Package 'cbaf'

October 15, 2018

biocViews Technology

Title Multiple automated functions for cbioportal.org

Version 1.2.0

Description This package contains functions that allow analysing and comparing various gene groups from different cancers/cancer subgroups easily. So far, it is compatible with RNA-seq, microRNA-seq, microarray and methylation datasets that are stored on cbioportal.org.

License Artistic-2.0

Encoding UTF-8

LazyData true

Imports BiocFileCache, RColorBrewer, cgdsr, genefilter, gplots, grDevices, stats, utils, xlsx

NeedsCompilation no

RoxygenNote 6.0.1

Suggests knitr, rmarkdown, BiocStyle

VignetteBuilder knitr

Collate 'cbaf-obtainMultipleStudies.R' 'cbaf-obtainOneStudy.R' 'cbaf-automatedStatistics.R' 'cbaf-availableData.R' 'cbaf-cleanDatabase.R' 'cbaf-heatmapOutput.R' 'cbaf-processMultipleStudies.R' 'cbaf-processOneStudy.R' 'cbaf-xlsxOutput.R'

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

git_branch RELEASE_3_7

git_last_commit b172715

git_last_commit_date 2018-04-30

Date/Publication 2018-10-15

Author Arman Shahrisa [aut, cre, cph], Maryam Tahmasebi Birgani [aut]

Maintainer Arman Shahrisa <shahrisa.arman@hotmail.com>

R topics documented:

`automatedStatistics` *Perform the requested statistics for various studies / subgroups of a study.*

Description

This function calculates frequency percentage, frequency ratio, mean value and median value of samples greather than specific cutoff in the selected study / subgroups of the study. Furthermore, it can looks for the five genes that contain the highest values in each study / study subgroup. It uses the data generated by obtainOneStudy()/obtainMultipleStudies() function.

Usage

```
automatedStatistics(submissionName, obtainedDataType =
  "multiple studies", calculate = c("frequencyPercentage", "frequencyRatio",
  "meanValue"), topGenes = TRUE, cutoff=NULL, round=TRUE)
```

Arguments

`submissionName` a character string containing name of interest. It is used for naming the process.

`obtainedDataType`
: a character string that specifies the type of input data produced by the previous function. Two options are availabe: `"single study"` for obtainOneStudy() and `"multiple studies"` for obtainMultipleStudies(). The function uses obtainedDataType and submissionName to construct the name of the BiocFileCach object and then finds the appropriate data inside it. Default value is multiple studies‘.

`calculate`
: a character vector that containes the statistical procedures users prefer the function to compute. The complete results can be obtained by `c("frequencyPercentage", "frequencyR`
This will tell the function to compute the following: `"frequencyPercentage"`, which is the percentge of samples having the value greather than specific cutoff divided by the total sample size for every study / study subgroup; `"frequency ratio"`, which shows the number of selected samples divided by the total number of samples that give the frequency percentage for every study / study subgroup. It shows the selected and total sample sizes.; `"Mean Value"`, that contains mean value of selected samples for each study; `"Median Value"`, which shows the median value of selected samples for each study. The default input is `calculate = c("frequencyPer`

topGenes — a logical value that, if set as TRUE, causes the function to create three data.frame that contain the five top genes for each cancer. To get all the three data.frames, "frequencyPercentage", "meanValue" and "MedianValue" must have been included for `calculate`.

cutoff — a number used to limit samples to those that are greather than this number (cutoff). The default value for methylation data is `0.6` while gene expression studies use default value of 2. For methylation studies, it is `observed/expected ratio`, for the rest, it is `"z-score"`. To change the cutoff to any desired number, change the option to `cutoff = desiredNumber` in which desiredNumber is the number of interest.

round — a logical value that, if set to be `TRUE`, will force the function to round all the calculated values to two decimal places. The default value is `TRUE`.

Details

Package:	cbaf
Type:	Package
Version:	1.1.4
Date:	2017-11-23
License:	Artistic-2.0

Value

A new section in the BiocFileCache object that was created by one of the obtainOneStudy() or obtainMultipleStudies() functions. It contains a list that contains some or all of the following statistical measurements for every gene group, based on what user has chosen: `Frequency.Percentage` , `Top.Genes.of.Frequency.Percentage`, `Frequency.Ratio`, `Mean.Value`, `Top.Genes.of.Mean.Value`, `Median`, `Top.Genes.of.Median`.

