Package ‘pathprint’

June 14, 2012

Type Package

Title Pathway fingerprinting for analysis of gene expression arrays

Description Algorithms to convert a gene expression array provided as an expression table or a GEO reference to a ’pathway fingerprint’; a vector of discrete ternary scores representing high (1), low(-1) or insignificant (0) expression in a suite of pathways

Author Gabriel Altschuler, Uma Saxena

Maintainer Gabriel Altschuler <galtschu@hsph.harvard.edu>

Version 1.2.3

Depends R (>= 2.10)

Suggests ALL

Date 2012-06-14

License GPL

LazyLoad yes

LazyData yes

R topics documented:

`chipframe` *Probe to Entrez Gene ID mappings*

Description

Probe to Entrez Gene ID mappings for platforms covered by the pathway fingerprint

Usage

```
chipframe
```

Format

List with an entry for each GEO ID covered by pathprint (31 elements), each containing a list

`ann` dataframe containing array probe to Entrez Gene ID mappings

`title` character, array name

Details

The platform IDs correspond to GEO references `http://www.ncbi.nlm.nih.gov/geo/`

Source

Annotations obtained from the AILUN server `http://ailun.stanford.edu/`

References

Chen R., Li L., and Butte AJ (2007) AILUN: Reannotating Gene Expression Data Automatically, *Nature Methods*, 4(11), 879

See Also

`customCDFAnn`

Examples

```
names(chipframe)
chipframe$GPL570$title
head(chipframe$GPL570$ann)
```

`consensusDistance` *Calculate a distribution of distances from a consensus fingerprint*

Description

Calculates the distance from a consensus for a series of pathway fingerprints, accounting only for significantly high or low (-1 or 1) pathways in the consensus

Usage

```
consensusDistance(consensus, fingerprintframe)
```

Arguments

`consensus`	consensus fingerprint
`fingerprintframe`	dataframe of sample fingerprints from which the distance will be calculated

Details

The consensus fingerprint can be calculated using `consensusFingerprint` or alternatively can be a single fingerprint vector

Value

A dataframe with rows corresponding to each sample contained in the fingerprintframe with the following columns

`distance`	Manhattan distance of sample from the consensus fingerprint, scaled by the maximum possible distance
`pvalue`	p-value representing the probabilty that the samples are not phenotypically matched. N.B. this is only valid when the fingerprint frame represents a sufficiently broad coverage of phenotypes, e.g. the GEO corpus. This p-value is based on an assumption that the distances are normally distributed

Author(s)

Gabriel Altschuler

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

See Also

`consensusFingerprint`

Examples

```
# search for pluripotent arrays
# create consensus fingerprint for pluripotent samples
pluripotent.consensus<-consensusFingerprint(
GEO.fingerprint.matrix[,pluripotents.frame$GSM], threshold=0.9)

# calculate distance from the pluripotent consensus
geo.pluripotentDistance<-consensusDistance(
pluripotent.consensus, GEO.fingerprint.matrix)

# plot histograms
par(mfcol = c(2,1), mar = c(0, 4, 4, 2))
geo.pluripotentDistance.hist<-hist(geo.pluripotentDistance[,"distance"],
nclass = 50, xlim = c(0,1), main = "Distance from pluripotent consensus")
par(mar = c(7, 4, 4, 2))
hist(geo.pluripotentDistance[pluripotents.frame$GSM, "distance"],
breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
main = "", xlab = "above: all GEO, below: curated pluripotent samples")
```

consensusFingerprint

Construct a consensus fingerprint

Description

Produces a pathway fingerprint that represents the consensus of a series of pathway fingerprints, according to a user-defined threshold

Usage

```
consensusFingerprint(fingerprintframe, threshold)
```

Arguments

`fingerprintframe`
matrix of fingerprints from which the consensus will be calculated

`threshold` threshold value (between 0 and 1)

Details

For each pathway the mean fingerprint score, m, is calculated, and the consensus defined as
+1 if m > threshold
-1 if m < threshold
0 otherwise

