Revision of Graphs with Relative Standard Deviation

James R. Hunter

08 December 2016

Load Modules

```
suppressMessages(library(tidyverse))
suppressPackageStartupMessages(library(purrr))
suppressPackageStartupMessages(library(knitr))
suppressPackageStartupMessages(library(RColorBrewer))
suppressPackageStartupMessages(library(DescTools))
options(scipen = 1000)
```

Load Data

```
load("mzdata.RData")
```

Fix RTC and PPC genes (give them numbers)

```
newnames <- c("RTC1", "RTC2", "RTC3", "PPC1", "PPC2", "PPC3")
mzdata$Symbol[91:96] <- newnames
```

Calculating Differences between EC and RP and IR and NR for both time periods

```
# Calculate difference using both recent and 12 month data
# Use absolute value of differences since direction doesn't count
grpdiffs <- mzdata %>%
            mutate(diffecrpri = abs(RPRI - ECRI)) %>%
            mutate(diffecrp12 = abs(RP12 - EC12)) %>%
            mutate(diffirnrri = abs(NRBT - IRBT)) %>%
            mutate(diffirnr12 = abs(NR12 - IR12)) %>%
            select(Symbol, diffecrpri, diffecrp12, diffirnrri, diffirnr12)
```

Make long version of data frame

```
mz_long <- mzdata %>%
             select(-1) %>% # use all cols except Position
             gather(key = gr, value = expr, 2:9) %>%
             mutate(term = substr(gr, 3, 4)) %>%
             mutate(gr = substr(gr, 1, 2))
```

Test if any of the group differences as a whole are significant

```
# Set up data frames
ecrpri <- mz_long %>%
          filter(gr %in% c("EC", "RP") & term == "RI")
ecrp12 <- mz_long %>%
          filter(gr %in% c("EC", "RP") & term == "12")
irnrbt <- mz_long %>%
          filter(gr %in% c("IR", "NR") & term == "BT")
irnr12 <- mz_long %>%
          filter(gr %in% c("IR", "NR") & term == "12")

# run t-tests
(t_ecrpri <- t.test(expr ~ gr, data = ecrpri))
```

```
##
##  Welch Two Sample t-test
##
## data:  expr by gr
## t = 0.66907, df = 175.03, p-value = 0.5043
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1706739  0.3457426
## sample estimates:
## mean in group EC mean in group RP
##         1.486662         1.399128
```

```
(t_ecrp12 <- t.test(expr ~ gr, data = ecrp12))
```

```
##
##  Welch Two Sample t-test
##
## data:  expr by gr
## t = -1.6371, df = 155.27, p-value = 0.1036
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.52708207  0.04936123
## sample estimates:
## mean in group EC mean in group RP
##         1.394566         1.633426
```

```
(t_irnrbt <- t.test(expr ~ gr, data = irnrbt))
```

```
##
##  Welch Two Sample t-test
##
## data:  expr by gr
## t = -0.13291, df = 185.01, p-value = 0.8944
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.159483  1.013114
## sample estimates:
## mean in group IR mean in group NR
##         1.973397         2.046581
```

```
(t_irnr12 <- t.test(expr ~ gr, data = irnr12))
```

```
##
##  Welch Two Sample t-test
##
## data:  expr by gr
## t = -1.5086, df = 124.47, p-value = 0.1339
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.307773  0.176477
## sample estimates:
## mean in group IR mean in group NR
##         1.664943         2.230591
```

None of the group by group tests are significant.

