Class 14: RNASeq Mini-Project

Yvonne Yu A16333006

Table of contents

A complete RNASeq analysis from counts to pathways and biological insight would be conducted.

Data Import

```
#Assigns the files to the object
counts <- read.csv("GSE37704_featurecounts.csv", row.names = 1)
metadata <- read.csv("GSE37704_metadata.csv")

#Visualizes the files
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
```

```
head(metadata)
```

```
         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
```

The following would delete the first column that identifies the length, and turn the counts into matrix.

```
#Deletes the first column as it is not a count
counts <- as.matrix(counts[,-1])
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
```

The following code removes the rows that has 0 counts through all of the samples. There was an identified of 15975 genes that remains after the removal.

```
#Determines the rows that don't have a sum of zero
x <- rowSums(counts) != 0

#Extracts those rows from the count dataset and assigns it to a new object
new_counts <- counts[x,]
head(new_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

```
#Compares the dimensions to ensure that the rows were removed
dim(counts)
```

```
[1] 19808     6
```

```
dim(new_counts)
```

```
[1] 15975     6
```

Setup for DESeq

Load in the necessary libraries for the project

```
library(DESeq2)
```

Running DESeq

DESeq analysis is conducted by creating the DESeq object and visual outputs the results.

```
#Formats the counts matrix
dds <- DESeqDataSetFromMatrix(countData=new_counts,
                              colData=metadata,
                              design=~condition)
```

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

```
#Assigns the DESeq object to dds
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: 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
```

Provides a summary of the results, contrasting based on the condition.

```
res <- results(dds, contrast=c("condition", "hoxa1_kd", "control_sirna"))
summary(res)
```

```
out of 15975 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 4349, 27%
LFC < 0 (down)     : 4396, 28%
outliers [1]       : 0, 0%
low counts [2]     : 1237, 7.7%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Save the file at current progress

```
write.csv(res, "myresults.csv")
```

Add gene annotation data (gene names etc.)

Adds the gene annotation data by adding the symbol, entrezID, and the gene name.

```
#Pulls the libraries that would be utilized
library("AnnotationDbi")
library("org.Hs.eg.db")
```

```
#Visualizes the columns that are identified in the `org.Hs.eg.db` package
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"
```

```
#Creates a column called "symbol" that would put the symbol of the gene
#based on the ENSEMBL id
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
```

```
#Creates a column called "entrez" that would put the entrezID of the gene
#based on the ENSEMBL id
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
```

```
#Creates a column called "name" that would put the Gene Name of the gene
#based on the ENSEMBL id
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>
ENSG00000279457   29.913579      0.1792571 0.3248216   0.551863 5.81042e-01
ENSG00000187634  183.229650      0.4264571 0.1402658   3.040350 2.36304e-03
```

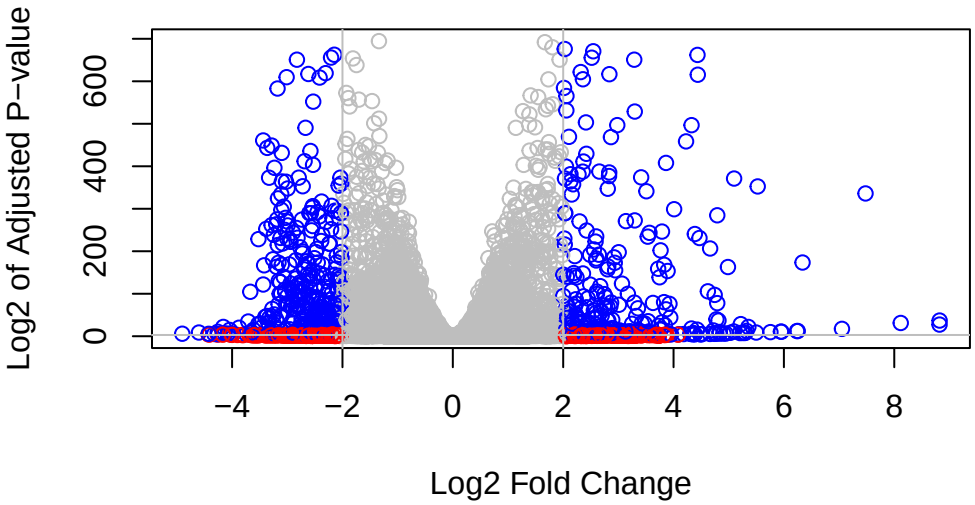
```
ENSG00000188976 1651.188076     -0.6927205 0.0548465 -12.630158 1.43990e-36
ENSG00000187961  209.637938      0.7297556 0.1318599   5.534326 3.12428e-08
ENSG00000187583   47.255123      0.0405765 0.2718928   0.149237 8.81366e-01
ENSG00000187642   11.979750      0.5428105 0.5215598   1.040744 2.97994e-01
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                   name
                  <numeric> <character> <character>            <character>
ENSG00000279457 6.86555e-01          NA          NA                     NA
ENSG00000187634 5.15718e-03      SAMD11      148398 sterile alpha motif ..
ENSG00000188976 1.76549e-35       NOC2L       26155 NOC2 like nucleolar ..
ENSG00000187961 1.13413e-07      KLHL17      339451 kelch like family me..
ENSG00000187583 9.19031e-01     PLEKHN1       84069 pleckstrin homology ..
ENSG00000187642 4.03379e-01       PERM1       84808 PPARGC1 and ESRR ind..
ENSG00000188290 1.30538e-24        HES4       57801 hes family bHLH tran..
ENSG00000187608 2.37452e-02       ISG15        9636 ISG15 ubiquitin like..
ENSG00000188157 4.21963e-16        AGRN      375790                 agrin
ENSG00000237330          NA      RNF223      401934 ring finger protein ..
```

