Package 'SLAPenrich'

March 14, 2017

Type Package

Title Sample-population Level Analysis of Pathway enrichments

Version 0.1

Date 2016-01-15

Author Francesco Iorio

Maintainer Francesco Iorio <iorio@ebi.ac.uk>

Description SLAPenrich implements a statistical framework to identify pathways that tend to be recurrently genomically altered across a the samples of a genomic dataset. Differently from traditional over-recurrence analyses, SLAPenrich does not require the genes belonging to a given pathway to be statistically enriched among those altered in the individual samples. Consistently with the mutual exclusivity principle, and differently from other proposed computational tools, our approach assumes that the functionality of a given pathway might be altered in an individual sample if at least one of its genes is genomically altered. The method accounts for the differences in the mutation rates between samples and the exonic lengths of the genes in the pathways. It statistically tests against the null hypothesis that no associations between a pathway and the disease population under study does exist, assessing analytically the divergence of the total number of samples with alterations in a given pathway from its expectation. Moreover, the used formalism allows SLAPenrich to perform differential enrichment analysis of pathway alterations across different clinically relevant sub-populations of samples. SLAPenrich also includes function to visualise the identified enriched pathway implementing a heuristic sorting to highlight mutual exclusivity trends among the pattern of alterations of the composing genes.

License MIT

Depends HGNChelper, igraph, pheatmap, poibin, stringr, qvalue, biomaRt

R topics documented:

`LUAD_CaseStudy` *Genomic event matrix derived from variants found in a cohort of 188 lung adenocarcinoma patients*

Description

A sparse integer matrix where column names are sample identifiers, and the row names official HUGO gene symbols. A non-zero entry in position i, j of this matrix indicates the number of somatic point mutations harbored by the $j-$sample in the $i-$gene. This matrix summarizes the somatic variants of a cohort of 188 lung adenocarcinoma patients of a public available dataset (see source).

Format

A named integer matrix with HUGO official gene symbols as row names and sample identifiers as column names: i.e. format: num [1:356, 1:163] 1 0 0 0 0 0 0 0 0 0 ...

- attr(*, "dimnames")=List of 2

..$: chr [1:356] "ABL1" "ABL2" "ACVR1B" "ACVR1C" ... ..$: chr [1:163] "16770" "16646" "17741" "16915" ...

Source

The dataset from which this matrix was derived has been studied in Ding et al, 2008. The variant annotations used to assemble this matrix are in Supplementary Table 2 of this publication (available at http://genome.wustl.edu/pub/supplemental/tsp_nature_2008/)

References

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455:1069-75.

See Also

`LUAD_CaseStudy_updatedGS`

LUAD_CaseStudy_clinicalInfos

Clinical informations for a cohort of 188 lung adenocarcinoma patients

Description

A named binary matrix where column names are clinical factors, and the row names sample identifiers. A non-zero entry in position i, j of this matrix indicates the for the $i-$sample the in the $j-$factor is positive. This matrix summarizes some clinical informations for a cohort of 188 lung adenocarcinoma patients of a public available dataset (see source). Particularly the smoking status of the patient-samples (former smoker, current smoker, never smoked, not available) and the bronchioalveolar carcinoma type (mucinous and not-mucinous). This dataset is paired with LUAD_CaseStudy, which summarises the somatic variants found in the same cohort of patients.

Format

A named binary matrix with sample identifiers as row names and clinical factor identifiers as column names: i.e. format: num [1:188, 1:6] 1 0 0 0 0 0 0 0 0 0 ... - attr(*, "dimnames")=List of 2 ..$: chr [1:188] "16530" "16594" "16596" "16600" ... ..$: chr [1:6] "SS_NotAvailable" "SS_Former" "SS_CurrentSmoker" "SS_Never" ...

Details

The sample identifiers on the rows match the column names of the integer matrix in LUAD_CaseStudy.

Source

The dataset from which this matrix was derived has been studied in Ding et al, 2008. The clinical informations used to assemble this matrix are in Supplementary Table 15 of this publication (available at http://genome.wustl.edu/pub/supplemental/tsp_nature_2008/)

References

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455:1069-75.

See Also

LUAD_CaseStudy

LUAD_CaseStudy_updatedGS

Genomic event matrix derived from variants found in a cohort of 188 lung adenocarcinoma patients, with updated gene names.

Description

A sparse integer matrix where column names are sample identifiers, and the row names official HUGO gene symbols. A non-zero entry in position i, j of this matrix indicates the number of somatic point mutations harbored by the $j-$sample in the $i-$gene. This matrix summarizes the somatic variants of a cohort of 188 lung adenocarcinoma patients of a public available dataset (see source). In this matrix the gene names are updated to recent HUGO nomenclatures.

Format

A named integer matrix with HUGO official gene symbols as row names and sample identifiers as column names: i.e. format: num [1:356, 1:163] 1 0 0 0 0 0 0 0 0 0 ...

- attr(*, "dimnames")=List of 2

..$: chr [1:356] "ABL1" "ABL2" "ACVR1B" "ACVR1C" ... ..$: chr [1:163] "16770" "16646" "17741" "16915" ...

Source

The dataset from which this matrix was derived has been studied in Ding et al, 2008. The variant annotations used to assemble this matrix are in Supplementary Table 2 of this publication (available at http://genome.wustl.edu/pub/supplemental/tsp_nature_2008/)

References

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455:1069-75.

See Also

LUAD_CaseStudy

SLAPE.all_genes_exonic_content_block_lengths_ensemble

Genome-wide total exonic block lenghts

Description

This object contains the total exnonic block lengths for all the genes.

Format

A genome-wide named numerical vector containing the total exonic block lengths for all the genes. Names of this vector are offical HUGO gene symbols. Please note that the name of this vector is `GECOBlengths`.

Details

This vector can be updated by using the `SLAPE.compute_gene_exon_content_block_lengths` function.

SLAPE.all_genes_exonic_lengths_ensemble *Genome-wide exone attributes and genomic coordinates*

Description

A data frame containing attributes and chromosomal coordinate of all the gene exones

Format

A data frame with 553609 rows (one for each exone) and the following columns

`ensembl_gene_id` String vector containing ensemble gene identifiers;

`external_gene_name` String vector containing gene names;

`exon_chrom_start` Numerical vector containing chromosomal start positions;

`exon_chrom_end` Numerical vector containing chromosomal end positions

Variable name: `GEA`.

Details

This list has been assembled by using functions from the `biomaRt` package and can be updated using the `SLAPE.update_exon_attributes`.

Note

This object is used by the `SLAPE.compute_gene_exon_content_block_lengths` and `SLAPE.gene_ecbl_length` function to compute the genome-wide total exonic block lengths or the total exonic block length of a given gene, respectively.

