

## Motifs, Domains and Colocation

### Learning objectives:

- Identify genes containing specific InterPro domains.
- Explore gene page sections related to orthology, domains and alpha-fold 3D structure prediction.
- Identify genes with specific protein motifs using regular expressions.
- Identify DNA motifs using regular expressions.
- Use the colocation tool to find genes by relative location to motifs.

### 1. Use InterPro domain searches to identify unannotated kinesin motor proteins.

**Note:** For this exercise use <http://giardiadb.org>

- a. Identify all genes annotated as hypothetical in all *Giardia* assemblages (genomes). Use the full text search and look for genes with the word “hypothetical” in their product names. There are two ways to do this:

i. Option 1:

1. Type the word hypothetical in the site search at the top of the page.
2. Filter your results first on genes,
3. Filter on product description.
4. Export your search results to a strategy by clicking on the blue button in the upper right.

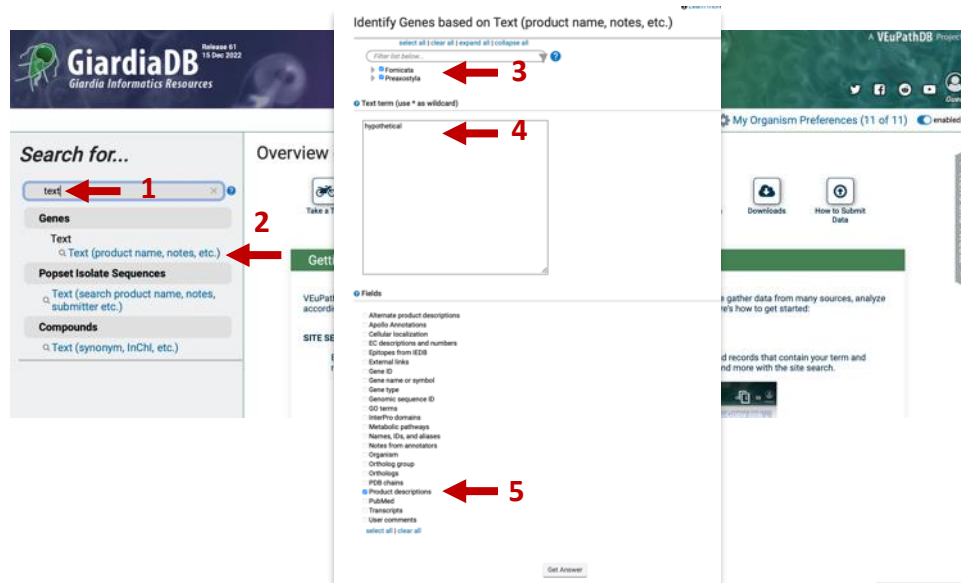
The screenshot shows the GiardiaDB website interface. At the top, there is a search bar with the word "hypothetical" entered. Below the search bar, there is a navigation menu with links: My Strategies, Searches, Tools, My Workspace, Data, About, Help, Contact Us. On the right side of the navigation menu, there are social media icons and a "Guest" user profile. Below the navigation menu, there is a section titled "Genes matching hypothetical". On the left side of this section, there is a "Filter results" panel. In this panel, the "Genes" filter is selected, and the "Product descriptions" filter is also selected. On the right side of the "Filter results" panel, there is a table showing the number of genes for each filter. The table has two columns: the filter name and the number of genes. The filters and their corresponding gene counts are: Cellular localization (236), InterPro domains (43), Orthologs (28,834), PDB chains (364), Preferred product description (15,096), Product descriptions (16,877), and User comments (32). The "Product descriptions" filter is highlighted with a red arrow labeled "3". In the top right corner of the "Genes matching hypothetical" section, there is a blue button labeled "Export as a Search Strategy to download or mine your results". This button is highlighted with a red arrow labeled "4".

