Overview of the *DMRcaller* package

Nicolae Radu Zabet*

October 27, 2020

Contents

1 Introduction

DNA methylation is an epigenetic modification of the DNA where a methyl group is added to the cytosine nucleotides. This modification is heritable, able to control the gene regulation and, in general, is associated with transcriptional gene silencing. While in mammals the DNA is predominantly methylated in CG context, in plants non-CG methylation (CHG and CHH, where H can be any of the A, C or T nucleotides) is also present and is important for the epigenetic regulation of transcription. Sequencing of bisulfite converted DNA has become the method of choice to determine genome wide methylation distribution. The *DMRcaller* package computes the set of Differentially Methylated Regions (DMRs) between two samples. *DMRcaller* will compute the differentially methylated regions from Whole Genome Bisulfite Sequencing (WGBS) or Reduced Representation Bisulfite Sequencing (RRBS) data. There are several tools able to call DMRs, but most work has been done in mammalian systems and, thus, they were designed to primarily call CG methylation.

2 Methods

The package computes the DMRs using the CX report files generated by Bismark (Krueger and Andrews, 2011), which contain the number of methylated and unmmethylated reads for each cytosine in the genome. The coverage at each position on the genome is not homogeneous and this makes it difficult to compute the differentially methylated cytosines. Here, we implemented three methods:

*e-mail: nzabet@essex.ac.uk, School of Biological Sciences, University of Essex, UK

- **noise_filter** where we use a kernel (Hebestreit et al., 2013) to smooth the number of methylated reads and the total number of reads (the *DMRcaller* package provides four kernels: `"uniform"`, `"triangular"`, `"gaussian"` and `"epanechnicov"`)
- **neighbourhood** where individual cytosines in a specific context are considered in the analysis without any smoothing
- **bins** where the genome is split into equal bins where all the reads are pooled together

The DMRs are then computed by performing a statistical test between the number of methylated reads and the total number of reads in the two conditions for each position, cytosine or bin. In particular, we implemented two statistical tests: (*i*) Fisher's exact test and (*ii*) the Score test. The former (Fisher's exact test) uses the `fisher.test` in the *stats* package.

The Score test is a statistical test of a simple null hypothesis that a parameter of interest is equal to some particular value. In our case, we are interested if the methylation levels in the two samples are equal or different. Given that m_1 is the number of methylated reads in condition 1, m_2 is the number of methylated reads in condition 2, n_1 is the total number of reads in condition 1 and n_2 is the total number of reads in condition 2, the Z-score of the Score test is

$$Z = \frac{(p_1 - p_2)\,\nu}{\sqrt{p(1-p)}} \tag{1}$$

where $p_1 = m_1/n_1$, $p_2 = m_2/n_2$,

$$p = \frac{m_1 + m_2}{n_1 + n_2} \quad \text{and} \quad \nu = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{2}$$

We then convert the Z-score to the p-value assuming a normal distribution and a two sided test.

```
pValue <- 2*pnorm(-abs(zScore))
```

Finally, for both statistical tests (Fisher's exact test and Score test), we adjust the p-values for multiple testing using Benjamini and Hochberg's method (Benjamini and Hochberg, 1995) to control the false discovery

```
pValue <- p.adjust(pValue, method="fdr")
```

The algorithm performs the statistical test for each position, cytosine or bin and then marks as DMRs all positions/cytosines/bins that satisfy the following three conditions:

- the difference in methylation levels between the two conditions is statistically significant according to the statistical test;
- the difference in methylation proportion between the two conditions is higher than a threshold value;
- the mean number of reads per cytosine is higher than a threshold.

To group adjacent DMRs, we run an iterative process, where neighbouring DMRs (within a certain distance of each other) are joined only if these three conditions are still met after joining the DMRs.

Finally, we filter the DMRs as follow

- Remove DMRs whose lengths are less than a minimum size.
- Remove DMRs with fewer cytosines than a threshold value.

For a set of potential DMRs (e.g. genes, transposable elements or CpG islands) the user can call the function `filterDMRs` where all reads in a set of provided regions are pooled together and then the algorithm performs the statistical test for each region.

3 Description

3.1 Data

Bismark (Krueger and Andrews, 2011) is a popular tool for methylation call on WGBS or RRBS data. *DMRcaller* takes as inputs the CX report files generated by Bismark and stores this data in a `GRanges` object. In the package, we included two CX report files that contain the methylation calls of WT and *met1-3 Arabidopsis thaliana* (Stroud et al., 2013). *MET1* gene encodes for the main DNA methyltransferase in *Arabidopsis thaliana* and the *met1-3* mutation results in a genome-wide loss of DNA methylation (mainly in CG context). Due to running time, we restricted the data and analysis to the first 1 Mb of the third chromosome of *A. thaliana*.

```
library(DMRcaller)

#load presaved data
data(methylationDataList)
```

To load a different dataset, one can use `readBismark` function, which takes as input the filename of the CX report file to be loaded.

```
# specify the Bismark CX report files
saveBismark(methylationDataList[["WT"]],
            "chr3test_a_thaliana_wt.CX_report")
saveBismark(methylationDataList[["met1-3"]],
            "chr3test_a_thaliana_met13.CX_report")

# load the data
methylationDataWT <- readBismark("chr3test_a_thaliana_wt.CX_report")
methylationDataMet13 <- readBismark("chr3test_a_thaliana_met13.CX_report")
methylationDataList <- GRangesList("WT" = methylationDataWT,
                                   "met1-3" = methylationDataMet13)
```

`methylationDataList` is a `GRangesList` object , where the `GRanges` elements contain four metadata columns

- **context** - the context of the Cytosine (CG, CHG or CHH)
- **readsM** - the number of methylated reads
- **readsN** - the total number of reads
- **trinucleotide_context** - the specific context of the cytosine (H is replaced by the actual nucleotide)

If the data consists of two or more replicates, these can be pooled together using the function `poolMethylationDatasets` or `poolTwoMethylationDatasets` (in the case of pooling only two datasets). The latter function (`poolTwoMethylationDatasets`) is useful when the datasets are large and creating a `GRangesList` object is not possible (e.g. the `GRanges` objects are two large).

```
# load the data
methylationDataAll <- poolMethylationDatasets(methylationDataList)

# In the case of 2 elements, this is equivalent to
methylationDataAll <- poolTwoMethylationDatasets(methylationDataList[[1]],
                                                 methylationDataList[[2]])
```

