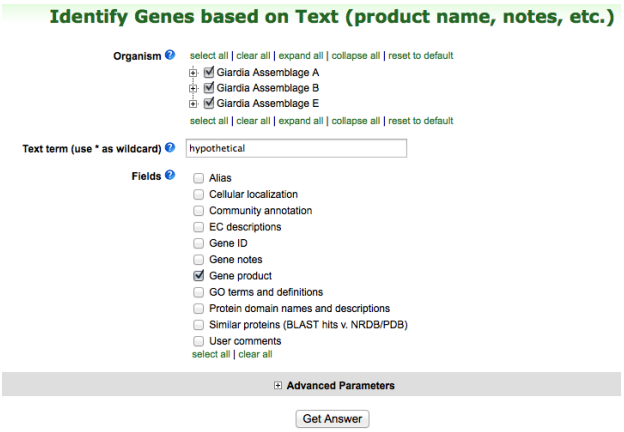
Motif Searches, Regular Expressions and Genomic Colocation

1. **Using InterPro domain searches to identify unannotated kinesin motor proteins.**

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

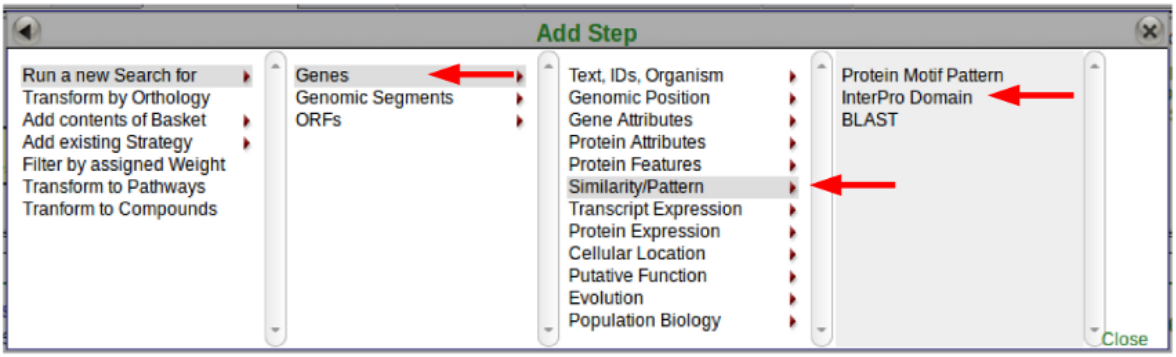
   a. Identify all genes annotated as hypothetical in all *Giardia* assemblages.

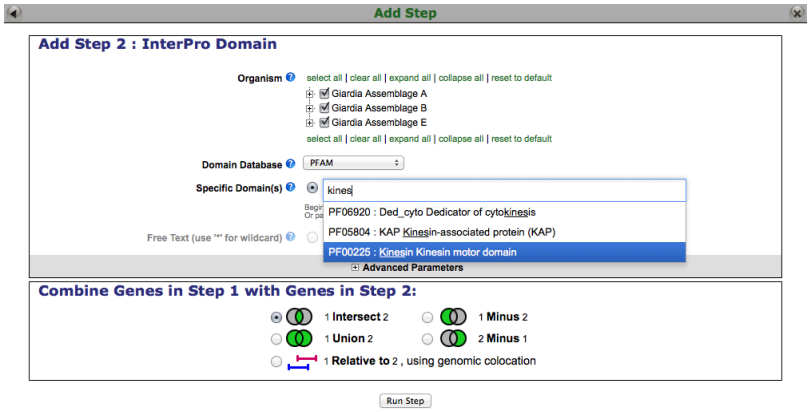
   (Hint: use the full text search and look for genes with the word "hypothetical" in their product names)



