Analyse d'expression différentielle pour un séquençage de l'ARN avec edgeR

-

12 décembre 2025

Table des matières

1 Résumé **1**

2 Introduction **1**

3 Méthodologie **2**

3.1 Introduction 2

3.2 Téléchargement des données 2

3.3 Création d'un *DGEList object* 2

3.4 Annotation 2

3.5 Filtrage des gènes 3

3.6 Normalisation 4

4 Résultats et discussions **4**

4.1 Data exploration 4

4.2 Design matrix 5

4.3 Dispersion Estimation 6

4.4 Differential expression 7

4.5 Gene Ontology Analysis 8

5 Conclusion **9**

6 Références **9**

1 Résumé

2 Introduction

Le séquençage de l'ARN (*RNA-seq*) est une méthode permettant de séquencer un ensemble de molécules d'ARN. Le séquençage d'ARN est utilisé pour déterminer quels segments d'ADN ont été transcrits en ARN et la quantité un gène est exprimé, afin de mieux comprendre la fonction des différents gènes (1).

edgeR est un package R conçu pour l'analyse d'expression différentielle des données de « *RNA-seq count* ». Il peut détecter des différences entre deux groupes ou plus quand au moins un des groupes a effectué des mesures répétées (2).

3 Méthodologie

3.1 Introduction

3.2 Téléchargement des données

Nous commonçons par charger les *libraries* pertinentes.

```
library(org.Hs.eg.db)
library(limma)
library(edgeR)
```

data taken from: https://www.doi.org

```
rawdata <- read.delim("TableS1.txt", check.names = FALSE, stringsAsFactors = FALSE)
head(rawdata)
```

```
##       RefSeqID    Symbol NbrOfExons   8N 8T   33N 33T   51N  51T
## 1    NM_182502 TMPRSS11B         10 2592  3  7805 321  3372    9
## 2    NM_003280     TNNC1          6 1684  0  1787   7  4894  559
## 3    NM_152381     XIRP2         10 9915 15 10396  48 23309 7181
## 4    NM_022438       MAL          3 2496  2  3585 239  1596    7
## 5 NM_001100112      MYH2         40 4389  7  7944  16  9262 1818
## 6    NM_017534      MYH2         40 4402  7  7943  16  9244 1815
```

3.3 Création d'un *DGEList object*

- create a DGEList object (https://www.rdocumentation.org/packages/edgeR/versions/3.14.0/topics/DGEList) to hold our read counts
- This object is used as a container for the counts, and for all the associated metadata eg. sample names, gene names and normalisation factors

```
y <- DGEList(counts=rawdata[,4:9], genes=rawdata[,1:3])
```

3.4 Annotation

```
idfound <- y$genes$RefSeqID %in% mappedRkeys(org.Hs.egREFSEQ)
y <- y[idfound,]
dim(y)
```

```
## [1] 15533     6
```

```
egREFSEQ <- toTable(org.Hs.egREFSEQ)
head(egREFSEQ)
```

```
##   gene_id    accession
## 1       1    NM_130786
## 2       1    NP_570602
## 3       2    NM_000014
## 4       2 NM_001347423
## 5       2 NM_001347424
## 6       2 NM_001347425
```

```
m <- match(y$genes$RefSeqID, egREFSEQ$accession)
y$genes$EntrezGene <- egREFSEQ$gene_id[m]
```

```
egSYMBOL <- toTable(org.Hs.egSYMBOL)
head(egSYMBOL)
```

```
##   gene_id   symbol
## 1       1     A1BG
## 2       2      A2M
## 3       9     NAT1
## 4      10     NAT2
## 5      11     NATP
## 6      12 SERPINA3
```

```
m <- match(y$genes$EntrezGene, egSYMBOL$gene_id)
y$genes$Symbol <- egSYMBOL$symbol[m]
head(y$genes)
```

```
##       RefSeqID    Symbol NbrOfExons EntrezGene
## 1    NM_182502 TMPRSS11B         10     132724
## 2    NM_003280     TNNC1          6       7134
## 3    NM_152381     XIRP2         10     129446
## 4    NM_022438       MAL          3       4118
## 5 NM_001100112      MYH2         40       4620
## 6    NM_017534      MYH2         40       4620
```

