Package ‘ENmix’

April 23, 2016

Version 1.4.1

Date 2015-09-09

Title Data preprocessing and quality control for Illumina HumanMethylation450 BeadChip

Type Package

Description Illumina HumanMethylation450 BeadChip array measurements have intrinsic levels of background noise that degrade methylation measurement. The ENmix package provides an efficient data pre-processing tool designed to reduce background noise and improve signal for DNA methylation estimation. The package utilizes a novel model-based background correction method, ENmix, that significantly improve accuracy and reproducibility of methylation measures. The data structure used by the ENmix package is compatible with several other related R packages, such as minfi, wateRmelon and ChAMP, providing straightforward integration of ENmix-corrected datasets for subsequent data analysis. The software is designed to support large scale data analysis, and provides multi-processor parallel computing wrappers for commonly used data preprocessing methods, including BMIQ probe design type bias correction and ComBat batch effect correction. In addition ENmix package has selectable complementary functions for efficient data visualization (such as data distribution plotting), quality control (identification and filtering of low quality data points, samples, probes, and outliers, along with imputation of missing values), inter-array normalization (3 different quantile normalizations), identification of probes with multimodal distributions due to SNPs and other factors, and exploration of data variance structure using principal component regression analysis plots. Together these provide a set of flexible and transparent tools for preprocessing of EWAS data in a computationally-efficient and user-friendly package.

Depends minfi,parallel,doParallel,Biobase (>= 2.17.8),foreach

Imports MASS,preprocessCore,wateRmelon,sva,geneplotter,impute

Suggests minfiData (>= 0.4.1), RPMM, RUnit, BiocGenerics

biocViews DNAMethylation, Preprocessing, QualityControl, TwoChannel, Microarray, OneChannel, MethylationArray, BatchEffect, Normalization, DataImport

License Artistic-2.0

NeedsCompilation no

Author Zongli Xu [cre, aut],
Liang Niu [aut],
Leping Li [ctb],
Jack Taylor [ctb]

Maintainer Zongli Xu <xuz@niehs.nih.gov>

R topics documented:

`bmiq.mc` *A multi-processor wrapper of BMIQ method*

Description

A multi-processor wrapper of BMIQ method. BMIQ is an intra-sample normalization procedure to correct the bias of Infinium 2 probe methylation beta values.

Usage

```
bmiq.mc(mdat, nCores = 1,...)
```

Arguments

`mdat`	An object of class `MethylSet`.
`nCores`	Number of cores used for computation.
`...`	See BMIQ in R package wateRmelon for more options.

Value

A data matrix of Methylation beta value.

Author(s)

Zongli Xu

References

Teschendorff AE et. al. *A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data*. Bioinformatics. 2013

See Also

See `BMIQ` in R package `watermelon` for model details

Examples

```
if(FALSE){
if (require(minfiData)) {
mdat=preprocessENmix(RGsetEx,bgParaEst="oob",nCores=6)
mdatq1=normalize.quantile.450k(mdat,method="quantile1")
beta=bmiq.mc(mdatq1,nCores=10)
}}
```

`ComBat.mc` *A multi-processor wrapper for ComBat method.*

Description

A multi-processor wrapper for ComBat method. ComBat is a method to adjust batch effect where the batch covariate is known.

Usage

```
ComBat.mc(dat,batch,nCores = 1,...)
```

Arguments

`dat`	A data matrix with column for samples and row for probe.
`batch`	Batch covariate (multiple batches allowed)
`nCores`	Number of cores will be used for computation
`...`	See `ComBat` in `sva` package for extra options

Value

A data matrix with the same dimension as input data, adjusted for batch effects. Warning: Values for multimodal distributed CpGs could be over-adjusted.

Author(s)

Zongli Xu

References

Johnson, WE, Rabinovic, A, and Li, C (2007). *Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics 8(1):118-127.*

See Also

See ComBat in sva package for details.

