPathDB resources allow you to combine results from different datatype searches using Boolean operators (e.g. Intersect, union, minus). Search strategies enable you to develop *in silico* experiments based on data from the species of interest or from other species (or strains) by leveraging orthology.



Getting started with your first search strategy

There are a few things to consider before developing a search strategy:

1. What is your question? Or what are you trying to find out? (overall strategy)
2. Can you break down your question into smaller components? (strategy steps)
3. What data or analyses can be used to answer the various components of your main question?
4. How will you combine the different components of your question? I.e. Which Boolean operators.

Example question

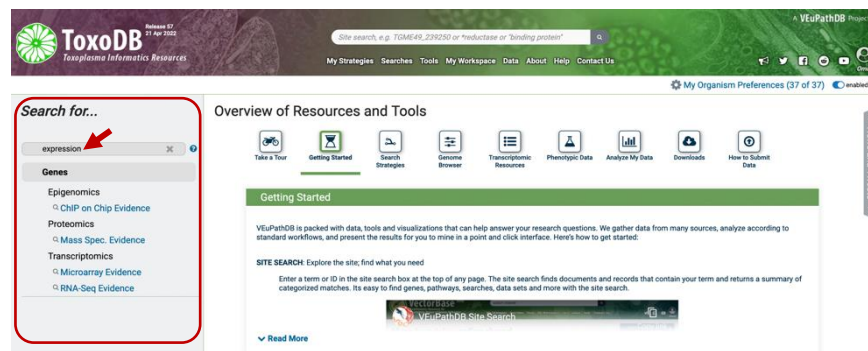
Big question: I would like to identify bradyzoite/tissue cyst specific therapeutic targets.

Let's break it down:

1. How do I identify genes whose expression is upregulated in bradyzoites?
 - a. Does upregulated mean a gene is not expressed in other stages?
 - b. How do you remove genes that are expressed in other stages?
2. Should I exclude expression from other stages? How can I do this?
3. How do I identify genes that have a specific type of variation?
4. How do I leverage orthology to define phyletic pattern?
5. What about essentiality? How do I find the genes that are important for parasite fitness?

Running your first search

1. Explore the data available in ToxoDB. What data can tell you about expression timing of genes? Expand the menu on the left-hand side of the home page and look for datatypes that would tell you about expression. Hint: try filtering the searches with a key work like “expression”.



2. Explore the RNA-Seq evidence data. Are there any experiments that tell you about bradyzoite expression? Try filtering the datasets using a keywords like “bradyzoite” or “differentiation”.

Identify Genes based on RNA-Seq Evidence

Filter Data Sets:

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

Organism	Data Set	Choose a Search
Besnoitia besnoiti strain Bb-Ger1	Tachyzoite and tissue cyst transcriptomes (Ramakrishnan et al.)	<input checked="" type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA
Emeria tenella strain Houghton	Life Cycle Stages Transcriptomes (Reid)	<input checked="" type="checkbox"/> FC <input type="checkbox"/> P
Emeria tenella strain Houghton	Gametocytes vs 2 Asexual Stages (Walker et al.)	<input checked="" type="checkbox"/> DE <input checked="" type="checkbox"/> FC <input type="checkbox"/> P
Emeria tenella strain Houghton	Transcriptome of E. tenella from infected chicken caecal tissues (Sandholt et al 2021)	<input checked="" type="checkbox"/> DE <input checked="" type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA
Emeria tenella strain Houghton	Emeria tenella transcriptome during infection in chicken macrophage-like cells (Sandholt et al 2021)	<input checked="" type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA
Neospora caninum Liverpool	Transcriptomes of virulent and avirulent N. caninum isolates during bovine infection (Horcajo et al.)	<input checked="" type="checkbox"/> DE <input checked="" type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA
Neospora caninum Liverpool	Tachyzoite Transcriptome Days 3 and 4 (Reid et al.)	<input checked="" type="checkbox"/> DE <input checked="" type="checkbox"/> FC <input type="checkbox"/> P
Neospora caninum Liverpool	Transcriptomes of virulent and avirulent N. caninum strains (Horcajo et al.)	<input checked="" type="checkbox"/> DE <input checked="" type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA
Toxoplasma gondii ME49	Tachyzoite Transcriptome Time Series (GT1) (Gregory)	<input checked="" type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA
Toxoplasma gondii ME49	Feline enterocyte, tachyzoite, bradyzoite stage transcriptome (Hehl, Ramakrishnan et al.)	<input checked="" type="checkbox"/> DE <input checked="" type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA
Toxoplasma gondii ME49	Bradyzoite in vivo transcriptome (M4) (Buchholz et al.)	<input type="checkbox"/> P
Toxoplasma gondii ME49	Tachyzoite Transcriptome Time Series (ME49) (Gregory)	<input checked="" type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA

- Select the fold change search for the “Feline enterocyte, tachyzoite, bradyzoite stage transcriptome (Hehl, Ramakrishnan et al.)”.

