Orthology and Gene Ontology

Learning objectives:

- Explore the orthology table on VEuPathDB gene pages
- Run and explore results of searches in OrthoMCL
- Leverage the phyletic pattern search
- Leverage the orthology transform tool
- Run and explore the results of a GO enrichment analysis
- Port GO enrichment results to Revigo

VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

Table of Contents

Introduction

1. Orthology and phyletics

Homologs are genes that share ancestry either by speciation (orthologs), gene duplication (paralogs), or gene transfer events (xenologs). Paralogs of a conserved gene may occur in a single species or strain. Conserved sequences in genomes can be used to infer evolutionary history (e.g., ribosomal sequences), their similarities and differences can be used to trace the divergence and evolution of organisms. Genes that share function by convergent evolution, but do not share ancestry are known as analogs.

Ortholog groups can also allow you to explore the potential functions of a gene, or group of genes across species. In pathogens like *Plasmodium falciparum*, ortholog groups might facilitate the identification of potential targets for drug or vaccine development.

For more detail on orthologs, paralogs and evolutionary genomics, read the review by Koonin[1].

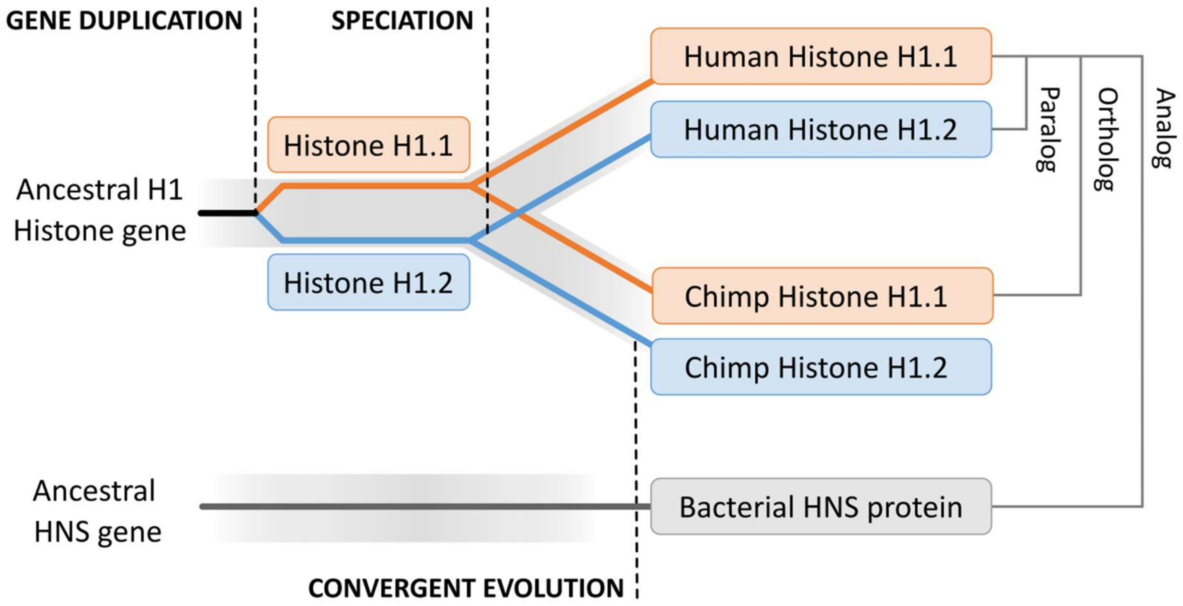


Figure 1. *Gene phylogeny (orange and blue) within species phylogeny (grey). Top shows an ancestral gene duplication event, producing two paralogs of the Histone H1 gene, producing H1.1 and H1.2. This is followed by a speciation event leading to Chimpanzee and Human Orthologs of the two genes. Bottom shows a gene with separate evolutionary origin that has evolved similar function to H1 Histones through convergent evolution, HNS (histone-like nucleoid-structuring protein). HNS is a bacterial analog to H1 Histone. Figure adapted from* this image *by Thomas Shafee (2018).*

2. OrthoMCL

OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences which not only share evolutionary history, but also share function. Thus, ortholog prediction is important in predicting the function of newly identified proteins. Detection of orthologs has become more widespread with the rapid progress in genome sequencing and the discovery of protein sequences [2,3].

OrthoMCL provides a database of ortholog groups with high degrees of functional conservation (e.g., they have consistent EC numbers, which link genes to specific products in metabolic pathways), making it a useful tool for functional annotation of genomes [4,5].

OrthoMCL identifies shared protein groups between species and is also capable of representing species specific gene expansion families. To achieve this, the OrthoMCL algorithm starts with reciprocal best BLAST hits within each proteome as potential in-paralog/recent paralog pairs and reciprocal best hits across any two proteomes as potential ortholog pairs. Related proteins are interlinked in a similarity graph. Then, MCL [6] is invoked to split mega-clusters - clusters that are just too big and uninformative. MCL clustering is based on weights between each pair of proteins, which are normalised by species to account for evolutionary distance. If you want to know more about how the MCL algorithm works have a look at this simplified explanation [7].

The organism specific orthology information garnered from our OrthoMCL analysis of VEuPathDB organisms is presented on gene pages and integrated into an Orthology Phylogenetic Profile search. They are available for anyone to explore or use for their own investigations. The OrthoMCL.org site offers a deep look into all data associated with the OrthoMCL results for orthology groups and proteins.

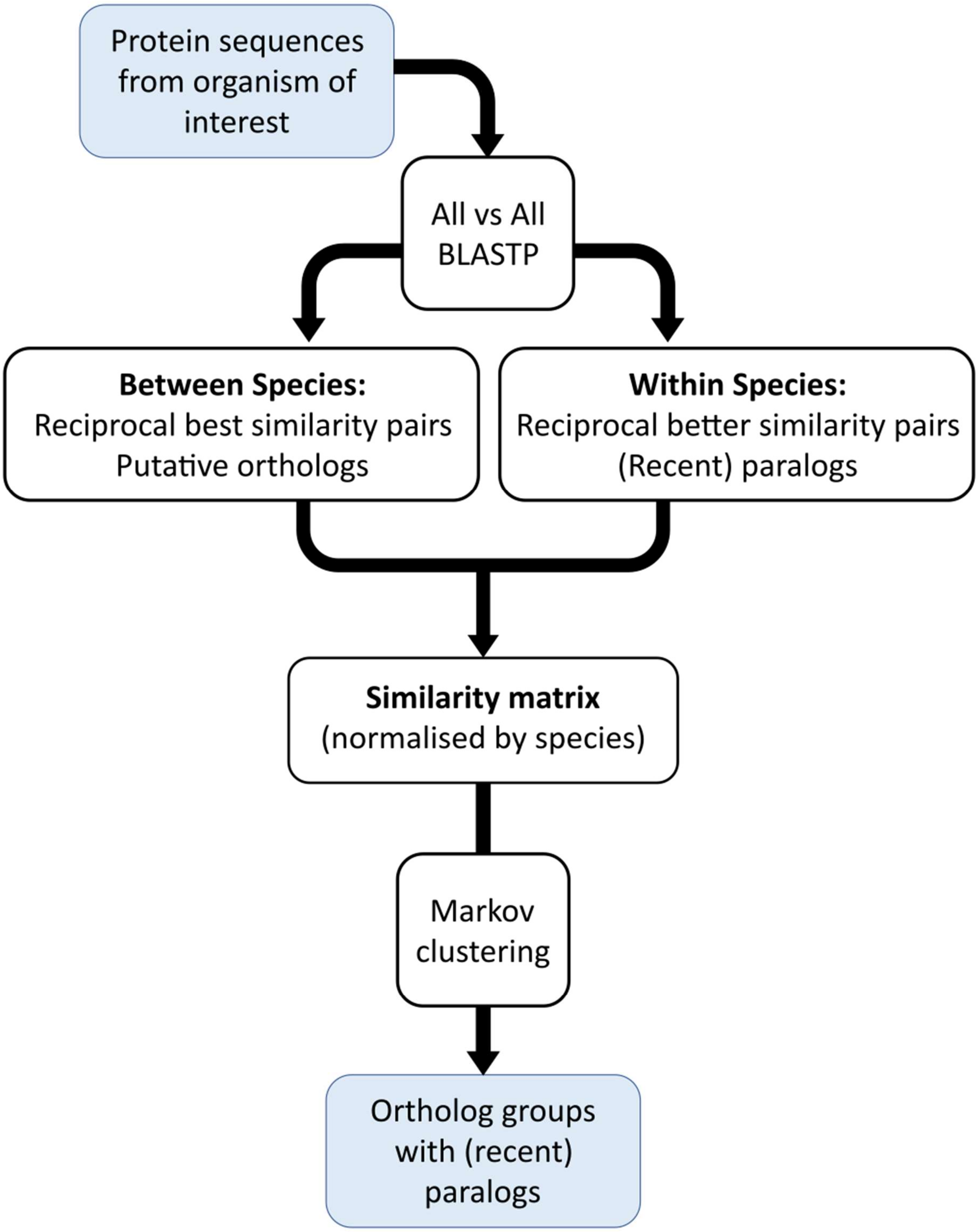
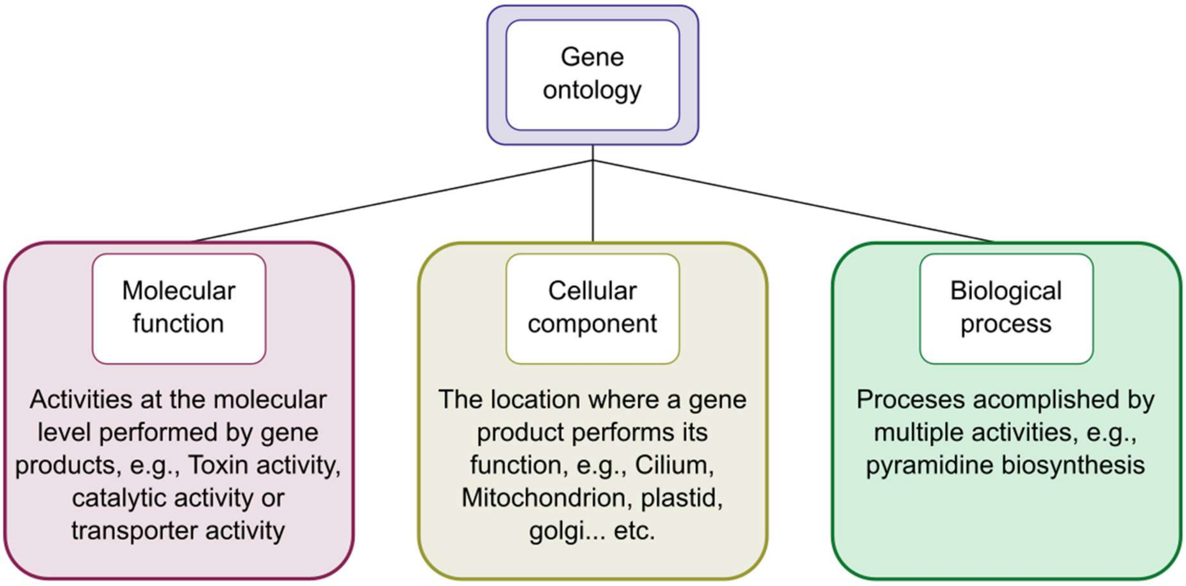

Protein sequences from organism of interest
All vs All BLASTP
Between Species:
Reciprocal best similarity pairs
Putative orthologs
Within Species:
Reciprocal better similarity pairs
(Recent) paralogs
Similarity matrix
(normalised by species)
Markov clustering
Ortholog groups with (recent) paralogs

Figure 2. *OrthoMCL's workflow.*

3. Gene Ontology

Ontologies are a controlled vocabulary of terms and concepts with relationships between them. Gene Ontology describes the knowledge of biological sciences and divides this knowledge into three broad categories: Molecular function, cellular component, and biological process.

