Package 'CRISPRcleanR'

December 9, 2017

Type Package

Title Unsupervised correction of gene independent cell responses to CRISPR-cas9 targeting

Version 0.1

Date 2017-09-11

Author Francesco Iorio

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

License GPL-2

Description CRISPRcleanR is an R package for identifying and correcting gene independent responses to CRISPRcas9 targeting, in genome-wide pooled sgRNA dropout screens. CRISPRcleanR uses an unsupervised approach based on the segmentation of single-guide RNA (sgRNA) fold change values across the genome, without making any assumption on the copy number status of the targeted genes. CRISPRcleanR reports sgRNA fold changes and normalised sgRNA read counts, and is therefore compatible with downstream analysis tools, and works with multiple sgRNA libraries.

Depends stringr, DNAcopy, pROC, stats, utils, grDevices, graphics, pracma

RoxygenNote 6.0.1

R topics documented:

`BAGEL_essential` *Reference Core fitness essential genes*

Description

A list of reference core fitness essential genes assembled from multiple RNAi studies used as classification template by the BAGEL algorithm to call gene depletion significance [1].

Usage

```
data(BAGEL_essential)
```

Format

A vector of strings containing HGNC symbols of reference core fitness essential genes.

References

[1] BAGEL: a computational framework for identifying essential genes from pooled library screens. Traver Hart and Jason Moffat. BMC Bioinformatics, 2016 vol. 17 p. 164.

See Also

`BAGEL_nonEssential`

Examples

```
data(BAGEL_essential)
head(BAGEL_essential)
```

BAGEL_nonEssential *Reference set of non essential genes*

Description

A list of reference non essential genes assembled from multiple RNAi studies used as classification template by the BAGEL algorithm to call gene depletion significance [1].

Usage

```
data(BAGEL_nonEssential)
```

Format

A vector of strings containing HGNC symbols of reference non essential genes.

References

[1] BAGEL: a computational framework for identifying essential genes from pooled library screens. Traver Hart and Jason Moffat. BMC Bioinformatics, 2016 vol. 17 p. 164.

See Also

BAGEL_essential

Examples

```
data(BAGEL_nonEssential)
head(BAGEL_nonEssential)
```

CCLE.gisticCNA *Genome-wide copy number data for 13 human cancer cell lines.*

Description

Genome-wide Gistic [1] scores quantifying copy number status across a subset of the cell lines in CL.subset that are used to assess CRISPRcleaneR results in [2].

Usage

```
data(CCLE.gisticCNA)
```

Format

A data frame with one observations per gene across 13 variables (one per cell line). Row names indicate HGNC gene symbols and column names indicate cell line COSMIC identifiers [3].

Source

This data frame has been derived from the tsv file downloadable at
http://www.cbioportal.org/study?id=cellline_ccle_broad#summary.
This has been obtained by processing Affymetrix SNP array data in the Cancer Cell Line Encyclopaedia [4] repository
(https://portals.broadinstitute.org/ccle_legacy/data/)

References

[1] Mermel CH, Schumacher SE, Hill B, et al. *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol. 2011;12(4):R41. doi: 10.1186/gb-2011-12-4-r41.

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting.
http://doi.org/10.1101/228189

[2] Forbes SA, Beare D, Boutselakis H, et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D777-D783,

[3] Barretina J, Caponigro G, Stransky N, et al. *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature. 2012 Mar 28;483(7391):603-7. doi: 10.1038/nature11003. Erratum in: Nature. 2012 Dec 13;492(7428):290.

Examples

```
data(CCLE.gisticCNA)
head(CCLE.gisticCNA)
```

`ccr.cleanChrm` *Identification and correction of genomic regions of equal log fold changes involving sgRNAs targeting a minimal number of genes within a given chromosome.*

Description

This function applies a circular binary segmentation algorithm [1, 2] to genomic-sorted log fold changes of all the sgRNAs targeting genes on the same chromosome. This procedure yields a sets of genomic regions of estimated equal sgRNAs' log fold changes, significantly differing on average from adjacent regions. If some of these regions fulfill certain criteria (detailed below) then they are deemed as responding to CRISPR-Cas9 targeting in a gene independent manner, i.e. they might be biased by local feature of the DNA) and their pattern of log fold changes is mean centered [3].

Usage

```
ccr.cleanChrm(gwSortedFCs,CHR,display=TRUE,label='',
              saveTO=NULL,min.ngenes=3,ignoredGenes=NULL)
```

Arguments

gwSortedFCs	A data frame containing genome-wide genomic-sorted sgRNAs' log fold changes. This data frame must include one named row per each sgRNAs and the following columns/headers: • CHR: the chromosome of the gene targeted by the sgRNA under consideration; • startp: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration; • endp: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration; • genes: the HGNC symbol of the gene targeted by the sgRNA under consideration; • avgFC: the log fold change of the sgRNA under consideration averaged across replicates; • BP: the genomic coordinate of the sgRNA defined as STARTpos+(ENDpos-STARTpos)/2. This can be generated using the ccr.logFCs2chromPos function, starting from a data frame containing sgRNAs' log fold changes generated by the ccr.NormfoldChanges function from raw sgRNAs' counts.
CHR	Numerical value indicating the chromosome to analyse and correct. X and Y chromosome must be indicated with 23 and 24, respectively.
display	A logical value indicating whether genomic plots showing the results of the biased regions' identification and their log fold change correction should be generated or not.
label	A string indicating the experiment name, used in the main title of the plots and for the name of the folder where results are saved.
saveTO	If different from NULL then it will contain the path where pdf files with then genomic plots showing the results of the biased regions' identification (and their log fold change correction) will be saved (within a folder named as defined in the label parameter).
min.ngenes	A numerical value (>0) specifying the minimal number of different genes that the set of sgRNAs within a region of estimated equal log fold changes should target in order for that region to be corrected, i.e. mean centered.
ignoredGenes	A vector of strings containing HGNC symbols of genes that should not be considered when computing the minimal number of different genes targeted by the sgRNAs in the same identified region of estimated equal log fold changes. This vector could contain, for example, a priori known essential genes. This parameter should be set to NULL for a completely unsupervised correction.

Value

A list containing two data frames. The first one (correctedFCs) contains a named row per each sgRNA and the following columns/header:

- CHR: the chromosome of the gene targeted by the sgRNA under consideration;
- startp: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;
- endp: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;

- genes: the HGNC symbol of the gene targeted by the sgRNA under consideration;
- avgFC: the log fold change of the sgRNA averaged across replicates;
- correction: the type of correction: 1 = increased, -1 = decreased;
- correctedFC: the corrected log fold change of the sgRNA

.
The second one (regions) contains the identified region of estimated equal log fold changes (one region per row) and the following columns/headers:

- CHR: the chromosome of the region under consideration;
- startp: the genomic coordinate of the starting position of the region under consideration;
- endp: the genomic coordinate of the ending position of the region under consideration;
- n.sgRNAs: the number of sgRNAs targeting sequences in the region under consideration;
- avg.logFC: the average log fold change of the sgRNAs targeting the region;
- guideIdx: the indexes range of the sgRNAs targeting the region under consideration as they appear in the gwSortedF Cs provided in input.

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

References

[1] Olshen, A. B., Venkatraman, E. S., Lucito, R., Wigler, M. (2004). *Circular binary segmentation for the analysis of array-based DNA copy number data*. Biostatistics 5: 557-572.

[2] Venkatraman, E. S., Olshen, A. B. (2007). *A faster circular binary segmentation algorithm for the analysis of array CGH data*. Bioinformatics 23: 657-63.

[3] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

See Also

ccr.logFCs2chromPos, ccr.NormfoldChanges

Examples

```
data(KY_Library_v1.0)

fn<-paste(system.file('extdata', package = 'CRISPRcleanR'),'/HT-29_counts.tsv',sep='')
normANDfcs<-ccr.NormfoldChanges(fn,min_reads=30,EXPname='Example',
                                libraryAnnotation=KY_Library_v1.0)

gwSortedFCs<-ccr.logFCs2chromPos(normANDfcs$logFCs,KY_Library_v1.0)

chr8cleaned<-ccr.cleanChrm(gwSortedFCs,8,display=TRUE,label='HT-29',
                           min.ngenes=3)
```

`ccr.correctCounts` *Correction of sgRNA treatment counts for gene independent responses to CRISPR-Cas9 targeting*

Description

This function applies an inverse transformation (described in ...) to CRISPRcleanR corrected sgRNAs' log fold changes and produces in output normalised corrected sgRNA counts (across treatments and control replicates), suitable for gene depletion/enrichment statistical testing via mean-variance modeling (for example through MAGeCK [1]*). *MAGeCK should be executed excluding initial normalisation, as the corrected sgRNA counts outputted by this function are already normalised.

Usage

```
ccr.correctCounts(CL,normalised_counts,
                  correctedFCs_and_segments,
                  libraryAnnotation,
                  minTargetedGenes=3,
                  OutDir='./',
                  ncontrols=1)
```

Arguments

`CL` A string specifying the name of the experiment. This will be used to compose names of files and folde where results will be saved.

