Benchmark data and software for assessing genome-wide CRISPR-Cas9 screening pipelines

Raffaele M. Iannuzzi, Ichcha Manipur, Clare Pacini, Fiona M. Behan, Mario R. Guarracino, Mathew J.

2022-08-22

Running Modalities

HT29benchmark is an R package available at: https://github.com/DepMap-Analytics/HT29benchmark.

This page contains instruction to quickly try the package. User manual and package documentation are available at https://github.com/DepMap-Analytics/HT29benchmark/blob/master/HT29benchmark.pdf.

R package: quick start

Install additional libraries from Bioconductor: + topGO, + clusterProfiler, + org.Hs.eg.db, + enrichplot

from CRAN: + VennDiagram, + data.table, + Kern.Smooth

Load libraries

```
# QC analysis
library(CRISPRcleanR)
library(HT29benchmark)

# additional for vignette production
library(data.table)
library(VennDiagram)
library(clusterProfiler)
library(enrichplot)
library(org.Hs.eg.db)
library(topGO)
library(RColorBrewer)
```

The package comes with built-in data objects containing the following Project Score data and sets of sgRNA guides used to perform the QC analysis:

- HT29R.GL_prSCORE_rCorr.RData

- HT29R.prSCORE_rCorr.RData
- HT29R.prSCORE_bkgr_screen_similarity.RData
- HT29R.prSCORE_bkgr_screen_similarity_HI.RData
- HT29R.prSCORE_bkgr_screen_similarity_sgRNA.RData
- HT29R.prSCORE_bkgr_screen_similarity_sgRNA_HI.RData
- HT29R.reproducible_GeneGuides.Rdata
- HT29R.consensus_GeneGuides.RData

Setup

The system will create a directory (i.e., the 'HT29R_resFolder') where the FCs (or rawCounts) will be downloaded and stored as well. A subdirectory will be created in the 'HT29R_resFolder' to store the saved plots (which is likely to happen if the 'saveToFig' parameter of the HT29R 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="")
```

Preprocessing of User-provided data

If the User have its own screening data and want to validate them using the QC pipeline we provide, he/she must specify the path to this file in the chunk below. This must be a tab delimited file (.tsv) with one row per sgRNA and columns/headers representing sgRNA identifiers, HGNC symbols of the genes targeted, sgRNAs' counts for controls and samples (see the 'ccr.NormfoldChanges' function for more details).