Value

Vector of consensus pathway fingerprint scores with names corresponding to pathways

Author(s)

Gabriel Altschuler

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

See Also

`consensusDistance`

Examples

```
# search for pluripotent arrays
# load fingerprint matrix and pluripotent reference

# create consensus fingerprint
pluripotent.consensus<-consensusFingerprint(
GEO.fingerprint.matrix[,pluripotents.frame$GSM], threshold=0.9)

# calculate distance from the pluripotent consensus
geo.pluripotentDistance<-consensusDistance(pluripotent.consensus,
GEO.fingerprint.matrix)

# plot histograms
par(mfcol = c(2,1), mar = c(0, 4, 4, 2))
geo.pluripotentDistance.hist<-hist(geo.pluripotentDistance[,"distance"],
nclass = 50, xlim = c(0,1), main = "Distance from pluripotent consensus")
par(mar = c(7, 4, 4, 2))
hist(geo.pluripotentDistance[pluripotents.frame$GSM, "distance"],
breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
main = "", xlab = "above: all GEO, below: curated pluripotent samples")


# annotate top 100 matches not in original seed with metadata
geo.pluripotentDistance.noSeed<-geo.pluripotentDistance[
!(rownames(geo.pluripotentDistance) %in% pluripotents.frame$GSM),
]

top.noSeed.meta<-GEO.metadata.matrix[
match(head(rownames(geo.pluripotentDistance.noSeed), 1000),
GEO.metadata.matrix$GSM),
]
head(top.noSeed.meta[,c("GSM", "GPL", "Source")],10)
```

`customCDFAnn` *Map probes to Entrez Gene IDs*

Description

Annotates an expression array with entrez gene IDs, averaging to resolve redundancies

Usage

```
customCDFAnn(data, ann)
```

Arguments

`data`	expression dataframe
`ann`	annotation dataframe

Details

Maps array probes to a unique list of Entrez Gene IDs. Rhe mean expression value is used for mutiple probes mapping to the same gene.

Value

Returns a dataframe with a column for each column of the input data and rownames as unique entrez IDs.

Author(s)

Gabriel Altschuler

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

See Also

`chipframe`, `single.chip.enrichment`

Examples

```
# load ALL dataset
require(ALL)
data(ALL)
annotation(ALL)

# The chip used was the Affymetrix Human Genome U95 Version 2 Array
# The correspending GEO ID is GPL8300

# Extract portion of the expression matrix
ALL.exprs<-exprs(ALL)
ALL.exprs.sub<-ALL.exprs[,1:5]

# Annotate with Entrez Gene IDs,
ALL.exprs.sub.entrez<-customCDFAnn(ALL.exprs.sub, chipframe$GPL8300$ann)
head(ALL.exprs.sub.entrez)
```

diffPathways *Detect differentially activated pathways between fingerprints*

Description

A function to return pathways consistently differentially expressed between two groups of pathway fingerprints

Usage

```
diffPathways(fingerprints, fac, threshold)
```

Arguments

`fingerprints`	matrix of fingerprints, the number of columns should correspond to the length of fac
`fac`	vector of characters or factors, in an order corresponding to the order of columns in the fingerprint matrix. Contains two levels, denoting the groups to be compared.
`threshold`	numeric, between 0 and 2 - the threshold at which to assign an average difference in pathway usage.

Details

The vector of factors must contain only two levels (or two unique values for a character vector).

Value

Returns a list of the rownames (i.e. pathways for the pathway fingerprint) corresponding to the rows for which the difference in the means between the two groups is greater than the threshold value. For a ternary fingerprint (-1,0,1), setting the threshold between 0.5 and 1 ensures that rownames are selected that differ across the majority of the arrays in the two groups. with values closer to 1 representing higher stringency. This can break down and allow false positives in the case where one group contains a significant but minority number of +1 and the other -1s.

