RNA seq mini project

Kelly_F

11/19/2021

The data for today's mini project comes from the following publication:

Trapnell C, Hendrickson DG, Sauvageau M, Goff L et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq". Nat Biotechnol 2013 Jan;31(1):46-53. PMID: 23222703

Workflow:

- Import counts data and metadata
- PCA analysis
- DESEQ analysis
- Volcano plot
- Annotation
- Pathway analysis

```
# Load packages
library(DESeq2)
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
## 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, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

## Loading required package: GenomicRanges

## Loading required package: GenomeInfoDb

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

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

```
##
## Attaching package: ’Biobase’
```

```
## The following object is masked from ’package:MatrixGenerics’:
##
##     rowMedians
```

```
## The following objects are masked from ’package:matrixStats’:
##
##     anyMissing, rowMedians
```

```
library(ggplot2)
library(AnnotationDbi)
#Import metadata and counts table
mdat <- read.csv("./GSE37704_metadata.csv")
head(mdat)
```

```
##          id     condition
## 1 SRR493366 control_sirna
## 2 SRR493367 control_sirna
## 3 SRR493368 control_sirna
## 4 SRR493369      hoxa1_kd
## 5 SRR493370      hoxa1_kd
## 6 SRR493371      hoxa1_kd
```

```
counts <- read.csv("./GSE37704_featurecounts.csv", row.names = 1)
head(counts)
```

```
##                 length SRR493366 SRR493367 SRR493368 SRR493369 SRR493370
## ENSG00000186092    918         0         0         0         0         0
## ENSG00000279928    718         0         0         0         0         0
## ENSG00000279457   1982        23        28        29        29        28
## ENSG00000278566    939         0         0         0         0         0
## ENSG00000273547    939         0         0         0         0         0
## ENSG00000187634   3214       124       123       205       207       212
##                 SRR493371
## ENSG00000186092         0
## ENSG00000279928         0
## ENSG00000279457        46
## ENSG00000278566         0
## ENSG00000273547         0
## ENSG00000187634       258
```

Modify counts table to remove “length” column

```
counts <- as.matrix(counts[,2:7])
head(counts)
```

```
##                 SRR493366 SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
## ENSG00000186092         0         0         0         0         0         0
## ENSG00000279928         0         0         0         0         0         0
## ENSG00000279457        23        28        29        29        28        46
## ENSG00000278566         0         0         0         0         0         0
## ENSG00000273547         0         0         0         0         0         0
## ENSG00000187634       124       123       205       207       212       258
```

Remove zeros from the counts table

```
counts = counts[(rowSums(counts)!=0), ]
head(counts)
```

```
##                 SRR493366 SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
## ENSG00000279457        23        28        29        29        28        46
## ENSG00000187634       124       123       205       207       212       258
## ENSG00000188976      1637      1831      2383      1226      1326      1504
## ENSG00000187961       120       153       180       236       255       357
## ENSG00000187583        24        48        65        44        48        64
## ENSG00000187642         4         9        16        14        16        16
```

```
nrow(counts)
```

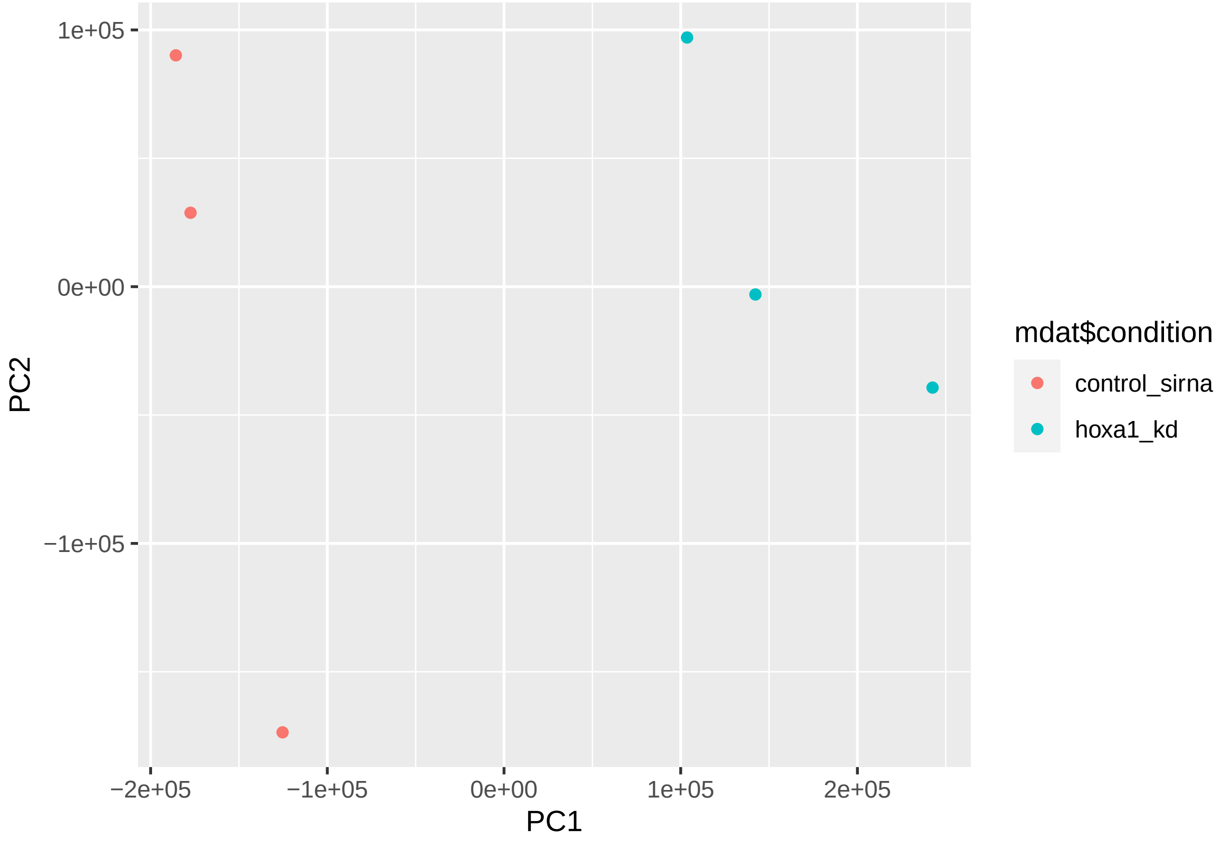
```
## [1] 15975
```

PCA

```
# Run PCA on counts data
pca <- prcomp(t(counts))
summary(pca)
```

```
## Importance of components:
##                              PC1       PC2       PC3       PC4      PC5
## Standard deviation     1.852e+05 1.001e+05 1.998e+04 6.886e+03 5.15e+03
## Proportion of Variance 7.659e-01 2.235e-01 8.920e-03 1.060e-03 5.90e-04
## Cumulative Proportion  7.659e-01 9.894e-01 9.983e-01 9.994e-01 1.00e+00
##                              PC6
## Standard deviation     9.558e-10
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

```r
# Save PCA values for graphing in dataframe
pca.data <- data.frame(pca$x)

# Generate plot
ggplot(pca.data, aes(PC1, PC2, col=mdat$condition)) +
  geom_point()
```



```r
# How to plot using base R:
#plot(pca$x[,1], pca$x[,2], pch=16, col=as.factor(mdat$condition))
```

DESEQ2 Analysis

```r
dds = DESeqDataSetFromMatrix(countData=counts,
                             colData=mdat,
                             design=~condition)
```

```
## Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
## design formula are characters, converting to factors
```

```r
dds = DESeq(dds)
```

```
## estimating size factors
## estimating dispersions
```

```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

```
## fitting model and testing
```

```r
dds
```

```
## class: DESeqDataSet
## dim: 15975 6
## metadata(1): version
## assays(4): counts mu H cooks
## rownames(15975): ENSG00000279457 ENSG00000187634 ... ENSG00000276345
##   ENSG00000271254
## rowData names(22): baseMean baseVar ... deviance maxCooks
## colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
## colData names(3): id condition sizeFactor
```

Store dds results

