Exploration of 3 ReCount datasets

April 30, 2015

Contents

1 PCA plots

Below I show that there are strong batch effects in Bottomly and some evidence of batch effect in Cheung. In the following code, I first subset the 3 datasets to a single biological condition (one species, one sex, or one population group). I then make a PCA plot using just the shifted log of normalized counts. I then repeat this whole process using the other group of the biological condition.

For Cheung and Montgomery/Pickrell, we do not have batch information for library preparation or for the day of sequencing, although these samples were likely processed in many batches as there are so many samples. For the data of Cheung, I have tried to find a proxy for batch by using the label of the Illumina machine, which is printed in the FASTQ file in the names of the reads. However this is just an approximation. Some samples on the same machine do tend to cluster, while others do not.

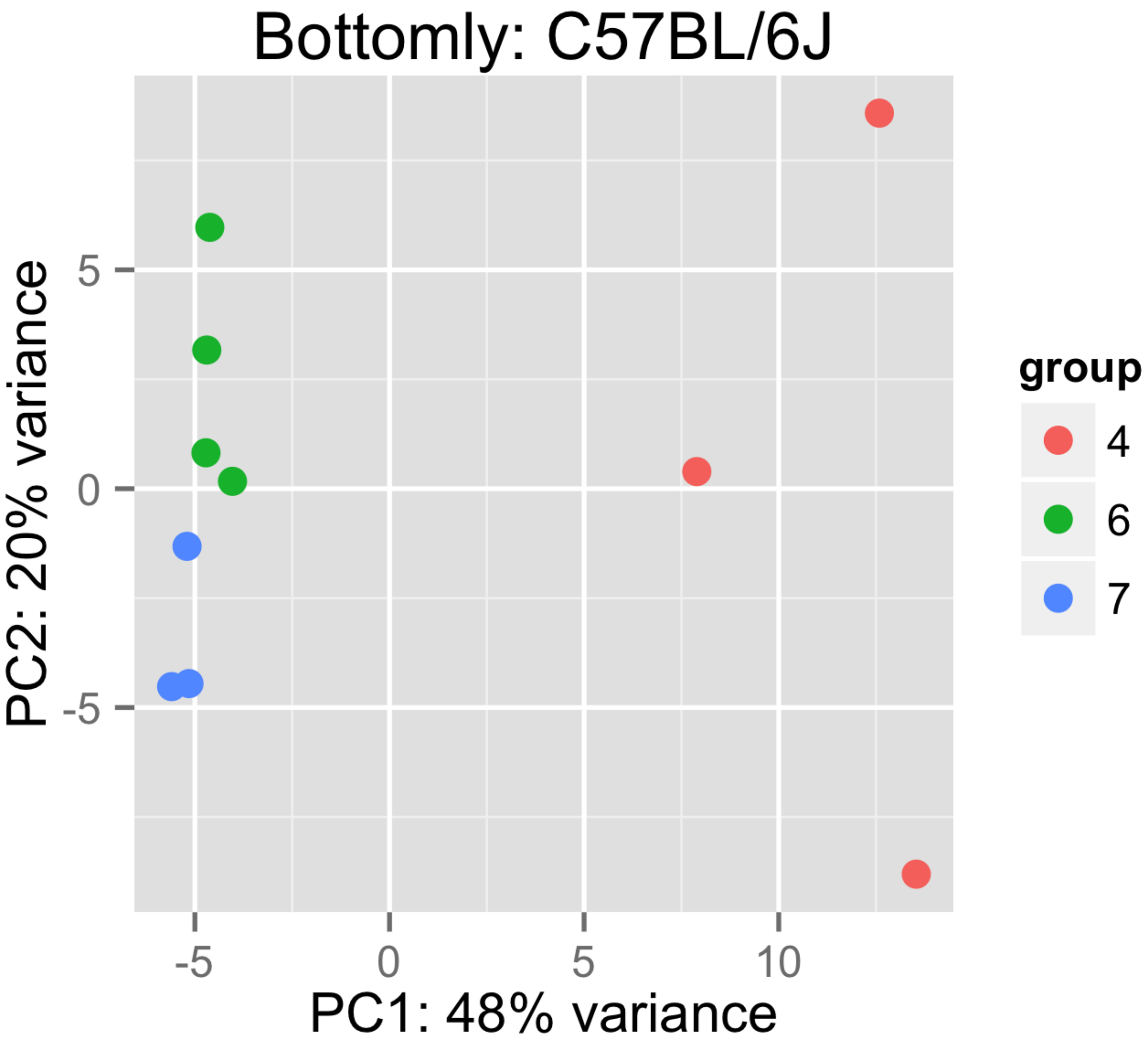
```
options(digits=3, width=100)
opts_chunk$set(tidy=FALSE, dev='png',
                fig.width=4, fig.height=4.5, fig.path="figure/rgraphics-",
                message=FALSE, error=FALSE, warning=FALSE, fig.align="center", dpi=360)
```