Identify Genes based on RNA-Seq Evidence

Filter Data Sets:

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

Organism	Data Set	Choose a Search
Toxoplasma gondii ME49	Feline enterocyte, tachyzoite, bradyzoite stage transcriptome (Hehl, Ramakrishnan et al.)	<input checked="" type="checkbox"/> DE <input checked="" type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA
Toxoplasma gondii ME49	Bradyzoite in vivo transcriptome (M4) (Buchholz et al.)	<input type="checkbox"/> P
Toxoplasma gondii ME49	Stage-specific RNA-sequencing in Toxoplasma gondii (Waldman et al.)	<input checked="" type="checkbox"/> DE <input checked="" type="checkbox"/> FC <input type="checkbox"/> P
Toxoplasma gondii ME49	Bradyzoite in vitro Transcriptome (ME49) (Sibley/Gregory)	<input type="checkbox"/> P
Toxoplasma gondii ME49	Mouse brain bradyzoite transcriptomes at 28, 90, 120 days post infection (Garfoot et al.)	<input checked="" type="checkbox"/> DE <input checked="" type="checkbox"/> FC <input type="checkbox"/> P

- Configure this search to identify genes that are upregulated by 2-fold in tissue cysts compared to all other stages (use average expression values).

Identify Genes based on T. gondii ME49 Feline enterocyte, tachyzoite, bradyzoite stage transcriptome RNA-Seq (fold change)

For the Experiment

☒ Feline enterocyte, tachyzoite, bradyzoite stage transcriptome toxo Transcriptomes of enteroepithelial stages - Sense

☐ Feline enterocyte, tachyzoite, bradyzoite stage transcriptome toxo Transcriptomes of enteroepithelial stages - Antisense

return ☒ Genes

that are

with a **Fold change**

between each gene's ☒ expression value

(or a Floor of reads ☐)

in the following **Reference Samples**

☒ EES1
☒ EES2
☒ EES3
☒ EES4
☒ EES5
☐ Tachyzoites
☐ Tissue cysts

select all | clear all

and its ☒ expression value

(or the Floor selected above)

in the following **Comparison Samples**

☐ EES3
☐ EES4
☐ EES5
☐ Tachyzoites
☒ Tissue cysts

select all | clear all

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

Up-regulated

For each gene, the search calculates:

$$\text{fold change} = \frac{\text{comparison expression value}}{\text{average expression value in reference}}$$

and returns genes when **fold change** ≥ 2 .

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

To narrow the window, use the maximum reference value. To broaden the window, use the maximum reference value.

5. Add a step and combine these results with results from another experiments containing bradyzoite samples. For example, try selecting the “Stage-specific RNA-sequencing in *Toxoplasma gondii* (Waldman et al.)”. Configure this search to identify genes that are upregulated by 2-fold when comparing 48hr bradyzoites to 24hr tachyzoites.
 - Did you use an intersect or a union operator? How would your results change if you use one or the other? Is one better than the other in this case?



6. How about combining this with data from a microarray experiment? Add a step, go to the microarray data section and select, for example, the “Bradyzoite Differentiation (3-day time series)(Pru) (Buchholz, Fritz and Boothroyd et al.)” experiment. Configure the search to identify genes that are upregulated by 2 fold between time 0 and the other time points.

Identify Genes based on *T. gondii* ME49 Bradyzoite Differentiation (3-day time series)(Pru) Microarray (fold change)

For the Experiment

☒ Bradyzoite Differentiation (3-day time series)(Pru)

return Genes

that are

with a **Fold change** \geq 2.0

between each gene's expression value

in the following **Reference Samples**

☒ 0 days

☐ 2 days

☐ 3 days

☐ 4 days

[select all](#) | [clear all](#)

and its expression value

in the following **Comparison Samples**

☐ 0 days

☒ 2 days

☒ 3 days

☒ 4 days

[select all](#) | [clear all](#)

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

Up-regulated

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** \geq 2.

