CRISPR Screen and Gene Expression Differential Analysis

Lianbo Yu, Yue Zhao, and Lang Li

2021-04-21

Contents

0.1 Introduction

CEDA is developed for analyzing read counts of single guide RNAs (sgRNAs) by CRISPR screen experiments. sgRNAs are synthetically generated from genes and each gene can generate multiple sgRNAs. CEDA models sgRNA counts at different levels of gene expression by multi-component normal mixtures and EM algorithms. Posterior estimates at sgRNA level are then summarized for each gene.

In this document, we use a MDA231 cell experiment as an example to demonstrate how to use CEDA to perform CRISPR screen data analysis.

0.2 Data and Normalization

Three samples of MDA231 cells were untreated at T0, and another three samples of MDA231 cells were treated with DMSO at T0. We are interested in detecting sgRNAs that are differentially changed by treatment.

sgRNA read counts along with a list of non-essential genes are stored in the dataset mda231 in CEDA.

```
library(CEDA)
data("mda231")
dim(mda231$sgRNA)
#> [1] 70855     9
length(mda231$neGene$Gene)
#> [1] 350
head(mda231$sgRNA)
#>                                 sgRNA     Gene DMSOa DMSOb DMSOc T0a  T0b T0c
#> 1 chr12:48578255-48578274_C12orf68_+ C12orf68   379   219   732 477  473 441
#> 2  chr9:91940499-91940518_SECISBP2_+ SECISBP2   320   380   497 671  744 583
#> 3   chr2:228678687-228678706_CCL20_-    CCL20   836   517   687 920 1008 896
#> 4         chr9:4740975-4740994_AK3_-      AK3  1324   558  1198 580  676 524
#> 5   chr4:147830251-147830270_TTC29_+    TTC29   745   276   376 291  287 254
#> 6     chr1:9658637-9658656_TMEM201_+  TMEM201   721   511   512 659  918 679
#>   exp.level.log2
#> 1     0.05693792
#> 2     2.50903958
#> 3     0.30585849
#> 4     3.65727414
#> 5     0.03231981
#> 6     2.44136661
```

sgRNA read counts needs to be normalized across sample replicates before formal analysis. Non-essential genes are assumed to have no change after DMSO treatment. Median normalization factors of the non-essential genes were used for normalizing sgRNA counts of all samples.

```
mda231.ne <- mda231$sgRNA[mda231$sgRNA$Gene %in% mda231$neGene$Gene,]
cols <- c(3:8)
mda231.norm <- medianNormalization(mda231$sgRNA[,cols], mda231.ne[,cols])[[2]]
```

0.3 Analysis

Our goal is to detect essential sgRNAs that have different count levels between conditions. R package limma was used to calculate log fold ratios between three untreated and three treated samples.

0.3.1 Calculating fold ratios

```
library(limma)
group <- gl(2,3,labels=c("Control","Baseline"))
design <- model.matrix(~  0 + group)
colnames(design) <- sapply(colnames(design),function(x) substr(x,6,nchar(x)))
contrast.matrix <- makeContrasts("Control-Baseline",levels=design)
limma.fit <- limma(log2(mda231.norm+1),design,contrast.matrix)
```

Then results from limma analysis were merged with sgRNA counts.

```
mda231.limma <- data.frame(mda231$sgRNA,limma.fit)
head(mda231.limma)
#>                                 sgRNA     Gene DMSOa DMSOb DMSOc T0a  T0b T0c
#> 1 chr12:48578255-48578274_C12orf68_+ C12orf68   379   219   732 477  473 441
#> 2  chr9:91940499-91940518_SECISBP2_+ SECISBP2   320   380   497 671  744 583
#> 3   chr2:228678687-228678706_CCL20_-    CCL20   836   517   687 920 1008 896
#> 4         chr9:4740975-4740994_AK3_-      AK3  1324   558  1198 580  676 524
#> 5   chr4:147830251-147830270_TTC29_+    TTC29   745   276   376 291  287 254
#> 6     chr1:9658637-9658656_TMEM201_+  TMEM201   721   511   512 659  918 679
#>   exp.level.log2        lfc        se          p
#> 1     0.05693792 -0.1956716 0.4062625 0.57786811
#> 2     2.50903958 -0.7136543 0.1756822 0.00504191
#> 3     0.30585849 -0.4534573 0.1416629 0.02218640
#> 4     3.65727414  0.7422065 0.3019398 0.02557574
#> 5     0.03231981  0.6609765 0.3844398 0.07906013
#> 6     2.44136661 -0.3328908 0.1946823 0.11712252
```

0.3.2 Fold ratios under the null hypotheses

Under the null hypothses, all sgRNAs levels are unchanged between the two conditions. To obtain fold ratios under the null, samples were permutated between two conditions, log fold ratios were obtained from limma analysis under each permutation.

```
betanull <- limmaPermutation(log2(mda231.norm+1),design,contrast.matrix,20)
theta0 <- sd(betanull)
theta0
#> [1] 0.4555191
```