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

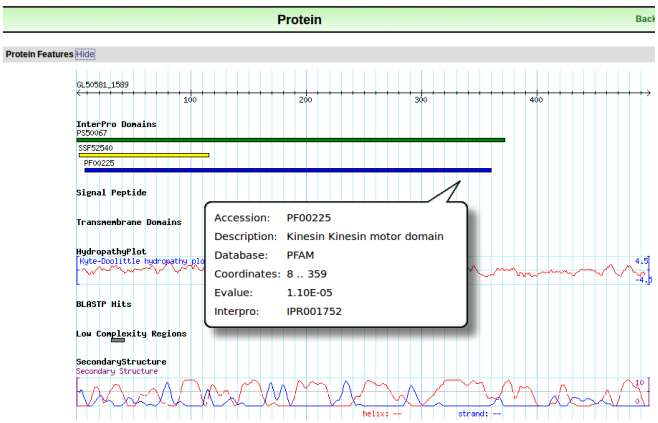
   (Hint: add a step to the strategy. Go to the "Interpro Domain" search under similarity/pattern, start typing the work kinesin and it should autocomplete.)





c. Go to the gene page for GL50581_1589 and look at the protein feature section. Does this look like a possible motor protein?

Hint: click on the ID for GL50581_1589 in the result table to go to the gene page. Scroll down to the protein section and mouse over the glyphs in the Protein Features graphic.

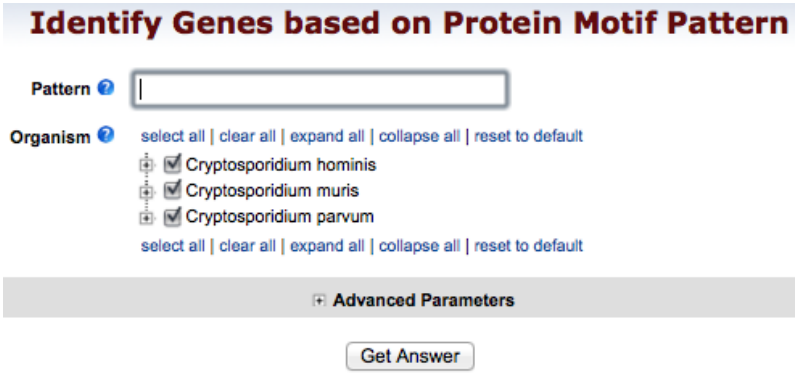


2. Using regular expressions to find motifs in CryptoDB: finding genes with the YXXΦ receptor signal motif

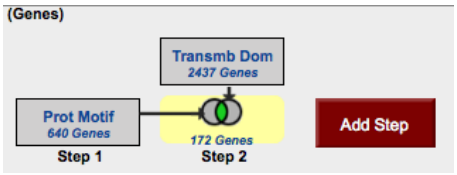
Note: for this exercise use http://cryptodb.org

a. 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.

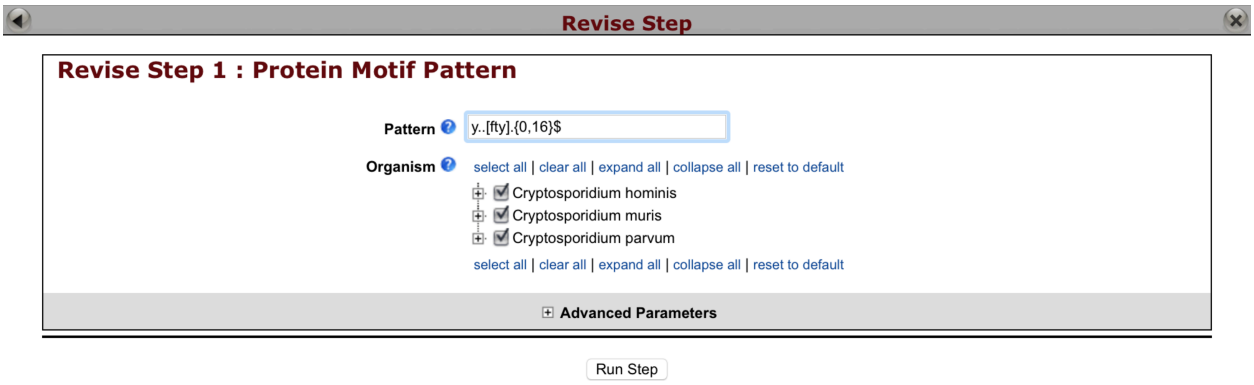
b. Use the "protein motif pattern" search to find all *Cryptosporidium* proteins that contain this motif anywhere in the terminal 10 amino acids of proteins. (hint: for your regular expression, remember that you want the first amino acid to be a tyrosine, followed any two amino acids, followed by any bulky hydrophobic amino acid (phenylalanine, tyrosine, threonine). Refer to regular expression tutorial if you need to).



