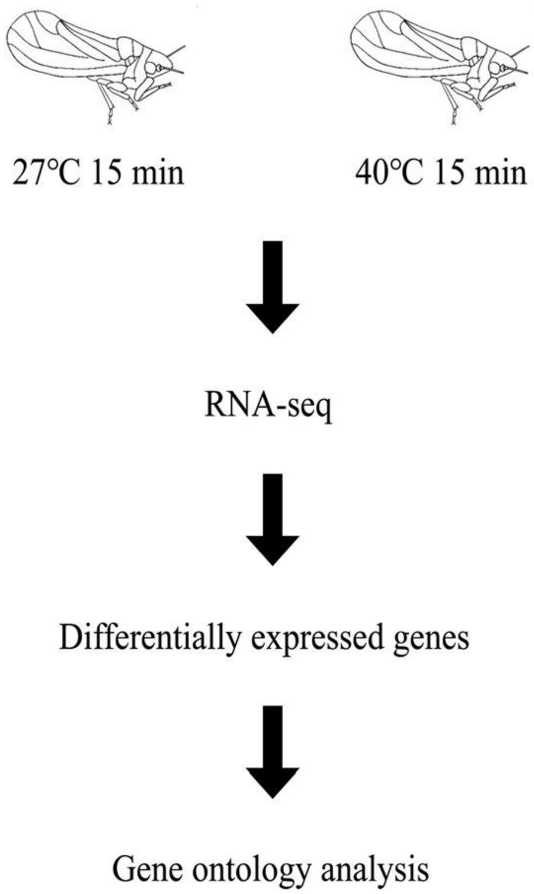
WEEK 15. PRACTICE

Differential expression analysis using DESeq2

Differentially expressed gene(DEG) analysis



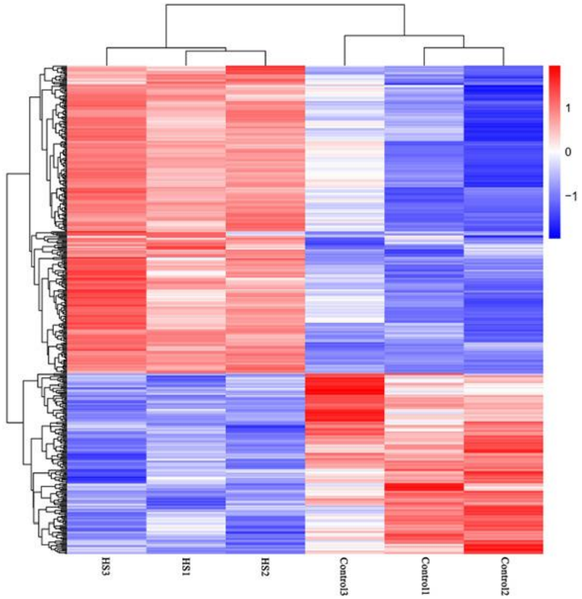
27℃ 15 min

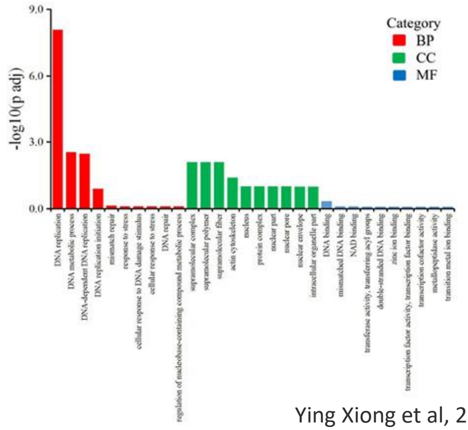
40℃ 15 min

RNA-seq

Differentially expressed genes

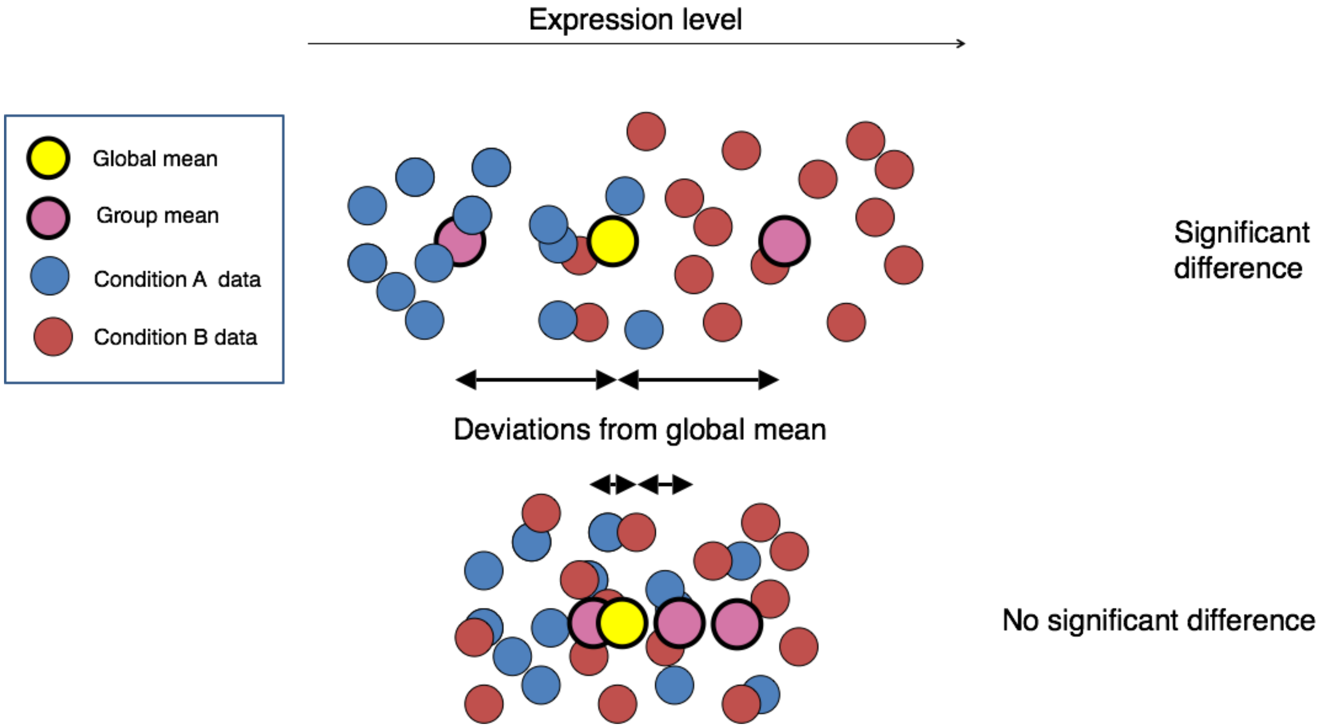
Gene ontology analysis





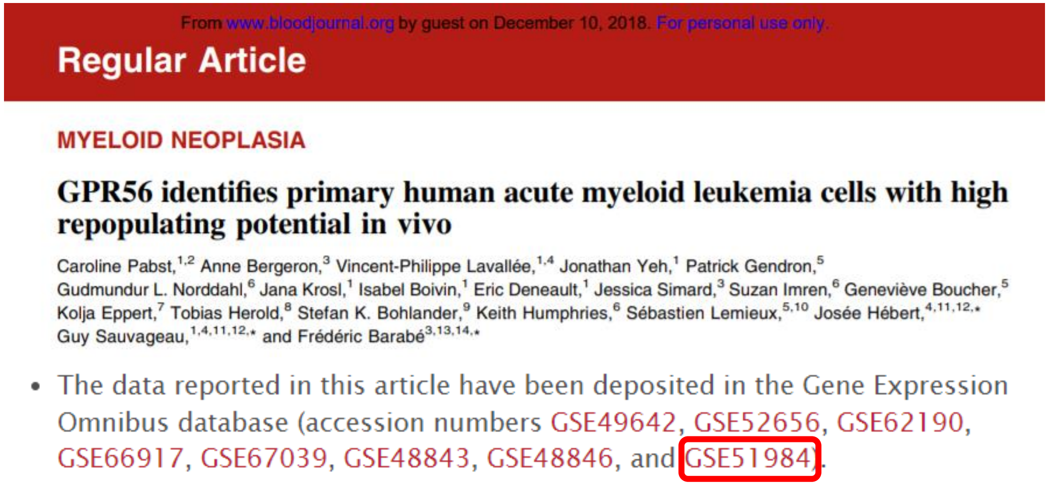
Ying Xiong et al, 2019

Differential expression analysis with DESeq2



https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html

Input data for this week



From www.bloodjournal.org by guest on December 10, 2018. For personal use only.

