10.1 BEDTools – Tutorial

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

- Explain the concept of DHSs
- Relate DHSs to DNA accessibility, transcription, and tissue specificity
- Recognize and understand the BED file format
- Use BEDTools intersect

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.

Before you begin

- Open a new terminal session from your JupyterHub (New > Terminal)
- Set the PWD to `/home/jovyan/Week.10/10.1.BEDTools/Tutorial.10.1`

Background

Section 10.1.3 of this tutorial will use data obtained from the paper:

> Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. (2011) *Genome Biology* **12:** R43

During development, the programming of different cell fates requires many epigenetic changes. An epigenetic change modifies gene expression without altering the nucleotide sequence. One such epigenetic change is the alteration of chromatin accessibility. As we learned in lecture, accessible chromatin can be bound by transcription factors, meaning that regions of chromatin that are open and closed determine which genes are expressed. This type of regulation is very important during development.

In Thomas *et al.*, the authors interrogated chromatin accessibility in the developing *Drosophila melanogaster* embryo. There are 17 stages of embryogenesis in *Drosophila*. DNase-seq was performed at stages 5, 9, 10, 11, and 14, in which the developing fly is undergoing gastrulation and then tissue differentiation. This resulted in BED files of accessible regions at each stage.

One region of the genome particularly relevant to development is the Hox locus. Hox genes are a set of transcription factors involved in proper developmental patterning. Each gene is required in a different part of the embryo (from head to abdomen) for proper development. During development Hox genes are expressed at a low level in early stages and expression increases as development progresses.

In the `Tutorial.10.1` directory, you will find the following files:

File	Description
`stage.5.DHS.bed`	Coordinates of DHSs in stage 5 *Drosophila* embryo
`stage.9.DHS.bed`	Coordinates of DHSs in stage 9 *Drosophila* embryo
`stage.11.DHS.bed`	Coordinates of DHSs in stage 11 *Drosophila* embryo
`hox.gene.promoters.bed`	Coordinates of promoter & surrounding region for 8 *Drosophila* Hox genes*

Data Sources:

* Coordinates generated using information from the UCSC Genome Browser (https://genome.ucsc.edu/)

10.1.1: Functional Genomics Assays

Question 1 (SELECT ALL THAT APPLY)

Which of the following statements about DNase I hypersensitive sites (DHSs) are true?

a. DHSs are only found in heterochromatin
b. DHSs are sensitive to cleavage of DNase I
c. DHSs mark sites of transcriptional regulation
d. DHSs are found at enhancers, but not promoters
e. DHSs are always present at the same locations in all cell types
f. DHSs can be identified using DNase-seq

Question 2

The protein ESR1 binds to DNA to regulate gene expression. What assay could you use to identify the genomic locations where ESR1 binds to DNA?

a. DNase-seq
b. CLIP-seq
c. ChIP-seq
d. ATAC-seq

Question 3 (SELECT ALL THAT APPLY)

The gene GATA4 is expressed in the artery, heart, liver, ovary, pancreas, stomach, and testis. Of the tissues listed below, in which would you expect to find a DHS at the GATA4 promoter?

a. Artery
b. Brain
c. Colon
d. Pancreas
e. Stomach
f. Uterus

10.1.2: BED Files

Question 4

Which of the following CANNOT be stored in a BED file?

a. The locations of all exons in the mouse genome
b. Binding site coordinates of the transcription factor SOX2 in K562 cells
c. DNA sequences of all human enhancers
d. DNase I hypersensitive sites in lymphocytes

Question 5 (SELECT ALL THAT APPLY)
Which columns/fields in a bed file are required?

a. Chromosome
b. Name
c. Strand
d. Stop position
e. Start position
f. Score

Question 6
Which of the following statements about BED files is FALSE?

a. Strand information can be included without including columns for name and score
b. You can use place holders for score or strand (like 0s or periods)
c. Columns must be included in a specific order
d. BED files can contain more than 6 columns

10.1.3: BEDTools Intersect

Question 7
If you are not currently there, change your directory so that the PWD is `/home/jovyan/Week.10/10.1.BLAST/Tutorial.10.1`. Install `bedtools` using `conda`.
In this directory you will find the files containing the DHSs for stages 5, 9, and 11 in *Drosophila* development, and the file containing the Hox gene promoter coordinates.

Use `bedtools intersect` to determine how many accessible sites overlap between stage 5 and stage 9 (use stage 5 for `-a`) and output the results to a file called `stage.5.and.9.bed`.
Fill in the command below to match the command you used (do not include what is already there!).

```
_________ > stage.5.and.9.bed
```

Question 8
How many overlapping intervals are there between stage 5 and 9?

Question 9
Use `bedtools intersect` to determine how many accessible sites overlap between stage 5 and stage 11. Which stage has more overlapping intervals with stage 5? (Enter 9 or 11 in the text box.)

Question 10

Would you expect stage 9 and 11 to have more overlapping intervals than stage 5 and 11? Use `bedtools intersect` to find out. How many overlapping intervals are there between stages 9 and 11?

Question 11

Use the `-u` option with `bedtools intersect` to determine which **unique** Hox gene promoter regions are accessible in stage 11. Output the result to standard output (not to a file).
What command did you use?

Question 12

Use the same type of command you used in question 8 to determine which unique Hox gene promoter regions are accessible in stage 5 and 9.
Which of the following statements is true?

a. None of the Hox gene promoters are accessible in stage 5
b. The most Hox gene promoters are accessible in stage 9
c. In stage 11, 4 Hox gene promoters are accessible
d. All the Hox gene promoters that are accessible in stage 11 are also accessible in stage 5

Question 13

Consider the following information:
Stage 5 embryos were collected at 2 hours 10 minutes, stage 9 at 3 hours 20 minutes, and stage 11 at 5 hours 40 minutes.
Based on your results from questions 9 and 10, which gene do you think is not expressed until after 3 and half hours of embryo development?

a. Ubx
b. lab
c. abd-B
d. None of them are expressed until after 5 hours