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

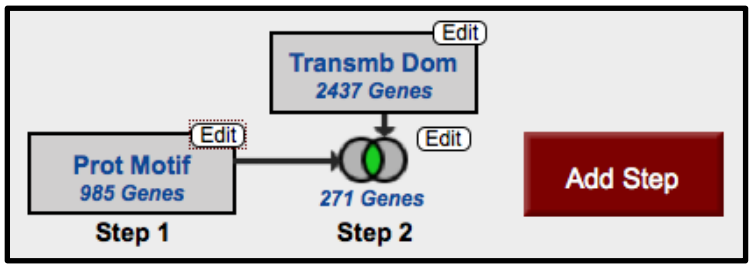


d. What would happen if you revise the first step (the motif pattern step) to include genes with the sorting motif in the C-terminal 20 amino acids? (hint: edit the first step and modify your regular expression).



Here is a saved strategy that provides you with the results of the above search:

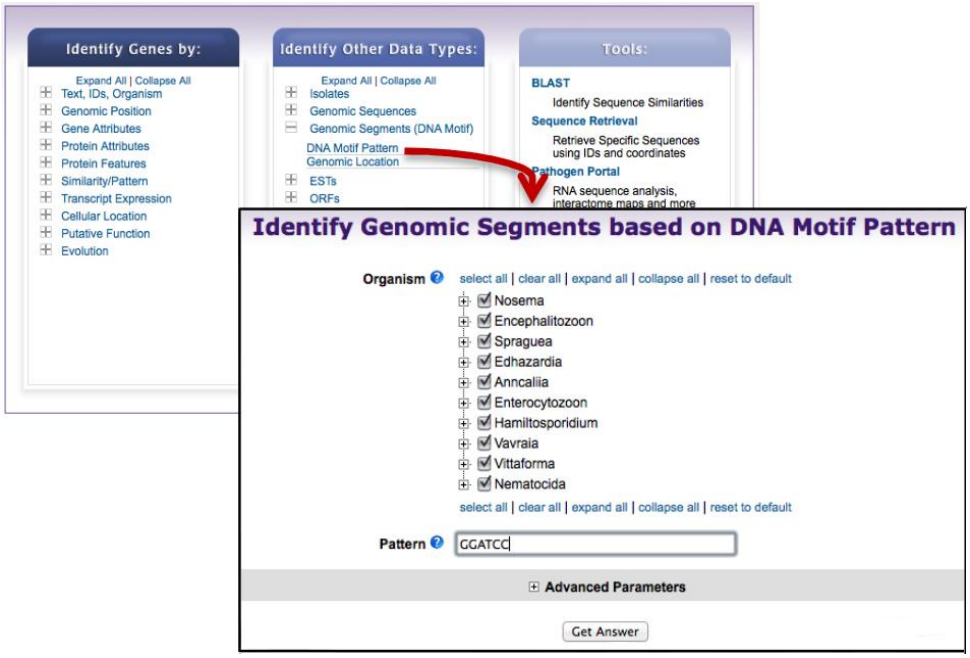
http://cryptodb.org/cryptodb/im.do?s=928309b4c1b9ef3f



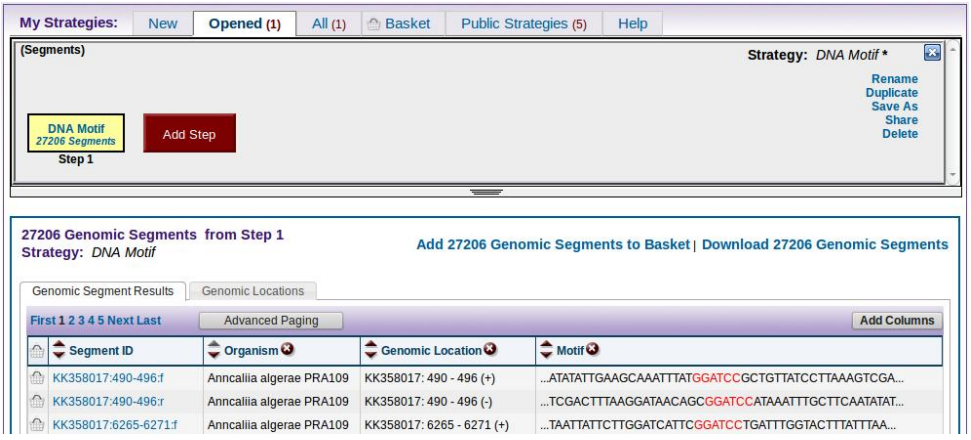
3. **Identification of specific DNA motifs.**
   For this exercise use http://microsporidiadb.org

