Package 'maskBAD'

January 19, 2025

Version 1.50.0

Title Masking probes with binding affinity differences

Author Michael Dannemann <michael_dannemann@eva.mpg.de>

Maintainer Michael Dannemann <michael_dannemann@eva.mpg.de>

Depends R (>= 2.10), gcrma (>= 2.27.1), affy

Suggests hgu95av2probe, hgu95av2cdf

Description Package includes functions to analyze and mask microarray expression data.

License GPL (>= 2)

biocViews Microarray

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

git_branch RELEASE_3_20

git_last_commit f9c4feb

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2025-01-19

Contents

exmask *Output object of the function mask*

Description

This data is the output object of the function `mask` for the AffyBatch object `newAffyBatch`.

Usage

```
exmask
```

Format

List of 1 or 2 objects.

Source

??

References

??

mask *Filtering/Masking expression data*

Description

Identyfying probes with binding affinity difference (BAD probes) between two groups of samples on the basis of expression data.

Usage

```
mask(affy,exprlist=NULL,useExpr=TRUE,ind,PM=FALSE,verbose=TRUE)
```

Arguments

`affy`	An object of class `AffyBatch`.
`exprlist`	A vector with probesetnames to be used. If NULL, all probesets are analyzed.
`useExpr`	Logical. If 'TRUE', only expressed genes (see Details) are used. If 'FALSE', all probes are analyzed.
`ind`	Numeric vector, with values 1 and 2, definining group assignement for samples in `affy`.
`PM`	Logical. If 'TRUE', only probes with a mean pm value greater than the mean mm value are used.
`verbose`	Logical. If 'TRUE', it writes out some messages indicating progress. If 'FALSE' nothing should be printed.

Details

The function mask identifies in expression data probes which binding affinity (BAD probes) differs between two groups of samples, e.g two species. The basic input data is AffyBatch object(expresssion data prepared using the function ReadAffy from the library Affy) and a vector defining group assignement of samples. As masking is based on expression values, only expressed probes should be used. As a default they are defined by the affy function mas5calls and condition of being expressed (having "P" value) in at least 90% of samples from each group, but any set of probesets might be submitted with exprlist argument. Probes are analyzed for difference in binding affinity between groups. Each probe is assigned a quality score, based on all pairwise probes' correlations within probesets (for details see vignette or paper). Probes' quality scores, their x and y coordinates on the microarray and the probeset names are stored in a matrix.

Value

A list of two objects will be returned.

probes	A data frame with x,y coordinates, quality score and probeset for each analyzed probe.
notUsed	If PM=TRUE: A vector with unused probes having a lower pm mean value than mm mean value.

Author(s)

Michael Dannemann, Michael Lachmann

References

Dannemann et al, The effects of probe binding affinity differences on gene expression measurements and how to deal with them. Bioinformatics 2009 \ Khaitovich et al,Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees, Science 2005

See Also

overlapExprExtMasks, prepareMaskedAffybatch, mas5calls, plotProbe

Examples

```
data(AffyBatch)
## we provide 20 samples (10 for both human and chimpanzee)
## the first 10 entries are chimpanzee samples the last 10 from human
ind.vec=rep(1:2,each=10)
## mask on AffyBatch with all genes
exmask <-
mask(newAffyBatch,ind=ind.vec,PM=TRUE,useExpr=FALSE)
```

newAffyBatch *AffyBatch with reduced genes*

Description

This data is an AffyBatch object with a subset of 100 genes with human chimpanzee data (cdf hgu95av2) - 10 individuals each.

Usage

```
newAffyBatch
```

Format

AffyBatch object

Source

??

References

Khaitovich et al.,Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees, Science 2005

newCdf *Object of type environment*

Description

The environment object is part of the masked object `newAffyBatch`.

Usage

```
newCdf
```

Format

Object of type environment

Source

??

References

??

overlapExprExtMasks *Error Analysis of Masking Results*

Description

Expression mask results for a range of cutoff values are compared with an external mask (for example a mask based on sequence data) and type 1 and type 2 errors are estimated.

Usage

```
overlapExprExtMasks(probes,seqdata,cutoffs="none",wilcox.ks=FALSE,sample=10,plotCutoffs=TRUE,ver
```

Arguments

probes	A matrix with 3 columns. The first and second column represent the x and y coordinates on the Microarray. The third column contains a quality entry for each probe, e.g. the quality score obtained from mask analysis.
seqdata	A matrix with 3 columns containing x, y coordinates and 0,1 entries in column 3, defining whether a probe has a sequence difference (0) or not.
cutoffs	A vector including all cutoff values for the quality scores of an expression mask that should be used for the error analysis. If no cutoffs are given (default is "none") the cutoffs are the quantiles of the quality scores starting from 0 to 1 in steps of 0.01.
wilcox.ks	Logical, default=FALSE element determining whether the Kolmogorow-Smirnow Test and Wilcoxon Rank Test analysis should be performed (see reference below).
sample	To compare the p value distribution with the Kolmogorow-Smirnow Test and Wilcoxon Rank Test for different cutoffs, the sampling option can be used to compute the quality score distribution for different cutoffs. This value indicates how often the sampling should be performed.
plotCutoffs	Logical, default=TRUE element determining whether the cutoffs should be drawn in the overlap plot.
verbose	Logical. If 'TRUE', it writes out some messages indicating progress. If 'FALSE' nothing should be printed.