`normalised_counts`
A data frame containing normalised sgRNAs' read counts, which can be computed using the `ccr.NormfoldChanges` function from raw sgRNAs' counts.

`correctedFCs_and_segments`
sgRNAs log fold changes corrected for gene independent responses, generated with the function `ccr.GWclean`.

`libraryAnnotation`
A data frame containing the sgRNAs' genome-wide annotations with at least a named row for each of the sgRNAs included in the foldchanges data frame provided in input. The following columns/headers should be present in this data frame (additional columns will be ignored):

- GENES: string vector containing the HGNC symbols of the genes targeted by the sgRNA under consideration;
- EXONE: string vector containing the gene exon targeted by the sgRNA under consideration (these should include the prefix "ex" followed by the exone number);
- CHRM: string vector the chromosome of the gene targeted by the sgRNA under consideration (X and Y chromosome should be specified as "X" and "Y");
- STRAND: string vector containing the strand targeted by the sgRNA under consideration ("+" or "-");
- STARTpop: numeric vector containing the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;

- ENDpos: numeric vector containing the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;

`minTargetedGenes` Minimanl number of different genes targeted by sgRNAs in a biased segment in order for the corresponding counts to be corrected (default = 3).

`OutDir` Path of the folder where results and plots will be saved.

`ncontrols` A numerical value indicating the number of control replicates (therefore columns to be considered as controls in the normalised counts).

Value

A data frame with one entry per sgRNA and individual columns for the control/treatment samples included in the normalised count data object specified by the `normalised_counts` parameter, and containing sgRNA counts corrected for gene independent responses to CRISPR-Cas9 targeting and median-ratio normalised.

Author(s)

Francesco Iorio (fi9323@gmail.com)

References

[1] Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., et al. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biology, 15(12), 554. [2] Hart, T., & Moffat, J. (2016). BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics, 17(1), 164.

See Also

`ccr.NormfoldChanges`, `ccr.GWclean`

Examples

```
## Loading sgRNA library annotation file
data(KY_Library_v1.0)

## Deriving the path of the file with the example dataset,
## from the mutagenesis of the EPLC-272H colorectal cancer cell line
fn<-paste(system.file('extdata', package = 'CRISPRcleanR'),
                      '/EPLC-272H_counts.tsv',sep='')

## Loading, median-normalizing and computing fold-changes for the example dataset
normANDfcs<-ccr.NormfoldChanges(fn,min_reads=30,
                                EXPname='EPLC-272H',
                                libraryAnnotation = KY_Library_v1.0)

## Genome-sorting of the fold changes
gwSortedFCs<-ccr.logFCs2chromPos(normANDfcs$logFCs,KY_Library_v1.0)

## Identifying and correcting biased sgRNAs' fold changes
correctedFCs<-ccr.GWclean(gwSortedFCs,display=FALSE,label='EPLC-272H')

## correcting individual sgRNA treatment counts

correctedCounts<-ccr.correctCounts('EPLC-272H',normANDfcs$norm_counts,
```

```
                    correctedFCs,
                    KY_Library_v1.0,
                    minTargetedGenes=3,
                    OutDir='./')

head(correctedCounts)
```

`ccr.geneMeanFCs` *Gene level log fold changes*

Description

This functions computes gene level log fold changes based on average log fold changes of targeting sgRNAs

Usage

```
ccr.geneMeanFCs(sgRNA_FCprofile, libraryAnnotation)
```

Arguments

`sgRNA_FCprofile`
: A named numerical vector containing the sgRNAs' log fold-changes, with names corresponding to sgRNAs identifiers.

`libraryAnnotation`
: A data frame containing the sgRNA library annotation (with same format of `KY_Library_v1.0`).

Value

A numerical vector containing gene average log fold-changes, with corresponding HGNC symbols as names.

Author(s)

Francesco Iorio (fi9323@gmail.com)

See Also

`KY_Library_v1.0`

Examples

```
## loading corrected sgRNAs log fold-changes and segment annotations for
## an example cell line (EPLC-272H)
data(EPLC.272HcorrectedFCs)


## loading sgRNA library annotation
data(KY_Library_v1.0)

## storing sgRNA log fold-changes in a named vector
FCs<-EPLC.272HcorrectedFCs$corrected_logFCs$avgFC
```

```
names(FCs)<-rownames(EPLC.272HcorrectedFCs$corrected_logFCs)

## computing gene level log fold-changes
geneFCs<-ccr.geneMeanFCs(FCs,KY_Library_v1.0)

head(geneFCs)
```

`ccr.genes2sgRNAs` *Targeting sgRNAs*

Description

This function returns the set of sgRNAs targeting the set of genes provided in input, in a given pooled library.

Usage

```
ccr.genes2sgRNAs(libraryAnnotation,genes)
```

Arguments

`libraryAnnotation`	A data frame with a named row for each sgRNA with the same format of `KY_Library_v1.0`
`genes`	A list of strings containing HGNC symbols

Value

A list of strings containing the identifiers of the sgRNAs targeting the inputted set of genes

Author(s)

Francesco Iorio (fi9323@gmail.com)

See Also

`KY_Library_v1.0`

Examples

```
## Loading an sgRNA pooled library annotation
data(KY_Library_v1.0)
## Loading an example set of genes
data(BAGEL_essential)

ccr.genes2sgRNAs(KY_Library_v1.0,BAGEL_essential)
```

`ccr.get.CCLEgisticSets` *CCLE gistic score gene sets*

Description

This function splits all the genes into 5 classes (-2, -1, 0, +1 and +2) based on the CNA Gistic [1] score observed in a given cell line.

Usage

```
ccr.get.CCLEgisticSets(cellLine,CCLE.gisticCNA=NULL)
```

Arguments

`cellLine`	A string specifying the name of a cell line (or a COSMIC identifier [2]);
`CCLE.gisticCNA`	Genome-wide Gistic [1] scores quantifying copy number status across cell lines with the same format of `CCLE.gisticCNA`. If NULL then this function uses the `CCLE.gisticCNA` builtin data frame, containing data for 13 cell lines of the 15 used in [3] to assess the performances of CRISPRcleanR.

Value

A named list of vectors with the following fields:

`gm2`	A vector of strings containing identifiers of sgRNAs targeting genes whit a Gistic score = -2 in the cell line under consideration;
`gm1`	A vector of strings containing identifiers of sgRNAs targeting genes whit a Gistic score = -1 in the cell line under consideration;
`gz`	A vector of strings containing identifiers of sgRNAs targeting genes whit a Gistic score = 0 in the cell line under consideration;
`gp1`	A vector of strings containing identifiers of sgRNAs targeting genes whit a Gistic score = +1 in the cell line under consideration;
`gp2`	A vector of strings containing identifiers of sgRNAs targeting genes whit a Gistic score = +2 in the cell line under consideration;

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

References

[1] Mermel CH, Schumacher SE, Hill B, et al. *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol. 2011;12(4):R41. doi: 10.1186/gb-2011-12-4-r41.

[2] Forbes SA, Beare D, Boutselakis H, et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D777-D783,

[3] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

See Also

ccr.get.gdsc1000.AMPgenes

Examples

```
GS<-ccr.get.CCLEgisticSets('HT-29')

head(GS$gm2)
head(GS$gm1)
head(GS$gz)
head(GS$gp1)
head(GS$gp2)
```

---

ccr.get.gdsc1000.AMPgenes

Copy number amplified genes in a given cell line from the GDSC1000

Description

This function takes in input the name (or the COSMIC identifier [1]) of a cell line included in the GDSC1000 project [2] and it identifies the genes that are copy number amplified (according to a user defined minimal copy number value) in that cell line, using gene level copy number data from the Genomics of Drug Sensitivity in 1,000 Cancer Cell lines (GDSC1000) [2].

Usage

```
ccr.get.gdsc1000.AMPgenes(cellLine, minCN = 8, exact = FALSE,
                          GDSC.geneLevCNA=NULL)
```

Arguments

cellLine	A string specifying the name of a cell line (or a COSMIC identifier [1]);
minCN	Lower threshold for the minimum copy number of any genomic segment containing coding sequence of a gene in order for it to be considered as copy number amplified.
exact	If TRUE, then those genes for which any genomic segment containing coding sequence has a minimum copy number equal to minCN are considered as copy number amplified.
GDSC.geneLevCNA	Genome-wide copy number data with the same format of GDSC.geneLevCNA. This can be assembled from the xls sheet specified in the source section [a] (containing data for the GDSC1000 cell lines). If NULL, then this function uses the data in the built in GDSC.geneLevCNA data frame, containing data derived from [a] for 15 cell lines used in [3] to assess the performances of CRISPRcleanR.

Value

A data frame, containing one row for each copy number amplified gene with the following columns:

`Gene`	HGNC symbol of the gene;
`minCN`	Minimum copy number of any genomic segment containing coding sequence of the gene in the cell line under consideration.

Author(s)

Francesco Iorio (fi9323@gmail.com)

Source

[a] `ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/release-6.0/Gene_level_CN.xlsx`.

