HT-29 CRISPR-Cas9 screen assessment

Raffaele M. Iannuzzi, Francesco Iorio

2022-09-06

Introduction

This document guides the user through (and shows results from) the execution of functions included in the *HT29benchmark* R package. The aim of these functions is to assess novel experimental pipelines for genome-wide CRISPR-Cas9 screens upon the execution of a single calibration screen of the HT-29 human colon cancer cell line (https://www.cellosaurus.org/CVCL_0320) employing a commercially available genome-wide library of single guide RNAs (the Sanger library) (Tzelepis et al. (2016)) (AddGene: 67989) and setting described in Behan et al. (2019).

The code provided can be executed to reproduce the figures of the manuscript describing the HT-29 reference dataset and the *HT29benchmark* package (currently under review) and to assess a user-provided calibration screens. Whether to employ the example data or novel input data for the quality assessment can be determined via commenting/un-commenting specific code portions as described in this document.

The user-provided screen is evaluated through different metrics, and outcomes are then contrasted with those obtained analysing a high-quality screen of the HT-29 cell line, screened in multiple batches with the Sanger library (the HT-29 reference dataset).

The HT29benchmark R package is available at: https://github.com/DepMap-Analytics/HT29benchmark; with user reference manual available at https://github.com/DepMap-Analytics/HT29benchmark/blob/master/HT29benchmark.pdf. The HT-29 reference dataset can be downaloded through a dedicated function of the *HT29benchmark* package, as well is available on FigShare (Behan M., Iorio, and Garnett J. (2022)).

Environment preparation and data retrival

The following code loads all required libraries installing the missing ones from Bioconductor and CRAN:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

toInstall<- setdiff(c("topGO", "clusterProfiler","org.Hs.eg.db","enrichplot"),
                    rownames(installed.packages()))

BiocManager::install(toInstall)

toInstall<- setdiff(c("VennDiagram", "data.table", "KernSmooth"),
                    rownames(installed.packages()))

install.packages(toInstall)
```

```
# Needed for the analysis
library(CRISPRcleanR)
library(HT29benchmark)
```

```
# Needed for the report
library(data.table)
library(VennDiagram)
library(clusterProfiler)
library(enrichplot)
library(org.Hs.eg.db)
library(topGO)
library(RColorBrewer)
```

The following code creates a directory (HT29R_resFolder) in the local folder, in which the HT-29 reference dataset (sgRNA depletion fold changes or sgRNA counts) are downloaded. A subdirectory is also created (USER) and used to save plots and other figures (if the saveToFig parameter of the HT29benchmark functions is set to 'TRUE').

```
dir.create('~/HT29R_resFolder/')
tmpDir <- path.expand('~/HT29R_resFolder/')
dir.create(paste(tmpDir, "USER/",sep=""))
resultsDir <- paste(tmpDir,"USER/", sep="")
```

The following code downloads the HT-29 reference dataset (i.e. sgRNA depletion log fold-changes from high-quality HT-29 screens) and stores it in the HT29_resFolder directory.
Additionally, individual reference screens file names are stored in the ref_fn variable

```
HT29R.downloadRefData(destFolder = tmpDir, whatToDownload = 'FCs')
```

```
ref_fn <- dir(tmpDir)
ref_fn <- grep('_foldChanges.Rdata',ref_fn,value=TRUE)
```

The code below downloads a demo screen in the 'HT29R_resFolder.' This encompasses data from a 6-replicates mid-quality screen of the HT-29 cell line, employing the Sanger library (Tzelepis et al. 2016) and using settings described in Behan et al. (2019), available on FigShare (Behan M., Iorio, and Garnett J. (2022)).
IMPORTANT: The following code should not be executed to perform the analysis of real user-data. See further code chunk.

```
URL <- 'https://figshare.com/ndownloader/files/36658530?private_link=5b2a579791c47a417474'
download.file(URL, destfile = paste0(tmpDir, '/Example_UserScreen.tsv'))
```

```
userDataPATH <- paste0(tmpDir, '/Example_UserScreen.tsv')
```

IMPORTANT: The following code should be un-commented and executed for the analysis of real user-data. The path to the real user dataset should be provided. This should be string specifying the path to a tsv file containing the raw sgRNA counts of a HT-29 calibration screen performed with the Sanger library (Tzelepis et al. 2016) and the experimental settings described in Behan et al. (2019)
The file should be tab delimited, it should contain one row per sgRNA and the following columns:

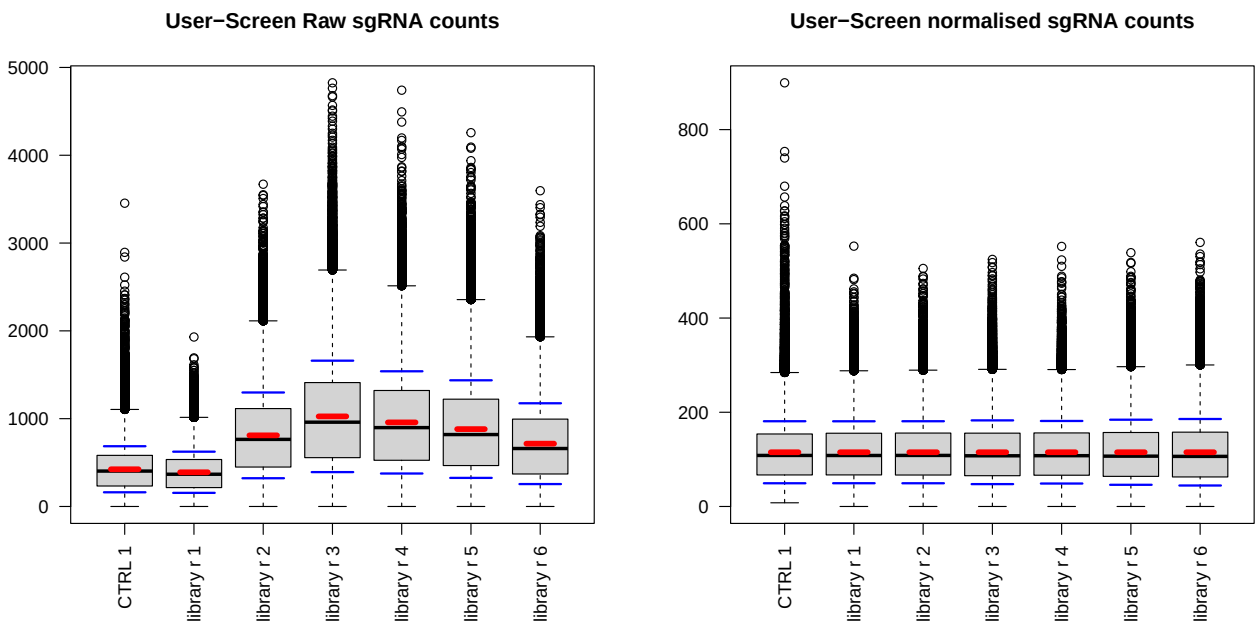
- sgRNA: column with alphanumerical identifiers of the sgRNA under consideration;
- gene: column with HGNC symbols of the genes targeted by the sgRNA under consideration;
- columns containing the sgRNAs' counts for the controls;
- columns for library transfected samples (one column per replicate).

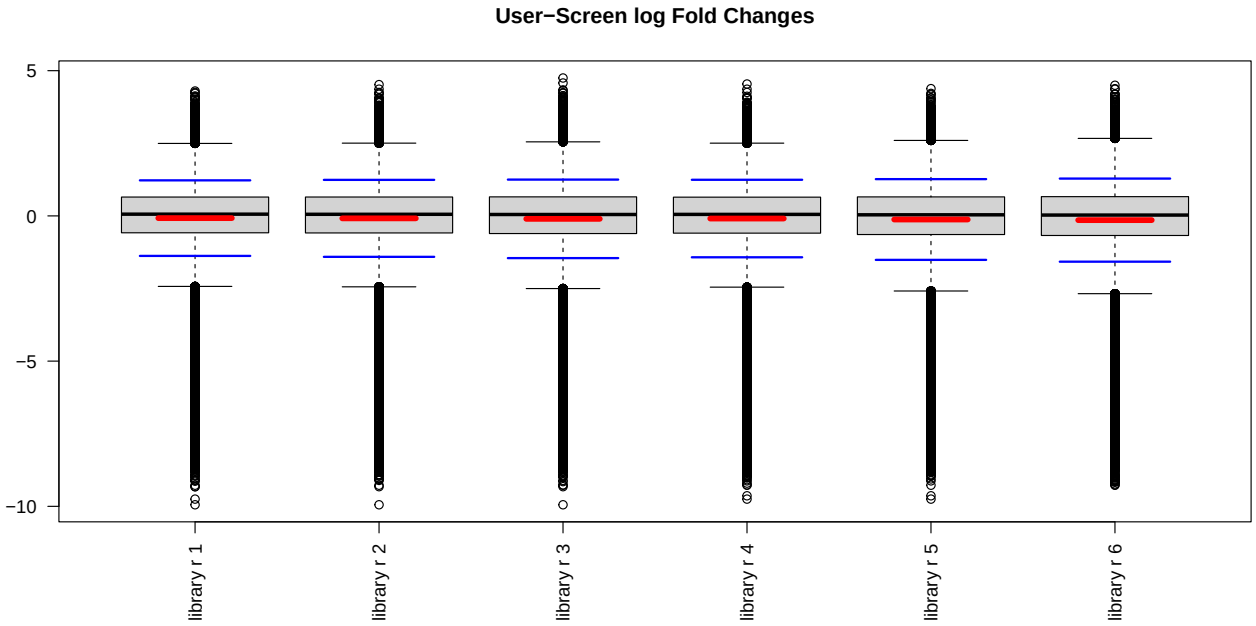
```
## userDataPATH <- 'PATH/TO/REAL-USER-DATA/userdata.tsv'
```

User data normalisation and computation of depletion log fold-changes

The following code normalises user-provided sgRNA counts, computes log fold-changes, stores results in the USER subfolder and produces the plots below with distributions of sgRNA counts pre-/post-normalisation and depletion log fold-changes across samples. The ccr.NormfoldChanges function comes from our previously published CRISPRcleanR package (Iorio et al. (2018)). Results are also stored in the UserData variable. This is a list containing two data frames respectively including the normalised sgRNAs' counts (norm_counts) and the sgRNAs' log fold-changes (logFCs). The first two columns in these data frames contain sgRNAs' identifiers and HGNC symbols of targeted genes, respectively.