Author(s)

Arman Shahrisa, <shahrisa.arman@hotmail.com> [maintainer, copyright holder]

Maryam Tahmasebi Birgani, <tahmasebi-ma@ajums.ac.ir>

Examples

```
genes <- list(K.demethylases = c("KDM1A", "KDM1B", "KDM2A", "KDM2B", "KDM3A",
 "KDM3B", "JMJD1C", "KDM4A"), K.methyltransferases = c("SUV39H1", "SUV39H2",
 "EHMT1", "EHMT2", "SETDB1", "SETDB2", "KMT2A", "KMT2A"))

obtainOneStudy(genes, "test", "Breast Invasive Carcinoma (TCGA, Cell 2015)",
"RNA-Seq", desiredCaseList = c(3,4))

automatedStatistics("test", obtainedDataType = "single study", calculate =
c("frequencyPercentage", "frequencyRatio"))
```

`availableData` *Check which Data types are available for each cancer study.*

Description

This function checks all the cancer studies that are registered in 'cbioportal.org' to examine whether or not they contain RNA-Seq, microRNA-Seq, microarray(mRNA), microarray(miRNA) and methylation data.

Usage

```
availableData(excelFileName)
```

Arguments

`excelFileName`	a character string that is required to name the output and, if requested, excel file.

Details

Package:	cbaf
Type:	Package
Version:	1.1.4
Date:	2017-11-23
License:	Artistic-2.0

Value

An excel file that contains all the cancer studies versus available data types

Author(s)

Arman Shahrisa, <shahrisa.arman@hotmail.com> [maintainer, copyright holder]

Maryam Tahmasebi Birgani, <tahmasebi-ma@ajums.ac.ir>

`cleanDatabase` *Clean the created database(s)*

Description

This function removes the created databases in the cbaf package directory. This helps users to obtain the fresh data from cbioportal.org.

Usage

```
cleanDatabase(databaseNames = NULL)
```

Arguments

databaseNames	a character vector that contains name of databases that will be removed. The default value in null.

Details

Package:	cbaf
Type:	Package
Version:	1.1.4
Date:	2017-11-23
License:	Artistic-2.0

Value

prints the number of removed databases.

Author(s)

Arman Shahrisa, <shahrisa.arman@hotmail.com> [maintainer, copyright holder]

Maryam Tahmasebi Birgani, <tahmasebi-ma@ajums.ac.ir>

Examples

```
cleanDatabase(databaseNames = "Whole")
```

---

heatmapOutput *Generate heatmaps for various studies/subgroups of a study.*

Description

This function can prepare heatmap for 'frequency percentage', 'mean value' and 'median value' data provided by automatedStatistics() function.

Usage

```
heatmapOutput(submissionName, shortenStudyNames = TRUE,
  geneLimit = FALSE, rankingMethod = "variation", heatmapFileFormat = "TIFF",
  resolution = 600, RowCex = "auto", ColCex = "auto",
  heatmapMargines = "auto", rowLabelsAngle = 0, columnLabelsAngle = 45,
  heatmapColor = "RdBu", reverseColor = TRUE, transposedHeatmap = FALSE,
  simplifyBy = FALSE, genesToDrop = FALSE)
```

