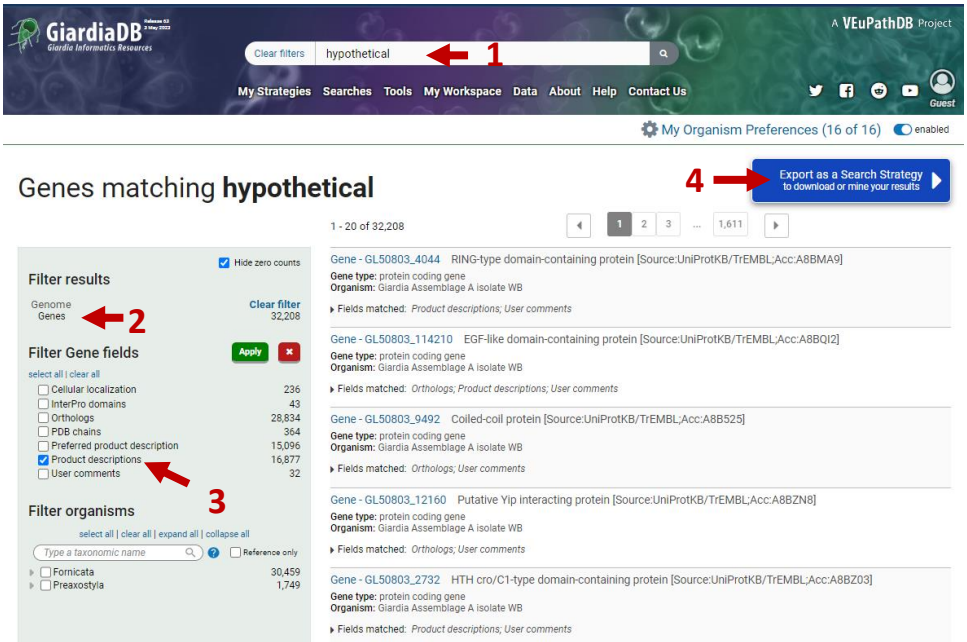
Motifs, Domains and Colocation

Learning objectives:

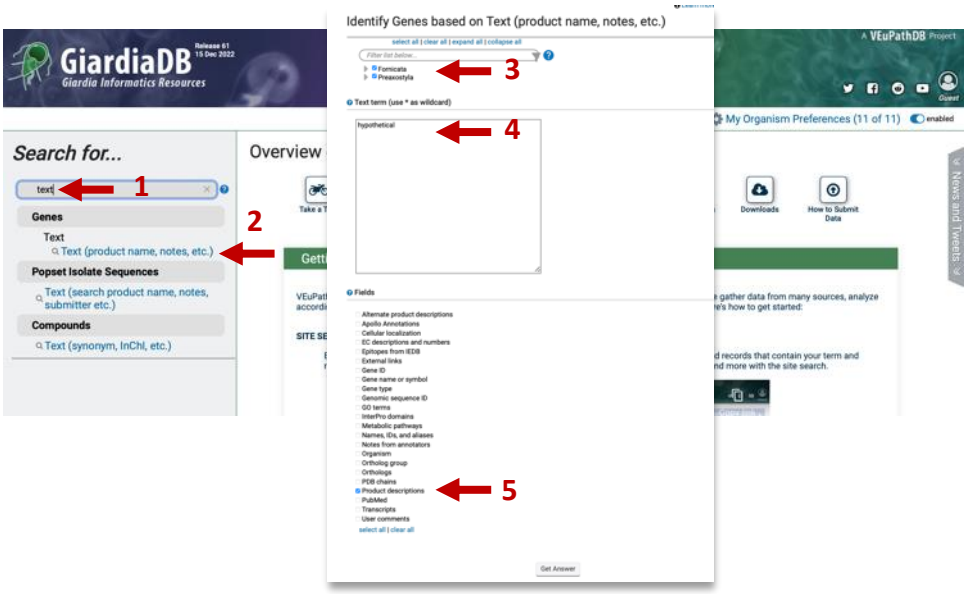
- Identify genes containing specific InterPro domains.
- Explore gene page sections related to orthology, domains and alpha-fold 3D structure prediction.
- Identify genes with specific protein motifs using regular expressions.
- Identify DNA motifs using regular expressions.
- Use the colocation tool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

