Package 'wateRmelon'

December 6, 2024

Type Package

Title Illumina DNA methylation array normalization and metrics

Version 2.12.0

Description 15 flavours of betas and three performance metrics, with methods for objects produced by methylumi and minfi packages.

License GPL-3

Depends R (>= 3.5.0), Biobase, limma, methods, matrixStats, methylumi, lumi, ROC, IlluminaHumanMethylation450kanno.ilmn12.hg19, illuminaio

Imports Biobase

Enhances minfi

Suggests RPMM, IlluminaHumanMethylationEPICanno.ilm10b2.hg19, BiocStyle, knitr, rmarkdown, IlluminaHumanMethylationEPICmanifest, irlba, FlowSorted.Blood.EPIC, FlowSorted.Blood.450k, preprocessCore

LazyLoad yes

biocViews DNAMethylation, Microarray, TwoChannel, Preprocessing, QualityControl

Collate as.methylumi.R bscon.R bscon_methy.R bscon_minfi.R getAnn.R oxyscale.R adaptRefQuantiles.R beta1.R Beta2M.R betaqn.R bgeq.R bgeqot.R bgeqq2.R bgeqqn.R BMIQ_1.1.R combo.R createAnnotation.R concatenateMatrices.R coRankedMatrices.R correctI.R correctII.R dataDetectPval2NA.R db1.R detectionPval.filter.R dfs2.R dfsfit.R dmrse.R dmrse_col.R dmrse_row.R dyebuy1.R dyebuy2.R dyebuy3.R dyebuy4.R estimateCellCounts.R estimateSex.R filterXY.R findAnnotationProbes.R gcoms.R gcose.R genki.R genkme.R genkus.R genotype.R getMethylumiBeta.R getQuantiles.R getSamples.R getsnp.R horv.R loadMethylumi2.R lumiMethyR2.R M2Beta.R melon.R nbBeadsFilter.R normalize.quantiles2.R normalizeIlluminaMethylation.R ot.R outlyx.R pasteque.R peak.correction.R pfilter.R pipelineIlluminaMethylation.batch.R pwod.R readEPIC.R preprocessIlluminaMethylation.R referenceQuantiles.R adjustedDasen.R adjustedFunnorm.R robustQuantileNorm_Illumina450K.probeCategories.R robustQuantileNorm_Illumina450K.R seabird.R sextest.R summits.R swan2.R uniqueAnnotationCategory.R qual.R uSexQN.R smokp.R

manifesto.R readAny.R AllGenerics.R x_methylumi.R y_minfi.R z_bigmelon.R

RoxygenNote 7.3.1

NeedsCompilation no

VignetteBuilder knitr

Encoding UTF-8

git_url https://git.bioconductor.org/packages/wateRmelon

git_branch RELEASE_3_20

git_last_commit 3f299e1

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-12-05

Author Leo C Schalkwyk [cre, aut],
Tyler J Gorrie-Stone [aut],
Ruth Pidsley [aut],
Chloe CY Wong [aut],
Nizar Touleimat [ctb],
Matthieu Defrance [ctb],
Andrew Teschendorff [ctb],
Jovana Maksimovic [ctb],
Louis Y El Khoury [ctb],
Yucheng Wang [ctb],
Alexandria Andrayas [ctb]

Maintainer Leo C Schalkwyk <lschal@essex.ac.uk>

Contents

`wateRmelon-package` *Illumina 450K arrays: normalization and performance metrics*

Description

Functions for calculating the index of DNA methylation proportion beta in 15 different ways, and three different ways of estimating data quality or normalization performance.

Details

Package:	wateRmelon
Type:	Package
Version:	1.0
Date:	2012-10-10
License:	GPL3

Author(s)

Leonard C Schalkwyk, Ruth Pidsley and Chloe Wong Maintainer: Who to complain to <leonard.schalkwyk@kcl.ac.uk>

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

.createAnnotation *Internal function to guess correct pData and retrieve using minfi*

Description

Internal function to guess correct pData and retrieve using minfi

Usage

```
.createAnnotation(object)
```

.getManifestString *Internal functions for Illumina i450 normalization functions*

Description

got and fot find the annotation column differentiating type I and type II assays in MethylSet (got) or MethyLumiSet (fot) objects. pop extracts columns from IlluminaHumanMethylation450k.db

Usage

```
.getManifestString(annotation)
```

Arguments

annotation	A string naming the array type
x	a MethyLumiSet
obj	a MethylSet
fd	a character vector of the desired annotation columns
rn	a character vector of the desired features

Details

got returns a character vector of 'I' and 'II', fot returns the index of the relevant column. pop returns a data frame

Author(s)

lschal@essex.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

adaptRefQuantiles *Functions from 450-pipeline (Touleimat & Tost)*

Description

These functions are part of the 450K pipeline (Touleimat and Tost, Epigenomics 2012 4:325). For freestanding use of the normalization function, a wrapper is provided, see `tost`

Author(s)

Nizar Touleimat, wrapper by Leonard.Schalkwyl@kcl.ac.uk

References

Touleimat N, Tost J: Complete pipeline for Infinium R Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 2012, 4:325-341

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

adjustedDasen *adjustedDasen*

Description

adjustedDasen utilizes dasen normliasation to normalise autosomal CpGs, and infers the sex chromosome linked CpGs by linear interpolation on corrected autosomal CpGs.

Usage

```
adjustedDasen(
  mns,
  uns,
  onetwo,
  chr,
  offset_fit = TRUE,
  cores = 1,
  ret2 = FALSE,
  fudge = 100,
  ...
)
```

Arguments

`mns`	matrix of methylated signal intensities, samples in column and probes in row.
`uns`	matrix of unmethylated signal intensities, samples in column and probes in row.
`onetwo`	character vector or factor of length nrow(mns) indicating assay type 'I' or 'II'.
`chr`	character vector stores the mapped chromosomes for all probes, e.g. chr <- c('1', 'X', '21', ..., 'Y').

offset_fit	logical (default is TRUE). To use dasen, set it TRUE; to use nasen, set it FALSE.
cores	an integer(e.g. 8) defines the number of cores to parallel processing. Default value is 1, set to -1 to use all available cores.
ret2	logical (default is FALSE), if TRUE, returns a list of intensities and betas instead of a naked matrix of betas.
fudge	default 100, a value added to total intensity to prevent denominators close to zero when calculating betas, e.g. betas <- mns / (mns + uns + fudge).
...	additional argument roco for dfsfit giving Sentrix rows and columns. This allows a background gradient model to be fit. This is split from data column names by default. roco=NULL disables model fitting (and speeds up processing), otherwise roco can be supplied as a character vector of strings like 'R01C01' (only 3rd and 6th characters used).

Value

a matrix of normalised beta values.

References

A data-driven approach to preprocessing Illumina 450K methylation array data, Pidsley et al, BMC Genomics.
interpolatedXY: a two-step strategy to normalise DNA methylation microarray data avoiding sex bias, Wang et al., 2021.

Examples

```
data(melon)
normalised_betas <- adjustedDasen(mns = methylated(melon), uns = unmethylated(melon), onetwo = fData(melon)[,f
## if input is an object of methylumiset or methylset
normalised_betas <- adjustedDasen(melon)
```

`adjustedFunnorm` *adjustedFunnorm*

Description

adjustedFunnorm utilizes functional normliasation to normalise autosomal CpGs, and infers the sex chromosome linked CpGs by linear interpolation on corrected autosomal CpGs.

Usage

```
adjustedFunnorm(
  rgSet,
  nPCs = 2,
  sex = NULL,
  bgCorr = TRUE,
  dyeCorr = TRUE,
  keepCN = TRUE,
  ratioConvert = TRUE,
  verbose = TRUE
)
```

Arguments

rgSet	An object of class "RGChannelSet".
nPCs	Number of principal components from the control probes PCA.
sex	An optional numeric vector containing the sex of the samples.
bgCorr	Should the NOOB background correction be done, prior to functional normalization (see "preprocessNoob")
dyeCorr	Should dye normalization be done as part of the NOOB background correction (see "preprocessNoob")?
keepCN	Should copy number estimates be kept around? Setting to 'FALSE' will decrease the size of the output object significantly.
ratioConvert	Should we run "ratioConvert", ie. should the output be a "GenomicRatioSet" or should it be kept as a "GenomicMethylSet"; the latter is for experts.
verbose	Should the function be verbose?

