CRISPhieRmix Manual

Timothy Daley

8/2/2018

Contents

Contents

1 Introduction to CRISPhieRmix

CRISPhieRmix is an R package for analysing large CRISPR interference and activation (CRISPRi/a) screens. CRISPhieRmix uses a hierarchical mixture model approach to identify genes that are unlikely to follow the null distribution. CRISPhieRmix assumes that genes follow a mixture of null genes (genes with no effect) and non-null genes (genes with a significant effect). All guides from null genes follow a common null distribution. The guides for the non-null genes, on the other hand, are assumed to follow a mixture distribution, where some guides are ineffective (possibly with little or no change in the target gene expression) and follow the null distribution and some guides have an effect and follow an alternative distribution.

CRISPhieRmix can be used with out without negative control guides. We find that negative control guides help to better model the null distribution, as in our experience the null distribution tends to have long tails, and this helps to control the false discovery rate. A critical assumption is that the negative control guides accurcately reflect the null distribution. This assumption can be violated in some screens, and we will discuss this later in depth.

If you have any issues with software, please create an issue in the github page https://github.com/timydaley/CRISPhieRmix. If you have any questions, please email me at tdaley@stanford.edu.

2 Input parameters for CRISPhieRmix

2.1 Input

- **x** log2 fold changes of guides targeting genes (required)
- **geneIds** gene ids corresponding to x (required)
- **negCtrl** log2 fold changes of negative control guides; if negCtrl is not included then CRISPhieRmix uses a normal hierachical mixture model
- **max_iter** maximum number of iterations for EM algorithm, default = 100
- **tol** tolerance for convergence of EM algorithm, default = 1e-10
- **pq** initial value of p·q, default = 0.1
- **mu** initial value of mu for the interesting genes, default = -4
- **sigma** initial value of sigma for the interesting genes, default = 1
- **nMesh** the number of points to use in numerical integration of posterior probabilities, default = 100
- **BIMODAL** boolean variable for BIMODAL mode to fit both positive and negative sides, used for cases such as Jost et al. 2017 where both sides are of interest.
- **VERBOSE** boolean variable for VERBOSE mode, default = FALSE
- **PLOT** boolean variable to produce plots, default = FALSE

2.2 Return

- **mixFit** a list containing the mixture fit for **x**
- **genes** a vector of genes from **geneIds**, in the same order as the following return values
- **locfdr** a vector of local false discovery rates, the posterior probablity a gene is null in the same order as **genes**
- **FDR** a vector of global false discovery rates in the same order as **genes**
- **genePosteriors** a vector of posterior probabilities that the genes are non-null (equal to 1 - **locfdr**), in the same order as **genes**; when **BIMODAL** is set to TRUE then both **negGenePosteriors** and **posGenePosteriors** are returned that give the posterior probability that a gene is negative and positive, respectively

3 Examples

3.1 A CRISPRi dropout screen

Gilbert et al. 2014 performed genome wide screens for ricin resistance and susceptibility. In table 2, sheet 1 the authors provide the results of the CRISPRi screen at the sgRNA level. This includes the growth phenotype gamma and the ricin phenotype rho. They were interested in the ricin phenotype, but we will look at the growth phenotype to identify genes that lead to decreased growth and can thus be considered essential. In the CRISPhieRmix paper we used log2 fold changes computed by DESeq2 for a fairer comparison of algorithms (since other algorithms such as MAGeCK work directly on the counts), but here we will use the scores computed by Gilbert et al to show the flexibility of the CRISPhieRmix approach.

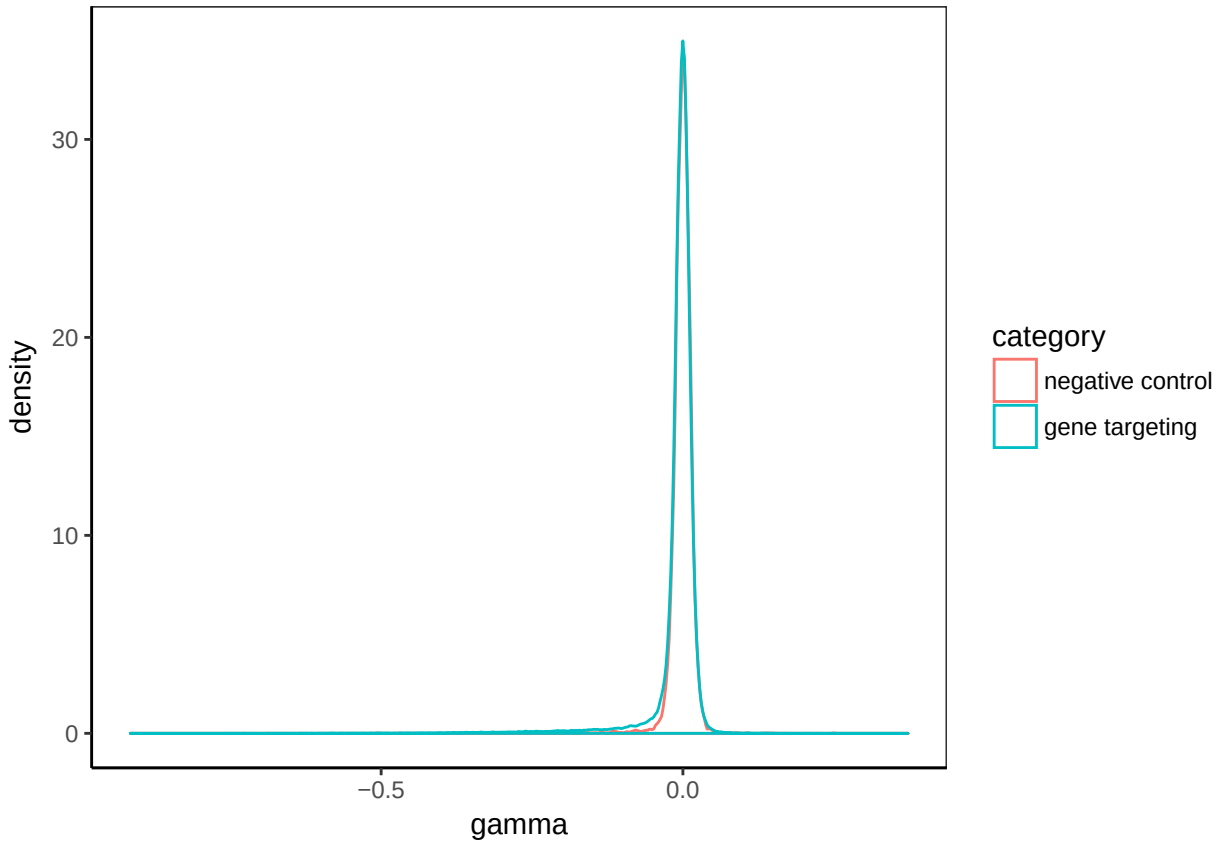
```
Gilbert2014Table2CRISPRi = read.table(file = "Gilbert2014Table2CRISPRi.txt", sep = "\t",
                                      header = TRUE)
head(Gilbert2014Table2CRISPRi)
```

```
##   gene sgRNA.ID Transcripts.targeted   Protospacer.sequence
## 1 A1BG   A1BG-1                  all  GCAAGAGAAAGACCACGAGCA
## 2 A1BG  A1BG-10                  all  GCGGGAACAGGAGCCTTACGG
## 3 A1BG   A1BG-2                  all  GTCTGCAGCAATGAGGCCCCA
## 4 A1BG   A1BG-3                  all   GCAGCCATATGTGAGTGCAG
## 5 A1BG   A1BG-4                  all GACATGATGGTCGCGCTCACTC
## 6 A1BG   A1BG-5                  all  GAATGGTGGGCCCAGGCCGGG
##   Growth.phenotype..gamma. CTx.DTA.phenotype..rho.
## 1             -0.008127700             0.011324965
## 2             -0.006738800             0.011671726
## 3             -0.009591072            -0.028942999
## 4             -0.017039814             0.046149024
## 5             -0.004868127             0.004907662
## 6             -0.012730560            -0.046096108
```

```
# identify the negative control guides
head(sort(table(Gilbert2014Table2CRISPRi$gene), decreasing = TRUE))
```