Filter	Count
Cellular localization	236
InterPro domains	43
Orthologs	28,834
PDB chains	364
Preferred product description	15,096
Product descriptions	16,877
User comments	32

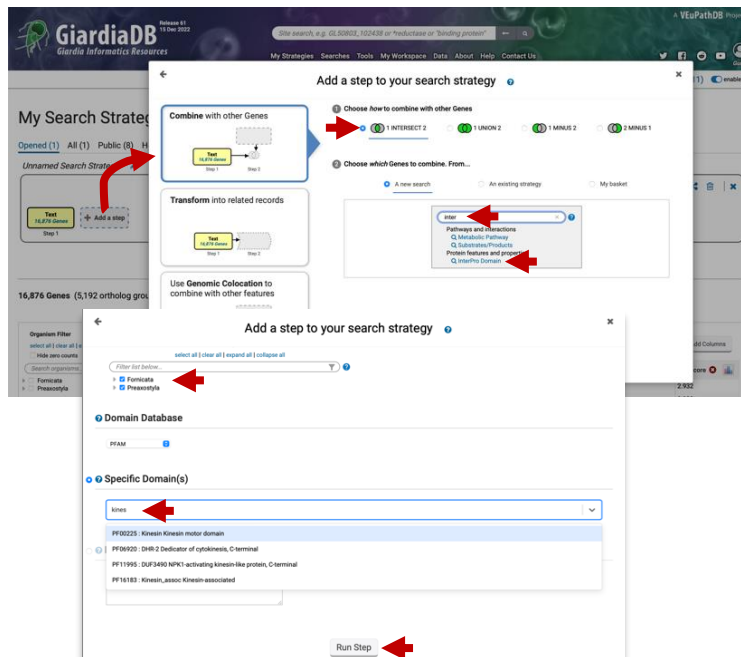
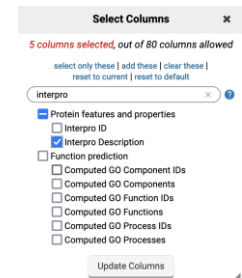
ii. Option 2:

1. Find the text search in the left menu.
2. Go to the Text search page
3. Select all organisms.
4. Type the word hypothetical in the search box.

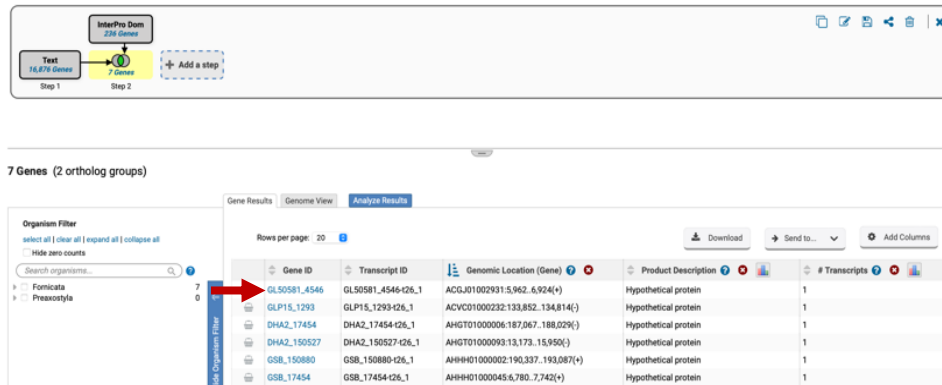
5. Select only product description in the fields section.
6. Click on get answer.



- b. Do any of these genes have InterPro domains? How can you find out?
  1. One way is to add a column for InterPro descriptions – Click on add column and select InterPro descriptions.
  2. Sort the column so you can see the InterPro descriptions at the top.
- c. How many of these hypothetical genes have a kinesin-motor protein PFAM domain?
  1. Add a step to the strategy. Go to the “Interpro Domain” search under ‘Protein features and properties’.
  2. Use the Specific Domains parameter to run a search for ‘kinesin-motor’ domain (PF00225). Start typing the word kinesin and it should autocomplete.

