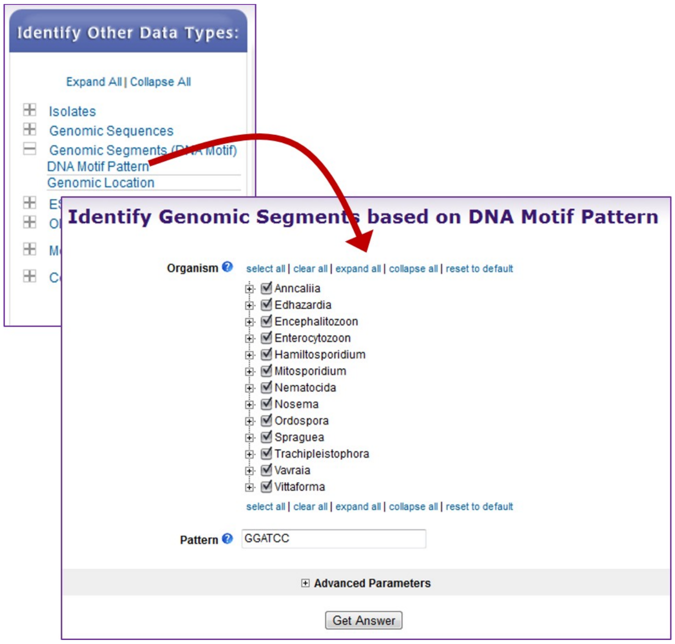
DNA Motifs and Genomic Colocation

Identify a specific DNA motif and colocate this motif with genes:
For this exercise use http://microsporidiadb.org

a. **Find all *Bam*HI restriction sites in all microsporidia genomic sequences available in MicrosporidiaDB.** The DNA motif search can be used to finds simple DNA motifs or complex motifs like transcription factor binding sites using regular expressions. A DNA restriction site can be defined by a DNA motif. For example the *Bam*H1 restriction site is GGATCC.


Identify Other Data Types:
Expand All | Collapse All
Isolates
Genomic Sequences
Genomic Segments (DNA Motif)
DNA Motif Pattern
Genomic Location
ES
Ol
M
C
select all | clear all | expand all | collapse all | reset to default
Advanced Parameters
Get Answer