Value

an object of class "GenomicRatioSet", unless "ratioConvert=FALSE" in which case an object of class "GenomicMethylSet".

References

Functional normalization of 450k methylation array data improves replication in large cancer studies, Fortin et al., 2014, Genome biology.
interpolatedXY: a two-step strategy to normalise DNA methylation microarray data avoiding sex bias, Wang et al., 2021.

Examples

```
## Not run:
GRset <- adjustedFunnorm(RGSet)

## End(Not run)
```

---

agep *Age Prediction from methylomic expression data*

Description

Predict age of samples using Horvaths Coefficients

Usage

```
agep(betas, coeff = NULL, method = c('horvath', 'hannum', 'phenoage', 'skinblood', 'lin', 'all'), n_
```

Arguments

betas	Matrix of betas or MethyLumiSet or MethylSet object.
coeff	If NULL, will default to whatever method is specified in method. If not NULL, the expected input should be a vector of coefficients and intercept
method	Currently: "horvath", "hannum", "phenoage", "skinblood", "lin" and "all", if "all" agep will seek to calculate ages using all methods else will use the method specified. Default is "horvath".
n_missing	Logical, additionally output the number of missing CpGs for each sample using the specified method or coeff list.
missing_probes	Logical, additionally output the names of missing CpGs for each sample using the specified method or coeff list.
...	To pass to arguments to downstream functions to specify adult.age

Value

Returns matrix of predicted ages per sample. With additional columns created whether n_missing or missing_probes are specified. If method is "all" then all ages will be provided in the same matrix output

Author(s)

Original Functions: Steve Horvath

wateRmelon Implementation: Tyler Gorrie-Stone, Leo Schalkwyk, Louis El Khoury

References

Horvath S: DNA methylation age of human tissues and cell types. Genome Biology 2013, 14:R115

Examples

```
data(melon)
agep(melon,coeff=NULL, method="all", n_missing=FALSE)
agep(melon,coeff=NULL, method="horvath", n_missing=TRUE)
```

---

as.methylumi-methods *Methods for Function* as.methylumi

Description

Returns a MethyLumiSet object populaed with the data provided. There are MethyLumiSet and MethylSet methods. In the default method, the data is all optional. Please note that for the results to be sane, mn, un, bn, and pv have to be in the same sample and feature order and the same size. The function does not currently do any checks!

Usage

```
# default method
as.methylumi (mn = NULL, un = NULL, bn = NULL, pv = NULL, qc = NULL, da = NULL,...)
```

Arguments

mn	matrix of methylated signal intensities, each column representing a sample (generic) or a MethyLumiSet, RGSet, or MethylSet object. Column names are used to get Sentrix row and column by default, see '...'.
un	matrix of unmethylated signal intensities, each column representing a sample (default method) or NULL when mn is an object containing methylated and unmethylated values
bn	matrix of precalculated betas, each column representing a sample
pv	matrix of detection p-values, each column representing a sample
da	annotation data frame, such as x@featureData@data #methylumi package. If NULL (the default), the `IlluminaHumanMethylation450kmanifest` package is used. See the fd argument
qc	control probe intensities: list of 2 matrices, Cy3 and Cy5, with rownames, such as produced by intensitiesByChannel(QCdata(x)) (methylumi package)
...	Other arguments such as a featureData object or optional assayData

Methods

`signature(mn = "MethylSet")` Coerces a MethylSet to a MethyLumiSet, and provides it with a set of featureData, which by default is just the chromosome and DESIGN (ie typeI or type II assay). Other data can be included using the `fd` argument, available data is listed by the function `getColumns()`

`signature(mn = "MethyLumiSet")` This is mainly useful for adding featureData as described under MethylSet above. MethyLumiSet objects produced by methylumiR have the full annotation, those from methylumIDAT do not, and functions such as `swan` require it

`signature(mn = "ANY") as.methylumi (mn = NULL, un = NULL, bn = NULL, pv = NULL, qc = NULL, da = NULL, fd = c("CHR", "DESIGN"))`

---

beadc — *Calculates the number of samples with bead count <3 for each probe in matrix of bead count values*

Description

Calculates the number of samples with bead count <3 for each probe in matrix of bead count values.

Usage

```
beadc(x)
```

Arguments

x	matrix of bead count values returned by the beadcount function

Value

Vector of number of samples with bead count <3 for each probe

Note

The beadc function is internal to the pfilter function

Author(s)

ruth.pidsley@kcl.ac.uk

References

[1] Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

---

beadcount *Creates matrix of beacounts from minfi data.*

Description

Creates matrix of beacounts from data read in using the minfi package. NAs represent probes with beadcount <3. An Extended RG Channel Set is required for this function to work.

Usage

```
beadcount(x)
```

Arguments

x	450K methylation data read in using minfi to create an Extended RG Channel Set

Value

A matrix of bead counts with bead counts <3 represented by NA for use in the pfilter function for quality control

Note

The beadcount function is internal to the pfilter function

Author(s)

Ruth.Pidsley@kcl.ac.uk

References

[1] Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

Beta2M *Internal functions for peak.correction (fuks)*

Description

Internal functions for peak.correction

Usage

```
Beta2M(B)
```

Arguments

B	a vector or matrix of beta values for conversion

Value

a vector or matrix of the same shape as the input

Author(s)

Matthieu Defrance <defrance@bigre.ulb.ac.be>

References

Dedeurwaerder S, Defrance M, Calonne E, Sotiriou C, Fuks F: Evaluation of the Infinium Methylation 450K technology . Epigenetics 2011, 3(6):771-784.

betaqn-exprmethy450-methods *Calculate normalized betas from exprmethy450 of Illumina 450K methylation arrays*

Description

Quantile normalize betas from exprmethy450 objects

Arguments

bn	An exprmethy450 object.
fudge	value added to total intensity to prevent denominators close to zero when calculating betas

Details

betaqn quantile normalizes betas

Value

exprmethy450 object of the same shape and order as bn.

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

[1] Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

[2] Dedeurwaerder S, Defrance M, Calonne E, Sotiriou C, Fuks F: Evaluation of the Infinium Methylation 450K technology . Epigenetics 2011, 3(6):771-784.

[3] Touleimat N, Tost J: Complete pipeline for Infinium R Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 2012, 4:325-341

[4] Maksimovic J, Gordon L, Oshlack A: SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome biology 2012, 13(6):R44

`BMIQ` *Beta-Mixture Quantile (BMIQ) Normalisation method for Illumina 450k arrays*

Description

BMIQ is an intra-sample normalisation procedure, correcting the bias of type-2 probe values. BMIQ uses a 3-step procedure: (i) fitting of a 3-state beta mixture model, (ii) transformation of state-membership probabilities of type2 probes into quantiles of the type1 distribution, and (iii) a conformal transformation for the hemi-methylated probes. Exact details can be found in the reference below.

Usage

```
BMIQ(beta.v, design.v, nL = 3, doH = TRUE, nfit = 50000, th1.v = c(0.2, 0.75), th2.v = NULL, niter = 5,
## S4 method for signature 'MethyLumiSet'
BMIQ(beta.v, nL=3, doH=TRUE, nfit=5000, th1.v=c(0.2,0.75), th2.v=NULL, niter=5, tol=0.001, plots=FA
CheckBMIQ(beta.v, design.v, pnbeta.v)
```

Arguments

`beta.v`	vector consisting of beta-values for a given sample, or a MethyLumiSet or MethylSet containing multiple samples. For the MethyLumiSet and MethylSet methods, this is the only required argument, and the function will be run on each sample.
`design.v`	corresponding vector specifying probe design type (1=type1,2=type2). This must be of the same length as beta.v and in the same order.
`nL`	number of states in beta mixture model. 3 by default. At present BMIQ only works for nL=3.
`doH`	perform normalisation for hemimethylated type2 probes. These are normalised using an empirical conformal transformation and also includes the left-tailed type2 methylated probes since these are not well described by a beta distribution. By default TRUE.

nfit	number of probes of a given design type to use for the fitting. Default is 50000. Smaller values (~10000) will make BMIQ run faster at the expense of a small loss in accuracy. For most applications, 5000 or 10000 is ok.
th1.v	thresholds used for the initialisation of the EM-algorithm, they should represent buest guesses for calling type1 probes hemi-methylated and methylated, and will be refined by the EM algorithm. Default values work well in most cases.
th2.v	thresholds used for the initialisation of the EM-algorithm, they should represent buest guesses for calling type2 probes hemi-methylated and methylated, and will be refined by the EM algorithm. By default this is null, and the thresholds are estimated based on th1.v and a modified PBC correction method.
niter	maximum number of EM iterations to do. This number should be large enough to yield good fits to the type1 distribution. By default 5.
tol	tolerance convergence threshold for EM algorithm. By default 0.001.
plots	logical specifying whether to plot the fits and normalised profiles out. By default TRUE.
sampleID	the ID of the sample being normalised.
pri	logical: print verbose progress information?
pnbeta.v	BMIQ normalised profile.

