Package 'CpGassoc'

January 20, 2025

Type Package

Title Association Between Methylation and a Phenotype of Interest

Version 2.70

Date 2024-07-15

Author Barfield, R., Conneely, K., Kilaru,V

Maintainer R Barfield <barfieldrichard8@gmail.com>

Description

Is designed to test for association between methylation at CpG sites across the genome and a phenotype of interest, adjusting for any relevant covariates. The package can perform standard analyses of large datasets very quickly with no need to impute the data. It can also handle mixed effects models with chip or batch entering the model as a random intercept. Also includes tools to apply quality control filters, perform permutation tests, and create QQ plots, manhattan plots, and scatterplots for individual CpG sites.

Depends R (>= 2.10), nlme, methods

Repository CRAN

License GPL (>= 2)

NeedsCompilation no

Date/Publication 2024-07-16 09:00:05 UTC

Contents

CpGassoc-package 2
Class cpg.perm 3
cpg 5
cpg.assoc 7
cpg.combine 10
cpg.everything 11
cpg.GC 12
cpg.perm 13
cpg.qc 15
cpg.work 16
design 19

manhattan 20
manhattan.reflect 21
Other CpGassoc Functions 23
Sample Data CpGassoc 24
scatterplot 25

Index **27**

`CpGassoc-package` *Association Between Methylation and a Phenotype of Interest*

Description

Is designed to test for association between methylation at CpG sites across the genome and a phenotype of interest, adjusting for any relevant covariates. The package can perform standard analyses of large datasets very quickly with no need to impute the data. It can also handle mixed effects models with chip or batch entering the model as a random intercept. Also includes tools to apply quality control filters, perform permutation tests, and create QQ plots, manhattan plots, and scatterplots for individual CpG sites.

Details

Package: CpGassoc Type: Package Title: Association between Methylation and a phenotype of interest Version: 2.70 Date: 20

CpGassoc is a suite of R functions designed to perform flexible analyses of methylation array data. The two main functions are `cpg.assoc` and `cpg.perm`. `cpg.assoc` will perform an association test between the CpG sites and the phenotype of interest. Covariates can be added to the model, and can be continuous or categorical in nature. `cpg.assoc` allows users to set their own false discovery rate threshold, to transform the beta values to log(beta/(1-beta)), and to subset if required. cpg.assoc can also fit a linear mixed effects model with a single random effect to control for possible technical difference due to batch or chip. `cpg.assoc` uses the Holm method to determine significance. The user can also specify an FDR method to determine significance based on the function `p.adjust`. `cpg.perm` performs the same tasks as `cpg.assoc` followed by a permutation test on the data, repeating the analysis multiple times after randomly permuting the main phenotype of interest. The user can specify the seed and the number of permutations. If over one hundred permutations are performed QQ plots can be created with empirical confidence intervals based on the permuted t-statistics. For more information see `plot.cpg.perm`. For more information on how to perform `cpg.assoc` or `cpg.perm` see their corresponding help pages. CpGassoc can also perform quality control (see `cpg.qc`).

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.assoc cpg.combine cpg.perm cpg.work plot.cpg scatterplot manhattan plot.cpg.perm cpg.qc

Examples

```
#Using cpg.assoc:
data(samplecpg,samplepheno,package="CpGassoc")
results<-cpg.assoc(samplecpg,samplepheno$weight,large.data=FALSE)
results

##Using cpg.perm:
Testperm<-cpg.perm(samplecpg[1:200,],samplepheno$weight,data.frame(samplepheno$Dose),
                seed=2314,nperm=10,large.data=FALSE)
Testperm


#For more examples go to those two pages main help pages.
```

Class cpg.perm *Methods for object of class* "cpg.perm".

Description

Methods and extra functions for class "cpg.perm". plot.cpg.perm creates a QQ plot based on the association p-values or t-statistics from the function cpg.perm.

Usage

```
  ## S3 method for class 'cpg.perm'
plot(x, save.plot = NULL, file.type = "pdf", popup.pdf = FALSE,
 main.title = NULL, eps.size = c(5, 5), tplot = FALSE, perm.ci = TRUE, classic = TRUE,
  gc.p.val = FALSE, gcdisplay = FALSE, ...)

  ## S3 method for class 'cpg.perm'
summary(object,...)

  ## S3 method for class 'cpg.perm'
print(x,...)

  ## S3 method for class 'cpg.perm'
sort(x,decreasing,...)
```

Arguments

x	Output from cpg.perm. Of class "cpg.perm".
save.plot	Name of the file for the plot to be saved to. If not specified, plot will not be saved.

file.type	Type of file to be saved. Can either be "pdf" or "eps". Selecting file.type="eps" will result in publication quality editable postscript files that can be opened by Adobe Illustrator or Photoshop.
popup.pdf	TRUE or FALSE. If creating a pdf file, this indicates if the plot should appear in a popup window as well. If running in a cluster-like environment, best to leave FALSE.
main.title	Main title to be put on the graph. If NULL one based on the analysis will be used.
eps.size	Vector indicating the size of .eps file (if creating one). Corresponds to the options horizontal and height in the postscript function.
tplot	Logical. If TRUE, ordered t-statistics will be plotted against their expected quantities. If FALSE (default), -log(p) will be plotted. If indep is a class variable this option will be ignored.
perm.ci	Logical. If TRUE, the confidence intervals computed will be from the permutated values, otherwise will be based on the theoretical values.
classic	Logical. If TRUE, a classic qq-plot will be generated, with all p-values plotted against predicted values (including significant). If FALSE Holm-significant CpG sites will not be used to compute expected quantiles and will be plotted separately.
gc.p.val	Logical. If true, plot will use the genomic control adjusted p-values.
gcdisplay	Logical.If true, plot will display the genomic control value in the legend.
object	Output of class "cpg.perm" from "cpg.perm".
decreasing	logical. Should the sort be increasing or decreasing? Not available for partial sorting.
...	Arguments to be passed to methods, such as graphical parameters.

