4.7 Bisulfite sequencing of mouse oocytes

4.7.1 Introduction

The bisulfite sequencing (BS-seq) data of this case study is described in Gahurova *et al.* [13]. The sequence and count data are publicly available from the Gene Expression Omnibus (GEO) at the series accession number GSE86297.

This study investigates the onset and progression of *de novo* methylation. Growing oocytes from pre-pubertal mouse ovaries (post-natal days 7-18) isolated and sorted into the following, non-overlapping size categories: 40-45, 50-55 and 60-65μm with two biological replicates in each. Methylation maps were generated by bisulfite conversion of oocyte DNA and Illumina sequencing. Reduced representation bisulfite sequencing (RRBS [27]) was applied for focusing coverage of CGIs and other GC-rich sequences in all three size classes of oocytes. RRBS reads were trimmed to remove poor quality calls and adapters using Trim Galore and mapped to the mouse genome GRCm38 assembly by Bismark [17]. This is summarized in the table below.

```
> library(edgeR)
> targets <- read.delim("targets.txt", stringsAsFactors=FALSE)
```

```
> targets
         GEO    Sample Group                                          File
1 GSM2299710 40-45um-A  40um GSM2299710_RRBS_40-45oocyte_LibA.cov.txt.gz
2 GSM2299711 40-45um-B  40um GSM2299711_RRBS_40-45oocyte_LibB.cov.txt.gz
3 GSM2299712 50-55um-A  50um GSM2299712_RRBS_50-55oocyte_LibA.cov.txt.gz
4 GSM2299713 50-55um-B  50um GSM2299713_RRBS_50-55oocyte_LibB.cov.txt.gz
5 GSM2299714 60-65um-A  60um GSM2299714_RRBS_60-65oocyte_LibA.cov.txt.gz
6 GSM2299715 60-65um-B  60um GSM2299715_RRBS_60-65oocyte_LibB.cov.txt.gz
```

4.7.2 Reading in the data

The Bismark outputs of the data include one coverage file of the methylation in CpG context for each sample. The coverage file for each of the six samples is available for download at GEO. The first six rows of the coverage output for the first sample are shown below.

```
> s1 <- read.delim(file="GSM2299710_RRBS_40-45oocyte_LibA.cov.txt.gz",
+     header=FALSE, nrows=6)
```

```
> s1
  V1      V2      V3   V4 V5   V6
1  6 3121266 3121266 0.00  0   17
2  6 3121296 3121296 0.00  0   17
3  6 3179319 3179319 1.28  1   77
4  6 3180316 3180316 4.55  1   21
5  6 3182928 3182928 4.33 22  486
6  6 3182937 3182937 5.37 61 1074
```

The six columns (from left to right) represent: chromosome, start position, end position, methylation proportion in percentage, number of methylated C's and number of unmethylated C's. Since the start and end positions of a CpG site from Bismark are the same, we can keep only one of them. The last two columns of counts are we will use for the analysis.

We read in the coverage files of all six samples using `readBismark2DGE`. A DGEList object is created using the count table, and the chromosome number and positions are used for annotation.

```
> files <- targets$File
> yall <- readBismark2DGE(files, sample.names=targets$Sample)
```

The edgeRpackage stores the counts and associated annotation in a DGEList object. There is a row for each CpG locus found in any of the files. There are columns of methylated and unmethylated counts for each sample. The chromosomes and genomic loci are stored in the `genes` component.

```
> yall

An object of class "DGEList"
$counts
          40-45um-A-Me 40-45um-A-Un 40-45um-B-Me 40-45um-B-Un 50-55um-A-Me
6-3121266            0           17            0            4            0
6-3121296            0           17            0            4            0
6-3179319            1           77            0           76            2
6-3180316            1           21            0            0            1
6-3182928           22          486            8          953            7
          50-55um-A-Un 50-55um-B-Me 50-55um-B-Un 60-65um-A-Me 60-65um-A-Un
6-3121266           17            0            0            3            3
6-3121296           16            0            0            0            6
6-3179319           52            0            7           10           43
6-3180316            7            0            0            2            4
6-3182928          714           32         1190           10          618
          60-65um-B-Me 60-65um-B-Un
6-3121266            0           11
6-3121296            0           11
6-3179319            3           30
6-3180316            1            0
6-3182928           12          651
2271667 more rows ...

$samples
             group lib.size norm.factors
40-45um-A-Me     1  1231757            1
40-45um-A-Un     1 36263318            1
40-45um-B-Me     1  1719267            1
40-45um-B-Un     1 55600556            1
50-55um-A-Me     1  2691638            1
7 more rows ...

$genes
          Chr   Locus
6-3121266   6 3121266
```

```
6-3121296   6 3121296
6-3179319   6 3179319
6-3180316   6 3180316
6-3182928   6 3182928
2271667 more rows ...
```

```
> dim(yall)
```

```
[1] 2271672      12
```

