Package 'IMA'

February 23, 2012

Type Package

Title IMA (Illumina Methylation Analyzer)

Version 3.1.1

Date 2011-09-13

Depends R (>= 2.13.0), methods, utils, stats, WriteXLS, limma, MASS,bioDist, preprocessCore

Author Dan Wang, Li Yan, Qiang Hu, Dominic J Smiraglia, Song Liu

Maintainer Dan Wang<wangdan412@gmail.com>

Description IMA (Illumina Methylation Analyzer) is a package designed to automate the pipeline for exploratory analysis and summarization of site-level and region-level methylation changes in epigenetic studies utilizing the 450K DNA methylation microarray.

License GPL-2

LazyLoad yes

R topics documented:

IMA-package *IMA(Illumina Methylation Analyzer) is a package designed to automate the pipeline for exploratory analysis and summarization of site-level and region-level methylation changes in epigenetic studies utilizing the 450K DNA methylation microarray*

Description

IMA automates the tasks commonly required for the exploratory analysis and summarization of epigenetic data sets utilizing the 450K DNA methylation microarray. The package makes use of Illumina methylation annotation for region definition, as well as several Bioconductor packages for various preprocessing and differential testing steps. There are two major differences between IMA and existing packages for Infinium methylation microarray analysis. First, instead of analyzing CpG site only, IMA provide both site-level and region-level methylation analysis. Second, instead of manually calling individual R functions at the command line, IMA provides a pipeline which automate the tasks commonly required for the exploratory analysis and summarization of 450K microarray data. The user can either run the pipeline with default setting or specify optional routes in the parameter file of pipeline.

The main purpose of developing IMA package is to provide a range of commonly used analysis options for potential users to perform exploratory analysis and summarization of 450K microarray data in an automatic way. It is the best interest for the users to consult experienced bioinformatician/statistician about which specific analysis option should be chosen for their 450k microarray data. Written in open source R environment, it provides the flexibility for users to adopt, extend and customize the functionality for their specific needs. It can be used as an automatic pipeline to analyze specific regions as well as specific sites for downstream functional exploration and hypothesis generation.

Details

Package:	IMA
Type:	Package
Version:	2.0
Date:	2011-11-17
License:	GPL
Lazyload:	yes

Author(s)

Dan Wang, Li Yang, Qiang Hu,Dominic J Smiraglia,Song Liu

Maintainer: Dan Wang <wangdan412@gmail.com>

See Also

IMA.methy450R,IMA.methy450PP,regionswrapper,testfunc,sitetest,indexregionfunc,annotfunc

Examples

```
###Users specify the data paths/paramters in the pipeline.R before
#run the comman below
###R --no-save < pipeline.R##This will automately produce the result
#of methylation change
###the pipeline.R file could be found here
###http://ima.r-forge.r-project.org/
```

---

annotfunc *Annotate siteIDs/regionIDs of interest*

Description

Provide annotation information for a list of site IDs/region IDs of interest

Usage

```
annotfunc(listtoannot, fullannot, filteredannot, fullIndexannot,
filteredIndexannot, category = c("site", "region"))
```

Arguments

`listtoannot`	a list of site IDs or region IDs of interest
`fullannot`	the annotation file for all 450k probeIDs
`filteredannot`	the annotation file for the probeIDs after preprocessing filtering by *IMA.methy450PP* function
`fullIndexannot`	Index annotation file containing the full probeIDs/siteIDs for each region, and this object is available as *fullannotInd.rda* file from http://www.rforge.net/IMA/fullannotInd.rda
`filteredIndexannot`	Index annotation file containing the probeIDs/siteIDs for each region after preprocessing filtering, and this object is returned by the *IMA.methy450PP* function
`category`	category = "site" will indicate site-level annotation, i.e. the "listtoannot" is a list of site-level IDs. Otherwise, the listtoannot is a list of region-level IDs

Value

an annotation matrix for the list of site IDs/region IDs of interest

Author(s)

