# Package ‘methylCC’

**May 30, 2024**

**Title**  Estimate the cell composition of whole blood in DNA methylation samples

**Version**  1.18.0

**Imports**  Biobase, GenomicRanges, IRanges, S4Vectors, dplyr, magrittr, minfi, bsseq, quadprog, plyranges, stats, utils, bumphunter, genefilter, methods, IlluminaHumanMethylation450kmanifest, IlluminaHumanMethylation450kanno.ilmn12.hg19

**Depends**  R (>= 3.6), FlowSorted.Blood.450k

**Suggests**  rmarkdown, knitr, testthat (>= 2.1.0), BiocGenerics, BiocStyle, tidyr, ggplot2

**Description**  A tool to estimate the cell composition of DNA methylation whole blood sample measured on any platform technology (microarray and sequencing).

**biocViews**  Microarray, Sequencing, DNAMethylation, MethylationArray, MethylSeq, WholeGenome

**VignetteBuilder**  knitr

**RoxygenNote**  6.1.1

**Encoding**  UTF-8

**License**  GPL-3

**BugReports**  [https://github.com/stephaniehicks/methylCC/](https://github.com/stephaniehicks/methylCC/)

**URL**  [https://github.com/stephaniehicks/methylCC/](https://github.com/stephaniehicks/methylCC/)

**git_url**  https://git.bioconductor.org/packages/methylCC

**git_branch**  RELEASE_3_19

**git_last_commit**  803b15d

**git_last_commit_date**  2024-04-30

**Repository**  Bioconductor 3.19

**Date/Publication**  2024-05-29

**Author**  Stephanie C. Hicks [aut, cre] (<https://orcid.org/0000-0002-7858-0231>), Rafael Irizarry [aut] (<https://orcid.org/0000-0002-3944-4309>)

**Maintainer**  Stephanie C. Hicks <shicks19@jhu.edu>
Contents

.extract_raw_data .................................................. 2
.find_dmrs ................................................................. 3
.initializeMLEs ............................................................ 4
.initialize_theta ........................................................... 5
.methylcc_engine .......................................................... 5
.methylcc_estep ............................................................. 6
.methylcc_mstep ............................................................ 7
.pick_target_positions .................................................... 7
.preprocess_estimatecc .................................................... 8
.splitit ................................................................. 8
.WFun ................................................................. 9
.cell_counts ............................................................... 9
.estimatecc ............................................................. 10
.estimatecc-class .......................................................... 12
.FlowSorted.Blood.450k.sub ............................................... 12
.offMethRegions .......................................................... 13
.onMethRegions ............................................................ 13

Index ......................................................................... 14

.extract_raw_data Extract raw data

Description

Extract the methylation values and GRanges objects

Usage

.extract_raw_data(object)

Arguments

object an object can be a RGChannelSet, GenomicMethylSet or BSseq object

Value

A list preprocessed objects from the RGChannelSet, GenomicMethylSet or BSseq objects to be used in .preprocess_estimatecc().
Description

This function uses the FlowSorted.Blood.450k whole blood reference methylomes with six cell types to identify differentially methylated regions.

Usage

```r
.find_dmrs(verbos = TRUE, gr_target = NULL, include_cpgs = FALSE,
include_dmrs = TRUE, num_cpgs = 50, num_regions = 50,
bumphunter_beta_cutoff = 0.2, dmr_up_cutoff = 0.5,
dmr_down_cutoff = 0.4, dmr_pval_cutoff = 1e-11,
cpg_pval_cutoff = 1e-08, cpg_up_dm_cutoff = 0,
cpg_down_dm_cutoff = 0, pairwise_comparison = FALSE,
mset_train_flow_sort = NULL)
```