Identify Genomic Segments based on DNA Motif Pattern

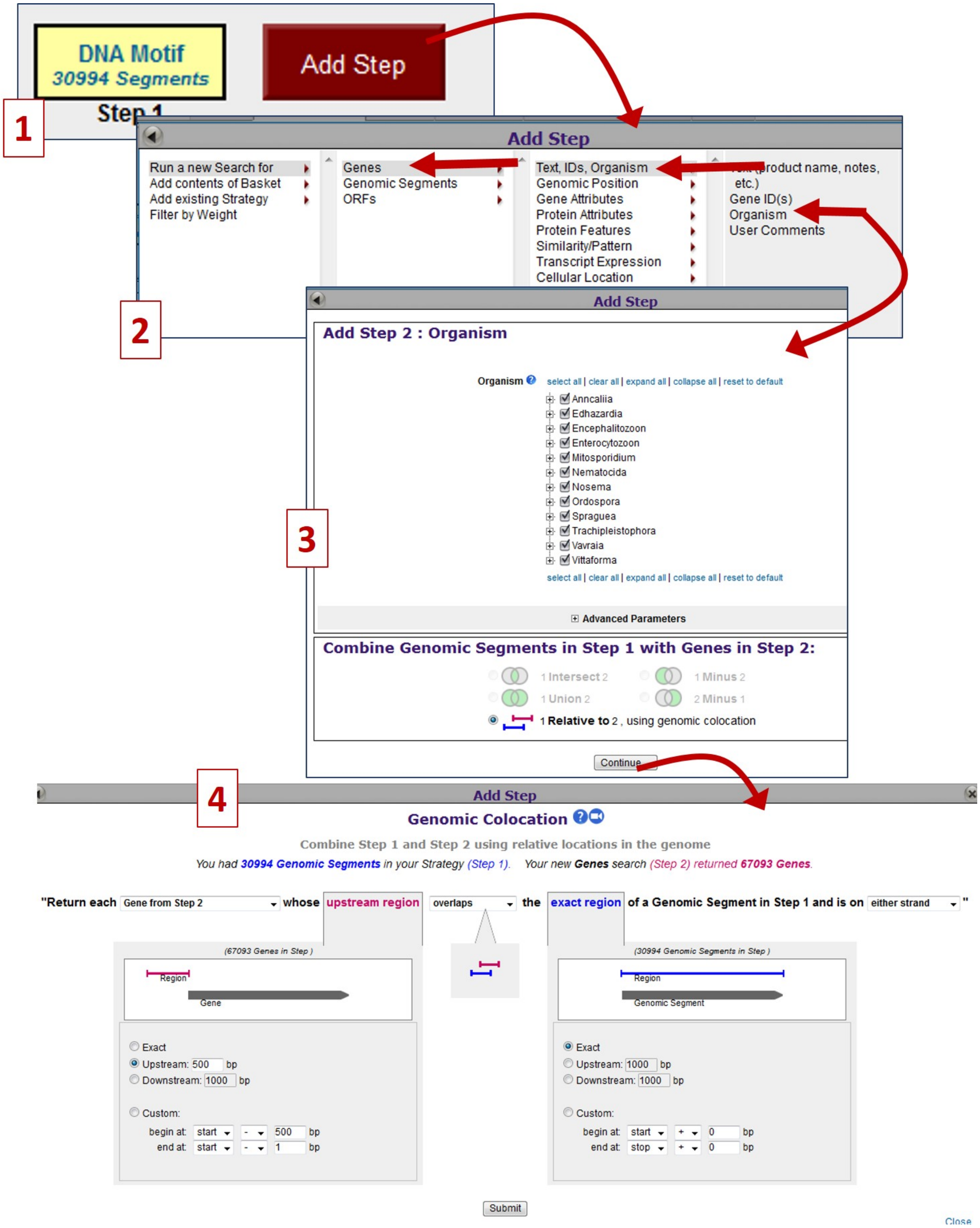
Organism
☑ Anncaliia
☑ Edhazardia
☑ Encephalitozoon
☑ Enterocytozoon
☑ Hamiltosporidium
☑ Mitosporidium
☑ Nematocida
☑ Nosema
☑ Ordospora
☑ Spraguea
☑ Trachipleistophora
☑ Vavraia
☑ Vittaforma

Pattern GGATCC

- 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.

b. **Find genes that have one of these *Bam*HI sites within 500 nucleotides upstream of their start.**
In 4a you found *Bam*HI sites anywhere in the genome. 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.



Add Step 2 : Organism

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

- ☑ Anncaliia
- ☑ Edhazardia
- ☑ Encephalitozoon
- ☑ Enterocytozoon
- ☑ Mitosporidium
- ☑ Nematocida
- ☑ Nosema
- ☑ Ordospora
- ☑ Spraguea
- ☑ Trachipleistophora
- ☑ Vavraia
- ☑ Vittaforma

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

⊞ Advanced Parameters

Combine Genomic Segments in Step 1 with Genes in Step 2:

○ 1 Intersect 2 ○ 1 Minus 2
○ 1 Union 2 ○ 2 Minus 1
● 1 Relative to 2 , using genomic colocation

Continue

Genomic Colocation

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

*You had **30994 Genomic Segments** in your Strategy (Step 1). Your new **Genes** search (Step 2) returned **67093 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 "

(67093 Genes in Step)

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

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

(30994 Genomic Segments in Step)

