

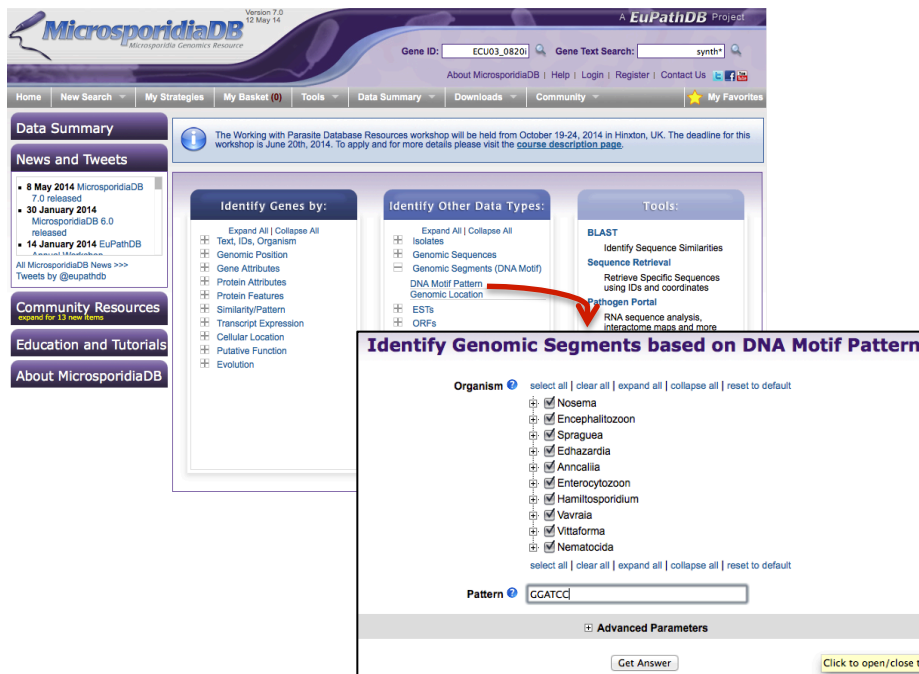
DNA motifs Exercise 7

7.1 Identification of specific DNA motifs.

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

- 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”.



- 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.

The screenshot shows the search results page for the DNA motif 'GGATCC'. The page title is '20628 Genomic Segments from Step 1'. The table below shows the results:

Segment ID	Organism	Genomic Location	Motif
CAIR01000013.1:1604-1610.f	Anncalia algerae Undeen	CAIR01000013.1:1604 - 1610 (+)	...AAACAAAGTTTACAACAGTGGATCCCACTGTTCTCCGACAC...
CAIR01000013.1:1604-1610.r	Anncalia algerae Undeen	CAIR01000013.1:1604 - 1610 (-)	...GTGTCGGAGGAACAGTGATGGATCCCACTGTTGAACTTTGTT...
CAIR01000037.1:501-507.f	Anncalia algerae Undeen	CAIR01000037.1:501 - 507 (+)	...TTATTATTATGCATTAAGTGGATCCCTTTTGCATAAATAAAAA...
CAIR01000037.1:501-507.r	Anncalia algerae Undeen	CAIR01000037.1:501 - 507 (-)	...TTTTAATTATGCAAAAAGGGATCCATTCAATGCATAAATAATA...
CAIR01000050.1:666-672.f	Anncalia algerae Undeen	CAIR01000050.1:666 - 672 (+)	...TTGTGTGGACGCTGTGTCAAGATCCCTTGAATAATTTGAGTGAT...

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.

1 Add Step

2 Genes

3 Text, IDs, Organism

4 1 Relative to 2, using genomic collocation

Segment ID	Organism
CAIR01000013.1:1604-1610.f	Anncaliia algerae CAIR01000013.1:1604-1610.f
CAIR01000013.1:1604-1610.r	Anncaliia algerae CAIR01000013.1:1604-1610.r
CAIR01000037.1:501-507.f	Anncaliia algerae CAIR01000037.1:501-507.f
CAIR01000037.1:501-507.r	Anncaliia algerae CAIR01000037.1:501-507.r
CAIR01000050.1:666-672.f	Anncaliia algerae CAIR01000050.1:666-672.f

Add Step 2 : Organism

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

- Encephalitozoon
- Enterocytozoon
- Nematocida
- Nosema
- Vavraia
- Vittaforma

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

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 collocation

Continue....