See Also

`SLAPE.update_exon_attributes`, `SLAPE.compute_gene_exon_content_block_lengths` `SLAPE.gene_ecbl_length`

SLAPE.analyse *Performing Sample population Level Analysis of Pathway alteration Enrichments (SLAPenrich).*

Usage

```
SLAPE.analyse(EM,
              show_progress = TRUE,
              correctionMethod = "fdr",
              NSAMPLES = 1,
              NGENES = 1,
              accExLength = TRUE,
              BACKGROUNDpopulation = NULL,
              PATH_COLLECTION,
              path_probability = "Bernoulli",
              GeneLenghts,
              RHO=NA)
```

Arguments

EM	A sparse binary matrix, or a sparse matrix with integer non-null entries. In this matrix the column names are sample identifiers, and the row names official HUGO gene symbols. A non-zero entry in position i, j of this matrix indicates the presence of a somatic mutations harbored by the $j-$sample in the $i-$gene. If the matrix contains integer entries then these values are deemed as the number of somatic point mutations harbored by a given sample in a given gene (these values will be considered if the analysis takes into account of the gene exonic lengths, or converted in binary values otherwise).
show_progress	Boolean parameter determining if a progress bar should be visualized during the execution of the analysis (default) or not.
correctionMethod	A string indicating which method should be used to correct pathway enrichment p-values for multiple hypothesis testing. Possible values for this parameter are all the values for the method parameter of the R p.adjust function, plus "qvalue" for the Storey-Tibshirani method (Storey and Tibshirani, 2003).
NSAMPLES	The minimal number of samples of EM in which at least one gene belonging to a given pathway should be mutated in order for that pathway to be tested for alteration enrichments at the sample population level.
NGENES	The minimal number of genes of a given pathway P that must be mutated in at least one sample of the EM in order for that pathway to be tested for alteration enrichments at the sample population level.
accExLength	Boolean parameter determining whether the sample-wise pathway alteration probability model should take into account of the total exonic block lengths of the genes in the pathways and analyzed dataset (default) or not (see details), when using an Hypergeometric model(as specified by the path_probability parameter). This parameter is neglected if a Bernoulli model is used for these probabilities instead of a Hypergeometric model.
BACKGROUNDpopulation	A string vector containing the official HUGO symbols of the gene background population used to compute the sample-wise pathway alteration probabilities (see details). If NULL (default) then the population of all the genes included in at least one pathway of the collection specified in the PATH_COLLECTION parameter
PATH_COLLECTION	A list containing the pathway collection to be tested on the EM dataset for alteration enrichments at the sample population level. Several collections are included in the package as data object. See for example SLAPE.PATHCOM_HUMAN or SLAPE.PATHCOM_HUMAN_nr_i_hu_2016 for a description of the fields required in this list.
path_probability	A string specifying which model should be used to compute the sample-wise pathway alteration probabilities (see details). Possible values for this parameter are "Bernoulli" (default) and "HyperGeom".
GeneLenghts	A named vector containing the genome-wide total exonic block lengths. Names of this vector are official HUGO gene symbol. This is available in the SLAPE.all_genes_exonic_cor object. An updated version of this vector can be assembled using the SLAPE.compute_gene_exon_co function.
RHO	The mutation rate to be used to compute the sample-wise pathway alteration probabilities with the Bernoulli model. If NA (default) then this value is estimated directly by the EM dataset. The value of this parameter is ignored if an

Hypergeometric model is used to compute these probabilities (this is specified by the path_probability parameter).

Details

This is the core analysis function implementing the statistical framework described in (Iorio et al, in preparation).

It takes in input a dataset (the parameter EM) stored in a sparse binary matrix, or a sparse matrix with integer non-null entries. In this matrix the columns correspond to samples, the rows correspond to genes and a non-zero entry indicate the presence of a somatic mutations harbored by a given sample in a given gene. If the matrix contains integer entries then they are deemed as the number of somatic point mutations harbored by a given sample in a given gene (these values will be considered if the analysis takes into account of the gene exonic lengths, or converted in binary values otherwise, see below).

For each pathway gene-set P in the pathway collection specified in the parameter PATH_COLLECTION, and an inputted genomic dataset (in EM), this function computes first of all a vector of probabilities $\pi = \{p_i\}$ quantifying how likely each sample is to harbor at least one somatic mutation in a gene belonging to P, by random chance.

These probabilities are computed (as specified by the parameter path_probability) by default using a Bernoulli model accounting for the total exonic block lengths of all the genes belonging to P, and the expected or observed background mutation rate ρ:

$$p_i = \Pr(X_i \geq 1) = 1 - \exp(-\rho k')$$

where $k' = \sum_{g \in P} l(g)$ is the sum of the exonic block length of all the genes g in pathway P, and ρ the background mutation rate, which can be estimated from the input dataset directly or set to established estimated values (such as $10^{-6}/\text{nucleotide}$) (Greenman et al, 2007; Stratton et al, 2009; Ding et al, 2008, Sjoblom et al, 2006), depending on the parameter RHO being set to its default NA value or not, and X_i is the total number of genes belonging to P that are mutated in sample s_i.

Alternatively, these probabilities can be computed through a complementary cumulative hypergeometric distribution evaluated at $X = 0$, and taking into account of the mutation burden of the samples n_i, the size of P in terms of number of genes k the number of genes in the background-population N

$$p_i = \Pr(X_i \geq 1) = \sum_{x=1}^{k} H(x, N, k, n_i)$$

where H is the probability mass function of a hypergeometric distribution with parameters x, N, k and n_i. In this case the parameter accExLength must be set to FALSE.

To take into account of the total exonic block lengths of the genes in P and the background population both then the parameter accExLength must be set to its default parameter (TRUE). In this case the parameter of H will be:

1. $N' = \sum_{g \in G} l(g)$: the sum of all the exonic content block lengths of all the genes in the background population G;
2. $k' = \sum_{g \in P} l(g)$: the sum of the exonic block lengths of all the genes in pathway P)
3. n_i : the total number of individual alterations involving genes of the pathway P in sample s_i).

In each case the gene background population G can be defined by the user (through the parameter BACKGROUNDpopulation) or assembled pooling together all the genes belonging to at least one pathway of the collection specified in PATH_COLLECTION.

After the vector of probabilities π has been computed, this function computes a pathway alteration score, quantifying the deviance of the number of samples in the datasets harbouring at least a somatic mutation in at least one gene of P, $O(P)$, from its random expectation $E(P)$:

$$A(P) = \log_{10} \frac{O(P)}{E(P)},$$

where $E(P) = \sum_i^m p_i$ and m is the number of samples in the anlyzed dataset.

Finally, this function computes an alteration score significance with a p-value against the null hypothesis: "O(P) is drawn from a Poisson binomial distribution with $\pi = \{p_i\}$ success probabilities". This comes from the observation that if there is no tendency for a given pathway to be recurrently mutated across m samples of the datasets then each of these samples can be considered as the observation of a single Bernoulli trial (in a series of m of them), where the event under consideration in the i-th trial is "At least one gene belonging to P is mutated in the i-th sample". The success probability of this event is given by p_i.

After alteration scores and corresponding significance have been assessed for all the pathways considered in the analysis, resulting p-values are corrected for multiple hypothesis testing with a user-defined method, specified by the parameter `correctionMethod`.

For each of the tested pathways an exclusive coverage score quantifying the tendency of the genes in a given pathway is also computed. This is quantified as the ration between the number of samples where a unique gene belonging to a given pathway is altered divided by the number of samples where at least one gene belonging to given pathway is altered.

