Package 'rsgcc'

February 20, 2015

Type Package

Title Gini methodology-based correlation and clustering analysis of microarray and RNA-Seq gene expression data

Version 1.0.6

Author Chuang Ma, Xiangfeng Wang

Maintainer Chuang Ma <chuangma2006@gmail.com>

URL http://www.cmbb.arizona.edu/

Depends R (>= 2.15.1), biwt, cairoDevice, fBasics, grDevices, gplots, gWidgets, gWidgetsRGtk2, minerva, parmigene, stringr, snowfall

Suggests bigmemory, ctc

Description This package provides functions for calculating associations between two genes with five correlation methods(e.g., the Gini correlation coefficient [GCC], the Pearson's product moment correlation coefficient [PCC], the Kendall tau rank correlation coefficient [KCC], the Spearman's rank correlation coefficient [SCC] and the Tukey's biweight correlation coefficient [BiWt], and three non-correlation methods (e.g., mutual information [MI] and the maximal information-based nonparametric exploration [MINE], and the euclidean distance [ED]). It can also been implemented to perform the correlation and clustering analysis of transcriptomic data profiled by microarray and RNA-Seq technologies. Additionally, this package can be further applied to construct gene co-expression networks (GCNs).

LazyLoad yes

License GPL (>= 2)

Date 2013-06-12

NeedsCompilation yes

Repository CRAN

Date/Publication 2013-06-18 07:40:43

R topics documented:

`rsgcc-package` *Gini methodology-based correlation and clustering analysis of microarray and RNA-Seq gene expression data*

Description

This package provides functions for calculating the Gini, the Pearson, the Spearman, the Kendall and Tukey's Biweight correlations, Compared to the other mentioned correlation methods, the GCC may perform better to detect regulatory relationships from gene expression data. In addition, the GCC also has some other advantageous merits, such as independent of distribution forms, more capable of detecting non-linear relationships, more tolerant to outliers and less dependence on sample size. For more information about these correlation methods, please refer to (Ma and Wang, 2012). This package also provides an graphical user interface (GUI) to perform clustering analysis of microarray and RNA-Seq data in a coherent step-by-step manner.

Details

Package:	rsgcc
Type:	Package
Version:	1.0.6
Date:	2013-06-12
License:	GPL(>=2)

Note

1) The implement of rsgcc requires several R packages developed by other developers(e.g., biwt, cairoDevice, fBasics, snowfall, grDevices, gplots, gWidgets, gWidgetsRGtk2, stringr, ctc). Please make sure that these packages have been successfully installed before loading the rsgcc package.

2) A general method to install a new package on the computer is to use the command:

install.package("packagename").

Some other methods can be found at http://math.usask.ca/~longhai/software/installrpkg.html.

For the installation of ctc package, please use the biocLite resource with the following commands:

source("http://bioconductor.org/biocLite.R")

biocLite("ctc")

3) To run the "rsgcc.gui()" function, please do remember to select the GUI toolkit "gWidgetsRGtk2".

4) Bug reports and suggestions/questions can be sent to Chuang Ma (chuangma2006@gmail.com) or Xiangfeng Wang (xwang1@cals.arizona.edu).

Author(s)

Chuang Ma, Xiangfeng Wang. Maintainer: Chuang Ma <chuangma2006@gmail.com>

References

Chuang Ma, Xiangfeng Wang. Application of the Gini correlation coefficient to infer regulatory relationships in transcriptome analysis. Plant Physiology, 2012, 160(1):192-203.

`adjacencymatrix` *adjacency matrix calculation*

Description

This function generates the adjacency matrix for network re-construction from gene expression data with different methods including Gini correlation (GCC), Pearson correlation (PCC), Spearman correlation (SCC), Kendall correlation (KCC),Tukey's biweight correlation coefficient (BiWt), mutual information (MI), and maximal information-based nonparametric exploration (MINE) statistic methods. Euclidean distance (ED) between two genes can also be calculated. It was implemented these methods in C language and parallel mode, and thus is greatly faster than the cor.matrix function.

Usage

```
adjacencymatrix(mat, genes.row = NULL, genes.col = NULL,
     method = c("GCC", "PCC", "SCC", "KCC", "BiWt", "MI", "MINE", "ED"),
     k = 3, cpus = 1,
     saveType = "matrix",
     backingpath = NULL,
```

```
    backingfile = "adj_mat",
    descriptorfile = "adj_desc",
    ... )
```

Arguments

mat	a data matrix containing gene expression dataset where rows defines for genes and columns for samples.
genes.row	if genes.row and genes.col are not NULL, a subset of genes will be selected for correlation calcuation and set as the rownames of adjacency matrix. Currently, doesn't work for BiWt and MINE
genes.col	if genes.row and genes.col are not NULL, a subset of genes will be selected for correlation calcuation and set as the colnames of adjacency matrix.Currently, doesn't work for BiWt and MINE
method	a method used to calculate the association between a pair of genes.
k	the number of nearest neighbors to be considered for estimating the mutual information. Must be less than the number of columns of mat, and only work for the mutual information(MI) method.
cpus	the number of cpus will be used for calcuation.
saveType	the type (matrix or bigmatrix) specified for the output.
backingpath	the path used to save big matrix. If it is NULL, current working directory will be used. Works only when the saveType is "bigmatrix".
backingfile	the file name of big matrix. Works only when the saveType is "bigmatrix".
descriptorfile	the description file of big matrix. Works only when the saveType is "bigmatrix".
...	Further parameters passed for MINE method. More information can be found in R package minerva.

