Class 14: Pathway Analysis from RNA-Seq Results

Katherine Quach (A18541014)

Table of contents

Background

Our Data for today comes from a HOX gene knock-out study

> 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

The authors report on differential analysis of lung fibroblasts in response to loss of the developmental transcription factor HOXA1.

Data Import

We have 2 key input files: counts and metadata.

```
library(DESeq2)
```

```
Loading required package: S4Vectors
```

```
Loading required package: stats4
```

```
Loading required package: BiocGenerics
```

```
Loading required package: generics
```

```

Attaching package: 'generics'
```

```
The following objects are masked from 'package:base':

    as.difftime, as.factor, as.ordered, intersect, is.element, setdiff,
    setequal, union
```

```

Attaching package: 'BiocGenerics'
```

```
The following objects are masked from 'package:stats':

    IQR, mad, sd, var, xtabs
```

```
The following objects are masked from 'package:base':

    anyDuplicated, aperm, append, as.data.frame, basename, cbind,
    colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
    get, grep, grepl, is.unsorted, lapply, Map, mapply, match, mget,
    order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
    rbind, Reduce, rownames, sapply, saveRDS, table, tapply, unique,
    unsplit, which.max, which.min
```

```

Attaching package: 'S4Vectors'
```

```
The following object is masked from 'package:utils':

    findMatches

The following objects are masked from 'package:base':

    expand.grid, I, unname

Loading required package: IRanges

Loading required package: GenomicRanges

Loading required package: Seqinfo

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
```

```r
metaFile <- "data/GSE37704_metadata.csv"
countFile <- "data/GSE37704_featurecounts.csv"

# Import metadata and take a peek
colData = read.csv("GSE37704_metadata.csv", row.names=1)
head(colData)
```

```
              condition
SRR493366 control_sirna
SRR493367 control_sirna
SRR493368 control_sirna
SRR493369      hoxa1_kd
SRR493370      hoxa1_kd
SRR493371      hoxa1_kd
```

```r
# Import countdata
countData = read.csv("GSE37704_featurecounts.csv", row.names=1)
head(countData)
```

```
                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
```

```
head(countData)
```

```
                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
```

Q. Complete the code below to remove the troublesome first column from countData

We need to remove the first length column from `countdata` to have a 1:1 corespondents with `colData` rows.

```
countData <- countData[,-1]
```

Q. Complete the code below to filter countData to exclude genes (i.e. rows) where we have 0 read count across all samples (i.e. columns).

```
rownames(colData) == colnames(countData)
```

```
[1] TRUE TRUE TRUE TRUE TRUE TRUE
```

Check and Tidy

```
library(DESeq2)
```

Setup for DESeq2

```
dds <- DESeqDataSetFromMatrix(countData = countData,
                              colData = colData,
                              design = ~condition)
```

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

Runing DESeq2

```
dds <- DESeq(dds)
```

```
estimating size factors
```

```
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```
fitting model and testing
```

```
dds
```

```
class: DESeqDataSet
dim: 19808 6
metadata(1): version
assays(4): counts mu H cooks
rownames(19808): ENSG00000186092 ENSG00000279928 ... ENSG00000277475
  ENSG00000268674
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(2): condition sizeFactor
```

```
resultsNames(dds)
```

```
[1] "Intercept"                           "condition_hoxa1_kd_vs_control_sirna"
```

Results

```
res <- results(dds)
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>
ENSG00000186092    0.0000             NA        NA        NA         NA
ENSG00000279928    0.0000             NA        NA        NA         NA
ENSG00000279457   29.9136       0.179257  0.324822  0.551863 0.58104205
ENSG00000278566    0.0000             NA        NA        NA         NA
ENSG00000273547    0.0000             NA        NA        NA         NA
ENSG00000187634  183.2296       0.426457  0.140266  3.040350 0.00236304
                     padj
                <numeric>
ENSG00000186092         NA
ENSG00000279928         NA
ENSG00000279457 0.68707978
ENSG00000278566         NA
ENSG00000273547         NA
ENSG00000187634 0.00516278
```

Q. Call the summary() function on your results to get a sense of how many genes are up or down-regulated at the default 0.1 p-value cutoff.

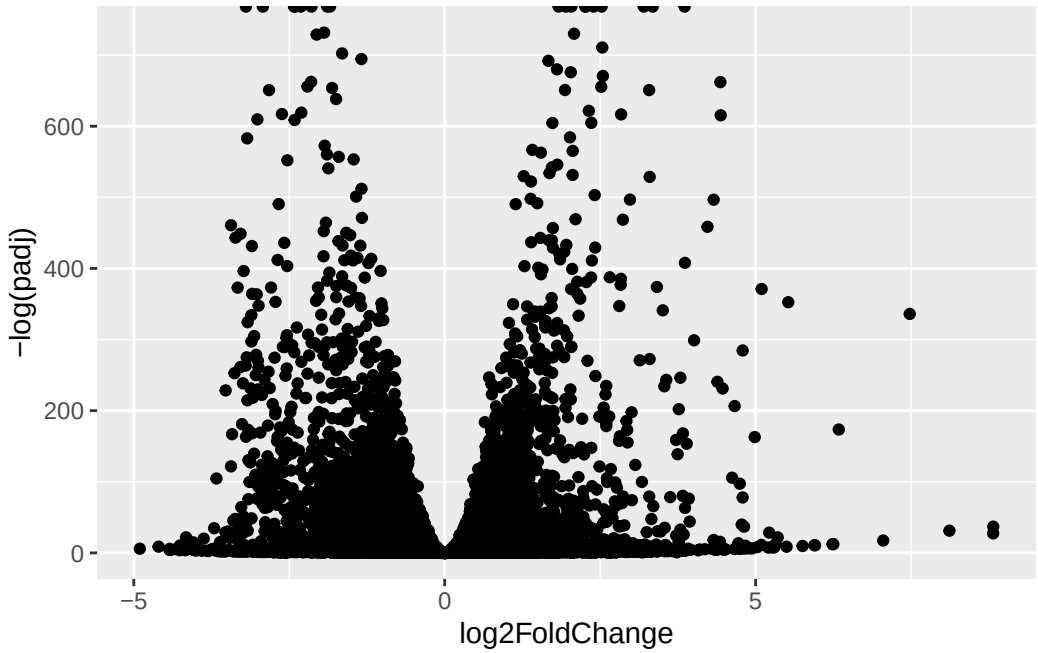
```
summary(res)
```

```
out of 15975 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 4349, 27%
LFC < 0 (down)     : 4393, 27%
outliers [1]       : 0, 0%
low counts [2]     : 1221, 7.6%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Volcano plot

