In silico CRISPR-Cas9 validation Report

Raffaele M. Iannuzzi @I0RI0 lab

16/08/2022

In order to generate this report, the system will use the following Project Score data embedded in the package:

- 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/')
```

```
## Warning in dir.create("~/HT29R_resFolder/"): '/Users/raffaele.iannuzzi/
## HT29R_resFolder' already exists
```

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

```
## Warning in dir.create(paste(tmpDir, "USER/", sep = "")): '/Users/
## raffaele.iannuzzi/HT29R_resFolder/USER' already exists
```

```
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 the HT29 manuscript for more details).
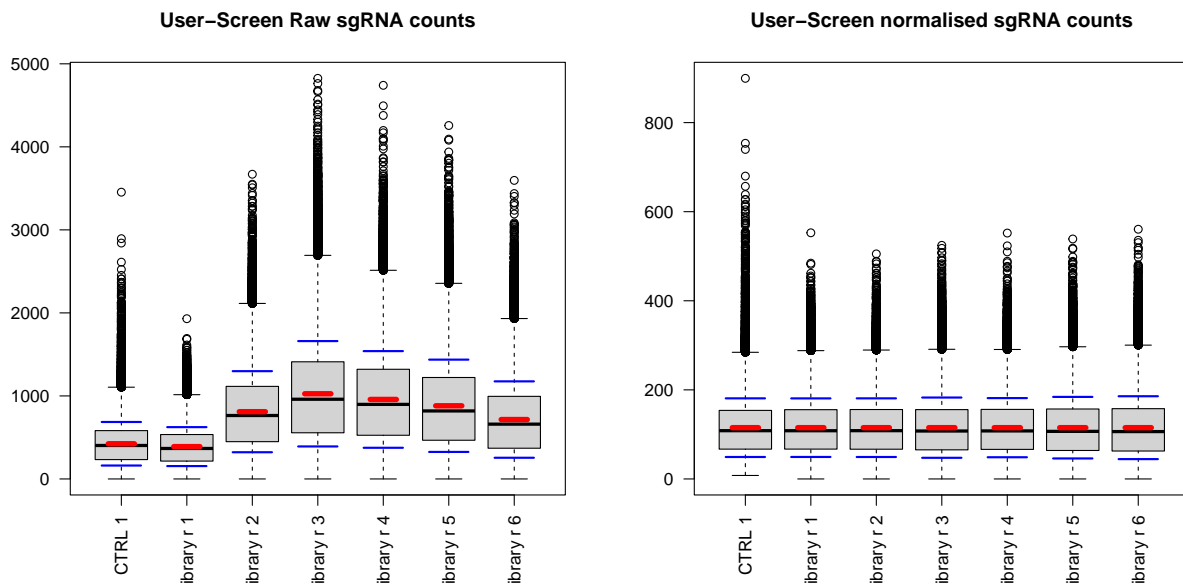
```
# comment these lines of codes if you have already provided your own data
URL <- 'https://figshare.com/ndownloader/files/36658530?private_link=5b2a579791c47a417474'
download.file(URL, destfile = paste0(tmpDir, '/Example_UserScreen.tsv'))
userData <- paste0(tmpDir, '/Example_UserScreen.tsv')
```
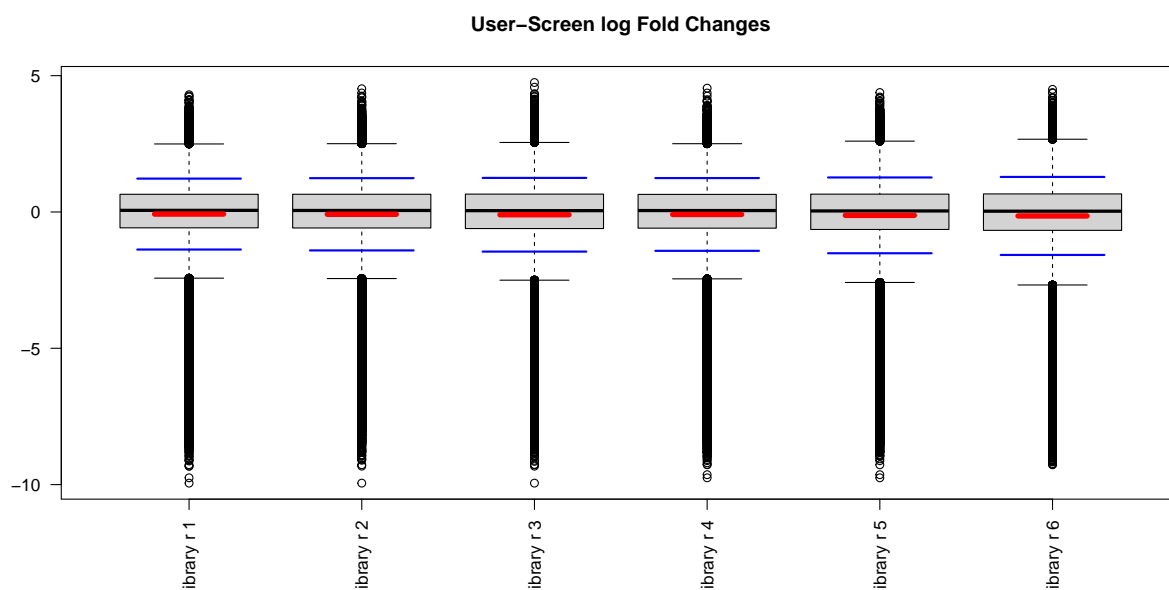
Instead, if the User have its own screening data and want to validate them using the QC pipeline we provide, he/she must specify the path to this file in the chunk below. This must be a tab delimited file (.tsv) with one row per sgRNA and columns/headers representing sgRNA identifiers, HGNC symbols of the genes targeted, sgRNAs' counts for controls and samples (see the 'ccr.NormfoldChanges' function for more details).

