微生信
Wei Sheng Xin

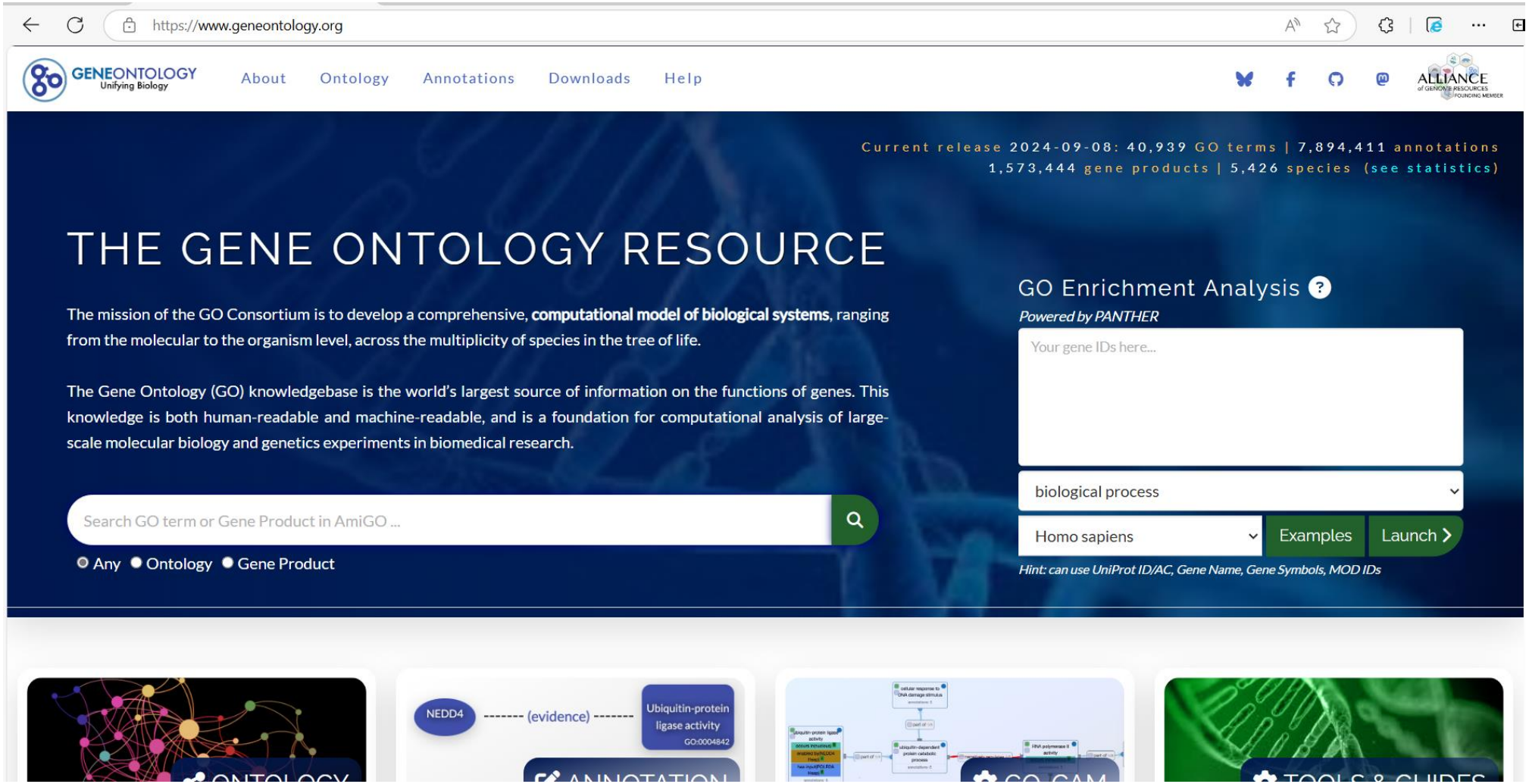
Clusterprofiler GO和KEGG富集分析R代码详解

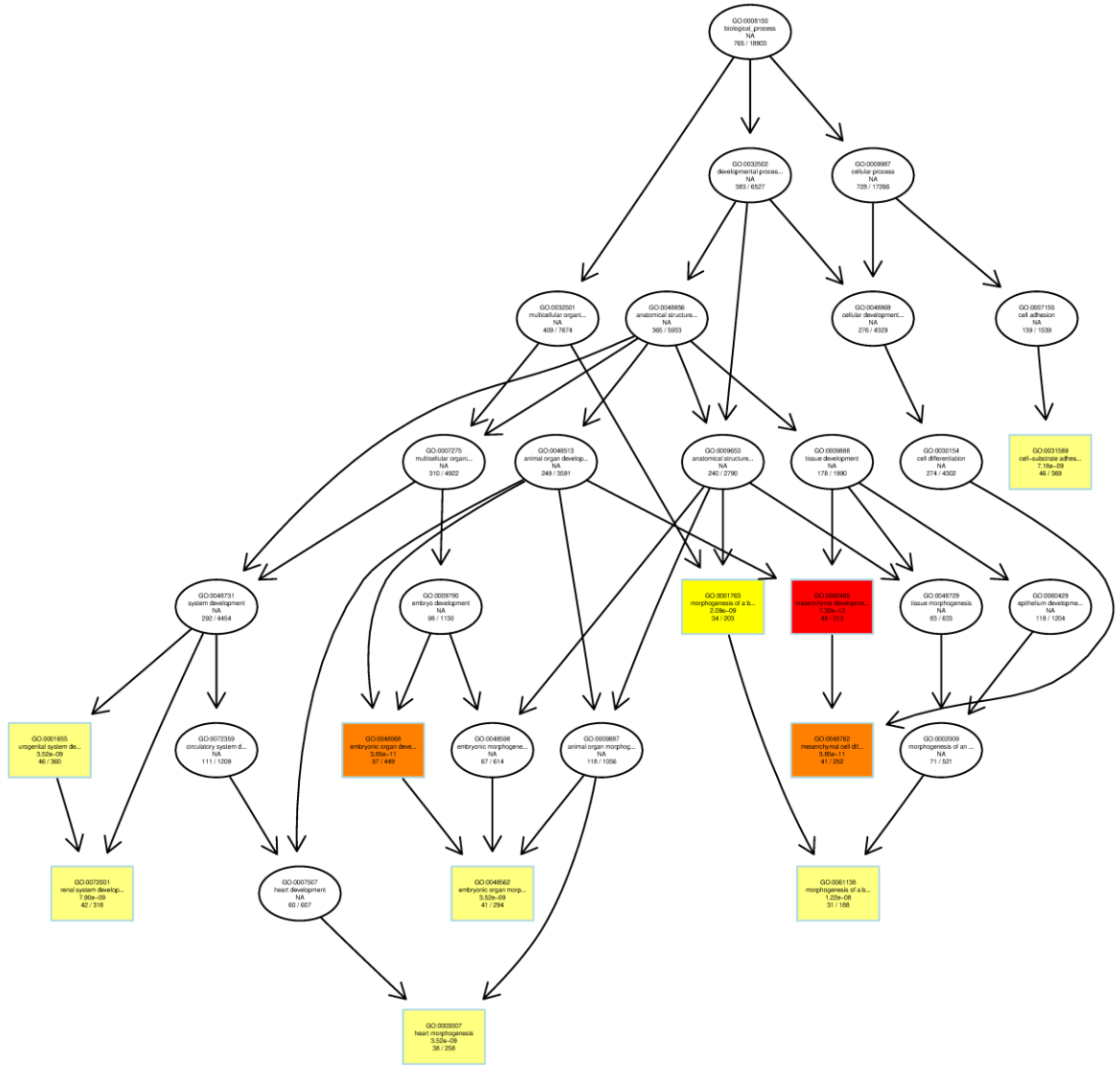
陈明杰 202411

Gene Ontology（基因本体论）简介

- Gene Ontology（GO）是一个在生物信息学领域广泛使用的知识体系，旨在提供一套标准化的词汇来描述基因和蛋白质在细胞中的功能。目的是创建一个动态的、受控的词汇表，以统一和标准化的方式来描述基因和蛋白质的作用。GO将基因和蛋白质的功能分为三个主要的本体论类别：
- 1、分子功能（Molecular Function，MF）：描述单个基因产物（如蛋白质或RNA）或其复合物在分子水平上的活动，例如“催化活性”或“转运蛋白活性”。
- 2、细胞组分（Cellular Component，CC）：描述基因产物在细胞内的位置，如线粒体、核糖体等。
- 3、生物过程（Biological Process，BP）：描述多个分子活动协同完成的生物学过程，如信号转导或代谢过程。
- GO的每个术语都有一个唯一的标识符（GO ID）和名称，并且与本体中的其他术语通过定义的关系相连。这些关系可以是“is_a”或“part_of”，形成了一个有向无环图（DAG）的结构。GO注释是将基因产物与GO术语相关联的过程，这对于理解基因的功能和进行基因表达分析至关重要。GO注释的结果可以用于多种分析，包括基因本体论富集分析，这是一种统计方法，用于确定在一组基因中哪些GO术语的出现频率显著高于随机预期，从而揭示基因集的生物学功能。

www.geneontology.org





KEGG （https://www.kegg.jp）

- KEGG（Kyoto Encyclopedia of Genes and Genomes，京都基因与基因组百科全书）是一个综合数据库资源，旨在从基因组和分子水平信息中理解生物系统的高层次功能和用途，例如细胞、生物体和生态系统。KEGG PATHWAY 是KEGG数据库中最重要且最常用的部分，它包含了大量由科研人员根据已有研究文献，通过手动绘制的通路图，代表着代谢过程、环境信息过程、细胞过程、生物系统、人类疾病和药物开发。

