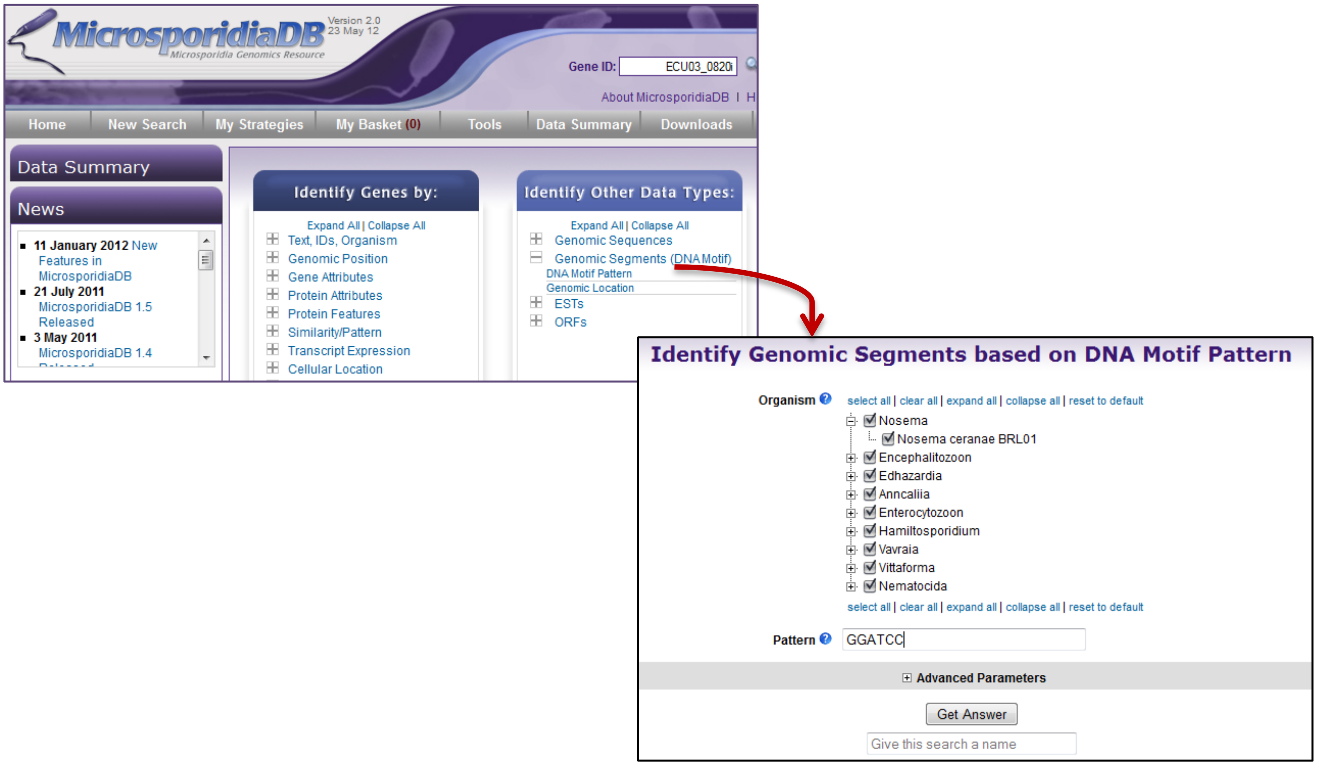
7.1 Identification of specific DNA motifs.

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".