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)
```


User−Screen Raw sgRNA counts   User−Screen normalised sgRNA counts

User−Screen log Fold Changes

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

```
## [1] "Downloading data for HT29_c903 experiment (1 of 6)"
## [1] "Downloading data for HT29_c904 experiment (2 of 6)"
## [1] "Downloading data for HT29_c905 experiment (3 of 6)"
## [1] "Downloading data for HT29_c906 experiment (4 of 6)"
## [1] "Downloading data for HT29_c907 experiment (5 of 6)"
## [1] "Downloading data for HT29_c908 experiment (6 of 6)"
```

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)
```
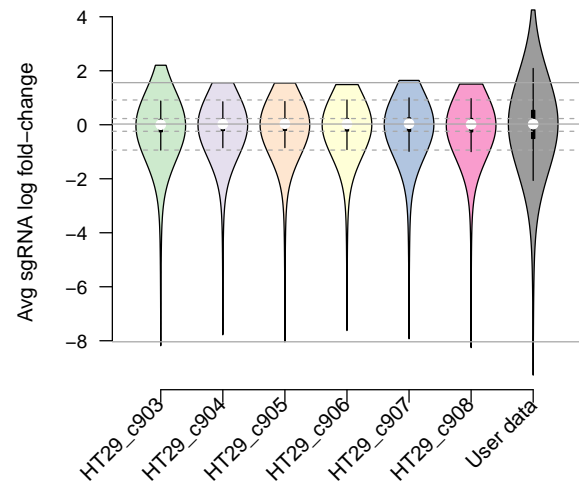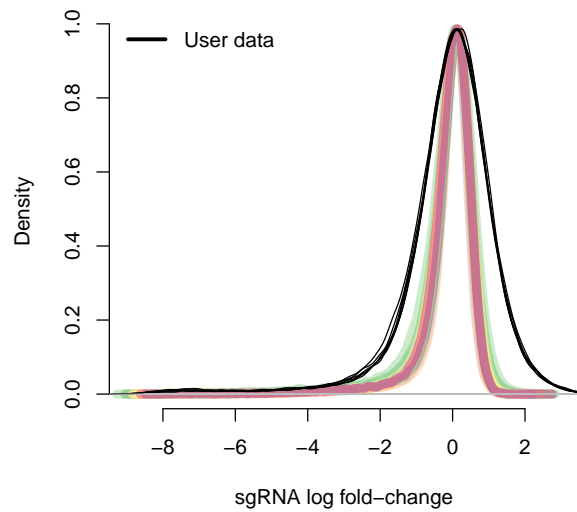
```
## HT29 sgRNAs logFCs statistics:
##
## Avg. Range: -7.97±0.099 ; 1.66±0.111
## Avg. Median: 0.025±0.006
## Avg. IQR range: -0.23±0.01 ; 0.22±0.01
## Avg. 10-90th perc range: -0.68±0.02 ; 0.39±0.01
## Avg. Skewness: -3.81±0.04
## Avg. Kurtosis: 19.53±0.55
##
## User screen sgRNA logFCs statistics:
##
## 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
```