Author(s)

Gabriel Altschuler

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

See Also

`exprs2fingerprint`, `consensusDistance`, `consensusFingerprint`

Examples

```
# Use ALL dataset as an example

require(ALL)
data(ALL)
annotation(ALL)

# The chip used was the Affymetrix Human Genome U95 Version 2 Array
# The corresponding GEO ID is GPL8300

# Analyze patients with ALL1/AF4 and BCR/ABL translocations
ALL.eset <- ALL[, ALL$mol.biol %in% c("BCR/ABL", "ALL1/AF4")]
ALL.exprs<-exprs(ALL.eset)

patient.type<-as.character(ALL$mol.biol[
ALL$mol.biol %in% c("BCR/ABL", "ALL1/AF4")])

# Process fingerprints
ALL.fingerprint<-exprs2fingerprint(exprs = ALL.exprs,
platform = "GPL8300",
species = "human",
progressBar = TRUE
)

color.map <- function(mol.biol) {
if (mol.biol=="ALL1/AF4") "#00FF00" else "#FF00FF"
}
patientcolors <- sapply(ALL$mol.biol[
ALL$mol.biol %in% c("BCR/ABL", "ALL1/AF4")],
function(x){
if (x == "ALL1/AF4") "#00FF00" else "#FF00FF"
})


# define differentially activated pathways between the two groups
signif.pathways<-diffPathways(ALL.fingerprint,
fac = patient.type,
threshold = 0.6)

# draw heatmap
heatmap(ALL.fingerprint[signif.pathways,],
ColSideColors = patientcolors,
col = c("blue", "white", "red"),
scale = "none", mar = c(10,20),
cexRow = 0.75)
title(sub = "Pathways differentially activated in patients
 with ALL1/AF4 (green) and BCR/ABL(purple) translocations",
      cex.sub = 0.75)
```

---

exprs2fingerprint *Create a pathway fingerprint from a gene expression table.*

Description

The function converts the gene expression values to a ternary matrix of pathway expression values, (-1,0,1) corresponding to (low, background, high). This is based on applying a pre-calculated threshold to pathway enrichment scores.

Usage

```
exprs2fingerprint(exprs, platform, species, progressBar = TRUE)
```

Arguments

`exprs`	matrix containing a probe expression table, can be one or more columns
`platform`	microarray platform GEO ID
`species`	character string to define the species of the experiment, see details.
`progressBar`	logical. If TRUE, a progress bar is displayed while the script is running

Details

exprs should be a matrix or dataframe of the expression values, with rownames containing probe names and colnames the experiment IDs. Platforms should be of the type listed in GEO (e.g. "GPL570"). Species can be full latin names
"Homo sapiens", "Mus musculus", "Rattus norvegicus", "Danio rerio", "Drosophila melanogaster", "Caenorhabditis elegans".
or corresponding common-use names
"human", "mouse", "rat", "zebrafish", "drosophila", "C.elegans".
The array is first annotated with Entrez Gene IDs using annotations contained in `chipframe`. Pathway expression scores are calculated by the mean-squared rank of the gene expression and normalized against the appropriate distribution for the given platform in the GEO corpus. There is a progressBar to track the script, can be set to FALSE for (possibly) marginally faster running

Value

Returns a dataframe containing the pathway fingerprint for each of column in the expression table. Rownames correspond to pathways and colnames to the experiment IDs.

Author(s)

Gabriel Altschuler

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

See Also

`consensusFingerprint`, `single.chip.enrichment`, `customCDFAnn`, `thresholdFingerprint`

Examples

```
# Use ALL dataset as an example
require(ALL)
data(ALL)
annotation(ALL)

# The chip used was the Affymetrix Human Genome U95 Version 2 Array
# The corresponding GEO ID is GPL8300

# Extract portion of the expression matrix
ALL.exprs<-exprs(ALL)
ALL.exprs.sub<-ALL.exprs[,1:5]

# Process fingerprints
ALL.fingerprint<-exprs2fingerprint(exprs = ALL.exprs.sub,
platform = "GPL8300",
species = "human",
progressBar = TRUE
)

head(ALL.fingerprint)
```