Identify Genomic Segments based on DNA Motif Pattern

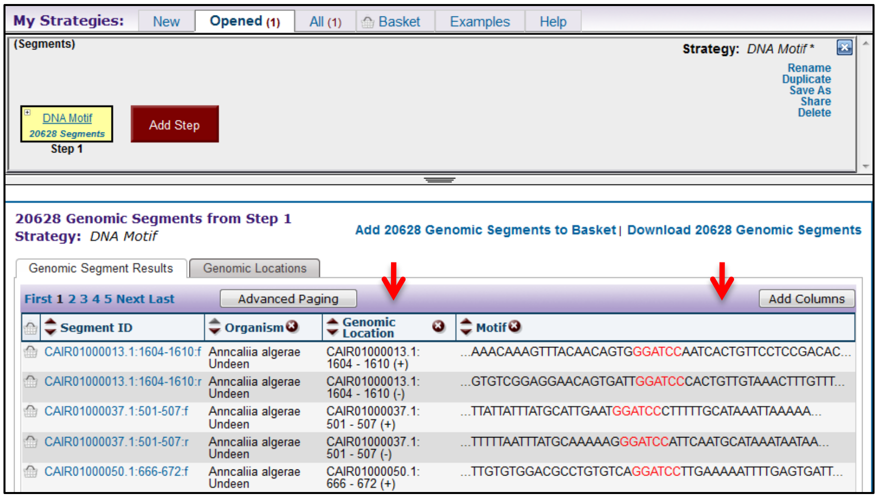
Organism: select all | clear all | expand all | collapse all | reset to default

- ☑ Nosema
  - ☑ Nosema ceranae BRL01
- ☑ Encephalitozoon
- ☑ Edhazardia
- ☑ Anncaliia
- ☑ Enterocytozoon
- ☑ Hamiltosporidium
- ☑ Vavraia
- ☑ Vittaforma
- ☑ Nematocida

select all | clear all | expand all | collapse all | reset to default

Pattern: GGATCC

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.

