GEOsuppl analysis of Series GSE112049

Hyun Yong Jin

April 26, 2019

```
#Gautam et al., Nature Immunology 2019
#RNAseq data was linked to the paper, GSE112049
#Look up Notch target genes in the dataset

installed.packages("GEOquery")
```

```
##      Package LibPath Version Priority Depends Imports LinkingTo Suggests
##      Enhances License License_is_FOSS License_restricts_use OS_type Archs
##      MD5sum NeedsCompilation Built
```

```
#install.packages("GEOquery") #Not available for 3.5.1
R.Version() #current version is 3.5.1, 2018-07-02
```

```
## $platform
## [1] "x86_64-w64-mingw32"
##
## $arch
## [1] "x86_64"
##
## $os
## [1] "mingw32"
##
## $system
## [1] "x86_64, mingw32"
##
## $status
## [1] ""
##
## $major
## [1] "3"
##
## $minor
## [1] "5.1"
##
## $year
## [1] "2018"
##
## $month
## [1] "07"
##
## $day
## [1] "02"
##
## $`svn rev`
## [1] "74947"
##
## $language
## [1] "R"
##
## $version.string
## [1] "R version 3.5.1 (2018-07-02)"
##
## $nickname
## [1] "Feather Spray"
```

```
#To install for R version 3.5
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("GEOquery", version = "3.8")
```

```
## Bioconductor version 3.8 (BiocManager 1.30.4), R 3.5.1 (2018-07-02)
```

```
## Installing package(s) 'GEOquery'
```

```
## package 'GEOquery' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\hjin02\AppData\Local\Temp\RtmpuMImVU\downloaded_packages
```

```
## installation path not writeable, unable to update packages: boot, class,
##   cluster, codetools, foreign, lattice, MASS, Matrix, mgcv, nlme, rpart,
##   survival
```

```
library(GEOquery)
```

```
## Warning: package 'GEOquery' was built under R version 3.5.2
```

```
## Loading required package: Biobase
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind,
##     colMeans, colnames, colSums, dirname, do.call, duplicated,
##     eval, evalq, Filter, Find, get, grep, grepl, intersect,
##     is.unsorted, lapply, lengths, Map, mapply, match, mget, order,
##     paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
##     Reduce, rowMeans, rownames, rowSums, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which, which.max,
##     which.min
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
## Setting options('download.file.method.GEOquery'='auto')
```

```
## Setting options('GEOquery.inmemory.gpl'=FALSE)
```

```
getGEOSuppFiles('GSE112049', fetch_files = FALSE) #Check to see what's in the s
upplement files
```

```
##                             fname
## 1 GSE112049_rnaseq_matrix.txt.gz
#
#
                   url
## 1 https://ftp.ncbi.nlm.nih.gov/geo/series/GSE112nnn/GSE112049/suppl//GSE1120
49_rnaseq_matrix.txt.gz
```

```
getGEOSuppFiles('GSE112049')
```

```
##                                                                    size
## C:/Users/hjin02/Documents/GSE112049/GSE112049_rnaseq_matrix.txt.gz 769171
##                                                                   isdir
## C:/Users/hjin02/Documents/GSE112049/GSE112049_rnaseq_matrix.txt.gz FALSE
##                                                                    mode
## C:/Users/hjin02/Documents/GSE112049/GSE112049_rnaseq_matrix.txt.gz  666
#
#
    mtime
## C:/Users/hjin02/Documents/GSE112049/GSE112049_rnaseq_matrix.txt.gz 2019-04-2
6 14:18:45
#
#
    ctime
## C:/Users/hjin02/Documents/GSE112049/GSE112049_rnaseq_matrix.txt.gz 2019-04-2
6 10:09:33
#
#
    atime
## C:/Users/hjin02/Documents/GSE112049/GSE112049_rnaseq_matrix.txt.gz 2019-04-2
6 10:09:33
##                                                                     exe
## C:/Users/hjin02/Documents/GSE112049/GSE112049_rnaseq_matrix.txt.gz  no
```

```
tab <- read.delim("GSE112049/GSE112049_rnaseq_matrix.txt.gz")


library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages ---------------------------------------- tidyverse
1.2.1 --
```

```
## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'tidyr' was built under R version 3.5.3
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.3
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
## Warning: package 'stringr' was built under R version 3.5.3
```

```
## Warning: package 'forcats' was built under R version 3.5.3
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflic
ts() --
## x dplyr::combine()    masks Biobase::combine(), BiocGenerics::combine()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
## x ggplot2::Position() masks BiocGenerics::Position(), base::Position()
```

```
tib <-as.tibble(tab) #Let's convert it to tibble format
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new se
mantics).