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.

7. Add a step and combine any genes that have mass spec evidence from the “Mouse brain bradyzoite proteomics time course (Garfoot et al.)” experiment.

Identify Genes based on Mass Spec. Evidence

Experiments and Samples

5 selected, out of 101

[select only these](#) | [add these](#) | [clear these](#)

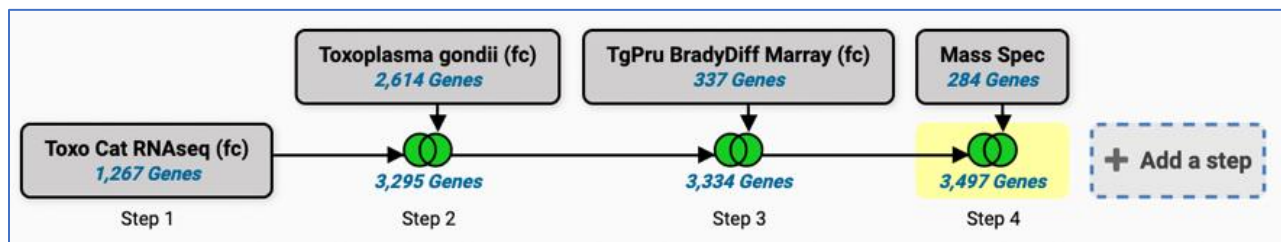
brady

- ☒ **Toxoplasma gondii**
 - ☒ **Toxoplasma gondii ME49**
 - ☒ Mouse brain bradyzoite proteomics time course (Garfoot et al.)
 - ☒ Bradyzoites 21 days
 - ☒ Bradyzoites 28 days
 - ☒ Bradyzoites 3 months
 - ☒ Bradyzoites 4 months
 - ☒ Bradyzoites 5 months

[select only these](#) | [add these](#) | [clear these](#)

Minimum Number of Unique Peptide Sequences

10



8. Now let's exclude any gene that is highly expressed in tachyzoite and sexual stages. To do this, add a step and select the **percentile** search for the "Feline enterocyte, tachyzoite, bradyzoite stage transcriptome (Hehl, Ramakrishnan et al.)" dataset. Configure the search to exclude any gene that is expressed in all stages except tissue cyst at 70 or higher percentile.

- What does the "Matches Any or All Selected Samples?" parameter do? Which option is more stringent, any or all?
- Which Boolean operator did you use?

Experiment

- ☒ Feline enterocyte, tachyzoite, bradyzoite stage transcriptome toxo Transcriptomes of enteroepithelial stages - Sense
- ☐ Feline enterocyte, tachyzoite, bradyzoite stage transcriptome toxo Transcriptomes of enteroepithelial stages - Antisense

Samples

- ☒ EES1
- ☒ EES2
- ☒ EES3
- ☒ EES4
- ☒ EES5
- ☒ Tachyzoites
- ☐ Tissue cysts

[select all](#) | [clear all](#)

Minimum expression percentile

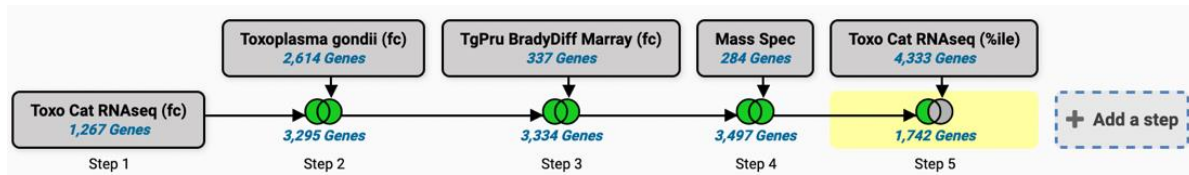
70

Maximum expression percentile

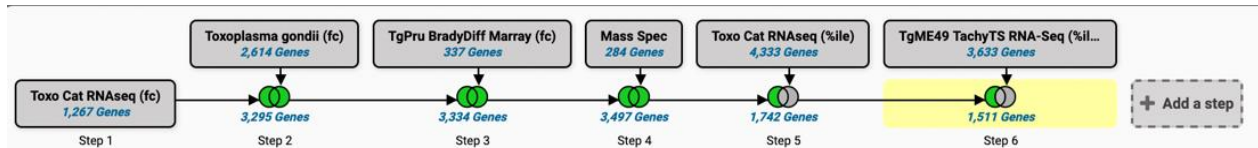
100

Matches Any or All Selected Samples?

any



9. You can exclude additional genes that show expression in stages you are not interested in from other experiments using the above method. Try this with another experiment – for example the “Tachyzoite Transcriptome Time Series (ME49) (Gregory)”.



