Package 'CancerMutationAnalysis'

March 25, 2013

Type Package

Title Cancer mutation analysis

Version 1.2.1

Author Giovanni Parmigiani, Simina M. Boca

Maintainer Simina M. Boca <simina.boca@nih.gov>

Imports AnnotationDbi, limma, methods, stats

Depends R (>= 2.10.0), qvalue

Suggests KEGG.db

Description This package implements gene and gene-set level analysis methods for somatic mutation studies of cancer. The gene-level methods distinguish between driver genes (which play an active role in tumorigenesis) and passenger genes (which are mutated in tumor samples, but have no role in tumorigenesis) and incorporate a two-stage study design. The gene-set methods implement a patient-oriented approach, which calculates gene-set scores for each sample, then combines them across samples; a gene-oriented approach which uses the Wilcoxon test is also provided for comparison.

biocViews Genetics, Bioinformatics, Software

License GPL (>= 2) + file LICENSE

LazyLoad yes

R topics documented:

BackRatesBreast *Data from the Wood et al. 2007 study: Background mutation rates*

Description

Background rates for somatic mutations used in the breast cancer portion of the Wood et al. 2007 study.

Usage

```
data(WoodBreast07)
```

Format

The background rates for somatic mutations used in the breast cancer portion of the Wood et al. study, broken down by mutation type. The object is a data frame, with the variables representing the 25 different mutation types, and the rows specifying whether the estimates of the backgrond rates are "Lower," "Median," or "Upper," as well as whether or not the rates are separately estimated for the prevalence screen (denoted by "SepPrev").

References

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* DOI:10.1126/science.1145720

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovBreast, GeneSampBreast, GeneAlterBreast

---

BackRatesColon *Data from the Wood et al. 2007 study: Background mutation rates*

Description

Background rates for somatic mutations used in the colon cancer portion of the Wood et al. 2007 study.

Usage

data(WoodColon07)

Format

The background rates for somatic mutations used in the colon cancer portion of the Wood et al. study, broken down by mutation type. The object is a data frame, with the variables representing the 25 different mutation types, and the rows specifying whether the estimates of the backgrond rates are "Lower," "Median," or "Upper," as well as whether or not the rates are separately estimated for the prevalence screen (denoted by "SepPrev").

References

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* DOI:10.1126/science.1145720

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovColon, GeneSampColon, GeneAlterBreast

---

BackRatesGBM *Data from the Parsons et al. 2008 study: Background mutation rates*

Description

Background rates for somatic mutations used in the Parsons et al. 2008 glioblastoma multiforme (GBM) study.

Usage

data(ParsonsGBM08)

Format

The background rates for somatic mutations used in the Parsons et al. GBM study, broken down by mutation type. The object is a data frame, with the variables representing the 25 different mutation types, and the rows specifying whether the estimates of the backgrond rates are "Upper," "Median," or "Lower."

References

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science.* DOI: 10.1126/science.1164382

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovGBM, GeneSampGBM, GeneAlterGBM

BackRatesMB *Data from the Parsons et al. 2011 study: Background mutation rates*

Description

Background rates for somatic mutations used in the Parsons et al. 2011 medulloblastoma (MB) study.

Usage

```
data(ParsonsMB11)
```

Format

The background rates for somatic mutations used in the Parsons et al. MB study, broken down by mutation type. The object is a data frame which has a single row, with the variables representing the 25 different mutation types.

References

Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin J, Boca SM, Carter H, Samayoa J, Bettegowda C, et al. The genetic landscape of the childhood cancer medulloblastoma. *Science.* DOI: 10.1126/science.1198056

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovMB, GeneSampMB, GeneAlterMB

BackRatesPancreas *Data from the Jones et al. 2008 study: Background mutation rates*

Description

Background rates for somatic mutations used in the Jones et al. 2008 pancreatic cancer study.

Usage

```
data(JonesPancreas08)
```

Format

The background rates for somatic mutations used in the Jones et al. pancreatic cancer study, broken down by mutation type. The object is a data frame, with the variables representing the 25 different mutation types, and the rows specifying whether the estimates of the backgrond rates are "Upper," "Median," or "Lower."

References

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science.* DOI: 10.1126/science.1164368

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovGBM, GeneSampGBM, GeneAlterGBM

---

cma.fdr *Gene-level Empirical Bayes (EB) false discovery rate (FDR) analysis for somatic mutations in cancer*

Description

Empirical Bayes estimates of the False Discovery Rate (FDR) and passenger probabilities in the analysis of somatic mutations in cancer.

Usage

```
cma.fdr(cma.alter,
        cma.cov,
        cma.samp,
        scores = c("CaMP", "logLRT"),
        passenger.rates = t(data.frame(.55*rep(1.0e-6,25))),
        allgenes=TRUE,
```

estimate.p0=FALSE,
p0.step=1,
p0=1,
eliminate.noval=FALSE,
filter.threshold=0,
filter.above=0,
filter.below=0,
filter.mutations=0,
aa=1e-10,
bb=1e-10,
priorH0=1-500/13020,
prior.a0=100,
prior.a1=5,
prior.fold=10,
M=2,
DiscOnly=FALSE,
PrevSamp="Sjoeblom06",
KnownCANGenes=NULL,
showFigure=FALSE,
cutoffFdr=0.1)