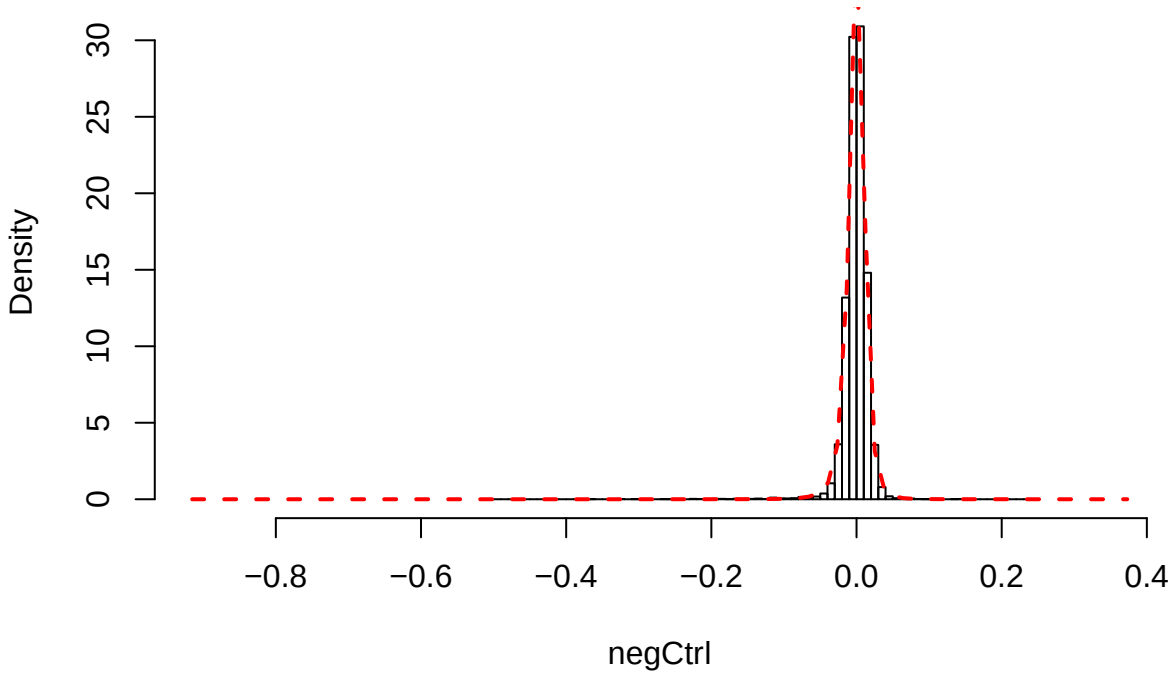
```
##
## negative_control             HLA            NKX2        C1QTNF9B
##            11219             193              85              45
##            STON1            SSPO
##               45              44
```

```
geneTargetingGuides = which(Gilbert2014Table2CRISPRi$gene != "negative_control")
negCtrlGuides = which(Gilbert2014Table2CRISPRi$gene == "negative_control")
gamma = Gilbert2014Table2CRISPRi$Growth.phenotype..gamma.[geneTargetingGuides]
negCtrl = Gilbert2014Table2CRISPRi$Growth.phenotype..gamma.[negCtrlGuides]
geneIds = Gilbert2014Table2CRISPRi$gene[geneTargetingGuides]
# need to remove the negative_control factor
geneIds = factor(geneIds, levels = unique(geneIds))
x = data.frame(gamma = c(gamma, negCtrl),
               category = c(rep("gene targeting", times = length(gamma)),
                            rep("negative control", times = length(negCtrl))))
x$category = factor(x$category, levels = c("negative control", "gene targeting"))
library(ggplot2)
ggplot(x, aes(x = gamma, colour = category)) + geom_density() + theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(), axis.line = element_lin
```