References

[1] Forbes SA, Beare D, Boutselakis H, et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D777-D783,

[2] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, et al. *A landscape of pharmacogenomic interactions in cancer* Cell 2016 Jul 28;166(3):740-54

[3] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

See Also

`ccr.get.CCLEgisticSets`

Examples

```
CNAgenes<-
    ccr.get.gdsc1000.AMPgenes('HT-29')
head(CNAgenes)
```

---

ccr.get.nonExpGenes *Non expressed genes in a given cell line*

Description

This function takes in input the name (or the COSMIC identifier [1]) of a cell line and it identifies genes that are not expressed (according to a user defined FPKM threshold) using a collection of RNAseq profile from [2].

Usage

```
ccr.get.nonExpGenes(cellLine, th = 0.05,
                    amplified = FALSE, minCN = 8,
                    RNAseq.fpkms=NULL)
```

Arguments

cellLine	A string specifying the name of a cell line (or a COSMIC identifier [1]);
th	Minimum FPKM value for a gene to be considered as expressed;
amplified	A logic value specifying whether the selected not expressed genes should be also copy number amplified function;
minCN	If amplified = TRUE, this parameter defines a lower threshold for the minimum copy number of any genomic segment containing coding sequence of a gene in order for it to be considered as copy number amplified.
RNAseq.fpkms	Genome-wide substitute reads with fragments per kilobase of exon per million reads mapped (FPKM) across cell lines. These can be derived from a comprehensive collection of RNAseq profiles described in [2]. The format must be the same of the RNAseq.fpkms builtin data frame. If NULL then this function uses the RNAseq.fpkms builtin data fram containing data for 15 cell lines used in [3] to assess CRISPRcleaneR results.

Value

A vector of string containing the HGNC symbols of non expressed (optionally copy number amplified) genes in the cell line under consideration.

Author(s)

Francesco Iorio (fi9323@gmail.com)

References

[1] Forbes SA, Beare D, Boutselakis H, et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D777-D783.

[2] Garcia-Alonso L, Iorio F, Matchan A, et al. *Transcription factor activities enhance markers of drug response in cancer* doi: https://doi.org/10.1101/129478

[3] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

See Also

ccr.get.gdsc1000.AMPgenes

Examples

```
ccr.get.nonExpGenes('HT-29',amplified = TRUE)
```

`ccr.GWclean` *Unsupervised identification and correction of gene independent cell responses to CRISPR-Cas9 targeting.*

Description

This function takes in input a genome-wide essentiality profile derived from a CRISPR-Cas9 experiment employing a pooled library of single guide RNAs (sgRNAs) targeting protein coding genes, which are transfected in an *in vitro* model stably expressing Cas9. The essentiality profile quantifies the loss/gain-of-fitness caused by each sgRNA-targeting, and it is expressed as log fold changes (logFCs) between the aboundance of the sgRNAs at an end point after cell purification and their aboundance in the plasmid pool used for viral production, or at an initial time point, or in any other control condition. A circular binary segmentation algorithm [1, 2] is applied by this function to the genome-wide pattern of logFCs provided in input, in order to identify genomic regions including sgRNAs with sufficiently equal logFC (and mean logFC sufficiently different from background) and targeting a minimal number of different genes. Assuming that it is very unlikely to observe the same loss/gain-of-fitness effect when targeting a large number of contiguous genes, if certain user-defined condition (detailed below) are met then the logFCs of such regions are deemed as biased by some local feature of the involved genomic segment (which could be, for example, copy number amplified [3]), and they are corrected, i.e. mean centered [4].

Usage

```
ccr.GWclean(gwSortedFCs,label='',display=TRUE,
            saveTO=NULL,ignoredGenes=NULL,min.ngenes=3)
```

Arguments

`gwSortedFCs` A data frame containing genome-wide genomic-sorted sgRNAs' log fold changes. This data frame must include one named row per each sgRNA and the following columns/headers:

- CHR: the chromosome of the gene targeted by the sgRNA under consideration;
- startp: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;
- endp: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;
- genes: the HGNC symbol of the gene targeted by the sgRNA under consideration;
- avgFC: the log fold change of the sgRNA under consideration averaged across replicates;
- BP: the genomic coordinate of the sgRNA defined as STARTpos+(ENDpos-STARTpos)/2.

This can be generated using the `ccr.logFCs2chromPos` function, starting from a data frame containing sgRNAs' log fold changes generated by the `ccr.NormfoldChanges` function (from raw sgRNAs' counts), from raw sgRNAs' counts.

label	A string indicating the experiment name. This is used to compose the main title of the plots generated by this function and the name of the folder where the results are saved.
display	A logical value indicating whether genomic plots showing the results of the biased regions' identification and their log fold change correction should be generated or not.
saveTO	If different from NULL then this parameter will contain the path where pdf files with then genomic plots showing the results of the biased regions' identification (and their log fold change correction) will be saved (within a folder named as defined in the label parameter).
ignoredGenes	A vector of strings containing HGNC symbols of genes that should not be considered when computing the minimal number of different genes targeted by sgRNAs in the same identified region of estimated equal log fold changes. This could contain, for example, a-priori known essential genes.
min.ngenes	A numerical value (>0) specifying the minimal number of different genes that the set of sgRNAs within a region of estimated equal logFCs should target in order for theri logFCs to be corrected, i.e. mean centered.

Value

A list containing two data frames and a vector of strings. The first data frame (corrected_logFCs) contains a named row per each sgRNA and the following columns/header:

- CHR: the chromosome of the gene targeted by the sgRNA under consideration;
- startp: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;
- endp: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;
- genes: the HGNC symbol of the gene targeted by the sgRNA under consideration;
- avgFC: the log fold change of the sgRNA averaged across replicates;
- correction: the type of correction: 1 = increased log fold change, -1 = decreased log fold change. 0 indicates no correction;
- correctedFC: the corrected log fold change of the sgRNA

.
The second data frame (segments) contains the identified region of estimated equal log fold changes (one region per row) and the following columns/headers:

- CHR: the chromosome of the gene targeted by the sgRNA under consideration;
- startp: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;
- endp: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;
- genes: the HGNC symbol of the gene targeted by the sgRNA under consideration;
- avgFC: the log fold change of the sgRNA averaged across replicates;
- correction: the type of correction: 1 = increased log fold change, -1 = decreased log fold change. 0 indicates no correction;
- correctedFC: the corrected log fold change of the sgRNA

.
The second data frame (segments) contains the identified region of estimated equal log fold changes (one region per row) and the following columns/headers:

- `CHR`: the chromosome of the region under consideration;
- `startp`: the genomic coordinate of the starting position of the region under consideration;
- `endp`: the genomic coordinate of the ending position of the region under consideration;
- `n.sgRNAs`: the number of sgRNAs targeting sequences in the region under consideration;
- `avg.logFC`: the average log fold change of the sgRNAs in the region;
- `guideIdx`: the indexes range of the sgRNAs targeting the region under consideration as they appear in the gwSortedF Cs provided in input.

The string of vectors (SORTED_sgRNAs) contains the sgRNAs' identifiers in the same order as they are reported in the gwSortedFCs input data frame, i.e. genome sorted.

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

References

[1] Olshen, A. B., Venkatraman, E. S., Lucito, R., Wigler, M. (2004). *Circular binary segmentation for the analysis of array-based DNA copy number data*. Biostatistics 5: 557-572. \

[2] Venkatraman, E. S., Olshen, A. B. (2007). *A faster circular binary segmentation algorithm for the analysis of array CGH data*. Bioinformatics 23: 657-63. \

[3] Andrew J. Aguirre, Robin M. Meyers, Barbara A. Weir, Francisca Vazquez, Cheng-Zhong Zhang, Uri Ben-David, April Cook, Gavin Ha, William F. Harrington, Mihir B. Doshi, Maria Kost-Alimova, Stanley Gill, Han Xu, Levi D. Ali, Guozhi Jiang, Sasha Pantel, Yenarae Lee, Amy Goodale, Andrew D. Cherniack, Coyin Oh, Gregory Kryukov, Glenn S. Cowley, Levi A. Garraway, Kimberly Stegmaier, Charles W. Roberts, Todd R. Golub, Matthew Meyerson, David E. Root, Aviad Tsherniak and William C. Hahn. *Genomic copy number dictates a gene-independent cell response to CRISPR-Cas9 targeting*. Cancer Discov June 3 2016 DOI: 10.1158/2159-8290.CD-16-0154

[4] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

See Also

`ccr.cleanChrm`

Examples

```
## Loading sgRNA library annotation file
data(KY_Library_v1.0)

## Deriving the path of the file with the example dataset,
## from the mutagenesis of the HT-29 colorectal cancer cell line
fn<-paste(system.file('extdata', package = 'CRISPRcleanR'),'/HT-29_counts.tsv',sep='')

## Loading, median-normalizing and computing fold-changes for the example dataset
normANDfcs<-ccr.NormfoldChanges(fn,min_reads=30,EXPname='HT-29',
```

```
                                libraryAnnotation = KY_Library_v1.0)

## Genome-sorting of the fold changes
gwSortedFCs<-ccr.logFCs2chromPos(normANDfcs$logFCs,KY_Library_v1.0)

## Identifying and correcting biased sgRNAs' fold changes
correctedFCs<-ccr.GWclean(gwSortedFCs,display=TRUE,label='HT-29')

## Visualising first five entries of the corrected fold changes
head(correctedFCs$corrected_logFCs)
```

`ccr.logFCs2chromPos` *Genomic sorting of sgRNAs' log fold changes.*

Description

This function maps genome-wide sgRNAs' log fold changes (averaged across replicates) on the genome and returns them sorted according to the position of their targeted region on the chromosomes.