```r
res <- results(dds, alpha= 0.05)
head(res)
```

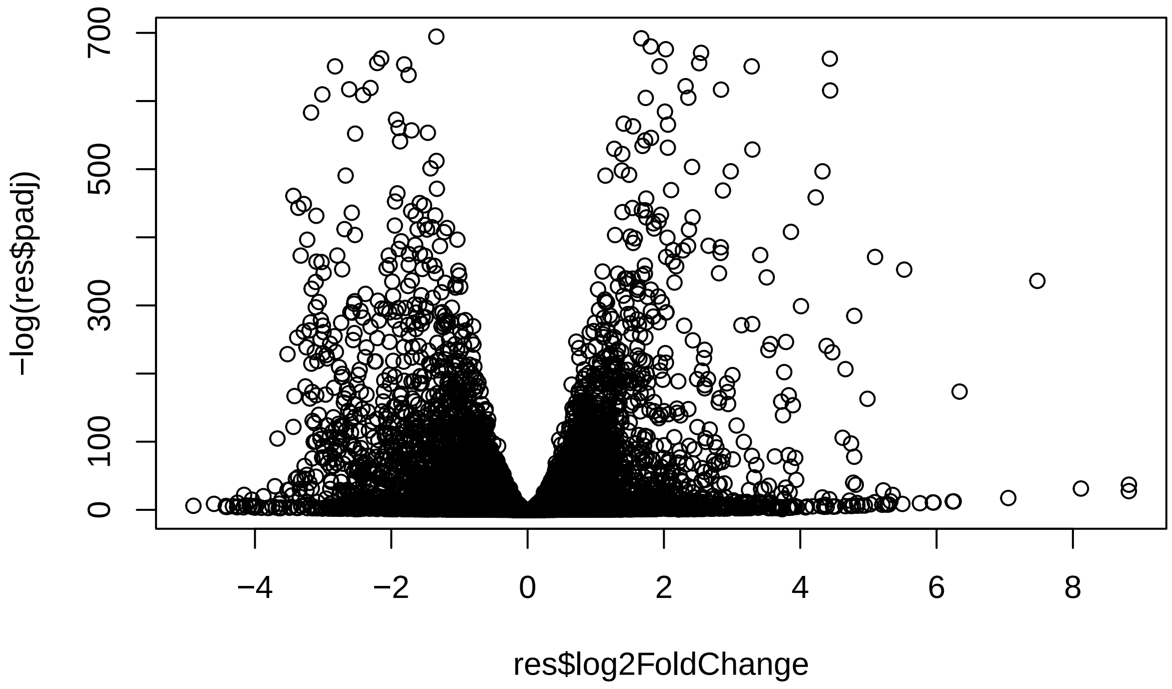
```
## log2 fold change (MLE): condition hoxa1 kd vs control sirna
## Wald test p-value: condition hoxa1 kd vs control sirna
## DataFrame with 6 rows and 6 columns
##                  baseMean log2FoldChange     lfcSE       stat      pvalue
##                 <numeric>      <numeric> <numeric>  <numeric>   <numeric>
## ENSG00000279457   29.9136      0.1792571 0.3248216   0.551863 5.81042e-01
## ENSG00000187634  183.2296      0.4264571 0.1402658   3.040350 2.36304e-03
## ENSG00000188976 1651.1881     -0.6927205 0.0548465 -12.630158 1.43990e-36
## ENSG00000187961  209.6379      0.7297556 0.1318599   5.534326 3.12428e-08
## ENSG00000187583   47.2551      0.0405765 0.2718928   0.149237 8.81366e-01
## ENSG00000187642   11.9798      0.5428105 0.5215598   1.040744 2.97994e-01
##                        padj
##                   <numeric>
## ENSG00000279457 6.73177e-01
## ENSG00000187634 4.93953e-03
## ENSG00000188976 1.69098e-35
## ENSG00000187961 1.08627e-07
## ENSG00000187583 9.14739e-01
## ENSG00000187642 3.90951e-01
```

```
summary(res)
```

```
##
## out of 15975 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 4043, 25%
## LFC < 0 (down)     : 4142, 26%
## outliers [1]       : 0, 0%
## low counts [2]     : 1859, 12%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

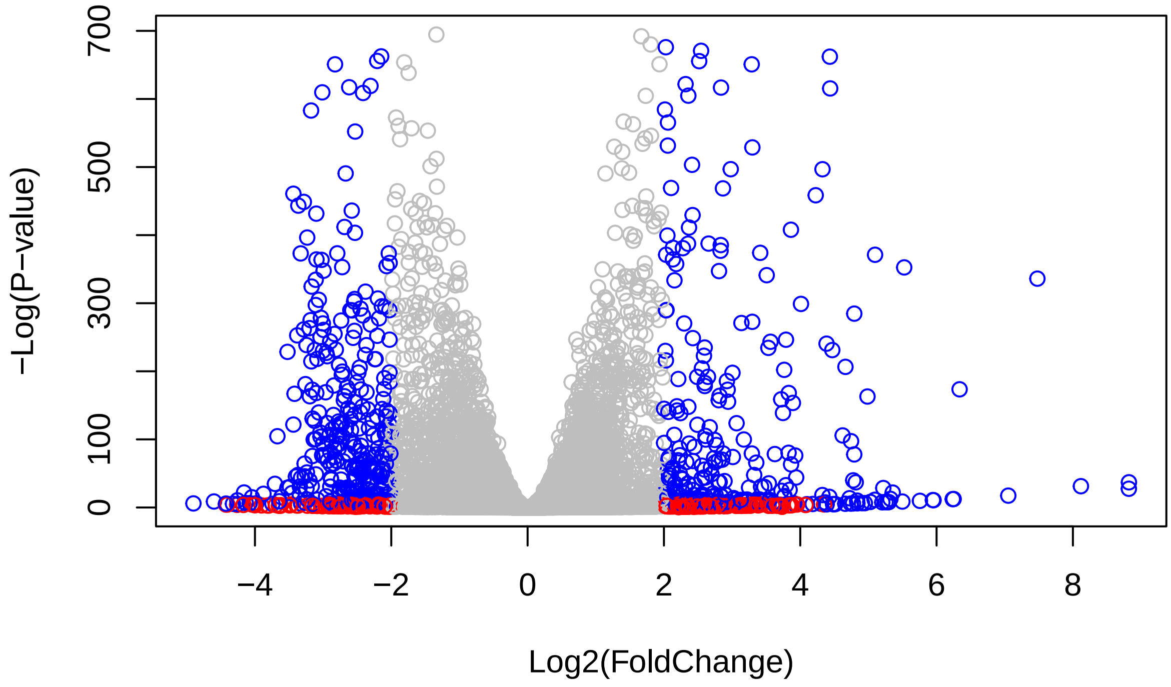
Visualize DESEQ2 results with volcano plot

```
plot(res$log2FoldChange, -log(res$padj) )
```



```
# Make a color vector for all genes
mycols <- rep("gray", nrow(res) )

# Color red the genes with absolute fold change above 2
mycols[ abs(res$log2FoldChange) > 2 ] <- "red"
```

```
# Color blue those with adjusted p-value less than 0.01
#  and absolute fold change more than 2
inds <- (res$padj < 0.01) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "blue"
plot( res$log2FoldChange, -log(res$padj), col=mycols, xlab="Log2(FoldChange)", ylab="-Log(P-value)")
```



Annotation

Add SYMBOL, ENTREZID, and GENENAME annotation to results table for downstream KEGG analysis

```
library("org.Hs.eg.db")
```

```
##
```

```
res$symbol <- mapIds(org.Hs.eg.db, # Annotation package
                     keys=row.names(res), # Our genenames
                     keytype="ENSEMBL",        # The format of our genenames
                     column="SYMBOL",          # The new format we want to add
                     multiVals="first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```r
res$entrez <- mapIds(org.Hs.eg.db, # Annotation package
                     keys=row.names(res), # Our genenames
                     keytype="ENSEMBL",        # The format of our genenames
                     column="ENTREZID",          # The new format we want to add
                     multiVals="first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```r
res$genenames <-mapIds(org.Hs.eg.db, # Annotation package
                     keys=row.names(res), # Our genenames
                     keytype="ENSEMBL",        # The format of our genenames
                     column="GENENAME",          # The new format we want to add
                     multiVals="first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```r
head(res, 10)
```

```
## log2 fold change (MLE): condition hoxa1 kd vs control sirna
## Wald test p-value: condition hoxa1 kd vs control sirna
## DataFrame with 10 rows and 9 columns
##                     baseMean log2FoldChange     lfcSE       stat      pvalue
##                    <numeric>      <numeric> <numeric>  <numeric>   <numeric>
## ENSG00000279457   29.913579      0.1792571 0.3248216   0.551863 5.81042e-01
## ENSG00000187634  183.229650      0.4264571 0.1402658   3.040350 2.36304e-03
## ENSG00000188976 1651.188076     -0.6927205 0.0548465 -12.630158 1.43990e-36
## ENSG00000187961  209.637938      0.7297556 0.1318599   5.534326 3.12428e-08
## ENSG00000187583   47.255123      0.0405765 0.2718928   0.149237 8.81366e-01
## ENSG00000187642   11.979750      0.5428105 0.5215598   1.040744 2.97994e-01
## ENSG00000188290  108.922128      2.0570638 0.1969053  10.446970 1.51282e-25
## ENSG00000187608  350.716868      0.2573837 0.1027266   2.505522 1.22271e-02
## ENSG00000188157 9128.439422      0.3899088 0.0467163   8.346304 7.04321e-17
## ENSG00000237330    0.158192      0.7859552 4.0804729   0.192614 8.47261e-01
##                        padj      symbol      entrez              genenames
##                   <numeric> <character> <character>            <character>
## ENSG00000279457 6.73177e-01      WASH9P   102723897 WAS protein family h..
## ENSG00000187634 4.93953e-03      SAMD11      148398 sterile alpha motif ..
## ENSG00000188976 1.69098e-35       NOC2L       26155 NOC2 like nucleolar ..
## ENSG00000187961 1.08627e-07      KLHL17      339451 kelch like family me..
## ENSG00000187583 9.14739e-01     PLEKHN1       84069 pleckstrin homology ..
## ENSG00000187642 3.90951e-01       PERM1       84808 PPARGC1 and ESRR ind..
## ENSG00000188290 1.25029e-24        HES4       57801 hes family bHLH tran..
## ENSG00000187608 2.27431e-02       ISG15        9636 ISG15 ubiquitin like..
## ENSG00000188157 4.04154e-16        AGRN      375790                  agrin
## ENSG00000237330          NA      RNF223      401934 ring finger protein ..
```