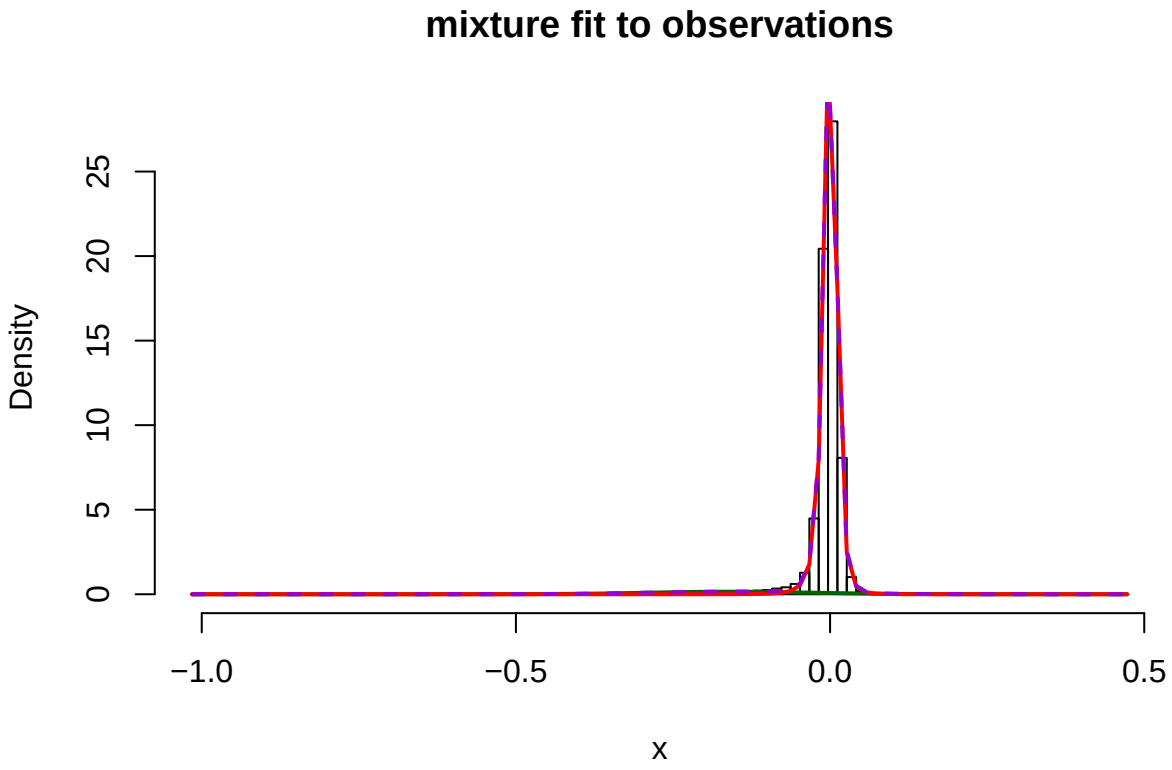


We see from the figure above that most of the gene targeting guides follow the same distribution as the negative control guides, but with a longer tail on the negative end. This is the signal due to cells dying as a result of the gene effects. We can now apply CRISPhieRmix to this data to identify genes that led to decreased growth or cell death.

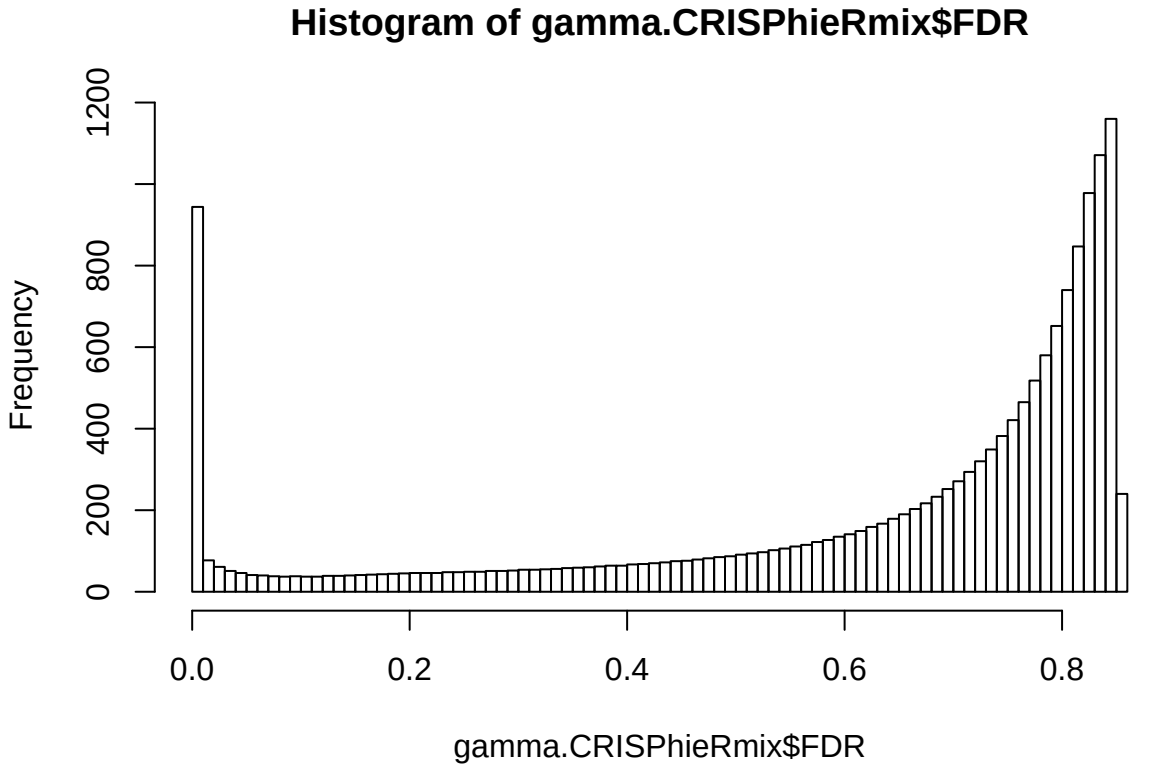
```
library(CRISPhieRmix)
gamma.CRISPhieRmix = CRISPhieRmix(x = gamma, geneIds = geneIds, negCtrl = negCtrl,
                                  mu = -0.2, sigma = 0.1, VERBOSE = TRUE, PLOT = TRUE)
```




```
## fit negative control distributions
## 2 groups
## EM converged
## mu =  -0.1575749
## sigma =  0.1315886
## pq =  0.04938287
```




```
hist(gamma.CRISPhieRmix$FDR, breaks = 100)
```




```
sum(gamma.CRISPhieRmix$FDR < 0.1)
```

```
## [1] 1373
```

