TCGA colon/adenocarcinoma data analysis

Uwe Schwartz

20 8 2017

Dependencies

```
library(multtest)
```

Load data

This markdown script contains code to compare weather identifed Evi/Wls non-canonical regulated genes are correlated with mRNA expression of Evi/Wls in colon cancer (TCGA date set, 2013). Level 3 microarray expression data was downloaded from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/).

```
# function to load TCGA data into R
loadTCGA<-function(path){
  norm.mx<-list.files(path)
  for(i in 1:length(norm.mx)){
    part<-read.delim(paste0(path,norm.mx[i]))
    if(i==1){
      exp.ma<- as.matrix(as.numeric(as.character(part$value)))
      rownames(exp.ma)<-as.character(part$gene.symbol)
      colnames(exp.ma)<- as.character(part$barcode[1])
    } else {
    exp.ma<- cbind(exp.ma, as.numeric(as.character(part$value)))
    colnames(exp.ma)[i]<-  as.character(part$barcode[1])
    }
  }
  return(exp.ma)
}

# load matched normal tissue data (n=12)
path.1<-paste0("data/TCGA/8957cb58-4ae8-44e4-b52f-0405cb175a85/",
               "Expression-Genes/UNC__AgilentG4502A_07_3/Level_3/")
ma.1<-loadTCGA(path.1)

# load unmatched normal tissue data (n=7)
path.2 <-paste0("data/TCGA//a304baf7-96d7-445c-b9bf-b475bcf3fa4e/",
                "Expression-Genes/UNC__AgilentG4502A_07_3/Level_3/")
ma.2<-loadTCGA(path.2)

# load tumor tissue data (n=155)
path.3 <-paste0("data/TCGA/21c03af6-e84d-469d-9798-e2563e0cce5a/",
                "Expression-Genes/UNC__AgilentG4502A_07_3/Level_3/")
ma.3<-loadTCGA(path.3)

# merge data sets,
exp.ma<-cbind(ma.1,ma.2,ma.3)
```

Differential expression analysis

Next we selected samples with high differences in EVI expression, which is involved in Wnt secretion. In this analysis we assume that in Samples with low EVI expression, complete Wnt secretion and hence signaling is impaired compared to samples with high EVI expression.

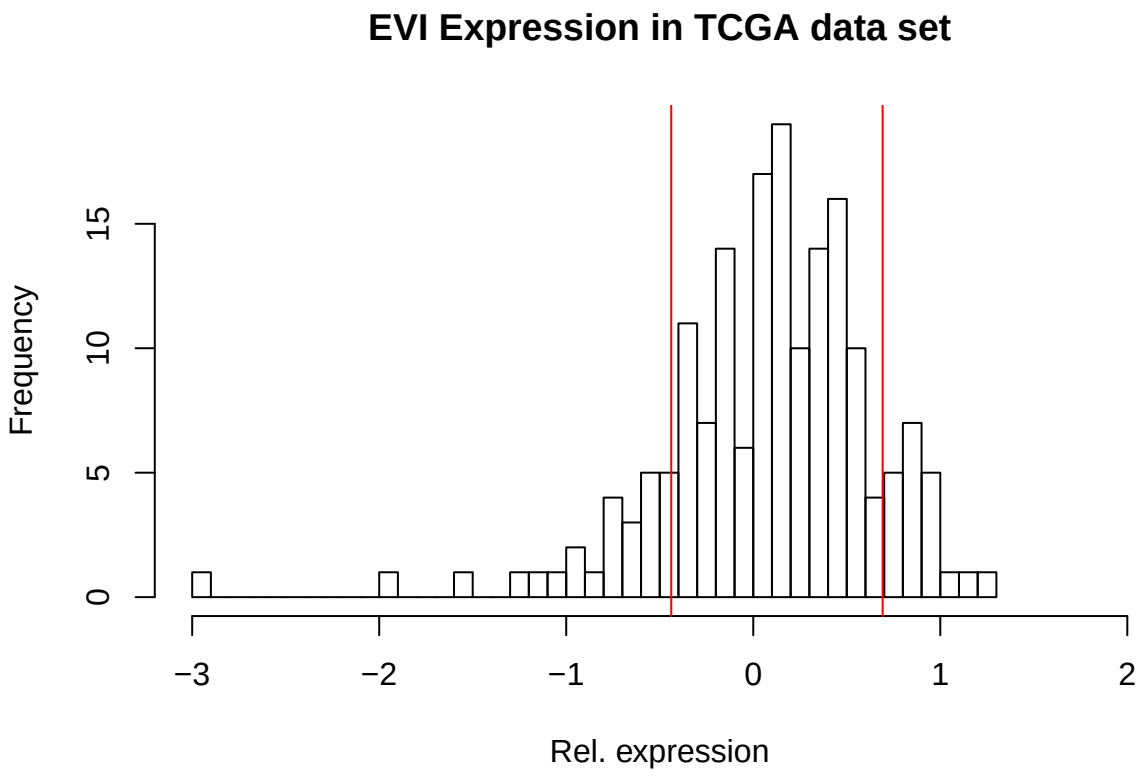
```
# select EVI (=GPR177)
evi.exp<-exp.ma[rownames(exp.ma)=="GPR177",]

# select samples, whose EVI expression differs more than 1 standard deviation from median
up.reg<-evi.exp>(median(evi.exp)+1*sd(evi.exp))
down.reg<-evi.exp<(median(evi.exp)-1*sd(evi.exp))

evi.up<-evi.exp[up.reg]
evi.down<-evi.exp[down.reg]
```

We can now visualize the distribution of Evi expression and the selection thresholds for further analysis. Thresholds are indicated by red lines.