```
library(ggplot2)

        ggplot(res) +
        aes(log2FoldChange,
        -log(padj)) +
        geom_point()
```

```
Warning: Removed 5054 rows containing missing values or values outside the scale range
(`geom_point()`).
```



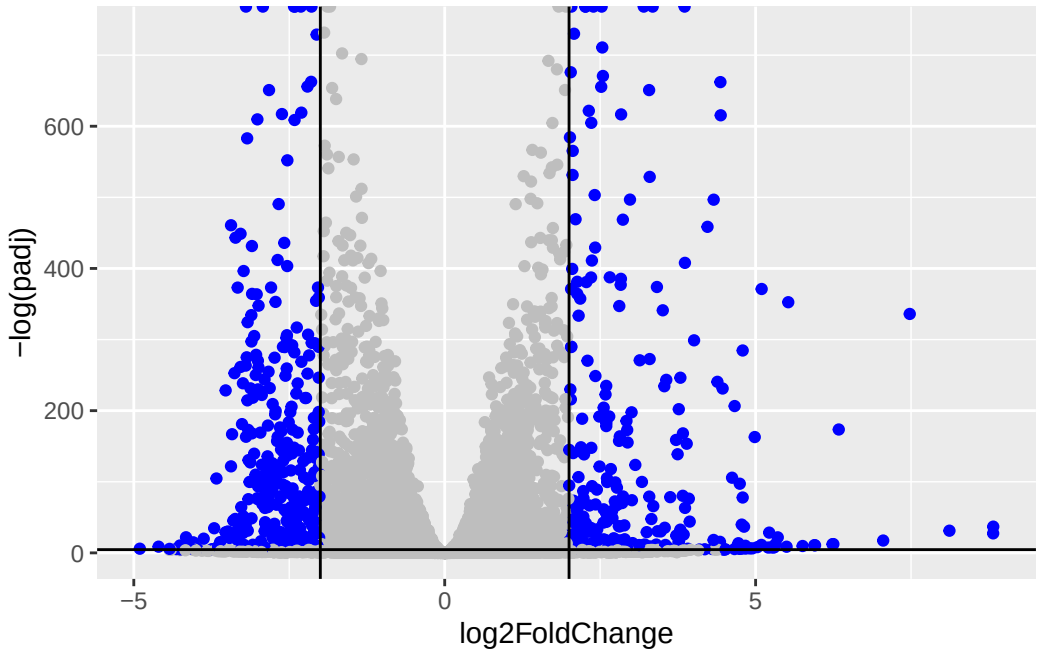
Q. Improve this plot by completing the below code, which adds color, axis labels and cutoff lines:

Let's add some color to this plot along with cutoff lines for fold-change and P-value

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

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

# Color gray those with adjusted p-value more than 0.01
mycols[res$padj > 0.01] <- "gray"
```

```
ggplot(res) +
        aes(log2FoldChange,
       -log(padj)) +
        geom_point(col = mycols) +
        geom_vline(xintercept = c(-2,2)) +
        geom_hline(yintercept = -log(0.01))
```

```
Warning: Removed 5054 rows containing missing values or values outside the scale range
(`geom_point()`).
```



Add annotation

Q. Use the mapIDs() function multiple times to add SYMBOL, ENTREZID and GENENAME annotation to our results by completing the code below.

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

```
columns(org.Hs.eg.db)
```

```
 [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
 [6] "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"  "GENENAME"
[11] "GENETYPE"     "GO"           "GOALL"        "IPI"          "MAP"
[16] "OMIM"         "ONTOLOGY"     "ONTOLOGYALL"  "PATH"         "PFAM"
[21] "PMID"         "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"
[26] "UNIPROT"
```

```
res$symbol = mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="SYMBOL",
                    multiVals="first")
```

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

```
res$entrez = mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="ENTREZID",
                    multiVals="first")
```

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

```
res$name =   mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="GENENAME",
                    multiVals="first")
```

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

```
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>
ENSG00000186092    0.0000             NA        NA         NA          NA
ENSG00000279928    0.0000             NA        NA         NA          NA
ENSG00000279457   29.9136      0.1792571 0.3248216   0.551863 5.81042e-01
ENSG00000278566    0.0000             NA        NA         NA          NA
ENSG00000273547    0.0000             NA        NA         NA          NA
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      symbol      entrez                   name
                  <numeric> <character> <character>            <character>
ENSG00000186092          NA       OR4F5       79501 olfactory receptor f..
ENSG00000279928          NA          NA          NA                    NA
ENSG00000279457 6.87080e-01          NA          NA                    NA
ENSG00000278566          NA          NA          NA                    NA
ENSG00000273547          NA          NA          NA                    NA
ENSG00000187634 5.16278e-03      SAMD11      148398 sterile alpha motif ..
ENSG00000188976 1.76741e-35       NOC2L       26155 NOC2 like nucleolar ..
ENSG00000187961 1.13536e-07      KLHL17      339451 kelch like family me..
ENSG00000187583 9.18988e-01     PLEKHN1       84069 pleckstrin homology ..
ENSG00000187642 4.03817e-01       PERM1       84808 PPARGC1 and ESRR ind..
```