Usage

```
ccr.logFCs2chromPos(foldchanges, libraryAnnotation)
```

Arguments

`foldchanges` A data frame containing genome-wide sgRNAs' log fold changes, one column per library transfection replicate, with first and second column containing the sgRNAs' identifiers and the HGNC symbols of the targeted genes, respectively. This can be generated from raw count files using the `ccr.NormfoldChanges` function.

`libraryAnnotation`
A data frame containing the sgRNAs' genome-wide annotations with at least a named row for each of the sgRNAs included in the foldchanges data frame provided in input. The following columns/headers should be present in this data frame (additional columns will be ignored):

- GENES: string vector containing the HGNC symbols of the genes targeted by the sgRNA under consideration;
- EXONE: string vector containing the gene exon targeted by the sgRNA under consideration (these should include the prefix "ex" followed by the exone number);
- CHRM: string vector the chromosome of the gene targeted by the sgRNA under consideration (X and Y chromosome should be specified as "X" and "Y");
- STRAND: string vector containing the strand targeted by the sgRNA under consideration ("+" or "-");
- STARTpop: numeric vector containing the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;
- ENDpos: numeric vector containing the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;

Additiol columns can be optionally included and will be ignored by this function. The annation for the genome-wide sgRNA library presented in [1] is included in the KY_Library_v1.0 data object, formatted as described above.

Value

A data frame with a named row per each sgRNA and the following columns/headers:

- CHR: the chromosome where the gene targeted by the sgRNA under consideration resides;
- startp: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;
- endp: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;
- avgFC: the log fold change of the sgRNA averaged across replicates;
- BP: the genomic coordinate of the sgRNA defined as STARTpos+(ENDpos-STARTpos)/2.

Author(s)

Francesco Iorio (iorio@ebi.ac.uk)

References

[1] Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, Mupo A, Grinkevich V, Li M, Mazan M, Gozdecka M, Onishi S, Cooper J, Patel M, McKerrell T, Chen B, Domingues AF, Gallipoli P, Teichmann S, Ponstingl H, McDermott U, Saez-Rodriguez J, Huntly BJP, Iorio F, Pina C, Vassiliou GS, Yusa K. *A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukaemia*. Cell Reports 2016 Oct 18;17(4):1193-1205

See Also

ccr.NormfoldChanges, KY_Library_v1.0

Examples

```
 data(KY_Library_v1.0)
fn<-paste(system.file('extdata', package = 'CRISPRcleanR'),'/A2058_counts.tsv',sep='')
 normANDfcs<-ccr.NormfoldChanges(fn,min_reads=30,
                                 EXPname='Example',
                                 libraryAnnotation=KY_Library_v1.0)
 mappedLogFCs<-ccr.logFCs2chromPos(normANDfcs$logFCs,KY_Library_v1.0)
 head(mappedLogFCs)
```

---

ccr.multDensPlot *Mutiple shaded density plot*

Description

This functions plots multiple distribution densities with solid colors for the curves and shaded colors for underlying areas.

Usage

```
ccr.multDensPlot(TOPLOT, COLS,
                 XLIMS, TITLE, LEGentries, XLAB)
```

Arguments

TOPLOT	A list of density object computed using the density function of the stats package.
COLS	A vector of colors of the same length of TOPLOT that are used to plot the density curves. Alpha-reduced versions of these colors are used to fill the underlying areas.
XLIMS	A vector of two numerical values optionally specifying x-axis limits (NULL by default).
TITLE	A string containing the plot title.
LEGentries	A vector of strings (one per each density in TOPLOT) specifying corresponding legend entries.
XLAB	A string containing the x-axis label.

Author(s)

Francesco Iorio (fi9323@gmail.com)

Examples

```
## generating random data
x <- rnorm(1000, 0, 0.5)
y <- rnorm(1000, 2, 0.4)
z <- rnorm(1000, -1, 1.5)


## assembling kernel estimated distributions into a list
ToPlot<-list(x=density(x),y=density(y),z=density(z))

## density visualisation
ccr.multDensPlot(ToPlot,COLS = c('red','blue','gray'),
    TITLE = 'example',LEGentries = c('x','y','z'),
    XLIMS = c(-5,3))
```

---

ccr.NormfoldChanges *Median-ratio normalisation of sgRNA counts and fold change computation*

Description

This function median-ratio normalises [1,2] sgRNAs' counts stored in a tsv file whose path is provided in input, to adjust for the effect of library size and read count distributions. It computes log fold changes of transfected library replicates versus controls (tipically the sgRNA counts in the plasmid). The output of this function is returned as a list, and it is also saved into two tsv files.

Usage

```
ccr.NormfoldChanges(filename, display=TRUE, saveToFig=FALSE,
                    outdir='./', min_reads=30, EXPname='',
                    libraryAnnotation, ncontrols=1)
```

Arguments

filename	A string specifying the path of a tsv file containing the raw sgRNA counts. This must be a tab delimited file with one row per sgRNA and the following columns/headers: • sgRNA: containing alphanumerical identifiers of the sgRNA under consideration; • gene: containing HGNC symbols of the genes targeted by the sgRNA under consideration; followed by the columns containing the sgRNAs' counts for the controls and columns for library trasfected samples.
display	A logic value specifying whether figures containing boxplots with the count values pre/post normalisation and log fold-changes should be visualised (TRUE, by default).
saveToFig	A logic value specifying whether figures containing boxplots with the count values pre/post normalisation and log fold-changes should be saved as pdf files (FALSE, by default). Setting this parameter to TRUE overrides the value of the display parameter.
outdir	Path of the directory where the normalised sgRNAs' counts and the log fold changes, as well as the pdf files (if the parameter saveToFig is set to TRUE), must be saved.
min_reads	This parameter defines a filter threshold value for sgRNAs, based on their average counts in the control sample. Specifically, it indicates the minimal number of counts that each individual sgRNA needs to have in the controls (on average) in order to be included in the output.
EXPname	A string specifying the name of the experiment. This will be used to compose main title of the generated figures and file names.
libraryAnnotation	A data frame containing the sgRNA annotations, with a named row for each sgRNA, and columns for targeted genes, genomic coordinates and possibly other informations. This should be formatted as the KY_Library_v1.0 data object containing the annotation of the sgRNA library presented in [3].
ncontrols	A numerical value indicating the number of control replicates (therefore columns to be considered as control counts after the first two, in the inputted tsv file).

Value

A list containing two data frames: for the normalised sgRNAs' counts (norm_counts) and the sgRNAs' log fold changes (logFCs) respectively. First two columns in these data frames contain sgRNAs' identifiers and HGNC symbols of targete gene, respectively.

Author(s)

Francesco Iorio (fi9323@gmail.com)

References

[1] Wang T, Wei JJ, Sabatini DM, Lander ES. *Genetic screens in human cells using the CRISPR-Cas9 system*. Science. 2014, 343: 80-84.
[2] Anders S, Huber W. *Differential expression analysis for sequence count data*. Genome Biol. 2010, 11: R106
[3] Tzelepis K, Koike-Yusa H, De Braekeleer E, et al *A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukaemia*. Cell Reports 2016 Oct 18;17(4):1193-1205

See Also

KY_Library_v1.0

Examples

```
## loading sgRNA library annotation
data(KY_Library_v1.0)

## derive path for an example dataset
fn<-paste(system.file('extdata', package = 'CRISPRcleanR'),'/HT-29_counts.tsv',sep='')

## sgRNAs' normalisation and computation of log fold-changes
normANDfcs<-ccr.NormfoldChanges(fn,
                                min_reads=30,
                                EXPname='Example',
                                libraryAnnotation=KY_Library_v1.0)

## inspecting first 5 entries of the data frames containing the
## normalised counts and the log fold-changes
head(normANDfcs$norm_counts)
head(normANDfcs$logFCs)
```

---

ccr.perf_distributions

CRISPRcleanR correction assessment: inspection of sgRNA log fold changes distributions

---

Description

This function creates distributions density plots of sgRNA log fold changes for defined sets of targeted genes prior/post CRISPRcleanR correction.