Value

A list containing the results of the sample-population level analysis of pathway alterations, containing the following items:

`pathway_id` A numerical vector containing one numerica value for each tested pathway. This is the index of the pathway in the vectors and lists of the pathway collection object specified in the `PATH_COLLECTION` input parameter;

`pathway_EM` The pathway-level alteration matrix: a binary matrix with pathway indexes on the rows and sample identifiers on the columns, and binary entries indicating whether a given pathway is genomically altered in a given sample. The column names are the same of the `EM` input dataset;

`pathway_Probability` The sample-wise pathway alteration probabilities. A numerical matrix with pathway indexes on the rows, sample identifiers on the columns where the generic entry i, j quantifies the likelyhood of the i-pathway to be genomically altered in the j-sample according to the model selected through the input parameter specifications (see Details and Arguments);

`pathway_mus` The expected number of samples in which each pathway is altered: a named vector in which each entry specifies for each pathway (whose index is the entry name) the expected number of samples where it is altered according to the model selected through the input parameter specifications;

`pathway_logOddRatios` The pathway alteration scores: a named vector in which each entry specifies for each pathway (whose index is the entry name) the $\log_{10}$ ratio between the number of samples in which that pathway is altered divided by the expected number of samples in which that pathway is altered (this is computed according to the model selected through the input parameter specifications);

`pathway_pvals` Significance of the pathway alteration scores: a named vector of p-values against the null hypothesis that there is no association between the disease population summarized in the `EM` input dataset and a pathway P. The entry-names of this vector are the pathway indexes;

`pathway_perc_fdr` Significance of the pathway alteration scores after correction for multiple hypotheses testing: a named vector of adjusted p-values (in the form of false discovery rate percentages) against the null hypothesis that there is no association between the disease population summarized in the `EM` input dataset and a pathway P. The entry-names of this vector are the pathway indexes.

`pathwayExclusive_coverage` The mutual exclusive coverage of each pathway: a named vector containing the mutual exclusivity coverage score of each pathway (see details). The names of this vector are the pathway indexes.

`pathway_individualEMs` The individual pathway alteration matrices. A named list with an entry for each tested pathway. Each of this entry is a binary matrix with all the genes belonging to the pathway under consideration on the rows, sample identifiers on the columns and binary entries specifing whether a given gene is altered in a given sample. The column names of these matrix are the same of the input `EM` dataset, whereas the entry-names of this list are the pathway indexes.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

References

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455:1069-75.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446:153-8.

Iorio F, Garcia-Alonso L, Buendia JE, Brammeld J, Wille DR, McDermott U, Saez-Rodriguez J. Identification and visualization of population-level enrichments of pathway alterations in cancer genomes with SLAPenrich. [In preparation]

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006;314:268-74.

Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. Nature Publishing Group; 2009;458:719-24.

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America. 2003;100:9440–5.

Examples

```
#Loading the Genomic Event data object derived from variants annotations
#identified in 188 Lung Adenocarcinoma patients (Ding et al, 2008)
data(LUAD_CaseStudy_updatedGS)

#Loading KEGG pathway gene-set collection data object obtained by post-processing
data(SLAPE.MSigDB_KEGG_hugoUpdated)

#Loading genome-wide total exonic block lengths
```

```
data(SLAPE.all_genes_exonic_content_block_lengths_ensemble)

#Running SLAPenrich analysis
PFPw<-SLAPE.analyse(EM = LUAD_CaseStudy_ugs,PATH_COLLECTION = KEGG_PATH,
                    show_progress = TRUE,
                    NSAMPLES = 1,
                    NGENES = 1,
                    accExLength = TRUE,
                    BACKGROUNDpopulation = rownames(LUAD_CaseStudy_ugs),
                    path_probability = 'Bernoulli',
                    GeneLenghts = GECOBLenghts)

#Show the top-10 SLAPenriched pathway with an exclusive coverage > 80%
unlist(KEGG_PATH$PATHWAY[PFPw$pathway_id[which(PFPw$pathway_exclusiveCoverage>80)]][1:10])
```

`SLAPE.check_and_fix_gs_Dataset`
Check and fix gene symbol names in a genomic event dataset

Description

This function checks that the row names of a genomic dataset are actually updated official gene symbols approved by the HUGO Gene Nomenclature Committee (HGNC) (http://www.genenames.org).

Usage

```
SLAPE.check_and_fix_gs_Dataset(Dataset, updated.hgnc.table)
```

Arguments

`Dataset`	An integer matrix modeling a genomic event dataset where row names are gene symbols and column names are sample identifiers. A non-null entry in the i, j position indicates the presence of a somatic mutation hosted by the i-th gene in the j-th sample (if the matrix is binary) or the number of point mutations hosted by the i-th gene in the j-th sample (if the matrix contains integers).
`updated.hgnc.table`	A data frame containing up-to-date approved HGNC symbols (Approved.Symbol variable) and their synonyms (Symbol variable). This is availalbe in the `SLAPE.hgnc.table` data object or can be created by downloading updated relevant information from the HUGO Gene Nomenclature Committee web-portal (http://www.genenames.org), using the function `SLAPE.update_HGNC_Table`.

Value

The integer matrix provided in input but with row names updated to the most recent approved gene symbol and rows with not found gene synonyms as names removed.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

See Also

SLAPE.update_HGNC_Table, SLAPE.hgnc.table

Examples

```
data(LUAD_CaseStudy)
data(SLAPE.hgnc.table)
updatedGeneSymbolsDataset<-SLAPE.check_and_fix_gs_Dataset(LUAD_CaseStudy,hgnc.table)
```

SLAPE.check_and_fix_path_collection
Check and fix gene symbol names in a collection of pathway gene sets.

Description

This function checks that gene identifiers contained in a pathway gene set collection are actually updated official gene symbols approved by the HUGO Gene Nomenclature Committee (HGNC) (http://www.genenames.org).

Usage

```
SLAPE.check_and_fix_path_collection(pathColl, updated.hgnc.table)
```

Arguments

pathColl	A list containing pathway gene-sets and annotations.
updated.hgnc.table	A data frame containing up-to-date approved HGNC symbols (Approved.Symbol variable) and their synonyms (Symbol variable). This is availalbe in the SLAPE.hgnc.table data object or can be created by downloading updated relevant information from the HUGO Gene Nomenclature Committee web-portal (http://www.genenames.org), using the function SLAPE.update_HGNC_Table.

Value

Pathway collection provided in input but with gene identifiers updated to the most recent approved gene symbols and not approved symbols removed.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

See Also

SLAPE.update_HGNC_Table, SLAPE.hgnc.table

Examples

```
data(SLAPE.PATHCOM_HUMAN)
data(SLAPE.hgnc.table)
updatedGeneSymbolsDataset<-
SLAPE.check_and_fix_path_collection(PATHCOM_HUMAN,hgnc.table)
```

SLAPE.compute_gene_exon_content_block_lengths
Computing genome-wide total exonic block lengths

Usage

```
SLAPE.compute_gene_exon_content_block_lengths(ExonAttributes)
```

Arguments

ExonAttributes	Dataframe containing genomic coordinates of all the exon for all the genes in the genome. This is available in the SLAPE.all_genes_exonic_lengths_ensemble data object.

Value

A genome-wide named vector containing the total exonic block lengths for all the genes. Names of this vector are offical HUGO gene symbols.

Note

The genome-wide exonic block lengths are precomputed and available in the SLAPE.all_genes_exonic_content_bloc data object, this function can be used to update this data object with the most-up-to-date information from Ensemble (using biomaRt functions).

See Also

SLAPE.all_genes_exonic_content_block_lengths_ensemble, SLAPE.all_genes_exonic_lengths_ensemble, SLAPE.compute_gene_exon_content_block_lengths,

SLAPE.core_components *Identification of core-component genes shared by multiple SLAPenriched pathways*

Description

This function identifies group of genes that are recurrently altered in the analysed dataset and that are shared by multiple SLAPenriched pathways, thus are putatively leading the enrichment scores. Additionally this function generates pdf files containing *pathway-membership heatmaps* showing to which pathway each of the genes in the core-component belongs to, together with barplots with alteration frequencies for all the genes in the core-components. Results are also stored in individual Robjects.

