STAT115 Homework 4: scRNA-seq and ChIP-seq

(your name)

2018-03-15

Part I. scRNA-seq

We will continue to analyze the scRNA-seq data from last HW. Let's take the top 10 principle components from the original expression matrix. We have provided you with a Seurat object located on Odyssey at /n/stat115/HW4/rand10Xneurons.rds on which PCA has already been performed. Please transfer the file to your local computer for the following analysis.

1. Run tSNE on the 10 principle components. Visualize the cells and their corresponding tSNE coordinates and comment on the number of cell clusters that become apparent from the visualization. On a 2D plot, does tSNE cluster differ from PCA clustering?

```
# Your code here
```

2. For GRADUATE students: If you run tSNE several times with different k.seed, are the number of clusters the same and are the clusters robust?

3. For GRADUATE students: Try different resolutions in tSNE and visualize these clusters on the tSNE graph. How does changing resolution have on the number of clustering and the number of cells assigned to each cluster?

4. Use resolution = 1.5, how many cells are assigned to each group? 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.

5. For GRADUATE students: Based on the expression characteristics of your cell clusters, provide putative biological annotation (e.g. Foxp1, Sp9 genes are high in striatopallidal projection neurons) to 4 of the clusters. This paper may serve as a good resource as well as the Allen Brain Atlas.

Part II. ChIP-seq

KLF5 is a transcription factor frequently amplified in gastric cancers. Since it is technically challenging to do transcription factor ChIP-seq directly in tumors, one study Chia et al, Gut 2014 conducted KLF5 ChIP-seq in gastric cancer cell lines. The authors also used siRNA to knockdown KLF5 in the same cell lines to compare the differential expression between KLF5 high and KLF5 low cells. We will learn ChIP-Seq data analysis with MACS and BETA, as well motif finding, target gene identification, and functional annotation. We want to integrate ChIP-seq and differential gene expression in order to answer some important questions:

- What are the direct targets of KLF5 binding in gastric cancer?

- What are the collaborating transcription factors of KLF5 in gastric cancer?
- Does KLF5 binding activate or repress gene expression?
- What are the functions of KLF5 in gastric cancer?

RMA and LIMMA are available in R Bioconductor. SamTools, BWA, MACS and BETA are available on Odyssey. You can use the following commands to load them: module load macs2/2.1.0.20140616-fasrc01 module load centos6/BETA_1.0.5

It will help to read the MACS README and BETA Nature Protocol paper:

- https://pypi.python.org/pypi/MACS2/2.0.10.09132012
- http://liulab.dfci.harvard.edu/publications/NatProtoc12_1728.pdf
- http://liulab.dfci.harvard.edu/publications/NatProtoc13_2502.pdf

The raw data for this study are available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1250891. The original study covered three different transcription factors in three different gastric cell lines, but we will only focus on KLF5 in YCC3 cell line. Needed data are available on Odyssey under /n/stat115/HW4/.

6. Identify differentially expressed transcripts between wild type vs. knockdown (in RefSeq) using LIMMA. Assume for simplicity that A) the data do need normalization and B) there is no batch effect. Make sure you use the right cdf file for this array type. Use appropriate cutoffs for fold-change and p-values/FDR when identifying differentially expressed transcripts.

7. For GRADUATE students: We sampled 1M ChIP-seq raw reads from one ChIP-seq data (/n/stat115/HW4/sample.1M.fastq). Run BWA on the sampled fastq data to map the reads to Hg38 human genome assembly. Report the commands, logs files, and a snapshot out the output (possibly using screenshots) to demonstrate your alignment procedure. What proportion of the subsetted reads successfully mapped?

8. 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). What % of reads are redundant in these two samples?

Hint: Read the manual of MACS2 (https://pypi.python.org/pypi/MACS2/2.0.10.09132012). It takes about ~10 minutes to run this on Odyssey.

9. In ChIP-Seq analysis, a bias often occurs in results when the number of reads in treatment and control are different. One solution for correcting the bias is to subset the sample with the larger number of reads to the same number as the treatment, which can be achieved using macs2 randsample. The samples in this experiment might not be too bad, but since it only takes about 1~2 minutes to run this on Odyssey, let's give it a try. Show the file sizes of the larger sample with more reads, before and after down sampling.

10. After removing the duplicated reads and sampling to balance reads in the ChIP-seq and control samples, we are ready to call the peaks using MACS2. How many KLF5 peaks have FDR < 0.05 and Fold change > 2?

Hint: Read the manual of MACS2. Run macs on Odyssey with enough memory (2GB) by using "srun –mem=2048 -p interact –pty" This will take ~an hour on Odyssey.

11. For GRADUATE students: Search for this sample in Cistrome DB (http://cistrome.org/db/#/). Comment on the overall quality of this particular KLF5 ChIP-seq sample.

12. When the total ChIP-seq peaks for factors (with FDR cutoff alone) are less than 10K, you can run BETA without filtering by fold change. Run BETA-basic to integrate KLF5 binding data with differential expression data to study KLF5 regulation function. Does TET1 function as a gene expression activator, repressor, or both? Support your answer with the summary plot from BETA.

Hint: Read the BETA Nature Protocol paper. In addition, BETA cannot run on PC. If serious error reports in server, we recommend running BETA on 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. Your input should be the peaks that you called in step (e) with the LIMMA output table in step a). You need to have an additional column from LIMMA to run BETA. In addition, think carefully: if you see genes down-regulated when a transcription factor is knocked down, does this factor function as an activator or repressor?

13. For GRADUATE students: What is the KLF5 binding motif? Based on the BETA-plus motif analysis, does KLF5 collaborate with other families of transcription factors? This will take ~ 1 hour on Cistrome, if your job is ahead of the other 50 students in the class, so please budget sufficient time.

Hint: Only consider motifs with the most stringent cutoff. If the list of significant motifs is long, consider the top 3 unique motifs.

14. For GRADUATE students: Transcription factors work in a complex called cofactors. Motif analysis from BETA suggests other families of TFs might collaborate with KLF5 to regulate its target genes. Could you look at the expression data to pinpoint the exact family member that is KLF5's collaborator?

Hint: If this transcription factor is a collaborator of KLF5, it is likely expressed higher than other family members, and also has correlated gene expression with KLF5 in tumors. If your differential expression data

is in RefSeq (e.g. NM_12345), you need to convert it into gene symbol (e.g. TP53).

15. Perform GO analysis on the BETA output of KLF5 top 200 direct target genes using DAVID. What are the functional enrichment of KLF5 direct target genes in gastric cancer?

Hint: TF target is not a binary Yes / No decision, but a strong / weak gradient. Pick a reasonable number of targets to run DAVID.

Submission

Please submit your solution directly on the canvas website. Please provide your code (.Rmd) and a pdf/html 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.

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.