Arguments

submissionName	a character string containing name of interest. It is used for naming the process.
shortenStudyNames	a logical vector. If the value is set as TRUE, function will try to remove the last part of the cancer names aiming to shorten them. The removed segment usually contains the name of scientific group that has conducted the experiment.
geneLimit	if large number of genes exist in at least one gene group, this option can be used to limit the number of genes that are shown on heatmap. For instance, geneLimit=50 will limit the heatmap to 50 genes that show the most variation across multiple study / study subgroups. The default value is FALSE.
rankingMethod	a character value that determines how genes will be ranked prior to drawing heatmap. "variation" orders the genes based on unique values in one or few cancer studies while "highValue" ranks the genes when they cotain high values in multiple / many cancer studies. This option is useful when number of genes are too much so that user has to limit the number of genes on heatmap by geneLimit.
heatmapFileFormat	This option enables the user to select the desired image file format of the heatmaps. The default value is "TIFF". Other suppoeted formats include "PNG", "BMP", and "JPG".
resolution	a number. This option can be used to adjust the resolution of the output heatmaps as 'dot per inch'. The defalut value is 600.
RowCex	a number that specifies letter size in heatmap row names, which ranges from 0 to 2. If RowCex = "auto", the function will automatically determine the best RowCex.
ColCex	a number that specifies letter size in heatmap column names, which ranges from 0 to 2. If ColCex = "auto", the function will automatically determine the best ColCex.
heatmapMargines	a numeric vector that is used to set heatmap margins. If heatmapMargines = "auto", the function will automatically determine the best possible margines. Otherwise, enter the desired margine as e.g. c(10,10.)
rowLabelsAngle	a number that determines the angle with which the gene names are shown in heatmaps. The default value is 0 degree.
columnLabelsAngle	a number that determines the angle with which the studies/study subgroups names are shown in heatmaps. The default value is 45 degree.
heatmapColor	a character string that defines heatmap color. The default value is 'RdBu'. 'RdGr' is also a popular color in genomic studies. To see the rest of colors, please type library(RColorBrewer) and then display.brewer.all().
reverseColor	a logical value that reverses the color gradiant for heatmap(s).
transposedHeatmap	a logical value that transposes heatmap rows to columns and vice versa.
simplifyBy	a number that tells the function to change the values smaller than that to zero. The purpose behind this option is to facilitate recognizing candidate genes. Therefore, it is not suited for publications. It has the same unit as cutoff.
genesToDrop	a character vector. Gene names within this vector will be omitted from heatmap.The default value is FALSE.

Details

Package:	cbaf
Type:	Package
Version:	1.1.4
Date:	2017-11-23
License:	Artistic-2.0

Value

Based on preference, three heatmaps for "Frequency.Percentage" , "Mean.Value" and "Median.value" can be generated. If more than one group of genes are entered, output for each group will be strored in a separate sub-directory.

Author(s)

Arman Shahrisa, <shahrisa.arman@hotmail.com> [maintainer, copyright holder]

Maryam Tahmasebi Birgani, <tahmasebi-ma@ajums.ac.ir>

Examples

```
genes <- list(K.demethylases = c("KDM1A", "KDM1B", "KDM2A", "KDM2B", "KDM3A",
 "KDM3B", "JMJD1C", "KDM4A"), K.methyltransferases = c("SUV39H1", "SUV39H2",
 "EHMT1", "EHMT2", "SETDB1", "SETDB2", "KMT2A", "KMT2A"))

obtainOneStudy(genes, "test", "Breast Invasive Carcinoma (TCGA, Cell 2015)",
"RNA-Seq", desiredCaseList = c(3,4))

automatedStatistics("test", obtainedDataType = "single study", calculate =
c("frequencyPercentage", "frequencyRatio"))

heatmapOutput(submissionName = "test")
```

---

obtainMultipleStudies *Obtain the requested data for various cancer studies.*

Description

This function Obtains the requested data for the given genes across multiple cancer studies. It can check whether or not all genes are included in cancer studies and, if not, looks for the alternative gene names.

Usage

```
obtainMultipleStudies(genesList, submissionName, studiesNames,
  desiredTechnique, cancerCode = FALSE, validateGenes = TRUE)
```

Arguments

genesList	a list that contains at least one gene group
submissionName	a character string containing name of interest. It is used for naming the process.
studiesNames	a character vector or a matrix that containes desired cancer names. The character vector containes standard names of cancer studies that can be found on cbioportal.org, such as "Acute Myeloid Leukemia (TCGA, NEJM 2013)". Alternatively, a matrix can used if users prefer user-defined cancer names. In this case, the first column of matrix comprises the standard cancer names while the second column must contain the desired cancer names.
desiredTechnique	a character string that is one of the following techniques: "RNA-Seq", "microRNA-Seq", "microarray.mRNA" , "microarray.microRNA" or "methylation".
cancerCode	a logical value that tells the function to use cbioportal abbreviated cancer names instead of complete cancer names, if set to be TRUE. For example, "laml_tcga_pub" is the abbreviated name for "Acute Myeloid Leukemia (TCGA, NEJM 2013)".
validateGenes	a logical value that, if set to be TRUE, function will check each cancer study to find whether or not each gene has a record. If a cancer study doesn't have a record for specific gene, it checks for alternative gene names that cbioportal might use instead of the given gene name.