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)
```

```
## 15 pair-wise replicate comparisons (out of 15) yield correlation scores greater or equal than Project
```
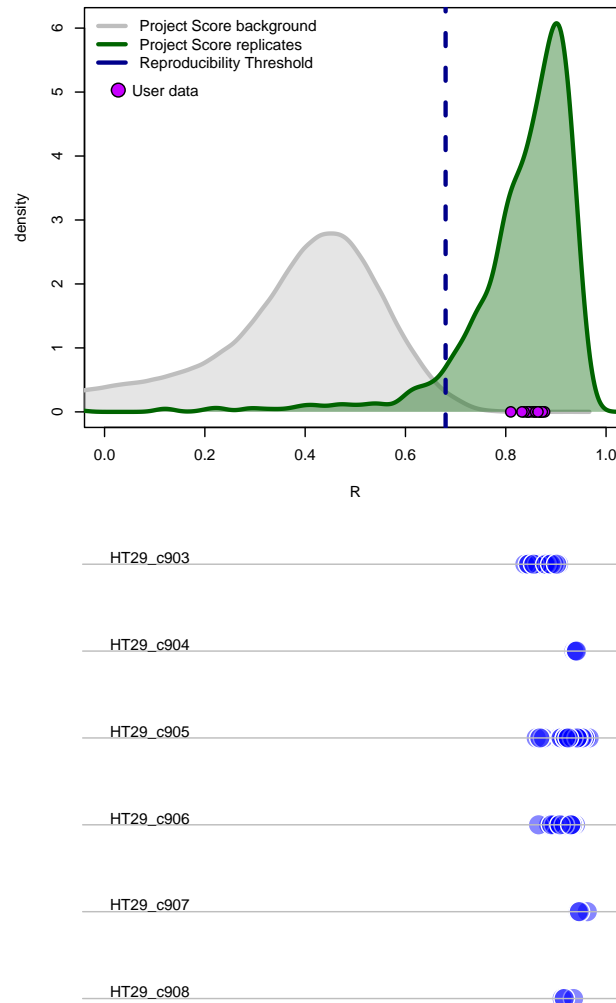

Screen Similarity

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,
```
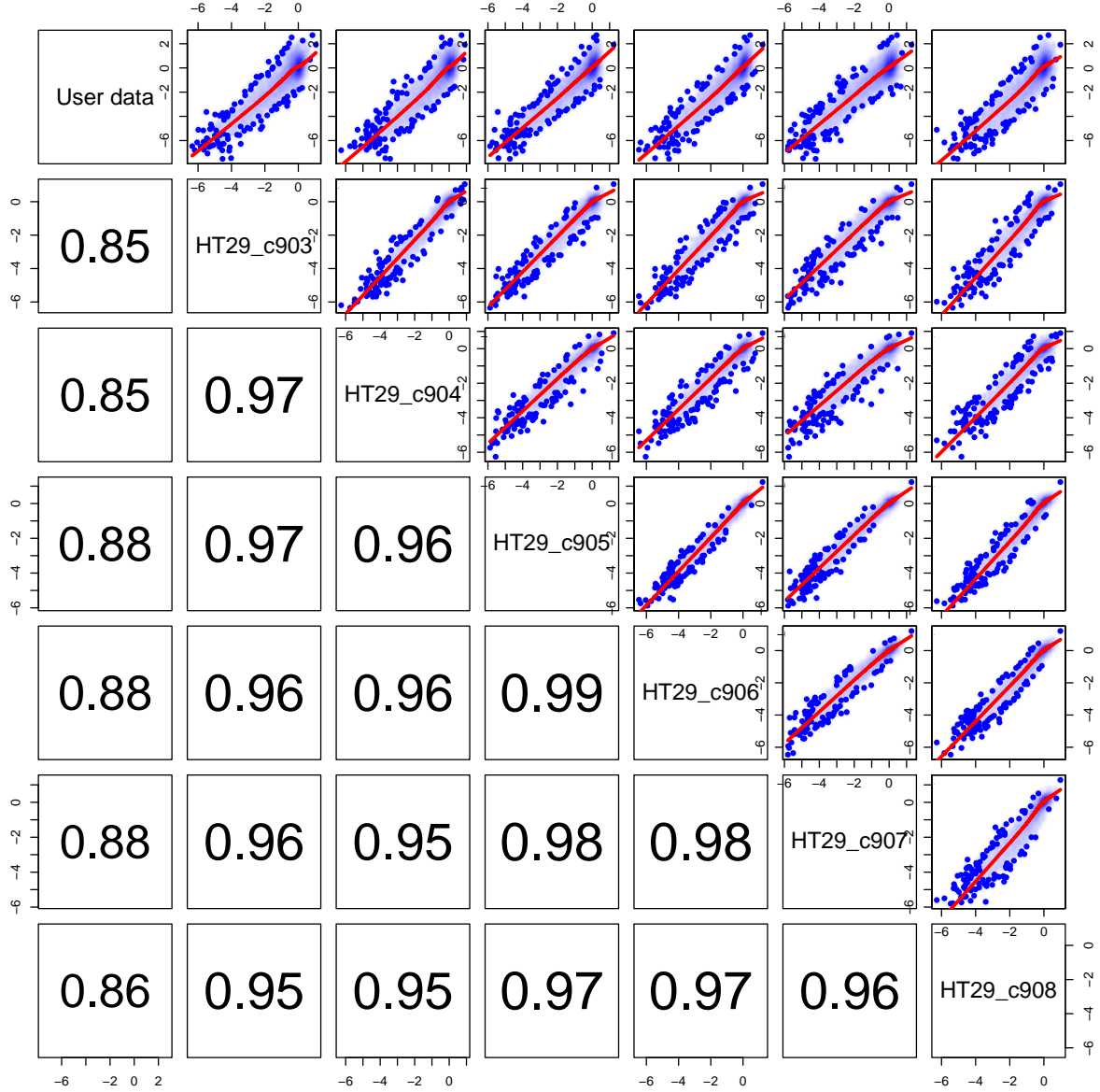
```
                         display = TRUE,
                         userFCs = expData$logFCs)
```