Details

Full details can be found in the reference below. Note: these functions require the RPMM package, not currently a dependency of the wateRmelon package.

Value

Default method: A list with following entries:

nbeta	the normalised beta-profile for the sample
class1	the assigned methylation state of type1 probes
class2	the assigned methylation state of type2 probes
av1	the mean beta-values for the nL states for type1 probes
av2	the mean beta-values for the nL states for type2 probes
hf	the estimated "Hubble" dilation factor used in the normalisation of hemi-methylated probes
th1	estimated thresholds for calling unmethylated and methylated type1 probes
th2	estimated thresholds for calling unmethylated and methylated type2 probes

MethyLumiSet method: A methyLumiSet object

Author(s)

Andrew Teschendorff, MethyLumiSet method by Leo Schalkwyk Leonard.Schalkwyk@kcl.ac.uk

References

Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A Beta-Mixture Quantile Normalisation method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. Bioinformatics. 2012 Nov 21.

Examples

```
# library(RPMM)
# data(melon)
# BMIQ(melon,nfit=100)
```

---

bscon *Calculate bisulphite conversion*

Description

Uses control data from Infinium HumanMethylation450 BeadChip to calculate bisulfite conversion for each array

Usage

```
bscon(x, ...)  # S4 methods exist for RGChannelSet and MethyLumiset objects
```

Arguments

x	IDAT or report files containing 450k data
...	current methods have no optional arguments

Details

This function uses the green and red channels reading of the type I and type II bisilfite conversion data to return the median bisulfite conversion percentage value for each array.

For the type I chemistry the beta values are calculated by dividing the first three probes of the green channel (C1, C2, C3) and the second three probes of the red channel (C4, C5, C6) by the sum of these probes and the unconverted probes of the green (U1, U2, U3) and the red (U4, U5, U6) channel.

The beta values from type II chemistry are calculated by dividing the methylated (red) channels by the sum of methylated (red) and unmethylated (green) channels.

Value

A vector of percentage values referring to the bisulfite conversion levels of each array.

Note

Updates to HumanMethylationEPIC manifest has seen the removal of control probes C6 and U6. This does not appear to grossly affect how function performs however we are considering alternative approaches to account for this.

Author(s)

Louis El Khoury (louis.el-khoury@essex.ac.uk), Eilis Hannon, Leonard Schalkwyk (lschal@essex.ac.uk)

Examples

```
library(wateRmelon)
data(melon)
bs <- bscon(melon)
bs
```

---

canno *canno - process csv manifest into annotation object for illumina methylation preprocessing*

Description

canno - process csv manifest into annotation object for illumina methylation preprocessing

Usage

```
canno(man = "EPIC-8v2-0_A1.csv", name = NULL)
```

Arguments

man	name

Details

This is based on the scripts shipped with minfi annotation packages. It is based on existing Illumina Human Methylation csv format manifests, but because reverse-engineered, may require updates to work on future products.

Value

IlluminaMethylationManifest

---

colnames-methods *Methods for Function* colnames *in Package* **wateRmelon**

Description

Methods for function colnames in package **wateRmelon**.

Methods:

signature(x = "MethyLumiSet") returns the sample names

combo *Combine MethyLumiSet objects*

Description

This is a wrapper for combining different MethyLumiSet objects.

Usage

```
combo(...)
```

Arguments

...	Eventually, any number of MethyLumiSet objects. Currently only guaranteed for 2 objects.

Details

This is a wrapper for `methylumi::combine`, which works around a name clash with a different combine function from the `gdata` package, and also a bug in `methylumi::combine`.

Value

a `MethyLumiSet`. The `assayData`, `QCdata`, `experimentData`, `protocolData` and `phenoData` are joined on `sampleName` . `featureData` and annotation are taken from the object given in the first argument

Note

the function uses `sampleNames` and gets rid of duplicates. Numeric sampleNames cause problems (and are a Bad Idea anyway). They should be turned into names with `make.names()` first.

Author(s)

Leo Schalkwyk <leonard.schalkwyk@kcl.ac.uk>

References

[1] Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

See Also

`as.methylumi`

Examples

```
library(wateRmelon)
data(melon)
## pretend we have two different data sets
melon
pelon <- melon
sampleNames(pelon) <- gsub('^6', 7, sampleNames(pelon))
combo(melon, pelon)
```

---

dasen *Calculate normalized betas from Illumina 450K methylation arrays*

Description

Multiple ways of calculating the index of methylation (beta) from methylated and unmethylated probe intensities used in Pidsley et al 2012. S4 methods exist where possible for MethyLumiSet, MethylSet, RGSet and exprmethy450 objects.

Usage

```
dasen ( mns, uns, onetwo, fudge = 100, ret2=FALSE, ... )
nasen ( mns, uns, onetwo, ret2=FALSE, fudge = 100, ... )
betaqn( bn )
naten ( mn,  un, fudge = 100, ret2=FALSE, ... )
naten ( mn,  un, fudge = 100, ret2=FALSE, ... )
nanet ( mn,  un, fudge = 100, ret2=FALSE, ... )
nanes ( mns, uns, onetwo, fudge = 100, ret2=FALSE, ... )
danes ( mn,  un,  onetwo, fudge = 100, ret2=FALSE, ... )
danet ( mn,  un,  onetwo, fudge = 100, ret2=FALSE, ... )
danen ( mns, uns, onetwo, fudge = 100, ret2=FALSE, ... )
daten1( mn,  un,  onetwo, fudge = 100, ret2=FALSE, ... )
daten2( mn,  un,  onetwo, fudge = 100, ret2=FALSE, ... )
tost  ( mn,  un,  da, pn )
fuks  ( data, anno)
swan  ( mn,  un,  qc, da=NULL, return.MethylSet=FALSE )
```

Arguments

mn, mns	matrix of methylated signal intensities, each column representing a sample (generic) or a MethyLumiSet, RGSet, or MethylSet object. Column names are used to get Sentrix row and column by default, see '...'.
un, uns	matrix of unmethylated signal intensities, each column representing a sample (default method) or NULL when mn is an object containing methylated and unmethylated values
bn, data	matrix of precalculated betas, each column representing a sample
onetwo	character vector or factor of length nrow(mn) indicating assay type 'I' or 'II'
pn	matrix of detection p-values, each column representing a sample

da, anno	annotation data frame, such as x@featureData@data #methylumi package. If NULL, the swan method requires the `IlluminaHumanMethylation450kmanifest` package.
qc	control probe intensities: list of 2 matrices, Cy3 and Cy5, with rownames, such as produced by intensitiesByChannel(QCdata(x)) #methylumi package
fudge	value added to total intensity to prevent denominators close to zero when calculating betas
return.MethylSet	if TRUE, returns a MethylSet object instead of a naked matrix of betas.
ret2	if TRUE, returns a list of intensities and betas instead of a naked matrix of betas.
...	additional argument roco for dfsfit giving Sentrix rows and columns. This allows a background gradient model to be fit. This is split from data column names by default. roco=NULL disables model fitting (and speeds up processing), otherwise roco can be supplied as a character vector of strings like 'R01C01' (only 3rd and 6th characters used).

Details

dasen same as nasen but type I and type II backgrounds are equalized first. This is our recommended method

betaqn quantile normalizes betas

naten quantile normalizes methylated and unmethylated intensities separately, then calculates betas

nanet quantile normalizes methylated and unmethylated intensities together, then calculates betas. This should equalize dye bias

nanes quantile normalizes methylated and unmethylated intensities separately, except for type II probes where methylated and unmethylated are normalized together. This should equalize dye bias without affecting type I probes which are not susceptible

danes same as nanes, except type I and type II background are equalized first

danet same as nanet, except type I and type II background are equalized first

danen background equalization only, no normalization

daten1 same as naten, except type I and type II background are equalized first (smoothed only for methylated)

daten2 same as naten, except type I and type II background are equalized first (smoothed for methylated an unmethylated)

nasen same as naten but type I and typeII intensities quantile normalized separately

tost method from Touleimat and Tost 2011

fuks method from Dedeurwaerder et al 2011. Peak correction only, no normalization

swan method from Maksimovic et al 2012

Value

a matrix (default method) or object of the same shape and order as the first argument containing betas.

