

# Package ‘betaHMM’

December 4, 2025

**Type** Package

**Title** A Hidden Markov Model Approach for Identifying Differentially Methylated Sites and Regions for Beta-Valued DNA Methylation Data

**Version** 1.7.0

**Description** A novel approach utilizing a homogeneous hidden Markov model. And effectively model untransformed beta values. To identify DMCs while considering the spatial. Correlation of the adjacent CpG sites.

**License** GPL-3

**Encoding** UTF-8

**Imports** stats, ggplot2, scales, methods, pROC, foreach, doParallel, parallel, cowplot, dplyr, tidyr, tidyselect, stringr, utils

**Depends** R (>= 4.3.0), SummarizedExperiment, S4Vectors, GenomicRanges

**RoxygenNote** 7.2.3

**biocViews** DNAMethylation, DifferentialMethylation, ImmunoOncology, BiomedicalInformatics, MethylationArray, Software, MultipleComparison, Sequencing, Spatial, Coverage, GeneTarget, HiddenMarkovModel, Microarray

**Suggests** rmarkdown, knitr, testthat (>= 3.0.0), BiocStyle

**VignetteBuilder** knitr

**Config/testthat/edition** 3

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

**git\_branch** devel

**git\_last\_commit** 3efe81d

**git\_last\_commit\_date** 2025-10-29

**Repository** Bioconductor 3.23

**Date/Publication** 2025-12-04

**Author** Koyel Majumdar [cre, aut] (ORCID: <https://orcid.org/0000-0001-6469-488X>),  
Romina Silva [aut],

Antoinette Sabrina Perry [aut],  
 Ronald William Watson [aut],  
 Isobel Claire Gorley [aut] (ORCID:  
<https://orcid.org/0000-0001-7713-681X>),  
 Thomas Brendan Murphy [aut] (ORCID:  
<https://orcid.org/0000-0002-5668-7046>),  
 Florence Jaffrezic [aut],  
 Andrea Rau [aut] (ORCID: <https://orcid.org/0000-0001-6469-488X>)

**Maintainer** Koyel Majumdar <koyelmajumdar.phdresearch@gmail.com>

## Contents

|                                        |           |
|----------------------------------------|-----------|
| annotatedData . . . . .                | 2         |
| annotation_data . . . . .              | 7         |
| AUC_DM_analysis . . . . .              | 8         |
| backward . . . . .                     | 9         |
| BaumWelch . . . . .                    | 10        |
| betaHMM . . . . .                      | 11        |
| betaHMMResults-class . . . . .         | 17        |
| betaHMMrun . . . . .                   | 18        |
| dmcResults-class . . . . .             | 20        |
| dmc_identification . . . . .           | 21        |
| dmc_identification_run . . . . .       | 23        |
| dmrResults-class . . . . .             | 25        |
| dmr_identification . . . . .           | 26        |
| dmr_identification_run . . . . .       | 27        |
| forward . . . . .                      | 29        |
| initialise_parameters . . . . .        | 29        |
| pca_methylation_data . . . . .         | 30        |
| plot . . . . .                         | 31        |
| sample_annotation_file . . . . .       | 34        |
| sample_methylation_file . . . . .      | 35        |
| summary . . . . .                      | 36        |
| threshold_identification . . . . .     | 37        |
| threshold_identification_run . . . . . | 39        |
| threshold_Results-class . . . . .      | 40        |
| Viterbi . . . . .                      | 41        |
| <b>Index</b>                           | <b>43</b> |

---

|               |                                           |
|---------------|-------------------------------------------|
| annotatedData | <i>Accessors for the betaHMM package.</i> |
|---------------|-------------------------------------------|

---

## Description

The accessor methods for accessing the betaHMMResults/ dmcResults/ dmrResults/ threshold\_Results metadata.

**Usage**

```
annotatedData(object)

K(object)

N(object)

R(object)

A(object)

phi(object)

treatment_group(object)

llk(object)

tau(object)

hidden_states(object)

chromosome_number(object)

AUC(object)

uncertainty(object)

model_parameters(object)

threshold(object)

## S4 method for signature 'betaHMMResults'
annotatedData(object)

## S4 method for signature 'threshold_Results'
annotatedData(object)

## S4 method for signature 'RangedSummarizedExperiment'
K(object)

## S4 method for signature 'betaHMMResults'
K(object)

## S4 method for signature 'dmcResults'
K(object)

## S4 method for signature 'threshold_Results'
K(object)
```

```
## S4 method for signature '`NULL`'  
K(object)  
  
## S4 method for signature 'RangedSummarizedExperiment'  
N(object)  
  
## S4 method for signature 'betaHMMResults'  
N(object)  
  
## S4 method for signature 'dmcResults'  
N(object)  
  
## S4 method for signature '`NULL`'  
N(object)  
  
## S4 method for signature 'RangedSummarizedExperiment'  
R(object)  
  
## S4 method for signature 'betaHMMResults'  
R(object)  
  
## S4 method for signature 'dmcResults'  
R(object)  
  
## S4 method for signature '`NULL`'  
R(object)  
  
## S4 method for signature 'RangedSummarizedExperiment'  
A(object)  
  
## S4 method for signature 'betaHMMResults'  
A(object)  
  
## S4 method for signature '`NULL`'  
A(object)  
  
## S4 method for signature 'RangedSummarizedExperiment'  
tau(object)  
  
## S4 method for signature 'betaHMMResults'  
tau(object)  
  
## S4 method for signature '`NULL`'  
tau(object)  
  
## S4 method for signature 'RangedSummarizedExperiment'  
treatment_group(object)
```

```
## S4 method for signature 'betaHMMResults'
treatment_group(object)

## S4 method for signature 'dmcResults'
treatment_group(object)

## S4 method for signature '`NULL`'
treatment_group(object)

## S4 method for signature 'RangedSummarizedExperiment'
llk(object)

## S4 method for signature 'betaHMMResults'
llk(object)

## S4 method for signature '`NULL`'
llk(object)

## S4 method for signature 'RangedSummarizedExperiment'
phi(object)

## S4 method for signature 'betaHMMResults'
phi(object)

## S4 method for signature 'threshold_Results'
phi(object)

## S4 method for signature '`NULL`'
phi(object)

## S4 method for signature 'RangedSummarizedExperiment'
hidden_states(object)

## S4 method for signature 'betaHMMResults'
hidden_states(object)

## S4 method for signature 'threshold_Results'
hidden_states(object)

## S4 method for signature '`NULL`'
hidden_states(object)

## S4 method for signature 'RangedSummarizedExperiment'
chromosome_number(object)

## S4 method for signature 'betaHMMResults'
chromosome_number(object)
```

```
## S4 method for signature 'dmrResults'
chromosome_number(object)

## S4 method for signature '`NULL`'
chromosome_number(object)

## S4 method for signature 'RangedSummarizedExperiment'
AUC(object)

## S4 method for signature 'dmcResults'
AUC(object)

## S4 method for signature '`NULL`'
AUC(object)

## S4 method for signature 'RangedSummarizedExperiment'
uncertainty(object)

## S4 method for signature 'dmcResults'
uncertainty(object)

## S4 method for signature '`NULL`'
uncertainty(object)

## S4 method for signature 'RangedSummarizedExperiment'
model_parameters(object)

## S4 method for signature 'threshold_Results'
model_parameters(object)

## S4 method for signature '`NULL`'
model_parameters(object)

## S4 method for signature 'RangedSummarizedExperiment'
threshold(object)

## S4 method for signature 'threshold_Results'
threshold(object)

## S4 method for signature '`NULL`'
threshold(object)
```

## Arguments

**object**                    a betaHMMResults/ dmcResults/ threshold\_Results object.

**Value**

Output varies depending on the method.

