Assignment 10

How to complete this assignment

- Read through the background information
- Go through each question in order and complete any tasks that are described in the question. **Note that some information in this file may not be available in the Quercus quiz.**
- 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 assignment quiz on Quercus and enter your answers to each question.
- You may only submit this quiz once, so be sure you answer all questions before submitting the quiz.

IMPORTANT NOTE

This assignment (like Tutorial 10.2) will be completed partially in the command line and partially in R. To complete this assignment, go through the steps to answer questions 1 through 7 using the command line. Once complete, open the R Markdown file: `Assignment.10.Rmd`
Follow the directions in the R Markdown file to complete the described analysis, and then answer the associated quiz questions (questions 8-14).

Before you begin

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

Mark breakdown

Part 1 – 7 questions – 9 marks
Part 2 – 7 questions – 11 marks

BACKGROUND

In this assignment you will be working with data from the paper:

> Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. (2016) *Nature Genetics* **48:** 510-518

Genome Wide Association (GWA) studies identify variants associated with a phenotype by genotyping a group of individuals who display the phenotype (case) and a group of individuals who do not display the phenotype (controls). The genotyping data is then used to identify variants that are associated with the phenotype by comparing allele frequencies in cases vs. controls. Based on the statistical test used to identify variants, each variant has a p-value quantifying the significance of the association.

Ellinghaus *et al.* performed a GWA study to identify variants that are associated with five chronic inflammatory diseases (CIDs): ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, and ulcerative colitis. The variants associated with CIDs are in the BED file `CID.variants.bed`. Each variant has a reference SNP (rs) identifier and a score. The scores in the file are -log10(p-values), thus the higher the score, the more significant the variant.

In this assignment you will investigate the locations of these variants in the genome and investigate the genes impacted by these variants.

File	Description
`CID_variants.bed`	Coordinates of variants identified to be associated with the five CIDs
`ENCODE_promoters.bed`	Coordinates of candidate human promoters identified by ENCODE*
`hg38_transcripts.bed`	Coordinates of human transcripts in the hg38 genome**
`hg38_CDSs.bed`	Coordinates of human coding sequences in hg38 genome
`GTEx_expression_data.txt`	Median expression levels (TPM) for 56,200 genes in 54 distinct tissue types

Data Sources

The GWAS Catalog was downloaded from UCSC Genome Browser (https://genome.ucsc.edu/) and filtered for variants from Ellinghaus *et al.* to generate the set of CID variants.
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 knownGenes by filtering for "basic"
GTEx expression data was downloaded directly from GTEx (https://gtexportal.org/). Data is from GTEx Analysis V8.

PART 1: Locations of Variants in the Genome (9 marks)

Question 1

(1 mark)

Identify the CID associated variants that are within coding sequence (`hg38_CDSs.bed`). Assign the output to a file named `CID_variants_in_CDSs.bed`. Use the `-wb` option as shown below. This will output the overlapping CDS information into the output, similarly to `bedtools closest`.

Fill in the command below to match the command you used (do not include what is already there!).

```
bedtools intersect -wb _________ > CID_variants_in_CDSs.bed
```

Question 2

(2 marks)

Identify the variant with the highest score that is contained within a coding sequence (`CID_variants_in_CDSs.bed`) and get the Ensembl ID of the gene it is contained within (the gene ID will be in column 13). Search the Ensembl ID in the UniProt search bar and filter the results by "Reviewed" on the right side of the results and view the result. Which of the following is the gene involved in:

a. alternative splicing regulation
b. actin organization
c. blood vessel development
d. gastrointestinal immunity

Question 3

(1 mark)

Use `bedtools intersect` to find the transcripts (`hg38_transcripts.bed`) that overlap each of the CID variants. For this task the command is provided for you:

```
bedtools intersect -wb -a CID_variants.bed -b
hg38_transcripts.bed > CID_variants_in_transcripts.bed
```

As in question 1, the `-wb` option adds the information from `hg38_genes.bed` so that the results will contain the transcript information, like `bedtools closest`.