3.5 Filtrage des gènes

```
o <- order(rowSums(y$counts))
y <- y[o,]
d <- duplicated(y$genes$Symbol)
y <- y[!d,]
nrow(y)
```

```
## [1] 10510
```

```
y$samples$lib.size <- colSums(y$counts)
```

```
rownames(y$counts) <- rownames(y$genes) <- y$genes$EntrezGene
y$genes$EntrezGene <- NULL
```

3.6 Normalisation

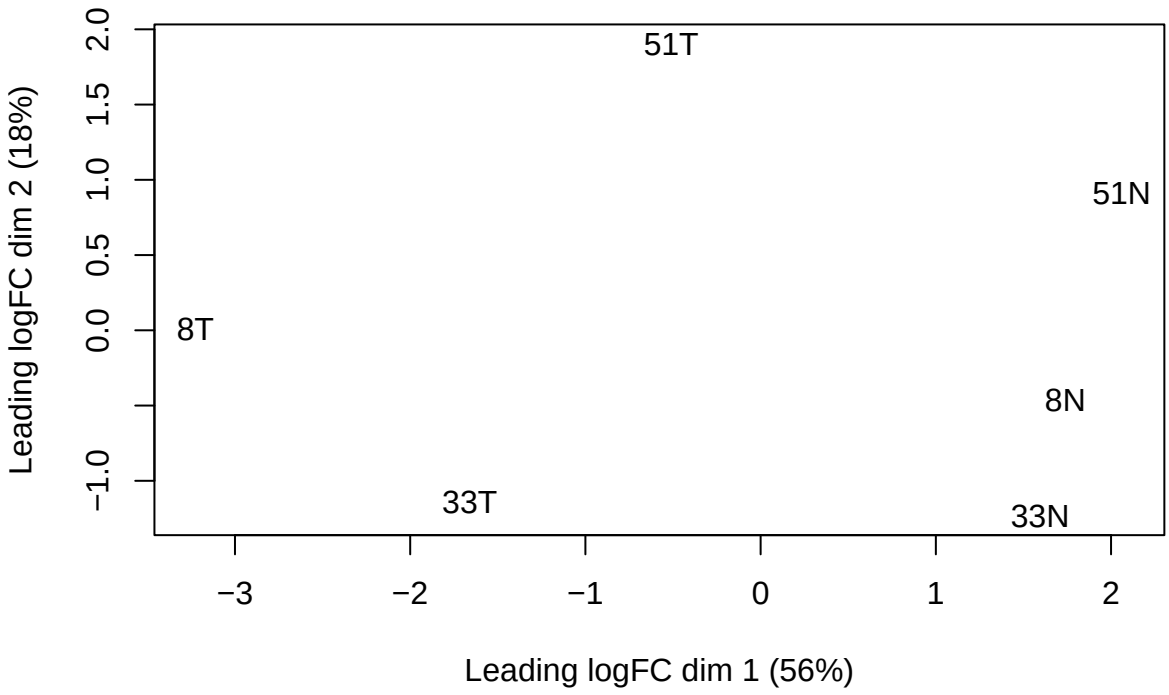
```
y <- normLibSizes(y)
y$samples
```

```
##     group lib.size norm.factors
## 8N      1  7397598    1.1542497
## 8T      1  7124442    1.0619357
## 33N     1 15260793    0.6556112
## 33T     1 13651143    0.9484143
## 51N     1 19318441    1.0892960
## 51T     1 14382783    1.2045134
```

4 Résultats et discussions

4.1 Data exploration

```
plotMDS(y)
```



4.2 Design matrix

```
Patient <- factor(c(8,8,33,33,51,51))
Tissue <- factor(c("N","T","N","T","N","T"))
data.frame(Sample=colnames(y),Patient,Tissue)
```

```
##   Sample Patient Tissue
## 1     8N       8      N
## 2     8T       8      T
## 3    33N      33      N
## 4    33T      33      T
## 5    51N      51      N
## 6    51T      51      T
```

```
design <- model.matrix(~Patient+Tissue)
rownames(design) <- colnames(y)
design
```

```
##     (Intercept) Patient33 Patient51 TissueT
## 8N            1         0         0       0
## 8T            1         0         0       1
## 33N           1         1         0       0
## 33T           1         1         0       1
```