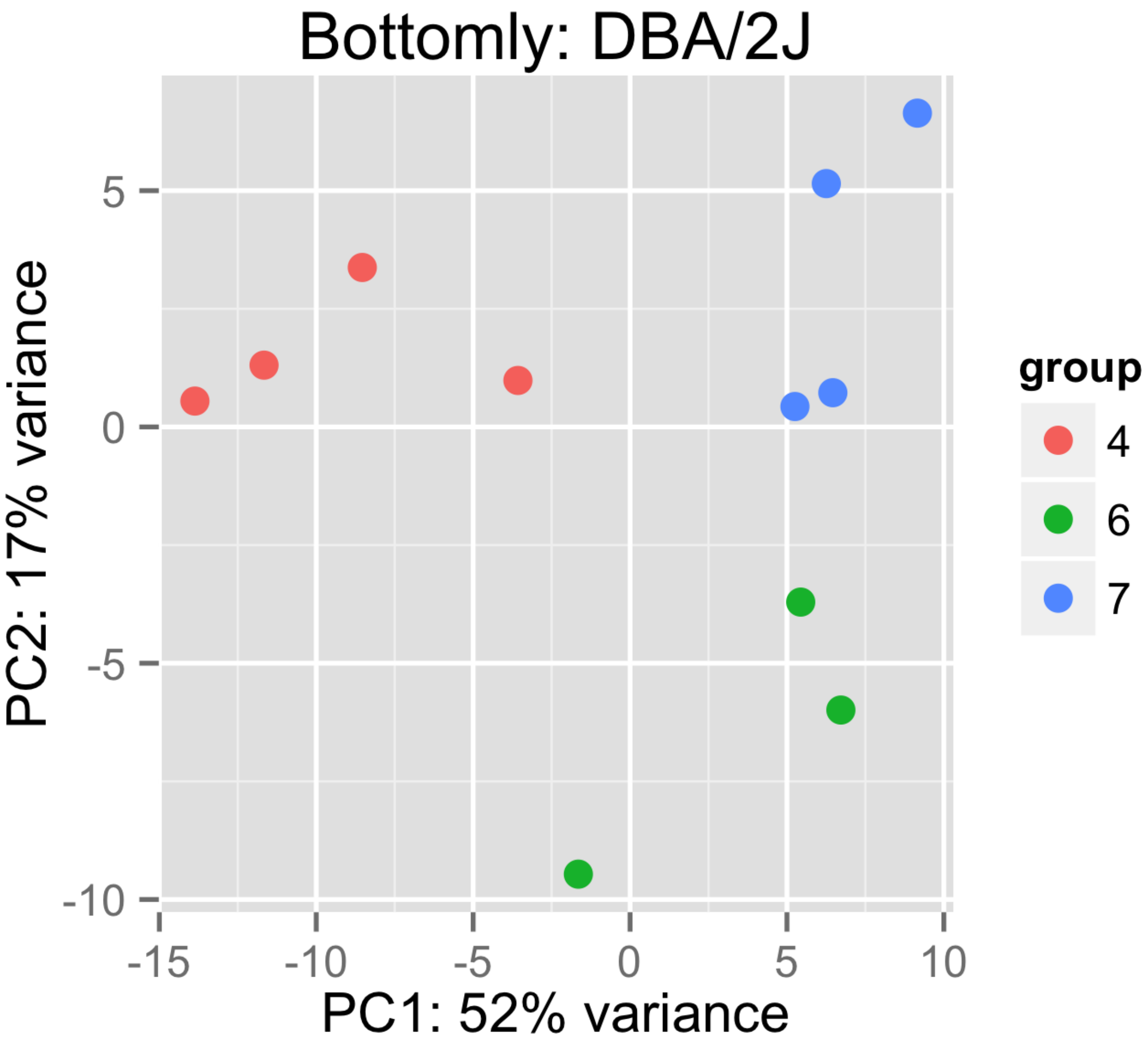
```
files <- c("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/cheung_eset.RData",
           "http://bowtie-bio.sourceforge.net/recount/ExpressionSets/montpick_eset.RData",
           "http://bowtie-bio.sourceforge.net/recount/ExpressionSets/bottomly_eset.RData")
for (file in files) if (!file.exists(basename(file))) download.file(file, basename(file))
for (file in files) load(basename(file))
library("DESeq2")
library("lsr")
library("Biobase")
library("ggplot2")
```

```
bottomly.eset <- bottomly.eset[ rowMeans(exprs(bottomly.eset)) >= 1,]
bottomly.eset$experiment.number <- factor(bottomly.eset$experiment.number)
dim(bottomly.eset)

## Features  Samples
##    11175       21

cheung.eset <- cheung.eset[ rowMeans(as.matrix(exprs(cheung.eset))) >= 1,]
cheung.batch <- read.delim("cheung_batch.txt", header=FALSE, stringsAsFactors=FALSE)
cheung.batch$machine <- factor(sapply(strsplit(cheung.batch$V2, ":"), `[`, 1))
cheung.batch$sub.machine <- factor(sapply(strsplit(as.character(cheung.batch$machine), "_")
cheung.batch$id <- paste0("NA",substr(cheung.batch$V1, 3, 8))
cheung.eset$machine <- cheung.batch$machine[match(colnames(cheung.eset), cheung.batch$id)]
cheung.eset$sub.machine <- cheung.batch$sub.machine[match(colnames(cheung.eset), cheung.bat
dim(cheung.eset)

## Features  Samples
##     9067       41

montpick.eset <- montpick.eset[ rowMeans(exprs(montpick.eset)) >= 1,]
dim(montpick.eset)

## Features  Samples
##     8124      129
```

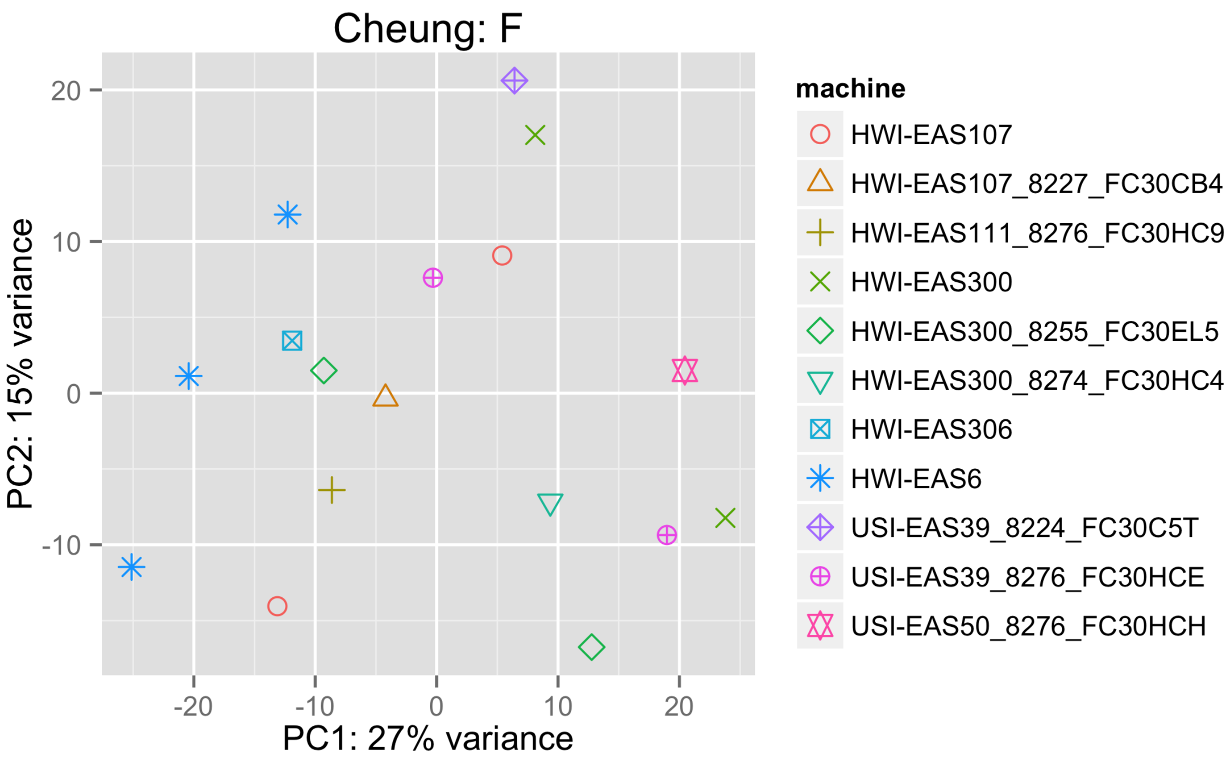
```
k <- 20
scale.shift.log <- function(x, k) {
  sf <- estimateSizeFactorsForMatrix(x)
  log2( t(t(x) / sf) + k )
}
```

