

lab14.0

Ilyas Darif

2025-02-20

Table of contents

Background	1
Data Import	2
Inspect and tidy data	2
Setup for DESeq	3
Run DESeq	4
Volcano plot of results	5
Gene annotation	7
Pathway analysis	9
Gene Ontology analysis	11

Background

The data for for hands-on session comes from GEO entry: GSE37704, which is associated with 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

The authors report on differential analysis of lung fibroblasts in response to loss of the developmental transcription factor HOXA1. Their results and others indicate that HOXA1 is required for lung fibroblast and HeLa cell cycle progression. In particular their analysis show that “loss of HOXA1 results in significant expression level changes in thousands of individual transcripts, along with isoform switching events in key regulators of the cell cycle”. For our session we have used their Sailfish gene-level estimated counts and hence are restricted to protein-coding genes only.

Data Import

```
counts <- read.csv("GSE37704_featurecounts.csv", row.names=1)
colData <- read.csv("GSE37704_metadata.csv")
```

Inspect and tidy data

Does the counts columns match the colData rows?

```
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					

```
colData$id
```

```
[1] "SRR493366" "SRR493367" "SRR493368" "SRR493369" "SRR493370" "SRR493371"
```

```
colnames(counts)
```

```
[1] "length"      "SRR493366" "SRR493367" "SRR493368" "SRR493369" "SRR493370"
[7] "SRR493371"
```

```
countData <- counts[, -1]
head(countData)
```

	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

check for matching count data and colData

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

```
      id condition
[1,] TRUE      FALSE
[2,] TRUE      FALSE
[3,] TRUE      FALSE
[4,] TRUE      FALSE
[5,] TRUE      FALSE
[6,] TRUE      FALSE
```

Q1. How many genes in total

```
nrow(countData)
```

```
[1] 19808
```

Q2 Filter to remove zero count genes (rows where there are zero counts in all columns). How many genes are left?

```
to.keep.inds <- rowSums(countData) > 0
```

```
new.counts <- countData[to.keep.inds,]
```

```
nrow(new.counts)
```

```
[1] 15975
```

Setup for DESeq

```
library(DESeq2)
```

Setup input object for DESeq

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

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

Run DESeq

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

estimating size factors

estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

final dispersion estimates

fitting model and testing

```
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>
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	6.86555e-01				
ENSG00000187634	5.15718e-03				
ENSG00000188976	1.76549e-35				
ENSG00000187961	1.13413e-07				
ENSG00000187583	9.19031e-01				
ENSG00000187642	4.03379e-01				

