# Chapter 4

# Case studies

## 4.1 RNA-Seq of oral carcinomas vs matched normal tissue

### 4.1.1 Introduction

This section provides a detailed analysis of data from a paired design RNA-seq experiment, featuring oral squamous cell carcinomas and matched normal tissue from three patients [40]. The aim of the analysis is to detect genes differentially expressed between tumor and normal tissue, adjusting for any differences between the patients. This provides an example of the GLM capabilities of edgeR.

RNA was sequenced on an Applied Biosystems SOLiD System 3.0 and reads mapped to the UCSC hg18 reference genome [40]. Read counts, summarised at the level of refSeq transcripts, are available in Table S1 of Tuch et al. [40].

### 4.1.2 Reading in the data

The read counts for the six individual libraries are stored in one tab-delimited file. To make this file, we downloaded Table S1 from Tuch et al. [40], deleted some unnecessary columns and edited the column headings slightly:

```
> 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
```

For easy manipulation, we put the data into a `DGEList` object:

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

### 4.1.3 Annotation

The study by Tuch et al. [40] was undertaken a few years ago, so not all of the RefSeq IDs provided by match RefSeq IDs currently in use. We retain only those transcripts with IDs in the current NCBI annotation, which is provided by the `org.HS.eg.db` package:

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

[1] 15537     6
```

We add Entrez Gene IDs to the annotation:

```
> 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]
```

Now use the Entrez Gene IDs to update the gene symbols:

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

  gene_id symbol
1       1   A1BG
2       2    A2M
3       3  A2MP1
4       9   NAT1
5      10   NAT2
6      11   NATP

> 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
```

### 4.1.4 Filtering and normalization

Different RefSeq transcripts for the same gene symbol count predominantly the same reads. So we keep one transcript for each gene symbol. We choose the transcript with highest overall count:

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

[1] 10512
```

Normally we would also filter lowly expressed genes. For this data, all transcripts already have at least 50 reads for all samples of at least one of the tissues types.

Recompute the library sizes:

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

Use Entrez Gene IDs as row names:

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

TMM normalization is applied to this dataset to account for compositional difference between the libraries.

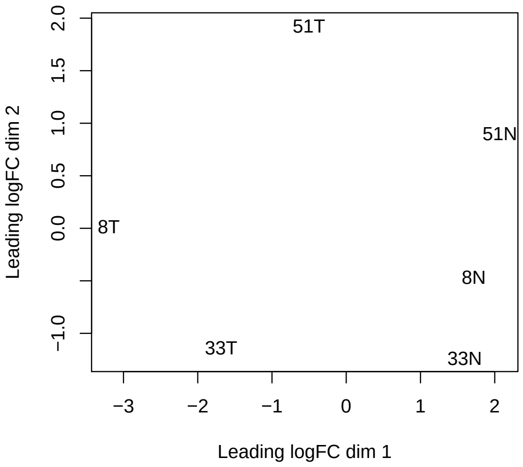
```
> y <- calcNormFactors(y)
> y$samples

    group lib.size norm.factors
8N      1  7989626        1.146
8T      1  7371161        1.086
33N     1 15754803        0.672
33T     1 14043438        0.973
51N     1 21540651        1.032
51T     1 15193446        1.190
```

### 4.1.5 Data exploration

The first step of an analysis should be to examine the samples for outliers and for other relationships. The function `plotMDS` produces a plot in which distances between samples correspond to leading biological coefficient of variation (BCV) between those samples:

```
> plotMDS(y)
```



In the plot, dimension 1 separates the tumor from the normal samples, while dimension 2 roughly corresponds to patient number. This confirms the paired nature of the samples. The tumor samples appear more heterogeneous than the normal samples.

### 4.1.6 The design matrix

Before we fit negative binomial GLMs, we need to define our design matrix based on the experimental design. Here we want to test for differential expression between tumour and normal tissues within patients, i.e. adjusting for differences between patients. In statistical terms, this is an additive linear model with patient as the blocking factor:

```
> 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"
```

This sort of additive model is appropriate for paired designs, or experiments with batch effects.

### 4.1.7 Estimating the dispersion

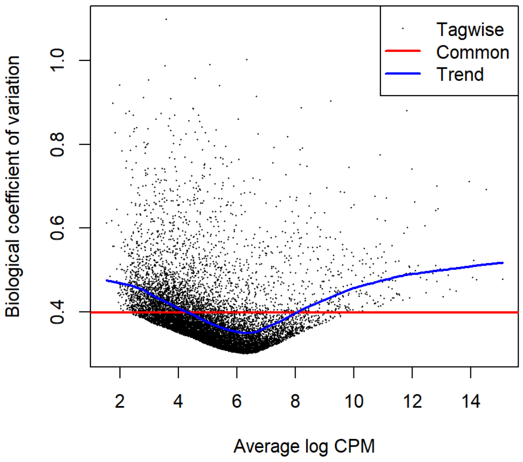
We estimate the NB dispersion for the dataset.