## This warning is displayed once per session.
```

```
tib
```

```
## # A tibble: 13,017 x 7
##    Gene        KO1    KO2    KO3    WT1    WT2    WT3
##    <fct>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Mrpl15   3.38    3.42   3.43   3.45   3.55   3.43
##  2 Lypla1   5.64    5.76   5.74   5.71   5.73   5.65
##  3 Gm19860 -0.380  -0.518 -0.610 -1.04  -1.13  -0.638
##  4 Tcea1    5.81    5.89   5.83   5.86   5.87   5.84
##  5 Atp6v1h  5.15    5.20   5.15   5.23   5.19   5.21
##  6 Rb1cc1   3.97    3.85   3.99   3.95   3.90   3.94
##  7 Pcmtd1   3.92    3.82   3.94   3.84   3.81   3.92
##  8 Rrs1     3.86    3.79   3.79   3.88   3.80   3.71
##  9 Mybl1    0.0702  0.277  0.294  0.154 -0.151 -0.112
## 10 Vcpip1   4.60    4.41   4.60   4.57   4.32   4.47
## # ... with 13,007 more rows
```

```
summary(tib)
```

```
##             Gene            KO1              KO2              KO3
##  0610007P14Rik:    1   Min.   :-2.322   Min.   :-2.3219   Min.   :-2.322
##  0610009B22Rik:    1   1st Qu.: 0.884   1st Qu.: 0.8422   1st Qu.: 0.909
##  0610009L18Rik:    1   Median : 3.193   Median : 3.1788   Median : 3.209
##  0610009O20Rik:    1   Mean   : 2.864   Mean   : 2.8507   Mean   : 2.878
##  0610010F05Rik:    1   3rd Qu.: 4.689   3rd Qu.: 4.6956   3rd Qu.: 4.696
##  0610010K14Rik:    1   Max.   :13.095   Max.   :13.0026   Max.   :12.962
##  (Other)      :13011
##       WT1               WT2               WT3
##  Min.   :-2.3219   Min.   :-2.3219   Min.   :-2.3219
##  1st Qu.: 0.8435   1st Qu.: 0.8003   1st Qu.: 0.8523
##  Median : 3.1779   Median : 3.1477   Median : 3.1613
##  Mean   : 2.8388   Mean   : 2.8196   Mean   : 2.8430
##  3rd Qu.: 4.6806   3rd Qu.: 4.6665   3rd Qu.: 4.6807
##  Max.   :12.8816   Max.   :12.8278   Max.   :12.8814
##
```

```
lin <- function(x, na.rm = FALS) (2^x) #Function to linerize each values

tiblin <- tib %>% mutate_at(vars(starts_with("W")), lin) %>% mutate_at(vars(sta
rts_with("K")), lin) #Convert log value to linear when header starts with W (fo
r WT) or K (for KO)
summary(tiblin)
```

```
##             Gene            KO1                KO2
##  0610007P14Rik:    1   Min.   :   0.200   Min.   :   0.200
##  0610009B22Rik:    1   1st Qu.:   1.845   1st Qu.:   1.793
##  0610009L18Rik:    1   Median :   9.148   Median :   9.056
##  0610009O20Rik:    1   Mean   :  40.082   Mean   :  40.147
##  0610010F05Rik:    1   3rd Qu.:  25.791   3rd Qu.:  25.913
##  0610010K14Rik:    1   Max.   :8752.593   Max.   :8207.060
##  (Other)      :13011
##       KO3                WT1                WT2
##  Min.   :   0.200   Min.   :   0.200   Min.   :   0.200
##  1st Qu.:   1.878   1st Qu.:   1.794   1st Qu.:   1.742
##  Median :   9.249   Median :   9.050   Median :   8.862
##  Mean   :  39.561   Mean   :  39.805   Mean   :  40.908
##  3rd Qu.:  25.915   3rd Qu.:  25.646   3rd Qu.:  25.396
##  Max.   :7981.718   Max.   :7546.438   Max.   :7270.121
##
##       WT3
##  Min.   :   0.200
##  1st Qu.:   1.805
##  Median :   8.947
##  Mean   :  40.460
##  3rd Qu.:  25.647
##  Max.   :7545.640
##
```

```
tibWT <- select(tiblin, starts_with("W"))
tibKO <- select(tiblin, starts_with("K"))

tibFC <- tiblin %>%
    mutate(KO.mean = rowMeans(tibKO)) %>%
    mutate(WT.mean = rowMeans(tibWT)) %>%
    mutate(FC = KO.mean/WT.mean)

library(broom) #For function tidy
```

```
## Warning: package 'broom' was built under R version 3.5.3
```

```
#for-loop t-test
testresults <- vector("list", nrow(tiblin)) #To make empty vector

for (j in seq(nrow(tiblin))) {
  testresults[[j]] <- tidy(t.test(as.data.frame(tibKO[j,]), as.data.frame(tibWT
[j,])))
}

head(testresults)
```

```
## [[1]]
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1   -0.484      10.7      11.1     -1.57   0.230      2.55    -1.57
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
##
## [[2]]
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1    0.455      52.4      52.0     0.293   0.785      3.63    -4.03
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
##
## [[3]]
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1    0.179     0.707     0.529      2.70  0.0694      3.19  -0.0255
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
##
## [[4]]
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1   -0.522      57.4      57.9    -0.534   0.639      2.38    -4.15
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
##
## [[5]]
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1    -1.20      35.9      37.1     -2.28  0.0960      3.42    -2.77
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
##
## [[6]]
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1   0.0795      15.3      15.3     0.168   0.880      2.43    -1.65
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
```

```
t_stats = do.call(rbind, testresults) #list to dataframe

big_data <-bind_cols(tibFC, t_stats)
glimpse(big_data)
```

