Open Peer Review

Discuss this article

Comments (0)

METHOD ARTICLE

Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR [version 1; referees: awaiting peer review]

Yunshun Chen¹,², Bhupinder Pal¹,², Jane E. Visvader¹,², Gordon K. Smyth³

¹Department of Medical Biology, The University of Melbourne, Melbourne, VIC, 3010, Australia
²The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, 3052, Australia
³School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC, 3010, Australia

Abstract

Studies in epigenetics have shown that DNA methylation is a key factor in regulating gene expression. Aberrant DNA methylation is often associated with DNA instability, which could lead to development of diseases such as cancer. DNA methylation typically occurs in CpG context. When located in a gene promoter, DNA methylation often acts to repress transcription and gene expression. The most commonly used technology of studying DNA methylation is bisulfite sequencing (BS-seq), which can be used to measure genomewide methylation levels on the single-nucleotide scale. Notably, BS-seq can also be combined with enrichment strategies, such as reduced representation bisulfite sequencing (RRBS), to target CpG-rich regions in order to save per-sample costs. A typical DNA methylation analysis involves identifying differentially methylated regions (DMRs) between different experimental conditions. Many statistical methods have been developed for finding DMRs in BS-seq data. In this workflow, we propose a novel approach of detecting DMRs using edgeR. By providing a complete analysis of RRBS profiles of epithelial populations in the mouse mammary gland, we will demonstrate that differential methylation analyses can fit into the existing pipelines specifically designed for RNA-seq differential expression studies.

In addition, the edgeR generalized linear model framework offers great flexibilities for complex experimental design, while still accounting for the biological variability. The analysis approach illustrated in this article can be applied to any BS-seq data that includes some replication, but it is especially appropriate for RRBS data with small numbers of biological replicates.
This article is included in the Bioconductor gateway.

Corresponding author: Gordon K. Smyth (smyth@wehi.edu.au)

Author roles: Chen Y: Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Pal B: Investigation, Writing – Review & Editing; Visvader JE: Funding Acquisition, Supervision, Writing – Review & Editing; Smyth GK: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Software, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Chen Y, Pal B, Visvader JE and Smyth GK. Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR [version 1; referees: awaiting peer review] F1000Research 2017, 6:2055 (doi: 10.12688/f1000research.13196.1)

Copyright: © 2017 Chen Y et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: This work was supported by the National Health and Medical Research Council (Fellowship 1058892 and Program 1054618 to G.K.S, Independent Research Institutes Infrastructure Support to the Walter and Eliza Hall Institute) and by a Victorian State Government Operational Infrastructure Support Grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Introduction**

Studies in the past have shown that DNA methylation, as an important epigenetic factor, plays a vital role in genomic imprinting, X-chromosome inactivation and regulation of gene expression. Aberrant DNA methylation is often correlated with DNA instability, which leads to development of diseases including imprinting disorders and cancer. In mammals, DNA methylation almost exclusively occurs at CpG sites, i.e. regions of DNA where a cytosine (C) is linked by a phosphate (p) and bond to a guanine (G) in the nucleotide sequence from 5’ to 3’. It has been found that 70% - 80% of CpG cytosines are methylated in mammals, regardless of the cell type. Unmethylated CpGs usually group together in clusters of regions known as CpG islands, which cover about 2% of the entire genome. Around 40% of mammalian genes and 70% of human genes have CpG islands enriched in their promoter regions. CpG methylation in gene promoters is generally associated with repression of transcription, and hence silencing of gene expression. When occurring at the promoters of tumor suppressor gene, DNA methylation could repress the tumour suppressors, leading to oncogenesis. In contrast, high levels of methylation have been observed in the gene body of highly expressed genes, which implies positive correlation between gene body methylation and gene expression.

Among numerous existing technologies, the most widely used method to investigate DNA methylation is bisulfite sequencing (BS-seq), which produces data on the single-nucleotide scale. Unmethylated cytosines (C) are converted to Uracils (U) by sodium bisulfite and then deaminated to thymines (T) during PCR amplification. Methylated Cs, on the other hand, remain intact after bisulfite treatment. The BS-seq technique can be used to measure genome-wide single-cytosine methylation levels by sequencing the entire genome. This strategy produces whole genome bisulfite sequencing (WGBS) data. However, the WGBS approach could be cost-prohibitive for species, such as human, with large genome. In addition, the fact that CpG islands reside in only 2% of the entire genome makes the WGBS approach inefficient when comparing a large number of samples.

To improve the efficiency and bring down the scale and cost of WGBS, enrichment strategies have been developed and combined with BS-seq to target a specific fraction of the genome. A common targeted approach is reduced representation bisulfite sequencing (RRBS) that targets CpG-rich regions. Under the RRBS strategy, small fragments that compose only 1% of the genome are generated using MspI digestion, which means fewer reads are required to obtain accurate sequencing. The RRBS approach can capture approximately 70% of gene promoters and 85% of CpG islands, while requiring only small quantities of input sample. In general, RRBS has great advantages in cost and efficiency when dealing with large scale data, whereas WGBS is more suitable for studies where all CpG islands or promoters across the entire genome are of interest.

The first step of analyzing BS-seq data is to align short reads to genome. The number of C-to-T conversions are then counted for all the mapped reads. A number of software tools have been developed for the purposes of read mapping and methylation calling of BS-seq data. Popular ones include Bismark, MethylCoder, BRAT, BSMap, BS-Seeker and Bismark. Most of the software tools rely on existing short read aligners, such as Bowtie.

Typical downstream DNA methylation studies often involve finding differentially methylated regions (DMRs) between different experimental conditions. A number of statistical methods and software packages have been developed for detecting DMRs using the BS-seq technology. methylkit and RnBeads implement Fisher’s Exact Test, which is a popular choice for two-group comparisons with no replicates. In the case of complex experimental designs, regression methods are widely used to model methylation levels or read counts. RnBeads offers a linear regression approach based on the moderated t-test and empirical Bayes method implemented in limma. BSmooth is another analysis pipeline that uses linear regression and empirical Bayes together with a local likelihood smoother. methylkit also has an option to apply logistic regression with overdispersion correction. Some other methods have been developed based on beta-binomial distribution to achieve better variance modelling. For example, DSS fits a Bayesian hierarchical beta-binomial model to BS-seq data and uses Wald tests to detect DMRs. Other software using beta-binomial model include BiSeq, MOABS and RADMeth.

