HT-29 CRISPR-Cas9 screen assessment

Raffaele M. Iannuzzi, Francesco Iorio

2022-09-03

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 user-provided calibration-screen data is evaluated through different metrics. Outcomes are then contrasted with those obtained when applying the same metrics to a high-quality dataset containing results from screening the HT-29 in multiple batches with the Sanger library (the HT-29 reference dataset).

Portions of and text and code in this document can be commented/removed to generated a quality-control assessment report.

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. A manuscript describing the HT-29 reference dataset and the *HT29benchmark* package is currently under review. 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 chunks load 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.

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

The code chunk 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 setting described in Behan et al. (2019) (see manuscript for more details), and it is 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 uncommented 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 calibration screen of the HT-29 cell line performed with the Sanger library (Tzelepis et al. 2016) and experimental setting described in Behan et al. (2019)
The format of this file should be as it follows. The file should be tab delimited, it should contain one row per sgRNA and the following columns/headers:

- sgRNA: column with alphanumerical identifiers of the sgRNA under consideration;
- gene: column with 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 transfected samples (one column per replicate).

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

The following code download sgRNA depletion fold-changes from the HT-29 reference dataset and stores them into the HT29_resFolder directory. Additionally, individual reference screens file-names are stored in the ref_fn variable

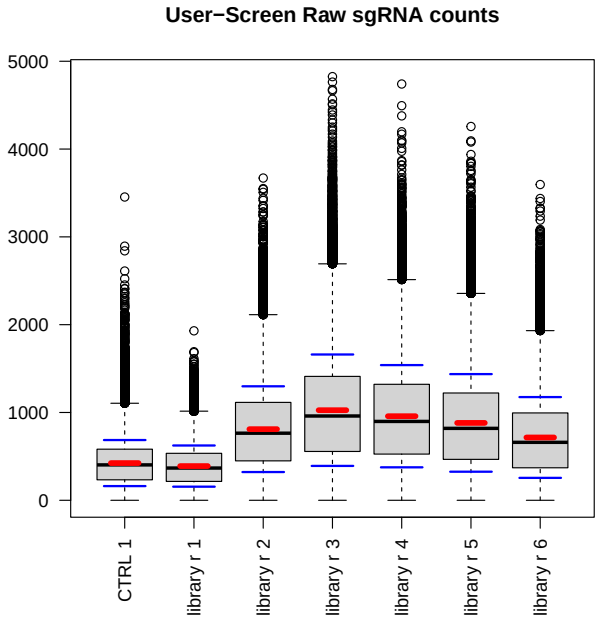
```
HT29R.downloadRefData(destFolder = tmpDir, whatToDownload = 'FCs')
```

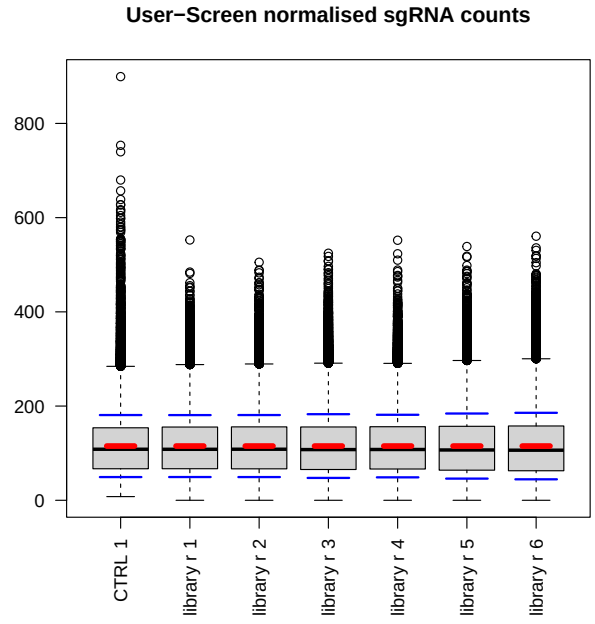
```
ref_fn <- dir(tmpDir)
ref_fn <- grep('_foldChanges.Rdata',ref_fn,value=TRUE)
```