Alternatively, one can use `readBismarkPool` to directly read a list of CX report files and pool them together.

```
# load the data
methylationDataAll <- readBismarkPool(c(file_wt, file_met13))
```

3.2 Low resolution profiles

The *DMRcaller* package also offers the possibility to visualise context specific global changes in the methylation profile. To achieve this, the user can call `plotMethylationProfileFromData` function, which computes the mean methylation proportion in tiling bins of fixed size; see Figure 1.

```
par(mar=c(4, 4, 3, 1)+0.1)
plotMethylationProfileFromData(methylationDataList[["WT"]],
                               methylationDataList[["met1-3"]],
                               conditionsNames = c("WT","met1-3"),
                               windowSize = 10000,
                               autoscale = FALSE,
                               context = c("CG"))

## Recompute regions...
## Computing low resolution profiles...
## Calculating methylation profile for  Chr3:101..999999  using a window of  10000  bp
## Calculating methylation profile for  Chr3:101..999999  using a window of  10000  bp
```

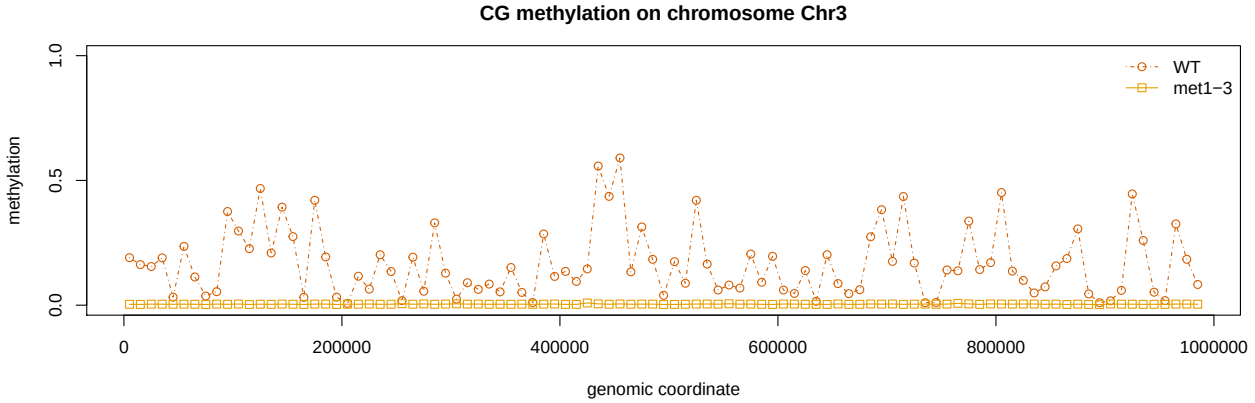

Figure 1: *Low resolution profile in CG context for WT and met1-3.*

Alternatively, for a finer control, the user can use `computeMethylationProfile` function to compute the methylation profile at certain locations on the genome. This function returns a `GRanges` object with four metadata columns

- **sumReadsM** - the number of methylated reads
- **sumReadsN** - the total number of reads
- **Proportion** - the proportion of methylated reads
- **context** - the context

One or more of these `GRanges` objects can be put in a `GRangesList` object which is then passed as a parameter to the `plotMethylationProfile` function.

```
regions <- GRanges(seqnames = Rle("Chr3"), ranges = IRanges(1,1E6))

# compute low resolution profile in 10 Kb windows
profileCGWT <- computeMethylationProfile(methylationDataList[["WT"]],
                                         regions,
                                         windowSize = 10000,
                                         context = "CG")

profileCGMet13 <- computeMethylationProfile(methylationDataList[["met1-3"]],
```

```
                                          regions,
                                          windowSize = 10000,
                                          context = "CG")

profilesCG <- GRangesList("WT" = profileCGWT, "met1-3" = profileCGMet13)

#plot the low resolution profile
par(mar=c(4, 4, 3, 1)+0.1)
par(mfrow=c(1,1))
plotMethylationProfile(profilesCG,
                       autoscale = FALSE,
                       labels = NULL,
                       title="CG methylation on Chromosome 3",
                       col=c("#D55E00","#E69F00"),
                       pch = c(1,0),
                       lty = c(4,1))
```

3.3 Coverage of the bisulfite sequencing data

The number of reads from the bisulfite sequencing can differ significantly between different locations on the genome in the sense that cytosines in the same context (including neighbouring cytosines) can display large variability in the coverage. To plot the coverage of the bisulfite sequencing datasets, one can use `plotMethylationDataCoverage` function which takes as input one or two datasets and the vector with the thresholds used to compute the proportion of cytosines with at least that many reads; see Figure 2.

Alternatively, the *DMRcaller* also provides the `computeMethylationDataCoverage` function which returns a numeric vector with the number or proportion of cytosines in a specific context that have at least a certain number of reads specified by the input vector `breaks`.

```
# compute the coverage in the two contexts
coverageCGWT <- computeMethylationDataCoverage(methylationDataList[["WT"]],
                                               context="CG",
                                               breaks = c(1,5,10,15))
```

3.4 Spatial correlation of methylation levels

Methylation levels are often spatially correlated and some methods to detect DMRs assume this spatial correlation. Nevertheless, different tissues, samples and even methylation context will display different levels of correlation. *DMRcaller* implements `plotMethylationDataSpatialCorrelation` function that plots the correlation of methylation levels as function of distance between cytosines. This function takes as input one or two datasets and the vector with the distances between cytosines; see Figure 2.