https://www.genome.jp/kegg/pathway.html

KEGG Databases Tools Auto annotation Kanehisa Lab

KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND NETWORK DISEASE DRUG

Select prefix
map Organism

Enter keywords
Go Help

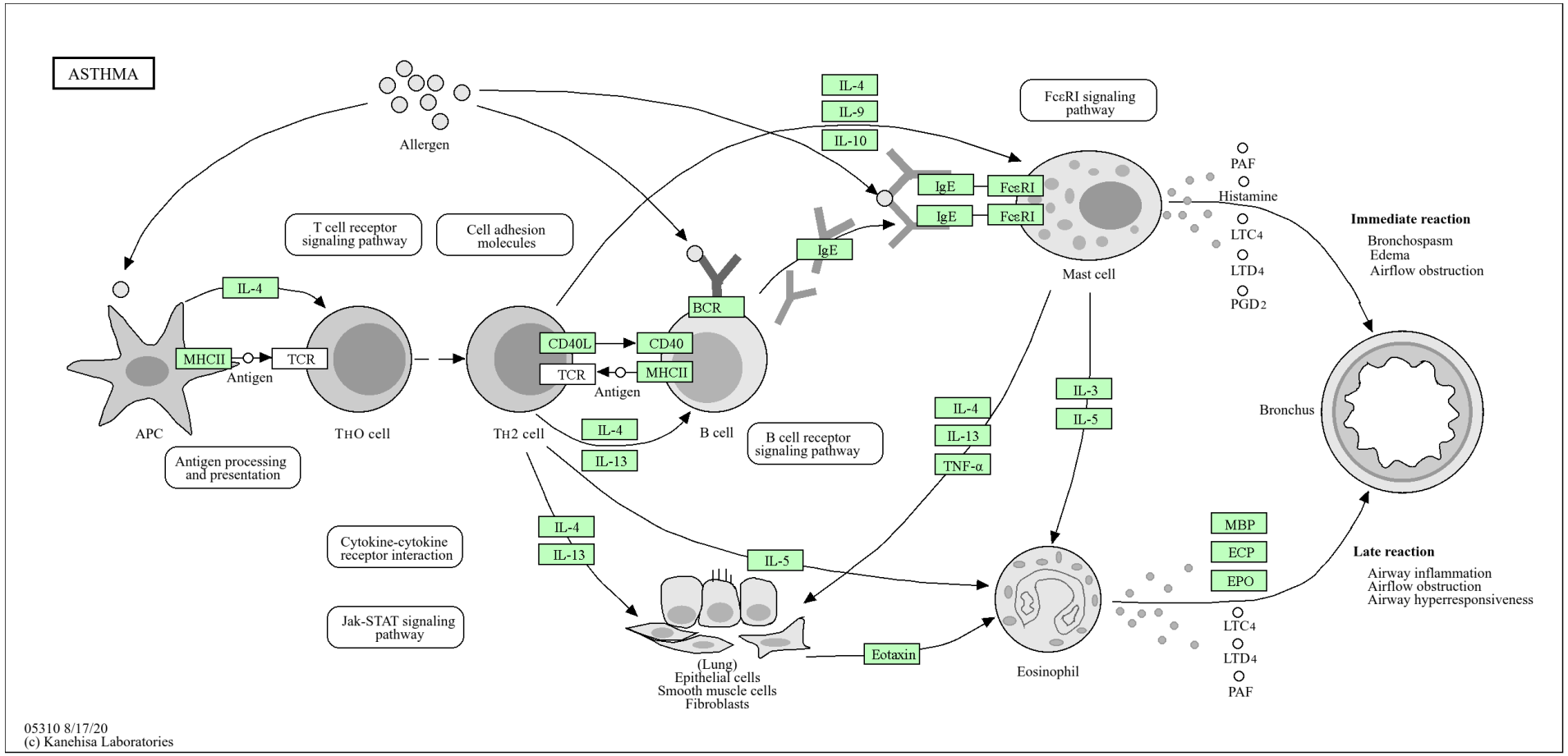
[New pathway maps | Update history]

