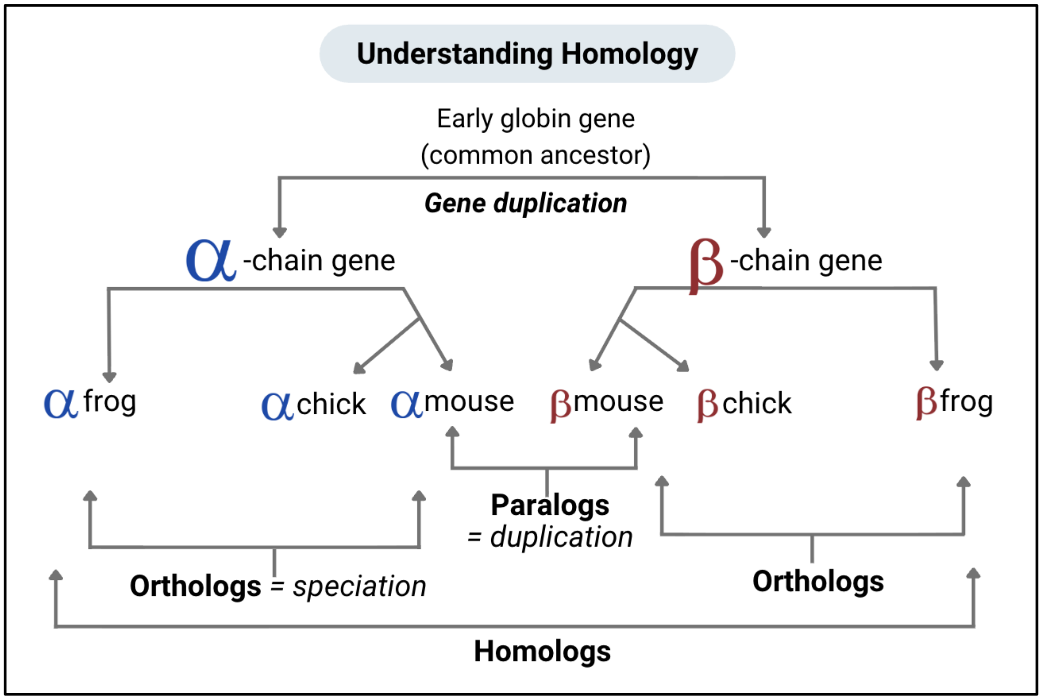
VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

Orthology on Gene Record Pages & OrthoMCL

Learning objectives

- Use orthology information on gene record pages to infer the function of a gene whose protein product is undefined
- Navigate individual ortholog group pages
- Examine 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 and information is collated on **ortholog group pages** (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**.

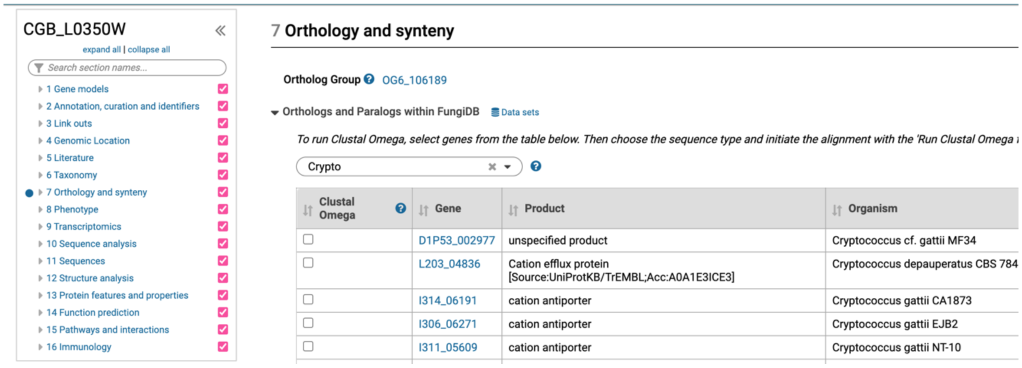
OrthoMCL contains two sets of genomes. A Core set of 149 genomes have been chosen as well annotated reference organisms that broadly represent the major branches of the tree of life. The OrthoMCL algorithm uses BLAST to calculate pairwise distances among all proteins in the 149 core genomes, normalizes the scores for sequence length and evolutionary distance, then uses MCL clustering (Dongen 2000; www.micans.org/mcl) 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) 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 page.