Value

value	a matrix (or big.matrix) recording the associations between the gene pairs.

Note

1) The mutural information estimation is based on k-nearest neighbor distance (Sales G and Romualdi C, 2012). Thus the parameter k only works for the mutual information method.

2) Two correlations can be produced by the GCC method by reciprocally using the rank and value information of one gene (or variable). Here the correlation with the maximum absolute values is selected for generating the adjacency matrix.

3) More information about the big.matrix can be found in bigmemory package.

References

[1] Ma C and Xiang XF. Application of the Gini correlation coefficient to infer regulatory relationships in transcriptome analysis, Plant Physiology, 2012, 160(1):192-203.

[2] Sales G and Romualdi C. Parmigene-a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics, 2012, 27:1876-1877.

[3] Davide Albanese, Michele Filosi, Roberto Visintainer, et al. minerva and minepy: a C engine for the MINE stuite and its R, Python and MATLAB wrappers. Bioinformatics, 2013, 29(3): 407-408.

[4] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, et al. Detecting novel associations in large data sets. Science, 2011, 334(6062): 1518-1524.

[5] Johanna Hardin, Aya Mitani, Leanne Hicks and Brian VanKoten. A robust measure of correlation between two genes on a microarray. BMC Bioinformatics, 2007, 8:220.

Examples

```
## Not run:
 mat = matrix(rnorm(180), nrow = 10)
 rownames(mat) <- c(1:10)
 colnames(mat) <- c(1:18)
 mat

 #using GCC method to compute the correlation of all gene paris
 adjacencymatrix( mat, method = "GCC", cpus = 2 )

 #using GCC method to compute the correlation of a subset of gene pairs
 adjacencymatrix( mat = mat, genes.row = c(1:5), genes.col = c(5:8), method = "GCC", cpus = 2 )



 #for MI method, k works here.
 adjacencymatrix( mat, method = "MI", k= 3)

## End(Not run)
```

`cor.matrix` *correlation calculation for a set of genes*

Description

This function provides five correlation methods (GCC, PCC, SCC, KCC and BiWt) to calculates the correlations between a set of genes.

Usage

```
cor.matrix(GEMatrix,
      cpus = 1,
      cormethod = c("GCC", "PCC", "SCC", "KCC", "BiWt"),
      style = c("all.pairs", "pairs.between", "adjacent.pairs", "one.pair"),
      var1.id = NA,
      var2.id = NA,
      pernum = 0,
      sigmethod = c("two.sided", "one.sided"),
      output = c("matrix", "paired"))
```

Arguments

GEMatrix	a data matrix containing the gene expression data of a set of genes. Each row of the GEMatrix corresponds to a gene, and each column corresponds to the expression level in a sample.
cpus	the number of cpus used for correlation calcluation.
cormethod	a character string that specifies a correlation method to be used for correlation calculation.
style	a character string that indicates the all or partial genes to be used for correlation calculation.
var1.id	a numeric vector specifying the row numbers of genes.
var2.id	a numeric vector specifying the row numbers of genes. Suppose the var1.id and var2.id are respectively c(1,2) and c(3,6), then the the correlation of gene pairs (G1,G3) and (G2,G6) will be calcuated. For styles of "pairs.between" and "one.pair", this parameter MUST be pre-defined. For the other styles, this parameter can be automatically defined by the program itself.
pernum	the number of permutation test used for calcluating statistical significance level (i.e., p-value) of correlations.
sigmethod	a character string ("two-sided" or "one-sided") that specifies the method used to compute p-value for permutation test.
output	a character string ("matrix" or "paired") that represents the output format of correlations. Specifiying the "matrix" will output two matrix for correlations and p-values, respectively. Specifiying the "paired" will output only one matrix, in which each row provides the information of gene pair, the correlation and p-value.

Details

Given a data matrix (e.g., microarray and RNA-Seq gene expression matrix), calculating correlation with GCC and other correlation methods for partial(or all) individuals (e.g., genes). The statistical significance (i.e., p-value) of each correlation is derived from the permutation test. Parallel computing options are also provided for speeding up the correlation calculation.

Value

A list with the following components:

corMatrix	correlation of gene pairs shown in matrix form. This data matrix is generated only when the output format "matrix" is specified.
pvalueMatrix	p-value of correlations shown in matrix form. This data matrix is generated only when the output format "matrix" is specified.
corpvalueMatrix	correlation and p-values listed in one form. This data matrix is generated only when the output format "paired" is indicated.

Note

(1) The rsgcc provides the RNA-Seq profiled expression level of 100 genes as a sample data to implement cor.matrix, cor.pari and other functions in the package. After running the command: data(rsgcc), the expression data of these genes will be loaded to the GEMatrix "rnaseq". The user can also load the GEMatrix from the gene expression file, which should be in a textual format of a gene expression matrix. An example of the gene expression file(e.g., "/home/rsgcc/geneExpFile.txt") is shown as follow:

sample1 sample2 sample3 sample4

gene1 45 65 77 75

gene2 75 78 83 39

gene3 2 11 10 6

Then the GEMatrix can be obtained by load this gene expression file with the command: x <- as.matrix(read.table("/home/rsgcc/geneExpFile.txt"))

(1) var1.id and var2.id should be defined with the numeric vector format for "pairs.between", or "one.pair" styles.