Screen similarity
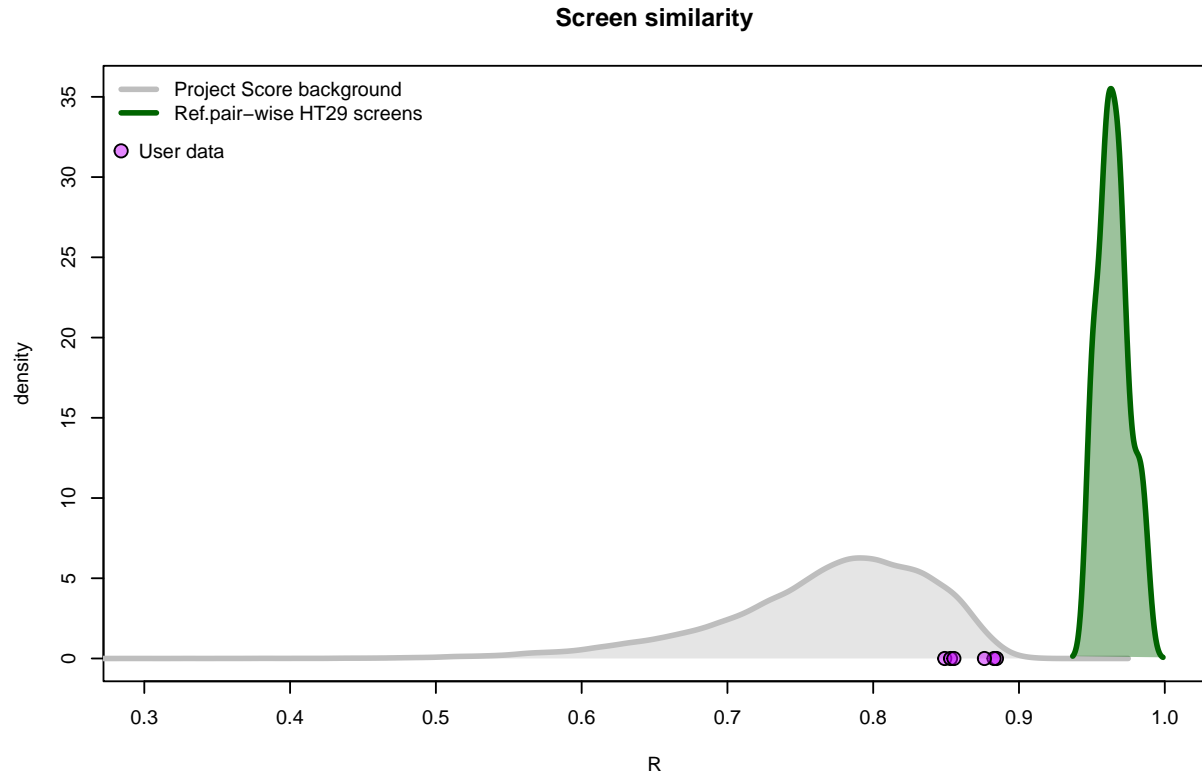
PRJ SCORE BACKGROUND vs REFERENCE = 1.2e−19
USER−DATA vs REFERENCE = 3.1e−06
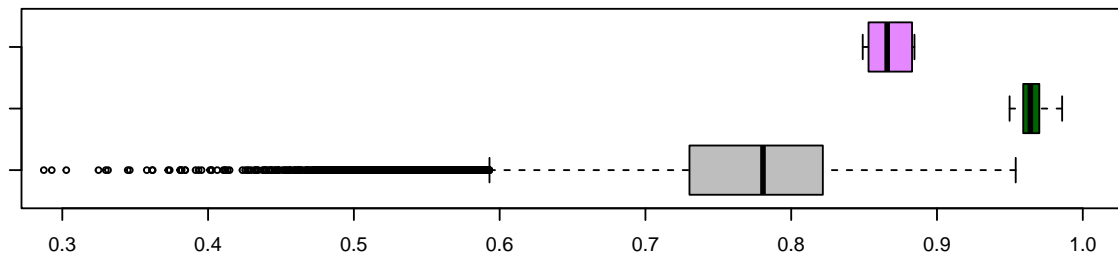
Pearson's correlation matrix

