

# Package ‘methylCC’

March 30, 2021

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

**Version** 1.4.0

**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** 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** CC BY 4.0

**BugReports** <https://github.com/stephaniehicks/methylCC/>

**URL** <https://github.com/stephaniehicks/methylCC/>

**git\_url** <https://git.bioconductor.org/packages/methylCC>

**git\_branch** RELEASE\_3\_12

**git\_last\_commit** f078962

**git\_last\_commit\_date** 2020-10-27

**Date/Publication** 2021-03-29

**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](mailto:shicks19@jhu.edu)>

## R topics documented:

.extract_raw_data . . . . .	2
.find_dmrs . . . . .	3
.initializeMLEs . . . . .	4
.initialize_theta . . . . .	5
.methylcc_engine . . . . .	5
.methylcc_estep . . . . .	6
.methylcc_mstep . . . . .	6
.pick_target_positions . . . . .	7
.preprocess_estimatecc . . . . .	7
.splitit . . . . .	8
.WFun . . . . .	8
cell_counts . . . . .	9
estimatecc . . . . .	9
estimatecc-class . . . . .	11
FlowSorted.Blood.450k.sub . . . . .	12
offMethRegions . . . . .	12
onMethRegions . . . . .	12
<b>Index</b>	<b>13</b>

---

.extract_raw_data	<i>Extract raw data</i>
-------------------	-------------------------

---

### 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().

`.find_dmrs`*Finding differentially methylated regions***Description**

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

**Usage**

```
.find_dmrs(verbose = 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**

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

<code>cpg_down_dm_cutoff</code>	A cutoff threshold for identifying differentially methylated CpGs that are not methylated in one cell type, but are methylated in the other cell types.
<code>pairwise_comparison</code>	TRUE/FAISE of whether all pairwise comparisons (e.g. methylated in Granulocytes and Monocytes, but not methylated in other cell types). Default if FALSE.
<code>mset_train_flow_sort</code>	Default is NULL. However, a user can provide a <code>MethylSet</code> object after processing the <code>FlowSorted.Blood.450k</code> dataset. The default normalization is <code>preprocessIllumina()</code> .

**Value**

A list of data frames and `GRanges` objects.

---

<code>.initializeMLEs</code>	<i>.initializeMLEs</i>
------------------------------	------------------------

---

**Description**

Helper functions to initialize MLEs in `estimatecc()`.

**Usage**

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

**Arguments**

<code>init_param_method</code>	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".
<code>n</code>	Number of samples
<code>K</code>	Number of cell types
<code>Ys</code>	observed methylation levels in samples provided by user of dimension $R \times n$
<code>Zs</code>	Cell type specific regions of dimension $R \times K$
<code>a0init</code>	Default NULL. Initial mean methylation level in unmethylated regions
<code>a1init</code>	Default NULL. Initial mean methylation level in methylated regions
<code>sig0init</code>	Default NULL. Initial var methylation level in unmethylated regions
<code>sig1init</code>	Default NULL. Initial var methylation level in methylated regions
<code>tauinit</code>	Default NULL. Initial var for measurement error

**Value**

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

---

*.initialize\_theta*      *.initialize\_theta*

---

**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*      *.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 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
- epsilon              Add here.
- max\_iter            Add here.

**Value**

A list of MLE estimates that is used in `estimatecc()`.

---

*.methyloc\_estep*      *Expectation step*

---

**Description**

Expectation step in EM algorithm for methyloc

**Usage**

```
.methyloc_estep(Ys, Zs, current_pi_mle, current_theta, meth_status = 0)
```

**Arguments**

<i>Ys</i>	observed methylation levels in samples provided by user of dimension $R \times n$
<i>Zs</i>	Cell type specific regions of dimension $R \times K$
<i>current_pi_mle</i>	cell composition MLE estimates of dimension $K \times n$
<i>current_theta</i>	other parameter estimates in EM algorithm
<i>meth_status</i>	Indicator function corresponding to regions that are unmethylated ( <i>meth_status</i> =0) or methylated ( <i>meth_status</i> =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 *.methyloc\_engine*()

---

*.methyloc\_mstep*      *Maximization step*

---

**Description**

Maximization step in EM Algorithm for methyloc

**Usage**