Part 1: Orthology on Gene Record Pages

1. For this exercise we will start at FungiDB. Go to the FungiDB gene record page for CGB_L0350W, a protein in *Cryptococcus gatti*. Although this gene is annotated as "hypothetical", examining orthologs and ontology information may suggest protein function.

2. Navigate to the *Orthology and Synteny* section on this page *(Hint: you can use the content navigation tool on the left pane to find this 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 more informative gene product descriptions?
   What function might CGB_L0350W have, based on this table?



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

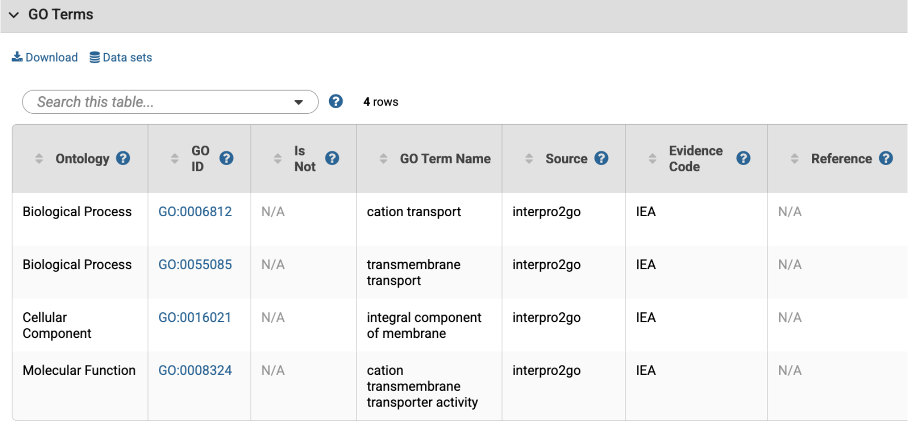
Orthologs and Paralogs within FungiDB   Data sets

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

Crypto

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

3. Move to the *Function prediction* section of the gene page and examine the GO Slim and GO Terms tables. What GO terms are represented here? Do they match with the function suggested by the orthology section? What is the source of the GO terms? What is the evidence code?
   Annotations can be assigned based on direct evidence, as from an experimental (EXP), 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.



GO Terms

Download   Data sets

Search this table...   4 rows

Ontology	GO ID	Is Not	GO Term Name	Source	Evidence Code	Reference
Biological Process	GO:0006812	N/A	cation transport	interpro2go	IEA	N/A
Biological Process	GO:0055085	N/A	transmembrane transport	interpro2go	IEA	N/A
Cellular Component	GO:0016021	N/A	integral component of membrane	interpro2go	IEA	N/A
Molecular Function	GO:0008324	N/A	cation transmembrane transporter activity	interpro2go	IEA	N/A

4. How could we learn about orthologs of this hypothetical protein from organisms outside FungiDB?

   We could look at its orthogroup in OrthoMCL.org. To do this, return to the *Orthology and Synteny* section of the gene page and look for the link next to Ortholog Group (OrthoMCL 7). Click the link “Search for CGB_L0350W”. Follow the link to get to the ortholog group page OG7_0001789.

   How many orthologs are shown on the orthogroup page? Is this different from the orthologs shown in the gene page (within FungiDB)?