10. Now that we have a list of genes that are upregulated in bradyzoites and are likely not highly expressed in other stages, let’s find out which of these have more than 10 non-synonymous SNPs. To do this, add a step and find the search for genes by SNP characteristics.

← Add a step to your search strategy ? ×

Combine with other Genes

TgME49 TachyTS RNA-Seq (%ile) 3,633 Genes

1,511 Genes

Step 6

Step 7

Transform into related records

TgME49 TachyTS RNA-Seq (%ile) 3,633 Genes

1,511 Genes

Step 6

Step 7

Use Genomic Colocation to combine with other features

TgME49 TachyTS RNA-Seq (%ile) 3,633 Genes

1,511 Genes

Step 6

Step 7

1 Choose how to combine with other Genes

☒ 6 INTERSECT 7 ☐ 6 UNION 7 ☐ 6 MINUS 7 ☐ 7 MINUS 6

2 Choose which Genes to combine. From...

☒ A new search ☐ An existing strategy ☐ My basket

vari

Genetic variation

- ☐ Copy Number (CNV)
- ☐ Copy Number Comparison (CNV)
- ☒ SNP Characteristics

- Configure the SNP search to find genes to select all samples aligned to *T. gondii* ME49.

Search for Genes by SNP Characteristics

The results will be  intersected with  the results of Step 6.

Organism

Toxoplasma gondii ME49

Set of Samples

65 Set of Samples Total

expand all | collapse all

Find a variable

Collection year

Country

obsolete_average mapping coverage

proportion mapped reads

Sample

Sample source

Organism under investigation

65 of 65 Set of Samples selected

Data Set

Data Set

Keep checked values at top

65 (100%) of 65 Set of Samples have data for this variable

<input checked="" type="checkbox"/> Data Set	Remaining Set of Samples	Set of Samples	Distribution	%
	65 (100%)	65 (100%)		
<input checked="" type="checkbox"/> Aligned genomic sequence reads - RH Strain	1 (2%)	1 (2%)		(100%)
<input checked="" type="checkbox"/> Aligned genomic sequence reads - White Paper Strains	62 (95%)	62 (95%)		(100%)
<input checked="" type="checkbox"/> Toxoplasma gondii ME49 Genome Sequence and Annotation	1 (2%)	1 (2%)		(100%)
<input checked="" type="checkbox"/> Toxoplasma gondii strain CZ clone H3 aligned genome sequence	1 (2%)	1 (2%)		(100%)

- Next set the percent isolates with a SNP call to 60, the SNP type to non-synonymous and the number of SNPs of this type to ≥ 10 .

Read frequency threshold

80%

Minor allele frequency \geq

0

Percent isolates with a base call \geq

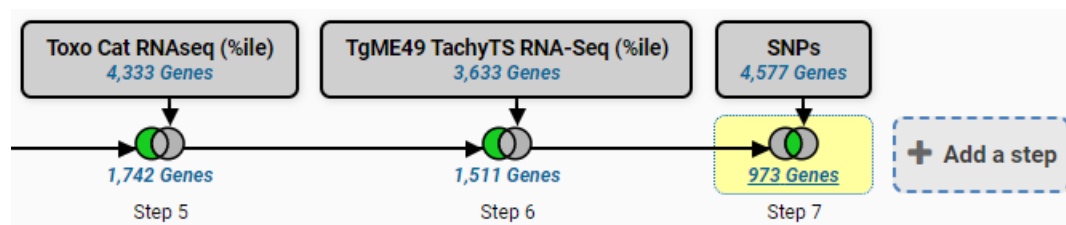
60

SNP Class

Non-Synonymous

Number of SNPs of above class \geq

10



11. Now let's determine how many of these genes do not have orthologs in mammals. Add a step and find the search called "Orthology Phylogenetic Profile".

