

Jaimee Beckett
Units 3-4 Graded Homework

1. There are two attached files: znf214_mrna.txt and znf214_genomic.txt. Use Splign to find the mRNA and CDS coordinates in the genomic DNA.
 - a. **mRNA locations:** 614~896, 18114~18260, 19414~21651
 - b. **CDS locations:** 18134~18260, 19414~21107

mRNA locations:

HOME SEARCH SITE MAP			Overview		Online		Download		Documentation		Contacts	
#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)				
1	mrna(+)	chromosome:GRCh38:11:6998718:7020968:-1(+)	614-21651	100.00	100.00	100.00	0.00	0.00				

[Graphics](#)
[Text](#)

#	Query	Subject	Idty	Len	Q.Start	Q.Fin	S.Start	S.Fin	Type	Details
+1	mrna	chromosome:GRCh38:11	1	283	1	283	614	896	<exon>GC	M283
+1	mrna	chromosome:GRCh38:11	1	147	284	430	18114	18260	AG<exon>GT	M147
+1	mrna	chromosome:GRCh38:11	1	2238	431	2668	19414	21651	AG<exon>	M2238

Help: for questions, comments, or bug reporting, please visit [NCBI Support Center](#)

Segment 2, the first CDS location:

Query Subject Span(bp) Coverage(%) Overall(%) Exon(%) CDS(%) In-frame(%)

1	mrna(+)	chromosome:GRCh38:11:6998718:7020968:-1(+)	614-21651	100.00	100.00	100.00	0.00	0.00
---	---------	--	-----------	--------	--------	--------	------	------

Graphics | Text

Model 1

Coverage	100.00%	CDS	0.00%	Mismatches and indels	0
Overall	100.00%	In-frame	0.00%	Exons (min/max/ave), bp	147 / 2238 / 889
Exon	100.00%	Primary transcript	2668 bp	Introns (min/max/ave), bp	1154 / 17218 / 9186

mrna (+)

1 chromosome:GRCh38:11:6998718:7020968:-1 (+)

614 21651

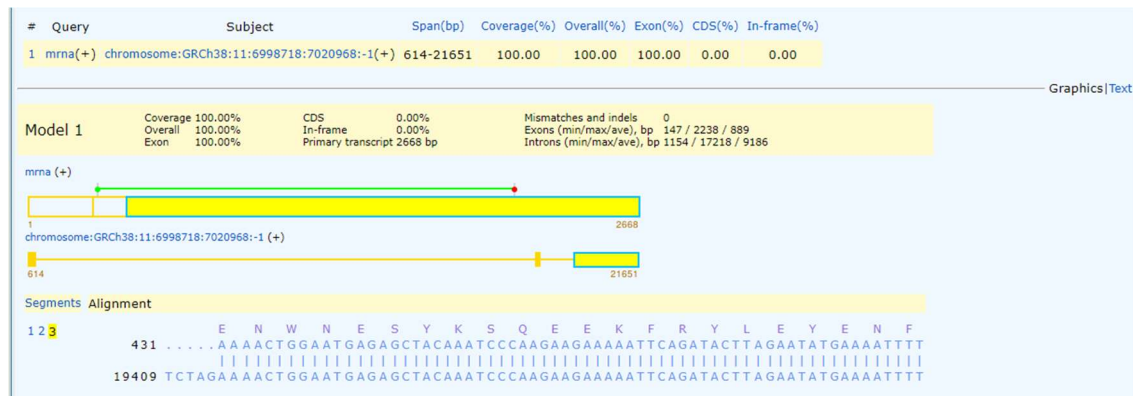
Segments Alignment

1 2 3

```

      M A V T F E D V T I I F T W E
284 . . . . . AAAGCCTGATCTTTGACCAGATGGCAGTAACATTGAAGATGTGACTATTATTTTACATGGGG
      |||
18109 CCTA GAAAGCCTGATCTTTGACCAGATGGCAGTAACATTGAAGATGTGACTATTATTTTACATGGGG
      |||
      E W K F L D S S Q K R L Y R E V M W E N Y T N
349 GAGT GGAATTCCTGGATTCTTCTCAAAAAAGACTCTACAGGGAGGTCATGTGGGAGAACTACACA AATG
      |||
18179 GAGT GGAATTCCTGGATTCTTCTCAAAAAAGACTCTACAGGGAGGTCATGTGGGAGAACTACACA AATG
      |||
      V M S V
419 TCAT GTCAGTAG . . . . .
      |||
18249 TCAT GTCAGTAGGTAA
  
```

Beginning of segment 3, start of 2nd CDS location:



End of segment 3, end of 2nd CDS location:



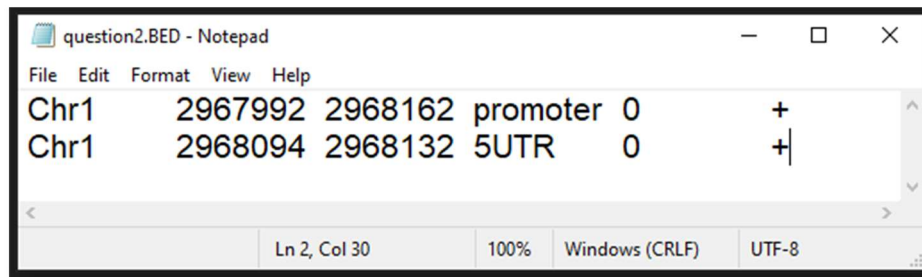
2. Create a BED6 file with 2 lines based on the attached paper (Takenaka_et_al-2015-FEBS_Journal.pdf). Figure 3 shows the location of transcription factor DdIR binding to the promoter region of the ddIR-ddl operon in *Brevibacillus brevis*. The chromosomal location of the ddIR CDS is 2968133..2969623. The zero-based BED6 file should contain the location information of two genomic regions:
 - a. The region bound by the DdIR transcription factor, which we will call the promoter. It is 170 bp in length, begins 140 nucleotides upstream from the start codon, and ends 29 nucleotides downstream from the start codon.
 - Begins **140 nucleotides upstream** from the start codon
 - Start site is 2968133
 - $2968133 - 140 = 2967993$
 - Compensation of 1bp for BED file format
 - $2967993 - 1 = 2967992$
 - Ends **29 nucleotides downstream** from the start codon
 - $2968133 + 29 = 2968162$
 - b. The 5' UTR, noting that the transcription start site, as predicted by BPROM, begins 38 nucleotides upstream from the start codon. The 5' UTR is defined as the region from the transcription start site through the nucleotide that immediately precedes the start codon.
 - Begins **38 nucleotides upstream** from the start codon
 - Start site is 2968133
 - $2968133 - 38 = 2968095$
 - Compensation of 1bp for BED file format

- $2968095 - 1 = 2968094$
- End site **immediately precedes the start codon**
 - $2968133 - 1 = 2968132$

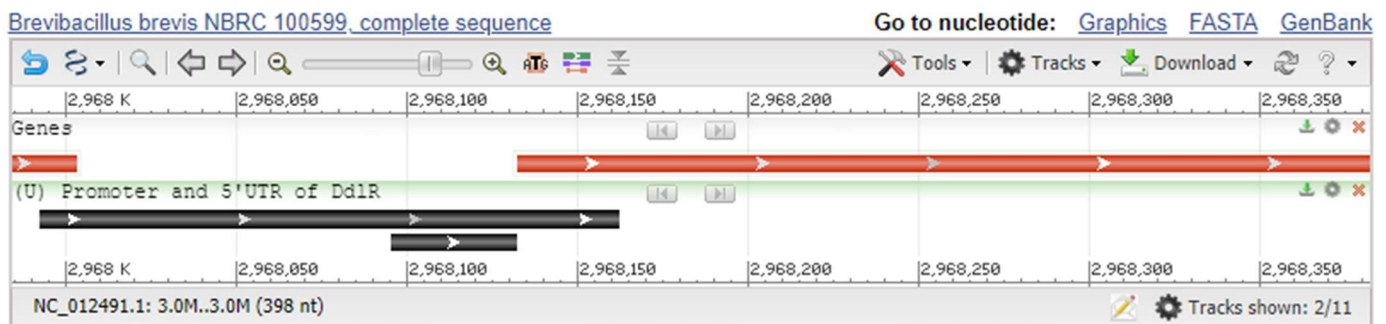
A BED6 file contains the following information separated by a tab: Chromosome, start, end, name, score, and strand. The Chromosome will always be 1 in this case since bacteria only have 1 chromosome. Finally, the direction is + because Figure 3 is pointing to the right.

3. Submit a screenshot of the BED6 from Problem 2. Using the NCBI Genome Browser for *Brevibacillus brevis* NBRC 100599, load your BED6 file. Take a screenshot showing the entire promoter, 5' UTR region, and CDS of DdlR. Be sure to zoom in so that these regions take up a majority of the shot.

Screenshot from Problem2:



Screenshot of BED6 file in NCBI Genome Browser:



4. Use the web-based Biomart in Ensembl to create a dataset and save it as a TSV, CSV, or XLS file. Use the following parameters to make the dataset:

Dataset:

Ensembl Genes 100 (**or the latest version**)

Mouse genes (GRCm38.p6) (**or the latest version**)

Filters:

Chromosome 11

Band E2 only

Transcript count >=7
Limit to genes with RefSeq protein (peptide) IDs only

Attributes:

Default attributes

Add "RefSeq Protein (peptide) ID"

Get all the results, export the results to a file, and submit the file.