Save annotated results

Q. Finally for this section let's reorder these results by adjusted p-value and save them to a CSV file in your current project directory.

```
write.csv(res, file = "results_annotated.csv")
```

Pathway Analysis

```
library(gage)
library(gageData)
library(pathview)

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"
```

```r
foldchanges = res$log2FoldChange
names(foldchanges) = res$entrez
head(foldchanges)
```

```
    79501      <NA>      <NA>      <NA>      <NA>    148398
       NA        NA 0.1792571        NA        NA 0.4264571
```

```
# Get the results
keggres <- gage(foldchanges, gsets=kegg.sets.hs)

attributes(keggres)
```

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

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

```
                                       p.geomean stat.mean        p.val
hsa04110 Cell cycle                   7.077982e-06 -4.432593 7.077982e-06
hsa03030 DNA replication              9.424076e-05 -3.951803 9.424076e-05
hsa03013 RNA transport                1.048017e-03 -3.112129 1.048017e-03
hsa04114 Oocyte meiosis               2.563806e-03 -2.827297 2.563806e-03
hsa03440 Homologous recombination     3.066756e-03 -2.852899 3.066756e-03
hsa00010 Glycolysis / Gluconeogenesis 4.360092e-03 -2.663825 4.360092e-03
                                            q.val set.size         exp1
hsa04110 Cell cycle                   0.001160789      124 7.077982e-06
hsa03030 DNA replication              0.007727742       36 9.424076e-05
hsa03013 RNA transport                0.057291598      149 1.048017e-03
hsa04114 Oocyte meiosis               0.100589607      112 2.563806e-03
hsa03440 Homologous recombination     0.100589607       28 3.066756e-03
hsa00010 Glycolysis / Gluconeogenesis 0.119175854       65 4.360092e-03
```

```
pathview(gene.data=foldchanges, pathway.id="hsa04110")
```

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

```
Info: Working in directory /Users/katherinequach/Desktop/BIMM 143/Week 7
```

```
Info: Writing image file hsa04110.pathview.png
```

```
# A different PDF based output of the same data
pathview(gene.data=foldchanges, pathway.id="hsa04110", kegg.native=FALSE)
```