Arguments

cma.alter	Data frame with somatic mutation information, broken down by gene, sample, screen, and mutation type. See GeneAlterBreast for an example.
cma.cov	Data frame with the total number of nucleotides "at risk" ("coverage"), broken down by gene, screen, and mutation type. See GeneCovBreast for an example.
cma.samp	Data frame with the number of samples analyzed, broken down by gene and screen. See GeneSampBreast for an example.
scores	Vector with the scores which are to be computed. It can include: CaMP (Cancer Mutation Prevalence score), logLRT (log Likelihood Ratio Test score), neglogPg, logLRT, logitBinomialPosteriorDriver, PoissonlogBF, PoissonPosterior, Poissonlmlik0, Poissonlmlik1
passenger.rates	Data frame of passenger mutation rates per nucleotide, by type, or "context". If two rows are present, the first refers to the Discovery screen and the second to the Prevalence screen.
allgenes	If TRUE, genes where no mutations were found are considered in the analysis.
estimate.p0	If TRUE, estimates the percent of genes with only passenger mutations. Requires allgenes=TRUE
p0.step	Size of bins of histograms in the distribution of scores, to use in estimating p0 if estimate.p0 = TRUE. All scores are in the log 10 scale.
p0	Proportion of genes with only passenger mutations. Only used if estimate.p0=FALSE
eliminate.noval	If TRUE, the genes which are not validated are eliminated from the analysis. Validated genes are those where at least one mutation was found in both the Discovery and Prevalence (or Validation) screens.
filter.threshold	This and the following three input control filtering of genes, allowing to exclude genes from analysis, by size and number of mutations. Different criteria can be set above and below this threshold. The threshold is a gene size in base pairs.
filter.above	Minimum number of mutations per Mb, applied to genes of size greater than threshold.size.

filter.below	Minimum number of mutations per Mb, applied to genes of size lower than threshold.size.
filter.mutations	Only consider genes whose total number of mutations is greater than or equal to filter.mutations.
aa	Hyperparameter of beta prior used in compute.binomial.posterior.
bb	Hyperparameter of beta prior used in compute.binomial.posterior.
priorH0	Prior probability of the null hypothesis, used to convert the BF in compute.poisson.BF to a posterior probability
prior.a0	Shape hyperparameter of gamma prior on passenger rates used in compute.poisson.BF
prior.a1	Shape hyperparameter of gamma prior on non-passenger rates used in compute.poisson.BF
prior.fold	Hyperparameter of gamma prior on non-passenger rates used compute.poisson.BF. The mean of the gamma is set so that the ratio of the mean to the passenger rate is the specified prior.fold in each type.
M	The number of null datasets generated to get the false discovery rates. Numbers on the order of 100 are recommended, but this will cause the function to run very slowly.
DiscOnly	If TRUE, only considers data from Discovery screen.
PrevSamp	If "Sjoeblom06", then the experimental design from Sjoeblom et al. or Wood et al. is used, namely, genes "pass" from the Discovery into the Prevalence (or Validation) screens if they are mutated at least once in the Discovery samples. If "Parsons11", the experimental design from Parsons et al. 2011 is approximated, namely, in the null datasets, a gene passes into the Prevalence screen if it is mutated at least once, and is found on a specified list of known cancer candidate (CAN) genes, or if it is mutated at least twice.
KnownCANGenes	Vector of known CAN genes, to be used if PrevSamp is not set to "Sjoeblom07".
showFigure	If TRUE, displays a figure for each score in scores, showing the right tail of the density of scores under the null, the right tail of the density of real scores as a rug (1-d) plot and the number of real genes and average number of null genes to the right of the cutoff chosen based on cutoffFdr.
cutoffFdr	If showFigure is set to TRUE, it gives the value at which we are interested in controlling the false discovery rate (Fdr). The corresponding score threshold is plotted on the figure, with the number of real genes greater than it and the average number of null genes greater than it specified. The estimated Fdr at that threshold is the ratio of the average number of null genes and the number of real genes, multiplied by p0, which is often taken to be 1.

Value

A list of data frames. Each gives a gene gene-by-gene significance for one of the score requested. The columns in each data frame are:

score	The score requested (e.g. the LRT).
F	Number of genes experimentally observed to give a larger score than the gene in question.
F0	Number of genes giving a larger score than the gene in question in datasets simulated from passenger mutation rates.

Fdr	The Empirical Bayes False Discovery Rate, as defined in Efron and Tibshirani 2002.
fdr	The Empirical Bayes Local False Discovery Rate, as defined in Efron and Tibshirani 2002.
p0	Scalar, Proportion of genes with only passenger mutations. Estimated or passed on from input (depending on whether estimate.p0 is TRUE

Author(s)

Giovanni Parmigiani, Simina M. Boca

References

Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*. DOI: 10.1002/gepi.1124

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data, 2007. http://www.bepress.com/jhubiostat/paper126/

Sjoeblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, et al. The consensus coding sequences of breast and colorectal cancers. *Science.* DOI: 10.1126/science.1133427

Wood LD, Parsons DW, Jones S, Lin J, Sjoeblom, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The Genomic Landscapes of Human Breast and Colorectal Cancer. *Science.* DOI: 10.1126/science.1145720

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science.* DOI: 10.1126/science.1164382

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science.* DOI: 10.1126/science.1164368

Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin J, Boca SM, Carter H, Samayoa J, Bettegowda C, et al. The genetic landscape of the childhood cancer medulloblastoma. *Science.* DOI: 10.1126/science.1198056

See Also

GeneCov, GeneSamp, GeneAlter, BackRates,cma.scores

Examples

```
data(ParsonsMB11)
set.seed(188310)
cma.fdr.out <- cma.fdr(cma.alter = GeneAlterMB,
                       cma.cov = GeneCovMB,
                       cma.samp = GeneSampMB,
                       allgenes = TRUE,
                       estimate.p0=FALSE,
                       eliminate.noval=FALSE,
                       filter.mutations=0,
                       M = 2)
names(cma.fdr.out)
```

cma.scores *Gene-level scores for the analysis of somatic point mutations in cancer*

Description

Computes various gene-level scores for the analysis of somatic point mutations in cancer.

