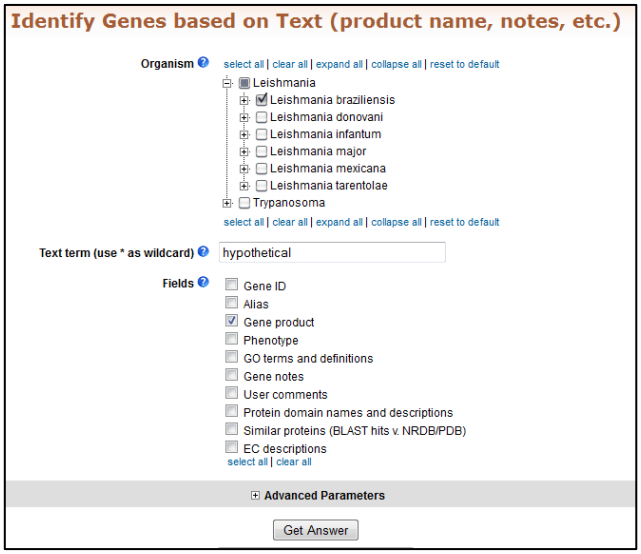
Motif Searches and Regular Expressions
Exercise 6

6.1 Using InterPro domain searches to identify unannotated kinesin motor proteins.

Note: For this exercise use http://tritrypdb.org

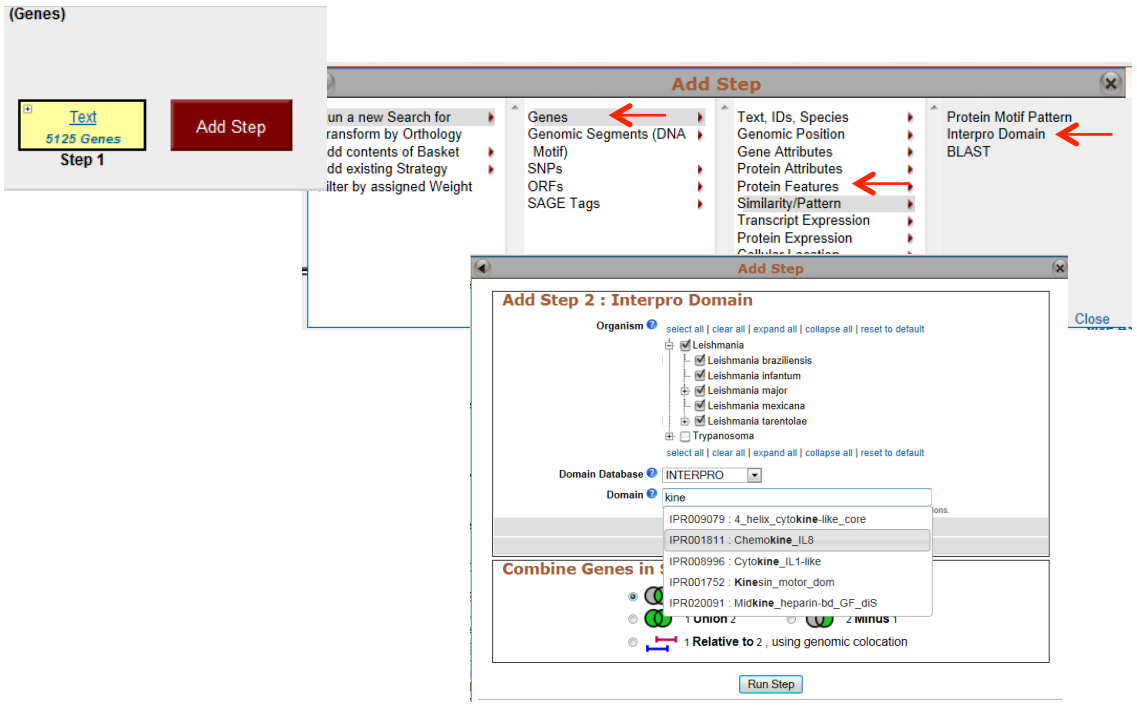
a. Identify all genes annotated as hypothetical in *L. braziliensis.*

Hint: use the full text search and look for genes with the word “hypothetical” in their product names