In this workflow, we demonstrate an edgeR approach of differential methylation analysis. edgeR is one of the most popular Bioconductor packages for assessing differential expression in RNA-seq data. It is based on the negative binomial (NB) distribution and it models the variation between biological replicates through the NB
dispersion parameter. Unlike other approaches to methylation sequencing data, the analysis explained in this workflow keeps the counts for methylated and unmethylated reads as separate observations. \textit{edgeR} linear models are used to fit the total read count (methylated plus unmethylated) at each genomic locus, in such a way that the proportion of methylated reads at each locus is modelled indirectly as an over-dispersed binomial distribution. This approach has a number of advantages. First, it allows the differential methylation analysis to be undertaken using existing \textit{edgeR} pipelines developed originally for RNA-seq differential expression analyses. The \textit{edgeR} generalized linear model (GLM) framework offers great flexibility for analysing complex experimental designs while still accounting for the biological variability. Second, keeping methylated and unmethylated read count as separate data observations allows the inherent variability of the data to be modeled more directly and perhaps more realistically. Differential methylation is assessed by likelihood ratio tests so we do not need to assume that the log-fold-changes or other coefficient estimators are normally distributed.

This article presents an analysis of an RRBS data set generated by the authors containing replicated RRBS profiles of basal and luminal cell populations from the mouse mammary epithelium. As with other articles in the Bioconductor Gateway series, our aim is to provide an example analysis with complete start to finish code. As with other Bioconductor workflow articles, we illustrate one analysis strategy in detail rather than comparing different pipelines. The analysis approach illustrated in this article can be applied to any BS-seq data that includes some replication, but is especially appropriate for RRBS data with small numbers of biological replicates. The results shown in this article were generated using Bioconductor Release 3.6.

**The NB linear modeling approach to BS-seq data**

**A small example**

To introduce the \textit{edgeR} linear modeling approach to BS-seq data, consider a genomic locus that has $m_a$ methylated and $u_a$ unmethylated reads in condition A and $m_b$ methylated and $u_b$ unmethylated reads in condition B. Our approach is to model all four counts as NB distributed with the same dispersion but different means. Suppose the data is as given in Table 1. If this were a complete dataset, then it could be analyzed in \textit{edgeR} as follows.

```r
> counts <- matrix(c(2,12,11,0),1,4)
> dimnames(counts) <- list("Locus", c("A_Me","A_Un","B_Me","B_Un"))
> counts
 A_Me A_Un B_Me B_Un
 Locus  2 12 11  0
> design <- cbind(Sample1 = c(1,1,0,0),
   Sample2 = c(0,0,1,1),
   A_MvsU = c(1,0,1,0),
   BvsA_MvsU = c(0,0,1,0))
> fit <- glmFit(counts, design, lib.size=c(100,100,100,100), dispersion=0.0247)
> lrt <- glmLRT(fit, coef="BvsA_MvsU")
> topTags(lrt)
Coefficient:  BvsA_MvsU
 logFC  logCPM  LR  PValue  FDR
Locus  8.99  16.3  20.7  5.27e-06  5.27e-06
```

**Table 1. A small example data set.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Condition</th>
<th>Methylated Count</th>
<th>Unmethylated Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>11</td>
<td>0</td>
</tr>
</tbody>
</table>
In this analysis, the first two coefficients are used to model the total number of reads (methylated or unmethylated) for samples 1 and 2, respectively. Coefficient 3 \((A_{MvsU})\) estimates the log ratio of methylated to unmethylated reads for sample 1, a quantity that can also be viewed as the logit proportion of methylated reads in sample 1. Coefficient 4 \((BvsA_{MvsU})\) estimates the difference in logit proportions of methylated reads between conditions B and A. The difference in logits is estimated here as 8.99 on the log2 scale. The P-value for differential methylation (B vs A) is \(P = 5.27 \times 10^{-6}\).

The dispersion parameter controls the degree of biological variability\(^2\). If we had set \text{dispersion}=0\) in the above code, then the above analysis would be exactly equivalent to a logistic binomial regression, with the methylated counts as responses and the total counts as sizes, and with a likelihood ratio test for a difference in proportions between conditions A and B. Positive values for the dispersion produce over-dispersion relative to the binomial distribution. We have set the dispersion here equal to the value that is estimated below for the mammary epithelial data.

In the above code, the two library sizes for each sample should be equal. Otherwise, the library size values are arbitrary and any settings would have lead to the same P-value.

**Relationship to beta-binomial modeling**

It is interesting to compare this approach with beta-binomial modeling. It is well known that if \(m\) and \(u\) are independent Poisson random variables with means \(\mu_m\) and \(\mu_u\), then the conditional distribution of \(m\) given \(m + u\) is binomial with success probability \(p = \mu_m / (\mu_m + \mu_u)\). If the Poisson means \(\mu_m\) and \(\mu_u\) themselves follow gamma distributions, then the marginal distributions of \(m\) and \(u\) are NB instead of Poisson. If the two NB distributions have different dispersions, and have expected values in inverse proportion to the dispersions, then the conditional distribution of \(m\) given \(m + u\) follows a beta-binomial distribution. The approach taken in this article is closely related to the beta-binomial approach but makes different and seemingly more natural assumptions about the NB distributions. We instead assume the two NB distributions to have the same dispersion but different means. The NB linear modeling approach allows the means and dispersions of the two NB distributions to be estimated separately, in concordance with the data instead of being artificially linked.

**Description of the biological experiment**

**Aim of the study**

The epithelium of the mammary gland exists in a highly dynamic state, undergoing dramatic morphogenetic changes during puberty, pregnancy, lactation, and regression\(^3\). Characterization of the lineage hierarchy of cells in the mammary epithelium is an important step toward understanding which cells are predisposed to oncogenesis. In this study, we profiled the methylation status of the two major functionally distinct epithelial compartments: basal and luminal cells. The basal cells were further divided into those showing high or low expression of the surface marker Itga5 as part of our investigation of heterogeneity within the basal compartment. We carried out global RRBS DNA methylation assays on two biological replicates of each of the three cell populations to determine whether the epigenetic machinery played a potential role in (i) differentiation of luminal cells from basal and (ii) any compartmentalization of the basal cells associated with Itga5.