Details

Package:	cbaf
Type:	Package
Version:	1.1.4
Date:	2017-11-23
License:	Artistic-2.0

Value

a BiocFileCach object that contains the obtained data without further processing. Name of the object is combination of bfc_ and submissionName. Inside it, there is a section for the obtained data, which is stored as a list. At first level, this list is subdivided into diferent groups based on the list of genes that user has given the function, then each gene group itself contains one matrix for every cancer study. Additonally, if validateGenes = TRUE, another section that contains gene validation results will be created in the BiocFileCach object.

Author(s)

Arman Shahrisa, <shahrisa.arman@hotmail.com> [maintainer, copyright holder]

Maryam Tahmasebi Birgani, <tahmasebi-ma@ajums.ac.ir>

Examples

```
genes <- list(K.demethylases = c("KDM1A", "KDM1B", "KDM2A", "KDM2B", "KDM3A",
 "KDM3B", "JMJD1C", "KDM4A"), K.methyltransferases = c("SUV39H1", "SUV39H2",
 "EHMT1", "EHMT2", "SETDB1", "SETDB2", "KMT2A", "KMT2A"))
```

```
studies <- c("Acute Myeloid Leukemia (TCGA, Provisional)",
"Adrenocortical Carcinoma (TCGA, Provisional)",
"Bladder Urothelial Carcinoma (TCGA, Provisional)",
"Brain Lower Grade Glioma (TCGA, Provisional)",
"Breast Invasive Carcinoma (TCGA, Provisional)")

obtainMultipleStudies(genes, "test2", studies, "RNA-Seq")
```

`obtainOneStudy` *Obtain the requested data for various subgroups of a cancer study.*

Description

This function Obtains the requested data for the given genes across multiple subgroups of a cancer. It can check whether or not all genes are included in subgroups of a cancer study and, if not, looks for the alternative gene names.

Usage

```
obtainOneStudy(genesList, submissionName, studyName, desiredTechnique,
  desiredCaseList = FALSE, validateGenes = TRUE)
```

Arguments

`genesList`	a list that contains at least one gene group
`submissionName`	a character string containing name of interest. It is used for naming the process.
`studyName`	a character string showing the desired cancer name. It is an standard cancer study name that can be found on cbioportal.org, such as `"Acute Myeloid Leukemia (TCGA, NEJM 20`
`desiredTechnique`	a character string that is one of the following techniques: `"RNA-Seq"`, `"microRNA-Seq"`, `"microarray.mRNA"` , `"microarray.microRNA"` or `"methylation"`.
`desiredCaseList`	a numeric vector that contains the index of desired cancer subgroups, assuming the user knows index of desired subgroups. If not, desiredCaseList is set to `"none"`, function will show the available subgroups and ask the user to enter the desired ones during the process. The default value is `"none"`.
`validateGenes`	a logical value that, if set to be 'TRUE', causes the function to check each cancer study to find whether or not each gene has a record. If a cancer doesn't have a record for specific gene, function looks for alternative gene names that cbioportal might use instead of the given gene name.

Details

Package:	cbaf
Type:	Package
Version:	1.1.4
Date:	2017-11-23
License:	Artistic-2.0

Value

a BiocFileCach object that contains the obtained data without further processing. Name of the object is combination of 'bfc_' and submissionName. Inside it, there is a section for the obtained data, which is stored as a list. At first level, this list is subdivided into diferent groups based on the list of genes that user has given the function, then each gene group itself contains one matrix for every study subgroup. Additonally, if validateGenes = TRUE, another section that contains gene validation results will be created in the BiocFileCach object.

Author(s)

Arman Shahrisa, <shahrisa.arman@hotmail.com> [maintainer, copyright holder]

Maryam Tahmasebi Birgani, <tahmasebi-ma@ajums.ac.ir>

Examples

```
genes <- list(K.demethylases = c("KDM1A", "KDM1B", "KDM2A", "KDM2B", "KDM3A",
 "KDM3B", "JMJD1C", "KDM4A"), K.methyltransferases = c("SUV39H1", "SUV39H2",
 "EHMT1", "EHMT2", "SETDB1", "SETDB2", "KMT2A", "KMT2A"))

obtainOneStudy(genes, "test", "Breast Invasive Carcinoma (TCGA, Cell 2015)",
"RNA-Seq", desiredCaseList = c(2,3,4,5))
```

---

processMultipleStudies

Check Expression/methylation Profile for various cancer studies.

