

## Table of Contents

<b>NIPA.....</b>	<b>1</b>
<b>Code availability.....</b>	<b>1</b>
hypergeometric tests.....	1
<b>User input to change .....</b>	<b>2</b>
Required Packages.....	2
User options .....	2
<b>Output .....</b>	<b>3</b>
<b>Gene Ontology (Enrichment using GOstats).....</b>	<b>3</b>
Output Tables .....	3
Enriched GO figures .....	3
GO Directed acyclic graph .....	4
<b>Pathways (KEGG).....</b>	<b>6</b>
<b>NIPA.report.txt: .....</b>	<b>7</b>

## NIPA

### Code availability

<https://github.com/ADAC-UoN/NIPA>

Code to determine enriched Gene Ontology and Pathways using KEGG using hypergeometric tests.

### hypergeometric tests

using phyper

phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)

x, q vector of quantiles representing the number of white balls drawn without replacement from an urn which contains both black and white balls.

m the number of white balls in the urn.

n the number of black balls in the urn.

k the number of balls drawn from the urn.

if

pop size : 5260 # total number of entrez gene in all pathways

sample size : 131 # total goi

Number of items in the pop that are classified as successes : 1998 # entrez in a particular pathway

Number of items in the sample that are classified as successes : 62 # goi in a particular pathway

e.g pathway 100 genes 10 are in goi list of size 400 universe = 20,000

phyper(1,100,20000-100,400, lower.tail=FALSE) = 0.597 = probability of finding this many or greater goi in pathway  
 phyper(80,100,20000-100,400, lower.tail=FALSE) = 4.603708e-122 = probability of finding this many or greater goi in pathway

## User input to change

Required Packages

See Bioconductor for instructions to install.

```
library(biomaRt)
library(pathview)
library(gage)
library(gageData)
library(ggplot2)
library(stringr)
library(dplyr)
library(RamiGO)
```

## User options

The code in lines 25-65 will need editing

```
#####
## Input Variables -- USER TO CHANGE [START]
## Check all or may fail.
#####
goi.column = 2          # if results are from analysis and are a column of a larger table give input column else will assume is column 1 or a single column as
goi.header = "yes"      # "yes" or "no" if header on file

species = "mouse"       #currently one of "mouse", "human", "rat", "pig", "zebrafish, cow, fly, sheep",
outfile.prefix <- "V3.vs.V2" # prefix attached to output files.

# colour pathways by expression fold change?
keggFC = "yes"          # yes or no. will colour enriched KEGG pathways by FC data [specify column below]
keggFC.col = 5          # if keggFC = yes specify column of input table with FC values assumes tab delimited

id.type = "Entrez"      # one of
                        # "ENSG" (ensembl gene),
                        # "ENST" (ensembl trasncript),
                        # "ENSP" (ensembl peptide),
                        # "Entrez"
                        # "Uniprot" (UniProt/SwissProt Accession)
                        # "Unigene"
                        # "Refseq_mrna" (RefSeq mRNA [e.g. NM_001195597])
                        # "Refseq_peptide" (RefSeq Protein ID [e.g. NP_001005353])
                        # "hgnc" (HGNC ID [e.g. LIS1])

# set variables for hypergeometric cutoff enrichment qval less than this and with greater or equal to minimum number of genes in pathway or GO term will be drawn
kegg.qval.cutoff = 0.05
GO.cutoff = 0.05        # qvalue cutoff
min.genes.cutoff = 2

# change below to determine which test to conduct.
doGO = "yes"            # yes or no. Run hypergeometric test to find enriched GO terms in BP, MF and CC category
doKEGG = "yes"          # yes or no. Run hypergeometric test to find and plot enriched KEGG pathways and visualise using PathView

#####
## Input Variables -- USER TO CHANGE [END]
#####
```