(2) To perform BiWt, the R package "biwt" should be installed in advance.

(3) To perform the parallel compution, the "snowfall" package in R should be installed in advance.

Author(s)

Chuang Ma, Xiangfeng Wang

See Also

cor.pair, onegcc, cor.test.

Examples

```
## Not run:
   data(rsgcc)        #load the sample data in rsgcc package
   x <- rnaseq[1:4,]  #construct a GEMatrix with the RNA-Seq data of the first four genes

   #run on one CPU for all the possible gene pairs in the GEMatrix "x".
   #do not cacluate the p-value of computated correlations.
   cor.matrix(x, cpus = 1,
              cormethod = "GCC", style = "all.pairs",
              pernum = 0, sigmethod = "two.sided",
              output = "matrix")

   #run on two CPUs, snowfall package should be properly installed.
   #cacluate the p-value of correlations with the 2000 permutation tests.
   #output the results in "paired" format.
   cor.matrix(x, cpus = 2,
              cormethod = "GCC", style = "all.pairs",
              pernum = 2000, sigmethod = "two.sided",
              output = "paired")
```

```
 #calcuate correlation on the pairs between the 1st, 2nd and 3rd genes in the GEMatrix "x".
 cor.matrix(x, cpus = 1,
            cormethod = "GCC", style = "pairs.between",
            var1.id = c(1:3), var2.id = c(1:3),
            pernum = 2000, sigmethod = "two.sided",
            output = "matrix")

#calcuate correlation on the adjacent genes ((G1,G2), (G2,G3), (G3,G4),...) in the GEMatrix "x".
 cor.matrix(x, cpus = 1,
            cormethod = "GCC", style = "adjacent.pairs",
            pernum = 2000, sigmethod = "two.sided",
            output = "matrix")


## End(Not run)
```

`cor.pair` *compute the correlation between two genes*

Description

This function can compute the correlation of a pair of genes with Gini correlation and four other correlation methods. The signficance level (p-value) of computed correlation can be estimated with the permutation test method.

Usage

```
cor.pair(idxvec,
      GEMatrix,
      rowORcol = c("row", "col"),
      cormethod = c("GCC", "PCC", "SCC", "KCC", "BiWt"),
      pernum = 0,
      sigmethod = c("two.sided", "one.sided"))
```

Arguments

`idxvec`	a numer vector containing two elements for the index of genes or samples in GEMatrix (e.g., c(1,2)).
`GEMatrix`	a data matrix containing numeric variables. Example: rows correspond to genes and columns to samples. This parameter is the same as the "GEMatrix" defined for cor.matrix.
`rowORcol`	a character string ("row" or "col") indicating gene expression data will be extracted by rows or columns for correlation calculation. "row": correlation between two genes. "col": correlaiton between two samples.
`cormethod`	a character string that specifies the correlation method to be used for correlation calculation.

pernum	the number of permutation test used for calcluating statistical significance level (i.e., p-value) of correlations.
sigmethod	a character string ("two-sided" or "one-sided") specifying the method used to compute p-value for permutation test.

Value

A list with the following components:

gcc.rankx	a Gini correlation produced by using the rank information of the first gene (i.e., the first element in idxvec).
gcc.ranky	a Gini correlation produced by using the rank information of the second gene (i.e., the second element in idxvec).
gcc.rankx.pvalue	p-value of gcc.rankx.
gcc.ranky.pvalue	p-value of gcc.ranky.
cor	the correlation produced by "PCC", "SCC", "KCC" or "BiWt".
pvalue	the p-value of cor.

Note

(1) To perform BiWt, the R package "biwt" should be installed in advance.

(2) When the cormethod is defined as "GCC", this function will output a list with four numeric elements: gcc.rankx, gcc.ranky, gcc.rankx.pvalue, gcc.ranky.pvalue. Otherwise, it will output a list with two elements (cor and p-value)

Author(s)

Chuang Ma, Xiangfeng Wang

See Also

onegcc, cor.matrix, gcc.corfinal.

Examples

```
   data(rsgcc)        #load the sample data in rsgcc package
   x <- rnaseq[1:4,]  #construct a GEMatrix with the RNA-Seq data of the first four genes

   #compute correlation between the 1st and 4th genes
   corpair <- cor.pair(c(1,4), GEMatrix = x, rowORcol = "row",
                       cormethod = "GCC", pernum = 0,
                       sigmethod = "two.sided")
## Not run:
   #compute correlation between the 1st and 4th genes,
   #the p-value of correlation will be estimated with 2,000 permuation test.
   corpair <- cor.pair(c(1,4), GEMatrix = x, rowORcol = "row",
```

```
                        cormethod = "GCC", pernum = 2000,
                        sigmethod = "two.sided")

    #compute correlation between the 1st and 4th samples
    corpair <- cor.pair(c(1,4), GEMatrix = x, rowORcol = "col",
                        cormethod = "GCC", pernum = 0,
                        sigmethod = "two.sided")

## End(Not run)
```

`data` *example of RNA-Seq gene expression data*

Description

RNA-Seq profiled gene expression data of 100 genes

Usage

```
data(rsgcc)
```

Examples

```
data(rsgcc)
x <- rnaseq[1:3,] #The first 3 genes in GEMatrix "rnaseq".
```

`gcc.corfinal` *get the final correlaiton and p-value of Gini method*

Description

Compare two correlations produced by GCC method for a gene pair, and choose one as the final output of GCC method.