**Sample preparation**

Inguinal mammary glands (minus lymph node) were harvested from FVB/N mice. All animal experiments were conducted using mice bred at and maintained in our animal facility, according to the Walter and Eliza Hall Institute of Medical Research Animal Ethics Committee guidelines. Epithelial cells were suspended and fluorescence-activated cell sorting (FACS) was used to isolate basal and luminal cell populations\(^4\). Genomic DNA (gDNA) was extracted from freshly sorted cells using the Qiagen DNeasy kit. Around 25ng gDNA input was subjected to DNA methylation analysis by BS-seq using the Ovation RRBS Methyl-seq kit from NuGEN. The process includes MspI digestion of gDNA, sequencing adapter ligation, end repair, bisulfite conversion, and PCR amplification to produce the final sequencing library. The Qiagen EpiTect Bisulfite kit was used for bisulfite-mediated conversion of unmethylated cytosines.
Experimental design
There are three groups of samples: luminal population, Itga5- basal population and Itga5+ basal population. Two biological replicates were collected for each group. This experimental design is summarized in the table below.

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

  Sample  Population  Description
  1  P6_1          P6          Luminal
  2  P6_4          P6          Luminal
  3  P7_2          P7  Basal_Itga5_neg
  4  P7_5          P7  Basal_Itga5_neg
  5  P8_3          P8  Basal_Itga5_pos
  6  P8_6          P8  Basal_Itga5_pos
```

The experiment has a simple one-way layout with three groups. A single grouping factor is made as follows:

```r
> Group <- factor(targets$Population)
> Group

[1] P6 P6 P7 P7 P8 P8
Levels: P6 P7 P8
```

The sequencing was carried out on the Illumina NextSeq 500 platform. About 30 million 75bp paired-end reads were generated for each sample.

Differential methylation analysis at CpG loci

Processing bisulfite sequencing FASTQ files
The first step of the analysis is to map the sequencing reads from the FASTQ files to the mouse genome and then perform methylation calls. Though many options are available, we use Bismark for read alignment and methylation calling. Bismark is one of the most popular software tools to perform alignments of bisulfite-treated sequencing reads to a genome of interest and perform methylation calls. It maps sequencing reads using the short read aligner Bowtie 1\(^1\) or alternatively Bowtie 2\(^2\).

To increase alignment rates and reduce false methylation calls, it is recommended to trim poor quality reads on sequence ends and remove adapters that can be potentially sequenced prior to the alignment. This is done using trim_galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). After that, Bismark version v0.13.0 is used to align the reads to the mouse genome mm10. The final methylation calls are made using bismark_methylation_extractor.

Downloading the data
The Bismark outputs include one coverage bed file of the methylation in CpG context for each sample. The coverage outputs from Bismark are available at http://bioinf.wehi.edu.au/edgeR/F1000Research2017/. Readers wishing to reproduce the analysis presented in this article can download the zipped coverage bed files produced by Bismark from the above link.

Bed files can be read into R using read.delim as for txt files. Each of the bed files has the following format:

```r
> P6_1 <- read.delim("P6_1.bismark.cov.gz", header=FALSE)
> head(P6_1)

 V1   V2   V3   V4   V5   V6
 1 chr6 3052156 3052156  87.9  51  7
 2 chr6 3052157 3052157  85.7  88  3
 3 chr6 3052246 3052246  99.0  54  1
 4 chr6 3052415 3052415  99.5  69  3
 5 chr6 3052416 3052416  99.0  54  0
 6 chr6 3052434 3052434  96.3  70  3
```
The columns in the bed file represent: V1: chromosome number; V2: start position of the CpG site; V3: end position of the CpG site; V4: methylation proportion; V5: number of methylated Cs; V6: number of unmethylated Cs.

Reading in the data
Since the start and end positions in the coverage outputs are identical for each CpG site, only one of them is needed for marking the location of each. We also ignore the methylation proportion as it can be directly calculated from the number of methylated and unmethylated Cs. The data can then be read into a list in R.

```r
> Sample <- targets$Sample
> fn <- paste0(Sample,".bismark.cov.gz")
> data <- list()
> for(i in 1:length(Sample)) {
+     data[[i]] <- read.delim(file=fn[i], header=FALSE)[,-(3:4)]
+     names(data[[i]]) <- c("Chr", "Position", "Meth", "Un")
+ }
```

The data object is a list containing six data frames, each of which represents one sample. The first and second columns of each data frame are the chromosome numbers and positions of all the CpG loci observed in that sample. The last two columns contain the numbers of methylated and unmethylated Cs detected at those loci. Since the number of reported CpG loci varies across different samples, care is required to combine the information from all the samples. We first obtain all unique CpG loci observed in at least one of the six samples. This is done by combining the chromosome number and position of each CpG site. Then we extract read counts of methylated and unmethylated Cs at these locations across all the samples and combine them into a count matrix.

```r
> position <- sapply(data, function(x) paste(x[,1], x[,2], sep="-"))
> position_all <- unique(unlist(position))
> counts <- matrix(0L, nrow=length(position_all), ncol=2*length(Sample))
> for(i in 1:length(Sample)) {
+     m <- match(position[[i]], position_all)
+     counts[m, c(2*i-1,2*i)] <- as.matrix(data[[i]][, 3:4])
+ }
```

The counts object is a matrix of integer counts with 12 columns, two for each sample. The odd number of columns contain the numbers of methylated Cs, whereas the even number of columns contain the numbers of unmethylated Cs. The genomic positions are used as the row names of the count matrix.

