

## Orthology and Ontologies

### Learning objectives:

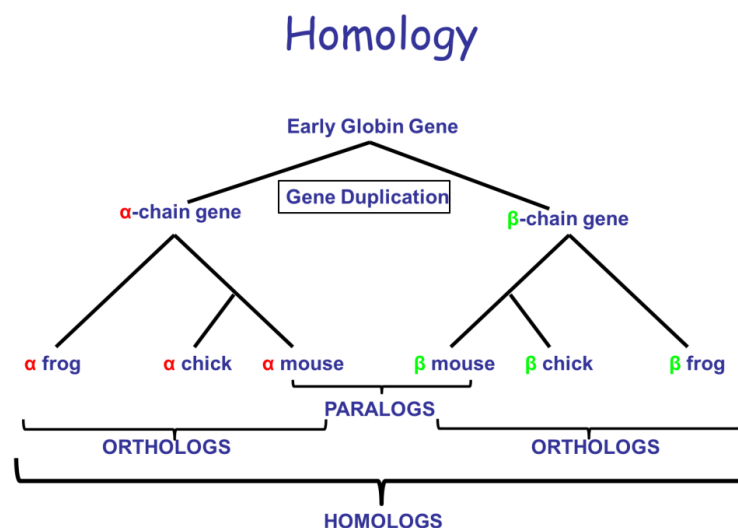
- Use orthology information on gene record pages to infer function on a gene whose protein product is undefined.
- Run phyletic pattern searches using check boxes or an expression.
- Combine searches using the strategy system.
- Explore individual ortholog group pages.
- Explore the group cluster graphs.
- 

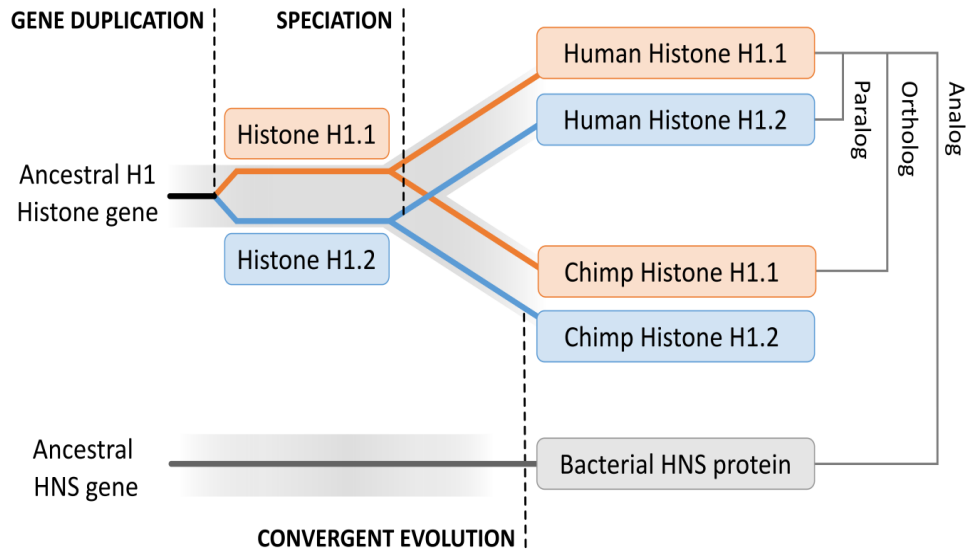
### About Orthology and Phyletics

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. A good place to start for more information about homology is [Koonin, EV. Orthologs, paralogs and evolution. Annu Rev Genet 2005.](#)

-





*Figure SEQ Figure \\* ARABIC 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).*

## About OrthoMCL

OrthoMCL is a genome-scale database that groups orthologous protein sequences across the tree of life. An orthogroup contains genes descended from a common ancestor by a process of duplication and speciation (see figure above), so a single orthogroup may contain both genes across different species with similar function and paralogs within a single species. Each protein in every OrthoMCL species is assigned to precisely one ortholog group (e.g. [OG6\\_162879](#)). Importantly, proteins within a single OrthoMCL group have been shown to display a high degree of functional conservation (e.g., a group's proteins have consistent EC numbers) ([Li et al. 2003](#)). Orthology is important in predicting the function of the rapidly increasing number of newly identified proteins produced by genome sequencing and the automated discovery of protein sequences ([Glover et al. 2019](#)). Within VEuPathDB, orthology can be used to transform a list of genes from one species into their closest equivalents in another species.

OrthoMCL contains two sets of genomes. A **Core** set of 150 genomes have been chosen as well annotated reference organisms that broadly represent the major branches of the tree of life. The OrthoMCL algorithm uses BLAST to calculate pairwise distances among all proteins in the 150 core genomes, normalizes the scores for sequence length and evolutionary distance, then uses MCL clustering ([Dongen 2000](#); [www.micans.org/mcl](http://www.micans.org/mcl)) to create orthogroups of similar proteins. All of the non-core VEuPathDB species (pathogens, hosts, and vectors) have been added as **Peripheral** organisms, in some cases including multiple strains and genome assemblies for the same species. All proteins from the Peripheral organisms are assigned to the most similar Core cluster by best BLAST score, but proteins that do not match any Core protein with an e-value better than  $1e^{-5}$  are set aside as

**Residuals.** Pairwise BLAST distances among all Residual proteins are computed and used for a second round of MCL clustering to create Residual groups (e.g. [OG6r20\\_100305](#))