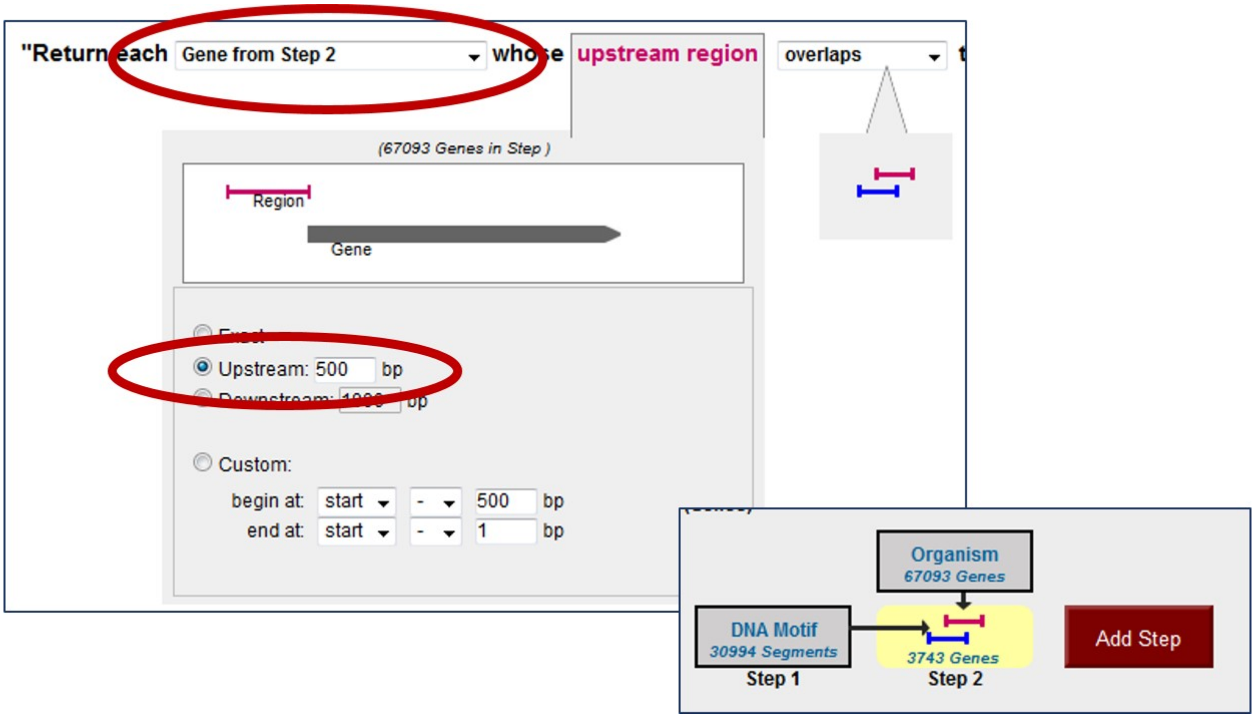
● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp

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

Submit

Close

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



"Return each Gene from Step 2 whose upstream region overlaps t

(67093 Genes in Step)

○ Exact

● Upstream: 500 bp

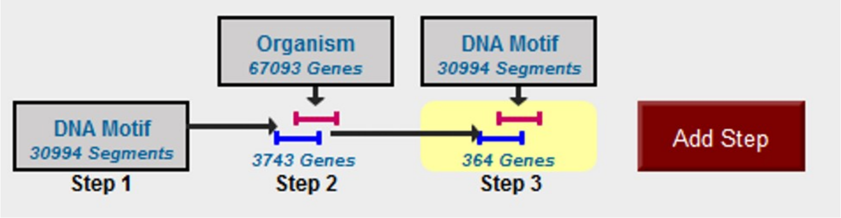
○ Downstream: 1000 bp

○ Custom:

begin at: start - 500 bp

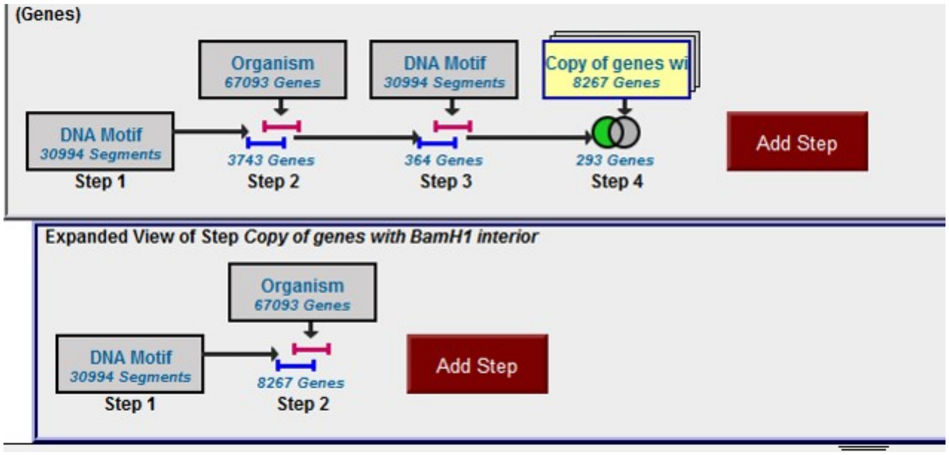
end at: start - 1 bp

c. **Using a similar sequence of steps as in 4b, define which of these genes also have a *Bam*HI site in their 500 nucleotide downstream region.**
Hint: add a search for the BamH1 site and collocate that with the genes that have a BamH1 site upstream of their start sites.