```
data('KY_Library_v1.0')
UserData <- ccr.NormfoldChanges(filename = userDataPATH,
                                Dframe = NULL,
                                min_reads = 30,
                                EXPname = "User-Screen",
                                libraryAnnotation = KY_Library_v1.0,
                                saveToFig = FALSE,
                                outdir = resultsDir,
                                display = TRUE)
```

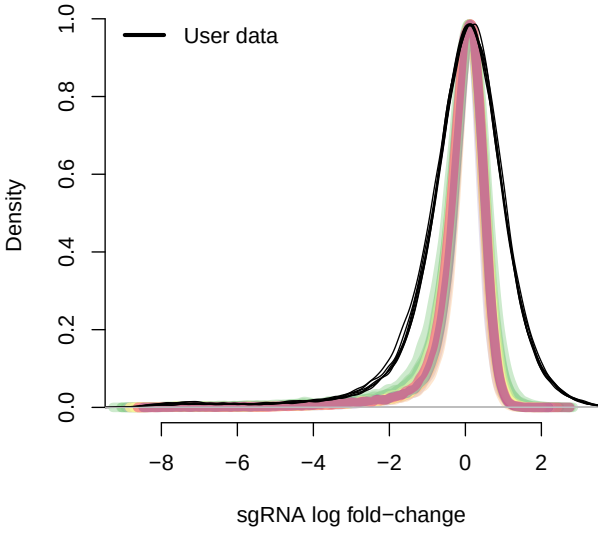


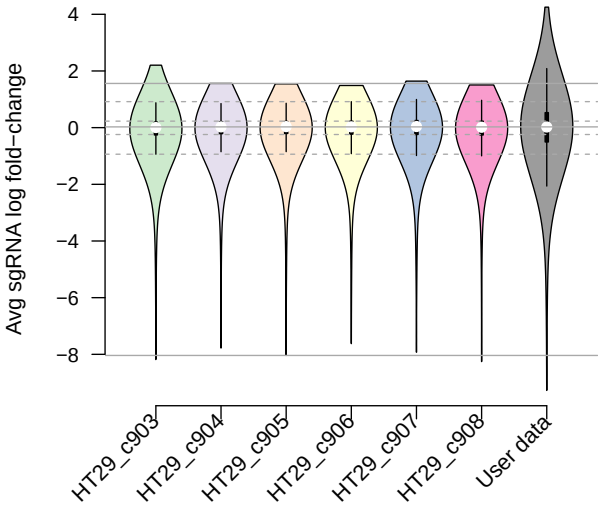


Inspection of sgRNA depletion log fold-change distributions

The following code allows visualising (through dedicated plots) distributions of sgRNA log fold-changes from the HT-29 reference dataset (across screens) as well as user screen data.
Average values and confidence intervals of these distribution are also plotted and can be compared.

```
HT29R.FCdistributions(refDataDir = tmpDir,
                      resDir = resultsDir,
                      userFCs = UserData$logFCs,
                      stats = TRUE,
                      saveToFig = FALSE,
                      display = TRUE)
```






```
## HT-29 reference dataset, sgRNAs logFCs statistics:
##
## Avg. Range: -7.97±0.099 ; 1.66±0.111
```

```
## Avg. Median: 0.025±0.006
## Avg. IQR range: -0.23±0.01 ; 0.22±0.01
## Avg. 10-90th perc range: -0.68±0.02 ; 0.39±0.01
## Avg. Skewness: -3.81±0.04
## Avg. Kurtosis: 19.53±0.55
##
## User screen, sgRNAs logFCs statistics:
##
## Range min: -9.273 ; Range max: 4.253
## Median: 0.022
## IQR min: -0.506 ; IQR max: 0.531
## 10th perc: -1.15 ; 90th perc: 1.017
## Skewness: -2.089
## Kurtosis: 8.673
```

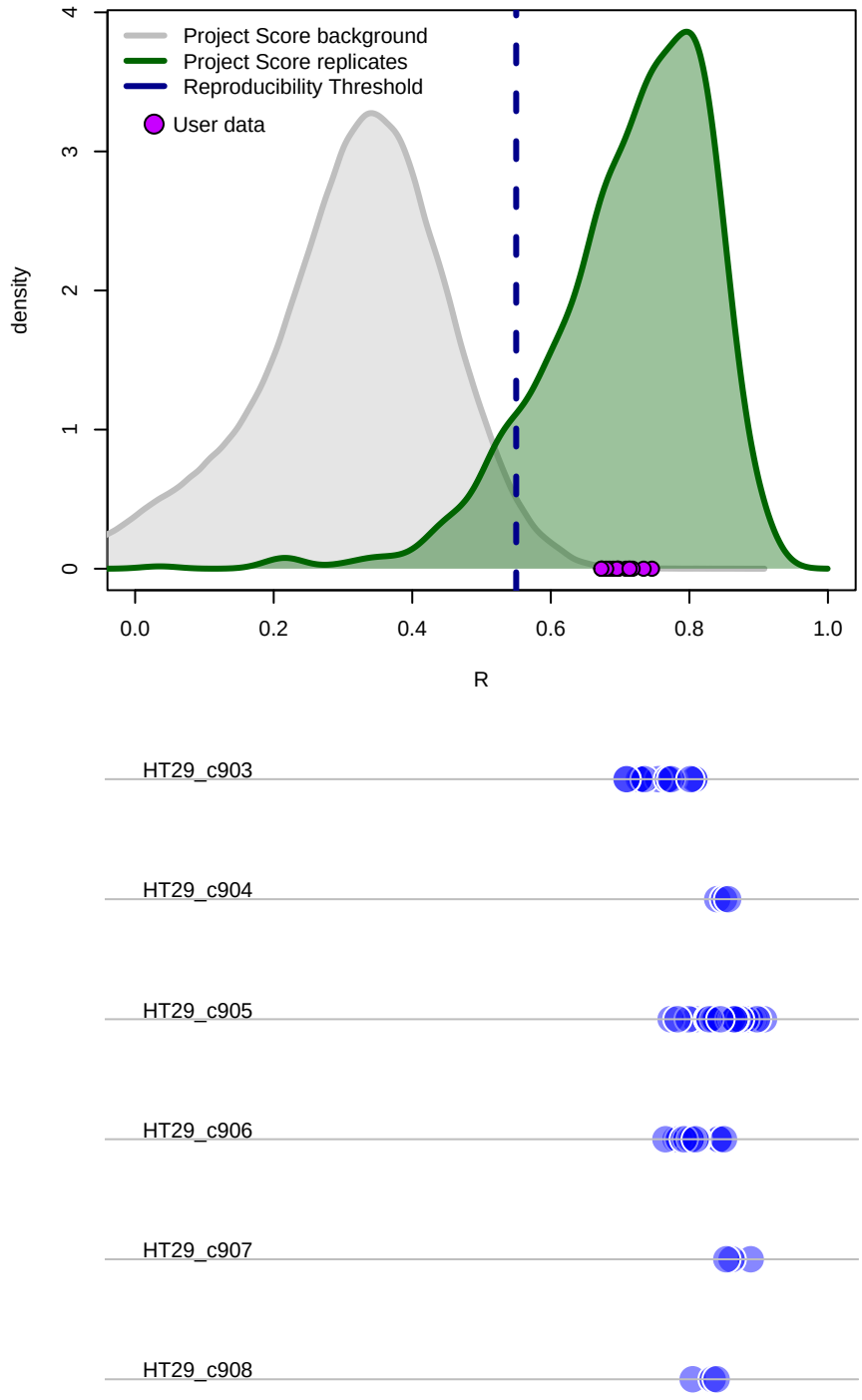
Intra-screen reproducibility

The following code evaluates and compares intra-screen replicate similarities for the HT-29 reference screen, as well as for the user-provided screen. Particularly, it shows Pearson's correlation scores computed between sgRNA depletion log fold-change profiles of replicates for each of the six HT-29 reference screens (blue dots), and between replicates of the user-provided screen (pink dots). These correlation scores are computed considering only depletion log fold-changes of highly reproducible/informative sgRNAs (defined as in Behan et al. (2019)). These scores are evaluated through comparing them with:

- correlation scores obtained comparing replicates of the same screen across > 200 cell lines (from Project Score (Behan et al. (2019), Dwane et al. (2021)), in green);
- correlation scores obtained from comparing each possible pair of replicates, regardless the screen (again from Project Score (Behan et al. (2019), Dwane et al. (2021)), in gray);
- a quality threshold, derived from these two distributions, as defined in Behan et al. (2019).