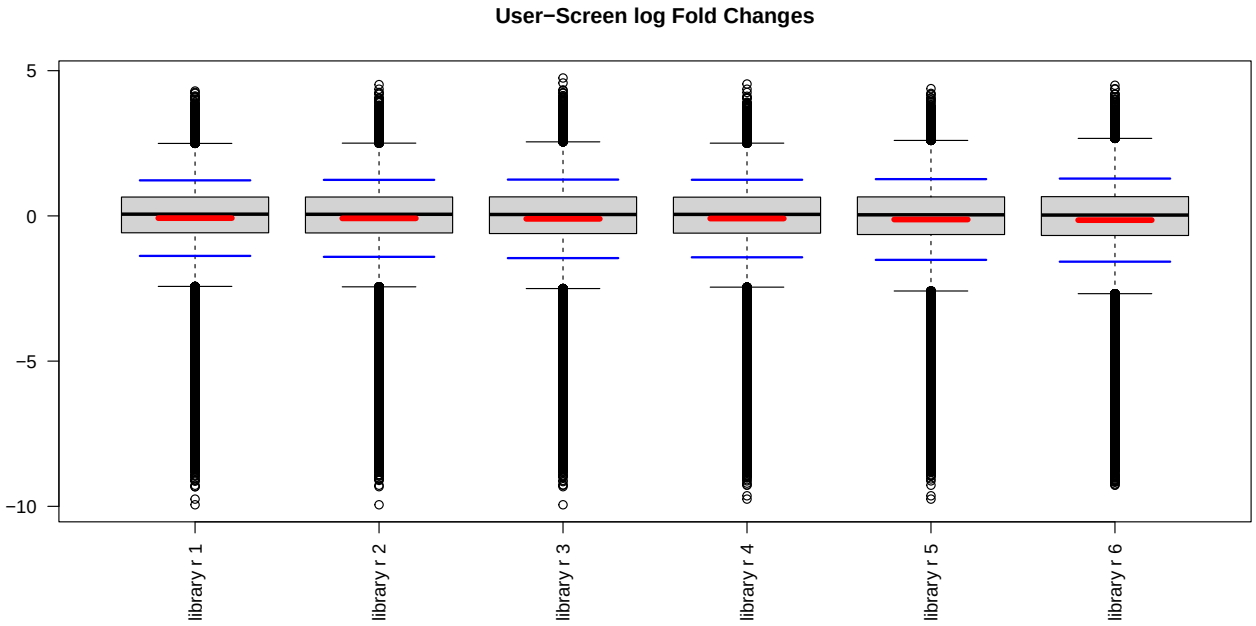
User data normalisation and computation of depletion log fold-changes

The following code normalises user-data 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 is 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). First two columns in these data frames contain sgRNAs' identifiers and HGNC symbols of targete gene, 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)
```





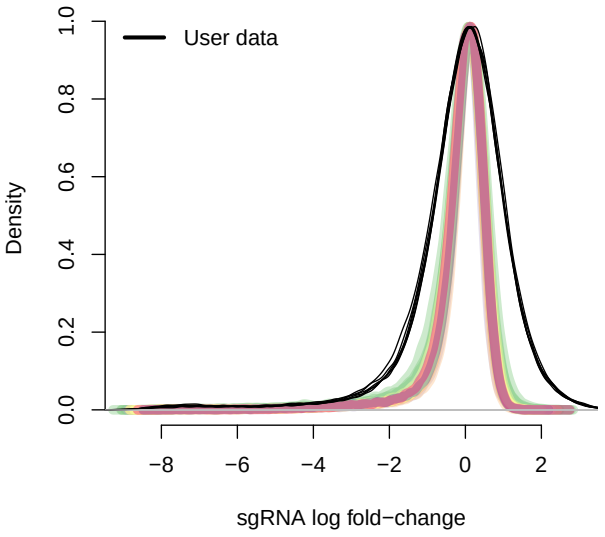


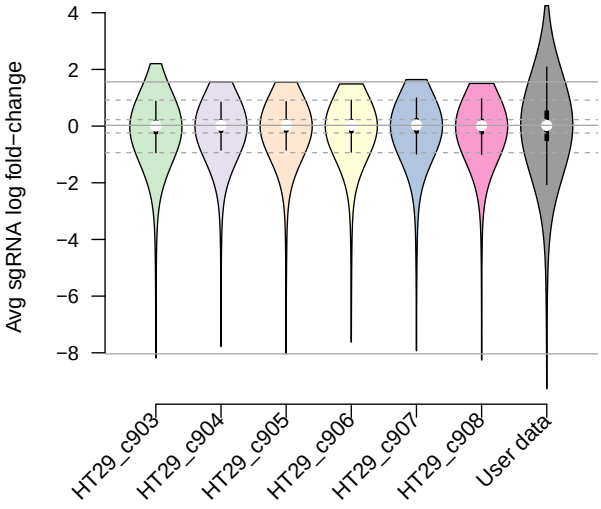
Inspection of sgRNA depletion log fold-change distributions

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 visualised 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:
```

