edgeR

2024-03-26

edgeR Tutorial

We are using raw read counts, not normalized counts as input to edgeR.

```
cancer_counts <- read.table(file="oral_carcinoma_counts_2018.txt", sep =
"\t", header = T)
head(cancer_counts)
```

```
##           ENSEMBL     N8     T8    N33    T33     N51    T51
## 1 ENSG00000251562 306305 330105 473438 309917  712348 633871
## 2 ENSG00000155657 328503   1204 206612   3178 1675945 191624
## 3 ENSG00000199753  62098  73284 581364 365430  205994 106640
## 4 ENSG00000275996  56374  56594 554127 260275  229356 135146
## 5 ENSG00000276788  95410 181223 394803 209376  249091 131438
## 6 ENSG00000171401 393801   2291 359693 106059  211919   1833
```

ID Mapping

Map the ENSEMBL gene identifiers to Entrez Gene IDs

```
mapping <-AnnotationDbi::select(org.Hs.eg.db, as.character(cancer_counts$ENSEMBL), keytype = "ENSEMBL",
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
head(mapping)
```

```
##           ENSEMBL ENTREZID
## 1 ENSG00000251562   378938
## 2 ENSG00000155657     7273
## 3 ENSG00000199753   692227
## 4 ENSG00000275996     9301
## 5 ENSG00000276788     9302
## 6 ENSG00000171401     3860
```

Remove duplicates

Use the duplicated function to deduplicate the rows in the mapping table

```
d <- duplicated(mapping$ENSEMBL)
sum(d)
```

```
## [1] 53
```

```
mapping <- mapping[!d,]
```

Merge mapping and cancer counts

```
cancer_counts <- merge(cancer_counts, mapping, by = "ENSEMBL")
head(cancer_counts)
```

```
##           ENSEMBL   N8  T8 N33 T33  N51 T51 ENTREZID
## 1 ENSG00000000003  217  79 264 267  240 339     7105
## 2 ENSG00000000419  135 376 294 487  332 456     8813
## 3 ENSG00000000457  126  95 276 207  391 303    57147
## 4 ENSG00000000460   54  76 133 151  179 154    55732
## 5 ENSG00000000971 1082  30 916 175 2546 608     3075
## 6 ENSG00000001036   67 105 138 122  424 199     2519
```

Remove missing data and duplicate ENSEMBL IDs

```
missing <- is.na(cancer_counts$ENTREZID)
sum(missing)
```

```
## [1] 22
```

```
cancer_counts <- cancer_counts[!missing,]
o <- order(rowSums(cancer_counts[,c(2:7)]), decreasing=TRUE)
cancer_counts <- cancer_counts[o,]
d2 <- duplicated(cancer_counts$ENTREZID)
sum(d2)
```

```
## [1] 0
```

```
cancer_counts <- cancer_counts[!d2,]
```

##Create new mapping table

```
mapping2 <- AnnotationDbi::select(org.Hs.eg.db, as.character(cancer_counts$ENTREZID),
keytype = "ENTREZID", column="SYMBOL")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```r
d3 <- duplicated(mapping2$ENTREZID)
sum(d3)
```

```
## [1] 0
```

```r
#Remove duplicates
mapping2 <- mapping2[!d3,]

#Merge the mapping2 and cancer counts table
cancer_counts <- merge(cancer_counts, mapping2, by= "ENTREZID")
head(cancer_counts)
```

```
##    ENTREZID         ENSEMBL  N8 T8  N33 T33 N51 T51      SYMBOL
## 1     10000 ENSG00000117020 207 54  131 148 823 222        AKT3
## 2     10001 ENSG00000133997 108 98  141 210 181 263        MED6
## 3 100033414 ENSG00000207001 152 15 1186 337 563 154  SNORD116-2
## 4 100033418 ENSG00000207442  66  5  600 188 243  67  SNORD116-6
## 5 100033420 ENSG00000207093 147 59  911 268 443 250  SNORD116-8
## 6 100033434 ENSG00000207375  56 15  994 387 102  58 SNORD116-23
```

```r
# Column 1 = ENTREZID
# Column 2= ENSEMBL
# Column 9 = Symbol
```