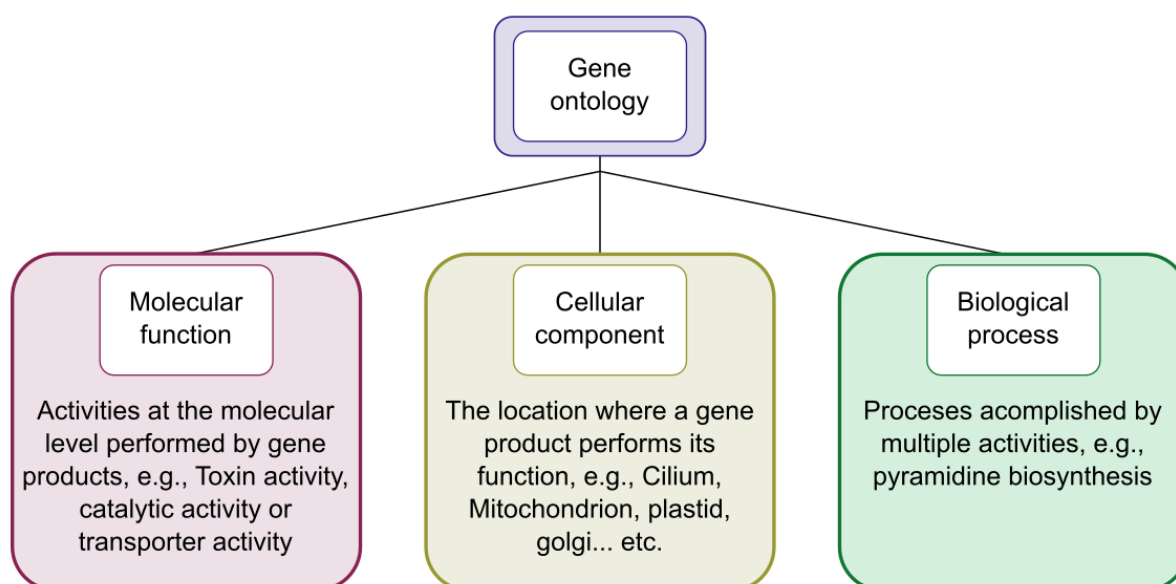
The OrthoMCL website offers the ability to explore ortholog groups by taxonomy, number of proteins or species, sequence similarity, EC numbers, PFAM domains, and text search of gene descriptions. Users can use the Ortholog Group or Protein queries in the grey Search box to the left or the Searches menu in the header bar, or just type a search term in the 'Site search' box above which will result in a list of proteins and groups to explore. In addition, users can use a VEuPathDB Galaxy workflow to map their own set of proteins (e.g. protein sequences derived from a genome sequence of an organism) to OrthoMCL groups. See the [Assign Proteins to Groups](#) page.

For more information, see the [About OrthoMCL](#) and [OrthoMCL FAQ](#) pages.

### About 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/>



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.

**1. Orthology and ontology information on gene record pages and in OrthoMCL. For this exercise we will start at [FungiDB](#).**

- a. Go to the FungiDB gene record page for CGB\_L0350W hypothetical protein CNBL0590, a protein in *Cryptococcus gattii*. Although this gene is annotated as hypothetical (meaning that the gene product is undefined), examining CGB\_L0350W orthologs and ontology information may inform protein function.
- b. Navigate to the 'Orthology and Synteny' section. The 'Orthologs and Paralogs within FungiDB' table shows the product descriptions and other data for genes within FungiDB that are part of the Ortholog Group for CGB\_L0350W. Does this gene have orthologs in other *Cryptococcus* species that have specific gene product descriptions?

**CGB\_L0350W** <<

expand all | collapse all

Search section names...

- 1 Gene models ☒
- 2 Annotation, curation and identifiers ☒
- 3 Link outs ☒
- 4 Genomic Location ☒
- 5 Literature ☒
- 6 Taxonomy ☒
- 7 Orthology and synteny ☒**
- 8 Phenotype ☒
- 9 Transcriptomics ☒
- 10 Sequence analysis ☒
- 11 Sequences ☒
- 12 Structure analysis ☒
- 13 Protein features and properties ☒
- 14 Function prediction ☒
- 15 Pathways and interactions ☒
- 16 Immunology ☒

### 7 Orthology and synteny

Ortholog Group OG6\_106189

Orthologs and Paralogs within FungiDB Data sets

To run Clustal Omega, select genes from the table below. Then choose the sequence type and initiate the alignment with the 'Run Clustal Omega' button.

Crypto X ?

Clustal Omega	Gene	Product	Organism
<input type="checkbox"/>	D1P53_002977	unspecified product	Cryptococcus cf. gattii MF34
<input type="checkbox"/>	L203_04836	Cation efflux protein [Source:UniProtKB/TrEMBL;Acc:A0A1E3ICE3]	Cryptococcus depauperatus CBS 784
<input type="checkbox"/>	I314_06191	cation antiporter	Cryptococcus gattii CA1873
<input type="checkbox"/>	I306_06271	cation antiporter	Cryptococcus gattii EJB2
<input type="checkbox"/>	I311_05609	cation antiporter	Cryptococcus gattii NT-10

- c. Move to the "Function prediction" section of the gene page and examine the GO Slim and GO Terms tables. Annotations can be assigned based on direct evidence, as from an experimental (EXP), or inferred from direct assay (IDA), etc. What does the IEA Evidence code mean? Visit <https://geneontology.org/docs/guide-go-evidence-codes/> to find out.

- d. Examine this gene's Ortholog Group on OMCL.org to learn about orthologs from organisms outside FungiDB. Return to the 'Orthology and Synteny' use the Ortholog Group link to visit this genes orthology group, OG6\_106189, at OrthoMCL. The OrthoMCL group page is divided into 5 sections.

**Phyletic distribution:** Numbers refer to the number of proteins in that organism or taxonomic group. In order to see organisms and taxonomic groups without proteins in this ortholog group, uncheck 'Hide zero counts.'

**Group summary:** This section provides a summary based on protein types. A core protein is from one of the 150 core species that were initially used to form 'core' groups. A peripheral protein is from a peripheral species whose entire proteome was mapped into the 'core' groups. Peripheral proteins that do not map into a 'core' group are placed into residuals groups.