Alternatively, the *DMRcaller* also provides the `computeMethylationDataSpatialCorrelation` function which returns a numeric vector with the correlation of methylation levels of cytosines separated by certain distances in a specific context.

```
# compute the coverage in the two contexts
correlation_CG_wt <- computeMethylationDataSpatialCorrelation(methylationDataList[["WT"]],
                                                              context="CG",
                                                              distances=c(1,10,100,1000,10000))
```

3.5 Calling DMRs

DMRcaller package provides `computeDMRs` function to call DMRs. The output of this function is a `GRanges` with 11 metadata columns.

```
# plot the coverage in the two contexts
par(mar=c(4, 4, 3, 1)+0.1)
plotMethylationDataCoverage(methylationDataList[["WT"]],
                            methylationDataList[["met1-3"]],
                            breaks = c(1,5,10,15),
                            regions = NULL,
                            conditionsNames=c("WT","met1-3"),
                            context = c("CHH"),
                            proportion = TRUE,
                            labels=LETTERS,
                            contextPerRow = FALSE)
```

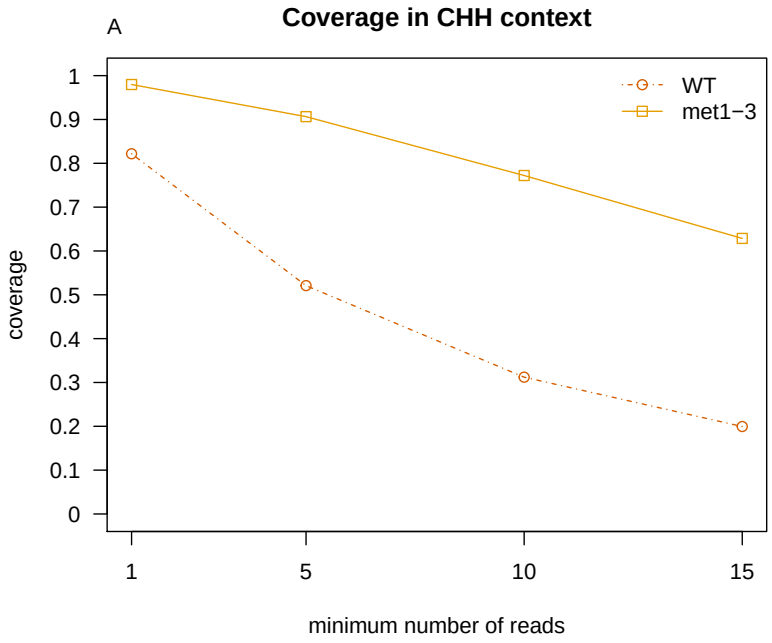

Figure 2: *Coverage*. For example, this figure shows that in WT only 30% of the cytosines in CHH context have at least 10 reads.

```
# compute the spatial correlation of methylation levels
plotMethylationDataSpatialCorrelation(methylationDataList[["WT"]],
                            distances = c(1,100,10000), regions = NULL,
                            conditionsNames = c("WT"),
                            context = c("CG"),
                            labels = LETTERS, col = NULL,
                            pch = c(1,0,16,2,15,17), lty = c(4,1,3,2,6,5),
                            contextPerRow = FALSE,
                            log = "x")

## Computing methylation levels correlation for distances of  1  bp
## Computing methylation levels correlation for distances of  100  bp
## Computing methylation levels correlation for distances of  10000  bp
```

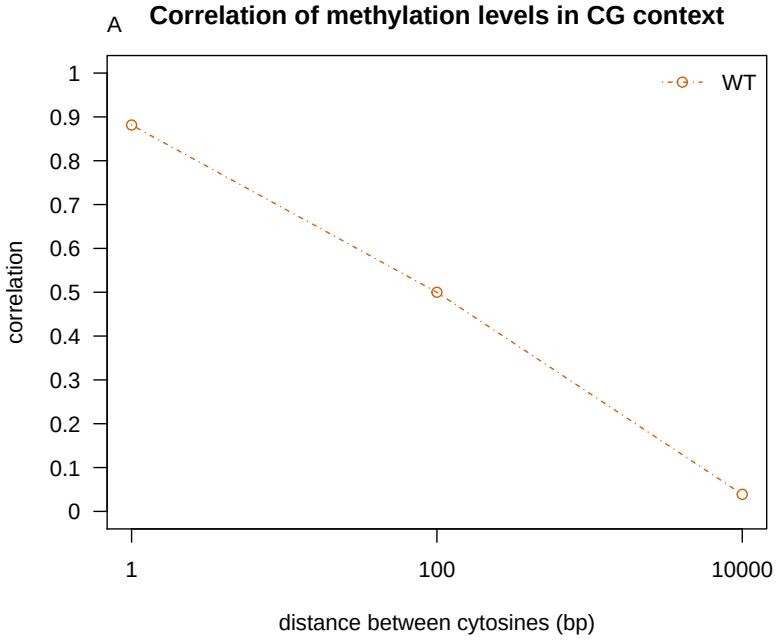

Figure 3: *Spatial correlation of methylation levels.*

- **direction** - a numeric value indicating whether the methylation was lost in the second condition compared to the first one (-1) or gained ($+1$)
- **context** - the context of the cytosine (CG, CHG or CHH)
- **sumReadsM1** - the number of methylated reads in the DMR in condition 1
- **sumReadsN1** - the total number of reads in the DMR in condition 1
- **proportion1** - the proportion of methylated reads in the DMR in condition 1
- **sumReadsM2** - the number of methylated reads in the DMR in condition 2
- **sumReadsN2** - the total number of reads in the DMR in condition 2
- **proportion2** - the proportion of methylated reads in the DMR in condition 2
- **cytosinesCount** - the number of cytosines in the DMR
- **pValue** - the adjusted p-value of the statistical test
- **regionType** - a character string indicating whether the methylation was lost in the second condition compared to the first one (`"loss"`) or gained (`"gain"`)

For predifined regions (e.g. genes, transposons or CpG islands) the user can call `filterDMRs` function to extract the list of regions that are differentially methylated. The output of this function is again a GRanges with the same 11 metadata columns.