Write results to file

```
# Reorder by adjusted p-value and write to local directroy
res = res[order(res$pvalue),]
write.csv(res, file="./deseq_results.csv")
```

Pathway Ananlysis

```
# Load packages
library(pathview)
```

```
## ##############################################################################
## Pathview is an open source software package distributed under GNU General
## Public License version 3 (GPLv3). Details of GPLv3 is available at
## http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are required to
## formally cite the original Pathview paper (not just mention it) in publications
## or products. For details, do citation("pathview") within R.
##
## The pathview downloads and uses KEGG data. Non-academic uses may require a KEGG
## license agreement (details at http://www.kegg.jp/kegg/legal.html).
## ##############################################################################
```

```
library(gage)
```

```
##
```

```
library(gageData)

data(kegg.sets.hs)
data(sigmet.idx.hs)

# Focus on signaling and metabolic pathways only
kegg.sets.hs = kegg.sets.hs[sigmet.idx.hs]

# Examine the first 3 pathways
head(kegg.sets.hs, 3)
```

```
## $‘hsa00232 Caffeine metabolism‘
## [1] "10"   "1544" "1548" "1549" "1553" "7498" "9"
##
## $‘hsa00983 Drug metabolism - other enzymes‘
##  [1] "10"     "1066"   "10720"  "10941"  "151531" "1548"   "1549"   "1551"
##  [9] "1553"   "1576"   "1577"   "1806"   "1807"   "1890"   "221223" "2990"
## [17] "3251"   "3614"   "3615"   "3704"   "51733"  "54490"  "54575"  "54576"
## [25] "54577"  "54578"  "54579"  "54600"  "54657"  "54658"  "54659"  "54963"
## [33] "574537" "64816"  "7083"   "7084"   "7172"   "7363"   "7364"   "7365"
## [41] "7366"   "7367"   "7371"   "7372"   "7378"   "7498"   "79799"  "83549"
## [49] "8824"   "8833"   "9"      "978"
##
## $‘hsa00230 Purine metabolism‘
##   [1] "100"    "10201"  "10606"  "10621"  "10622"  "10623"  "107"    "10714" 
##   [9] "108"    "10846"  "109"    "111"    "11128"  "11164"  "112"    "113"   
##  [17] "114"    "115"    "122481" "122622" "124583" "132"    "158"    "159"   
##  [25] "1633"   "171568" "1716"   "196883" "203"    "204"    "205"    "221823"
##  [33] "2272"   "22978"  "23649"  "246721" "25885"  "2618"   "26289"  "270"   
##  [41] "271"    "27115"  "272"    "2766"   "2977"   "2982"   "2983"   "2984"  
##  [49] "2986"   "2987"   "29922"  "3000"   "30833"  "30834"  "318"    "3251"  
##  [57] "353"    "3614"   "3615"   "3704"   "377841" "471"    "4830"   "4831"  
##  [65] "4832"   "4833"   "4860"   "4881"   "4882"   "4907"   "50484"  "50940" 
##  [73] "51082"  "51251"  "51292"  "5136"   "5137"   "5138"   "5139"   "5140"  
##  [81] "5141"   "5142"   "5143"   "5144"   "5145"   "5146"   "5147"   "5148"  
##  [89] "5149"   "5150"   "5151"   "5152"   "5153"   "5158"   "5167"   "5169"  
##  [97] "51728"  "5198"   "5236"   "5313"   "5315"   "53343"  "54107"  "5422"  
## [105] "5424"   "5425"   "5426"   "5427"   "5430"   "5431"   "5432"   "5433"  
## [113] "5434"   "5435"   "5436"   "5437"   "5438"   "5439"   "5440"   "5441"  
## [121] "5471"   "548644" "55276"  "5557"   "5558"   "55703"  "55811"  "55821" 
## [129] "5631"   "5634"   "56655"  "56953"  "56985"  "57804"  "58497"  "6240"  
## [137] "6241"   "64425"  "646625" "654364" "661"    "7498"   "8382"   "84172" 
## [145] "84265"  "84284"  "84618"  "8622"   "8654"   "87178"  "8833"   "9060"  
## [153] "9061"   "93034"  "953"    "9533"   "954"    "955"    "956"    "957"   
## [161] "9583"   "9615"
```

```
# Create named vector of fold changes for input to gauge function
foldchanges <-  res$log2FoldChange
names(foldchanges) <- res$entrez
head(foldchanges)
```

```
##      1266     54855      1465     51232      2034      2317 
## -2.422719  3.201955 -2.313738 -2.059631 -1.888019 -1.649792
```

```
# Run Pathview
keggres = gage(foldchanges, gsets=kegg.sets.hs)

# Look at object returned from gauge
attributes(keggres)
```

```
## $names
## [1] "greater" "less"    "stats"
```

```
# Look at the first few down pathways
head(keggres$less)
```

```
##                                          p.geomean stat.mean        p.val
## hsa04110 Cell cycle                   8.995727e-06 -4.378644 8.995727e-06
## hsa03030 DNA replication              9.424076e-05 -3.951803 9.424076e-05
## hsa03013 RNA transport                1.375901e-03 -3.028500 1.375901e-03
## hsa03440 Homologous recombination     3.066756e-03 -2.852899 3.066756e-03
## hsa04114 Oocyte meiosis               3.784520e-03 -2.698128 3.784520e-03
## hsa00010 Glycolysis / Gluconeogenesis 8.961413e-03 -2.405398 8.961413e-03
##                                             q.val set.size         exp1
## hsa04110 Cell cycle                   0.001448312      121 8.995727e-06
## hsa03030 DNA replication              0.007586381       36 9.424076e-05
## hsa03013 RNA transport                0.073840037      144 1.375901e-03
## hsa03440 Homologous recombination     0.121861535       28 3.066756e-03
## hsa04114 Oocyte meiosis               0.121861535      102 3.784520e-03
## hsa00010 Glycolysis / Gluconeogenesis 0.212222694       53 8.961413e-03
```

```
# Look at the first few up pathways
head(keggres$greater)
```

