

Class 14: RNASeq Mini Project

Christopher Brockie (PID: A16280405)

About our input data

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

Data Import

First we must call the DESeq package and load our data files

```
#Call DESeq2
library(DESeq2)

#Read and assign data files
metaFile <- "GSE37704_metadata.csv"
countFile <- "GSE37704_featurecounts.csv"

# Import files
colData = read.csv(metaFile, row.names=1)
countData = read.csv(countFile, row.names=1)
```

Remember that we need the `countData` and `colData` files to match up so we will need to remove that odd first column in `countData` namely `countData$length`.

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

```
# Note we need to remove the odd first $length col
countData <- as.matrix(countData[,-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

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).

```
# Filter count data where you have 0 read count across all samples.
countData = countData[rowSums(countData != 0) > 0, ]
head(countData)
```

	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

Q. How many genes do we have left?

```
nrow(countData)
```

```
[1] 15975
```

DESeq setup and analysis

Now lets setup the DESeqDataSet object required for the DESeq() function and then run the DESeq pipeline.

```
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

```
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(2): condition sizeFactor
```

Next, get results for the HoxA1 knockdown versus control siRNA (remember that these were labeled as “hoxa1_kd” and “control_sirna” in our original colData metaFile input to DESeq, you can check this above and by running `resultsNames(dds)` command).

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

```
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				

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.

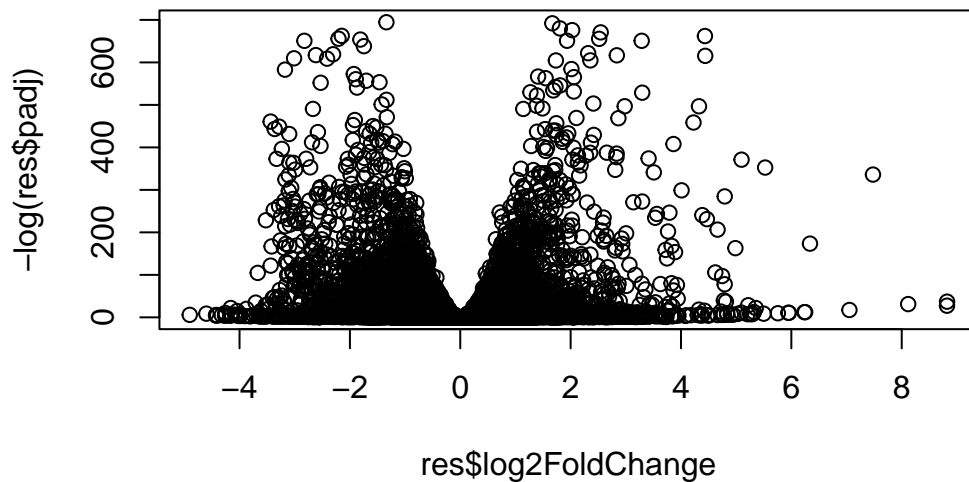
```
summary(res)
```

```
out of 15975 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)      : 4349, 27%
LFC < 0 (down)    : 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
```

Volcano plot

Now we will make a volcano plot, it's a plot of log2 fold change vs -log adjusted p-value.

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



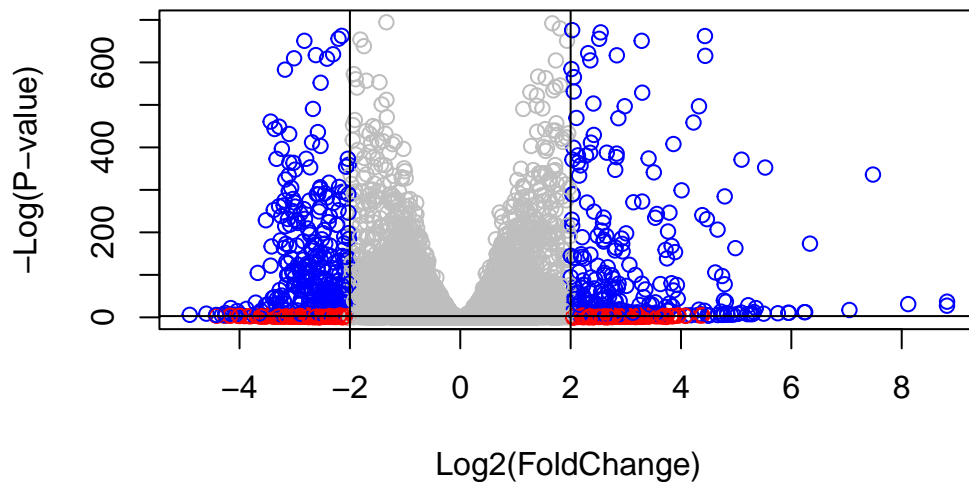
Q. Improve this plot by completing the below code, which adds color and axis labels