`genesets` *Names of genesets used in pathprint*

Description

An index to the genesets used in pathprint for each species, referenced by common and latin name

Usage

```
genesets
```

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
genesets
```

GEO.fingerprint.matrix

Matrix of GEO fingerprints

Description

A matrix containing Pathway Fingerprint vectors have been pre-calculated for ~160,000 publicly available arrays from the GEO corpus, spanning 6 species and 31 platforms

Usage

```
GEO.fingerprint.matrix
```

Source

Primary data was retrieved from http://www.ncbi.nlm.nih.gov/geo/

References

Barrett et al. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic acids research* (2007) vol. 35 (Database issue) pp. D760-5
Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
# create consensus fingerprint for pluripotent samples
pluripotent.consensus<-consensusFingerprint(
GEO.fingerprint.matrix[,pluripotents.frame$GSM], threshold=0.9)

# calculate distance from the pluripotent consensus
geo.pluripotentDistance<-consensusDistance(
pluripotent.consensus, GEO.fingerprint.matrix)

# plot histograms
par(mfcol = c(2,1), mar = c(0, 4, 4, 2))
geo.pluripotentDistance.hist<-hist(geo.pluripotentDistance[,"distance"],
nclass = 50, xlim = c(0,1), main = "Distance from pluripotent consensus")
par(mar = c(7, 4, 4, 2))
hist(geo.pluripotentDistance[pluripotents.frame$GSM, "distance"],
breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
main = "", xlab = "above: all GEO, below: pluripotent samples")


# annotate top 100 matches not in original seed with metadata
geo.pluripotentDistance.noSeed<-geo.pluripotentDistance[
!(rownames(geo.pluripotentDistance) %in% pluripotents.frame$GSM),
]


top.noSeed.meta<-GEO.metadata.matrix[
match(head(rownames(geo.pluripotentDistance.noSeed), 1000),
GEO.metadata.matrix$GSM),
]
```

```
head(top.noSeed.meta[,c("GSM", "GPL", "Source")],10)
```

`GEO.metadata.matrix` *Matrix of GEO metadata*

Description

Metadata associated with samples contained in the GEO fingerprint matrix, it includes experiment IDs, platform, species and a selection of the record description provided by the GEO database

Usage

```
GEO.metadata.matrix
```

Format

A data frame with 158487 observations on the following 7 variables.

`GSM` GEO sample ID

`GSE` GEO series ID

`GPL` GEO platform ID

`Species` GEO description - Species

`Title` GEO description - Title

`Source` GEO description - Source

`Characteristics` GEO description - Characteristic

Source

http://www.ncbi.nlm.nih.gov/geo/

References

Barrett et al. NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucleic acids research (2007) vol. 35 (Database issue) pp. D760-5
Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
# create consensus fingerprint for pluripotent samples
pluripotent.consensus<-consensusFingerprint(
GEO.fingerprint.matrix[,pluripotents.frame$GSM], threshold=0.9)

# calculate distance from the pluripotent consensus
geo.pluripotentDistance<-consensusDistance(
pluripotent.consensus, GEO.fingerprint.matrix)

# plot histograms
par(mfcol = c(2,1), mar = c(0, 4, 4, 2))
geo.pluripotentDistance.hist<-hist(geo.pluripotentDistance[,"distance"],
nclass = 50, xlim = c(0,1), main = "Distance from pluripotent consensus")
par(mar = c(7, 4, 4, 2))
hist(geo.pluripotentDistance[pluripotents.frame$GSM, "distance"],
breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
main = "", xlab = "above: all GEO, below: pluripotent samples")


