CRISPRcleanR: An R package for unsupervised identification and correction of gene independent cell responses to CRISPR-cas9 targeting

Francesco Iorio, fi1@sanger.ac.uk

December 3, 2017

1 Quick start

1.1 Installation

First, you need to install and load the devtools package. You can do this from CRAN. Invoke R and then type.

```
install.packages("devtools")
library(devtools)
```

Secondly, install the CRISPRcleanR with the following command:

```
install_github("francescojm/CRISPRcleanR")
```

1.2 Raw sgRNA count median-ratio normalisation and computation of sgRNAs' log fold-changes

Load the package.

```
    library(CRISPRcleanR)

## Loading required package:  stringr
## Loading required package:  DNAcopy
```

Step 1: Load your sgRNA library annotation. In this example we will use a built in data frame containing the annotation of the SANGER v1.0 library [1]:

```
data(KY_Library_v1.0)
```

To use your own library annotation you will have to put it in a data frame with the same format of the `KY_Library_v1.0` data frame (detailed in the corresponding entry of the reference manual of the CRISPRcleanR package).

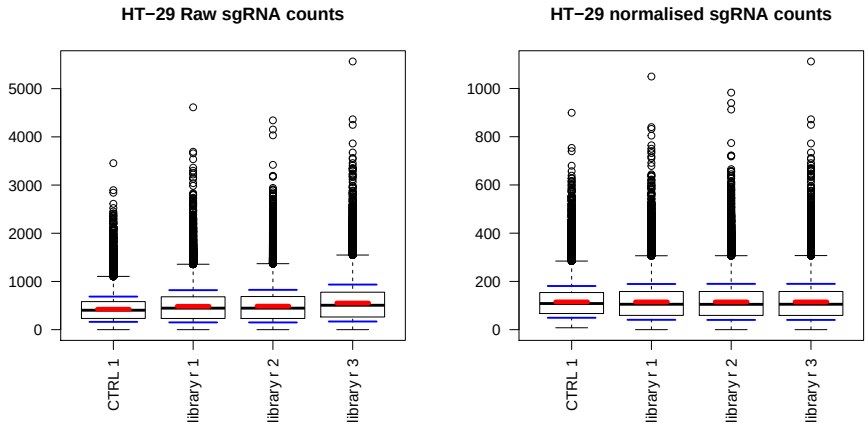
Step 2: Store the path of the tsv file containing your sgRNAs' raw counts in a temporare variable. In this example we will use counts generated upon a CRISPR-Cas9 pooled drop-out screen (described in [2]) built in this package.

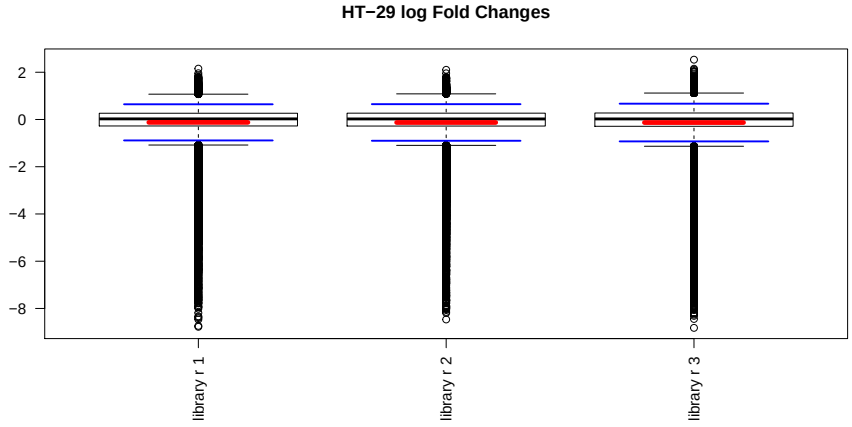
```
fn<-paste(system.file('extdata',package = 'CRISPRcleanR'),
          '/HT-29_counts.tsv',sep='')
```

The tsv with the sgRNAs' raw counts must be formatted as specified in the reference manual entry for function `ccr.NormfoldChanges`.