**Examples**

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

|                 |                                       |
|-----------------|---------------------------------------|
| annotation_data | <i>MethylationEPIC manifest data.</i> |
|-----------------|---------------------------------------|

---

**Description**

A dataset containing a subset of the manifest data from the Illumina MethylationEPIC beadchip array. A subset of the complete dataset has been uploaded in the package for testing purpose. The complete dataset is available on [GitHub](#).

**Usage**

```
data(annotation_data)
```

**Format**

A data frame with 100 rows and 9 columns.

- IlmnID: The unique identifier from the Illumina CG database, i.e. the probe ID.
- Genome\_Build: The genome build referenced by the Infinium MethylationEPIC manifest.
- CHR: The chromosome containing the CpG (Genome\_Build = 37).
- MAPINFO: The chromosomal coordinates of the CpG sites.
- UCSC\_RefGene\_Name: The target gene name(s), from the UCSC database. Note: multiple listings of the same gene name indicate splice variants.
- UCSC\_RefGene\_Accession: The UCSC accession numbers of the target transcripts. Accession numbers are in the same order as the target gene transcripts.
- UCSC\_RefGene\_Group: Gene region feature category describing the CpG position, from UCSC. Features are listed in the same order as the target gene transcripts.
- UCSC\_CpG\_Islands\_Name: The chromosomal coordinates of the CpG Island from UCSC.
- Relation\_to\_UCSC\_CpG\_Island: The location of the CpG relative to the CpG island.

**Value**

A data frame containing the array design for Illumina's Human Methylation EPIC microarray for the chromosome 7. Based on the v1.0b2 version of the manifest file.

**See Also**

[pca\\_methylation\\_data](#)

---

|                 |                                                                                                    |
|-----------------|----------------------------------------------------------------------------------------------------|
| AUC_DM_analysis | <i>Area under the curve method for calculating dissimilarities between estimated distributions</i> |
|-----------------|----------------------------------------------------------------------------------------------------|

---

**Description**

The function is used to calculate the dissimilarity between the cumulative distributions estimated in each hidden state.

**Usage**

```
AUC_DM_analysis(M, N, R, K, tau, A, phi)
```

**Arguments**

|     |                                                                                                            |
|-----|------------------------------------------------------------------------------------------------------------|
| M   | Number of methylation states to be identified in a single DNA sample.                                      |
| N   | Number of patients or DNA sample replicates collected for each treatment group.                            |
| R   | Number of treatment groups (For. eg: Benign and Tumour).                                                   |
| K   | Number of hidden states estimated.                                                                         |
| tau | The initial distribution for hidden states in the betaHMM model.                                           |
| A   | The transition matrix for hidden states in the betaHMM model.                                              |
| phi | The shape parameters estimated for the observed data and the estimated hidden states in the betaHMM model. |

**Value**

A dataframe returning the hidden state, the AUC value and the distributions which resulted in highest AUC value calculated for the corresponding hidden state.

backward

*Backward algorithm***Description**

Computes the backward probabilities for  $K$  hidden states using the initial distribution, transition matrix and distribution parameters.

**Usage**

```
backward(probabilities, trained_params)
```

**Arguments**

|                |                                                                                                                                                             |
|----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| probabilities  | A matrix containing the marginal probability distribution of the $K$ hidden states within the Markov model.                                                 |
| trained_params | A list containing the initialised shape parameters for the betaHMM model (initial distribution, transition matrix, shape parameters for the observed data). |

**Value**

A list containing:

- log\_beta - A matrix of dimension  $C * K$  (where  $C$  is the number of CpG sites and  $K$  is the number of hidden states) containing the logarithmized backward probabilities.
- scaled\_logL - The log-likelihood calculated using the backward probabilities.

---

BaumWelch

*HMM parameter estimation using Baum-Welch algorithm*


---

### Description

The function determines the parameters of a homogeneous beta hidden Markov model (betaHMM), wherein the Baum-Welch algorithm constitutes a variant of the EM (Estimation-Maximization) procedure.

### Usage

```
BaumWelch(
  data,
  trained_params = NULL,
  K,
  M,
  N,
  R,
  seed = NULL,
  iterations = 100
)
```

### Arguments

|                             |                                                                                                                                                                                                                                                                                                                                                                |
|-----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>data</code>           | A dataframe of dimension $C \times (N \times R)$ containing methylation values for $C$ CpG sites from $R$ treatment groups each having $N$ DNA samples.                                                                                                                                                                                                        |
| <code>trained_params</code> | A list containing the initialized model parameters: <ul style="list-style-type: none"> <li>• <code>A</code> - The transition matrix for the betaHMM model.</li> <li>• <code>tau</code> - The initial distribution for the betaHMM model.</li> <li>• <code>phi</code> - The shape parameters for the observation sequence data in the betaHMM model.</li> </ul> |
| <code>K</code>              | The number of hidden states identified using the betaHMM model.                                                                                                                                                                                                                                                                                                |
| <code>M</code>              | Number of methylation states to be identified in a single DNA sample.                                                                                                                                                                                                                                                                                          |
| <code>N</code>              | Number of DNA samples (patients/replicates) collected for each treatment group.                                                                                                                                                                                                                                                                                |
| <code>R</code>              | Number of treatment groups (For. eg: Benign and Tumour).                                                                                                                                                                                                                                                                                                       |
| <code>seed</code>           | Seed to allow for reproducibility (default = NULL).                                                                                                                                                                                                                                                                                                            |
| <code>iterations</code>     | Number of iterations for algorithm convergence.                                                                                                                                                                                                                                                                                                                |

### Value

A list containing:

- `A` - The transition matrix estimated for the betaHMM model.
- `tau` - The initial distribution estimated for the betaHMM model.

- `phi` - The shape parameters estimated for the observed data in the betaHMM model.
- `log_vec` - A vector containing the log-likelihood values calculated for each iteration of the algorithm.
- `z` - A matrix of dimension  $C \times K$  containing the conditional posterior probability of each CpG site belonging to each of the  $K$  hidden states.

---

betaHMM

HMM for beta valued DNA data

---

## Description

This is the primary user interface for the betaHMM function. Generic S4 methods are implemented to estimate the parameters of a homogeneous hidden Markov model for the beta valued DNA methylation data. The supported classes are `matrix`, `data.frame`, `RangedSummarizedExperiment` and `GRanges`. The output of betaHMM method is an S4 object of class `betaHMMResults`.