Dan Wang, Li Yan, Qiang Hu, Dominic J Smiraglia, Song Liu

Maintainer: Dan Wang <wangdan412@gmail.com>

See Also

IMA.methy450R,IMA.methy450PP,regionswrapper,testfunc,sitetest,indexregionfunc,

Examples

```
###Region level annotation
# >load("./fullannotInd")
# >beta = dataf@bmatrix ##dataf is returned by IMA.methy450PP()
# >betar <- indexregionfunc(indexlist=dataf@TSS1500Ind,beta=beta,indexmethod="mean")
# >TSS1500testALL = testfunc(eset = betar,testmethod="limma",Padj="BH",concov="OFF",
# +groupinfo = dataf2@groupinfo,gcase ="g1",gcontrol="g2",paired = TRUE)
# >TSS1500test = outputDMfunc(TSS1500testALL,rawpcut=0.05,adjustpcut=0.05,betadiffcut=0.14)
# >listtoannot = rownames(TSS1500test)[1:1000]
# >fullIndexannot = TSS1500Ind
# >temp = annotfunc(listtoannot,fullannot,filteredannot,fullIndexannot,
# >filteredIndexannot,category = "region")
  ####site level annotation
# >listtoannot = rownames(sitetest[1:1000,])
# >temp = annotfunc(listtoannot,fullannot,filteredannot,category = "site")
```

IMA.methy450PP *Data preprocessing and quality control*

Description

It allows user to choose several filtering steps or modify filtering criteria for specific quality control purpose. These include whether or not to filter probes based on detection P-value; whether or not to remove the loci from the X or Y chromosome, or both; whether or not to perform peak transformation, whether or not to transfer the raw β value using either arcsine square root or logit; whether or not to perform quantile normalization; whether or not to remove the loci containing missing β values; whether or not to filter out loci whose methylation level are measured by probes containing SNP(s) at/near the targeted CpG site. The user can choose the preprocessing routes and corresponding cutoffs in the argument of this function.

Usage

```
IMA.methy450PP(data, na.omit = TRUE,peakcorrection = FALSE,normalization = FALSE,
transfm =c(FALSE,"arcsinsqr", "logit"),samplefilterdetectP = c(FALSE, 1e-05),
samplefilterperc = 0.75, sitefilterdetectP = c(FALSE, 0.05),sitefilterperc
= 0.75,locidiff = c(FALSE, 0.01),locidiffgroup = list("g1","g2"), XYchrom = c(FALSE,"X","Y",
c("X","Y")), snpfilter = c(FALSE,"snpsites.txt"))
```

Arguments

data	an exprmethy450 class returned by the IMA.methy450R function
na.omit	if TRUE remove the sites containing missing value
peakcorrection	if TRUE, peak correction is performed based on the paper by sarah Dedeurwaerder et al.

`normalization` if TRUE, quantile normalization performed

`transfm` if FALSE, no transfm is performed, "arcsinsqr":arcsine square root transformation on β value is performed, "logit":logit transformation on β is performed

`samplefilterdetectP`
Default is false, i.e, no sample filtering by detection P-value. Otherwise, choose the cut off of detection P-value.

`samplefilterperc`
Keep the samples having at least specified percentage of sites with detection P-value less than the **samplefilterdetectP**.

`sitefilterdetectP`
Default is false, i.e. no site filtering by detection p-value. Otherwise, choose the cut off of detection P-value.

`sitefilterperc` Remove the sites having specified percentage of samples with detection P-value greater than **sitefilterdetectP**.

`locidiff` if FALSE, don't filter sites by the difference of group β value. Otherwise, remove the sites with β value difference greater than the specified value.

`locidiffgroup` specify which two groups are considered to check the loci difference if locidiff is not true

`XYchrom` if "X", remove the sites on chromosome X , if "Y", remove the sites on chromosome Y, if c("X","Y"), remove both on chromosome X and Y.

`snpfilter` if FALSE, keep the loci whose methylation level are measured by probes containing SNP(s) at/near the targeted CpG site; otherwise filter out the list of snp-containing loci by specifying the snp file name and location

Details

It allows user to choose several filtering steps or modify filtering criteria for specific quality control purpose. By default, IMA will filter out loci with missing β value, from the X chromosome or with median detection P-value greater than 0.05. Users can choose to filter out loci whose methylation level are measured by probes containing SNP(s) at/near the targeted CpG site. The option for sample level quality control is also provided. Although the raw β values will be analyzed as recommended by Illumina, users can choose Arcsine square root transformation when modeling the methylation level as the response in a linear model. Logit transformation is also available as an option. The default setting in IMA package for preprocessing is that no normalization will be performed. Although quantile normalization is available as an alternative preprocessing option, it should be pointed out that several literatures show that quantile normalization does not remove unwanted technical variation between samples in methylation analysis.