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

```
Warning: reconcile groups sharing member nodes!

     [,1] [,2]
[1,] "9"  "300"
[2,] "9"  "306"

Info: Working in directory /Users/katherinequach/Desktop/BIMM 143/Week 7

Info: Writing image file hsa04110.pathview.pdf
```

```
# Focus on top 5 upregulated pathways here for demo purposes only
keggrespathways <- rownames(keggres$greater)[1:5]

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

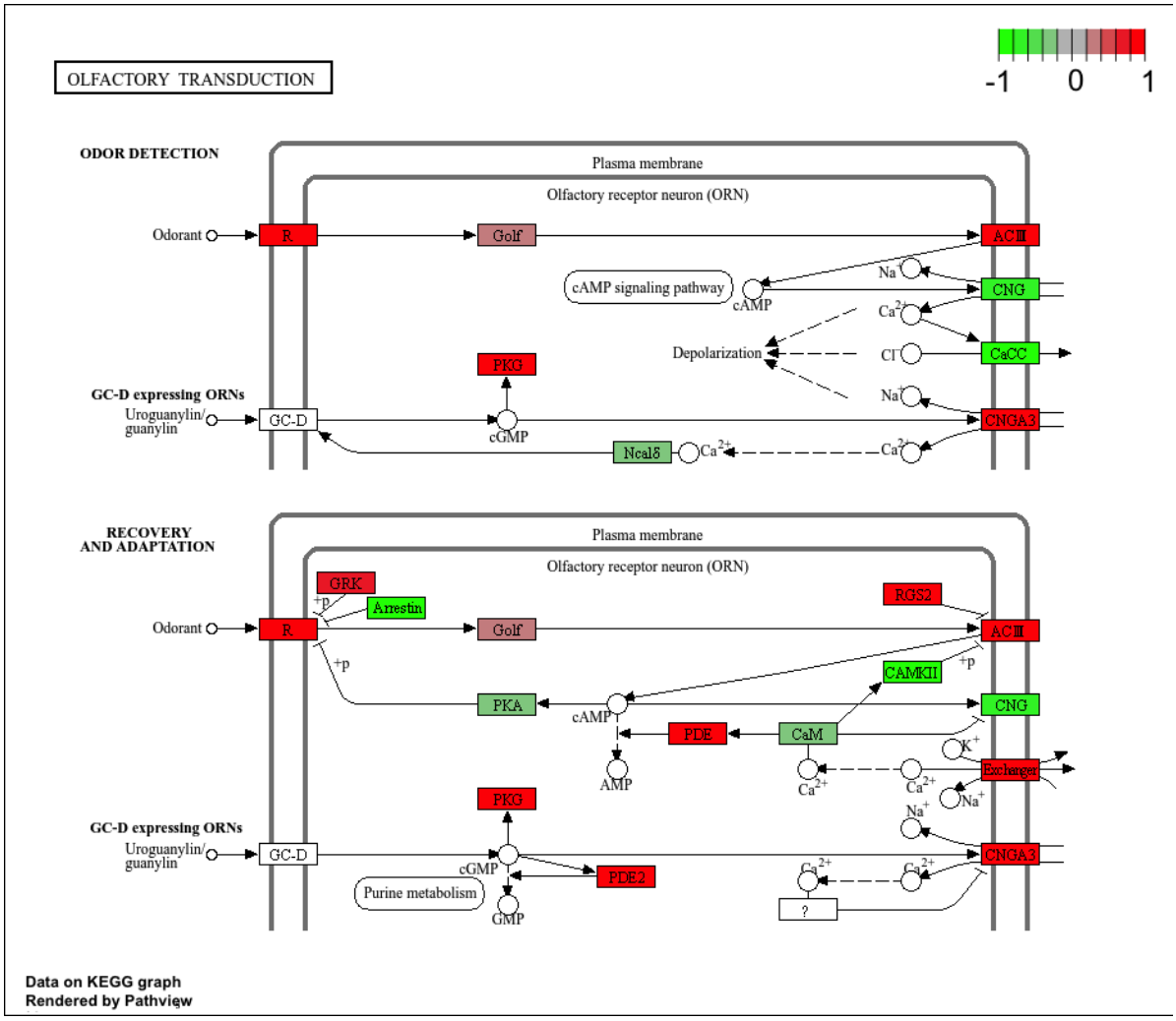
```
[1] "hsa04740" "hsa04640" "hsa00140" "hsa04630" "hsa04976"
```

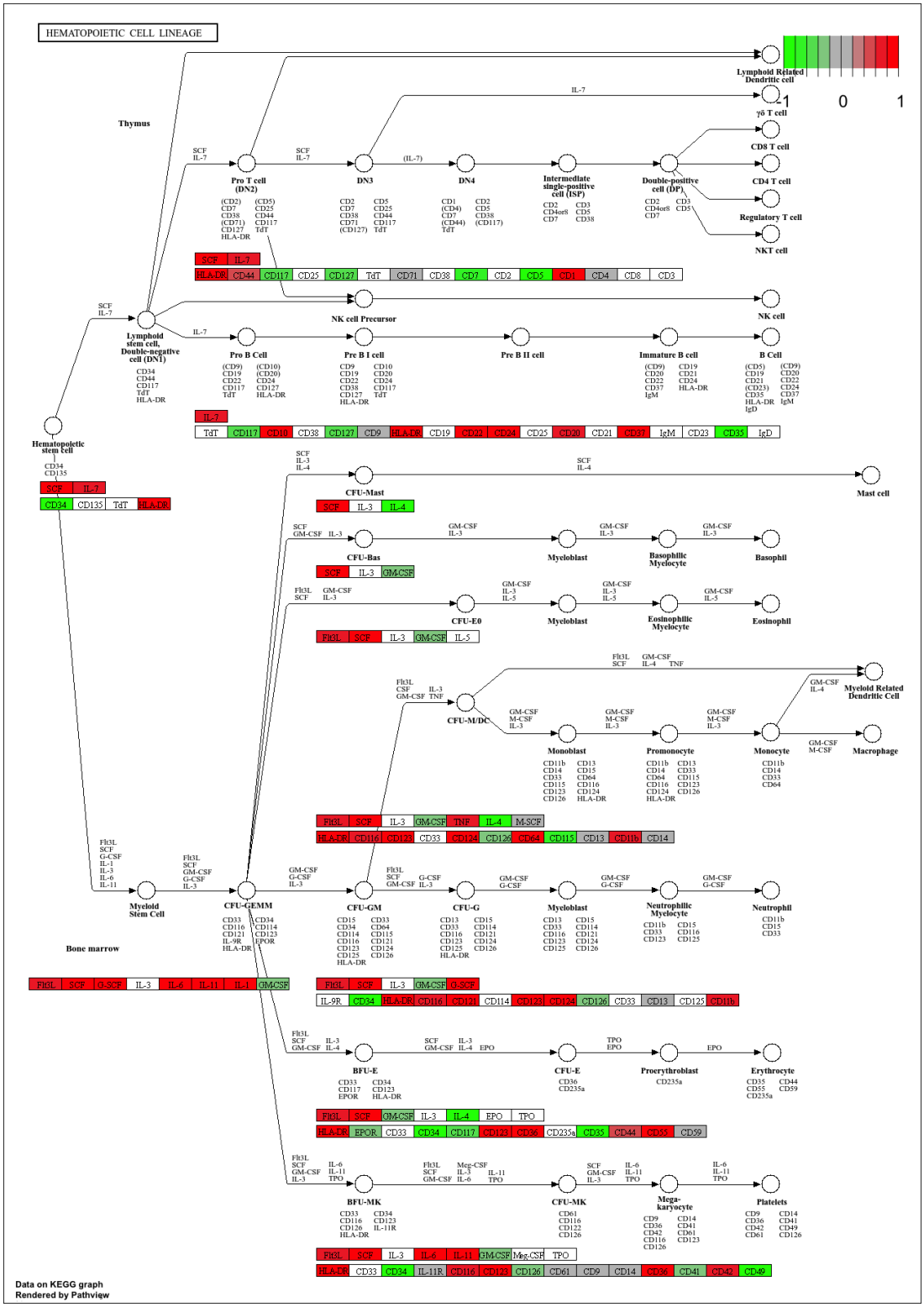
```
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