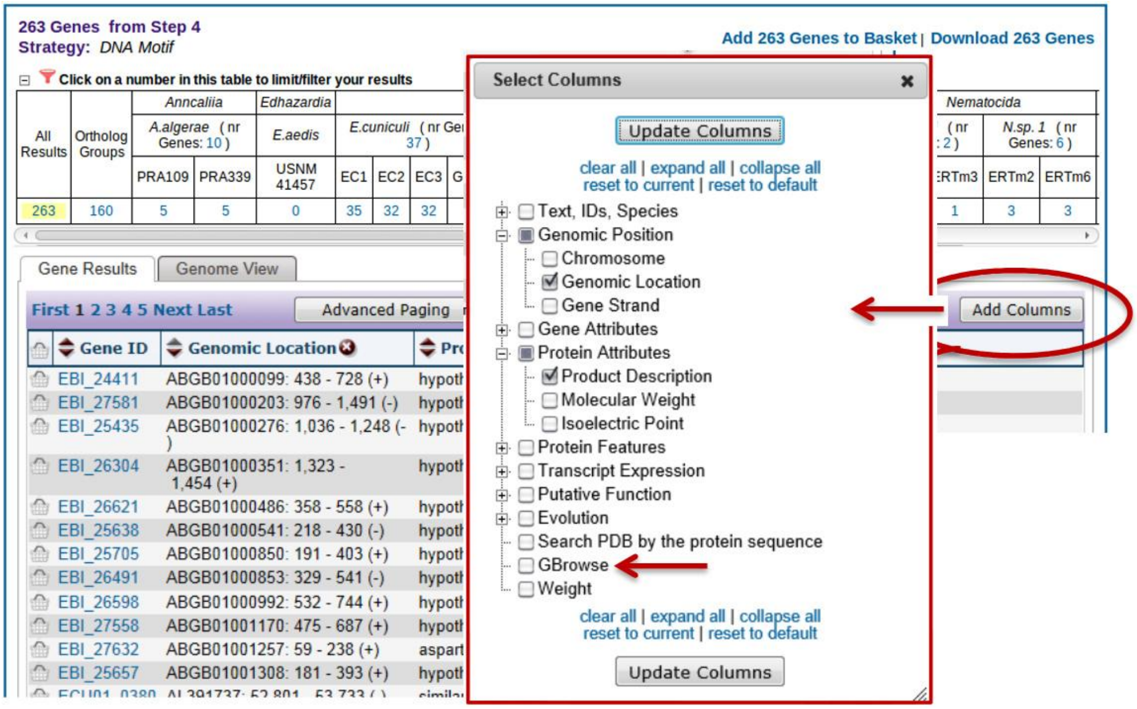
Step 1 = BamH1 anywhere in genome
Step 2 = all genes
2 (1+2) = genes with BamH1 1000bp upstream
Step 3 = BamH1 anywhere in genome
3 (2+3) = genes from 2 with BamH1 1000bp downstream

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



263 Genes from Step 4
Strategy: *DNA Motif*

Add 263 Genes to Basket | Download 263 Genes

⊟ Click on a number in this table to limit/filter your results

All Results	Ortholog Groups	*Anncaliia*		*Edhazardia*					*Nematocida*		
		A.algerae (nr Genes: 10)		*E.aedis*	*E.cuniculi* (nr Ge 37)				(nr : 2)	*N.sp. 1* (nr Genes: 6)	
		PRA109	PRA339	USNM 41457	EC1	EC2	EC3	G	ERTm3	ERTm2	ERTm6
263	160	5	5	0	35	32	32		1	3	3