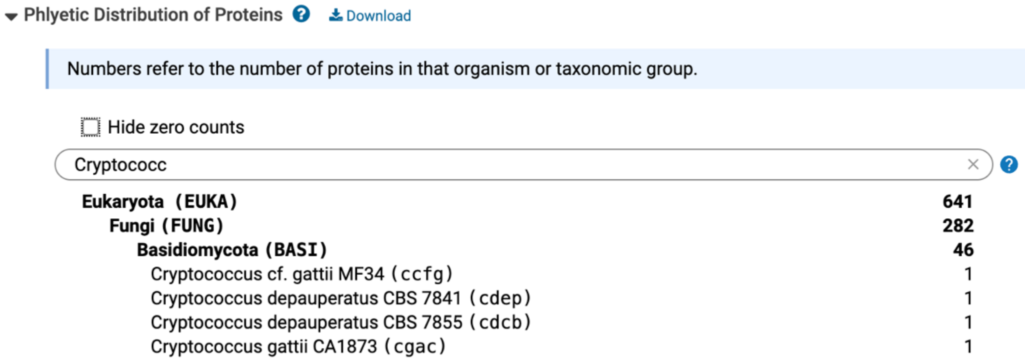
Part 2: Orthogroup Pages on OrthoMCL

Look at the keywords in the header section for the ortholog group page OG7_0001789.
What functions are mentioned?

We will explore the different sections of the OrthoMCL group page.

1. **Phyletic distribution**: Numbers in the table refer to the number of proteins in that organism or taxonomic group.

   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

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

□ Hide zero counts

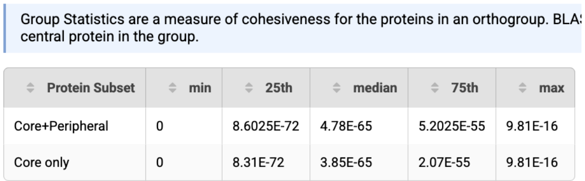
Cryptococc

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

   Is this gene found in both Ascomycetes and Basidomycetes?

   Does this protein have orthologs in Archaea and Bacteria?

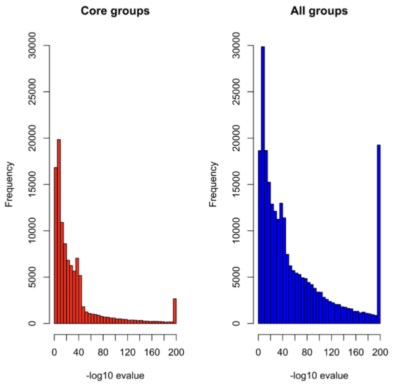
2. **Group summary**: This section provides some statistics about the internal dispersion of proteins in the group. BLAST e-values are presented to evaluate group cohesiveness, asking: “How strongly and consistently do the proteins in this orthogroup match each other?”



Group Statistics are a measure of cohesiveness for the proteins in an orthogroup. BLAS
central protein in the group.

Protein Subset	min	25th	median	75th	max
Core+Peripheral	0	8.6025E-72	4.78E-65	5.2025E-55	9.81E-16
Core only	0	8.31E-72	3.85E-65	2.07E-55	9.81E-16

   The graphs tell you about all the groups in OrthoMCL. The distribution of e-values for orthogroups representing the core proteomes is shown in red and for core and peripherals together is shown in blue.

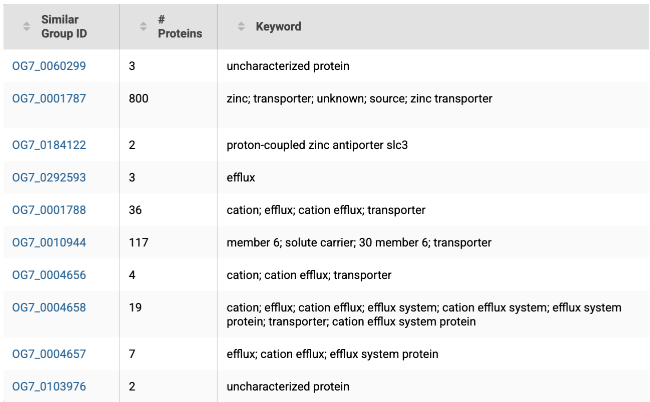


The histograms show the distribution of the median percent identity within Core (red) and All (blue) ortholog groups.