```
HT29R.evaluateReps(refDataDir = tmpDir,
                   resDir = resultsDir,
                   userFCs = UserData$logFCs,
                   geneLevel = FALSE,
                   display = TRUE,
                   saveToFig = FALSE)
```

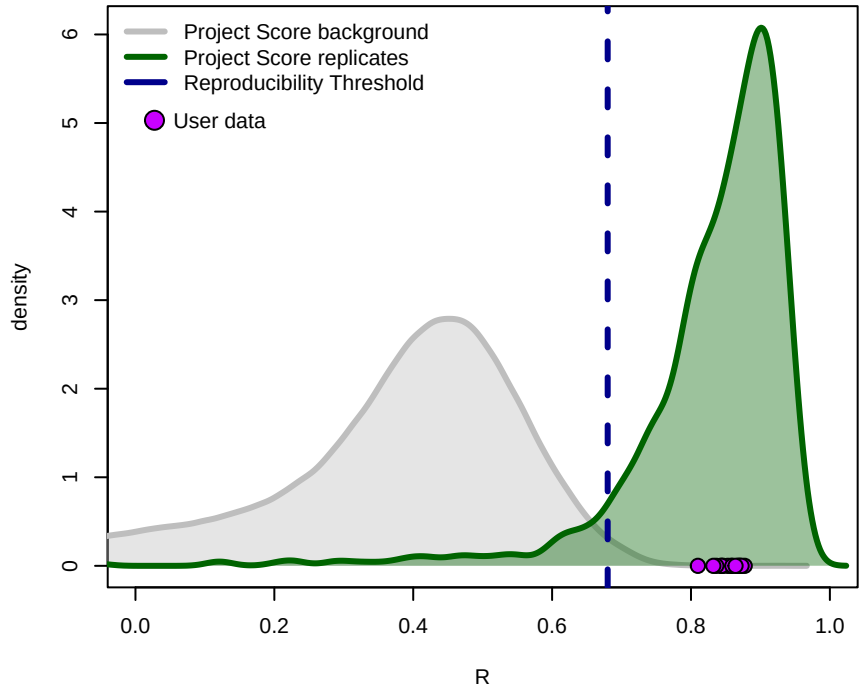
```
## User screen results:
## 15 pair-wise replicate comparisons (out of 15) yield correlation scores greater or
## equal than Project Score QC threshold.
## 0 pair-wise replicate comparisons (out of 15) yield correlation scores lower
## than Project Score QC threshold
```

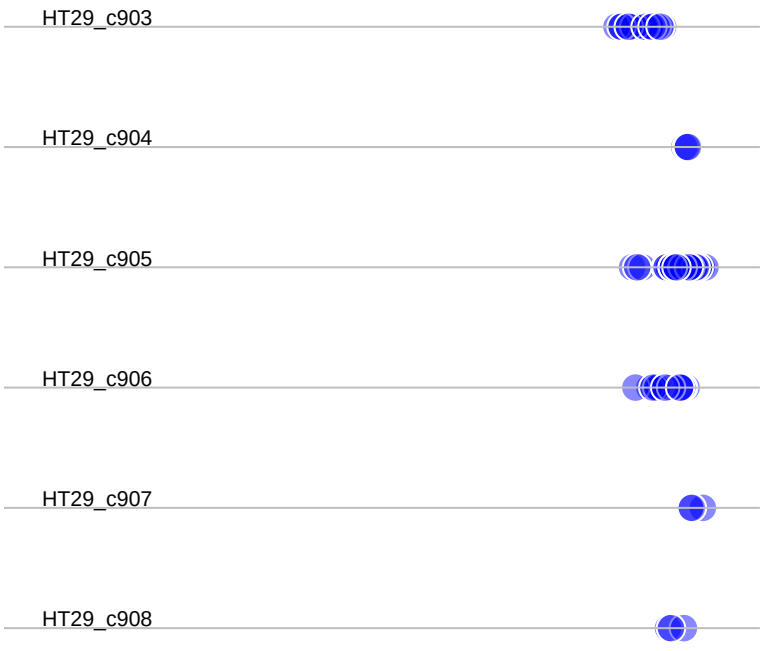


The following code chunk performs the same comparisons of the previous one, but this time at the gene level, collapsing sgRNA depletion log fold-changes by averaging on a targeted gene basis.

```
HT29R.evaluateReps(refDataDir = tmpDir,
                   resDir = resultsDir,
                   userFCs = UserData$logFCs,
                   geneLevel = TRUE,
                   display = TRUE,
                   saveToFig = FALSE)
```

```
## User screen results:
## 15 pair-wise replicate comparisons (out of 15) yield correlation scores greater or
## equal than Project Score QC threshold.
## 0 pair-wise replicate comparisons (out of 15) yield correlation scores lower
## than Project Score QC threshold
```





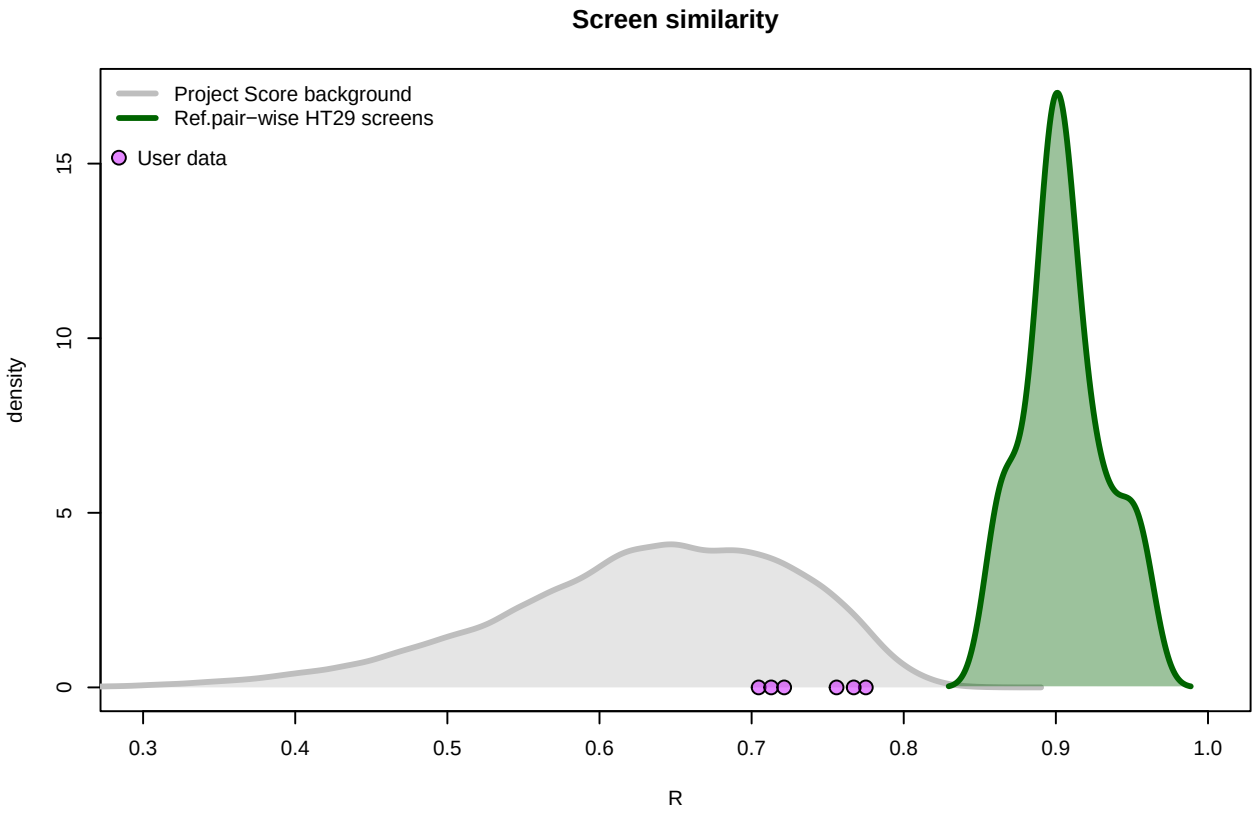
Inter-screen similarity

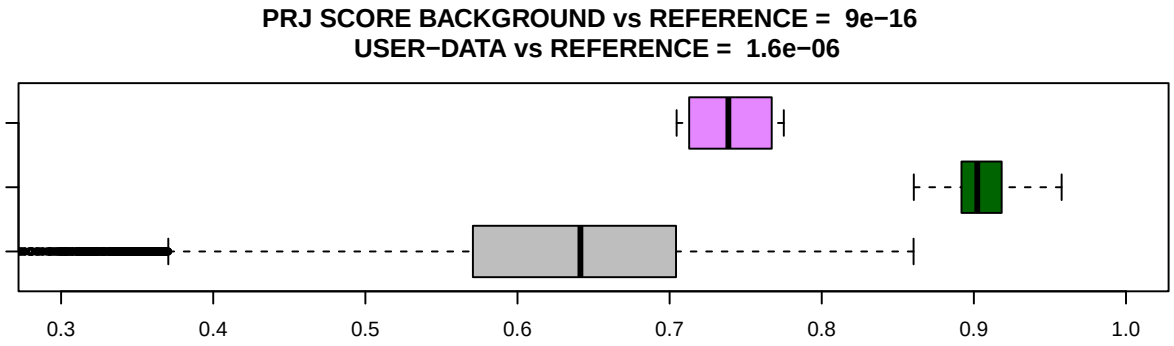
Inter-screen similarities are then evaluated for the HT-29 reference screens (green distribution), as well as for the user-provided screen (pink dots). Particularly, figures show Pearson's correlation scores computed

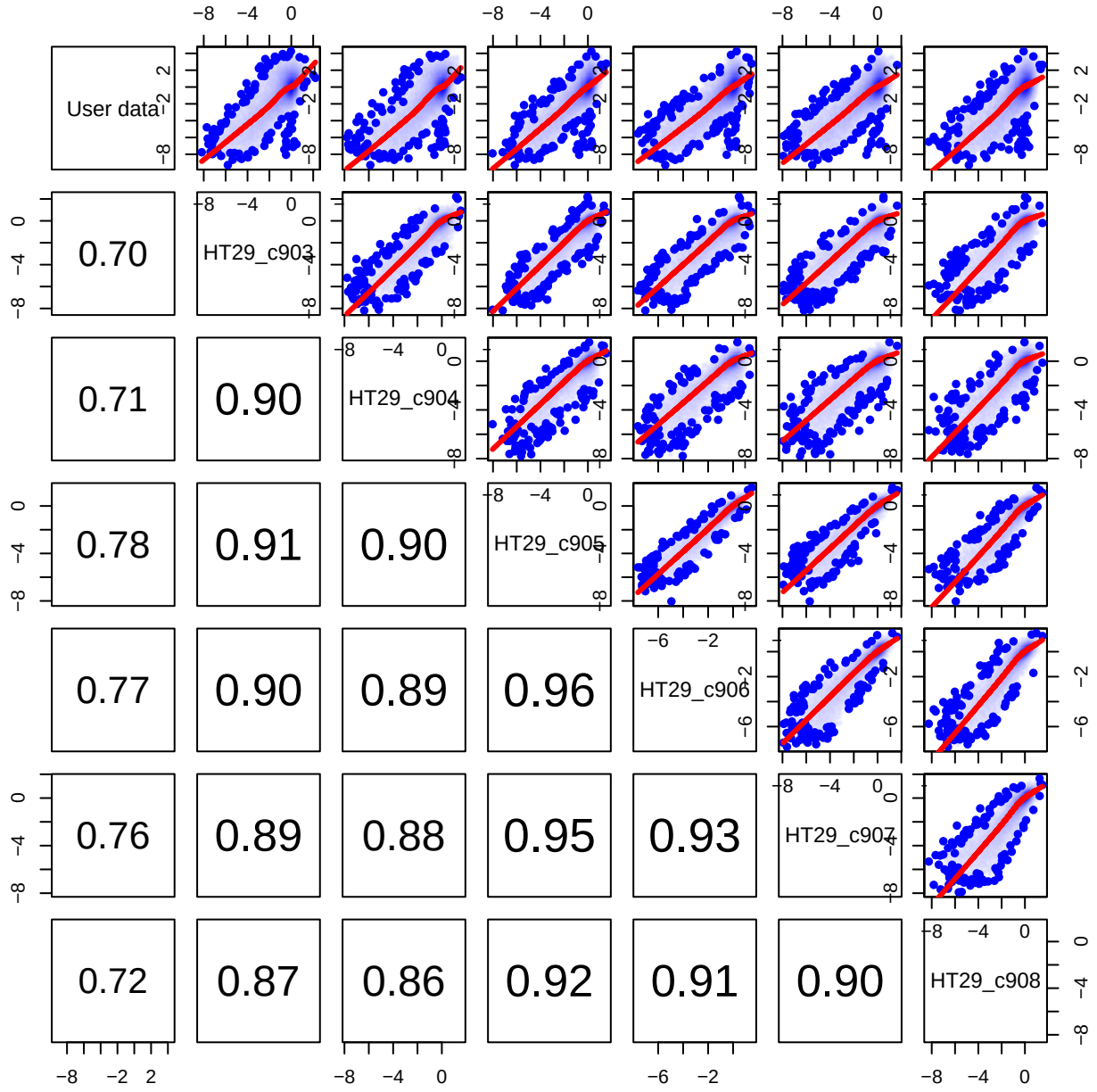
between replicate-averaged sgRNA depletion log fold-change profiles for each of the six HT-29 reference screens (green), and between the user-provided screen with each reference screen (pink). The expected random scores are computed between each possible pair of screens in Project Score (grey). P-values are obtained with a two-sided t-test comparing background and reference distributions or reference and user-data scores. Scatterplots display a detailed correlation matrix of pairwise Pearson's correlation scores between different HT-29 reference screens and between user-provided versus HT-29 reference screens.

The "geneGuides" parameter in the HT29.expSimilarity function defines whether all sgRNAs or only highly informative sgRNAs (defined as in Behan et al. (2019)) will be used for this assessment (geneGuides="All" or geneGuides="HI," respectively). In this example, all guides have been used.