The peak near zero indicates the genes that are likely to be essential. Let's look at how much these genes overlap the gold standard list of core constitutive genes and non-essential genes of Hart et al. 2014.

```
ConstitutiveCoreEssentialGenes = scan("ConstitutiveCoreEssentialGenes.txt", what = character())
length(ConstitutiveCoreEssentialGenes)
```

```
## [1] 217
```

```
length(intersect(ConstitutiveCoreEssentialGenes,
                 gamma.CRISPhieRmix$genes[which(gamma.CRISPhieRmix$FDR < 0.1)]))
```

```
## [1] 170
```

```
NonEssentialGenes = scan("NonEssentialGenes.txt", what = character())
length(NonEssentialGenes)
```

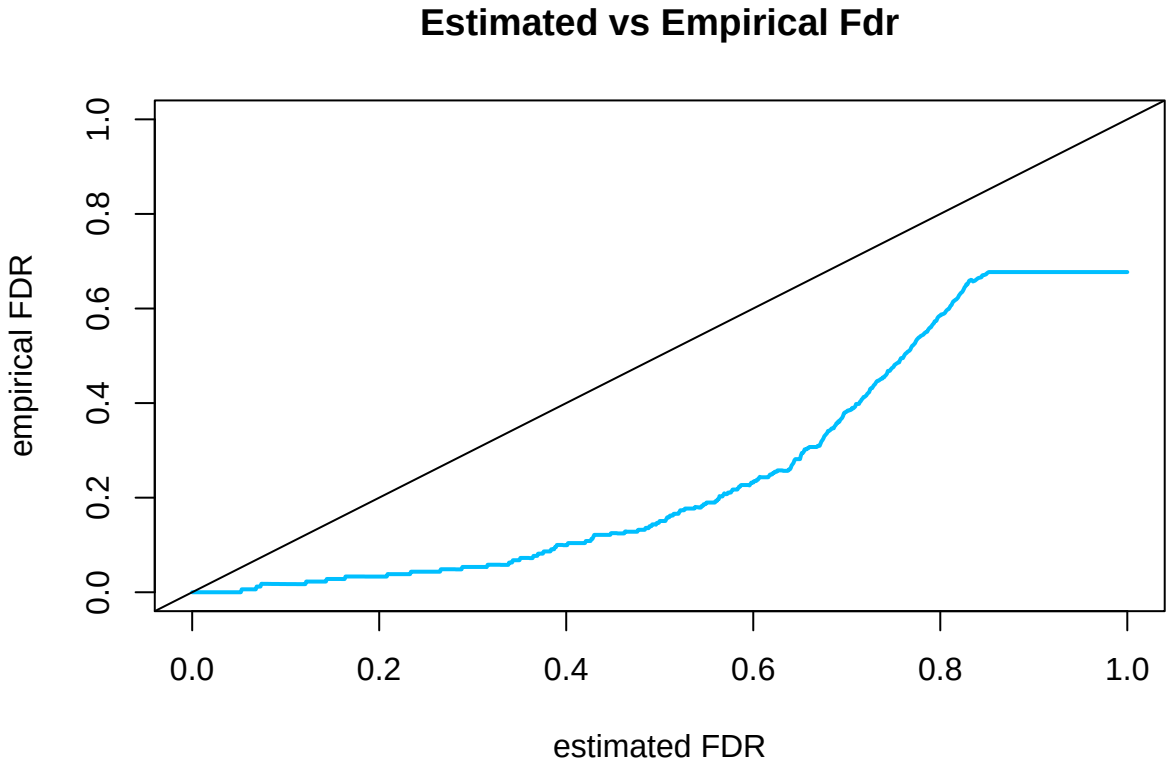
```
## [1] 927
```

```
length(intersect(NonEssentialGenes, gamma.CRISPhieRmix$genes[which(gamma.CRISPhieRmix$FDR < 0.1)]))
```

```
## [1] 3
```

From this we can estimate the empirical false discovery rate at FDR level 0.1 as $3/1373 \approx 0.002$ and an empirical true positive rate as $170/217 \approx 0.78$. The empirical vs estimated FDR curve is plotted below.

```
EssentialGenes = data.frame(gene = factor(c(sapply(ConstitutiveCoreEssentialGenes, toString),
                                            sapply(NonEssentialGenes, toString))),
                            essential = c(rep(1, times = length(ConstitutiveCoreEssentialGenes)),
                                          rep(0, times = length(NonEssentialGenes))))
EssentialGenes = EssentialGenes[which(EssentialGenes$gene %in% gamma.CRISPhieRmix$genes), ]
gamma.CRISPhieRmixEssential = data.frame(genes = gamma.CRISPhieRmix$genes, FDR = gamma.CRISPhieRmix$FDR)
gamma.CRISPhieRmixEssential = gamma.CRISPhieRmixEssential[which(gamma.CRISPhieRmixEssential$genes %in% E
gamma.CRISPhieRmixEssential = gamma.CRISPhieRmixEssential[match(EssentialGenes$gene, gamma.CRISPhieRmixE
fdr.curve <- function(thresh, fdrs, baseline){
  w = which(fdrs < thresh)
  if(length(w) > 0){
    return(sum(1 - baseline[w])/length(w))
  }
  else{
    return(NA)
  }
}
s = seq(from = 0, to = 1, length = 1001)
gamma.CRISPhieRmixFdrCurve = sapply(s, function(t) fdr.curve(t, gamma.CRISPhieRmixEssential$FDR,
                                                             EssentialGenes$essential))
plot(c(0, s[!is.na(gamma.CRISPhieRmixFdrCurve)]), c(0, gamma.CRISPhieRmixFdrCurve[!is.na(gamma.CRISPhieR
     ylab = "empirical FDR", main = "Estimated vs Empirical Fdr", xlim = c(0, 1), ylim = c(0, 1),
     lwd = 2, col = "deepskyblue")
abline(0, 1)
```