Arguments

- `verbose` TRUE/FALSE argument specifying if verbose messages should be returned or not. Default is TRUE.
- `gr_target` Default is NULL. However, the user can provide a GRanges object from the object in estimatecc. Before starting the procedure to find differentially methylated regions, the intersection of the `gr_target` and GRanges object from the reference methylomes (FlowSorted.Blood.450k).
- `include_cpgs` TRUE/FALSE. Should individual CpGs be returned. Default is FALSE.
- `include_dmrs` TRUE/FALSE. Should differentially methylated regions be returned. Default is TRUE. User can turn this to FALSE and search for only CpGs.
- `num_cpgs` The max number of CpGs to return for each cell type. Default is 50.
- `num_regions` The max number of DMRs to return for each cell type. Default is 50.
- `bumphunter_beta_cutoff` The cutoff threshold in bumphunter() in the bumphunter package.
- `dmr_up_cutoff` A cutoff threshold for identifying DMRs that are methylated in one cell type, but not in the other cell types.
- `dmr_down_cutoff` A cutoff threshold for identifying DMRs that are not methylated in one cell type, but methylated in the other cell types.
- `dmr_pval_cutoff` A cutoff threshold for the p-values when identifying DMRs that are methylated in one cell type, but not in the other cell types (or vice versa).
- `cpg_pval_cutoff` A cutoff threshold for the p-values when identifying differentially methylated CpGs that are methylated in one cell type, but not in the other cell types (or vice versa).
.initializeMLEs

Initialize MLEs

Description
Helper functions to initialize MLEs in estimatecc().

Usage
.initializeMLEs(init_param_method, n, K, Ys, Zs, a0init, a1init, sig0init, sig1init, tauinit)

Arguments

init_param_method
method to initialize parameter estimates. Choose between "random" (randomly sample) or "known_regions" (uses unmethylated and methylated regions that were identified based on Reinus et al. (2012) cell sorted data.). Defaults to "random".

n
Number of samples

K
Number of cell types

Ys
observed methylation levels in samples provided by user of dimension R x n

Zs
Cell type specific regions of dimension R x K

a0init
Default NULL. Initial mean methylation level in unmethylated regions

a1init
Default NULL. Initial mean methylation level in methylated regions

sig0init
Default NULL. Initial var methylation level in unmethylated regions

sig1init
Default NULL. Initial var methylation level in methylated regions

tauinit
Default NULL. Initial var for measurement error

Values
A list of data frames and GRanges objects.
.initialize_theta

Value
A list of MLE estimates to be used in estimatecc().

Description
Creates a container with initial theta parameter estimates

Usage
.initialize_theta(n, K, alpha0 = NULL, alpha1 = NULL, sig0 = NULL, 
sig1 = NULL, tau = NULL)

Arguments
- n: Number of samples
- K: Number of cell types
- alpha0: Default NULL. Initial mean methylation level in unmethylated regions
- alpha1: Default NULL. Initial mean methylation level in methylated regions
- sig0: Default NULL. Initial var methylation level in unmethylated regions
- sig1: Default NULL. Initial var methylation level in methylated regions
- tau: Default NULL. Initial var for measurement error

Value
A data frame with initial parameter estimates to be used in .initializeMLEs().

.methylcc_engine

Description
Helper function for estimatecc

Usage
.methylcc_engine(Ys, Zs, current_pi_mle, current_theta, epsilon, max_iter)
Arguments

- **Ys**: observed methylation levels in samples provided by user of dimension \( R \times n \)
- **Zs**: Cell type specific regions of dimension \( R \times K \)
- **current_pi_mle**: cell composition MLE estimates of dimension \( K \times n \)
- **current_theta**: other parameter estimates in EM algorithm
- **epsilon**: Add here.
- **max_iter**: Add here.

Value

A list of MLE estimates that is used in \texttt{estimatecc()}.

Description

Expectation step in EM algorithm for methylCC

Usage

\[
\text{.methylcc_estep}(Ys, Zs, \text{current_pi_mle}, \text{current_theta}, \text{meth_status} = 0)
\]

Arguments

- **Ys**: observed methylation levels in samples provided by user of dimension \( R \times n \)
- **Zs**: Cell type specific regions of dimension \( R \times K \)
- **current_pi_mle**: cell composition MLE estimates of dimension \( K \times n \)
- **current_theta**: other parameter estimates in EM algorithm
- **meth_status**: Indicator function corresponding to regions that are unmethylated (\( \text{meth_status}=0 \)) or methylated (\( \text{meth_status}=1 \))