```
# 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(
abline(v=c(-2,2))
abline(h=-log(0.05))
```



Adding gene annotation

Since we mapped and counted against the Ensembl annotation, our results only have information about Ensembl gene IDs. However, our pathway analysis downstream will use KEGG pathways, and genes in KEGG pathways are annotated with Entrez gene IDs. We need to add them.

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

```
library("AnnotationDbi")
```

Warning: package 'AnnotationDbi' was built under R version 4.3.2

```
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>
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 ..	

Save 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="myresults.csv")
```

Geneset enrichment

I will use KEGG and GO...

```
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"

```

Make my input vector of importance for `gage()`

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

```
<NA>      148398      26155      339451      84069      84808
0.17925708 0.42645712 -0.69272046 0.72975561 0.04057653 0.54281049
```

Run `gage()` with `kegg.sets.hs`

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

```
# Look at the first few down (less) pathways
head(keggres$less, 3)
```

		p.geomean	stat.mean	p.val	q.val
hsa04110	Cell cycle	8.995727e-06	-4.378644	8.995727e-06	0.001448312
hsa03030	DNA replication	9.424076e-05	-3.951803	9.424076e-05	0.007586381
hsa03013	RNA transport	1.246882e-03	-3.059466	1.246882e-03	0.066915974

		set.size	exp1
hsa04110	Cell cycle	121	8.995727e-06
hsa03030	DNA replication	36	9.424076e-05
hsa03013	RNA transport	144	1.246882e-03

The top two here (hsa04110 and hsa03030) appear to be the main sets picked out. I will now use `pathview()` to pull these pathways and color up my genes that intersect with these two pathways

```
pathview(gene.data=foldchanges, pathway.id="hsa04110")
pathview(gene.data=foldchanges, pathway.id="hsa03030")
```

And insert into my report here:

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