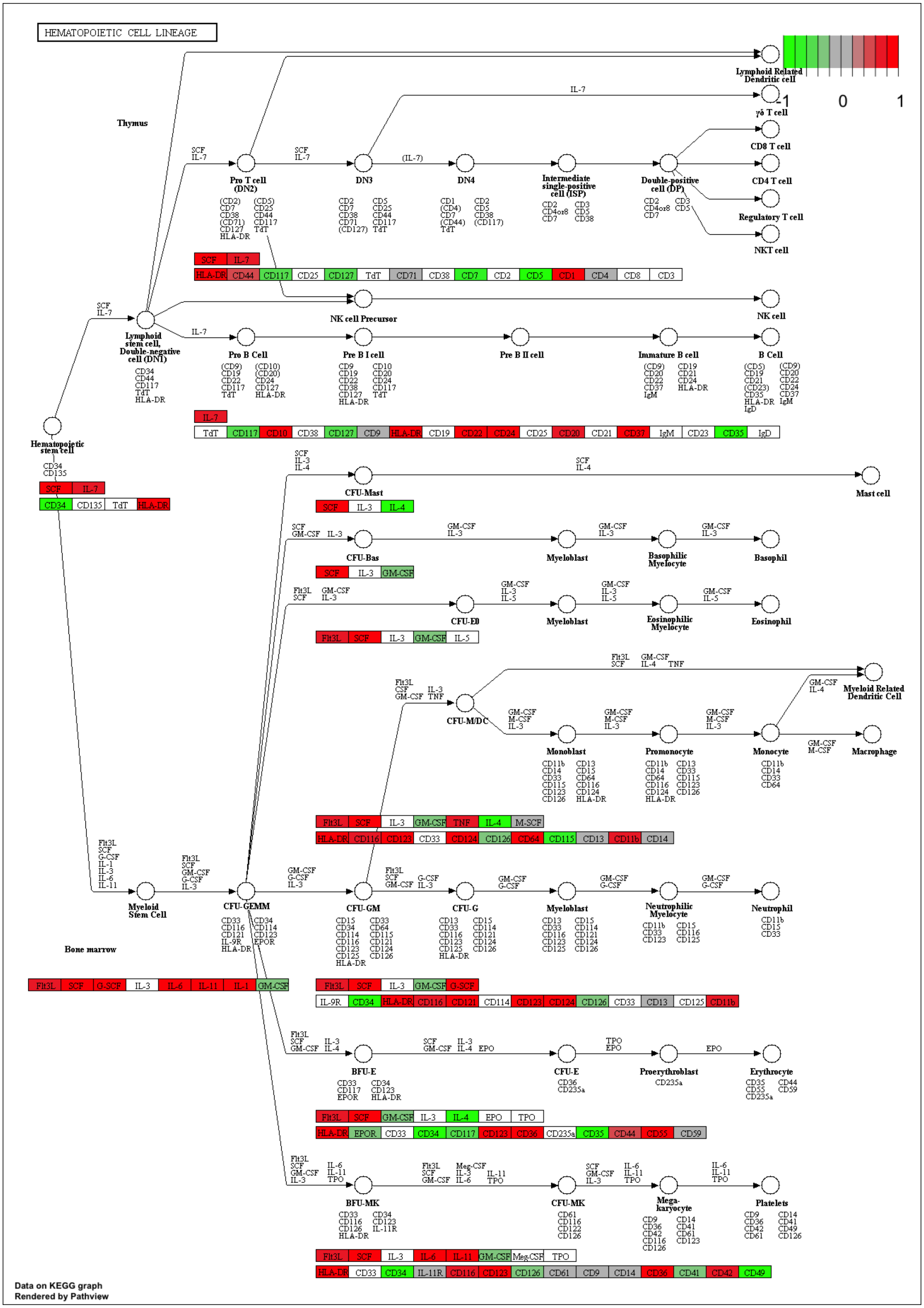
```
##                                          p.geomean stat.mean       p.val
## hsa04640 Hematopoietic cell lineage   0.002822776  2.833362 0.002822776
## hsa04630 Jak-STAT signaling pathway   0.005202070  2.585673 0.005202070
## hsa00140 Steroid hormone biosynthesis 0.007255099  2.526744 0.007255099
## hsa04142 Lysosome                     0.010107392  2.338364 0.010107392
## hsa04330 Notch signaling pathway      0.018747253  2.111725 0.018747253
## hsa04916 Melanogenesis                0.019399766  2.081927 0.019399766
##                                           q.val set.size        exp1
## hsa04640 Hematopoietic cell lineage   0.3893570       55 0.002822776
## hsa04630 Jak-STAT signaling pathway   0.3893570      109 0.005202070
## hsa00140 Steroid hormone biosynthesis 0.3893570       31 0.007255099
## hsa04142 Lysosome                     0.4068225      118 0.010107392
## hsa04330 Notch signaling pathway      0.4391731       46 0.018747253
## hsa04916 Melanogenesis                0.4391731       90 0.019399766
```

```
# Investigate top "up" pathway with pathview()
pathview(gene.data=foldchanges, pathway.id="hsa04640")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory /Volumes/GoogleDrive/My Drive/GitHub/bggn_213/11_19_21_RNAseq_mini_projec
```

```
## Info: Writing image file hsa04640.pathview.png
```