Pathway Maps

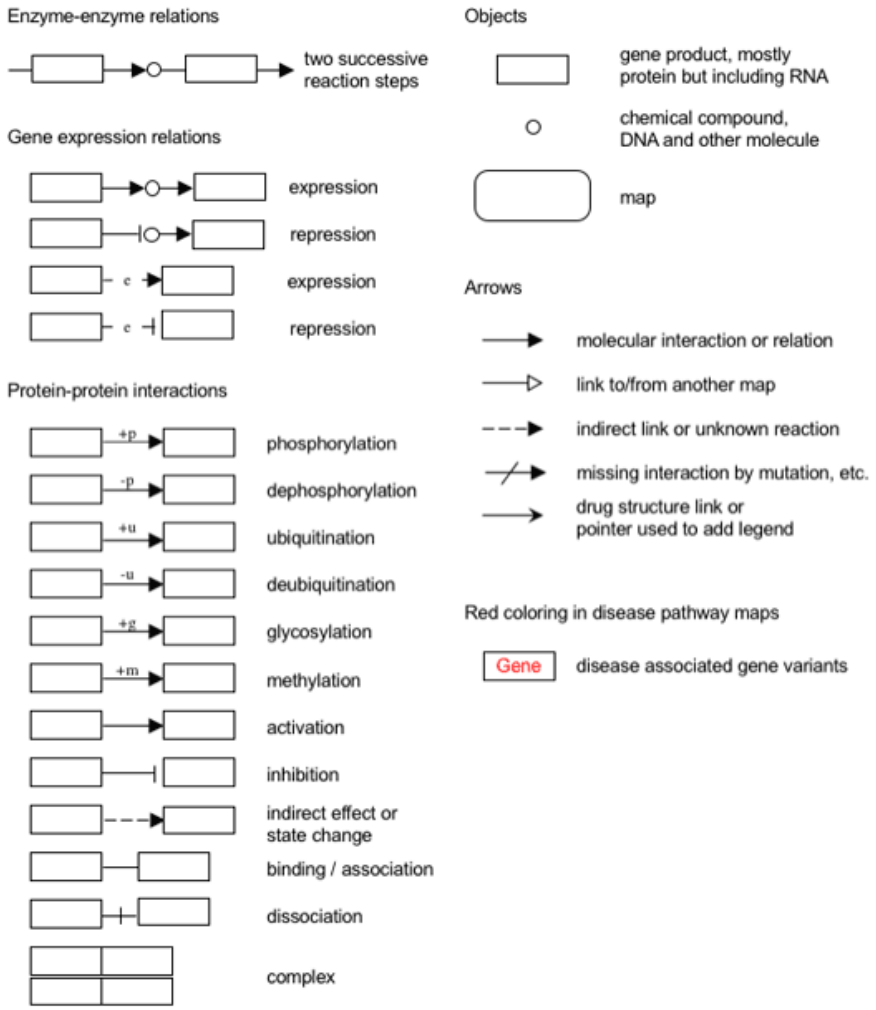
KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:

1. **Metabolism**
   Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan
   Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human Diseases**
7. **Drug Development**

The pathway map viewer linked from this page is a part of KEGG Web Apps and contains features of KEGG mapping.


ASTHMA
Allergen
IL-4
IL-9
IL-10
FcεRI signaling pathway
IgE
FcεRI
IgE
FcεRI
Mast cell
PAF
Histamine
LTC4
LTD4
PGD2
Immediate reaction
Bronchospasm
Edema
Airflow obstruction
T cell receptor signaling pathway
Cell adhesion molecules
IgE
BCR
IL-4
MHCII
Antigen
TCR
APC
TH0 cell
CD40L
CD40
TCR
MHCII
Antigen
TH2 cell
B cell
B cell receptor signaling pathway
IL-4
IL-13
IL-4
IL-13
TNF-α
IL-3
IL-5
Bronchus
Antigen processing and presentation
Cytokine-cytokine receptor interaction
IL-4
IL-13
IL-5
Jak-STAT signaling pathway
Eotaxin
(Lung)
Epithelial cells
Smooth muscle cells
Fibroblasts
Eosinophil
MBP
ECP
EPO
LTC4
LTD4
PAF
Late reaction
Airway inflammation
Airflow obstruction
Airway hyperresponsiveness
05310 8/17/20
(c) Kanehisa Laboratories

通路图上各元素说明


Enzyme-enzyme relations
two successive reaction steps
Gene expression relations
expression
repression
e
expression
e
repression
Protein-protein interactions
+p
phosphorylation
-p
dephosphorylation
+u
ubiquitination
-u
deubiquitination
+g
glycosylation
+m
methylation
activation
inhibition
indirect effect or state change
binding / association
dissociation
complex
Objects
gene product, mostly protein but including RNA
chemical compound, DNA and other molecule
map
Arrows
molecular interaction or relation
link to/from another map
indirect link or unknown reaction
missing interaction by mutation, etc.
drug structure link or pointer used to add legend
Red coloring in disease pathway maps
Gene
disease associated gene variants


酶-酶关系
两个连续的反应步骤
基因表达关系
表达
抑制
e
表达
e
抑制
蛋白-蛋白相互作用
+p
磷酸化
-p
去磷酸化
+u
泛素化
-u
去泛素化
+g
糖基化
+m
甲基化
激活
抑制
间接效应或状态变化
结合/关联
解离
复合物
对象
基因产物，主要是蛋白，但包括RNA
化合物，DNA和其他分子
其他通路图
箭头
分子相互作用或相关
链接到其他通路或从其他通路链接过来
间接链接或未知反应
由于突变等缺失相互作用
药物结构链接或用于添加图例的指针
疾病通路图中的红字
Gene
疾病相关的基因变异