Details

The function overlapExprExtMasks compares expression mask results with an external (for example sequence-based) mask and might help to choose a quality score cutoff for masking probes.

Value

A list of five objects will be returned.

type1	A vector of the type 1 error for each cutoff.
type2	A vector of the type 2 error for each cutoff.
confT1	A matrix with the upper (column 1) and lower (column 2) confidence intervals for the type 1 error.

confT2	A matrix with the upper (column 1) and lower (column 2) confidence intervals for the type 2 error.
ksP	If wilcox.ks is 'TRUE', a vector of quality scores from a two sample Kolmogorov-Smirnov comparing distributions of quality score for probes designated as BAD and not in external mask.
wilcoxonP	If wilcox.ks is 'TRUE', a vector of quality scores from a two sample Wilcoxon rank test comparing distributions of quality score for probes designated as BAD and not in external mask.
ksBoot	For each cutoff sample(default=10) times cutoff values for the Kolmogorov-Smirnov test will be generated.
wilcoxBoot	For each cutoff sample(default=10) times cutoff values for the wilcoxon rank sum test will be generated.
cutoffs	List of cutoffs used for the error analysis
testCutoffs	If wilcox.ks is 'TRUE', a list with cutoff information will be provided. The first list entry includes all cutoffs used in the two sample Kolmogorov-Smirnov test and the two sample wilcoxon rank sum test analysis will be produced. A cutoff can appear sample(default=10) times. In theory there should be sample times the number of cutoff values entries in this vector, but usually there are fewer entries, because for certain cutoff values, it is not possible to calculate the exact p value in one of the tests. The second list entry transforms the cutoffs in ranks and can be used for the plotting of the test results.

Author(s)

Michael Dannemann

References

Dannemann et al, The effects of probe binding affinity differences on gene expression measurements and how to deal with them. Bioinformatics 2009

See Also

mask, prepareMaskedAffybatch, plotProbe

Examples

```
## loading mask on all genes (exmask1) of the same dataset
data(exmask)
overlapExSeq <- overlapExprExtMasks(exmask$probes[,1:3],sequenceMask[,c(1,2,4)])

## plot results
plot(overlapExSeq$type1,overlapExSeq$type2,type="l",col="red",
    main="Overlap expression based mask - sequence based mask",xlab="Type 1",ylab="Type 2")
abline(1,-1,col="gray")

## performing wilcoxon rank sum test and Kolmogorov-Smirnov test on
## expression mask with all genes (exmask)
overlapTests <-
  overlapExprExtMasks(exmask$probes[,1:3],sequenceMask[,c(1,2,4)],wilcox.ks=TRUE)
layout(matrix(1:2,ncol=1))
plot(overlapTests$testCutoff[[1]],overlapTests$ksBoot,col="red",main="Kolmogorov-Smirnov Test",xlab="Qualit
     ylab="p value (Kolmogorov-Smirnov Test)",ylim=c(0,1),pch=16,xaxt="n")
```

```
axis(1,at=1:length(unique(overlapTests$testCutoff[[2]])),labels=signif(unique(overlapTests$testCutoff[[2]]),
lines(which(unique(overlapTests$testCutoff[[2]]) %in% overlapTests$testCutoff[[2]]),overlapTests$ksP[!is.na
plot(overlapTests$testCutoff[[1]],overlapTests$wilcoxonBoot,col="green",main="Wilcoxon Rank Sum Test",xlab="
     ylab="p value (Wilcoxon Rank Sum Test)",ylim=c(0,1),pch=16,xaxt="n")
axis(1,at=1:length(unique(overlapTests$testCutoff[[2]])),labels=signif(unique(overlapTests$testCutoff[[2]]),
lines(which(unique(overlapTests$testCutoff[[2]]) %in% overlapTests$testCutoff[[2]]),overlapTests$wilcoxonP[
```

---

plotProbe *Plot probes*

Description

Pairwise plot probes of a probeset.