Note

Empirical confidence intervals will be computed only if there are a hundred or more permutations. Otherwise the theoretical confidence intervals will be plotted.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.perm plot.cpg scatterplot manhattan cpg.assoc

Examples

```
data(samplecpg,samplepheno,package="CpGassoc")

#The qq plot:
Testperm<-cpg.perm(samplecpg[1:300,],samplepheno$weight,seed=2314,nperm=10,large.data=FALSE)
plot(Testperm)
```

```
#The t-statistic plot from cpg.perm has confidence intervals since we were allowed
#to perform permutations on the T-values.
plot(Testperm,tplot=TRUE)
#If there was 100 or more permutations, there would be emperical confidence intervals.

#Getting an example of the non classic QQ plot
plot(Testperm,classic=FALSE)


###Now for Sort
head(sort(Testperm)$results)
head(Testperm$results)
```

---

cpg *Methods for object of class* "cpg"

Description

Methods and extra functions for class "cpg". plot.cpg creates a QQ plot based on the association p-values or t-statistics from the function cpg.assoc.

Usage

```
  ## S3 method for class 'cpg'
plot(x, save.plot = NULL, file.type = "pdf", popup.pdf = FALSE,
               tplot = FALSE, classic = TRUE,main.title = NULL, eps.size = c(5, 5),
                      gc.p.val = FALSE, gcdisplay = FALSE, ...)

  ## S3 method for class 'cpg'
summary(object,...)

  ## S3 method for class 'cpg'
print(x,...)

  ## S3 method for class 'cpg'
sort(x,decreasing,...)
```

Arguments

x	Output of class "cpg" from cpg.assoc or cpg.work.
save.plot	Name of the file for the plot to be saved to. If not specified, plot will not be saved.
file.type	Type of file to be saved. Can either be "pdf" or "eps". Selecting file.type="eps" will result in publication quality editable postscript files that can be opened by Adobe Illustrator or Photoshop.

popup.pdf	TRUE or FALSE. If creating a pdf file, this indicates if the plot should appear in a popup window as well. If running in a cluster-like environment, best to leave FALSE.
tplot	Logical. If TRUE, t-statistics will be plotted vs. their expected quantiles. If FALSE (default), -log(p) will be plotted. (Note: if class(x$indep)=='factor' this option will be ignored.)
classic	Logical. If TRUE, a classic qq-plot will be generated, with all p-values plotted against predicted values (including significant). If FALSE Holm-significant CpG sites will not be used to compute expected quantiles and will be plotted separately.
main.title	Main title to be put on the graph. If NULL one based on the analysis will be used.
eps.size	Vector indicating the size of .eps file (if creating one). Corresponds to the options horizontal and height in the postscript function.
gc.p.val	Logical. If true, plot will use the genomic control adjusted p-values.
gcdisplay	Logical.If true, plot will display the genomic control value in the legend.
object	Output of class "cpg" from cpg.assoc or cpg.work.
decreasing	logical. Should the sort be increasing or decreasing? Not available for partial sorting.
...	Arguments to be passed to methods, such as graphical parameters.

Value

sort.cpg returns an item of class "cpg" that is sorted by p-value. summary.cpg creates a qq-plot based on the data, and scatterplots or boxplots for the top sites.

Note

Plots with empirical confidence intervals based on permutation tests can be obtained from cpg.perm. See plot.cpg.perm for more info.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.perm cpg.assoc scatterplot manhattan plot.cpg.perm

Examples

```
##QQ Plot:
data(samplecpg,samplepheno,package="CpGassoc")
test<-cpg.assoc(samplecpg,samplepheno$weight,large.data=FALSE)
plot(test)
##t-statistic plot:
plot(test,tplot=TRUE)
```

```
#Getting our plot:
plot(test,classic=FALSE)


##Now an example of sort
head(sort(test)$results)

##Summary
summary(test)
```

---

cpg.assoc *Association Analysis Between Methylation Beta Values and Phenotype of Interest*

Description

Association Analysis Between Methylation Beta Values and Phenotype of Interest.

Usage

```
cpg.assoc(beta.val, indep, covariates = NULL, data = NULL, logit.transform = FALSE,
chip.id = NULL, subset = NULL, random = FALSE, fdr.cutoff = 0.05, large.data = FALSE,
fdr.method = "BH", logitperm = FALSE,return.data=FALSE)
```

Arguments

beta.val	A vector, matrix, or data frame containing the beta values of interest (1 row per CpG site, 1 column per individual).
indep	A vector containing the variable to be tested for association. cpg.assoc will evaluate the association between the beta values (dependent variable) and indep (independent variable).
covariates	A data frame consisting of additional covariates to be included in the model. covariates can also be specified as a matrix if it takes the form of a model matrix with no intercept column, or can be specified as a vector if there is only one covariate of interest. Can also be a formula(e.g. ~cov1+cov2).
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from the environment from which cpg.assoc is called.
logit.transform	logical. If TRUE, the logit transform of the beta values log(beta.val/(1-beta.val)) will be used. Any values equal to zero or one will be set to the next smallest or next largest value respectively; values <0 or >1 will be set to NA.
chip.id	An optional vector containing chip, batch identities, or other categorical factor of interest to the researcher. If specified, chip id will be included as a factor in the model.