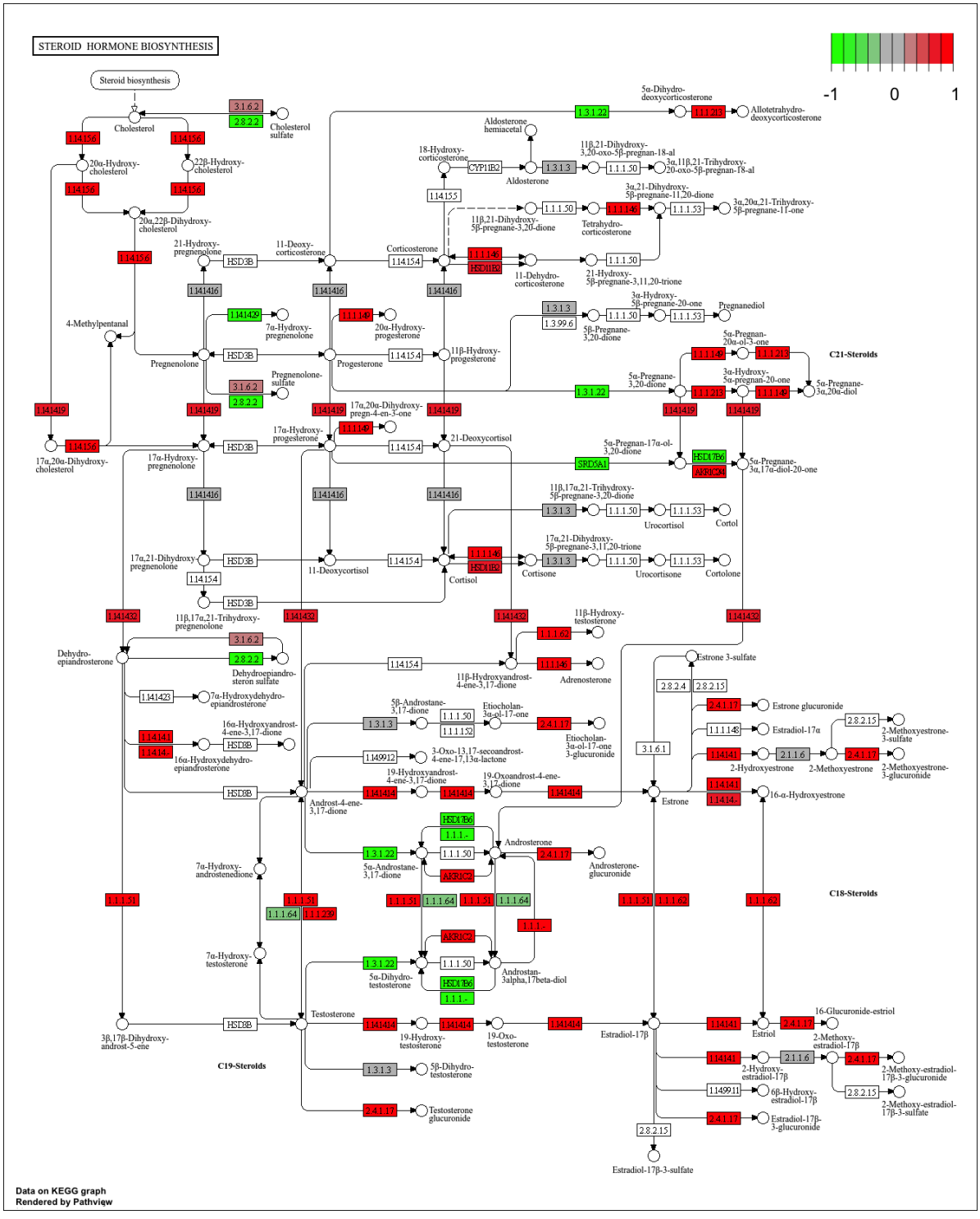
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/katherinequach/Desktop/BIMM 143/Week 7

Info: Writing image file hsa04740.pathview.png

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

Info: Working in directory /Users/katherinequach/Desktop/BIMM 143/Week 7

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory /Users/katherinequach/Desktop/BIMM 143/Week 7
```

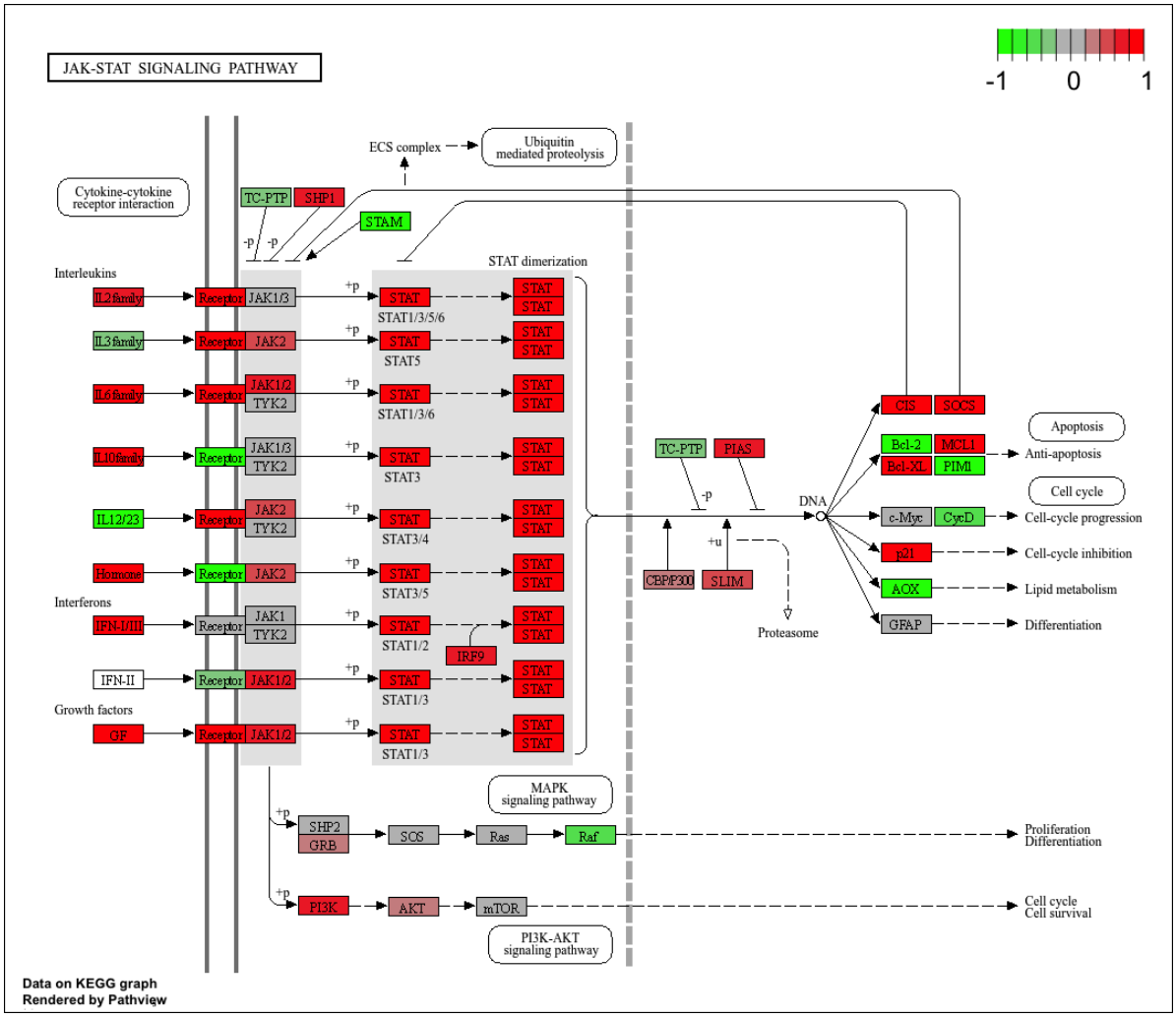
```
Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory /Users/katherinequach/Desktop/BIMM 143/Week 7