## Options

goi.column	number of column with gene identifier in starts form 1. If a simple list, keep as = 1.
goi.header = "yes"	# "yes" or "no" if header on file
species = "mouse"	#currently one of "mouse", "human", "rat", "pig", "zebrafish, cow, fly, sheep",
outfile.prefix <- "ADAC.test"	prefix attached to output files. Change to user requirements.
keggFC = "yes"	yes or no. If yes will colour enriched KEGG pathways by FC data [specify column below]

keggFC.col = 5	if keggFC = yes specify column of input table with FC values assumes tab delimited
id.type = "Entrez"	change to correct accepted identifier type
kegg.qval.cutoff = 0.05	qvalue cutoff for KEGG enrichment analysis
GO.cutoff = 0.05	qvalue cutoff for GO enrichment analysis
min.genes.cutoff = 2	Minimum number of genes in category to be reported
doGO = "yes"	# yes or no. Run hypergeometric test to find enriched GO terms in BP, MF and CC category
doKEGG = "yes"	# yes or no. Run hypergeometric test to find and plot enriched KEGG pathways and visualise using PathView

## Output

### Gene Ontology (Enrichment using GOSTats)

#### Output Tables

GO enriched terms GO.BP.table, GO.CC.table, GO.MF.table for "Biological Process", "Cellular Component" and "Molecular Function" respectively. e.g. for Biological Process Columns are

GO ID,

All\_gene\_count (number of genes in GO term)

GOI\_gene\_count (number of genes in GO term also in user input "genes of interest" list)

Pval (p value from hypergeometric test)

Qval (Benjamini-Hochberg corrected pvalue)

This is a tab-delimited text file that can open in excel or similar.

A	B	C	D	E	F	G	H
GO_ID	ALL_gene_count	GOI_gene_ids	GOI_gene_count	GO_Name	pval	qval	
GO:0006955	233	21825 56619 14	47	immune response	1.32E-54	3.14E-51	
GO:0009615	83	16819 234311 1	29	response to virus	6.90E-43	1.63E-39	
GO:0071346	69	55932 229900 2	27	cellular response to interferon-gamma	7.62E-42	1.81E-38	
GO:0035458	38	55932 14469 14	21	cellular response to interferon-beta	1.08E-37	2.56E-34	
GO:0071222	136	16819 14469 20	30	cellular response to lipopolysaccharide	3.89E-37	9.22E-34	
GO:0006935	130	20296 20306 20	27	chemotaxis	6.54E-33	1.55E-29	
GO:0043123	145	56489 142980 7	24	positive regulation of I-kappaB kinase/NF-kappaB signaling	6.13E-27	1.45E-23	
GO:0006915	536	16819 14939 17	38	apoptotic process	9.07E-27	2.15E-23	
GO:0032760	51	142980 20296 1	17	positive regulation of tumor necrosis factor production	5.02E-26	1.19E-22	
GO:0045071	32	671535 80287 7	14	negative regulation of viral genome replication	3.26E-24	7.73E-21	
GO:0032496	199	20296 11535 18	24	response to lipopolysaccharide	2.15E-23	5.09E-20	
GO:0050830	88	55932 229900 1	18	defense response to Gram-positive bacterium	8.77E-23	2.08E-19	
GO:0043065	294	23882 142980 1	27	positive regulation of apoptotic process	1.06E-22	2.52E-19	