Usage

```
SLAPE.core_components(PFP, EM, PATH = "./", fdrth = Inf, exclcovth = 0,
                      PATH_COLLECTION)
```

Arguments

PFP	A list containing the SLAPenrich analysis results outputted by the SLAPE.analyse function while analysing the genomic dataset summarised by the EM genomic event matrix.
EM	A sparse binary matrix, or a sparse matrix with integer non-null entries. In this matrix the column names are sample identifiers, and the row names official HUGO gene symbols. A non-zero entry in position i, j of this matrix indicates the presence of a somatic mutations harbored by the $j-$sample in the $i-$gene. This matrix must be the same that has been inputted to the SLAPE.analyse function to produce the results in the PFP list.
PATH	String specifiying the path of the directory where the pdf file shoud be created.
fdrth	The false discovery rate threshold that should be used to select SLAPenriched pathways from the PFP list (percentage). By default in the PFP with an FDR < 5% are selected.
exclcovth	The mutual exclusivity coverage threshold that should be used to select SLAPenriched pathways from the PFP list (percentage). By default in the PFP with an exclusive coverage > 50% are selected.
PATH_COLLECTION	The pathway collection that has been tested against the EM dataset with the SLAPE.analyse function to produce the PFP list of results.

Details

To identify shared core-components across significantly enriched pathways, the set of enriched pathways and their composing genes are modeled as a bipartite network, in which the first set of nodes contains one element per enriched pathway and the second set contains one element per each of the genes that are included in at least one of the enriched pathways.

In this network, a pathway node is connected with an edge to each of its composing gene nodes.

The resulting bipartite network is then mined for communities, i.e. groups of densely interconnected nodes by using a fast community detection algorithm based on a greedy strategy (Newman, 2004).

The resulting communities are finally saved into pdf files containing heatmaps where nodes in the first set (pathways) are on the columns by columns, nodes in the second set (genes) are on the rows and a not-empty cell in position i,j indicates that the i-th gene belongs to the j-th pathway.

Note

This function makes use of the `fastgreedy.community` function of the `igraph` R package (Csardi and Nepusz, 2006).

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

References

Newman MEJ. Fast algorithm for detecting community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;69:066133.

Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. 2006;1695:38

See Also

SLAPE.analyse

Examples

```
#Loading the Genomic Event data object derived from variants annotations
#identified in 188 Lung Adenocarcinoma patients (Ding et al, 2008)
data(LUAD_CaseStudy_updatedGS)

#Loading KEGG pathway gene-set collection data object
data(SLAPE.MSigDB_KEGG_hugoUpdated)

#Loading genome-wide total exonic block lengths
data(SLAPE.all_genes_exonic_content_block_lengths_ensemble)

#Running SLAPenrich analysis
PFPw<-SLAPE.analyse(EM = LUAD_CaseStudy_ugs,PATH_COLLECTION = KEGG_PATH,
                    show_progress = TRUE,
                    NSAMPLES = 1,
                    NGENES = 1,
                    accExLength = TRUE,
                    BACKGROUNDpopulation = rownames(LUAD_CaseStudy_ugs),
                    path_probability = 'Bernoulli',
                    GeneLenghts = GECOBLenghts)

#Generating pdf files containing heatmaps of the core-components
#of SLAPenriched pathway with an FDR < 5% and exclusive coverage > 80%.
#The pdf files are saved in the current working directory.
SLAPE.core_components(PFP=PFPw,
                      EM=LUAD_CaseStudy_ugs,
                      PATH='./LUAD_coreComponents_',
                      fdrth = 5,
                      exclcovth = 50,
                      PATH_COLLECTION = KEGG_PATH)
```

SLAPE.diff_SLAPE_analysis

Differential SLAPenrichment analysis

Description

This function allows the identification of pathways that are differentially enriched across two sub-populations of samples of the same input dataset. Similarly to differential gene expression analysis, the two sub-populations to be contrasted are defined through a contrast matrix.

Usage

```
SLAPE.diff_SLAPE_analysis(EM, contrastMatrix,
                          positiveCondition, negativeCondition,
                          show_progress = TRUE, display = TRUE,
                          correctionMethod = "fdr",
                          path_probability = "Bernoulli",
                          NSAMPLES = 1, NGENES = 1,
```

```
                        accExLength = TRUE,
                        BACKGROUNDpopulation = NULL,
                        GeneLenghts,
                        PATH_COLLECTION,
                        SLAPE.FDRth = 5, PATH = "./")
```

Arguments

EM	A sparse binary matrix, or a sparse matrix with integer non-null entries. In this matrix the column names are sample identifiers, and the row names official HUGO gene symbols. A non-zero entry in position i, j of this matrix indicates the presence of a somatic mutations harbored by the $j-$sample in the $i-$gene. If the matrix contains integer entries then these values are deemed as the number of somatic point mutations harbored by a given sample in a given gene (these values will be considered if the analysis takes into account of the gene exonic lengths, or converted in binary values otherwise).
contrastMatrix	A binary matrix specifying which sample is included in which sub-population. The row names of this matrix are sample identifiers (and must match the column names of the EM dataset). The column names indicate the sub-population identifiers. A 1 in the position i, j of such a matrix indicates that the $i-$th sample is included in the sub-population corresponding to the $j-$th condition.
positiveCondition	String indicating one of the two sub-populations of samples to be contrasted (the positive population). It should match a column header of the contrastMatrix.
negativeCondition	String indicating one of the two sub-populations of samples be contrasted (the negative population). It should match a column header of the contrastMatrix.
show_progress	Boolean parameter determining if a progress bar should be visualized during the execution of the analysis (default) or not.
display	Boolean parameter determining if result figures should be displayed and saved.
correctionMethod	A string indicating which method should be used to correct pathway enrichment p-values for multiple hypothesis testing, in the two individual SLAPenrich analyses (therefore SLAPE.analyse calls). Possible values for this parameter are all the values for the method parameter of the R p.adjust function, plus "qvalue" for the Storey -Tibshirani method (Storey and Tibshirani, 2003).
path_probability	A string specifying which model should be used to compute the sample-wise pathway alteration probabilities, in the two individual SLAPenrich analyses (therefore SLAPE.analyse calls). Possible values for this parameter are "Bernoulli" (default) and "HyperGeom".
NSAMPLES	The minimal number of samples of EM in which at least one gene belonging to a given pathway should be mutated in order for that pathway to be tested for alteration enrichments at the sample population level, in the two individual SLAPenrich analyses (therefore SLAPE.analyse calls).
NGENES	The minimal number of genes of a given pathway P that must be mutated in at least one sample of the EM in order for that pathway to be tested for alteration enrichments at the sample population level, in the two individual SLAPenrich analyses (therefore SLAPE.analyse calls).

accExLength
: Boolean parameter determining whether the sample-wise pathway alteration probability model in the two individual SLAPenrich analyses (therefore SLAPE.analyse calls) should take into account of the total exonic block lengths of the genes in the pathways and analyzed dataset (default) or not (see details), when using an Hypergeometric model(as specified by the path_probability parameter). This parameter is neglected if a Bernoulli model is used for these probabilities instead of a Hypergeometric model.

BACKGROUNDpopulation
: A string vector containing the official HUGO symbols of the gene background population used to compute the sample-wise pathway alteration probabilities in the two individual SLAPenrich analyses (therefore SLAPE.analyse calls). If NULL (default) then the population of all the genes included in at least one pathway of the collection specified in the PATH_COLLECTION parameter

GeneLenghts
: A named vector containing the genome-wide total exonic block lengths to be used in the two individual SLAPenrich analyses (therefore SLAPE.analyse calls).Names of this vector are official HUGO gene symbol. This is available in the SLAPE.all_genes_exonic_cor object. An updated version of this vector can be assembled using the SLAPE.compute_gene_exon_co function.