## Usage

```
betaHMM(methylation_data, annotation_file, ...)
```

```
## S4 method for signature 'matrix,matrix'
```

```
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)
```

```
## S4 method for signature 'data.frame,data.frame'
```

```
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)
```

```
)

## S4 method for signature
## 'RangedSummarizedExperiment,RangedSummarizedExperiment'
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)

## S4 method for signature 'GRanges,GRanges'
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)

## S4 method for signature 'RangedSummarizedExperiment,matrix'
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)

## S4 method for signature 'RangedSummarizedExperiment,data.frame'
betaHMM(
  methylation_data,
```

```
    annotation_file,
    M = 3,
    N = 4,
    R = 2,
    treatment_group = NULL,
    parallel_process = FALSE,
    seed = NULL,
    iterations = 100,
    ...
)

## S4 method for signature 'RangedSummarizedExperiment,GRanges'
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)

## S4 method for signature 'matrix,RangedSummarizedExperiment'
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)

## S4 method for signature 'data.frame,RangedSummarizedExperiment'
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
```

```
    seed = NULL,
    iterations = 100,
    ...
)

## S4 method for signature 'data.frame,GRanges'
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)

## S4 method for signature 'matrix,data.frame'
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)

## S4 method for signature 'matrix,GRanges'
betaHMM(
  methylation_data,
  annotation_file,
  M = 3,
  N = 4,
  R = 2,
  treatment_group = NULL,
  parallel_process = FALSE,
  seed = NULL,
  iterations = 100,
  ...
)

## S4 method for signature 'GRanges,matrix'
```

```
betaHMM(  
  methylation_data,  
  annotation_file,  
  M = 3,  
  N = 4,  
  R = 2,  
  treatment_group = NULL,  
  parallel_process = FALSE,  
  seed = NULL,  
  iterations = 100,  
  ...  
)  
  
## S4 method for signature 'GRanges,data.frame'  
betaHMM(  
  methylation_data,  
  annotation_file,  
  M = 3,  
  N = 4,  
  R = 2,  
  treatment_group = NULL,  
  parallel_process = FALSE,  
  seed = NULL,  
  iterations = 100,  
  ...  
)  
  
## S4 method for signature 'GRanges,RangedSummarizedExperiment'  
betaHMM(  
  methylation_data,  
  annotation_file,  
  M = 3,  
  N = 4,  
  R = 2,  
  treatment_group = NULL,  
  parallel_process = FALSE,  
  seed = NULL,  
  iterations = 100,  
  ...  
)  
  
## S4 method for signature 'data.frame,matrix'  
betaHMM(  
  methylation_data,  
  annotation_file,  
  M = 3,  
  N = 4,  
  R = 2,
```

```

    treatment_group = NULL,
    parallel_process = FALSE,
    seed = NULL,
    iterations = 100,
    ...
)

```

## Arguments

|                  |                                                                                                                                                                                                                                                                                         |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| methylation_data | A dataframe of dimension $(C \times (N \times R)) + 1$ containing methylation values for $C$ CpG sites from $R$ treatment groups each having $N$ DNA samples and the IllumID for each CpG site. Maybe provided as a matrix or data.frame, GRanges or RangedSummarizedExperiment object. |
| annotation_file  | A dataframe containing the EPIC methylation annotation file. Maybe provided as a matrix or data.frame, GRanges or RangedSummarizedExperiment object.                                                                                                                                    |
| ...              | Extra arguments                                                                                                                                                                                                                                                                         |
| M                | Number of methylation states to be identified in a single DNA sample.                                                                                                                                                                                                                   |
| N                | Number of DNA samples (patients/replicates) collected for each treatment group.                                                                                                                                                                                                         |
| R                | Number of treatment groups (For. eg: Benign and Tumour).                                                                                                                                                                                                                                |
| treatment_group  | The names of each treatment groups/ conditions. If no value is passed then default values of sample names, e.g. Sample 1, Sample 2, etc are used as legend text (default = NULL).                                                                                                       |
| parallel_process | The 'TRUE' option results in parallel processing of the models for each chromosome for increased computational efficiency. The default option has been set as 'FALSE' due to package testing limitations.                                                                               |
| seed             | Seed to allow for reproducibility (default = NULL).                                                                                                                                                                                                                                     |
| iterations       | Number of iterations for algorithm convergence (default=100).                                                                                                                                                                                                                           |

## Value

An S4 object of class `betaHMMResults`, where conditional probabilities of each CpG site belonging to a hidden state is stored as a `SimpleList` of assay data, and the corresponding estimated model parameters, log-likelihood values, and most probable hidden state sequence for each chromosome are stored as metadata.

## Author(s)

Koyel Majumdar

## Examples

```

## Use simulated data for the betaHMM workflow example
set.seed(12345)

```

```
## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

betaHMMResults-class    *betaHMMResults object and constructor*

---

## Description

betaHMMResults is a subclass of RangedSummarizedExperiment, used to store the betaHMM results as well as the annotated data useful for plotting.

## Usage

```
betaHMMResults(SummarizedExperiment, annotatedData)
```

## Arguments

SummarizedExperiment  
                                   a RangedSummarizedExperiment of betaHMM results

annotatedData    The annotated data passed as an input argument to the betaHMM package.

## Details

This constructor function would not typically be used by "end users". This simple class extends the RangedSummarizedExperiment class of the SummarizedExperiment package to allow other packages to write methods for results objects from the betaHMM package. It is used by to wrap up the results table.

**Value**

a betaHMMResults object

**Examples**

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

betaHMMrun

*The betaHMM model parameter estimation function*


---

**Description**

A homogeneous hidden Markov model for the beta valued DNA methylation data.

**Usage**

```
betaHMMrun(
  methylation_data,
  annotation_file,
  M,
  N,
  R,
  treatment_group = NULL,
  parallel_process = FALSE,
```

```

    seed = NULL,
    iterations = 100,
    ...
)

```

## Arguments

|                  |                                                                                                                                                                                                           |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| methylation_data | A dataframe of dimension $(C \times (N \times R)) + 1$ containing methylation values for $C$ CpG sites from $R$ treatment groups each having $N$ DNA samples and the IlmnID for each CpG site.            |
| annotation_file  | A dataframe containing the EPIC methylation annotation file.                                                                                                                                              |
| M                | Number of methylation states to be identified in a single DNA sample.                                                                                                                                     |
| N                | Number of DNA samples (patients/replicates) collected for each treatment group.                                                                                                                           |
| R                | Number of treatment groups (For. eg: Benign and Tumour).                                                                                                                                                  |
| treatment_group  | The names of each treatment groups/ conditions. If no value is passed then default values of sample names, e.g. Sample 1, Sample 2, etc are used as legend text (default = NULL).                         |
| parallel_process | The 'TRUE' option results in parallel processing of the models for each chromosome for increased computational efficiency. The default option has been set as 'FALSE' due to package testing limitations. |
| seed             | Seed to allow for reproducibility (default = NULL).                                                                                                                                                       |
| iterations       | Number of iterations for algorithm convergence (default=100).                                                                                                                                             |
| ...              | Extra arguments                                                                                                                                                                                           |

## Details

The betaHMM function employs initially set parameters (utilizing a basic 3-state beta hidden Markov model) to estimate the parameters of the homogeneous hidden Markov model, adapted for beta-valued DNA methylation data, through implementation of the Baum-Welch algorithm. Subsequently, the derived parameters are utilized to ascertain the most probable sequence of hidden states using the Viterbi algorithm.

## Value

The function returns an object of the [betaHMMResults](#) class which contains a SimpleList of assay data containing the posterior probability of each CpG site belonging to each of the  $K$  hidden states and the following values as metadata:

- K - The number of hidden states identified using the betaHMM model.
- C - The number of CpG sites analysed using the betaHMM model.
- N - The number of DNA samples corresponding to each treatment group analysed using the betaHMM model.

- R - The number of treatment groups analysed using the betaHMM model.
- A - The transition matrix estimated for the betaHMM model.
- tau - The initial distribution estimated for the betaHMM model.
- treatment\_group - The names of the treatment groups/conditions analysed.
- phi - The shape parameters estimated for the observed data in the betaHMM model.
- llk - A vector containing the log-likelihood values calculated for each iteration of the algorithm.
- hidden\_states - The vector containing the estimated hidden states for each CpG sites.

### Examples

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

dmcResults-class

*dmcResults object and constructor*


---

### Description

dmcResults is a subclass of RangedSummarizedExperiment, used to store the DMCs identified.

### Usage