Volcano plot of results

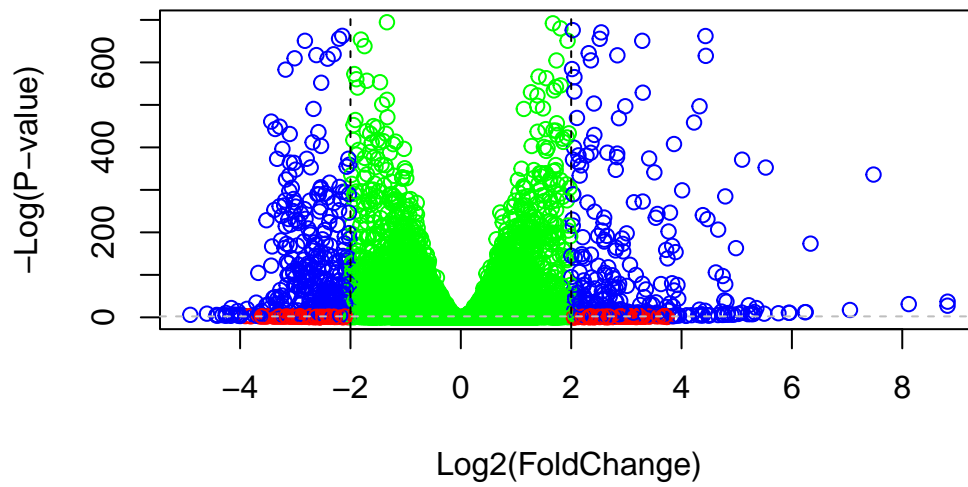
```
library(ggplot2)
```

```
# Setup our custom point color vector
mycols <- rep("green", nrow(res))
mycols[ abs(res$log2FoldChange) > 2 ] <- "red"

inds <- (res$padj < 0.05) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "blue"

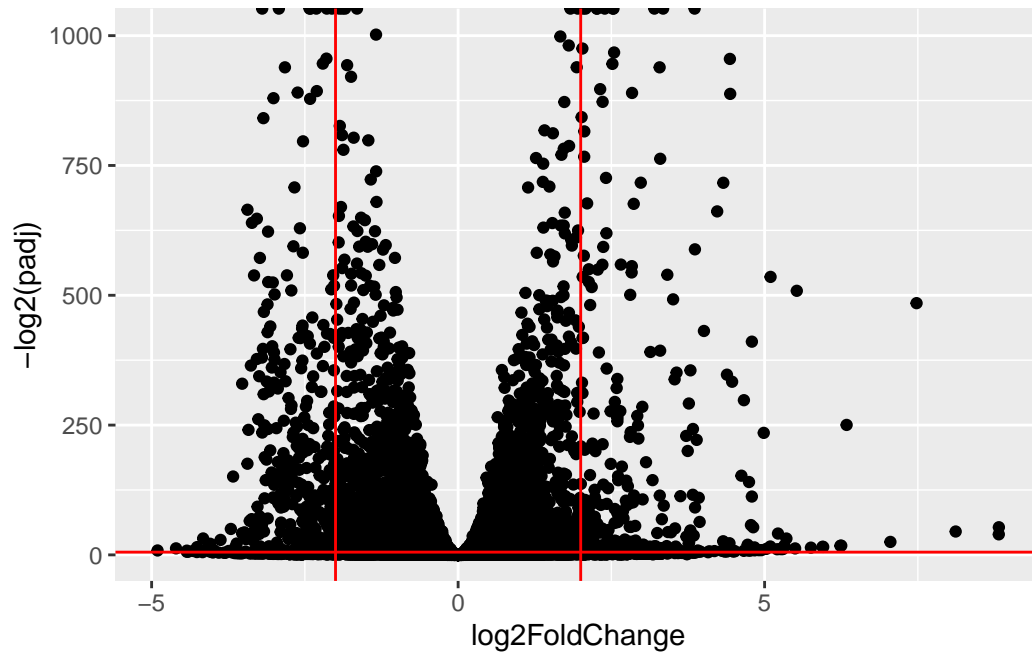
# Volcano plot with custom colors
plot( res$log2FoldChange, -log(res$padj),
      col=mycols, ylab="-Log(P-value)", xlab="Log2(FoldChange)" )

# Cut-off lines
abline(v=c(-2,2), col="black", lty=2)
abline(h=-log(0.1), col="gray", lty=2)
```



```
ggplot(res) +  
  aes(log2FoldChange, -log2(padj)) +  
  geom_point() +  
  geom_vline(xintercept = c(-2,2), col="red") +  
  geom_hline(yintercept = -log(0.005), col="red")
```

Warning: Removed 1237 rows containing missing values or values outside the scale range (`geom_point()`).



Gene annotation

```
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"				

Add gene SYMBOL and ENTREZID

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

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

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

'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 8 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		
	<numeric>	<character>	<character>		
ENSG00000279457	6.86555e-01	NA	NA		
ENSG00000187634	5.15718e-03	SAMD11	148398		
ENSG00000188976	1.76549e-35	NOC2L	26155		
ENSG00000187961	1.13413e-07	KLHL17	339451		
ENSG00000187583	9.19031e-01	PLEKHN1	84069		
ENSG00000187642	4.03379e-01	PERM1	84808		
ENSG00000188290	1.30538e-24	HES4	57801		

ENSG00000187608	2.37452e-02	ISG15	9636
ENSG00000188157	4.21963e-16	AGRN	375790
ENSG00000237330	NA	RNF223	401934

Pathway analysis