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)
```
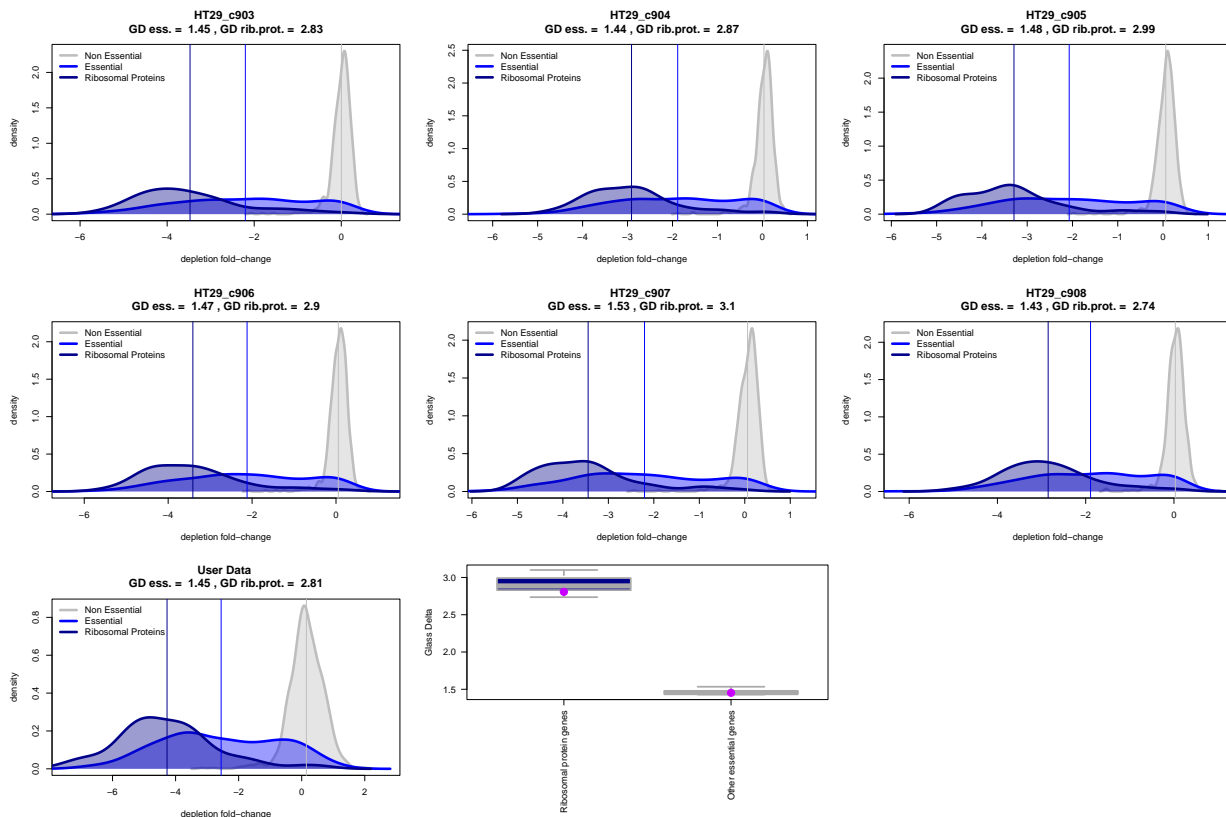

ROC analysis

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.

All curves