subset	An optional logical vector specifying a subset of observations to be used in the fitting process.
random	Logical. If TRUE, 'chip.id' will be included in the model as a random effect, and a random intercept model will be fitted. If FALSE, 'chip.id' will be included in the model as an ordinary categorical covariate, for a much faster analysis.
fdr.cutoff	The desired FDR threshold. The default setting is .05. The set of CpG sites with FDR < 'fdr.cutoff' will be labeled as significant.
large.data	Logical. Enables analyses of large datasets. When large.data=TRUE, cpg.assoc avoids memory problems by performing the analysis in chunks. Note: this option no longer works within windows systems. Based on reading max memory allowable by the system. Defaults to False.
fdr.method	Character. Method used to calculate False Discovery Rate. Choices include any of the methods available in p.adjust(). The default method is "BH" for the Benjamini & Hochberg method.
logitperm	Logical. For internal use only.
return.data	Logical. cpg.assoc can return dataframes containing the the variable of interest, covariates, and the chip id (if present). Defaults to FALSE. Set to TRUE if plan on using the downstream scatterplot functions).

Details

cpg.assoc is designed to test for association between an independent variable and methylation at a number of CpG sites, with the option to include additional covariates and factors.
cpg.assoc assesses significance with the Holm (step-down Bonferroni) and FDR methods.

If class(indep)='factor', cpg.assoc will perform an ANOVA test of the variable conditional on the covariates specified. Covariates, if entered, should be in the form of a data frame, matrix, or vector. For example, covariates=data.frame(weight,age,factor(city)). The data frame can also be specified prior to calling cpg.assoc. The covariates should either be vectors or columns of a matrix or data.frame.
cpg.assoc is also designed to deal with large data sets. Setting large.data=TRUE will make cpg.assoc split up the data in chunks. Other option is to use cpg.combine and split up oneself.

Value

cpg.assoc will return an object of class "cpg". The functions summary and plot can be called to get a summary of results and to create QQ plots.

results	A data frame consisting of the t or F statistics and P-values for each CpG site, as well as indicators of Holm and FDR significance. CpG sites will be in the same order as the original input, but the sort() function can be used directly on the cpg.assoc object to sort CpG sites by p-value.
Holm.sig	A list of sites that met criteria for Holm significance.
FDR.sig	A data.frame of the CpG sites that were significant by the FDR method specified.

info	A data frame consisting of the minimum P-value observed, the FDR method that was used, the phenotype of interest, the number of covariates in the model, the name of the matrix or data frame the methylation beta values were taken from, the FDR cutoff value and whether a mixed effects analysis was performed.
indep	If return.data=T, the independent variable that was tested for association.
covariates	If return.data=T, data.frame or matrix of covariates, if specified (otherwise NULL).
chip	If return.data=T, chip.id vector, if specified (otherwise NULL).
coefficients	A data frame consisting of the degrees of freedom, and if object is continous the intercept effect adjusted for possible covariates in the model, the estimated effect size, and the standard error. The degrees of freedom is used in plot.cpg to compute the genomic inflation factors.

Author(s)

Barfield, R.; Conneely, K.; Kilaru,V.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.work cpg.perm plot.cpg scatterplot cpg.combine manhattan plot.cpg.perm sort.cpg.perm sort.cpg cpg.qc cpg.GC

Examples

```
 # Sample output from CpGassoc
data(samplecpg,samplepheno,package="CpGassoc")
results<-cpg.assoc(samplecpg,samplepheno$weight,large.data=FALSE)
results
#Analysis with covariates. There are multiple ways to do this. One can define the
#dataframe prior or do it in the function call.
test<-cpg.assoc(samplecpg,samplepheno$weight,data.frame(samplepheno$Distance,
samplepheno$Dose),large.data=FALSE)
# or
covar<-data.frame(samplepheno$Distance,samplepheno$Dose)
test2<-cpg.assoc(samplecpg,samplepheno$weight,covar,large.data=FALSE)


#Doing a mixed effects model. This does take more time, so we will do a subset of
#the samplecpg
randtest<-cpg.assoc(samplecpg[1:10,],samplepheno$weight,chip.id=samplepheno$chip,
random=TRUE,large.data=FALSE)
```

cpg.combine *Combine various objects of class "cpg"*

Description

Takes a list containing objects of class "cpg" and combines them into one cpg item. Assumes that there are no repeated CpG sites between the various objects (i.e. analysis wasn't performed on the same sites twice).

Usage

```
cpg.combine(allvalues, fdr.method="BH",fdr.cutoff=.05,return.data=FALSE)
```

Arguments

allvalues	A list containing the "cpg" objects that are desired to be consolidated.
fdr.method	FDR method that user wants to use. For options see the cpg.assoc help page.
fdr.cutoff	The desired FDR threshold. The default setting is .05. The set of CpG sites with FDR < fdr.cutoff will be labeled as significant.
return.data	Logical. cpg.assoc can return dataframes containing the the variable of interest, covariates, and the chip id (if present). Defaults to FALSE. Set to TRUE if plan on using the downstream scaterrplot functions).

Value

info.data	An object of class "cpg" that is the consolidated version of the objects of class cpg that were passed in.

Note

This is designed to be used by cpg.assoc when it does analysis on large data sets or by the user if they split up the analysis by chromosome or some other such partition.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.assoc cpg.perm cpg.work plot.cpg scatterplot manhattan plot.cpg.perm

Examples

```
data(samplecpg,samplepheno,package="CpGassoc")
test1<-cpg.assoc(samplecpg[1:100,],samplepheno$weight,large.data=FALSE)
test2<-cpg.assoc(samplecpg[101:200,],samplepheno$weight,large.data=FALSE)
bigtest<-list(test1,test2)
overall<-cpg.combine(bigtest)
overall
```

---

cpg.everything *Multi-Task function*

Description

A function designed to do a group of smaller functions required for the cpg.assoc

Usage

```
cpg.everything(x, ...)
```

Arguments

x	Just a generic object
...	Arguments to be passed to methods.

