Quantification of DNA Methylation and Epimutations from Bisulfite Sequencing Data – the BEAT package

Kemal Akman[1], Achim Tresch

Max-Planck-Institute for Plant Breeding Research
Cologne, Germany
[1]akman@mpipz.mpg.de

May 3, 2016

Abstract

The following example illustrates a standard use case of the BEAT package, which is used for processing and modeling methylated and unmethylated counts of CG positions from Bisulfite sequencing (BS-Seq) for determination of epimutation rates, i.e. the rate of change in DNA methylation status at CG sites between a reference multi-cell sample and single-cell samples.

The input for the package BEAT consists of count data in the form of counts for unmethylated and methylated cytosines per genomic position, which are then grouped into genomic regions of sufficient coverage in order to allow for low-coverage samples to be analyzed. Methylation rates of each region are then modeled using a binomial mixture model in order to adjust for experimental bias, which arises mainly out of incomplete bisulfite conversion (resulting in unmethylated CG positions falsely appearing to be methylated) and sequencing errors (resulting in random errors in methylation status, including methylated CG positions falsely appearing as unmethylated).

This vignette explains the use of the package. For a detailed exposition of the statistical method, please see our upcoming paper.

Contents

1 Preamble

1.1 Biological Background

Bisulfite sequencing (BS-Seq) is a sequence-based method to accurately detect DNA methylation at specific loci, which involves treating DNA with sodium bisulfite [**?**]. Bisulfite conversion of unmethylated cytosines into uracil is a relatively simple chemical reaction which has now become a standard in DNA methylation profiling. The key advantage of this method is accuracy, as the degree of methylation at each cytosine can be quantified with great precision [**?**].

Most currently available techniques for determination of DNA methylation can only measure average values obtained from cell populations as a whole, requiring at least 30 ng of DNA, i.e. the equivalent of about 6000 cells [**?**]. Since these population approaches cannot account for cell- and position-specific differences in DNA methylation, which are termed epimutations [**?**], they are unsuitable for the characterization of cellular heterogeneity. However, this heterogeneity plays an important role in differentiation and development, stem cell re-programming, in diseases such as cancer and aging [**?**]. Developing single-cell approaches for measuring DNA methylation is not only be vital to fully understand individual cell-specific changes and complexity of tissue microenvironments, but also for the analysis of clinical samples, such as circulating tumor cells or needle biopsies, when the amount of material is often limited. With the BEAT package, we present a pipeline for the computational analysis part of such single-cell BS-Seq experiments.

1.2 Package Functionality

This vignette will illustrate the complete work flow of a DNA methylation analysis, which includes model-based estimation of regional methylation rates from observed BS-Seq methylation counts, calling of methylation status, and the comparison of samples to determine regional differences in methylation status, i.e., epimutation calling. These three steps are implemented by the three public functions of BEAT, namely *positions_to_regions*, *generate_results* and *epimutation_calls*, respectively. The functionality of BEAT will be demonstrated using sample data consisting of the first chromosome of mouse liver hepatocytes, one reference sample ('reference.positions.csv') and one single-cell sample to compare against ('sample.positions.csv') The main challenges in DNA methylation analysis are bias correction, modeling of sampling variance (shot noise due to low count numbers) and assessing the significance of changes.

2 Input format

The package BEAT expects as input one csv per sample with counts for unmethylated and methylated cytosines per genomic position. Such data can be obtained from the output of BS-Seq mapping tools such as bismark (see http://www.bioinformatics.babraham.ac.uk/projects/bismark/) followed by simple script-based data processing. BS-Seq data for methylated- and unmethylated counts per genomic position needs to be formatted into a data.frame object with the columns: 'chr', 'pos', 'meth', 'unmeth' (signifying chromosome name, chromosomal position, as well as methylated- and unmethylated cytosine counts at that chromosomal position).

```
> library(BEAT)
> localpath <- system.file('extdata', package = 'BEAT')
> positions <- read.csv(file.path(localpath, "sample.positions.csv"))
> head(positions)
```

```
   chr       pos meth unmeth
1 chr1 100001310    1      0
2 chr1 100002648    0      1
3 chr1 100002688    0      1
4 chr1 100002802    1      0
5 chr1 100004564    0      1
6 chr1 100004606    0      1
```