We remove the mitochondrial genes as they are usually of less interest.

```
> table(yall$genes$Chr)
```

```
     6      9     17      1      3     13     10      2      4      5     11
111377 120649 101606 140819 108466  95196 116980 173357 157628 159979 161754
    18     16      7      8     14     19      X     12     15      Y     MT
 71737  70964 140225 130786  84974  70614  58361  95580  99646    662    312
```

```
> yall <- yall[yall$genes$Chr!="MT", ]
```

For convenience, we sort the DGEList so that all loci are in genomic order, from chromosome 1 to chromosome Y.

```
> ChrNames <- c(1:19,"X","Y")
> yall$genes$Chr <- factor(yall$genes$Chr, levels=ChrNames)
> o <- order(yall$genes$Chr, yall$genes$Locus)
> yall <- yall[o,]
```

We now annotate the CpG loci with the identity of the nearest gene. We search for the gene transcriptional start site (TSS) closest to each our CpGs:

```
> TSS <- nearestTSS(yall$genes$Chr, yall$genes$Locus, species="Mm")
> yall$genes$EntrezID <- TSS$gene_id
> yall$genes$Symbol <- TSS$symbol
> yall$genes$Strand <- TSS$strand
> yall$genes$Distance <- TSS$distance
> yall$genes$Width <- TSS$width
> head(yall$genes)
```

```
          Chr   Locus EntrezID Symbol Strand Distance  Width
1-3003886   1 3003886   497097   Xkr4      -  -667612 457017
1-3003899   1 3003899   497097   Xkr4      -  -667599 457017
1-3020877   1 3020877   497097   Xkr4      -  -650621 457017
1-3020891   1 3020891   497097   Xkr4      -  -650607 457017
1-3020946   1 3020946   497097   Xkr4      -  -650552 457017
1-3020988   1 3020988   497097   Xkr4      -  -650510 457017
```

Here `EntrezID`, `Symbol`, `Strand` and `Width` are the Entrez Gene ID, symbol, strand and width of the nearest gene. `Distance` is the genomic distance from the CpG to the TSS. Positive values means the TSS is downstream of the CpG and negative values means the TSS is upstream.

4.7.3 Filtering and normalization

We now turn to statistical analysis of differential methylation. Our first analysis will be for individual CpG loci.

CpG loci that have low coverage are removed prior to downstream analysis as they provide little information for assessing methylation levels. We sum up the counts of methylated and unmethylated reads to get the total read coverage at each CpG site for each sample:

```
> Methylation <- gl(2,1,ncol(yall), labels=c("Me","Un"))
> Me <- yall$counts[, Methylation=="Me"]
> Un <- yall$counts[, Methylation=="Un"]
> Coverage <- Me + Un
> head(Coverage)
          40-45um-A-Me 40-45um-B-Me 50-55um-A-Me 50-55um-B-Me 60-65um-A-Me
1-3003886            0            0            0            0            3
1-3003899            0            0            0            0            3
1-3020877           84           77          114           21           86
1-3020891           84           78          116           21           86
1-3020946          146          369          210          165          195
1-3020988           38           91           60           94           50
          60-65um-B-Me
1-3003886            0
1-3003899            0
1-3020877           57
1-3020891           57
1-3020946          168
1-3020988           25
```

As a conservative rule of thumb, we require a CpG site to have a total count (both methylated and unmethylated) of at least 8 in every sample before it is considered in the study.

```
> HasCoverage <- rowSums(Coverage >= 8) == 6
```

This filtering criterion could be relaxed somewhat in principle but the number of CpGs kept in the analysis is large enough for our purposes.

