Package 'CELLector'

March 29, 2018

Type Package

Title Genomics guided selection of cancer cell lines

Version 0.2.0

Author Hanna Najgebauer and Francesco Iorio

Maintainer Francesco Iorio <francesco.iorio@sanger.ac.uk>

Description Functions to select the most relevant cancer cell lines to be included in a new in-vitro study, in a genomic-guided fashion. CELLector combines methods from graph theory and market basket analysis; it leverages tumour genomics data to explore, rank, and select optimal cell line models in a user-friendly way, enabling scientists to make appropriate and informed choices about model inclusion/exclusion in retrospective analyses and future studies. Additionally, it allows the selection of models within user-defined contexts, for example, by focusing on genomic alterations occurring in biological pathways of interest or considering only predetermined sub-cohorts of cancer patients. Finally, CELLector identifies combinations of molecular alterations underlying disease subtypes currently lacking representative cell lines, providing guidance for the future development of new cancer models.

License MIT

Encoding UTF-8

LazyData true

Depends arules, dplyr, stringr, data.tree, sunburstR, igraph, collapsibleTree

R topics documented:

CELLector.Build_Search_Space 2
CELLector.CellLine.BEMs 3
CELLector.CFEs 5
CELLector.CFEs.CNAid_decode 6
CELLector.CFEs.CNAid_mapping 7
CELLector.cna_look_up 8
CELLector.HCCancerDrivers 9
CELLector.mostSupported_CFEs 9
CELLector.MSIstatus 10
CELLector.Pathway_CFEs 11
CELLector.PrimTum.BEMs 12

Index **13**

CELLector.Build_Search_Space

CELLector search space construction

Description

This function assembles a user defined CELLector search space analysing genomic data from a larg cohort of cancer patients (spcified in input). It identifies recurrent subtypes with matched genomic signatures (as combination of cancer functional events (CFEs), defined in [1]), linking them into a hierarchical structure shaped as a a binary three with a corresponding navigable table, as detailed in [2].

Usage

```
CELLector.Build_Search_Space(ctumours,
                             cancerType,
                             minlen = 1,
                             verbose = TRUE,
                             mutOnly = FALSE,
                             cnaOnly = FALSE,
                             minGlobSupp = 0.01,
                             FeatureToExclude = NULL,
                             pathway_CFEs = NULL,
                             pathwayFocused = NULL,
                             subCohortDefinition = NULL,
                             NegativeDefinition = FALSE,
                             cnaIdMap,
                             cnaIdDecode,
                             cdg)
```

Arguments

ctumours	A binary event matrix (BEM) modeling a cohort of cancer patients. With cancer functional events (CFEs) on the columns and sample identifers on the rows. See CELLector.PrimTum.BEMs for further details
cancerType	The cancer type under consideration (spcified via a TCGA label): currently available types = *BLCA, BRCA, COREAD, GBM, HNSC, KIRC, LAML, LGG, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA, UCEC*
minlen	The minimal length of the genomic signatures (of how many indivudal CFE it needs to be composed) in order to be considered in the analysis (1 by default)
verbose	A boolean argument specifying whether step-by-step information on the algorithm progression should be displayed run-time
mutOnly	A boolean argument specifying wether only CFE involving somatic mutations should be considered in the analysis. If the cnaOnly argument is equal to TRUE then this must be FALSE (default value)
cnaOnly	A boolean argument specifying wether only CFE involving copy number alterations (CNAs) of chromosomal segments that are recurrently CN altered should be considered in the analysis. If the mutOnly argument is equal to TRUE then this must be FALSE (default value)

`minGlobSupp` Minimal size of the outpputted subtypes, as ratio with respect to the whole cohort of patients

`FeatureToExclude`
A string (or a vector of strings) with identifiers of CFEs that should be ignored

`pathway_CFEs` TO BE CONTINUED

`pathwayFocused`

`subCohortDefinition`

`NegativeDefinition`

`cnaIdMap`

`cnaIdDecode`

`cdg`

Details

Starting from an initial cohort of patients affected by a given cancer type and modeled by the inputted binary event matrix (BEM), the most frequent alteration or set of molecular alterations (depending on the `minlen` argument) with the largest support (the subpopulation of patients in which these alterations occur simultaneously) is identified using the `eclat` function of the `arules` R package.

