STAT115 Homework 3: RNA-seq and scRNA-seq

(your name)

2018-02-20

Recommendation to students: start on Part II and Part III first!

Part I. RNA-seq

For this HW, we will use the RNA-seq data from the ENCODE project. There are two RNA-seq samples of HepG2 with U2AF1 knock down and control, each with 2 replicates. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE88002 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE88226 The raw FASTQ files are available on Odyssey at:

1. Normally for paired-end RNA-seq data, each sample will have two separate FASTQ files, with line-by-line correspondence to the two reads from the same fragment. Use STAR (Dobin et al, Bioinformatics 2012) to map the reads to the reference genome, available on Odyssey at /n/stat115/HW3/STARIndex. Note this folder contains the STAR index and you do not need to generate your own. The mapping could take a long time, so we have created two FASTQ files in one sample with only 5M fragments for you to run STAR instead of the full data (/n/stat115/HW3/FastqData/ENCFF500PDO_sub.fastq,/n/stat115/HW3/FastqData/ENC Generate the output in standard SAM format. How many fragments out of the 5M raw data are mappable? How long did it take?

```
# Your code here
```

2. Traditional read mappers align reads to a reference genome using various algorithms however they are relatively slow compared to the newer pseudo-mapping techniques which only align to a RefSeq transcriptome. These pipelines greatly simplify the process of going from FASTQ to read counts on genes and are much, much faster. Use Salmon to pseudo-align the 5M fragments to the human reference transcriptome available at /n/stat115/transcriptome/Homo_sapiens.GRCh38.cdna.all.fa. Note that you will need to first create a Salmon index of the transcriptome and then quantify the sample abundances to this index. Please include `--gcBias` option. How does Salmon compare to STAR in runtime? What is the gene with the highest TPM in the sample?

Hint: https://www.bioconductor.org/help/workflows/rnaseqGene/

```
# Your code here
```

3. Now run Salmon on the full data to generate the quant.sf files, then run DESeq2 (Love et al, Genome Biol 2014) to find the differentially expressed transcripts at the $\alpha = .01$ significance level. How many RefSeq transcripts are up vs down-regulated U2AF1? Provide a MA plot displaying the differential expression. You will not need to generate a new index and may use the same one from the previous question.

Hint: https://www.bioconductor.org/help/workflows/rnaseqGene/.

```
# Your code here
```

```
# Your code here
```

4. Use some GO analysis tool to examine whether U2AF1 regulates some specific functions / processes / pathways.

5. For GRADUATE students: DESeq2 can optionally aggregate the differential expression at either transcript (no aggregate) or gene level (aggregate). How do the DE results compare between running the pipeline with/without aggregation to gene level, in terms of the genes / transcripts called and the GO categories?

```
# Your code here
```

6. For GRADUATE students: Are the same genes/txs called DE if you use TPMs instead of counts (i.e. abundances versus counts)? Explain the difference between TPMs and counts.

```
# Your code here
```

Part II. 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.

7. Describe the composition of the raw dataset (i.e. number of genes, number of samples, and dropout rate).

```
library(Seurat)
```

```
## Warning: package 'Seurat' was built under R version 3.4.3
```

```
## Warning: package 'cowplot' was built under R version 3.4.3
```

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

```r
neurons <- readRDS("data/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)),]

# Your code...
```

8. We want to filter weakly detected cells and lowly expressed genes. Let's keep all genes expressed in >= 3 cells, and all cells with >= 200 detected genes. How do these summary statistics change?

```r
# Your code here
```

9. 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 (http://satijalab.org/seurat/pbmc3k_tutorial_1_4.html). Remove the unwanted source of variation from mitochondria genes.

```r
# Your code here
```

10. Perform PCA on the dataset after the filtering. Provide summary plots, statistics, and tables to show 1) how many PCs are statistically significant, 2) which genes contribute to which principle components, and 3) how much variability is explained in these top components.

```r
# Your code here
```

11. For GRADUATE students: Determine which PCs are heavily weighted by cell cycle genes. Provide plots and other quantitative arguments to support your case. How to correct the unwanted variation from cell cycle genes?

```r
# Your code here
```

Our scRNA-seq saga will continue in HW4...

Part III. Python programming

For Part I, we ask you to use STAR to map 5M fragments and generate a SAM file. Write a python program to calculate the average insert size (distance between the 5' ends of the +strand and -strand reads) of the top 1000 fragments where both ends are mappable and their mapped distance are between

50 and 2000bp. If the distance between the two reads are out of range, skip to the next fragment. https://samtools.github.io/hts-specs/SAMv1.pdf

Submission

Please submit your solution directly on the canvas website. Please 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.