We also filter out CpGs that are never methylated or always methylated as they provide no information about differential methylation:

```
> HasBoth <- rowSums(Me) > 0 & rowSums(Un) > 0
> table(HasCoverage, HasBoth)

           HasBoth
HasCoverage   FALSE    TRUE
      FALSE 1601772  295891
      TRUE   118785  254912
```

The DGEList object is subsetted to retain only the non-filtered loci:

```
> y <- yall[HasCoverage & HasBoth,, keep.lib.sizes=FALSE]
```

A key difference between BS-seq and other sequencing data is that the pair of libraries holding the methylated and unmethylated reads for a particular sample are treated as a unit. To ensure that the methylated and unmethylated reads for the same sample are treated on the same scale, we need to set the library sizes to be equal for each pair of libraries. We set the library sizes for each sample to be the average of the total read counts for the methylated and unmethylated libraries:

```
> TotalLibSize <- y$samples$lib.size[Methylation=="Me"] +
+                 y$samples$lib.size[Methylation=="Un"]
> y$samples$lib.size <- rep(TotalLibSize, each=2)
> y$samples
```

```
             group lib.size norm.factors
40-45um-A-Me     1 20854816            1
40-45um-A-Un     1 20854816            1
40-45um-B-Me     1 39584537            1
40-45um-B-Un     1 39584537            1
50-55um-A-Me     1 22644990            1
50-55um-A-Un     1 22644990            1
50-55um-B-Me     1 25264124            1
50-55um-B-Un     1 25264124            1
60-65um-A-Me     1 18974220            1
60-65um-A-Un     1 18974220            1
60-65um-B-Me     1 20462334            1
60-65um-B-Un     1 20462334            1
```

Other normalization methods developed for RNA-seq data are not required for BS-seq data.

4.7.4 Data exploration

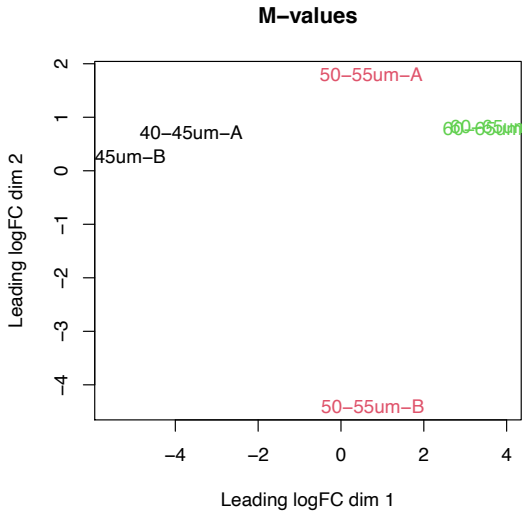
The data can be explored by generating multi-dimensional scaling (MDS) plots on the methylation level (M-value) of the CpG sites. The M-value is calcualted by the log of the ratio of methylated and unmethylated C's, which is equivalent to the difference between methylated and unmethylated C's on the log-scale [8]. A prior count of 2 is added to avoid logarithms of zero.

```
> Me <- y$counts[, Methylation=="Me"]
> Un <- y$counts[, Methylation=="Un"]
> M <- log2(Me + 2) - log2(Un + 2)
> colnames(M) <- targets$Sample
```

Here `M` contains the empirical logit methylation level for each CpG site in each sample. We have used a prior count of 2 to avoid logarithms of zero.

Now we can generate a multi-dimensional scaling (MDS) plot to explore the overall differences between the methylation levels of the different samples.

```
> plotMDS(M, col=rep(1:3, each=2), main="M-values")
```



Replicate samples cluster together within the 40-45 and 60-65μm categories but are far apart in the 50-55μm group. The plot also indicates a huge difference in methylation level between the 40-45 and 60-65μm groups.