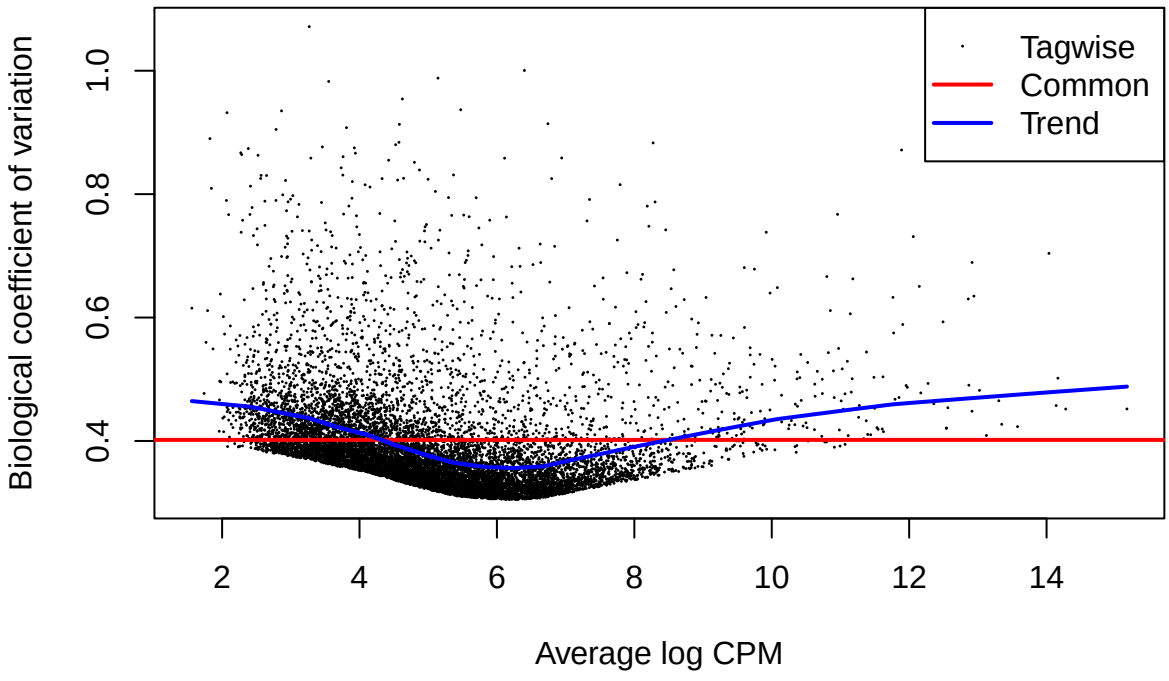
```
## 51N           1         0         1       0
## 51T           1         0         1       1
## attr(,"assign")
## [1] 0 1 1 2
## attr(,"contrasts")
## attr(,"contrasts")$Patient
## [1] "contr.treatment"
##
## attr(,"contrasts")$Tissue
## [1] "contr.treatment"
```

4.3 Dispersion Estimation

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

```
## [1] 0.1613756
```

```
plotBCV(y)
```