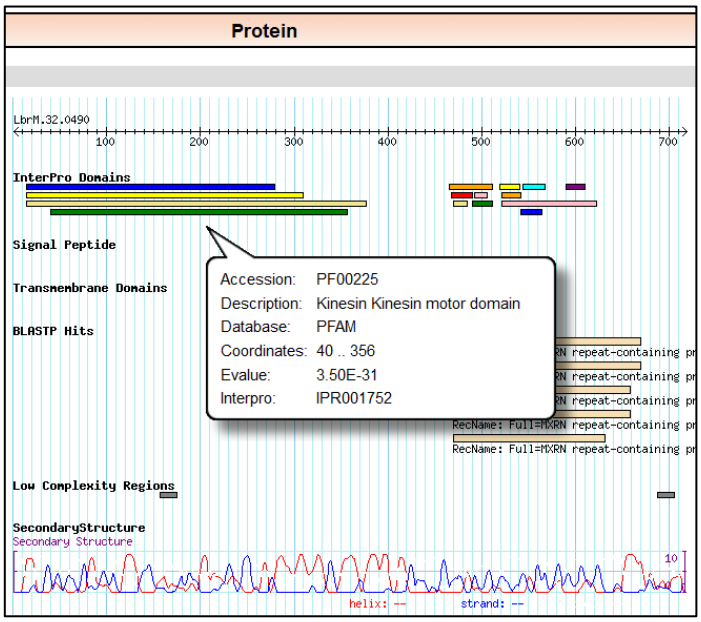


b. How many of these hypothetical genes have a kinesin-motor protein InterPro domain?

Hint: add a step to the strategy. Go to the “Interpro Domain” search under similarity/pattern, start typing the work kinesin and it should autocomplete.



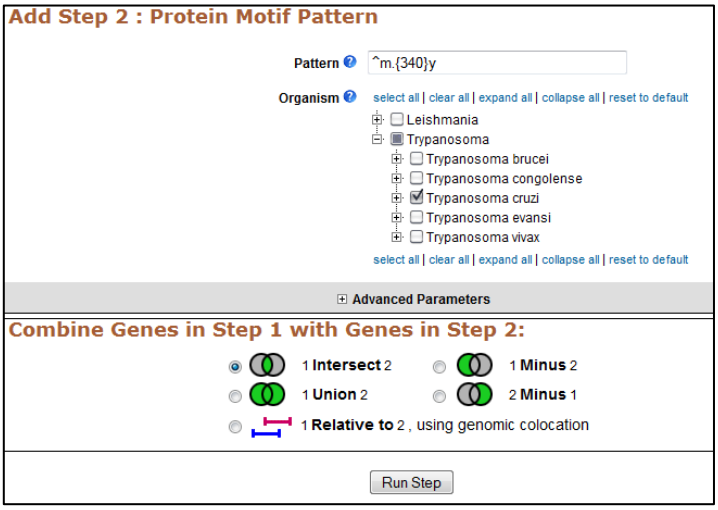
c. Go to the gene page for LbrM.32.0490 and look at the protein feature section. Does this look like a possible motor protein? Hint: click on the ID for LbrM.32.0490 in the result table to go to the gene page. Mouse over the glyphs in the Protein Features graphic.



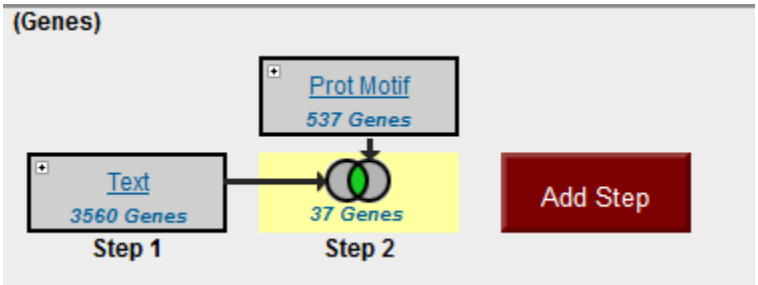
6.2 Using regular expressions to find motifs in TriTypDB: finding active trans-sialidases in *T. cruzi.*

a. *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”, you return over 3500 genes among the strains in the database!!! Try this and see what you get.

b. However, not all of these 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 motif search step to the text search in ‘a’ to identify only the active trans-sialidases.
   - Hint: for your regular expression, remember that you want the first amino acid to be a methionine, followed by 340 of any amino acid, followed by a tyrosine ‘Y’. Refer to regular expression tutorial if you need to.