```
.methyloc_mstep(Ys, Zs, current_pi_mle, current_theta, estep0, estep1)
```

**Arguments**

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

**Value**

A list of the updated MLEs (or the M-Step in the EM algorithm) used in *.methyloc\_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

### **Value**

A list of `GRanges` objects to be used in `.preprocess_estimatecc()`

---

*.preprocess\_estimatecc*  
*.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 a `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

---

<code>.splitit</code>	<code>.splitit</code>
-----------------------	-----------------------

---

**Description**

helper function to split along a variable

**Usage**

```
.splitit(x)
```

**Arguments**

`x` a vector

**Value**

A list to be used in find\_dmrs()

---

<code>.WFun</code>	<i>Helper function to take the product of Z and cell composition estimates</i>
--------------------	--

---

**Description**

Helper function which is the product of Z and pi\_mle

**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	<i>Generic function that returns the cell composition estimates</i>
-------------	---

---

**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.

**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	<i>Estimate cell composition from DNAm data</i>
------------	---

---

**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**

<code>object</code>	an object can be a <code>RGChannelSet</code> , <code>GenomicMethylSet</code> or <code>BSseq</code> object
<code>find_dmrs_object</code>	If the user would like to supply different differentially methylated regions, they can use the output from the <code>find_dmrs</code> function to supply different regions to <code>estimatecc</code> .
<code>verbose</code>	TRUE/FALSE argument specifying if verbose messages should be returned or not. Default is TRUE.
<code>epsilon</code>	Threshold for EM algorithm to check for convergence. Default is 0.01.
<code>max_iter</code>	Maximum number of iterations for EM algorithm. Default is 100 iterations.
<code>take_intersection</code>	TRUE/FALSE asking if only the CpGs included in <code>object</code> should be used to find DMRs. Default is FALSE.
<code>include_cpgs</code>	TRUE/FALSE. Should individual CpGs be returned. Default is FALSE.
<code>include_dmrs</code>	TRUE/FALSE. Should differentially methylated regions be returned. Default is TRUE.
<code>init_param_method</code>	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".
<code>a0init</code>	Default NULL. Initial mean methylation level in unmethylated regions
<code>a1init</code>	Default NULL. Initial mean methylation level in methylated regions
<code>sig0init</code>	Default NULL. Initial var methylation level in unmethylated regions
<code>sig1init</code>	Default NULL. Initial var methylation level in methylated regions
<code>tauinit</code>	Default NULL. Initial var for measurement error
<code>demo</code>	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	<i>the estimatecc class</i>
------------------	-----------------------------

---

## 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

*Unmethylated regions for all celltypes*

---

### **Description**

This is the script used to create the offMethRegions data set. The purpose is use in the estimate\_cc() function.

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

### **Format**

add more here.

---

onMethRegions

*Methylated regions for all celltypes*

---

### **Description**

This is the script used to create the onMethRegions data set. The purpose is use in the estimate\_cc() function.

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

### **Format**

add more here.

# Index

`.WFun`, 8  
`.extract_raw_data`, 2  
`.find_dmrs`, 3  
`.initializeMLEs`, 4  
`.initialize_theta`, 5  
`.methylcc_engine`, 5  
`.methylcc_estep`, 6  
`.methylcc_mstep`, 6  
`.pick_target_positions`, 7  
`.preprocess_estimatecc`, 7  
`.splitit`, 8  
  
`cell_counts`, 9  
`cell_counts, estimatecc-method`  
    (`cell_counts`), 9  
  
`estimatecc`, 9  
`estimatecc-class`, 11  
  
`FlowSorted.Blood.450k.sub`, 12  
  
`offMethRegions`, 12  
`onMethRegions`, 12