4.7.5 The design matrix

One aim of this study is to identify differentially methylated (DM) loci between the different cell populations. In edgeR, this can be done by fitting linear models under a specified design matrix and testing for corresponding coefficients or contrasts. A basic sample-level design matrix can be made as follows:

```
> designSL <- model.matrix(~0+Group, data=targets)
> designSL

  Group40um Group50um Group60um
1         1         0         0
2         1         0         0
3         0         1         0
4         0         1         0
5         0         0         1
6         0         0         1
attr(,"assign")
[1] 1 1 1
attr(,"contrasts")
attr(,"contrasts")$Group
[1] "contr.treatment"
```

The we expand this to the full design matrix modeling the sample and methylation effects:

```
> design <- modelMatrixMeth(designSL)
> design

   Sample1 Sample2 Sample3 Sample4 Sample5 Sample6 Group40um Group50um
```

```
1        1       0       0       0       0       0         1         0
2        1       0       0       0       0       0         0         0
3        0       1       0       0       0       0         1         0
4        0       1       0       0       0       0         0         0
5        0       0       1       0       0       0         0         1
6        0       0       1       0       0       0         0         0
7        0       0       0       1       0       0         0         1
8        0       0       0       1       0       0         0         0
9        0       0       0       0       1       0         0         0
10       0       0       0       0       1       0         0         0
11       0       0       0       0       0       1         0         0
12       0       0       0       0       0       1         0         0
   Group60um
1          0
2          0
3          0
4          0
5          0
6          0
7          0
8          0
9          1
10         0
11         1
12         0
```

The first six columns represent the sample coverage effects. The last three columns represent the methylation levels (in logit units) in the three groups.

4.7.6 Estimating the dispersion

For simplicity, we only consider the CpG methylation in chromosome 1. We subset the coverage files so that they only contain methylation information of the first chromosome.

```
> y1 <- y[y$genes$Chr==1, ]
```

We estimate the NB dispersion for each CpG site using the `estimateDisp` function. The mean-dispersion relationship of BS-seq data has been studied in the past and no apparent mean-dispersion trend was observed [10]. Therefore, we would not consider a mean-dependent dispersion trend for BS-seq methylation data.

```
> y1 <- estimateDisp(y1, design=design, trend="none")
> y1$common.dispersion

[1] 0.384

> summary(y1$prior.df)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    Inf     Inf     Inf     Inf     Inf     Inf
```

The estimated prior degrees of freedom are infinite for all the CpGs, which implies all the CpG-wise dispersions are exactly the same as the common dispersion. A BCV plot is often useful to visualize the dispersion estimates, but it is not informative in this case.

4.7.7 Differential methylation analysis at CpG loci

Then we can proceed to testing for differentially methylated CpG sites between different groups. We fit NB GLMs for all the CpG loci.

```
> fit <- glmFit(y1, design)
```

We identify differentially methylated CpG loci between the 40-45 and 60-65μm group using the likelihood-ratio test. The contrast corresponding to this comparison is constructed using the `makeContrasts` function.

```
> contr <- makeContrasts(
+     Group60vs40 = Group60um - Group40um, levels=design)
> lrt <- glmLRT(fit, contrast=contr)
```

The top set of most significant DMRs can be examined with `topTags`. Here, positive log-fold changes represent CpG sites that have higher methylation level in the 60-65μm group compared to the 40-45μm group. Multiplicity correction is performed by applying the Benjamini-Hochberg method on the p-values, to control the false discovery rate (FDR).