**List of Proteins:** Lists all proteins in the ortholog group. The table also is a tool for running a Clustal Omega analysis of selected proteins.

**PFam domains:** Provides a list of PFam domains that are shared by the proteins within the ortholog group. PFam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. PFam domains are domains (consensus sequences) associated with protein families that define protein function.

**Cluster Graph:** Provides a dynamic network visualization of the ortholog group where nodes are proteins and edges can display orthologs, inparalogs, coorthologs, peripheral-core relationships, peripheral-peripheral relationships, or other similarities.

- e. Do all *Cryptococcus* species currently integrated in FungiDB contain this protein?

## 1 Phyletic distribution

▼ Phlyetic Distribution of Proteins ? Download

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

→ ☐ Hide zero counts

Cryptococc	×	?
<b>Eukaryota (EUKA)</b>	<b>294</b>	
<b>Fungi (FUNG)</b>	<b>140</b>	
<b>Basidiomycota (BASI)</b>	<b>33</b>	
Cryptococcus cf. gattii MF34 (ccfg)	1	
Cryptococcus depauperatus CBS 7841 (cdep)	1	
Cryptococcus depauperatus CBS 7855 (cdcb)	1	
Cryptococcus gattii CA1873 (cgac)	1	

- f. Is this gene found in both Ascomycetes and Basidomycetes?  
g. Does this protein have othologs in Archaea and Bacteria?

h. What is the most common PFAM domain associated with the proteins in this group?

▼ PFam Legend [Download](#)

Accession	Symbol	Description	Count	Legend
PF01545	Cation_efflux	Cation efflux family	290	
PF03645	Tctex-1	Tctex-1 family	2	
PF03102	NeuB	NeuB family	1	
PF01423	LSM	LSM domain	1	

i. Create a protein alignment for *Cryptococcus* genes. Use the 'List of All Proteins' table to run the Clustal Omega analysis on all *Cryptococcus* proteins.

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

6 rows (filtered from a total of 286)

Clustal Omega	Accession	Description	Organism	Taxonomy
<input checked="" type="checkbox"/>	ccfglD1P53_002977	unknown	Cryptococcus cf. gattii MF34	Fungi
<input checked="" type="checkbox"/>	cdepIL203_04836	Cation efflux protein [Source:UniProtKB/TrEMBL;Acc:A0A1E3ICE3]	Cryptococcus depauperatus CBS 7841	Fungi
<input checked="" type="checkbox"/>	cdcbIL204_05931	Cation:cation antiporter [Source:UniProtKB/TrEMBL;Acc:A0A1E3INZ6]	Cryptococcus depauperatus CBS 7855	Fungi
<input checked="" type="checkbox"/>	cgacIL314_06191	unknown	Cryptococcus gattii CA1873	Fungi
<input checked="" type="checkbox"/>	cgaelI306_06271	Unplaced genomic scaffold supercont1.232, whole genome shotgun sequence [Source:UniProtKB/TrEMBL;Acc:A0A0D0Y1M6]	Cryptococcus gattii EJB2	Fungi
<input checked="" type="checkbox"/>	cganIL311_05609	unknown	Cryptococcus gattii NT-10	Fungi

Please note: selecting a large number of proteins will take several minutes to align.

Output format:

j. Go to OG6\_129371 record page and open the cluster graph. (since this group has fewer members than OG6\_106189, the cluster graph display is more responsive)

Edge Type selection. All on by default. Hover over the selection boxes to highlight edges in red.

E-value slider: Increase the relationship strength by increasing the E-value cutoff for display.

Sequence List: table of all proteins in this ortholog group

Node details: click a cluster node to see more data in the side panel.

Cluster Graph: OG6\_129371 (20 proteins) ?

Back to Group page

Edge Options

Edge Type

☒ Ortholog ☒ Coortholog

☒ Inparalog ☒ Peripheral-Core

☒ Peripheral-Peripheral ☒ Other Similarities

E-Value Cutoff

Max E-Value: 1E -5

Node Options

Show Nodes By

☐ Taxa ☐ EC Numbers ☐ Pfam Domains

☒ Core/Peripheral

The core and peripheral proteins are colored as shown below

☒ Core (8) ☐ Peripheral (12)

Sequence List

Node Details

Showing 20 of 20 Organisms (20 Core and 0 Peripheral)

Accession	Taxon	Length	Description
acas-oid ACA1_146440	acas-oid	321	methyltransferase dom
acas ACA1_146440	acas	321	Methyltransf_11 domai
acid C1FA49	acid	267	Methyltransf_11 domai
bant A0A0F7RG63	bant	258	Methyltransf_11 domai
bsub O31474	bsub	253	Uncharacterized methy
cnit E4TK29	cnit	244	Methyltransf_11 domai
ecol P30866	ecol	207	Uncharacterized protei
fpro FPRO_08976	fpro	164	unknown
ftul Q5NGZ9	ftul	258	Methyltransf_11 domai
pgra-oid PGTG_08755	pgra-oid	1278	hypothetical protein
pgra PGTG_08755	pgra	1278	Methyltransf_25 domai
psae G3XCZ7	psae	187	Methyltransf_11 domai
psor VP01_458g1	psor	96	unknown
psor VP01_458g2	psor	1211	Methyltransf_25 domai
pstr PSTT_12989	pstr	1460	Methyltransferase dom
ptri PTTG_07820	ptri	1290	Methyltransf_25 domai
shfl A0A0H2VSZ8	shfl	206	Methyltransf_11 domai
vbra-oid Vbra_13896	vbra-oid	512	Ubiquinone/menaquinc
vbra Vbra_13896	vbra	512	Methyltransf_11 domai
yepi Q0WC68	yepi	173	Methyltransf_11 domai

Show Nodes by: change the meaning of the nodes in the display.