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

[1] Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

[2] Dedeurwaerder S, Defrance M, Calonne E, Sotiriou C, Fuks F: Evaluation of the Infinium Methylation 450K technology . Epigenetics 2011, 3(6):771-784.

[3] Touleimat N, Tost J: Complete pipeline for Infinium R Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 2012, 4:325-341.

[4] Maksimovic J, Gordon L, Oshlack A: SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome biology 2012, 13(6):R44

See Also

pfilter, as.methylumi

Examples

```
#MethyLumiSet method
data(melon)
melon.dasen <- dasen(melon)
```

`dasen-methods` *Calculate normalized betas from MethyLumiSets of Illumina 450K methylation arrays*

Description

Multiple ways of calculating the index of methylation (beta) from methylated and unmethylated probe intensities used in Pidsley et al 2012.

Arguments

mn, mns, data, bn	A MethyLumiSet object. Sample names names are used to get Sentrix row and column by default, see '...'.
fudge	value added to total intensity to prevent denominators close to zero when calculating betas
...	additional argument roco for dfsfit giving Sentrix rows and columns. This allows a background gradient model to be fit. This is split from data column names by default. roco=NULL disables model fitting (and speeds up processing), otherwise roco can be supplied as a character vector of strings like 'R01C01' (only 3rd and 6th characters used).

Details

dasen same as nasen but type I and type II backgrounds are normalized first. This is our recommended method

betaqn quantile normalizes betas

naten quantile normalizes methylated and unmethylated intensities separately, then calculates betas

nanet quantile normalizes methylated and unmethylated intensities together, then calculates betas. This should equalize dye bias.

nanes quantile normalizes methylated and unmethylated intensities separately, except for type II probes where methylated and unmethylated are normalized together. This should equalize dye bias without affecting type I probes which are not susceptible.

danes same as nanes, except typeI and type II background are equalised first.

danet same as nanet, except typeI and type II background are equalised first.

danen background equalisation only, no normalization

daten1 same as naten, except typeI and type II background are equalised first (smoothed only for methylated)

daten2 same as naten, except typeI and type II background are equalised first (smoothed for methylated an unmethylated)

nasen same as naten but typeI and typeII intensities quantile normalized separately

tost method from Touleimat and Tost 2011

fuks method from Dedeurwaerder et al 2011. Peak correction only, no normalization

swan method from Maksimovic et al 2012

Value

a matrix (default method) or object of the same shape and order as the first argument containing betas.

methods

dasen (mns, fudge = 100, ...) nasen (mns, fudge = 100) betaqn(bn) naten (mn, fudge = 100) naten (mn, fudge = 100) nanet (mn, fudge = 100) nanes (mns,fudge = 100) danes (mn, fudge = 100, ...) danet (mn, fudge = 100, ...) danen (mns,fudge = 100, ...) daten1(mn, fudge = 100, ...) daten2(mn, fudge = 100, ...) tost (mn) fuks (data) swan (mn)

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

[1] Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

[2] Dedeurwaerder S, Defrance M, Calonne E, Sotiriou C, Fuks F: Evaluation of the Infinium Methylation 450K technology . Epigenetics 2011, 3(6):771-784.

[3] Touleimat N, Tost J: Complete pipeline for Infinium R Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 2012, 4:325-341

[4] Maksimovic J, Gordon L, Oshlack A: SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome biology 2012, 13(6):R44

dasen-minfi-methods *Calculate normalized betas from Illumina 450K methylation arrays*

Description

Multiple ways of calculating the index of methylation (beta) from methylated and unmethylated probe intensities used in Pidsley et al 2012.

Arguments

mn, mns	matrix of methylated signal intensities, each column representing a sample (default method), or an object for which a method is available. Column names are used to get Sentrix row and column by default, see '...'.
un, uns	matrix of unmethylated signal intensities, each column representing a sample (default method) or NULL when mn is an object containing methylated and unmethylated values
bn, data	matrix of precalculated betas, each column representing a sample
onetwo	character vector or factor of length nrow(mn) indicating assay type 'I' or 'II'
da, anno	annotation data frame, such as x@featureData@data #methylumi package
qc	control probe intensities: list of 2 matrices, Cy3 and Cy5, with rownames, such as produced by intensitiesByChannel(QCdata(x)) #methylumi package
fudge	value added to total intensity to prevent denominators close to zero when calculating betas
...	additional argument roco for dfsfit giving Sentrix rows and columns. This allows a background gradient model to be fit. This is split from data column names by default. roco=NULL disables model fitting (and speeds up processing), otherwise roco can be supplied as a character vector of strings like 'R01C01' (only 3rd and 6th characters used).

Details

dasen same as nasen but type I and type II backgrounds are normalized first. This is our recommended method

betaqn quantile normalizes betas

naten quantile normalizes methylated and unmethylated intensities separately, then calculates betas

nanet quantile normalizes methylated and unmethylated intensities together, then calculates betas. This should equalize dye bias.

nanes quantile normalizes methylated and unmethylated intensities separately, except for type II probes where methylated and unmethylated are normalized together. This should equalize dye bias without affecting type I probes which are not susceptible.

danes same as nanes, except typeI and type II background are equalised first.

danet same as nanet, except typeI and type II background are equalised first.

danen background equalisation only, no normalization

daten1 same as naten, except typeI and type II background are equalised first (smoothed only for methylated)

daten2 same as naten, except typeI and type II background are equalised first (smoothed for methylated an unmethylated)

nasen same as naten but typeI and typeII intensities quantile normalized separately

tost method from Touleimat and Tost 2011

fuks method from Dedeurwaerder et al 2011. Peak correction only, no normalization

swan method from Maksimovic et al 2012

Value

a matrix of betas is returned by the MethySet and RGChannelSet methods because they do not have a defined slot for betas.

methods

dasen (mns, uns, onetwo, fudge = 100, ...) nasen (mns, uns, onetwo, fudge = 100) betaqn(bn) naten (mn, un, fudge = 100) naten (mn, un, fudge = 100) nanet (mn, un, fudge = 100) nanes (mns, uns, onetwo, fudge = 100) danes (mn, un, onetwo, fudge = 100, ...) danet (mn, un, onetwo, fudge = 100, ...) danen (mns, uns, onetwo, fudge = 100, ...) daten1(mn, un, onetwo, fudge = 100, ...) daten2(mn, un, onetwo, fudge = 100, ...) tost (mn, un, da, pn) fuks (data, anno) swan (mn, un, qc)

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

[1] Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

[2] Dedeurwaerder S, Defrance M, Calonne E, Sotiriou C, Fuks F: Evaluation of the Infinium Methylation 450K technology . Epigenetics 2011, 3(6):771-784.

[3] Touleimat N, Tost J: Complete pipeline for Infinium R Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 2012, 4:325-341)

[4] Maksimovic J, Gordon L, Oshlack A: SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome biology 2012, 13(6):R44

---

db1 *Internal wateRmelon functions for calculating betas*

Description

db1 is used for quantile normalizing methylated together with unmethylated (dye bias methods nanet, nanes, danes and danet. dfs* functions are used for smoothing the background equalization in methods whose names start with d (daten etc).

Usage

```
db1(mn, un)
dfsfit(mn, onetwo, roco=substring(colnames(mn), regexpr("R0[1-9]C0[1-9]", colnames(mn))), ... )

dfs2(x, onetwo)
```

Arguments

mn, x	matrix of methylated signal intensities, each column representing a sample (default method), or an object for which a method is available. For dfsfit and dfs2 this can also be a matrix of unmethylated intensities.
un	matrix of unmethylated signal intensities, each column representing a sample (default method) or NULL when mn is an object containing methylated and unmethylated values.
onetwo	character vector or factor of length nrow(mn) indicating assay type 'I' or 'II'
roco	roco for dfsfit giving Sentrix rows and columns. This allows a background gradient model to be fit. This is split from data column names by default. roco=NULL disables model fitting (and speeds up processing), otherwise roco can be supplied as a character vector of strings like 'R01C01' (3rd and 6th characters used).
...	no additional arguments currently used

Details

db1 - quantile normalizes methylated against unmethylated (basic function for dyebuy* dye bias methods). dfsfit - corrects the difference in backgrounds between type I and type II assays and fits a linear model to Sentrix rows and columns if these are available to improve precision where there is a background gradient. dfs2 - finds the difference between type I and type II assay backgrounds for one or more samples.