If you need help, you can go to this sample strategy below to see the answer:
http://tritrypdb.org/tritrypdb/im.do?s=a905e36f634f7b42

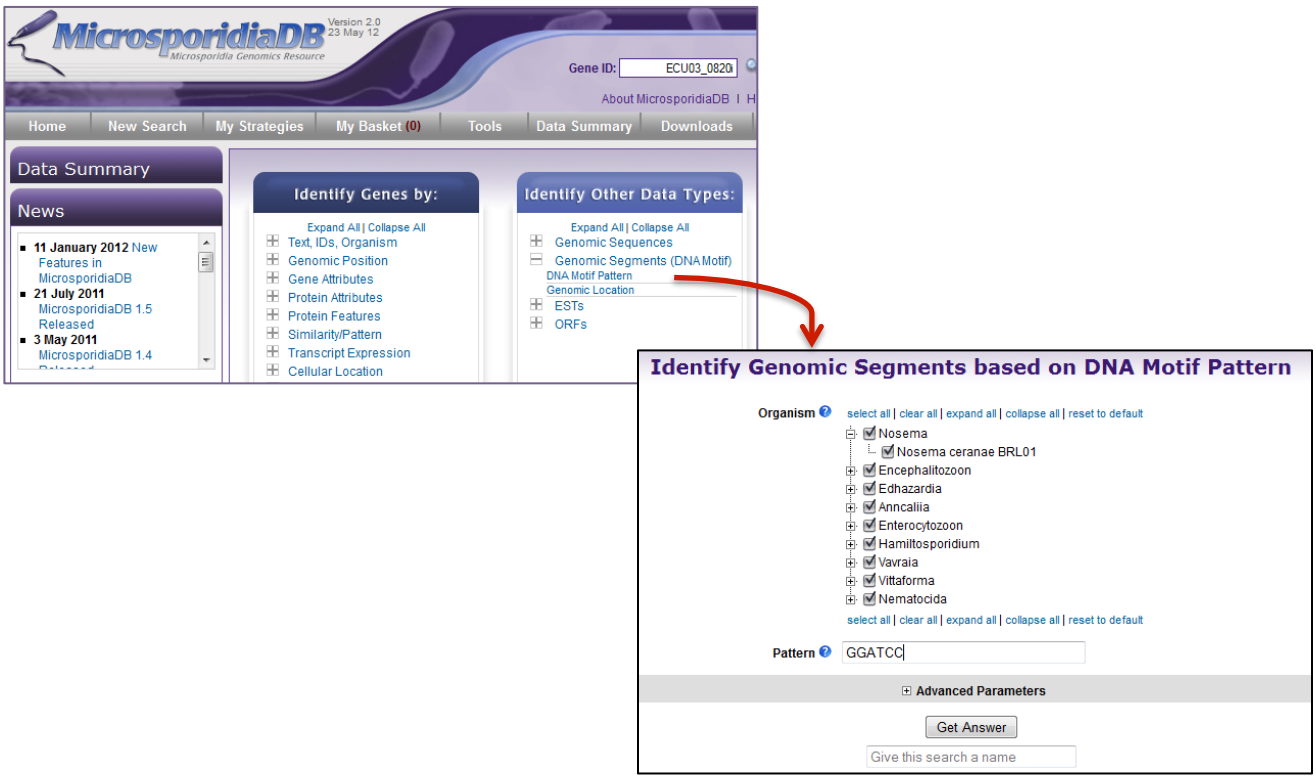


6.3 Identification of specific DNA motifs.

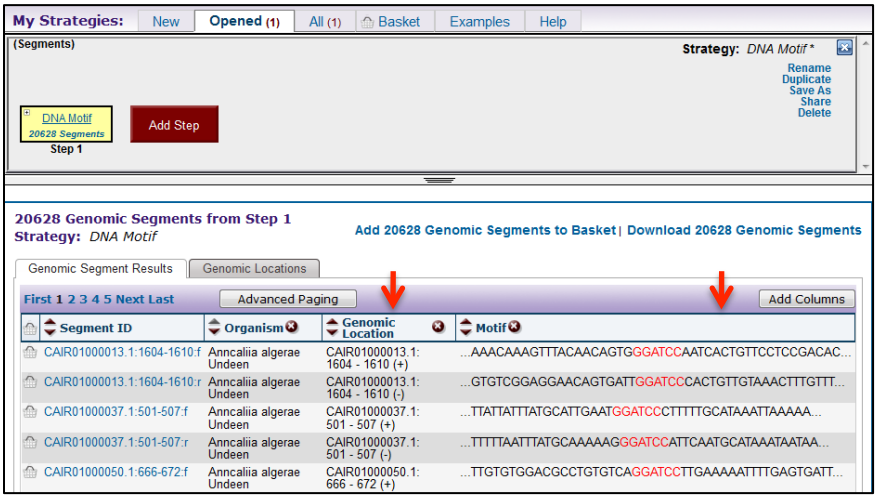
Note: For this exercise use http://microsporidiadb.org