Usage

```
cma.scores(cma.alter = NULL,
        cma.cov,
        cma.samp,
        scores = c("CaMP", "logLRT"),
        cma.data = NULL,
        coverage = NULL,
   passenger.rates = t(data.frame(0.55*rep(1.0e-6,25))),
        allow.separate.rates = TRUE,
        filter.above=0,
        filter.below=0,
        filter.threshold=0,
   filter.mutations=0,
        aa=1e-10,
        bb=1e-10,
        priorH0=1-300/13020,
        prior.a0=100,
        prior.a1=5,
        prior.fold=10)
```

Arguments

cma.alter	Data frame with somatic mutation information, broken down by gene, sample, screen, and mutation type. See GeneAlterBreast for an example.
cma.cov	Data frame with the total number of nucleotides "at risk" ("coverage"), broken down by gene, screen, and mutation type. See GeneCovBreast for an example.
cma.samp	Data frame with the number of samples analyzed, broken down by gene and screen. See GeneSampBreast for an example.
scores	Vector with the scores which are to be computed. It can include: CaMP (Cancer Mutation Prevalence score), logLRT (log Likelihood Ratio Test score), neglogPg, logLRT, logitBinomialPosteriorDriver, PoissonlogBF, PoissonPosterior, Poissonlmlik0, Poissonlmlik1
cma.data	Provided for back-compatibility and internal operations. cma.data and coverage objects were used in prior versons of this package, and may be specified instead of cma.alter, cma.cov, and cma.samp.
coverage	Provided for back-compatibility and internal operations. cma.data and coverage objects were used in prior versons of this package, and may be specified instead of cma.alter, cma.cov, and cma.samp.
passenger.rates	Data frame of "passenger" (or "background") mutation rates per nucleotide, by type, or "context". If two rows are present, the first refers to the Discovery screen and the second to the Prevalence screen.

allow.separate.rates
: If TRUE, allows for use separate rates for Discovery and Prevalence screens.

filter.threshold
: This and the following three input control filtering of genes, allowing to exclude genes from analysis, by size and number of mutations. Different criteria can be set above and below this threshold. The threshold is a gene size in base pairs.

filter.above
: Minimum number of mutations per Mb, applied to genes of size greater than threshold.size.

filter.below
: Minimum number of mutations per Mb, applied to genes of size lower than threshold.size.

filter.mutations
: Only consider genes whose total number of mutations is greater than or equal to filter.mutations.

aa
: Hyperparameter of beta prior used in compute.binomial.posterior.

bb
: Hyperparameter of beta prior used in compute.binomial.posterior

priorH0
: Prior probability of the null hypothesis, used to convert the BF in compute.poisson.BF to a posterior probability

prior.a0
: Shape hyperparameter of gamma prior on passenger rates used in compute.poisson.BF

prior.a1
: Shape hyperparameter of gamma prior on non-passenger rates used in compute.poisson.BF

prior.fold
: Hyperparameter of gamma prior on non-passenger rates used compute.poisson.BF. The mean of the gamma is set so that the ratio of the mean to the passenger rate is the specified prior.fold in each type.

Details

The scores computed by this function are relevant for two stage experiments like the one in the Sjoeblom et al. article. In this design genes are sequenced in a first "Discovery" sample. A non-random set of genes is then also sequenced in a subsequent "Prevalence" (or "Validation") screen. For instance, in Sjoeblom et al. and Wood et al., genes "pass" the Discovery screen if they are mutated at least once in it. The goal of this tool is to facilitate reanalysis of the Sjoeblom et al. 2006, Wood et al. 2007, Jones et al. 2008, Parsons et al. 2008, and Parsons et al. 2011 datasets. Application to other projects requires a detailed understanding of these projects.

Value

A data frame giving gene-by-gene values for each score. The columns in this data frame are:

CaMP
: The CaMP score of Sjoeblom and colleagues.

neglogPg
: The negative log10 of Pg, where Pg represents the probability that a gene has its exact observed mutation profile under the null, i.e. assuming the given passenger rates.

logLRT
: The log10 of the likelihood ratio test (LRT).

logitBinomialPosteriorDriver
: logit of the posterior probability that a gene's mutation rates above the specified passenger rates using a binomial model

PoissonlogBF
: The log10 of the Bayes Factor (BF) using a Poisson-Gamma model.

PoissonPosterior
: The posterior probability that a given gene is a driver, using a Poisson-Gamma model.

Poissonlmlik0
: Marginal likelihood under the null hypothesis in the Poisson-Gamma model

Poissonlmlik1
: Marginal likelihood under the alternative hypothesis in the Poisson-Gamma model

Author(s)

Giovanni Parmigiani, Simina M. Boca

References

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

Sjoeblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, et al. The consensus coding sequences of breast and colorectal cancers. *Science.* DOI: 10.1126/science.1133427

Wood LD, Parsons DW, Jones S, Lin J, Sjoeblom, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The Genomic Landscapes of Human Breast and Colorectal Cancer. *Science.* DOI: 10.1126/science.1145720

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science.* DOI: 10.1126/science.1164382

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science.* DOI: 10.1126/science.1164368

Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin J, Boca SM, Carter H, Samayoa J, Bettegowda C, et al. The genetic landscape of the childhood cancer medulloblastoma. *Science.* DOI: 10.1126/science.1198056

See Also

GeneCov, GeneSamp, GeneAlter, BackRates, cma.set.stat

Examples

```
data(ParsonsGBM08)
ScoresGBM <- cma.scores(cma.alter = GeneAlterGBM,
                        cma.cov = GeneCovGBM,
                        cma.samp = GeneSampGBM)
```

---

cma.set.sim *Simulates data and performs gene-set analysis methods on the simulated datasets.*

Description

This function simulates data under the passenger or permutation null, either under the null or including spiked-in gene-sets. It then calculates the p-values and q-values for all the selected gene-set analysis methods.

Usage

```
cma.set.sim(cma.alter,
            cma.cov,
            cma.samp,
  GeneSets,
  passenger.rates = t(data.frame(0.55*rep(1.0e-6,25))),
  ID2name=NULL,
  BH = TRUE,
  nr.iter,
  pass.null = FALSE,
  perc.samples = NULL,
  spiked.set.sizes = NULL,
  gene.method = FALSE,
  perm.null.method = TRUE,
  perm.null.het.method = FALSE,
  pass.null.method = FALSE,
  pass.null.het.method = FALSE,
  show.iter,
  KnownMountains = c("EGFR","SMAD4","KRAS",
  "TP53","CDKN2A","MYC","MYCN","PTEN","RB1"),
  exclude.mountains=TRUE,
            verbose=TRUE)
```