a. Find all *Bam*HI 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: *Bam*HI = GGATCC and the DNA motif search is under the heading "Genomic Segments".



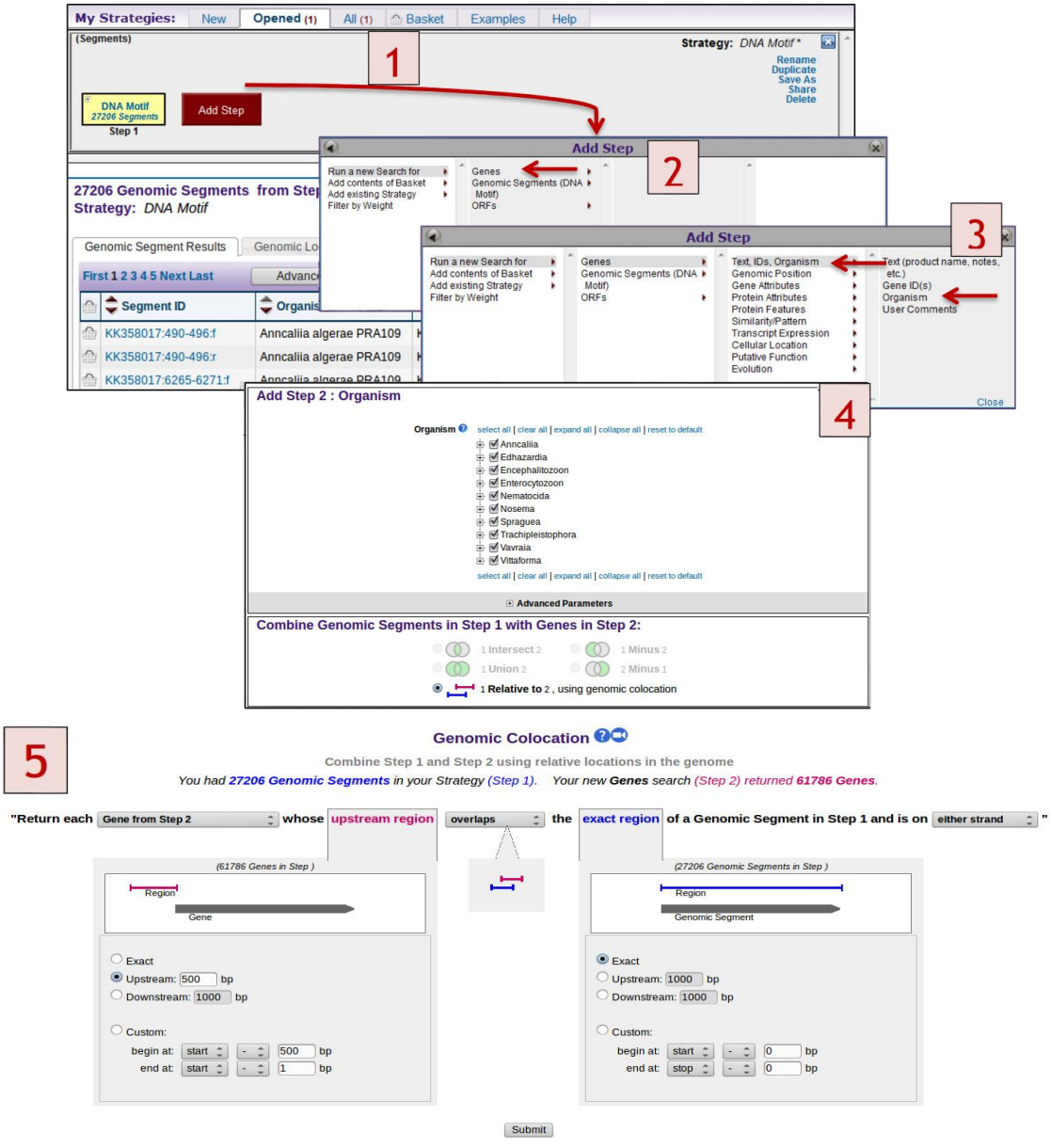
b. How many times does the *Bam*HI site occur in the genomes you searched? Take a look at your results; notice the Genomic location and the Motif columns.