Step 3: Performing median-ratio normalisation of raw counts and computing sgRNAs' log fold-changes. In this example we will exclude sgRNAs with less than 30 reads in the plasmid sample.

```
normANDfcs<-ccr.NormfoldChanges(fn,
                                min_reads=30,
                                EXPname='HT-29',
                                libraryAnnotation=KY_Library_v1.0)
```





This function returns a list of two data frames, respectively with normalised counts and log fold-changes, and it saves the as Robject in the directory whose path is specified with the parameter `outdir` (set to `'./'` by default).

```
head(normANDfcs$norm_counts)

##                                        sgRNA gene ERS717283.plasmid
## 1 A1BG_CCDS12976.1_ex3_19:58862927-58862950:-_5-1 A1BG         292.14621
## 2 A1BG_CCDS12976.1_ex4_19:58863655-58863678:+_5-2 A1BG         151.02032
## 3 A1BG_CCDS12976.1_ex4_19:58863697-58863720:-_5-3 A1BG         209.08503
## 4 A1BG_CCDS12976.1_ex4_19:58863866-58863889:+_5-4 A1BG         110.40106
## 5 A1BG_CCDS12976.1_ex5_19:58864367-58864390:-_5-5 A1BG          95.81979
## 6  A1CF_CCDS7241.1_ex6_10:52588014-52588037:-_5-1 A1CF          60.92889
##   HT29_c904R1 HT29_c904R2 HT29_c904R3
## 1   308.05192   354.89835   305.56806
## 2   145.38048   113.16912   166.47364
## 3   280.97793   203.70441   223.03071
## 4    80.31191    64.05372    78.74023
## 5    78.71932   123.80701   103.32157
## 6    47.77762    76.50232    59.75464

head(normANDfcs$logFCs)

##                                        sgRNA gene HT29_c904R1
## 1 A1BG_CCDS12976.1_ex3_19:58862927-58862950:-_5-1 A1BG  0.07635566
## 2 A1BG_CCDS12976.1_ex4_19:58863655-58863678:+_5-2 A1BG -0.05472442
## 3 A1BG_CCDS12976.1_ex4_19:58863697-58863720:-_5-3 A1BG  0.42548611
## 4 A1BG_CCDS12976.1_ex4_19:58863866-58863889:+_5-4 A1BG -0.45663336
## 5 A1BG_CCDS12976.1_ex5_19:58864367-58864390:-_5-5 A1BG -0.28197989
## 6  A1CF_CCDS7241.1_ex6_10:52588014-52588037:-_5-1 A1CF -0.34756262
##   HT29_c904R2 HT29_c904R3
## 1  0.28027938  0.06469488
## 2 -0.41467098  0.14010902
## 3 -0.03752168  0.09293733
## 4 -0.78070106 -0.48496821
## 5  0.36800353  0.10820208
## 6  0.32598467 -0.02784489
```

IMPORTANT: if there are control replicates in your sgRNAs count file their number must be specified by in the parameter `ncontrols` (equal to 1 by default) of the `ccr.NormfoldChanges` function.

1.3 Genome sorting of sgRNAs' log fold-changes and their correction for gene independent responses to CRISPR-Cas9 targeting

Step 1: Map genome-wide sgRNAs' log fold changes (averaged across replicates) on the genome, sorted according to their positions of the targeted region on the chromosomes.

```
gwSortedFCs<-
    ccr.logFCs2chromPos(normANDfcs$logFCs,KY_Library_v1.0)
```

```
head(gwSortedFCs)
```

```
##                                     CHR startp   endp  genes
## SAMD11_CCDS2.2_ex3_1:871254-871277:+_5-1   1 871254 871277 SAMD11
## SAMD11_CCDS2.2_ex4_1:874451-874474:-_5-2   1 874451 874474 SAMD11
## SAMD11_CCDS2.2_ex4_1:874487-874510:+_5-3   1 874487 874510 SAMD11
## SAMD11_CCDS2.2_ex5_1:874693-874716:+_5-4   1 874693 874716 SAMD11
## SAMD11_CCDS2.2_ex6_1:876601-876624:-_5-5   1 876601 876624 SAMD11
## NOC2L_CCDS3.1_ex8_1:887388-887411:+_5-1    1 887388 887411  NOC2L
##                                        avgFC       BP
## SAMD11_CCDS2.2_ex3_1:871254-871277:+_5-1 -0.12965287 871265.5
## SAMD11_CCDS2.2_ex4_1:874451-874474:-_5-2  0.09329615 874462.5
## SAMD11_CCDS2.2_ex4_1:874487-874510:+_5-3  0.25286616 874498.5
## SAMD11_CCDS2.2_ex5_1:874693-874716:+_5-4 -0.05128489 874704.5
## SAMD11_CCDS2.2_ex6_1:876601-876624:-_5-5 -0.02110076 876612.5
## NOC2L_CCDS3.1_ex8_1:887388-887411:+_5-1  -1.27571756 887399.5
```