4.4 Differential expression

```
fit <- glmFit(y, design)
```

```
lrt <- glmLRT(fit)
topTags(lrt)
```

```
## Coefficient:  TissueT
##         RefSeqID   Symbol NbrOfExons     logFC   logCPM        LR       PValue
## 5737   NM_000959    PTGFR          3 -5.201023 4.822267 100.46226 1.206744e-23
## 5744   NM_198966    PTHLH          4  3.881888 5.820335  84.29578 4.260245e-20
## 1288   NM_001847   COL4A6         45  3.710900 5.709635  78.26706 9.001057e-19
## 10351  NM_007168    ABCA8         38 -3.996432 5.022051  77.53105 1.306522e-18
## 5837   NM_005609     PYGM         20 -5.495113 6.075033  74.74014 5.369333e-18
## 487    NM_173201   ATP2A1         22 -4.623578 6.040006  73.58113 9.658372e-18
## 27179  NM_014440    IL36A          4 -6.178402 5.486198  72.16644 1.977909e-17
## 196374 NM_173352    KRT78          9 -4.258876 7.697534  70.84623 3.861808e-17
## 6387   NM_199168   CXCL12          3 -3.717669 5.864198  68.91229 1.029414e-16
## 83699  NM_031469 SH3BGRL2          4 -3.947822 5.622290  68.36318 1.359935e-16
##                 FDR
## 5737   1.268288e-19
## 5744   2.238759e-16
## 1288   3.153370e-15
## 10351  3.432885e-15
## 5837   1.128634e-14
## 487    1.691825e-14
## 27179  2.969689e-14
## 196374 5.073451e-14
## 6387   1.202127e-13
## 83699  1.429291e-13
```

```
colnames(design)
```

```
## [1] "(Intercept)" "Patient33"   "Patient51"   "TissueT"
```

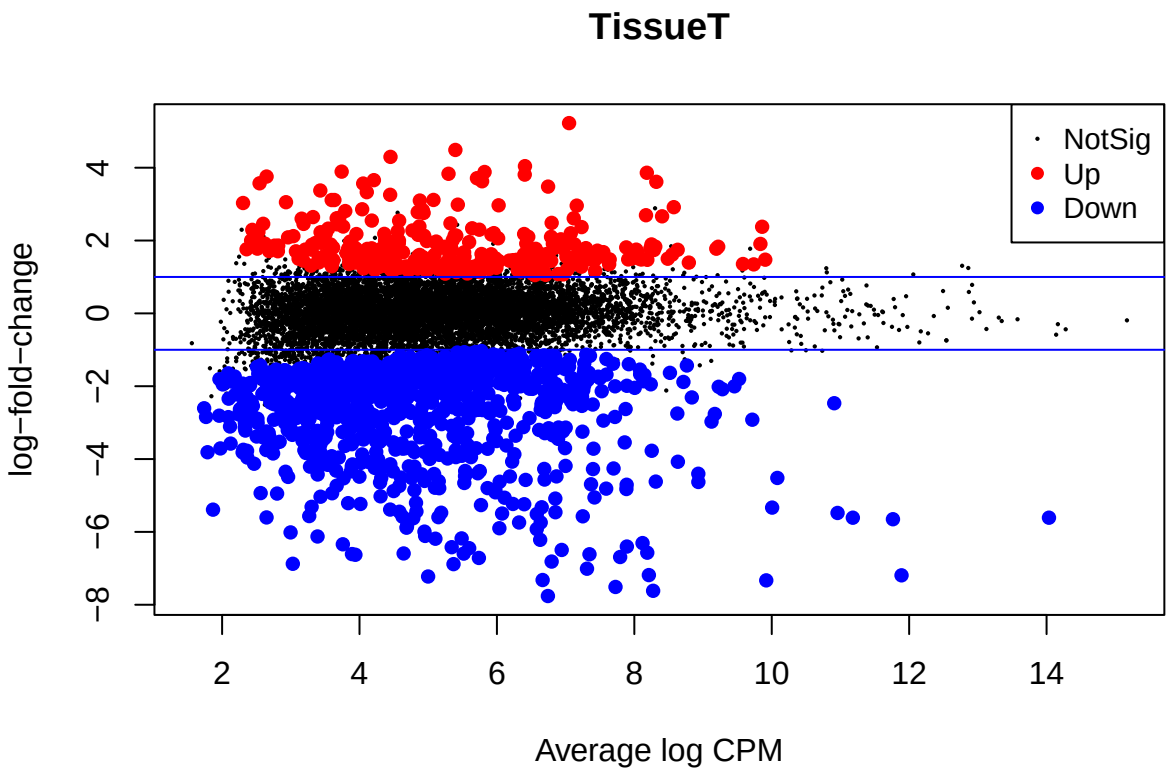
```
o <- order(lrt$table$PValue)
cpm(y)[o[1:10],]
```

```
##                8N          8T       33N        33T        51N          51T
## 5737    53.286956   0.9252284  28.28544   0.926860  82.448258   2.59751429
## 5744     5.387253  78.1157149  10.59455 133.854037   5.940076 104.07373912
## 1288    11.945647 137.0659837   6.19681  96.470682   4.514458  57.26075940
## 10351   56.449039   3.3043873  41.77849   2.239912  83.636273   6.34947937
## 5837   163.842749   2.9078608 126.63482   1.235813 103.167244   5.94542159
## 487    114.537676   3.3043873 155.12015   4.016393 107.634181   9.29332890
## 27179   42.980907   1.3217549 182.30616   3.475725  38.111529   0.05772254
## 196374 399.125153  21.9411314 615.48318  50.436634 153.206446   4.73324826
## 6387    63.827233   3.0400363  68.46476   6.179067 188.514259  17.95170985
## 83699  103.177600   5.4191951 124.03615   5.715637  50.894573   5.65680889
```

```
summary(decideTests(lrt))
```

```
##        TissueT
## Down       946
## NotSig    9243
## Up         321
```

```
plotMD(lrt)
abline(h=c(-1, 1), col="blue")
```