Based on this, the cohort of patients is split into two subpopulations depending on the collective presence or absence of the identified alterations. This process is then executed recursively on the two resulting subpopulations and it continues until all the alteration sets (with a support of minimal size, as specified in the `minGlobSupp` argument) are identified.

Each of the alterations sets identified through this recursive process is stored in a tree node. Linking nodes identified in adjacent recursions yields a binary tree: the CELLector search space. Each individual path (from the root to a node) of this tree defines a rule (signature), represented as a logic AND of multiple terms (which can be also negated), one per each node in the path. If the genome of a given patient in the analysed cohort satisfies the rule then it is contained in the subpopulation represented by the terminal node of that path. Collectively, all the paths in the search space provide a representation of the spectrum of combinations of molecular alterations observed in a given cancer type, and their clinical prevalence in the analysed patient population.

See Also

`CELLector.PrimTum.BEMs`,

`CELLector.CellLine.BEMs`

Cell Lines' Binary Event Matrices

Description

A list containing 16 data frames (one for cancer type), identified through TCGA labels. Each of these data frames contains cell lines' *binary event matrices* (BEMs) with the status (presence/absence) of *cancer functional events* (CFEs) as defined in [1].

Usage

```
data(CELLector.CellLine.BEMs)
```

Format

A named list of data frames (with TCGA cancer type labels as names). Each of these data frames contains two columns with COSMIC [2] identiefiers and names of cell lines (one per row), respectively, and then binary entries indicating the status of each CFEs (one per column) across cell lines.

Details

BEMs for cell lines from the Genomics of Drug Sensitivity in Cancer (GDSC1000, [1]) panel. Data is available for cell lines matching one among 16 different TCGA cancer types: *BLCA, BRCA, COREAD, GBM, HNSC, KIRC, LAML, LGG, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA, UCEC*.

A decoding table for these labels is available at Each data frame contains cell lines on the rows (with COSMIC identifiers and names, respectively on first and second column) and then a binary matrix with a CFE per column and entries indicating the presence/absence of a given CFE in a given cell line.

Gene symbols as column names indicate high confidence cancer driver genes and the entries in the corresponding columns indicate the presence/absence of somatic mutations. Column names with *cna* as prefix indicate chromosomal segments that are recurrently copy number altered in cancer (RACSs, defined in [1]). A list with all the considered CFEs is available in the `CELLector.CFEs` data object. A decoding table for the RACSs is available in the `CELLector.CFEs.CNAid_decode`, with the mapping realised by the values in the `CNA_identifier` column.

Please note that the same RACS identifier across multiple cancer types might indicate different chromosomal regions, therefore in order to be decode it should be considered jointly with the TCGA label of the data frame it has been extracted from.

References

[1] Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754 (2016).

[2] Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43, D805–11 (2015).

See Also

`CELLector.PrimTum.BEMs`, `CELLector.CFEs`, `CELLector.CFEs.CNAid_decode`

Examples

```
data(CELLector.CellLine.BEMs)
CELLector.CellLine.BEMs$COREAD[1:10,
                               c("COSMIC_identifier","CellLine",'BRAF','KRAS','cna27')]
```

CELLector.CFEs *Cancer Functional Events*

Description

Identifiers of cancer functional events (CFEs, i.e. somatic mutations in high confidence cancer driver genes or chromosomal regions of recurrent copy number amplification/deletion) from [1], which are also present in the binary event matrices of the cell lines and the primary tumours considered in this version of CELLector.

Usage

```
data("CELLector.CFEs")
```

Format

A vector of strings with one entry per identifier.

Details

Gene symbols indicate somatic mutations in igh confidence cancer driver genes and entries with *cna* prefix indicate chromosomal segments that are recurrently copy number altered in cancer (RACSs), both defined in [1].

A decoding table for the RACSs is available in the CELLector.CFEs.CNAid_decode, with the mapping realised by the values in the CNA_identifier column.