Determine Genes that Are Significantly Different between EC and RP and IR and NR

DeltaDeltaCT (2^-DDCT)

```
# read file from Excel
ecrpriddct <- read_csv("ecrpriddct.csv")
```

```
## Parsed with column specification:
## cols(
##   Symbol = col_character(),
##   elite_1010V1 = col_double(),
##   elite_1034V2 = col_double(),
##   elite_1073V1 = col_double(),
##   elite_1122V1 = col_double(),
##   elite_1168V1 = col_double(),
##   rapid_1120V1 = col_double(),
##   rapid_1141V1 = col_double(),
##   rapid_2041V1 = col_double(),
##   rapid_2044V1 = col_double(),
##   rapid_2046V1 = col_double()
## )
```

```
ecrp12ddct <- read_csv("ecrp12ddct.csv")
```

```
## Parsed with column specification:
## cols(
##   Symbol = col_character(),
##   elite_1010V1 = col_double(),
##   elite_1034V2 = col_double(),
##   elite_1073V1 = col_double(),
##   elite_1122V1 = col_double(),
##   elite_1168V1 = col_double(),
##   rapid_1120V5 = col_double(),
##   rapid_1141V5 = col_double(),
##   rapid_2041V5 = col_double(),
##   rapid_2044V5 = col_double(),
##   rapid_2046V5 = col_double()
```

```
## )
```

```
irnrbtddct <- read_csv("irnrbtddct.csv")
```

```
## Parsed with column specification:
## cols(
##   Symbol = col_character(),
##   RI_1046V7 = col_double(),
##   RI_1087V7 = col_double(),
##   RI_1117V6 = col_double(),
##   RI_2028V9 = col_double(),
##   RI_2033V6 = col_double(),
##   NRI_1050V1 = col_double(),
##   NRI_1097V1 = col_double(),
##   NRI_2016V1 = col_double(),
##   NRI_2020V5 = col_double(),
##   NRI_2042V4 = col_double()
## )
```

```
irnr12ddct <- read_csv("irnr12ddct.csv")
```

```
## Parsed with column specification:
## cols(
##   Symbol = col_character(),
##   RI_1046V7 = col_double(),
##   RI_1087V7 = col_double(),
##   RI_1117V6 = col_double(),
##   RI_2028V9 = col_double(),
##   RI_2033V6 = col_double(),
##   NRI_1050V1 = col_double(),
##   NRI_1097V1 = col_double(),
##   NRI_2016V1 = col_double(),
##   NRI_2020V5 = col_double(),
##   NRI_2042V4 = col_double()
## )
```

Function to fix revised Symbols per MZ

```
fixSymbol <- function(filename) {
  filename$Symbol[filename$Symbol == "ELANE"] <- "ELA2"
  filename$Symbol[filename$Symbol == "SERPINC1"] <- "ATIII"
  filename$Symbol[filename$Symbol == "TNFSF10"] <- "TRAIL"
  filename$Symbol[filename$Symbol == "TNFRSF1B"] <- "TNFR2"
  # now the GDC gene needs to be renamed HGDC to be consistent with mzdata
  filename$Symbol[filename$Symbol == "GDC"] <- "HGDC"
  return(filename)
  }
```

Execute fix

```
ecrpriddct <- fixSymbol(ecrpriddct)
ecrp12ddct <- fixSymbol(ecrp12ddct)
```

```
irnrbtddct <- fixSymbol(irnrbtddct)
irnr12ddct <- fixSymbol(irnr12ddct)
```

Measure significance via a t parametric test of differences

```
genes <- ecrpriddct$Symbol
```

Function for processing files to test significance

```
processGroupFiles <- function(input) {
  patients <- colnames(input[2:11])

# set up data frame of t-tests
  ttdf <- data_frame(Symbol = genes, tstat = 0.0, pval = 0.0, lowerci = 0.0,
                     upperci = 0.0, sd1 = 0.0, sd2 = 0.0)

# transform matrix to put genes as variables and patients as observations
  x <- as_data_frame(t(as.matrix(input[,2:11])))
  colnames(x) <- genes

# set up to loop t-tests over genes
  for (i in 1:length(genes)) {
  # set up data frame for this gene from transformed matrix
  # columns 1:5 are group 1 and cols 6:10 are group 2
  # get group 1 data
    g1 <- x %>% slice(1:5) %>% select(i)
    g1 <- unlist(g1)

  # get group 2 data
    g2 <- x %>% slice(6:10) %>% select(i)
    g2 <- unlist(g2)

  # Calculate standard deviations for both groups and place in data frame
    ttdf$sd1[i] <- sd(g1, na.rm = TRUE)
    ttdf$sd2[i] <- sd(g2, na.rm = TRUE)

  # run t.test and load values in data frame
  xt <- t.test(g1, g2)
  ttdf$tstat[i] <- xt$statistic
  ttdf$pval[i] <- xt$p.value
  ttdf$lowerci[i] <- xt$conf.int[1]
  ttdf$upperci[i] <- xt$conf.int[2]
  }
rm(x)
return(ttdf)
}
```