3 Configuration and parameters

BEAT can process multiple samples at a time. For each sample specified under *'sampNames'*, the package BEAT reads this data frame from a csv from the working directory, which is specified by the parameter *'localpath'*. The file should be called <samplename>.positions.csv and should contain the aforementioned csv under the name 'positions'. Result files written by BEAT will have the names <samplename>.results.RData. A distinction is made between two samples whose methylation status is compared against that of the reference. For each sample, the assignment of reference or non-reference status to samples is done via the vector *'is.reference'*, which is indexed by the *'sampNames'* vector and contains one TRUE entry for the reference and one to many FALSE entries for samples to be compared against the reference. Our example data set contains two files, one named 'reference' for a reference consisting of a mixture of cells, and one named 'sample' for a single-cell sample to be compared against the reference.

```
> sampNames <- c('reference','sample')
> sampNames

[1] "reference" "sample"

> is.reference <- c(TRUE, FALSE)
```

For the modeling part of BEAT, Bisulfite conversion rates have to be specified per sample using the vector *'convrates'*, which is indexed by *'sampNames'*. These conversion rates need to be specified manually by the user. Practically, for mammalian somatic cell samples, these rates can be estimated for each sample by looking at the non-CpG methylation rate per sample and using the inverse value as estimated CpG methylation rate, because non-CpG methylation in these types of cells is expected to be near zero. For example, if non-CpG methylation in a sequenced sample was measured to be 0.1 then BS-conversion rate for that sample would be set to 0.9 when near-zero non-CpG methylation can be expected, as is the case in somatic mammalian cells.

```
> pplus <- c(0.2, 0.5)
> convrates <- 1 - pplus
```

The aforementioned values are then set in a parameter object, which is used throughout the further work flow in this package. It provides the following additional options:

- 1-*convrates* represents the fraction of unmethylated counts that are falsely called as methylated due to incomplete BS-conversion. This parameter is also referred to as 'pplus' in the statistical model.
- *'pminus'* represents the fraction of methylated counts that are falsely called as unmethylated.
- *'regionSize'* is the size of regions into which genomic positions are grouped. In single cells, the number of positions with sufficient coverage for reliable statistical predictions may be very small. Therefore, our pipeline applies epimutation calling to regions instead of single positions. Single positions are pooled into regions of appropriate size, i.e., regions containing sufficiently many CpG positions that have a positive read count number in both the reference and the single cell sample (.shared. CpG positions). Our method has then sufficient power to reliably detect epimutation events affecting these regions. For a detailed description of pooling positions into regions, see section **??**.

- After pooling CG positions to regions, there may still be regions with low count numbers that do not allow for reliable downstream analysis. Regions with less than *'minCounts'* counts will be removed from further processing, potentially saving significant processing time in further analysis steps.

The following parameters are of minor importance, for a first analysis, they can be left at their pre-set values.

- *'verbose'* is an option that prints additional information during computation when set to TRUE. *'computeRegions'* is an option that will recompute the regions from given positional input if set to TRUE; otherwise, it will depend on existing region files already present in the *'localpath'* directory (file names ending in regions.RData).
- *'computeMatrices'* is an option that will recompute the model data from given regions if set to TRUE; otherwise, it will depend on existing model output files already present in the *'localpath'* directory (file names ending in convMat.RData and results.RData.
- *'writeEpicallMatrix'* is an option that can be set to TRUE to generate epimutation calling output in the form of matrices (one row per genomic position, output format is CSV).

```
> params <- makeParams(localpath, sampNames, convrates,
+ is.reference, pminus = 0.2, regionSize = 10000, minCounts = 5)
> params

$localpath
[1] "/private/tmp/RtmphnBwEO/Rinst14ea0b93265c/BEAT/extdata"

$sampNames
[1] "reference" "sample"

$pplus
reference    sample
      0.2       0.5

$is.reference
reference    sample
     TRUE     FALSE

$pminus
reference    sample
      0.2       0.2

$regionSize
[1] 10000

$minCounts
[1] 5
```

```
$verbose
[1] TRUE

$computeRegions
[1] TRUE

$computeMatrices
[1] TRUE

$writeEpicallMatrix
[1] TRUE
```