Usage

```
gcc.corfinal(gcccor)
```

Arguments

`gcccor`	a list output by cor.pair function for GCC method.

Details

If the p-value is "NA", the correlation with absolute maximum value is selected; otherwise, the correlation with lower p-value is chosen.

Value

gcc.fcor	the final correlation of GCC.
gcc.fpavlue	the final pvalue of correlation.

Author(s)

Chuang Ma, Xiangfeng Wang

See Also

onegcc, cor.pair.

Examples

```
## Not run:
   data(rsgcc)
   x <- rnaseq[1:4,]

   #compute correlation between 1th and 4th genes
   #significance level of the computed correlation
   #is calcuated with 200 permutation tests.
   corpair <- cor.pair(c(1,4), GEMatrix = x, rowORcol = "row",
                       cormethod = "GCC", pernum = 200,
                       sigmethod = "two.sided")

   #get the final correlation and p-value of GCC method
   gcc.corfinal(corpair)

## End(Not run)
```

---

gcc.dist *compute distance matrix for hierarchical clustering*

Description

This function computes the distance between the rows of a data matrix with the specified distance method.

Usage

```
gcc.dist(x,
      cpus = 1,
      method = c("GCC", "PCC", "SCC", "KCC", "BiWt", "MI", "MINE", "ED"),
      distancemethod = c("Raw", "Abs", "Sqr"))
```

Arguments

x	a data matrix containing numeric variables, which is the same as the "GEMatrix" defined in the cor.matrix function.
cpus	the number of cpus used for computation.
method	a character string indicating the method to be used to calculate the associations.
distancemethod	a character string indicating the distance method to be used. Currently, three distance methods are available, include: "Raw" (1-cor)", "Abs" (1-|cor|), and "Sqr" (1-|cor|^2).

Value

A list with the following components:

dist	a data matrix containing the distances between different genes.
pairmatrix	a data matrix including the correlation between different genes.

Author(s)

Chuang Ma, Xiangfeng Wang

See Also

cor.matrix, gcc.hclust, gcc.tsheatmap.

Examples

```
data(rsgcc)
x <- rnaseq[1:10,]
gcc.dist(x, method = "GCC", distancemethod = "Raw", cpus = 1)
```

---

gcc.hclust *hierarchical cluster*

Description

Hierarchical cluster analysis of microarrany and RNA-Seq gene expression data with Gini correlation and four other correlation methods.

Usage

```
gcc.hclust(x,
      cpus = 1,
      method = c("GCC", "PCC", "SCC", "KCC", "BiWt", "MI", "MINE", "ED"),
      distancemethod = c("Raw", "Abs", "Sqr"),
      clustermethod = c("complete", "average", "median",
                        "centroid", "mcquitty", "single", "ward"))
```

Arguments

x	a data matrix containing numeric variables, which is the same as the GEMatrix defined in the cor.matrix function.
cpus	the number of cpus used for computation.
method	a character string indicating the method to be used to calculate the associations.
distancemethod	a character string specifying the distance method to be used. Currently, three distance methods are available, include: "Raw" (1-cor)", "Abs" (1-|cor|), and "Sqr" (1-|cor|^2).
clustermethod	the distance measure to be used. This must be one of "complete", "average", "median", "centroid", "mcquitty", "single", or "ward".

Details

This function generate the cluster tree with different distance measures for clustering analysis of microarray and RNA-Seq gene expression data by integrating the hclust function of stats package in R (http://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html). Similar to the hclust, the values output by gccdist can be directly used to plot cluster trees with plot function.

Value

A list with the following components:

hc	an object describes the tree information produced by the clustering process. This object is also a list with five components: "merge" is a numeric matrix with n-1 rows and 2 columns. n is the number of used individuals (e.g., genes). Row i describes the merging of clusters at step i of the clustering. "order" is a vector giving the order of individuals for tree cluster plotting. "height" is a vector with n-1 numeric values associated with the distance measure for the particular cluster method. "labels" are labels of the individuals being clustered. "method" is the distance measure used for cluster analysis. See details for the description in hclust function of stats package.
dist	a data matrix containing the distances between different genes.
pairmatrix	a data matrix including the correlation between different genes.

Author(s)

Chuang Ma, Xiangfeng Wang

Examples

```
## Not run:

   #obtain gene expression data of 10 genes.
   data(rsgcc)
   x <- rnaseq[1:10,]

   #hierarchical clustering analysis of these 10 genes with GCC method
```

```
   hc <- gcc.hclust(x, cpu=1, method = "GCC",
                    distancemethod = "Raw", clustermethod = "complete")

   #plot cluster tree
   plot(hc$hc)

## End(Not run)
```

---

gcc.heatmap *heat map*

Description

The heat map is a color imange representing the data in the a matrix. The dendrogram information are usually added to the left side and/or to the top for displaying the clustering information.