富集分析

- 富集分析（Enrichment Analysis）是一种统计方法，用于确定在一组感兴趣的基因、蛋白质或其他类型的生物分子中，哪些生物学过程、分子功能或细胞组分（如GO术语）显著地过度表达或富集
- **富集分析的主要步骤包括：**

1. **定义感兴趣的基因集**：这通常是实验数据中显著变化的基因，或者是根据某种生物学标准选择的基因
2. **选择参考基因集**：这是整个基因组或蛋白质组的背景集，用于与感兴趣的基因集进行比较
3. **选择适当的统计测试**：常用的方法包括超几何分布测试、卡方检验、Fisher精确检验等
4. **计算富集比**：这是感兴趣的基因集中特定类别的基因数与参考基因集中该类别基因数的比例
5. **调整多重比较**：由于富集分析通常涉及多个假设测试，因此需要调整多重比较以控制假阳性率，常用的方法包括Bonferroni校正、FDR（False Discovery Rate）控制等
6. **解释结果**：分析哪些生物学过程或功能类别显著富集，这有助于揭示基因集背后的生物学意义

代码详解（GO，Pathway通用）

安装

加载R包

读取基因列表

名字转换

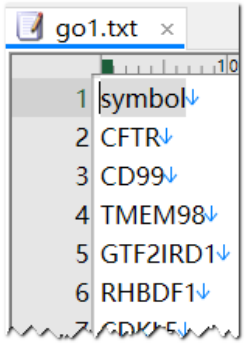
```r
# 安装包
# options(BioC_mirror="https://mirrors.tuna.tsinghua.edu.cn/bioconductor")
# BiocManager::install('org.Hs.eg.db') # human
# BiocManager::install('org.Mm.eg.db') # mouse
# BiocManager::install('org.Rn.eg.db') # rat


# 加载R包
library(clusterProfiler)  # 功能富集分析
library(org.Hs.eg.db)     # 人类基因注释数据库
library(enrichplot)       # 富集结果可视化
library(ggplot2)          # 绘图
library(topGO)            # DAG网络图


# 读取差异基因txt
data = read.table("go1.txt", header=T, quote="")
deg_genes = data$symbol


# 将symbol转成entrezid
gene_symbol <- bitr(deg_genes,  #感兴趣的基因集
                    fromType="SYMBOL",    #输入ID的类型
                    toType=c("ENTREZID"),    #输出ID的类型，可为多个
                    OrgDb="org.Hs.eg.db")  #物种注释数据库

gene <- gene_symbol[,2]
# 数据准备结束
```




```
> head(gene_symbol)
    SYMBOL ENTREZID
1     CFTR     1080
2     CD99     4267
3   TMEM98    26022
4 GTF2IRD1     9569
5   RHBDF1    64285
6    CDKL5     6792
```

富集分析

```r
# 富集分析
#############BP

BP <- enrichGO(gene = gene,   #基因列表(转换的ID)
               keyType = "ENTREZID",   #指定的基因ID类型，默认为ENTREZID
               OrgDb=org.Hs.eg.db,   #物种对应的org包
               ont = "BP",    # BP生物学过程
               pvalueCutoff = 1,   #p值阈值
               pAdjustMethod = "fdr",   #多重假设检验校正方式
               minGSSize = 10,    #注释的最小基因集，默认为10
               maxGSSize = 500,   #注释的最大基因集，默认为500
               qvalueCutoff = 1,   #q值阈值
               readable = TRUE)   #基因ID转换为基因名
```

结果

```r
# 结果导出
write.table(BP@result, file='BP.txt', sep='\t', row.names=F, quote=F)
```

气泡图

```r
# 气泡图
pdf("BP.dot.pdf", width=8, height=6)

dotplot(BP,   #GO富集分析结果
        x = "GeneRatio",   #横坐标,默认GeneRatio ,也可以为Count
        color = "p.adjust",   # 右纵坐标,默认p.adjust,也可以为pvalue和qvalue
        showCategory = 20,    # 展示前20个点，默认为10个
        label_format = 100,   # 不要换行
        size = NULL,          # 点的大小
        title = "BP_dotplot" # 设置图片的标题
        )

dev.off()
```

c('term1', 'term2', 'term3') # 手工挑选条目绘图

765	输入基因数
18903	数据库中全部基因数（背景）
48	输入基因数 ∩ 某条目中的基因数
313	某条目中的基因数

多种算法 → P值 → 多重检验 → p.adj

Enrichment Score

-log10(p)