To learn more about Gene Ontology, please visit: http://geneontology.org/docs/ontology-documentation/


Gene ontology
Molecular function
Activities at the molecular level performed by gene products, e.g., Toxin activity, catalytic activity or transporter activity
Cellular component
The location where a gene product performs its function, e.g., Cilium, Mitochondrion, plastid, golgi... etc.
Biological process
Proceses acomplished by multiple activities, e.g., pyramidine biosynthesis

A gene can be assigned a GO term either manually (by an annotator or curator when they evaluate experimental evidence from a publication) or computationally (based on the GO terms of genes that share sequence or functional domains). The origin of the assignment is documented; some researchers believe that manually assigned functional annotations are more accurate than those that are electronically transferred since a researcher has reviewed the manually annotated assignments. GO terms can be used to test whether your set of genes are enriched for a molecular function, cellular component, or biological process.

For example: A researcher performs a proteomics experiment on a protein fraction collected during an antimalarial treatment and identifies 100 proteins in total. When they examine the GO terms assigned to the gene set corresponding to the proteome, they see that 25 genes are assigned GO:0016301, kinase activity. Out of 5000 genes in the genome, only 100 are assigned GO:0016301. There is an overrepresentation of GO:0016301 in the researcher’s proteome which is ‘enriched’ for kinase activity.

A standard enrichment determination method employs Fisher's exact test, a statistical test that evaluates a 2x2 contingency table (in this case, the number of genes in my set versus number of genes from the same genome not in my set, and number of genes with GO term X versus number of genes without term X). This test produces a p-value between 0 and 1, where $p \leq 0.05$ is considered significant (that is, less than 5% probability that the enrichment is due to chance).

However, the test is performed for each of the 100s of GO terms, increasing the chances that a GO term will be incorrectly considered enriched (a false positive, or type I, error). Thus, the original p-value must be adjusted for so-called multiple hypothesis testing, resulting in an adjusted p-value such as the Benjamini-Hochberg false discovery rate (FDR) or Bonferroni adjusted p-value.

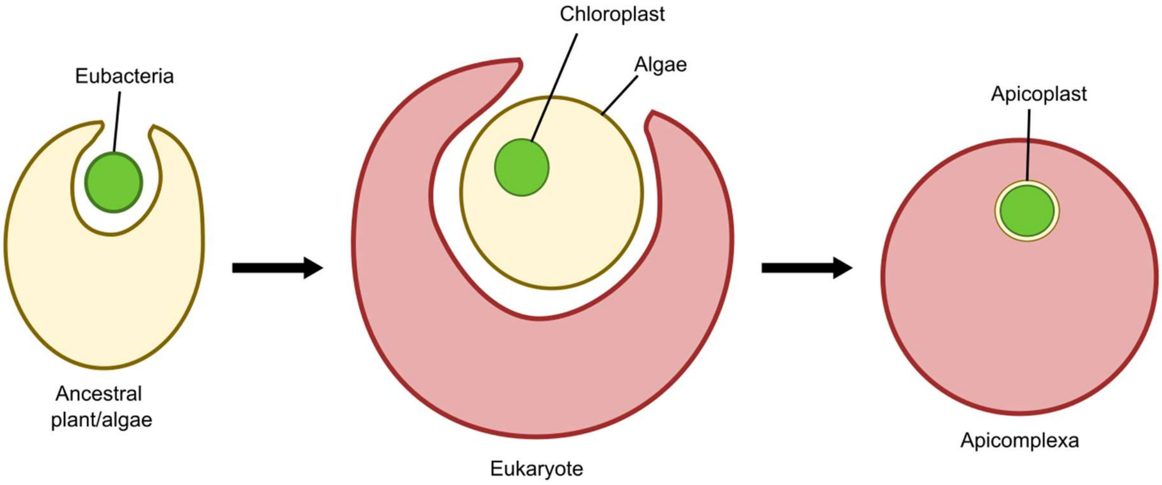
Exercises

1. Using the orthology transform tool to identify apicoplast targeted genes in *Toxoplasma* and *Neospora*.

Note: For this exercise use http://veupathdb.org

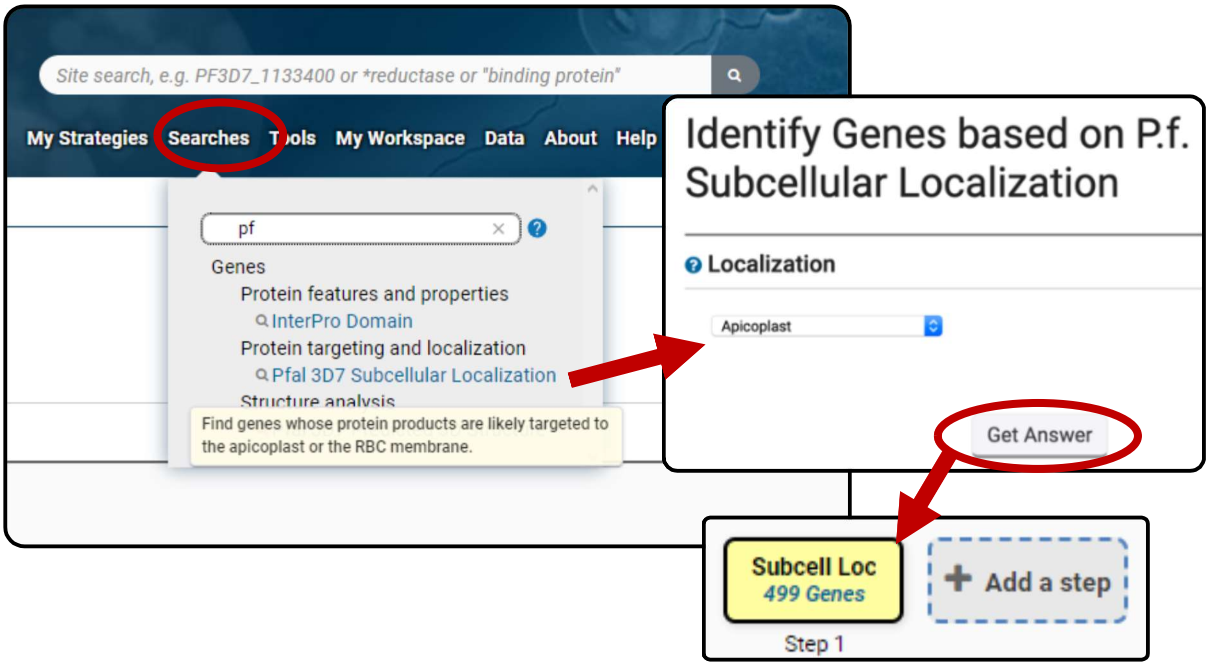
What is an apicoplast?

The apicoplast likely became encased in four membranes via a double endosymbiotic event. The chloroplast arose by engulfment of a cyanobacteria by a plant/algae ancestor. An alga was then engulfed by the ancestor of all apicomplexans. Thus, an apicoplast organelle arose with four membranes.



1. Start by finding genes in *Plasmodium* that are predicted to target the apicoplast.

   Hint: Navigate to the Pfal 3D7 Subcellular Localization search for Apicoplast. You can filter the type of search by text query.



Site search, e.g. PF3D7_1133400 or *reductase or "binding protein"
My Strategies
Searches
Tools
My Workspace
Data
About
Help
pf
Genes
Protein features and properties
InterPro Domain
Protein targeting and localization
Pfal 3D7 Subcellular Localization
Structure analysis

Find genes whose protein products are likely targeted to the apicoplast or the RBC membrane.

Identify Genes based on P.f. Subcellular Localization

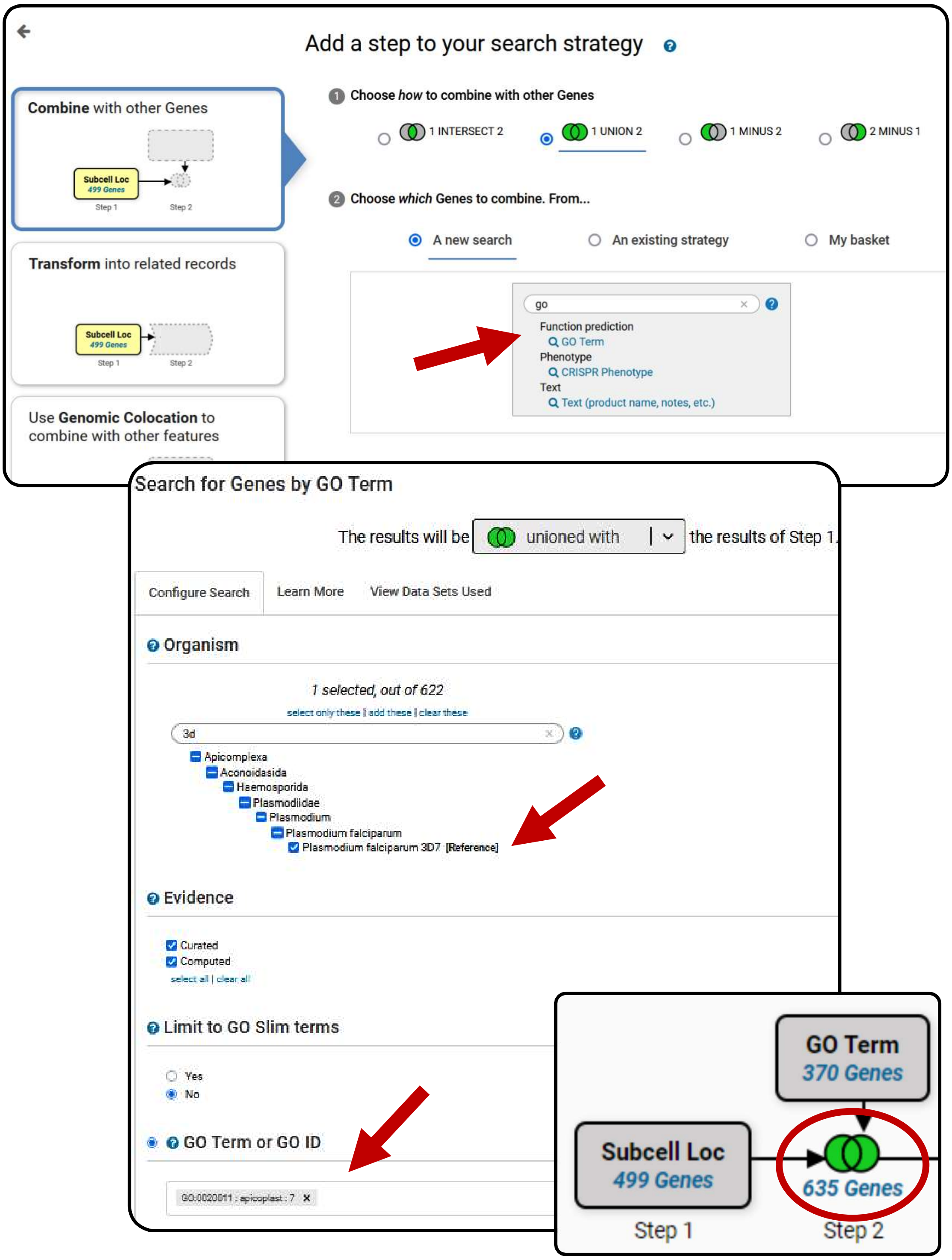
Localization: Apicoplast

Get Answer
Subcell Loc
499 Genes
Step 1
Add a step

2. You can further expand your list of potentially Apicoplast targeted proteins by adding a GO terms search strategy for the term “apicoplast” or the GO ID: “GO:0020011” in *P. falciparum* 3D7.

   Hint: click on add step the go to the function prediction category and select the GO term search.

   a. Which Boolean operation did you use? Union or intersect?



Add a step to your search strategy

Combine with other Genes

Transform into related records

Use **Genomic Colocation** to combine with other features

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

go

Function prediction
GO Term
Phenotype
CRISPR Phenotype
Text
Text (product name, notes, etc.)

Search for Genes by GO Term

The results will be unioned with the results of Step 1.