Below we present examples of calling both functions.

```
chr_local <- GRanges(seqnames = Rle("Chr3"), ranges = IRanges(5E5,6E5))

# compute the DMRs in CG context with noise_filter method
DMRsNoiseFilterCG <- computeDMRs(methylationDataList[["WT"]],
                      methylationDataList[["met1-3"]],
                      regions = chr_local,
                      context = "CG",
                      method = "noise_filter",
                      windowSize = 100,
                      kernelFunction = "triangular",
                      test = "score",
                      pValueThreshold = 0.01,
                      minCytosinesCount = 4,
                      minProportionDifference = 0.4,
                      minGap = 0,
                      minSize = 50,
                      minReadsPerCytosine = 4,
                      cores = 1)

## Parameters checking ...
## Extract methylation in the corresponding context
## Computing DMRs at  Chr3:500000..600000
## Calculating interpolations...
## Identifying DMRs...
## Analysed reads inside DMRs
## Merge DMRs iteratively
## Filter DMRs

print(DMRsNoiseFilterCG)
```

```
## GRanges object with 60 ranges and 11 metadata columns:
##        seqnames        ranges strand | direction     context sumReadsM1
##           <Rle>     <IRanges>  <Rle> | <numeric> <character>  <numeric>
##    [1]     Chr3 503043-503148      * |        -1          CG        299
##    [2]     Chr3 503390-503542      * |        -1          CG        158
##    [3]     Chr3 503612-503901      * |        -1          CG        342
##    [4]     Chr3 504042-504093      * |        -1          CG         59
##    [5]     Chr3 504255-504348      * |        -1          CG        265
##    ...      ...           ...    ... .       ...         ...        ...
##   [56]     Chr3 593906-594076      * |        -1          CG        216
##   [57]     Chr3 594128-594214      * |        -1          CG         27
##   [58]     Chr3 594285-594385      * |        -1          CG        128
##   [59]     Chr3 599027-599107      * |        -1          CG         57
##   [60]     Chr3 599509-599634      * |        -1          CG        168
##        sumReadsN1 proportion1 sumReadsM2 sumReadsN2 proportion2 cytosinesCount
##         <numeric>   <numeric>  <numeric>  <numeric>   <numeric>      <numeric>
##    [1]        365    0.819178          0        419  0.00000000             10
##    [2]        198    0.797980          0        414  0.00000000             12
##    [3]        442    0.773756          3        807  0.00371747             25
##    [4]         86    0.686047          0        249  0.00000000              6
##    [5]        351    0.754986          0        412  0.00000000             12
##    ...        ...         ...        ...        ...         ...            ...
##   [56]        253    0.853755          1        648  0.00154321             16
##   [57]         45    0.600000          2        107  0.01869159              4
##   [58]        149    0.859060          0        258  0.00000000              4
##   [59]        111    0.513514          3        154  0.01948052              4
##   [60]        219    0.767123          0        201  0.00000000              8
##              pValue  regionType
##           <numeric> <character>
##    [1] 4.18251e-122        loss
##    [2]  1.86673e-98        loss
##    [3] 4.84005e-185        loss
##    [4]  7.28326e-47        loss
##    [5] 3.75353e-105        loss
##    ...          ...         ...
##   [56] 2.23081e-158        loss
##   [57]  8.30991e-17        loss
##   [58]  5.30109e-72        loss
##   [59]  2.60868e-21        loss
##   [60]  1.19821e-57        loss
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

```
# compute the DMRs in CG context with neighbourhood method
DMRsNeighbourhoodCG <- computeDMRs(methylationDataList[["WT"]],
                                   methylationDataList[["met1-3"]],
                                   regions = chr_local,
                                   context = "CG",
                                   method = "neighbourhood",
                                   test = "score",
                                   pValueThreshold = 0.01,
                                   minCytosinesCount = 4,
                                   minProportionDifference = 0.4,
                                   minGap = 200,
                                   minSize = 1,
```

```
                                    minReadsPerCytosine = 4,
                                    cores = 1)

## Parameters checking ...
## Extract methylation in the corresponding context
## Computing DMRs
## Merge DMRs iteratively
## Filter DMRs

print(DMRsNeighbourhoodCG)

## GRanges object with 34 ranges and 16 metadata columns:
##        seqnames        ranges strand |  context trinucleotide_context   readsM1
##           <Rle>     <IRanges>  <Rle> | <factor>              <factor> <integer>
##    [1]     Chr3 503058-503853      * |       CG                   CGG        96
##    [2]     Chr3 504058-504069      * |       CG                   CGG        22
##    [3]     Chr3 504292-504490      * |       CG                   CGA        35
##    [4]     Chr3 506440-506776      * |       CG                   CGT        28
##    [5]     Chr3 507119-507480      * |       CG                   CGA         6
##    ...      ...           ...    ... .      ...                   ...       ...
##   [30]     Chr3 588591-588633      * |       CG                   CGC        11
##   [31]     Chr3 591681-591790      * |       CG                   CGT        25
##   [32]     Chr3 593736-594337      * |       CG                   CGA        65
##   [33]     Chr3 598934-599219      * |       CG                   CGT         6
##   [34]     Chr3 599556-599586      * |       CG                   CGA        46
##          readsN1   readsM2   readsN2       pValue sumReadsM1 sumReadsN1
##        <integer> <integer> <integer>    <numeric>  <numeric>  <numeric>
##    [1]       139         0       117  0.00000e+00        806       1217
##    [2]        25         0        48  1.55222e-42         56         78
##    [3]        39         0        42 6.07831e-180        389        584
##    [4]        42         0        59 2.59969e-108        228        370
##    [5]         9         0        21 1.10736e-102        225        415
##    ...       ...       ...       ...          ...        ...        ...
##   [30]        12         0        38 1.39683e-145        353        426
##   [31]        31         2        59 1.27477e-143        268        296
##   [32]        75         1        56  0.00000e+00        659       1068
##   [33]         6         0        15  2.41968e-39         83        178
##   [34]        55         0        63  4.16663e-57        166        217
##        proportion1 sumReadsM2 sumReadsN2 proportion2 cytosinesCount direction
##          <numeric>  <numeric>  <numeric>   <numeric>      <numeric> <numeric>
##    [1]    0.662284          4       2205  0.00181406             65        -1
##    [2]    0.717949          0        210  0.00000000              5        -1
##    [3]    0.666096          0        911  0.00000000             24        -1
##    [4]    0.616216          3        629  0.00476948             20        -1
##    [5]    0.542169          1        687  0.00145560             22        -1
##    ...         ...        ...        ...         ...            ...       ...
##   [30]    0.828638          4        509  0.00785855              8        -1
##   [31]    0.905405          4        486  0.00823045             14        -1
##   [32]    0.617041          6       1817  0.00330215             41        -1
##   [33]    0.466292          3        331  0.00906344             13        -1
##   [34]    0.764977          0        200  0.00000000              7        -1
##         regionType
##        <character>
##    [1]        loss
##    [2]        loss
##    [3]        loss
##    [4]        loss
```