PATH_COLLECTION
: A list containing the pathway collection to be tested on the EM dataset for alteration enrichments at the sample population level, in the two individual SLAPenrich analyses (therefore SLAPE.analyse calls). Several collections are included in the package as data object. See for example SLAPE.PATHCOM_HUMAN or SLAPE.PATHCOM_HUMAN_nr or SLAPE.PATHCOM_HUMAN_nr_i_hu_2016 for a description of the fields required in this list.

SLAPE.FDRth
: The false discovery rate threshold to be considered for selecting significant pathways in at least one of the two individual SLAPenrich analyses (therefore SLAPE.analyse calls), and for which differential enrichment scores should be computed.

PATH
: String specifiying the path of the directory where the pdf file containing results shoud be created.

Details

This function first performs two individual SLAPenrichment analyses using the SLAPE.analyse function, on the user-defined two sub-populations of samples, yielding two lists of results. The pathways that are significantly enriched in at least one of the two result lists (according to a user defined false discovery rate (FDR) threshold) are then selected and, for each of them, a differential enrichment score is computed as:

$$\Delta_{A,B}(P) = -\log_{10} \mathrm{FDR}_A(P) + \log_{10} \mathrm{FDR}_B(P),$$

where A and B are the two contrasted sub-populations (respectively, positive and negative) and FDR_A and FDR_B are the two SLAPenrichment FDRs obtained in the two corresponding individual analyses.

Results are visualised at the level of the inputted alterations across the two contrasted population, on the domain of the differentially enriched pathways as well has heatmaps and barplots of the differential enrichment scores. Visualisations are saved into a set of pdf files.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

References

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America. 2003;100:9440–5.

Examples

```
#Loading the Genomic Event data object derived from variants annotations
#identified in 188 Lung Adenocarcinoma patients (Ding et al, 2008)
data(LUAD_CaseStudy_updatedGS)

#Loading KEGG pathway gene-set collection data object
data(SLAPE.PATHCOM_HUMAN_nr_i_hu_2016)

#Loading genome-wide total exonic block lengths
data(SLAPE.all_genes_exonic_content_block_lengths_ensemble)

#Loading clinical infos for 188 Lung Adenocarcinoma patients
#(Ding et al, 2008)
data(LUAD_CaseStudy_clinicalInfos)

#Performing differential SLAPenrichment analysis comparing
#Smokers Vs. Non Smokers. Pdf files with result figures are saved in
#the current working directory
RES<-
   SLAPE.diff_SLAPE_analysis(EM = LUAD_CaseStudy_ugs,contrastMatrix = LUAD_CaseStudy_clinicalInfos,
                             BACKGROUNDpopulation = rownames(LUAD_CaseStudy_ugs),
                             SLAPE.FDRth = 5,display = TRUE,
                             positiveCondition = 'SS_CurrentSmoker',
                             negativeCondition = 'SS_Never',
                             PATH_COLLECTION = PATHCOM_HUMAN,
                             GeneLenghts = GECOBLenghts,
                             PATH = './')

#Showing the top-10 most differentially enriched pathways in the Smokers population
RES[1:10,]
```

SLAPE.gene_ecbl_length

Computing the total exonic block length of a given gene

Usage

```
SLAPE.gene_ecbl_length(ExonAttributes, GENE)
```

Arguments

ExonAttributes	Dataframe containing genomic coordinates of all the exon for all the genes in the genome. This is available in the SLAPE.all_genes_exonic_lengths_ensemble data object.
GENE	A official HUGO gene symbol.

Value

An integer value specifying the total exonic block length of the genes specified in GENE.

Note

All the genome-wide exonic block lengths are precomputed and available in the SLAPE.all_genes_exonic_content_bl data object. This data object can be updated using the SLAPE.update_exon_attributes function.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

See Also

SLAPE.all_genes_exonic_lengths_ensemble, SLAPE.all_genes_exonic_content_block_lengths_ensemble

---

SLAPE.heuristic_mut_ex_sorting

*Run-minimizing permutation of rows and colums of a binary matrix highlighting trends of mutual exclusivity (*mutual exclusivity sorting*)*

Description

The set of somatic mutations of a cancer genomic dataset can be easily modeled as a binary matrix, where columns indicate samples, and rows indicate genes (or vice-versa) and non-zero entries indicate the presence of somatic mutations in given gene/sample combinations. In a binary matrix, a run is a sequence of consecutive non-zero entries. Reordering the rows and the columns of a binary matrix modeling a sub-set of genes and samples of a genomic dataset in a way that the number of runs on its rows, as well as its column-wise marginal totals, are minimized is an effective way to highlight patterns of mutual exclusivity among the runs of different rows, therefore among the groups of samples harbouring mutations in at least one gene of the considered sub-set. Here we call such permutations of rows and columns a *mutual exclusivity sorting* of the inputted binary matrix.

Usage

```
SLAPE.heuristic_mut_ex_sorting(mutPatterns)
```

Arguments

mutPatterns	A generic binary matrix with row and column names.

Details

This function implements a heruistic to perform the *mutual exclusivity sorting* of a binary matrix in order to minimize the number row-wise runs and the column-wise marginal totals. To make the description of the algorithm simpler we will assume that the inputted binary matrix models a genomic dataset with samples on the columns and genes on the rows (as detailed in the description).

In the initial step of the algorithm all the samples and all the genes in the input matrix are declared *uncovered* and an empty vector is initialized: this is the list of *covered* genes G.

Then the algorithm proceeds through a series of iterations until the sets of uncovered genes and uncovered samples are both empty.

In each of these iterations a *best in class gene* is identified. This is the uncovered gene with the maximal *exclusive coverage*, which is defined as the number of uncovered samples in which this gene is mutated minus the number of samples in which at least another uncovered gene is mutated.

Finally, the identified best in class gene is removed from the set of the uncovered genes, it is attached to G, and the set of samples in which it is mutated are removed from the set of the uncovered samples.

After these iterations have been executed, an empty vector of samples L is initialized, all the samples of the dataset are labeled again as uncovered, an empty vector of samples S is initialized.

Then for each of the best in class gene g (in the same order as they appear in G) and until there are uncovered samples, the uncovered samples in which g is mutated are sorted according to the exclusive coverage of g across them (in decreasing ordered), they are labeled as covered samples and attached in the resulting order to L.

To obtain the final mutual-exclusivity sorting of the initial dataset, the corresponding inputted binary matrix is rearranged by permuting the genes/rows in the same order as they appear in G and the samples/columns in the same order as they appear in L.