Usage

```
gcc.heatmap(x,
      cpus = 1,
      method = c("GCC", "PCC", "SCC", "KCC", "BiWt", "MI", "MINE", "ED"),
      distancemethod = c("Raw", "Abs", "Sqr"),
      clustermethod = c("complete", "average", "median",
                        "centroid", "mcquitty", "single", "ward"),

      #hcdata output by gcc.tsheatmap function
      rowhcdata = NULL,
      colhcdata = NULL,

      keynote = "FPKM",

      ## dendrogram control
      symm = FALSE,
      Rowv = TRUE,
      Colv = if (symm) "Rowv" else TRUE,
      dendrogram = c("both", "row", "column", "none"),

      ## data scaling
      scale = c("none", "row", "column"),
      na.rm = TRUE,
      revC = identical(Colv, "Rowv"),
      add.expr,

      #break points for binning values in x
      breaks = 16,
      quanbreaks = TRUE,
      symbreaks = min(x < 0, na.rm = TRUE) || scale != "none",
```

```
#colors
colrange = c("green", "black", "red"),

colsep,
rowsep,
sepcolor = "white",
sepwidth = c(0.05, 0.05),
cellnote,
notecex = 1,
notecol = "cyan",
na.color = par("bg"),
trace = c("none", "column", "row", "both"),
tracecol = "cyan",
hline = median(breaks),
vline = median(breaks),
linecol = tracecol,
margins = c(5, 5),
ColSideColors,
RowSideColors,
cexRow = 0.2 + 1/log10(dim(x)[1]),
cexCol = 0.2 + 1/log10(dim(x)[2]),
labRow = NULL,
labCol = NULL,

#color key
key = TRUE,
keysize = 0.65,
density.info = c("none", "histogram", "density"),
denscol = tracecol,
symkey = min(x < 0, na.rm = TRUE) || symbreaks,
densadj = 0.25,

#image information
main = NULL,
xlab = NULL,
ylab = NULL,
lmat = NULL,
lhei = NULL,
lwid = NULL,
...)
```

Arguments

x	a data matrix containing numeric variables. Example: rows may correspond to genes and columns to samples.
cpus	the number of cpus used for correlaiton calcluation. snowfall package in R

	needed to be installed in advance.
method	a character string that specifies a correlation method to be used for association calculation.
distancemethod	a character string specifying the distance method to be used. Currently, three distance methods are available, include: "Raw" (1-cor)", "Abs" (1-|cor|), and "Sqr" (1-|cor|^2).
clustermethod	the distance measure to be used. This must be one of "complete", "average", "median", "centroid", "mcquitty", "single", or "ward".
rowhcdata	the object of class hc generated from gcc.hclust for rows in x.
colhcdata	the object of class hc generated from gcc.hclust for columns in x.
keynote	a character string indicating the lable of color key.
symm	logical indicating if x should be treated as a symmetrical matrix.
Rowv	logical determining if the row dendrogram should be reordered.
Colv	logical determining if the columns dendrogram should be reordered.
dendrogram	a character string indicating whether to draw the "none", "row", "column", "both" dendrograms.
scale	a character string specifying if the data values would be centered and scaled by rows or by columns, or none.
na.rm	logical indicating whether the Nas should be eliminated.
revC	logical indicating if the column order should be reversed for plotting.
add.expr	expression that will be evaluated after the call to image. Can be used to add components to the plot.
breaks	(optional)Either a integer number specifying the break points to be used, or a numeric vector indicating the splitting points for binning x into colors.
quanbreaks	logical indicating if the splitting points for binning x in quantile scale. For instance, if quanbreaks is TRUE, the breaks would be quantile(unique(c(x)), probs = seq(0, 1, length = breaks), na.rm = TRUE).
symbreaks	Boolean indicating whether breaks should be made symmetric about 0. This option works when the quanbreaks is FALSE.
colrange	colors used for the image. It could be a function(i.e., heat.colors) or a vector of colors with at leat two elements (e.g., c("green", "black", "red")).
colsep	(optional) vectors of integers indicating which columns should be seperated from the preceding columans by a narrow space of color sepcor.
rowsep	(optional) vectors of integers indicating which rows should be seperated from the preceding rows by a narrow space of color sepcor.
sepcolor	(optional) color used to seperate rows or columns.
sepwidth	(optional) A numeric vector containing two elements giving the width (colsep) or height (rowsep) for the seperation of columns or rows.
cellnote	(optional)a matrix of character strings which will be placed within each color cell.
notecex	(optional)numeric scaling factor for cellnot itmes.