- k. Hover over the Peripheral-Core and Peripheral-Peripheral boxes to visualize those relationships.
- l. At what E-value cutoff do we lose all relationships in the cluster? How many group members are left at this E-value?
- m. How many EC numbers are associated with the group?
- n. Notice that the cluster has upper and lower groups. Can you find a trait that separates these groups?

2. **Find ortholog groups with specific phyletic patterns.** Phyletic Pattern Expression is a flexible and powerful way to identify ortholog groups with a certain conservation pattern. The pattern is used to identify groups based on whether proteins from specific taxa are present or absent.
  - a. Go to the Phyletic Pattern search in OrthoMCL.org



- b. Find ortholog groups that are Eukaryota specific. Arrange the taxonomic tree to include groups with proteins from Eukaryota but exclude proteins from Archaea and Bacteria. (Notice that using the tree creates an expression above. It's also possible to ignore the tree and write an expression for your phyletic pattern. See more information in the Learn tab for the search)

Expression:  Get Answer

**Key:** ● = no constraints | ✔ = must be in group | ✔ = at least one subtaxon must be in group | ✖ = must not be in group | \* = mixture of constraints

[expand all](#) | [collapse all](#)

- ▼ \* Root (ALL)
  - ▼ ● Eukaryota (EUKA)
    - ▶ ● Alveolates (ALVE)
    - ▶ ● Amoebozoa (AMOE)
    - ▶ ● Euglenozoa (EUGL)
    - ▶ ● Fungi (FUNG)
    - ▶ ● Metazoa (META)
    - ▶ ● Other Eukaryota (OEUK)
    - ▶ ● Viridiplantae (VIRI)
  - ▶ ✖ Archaea (ARCH)
    - ▶ ✖ Nitrosopumilus maritimus (strain SCM1) (nmar)
    - ▶ ✖ Crenarchaeota (CREN)
    - ▶ ✖ Euryarchaeota (EURY)
    - ▶ ✖ Korarchaeota (KORA)
    - ▶ ✖ Nanoarchaeota (NANO)
  - ▶ ✖ Bacteria (BACT)
    - ▶ ✖ Firmicutes (FIRM)
    - ▶ ✖ Other Bacteria (OBAC)
    - ▶ ✖ Proteobacteria (PROT)

Phyletic  
928,122 Ortholog Groups + Add a step

Step 1

Get Answer



Configure Search

Learn More

## Description

Find Ortholog Groups by phyletic pattern.

Phyletic Pattern Expression is a flexible and powerful way to identify ortholog groups with a certain conservation pattern. Proteins from specific taxa are present or absent. Also, the pattern finds groups with a certain copy number (

### Examples

These expressions find ortholog groups in which...

<b>hsap&gt;=5</b>	there are five or more human sequences
<b>hsap+ecol=2T</b>	both human and E. coli are present.
<b>hsap+ecol=1T</b>	only one species of human or E. coli is present.

- c. Find all groups that contain orthologs from at least one species of Ascomycota fungi (1T) but not from bacteria, archaea or metazoan (0T).

Expression: ASCO>=1T AND META=0T AND ARCH=0T AND BACT=0T

Get Answer

Key: ● = no constraints | ✔ = must be in group | ✔ = at least one subtaxon must be in group | ✖ = must not be in group | \* = mixture of constraints

expand all | collapse all

Type a taxonomic name

- Root (ALL)
  - Eukaryota (EUKA)
    - Alveolates (ALVE)
    - Amoebozoa (AMOE)
    - Euglenozoa (EUGL)
    - Fungi (FUNG)
      - Allomyces macrogynus ATCC 38327 (amac)
      - Catenaria anguillulae PL171 (cang)
      - Conidiobolus coronatus (strain ATCC 28846 / CBS 209.66 / NRRL 28638) (Delacroixia coronata) (ccor)
      - Rozella allomycis (strain CSF55) (rall)
      - Ascomycota (ASCO)
      - Basidiomycota (BASI)
      - Chytridiomycota (CHYT)
      - Microsporidia (MICR)
      - Mucoromycota (MUCO)
    - Metazoa (META)
    - Other Eukaryota (OEUK)
    - Viridiplantae (VIRI)
  - Archaea (ARCH)
    - Nitrosopumilus maritimus (strain SCM1) (nmar)
    - Crenarchaeota (CREN)
    - Euryarchaeota (EURY)
    - Korarchaeota (KORA)
    - Nanoarchaeota (NANO)
  - Bacteria (BACT)
    - Firmicutes (FIRM)
    - Other Bacteria (OBAC)
    - Proteobacteria (PROT)

Phyletic  
197,316 Ortholog Groups

Step 1

+ Add a step

Get Answer

- d. Interpret your results. Sort the result table by the Alveolata column (descending) and hover over the the Alveolata cell in the first row. Can you tell the distribution of Alveolata in the group?

Phylectic  
197,316 Ortholog Groups

Step 1

+ Add a step

197,316 Ortholog Groups

Revise this search

Ortholog Group Results

1 2 3 ... 9,866

Rows per page: 20

Ortholog Group	Total Number Proteins	Archaea	Bacteria	Alveolata	Amoeba	Euglenozoa
OG6_111091	243	0 / 27 (0%)	0 / 47 (0%)	140 / 141 (99%)	0 / 16 (0%)	0 / 74 (0%)
OG6_119698	146	0 / 27 (0%)	0 / 47 (0%)	138 / 141 (98%)	0 / 16 (0%)	0 / 74 (0%)
OG6_105220	677	0 / 27 (0%)	0 / 47 (0%)	132 / 141 (94%)	16 / 16 (100%)	73 / 74 (99%)
OG6_122551	154	0 / 27 (0%)	0 / 47 (0%)	114 / 141 (81%)	2 / 16 (13%)	0 / 74 (0%)
OG6_108921	420	0 / 27 (0%)	0 / 47 (0%)	84 / 141 (60%)	13 / 16 (81%)	0 / 74 (0%)
OG6_106907	492	0 / 27 (0%)	0 / 47 (0%)	136 / 141 (96%)	0 / 16 (0%)	0 / 74 (0%)
OG6_116220	171	0 / 27 (0%)	0 / 47 (0%)	136 / 141 (96%)	0 / 16 (0%)	0 / 74 (0%)

ALVEOLATA

Ciliates:

2 / 2

Apicomplexa

Haemosporida:

63 / 63

Coccidia:

50 / 51

Piroplasmida:

18 / 18

Other apicomplexa:

4 / 4

Other alveolata:

3 / 3

- e. Revise your search to find groups that: do not contain orthologs from Alveolates, Amoebozoa, Archaea, Bacteria and Ascomycota but contain at least one ortholog group from *Mucor circinelloides* f. *circinelloides* 1006PhL (mcic) AND *Mucor circinelloides* f. *lusitanicus* CBS 277.49 (mcir).