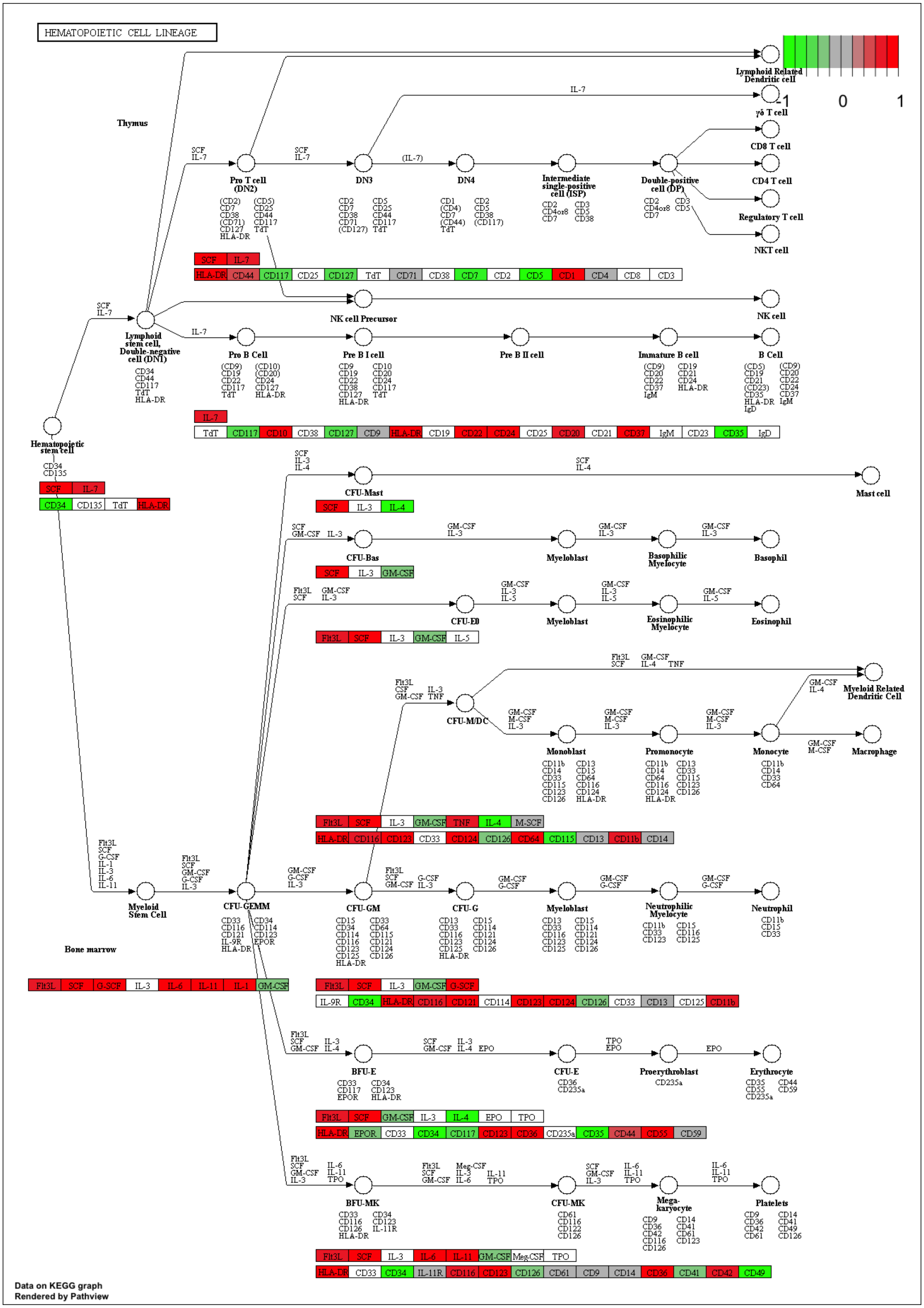

HEMATOPOIETIC CELL LINEAGE
-1
0
1
Thymus
Lymphoid Related Dendritic cell
IL-7
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
SCF IL-7
Pro T cell (DN2)
(CD2) (CD5) CD7 CD25 CD38 CD44 (CD71) CD117 CD127 TdT HLA-DR
DN3
CD2 CD5 CD7 CD25 CD38 CD44 CD71 CD117 (CD127) TdT
(IL-7)
DN4
CD1 CD2 (CD4) CD5 CD7 CD38 (CD44) (CD117) TdT
Intermediate single-positive cell (ISP)
CD2 CD3 CD4or8 CD5 CD7 CD38
Double-positive cell (DP)
CD2 CD3 CD4or8 CD5 CD7
SCF
IL-7
HLA-DR
CD44
CD117
CD25
CD127
TdT
CD71
CD38
CD7
CD2
CD5
CD1
CD4
CD8
CD3
SCF IL-7
NK cell Precursor
NK cell
Lymphoid stem cell, Double-negative cell (DN1)
CD34 CD44 CD117 TdT HLA-DR
IL-7
Pro B Cell
(CD9) (CD10) CD19 (CD20) CD22 CD24 CD117 CD127 TdT HLA-DR
Pre B I cell
CD9 CD10 CD19 CD20 CD22 CD24 CD38 CD117 CD127 TdT HLA-DR
Pre B II cell
Immature B cell
(CD9) CD19 CD20 CD21 CD22 CD24 CD37 HLA-DR IgM
B Cell
(CD5) (CD9) CD19 CD20 CD21 CD22 (CD23) CD24 CD35 CD37 HLA-DR IgM IgD
IL-7
TdT
CD117
CD10
CD38
CD127
CD9
HLA-DR
CD19
CD22
CD24
CD25
CD20
CD21
CD37
IgM
CD23
CD35
IgD
Hematopoietic stem cell
CD34 CD135
SCF
IL-7
CD34
CD135
TdT
HLA-DR
SCF IL-3 IL-4
CFU-Mast
SCF IL-4
Mast cell
SCF
IL-3
IL-4
SCF GM-CSF IL-3
CFU-Bas
GM-CSF IL-3
Myeloblast
GM-CSF IL-3
Basophilic Myelocyte
GM-CSF IL-3
Basophil
SCF
IL-3
GM-CSF
Flt3L SCF GM-CSF IL-3
CFU-E0
GM-CSF IL-3 IL-5
Myeloblast
GM-CSF IL-3 IL-5
Eosinophilic Myelocyte
GM-CSF IL-5
Eosinophil
Flt3L
SCF
IL-3
GM-CSF
IL-5
Flt3L SCF GM-CSF IL-4 TNF
Myeloid Related Dendritic Cell
GM-CSF IL-4
Flt3L CSF GM-CSF IL-3 TNF
CFU-M/DC
GM-CSF M-CSF IL-3
Monoblast
CD11b CD13 CD14 CD15 CD33 CD64 CD115 CD116 CD123 CD124 CD126 HLA-DR
GM-CSF M-CSF IL-3
Promonocyte
CD11b CD13 CD14 CD33 CD64 CD115 CD116 CD123 CD124 CD126 HLA-DR
GM-CSF M-CSF IL-3
Monocyte
CD11b CD14 CD33 CD64
GM-CSF M-CSF
Macrophage
Flt3L
SCF
IL-3
GM-CSF
TNF
IL-4
M-SCF
HLA-DR
CD116
CD123
CD33
CD124
CD126
CD64
CD115
CD13
CD11b
CD14
Flt3L SCF G-CSF IL-1 IL-3 IL-6 IL-11
Myeloid Stem Cell
Flt3L SCF GM-CSF G-CSF IL-3
CFU-GEMM
CD33 CD34 CD116 CD114 CD121 CD123 IL-9R EPOR HLA-DR
Bone marrow
Flt3L
SCF
G-SCF
IL-3
IL-6
IL-11
IL-1
GM-CSF
GM-CSF G-CSF IL-3
CFU-GM
CD15 CD33 CD34 CD64 CD114 CD115 CD116 CD121 CD123 CD124 CD125 CD126 HLA-DR
Flt3L SCF GM-CSF G-CSF IL-3
CFU-G
CD13 CD15 CD33 CD114 CD116 CD121 CD123 CD124 CD125 CD126 HLA-DR
GM-CSF G-CSF
Myeloblast
CD13 CD15 CD33 CD114 CD116 CD121 CD123 CD124 CD125 CD126
GM-CSF G-CSF
Neutrophilic Myelocyte
CD11b CD15 CD33 CD116 CD123 CD125
GM-CSF G-CSF
Neutrophil
CD11b CD15 CD33
Flt3L
SCF
IL-3
GM-CSF
G-SCF
IL-9R
CD34
HLA-DR
CD116
CD121
CD114
CD123
CD124
CD126
CD33
CD13
CD125
CD11b
Flt3L SCF GM-CSF IL-3 IL-4
BFU-E
CD33 CD34 CD117 CD123 EPOR HLA-DR
SCF GM-CSF IL-3 IL-4 EPO
CFU-E
CD36 CD235a
TPO EPO
Proerythroblast
CD235a
EPO
Erythrocyte
CD35 CD44 CD55 CD59 CD235a
Flt3L
SCF
GM-CSF
IL-3
IL-4
EPO
TPO
HLA-DR
EPOR
CD33
CD34
CD117
CD123
CD36
CD235a
CD35
CD44
CD55
CD59
Flt3L SCF GM-CSF IL-3 IL-6 IL-11 TPO
BFU-MK
CD33 CD34 CD116 CD123 CD126 IL-11R HLA-DR
Flt3L SCF GM-CSF Meg-CSF IL-3 IL-6 IL-11 TPO
CFU-MK
CD61 CD116 CD122 CD126
SCF GM-CSF IL-3 IL-6 IL-11 TPO
Mega-karyocyte
CD9 CD14 CD36 CD41 CD42 CD61 CD116 CD123 CD126
IL-6 IL-11 TPO
Platelets
CD9 CD14 CD36 CD41 CD42 CD49 CD61 CD126
Flt3L
SCF
IL-3
IL-6
IL-11
GM-CSF
Meg-CSF
TPO
HLA-DR
CD33
CD34
IL-11R
CD116
CD123
CD126
CD61
CD9
CD14
CD36
CD41
CD42
CD49
Data on KEGG graph
Rendered by Pathview

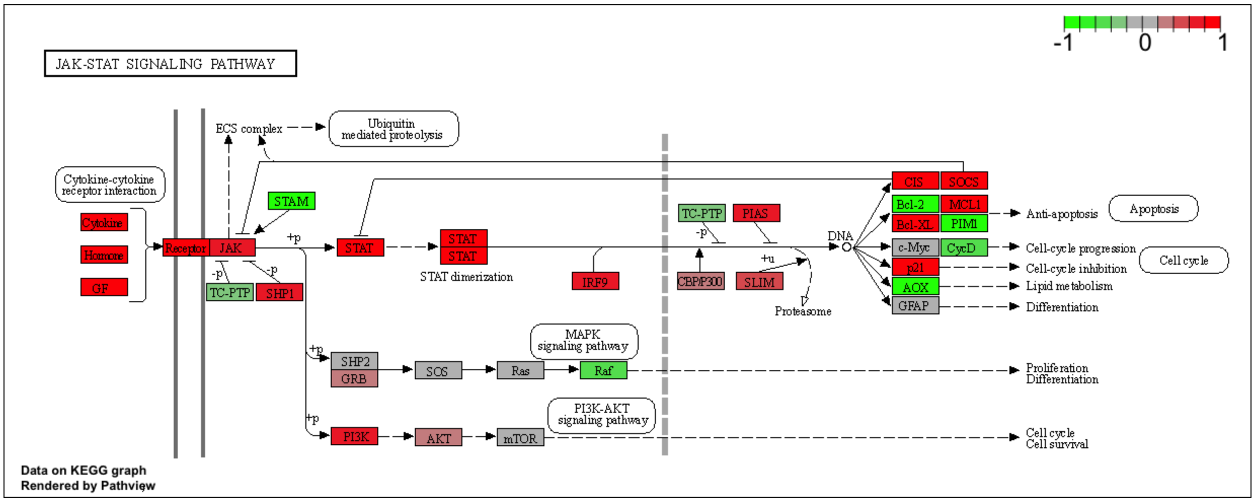
```r
# Generate visualization with the top 5 upregulated pathways
keggrespathways <- rownames(keggres$greater)[1:5]

# Extract the 8 character long IDs part of each string
keggresids = substr(keggrespathways, start=1, stop=8)
keggresids
```

