10.2 BEDTools Applications – Tutorial

At the end of this tutorial you should be able to:

- Use BEDTools subtract, closest, and sort
- Use and interpret BED files containing Ensembl IDs
- Analyze the results of BEDTools analysis in R

How to complete this tutorial

- Go through each question in order and complete any tasks that are described in the question.
- As you complete the questions, mark your answer to each question.
- Questions will be either:
    - multiple-choice questions that require you to provide either a single answer or to select multiple answers
    - questions that require a short text answer
- Open the associated quiz on Quercus and enter your answers to each question to verify that you completed the tutorial questions correctly.
- Alternatively, open the Quercus quiz when you start the tutorial and verify your answers as you complete the tutorial. **Note that there may be some information that is in this file that is not in the Quercus quiz!**
- The answers along with the complete set of commands you should use throughout the tutorial will be released at the end of the week.

IMPORTANT NOTE

This tutorial will be completed partially in the command line and partially in R. To complete this tutorial, go through the steps to answer questions 1 through 10 using the command line. Once complete, open the R Markdown file: `Tutorial.10.2.Rmd`
In contrast to previous weeks, Otter Grader will not be used to mark the R portion. Instead, follow the directions in the R Markdown file to complete the described analysis, and then answer the associated quiz questions in Quercus (questions 11-15).

Before you begin

- Open a new terminal session from your JupyterHub (New > Terminal)
- Set the PWD to `/home/jovyan/Week.10/10.2.BEDTools/Tutorial.10.2`

Background

In this tutorial we will investigate the genes regulated by the transcription factor HNF4A (Ensembl Gene ID: ENSG00000101076). Using HNF4A ChIP-seq data from 3 different experiments, we will identify human promoters that are likely bound by HNF4A, predict the genes these promoters regulate, and use the expression levels of these genes across tissue types to predict the tissue in which HNF4A has the strongest effect on the regulation of transcription.

In the `Tutorial.10.2` directory, you will find the following files (All files have coordinates corresponding to the hg38 human genome assembly):

File	Description
`HNF4A_ChIP_exp1.bed`	Coordinates of HNF4A binding sites identified by ChIP-seq with a score = 1000 (ENCFF072CXB)
`HNF4A_ChIP_exp1.bed`	Coordinates of HNF4A binding sites identified by ChIP-seq with a score = 1000 (ENCFF837QHJ)
`HNF4A_ChIP_exp1.bed`	Coordinates of HNF4A binding sites identified by ChIP-seq with a score = 1000 (ENCFF905JAC)
`ENCODE_promoters.bed`	Coordinates of candidate human promoters identified by ENCODE*
`hg38_transcripts.bed`	Coordinates of human transcripts in the hg38 genome**
`GTEx_expression_data.txt`	Median expression levels (TPM) for 56,200 genes in 54 distinct tissue types***

All BED files were data downloaded from the UCSC Genome Browser (https://genome.ucsc.edu/) and have coordinates corresponding to the hg38 human genome assembly.
* `ENCODE_promoters.bed` was generated from ENCODE_cCREs.bed by filtering for "PLS"
** `hg38_transcripts.bed` was generated from GENCODE V8 knownGenes by filtering for "basic"
*** GTEx expression data was downloaded directly from GTEx (https://gtexportal.org/). Data is from GTEx Analysis V8.

10.2.1: BEDTools Subtract

Question 1
Intersect the HNF4A binding sites in ChIP-seq experiments 1 and 2 (`HNF4A_ChIP_exp1.bed` & `HNF4A_ChIP_exp2.bed`) to get the regions that are identified in both experiments. Output the result to a file named `HNF4A_ChIP_exp12.bed`.
Fill in the command below to match the command you used (do not include what is already there!).

```
_________ > HNF4A_ChIP_exp12.bed
```

Question 2
Intersect the HNF4A binding site regions that are in both experiments 1 and 2 (`HNF4A_ChIP_exp12.bed`) with the binding sites from ChIP-seq experiment 3 (`HNF4A_ChIP_exp3.bed`) to get the regions that overlap in all three experiments. Output the result to a file named `HNF4A_ChIP_exp123.bed`.
How many binding site intervals are there in `HNF4A_ChIP_exp123.bed`?

Question 3
The HNF4A ChIP-seq experiment 1 identified the fewest HNF4A binding sites of the three ChIP-seq experiments. How many binding site regions identified in ChIP-seq experiment 1 are not in ChIP-seq experiment 2 or ChIP-seq experiment 3?