```
library(gage)
```

```
library(gageData)
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).
```

```
#####
```

Input vector for gage()

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

Load up the KEGG genesets

```
data(kegg.sets.hs)
```

Run pathway analysis

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

```
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	1.246882e-03	-3.059466
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	1.246882e-03	0.065461279
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	1.246882e-03
hsa03440 Homologous recombination	28	3.066756e-03
hsa04114 Oocyte meiosis	102	3.784520e-03

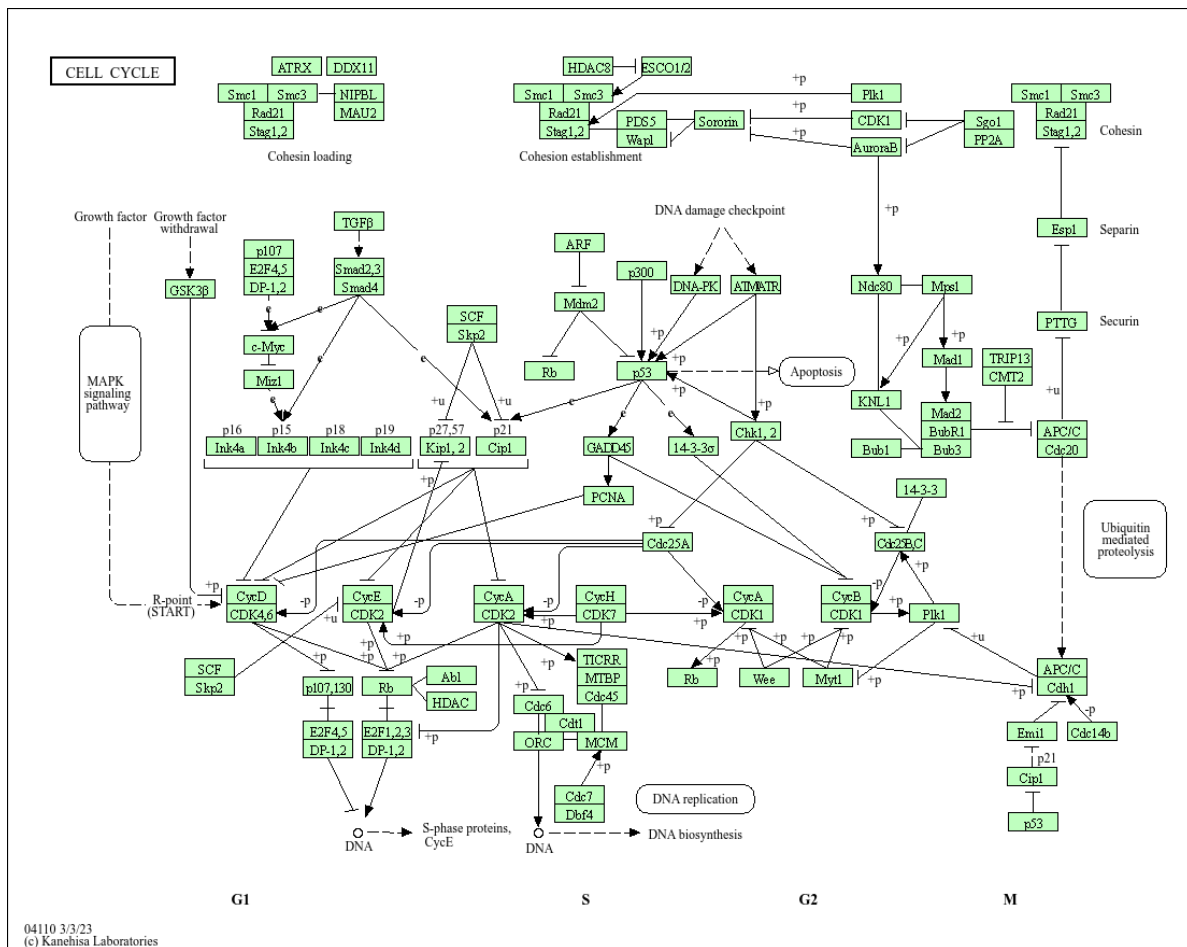
Cell cycle figure

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

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

Info: Working in directory C:/Users/ilyas/OneDrive/Desktop/bimm143Wlstuff/class14.0