If you are getting frustrated trying to figure this one out, you have a right to be! If your results look different, hover over the search step and click to revise the parameter search. The cool thing about OrthoMCL is that has an added feature that allows you to enter an expression to define the phyletic pattern. This option provides additional flexibility. Try to figure out what expression to use to before peaking at the next page. (hint: start by assigning the “do not contain” parameter (x) using check boxes to Alveolates, Amoebozoa, archaea, bacteria and Ascomycetes. Next, use the expression window to add “AND” followed by specific criteria for *Mucor* spp. Use the learn more tab for more information. If you ran a search using just check boxes, the search will be configured to look for groups that do not contain orthologs from Alveolates, Amoebozoa, archaea, bacteria and Ascomycetes but do contain ortholog groups from all Mucoromycota.

Expression:

**Key:** ● = no constraints | ✔ = must be in group | ✔ = at least one subtaxon must be in group | ✖ = must not be in group | \* = mixture of constraints

expand all | collapse all

Type a taxonomic name

- \* Root (ALL)
  - \* Eukaryota (EUKA)
    - ✖ Alveolates (ALVE)
    - ✖ Amoebozoa (AMOE)
    - Euglenozoa (EUGL)
    - \* Fungi (FUNG)
      - Allomyces macrogynus ATCC 38327 (amac)
      - Catenaria anguillulae PL171 (cang)
      - Conidiobolus coronatus (strain ATCC 28846 / CBS 209.66 / NRRL 28638) (con)
      - Rozella allomycis (strain CSF55) (ra11)
      - ✖ Ascomycota (ASCO)
      - Basidiomycota (BASI)
      - Chytridiomycota (CHYT)
      - Microsporidia (MICR)
      - ▼ Mucoromycota (MUCO)
        - Lichtheimia corymbifera JMRC:FSU:9682 (lcor)
        - **Mucor circinelloides 1006PhL (mcic)**
        - **Mucor lusitanicus CBS 277.49 (mcir)**
        - Phycomyces blakesleeanus NRRL 1555(-) (pbla)
        - Rhizophagus irregularis A1 (DAOM-664342) (rira)
        - Rhizophagus irregularis C2 (rirc)
        - Rhizophagus irregularis DAOM 181602=DAOM 197198 (rhiz)
        - Rhizophagus irregularis DAOM 181602=DAOM 197198 (old build 2018-01-30) (rhiz-old)
        - Rhizopus delemar RA 99-880 (rde1)
        - Rhizopus delemar RA 99-880 (old build 2015-03-23) (rde1-old)
        - Rhizopus microsporus var. microsporus ATCC 52814 (rmma)
    - Metazoa (META)
    - Other Eukaryota (OEUK)
    - Viridiplantae (VIRI)
    - ✖ Archaea (ARCH)
    - ✖ Bacteria (BACT)

This is as far as you can get using the tree since the button beside mcic and mcir do not accommodate the yellow checks. You must alter the Expression with specific

Phyletic  
1,564 Ortholog Groups

+ Add a step


Step 1

ALVE=0T AND AMOE=0T AND ASCO=0T AND ARCH=0T AND BACT=0T AND mcic=1T and mcir=1T

<https://orthomcl.org/orthomcl/app/workspace/strategies/import/c1883ab75f86053d>

Useful information:

All VEuPathDB genomics sites (e.g., FungiDB) have an integrated phyletic pattern search that uses OrthoMCL to return lists of genes. For example, you use the “Orthology Phylogenetic Profile” search to identify genes in your organism of interest that are restricted in their profile. For example, you frequently want to identify genes that are conserved among organisms in your genus of interest but not present in the host as these genes may make good drug targets or vaccine candidates.



**FungiDB**  
Fungal & Oomycete Informatics Resources

Search for...

phyl

Genes

Orthology and synteny

Orthology Phylogenetic Profile

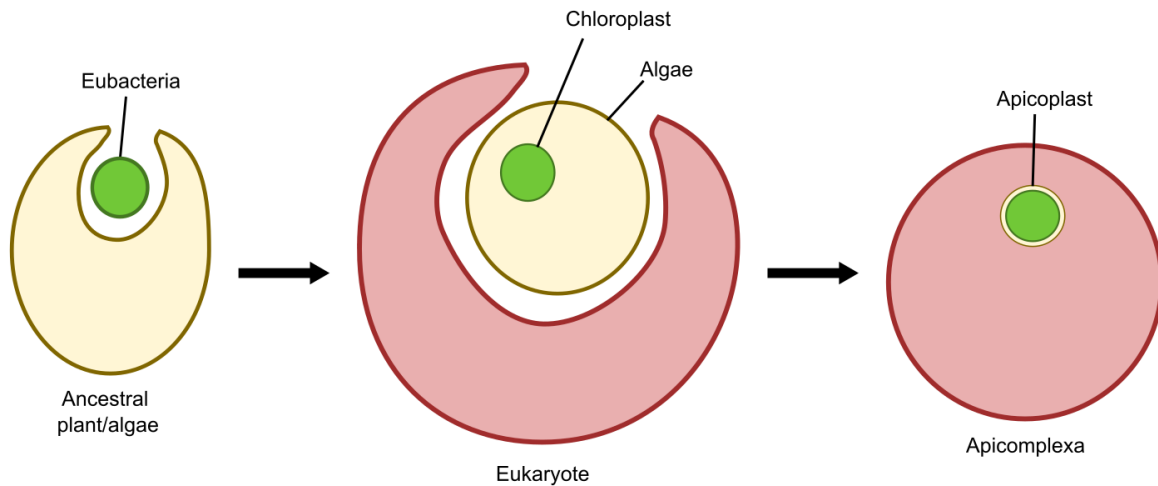
Phylogenetic

Find genes that have an orthology-based phylogenetic profile that you specify.