a. Find all BamHI restriction sites in all microsporidia genomic sequences available in MicrosporidiaDB. Note: you can use the DNA motif search to find complex motifs like transcription factor binding sites using regular expressions.

Hint: BamHI = GGATCC and the DNA motif search is under the heading "Genomic Segments".



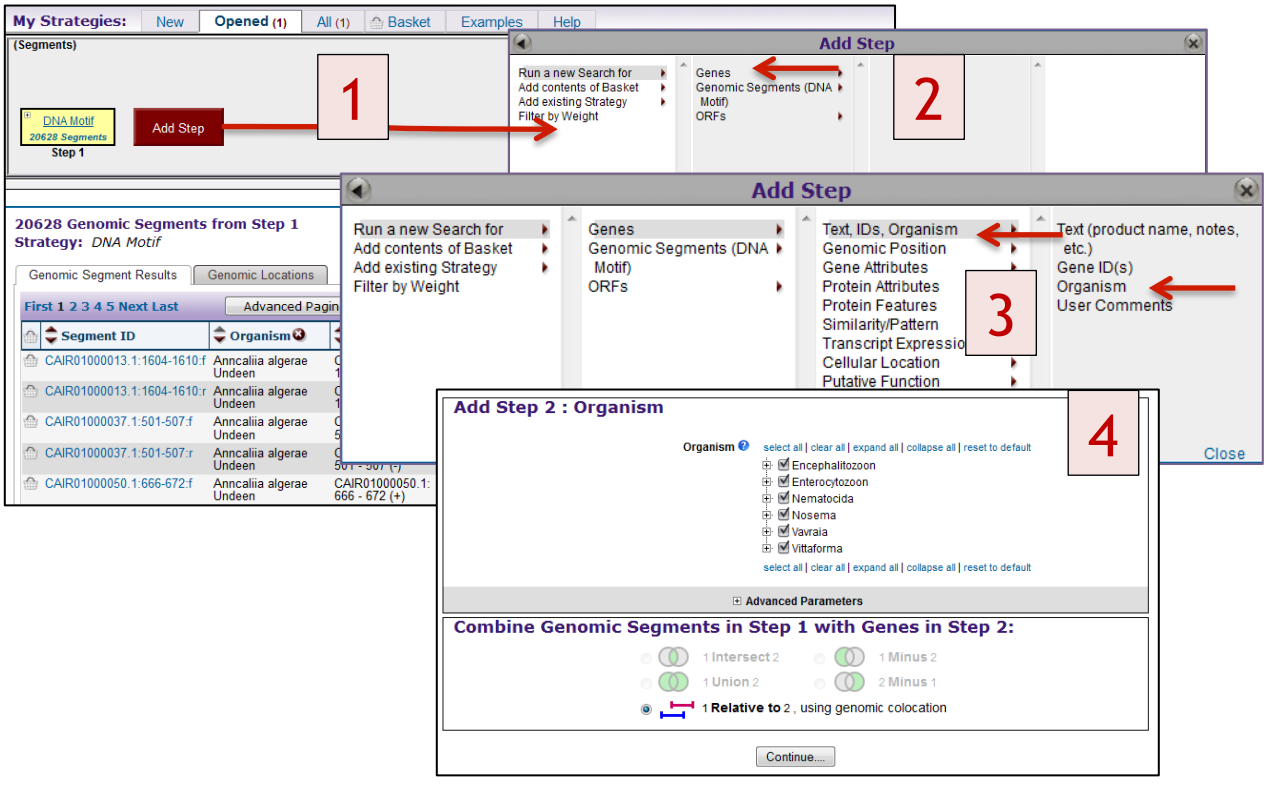
b. How many times does the BamHI site occur in the genomes you searched? Take a look at your results; notice the Genomic location and the Motif columns.