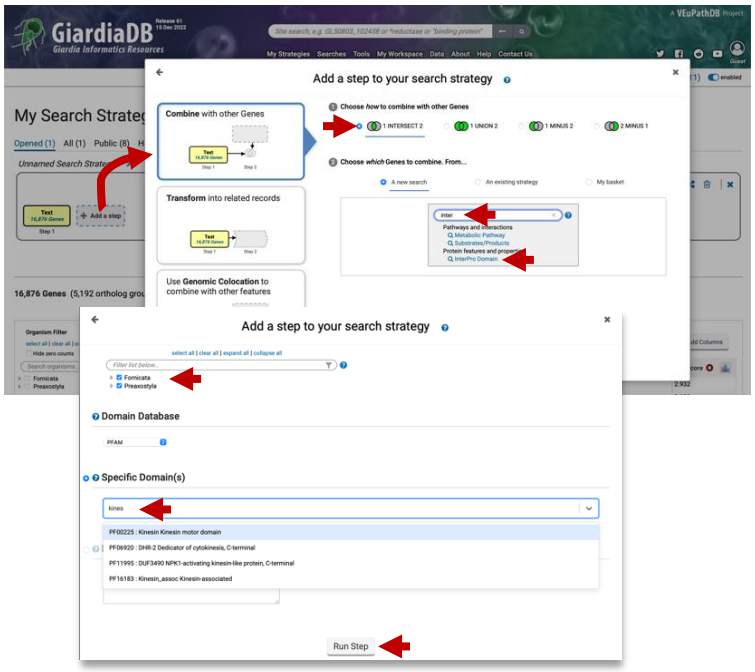


      ii. Option 2:
         1. Find the text search in the left menu.
         2. Go to the Text search page
         3. Select all organisms.

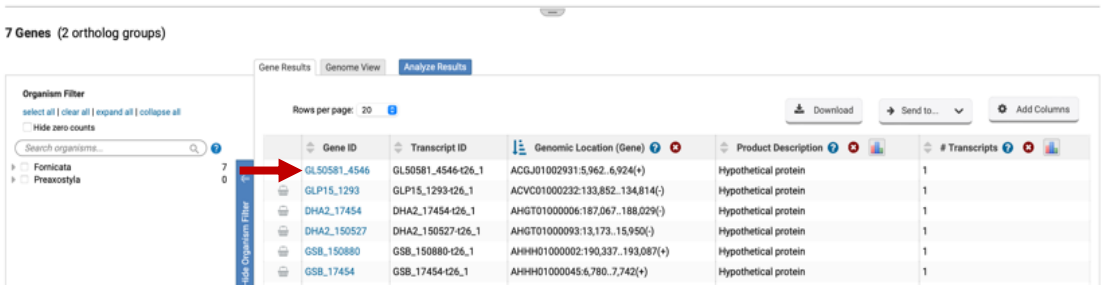
4. Type the word hypothetical in the search box.
5. Select only product description in the fields section.
6. Click on get answer.



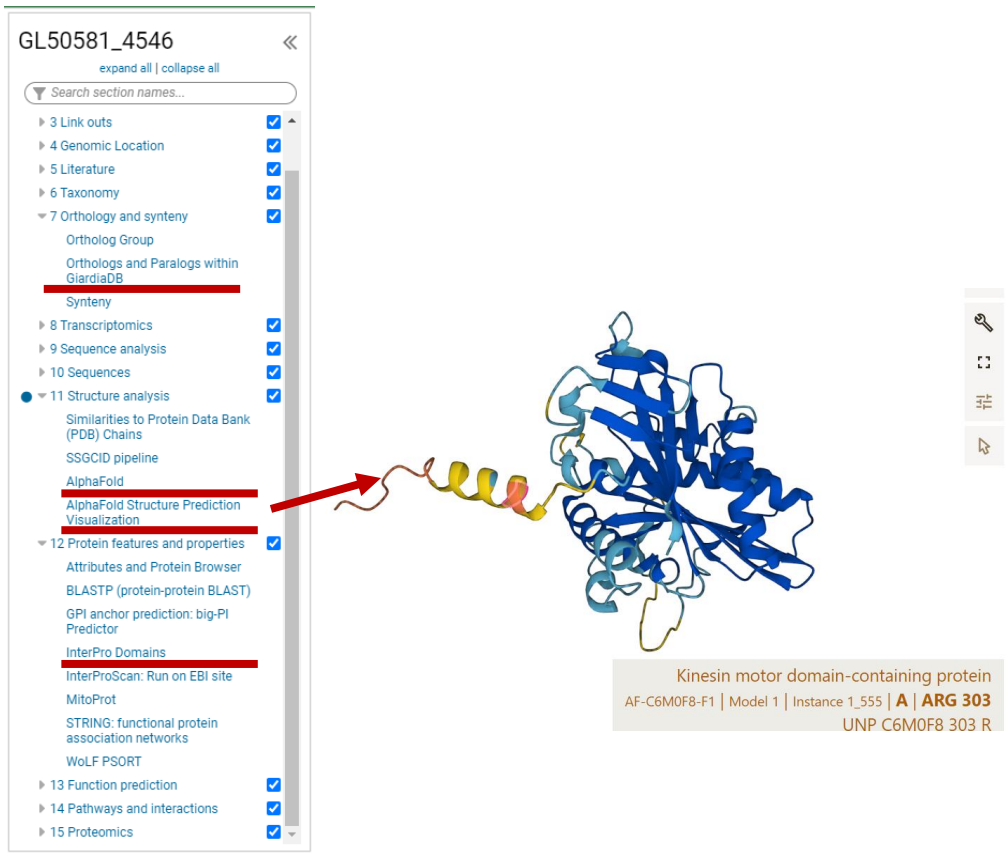
b. How many of these hypothetical genes have a kinesin-motor protein PFAM domain?
   1. Add a step to the strategy. Go to the "Interpro Domain" search under 'Protein features and properties'.
   2. Use the Specific Domains parameter to run a search for 'kinestin-motor' domain (PF00225). Start typing the work kinesin and it should autocomplete.