Info: Writing image file hsa04110.pathview.png



DNA Replication Figure

```
pathview(foldchanges, pathway.id = "hsa03030")
```

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

Info: Working in directory C:/Users/ilyas/OneDrive/Desktop/bimm143Wistuff/class14.0

Info: Writing image file hsa03030.pathview.png

Gene Ontology analysis

Run pathway analysis with GO

```

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
G0:0007156 homophilic cell adhesion	8.519724e-05	3.824205	8.519724e-05
G0:0002009 morphogenesis of an epithelium	1.396681e-04	3.653886	1.396681e-04
G0:0048729 tissue morphogenesis	1.432451e-04	3.643242	1.432451e-04
G0:0007610 behavior	1.925222e-04	3.565432	1.925222e-04
G0:0060562 epithelial tube morphogenesis	5.932837e-04	3.261376	5.932837e-04
G0:0035295 tube development	5.953254e-04	3.253665	5.953254e-04

	q.val	set.size	expl
G0:0007156 homophilic cell adhesion	0.1952430	113	8.519724e-05
G0:0002009 morphogenesis of an epithelium	0.1952430	339	1.396681e-04
G0:0048729 tissue morphogenesis	0.1952430	424	1.432451e-04
G0:0007610 behavior	0.1968058	426	1.925222e-04
G0:0060562 epithelial tube morphogenesis	0.3566193	257	5.932837e-04
G0:0035295 tube development	0.3566193	391	5.953254e-04

\$less

	p.geomean	stat.mean	p.val
G0:0048285 organelle fission	1.536227e-15	-8.063910	1.536227e-15
G0:0000280 nuclear division	4.286961e-15	-7.939217	4.286961e-15
G0:0007067 mitosis	4.286961e-15	-7.939217	4.286961e-15
G0:0000087 M phase of mitotic cell cycle	1.169934e-14	-7.797496	1.169934e-14
G0:0007059 chromosome segregation	2.028624e-11	-6.878340	2.028624e-11
G0:0000236 mitotic prometaphase	1.729553e-10	-6.695966	1.729553e-10

	q.val	set.size	expl
G0:0048285 organelle fission	5.843127e-12	376	1.536227e-15
G0:0000280 nuclear division	5.843127e-12	352	4.286961e-15
G0:0007067 mitosis	5.843127e-12	352	4.286961e-15
G0:0000087 M phase of mitotic cell cycle	1.195965e-11	362	1.169934e-14
G0:0007059 chromosome segregation	1.659009e-08	142	2.028624e-11
G0:0000236 mitotic prometaphase	1.178690e-07	84	1.729553e-10

\$stats

	stat.mean	exp1
G0:0007156 homophilic cell adhesion	3.824205	3.824205
G0:0002009 morphogenesis of an epithelium	3.653886	3.653886
G0:0048729 tissue morphogenesis	3.643242	3.643242
G0:0007610 behavior	3.565432	3.565432
G0:0060562 epithelial tube morphogenesis	3.261376	3.261376
G0:0035295 tube development	3.253665	3.253665