4 Pooling of CG counts into regions

The supplied methylation counts for individual CG positions are grouped into regions by BEAT for modeling according to the specified *'regionSize'* and *'minCount'* parameters. The function positions_to_regions takes as input the samples as csv objects located under <samplename>.positions.csv in the working directory, as discussed above. The function saves a list of data.frames of resulting counts per genomic regions in the current working directory under <samplename>.regions.<regionSize>.<min with samplename, regionSize and minCounts replaced by the sample name and the respective parameters given. Each data.frame object contains a list of genomic regions covered by the given samples, consisting of the columns: 'chr', 'start', 'stop', 'meth', 'unmeth' (signifying chromosome name, start of region by chromosomal position, end of region by chromosomal position, as well as methylated- and unmethylated cytosine counts at that chromosomal position).

```
> positions_to_regions(params)

Sample: reference.positions.csv regionSize: 10000 minCounts: 5
Processed reference.positions.csv, yielding 16728 regions of 10000 nt
Sample: sample.positions.csv regionSize: 10000 minCounts: 5
Processed sample.positions.csv, yielding 15948 regions of 10000 nt
```

5 Statistical modeling

Most of details about the statistical model explained in this section is also explained in our upcoming paper on the computational analysis for single-cell BS-Seq analysis. We repeat the statistical details in this section for the sake of completeness of the description of the computational and statistical details of our pipeline. Taking as input the region-based counts computed in the step above, the underlying model of the BEAT package then computes corrected counts, taking into account especially incomplete conversion rates (taken from the *'convrates'* parameter) and the estimated sequencing error (specified as the *'pminus'* parameter). The model is described briefly as follows.
We have derived a Bayesian statistical model that gives detailed information about the methylation rate in a region of multiple CpG positions which is described below. Apart from estimating the methylation rate, it provides measures of confidence for this estimate, it can test regions for high or low methylation. On the basis of of these tests it is later possible to give a precise

definition of a regional epimutation event. For multi-cell samples, we assume that all counts at a single CpG position were obtained from pairwise different bisulfite converted DNA template strands and represent independent observations. This certainly holds in good approximation, because the number of available DNA template strands typically supersedes the read coverage at this position by far. For single cell samples, we encounter the opposite situation: There are at most two template DNA strands available, and for many CpG positions this number is reduced further through DNA degradation. Multiple reads covering one CpG position are therefore highly dependent. We combine multiple counts at one position to one single (non-)methylation call. For different CpG position, these calls are then independent observations. First, fix one region, i.e. some set of CpG positions. The number of counts at a given position is the number of reads mapping to that position. Let n denote the total number of counts at all CpG positions in the given region, and let k (respectively $n-k$) of them indicate methylation (respectively non-methylation). Let r be the (unknown) methylation rate at the given position. Then, assuming independence of the single counts as mentioned above, the actual number j of counts originating from methylated CpGs in this region follows a binomial distribution,