Configure Search | Learn More | View Data Sets Used

Organism

1 selected, out of 622

select only these | add these | clear these

3d

- Apicomplexa
  - Aconoidasida
    - Haemosporida
      - Plasmodiidae
        - Plasmodium
          - Plasmodium falciparum
            - ☑ Plasmodium falciparum 3D7 [Reference]

Evidence

☑ Curated
☑ Computed
select all | clear all

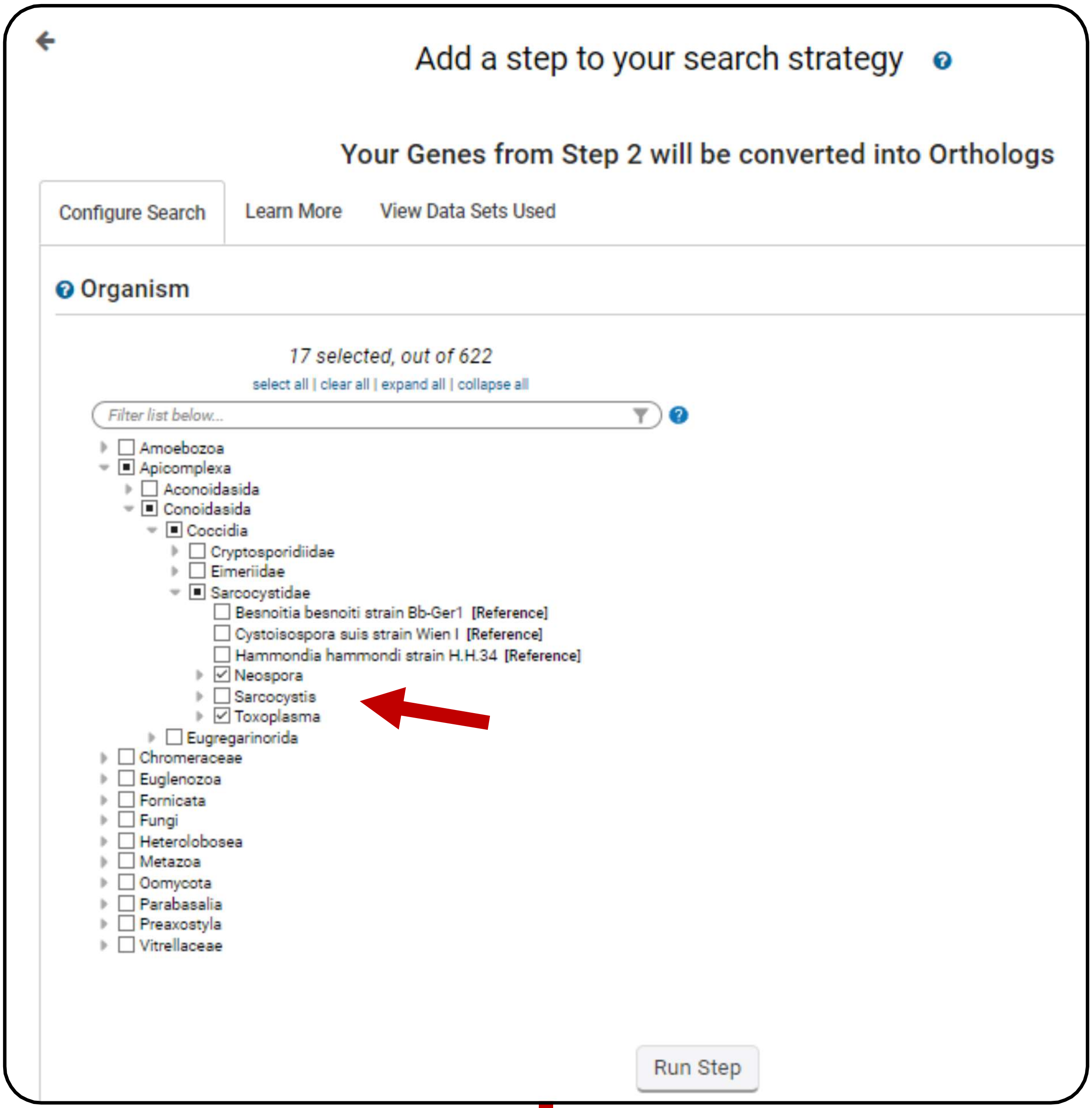
Limit to GO Slim terms

○ Yes
● No

GO Term or GO ID

GO:0020011 : apicoplast : 7 ✕

3. Add a step to your strategy that transforms the results with *Toxoplasma* and *Neospora* orthologs.



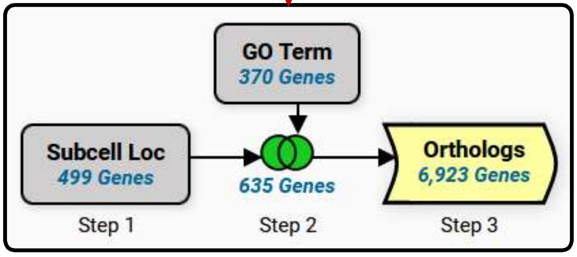
Add a step to your search strategy

Your Genes from Step 2 will be converted into Orthologs

Organism

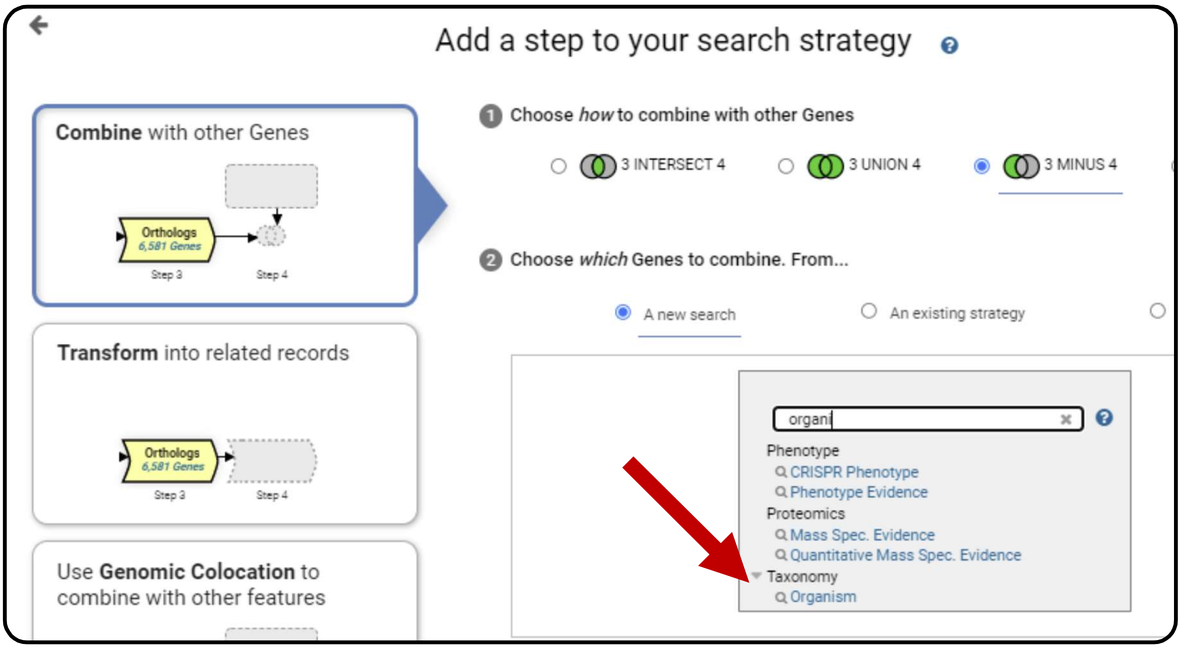
17 selected, out of 622

- □ Amoebozoa
- ■ Apicomplexa
  - □ Aconoidasida
  - ■ Conoidasida
    - ■ Coccidia
      - □ Cryptosporidiidae
      - □ Eimeriidae
      - ■ Sarcocystidae
        - □ Besnoitia besnoiti strain Bb-Ger1 [Reference]
        - □ Cystoisospora suis strain Wien I [Reference]
        - □ Hammondia hammondi strain H.H.34 [Reference]
        - ☑ Neospora
        - □ Sarcocystis
        - ☑ Toxoplasma
    - □ Eugregarinorida
- □ Chromeraceae
- □ Euglenozoa
- □ Fornicata
- □ Fungi
- □ Heterolobosea
- □ Metazoa
- □ Oomycota
- □ Parabasalia
- □ Preaxostyla
- □ Vitrellaceae



4. Although *Cryptosporidium* is an apicomplexan parasite it has lost its apicoplast! Can you use this fact to refine your results from the above search?
    a. First pull up all *Cryptosporidium* genes with the Genes by Taxonomy search and then transform these back to their *Toxoplasma* and *Neospora* orthologs for the subtraction to complete. Think about what kind of intersection you should be using!

Hint: try subtracting out any orthologs present in *Cryptosporidium*. You will need to use a nested strategy.



Add a step to your search strategy

Combine with other Genes

Transform into related records

Use **Genomic Colocation** to combine with other features

1 Choose *how* to combine with other Genes

○ 3 INTERSECT 4 ○ 3 UNION 4 ● 3 MINUS 4

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

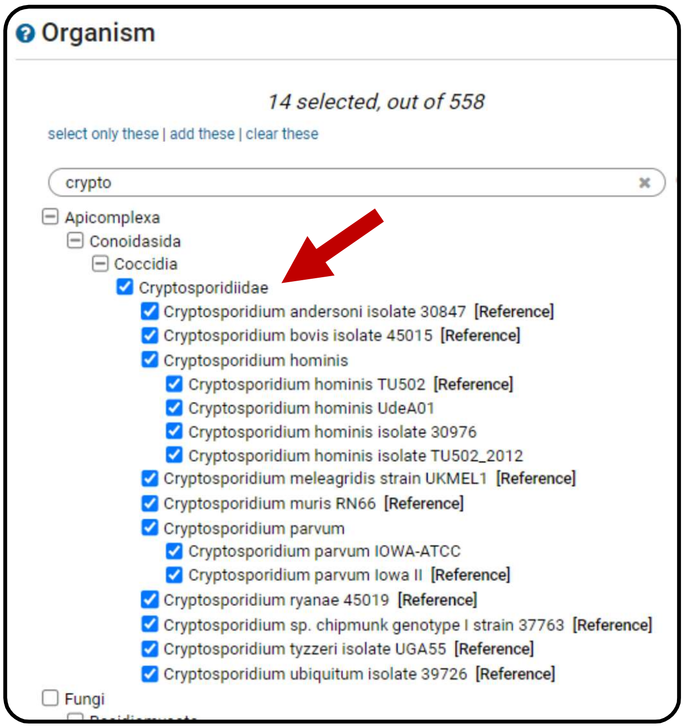
● A new search ○ An existing strategy ○

organi

Phenotype
- CRISPR Phenotype
- Phenotype Evidence

Proteomics
- Mass Spec. Evidence
- Quantitative Mass Spec. Evidence

Taxonomy
- Organism



Organism

14 selected, out of 558

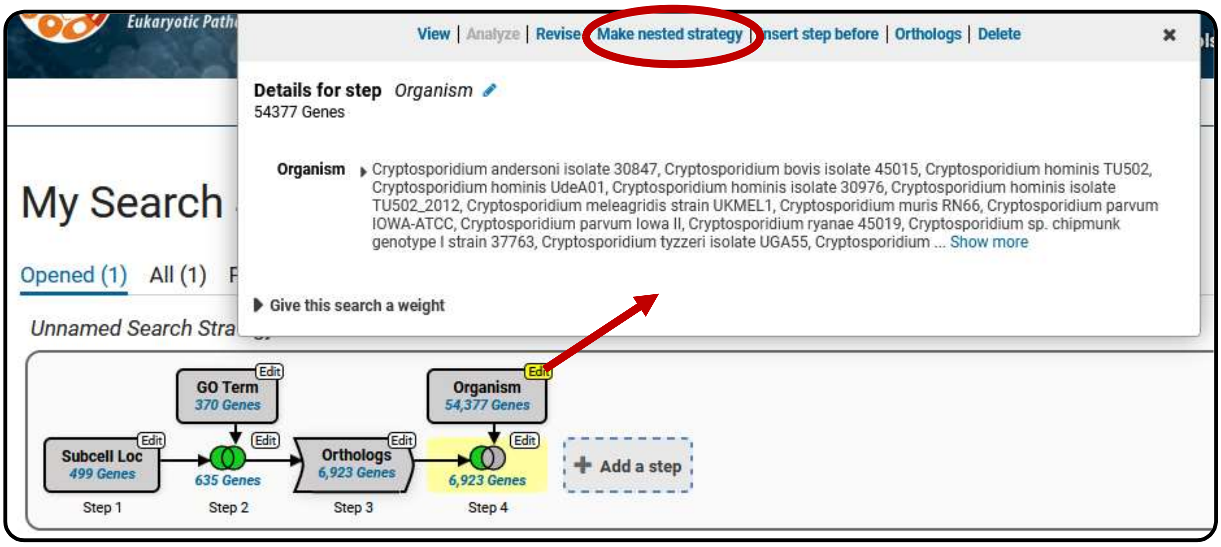
select only these | add these | clear these