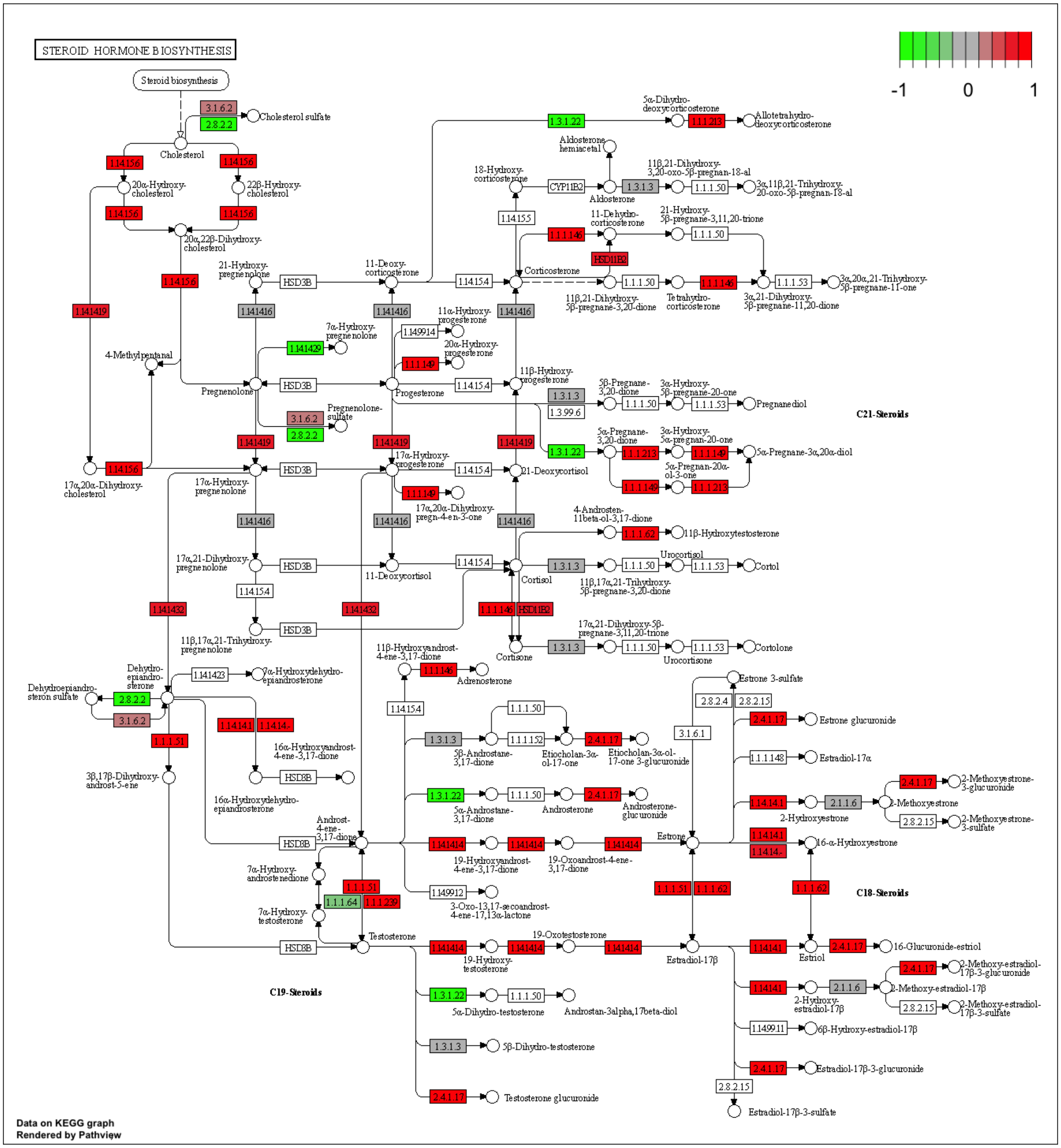
```
## [1] "hsa04640" "hsa04630" "hsa00140" "hsa04142" "hsa04330"
```

```r
# Draw plots
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

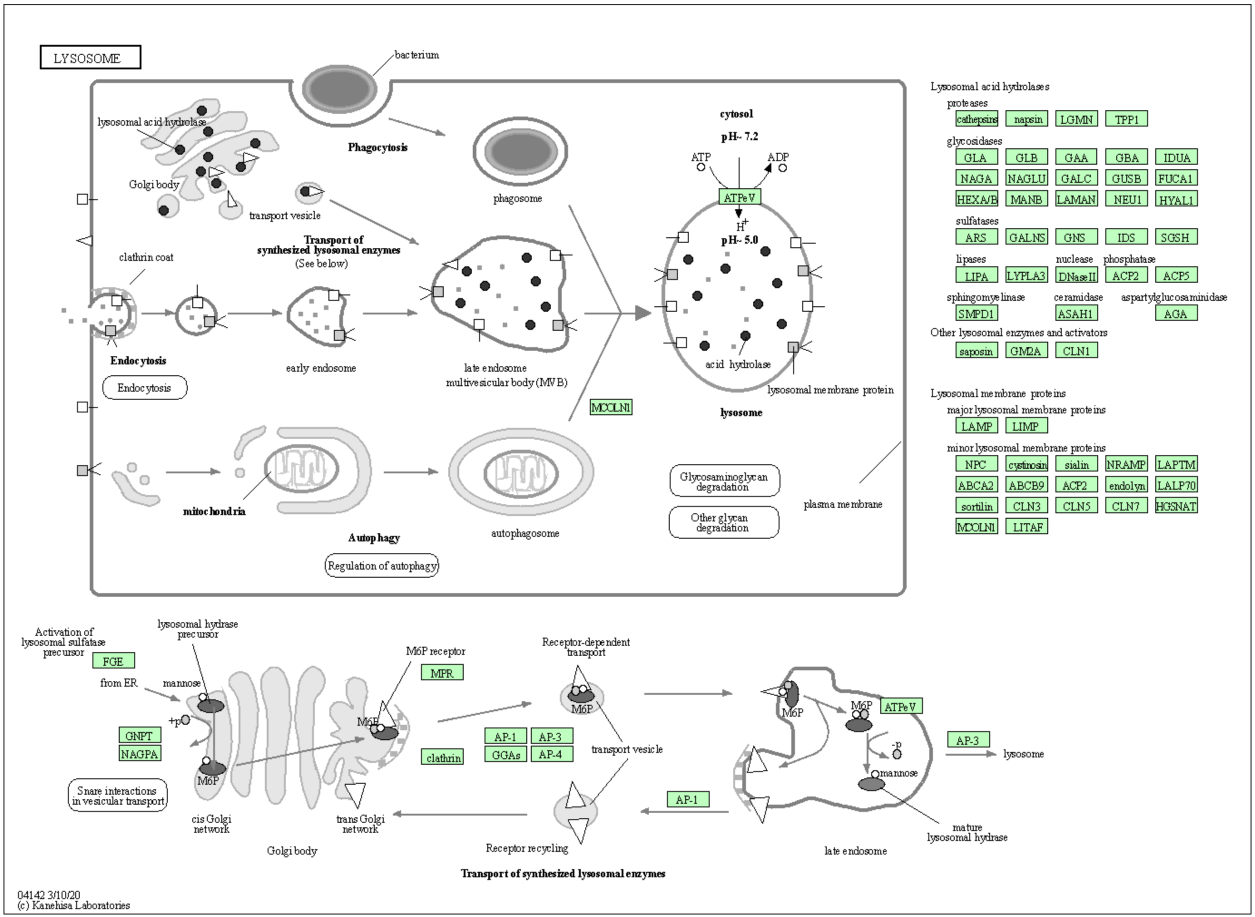
```
## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Volumes/GoogleDrive/My Drive/GitHub/bggn_213/11_19_21_RNAseq_mini_projec

## Info: Writing image file hsa04640.pathview.png

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Volumes/GoogleDrive/My Drive/GitHub/bggn_213/11_19_21_RNAseq_mini_projec

## Info: Writing image file hsa04630.pathview.png

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Volumes/GoogleDrive/My Drive/GitHub/bggn_213/11_19_21_RNAseq_mini_projec

## Info: Writing image file hsa00140.pathview.png

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Volumes/GoogleDrive/My Drive/GitHub/bggn_213/11_19_21_RNAseq_mini_projec

## Info: Writing image file hsa04142.pathview.png

## Info: some node width is different from others, and hence adjusted!

## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Volumes/GoogleDrive/My Drive/GitHub/bggn_213/11_19_21_RNAseq_mini_projec