Value

The inputted binary matrix with rows and columns permuted by the heuristic mutual exclusivity sorting algorithm described in the details.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

Examples

```
 #Generating a random sparse binary matrix
dataset<-matrix(0,10,20,dimnames = list(paste('g',1:10,sep=''),paste('s',1:20,sep='')))
 dataset[sample(sample(200,30))]<-1

 #Visualising a heatmap of the original dataset,
 #with blue cells indicating non-null entries
 pheatmap(dataset,cluster_rows = FALSE,cluster_cols = FALSE,
      legend = FALSE,main='Original dataset',col=c('white','blue'))

 #wait
 cat ("Press [enter] to continue")
 line <- readline()

 #Mutual exclusivity sorting the binary matrix
 me_sorted_dataset<-SLAPE.heuristic_mut_ex_sorting(dataset)

 #Visualising a heatmap of the mutual exclusivity sorted dataset,
 #with blue cells indicating non-null entries
 pheatmap(me_sorted_dataset,cluster_rows = FALSE,cluster_cols = FALSE,
          legend = FALSE,main='Original dataset',col=c('white','blue'))
```

SLAPE.hgnc.table *HUGO gene symbols and their previous synonyms (up to February 2016)*

Usage

```
data("SLAPE.hgnc.table")
```

Format

A data frame with approved HUGO gene symbols in one column `Approved.Symbol` and their previously approved synonyms `Symbol` in another column (up to February 2016). Variable Name: `hgnc.table`.

Note

This object can be updated to a more recent version using the `SLAPE.update_HGNC_Table` function.

Source

HUGO Gene Nomenclature Committee web-portal (http://www.genenames.org)

See Also

`SLAPE.update_HGNC_Table`

Examples

```
data(SLAPE.hgnc.table)
## maybe str(SLAPE.hgnc.table_20160210) ; plot(SLAPE.hgnc.table_20160210) ...
```

SLAPE.MSigDB_KEGG_hugoUpdated

Collection of KEGG pahtway gene-sets from the Molecular Signature Database post-processed to update composing gene name to recent HUGO gene symbols

---

Description

A list containing KEGG pathway gene-sets downloaded from the Molecular Signature Database (Subramanian et al, 2005) and post-processed to update gene names to recent HUGO nomenclature.

Format

A list containing the following items:

`PATHWAY` A string vector in which the i-th entry contains the KEGG pathway name for the i-th pathway gene-set;

`SOURCE` A string vector in which the i-th entry contains the source of the i-th pathway;

`HGNC_SYMBOL` A list in which the i-th element is a string vector containing the official HUGO symbols of the genes belonging to the i-th pathway or multiple merged pathways, updated to recent nomenclature;

`Ngenes` An integer vector in which the i-th element is the number of genes belonging to the i-th pathway;

`backGround` A string vector containing the HUGO symbols of all the genes belonging to at least one pathway;

Please note that the name of this list is `KEGG_PATH`.

Details

All the pathway gene sets contained in this object are updated to recent official HUGO gene nomenclatures, using the informations contained in the SLAPE.hgnc.table data object (which can be itself updated using the dedicated function SLAPE.update_HGNC_Table

Source

This list was assembled from a collection of KEGG pathway gene sets from the Molecular Signature Databse (Subramanian et al, 2005) (http://http://software.broadinstitute.org/gsea/msigdb).

References

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50

SLAPE.PATHCOM_HUMAN *Collection of pahtway gene-sets from Pathway-Commons*

Description

A list containing pathway gene-sets from multiple public resources, downloaded from Pathway-Commons.

Format

A list containing the following items:

PATHWAY A string vector in which the i-th entry contains the Pathway-Commons name of the i-th pathway;

SOURCE A string vector in which the i-th entry contains the Pathway-Commons description of the source of the i-th pathway;

UNIPROTID A list in which the i-th element is a string vector containing the uniprot identifiers of the genes belonging to the i-th pathway;

HGNC_SYMBOL A list in which the i-th element is a string vector containing the official HUGO symbols of the genes belonging to the i-th pathway;

Ngenes An integer vector in which the i-th element is the number of genes belonging to the i-th pathway;

backGround A string vector containing the HUGO symbols of all the genes belonging to at least one pathway

miniSOURCE A string vector in which the i-th entry contains the name of the source of the i-th pathway (panther, humancyc, pid or reactome);

includesTP53 A boolean vector whose i-th is TRUE if the i-th pathway contains TP53.

Please note that the name of this list is PATHCOM_HUMAN.

Source

This list was assembled from the collection of pathway gene sets from the Pathway-Commons data portal (v4-201311) (Cerami et al, 2011) (http://www.pathwaycommons.org/archives/PC2/v4-201311/).

References

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39:D685-90

SLAPE.PATHCOM_HUMAN_nr_i_hu_2014

Collection of pahtway gene-sets from Pathway-Commons (v2014) post-processed for redundancy reduction and to update composing gene name to recent HUGO gene symbols

---

Description

A list containing pathway gene-sets from multiple public resources, downloaded from Pathway-Commons and post-processed to reduce their overlaps (see details) and update gene names.

Format

A list containing the following items:

`PATHWAY` A string vector in which the i-th entry contains the Pathway-Commons name, or multiple Pathway-Commons name joined (separated by '//'), for the i-th pathway gene-set (or gene-set resulting from merging multiple pathways, see details);

`SOURCE` A string vector in which the i-th entry contains the Pathway-Commons description of the source of the i-th pathway, or sources of multiple merged pathways;

`UNIPROTID` A list in which the i-th element is a string vector containing the uniprot identifiers of the genes belonging to the i-th pathway;

`HGNC_SYMBOL` A list in which the i-th element is a string vector containing the official HUGO symbols of the genes belonging to the i-th pathway or multiple merged pathways, differently from `SLAPE.PATHCOM_HUMAN_nr_i_hu` in this object these symbols are updated to recent nomenclature;

`Ngenes` An integer vector in which the i-th element is the number of genes belonging to the i-th pathway;

`backGround` A string vector containing the HUGO symbols of all the genes belonging to at least one pathway;

`miniSOURCE` A string vector in which the i-th entry contains the name of the source of the i-th pathway (panther, humancyc, pid or reactome);

`includesTP53` A boolean vector whose i-th is `TRUE` if the i-th pathway contains TP53.

Please note that the name of this list is `PATHCOM_HUMAN`.

Details

This object was assembled from a collection of pathway gene sets from the Pathway Commons data portal. From this collection gene sets containing less than 4 genes were discarded. Additionally, in order to remove redundancies those gene sets i) corresponding to the same pathway across different resources or ii) with a large overlap (Jaccard index (J) > 0.8, as detailed below) were merged together by intersecting them. The gene sets resulting from these compressions were then added to the collection (with a joint pathway label) and those participating in at least one of these merging were discarded. The final collection resulting from this pre-processing is composed by 1,636 gene sets, for a total amount of 8,056 unique genes included in at least one gene set. Given two gene sets P_1 and P_2 the corresponding $J(P_1, P_2)$ is defined as:

$$J(P_1, P_2) = \frac{|P_1 \cap P_2|}{|P_1 \cup P_2|}$$

.

Additionally, all the pathway gene sets contained in this object are updated to recent official HUGO gene nomenclatures, using the informations contained in the `SLAPE.hgnc.table` data object (which can be itself updated using the dedicated function `SLAPE.update_HGNC_Table`).

Source

This list was assembled from the collection of pathway gene sets from the Pathway-Commons data portal (v4-201311) (Cerami et al, 2011) (http://www.pathwaycommons.org/archives/PC2/v4/).

References

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39:D685-90

See Also

SLAPE.PATHCOM_HUMAN, SLAPE.update_HGNC_Table

`SLAPE.PATHCOM_HUMAN_nr_i_hu_2016`

Collection of pahtway gene-sets from Pathway-Commons (v2016) post-processed for redundancy reduction and to update composing gene name to recent HUGO gene symbols

---

Description

A list containing pathway gene-sets from multiple public resources, downloaded from Pathway-Commons and post-processed to reduce their overlaps (see details) and update gene names.