**3. Identify apicoplast targeted genes in *Toxoplasma* and *Neurospora*.** Note: For this exercise use <https://veupathdb.org/veupathdb/app>

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 algae was then engulfed by the ancestor of all apicomplexans. Thus, an apicoplast organelle arose with four membranes.



- a. 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 types of search by text query.*

The screenshot shows the VeupathDB web interface. At the top, the 'Searches' tab is highlighted with a red circle. Below the search bar, a dropdown menu for the query 'pf' is open, with 'Pfal 3D7 Subcellular Localization' selected and indicated by a red arrow. On the right, the 'Identify Genes based on P.f. Subcellular Localization' section shows the 'Localization' filter set to 'Apicoplast'. The 'Get Answer' button is circled in red. Below this, a yellow box displays 'Subcell Loc 499 Genes' for 'Step 1', and a dashed box shows '+ Add a step'.

- b. Expand your list of potentially Apicoplast targeted proteins by adding a GO terms search for the term “apicoplast” or the GO ID: “GO:0020011” in *P. falciparum* 3D7 (Which Boolean operation should you use? Union or intersect?)

← Add a step to your search strategy ?

**Combine with other Genes**

Step 1 Step 2

**Transform into related records**

Step 1 Step 2

**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

- 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 | v 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**

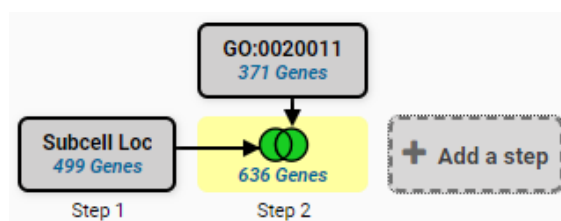
☒ Curated  
☒ Computed  
 select all | clear all

**Limit to GO Slim terms**

☐ Yes  
☒ No

**GO Term or GO ID**

GO:0020011 : apicoplast : 7 X



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

**Transform into related records**

GO:0020011  
371 Genes

636 Genes  
Step 2

Step 3

**Organism**

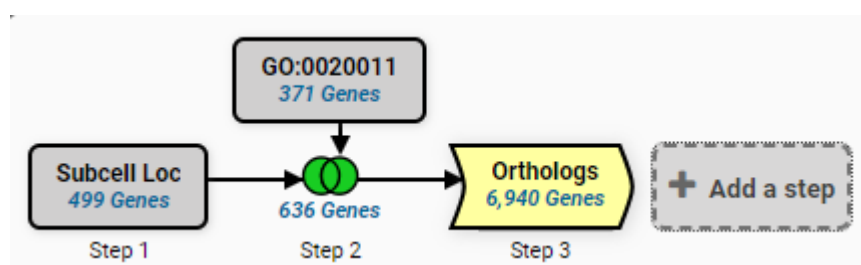
17 selected, out of 675  
select all | clear all | expand all | collapse all

Filter list below...

☐ Reference only

- ☐ 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

Run Step



- d. Although *Cryptosporidium* is an apicomplexan parasite it has lost its apicoplast! Use this fact to refine your results from the above search and remove genes that also have orthologs in *Cryptosporidium*.

Hint: try subtracting out any orthologs present in *Cryptosporidium*. You will need to use a nested strategy. First retrieve all *Cryptosporidium* genes with the Genes by Taxonomy search and then transform these to their *Toxoplasma* and *Neospora* orthologs for the subtraction to complete. Think about what kind of intersection you should be using!

← Add a step to your search strategy ?

**Combine with other Genes**

Step 3 Step 4

**Transform into related records**

Step 3 Step 4

**1 Choose how to combine with other Genes**

☐ 3 INTERSECT 4
 ☐ 3 UNION 4
 ☒ 3 MINUS 4
 ☐ 4 MINUS 3

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

☒ A new search
 ☐ An existing strategy
 ☐ My basket

- Taxonomy
- Q Organism



## Add a step to your search strategy

The results will be subtracted from | v the resu

Configure Search Learn More View Data Sets Used

Reset values to default

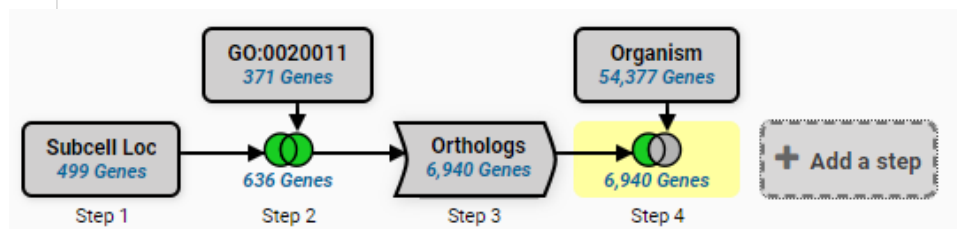
**Organism**

14 selected, out of 675

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

? ☐ Reference only

- ☐ 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]