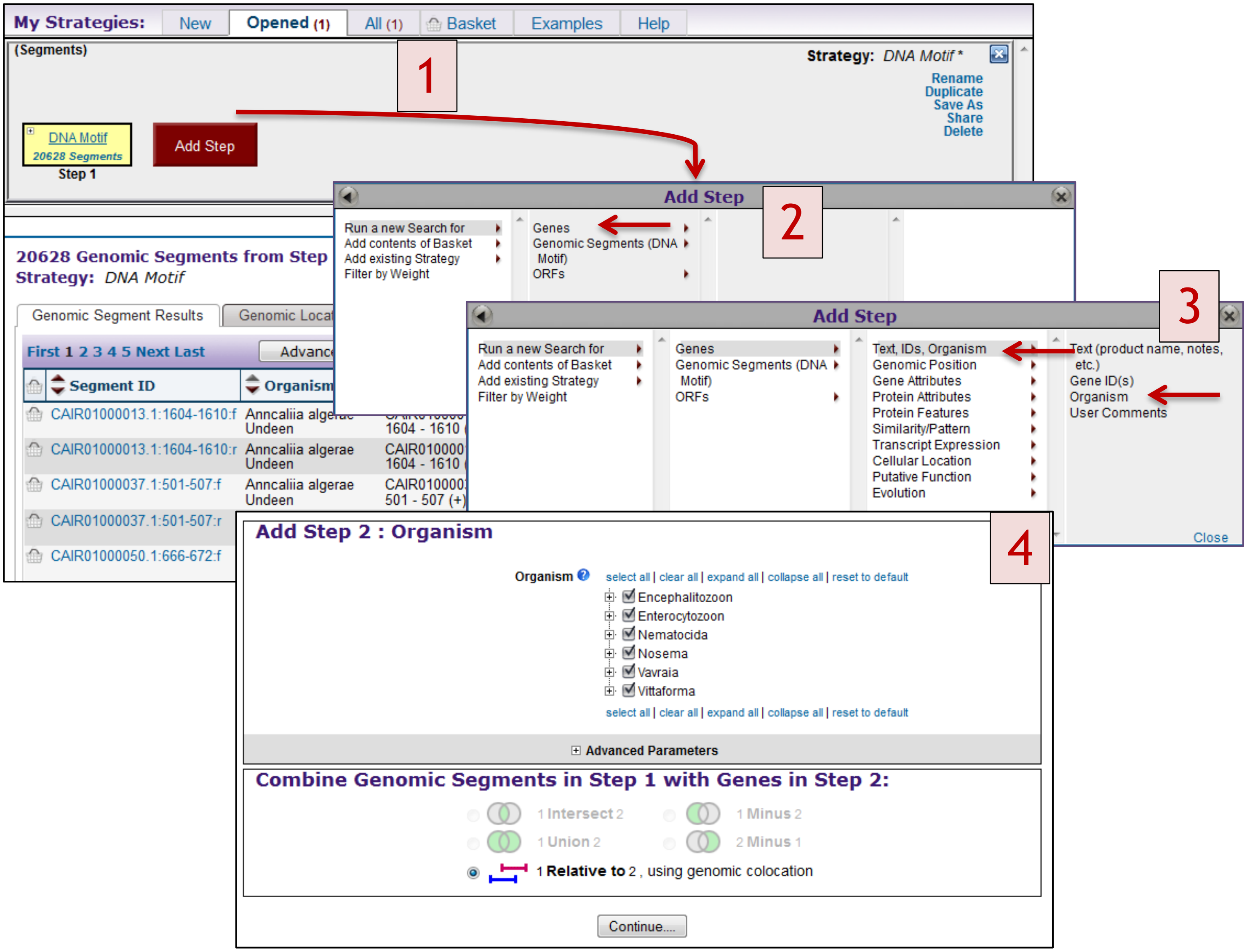


20628 Genomic Segments from Step 1
Strategy: *DNA Motif*

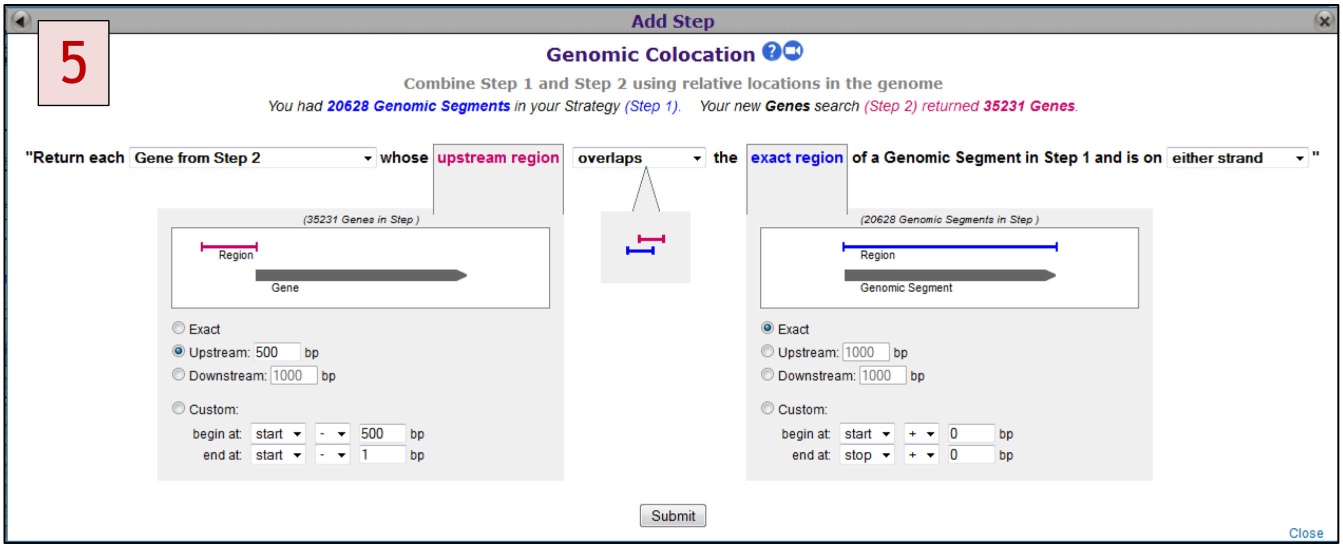
Segment ID	Organism	Genomic Location	Motif
CAIR01000013.1:1604-1610:f	Anncaliia algerae Undeen	CAIR01000013.1: 1604 - 1610 (+)	...AAACAAAGTTTACAACAGTGGGATCCAATCACTGTTCCTCCGACAC...
CAIR01000013.1:1604-1610:r	Anncaliia algerae Undeen	CAIR01000013.1: 1604 - 1610 (-)	...GTGTCGGAGGAACAGTGATTGGATCCCACTGTTGTAAACTTTGTTT...
CAIR01000037.1:501-507:f	Anncaliia algerae Undeen	CAIR01000037.1: 501 - 507 (+)	...TTATTATTTATGCATTGAATGGATCCCTTTTTGCATAAATTAAAAA...
CAIR01000037.1:501-507:r	Anncaliia algerae Undeen	CAIR01000037.1: 501 - 507 (-)	...TTTTTAATTTATGCAAAAAGGGATCCATTCAATGCATAAATAATAA...
CAIR01000050.1:666-672:f	Anncaliia algerae Undeen	CAIR01000050.1: 666 - 672 (+)	...TTGTGTGGACGCCTGTGTCAGGATCCTTGAAAAATTTTGAGTGATT...