MDS plot

```r
y <- DGEList(counts=cancer_counts[,3:8], genes=cancer_counts[,c(1:2,9)])
head(y$genes)
```

```
##    ENTREZID         ENSEMBL      SYMBOL
## 1     10000 ENSG00000117020        AKT3
## 2     10001 ENSG00000133997        MED6
## 3 100033414 ENSG00000207001  SNORD116-2
## 4 100033418 ENSG00000207442  SNORD116-6
## 5 100033420 ENSG00000207093  SNORD116-8
## 6 100033434 ENSG00000207375 SNORD116-23
```

```r
head(y$samples)
```

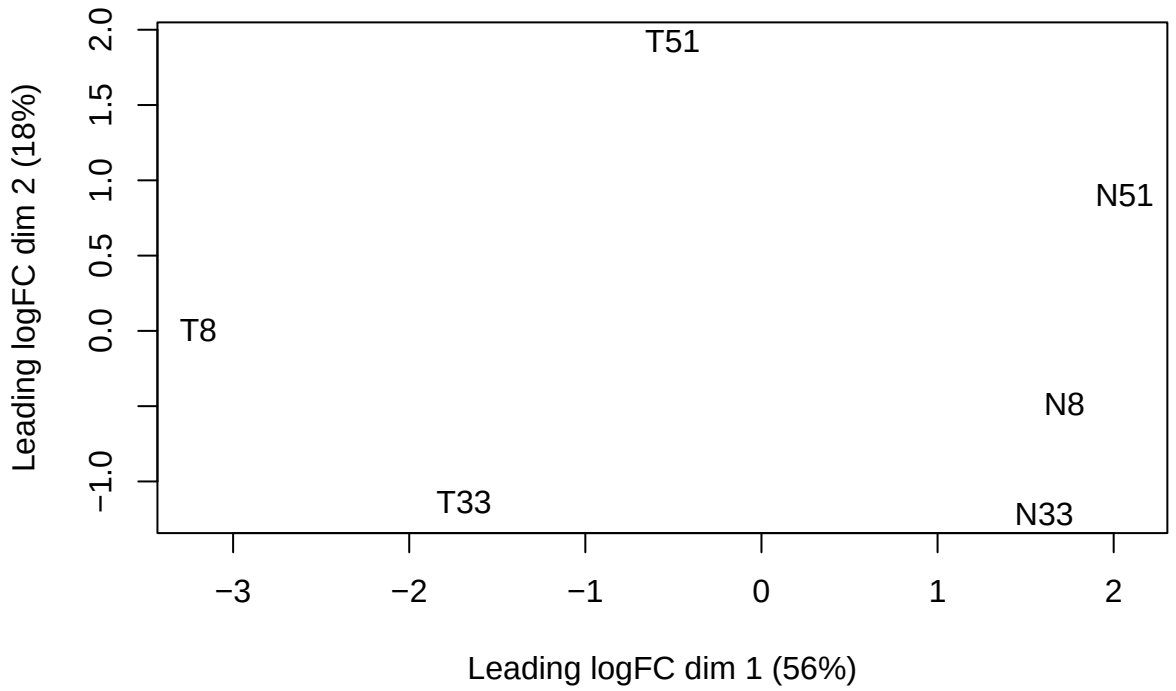
```
##     group lib.size norm.factors
## N8      1  7781155            1
## T8      1  7064702            1
## N33     1 14535010            1
## T33     1 12937623            1
## N51     1 21006218            1
## T51     1 14641637            1
```

```
head(y$counts)
```

```
##    N8 T8  N33 T33 N51 T51
## 1 207 54  131 148 823 222
## 2 108 98  141 210 181 263
## 3 152 15 1186 337 563 154
## 4  66  5  600 188 243  67
## 5 147 59  911 268 443 250
## 6  56 15  994 387 102  58
```

```
rownames(y$counts) <- rownames(y$genes) <- y$genes$ENTREZID

y$genes$ENTREZID <- NULL
y <- calcNormFactors(y)
y$samples$group = c("N", "T", "N", "T", "N", "T")
plotMDS(y)
```