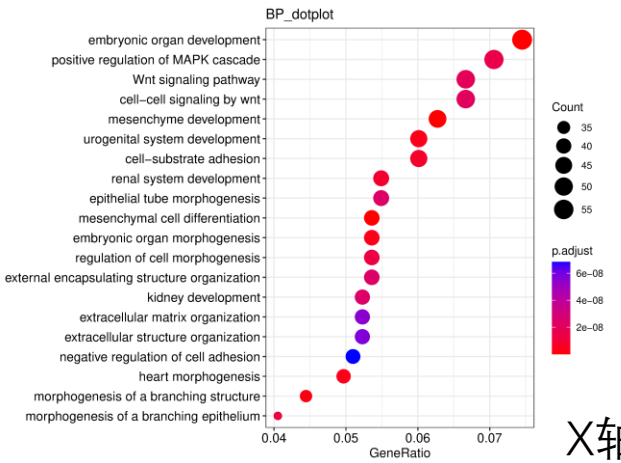
	A	B	C	D	E	F	G	H	I
1	ID	Description	GeneRati	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
2	GO:0060485	mesenchyme develop	48/765	313/1890	1.47E-15	7.32E-12	5.62E-12	ISL1/IGF	48
3	GO:0048568	embryonic organ de	57/765	449/1890	1.65E-14	3.85E-11	2.96E-11	KITLG/PR	57
4	GO:0048762	mesenchymal cell d	41/765	252/1890	2.32E-14	3.85E-11	2.96E-11	ISL1/IGF	41
5	GO:0001763	morphogenesis of a	34/765	203/1890	1.68E-12	2.09E-09	1.61E-09	PRDM1/AB	34
6	GO:0048562	embryonic organ mo	41/765	294/1890	4.3E-12	3.52E-09	2.7E-09	ITGA8/ME	41

GO条目ID　GO条目名字　P，校正p，q统计值　具体重叠基因情况

条目基因/背景基因

重叠基因/输入基因

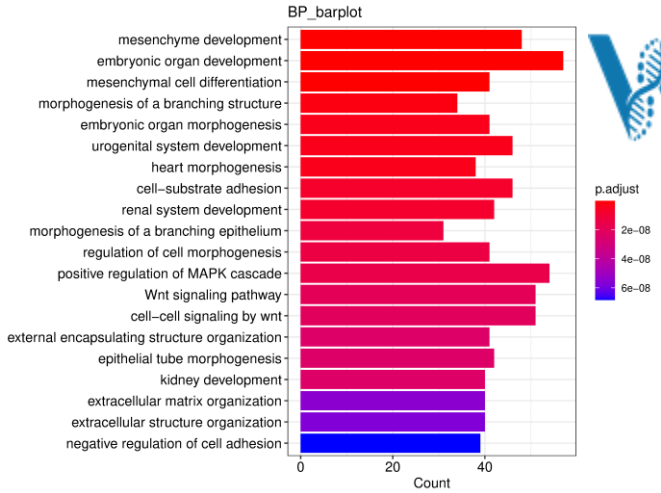
Rich factor 比值



X轴多变

Bar图

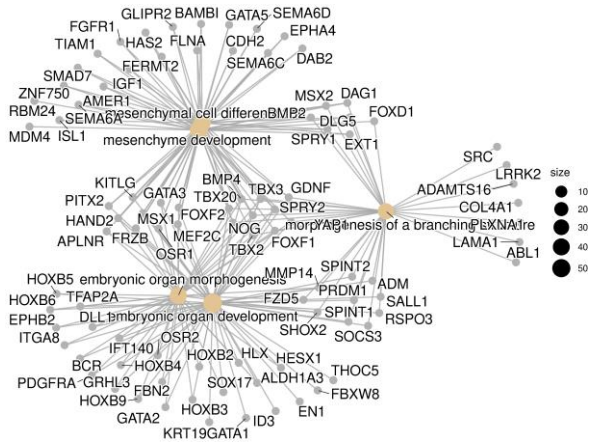
```
barplot(BP,  #GO富集分析结果
        x = "Count",          #横坐标,默认Count,也可以为GeneRation
        color = "p.adjust",   #右纵坐标,默认p.adjust,也可以为pvalue和qvalue
        showCategory = 20,    #展示前20个，默认为10个
        label_format = 100,   # 不要换行
        size = NULL,
        title = "BP_barplot"   #设置图片的标题
        )
dev.off()
```



Cnet图

```
# 绘制GO富集分析的网络图
pdf('BP.cnet.pdf', width=8, height=6)
cnetplot(BP,
        showCategory = 5,
        foldChange = NULL, # color.params = list(foldChange = your_value)
        node_label = "all")
dev.off()
```