notecol	(optional)character string specifying the color of cellnote text. Default to "green".
na.color	color to be used for missing value (NA). Defaults to the plot background color.
trace	character string indicating a solid "trace" lined should be drawn across "rows", or "column", or "both" or "none".
tracecol	color for trace
hline	vector of values whithin cells where horizontal lines should be drawn with line col.
vline	vector of values whithin cells where vertical lines should be drawn with line col.
linecol	color for hline and vline.
margins	a numeric vector containing 2 elements specifying the margins for column and row names, respectively. See (par(mar=*)).
ColSideColors	(optional)character string of colors for annotating the columns of heat map.
RowSideColors	(optional)character string of colors for annotating the rows of heat map.
cexRow	cex.axis for the row lables.
cexCol	cex.axis for the column lables.
labRow	character strings indicating the lables of rows. Default to rownames(x)
labCol	character strings indicating the lables of columns. Default to colnames(x)
key	logical indicating whether the color key would be draw.
keysize	numeric value specifying the size of color key.
density.info	character string indicating whether to superimpose a "histogram", a "density" plot, or not plot("none") on the color-key.
denscol	character string giving the color for the density display specified by density.info, defaults to the same value as tracecol.
symkey	Boolean indicating whether the color key should be made symmetric about 0. Defaults to TRUE if the data includes negative values and to FALSE otherwise.
densadj	Numeric scaling value for tuning the kernel width when a density plot is drawn on the color key. Default to 0.25.
main	main title. defaults to none.
xlab	x-axis label. defaults to none.
ylab	y-axis label. defaults to none.
lmat	position matrix for visual layout. See details from the help page of heatmap.2.
lhei	column height for visual layout. See details from the help page of heatmap.2
lwid	column width for visual layout. See details from the help page of heatmap.2
...	additonal arguments passed on to image.

Details

This function plots the heat map of microarray and RNA-Seq gene expression data by modifying the scripts of heatmap.2 in R. The main modifications include: (1) designing several distance measures derived from Gini correlation and other correlation methods; (2) providing the option of quanbreaks for RNA-Seq data.

Value

A list with the following components:

retval	a list with components of "rowInd" (row index of heat map from x), "colInd" (column index of heat map from x), "call" (the match call), "carpet" (reordered and scaled 'x' values used generate the main 'carpet'), "rowDendrogram" (row dendrogram), "colDendrogram" (column dendrogram), "breaks" (break points for binning x), "col" (colors used), and "colorTable" (a data frame providing the lower and upper bound and color for each bin).
hcr	the values returned from gcc.hclust function for clustering individuals (e.g., genes) in row direction
hcc	the values returned from gcc.hclust function for clustering individuals (e.g., genes) in column direction

Note

This function clusters microarray and RNA-Seq gene expression data and plot heatmap by refining heatmap.2 function in gplots package. Therefore, most parameters and output values are defined similarly as those in heatmap.2.

Author(s)

Chuang Ma, Xiangfeng Wang

See Also

gcc.dist, cor.matrix, gcc.hclust, gcc.tsheatmap.

Examples

```
## Not run:
   data(rsgcc)
   x <- rnaseq[1:50,]
   ghm <- gcc.heatmap(x, cpus = 1, method = "GCC",
               distancemethod = "Raw", clustermethod = "complete", labRow = "")

## End(Not run)
```

---

gcc.tsheatmap *correlaiton and clustering analysis of tissue-specific genes*

Description

This function performs the correlaiton and clustering analysis of tissue-specific genes with expression data generated from microarray and RNA-Seq experiments.

Usage

```
gcc.tsheatmap(x,

      cpus = 1,

      ## correlation method
      method = c("GCC", "PCC", "SCC", "KCC", "BiWt", "MI", "MINE", "ED"),

      distancemethod = c("Raw", "Abs", "Sqr"),

      #cluster method
      clustermethod = c("complete", "average", "median",
                        "centroid", "mcquitty", "single", "ward"),

      #hcdata by output gcc.tsheatmap
      rowhcdata = NULL,
      colhcdata = NULL,


      keynote = "FPKM",

      ## dendrogram control
      symm = FALSE,

      ## data scaling
      scale = c("none","row", "column"),
      na.rm=TRUE,

      ## image plot
      revC = identical(Colv, "Rowv"),
      add.expr,

      ## mapping data to colors
      breaks,
      symbreaks=min(x < 0, na.rm=TRUE) || scale!="none",

      ## colors
      colrange = c("yellow", "red"),

      tissuecol= "heat.colors",

      ## block sepration
      colsep = 0.15,
      rowsep,
      sepcolor="white",
      sepwidth=c(0.05,0.05),

      ## level trace
```

```
trace=c("none","column","row","both"),
tracecol="cyan",
hline=median(breaks),
vline=median(breaks),
linecol=tracecol,

## plot margins
margins = c(5, 5),

## plot labels
main = NULL,
xlab = NULL,
ylab = NULL,

## plot layout
lmat = NULL,
lhei = NULL,
lwid = NULL,

## extras
...)
```

Arguments

x	a data matrix containing numeric variables. Example: rows may correspond to genes and columns to samples.
cpus	the number of cpus used for correlaiton calcluation. snowfall package in R needed to be installed in advance.
method	a character string that specifies a correlation method to be used for association calculation.
distancemethod	a character string specifying the distance method to be used. Currently, three distance methods are available, include: "Raw" (1-cor)", "Abs" (1-|cor|), and "Sqr" (1-|cor|^2).
clustermethod	the distance measure to be used. This must be one of "complete", "average", "median", "centroid", "mcquitty", "single", or "ward".
rowhcdata	the object of class hc generated from gcc.hclust for rows in x.
colhcdata	the object of class hc generated from gcc.hclust for columns in x.
keynote	a character string indicating the lable of color key.
symm	logical indicating if x should be treated as a symmetrical matrix.
scale	a character string specifying if the data values would be centered and scaled by rows or by columns, or none.
na.rm	logical indicating whether the Nas should be eliminated.
revC	logical indicating if the column order should be reversed for plotting.
add.expr	expression that will be evaluated after the call to image.