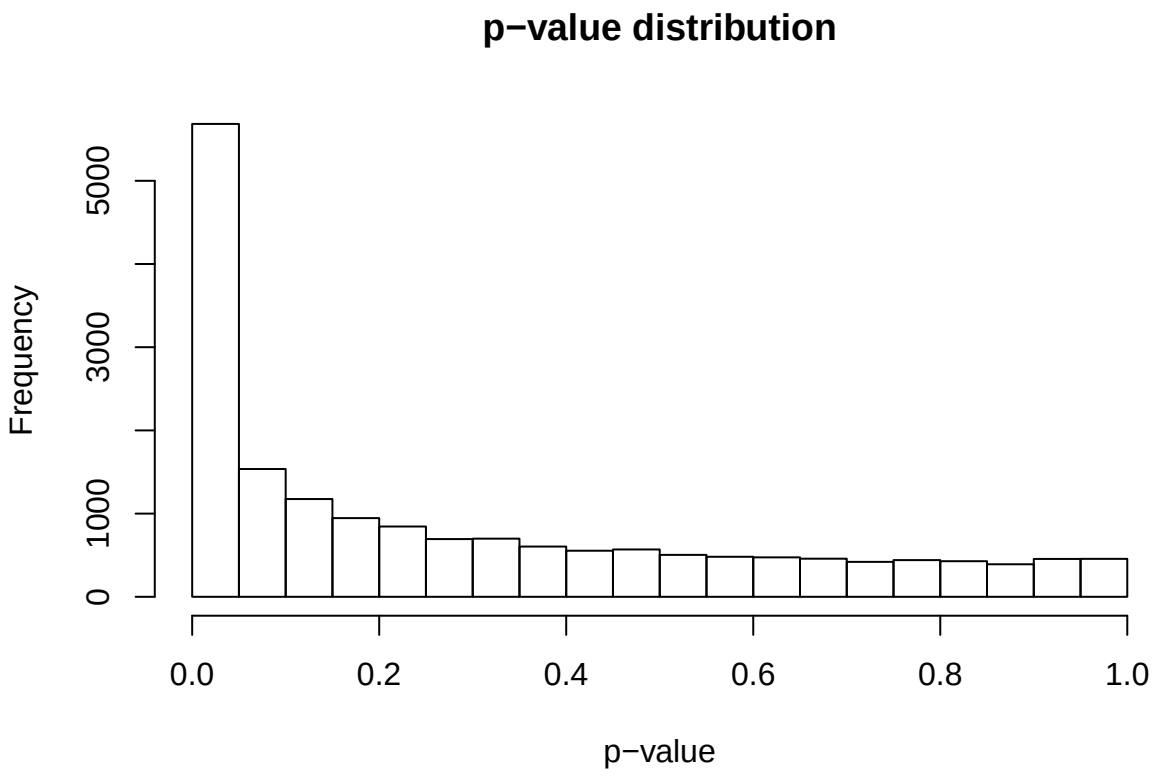
```
hist(evi.exp, breaks=30, xlim=c(-3,2), xlab="Rel. expression",
     main="EVI Expression in TCGA data set")
abline(v=median(evi.exp)+c(1,-1)*sd(evi.exp), col="red")
```



Statistical test

We use a t-test statistic to identify differentially expressed genes between low EVI samples and high EVI samples. Finally p-values are corrected for multiple testing.

```
ttest.data<-exp.ma[,c(names(evi.up), names(evi.down))]
label<- c(rep(1,length(evi.up)), rep(0, length(evi.down)))
# T-test
tStat<- mt.teststat(ttest.data, classlabel=label, test="t")
# get p-value
ttest.p<-2*pt(-abs(tStat), df=(ncol(ttest.data)-2))
# plot p-value distribution
hist(ttest.p, breaks=20, xlab="p-value", main="p-value distribution")
```




```
# adjust p-Value for multiple testing by Benjamini-Hochberg
pAdjusted <- mt.rawp2adjp(ttest.p, proc = c("BH"))
```

In the last step a final table is generated comprised of within-group median expression levels, the between-group fold change and the raw/adjusted p-values.

```
# assign gene names
pAdj<-(pAdjusted$adjp[order(pAdjusted$index),])
rownames(pAdj)<- rownames(ttest.data)

#calculate logFC
low.EVI.median<-apply(ttest.data[,label==0], 1, median)
high.EVI.median<-apply(ttest.data[,label==1], 1, median)
# calculate fold change
fc.high_low<- 2**(low.EVI.median-high.EVI.median)
# final table
table.EVI<-cbind(low.EVI.median, high.EVI.median, fc.high_low, pAdj)
colnames(table.EVI)[1:3]<- c("low EVI - TCGA", "high EVI - TCGA",
                             "FoldChange [lowEVI/highEVI]")
```

Session info

```
sessionInfo()
```

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.12.6 (Sierra)
##
## locale:
## [1] de_DE.UTF-8/de_DE.UTF-8/de_DE.UTF-8/C/de_DE.UTF-8/de_DE.UTF-8
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] multtest_2.28.0     Biobase_2.32.0      BiocGenerics_0.18.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.12    lattice_0.20-35 digest_0.6.12   rprojroot_1.2
##  [5] MASS_7.3-47     grid_3.3.1      backports_1.1.0 stats4_3.3.1
##  [9] magrittr_1.5    evaluate_0.10.1 stringi_1.1.5   Matrix_1.2-11
## [13] rmarkdown_1.6   splines_3.3.1   tools_3.3.1     stringr_1.2.0
## [17] yaml_2.1.14     survival_2.41-3 htmltools_0.3.6 knitr_1.17
```