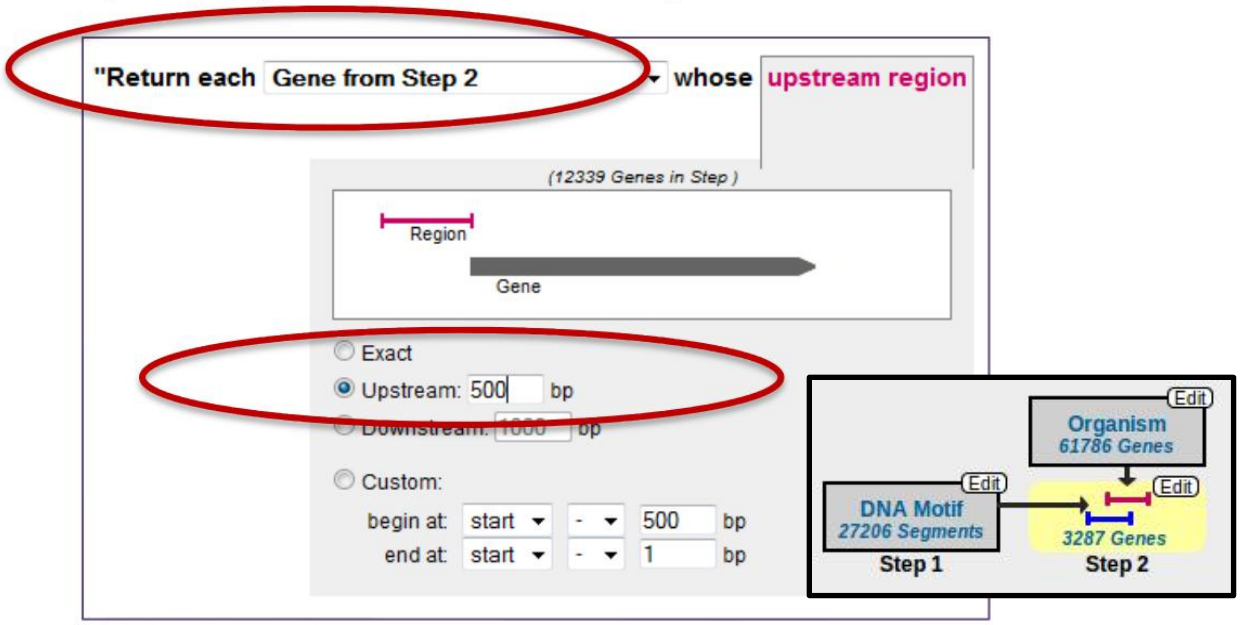
c. Find genes that have one of these *Bam*HI sites within 500 nucleotides upstream of their start.

In section 1 you found *Bam*HI sites, but now you are looking for genes that have one of these sites located within 500 nucleotides upstream of their start.

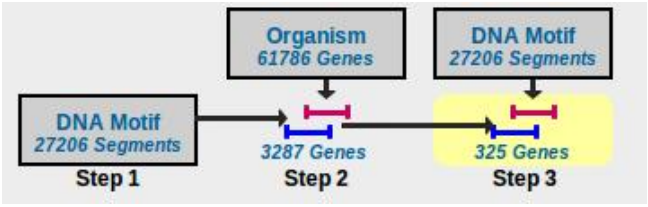
Hint: You can achieve this by running a genomic collocation search that defines the genomic relationship between the *Bam*HI sites and genes. Add a "Genes by Organism" step to the motif search and select the "1 relative to 2, using genomic locations" option.