```
RES<-HT29R.expSimilarity(refDataDir = tmpDir,
                         resDir = resultsDir,
                         geneGuides = "All",
                         geneLevel = FALSE,
                         Rscore = TRUE,
                         saveToFig = FALSE,
                         display = TRUE,
                         userFCs = UserData$logFCs)
```

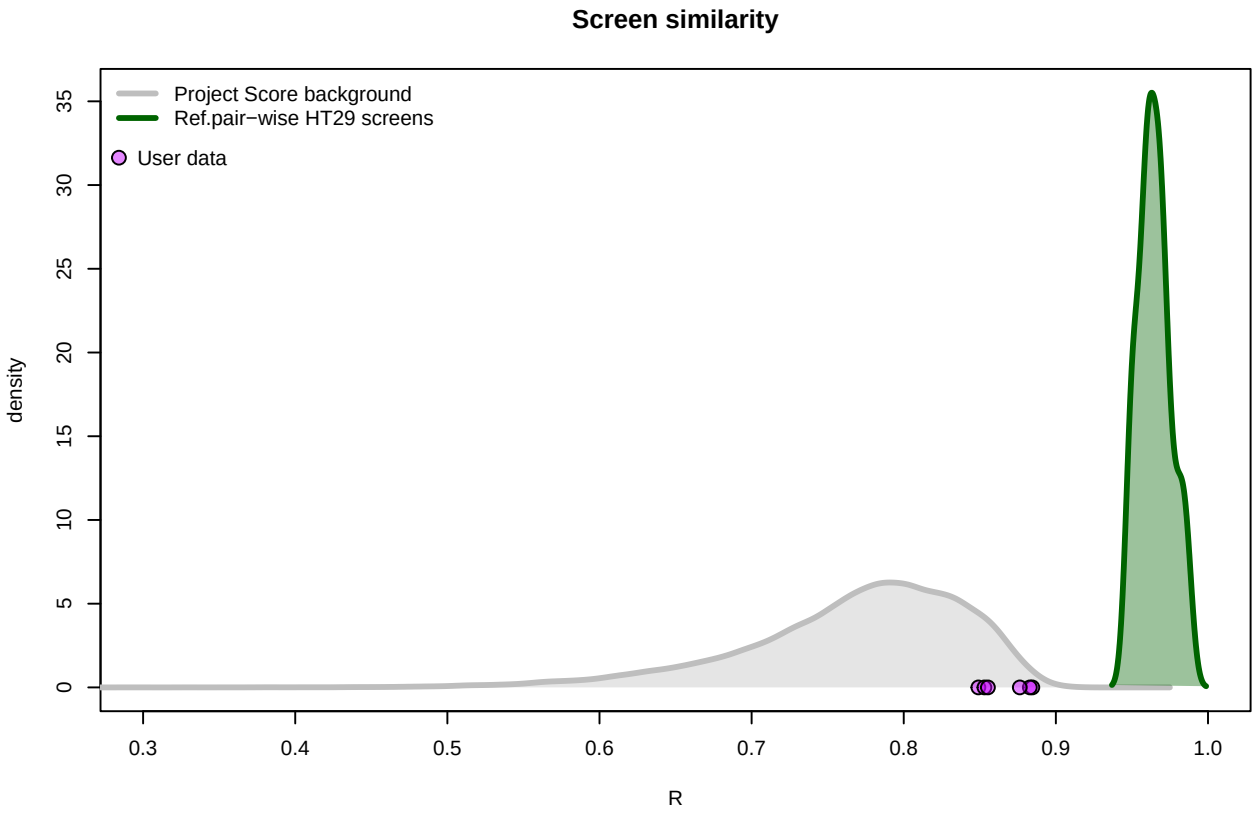

Screen similarity
Project Score background
Ref.pair−wise HT29 screens
User data
density
0
5
10
15
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
R

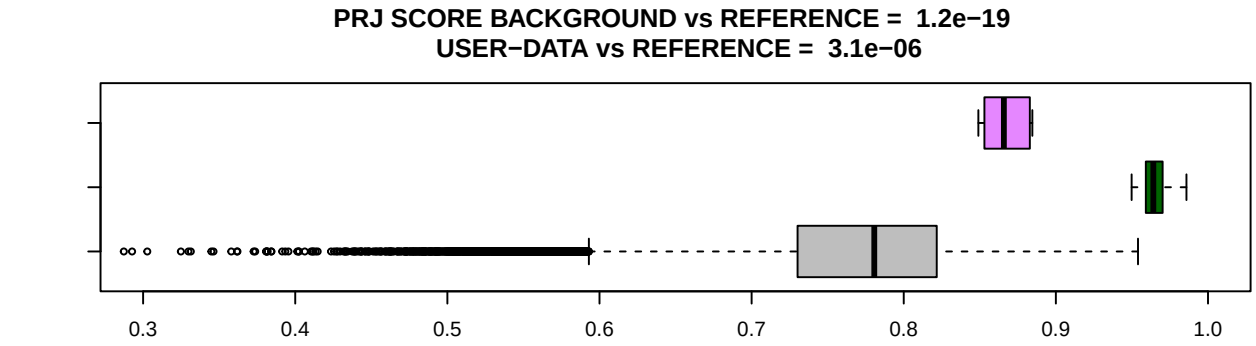

PRJ SCORE BACKGROUND vs REFERENCE = 9e−16
USER−DATA vs REFERENCE = 1.6e−06
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

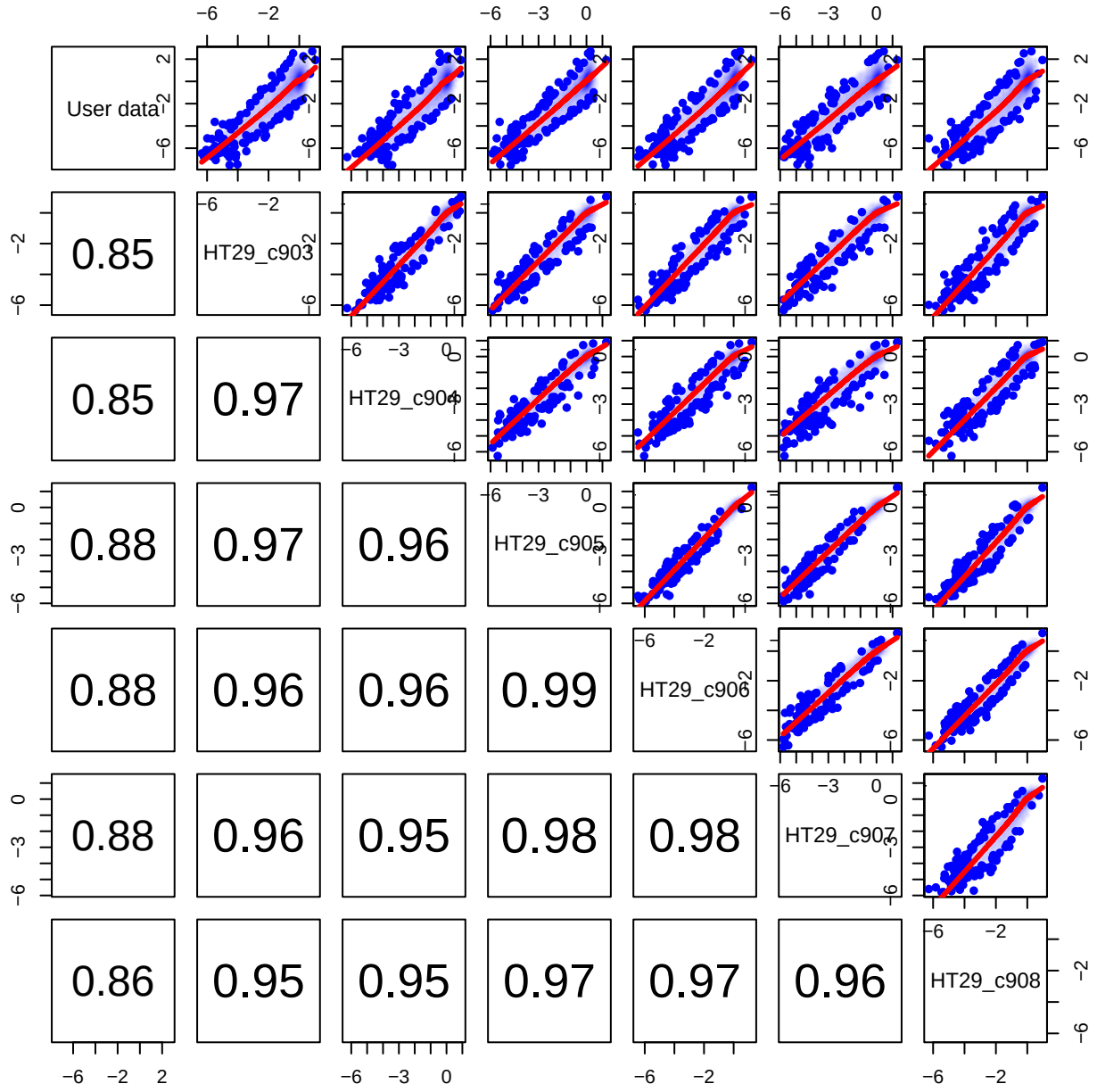


Similar plots can be generated changing the "geneLevel" parameter. With geneLevel=TRUE, a gene level analysis is performed.