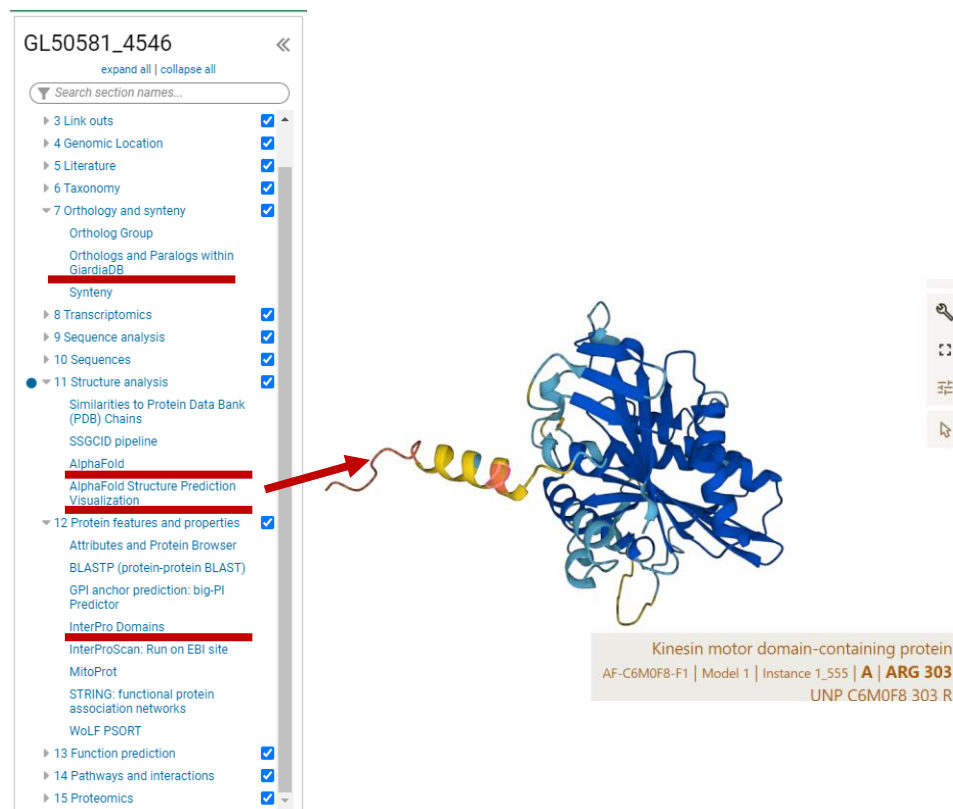


- Go to the gene page for GL50581\_4546 by clicking on ID link in the result table. Explore the gene page.



Gene ID	Transcript ID	Genomic Location (Gene)	Product Description	# Transcripts
GL50581_4546	GL50581_4546-26_1	ACGJ01002931:5,962..6,924(+)	Hypothetical protein	1
GLP15_1293	GLP15_1293-26_1	ACVC01000232:133,852..134,814(+)	Hypothetical protein	1
DHA2_17454	DHA2_17454-26_1	AHG01000006:187,067..188,029(+)	Hypothetical protein	1
DHA2_150527	DHA2_150527-26_1	AHG01000093:13,173..15,950(+)	Hypothetical protein	1
GSB_150880	GSB_150880-26_1	AHH01000002:190,337..193,087(+)	Hypothetical protein	1
GSB_17454	GSB_17454-26_1	AHH01000045:6,780..7,742(+)	Hypothetical protein	1

- What might you conclude about the possible function of this protein? In particular examine
  - the table called “Orthologs and Paralogs within GiardiaDB”
  - the “Interpro Domains” table in the protein features and properties section and
  - the AlphaFold prediction table and graphic. Click the links in the table and hover over the secondary structures in the image.



GL50581\_4546

expand all | collapse all

Search section names...

- 3 Link outs
- 4 Genomic Location
- 5 Literature
- 6 Taxonomy
- 7 Orthology and synteny
  - Ortholog Group
  - Orthologs and Paralogs within GiardiaDB
  - Synteny
- 8 Transcriptomics
- 9 Sequence analysis
- 10 Sequences
- 11 Structure analysis
  - Similarities to Protein Data Bank (PDB) Chains
  - SSGCI pipeline
  - AlphaFold
  - AlphaFold Structure Prediction Visualization
- 12 Protein features and properties
  - Attributes and Protein Browser
  - BLASTP (protein-protein BLAST)
  - GPI anchor prediction: big-PI Predictor
  - InterPro Domains
  - InterProScan: Run on EBI site
  - MitoProt
  - STRING: functional protein association networks
  - WoLF PSORT
- 13 Function prediction
- 14 Pathways and interactions
- 15 Proteomics

Kinesin motor domain-containing protein  
AF-C6M0F8-F1 | Model 1 | Instance 1\_555 | A | ARG 303  
UNP C6M0F8 303 R

2. Use regular expressions to find motifs in TriTrypDB: finding active trans-sialidases in *T. cruzi*.  
Note: for this exercise use <http://tritrypdb.org>

1. *T. cruzi* has an expanded family of trans-sialidases. In fact, if you run a text search for any gene with the word “trans-sialidase” in the product description of all *T. cruzi* strains, you return over 13,000 genes among the strains in the database!!! Try this and see what you get.
2. Not all of the genes returned in (a) are predicted to be active. It is known that active trans-sialidases have a signature tyrosine (Y) at position 342 in their amino acid sequence. Add a Protein Motif Pattern search step to the text search in ‘a’ to identify only the active trans-sialidases.