Execute Files for Each of 4 Graphs

```
ecrprisig <- processGroupFiles(ecrpriddct) %>% filter(pval < 0.05) %>% arrange(pval)
ecrp12sig <- processGroupFiles(ecrp12ddct) %>% filter(pval < 0.05) %>% arrange(pval)
```

```
irnrbtsig <- processGroupFiles(irnrbtddct) %>% filter(pval < 0.05) %>% arrange(pval)
irnr12sig <- processGroupFiles(irnr12ddct) %>% filter(pval < 0.1) %>% arrange(pval)
## NB: irnr12 had no sig genes below p < 0.05; alpha changed to 0.1
```

Identify genes that are significant for each graph

```
ecrpriGenesig <- ecrprisig$Symbol
ecrp12Genesig <- ecrp12sig$Symbol
irnrbtGenesig <- irnrbtsig$Symbol
irnr12Genesig <- irnr12sig$Symbol
```

Function to Set up Graph joining expression data and significance test data

```
graphSetUp <- function(genessig, gr1, gr2, t, sigfile) {

  tempLong <- mz_long %>%
              filter(Symbol %in% genessig & gr %in% c(gr1, gr2) & term == t)
  temp <- sigfile %>%
      left_join(tempLong, by = "Symbol") %>%
      mutate(stddev = ifelse(gr == gr1, sd1, sd2)) %>%
      mutate(lower = expr - stddev) %>%
      mutate(upper = expr + stddev) %>%
      mutate(cv = 100*stddev/expr) %>% # coefficient of variation in pct form
      select(Symbol, gr, expr, term, pval, stddev, cv, lower, upper)

  for (i in seq(1, nrow(temp), by = 2)) {
    temp$lower[i : (i + 1)] <- mean(temp$lower[c(i, i + 1)])
    temp$upper[i : (i + 1)] <- mean(temp$upper[c(i, i + 1)])
  }
 return(temp)
}
```

Execute graphSetUp to prepare the 4 graphs

```
ecrprigrdb <- graphSetUp(ecrpriGenesig, "EC", "RP", "RI", sigfile = ecrprisig)
ecrp12grdb <- graphSetUp(ecrp12Genesig, "EC", "RP", "12", sigfile = ecrp12sig)
irnrbtgrdb <- graphSetUp(irnrbtGenesig, "IR", "NR", "BT", sigfile = irnrbtsig)
irnr12grdb <- graphSetUp(irnr12Genesig, "IR", "NR", "12", sigfile = irnr12sig)
```

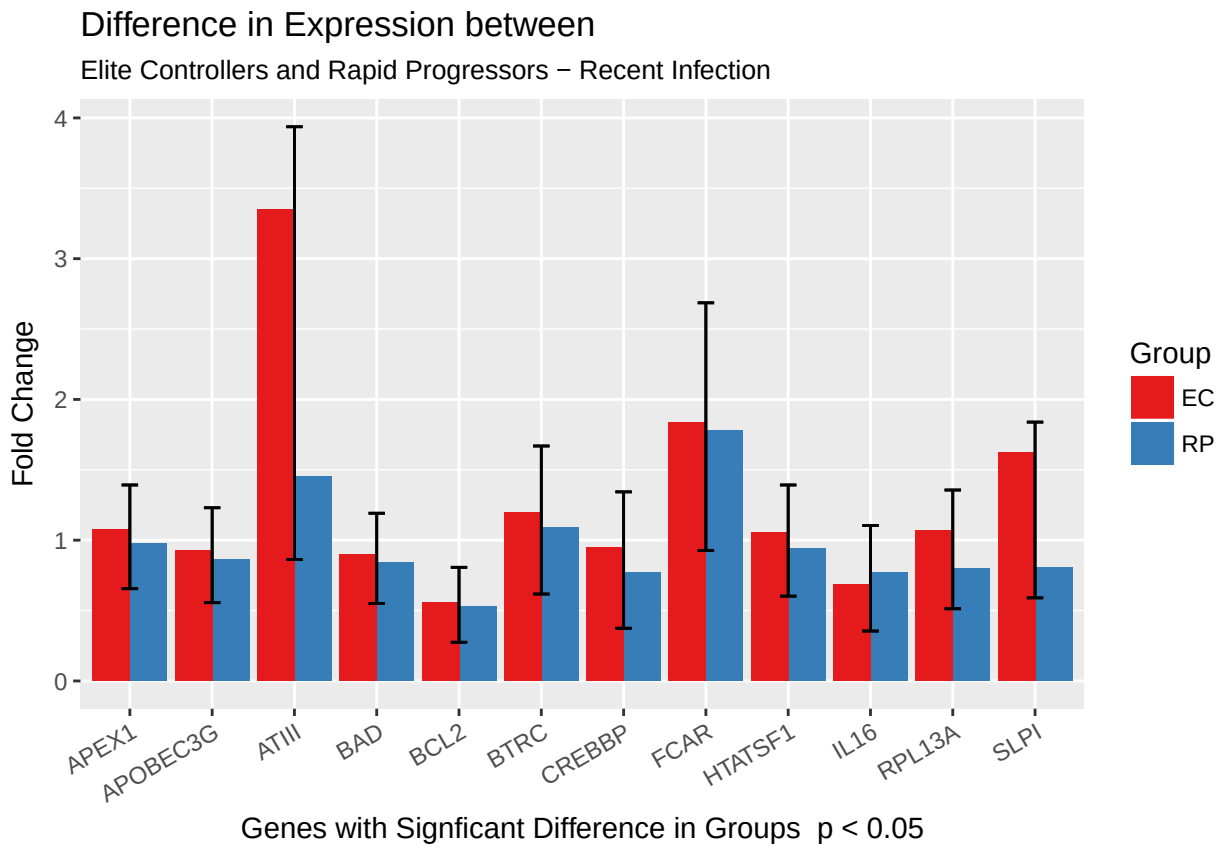
Setup graph function for paper

```
makeBarGraph <- function(infile, alpha, grouptext, term) {
  xtext <- paste0("Genes with Signficant Difference in Groups  p < ", alpha)
  sttext <- paste0(grouptext, " - ", term)
  x <- ggplot(data = infile, aes(x = Symbol, y = expr, fill = gr))
  x <- x + scale_fill_brewer(palette = "Set1")
  x <- x + geom_bar(stat = "identity", position = "dodge")
```

