

# Package ‘MethylMix’

January 1, 2025

**Title** MethylMix: Identifying methylation driven cancer genes

**Version** 2.37.0

**Description** MethylMix is an algorithm implemented to identify hyper and hypomethylated genes for a disease. MethylMix is based on a beta mixture model to identify methylation states and compares them with the normal DNA methylation state. MethylMix uses a novel statistic, the Differential Methylation value or DM-value defined as the difference of a methylation state with the normal methylation state. Finally, matched gene expression data is used to identify, besides differential, functional methylation states by focusing on methylation changes that effect gene expression. References:

Gevaert O. MethylMix: an R package for identifying DNA methylation-driven genes. *Bioinformatics* (Oxford, England). 2015;31(11):1839-41. doi:10.1093/bioinformatics/btv020.

Gevaert O, Tibshirani R, Plevritis SK. Pancancer analysis of DNA methylation-driven genes using MethylMix. *Genome Biology*. 2015;16(1):17. doi:10.1186/s13059-014-0579-8.

**Depends** R (>= 3.2.0)

**License** GPL-2

**Encoding** UTF-8

**LazyData** true

**Author** Olivier Gevaert

**Maintainer** Olivier Gevaert <olivier.gevaert@gmail.com>

**Type** Package

**Date** 2018-07-13

**Imports** foreach, RPMM, RColorBrewer, ggplot2, RCurl, impute, data.table, limma, R.matlab, digest

**Suggests** BiocStyle, doParallel, testthat, knitr, rmarkdown

**biocViews**

DNAMethylation,StatisticalMethod,DifferentialMethylation,GeneRegulation,GeneExpression,MethylationArray,Different

**RoxygenNote** 6.0.1

**VignetteBuilder** knitr

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

**git\_branch** devel

**git\_last\_commit** 6007130

**git\_last\_commit\_date** 2024-10-29

**Repository** Bioconductor 3.21

**Date/Publication** 2025-01-01

## Contents

BatchData . . . . .	3
betaEst_2 . . . . .	3
blc_2 . . . . .	4
ClusterProbes . . . . .	4
ComBat_NoFiles . . . . .	5
combineForEachOutput . . . . .	6
Download_DNAMethylation . . . . .	7
Download_GeneExpression . . . . .	8
GEcancer . . . . .	9
GetData . . . . .	10
get_firehoseData . . . . .	11
METcancer . . . . .	12
MethylMix . . . . .	13
MethylMix_MixtureModel . . . . .	14
MethylMix_ModelGeneExpression . . . . .	15
MethylMix_ModelSingleGene . . . . .	16
MethylMix_PlotModel . . . . .	17
MethylMix_Predict . . . . .	19
MethylMix_RemoveFlipOver . . . . .	20
METnormal . . . . .	20
predictOneGene . . . . .	21
Preprocess_CancerSite_Methylation27k . . . . .	22
Preprocess_CancerSite_Methylation450k . . . . .	22
Preprocess_DNAMethylation . . . . .	23
Preprocess_GeneExpression . . . . .	24
Preprocess_MAdata_Cancer . . . . .	25
Preprocess_MAdata_Normal . . . . .	26
ProbeAnnotation . . . . .	27
SNPprobes . . . . .	27
TCGA_BatchCorrection_MolecularData . . . . .	27
TCGA_GENERIC_BatchCorrection . . . . .	28
TCGA_GENERIC_CheckBatchEffect . . . . .	28
TCGA_GENERIC_CleanUpSampleNames . . . . .	29
TCGA_GENERIC_GetSampleGroups . . . . .	29
TCGA_GENERIC_LoadIlluminaMethylationData . . . . .	30
TCGA_GENERIC_MergeData . . . . .	30

BatchData

TCGA\_GENERIC\_MET\_ClusterProbes\_Helper\_ClusterGenes\_with\_hclust

TCGA\_Load\_MolecularData

TCGA\_Process\_EstimateMissingValues

Index

3

31

31

32

33

BatchData	<i>BatchData data set</i>
-----------	---------------------------

Description

Data set with batch number for TCGA samples.

betaEst_2	<i>The betaEst_2 function</i>
-----------	-------------------------------

Description

Internal. Estimates a beta distribution via Maximum Likelihood. Adapted from RPMM package.

Usage

betaEst\_2(Y, w, weights)

Arguments

- Y

data vector.
- w

posterior weights.
- weights

Case weights.

Value

(a,b) parameters.

---

blc_2	<i>The blc_2 function</i>
-------	---------------------------

---

**Description**

Internal. Fits a beta mixture model for any number of classes. Adapted from RPMM package.

**Usage**

```
blc_2(Y, w, maxiter = 25, tol = 1e-06, weights = NULL, verbose = TRUE)
```

**Arguments**

Y	Data matrix (n x j) on which to perform clustering.
w	Initial weight matrix (n x k) representing classification.
maxiter	Maximum number of EM iterations.
tol	Convergence tolerance.
weights	Case weights.
verbose	Verbose output.

**Value**

A list of parameters representing mixture model fit, including posterior weights and log-likelihood.

---

ClusterProbes	<i>The ClusterProbes function</i>
---------------	-----------------------------------

---

**Description**

This function uses the annotation for Illumina methylation arrays to map each probe to a gene. Then, for each gene, it clusters all its CpG sites using hierarchical clustering and Pearson correlation as distance and complete linkage. If data for normal samples is provided, only overlapping probes between cancer and normal samples are used. Probes with SNPs are removed. This function is prepared to run in parallel if the user registers a parallel structure, otherwise it runs sequentially. This function also cleans up the sample names, converting them to the 12 digit format.

**Usage**

```
ClusterProbes(MET_Cancer, MET_Normal, CorThreshold = 0.4)
```

**Arguments**

MET_Cancer	data matrix for cancer samples.
MET_Normal	data matrix for normal samples.
CorThreshold	correlation threshold for cutting the clusters.

**Value**

List with the clustered data sets and the mapping between probes and genes.

**Examples**

```
## Not run:

# Optional register cluster to run in parallel
library(doParallel)
cl <- makeCluster(5)
registerDoParallel(cl)

# Methylation data for ovarian cancer
cancerSite <- "OV"
targetDirectory <- paste0(getwd(), "/")

# Downloading methylation data
METdirectories <- Download_DNAMethylation(cancerSite, targetDirectory, TRUE)

# Processing methylation data
METProcessedData <- Preprocess_DNAMethylation(cancerSite, METdirectories)

# Saving methylation processed data
saveRDS(METProcessedData, file = paste0(targetDirectory, "MET_", cancerSite, "_Processed.rds"))

# Clustering methylation data
res <- ClusterProbes(METProcessedData[[1]], METProcessedData[[2]])

# Saving methylation clustered data
toSave <- list(METcancer = res[[1]], METnormal = res[[2]], ProbeMapping = res$ProbeMapping)
saveRDS(toSave, file = paste0(targetDirectory, "MET_", cancerSite, "_Clustered.rds"))

stopCluster(cl)

## End(Not run)
```

---

ComBat\_NoFiles*The ComBat\_NoFiles function*

---

**Description**

Internal. Performs batch correction.

**Usage**

```
ComBat_NoFiles(dat, saminfo, type = "txt", write = F, covariates = "all",
  par.prior = F, filter = F, skip = 0, prior.plots = T)
```

**Arguments**

<code>dat</code>	<code>dat</code>
<code>saminfo</code>	<code>saminfo</code>
<code>type</code>	currently supports two data file types 'txt' for a tab-delimited text file and 'csv' for an Excel .csv file (sometimes R handles the .csv file better, so use this if you have problems with a .txt file!).
<code>write</code>	if 'T' ComBat writes adjusted data to a file, and if 'F' and ComBat outputs the adjusted data matrix if 'F' (so assign it to an object! i.e. <code>NewData &lt;- ComBat('my expression.xls','Sample info file.txt', write=F)</code> ).
<code>covariates</code>	'covariates=all' will use all of the columns in your sample info file in the modeling (except array/sample name), if you only want use a some of the columns in your sample info file, specify these columns here as a vector (you must include the Batch column in this list).
<code>par.prior</code>	if 'T' uses the parametric adjustments, if 'F' uses the nonparametric adjustments— if you are unsure what to use, try the parametric adjustments (they run faster) and check the plots to see if these priors are reasonable.
<code>filter</code>	'filter=value' filters the genes with absent calls in > 1-value of the samples. The default here (as well as in dchip) is .8. Filter if you can as the EB adjustments work better after filtering. Filter must be numeric if your expression index file contains presence/absence calls (but you can set it >1 if you don't want to filter any genes) and must be 'F' if your data doesn't have presence/absence calls;
<code>skip</code>	is the number of columns that contain probe names and gene information, so 'skip=5' implies the first expression values are in column 6
<code>prior.plots</code>	if true will give prior plots with black as a kernel estimate of the empirical batch effect density and red as the parametric estimate.

**Value**

Results.

---

`combineForEachOutput`    *The combineForEachOutput function*

---

**Description**

Internal. Function to combine results from the foreach loop.

**Usage**

```
combineForEachOutput(out1, out2)
```

**Arguments**

<code>out1</code>	result from one foreach loop.
<code>out2</code>	result from another foreach loop.

**Value**

List with the combined results.

---

Download\_DNAmethylation

*The Download\_DNAmethylation function*


---

**Description**

Downloads DNA methylation data from TCGA.

**Usage**

```
Download_DNAmethylation(CancerSite, TargetDirectory, downloadData = TRUE)
```

**Arguments**

CancerSite	character of length 1 with TCGA cancer code.
TargetDirectory	character with directory where a folder for downloaded files will be created.
downloadData	logical indicating if data should be downloaded (default: TRUE). If false, the url of the desired data is returned.

**Value**

list with paths to downloaded files for both 27k and 450k methylation data.

**Examples**

```
## Not run:

# Optional register cluster to run in parallel
library(doParallel)
cl <- makeCluster(5)
registerDoParallel(cl)

# Methylation data for ovarian cancer
cancerSite <- "OV"
targetDirectory <- paste0(getwd(), "/")

# Downloading methylation data
METdirectories <- Download_DNAmethylation(cancerSite, targetDirectory, TRUE)

# Processing methylation data
METProcessedData <- Preprocess_DNAmethylation(cancerSite, METdirectories)

# Saving methylation processed data
saveRDS(METProcessedData, file = paste0(targetDirectory, "MET_", cancerSite, "_Processed.rds"))
```

```
# Clustering methylation data
res <- ClusterProbes(METProcessedData[[1]], METProcessedData[[2]])

# Saving methylation clustered data
toSave <- list(METcancer = res[[1]], METnormal = res[[2]], ProbeMapping = res$ProbeMapping)
saveRDS(toSave, file = paste0(targetDirectory, "MET_", cancerSite, "_Clustered.rds"))

stopCluster(cl)

## End(Not run)
```

---

Download\_GeneExpression

*The Download\_GeneExpression function*

---

## Description

Downloads gene expression data from TCGA.

## Usage

```
Download_GeneExpression(CancerSite, TargetDirectory, downloadData = TRUE)
```

## Arguments

CancerSite	character of length 1 with TCGA cancer code.
TargetDirectory	character with directory where a folder for downloaded files will be created.
downloadData	logical indicating if data should be downloaded (default: TRUE). If false, the url of the desired data is returned.

## Details

This function downloads RNAseq data (file tag "mRNAseq\_Preprocess.Level\_3"), with the exception for OV and GBM, for which micro array data is downloaded since there is not enough RNAseq data

## Value

list with paths to downloaded files for both 27k and 450k methylation data.



**Examples**

```
## Not run:

# Optional register cluster to run in parallel
library(doParallel)
cl <- makeCluster(5)
registerDoParallel(cl)

# Gene expression data for ovarian cancer
cancerSite <- "OV"
targetDirectory <- paste0(getwd(), "/")

# Downloading gene expression data
GEDirectories <- Download_GeneExpression(cancerSite, targetDirectory, TRUE)

# Processing gene expression data
GEProcessedData <- Preprocess_GeneExpression(cancerSite, GEDirectories)

# Saving gene expression processed data
saveRDS(GEProcessedData, file = paste0(targetDirectory, "GE_", cancerSite, "_Processed.rds"))

stopCluster(cl)

## End(Not run)
```

GEcancer

*Cancer Gene expression data of glioblastoma patients from the TCGA project*

**Description**

Cancer Gene expression data of glioblastoma patients from the TCGA project. A set of 14 genes that have been shown in the literature to be involved in differential methylation in glioblastoma were selected as an example to try out MethyLMix.

**Usage**

```
data(GEcancer)
```

**Format**

A numeric matrix with 14 rows (genes) and 251 columns (samples).

**References**

Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008 Oct 23; 455(7216):1061-8. doi: 10.1038/nature07385. Epub 2008 Sep 4. Erratum in: *Nature*. 2013 Feb 28;494(7438):506. PubMed PMID: 18772890; PubMed Central PMCID: PMC2671642.

**See Also**

TCGA: The Cancer Genome Atlas: <http://cancergenome.nih.gov/>

---

GetData	<i>The GetData function</i>
---------	-----------------------------

---

**Description**

This function wraps the functions for downloading and pre-processing DNA methylation and gene expression data, as well as for clustering CpG probes.

**Usage**

```
GetData(cancerSite, targetDirectory)
```

**Arguments**

`cancerSite`        character of length 1 with TCGA cancer code.  
`targetDirectory`    character with directory where a folder for downloaded files will be created.

**Details**

Pre-process of DNA methylation data includes eliminating samples and genes with too many NAs, imputing NAs, and doing Batch correction. If there is both 27k and 450k data, and both data sets have more than 50 samples, we combine the data sets, by reducing the 450k data to the probes present in the 27k data, and bath correction is performed again to the combined data set. If there are samples with both 27k and 450k data, the 450k data is used and the 27k data is discarded, before the step mentioned above. If the 27k or the 450k data does not have more than 50 samples, we use the one with the greatest number of samples, we do not combine the data sets.

For gene expression, this function downloads RNAseq data (file tag "mRNAseq\_Preprocess.Level\_3"), with the exception for OV and GBM, for which micro array data is downloaded since there is not enough RNAseq data. Pre-process of gene expression data includes eliminating samples and genes with too many NAs, imputing NAs, and doing Batch correction.

For the clustering of the CpG probes, this function uses the annotation for Illumina methylation arrays to map each probe to a gene. Then, for each gene, it clusters all its CpG sites using hierarchical clustering and Pearson correlation as distance and complete linkage. If data for normal samples is provided, only overlapping probes between cancer and normal samples are used. Probes with SNPs are removed.

This function is prepared to run in parallel if the user registers a parallel structure, otherwise it runs sequentially.

This function also cleans up the sample names, converting them to the 12 digit format.

**Value**

The following files will be created in target directory:

- gdac: a folder with the raw data downloaded from TCGA.
- MET\_CancerSite\_Processed.rds: processed methylation data at the CpG sites level (not clustered).
- GE\_CancerSite\_Processed.rds: processed gene expression data.
- data\_CancerSite.rds: list with both gene expression and methylation data. Methylation data is clustered and presented at the gene level. A matrix with the mapping from CpG sites to genes is included.

**Examples**

```
## Not run:
# Get data for ovarian cancer
cancerSite <- "OV"
targetDirectory <- paste0(getwd(), "/")
GetData(cancerSite, targetDirectory)

# Optional register cluster to run in parallel
library(doParallel)
cl <- makeCluster(5)
registerDoParallel(cl)

cancerSite <- "OV"
targetDirectory <- paste0(getwd(), "/")
GetData(cancerSite, targetDirectory)

stopCluster(cl)

## End(Not run)
```

---

get_firehoseData	<i>The get_firehoseData function</i>
------------------	--------------------------------------

---

**Description**

Gets data from TCGA's firehose.

**Usage**

```
get_firehoseData(downloadData = TRUE, saveDir = "./",
  TCGA_acronym_uppercase = "LUAD", dataType = "stddata",
  dataFileTag = "mRNAseq_Preprocess.Level_3", FFPE = FALSE,
  fileType = "tar.gz", gdacURL = "http://gdac.broadinstitute.org/runs/",
  untarUngzip = TRUE, printDisease_abbr = FALSE)
```

Arguments

- downloadData      logical indicating if data should be downloaded (default: TRUE). If false, the url of the desired data is returned.
- saveDir            path to directory to save downloaded files.
- TCGA\_acronym\_uppercase      TCGA's cancer site code.
- dataType          type of data in TCGA (default: "stddata").
- dataFileTag       name of the file to be downloaded (the default is to download RNAseq data, but this can be changed to download other data).
- FFPE               logical indicating if FFPE data should be downloaded (default: FALSE).
- fileType           type of downloaded file (default: "fileType", other type not admitted at the moment).
- gdacURL           gdac url.
- untarUngzip       logical indicating if the gzip file downloaded should be untarred (default: TRUE).
- printDisease\_abbr      if TRUE data is not downloaded but all the possible cancer sites codes are shown (default: FALSE).

Value

DownloadedFile path to directory with downloaded files.

---

METcancer	<i>DNA methylation data from cancer tissue from glioblastoma patients from the TCGA project</i>
-----------	---

---

Description

Cancer Gene expression data of glioblastoma patients from the TCGA project. A set of 14 genes that have been shown in the literature to be involved in differential methylation in glioblastoma were selected as an example to try out MethyLMix.

Usage

data(METcancer)

Format

A numeric matrix with 14 rows (genes) and 251 columns (samples).

References

Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23; 455(7216):1061-8. doi: 10.1038/nature07385. Epub 2008 Sep 4. Erratum in: Nature. 2013 Feb 28;494(7438):506. PubMed PMID: 18772890; PubMed Central PMCID: PMC2671642.

**See Also**

TCGA: The Cancer Genome Atlas: <http://cancergenome.nih.gov/>

---

MethylMix

*MethylMix: Mixture model for DNA methylation data in cancer.*


---

**Description**

MethylMix identifies DNA methylation driven genes by modeling DNA methylation data in cancer vs. normal and looking for homogeneous subpopulations. In addition matched gene expression data (e.g. from microarray technology or RNA sequencing) is used to identify functional DNA methylation events by requiring a negative correlation between methylation and gene expression of a particular gene. See references below.

**Usage**

```
MethylMix(METcancer, GEcancer, METnormal = NULL, listOfGenes = NULL,
  filter = TRUE, NoNormalMode = FALSE, OutputRoot = "")
```

**Arguments**

METcancer	Matrix with the methylation data of cancer tissue with genes in rows and samples in columns.
GEcancer	Gene expression data for cancer tissue with genes in rows and samples in columns.
METnormal	Matrix with the normal methylation data of the same genes as in METcancer. Again genes in rows and samples in columns. The samples do not have to match with the cancer data. If this argument is NULL, MethylMix will run without comparing to normal samples.
listOfGenes	Vector with genes names to be evaluated, names must coincide with the names of the rows of METcancer.
filter	Logical indicating if the linear regression to select genes with significant negative relation between methylation and gene expression should be performed (default: TRUE).
NoNormalMode	Logical indicating if the methylation states found in the cancer samples should be compared to the normal samples (default: FALSE).
OutputRoot	Path to store the MethylMix results object.

**Value**

MethylMixResults is a list with the following components:

MethylationDrivers

Genes identified as transcriptionally predictive and differentially methylated by MethylMix.

NrComponents

The number of methylation states found for each driver gene.

MixtureStates	A list with the DM-values for each driver gene. Differential Methylation values (DM-values) are defined as the difference between the methylation mean in one mixture component of cancer samples and the methylation mean in the normal samples, for a given gene.
MethylationStates	Matrix with DM-values for all driver genes (rows) and all samples (columns).
Classifications	Matrix with integers indicating to which mixture component each cancer sample was assigned to, for each gene.
Models	Beta mixture model parameters for each driver gene.

## References

- Gevaert O. **MethylMix: an R package for identifying DNA methylation-driven genes**. Bioinformatics (Oxford, England). 2015;31(11):1839-41. doi:10.1093/bioinformatics/btv020.
- Gevaert O, Tibshirani R, Plevritis SK. **Pancancer analysis of DNA methylation-driven genes using MethylMix**. Genome Biology. 2015;16(1):17. doi:10.1186/s13059-014-0579-8.
- Pierre-Louis Cedoz, Marcos Prunello, Kevin Brennan, Olivier Gevaert; MethylMix 2.0: an R package for identifying DNA methylation genes. Bioinformatics. doi:10.1093/bioinformatics/bty156.

## Examples

```
# load the three data sets needed for MethylMix
data(METcancer)
data(METnormal)
data(GEcancer)

# run MethylMix on a small set of example data
MethylMixResults <- MethylMix(METcancer, GEcancer, METnormal)

## Not run:
# run in parallel
library(doParallel)
cl <- makeCluster(5)
registerDoParallel(cl)
MethylMixResults <- MethylMix(METcancer, GEcancer, METnormal)
stopCluster(cl)

## End(Not run)
```

**Description**

Internal. Prepares all the structures to store the results and calls in a foreach loop a function that fits the mixture model in each gene.

**Usage**

```
MethylMix_MixtureModel(METcancer, METnormal = NULL, FunctionalGenes,
  NoNormalMode = FALSE)
```

**Arguments**

METcancer	matrix with methylation data for cancer samples (genes in rows, samples in columns).
METnormal	matrix with methylation data for normal samples (genes in rows, samples in columns). If NULL no comparison to normal samples will be done.
FunctionalGenes	vector with genes names to be considered for the mixture models.
NoNormalMode	logical, if TRUE no comparison to normal samples is performed. Defaults to FALSE.

**Value**

MethylationStates matrix of DM values, with driver genes in the rows and samples in the columns.

NrComponents matrix with the number of components identified for each driver gene.

Models list with the mixture model fitted for each driver gene.

MethylationDrivers character vector with the genes found by MethylMix as differentially methylated and transcriptionally predictive (driver genes).

MixtureStates a list with a matrix for each driver gene containing the DM values.

Classifications a vector indicating to which component each sample was assigned.

---

MethylMix\_ModelGeneExpression

*The MethylMix\_ModelGeneExpression function*

---

**Description**

Model gene expression as a function of gene expression with a simple linear regression model. Genes with a significant negative linear association between DNA methylation and gene expression are returned.

**Usage**

```
MethylMix_ModelGeneExpression(METcancer, GECancer, CovariateData = NULL)
```

**Arguments**

METcancer	matrix with methylation data for cancer samples (genes in rows, samples in columns).
GEcancer	matrix with gene expression data for cancer samples (genes in rows, samples in columns).
CovariateData	vector (numeric or character) indicating a covariate to be included in the model to adjust for it. Not used in an standard run of MethylMix. It can be used if samples can from different tissue type, for example.

**Value**

vector with the names of the genes for which there is a significant linear and negative association between methylation and gene expression.

**Examples**

```
# load data sets
data(METcancer)
data(GEcancer)

# model gene expression
MethylMixResults <- MethylMix_ModelGeneExpression(METcancer, GEcancer)
```

---

MethylMix\_ModelSingleGene

*The MethylMix\_ModelSingleGene function*

---

**Description**

Internal. For a given gene, this function fits the mixture model, selects the number of components and defines the respective methylation states.

**Usage**

```
MethylMix_ModelSingleGene(GeneName, METdataVector, METdataNormalVector = NULL,
  NoNormalMode = FALSE, maxComp = 3, PvalueThreshold = 0.01,
  MeanDifferenceTreshold = 0.1, minSamplesPerGroup = 1)
```

**Arguments**

GeneName	character string with the name of the gene to model
METdataVector	vector with methylation data for cancer samples.
METdataNormalVector	vector with methylation data for normal samples. It can be NULL and then no normal mode will be used.



NoNormalMode	logical, if TRUE no comparison to normal samples is performed. Defaults to FALSE.
maxComp	maximum number of mixture components admitted in the model (3 by default).
PvalueThreshold	threshold to consider results significant.
MeanDifferenceThreshold	threshold in beta value scale from which two methylation means are considered different.
minSamplesPerGroup	minimum number of samples required to belong to a new mixture component in order to accept it. Default is 1 (not used). If -1, each component has to have at least 5% of all cancer samples.

### Details

maxComp, PvalueThreshold, METDiffThreshold, minSamplesPerGroup are arguments for this function but are fixed in their default values for the user because they are not available in the main MethylMix function, to keep it simple. It would be easy to make them available to the user if we want to.

### Value

NrComponents number of components identified.  
 Models an object with the parameters of the model fitted.  
 MethylationStates vector with DM values for each sample.  
 MixtureStates vector with DMvalues for each component.  
 Classifications a vector indicating to which component each sample was assigned.  
 FlipOverState FlipOverState

---

MethylMix_PlotModel	<i>The MethylMix_PlotModel function.</i>
---------------------	--

---

### Description

Produces plots to represent MethylMix's output.

### Usage

```
MethylMix_PlotModel(GeneName, MixtureModelResults, METcancer, GEcancer = NULL,
  METnormal = NULL)
```

**Arguments**

GeneName	Name of the gene for which to create a MethylMix plot.
MixtureModelResults	List returned by MethylMix function.
METcancer	Matrix with the methylation data of cancer tissue with genes in rows and samples in columns.
GEcancer	Gene expression data for cancer tissue with genes in rows and samples in columns (optional).
METnormal	Matrix with the normal methylation data of the same genes as in METcancer (optional). Again genes in rows and samples in columns.

**Value**

a list with MethylMix plots, a histogram of the methylation data (MixtureModelPlot) and a scatter-plot between DNA methylation and gene expression (CorrelationPlot, is NULL if gene expression data is not provided). Both plots show the different mixture components identified.

**Examples**

```
# Load the three data sets needed for MethylMix
data(METcancer)
data(METnormal)
data(GEcancer)

# Run methylmix on a small set of example data
MethylMixResults <- MethylMix(METcancer, GEcancer, METnormal)

# Plot the most famous methylated gene for glioblastoma
MethylMix_PlotModel("MGMT", MethylMixResults, METcancer)

# Plot MGMT also with its normal methylation variation
MethylMix_PlotModel("MGMT", MethylMixResults, METcancer, METnormal = METnormal)

# Plot a MethylMix model for another gene
MethylMix_PlotModel("ZNF217", MethylMixResults, METcancer, METnormal = METnormal)

# Also plot the inverse correlation with gene expression (creates two separate plots)
MethylMix_PlotModel("MGMT", MethylMixResults, METcancer, GEcancer, METnormal)

# Plot all functional and differential genes
for (gene in MethylMixResults$MethylationDrivers) {
  MethylMix_PlotModel(gene, MethylMixResults, METcancer, METnormal = METnormal)
}
```

---

MethylMix_Predict	<i>The MethylMix_Predict function</i>
-------------------	---------------------------------------

---

## Description

Given a new data set with methylation data, this function predicts the mixture component for each new sample and driver gene. Predictions are based on posterior probabilities calculated with MethylMix's fitted mixture model.

## Usage

```
MethylMix_Predict(newBetaValuesMatrix, MethylMixResult)
```

## Arguments

`newBetaValuesMatrix`

Matrix with new observations for prediction, genes/cpg sites in rows, samples in columns. Although this new matrix can have a different number of genes/cpg sites than the one provided as METcancer when running MethylMix, naming of genes/cpg sites should be the same.

`MethylMixResult`

Output object from MethylMix

## Value

A matrix with predictions (indices of mixture component), driver genes in rows, new samples in columns

## Examples

```
# load the three data sets needed for MethylMix
data(METcancer)
data(METnormal)
data(GEcancer)

# run MethylMix on a small set of example data
MethylMixResults <- MethylMix(METcancer, GEcancer, METnormal)
# toy example new data, of same dimension of original METcancer data
newMETData <- matrix(runif(length(METcancer)), nrow = nrow(METcancer))
rownames(newMETData) <- rownames(METcancer)
colnames(newMETData) <- paste0("sample", 1:ncol(METcancer))
predictions <- MethylMix_Predict(newMETData, MethylMixResults)
```

---

MethylMix\_RemoveFlipOver

*The MethylMix\_RemoveFlipOver function*


---

### Description

Internal. The estimated densities for each beta component can overlap, generating samples that look like being separated from their group. This function re classifies such samples.

### Usage

```
MethylMix_RemoveFlipOver(OrigOrder, MethylationState, classification,
  METdataVector, NrComponents, UseTrainedFlipOver = FALSE,
  FlipOverState = 0)
```

### Arguments

OrigOrder	order of sorted values in the methylation vector.
MethylationState	methylation states for this gene.
classification	vector with integers indicating to wich component each sample was classified into.
METdataVector	vector with methylation values from the cancer samples.
NrComponents	number of components in this gene.
UseTrainedFlipOver	.
FlipOverState	.

### Value

Corrected vectors with methylation states and classification.

---

METnormal

*DNA methylation data from normal tissue from glioblastoma patients*


---

### Description

Normal tissue DNA methylation data of glioblastoma patients. These normal tissue samples were run on the same platform and are described in the publication referenced below.

### Usage

```
data(METnormal)
```

**Format**

A numeric matrix with 14 rows (genes) and 4 columns (samples).

**References**

Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloso CE, Sulman EP, Bhat KP, Verhaak RG, Hoadley KA, Hayes DN, Perou CM, Schmidt HK, Ding L, Wilson RK, Van Den Berg D, Shen H, Bengtsson H, Neuvial P, Cope LM, Buckley J, Herman JG, Baylin SB, Laird PW, Aldape K; Cancer Genome Atlas Research Network. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*. 2010 May 18;17(5):510-22. doi: 10.1016/j.ccr.2010.03.017. Epub 2010 Apr 15. PubMed PMID: 20399149; PubMed Central PMCID: PMC2872684

---

predictOneGene	<i>The predictOneGene function</i>
----------------	------------------------------------

---

**Description**

Auxiliar function. Given a new vector of beta values, this function calculates a matrix with posterior prob of belonging to each mixture component (columns) for each new beta value (rows), and return the number of the mixture component with highest posterior probabilit

**Usage**

```
predictOneGene(newVector, mixtureModel)
```

**Arguments**

newVector	vector with new beta values
mixtureModel	beta mixture model object for the gene being evaluated.

**Value**

A matrix with predictions (indices of mixture component), driver genes in rows, new samples in columns

---

`Preprocess_CancerSite_Methylation27k`*The Preprocess\_CancerSite\_Methylation27k function*

---

**Description**

Internal. Pre-processes DNA methylation data from TCGA from Illymina 27k arrays.

**Usage**

```
Preprocess_CancerSite_Methylation27k(CancerSite, METdirectory,  
  MissingValueThreshold = 0.2)
```

**Arguments**

<code>CancerSite</code>	character of length 1 with TCGA cancer code.
<code>METdirectory</code>	character with directory where a folder for downloaded files will be created. Can be the object returned by the <code>Download_DNAMethylation</code> function.
<code>MissingValueThreshold</code>	threshold for removing samples or genes with missing values.

**Value**

List with pre processed methylation data for cancer and normal samples.

---

`Preprocess_CancerSite_Methylation450k`*The Preprocess\_CancerSite\_Methylation450k function*

---

**Description**

Internal. Pre-processes DNA methylation data from TCGA from Illymina 450k arrays.

**Usage**

```
Preprocess_CancerSite_Methylation450k(CancerSite, METdirectory,  
  MissingValueThreshold = 0.2)
```

**Arguments**

<code>CancerSite</code>	character of length 1 with TCGA cancer code.
<code>METdirectory</code>	character with directory where a folder for downloaded files will be created. Can be the object returned by the <code>Download_DNAMethylation</code> function.
<code>MissingValueThreshold</code>	threshold for removing samples or genes with missing values.

**Value**

List with pre processed methylation data for cancer and normal samples.

---

```
Preprocess_DNAMethylation
```

*The Preprocess\_DNAMethylation function*

---

**Description**

Pre-processes DNA methylation data from TCGA.

**Usage**

```
Preprocess_DNAMethylation(CancerSite, METdirectories,
  MissingValueThreshold = 0.2)
```

**Arguments**

**CancerSite** character of length 1 with TCGA cancer code.

**METdirectories** character vector with directories with the downloaded data. It can be the object returned by the Download\_DNAMethylation function.

**MissingValueThreshold** threshold for removing samples or genes with missing values.

**Details**

Pre-process includes eliminating samples and genes with too many NAs, imputing NAs, and doing Batch correction. If there is both 27k and 450k data, and both data sets have more than 50 samples, we combine the data sets, by reducing the 450k data to the probes present in the 27k data, and bath correction is performed again to the combined data set. If there are samples with both 27k and 450k data, the 450k data is used and the 27k data is discarded, before the step mentioned above. If the 27k or the 450k data does not have more than 50 samples, we use the one with the greatest number of samples, we do not combine the data sets.

**Value**

List with the pre-processed data matrix for cancer and normal samples.

**Examples**

```
## Not run:

# Optional register cluster to run in parallel
library(doParallel)
cl <- makeCluster(5)
registerDoParallel(cl)
```

```

# Methylation data for ovarian cancer
cancerSite <- "OV"
targetDirectory <- paste0(getwd(), "/")

# Downloading methylation data
METdirectories <- Download_DNAMethylation(cancerSite, targetDirectory, TRUE)

# Processing methylation data
METProcessedData <- Preprocess_DNAMethylation(cancerSite, METdirectories)

# Saving methylation processed data
saveRDS(METProcessedData, file = paste0(targetDirectory, "MET_", cancerSite, "_Processed.rds"))

# Clustering methylation data
res <- ClusterProbes(METProcessedData[[1]], METProcessedData[[2]])

# Saving methylation clustered data
toSave <- list(METcancer = res[[1]], METnormal = res[[2]], ProbeMapping = res$ProbeMapping)
saveRDS(toSave, file = paste0(targetDirectory, "MET_", cancerSite, "_Clustered.rds"))

stopCluster(cl)

## End(Not run)

```

---

Preprocess\_GeneExpression

*The Preprocess\_GeneExpression function*


---

## Description

Pre-processes gene expression data from TCGA.

## Usage

```

Preprocess_GeneExpression(CancerSite, MAdirectories,
  MissingValueThresholdGene = 0.3, MissingValueThresholdSample = 0.1)

```

## Arguments

CancerSite	character of length 1 with TCGA cancer code.
MAdirectories	character vector with directories with the downloaded data. It can be the object returned by the Download_DNAMethylation function.
MissingValueThresholdGene	threshold for missing values per gene. Genes with a percentage of NAs greater than this threshold are removed. Default is 0.3.
MissingValueThresholdSample	threshold for missing values per sample. Samples with a percentage of NAs greater than this threshold are removed. Default is 0.1.



**Details**

Pre-process includes eliminating samples and genes with too many NAs, imputing NAs, and doing Batch correction.

**Value**

List with the pre-processed data matrix for cancer and normal samples.

**Examples**

```
## Not run:

# Optional register cluster to run in parallel
library(doParallel)
cl <- makeCluster(5)
registerDoParallel(cl)

# Gene expression data for ovarian cancer
cancerSite <- "OV"
targetDirectory <- paste0(getwd(), "/")

# Downloading gene expression data
GEDirectories <- Download_GeneExpression(cancerSite, targetDirectory, TRUE)

# Processing gene expression data
GEProcessedData <- Preprocess_GeneExpression(cancerSite, GEDirectories)

# Saving gene expression processed data
saveRDS(GEProcessedData, file = paste0(targetDirectory, "GE_", cancerSite, "_Processed.rds"))

stopCluster(cl)

## End(Not run)
```

---

Preprocess\_MAdata\_Cancer

*The Preprocess\_MAdata\_Cancer function*

---

**Description**

Internal. Pre-process gene expression data for cancer samples.

**Usage**

```
Preprocess_MAdata_Cancer(CancerSite, Directory, File,
  MissingValueThresholdGene = 0.3, MissingValueThresholdSample = 0.1)
```

**Arguments**

CancerSite	TCGA code for the cancer site.
Directory	Directory.
File	File.
MissingValueThresholdGene	threshold for missing values per gene. Genes with a percentage of NAs greater than this threshold are removed. Default is 0.3.
MissingValueThresholdSample	threshold for missing values per sample. Samples with a percentage of NAs greater than this threshold are removed. Default is 0.1.

**Value**

The data matrix.

---

Preprocess\_MAdata\_Normal

*The Preprocess\_MAdata\_Normal function*

---

**Description**

Internal. Pre-process gene expression data for normal samples.

**Usage**

```
Preprocess_MAdata_Normal(CancerSite, Directory, File,
  MissingValueThresholdGene = 0.3, MissingValueThresholdSample = 0.1)
```

**Arguments**

CancerSite	TCGA code for the cancer site.
Directory	Directory.
File	File.
MissingValueThresholdGene	threshold for missing values per gene. Genes with a percentage of NAs greater than this threshold are removed. Default is 0.3.
MissingValueThresholdSample	threshold for missing values per sample. Samples with a percentage of NAs greater than this threshold are removed. Default is 0.1.

**Value**

The data matrix.

---

ProbeAnnotation	<i>ProbeAnnotation data set</i>
-----------------	---------------------------------

---

**Description**

Data set with annotation from Illumina methylatin arrays mapping CpG sites to genes.

---

SNPprobes	<i>SNPprobes data set</i>
-----------	---------------------------

---

**Description**

Vector with probes with SNPs.

---

TCGA_BatchCorrection_MolecularData	<i>The TCGA_BatchCorrection_MolecularData function</i>
------------------------------------	--

---

**Description**

Internal. Wrapper to perform batch correction.

**Usage**

TCGA\_BatchCorrection\_MolecularData(GEN\_Data, BatchData, MinInBatch)

**Arguments**

- |            |   |
|------------|---|
| GEN_Data   | matrix with data to be corrected for batch effects. |
| BatchData  | Batch data.   |
| MinInBatch | minimum number of samples per batch.                |

**Value**

The corrected data matrix.

---

TCGA\_GENERIC\_BatchCorrection

*The TCGA\_GENERIC\_BatchCorrection function*

---

**Description**

Internal. Performs batch correction.

**Usage**

TCGA\_GENERIC\_BatchCorrection(GEN\_Data, BatchData)

**Arguments**

GEN_Data	matrix with data to be corrected for batch effects.
BatchData	Batch data.

**Value**

The corrected data matrix.

---

TCGA\_GENERIC\_CheckBatchEffect

*The TCGA\_GENERIC\_CheckBatchEffect function*

---

**Description**

Internal. Checks if batch correction is needed.

**Usage**

TCGA\_GENERIC\_CheckBatchEffect(GEN\_Data, BatchData)

**Arguments**

GEN_Data	matrix with data to be corrected for batch effects.
BatchData	Batch data.

**Value**

list with results.

---

`TCGA_GENERIC_CleanUpSampleNames`*The TCGA\_GENERIC\_CleanUpSampleNames function*

---

**Description**

Internal. Cleans the samples IDs into the 12 digit format and removes doubles.

**Usage**

```
TCGA_GENERIC_CleanUpSampleNames(GEN_Data, IDlength = 12)
```

**Arguments**

<code>GEN_Data</code>	data matrix.
<code>IDlength</code>	length of samples ID.

**Value**

data matrix with cleaned sample names.

---

`TCGA_GENERIC_GetSampleGroups`*The TCGA\_GENERIC\_GetSampleGroups function*

---

**Description**

Internal. Looks for the group of the samples (normal/cancer).

**Usage**

```
TCGA_GENERIC_GetSampleGroups(SampleNames)
```

**Arguments**

<code>SampleNames</code>	vector with sample names.
--------------------------	---------------------------

**Value**

a list.

---

`TCGA_GENERIC_LoadIlluminaMethylationData`*The TCGA\_GENERIC\_LoadIlluminaMethylationData function*

---

**Description**

Internal. Read in an illumina methylation file with the following format: header row with sample labels, 2nd header row with 4 columns per sample: beta-value, geneSymbol, chromosome and GenomicCoordinate. The first column has the probe names.

**Usage**

```
TCGA_GENERIC_LoadIlluminaMethylationData(Filename)
```

**Arguments**

Filename            name of the file with the data.

**Value**

methylation data.

---

`TCGA_GENERIC_MergeData`*The TCGA\_GENERIC\_MergeData function*

---

**Description**

Internal.

**Usage**

```
TCGA_GENERIC_MergeData(NewIDListUnique, DataMatrix)
```

**Arguments**

NewIDListUnique            unique rownames of data.

DataMatrix            data matrix.

**Value**

data matrix.

---

`TCGA_GENERIC_MET_ClusterProbes_Helper_ClusterGenes_with_hclust`*The TCGA\_GENERIC\_MET\_ClusterProbes\_Helper\_ClusterGenes\_with\_hclust function*

---

**Description**

Internal. Cluster probes into genes.

**Usage**

```
TCGA_GENERIC_MET_ClusterProbes_Helper_ClusterGenes_with_hclust(Gene,  
  ProbeAnnotation, MET_Cancer, MET_Normal = NULL, CorThreshold = 0.4)
```

**Arguments**

Gene	gene.
ProbeAnnotation	data set matching probes to genes.
MET_Cancer	data matrix for cancer samples.
MET_Normal	data matrix for normal samples.
CorThreshold	correlation threshold for cutting the clusters.

**Value**

List with the clustered data sets and the mapping between probes and genes.

---

`TCGA_Load_MolecularData`*The TCGA\_Load\_MolecularData function*

---

**Description**

Internal. Reads in gene expression data. Deletes samples and genes with more NAs than the respective thresholds. Imputes other NAs values.

**Usage**

```
TCGA_Load_MolecularData(Filename, MissingValueThresholdGene = 0.3,  
  MissingValueThresholdSample = 0.1)
```

**Arguments**

Filename            name of the file with the data.

MissingValueThresholdGene            threshold for missing values per gene. Genes with a percentage of NAs greater than this threshold are removed. Default is 0.3.

MissingValueThresholdSample            threshold for missing values per sample. Samples with a percentage of NAs greater than this threshold are removed. Default is 0.1.

**Value**

gene expression data.

---

TCGA\_Process\_EstimateMissingValues

*The TCGA\_Process\_EstimateMissingValues function*

---

**Description**

Internal. Removes patients and genes with more missing values than the MissingValueThreshold, and imputes remaining missing values using Tibshirani's KNN method.

**Usage**

```
TCGA_Process_EstimateMissingValues(MET_Data, MissingValueThreshold = 0.2)
```

**Arguments**

MET\_Data            data matrix.

MissingValueThreshold            threshold for removing samples and genes with too many missing values.

**Value**

the data set with imputed values and possibly some genes or samples deleted.



# Index

- \* **cluster**
  - GetData, [10](#)
- \* **cluster\_probes**
  - ClusterProbes, [4](#)
- \* **datasets**
  - BatchData, [3](#)
  - GEcancer, [9](#)
  - METcancer, [12](#)
  - METnormal, [20](#)
  - ProbeAnnotation, [27](#)
  - SNPprobes, [27](#)
- \* **download**
  - Download\_DNAMethylation, [7](#)
  - Download\_GeneExpression, [8](#)
  - GetData, [10](#)
- \* **internal**
  - betaEst\_2, [3](#)
  - blc\_2, [4](#)
  - ComBat\_NoFiles, [5](#)
  - combineForEachOutput, [6](#)
  - get\_firehoseData, [11](#)
  - MethylMix\_MixtureModel, [14](#)
  - MethylMix\_ModelSingleGene, [16](#)
  - MethylMix\_RemoveFlipOver, [20](#)
  - Preprocess\_CancerSite\_Methylation27k, [22](#)
  - Preprocess\_CancerSite\_Methylation450k, [22](#)
  - Preprocess\_MAdata\_Cancer, [25](#)
  - Preprocess\_MAdata\_Normal, [26](#)
  - TCGA\_BatchCorrection\_MolecularData, [27](#)
  - TCGA\_GENERIC\_BatchCorrection, [28](#)
  - TCGA\_GENERIC\_CheckBatchEffect, [28](#)
  - TCGA\_GENERIC\_CleanUpSampleNames, [29](#)
  - TCGA\_GENERIC\_GetSampleGroups, [29](#)
  - TCGA\_GENERIC\_LoadIlluminaMethylationData, [30](#)
  - TCGA\_GENERIC\_MergeData, [30](#)
  - TCGA\_GENERIC\_MET\_ClusterProbes\_Helper\_ClusterGenes\_wit, [31](#)
  - TCGA\_Load\_MolecularData, [31](#)
  - TCGA\_Process\_EstimateMissingValues, [32](#)
- \* **preprocess**
  - GetData, [10](#)
  - Preprocess\_DNAMethylation, [23](#)
  - Preprocess\_GeneExpression, [24](#)
- BatchData, [3](#)
- betaEst\_2, [3](#)
- blc\_2, [4](#)
- ClusterProbes, [4](#)
- ComBat\_NoFiles, [5](#)
- combineForEachOutput, [6](#)
- Download\_DNAMethylation, [7](#)
- Download\_GeneExpression, [8](#)
- GEcancer, [9](#)
- get\_firehoseData, [11](#)
- GetData, [10](#)
- METcancer, [12](#)
- MethylMix, [13](#)
- MethylMix\_MixtureModel, [14](#)
- MethylMix\_ModelGeneExpression, [15](#)
- MethylMix\_ModelSingleGene, [16](#)
- MethylMix\_PlotModel, [17](#)
- MethylMix\_Predict, [19](#)
- MethylMix\_RemoveFlipOver, [20](#)
- METnormal, [20](#)
- predictOneGene, [21](#)
- Preprocess\_CancerSite\_Methylation27k, [22](#)
- Preprocess\_CancerSite\_Methylation450k, [22](#)

Preprocess\_DNAMethylation, [23](#)  
Preprocess\_GeneExpression, [24](#)  
Preprocess\_MAdata\_Cancer, [25](#)  
Preprocess\_MAdata\_Normal, [26](#)  
ProbeAnnotation, [27](#)  
  
SNPprobes, [27](#)  
  
TCGA\_BatchCorrection\_MolecularData, [27](#)  
TCGA\_GENERIC\_BatchCorrection, [28](#)  
TCGA\_GENERIC\_CheckBatchEffect, [28](#)  
TCGA\_GENERIC\_CleanUpSampleNames, [29](#)  
TCGA\_GENERIC\_GetSampleGroups, [29](#)  
TCGA\_GENERIC\_LoadIlluminaMethylationData,  
    [30](#)  
TCGA\_GENERIC\_MergeData, [30](#)  
TCGA\_GENERIC\_MET\_ClusterProbes\_Helper\_ClusterGenes\_with\_hclust,  
    [31](#)  
TCGA\_Load\_MolecularData, [31](#)  
TCGA\_Process\_EstimateMissingValues, [32](#)