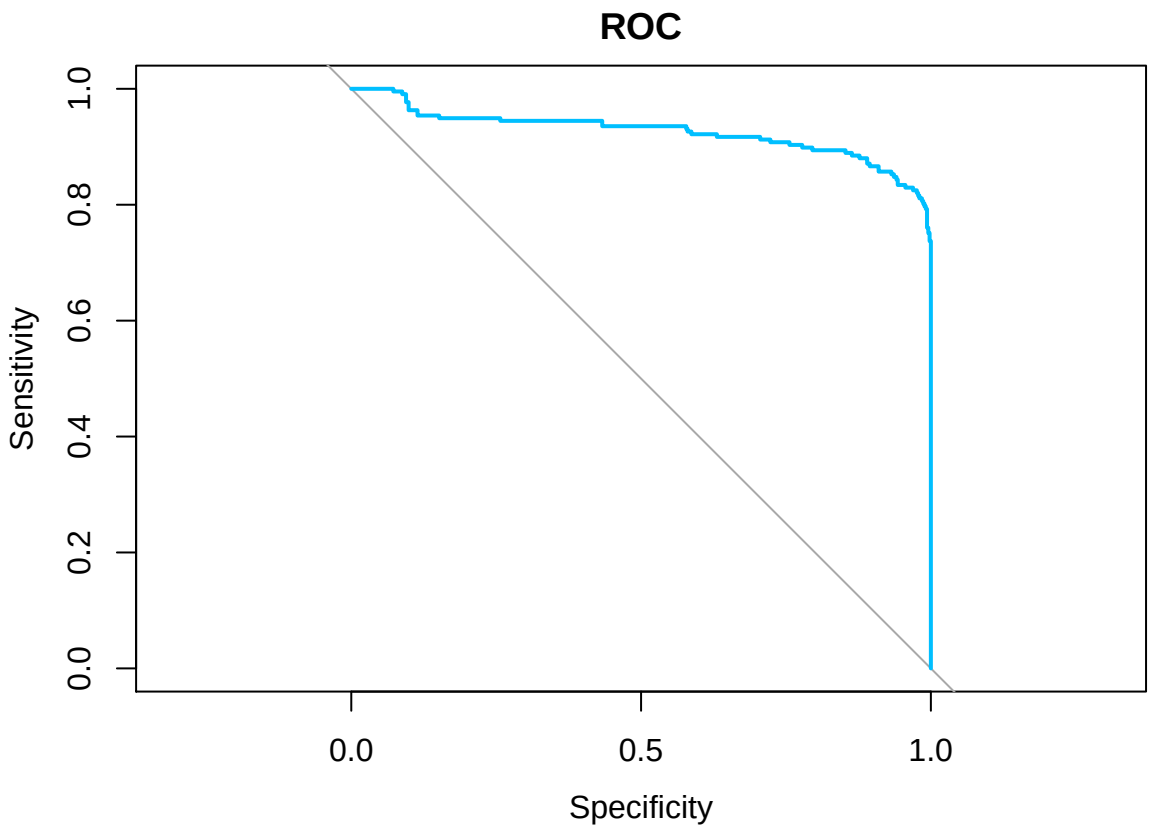


The receiver operator curve based on the above annotated genes is given below.

```
gamma.CRISPhieRmixROC = pROC::roc(EssentialGenes$essential,
                                  gamma.CRISPhieRmixEssential$FDR, auc = TRUE)
gamma.CRISPhieRmixROC
```

```
##
## Call:
## roc.default(response = EssentialGenes$essential, predictor = gamma.CRISPhieRmixEssential$FDR,     au
##
## Data: gamma.CRISPhieRmixEssential$FDR in 455 controls (EssentialGenes$essential 0) > 217 cases (Esse
## Area under the curve: 0.9259
```

```
plot(gamma.CRISPhieRmixROC, col = "deepskyblue", lwd = 2, xlim = c(0, 1), ylim = c(0, 1), main = "ROC")
```


ROC
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
Specificity

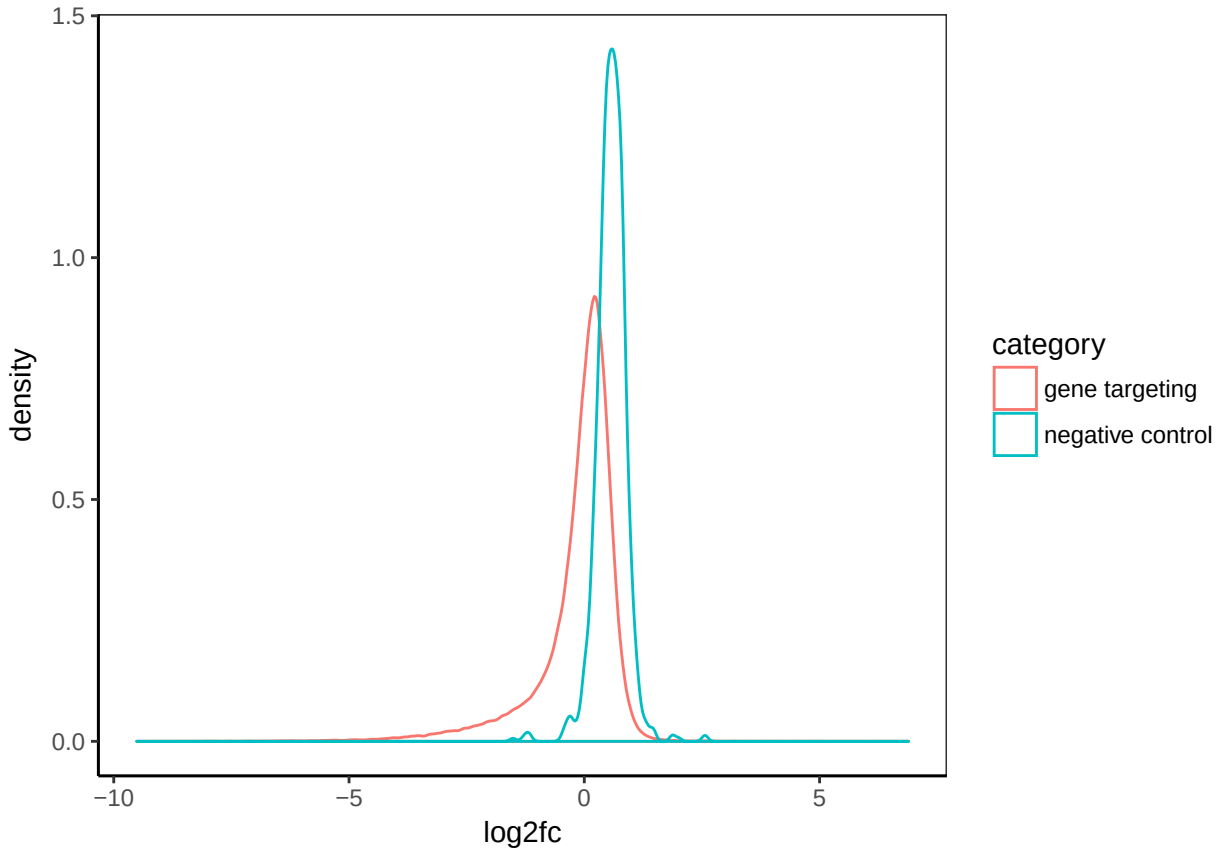
3.2 Checking the negative control distribution