Value

db1 - a list of 2 matrices of intensities, methylated and unmethylated dfsfit - a matrix of adjusted intensities dfs2 - a background offset value

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

dmrse *Standard error of iDMR 450k array DNA methylation features*

Description

Imprinting differentially methylated regions (iDMRs) are expected to be approximately half methylated, as is observed at the 227 probes in known iDMRs. These functions calculate measures of dispersion for the beta values at these CpG sites, of which the most useful is dmrse_row, which is the between-sample standard error.

Usage

```
dmrse(betas, idmr = iDMR())
dmrse_col(betas, idmr = iDMR())
dmrse_row(betas, idmr = iDMR())
```

Arguments

betas	a matrix of betas (default method), a MethyLumiSet object (methylumi package), a MethylSet or RGChannelSet object (minfi package) or a exprmethy450 object (IMA package).
idmr	a character vector of iDMR probe names such as returned by iDMR()

Value

return a standard error of the mean of betas for all samples and iDMR probes (dmrse) or the standard error of the mean for just the between sample component(dmrse_row) or between probe(dmrse_col) component.

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

See Also

seabi, a sex-difference metric, and genki, based on SNPs.

Examples

```
#MethyLumiSet method
   data(melon)
   dmrse(melon)

#MethyLumiSet method after normalization
   melon.dasen <- dasen(melon)
   dmrse(melon.dasen)
```

`dmrse-methods` *Methods for Function* `dmrse` *in Package* **wateRmelon**

Description

Methods for function `dmrse`, `dmrse_row` and `dmrse_col` in package **wateRmelon**. Please see `dmrse` for details of the calculation of this standard-error performance metric.

Methods:

`signature(betas = "exprmethy450")` all of the methods simply extract betas from the data object (which can be a `exprmethy450`, `MethylSet`, `MethyLumiSe`, or `RGChannelSet`) and calculate the metric.

`epicv2clean.default` *Strip and subset EPICv2 data to work with legacy data and methods*

Description

Returns an object with rownames stripped of the EPICv2 suffixes, duplicate probes are omitted.

Usage

```
## Default S3 method:
epicv2clean(x)
```

Details

EPICv2 manifests contain a few thousand probes with up to 10 replicate syntheses. To accomodate this a modified naming scheme is used, so none of the probe names match those on the EPIC and previous arrays (even though most of the probes are the same sequence and presumably simiar performance).

This simple function relies on the rowname and subsetting methods and will work for matrix, dataframe, MethyLumiSet, or MethylSet objects, and there is a method for gds (bigmelon) objects.

`estimateCellCounts` *Cell Proportion Estimation using wateRmelon*

Description

Estimates relative proportion of pure cell types within a sample, mostly identical to `estimateCellCounts`. References for both 450k and EPIC array are available. However 450k reference can be used on EPIC data by specifying the reference platform. Additionally a measure of error is calculated as a means of quality control.

Usage

```
estimateCellCounts.wmln(
    object,
    referencePlatform = c("IlluminaHumanMethylation450k",
        "IlluminaHumanMethylationEPIC",
        "IlluminaHumanMethylation27k"),
    mn = NULL,
    un = NULL,
    bn = NULL,
    perc = 1,
    compositeCellType = "Blood",
    probeSelect = "auto",
    cellTypes = c("CD8T","CD4T","NK","Bcell","Mono","Gran"),
    returnAll = FALSE,
    meanPlot = FALSE,
    verbose=TRUE,
    ...)
```

Arguments

object	An object of class methylumiset, which contains (un)normalised methylated and unmethylated intensities
mn	if NULL will call methylated(object), otherwise can be given matrix of identical dimensions to object.
un	if NULL will call unmethylated(object), otherwise can be given matrix of identical dimensions to object.
bn	if NULL will call betas(object), otherwise can be given matrix of identical dimensions to object.
perc	Percentage of query-samples to use to normalise reference dataset. This should be 1 unless using a very large data-set then lowering this will allow for an increase in performance
compositeCellType	Which composite cell type is being deconvoluted. Should be either "Blood", "CordBlood", or "DLPFC"
probeSelect	How should probes be selected to distinguish cell types? Options include "both", which selects an equal number (50) of probes (with F-stat p-value < 1E-8) with the greatest magnitude of effect from the hyper- and hypo-methylated sides, and "any", which selects the 100 probes (with F-stat p-value < 1E-8) with the greatest magnitude of difference regardless of direction of effect. Default input "auto" will use "any" for cord blood and "both" otherwise, in line with previous versions of this function and/or our recommendations. Please see the references for more details.
cellTypes	Which cell types, from the reference object, should be we use for the deconvolution? See details.
referencePlatform	The platform for the reference dataset; if the input rgSet belongs to another platform, it will be converted using convertArray.
returnAll	Should the composition table and the normalized user supplied data be return?
verbose	Should the function be verbose?

meanPlot	Whether to plots the average DNA methylation across the cell-type discrimating probes within the mixed and sorted samples.
...	Other arguments, i.e arguments passed to plots

Details

See estimateCellCounts for more information regarding the exact details. estimateCellCounts.wmln differs slightly, as it will impose the quantiles of type I and II probes onto the reference Dataset rather than normalising the two together. This is 1) More memory efficient and 2) Faster - due to not having to normalise out a very small effect the other 60 samples from the reference set will have on the remaining quantiles.

Optionally, a proportion of samples can be used to derive quantiles when there are more than 1000 samples in a dataset, this will further increase performance of the code at a cost of precision. If data is pre-normalised a minimum of two samples are required.

---

estimateSex *Predict sex by using robust sex-related CpG sites on ChrX and ChrY*

Description

Predict sex by using robust sex-related CpG sites on ChrX and ChrY

Usage

```
estimateSex(betas, do_plot = FALSE)
```

Arguments

betas	A matrix with sample IDs as column names, and probe names as row names, ideally: beta = M / (M + U + 100). Take a look at an example betas with: "data(melon); print(betas(melon)[1:10, 1:3])".
do_plot	logical. Should plot the predicted results? Default: FALSE

Value

dataframe contains predicted sex information.

Author(s)

Wang, Yucheng, et al. "DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy." BMC genomics 22.1 (2021): 1-11.

Examples

```
data(melon)
pred_XY <- estimateSex(betas(melon), do_plot=TRUE)
```

genki *SNP derived performance metrics for Illumina 450K DNA methylation arrays.*

Description

A very simple genotype calling by one-dimensional K-means clustering is performed on each SNP, and for those SNPs where there are three genotypes, the squared deviations are summed for each genotype (similar to a standard deviation for each of allele A homozygote, heterozygote and allele B homozygote). By default these are further divided by the square root of the number of samples to get a standard error-like statistic.

Usage

```
genki(bn, g = getsnp(rownames(bn)), se = TRUE)
```

Arguments

bn	a matrix of beta values(default method), a `MethyLumiSet` object (`methylumi` package), a `MethylSet` or `RGChannelSet` object (`minfi` package) or a `exprmethy450` object (`IMA` package).
g	vector of SNP names
se	TRUE or FALSE specifies whether to calculate the standard error-like statistic

Details

There are 65 well-behaved SNP genotyping probes included on the array. These each produce a distribution of betas with tight peaks for the three possible genotypes, which will be broadened by technical variation between samples. The spread of the peaks is thus usable as a performance metric.

Value

a vector of 3 values for the dispersion of the three genotype peaks (AA, AB, BB : low, medium and high beta values)

Note

Corrected `RGChannelSet` methods - 12/10/2015

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

Examples

```
#MethyLumiSet method
   data(melon)
   genki(melon)

#MethyLumiSet method after normalization
   melon.dasen <- dasen(melon)
   genki(melon.dasen)
```

---

genki-methods *Methods for Function* genki *in Package* **wateRmelon**

Description

Methods for function genki in package **wateRmelon**. Please see genki for details of the calculation of this standard-error performance metric.

Methods:

signature(betas = "exprmethy450") all of the methods simply extract betas from the data object (which can be a exprmethy450, MethylSet, MethyLumiSe, or RGChannelSet) and calculate the metric.