breaks	(optional)Either a integer number specifying the break points to be used, or a numeric vector indicating the splitting points for binning x into colors.
symbreaks	Boolean indicating whether breaks should be made symmetric about 0. This option works if the quanbreaks is FALSE.
colrange	colors used for the image. It could be a function(i.e., heat.colors) or a vector of colors with at leat two elements (e.g., c("green", "black", "red")).
tissuecol	colors for tissues. tissuecol could be a function(i.e., heat.colors) or a vector of colors for different tissues.
colsep	(optional) vectors of integers indicating which columns should be seperated from the preceding columans by a narrow space of color sepcor.
rowsep	(optional) vectors of integers indicating which rows should be seperated from the preceding rows by a narrow space of color sepcor.
sepcolor	(optional) color used to seperate rows or columns.
sepwidth	(optional) A numeric vector containing two elements giving the width (colsep) or height (rowsep) for the seperation of columns or rows.
trace	character string indicating a solid "trace" lined should be drawn across "rows", or "column", or "both" or "none".
tracecol	color for trace
hline	vector of values whithin cells where horizontal lines should be drawn with line col.
vline	vector of values whithin cells where vertical lines should be drawn with line col.
linecol	color for hline and vline.
margins	a numeric vector containing 2 elements specifying the margins for column and row names, respectively. See (par(mar=*)).
main	main title. defaults to none.
xlab	x-axis label. defaults to none.
ylab	y-axis label. defaults to none.
lmat	position matrix for visual layout.
lhei	column height for visual layout.
lwid	column width for visual layout. For instance, lwid = c(0.5, 0.05, 0.01, 0.5, 0.01, 0.05, 0.5)
...	additonal arguments passed on to image.

Value

A list with the following components:

retval	a list with components of "rowInd" (row index of heat map from x), "colInd" (column index of heat map from x), "call" (the match call), "carpet" (reordered and scaled 'x' values used generate the main 'carpet'), "rowDendrogram" (row dendrogram), "colDendrogram" (column dendrogram), "breaks" (break points for binning x), "col" (colors used), and "colorTable" (a data frame providing the lower and upper bound and color for each bin).

hcr	the values returned from gcc.hclust function for clustering individuals (e.g., genes) in row direction
hcc	the values returned from gcc.hclust function for clustering individuals (e.g., genes) in column direction

Author(s)

Chuang Ma, Xiangfeng Wang

See Also

gcc.dist, cor.matrix, gcc.hclust, gcc.tsheatmap.

Examples

```
## Not run:
   data(rsgcc)

   #get expression matrix of tissue-specific genes
   tsRes <- getsgene(rnaseq, tsThreshold = 0.75, MeanOrMax = "Max", Fraction = TRUE)

   #heat map of tissue-specific genes
   thm <- gcc.tsheatmap(tsRes$tsgene, cpus = 1, method = "GCC",
                    distancemethod = "Raw", clustermethod = "complete")

## End(Not run)
```

---

getsgene *identify tissue(or condtion)-specific genes*

Description

This function identifies tissue(or condition)- specific genes by considering the difference between the mean expression value of one tissue and the max expression value of other tissue.

Usage

```
getsgene(x, Log = FALSE, Base = 2, AddOne = FALSE,
         tsThreshold = 0.95, MeanOrMax = "Mean", Fraction = TRUE)
```

Arguments

x	a numeric matrix containing gene expression value. The column labels are samples names. For two samples from the same tissue T, their names should be assigned as T.1 and T.2, respectively.
Log	logical indicating whether the gene expression value would be log-transformed.
Base	a numeric value specifying the base of logarithm.

AddOne	logical indicating if add one for avoding the problem of log-zero.
tsThreshold	a numeric value giving the threshold of tissue specificity score. The tissue specificity score is 1, if the gene is only expressed in one tissue. Otherwise, the tissue specificity socre will be smaller than 1.
MeanOrMax	character "Mean" or "Max" indicate the mean or maximal expression value will be calculated for the tissue of interest.
Fraction	logical indicating whether the gene expression value would be scaled across tissues.

Details

The tissue specificity score is calculated with the formula 1-min(R(1), R(2), ..., R(i),..., R(n)), where R(i) = M(i)/E(i), E(i) is the mean or maximal expression value of tissue i, and M(i) is the maximal expression value of other tissues. If the tissue specificity score higher than tsThreshold, then the gene is considered as tissue specifically expressed.

If Fraction is TRUE, the expression values of a gene is scaled accorss the tissues with the formula e(i)/(e(1)+e(2)+...+e(n)). e(i) is the expression value of the consider gene in ith sample.

Value

A list with following components:

csGenes	a data matrix containing expression vlaues of tissue specific genes.
csScoreMat	a data matrix with three columns containg the gene index information from x, tissue specificity score and the tissue information with the tissue specificity score.

Author(s)

Chuang Ma, Xiangfeng Wang.

References

[1] Chuang Ma, Xiangfeng Wang. Machine learning-based differential network analysis of transcriptomic data: a case study of stress-responsive gene expression in Arabidopsis thaliana. 2013 (Submitted).

Examples

```
## Not run:
   data(rsgcc)
   tsRes <- getsgene(rnaseq, tsThreshold = 0.75, MeanOrMax = "Mean", Fraction = TRUE)

## End(Not run)
```

onegcc *compute one Gini correlation coefficient*

Description

onegcc calcluates one Gini correlation coefficient with rank information of the first variable.