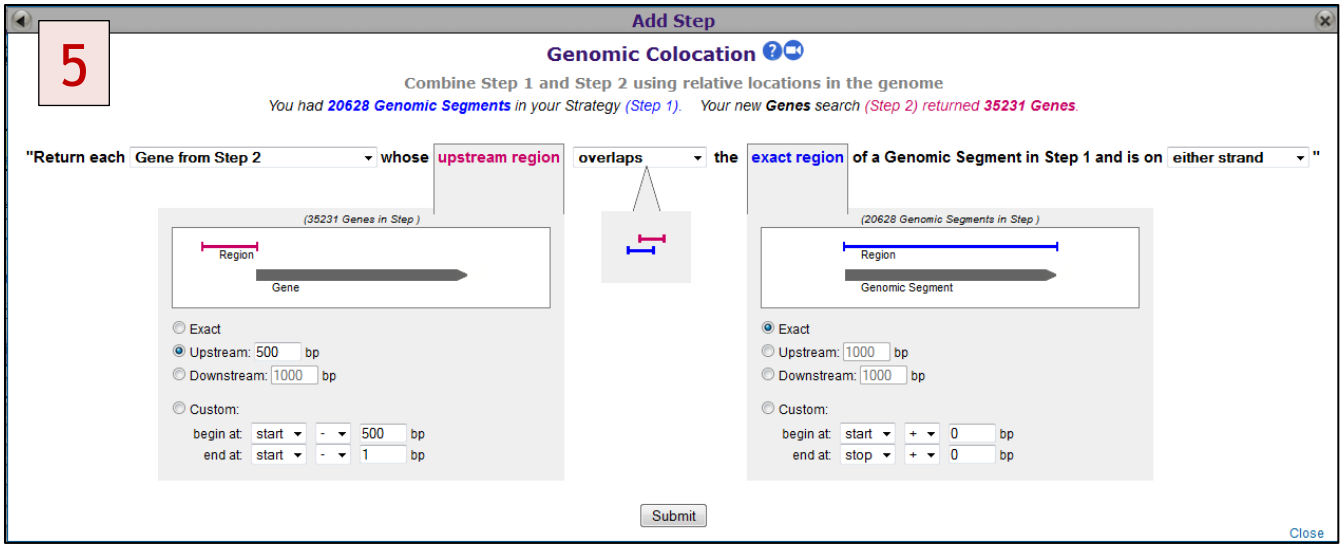


6.4 Find genes that have one of these BamHI sites within 500 nucleotides upstream of their start.

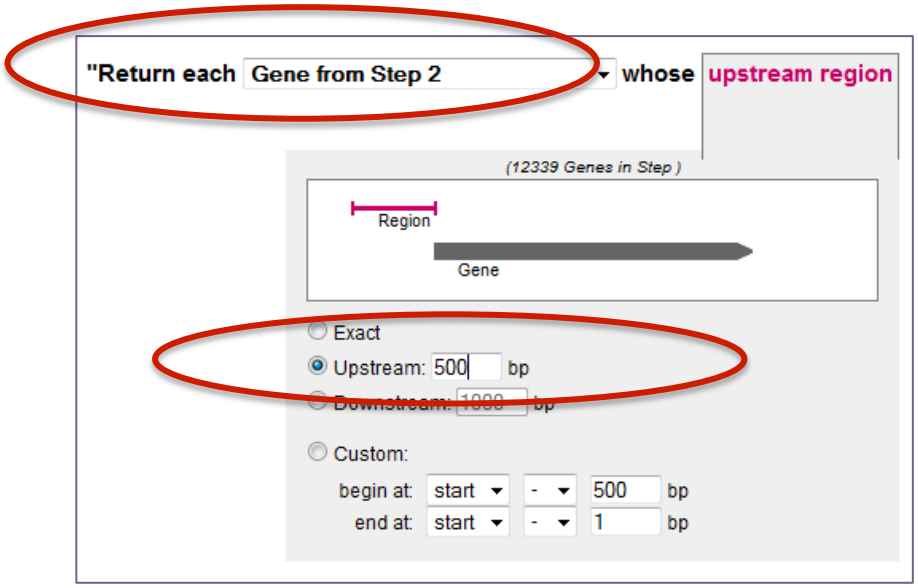
In the section 7.3 you found BamHI sites, but now you are looking for genes that have one of these sites located within 500 nucleotides upstream of their start.