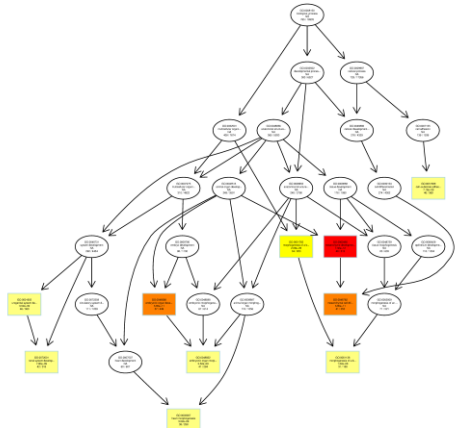
输入带log2fc时，基因点根据log2fc上色



DAG图

```
# 绘制DAG图
pdf('BP.dag.pdf', width=8, height=8)

plotGOgraph(BP,  #输出enrichGO或gseGO的有向无环图(与输入的对象对应)
            firstSigNodes = 10,  #显著性节点的个数,默认10个
            useInfo = "all",
            sigForAll = T,  #是否在所有节点展示score/p-value
            useFullNames = T,  #是否使用全称
            )
dev.off()
```



代码详解（GO，Pathway通用）

安装

加载R包

读取基因列表

名字转换

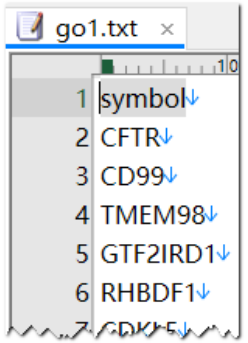
```r
# 安装包
# options(BioC_mirror="https://mirrors.tuna.tsinghua.edu.cn/bioconductor")
# BiocManager::install('org.Hs.eg.db') # human
# BiocManager::install('org.Mm.eg.db') # mouse
# BiocManager::install('org.Rn.eg.db') # rat


# 加载R包
library(clusterProfiler)  # 功能富集分析
library(org.Hs.eg.db)     # 人类基因注释数据库
library(enrichplot)       # 富集结果可视化
library(ggplot2)          # 绘图
library(topGO)            # DAG网络图


# 读取差异基因txt
data = read.table("go1.txt", header=T, quote="")
deg_genes = data$symbol


# 将symbol转成entrezid
gene_symbol <- bitr(deg_genes,  #感兴趣的基因集
                    fromType="SYMBOL",    #输入ID的类型
                    toType=c("ENTREZID"),    #输出ID的类型，可为多个
                    OrgDb="org.Hs.eg.db")  #物种注释数据库

gene <- gene_symbol[,2]
# 数据准备结束
```




```
> head(gene_symbol)
    SYMBOL ENTREZID
1     CFTR     1080
2     CD99     4267
3   TMEM98    26022
4 GTF2IRD1     9569
5   RHBDF1    64285
6    CDKL5     6792
```

Pathway富集分析（与GO类似）

```
############ KEGG
KEGG<- enrichKEGG(gene = gene,        #基因列表(同GO)
                  organism = "hsa",  #物种
                  keyType = "kegg",  #指定的基因ID类型，默认为kegg
                  minGSSize = 10,
                  maxGSSize = 500,
                  pvalueCutoff = 1,
                  pAdjustMethod = "fdr",
                  qvalueCutoff = 1,
                  use_internal_data = TRUE # 使用本地KEGG.db库
                  )

KEGG <- setReadable(KEGG, OrgDb = org.Hs.eg.db, keyType="ENTREZID")  # 将entrezid转成symbol

# 如果使用本地KEGG.db库
#all_path = as.data.frame(KEGG.db::KEGGPATHID2NAME)
#KEGG@result$Description = all_path$path_name[match(KEGG@result$ID, all_path$path_id)]

# 导出结果
write.table(KEGG@result, file='KEGG.txt', sep='\t', row.names=F, quote=F)

# 气泡图
pdf('KEGG.dot.pdf', width=6, height=6)
dotplot(KEGG,   #GO富集分析结果
        x = "GeneRatio",   #横坐标,默认GeneRatio,也可以为Count
        color = "p.adjust",   #右纵坐标,默认p.adjust,也可以为pvalue和qvalue
        showCategory = 20,   #展示前20个点，默认为10个
        label_format = 100, # 不要换行
        size = NULL, #点的大小
        title = "KEGG dotplot"   #设置图片的标题
        )
dev.off()
```

```
# R4.3+下面创建KEGG.db
# remotes::install_github("YuLab-SMU/createKEGGdb")
library(createKEGGdb)
species <-c('hsa','mmu','rno')
create_kegg_db(species)
# 可以装在其他低或高版本R中
install.packages("KEGG.db_1.0.tar.gz", type="source")
```

use_internal_data参数	优点	缺点
T（本地KEGG.db库）	运行快，稳定（结果保持不变）	非实时更新
F（联网KEGG）	实时更新（若间隔时间较长，则前后两次的结果可能会不一致）	运行慢 不稳定

