Ratio distribution plots and Scatter plots

Xiaowen Shi, Chen Chen

09/18/2018

1. Ratio distribution plots

A random RNA-Seq dataset was generated for the demonstration, composed of 3 biological replicates for the experimental group and the control group. Counts for 300 cis genes and 1700 trans genes are shown.

```
set.seed(2018)
cis_expect_peak = 1.2
trans_expect_peak = 0.99
count_mean = 200
count_sd = 100
num_trans = 1700
num_cis = 300

expect_reads <- round(abs(rnorm(n = num_trans+num_cis,mean = count_mean,sd = count_sd)))
expect_reads_treatment <- round(abs(c(expect_reads[1:num_cis]*rnorm(num_cis,cis_expect_peak,cis_expect_peak/4),expect_reads[(num_cis+1):length(expect_reads)]*rnorm(num_trans,trans_expect_peak,trans_expect_peak/4))))
counts <- NULL
for (i in 1:3) {
  counts <- cbind(counts,rpois(n =  num_trans+num_cis,lambda = expect_reads))
}
for (i in 1:3) {
  counts <- cbind(counts,rpois(n =  num_trans+num_cis,lambda = expect_reads_treatment))
}
colnames(counts) <- c('c1','c2','c3','t1','t2','t3') #3 controls and 3 treatments
cis_genes <- paste0('cis_gene',1:num_cis)
trans_genes <- paste0('trans_gene',1:num_trans)
rownames(counts) <- c(cis_genes,trans_genes)
gene_length <- abs(round(rnorm(n = 2000,mean = 800,sd = 400)))
head(counts)
```

```
##            c1  c2  c3  t1  t2  t3
## cis_gene1 161 175 179 228 225 217
## cis_gene2  43  40  41  48  49  55
## cis_gene3 167 178 176 141 158 138
## cis_gene4 228 228 222 434 379 414
## cis_gene5 383 355 361 266 290 276
## cis_gene6 186 170 152 257 223 242
```

a.Normalization of read counts(with rpkm() in edgeR).

```
library(edgeR)
```

```
## Loading required package: limma
```

```
source('plot_utils.R')
rpkm_data <- rpkm(counts,gene.length=gene_length)
head(rpkm_data)
```

```
##                   c1         c2         c3         t1         t2
## cis_gene1 2512.80273 2715.53795 2788.47502 3465.33941 3422.67421
## cis_gene2   75.34904   69.68745   71.70911   81.90848   83.68659
## cis_gene3  818.00791  866.85477  860.46836  672.57165  754.30787
## cis_gene4  616.70508  613.14450  599.34497 1143.17001  999.15396
## cis_gene5  874.11757  805.53560  822.35528  591.19551  645.08894
## cis_gene6  526.33030  478.27718  429.30923  708.20170  615.03640
##                   t3
## cis_gene1 3291.64378
## cis_gene2   93.66827
## cis_gene3  656.96267
## cis_gene4 1088.33751
## cis_gene5  612.21045
## cis_gene6  665.55105
```

b.Remove lowly expressed genes.

```
rpkm_data_filtered <- rpkm_data[apply(rpkm_data, 1, function(x)sum(x>0))>3,]
head(rpkm_data_filtered)
```

```
##                   c1         c2         c3         t1         t2
## cis_gene1 2512.80273 2715.53795 2788.47502 3465.33941 3422.67421
## cis_gene2   75.34904   69.68745   71.70911   81.90848   83.68659
## cis_gene3  818.00791  866.85477  860.46836  672.57165  754.30787
## cis_gene4  616.70508  613.14450  599.34497 1143.17001  999.15396
## cis_gene5  874.11757  805.53560  822.35528  591.19551  645.08894
## cis_gene6  526.33030  478.27718  429.30923  708.20170  615.03640
##                   t3
## cis_gene1 3291.64378
## cis_gene2   93.66827
## cis_gene3  656.96267
## cis_gene4 1088.33751
## cis_gene5  612.21045
## cis_gene6  665.55105
```

c.Compute the mean of normalized counts.

```
mean_control <- apply(rpkm_data_filtered[,1:3], 1, mean)
mean_treatment <- apply(rpkm_data_filtered[,4:6], 1, mean)
mean_control[mean_control==0] <- 10e-6
```