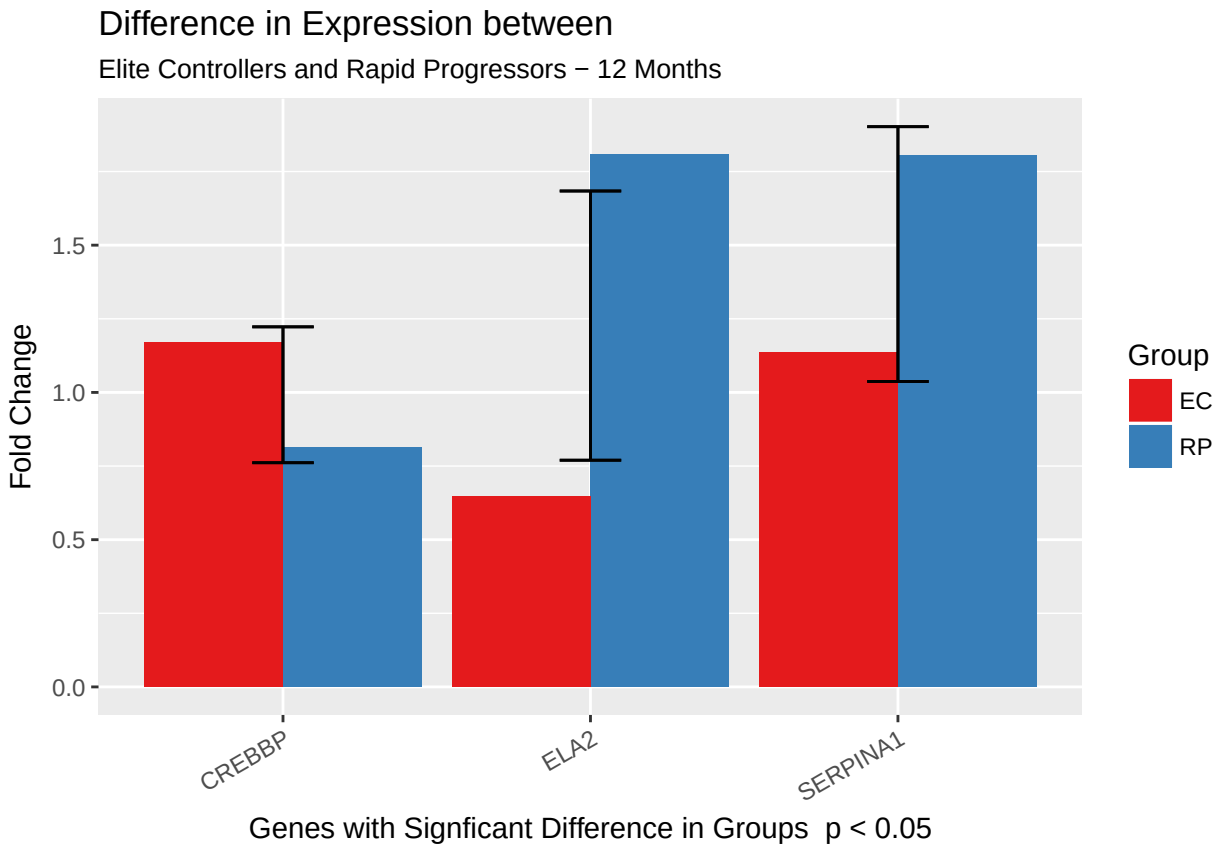
```
  x <- x + geom_errorbar(aes(x = Symbol, ymin = lower, ymax = upper), width = 0.2)
  x <- x + xlab(xtext)
  x <- x + ylab("Fold Change")
  x <- x + labs(title = "Difference in Expression between", subtitle = sttext)
  x <- x + labs(fill = "Group")
  x <- x + theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1))
  return(x)
}
```