```
## Focus on top 5 down-regulated pathways
keggrespathways <- rownames(keggres$less)[1:5]
```

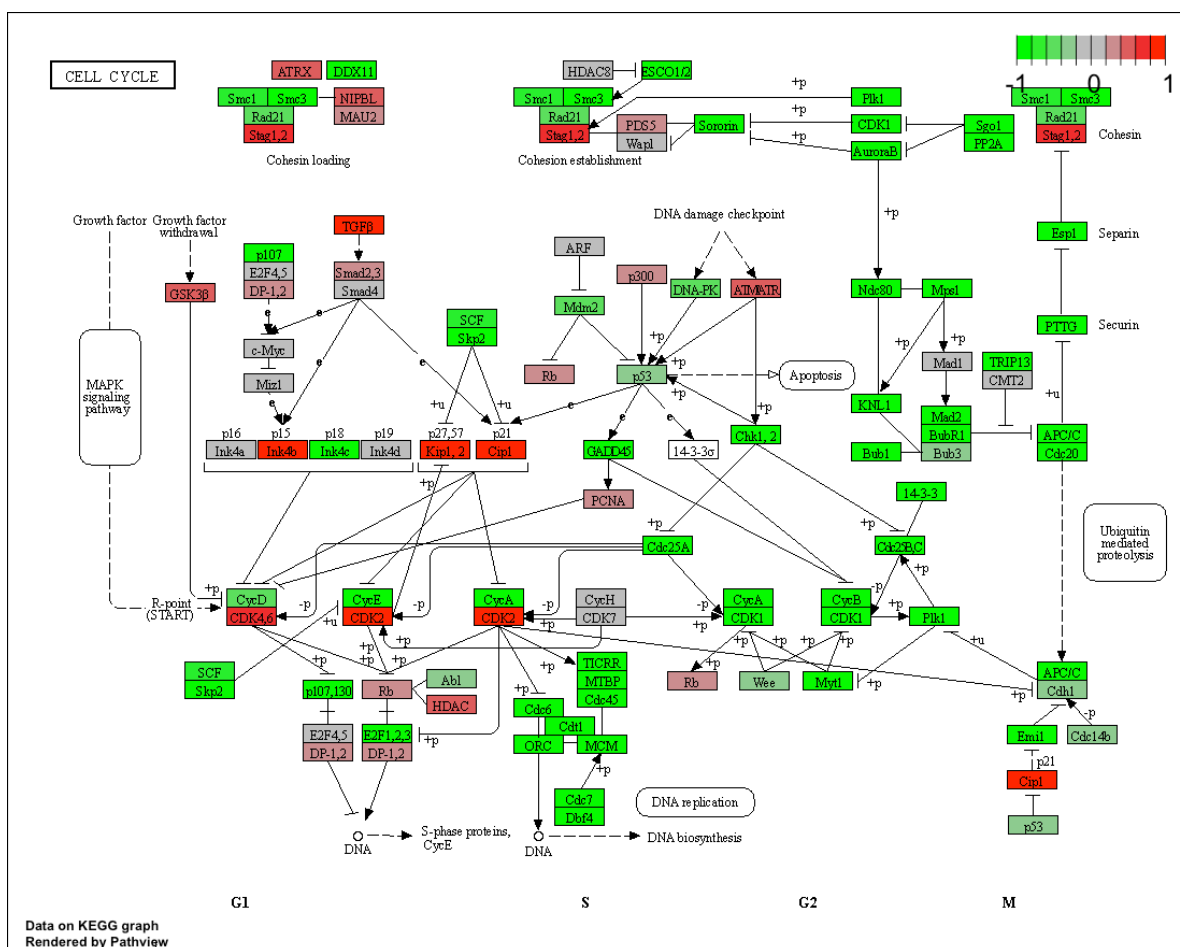