# annotate top 100 matches not in original seed with metadata
geo.pluripotentDistance.noSeed<-geo.pluripotentDistance[
!(rownames(geo.pluripotentDistance) %in% pluripotents.frame$GSM),
]


top.noSeed.meta<-GEO.metadata.matrix[
match(head(rownames(geo.pluripotentDistance.noSeed), 1000),
GEO.metadata.matrix$GSM),
]
head(top.noSeed.meta[,c("GSM", "GPL", "Source")],10)
```

hyperPathway *Produces a list of pathways enrichments calculated using the hypergeometric distribution*

Description

A function that returns a list enrichment statistics for a gene list in a set of pathways or modules

Usage

```
hyperPathway(genelist, geneset, Nchip)
```

Arguments

genelist	
geneset	list of pathways or genesets over which to assess statistic
Nchip	The background number of genes on which to base the hypergeometric distribution

Author(s)

Gabriel Altschuler

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

pathprint *Pathway fingerprinting for analysis of gene expression arrays*

Description

Algorithms to convert a gene expression array provided as an expression table to a 'pathway fingerprint'. The pathway fingerprint provides an unbiased, consistent annotation of expression data as a molecular phenotype, represented by activation status in 633 pathways. This is a vector of discrete ternary scores to represent high (1), low(-1) or insignificant (0) expression in a suite of pathways. Systematic definition of these functional relationships provides a tool for searching a pathway activation map of gene expression spanning species and technologies.

Details

Package:	pathprint
Type:	Package
Version:	1.2.3
Date:	2012-06-14
License:	GPL
LazyLoad:	yes

Author(s)

Gabriel Altschuler, Uma Saxena.
Maintainer: Gabriel Altschuler <galtschu@hsph.harvard.edu>

References

Pathway activation mapping for large scale comparison of gene expression, Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

See Also

exprs2fingerprint, consensusFingerprint, consensusDistance

Examples

```
# Use fingerprints to analyze the ALL dataset
require(ALL)
data(ALL)
annotation(ALL)

# The chip used was the Affymetrix Human Genome U95 Version 2 Array
# The corresponding GEO ID is GPL8300

# Extract portion of the expression matrix
ALL.exprs<-exprs(ALL)
ALL.exprs.sub<-ALL.exprs[,1:5]

# Process fingerprints
ALL.fingerprint<-exprs2fingerprint(exprs = ALL.exprs.sub,
platform = "GPL8300",
species = "human",
progressBar = TRUE
)

head(ALL.fingerprint)


####
# Construct consensus fingerprint based on pluripotent records
# Use this consensus to find similar arrays

pluripotent.consensus<-consensusFingerprint(
GEO.fingerprint.matrix[,pluripotents.frame$GSM], threshold=0.9)

# calculate distance from the pluripotent consensus
geo.pluripotentDistance<-consensusDistance(
pluripotent.consensus, GEO.fingerprint.matrix)

# plot histograms
par(mfcol = c(2,1), mar = c(0, 4, 4, 2))
geo.pluripotentDistance.hist<-hist(geo.pluripotentDistance[,"distance"],
nclass = 50, xlim = c(0,1), main = "Distance from pluripotent consensus")

par(mar = c(7, 4, 4, 2))
hist(geo.pluripotentDistance[pluripotents.frame$GSM, "distance"],
breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
main = "", xlab = "above: all GEO, below: pluripotent samples")


# annotate top 100 matches not in original seed with metadata
geo.pluripotentDistance.noSeed<-geo.pluripotentDistance[
!(rownames(geo.pluripotentDistance) %in% pluripotents.frame$GSM),
]

top.noSeed.meta<-GEO.metadata.matrix[
match(head(rownames(geo.pluripotentDistance.noSeed), 1000),
GEO.metadata.matrix$GSM),
]
head(top.noSeed.meta[,c("GSM", "GPL", "Source")],10)
```

---

pathprint.Ce.gs *Pathprint genesets - C. elegans*

Description

Pathways and genesets used by pathprint for *C. elegans* arrays, referenced by Entrez Gene ID

Usage