```
head(keggres$greater)
```

	p.geomean	stat.mean
hsa04060 Cytokine-cytokine receptor interaction	9.131044e-06	4.358967
hsa05323 Rheumatoid arthritis	1.809824e-04	3.666793
hsa05146 Amoebiasis	1.313400e-03	3.052596
hsa05332 Graft-versus-host disease	2.605234e-03	2.948229
hsa04640 Hematopoietic cell lineage	2.822776e-03	2.833362
hsa04630 Jak-STAT signaling pathway	5.202070e-03	2.585673

	p.val	q.val
hsa04060 Cytokine-cytokine receptor interaction	9.131044e-06	0.001917519
hsa05323 Rheumatoid arthritis	1.809824e-04	0.019003147
hsa05146 Amoebiasis	1.313400e-03	0.091937999
hsa05332 Graft-versus-host disease	2.605234e-03	0.118556573
hsa04640 Hematopoietic cell lineage	2.822776e-03	0.118556573
hsa04630 Jak-STAT signaling pathway	5.202070e-03	0.182072434

	set.size	exp1
hsa04060 Cytokine-cytokine receptor interaction	177	9.131044e-06
hsa05323 Rheumatoid arthritis	72	1.809824e-04
hsa05146 Amoebiasis	94	1.313400e-03
hsa05332 Graft-versus-host disease	22	2.605234e-03
hsa04640 Hematopoietic cell lineage	55	2.822776e-03
hsa04630 Jak-STAT signaling pathway	109	5.202070e-03

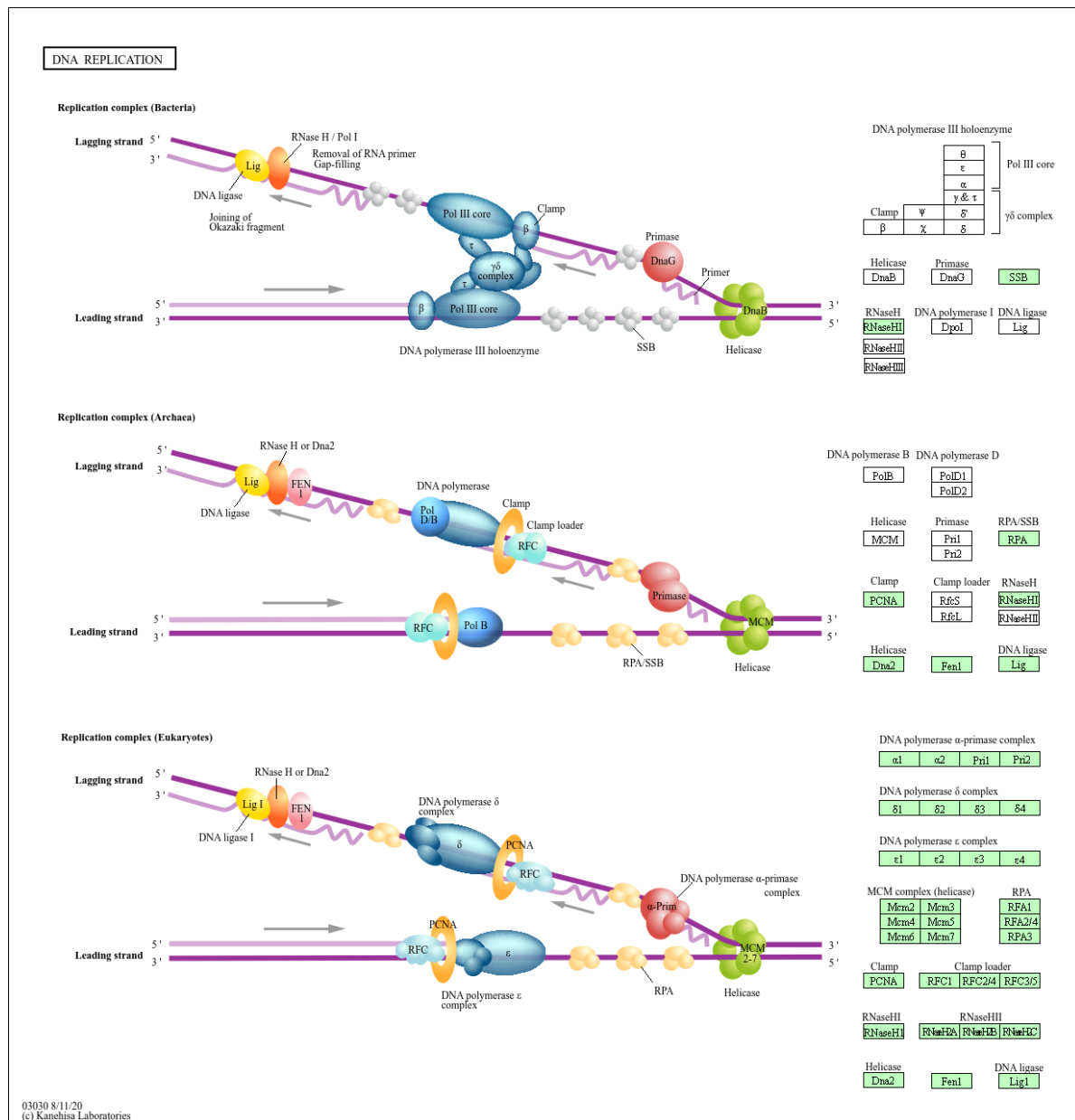
DNA Replication Figure