$$P(j \mid n, r) = Bin(j; n, r) \tag{1}$$

Let the false positive rate p_+ be the global rate of false methylation counts, which is identical to the non-conversion rate of non-methylated cytosines. Conversely, let the false negative rate p_- be the global rate of false non-methylation counts, which is identical to the inappropriate conversion rate of methylated cytosines. One can find an upper bound for p_+ by considering all methylation counts at non-CpG positions as false positives (resulting from non-conversion of presumably unmethylated cytosines). This leads to an estimate of $p_+ = 0.2$, 0.51, 0.41, 0.44, 0.39 in the experiments Liver, H1, H2, H3 and H4, respectively. Bear in mind that in single cell bisulfite experiments, the limited DNA amount requires a particularly mild bisulfite treatment, which increases the false positive rate relative to standard bisulfite sequencing procedures. In the literature, false negative rates were not described, an estimate of $p_- = 0.01$ is reported[9]. We chose a conservative value of $p_- = 0.2$, which takes into account potential errors originating from mapping artifacts or sequencing errors. Due to failed or inappropriate conversion, the number k of counts indicating methylation differs from the actual number j of counts originating from methylated CpGs. Given the true number of methylation counts j, the the observed methylation counts k are the sum of the number m of correctly identified methylations and the number $k-m$ if incorrectly identified methylations (false positives). Hence, the probability distribution of k is a convolution of two binomial distributions,

$$\begin{aligned} P(k \mid j, n;\, p_+, p_-) &= \sum_{m=0}^{k} P(m \mid j, 1-p_-) \cdot P(k-m \mid n-j, p_+) \\ &= \sum_{m=0}^{k} \underbrace{Bin(m; j, 1-p_-)}_{=:C^1_{m,j}} \cdot \underbrace{Bin(k-m; n-j, p_+)}_{=:C^2_{n-j,k-m}} \end{aligned} \tag{2}$$

In (**??**), we use the convention that $Bin(m; j, p) = 0$ whenever $m > j$. Thus, given n reads, k

methylation counts, the likelihood function for r is a mixture of Bionomial distributions,

$$\begin{aligned} P(k \mid n, r; p_+, p_-) &= \sum_{j=0}^{n} P(k, j \mid n, r, p_+, p_-) \\ &= \sum_{j=0}^{n} P(k \mid j, n, r, p_+, p_-) \cdot P(j \mid n, r, p_+, p_-) \\ &= \sum_{j=0}^{n} P(k \mid j, n, p_+, p_-) \cdot P(j \mid n, r) \\ &\overset{(??,??)}{=} \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot Bin(j; n, r) \end{aligned} \tag{3}$$

In our Bayesian approach, we furthermore need to specify a prior for r to calculate the posterior distribution of r. Recall the beta distribution(s), which is a 2-parameter family of continuous probability distributions defined the unit interval $[0, 1]$,

$$Beta(r; \alpha, \beta) \propto r^{\alpha-1}(1-r)^{\beta-1} \ , \text{ for } \alpha, \beta > 0, \ r \in (0, 1) \ ,$$

We assume that a fraction of λ_m postions are essentially methylated, i.e., and that their rate r follows a $Beta(r; \alpha = r_m \cdot w_m, \beta_m = (1 - r_m) \cdot w_m)$ distribution, having an expectation value for r of $\frac{\alpha_m}{\alpha_m + \beta_m} = r_m$. Here, we set $r_m = 0.7$. The additional parameter w_m weights the strength of the prior relative to the strength of the likelihood. Since (the confidence into/ the knowledge about) our prior distribution of methylation rates is rather weak, we want our procedure to be strongly data-driven, therefore we choose a low w_m, $w_m = 0.5$. A fraction of $\lambda_u = 1 - \lambda_m$ is essentially unmethylated, and their rate is assumed to follow a $Beta(r; \alpha_u = r_u \cdot w_u, \beta_u = (1 - r_u) \cdot w_u)$ distribution, having an expectation value for r of $\frac{\alpha_u}{\alpha_u + \beta_u} = r_u$, where we set $r_u = 0.2$ and $w_u = 0.5$. Thus, the prior distribution $\pi(r)$ is a 2-Beta mixture distribution,

$$\pi(r; \alpha_m, \beta_m, \alpha_u, \beta_u, \lambda_m) = \sum_{s \in \{m,u\}} \lambda_s Beta(r; \alpha_s, \beta_s) \tag{4}$$