Regular Article

MYELOID NEOPLASIA

GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo

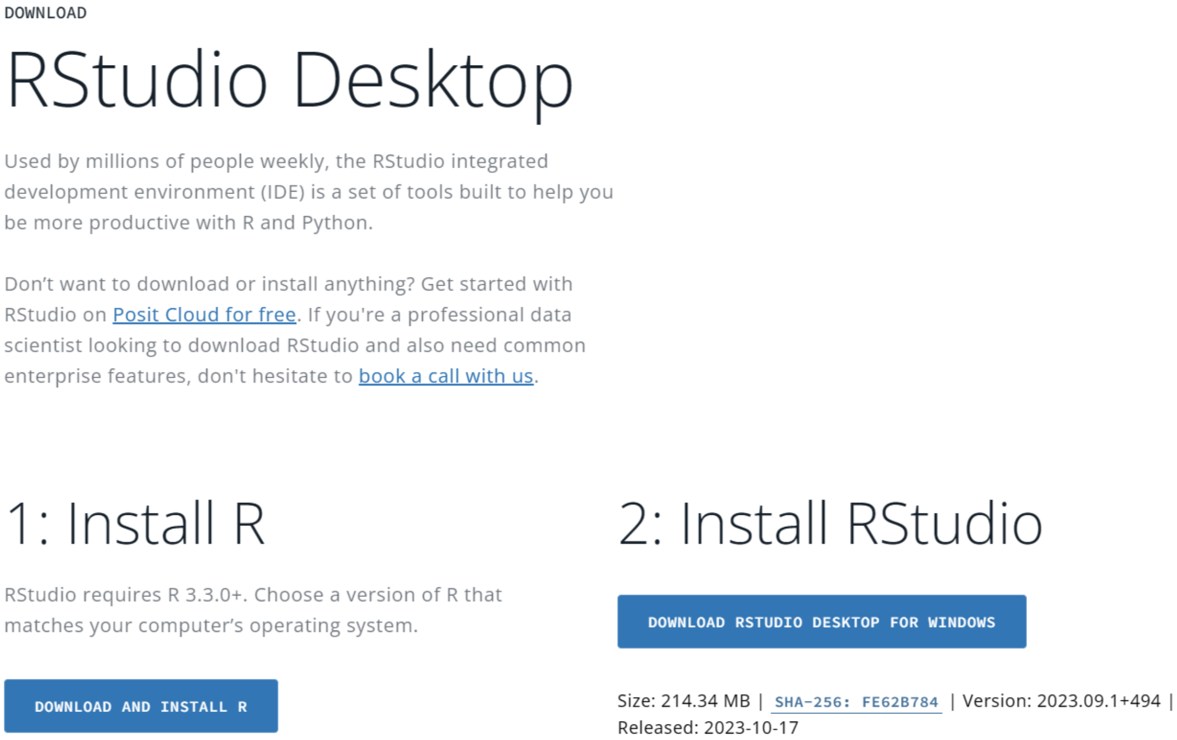
Caroline Pabst,[1,2] Anne Bergeron,[3] Vincent-Philippe Lavallée,[1,4] Jonathan Yeh,[1] Patrick Gendron,[5] Gudmundur L. Norddahl,[6] Jana Krosl,[1] Isabel Boivin,[1] Eric Deneault,[1] Jessica Simard,[3] Suzan Imren,[6] Geneviève Boucher,[5] Kolja Eppert,[7] Tobias Herold,[8] Stefan K. Bohlander,[9] Keith Humphries,[6] Sébastien Lemieux,[5,10] Josée Hébert,[4,11,12,*] Guy Sauvageau,[1,4,11,12,*] and Frédéric Barabé[3,13,14,*]

- The data reported in this article have been deposited in the Gene Expression Omnibus database (accession numbers GSE49642, GSE52656, GSE62190, GSE66917, GSE67039, GSE48843, GSE48846, and GSE51984).

- RNA-seq of T-cells and B-cells (5 replicates for each cell types)
  - the genes whose variances of RNA-seq expression values (FPKM) among the samples are high were selected (501 genes)

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51984

install R studio



DOWNLOAD

RStudio Desktop

Used by millions of people weekly, the RStudio integrated development environment (IDE) is a set of tools built to help you be more productive with R and Python.

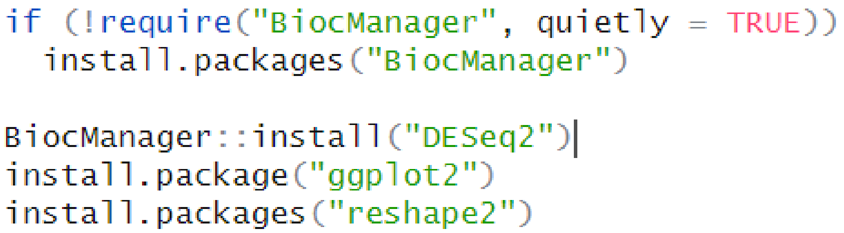
Don’t want to download or install anything? Get started with RStudio on Posit Cloud for free. If you're a professional data scientist looking to download RStudio and also need common enterprise features, don't hesitate to book a call with us.

1: Install R

RStudio requires R 3.3.0+. Choose a version of R that matches your computer’s operating system.

2: Install RStudio

Size: 214.34 MB | SHA-256: FE62B784 | Version: 2023.09.1+494 | Released: 2023-10-17



```
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("DESeq2")
install.package("ggplot2")
install.packages("reshape2")
```

Identification of DEGs using DESeq2

- DESeq2

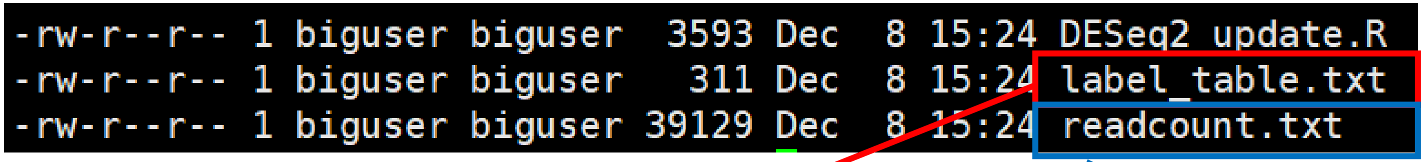
  - test for differential expression based on a model using the negative binomial distribution

- Usage (in R):

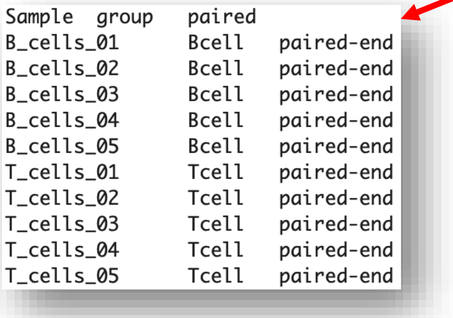
```
dds <- DESeqDataSetFromMatrix(countData = <read count table>,
                              colData = <label table>,
                              design = ~ <group attribute in label>)
dds <- DESeq(dds)
```