```
RES<-HT29R.expSimilarity(refDataDir = tmpDir,
                         resDir = resultsDir,
                         geneGuides = "All",
                         geneLevel = TRUE,
                         Rscore = TRUE,
                         saveToFig = FALSE,
                         display = TRUE,
                         userFCs = UserData$logFCs)
```


Screen similarity
Project Score background
Ref.pair−wise HT29 screens
User data
density
0
5
10
15
20
25
30
35
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
R


PRJ SCORE BACKGROUND vs REFERENCE = 1.2e−19
USER−DATA vs REFERENCE = 3.1e−06
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0



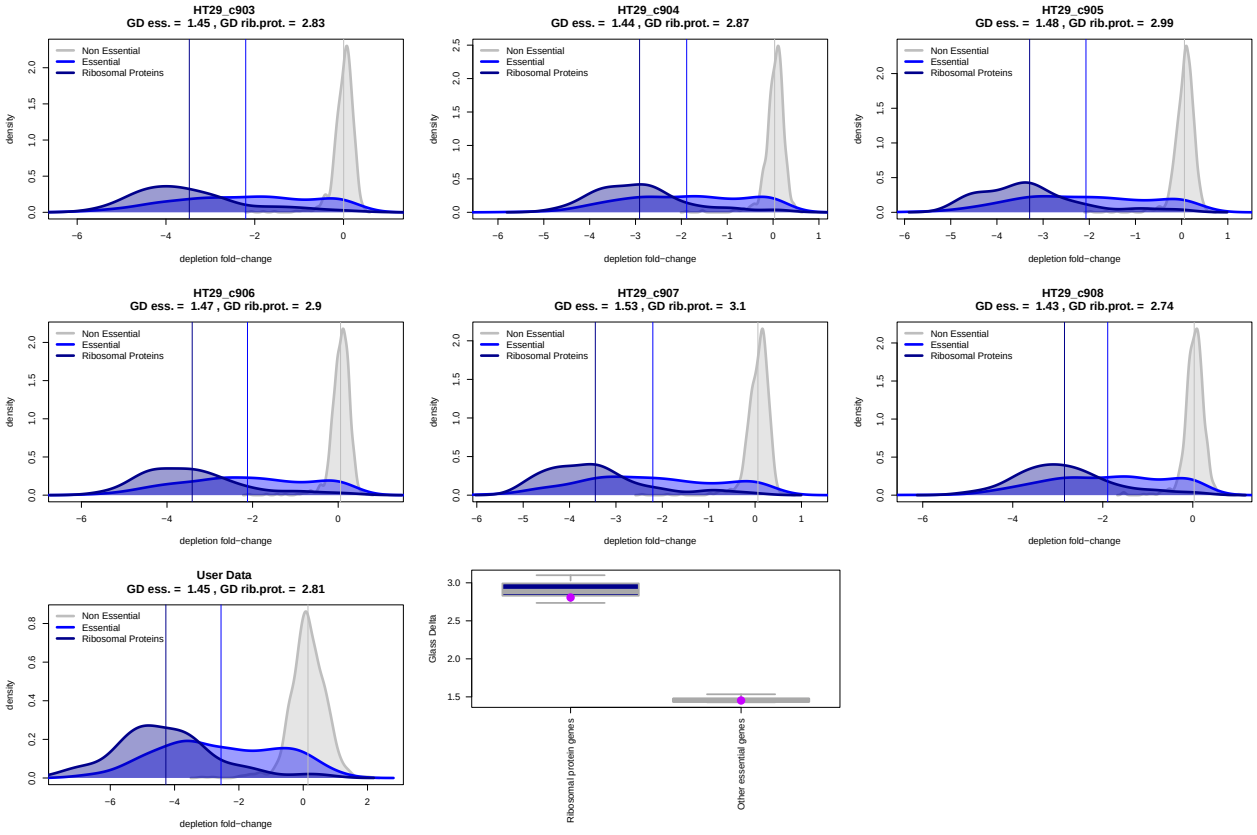
Phenotype intensity

A high-quality screen is expected to retrieve known essential genes as highly essential (i.e. with low logFC). Hence, both reference screens and user-provided data are assessed in this respect, using previosuly-defined essential and non-essential genes (Hart and Moffat (2016)), and genes coding for ribosomal proteins, fundamental components of cells' machinery.

Distributions of depletion log fold-changes for gene subsets are plotted for each reference HT-29 screen and for the user-provided screen. Vertical lines indicate mean logFCs for each distribution. The Glass's Δ (GD) distances between distributions for reference essential genes or ribosomal protein genes with respect to non-essential genes are reported at the top of each plot.