```
dmcResults(SummarizedExperiment)
```

**Arguments**

SummarizedExperiment

a RangedSummarizedExperiment of dmcResults results.

**Details**

This constructor function would not typically be used by "end users". This simple class extends the RangedSummarizedExperiment class of the SummarizedExperiment package to allow other packages to write methods for results objects from the `dmc_identification` function. It is used by to wrap up the results table.

**Value**a `dmcResults` object**Examples**

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

## Description

This is the primary user interface for the `dmc_identification` function. Generic S4 methods are implemented to identify the DMCs from the estimated betaHMM model parameters for each chromosome. The supported class is a `betaHMMResults` object. The output is an S4 object of class `dmcResults`.

## Usage

```
dmc_identification(betaHMM_object, ...)

## S4 method for signature 'betaHMMResults'
dmc_identification(
  betaHMM_object,
  AUC_threshold = 0.8,
  uncertainty_threshold = 0.2,
  ...
)
```

## Arguments

`betaHMM_object` An S4 object of class `betaHMMResults`

`...` extra arguments

`AUC_threshold` The threshold for AUC metric for each chromosome.

`uncertainty_threshold` The threshold for uncertainty of belonging to a particular hidden state, for each chromosome.

## Value

An S4 object of class `dmcResults`.

## See Also

`betaHMM`

## Examples

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))
```

```
## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

dmc\_identification\_run

*DMC identification from estimated betaHMM model parameters*

---

## Description

The dissimilarities between the cumulative distributions calculated for each hidden state are determined through employment of the area-under-curve (AUC) technique. By incorporating user-defined threshold values for AUC alongside the associated uncertainties in membership within that hidden state, the aim is to pinpoint the most distinctively methylated states. This process facilitates the identification of CpGs that exhibit the most notable differential methylation, guided by the predefined threshold criteria.

## Usage

```
dmc_identification_run(
  betaHMM_object,
  AUC_threshold = 0.8,
  uncertainty_threshold = 0.2,
  ...
)
```

## Arguments

**betaHMM\_object** A [betaHMMResults](#) object.

**AUC\_threshold** The threshold for AUC metric for each chromosome.

**uncertainty\_threshold** The threshold for uncertainty of belonging to a particular hidden state, for each chromosome.

**...** extra arguments

**Value**

The function returns an object of the `dmcResults` class which contains a SimpleList of assay data which contains the following values:

- CHR Chromosome number
- MAPINFO Mapinfo
- IlmnID IlmnID
- N\*R columns containing methylation states
- hidden\_state The assigned hidden\_state
- DMC The value is 1 if the CpG is a DMC else 0.

The object contains the following values as the metadata:

- A list containing the AUC values for K hidden states for each chromosome and the conditions compared which resulted in the highest AUC value when more than 2 conditions are compared.
- A list containing the conditional probability values for K hidden states for each chromosome.
- The treatment group labels.
- K The number of hidden states estimated.
- N The number of DNA replicates/patients for each treatment group.
- R The number of treatment groups to be compared.

**Examples**

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

|                  |                                          |
|------------------|------------------------------------------|
| dmrResults-class | <i>dmrResults object and constructor</i> |
|------------------|------------------------------------------|

---

## Description

dmrResults is a subclass of RangedSummarizedExperiment, used to store the DMRs identified.

## Usage

```
dmrResults(SummarizedExperiment)
```

## Arguments

SummarizedExperiment  
a dmrResults results.

## Details

This constructor function would not typically be used by "end users". This simple class extends the RangedSummarizedExperiment class of the SummarizedExperiment package to allow other packages to write methods for results objects from the [dmr\\_identification](#) function. It is used by to wrap up the results table.

## Value

a [dmrResults](#) object

## Examples

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
```

```
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

|                    |                                                |
|--------------------|------------------------------------------------|
| dmr_identification | <i>DMR identification from DMCs identified</i> |
|--------------------|------------------------------------------------|

---

## Description

This is the primary user interface for the [dmr\\_identification](#) function. Generic S4 methods are implemented to identify the DMRs from the DMCs identified in each chromosome. The supported classes are `data.frame` and [dmcResults](#) object. The output is an S4 object of class [dmrResults](#).

## Usage

```
dmr_identification(dmc_identification_object, ...)

## S4 method for signature 'dmcResults'
dmr_identification(dmc_identification_object, DMC_count = 2, ...)

## S4 method for signature 'matrix'
dmr_identification(dmc_identification_object, DMC_count = 2, ...)

## S4 method for signature 'data.frame'
dmr_identification(dmc_identification_object, DMC_count = 2, ...)
```

## Arguments

|                                        |                                                                                          |
|----------------------------------------|------------------------------------------------------------------------------------------|
| <code>dmc_identification_object</code> | a <a href="#">dmcResults</a> object or the assay data from the <code>dmcResults</code> . |
| <code>...</code>                       | extra arguments                                                                          |
| <code>DMC_count</code>                 | The minimal number of consecutive CpGs in a DMR.                                         |

## Value

An S4 object of class [dmrResults](#) where the CpG site information for each DMR is stored as a `SimpleList` of assay data and the chromosomes analysed by the model is stored as the metadata.

## See Also

[betaHMM](#)

**Examples**

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

dmr\_identification\_run

*DMR identification from DMCs identified*


---

**Description**

Function to identify the DMRs from the DMCs identified in each chromosome.

**Usage**