```r
> rownames(counts) <- position_all
> Sample2 <- rep(Sample, each=2)
> Sample2 <- factor(Sample2)
> Meth <- rep(c("Me","Un"), length(Sample))
> Meth <- factor(Meth, levels=c("Un","Me"))
> colnames(counts) <- paste(Sample2, Meth, sep="-")
> head(counts)
```

<table>
<thead>
<tr>
<th></th>
<th>P6_1-Me</th>
<th>P6_1-Un</th>
<th>P6_4-Me</th>
<th>P6_4-Un</th>
<th>P7_2-Me</th>
<th>P7_2-Un</th>
<th>P7_5-Me</th>
<th>P7_5-Un</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr6-3052156</td>
<td>51</td>
<td>7</td>
<td>62</td>
<td>13</td>
<td>48</td>
<td>3</td>
<td>31</td>
<td>8</td>
</tr>
<tr>
<td>chr6-3052157</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>chr6-3052246</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>chr6-3052415</td>
<td>57</td>
<td>0</td>
<td>75</td>
<td>1</td>
<td>50</td>
<td>1</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>chr6-3052416</td>
<td>7</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>chr6-3052434</td>
<td>54</td>
<td>3</td>
<td>72</td>
<td>4</td>
<td>48</td>
<td>3</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>chr6-3052156</td>
<td>40</td>
<td>9</td>
<td>28</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr6-3052157</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr6-3052246</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr6-3052415</td>
<td>46</td>
<td>1</td>
<td>36</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr6-3052416</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr6-3052434</td>
<td>47</td>
<td>0</td>
<td>36</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We then proceed to the \textit{edgeR} analysis of the methylation data. The \textit{edgeR} package stores data in a simple list-based data object called a \	exttt{DGEList}. We first create a \texttt{DGEList} object using the count matrix generated before. The information of CpG sites is converted into a data frame and stored in the \texttt{genes} component of the \texttt{DGEList} object.

\begin{verbatim}
> library(edgeR)
> options(digits=3)
> Chr <- gsub("-.*$", ",", position_all)
> Position <- gsub("^.*-", ",", position_all)
> Genes <- data.frame(Chr=Chr, Position=Position)
> y <- DGEList(counts, genes=Genes, group=rep(Group,each=2))
\end{verbatim}

**Filtering to remove low counts**

We first sum up the read counts of both methylated and unmethylated Cs at each CpG site within each sample.

\begin{verbatim}
> counts_total <- t(rowsum(t(counts), Sample2))
> head(counts_total)

P6_1 P6_4 P7_2 P7_5 P8_3 P8_6
chr6-3052156 58 75 51 39 49 38
chr6-3052157 7 5 0 4 0 2
chr6-3052246 1 0 0 2 2 2
chr6-3052415 57 76 51 37 47 36
chr6-3052416 7 5 0 4 0 2
chr6-3052434 57 76 51 37 47 36
\end{verbatim}

CpG loci that have very low counts across all the samples shall be removed prior to downstream analysis as they provide little information for assessing methylation levels. As a rule of thumb, we require a CpG site to have a total count (both methylated and unmethylated) of at least 10 across all the samples before it is considered in the study.

\begin{verbatim}
> keep <- rowSums(counts_total >= 10) == 6
> table(keep)

keep
 FALSE TRUE
3139160 398926
\end{verbatim}

The \texttt{DGEList} object is subsetted to retain only the non-filtered loci:

\begin{verbatim}
> y <- y[keep,, keep.lib.sizes=FALSE]
\end{verbatim}

The option \texttt{keep.lib.sizes=FALSE} causes the library sizes to be recomputed after the filtering. This is generally recommended, although the effect on the downstream analysis is usually small.

**Normalization**

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:

\begin{verbatim}
> TotalReadCount <- colMeans(matrix(y$samples$lib.size, nrow=2, ncol=6))
> y$samples$lib.size <- rep(TotalReadCount, each=2)
> y$samples
\end{verbatim}
Other normalization methods developed for RNA-seq data, such as TMM\textsuperscript{32}, are not required for BS-seq data.

**Exploring differences between samples**

In DNA methylation studies, methylation levels are of most interest. For Illumina methylation assay, two common measurements of methylation levels are $\beta$-values and M-values, which are defined as $\beta = \frac{M}{M+U}$ and $M-value = \log_{2} \left( \frac{M}{U} \right)$ where $M$ and $U$ denote the methylated and unmethylated intensity\textsuperscript{33}. Here we adopt the same idea and extend the two measurements to BS-seq data. That is, denote the methylated and unmethylated Cs by $M$ and $U$ respectively, and define the $\beta$-values and M-values in the same way as above.

In practice, for a particular CpG site in one sample, the M-value can be computed by subtracting the log2 count-per-million (CPM) of the unmethylated Cs from that of the methylated Cs. This is equivalent to the calculation of the defined M-values as the library sizes are set to be the same for each pair of methylated and unmethylated columns and they cancel each other out in the subtraction. A prior count of 2 is added to the calculation of log2-CPM to avoid undefined values and to reduce the variability of M-values for CpG sites with low counts. The calculation of $\beta$-value is straightforward though a small offset may also be added to the calculation.

\begin{verbatim}
> Beta <- y$counts[, Meth=="Me"] / counts_total[keep, ]
> logCPM <- cpm(y, log=TRUE, prior.count=2)
> M <- logCPM[, Meth=="Me"] - logCPM[, Meth=="Un"]
> colnames(Beta) <- colnames(M) <- Sample
\end{verbatim}

The outputs $\beta$ and $M$ are numeric matrices with six columns, each of which contains the $\beta$-values or M-values calculated at each CpG site in one sample. Then we can generate multi-dimensional scaling (MDS) plots to explore the overall differences between the methylation levels of the different samples. Here we decorate the MDS plots to indicate the cell groups:

\begin{verbatim}
> par(mfrow=c(1,2))
> plotMDS(Beta, col=rep(1:3, each=2), main="Beta-values")
> plotMDS(M, col=rep(1:3, each=2), main="M-values")
\end{verbatim}

Figure 1 shows the resulting plots. In these plots, the distance between each pair of samples represents the average log-fold change between the samples for the top most differentially methylated CpG loci between that pair of samples. (We call this average the leading log-fold change.) The two replicate samples from the luminal population (P6) are seen to be well separated from the four basal samples (populations P7 and P8).
Design matrix
One aim of this study is to identify differentially methylated regions (DMRs) between different groups. In `edgeR`, this can be done by fitting linear models under a specified design matrix and testing for corresponding coefficients or contrasts. Here, a design matrix is constructed as follows:

```r
> design <- model.matrix(~ Sample2 + Meth)
> colnames(design) <- gsub("Sample2", "", colnames(design))
> colnames(design) <- gsub("Meth", "", colnames(design))
> colnames(design)[1] <- "Int"
> design <- cbind(design,
+   Me2=rep(0, 12),
+   Me3=rep(0, 12))