Details

A function created to do a bunch of much smaller tasks within **CpGassoc** based on the class of x.

Note

cpg.everything is designed to perform a multitude of smaller tasks for cpg.assoc that do not warrant a full help page. These include:warnings, getting the name of the independent variable, designing the random function to be used, and getting the names for the values returned.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.assoc cpg.perm plot.cpg scatterplot manhattan plot.cpg.perm

Examples

```
#Has four methods:character,complex, numeric/matrix, and logical
#They correspond to getting the indep variable name, warnings, getting the random function,
#and getting the names for the values returned. For the design of these functions see
#the R code.
```

cpg.GC *cpg.GC and methods for output of function*

Description

cpg.GC accepts an object of class "cpg.perm" or "cpg" and returns information regarding Holm and FDR-significance of the GC (genomic control) adjusted test statistics. For "cpg.perm" will return permutation p-values based on the GC-adjusted values from each permutation.

Usage

```
   cpg.GC(x)

  ## S3 method for class 'cpg.gc'
print(x,...)

  ## S3 method for class 'cpg.perm.gc'
print(x,...)
```

Arguments

x	Object of class "cpg.perm" or "cpg".
...	Arguments to be passed to methods, such as graphical parameters.

Details

cpg.GC will display the number of Holm and FDR-significant sites using the genomic control adjusted p-values test statistics. It will also display the estimated genomic control inflation factor.

Value

cpg.GC returns an object of class "cpg.gc" or "cpg.perm.gc"

gc.results	Matrix consisting of GC-adjusted test statistics for each CpG site. Similar to the results output of cpg.assoc.
gc.info	Data frame with information on the number of Holm and FDR significant sites. Will also have the genomic control inflation estimate. Objects from "cpg.perm" will also have information concerning the permutation p-values.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.work cpg.perm plot.cpg scatterplot cpg.combine manhattan plot.cpg.perm sort.cpg.perm sort.cpg cpg.qc cpg.assoc

Examples

```
data(samplecpg,samplepheno,package="CpGassoc")
results<-cpg.assoc(samplecpg,samplepheno$weight,large.data=FALSE)

cpg.GC(results)
##If the genomic inflation factor is less than one there is no need for adjustment
```

---

cpg.perm — *Perform a Permutation Test of the Association Between Methylation and a Phenotype of Interest*

Description

Calls cpg.assoc to get the observed P-values from the study and then performs a user-specified number of permutations to calculate an emperical p-value. In addition to the same test statistics computed by cpg.assoc, cpg.perm will compute the permutation p-values for the observed p-value, the number of Holm significant sites, and the number of FDR significant sites.

Usage

```
cpg.perm(beta.values, indep, covariates = NULL, nperm, data = NULL, seed = NULL,
logit.transform = FALSE, chip.id = NULL, subset = NULL, random = FALSE,
fdr.cutoff = 0.05, fdr.method = "BH",large.data=FALSE,return.data=FALSE)
```

Arguments

beta.values	A vector, matrix, or data frame containing the beta values of interest (1 row per CpG site, 1 column per individual).
indep	A vector containing the main variable of interest. cpg.assoc will evaluate the association between indep and the beta values.
covariates	A data frame consisting of the covariates of interest. covariates can also be a matrix if it is a model matrix minus the intercept column. It can also be a vector if there is only one covariate of interest. Can also be a formula(e.g. ~cov1+cov2).
nperm	The number of permutations to be performed.
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from the environment from which cpg.perm is called.
seed	The required seed for random number generation. If not input, will use R's internal seed.

`logit.transform`
: logical. If `TRUE`, the logit transform of the beta values log(beta.val/(1-beta.val)) will be used. Any values equal to zero or one will be set to the next smallest or next largest value respectively; values <0 or >1 will be set to NA.

`chip.id`
: An optional vector containing chip, batch identities, or other categorical factor of interest to the researcher. If specified, chip id will be included as a factor in the model.

`subset`
: An optional logical vector specifying a subset of observations to be used in the fitting process.

`random`
: logical. If `TRUE`, the 'chip.id' will be processed as a random effect, and a random intercept model will be fitted.

`fdr.cutoff`
: The threshold at which to compare the FDR values. The default setting is .05. Any FDR values less than .05 will be considered significant.

`fdr.method`
: Character. Method used to calculate False Discovery Rate. Can be any of the methods listed in `p.adjust`. The default method is "BH" for the Benjamini & Hochberg method.

`large.data`
: Logical. Enables analyses of large datasets. When `large.data=TRUE`, `cpg.assoc` avoids memory problems by performing the analysis in chunks. Note: this option no longer works within windows systems. Based on reading max memory allowable by the system. Defaults to False.

`return.data`
: Logical. cpg.assoc can return dataframes containing the the variable of interest, covariates, and the chip id (if present). Defaults to FALSE. Set to TRUE if plan on using the downstream scatterplot functions).

Value

The item returned will be of class `"cpg.perm"`. It will contain all of the values of class cpg (`cpg.assoc`) and a few more:

`permutation.matrix`
: A matrix consisting of the minimum observed P-value, the number of Holm significant CpG sites, and the number of FDR significant sites for each permutation.

`gc.permutation.matrix`
: Similar to the permutation.matrix only in relation to the genomic control adjusted p-values.

`perm.p.values`
: A data frame consisting of the permutation P-values, and the number of permutations performed.

`perm.tstat`
: If one hundred or more permutations were performed and indep is a continuous variable, consists of the quantile .025 and .975 of observed t-statistics for each permutation, ordered from smallest to largest. perm.tstat is used by `plot.cpg.perm` to compute the confidence intervals for the QQ plot of t-statistics. Otherwise `NULL`.