#### Enriched GO figures

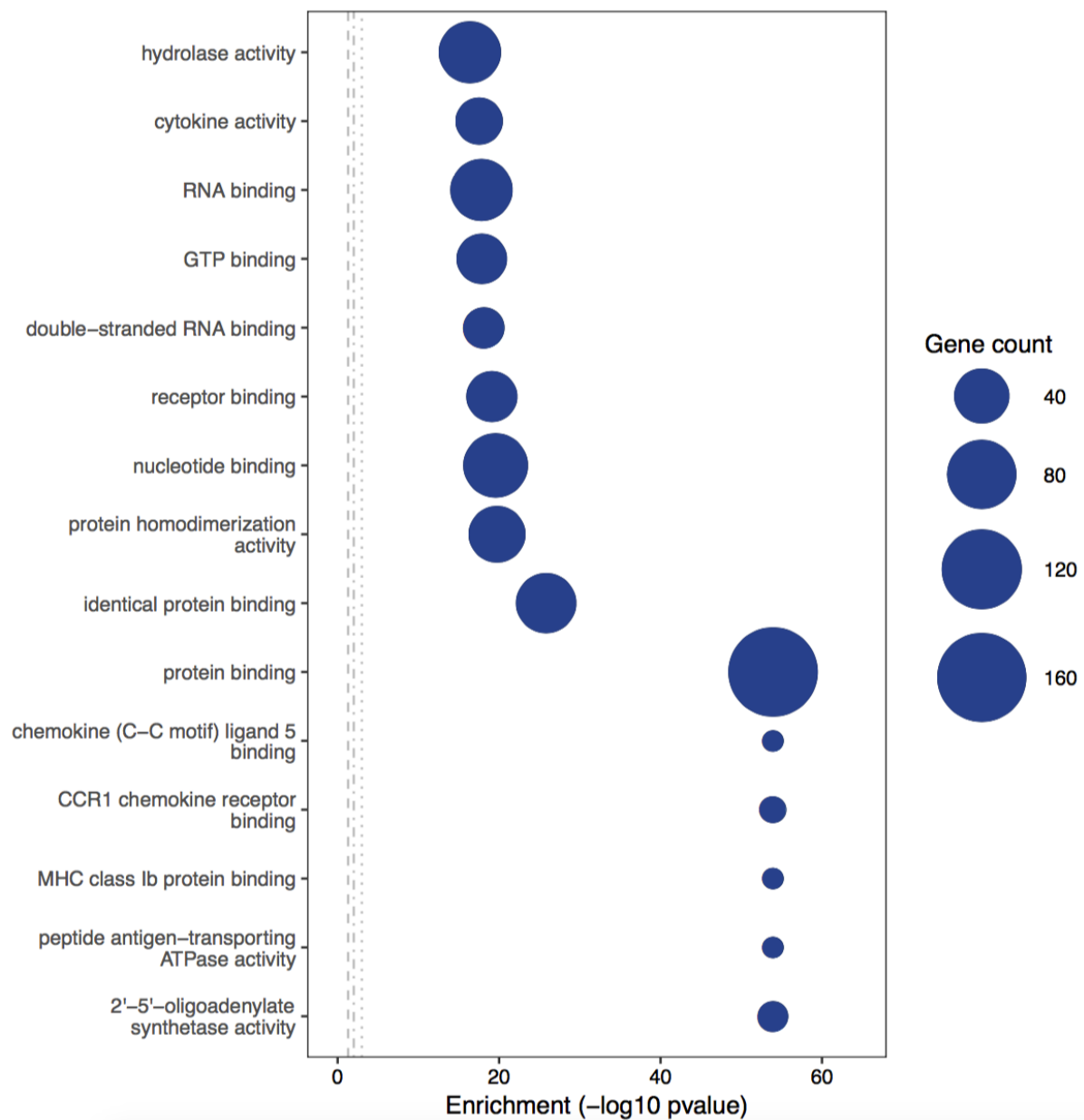
GO.BP.Significant.enrichment.plot.pdf, GO.MF.Significant.enrichment.plot.pdf and

GO.CC.Significant.enrichment.plot.pdf

Example below

x axis = Enrichment (-log10 pvalue)

The number of genes in each term are shown by size circle (bigger circles reflect more GOI in the GO terms).