---

genkme *Internal functions for genotype-based normalization metrics*

Description

genkme - genotype calling with 1d k-means

genkus - apply genkme to available SNPs

getsnp - grep the rs-numbered probes

gcose - calculate between-sample SNP standard error

gcoms - calculate between-sample SNP mean-squared deviation

Usage

```
genkme(y, peaks = c(0.2, 0.5, 0.8))
```

Arguments

y	a vector or matrix of numeric values (betas, between 0 and 1)
peaks	initial values for cluster positions

Details

see genki

Value

see genki

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

---

got *Internal functions for Illumina i450 normalization functions*

Description

got and fot find the annotation column differentiating type I and type II assays in MethylSet (got) or MethyLumiSet (fot) objects. pop extracts columns from IlluminaHumanMethylation450k.db

Usage

```
got(obj)
fot(x)
```

Arguments

x	a MethyLumiSet
obj	a MethylSet

Details

got returns a character vector of 'I' and 'II', fot returns the index of the relevant column. pop returns a data frame

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

idet — *idet - identify idats by a hash of the addresses*

Description

idet - identify idats by a hash of the addresses

Usage

```
idet(idat)
```

Arguments

idat	can be an idat file or the list produced by reading one with readIDAT()

Details

this function is a response to the fact that IlluminaHumanMethylationEPIC and EPICv2 idats both have the ChipType "BeadChip 8x5" but different manifests. They did have different numbers of addresses. Subsequently we have had some confusion.... This hash (certainly taken together with the ChipType) should be bomb proof as an identifier. Warning: this is slow.

Value

three strings: ChipType, an md5 hash of the MidBlock (address vector) and if known, the annotation name

iDMR — *Imprinting differentially methylated region probes of Illumina 450 arrays*

Description

A character vector of 227 propes on the Illumina 450k methylation array

Usage

```
data(iDMR)
```

Format

The format is: chr [1:227] "cg00000029" "cg00155882" "cg00576435" "cg00702231" "cg00765653" "cg00766368" ...

Source

DMR coordinates from https://atlas.genetics.kcl.ac.uk/

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

Examples

```
data(iDMR)
## maybe str(iDMR) ; plot(iDMR) ...
```

`melon` *Small MethyLumi data set for examples and testing*

Description

This object was derived using `methylumiR` on an edited GenomeStudio file containing a small subset of features. It works with all of the wateRmelon package beta functions (see `dasen` and metrics (see `genki`, `seabi`, and `dmrse_col`) except for `swan`.

Usage

```
data(melon)
```

Format

MethyLumiSet with assayData containing 3363 features, 12 samples

Source

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

Examples

```
library(methylumi)
data(melon)
boxplot(log(methylated(melon)), las=2)
## maybe str(melon) ; plot(melon) ...
```

`metrics` *Calculate a full set of 450K normalization/performance metrics*

Description

Calculate X-chromosome, SNP and imprinting DMR metrics for a matrix of betas from an Illumina 450K Human DNA methylation array. Requires precalculated t-test p-values for sex differences, a list of X-chromosome features and of imprinting DMR features.

Usage

```
metrics(betas, pv, X, idmr = iDMR, subset = NULL)
```

Arguments

betas	a matrix of betas, each row representing a probe, each column a sample
pv	a vector of p-values such as produced by sextest, one per row of betas
X	a logical vector of the same length as pv, indicating whether each probe is mapped to the X-chromosome
idmr	a character vector of probe names known to be in imprinting DMRs. Can be obtained with iDMR() or data(iDMR)
subset	index or character vector giving a subset of betas to be tested

Value

dmrse_row	see dmrse_row
dmrse_col	see dmrse_col
dmrse	see dmrse
gcoms_a	see genki
gcose_a	see genki
gcoms_b	see genki
gcose_b	see genki
gcoms_c	see genki
gcose_c	see genki
seabird	see seabi

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

Examples

```
data(melon)
melon.dasen <- dasen(melon)
bn <-betas(melon.dasen)
X <- melon.dasen@featureData@data$CHR=='X'
data(iDMR)
sex <- pData(melon.dasen)$sex
pv <- sextest(bn,sex)
melon.metrics <- metrics(bn, pv, X, idmr = iDMR, subset = NULL)
```

NChannelSetToMethyLumiSet2 *For internal use, is read using minfi-like machinery and then preprocessed into the more flexible and convenient methylumi object used by wateRmelon/bigmelon*

Description

For internal use, is read using minfi-like machinery and then preprocessed into the more flexible and convenient methylumi object used by wateRmelon/bigmelon

Usage

```
NChannelSetToMethyLumiSet2(
  NChannelSet,
  parallel = F,
  pval = 0.05,
  n = F,
  n.sd = F,
  oob = T,
  to = TRUE
)
```

Arguments

NChannelSet	an NChannelSet (raw red and green values not yet mapped to Illumina IDs/CpG names
parallel	no effect, included for future parallelisation
pval	detection pval threshold for filtering. Inactivated.
n	keep nbeads data (min of m & u)
n.sd	process SD of U and M (not currently implemented)
oob	keep out-of-band signals
to	does chip have type I and II probes?

Value

A methylumi object with betas, U and M, optionally additional data

outlyx *Identify Outliers within Methylumi and Minfi packaged objects*

Description

Seeks to identify outliers based on multiple (currently 2) outlier detection methods for `methylumi` and `minfi` packaged objects.

Usage

```
outlyx(x, iqr=TRUE, iqrP=2, pc=1,
       mv=TRUE, mvP=0.15, plot=TRUE, ...)
```

Arguments

x	A MethyLumiSet, MethylSet, RGChannelSet object or matrix containing raw betas.
iqr	If TRUE, the outliers based on interquartile ranges will be determined
iqrP	The number of interquartile ranges outliers are to be identified from designated principle component.
pc	Desired principal component for outlier identification - only used if other principal components want to be discriminated, only used for IQR outlier detection.
mv	If TRUE, the outliers will detected using pcout
mvP	Arbitrary cut-off point for identifying outliers via pcout
plot	If TRUE, alongside regular output, a plot will be constructed displaying relative 'location' of each sample. Outliers are those that fall within the highlighted regions.
...	Additional arguments passed to pcout

Value

Returns a dataframe of TRUE/FALSE per sample where TRUE is outlying. Dataframe contains 3 columns, the first column (iqr) denotes samples which are outlying according to IQR on Principal component 1, the second column (mv) denotes outliers according to mahalanobis distances. And the third column (outliers) denotes samples that are TRUE in the first two columns.

Note

May perform poorly on normalized data

Author(s)

Tyler Gorrie-Stone - tgorri@essex.ac.uk

Examples

```
library(wateRmelon)
data(melon)
outliers <- outlyx(melon,iqr=TRUE, iqrP=2, pc=1,
                   mv=TRUE, mvP=0.15, plot=TRUE)
```

outlyx-methods *Methods for Function* outlyx *in Package* **wateRmelon**

Description

Methods for function outlyx, please see outlyx for details of how function performs.

Methods

signature(x = "MethyLumiSet") all of the methods simply extract betas from the data object (which can be a MethylSet, MethyLumiSet, or RGChannelSet) and calculates the outliers.

pfilter *Basic data filtering for Illumina 450 methylation data*

Description

The pfilter function filters data sets based on bead count and detection p-values. The user can set their own thresholds or use the default pfilter settings. pfilter will take data matrices of beta values, signal intensities and annotation data, but will also take methylumi (MethyLumiSet) or minfi (RGChannelSetExtended) objects. However it has come to our attention that data read in using the various packages and input methods will give subtly variable data output as they calculate detection p-value and beta values differently, and do/don?t give information about beadcount. The pfilter function does not correct for this, but simply uses the detection p-value and bead count provided by each package.