```
# 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)
```
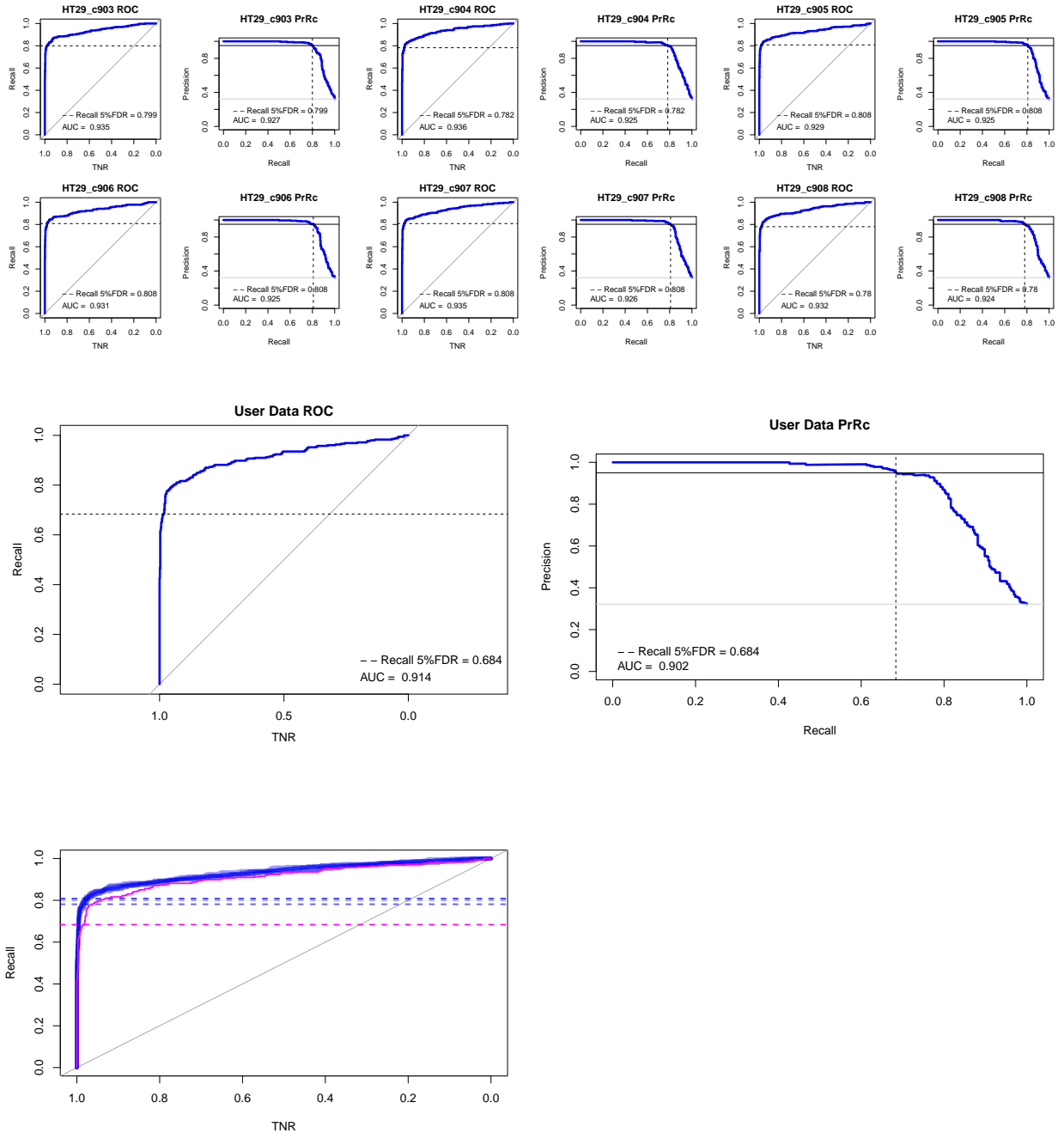
HT-29-specifics genes at 5% FDR

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)