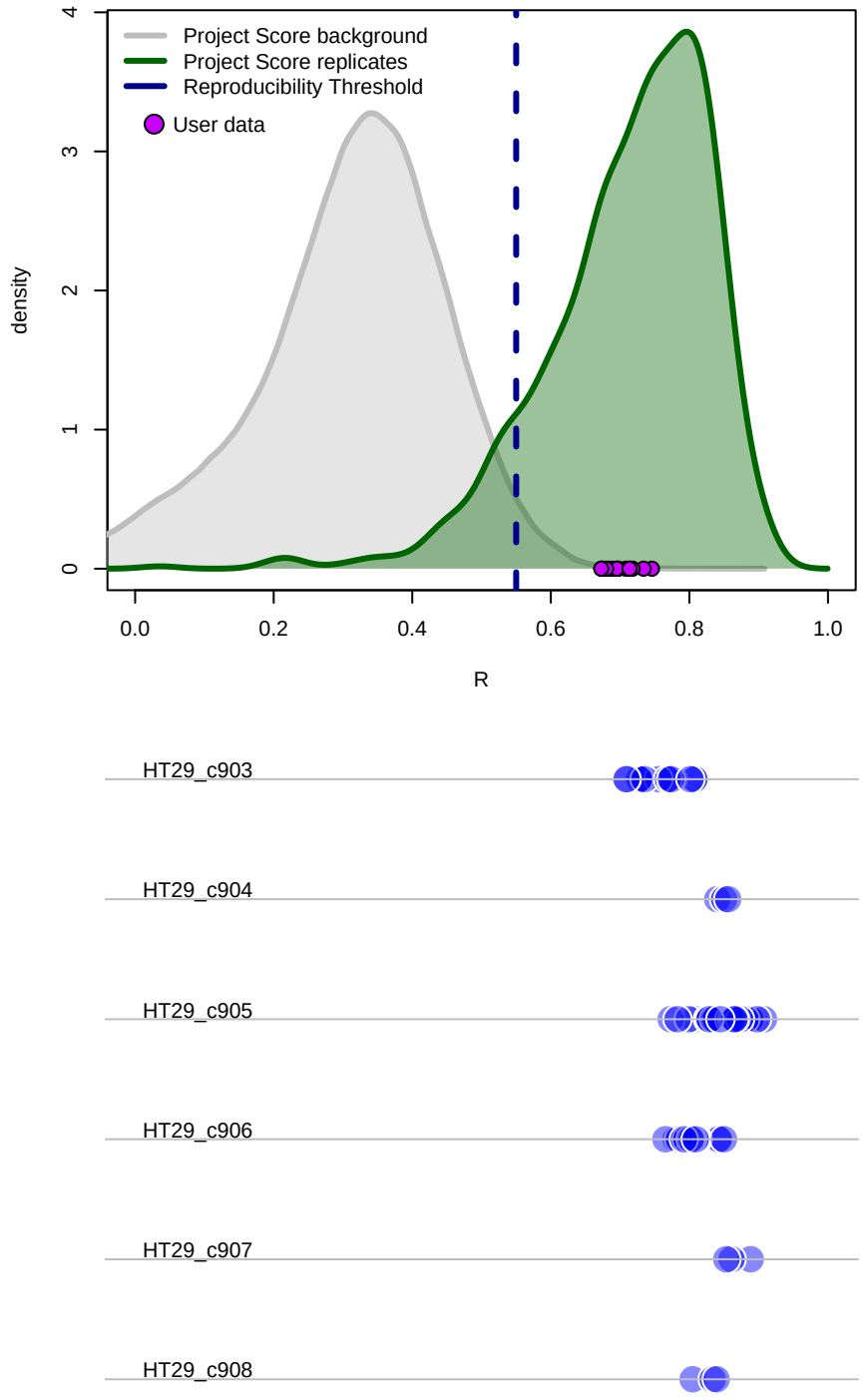
```
##
## Avg. Range: -7.97±0.099 ; 1.66±0.111
## Avg. Median: 0.025±0.006
## Avg. IQR range: -0.23±0.01 ; 0.22±0.01
## Avg. 10-90th perc range: -0.68±0.02 ; 0.39±0.01
## Avg. Skewness: -3.81±0.04
## Avg. Kurtosis: 19.53±0.55
##
## User screen, sgRNAs logFCs statistics:
##
## Range min: -9.273 ; Range 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 assessment

The following code evaluates and compares inter-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 only (defined as in Behan et al. (2019)). These scores are compared with correlation scores from comparing replicates of the same screen across > 200 cell lines (from Project Score (Behan et al. (2019), Dwane et al. (2021)), in green), those 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) and 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)
```

