Statistics For Genomics

Assignment 1

Martin Skarzynski

2018-04-26

Statistics For Genomics – Homework 1

Instructions

Using a microarray dataset, conduct a differential expression analysis. This includes completing the following tasks

- A description of the biological problem.
- Preprocessing the data, and examining if the preprocessing appears effective.
- Do differential expression.
- Translate your statistical findings into biological conclusions.

The end result will be a ~3 page report plus up to 5 display items. R code to reproduce your results should be included and commented to such an extend that a reviewer may be able to check whether the analysis was correct. I suggest commenting at the level of "The following code produces Figure X", or "Here we normalize the data using X", not at the level of single lines. The material should be made available as a single PDF.

I expect you will need to make more than 5 figures in the course of this work, but only include figures that are "interesting" and "worth mentioning". Shorter text (but not too short) is in general better. I expect that every single conclusion is explictely stated, including what you conclude based on the figures, ie. do not just say "Figure X is an MA plot", but "Figure X is an MA plot with an added lowess line. The line shows that the fold change estimates are symmetric around zero." These reports will be assessed using peer review, so write to the level of your fellow students. We will assess the following criteria

- The ability to do a correct analysis.
- The ability to translate between statistics and biology.

For microarray data, I suggest looking at NCBI GEO for data. You can utilize the Bioconductor package GEOquery to programmatically obtain the data in R. You are not allowed to use data described in the Limma Users Guide.

Introduction

The title of the dataset I chose is:

Gene expression profile of human monocytes stimulated with all-trans retinoic acid (ATRA) or 1,25a-dihydroxyvitamin D3 (1,25D3)

The dataset is available on the Gene Expression Omnibus (GEO) website

GEO is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community.

This dataset was published in a paper by Wheelwright et al. in 2014:

Wheelwright M, Kim EW, Inkeles MS, De Leon A et al. All-trans retinoic acid-triggered antimicrobial activity against Mycobacterium tuberculosis is dependent on NPC2. J Immunol 2014 Mar 1;192(5):2280-90. PMID: 24501203

The authors treated human monocytes with all-trans retinoic acid (ATRA) or 1,25a-dihydroxyvitamin D3 (1,25D3) and measured their gene expression using microarray.

Data Analysis

```
#Install the Bioconductor libraries
#source("http://bioconductor.org/biocLite.R")
#biocLite("GEOquery")
#biocLite("limma")
#biocLite("Biobase")
#biocLite("affy")
```

```
#Load the Bioconductor libraries
library(GEOquery)
library(Biobase)
library(limma)
#library(affy)
```

```
#Load the dataset using the getGEO() Command
gset <- getGEO("GSE46268", GSEMatrix =TRUE)
```

```
#Inspect data
class(gset)
```

```
## [1] "list"
```

```
length(gset)
```

```
## [1] 1
```

```
names(gset)
```

```
## [1] "GSE46268_series_matrix.txt.gz"
```

```
#Unpack data
idx <- length(gset) # 1
gset <- gset[[idx]]
```

```
#Inspect data
class(gset)
```

```
## [1] "ExpressionSet"
## attr(,"package")
## [1] "Biobase"
```

```
length(gset)
```

```
## [1] 1
```

```
names(gset)
```

```
## NULL
```

```
slotNames(gset)
```

```
## [1] "experimentData"    "assayData"         "phenoData"
## [4] "featureData"       "annotation"        "protocolData"
## [7] ".__classVersion__"
```

```
#Explore the new dataset structure
## Dataset dimensions:
dim(gset)
```

```
## Features  Samples
##    54675       12
```

```
#The features are the number of genes on the array (54675). There are 12 samples.
```

```
#Dataset structure
#str(gset)
```

```
#Inspect list of samples with their associated attributes and their treatment (phenotypic data)
#pData(phenoData(gset))
#Get phenotype data dimensions
dim(pData(phenoData(gset)))
```

```
## [1] 12 42
```

```
#Get first 5 column names
colnames(pData(phenoData(gset)))[1:5]
```

```
## [1] "title"            "geo_accession"    "status"
## [4] "submission_date"  "last_update_date"
```

```
# group names for all samples in a series
sml <- c("G0","G0","G0","G0","G1","G1","G1","G1","G2","G2","G2","G2")
# order samples by group
ex <- exprs(gset)[ , order(sml)]
sml <- sml[order(sml)]
fl <- as.factor(sml)
labels <- c("control","retinoic","D3")
```

