­Spring 2017 STAT115/215 BIO/BST282

Due: 3/19/2017 midnight

**HOMEWORK 3: scRNA-seq, ChIP-seq, epigenetics, Python**

**Part I. Single cell RNA-seq**

For this exercise, we will be analyzing a single cell RNA-Seq dataset of mouse brain (Cortex, hippocampus, and subventricular zone) from the 10X Genomics platform. The full dataset consists of nearly 1.3M single cells, but for this assignment, we’ll consider a random subset of these cells. A full description of the data is available: <https://support.10xgenomics.com/single-cell/datasets/1M_neurons>.

1. Describe the composition of the raw dataset (i.e. number of genes, number of samples, and dropout rate). After filtering against weakly detected cells and lowly expressed genes using reasonable parameters, how do these summary statistics change?
2. What proportion of the counts from your filtered dataset map to mitochondrial genes? Compare these values to other mitochondrial read distributions in the PBMC dataset shown in lab and in the Seurat vignette. If you determine that mitochondrial reads represent a source of unwanted variation in the filtered data, use techniques discussed in lab to remove this unwanted source of variation.
3. Perform linear dimensionality reduction (PCA) on the filtered dataset. Provide summary plots, statistics, and tables to show A) how many PCs are statistically significant, B) which genes contribute to which principle components, and C) how much variability is explained in these top components. Compare the variability in the top PCs to other scRNA-Seq datasets.

a. **[Graduate students]** Determine which PCs are heavily weighted by cell cycle genes. Provide plots and other quantitative arguments to support your argument. Assuming that cell cycle is a source of unwanted variation in the data, how could you correct for it?

1. Perform a non-linear dimensionality reduction (tSNE) using the principle components as features. Visualize the cells and their corresponding tSNE coordinates and comment on the number of cell clusters that become apparent from the visualization. Are the number of clusters that form robust when rerunning tSNE?
2. Using the principle components as features, perform a clustering algorithm of your choice (either supervised or unsupervised) to uncover potential subpopulations in this data. How many cells become assigned to each group? Visualize these clusters on the tSNE graph.
3. Using differential expression analyses between clusters, identify putative biomarkers for each cell subpopulation. Visualize the gene expression values of these potential markers on your tSNE coordinates. Comment on any cluster heterogeneity or rare subpopulation characteristics based on these gene expression values.

a. **[Graduate students]** Based on the data-driven characteristics of your cell clusters, provide a putative biological annotation (e.g. hippocampal cells) to the identified populations. This paper may serve as a good resource: <http://science.sciencemag.org/content/347/6226/1138> as well as the Allen Brain Atlas: <http://mouse.brain-map.org>

1. Seurat is one of many analysis packages for scRNA-Seq. As many of these frameworks are very young, what feedback do you have to improve the user experience of single cell analyses?

**Part II. ChIP-seq and Epigenetics**

For this section, we will analyze ChIP-Seq data analysis with MACS and consider an integratie model--BETA. This provides a highly integrative methodology which includes ChIP-seq, epigenetics, motif finding, target gene identification, and functional annotation.

RMA and LIMMA are available in R Bioconductor. MACS and BETA are available on Odyssey. Make sure to read the MACS README and BETA *Nature Protocol* paper:

<https://pypi.python.org/pypi/MACS2/2.0.10.09132012>

<http://www.nature.com/nprot/journal/v8/n12/full/nprot.2013.150.html>

You can use the following commands to load them:

module load bio/MACS-2.0.10.20130306

module load centos6/BETA\_1.0.5

**The data required for this question are listed below, and available on Odyssey folder “/n/stat115/hws/3”.**

**Chip-seq data:**

TET1 is the enzyme that promote DNA demethylation by converting 5-methylcytosine to 5-hydroxymethylcytosine (5hmC). Theoretically this process should activate gene expression, but recent literature and data might shed lights on a more complex regulation mechanism. Here we provide TET1 ChIP-seq data in mouse embryonic stem cell:

mES\_TET1.bam (BAM file for the ChIP-seq data for TET1, GSM706672)

mES\_Control.bam (BAM file for Control data, GSM706673)

Please note: in labs, you were taught how to run these in Mm9 genome assembly, but for HW3 the data we will ask you to try mapping it to the Mm10 genome assembly. Also the BAM files are mapped to Mm10.

**Expression data:**

Folder “ctrl\_vs\_KD” holds raw data for microarray. The expression data are from two conditions, which are wild type mouse embryonic stem cell (ES) and ES with TET1 knockdown by siRNA (a RNA interference technique). The phenotype of expression data:

GSM84606 Control ES cells 96hr, rep1

GSM846064 Control ES cells 96hr, rep2

GSM846065 Tet1-KD ES cells siRNA #1 96hr, rep1

GSM846066 Tet1-KD ES cells siRNA #1 96hr, rep2

GSM846067 Tet1-KD ES cells siRNA #2 96hr, rep1

GSM846068 Tet1-KD ES cells siRNA #2 96hr, rep2

1. Identify differentially expressed transcripts between wild type vs. knockdown (in RefSeq) using LIMMA.

Hint: For the annotation cdf file, you can download from <http://mbni.org/customcdf/21.0.0/refseq.download/mouse4302mmrefseqcdf_21.0.0.tar.gz> and install it in R.