5

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: bp

Downstream: bp

Custom:

begin at: bp

end at: bp

(20628 Genomic Segments in Step)

Exact

Upstream: bp

Downstream: bp

Custom:

begin at: bp

end at: 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: bp

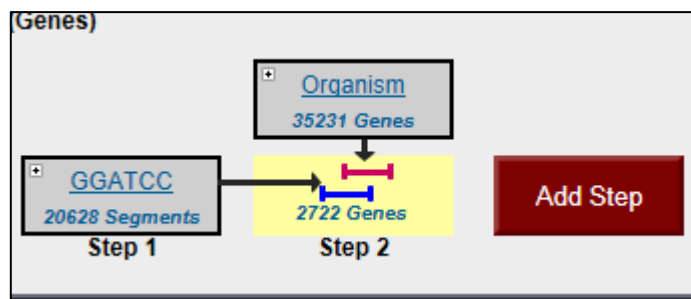
Downstream: bp

Custom:

begin at: bp

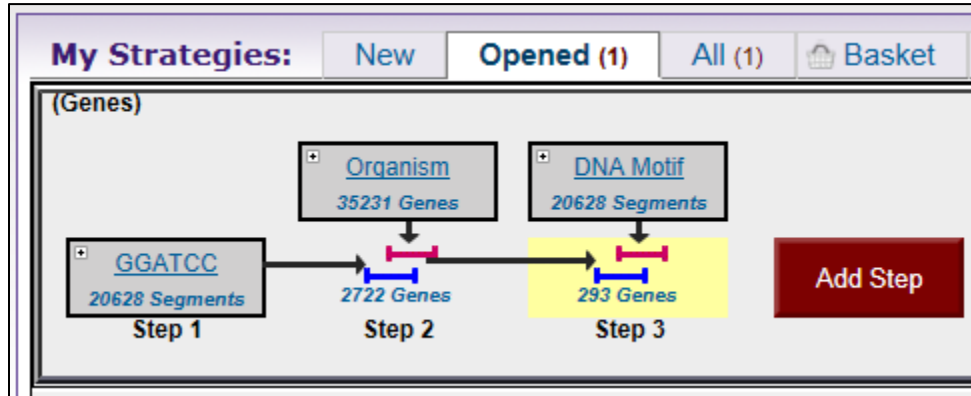
end at: bp

How many genes did you get?



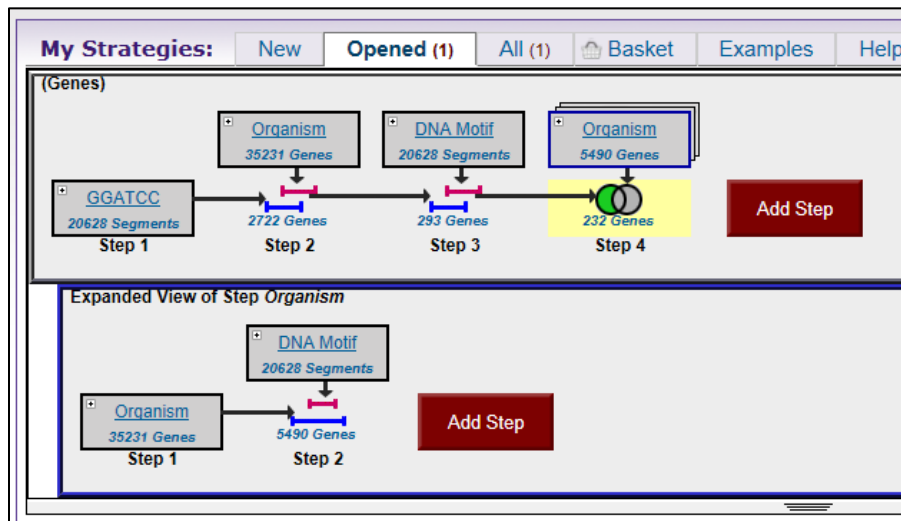
7.3 Using a similar sequence of steps as in part 7.2, define which of these genes also have a BamHI site in their 500 nucleotide downstream region.