- I used the datasets Ensembl Genes 102 (from the archive site) and Mouse Genes (GRCm38.p6). Screenshots of the results are shown below and the results file "mart_export.txt" is submitted.

Archive! Ensembl BioMart | Downloads | Help & Docs | Blog

Search all species...

Dataset 64 / 56305 Genes
Mouse genes (GRCm38.p6)
Filters
Chromosome/scaffold: 11
Band Start: E2
Band End: E2
Transcript count >= 7
With RefSeq peptide ID(s): Only
Attributes
Gene stable ID
Gene stable ID version
Transcript stable ID
Transcript stable ID version
RefSeq peptide ID
Dataset
[None Selected]

Export all results to File CSV Unique results only Go

Email notification to

View 10 rows as HTML Unique results only

Gene stable ID	Gene stable ID version	Transcript stable ID	Transcript stable ID version	RefSeq peptide ID
ENSMUSG00000034566	ENSMUSG00000034566.10	ENSMUST00000043931	ENSMUST00000043931.8	NP_082138
ENSMUSG00000025137	ENSMUSG00000025137.15	ENSMUST00000026129	ENSMUST00000026129.15	NP_001349930
ENSMUSG00000025137	ENSMUSG00000025137.15	ENSMUST00000026129	ENSMUST00000026129.15	NP_077191
ENSMUSG00000025137	ENSMUSG00000025137.15	ENSMUST00000106188	ENSMUST00000106188.3	NP_001334544
ENSMUSG00000020770	ENSMUSG00000020770.13	ENSMUST00000021116	ENSMUST00000021116.11	NP_766157
ENSMUSG00000020770	ENSMUSG00000020770.13	ENSMUST00000106452	ENSMUST00000106452.1	NP_001272935
ENSMUSG00000025138	ENSMUSG00000025138.14	ENSMUST00000080202	ENSMUST00000080202.11	NP_001350368
ENSMUSG00000025138	ENSMUSG00000025138.14	ENSMUST00000080202	ENSMUST00000080202.11	NP_694696
ENSMUSG00000045775	ENSMUSG00000045775.15	ENSMUST00000106532	ENSMUST00000106532.3	NP_001346537
ENSMUSG00000045775	ENSMUSG00000045775.15	ENSMUST00000092445	ENSMUST00000092445.11	NP_001346535

mart_export.txt - Notepad

File Edit Format View Help

Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version, RefSeq peptide ID

ENSMUSG00000034566,ENSMUSG00000034566.10,ENSMUST00000043931,ENSMUST00000043931.8,NP_082138

ENSMUSG00000025137,ENSMUSG00000025137.15,ENSMUST00000026129,ENSMUST00000026129.15,NP_001349930

ENSMUSG00000025137,ENSMUSG00000025137.15,ENSMUST00000026129,ENSMUST00000026129.15,NP_077191

ENSMUSG00000025137,ENSMUSG00000025137.15,ENSMUST00000106188,ENSMUST00000106188.3,NP_001334544

ENSMUSG00000020770,ENSMUSG00000020770.13,ENSMUST00000021116,ENSMUST00000021116.11,NP_766157

ENSMUSG00000020770,ENSMUSG00000020770.13,ENSMUST00000106452,ENSMUST00000106452.1,NP_001272935

ENSMUSG00000025138,ENSMUSG00000025138.14,ENSMUST00000080202,ENSMUST00000080202.11,NP_001350368

ENSMUSG00000025138,ENSMUSG00000025138.14,ENSMUST00000080202,ENSMUST00000080202.11,NP_694696

ENSMUSG00000045775,ENSMUSG00000045775.15,ENSMUST00000106532,ENSMUST00000106532.3,NP_001346537

ENSMUSG00000045775,ENSMUSG00000045775.15,ENSMUST00000092445,ENSMUST00000092445.11,NP_001346535

ENSMUSG00000045775,ENSMUSG00000045775.15,ENSMUST00000092445,ENSMUST00000092445.11,NP_001074403

ENSMUSG00000057948,ENSMUSG00000057948.12,ENSMUST00000075036,ENSMUST00000075036.8,NP_001009573

ENSMUSG00000020773,ENSMUSG00000020773.11,ENSMUST00000106441,ENSMUST00000106441.7,NP_766158

ENSMUSG00000020773,ENSMUSG00000020773.11,ENSMUST00000021120,ENSMUST00000021120.5,NP_001192010

ENSMUSG00000020776,ENSMUSG00000020776.18,ENSMUST00000103031,ENSMUST00000103031.7,NP_001351008

ENSMUSG00000020776,ENSMUSG00000020776.18,ENSMUST00000103031,ENSMUST00000103031.7,NP_001351007