Arguments

cma.alter	Data frame with somatic mutation information, broken down by gene, sample, screen, and mutation type. See GeneAlterBreast for an example.
cma.cov	Data frame with the total number of nucleotides "at risk" ("coverage"), broken down by gene, screen, and mutation type. See GeneCovBreast for an example.
cma.samp	Data frame with the number of samples analyzed, broken down by gene and screen. See GeneSampBreast for an example.
GeneSets	An object which annotates genes to gene-sets; it can either be a list with each component representing a set, or an object of the class AnnDbBimap.
passenger.rates	Data frame with 1 row and 25 columns, of passenger mutation rates per nucleotide, by type, or "context". Columns denote types and must be in the same order as the first 25 columns in the MutationsBrain objects.
ID2name	Vector mapping the gene identifiers used in the GeneSets object to the gene names used in the other objects; if they are the same, this parameter is not needed. See EntrezID2Name for an example.
BH	If set to TRUE, uses the Benjamini-Hochberg method to get q-values; if set to FALSE, uses the Storey method from the qvalue package.
nr.iter	The number of iterations to be simulated.
pass.null	If set to true TRUE, implements the passenger null hypothesis, using the rates from passenger.rates; otherwise, implements the permutation null, permuting mutational events.
perc.samples	Vector representing the probabilities of the spiked-in gene-sets being altered in any given sample, as percentages; for example perc.samples = c(75, 90) means that these probabilities are 0.75 and 0.90.

spiked.set.sizes Vector representing the sizes, in genes, of the spiked-in gene-sets; for example, if perc.samples = c(75, 90) and spiked.set.sizes = c(50, 100), there would be 4 spiked-in sets, one with 50 genes and probability of being altered of 0.75 in each sample, one with 50 genes and probability of being altered of 0.90 in each sample, one with 100 genes and probability of being altered of 0.75 in each sample, and one with 100 genes and probability of being altered of 0.90 in each sample.

gene.method If set to TRUE, implements gene-oriented method.

perm.null.method
If set to TRUE, implements patient-oriented method with permutation null and no heterogeneity.

perm.null.het.method
If set to TRUE, implements patient-oriented method with permutation null and heterogeneity.

pass.null.method
If set to TRUE, implements patient-oriented method with passenger null and no heterogeneity.

pass.null.het.method
If set to TRUE, implements patient-oriented method with passenger null and heterogeneity.

show.iter If set to TRUE and verbose is also set to TRUE, shows what simulation is currently running.

KnownMountains
Vector of genes to be excluded from the permutation null simulations if exclude.mountains = TRUE

exclude.mountains
If set to TRUE, excludes the genes in KnownMountains.

verbose If TRUE, prints intermediate messages.

Value

An object of the class SetMethodsSims. See SetMethodsSims for more details.

Author(s)

Simina M. Boca, Giovanni Parmigiani.

References

Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene-set analysis for cancer mutation data. *Genome Biology.* DOI: 10.1186/gb-2010-11-11-r112

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, DOI: 10.2307/2346101

Storey JD and Tibshirani R. Statistical significance for genome-wide experimens. *Proceedings of the National Academy of Sciences.* DOI: 10.1073/pnas.1530509100

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science.* DOI: 10.1126/science.1164382

Wood LD, Parsons DW, Jones S, Lin J, Sjoeblom, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The Genomic Landscapes of Human Breast and Colorectal Cancer. *Science.* DOI: 10.1126/science.1145720

See Also

SetMethodsSims-class, CoverageBrain, EventsBySampleBrain, GeneSizes08, MutationsBrain, ID2name, cma.set.stat, extract.sims.method, combine.sims

Examples

```
##Note that this takes a few minutes to run:
library(KEGG.db)
data(ParsonsGBM08)
data(EntrezID2Name)

setIDs <- c("hsa00250", "hsa05213")
set.seed(831984)
ResultsSim <-
   cma.set.sim(cma.alter = GeneAlterGBM,
               cma.cov = GeneCovGBM,
               cma.samp = GeneSampGBM,
               GeneSets =  KEGGPATHID2EXTID[setIDs],
               ID2name = EntrezID2Name,
               nr.iter = 2,
               pass.null = TRUE,
               perc.samples = c(75, 95),
               spiked.set.sizes = 50,
               perm.null.method = TRUE,
               pass.null.method = TRUE)

ResultsSim
```

---

cma.set.stat *Implements gene-set analysis methods.*

Description

This function implements the gene-set analysis methods. It returns a data-frame with p-values and q-values for all the methods selected.

Usage

```
cma.set.stat(cma.alter,
               cma.cov,
               cma.samp,
               GeneSets,
               ID2name=NULL,
   Scores,
   passenger.rates = t(data.frame(0.55*rep(1.0e-6,25))),
   BH = TRUE,
   gene.method = FALSE,
   perm.null.method = TRUE,
```

```
perm.null.het.method = FALSE,
pass.null.method = FALSE,
pass.null.het.method = FALSE,
            score = "logLRT",
            verbose = TRUE)
```

Arguments

cma.alter	Data frame with somatic mutation information, broken down by gene, sample, screen, and mutation type. See GeneAlterBreast for an example.
cma.cov	Data frame with the total number of nucleotides "at risk" ("coverage"), broken down by gene, screen, and mutation type. See GeneCovBreast for an example.
cma.samp	Data frame with the number of samples analyzed, broken down by gene and screen. See GeneSampBreast for an example.
GeneSets	An object which annotates genes to gene-sets; it can either be a list with each component representing a set, or an object of the class AnnDbBimap.
ID2name	Vector mapping the gene identifiers used in the GeneSets object to the gene names used in the other objects; if they are the same, this parameter is not needed. See EntrezID2Name for an example.
Scores	Data frame of gene scores. The logLRT scores are used for the gene.method option. It can be the output of cma.scores. If the gene.method option is set to FALSE, this parameter is not needed.
passenger.rates	Data frame with 1 row and 25 columns, of passenger mutation rates per nucleotide, by type, or "context". Columns denote types and must be in the same order as the first 25 columns in the MutationsBrain objects.
BH	If set to TRUE, uses the Benjamini-Hochberg method to get q-values; if set to FALSE, uses the Storey method from the qvalue package.
gene.method	If set to TRUE, implements gene-oriented method.
perm.null.method	If set to TRUE, implements patient-oriented method with permutation null and no heterogeneity.
perm.null.het.method	If set to TRUE, implements patient-oriented method with permutation null and heterogeneity.
pass.null.method	If set to TRUE, implements patient-oriented method with passenger null and no heterogeneity.
pass.null.het.method	If set to TRUE, implements patient-oriented method with passenger null and heterogeneity.
score	Can be any of the scores which result from cma.scores. Specifies the gene-scoring mechanism used in the gene-oriented method.
verbose	If TRUE, prints intermediate messages.