7.2 Find genes that have one of these BamHI sites within 500 nucleotides upstream of their start.
In the section 7.1 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.



5



Add Step

Genomic Colocation

Combine Step 1 and Step 2 using relative locations in the genome

*You had **20628 Genomic Segments** in your Strategy (Step 1). Your new **Genes** search (Step 2) returned **35231 Genes**.*

"Return each **Gene from Step 2** whose **upstream region** **overlaps** the **exact region** of a Genomic Segment in Step 1 and is on **either strand** "

(35231 Genes in Step)

○ Exact
● Upstream: 500 bp
○ Downstream: 1000 bp

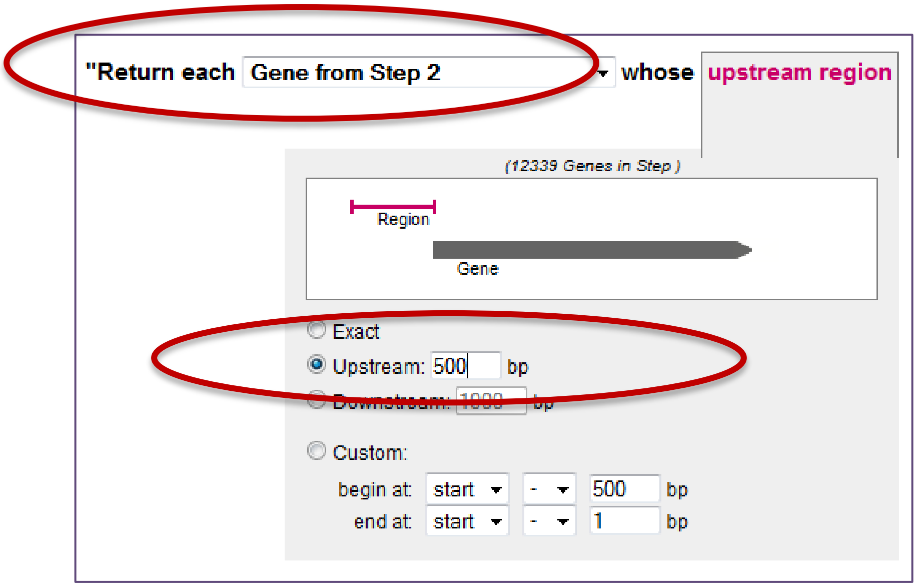
○ Custom:
begin at: start - 500 bp
end at: start - 1 bp

(20628 Genomic Segments in Step)

● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp

○ Custom:
begin at: start + 0 bp
end at: stop + 0 bp

How did you modify the location relative to genes?