> design
Int P6_4 P7_2 P7_5 P8_3 P8_6 Me Me2 Me3
 1  1  0  0  0  0  0  1  0  0
 2  1  0  0  0  0  0  0  0  0
 3  1  1  0  0  0  0  1  0  0
 4  1  1  0  0  0  0  0  0  0
 5  1  0  1  0  0  0  1  1  0
 6  1  0  1  0  0  0  0  0  0
 7  1  0  0  1  0  0  1  1  0
 8  1  0  0  1  0  0  0  0  0
 9  1  0  0  0  1  0  1  0  1
10  1  0  0  0  1  0  0  0  0
11  1  0  0  0  0  1  1  0  1
12  1  0  0  0  0  1  0  0  0
```

The first six columns represent the sample effect. It accounts for the fact that each pair of columns of the count matrix are from one of the six samples. The 7th column “Me” represents the methylation level (in M-value)

![Figure 1. The MDS plots of the methylation levels of the data set.](image)

Methylation levels are measured in beta values (left) and M-values (right). Samples are separated by the cell population in the first dimension in both MDS plots.
in the P6 group. The 8th column “Me2” represents the difference in methylation level between the P7 and P6 groups. Finally, the last column “Me3” represents the difference in methylation level between the P8 and P6 groups.

**Dispersion estimation**

With the design matrix specified, we can now proceed to the standard *edgeR* pipeline and analyze the data in the same way as for RNA-seq data. Similar to the RNA-seq data, the variability between biological replicates has also been observed in bisulfite sequencing data. This variability can be captured by the NB dispersion parameter under the generalized linear model (GLM) framework in *edgeR*.

The mean-dispersion relationship of BS-seq data has been studied in the past and no apparent mean-dispersion trend was observed\(^2\). This is also verified through our own practice. Therefore, we would not consider a mean-dependent dispersion trend as we normally would for RNA-seq data. A common dispersion estimate for all the loci, as well as an empirical Bayes moderated dispersion for each individual locus, can be obtained from the `estimateDisp` function in *edgeR*:

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

[1] 0.0247
```

```r
> summary(y$prior.df)

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

This returns a DGEList object with additional components (common.dispersion and tagwise.dispersion) added to hold the estimated dispersions. Here the estimation of trended dispersion has been turned off by setting `trend="none"`. For this data, the estimated prior degrees of freedom (df) are infinite for all the loci, 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.

**Testing for differentially methylated CpG loci**

We first fit NB GLMs for all the CpG loci using the `glmFit` function in *edgeR*.

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

Then we can proceed to testing for differentially methylated CpG sites between different populations. One of the most interesting comparisons is between the basal (P7 and P8) and luminal (P6) groups. The contrast corresponding to any specified comparison can be constructed conveniently using the `makeContrasts` function:

```r
> contr <- makeContrasts(BvsL=0.5*(Me2+Me3), levels=design)
```

The actual testing is performed using likelihood ratio tests (LRT) in *edgeR*.

```r
> lrt <- glmLRT(fit, contrast=contr)
```

The top set of most differentially methylated (DM) CpG sites can be viewed with `topTags`:

```r
> topTags(lrt)
```

| Coefficient: 0.5*Me2 0.5*Me3 | Chr | Position logFC logCPM LR PValue FDR |
|-----------------------------|-----|----------------------|-----------------|-----------------|--------------------------|
| chr13-45709467              | chr13 | 45709467  -7.60    | 3.06 342 2.71e-76 9.62e-71 |
| chr16-76326604              | chr16 | 76326604  8.87     | 2.60 341 4.82e-76 9.62e-71 |
| chr10-40387375              | chr10 | 40387375  8.13     | 2.85 333 2.48e-74 3.30e-69 |
Here positive log-fold changes represent CpG sites that have higher methylation level in the basal population compared to the luminal population. The Benjamini-Hochberg multiple testing correction is applied to control the false discovery rate (FDR).

The total number of DM CpG sites identified at an FDR of 5% can be shown with `decideTestsDGE`. There are in fact more than 50,000 differentially methylated CpGs in this comparison:

```r
> summary(decideTests(lrt))

  0.5*Me2  0.5*Me3
-1   35891
 0  344846
 1   18189
```

The differential methylation results can be visualized with an MD plot (see Figure 2):

```r
> plotMD(lrt)
```

**Figure 2.** MD plot showing the log-fold change of the methylation level and average abundance of each CpG site. Significantly up and down methylated CpGs are highlighted in red and blue, respectively.
The logFC of the methylation level for each CpG site is plotted against the average abundance in log2-CPM. Significantly differentially methylated CpGs are highlighted.

**Differential methylation in gene promoters**

**Pre-defined gene promoters**

The majority of CpGs are methylated in mammals. On the other hand, unmethylated CpGs tend to group into clusters of CpG islands, which are often enriched in gene promoters. CpG methylation in promoter regions is often associated with silencing of transcription and gene expression\(^5\). Therefore it is 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. The genomic locations and their associated annotations of the promoters can be obtained using the `TxDb.Mmusculus.UCSC.mm10.knownGene` package.

```r
> library(TxDB.Mmusculus.UCSC.mm10.knownGene)
> genes_Mm <- genes(TxDB.Mmusculus.UCSC.mm10.knownGene)
> pr <- promoters(genes_Mm, upstream=2000, downstream=1000)
```

```
GRanges object with 24044 ranges and 1 metadata column:

<table>
<thead>
<tr>
<th>seqnames</th>
<th>ranges</th>
<th>strand</th>
<th>gene_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr9</td>
<td>[21074497, 21077496]</td>
<td>-</td>
<td>100009600</td>
</tr>
<tr>
<td>chr7</td>
<td>[84963010, 84966009]</td>
<td>-</td>
<td>100009609</td>
</tr>
<tr>
<td>chr10</td>
<td>[77709446, 77712445]</td>
<td>+</td>
<td>100009614</td>
</tr>
<tr>
<td>chr11</td>
<td>[45806083, 45809082]</td>
<td>+</td>
<td>100009664</td>
</tr>
<tr>
<td>chr4</td>
<td>[144161652, 144164651]</td>
<td>-</td>
<td>100012</td>
</tr>
</tbody>
</table>

... ... ... ... ...