Input files for the DESeq2

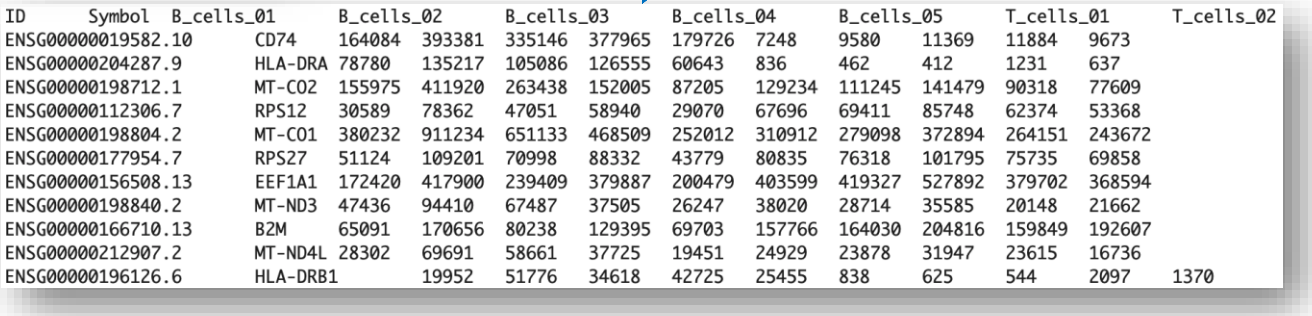
cp /home/biguser/tutor/Week15/inputTables/* .



```
-rw-r--r-- 1 biguser biguser  3593 Dec  8 15:24 DESeq2_update.R
-rw-r--r-- 1 biguser biguser   311 Dec  8 15:24 label_table.txt
-rw-r--r-- 1 biguser biguser 39129 Dec  8 15:24 readcount.txt
```

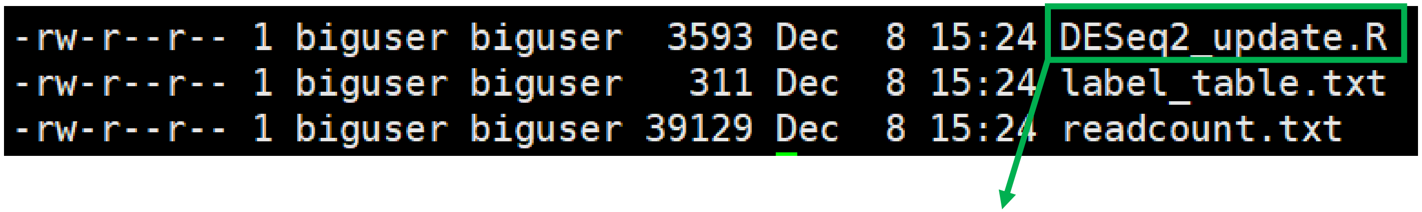


```
Sample  group   paired
B_cells_01      Bcell   paired-end
B_cells_02      Bcell   paired-end
B_cells_03      Bcell   paired-end
B_cells_04      Bcell   paired-end
B_cells_05      Bcell   paired-end
T_cells_01      Tcell   paired-end
T_cells_02      Tcell   paired-end
T_cells_03      Tcell   paired-end
T_cells_04      Tcell   paired-end
T_cells_05      Tcell   paired-end
```

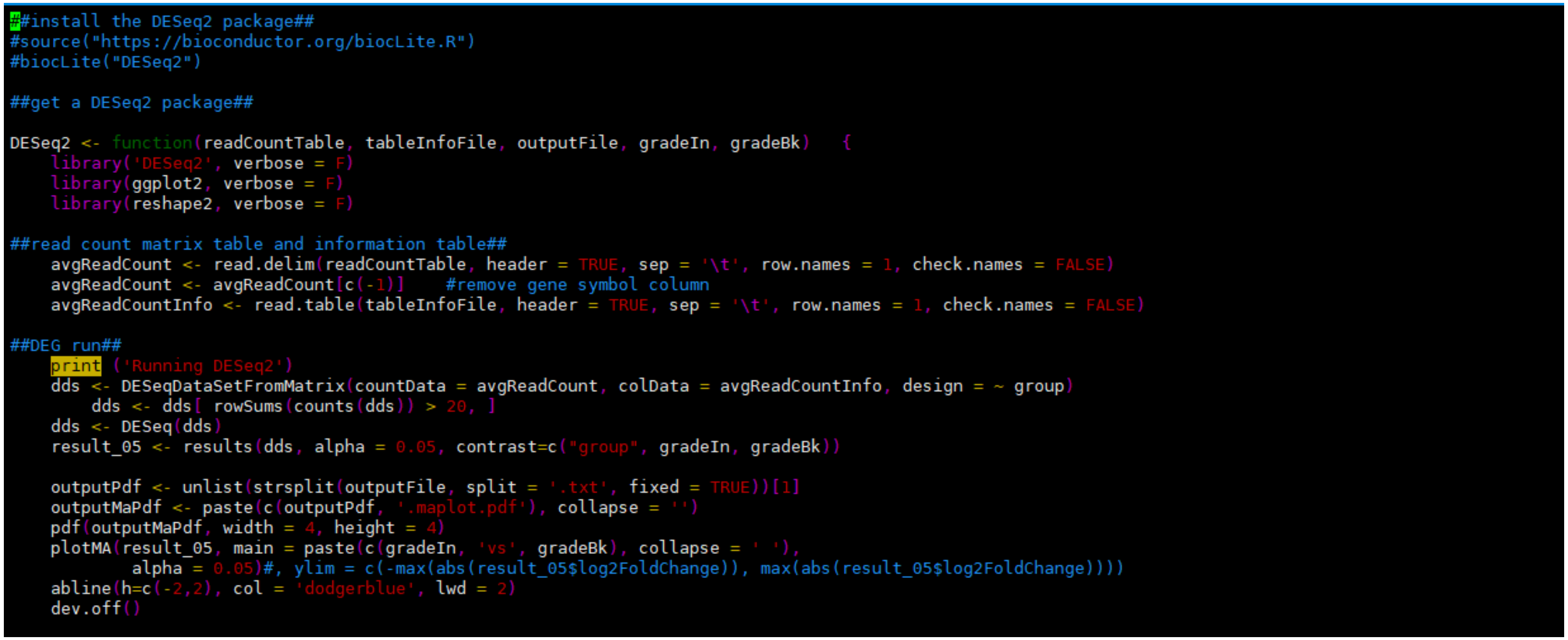


```
ID      Symbol  B_cells_01      B_cells_02      B_cells_03      B_cells_04      B_cells_05      T_cells_01      T_cells_02
ENSG00000019582.10      CD74    164084  393381  335146  377965  179726  7248    9580    11369   11884   9673
ENSG00000204287.9       HLA-DRA 78780   135217  105086  126555  60643   836     462     412     1231    637
ENSG00000198712.1       MT-CO2  155975  411920  263438  152005  87205   129234  111245  141479  90318   77609
ENSG00000112306.7       RPS12   30589   78362   47051   58940   29070   67696   69411   85748   62374   53368
ENSG00000198804.2       MT-CO1  380232  911234  651133  468509  252012  310912  279098  372894  264151  243672
ENSG00000177954.7       RPS27   51124   109201  70998   88332   43779   80835   76318   101795  75735   69858
ENSG00000156508.13      EEF1A1  172420  417900  239409  379887  200479  403599  419327  527892  379702  368594
ENSG00000198840.2       MT-ND3  47436   94410   67487   37505   26247   38020   28714   35585   20148   21662
ENSG00000166710.13      B2M     65091   170656  80238   129395  69703   157766  164030  204816  159849  192607
ENSG00000212907.2       MT-ND4L 28302   69691   58661   37725   19451   24929   23878   31947   23615   16736
ENSG00000196126.6       HLA-DRB1        19952   51776   34618   42725   25455   838     625     544     2097    1370
```