3. Go to the gene page for GL50581_4546 by clicking on ID link in the result table. Explore the gene page.

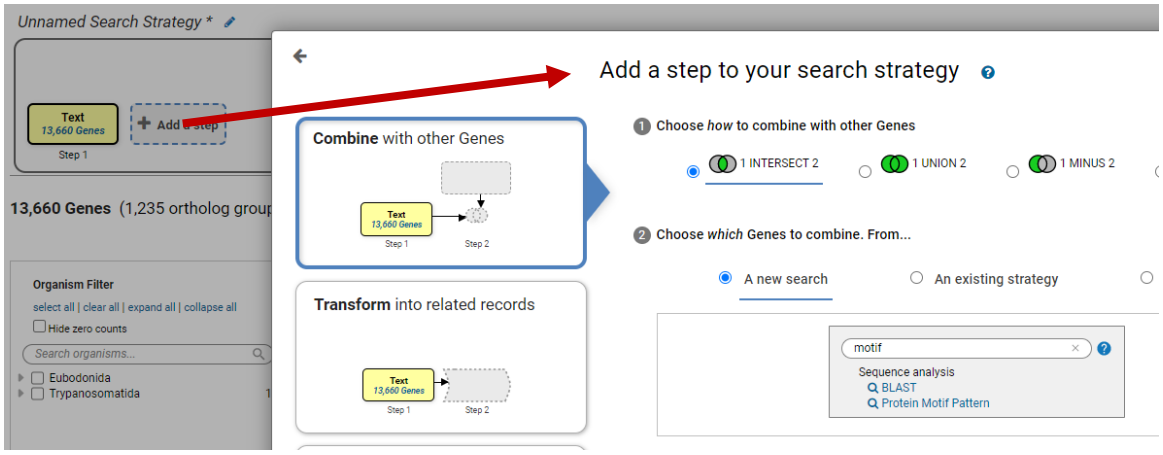


4. What might you conclude about the possible function of this protein? In particular examine
   i. the table called "Orthologs and Paralogs within GiardiaDB"
   ii. the "Interpro Domains" table in the protein features and properties section and
   iii. the AlphaFold prediction table and graphic. Click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

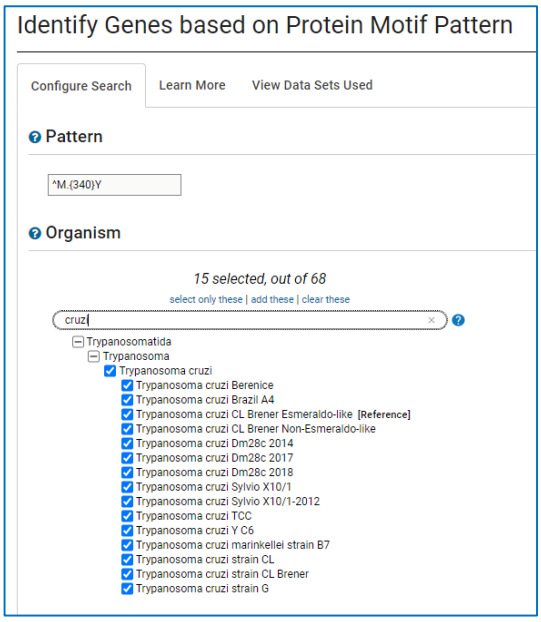


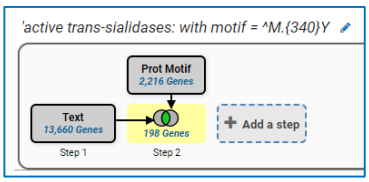
2. **Use regular expressions to find motifs in TriTrypDB: finding active trans-sialidases in *T. cruzi*. Note: for this exercise use http://tritrypdb.org**