Description

This function Obtains the requested data for the given genes across multiple cancer studie. It can check whether or not all genes are included in cancer studies and and, if not, looks for the alternative gene names. Then it calculates frequency percentage, frequency ratio, mean value and median value of samples greather than specific value in the selected cancer studies. Furthermore, it looks for the five genes that comprise the highest values in each cancer study.

Usage

```
processMultipleStudies(genesList, submissionName, studiesNames,
  desiredTechnique, cancerCode = FALSE, validateGenes = TRUE, calculate =
  c("frequencyPercentage", "frequencyRatio", "meanValue"), cutoff=NULL,
  round=TRUE, topGenes = TRUE, shortenStudyNames = TRUE, geneLimit = FALSE,
  rankingMethod = "variation", heatmapFileFormat = "TIFF", resolution = 600,
  RowCex = "auto", ColCex = "auto", heatmapMargines = "auto",
  rowLabelsAngle = 0, columnLabelsAngle = 45, heatmapColor = "RdBu",
  reverseColor = TRUE, transposedHeatmap = FALSE, simplifyBy = FALSE,
  genesToDrop = FALSE)
```

Arguments

genesList	a list that contains at least one gene group
submissionName	a character string containing name of interest. It is used for naming the process.
studiesNames	a character vector or a matrix that containes desired cancer names. The character vector containes standard names of cancer studies that can be found on cbioportal.org, such as "Acute Myeloid Leukemia (TCGA, NEJM 2013)". Alternatively, a matrix can be used if users prefer user-defined cancer names. In this case, the first column of matrix comprises the standard cancer names while the second column must contain the desired cancer names.
desiredTechnique	a character string that is one of the following techniques: "RNA-Seq", "microRNA-Seq", "microarray.mRNA" , "microarray.microRNA" or "methylation".
cancerCode	a logical value that tells the function to use cbioportal abbreviated cancer names instead of complete cancer names, if set to be "TRUE". For example, "laml_tcga_pub" is the abbreviated name for "Acute Myeloid Leukemia (TCGA, NEJM 2013)".
validateGenes	a logical value that, if set to be TRUE, causes the function to check each cancer study to find whether or not each gene has a record. If a cancer doesn't have a record for specific gene, function looks for alternative gene names that cbioportal might use instead of the given gene name.
calculate	a character vector that containes the statistical procedures users prefer the function to compute. The complete results can be obtained by c("frequencyPercentage", "frequencyR This will tell the function to compute the following: "frequencyPercentage", which is the percentge of samples having the value greather than specific cutoff divided by the total sample size for every study / study subgroup; "frequency ratio", which shows the number of selected samples divided by the total number of samples that give the frequency percentage for every study / study subgroup. It shows the selected and total sample sizes.; "Mean Value", that contains mean value of selected samples for each study; "Median Value", which shows the median value of selected samples for each study. The default input is calculate = c("frequencyPer
cutoff	a number used to limit samples to those that are greather than this number (cutoff). The default value for methylation data is 0.6 while gene expression studies use default value of 2. For methylation studies, it is observed/expected ratio, for the rest, it is "z-score". To change the cutoff to any desired number, change the option to cutoff = desiredNumber in which desiredNumber is the number of interest.
round	a logical value that, if set to be TRUE, will force the function to round all the calculated values to two decimal places. The default value is TRUE.
topGenes	a logical value that, if set as TRUE, causes the function to create three dataframes that contain the five top genes for each cancer. To get all the three data.frames, "frequencyPercentage", "meanValue" and "medianValue" must have been included for calculate.
shortenStudyNames	a logical vector. If the value is set as TRUE, function will try to remove the last part of the cancer names aiming to shorten them. The removed segment usually contains the name of scientific group that has conducted the experiment.
geneLimit	if large number of genes exist in at least one gene group, this option can be used to limit the number of genes that are shown on heatmap. For instance, geneLimit=50 will limit the heatmap to 50 genes showing the most variation across multiple study / study subgroups. The default value is FALSE.

`rankingMethod` a character value that determines how genes will be ranked prior to drawing heatmap. `"variation"` orders the genes based on unique values in one or few cancer studies while `"highValue"` ranks the genes when they cotain high values in multiple / many cancer studies. This option is useful when number of genes are too much so that user has to limit the number of genes on heatmap by `geneLimit`.