Examples

```
if(FALSE){
if (require(minfiData)) {
mdat=preprocessENmix(RGsetEx,bgParaEst="oob",nCores=6)
mdat=normalize.quantile.450k(mdat,method="quantile1")
beta=bmiq.mc(mdat,nCores=10)
batch=factor(pData(mdat)$Slide)
betaC=ComBat.mc(beta,batch,nCores=6,mod=NULL)
}}
```

multifreqpoly *Frequency polygon plot to display data distribution.*

Description

Produce Frequency polygon plot for each column of a numeric data matrix.

Usage

```
multifreqpoly(mat, nbreaks=100, col=1:ncol(mat), xlab="",
              ylab="Frequency",legend = list(x = "top", fill=col,
              legend = if(is.null(colnames(mat))) paste(1:ncol(mat))
              else colnames(mat)),...)
```

Arguments

mat	A numeric matrix
nbreaks	The number of bins for frequency counting
col	Line plot color code, the length should be equal to the number of columns in mat
xlab	x-axis lable
ylab	y-axis lable
legend	A list of arguments that get passed to the function "legend"
...	Further arguments that get passed to the function "plot"

Value

Frequency polygon plot.

Author(s)

Zongli Xu

References

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor, *ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip*. Nucleic Acids Research 2015.

Examples

```
if(FALSE){
if (require(minfiData)) {
mdat <- preprocessRaw(RGsetEx)
beta=getBeta(mdat, "Illumina")
multifreqpoly(beta,col=rep("black",ncol(beta)))
}}
```

nmode.mc *Estimating number of mode in methylaion data for each probe.*

Description

Due to SNPs in CpG probe region or other unknow factors, methylation beta values for some CpGs have multimodal distribution. This function is to identify this type of probes with obovious multimoal distribution.

Usage

```
nmode.mc(x, minN = 3, modedist=0.2, nCores = 1)
```

Arguments

`x`	A methylation beta value matrix with row for probes and column for samples.
`minN`	Minimum number of data points at each cluster
`modedist`	Minimum mode distance
`nCores`	Number of cores used for computation

Details

This function used an empirical approach to estimate number of mode in methylation beta value for each CpG probe. By default, the function requires the distance between modes have to be greater than 0.2 in methylation beta value, and each mode clusters should has at least 3 data points or 5% of data points whichever is greater.

Value

A vector of integers

Author(s)

Zongli Xu

References

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor, *ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip.* Nucleic Acids Research 2015

Examples

```
if(FALSE){
if (require(minfiData)) {
mdat <- preprocessRaw(RGsetEx)
beta=getBeta(mdat, "Illumina")
nmode=nmode.mc(beta, minN = 3,modedist=0.2, nCores = 5)
}}
```

---

normalize.quantile.450k

Quantile normalization.

Description

Quantile normalization of methylation intensity data across samples for Illumina Infinium HumanMethylation450 BeadChip.

Usage

```
normalize.quantile.450k(mdat, method = "quantile1")
```

Arguments

mdat	An object of class MethylSet.
method	Quantile normalization method. This should be one of the following strings: "quantile1", "quantile2", or "quantile3".

Details

By default, method = "quantile1" will separately quantile normalize Methylated or Unmethylated intensities for Infinium I or II probes. The "quantile2" will quantile normalize combined Methylated or Unmethylated intensities for Infinium I or II probes. The "quantile3" will quantile normalize combined Methylated or Unmethylated intensities for Infinium I and II probes together.

Value

An object of class MethylSet.

Author(s)

Zongli Xu

References

Pidsley, R., CC, Y.W., Volta, M., Lunnon, K., Mill, J. and Schalkwyk, L.C. (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. BMC genomics, 14, 293.

Examples

```
if(FALSE){
if (require(minfiData)) {
mdat=preprocessENmix(RGsetEx,bgParaEst=”oob”,nCores=6)
mdatq1=normalize.quantile.450k(mdat,method=”quantile1”)
}}
```

pcrplot *Principal component regression plot*

Description

First, principal component analysis will be performed in the standadized input data matrix (standadized for each row/CpG), and then the specified number of top principal components (that explain most data variation) will be used to perform linear regression with each specified variables. Regression P values will be plotted for exploration of methylation data variance structure or identification of possible confounding variables for association analysis.

Usage

```
pcrplot(beta, cov,npc=50)
```

Arguments

`beta` A methylation beta value matrix with row for probes and column for samples.