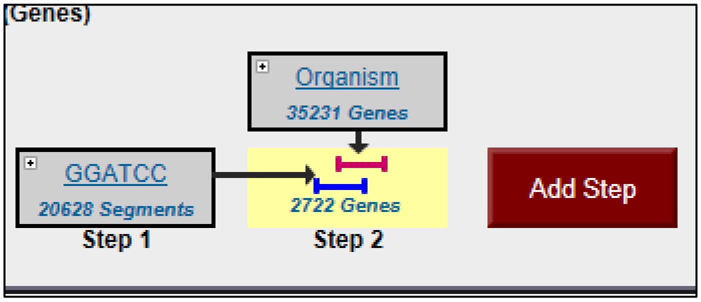
"Return each **Gene from Step 2** whose **upstream region**

(12339 Genes in Step)

○ Exact
● Upstream: 500 bp
○ Downstream: 1000 bp

○ Custom:
begin at: start - 500 bp
end at: start - 1 bp

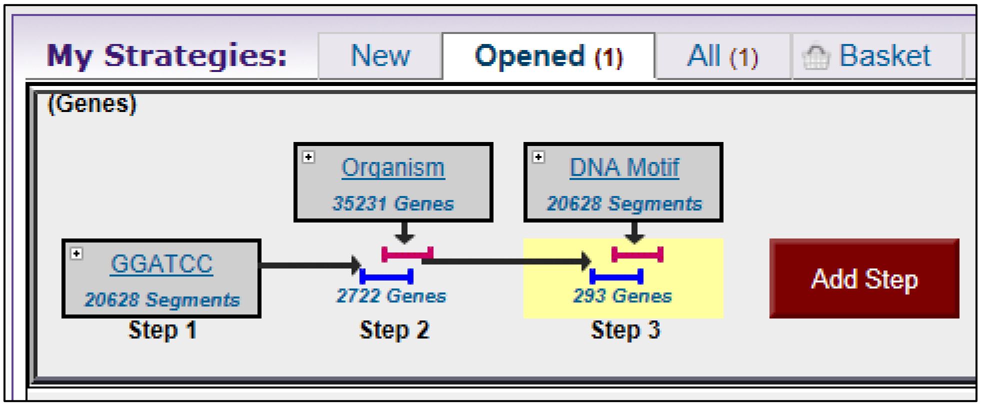
How many genes did you get?



(Genes)

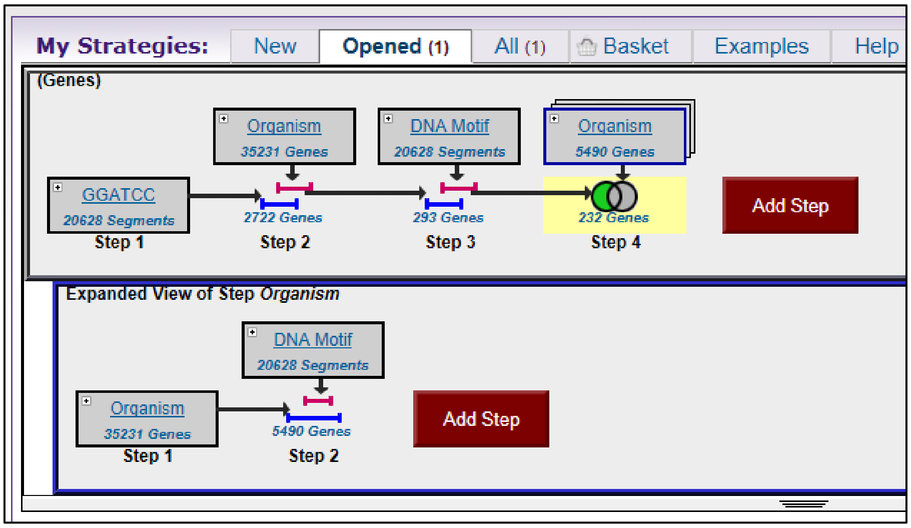
7.3 Using a similar sequence of steps as in part 7.2, define which of these genes also have a BamHI site in their 500 nucleotide downstream region.