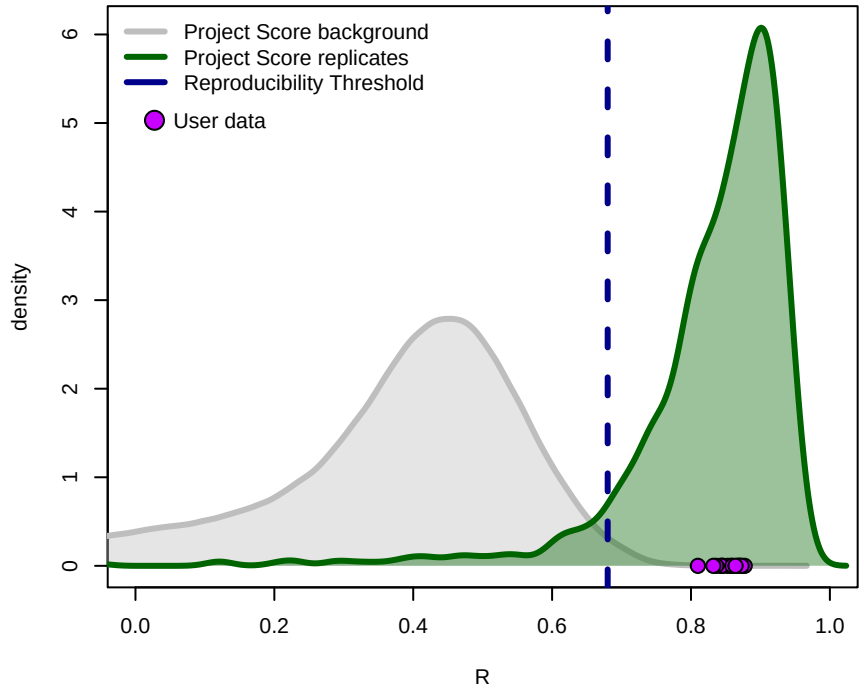
```
## User screen results:
## 15 pair-wise replicate comparisons (out of 15) yield correlation scores greater or
## equal than Project Score QC threshold.
## 0 pair-wise replicate comparisons (out of 15) yield correlation scores lower
## than Project Score QC threshold
```

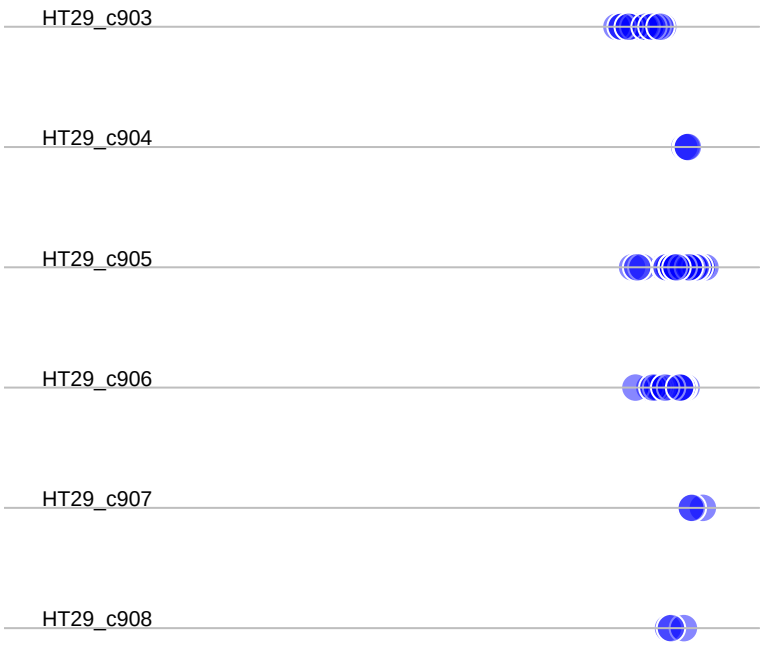


The following code chunk performs the same comparisons of the previous one, but this time at the gene level, collapsing sgRNA depletion log fold-changes by averaging on a targeted gene basis.

```
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
```





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.

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.