Dispersion and BCV

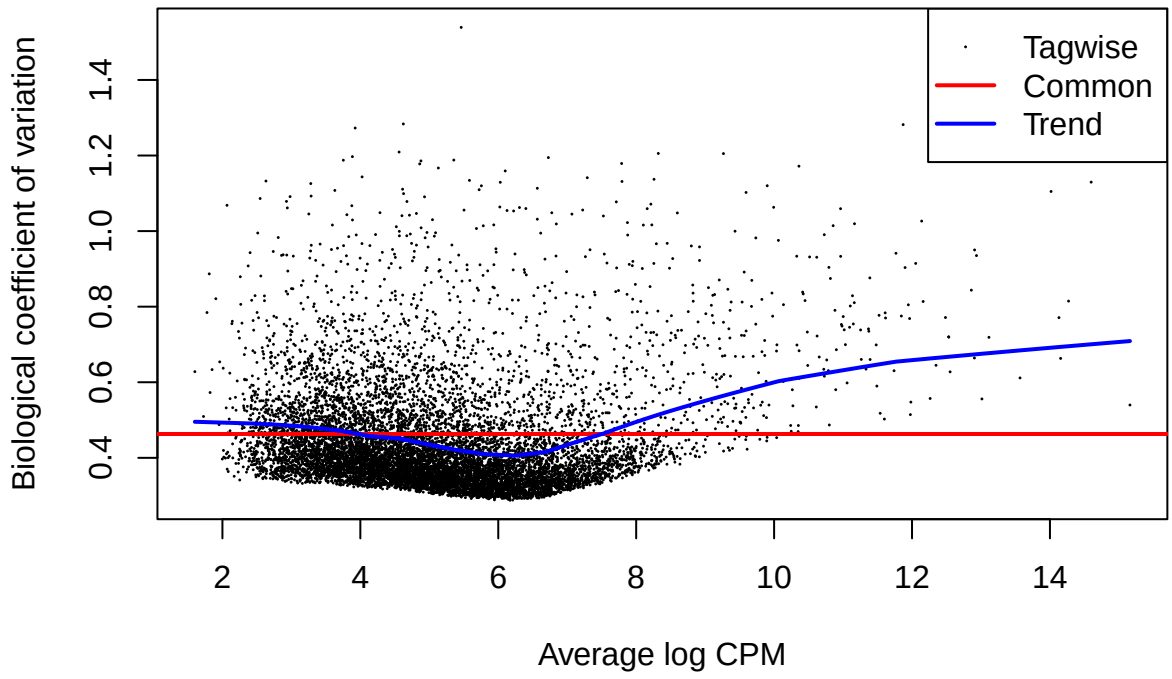
```
y <- estimateDisp(y)
```

```
## Using classic mode.
```

```
y$common.dispersion
```

```
## [1] 0.2144357
```

```
#There is little variability between replicates.
plotBCV(y)
```