```
userData <- 'path/to/data'
```

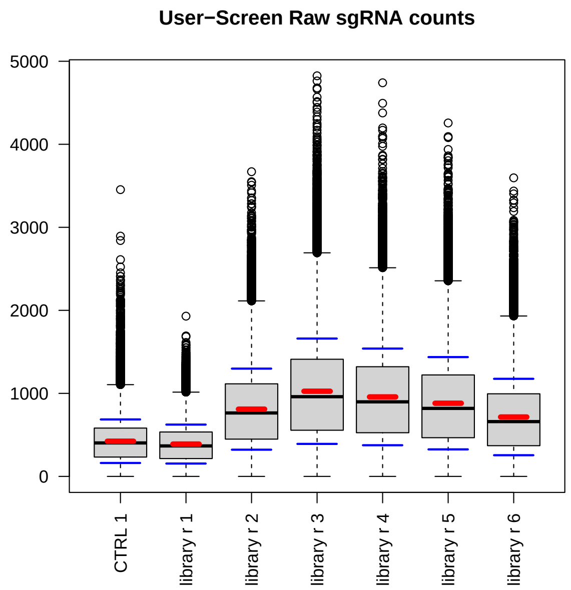
The chunk below downloads our example User screen in the 'HT29R_resFolder'. We have used 6 replicates of a lower quality recessive screen to validate our QC pipeline (see manuscript for more details).

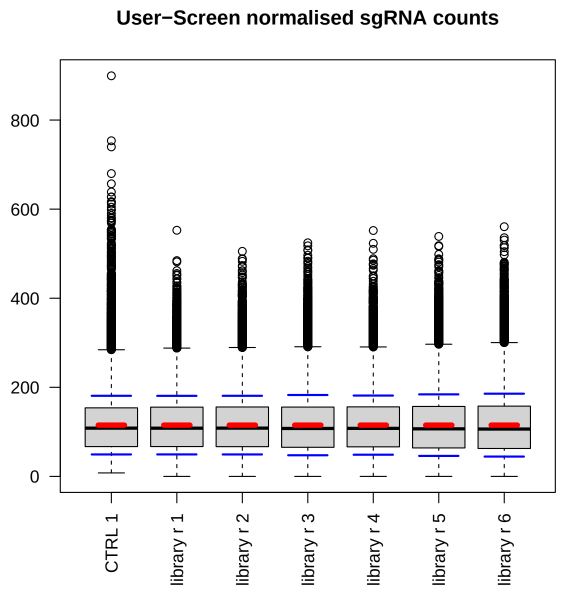
```
# comment these lines of codes if you have already provided your own data
URL <- 'https://figshare.com/ndownloader/files/36658530?private_link=5b2a579791c47a417474'
download.file(URL, destfile = paste0(tmpDir, '/Example_UserScreen.tsv'))
userData <- paste0(tmpDir, '/Example_UserScreen.tsv')
```

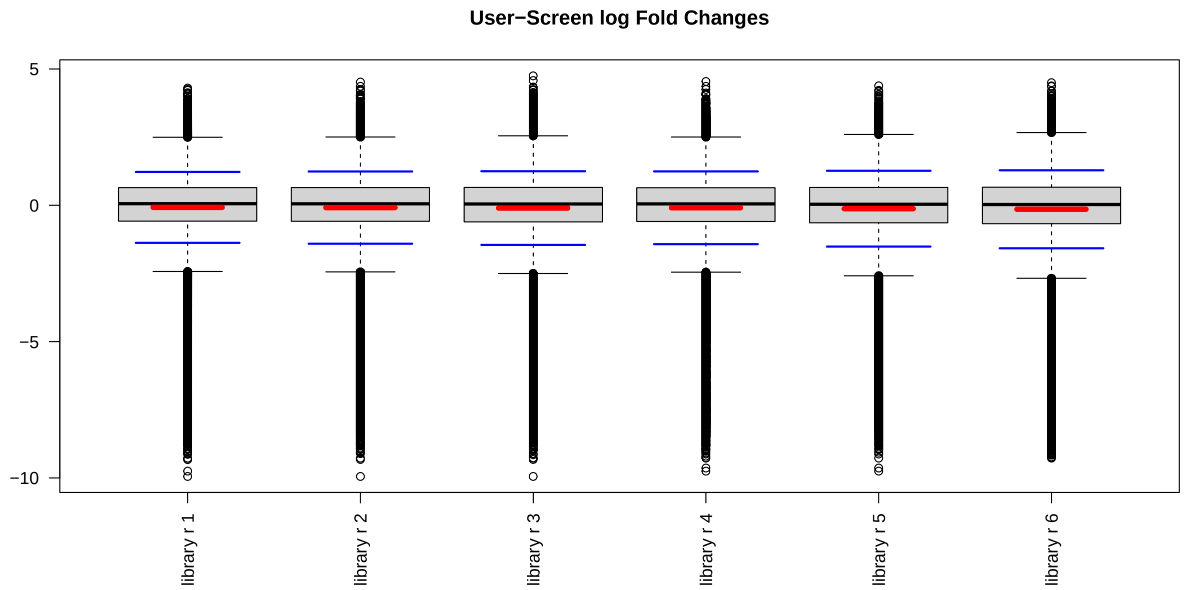
Normalize sgRNA counts and store computed User log fold-changes in 'expData'

Create a list of two elements: (i) 'norm_counts' and (ii) 'logFCs' dataframes.

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





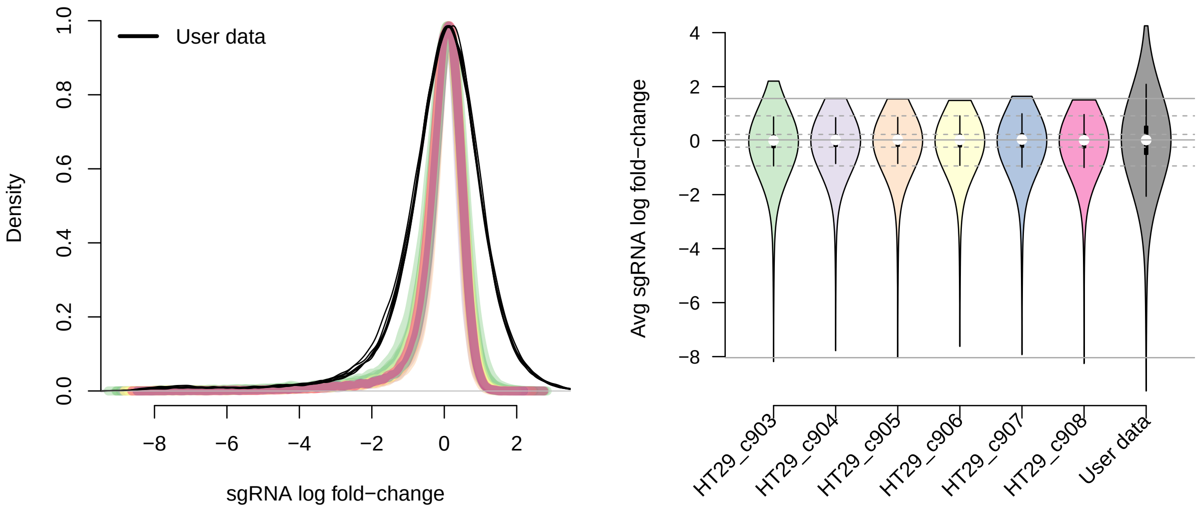