   The e-values in the table tell you the statistics for this orthogroup. Look at the median e-value in the table and compare it to the graphs for all orthogroups. Is this group tighter or more dispersed than most OrthoMCL groups?

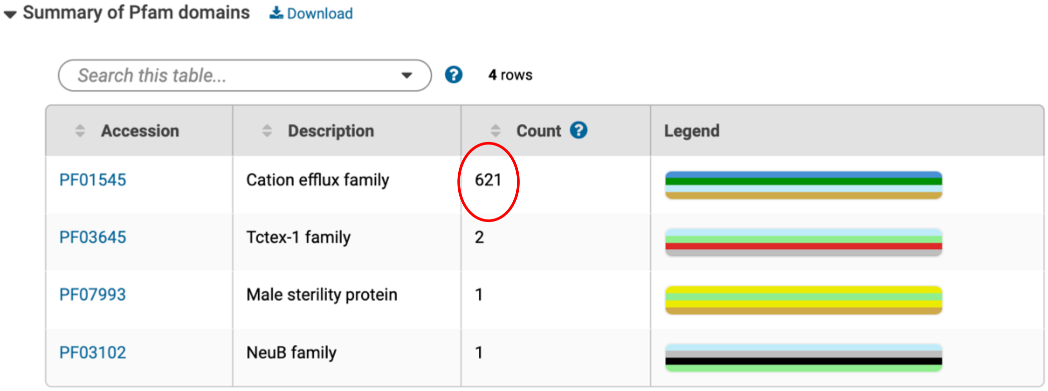
   The median e-value in the table tells you that this group has a better e-value than a large percentage of the groups, so it should be a pretty cohesive group. High cohesiveness suggests that the proteins are true orthologs with conserved function and that the group is well-defined phylogenetically.

3. The **Similar Groups** section provides additional information from closely related orthogroups that may have useful annotations. Do the keywords in this “similar groups” table match the keywords for this orthogroup?



Similar Group ID	# Proteins	Keyword
OG7_0060299	3	uncharacterized protein
OG7_0001787	800	zinc; transporter; unknown; source; zinc transporter
OG7_0184122	2	proton-coupled zinc antiporter slc3
OG7_0292593	3	efflux
OG7_0001788	36	cation; efflux; cation efflux; transporter
OG7_0010944	117	member 6; solute carrier; 30 member 6; transporter
OG7_0004656	4	cation; cation efflux; transporter
OG7_0004658	19	cation; efflux; cation efflux; efflux system; cation efflux system; efflux system protein; transporter; cation efflux system protein
OG7_0004657	7	efflux; cation efflux; efflux system protein
OG7_0103976	2	uncharacterized protein

4. **Summary of Pfam domains**: Provides a list of Pfam domains that are found in the proteins within the ortholog group. Pfam is a database of protein families that includes functional descriptions and multiple sequence alignments of conserved domains generated using Hidden Markov Models. What is the most common Pfam domain associated with the proteins in this group?