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



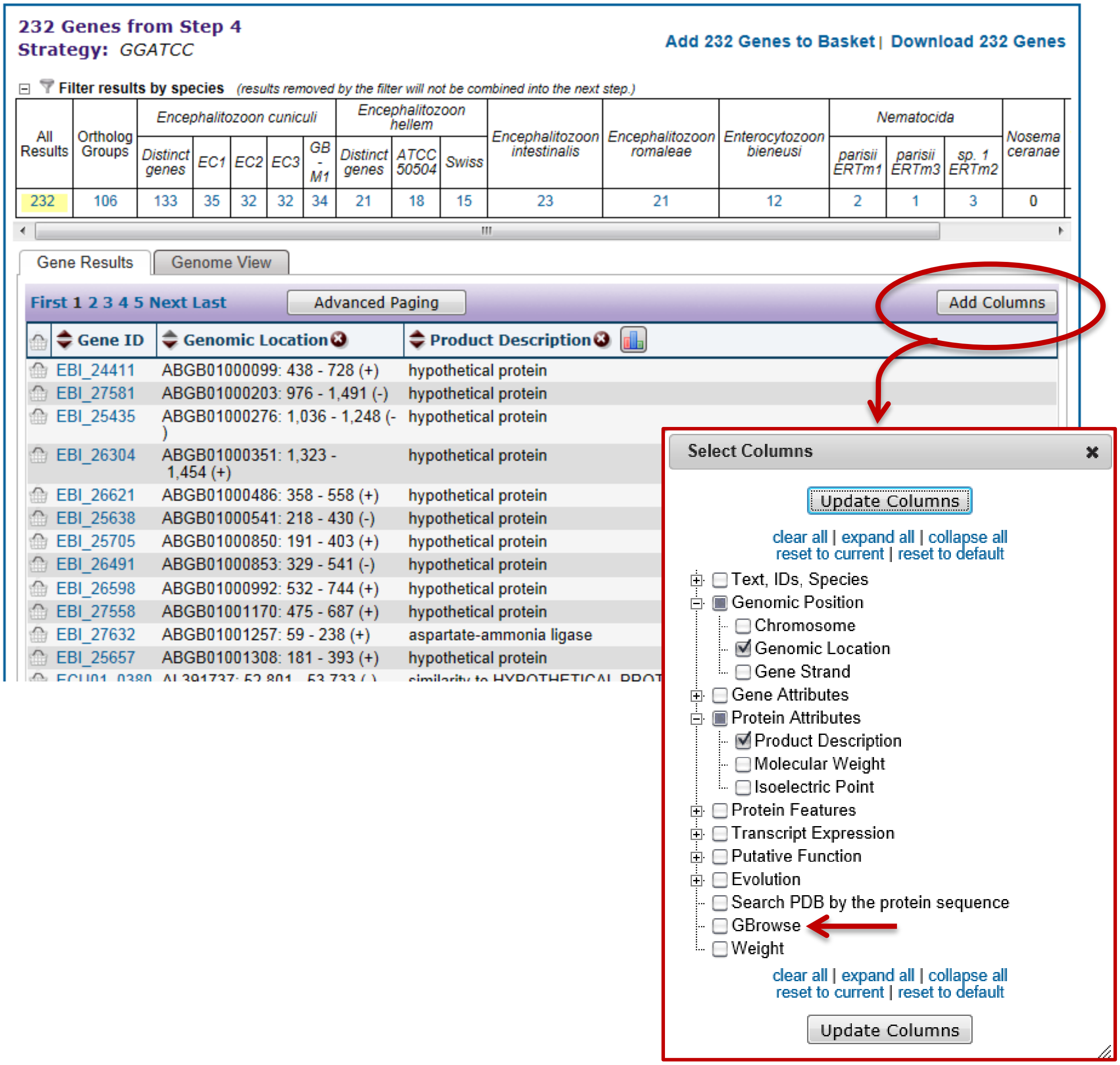
7.4 Taking this a step further, define which of these genes do **NOT** contain a BamHI site within them.

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.



232 Genes from Step 4
Strategy: *GGATCC*

Add 232 Genes to Basket | Download 232 Genes

⊟ Filter results by species *(results removed by the filter will not be combined into the next step.)*

All Results	Ortholog Groups	*Encephalitozoon cuniculi*					*Encephalitozoon hellem*			*Encephalitozoon intestinalis*	*Encephalitozoon romaleae*	*Enterocytozoon bieneusi*	*Nematocida*			*Nosema ceranae*
		Distinct genes	*EC1*	*EC2*	*EC3*	*GB-M1*	*Distinct genes*	*ATCC 50504*	*Swiss*				*parisii ERTm1*	*parisii ERTm3*	*sp. 1 ERTm2*	
232	106	133	35	32	32	34	21	18	15	23	21	12	2	1	3	0

Gene Results | Genome View

First 1 2 3 4 5 Next Last | Advanced Paging | Add Columns

Gene ID	Genomic Location	Product Description
EBI_24411	ABGB01000099: 438 - 728 (+)	hypothetical protein
EBI_27581	ABGB01000203: 976 - 1,491 (-)	hypothetical protein
EBI_25435	ABGB01000276: 1,036 - 1,248 (-)	hypothetical protein
EBI_26304	ABGB01000351: 1,323 - 1,454 (+)	hypothetical protein
EBI_26621	ABGB01000486: 358 - 558 (+)	hypothetical protein
EBI_25638	ABGB01000541: 218 - 430 (-)	hypothetical protein
EBI_25705	ABGB01000850: 191 - 403 (+)	hypothetical protein
EBI_26491	ABGB01000853: 329 - 541 (-)	hypothetical protein
EBI_26598	ABGB01000992: 532 - 744 (+)	hypothetical protein
EBI_27558	ABGB01001170: 475 - 687 (+)	hypothetical protein
EBI_27632	ABGB01001257: 59 - 238 (+)	aspartate-ammonia ligase
EBI_25657	ABGB01001308: 181 - 393 (+)	hypothetical protein
ECU01_0380	AL391737: 52,801 - 53,733 (-)	similarity to HYPOTHETICAL PROT

Select Columns ✖

Update Columns

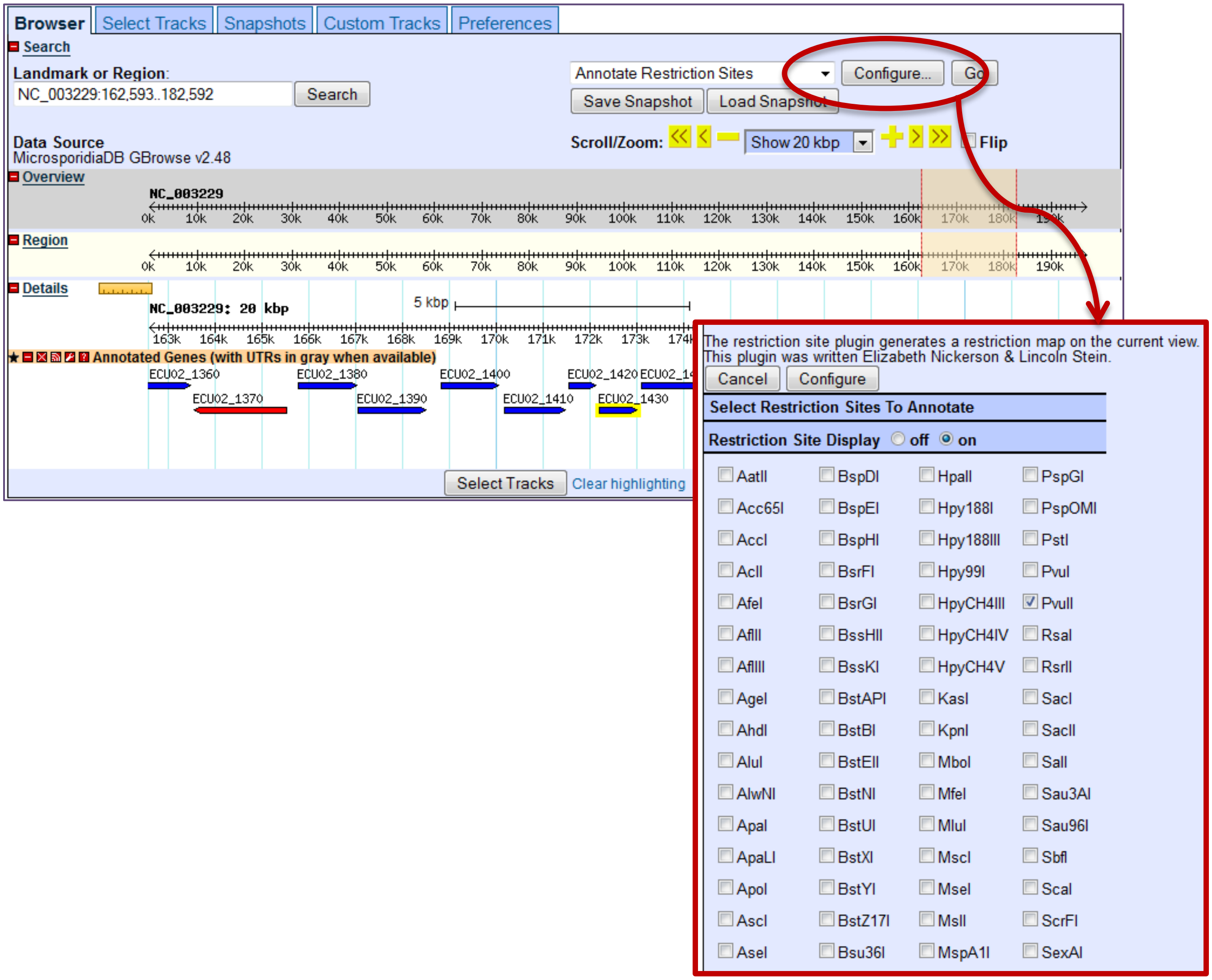
clear all | expand all | collapse all
reset to current | reset to default