Please note that the same RACS identifier across multiple cancer types might indicate different chromosomal regions, therefore in order to be decode it should be considered jointly with the TCGA label of the data frame it has been extracted from.

References

[1] Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754 (2016).

See Also

CELLector.PrimTum.BEMs, CELLector.CellLine.BEMs,

CELLector.CFEs, CELLector.CFEs.CNAid_decode

Examples

```
data(CELLector.CFEs)
head(CELLector.CFEs)
```

CELLector.CFEs.CNAid_decode *Decoding table for copy number alteration cancer functional events*

Description

A table with identifiers of cancer functional events (CFEs) involving chromosomal regions of recurrent copy number alterations (RACSs, as defined by [1], i.e. identified throgh ADMIRE [2]) and their annotation.

Usage

```
data("CELLector.CFEs.CNAid_decode")
```

Format

A data frame with 731 observations (one for each CNA CFE) on the following 15 variables.

`Identifier` The RACS identifer, as defined in [1]

`CancerType` A TCGA label indicating the cancer type where the RACS has been identified (via ADMIRE [2])

`Recurrent` A string specifying whether the RACS under consideration is frequently amplified (value = `Amplification`) or deleted) (value = `deleted`)

`chr` Chromosome number of the RACS

`start` Starting position of the RACS

`stop` Ending position of the RACS

`nGenes` Number of protein coding genes included in the RACS

`locus` Genomic locus of the RACS

`ContainedGenes` A string with comma separated symbols of the genes included in the RACS

`CNA_Identifier` A string containing the identifer of the RACS as it appears in the Binary Event Matrix (BEM) of the cancer type specified in the `CancerType` field included in the `CELLector.CellLine.BEMs` and the `CELLector.PrimTum.BEMs` data objects

Details

This data frame contains a comprehensive annotation of the CFEs involving RACSs appearing in the BEMs of cell lines and primary tumours, contained in the `CELLector.CellLine.BEMs` and the `CELLector.PrimTum.BEMs` data objects. Please note that the same RACS identifier across multiple cancer types might indicate different chromosomal regions, therefore in order to be decode it should be considered jointly with the TCGA label of the data frame it has been extracted from.

This table is used by the `CELLector.cna_look_up` function to decode the identifier of CFE involving a RACS.

References

[1] Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754 (2016).

[2] van Dyk, E., Reinders, M. J. T. & Wessels, L. F. A. A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control. Nucleic Acids Res. 41, e100 (2013).

See Also

CELLector.CellLine.BEMs, CELLector.PrimTum.BEMs, CELLector.cna_look_up

Examples

```
data(CELLector.CFEs.CNAid_decode)
haed(CELLector.CFEs.CNAid_decode)

data(CELLector.CellLine.BEMs)
colnames(CELLector.CellLine.BEMs$COREAD)[8]

CELLector.cna_look_up(cna_ID = colnames(CELLector.CellLine.BEMs$COREAD)[8],
                      TCGALabel = 'COREAD',
                      cnaId_decode = CELLector.CFEs.CNAid_decode)
```

CELLector.CFEs.CNAid_mapping

Pan-Cancer/Cancer-Specific RACSs map.

Description

A data frame mapping chromosomal regions of recurrent copy number amplifications/deletions in cancer (RACSs, as defined in [1]) identified via ADMIRE [2] in the context of specific cancer types to PanCancer RACSs.

Usage

```
data("CELLector.CFEs.CNAid_mapping")
```

Format

A data frame with 425 observations (one for each PanCancer RACS) and a column for each of 27 different cancer types (specified by TCGA labels). The entry in position *i,j* contains the identifier of the *i*th PanCancer RACS in the context of the *j*th cancer type (where available).

References

[1] Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754 (2016).

[2] van Dyk, E., Reinders, M. J. T. & Wessels, L. F. A. A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control. Nucleic Acids Res. 41, e100 (2013).

Examples

```
data(CELLector.CFEs.CNAid_mapping)
head(CELLector.CFEs.CNAid_mapping)
```

CELLector.cna_look_up *Decoding identifers of chromosomal regions of recurrent Copy Number Alterations*

Description

This functions shows the annotation for a chromosomal region of recurrent copy number alterations (RACS) as defined in [1].