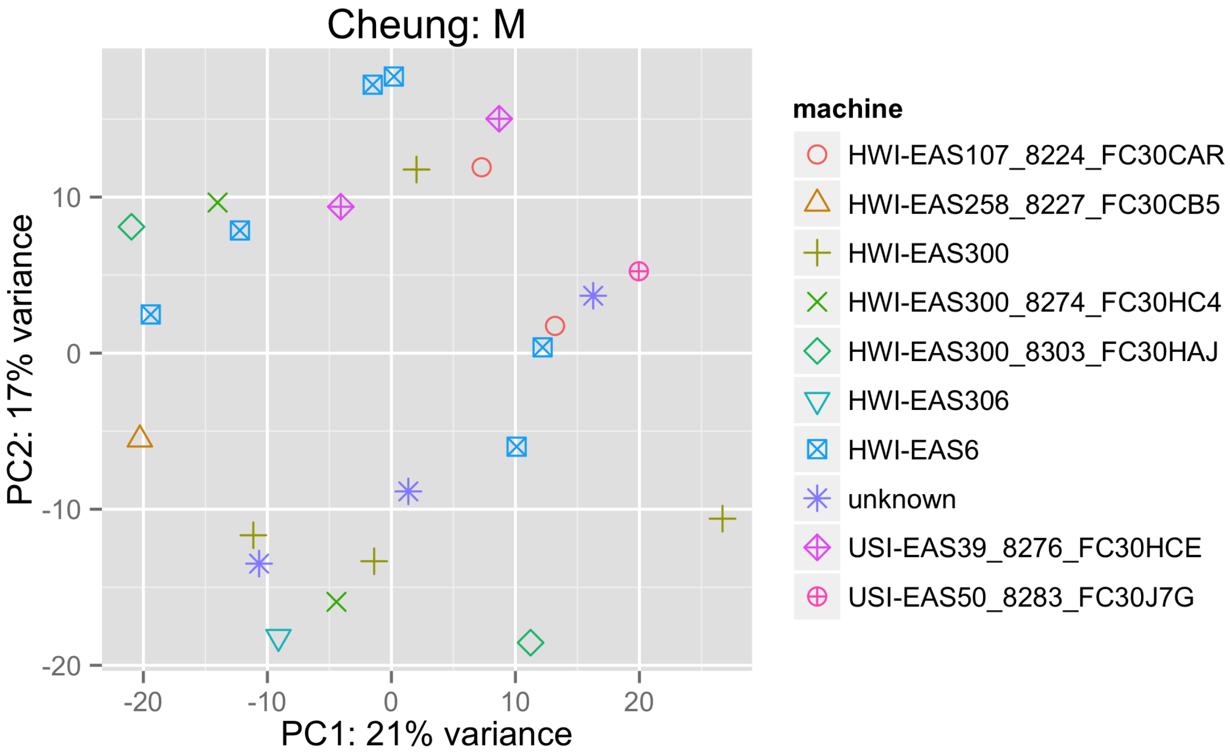
```
for (i in 1:2) {
  group <- levels(bottomly.eset$strain)[i]
  bottomly <- bottomly.eset[,bottomly.eset$strain == group]
  # just log transform, assign DESeqTransform class for plotPCA method
  # note: no NB-based transformation here
  bottomly.se <- DESeqTransform(SummarizedExperiment(assay=scale.shift.log(exprs(bottomly)
                                                     colData=DataFrame(pData(bottomly))))
  p <- plotPCA(bottomly.se, intgroup="experiment.number") + ggtitle(paste("Bottomly:",group
  print(p)
}
```


Bottomly: C57BL/6J
PC2: 20% variance
5
0
-5
-5
0
5
10
PC1: 48% variance
group
4
6
7