- ☐ Text, IDs, Species
- ■ Genomic Position
  - ☐ Chromosome
  - ☑ Genomic Location
  - ☐ Gene Strand
- ☐ Gene Attributes
- ■ Protein Attributes
  - ☑ Product Description
  - ☐ Molecular Weight
  - ☐ Isoelectric Point
- ☐ Protein Features
- ☐ Transcript Expression
- ☐ Putative Function
- ☐ Evolution
- ☐ Search PDB by the protein sequence
- ☐ GBrowse
- ☐ Weight

clear all | expand all | collapse all
reset to current | reset to default

Update Columns

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.



Landmark or Region: NC_003229:162,593..182,592

Data Source
MicrosporidiaDB GBrowse v2.48

Annotate Restriction Sites

Show 20 kbp

The restriction site plugin generates a restriction map on the current view. This plugin was written Elizabeth Nickerson & Lincoln Stein.

Select Restriction Sites To Annotate

Restriction Site Display ○ off ● on

☐ AatII
☐ Acc65I
☐ AccI
☐ AclI
☐ AfeI
☐ AflII
☐ AflIII
☐ AgeI
☐ AhdI
☐ AluI
☐ AlwNI
☐ ApaI
☐ ApaLI
☐ ApoI
☐ AscI
☐ AseI
☐ BspDI
☐ BspEI
☐ BspHI
☐ BsrFI
☐ BsrGI
☐ BssHII
☐ BssKI
☐ BstAPI
☐ BstBI
☐ BstEII
☐ BstNI
☐ BstUI
☐ BstXI
☐ BstYI
☐ BstZ17I
☐ Bsu36I
☐ HpaII
☐ Hpy188I
☐ Hpy188III
☐ Hpy99I
☐ HpyCH4III
☐ HpyCH4IV
☐ HpyCH4V
☐ KasI
☐ KpnI
☐ MboI
☐ MfeI
☐ MluI
☐ MscI
☐ MseI
☐ MslI
☐ MspA1I
☐ PspGI
☐ PspOMI
☐ PstI
☐ PvuI
☑ PvuII
☐ RsaI
☐ RsrII
☐ SacI
☐ SacII
☐ SalI
☐ Sau3AI
☐ Sau96I
☐ SbfI
☐ ScaI
☐ ScrFI
☐ SexAI