crypto

- Apicomplexa
  - Conoidasida
    - Coccidia
      - ☑ Cryptosporidiidae
        - ☑ Cryptosporidium andersoni isolate 30847 [Reference]
        - ☑ Cryptosporidium bovis isolate 45015 [Reference]
        - ☑ Cryptosporidium hominis
          - ☑ Cryptosporidium hominis TU502 [Reference]
          - ☑ Cryptosporidium hominis UdeA01
          - ☑ Cryptosporidium hominis isolate 30976
          - ☑ Cryptosporidium hominis isolate TU502_2012
        - ☑ Cryptosporidium meleagridis strain UKMEL1 [Reference]
        - ☑ Cryptosporidium muris RN66 [Reference]
        - ☑ Cryptosporidium parvum
          - ☑ Cryptosporidium parvum IOWA-ATCC
          - ☑ Cryptosporidium parvum Iowa II [Reference]
        - ☑ Cryptosporidium ryanae 45019 [Reference]
        - ☑ Cryptosporidium sp. chipmunk genotype I strain 37763 [Reference]
        - ☑ Cryptosporidium tyzzeri isolate UGA55 [Reference]
        - ☑ Cryptosporidium ubiquitum isolate 39726 [Reference]
- □ Fungi

How to make a nested strategy

Click edit > nested strategy

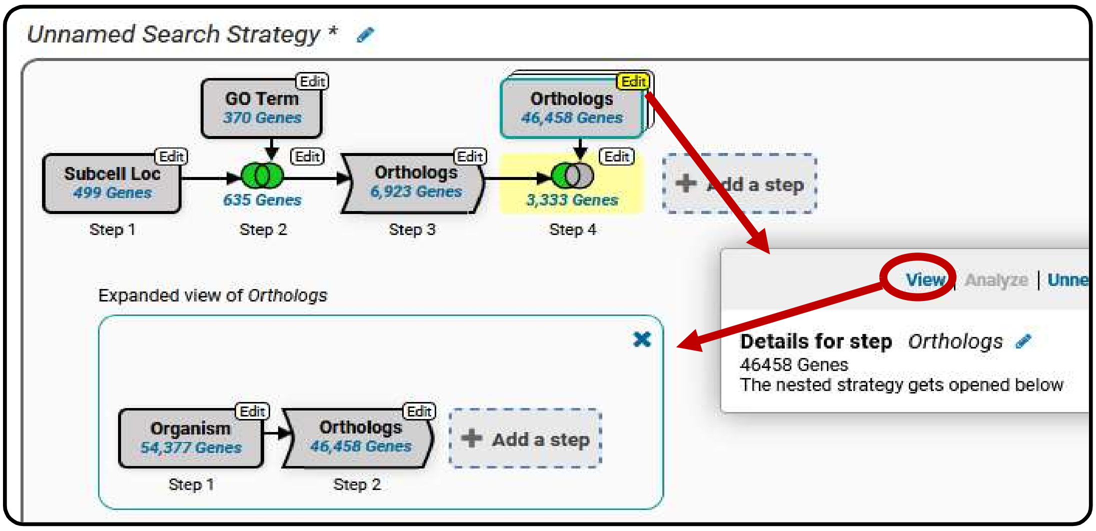


Details for step *Organism*
54377 Genes

Organism Cryptosporidium andersoni isolate 30847, Cryptosporidium bovis isolate 45015, Cryptosporidium hominis TU502, Cryptosporidium hominis UdeA01, Cryptosporidium hominis isolate 30976, Cryptosporidium hominis isolate TU502_2012, Cryptosporidium meleagridis strain UKMEL1, Cryptosporidium muris RN66, Cryptosporidium parvum IOWA-ATCC, Cryptosporidium parvum Iowa II, Cryptosporidium ryanae 45019, Cryptosporidium sp. chipmunk genotype I strain 37763, Cryptosporidium tyzzeri isolate UGA55, Cryptosporidium ... Show more

▸ Give this search a weight

Then to view and edit the nested strategies, for instance, to add an ortholog transformation step for *Neospora* and *Toxoplasma*, click edit > view



Details for step *Orthologs*
46458 Genes
The nested strategy gets opened below

This leaves you with apicoplast specific genes for *Toxoplasma* and *Neospora* that you could target in future research.

2. GO enrichment analysis

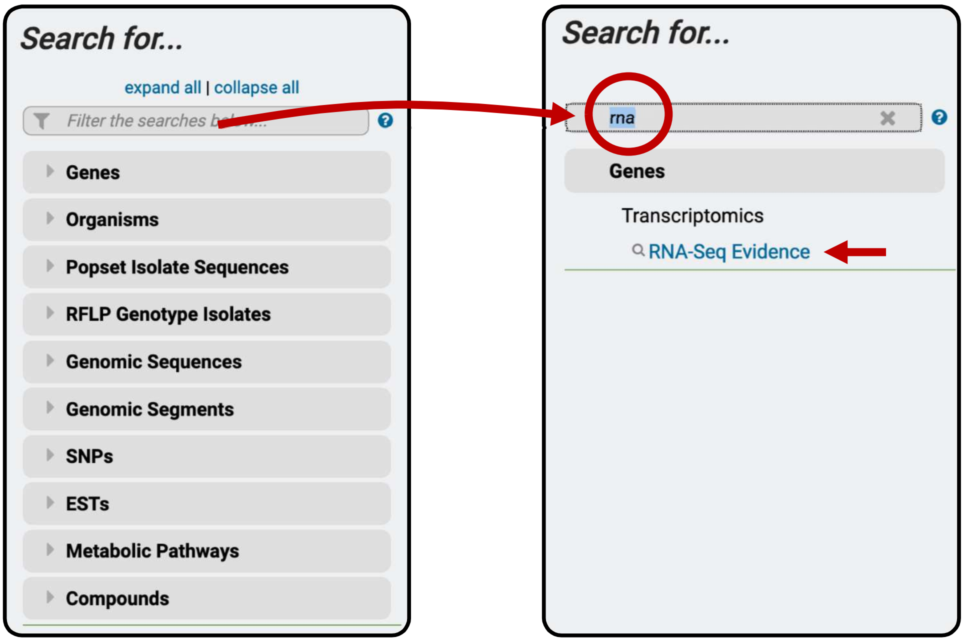
GO term enrichment analysis can be carried out at any stage in any search strategy to see what categories of genes are most common. This function is currently available on every organism site except for the VEuPathDB main site (this is due to the underlying complexities of the website).

Pick your favorite database (e.g. plasmodb, toxodb) and have a play around with GO terms or follow along with this example.

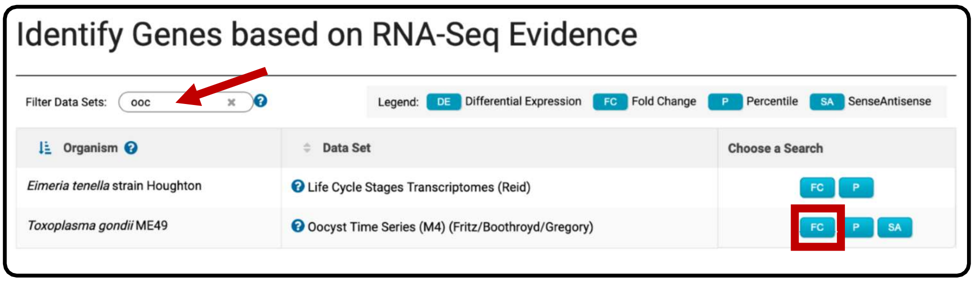
Retrieving RNA-seq evidence

To run a GO enrichment analysis, you need a list of genes to test. This can be a list of gene IDs from your experimental results (upload them with the ID search) or a gene list resulting from a search you conducted on a VEuPathDB website. For this example, in ToxoDB, we will identify genes that are differentially regulated over time.

1. Navigate to the RNA-Seq searches and find the data set called “**Oocyst Time Series (M4)**” from Fritz et al. A quick way of getting to the RNA-Seq searches is to type ‘rna’ in the filter box on the left of the home page and click on the RNA-Seq Evidence link. See image below.



2. The RNA-Seq evidence page includes a list of all data sets that are loaded in the website. To quickly find a dataset, you can start typing key words in the “Filter Data Sets” box. For example, start typing the word “oocyst”.



Identify Genes based on RNA-Seq Evidence

Filter Data Sets: ooc

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

Organism	Data Set	Choose a Search
Eimeria tenella strain Houghton	Life Cycle Stages Transcriptomes (Reid)	FC P
Toxoplasma gondii ME49	Oocyst Time Series (M4) (Fritz/Boothroyd/Gregory)	FC P SA

3. Once you find the data set of interest, choose the fold-change (FC) search. For this exercise, identify genes that are upregulated by 20-fold in days 4 and 10 compared to the day 0 time point. Parameters to set:

   1 - Up-regulated

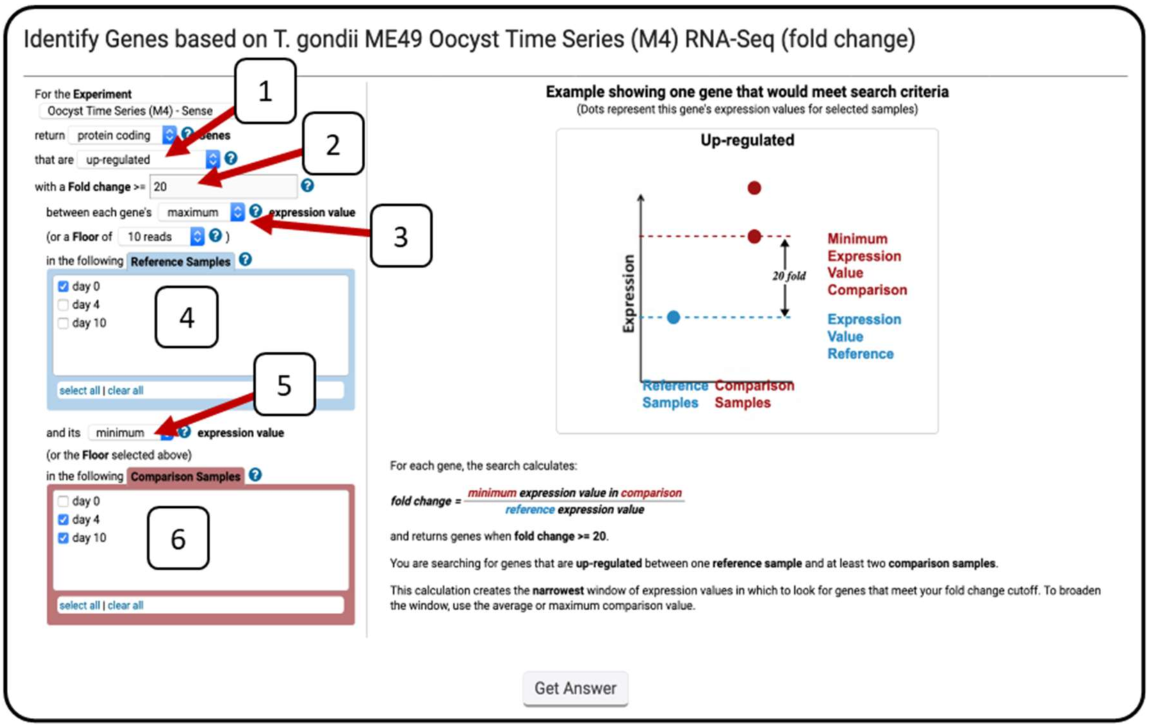
   2 - 20-fold

   3 - Maximum

   4 - Day 0

   5 - Minimum

   6 - Day 4 and 10



Identify Genes based on T. gondii ME49 Oocyst Time Series (M4) RNA-Seq (fold change)

For the **Experiment** Oocyst Time Series (M4) - Sense

return protein coding **Genes**

that are up-regulated

with a **Fold change** >= 20

between each gene's maximum **expression value**

(or a **Floor** of 10 reads)

in the following **Reference Samples**

☑ day 0
□ day 4
□ day 10

select all | clear all

and its minimum **expression value**

(or the **Floor** selected above)

in the following **Comparison Samples**

□ day 0
☑ day 4
☑ day 10

select all | clear all

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

For each gene, the search calculates:

fold change = *minimum expression value in comparison* / *reference expression value*

and returns genes when **fold change >= 20**.