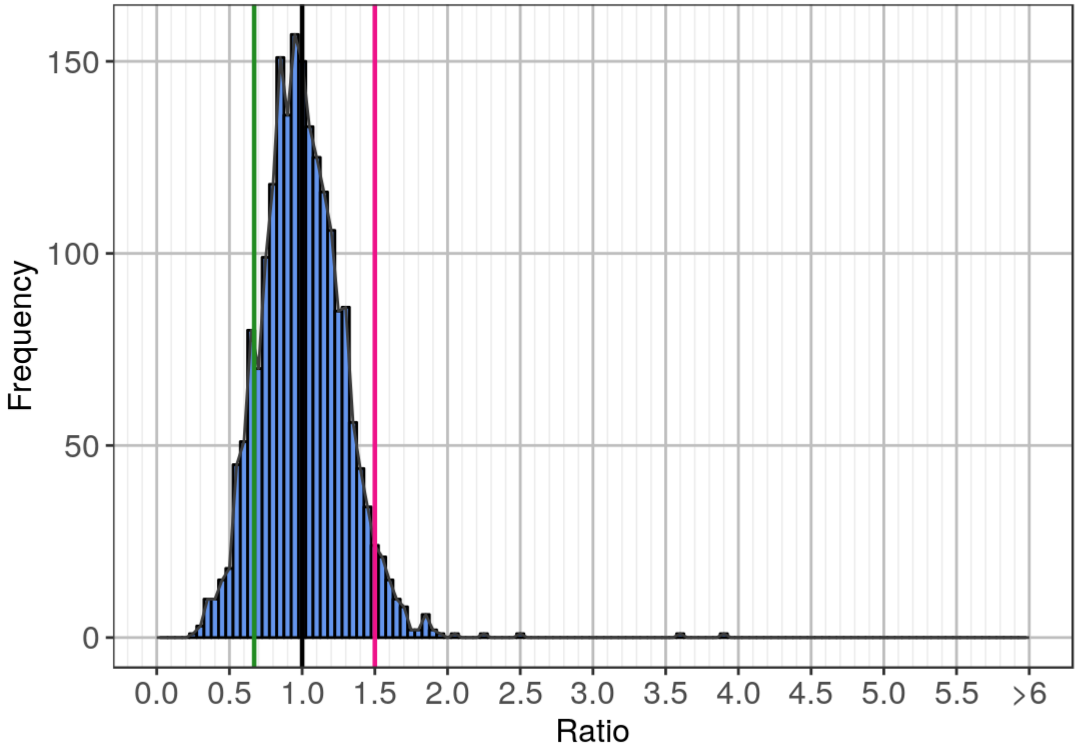
Calculate the ratio of each gene.

```
r <- mean_treatment/mean_control
```

d.Generate a histogram using ggplot2 package.

```
distribution_plot <- plot_distribution(r,title_name = '',left_line = 0.67,right_line =
1.5)
plot(distribution_plot)
```

```
## Warning: Removed 2 rows containing missing values (geom_path).
```