```
> y <- estimateDisp(y, design, robust=TRUE)
> y$common.dispersion
[1] 0.159
```

The square root of the common dispersion gives the coefficient of variation of biological variation. Here the common dispersion is found to be 0.159, so the coefficient of biological variation is around 0.4.

The dispersion estimates can be viewed in a BCV plot:

```
> plotBCV(y)
```



### 4.1.8 Differential expression

Now proceed to determine differentially expressed genes. Fit genewise glms:

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

Conduct likelihood ratio tests for tumour vs normal tissue differences and show the top genes:

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

Coefficient:  TissueT
           RefSeqID   Symbol NbrOfExons logFC logCPM   LR   PValue      FDR
5737   NM_001039585    PTGFR          4 -5.18   4.74 98.7 2.96e-23 3.11e-19
5744      NM_002820    PTHLH          4  3.97   6.21 92.2 7.99e-22 4.20e-18
3479   NM_001111283     IGF1          5 -3.99   5.71 86.5 1.38e-20 4.84e-17
1288      NM_033641   COL4A6         45  3.66   5.72 77.5 1.30e-18 3.41e-15
10351     NM_007168    ABCA8         38 -3.98   4.94 75.9 2.96e-18 6.23e-15
5837      NM_005609     PYGM         20 -5.48   5.99 75.4 3.93e-18 6.88e-15
487       NM_004320   ATP2A1         23 -4.62   5.96 74.8 5.21e-18 7.83e-15
27179     NM_014440    IL36A          4 -6.17   5.40 72.2 1.95e-17 2.56e-14
196374    NM_173352    KRT78          9 -4.25   7.61 70.8 3.95e-17 4.61e-14
83699     NM_031469 SH3BGRL2          4 -3.93   5.53 67.8 1.84e-16 1.93e-13
```

Note that `glmLRT` has conducted a test for the last coefficient in the linear model, which we can see is the tumor vs normal tissue effect:

```
> colnames(design)

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

The genewise tests are for tumor vs normal differential expression, adjusting for baseline differences between the three patients. The tests can be viewed as analogous to paired $t$-tests. The top DE tags have tiny $p$-values and FDR values, as well as large fold changes.

Here's a closer look at the counts-per-million in individual samples for the top genes:

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

           8N      8T    33N     33T    51N      51T
5737    49.69   0.875  27.10   0.878  78.11   2.5433
5744     7.32  95.851  11.80 204.166   6.88 116.3276
3479    50.24   3.124  32.39   1.902 211.60  14.2092
1288    12.12 140.215   6.33  94.438   4.86  56.8369
10351   52.64   3.124  39.47   2.121  79.19   6.0818
5837   152.79   2.749 119.63   1.170  97.68   5.6947
487    107.90   3.124 147.11   3.804 102.81   8.9015
27179   40.08   1.250 172.22   3.292  36.08   0.0553
196374 372.19  20.745 581.44  47.768 145.06   4.5337
83699   96.21   5.124 117.18   5.413  48.19   5.4183
```

We see that all the top genes have consistent tumour vs normal changes for the three patients.

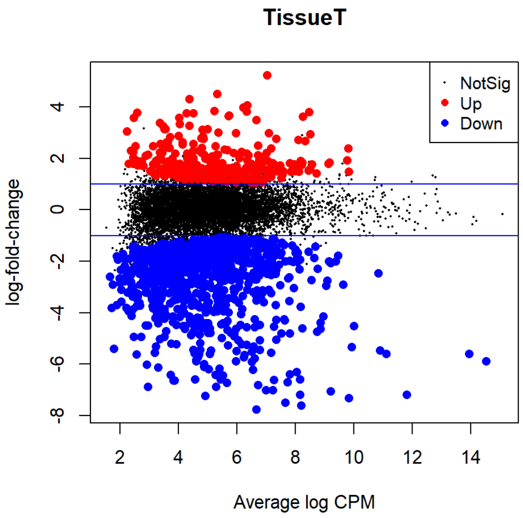
The total number of differentially expressed genes at 5% FDR is given by:

```
> summary(decideTests(lrt))

       TissueT
Down       938
NotSig    9243
Up         331
```

Plot log-fold change against log-counts per million, with DE genes highlighted:

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



The blue lines indicate 2-fold changes.