Add a step to your search strategy

Combine with other Genes

Step 7: 4,877 Genes
Step 8: 973 Genes

Transform into related records

Step 7: 4,877 Genes
Step 8: 973 Genes

Use Genomic Colocation to combine with other features

Step 7: 4,877 Genes
Step 8: 973 Genes

1 Choose how to combine with other Genes

☒ 7 INTERSECT 8 ☐ 7 UNION 8 ☐ 7 MINUS 8 ☐ 8 MINUS 7

2 Choose which Genes to combine. From...

☒ A new search ☐ An existing strategy ☐ My basket

Search: phyl
Results: Orthology and synteny, Orthology Phylogenetic Profile

- There are different ways to configure this search depending on which Boolean operator you use. If you use the intersect operator, then configure the search to return all genes in ToxoDB that do not have orthologs in mammals.

select all | clear all | expand all | collapse all

Filter list below...

☒ Eimeriidae
☒ Sarcocystidae

select all | clear all | expand all | collapse all

Select orthology profile

Click on ☒ to determine which organisms to include or exclude in the orthology profile.
(☐ = no constraints | ☒ = must be in group | ☒ = must not be in group | ☒ = mixture of constraints)

☒ All Organisms expand all | collapse all

☒ Bacteria (BACT)

☒ Firmicutes (FIRM)

☒ Proteobacteria (PROT)

☒ Other Bacteria (OBAC)

☒ Archaea (ARCH)

☒ Nitrosopumilus maritimus (strain SCM1) (nmar)

☒ Euryarchaeota (EURY)

☒ Crenarchaeota (CREN)

☒ Nanoarchaeota (NANO)

☒ Korarchaeota (KORA)

☒ Eukaryota (EUKA)

☒ Alveolates (ALVE)

☒ Amoebozoa (AMOE)

☒ Euglenozoa (EUGL)

☒ Viridiplantae (VIRI)

☒ Fungi (FUNG)

☒ Metazoa (META)

☒ Nematodes (NEMA)

☒ Arthropoda (ARTH)

☒ Chordata (CHOR)

☒ Branchiostoma floridae (Florida lancelet) (Amphioxus) (bflo)

☒ Xenopus tropicalis (Western clawed frog) (Silurana tropicalis) (xtro)

☒ Actinopterygii (ACTI)

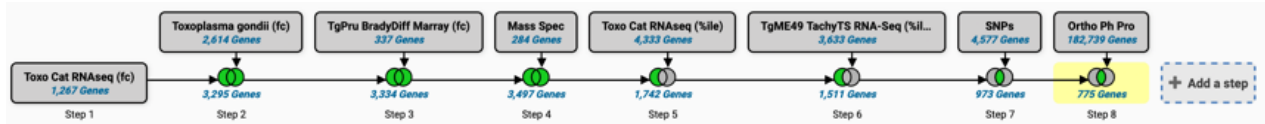
☒ Aves (AVES)

☒ Mammalia (MAMM)

☒ Tunicates (TUNI)



☒ Other Metazoa (OMET)

☒ Other Eukaryota (OEUK)



12. As a final step let's determine which of these genes are essential based on the genome wide CRISPR screen from the Lourido lab. Add a step and find the CRISPR phenotype search. Set the phenotype score \leq to -2.

Search for Genes by CRISPR Phenotype

The results will be  intersected with  the results of Step 8.

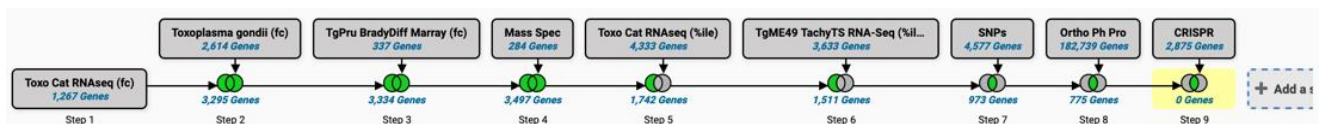
 Phenotype Score \geq

-6.89

 Phenotype Score \leq

-2 

Run Step



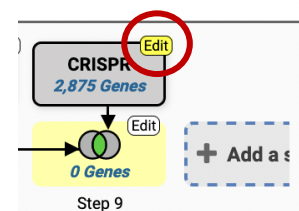
- How many results did you get? Is this surprising? Why do you think you got 0 results?
 - How can you get over the problem observed above? Is there a tool that would allow you to convert *T. gondii* GT1 genes to *T. gondii* ME49 genes?
13. Hover over the CRISPR step and click on the edit icon. In the popup click on the “orthologs” option and select ME49 from the list of organisms to transform to.
- Did this improve the results?