R code to run DESeq2

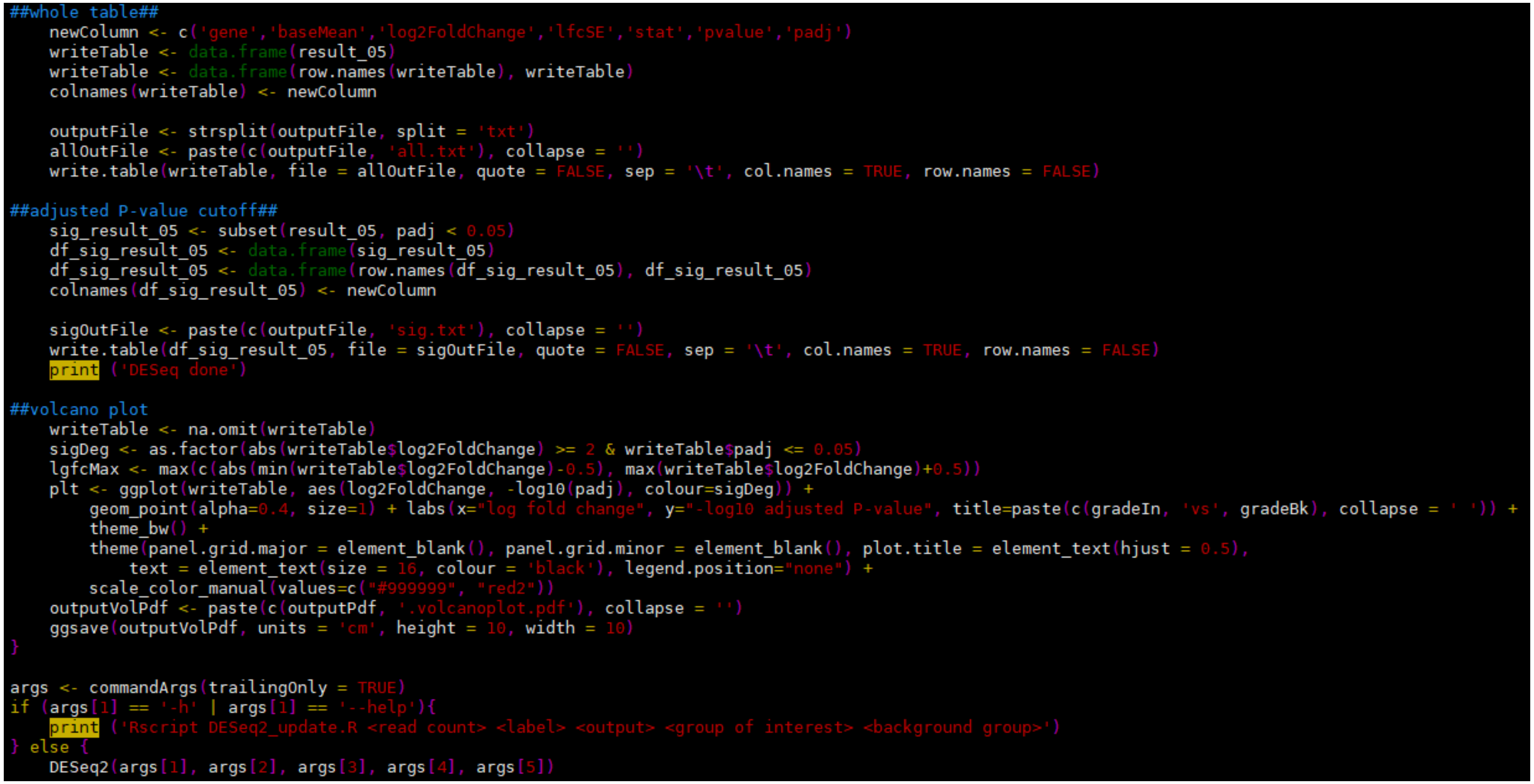


```
-rw-r--r-- 1 biguser biguser  3593 Dec  8 15:24 DESeq2_update.R
-rw-r--r-- 1 biguser biguser   311 Dec  8 15:24 label_table.txt
-rw-r--r-- 1 biguser biguser 39129 Dec  8 15:24 readcount.txt
```



```
##install the DESeq2 package##
#source("https://bioconductor.org/biocLite.R")
#biocLite("DESeq2")

##get a DESeq2 package##

DESeq2 <- function(readCountTable, tableInfoFile, outputFile, gradeIn, gradeBk)   {
    library('DESeq2', verbose = F)
    library(ggplot2, verbose = F)
    library(reshape2, verbose = F)

##read count matrix table and information table##
    avgReadCount <- read.delim(readCountTable, header = TRUE, sep = '\t', row.names = 1, check.names = FALSE)
    avgReadCount <- avgReadCount[c(-1)]    #remove gene symbol column
    avgReadCountInfo <- read.table(tableInfoFile, header = TRUE, sep = '\t', row.names = 1, check.names = FALSE)

##DEG run##
    print ('Running DESeq2')
    dds <- DESeqDataSetFromMatrix(countData = avgReadCount, colData = avgReadCountInfo, design = ~ group)
        dds <- dds[ rowSums(counts(dds)) > 20, ]
    dds <- DESeq(dds)
    result_05 <- results(dds, alpha = 0.05, contrast=c("group", gradeIn, gradeBk))

    outputPdf <- unlist(strsplit(outputFile, split = '.txt', fixed = TRUE))[1]
    outputMaPdf <- paste(c(outputPdf, '.maplot.pdf'), collapse = '')
    pdf(outputMaPdf, width = 4, height = 4)
    plotMA(result_05, main = paste(c(gradeIn, 'vs', gradeBk), collapse = ' '),
            alpha = 0.05)#, ylim = c(-max(abs(result_05$log2FoldChange)), max(abs(result_05$log2FoldChange))))
    abline(h=c(-2,2), col = 'dodgerblue', lwd = 2)
    dev.off()
```

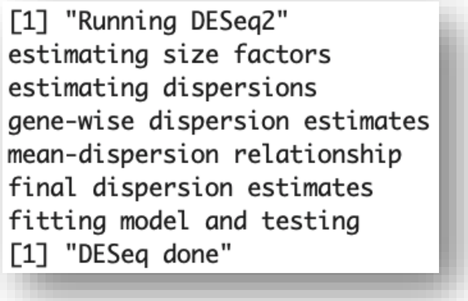
R code to run DESeq2



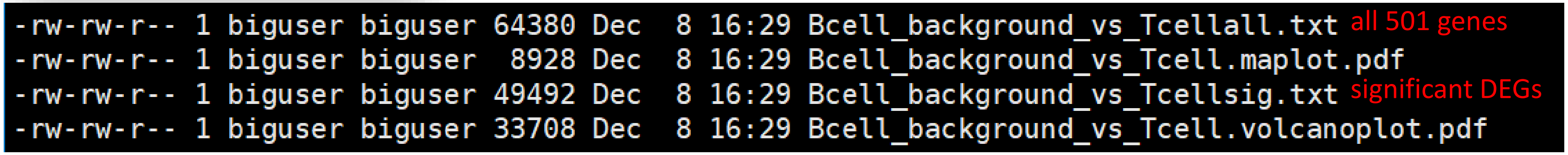
```
##whole table##
    newColumn <- c('gene','baseMean','log2FoldChange','lfcSE','stat','pvalue','padj')
    writeTable <- data.frame(result_05)
    writeTable <- data.frame(row.names(writeTable), writeTable)
    colnames(writeTable) <- newColumn

    outputFile <- strsplit(outputFile, split = 'txt')
    allOutFile <- paste(c(outputFile, 'all.txt'), collapse = '')
    write.table(writeTable, file = allOutFile, quote = FALSE, sep = '\t', col.names = TRUE, row.names = FALSE)

##adjusted P-value cutoff##
    sig_result_05 <- subset(result_05, padj < 0.05)
    df_sig_result_05 <- data.frame(sig_result_05)
    df_sig_result_05 <- data.frame(row.names(df_sig_result_05), df_sig_result_05)
    colnames(df_sig_result_05) <- newColumn

    sigOutFile <- paste(c(outputFile, 'sig.txt'), collapse = '')
    write.table(df_sig_result_05, file = sigOutFile, quote = FALSE, sep = '\t', col.names = TRUE, row.names = FALSE)
    print ('DESeq done')

##volcano plot
    writeTable <- na.omit(writeTable)
    sigDeg <- as.factor(abs(writeTable$log2FoldChange) >= 2 & writeTable$padj <= 0.05)
    lgfcMax <- max(c(abs(min(writeTable$log2FoldChange)-0.5), max(writeTable$log2FoldChange)+0.5))
    plt <- ggplot(writeTable, aes(log2FoldChange, -log10(padj), colour=sigDeg)) +
        geom_point(alpha=0.4, size=1) + labs(x="log fold change", y="-log10 adjusted P-value", title=paste(c(gradeIn, 'vs', gradeBk), collapse = ' ')) +
        theme_bw() +
        theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(), plot.title = element_text(hjust = 0.5),
            text = element_text(size = 16, colour = 'black'), legend.position="none") +
        scale_color_manual(values=c("#999999", "red2"))
    outputVolPdf <- paste(c(outputPdf, '.volcanoplot.pdf'), collapse = '')
    ggsave(outputVolPdf, units = 'cm', height = 10, width = 10)
}

args <- commandArgs(trailingOnly = TRUE)
if (args[1] == '-h' | args[1] == '--help'){
    print ('Rscript DESeq2_update.R <read count> <label> <output> <group of interest> <background group>')
} else {
    DESeq2(args[1], args[2], args[3], args[4], args[5])
```