`heatmapFileFormat`
This option enables the user to select the desired image file format of the heatmaps. The default value is `"TIFF"`. Other suppoeted formats include `"BMP"`, `"JPG"`, and `"PNG"`.

`resolution` a number. This option can be used to adjust the resolution of the output heatmaps as 'dot per inch'. The defalut value is 600.

`RowCex` a number that specifies letter size in heatmap row names, which ranges from 0 to 2. If `RowCex = "auto"`, the function will automatically determine the best RowCex.

`ColCex` a number that specifies letter size in heatmap column names, which ranges from 0 to 2. If `ColCex = "auto"`, the function will automatically determine the best ColCex.

`heatmapMargines`
a numeric vector that is used to set heatmap margins. If `heatmapMargines = "auto"`, the function will automatically determine the best possible margines. Otherwise, enter the desired margine as e.g. c(10,10.)

`rowLabelsAngle` a number that determines the angle with which the gene names are shown in heatmaps. The default value is 0 degree.

`columnLabelsAngle`
a number that determines the angle with which the studies/study subgroups names are shown in heatmaps. The default value is 45 degree.

`heatmapColor` a character string that defines heatmap color. The default value is `'RdBu'`. `'RdGr'` is also a popular color in genomic studies. To see the rest of colors, please type `library(RColorBrewer)` and then `display.brewer.all()`.

`reverseColor` a logical value that reverses the color gradiant for heatmap(s).

`transposedHeatmap`
a logical value that transposes heatmap rows to columns and vice versa.

`simplifyBy` a number that tells the function to change the values smaller than that to zero. The purpose behind this option is to facilitate recognizing candidate genes. Therefore, it is not suited for publications. It has the same unit as `cutoff`.

`genesToDrop` a character vector. Gene names within this vector will be omitted from heatmap.The default value is `FALSE`.

Details

Package:	cbaf
Type:	Package
Version:	1.1.4
Date:	2017-11-23
License:	Artistic-2.0

Value

a BiocFileCache object that containes some or all of the following groups, based on what user has chosen: obtainedData, validationResults, frequencyPercentage, Top.Genes.of.Frequency.Percentage, frequencyRatio, meanValue, Top.Genes.of.Mean.Value, medianValue, Top.Genes.of.Median.Value. It also saves these results in one excel files for convenience. Based on preference, three heatmaps for frequency percentage, mean value and median can be generated. If more than one group of genes is entered, output for each group will be strored in a separate sub-directory.

Author(s)

Arman Shahrisa, <shahrisa.arman@hotmail.com> [maintainer, copyright holder]

Maryam Tahmasebi Birgani, <tahmasebi-ma@ajums.ac.ir>

Examples

```
genes <- list(K.demethylases = c("KDM1A", "KDM1B", "KDM2A", "KDM2B", "KDM3A",
 "KDM3B", "JMJD1C", "KDM4A"), K.methyltransferases = c("SUV39H1", "SUV39H2",
 "EHMT1", "EHMT2", "SETDB1", "SETDB2", "KMT2A", "KMT2A"))

studies <- c("Acute Myeloid Leukemia (TCGA, Provisional)",
"Adrenocortical Carcinoma (TCGA, Provisional)",
"Bladder Urothelial Carcinoma (TCGA, Provisional)",
"Brain Lower Grade Glioma (TCGA, Provisional)",
"Breast Invasive Carcinoma (TCGA, Provisional)")

processMultipleStudies(genes, "test2", studies, "RNA-Seq",
calculate = c("frequencyPercentage", "frequencyRatio"), heatmapMargines =
c(16,10), RowCex = 1, ColCex = 1)
```

---

processOneStudy *Check Expression/methylation Profile for various subgroups of a cancer study.*

Description

This function Obtains the requested data for the given genes across multiple subgroups of a cancer. It can check whether or not all genes are included in subgroups of a cancer study and, if not, looks for the alternative gene names. Then it calculates frequency percentage, frequency ratio, mean value and median value of samples greather than specific value in the selected subgroups of the cancer. Furthermore, it looks for the five genes that comprise the highest values in each cancer study subgroup.