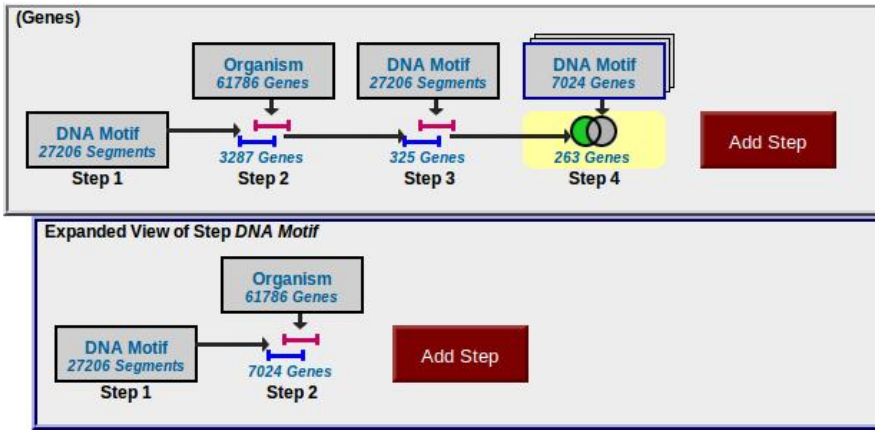
How did you modify the location relative to genes? How many genes did you get?



d. Using a similar sequence of steps as in part 2, define which of these genes also have a *Bam*HI 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.



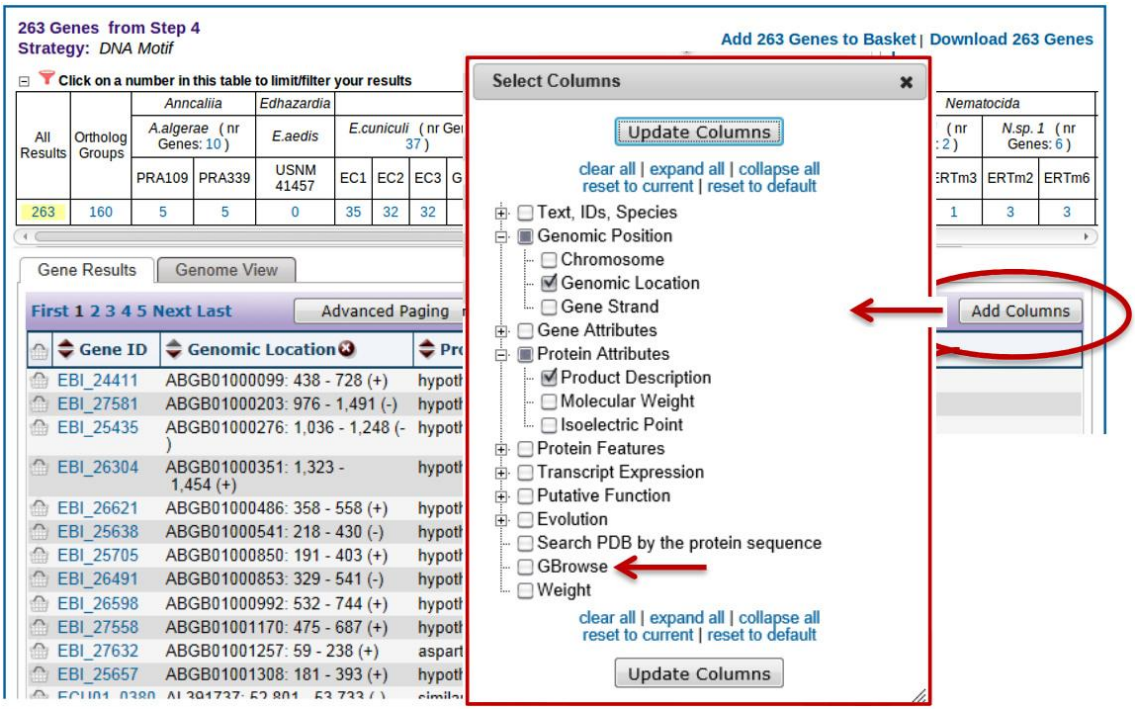
e. Taking this a step further, define which of these genes do NOT contain a *Bam*HI site within them.



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 *Bam*HI 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.