Usage

```
pfilter(mn, un, bn, da, pn, bc, perCount=NULL, pnthresh = NULL, perc = NULL,
pthresh = NULL,logical.return=FALSE)
```

Arguments

mn	matrix of methylated signal intensities, each column representing a sample (default method), or an object for which a method is available e.g MethyLumiSet or RGChannelSetExtended. N.B. Bead count filtering will not work unless data read in as an RGGhannelSetExtended rather than an RGChannelSet.
un	matrix of unmethylated signal intensities, each column representing a sample (default method) or NULL when mn is a MethyLumiSet or RGChannelSetExtended object
bn	matrix of precalculated betas, each column representing a sample, or NULL when mn is a MethyLumiSet or RGChannelSetExtended object
da	annotation data frame, such as x@featureData@data #methylumi package, or NULL when mn is a MethyLumiSet or RGChannelSetExtended object
pn	matrix of detection p-values, each column representing a sample, a MethyLumiSet or RGChannelSetExtended object

bc	matrix of arbitrary values, each column representing a sample and eeach row representing a probe, in which "NA" represents beadcount <3, or NULL when mn is a MethyLumiSet or RGChannelSetExtended object
perCount	remove sites having this percentage of samples with a beadcount <3, default set to 5
pnthresh	cutoff for detection p-value, default set to 0.05
perc	remove samples having this percentage of sites with a detection p-value greater than pnthresh, default set to 1
pthresh	remove sites having this percentage of samples with a detection p-value greater than pnthresh, default set to 1
logical.return	If it is TRUE, FALSE or TRUE is returned to indicate success

Value

a filtered MethyLumiSet or

a list of the filtered matrices:

mn : methylated intensities

un : unmethylated intensities

bn : betas

da : feature data

or

a filtered MethylSet object.

Methods

`signature(mn = "MethyLumiSet")` This is used for performing the pfilter method on MethyLumiSet objects produced by methylumiR.

`signature(mn = "RGChannelSetExtended")` This is used for performing the pfilter method on RGChannelSetExtended objects produced by minfi.

Note

Adjusted `RGChannelSetExtended` methods - 12/10/2015 Now outputs a MethylSet object using preprocessRaw from minfi.

Author(s)

Ruth.Pidsley@kcl.ac.uk

References

[1] Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

Examples

```
# MethyLumiSet method
data(melon)
melon.pf <- pfilter(melon)
```

pwod *Probe-Wise Outlier Detection*

Description

'P'robe-'W'ise 'O'utlier 'D'etection via interquartile ranges.

Usage

```
pwod(object, mul=4)
```

Arguments

object	MethyLumiSet, RGChannelSet, MethylSet object or matrix containing betas.
mul	Number of interquartile ranges used to determine outlying probes. Default is 4 to ensure only very obvious outliers are removed.

Details

Detects outlying probes across arrays in methylumi and minfi objects. Outliers are probable low MAF/SNP heterozygotes.

Value

Returns supplied beta matrix with outlying probes coerced to NA

Author(s)

Tyler Gorrie-Stone - tgorri@essex.ac.uk

Examples

```
library(wateRmelon)
data(melon)
cattle <- betas(melon)
new.betas <- pwod(cattle, mul=4)
```

pwod-methods *Methods for Function* pwod *in Package* **wateRmelon**

Description

Methods for function pwod, please see pwod for details of how function performs.

Methods

signature(object = "MethyLumiSet") all of the methods simply extract betas from the data object (which can be a MethylSet, MethyLumiSet, or RGChannelSet) and calculates the outliers.

qual *A measure of Normalization Violence*

Description

Calculates 4 metrics to assess the degree of difference between normalized and raw betas.

Usage

```
qual(norm,raw)
```

Arguments

norm	Matrix of normalized betas
raw	Matrix of raw betas

Value

Returns data.frame containing rmsd, sdd, sadd and srms for each sample (columns) in supplied matrices.

Author(s)

Leo Schalkwyk

Examples

```
library(wateRmelon)
data(melon)
d.melon <- dasen(melon)
raw.bet <- betas(melon)
norm.bet <- betas(d.melon)
qual(norm=norm.bet, raw=raw.bet)
```

read.manifest *read.manifest - read in csv format Illumina chip manifest files*

Description

read.manifest - read in csv format Illumina chip manifest files

Usage

```
read.manifest(file)
```

Arguments

file

Details

This function is probably not much use for calling directly. It mostly exists to be called by canno.

Value

a list of of dataframes of data prepared for making IlluminaMethylationManifest

---

`readEPIC` *readEPIC*

Description

Reads Epic arrays from raw idats into `MethyLumiSet` objects from directory.

Usage

```
readEPIC(idatPath, barcodes=NULL, pdat=NULL,parallel=F,n=T,oob=F,force=F, ...)
```

Arguments

`idatPath`	Path directory where .idat files are located. `readEPIC` looks in the specified path and converts all .idats within path to relevant barcodes, which is then passed to a modified version of `methylumIDAT` to parse all idats present in the specified directory.
`barcodes`	If NULL, function will search supplied argument in "idatPath" for all idats within directory. If given a vector of barcodes, readEPIC will search for those specific barcodes within the idatPath supplied.
`parallel`	If TRUE, an attempt will be made to process using multiple cores on a multicore machine.
`pdat`	A data.frame describing the samples. A special column named "barcodes" can be used to specify the barcodes to be read when using `methylumIDATepic`. See `methylumIDAT for usage`
`n`	If TRUE, beadcounts from .idats will be included in final object
`oob`	If TRUE, out-of-band (OOB) or opposite-channel signals will be kept
`force`	If TRUE, will combine EPIC IDATs read with differing dmaps
`...`	Additional arguments passed to `methylumIDAT`

Details

Read a set of .idat files within a file directory and return a `MethylumiSet` object.

Value

A `MethyLumiSet` object.

Note

Contains heavily modified version of `methylumIDAT` and other accessory functions used to construct a `MethylumiSet` object, specifically tailored for EPIC arrays. `readEPIC` can also handle 450k and 27k arrays as `methylumIDAT` functionality for these platforms remains unchanged.

Alternatively it is possible to invoke `methylumIDATepic` to use the modified version `methylumIDAT`, which has similar usage.

EPIC manifest has since been updated to B4 version, which has notably fewer probes than previous manifests. It is entirely possible that we will migrate to the manifest packages available on BioConductor and allow for versioning control.

Author(s)

Tyler Gorrie-Stone - tgorri@essex.ac.uk

References

methylumi

Examples

```
#Ficticious file pathway
# path <- "Data/Experiment/Idatlocation"
# data <- readEPIC(path, barcodes = NULL oob=F, n=T)
```

---

readPepo *readPepo - read (any kind of) Illumina DNA methylation array idat files into a methylumi object*

Description

Pepo is a botanical term for any melon-like fruit. Given the appropriate manifest file, this function should be able to read any of the Illumina Infinium DNAm arrays including 450k, EPIC, EPIC2, MSA and Mouse.

Usage

```
readPepo(
  idatdir = ".",
  filelist = NULL,
  barcodelist = NULL,
  manifest = NULL,
  parallel = F,
  n = F,
  pdat = NULL,
  oob = F,
  two = TRUE
)
```

Arguments

idatdir	the directory with the idatfiles. Currently only handle one directory.
filelist	optional list of idat files to process.
barcodelist	optional list of barcodes to process.
manifest	name of a IlluminaMethylationManifest object or a csv format manifest. If missing, will run idet() on one of the idat files.
parallel	try to use multiple cores.
n	keep beadcounts.
pdat	optional data.frame describing the samples.
two	are there two different assay types (true of human methylation arrays except 27k)
keep	out-of-band (OOB) or opposite-channel signals

Value

A 'MethyLumiSet' object.

---

seabi *Calculate a performance metric based on male-female differences for Illumina methylation 450K arrays*

Description

Calculates an area under ROC curve - based metric for Illumina 450K data using a t-test for male-female difference as the predictor for X-chromosome location of probes. The metric is 1-area so that small values indicate good performance, to match our other, standard error based metrics gcose and dmrse. Note that this requires both male and female samples of known sex and can be slow to compute due to running a t-test on every probe.

Usage

```
seabi(bn, stop = 1, sex, X)
```

Arguments

bn	a matrix of betas (default method) or an object containing betas i.e. a MethyLumiSet object (methylumi package), a MethylSet or RGChannelSet object (minfi package) or a exprmethy450 object (IMA package).
stop	partial area under curve is calculated if stop value <1 is provided
sex	a factor giving the sex of each sample (column)
X	a logical vector of length equal to the number of probes, true for features mapped to X-chromosome

Value

a value between 0 and 1. values close to zero indicate high data quality as judged by the ability to discriminate male from female X-chromosome DNA methylation.

Author(s)

leonard.schalkwyk@kcl.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

Examples

```
   library(methylumi)
   data(melon)
   sex  <- pData(melon)$sex
   X    <- melon@featureData@data$CHR=='X'
   seabi(betas(melon), sex=sex, X=X)

# methylumi method
   seabi(melon, sex=sex, X=X)
```

---

seabi-methods *Methods for Function* seabi *in Package* **wateRmelon**

Description

Methods for function seabi in package **wateRmelon**. Please see seabi for details of the calculation of this ROC AUC performance metric.