Usage

```
ccr.perf_distributions(cellLine, correctedFCs,
                       GDSC.geneLevCNA = NULL,
                       CCLE.gisticCNA = NULL,
                       RNAseq.fpkms = NULL,
                       minCNs = c(8, 10),
                       libraryAnnotation)
```

Arguments

cellLine	A string specifying the name of a cell line (or a COSMIC identifier [1]);
correctedFCs	sgRNAs log fold changes corrected for gene independent responses to CRISPR-Cas9 targeting, generated with the function ccr.GWclean (first data frame included in the list outputted by ccr.GWclean, i.e. corrected_logFCs).
GDSC.geneLevCNA	Genome-wide copy number data with the same format of GDSC.geneLevCNA. This can be assembled from the xls sheet specified in the source section [a] (containing data for the GDSC1000 cell lines). If NULL, then this function uses the built in GDSC.geneLevCNA data frame, containing data derived from [a] for 15 cell lines used in [2] to assess the performances of CRISPRcleanR.
CCLE.gisticCNA	Genome-wide Gistic [3] scores quantifying copy number status across cell lines with the same format of CCLE.gisticCNA. If NULL then this function uses the CCLE.gisticCNA builtin data frame, containing data for 13 cell lines of the 15 used in [2] to assess the performances of CRISPRcleanR.
RNAseq.fpkms	Genome-wide substitute reads with fragments per kilobase of exon per million reads mapped (FPKM) across cell lines. These can be derived from a comprehensive collection of RNAseq profiles described in [4]. The format must be the same of the RNAseq.fpkms builtin data frame. If NULL then this function uses the RNAseq.fpkms builtin data fram containing data for 15 cell lines used in [2] to assess CRISPRcleaneR results.
minCNs	A numerical vector with two entries specifying the minimal copy number for a gene in order to be considered amplified based on the data in GDSC.geneLevCNA. These two values can be 2, 4, 8 or 10.
libraryAnnotation	The sgRNA library annotations formatted as specified in the reference manual entry of the KY_Library_v1.0 built in library.

Details

This function generates 4 sets of plots. They contains log fold change distributions density plots prior/post CRISPRcleanR correction respectively for

- (i) Copy number amplified genes according to the data in GDSC.geneLevCNA based on the two threshold values specified in minCNs;
- (ii) Copy number amplified genes according to the data in CCLE.gisticCNA (gistic score = +2);
- (iii) Copy number amplified non expressed genes according to the data in GDSC.geneLevCNA based on the two threshold values specified in minCNs, and the data in RNAseq.fpkms (FPKM < 0.05);
- (iv) reference sets of core fitness essential genes from MSigDB [5] (included in the builtin vectors EssGenes.DNA_REPLICATION_cons, EssGenes.KEGG_rna_polymerase, EssGenes.PROTEASOME_cons, EssGenes.ribosomalProteins, EssGenes.SPLICEOSOME_cons, and reference core-fitness-essential and non-essential genes assembled from multiple RNAi studies used as classification template by the BAGEL algorithm to call gene depletion significance [6] (BAGEL_essential, BAGEL_nonEssential).

Author(s)

Francesco Iorio (fi9323@gmail.com)

Source

[a] ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/release-6.0/Gene_level_CN.xlsx.

References

[1] Forbes SA, Beare D, Boutselakis H, et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D777-D783.

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

[3] Mermel CH, Schumacher SE, Hill B, et al. *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol. 2011;12(4):R41. doi: 10.1186/gb-2011-12-4-r41.

[4] Garcia-Alonso L, Iorio F, Matchan A, et al. *Transcription factor activities enhance markers of drug response in cancer* doi: https://doi.org/10.1101/129478

[5] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545-15550. http://doi.org/10.1073/pnas.0506580102

[6] BAGEL: a computational framework for identifying essential genes from pooled library screens. Traver Hart and Jason Moffat. BMC Bioinformatics, 2016 vol. 17 p. 164.

See Also

KY_Library_v1.0, ccr.GWclean,
GDSC.geneLevCNA, CCLE.gisticCNA, RNAseq.fpkms,
EssGenes.DNA_REPLICATION_cons, EssGenes.KEGG_rna_polymerase, EssGenes.PROTEASOME_cons,
EssGenes.ribosomalProteins, EssGenes.SPLICEOSOME_cons
BAGEL_essential, BAGEL_nonEssential

Examples

```
## loading corrected sgRNAs log fold-changes and segment annotations for an example
## cell line (HT-29)
data(HT.29correctedFCs)

## loading library annotation
data(KY_Library_v1.0)

## inpecting sgRNA log fold change distributions prior/post CRISPRcleanR correction
ccr.perf_distributions('HT-29',HT.29correctedFCs$corrected_logFCs,
                       libraryAnnotation = KY_Library_v1.0)
```

`ccr.perf_statTests` *CRISPRcleanR correction assessment: Statistical tests*

Description

This function tests the log fold changes of sgRNAs targeting different sets of genes for statistically significant differences with respect to background pre and post CRISPRcleanR correction, creating two sets of boxplots with outcomes and outputting statistical indicators.

Usage

```
ccr.perf_statTests(cellLine, libraryAnnotation, correctedFCs,
                   outDir = "./",
                   GDSC.geneLevCNA = NULL,
                   CCLE.gisticCNA = NULL,
                   RNAseq.fpkms = NULL)
```

Arguments

`cellLine`	A string specifying the name of a cell line (or a COSMIC identifier [1]);
`libraryAnnotation`	The sgRNA library annotations formatted as specified in the reference manual entry of the `KY_Library_v1.0` built in library.
`correctedFCs`	sgRNAs log fold changes corrected for gene independent responses to CRISPR-Cas9 targeting, generated with the function `ccr.GWclean` (first data frame included in the list outputted by `ccr.GWclean`, i.e. `corrected_logFCs`).
`outDir`	The path of the folder where the boxplot will be saved.
`GDSC.geneLevCNA`	Genome-wide copy number data with the same format of `GDSC.geneLevCNA`. This can be assembled from the xls sheet specified in the source section [a] (containing data for the GDSC1000 cell lines). If NULL, then this function uses the built in `GDSC.geneLevCNA` data frame, containing data derived from [a] for 15 cell lines used in [2] to assess the performances of CRISPRcleanR.
`CCLE.gisticCNA`	Genome-wide Gistic [3] scores quantifying copy number status across cell lines with the same format of `CCLE.gisticCNA`. If NULL then this function uses the `CCLE.gisticCNA` builtin data frame, containing data for 13 cell lines of the 15 used in [2] to assess the performances of CRISPRcleanR.
`RNAseq.fpkms`	Genome-wide substitute reads with fragments per kilobase of exon per million reads mapped (FPKM) across cell lines. These can be derived from a comprehensive collection of RNAseq profiles described in [4]. The format must be the same of the `RNAseq.fpkms` builtin data frame. If NULL then this function uses the `RNAseq.fpkms` builtin data fram containing data for 15 cell lines used in [2] to assess CRISPRcleaneR results.

Details

This functions assess the statistical difference pre/post CRISPRcleanR correction of log fold changes for sgRNAs targeting respectively:

- copy number (CN) deleted genes according to the GDSC1000 repository

- CN deleted genes (gistic score = -2) according to the CCLE repository
- non expressed genes (FPKM < 0.05)
- genes with gistic score = 1
- genes with gistic score = 2
- non espressed genes (FPKM < 0.05) with gistic score = 1
- non espressed genes (FPKM < 0.05) with gistic score = 2
- genes with minimal CN = 2, according to the GDSC1000
- genes with minimal CN = 4, according to the GDSC1000
- genes with minimal CN = 8, according to the GDSC1000
- genes with minimal CN = 10, according to the GDSC1000
- non expressed genes (FPKM < 0.05) with minimal CN = 2, according to the GDSC1000
- non expressed genes (FPKM < 0.05) with minimal CN = 4, according to the GDSC1000
- non expressed genes (FPKM < 0.05) with minimal CN = 8, according to the GDSC1000
- non expressed genes (FPKM < 0.05) with minimal CN = 10, according to the GDSC1000
- core fitness essential genes, assembling signatures from MsigDB [5], included in the builtin vectors `EssGenes.DNA_REPLICATION_cons`, `EssGenes.KEGG_rna_polymerase`, `EssGenes.PROTEASOME_cons`, `EssGenes.ribosomalProteins`, `EssGenes.SPLICEOSOME_cons`
- Reference core fitness essential genes assembled from multiple RNAi studies used as classification template by the BAGEL algorithm to call gene depletion significance [6] (`BAGEL_essential`)
- Reference core fitness essential genes assembled from multiple RNAi studies used as classification template by the BAGEL algorithm to call gene depletion significance [6] after the removal core fitness essential genes from MsigDB [5]
- Reference non essential genes assembled from multiple RNAi studies used as classification template by the BAGEL algorithm to call gene depletion significance [6] (`BAGEL_nonEssential`)

Value

A list of three named 2x19 matrices, with one entry per statistical test, rows indicating pre/post CRISPRcleanR correction sgRNAs' log fold changes and one column per each tested gene set. In each matrix the entries contains, respectively

`PVALS`	Pvalue resulting from a Student's t-test assessing the differences between sgRNAs log fold changes pre (first row) and post (second row) CRISPRcleanR correction with respect to background
`SIGNS`	The sign of the difference (1 = mean log fold change of the tested set larger that the mean of the background population, -1 = mean log fold change of the tested set smaller than the mean of the background population)
`EFFsizes`	Effect size (computing via the Cohen's D): difference of the means / pooled standard deviation.

Author(s)

Francesco Iorio (fi9323@gmail.com)

Source

[a] ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/release-6.0/Gene_level_CN.xlsx.