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory /Users/katherinequach/Desktop/BIMM 143/Week 7

Info: Writing image file hsa04976.pathview.png
```

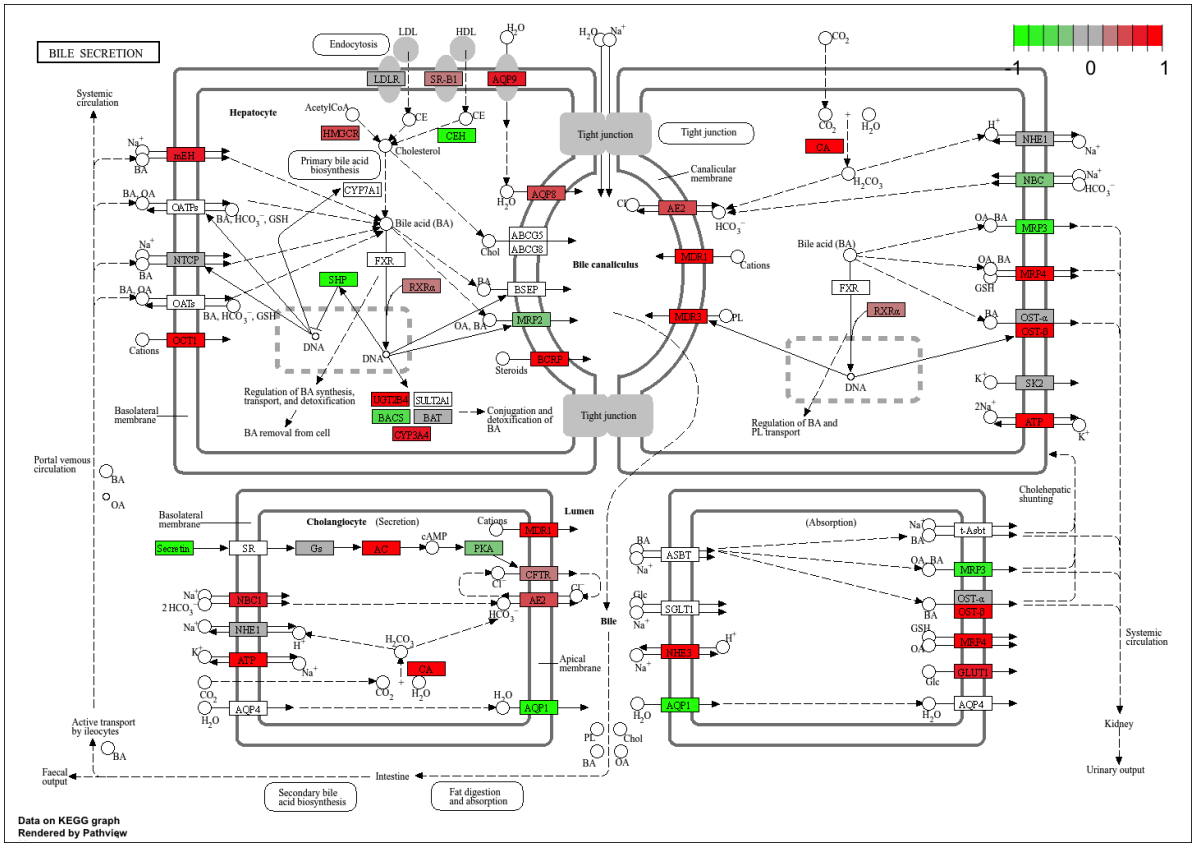
Q. Can you do the same procedure as above to plot the pathview figures for the top 5 down-regulated pathways?


OLFACTORY TRANSDUCTION
-1
0
1
ODOR DETECTION
Plasma membrane
Olfactory receptor neuron (ORN)
Odorant
R
Golf
ACIII
cAMP signaling pathway
cAMP
Na+
CNG
Ca2+
Depolarization
Cl-
CaCC
PKG
GC-D expressing ORNs
Uroguanylin/
guanylin
GC-D
cGMP
Na+
CNGA3
Ncal6
Ca2+
RECOVERY
AND ADAPTATION
Plasma membrane
Olfactory receptor neuron (ORN)
GRK
+p
Arrestin
RGS2
Odorant
R
Golf
ACIII
+p
CAMKII
+p
PKA
cAMP
CNG
PDE
CaM
K+
AMP
Ca2+
Exchanger
Ca2+
Na+
Na+
PKG
GC-D expressing ORNs
Uroguanylin/
guanylin
GC-D
cGMP
PDE2
CNGA3
Ca2+
Ca2+
Purine metabolism
GMP
?
Data on KEGG graph
Rendered by Pathview


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


STEROID HORMONE BIOSYNTHESIS
-1
0
1
Steroid biosynthesis
Cholesterol
Cholesterol sulfate
20α-Hydroxy-cholesterol
22β-Hydroxy-cholesterol
20α,22β-Dihydroxy-cholesterol
4-Methylpentanal
Pregnenolone
Pregnenolone-sulfate
21-Hydroxy-pregnenolone
11-Deoxy-corticosterone
Corticosterone
11-Dehydro-corticosterone
Tetrahydro-corticosterone
11β,21-Dihydroxy-5β-pregnane-3,20-dione
21-Hydroxy-5β-pregnane-3,11,20-trione
18-Hydroxy-corticosterone
Aldosterone
Aldosterone hemiacetal
5α-Dihydro-deoxycorticosterone
Allotetrahydro-deoxycorticosterone
11β,21-Dihydroxy-3,20-oxo-5β-pregnan-18-al
3α,11β,21-Trihydroxy-20-oxo-5β-pregnan-18-al
3α,21-Dihydroxy-5β-pregnane-11,20-dione
3α,20α,21-Trihydroxy-5β-pregnane-11-one
3α-Hydroxy-5β-pregnane-20-one
Pregnanediol
5β-Pregnane-3,20-dione
5α-Pregnan-20α-ol-3-one
5α-Pregnane-3,20-dione
3α-Hydroxy-5α-pregnan-20-one
5α-Pregnane-3α,20α-diol
C21-Steroids
7α-Hydroxy-pregnenolone
20α-Hydroxy-progesterone
Progesterone
11β-Hydroxy-progesterone
17α,20α-Dihydroxy-pregn-4-en-3-one
17α-Hydroxy-progesterone
21-Deoxycortisol
17α,20α-Dihydroxy-cholesterol
17α-Hydroxy-pregnenolone
5α-Pregnan-17α-ol-3,20-dione
5α-Pregnane-3α,17α-diol-20-one
11β,17α,21-Trihydroxy-5β-pregnane-3,20-dione
Urocortisol
Cortol
17α,21-Dihydroxy-pregnenolone
11-Deoxycortisol
Cortisol
17α,21-Dihydroxy-5β-pregnane-3,11,20-trione
Cortisone
Urocortisone
Cortolone