Hint: You can achieve this by running a genomic collocation search that defines the genomic relationship between the BamHI sites and genes. Add a “Genes by Organism” step to the motif search and select the “1 relative to 2, using genomic locations” option.

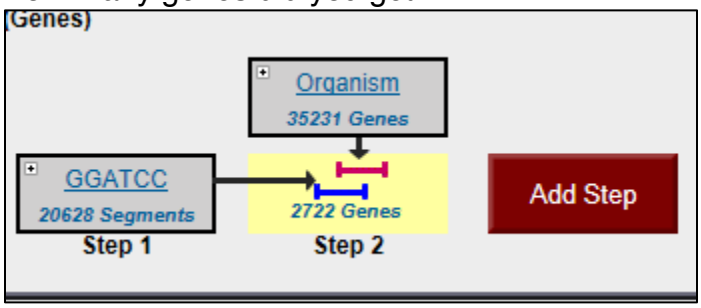




How did you modify the location relative to genes?

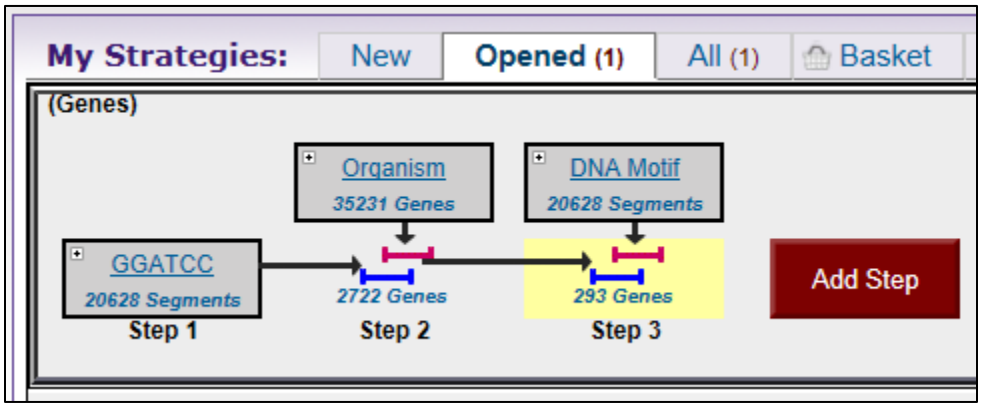


How many genes did you get?



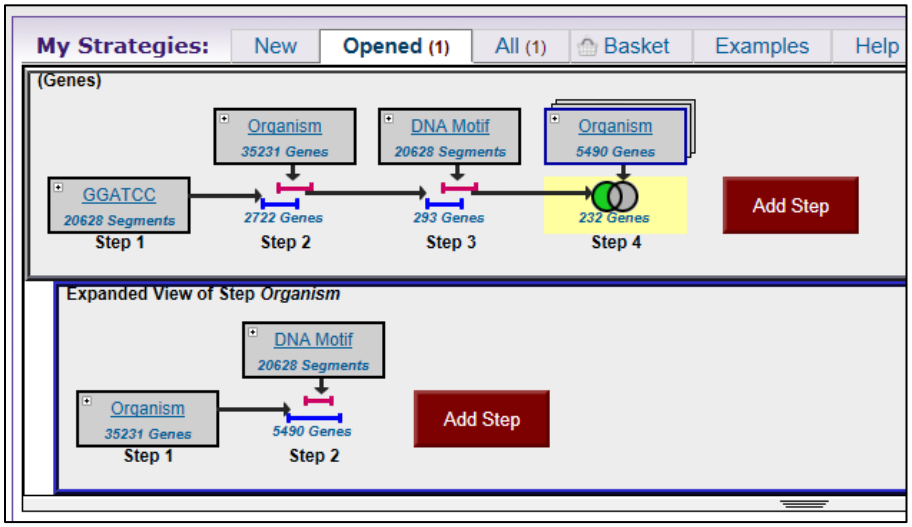
6.5 Using a similar sequence of steps as in part 7.4, define which of these genes also have a BamHI site in their 500 nucleotide downstream region.

Hint: after you click on add step you will have to select DNA motif search and select the genomic collocation option.