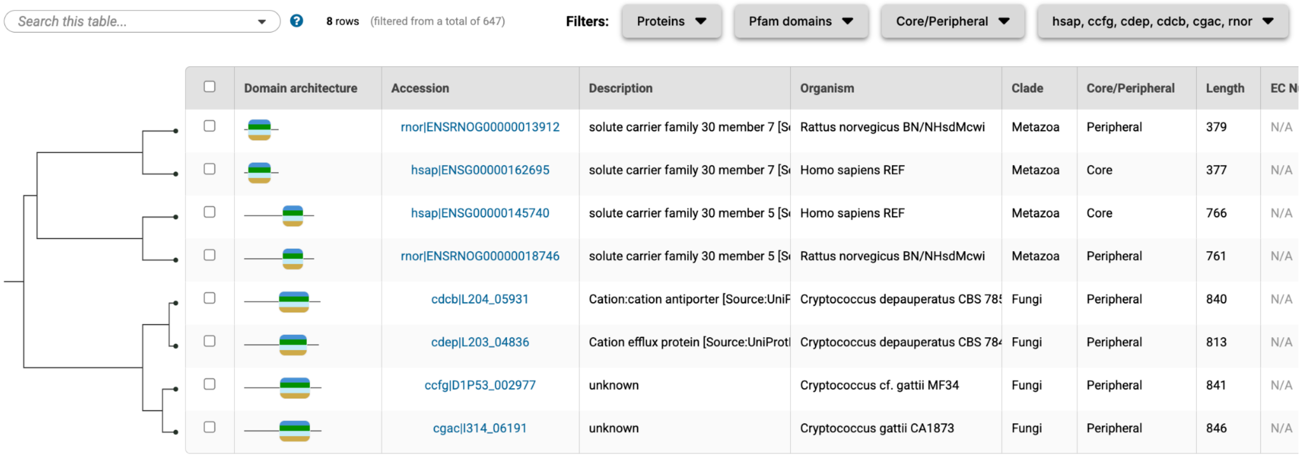
▾ Summary of Pfam domains Download

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	

5. **List of Proteins with Phylogenetic tree**: Lists all proteins in the ortholog group. 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 focus on specific taxonomic groups or direct comparisons among proteins of interest.
Look at the phylogenetic tree. Use the Organism filter to limit the tree to just *Homo sapiens* genes and the first 4 *Cryptococcus* species. Humans have two gene copies in this ortholog group while most fungi in the table 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 humans each have close, functionally similar orthologs in rats.

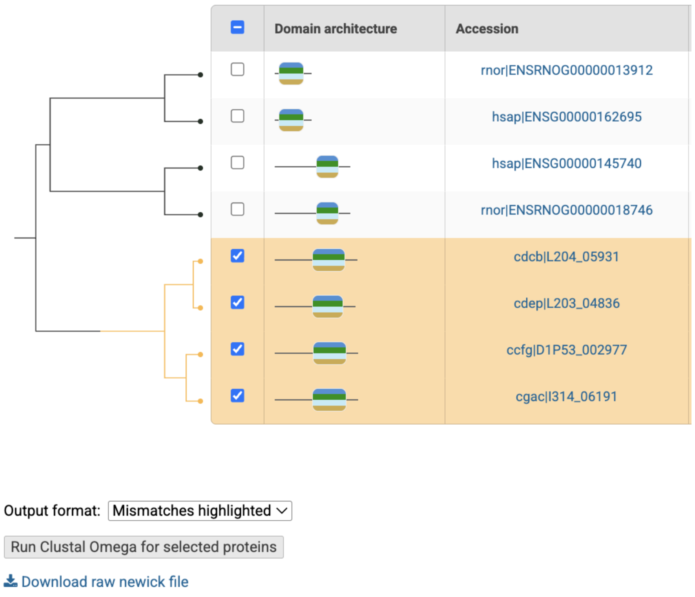


Search this table... 8 rows (filtered from a total of 647) Filters: Proteins ▾ Pfam domains ▾ Core/Peripheral ▾ hsap, ccfg, cdep, cdcb, cgac, rnor ▾