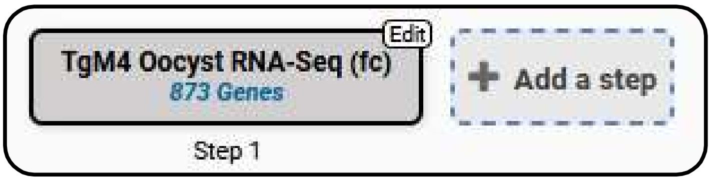
You are searching for genes that are **up-regulated** between one **reference sample** and at least two **comparison samples**.

This calculation creates the **narrowest** window of expression values in which to look for genes that meet your fold change cutoff. To broaden the window, use the average or maximum comparison value.

Get Answer

4. Click "Get Answer" to initiate the search. This will return a one-step search strategy.

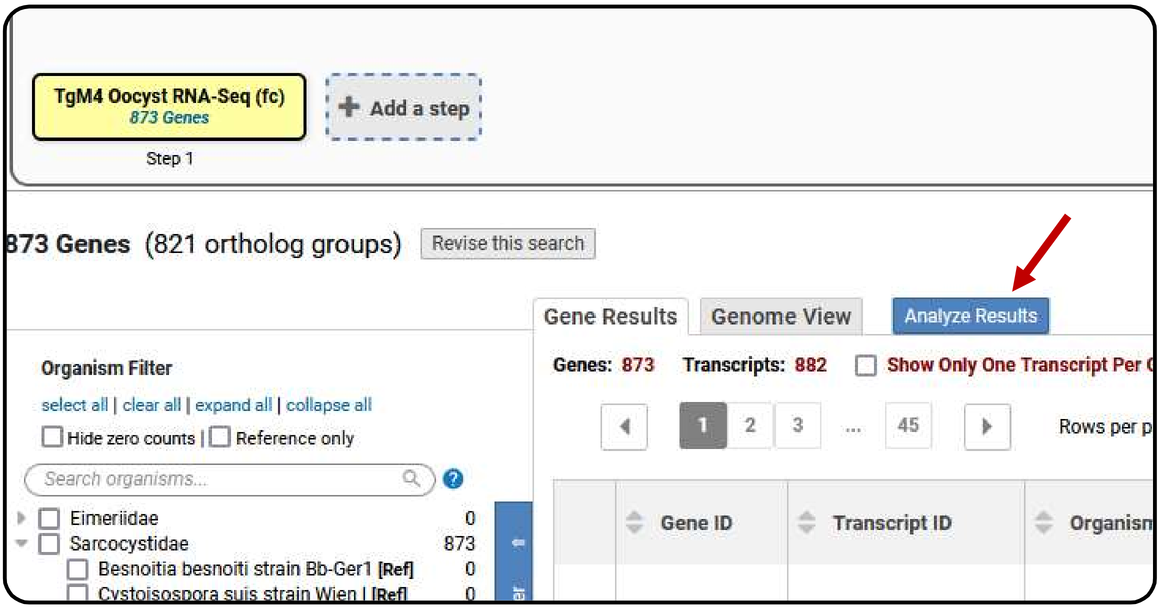
   a. How many genes did you get?



TgM4 Oocyst RNA-Seq (fc)
873 Genes
Edit
Step 1
Add a step

Run the GO enrichment analysis

1. Click on the Analyze Results tab just above the list of genes (arrow in image below) to open the enrichment tools.
    a. Besides GO enrichment, what other analyses are available?

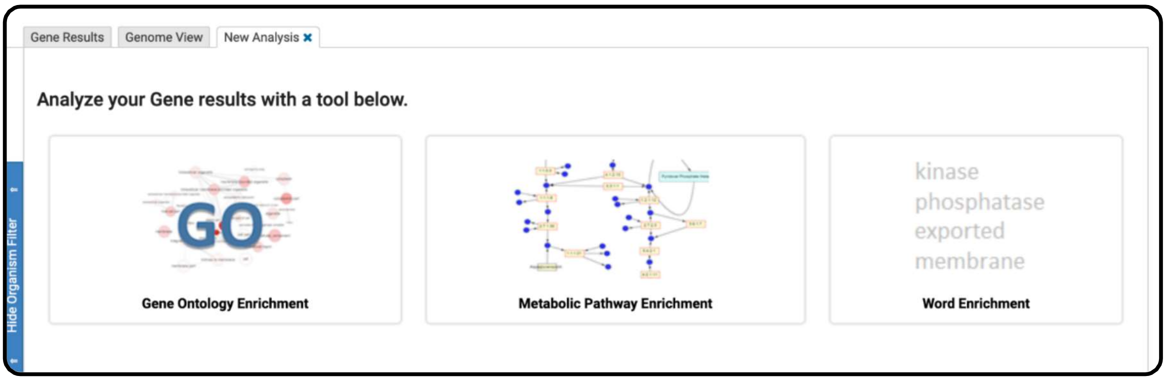


873 Genes (821 ortholog groups)

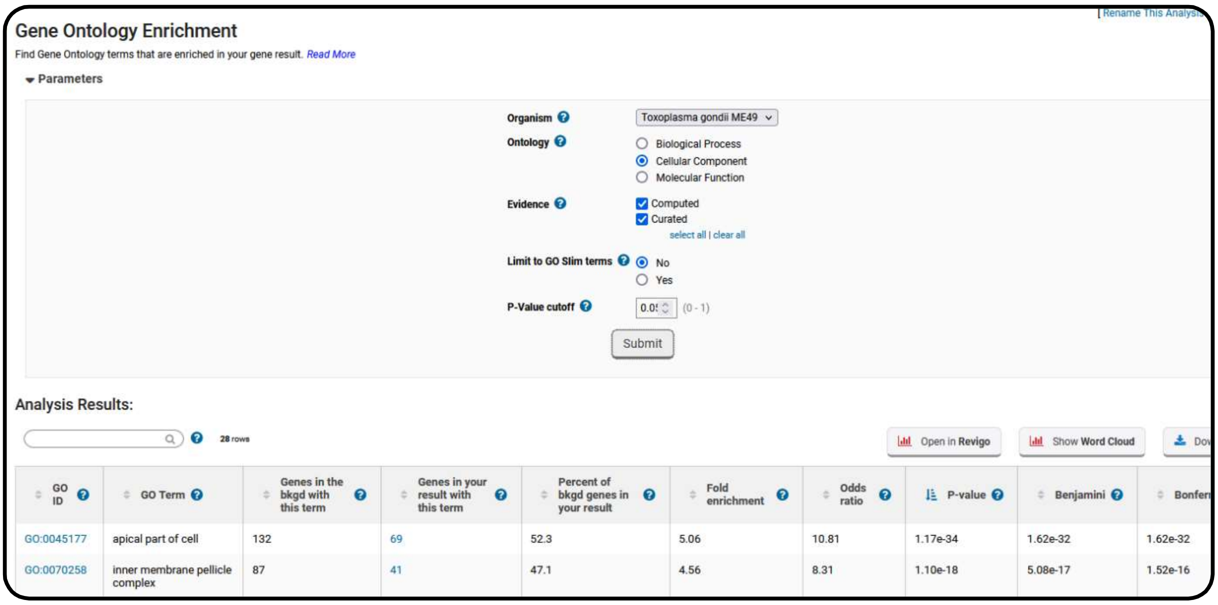
Genes: 873 Transcripts: 882

□ Eimeriidae	0
□ Sarcocystidae	873
□ Besnoitia besnoiti strain Bb-Ger1 [Ref]	0
□ Cystoisospora suis strain Wien I [Ref]	0

2. Click on the GO enrichment option. This will reveal the parameters that you can modify. For the purpose of this exercise, keep all the defaults and click on “Submit”.
    - Organism = T. gondii ME49
    - Ontology = Cellular Component
    - Evidence = Computed and Curated
    - Limit to GO Slim terms? = NO



Analyze your Gene results with a tool below.



[Rename This Analysis

Gene Ontology Enrichment

Find Gene Ontology terms that are enriched in your gene result. Read More

Parameters

Organism: Toxoplasma gondii ME49

Ontology: ○ Biological Process ● Cellular Component ○ Molecular Function

Evidence: ☑ Computed ☑ Curated

select all | clear all

Limit to GO Slim terms: ● No ○ Yes

P-Value cutoff: 0.0! (0 - 1)

Submit

Analysis Results:

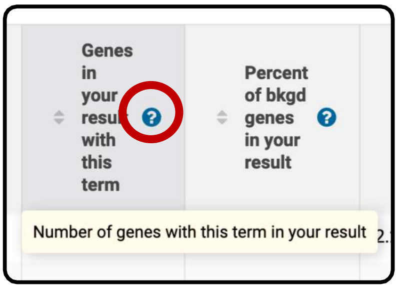
28 rows

GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonfer
GO:0045177	apical part of cell	132	69	52.3	5.06	10.81	1.17e-34	1.62e-32	1.62e-32
GO:0070258	inner membrane pellicle complex	87	41	47.1	4.56	8.31	1.10e-18	5.08e-17	1.52e-16

Explore your results of your analysis

1. What is the top enriched GO term from this analysis?
2. Does this make sense for an enrichment analysis of the cellular component of your Oocyst expressed genes? Notice that the p-value with Benjamini or Bonferroni correction is very low.
3. What do each of the columns in the analysis table represent?