Value

List of expected value of the first two moments of the random effects (or the E-Step in the EM algorithm) used in \texttt{.methylcc_engine()}.
.methylcc_mstep

Maximization step

Description
Maximization step in EM Algorithm for methylCC

Usage
.methylcc_mstep(Ys, Zs, current_pi_mle, current_theta, estep0, estep1)

Arguments
Ys observed methylation levels in samples provided by user of dimension R x n
Zs Cell type specific regions of dimension R x K
current_pi_mle cell composition MLE estimates of dimension K x n
current_theta other parameter estimates in EM algorithm
estep0 Results from expectation step for unmethylated regions
estep1 Results from expectation step for methylated regions

Value
A list of the updated MLEs (or the M-Step in the EM algorithm) used in .methylcc_engine()

.pick_target_positions

Pick target positions

Description
Pick probes from target data using the indices in dmp_regions

Usage
.pick_target_positions(target_granges, target_object = NULL,
                        target_cvg = NULL, dmp_regions)

Arguments
target_granges add more here.
target_object an optional argument which contains the meta-data for target_granges. If
target_granges already contains the meta-data, do not need to supply target_object.
target_cvg coverage reads for the target object
dmp_regions differentially methylated regions
.preprocess_estimatecc

Description

This function preprocesses the data before the estimatecc() function

Usage

.preprocess_estimatecc(object, verbose = TRUE,
    init_param_method = "random",
    celltype_specific_dmrs = celltype_specific_dmrs)

Arguments

object an object can be an RGChannelSet, GenomicMethylSet or BSseq object
verbose TRUE/FALSE argument specifying if verbose messages should be returned or not. Default is TRUE.
init_param_method method to initialize parameter estimates. Choose between "random" (randomly sample) or "known_regions" (uses unmethylated and methylated regions that were identified based on Reinus et al. (2012) cell sorted data.). Defaults to "random".
celltype_specific_dmrs cell type specific differentially methylated regions (DMRs).

Value

A list of object to be used in estimatecc

.splitit

Description

helper function to split along a variable

Usage

.splitit(x)
Arguments

x a vector

Value

A list to be used in find_dmr.s()

Description

Helper function to take the product of Z and cell composition estimates

Usage

.WFun(Zs, pi_mle)

Arguments

Zs Cell type specific regions of dimension R x K
pi_mle cell composition MLE estimates

Value

A list of output after taking the product of Z and cell composition mle estimates to be used in .methylcc_estep().

cell_counts Generic function that returns the cell composition estimates

Description

Given a estimatecc object, this function returns the cell composition estimates

Accessors for the 'cell_counts' slot of a estimatecc object.

Usage

cell_counts(object)

## S4 method for signature 'estimatecc'
cell_counts(object)

Arguments

object an object of class estimatecc.
estimatecc

Value

Returns the cell composition estimates

Examples

# This is a reduced version of the FlowSorted.Blood.450k
# dataset available by using BiocManager::install("FlowSorted.Blood.450k"),
# but for purposes of the example, we use the smaller version
# and we set \code{demo=TRUE}. For any case outside of this example for
# the package, you should set \code{demo=FALSE} (the default).

dir <- system.file("data", package="methylCC")
files <- file.path(dir, "FlowSorted.Blood.450k.sub.RData")
if(file.exists(files)){
  load(file = files)
  set.seed(12345)
  est <- estimatecc(object = FlowSorted.Blood.450k.sub, demo = TRUE)
  cell_counts(est)
}

estimatecc

Estimate cell composition from DNAm data

Description

Estimate cell composition from DNAm data

Usage

estimatecc(object, find_dmrs_object = NULL, verbose = TRUE,
epsilon = 0.01, max_iter = 100, take_intersection = FALSE,
include_cpgs = FALSE, include_dmrs = TRUE,
init_param_method = "random", a0init = NULL, a1init = NULL,
sig0init = NULL, sig1init = NULL, tauinit = NULL, demo = FALSE)