```
pathprint.Ce.gs
```

Details

Gene sets were inferred by homology from the human genesets, pathprint.Hs.gs, using the HomoloGene database, www.ncbi.nlm.nih.gov/homologene

Source

O. Hofmann

References

Sayers et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research (2011) vol. 39 (Database issue) pp. D38-51
Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
pathprint.Ce.gs[grep("ZN175_7728", names(pathprint.Ce.gs))]
```

pathprint.Dm.gs *Pathprint genesets - D. melanogaster*

Description

Pathways and genesets used by pathprint for *D. melanogaster* arrays, referenced by Entrez Gene ID

Usage

```
pathprint.Dm.gs
```

Details

Gene sets were inferred by homology from the human genesets, pathprint.Hs.gs, using the HomoloGene database, www.ncbi.nlm.nih.gov/homologene

Source

O. Hofmann

References

Sayers et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research (2011) vol. 39 (Database issue) pp. D38-51
Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
pathprint.Dm.gs[grep("ZN175_7728", names(pathprint.Dm.gs))]
```

`pathprint.Dr.gs` *Pathprint genesets - D. rerio*

Description

Pathways and genesets used by pathprint for *D. rerio* arrays, referenced by Entrez Gene ID

Usage

```
pathprint.Dr.gs
```

Details

Gene sets were inferred by homology from the human genesets, `pathprint.Hs.gs`, using the HomoloGene database, `www.ncbi.nlm.nih.gov/homologene`

Source

O. Hofmann

References

Sayers et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research (2011) vol. 39 (Database issue) pp. D38-51
Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
pathprint.Dr.gs[grep("ZN175_7728", names(pathprint.Dr.gs))]
```

`pathprint.Hs.gs` *Pathprint genesets - H. sapiens*

Description

Pathways and genesets used by pathprint for *H.sapiens* arrays, referenced by Entrez Gene ID

Usage

```
pathprint.Hs.gs
```

Source

O. Hofmann

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
pathprint.Hs.gs[grep("ZN175_7728", names(pathprint.Hs.gs))]
```

pathprint.Mm.gs *Pathprint genesets - M. musculus*

Description

Pathways and genesets used by pathprint for *M. musculus arrays*, referenced by Entrez Gene ID

Usage

```
pathprint.Mm.gs
```

Details

Gene sets were inferred by homology from the human genesets, `pathprint.Hs.gs`, using the HomoloGene database, `www.ncbi.nlm.nih.gov/homologene`

Source

O. Hofmann

References

Sayers et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research (2011) vol. 39 (Database issue) pp. D38-51
Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
pathprint.Mm.gs[grep("ZN175_7728", names(pathprint.Mm.gs))]
```

pathprint.Rn.gs *Pathprint genesets - R. norvegicus*

Description

Pathways and genesets used by pathprint for *R. norvegicus* arrays, referenced by Entrez Gene ID

Usage

```
pathprint.Rn.gs
```

Details

Gene sets were inferred by homology from the human genesets, `pathprint.Hs.gs`, using the HomoloGene database, `www.ncbi.nlm.nih.gov/homologene`

Source

O. Hofmann

References

Sayers et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research (2011) vol. 39 (Database issue) pp. D38-51
Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
pathprint.Rn.gs[grep("ZN175_7728", names(pathprint.Rn.gs))]
```

`platform.thresholds` *Pathway fingerprint threshold values*

Description

Ternary threshold values for conversion of continuous geneset enrichment scores to discrete Pathway Fingerprint scores - high (1), mid (0), low (-1) for each geneset and platform covered by the Pathway Fingerprint.

Usage

```
platform.thresholds
```

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

Examples

```
head(platform.thresholds[[1]])
```

`pluripotents.frame` *Manually curated list of pluripotent arrays*

Description

A manually compiled list of pluripotent arrays (induced pluripotent cells and embryonic stem cells) together with their GEO IDs and descriptions

Usage

```
pluripotents.frame
```

Format

A data frame with 278 observations on the following 5 variables.

`GSM` GEO sample ID

`GSE` GEO series ID

`GPL` GEO platform ID

`source` GEO description - Source

`Characteristics` GEO description - Characteristic

Source

http://www.ncbi.nlm.nih.gov/geo/

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

See Also

`consensusDistance`, `consensusFingerprint`

Examples