   ***Hint**: move your mouse over the question mark next to each column header*



Genes in your resu with this term
Percent of bkgd genes in your result

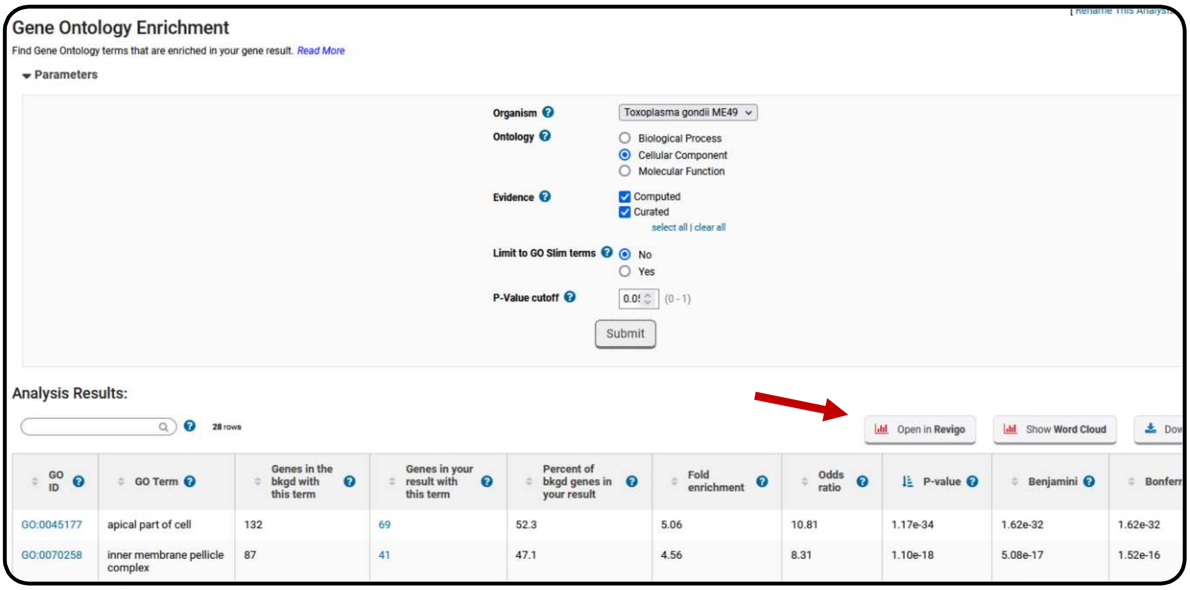
Number of genes with this term in your result

- *Fold enrichment* -The ratio of the proportion of genes in the list of interest with a specific GO term over the proportion of genes in the background with that term.
- *Odds ratio* -The odds of the GO term appearing in the gene list are the same as that for the background list.
- *P-value* –The null hypothesis or the probability of getting a result that is equal or greater than what was observed.
- *Benjamini-Hochburg false discovery rate* – A method for controlling false discovery rates for type 1 errors.

- *Bonferroni adjusted P-values* - A method for correcting significance based on multiple comparisons.

Port your results to Revigo

1. Click the Open in Revigo button to port the results to Revigo, the Reduce and Visualize Ontology tool.



Gene Ontology Enrichment

Find Gene Ontology terms that are enriched in your gene result. *Read More*

▾ Parameters

Organism: Toxoplasma gondii ME49

Ontology: ○ Biological Process ● Cellular Component ○ Molecular Function

Evidence: ☑ Computed ☑ Curated (select all | clear all)

Limit to GO Slim terms: ● No ○ Yes

P-Value cutoff: 0.0! (0 - 1)

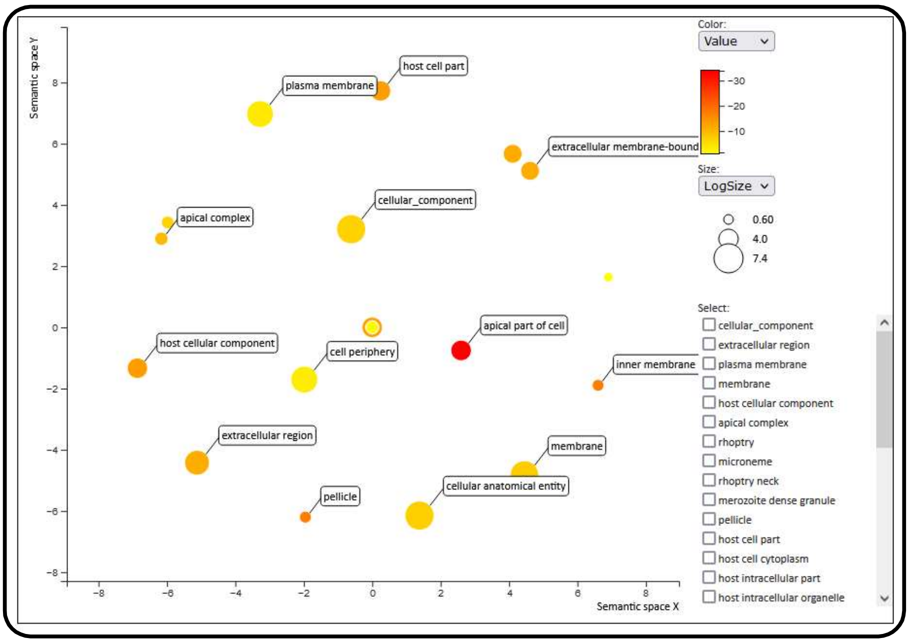
Submit

Analysis Results:

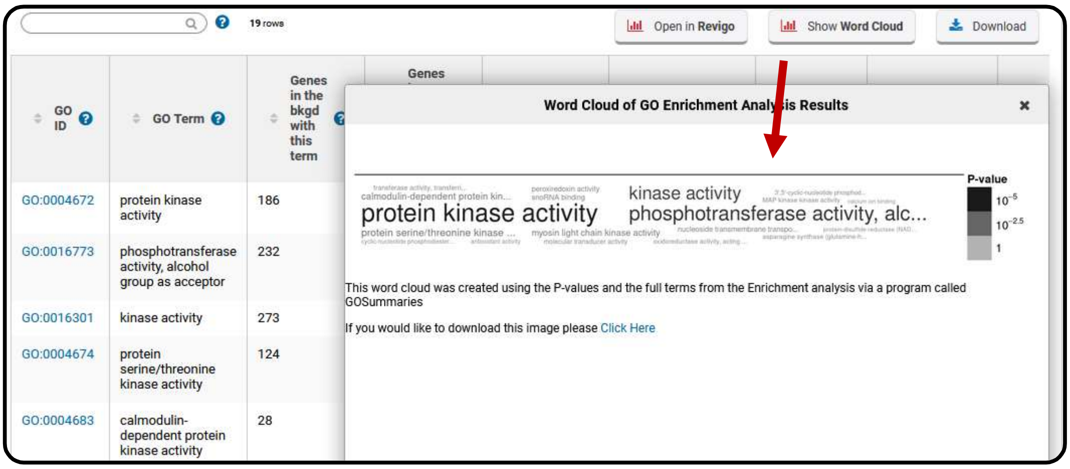
28 rows

GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferr
GO:0045177	apical part of cell	132	69	52.3	5.06	10.81	1.17e-34	1.62e-32	1.62e-32
GO:0070258	inner membrane pellicle complex	87	41	47.1	4.56	8.31	1.10e-18	5.08e-17	1.52e-16

2. Once at Revigo, you may need to scroll down to click Start Revigo to run the analysis with default parameters. Revigo provides a scatterplot and table, an integrative map and a tree map to supplement the table provided in the VEuPathDB site. See the Revigo publication for more information.



3. Try rerunning the GO enrichment analysis, but this time select the Molecular Function ontology.
    a. What is the top enriched GO term?
    b. What is the p-value for the enrichment?
    c. Do you have more or less confidence than in the last search that this function is enriched in your gene set?
4. Click on the “Word Cloud” button above the Molecular Function analysis results. What type of analysis is this? What information can you (See image below).



19 rows
Open in Revigo
Show Word Cloud
Download
Genes
×
transferase activity, transferri...
calmodulin-dependent protein kin...
peroxiredoxin activity
snoRNA binding
kinase activity
3',5'-cyclic-nucleotide phosphod...
MAP kinase kinase activity
calcium ion binding
protein kinase activity
phosphotransferase activity, alc...
protein serine/threonine kinase ...
myosin light chain kinase activity
nucleoside transmembrane transpo...
protein-disulfide reductase (NAD...
asparagine synthase (glutamine-h...
cyclic-nucleotide phosphodiester...
antioxidant activity
molecular transducer activity
oxidoreductase activity, acting ...
P-value
10^{-5}
$10^{-2.5}$
1

GO ID	GO Term	Genes in the bkgd with this term
GO:0004672	protein kinase activity	186
GO:0016773	phosphotransferase activity, alcohol group as acceptor	232
GO:0016301	kinase activity	273
GO:0004674	protein serine/threonine kinase activity	124
GO:0004683	calmodulin-dependent protein kinase activity	28

Word Cloud of GO Enrichment Analysis Results

This word cloud was created using the P-values and the full terms from the Enrichment analysis via a program called GOSummaries

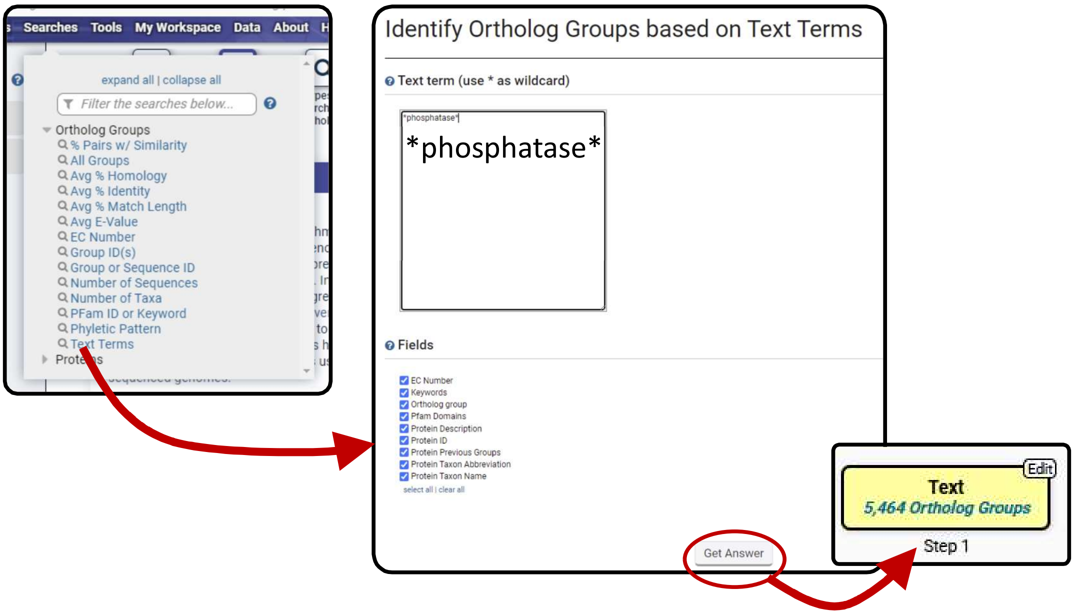
If you would like to download this image please Click Here

3. Search orthologs and filter by phyletics in OrthoMCL (optional)

Note: Use http://orthomcl.org for this exercise.

Find all plant proteins that are likely phosphatases that do not have orthologs outside of plants.

1. Use the text search to find OrthoMCL groups that contain the word “*phosphatase*”. Note that the search should be run without the quotation marks but with the asterisks.



Identify Ortholog Groups based on Text Terms

Text term (use * as wildcard)

phosphatase

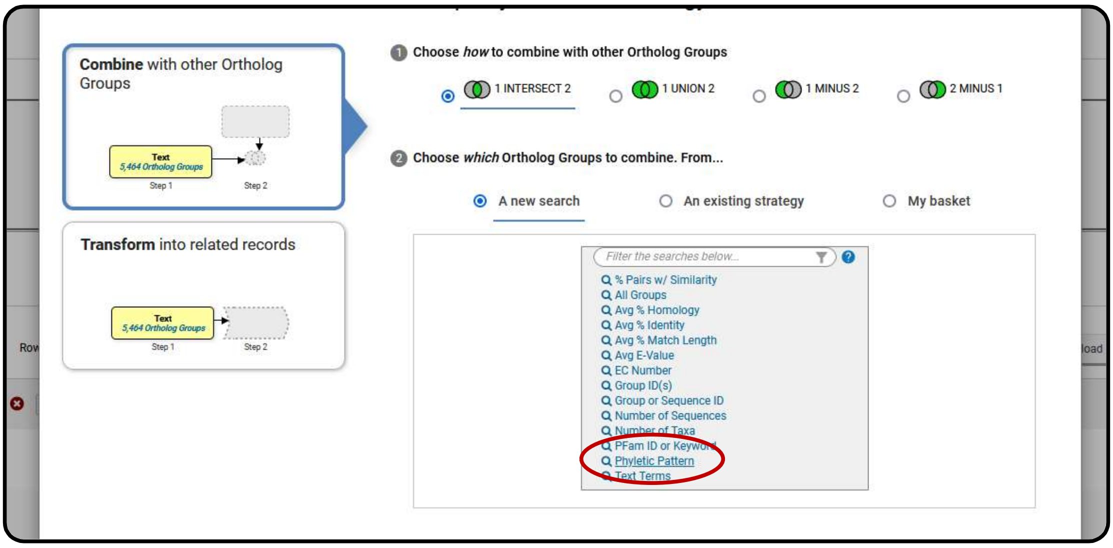
Fields

☑ EC Number
☑ Keywords
☑ Ortholog group
☑ Pfam Domains
☑ Protein Description
☑ Protein ID
☑ Protein Previous Groups
☑ Protein Taxon Abbreviation
☑ Protein Taxon Name

select all | clear all

2. Add a step and run a phyletic pattern search for groups that contain any plant protein but do not contain any other organism outside plants.