Value

This function will return a methy450batch class including:

`bmatrix` a matrix of β value for individual sites

`detectP` a matrix of detection p-value for individual sites

`annot` a matrix of annotation information for individual sites

`groupinfo` a list of sample ID and phenotype of each sample

TSS1500Ind	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the TSS1500 region of each gene
TSS200Ind	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the TSS200 region of each gene
UTR5Ind	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the 5' UTR region of each gene
EXON1Ind	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the 1st EXON of each gene
UTR3Ind	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the 3' UTR region of each gene
GENEBODYInd	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the gene body region of each gene
ISLANDInd	two lists of IDS - SID (site IDs) and PID (Position IDs) belonging to the ISLAND region of each UCSC_CPG_ISLAND
NSHOREInd	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the N Shore region of each UCSC_CPG_ISLAND
SSHOREInd	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the S Shore region of each UCSC_CPG_ISLAND
NSHELFInd	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the N Shelf region of each UCSC_CPG_ISLAND
SSHELFInd	two lists of IDs - SID (site IDs) and PID (Position IDs) belonging to the S Shelf region of each UCSC_CPG_ISLAND

Author(s)

Dan Wang, Li Yan, Qiang Hu, Dominic J Smiraglia, Song Liu

Maintainer: Dan Wang <wangdan412@gmail.com>

See Also

IMA.methy450R,regionswrapper,testfunc,sitetest,indexregionfunc,annotfunc

Examples

```
setwd(system.file("extdata", package="IMA"))
MethyFileName = "SampleMethFinalReport.txt"
PhenoFileName = "SamplePhenotype.txt"
data = IMA.methy450R(file = MethyFileName,columnGrepPattern=list(beta=".AVG_Beta",detectp=".Detection.Pval"),gro
dataf = IMA.methy450PP(data,na.omit = TRUE,normalization=FALSE,transfm =FALSE,peakcorrection = TRUE, samplefilter
```

IMA.methy450R *Load methylation 450k data*

Description

This function will load the methylation 450k data. The input information for the package consists basically of two files containing $beta$-value methylation data (including annotation) produced by BeadStudio or GenomeStudio, and sample phenotype data prepared by the user. *IMA.methy450R* function loads the input files with a single command described below and an *exprmethy450* object will be created, which includes the following features: *beta* value matrix, locus annotation, detection P-value and sample phenotype information. Besides, basic quality control information will be outputted in the QC.pdf, which include unsupervised sample clustering using all loci, boxplot for beta value of each sample, and barplot showing the percent of loci with detection P-value smaller than 1e-5 in each sample.

Usage

```
IMA.methy450R(fileName, columnGrepPattern = list(beta = ".AVG_Beta",
 detectp = ".Detection.Pval"), groupfile)
```

Arguments

`fileName`	This is the input file containing the 450k methylation data including β value, detection p-value and loci annotation information. This file can be be produced from the illumina BeadStudio or GenomeStudio software.
`columnGrepPattern`	Specify the columns with corresponding characters within the 450k methylation data file produced by illumina BeadStudio or GenomeStudio software. These columns will be loaded into the function.
`groupfile`	This is the input file containing the phenotype data prepared by the users.

Value

This function will return an *exprmethy450* class including these features: a β value matrix, an annotation matrix, a detection p-value matrix and a list of sample phenotype information. A QC.pdf file including the basic quality control information will be generated.

`bmatrix`	the β value matrix for each site in each sample
`detectP`	a matrix of detection p-value for each site in each sample
`annot`	a matrix of annotation information for each targeted site
`groupinfo`	a matrix of sample ID and phenotype information for each sample

Author(s)

Dan Wang, Li Yan, Qiang Hu,Dominic J Smiraglia,Song Liu

Maintainer: Dan Wang <wangdan412@gmail.com>

See Also

IMA.methy450PP,regionswrapper,testfunc,sitetest,indexregionfunc,annotfunc

Examples

```
setwd(system.file("extdata", package="IMA"))
MethyFileName = "SampleMethFinalReport.txt"
PhenoFileName = "SamplePhenotype.txt"
data = IMA.methy450R(fileName = MethyFileName,columnGrepPattern=list(beta=".AVG_Beta",detectp=".Detection.Pval")
```

`indexregionfunc` *Region-level Methylation Index Calculation*

Description

For each specific region of a gene, IMA will collect the loci within it and derive an index of overall region methylation value. Currently, there are three different index metrics implemented in IMA: mean, median, and Tukey's Biweight robust average. By default, the mean *beta* values will be used as the region's methylation index for further analysis.

Usage

```
indexregionfunc(indexlist, beta, indexmethod = c("mean", "median", "tbrm"))
```

Arguments

`indexlist`	specify the region(s) of interest
`beta`	β value matrix for the 450K microarray
`indexmethod`	The methods available to derive an index of overall region methylation value for each region. The options are mean, median, and Tukey's Biweight robust average.