The data shown here is taken from Wang et al, 2015. The authors performed a genome-wide screen for essential genes.

```
Wang2015counts = read.table(file = "aac7041_SM_Table_S2.txt", sep = "\t", header = TRUE)
Wang2015counts = Wang2015counts[-which(rowSums(Wang2015counts[ ,-c(1)]) == 0), ]
which.negCtrl = which(startsWith(sapply(Wang2015counts$sgRNA, toString), "CTRL"))
geneIds = sapply(Wang2015counts$sgRNA[-which.negCtrl],
                 function(g) unlist(strsplit(toString(g), split = "_"))[1])
geneIds = sapply(geneIds, function(g) substring(g, first = 3))
geneIds = factor(geneIds, levels = unique(geneIds))

counts = Wang2015counts[ , -c(1)]
colData = data.frame(cellType = sapply(colnames(counts),
                                       function(x) unlist(strsplit(toString(x), split = ".",
                                                                   fixed = TRUE))[1]),
                     condition = factor(rep(c(0, 1), times = 5)))
rownames(colData) = colnames(counts)
Wang2015DESeq = DESeq2::DESeqDataSetFromMatrix(countData = counts,
                                               colData = colData, design = ~ condition)
Wang2015DESeq = DESeq2::DESeq(Wang2015DESeq)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing
```

```
Wang2015DESeq = DESeq2::results(Wang2015DESeq)
log2fc = Wang2015DESeq$log2FoldChange
log2fc.negCtrl = log2fc[which.negCtrl]
log2fc.geneTargeting = log2fc[-which.negCtrl]
library(ggplot2)
x = data.frame(log2fc = log2fc, category = c(rep("negative control",
                                                 times = length(which.negCtrl)),
                                             rep("gene targeting", times = length(log2fc.geneTargeting))
ggplot(x, aes(x = log2fc, colour = category)) + geom_density() + theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(), axis.line = element_lin
```



As we can see, the distribution of the negative control guides do not line up with the central peak of the gene targeting guides. This indicates that the distribution of the negative control guides does not reflect the null genes. In this case, using the negative control guides to estimate the null will likely lead to very incorrect inferences and is not suggested. It's unclear why the negative control guides do not look like the central peak. If this is the case in your experiment, it is worth investigating why.

3.3 Using a normal hierarchical mixture to rank genes without negative control guides

We'll look at ranking the genes without negative control guides for this data.

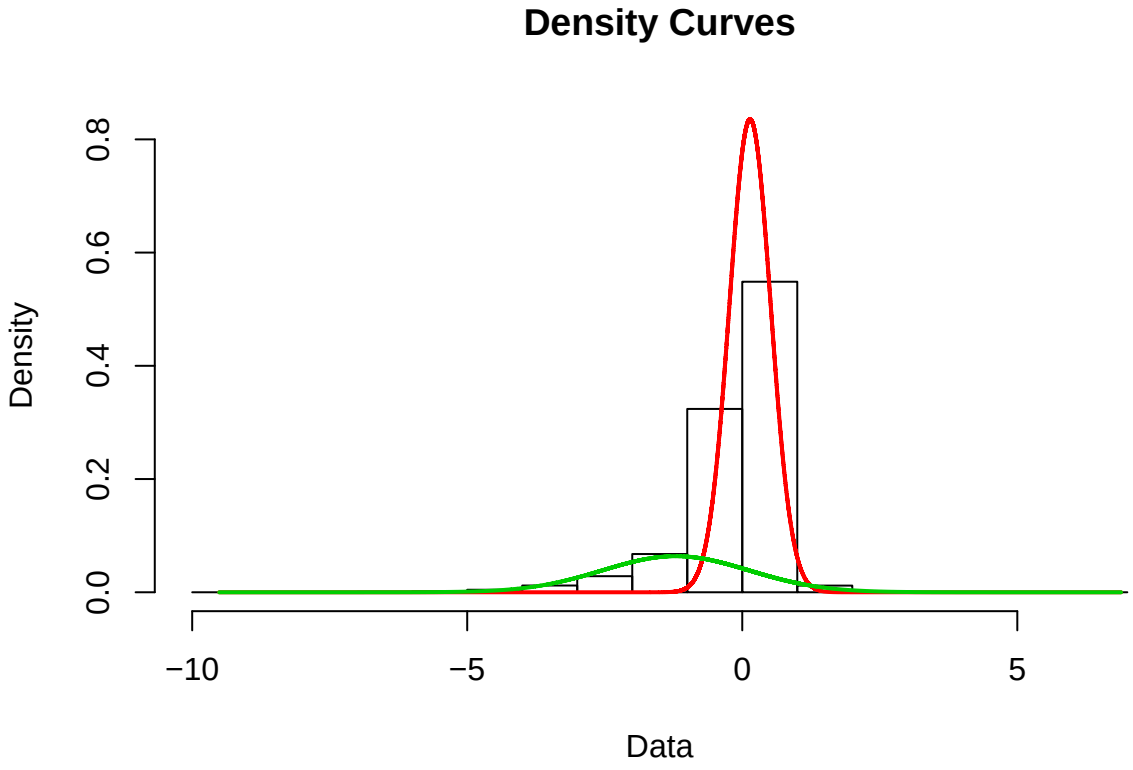
```
Wang2015NormalMix = CRISPhieRmix(x = log2fc.geneTargeting, geneIds = geneIds, PLOT = TRUE, VERBOSE = TRU
```

```
## no negative controls provided, fitting hierarchical normal model
```

```
## Loading required package: mixtools
```

```
## mixtools package, version 1.1.0, Released 2017-03-10
## This package is based upon work supported by the National Science Foundation under Grant No. SES-0518
```