```
dmr_identification_run(
  dmc_identification_object,
  DMC_count = 2,
  parallel_process = FALSE,
  ...
)
```

**Arguments**

dmc\_identification\_object  
a dmcResults object or the assay data from the dmcResults.

DMC\_count        The minimal number of consecutive CpGs in a DMR.

parallel\_process        The 'TRUE' option results in parallel processing of the DMCs from each chromosome for increased computational efficiency. The default option has been set as 'FALSE' due to package testing limitations.

...                extra arguments

### Value

A [dmrResults](#) object containing a SimpleList of assay data containing the following data:

- start\_CpG - The starting CpG site IlmnID in the particular DMR
- end\_CpG - The ending CpG site IlmnID in the particular DMR
- DMR\_size - Number of CPG sites identified in the DMR
- chr\_dmr - The chromosome corresponding to the CpG sites in the DMR.
- map\_start - MAPINFO of starting CpG site in the particular DMR
- map\_end - MAPINFO of the ending CpG site in the particular DMR

The object also returns the chromosomes analysed by the betaHMM model as the metadata.

### Examples

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

|         |                          |
|---------|--------------------------|
| forward | <i>Forward algorithm</i> |
|---------|--------------------------|

---

**Description**

Computes the forward probabilities for  $K$  hidden states using the initial distribution, transition matrix and distribution parameters.

**Usage**

```
forward(probabilities, trained_params)
```

**Arguments**

|                |                                                                                                                                                             |
|----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| probabilities  | A matrix containing the marginal probability distribution of the $K$ states within the Markov model.                                                        |
| trained_params | A list containing the initialised shape parameters for the betaHMM model (initial distribution, transition matrix, shape parameters for the observed data). |

**Value**

A list containing:

- log\_alpha - A  $C \times K$  matrix (where  $C$  is the number of CpG sites and  $K$  is the number of hidden states) containing the logarithmized forward probabilities.
- scaled\_logL - The log-likelihood calculated using the forward probabilities.

---

|                       |                                                  |
|-----------------------|--------------------------------------------------|
| initialise_parameters | <i>Initialising the betaHMM model parameters</i> |
|-----------------------|--------------------------------------------------|

---

**Description**

Initialise the betaHMM model parameters.

**Usage**

```
initialise_parameters(data, M, N, R, seed = NULL)
```

**Arguments**

|      |                                                                                                                                                                                            |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| data | A dataframe of dimension $C \times (N \times R)$ containing methylation values for $C$ CpG sites from $R$ treatment groups each having $N$ replicates or each collected from $N$ patients. |
| M    | Number of methylation states to be identified in a DNA sample.                                                                                                                             |
| N    | Number of patients or DNA sample replicates collected for each treatment group.                                                                                                            |
| R    | Number of treatment groups (For. eg: Benign and Tumour).                                                                                                                                   |
| seed | Seed to allow for reproducibility (default = NULL).                                                                                                                                        |

**Details**

Computes the shape parameters using the `threshold_function` to initialise the shape parameters of the model.

**Value**

A list containing:

- A - The transition matrix for the betaHMM model.
- tau - The initial distribution for the betaHMM model.
- phi - The shape parameters for the observation sequence data in the betaHMM model.

---

pca\_methylation\_data    *Simulated DNA methylation data*

---

**Description**

A subset of the dataset containing beta methylation values from  $R = 2$  sample types (Benign and Tumour), collected from  $N = 4$  patients from the a prostate cancer study. The dataset contains methylation values corresponding to chromosome 7.

**Usage**

```
data(pca_methylation_data)
```

**Format**

A data frame with 38672 rows and 9 columns. The data contain no missing values.

- IlmnID: The unique identifier from the Illumina CG database, i.e. the probe ID.
- Benign\_Patient\_1: Methylation values from benign tissue from patient 1.
- Benign\_Patient\_2: Methylation values from benign tissue from patient 2.
- Benign\_Patient\_3: Methylation values from benign tissue from patient 3.
- Benign\_Patient\_4: Methylation values from benign tissue from patient 4.
- Tumour\_Patient\_1: Methylation values from tumor tissue from patient 1.
- Tumour\_Patient\_2: Methylation values from tumor tissue from patient 2.
- Tumour\_Patient\_3: Methylation values from tumor tissue from patient 3.
- Tumour\_Patient\_4: Methylation values from tumor tissue from patient 4.

**Details**

The array data were then normalized and probes located outside of CpG sites and on the sex chromosome were filtered out. The CpG sites with missing values were removed from the resulting dataset. A subset of the complete dataset has been uploaded in the package for testing purposes. The complete dataset is available on [GitHub](#).

**Value**

A data frame containing a subset of methylation data from real study.

**See Also**

[annotation\\_data](#)

---

|      |                                                                                                |
|------|------------------------------------------------------------------------------------------------|
| plot | <i>Visualize results from betaHMM/ dmc_identification/ threshold_identification functions.</i> |
|------|------------------------------------------------------------------------------------------------|

---

**Description**

Visualize results from betaHMM/ dmc\_identification/ threshold\_identification functions.

**Usage**

```
plot(x, ...)
```

## S4 method for signature 'betaHMMResults'

```
plot(
  x,
  chromosome = NULL,
  what = c("fitted density", "kernel density", "uncertainty"),
  treatment_group = NULL,
  AUC = NULL,
  uncertainty_threshold = 0.2,
  title = NULL,
  ...
)
```

## S4 method for signature 'dmcResults'

```
plot(
  x,
  start_CpG = NULL,
  end_CpG = NULL,
  treatment_group = NULL,
  N = NULL,
  title = NULL,
  ...
)
```

## S4 method for signature 'threshold\_Results'

```
plot(x, plot_threshold = TRUE, title = NULL, ...)
```

**Arguments**

|                                    |                                                                                                                                                                                          |
|------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>x</code>                     | An object of class <code>betaHMMResults/ dmcResults/ threshold_Results</code> object.                                                                                                    |
| <code>...</code>                   | Other graphics parameters.                                                                                                                                                               |
| <code>chromosome</code>            | The chromosome number for which the plot is to be displayed.                                                                                                                             |
| <code>what</code>                  | The different plots that can be obtained are either 'fitted density', 'kernel density' or 'uncertainty' (default = 'fitted density').                                                    |
| <code>treatment_group</code>       | The names of the different treatment groups to be displayed in the plot. If no value is passed then the sample names estimated by the <code>betaHMM</code> function are used.            |
| <code>AUC</code>                   | The AUC values for that chromosome.                                                                                                                                                      |
| <code>uncertainty_threshold</code> | The uncertainty threshold value used for DMC identification.                                                                                                                             |
| <code>title</code>                 | The title that the user wants to display. If no title is to be displayed the default is 'NULL'.                                                                                          |
| <code>start_CpG</code>             | The IlmnID of starting CpG site when plotting the DMCs.                                                                                                                                  |
| <code>end_CpG</code>               | The IlmnID of ending CpG site/ the total number of CpGs to be plotted excluding the starting CpG site when plotting the DMCs.                                                            |
| <code>N</code>                     | The number of DNA samples corresponding to each treatment group analysed using the <code>betaHMM</code> model. If 'NULL', the value from <code>betaHMMResults</code> object is selected. |
| <code>plot_threshold</code>        | The "TRUE" option displays the threshold points in the graph for the 3 state <code>betaHMM</code> model (default = "FALSE").                                                             |

**Value**

This function displays the following plots as requested by the user when analysing the `betaHMMResults` output:

- fitted density estimates - Plot showing the fitted density estimates of the clustering solution under the optimal model selected.
- kernel density estimates - Plot showing the kernel density estimates of the clustering solution under the optimal model selected.
- uncertainty - A boxplot showing the uncertainties in the hidden state estimation.

The function displays the DMCs and DMRs plot from the `dmcResults` object.

The function displays the plot for the estimated shape parameters and threshold for the methylation states in a single DNA treatment condition from the `threshold_Results` object.

**Author(s)**

Koyel Majumdar

**Examples**

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")

## Use simulated data for the betaHMM workflow example
set.seed(12345)

## read files
data(sample_methylation_file)
data(sample_annotation_file)
# Run betaHMM function
beta_out <- betaHMM(sample_methylation_file[1:50,],
                    sample_annotation_file[1:50,],
                    M = 3, N = 4, R = 2, iterations=2,
                    parallel_process = FALSE, seed = 12345,
                    treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")
```

---

`sample_annotation_file`*MethylationEPIC manifest data.*

---

## Description

A dataset containing a subset of the manifest data from the Illumina MethylationEPIC beadchip array. A subset of the complete dataset has been uploaded in the package for testing purpose. The complete dataset is available on [GitHub](#).