Usage

```
CELLector.cna_look_up(cna_ID, cnaId_decode, TCGALabel)
```

Arguments

cna_ID	A string containin the RACS identifier. Full list available in the CELLector.CFEs object.
cnaId_decode	A data frame containing the RACSs' annotation, available in the CELLector.CFEs.CNAid_decode object
TCGALabel	A TCGA label indicating the cancer type under consideration: *BLCA, BRCA, COREAD, GBM, HNSC, KIRC, LAML, LGG, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA, UCEC* available in this version.

Value

A data frame with a single line containing the annotation of the RACS indicated in input.

Author(s)

Hanna Najgebauer and Francesco Iorio

References

[1] Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754 (2016).

See Also

CELLector.CFEs,

CELLector.CFEs.CNAid_decode

Examples

```
CELLector.cna_look_up(cna_ID='cna26',
                      cnaId_decode = CELLector.CFEs.CNAid_decode,
                      TCGALabel = 'BRCA')
```

CELLector.HCCancerDrivers

High Confidence Cancer Driver genes

Description

A list of high confidence cancer driver genes from [1]

Usage

```
data("CELLector.HCCancerDrivers")
```

Format

A vector of strings with one entry per cancer gene.

References

[1] Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754 (2016).

Examples

```
data(CELLector.HCCancerDrivers)
## maybe str(CELLector.HCCancerDrivers) ; plot(CELLector.HCCancerDrivers) ...
```

CELLector.mostSupported_CFEs

Most recurrent combinations of Cancer Functional Events

Description

This function identifies the most frequent combination of cancer functional events (CFEs) in a large cohort of cancer patients.

Usage

```
CELLector.mostSupported_CFEs(transactions,
                             minSupport = 0.05,
                             minlen = 1,
                             maxLen = 10)
```

Arguments

`transactions`	A named binary matrix with CFEs on the rows, samples on the columns and entries specifying the presence/absence of a given CFE in a given sample: the *transactions* object.
`minSupport`	The minimal support that a combination of CFEs must have, i.e. the minimal ratio of samples in which the CFEs must be observed simoultanously, in order to be considered in the analysis.

`minlen` The minimal length of a combination of CFEs (of how many indivudal CFE it needs to be composed) in order to be considered in the analysis (1 by default).

`maxLen` The maximal length of a combination of CFEs (the maximal number of indivudal CFEs) in order to be considered in the analysis (10 by default).

Details

This function uses the *eclat* function from the R package *arules*.

Value

A list with the following fields:

`MSIS` A string or a vector of strings (depending on the argument `minlen`) specifying the CFE (or the combination of individual CFEs) that is the most frequently observed (simultaneously across the samples in input)

`SUPPORT` The ratio of samples where the combination of CFEs in `MSIS` is obaserved on the total number of samples, i.e. number of columns in the `transactions` argument

`absSUPPORT` The number of samples where the combination of CFEs in `MSIS` is obaserved

`supportingSamples` The identifiers of the samples supporting `MSIS`, i.e. the names of the columns of *transactions*, in which the entries corresponding to `MSIS` rows are equal to 1.

Author(s)

Hanna Najgebauer and Francesco Iorio

References

Najgebauer et al., CELLector: Genomics Guided Selection of Cancer in vitro Models. doi:10.1101/275032

Examples

```
data(CELLector.PrimTum.BEMs)
CELLector.mostSupported_CFEs(transactions = t(CELLector.PrimTum.BEMs$COREAD),
                             minlen = 2)
```

CELLector.MSIstatus *Cell lines' Microsatellite status*

Description

The microsatellite status of the cell lines in the CELLector collection, which can be stable (MSI-S), lowly instable (MSI-L), or highly instable (MSI-H) from [1]

Usage

```
data("CELLector.MSIstatus")
```

Format

A named vector of string with one entry per cell lines (with COSMIC [2] identifiers as names) specigying the MSI status of each cell line as detailed in the description above.