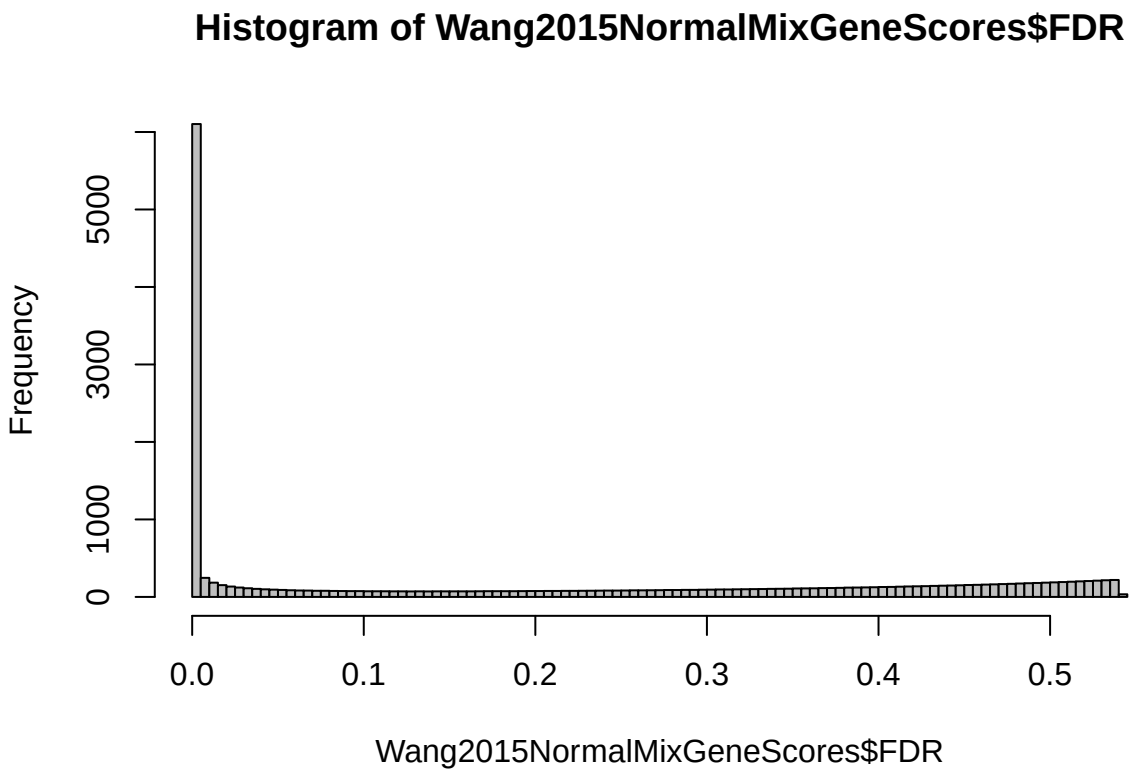
```
## number of iterations= 53
```




```
Wang2015NormalMixGeneScores = data.frame(genes = Wang2015NormalMix$genes, FDR = Wang2015NormalMix$FDR)
head(Wang2015NormalMixGeneScores[order(Wang2015NormalMixGeneScores$FDR, decreasing = FALSE), ], 20)
```

```
##        genes FDR
## 7       AAAS   0
## 18      AAMP   0
## 22      AARS   0
## 23     AARS2   0
## 26  AASDHPPT   0
## 28      AATF   0
## 49     ABCB7   0
## 51     ABCB9   0
## 67     ABCE1   0
## 68     ABCF1   0
## 78    ABHD11   0
## 86   ABHD16B   0
## 90     ABHD2   0
## 100     ABL1   0
## 102   ABLIM1   0
## 108     ABT1   0
## 117    ACAD8   0
## 130    ACBD3   0
## 137      ACD   0
## 138      ACE   0
```

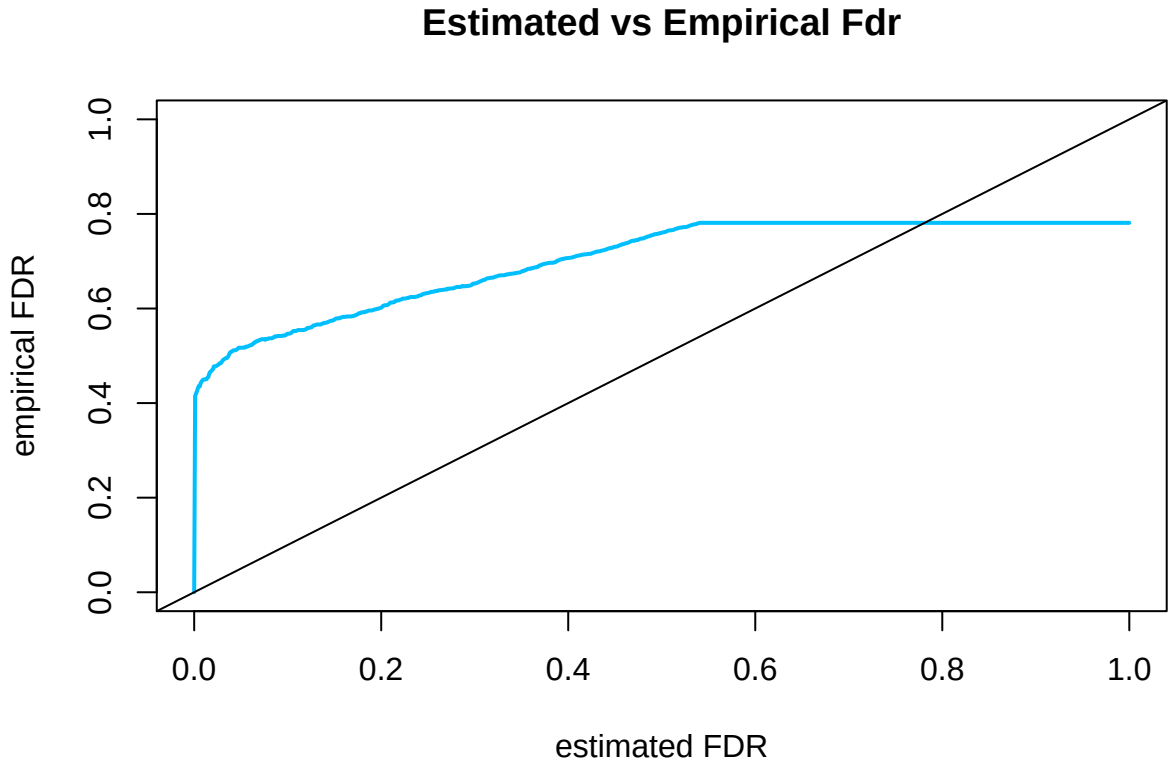
```
hist(Wang2015NormalMixGeneScores$FDR, breaks = 100, col = "grey")
```




```
sum(Wang2015NormalMixGeneScores$FDR == 0)
```

```
## [1] 2777
```