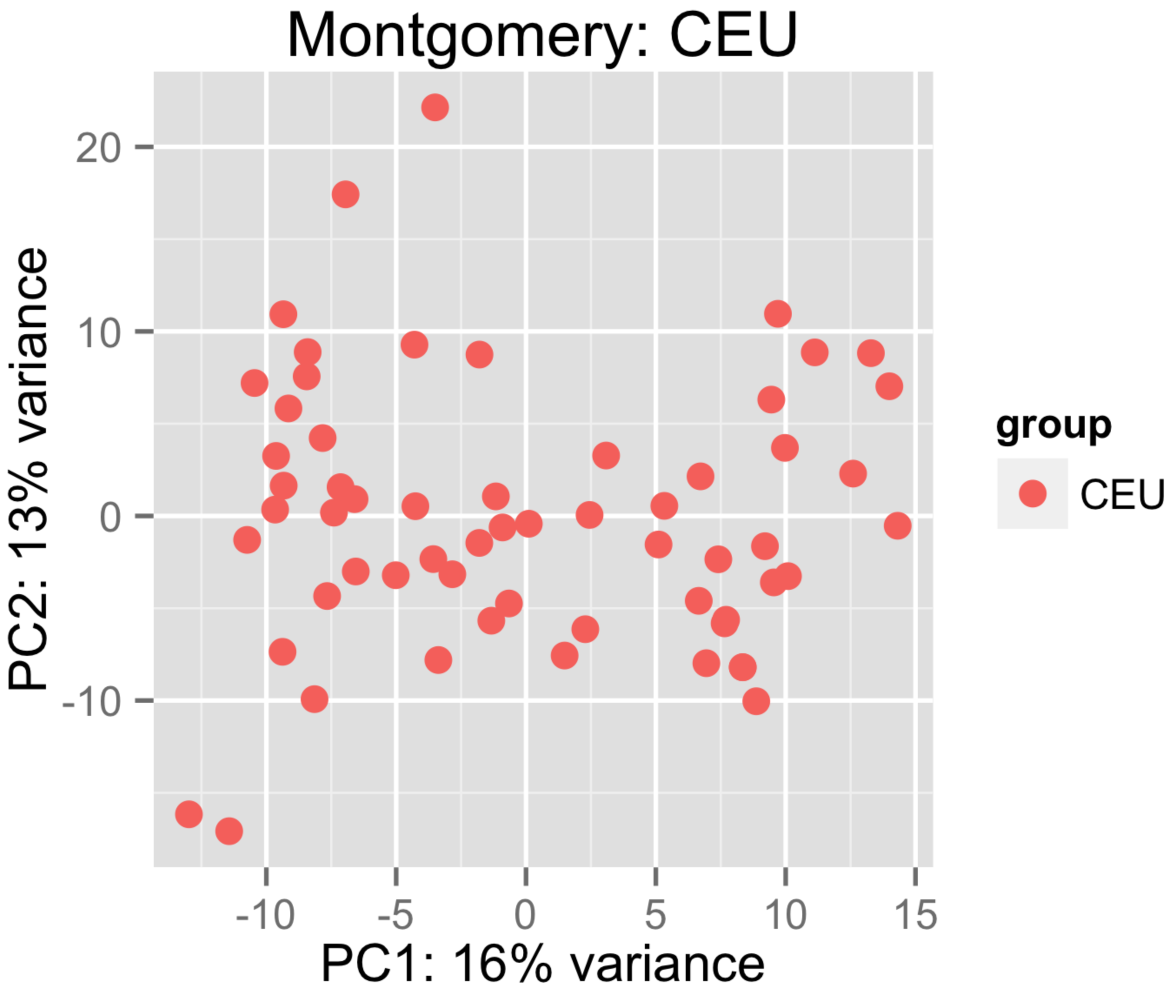

Bottomly: DBA/2J
PC2: 17% variance
PC1: 52% variance
5
0
-5
-10
-15
-10
-5
0
5
10
group
4
6
7

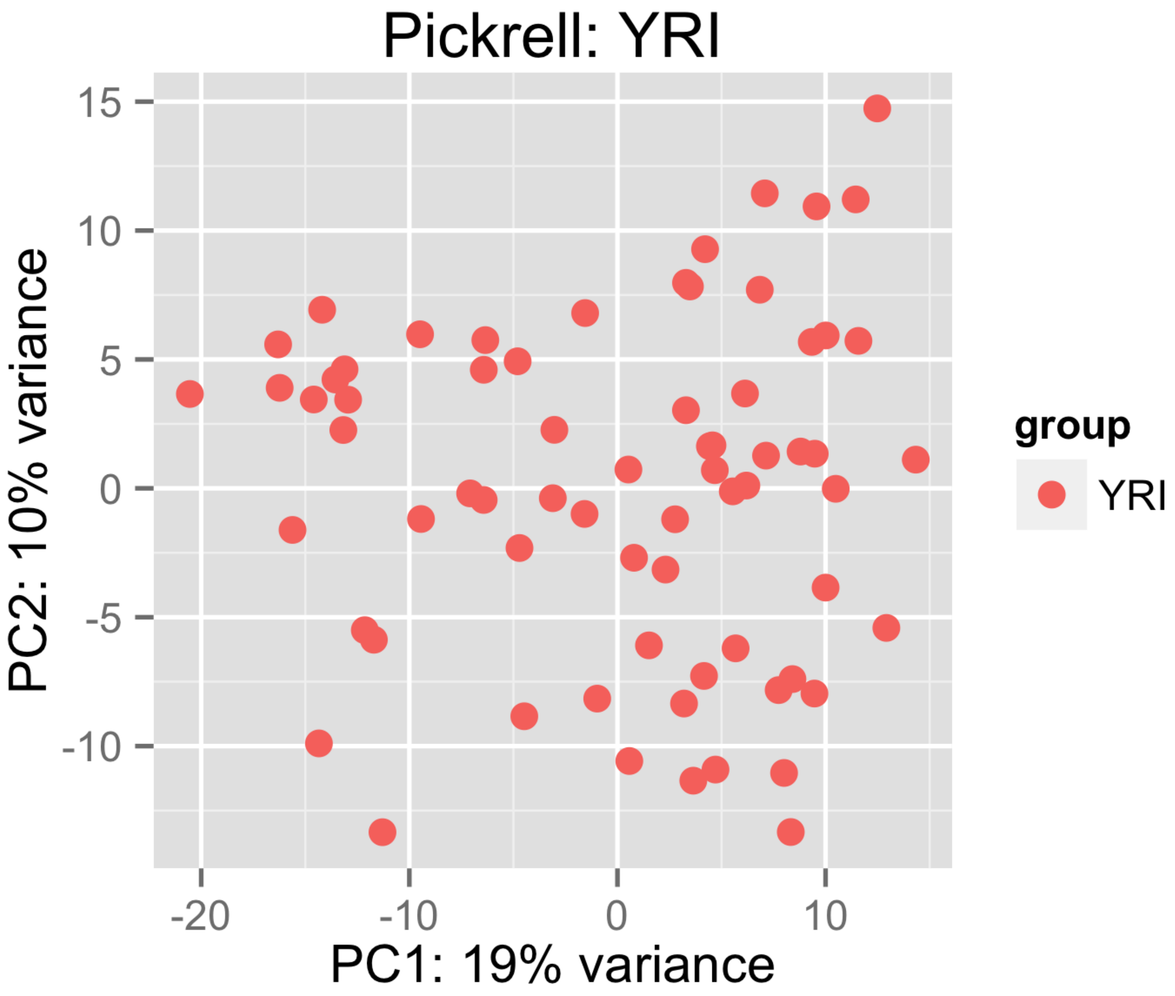
```
for (i in 1:2) {
  group <- levels(cheung.eset$gender)[i]
  cheung <- cheung.eset[,cheung.eset$gender == group]
  # just log transform, assign DESeqTransform class for plotPCA method
  # note: no NB-based transformation here
  cheung.se <- DESeqTransform(SummarizedExperiment(assay=scale.shift.log(as.matrix(exprs(ch
                                                 colData=DataFrame(pData(cheung))))
  data <- plotPCA(cheung.se, intgroup="machine", returnData=TRUE)
  p <- ggplot(data, aes(PC1, PC2, col=machine, shape=machine)) +
    scale_shape_manual(values=seq_len(nlevels(cheung.se$machine))) +
    geom_point(size=3) +
    ggtitle(paste("Cheung:",group)) +
    xlab(paste0("PC1: ",100*round(attr(data, "percentVar")[1],2),"% variance")) +
    ylab(paste0("PC2: ",100*round(attr(data, "percentVar")[2],2),"% variance"))
  print(p)
}
```