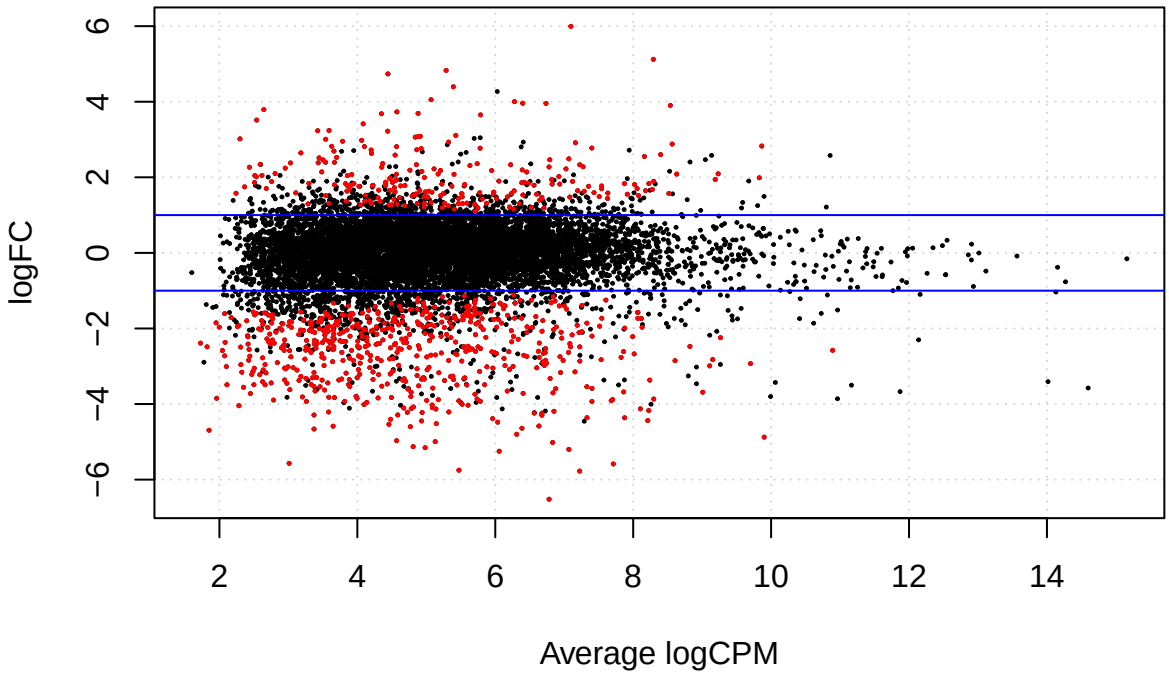


Differentially expressed genes—exact test

```
et <- exactTest(y, pair=c("N","T"))
summary(de<-decideTestsDGE(et))
```

```
##          T-N
## Down     571
## NotSig 9650
## Up       222
```

```
detags <- rownames(y)[as.logical(de)]
plotSmear(et, de.tags=detags)
abline(h=c(-1, 1), col="blue")
```




```
diffExpGenes <- topTags(et, n=1000, p.value = 0.05)
head(diffExpGenes)
```

```
## Comparison of groups:  T-N
##                ENSEMBL   SYMBOL     logFC   logCPM       PValue          FDR
## 487   ENSG00000196296   ATP2A1 -4.479821 6.034715 8.733434e-16 7.059757e-12
## 83699 ENSG00000198478 SH3BGRL2 -4.030906 5.606182 1.352055e-15 7.059757e-12
## 5837  ENSG00000068976     PYGM -5.249132 6.057221 2.119041e-15 7.376381e-12
## 5744  ENSG00000087494    PTHLH  4.002838 6.278230 5.285440e-15 1.379896e-11
## 23328 ENSG00000111961    SASH1 -3.323771 6.796001 9.846236e-14 2.056485e-10
## 5737  ENSG00000122420    PTGFR -5.126920 4.809714 3.265969e-13 5.684418e-10
```

```
write.table(diffExpGenes$table, file="tumor_v_normal_exactTest.txt", sep =
"\t", row.names=TRUE, col.names=NA)
```

Differentially expressed genes—generalized linear model

```
Patient <- factor(c(8,8,33,33,51,51))
Tissue <- factor(c("N","T","N","T","N","T"))
design <- model.matrix(~Patient+Tissue)
rownames(design) <- colnames(y)
design
```

```
##     (Intercept) Patient33 Patient51 TissueT
## N8            1         0         0       0
## T8            1         0         0       1
## N33           1         1         0       0
## T33           1         1         0       1
## N51           1         0         1       0
## T51           1         0         1       1
## attr(,"assign")
## [1] 0 1 1 2
## attr(,"contrasts")
## attr(,"contrasts")$Patient
## [1] "contr.treatment"
##
## attr(,"contrasts")$Tissue
## [1] "contr.treatment"
```

```
y <- estimateDisp(y, design, robust=TRUE)
y$common.dispersion
```

```
## [1] 0.1589545
```

```
#When design matrix was taken into account, the common dispersion decreased. This may be due to incorpo
```

```
fit <- glmFit(y, design)
lrt <- glmLRT(fit, coef=4)
summary(de2 <- decideTestsDGE(lrt))
```

```
##        TissueT
## Down       936
## NotSig    9175
## Up         332
```

```
diffExpGenes2 <- topTags(lrt, n=1000, p.value = 0.05)
head(diffExpGenes2$table)
```

```
##               ENSEMBL SYMBOL     logFC   logCPM       LR       PValue
## 5737  ENSG00000122420  PTGFR -5.181775 4.810518 98.68573 2.959271e-23
## 5744  ENSG00000087494  PTHLH  3.970101 6.278331 92.68019 6.146564e-22
## 3479  ENSG00000017427   IGF1 -3.987989 5.784339 86.75831 1.226273e-20
## 1288  ENSG00000197565 COL4A6  3.656176 5.786348 77.82939 1.123360e-18
## 10351 ENSG00000141338  ABCA8 -3.982850 5.006862 75.90194 2.981067e-18
## 5837  ENSG00000068976   PYGM -5.480473 6.057091 75.34552 3.951446e-18
##                FDR
## 5737  3.090366e-19
## 5744  3.209428e-18
## 3479  4.268655e-17
## 1288  2.932812e-15
## 10351 6.226256e-15
## 5837  6.877493e-15
```