```
pathview(foldchanges, pathway.id = "hsa03030")
```

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

Info: Working in directory C:/Users/ilyas/OneDrive/Desktop/bimm143Wistuff/class14.0

Info: Writing image file hsa03030.pathview.png



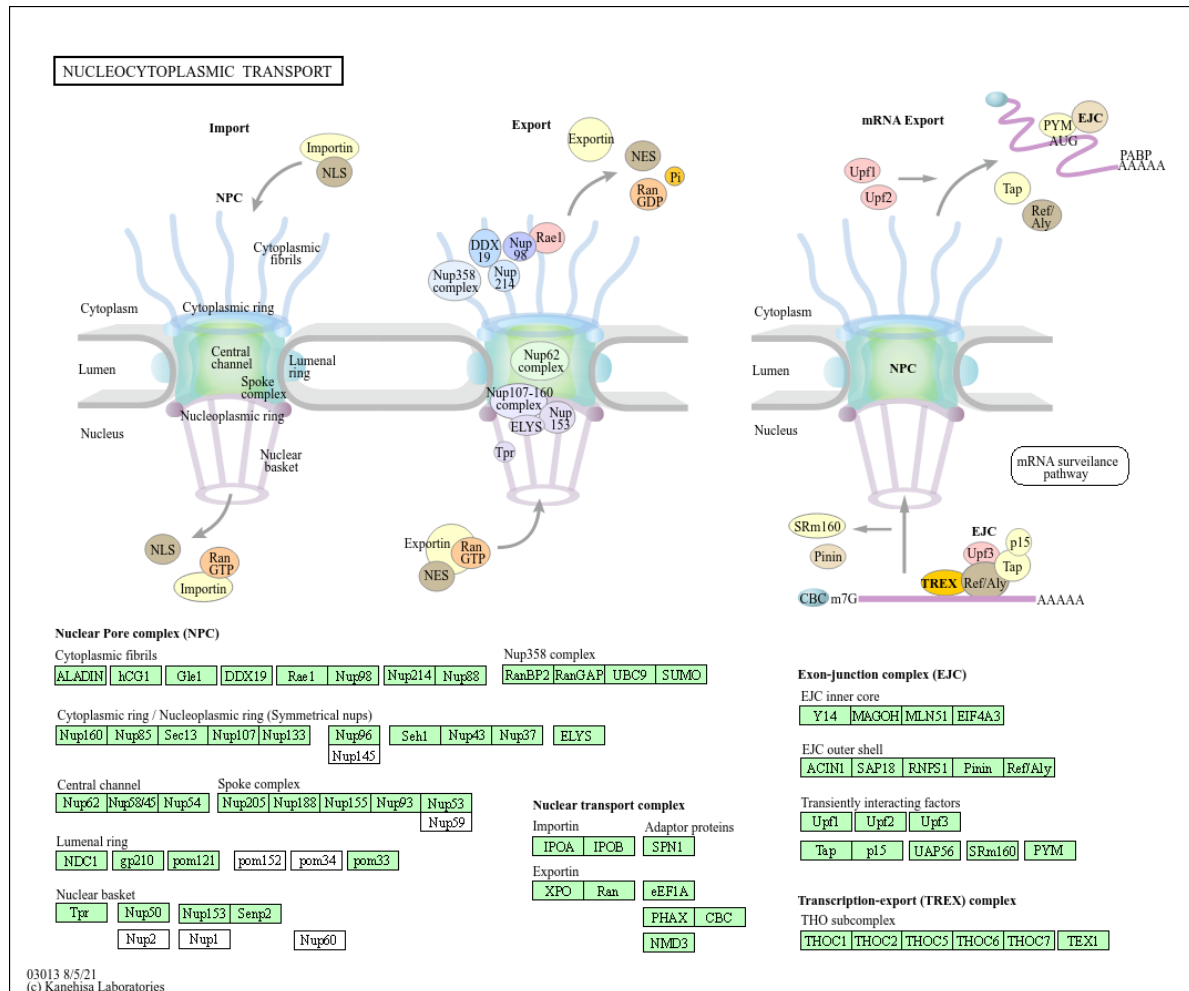
RNA transport figure