`perm.pval`
: If one hundred or more permutations were performed, consists of the observed p-values for each permutation, ordered from smallest to largest. perm.pval is usd by `plot.cpg.perm` to compute the confidence intervals for the QQ plot of the p-values. Otherwise NULL.

Author(s)

Barfield, R.; Conneely, K.; Kilaru,V.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.assoc cpg.work plot.cpg scatterplot cpg.combine manhattan plot.cpg.perm sort.cpg.perm sort.cpg cpg.qc

Examples

```
##Loading the data
data(samplecpg,samplepheno,package="CpGassoc")

#Performing a permutation 10 times
Testperm<-cpg.perm(samplecpg[1:200,],samplepheno$weight,seed=2314,nperm=10,large.data=FALSE)
Testperm
#All the contents of CpGassoc are included in the output from Testperm

#summary function works on objects of class cpg.perm
summary(Testperm)
```

---

cpg.qc *Performs quality control on Illumina data.*

Description

cpg.qc is designed to perform quality control on Illumina data prior to analysis. In addition to the matrix of beta values, this function requires as input matrices of Signal A, Signal B, and detection p-values. It can also set to NA datapoints with detection p-values exceeding a user-specified cutoff, and can remove samples or sites that have a missing rate above a user-specified value. Finally, users can opt to compute beta values as M/(U+M) or M/(U+M+100). Illumina suggested previous array versions use a 2000 signal value as a possible cutoff, but this is not appropriate for EPICv2 and beyond. Older versions of CpGassoc use the 2000 signal value cutoff. Default is now set to 0.

Usage

```
cpg.qc(beta.orig,siga,sigb,pval,p.cutoff=.001,cpg.miss=NULL,sample.miss=NULL,
constant100=FALSE,sig.return=FALSE,low.sig.remove=FALSE,low.sig.cutoff=0)
```

Arguments

beta.orig	The original beta values matrix.
siga	The unmethylated signals matrix.
sigb	The methylated signals matrix.

pval	A matrix of detection p-values. pval should have the same dimension as the beta values and signals: one row for each site and one column for each individual.
p.cutoff	The user-specified cutoff for detection p-values (default=.001).
cpg.miss	Optional cutoff value. If specified, cpg.qc will remove cpg sites where the proportion of missing values exceeds this cutoff.
sample.miss	Optional cutoff value. If specified, cpg.qc will remove samples where the proportion of missing values exceeds this cutoff.
constant100	Logical. If true, the new beta values will be calculated as M/(U+M+100); if false (default) they will be calculated as M/(U+M).
sig.return	Logical. If true, cpg.qc returns a list with the betas and the qced signal data as well.
low.sig.remove	Logical. If true, cpg.qc will remove samples that have low intensity (mean signal intensity less than half of the overall median or low.sig.cutoff).
low.sig.cutoff	Numeric. Value to be used. by low.sig.cutoff

Details

It is important that all the matrices listed above ('pval', 'siga', 'sigb', 'beta.orig') are ordered similarly with respect to samples and CpG sites.

Value

cpg.qc returns a new matrix of beta values that has been subjected to the specified quality control filters. This matrix can be input directly into cpg.assoc.

Author(s)

Barfield, R.; Conneely, K.; Kilaru,V.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.work cpg.perm plot.cpg scatterplot cpg.combine cpg.assoc

---

cpg.work *Does the analysis between the CpG sites and phenotype of interest*

Description

Association Analysis Between Methylation Beta Values and Phenotype of Interest. This function contains the code that does the brunt of the work for cpg.assoc and cpg.perm.

Usage

```
cpg.work(beta.values, indep, covariates = NULL, data = NULL, logit.transform = FALSE,
chip.id = NULL, subset = NULL, random = FALSE, fdr.cutoff = 0.05, callarge = FALSE,
fdr.method = "BH", logitperm = FALSE,big.split=FALSE,return.data=FALSE)
```

Arguments

beta.values	A vector, matrix, or data frame containing the beta values of interest (1 row per CpG site, 1 column per individual).
indep	A vector containing the main variable of interest. cpg.work will evaluate the association between indep and the beta values.
covariates	A data frame consisting of the covariates of interest. covariates can also be a matrix if it is a model matrix minus the intercept column. It can also be a vector if there is only one covariate of interest. Can also be a formula (e.g. ~cov1+cov2).
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from the environment from which cpg.work is called.
logit.transform	logical. If TRUE, the logit transform of the beta values log(beta.val/(1-beta.val)) will be used. Any values equal to zero or one will be set to the next smallest or next largest value, respectively; values <0 or >1 will be set to NA.
chip.id	An optional vector containing chip, batch identities, or other categorical factor of interest to the researcher. If specified, chip id will be included as a factor in the model.
subset	an optional logical vector specifying a subset of observations to be used in the fitting process.
random	logical. If TRUE, the 'chip.id' will be included in the model as a random effect, and a random intercept model will be fitted. If FALSE, 'chip.id' will be included in the model as an ordinary categorical covariate, for a much faster analysis.
fdr.cutoff	The threshold at which to compare the FDR values. The default setting is .05. Any FDR values less than .05 will be considered significant.
callarge	logical. Used by cpg.assoc when it calls cpg.work. If TRUE it means that beta.values is actually split up from a larger data set and that memory.limit may be a problem. This tells cpg.work to perform more rm() and gc() to clear up space.
fdr.method	Character.Method used to calculate False Discovery Rate. Can be any of the methods listed in p.adjust. The default method is "BH" for the Benjamini & Hochberg method.
logitperm	Passes from cpg.perm when permutation test is performed. Stops from future checks involving the logistic transformation.
big.split	Passes from cpg.assoc. Internal flag to inform cpg.work that the large data did not need to be split up.
return.data	Logical. cpg.assoc can return dataframes containing the the variable of interest, covariates, and the chip id (if present). Defaults to FALSE. Set to TRUE if plan on using the downstream scatterplot functions).