Gene Results | Genome View

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

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

Select Columns ✖

Update Columns

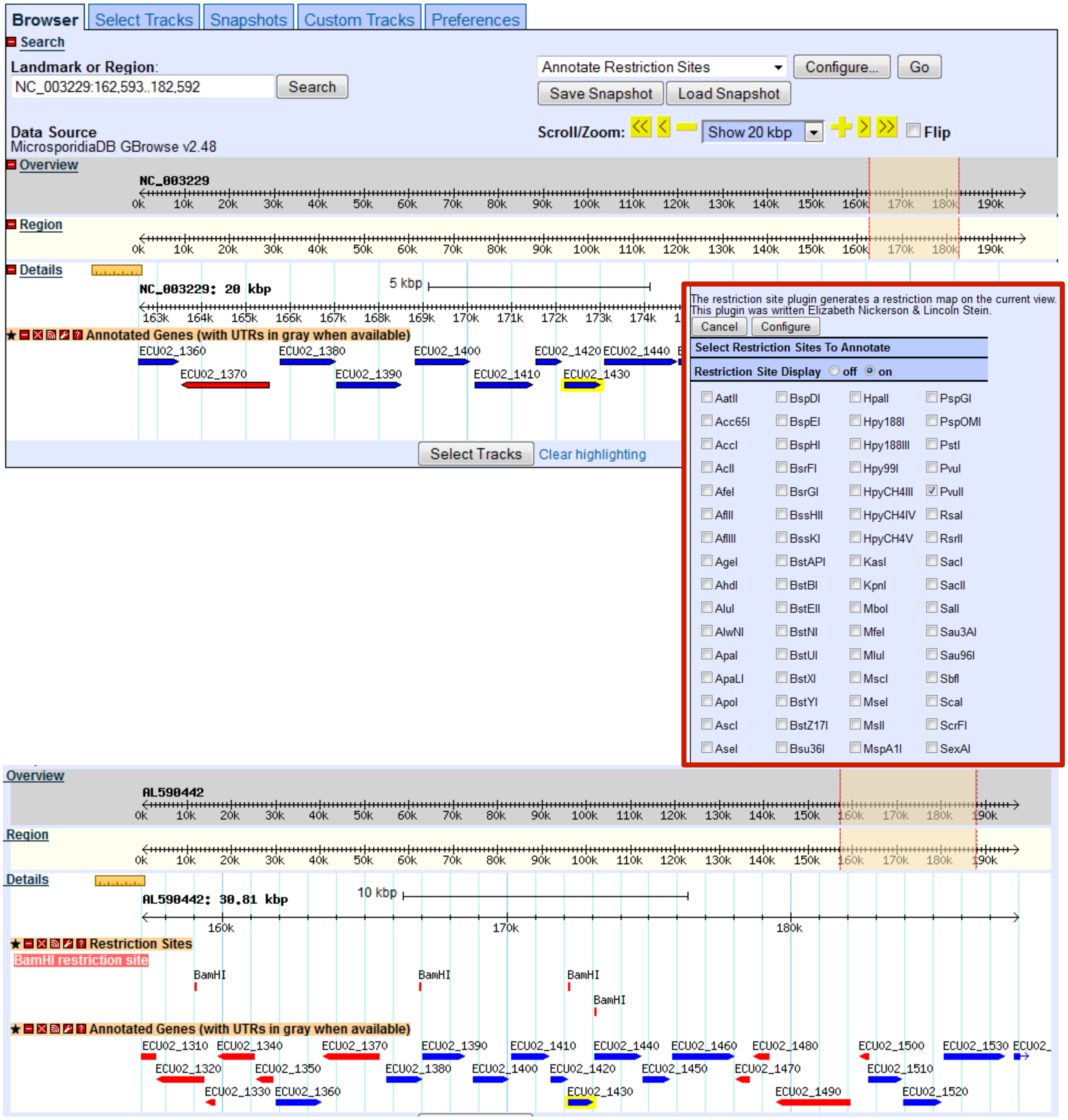
clear all | expand all | collapse all
reset to current | reset to default

- ⊞ □ 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 display 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

Annotate Restriction Sites

Data Source MicrosporidiaDB GBrowse v2.48

Scroll/Zoom: Show 20 kbp ☐ Flip

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

Cancel Configure

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