```
head(pluripotents.frame)

# Use pathway fingerprints to search for
# additional pluripotent arrays across GEO
# create consensus pluripotent fingerprint
pluripotent.consensus<-consensusFingerprint(
GEO.fingerprint.matrix[,pluripotents.frame$GSM], threshold=0.9)

# calculate distance from the pluripotent consensus
geo.pluripotentDistance<-consensusDistance(
pluripotent.consensus, GEO.fingerprint.matrix)

# plot histograms
par(mfcol = c(2,1), mar = c(0, 4, 4, 2))
geo.pluripotentDistance.hist<-hist(geo.pluripotentDistance[,"distance"],
nclass = 50, xlim = c(0,1), main = "Distance from pluripotent consensus")
par(mar = c(7, 4, 4, 2))
hist(geo.pluripotentDistance[pluripotents.frame$GSM, "distance"],
breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
main = "", xlab = "above: all GEO, below: pluripotent samples")

# annotate top 100 matches not in original seed with metadata
geo.pluripotentDistance.noSeed<-geo.pluripotentDistance[
!(rownames(geo.pluripotentDistance) %in% pluripotents.frame$GSM),
]

top.noSeed.meta<-GEO.metadata.matrix[
match(head(rownames(geo.pluripotentDistance.noSeed), 1000),
GEO.metadata.matrix$GSM),
]
head(top.noSeed.meta[,c("GSM", "GPL", "Source")],10)
```

single.chip.enrichment
Calculate enrichment of a list of genesets in an array

Description

Function to assess enrichment of gene sets in an array or matrix of arrays using various summary statistics

Usage

```
single.chip.enrichment(exprs,
geneset,
transformation = "rank",
statistic = "mean",
normalizedScore = FALSE,
progressBar = TRUE)
```

Arguments

`exprs` An expression matrix, rownames correspond to gene ids used in the list of genesets

`geneset` list of pathways or genesets over which to assess statistic

`transformation` Initial transformation applied to each column of exprs, can be one of "rank", "squared.rank" or "log.rank"

`statistic` Summary statistic to be applied, either "mean" or "median"

`normalizedScore` Logical. If statistic = "mean" and normalizedScore = TRUE, option to calculate a parametric significance score based on the expected distribution of scores. Other summary statistics currently not supported

`progressBar` Logical. Shows progress of script, good to check running okay, set to FALSE for possible faster running

Details

This is the worker function for exprs2fingerprint, in conjuction with an exprs based on Entrez Gene IDs and the standard pathprint genesets e.g. `pathprint.Hs.gs`. The (un-normalized) results are passed onto thresholdFingerprint to produce the Pathway Fingerprint scores

Value

Matrix containing pathway enrichment scores for each sample in the exprs input matrix. Rownames are genesets and colnames are the columns of the exprs matrix.

Author(s)

Gabriel Altschuler

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

See Also

exprs2fingerprint

Examples

```
# Compare continuous pathway enrichment values to Pathway Fingerprint scores

# Use ALL dataset as an example

require(ALL)
data(ALL)
annotation(ALL)

# The chip used was the Affymetrix Human Genome U95 Version 2 Array
# The corresponding GEO ID is GPL8300

# Analyze patients with ALL1/AF4 and BCR/ABL translocations
ALL.eset <- ALL[, ALL$mol.biol %in% c("BCR/ABL", "ALL1/AF4")]
ALL.exprs<-exprs(ALL.eset)

patient.type<-as.character(ALL$mol.biol[
ALL$mol.biol %in% c("BCR/ABL", "ALL1/AF4")])

