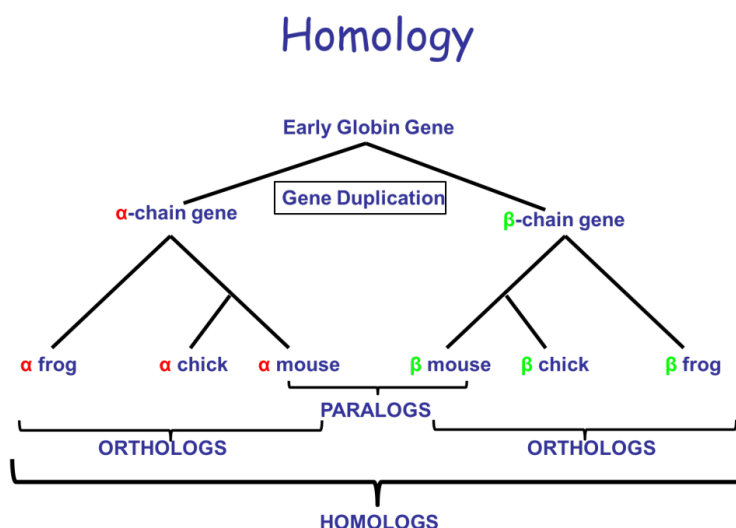


## Exploring Orthology in VEuPathDB

### Learning objectives:

- Use orthology information on gene record pages to infer function of a gene whose protein product is undefined.
- Explore individual ortholog group pages.
- Explore the group phylogenetic tree.



### 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. [OG7\\_0007567](#)). 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 149 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 149 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. [OGR7\\_0007343](#))

The OrthoMCL website ([orthomcl.org](#)), now available in a updated beta version, <https://beta.orthomcl.org/orthomcl.b7/app>, 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 map their own set of proteins (e.g. protein sequences derived from a genome sequence of an organism) to OrthoMCL groups. See the [Map Proteins to OrthoMCL](#) tool.

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

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

- 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.
- 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.

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

- 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.
- Examine this gene's Ortholog Group on OrthoMCL to learn about orthologs from organisms outside FungiDB. Return to the 'Orthology and Synteny' section of the gene page and observe the Ortholog Group link, OG6\_106189. For the purposes of this


section of the exercise, go to the Ortholog Group page on [beta.orthomcl.org](https://beta.orthomcl.org/OG7_0001789), [OG7\\_0001789](https://beta.orthomcl.org/OG7_0001789)

Look at the keywords in the header section for this group - what functions are mentioned?

The OrthoMCL group page is divided into 4 sections.



1. **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.'
2. **Group summary:** This section provides some statistics about the internal dispersion of proteins in the group. Is this group tighter or more dispersed than most OrthoMCL groups? The **Similar Groups** section provides additional information from closely related orthogroups that may have useful annotations.
3. **Summary of 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.
4. **List of Proteins with Phylogenetic tree:** Lists all proteins in the ortholog group. The table also is a tool for running a Clustal Omega analysis of selected proteins. The phylogenetic tree provides a dynamic visualization of the ortholog group which indicates patterns of speciation and gene duplication as well as the distribution of functional domains across the proteins in the group. Filters allow direct comparisons across species of interest.

**Phyletic distribution:** Do all *Cryptococcus* species currently integrated in FungiDB contain this protein (uncheck the Hide zero counts button)? How many copies (paralogs) are found in each genome?

▼ Phyletic Distribution of Proteins  [Download](#)

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

☐ Hide zero counts

<b>Eukaryota (EUKA)</b>	<b>641</b>
<b>Fungi (FUNG)</b>	<b>282</b>
<b>Basidiomycota (BASI)</b>	<b>46</b>
Cryptococcus cf. gattii MF34 (ccfg)	1
Cryptococcus depauperatus CBS 7841 (cdep)	1
Cryptococcus depauperatus CBS 7855 (cdcb)	1
Cryptococcus gattii CA1873 (cgac)	1