Cheung: M
PC2: 17% variance
PC1: 21% variance
10
0
-10
-20
-20
-10
0
10
20
machine
HWI-EAS107_8224_FC30CAR
HWI-EAS258_8227_FC30CB5
HWI-EAS300
HWI-EAS300_8274_FC30HC4
HWI-EAS300_8303_FC30HAJ
HWI-EAS306
HWI-EAS6
unknown
USI-EAS39_8276_FC30HCE
USI-EAS50_8283_FC30J7G

```
for (i in 1:2) {
  group <- levels(montpick.eset$population)[i]
  mont <- montpick.eset[,montpick.eset$population == group]
  study <- mont$study[1]
  # just log transform, assign DESeqTransform class for plotPCA method
  # note: no NB-based transformation here
  mont.se <- DESeqTransform(SummarizedExperiment(assay=scale.shift.log(exprs(mont),k),
                                                 colData=DataFrame(pData(mont))))
  p <- plotPCA(mont.se, intgroup="population") + ggtitle(paste0(study,": ",group))
  print(p)
}
```





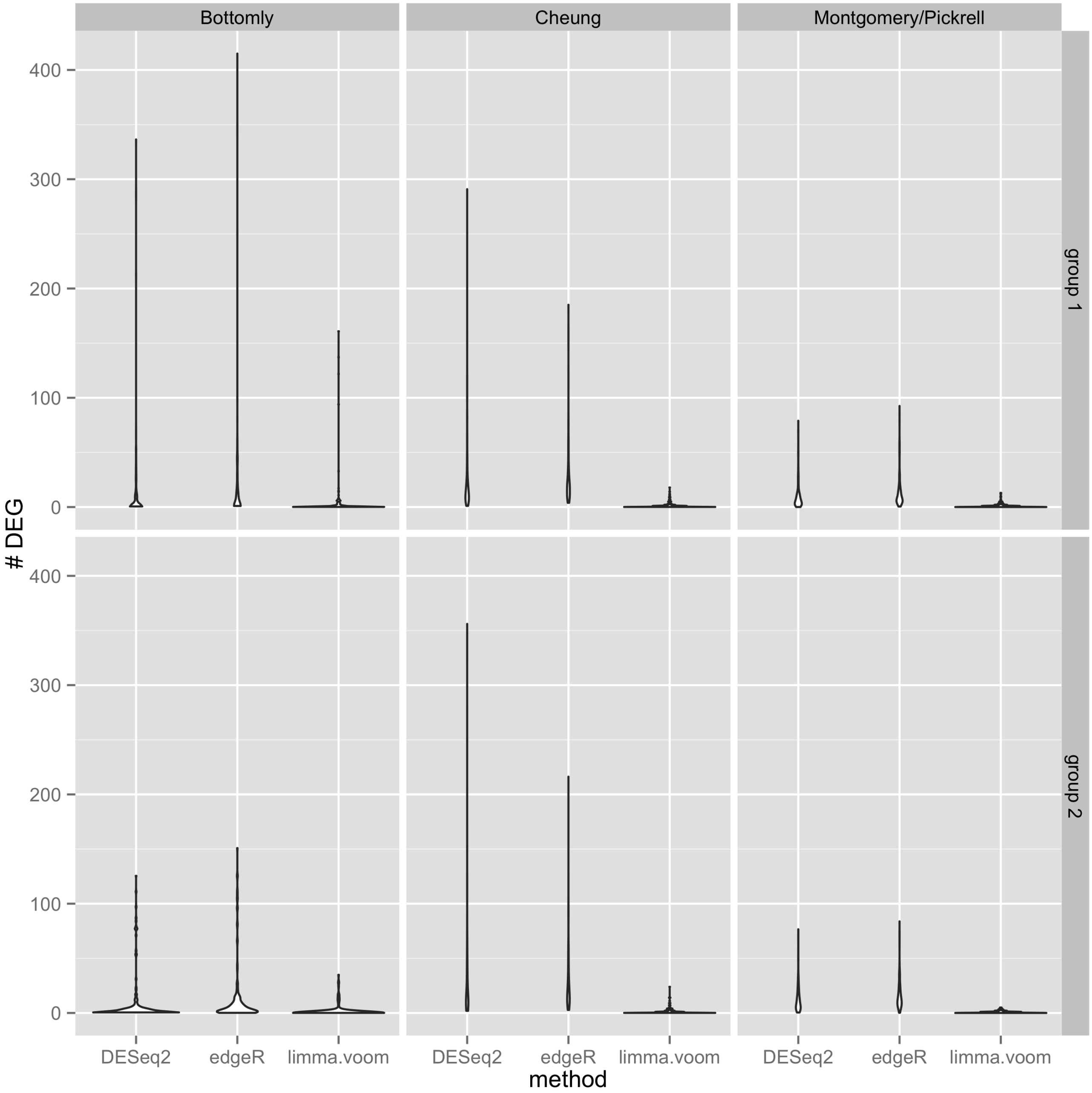
2 Permutation sum of p-value distributions

The following loads in the sum of p-values less than 10^{-4} for 200 random permutations. The code for this is available in `simulations.R`.