Downloading reference sgRNA depletion fold-changes from high-quality HT-29 screens into 'HT29_resFolder' directory

sgRNAs depletion statistics

Average parameters and confidence intervals of the distribution of sgRNA log fold-changes observed when screening HT-29 with reagent and experimental settings described in Behan et al. 2019. If provided, User data statistics will be also shown.

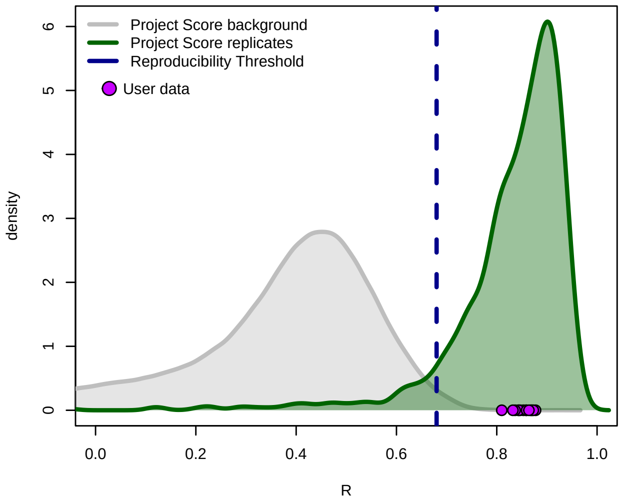
```
HT29R.FCdistributions(refDataDir = tmpDir,
                      resDir = resultsDir,
                      userFCs = expData$logFCs,
                      stats = TRUE,
                      saveToFig = FALSE,
                      display = TRUE)
```

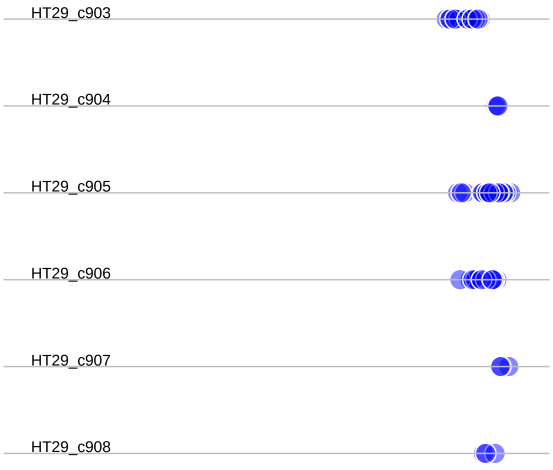


Screen reproducibility

Kernel distributions of the pair-wise R scores between replicates computed on log fold-changes at both sgRNAs-level and gene-level across all cells (Background in grey) and for any pair of randomly selected cell lines (Expectation in green). The User can check if its pair-wise R scores (in pink) are reaching the significant threshold while comparing them with the pair-wise R scores of the HT-29 experiments replicates (in blue)

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



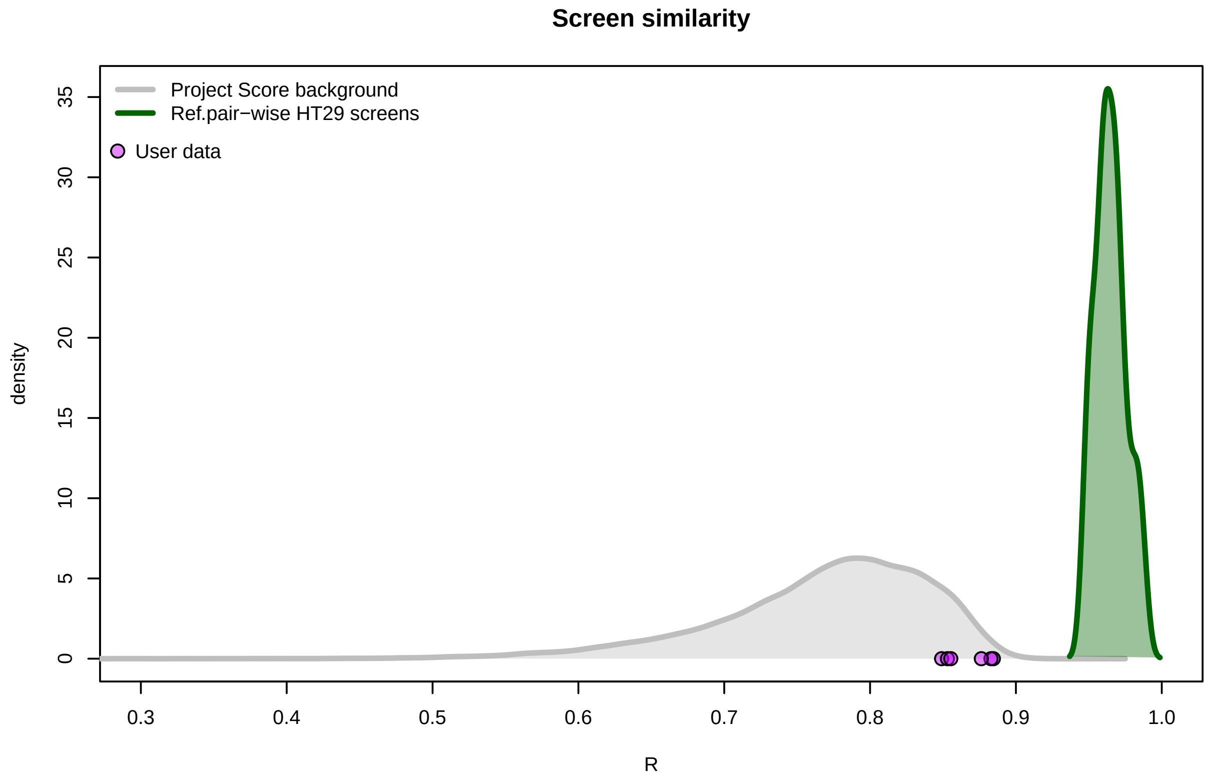


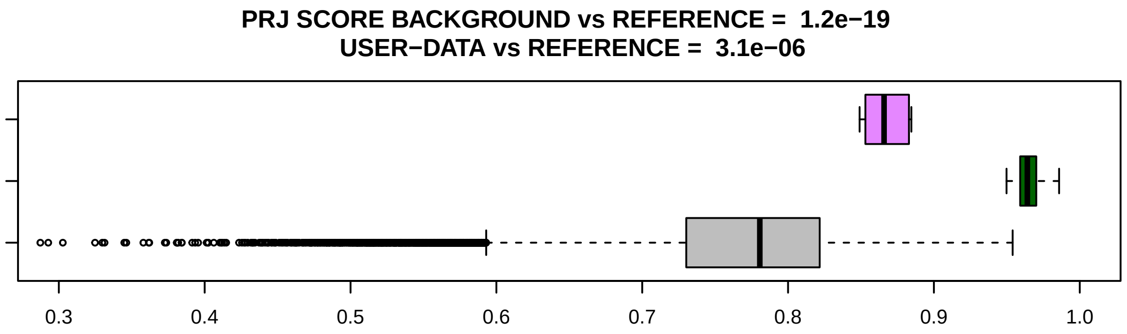
Screen Similarity

Project Score comparison

Kernel distributions of Project Score background and HT29 screens and correlation scores for User data with P-values for two-tailed Student's t-test highlighted in bold.

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

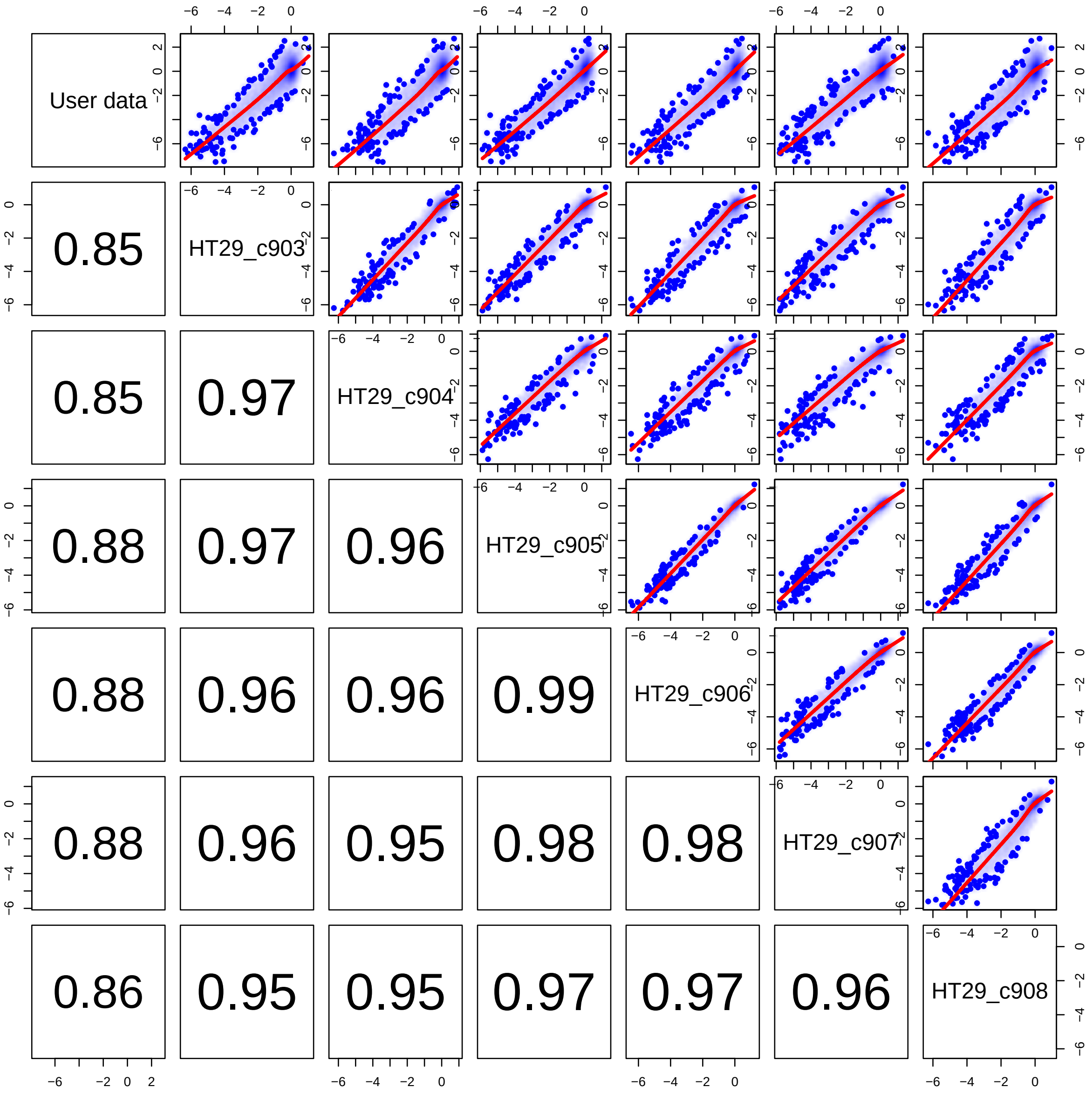




Pearson's correlation matrix

The Pearson's correlation matrix computed on log fold-changes on HT29 vs User screen experiments averaged across technical replicates.