```
## Observations: 13,017
## Variables: 20
## $ Gene        <fct> Mrpl15, Lypla1, Gm19860, Tcea1, Atp6v1h, Rb1cc1, P...
## $ KO1         <dbl> 10.4389198, 49.9261309, 0.7684682, 56.0007348, 35....
## $ KO2         <dbl> 10.7268275, 54.0140733, 0.6984076, 59.1800863, 36....
## $ KO3         <dbl> 10.7909978, 53.3007999, 0.6552064, 57.0756944, 35....
## $ WT1         <dbl> 10.9130663, 52.4568042, 0.4855006, 57.9796644, 37....
## $ WT2         <dbl> 11.7086522, 53.2164299, 0.4575494, 58.4202735, 36....
## $ WT3         <dbl> 10.7861916, 50.2035600, 0.6424909, 57.4218954, 37....
## $ KO.mean     <dbl> 10.6522484, 52.4136680, 0.7073607, 57.4188385, 35....
## $ WT.mean     <dbl> 11.1359700, 51.9589314, 0.5285136, 57.9406111, 37....
## $ FC          <dbl> 0.9565622, 1.0087518, 1.3383964, 0.9909947, 0.9675...
## $ estimate    <dbl> -0.48372166, 0.45473664, 0.17884710, -0.52177260, ...
## $ estimate1   <dbl> 10.6522484, 52.4136680, 0.7073607, 57.4188385, 35....
## $ estimate2   <dbl> 11.1359700, 51.9589314, 0.5285136, 57.9406111, 37....
## $ statistic   <dbl> -1.56896205, 0.29305693, 2.69564859, -0.53385769, ...
## $ p.value     <dbl> 0.23011685, 0.78546480, 0.06937386, 0.63918548, 0....
## $ parameter   <dbl> 2.551672, 3.628053, 3.186717, 2.379388, 3.416183, ...
## $ conf.low    <dbl> -1.57000743, -4.03338126, -0.02546822, -4.14553105...
## $ conf.high   <dbl> 0.60256412, 4.94285454, 0.38316242, 3.10198586, 0....
## $ method      <chr> "Welch Two Sample t-test", "Welch Two Sample t-tes...
## $ alternative <chr> "two.sided", "two.sided", "two.sided", "two.sided"...
```

```
head(big_data)
```

```
## # A tibble: 6 x 20
##   Gene     KO1    KO2    KO3    WT1    WT2    WT3 KO.mean WT.mean    FC
##   <fct>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>   <dbl> <dbl>
## 1 Mrpl~ 10.4   10.7   10.8   10.9   11.7   10.8    10.7    11.1   0.957
## 2 Lypl~ 49.9   54.0   53.3   52.5   53.2   50.2    52.4    52.0   1.01
## 3 Gm19~  0.768  0.698  0.655  0.486  0.458  0.642   0.707   0.529 1.34
## 4 Tcea1 56.0   59.2   57.1   58.0   58.4   57.4    57.4    57.9   0.991
## 5 Atp6~ 35.4   36.7   35.4   37.5   36.6   37.1    35.9    37.1   0.968
## 6 Rb1c~ 15.7   14.4   15.9   15.5   15.0   15.4    15.3    15.3   1.01
## # ... with 10 more variables: estimate <dbl>, estimate1 <dbl>,
## #   estimate2 <dbl>, statistic <dbl>, p.value <dbl>, parameter <dbl>,
## #   conf.low <dbl>, conf.high <dbl>, method <chr>, alternative <chr>
```

```
write.csv(big_data, file="GSE112049_all_data.csv") #write all data


#Select genes with high fold changes
signif <- big_data %>%