My Organism Preferences (785 of 785) disabled

## My Search Strategies

Opened (1) All (1) Public (11) Help

Unnamed Search 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

+ Ad Give this search a weight

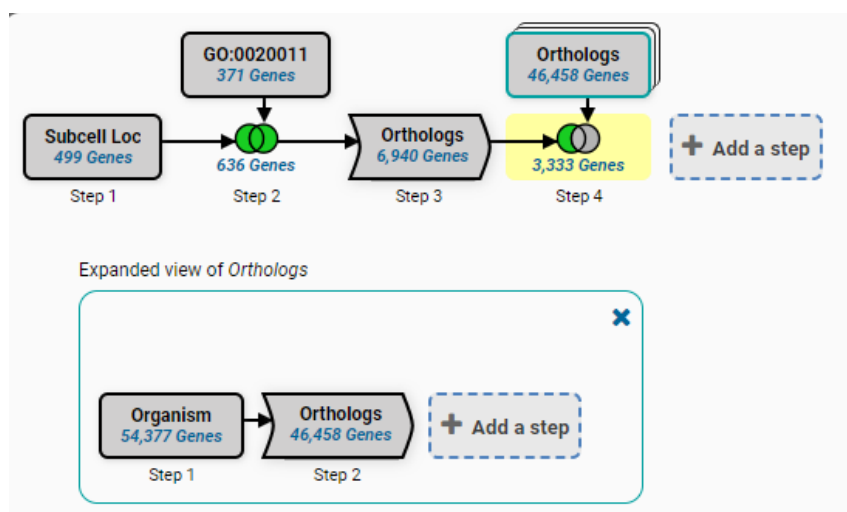
### Organism

17 selected, out of 675

select all | clear all | expand all | collapse all

Filter list below...

- ☐ Amoebozoa
- ☒ Apicomplexa
  - ☐ Aconoidasida
  - ☒ Conoidasida
    - ☒ Coccidia
      - ☐ Cryptosporidiidae
      - ☐ Elmeriidae
      - ☒ 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



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

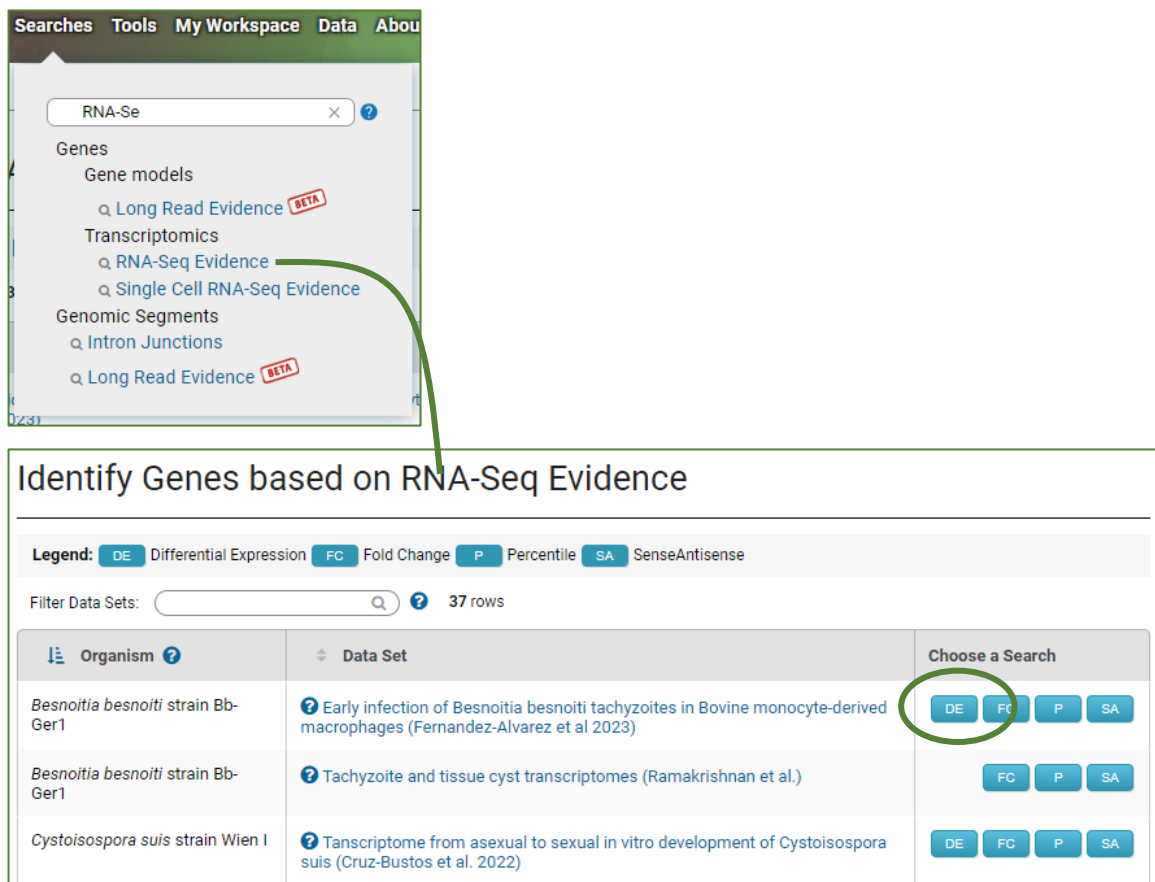
<https://veupathdb.org/veupathdb/app/workspace/strategies/import/543f14bfab645f7e>



4. Determine functional enrichment of the set of genes that are upregulated at least 2-fold in an early infection of *Besnoitia besnoiti* tachyzoites in Bovine monocyte-derived macrophage. For this exercise use [ToxoDB](#).

ToxoDB has RNA-sequence data from a study of the early interaction between *B. besnoiti* tachyzoites and primary bovine monocyte-derived macrophages *in vitro*. [Data set record](#). Dual transcriptomic profiling of *B. besnoiti* tachyzoites and *Bos taurus* macrophages was conducted at early infection 4 h and 8 h post infection by high-throughput RNA sequencing. Bovine macrophages inoculated with heat-killed tachyzoites (MO-hkBb) and non-infected macrophages (MO) were used as controls. In this exercise we will find a list of genes that are differentially expressed based on the RNA Seq data and investigate the functional enrichment (if any) for the gene set.