`cov` A data frame of covariates. Categorical variables should be converted to factors.

`npc` The number of top principal components to plot

Value

A jpeg figure "svdscreeplot.jpg" to show the variations explained by each principal component.

A jpeg figure "pcr_diag.jpg" to show association strength between principal components and covariates with cell colors indicating different levels of association P values.

Author(s)

Zongli Xu

References

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor, *ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip*. Nucleic Acids Research 2015

Examples

```
if(FALSE){
if (require(minfiData)) {
mdat <- preprocessRaw(RGsetEx)
beta=getBeta(mdat, "Illumina")
group=pData(mdat)$Sample_Group
slide=factor(pData(mdat)$Slide)
cov=data.frame(group,slide)
pcrplot(beta,cov,npc=6)
}}
```

---

plotCtrl *Plot internal controls of 450K BeadChip.*

Description

Intensity data are ploted for all internal control probe types on the Illumina Infinium HumanMethylation450 BeadChip. These figures can be used to check data quality and experimental procedures.

Usage

```
plotCtrl(rgSet,IDorder=NULL)
```

Arguments

`rgSet`	An object of class `RGChannelSet`.
`IDorder`	A list of sample ids in the order user specified. The list can be a subset of the samples in input dataset. If an id list is provided, all plots will be produced in the order of the list. The default is NULL.

Value

A set of jpeg figures.

Author(s)

Zongli Xu

References

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor, *ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip*. Nucleic Acids Research 2015.

Examples

```
if(FALSE){
if (require(minfiData)) {
pinfo=pData(RGsetEx)
IDorder=rownames(pinfo)[order(pinfo$Slide,pinfo$Array)]
plotCtrl(RGsetEx,IDorder)
}}
```

`preprocessENmix` *The ENmix background correction for HumanMethylation450k BeadChip*

Description

ENmix models methylation signal intensities with a flexible exponential-normal mixture distribution, and models background noise with a truncated normal distribution. ENmix will split 450k BeadChip intensity data into 6 parts and separately model methylated and unmethylated intensities, 2 different color channels and 2 different probe designs.

Usage

```
preprocessENmix(rgSet, bgParaEst = "oob", dyeCorr=TRUE, QCinfo=NULL,
                exSample=NULL, exCpG=NULL, nCores = 2)
```

Arguments

rgSet	An object of class RGChannelSetExtended, RGChannelSet or MethylSet.
bgParaEst	Optional method to estimate background normal distribution parameters. This must be one of the strings: "oob","est", or "neg".
dyeCorr	Dye bias correction: "TRUE" or "FALSE"
QCinfo	If QCinfo object from function QCinfo() was provided, low quality samples and CpGs will be excluded before background correction.
exSample	User specified sample list to be excluded before background correction
exCpG	User specified probe list to be excluded before background correction
nCores	Number of cores will be used for computation

Details

By default, ENmix will use out-of-band Infinium I intensities ("oob") to estimate normal distribution parameters to model background noise. Option "est" will use combined methylated and unmethylated intensities to estimate background distribution parameters separately for each color channel and each probe type. Option "neg" will use 600 chip internal controls probes to estimate background distribution parameters. If rgSet if a `MethylSet`, then only option "est" can be selected.

Value

An object of class `MethylSet`

Author(s)

Zongli Xu and Liang Niu

References

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor, ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Research 2015.

See Also

Package `minfi` for classes `RGChannelSet` and `MethylSet`

Examples

```
if(FALSE){
if (require(minfiData)) {
mdat=preprocessENmix(RGsetEx,bgParaEst="oob",nCores=6)
}}
```

`QCinfo` *QC information.*

Description

Extract informations for data quanlity controls: detection P values and number of beads for each call of methylation beta value.

Usage

```
QCinfo(rgSet, detPthre=0.05, nbthre=3, samplethre=0.01, CpGthre=0.05,
       bisulthre=NULL, outlier=TRUE, distplot=TRUE)
```

Arguments

`rgSet` An object of class `RGChannelSetExtended`.