1. ID Mapping. You have a data frame in R called "counts" that contains gene symbols in the first column called "SYMBOL" and integer read counts from six human samples in the subsequent columns. You want to map the gene symbols to Ensembl IDs.

a. Show the R command you would use to create a table that maps between the symbols and the Ensembl IDs.

mapping <- select(org.Hs.eg.db, as.character(counts$SYMBOL), keytype = "SYMBOL", column="ENSEMBL")

b. Show the R commands you would use to determine whether there are any Ensembl IDs in your list that map to multiple gene symbols.

d <- duplicated(mapping$SYMBOL)

sum(d)

c. Show the R commands you would use to determine whether there are gene symbols that do not have a corresponding Ensembl ID.

counts <- merge(counts, mapping, by = "SYMBOL")

missing <- is.na(counts$ENSEMBL)

sum(missing)

2. MDS plot.

a. Provide a screenshot of your MDS plot.

Provided above.

b. What sample characteristic is distinguished by dimension 1 (horizontal axis)?

Tumor

c. What sample characteristic is distinguished by dimension 2 (vertical axis)?

Patient

3. Dispersion and BCV.

a. What was the common dispersion for the exact test? For the generalized linear model?

The common dispersion was 0.21 for the exact test and 0.16 for the linear model.

b. Provide a screenshot of your BCV plot for the exact test.

Provided above.

c. Is the common BCV in the range of what you would expect given the nature of the experiment? Why or why not?

Considering these are human samples, I would expect that the common BCV to be near or above 0.4, indicating that there is on average 40% variability in the expression of genes across diseased and healthy groups. We also see an increase of BVC as counts increase which I would not expect.

4. Differentially expressed genes—exact test

a. How many significantly up-regulated, significantly down-regulated, and non- significant genes did you find?

Down-regulated = 571 Not Significant = 9650 Up-regulated = 222

b. Provide a screenshot of your log-fold change vs. average log CPM plot.

Provided above.

c. In what range(s) of fold-change do most of the significantly differentially expressed genes lie? 1:6 for up-regulated and -1:-7 for doqn-regulated

d. What were the top five most differentially regulated genes? Were they up- or down-regulated? ATP2A1 - Down-regulated SH3BGRL2 - Down-regulated PYGM - Down-regulated PTHLH - Up-regulated SASH1 - Down-regulated

5. Differentially expressed genes—generalized linear model

a. Provide a screenshot of your design matrix. Provided above.

b. How many significantly up-regulated, significantly down-regulated, and non- significant genes did you find?

Down-regulated = 936 Not Significant = 9175 Up-regulated = 332

c. What were the top five most differentially regulated genes? Were they up- or down-regulated? PTGFR - Down-regulated PTHLH - Up-regulated IGF1 - Down-regulated COL4A6 - Up-regulated ABCA8 - Down-regulated

d. Were there any genes in the top five that did not appear in the top five of the exact test? If so, what p-value and rank did these genes have in the exact test? ABCA8 is ranked 8th with pval 2.04E-11, PRGFR is ranked 6th with pval 3.27E-13, IGF1 is ranked 262nd with pval 9.51E-05 and COL4A6 is ranked 16th with pval 1.59E-10.

e. Show how you would modify the R command lrt <- glmLRT(fit, coef=4), so that you are testing the patient-specific effect of Patient 51 rather than the tumor-specific effect. (You don't have to carry out this analysis—just show >Patient <- factor(c(8,8,33,33,51,51))

```
design <- model.matrix(~Patient)

rownames(design) <- colnames(y)

design

y <- estimateDisp(y, design, robust=TRUE)

y$common.dispersion

fit <- glmFit(y, design)

lrt <- glmLRT(fit, coef=3)

summary(de2 <- decideTestsDGE(lrt))

diffExpGenes2 <- topTags(lrt, n=1000, p.value = 0.05)
```