### 4.1.9 Gene ontology analysis

We perform a gene ontology analysis focusing on the ontology of biological process (BP). The genes up-regulated in the tumors tend to be associated with cell differentiation, cell migration and tissue morphogenesis:

```
> go <- goana(lrt)
> topGO(go, ont="BP", sort="Up", n=30, truncate=30)

                                     Term Ont    N  Up Down     P.Up   P.Down
GO:0048699          generation of neurons  BP  922  69  101 5.11e-12 1.55e-02
GO:0040011                     locomotion  BP 1100  76  163 1.76e-11 9.55e-12
GO:0022008                   neurogenesis  BP  991  71  108 1.84e-11 1.44e-02
GO:0009888             tissue development  BP 1242  81  192 6.00e-11 9.94e-16
GO:0016477                 cell migration  BP  941  67  147 1.03e-10 2.13e-12
GO:0006928 movement of cell or subcell...  BP 1282  82  194 1.18e-10 6.34e-15
GO:0022610            biological adhesion  BP  882  64  156 1.35e-10 4.49e-18
GO:0030154           cell differentiation  BP 2434 127  306 2.08e-10 2.33e-12
GO:0048870                  cell motility  BP 1004  69  151 2.45e-10 2.09e-11
```

```
GO:0051674           localization of cell  BP 1004  69   151 2.45e-10 2.09e-11
GO:0048869 cellular developmental proc...  BP 2494 129   310 2.46e-10 6.71e-12
GO:0007155                  cell adhesion  BP  877  63   155 2.95e-10 6.32e-18
GO:0030182          neuron differentiation  BP  824  60    93 5.03e-10 9.34e-03
GO:0030198 extracellular matrix organi...  BP  261  30    53 7.08e-10 7.83e-09
GO:0043062 extracellular structure org...  BP  261  30    53 7.08e-10 7.83e-09
GO:0007399     nervous system development  BP 1410  84   148 2.28e-09 1.61e-02
GO:0048513       animal organ development  BP 2123 111   286 5.36e-09 2.67e-15
GO:0009653 anatomical structure morpho...  BP 1690  94   237 6.54e-09 2.16e-14
GO:0048468               cell development  BP 1301  78   182 7.84e-09 7.39e-11
GO:0060429          epithelium development  BP  765  54    94 1.28e-08 7.09e-04
GO:0007275 multicellular organism deve...  BP 3149 146   380 2.24e-08 3.46e-13
GO:0048731             system development  BP 2853 135   361 3.36e-08 1.43e-15
GO:0043588               skin development  BP  225  25    33 3.91e-08 2.97e-03
GO:0030155    regulation of cell adhesion  BP  481  39    76 4.75e-08 4.57e-07
GO:0008544          epidermis development  BP  246  26    37 5.80e-08 1.08e-03
GO:0008283 cell population proliferati...  BP 1178  70   146 8.22e-08 1.33e-05
GO:0009887     animal organ morphogenesis  BP  633  45    87 2.02e-07 2.37e-05
GO:0030030 cell projection organizatio...  BP  995  61    94 2.31e-07 2.88e-01
GO:0120036 plasma membrane bounded cel...  BP  972  60    91 2.34e-07 3.24e-01
GO:0048856 anatomical structure develo...  BP 3434 152   414 2.34e-07 1.02e-14
```

### 4.1.10 Setup

This analysis was conducted on:

```
> sessionInfo()

R version 4.0.3 (2020-10-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 16299)

Matrix products: default

Random number generation:
 RNG:     Mersenne-Twister
 Normal:  Inversion
 Sample:  Rounding

locale:
[1] LC_COLLATE=English_Australia.1252  LC_CTYPE=English_Australia.1252
[3] LC_MONETARY=English_Australia.1252 LC_NUMERIC=C
[5] LC_TIME=English_Australia.1252

attached base packages:
[1] parallel  stats4    stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] org.Hs.eg.db_3.12.0  AnnotationDbi_1.51.3 IRanges_2.23.10
 [4] S4Vectors_0.27.14     Biobase_2.49.1        BiocGenerics_0.35.4
 [7] edgeR_3.31.5          limma_3.45.18         knitr_1.30
[10] BiocStyle_2.17.1

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.5          compiler_4.0.3      BiocManager_1.30.10
 [4] highr_0.8           tools_4.0.3         digest_0.6.25
 [7] bit_4.0.4           statmod_1.4.34      evaluate_0.14
[10] RSQLite_2.2.1       memoise_1.1.0       lattice_0.20-41
[13] pkgconfig_2.0.3     rlang_0.4.8         DBI_1.1.0
[16] yaml_2.2.1          xfun_0.18           stringr_1.4.0
[19] vctrs_0.3.4         locfit_1.5-9.4      bit64_4.0.5
[22] grid_4.0.3          rmarkdown_2.4       GO.db_3.12.0
[25] blob_1.2.1          magrittr_1.5        htmltools_0.5.0
[28] splines_4.0.3       stringi_1.5.3
```

## 4.2 RNA-Seq of pathogen inoculated arabidopsis with batch effects

### 4.2.1 Introduction

This case study re-analyses Arabidopsis thaliana RNA-Seq data described by Cumbie et al. [6]. Summarized count data is available as a data object in the CRAN package NBPSeq comparing ΔhrcC challenged and mock-inoculated samples [6]. Samples were collected in three batches, and adjustment for batch effects proves to be important. The aim of the analysis therefore is to detect genes differentially expressed in response to ΔhrcC challenge, while correcting for any differences between the batches.

### 4.2.2 RNA samples

*Pseudomonas syringae* is a bacterium often used to study plant reactions to pathogens. In this experiment, six-week old Arabidopsis plants were inoculated with the ΔhrcC mutant of *P. syringae*, after which total RNA was extracted from leaves. Control plants were inoculated with a mock pathogen.

Three biological replicates of the experiment were conducted at separate times and using independently grown plants and bacteria.

The six RNA samples were sequenced one per lane on an Illumina Genome Analyzer. Reads were aligned and summarized per gene using GENE-counter. The reference genome was derived from the TAIR9 genome release (www.arabidopsis.org).