   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.



   c. 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.
   d. https://tritrypdb.org/tritrypdb/app/workspace/strategies/import/790a20c5832fd4f6

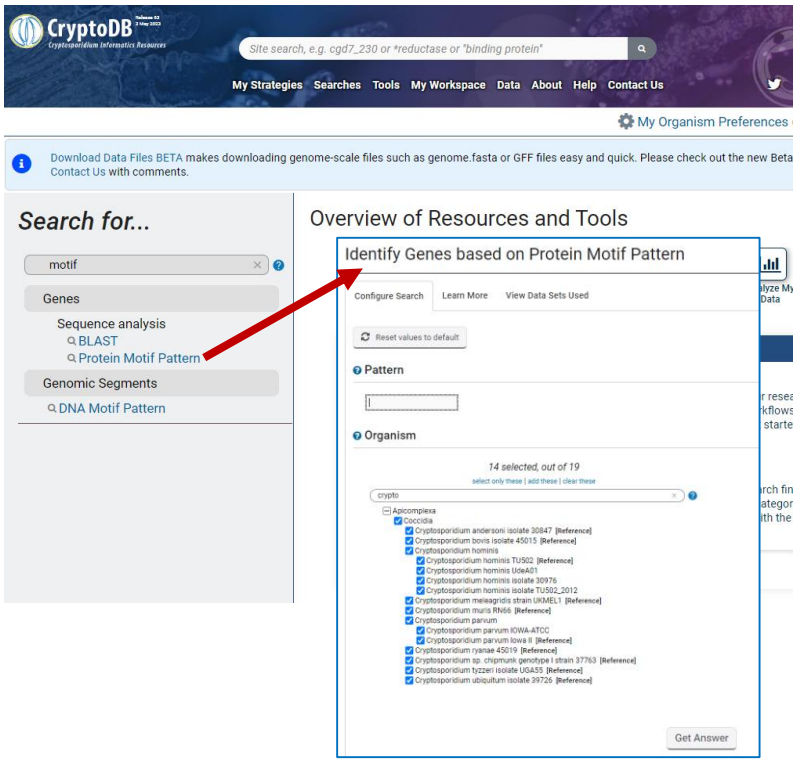




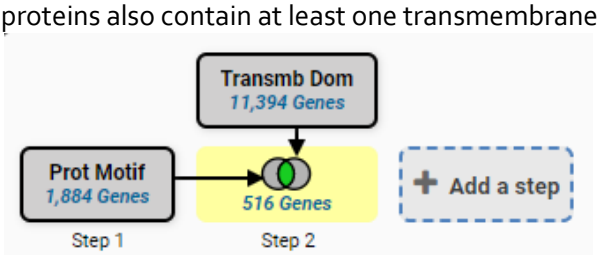
3. **Find Cryptosporidium genes with the YXXΦ receptor signal motif. Note: for this exercise use http://cryptodb.org**

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.



b. How many of these proteins also contain at least one transmembrane domain. Strategy