☐	Domain architecture	Accession	Description	Organism	Clade	Core/Peripheral	Length	EC N
☐		rnor|ENSRNOG00000013912	solute carrier family 30 member 7 [S	Rattus norvegicus BN/NHsdMcwi	Metazoa	Peripheral	379	N/A
☐		hsap|ENSG00000162695	solute carrier family 30 member 7 [S	Homo sapiens REF	Metazoa	Core	377	N/A
☐		hsap|ENSG00000145740	solute carrier family 30 member 5 [S	Homo sapiens REF	Metazoa	Core	766	N/A
☐		rnor|ENSRNOG00000018746	solute carrier family 30 member 5 [S	Rattus norvegicus BN/NHsdMcwi	Metazoa	Peripheral	761	N/A
☐		cdcb|L204_05931	Cation:cation antiporter [Source:UniP	Cryptococcus depauperatus CBS 78	Fungi	Peripheral	840	N/A
☐		cdep|L203_04836	Cation efflux protein [Source:UniProt	Cryptococcus depauperatus CBS 78	Fungi	Peripheral	813	N/A
☐		ccfg|D1P53_002977	unknown	Cryptococcus cf. gattii MF34	Fungi	Peripheral	841	N/A
☐		cgac|I314_06191	unknown	Cryptococcus gattii CA1873	Fungi	Peripheral	846	N/A

6. The table also has a tool for running a **Clustal Omega analysis** of selected proteins. You may want to create a protein alignment for *Cryptococcus* genes to identify conserved regions, compare species, detect important residues or mutations, improve functional annotation, or support broader evolutionary or comparative genomic analyses.

   To do this, select *Cryptococcus* proteins in the table above and run Clustal Omega analysis on them.



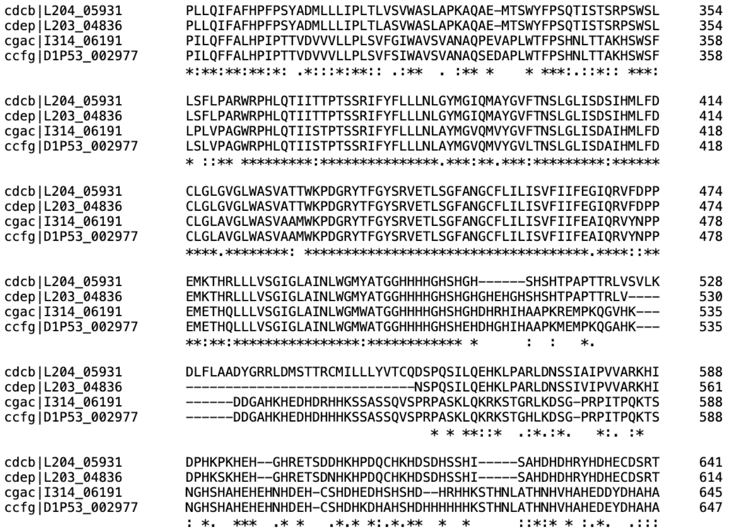
⊟	Domain architecture	Accession
□		rnor|ENSRNOG00000013912
□		hsap|ENSG00000162695
□		hsap|ENSG00000145740
□		rnor|ENSRNOG00000018746
☑		cdcb|L204_05931
☑		cdep|L203_04836
☑		ccfg|D1P53_002977
☑		cgac|I314_06191

Output format: Mismatches highlighted

   Note that you can also download a raw newick file of the phylogenetic tree. A **Newick file** is a plain-text format that represents phylogenetic trees using parentheses, commas, and optional branch lengths. It provides a compact way to encode evolutionary relationships so they can be analyzed or visualized by phylogenetic software.

7. The **alignment** will open in a new browser tab. Can you identify the conserved motifs? These are indicated by asterisks (*)