Figure 1: Cell cycle genes

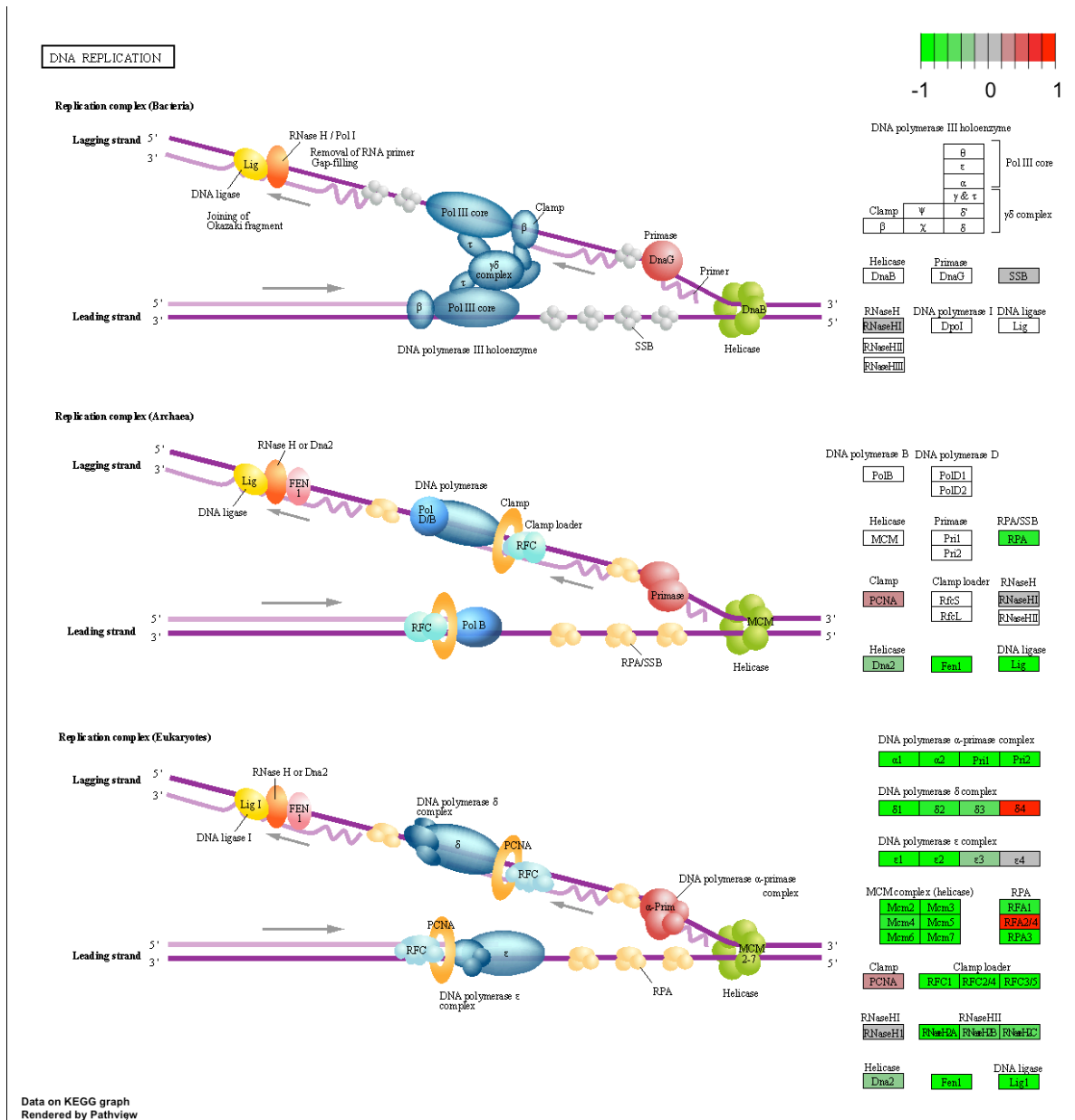


Figure 2: DNA replication genes

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

