STAT115 Homework 2

(your name)

2018-02-04

Part I. Expression Index and Differential Expression

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 (please 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: http://homer.salk.edu/homer/basicTutorial/affymetrix.html.

Also, please watch this video about batch effect: http://www.youtube.com/watch?v=z3vqrkRGSLI.

Now let's analyze the data in this Science paper:

0. 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(affy)
# etc.
```

1. 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 you can specify the file names using the argument filenames in ReadAffy. Load the data in R. Draw pairwise MA plot of the raw probe values for the 3 ABL in control samples. Do the raw data need normalization?

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

2. 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. 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?

3. 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 genes are differentially expressed at this cutoff? Count in terms of gene symbol (e.g. TP53) instead of transcript like (e.g. NM_000546)?

4. Draw a hierarchical clustering of the 12 samples. Does the clustering suggest the presence of batch effect? Hint: use functions dist, hclust, plot.

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

6. FOR GRADUATES: Run K-means clustering of differentially expressed genes across all 12 samples. Experiment with different K (there may not be a correct answer here so just explore and explain your reasoning). Hint: function kmeans.

7. Run the four list of differential genes (up / down, LNCaP / ABL) separately on DAVID (http://david.abcc.ncifcrf.gov/, you might want to read their Nat Prot 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?

8. FOR GRADUATES: Try Gene Set Enrichment analysis (http://www.broadinstitute.org/gsea/index.jsp) on the siEZH2 experiments in LNCaP and ABL separately. Do the two cell lines differ in the enriched gene sets?

Note: In real data analysis, after expression index calculation, it is a good practice do clustering analysis to check for and correct potential batch effect before proceeding to differential expression analysis and GO analysis, even if are able to find sufficient number of differential genes before batch removal.

Part II: Microarray Clustering and Classification

The sample data is in file "taylor2010_data.txt" included in this homework. This dataset has expression profiles of 23974 genes in 27 normal samples, 129 primary cancer samples, 18 metastasized cancer samples, and 5 unknown samples. Assume the data has been normalized and summarized to expression index.

The skeleton R code is in HW2_1.R. Please fill in missing R code for each question, and attach this R file with homework submission.

```r
taylor <- as.matrix(read.csv("data/taylor2010_data.txt", sep="\t",row.names=1))
index_normal <- grepl("N.P", colnames(taylor))
index_primary <- grepl("P.P", colnames(taylor))
index_met <- grepl("M.P", colnames(taylor))
n_normal <- sum(index_normal);
n_primary = sum(index_primary);
n_met = sum(index_met);

# class label (design vector)
taylor_classes = c(rep(0,n_normal),rep(1,n_primary),rep(2,n_met));

# train (known type samples), and test (unknown type samples)
train <- taylor[,1:174];
test <- taylor[,175:179];

# colors for plotting
cols = c(taylor_classes+2, 1,1,1,1,1)
```

```
tumortype_class <- factor(taylor_classes, levels = 0:2,
                          labels = c("Normal", "Primary", "Metastasized"))

train_samps <- 1:174
test_samps <- 175:179
```

1. For the 174 samples with known type (normal, primary, metastasized), use LIMMA to find the differentially expressed genes with fold change threshold 1.5, and adjusted p-value threshold 0.05. How many differentially expressed genes are there? Hint: the design vector consists of type indicator for the 174 samples. For example, 0 for normal, 1 for primary, and 2 for metastasized.

2. Draw k-means clustering on all samples using the differentially expressed genes. Do the samples cluster according to disease status?

3. Do PCA biplot on the samples with differentially expressed genes genes, and use 4 different colors to distinguish the 4 types of samples (normal, primary, metastasized and unknown). Do the samples from different groups look separable? Hint: use the PCA ggplot R code, also function legend is useful. (http://docs.ggplot2.org/0.9.3.1/geom_point.html)

4. FOR GRADUATES: What percent of variation in the data is captured in the first two principle components? How many principle components do we need to capture 85% of the variation in the data? R Hint: use function prcomp.

5. Based on the PCA biplot, can you classify the 5 unknown samples? Put the PCA biplot in your HW write-up, and indicate which unknown sample should be classified into which known type (normal, primary, metastasized). Do you have different confidence for each unknown sample?

6. FOR GRADUATES: Use PCA on all samples and all the genes (instead of the differentially expressed genes) for sample classification. Compare to your results in the previous question. Which PCA plot looks better and why?

7. Run KNN (try K = 1, 3 and 5) on the differential genes and all the samples, and predict the unknown samples based on the 174 labeled samples. Hint: use the library `class` and function `knn`.

```
library(class)
```

8. Run SVM (try a linear kernel) on the differential genes and all the samples, and predict the unknown samples based on the 174 labeled samples. Hint: use the library `e1071` and function `svm`.

```
library(e1071)
```

9. FOR GRADUATES: Implement a 3-fold cross validation on your SVM classifier, based on the 174 samples with known labels. What is your average (of 3) classification error rate on the training data?

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.