References

[1] Forbes SA, Beare D, Boutselakis H, et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D777-D783.

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

[3] Mermel CH, Schumacher SE, Hill B, et al. *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol. 2011;12(4):R41. doi: 10.1186/gb-2011-12-4-r41.

[4] Garcia-Alonso L, Iorio F, Matchan A, et al. *Transcription factor activities enhance markers of drug response in cancer* doi: https://doi.org/10.1101/129478

[5] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545-15550. http://doi.org/10.1073/pnas.0506580102

[6] BAGEL: a computational framework for identifying essential genes from pooled library screens. Traver Hart and Jason Moffat. BMC Bioinformatics, 2016 vol. 17 p. 164.

See Also

KY_Library_v1.0, ccr.GWclean,
GDSC.geneLevCNA, CCLE.gisticCNA, RNAseq.fpkms,
EssGenes.DNA_REPLICATION_cons, EssGenes.KEGG_rna_polymerase, EssGenes.PROTEASOME_cons,
EssGenes.ribosomalProteins, EssGenes.SPLICEOSOME_cons
BAGEL_essential, BAGEL_nonEssential

Examples

```
## loading corrected sgRNAs log fold-changes and segment annotations for an example
## cell line (EPLC-272H)
data(EPLC.272HcorrectedFCs)

## loading library annotation
data(KY_Library_v1.0)

## Evaluate correction effects. Boxplots will be saved in EPLC-272H.pdf
## in the current directory
RES<-ccr.perf_statTests('EPLC-272H',libraryAnnotation = KY_Library_v1.0,
                  correctedFCs = EPLC.272HcorrectedFCs$corrected_logFCs)
RES$PVALS
RES$EFFsizes
```

`ccr.PrecisionRecallCurve` *Classification performances of reference sets of genes (or sgRNAs) based on depletion log fold-changes*

Description

This functions computes Precision/Recall (PR) curve, area under the PR curve and (optionally) Recall at fixed false discovery rate (and corresponding log fold change threshold) when classifying reference sets of genes (or sgRNAs) based on their depletion log fold-changes

Usage

```
ccr.PrecisionRecallCurve(FCsprofile,
                         positives,
                         negatives,
                         display = TRUE,
                         FDRth = NULL)
```

Arguments

`FCsprofile`	A numerical vector containing gene average depletion log fold changes (or sgRNAs' depletion log fold changes) with names corresponding to HGNC symbols (or sgRNAs' identifiers).
`positives`	A vector of strings containing a reference set of positive cases: HGNC symbols of essential genes or identifiers of their targeting sgRNAs. This must be a subset of `FCsprofile` names, disjointed from `negatives`.
`negatives`	A vector of strings containing a reference set of negative cases: HGNC symbols of essential genes or identifiers of their targeting sgRNAs. This must be a subset of `FCsprofile` names, disjointed from `positives`.
`display`	A logical parameter specifying if a plot containing the computed precision/recall curve with ROC indicators should be plotted (default = TRUE).
`FDRth`	If different from NULL, will be a numerical value >=0 and <=1 specifying the false discovery rate threshold at which fixed recall will be computed. In this case, if the `display` parameter is TRUE, an orizontal dashed line will be added to the plot at the resulting recall and its value will be visualised in the legend.

Value

A list containint three numerical variable `AUC`, `Recall`, and `sigthreshold` indicating the area under the precision/recall curve and (if `FDRth` is not NULL) the recall at the specifying false discovery rate and the corresponding log fold change threshold (both equal to NULL, if `FDRth` is NULL), respectively.

Author(s)

Francesco Iorio (fi9323@gmail.com)

See Also

`BAGEL_essential`, `BAGEL_nonEssential`, `ccr.genes2sgRNAs`, `ccr.VisDepAndSig`

Examples

```
## loading corrected sgRNAs log fold-changes and segment annotations for an example
## cell line (EPLC-272H)
data(EPLC.272HcorrectedFCs)

## loading reference sets of essential and non-essential genes
data(BAGEL_essential)
data(BAGEL_nonEssential)

## loading library annotation
data(KY_Library_v1.0)

## storing sgRNA log fold-changes in a named vector
FCs<-EPLC.272HcorrectedFCs$corrected_logFCs$avgFC
names(FCs)<-rownames(EPLC.272HcorrectedFCs$corrected_logFCs)

## deriving sgRNAs targeting essential and non-essential genes (respectively)
BAGEL_essential_sgRNAs<-ccr.genes2sgRNAs(KY_Library_v1.0,BAGEL_essential)
BAGEL_nonEssential_sgRNAs<-ccr.genes2sgRNAs(KY_Library_v1.0,BAGEL_nonEssential)

## computing classification performances at the sgRNA level
ccr.PrecisionRecallCurve(FCs,BAGEL_essential_sgRNAs,BAGEL_nonEssential_sgRNAs)

## computing gene level log fold-changes
geneFCs<-ccr.geneMeanFCs(FCs,KY_Library_v1.0)

## computing classification performances at the sgRNA level, with Recall at 5% FDR
ccr.PrecisionRecallCurve(geneFCs,BAGEL_essential,BAGEL_nonEssential,FDRth = 0.05)
```

`ccr.RecallCurves` *CRISPRcleanR correction assessment: Recall curve inspection*

Description

This function creates plots with Recall curve outcomes (a it computes areas under the Recall curves) resulting from classifying defined sets of sgRNAs (respectively genes) based on their log fold change (respectively log fold changes averaged across targeting sgRNAs).

Usage

```
ccr.RecallCurves(cellLine, correctedFCs, GDSC.geneLevCNA = NULL,
                 RNAseq.fpkms = NULL, minCN = 8, libraryAnnotation,
                 GeneLev = FALSE)
```

Arguments

`cellLine`	A string specifying the name of a cell line (or a COSMIC identifier [1]);
`correctedFCs`	sgRNAs log fold changes corrected for gene independent responses to CRISPR-Cas9 targeting, generated with the function `ccr.GWclean` (first data frame included in the list outputted by `ccr.GWclean`, i.e. `corrected_logFCs`).

GDSC.geneLevCNA
: Genome-wide copy number data with the same format of GDSC.geneLevCNA. This can be assembled from the xls sheet specified in the source section [a] (containing data for the GDSC1000 cell lines). If NULL, then this function uses the built in GDSC.geneLevCNA data frame, containing data derived from [a] for 15 cell lines used in [2] to assess the performances of CRISPRcleanR.

RNAseq.fpkms
: Genome-wide substitute reads with fragments per kilobase of exon per million reads mapped (FPKM) across cell lines. These can be derived from a comprehensive collection of RNAseq profiles described in [4]. The format must be the same of the RNAseq.fpkms builtin data frame. If NULL then this function uses the RNAseq.fpkms builtin data fram containing data for 15 cell lines used in [2] to assess CRISPRcleaneR results.

minCN
: A numerical value specifying the minimal copy number for a gene in order to be considered amplified based on the data in GDSC.geneLevCNA. This value can be 2, 4, 8 or 10.

libraryAnnotation
: The sgRNA library annotations formatted as specified in the reference manual entry of the KY_Library_v1.0 built in library.

GeneLev
: A logical value specifying if the Recall should be computed at level of genes. In this case average gene log fold changes are computed from the inputted corrected log fold changes across targeting sgRNAs.

Details

This function generates 2 plots, showing Recall curves resulting from classifying the following 4 sets of sgRNAs (or Genes, depending on the parameter GeneLev, based on their log fold changes (or log fold changes averaged across targeting guides):

- (i) Copy number amplified genes according to the data in GDSC.geneLevCNA based on the threshold value specified in minCNs;
- (ii) Copy number amplified non expressed genes according to the data in GDSC.geneLevCNA based on the threshold value specified in minCNs, and the data in RNAseq.fpkms (FPKM < 0.05);
- (iv) reference sets of core-fitness-essential and non-essential genes assembled from multiple RNAi studies used as classification template by the BAGEL algorithm to call gene depletion significance [5]
(BAGEL_essential, BAGEL_nonEssential).

Author(s)

Francesco Iorio (fi9323@gmail.com)

Source

[a] ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/release-6.0/Gene_level_CN.xlsx.

References

[1] Forbes SA, Beare D, Boutselakis H, et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D777-D783.

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

[3] Mermel CH, Schumacher SE, Hill B, et al. *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol. 2011;12(4):R41. doi: 10.1186/gb-2011-12-4-r41.

[4] Garcia-Alonso L, Iorio F, Matchan A, et al. *Transcription factor activities enhance markers of drug response in cancer* doi: https://doi.org/10.1101/129478

[5] BAGEL: a computational framework for identifying essential genes from pooled library screens. Traver Hart and Jason Moffat. BMC Bioinformatics, 2016 vol. 17 p. 164.