```
##    [5]        loss
##    ...         ...
##   [30]        loss
##   [31]        loss
##   [32]        loss
##   [33]        loss
##   [34]        loss
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

```
# compute the DMRs in CG context with bins method
DMRsBinsCG <- computeDMRs(methylationDataList[["WT"]],
                          methylationDataList[["met1-3"]],
                          regions = chr_local,
                          context = "CG",
                          method = "bins",
                          binSize = 100,
                          test = "score",
                          pValueThreshold = 0.01,
                          minCytosinesCount = 4,
                          minProportionDifference = 0.4,
                          minGap = 200,
                          minSize = 50,
                          minReadsPerCytosine = 4,
                          cores = 1)

## Parameters checking ...
## Extract methylation in the corresponding context
## Computing DMRs at  Chr3:500000..600000
## Count inside each bin...
## Filter the bins...
## Identifying DMRs...
## Merge adjacent DMRs
## Merge DMRs iteratively
## Filter DMRs

print(DMRsBinsCG)

## GRanges object with 40 ranges and 11 metadata columns:
##        seqnames        ranges strand | sumReadsM1 sumReadsN1 proportion1
##           <Rle>     <IRanges>  <Rle> |  <numeric>  <numeric>   <numeric>
##    [1]     Chr3 503000-503199      * |        299        731    0.409029
##    [2]     Chr3 503100-503199      * |         13         28    0.464286
##    [3]     Chr3 503400-503499      * |        158        198    0.797980
##    [4]     Chr3 503400-504499      * |        959       1674    0.572879
##    [5]     Chr3 506400-506699      * |        182        321    0.566978
##    ...      ...           ...    ... .        ...        ...         ...
##   [36]     Chr3 593700-594399      * |        660       1151    0.573414
##   [37]     Chr3 599000-599299      * |         77        184    0.418478
##   [38]     Chr3 599200-599299      * |         20         35    0.571429
##   [39]     Chr3 599500-599599      * |        168        219    0.767123
##   [40]     Chr3 599500-599599      * |        168        219    0.767123
##        sumReadsM2 sumReadsN2 proportion2 cytosinesCount     context direction
##         <numeric>  <numeric>   <numeric>      <numeric> <character> <numeric>
##    [1]          1        776 0.001288660             17          CG        -1
##    [2]          0        100 0.000000000              4          CG        -1
```

```
##    [3]          0        414 0.000000000             12          CG        -1
##    [4]          3       3183 0.000942507             90          CG        -1
##    [5]          3        546 0.005494505             18          CG        -1
##    ...        ...        ...         ...            ...         ...       ...
##   [36]          7       1907  0.00367069             44          CG        -1
##   [37]          3        356  0.00842697             14          CG        -1
##   [38]          0        107  0.00000000              6          CG        -1
##   [39]          0        201  0.00000000              8          CG        -1
##   [40]          0        201  0.00000000              8          CG        -1
##              pValue  regionType
##           <numeric> <character>
##    [1]  4.74723e-87        loss
##    [2]  6.54241e-13        loss
##    [3]  1.65932e-98        loss
##    [4]  0.00000e+00        loss
##    [5]  2.63625e-84        loss
##    ...          ...         ...
##   [36] 2.64404e-298        loss
##   [37]  5.89785e-37        loss
##   [38]  3.36772e-17        loss
##   [39]  1.19821e-57        loss
##   [40]  1.19821e-57        loss
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

```
# load the gene annotation data
data(GEs)

#select the genes
genes <- GEs[which(GEs$type == "gene")]

# compute the DMRs in CG context over genes
DMRsGenesCG <- filterDMRs(methylationDataList[["WT"]],
                          methylationDataList[["met1-3"]],
                          potentialDMRs = genes[overlapsAny(genes, chr_local)],
                          context = "CG",
                          test = "score",
                          pValueThreshold = 0.01,
                          minCytosinesCount = 4,
                          minProportionDifference = 0.4,
                          minReadsPerCytosine = 3,
                          cores = 1)

## Parameters checking ...
## Extract methylation in the corresponding context
## Computing DMRs at  Chr3:101..999999
## Selecting data...
## Identifying DMRs...

print(DMRsGenesCG)

## GRanges object with 3 ranges and 21 metadata columns:
##       seqnames        ranges strand |   source     type     score     phase
##          <Rle>     <IRanges>  <Rle> | <factor> <factor> <numeric> <integer>
##   [1]     Chr3 576378-579559      + |   TAIR10     gene        NA      <NA>
##   [2]     Chr3 528574-532582      - |   TAIR10     gene        NA      <NA>
```

```
##   [3]     Chr3 570134-572345      - |   TAIR10     gene        NA      <NA>
##                ID        Name                Note          Parent       Index
##       <character> <character>     <CharacterList> <CharacterList> <character>
##   [1]   AT3G02680   AT3G02680 protein_coding_gene                         <NA>
##   [2]   AT3G02530   AT3G02530 protein_coding_gene                         <NA>
##   [3]   AT3G02660   AT3G02660 protein_coding_gene                         <NA>
##       Derives_from           Alias sumReadsM1 sumReadsN1 proportion1 sumReadsM2
##        <character> <CharacterList>  <numeric>  <numeric>   <numeric>  <numeric>
##   [1]         <NA>                       1106       2379    0.464901         14
##   [2]         <NA>                       1150       2756    0.417271         30
##   [3]         <NA>                        874       1848    0.472944         10
##       sumReadsN2 proportion2 cytosinesCount    pValue  regionType direction
##        <numeric>   <numeric>      <numeric> <numeric> <character> <numeric>
##   [1]       4546  0.00307963            142         0        loss        -1
##   [2]       6207  0.00483325            171         0        loss        -1
##   [3]       3487  0.00286779            104         0        loss        -1
##   -------
##   seqinfo: 7 sequences from an unspecified genome; no seqlengths
```

3.6 Merge DMRs

Finally, for merging adjacent DMRs, *DMRcaller* provides the function `mergeDMRsIteratively` which can be used as follows:

```
DMRsNoiseFilterCGMerged <- mergeDMRsIteratively(DMRsNoiseFilterCG,
                                                minGap = 200,
                                                respectSigns = TRUE,
                                                methylationDataList[["WT"]],
                                                methylationDataList[["met1-3"]],
                                                context = "CG",
                                                minProportionDifference = 0.4,
                                                minReadsPerCytosine = 4,
                                                pValueThreshold = 0.01,
                                                test="score")

## Parameters checking ...
## Merge DMRs iteratively ...

print(DMRsNoiseFilterCGMerged)