References

[1] Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754 (2016).

[2] Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43, D805–11 (2015

Examples

```
data(CELLector.MSIstatus)
head(CELLector.MSIstatus)
```

CELLector.Pathway_CFEs

Cancer functional events in biological pathways

Description

Lists of cancer functional events (CFEs) from [1] involving genes in 14 key cancer biological pathways

Usage

```
data("CELLector.Pathway_CFEs")
```

Format

Named list of string vectors, whose elements are CFEs involving genes in a fixed biological pathway.

References

[1] Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754 (2016).

Examples

```
data(CELLector.Pathway_CFEs)
CELLector.Pathway_CFEs$`RAS-RAF-MEK-ERK / JNK signaling`
```

`CELLector.PrimTum.BEMs`

Primary Tumours' Binary Event Matrices

Description

A list containing 16 data frames (one for cancer type), identified through TCGA labels. Each of these data frames contains primary tumours' *binary event matrices* (BEMs) with the status (presence/absence) of *cancer functional events* (CFEs) as defined in [1].

Usage

```
data("CELLector.PrimTum.BEMs")
```

Format

A named list of binary matrices (with TCGA cancer type labels as names). The entries of each of these matrices indicate the status (Present/Absent) of each CFE (one per row) across primary tumors samples (one per column).

Details

BEMs of primary tumours from the Genomics of Drug Sensitivity in Cancer (GDSC1000, [1]) study. Data is available for 16 different TCGA cancer types: *BLCA, BRCA, COREAD, GBM, HNSC, KIRC, LAML, LGG, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA, UCEC*.

A decoding table for these labels is available at Each data frame contains primary tumour samples on the columns and CFEs on the rows, with entries indicating the presence/absence of a given CFE in a given primary tumour sample.

Gene symbols as row names indicate high confidence cancer driver genes and the entries in the corresponding rows indicate the presence/absence of somatic mutations. Row names with *cna* as prefix indicate chromosomal segments that are recurrently copy number altered in cancer (RACSs, defined in [1]). A list with all the considered CFEs is available in the `CELLector.CFEs` data object. A decoding table for the RACSs is available in the `CELLector.CFEs.CNAid_decode`, with the mapping realised by the values in the `CNA_identifier` column.

Please note that the same RACS identifier across multiple cancer types might indicate different chromosomal regions, therefore in order to be decode it should be considered jointly with the TCGA label of the data frame it has been extracted from.

References

[1] Iorio, F. et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740–754 (2016).

See Also

`CELLector.CellLine.BEMs`, `CELLector.CFEs`, `CELLector.CFEs.CNAid_decode`

Examples

```
data(CELLector.PrimTum.BEMs)
CELLector.PrimTum.BEMs$COREAD[c('BRAF','KRAS','cna27'),1:10]
```

Index

∗Topic **\textasciitildekwd1**
CELLector.Build_Search_Space, 2
∗Topic **\textasciitildekwd2**
CELLector.Build_Search_Space, 2
∗Topic **analysis**
CELLector.mostSupported_CFEs, 9
∗Topic **annotation/decoding**
CELLector.cna_look_up, 8
∗Topic **datasets**
CELLector.CellLine.BEMs, 3
CELLector.CFEs, 5
CELLector.CFEs.CNAid_decode, 6
CELLector.CFEs.CNAid_mapping, 7
CELLector.HCCancerDrivers, 9
CELLector.MSIstatus, 10
CELLector.Pathway_CFEs, 11
CELLector.PrimTum.BEMs, 12

CELLector.Build_Search_Space, 2
CELLector.CellLine.BEMs, 3, *5*, *12*
CELLector.CFEs, *4, 5*, 5, *8*, *12*
CELLector.CFEs.CNAid_decode, *4, 5*, 6, *8*, *12*
CELLector.CFEs.CNAid_mapping, 7
CELLector.cna_look_up, *6*, 8
CELLector.HCCancerDrivers, 9
CELLector.mostSupported_CFEs, 9
CELLector.MSIstatus, 10
CELLector.Pathway_CFEs, 11
CELLector.PrimTum.BEMs, *2–5*, 12