Biological vs. Technical Variability

Contents

1 Variability in data

Some variability in the data is indeed biological variability, while some are due to technical stuff. In this session we want to check for the difference between the two.

The sample data includes RNA from 12 randomly selected mice from two strains, and two pools with the RNA from all twelve mice from each of the two strains.

```
# Import packages and sample data
library(Biobase)
library(maPooling)
data(maPooling)
```

```
# Extract and illustrate pheno data to know which mice were included in which samples
pd <- pData(maPooling)
pd <- rbind(as.numeric(grepl("b", colnames(pd))), pd)
rownames(pd)[1] <- "strain"
```

Note: Each row represents a sample and the columns are the mice. The first row represents the strain. A "1" in cell i,j indicates that RNA from mouse j was included in sample i.

```
# Identifying pooled data rows
pooled <- data.frame(which(rowSums(pd[-1,]) == 12))
pooled <- cbind(as.numeric(grepl("b", rownames(pooled))), pooled)
colnames(pooled) <- c("strain" , "index")
pooled
```

```
##       strain index
## aq         0    25
## aqtr1      0    26
## aqtr2      0    27
## aqtr3      0    28
## bq         1    53
## bqtr1      1    54
## bqtr2      1    55
## bqtr3      1    56
```

```
# Compare the mean expression between groups for all genes
pooled_y <- exprs(maPooling[, rownames(pooled)])
pooled_g <- factor(pooled[, 1])

# t-test
library(genefilter)
pooled_tt <- rowttests(pooled_y, pooled_g)

# Check the p-values for the first five genes
five_genes_pooled <-
  data.frame(cbind(rownames(pooled_tt)[1:5], pooled_tt$p.value[1:5]))
colnames(five_genes_pooled) <- c("gene_id", "p-value")
five_genes_pooled
```

```
##      gene_id            p-value
## 1 1367452_at  0.114082582805314
## 2 1367453_at 0.0350608093107877
## 3 1367454_at  0.389086844408676
## 4 1367455_at  0.505790072963956
## 5 1367456_at  0.429254146958914
```

```r
# Identifying individual data rows
individuals <- data.frame(which(rowSums(pd[-1,]) == 1))
individuals <- cbind(as.numeric(grepl("b", rownames(individuals))), individuals)
colnames(individuals) <- c("strain" , "index")
individuals
```

```
##       strain index
## a10        0     1
## a11        0     4
## a12        0     5
## a14        0     8
## a2         0     9
## a3         0    12
## a3tr1      0    13
## a3tr2      0    14
## a4         0    15
## a5         0    17
## a6         0    19
## a7         0    21
## a8         0    22
## a9         0    24
## b10        1    29
## b11        1    32
## b12        1    33
## b13        1    36
## b14        1    38
## b15        1    39
## b2         1    40
## b3         1    43
## b3tr1      1    44
## b3tr2      1    45
## b5         1    46
## b6         1    48
## b8         1    50
## b9         1    52
```

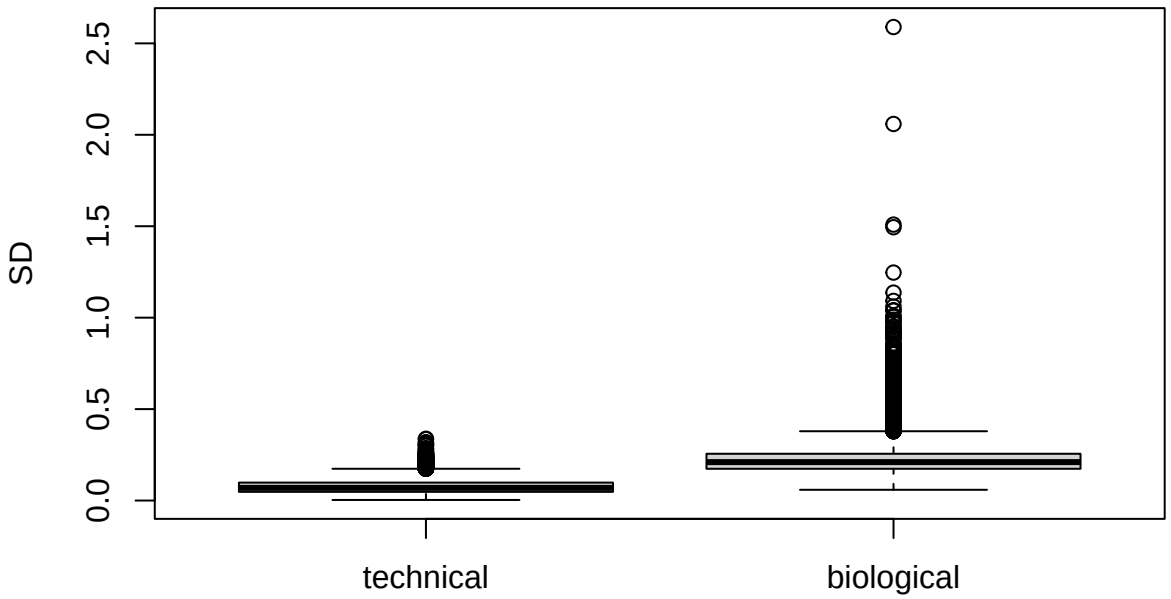
```r
# Remove samples including technical replicates (tr)
individuals[-c(grep("tr", rownames(individuals))),]
```

```
##     strain index
## a10      0     1
## a11      0     4
## a12      0     5
## a14      0     8
## a2       0     9
## a3       0    12
## a4       0    15
## a5       0    17
## a6       0    19
## a7       0    21
## a8       0    22
## a9       0    24
## b10      1    29
## b11      1    32
## b12      1    33
```

```
## b13      1    36
## b14      1    38
## b15      1    39
## b2       1    40
## b3       1    43
## b5       1    46
## b6       1    48
## b8       1    50
## b9       1    52
```

```
ind_y <- exprs(maPooling[, rownames(individuals)])
ind_g <- factor(individuals[, 1])
```

```
# Compare variabilities
technicalsd <- rowSds(pooled_y[, pooled_g == 0])
biologicalsd <- rowSds(ind_y[, ind_g == 0])
boxplot(
  technicalsd,
  biologicalsd,
  names = c("technical", "biological"),
  ylab = "SD"
)
```




```
# Compare the mean expression between groups for all genes
# t-test
ind_tt <- rowttests(ind_y, ind_g)

# Check the p-values for the first five genes
five_genes_ind <- data.frame(cbind(rownames(ind_tt)[1:5],
                                   ind_tt$p.value[1:5]))
colnames(five_genes_ind) <- c("gene_id", "p-value")
five_genes_ind
```

```
##      gene_id           p-value
## 1 1367452_at 0.311833234458842
## 2 1367453_at 0.566594933250323
```

```
## 3 1367454_at 0.931460533492937
## 4 1367455_at 0.235577228169978
## 5 1367456_at  0.96660255016996
```

```
# Compare p-values between the two models
five_genes <- data.frame(cbind(five_genes_pooled[, 2], five_genes_ind[, 2]))
colnames(five_genes) <- c("pooled", "individuals")
five_genes
```

```
##               pooled       individuals
## 1  0.114082582805314 0.311833234458842
## 2 0.0350608093107877 0.566594933250323
## 3  0.389086844408676 0.931460533492937
## 4  0.505790072963956 0.235577228169978
## 5  0.429254146958914  0.96660255016996
```