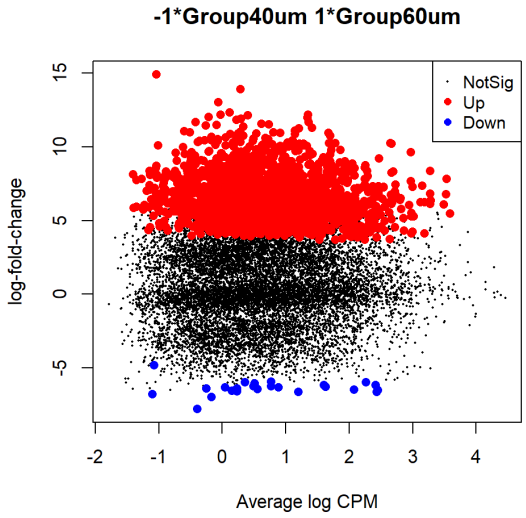
```
> topTags(lrt)

Coefficient:  -1*Group40um 1*Group60um
            Chr     Locus EntrezID        Symbol Strand Distance  Width logFC
1-172206751   1 172206751    18611        Pea15a      -      -53  10077  13.9
1-141992739   1 141992739    75910 4930590L20Rik      -  1336337  86227  11.4
1-131987595   1 131987595   212980       Slc45a3      +   -16986  12364  10.8
1-169954561   1 169954561    15490       Hsd17b7      -   -14644  19669  12.2
1-74571516    1  74571516    77264        Zfp142      -   -16512  21603  13.0
1-36499377    1  36499377    94218         Cnnm3      +    12490  16370  14.9
1-89533694    1  89533694   347722         Agap1      +   -78883 440472  12.0
1-172206570   1 172206570    18611        Pea15a      -     -234  10077  10.3
1-75475455    1  75475455    74241          Chpf      -    -4016   4903  12.3
1-51978650    1  51978650    20849         Stat4      +     8498 120042  12.2
             logCPM   LR   PValue      FDR
1-172206751  0.2784 46.5 9.32e-12 1.33e-07
1-141992739  0.3304 41.9 9.59e-11 5.43e-07
1-131987595  1.6943 41.6 1.14e-10 5.43e-07
1-169954561  1.3471 40.8 1.73e-10 6.14e-07
1-74571516  -0.0658 40.0 2.60e-10 7.41e-07
1-36499377  -1.0398 39.0 4.22e-10 8.08e-07
1-89533694   1.3383 38.9 4.48e-10 8.08e-07
1-172206570  1.6996 38.7 5.05e-10 8.08e-07
1-75475455   0.1106 38.6 5.11e-10 8.08e-07
1-51978650   0.4010 38.2 6.25e-10 8.90e-07
```

The total number of DMRs in each direction at a FDR of 5% can be examined with `decideTests`.

```
> summary(decideTests(lrt))
       -1*Group40um 1*Group60um
Down                         24
NotSig                    12473
Up                         1738
```

The differential methylation results can be visualized using an MD plot. The difference of the M-value for each CpG site is plotted against the average abundance of that CpG site. Significantly DMRs at a FDR of 5% are highlighted.

```
> plotMD(lrt)
```



It can be seen that most of the DMRs have higher methylation levels in 60-65μm group compared to the 40-45μm group. This is consistent with the findings in Gahurova *et al.* [13].

4.7.8 Summarizing counts in promoter regions

It is usually of great biological interest to examine the methylation level within the gene promoter regions. For simplicity, we define the promoter of a gene as the region from 2kb upstream to 1kb downstream of the transcription start site of that gene. We then subset the CpGs to those contained in a promoter region.

```
> InPromoter <- yall$genes$Distance >= -1000 & yall$genes$Distance <= 2000
> yIP <- yall[InPromoter,,keep.lib.sizes=FALSE]
```

We compute the total counts for each gene promoter:

```
> ypr <- rowsum(yIP, yIP$genes$EntrezID, reorder=FALSE)
> ypr$genes$EntrezID <- NULL
```

The integer matrix `ypr$counts` contains the total numbers of methylated and unmethylated CpGs observed within the promoter of each gene.

Filtering is performed in the same way as before. We sum up the read counts of both methylated and unmethylated Cs at each gene promoter within each sample.

```
> Mepr <- ypr$counts[,Methylation=="Me"]
> Unpr <- ypr$counts[,Methylation=="Un"]
> Coveragepr <- Mepr + Unpr
```

Since each row represents a 3,000-bps-wide promoter region that contains multiple CpG sites, we would expect less filtering than before.

```
> HasCoveragepr <- rowSums(Coveragepr >= 8) == 6
> HasBothpr <- rowSums(Mepr) > 0 & rowSums(Unpr) > 0
> table(HasCoveragepr, HasBothpr)

             HasBothpr
HasCoveragepr FALSE  TRUE
        FALSE  3658  3056
        TRUE     85 15047

> ypr <- ypr[HasCoveragepr & HasBothpr,,keep.lib.sizes=FALSE]
```