10.2.2: Ensembl IDs & Transcript Isoforms

Question 4
Which of the following statements are true about Ensembl IDs?

a. Human transcripts have IDs that start with "ENST".
b. Human genes have IDs that start with "ENMUSG".
c. A gene ID will always be associated with only one transcript ID.
d. IDs can end with version numbers

Question 5
Which of the following statements are true about transcript isoforms?

a. Transcript isoforms associated with different genes cannot overlap
b. Transcript isoforms associated with same gene generally overlap
c. Transcript isoforms can differ in length
d. BEDTools commands used with the set of transcript isoforms can return multiple results

10.2.3: BEDTools Closest

Question 6
Intersect the HNF4A binding sites that are present in all three ChIP-seq experiments with the set of ENCODE promoters (`ENCODE_promoters.bed`) to get the UNIQUE promoters that are overlapped by one or more binding site. Output the result to a file named `HNF4A_ChIP_exp123_promoters.bed`.
Fill in the command below to match the command you used (do not include what is already there!).

```
bedtools intersect _________ > HNF4A_ChIP_exp123_promoters.bed
```

Question 7
Sort the file `HNF4A_ChIP_exp123_promoters.bed` and save the result to a file called `HNF4A_ChIP_exp123_promoters_sorted.bed`.
Fill in the command below to match the command you used (do not include what is already there!).

```
_________ > HNF4A_ChIP_exp123_promoters_sorted.bed
```

Question 8
Find the transcript(s) (`hg38_transcripts.bed`) that is(are) closest to each promoter and save the output to a file called `HNF4A_promoters_closest_transcript.bed`. Remember to use the necessary options so that `bedtools closest` will return the closest transcript downstream of the promoter, as well as the distance between the promoter and the transcript.
Fill in the command below to match the command you used (do not include what is already there!).

```
bedtools _________ > HNF4A_promoters_closest_transcript.bed
```

Question 9
Count the number of lines in the files `HNF4A_ChIP_exp123_promoters_sorted.bed` and `HNF4A_promoters_closest_transcript.bed`. How do you explain the discrepancy in the number of intervals in each file?
(**Hint:** View all the lines in `HNF4A_promoters_closest_transcript.bed` that contain the gene ID "ENSG00000250722".)

a. Genes can have multiple transcript isoforms, so if a promoter is equally close to all isoforms, one line will be returned for each
b. `bedtools closest` returns the 3 closest intervals by default
c. In the human genome, many different genes start at the exact same base in the genome, so if a promoter is equally close to all genes, one line will be returned for each
d. `bedtools closest` is known to make errors

Question 10

Use the `cut` command to get the columns containing the promoter ID, the gene ID, the gene type and the distance between the gene and the promoter (columns 4, 13, 14, & 15) from `HNF4A_promoters_closest_transcript.bed`, get all the unique lines (use commands we learned in the command line week), and save it to a file called `HNF4A_regulated_genes.bed`. Your file should have 125 lines.

Before moving onto the next section of the tutorial, confirm that your file has the correct data. Which of the following matches the last line in your file?

a. `EH38E1321662 ENSG00000272510 nonCoding -18982`
b. `EH38E2737681 ENSG00000176058 coding -20`
c. `EH38E1902956 ENSG00000265758 nonCoding -29167`
d. `EH38E2737022 ENSG00000169692 coding -6312`

10.2.4: Analysis of Results in R

To answer the following questions, you will need to open the file `Tutorial.10.2.Rmd` and perform the analysis described in the file. These questions in also in the .Rmd file for your reference.

Question 11

How many of the genes in `HNFA_regulated_genes` are **coding** genes within 300 base pairs of the closest promoter?

Question 12

In which tissue type is HNF4A most highly expressed? (Ensembl Gene ID: ENSG00000101076)

Type the tissue name as it is written in the column names of the `GTEx_data` data frame. For example: Adrenal.Gland

Question 13

Based on the heatmap of tissue expression, in which tissue are most genes most highly expressed? Type the tissue name as it is written in the column names of the `GTEx_data` data frame. For example: Adrenal.Gland

Question 14

HNF4A is expressed the least in the putamen ("Brain.Putamen") (0.00413535). Perform a t-test to determine if the set of genes likely regulated by HNF4A have a different expression distribution in liver vs. putamen tissues. Are the distributions significantly different (p-value < 0.05)?

a. Yes, they are significantly different
b. No, they are not significantly different

Question 15

Examine the plots of liver tissue vs. brain putamen expression values for HNF4A regulated genes and liver tissue vs transverse colon expression values for HNF4A regulated genes. Which of the two tissues has more of these genes expressed at a higher level than they are expressed in liver tissue?

a. Brain.Putamen
b. Colon.Transverse