Package 'SLAPenrich'

December 8, 2016

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

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.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(PATHCHOM_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.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.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.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.20140608_PATHCOM_HUMAN_nonredundant_intersection_hugoUpdated 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_20160210` 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.20140608_PATHCOM_HUMAN_nonredundant_intersection_hugoUpdated` 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_20160210` 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.20140608_PATHCOM_HUMAN`, `SLAPE.update_HGNC_Table`

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.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

Index