0.3.3 Fitting three-component mixture models

A three-component mixture model is assumed for log fold ratios at different level of gene expression. Empirical Bayes method was employed to estimate paremeters of the mixtures and posterior means were obtained for

estimating actual log fold ratios between the two conditions. P-values of sgRNAs were then calculated by permutation method.

```
nmm.fit <- normalMM(mda231.limma,theta0)
```

Results from the mixture model were shown in Figure 1. False discovery rate of 0.05 was used for declaring significant changes in red color between the two conditions for sgRNAs.

```
scatterPlot(nmm.fit$data,fdr=0.05,xlim(-0.5,12),ylim(-8,5))
```

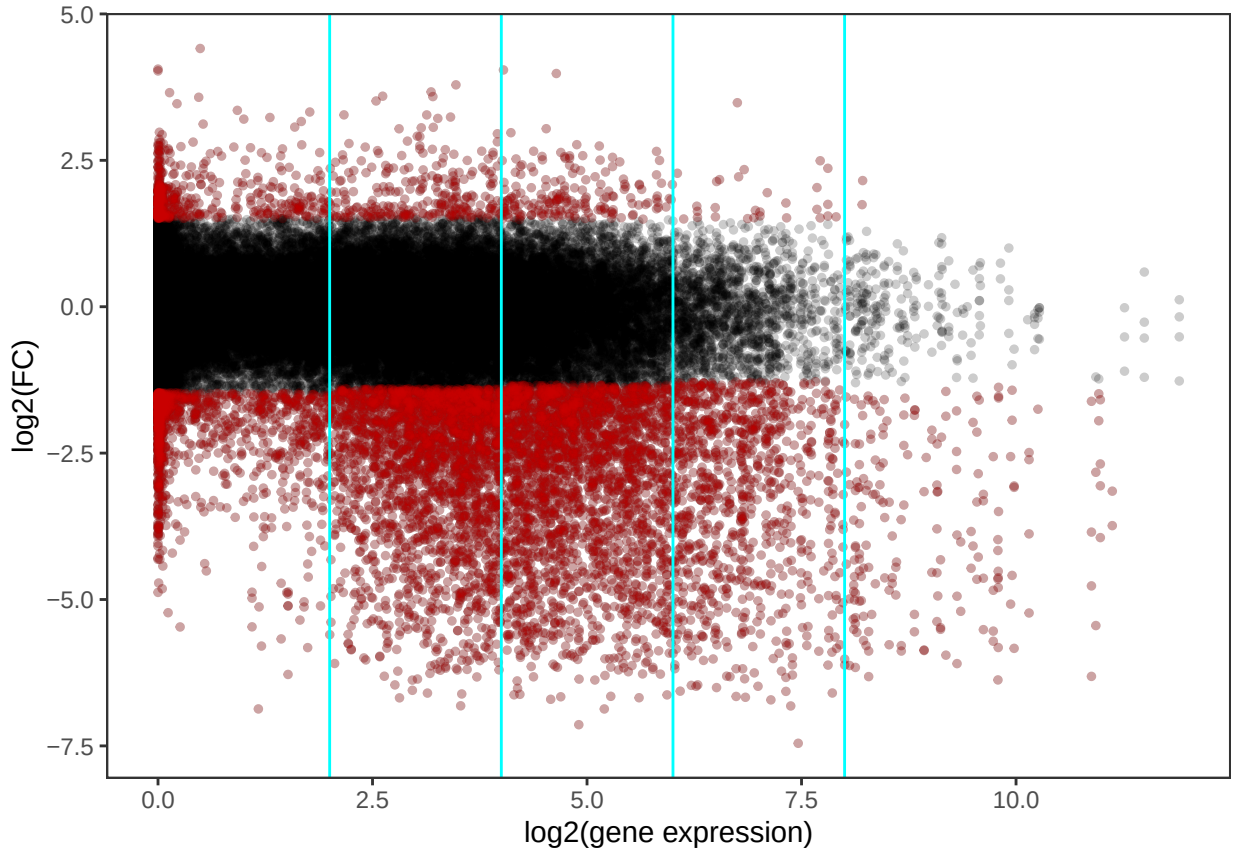

Figure 1: Log fold ratios of sgRNAs vs. gene expression level

0.3.4 Gene level summarization

From the p-values of sgRNAs, gene level p-values were obtained by using modified robust rank aggregation method (alpha-RRA). Log fold ratios were also summarized at gene level.

```
mda231.nmm <- nmm.fit[[1]]
p.gene <- calculateGenePval(exp(mda231.nmm$log_p), mda231.nmm$Gene, 0.05)
fdr.gene <- stats::p.adjust(p.gene$pvalue, method = "fdr")
lfc.gene <- calculateGeneLFC(mda231.nmm$lfc, mda231.nmm$Gene)
```