1. Since aligning the full .fastq file would be time-intensive, a subset of ~1M raw reads from the wild type TET1 ChIP-seq were extracted and saved as “TET1\_mES.1M.fastq.” Align this fastq file using BWA (and produce a corresponding .bam file. The mm10 index library for BWA is available under the **“/n/stat115/hws/3”** directory in Odyssey
2. In ChIP-Seq experiments, when sequencing library preparation involves a PCR amplification step, it is common to observe multiple reads where identical nucleotide sequences are disproportionally represented in the final results. Thus, it may be necessary to perform a duplicate read removal step, which flags identical reads and subsequently removes them from the dataset. Run this separate for treatment and control samples (macs2 filterdup).

**Hint:** You may find the manual of MACS2 (<https://pypi.python.org/pypi/MACS2/2.0.10.09132012)> useful .   
The peak calling takes about 5~10 minutes to run this on Odyssey.

1. In ChIP-Seq analysis, there may be bias in results when the number of reads in treatment and control are different. In assuming, in treatment has ~8 million reads and control has ~11 million reads (after removing the duplicated reads). One solution for correcting the bias is to sample down the control to the same number of reads as the treatment. We can do it with MACS2. Sample down control data using MACS2 (macs2 randsample). After removing the duplications in treatment and control, and sampling down the control reads, we are ready to call the peaks using MACS2.

**Hint:** Sample down the control to the same number of reads as treatment (8757629 reads). Read the manual of MACS2. It takes about 1~2 minutes to run this on Odyssey. It takes about 45~90 minutes to run this on Odyssey. Run macs on Odyssey with enough memory (2GB) by using “srun --mem=2048 -p interact –pty”

f) Integrate TET1 binding data with differential expression data to study TET1 regulation function.

Does TET1 function as a gene expression activator, repressor, or both?

What motifs are enriched in the TET1 ChIP-seq peaks?

**Hint:** Read the BETA *Nature Protocol* paper. You need to have an additional column from LIMMA to run BETA. In addition, BETA cannot run on PC. If serious error reports in server, we recommend to run BETA in CistromeAP (<http://cistrome.org/ap/)> where you can find BETA basic/plus/minus in “Integrative Analysis”. Please take care in website about setting species and columns for gene id, fold-change, and FDR/adj.p.value.

g) **For graduate students:** Transcription factor always works as a complex which called cofactors. Motif analysis from BETA suggests that some factors’ motif are there. From the comparison of differential expression between knockdown and wild type, could you find which could be potential cofactors?

**Hint:** If this TF is cooperative with TET1, it will need to be expressed much higher at the time when TET1 is on the DNA, and motif could be probably found there. The list of motif factors in Cistrome database can be found here: /n/stat115/hws/3/motif\_factor.txt.

h) Perform gene ontology enrichment analysis on the output genes from BETA using DAVID. What genes does TET1 regulate (enriched GO terms or pathways)? What are the p-values and FDR for the most significant enrichments?

**Part III: Python Coding**

From UCSC download page (http://hgdownload.soe.ucsc.edu/downloads.html), download the mouse RefSeq (the expression and ChIP-seq are both mouse data) annotation table (find the file refGene.txt.gz for mm10). To understand the columns in this file, check the query annotation at http://hgdownload.soe.ucsc.edu/goldenPath/mm10/database/refGene.sql.

1. Write a small python script to make a file which contain each first exon in Chr13 from the refGene.txt. The file format could be: 1# column for chromosome, 2# column for start site, and 3# for end site.
2. Using 5pm HDAC3 peak information and file generated in step a, can you write a further python program to calculate that how many binding sites are bound in first exon region of Chr13 refGene.

Hint: if no chr13 binding sites are found in your results, please select a chromosome personally to do this program, and label in your answer.

**Rules for submitting the homework:**

Please submit your solution directly on the canvas website. Please provide your code, and a **pdf** file for your final write-up. Please pay attention to the clarity and cleanness of your homework. Page numbers and figure or table numbers are highly recommended for easier reference.

If you have any questions about the HW, please post them on the Canvas HW3 discussion site, instead of emails. The teaching fellows will grade your homework and give the grades with feedback through canvas within one week after the due date. Some of the questions might not have a unique or optimal solution. TFs will grade those according to your creativity and effort on exploration, especially in the graduate-level questions.