Value

A data frame, with the rows representing set names and the columns representing the p-values and q-values corresponding to the different methods.

Author(s)

Simina M. Boca, Giovanni Parmigiani, Luigi Marchionni, Michael A. Newton.

References

Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene-set analysis for cancer mutation data. *Genome Biology.* DOI: 10.1186/gb-2010-11-11-r112

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B.* DOI: 10.2307/2346101

Storey JD and Tibshirani R. Statistical significance for genome-wide experiments. *Proceedings of the National Academy of Sciences.* DOI: 10.1073/pnas.1530509100

Schaeffer EM, Marchionni L, Huang Z, Simons B, Blackman A, Yu W, Parmigiani G, Berman DM. Androgen-induced programs for prostate epithelial growth and invasion arise in embryogenesis and are reactivated in cancer. *Oncogene.* DOI: 10.1038/onc.2008.327

Thomas MA, Taub AE. Calculating binomial probabilities when the trial probabilities are unequal. *Journal of Statistical Computation and Simulation.* DOI: 10.1080/00949658208810534

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science.* DOI: 10.1126/science.1164382

Wood LD, Parsons DW, Jones S, Lin J, Sjoeblom, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The Genomic Landscapes of Human Breast and Colorectal Cancer. *Science.* DOI: 10.1126/science.1145720

See Also

GeneCov, GeneSamp, GeneAlter, BackRates, cma.scores, cma.set.sim

Examples

```
library(KEGG.db)
data(ParsonsGBM08)
data(EntrezID2Name)

setIDs <- c("hsa00250", "hsa05213")
SetResults <- cma.set.stat(cma.alter = GeneAlterGBM,
                           cma.cov = GeneCovGBM,
                           cma.samp = GeneSampGBM,
                           GeneSets =  KEGGPATHID2EXTID[setIDs],
                           ID2name = EntrezID2Name,
                           perm.null.method = TRUE,
                           pass.null.method = TRUE)

SetResults
```

combine.sims *Combines two SetMethodSims objects.*

Description

This function is used to combine two SetMethodSims objects, which have the results from simulated datasets, provided that the values for pass.null, perc.samples, and spiked.set.sizes match up when the objects are generated with the sim.data.p.values function.

Usage

combine.sims(obj1, obj2)

Arguments

obj1	Object of the class SetMethodsSims.
obj2	Object of the class SetMethodsSims.

Value

An object of the class SetMethodsSims. See SetMethodsSims for more details.

Author(s)

Simina M. Boca, Giovanni Parmigiani.

References

Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene-set analysis for cancer mutation data. *Genome Biology.* DOI:10.1186/gb-2010-11-11-r112

See Also

SetMethodsSims-class, cma.set.sim

Examples

```
## Not run:
##Note that this takes a few minutes to run:
library(KEGG.db)
data(ParsonsGBM08)
data(EntrezID2Name)

setIDs <- c("hsa00250", "hsa05213")
set.seed(831984)
ResultsSim <-
   sim.data.p.values(cma.alter = GeneAlterGBM,
                     cma.cov = GeneCovGBM,
                     cma.samp = GeneSampGBM,
                     GeneSets =  KEGGPATHID2EXTID[setIDs],
                     ID2name = EntrezID2Name,
                     nr.iter = 2,
                     pass.null = TRUE,
```

```
                perc.samples = c(75, 95),
                spiked.set.sizes = 50,
                perm.null.method = TRUE,
                pass.null.method = TRUE)

ResultsSim

combine.sims(ResultsSim, ResultsSim)

## End(Not run)
```

EntrezID2Name *Map of gene IDs to gene names*

Description

Entrez gene identifiers used in the KEGG.db package are mapped to gene names.

Usage

```
data(EntrezID2Name)
```

Format

Vector having as names the Entrez gene identifiers used in the KEGG.db package and as entries the gene names used in the various data objects available.

References

ftp://ftp.genome.ad.jp/pub/kegg/pathways

See Also

cma.set.stat, cma.set.sim

---

extract.sims.method *Extracts the p-values or q-values from a SetMethodsSims object for a specific method.*

Description

This function is used to obtain a single data frame with the p-values or q-values from one of the specific gene-set analysis methods, from a SetMethodsSims object which has the results from simulated datasets.

Usage

```
extract.sims.method(object, method)
```

Arguments

object	Object of the class SetMethodsSims.
method	Character string giving the method used for extraction, and whether p-values or q-values are extracted. The string should be one of the column names of the data frame resulting from the cma.set.stat function.

Value

An object of the class SetMethodsSims. See SetMethodsSims for more details.

Author(s)

Simina M. Boca, Giovanni Parmigiani.

References

Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene-set analysis for cancer mutation data. *Genome Biology.* DOI:10.1186/gb-2010-11-11-r112

See Also

SetMethodsSims-class, cma.set.sim, cma.set.stat

Examples

```
## Not run:
##Note that this takes a few minutes to run:
library(KEGG.db)
data(ParsonsGBM08)
data(EntrezID2Name)

setIDs <- c("hsa00250", "hsa05213")
set.seed(831984)
ResultsSim <-
   sim.data.p.values(cma.alter = GeneAlterGBM,
                     cma.cov = GeneCovGBM,
                     cma.samp = GeneSampGBM,
                     GeneSets =  KEGGPATHID2EXTID[setIDs],
                     ID2name = EntrezID2Name,
                     nr.iter = 2,
                     pass.null = TRUE,
                     perc.samples = c(75, 95),
                     spiked.set.sizes = 50,
                     perm.null.method = TRUE,
                     pass.null.method = TRUE)

ResultsSim

extract.sims.method(ResultsSim, "p.values.perm.null")

## End(Not run)
```

GeneAlterBreast — *Data from the Wood et al. 2007 and Sjoeblom et al. 2006 studies: Alterations for every gene and sample*

Description

Somatic alterations for each gene and tumor sample from the breast cancer portion of the Wood et al. 2007 and Sjoeblom et al. 2006 studies.