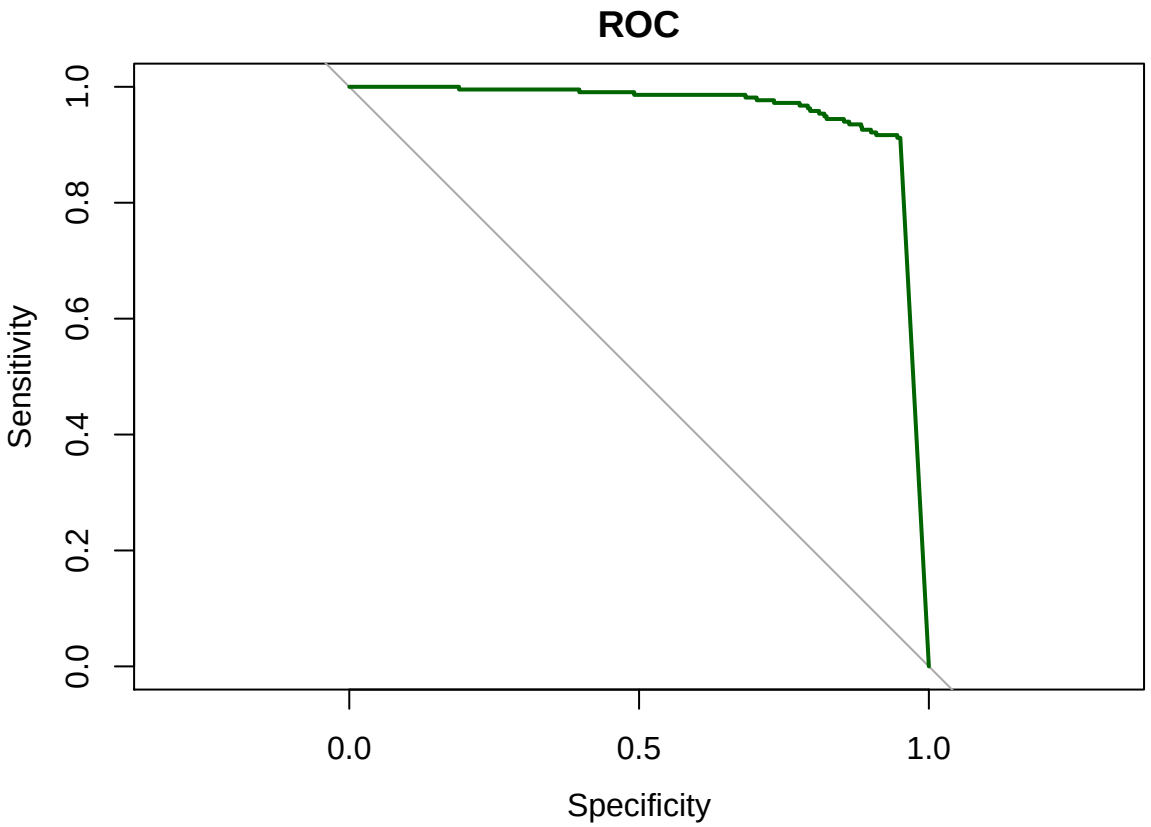
```
EssentialGenes = data.frame(gene = factor(c(sapply(ConstitutiveCoreEssentialGenes, toString),
                                            sapply(NonEssentialGenes, toString))),
                            essential = c(rep(1, times = length(ConstitutiveCoreEssentialGenes)),
                                          rep(0, times = length(NonEssentialGenes))))
EssentialGenes = EssentialGenes[which(EssentialGenes$gene %in% Wang2015NormalMixGeneScores$genes), ]
Wang2015NormalMixGeneScoresEssential = Wang2015NormalMixGeneScores[which(Wang2015NormalMixGeneScores$ge
Wang2015NormalMixGeneScoresEssential = Wang2015NormalMixGeneScoresEssential[match(EssentialGenes$gene,

Wang2015NormalMixGeneScoresEssentialFdrCurve = sapply(s, function(t) fdr.curve(t,
                                                                                Wang2015NormalMixGeneScoresEssentia
                                                                              EssentialGenes$essential))
plot(c(0, s[!is.na(Wang2015NormalMixGeneScoresEssentialFdrCurve)]), c(0, Wang2015NormalMixGeneScoresEss
     ylab = "empirical FDR", main = "Estimated vs Empirical Fdr", xlim = c(0, 1), ylim = c(0, 1),
     lwd = 2, col = "deepskyblue")
abline(0, 1)
```




```
Wang2015NormalMixGeneScoresEssentialROC = pROC::roc(EssentialGenes$essential,
                                  Wang2015NormalMixGeneScoresEssential$FDR, auc = TRUE)
Wang2015NormalMixGeneScoresEssentialROC
```

```
##
## Call:
## roc.default(response = EssentialGenes$essential, predictor = Wang2015NormalMixGeneScoresEssential$FDR
##
## Data: Wang2015NormalMixGeneScoresEssential$FDR in 771 controls (EssentialGenes$essential 0) > 216 cas
## Area under the curve: 0.9555
```

```
plot(Wang2015NormalMixGeneScoresEssentialROC, col = "darkgreen", lwd = 2, xlim = c(0, 1), ylim = c(0, 1
```



As we can see, the normal hierarchical mixture model is calling way too many genes as significant, but does a good job at ranking the essential genes ahead of the non-essential genes. We see that it calls way too many genes are essential at any level of FDR. This will be problematic for most applications, as we can't test all the genes called positive.