## Info: Writing image file hsa04330.pathview.png
```

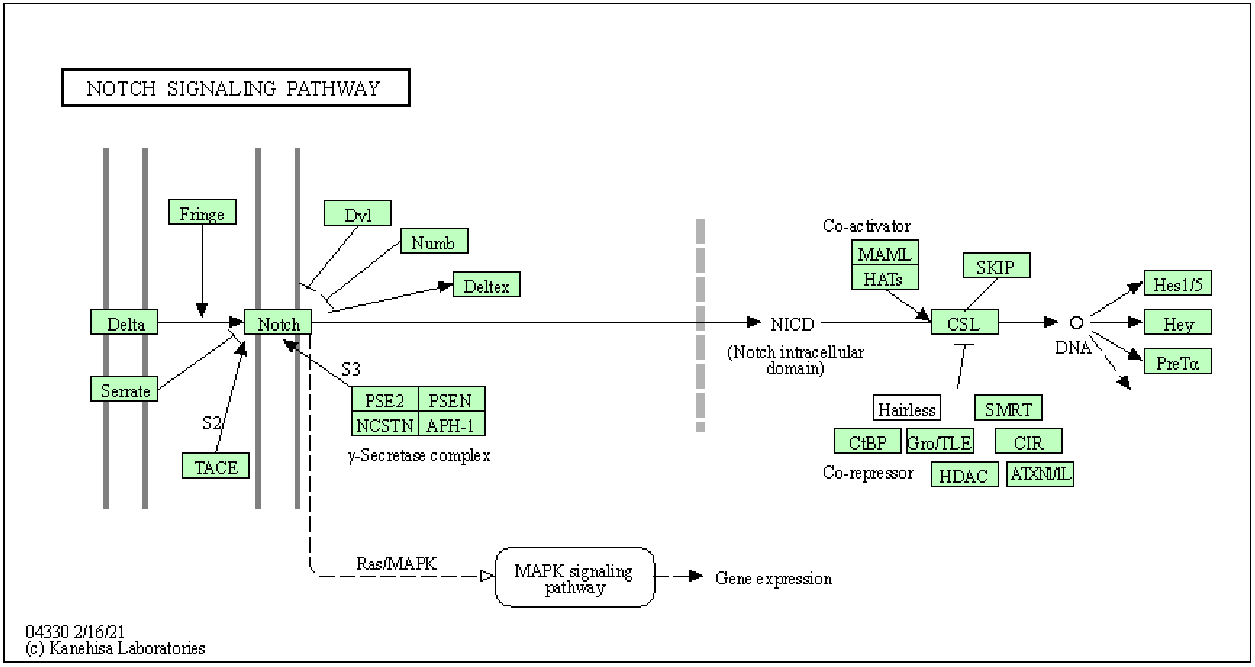

HEMATOPOIETIC CELL LINEAGE
-1
0
1
Thymus
Bone marrow
Hematopoietic stem cell
CD34 CD135
SCF IL-7
CD34 CD135 TdT HLA-DR
Lymphoid stem cell, Double-negative cell (DN1)
CD34 CD44 CD117 TdT HLA-DR
Pro T cell (DN2)
(CD2) CD7 CD38 (CD71) CD127 HLA-DR (CD5) CD25 CD44 CD117 TdT
DN3
CD2 CD7 CD38 CD71 (CD127) CD5 CD25 CD44 CD117 TdT
DN4
CD1 (CD4) CD7 (CD44) TdT CD2 CD5 CD38 (CD117)
Intermediate single-positive cell (ISP)
CD2 CD4or8 CD7 CD3 CD5 CD38
Double-positive cell (DP)
CD2 CD4or8 CD7 CD3 CD5
Lymphoid Related Dendritic cell
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
IL-7
(IL-7)
SCF IL-7
SCF IL-7 HLA-DR CD44 CD117 CD25 CD127 TdT CD71 CD38 CD7 CD2 CD5 CD1 CD4 CD8 CD3
NK cell Precursor
NK cell
Pro B Cell
(CD9) CD19 CD22 CD117 TdT (CD10) (CD20) CD24 CD127 HLA-DR
Pre B I cell
CD9 CD19 CD22 CD38 CD127 HLA-DR CD10 CD20 CD24 CD117 TdT
Pre B II cell
Immature B cell
(CD9) CD20 CD22 CD37 IgM CD19 CD21 CD24 HLA-DR
B Cell
(CD5) CD19 CD21 (CD23) CD35 HLA-DR IgD (CD9) CD20 CD22 CD24 CD37 IgM
IL-7
TdT CD117 CD10 CD38 CD127 CD9 HLA-DR CD19 CD22 CD24 CD25 CD20 CD21 CD37 IgM CD23 CD35 IgD
CFU-Mast
Mast cell
SCF IL-3 IL-4
SCF IL-4
SCF IL-3 IL-4
CFU-Bas
Myeloblast
Basophilic Myelocyte
Basophil
SCF GM-CSF IL-3
GM-CSF IL-3
SCF IL-3 GM-CSF
CFU-E0
Eosinophilic Myelocyte
Eosinophil
Flt3L SCF GM-CSF IL-3
GM-CSF IL-3 IL-5
GM-CSF IL-5
Flt3L SCF IL-3 GM-CSF IL-5
CFU-M/DC
Flt3L CSF GM-CSF IL-3 TNF
Flt3L SCF GM-CSF IL-4 TNF
Myeloid Related Dendritic Cell
Monoblast
CD11b CD14 CD33 CD115 CD123 CD126 CD13 CD15 CD64 CD116 CD124 HLA-DR
Promonocyte
CD11b CD14 CD64 CD116 CD124 HLA-DR CD13 CD33 CD115 CD123 CD126
Monocyte
CD11b CD14 CD33 CD64
Macrophage
GM-CSF M-CSF IL-3
GM-CSF IL-4
GM-CSF M-CSF
Flt3L SCF IL-3 GM-CSF TNF IL-4 M-SCF
HLA-DR CD116 CD123 CD33 CD124 CD126 CD64 CD115 CD13 CD11b CD14
Myeloid Stem Cell
Flt3L SCF G-CSF IL-1 IL-3 IL-6 IL-11
Flt3L SCF GM-CSF G-CSF IL-3
CFU-GEMM
CD33 CD116 CD121 IL-9R HLA-DR CD34 CD114 CD123 EPOR
Flt3L SCF G-SCF IL-3 IL-6 IL-11 IL-1 GM-CSF
CFU-GM
CD15 CD34 CD114 CD116 CD123 CD125 HLA-DR CD33 CD64 CD115 CD121 CD124 CD126
CFU-G
CD13 CD33 CD116 CD123 CD125 HLA-DR CD15 CD114 CD121 CD124 CD126
Myeloblast
CD13 CD33 CD116 CD123 CD125 CD15 CD114 CD121 CD124 CD126
Neutrophilic Myelocyte
CD11b CD33 CD123 CD15 CD116 CD125
Neutrophil
CD11b CD15 CD33
GM-CSF G-CSF IL-3
Flt3L SCF GM-CSF G-CSF IL-3
GM-CSF G-CSF
Flt3L SCF IL-3 GM-CSF G-SCF
IL-9R CD34 HLA-DR CD116 CD121 CD114 CD123 CD124 CD126 CD33 CD13 CD125 CD11b
BFU-E
CD33 CD117 EPOR CD34 CD123 HLA-DR
CFU-E
CD36 CD235a
Proerythroblast
CD235a
Erythrocyte
CD35 CD55 CD235a CD44 CD59
Flt3L SCF GM-CSF IL-3 IL-4
SCF GM-CSF IL-3 IL-4 EPO
TPO EPO
EPO
Flt3L SCF GM-CSF IL-3 IL-4 EPO TPO
HLA-DR EPOR CD33 CD34 CD117 CD123 CD36 CD235a CD35 CD44 CD55 CD59
BFU-MK
CD33 CD116 CD126 HLA-DR CD34 CD123 IL-11R
CFU-MK
CD61 CD116 CD122 CD126
Mega-karyocyte
CD9 CD36 CD42 CD116 CD126 CD14 CD41 CD61 CD123
Platelets
CD9 CD36 CD42 CD61 CD14 CD41 CD49 CD126
Flt3L SCF GM-CSF IL-3 IL-6 IL-11 TPO
Flt3L SCF GM-CSF Meg-CSF IL-3 IL-6 IL-11 TPO
SCF GM-CSF IL-3 IL-6 IL-11 TPO
IL-6 IL-11 TPO
Flt3L SCF IL-3 IL-6 IL-11 GM-CSF Meg-CSF TPO
HLA-DR CD33 CD34 IL-11R CD116 CD123 CD126 CD61 CD9 CD14 CD36 CD41 CD42 CD49
Data on KEGG graph
Rendered by Pathview


-1
0
1
JAK-STAT SIGNALING PATHWAY
ECS complex
Ubiquitin mediated proteolysis
Cytokine-cytokine receptor interaction
Cytokine
Hormone
GF
Receptor
JAK
STAM
+p
STAT
STAT
STAT
STAT dimerization
-p
TC-PTP
-p
SHP1
IRF9
TC-PTP
-p
PIAS
CBP/P300
SLIM
+u
Proteasome
DNA
CIS
SOCS
Bcl-2
MCL1
Bcl-XL
PIM1
c-Myc
CycD
p21
AOX
GFAP
Anti-apoptosis
Apoptosis
Cell-cycle progression
Cell-cycle inhibition
Cell cycle
Lipid metabolism
Differentiation
+p
SHP2
GRB
SOS
Ras
Raf
MAPK signaling pathway
Proliferation
Differentiation
+p
PI3K
AKT
mTOR
PI3K-AKT signaling pathway
Cell cycle
Cell survival
Data on KEGG graph
Rendered by Pathview



STEROID HORMONE BIOSYNTHESIS
-1
0
1
Steroid biosynthesis
3.1.6.2
2.8.2.2
Cholesterol sulfate
Cholesterol
1.14.15.6
20α-Hydroxy-cholesterol
22β-Hydroxy-cholesterol
20α,22β-Dihydroxy-cholesterol
1.14.14.19
4-Methylpentanal
Pregnenolone
21-Hydroxy-pregnenolone
HSD3B
11-Deoxy-corticosterone
1.14.14.16
1.14.14.29
7α-Hydroxy-pregnenolone
Pregnenolone-sulfate
17α,20α-Dihydroxy-cholesterol
17α-Hydroxy-pregnenolone
17α,21-Dihydroxy-pregnenolone
11β,17α,21-Trihydroxy-pregnenolone
1.14.14.32
1.14.15.4
Progesterone
11α-Hydroxy-progesterone
1.14.99.14
20α-Hydroxy-progesterone
1.1.1.149
17α-Hydroxy-progesterone
17α,20α-Dihydroxy-pregn-4-en-3-one
11-Deoxycortisol
11β-Hydroxy-progesterone
Corticosterone
18-Hydroxy-corticosterone
CYP11B2
Aldosterone
Aldosterone hemiacetal
1.14.15.5
11-Dehydro-corticosterone
HSD11B2
1.1.1.146
11β,21-Dihydroxy-5β-pregnane-3,20-dione
Tetrahydro-corticosterone
3α,21-Dihydroxy-5β-pregnane-11,20-dione
1.1.1.53
3α,20α,21-Trihydroxy-5β-pregnane-11-one
21-Hydroxy-5β-pregnane-3,11,20-trione
1.1.1.50
11β,21-Dihydroxy-3,20-oxo-5β-pregnan-18-al
3α,11β,21-Trihydroxy-20-oxo-5β-pregnan-18-al
1.3.1.3
1.3.1.22
5α-Dihydro-deoxycorticosterone
1.1.1.213
Allotetrahydro-deoxycorticosterone
5β-Pregnane-3,20-dione
3α-Hydroxy-5β-pregnane-20-one
Pregnanediol
1.3.99.6
5α-Pregnane-3,20-dione
3α-Hydroxy-5α-pregnan-20-one
5α-Pregnane-3α,20α-diol
5α-Pregnan-20α-ol-3-one
C21-Steroids
21-Deoxycortisol
4-Androsten-11beta-ol-3,17-dione
1.1.1.62
11β-Hydroxytestosterone
Cortisol
Urocortisol
Cortol
11β,17α,21-Trihydroxy-5β-pregnane-3,20-dione
Cortisone
17α,21-Dihydroxy-5β-pregnane-3,11,20-trione
Urocortisone
Cortolone
11β-Hydroxyandrost-4-ene-3,17-dione
Adrenosterone
Dehydroepiandrosteron sulfate
Dehydro-epiandro-sterone
1.14.14.23
7α-Hydroxydehydro-epiandrosterone
1.1.1.51
1.14.14.1
1.14.14.-
16α-Hydroxyandrost-4-ene-3,17-dione
3β,17β-Dihydroxy-androst-5-ene
16α-Hydroxydehydro-epiandrosterone
Androst-4-ene-3,17-dione
7α-Hydroxy-androstenedione
7α-Hydroxy-testosterone
1.1.1.64
1.1.1.239
Testosterone
C19-Steroids
5β-Androstane-3,17-dione
1.1.1.152
Etiocholan-3α-ol-17-one
2.4.1.17
Etiocholan-3α-ol-17-one 3-glucuronide
5α-Androstane-3,17-dione
Androsterone
Androsterone-glucuronide
19-Hydroxyandrost-4-ene-3,17-dione
19-Oxoandrost-4-ene-3,17-dione
1.14.99.12
3-Oxo-13,17-secoandrost-4-ene-17,13α-lactone
19-Hydroxy-testosterone
19-Oxotestosterone
5α-Dihydro-testosterone
Androstan-3alpha,17beta-diol
5β-Dihydro-testosterone
Testosterone glucuronide
Estrone 3-sulfate
2.8.2.4
2.8.2.15
3.1.6.1
Estrone glucuronide
1.1.1.148
Estradiol-17α
Estrone
2-Hydroxyestrone
2.1.1.6
2-Methoxyestrone
2-Methoxyestrone-3-glucuronide
2-Methoxyestrone-3-sulfate
16-α-Hydroxyestrone
C18-Steroids
Estradiol-17β
Estriol
16-Glucuronide-estriol
2-Hydroxy-estradiol-17β
2-Methoxy-estradiol-17β
2-Methoxy-estradiol-17β-3-glucuronide
2-Methoxy-estradiol-17β-3-sulfate
1.14.99.11
6β-Hydroxy-estradiol-17β
Estradiol-17β-3-glucuronide
Estradiol-17β-3-sulfate
Data on KEGG graph
Rendered by Pathview