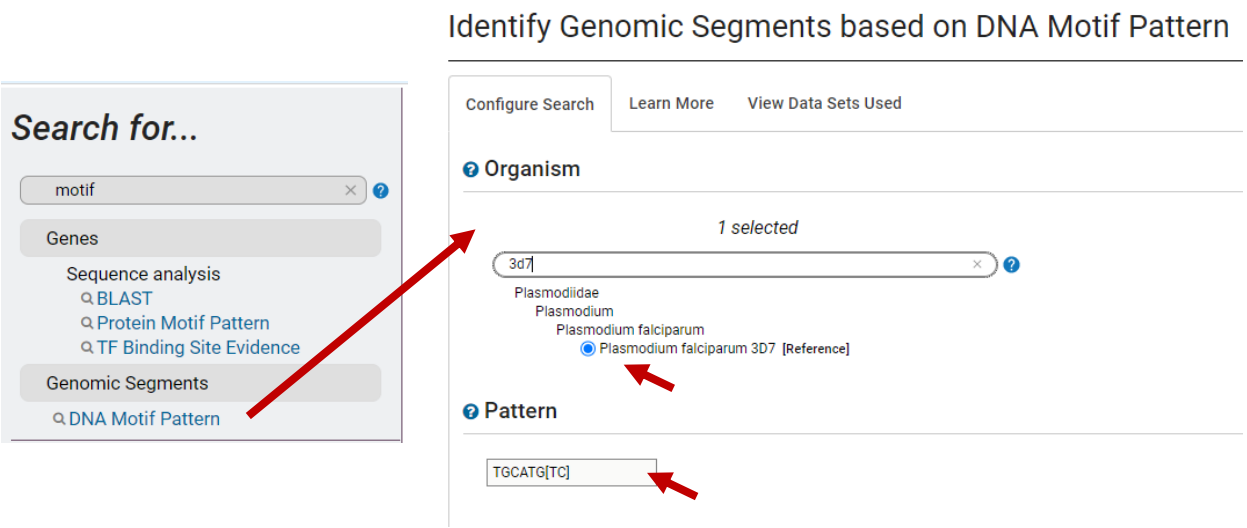


c. What would happen if you revise the first step (the motif pattern step) to include genes with the sorting motif in the C-terminal 20 amino acids? (hint: edit the first step and modify your regular expression). Here is the search strategy for this but don't click on it until you have tried this yourself: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

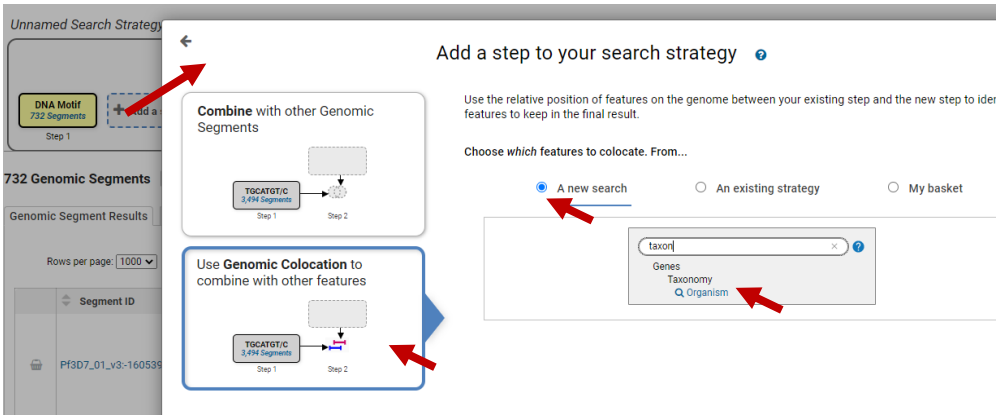
4. **Find Plasmodium genes downstream of a AP2 binding motif. For this exercise use: https://PlasmoDB.org**

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.

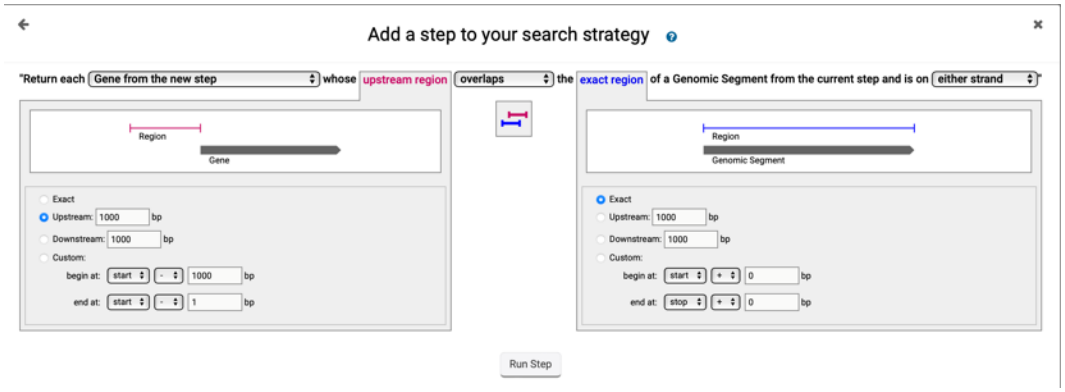
a. Find the TGCATGT/C DNA motif in the *P. falciparum* 3D7 genome.
   1. Select the "Search for genomic segments (DNA motif)" menu from the Search menu and look for TGCATGT/C in *P. falciparum* 3D7 .



   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 withing 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*.



b. Set up the colocation 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.*



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.