Details

cpg.work does the analysis between the methylation and the phenotype of interest. It is called by cpg.assoc to do the brunt of the work. It can be called itself with the same input as cpg.assoc, it just cannot handle large data sets.

Value

cpg.work will return an object of class "cpg". The functions summary and plot can be called to get a summary of results and to create QQ plots. The output is in the same order as the original input. To sort it by p-value, use the sort function.

results	A data frame consisting of the statistics and P-values for each CpG site. Also has the adjusted p-value based on the fdr.method and whether the site was Holm significant.
Holm.sig	A list of sites that met criteria for Holm significance.
FDR.sig	A data.frame of the sites that were FDR significant by the fdr method.
info	A data frame consisting of the minimum P-value observed, the fdr method used, what the phenotype of interest was, and the number of covariates in the model.
indep	If return.data=T, the independent variable that was tested for association.
covariates	If return.data=T, data.frame or matrix of covariates, if specified (otherwise NULL).
chip	If return.data=T, chip.id vector, if specified (otherwise NULL).
coefficients	A data frame consisting of the degrees of freedom, and if object is continuous the intercept effect adjusted for possible covariates in the model, the estimated effect size, and the standard error. The degrees of freedom is used in plot.cpg to compute the genomic inflation factors.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.perm cpg.assoc plot.cpg scatterplot cpg.combine manhattan plot.cpg.perm sort.cpg.perm sort.cpg cpg.qc cpg.GC

Examples

```
##See the examples listed in cpg.assoc for ways in which to use cpg.work.
##Just change the cpg.assoc to cpg.work.
```

design *Create full and reduced design matrices for the cpg.assoc function.*

Description

Designed to be used by cpg.assoc and cpg.perm. Creates a full and reduced design matrices.

Usage

```
design(covariates, indep, chip.id, random)
```

Arguments

covariates	A data frame consisting of the covariates of interest. covariates can also be a matrix if it is a model matrix minus the intercept column. It can also be a vector if there is only one covariate of interest. If no covariates must be specified as NULL.
indep	A vector containing the main variable of interest. cpg.assoc will evaluate the association between indep and the beta values.
chip.id	An optional vector containing chip or batch identities. If specified, chip.id will be included as a factor in the model.
random	Is the model going to be a mixed effects. If so, chip.id will not be included in the design matrices.

Value

Returns a list containing the full and reduced design matrices.

full	The full design matrix.
reduced	The reduced design matrix.

Note

The design function is designed to be used exclusively by the cpg.assoc and cpg.perm functions.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.assoc cpg.perm cpg.work plot.cpg scatterplot cpg.combine manhattan plot.cpg.perm

Examples

```
data(samplecpg,samplepheno,package="CpGassoc")
#Example where there are covariates:
covar<-data.frame(samplepheno$weight,samplepheno$Distance)
test<-design(covar,samplepheno$SBP,samplepheno$chip,FALSE)
dim(test$full)
dim(test$reduced)
test$reduced[1:5,1:5]
test$full[1:5,1:5]
#When no covariates or chip.id:
test2<-design(NULL,samplepheno$SBP,NULL,FALSE)
dim(test2$full)
dim(test2$reduced)
```

---

manhattan *Create a manhattan plot*

Description

This function will produce a manhattan plot for the observed P-values from a object of class "cpg" or "cpg.perm".

Usage

```
manhattan(x, cpgname, chr, pos, save.plot = NULL, file.type="pdf",
popup.pdf = FALSE, eps.size = c(15, 5), main.title = NULL, cpg.labels = NULL,
chr.list = NULL, color.list = NULL, point.size = NULL, ...)
```

Arguments

x	Object of class "cpg" or "cpg.perm".
cpgname	A vector consisting of the labels for each CpG site.
chr	A vector consisting of the chromosome number for each CpG site.
pos	The map position of each CpG site within its chromosome.
save.plot	Name of the file for the plot to be saved to. If not specified, plot will not be saved.
file.type	Type of file to be saved. Can either be "pdf" or "eps". Selecting file.type="eps" will result in publication quality editable postscript files that can be opened by Adobe Illustrator or Photoshop.
popup.pdf	TRUE or FALSE. If creating a pdf file, this indicates if the plot should appear in a popup window as well. If running in a cluster-like environment, best to leave FALSE.
eps.size	Vector indicating the size of .eps file (if creating one). Corresponds to horizontal and height.
main.title	Main title to be put on the graph. If NULL one based on the analysis will be used.

cpg.labels	A character scalar of either "FDR" or "HOLM" which will label the significant sites on the manhattan plot.
chr.list	A vector listing the chromosomes to be plotted (all available chromosomes are plotted by default). The X and Y chromosomes can be denoted by 23 and 24.
color.list	A vector of custom colors to be used for each chromosomes in the manhattan plot.
point.size	The size of the points in the manhattan plot, if NULL, default to our default, where significant probes have different sizes.
...	Arguments to be passed to methods, such as graphical parameters.

Note

'cpgname', 'chr', and 'pos' must be sorted in the same order, so that the first cpgname[1] corresponds to chr[1] and pos[1], and so on.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.perm plot.cpg scatterplot cpg.assoc plot.cpg.perm manhattan.reflect

Examples

```
#Doing a Manhattan plot. First load the data:
data(samplecpg,samplepheno,annotation,package="CpGassoc")

examplemanhat<-cpg.assoc(samplecpg,samplepheno$Disease,large.data=FALSE)

manhattan(examplemanhat,annotation$TargetID,annotation$CHR,annotation$MAPINFO)
```

---

manhattan.reflect *Create a Reflective Manhattan plot*

Description

This function will produce a reflective manhattan plot for the observed P-values from an object of class "cpg" or "cpg.perm". The original analysis needs to be performed on a continuous variable (need T-statistics).