The pragmatic reason for choosing a Beta mixture as a prior distribution is the fact that the Beta distribution is the conjugate prior of the Binomial distribution[15], such that for some normalizing constant $D^{\alpha,\beta}_{j,n}$,

$$Bin(j; n, r) \cdot Beta(r; \alpha, \beta) = D^{\alpha,\beta}_{j,n} \cdot Beta(r; ; j + \alpha, n - j + \beta) \tag{5}$$

By virtue of Equation (**??**), we can write down the posterior distribution of r analytically (Equation **??**). This has the advantage that we can answer all questions on the posterior distribution of r efficiently and up to an arbitrary precision. Efficiency is an issue, because we need to calculate posterior distributions for all regions, which can easily amount to millions.

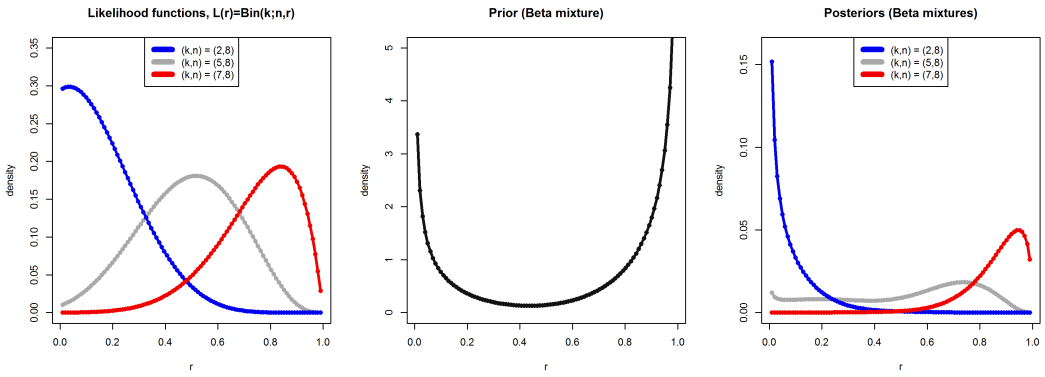

Figure 1: Plot of the likelihood functions for three different observations (k, n) (left), the Beta mixture prior distribution (middle) and the corresponding three posteriors (right). The number n of counts is set to 8, of which $k = 2$ (blue), $k = 5$ (grey) and $k = 7$ (red) are methylation counts. The unknown parameter $p_+$was determined empirically from the false non-CpG methylation, which reflects the incomplete conversion rate, as follows: L1: 0.2, H1: 0.51, H2: 0.41, H3: 0.44, H4: 0.39. $p_-$was set to 0.2 as a very robust choice. The beta mixture prior was set as described in the text.

$$
\begin{aligned}
&P(r \mid k, n;\, p_+, p_-; \alpha_m, \beta_m, \alpha_u, \beta_u, \lambda_m) \\
&\quad = \; N^{-1} \cdot P(k \mid n, r;\, p_+, p_-) \cdot \pi(r; \alpha_m, \beta_m, \alpha_u, \beta_u) \qquad (6) \\
&\quad \overset{(??,??)}{=} \; N^{-1} \cdot \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot Bin(j; n, r) \cdot \sum_{s \in \{m,u\}} \lambda_s Beta(r; \alpha_s, \beta_s) \\
&\quad \overset{(??)}{=} \; N^{-1} \cdot \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot \left(\sum_{s \in \{m,u\}} \lambda_s D^{\alpha_s,\beta_s}_{j,n} Beta(r; j + \alpha_s, n - j + \beta_s) \right) \qquad (7)
\end{aligned}
$$

In the above equation, N is a normalization constant,

$$
N = \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot \sum_{s} \lambda_s D^{\alpha_s,\beta_s}_{j,n} \qquad (8)
$$

The ingredients for the construction of the posterior distribution are visualized in Figure (**??**).