```
layout(matrix(1, nrow=1, ncol=1, byrow=TRUE))

HT29R.PhenoIntensity(refDataDir = tmpDir,
                     resDir = resultsDir,
```

```
                userFCs = UserData$logFCs,
                geneLevel = TRUE,
                saveToFig = FALSE,
                display = TRUE)
```



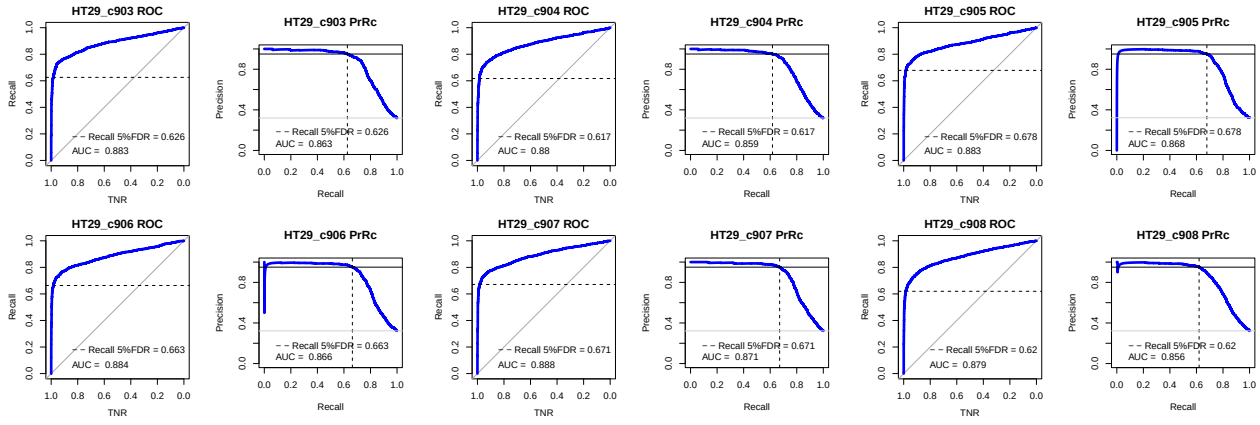
ROC analysis

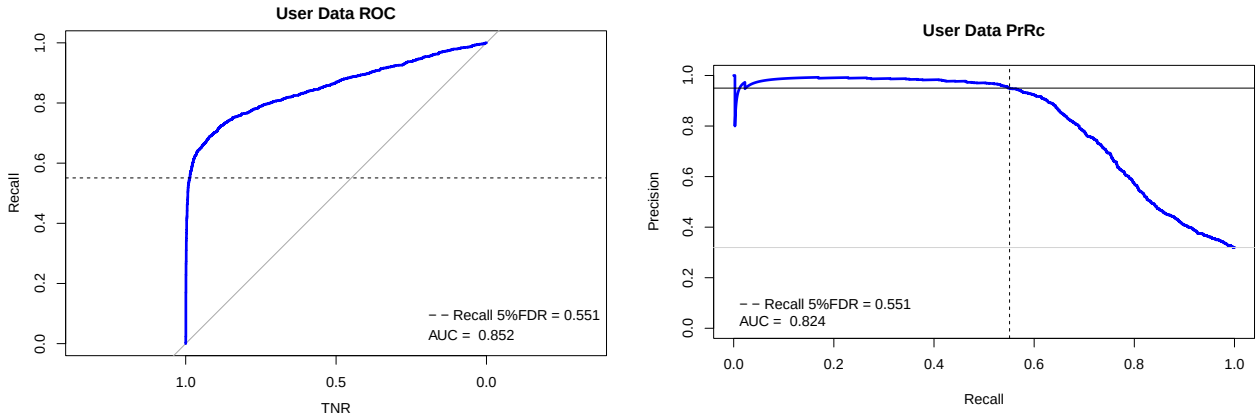
Receiver Operating Curve (ROC) and Precision/Recall (PrRc) curves obtained when classifying genes based on their depletion log fold-change according to previously defined lists of essential and non-essential genes (Hart and Moffat (2016)). If geneLevel is TRUE, the user must provide the lists of known essential and non-essential gene HGNC symbols, otherwise sgRNA identifiers must be provided. The conversion from gene names to sgRNA IDs can be performed with the ccr.genes2sgRNAs function.

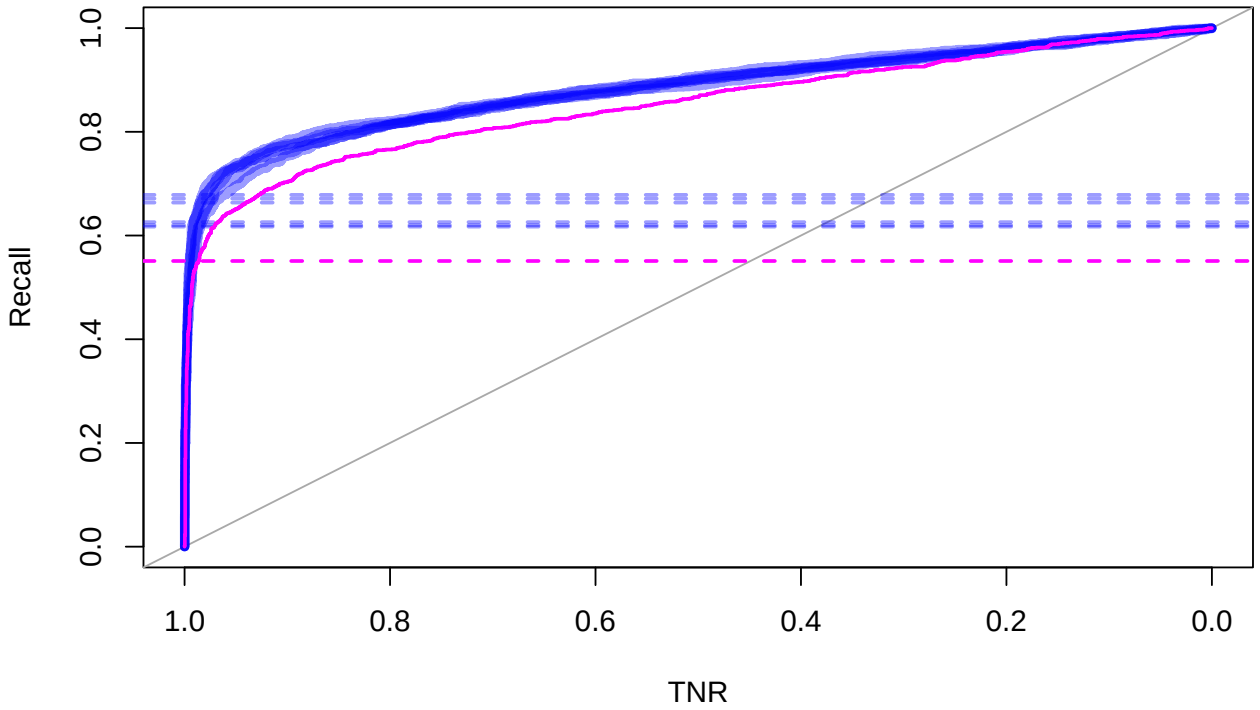
At the sgRNA level:

```
# uncomment if geneLevel = TRUE
data("BAGEL_essential")
data("BAGEL_nonEssential")

Essential_sgRNAs <- ccr.genes2sgRNAs(KY_Library_v1.0, BAGEL_essential)
nonEssential_sgRNAs <- ccr.genes2sgRNAs(KY_Library_v1.0, BAGEL_nonEssential)

HT29R.ROCanalysis(refDataDir = tmpDir,
                  positives = Essential_sgRNAs,
                  negatives = nonEssential_sgRNAs,
                  userFCs = UserData$logFCs,
                  geneLevel = FALSE,
```

