Homework due May 13, 2021 23:01 +03

For the dataset we have been working with, models do not help due to the almost perfect confounding. This is one reason we created the subset dataset:

```
library(GSE5859Subset)
data(GSE5859Subset)
```

Here we purposely confounded month and group (sex) but not completely:

```
sex = sampleInfo$group
month = factor( format(sampleInfo$date,"%m"))
table( sampleInfo$group, month)
```

Modeling Batch Effects Exercises #1

1/1 point (graded)

Using the functions `rowttests()` and `qvalue()` compare the two groups, in this case males and females coded in `sex`. Because this is a smaller dataset, which decreases our power, we will use a more lenient FDR cut-off of 10%.

How many gene have q-values less than 0.1?

59

✓ **Answer:** 59

59

Explanation

```
library(qvalue)
library(genefilter)
pvals = rowttests(geneExpression,factor(sampleInfo$g))$p.va
qvals = qvalue(pvals)$qvalues
sum(qvals<0.1)
```

Submit

You have used 1 of 5 attempts

ℹ Answers are displayed within the problem

Modeling Batch Effects Exercises #2

1/1 point (graded)

Note that `sampleInfo$group` here represents males and females. Thus we expect differences to be on chrY and, for genes that escape inactivation, chrX. Note that we do not expect many autosomal genes to be different between males and females. This gives us an opportunity to evaluate false and true positives with experimental data. For example, we evaluate results using the proportion genes of the list that are on chrX or chrY.

For the list of genes with q<0.1 calculated in Modeling Batch Effects Exercises #1, what proportion of genes are on chrX or chrY?

0.3389831 ✓ **Answer:** 0.3389831

0.3389831

Explanation

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

Submit You have used 1 of 5 attempts

ℹ Answers are displayed within the problem

Modeling Batch Effects Exercises #3

1/1 point (graded)

Now for the autosomal genes (not on chrX and chrY) for which q-value < 0.1 perform a t-test comparing samples processed in June to those processed in October.

What proportion of these have p-values < 0.05?

0.8717949

✓ **Answer:** 0.8717949

0.8717949

Explanation

```
index = which(qvals<0.1 & !geneAnnotation$CHR%in%c("chrX","
month = factor( format(sampleInfo$date,"%m"))
pvals = rowttests(geneExpression[index,],month)$p.value
mean(pvals<0.05)
```

Submit

You have used 1 of 5 attempts

Answers are displayed within the problem

Modeling Batch Effects Exercises #4

1/1 point (graded)

The above result shows that the great majority of the autosomal genes show differences due to processing data. This provides further evidence that confounding is resulting in false positives. So we are going to try to model the month effect to better estimate the sex effect. We are going to use a linear model.

Which of the following creates the appropriate design matrix?

○ `X = model.matrix(~sex+ethnicity)`

○ `X = cbind(sex,as.numeric(month))`

● `X = model.matrix(~sex+month)`

○ It can't be done with one line

✓

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Modeling Batch Effects Exercises #5

1/1 point (graded)

Now use the `X` defined above to fit a regression model using `lm` for each gene. Note that you can obtain p-values for estimated parameters using `summary()`. Here is an example:

```
X = model.matrix(~sex+month)
i = 234
y = geneExpression[i,]
fit = lm(y~X-1)
summary(fit)$coef
```

How many of the q-values for the group comparison are <0.1 now?

17 ✓ **Answer:** 17

17

Explanation

```
library(qvalue)
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)
```

Note the big drop from what we obtained without the correction.
Also note that this code is suboptimal. Later we will learn faster ways of making this computation.

Submit You have used 1 of 2

attempts

ℹ Answers are displayed within the problem

Modeling Batch Effects Exercises #6

1/1 point (graded)

With this new list, what proportion of these are chrX and chrY?

0.8823528 ✓ **Answer:** 0.8823529

0.8823528

Explanation

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

Note the big improvement.

Submit You have used 1 of 5 attempts

ℹ Answers are displayed within the problem

Modeling Batch Effects Exercises #7

1/1 point (graded)

Now, from the linear model in Modeling Batch Effects Exercises #6, extract the p-values related to the coefficient representing the October versus June differences using the same linear model.

How many of the q-values for the month comparison are < 0.1 now?

3170 ✓ **Answer:** 3170

3170

Explanation

```
pvals = sapply(1:nrow(geneExpression),function(i){
  y = geneExpression[i,]
  fit = lm(y~X)
  summary(fit)$coef[3,4]
})
qvals = qvalue(pvals)$qvalue
sum(qvals<0.1)
```

Note that this approach is basically the approach implemented by ComBat.

Submit

You have used 1 of 5 attempts

Answers are displayed within the problem