Usage

```
processOneStudy(genesList, submissionName, studyName, desiredTechnique
  , desiredCaseList = FALSE, validateGenes = TRUE, calculate =
  c("frequencyPercentage", "frequencyRatio", "meanValue"), cutoff=NULL,
  round=TRUE, topGenes = TRUE, shortenStudyNames = TRUE, geneLimit = FALSE,
  rankingMethod = "variation", heatmapFileFormat = "TIFF", resolution = 600,
  RowCex = "auto", ColCex = "auto", heatmapMargines = "auto",
  rowLabelsAngle = 0, columnLabelsAngle = 45, heatmapColor = "RdBu",
```

```
  reverseColor = TRUE, transposedHeatmap = FALSE, simplifyBy = FALSE,
  genesToDrop = FALSE)
```

Arguments

genesList	a list that contains at least one gene group
submissionName	a character string containing name of interest. It is used for naming the process.
studyName	a character string showing the desired cancer name. It is an standard cancer study name that can be found on cbioportal.org, such as "Acute Myeloid Leukemia (TCGA, NEJM 20
desiredTechnique	a character string that is one of the following techniques: "RNA-Seq", "microRNA-Seq", "microarray.mRNA" , "microarray.microRNA" or "methylation".
desiredCaseList	a numeric vector that contains the index of desired cancer subgroups, assuming the user knows index of desired subgroups. If not, desiredCaseList is set to "none", function will show the available subgroups and ask the user to enter the desired ones during the process. The default value is "none".
validateGenes	a logical value that, if set to be TRUE, causes the function to check each cancer study to find whether or not each gene has a record. If a cancer doesn't have a record for specific gene, function looks for alternative gene names that cbioportal might use instead of the given gene name.
calculate	a character vector that containes the statistical procedures users prefer the function to compute. The complete results can be obtained by c("frequencyPercentage", "frequencyl This will tell the function to compute the following: "frequencyPercentage", which is the percentge of samples having the value greather than specific cutoff divided by the total sample size for every study / study subgroup; "frequency ratio", which shows the number of selected samples divided by the total number of samples that give the frequency percentage for every study / study subgroup. It shows the selected and total sample sizes.; "Mean Value", that contains mean value of selected samples for each study; "Median Value", which shows the median value of selected samples for each study. The default input is calculate = c("frequencyPer
cutoff	a number used to limit samples to those that are greather than specific number (cutoff). The default value for methylation data is 0.6 while gene expression studies use default value of 2. For methylation studies, it is observed/expected ratio, for the rest, it is "z-score". To change the cutoff to any desired number, change the option to cutoff = desiredNumber, in which desiredNumber is the number of interest.
round	a logical value that, if set to be TRUE, will force the function to round all the calculated values to two decimal places. The default value is TRUE.
topGenes	a logical value that, if set as TRUE, causes the function to create three dataframes that contain the five top genes for each cancer. To get all the three dataframes, "frequencyPercentage", "meanValue" and "medianValue" must have been included for "calculate".
shortenStudyNames	a logical vector. If the value is set as TRUE, function will try to remove the last part of the cancer names aiming to shorten them. The removed segment usually contains the name of scientific group that has conducted the experiment.
geneLimit	if large number of genes exist in at least one gene group, this option can be used to limit the number of genes that are shown on heatmap. For instance, geneLimit=50 will limit the heatmap to 50 genes showing the most variation across multiple study / study subgroups. The default value is none.

`rankingMethod` a character value that determines how genes will be ranked prior to drawing heatmap. `"variation"` orders the genes based on unique values in one or few cancer studies while `"highValue"` ranks the genes when they cotain high values in multiple / many cancer studies. This option is useful when number of genes are too much so that user has to limit the number of genes on heatmap by `geneLimit`.

`heatmapFileFormat`
This option enables the user to select the desired image file format of the heatmaps. The default value is `"TIFF"`. Other suppoeted formats include `"PNG"`, `"BMP"`, and `"JPG"`.

`resolution` a number. This option can be used to adjust the resolution of the output heatmaps as 'dot per inch'. The defalut value is 600.

`RowCex` a number that specifies letter size in heatmap row names, which ranges from 0 to 2. If `RowCex = "auto"`, the function will automatically determine the best RowCex.

`ColCex` a number that specifies letter size in heatmap column names, which ranges from 0 to 2. If `ColCex = "auto"`, the function will automatically determine the best ColCex.