Usage

```
plotProbe(affy,probeset,probe=NA,probeXY=NA,scan=TRUE,ind,exmask="none",seqmask="none",names=FAL
```

Arguments

affy	An object of class AffyBatch.
probeset	Probe set name (Affymetrix ID).
probe	Number of the main probe.
probeXY	If probe is NA the x and y coordinates of the main probe can be given in the format 'x.y'.
scan	If scan is 'TRUE', each probewise comparison of the probe against all other probes in this probeset will be performed seperately. If scan is 'FALSE', all plots will be plotted in one layout. The layout has 3 columns. If the number of remaining probes that the probe should be compared with is not a multiple of 3, the number of probes will be reduced to the next lower multiple of 3.
ind	Numeric vector, with values 1 and 2, definining group assignement for samples in affy.
exmask	Optional: an expression mask object for this affy batch. Data frame with probe information, for example first element of the output of function mask. Should contain: column 1: probe x-coordinate, column 2:probe y coordinate, column 3: probeset, column 4: quality score: values to based filtering on, probes with values smaller than cutoff are discarded.
seqmask	Optional: a sequence mask object for this mask.
names	If 'TRUE' , the sample names are plotted to identify each individual.

Details

The function plotProbe plots single probe against all other probes of its probe set. The information from the expression based mask, the sequence based mask and the test for the two plotted probes is shown.

Author(s)

Michael Dannemann

References

Dannemann et al, The effects of probe binding affinity differences on gene expression measurements and how to deal with them. Bioinformatics 2009

See Also

mask, overlapExprExtMasks, prepareMaskedAffybatch

Examples

```
data(exmask)
data(AffyBatch)
## plot for one probe comparisons with other probes of the probeset
## for a random probeset
availableProbesets <- as.character(unique(exmask$probes[,4]))
availableProbesets
## scan the plots
## Not run: plotProbe(affy=newAffyBatch,probeset=availableProbesets[22],probe=5,scan=TRUE,ind=rep(1:2,each=
## scan with names=TRUE
## Not run: plotProbe(affy=newAffyBatch,probeset=availableProbesets[22],probe=5,scan=TRUE,ind=rep(1:2,each=
## plot with given x y information
## Not run: plotProbe(affy=newAffyBatch,probeset=availableProbesets[22],probeXY="313.415",scan=TRUE,ind=rep
## all plots in one layout
plotProbe(affy=newAffyBatch,probeset=availableProbesets[22],probe=5,scan=FALSE,ind=rep(1:2,each=10),exmask=
```

`prepareMaskedAffybatch` *Creating a new CDF*

Description

Create a new affyBatch, withprobes and probesetsdefined by mask.

Usage

```
prepareMaskedAffybatch(affy,cdfTablePath,exmask="none",cdfName="new_cdf",exclude=NA,cutoff=0.2)
```

Arguments

affy	An object of class AffyBatch.
cdfTablePath	Location of the probe information table. This is a plain text file with probes to build new cdf. It should contain 3 or 5 columns. Column 1: Probeset ID. Column 2: probe x-coordinate. Column 3: probes y-coordinate. Optional column 4: Mismatch probe x-coordinate. Optional column 5: Mismatch probe y coordinate.
exmask	Data frame with probe information, for example first element of the output of function mask. Should contain: column 1: probe x-coordinate, column 2:probe y coordinate, column 3 :probeset, column 4: quality score: values to based filtering on, probes with values smaller than cutoff are discarded.
cdfName	Name for the new CDF.

cutoff	With mask.object, defines the minimum quality score necessary for a probe to qualify to the new cdf.
exclude	Default 'NA'. If exclude set to a number>0, probesets with less than 'exclude' probes remaining after masking are excluded from the new affyBatch object.

Details

The function prepareMaskedAffybatch creates a new affyBatch including only the probes remaining after masking. Set of probes might be defined by a txt file, with cdfTablePath argument, or by a data frame mask.object and cutoff the probes have to exceed to be used in the new cdf.

Value

newAffyBatch	A list with an affyBatch object and an environment for the new CDF identifier.

Author(s)

Michael Lachmann, Mehmet Somel, Michael Dannemann, Anna Lorenc

References

Dannemann et al, The effects of probe binding affinity differences on gene expression measurements and how to deal with them. Bioinformatics 2009

See Also

mask, overlapExprExtMasks, plotProbe

Examples

```
## prepare new affy batch after masking
## using the expression mask object from the example of the mask function
data(AffyBatch)
data(exmask)
## AffyBatch object before masking
newAffyBatch
affyBatchAfterMasking <-
   prepareMaskedAffybatch(affy=newAffyBatch,exmask=exmask$probes)
## AffyBatch object after masking
affyBatchAfterMasking
```

---

sequenceMask *Object containing sequence information for probes.*

Description

This data is a table with information about sequence difference between human and chimpanzee for all available probes.

Usage

```
sequenceMask
```

Format

data.frame.

Source

??

References

??

Index