## Usage

```
data(sample_annotation_file)
```

## Format

A data frame with 100 rows and 9 columns.

- `IlmnID`: The unique identifier from the Illumina CG database, i.e. the probe ID.
- `Genome_Build`: The genome build referenced by the Infinium MethylationEPIC manifest.
- `CHR`: The chromosome containing the CpG (`Genome_Build = 37`).
- `MAPINFO`: The chromosomal coordinates of the CpG sites.
- `UCSC_RefGene_Name`: The target gene name(s), from the UCSC database. Note: multiple listings of the same gene name indicate splice variants.
- `UCSC_RefGene_Accession`: The UCSC accession numbers of the target transcripts. Accession numbers are in the same order as the target gene transcripts.
- `UCSC_RefGene_Group`: Gene region feature category describing the CpG position, from UCSC. Features are listed in the same order as the target gene transcripts.
- `UCSC_CpG_Islands_Name`: The chromosomal coordinates of the CpG Island from UCSC.
- `Relation_to_UCSC_CpG_Island`: The location of the CpG relative to the CpG island.

## Value

A data frame containing the array design for Illumina's Human Methylation EPIC microarray for the simulated CpG sites. Based on the v1.0b2 version of the manifest file.

## See Also

[sample\\_methylation\\_file](#)

---

`sample_methylation_file`*Simulated DNA methylation data*

---

**Description**

A dataset containing simulated beta methylation values from  $R = 2$  sample types (Benign and Tumour), collected from  $N = 4$  patients.

**Usage**

```
data(sample_methylation_file)
```

**Format**

A data frame with 100 rows and 9 columns. The data contain no missing values.

- IlmnID: The unique identifier from the Illumina CG database, i.e. the probe ID.
- Benign\_Patient\_1: Methylation values from benign tissue from patient 1.
- Benign\_Patient\_2: Methylation values from benign tissue from patient 2.
- Benign\_Patient\_3: Methylation values from benign tissue from patient 3.
- Benign\_Patient\_4: Methylation values from benign tissue from patient 4.
- Tumour\_Patient\_1: Methylation values from tumor tissue from patient 1.
- Tumour\_Patient\_2: Methylation values from tumor tissue from patient 2.
- Tumour\_Patient\_3: Methylation values from tumor tissue from patient 3.
- Tumour\_Patient\_4: Methylation values from tumor tissue from patient 4.

**Details**

The array data were then normalized and probes located outside of CpG sites and on the sex chromosome were filtered out. The CpG sites with missing values were removed from the resulting dataset. A subset of the complete dataset has been uploaded in the package for testing purposes. The complete dataset is available on [GitHub](#).

**Value**

The data frame containing simulated methylation values.

**See Also**

[sample\\_annotation\\_file](#)

---

**summary***Summarize results from the three betaHMM workflow functions*

---

**Description**

A function to summarize the `betaHMMResults`, `dmcResults` or `dmrResults` objects.

**Usage**

```
summary(object, ...)  
  
## S4 method for signature 'betaHMMResults'  
summary(object, ...)  
  
## S4 method for signature 'dmcResults'  
summary(object, ...)  
  
## S4 method for signature 'dmrResults'  
summary(object, ...)
```

**Arguments**

|                     |                                                                      |
|---------------------|----------------------------------------------------------------------|
| <code>object</code> | An object of class 'betaHMMResults' or 'dmcResults' or 'dmrResults'. |
| <code>...</code>    | Additional arguments                                                 |

**Value**

Summary of the 'betaHMMResults' or 'dmcResults' or 'dmrResults' object.

**Author(s)**

Koyel majumdar

**See Also**

[betaHMM](#), [dmc\\_identification](#), [dmr\\_identification](#)

**Examples**

```
## Use simulated data for the betaHMM workflow example  
set.seed(12345)  
  
## read files  
data(sample_methylation_file)  
data(sample_annotation_file)  
# Run betaHMM function  
beta_out <- betaHMM(sample_methylation_file[1:50,],  
                    sample_annotation_file[1:50,],
```

```

M = 3, N = 4, R = 2, iterations=2,
parallel_process = FALSE, seed = 12345,
treatment_group = c("Benign", "Tumour"))

## Run dmc_identification function
dmc_out <- dmc_identification(beta_out)

# Run dmr_identification function
dmr_out <- dmr_identification(dmc_out, parallel_process = FALSE)

# Plot functions
# Get the AUC values calculated for each hidden state
AUC_chr <- AUC(dmc_out)

## plot the uncertainty for each hidden state
plot(beta_out, chromosome = "1", what = "uncertainty")

```

---

threshold\_identification

*HMM for beta valued DNA data for a single treatment condition*

---

## Description

The supported classes are matrix and data.frame. The output of threshold\_identification is an S4 object of class threshold\_Results.

## Usage

```

threshold_identification(object1, ...)

## S4 method for signature 'matrix'
threshold_identification(
  object1,
  package_workflow = TRUE,
  annotation_file = NULL,
  M = 3,
  N = 4,
  parameter_estimation_only = FALSE,
  seed = NULL,
  ...
)

## S4 method for signature 'data.frame'
threshold_identification(
  object1,
  package_workflow = TRUE,
  annotation_file = NULL,

```

```

M = 3,
N = 4,
parameter_estimation_only = FALSE,
seed = NULL,
...
)

```

### Arguments

|                           |                                                                                                                                                   |
|---------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| object1                   | Methylation data and IlmnID. Maybe provided as a matrix or dataframe.                                                                             |
| ...                       | Extra parameters.                                                                                                                                 |
| package_workflow          | Flag set to TRUE if method called from package workflow. If set to FALSE then the parameter annotation_file needs to be supplied to the function. |
| annotation_file           | A dataframe containing the EPIC methylation annotation file.                                                                                      |
| M                         | Number of methylation states to be identified in a single DNA sample.                                                                             |
| N                         | Number of DNA samples (patients/replicates) collected for each treatment group.                                                                   |
| parameter_estimation_only | If only model parameters are to be estimated then value is TRUE else FALSE.                                                                       |
| seed                      | Seed to allow for reproducibility (default = NULL).                                                                                               |

### Value

An S4 object of class threshold\_Results.