Usage

```
manhattan.reflect(x, cpgname, chr, pos, save.plot = NULL, file.type="pdf",
popup.pdf = FALSE, eps.size = c(15, 5), main.title = NULL, cpg.labels = NULL,
chr.list = NULL, color.list = NULL, fdr.cutoff=NULL, point.size=NULL, ...)
```

Arguments

x	Object of class "cpg" or "cpg.perm".
cpgname	A vector consisting of the labels for each CpG site.
chr	A vector consisting of the chromosome number for each CpG site.
pos	The map position of each CpG site within its chromosome.
save.plot	Name of the file for the plot to be saved to. If not specified, plot will not be saved.
file.type	Type of file to be saved. Can either be "pdf" or "eps". Selecting file.type="eps" will result in publication quality editable postscript files that can be opened by Adobe Illustrator or Photoshop.
popup.pdf	TRUE or FALSE. If creating a pdf file, this indicates if the plot should appear in a popup window as well. If running in a cluster-like environment, best to leave FALSE.
eps.size	Vector indicating the size of .eps file (if creating one). Corresponds to horizontal and height.
main.title	Main title to be put on the graph. If NULL one based on the analysis will be used.
cpg.labels	A character scalar of either "FDR" or "HOLM" which will label the significant sites on the manhattan plot.
chr.list	A vector listing the chromosomes to be plotted (all available chromosomes are plotted by default). The X and Y chromosomes can be denoted by 23 and 24.
color.list	A vector of custom colors to be used for each chromosomes in the manhattan plot.
fdr.cutoff	A numeric scalar between 0 and 1 to indicate what to consider FDR significant. Defaults to NULL.
point.size	The size of the points in the manhattan plot, if NULL, default to our default, where significant probes have different sizes.
...	Arguments to be passed to methods, such as graphical parameters.

Note

'cpgname', 'chr', and 'pos' must be sorted in the same order, so that the first cpgname[1] corresponds to chr[1] and pos[1], and so on.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.perm plot.cpg scatterplot cpg.assoc plot.cpg.perm manhattan

Examples

```
#Doing a Manhattan plot. First load the data:
data(samplecpg,samplepheno,annotation,package="CpGassoc")

examplemanhat<-cpg.assoc(samplecpg,samplepheno$Disease,large.data=FALSE)

manhattan.reflect(examplemanhat,annotation$TargetID,annotation$CHR,annotation$MAPINFO)
```

Other CpGassoc Functions

Information on miscellaneous other functions

Description

`cpg.length` compares the dimensions of the covariates, the independent phenotype, chip, and the matrix of beta values from `cpg.assoc` or `cpg.perm`. If the number of individuals does not match up. Stops the code.

`pointsizefunction` simply scales the size of the points for the qq-plot for `plot.cpg` and `plot.cpg.perm`

`cpgassocsummary` is used by `cpg.work` when the phenotypes contains such small amount of variance such that matrix methods can not be used. Takes an object of class `"aov"` or `"mlm"` and gets the test statistics from these objects while avoiding doing a loop or the list structures which is typically done if one does summary on one of these objects.

Usage

```
cpg.length(indep, numpatients, covariates, chip.id)

pointsizefunction(x)

cpgassocsummary(object)
```

Arguments

`indep`	object `indep` from `cpg.assoc`, `cpg.work` or `cpg.perm`.
`numpatients`	The number of column of the object beta.values in `cpg.work` or `cpg.perm` or number of columns of beta.val in `cpg.assoc`
`covariates`	object covariate from `cpg.assoc`, `cpg.work` or `cpg.perm`.
`chip.id`	object `chip.id` from `cpg.assoc`, `cpg.work` or `cpg.perm`.
`x`	the p-values from the object `"cpg"` or `"cpg.perm"`
`object`	An object of `"aov"` or `"mlm"`

Details

`cpg.length` stops the functions if the dimensions do not add up. Used Internally by **CpGassoc**

`pointsizefunction` simply returns a vector of equal length of x to be used in plot for the cex option.

`cpgassocsummary` returns a matrix of columns two and three of the `results` value of an `"cpg"` or `"cpg.perm"` object. This is for the cpg sites with non-missing data

See Also

`cpg.perm` `cpg.assoc` `scatterplot` `manhattan` `plot.cpg.perm`

Sample Data CpGassoc *Sample data from* **CpGassoc**

Description

samplecpg Matrix containing sample (fake) methylation data for 258 individuals over 1228 CpG sites

samplepheno Matrix with phenotype info for the 258 individuals

annotation Matrix with annotation information for the 1228 CpG sites

Usage

```
data(samplecpg)
```

Format

samplecpg Matrix, dimensions 1228 x 258

samplepheno Matrix, dimensions 258 x 6. Phenotype info on `"Dose"`, `"SBP"`, `"Distance"`, `"weight"`, `"Disease"`, and `"chip"`

annotation Matrix, dimensions 1228 x 3. Header names are `"TargetID"`, `"CHR"`, and `"MAPINFO"`

See Also

`cpg.perm` `cpg.assoc` `scatterplot` `manhattan` `plot.cpg.perm`

Examples

```
##See help pages for other functions for usage of these datasets
```

`scatterplot` *Plot beta values of individual CpG sites against the independent variable.*

Description

Plot beta values of individual CpG sites against the independent variable. Can create scatterplots and boxplots. If scatterplots the intercept will be adjusted for any covariates that were included in the model. Only available if return.data was set to T.