Make the graphs

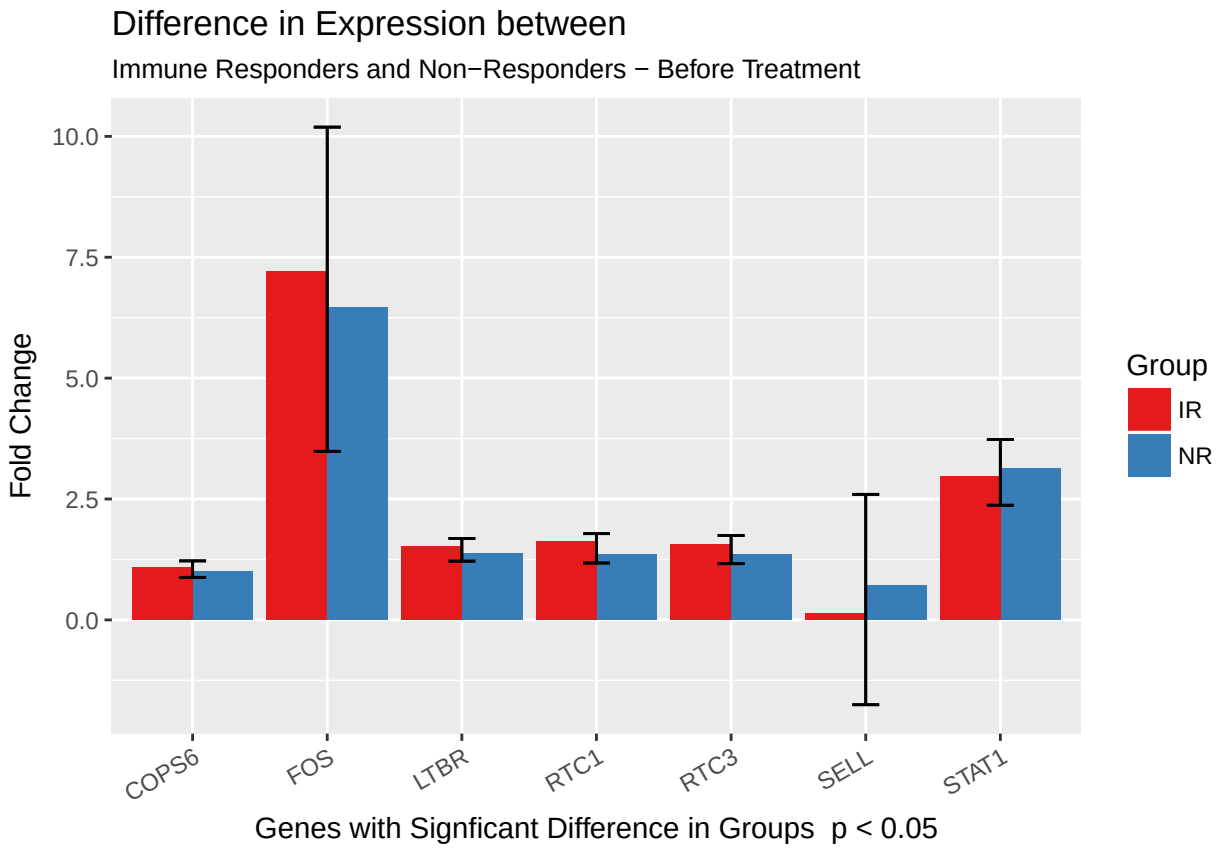
```
graf1 <- makeBarGraph(ecrprigrdb, "0.05", "Elite Controllers and Rapid Progressors",
                      "Recent Infection")
graf1
```