```
[1] "hsa04110" "hsa03030" "hsa03013" "hsa03440" "hsa04114"
```

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

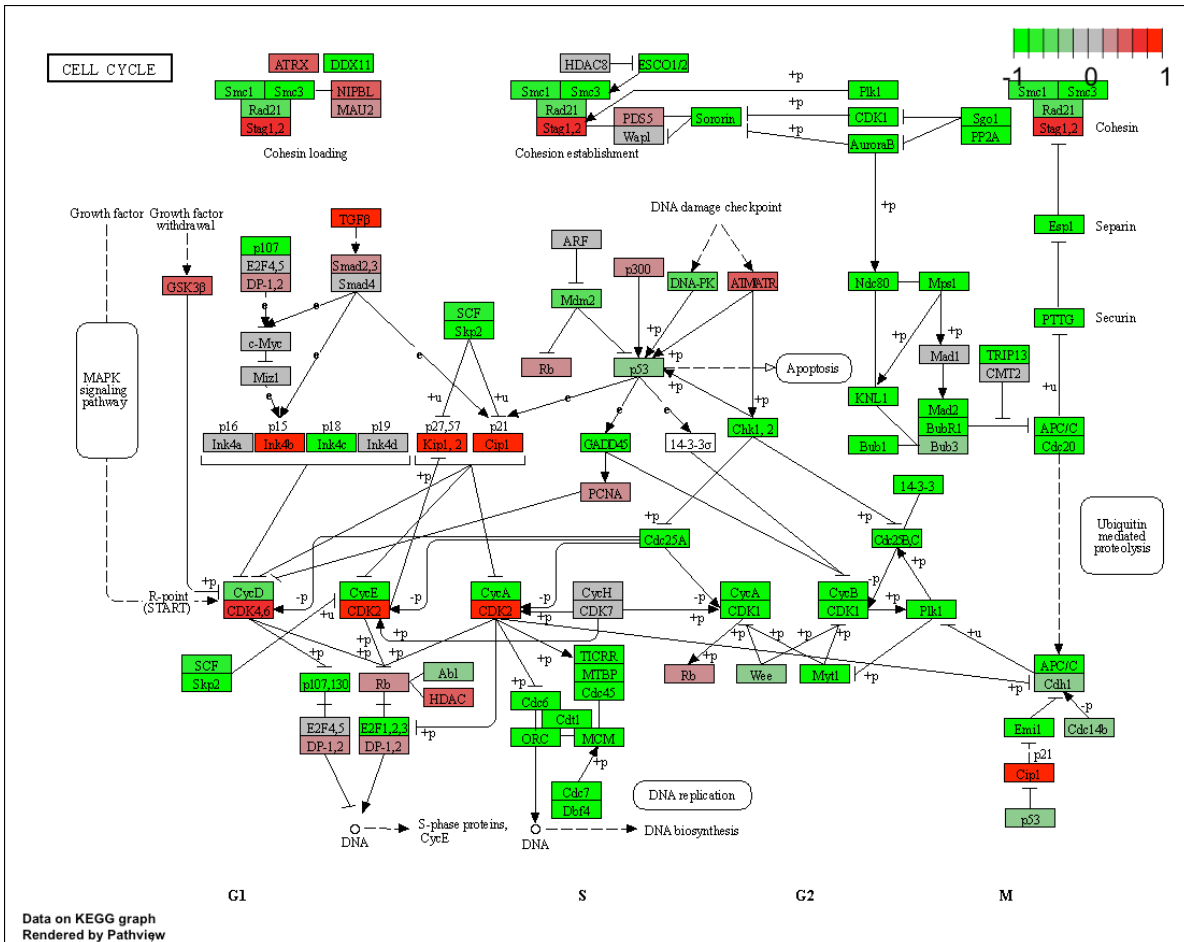


Figure 3: Top 5 down-regulated pathways

Gene Ontology

We can also do a similar procedure with gene ontology. Similar to above, `go.sets.hs` has all GO terms. `go.subs.hs` is a named list containing indexes for the BP, CC, and MF ontologies. Let's focus on BP (a.k.a Biological Process) here.