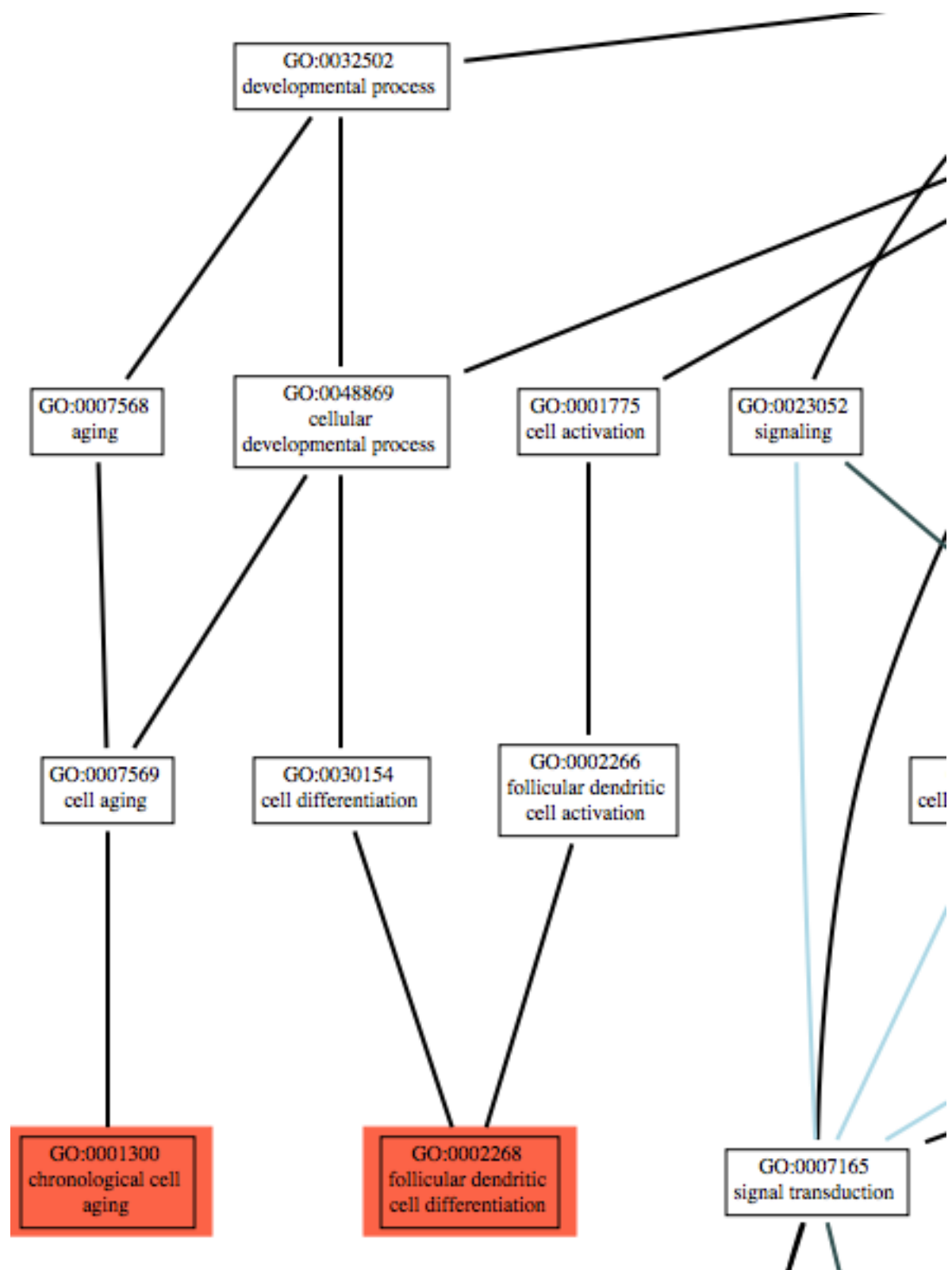


#### GO Directed acyclic graph

The relationships of enriched GO terms are shown in the directed acyclic graph in svg format e.g. GO.BP.top.DAG.svg.

Figures are generated using getAmigoTree see

<https://www.rdocumentation.org/packages/RamiGO/versions/1.18.0/topics/getAmigoTree>



(KEGG.enrichment.analysis.results.table) pathview output e.g mmu03010.pathview.png: red shows entities in user input list which are present in enriched pathway

[illegible]

6