- a. Run a differential expression search looking for genes that are upregulated at least 2-fold at 8 hours post infection compared to 4 hours with  $p < 0.001$ . Navigate to the RNA-seq searches in ToxoDB and choose the DE search for **Early infection of *Besnoitia besnoiti* tachyzoites in Bovine monocyte-derived macrophages (Fernandez-Alvarez et al 2023)**.



**Searches** Tools My Workspace Data About

RNA-Se

Genes

Gene models

q Long Read Evidence BETA

Transcriptomics

q RNA-Seq Evidence

q Single Cell RNA-Seq Evidence

Genomic Segments

q Intron Junctions

q Long Read Evidence BETA

**Identify Genes based on RNA-Seq Evidence**

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

Filter Data Sets: 37 rows

Organism	Data Set	Choose a Search
<i>Besnoitia besnoiti</i> strain Bb-Ger1	Early infection of <i>Besnoitia besnoiti</i> tachyzoites in Bovine monocyte-derived macrophages (Fernandez-Alvarez et al 2023)	DE FC P SA
<i>Besnoitia besnoiti</i> strain Bb-Ger1	Tachyzoite and tissue cyst transcriptomes (Ramakrishnan et al.)	FC P SA
<i>Cystoisospora suis</i> strain Wien I	Transcriptome from asexual to sexual in vitro development of <i>Cystoisospora suis</i> (Cruz-Bustos et al. 2022)	DE FC P SA

- b. Arrange the search to return genes upregulated at least 2-fold ( $p < 0.001$ ) in the 8hr Bbes in MO sample compared to 4hr Bbes in MO.

**Experiment**

☒ Early infection of *Besnoitia besnoiti* tachyzoites in Bovine monocyte-derived macrophages - Sense  
☐ Early infection of *Besnoitia besnoiti* tachyzoites in Bovine monocyte-derived macrophages - Antisense

**Reference Sample**

☒ 4 hpi Bbes in MO  
☐ 8 hpi Bbes in MO

**Comparator Sample**

☐ 4 hpi Bbes in MO  
☒ 8 hpi Bbes in MO

**Direction**

up-regulated

**fold difference >=**

2

**adjusted P value less than or equal to**

0.001

Get Answer

c. From the result page, choose the Analyze Results tab to go to the Enrichment Tool.

**Bbes infection in Bbos macrop...**  
135 Genes

Step 1

135 Genes (118 ortholog groups) [Revise this search](#)

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

Rows per page: 20

Gene ID	Transcript ID	Organism	Product Description	Fold Change	Adjusted P
BESB_010080	mrna.BESB_010080	<i>Besnoitia besnoiti</i> strain Bb-Ger1	rhostry kinase family protein ROP37 (incomplete catalytic triad)	7.28	5.8381
BESB_012870	mrna.BESB_012870	<i>Besnoitia besnoiti</i> strain Bb-Ger1	hypothetical protein	6.83	3.3591
BESB_073760	mrna.BESB_073760	<i>Besnoitia besnoiti</i> strain Bb-Ger1	hypothetical protein	5.71	4.1921
BESB_006260	mrna.BESB_006260	<i>Besnoitia besnoiti</i> strain Bb-Ger1	alpha amylase, catalytic domain-containing protein	5.36	1.1281
BESB_020840	mrna.BESB_020840	<i>Besnoitia besnoiti</i> strain Bb-Ger1	hypothetical protein	5.27	5.1741






Analyze your Gene results with a tool below.

**GO**  
Gene Ontology Enrichment







**Metabolic Pathway Enrichment**

**kinase  
phosphatase  
exported  
membrane**  
Word Enrichment

- d. Run a GO enrichment analysis on the Biological Process ontology terms associated with your gene list.

Organism  Besnoitia besnoiti strain Bb-Ger1  
 Ontology    
☒ Biological Process  
☐ Cellular Component  
☐ Molecular Function  
 Evidence    
☒ Computed  
☒ Curated  
[select all](#) | [clear all](#)  
 Limit to GO Slim terms    
☒ No  
☐ Yes  
 P-Value cutoff  0.05 (0 - 1)

Analysis Results: 27 rows [Open in Revigo](#) [Show Word Cloud](#) [Download](#)

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 	Bonferroni 
GO:0006468	protein phosphorylation	200	19	9.5	5.75	7.78	3.02e-10	2.74e-8	5.25e-8
GO:0016310	phosphorylation	225	20	8.9	5.38	7.30	3.15e-10	2.74e-8	5.49e-8
GO:0006796	phosphate-containing compound metabolic process	372	23	6.2	3.74	5.02	1.57e-8	7.17e-7	2.72e-6
GO:0006793	phosphorus metabolic process	373	23	6.2	3.73	5.01	1.65e-8	7.17e-7	2.87e-6
GO:0036211	protein modification process	363	22	6.1	3.67	4.84	5.03e-8	1.46e-6	8.75e-6

- e. What is the top enriched GO term from this analysis? Does this make sense for an enrichment analysis of the biological processes associated with your differentially expressed genes? Notice that the p-value with Benjamini or Bonferroni correction is very low. What do each of the columns in the analysis table represent? Hint: move your mouse over the question mark next to each column header

- 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.

- f. Click on the link in the 'Genes in your result with this term' column. This will create a one-step strategy that returns only the genes with this GO term.
- g. Perform the enrichment analysis on the other ontologies. Its possible to modify the paramaters of the current enrichment analysis but you will overwrite your current results. To run a new analysis and save the old, start with the Analyze Results tab.

- h. Is there a cellular location or molecular function that is enriched? Do these support the biological process enrichments?