View the lines in CID_variants_in_transcripts.bed that show which transcripts are overlapped by the SNP rs2476601. Which of the following explains the 7 lines containing overlaps with rs2476601?

a. SNP rs2476601 overlaps 3 genes, one with 2 isoforms, one with 1 isoform, and one with 4 isoforms
b. SNP rs2476601 overlaps 2 genes, one with 1 isoform and one with 6 isoforms
c. SNP rs2476601 overlaps 3 genes, two with 3 isoforms and one with 1 isoform
d. SNP rs2476601 overlaps 2 genes, one with 2 isoforms and one with 5 isoforms

Question 4
(1 mark)
Use a BEDTools command to generate a file called `CID_variants_not_in_transcript.bed` that contains CID variants that are NOT overlapped by any transcripts. How many variants are not overlapped by transcripts?

Question 5
(1 mark)
Use a BEDTools command to generate a file called `CID_variants_not_in_transcript_in_promoter.bed` that contains CID variants that are NOT overlapped by any transcripts, but ARE overlapped by an ENCODE candidate promoter. How many variants are in a promoter, but not a transcript?

Question 6
(1 mark)
Extract columns 4, 5, 13, 14, and 15 from `CID_variants_in_transcripts.bed`, sort the results, get the unique lines, and save it to a file called `CID_variants_and_genes.txt`. This file will now contain only row for each gene overlapped by each variant as all of the isoforms will be collapsed into one line.
How many lines are in this file?

Question 7
(2 marks)
In columns 3 and 4 of `CID_variants_and_genes.txt` are the Ensembl gene ID and the type of gene ("coding", "nonCoding", or "pseudo"). Identify the Ensembl gene ID of the coding gene with the most overlapping SNPs. Search the gene ID on UniProt. Which of the following is the gene a component of?

a. The CCR4-NOT complex
b. The interleukin-23 receptor
c. The spliceosome
d. The NEXT complex

PART 2: Analysis of Genes Containing CID Variants (11 marks)

To answer the following questions, you will need to open the file `Assignment.10.Rmd` and perform the analysis described in the file.

Question 8

(1 mark)
Based on the bar plot displaying the count of coding, non-coding, and pseudogenes that are overlapped by CID variant, which of the three gene types is the least common in the data frame?

a. Coding genes
b. Non-coding genes
c. Pseudogenes

Question 9

(1 mark)
Which tissue has the highest mean expression value for the genes that contain CID variants?
Type the tissue name as it is written in the column names of the `GTEx_data` data frame.
For example: Adrenal.Gland

Question 10

(2 marks)
Based on the heatmap of the CID variant genes expression across tissue types, which tissue has the largest number of genes that are expressed most highly in that tissue?
Type the tissue name as it is written in the column names of the `GTEx_data` data frame.
For example: Adrenal.Gland

Question 11

(2 marks)
How many of the CID variants with a score above the median variant score overlap a coding gene?

Question 12

(2 marks)
Based on the heatmap of tissue expression, now that the data set is reduced to variants with high scores in coding genes, which tissue has the largest number of genes that are expressed most highly in that tissue?
Type the tissue name as it is written in the column names of the `GTEx_data` data frame.
For example: Adrenal.Gland

Question 13
(1 mark)
What percentage of the **coding genes with high scoring variants** are **inflammatory response genes**?

a. 4%
b. 17%
c. 39%
d. 57%

Question 14
(2 marks)
In questions 10 & 12 you should have identified two different tissues with the largest number of genes that are expressed most highly. We will refer to these two tissues as tissue A (tissue from question 10) & tissue B (tissue from question 12).

Use a t-test to determine if the expression patterns of the inflammatory response coding genes with high scoring variants are significantly different in these two tissues. Is there a significant difference?

a. Yes, there is a significant difference
b. No, there is not a significant difference