```
pathview(foldchanges, pathway.id = "hsa03013")
```

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

Info: Working in directory C:/Users/ilyas/OneDrive/Desktop/bimm143Wistuff/class14.0

Info: Writing image file hsa03013.pathview.png



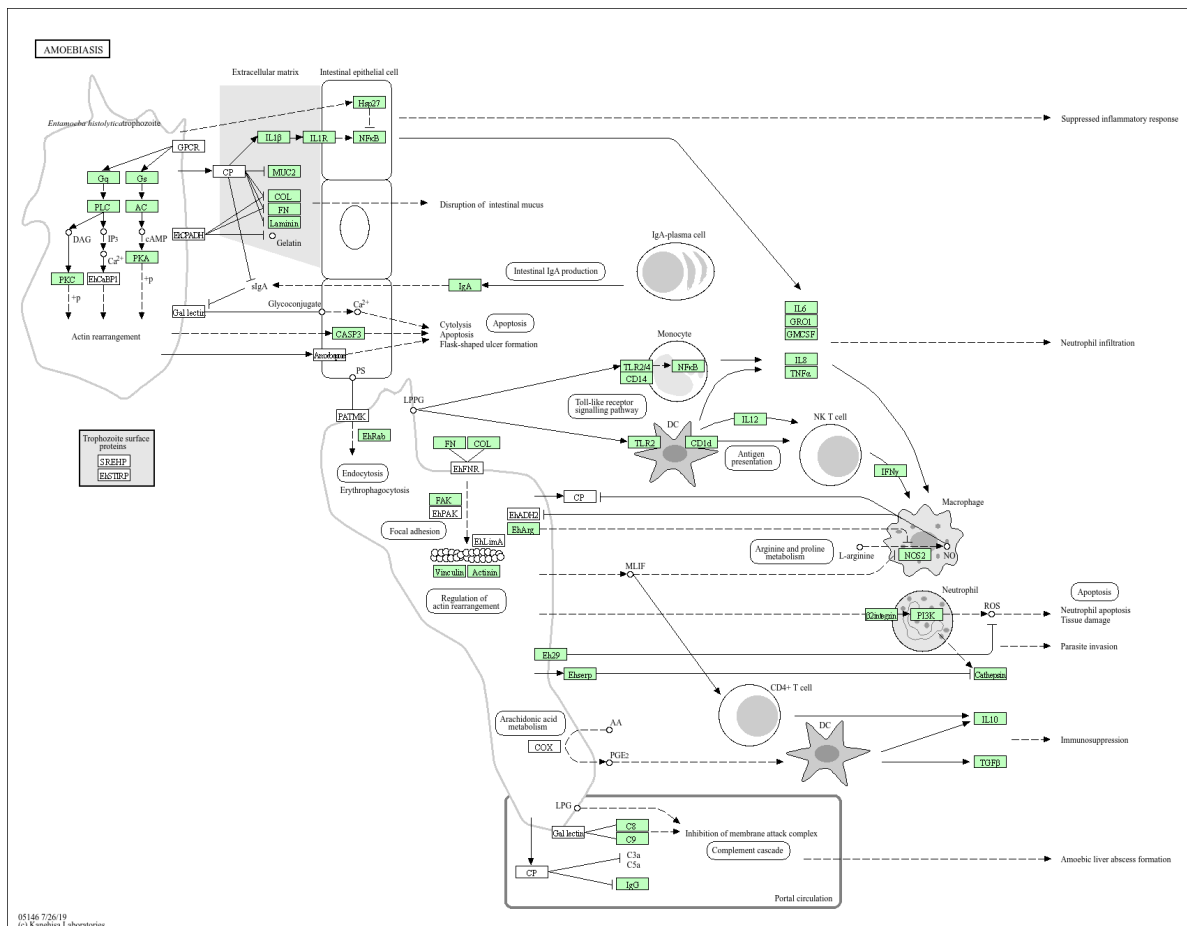
Amoebiasis figure

