Homework due May 13, 2021 23:01 +03

Load the data for this gene expression dataset:

```
library(Biobase)
library(GSE5859)
data(GSE5859)
```

Note that this is the original dataset from which we selected the subset used in `GSE5859Subset`. You can obtain it from the genomicsclass GitHub repository:

```
library(devtools)
install_github("genomicsclass/GSE5859")
```

We can extract the gene expression data and sample information table using the Bioconductor functions `exprs()` and `pData()` like this:

```
geneExpression = exprs(e)
sampleInfo = pData(e)
```

Note that we will learn much more about how these Bioconductor functions work in later courses.

Confounding in Genomics Exercises #1

1/1 point (graded)

Familiarize yourself with the `sampleInfo` table. Note that some samples were processed at different times. This is an extraneous variable and should not affect the values in `geneExpression`. However, as we have seen in previous analyses, it does appear to have an effect, so we will explore this here.

You can extract the year from each date like this:

```
year = format(sampleInfo$date,"%y")
```

Note there are

```
length( unique(year) )
```

unique years for which we have data.

For how many of these years do we have more than one ethnicity represented?

2 ✔ **Answer:** 2

2

Explanation

```
tab = table(year,sampleInfo$ethnicity)
print(tab)
x = rowSums( tab != 0)
sum( x >= 2)
```

Submit You have used 2 of 5 attempts

Answers are displayed within the problem

Confounding in Genomics Exercises #2

1/1 point (graded)

Repeat the above exercise but now instead of year consider the month as well. Specifically, instead of the `year` variable defined above, use:

```
month.year = format(sampleInfo$date,"%m%y")
```

For what **proportion** of these `month.year` values do we have more than one ethnicity represented?

0.04761905 ✔ **Answer:** 0.04761905

0.04761905

Explanation

```
tab = table(month.year,sampleInfo$ethnicity)
print(tab)
x = rowSums( tab != 0)
mean( x >= 2)
```

Note that this implies that `month.year` and ethnicity are almost completely confounded. This means that it is hard to separate effects due to date from effects due to our outcome of interest.

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Confounding in Genomics Exercises #3

2/2 points (graded)

Perform a t-test (use `rowttests()` from the **genefilter** package) comparing CEU samples processed in 2002 to those processed in 2003. Then use the **qvalue** package to obtain q-values for each gene.

How many genes have q-values < 0.05?

4308

✔ **Answer:** 4308

4308

Explanation

```
library(qvalue)
library(genefilter)
year = factor( format(sampleInfo$date,"%y") )
index = which(year%in% c("02","03") & sampleInfo$ethnicity=
year = droplevels(year[index])
pval = rowttests(geneExpression[ ,index], year)$p.value
qval = qvalue(pval)
sum(qval$qvalue < 0.05)
```

What is the estimate of `pi0` provided by `qvalue()`?

0.3628642 ✔ **Answer:** 0.3628642

0.3628642

Explanation

```
print( qvalue(pval)$pi0 )
```

Note that the estimated percentage of genes that are differentially expressed is above 30%. This is one way to show the magnitude of the effect processing date has on the measurements.

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Confounding in Genomics Exercises #4

1/1 point (graded)

Now perform a t-test (use `rowttests()`) comparing CEU samples processed in 2003 to CEU samples processed in 2004. Then use the **qvalue** package to obtain q-values for each gene.

How many genes have q-values < 0.05?

2463 ✔ **Answer:** 2463

2463

Explanation

```
library(qvalue)
library(genefilter)
year = factor( format(sampleInfo$date,"%y") )
index = which(year%in% c("03","04")  & sampleInfo$ethnicity=="CEU")
year = droplevels(year[index])
pval = rowttests(geneExpression[ ,index], year)$p.value
qval = qvalue(pval)
sum(qval$qvalue < 0.05)
```

Here we confirm the processing date has an effect on our measurements.

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Confounding in Genomics Exercises #5

1/1 point (graded)

Now we are going to compare ethnicities as was done in the original publication in which these data were first presented. Use the `rowttests()` function to compare the ASN population to the CEU population. Once again, use the `qvalue()` function to obtain q-values.

How many genes have q-values < 0.05?

7217 ✔ **Answer:** 7217

7217

Explanation

```
library(qvalue)
library(genefilter)
index = which(sampleInfo$ethnicity%in% c("CEU","ASN"))
g = droplevels(sampleInfo$ethnicity[index])
pval = rowttests(geneExpression[ ,index], g)$p.value
qval = qvalue(pval)
sum(qval$qvalue < 0.05)
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Confounding in Genomics Exercises #6

1/1 point (graded)

Note that over 80% of genes are called differentially expressed between ethnic groups. However, due to the confounding with processing date, we need to confirm these differences are actually due to ethnicity. This will not be easy due to the almost perfect confounding. However, above we noted that two groups were represented in 2005. Just like we stratified by majors to remove the "major effect" in our admissions example, here we can stratify by year and perform a t-test comparing ASN and CEU, but only for samples processed in 2005.

How many genes have q-values < 0.05?

560 ✔ **Answer:** 560

560

Explanation

```
library(qvalue)
library(genefilter)
year = factor( format(sampleInfo$date,"%y") )
index = which(sampleInfo$ethnicity%in% c("CEU","ASN") & year=="05")
g = droplevels(sampleInfo$ethnicity[index])
pval = rowttests(geneExpression[ ,index], g)$p.value
qval = qvalue(pval)
sum(qval$qvalue < 0.05)
```

Note the dramatic drop in the number of genes with q-value < 0.05 when we fix the year. However, the sample size is much smaller in this latest analysis which means we have less power:

```
table(sampleInfo$ethnicity[index])
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Confounding in Genomics Exercises #7

1/1 point (graded)

To provide a more balanced comparison, we repeat the analysis but now by taking 3 random CEU samples from 2002. Repeat the analysis above but comparing the ASN from 2005 to three random CEU samples from 2002. Set the seed at 3, `set.seed(3)`, before random sampling.

How many genes have q-values < 0.05?

3943

✔ **Answer:** 3943

3943

Explanation

```
library(qvalue)
library(genefilter)
year = factor( format(sampleInfo$date,"%y") )
index1 = which(sampleInfo$ethnicity=="ASN" & year=="05")
set.seed(3)
index2 = sample( which(sampleInfo$ethnicity == "CEU" & year=
index = c( index1, index2)
g = droplevels(sampleInfo$ethnicity[index])
pval = rowttests(geneExpression[ ,index], g)$p.value
qval = qvalue(pval)
sum(qval$qvalue < 0.05)
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem