Package ‘MethylMix’

April 14, 2017

Title MethylMix: Identifying methylation driven cancer genes

Version 2.0.0

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 0. 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 2017-01-13

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,Diffe

RoxygenNote 5.0.1

VignetteBuilder knitr

NeedsCompilation no

R topics documented:

`BatchData` *BatchData data set*

Description

Data set with batch number for TCGA samples.

`ClusterProbes` *The ClusterProbes function*

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 hierchical 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.

`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 *The GetData function*

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 hierchical 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)
```

METcancer *DNA methylation data from cancer tissue from 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(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 significative linear 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 0. 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.

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)
```

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_PlotModel *The MethylMix_PlotModel function.*

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 scatterplot 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)
}
```

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, Pelloski 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

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)
```

ProbeAnnotation *ProbeAnnotation data set*

Description

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

SNPprobes *SNPprobes data set*

Description

Vector with probes with SNPs.

Index