## GRanges object with 37 ranges and 11 metadata columns:
##        seqnames        ranges strand | direction     context sumReadsM1
##           <Rle>     <IRanges>  <Rle> | <numeric> <character>  <numeric>
##    [1]     Chr3 503043-503148      * |        -1          CG        299
##    [2]     Chr3 503390-504509      * |        -1          CG        959
##    [3]     Chr3 506392-506723      * |        -1          CG        182
##    [4]     Chr3 507286-507422      * |        -1          CG        153
##    [5]     Chr3 514791-514891      * |        -1          CG        560
##    ...      ...            ...    ... .       ...         ...        ...
##   [33]     Chr3 588556-588681      * |        -1          CG        355
##   [34]     Chr3 591657-591828      * |        -1          CG        268
##   [35]     Chr3 593709-594385      * |        -1          CG        659
##   [36]     Chr3 599027-599107      * |        -1          CG         57
##   [37]     Chr3 599509-599634      * |        -1          CG        168
##        sumReadsN1 proportion1 sumReadsM2 sumReadsN2 proportion2 cytosinesCount
##         <numeric>   <numeric>  <numeric>  <numeric>   <numeric>      <numeric>
```

```
##    [1]        365    0.819178          0        419 0.000000000             10
##    [2]       1674    0.572879          3       3183 0.000942507             90
##    [3]        321    0.566978          3        546 0.005494505             18
##    [4]        217    0.705069          0        322 0.000000000              8
##    [5]        760    0.736842          1        680 0.001470588             10
##    ...        ...         ...        ...        ...         ...            ...
##   [33]        458    0.775109          5        605  0.00826446             10
##   [34]        321    0.834891          4        540  0.00740741             16
##   [35]       1068    0.617041          6       1827  0.00328407             42
##   [36]        111    0.513514          3        154  0.01948052              4
##   [37]        219    0.767123          0        201  0.00000000              8
##               pValue  regionType
##            <numeric> <character>
##    [1] 4.18251e-122        loss
##    [2]  0.00000e+00        loss
##    [3]  1.44994e-84        loss
##    [4]  1.20961e-70        loss
##    [5] 2.03906e-178        loss
##    ...          ...         ...
##   [33] 4.02207e-150        loss
##   [34] 4.83924e-140        loss
##   [35]  0.00000e+00        loss
##   [36]  2.60868e-21        loss
##   [37]  1.19821e-57        loss
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

Note that two neighbouring DMRs will be merged if all the conditions below are met

- they are within a distance from each other smaller than minGap
- the difference in methylation levels between the two conditions is statistically significant according to the statistical test when the two DMRs are joined
- the difference in methylation proportion between the two conditions is higher than a threshold value when the two DMRs are joined
- the number of reads per cytosine is higher than a threshold when the two DMRs are joined

3.7 Extract methylation data in regions

`analyseReadsInsideRegionsForCondition` function can extract additional information in a set of genomic regions (including DMRs) from any `methylationData` object. For example, to establish a link between the CG and CHH methylation, one might want to extract the number of methylated reads and the total number of reads in CHH context inside DMRs called in CG context.

```
#retrive the number of reads in CHH context in WT in CG DMRs
DMRsNoiseFilterCGreadsCHH <- analyseReadsInsideRegionsForCondition(
                             DMRsNoiseFilterCGMerged,
                             methylationDataList[["WT"]], context = "CHH",
                             label = "WT")

## Parameters checking ...
## Extract methylation levels in corresponding context ...
## Compute reads inside each region ...

print(DMRsNoiseFilterCGreadsCHH)
```

```
## GRanges object with 37 ranges and 15 metadata columns:
##        seqnames        ranges strand | direction     context sumReadsM1
##           <Rle>     <IRanges>  <Rle> | <numeric> <character>  <numeric>
##    [1]     Chr3 503043-503148      * |        -1          CG        299
##    [2]     Chr3 503390-504509      * |        -1          CG        959
##    [3]     Chr3 506392-506723      * |        -1          CG        182
##    [4]     Chr3 507286-507422      * |        -1          CG        153
##    [5]     Chr3 514791-514891      * |        -1          CG        560
##    ...      ...           ...    ... .       ...         ...        ...
##   [33]     Chr3 588556-588681      * |        -1          CG        355
##   [34]     Chr3 591657-591828      * |        -1          CG        268
##   [35]     Chr3 593709-594385      * |        -1          CG        659
##   [36]     Chr3 599027-599107      * |        -1          CG         57
##   [37]     Chr3 599509-599634      * |        -1          CG        168
##        sumReadsN1 proportion1 sumReadsM2 sumReadsN2 proportion2 cytosinesCount
##         <numeric>   <numeric>  <numeric>  <numeric>   <numeric>      <numeric>
##    [1]        365    0.819178          0        419 0.000000000             10
##    [2]       1674    0.572879          3       3183 0.000942507             90
##    [3]        321    0.566978          3        546 0.005494505             18
##    [4]        217    0.705069          0        322 0.000000000              8
##    [5]        760    0.736842          1        680 0.001470588             10
##    ...        ...         ...        ...        ...         ...            ...
##   [33]        458    0.775109          5        605  0.00826446             10
##   [34]        321    0.834891          4        540  0.00740741             16
##   [35]       1068    0.617041          6       1827  0.00328407             42
##   [36]        111    0.513514          3        154  0.01948052              4
##   [37]        219    0.767123          0        201  0.00000000              8
##              pValue  regionType sumReadsMWTCHH sumReadsNWTCHH proportionWTCHH
##           <numeric> <character>      <numeric>      <numeric>       <numeric>
##    [1] 4.18251e-122        loss              0            303      0.00000000
##    [2]  0.00000e+00        loss             99           3323      0.02979236
##    [3]  1.44994e-84        loss             10           1047      0.00955110
##    [4]  1.20961e-70        loss              0            571      0.00000000
##    [5] 2.03906e-178        loss              1            665      0.00150376
##    ...          ...         ...            ...            ...             ...
##   [33] 4.02207e-150        loss             12            672      0.01785714
##   [34] 4.83924e-140        loss              6            792      0.00757576
##   [35]  0.00000e+00        loss             29           2560      0.01132812
##   [36]  2.60868e-21        loss              0            193      0.00000000
##   [37]  1.19821e-57        loss              1            206      0.00485437
##        cytosinesCountCHH
##                <numeric>
##    [1]                27
##    [2]               309
##    [3]                90
##    [4]                33
##    [5]                32
##    ...               ...
##   [33]                32
##   [34]                52
##   [35]               196
##   [36]                23
##   [37]                36
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

3.8 Plotting the distribution of DMRs

Sometimes, it is useful to obtain the distribution of the DMRs over the chromosomes. The *DMRcaller* provides the `computeOverlapProfile` function, which computes this distribution. The `GRanges` object generated by this function can then be added to a `GRangesList` object, which can be plotted using `plotOverlapProfile` function; see Figure 4. Additionally, the `plotOverlapProfile` function allows the user to specify two `GRangesList`, thus, allowing the plotting of distributions of hypo or hyper methylated DMRs separately.