See Also

KY_Library_v1.0, ccr.GWclean,
GDSC.geneLevCNA, RNAseq.fpkms,
BAGEL_essential, BAGEL_nonEssential

Examples

```
## loading corrected sgRNAs log fold-changes and segment annotations for an example
## cell line (EPLC-272H)
data(EPLC.272HcorrectedFCs)

## loading library annotation
data(KY_Library_v1.0)

## Creating recall curve plots and computing corresponding underlying area
## at the level of sgRNAs
ccr.RecallCurves('EPLC-272H',EPLC.272HcorrectedFCs$corrected_logFCs,
                 libraryAnnotation=KY_Library_v1.0)

## Creating recall curve plots and computing corresponding underlying area
## at the gene level
ccr.RecallCurves('EPLC-272H',EPLC.272HcorrectedFCs$corrected_logFCs,
                 libraryAnnotation=KY_Library_v1.0,GeneLev = TRUE)
```

---

ccr.VisDepAndSig *Depletion profile visualisation with genes signatures superimposed and recall*

Description

This functions ranks the gene (or sgRNAs) log fold changes. Based on this it determines a log fold change threshold based on a user defined false discovery rate when classifying two gene (sgRNA) positive/negative references sets (tipically core-fitness-essential and non-essential genes), and it computes the Recall (or True Positive Rate) of genes in other user defined sets at the determined threshold. It produces a plot where the log fold changes are visualised alongside the rank positions of the genes included in the inputted sets and, their recall and the determined FDR threshold.

Usage

```
ccr.VisDepAndSig(FCsprofile,SIGNATURES,TITLE='',
                 pIs=NULL,nIs=NULL,
                 th=0.05,plotFCprofile=TRUE)
```

Arguments

FCsprofile	A numerical vector containing gene average depletion log fold changes (or sgRNAs' depletion log fold changes) with names corresponding to HGNC symbols (or sgRNAs' identifiers).
SIGNATURES	A named list of vectors containing HGNC gene symbols. Two of these lists are used as classification template (respectively for positive and negative cases) to determine a log fold-change threshold providing a user defined classification false discovery rate.
TITLE	A string specifiying the title of the plot.
pIs	The index position of the signature that contains the positive cases of the classification template.
nIs	The index position of the signature that contains the negative cases of the classification template.
th	A numerical value specifying the desired classification false discovery rate (this must be a real number between 0 and 1).
plotFCprofile	A logic value specifying whether the log fold changes should be plotted.

Value

A named numerical vector containing recall scores for all the inputted signatures at the computed false discovery rate threshold for log fold-changes.

Author(s)

Francesco Iorio (iorio@gmail.com)

See Also

ccr.PrecisionRecallCurve

Examples

```
## loading corrected sgRNAs log fold-changes and segment annotations
## for an example cell line (EPLC-272H)
data(EPLC.272HcorrectedFCs)

## loading reference sets of essential and non-essential genes
data(BAGEL_essential)
data(BAGEL_nonEssential)

## loading other sets of core fitness genes
data(EssGenes.ribosomalProteins)
data(EssGenes.DNA_REPLICATION_cons)
data(EssGenes.KEGG_rna_polymerase)
data(EssGenes.PROTEASOME_cons)
data(EssGenes.SPLICEOSOME_cons)
```

```
## storing the sgRNA log fold changes into a name vector
FCs<-EPLC.272HcorrectedFCs$corrected_logFCs$avgFC
names(FCs)<-rownames(EPLC.272HcorrectedFCs$corrected_logFCs)

## loading sgRNA library annotation
data(KY_Library_v1.0)

## computing gene average log fold changes
FCs<-ccr.geneMeanFCs(FCs,KY_Library_v1.0)


## Assembling a named list with all the considered gene sets
SIGNATURES<-list(Ribosomal_Proteins=EssGenes.ribosomalProteins,
                 DNA_Replication = EssGenes.DNA_REPLICATION_cons,
                 RNA_polymerase = EssGenes.KEGG_rna_polymerase,
                 Proteasome = EssGenes.PROTEASOME_cons,
                 Spliceosome = EssGenes.SPLICEOSOME_cons,
                 CFE=BAGEL_essential,
                 non_essential=BAGEL_nonEssential)

## Visualising log fold change profile with superimposed signatures specifying
## that the reference gene sets are in positions 6 and 7
Recall_scores<-ccr.VisDepAndSig(FCsprofile = FCs,
                            SIGNATURES = SIGNATURES,
                            TITLE = 'EPLC-272H',
                            pIs = 6,
                            nIs = 7)

Recall_scores
```

`CL.subset` *COSMIC identifiers of 15 immortalised human cancer cell lines*

Description

COSMIC identifiers [1] of 15 cell lines included in the GDSC1000 panel [2] that are used in [3] to assess CRISPRcleaneR results.

Usage

```
data(CL.subset)
```

Format

A vector of strings.

References

[1] Forbes SA, Beare D, Boutselakis H, et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D777-D783,

[2] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, et al. *A landscape of pharmacogenomic interactions in cancer* Cell 2016 Jul 28;166(3):740-54

[3] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

Examples

```
data(CL.subset)

## Loading annotation for the GDSC1000 cell lines
data(GDSC.CL_annotation)

## Visualising annotation
GDSC.CL_annotation[CL.subset,]
```

`EPLC.272HcorrectedFCs` *CRISPRcleanR corrected data for an example cell line*

Description

This list contains corrected sgRNAs log fold-changes and segment annotations for an example cell line (EPLC-272H), obtained using the `ccr.GWclean` function, as detailed in its reference manual entry `ccr.GWclean`.

Usage

```
data("EPLC.272HcorrectedFCs")
```

Format

A list containing two data frames and a vector of strings. The first data frame (corrected_logFCs) contains a named row per each sgRNA and the following columns/header:

- `CHR`: the chromosome of the gene targeted by the sgRNA under consideration;
- `startp`: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;
- `endp`: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;
- `genes`: the HGNC symbol of the gene targeted by the sgRNA under consideration;
- `avgFC`: the log fold change of the sgRNA averaged across replicates;
- `correction`: the type of correction: 1 = increased log fold change, -1 = decreased log fold change. 0 indicates no correction;
- `correctedFC`: the corrected log fold change of the sgRNA

.

The second data frame (segments) contains the identified region of estimated equal log fold changes (one region per row) and the following columns/headers:

- `CHR`: the chromosome of the gene targeted by the sgRNA under consideration;
- `startp`: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;

- endp: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;
- genes: the HGNC symbol of the gene targeted by the sgRNA under consideration;
- avgFC: the log fold change of the sgRNA averaged across replicates;
- correction: the type of correction: 1 = increased log fold change, -1 = decreased log fold change. 0 indicates no correction;
- correctedFC: the corrected log fold change of the sgRNA

.

The second data frame (segments) contains the identified region of estimated equal log fold changes (one region per row) and the following columns/headers:

- CHR: the chromosome of the region under consideration;
- startp: the genomic coordinate of the starting position of the region under consideration;
- endp: the genomic coordinate of the ending position of the region under consideration;
- n.sgRNAs: the number of sgRNAs targeting sequences in the region under consideration;
- avg.logFC: the average log fold change of the sgRNAs in the region;
- guideIdx: the indexes range of the sgRNAs targeting the region under consideration as they appear in the gwSortedF Cs provided in input.

The string of vectors (SORTED_sgRNAs) contains the sgRNAs' identifiers in the same order as they are reported in the gwSortedFCs data frame inputted to the ccr.GWclean function.

Examples

```
data(EPLC.272HcorrectedFCs)
head(EPLC.272HcorrectedFCs$corrected_logFCs)
head(EPLC.272HcorrectedFCs$segments)
head(EPLC.272HcorrectedFCs$SORTED_sgRNAs)
```

EssGenes.DNA_REPLICATION_cons

Core Fitness essential genes involved in DNA replication

Description

List of core fitness essential genes involved in DNA replication assembled by merging together multilpe DNA replication signatures from MSigDB [1] as detailed in [2].

Usage

```
data("EssGenes.DNA_REPLICATION_cons")
```

Format

A vector of strings containing HGNC symbols.

References

[1] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545-15550. http://doi.org/10.1073/pnas.0506580102

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

Examples

```
data(EssGenes.DNA_REPLICATION_cons)
head(EssGenes.DNA_REPLICATION_cons)
```

`EssGenes.KEGG_rna_polymerase`

Core Fitness essential rna polymerase genes

Description

List of core fitness essential rna polymerase genes downloaded from MSigDB [1].

Usage

```
data("EssGenes.KEGG_rna_polymerase")
```

Format

A vector of strings containing HGNC symbols.

References

[1] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545-15550. http://doi.org/10.1073/pnas.0506580102

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

Examples

```
data(EssGenes.KEGG_rna_polymerase)
head(EssGenes.KEGG_rna_polymerase)
```

EssGenes.PROTEASOME_cons

Core Fitness essential proteasome genes

Description

List of core fitness essential proteasome genes assembled by merging together multilpe DNA replication signatures from MSigDB [1] as detailed in [2].

Usage

```
data("EssGenes.PROTEASOME_cons")
```

Format

A vector of strings containing HGNC symbols.

References

[1] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545-15550. http://doi.org/10.1073/pnas.0506580102

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

Examples

```
data(EssGenes.PROTEASOME_cons)
head(EssGenes.PROTEASOME_cons)
```

EssGenes.ribosomalProteins

Core Fitness essential genes coding for ribosomal proteins

Description

List of core fitness essential coding for ribosomal proteins curated from [1].