Figures

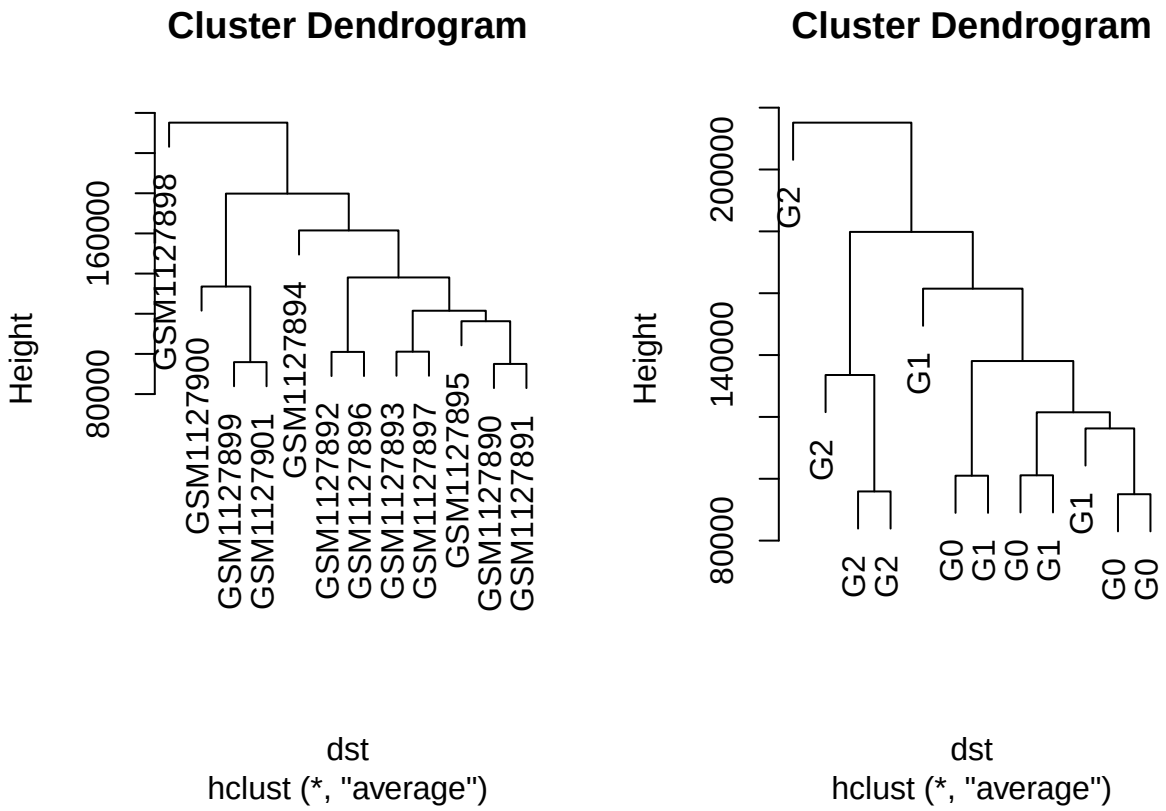
```
# set parameters and draw the plot
palette(c("#dfeaf4","#f4dfdf","#f2cb98", "#AABBCC"))
dev.new(width=4+dim(gset)[[2]]/5, height=6)
par(mar=c(2+round(max(nchar(sampleNames(gset)))/2),4,2,1))
title <- paste ("GSE46268", '/', annotation(gset), " selected samples", sep ='')
boxplot(ex, boxwex=0.6, notch=T, main=title, outline=FALSE, las=2, col=fl)
legend("topleft", labels, fill=palette(), bty="n")
```

The boxplots all look very similar between the groups.

The analyses above utilize elements of GEO2R, which is described on the GEO web site.

Hierarchical Clustering

```
# calculate a distance matrix between each sample (each array)
dst <- dist(t(exprs(gset)))
# Hierarchical cluster analysis on above distance matrix
hh <- hclust(dst, method="average")
#plot the tree by sample name or by group name using the fl object created previously:
# We will plot both of them on the same plot
par(mfrow=c(1,2))
# plot default is by sample name
plot(hh)
# label sample by group
plot(hh, label=fl)
```



The groups are labeled G0, G1 and G2. These labels correspond to the treatments:

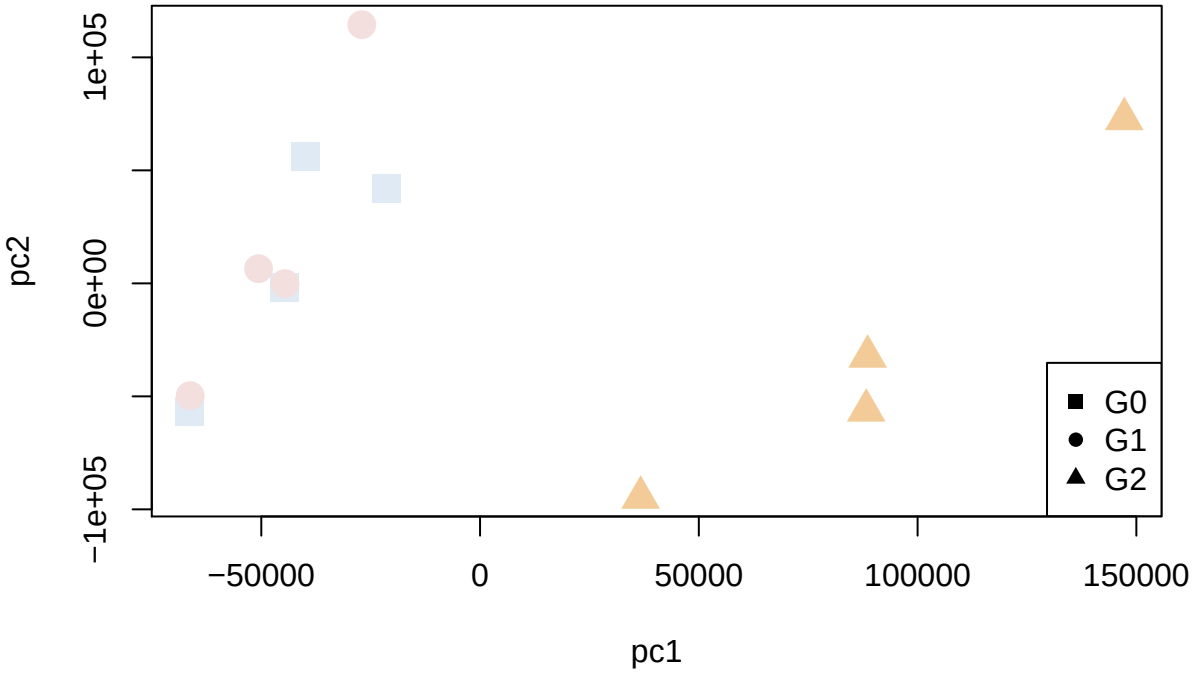
- 1,25a-dihydroxyvitamin D3 (G0),
- all-trans retinoic acid (G1),
- and control (G2).

Hierarchical Clustering Plot Conclusions

The hierarchical clustering method has trouble separating the Group1 (G1) and Group 0 (G0), but Group 2 is very distinct. This is to not all together surprising because the control is different than treatment, but I would not expect Vitamin D and ATRA to have similar effects.

Principal Component Analysis

```
par(mfrow=c(1,1))
#Principal Component Analysis
PC=prcomp(t(exprs(gset)))
scores = predict(PC)
# extract PC1 and PC2
pc1 <- scores[ ,1]
pc2 <- scores[ ,2]
# Create a vector of number for choosing the plot symbol
# We add 14 to reach the symbols that are filled
shape <- as.numeric(fl) + 14
# plot for the first 2 principal components
plot(pc1, pc2, col=fl, pch=shape, cex=2)
# legend("topright", pch=unique(shape), paste(unique(fl)))
# add a legend
legend("bottomright", pch=unique(shape), paste(unique(fl)))
```



PCA Plot Conclusions

The Principal Component Analysis (PCA) method also has trouble separating the Group1 (G1) and Group 0 (G0), but Group 2 is, as before, very distinct. It may not be possible to separate out the effects of Vitamin D and ATRA in these samples.

```
design.matrix <- model.matrix(~sml)
fit <- lmFit(object = ex, design = design.matrix)
head(fit$coefficients)
```

```
##           (Intercept)      smlG1      smlG2
## 1007_s_at    739.0033 -10.32250 -147.27425
## 1053_at      644.7845   8.51400 -172.87450
```