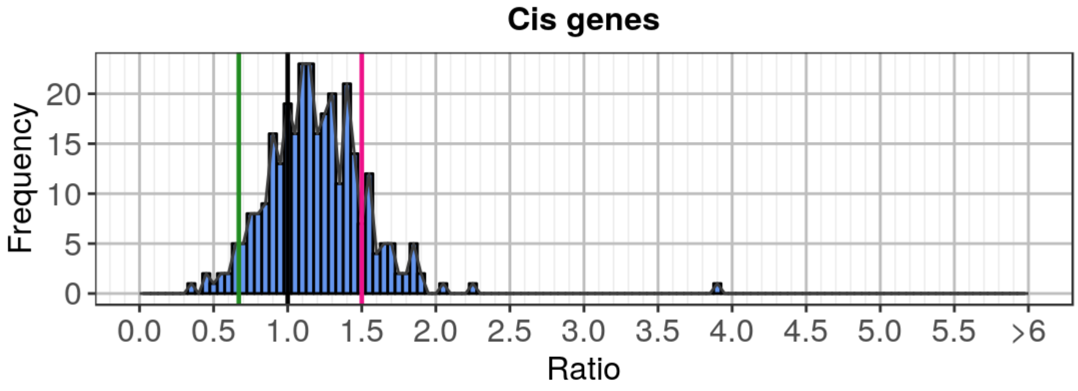


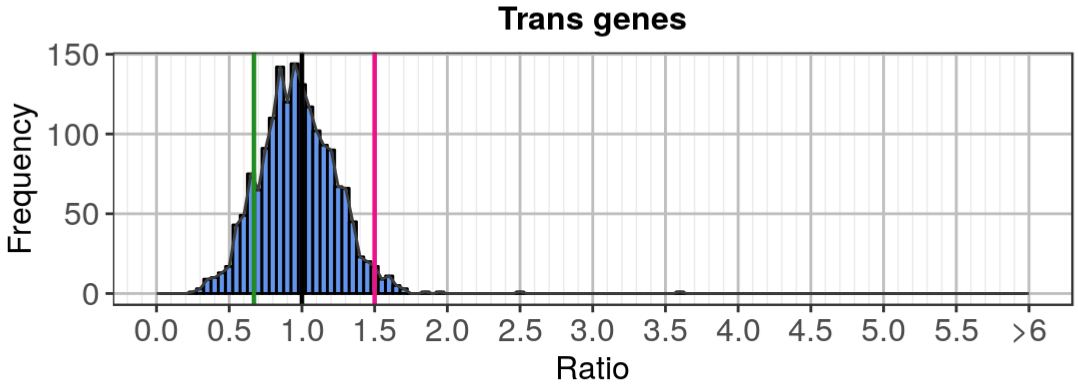
e.Plot cis and trans genes seperately.

```
distribution_plot2 <- plot_distribution_pairs(r,cis_genes,title_name = '',left_line = 0.
67,right_line = 1.5,max_ratio = 6)
```

```
## Warning: Removed 2 rows containing missing values (geom_path).

## Warning: Removed 2 rows containing missing values (geom_path).
```





2.Scatter plots

a.Perform differential gene expression analysis using edgeR.

```
group_table <- data.frame('group'=rep('experiment1',6),
                          'condition'=c(rep('control',3),rep('treatment',3)),
                          'control'=c(rep(T,3),rep(F,3)))
rownames(group_table) <- colnames(counts)
de <- edgeR_wrapper(cnt = counts,grp_table = group_table)
de1 <- de$experiment1.treatment
head(de1)
```

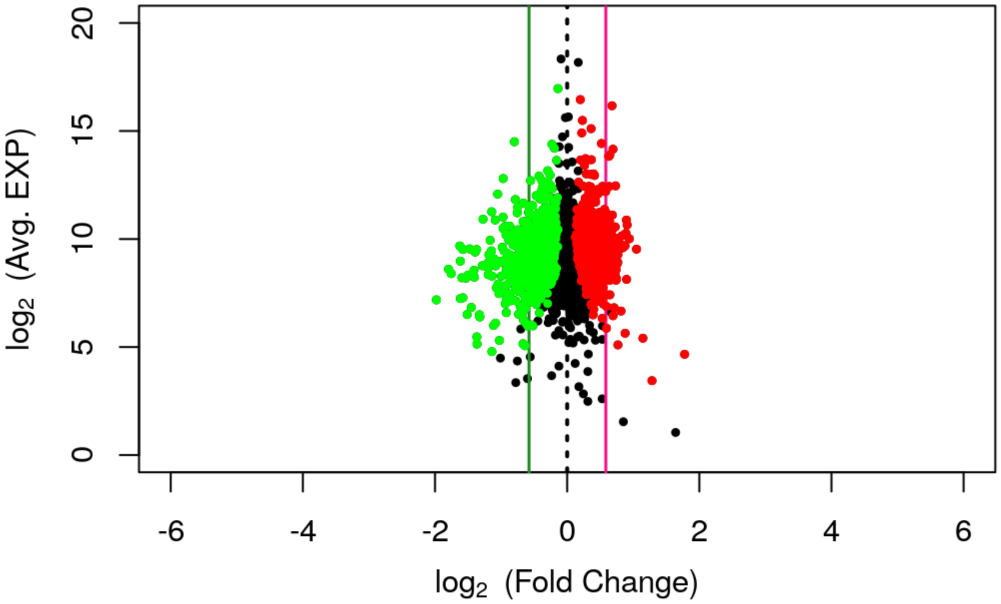
```
##                  logFC   logCPM         LR       PValue          FDR
## cis_gene1  0.3442511 8.948089  16.725063 4.320643e-05 1.315264e-04
## cis_gene2  0.2579085 6.892871   2.204484 1.376099e-01 1.939533e-01
## cis_gene3 -0.2885402 8.644853   9.556758 1.992146e-03 4.441798e-03
## cis_gene4  0.8203142 9.627311 133.726057 6.271893e-31 3.919933e-29
## cis_gene5 -0.4364816 9.647190  43.779215 3.675881e-11 2.712828e-10
## cis_gene6  0.4717456 9.001301  28.459975 9.565564e-08 4.428502e-07
```

b.Compute mean of normalized counts.

```
avg_expre <- apply(rpkm_data, 1, mean)
```

c.Generate a scatter plot.