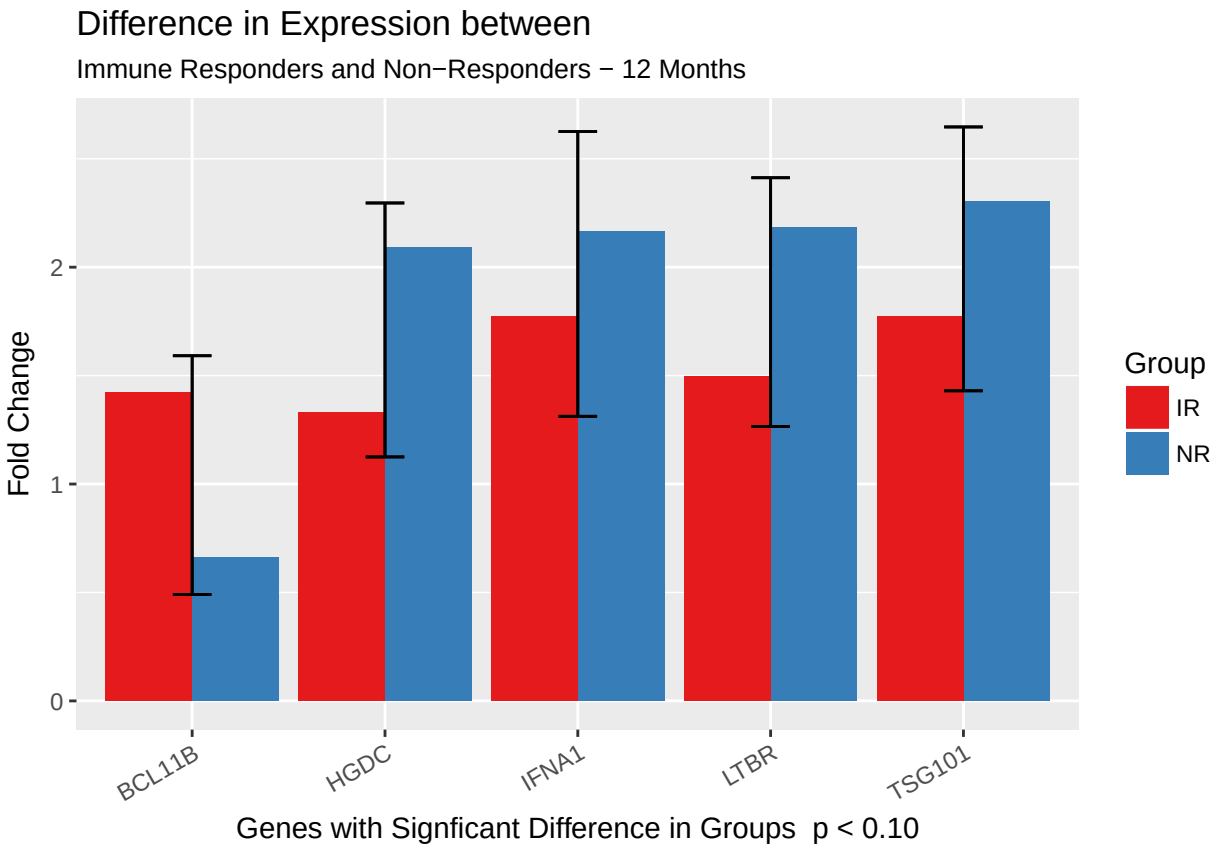

```
graf2 <- makeBarGraph(ecrp12grdb, "0.05", "Elite Controllers and Rapid Progressors",
                      "12 Months")
graf2
```




```
graf3 <- makeBarGraph(irnrbtgrdb, "0.05", "Immune Responders and Non-Responders",
                      "Before Treatment")
graf3
```




```
graf4 <- makeBarGraph(irnr12grdb, "0.10", "Immune Responders and Non-Responders",
                      "12 Months")
graf4
```


Difference in Expression between
Immune Responders and Non−Responders − 12 Months
Fold Change
2
1
0
BCL11B
HGDC
IFNA1
LTBR
TSG101
Genes with Signficant Difference in Groups p < 0.10
Group
IR
NR