```
                saveToFig = FALSE,
                display = TRUE)
```

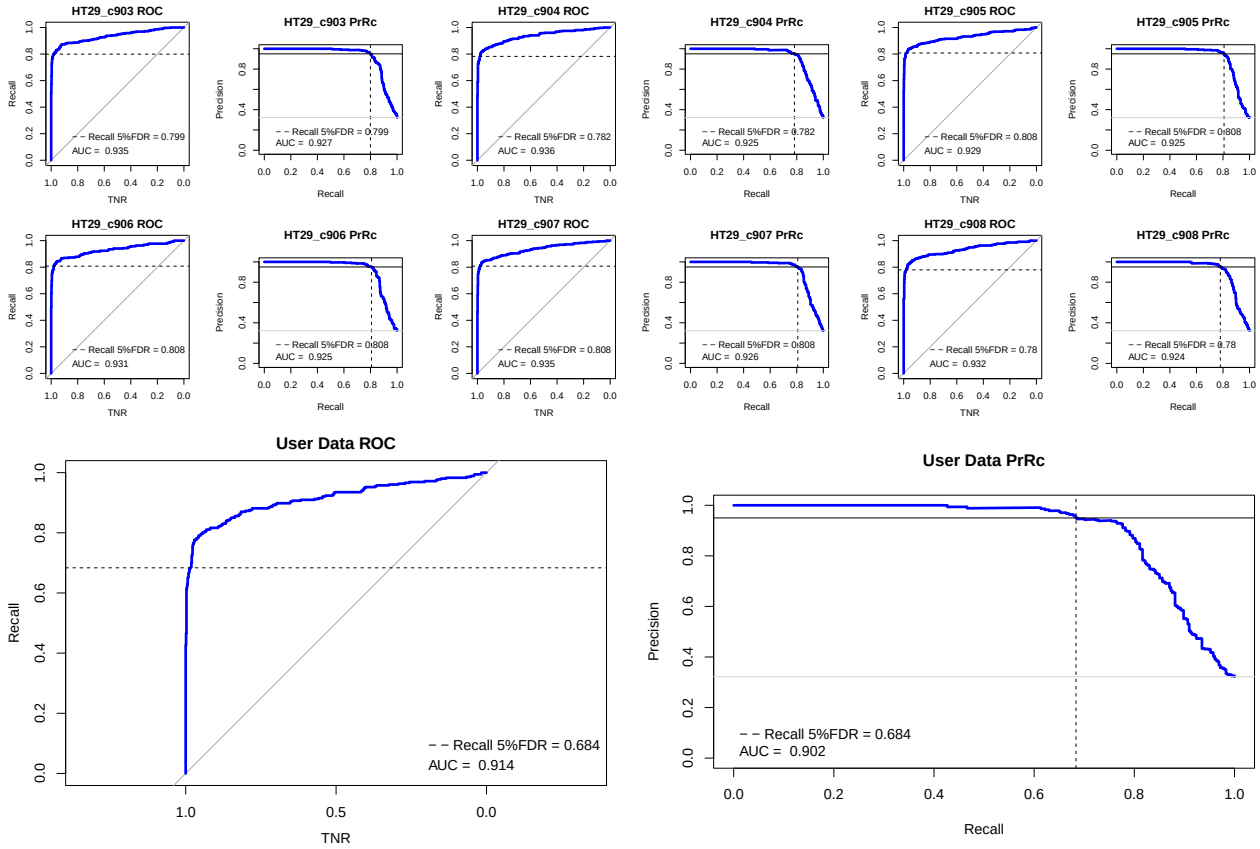


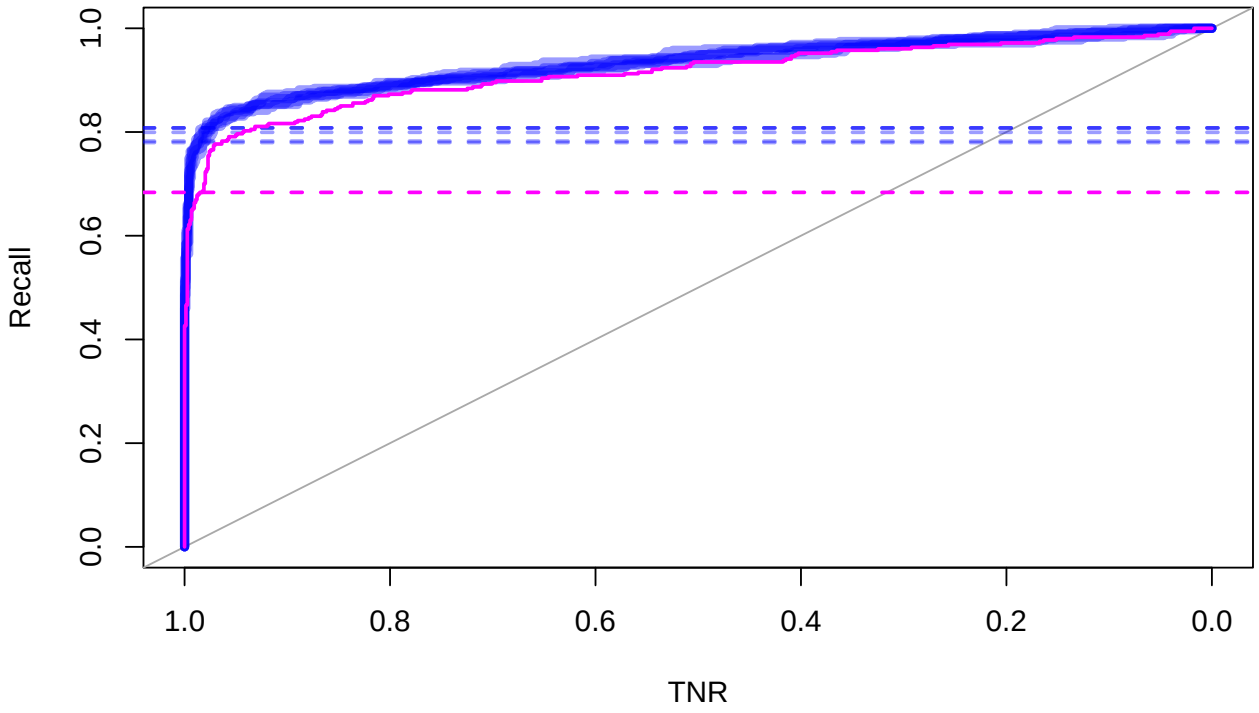




At the gene level:

```
HT29R.ROCanalysis(refDataDir = tmpDir,
                  positives = BAGEL_essential,
                  negatives = BAGEL_nonEssential,
                  userFCs = UserData$logFCs,
                  geneLevel = TRUE,
                  saveToFig = FALSE,
                  display = TRUE)
```





HT-29-specific fitness genes and their characterisation

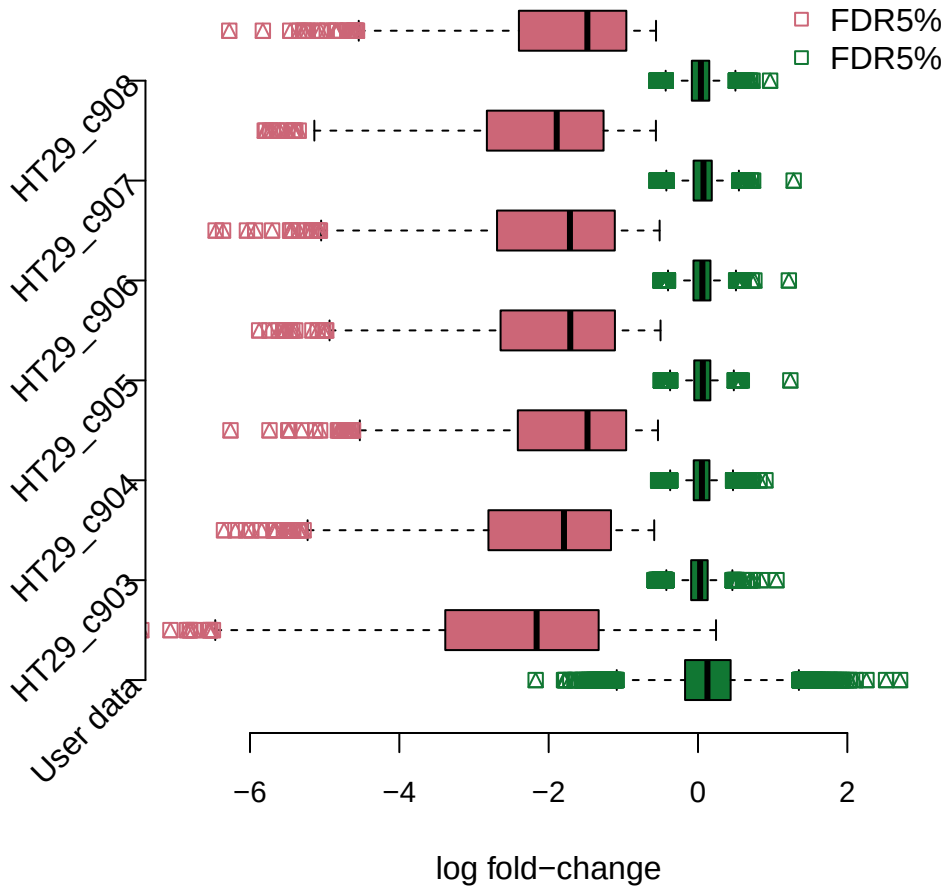
We assembled a list of genes that are significantly depleted across all the reference HT-29 screens and can thus be used to evaluate new user-provided HT-29 screens. For each reference HT-29 screen we identified a set of genes significantly depleted at a 5% FDR (False Discovery Rate) and its complement, i.e. a set of genes not significantly depleted, using reference sets of essential and non-essential genes (Hart and Moffat (2016)).

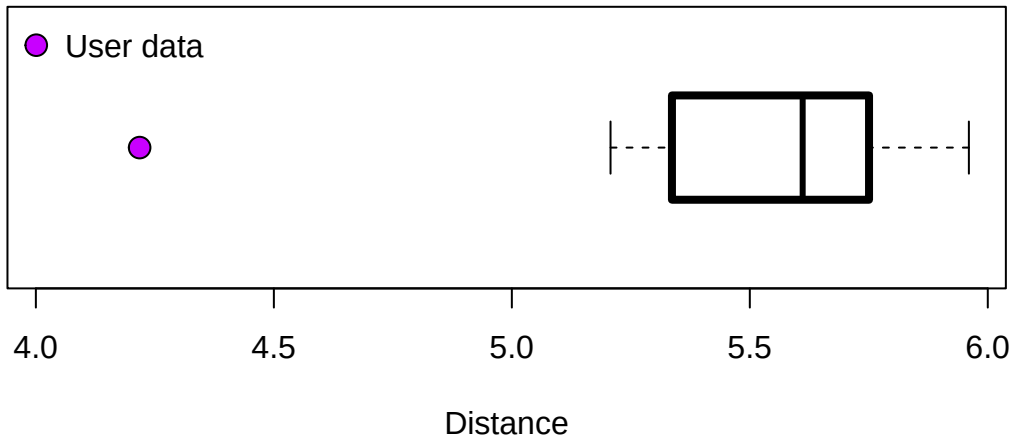
HT-29-specific essential genes are defined as the intersection between all lists of depleted genes across each HT-29 reference screen. Conversely, HT-29-specific non-essential genes are defined as the intersection between all lists of genes not significantly depleted across each HT-29 reference screen.

Below, the log fold-changes for essential and non-essential genes defined above are shown for each reference screen. Distribution of distances between HT-29-specific positive and negative essential genes, quantified through Cohen's d, is shown for the reference HT-29 screens (boxplot) and for user-provided data (pink dot).