```
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	exp1
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	exp1
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

Reactome Analysis

We can use reactome as either its (original) R package or via its newer online web-server. The latter has some potentially useful pathway viewing functionality so let's try it out online (<https://reactome.org/>)

To use it online we need a list of significant genes at the $\alpha < 0.05$ level as a plain text file. We can make this in R like this:

```
sig_genes <- res[res$padj <= 0.05 & !is.na(res$padj), "symbol"]

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

Now upload this file to the reactome website (<https://reactome.org/PathwayBrowser/#TOOL=AT>)

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 mitotic cell cycle pathway has the most significant p-value at 5.28E-4. Some of the most significant pathways match the KEGG results, but some do not. This may be because some of the nonmatching KEGG results are involved in the cell cycle, but are not explicitly linked to it.

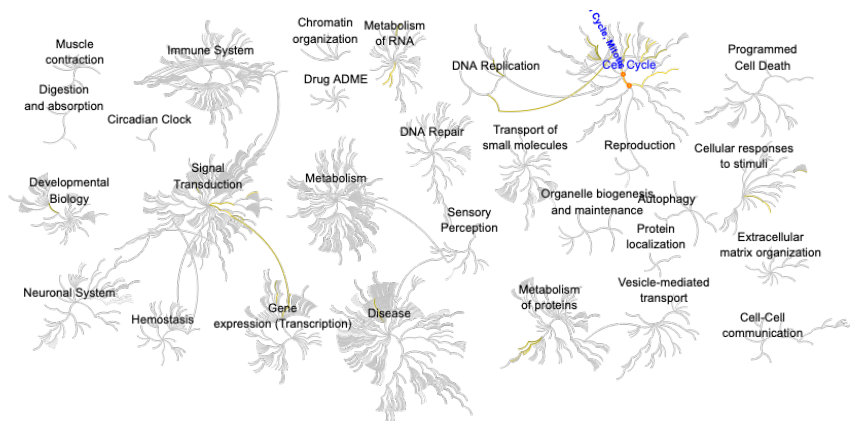


Figure 4: Highlighted significant pathways from Reactome pathway browser

GO Online

Gene Set Gene Ontology (GO) Enrichment is a method to determine over-represented or under-represented GO terms for a given set of genes. GO terms are formal structured controlled vocabularies (ontologies) for gene products in terms of their biological function. The goal of this analysis is to determine the biological process the given set of genes are associated with.

Upload the significant gene list into the Gene Set GO Enrichment website (<http://www.geneontology.org/page/go-enrichment-analysis>)

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 most significant pathway appears to be “regulation of cell migration involved in sprouting angiogenesis”. This can be loosely matched with the KEGG results involving morphogenesis, but does not exactly match the results. This difference could be due to the genes from the KEGG results participating in a wide variety of cellular processes, resulting in enrichment GO to yield wider and more specific results.

	Homo sapiens (REF)		upload_1 (▼ Hierarchy NEW! ?)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
regulation of cell migration involved in sprouting angiogenesis	38	33	15.11	2.18	+	1.37E-03	3.05E-02
↳regulation of blood vessel endothelial cell migration	91	69	36.18	1.91	+	6.90E-05	2.25E-03
↳regulation of endothelial cell migration	169	119	67.18	1.77	+	2.71E-06	1.23E-04
↳regulation of epithelial cell migration	229	156	91.04	1.71	+	3.52E-07	1.84E-05
↳regulation of multicellular organismal process	2975	1470	1182.66	1.24	+	2.05E-13	2.52E-11
↳regulation of biological process	11752	5297	4671.81	1.13	+	6.57E-33	3.52E-30
↳biological regulation	12172	5456	4838.77	1.13	+	1.11E-32	5.74E-30
↳regulation of cell migration	932	544	370.50	1.47	+	8.07E-13	9.28E-11
↳regulation of cell motility	990	565	393.56	1.44	+	4.37E-12	4.55E-10
↳regulation of locomotion	1035	580	411.45	1.41	+	2.28E-11	2.17E-09
↳regulation of cellular process	10986	4998	4367.30	1.14	+	1.26E-32	6.33E-30
platelet-derived growth factor receptor signaling pathway	36	31	14.31	2.17	+	2.51E-03	4.97E-02
↳transmembrane receptor protein tyrosine kinase signaling pathway	423	255	168.16	1.52	+	2.13E-07	1.15E-05
↳enzyme-linked receptor protein signaling pathway	639	365	254.02	1.44	+	3.18E-08	2.00E-06
↳cell surface receptor signaling pathway	2090	955	830.84	1.15	+	1.85E-04	5.31E-03
↳cellular response to stimulus	6449	2871	2563.69	1.12	+	9.59E-10	7.64E-08
↳cellular process	14613	6588	5809.15	1.13	+	1.02E-63	5.27E-60
↳response to stimulus	8207	3482	3262.55	1.07	+	3.13E-05	1.11E-03
negative regulation of metaphase/anaphase transition of cell cycle	35	30	13.91	2.16	+	2.32E-03	4.68E-02
↳negative regulation of cell cycle phase transition	228	159	90.64	1.75	+	8.14E-08	4.83E-06
↳regulation of cell cycle phase transition	434	293	172.53	1.70	+	3.65E-12	3.86E-10
↳regulation of cell cycle process	725	470	288.21	1.63	+	1.44E-16	2.40E-14

Figure 5: GO Enrichment Results from Significant Genes