Step 2: Identify and correct biased sgRNAs' log fold-changes putatively due to gene independent responses to CRISPR-Cas9 targeting (this function calls iteratively the `ccr.cleanChrm` function, which performs the correction in each chromosome individually). In this example we are using a completely unsuerpvised approach and correcting chromosomal segments of equal sgRNA log fold-changes if they include sgRNAs targeting at least 3 different genes, and without making any assumption on gene essentiality nor knowing *a priori* the copy number status of the included genes [2].

```
correctedFCs<-ccr.GWclean(gwSortedFCs,display=TRUE,label='HT-29')
```

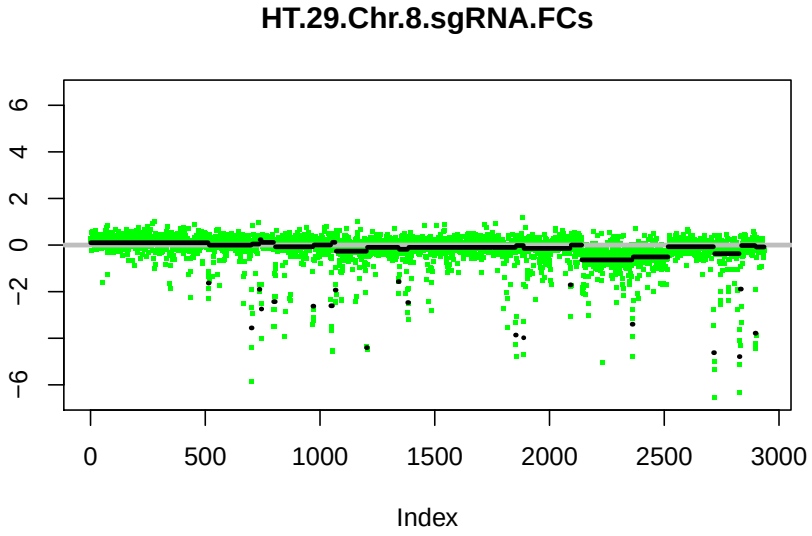
The corrected sgRNAs fold-changes are returned in a list (as a data frame), together with another data frame with annotation of the identified segments and a vector of strings containing all the sgRNAs identifier genome-sorted.

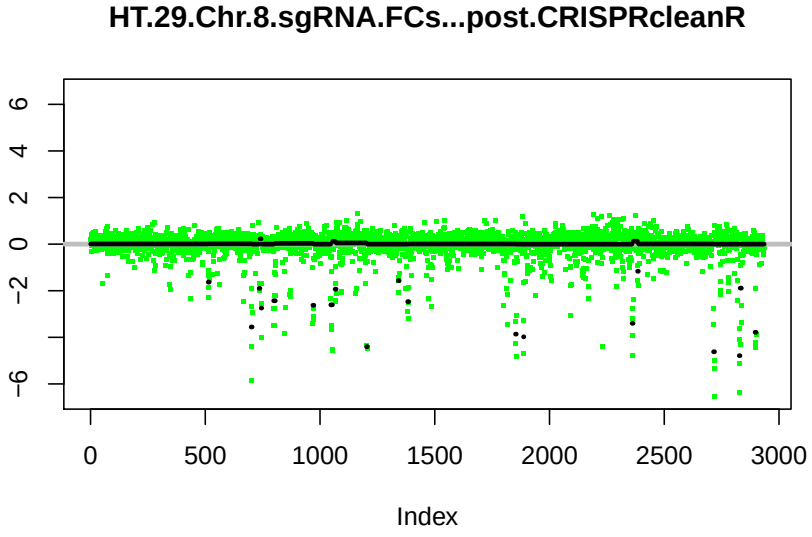
```
    head(correctedFCs$corrected_logFCs)
```

```
##                                     CHR startp   endp  genes
## SAMD11_CCDS2.2_ex3_1:871254-871277:+_5-1   1 871254 871277 SAMD11
## SAMD11_CCDS2.2_ex4_1:874451-874474:-_5-2   1 874451 874474 SAMD11
## SAMD11_CCDS2.2_ex4_1:874487-874510:+_5-3   1 874487 874510 SAMD11
## SAMD11_CCDS2.2_ex5_1:874693-874716:+_5-4   1 874693 874716 SAMD11
## SAMD11_CCDS2.2_ex6_1:876601-876624:-_5-5   1 876601 876624 SAMD11
## NOC2L_CCDS3.1_ex8_1:887388-887411:+_5-1    1 887388 887411  NOC2L
##                                        avgFC       BP correction
## SAMD11_CCDS2.2_ex3_1:871254-871277:+_5-1 -0.12965287 871265.5          0
## SAMD11_CCDS2.2_ex4_1:874451-874474:-_5-2  0.09329615 874462.5          0
## SAMD11_CCDS2.2_ex4_1:874487-874510:+_5-3  0.25286616 874498.5          0
## SAMD11_CCDS2.2_ex5_1:874693-874716:+_5-4 -0.05128489 874704.5          0
## SAMD11_CCDS2.2_ex6_1:876601-876624:-_5-5 -0.02110076 876612.5          0
## NOC2L_CCDS3.1_ex8_1:887388-887411:+_5-1  -1.27571756 887399.5          0
##                                     correctedFC
## SAMD11_CCDS2.2_ex3_1:871254-871277:+_5-1 -0.12965287
## SAMD11_CCDS2.2_ex4_1:874451-874474:-_5-2  0.09329615
## SAMD11_CCDS2.2_ex4_1:874487-874510:+_5-3  0.25286616
## SAMD11_CCDS2.2_ex5_1:874693-874716:+_5-4 -0.05128489
## SAMD11_CCDS2.2_ex6_1:876601-876624:-_5-5 -0.02110076
## NOC2L_CCDS3.1_ex8_1:887388-887411:+_5-1  -1.27571756
```