   Hint: make sure everything has a red x on it except for plants (Viridiplantae (VIRI)), which should be a grey circle.

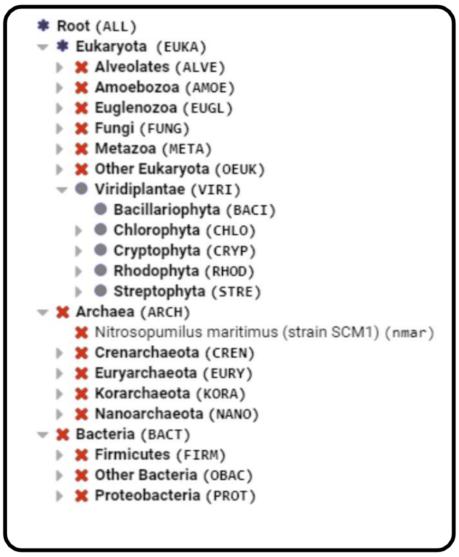


1 Choose *how* to combine with other Ortholog Groups

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

2 Choose *which* Ortholog Groups to combine. From...

● A new search ○ An existing strategy ○ My basket

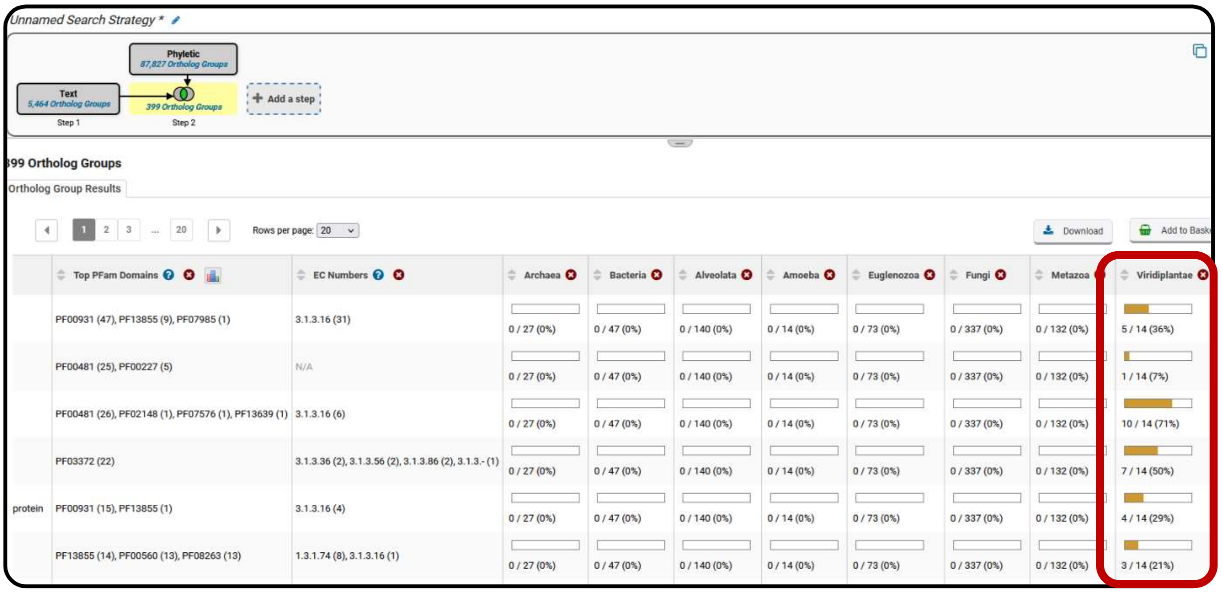


- ✱ Root (ALL)
  - ✱ Eukaryota (EUKA)
    - ✖ Alveolates (ALVE)
    - ✖ Amoebozoa (AMOE)
    - ✖ Euglenozoa (EUGL)
    - ✖ Fungi (FUNG)
    - ✖ Metazoa (META)
    - ✖ Other Eukaryota (OEUK)
    - ● Viridiplantae (VIRI)
      - ● Bacillariophyta (BACI)
      - ● Chlorophyta (CHLO)
      - ● Cryptophyta (CRYP)
      - ● Rhodophyta (RHOD)
      - ● Streptophyta (STRE)
  - ✖ Archaea (ARCH)
    - ✖ Nitrosopumilus maritimus (strain SCM1) (nmar)
    - ✖ Crenarchaeota (CREN)
    - ✖ Euryarchaeota (EURY)
    - ✖ Korarchaeota (KORA)
    - ✖ Nanoarchaeota (NANO)
  - ✖ Bacteria (BACT)
    - ✖ Firmicutes (FIRM)
    - ✖ Other Bacteria (OBAC)
    - ✖ Proteobacteria (PROT)

3. Examine your results.

   a. How many groups were returned by the search?

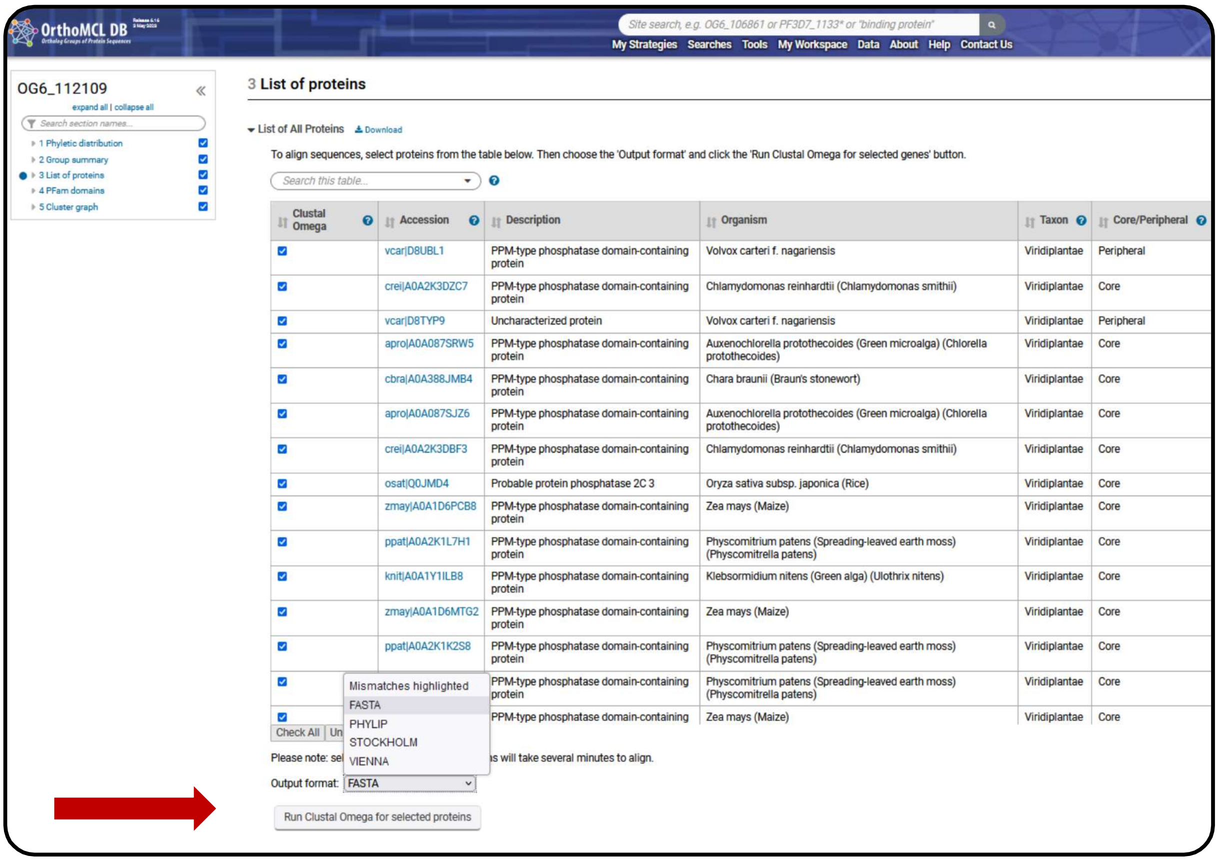
   b. What is the distribution of plant proteins in each orthology group? (use the Add Columns tool to turn on the Viridiplantae column if it is not already on)



399 Ortholog Groups

	Top PFam Domains	EC Numbers	Archaea	Bacteria	Alveolata	Amoeba	Euglenozoa	Fungi	Metazoa	Viridiplantae
	PF00931 (47), PF13855 (9), PF07985 (1)	3.1.3.16 (31)	0 / 27 (0%)	0 / 47 (0%)	0 / 140 (0%)	0 / 14 (0%)	0 / 73 (0%)	0 / 337 (0%)	0 / 132 (0%)	5 / 14 (36%)
	PF00481 (25), PF00227 (5)	N/A	0 / 27 (0%)	0 / 47 (0%)	0 / 140 (0%)	0 / 14 (0%)	0 / 73 (0%)	0 / 337 (0%)	0 / 132 (0%)	1 / 14 (7%)
	PF00481 (26), PF02148 (1), PF07576 (1), PF13639 (1)	3.1.3.16 (6)	0 / 27 (0%)	0 / 47 (0%)	0 / 140 (0%)	0 / 14 (0%)	0 / 73 (0%)	0 / 337 (0%)	0 / 132 (0%)	10 / 14 (71%)
	PF03372 (22)	3.1.3.36 (2), 3.1.3.56 (2), 3.1.3.86 (2), 3.1.3.- (1)	0 / 27 (0%)	0 / 47 (0%)	0 / 140 (0%)	0 / 14 (0%)	0 / 73 (0%)	0 / 337 (0%)	0 / 132 (0%)	7 / 14 (50%)
protein	PF00931 (15), PF13855 (1)	3.1.3.16 (4)	0 / 27 (0%)	0 / 47 (0%)	0 / 140 (0%)	0 / 14 (0%)	0 / 73 (0%)	0 / 337 (0%)	0 / 132 (0%)	4 / 14 (29%)
	PF13855 (14), PF00560 (13), PF08263 (13)	1.3.1.74 (8), 3.1.3.16 (1)	0 / 27 (0%)	0 / 47 (0%)	0 / 140 (0%)	0 / 14 (0%)	0 / 73 (0%)	0 / 337 (0%)	0 / 132 (0%)	3 / 14 (21%)

4. Next, you can run a multiple sequence alignment for OG6_112109. Click on the group ID in your result table and navigate to the List of Proteins section of the group page. The Clustal Omega tool is integrated into the table. There are several formats available for the Clustal output, making it easy to take these results to other visualization programs.



3 List of proteins

List of All Proteins Download

To align sequences, select proteins from the table below. Then choose the 'Output format' and click the 'Run Clustal Omega for selected genes' button.

