

Class 14: RNA-Seq Mini Project

Josefina O'Toole (PID: A16978557)

Table of contents

Background	1
Data Import	2
Inspect and tidy data	2
Setup for DESeq	4
Run DESeq	6
Volcano plot of results	7
Gene annotation	8
Pathway Analysis	10
Gene Ontology analysis	15

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")
```

```
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					

Inspect and tidy data

Does the counts columns match the colData rows?

```
head(colData)
```

	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

```
colnames(counts)
```

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

The fix here looks to be removing the first “length” column from counts:

```
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 countData and colData

```
colnames(countData) == colData$id
```

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

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?

```
table(rowSums(countData) == 0)
```

```
FALSE  TRUE
15975  3833
```

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

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

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

```
[1] 15975
```

Setup for DESeq

```
library(DESeq2)
```

```
Loading required package: S4Vectors
```

```
Loading required package: stats4
```

```
Loading required package: BiocGenerics
```

```
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, intersect, is.unsorted, lapply, Map, mapply,  
  match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,  
  Position, rank, rbind, Reduce, rownames, sapply, saveRDS, setdiff,  
  table, tapply, union, 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: GenomeInfoDb

Loading required package: SummarizedExperiment

Loading required package: MatrixGenerics

Loading required package: matrixStats

Attaching package: 'MatrixGenerics'

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

colAlls, colAnyNAs, colAnys, colAvgPerRowSet, 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, rowAvgPerColSet,
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

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	1.43989e-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	0.5215599	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				

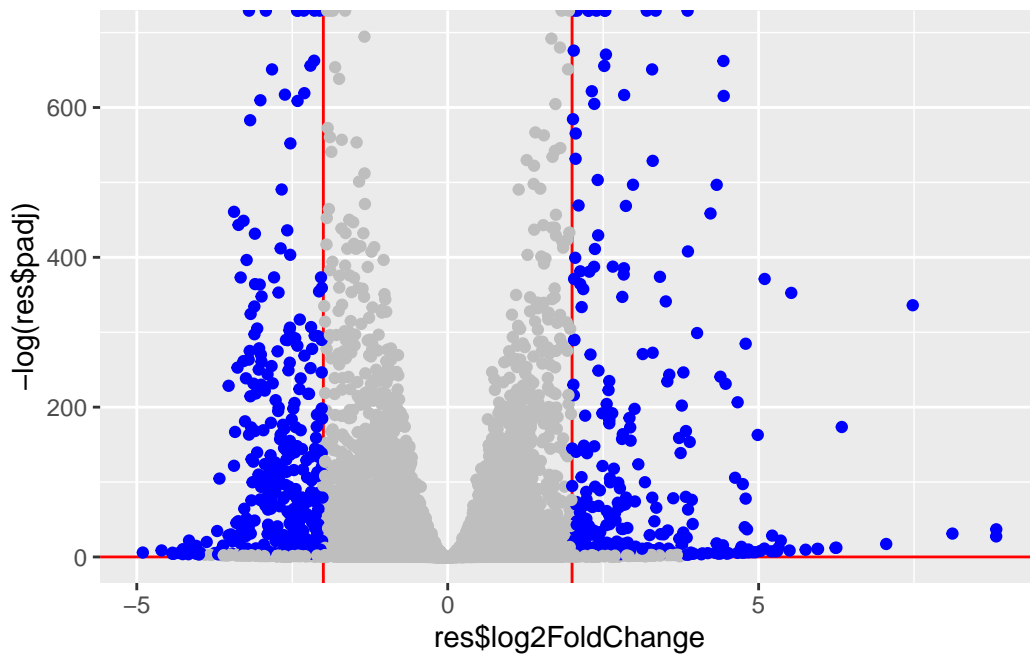
Volcano plot of results

```
library(ggplot2)
```

```
mycols <- rep("gray", nrow(res))
mycols[res$log2FoldChange >= 2] <- "blue"
mycols[res$log2FoldChange <= -2] <- "blue"
mycols[res$padj > 0.05] <- "gray"
```

```
ggplot(res) +
  aes(res$log2FoldChange, -log(res$padj)) +
  geom_vline(xintercept= c(-2,2), col="red") +
  geom_hline(yintercept=0.05, col="red") +
  geom_point(col=mycols)
```

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",
                    multiVals="first")
```

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

```
res$entrez = mapIds(org.Hs.eg.db,
                    keys=rownames(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="ENTREZID",
#                  multiVals="first")

#head(res, 10)
```

```
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 8 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	1.43989e-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	0.5215599	1.040744	2.97994e-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		