```
# Can also complete the same steps for the top 5 down regulated pathways
#keggrespathways <- rownames(keggres$less)[1:5]
```

```
# Extract the 8 character long IDs part of each string
#keggresids = substr(keggrespathways, start=1, stop=8)
#keggresids

# Draw plots
#pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

Gene Ontology (GO)

Repeat for gene ontology biological process

```
data(go.sets.hs)
data(go.subs.hs)

# Focus on Biological Process subset of GO
gobpsets = go.sets.hs[go.subs.hs$BP]

gobpres = gage(foldchanges, gsets=gobpsets, same.dir=TRUE)

lapply(gobpres, head)
```

```
## $greater
##                                              p.geomean stat.mean        p.val
## GO:0007156 homophilic cell adhesion       8.519724e-05  3.824205 8.519724e-05
## GO:0002009 morphogenesis of an epithelium 1.396681e-04  3.653886 1.396681e-04
## GO:0048729 tissue morphogenesis           1.432451e-04  3.643242 1.432451e-04
## GO:0007610 behavior                       2.195494e-04  3.530241 2.195494e-04
## GO:0060562 epithelial tube morphogenesis  5.932837e-04  3.261376 5.932837e-04
## GO:0035295 tube development               5.953254e-04  3.253665 5.953254e-04
##                                               q.val set.size         exp1
## GO:0007156 homophilic cell adhesion       0.1951953      113 8.519724e-05
## GO:0002009 morphogenesis of an epithelium 0.1951953      339 1.396681e-04
## GO:0048729 tissue morphogenesis           0.1951953      424 1.432451e-04
## GO:0007610 behavior                       0.2243795      427 2.195494e-04
## GO:0060562 epithelial tube morphogenesis  0.3711390      257 5.932837e-04
## GO:0035295 tube development               0.3711390      391 5.953254e-04
##
## $less
##                                             p.geomean stat.mean        p.val
## GO:0048285 organelle fission             1.536227e-15 -8.063910 1.536227e-15
## GO:0000280 nuclear division              4.286961e-15 -7.939217 4.286961e-15
## GO:0007067 mitosis                       4.286961e-15 -7.939217 4.286961e-15
## GO:0000087 M phase of mitotic cell cycle 1.169934e-14 -7.797496 1.169934e-14
## GO:0007059 chromosome segregation        2.028624e-11 -6.878340 2.028624e-11
## GO:0000236 mitotic prometaphase          1.729553e-10 -6.695966 1.729553e-10
##                                                 q.val set.size         exp1
## GO:0048285 organelle fission             5.841698e-12      376 1.536227e-15
## GO:0000280 nuclear division              5.841698e-12      352 4.286961e-15
## GO:0007067 mitosis                       5.841698e-12      352 4.286961e-15
## GO:0000087 M phase of mitotic cell cycle 1.195672e-11      362 1.169934e-14
## GO:0007059 chromosome segregation        1.658603e-08      142 2.028624e-11
## GO:0000236 mitotic prometaphase          1.178402e-07       84 1.729553e-10
##
## $stats
##                                           stat.mean     exp1
## GO:0007156 homophilic cell adhesion        3.824205 3.824205
## GO:0002009 morphogenesis of an epithelium  3.653886 3.653886
## GO:0048729 tissue morphogenesis            3.643242 3.643242
## GO:0007610 behavior                        3.530241 3.530241
## GO:0060562 epithelial tube morphogenesis   3.261376 3.261376
## GO:0035295 tube development                3.253665 3.253665
```

```
# Could graph gobpres results and create visualizations if desired...
```

Reactome

Conduct over-representation enrichment analysis and pathway-topology analysis with Reactome

```
# Create list of significant genes at alpha=0.05
sig_genes <- res[res$padj <= 0.05 & !is.na(res$padj), "symbol"]
print(paste("Total number of significant genes:", length(sig_genes)))
```

```
## [1] "Total number of significant genes: 8185"
```

```
write.table(sig_genes, file="significant_genes.txt", row.names=FALSE, col.names=FALSE, quote=FALSE)
```

> Q: What pathway has the most significant "Entities p-value"? Do the most significant pathways listed match your previous KEGG results? What factors could cause differences between the two methods?

"Endosomal/Vacuolar pathway". No, it is not the same. Although there is some overlap, the differences between the two methods may result from the database being used to search terms. Also, the goals of the two analyses have different end goals.