```
pathview(foldchanges, pathway.id = "hsa05146")
```

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

Info: Working in directory C:/Users/ilyas/OneDrive/Desktop/bimm143Wistuff/class14.0

Info: Writing image file hsa05146.pathview.png



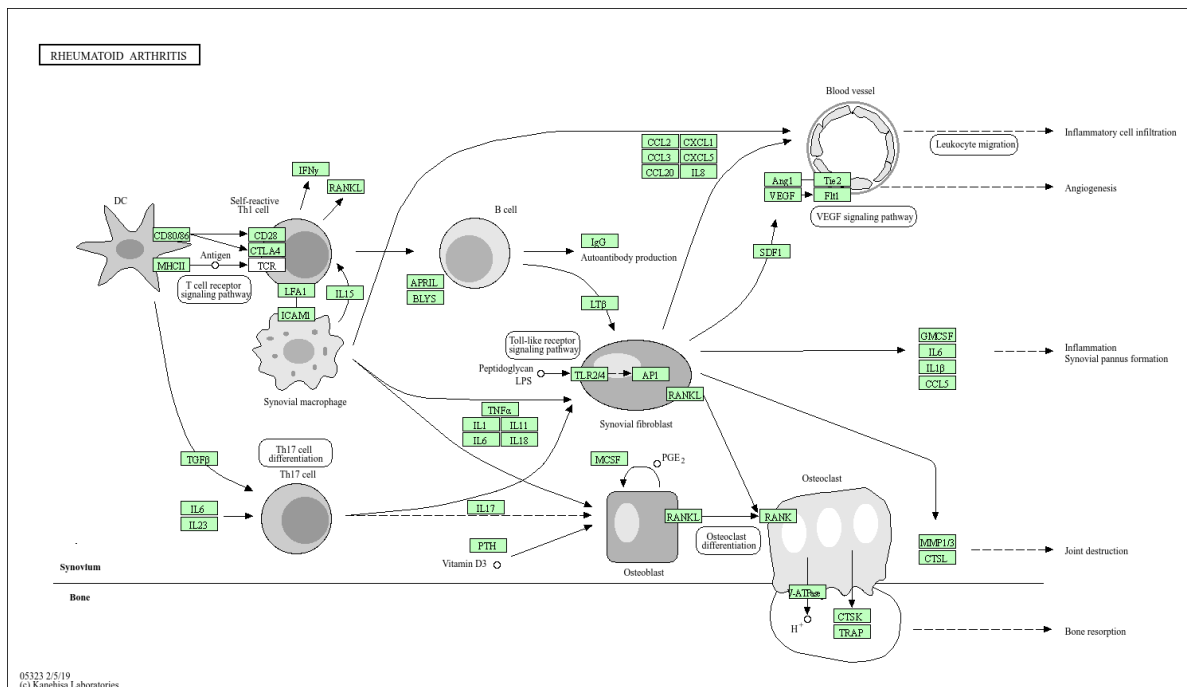
Rheumatoid arthritis figure

```
pathview(foldchanges, pathway.id = "hsa05323")
```

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

Info: Working in directory C:/Users/ilyas/OneDrive/Desktop/bimm143Wistuff/class14.0

Info: Writing image file hsa05323.pathview.png



```
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=)
```

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? - Cell Cycle it does match the kegg results top value matches the previous go looks at gene function at a standard basic is just a basic way of looking at the pathways kegg looks at gene interaction through a biological pathway in a more complex way