Usage

```
data(WoodBreast07)
```

Format

The somatic mutations in the breast cancer portions of the Wood et al. and Sjoeblom et al. studies, broken down by *gene*, *type* (point mutation, amplification, or deletion), *sample*, *screen* (Discovery or Prevalence), and, for point mutations, *mutation type*, composed of the wild type nucleotide, its context, and the mutated nucleotide. The object is a data frame, with the variables: Gene, Type, Sample, Screen, WTNuc (wild type nucleotide), Context, and MutNuc (mutated nucleotide). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). The three possible values for Type are Mut (point mutations), Amp (large amplifications), and Del (large deletions.) Indels have a "" entry for WTNuc, an "All" entry for Context, and a "ins.del" entry for MutNuc. Large amplifications and deletions have "" entries for WTNuc, Context, and MutNuc. For this study, only point mutation are available.

References

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* DOI:10.1126/science.1145720

Sjoeblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, et al. The consensus coding sequences of human breast and colorectal cancers. *Science.* DOI: 10.1126/science.1133427

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovBreast, GeneSampBreast

GeneAlterColon *Data from the Wood et al. 2007 and Sjoeblom et al. 2006 studies: Alterations for every gene and sample*

Description

Somatic alterations for each gene and tumor sample from the colon cancer portion of the Wood et al. 2007 and Sjoeblom et al. 2006 studies.

Usage

```
data(WoodColon07)
```

Format

The somatic mutations in the colon cancer portions of the Wood et al. and Sjoeblom et al. studies, broken down by *gene*, *type* (point mutation, amplification, or deletion), *sample*, *screen* (Discovery or Prevalence), and, for point mutations, *mutation type*, composed of the wild type nucleotide, its context, and the mutated nucleotide. The object is a data frame, with the variables: Gene, Type, Sample, Screen, WTNuc (wild type nucleotide), Context, and MutNuc (mutated nucleotide). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). The three possible values for Type are Mut (point mutations), Amp (large amplifications), and Del (large deletions.) Indels have a "" entry for WTNuc, an "All" entry for Context, and a "ins.del" entry for MutNuc. Large amplifications and deletions have "" entries for WTNuc, Context, and MutNuc. For this study, only point mutation are available.

References

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* DOI:10.1126/science.1145720

Sjoeblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, et al. The consensus coding sequences of human breast and colorectal cancers. *Science.* DOI: 10.1126/science.1133427

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovColon, GeneSampColon

GeneAlterGBM *Data from the Parsons et al. 2008 study: Alterations for every gene and sample*

Description

Somatic alterations for each gene and tumor sample from the Parsons et al. 2008 glioblastoma multiforme (GBM) study.

Usage

```
data(ParsonsGBM08)
```

Format

The somatic mutations in the GBM study from Parsons et al., broken down by *gene*, *type* (point mutation, amplification, or deletion), *sample*, *screen* (Discovery or Prevalence), and, for point mutations, *mutation type*, composed of the wild type nucleotide, its context, and the mutated nucleotide. The object is a data frame, with the variables: Gene, Type, Sample, Screen, WTNuc (wild type nucleotide), Context, and MutNuc (mutated nucleotide). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). For this study, only the Discovery screen is considered. The three possible values for Type are Mut (point mutations), Amp (large amplifications), and Del (large deletions.) Indels have a "" entry for WTNuc, an "All" entry for Context, and a "ins.del" entry for MutNuc. Large amplifications and deletions have "" entries for WTNuc, Context, and MutNuc.

References

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science.* DOI: 10.1126/science.1164382

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovGBM, GeneSampGBM

GeneAlterMB *Data from the Parsons et al. 2011 study: Alterations for every gene and sample*

Description

Somatic alterations for each gene and tumor sample from the Parsons et al. 2011 medulloblastoma (MB) study.

Usage

```
data(ParsonsMB11)
```

Format

The somatic mutations in the MB study from Parsons et al., broken down by *gene*, *type* (point mutation, amplification, or deletion), *sample*, *screen* (Discovery or Prevalence), and, for point mutations, *mutation type*, composed of the wild type nucleotide, its context, and the mutated nucleotide. The object is a data frame, with the variables: Gene, Type, Sample, Screen, WTNuc (wild type nucleotide), Context, and MutNuc (mutated nucleotide). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). The three possible values for Type are Mut (point mutations), Amp (large amplifications), and Del (large deletions.) Indels have a "" entry for WTNuc, an "All" entry for Context, and a "ins.del" entry for MutNuc. Large amplifications and deletions have "" entries for WTNuc, Context, and MutNuc.

References

Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin J, Boca SM, Carter H, Samayoa J, Bettegowda C, et al. The genetic landscape of the childhood cancer medulloblastoma. *Science.* DOI: 10.1126/science.1198056

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovMB, GeneSampMB

GeneAlterPancreas *Data from the Jones et al. 2008 study: Alterations for every gene and sample*

Description

Somatic alterations for each gene and tumor sample from the Jones et al. 2008 pancreatic cancer study.

Usage

```
data(JonesPancreas08)
```

Format

The somatic mutations in the pancreatic cancer study from Jones et al., broken down by *gene*, *type* (point mutation, amplification, or deletion), *sample*, *screen* (Discovery or Prevalence), and, for point mutations, *mutation type*, composed of the wild type nucleotide, its context, and the mutated nucleotide. The object is a data frame, with the variables: Gene, Type, Sample, Screen, WTNuc (wild type nucleotide), Context, and MutNuc (mutated nucleotide). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). For this study, only the Discovery screen is considered. The three possible values for Type are Mut (point mutations), Amp (large amplifications), and Del (large deletions.) Indels have a "" entry for WTNuc, an "All" entry for Context, and a "ins.del" entry for MutNuc. Large amplifications and deletions have "" entries for WTNuc, Context, and MutNuc.