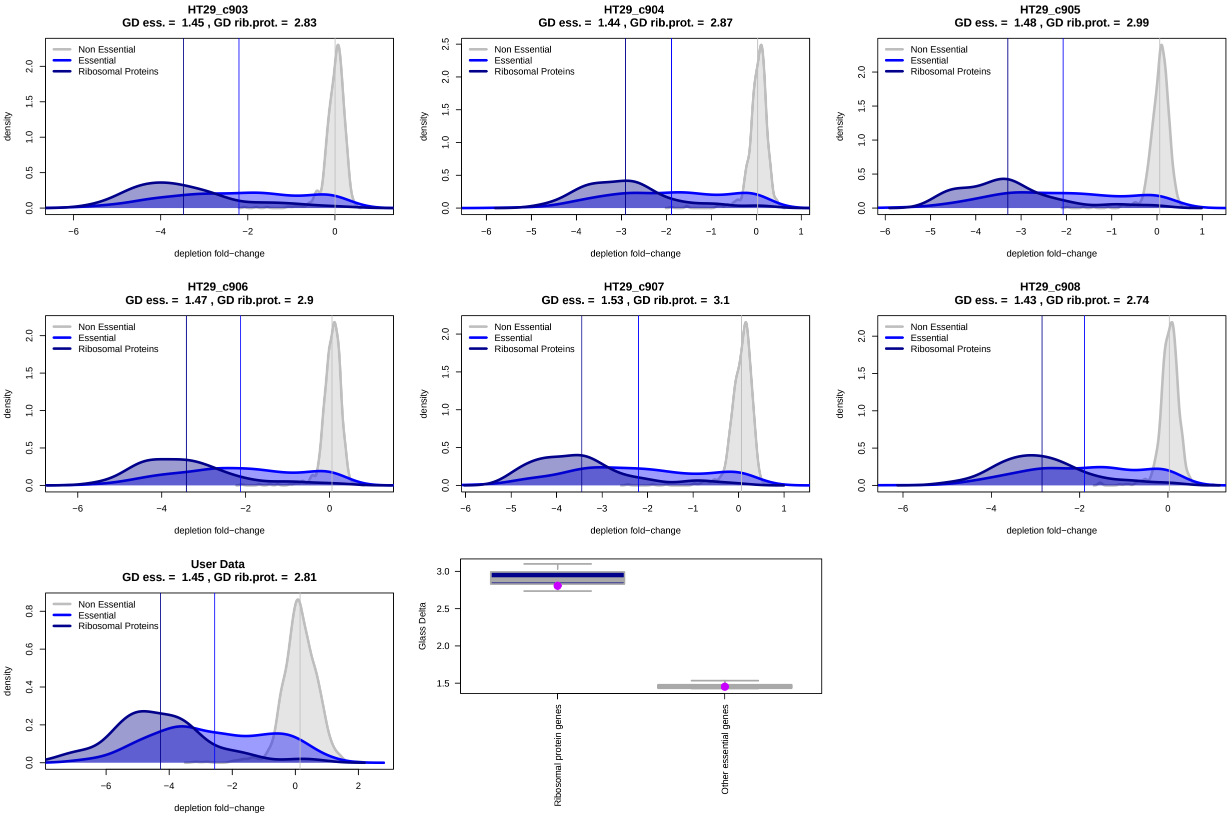
```
pairs(RES,lower.panel = panel.cor, upper.panel = my.panelSmooth)
```



Performance assessment: measuring phenotype intensity

Kernel distributions of the depletion signal observed in both HT29 reference and User screens evaluating median log fold-changes are shown. The distances between distributions for reference non-essential (N), essential (E) and ribosomal (R) protein genes are depicted as straight lines with the respective color codes. Glass's Delta (GD) score of the ribosomal protein (R) and essential genes (E) are computed for both HT29 reference and User data (in pink).

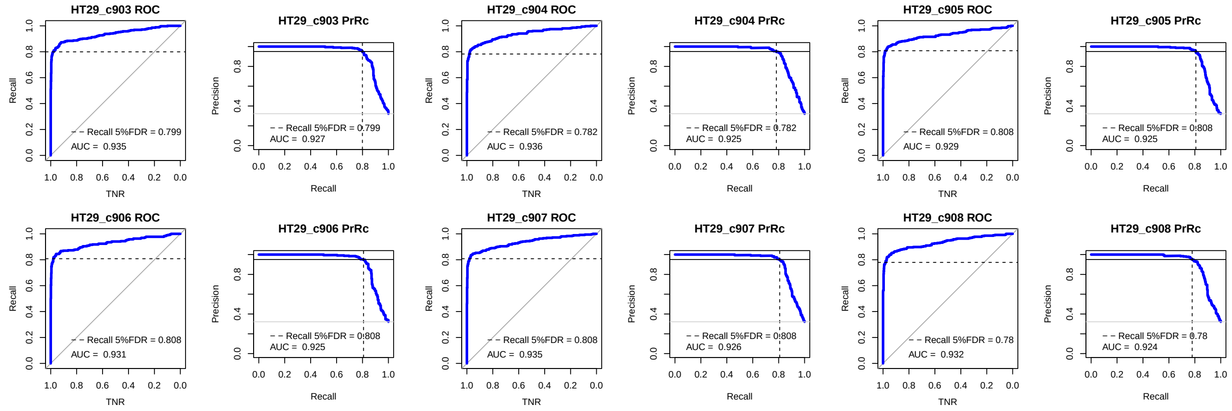
```
layout(matrix(c(1:6), nrow=2, ncol=3, byrow=TRUE))
HT29R.PhenoIntensity(refDataDir = tmpDir,
                     resDir = resultsDir,
                     userFCs = expData$logFCs,
                     geneLevel = TRUE,
                     saveToFig = FALSE,
