Homework due May 13, 2021 23:01 +03

In this section we will use the `sva()` function in the **sva** package and apply it to the following data:

```
library(sva)
library(Biobase)
library(GSE5859Subset)
data(GSE5859Subset)
```

SVA Exercises #1

1/1 point (graded)

In the previous section we estimated factors using PCA. But we noted that the first factor was correlated with our outcome of interest:

```
s <- svd(geneExpression-rowMeans(geneExpression))
cor(sampleInfo$group,s$v[,1])
```

As in the previous questions we are interested in finding genes that are differentially expressed between the two groups (males and females in this case). Here we learn to use SVA to estimate these effects while using a factor analysis approach to account for batch effects.

The `svafit()` function estimates factors, but downweighting the genes that appear to correlate with the outcome of interest. It also tries to estimate the number of factors and returns the estimated factors like this:

```
sex = sampleInfo$group
mod = model.matrix(~sex)
svafit = sva(geneExpression,mod)
head(svafit$sv)
```

Note that the resulting estimated factors are not that different from the PCs:

```
for(i in 1:ncol(svafit$sv)){
  print( cor(s$v[,i],svafit$sv[,i]) )
}
```

Now fit a linear model to estimate the difference between males and females for each gene but that instead of accounting for batch effects using `month` it includes the factors estimated by `sva` in the model. Use the `qvalue()` function to estimate q-values.

How many genes have q-value < 0.1?

13

✓ **Answer:** 13

13

Explanation

```
library(qvalue)
library(sva)
X= model.matrix(~sex+svafit$sv)
pvals = sapply(1:nrow(geneExpression),function(i){
  y = geneExpression[i,]
  fit = lm(y~X-1)
  summary(fit)$coef[2,4]
})
qvals = qvalue(pvals)$qvalue
sum(qvals<0.1)
```

Submit

You have used 1 of 5 attempts

Answers are displayed within the problem

SVA Exercises #2

1/1 point (graded)

What proportion of the genes from SVA Exercises #1 are from chrY or chrX?

0.9230769

✓ **Answer:** 0.9230769

0.9230769

Explanation

```
index = geneAnnotation$CHR[qvals<0.1]%in%c("chrX","chrY")
mean(index)
```

Remember that we should always perform exploratory data analysis to check problems. For example, before reporting this list of genes we could look at a volcano plot like this:

```
res = sapply(1:nrow(geneExpression),function(i){
    y = geneExpression[i,]
    fit = lm(y~X-1)
    summary(fit)$coef[2,c(1,4)]
})

qvals = qvalue(res[2,])$qvalue
pcutoff = max( res[2,qvals < .1] )
library(rafalib)
mypar2(1,1)

plot(res[1,],-log10(res[2,]),xlab="M",ylab="log10 p-value")

ind = which(geneAnnotation$CHR=="chrY")
points(res[1,ind],-log10(res[2,ind]),col=1,pch=16)

ind = which(geneAnnotation$CHR=="chrX")
points(res[1,ind],-log10(res[2,ind]),col=2,pch=16)

abline(h=-log10(pcutoff))
legend("bottomleft",c("chrX","chrY"),col=c(2,1),pch=16)
```

Note that there are six genes (five on chrY and one on chrX) that stand out as having large effects and small q-values.

Submit

You have used 1 of 5 attempts

ℹ Answers are displayed within the problem