Same as before, we do not perform normalization but set the library sizes for each sample to be the average of the total read counts for the methylated and unmethylated libraries.

```
> TotalLibSizepr <- 0.5*ypr$samples$lib.size[Methylation=="Me"] +
+                   0.5*ypr$samples$lib.size[Methylation=="Un"]
> ypr$samples$lib.size <- rep(TotalLibSizepr, each=2)
> ypr$samples

             group lib.size norm.factors
40-45um-A-Me     1  8016393            1
40-45um-A-Un     1  8016393            1
40-45um-B-Me     1 11769409            1
40-45um-B-Un     1 11769409            1
50-55um-A-Me     1  9989941            1
50-55um-A-Un     1  9989941            1
50-55um-B-Me     1  8507400            1
50-55um-B-Un     1  8507400            1
60-65um-A-Me     1  8090161            1
60-65um-A-Un     1  8090161            1
60-65um-B-Me     1  6500575            1
60-65um-B-Un     1  6500575            1
```

4.7.9 Differential methylation in gene promoters

We estimate the NB dispersions using the `estimateDisp` function. For the same reason, we do not consider a mean-dependent dispersion trend as we normally would for RNA-seq data.

```
> ypr <- estimateDisp(ypr, design, trend="none")
> ypr$common.dispersion

[1] 0.243
```

```
> ypr$prior.df
[1] 10.4
```

We fit NB GLMs for all the gene promoters using `glmFit`.

```
> fitpr <- glmFit(ypr, design)
```

Then we can proceed to testing for differential methylation in gene promoter regions between different populations. Suppose the comparison of interest is the same as before. The same contrast can be used for the testing.

```
> lrtpr <- glmLRT(fitpr, contrast=contr)
```

The top set of most differentially methylated gene promoters can be viewed with `topTags`:

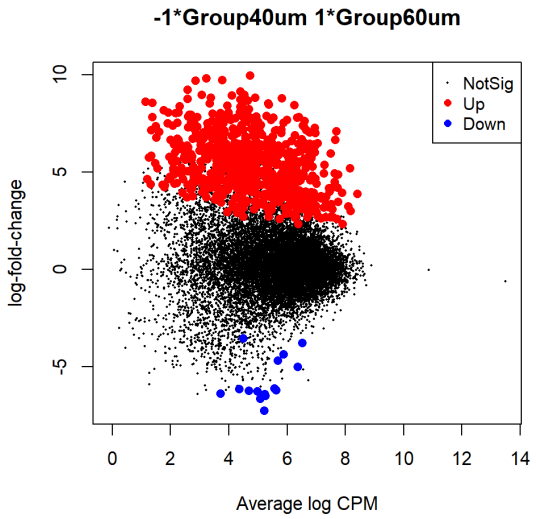
```
> topTags(lrtpr, n=20)
Coefficient:  -1*Group40um 1*Group60um
          Chr        Symbol Strand logFC logCPM   LR   PValue      FDR
78102      15 8430426J06Rik      -  7.80   5.53 84.5 3.87e-20 5.82e-16
210274      7        Shank2      +  7.32   6.56 79.6 4.63e-19 3.48e-15
100038353  18       Gm10532      +  7.76   4.79 74.9 4.83e-18 2.42e-14
102465670  11       Mir7115      +  8.23   4.50 72.3 1.87e-17 7.03e-14
15552       4         Htr1d      +  7.02   6.96 68.7 1.13e-16 2.97e-13
246257     11         Ovca2      -  7.62   6.60 68.6 1.18e-16 2.97e-13
30841       5         Kdm2b      -  6.64   7.64 67.7 1.89e-16 4.07e-13
226527      1        Cryzl2      +  8.97   4.54 66.6 3.35e-16 6.29e-13
114483644  17       Gm51291      +  8.80   5.81 66.3 3.97e-16 6.63e-13
20410      14        Sorbs3      -  6.43   6.92 64.0 1.25e-15 1.87e-12
104184     11          Blmh      +  7.53   5.87 62.0 3.46e-15 4.70e-12
102466776  17       Mir6966      +  7.41   4.43 61.7 3.96e-15 4.70e-12
102466209  14       Mir6947      +  6.92   5.13 61.7 4.06e-15 4.70e-12
72446       2         Prr5l      -  7.39   6.54 61.4 4.76e-15 5.12e-12
217198     11       Plekhh3      -  7.03   6.76 59.7 1.10e-14 1.10e-11
18611       1        Pea15a      -  7.18   5.79 58.8 1.73e-14 1.62e-11
237336     10         Tbpl1      -  7.09   7.69 58.6 1.96e-14 1.74e-11
75480       2 1700003F12Rik      +  9.15   4.39 58.3 2.23e-14 1.84e-11
19894       5         Rph3a      -  6.60   6.14 58.2 2.33e-14 1.84e-11
212307     18        Mapre2      +  6.36   6.96 58.1 2.49e-14 1.88e-11
```