```
# compute the distribution of DMRs
hotspots <- computeOverlapProfile(DMRsNoiseFilterCGMerged, chr_local,
                                  windowSize=5000, binary=TRUE)

## Calculating overlaps for  Chr3:500000..600000  using a window of  5000  bp

# plot the distribution of DMRs
plotOverlapProfile(GRangesList("Chr3"=hotspots))
```

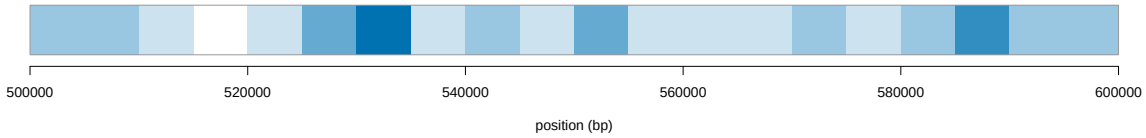

Figure 4: *Distribution of DMRs.* Darker colour indicates higher density, while lighter colour lower density.

3.9 Plotting profiles with DMRs

Finally, *DMRcaller* package also provides a function to plot methylation profiles at a specific location on the genome. To plot the methylation profile the user needs to call the `plotLocalMethylationProfile` function; see Figure 5.

4 Parallel computation

Computing the DMRs can be computationally intensive. For example, in the case of *A. thaliana* (with a genome of ≈ 130 *Mb*), it can take several hours to compute the DMRs depending on the method used and on the number of DMRs. To speed up computations, *DMRcaller* supports parallel computing of DMRs using the package *parallel*, but parallel computation is currently not supported on Windows.

The five functions used for computing and filtering the DMRs (`computeDMRs`, `filterDMRs`, `mergeDMRsIteratively` and `analyseReadsInsideRegionsForCondition`) accept the parameter `cores`, which specifies the number of cores that can be used when performing the corresponding computations. When using 10 cores, it can take between 10 and 30 minutes to compute the DMRs in *A. thaliana* depending on the selected parameters.

5 Analysis of biological replicates

The package also contains a set of functions for the analysis of multiple biological replicates.

The synthetic dataset is made by 300 different cytosines, extracted from those present in the *A. thaliana* dataset. The value for **readsN** are created using the function `rnorm`, while the values for **readsM** are generated using the function `rbinom`. The probabilities used are 0.1 in the external region and 0.8 in the central region. In this way a DMR should be detected in the central region of the synthetic dataset.

The difference in proportion is plotted in figure 6

The DMRs are computed using the function `computeDMRsReplicates`, which uses beta regression (Ferrari and Cribari-Neto, 2004) to detect differential methylation.

```
# select a 20 Kb location on the Chr3
chr3Reg <- GRanges(seqnames = Rle("Chr3"), ranges = IRanges(510000,530000))

# create a list with all DMRs
DMRsCGList <- list("noise filter" = DMRsNoiseFilterCGMerged,
                   "neighbourhood" = DMRsNeighbourhoodCG,
                   "bins" = DMRsBinsCG,
                   "genes" = DMRsGenesCG)
# plot the local profile
par(cex=0.9)
par(mar=c(4, 4, 3, 1)+0.1)
plotLocalMethylationProfile(methylationDataList[["WT"]],
                            methylationDataList[["met1-3"]],
                            chr3Reg,
                            DMRsCGList,
                            conditionsNames = c("WT", "met1-3"),
                            GEs,
                            windowSize = 300,
                            main="CG methylation")
```

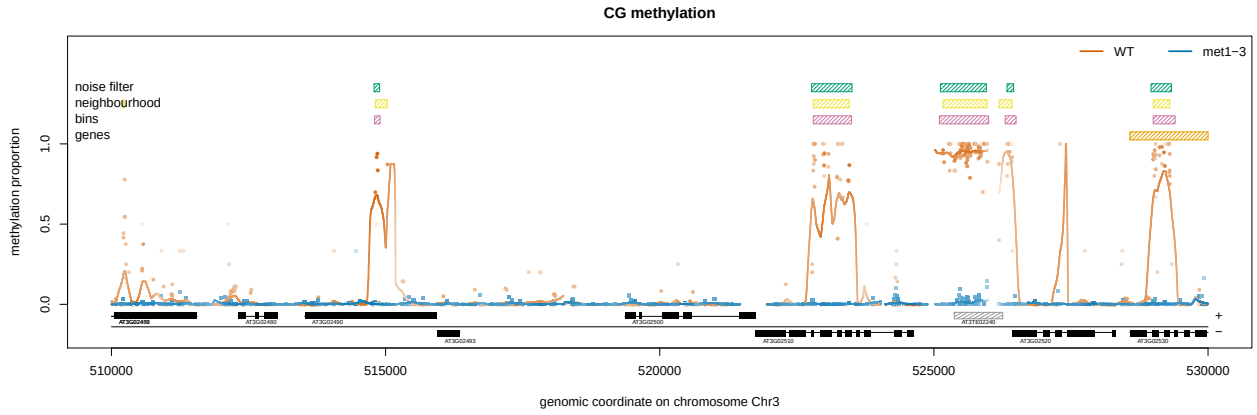

Figure 5: *Local methylation profile.* The points on the graph represent methylation proportion of individual cytosines, their colour (red or blue) which sample they belong to and the intensity of the the colour how many reads that particular cytosine had. This means that darker colours indicate stronger evidence that the corresponding cytosine has the corresponding methylation proportion, while lighter colours indicate a weaker evidence. The solid lines represent the smoothed profiles and the intensity of the colour the coverage at the corresponding position (darker colours indicate more reads while lighter ones less reads). The boxes on top represent the DMRs, where a filled box will represent a DMR which gained methylation while a box with a pattern represent a DMR that lost methylation. The DMRs need to have a metadata column `regionType` which can be either `"gain"` (where there is more methylation in condition 2 compared to condition 1) or `"loss"` (where there is less methylation in condition 2 compared to condition 1). In case this metadata column is missing all DMRs are drawn using filled boxes. Finally, we also allow annotation of the DNA sequence. We represent by black boxes all the exons, which are joined by a horizontal black line, thus, marking the full body of the gene. With grey boxes we mark the transposable elements. Both for genes and transposable elements we plot them over a mid line if they are on the positive strand and under the mid line if they are on the negative strand.