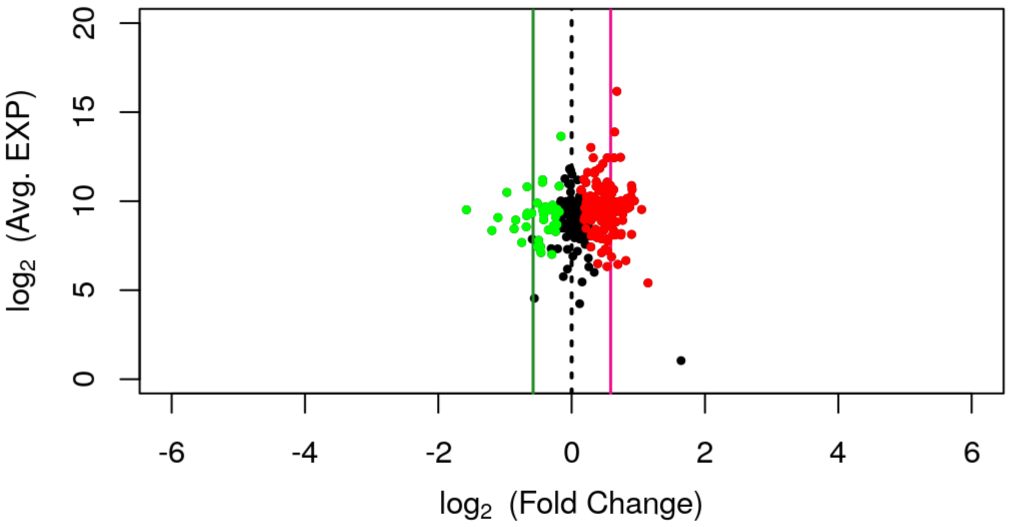
```
plot_scatter(Fold_Change = 2^de1$logFC,Expre = avg_expre,P_Value = de1$FDR,left_line =
0.67,right_line = 1.5)
```



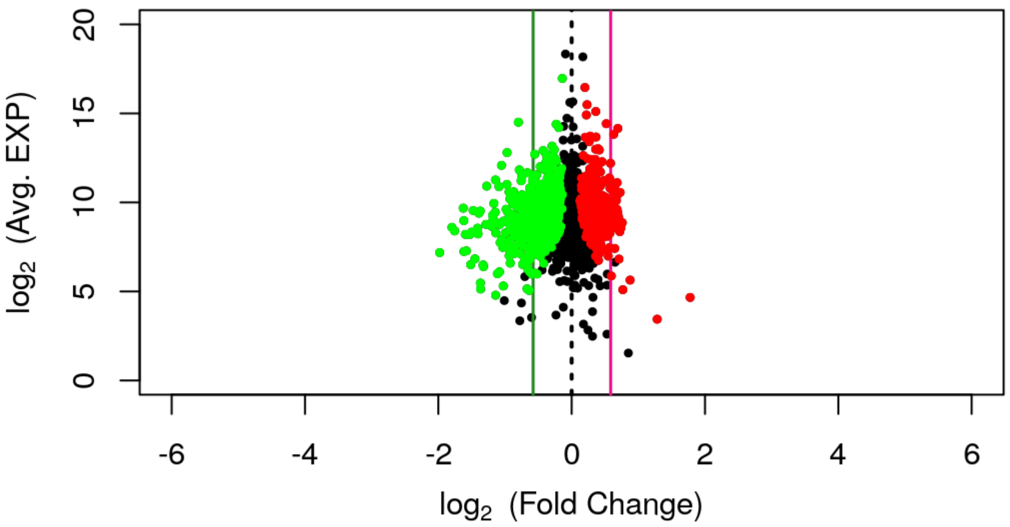
d.Plot cis and trans genes seperately.

```
fc <- 2^de1$logFC
pval <- de1$FDR
names(fc) <- names(pval) <- names(avg_expre)
plot_scatter_pairs(Fold_Change = fc,Expre = avg_expre,P_Value = pval,cis_genes,left_line
 = 0.67,right_line = 1.5)
```

Cis genes



Trans genes


log$_2$ (Avg. EXP)
log$_2$ (Fold Change)