Running DESeq2

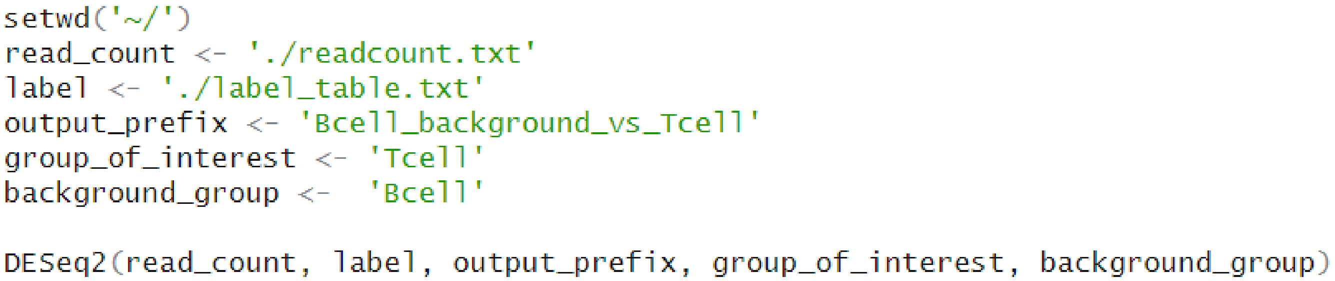
```
$Rscript DESeq2_update.R .readcount.txt label_table.txt
Bcell_background_vs_Tcell Tcell Bcell
```



```
[1] "Running DESeq2"
estimating size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
[1] "DESeq done"
```

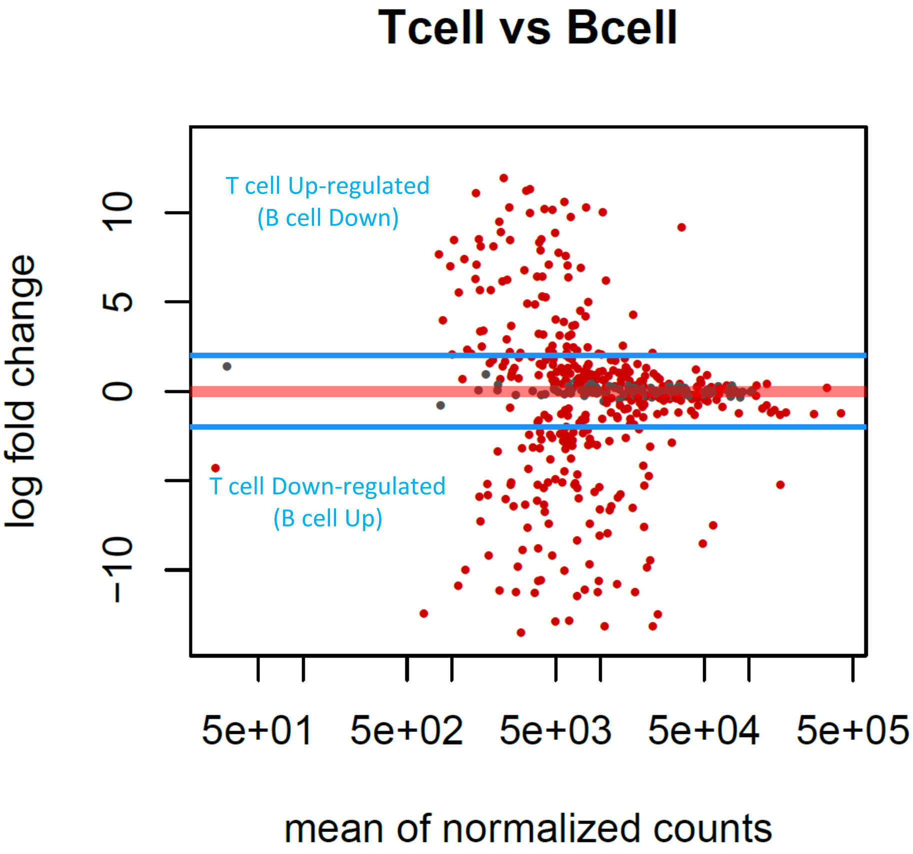


```
-rw-rw-r-- 1 biguser biguser 64380 Dec  8 16:29 Bcell_background_vs_Tcellall.txt all 501 genes
-rw-rw-r-- 1 biguser biguser  8928 Dec  8 16:29 Bcell_background_vs_Tcell.maplot.pdf
-rw-rw-r-- 1 biguser biguser 49492 Dec  8 16:29 Bcell_background_vs_Tcellsig.txt significant DEGs
-rw-rw-r-- 1 biguser biguser 33708 Dec  8 16:29 Bcell_background_vs_Tcell.volcanoplot.pdf
```



```
setwd('~/')
read_count <- './readcount.txt'
label <- './label_table.txt'
output_prefix <- 'Bcell_background_vs_Tcell'
group_of_interest <- 'Tcell'
background_group <-  'Bcell'

DESeq2(read_count, label, output_prefix, group_of_interest, background_group)
```

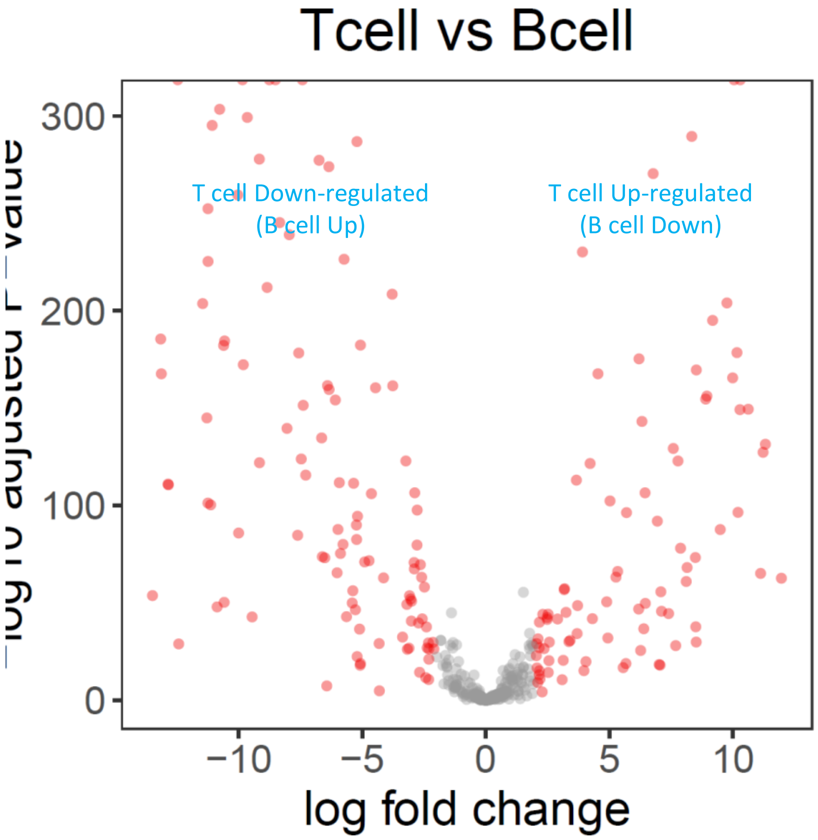
Output plots for DESeq2



Tcell vs Bcell

T cell Up-regulated
(B cell Down)