Value

return a matrix of the index of overall methylation value for the region(s) of interest

Author(s)

Dan Wang; Li Yan; Qiang Hu;Dominic J Smiraglia; Song Liu

Maintainer: Dan Wang <wangdan412@gmail.com>

Examples

```
# >beta = dataf@bmatrix;##dataf is returned by IMA.methy450PP() function
# >betar =indexregionfunc(indexlist=dataf@TSS1500Ind,beta=beta,indexmethod="mean")
```

outputDMfunc *Output the differentially methylated sites/regions*

Description

This function will output the differential testing results with user-specified significance criteria.

Usage

```
outputDMfunc(out, rawpcut = 0.05, adjustpcut = 0.05, betadiffcut = 0.14)
```

Arguments

out	This is the output by function of *sitetest*, *testfunc* or *regionwrapper* with three columns for each site/region: p-value, adjusted p-value and β value difference.
rawpcut	either null or define the output cut-off for the raw p-value
adjustpcut	either null or define the output cut-off for the adjusted p-value
betadiffcut	either null or define the output cut-off for the β value difference between two groups.

Value

return a matrix with "P-Value", "Adjust Pval","beta-Difference" in separate columns for each site/region satisfying the specified significance criteria.

Author(s)

Dan Wang, Li Yan, Qiang Hu, Dominic J Smiraglia, Song Liu

Maintainer: Dan Wang <wangdan412@gmail.com>

See Also

IMA.methy450R,regionswrapper,IMA.methy450PP,testfunc,indexregionfunc

Examples

```
# > sitetest = sitetest(dataf,gcase="g2",gcontrol="g1",testmethod = "mean",
# + Padj="BH",rawpcut = NULL,adjustpcut =NULL,betadiffcut = NULL)
# > sitetest = outputDMfunc(sitetest,rawpcut = 0.05,adjustpcut =0.05,
# + betadiffcut = 0.14)
### The list of loci with adjusted p-value less than 0.05 and beta value
### difference at least 0.14 will be outputted
```

`regionswrapper` *Differential methylation testing on all 11 categories of annotated region*

Description

This function will output all the differential methylation testing results for the 11 categories of annotated regions, and store the results in the separated sheets of an excel file.

Usage

```
regionswrapper(dataf, indexmethod = c("mean", "median", "tbrm"), gcase = "g2",
gcontrol = "g1", testmethod = c("wilcox", "limma", "pooled", "satterthwaite"),
Padj = "BH", concov = c("OFF", "ON"), paired = FALSE, list11excel, list11Rdata,
rawpcut = NULL, adjustpcut = NULL, betadiffcut = NULL)
```

Arguments

`dataf`	a *methy450batch* class returned by the *IMA.methy450PP* function
`indexmethod`	the methods available to derive an index of overall methylation level for each region
`gcase`	the string names of case group/groups in the sample.txt file
`gcontrol`	the string names of control group/groups in the sample.txt file
`testmethod`	the available methods include wilcoxon rank-sum test (default), Student's t-test ("pooled" or "satterthwaite") or empirical Bayes statistics.
`Padj`	the methods available for multiple testing correction. Users could choose any method provided in the *p.adjust* function of R *stat* package.
`concov`	if "ON", covariates is continuous variable, and the linear model would be used for testing the association between methylation level and continuous phenotype (e.g., age).
`paired`	if TRUE, the test method would change to the corresponding paired-test method
`list11excel`	the name of output excel file containing the differential testing results for each annotated region
`list11Rdata`	the name of output Rdata containing the differential testing result for each annotated region
`rawpcut`	either null or the output cut-off for the raw p-value
`adjustpcut`	either null or the output cut-off for the adjusted p-value
`betadiffcut`	either null or the output cut-off for the $beta$ value difference between two groups

Value

Return an excel file containing the differential testing result for each annotated region as well as an Rdata file containing the differential testing result for each annotated region

Author(s)

Dan Wang, Li Yan, Qiang Hu, Dominic J Smiraglia, Song Liu

Maintainer: Dan Wang <wangdan412@gmail.com>

See Also

IMA.methy450PP,IMA.methy450R,testfunc,sitetest,indexregionfunc,annotfunc

Examples

```
setwd(system.file("extdata", package="IMA"))
MethyFileName = "SampleMethFinalReport.txt"
PhenoFileName = "SamplePhenotype.txt"
data = IMA.methy450R(fileName = MethyFileName,columnGrepPattern=list(beta=".AVG_Beta",detectp=".Detection.Pval"
dataf = IMA.methy450PP(data,na.omit = TRUE,normalization=FALSE,transfm = FALSE,samplefilterdetectP = 1e-5,samplef
regionswrapper(dataf,indexmethod="mean",gcase="g2",gcontrol="g1",testmethod="limma",Padj="BH",concov="OFF",pai
```