Unnamed Search Strategy \*

Text 13,660 Genes Step 1

+ Add a step

13,660 Genes (1,235 ortholog groups)

Organism Filter  
select all | clear all | expand all | collapse all  
Hide zero counts  
Search organisms...  
Eubodonida  
Trypanosomatida

Add a step to your search strategy

1 Choose how to combine with other Genes  
☒ 1 INTERSECT 2 ☐ 1 UNION 2 ☐ 1 MINUS 2

2 Choose which Genes to combine. From...  
☒ A new search ☐ An existing strategy

motif  
Sequence analysis  
BLAST  
Protein Motif Pattern

- d. Write a regular expression that defines a protein sequence that starts with a methionine, and is followed by 340 of any amino acids, followed by a tyrosine ‘Y’. Refer to the Learn More tab on the search page for the [regular expression tutorial](#) linked there if you need to.
- e. <https://tritrypdb.org/tritrypdb/app/workspace/strategies/import/790a20c5832fd4f6>

Identify Genes based on Protein Motif Pattern

Configure Search Learn More View Data Sets Used

Pattern  
\*M.{340}Y

Organism  
15 selected, out of 68  
select only these | add these | clear these  
cruzi  
Trypanosomatida  
Trypanosoma  
Trypanosoma cruzi  
Trypanosoma cruzi Berenice  
Trypanosoma cruzi Brazil A4  
Trypanosoma cruzi CL Brener Esmeraldo-like [Reference]  
Trypanosoma cruzi CL Brener Non-Esmeraldo-like  
Trypanosoma cruzi Dm28c 2014  
Trypanosoma cruzi Dm28c 2017  
Trypanosoma cruzi Dm28c 2018  
Trypanosoma cruzi Sylvio X10/1  
Trypanosoma cruzi Sylvio X10/1-2012  
Trypanosoma cruzi TCC  
Trypanosoma cruzi Y C6  
Trypanosoma cruzi marinkellei strain B7  
Trypanosoma cruzi strain CL  
Trypanosoma cruzi strain CL Brener  
Trypanosoma cruzi strain G

'active trans-sialidases: with motif = \*M.{340}Y'

Prot Motif 2,216 Genes  
Text 13,660 Genes 198 Genes  
Step 1 Step 2  
+ Add a step

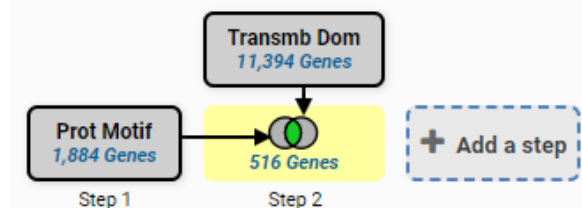
3. Find *Cryptosporidium* genes with the YXXΦ receptor signal motif. Note: for this exercise use <http://cryptodb.org>

The YXXΦ (Y=tyrosine, X=any amino acid, Φ=bulky hydrophobic [phenylalanine, tyrosine, threonine]) motif is conserved in many eukaryotic membrane proteins that are recognized by adaptor proteins for sorting in the endosomal/lysosomal pathway. This motif is typically located in the c-terminal end of the protein. **\*\*\*Note: do not look for the Φ symbol on your keyboard – this will not work. Rather you should use the amino acid symbols.**

- a. Use the “Protein Motif Pattern” search to find all *Cryptosporidium* proteins that contain this motif anywhere in the **terminal 10 amino acids of proteins**. (hint: for your regular expression, remember that you want the first amino acid to be a tyrosine, followed any two amino acids, followed by any bulky hydrophobic amino acid (phenylalanine, tyrosine, threonine). Refer to the Learn More tab of the Protein Motif Pattern search or the [regular expression tutorial](#) linked there if you need to.

The screenshot shows the CryptoDB website interface. On the left, under 'Search for...', the 'Protein Motif Pattern' option is selected. On the right, the 'Identify Genes based on Protein Motif Pattern' panel is active, showing a list of 14 selected genes out of 19 total. The genes listed include various *Cryptosporidium* species and strains, such as *Cryptosporidium andersoni*, *Cryptosporidium bovis*, *Cryptosporidium hominis*, *Cryptosporidium parvum*, and *Cryptosporidium tyzzeri*.

