Project description - 2021sp

Objective:
The goal of the project is to develop your problem-solving skills through a real bioinformatics application. Two components are very important in bioinformatics research, self-learning and interdisciplinary collaboration. You will experience both by working on this project. You will find that you have to self-learn materials and bioinformatics tools that are not covered in class. While this may be challenging at the beginning, it will benefit you greatly in the future. Previous students in this class have found this experience very rewarding.

Team:
Each team should consist of 3-4 people across disciplines. Please sign up your team at **here (https://docs.google.com/spreadsheets/d/1S4YNf-XhI5NsUEHW2Ojv_qwjDHxWuW07ApBJC97FuE4/edit#gid=0)** by March 1.

Project Description:

Each team will have a 15-min presentation and submit a report. The presentation will be in the last week of class and all team members need participate. The report is due by the end of the exam week.

Scope and Grading:
- For a 3-person team, you need finish the whole analysis for at least two pairs of cells or tissues
- For a 4-person team, you need finish the whole analysis for at least three pairs of cells or tissues
- Your score is based on both the quality and the comprehensiveness of your analysis. That is, with the same quality, the more pairs you exceed your target, the higher your score will be. Also, before you advance to do more pairs than your assigned number, please first complete the ATAC-seq analysis.

This project will study the differentiation process in hemopoiesis, i.e. the formation of blood cellular components. In particular, we will look into the difference in terms of gene expression (measured by RNA-seq data) and chromatin accessibility (measured by ATAC-seq data) in several cell lines involved in this process, and integrate these two pieces of information together. You will use some statistical and computational tools to identify differential patterns in both gene expression and chromatin accessibility (using Limma-voom or DEseq2), do GO-term analysis to find functions of differentially expressed genes, and perform clustering analysis to find genes with similar patterns.

The datasets are provided by Dr. Hardison's lab. Four cell lines (HSC, CMP, CFUE, and ERY) in the differentiation process are provided. The relationship between these cells can be found **here (https://psu.instructure.com/courses/2098115/files/119108918/download?wrap=1)** ⬇ **(https://psu.instructure.com/courses/2098115/files/119108918/download?download_frd=1)**.

For each cell line, there are two biological replicates. For each replicate sample, the RNA-seq data were generated from two experimental protocols. You will use Limma voom and DEseq2 to perform differentiation analysis, and compare results.

Questions to answer:

We are interested in the following questions:

About RNA-seq data:

1. What genes are differentially expressed across each pair of cell lines?
2. What are the functions of the genes with differential expression patterns?
3. How consistent are the results between DEseq2 and limma voom?
4. Construct a hierarchical tree using all the RNA-seq data, and perform clustering analysis. Describe the relationship based on your results.

About ATAC-seq data:

5. What regions have differential chromatin patterns across each pair of cell lines? What are the genes near these regions?
6. How are differential chromatin patterns related to the expression patterns of nearby genes?
7. What are the functions of the genes with differential chromatin patterns?
8. Construct a hierarchical tree using all the ATAC-seq data and use clustering analysis to explore the pattern of cell-line specified genes. Do you get the same structure as the tree from RNA-seq data?

Scope and grading

Ideally, we hope one team can answer all the above questions by doing pairwise comparison for all the cell lines. But due to time limit, you may not achieve all of them for all the cell lines. So I give each group an order of the cell lines to proceed. Please follow the order when you proceed.

- Start with the first pair of cell lines that are assigned to you.
- For your 2nd pair and beyond, you only need use the software that performs better in the first pair to do the analysis.
- You may go beyond the questions above. I include a paper (**here** (http://www.nature.com/nature/journal/v534/n7609/pdf/nature18606.pdf)) that uses RNA-seq and ATAC-seq. You may check to see what other analyses you can do with the data.

The assignment list is **here** (https://docs.google.com/spreadsheets/d/1S4YNf-XhI5NsUEHW2Ojv_qwjDHxWuW07ApBJC97FuE4/edit#gid=0) .

Download Data

1. Data:

You need both RNA-seq and ATAC-seq data for your analysis.

RNA-seq data is for analyzing gene expression patterns, and ATAC-seq is for analyzing chromatin patterns. The RNA-seq data was prepared by two experimental methods, TotalScript and ScriptSeq. Two replicates are available for each assay. You only need use ScriptSeq data.

2. The ID list (**ENCODE ID file** (https://psu.instructure.com/courses/2098115/files/119108875

/download?wrap=1) ↓ (https://psu.instructure.com/courses/2098115/files/119108875/download?download_frd=1)) consists of the IDs of the data files for each cell line. You need use these IDs to download the data from the ENCODE portal.
The first column in this list consists of .bam files and the second column consists of .tsv files (RNA-seq) or .bigBed files (ATAC-seq). The two rows for each assay are two replicates. You need both replicates to run differentiation analysis.
You may use the .bam file or the processed file (.tsv or .bigBed) for your project.
The .tsv file consists of RNA-seq expression levels extracted by RSEM and the .bigBed file consists of ATAC-seq peaks identified by HOMER.
(Alternatively, you may work with .bam file. Be aware that more data processing steps are needed to process .bam file and you need figure out how to do these additional steps yourself. -- I do not recommend you to do this for this project unless you are very proficient)

3. Download the data from ENCODE portal by following the format below
https://www.encodeproject.org/files/<id>/@@download/<id>.<filetag>
For example, to download ENCFF247FEJ.tsv file, use
https://www.encodeproject.org/files/ENCFF247FEJ/@@download/ENCFF247FEJ.tsv