   Blocks of **** (conserved stretches) can indicate active sites, binding motifs, protein domains, structural cores, etc. These regions are often functionally important across species.



```
cdcb|L204_05931        PLLQIFAFHPFPSYADMLLLIPLTLVSVWASLAPKAQAE-MTSWYFPSQTISTSRPSWSL     354
cdep|L203_04836        PLLQIFAFHPFPSYADMLLLIPLTLASVWASLAPKAQAE-MTSWYFPSQTISTSRPSWSL     354
cgac|I314_06191        PILQFFALHPIPTTVDVVVLLPLSVFGIWAVSVANAQPEVAPLWTFPSHNLTTAKHSWSF     358
ccfg|D1P53_002977      PILQFFALHPIPTTVDVVVLLPLSVFSIWAVSVANAQSEDAPLWTFPSHNLTTAKHSWSF     358
                       *:**:**:**:*: .*:::*:**:: .:**  . :** *    * ***:.::*:: ***:

cdcb|L204_05931        LSFLPARWRPHLQTIITTPTSSRIFYFLLLNLGYMGIQMAYGVFTNSLGLISDSIHMLFD     414
cdep|L203_04836        LSFLPARWRPHLQTIITTPTSSRIFYFLLLNLGYMGIQMAYGVFTNSLGLISDSIHMLFD     414
cgac|I314_06191        LPLVPAGWRPHLQTIISTPTSSRIFYFLLLNLAYMGVQMVYGVFTNSLGLISDAIHMLFD     418
ccfg|D1P53_002977      LSLVPAGWRPHLQTIISTPTSSRIFYFLLLNLAYMGVQMVYGVLTNSLGLISDAIHMLFD     418
                       * ::** *********:***************.***:**.***:*********:******

cdcb|L204_05931        CLGLGVGLWASVATTWKPDGRYTFGYSRVETLSGFANGCFLILISVFIIFEGIQRVFDPP     474
cdep|L203_04836        CLGLGVGLWASVATTWKPDGRYTFGYSRVETLSGFANGCFLILISVFIIFEGIQRVFDPP     474
cgac|I314_06191        CLGLAVGLWASVAAMWKPDGRYTFGYSRVETLSGFANGCFLILISVFIIFEAIQRVYNPP     478
ccfg|D1P53_002977      CLGLAVGLWASVAAMWKPDGRYTFGYSRVETLSGFANGCFLILISVFIIFEAIQRVYNPP     478
                       ****.********: ************************************.****::**

cdcb|L204_05931        EMKTHRLLLVSGIGLAINLWGMYATGGHHHHGHSHGH------SHSHTPAPTTRLVSVLK     528
cdep|L203_04836        EMKTHRLLLVSGIGLAINLWGMYATGGHHHHGHSHGHGHEHGHSHSHTPAPTTRLV----     530
cgac|I314_06191        EMETHQLLLVSGIGLAINLWGMWATGGHHHHGHSHGHDHRHIHAAPKREMPKQGVHK---     535
ccfg|D1P53_002977      EMETHQLLLVSGIGLAINLWGMWATGGHHHHGHSHEHDHGHIHAAPKMEMPKQGAHK---     535
                       **:**:****************:************ *      :  :   *.

cdcb|L204_05931        DLFLAADYGRRLDMSTTRCMILLLYVTCQDSPQSILQEHKLPARLDNSSIAIPVVARKHI     588
cdep|L203_04836        -----------------------------NSPQSILQEHKLPARLDNSSIVIPVVARKHI     561
cgac|I314_06191        ------DDGAHKHEDHDRHHKSSASSQVSPRPASKLQKRKSTGRLKDSG-PRPITPQKTS     588
ccfg|D1P53_002977      ------DDGAHKHEDHDHHHKSSASSQVSPRPASKLQKRKSTGHLKDSG-PRPITPQKTS     588
                                                      * * **::*  .:*.:*.   *:. :*

cdcb|L204_05931        DPHKPKHEH--GHRETSDDHKHPDQCHKHDSDHSSHI-----SAHDHDHRYHDHECDSRT     641
cdep|L203_04836        DPHKSKHEH--GHRETSDNHKHPDQCHKHDSDHSSHI-----SAHDHDHRYHDHECDSRT     614
cgac|I314_06191        NGHSHAHEHEHNHDEH-CSHDHEDHSHSHD--HRHHKSTHNLATHNHVHAHEDDYDHAHA     645
ccfg|D1P53_002977      NGHSHAHEHEHNHDEH-CSHDHKDHAHSHDHHHHHHKSTHNLATHNHVHAHEDEYDHAHA     647
                       : *.  ***  .* *   .*.* *:.*.**  *  *      ::*:* * :.*.  .:::
```