`detPthre` Detection P value threshold to identify low quality data point

`nbthre` Number of bead threshold to identify low quality data point

`samplethre` Threshold to identify low quality samples, the percentage of low quality methylation data points across probes for each sample

`CpGthre` Threshold to identify low quality probes, percentage of low quality methylation data points across samples for each probe `bisulthre` Threshold of bisulfite intensity for identification of low quality samples. By default, Mean - 3 x SD of sample bisufite control intensities will be used as the threshold.

`outlier` If TRUE, outlier samples in total intensity or beta value distribution will be idenfied and classified as bad samples.

`distplot` TRUE or FALSE, whether to produce beta value distribution plots before and after QC.

Value

detP: a matrix of detection P values

nbead: a matrix for number of beads

bisul: a vector of averaged intensities for bisulfite conversion controls

badsample: a list of low quality or outlier samples

badCpG: a list of low quality CpGs

Figure "qc_sample.jpg": scatter plot for Percent of low quality data per sample and Average bisulfite conversion intensity

Figure "qc_CpG.jpg": histogram for Percent of low quality data per CpG.

Author(s)

Zongli Xu

References

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor, *ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip*. Nucleic Acids Research 2015.

Examples

```
if(FALSE){
if (require(minfiData)) {
sheet <- read.450k.sheet(file.path(find.package("minfiData"),"extdata"), pattern = "csv$")
rgSet <- read.450k.exp(targets = sheet,extended = TRUE)
qcscore<-QCinfo(rgSet)
}}
```

`rm.outlier` *Filtering out outlier and/or low quality values*

Description

Setting outliers as missing value. Outlier was defined as value smaller than 3 times IQR from the lower quartile or larger than 3 times IQR from the upper quartile. If data quality information were provided, low quality data points will be set to missing first before looking for outliers. If specified, imputation will be performed using k-nearest neighbors method to impute all missing values.

Usage

```
rm.outlier(mat,byrow=TRUE,qcscore=NULL,detPthre=0.05,nbthre=3,
           rmcr=FALSE,rthre=0.05,cthre=0.05,impute=FALSE,
           imputebyrow=TRUE,...)
```

Arguments

`mat` An numeric matirx

`byrow` TRUE: Looking for outliers row by row, or FALSE: column by column.

`qcscore` If the data quality infomation (the output from function QCinfo) were provied, low quality data points as defined by detection p value threshold (detPthre) or number of bead threshold (nbthre) will be set to missing.

`detPthre` Detection P value threshold to define low qualitye data points, detPthre=0.05 in default.

`nbthre` Number of beads threshold define low qualitye data points, nbthre=3 in default.

`rmcr` TRUE: excluded rows and columns with too many missing values as defined by rthre and cthre. FALSE is in default

`rthre` Minimum of percentage of missing values for a row to be excluded

`cthre` Minimum of percentage of missing values for a column to be excluded

`impute` Whether to impute missing values. If TRUE, k-nearest neighbors methods will used for imputation. FALSE is in default. Warning: imputed values for multimodal distributed CpGs may not be correct.

`imputebyrow` TRUE: impute missing values using similar values in row, or FALSE: in column

`...` Arguments to be passed to the function impute.knn in R package "impute"

Value

An numeric matrix of same dimention as the input matrix.

Author(s)

Zongli Xu

References

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor, *ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip*. Nucleic Acids Research 2015.

Examples

```
if(FALSE){
if (require(minfiData)) {
sheet <- read.450k.sheet(file.path(find.package("minfiData"),"extdata"), pattern = "csv$")
rgSet <- read.450k.exp(targets = sheet,extended = TRUE)
qcscore<-QCinfo(rgSet)
mdat <- preprocessRaw(rgSet)
beta=getBeta(mdat, "Illumina")
#filter out outliers
b1=rm.outlier(beta)
#filter out low quality and outlier values
b2=rm.outlier(beta,qcscore=qcscore)
#filter out low quality and outlier values, remove rows and columns with too many missing values
b3=rm.outlier(beta,qcscore=qcscore,rmcr=TRUE)
#filter out low quality and outlier values, remove rows and columns with too many missing values, and then do impu
b3=rm.outlier(beta,qcscore=qcscore,rmcr=TRUE,impute=TRUE)
}}
```

Index