11β,17α,21-Trihydroxy-pregnenolone
Dehydro-epiandrosterone
Dehydroepiandro-steron sulfate
7α-Hydroxydehydro-epiandrosterone
16α-Hydroxydehydro-epiandrosterone
11β-Hydroxy-testosterone
11β-Hydroxyandrost-4-ene-3,17-dione
Adrenosterone
Estrone 3-sulfate
5β-Androstane-3,17-dione
Etiocholan-3α-ol-17-one
Etiocholan-3α-ol-17-one 3-glucuronide
3-Oxo-13,17-secoandrost-4-ene-17,13α-lactone
19-Hydroxyandrost-4-ene-3,17-dione
19-Oxoandrost-4-ene-3,17-dione
Androst-4-ene-3,17-dione
Estrone
Estrone glucuronide
Estradiol-17α
2-Hydroxyestrone
2-Methoxyestrone
2-Methoxyestrone-3-sulfate
2-Methoxyestrone-3-glucuronide
16-α-Hydroxyestrone
Androsterone
Androsterone-glucuronide
5α-Androstane-3,17-dione
7α-Hydroxy-androstenedione
C18-Steroids
7α-Hydroxy-testosterone
5α-Dihydro-testosterone
Androstan-3alpha,17beta-diol
Testosterone
3β,17β-Dihydroxy-androst-5-ene
19-Hydroxy-testosterone
19-Oxo-testosterone
Estradiol-17β
Estriol
16-Glucuronide-estriol
2-Methoxy-estradiol-17β
2-Hydroxy-estradiol-17β
2-Methoxy-estradiol-17β-3-glucuronide
6β-Hydroxy-estradiol-17β
Estradiol-17β-3-glucuronide
2-Methoxy-estradiol-17β-3-sulfate
Estradiol-17β-3-sulfate
C19-Steroids
5β-Dihydro-testosterone
Testosterone glucuronide
Data on KEGG graph
Rendered by Pathview


JAK-STAT SIGNALING PATHWAY
-1
0
1
ECS complex
Ubiquitin mediated proteolysis
Cytokine-cytokine receptor interaction
TC-PTP
SHP1
STAM
-p
STAT dimerization
Interleukins
IL2 family
IL3 family
IL6 family
IL10 family
IL12/23
Hormone
Interferons
IFN-I/III
IFN-II
Growth factors
GF
Receptor
JAK1/3
JAK2
JAK1/2
TYK2
JAK1
+p
STAT
STAT1/3/5/6
STAT5
STAT1/3/6
STAT3
STAT3/4
STAT3/5
STAT1/2
STAT1/3
IRF9
CBP/P300
PIAS
+u
SLIM
Proteasome
DNA
CIS
SOCS
Bcl-2
MCL1
Bcl-XL
PIM1
Apoptosis
Anti-apoptosis
Cell cycle
c-Myc
CycD
Cell-cycle progression
p21
Cell-cycle inhibition
AOX
Lipid metabolism
GFAP
Differentiation
MAPK signaling pathway
SHP2
GRB
SOS
Ras
Raf
Proliferation
Differentiation
PI3K
AKT
mTOR
Cell cycle
Cell survival
PI3K-AKT signaling pathway
Data on KEGG graph
Rendered by Pathview



Gene Ontology (GO) Analysis

```
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)

lapply(gobpres, head)
```

```
$greater
                                          p.geomean stat.mean        p.val
GO:0007156 homophilic cell adhesion      1.734864e-05  4.210777 1.734864e-
05
```