Results visualization

The Following would create a volcano plot, with the addition of the cut off lines and color coding the significant dataplots.

```r
#Makes all of the datapoints grey
mycols <- rep("grey", nrow(res))

#Identifies the locations of the plots that have a log2fold change that is
#greater than 2 or less than -2 (in red)
mycols[abs(res$log2FoldChange) >= 2 ] <- "red"

#Identifies the locations of the plots that have a log2fold change that is
#greater than 2 or less than -2, and has a p-value less than 0.05 (in blue)
mycols[-log(res$padj) >= -log2(0.05) & abs(res$log2FoldChange) >=2] <- "blue"

#Plots the graph
plot(res$log2FoldChange, -log(res$padj),
     col = mycols, ylab = "Log2 of Adjusted P-value",
     xlab = "Log2 Fold Change")
```

```r
#Creates the lines at -2 and 2 (which is basically the cut off for significant points)
abline(v = -2, col = "gray")
abline(v = 2, col = "gray")

#Cut off for the p-value less than 0.05
abline(h = -log(0.05), col = "gray")
```



Save our Results

```r
#Orders the res dataset based on the p-value
res <- res[order(res$pvalue),]

#Saves the res file into a csv file
write.csv(res, file = "deseq_results.csv")
```

Pathway analysis (KEGG, GO, Reactome)

KEGG

Load the libraries that are needed for the pathway analysis

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

```
data(kegg.sets.hs)

# Examine the first 3 pathways
head(kegg.sets.hs, 2)
```

```
$`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"
```

```
#Identifies the entrez ids with the log2foldchange values
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
```

```
#utilizes the gage analysis to pull the more significant of the pathways
keggres <- gage(foldchanges, gsets=kegg.sets.hs)
attributes(keggres)
```

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

```
head(keggres$less)
```

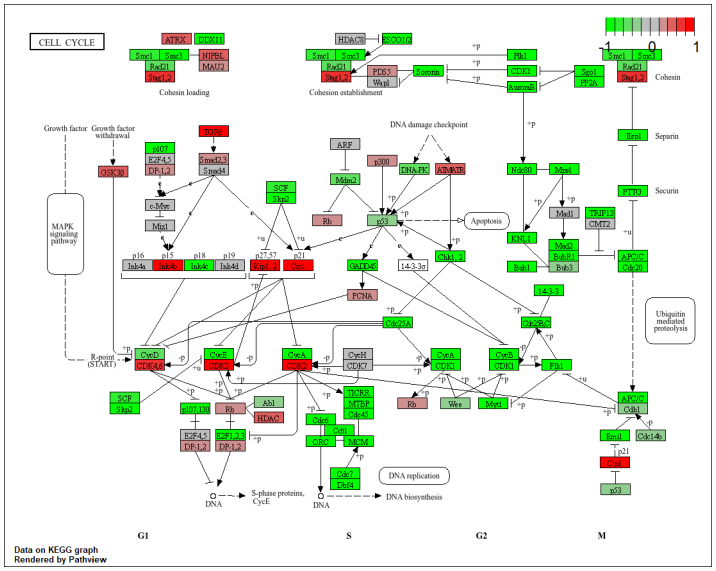
```
                                                   p.geomean stat.mean
hsa04110 Cell cycle                                8.995727e-06 -4.378644
hsa03030 DNA replication                           9.424076e-05 -3.951803
hsa05130 Pathogenic Escherichia coli infection 1.405864e-04 -3.765330
hsa03013 RNA transport                             1.375901e-03 -3.028500
hsa03440 Homologous recombination                  3.066756e-03 -2.852899
hsa04114 Oocyte meiosis                            3.784520e-03 -2.698128
                                                          p.val       q.val
hsa04110 Cell cycle                                8.995727e-06 0.001889103
hsa03030 DNA replication                           9.424076e-05 0.009841047
hsa05130 Pathogenic Escherichia coli infection 1.405864e-04 0.009841047
hsa03013 RNA transport                             1.375901e-03 0.072234819
hsa03440 Homologous recombination                  3.066756e-03 0.128803765
hsa04114 Oocyte meiosis                            3.784520e-03 0.132458191
                                                   set.size         exp1
hsa04110 Cell cycle                                     121 8.995727e-06
hsa03030 DNA replication                                 36 9.424076e-05
hsa05130 Pathogenic Escherichia coli infection           53 1.405864e-04
hsa03013 RNA transport                                  144 1.375901e-03
hsa03440 Homologous recombination                        28 3.066756e-03
hsa04114 Oocyte meiosis                                 102 3.784520e-03
```

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

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

```
Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14
```

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



Identifies the top 5 up regulated pathways

```
#Extracting the pathways for the top 5 up-regulated pathways
up_keggrespathways <- rownames(keggres$greater)[1:5]

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

```
[1] "hsa04060" "hsa05323" "hsa05146" "hsa05332" "hsa04640"
```

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

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

Info: Writing image file hsa04060.pathview.png

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

Info: Writing image file hsa05323.pathview.png
```

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

Info: Writing image file hsa05146.pathview.png

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

Info: Writing image file hsa05332.pathview.png

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

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

The following figures are pathways that are found to up-regulate.