Clustal Omega	Accession	Description	Organism	Taxon	Core/Peripheral
☑	vcar|D8UBL1	PPM-type phosphatase domain-containing protein	Volvox carteri f. nagariensis	Viridiplantae	Peripheral
☑	crei|A0A2K3DZC7	PPM-type phosphatase domain-containing protein	Chlamydomonas reinhardtii (Chlamydomonas smithii)	Viridiplantae	Core
☑	vcar|D8TYP9	Uncharacterized protein	Volvox carteri f. nagariensis	Viridiplantae	Peripheral
☑	apro|A0A087SRW5	PPM-type phosphatase domain-containing protein	Auxenochlorella protothecoides (Green microalga) (Chlorella protothecoides)	Viridiplantae	Core
☑	cbra|A0A388JMB4	PPM-type phosphatase domain-containing protein	Chara braunii (Braun's stonewort)	Viridiplantae	Core
☑	apro|A0A087SJZ6	PPM-type phosphatase domain-containing protein	Auxenochlorella protothecoides (Green microalga) (Chlorella protothecoides)	Viridiplantae	Core
☑	crei|A0A2K3DBF3	PPM-type phosphatase domain-containing protein	Chlamydomonas reinhardtii (Chlamydomonas smithii)	Viridiplantae	Core
☑	osat|Q0JMD4	Probable protein phosphatase 2C 3	Oryza sativa subsp. japonica (Rice)	Viridiplantae	Core
☑	zmay|A0A1D6PCB8	PPM-type phosphatase domain-containing protein	Zea mays (Maize)	Viridiplantae	Core
☑	ppat|A0A2K1L7H1	PPM-type phosphatase domain-containing protein	Physcomitrium patens (Spreading-leaved earth moss) (Physcomitrella patens)	Viridiplantae	Core
☑	knit|A0A1Y1ILB8	PPM-type phosphatase domain-containing protein	Klebsormidium nitens (Green alga) (Ulothrix nitens)	Viridiplantae	Core
☑	zmay|A0A1D6MTG2	PPM-type phosphatase domain-containing protein	Zea mays (Maize)	Viridiplantae	Core
☑	ppat|A0A2K1K2S8	PPM-type phosphatase domain-containing protein	Physcomitrium patens (Spreading-leaved earth moss) (Physcomitrella patens)	Viridiplantae	Core
☑		PPM-type phosphatase domain-containing protein	Physcomitrium patens (Spreading-leaved earth moss) (Physcomitrella patens)	Viridiplantae	Core
☑		PPM-type phosphatase domain-containing	Zea mays (Maize)	Viridiplantae	Core

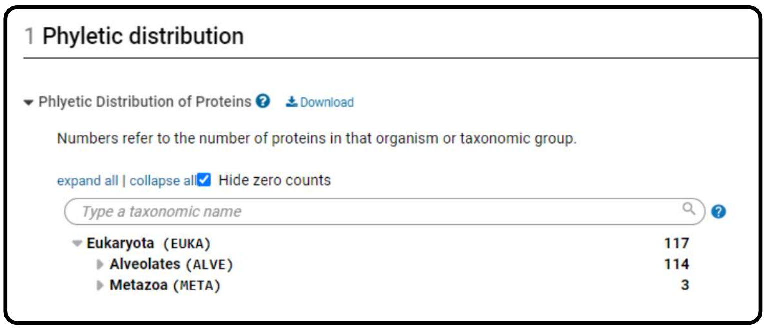
Please note: sel s will take several minutes to align.

Output format: FASTA

4. Explore a specific OrthoMCL group - examining the cluster graph (optional)

Note: Use http://orthomcl.org for this exercise.

1. Visit the OrthoMCL group OG6_131670. Use the site search to navigate to OG6_131670.
2. Examine the Phyletic Distribution.
   a. What is the phylogenetic distribution of the members of this group? The distribution is presented as a tree. Expand the tree to view the distribution.



1 Phyletic distribution

▾ Phyletic Distribution of Proteins ? Download

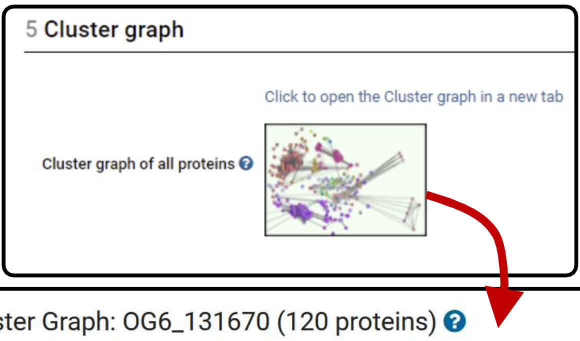
Numbers refer to the number of proteins in that organism or taxonomic group.

expand all | collapse all ☑ Hide zero counts

Type a taxonomic name

▾ Eukaryota (EUKA)	117
▸ Alveolates (ALVE)	114
▸ Metazoa (META)	3

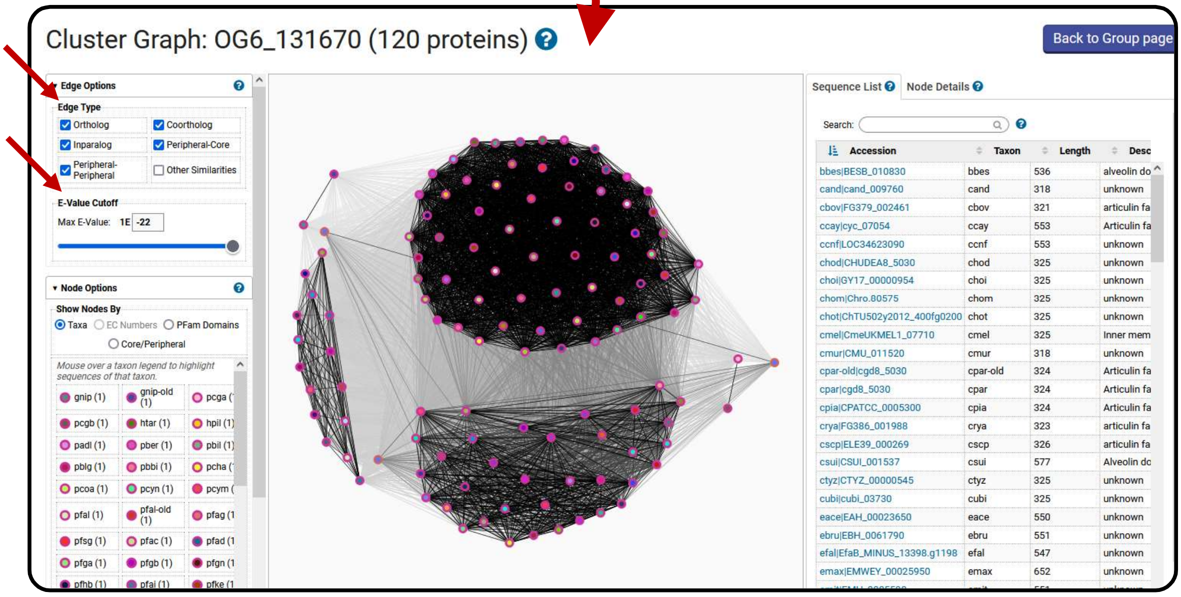
3. Navigate to the Cluster graph tab. Modify the E-value cutoff slider. What happens when you increase or decrease the E-value? Can you identify subclusters of orthologs? The view of the graph can be changed using the Edge type options and the Node options.



5 Cluster graph

Click to open the Cluster graph in a new tab

Cluster graph of all proteins ?



Cluster Graph: OG6_131670 (120 proteins) ?

▾ Edge Options ?

Edge Type

☑ Ortholog ☑ Coortholog

☑ Inparalog ☑ Peripheral-Core

☑ Peripheral-Peripheral ☐ Other Similarities

E-Value Cutoff

Max E-Value: 1E -22

▾ Node Options ?

Show Nodes By

● Taxa ○ EC Numbers ○ PFam Domains ○ Core/Peripheral

Mouse over a taxon legend to highlight sequences of that taxon.

gnip (1)	gnip-old (1)	pcga (
pcgb (1)	htar (1)	hpil (1)
padl (1)	pber (1)	pbil (1)
pblg (1)	pbbi (1)	pcha (
pcoa (1)	pcyn (1)	pcym (
pfal (1)	pfal-old (1)	pfag (1
pfsg (1)	pfac (1)	pfad (1
pfga (1)	pfgb (1)	pfgn (1
pfhb (1)	pfai (1)	pfke (1

Sequence List ? Node Details ?

Search: ?

Accession	Taxon	Length	Desc
bbes|BESB_010830	bbes	536	alveolin do
cand|cand_009760	cand	318	unknown
cbov|FG379_002461	cbov	321	articulin fa
ccay|cyc_07054	ccay	553	Articulin fa
ccnf|LOC34623090	ccnf	553	unknown
chod|CHUDEA8_5030	chod	325	unknown
choi|GY17_00000954	choi	325	unknown
chom|Chro.80575	chom	325	unknown
chot|ChTU502y2012_400fg0200	chot	325	unknown
cmel|CmeUKMEL1_07710	cmel	325	Inner mem
cmur|CMU_011520	cmur	318	unknown
cpar-old|cgd8_5030	cpar-old	324	Articulin fa
cpar|cgd8_5030	cpar	324	Articulin fa
cpia|CPATCC_0005300	cpia	324	Articulin fa
crya|FG386_001988	crya	323	articulin fa
cscp|ELE39_000269	cscp	326	articulin fa
csui|CSUI_001537	csui	577	Alveolin do
ctyz|CTYZ_00000545	ctyz	325	unknown
cubi|cubi_03730	cubi	325	unknown
eace|EAH_00023650	eace	550	unknown
ebru|EBH_0061790	ebru	551	unknown
efal|EfaB_MINUS_13398.g1198	efal	547	unknown
emax|EMWEY_00025950	emax	652	unknown

Resources

These are the operators you need to be aware of for these exercises. Don't forget to refer to the search strategies help sheets for more in-depth help!

Operator	Combined search will contain:
1 INTERSECT 2	IDs common between both lists
1 UNION 2	All IDs from both lists
1 MINUS 2	IDs only in list 1
2 MINUS 1	IDs only in list 2

Gene Ontology - http://geneontology.org/docs/ontology-documentation/

Enzyme Commission numbers - https://www.qmul.ac.uk/sbcs/iubmb/enzyme/

More info on Fischer's exact test - http://www.biostathandbook.com/fishers.html

Fisher's Exact Test and the Hypergeometric Distribution (the M&M example) - https://youtu.be/udyAvvaMjfM

Odds ratios - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/

False discovery rates and P value correction - http://brainder.org/2011/09/05/fdr-corrected-fdr-adjusted-p-values/

GO Slim - http://www-legacy.geneontology.org/GO.slims.shtml

REVIGO - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021800

References

1. Koonin, E. V. Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.* **39**, 309–338 (2005).
2. Glover, N. *et al.* Advances and Applications in the Quest for Orthologs. *Mol. Biol. Evol.* **36**, 2157–2164 (2019).
3. Linard, B. *et al.* Ten Years of Collaborative Progress in the Quest for Orthologs. *Mol. Biol. Evol.* **38**, 3033–3045 (2021).
4. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **13**, 2178–2189 (2003).
5. Fischer, S. *et al.* Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. *Curr. Protoc. Bioinforma.* **35**, (2011).
6. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* **30**, 121–141 (2008).
7. Jagota, A. Markov Clustering Algorithm. *Towards Data Science* https://towardsdatascience.com/markov-clustering-algorithm-577168dad475 (2020).

Glossary

Clusters of Orthologous Genes (COGs), phylogenetic classification of proteins encoded in complete genomes.

Co-orthology, recent descent and duplication

EC numbers, Enzyme Commission number used to classify enzymes based on the chemical reactions that they classify.

Homologs, genes that share ancestry either by speciation (orthologs), gene duplication (paralogs), or gene transfer events (xenologs)

In-paralog, recent duplication.

Markov Clustering Algorthm (MCL),

Orthology is the study of genes across species that are conserved over evolutionary time.

Ortholog, homologous genes resulting from gene duplication.

Paralog, homologous genes resulting from speciation.

Phyletics, or **phylogenetics**, is the study of patterns in evolutionary history across species.

Reciprocal best hits, where sequences from two different genomes find each other as the best scoring match

Xenolog, homologous genes resulting from a horizontal or lateral gene transfer event