Package 'HT29benchmark'

January 10, 2020

Type Package

Title HT29 CRISPR-Cas9 pooled screen data and metrics to benchmark experimental pipelines

Version 0.1.0

Author Ichcha Manipur & Francesco Iorio

Maintainer Francesco Iorio <Fi9323@gmail.com>

Description R package for benchmarking genome-wide CRISPR-Cas9 knock-out viability screening pipelines making use of a reference dataset from six high-quality screens of the HT29 cell line

License GPL-2

Depends CRISPRcleanR, stringr, RColorBrewer

Encoding UTF-8

LazyData true

R topics documented:

`HT29R.download_ref_dataset`

Download reference HT-29 screens data

Description

This function allows downloading reference datasets from high-quality CRISPR-Cas9 pooled screens of the HT-29 cell line with the KY sgRNA library[1]. This data has been generated through the experimental pipeline described in [2] and it is also public aviable on the Project Score web-site (https://score.depmap.sanger.ac.uk/downloads), part of the Cancer Dependency Map portfolio of tools and resources at the Wellcome Sanger Institute (https://depmap.sanger.ac.uk/).

Usage

```
H29R.download_ref_dataset(
      whatToDownload = "FCs",
      destFolder = "./",
    dataRepoURL = "https://cog.sanger.ac.uk/cmp/downloads/crispr_cas9_benchmark/",
      expNames=c("HT29_c903","HT29_c904","HT29_c905",
                 "HT29_c906","HT29_c907","HT29_c908"))
```

Arguments

whatToDownload	String parameter specifying what type of data to dowload. Possible values are "rawCounts" for plain .tsv files containing raw sgRNA countsare or "FCs" (default) for R objects containing sgRNA normalised depletion fold-changes: data frames in which the first two columns contain sgRNAs' identifiers and HGNC symbols of targete gene, followed by one column per screen replicate containing sgRNAs' fold-changes;
destFolder	String specifying where the dataset should be saved;
dataRepoURL	The URL of the data repository;
expNames	A vector of strings specifying the experiment names for the dataset to download.

Author(s)

Ichcha Manipur & Francesco Iorio (fi1@sanger.ac.uk)

References

[1] Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, et al. A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. Cell Rep. 2016;17:1193–205.

[2] Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature. 2019;568:511–6.

Examples

```
#### creating a temporary directory
dir.create('tempDir')

#### downloading reference sgRNA depletion fold-changes from high-quality
#### HT-29 screens into the temporary directory
HT29R.download_ref_dataset(destFolder = 'tempDir')
```

HT29R.expNames *Benchmark Experiment Names*

Description

Labels of individual HT-29 screen experiments

Usage

```
data("HT29R.expNames")
```

Format

A vector of 6 strings, containing each the name of one experiment.

Examples

```
data(HT29R.expNames)
print(HT29R.expNames)
```

HT29R.prSCORE_rCorr_Reprod

Pair-wise screen replicate correlations and background correlations from Project Score.

Description

Correlation scores obtained by comparing profiles of KY-Library[1] specific informative/reproducible sgRNAs depletion fold-changes between replicates of the same experiment, and between all possible pairs of individual replicates across experiments, from Project Score[2],

Usage

```
data("HT29R.replicateCountCorrelationReprod")
```

Format

A list of two numerical vectors:

`BGscores` a numeric vector with 882774 entries, containing the correlation scores obtained by comparing profiles of KY-Library[1] specific informative/reproducible sgRNAs depletion fold-changes between all possible pairs of individual replicates across experiments, i.e. background correlation.

`REPscores` a numeric vector with 1766 entries, containing the correlation scores obtained by comparing profiles of KY-Library[1] specific informative/reproducible sgRNAs depletion fold-changes between replicates of the same experiment.

References

[1] Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, et al. A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. Cell Rep. 2016;17:1193–205.

[2] Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature. 2019;568:511–6.

Examples

```
library(CRISPRcleanR)
data(HT29R.prSCORE_rCorr_Reprod)

ccr.multDensPlot(
  list(density(HT29R.prSCORE_rCorr_Reprod$BGscores),
       density(HT29R.prSCORE_rCorr_Reprod$REPscores)),
```

```
  XLIMS = c(0,1),
  TITLE = 'Observed vs. Expected replicate correlations from project Score\n',
  COLS = c('gray','darkgreen'),LEGentries = c('expected','observed'),XLAB='R')
```

`HT29R.replicateCorr_Pscore`
Screen reproducibility assessment using Project Score criteria

---

Usage

```
HT29R.replicateCorr_Pscore(refDataDir = "./",
                           resDir = "./",
                           userFCs = NULL)
```

Arguments

`refDataDir`	Reference HT29 dataset directory: a string specifying the location of the processed HT29 reference dataset.
`resDir`	Output directory: a string specifying the directory where the output of this function (a pdf file with multiple plots) should be saved.
`userFCs`	Data from a user performed screen: A data frame with the same format of the R ojbects composing the reference dataset, i.e. first two columns containing sgRNAs' identifiers and HGNC symbols of targete gene (headers = sgRNA and gene, respectivley), followed by one column per screen replicate containing sgRNAs' fold-changes.