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 Add 232 Genes to Basket | Download 232 Genes
Strategy: GGATCC

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

Select Columns x

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.

The screenshot shows the GBrowse interface with the 'Annotate Restriction Sites' configuration dialog open. The 'Configure...' button is circled in red, and an arrow points from it to the configuration window. The configuration window has a title bar that reads: 'The restriction site plugin generates a restriction map on the current view. This plugin was written Elizabeth Nickerson & Lincoln Stein.' Below the title bar are 'Cancel' and 'Configure' buttons. The main section is titled 'Select Restriction Sites To Annotate' and includes a 'Restriction Site Display' section with radio buttons for 'off' and 'on' (selected). Below this is a grid of checkboxes for various restriction enzymes.

Select Restriction Sites To Annotate			
Restriction Site Display <input type="radio"/> off <input checked="" type="radio"/> on			
<input type="checkbox"/> AatII	<input type="checkbox"/> BspDI	<input type="checkbox"/> HpaII	<input type="checkbox"/> PspGI
<input type="checkbox"/> Acc65I	<input type="checkbox"/> BspEI	<input type="checkbox"/> Hpy188I	<input type="checkbox"/> PspOMI
<input type="checkbox"/> AccI	<input type="checkbox"/> BspHI	<input type="checkbox"/> Hpy188III	<input type="checkbox"/> PstI
<input type="checkbox"/> AclI	<input type="checkbox"/> BsrFI	<input type="checkbox"/> Hpy99I	<input type="checkbox"/> PvuI
<input type="checkbox"/> AfeI	<input type="checkbox"/> BsrGI	<input type="checkbox"/> HpyCH4III	<input checked="" type="checkbox"/> PvuII
<input type="checkbox"/> AflII	<input type="checkbox"/> BssHII	<input type="checkbox"/> HpyCH4IV	<input type="checkbox"/> RsaI
<input type="checkbox"/> AflIII	<input type="checkbox"/> BssKI	<input type="checkbox"/> HpyCH4V	<input type="checkbox"/> RsrII
<input type="checkbox"/> AgeI	<input type="checkbox"/> BstAPI	<input type="checkbox"/> KasI	<input type="checkbox"/> SacI
<input type="checkbox"/> AhdI	<input type="checkbox"/> BstBI	<input type="checkbox"/> KpnI	<input type="checkbox"/> SacII
<input type="checkbox"/> AluI	<input type="checkbox"/> BstEII	<input type="checkbox"/> MboI	<input type="checkbox"/> SalI
<input type="checkbox"/> AlwNI	<input type="checkbox"/> BstNI	<input type="checkbox"/> MfeI	<input type="checkbox"/> Sau3AI
<input type="checkbox"/> ApaI	<input type="checkbox"/> BstUI	<input type="checkbox"/> MluI	<input type="checkbox"/> Sau96I
<input type="checkbox"/> ApaLI	<input type="checkbox"/> BstXI	<input type="checkbox"/> MscI	<input type="checkbox"/> SbfI
<input type="checkbox"/> ApoI	<input type="checkbox"/> BstYI	<input type="checkbox"/> MseI	<input type="checkbox"/> Scal
<input type="checkbox"/> AscI	<input type="checkbox"/> BstZ17I	<input type="checkbox"/> MslI	<input type="checkbox"/> ScrFI
<input type="checkbox"/> AseI	<input type="checkbox"/> Bsu36I	<input type="checkbox"/> MspA1I	<input type="checkbox"/> SexAI