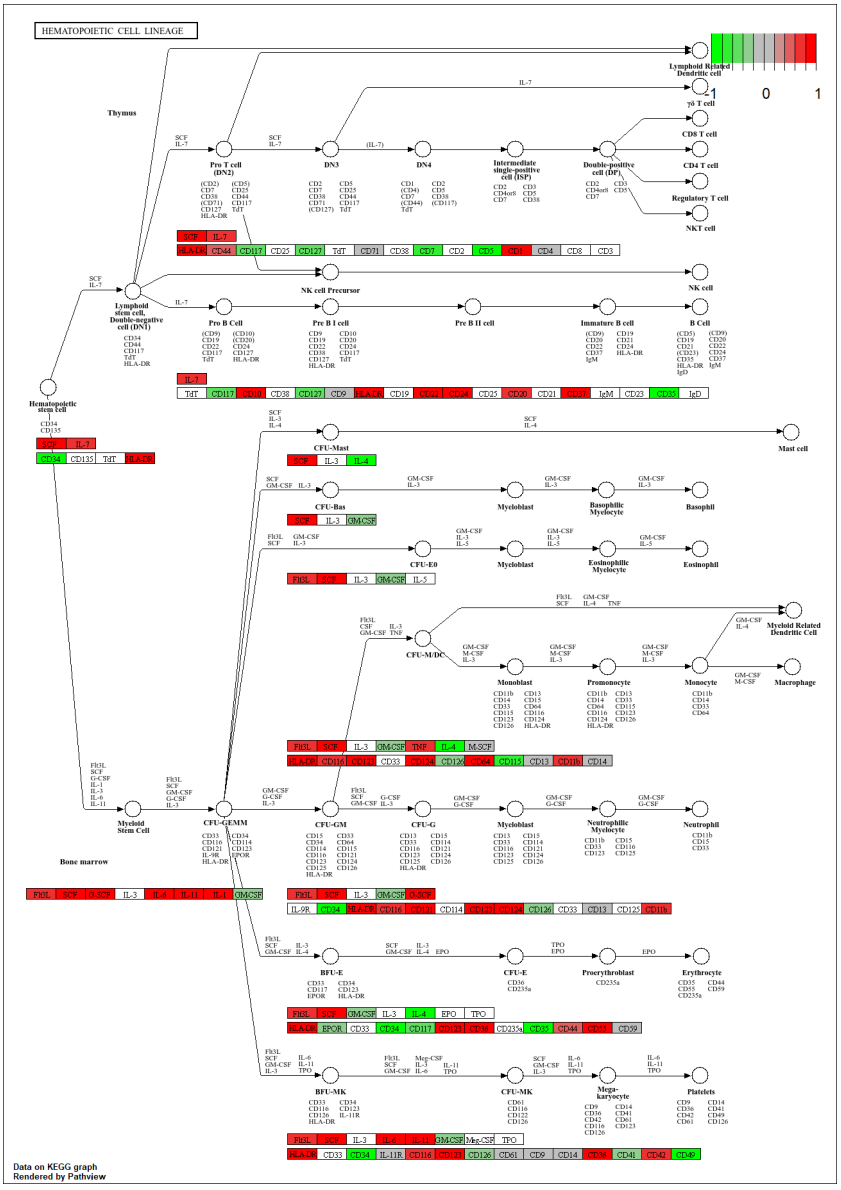


Figure 1: HSA04640

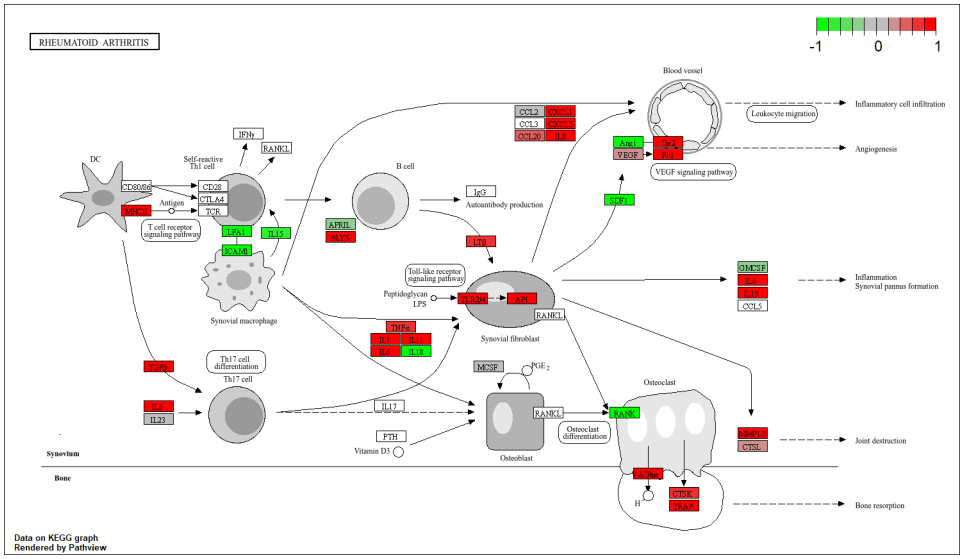


Figure 2: hsa05323

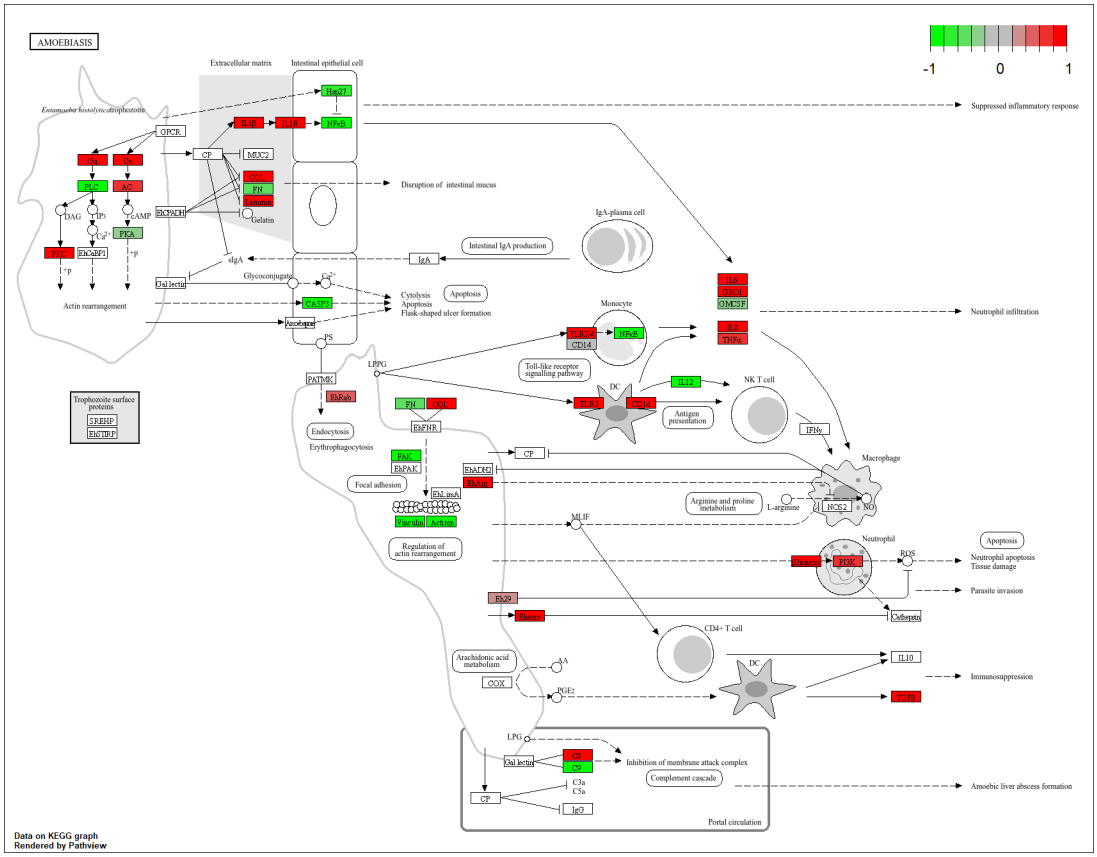


Figure 3: hsa05146

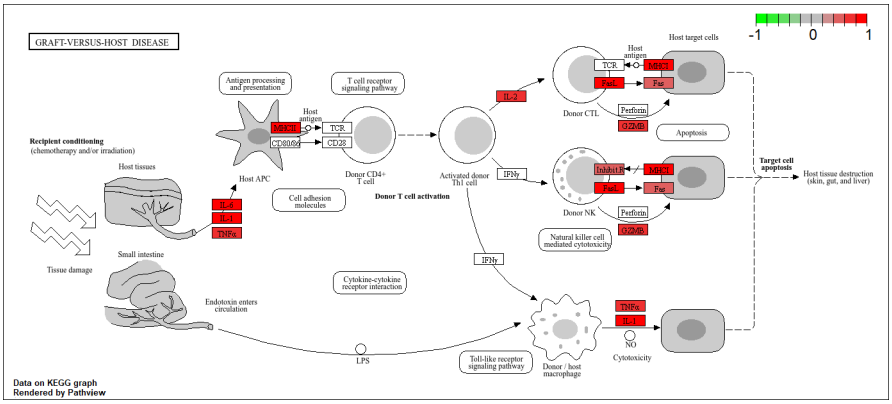
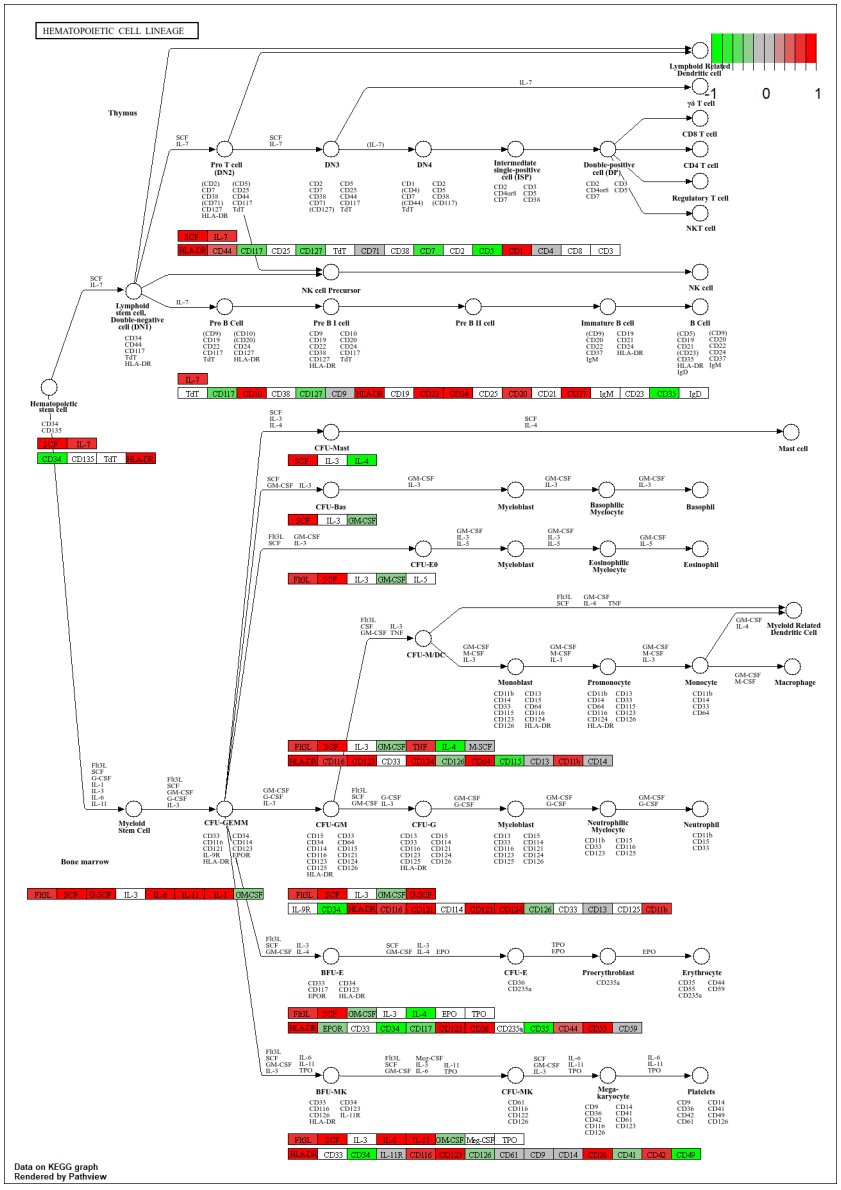


Figure 4: hsa05332



Identifies the Top 5 Down Regulated Pathways

```
#Extracting the pathways for the top 5 down-regulated pathways
low_keggrespathways <- rownames(keggres$less)[1:5]

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

```
[1] "hsa04110" "hsa03030" "hsa05130" "hsa03013" "hsa03440"
```

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

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

Info: Writing image file hsa04110.pathview.png

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

Info: Writing image file hsa03030.pathview.png

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

Info: Writing image file hsa05130.pathview.png

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

Info: Writing image file hsa03013.pathview.png

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

Info: Working in directory C:/Users/Yuyvo/OneDrive/Documents/BIMM143/Class14

Info: Writing image file hsa03440.pathview.png
```