```
res <- HT29R.FDRconsensus(refDataDir = tmpDir,
                          resDir = resultsDir,
                          userFCs = UserData$logFCs,
                          distance = "Cohen's",
                          FDRth = 0.05,
                          saveToFig = FALSE,
                          display = TRUE)
```

```
## Using group as id variables
```

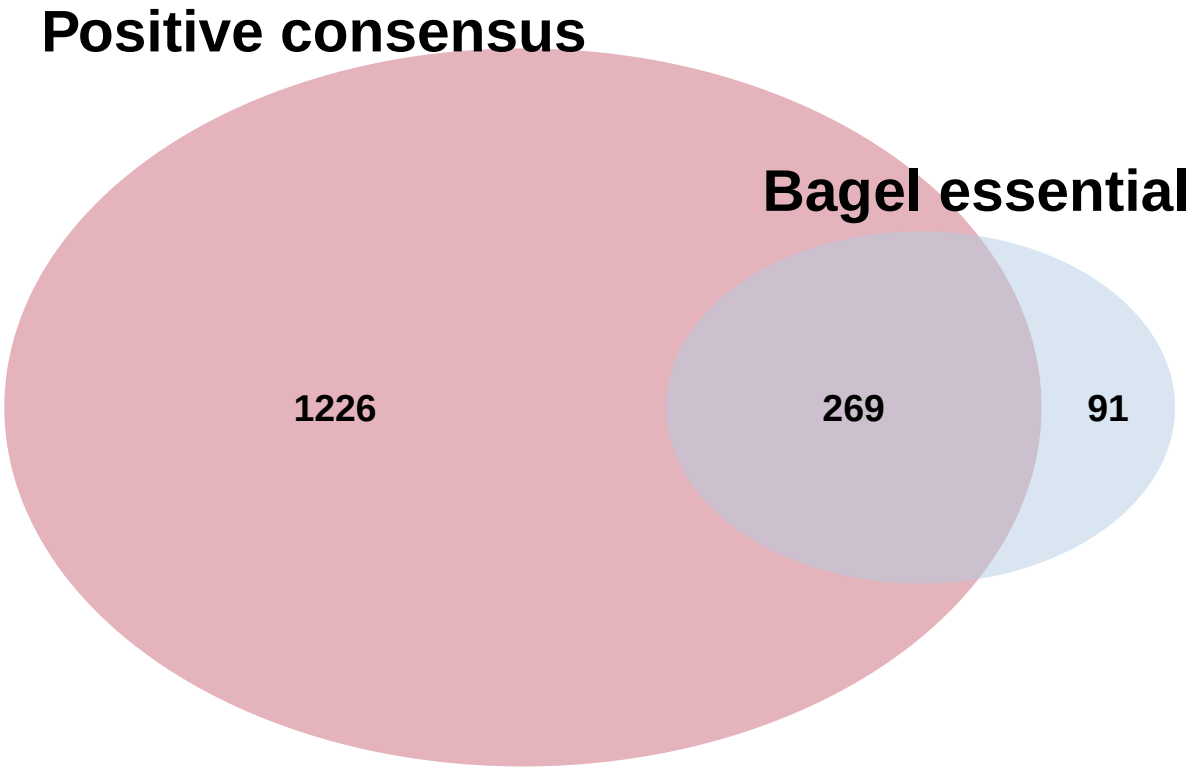




Overlap between the HT-29 positive consensus and prior known essential genes (Hart and Moffat (2016)) shows a significant enrichment.

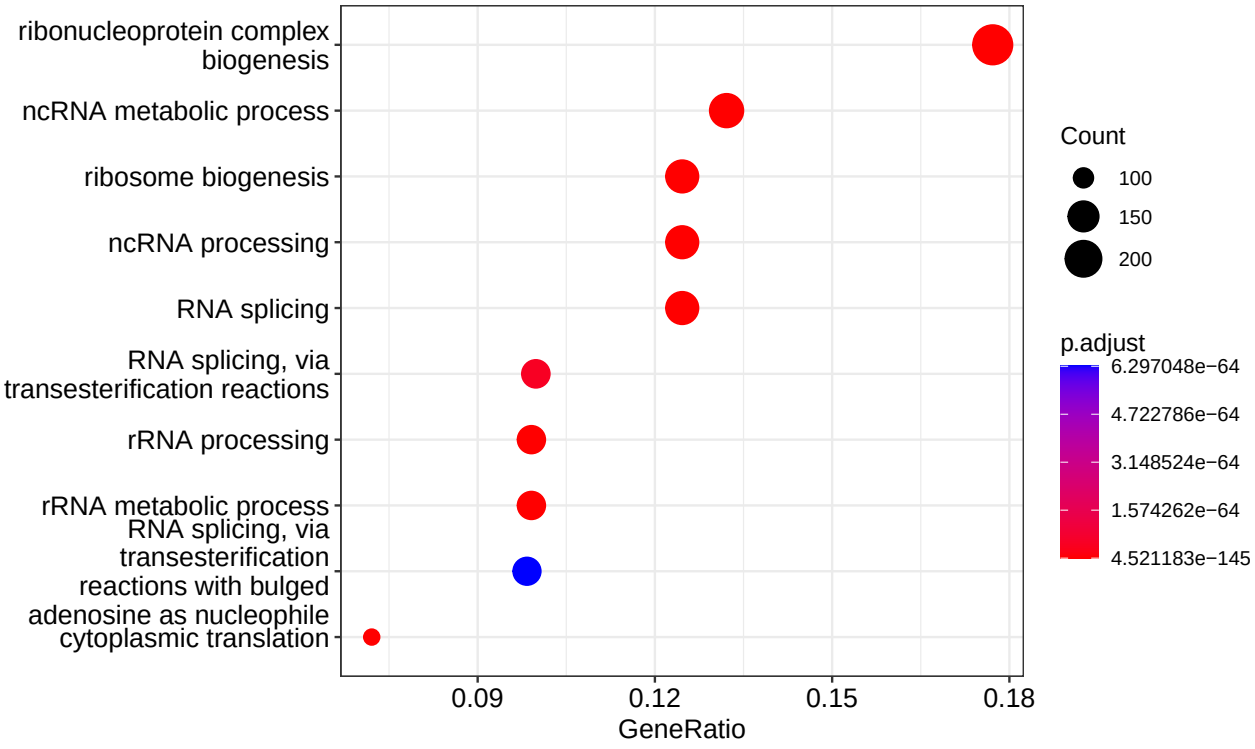
```
HT29R.FDRenrichment(consensus = res$POS, background=res$Universe, labels = BAGEL_essential)
```

Fisher's exact test:7.1146537498882e−221



Top 10 Gene Ontology categories (Biological Process, BP) enriched for the HT-29 Positive Consensus. The function annFUN.org retrieves genes' annotation to biological processes (BPs), returning a list with BPs and corresponding gene symbols. To test the enrichment, all the genes assigned to at least one BP category are employed as part of the universe, and the enrichment is computed through the enrichGO function from the enrichr package.

```
BPmapping <- annFUN.org("BP", mapping = "org.Hs.eg.db", ID = "symbol")
genesUniverse <- unique(unlist(BPmapping))

GO <- enrichGO(gene = res$POS,
               keyType = "SYMBOL",
               universe = genesUniverse,
               ont="BP",
               OrgDb = "org.Hs.eg.db")

dotplot(GO, showCategory=10)
```



References

Behan, Fiona M, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M Beaver, Giorgia Migliardi, Rita Santos, et al. 2019. "Prioritization of Cancer Therapeutic Targets Using CRISPR-Cas9 Screens." *Nature* 568 (7753): 511–16.

Behan M., Fiona, Francesco Iorio, and Garnett Garnett J. 2022. "Ht29 Reference Dataset. Figshare. Dataset. Https://Doi.org/10.6084/M9.figshare.20480544."

Dwane, Lisa, Fiona M Behan, Emanuel Gonçalves, Howard Lightfoot, Wanjuan Yang, Dieudonne van der Meer, Rebecca Shepherd, Miguel Pignatelli, Francesco Iorio, and Mathew J Garnett. 2021. "Project Score Database: A Resource for Investigating Cancer Cell Dependencies and Prioritizing Therapeutic Targets." *Nucleic Acids Res.* 49 (D1): D1365–72.

Hart, Traver, and Jason Moffat. 2016. "BAGEL: A Computational Framework for Identifying Essential Genes from Pooled Library Screens." *BMC Bioinformatics* 17 (April): 164.

Iorio, Francesco, Fiona M Behan, Emanuel Gonçalves, Shriram G Bhosle, Elisabeth Chen, Rebecca Shepherd, Charlotte Beaver, et al. 2018. "Unsupervised Correction of Gene-Independent Cell Responses to CRISPR-Cas9 Targeting." *BMC Genomics* 19 (1): 604.

Tzelepis, Konstantinos, Hiroko Koike-Yusa, Etienne De Braekeleer, Yilong Li, Emmanouil Metzakopian, Oliver M Dovey, Annalisa Mupo, et al. 2016. "A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia." *Cell Rep.* 17 (4): 1193–1205.