Usage

```
data("EssGenes.KEGG_rna_polymerase")
```

Format

A vector of strings containing HGNC symbols.

References

[1] Yoshihama, M. et al. The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. Genome Res. 12, 379-390 (2002)

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

Examples

```
data(EssGenes.ribosomalProteins)
head(EssGenes.ribosomalProteins)
```

`EssGenes.SPLICEOSOME_cons`

Core Fitness essential spliceosome genes

Description

List of core fitness essential spliceosome genes assembled by merging together multilpe DNA replication signatures from MSigDB [1] as detailed in [2].

Usage

```
data("EssGenes.SPLICEOSOME_cons")
```

Format

A vector of strings containing HGNC symbols.

References

[1] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 1554515550. http://doi.org/10.1073/pnas.0506580102

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

Examples

```
data(EssGenes.SPLICEOSOME_cons)
head(EssGenes.SPLICEOSOME_cons)
```

GDSC.CL_annotation *Tissue type and other annotations for 1,001 human cancer cell lines*

Description

Tissue type and other annotations for 1,001 human cancer cell lines

Usage

```
data(GDSC.CL_annotation)
```

Format

A data frame with 1,001 observations of the following 7 variables.

`CL.name` Cell line name;

`COSMIC.ID` Cosmic identifier of the cell line;

`GDSC.description_1` Tissue descriptor (Genomics of Drug Sensitivity in Cancer - Level 1);

`GDSC_description_2` Tissue descriptor (Genomics of Drug Sensitivity in Cancer - Level 2);

`'TCGA type'` Manaually curated matched TCGA cancer type;

`MMR` Microsatellite instability status (MSI-S = Stable, MSI-L = Instable, MSI-H = highly-Instable).

Source

This data frame has been derived from the xls table available at http://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources//Data/suppData/TableS1E.xlsx.

References

[1] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, et al. A landscape of pharmacogenomic interactions in cancer Cell 2016 Jul 28;166(3):740-54

Examples

```
data(GDSC.CL_annotation)
head(GDSC.CL_annotation)
```

GDSC.geneLevCNA *Genome-wide copy number data for 15 human cancer cell lines.*

Description

Genome-wide copy number data derived from PICNIC analysis of Affymetrix SNP6 segmentation data (EGAS00001000978, part of the Genomics of Drug Sensitivity in 1,000 Cancer Cell Lines (GDSC1000) panel [1]) for 15 cell lines used in [2] to assess CRISPRcleaneR results.

Usage

```
data(GDSC.geneLevCNA)
```

Format

A data frame with HGNC gene symbols on the row cancer cell lines' cosmic identifiers on the columns. The entry in position *i,j* indicates the copy number status of gene *i* in cell line *j*.

Details

Each entry of the data frame is a string made of four comma seperated peices of data (`n1,n2,n3,n4`), hyphen (`-`) is used when the corresponding data is unknown.

The four values indicate:

- `n1`: Maximum copy number of any genomic segment containing coding sequence of the gene (`-1` indicates a value could not be assigned).
- `n2`: Minimum copy number of any genomic segment containing coding sequence of the gene (`-1` indicates a value could not be assigned).
- `n3`: Zygosity - (`H`) if all segments containing gene sequence are heterozygous, (`L`) if any segment containing coding sequence has LOH, (`0`) if the complete coding sequence of the gene falls within a homozygous deletion.
- `n4`: Disruption (`D`) if the gene spans more than 1 genomic segment (`-`) if no disruption occures.

Source

This data frame has been derived from the xls table available at `ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/release-6.0/Gene_level_CN.xlsx`.

References

[1] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, et al. *A landscape of pharmacogenomic interactions in cancer* Cell 2016 Jul 28;166(3):740-54

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

Examples

```
data(GDSC.geneLevCNA)
GDSC.geneLevCNA[1:10,1:10]
```

HT.29correctedFCs *CRISPRcleanR corrected data for an example cell line*

Description

This list contains corrected sgRNAs log fold-changes and segment annotations for an example cell line (HT-29), obtained using the `ccr.GWclean` function, as detailed in its reference manual entry `ccr.GWclean`.

Usage

```
data("HT.29correctedFCs")
```

Format

A list containing two data frames and a vector of strings. The first data frame (corrected_logFCs) contains a named row per each sgRNA and the following columns/header:

- `CHR`: the chromosome of the gene targeted by the sgRNA under consideration;
- `startp`: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;
- `endp`: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;
- `genes`: the HGNC symbol of the gene targeted by the sgRNA under consideration;
- `avgFC`: the log fold change of the sgRNA averaged across replicates;
- `correction`: the type of correction: 1 = increased log fold change, -1 = decreased log fold change. 0 indicates no correction;
- `correctedFC`: the corrected log fold change of the sgRNA

.
The second data frame (segments) contains the identified region of estimated equal log fold changes (one region per row) and the following columns/headers:

- `CHR`: the chromosome of the gene targeted by the sgRNA under consideration;
- `startp`: the genomic coordinate of the starting position of the region targeted by the sgRNA under consideration;
- `endp`: the genomic coordinate of the ending position of the region targeted by the sgRNA under consideration;
- `genes`: the HGNC symbol of the gene targeted by the sgRNA under consideration;
- `avgFC`: the log fold change of the sgRNA averaged across replicates;
- `correction`: the type of correction: 1 = increased log fold change, -1 = decreased log fold change. 0 indicates no correction;
- `correctedFC`: the corrected log fold change of the sgRNA

.
The second data frame (segments) contains the identified region of estimated equal log fold changes (one region per row) and the following columns/headers:

- `CHR`: the chromosome of the region under consideration;
- `startp`: the genomic coordinate of the starting position of the region under consideration;
- `endp`: the genomic coordinate of the ending position of the region under consideration;
- `n.sgRNAs`: the number of sgRNAs targeting sequences in the region under consideration;
- `avg.logFC`: the average log fold change of the sgRNAs in the region;
- `guideIdx`: the indexes range of the sgRNAs targeting the region under consideration as they appear in the gwSortedF Cs provided in input.

The string of vectors (SORTED_sgRNAs) contains the sgRNAs' identifiers in the same order as they are reported in the gwSortedFCs data frame inputted to the `ccr.GWclean` function.

Examples

```
data(HT.29correctedFCs)
head(HT.29correctedFCs$corrected_logFCs)
head(HT.29correctedFCs$segments)
head(HT.29correctedFCs$SORTED_sgRNAs)
```

KY_Library_v1.0 *sgRNAs' genome-wide annotation for the Sanger sgRNA pooled Library v1.0*

Description

A data frame with a named row for each sgRNA of the Sanger sgRNA pooled library presented in [1] including annotations such as targeted genes, and genomic coordinates.

Usage

```
data("KY_Library_v1.0")
```

Format

A a row named data frame with 90709 observations (one for each sgRNA) of the following 7 variables.

`CODE` alphanumerical identifier of the sgRNAs;

`GENES` targeted gene;

`EXONE` exone of the targeted genomic region (string with 'ex' prefix followed by the exone number);

`CHRM` chromosome of where the targeted region resides (string)

`STRAND` targeted DNA strand ('+' or '-')

`STARTpos` starting genomic coordinate of the targeted genomic region (numeric);

`ENDpos` ending genomic coordinate of the targeted genomic region (numeric).

References

[1] Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, Mupo A, Grinkevich V, Li M, Mazan M, Gozdecka M, Onishi S, Cooper J, Patel M, McKerrell T, Chen B, Domingues AF, Gallipoli P, Teichmann S, Ponstingl H, McDermott U, Saez-Rodriguez J, Huntly BJP, Iorio F, Pina C, Vassiliou GS, Yusa K. *A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukaemia*. Cell Reports 2016 Oct 18;17(4):1193-1205

Examples

```
data(KY_Library_v1.0)
head(KY_Library_v1.0)
```

RNAseq.fpkms *RNAseq derived genome-wide basal expression profiles for 15 cell lines.*

Description

Genome-wide substitute reads with fragments per kilobase of exon per million reads mapped (FPKM) for the 15 cell lines specified in CL.subset, derived from a comprehensive collection of RNAseq profiles described in [1] and used in [2] to assess CRISPRcleaneR results.

Usage

```
data(RNAseq.fpkms)
```

Format

A data frame with one bservations per gene and one variable per cell line. Row names indicates HGNC symbols and column names indicate cell line COSMIC identifiers [3].

References

[1] Garcia-Alonso L, Iorio F, Matchan A, et al. *Transcription factor activities enhance markers of drug response in cancer* doi: https://doi.org/10.1101/129478

[2] Iorio, F., Behan, F. M., Goncalves, E., Beaver, C., Ansari, R., Pooley, R., et al. (n.d.). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. http://doi.org/10.1101/228189

[3] Forbes SA, Beare D, Boutselakis H, et al. *COSMIC: somatic cancer genetics at high-resolution* Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D777-D783,

See Also

CL.subset

Examples

```
data(RNAseq.fpkms)
head(RNAseq.fpkms)
```

Index