T cell Down-regulated
(B cell Up)

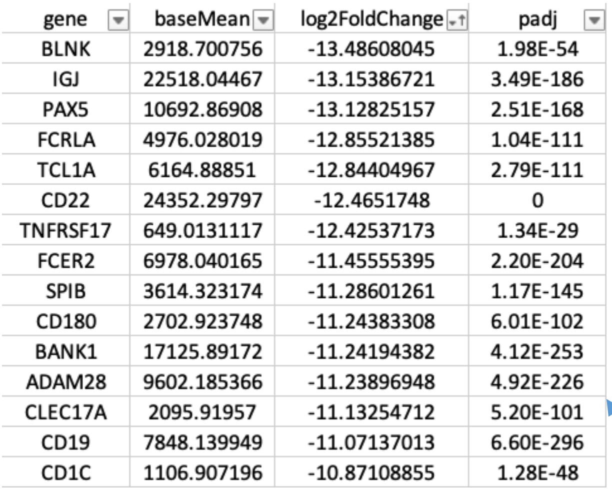


Tcell vs Bcell

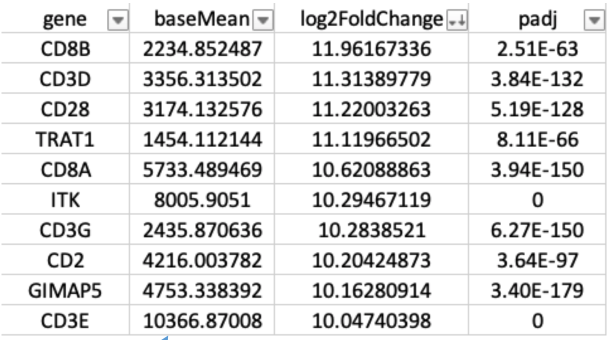
T cell Down-regulated
(B cell Up)

T cell Up-regulated
(B cell Down)

Output files for DESeq2

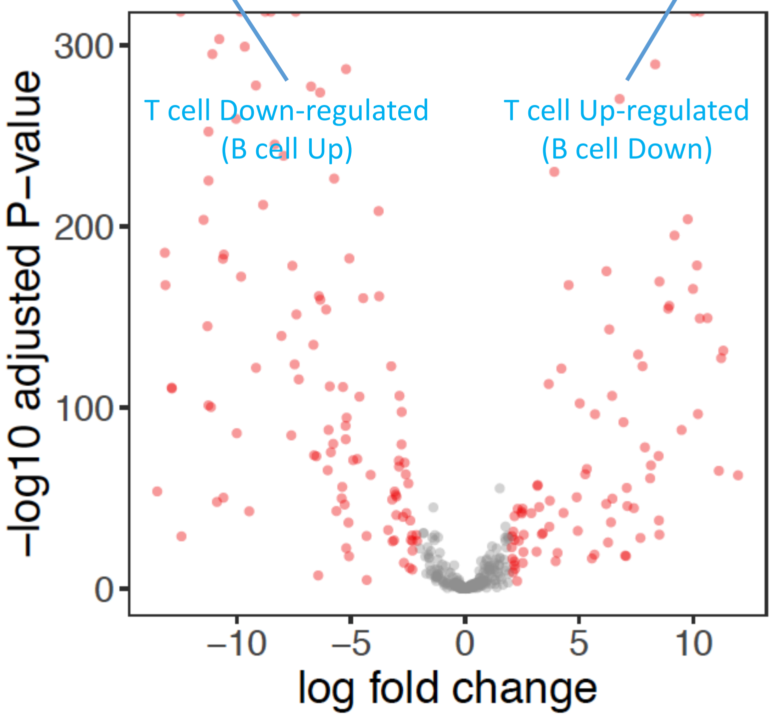


gene	baseMean	log2FoldChange	padj
BLNK	2918.700756	-13.48608045	1.98E-54
IGJ	22518.04467	-13.15386721	3.49E-186
PAX5	10692.86908	-13.12825157	2.51E-168
FCRLA	4976.028019	-12.85521385	1.04E-111
TCL1A	6164.88851	-12.84404967	2.79E-111
CD22	24352.29797	-12.4651748	0
TNFRSF17	649.0131117	-12.42537173	1.34E-29
FCER2	6978.040165	-11.45555395	2.20E-204
SPIB	3614.323174	-11.28601261	1.17E-145
CD180	2702.923748	-11.24383308	6.01E-102
BANK1	17125.89172	-11.24194382	4.12E-253
ADAM28	9602.185366	-11.23896948	4.92E-226
CLEC17A	2095.91957	-11.13254712	5.20E-101
CD19	7848.139949	-11.07137013	6.60E-296
CD1C	1106.907196	-10.87108855	1.28E-48



gene	baseMean	log2FoldChange	padj
CD8B	2234.852487	11.96167336	2.51E-63
CD3D	3356.313502	11.31389779	3.84E-132
CD28	3174.132576	11.22003263	5.19E-128
TRAT1	1454.112144	11.11966502	8.11E-66
CD8A	5733.489469	10.62088863	3.94E-150
ITK	8005.9051	10.29467119	0
CD3G	2435.870636	10.2838521	6.27E-150
CD2	4216.003782	10.20424873	3.64E-97
GIMAP5	4753.338392	10.16280914	3.40E-179
CD3E	10366.87008	10.04740398	0