- b. How many of these proteins also contain at least one transmembrane domain. [Strategy](#)



- c. What would happen if you revise the first step (the motif pattern step) to include genes with the sorting motif in the C-terminal 20 amino acids? (hint: edit the first step and modify your regular expression). Here is the search strategy for this but don't click on it until you have tried this yourself: [Strategy](#)

4. Find *Plasmodium* genes downstream of a AP2 binding motif. For this exercise use: <https://PlasmoDB.org>

The *Plasmodium* Ap2-EXP (Pf3D7\_1466400) is predicted to bind to the DNA motif TGCATGT/C (T/C means either a T or C). Use this motif to find all *Plasmodium falciparum* 3D7 genes located within 1000 nucleotides of this motif.

- a. Find the TGCATGT/C DNA motif in the *P. falciparum* 3D7 genome.
1. Select the "Search for genomic segments (DNA motif)" menu from the Search menu and look for TGCATGT/C in *P. falciparum* 3D7 .

The screenshot displays the PlasmoDB search interface. On the left, a 'Search for...' sidebar shows a 'motif' search bar and a 'Genomic Segments' section with a 'DNA Motif Pattern' link. A red arrow points from this link to the 'Organism' section on the right. The 'Organism' section shows a search bar with '3d7' entered, and a list of results including 'Plasmodium falciparum 3D7 [Reference]', which is selected. Another red arrow points from this selection to the 'Pattern' section. The 'Pattern' section has a search bar containing 'TGCATG[TC]', with a third red arrow pointing to it.

### Identify Genomic Segments based on DNA Motif Pattern

Configure Search   Learn More   View Data Sets Used

**Organism**

1 selected

3d7

Plasmodiidae  
Plasmodium  
Plasmodium falciparum  
● Plasmodium falciparum 3D7 [Reference]

**Pattern**

TGCATG[TC]

2. How did you write the pattern? Note that you cannot use T/C to indicate either a T or a C. See the description in the Learn More tab for additional help and hints.
3. Once you have identified all the TGCATGT/C motifs in the *P. falciparum* 3D7 genome, can you find all the genes that are within 1000 nucleotides of the motif (on the 5' end)?
  - a. VEuPathDB offers a colocation function to identify genomic features within a specified distance of each other. Add a step to your motif search to identify all genes from *P. falciparum* 3D7. You will notice that the only way to do this is to select the colocation option in the add step popup: Click "Add Step". Choose the genomic colocation option then select the organism search under taxonomy. You will run a search for all genes in *P. falciparum* 3D7.

Unnamed Search Strategy

**DNA Motif**  
732 Segments  
Step 1

**732 Genomic Segments**

Genomic Segment Results

Rows per page: 1000

Segment ID

PF3D7\_01\_v3-160539

**Add a step to your search strategy**

Use the relative position of features on the genome between your existing step and the new step to filter features to keep in the final result.

Choose which features to collocate. From...

☒ A new search ☐ An existing strategy ☐ My basket

Search: taxon

Genes  
Taxonomy  
Organism

**Combine with other Genomic Segments**

Step 1: TGCATGT/C 3,494 Segments  
Step 2: [Diagram showing overlap]

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

Step 1: TGCATGT/C 3,494 Segments  
Step 2: [Diagram showing overlap]

- b. Set up the collocation using the following guidelines: *Return each **gene** from the new step whose upstream region (1000bp) overlaps the exact region of a Genomic Segment in Step1 (TGCATGT/C) and is on either strand.*

**Add a step to your search strategy**

"Return each **Gene from the new step** whose **upstream region** overlaps the **exact region** of a Genomic Segment from the current step and is on **either strand**"

**Region**  
Gene

☐ Exact  
☒ Upstream: 1000 bp  
☐ Downstream: 1000 bp  
☐ Custom:  
begin at: start - 1000 bp  
end at: start + 1 bp

**Region**  
Genomic Segment

☒ Exact  
☐ Upstream: 1000 bp  
☐ Downstream: 1000 bp  
☐ Custom:  
begin at: start + 0 bp  
end at: stop + 0 bp

Run Step

4. Explore your results. What kinds of genes did you identify? Try doing a GO enrichment on your results to see if there is an enrichment of certain types of functions.

