STAT115 Homework 1

(your name)

2017-01-21

Part I– Microarrays

In this part, we are going to analyze a microarray gene expression dataset from the following paper using the methods learned from the lecture: Xu et al., **Science** 2012. *EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent.* PMID:23239736

The expression data is available at GEO under GSE39461, and for this HW we only need the first 12 samples. There are two prostate cancer cell lines used: LNCaP and ABL (ignore the "DHT" and "VEH" labels). To see the function of EZH2 gene, the authors knocked down EZH2 (siEZH2) in the two cell lines and examined the genes that are differentially expressed compared to the control, and generated 3 replicates for each condition. They are also interested in finding whether the genes regulated by EZH2 are similar or different in the LNCaP and ABL cell lines.

First, take a look at the following quick tutorial about Affymetrix array analysis. Also, please watch this video about batch effect.

1. Make sure BioConductor and all the modules you will need are installed, and include them in your R code.

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("affy")
# etc.
```

```
library(knitr)
library(affy)
```

2. Download the needed CEL files (GSM969458 to GSM969469) to your cwd. Note your cwd needs to be the same as where your CEL files are, or specify the file names using the argument filenames in ReadAffy. Load the data in R. *Note: if you've cloned this file from the Git repository, the files are located in the data subdirectory.*

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

3. Draw pairwise MA plot of the raw probe values for the 3 ABL in control samples. Do the raw data need normalization?

4. Use RMA, which includes background correction, quantile normalization, and expression index estimation, to obtain the expression level of each gene. This will generate an expression matrix, where genes are in rows and samples are in columns. (Hint: You may also try different bc, qc and ei methods by using expresso function)

5. What are the assumptions behind RMA qnorm? Draw pairwise MA plot on the expression index for the 3 ABL control samples after RMA. Is the RMA normalization successful?

G1. [GRADUATE STUDENTS] Can you replace the original CDF file so that RMA outputs expression index for RefSeq (e.g. NM_000546) vs Gene Symbol (e.g. TP53)?

6. Use LIMMA, find the differentially expressed genes between siEZH2 and control in LNCaP cells, and repe at the same in ABL cells. Use false discovery rate (FDR) 0.05 and fold change (FC) 1.5 as cutoff to filter the final result list. How many significant genes (if you are using gene symbol like TP53) or transcripts (if you are using RefSeq like NM_000546) did you find in each comparison? You can either do differential genes or transcripts, but make sure to specify this clearly in your answer.

7. Use hierarchical clustering to cluster the 12 samples. Does the clustering suggest the presence of batch effect? Why?

8. Use ComBat (http://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html) to adjust for batch effects and provide evidence that the batch effects are successfully adjusted.

9. Repeat the differential expression analysis using LIMMA, FDR 0.05 and FC 1.5. Are there significant genes reported?

10. Use Venn diagram to show siEZH2 up genes in LNCaP and ABL (show 3 numbers, genes that overlap, unique to LNCaP, or unique to ABL?) and the down genes in the two cells.

11. Use the union of differential (up or down) genes in LNCaP and ABL, and the expression index of all 12 samples, cluster the samples using hierarchical clustering. Draw a heatmap to show the clustering results.

G2. [GRADUATE STUDENTS] If you use single, complete or average linkage for hierarchical clustering of the genes, do the clustering heatmap look different?

12. Run the four list of differential genes (up / down, LNCaP / ABL) separately on DAVID (you might want to read their Nature Protocols tutorial) to see whether the genes in each list are enriched in specific biological process, pathways, etc. What's in common and what's the most significant difference in EZH2 regulated genes between LNCaP and ABL?

G3. [GRADUATE STUDENTS] Try Gene Set Enrichment analysis on the siEZH2 experiments in LNCaP and ABL separately. Do the two cell lines differ in the enriched gene sets?

G4. [GRADUATE STUDENTS] [Surrogate Variable Analysis](http://www.bioconductor.org/packages/release/bioc/html/sva.html) is designed to identify and correct for potential batch effect even when the batch information or the confounding factors are not known a priori. Run SVA on the original expression data (without COMBAT and without giving the known batch information). Can SVA correctly identify the batches and correct it? How does the batch correction compare with COMBAT?

Part II– Python Programming

In this part, we will do a python programing exercise, so that you can start to learn and use python as soon as possible. The following code is intended to translate DNA sequence to amino acid sequence, please complete it and use the data below to test your result.

Python code:

```
dna_seq = 'GCGTTTGACCGCGCTTGGGTGGCCTGGGACCCTGTGGGAGGCTTCCCCGGCGCCGAGAGCCCTGGCTGACGGCTGATGGGGAGGAGCCGGCG

protein_seq = 'MGRSRRAEKATGSPVPSPARDRCGKPGGASAGPAERTSEVKSLVYLPLGAGLGPQPLP_'

# Hint: can keep everything self-contained in your
# .Rmd using a different language engine: https://yihui.name/knitr/demo/engines/
def translate_DNA(sequence):
    start_codon = 'ATG'
```

```
    stop_codons = ('TAA', 'TAG', 'TGA')
    codontable = {
        'ATA':'I', 'ATC':'I', 'ATT':'I', 'ATG':'M',
        'ACA':'T', 'ACC':'T', 'ACG':'T', 'ACT':'T',
        'AAC':'N', 'AAT':'N', 'AAA':'K', 'AAG':'K',
        'AGC':'S', 'AGT':'S', 'AGA':'R', 'AGG':'R',
        'CTA':'L', 'CTC':'L', 'CTG':'L', 'CTT':'L',
        'CCA':'P', 'CCC':'P', 'CCG':'P', 'CCT':'P',
        'CAC':'H', 'CAT':'H', 'CAA':'Q', 'CAG':'Q',
        'CGA':'R', 'CGC':'R', 'CGG':'R', 'CGT':'R',
        'GTA':'V', 'GTC':'V', 'GTG':'V', 'GTT':'V',
        'GCA':'A', 'GCC':'A', 'GCG':'A', 'GCT':'A',
        'GAC':'D', 'GAT':'D', 'GAA':'E', 'GAG':'E',
        'GGA':'G', 'GGC':'G', 'GGG':'G', 'GGT':'G',
        'TCA':'S', 'TCC':'S', 'TCG':'S', 'TCT':'S',
        'TTC':'F', 'TTT':'F', 'TTA':'L', 'TTG':'L',
        'TAC':'Y', 'TAT':'Y', 'TAA':'_', 'TAG':'_',
        'TGC':'C', 'TGT':'C', 'TGA':'_', 'TGG':'W'
    }

    start = sequence.find(start_codon)
    codons = [] #Create a codon list to store codons generated from coding seq.
    # for i in range(start, len(sequence), 3):
    # Finish the loop or write your own code to find the coding sequence which should
    # start at the first start codon and stop at the occurrence of any stop codon.

    protein_sequence = ''.join([codontable[codon] for codon in codons]) #Translate condons to protein s
    return "{0}_".format(protein_sequence)
print(protein_seq)
```

```
## MGRSRRAEKATGSPVPSPARDRCGKPGGASAGPAERTSEVKSLVYLPLGAGLGPQPLP_
```

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.