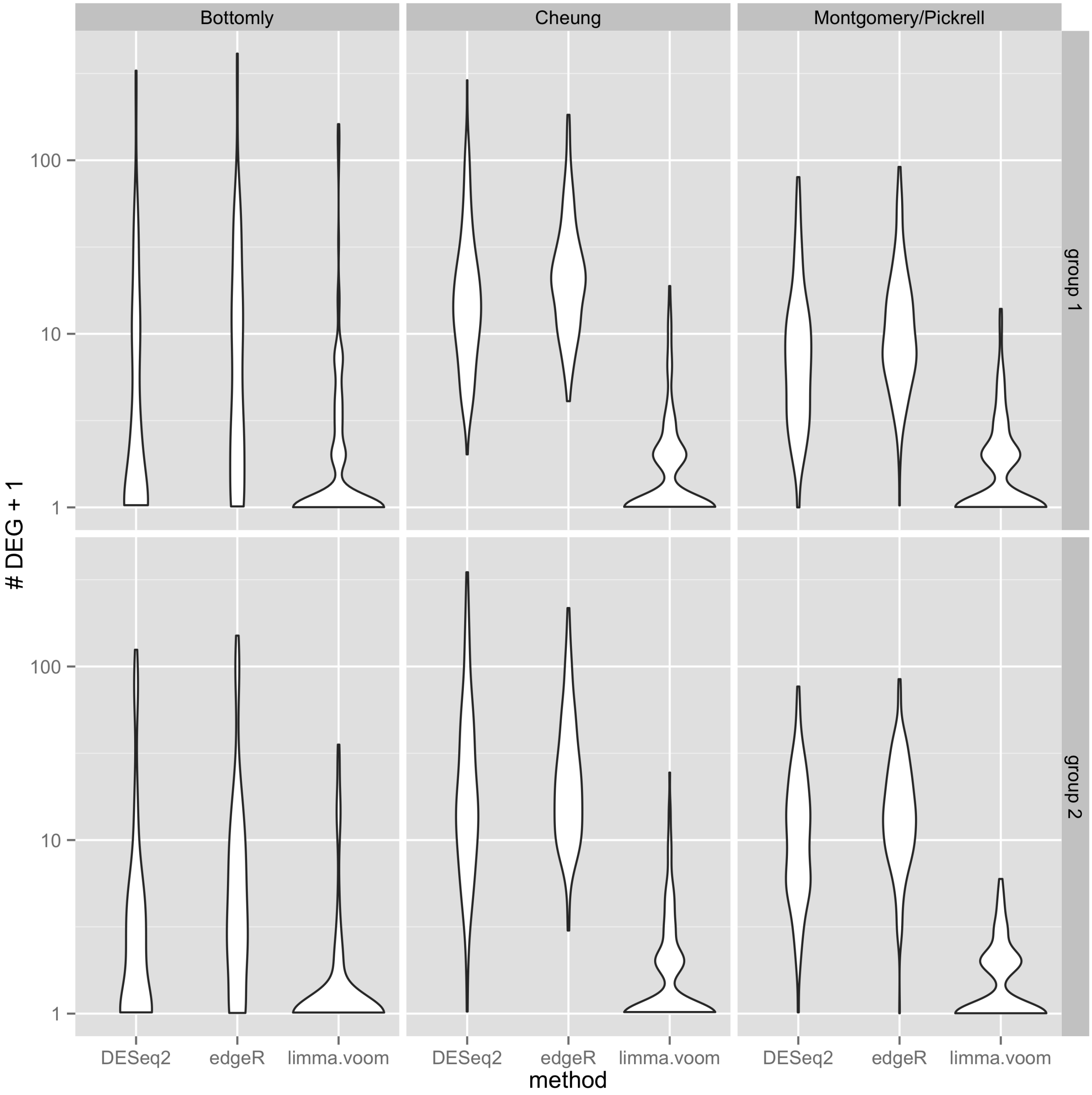
```
load("results.rda")
library("reshape")
```

```
meltFun <- function(res, group, dataset) {
  data <- melt(res[,1:3])
  colnames(data) <- c("method","nDEG")
  data$group <- group
  data$dataset <- dataset
  data
}
```

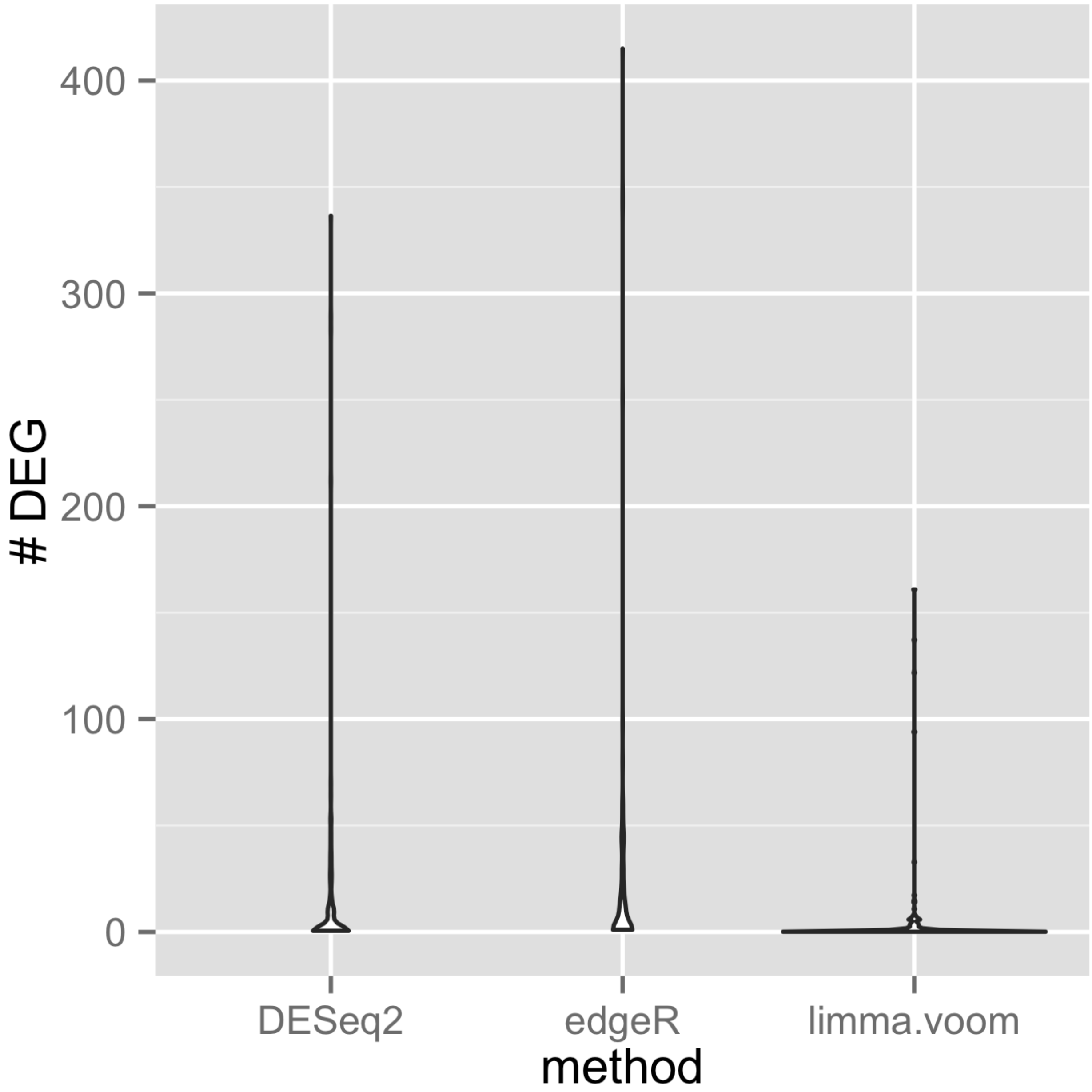
```
data <- list()
data[[1]] <- meltFun(res.b1, "group 1", "Bottomly")
data[[2]] <- meltFun(res.b2, "group 2", "Bottomly")
data[[3]] <- meltFun(res.c1, "group 1", "Cheung")
data[[4]] <- meltFun(res.c2, "group 2", "Cheung")
data[[5]] <- meltFun(res.m1, "group 1", "Montgomery/Pickrell")
data[[6]] <- meltFun(res.m2, "group 2", "Montgomery/Pickrell")
data.df <- do.call(rbind, data)
ggplot(data.df, aes(method, nDEG)) + geom_violin() +
  facet_grid(group ~ dataset) + ylab("# DEG")
```



```
ggplot(data.df, aes(method, nDEG+1)) + geom_violin() +
  scale_y_log10() + facet_grid(group ~ dataset) + ylab("# DEG + 1")
```



```
bottomly.df <- data.df[data.df$group == "group 1" & data.df$dataset == "Bottomly",]
ggplot(bottomly.df, aes(method, nDEG)) + geom_violin() + ylab("# DEG")
```