`sitetest` *Site-level Differential Methylation Analysis*

Description

For each specific site, Wilcoxon rank-sum test (default), Student's t-test ("pooled" or "satterthwaite") and empirical Bayes statistics are available for inference in differential testing. Robust linear regression is available as an option to infer methylation change associated with continuous variable (e.g., age). A variety of multiple testing correction algorithms is available, including conservative Bonferroni correction and more liberal false discovery rate control. Users can specify the significance criteria in the parameter file.

Usage

```
sitetest(dataf, gcase = "g2", gcontrol = "g1", testmethod = c("wilcox", "limma",
"pooled", "satterthwaite"), Padj = "BH", concov = "OFF", rawpcut = NULL,
adjustpcut = NULL, betadiffcut = NULL, paired = FALSE)
```

Arguments

`dataf`	a *methy450batch* class returned by the *IMA.methy450PP* function
`gcase`	the string names of case group/groups in the Phenotype.txt file
`gcontrol`	the string names of control group/groups in the Phenotype.txt file
`testmethod`	Wilcoxon rank-sum test (default), Student's t-test ("pooled" or "satterthwaite"), or empirical Bayes statistics.
`Padj`	The methods available for multiple testing correction. Users could choose any method provided in the *p.adjust* function of R *stat* package.

concov	if "ON", covariates is continuous variable, and the linear model would be used for testing the association between methylation level and continuous phenotype (e.g., age).
rawpcut	either null or the output cut-off for the raw p-value
adjustpcut	either null or the output cut-off for the adjusted p-value
betadiffcut	either null or the output cut-off for the $beta$ value difference between two groups
paired	if TRUE, the test method would change to the corresponding paired-test method

Value

return a matrix with "P-Value", "Adjust Pval", "beta-Difference" in separate columns for each site

Author(s)

Dan Wang, Li Yan, Qiang Hu, Dominic J Smiraglia, Song Liu

Maintainer: Dan Wang <wangdan412@gmail.com>

See Also

IMA.methy450R,regionswrapper,IMA.methy450PP,testfunc,indexregionfunc

Examples

```
# >sitetest = sitetest(dataf,gcase="g2",gcontrol="g1",testmethod = "mean",
# +Padj="BH",rawpcut = NULL,adjustpcut =NULL,betadiffcut = NULL,paired=FALSE)
```

---

testfunc *Differential methylation testing function*

Description

Choose the testing function for differential methylation inference

Usage

```
testfunc(eset, concov = c("ON", "OFF"), testmethod = c("wilcox","limma", "pooled",
"satterthwaite", "paired"), Padj = c("holm", "hochberg", "hommel","bonferroni",
"BH", "BY","fdr", "none"), groupinfo, gcase = "g1", gcontrol = "g2", paired = FALSE)
```

Arguments

eset	a $beta$ value matrix
concov	if "ON", covariates is continuous variable, and "grouplev" should be the continuous values.
testmethod	Wilcoxon rank-sum test (default), Student's t-test ("pooled" or "satterthwaite"), or empirical Bayes statistics.
Padj	The methods available for multiple testing correction. Users could choose any method provided in the *p.adjust* function of R *stat* package.
groupinfo	phenotype information for each sample
gcase	the string names of case group/groups in the sample.txt file
gcontrol	the string names of control group/groups in the sample.txt file
paired	if TRUE, the test method would change to the corresponding paired-test method

Value

return a matrix with "P-Value", "Adjust Pval","beta-Difference" in separate columns for each site or annotated region

Author(s)

Dan Wang, Li Yan, Qiang Hu, Dominic J Smiraglia, Song Liu

Maintainer: Dan Wang <wangdan412@gmail.com>

See Also

IMA.methy450R,regionswrapper,IMA.methy450PP,sitetest,indexregionfunc

Examples

```
# > beta = dataf@bmatrix;##dataf returned by IMA.methy450PP()
# > betar= indexregionfunc(indexlist=dataf@TSS1500Ind,beta=beta,indexmethod="mean")
# > TSS1500testALL = testfunc(eset =betar,testmethod="limma",Padj="BH",concov="OFF",
# + groupinfo = dataf2@groupinfo,gcase ="g1",gcontrol="g2",paired = TRUE)
```

Index