### Examples

```

## Use simulated data for the betaHMM workflow example
set.seed(12345)
library(betaHMM)

## read files
data(sample_methylation_file)
head(sample_methylation_file)
data(sample_annotation_file)
head(sample_annotation_file)
##merge data
df=merge(sample_annotation_file[,c('IlmnID','CHR','MAPINFO')],
sample_methylation_file,by='IlmnID')

## sort data
df=df[order(df$CHR,df$MAPINFO),]
thr_out=threshold_identification(df[,c(1,4:7)],package_workflow=TRUE,M=3,4,
parameter_estimation_only=TRUE,seed=12345)

```

---

```
threshold_identification_run
```

*Threshold identification function*

---

## Description

HMM for beta valued DNA data for a single treatment condition

## Usage

```
threshold_identification_run(
  data,
  package_workflow = TRUE,
  annotation_file = NULL,
  M,
  N,
  parameter_estimation_only = FALSE,
  seed = NULL,
  ...
)
```

## Arguments

|                                        |                                                                                                                                                                |
|----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>data</code>                      | Methylation data and IlmnID. Maybe provided as a matrix or dataframe.                                                                                          |
| <code>package_workflow</code>          | Flag set to TRUE if method called from package workflow. If set to FALSE then the parameter <code>annotation_file</code> needs to be supplied to the function. |
| <code>annotation_file</code>           | A dataframe containing the EPIC methylation annotation file.                                                                                                   |
| <code>M</code>                         | Number of methylation states to be identified in a single DNA sample.                                                                                          |
| <code>N</code>                         | Number of DNA samples (patients/replicates) collected for each treatment group.                                                                                |
| <code>parameter_estimation_only</code> | If only model parameters are to be estimated then value is TRUE else FALSE.                                                                                    |
| <code>seed</code>                      | Seed to allow for reproducibility (default = NULL).                                                                                                            |
| <code>...</code>                       | Extra parameters.                                                                                                                                              |

## Value

An S4 object of class `threshold_Results`, where conditional probabilities of each CpG site belonging to a one of the  $M$  methylation states is stored as a SimpleList of assay data, and the corresponding estimated model parameters, the thresholds and most probable hidden state sequence for each chromosome are stored as metadata.

## Examples

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)
library(betaHMM)

## read files
data(sample_methylation_file)
head(sample_methylation_file)
data(sample_annotation_file)
head(sample_annotation_file)
##merge data
df=merge(sample_annotation_file[,c('IlmnID', 'CHR', 'MAPINFO')],
sample_methylation_file,by='IlmnID')

## sort data
df=df[order(df$CHR,df$MAPINFO),]
thr_out=threshold_identification(df[,c(1,4:7)],package_workflow=TRUE,M=3,4,
parameter_estimation_only=TRUE,seed=12345)
```

---

threshold\_Results-class

*threshold\_Results object and constructor*

---

## Description

threshold\_Results is a subclass of RangedSummarizedExperiment, used to store the threshold\_identification results as well as the annotated data useful for plotting.

## Usage

```
threshold_Results(SummarizedExperiment, annotatedData)
```

## Arguments

|                      |                                                                                          |
|----------------------|------------------------------------------------------------------------------------------|
| SummarizedExperiment | a RangedSummarizedExperiment of threshold_Results object.                                |
| annotatedData        | The annotated data passed as an input argument to the threshold_identification function. |

## Details

This constructor function would not typically be used by "end users". This simple class extends the RangedSummarizedExperiment class of the SummarizedExperiment package to allow other packages to write methods for results objects from the [threshold\\_identification](#) function. It is used by to wrap up the results table.

**Value**

a `threshold_Results` object

**Examples**

```
## Use simulated data for the betaHMM workflow example
set.seed(12345)
library(betaHMM)

## read files
data(sample_methylation_file)
head(sample_methylation_file)
data(sample_annotation_file)
head(sample_annotation_file)
##merge data
df=merge(sample_annotation_file[,c('IlmnID', 'CHR', 'MAPINFO')],
sample_methylation_file,by='IlmnID')

## sort data
df=df[order(df$CHR,df$MAPINFO),]
thr_out=threshold_identification(df[,c(1,4:7)],package_workflow=TRUE,M=3,4,
parameter_estimation_only=TRUE,seed=12345)
```

---

Viterbi

---

*Estimating the hidden states using Viterbi algorithm*


---

**Description**

The Viterbi algorithm is used to estimate the most probable sequence of the hidden states utilizing the betaHMM model parameters estimated by the Baum-Welch algorithm.

**Usage**