                     display = TRUE)
```

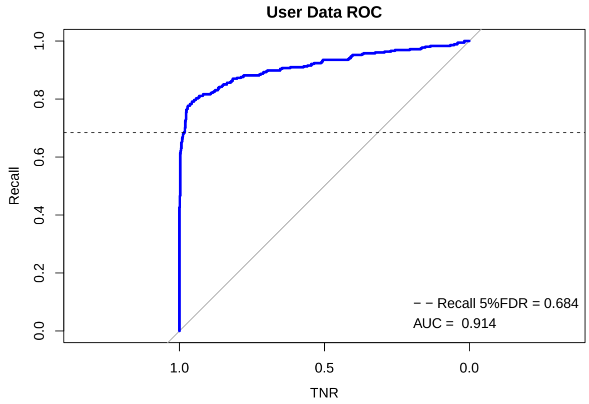


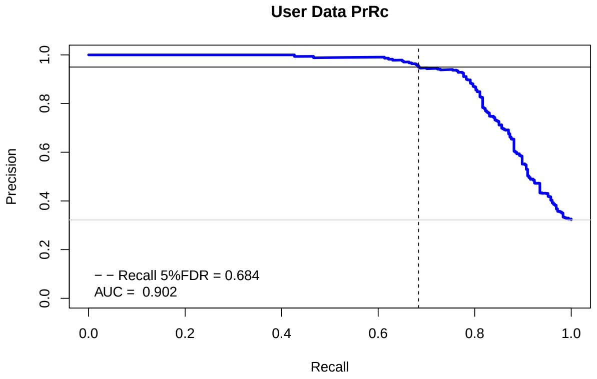
ROC analysis

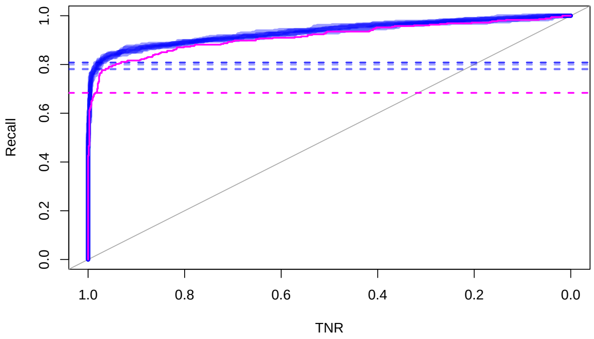
ROC and Precision/Recall curves obtained when classifying the fitness genes found for each HT-29 and User experiments at 5 % of FDR of Bagel essential and non-essential sgRNAs (geneLevel = FALSE) or genes (geneLevel = TRUE). If geneLevel is TRUE, User must provide the lists of Bagel essential and non-essential genes without converting it to sgRNAs.

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

# uncomment if geneLevel = FALSE
# 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 = BAGEL_essential,
                  negatives = BAGEL_nonEssential,
                  userFCs = expData$logFCs,
                  geneLevel = TRUE,
                  saveToFig = FALSE,
                  display = TRUE)
```