# Process fingerprints
ALL.fingerprint<-exprs2fingerprint(exprs = ALL.exprs,
platform = "GPL8300",
species = "human",
progressBar = TRUE
)

color.map <- function(mol.biol) {
if (mol.biol=="ALL1/AF4") "#00FF00" else "#FF00FF"
}
patientcolors <- sapply(ALL$mol.biol[
ALL$mol.biol %in% c("BCR/ABL", "ALL1/AF4")],
function(x){
if (x == "ALL1/AF4") "#00FF00" else "#FF00FF"
})


# define list of differentially activated pathways between the two groups
signif.pathways<-diffPathways(ALL.fingerprint,
fac = patient.type,
threshold = 0.6)

# draw heatmap
heatmap(ALL.fingerprint[signif.pathways,],
ColSideColors = patientcolors,
col = c("blue", "white", "red"),
scale = "none", mar = c(10,20),
cexRow = 0.75)
title(sub = "Pathways differentially activated in patients
 with ALL1/AF4 (green) and BCR/ABL(purple) translocations",
      cex.sub = 0.75)


######
# Compare to continous values
ALL.exprs.entrez <- customCDFAnn(ALL.exprs, chipframe$GPL8300$ann)
ALL.enrichment <- single.chip.enrichment(exprs = ALL.exprs.entrez,
geneset = pathprint.Hs.gs,
transformation = "squared.rank",
statistic = "mean",
normalizedScore = FALSE,
progressBar = TRUE
)

heatmap(ALL.enrichment[signif.pathways,],
ColSideColors = patientcolors,
col = colorRampPalette(c("blue", "white", "red"))(100),
scale = "row", mar = c(10,20),
cexRow = 0.75)
title(sub = "Continuous pathway enrichment scores for patients
 with ALL1/AF4 (green) and BCR/ABL(purple) translocations",
      cex.sub = 0.75)
```

thresholdFingerprint *Apply threshold values to produce a Pathway Fingerprint*

Description

Function to produce ternary threshold values, Pathway Fingerprint scores, from continuous geneset enrichment values. Returns ternary scores for each pathway, high (1), mid (0), low (-1)

Usage

```
thresholdFingerprint(SCE, platform)
```

Arguments

SCE	Pathway enrichment matrix from single.chip.enrichment
platform	GEO platform ID for array used

Details

The thresholds have been pre-calculated and optimized against a panel of tissue samples (see ref).

Value

Matrix containing ternary scores for each sample in the SCE input matrix. Rownames are genesets and colnames are the columns of the SCE matrix.

Author(s)

Gabriel Altshuler

References

Pathway activation mapping for large scale comparison of gene expression Altschuler G, Hofmann O, Payne R, Kalatskaya I, Stein L, Cai T, Hide W *in preparation*

See Also

exprs2fingerprint, platform.thresholds

Examples

```
# Comparing workflows

# 1. Pathway Fingerprint scores from exprs2fingerprint

# Use ALL dataset as an example

require(ALL)
data(ALL)
annotation(ALL)

# The chip used was the Affymetrix Human Genome U95 Version 2 Array
# The corresponding GEO ID is GPL8300

# Analyze patients with ALL1/AF4 and BCR/ABL translocations
ALL.eset <- ALL[,1:5]
ALL.exprs<-exprs(ALL.eset)
# Process fingerprints
ALL.fingerprint<-exprs2fingerprint(exprs = ALL.exprs,
platform = "GPL8300",
species = "human",
progressBar = TRUE
)

# 2. Thresholded pathway enrichment values

# Annotate
ALL.exprs.entrez <- customCDFAnn(ALL.exprs, chipframe$GPL8300$ann)

# Pathway enrichment
ALL.enrichment <- single.chip.enrichment(exprs = ALL.exprs.entrez,
geneset = pathprint.Hs.gs,
transformation = "squared.rank",
statistic = "mean",
normalizedScore = FALSE,
progressBar = TRUE
)

# Threshold
ALL.enrichment.threshold <- thresholdFingerprint(
ALL.enrichment, "GPL8300")

# Compare 1. and 2.
```

```
all.equal(ALL.enrichment.threshold, ALL.fingerprint)
```

Index