3 Check the arrangement of top permutations by number of DEG

```
i <- 1
bottomly <- bottomly.eset[,bottomly.eset$strain == levels(bottomly.eset$strain)[i]]
cheung <- cheung.eset[,cheung.eset$gender == levels(cheung.eset$gender)[i]]
mont <- montpick.eset[,montpick.eset$population == levels(montpick.eset$population)[i]]
m <- 6
set.seed(1) # master seed
```

```
nsim <- 200
seeds <- round(runif(nsim,1,1e6))
top.seeds <- head(order(-res.b1$DESeq2),10)
nDEG <- sort(res.b1$DESeq2, decreasing=TRUE)
perm.details <- data.frame(comparison=character(),nDEG=numeric(), stringsAsFactors=FALSE)
for (i in seq_along(top.seeds)) {
  s <- top.seeds[i]
  set.seed(seeds[s])
  sub <- sample(ncol(bottomly), m, FALSE)
  exp.num <- bottomly$experiment.number[sub]
  perm.format <- paste0("[",paste(sort(exp.num[1:3]),collapse=" "),"] vs [",paste(sort(exp
  perm.details[i,] <- c(perm.format, nDEG[i])
}
colnames(perm.details) <- c("comparison", "number of DEG")
perm.details

##              comparison number of DEG
## 1  [6 7 7] vs [4 4 4]           337
## 2  [4 4 4] vs [6 6 7]           289
## 3  [4 4 4] vs [6 6 7]           283
## 4  [6 6 7] vs [4 4 4]           212
## 5  [7 7 7] vs [4 4 6]            70
## 6  [6 7 7] vs [4 4 6]            64
## 7  [4 4 6] vs [6 6 7]            53
## 8  [6 6 7] vs [4 4 6]            53
## 9  [6 6 7] vs [4 4 6]            53
## 10 [6 7 7] vs [4 4 6]            48
```

```
\begin{kframe}
\begin{alltt}
\hlkwd{kable}\hlstd{(perm.details,} \hlkwc{format}\hlstd{=}\hlstr{"markdown"}\hlstd{)}
\end{alltt}
\end{kframe}

|comparison         |number of DEG |
|:------------------|:-------------|
|[6 7 7] vs [4 4 4] |337           |
|[4 4 4] vs [6 6 7] |289           |
|[4 4 4] vs [6 6 7] |283           |
|[6 6 7] vs [4 4 4] |212           |
|[7 7 7] vs [4 4 6] |70            |
|[6 7 7] vs [4 4 6] |64            |
|[4 4 6] vs [6 6 7] |53            |
|[6 6 7] vs [4 4 6] |53            |
|[6 6 7] vs [4 4 6] |53            |
|[6 7 7] vs [4 4 6] |48            |
```

