STAT115 Homework 3

(your name)

2017-03-03

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 here.

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?

```
library(Seurat)
library(Matrix)
library(dplyr)

neurons <- readRDS("data/Part1/rand10Xneurons.rds")

# Type cast the rownames/colnames of the raw data
rownames(neurons@raw.data) <- as.character(rownames(neurons@raw.data))
colnames(neurons@raw.data) <- as.character(colnames(neurons@raw.data))

# Remove duplicate row names (there's ~6 in the data; should be useful below)
neurons@raw.data <- neurons@raw.data[!duplicated(rownames(neurons@raw.data)),]
```

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.

G1. [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?

4. 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?

5. 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.

6. 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.

G2. [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 as well as the Allen Brain Atlas.

7. 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 integrative model–BETA. This software provides a multi-faceted methodology where ChIP-seq, epigenetics, motif finding, target gene identification, and functional annotation are all integrated.

From Lab 1, you should have RMA and LIMMA installed. MACS should be ran on Odyssey. A good starting place may be the MACS README and BETA Nature Protocol paper.

You should be able to load the MACS software in Odyssey using the following:

```
macs2/2.1.0.20140616-fasrc01
```

And then verify that the module has been loaded:

```
macs2 --help
```

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 promotes DNA demethylation by converting 5-methylcytosine to 5-hydroxymethylcytosine (5hmC). Under the prevailing paradigm (methylation silences gene expression), the conversion to 5hmC should activate gene expression. However, recent studies suggests a more complex regulatory mechanism. Here we provide TET1 ChIP-seq data in mouse embryonic stem cell:

`/n/stat115/hws/3/2017/TET1treatment.bam` (BAM file for the ChIP-seq data for TET1, GSM706672)
`/n/stat115/hws/3/2017/control.bam` (BAM file for ChIP-Seq control data, GSM706673)

Gene Expression Data

The expression data are from two conditions, which are wild type mouse embryonic stem cell (ES) and ES with TET1 knockdown by siRNA. The phenotype of expression data:

```
GSM846063   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
```

Note that these are already present in the git repository.

8. Identify differentially expressed transcripts between wild type vs. knockdown using LIMMA. Assume for simplicity that A) the data do need normalization and B) there is no batch effect. Use appropriate cutoffs for fold-change and p-values/FDR when identifying differentially expressed transcripts.

Note: You will need to export the full LIMMA results table to complete the later stages of the assignment. Use the `write.table` function after annotating your data frame with RefSeq IDs. This Bioc package should help with annotation.

```
library(affy)
celFiles <- list.celfiles(path = "data/Part2", full.names=TRUE)
data.affy <- ReadAffy(filenames = celFiles)
```

Note: like you did in Lab 1, make sure to convert your Affymetrix Probe IDs to a more appropriate format (RefSeq).

9. Since aligning the full .fastq file would be time-intensive, a subset of ~1M raw reads from ta ChIP-seq experiment were extracted and saved as `/n/stat115/hws/3/2017/sample.1M.fastq`. Align this fastq file using BWA to the mm10 reference genome. 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?
The mm10 index library for BWA is available under the `/n/stat115/hws/3` directory in Odyssey

10. In ChIP-Seq experiments, when library preparation involves a PCR amplification step, the observed sequences often have 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. Using the `macs2 filterdup` option, report the duplicate rates for both the treatment and control samples. Which of the two libraries is seemingly more biased due to the PCR amplification?

11. 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. Show the file sizes of the sample with more reads before and after downsampling. Using the files with equivalent numbers of deduplicated reads, call peaks for the TET2 treatment sample.

12. Integrate TET1 binding data with differential expression data to study TET1 regulation function by running BETA. Does TET1 function as a gene expression activator, repressor, or both? Support your answer with the summary plot from BETA. How many instances of activation and repression do you see?
Hint: you can run BETA on CistromeAP Galaxy instance (http://cistrome.org/ap/) under the "integrative analysis" option. Your input should be the peaks that you called in 11 with the differential expression table in step 8.

G3. [Graduate students] Transcription factor often regulate gene expression programs in multi-teared complex called co-factors. The motif analysis from BETA may suggest the presence of particular cofactors. From the comparison of differential expression between knockdown and wild type, identify putative pairs of transcription factors that may be cofactors. Support your reasoning with figures and analyses.
Hint: Graduate Students need to run BETA-plus to get the motif information. Running BETA-Basic is sufficient to answer 12 alone. There is a fairly significant difference in execution time, so allow for nearly approximately 1 hour when running the BETA-plus.The default input values should be fine noting which columns correspond to the summary statistics from the LIMMA output.
Hint 2: (How CL would solve this). Take all motifs that have a fairly stringent statistical significance threshold (on the order of 20 or so motifs) and note which genes they regulate (up or down). Find the corresponding TFs in the expression data from 8 (may have to do gene ID conversion). If two TFs are differentially expressed (from LIMMA) and are targeting the same sets of genes (from BETA), then one could infer that they are cofactors.

13. Perform gene ontology enrichment analysis on the output genes from BETA using DAVID. What pathways does TET1 regulate (enriched GO terms or pathways)? What are the p-values

and FDR for the most significant enrichments?
Hint: use screenshots to build a narrative.

G4. [Graduate students] With the increasing availability of integrated datasets that include epigenetics and RNA (through RNA-Seq or microarray), an exciting area of bioinformatics research involves integrative modeling of multiple datasets. BETA provides one such integrative workflow. Discuss the relative merits of the BETA approach compared to (at least) one other analysis platform that integrates epigenetic and transcriptomic data. Provide the journal article source of the tool that you are evaluating.

Part 3– Python Programming

From UCSC download page, find the mm10 RefSeq annotation table (the file refGene.txt.gz is contained in the Git repository). The annotation for each column is available here.

14. Write a python script to read in the refGene.txt file and output all the first exon of each gene on chromosome 13. Your output file should have 3 columns, the chromosome (each element should be "13"), the start base pair, and the end base pair. In addition to providing your code, report how many exons result from this parsing.
Hint: you don't need to have your python script execute at run time; simply show the script in this file.

```
# Write python code here
```

The number of lines in the resulting file where: `this many`.

15. Using the peak information and your file generated in 14, determine how many overlaps are present between between OCT4 transcription factor binding sites (file in `data/mESC_OCT4_chr13pad.bed`) and chromosome 13 first exons. Once again, show your code and the number of overlapping instances.

Submission

Please submit your solution directly on the canvas website. Provide your code (.Rmd) 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.

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.**