```
GO:0048729 tissue morphogenesis            5.407952e-05  3.888470 5.407952e-
05
GO:0002009 morphogenesis of an epithelium  5.727599e-05  3.878706 5.727599e-
05
GO:0030855 epithelial cell differentiation 2.053700e-04  3.554776 2.053700e-
04
GO:0060562 epithelial tube morphogenesis   2.927804e-04  3.458463 2.927804e-
04
GO:0048598 embryonic morphogenesis         2.959270e-04  3.446527 2.959270e-
04
                                                 q.val set.size         exp1
GO:0007156 homophilic cell adhesion        0.07584825      137 1.734864e-05
GO:0048729 tissue morphogenesis            0.08347021      483 5.407952e-05
GO:0002009 morphogenesis of an epithelium  0.08347021      382 5.727599e-05
GO:0030855 epithelial cell differentiation 0.16449701      299 2.053700e-04
GO:0060562 epithelial tube morphogenesis   0.16449701      289 2.927804e-04
GO:0048598 embryonic morphogenesis         0.16449701      498 2.959270e-04

$less
                                           p.geomean stat.mean        p.val
GO:0048285 organelle fission             6.386337e-16 -8.175381 6.386337e-16
GO:0000280 nuclear division              1.726380e-15 -8.056666 1.726380e-15
GO:0007067 mitosis                       1.726380e-15 -8.056666 1.726380e-15
GO:0000087 M phase of mitotic cell cycle 4.593581e-15 -7.919909 4.593581e-15
GO:0007059 chromosome segregation        9.576332e-12 -6.994852 9.576332e-12
GO:0051301 cell division                 8.718528e-11 -6.455491 8.718528e-11
                                                 q.val set.size         exp1
GO:0048285 organelle fission             2.515911e-12      386 6.386337e-16
GO:0000280 nuclear division              2.515911e-12      362 1.726380e-15
GO:0007067 mitosis                       2.515911e-12      362 1.726380e-15
GO:0000087 M phase of mitotic cell cycle 5.020784e-12      373 4.593581e-15
GO:0007059 chromosome segregation        8.373545e-09      146 9.576332e-12
GO:0051301 cell division                 6.352901e-08      479 8.718528e-11

$stats
                                         stat.mean     exp1
GO:0007156 homophilic cell adhesion       4.210777 4.210777
GO:0048729 tissue morphogenesis           3.888470 3.888470
GO:0002009 morphogenesis of an epithelium 3.878706 3.878706
GO:0030855 epithelial cell differentiation 3.554776 3.554776
GO:0060562 epithelial tube morphogenesis  3.458463 3.458463
GO:0048598 embryonic morphogenesis        3.446527 3.446527
```

Reactome Analysis

```
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: 8146"
```

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

> 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?

The pathway with the most significant "Entities p-value" is Response of EIF2AK4 (GCN2) to amino acid deficiency with a p-value of 7.1E-3. The most significant pathways (e.g. Cell Cycle Mitotic, Cell Cycle, Cell Cycle Checkpoints, etc.) don't match the previous KEGG results. Factors that would've caused these differences between the 2 methods is data annotation differences, pathway definitions, and gene coverage.

```
sessionInfo()
```

```
R version 4.5.2 (2025-10-31)
Platform: x86_64-apple-darwin20
Running under: macOS Monterey 12.7.6

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.5-x86_64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.5-x86_64/Resources/lib/libRlapack.dylib;

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: America/Los_Angeles
tzcode source: internal

attached base packages:
[1] stats4    stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
```

```
 [1] pathview_1.50.0             gageData_2.48.0
 [3] gage_2.60.0                 org.Hs.eg.db_3.22.0
 [5] AnnotationDbi_1.72.0        ggplot2_4.0.2
 [7] DESeq2_1.50.2               SummarizedExperiment_1.40.0
 [9] Biobase_2.70.0              MatrixGenerics_1.22.0
[11] matrixStats_1.5.0           GenomicRanges_1.62.1
[13] Seqinfo_1.0.0               IRanges_2.44.0
[15] S4Vectors_0.48.0            BiocGenerics_0.56.0
[17] generics_0.1.4

loaded via a namespace (and not attached):
 [1] KEGGREST_1.50.0     gtable_0.3.6        xfun_0.56
 [4] lattice_0.22-9      bitops_1.0-9        vctrs_0.7.1
 [7] tools_4.5.2         parallel_4.5.2      tibble_3.3.1
[10] RSQLite_2.4.6       blob_1.3.0          pkgconfig_2.0.3
[13] Matrix_1.7-4        RColorBrewer_1.1-3  S7_0.2.1
[16] graph_1.88.1        lifecycle_1.0.5     compiler_4.5.2
[19] farver_2.1.2        Biostrings_2.78.0   codetools_0.2-20
[22] htmltools_0.5.9     RCurl_1.98-1.17     yaml_2.3.12
[25] GO.db_3.22.0        pillar_1.11.1       crayon_1.5.3
[28] BiocParallel_1.44.0 DelayedArray_0.36.0 cachem_1.1.0
[31] abind_1.4-8         tidyselect_1.2.1    locfit_1.5-9.12
[34] digest_0.6.39       dplyr_1.2.0         labeling_0.4.3
[37] fastmap_1.2.0       grid_4.5.2          cli_3.6.5
[40] SparseArray_1.10.8  magrittr_2.0.4      S4Arrays_1.10.1
[43] XML_3.99-0.22       withr_3.0.2         scales_1.4.0
[46] bit64_4.6.0-1       rmarkdown_2.30      XVector_0.50.0
[49] httr_1.4.8          bit_4.6.0           otel_0.2.0
[52] png_0.1-8           memoise_2.0.1       evaluate_1.0.5
[55] knitr_1.51          rlang_1.1.7         Rcpp_1.1.1
[58] glue_1.8.0          DBI_1.2.3           Rgraphviz_2.54.0
[61] KEGGgraph_1.70.0    jsonlite_2.0.0      R6_2.6.1
```