Details on how the data frame with the corrected sgRNAs fold-changes should be interpreted can be found in the entry of the `ccr.GWclean` function in the package reference manual.

This function also produces one plot per chromosome, with segments of sgRNAs' equal log fold-changes before and after the correction. An example of these plot is reported below (chromosome 8, in HT-29 with the region containing *MYC* highly biased toward consistent negative fold-changes)





1.4 Correcting sgRNAs' treatment counts for mean-variance modeling

In order to apply the inverse transformation described in [2], thus to derive corrected normalised sgRNAs' treatment counts from corrected log fold-changes, it is sufficient to run the function `ccr.correctCounts` as follows:

```
correctedCounts<-ccr.correctCounts('HT-29',
                                   normANDfcs$norm_counts,
                                   correctedFCs,
                                   KY_Library_v1.0,
                                   minTargetedGenes=3,
                                   OutDir='./')
```

With the plasmid counts, are suitable for mean-variance modeling approach (such that implemented in MAGeCK[3]).

```
head(correctedCounts)
```

```
##                                                sgRNA gene ERS717283.plasmid
## 1 A1BG_CCDS12976.1_ex3_19:58862927-58862950:-_5-1 A1BG          292.14621
## 2 A1BG_CCDS12976.1_ex4_19:58863655-58863678:+_5-2 A1BG          151.02032
## 3 A1BG_CCDS12976.1_ex4_19:58863697-58863720:-_5-3 A1BG          209.08503
## 4 A1BG_CCDS12976.1_ex4_19:58863866-58863889:+_5-4 A1BG          110.40106
## 5 A1BG_CCDS12976.1_ex5_19:58864367-58864390:-_5-5 A1BG           95.81979
## 6  A1CF_CCDS7241.1_ex6_10:52588014-52588037:-_5-1 A1CF           60.92889
##   HT29_c904R1 HT29_c904R2 HT29_c904R3
## 1   309.77863   356.73522   307.28892
## 2   144.74469   112.89328   165.60212
## 3   280.34458   203.51866   222.73301
## 4    80.64680    64.52159    79.08797
## 5    78.22697   122.47062   102.36866
## 6    45.21299    71.83850    56.31473
```

This function also saves the correctedCounts as Rdata object at the location specified by the parameter `OutDir`. To run MAGeCK, using these corrected sgRNAs' counts you will need to save them as a tsv file first:

```
write.table(correctedCounts,
            quote=FALSE,
            row.names = FALSE,
            sep='\t',
            file='./HT-29_mgk_input_corrected.tsv')
```

then use this file as input for MAGeCK.

IMPORTANT: the corrected sgRNAs' count are already median-normalised therefore, when executing MAGeCK, the parameter `--norm-method` should be set to `none`.

2 Visualisation and assessment of Results

Coming Soon

References

[1] Konstantinos Tzelepis et al. "A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia." In: *Cell reports* 17.4 (Oct. 2016), pp. 1193–1205.

[2] Francesco Iorio et al. "Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting". In: *revision* 0.0 (), pp. 0–0.

[3] Wei Li et al. "MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens." In: *Genome Biology* 15.12 (2014), p. 554.