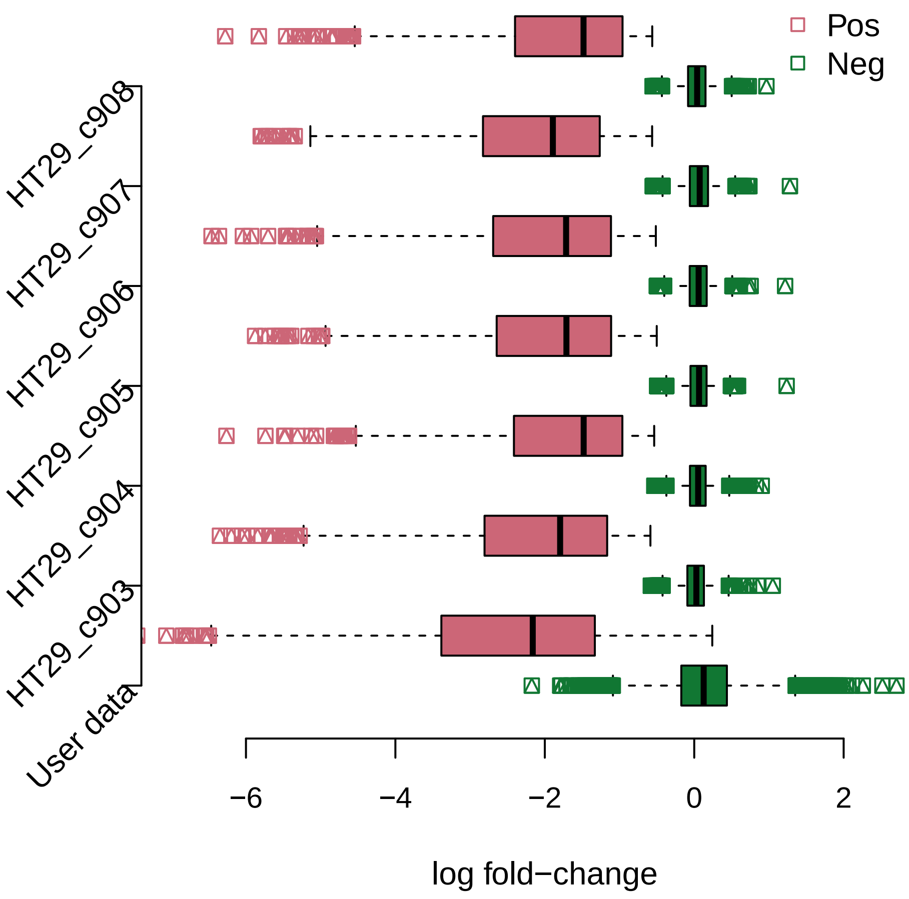


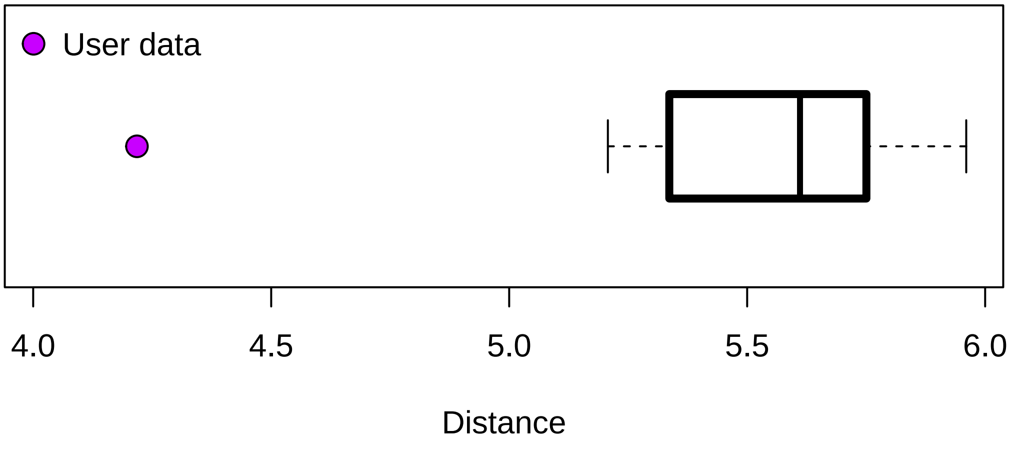
HT-29-specifics genes at 5% FDR

Boxplots of the depletion signal carried by the Positive (in red) and Negative genes' consensus (in green) with the respective measure of distance between the two distributions. The User can choose the effect size to compute (Cohen's d or Glass Delta). The function return the list of HT-29-specific genes at 5% FDR (POS), the Negative consensus of genes at 5 % FDR (NEG) plus the gene Universe to compute the Fisher's exact test.

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

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


Pos
Neg
HT29_c908
HT29_c907
HT29_c906
HT29_c905
HT29_c904
HT29_c903
User data
−6
−4
−2
0
2
log fold−change