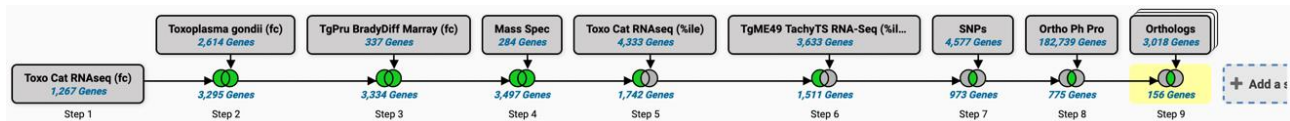
View | Analyze | Revise | Make nested strategy | Insert step before | **Orthologs** | Delete

Details for step CRISPR 
2875 Genes

Phenotype Score \geq -6.89
Phenotype Score \leq -2

► Give this search a weight





14. Explore the genes in your result list. Are there any interesting genes that you might pursue further in the lab?
15. How many hypothetical genes are in your results? A quick way to find out is to click on the graph icon in the Product Description column heading. This generates and



interactive word cloud. Hover over the word hypothetical to see the number.

16. What can you do to figure out what some of these hypothetical genes do? Here are a couple of suggestions:

- Add a column for InterPro domain descriptions. Click on add column, search for InterPro and add the appropriate column. Did this give you an idea for possible functions for some of the hypothetical genes?

- Add another column for the hyper_LOPIT subcellular localization data. Add the column called “Predicted Location (TAGM-MAP)”. Did this reveal some possible clues about the some of the other hypotheticals?

17. Do your results contain an enrichment of certain functions? Try some of the analysis tools available in the analysis tab.

Unnamed Search Strategy

Step 1: Toxo Cat RNAseq (fc) 2,247 Genes
 Step 2: Toxoplasma gondii (fc) 2,247 Genes
 Step 3: TyGru BradyDyff Murray (fc) 2,247 Genes
 Step 4: Mass Spec 284 Genes
 Step 5: Toxo Cat RNAseq (%ile) 1,742 Genes
 Step 6: TyME49 TachyTS RNA-Seq (%ile) 1,711 Genes
 Step 7: SNPs 4,877 Genes
 Step 8: Orbo Ph Pro 162,729 Genes
 Step 9: Orthologs 8,018 Genes

156 Genes (151 orthologs)

Gene Results | Genome V | **New Analysis**

Analyze your Gene results with a tool below.

Gene Ontology Enrichment | Metabolic Pathway Enrichment | Word Enrichment

Organism: Toxoplasma gondii ME49
 Ontology: ☐ Biological Process, ☐ Cellular Component, ☒ Molecular Function
 Evidence: ☒ Computed, ☒ Curated
 Limit to GO Slim terms: ☐ No, ☐ Yes
 P-Value cutoff: 0.01 (0 - 1)

Submit

Analysis Results: 56 rows

GO ID	GO Term	Genes in the bgld with this term	Genes in your result with this term	Percent of bgld genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0140399	ABC-type transporter activity	22	5	22.7	14.54	19.76	1.75e-5	2.73e-3	2.73e-3
GO:0140457	ATP-dependent activity	252	13	5.2	3.30	3.93	1.16e-4	7.96e-3	1.81e-2
GO:0042626	ATPase-coupled transmembrane transporter activity	53	6	11.3	7.24	8.64	1.53e-4	7.96e-3	2.39e-2
GO:1901363	heterocyclic compound binding	1334	35	2.6	1.68	2.29	3.59e-4	8.06e-3	5.59e-2
	hypothetical protein				N/A			N/A	N/A
	hypothetical protein				N/A			N/A	N/A
	hypothetical protein				N/A			mitochondrion - soluble	mitochondrion - soluble
	hypothetical protein				N/A			nucleus - chromatin	nucleus - chromatin
	hypothetical protein				N/A			N/A	N/A
	hypothetical protein				N/A			N/A	N/A
	hypothetical protein				N/A			dense granules	PM - peripheral 2

Link to strategy:

<https://toxodb.org/toxo/app/workspace/strategies/import/841229492ad74899>