Format

A list containing the following items:

`PATHWAY` A string vector in which the i-th entry contains the Pathway-Commons name, or multiple Pathway-Commons name joined (separated by '//'), for the i-th pathway gene-set (or gene-set resulting from merging multiple pathways, see details);

SOURCE A string vector in which the i-th entry contains the Pathway-Commons description of the source of the i-th pathway, or sources of multiple merged pathways;

UNIPROTID A list in which the i-th element is a string vector containing the uniprot identifiers of the genes belonging to the i-th pathway;

HGNC_SYMBOL A list in which the i-th element is a string vector containing the official HUGO symbols of the genes belonging to the i-th pathway or multiple merged pathways, differently from SLAPE.PATHCOM_HUMAN_nr_i_hu in this object these symbols are updated to recent nomenclature;

Ngenes An integer vector in which the i-th element is the number of genes belonging to the i-th pathway;

backGround A string vector containing the HUGO symbols of all the genes belonging to at least one pathway;

miniSOURCE A string vector in which the i-th entry contains the name of the source of the i-th pathway (panther, humancyc, pid or reactome);

includesTP53 A boolean vector whose i-th is TRUE if the i-th pathway contains TP53.

Please note that the name of this list is PATHCOM_HUMAN.

Details

This object was assembled from a collection of pathway gene sets from the Pathway Commons data portal. From this collection gene sets containing less than 4 genes were discarded. Additionally, in order to remove redundancies those gene sets i) corresponding to the same pathway across different resources or ii) with a large overlap (Jaccard index $(J) > 0.8$, as detailed below) were merged together by intersecting them. The gene sets resulting from these compressions were then added to the collection (with a joint pathway label) and those participating in at least one of these merging were discarded. The final collection resulting from this pre-processing is composed by 1,636 gene sets, for a total amount of 8,056 unique genes included in at least one gene set. Given two gene sets P_1 and P_2 the corresponding $J(P_1, P_2)$ is defined as:

$$J(P_1, P_2) = \frac{|P_1 \cap P_2|}{|P_1 \cup P_2|}$$

.

Additionally, all the pathway gene sets contained in this object are updated to recent official HUGO gene nomenclatures, using the informations contained in the SLAPE.hgnc.table data object (which can be itself updated using the dedicated function SLAPE.update_HGNC_Table).

Source

This list was assembled from the collection of pathway gene sets from the Pathway-Commons data portal (v4-201311) (Cerami et al, 2011) (http://www.pathwaycommons.org/archives/PC2/v8/).

References

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39:D685-90

See Also

SLAPE.PATHCOM_HUMAN, SLAPE.update_HGNC_Table

SLAPE.pathvis *Generatig a heatmap of the alteration matrix of a SLAPenriched pathway and barplots with statistical scores*

Description

This function generates a pdf file containing a heatmap of the alteration matrix of a SLAPenriched pathway across the samples of the analysed dataset, after permuting rows and columns to highlight trend of mutual exclusivity in the alteration-patterns.

Additionally, it generates, a pdf file with three barplots indicating, respectively: (i) the number of mutated genes across samples; (ii) the probabilities of the pathway under consideration to be altered across samples, together with the expected number of samples with alteration in the pathway under consideration; (iii) The observed pathway alteration status across samples, together with the observed total number of samples with alteration in the pathway under consideration.

Usage

```
SLAPE.pathvis(EM, PFP, Id, prefName = NULL, PATH = "./", PATH_COLLECTION)
```

Arguments

EM	A sparse binary matrix, or a sparse matrix with integer non-null entries. In this matrix the column names are sample identifiers, and the row names official HUGO gene symbols. A non-zero entry in position i, j of this matrix indicates the presence of a somatic mutations harbored by the $j-$sample in the $i-$gene. This matrix must be the same that has been inputted to the SLAPE.analyse function to produce the results in the PFP list.
PFP	A list containing the SLAPenrich analysis results outputted by the SLAPE.analyse function while analysing the genomic dataset summarised by the EM genomic event matrix.
Id	The index of the pathway for which the pdf files should be produced in the vectors/lists of the pathway collection set specified in PATH_COLLECTION.
prefName	A string specifing the prefix to be added to the pdf file name (NULL by default).
PATH	String specifiying the path of the pdf file to be created (including its name).
PATH_COLLECTION	The pathway collection that has been tested against the EM dataset with the SLAPE.analyse function to produce the PFP list of results.

Note

This function makes use of the SLAPE.heuristic_mut_ex_sorting function to realise the mutual exclusivity sorting of the pathway alteration matrix, and it is iteratively used by SLAPE.serialPathVis to generate pdf files with heatmaps and barplots for all the SLAPenriched pathways found in a list of results generated by SLAPE.analyse.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

See Also

SLAPE.analyse, SLAPE.serialPathVis

Examples

```
#Loading the Genomic Event data object derived from variants annotations
#identified in 188 Lung Adenocarcinoma patients (Ding et al, 2008)
data(LUAD_CaseStudy_updatedGS)

#Loading KEGG pathway gene-set collection data object
data(SLAPE.MSigDB_KEGG_hugoUpdated)

#Loading genome-wide total exonic block lengths
data(SLAPE.all_genes_exonic_content_block_lengths_ensemble)

#Running SLAPenrich analysis
PFPw<-SLAPE.analyse(EM = LUAD_CaseStudy_ugs,
                PATH_COLLECTION = KEGG_PATH,
                BACKGROUNDpopulation = rownames(LUAD_CaseStudy_ugs),
                GeneLenghts = GECOBLenghts)

#Generating a pdf file containing a heatmap of the mutual-exclusivity
#sorted pathway alteration matrix, for an SLAPenriched pathway with
#an exclusive coverage > 80%, and barplots with statistical scores.
#The pdf is saved in a file with \code{Example_} as prefix in its name,
#in the current working directory.
SLAPE.pathvis(EM = LUAD_CaseStudy_ugs,PFP = PFPw,
    Id = PFPw$pathway_id[which(PFPw$pathway_exclusiveCoverage>80)[1]],
    prefName = 'Example_',
    PATH = './',PATH_COLLECTION = KEGG_PATH)
```

SLAPE.readDataset *Reading genomic evant dataset*

Description

This function reads a genomic dataset from a csv file and it stores it into an integer matrix. Row names of this matrix are official gene symbols and column names are sample identifiers. A non-zero entry in the i, j position indicates the presence of somatic mutation hosted by the i-th gene in the j-th sample (if the matrix is binary) or the number of point mutations hosted by the i-th gene in the j-th sample (if the matrix contains integers).

Usage

```
SLAPE.readDataset(filename)
```

Arguments

filename	The path of the csv file to be read and stored in the genomic event matrix.

Value

An integer matrix modeling a genomic event dataset. Row names of this matrix are official gene symbols and column names are sample identifiers. A non-zero entry in the i, j position indicates the presence of a somatic mutation hosted by the i-th gene in the j-th sample (if the matrix is binary) or the number of point mutations hosted by the i-th gene in the j-th sample (if the matrix contains integers).

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

`SLAPE.serialPathVis` *Systematic generation of heatmaps of the alteration matrices for SLAPenriched pathways and barplots with statistical scores*

Description

This function generates pdf files containing heatmaps of the alteration matrices for SLAPenriched pathways (with user-defined enrichment FDR and exclusive coverage) across the samples of the analysed dataset, after permuting rows and columns to highlight trend of mutual exclusivity in the alteration-patterns.