5.1 Methylation statistics derived from read counts

For each region under consideration, we obtain an individual posterior distribution $P(r \mid k, n, p_+, p_-)$. With this posterior at hand, it is an easy task to calculate the expected methylation rate $\hat{r}$ in the corresponding region,

$$
\hat{r} = \int_0^1 r \cdot P(r \mid k, n, p_+, p_-)\, dr \qquad (9)
$$

It is customary to provide a Bayesian measure of uncertainty of this estimate, a so-called credible interval. A credible interval is an interval which contains the estimate $(\hat{r})$ and in which a prescribed probability mass of the posterior is located. One can construct a 90% credible interval $[m, M]$ as

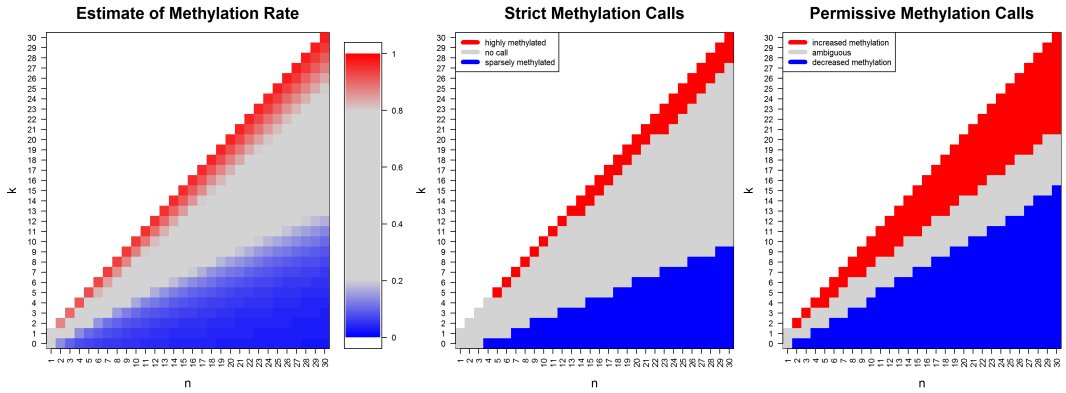

Figure 2: Illustration of the the results of our statistical modeling applied to regions of size $d = 1000$ in the Liver sample. In each plot, n (on the x-axis) denotes the total number of counts mapping to that region, of which k (on the y-axis) are counts indicating methylation. Left: Using the Liver-specific estimates of the false positive rate $p_+ = 0.2$ and the false negative rate $p_- = 0.1$ and the methylation prior in Equation (**??**), we obtain for each admissible pair (k, n) a methylation rate estimates $\hat{r}$ from Equation (**??**). Colors correspond to methylation rate, ranging from deep blue (zero methylation) to deep red (full methylation). Middle: The red respectively blue area defines the pairs (k, n) which satisfy our criteria for high respectively sparse methylation. Right: The red respectively blue area defines the pairs (k, n) which satisfy our criteria for increased respectively decreased methylation. Note that strict methylation calls are only made when at least $n = 5$ counts were observed.

the shortest interval containing $\hat{r}$ such that $P(r \in [n, M] \mid k, n, p_+, p_-) = 0.9$. Moreover, we call a region *highly methylated* if

$$P(r > 0.7 \mid k, n, p_+, p_-) > c \tag{10}$$

for some stringency level c which we set to 0.75 here. The false negative methylation calling rates were set to $p_- = 0.1$ for all samples, and the false positive calling rates were determined by $p_+ = 1 -$ CH methylation rate for each sample separately. A region is said to show *increased methylation* if

$$P(r > 0.5 \mid k, n, p_+, p_-) > c$$

Analogously, a region is called *sparsely methylated* if

$$P(r < 0.3 \mid k, n, p_+, p_-) > c$$

and a region with *decreased methylation* satisfies

$$P(r < 0.5 \mid k, n, p_+, p_-) > c \tag{11}$$

By definition, any highly methylated region has increased methylation, and every sparsely methylated region shows decreased methylation. For $c > 0.5$, high and sparse methylation calls are mutually exclusive. Regions that are neither highly nor sparsely methylated are called *ambiguous*. The time-critical step is the calculation of the region-specific posterior distribution $P(r \mid k, n, p_+, p_-)$, and the quantities related to it (Equations **??**-**??**). Since k and n vary for each region, and the number of regions is large, we save a lot of time by pre-calculating all required quantities for a