99889  chr3  [85886519, 85889518]  -  | 99889
99890  chr3  [110250000, 110252999]  -  | 99890
99899  chr3  [151748960, 151751959]  -  | 99899
99929  chr3  [65526444, 65529446]   +  | 99929
99982  chr4  [136601724, 136604723]  -  | 99982

seqinfo: 66 sequences (1 circular) from mm10 genome
```

Here, `pr` is a GRanges class object that contains the genomic ranges of the promoters of all the known mouse genes in the annotation package.

**Summarizing counts in promoter regions**

We create another GRanges class object `sites`, which contains the genomic locations of all the observed CpG sites.

```r
> Position <- as.numeric(Position)
> sites <- GRanges(seqnames=Chr, ranges=IRanges(start=Position, end=Position))
```

Then we find the overlaps between the gene promoter regions and all the CpG sites in the data using `findOverlaps`.
> olap <- findOverlaps(query=pr, subject=sites)
> olap

Hits object with 1522464 hits and 0 metadata columns:

```
queryHits subjectHits
<integer>    <integer>
[1]        3    2493045
[2]        3    2493046
[3]        3    2493047
[4]        3    2493048
[5]        6    1898041
    ...    ...      ...
[1522460]  24044  3077832
[1522461]  24044  3317008
[1522462]  24044  3317009
[1522463]  24044  3317010
[1522464]  24044  3434601
```
queryLength: 24044 / subjectLength: 3538086

The `queryHits` component of `olap` marks the indices of the promoter region as in `pr`, whereas the `subjectHits` component contains the indices of the CpG sites as in `sites` that overlap with the corresponding promoter regions.

The numbers of methylated and unmethylated CpGs overlapping with gene promoters are summed up for each promoter.

```r
> counts2 <- counts[subjectHits{olap}, ]
> counts2 <- rowsum(counts2, queryHits{olap})
```

The integer matrix `counts2` contains the total numbers of methylated and unmethylated CpGs observed within the promoter of each gene. Same as before, `counts2` has 12 columns, two for each sample. The odd number of columns contain the numbers of methylated Cs, whereas the even number of columns contain the numbers of unmethylated Cs. The only difference is that each row of `counts2` now represents a gene promoter instead of an individual CpG site.

The gene symbol information can be added to the annotation using the `org.Mm.eg.db` package. A `DGEList` object is created for the downstream `edgeR` analysis.

```r
> ind <- as.numeric(rownames(counts2))
> rownames(counts2) <- pr$gene_id[ind]
> library(org.Mm.eg.db)
> anno <- select(org.Mm.eg.db, keys=pr$gene_id, columns="SYMBOL",
+ keytype="ENTREZID")
> anno <- data.frame(Symbol=anno$SYMBOL[ind])
> y2 <- DGEList(counts2, genes=anno, group=rep(Group, each=2))
```

We sum up the read counts of both methylated and unmethylated Cs at each CpG sites within each sample.

```r
> counts2_total <- t(rowsum(t(counts2), Sample2))
```

---

Page 14 of 27
Filtering to remove low counts
Filtering is performed in the same way as before. Since each row represents a 3,000-bp-wide promoter region that contains multiple CpG sites, we would expect less filtering than before.

```r
counts2_total <- rowSums(counts2_total) >= 10
keep2 <- table(counts2_total)
keep2
# FALSE  TRUE
# 1754 16790

y2 <- y2[keep2, , 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.

```r
TotalReadCount2 <- colMeans(matrix(y2$counts, nrow=2, ncol=6)
> TotalReadCount2
# P6_1-Me  P6 12474999            1
# P6_1-Un  P6 12474999            1
# P6_4-Me  P6 12579436            1
# P6_4-Un  P6 12579436            1
# P7_2-Me  P7  5110397            1
# P7_2-Un  P7  5110397            1
# P7_5-Me  P7 11189796            1
# P7_5-Un  P7 11189796            1
# P8_3-Me  P8  4123987            1
# P8_3-Un  P8  4123987            1
# P8_6-Me  P8  3562239            1
# P8_6-Un  P8  3562239            1
```

Exploring differences between samples
Same as before, we measure the methylation levels of gene promoter regions using both $\beta$-values and M-values. A prior count of 2 is added to the calculation of log2-CPM to avoid undefined values and to reduce the variability of M-values for gene promoters with low counts. Then MDS plots are produced to examine the overall differences between the methylation levels of the different samples.

```r
Beta2 <- y2$counts[, Meth=="Me"] / counts2_total[keep2, ]
logCPM2 <- cpm(y2, log=TRUE, prior.count=2)
M2 <- logCPM2[, Meth=="Me"] - logCPM2[, Meth=="Un"]
colnames(Beta2) <- colnames(M2) <- Sample
par(mfrow=c(1,2))
plotMDS(Beta2, col=rep(1:3, each=2), main="Beta-values")
plotMDS(M2, col=rep(1:3, each=2), main="M-values")
```

The resulting Figure 3 shows that the two replicate samples from the luminal population (P6) are well separated from the four replicate samples from the basal population (P7 and P8).
Dispersion estimation
We estimate the NB dispersions using the `estimateDisp` function in `edgeR`. For the same reason, we do not consider a mean-dependent dispersion trend as we normally would for RNA-seq data.

```r
> y2 <- estimateDisp(y2, design=design, trend="none", robust=TRUE)
> y2$common.dispersion
[1] 0.0301
> summary(y2$prior.df)

  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
  10.2   10.2   10.3    10.3   10.4    10.4
```

The dispersion estimates can be visualized with a BCV plot (see Figure 4):

```r
> plotBCV(y2)
```

Testing for differential methylation in gene promoters
We first fit NB GLMs for all the gene promoters using `glmFit`.

```r
> fit2 <- glmFit(y2, design)
```

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

```r
> lrt2 <- glmLRT(fit2, contrast=contr)
```
The top set of most differentially methylated gene promoters can be viewed with `topTags`:

```r
> topTags(lrt2)
```

Here positive log-fold changes represent gene promoters that have higher methylation level in the basal population compared to the luminal population. The Benjamini-Hochberg multiple testing correction is applied to control the false discovery rate (FDR).
The total number of DM gene promoters identified at an FDR of 5% can be shown with `decideTestsDGE`. There are in fact about 1,200 differentially methylated gene promoters in this comparison:

```r
> summary(decideTests(lrt2))

               0.5*Me2  0.5*Me3
              -1       817
               0     15617
               1       356
```

The differential methylation results can be visualized with an MD plot (see Figure 5):

```r
> plotMD(lrt2)
```

**Correlate with RNA-seq profiles**

**RNA-seq profiles of mouse epithelium**

To show that DNA methylation (particularly in the promoter regions) represses gene expression, we relate the differential methylation results to the gene expression profiles of the RNA-Seq data. The RNA-seq data used here is from a study of the epithelial cell lineage in the mouse mammary gland\(^3\), in which the expression profiles of basal stem-cell enriched cells and committed luminal cells in the mammary glands of virgin, pregnant and lactating mice were examined. The complete differential expression analysis of the data is described in Chen et al.\(^3\).