References

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science.* DOI: 10.1126/science.1164368

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneCovPancreas, GeneSampPancreas

---

GeneCovBreast *Data from the Wood et al. 2007 and Sjoeblom et al. 2006 studies: Total number of nucleotides "at risk" ("coverage")*

Description

Total numbers of nucleotides "at risk" that were successfully sequenced in RefSeq genes in the breast cancer portion of the Wood et al. 2007 and Sjoeblom et al. 2006 studies.

Usage

```
data(WoodBreast07)
```

Format

Total number of nucleotides available for mutations ("coverage") in the breast cancer portion of the Wood et al. and Sjoeblom et al. studies, broken down by *gene*, *screen* (Discovery or Prevalence), and *mutation type*, composed of the wild type nucleotide and its context. The object is a data frame, with the variables: Gene, Screen, WTNuc (wild type nucleotide), Context, and Coverage. The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). The nucleotides availables for indels are all the successfully sequenced nucleotides in a gene; the corresponding rows have a "" entry for WTNuc and an "All" entry for "Context." The nucleotides availables for other mutations are excluding nucleotides which can only give rise to synonymous mutations.

References

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* DOI:10.1126/science.1145720

Sjoeblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, et al. The consensus coding sequences of human breast and colorectal cancers. *Science.* DOI: 10.1126/science.1133427

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterBreast, GeneSampBreast

GeneCovColon — *Data from the Wood et al. 2007 and Sjoeblom et al. 2006 studies: Total number of nucleotides "at risk" ("coverage")*

Description

Total numbers of nucleotides "at risk" that were successfully sequenced in RefSeq genes in the colon cancer portion of the Wood et al. 2007 and Sjoeblom et al. 2006 studies.

Usage

```
data(WoodColon07)
```

Format

Total number of nucleotides available for mutations ("coverage") in the colon cancer portion of the Wood et al. and Sjoeblom et al. studies, broken down by *gene*, *screen* (Discovery or Prevalence), and *mutation type*, composed of the wild type nucleotide and its context. The object is a data frame, with the variables: Gene, Screen, WTNuc (wild type nucleotide), Context, and Coverage. The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). The nucleotides availables for indels are all the successfully sequenced nucleotides in a gene; the corresponding rows have a "" entry for WTNuc and an "All" entry for "Context." The nucleotides availables for other mutations are excluding nucleotides which can only give rise to synonymous mutations.

References

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* DOI:10.1126/science.1145720

Sjoeblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, et al. The consensus coding sequences of human breast and colorectal cancers. *Science.* DOI: 10.1126/science.1133427

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler WK, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterColon, GeneSampColon

GeneCovGBM — *Data from the Parsons et al. 2008 study: Total number of nucleotides "at risk" ("coverage")*

Description

Total numbers of nucleotides "at risk" that were successfully sequenced in RefSeq genes in the Parsons et al. 2008 glioblastoma multiforme (GBM) study.

Usage

data(ParsonsGBM08)

Format

Total number of nucleotides available for mutations ("coverage") in the GBM study from Parsons et al., broken down by *gene*, *screen* (Discovery or Prevalence), and *mutation type*, composed of the wild type nucleotide and its context. The object is a data frame, with the variables: Gene, Screen, WTNuc (wild type nucleotide), Context, and Coverage. The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). For this study, only the Discovery screen is considered. The nucleotides availables for indels are all the successfully sequenced nucleotides in a gene; the corresponding rows have a "" entry for WTNuc and an "All" entry for "Context." The nucleotides availables for other mutations are excluding nucleotides which can only give rise to synonymous mutations.

References

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science.* DOI: 10.1126/science.1164382

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterGBM, GeneSampGBM

GeneCovMB — *Data from the Parsons et al. 2011 study: Total number of nucleotides "at risk" ("coverage")*

Description

Total numbers of nucleotides "at risk" that were successfully sequenced in RefSeq genes in the Parsons et al. 2011 medulloblastoma (MB) study.

Usage

data(ParsonsMB11)

Format

Total number of nucleotides available for mutations ("coverage") in the MB study from Parsons et al., broken down by *gene*, *screen* (Discovery or Prevalence), and *mutation type*, composed of the wild type nucleotide and its context. The object is a data frame, with the variables: Gene, Screen, WTNuc (wild type nucleotide), Context, and Coverage. The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). The nucleotides availables for indels are all the successfully sequenced nucleotides in a gene; the corresponding rows have a "" entry for WTNuc and an "All" entry for "Context." The nucleotides availables for other mutations are excluding nucleotides which can only give rise to synonymous mutations.

References

Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin J, Boca SM, Carter H, Samayoa J, Bettegowda C, et al. The genetic landscape of the childhood cancer medulloblastoma. *Science.* DOI: 10.1126/science.1198056

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterMB, GeneSampMB

GeneCovPancreas *Data from the Jones et al. 2008 study: Total number of nucleotides "at risk" ("coverage")*

Description

Total numbers of nucleotides "at risk" that were successfully sequenced in RefSeq genes in the Jones et al. 2008 pancreatic cancer study.

Usage

```
data(JonesPancreas08)
```

Format

Total number of nucleotides available for mutations ("coverage") in the pancreatic cancer study from Jones et al., broken down by *gene*, *screen* (Discovery or Prevalence), and *mutation type*, composed of the wild type nucleotide and its context. The object is a data frame, with the variables: Gene, Screen, WTNuc (wild type nucleotide), Context, and Coverage. The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence"). For this study, only the Discovery screen is considered. The nucleotides availables for indels are all the successfully sequenced nucleotides in a gene; the corresponding rows have a "" entry for WTNuc and an "All" entry for "Context." The nucleotides availables for other mutations are excluding nucleotides which can only give rise to synonymous mutations.

References

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science.* DOI: 10.1126/science.1164368

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterPancreas, GeneSampPancreas

GeneID2Name11 *Map of gene IDs to gene names for the Parsons et al. 2011 medulloblastoma (MB) study*

Description

Gene identifiers used in the Parsons et al. 2011 MB study are mapped to gene names.