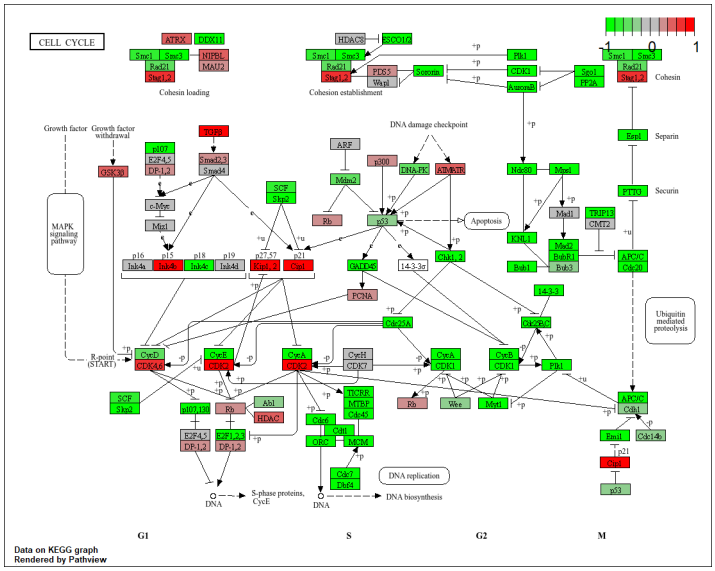


Figure 5: hsa04110

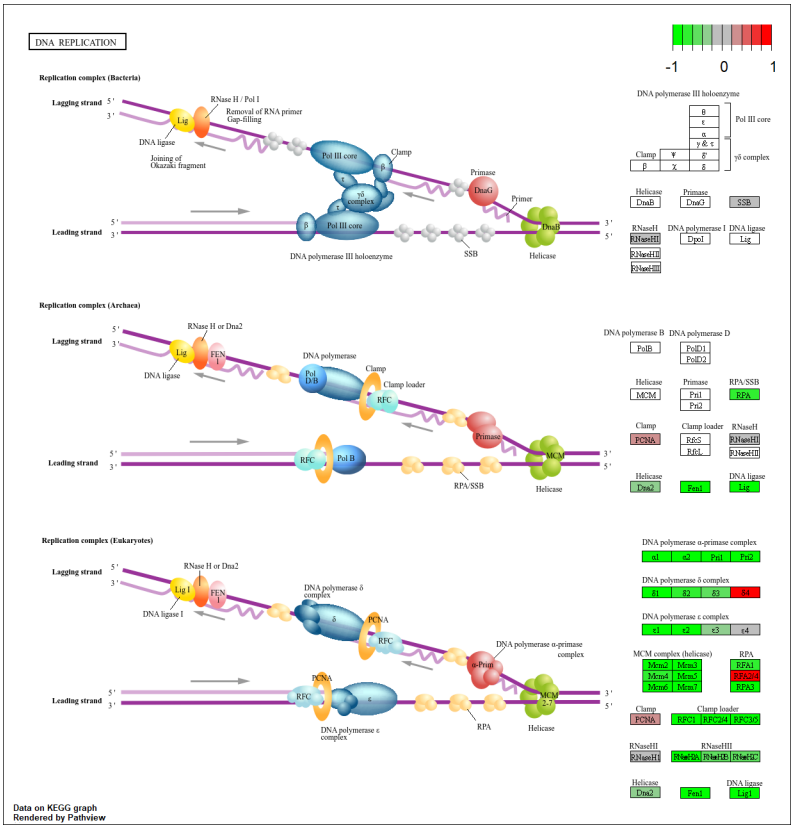


Figure 6: hsa03030

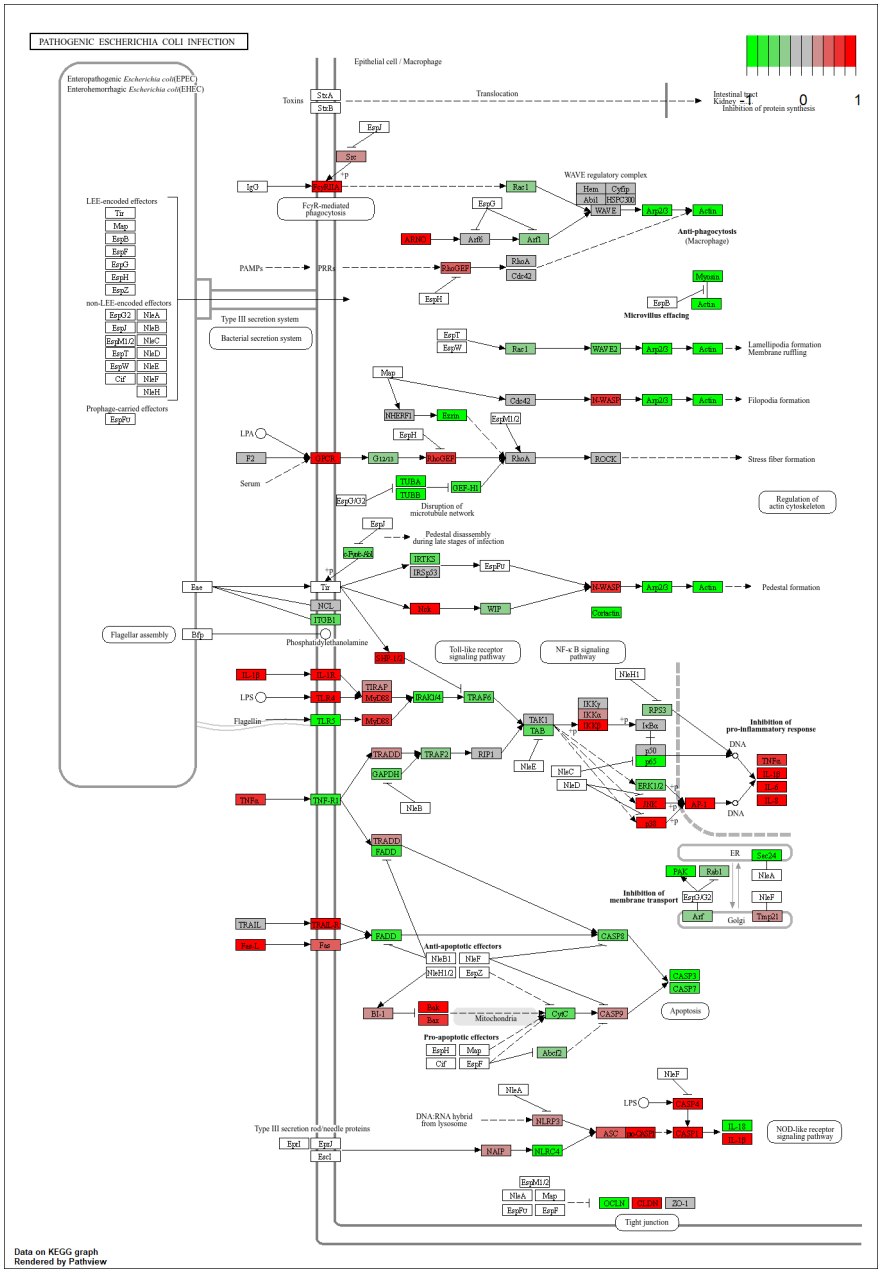


Figure 7: hsa05130

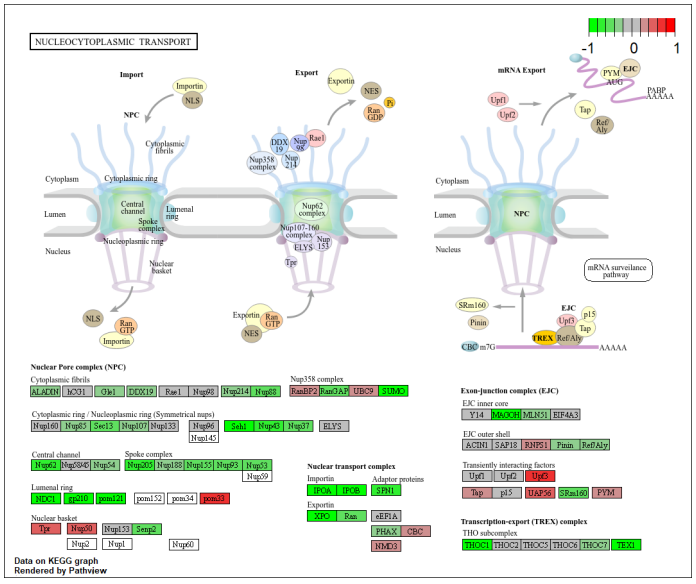


Figure 8: hsa03013

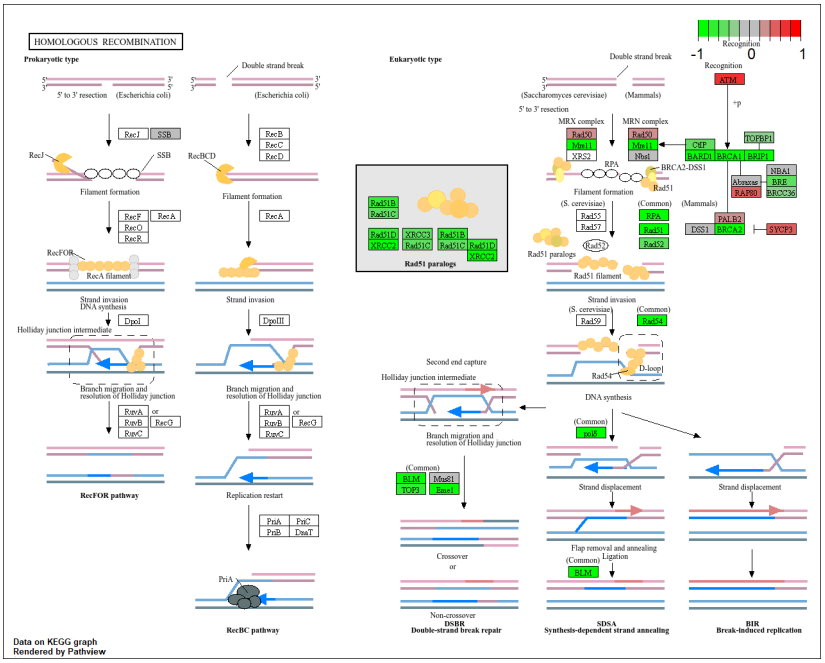


Figure 9: hsa03440

GO

The following code chunk utilizes GO to identify significant pathways.

```
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)
```

```
head(gobpres$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             5.843127e-12      376 1.536227e-15
GO:0000280 nuclear division              5.843127e-12      352 4.286961e-15
GO:0007067 mitosis                       5.843127e-12      352 4.286961e-15
GO:0000087 M phase of mitotic cell cycle 1.195965e-11      362 1.169934e-14
GO:0007059 chromosome segregation        1.659009e-08      142 2.028624e-11
GO:0000236 mitotic prometaphase          1.178690e-07       84 1.729553e-10
```

Reactome

The utilization of reactome is possible through an R package or through the online version, which allows for a more user friendly digital workflow on interactive visualization features. The following would be utilizing the web version.

First the creation of significant genes is necessary.

```
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: 8147"
```