```
# loading synthetic data
data("syntheticDataReplicates")

# create vector with colours for plotting
cbbPalette <- c("#000000", "#E69F00", "#56B4E9", "#009E73", "#F0E442",
                "#0072B2", "#D55E00", "#CC79A7")

# plotting the difference in proportions
plot(start(methylationData), methylationData$readsM1/methylationData$readsN1,
     ylim=c(0,1), col=cbbPalette[2], xlab="Position in Chr3 (bp)",
     ylab="Methylation proportion")
points(start(methylationData), methylationData$readsM2/methylationData$readsN2,
       col=cbbPalette[7], pch=4)
points(start(methylationData), methylationData$readsM3/methylationData$readsN3,
       col=cbbPalette[3], pch=2)
points(start(methylationData), methylationData$readsM4/methylationData$readsN4,
       col=cbbPalette[6], pch=3)
legend(x = "topleft", legend=c("Treatment 1", "Treatment 2", "Control 1",
                               "Control 2"), pch=c(1,4,2,3),
       col=cbbPalette[c(2,7,3,6)], bty="n", cex=1.0)
```

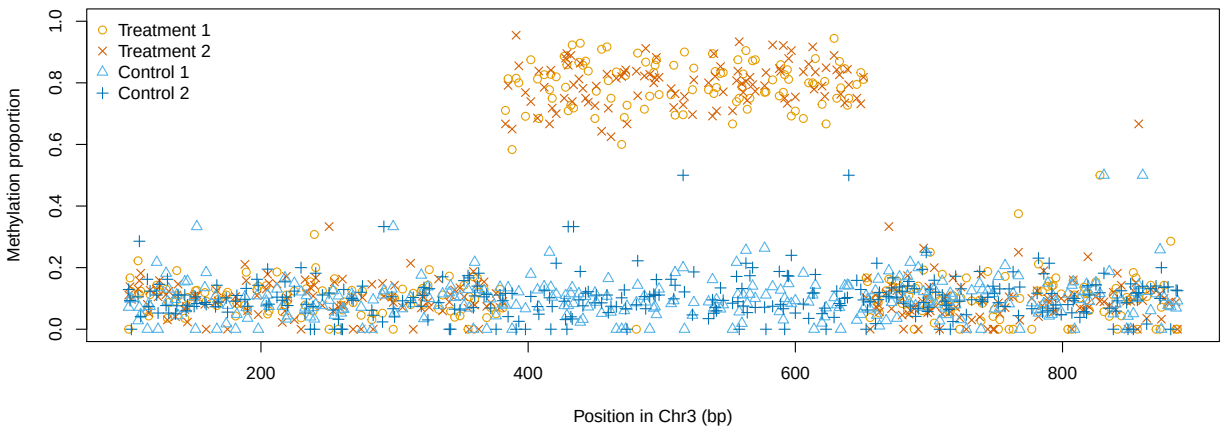

Figure 6: *Methylation proportions in the synthetic dataset.*

```r
# loading betareg library to allow using computeDMRsReplicates function
library(betareg)

# creating condition vector
condition <- c("a", "a", "b", "b")

# computing DMRs using the neighbourhood method
DMRsReplicatesBins <- computeDMRsReplicates(methylationData = methylationData,
                                            condition = condition,
                                            regions = NULL,
                                            context = "CG",
                                            method = "bins",
                                            binSize = 100,
                                            test = "betareg",
                                            pseudocountM = 1,
                                            pseudocountN = 2,
                                            pValueThreshold = 0.01,
                                            minCytosinesCount = 4,
                                            minProportionDifference = 0.4,
                                            minGap = 0,
                                            minSize = 50,
                                            minReadsPerCytosine = 4,
                                            cores = 1)

## Parameters checking ...
## Extract methylation in the corresponding context
## Computing DMRs at  Chr3:101..886
## Count inside each bin...
## Filter the bins...
## Identifying DMRs...
## Merge adjacent DMRs
## Merge DMRs iteratively
## Filter DMRs

print(DMRsReplicatesBins)

## GRanges object with 2 ranges and 11 metadata columns:
##       seqnames    ranges strand | sumReadsM1 sumReadsN1 proportion1 sumReadsM2
##          <Rle> <IRanges>  <Rle> |  <numeric>  <numeric>   <numeric>  <numeric>
##   [1]     Chr3   401-500      * |        436        546    0.797445         61
##   [2]     Chr3   501-600      * |        419        521    0.803059         42
##       sumReadsN2 proportion2 cytosinesCount     context direction    pValue
##        <numeric>   <numeric>      <numeric> <character> <numeric> <numeric>
##   [1]        596    0.103679              6          CG        -1         0
##   [2]        411    0.104116              4          CG        -1         0
##        regionType
##       <character>
##   [1]        loss
##   [2]        loss
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

6 Session information

```
sessionInfo()

## R version 4.0.3 (2020-10-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.5 LTS
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.12-bioc/R/lib/libRblas.so
## LAPACK: /home/biocbuild/bbs-3.12-bioc/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
## [1] betareg_3.1-3        DMRcaller_1.22.0     GenomicRanges_1.42.0
## [4] GenomeInfoDb_1.26.0  IRanges_2.24.0       S4Vectors_0.28.0
## [7] BiocGenerics_0.36.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.5             flexmix_2.3-17         Formula_1.2-4
##  [4] knitr_1.30             XVector_0.30.0         magrittr_1.5
##  [7] zlibbioc_1.36.0        RcppRoll_0.3.0         lattice_0.20-41
## [10] stringr_1.4.0          highr_0.8              tools_4.0.3
## [13] nnet_7.3-14            grid_4.0.3             xfun_0.18
## [16] modeltools_0.2-23      lmtest_0.9-38          GenomeInfoDbData_1.2.4
## [19] bitops_1.0-6           RCurl_1.98-1.2         evaluate_0.14
## [22] sandwich_3.0-0         stringi_1.5.3          compiler_4.0.3
## [25] zoo_1.8-8
```

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Ferrari, S. and Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7):799–815.

Hebestreit, K., Dugas, M., and Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653.

Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572.

Stroud, H., Greenberg, M. V., Feng, S., Bernatavichute, Y., and Jacobsen, S. E. (2013). Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. *Cell*, 152(1-2):352–364.