The total number of DM gene promoters identified at an FDR of 5% can be shown with `decideTests`.

```
> summary(decideTests(lrtpr))
       -1*Group40um 1*Group60um
Down                         15
NotSig                    14334
Up                          698
```

The differential methylation results can be visualized with an MD plot.

```
> plotMD(lrtpr)
```



4.7.10 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] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] edgeR_3.31.5     limma_3.45.18    knitr_1.30       BiocStyle_2.17.1

loaded via a namespace (and not attached):
```

```
 [1] Rcpp_1.0.5           compiler_4.0.3       pillar_1.4.6
 [4] BiocManager_1.30.10  highr_0.8            tools_4.0.3
 [7] digest_0.6.25        bit_4.0.4            evaluate_0.14
[10] RSQLite_2.2.1        memoise_1.1.0        lifecycle_0.2.0
[13] tibble_3.0.4         lattice_0.20-41      pkgconfig_2.0.3
[16] rlang_0.4.8          DBI_1.1.0            parallel_4.0.3
[19] yaml_2.2.1           xfun_0.18            org.Mm.eg.db_3.12.0
[22] stringr_1.4.0        IRanges_2.23.10      S4Vectors_0.27.14
[25] vctrs_0.3.4          hms_0.5.3            locfit_1.5-9.4
[28] stats4_4.0.3         bit64_4.0.5          grid_4.0.3
[31] Biobase_2.49.1       R6_2.4.1             AnnotationDbi_1.51.3
[34] rmarkdown_2.4        readr_1.4.0          blob_1.2.1
[37] magrittr_1.5         htmltools_0.5.0      ellipsis_0.3.1
[40] BiocGenerics_0.35.4  stringi_1.5.3        crayon_1.3.4
```

4.8 Time course RNA-seq experiments of Drosophila melanogaster

4.8.1 Introduction

The data for this case study was generated by Graveley *et al.* [14] and was previously analyzed by Law *et al.* [18] using polynomial regression. Here we reanalyze the data using smoothing splines to illustrate a general approach that can be taken to time-course data with many time points. The approach taken here does not require biological replicates at each time point — we can instead estimate the magnitude of biological variation from the smoothness or otherwise of the time-course expression trend for each gene.

Graveley *et al.* conducted RNA-seq to examine the dynamics of gene expression throughout developmental stages of the common fruit fly (*Drosophila melanogaster*). 30 whole-animal samples representing 27 distinct stages of development were used for sequencing. These included 12 embryonic samples collected at 2-hour intervals from 0–2 hours to 22–24 hours and also six larval, six pupal and three sexed adult stages at 1, 5 and 30 days after eclosion. Each biological sample was sequenced several times and we view these as technical replicates. Here we analyze only the data from the 12 embryonic stages.

RNA-seq read counts for this data are available from the ReCount [11] at . http://bowtie-bio.sourceforge.net. The table of read counts can be read into R directly from the ReCount website by

```
> CountFile <- paste("http://bowtie-bio.sourceforge.net/recount",
+                    "countTables",
+                    "modencodefly_count_table.txt", sep="/")
> Counts <- read.delim(CountFile, row.names=1)
```

The sample information can be read by

```
> SampleFile <- paste("http://bowtie-bio.sourceforge.net/recount",
+                     "phenotypeTables",
+                     "modencodefly_phenodata.txt", sep="/")
```