3.1 Mean vs median number of DEG

```
\begin{kframe}
\begin{alltt}
\hlkwd{kable}\hlstd{(}\hlkwd{summary}\hlstd{(res.b1[,}\hlnum{1}\hlopt{:}\hlnum{4}\hlstd{])
\end{alltt}
\end{kframe}
```

```
|   |    DESeq2   |    edgeR    |  limma.voom   |   nonzero     |
|:--|:------------|:------------|:--------------|:--------------|
|   |Min.   :  0  |Min.   :  0  |Min.   :  0.0  |Min.   :11153  |
|   |1st Qu.:  0  |1st Qu.:  1  |1st Qu.:  0.0  |1st Qu.:11159  |
|   |Median :  1  |Median :  4  |Median :  0.0  |Median :11161  |
|   |Mean   : 12  |Mean   : 18  |Mean   :  3.9  |Mean   :11161  |
|   |3rd Qu.:  9  |3rd Qu.: 15  |3rd Qu.:  1.0  |3rd Qu.:11163  |
|   |Max.   :337  |Max.   :416  |Max.   :161.0  |Max.   :11167  |
```

```
summary(res.b1)

##      DESeq2          edgeR        limma.voom       nonzero          subset
##  Min.   :  0   Min.   :  0   Min.   :  0.0   Min.   :11153   Length:200
##  1st Qu.:  0   1st Qu.:  1   1st Qu.:  0.0   1st Qu.:11159   Class :character
##  Median :  1   Median :  4   Median :  0.0   Median :11161   Mode  :character
##  Mean   : 12   Mean   : 18   Mean   :  3.9   Mean   :11161
##  3rd Qu.:  9   3rd Qu.: 15   3rd Qu.:  1.0   3rd Qu.:11163
##  Max.   :337   Max.   :416   Max.   :161.0   Max.   :11167

summary(res.b2)

##      DESeq2            edgeR          limma.voom       nonzero          subset
##  Min.   :  0.0   Min.   :  0.0   Min.   : 0.0   Min.   :11158   Length:200
##  1st Qu.:  0.0   1st Qu.:  1.0   1st Qu.: 0.0   1st Qu.:11164   Class :character
##  Median :  1.0   Median :  3.0   Median : 0.0   Median :11165   Mode  :character
##  Mean   :  7.5   Mean   : 11.4   Mean   : 1.4   Mean   :11165
##  3rd Qu.:  4.0   3rd Qu.:  9.0   3rd Qu.: 0.0   3rd Qu.:11166
##  Max.   :126.0   Max.   :151.0   Max.   :35.0   Max.   :11169

summary(res.c1)

##      DESeq2          edgeR          limma.voom        nonzero         subset
##  Min.   :  1   Min.   :  3.0   Min.   : 0.00   Min.   :9016   Length:200
##  1st Qu.:  7   1st Qu.: 11.0   1st Qu.: 0.00   1st Qu.:9042   Class :character
##  Median : 13   Median : 19.5   Median : 0.00   Median :9047   Mode  :character
##  Mean   : 23   Mean   : 28.0   Mean   : 1.08   Mean   :9046
##  3rd Qu.: 25   3rd Qu.: 31.0   3rd Qu.: 1.00   3rd Qu.:9052
##  Max.   :291   Max.   :186.0   Max.   :18.00   Max.   :9061
```

```
summary(res.c2)

##      DESeq2        edgeR         limma.voom       nonzero         subset
##  Min.   :  0   Min.   :  2.0   Min.   : 0.00   Min.   :9025   Length:200
##  1st Qu.:  7   1st Qu.: 10.0   1st Qu.: 0.00   1st Qu.:9050   Class :character
##  Median : 15   Median : 19.0   Median : 0.00   Median :9055   Mode  :character
##  Mean   : 35   Mean   : 31.4   Mean   : 1.18   Mean   :9053
##  3rd Qu.: 39   3rd Qu.: 38.2   3rd Qu.: 1.00   3rd Qu.:9059
##  Max.   :358   Max.   :217.0   Max.   :24.00   Max.   :9065

summary(res.m1)

##      DESeq2         edgeR        limma.voom       nonzero         subset
##  Min.   : 0.0   Min.   : 0.0   Min.   : 0.00   Min.   :7943   Length:200
##  1st Qu.: 2.0   1st Qu.: 5.0   1st Qu.: 0.00   1st Qu.:8044   Class :character
##  Median : 6.0   Median : 8.5   Median : 0.00   Median :8082   Mode  :character
##  Mean   :10.2   Mean   :13.5   Mean   : 0.99   Mean   :8070
##  3rd Qu.:12.0   3rd Qu.:16.0   3rd Qu.: 1.00   3rd Qu.:8100
##  Max.   :79.0   Max.   :93.0   Max.   :13.00   Max.   :8118

summary(res.m2)

##      DESeq2       edgeR        limma.voom       nonzero         subset
##  Min.   : 0   Min.   : 0.0   Min.   :0.00   Min.   :8102   Length:200
##  1st Qu.: 4   1st Qu.: 8.0   1st Qu.:0.00   1st Qu.:8114   Class :character
##  Median : 9   Median :13.0   Median :0.00   Median :8116   Mode  :character
##  Mean   :13   Mean   :16.5   Mean   :0.66   Mean   :8116
##  3rd Qu.:18   3rd Qu.:21.0   3rd Qu.:1.00   3rd Qu.:8118
##  Max.   :77   Max.   :84.0   Max.   :5.00   Max.   :8122
```

```
# session info for the PCA plots
toLatex(sessionInfo())
```

- R version 3.2.0 Patched (2015-04-26 r68264), `x86_64-apple-darwin10.8.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: Biobase 2.28.0, BiocGenerics 0.14.0, BiocInstaller 1.18.1, DESeq2 1.8.0, devtools 1.7.0, GenomeInfoDb 1.4.0, GenomicRanges 1.20.3, ggplot2 1.0.1, IRanges 2.2.1, knitr 1.10, lsr 0.5, Rcpp 0.11.5, RcppArmadillo 0.5.000.0, reshape 0.8.5, S4Vectors 0.6.0, testthat 0.9.1
- Loaded via a namespace (and not attached): acepack 1.3-3.3, annotate 1.46.0, AnnotationDbi 1.30.1, BiocParallel 1.2.1, cluster 2.0.1, codetools 0.2-11, colorspace 1.2-6, compiler 3.2.0, DBI 0.3.1, digest 0.6.8, evaluate 0.7, foreign 0.8-63, formatR 1.2,

Formula 1.2-1, futile.logger 1.4.1, futile.options 1.0.0, genefilter 1.50.0, geneplotter 1.46.0, grid 3.2.0, gtable 0.1.2, highr 0.5, Hmisc 3.15-0, labeling 0.3, lambda.r 1.1.7, lattice 0.20-31, latticeExtra 0.6-26, locfit 1.5-9.1, MASS 7.3-40, munsell 0.4.2, nnet 7.3-9, plyr 1.8.2, proto 0.3-10, RColorBrewer 1.1-2, reshape2 1.4.1, rpart 4.1-9, RSQLite 1.0.0, scales 0.2.4, splines 3.2.0, stringr 0.6.2, survival 2.38-1, tools 3.2.0, XML 3.98-1.1, xtable 1.7-4, XVector 0.8.0

```r
# session info for the p-value generation
toLatex(session.info)
```

- R version 3.1.2 (2014-10-31), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=en_US.UTF-8`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, splines, stats, stats4, utils
- Other packages: Biobase 2.26.0, BiocGenerics 0.12.1, BiocInstaller 1.16.2, BiocParallel 1.0.3, DESeq2 1.6.3, devtools 1.7.0, edgeR 3.8.6, GenomeInfoDb 1.2.4, GenomicRanges 1.18.4, IRanges 2.0.1, knitr 1.9, limma 3.22.7, Rcpp 0.11.5, RcppArmadillo 0.4.650.1.1, S4Vectors 0.4.0
- Loaded via a namespace (and not attached): acepack 1.3-3.3, annotate 1.44.0, AnnotationDbi 1.28.2, base64enc 0.1-2, BatchJobs 1.6, BBmisc 1.9, brew 1.0-6, checkmate 1.5.1, cluster 2.0.1, codetools 0.2-11, colorspace 1.2-6, DBI 0.3.1, digest 0.6.8, evaluate 0.5.5, fail 1.2, foreach 1.4.2, foreign 0.8-63, formatR 1.0, Formula 1.2-0, genefilter 1.48.1, geneplotter 1.44.0, ggplot2 1.0.1, grid 3.1.2, gtable 0.1.2, Hmisc 3.15-0, iterators 1.0.7, lattice 0.20-30, latticeExtra 0.6-26, locfit 1.5-9.1, MASS 7.3-39, munsell 0.4.2, nnet 7.3-9, plyr 1.8.1, proto 0.3-10, RColorBrewer 1.1-2, reshape2 1.4.1, rpart 4.1-9, RSQLite 1.0.0, scales 0.2.4, sendmailR 1.2-1, stringr 0.6.2, survival 2.38-1, tools 3.1.2, XML 3.98-1.1, xtable 1.7-4, XVector 0.6.0