Arguments

object an object can be a RGChannelSet, GenomicMethylSet or BSseq object
find_dmrs_object If the user would like to supply different differentially methylated regions, they
can use the output from the find_dmrs function to supply different regions to estimatecc.
verbose TRUE/FALSE argument specifying if verbose messages should be returned or
not. Default is TRUE.
epsilon Threshold for EM algorithm to check for convergence. Default is 0.01.
max_iter Maximum number of iterations for EM algorithm. Default is 100 iterations.
take_intersection TRUE/FALSE asking if only the CpGs included in object should be used to find DMRs. Default is FALSE.
include_cpgs TRUE/FALSE. Should individual CpGs be returned. Default is FALSE.
include_dmrs TRUE/FALSE. Should differentially methylated regions be returned. Default is TRUE.
init_param_method method to initialize parameter estimates. Choose between "random" (randomly sample) or "known_regions" (uses unmethyalted and methylated regions that were identified based on Reinus et al. (2012) cell sorted data.). Defaults to "random".
a0init Default NULL. Initial mean methylation level in unmethylated regions
a1init Default NULL. Initial mean methylation level in methylated regions
sig0init Default NULL. Initial var methylation level in unmethylated regions
sig1init Default NULL. Initial var methylation level in methylated regions
tauinit Default NULL. Initial var for measurement error
demo TRUE/FALSE. Should the function be used in demo mode to shorten examples in package. Defaults to FALSE.

Value
A object of the class estimatecc that contains information about the cell composition estimation (in the summary slot) and the cell composition estimates themselves (in the cell_counts slot).

Examples

# This is a reduced version of the FlowSorted.Blood.450k
# dataset available by using BiocManager::install("FlowSorted.Blood.450k),
# but for purposes of the example, we use the smaller version
# and we set \code{demo=TRUE}. For any case outside of this example for
# the package, you should set \code{demo=FALSE} (the default).

dir <- system.file("data", package="methylCC")
files <- file.path(dir, "FlowSorted.Blood.450k.sub.RData")
if(file.exists(files)){
  load(file = files)
  set.seed(12345)
  est <- estimatecc(object = FlowSorted.Blood.450k.sub, demo = TRUE)
  cell_counts(est)
}
estimatecc-class

the estimatecc class

Description

Objects of this class store all the values needed information to work with a estimatecc object

Value

summary returns the summary information about the cell composition estimate procedure and cell_counts returns the cell composition estimates

Slots

summary information about the samples and regions used to estimate cell composition
cell_counts cell composition estimates

Examples

# This is a reduced version of the FlowSorted.Blood.450k
# dataset available by using BiocManager::install("FlowSorted.Blood.450k"),
# but for purposes of the example, we use the smaller version
# and we set \code{demo=TRUE}. For any case outside of this example for
# the package, you should set \code{demo=FALSE} (the default).

dir <- system.file("data", package="methylCC")
files <- file.path(dir, "FlowSorted.Blood.450k.sub.RData")
if(file.exists(files)){
  load(file = files)
  set.seed(12345)
est <- estimatecc(object = FlowSorted.Blood.450k.sub, demo = TRUE)
cell_counts(est)
}

FlowSorted.Blood.450k.sub

A reduced size of the FlowSorted.Blood.450k dataset

Description

A reduced size of the FlowSorted.Blood.450k dataset

The object was created using the script in /inst and located in the /data folder.

Format

A RGset object with 2e5 rows (probes) and 6 columns (whole blood samples).
### offMethRegions

<table>
<thead>
<tr>
<th>Description</th>
<th>Unmethylated regions for all celltypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is the script used to create the offMethRegions data set. The purpose is use in the <code>estimate_cc()</code> function.</td>
<td></td>
</tr>
<tr>
<td>The object was created using the script in /inst and located in the /data folder.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>add more here.</td>
</tr>
</tbody>
</table>

### onMethRegions

<table>
<thead>
<tr>
<th>Description</th>
<th>Methylated regions for all celltypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is the script used to create the onMethRegions data set. The purpose is use in the <code>estimate_cc()</code> function.</td>
<td></td>
</tr>
<tr>
<td>The object was created using the script in /inst and located in the /data folder.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>add more here.</td>
</tr>
</tbody>
</table>