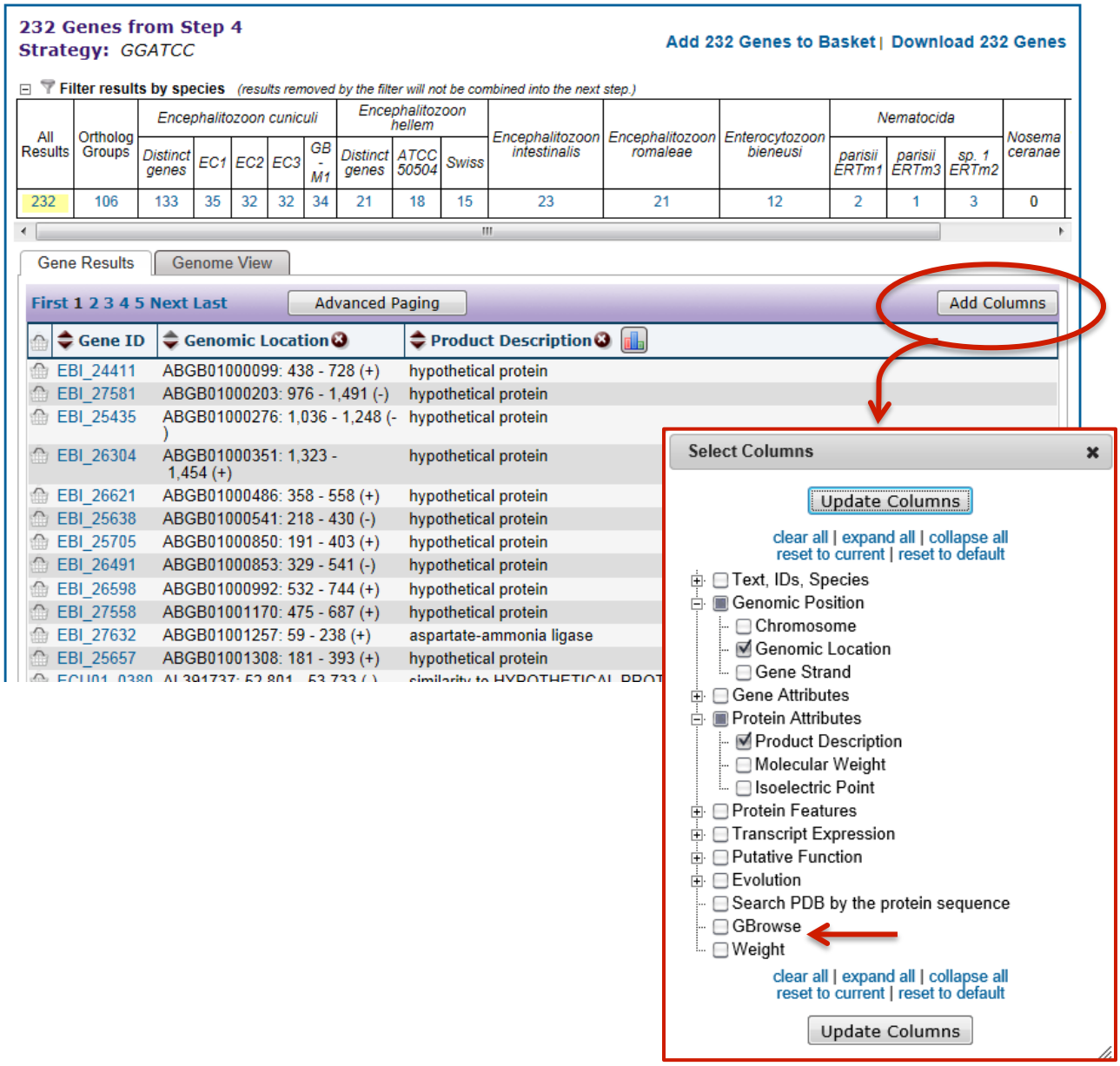
6.6 Taking this a step further, define which of these genes do **NOT** contain a BamHI site within them.

Hint: you will have to use a nested strategy.



Look at your results. Do they make sense? Confirm your results by looking at one of the genes in Gbrowse and showing BamHI restriction sites.

Note: you can add a column to any result table that allows you to go directly to GBrowse at the genomic coordinates of any ID in your result list. Click on the Add Columns button.



Note: you can configure restriction sites by clicking on the configure button in GBrowse and selecting the restriction sites you would like to display. To view restriction sites, the "Restriction Sites" data track must be turned on. Go to the "Select Tracks" page and click "Restriction Sites" under the "Analysis" section.