4.5 Gene Ontology Analysis

```
go <- goana(lrt)
topGO(go, ont="BP", sort="Up", n=15, truncate=45)
```

```
##                                          Term Ont    N  Up Down         P.Up
## GO:0022008                      neurogenesis  BP 1083  74  120 1.277061e-11
## GO:0009888                tissue development  BP 1294  82  199 3.631218e-11
## GO:0007399         nervous system development  BP 1556  92  162 7.160014e-11
## GO:0007155                       cell adhesion  BP  945  63  160 1.687833e-09
## GO:0048513           animal organ development  BP 1850  99  267 2.914578e-09
## GO:0048731                 system development  BP 2433 120  309 4.077632e-09
```

```
## GO:0060429             epithelium development  BP  771  54   96 5.630044e-09
## GO:0007275 multicellular organism development  BP 2857 134  336 7.638940e-09
## GO:0048699              generation of neurons  BP  913  60  103 7.739626e-09
## GO:0030154               cell differentiation  BP 2603 125  324 8.483777e-09
## GO:0048869     cellular developmental process  BP 2604 125  324 8.695071e-09
## GO:0008544              epidermis development  BP  258  27   34 2.071778e-08
## GO:0016477                    cell migration  BP 1012  63  156 2.439558e-08
## GO:0048870                     cell motility  BP 1085  66  161 2.588447e-08
## GO:0009653 anatomical structure morphogenesis  BP 1697  90  249 3.327324e-08
##                  P.Down
## GO:0022008 7.917927e-03
## GO:0009888 1.150777e-15
## GO:0007399 2.115949e-02
## GO:0007155 2.547931e-16
## GO:0048513 1.359572e-17
## GO:0048731 1.454654e-12
## GO:0060429 5.315945e-04
## GO:0007275 2.317471e-09
## GO:0048699 8.254978e-03
## GO:0030154 4.414787e-12
## GO:0048869 4.649368e-12
## GO:0008544 1.515137e-02
## GO:0016477 2.449541e-12
## GO:0048870 2.405124e-11
## GO:0009653 3.376325e-17
```

5 Conclusion

6 Références

1. Ian C. Nova PhD. RNA-seq (RNA sequencing) [Internet]. 2025. Available from: https://www.genome.gov/genetics-glossary/RNA-seq

2. Robinson MD SGK McCarthy DJ. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. [Internet]. 2009. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2796818/

3. Chen Y, McCarthy D, Baldoni P, Ritchie M, Robinson M, Smyth G. edgeR: Differential analysis of sequence read count data. User's guide [Internet]. 2025. Available from: https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

4. Maglott D PK Ostell J. Entrez gene: Gene-centered information at NCBI [Internet]. 2007. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC1761442/