Details

This function computes correlation scores between each pair of screen replicates for the HT29 reference dataset as well as for used defined data. This is performed as for the Project Score data [1]: considering only a set of 838 most informative sgRNAs (in the `HT29R.reproducible_GeneGuides object`), defined as those targeting the same genes and with an average pairwise Pearson's correlation > 0.6 between corresponding patterns of depletion fold-changes (FCs) across hundreds of screened cell lines. Per construction, the depletion patterns of these sgRNAs are both reproducible and informative (as they involve genes carrying an actual fitness signal). Computing correlation scores between replicates of the same screen on the domain of these sgRNAs only allows estimating a null distribution of replicate correlations and computing a reproducibility threshold defined as the minimal correlation score that should be observed between replicates of the same screen (R = 0.68, using the `HT29R.prSCORE_rCorr_Reprod` object). This is performed as genome-wide correlation scores computed between replicates of the same CRISPR-Cas9 pooled genome-wide viability screen are generally always very high and indistiguishable from expectation due to only a small percentage of genes exerting an effect on cellular fitness upon knock-out.

This function produces a pdf file (named `RepCor_Vs_PrScore.pdf`) in the specified directory with a plot of the expected/observed distributions of replicate pair-wise correlation scores and reproducibility threshold from the Project Score dataset. Below this plots the pair-wise correlation scores computed between replicates of the HT29 reference dataset are also plotted. If the user provides data from his own screen (trough the `userFCs` parameter) then pair-wise correlation scores computed between replicates of this screens are overlaied on the distribution plot. In this case, a matrix with these scores and a specification of how many of them are equal or greater than the replication threshold are printed in the consolle.

Author(s)

Ichcha Manipur & Francesco Iorio (fi1@sanger.ac.uk)

References

[1] Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature. 2019;568:511–6.

See Also

H29R.download_ref_dataset, HT29R.reproducible_GeneGuides, HT29R.prSCORE_rCorr_Reprod

Examples

```
## Creating a temporary folder to store the HT29 reference dataset
## and the pdf created by this function
dir.create('tmpdir')

## Downloading the HT29 reference dataset in the temporary folder
HT29R.download_ref_dataset(destFolder = 'tmpdir')

## Loading CRISPRcleanR library to use example screen data
library(CRISPRcleanR)

## 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 library Annotation
data('KY_Library_v1.0')

## Loading, median-normalizing and computing fold-changes for the example dataset
normANDfcs<-ccr.NormfoldChanges(fn,min_reads=30,EXPname='ExampleScreen',
                                libraryAnnotation = KY_Library_v1.0,
                                display = FALSE)
ExampleScreen<-normANDfcs$logFCs

## Evaluating screen reproducibility of HT29 reference and user defined data
## both, using Project Score criteria
HT29R.replicateCorr_Pscore(refDataDir = 'tmpDir',resDir = 'tmpDir',userFCs = ExampleScreen)

## Checking results
system2('open', args = 'tmpdir/RepCor_Vs_PrScore.pdf', wait = FALSE)

## Removing Example dataset processed files
file.remove('ExampleScreen_foldChanges.Rdata')
file.remove('ExampleScreen_normCounts.Rdata')
```

HT29R.reproducible_GeneGuides

Library specific informative/reproducible sgRNAs.

Description

838 KY-library[1] specific informative/reproducible sgRNAs (targeting 308 genes) for evaluating CRISPR-Cas9 pooled genome-wide viability screen replicates.

Usage

```
data(HT29R.reproducible_GeneGuides)
```

Format

A vector of strings with entries corresponding to sgRNA identifiers.

Details

Genome-wide correlation scores computed between replicates of the same CRISPR-Cas9 pooled genome-wide viability screen are generally always very high and indistiguishable from expectation due to only a small percentage of genes exerting an effect on cellular fitness upon knock-out. In [2] we have selected a set of 838 most informative sgRNAs, defined as those targeting the same genes and with an average pairwise Pearson's correlation > 0.6 between corresponding patterns of depletion fold-changes (FCs) across hundreds of screened cell lines. Per construction, the depletion patterns of these sgRNAs are both reproducible and informative (as they involve genes carrying an actual fitness signal). Computing correlation scores between replicates of the same screen on the domain of these sgRNAs only allowed the estimation of a null distribution of replicate correlations and computing a reproducibility threshold defined as the minimal correlation score that should be observed between replicates of the same screen (R = 0.68).

References

[1] Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, et al. A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. Cell Rep. 2016;17:1193–205.

[2] Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature. 2019;568:511–6.

Examples

```
data(HT29R.reproducible_GeneGuides)
head(HT29R.reproducible_GeneGuides)
```

Index