set of values $n = 1, ..., 45$, $k = 0, ..., n$. The statistics for, on average, 80% of all regions can then be looked up and do not need to be re-computed. The running times for $d = 250, 500, ...$ on the mouse genome took less than a minute plus $t = 25$ min for the pre-computing of each of the five samples, which did not vary substantially with region size.
The model generates posterior probabilities for each pair of methylated and total count values given the conversion rate and sequencing error. These posteriors are saved in the working directory for each sample in an R object under the file name <samplename>.convMat.<pminus>.<pplus>.RData, while the corrected counts along with corrected methylation rate estimates and other model data are saved as data.frames under the file name <samplename>.results.<pminus>.<pplus>.RData, where pminus and pplus are replaced by the respective parameters. For each sample, the model computes a data.frame consisting of the columns: 'chr', 'start', 'stop', 'meth', 'unmeth', 'methstate' and 'methest', signifying chromosome name, start of region by chromosomal position, end of region by chromosomal position, methylated- and unmethylated cytosine counts at that chromosomal position, methylation state (which is 1 for regions called methylated, -1 for unmethylated regions, and 0 for regions that are neither) and methylation estimate (i.e., the model-estimated methylation level for that region). The object contains further columns which are used for internal procedures and irrelevant for the package user.

```
> generate_results(params)

Sample: reference
Generating conversion matrix...
Adding methylation states and 9 matrices...
Sample: sample
Generating conversion matrix...
Adding methylation states and 9 matrices...
```

6 Epimutation calling

Finally, epimutations are called by comparing two results objects from the previous step.
The function 'epimutation_calls' compares one sample to a reference sample. A methylating epimutation is called for each genomic region common to both samples when it was called as unmethylated in the reference and as methylated in the other sample by the model. Accordingly, the region is called as demethylating epimutation when called as methylated in the reference and as unmethylated in the other sample. Epimutation rates are then computed as the frequency of epimutation events in relation to the total regions shared between each sample and the reference. The resulting epimutations are saved as data.frames, for each sample one object for methylating epimutations and one for demethylating epimutations, in the working directory as<sampleName>.methEpicalls.<regionSize>.<minCounts>.p+=<pplus>.p-=<pminus>.RData and <sampleName>.demethEpicalls.<regionSize>.<minCounts>.p+=<pplus>.p-=<pminus>.RData, respectively. Each data.frame object contains a list of genomic regions covered by the given samples, consisting of the columns: 'chr', 'pos', 'endpos', 'meth', 'unmeth', 'methstate', signifying chromosome name, start of region by chromosomal position, end of region by chromosomal position, methylated- and unmethylated cytosine counts at that chromosomal position and the methylation state, which is 1 for methylated positions and -1 for unmethylated positions.

```
> epiCalls <- epimutation_calls(params)

Statistics for sample: sample (min. coverage: 5 reads/site)
============================================================
Total shared CG sites(=regions!) between sample and reference/CTRL: 14492
Median CG sites of these total shared regions (REFERENCE): 23
Median CG sites of these total shared regions (SINGLE CELL): 13
Control has  10164  fully methylated and  967  fully unmethylated sites
Methylating Epimutation Rate: 0.00283
Demethylating Epimutation Rate: 0.47661
Total Epimutation Rate: 0.47944
Fully meth in single: 606, Fully unmeth in single: 9985
Written epi-methylation calls to matrix:
sample.methEpicalls.10000.5p+=0.5p-=0.2.RData
Written epi-demethylation calls to matrix:
sample.demethEpicalls.10000.5p+=0.5p-=0.2.RData

> head(epiCalls$methSites$singlecell,3)

NULL

> head(epiCalls$demethSites$singlecell,3)

NULL
```