Methods

signature(betas = "exprmethy450") all of the methods simply extract betas from the data object (which can be a exprmethy450, MethylSet, MethyLumiSe, or RGChannelSet) and calculate the metric. All the methods also require a factor differentiating male from female samples.

---

seabird *Calculate ROC area-under-curve for X-chromosome sex differences (internal function for calculating the seabi metric)*

Description

This is a wrapper for the prediction and performance functions from the ROCR package that takes a vector of p-values and a vector of true or false for being on the X. See seabi function which does everything.

Usage

```
seabird(pr, stop = 1, X)
```

Arguments

pr	a vector of p-values, such as calculated by seabird
stop	fraction for partial area under curve. For example 0.1 gives you the area for the lowest 10% of p-values.
X	logical vector the same length as pv, true for features mapped to X-chromosome

Value

Returns an area value between 0 and 1, where 1 is the best possible performance.

Author(s)

Leonard C Schalkwyk 2012 Leonard.Schalkwyk@kcl.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

---

sextest *Test Illumina methylation 450K array probes for sex difference (internal function for calculating* seabi *performance metric)*

Description

This is a wrapper for lm which does the equivalent of a Student t-test for difference in betas between males and females for each row of a matrix of betas.

Usage

```
sextest(betas, sex, ...)
```

Arguments

betas	a matrix of betas, each row is a probe, each column a sample
sex	a factor with 2 levels for male and female
...	additional arguments to be passed to lm

Value

Returns a vector of p-values of length equal to the number of rows of betas

Author(s)

Leonard.Schalkwyk@kcl.ac.uk

References

Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: A data-driven approach to preprocessing Illumina 450K methylation array data (submitted)

See Also

seabi seabird

Examples

```
#MethyLumiSet method
  data(melon)
  sex  <- pData(melon)$sex
  melon.sextest<-sextest(betas(melon),sex)


#MethyLumiSet method with quality control step
  data(melon)
  melon.dasen <- dasen(melon)
  sex  <- pData(melon.dasen)$sex
  melon.sextest<-sextest(betas(melon.dasen),sex)
```

---

smokp *Smoking Prediction from methylomic expression data*

Description

Predict smoking from samples using various methods

Usage

```
smokp(betas, method, sst)
```

Arguments

betas	Matrix of betas or MethyLumiSet or MethylSet object. Rows are Illumina IDs referring to CpG sites and Columns refer to samples or participants.
method	Currently: 'AHRR', 'McCartney', 'Maas', 'Sugden', 'Teschendorff', 'Yu', 'Gao', 'Yang', 'Zhang', 'Wen', 'Langdon', 'SSt', 'Packyears', 'Cessation', and 'All'. If "All" smokp will seek to predict smoking using all methods else will use the method specified. Default is "SSt". If 'Teschendorff', 'Yu', 'Gao', 'Yang' or 'Langdon' specified then smoking status is required.
sst	Named vector describing smoking status, coded as 'Current', 'Former', or 'Never', of participant for each sample (where names match rownames of betas).

Value

Returns data frame of predicted smoking per sample.

Author(s)

Original Functions: See References.

wateRmelon: Tyler Gorrie-Stone, Leo Schalkwyk, Louis El Khoury

smokp: Alexandria Andrayas

References

Philibert, R.A., Beach, S.R. and Brody, G.H. Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. Epigenetics, 7:11, 2012, 1331-1338.

Philibert, R., Hollenbeck, N., Andersen, E., McElroy, S., Wilson, S., Vercande, K., Beach, S.R., Osborn, T., Gerrard, M., Gibbons, F.X. and Wang, K. Reversion of AHRR demethylation is a quantitative biomarker of smoking cessation. Frontiers in psychiatry, 2016, 7, 55.

Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J., Peters, A. and Strauch, K., Tobacco smoking leads to extensive genome-wide changes in DNA methylation. 2013, PloS one, 2013, 8:5, e63812.

Elliott, H.R., Tillin, T., McArdle, W.L., Ho, K., Duggirala, A., Frayling, T.M., Smith, G.D., Hughes, A.D., Chaturvedi, N. and Relton, C.L. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clinical epigenetics, 2014, 6:1, 1-10.

Teschendorff, A.E., Yang, Z., Wong, A., Pipinikas, C.P., Jiao, Y., Jones, A., Anjum, S., Hardy, R., Salvesen, H.B., Thirlwell, C. and Janes, S.M. Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. JAMA oncology, 2015, 1:4, 476-485.

Zhang, Y., Florath, I., Saum, K. U., & Brenner, H. Self-reported smoking, serum cotinine, and blood DNA methylation. Environmental research, 2016, 146, 395-403.

Gao, X., Zhang, Y., Breitling, L.P. and Brenner, H. Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration. Oncotarget, 2016, 7:30, 46878.

Zhang, Y., Schöttker, B., Florath, I., Stock, C., Butterbach, K., Holleczek, B. & Brenner, H. Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality. Environmental health perspectives, 2016, 124:1, 67-74.

McCartney, D.L., Hillary, R.F., Stevenson, A.J., Ritchie, S.J., Walker, R.M., Zhang, Q., Morris, S.W., Bermingham, M.L., Campbell, A., Murray, A.D. and Whalley, H.C. Epigenetic prediction of complex traits and death. Genome biology, 2018, 19:1, 1-11.

Joehanes, R., Just, A.C., Marioni, R.E., Pilling, L.C., Reynolds, L.M., Mandaviya, P.R., Guan, W., Xu, T., Elks, C.E., Aslibekyan, S. and Moreno-Macias, H. Epigenetic signatures of cigarette smoking. Circulation: cardiovascular genetics, 2016, 9:5, 436-447.

Sugden, K., Hannon, E.J., Arseneault, L., Belsky, D.W., Broadbent, J.M., Corcoran, D.L., Hancox, R.J., Houts, R.M., Moffitt, T.E., Poulton, R. and Prinz, J.A. Establishing a generalized polyepigenetic biomarker for tobacco smoking. Translational psychiatry, 2019, 9:1, 1-12.

Gao, X., Jia, M., Zhang, Y., Breitling, L.P. and Brenner, H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clinical epigenetics, 2015, 7, 1-10.

Yu, H., Raut, J.R., Schöttker, B., Holleczek, B., Zhang, Y. and Brenner, H. Individual and joint contributions of genetic and methylation risk scores for enhancing lung cancer risk stratification: data from a population-based cohort in Germany. Clinical epigenetics, 2020, 12:1, 1-11.

Yang, Y., Gao, X., Just, A.C., Colicino, E., Wang, C., Coull, B.A., Hou, L., Zheng, Y., Vokonas, P., Schwartz, J. and Baccarelli, A.A. Smoking-related DNA methylation is associated with DNA methylation phenotypic age acceleration: The veterans affairs normative aging study. International journal of environmental research and public health, 2019, 16:13, 2356.

Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S., & Ollikainen, M. EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data. Epigenomics, 2019, 11:13, 1469-1486.

Maas, S.C., Vidaki, A., Wilson, R., Teumer, A., Liu, F., van Meurs, J.B., Uitterlinden, A.G., Boomsma, D.I., de Geus, E.J., Willemsen, G. and van Dongen, J. Validated inference of smoking

habits from blood with a finite DNA methylation marker set. European journal of epidemiology, 2019, 34, 1055-1074.

Langdon, R.J., Yousefi, P., Relton, C.L. and Suderman, M.J. Epigenetic modelling of former, current and never smokers. Clinical Epigenetics, 2021, 13, 1-13.

Wen, D., Shi, J., Liu, Y., He, W., Qu, W., Wang, C., Xing, H., Cao, Y., Li, J. and Zha, L. DNA methylation analysis for smoking status prediction in the Chinese population based on the methylation-sensitive single-nucleotide primer extension method. Forensic Science International, 2022, 339, 111412.

Examples

```
data(melon)
# note, melon is not a complete dataset, does not work with all methods
smokp(melon, method="McCartney", sst=NULL)
```

`wm_internal` *Internal functions for readEPIC and other wateRmelon functions introduced in v 1.13.1*

Description

few if any functions of interest to users

Usage

```
DataToNChannelSet2(mats, chans = c(Cy3 = "GRN", Cy5 = "RED"), parallel = F, protocol.data = F, IDAT =
```

Arguments

`mats`

`chans`

`parallel`

`protocol.data`

`IDAT`

`force`

Index