Usage

```
onegcc(x, y)
```

Arguments

x	a numeric vector.
y	a numeric vector with the same length of x.

Details

This is a generic function cacluating correlation with rank information of the first variable and the actual value information of the second variable.

Value

Gini correlation coefficient (a numeric value ranged from -1.0 to 1.0).

Author(s)

Chuang Ma, Xiangfeng Wang

Examples

```
data(rsgcc)
x <- rnaseq[1:10,]     #Just use a small subset of RNA-Seq data
onegcc(x[1,], x[2,])   # generate one correlaiton for one gene pair
onegcc(x[2,], x[1,])   # generate the other correlaiton for the same gene pair
```

rsgcc.gui *graphical user interface (GUI) of rsgcc package*

Description

This function provides a graphical user interface (GUI) to perform the correlation and clustering analysis via a series of mouse actions without command-line based R programmming. The output of clustering informaiton in "CDT" format can be further visulized and analyzed by TreeView program.

Usage

```
rsgcc.gui(margins = c(1, 1), labRow = "", labCol = "",
          lwid = c(0.5, 0.05, 0.01, 0.5, 0.01, 0.05, 0.5),
          keynote = "FPKM")
```

Arguments

margins	a numeric vector containing 2 elements specifying the margins for heat map. See (par(mar=*)).
labRow	character strings indicating the lables of rows. Default to rownames(x).
labCol	character strings indicating the lables of columns. Default to colnames(x).
lwid	column width for visual layout.
keynote	a character string indicating the lable of color key.

Details

For heat map of ts-genes, rsgcc will run the gcc.tsheatmap function. the lwid could be c(0.5, 0.05, 0.01, 0.5, 0.01, 0.05, 0.5). The 2nd, 4th and 6th elements are column widths of color tissue bar, heat map and color key bar. The 3rd and 5th are the widths of separation for these three figures. The The first and last elements are the widths of "blank region" for displaying the labels of tissue and key bar. If the option "Find and cluster ts-genes" is not selected, rsgcc will call the gcc.heatmap for clustering analysis. In this case, the lwid should be a numeric vector with two elements(e.g.,c(0.65,4)). Here 0.65 is the width of column for row dendrogram. 4 is the width of heat map.

Note

rsgcc.gui is built upon gWidgets package. Make sure "gWidgets" and "gWidgetsRGtk2" package is properly installed.

The following is a guide of using the rsgcc GUI.

Step 1: Select a gene expression data file and load expression data with read.table or read.csv, which is decided by the program itself according to the file suffix (".txt", ".csv", or nothing). Each row of the table is one gene, and each column is the expression data of one sample. The column names are sample IDs indiciating the tissue informaiton (i.e., "T1.1", "T1.2", "T2"). After the data is already loaded, you can click the option ("Display loaded data") to display the gene expression datain a

new window. If the tissue-specific genes are interested for the clustering analysis, please select the option ("Find and cluster ts-genes") and specify the threshold of tissue-specificity score (Default 0.95. The threshold should be smaller than 1.0; the <ENTER> MUST be pressed to confirm the change).

Step 2: Selecte a correlation method. Default: Gini correlation.

Step 3: Specify a distance measure. Default: raw correlation (1-coef).

Step 4: Choose a cluster method. In current version, rsgcc includes seven cluster methods. More information about these cluster methods can be found at the help page of hclust function.

Step 5: Set a integer for the number of CPUs to be used. The snowfall package in R is needed for the paralleled computing to speed up the calculation of correlation coefficients. After you change the number, the <ENTER> MUST be pressed to confirm the change.

Step 6: Press the button "Start to run" to perform the correlation and clustering analysis of gene expression data. A heat map will be visualized at the right region of interface if the task is finished.

Step 7: Three bars can be slided to adjust colors in heat map.

Step 8: After the appropriate colors have already been determined, you can "save correlation and cluster data". The correlations will be output to a file with three columns (gene1, gene2, correlation). The cluster information will be output into three files (the suffix are ".atr", ".gtr" and ".cdt") for visualizing and analyzing by Treeview program. A pdf file will also be generated for heat map. All these files in the same directory of gene expression data file.

Author(s)

Chuang Ma, Xiangfeng Wang

See Also

gcc.tsheatmap, gcc.heatmap.

Examples

```
## Not run:
   library("gWidgetsRGtk2")

   library(rsgcc)

   ## the GUI of rsgcc will show up after the GUI toolkit "gWidgetsRGtk2" is selected.
   rsgcc.gui()

## End(Not run)
```

uniqueTissues *get tissue information*

Description

This function reads the sample names of genes and get unique tissue information for further tissue-specific genes finding and clustering.

Usage

```
uniqueTissues(x)
```

Arguments

x	a numeric matrix containing gene expression value. The column labels are samples names. For two samples from the same tissue T, their names should be assigned as T.1 and T.2, respectively.

Value

A data matrix in which the elements is 0 (the sample not from the tissue) or 1 (the sample from the tissue)

Author(s)

Chuang Ma, Xiangfeng Wang

See Also

getsgene, gcc.tsheatmap.

Examples

```
## Not run:
   data(rsgcc)
   x <- rnaseq
   uniqueTissues(x)

## End(Not run)
```

Index