```
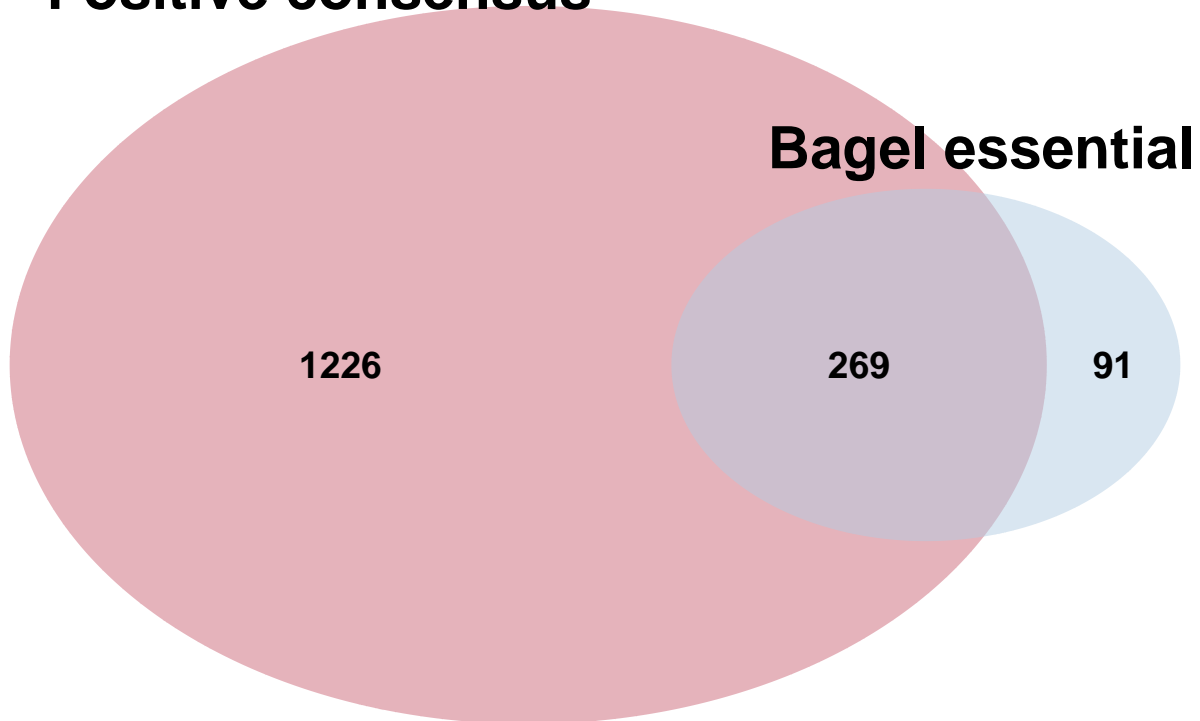
```
## Using group as id variables
```
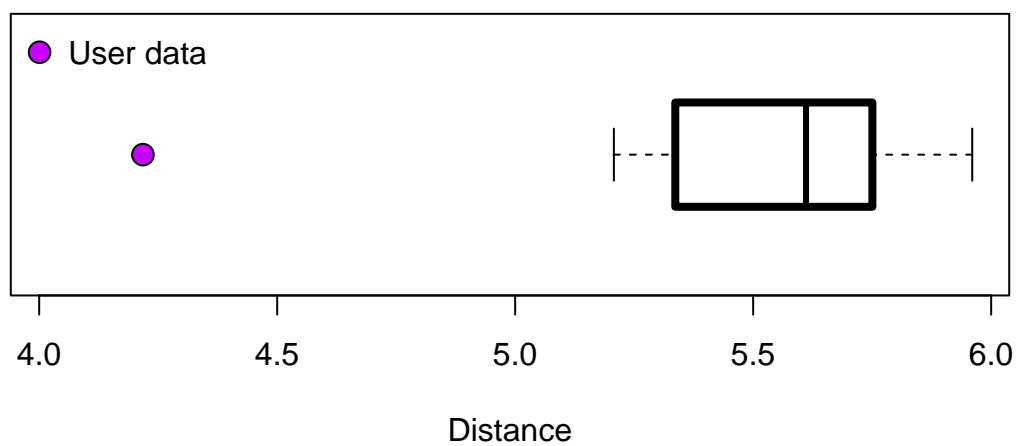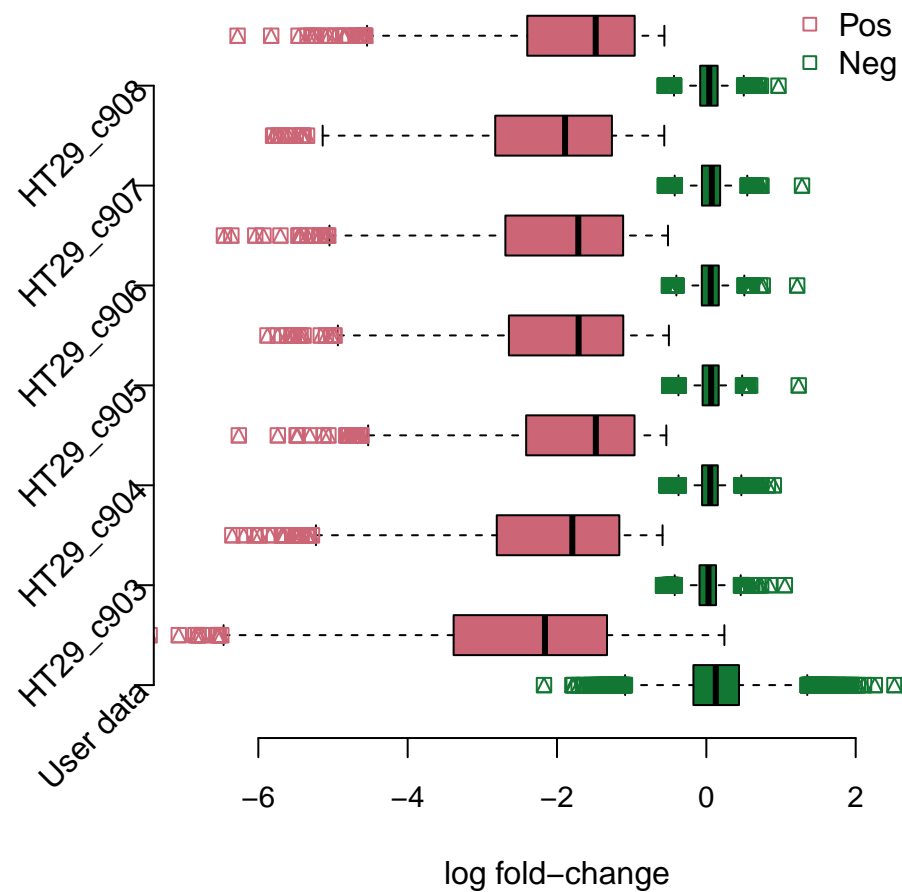
FDR enrichment

Venn diagram

Two-sided Fisher's exact test

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

Fisher's exact test:7.1146537498882e−221

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)
```