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

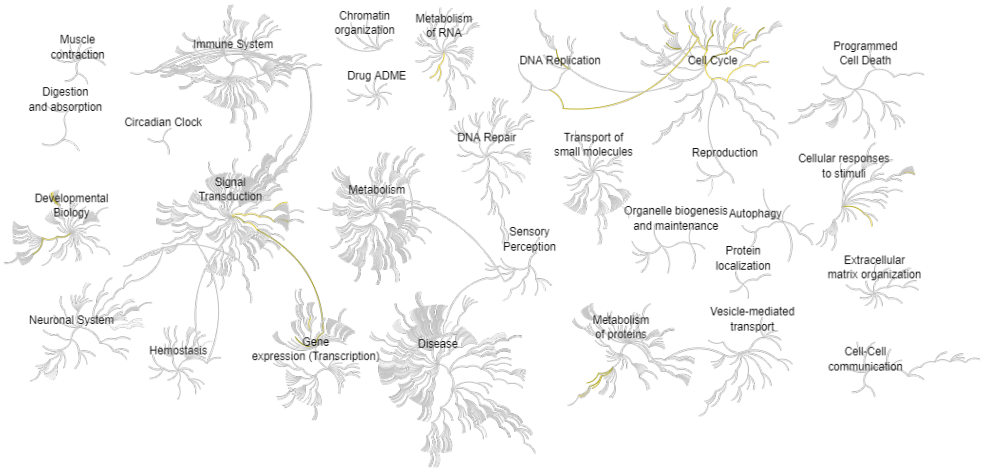


Figure 10: Output for the Reactome

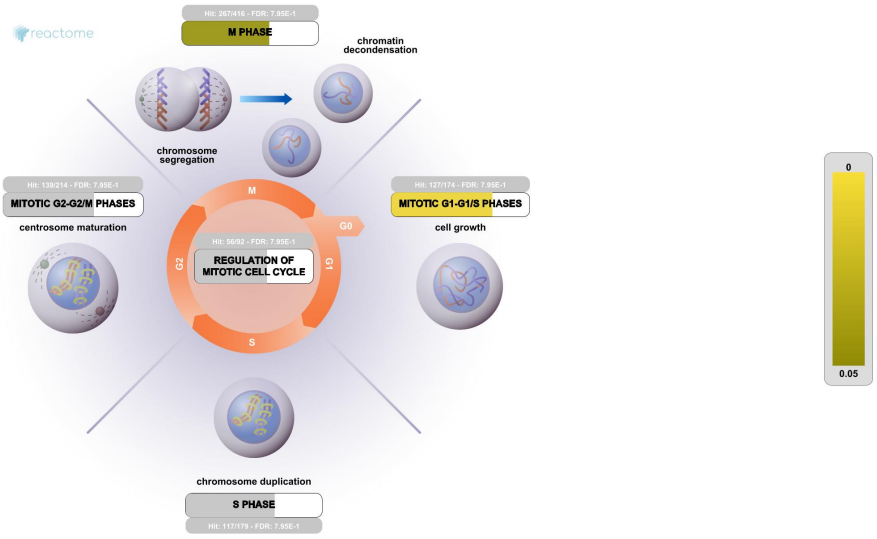


Figure 11: Pathway of the Most Significant Entities P-value

The most significant pathway that was identified through Reactome was the Cell Cycle, which

is similar to what was identified through KEGG.

GO Online Results

The following outputs the reuslts that was found from using the GO Online Pathway.

	Homo sapiens (REF)	upload_1 (Hierarchy) NEW!				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	▼ FDR
regulation of actin filament depolymerization	53	31	21.07	1.47	+	7.07E-03	4.99E-02
hematopoietic progenitor cell differentiation	106	56	42.14	1.33	+	7.02E-03	4.95E-02
cellular response to biotic stimulus	233	113	92.63	1.22	+	6.99E-03	4.94E-02
endoderm development	84	46	33.39	1.38	+	6.96E-03	4.91E-02
potassium ion transport	178	53	70.76	.75	-	6.93E-03	4.90E-02
meiotic cell cycle	256	123	101.77	1.21	+	6.92E-03	4.89E-02
positive regulation of telomerase activity	33	21	13.12	1.60	+	6.85E-03	4.85E-02
regulation of oligodendrocyte differentiation	47	28	18.68	1.50	+	6.84E-03	4.85E-02
cytokine production	29	19	11.53	1.65	+	6.82E-03	4.85E-02
regulation of intrinsic apoptotic signaling pathway by p53 class mediator	33	21	13.12	1.60	+	6.85E-03	4.85E-02
positive regulation of gluconeogenesis	18	13	7.16	1.82	+	6.79E-03	4.85E-02
membrane lipid catabolic process	47	28	18.68	1.50	+	6.84E-03	4.84E-02
regulation of astrocyte differentiation	29	19	11.53	1.65	+	6.82E-03	4.84E-02
regulation of spindle assembly	33	21	13.12	1.60	+	6.85E-03	4.84E-02
nucleobase biosynthetic process	18	13	7.16	1.82	+	6.79E-03	4.84E-02
negative regulation of stress fiber assembly	29	19	11.53	1.65	+	6.82E-03	4.84E-02
heart trabecula morphogenesis	33	21	13.12	1.60	+	6.85E-03	4.84E-02
nucleophagy	18	13	7.16	1.82	+	6.79E-03	4.84E-02
regulation of androgen receptor signaling pathway	29	19	11.53	1.65	+	6.82E-03	4.84E-02
cellular response to cholesterol	18	13	7.16	1.82	+	6.79E-03	4.84E-02
cardiac atrium morphogenesis	29	19	11.53	1.65	+	6.82E-03	4.84E-02
cellular response to prostaglandin E stimulus	18	13	7.16	1.82	+	6.79E-03	4.84E-02
anatomical structure arrangement	18	13	7.16	1.82	+	6.79E-03	4.83E-02
clathrin coat assembly	18	13	7.16	1.82	+	6.79E-03	4.83E-02
hippo signaling	18	13	7.16	1.82	+	6.79E-03	4.83E-02
piecemeal microautophagy of the nucleus	18	13	7.16	1.82	+	6.79E-03	4.83E-02
regulation of plasma membrane organization	18	13	7.16	1.82	+	6.79E-03	4.83E-02
tetrahydrofolate metabolic process	18	13	7.16	1.82	+	6.79E-03	4.82E-02
negative regulation of intracellular steroid hormone receptor signaling pathway	37	23	14.71	1.56	+	6.70E-03	4.79E-02
negative regulation of striated muscle cell differentiation	37	23	14.71	1.56	+	6.70E-03	4.79E-02
lamellipodium assembly	37	23	14.71	1.56	+	6.70E-03	4.79E-02
cardiac atrium development	37	23	14.71	1.56	+	6.70E-03	4.78E-02
regulation of carbohydrate biosynthetic process	97	52	38.56	1.35	+	6.59E-03	4.72E-02

From the utilization of GO, it was identified that the regulation of the actin filament polymerization was found to be the most significant, through FDR calculated p-value.