![MD plot showing the log-fold change of the methylation level and average abundance of CpG sites in each gene promoter.](image)

Significantly up and down methylated gene promoters are highlighted in red and blue, respectively.
The RNA-seq data is stored in the format of a DGEList object \texttt{y_rna} and saved in a RData file \texttt{rna.RData}. The object \texttt{y_rna} contains the count matrix, sample information, gene annotation, design matrix and dispersion estimates of the RNA-seq data. The gene filtering, normalization and dispersion estimation were performed in the same way as described in Chen et al. The DE analysis between the basal and luminal in the virgin mice was performed using \texttt{glmTreat} with a fold-change threshold of 3. The results are saved in the spreadsheet \texttt{BvsL-fc3.csv}. Both \texttt{rna.RData} and \texttt{BvsL-fc3.csv} are available for download at \url{http://bioinf.wehi.edu.au/edgeR/F1000Research2017/}.

We load the RData file and read in the DE results from the spreadsheet.

```r
> load("rna.RData")
> y_rna

An object of class "DGEList"

$counts
 MCL1.DG MCL1.DH MCL1.DI MCL1.DJ MCL1.DK MCL1.DL MCL1.LA MCL1.LB MCL1.LC
497097  438   300   65  237  354   287     0     0     0
20671   106   182   82  105   43    82    16    25    18
27395   309   234  337  300  290  270  560  464  489
18777   652   515  948  935  928   791   826  862  668
21399  1604  1495 1721 1317 1334  1258 1066 1334 1258
15636 more rows ...

$samples

 group   lib.size  norm.factors
 MCL1.DG    B.virgin  23137472  1.23
 MCL1.DH    B.virgin  21687755  1.21
 MCL1.DI    B.pregnant  23974787  1.13
 MCL1.DJ    B.pregnant  22545375  1.07
 MCL1.DK    B.lactating  21420532  1.04
7 more rows ...

$genes

 Length Symbol
 497097  3634    Xkr4
 20671  3130    Sox17
 27395  4203   Mrpl15
 18777  2433   Lypla1
 21399  2847    Tcea1
15636 more rows ...

$design

 B.lactating B.pregnant B.virgin L.lactating L.pregnant L.virgin
 1  0  0  1  0  0  0
 2  0  0  1  0  0  0
 3  0  1  0  0  0  0
$\text{common.dispersion}$

[1] 0.0134

$trended.dispersion$

[1] 0.02086 0.03012 0.01303 0.01007 0.00957

15636 more elements ...

$\text{tagwise.dispersion}$

[1] 0.13795 0.08336 0.01387 0.00678 0.00631

15636 more elements ...

$\text{AveLogCPM}$

[1] 2.58 1.32 4.00 5.06 5.64

15636 more elements ...

$trend.method$

[1] "locfit"

$prior.df$

[1] 4.68 6.08 6.77 6.77 6.77

15636 more elements ...

$prior.n$

[1] 0.78 1.01 1.13 1.13 1.13

15636 more elements ...

$\text{span}$

[1] 0.292

> rna_DE <- read.csv("BvsL-fc3.csv", row.names="GeneID")

> head(rna_DE)

<table>
<thead>
<tr>
<th>Length</th>
<th>Symbol</th>
<th>logFC</th>
<th>unshrunk.logFC</th>
<th>logCPM</th>
<th>PValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>24117</td>
<td>2242</td>
<td>9.15</td>
<td>9.18</td>
<td>6.77</td>
<td>1.79e-15</td>
<td>1.67e-11</td>
</tr>
<tr>
<td>69538</td>
<td>5264</td>
<td>7.35</td>
<td>7.36</td>
<td>7.66</td>
<td>2.67e-15</td>
<td>1.67e-11</td>
</tr>
<tr>
<td>55987</td>
<td>3506</td>
<td>8.15</td>
<td>8.18</td>
<td>6.01</td>
<td>1.67e-15</td>
<td>1.67e-11</td>
</tr>
<tr>
<td>12293</td>
<td>4933</td>
<td>8.30</td>
<td>8.31</td>
<td>7.81</td>
<td>2.74e-15</td>
<td>3.51e-11</td>
</tr>
<tr>
<td>12560</td>
<td>3995</td>
<td>6.98</td>
<td>6.98</td>
<td>7.54</td>
<td>3.51e-11</td>
<td>3.51e-11</td>
</tr>
<tr>
<td>110308</td>
<td>2190</td>
<td>8.94</td>
<td>8.94</td>
<td>10.27</td>
<td>3.51e-11</td>
<td>3.51e-11</td>
</tr>
</tbody>
</table>

**Correlation between the two datasets**

We select the genes of which the promoters are significantly DM (FDR < 0.05) and examine their expression level in the RNA-Seq data. A data frame object lfc is created to store the gene information, log-fold change of methylation level and log-fold change of gene expression of the selected genes.

> tp <- topTags(lrt2, n=Inf, p=0.05)$table
>
m <- match(row.names(tp), row.names(rna_DE))
>
lfc <- tp[,c("Symbol","logFC")]
>
names(lfc)[2] <- "ME"
>
lfc$RNA <- rna_DE$logFC[m]
> lfc <- lfc[!is.na(lfc$RNA), ]
> head(lfc)

       Symbol ME  RNA
16924     Lnx1 6.85 -2.27
238161  Akap6 5.26  3.23
64082  Popdc2 -4.91  7.67
11601  Angpt2 -5.58  2.10
12740   Cldn4 5.56 -5.10
387514 Tas2r143 -4.52  3.10

The Pearson correlation coefficient between the two log-fold changes of the selected genes is estimated. The result shows high negative correlation between gene expression and methylation in gene promoters.

> cor(lfc$ME, lfc$RNA)

[1] -0.47

The log-fold changes of the selected genes from the two datasets are plotted against each other for visualization (see Figure 6):

> plot(lfc$ME, lfc$RNA, main="Basal vs Luminal", xlab="log-FC Methylation", +
     ylab="log-FC Gene Expression", pch=16, cex=0.8, col="gray30")
> abline(h=0, v=0, col="gray10", lty=2, lwd=2)

**Figure 6.** Scatter plot of the log-fold changes of methylation levels in gene promoters (x-axis) vs the log fold-changes of gene expression (y-axis). The plot shows results for the genes of which the promoters are significantly differentially methylated between basal and luminal. The red line shows the least squares line with zero intercept. A strong negative correlation is observed.
The horizontal axis of the scatterplot shows the log-fold change in methylation level for each gene while the vertical axis shows the log-fold change in expression. To assess the correlation, we fit a least squares regression line through the origin and compute the p-value:

```r
> u <- lm(lfc$RNA ~ 0 + lfc$ME)
> coef(summary(u))

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>Std. Error</td>
</tr>
<tr>
<td>lfc$ME</td>
<td>-0.739</td>
</tr>
</tbody>
</table>
```

The negative association is highly significant ($P = 10^{-47}$). The last line of code adds the regression line to the plot (Figure 6).

**Gene set testing**

A rotation gene set test can be performed to further examine the relationship between gene expression and methylation in gene promoters. This is to test whether the set of genes (i.e., genes of which the promoters are differentially methylated) are differentially expressed (DE) and in which direction they are DE.

The indices are made by matching the Entrez Gene Ids between the two datasets. The log-fold changes of methylation level in gene promoters are used as weights for those genes. The test is conducted using the `fry` function in `edgeR`. The contrast is set to compare basal with luminal in virgin mice.

```r
> ME <- data.frame(GeneID=row.names(lfc), weights=lfc$ME)
> fry(y_rna, index=ME, design=y_rna$design, contrast=c(0,0,1,0,0,-1))
```

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NGenes</td>
<td>Direction</td>
<td>PValue</td>
<td>PValue.Mixed</td>
</tr>
<tr>
<td>set1</td>
<td>Down</td>
<td>1.51e-09</td>
<td>7.17e-11</td>
</tr>
</tbody>
</table>

The small PValue indicates the significant testing result. The result Down in the Direction column indicates negative correlation between the methylation and gene expression.

We can visualize the gene set results with a barcode plot (see Figure 7):

```r
> m <- match(row.names(rna_DE), row.names(tp))
> gw <- tp$logFC[m]
> gw[is.na(gw)] <- 0
> barcodeplot(rna_DE$logFC, gene.weights=gw, labels=c("Luminal","Basal"),
+ main="Basal vs Luminal")
> legend("topright", col=c("red","blue"), lty=1, lwd=2,
+ legend=c("Up-methylation in Basal", "Up-methylation in Luminal") )
```

In the barcode plot, genes are sorted left to right according to expression changes. Genes up-regulated in luminal are on the left and genes up-regulated in basal are on the right. The x-axis shows the expression log2-fold change between basal and luminal. The vertical red bars indicate genes up-methylated in basal and vertical blue bars indicate genes down-methylated in basal. The variable-height vertical bars represent the methylation log-fold changes. The red and blue worms measure relative enrichment, showing that increased methylation is associated with decreased regulation and down-methylation is associated with up-regulation. In other words, there is a negative association between methylation of promoter regions and expression of the corresponding gene.
Figure 7. Barcode plot showing strong negative correlation between gene expression and DNA methylation in gene promoters.

Packages used

This workflow depends on various packages from version 3.6 of the Bioconductor project, running on R version 3.4.0 or higher. Most of the workflow also works with Bioconductor 3.5, but the code in the last section (Correlate with RNA-seq samples) requires some minor changes for use with Bioconductor 3.5 because the earlier version of topTags did not preserve row names in the output table. A complete list of the packages used for this workflow is shown below:

> sessionInfo()

R version 3.4.2 (2017-09-28)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 15063)

Matrix products: default

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] stats4 parallel stats graphics grDevices utils datasets methods
[9] base
other attached packages:
[1] org.Mm.eg.db_3.4.2
[2] TxDb.Mmusculus.UCSC.mm10.knownGene_3.4.0
[3] GenomicFeatures_1.30.0
[4] AnnotationDbi_1.40.0
[5] Biobase_2.38.0
[6] GenomicRanges_1.30.0
[8] IRanges_2.12.0
[9] S4Vectors_0.16.0
[10] BiocGenerics_0.24.0
[12] limma_3.34.0
[13] knitr_1.17

loaded via a namespace (and not attached):
[1] Rcpp_0.12.13               compiler_3.4.2
[3] highr_0.6                  XVector_0.18.0
[5] prettyunits_1.0.2          bitops_1.0-6
[7] tools_3.4.2                zlibbioc_1.24.0
[9] progress_1.1.2             statmod_1.4.30
[11] biomaRt_2.34.0             digest_0.6.12
[13] bit_1.1-12                 RSQLite_2.0
[15] evaluate_0.10.1            memoise_1.1.0
[17] tibble_1.3.4               lattice_0.20-35
[19] pkgconfig_2.0.1            rlang_0.1.4
[21] Matrix_1.2-11              DelayedArray_0.4.1
[23] DBI_0.7                    GenomeInfoDbData_0.99.1
[25] rtracklayer_1.38.0          stringr_1.2.0
[27] Biostrings_2.46.0          locfit_1.5-9.1
[29] bit64_0.9-7                 grid_3.4.2
[31] R6_2.2.2                    BiocParallel_1.12.0
[33] XML_3.98-1.9                RMySQL_0.10.13
[35] blob_1.1.0                   magrittr_1.5
[37] matrixStats_0.52.2         GenomicAlignments_1.14.0
[39] Rsamtools_1.30.0           SummarizedExperiment_1.8.0
[41] assertthat_0.2.0           stringi_1.1.5
[43] RCurl_1.95-4.8

Data and software availability
All data and supporting files used in this workflow are available from: http://bioinf.wehi.edu.au/edgeR/F1000Research2017

Archived code/data as at time of publication: http://doi.org/10.5281/zenodo.1052871

All software used is publicly available as part of Bioconductor 3.6.

Competing interests
No competing interests were disclosed.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments
The authors thank Andrew Keniry for help on Bismark.

References


The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com