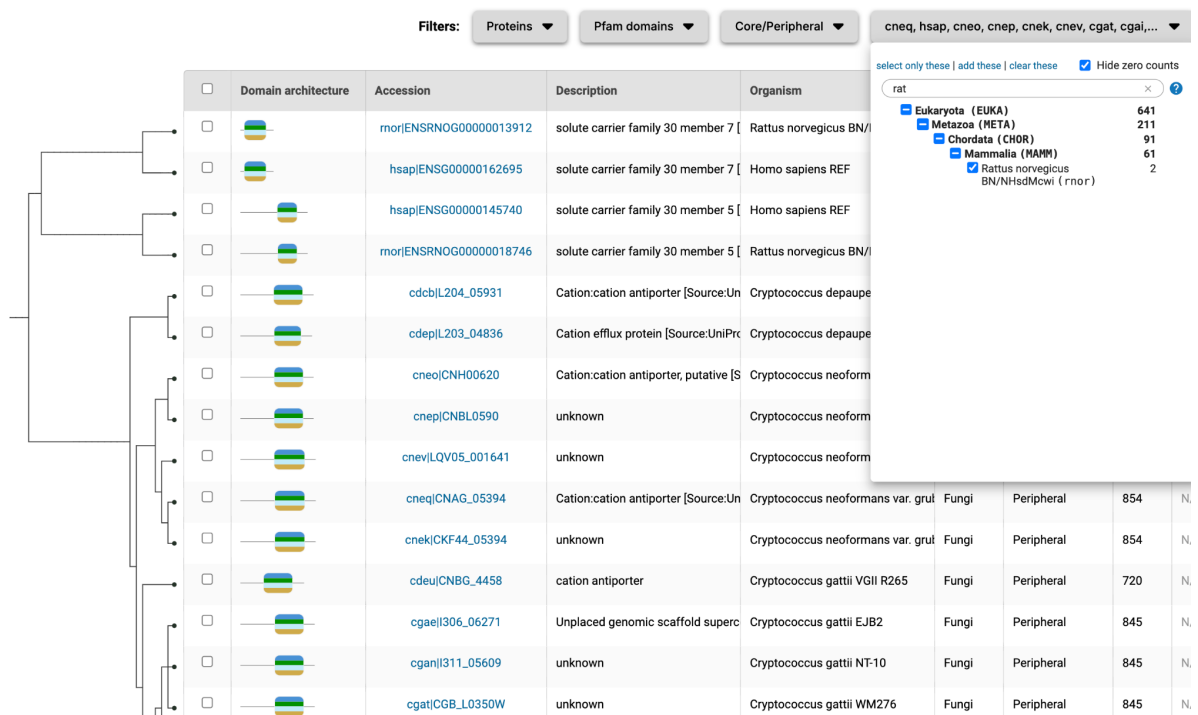
- e. Is this gene found in both Ascomycetes and Basidiomycetes?
- f. Does this protein have orthologs in Archaea and Bacteria?
- g. What is the most common Pfam domain associated with the proteins in this group?

▼ Summary of Pfam domains [Download](#)

Search this table... 4 rows

Accession	Description	Count	Legend
PF01545	Cation efflux family	621	
PF03645	Tctex-1 family	2	
PF07993	Male sterility protein	1	
PF03102	NeuB family	1	

- h. Look at the phylogenetic tree. Use the **Organism** filter to limit the tree to just *Cryptococcus* and *Homo sapiens* genes. Humans have two gene copies in this ortholog group while most fungi have just one. Is one of the two human copies closer to the fungal gene? Add *Rattus norvegicus* to the tree. Now you can see that there has been a gene duplication in the mammal clade, so that the two copies in human each have close functionally similar orthologs in rat.



- 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.

#### 4 List of proteins

▼ List of All Proteins [Download](#)

This section features a Clustal Omega alignment tool (max 1000 sequences) and a tree visualization of proteins in this group. The proteins can be filtered in a number of ways,

- Filter via text search on any/specific columns in the table.

Note: The ortholog group's phylogeny has been pruned to display only the currently filtered proteins. This may differ from a tree constructed *de novo* using only these sequences.

cryptococ X 14 rows (filtered from a total of 647)

Filters: Prot

<input checked="" type="checkbox"/>	Domain architecture	Accession	Description	Organism
<input checked="" type="checkbox"/>		cdcblL204_05931	Cation:cation antiporter [Source:UniProtKB/TrEMBL;Acc:ADA1	Cryptococcus depauperatus CBS 7855
<input checked="" type="checkbox"/>		cdeplL203_04836	Cation efflux protein [Source:UniProtKB/TrEMBL;Acc:ADA1E3]	Cryptococcus depauperatus CBS 7841
<input checked="" type="checkbox"/>		cneoCNH00620	Cation:cation antiporter, putative [Source:UniProtKB/TrEMBL/	Cryptococcus neoformans var. neoformans JEC21
<input checked="" type="checkbox"/>		cneplCNBL0590	unknown	Cryptococcus neoformans var. neoformans B-3501A
<input checked="" type="checkbox"/>		cnevlQV05_001641	unknown	Cryptococcus neoformans strain VNII
<input checked="" type="checkbox"/>		cneqCNAG_05394	Cation:cation antiporter [Source:UniProtKB/TrEMBL;Acc:J9VZ	Cryptococcus neoformans var. grubii H99
<input checked="" type="checkbox"/>		cnekCKF44_05394	unknown	Cryptococcus neoformans var. grubii KN99
<input checked="" type="checkbox"/>		cdeuCNBG_4458	cation antiporter	Cryptococcus gattii R265
<input checked="" type="checkbox"/>		cgael306_06271	Unplaced genomic scaffold supercont1.232, whole genome s	Cryptococcus gattii EJB2
<input checked="" type="checkbox"/>		cganl311_05609	unknown	Cryptococcus gattii NT-10
<input checked="" type="checkbox"/>		cgatlCGB_L0350W	unknown	Cryptococcus gattii WM276
<input checked="" type="checkbox"/>		cgall308_06163	Unplaced genomic scaffold supercont2.21, whole genome sh	Cryptococcus gattii VGIV IND107
<input checked="" type="checkbox"/>		ccfplD1PS3_002977	unknown	Cryptococcus cf. gattii MF34
<input checked="" type="checkbox"/>		cgacil314_06191	unknown	Cryptococcus gattii CA1873

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

Output format: Mismatches highlighted

[Run Clustal Omega for selected proteins](#)