Usage

```
scatterplot(x, cpg.rank = NULL, cpg.name = NULL, save.plot = NULL, file.type="pdf",
eps.size = c(5, 5), popup.pdf = FALSE, beta.values = NULL,
user.indep=NULL,main.title=NULL, ...)
```

Arguments

`x`	Object of class `"cpg"` or `"cpg.perm"`.
`cpg.rank`	A vector listing the rank of sites to be plotted. The rank is based on the ordered p-values.
`cpg.name`	A character vector containing the names of CpG sites to be plotted against the phenotype of interest. This option is ignored if 'cpg.rank' is specified.
`save.plot`	Prefix of the filename for the plot(s) to be saved to. If specified, plot filenames will be created by appending this prefix to either cpg.rank or cpg.name. If not specified, plot will not be saved.
`file.type`	Type of file to be saved. Can either be `"pdf"` or `"eps"`. Selecting `file.type="eps"` will result in publication quality editable postscript files that can be opened by Adobe Illustrator or Photoshop.
`eps.size`	Vector indicating the size of .eps file (if creating one). Corresponds to horizontal and height.
`popup.pdf`	`TRUE` or `FALSE`. If creating a pdf file, this indicates if the plot should appear in a popup window as well. If running in a cluster-like environment, best to leave `FALSE`.
`beta.values`	If the object has been renamed (i.e. `x$info$betainfo` is no longer in `ls(.GlobalEnv)`) then specify the new object here.
`user.indep`	Default `NULL`. If return.data=F in run, scatterplot will not work. Pass in samplepheno here. Must be in same order as samplecpg.
`main.title`	Main title to be put on the graph. If `NULL` one based on the analysis will be used.
`...`	Arguments to be passed to methods, such as graphical parameters.

Details

An unlimited number of CpG sites can be selected for plotting by specifying either 'cpg.rank' or 'cpg.name', as shown in the Examples below. Note that only one of these options is needed; if both are entered, 'cpg.rank' will be used.

Author(s)

Barfield, R.; Kilaru,V.; Conneely, K.
Maintainer: R. Barfield: <barfieldrichard8@gmail.com>

See Also

cpg.assoc plot.cpg manhattan cpg.perm plot.cpg.perm

Examples

```
#Load the data:
data(samplecpg,samplepheno,package="CpGassoc")

test<-cpg.assoc(samplecpg,samplepheno$weight,large.data=FALSE,return.data=TRUE)
##Using rank, will plot the top three sites in order of significance:
scatterplot(test,c(1:3),user.indep=sampleoheno$weight)
##Using name, specify three sites:
scatterplot(test,cpg.name=c("CpG1182","CpG1000","CpG42"))

##Plotting something that is categorical in nature:
test2<-cpg.assoc(samplecpg[1:200,],factor(samplepheno$Disease),large.data=FALSE,return.data=TRUE)
scatterplot(test2,c(2),beta.values=samplecpg[1:200,],
user.indep=sampleoheno$weight)
```

Index

∗ **datasets**
Sample Data CpGassoc, 24
∗ **package**
CpGassoc-package, 2

annotation *(*Sample Data CpGassoc*)*, 24

Class cpg.perm, 3
cpg, 5
cpg.assoc, *3, 4*, *6*, 7, *10, 11*, *13–16*, *18, 19*, *21, 22*, *24*, *26*
cpg.combine, *3*, *9*, 10, *13*, *15, 16*, *18, 19*
cpg.everything, 11
cpg.GC, *9*, 12, *18*
cpg.length *(*Other CpGassoc Functions*)*, 23
cpg.perm, *3, 4*, *6*, *9–11*, *13*, 13, *16*, *18, 19*, *21, 22*, *24*, *26*
cpg.qc, *2, 3*, *9*, *13*, *15*, 15, *18*
cpg.work, *3*, *9, 10*, *13*, *15, 16*, 16, *19*
CpGassoc *(*CpGassoc-package*)*, 2
CpGassoc-package, 2
cpgassocsummary *(*Other CpGassoc Functions*)*, 23

design, 19

manhattan, *3, 4*, *6*, *9–11*, *13*, *15*, *18, 19*, 20, *22*, *24*, *26*
manhattan.reflect, *21*, 21

Other CpGassoc Functions, 23

p.adjust, *14*, *17*
plot.cpg, *3, 4*, *9–11*, *13*, *15, 16*, *18, 19*, *21, 22*, *26*
plot.cpg *(*cpg*)*, 5
plot.cpg.perm, *2, 3*, *6*, *9–11*, *13*, *15*, *18, 19*, *21, 22*, *24*, *26*
plot.cpg.perm *(*Class cpg.perm*)*, 3
pointsizefunction *(*Other CpGassoc Functions*)*, 23
print.cpg *(*cpg*)*, 5
print.cpg.gc *(*cpg.GC*)*, 12
print.cpg.perm *(*Class cpg.perm*)*, 3
print.cpg.perm.gc *(*cpg.GC*)*, 12

Sample Data CpGassoc, 24
samplecpg *(*Sample Data CpGassoc*)*, 24
samplepheno *(*Sample Data CpGassoc*)*, 24
scatterplot, *3, 4*, *6*, *9–11*, *13*, *15, 16*, *18, 19*, *21, 22*, *24*, 25
sort.cpg, *9*, *13*, *15*, *18*
sort.cpg *(*cpg*)*, 5
sort.cpg.perm, *9*, *13*, *15*, *18*
sort.cpg.perm *(*Class cpg.perm*)*, 3
summary.cpg *(*cpg*)*, 5
summary.cpg.perm *(*Class cpg.perm*)*, 3