  select(Gene, KO.mean, WT.mean, FC, p.value) %>%
  filter(p.value<0.01 & (FC>2 | FC<1/2)) %>%
  arrange(desc(FC))

signif
```

```
## # A tibble: 94 x 5
##    Gene      KO.mean WT.mean    FC    p.value
##    <fct>       <dbl>   <dbl> <dbl>      <dbl>
##  1 Gm8210      28.6    0.449 63.9  0.000604
##  2 Gm13841     92.3    3.48  26.5  0.00208
##  3 Tma7-ps      6.87   0.404 17.0  0.00599
##  4 Batf3        3.31   0.413  8.03 0.00332
##  5 Tmprss13     6.11   0.859  7.12 0.000617
##  6 Kcnip3       1.16   0.212  5.45 0.00421
##  7 Aldh2        7.46   1.60   4.68 0.000236
##  8 Rps19-ps4    7.41   1.61   4.60 0.00198
##  9 Xcl1        17.1    4.20   4.08 0.00510
## 10 Vwa5a        6.75   1.73   3.89 0.00000416
## # ... with 84 more rows
```

```
write.csv(signif, file="GSE112049_significant_2fold_only.csv")



#Focus on Notch and other interesting genes, align based on FC
#There should be more elegant ways of doing it.


selectgenes <-
  big_data %>%
  select(Gene, KO.mean, WT.mean, FC, p.value) %>%
  filter(Gene == "Notch1"| Gene=="Notch2" | Gene=="Hes1" | Gene=="Hes2" | Gene=
="Nrarp"| Gene=="Dtx1"| Gene == "Rbpj" |Gene =="Actb" | Gene=="Id3"| Gene=="Fox
o1" | Gene=="Tcf7" | Gene=="Cxcr5" | Gene=="Il7r" | Gene=="Ctla4" | Gene=="pcdc
1" | Gene=="Havcr2" | Gene=="Klrg1" | Gene=="Zeb2" | Gene=="Bcl2" | Gene=="My
b") %>%
  arrange(desc(FC))

selectgenes
```

```
## # A tibble: 18 x 5
##    Gene    KO.mean  WT.mean    FC  p.value
##    <fct>     <dbl>    <dbl> <dbl>    <dbl>
##  1 Cxcr5     3.89     1.54  2.53  0.00468
##  2 Klrg1    25.8     10.5   2.46  0.0110
##  3 Zeb2     18.1      8.12  2.23  0.00322
##  4 Myb      37.2     17.8   2.09  0.00661
##  5 Havcr2   12.5      6.68  1.87  0.0356
##  6 Nrarp     9.18     6.03  1.52  0.0418
##  7 Notch2   15.2     13.1   1.16  0.0892
##  8 Rbpj      7.87     7.17  1.10  0.0590
##  9 Foxo1    44.3     40.5   1.09  0.166
## 10 Notch1   23.4     21.6   1.09  0.298
## 11 Ctla4    49.1     45.2   1.09  0.819
## 12 Actb   2182.    2207.    0.989 0.903
## 13 Dtx1     99.3    104.    0.953 0.313
## 14 Id3      14.2     15.0   0.952 0.775
## 15 Hes1      0.330    0.370 0.891 0.386
## 16 Tcf7    446.     536.    0.832 0.0177
## 17 Bcl2      3.13     4.72  0.662 0.000310
## 18 Il7r     45.4    101.    0.450 0.000975
```

```
write.csv(selectgenes, file="GSE112049_manually_selected_genes.csv")

#Bcl2 and Tcf7 downregulation of Bcl2 upregulatoin in KO cells were recapitulat
ed but Myb levels were even higher in KO cells.
#Myb floxed allele is targeted for exon II only. Likely detection of non-funtio
nal Myb transcript accumulated in KO. Bender et al., NI 2004
```