```
Viterbi(data, M, N, R, tau, A, phi, K)
```

**Arguments**

|      |                                                                                                                                                                                            |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| data | A dataframe of dimension $C \times (N \times R)$ containing methylation values for $C$ CpG sites from $R$ treatment groups each having $N$ replicates or each collected from $N$ patients. |
| M    | Number of methylation states to be identified in a single DNA sample.                                                                                                                      |
| N    | Number of DNA samples (patients/replicates) collected for each treatment group.                                                                                                            |
| R    | Number of treatment groups (For. eg: Benign and Tumour).                                                                                                                                   |
| tau  | The initial distribution for the betaHMM model.                                                                                                                                            |
| A    | The transition matrix for the betaHMM model.                                                                                                                                               |
| phi  | The shape parameters for the observation sequence data in the betaHMM model.                                                                                                               |
| K    | The number of hidden states identified using the betaHMM model.                                                                                                                            |

**Value**

A vector containing the most probable sequence of the hidden states of the betaHMM model.

# Index

- \* **datasets**
  - annotation\_data, [7](#)
  - pca\_methylation\_data, [30](#)
  - sample\_annotation\_file, [34](#)
  - sample\_methylation\_file, [35](#)
- \* **internal**
  - AUC\_DM\_analysis, [8](#)
  - backward, [9](#)
  - BaumWelch, [10](#)
  - forward, [29](#)
  - initialise\_parameters, [29](#)
  - Viterbi, [41](#)
- \* **methods**
  - betaHMM, [11](#)
  - summary, [36](#)
- A (annotatedData), [2](#)
- A, betaHMMResults-method
  - (annotatedData), [2](#)
- A, NULL-method (annotatedData), [2](#)
- A, RangedSummarizedExperiment-method
  - (annotatedData), [2](#)
- annotatedData, [2](#)
- annotatedData, betaHMMResults-method
  - (annotatedData), [2](#)
- annotatedData, threshold\_Results-method
  - (annotatedData), [2](#)
- annotation\_data, [7](#), [31](#)
- AUC (annotatedData), [2](#)
- AUC, dmcResults-method (annotatedData), [2](#)
- AUC, NULL-method (annotatedData), [2](#)
- AUC, RangedSummarizedExperiment-method
  - (annotatedData), [2](#)
- AUC\_DM\_analysis, [8](#)
- backward, [9](#)
- BaumWelch, [10](#)
- betaHMM, [11](#), [22](#), [26](#), [36](#)
- betaHMM, data.frame, data.frame-method
  - (betaHMM), [11](#)

- betaHMM, data.frame, GRanges-method
  - (betaHMM), [11](#)
- betaHMM, data.frame, matrix-method
  - (betaHMM), [11](#)
- betaHMM, data.frame, RangedSummarizedExperiment-method
  - (betaHMM), [11](#)
- betaHMM, data.frame, RangedSummarizedExperiment.-method
  - (betaHMM), [11](#)
- betaHMM, GRanges, data.frame-method
  - (betaHMM), [11](#)
- betaHMM, GRanges, GRanges-method
  - (betaHMM), [11](#)
- betaHMM, GRanges, matrix-method
  - (betaHMM), [11](#)
- betaHMM, GRanges, RangedSummarizedExperiment-method
  - (betaHMM), [11](#)
- betaHMM, matrix, data.frame-method
  - (betaHMM), [11](#)
- betaHMM, matrix, GRanges-method
  - (betaHMM), [11](#)
- betaHMM, matrix, matrix-method (betaHMM),  
[11](#)
- betaHMM, matrix, RangedSummarizedExperiment-method
  - (betaHMM), [11](#)
- betaHMM, matrix, RangedSummarizedExperiment.-method
  - (betaHMM), [11](#)
- betaHMM, RangedSummarizedExperiment, data.frame-method
  - (betaHMM), [11](#)
- betaHMM, RangedSummarizedExperiment, GRanges-method
  - (betaHMM), [11](#)
- betaHMM, RangedSummarizedExperiment, matrix-method
  - (betaHMM), [11](#)
- betaHMM, RangedSummarizedExperiment, RangedSummarizedExperiment-method
  - (betaHMM), [11](#)
- betaHMM, RangedSummarizedExperiment, RangedSummarizedExperiment-method
  - (betaHMM), [11](#)
- betaHMM-methods (betaHMM), [11](#)
- betaHMMResults, [19](#), [22](#), [23](#), [32](#)
- betaHMMResults (betaHMMResults-class),

- 17
- betaHMMResults-class, 17
- betaHMMrun, 18
- chromosome\_number (annotatedData), 2
- chromosome\_number, betaHMMResults-method (annotatedData), 2
- chromosome\_number, dmrResults-method (annotatedData), 2
- chromosome\_number, NULL-method (annotatedData), 2
- chromosome\_number, RangedSummarizedExperiment-method (annotatedData), 2
- dmc\_identification, 21, 21, 22, 36
- dmc\_identification, betaHMMResults-method (dmc\_identification), 21
- dmc\_identification-methods (dmc\_identification), 21
- dmc\_identification\_run, 23
- dmcResults, 21, 22, 24, 26, 32
- dmcResults (dmcResults-class), 20
- dmcResults-class, 20
- dmr\_identification, 25, 26, 26, 36
- dmr\_identification, data.frame-method (dmr\_identification), 26
- dmr\_identification, dmcResults-method (dmr\_identification), 26
- dmr\_identification, matrix-method (dmr\_identification), 26
- dmr\_identification-methods (dmr\_identification), 26
- dmr\_identification\_run, 27
- dmrResults, 25, 26, 28
- dmrResults (dmrResults-class), 25
- dmrResults-class, 25
- forward, 29
- hidden\_states (annotatedData), 2
- hidden\_states, betaHMMResults-method (annotatedData), 2
- hidden\_states, NULL-method (annotatedData), 2
- hidden\_states, RangedSummarizedExperiment-method (annotatedData), 2
- hidden\_states, threshold\_Results-method (annotatedData), 2
- initialise\_parameters, 29
- K (annotatedData), 2
- K, betaHMMResults-method (annotatedData), 2
- K, dmcResults-method (annotatedData), 2
- K, NULL-method (annotatedData), 2
- K, RangedSummarizedExperiment-method (annotatedData), 2
- K, threshold\_Results-method (annotatedData), 2
- llk (annotatedData), 2
- llk, betaHMMResults-method (annotatedData), 2
- llk, NULL-method (annotatedData), 2
- llk, RangedSummarizedExperiment-method (annotatedData), 2
- model\_parameters (annotatedData), 2
- model\_parameters, NULL-method (annotatedData), 2
- model\_parameters, RangedSummarizedExperiment-method (annotatedData), 2
- model\_parameters, threshold\_Results-method (annotatedData), 2
- N (annotatedData), 2
- N, betaHMMResults-method (annotatedData), 2
- N, dmcResults-method (annotatedData), 2
- N, NULL-method (annotatedData), 2
- N, RangedSummarizedExperiment-method (annotatedData), 2
- pca\_methylation\_data, 8, 30
- phi (annotatedData), 2
- phi, betaHMMResults-method (annotatedData), 2
- phi, NULL-method (annotatedData), 2
- phi, RangedSummarizedExperiment-method (annotatedData), 2
- phi, threshold\_Results-method (annotatedData), 2
- plot, 31
- plot, betaHMMResults-method (plot), 31
- plot, dmcResults-method (plot), 31
- plot, threshold\_Results-method (plot), 31
- plot-methods (plot), 31
- R (annotatedData), 2

R, betaHMMResults-method  
     (annotatedData), 2  
 R, dmcResults-method (annotatedData), 2  
 R, NULL-method (annotatedData), 2  
 R, RangedSummarizedExperiment-method  
     (annotatedData), 2  
  
 sample\_annotation\_file, 34, 35  
 sample\_methylation\_file, 34, 35  
 summary, 36  
 summary, betaHMMResults-method  
     (summary), 36  
 summary, dmcResults-method (summary), 36  
 summary, dmrResults-method (summary), 36  
 summary-methods (summary), 36  
  
 tau (annotatedData), 2  
 tau, betaHMMResults-method  
     (annotatedData), 2  
 tau, NULL-method (annotatedData), 2  
 tau, RangedSummarizedExperiment-method  
     (annotatedData), 2  
 threshold (annotatedData), 2  
 threshold, NULL-method (annotatedData), 2  
 threshold, RangedSummarizedExperiment-method  
     (annotatedData), 2  
 threshold, threshold\_Results-method  
     (annotatedData), 2  
 threshold\_identification, 37, 40  
 threshold\_identification, data.frame-method  
     (threshold\_identification), 37  
 threshold\_identification, matrix-method  
     (threshold\_identification), 37  
 threshold\_identification-methods  
     (threshold\_identification), 37  
 threshold\_identification\_run, 39  
 threshold\_Results, 32, 41  
 threshold\_Results  
     (threshold\_Results-class), 40  
 threshold\_Results-class, 40  
 treatment\_group (annotatedData), 2  
 treatment\_group, betaHMMResults-method  
     (annotatedData), 2  
 treatment\_group, dmcResults-method  
     (annotatedData), 2  
 treatment\_group, NULL-method  
     (annotatedData), 2  
 treatment\_group, RangedSummarizedExperiment-method  
     (annotatedData), 2  
  
 uncertainty (annotatedData), 2  
 uncertainty, dmcResults-method  
     (annotatedData), 2  
 uncertainty, NULL-method  
     (annotatedData), 2  
 uncertainty, RangedSummarizedExperiment-method  
     (annotatedData), 2  
  
 Viterbi, 41