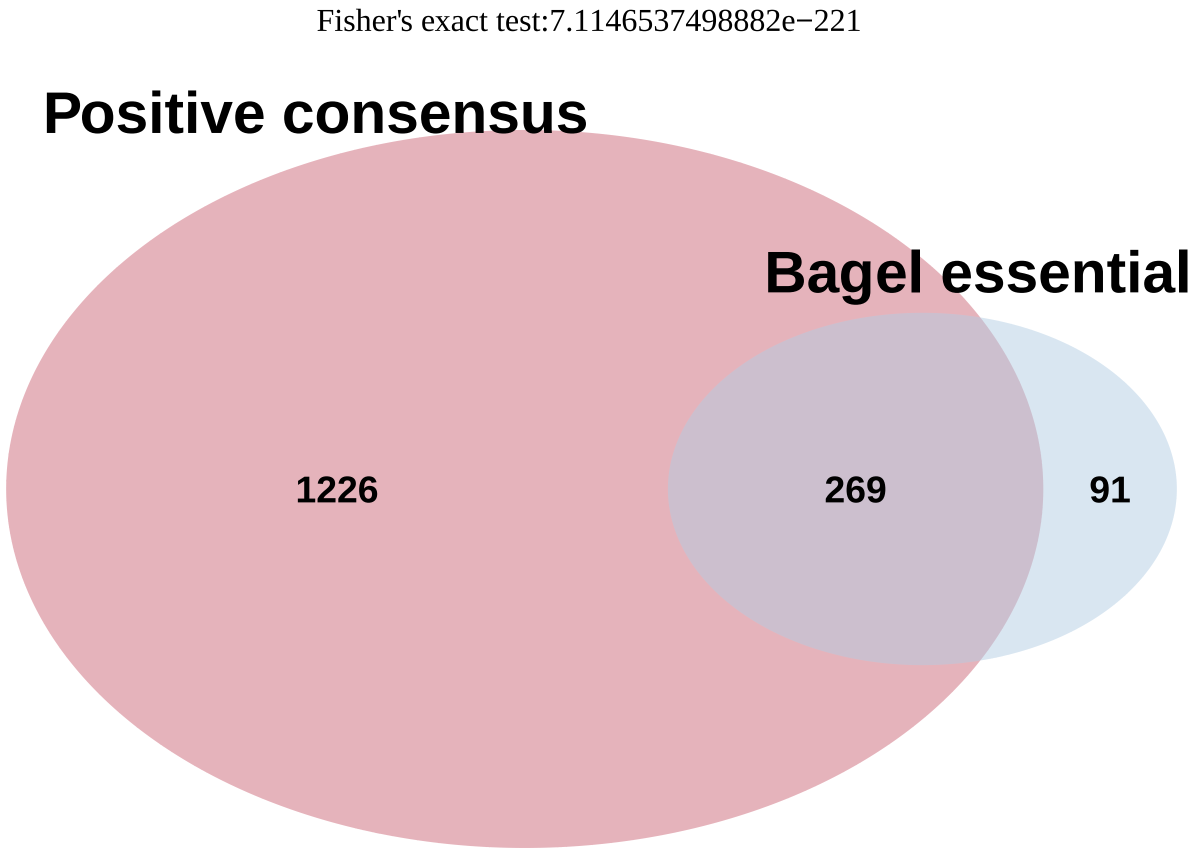

User data
4.0
4.5
5.0
5.5
6.0
Distance

FDR enrichment

Venn diagram

Two-sided Fisher's exact test

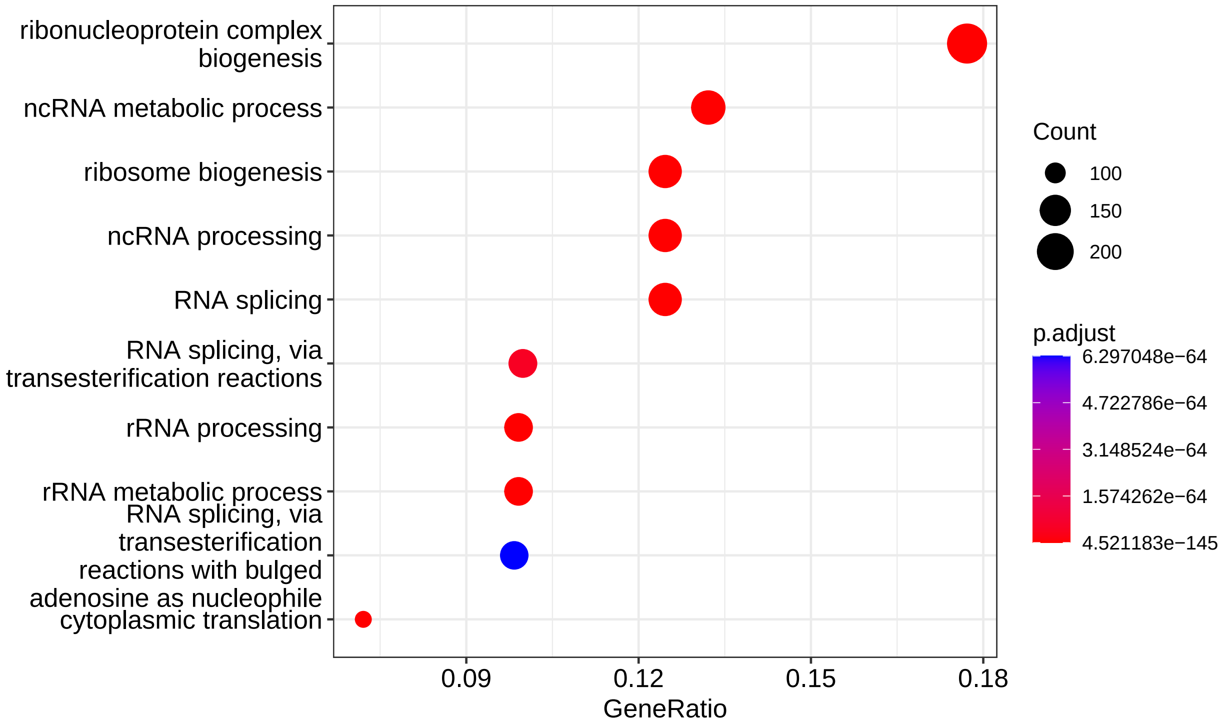
```
HT29R.FDRenrichment(consensus = res$POS, background=res$Universe, labels = BAGEL_essential)
```



GO analysis

Top 10 Gene Ontology categories (Biological Process, BP) enriched for the HT-29 Positive Consensus

```
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)
```


ribonucleoprotein complex biogenesis
ncRNA metabolic process
ribosome biogenesis
ncRNA processing
RNA splicing
RNA splicing, via transesterification reactions
rRNA processing
rRNA metabolic process
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
cytoplasmic translation
0.09
0.12
0.15
0.18
GeneRatio
Count
100
150
200
p.adjust
6.297048e−64
4.722786e−64
3.148524e−64
1.574262e−64
4.521183e−145