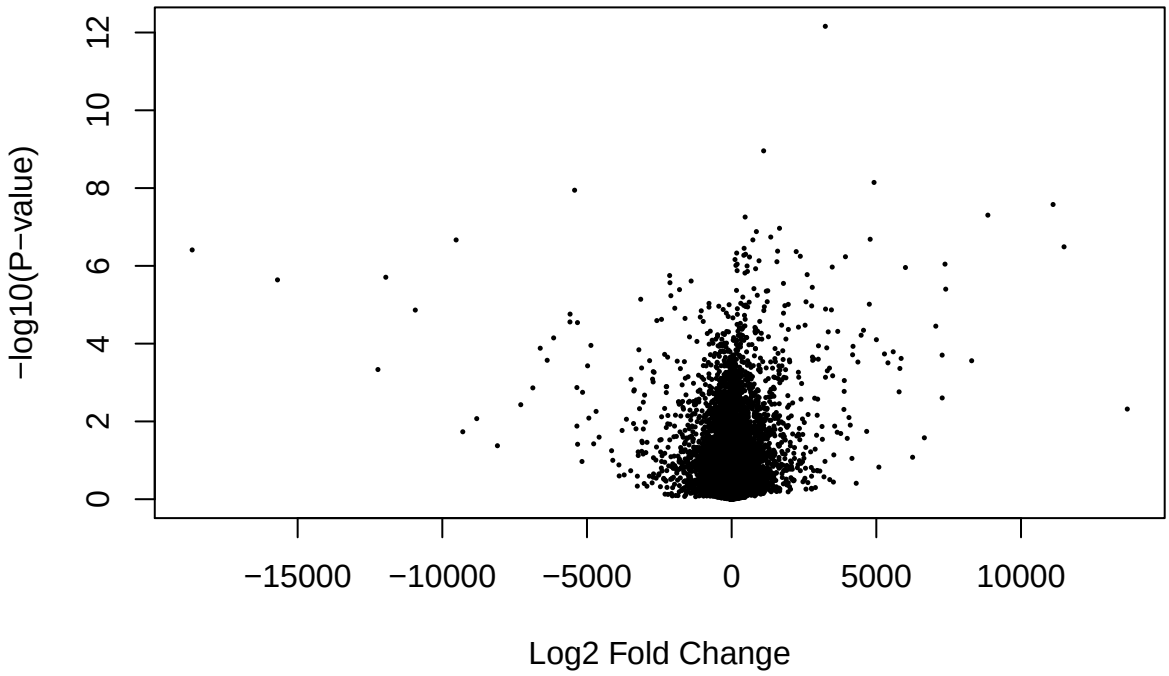
```
## 117_at      1764.7650 -355.76750 -895.33125
## 121_at      1598.8000 -375.46250 -343.46550
## 1255_g_at     48.3441   -1.85105  -14.07575
## 1294_at      609.6133   43.47225 -112.34700
```

```
beta <- fit$coefficients[,2]
s <- fit$stdev.unscaled[,2]*fit$sigma
n <- 4
t <- (beta*sqrt(n))/s
fit2 <- eBayes(fit)
names(fit2)
```

```
##  [1] "coefficients"     "rank"             "assign"
##  [4] "qr"               "df.residual"      "sigma"
##  [7] "cov.coefficients" "stdev.unscaled"   "pivot"
## [10] "Amean"            "method"           "design"
## [13] "df.prior"         "s2.prior"         "var.prior"
## [16] "proportion"       "s2.post"          "t"
## [19] "df.total"         "p.value"          "lods"
## [22] "F"                "F.p.value"
```

Volcano Plot

```
#p <- fit2$p.value[,2]
#plot(fit2$coefficients[,2], -log2(p), pch=20, cex=0.3)
volcanoplot(fit2, coef=2)
```



Volcano Plot Conclusions

My volcano plot looks a little strange in that there are very few highly significant (p-value) genes with large differences (Fold Change). It is possible that the limited number of differentially expressed genes are allowing for separation of the treated groups from the control, but are too few or too similar between treatment groups to allow for the hierarchical clustering and Principal Component Analysis to distinguish between all-trans retinoic acid (ATRA) or 1,25a-dihydroxyvitamin D3 (1,25D3). The other possible interpretation is that the treatments have similar effects. I think this second possibility is unlikely to be true, because all-trans retinoic acid (ATRA) or 1,25a-dihydroxyvitamin D3 (1,25D3) bind different nuclear receptors. ATRA binds retinoid X receptor (RXR) and retinoic acid receptor (RAR), whereas 1,25a-dihydroxyvitamin D3 (1,25D3) binds the vitamin D receptor (VDR). Therefore, I believe that the similarity between the treatment groups in my analysis is a limitation of the data or the methods I employed.