`heatmapMargines`
a numeric vector that is used to set heatmap margins. If `heatmapMargines = "auto"`, the function will automatically determine the best possible margines. Otherwise, enter the desired margine as e.g. c(10,10.)

`rowLabelsAngle` a number that determines the angle with which the gene names are shown in heatmaps. The default value is 0 degree.

`columnLabelsAngle`
a number that determines the angle with which the studies/study subgroups names are shown in heatmaps. The default value is 45 degree.

`heatmapColor` a character string that defines heatmap color. The default value is `'RdBu'`. `'RdGr'` is also a popular color in genomic studies. To see the rest of colors, please type `library(RColorBrewer)` and then `display.brewer.all()`.

`reverseColor` a logical value that reverses the color gradiant for heatmap(s).

`transposedHeatmap`
a logical value that transposes heatmap rows to columns and vice versa.

`simplifyBy` a number that tells the function to change the values smaller than that to zero. The purpose behind this option is to facilitate recognizing candidate genes. Therefore, it is not suited for publications. It has the same unit as `cutoff`.

`genesToDrop` a character vector. Gene names within this vector will be omitted from heatmap.The default value is `FALSE`.

Details

Package:	cbaf
Type:	Package
Version:	1.1.4
Date:	2017-11-23
License:	Artistic-2.0

Value

a BiocFileCache object that containes some or all of the following groups, based on what user has chosen: `ObtainedData`, `validationResults`, `frequencyPercentage`, `Top.Genes.of.Frequency.Percentage`, `frequencyRatio`, `meanValue`, `Top.Genes.of.Mean.Value`, `medianValue`, `Top.Genes.of.Median.Value`. It also saves these results in one excel files for convenience. Based on preference, three heatmaps for frequency percentage, mean value and median can be generated. If more than one group of genes is entered, output for each group will be strored in a separate sub-directory.

Author(s)

Arman Shahrisa, <shahrisa.arman@hotmail.com> [maintainer, copyright holder]

Maryam Tahmasebi Birgani, <tahmasebi-ma@ajums.ac.ir>

Examples

```
genes <- list(K.demethylases = c("KDM1A", "KDM1B", "KDM2A", "KDM2B", "KDM3A",
 "KDM3B", "JMJD1C", "KDM4A"), K.methyltransferases = c("SUV39H1", "SUV39H2",
 "EHMT1", "EHMT2", "SETDB1", "SETDB2", "KMT2A", "KMT2A"))

processOneStudy(genes, "test", "Breast Invasive Carcinoma (TCGA, Cell 2015)",
"RNA-Seq", desiredCaseList = c(2,3,4,5), calculate = c("frequencyPercentage",
"frequencyRatio"), heatmapMargines = c(16, 10), RowCex = 1, ColCex = 1)
```

`xlsxOutput` *Generate excel output for various studies/subgroups of a study.*

Description

This function generates excel files containing gene validation and all selected statistical methods. It uses outputs of obtainOneStudy()/obtainMultipleStudies() and automatedStatistics() functions.

Usage

```
xlsxOutput(submissionName)
```

Arguments

`submissionName`	a character string containing name of interest. It is used for naming the process.

Details

Package:	cbaf
Type:	Package
Version:	1.1.4
Date:	2017-11-23
License:	Artistic-2.0

Value

It generates one excel file for each gene group. This excel file contains output of automatedStatistics() and validation result from output of either obtainOneStudy() or obtainMultipleStudies().

Author(s)

Arman Shahrisa, <shahrisa.arman@hotmail.com> [maintainer, copyright holder]

Maryam Tahmasebi Birgani, <tahmasebi-ma@ajums.ac.ir>

Examples

```
genes <- list(K.demethylases = c("KDM1A", "KDM1B", "KDM2A", "KDM2B", "KDM3A",
 "KDM3B", "JMJD1C", "KDM4A"), K.methyltransferases = c("SUV39H1", "SUV39H2",
 "EHMT1", "EHMT2", "SETDB1", "SETDB2", "KMT2A", "KMT2A"))

obtainOneStudy(genes, "test", "Breast Invasive Carcinoma (TCGA, Cell 2015)",
"RNA-Seq", desiredCaseList = c(3,4))

automatedStatistics("test", obtainedDataType = "single study", calculate =
c("frequencyPercentage", "frequencyRatio"))

xlsxOutput("test")
```

Index