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
head(foldchanges)
```

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

Load up the KEGG genesets

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

Run pathway analysis

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

```
head(keggres$less, 3)
```

		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

		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

		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

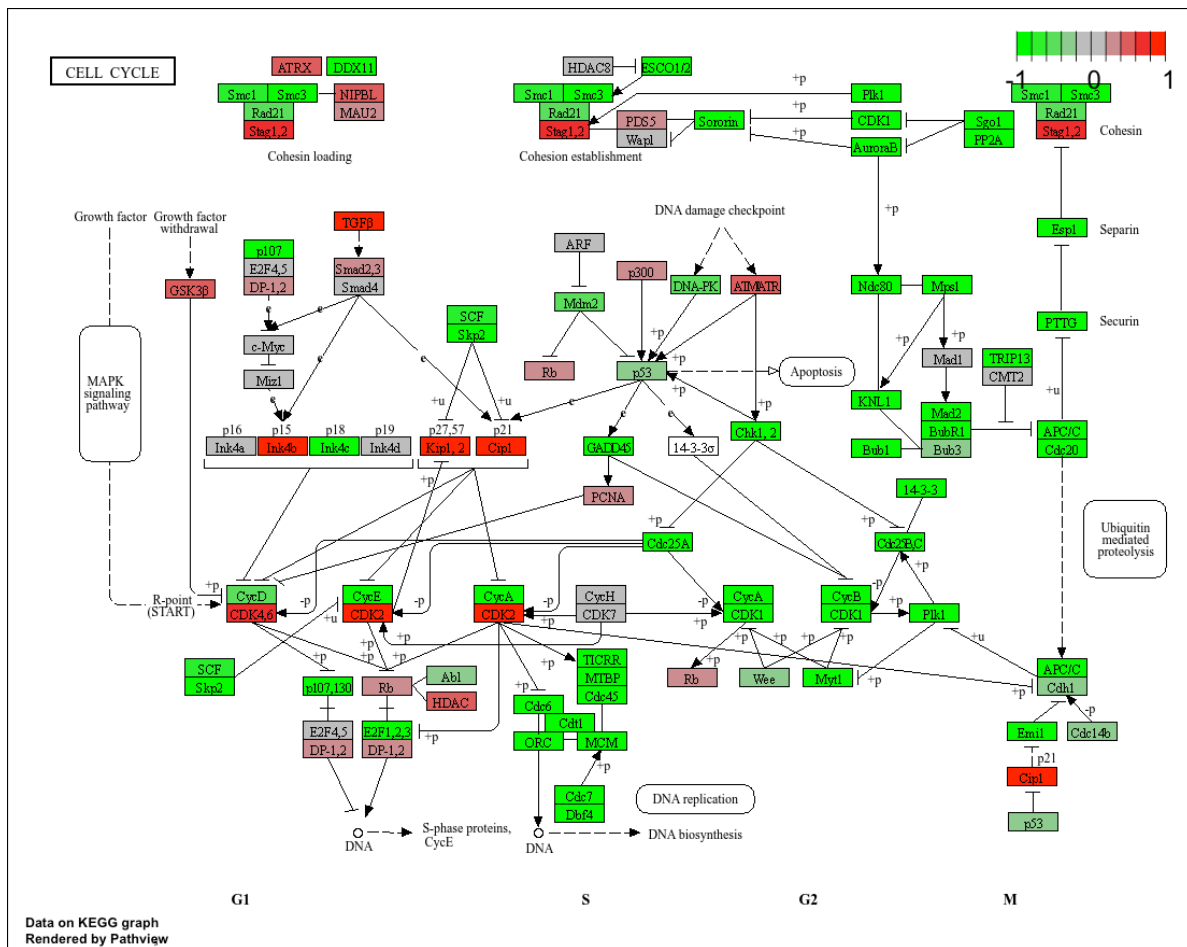
Cell cycle figure

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

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

Info: Working in directory /Users/josefinaotoole/Desktop/BIMM143_BioinformaticsLab/Class 14:

Info: Writing image file hsa04110.pathview.png



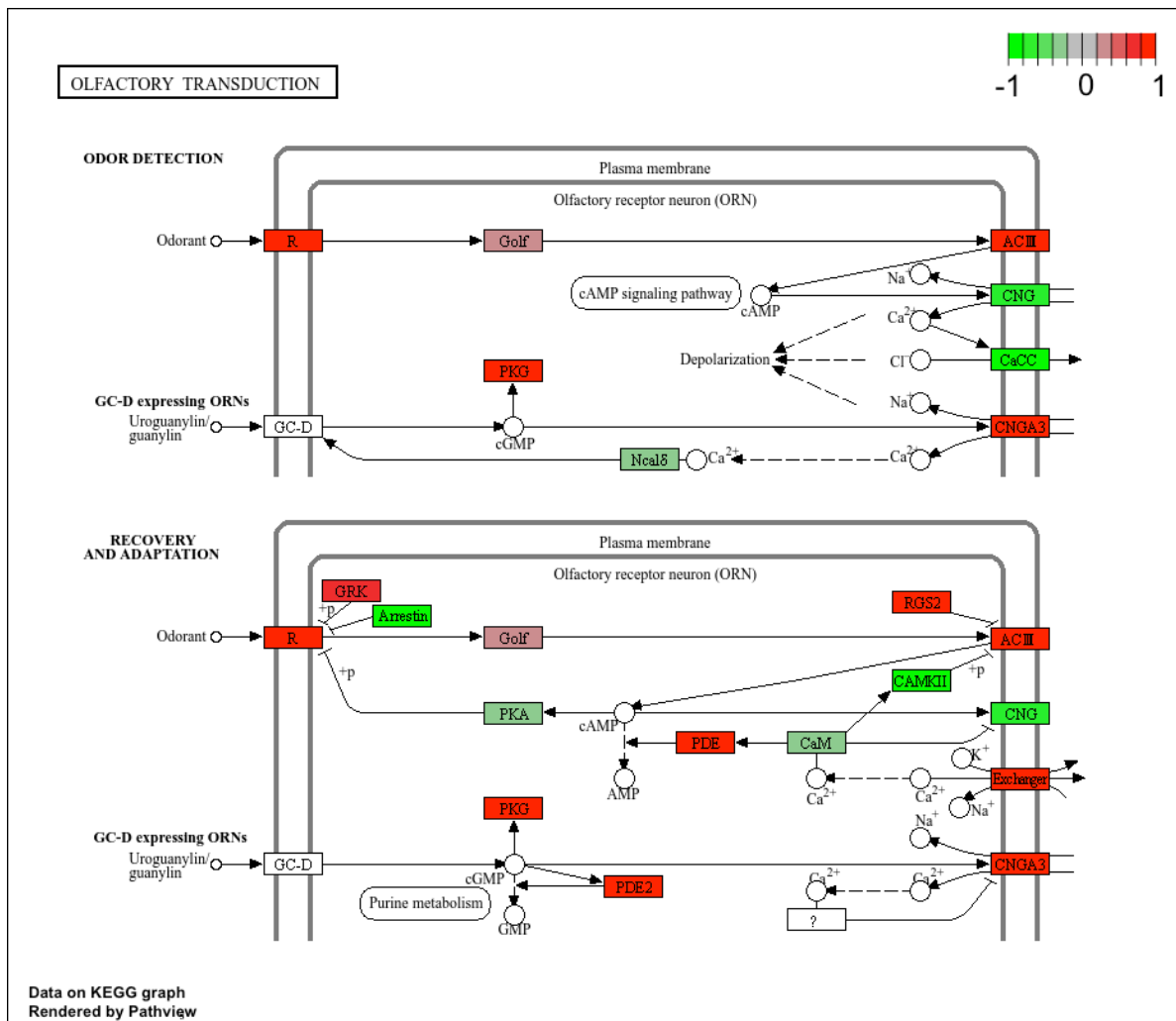
Olfactory Transduction

```
pathview(foldchanges, pathway.id = "hsa04740")
```

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

Info: Working in directory /Users/josefinaotoole/Desktop/BIMM143_BioinformaticsLab/Class 14:

Info: Writing image file hsa04740.pathview.png



Hematopoietic Cell Lineage

```
pathview(foldchanges, pathway.id = "hsa04640")
```

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

Info: Working in directory /Users/josefinaotoole/Desktop/BIMM143_BioinformaticsLab/Class 14:

Info: Writing image file hsa04640.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)

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	5.841698e-12	376	1.536227e-15
GO:0000280 nuclear division	5.841698e-12	352	4.286961e-15
GO:0007067 mitosis	5.841698e-12	352	4.286961e-15
GO:0000087 M phase of mitotic cell cycle	1.195672e-11	362	1.169934e-14
GO:0007059 chromosome segregation	1.658603e-08	142	2.028624e-11
GO:0000236 mitotic prometaphase	1.178402e-07	84	1.729553e-10