Tcell vs Bcell

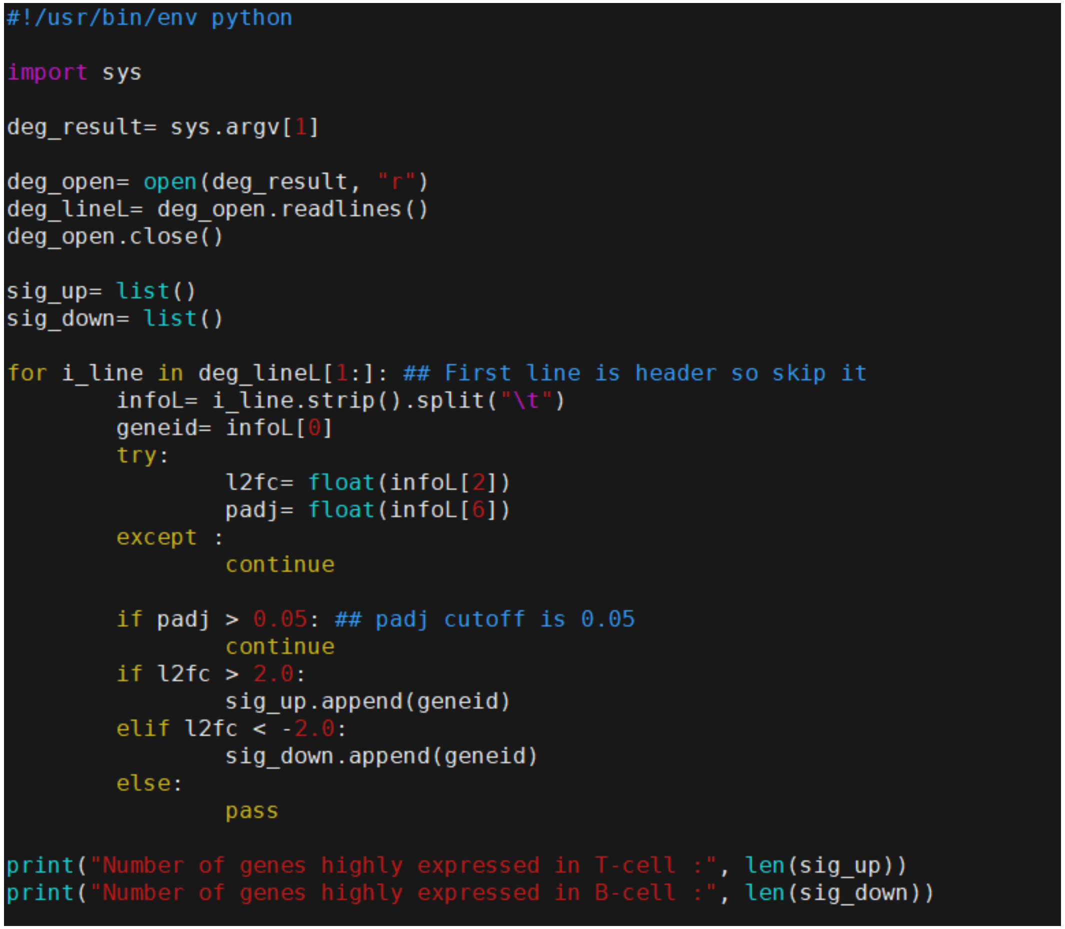


T cell Down-regulated
(B cell Up)

T cell Up-regulated
(B cell Down)

Python script to extract significant DEGs

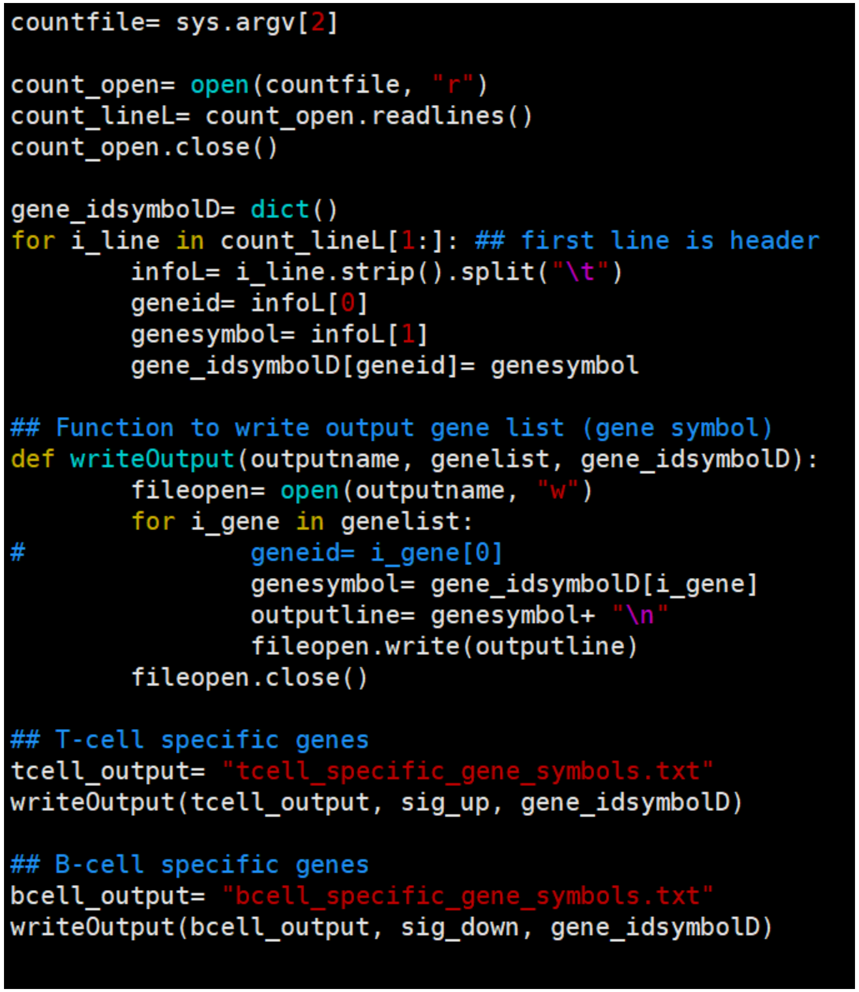
```
$ python getSigGenes.py
Bcell_background_vs_Tcellall.txt
readcount.txt
```



```python
#!/usr/bin/env python

import sys

deg_result= sys.argv[1]

deg_open= open(deg_result, "r")
deg_lineL= deg_open.readlines()
deg_open.close()

sig_up= list()
sig_down= list()

for i_line in deg_lineL[1:]: ## First line is header so skip it
        infoL= i_line.strip().split("\t")
        geneid= infoL[0]
        try:
                l2fc= float(infoL[2])
                padj= float(infoL[6])
        except :
                continue

        if padj > 0.05: ## padj cutoff is 0.05
                continue
        if l2fc > 2.0:
                sig_up.append(geneid)
        elif l2fc < -2.0:
                sig_down.append(geneid)
        else:
                pass

print("Number of genes highly expressed in T-cell :", len(sig_up))
print("Number of genes highly expressed in B-cell :", len(sig_down))
```

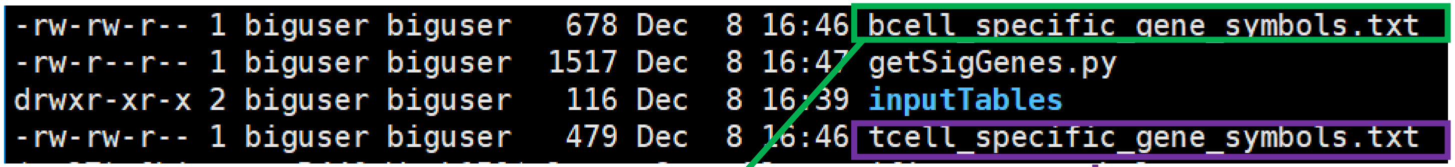
```
Number of genes highly expressed in T-cell : 85
Number of genes highly expressed in B-cell : 106
```

Python script to extract significant DEGs

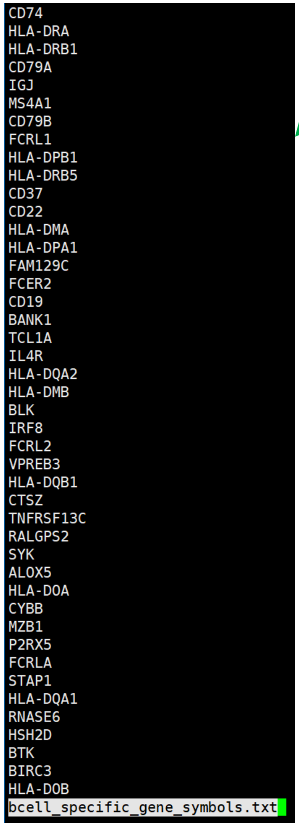


```
countfile= sys.argv[2]

count_open= open(countfile, "r")
count_lineL= count_open.readlines()
count_open.close()

gene_idsymbolD= dict()
for i_line in count_lineL[1:]: ## first line is header
        infoL= i_line.strip().split("\t")
        geneid= infoL[0]
        genesymbol= infoL[1]
        gene_idsymbolD[geneid]= genesymbol

## Function to write output gene list (gene symbol)
def writeOutput(outputname, genelist, gene_idsymbolD):
        fileopen= open(outputname, "w")
        for i_gene in genelist:
#               geneid= i_gene[0]
                genesymbol= gene_idsymbolD[i_gene]
                outputline= genesymbol+ "\n"
                fileopen.write(outputline)
        fileopen.close()

## T-cell specific genes
tcell_output= "tcell_specific_gene_symbols.txt"
writeOutput(tcell_output, sig_up, gene_idsymbolD)

## B-cell specific genes
bcell_output= "bcell_specific_gene_symbols.txt"
writeOutput(bcell_output, sig_down, gene_idsymbolD)
```