Additionally, it generates, a pdf files with three barplots indicating, for each pathway, respectively: (i) the number of mutated genes across samples; (ii) the probabilities of the pathway under consideration to be altered across samples, together with the expected number of samples with alteration in the pathway under consideration; (iii) The observed pathway alteration status across samples, together with the observed total number of samples with alteration in the pathway under consideration.

Usage

```
SLAPE.serialPathVis(EM, PFP, fdrth = 5, exCovTh = 50, PATH = "./", PATH_COLLECTION)
```

Arguments

`EM`	A sparse binary matrix, or a sparse matrix with integer non-null entries. In this matrix the column names are sample identifiers, and the row names official HUGO gene symbols. A non-zero entry in position i, j of this matrix indicates the presence of a somatic mutations harbored by the $j-$sample in the $i-$gene. This matrix must be the same that has been inputted to the `SLAPE.analyse` function to produce the results in the `PFP` list.
`PFP`	A list containing the SLAPenrich analysis results outputted by the `SLAPE.analyse` function while analysing the genomic dataset summarised by the `EM` genomic event matrix.
`fdrth`	The false discovery rate threshold that should be used to select SLAPenriched pathways from the `PFP` list (percentage). By default in the `PFP` with an FDR < 5% are selected.
`exCovTh`	The mutual exclusivity coverage threshold that should be used to select SLAPenriched pathways from the `PFP` list (percentage). By default in the `PFP` with an exclusive coverage > 50% are selected.

PATH String specifiying the path of the directory where the pdf file shoud be created.

PATH_COLLECTION
The pathway collection that has been tested against the EM dataset with the SLAPE.analyse function to produce the PFP list of results.

Note

This function calls iteratively the SLAPE.pathvis function.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

See Also

SLAPE.analyse, SLAPE.pathvis

Examples

```
#Loading the Genomic Event data object derived from variants annotations
#identified in 188 Lung Adenocarcinoma patients (Ding et al, 2008)
data(LUAD_CaseStudy_updatedGS)

#Loading KEGG pathway gene-set collection data object
data(SLAPE.MSigDB_KEGG_hugoUpdated)

#Loading genome-wide total exonic block lengths
data(SLAPE.all_genes_exonic_content_block_lengths_ensemble)

#Running SLAPenrich analysis
PFPw<-SLAPE.analyse(EM = LUAD_CaseStudy_ugs,
                PATH_COLLECTION = KEGG_PATH,
                BACKGROUNDpopulation = rownames(LUAD_CaseStudy_ugs),
                GeneLenghts = GECOBLenghts)

#Generating pdf files containing heatmaps of the mutual-exclusivity
#sorted pathway alteration matrices, for an SLAPenriched pathway with
#an exclusive coverage > 80%, and barplots with statistical scores.
#The pdf files are saved in the current working directory.
SLAPE.serialPathVis(EM = LUAD_CaseStudy_ugs,PFP = PFPw,
                exCovTh = 80,fdrth = 5,
                PATH = './',PATH_COLLECTION = KEGG_PATH)
```

SLAPE.update_exon_attributes

Creating an updated gene exon attributes data object

Usage

```
SLAPE.update_exon_attributes()
```

Value

A dataframe containing updated genomic coordinates of all the exon for all the genes in the genome, from Ensemble.

Note

A dataframe containing genomic coordinates of all the exon for all the genes in the genome, from Ensemble is precomputed and available in the SLAPE.all_genes_exonic_lengths_ensemble data object.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

See Also

SLAPE.all_genes_exonic_content_block_lengths_ensemble, SLAPE.all_genes_exonic_lengths_ensemble, SLAPE.compute_gene_exon_content_block_lengths, SLAPE.gene_ecbl_length

---

SLAPE.update_HGNC_Table

Updating the R data object containing a HUGO approved catalogue of gene symbols and synomyms

---

Description

This function creates a data frame containing up-to-date HUGO Gene Nomenclature Committee (HGNC) approved symbols and their synonyms downloading updated relevant information from the HUGO Gene Nomenclature Committee web-portal (http://www.genenames.org). This table is used by the SLAPE.check_and_fix_gs_Dataset and SLAPE.check_and_fix_path_collection functions to check and update the gene symbol identifiers of a integer genomic event matrix. A precomputed data frame (created on February 2016) is available in the SLAPE.hgnc.table data object.

Usage

```
SLAPE.update_HGNC_Table()
```

Value

A data frame with updated approved HUGO gene symbols in one column and their previously approved synonyms in another column.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

See Also

SLAPE.hgnc.table

SLAPE.check_and_fix_path_collection

SLAPE.check_and_fix_gs_Dataset

SLAPE.write.table *Writing SLAPenrich results in a csv file*

Description

This function takes in input a list of results outputted by the SLAPE.analyse, selects enriched pathways according to user-defined significance and mutual exclusivity coverage thresholds and creates an easy to explore csv file with selected enriched pathways.

Usage

```
SLAPE.write.table(PFP,
                  EM,
                  filename = "",
                  fdrth = Inf,
                  exclcovth = 0,
                  PATH_COLLECTION,
                  GeneLenghts)
```

Arguments

PFP	A list containing the SLAPenrich analysis results outputted by the SLAPE.analyse function while analysing the genomic dataset summarised by the EM genomic event matrix.
EM	A sparse binary matrix, or a sparse matrix with integer non-null entries. In this matrix the column names are sample identifiers, and the row names official HUGO gene symbols. A non-zero entry in position i, j of this matrix indicates the presence of a somatic mutations harbored by the $j-$sample in the $i-$gene. This matrix must be the same that has been inputted to the SLAPE.analyse function to produce the results in the PFP list.
filename	String specifiying the path of the csv file to be created (including its name).
fdrth	The false discovery rate threshold that should be used to select SLAPenriched pathways from the PFP list (percentage). By default all the pathway included in the PFP list are selected.
exclcovth	The mutual exclusivity coverage threshold that should be used to select SLAPenriched pathways from the PFP list (percentage). By default all the pathway included in the PFP list are selected.
PATH_COLLECTION	The pathway collection that has been tested against the EM dataset with the SLAPE.analyse function to produce the PFP list of results.
GeneLenghts	The named vector containing the genome-wide total exonic block lengths that has been used by the SLAPE.analyse function to produce the PFP list of results.

Author(s)

Francesco Iorio - iorio@ebi.ac.uk

See Also

SLAPE.analyse

Examples

```
#Loading the Genomic Event data object derived from variants annotations
#identified in 188 Lung Adenocarcinoma patients (Ding et al, 2008)
data(LUAD_CaseStudy_updatedGS)

#Loading KEGG pathway gene-set collection data object
data(SLAPE.MSigDB_KEGG_hugoUpdated)

#Loading genome-wide total exonic block lengths
data(SLAPE.all_genes_exonic_content_block_lengths_ensemble)

#Running SLAPenrich analysis
PFPw<-SLAPE.analyse(EM = LUAD_CaseStudy_ugs,
              PATH_COLLECTION = KEGG_PATH,
              BACKGROUNDpopulation = rownames(LUAD_CaseStudy_ugs),
              GeneLenghts = GECOBLenghts)

#Selecting pathway enriched at a 5% FDR,
#that have a 50% mutual exclusivity coverage and writing them
#in a csv file
SLAPE.write.table(PFP = PFPw,
                  EM = LUAD_CaseStudy_ugs,
                  filename = "./LungDS_KEGG_enrichments.csv",
                  fdrth=5, exclcovth = 50, PATH_COLLECTION = KEGG_PATH,
                  GeneLenghts = GECOBLenghts)
```

Index