Usage

```
data(GeneID2Name11)
```

Format

Vector having as names gene identifiers and as entries the gene names used in the various data objects available.

See Also

GeneAlterMB, GeneCovMB, GeneSampMB

GeneSampBreast *Data from the Wood et al. 2007 and Sjoeblom et al. 2006 studies: Number of samples for each gene and screen type*

Description

Number of samples analyzed for each gene and screen type from the breast cancer portion of the Wood et al. 2007 and Sjoeblom et al. 2006 studies.

Usage

```
data(WoodBreast07)
```

Format

The number of samples in the breast cancer portions of the Wood et al. and Sjoeblom et al. studies, broken down by *gene* and *screen* (Discovery and Prevalence). The object is a data frame, with the variables: Gene, Screen, and NrSamp (number of samples). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence").

References

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* DOI:10.1126/science.1145720

Sjoeblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, et al. The consensus coding sequences of human breast and colorectal cancers. *Science.* DOI: 10.1126/science.1133427

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterBreast, GeneCovBreast

---

GeneSampColon — *Data from the Wood et al. 2007 study: Number of samples for each gene and screen type*

Description

Number of samples analyzed for each gene and screen type from the colon cancer portion of the Wood et al. 2007 and Sjoeblom et al. 2006 studies.

Usage

data(WoodColon07)

Format

The number of samples in the colon cancer portions of the Wood et al. and Sjoeblom et al. studies, broken down by *gene* and *screen* (Discovery and Prevalence). The object is a data frame, with the variables: Gene, Screen, and NrSamp (number of samples). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence").

References

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* DOI:10.1126/science.1145720

Sjoeblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, et al. The consensus coding sequences of human breast and colorectal cancers. *Science.* DOI: 10.1126/science.1133427

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterColon, GeneCovColon

---

GeneSampGBM — *Data from the Parsons et al. 2008 study: Number of samples for each gene and screen type*

Description

Number of samples analyzed for each gene and screen type from the Parsons et al. 2008 glioblastoma multiforme (GBM) study.

Usage

data(ParsonsGBM08)

Format

The number of samples in the GBM study from Parsons et al., broken down by *gene* and *screen* (Discovery and Prevalence). The object is a data frame, with the variables: Gene, Screen, and NrSamp (number of samples). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence").

References

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science.* DOI: 10.1126/science.1164382

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterGBM, GeneCovGBM

GeneSampMB *Data from the Parsons et al. 2011 study: Number of samples for each gene and screen type*

Description

Number of samples analyzed for each gene and screen type from the Parsons et al. 2011 medulloblastoma (MB) study.

Usage

data(ParsonsMB11)

Format

The number of samples in the MB study from Parsons et al., broken down by *gene* and *screen* (Discovery and Prevalence). The object is a data frame, with the variables: Gene, Screen, and NrSamp (number of samples). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence").

References

Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin J, Boca SM, Carter H, Samayoa J, Bettegowda C, et al. The genetic landscape of the childhood cancer medulloblastoma. *Science.* DOI: 10.1126/science.1198056

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterMB, GeneCovMB

GeneSampPancreas *Data from the Jones et al. 2008 study: Number of samples for each gene and screen type*

Description

Number of samples analyzed for each gene and screen type from the Jones et al. 2008 pancreatic cancer study.

Usage

data(JonesPancreas08)

Format

The number of samples in the pancreatic cancer study from Jones et al., broken down by *gene* and *screen* (Discovery and Prevalence). The object is a data frame, with the variables: Gene, Screen, and NrSamp (number of samples). The two possible values for Screen are Disc ("Discovery") and Prev ("Prevalence").

References

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science.* DOI: 10.1126/science.1164368

Parmigiani G, Lin J, Boca S, Sjoeblom T, Kinzler KW, Velculescu VE, Vogelstein B. Statistical methods for the analysis of cancer genome sequencing data. http://www.bepress.com/jhubiostat/paper126/

See Also

cma.scores, cma.fdr, cma.set.stat, cma.set.sim, SimMethodsSims-class, GeneAlterPancreas, GeneCovPancreas

---

SetMethodsSims-class *Class representation for depositing output from simulations.*

Description

Stores results from the sim.data.p.values function.

Objects from the class

New objects can be created using calls of the form new("SetMethodsSims", null.dist, perc.samples, spiked.set.sizes,

Slots

null.dist: Object of class "character". Can be either "Passenger null" or "Permutation null," depending on what method is used to get the null data.

perc.samples: Object of class "numeric". Vector representing the probabilities of the spiked-in gene-sets being altered in any given sample, as percentages; for example perc.samples = c(75, 90) means that these probabilities are 0.75 and 0.90.

spiked.set.sizes: Object of class "numeric". Vector representing the sizes, in genes, of the spiked-in gene-sets; for example, if perc.samples = c(75, 90) and spiked.set.sizes = c(50, 100), there would be 4 spiked-in sets, one with 50 genes and probability of being altered of 0.75 in each sample, one with 50 genes and probability of being altered of 0.90 in each sample, one with 100 genes and probability of being altered of 0.75 in each sample, and one with 100 genes and probability of being altered of 0.90 in each sample.

GeneSets: Object of class "list". The entries of the list correspond to gene-sets and give the genes annotated to them.

cma.alter: Object of class "list". The entries of the list are objects similar to the GeneAlter objects and correspond to the simulation iterations.

cma.cov: Object of class "list". The entries of the list are objects similar to the GeneCov objects and correspond to the simulation iterations.

cma.samp: Object of class "list". The entries of the list are objects similar to the GeneSamp objects and correspond to the simulation iterations.

Scores: Object of class "list". The entries of this list are the output of cma.scores and correspond to the simulation iterations.

results: Object of class "list". The entries of this list are the output of cma.set.stat and correspond to the simulation iterations.

Methods

show signature(object = "SetMethodsSims")

Author(s)

Simina M. Boca, Giovanni Parmigiani.

References

Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene-set analysis for cancer mutation data. *Genome Biology.* DOI: 10.1186/gb-2010-11-11-r112

See Also

GeneCov, GeneSamp, GeneAlter, cma.set.sim, cma.set.stat, combine.sims, extract.sims.method

Index