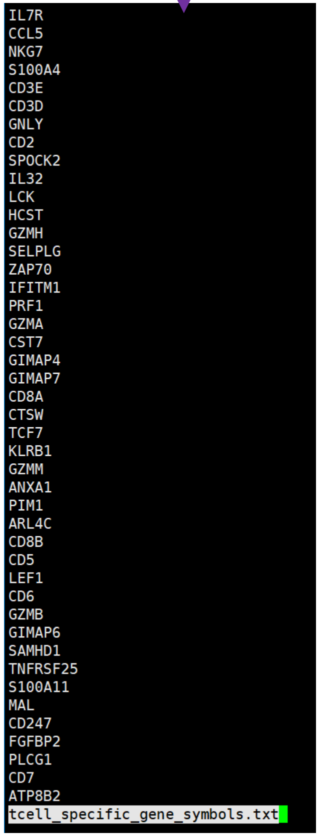
List of genes (symbols) that are significantly, differentially expressed



```
-rw-rw-r-- 1 biguser biguser   678 Dec  8 16:46 bcell_specific_gene_symbols.txt
-rw-r--r-- 1 biguser biguser  1517 Dec  8 16:47 getSigGenes.py
drwxr-xr-x 2 biguser biguser   116 Dec  8 16:39 inputTables
-rw-rw-r-- 1 biguser biguser   479 Dec  8 16:46 tcell_specific_gene_symbols.txt
```

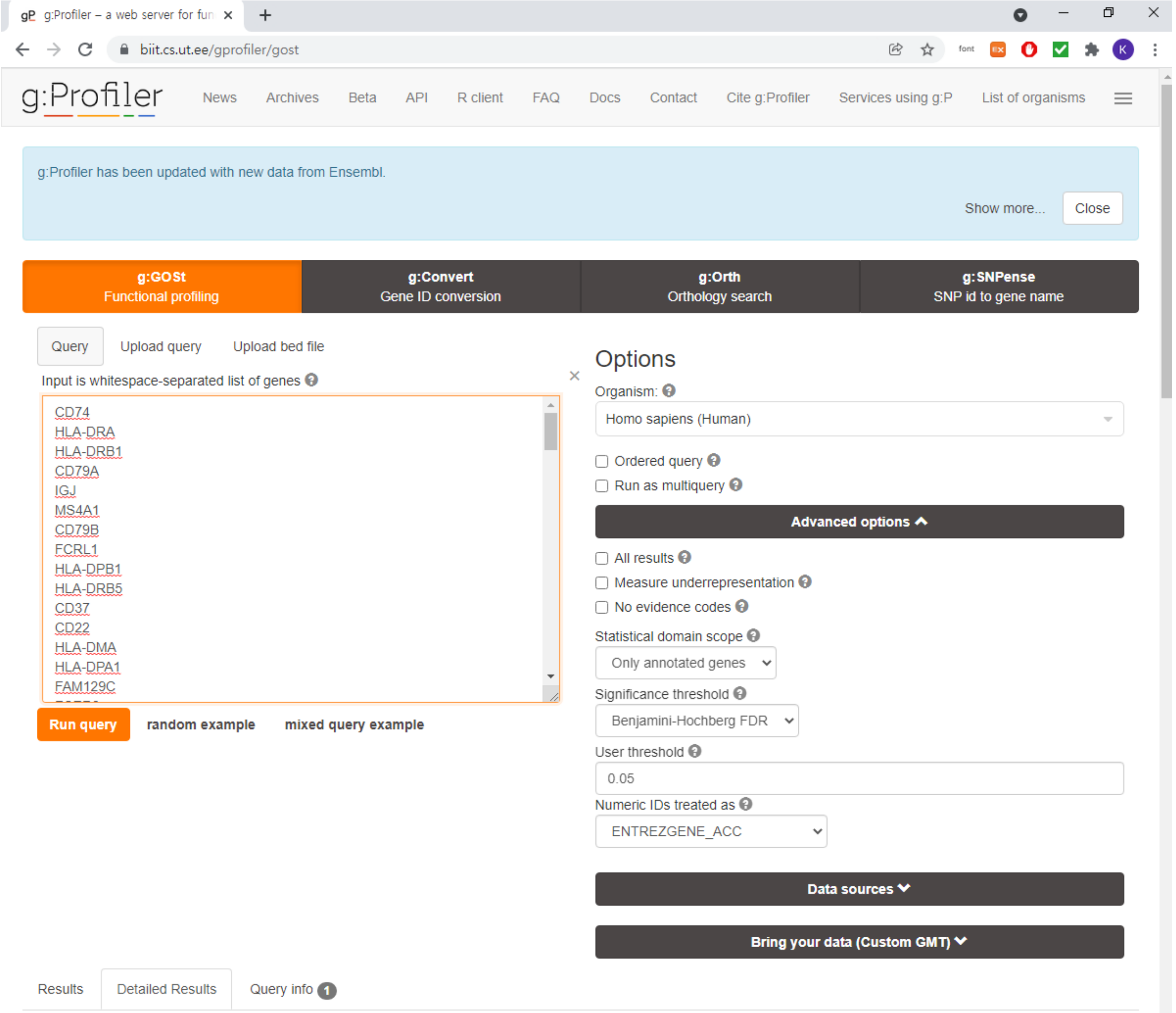


```
CD74
HLA-DRA
HLA-DRB1
CD79A
IGJ
MS4A1
CD79B
FCRL1
HLA-DPB1
HLA-DRB5
CD37
CD22
HLA-DMA
HLA-DPA1
FAM129C
FCER2
CD19
BANK1
TCL1A
IL4R
HLA-DQA2
HLA-DMB
BLK
IRF8
FCRL2
VPREB3
HLA-DQB1
CTSZ
TNFRSF13C
RALGPS2
SYK
ALOX5
HLA-DOA
CYBB
MZB1
P2RX5
FCRLA
STAP1
HLA-DQA1
RNASE6
HSH2D
BTK
BIRC3
HLA-DOB
bcell_specific_gene_symbols.txt
```



```
IL7R
CCL5
NKG7
S100A4
CD3E
CD3D
GNLY
CD2
SPOCK2
IL32
LCK
HCST
GZMH
SELPLG
ZAP70
IFITM1
PRF1
GZMA
CST7
GIMAP4
GIMAP7
CD8A
CTSW
TCF7
KLRB1
GZMM
ANXA1
PIM1
ARL4C
CD8B
CD5
LEF1
CD6
GZMB
GIMAP6
SAMHD1
TNFRSF25
S100A11
MAL
CD247
FGFBP2
PLCG1
CD7
ATP8B2
tcell_specific_gene_symbols.txt
```

Biological signatures associated with list of genes



gP g:Profiler – a web server for fun

biit.cs.ut.ee/gprofiler/gost

g:Profiler has been updated with new data from Ensembl.

Input is whitespace-separated list of genes

CD74
HLA-DRA
HLA-DRB1
CD79A
IGJ
MS4A1
CD79B
FCRL1
HLA-DPB1
HLA-DRB5
CD37
CD22
HLA-DMA
HLA-DPA1
FAM129C

Options

Organism: Homo sapiens (Human)

□ Ordered query

□ Run as multiquery

□ All results

□ Measure underrepresentation

□ No evidence codes

Statistical domain scope: Only annotated genes

Significance threshold: Benjamini-Hochberg FDR

User threshold: 0.05

Numeric IDs treated as: ENTREZGENE_ACC