使用本地版，加上这两句
否则description列是空的

Pathview添加颜色

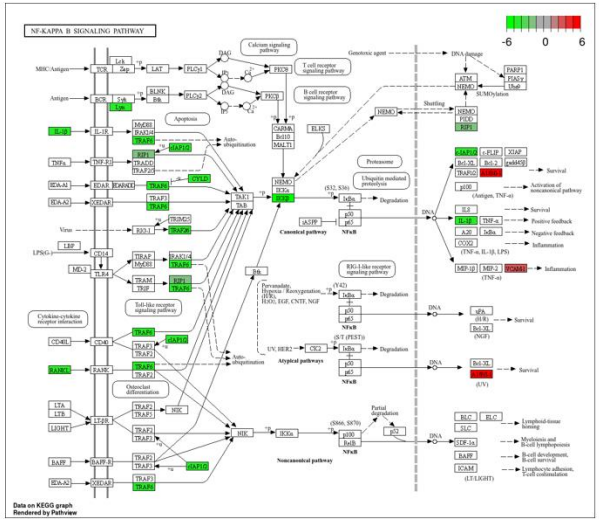
```r
# 安装
# BiocManager::install("pathview")

# 载入包
library(pathview)
options(bitmapType='cairo')

# 读取数据
data = read.table('test.txt',header=TRUE,row.names=1,sep="\t",check.names = FALSE,quote = "")

# 联网下载，并绘图
pv.out <- pathview(gene.data = data, pathway.id = "04062",  # pathway id
species="hsa",  # 物种
out.suffix = "weishengxin", # 输出文件名后缀
limit=list(gene=c(-8, 8)),  # colorbar的范围
kegg.native = T,
low = list(gene = "#00ff00"), mid =list(gene = "#acacac"), high = list(gene = "#ff0000")) # 设置颜色
```

	A	B
1	hsa04064	log2fc
2	3553	-4.40927
3	4067	-3.6447
4	7189	-4.09347
5	8737	-2.07215
6	329	-3.91764
7	1540	-4.52864
8	8600	-4.24196
9	3551	-5.82876
10	597	5.182839
11	7412	2.488494



https://www.genome.jp/dbget-bin/www_bget?pathway+hsa04064

其他物种的富集分析

- agriGOv2 （植物：https://systemsbiology.cau.edu.cn/agriGOv2/）
- Worm（线虫：https://wormbase.org/tools/enrichment/tea/tea.cgi）
- KOBAS (http://bioinfo.org/kobas，视频 https://www.bilibili.com/video/BV15c41137Du/)
- DAVID (https://david.ncifcrf.gov)
- Enrichr (https://maayanlab.cloud/Enrichr)
- Panther (https://www.pantherdb.org)

特殊物种（自己准备注释文件）

```
TCEANC2	GO:0005634
TCEANC2	GO:0006351
MEX3A	GO:0003676
MEX3A	GO:0003723
MEX3A	GO:0046872
MEX3A	GO:0005634
MEX3A	GO:0005737
MEX3A	GO:0005829
MEX3A	GO:0000932
U6	GO:0000244
U6	GO:0000353
U6	GO:0046540
U6	GO:0030621
U6	GO:0005688
U1	GO:0030627
U1	GO:0000395
U1	GO:0005685
```

基因-GOID

```
GO:0006396	RNA processing	biological_process
GO:0005730	nucleolus	cellular_component
GO:0005634	nucleus	cellular_component
GO:0006351	DNA-templated transcription	biological_process
GO:0003676	nucleic acid binding	molecular_function
GO:0003723	RNA binding	molecular_function
GO:0046872	metal ion binding	molecular_function
GO:0005737	cytoplasm	cellular_component
GO:0005829	cytosol	cellular_component
GO:0000932	P-body	cellular_component
```

GOID – 名字

```
go_rich <- enricher(gene = gene_select,
                    TERM2GENE = go_anno[c('ID','gene_id')],
                    TERM2NAME = go_anno[c('ID','Description')],
                    pvalueCutoff = 1,
                    pAdjustMethod = 'BH',
                    qvalueCutoff = 1,
                    minGSSize = 10,
                    maxGSSize = 200)
```

结果略有差异，该用哪个？

- 数据库版本不太一样（例如背景基因数）
- 算法，阈值，不太一样（fisher，超几何，卡方等）
- 但是基本上还是一样的
- 哪个对你更有利用哪个吧

```
BP <- enrichGO(gene = gene,  #基因列表(转换的ID)
               keyType = "ENTREZID",  #指定的基因ID类型，默认为ENTREZID
               OrgDb=org.Hs.eg.db,  #物种对应的org包
               ont = "BP",   # BP生物学过程
               pvalueCutoff = 1,  #p值阈值
               pAdjustMethod = "fdr",  #多重假设检验校正方式
               minGSSize = 10,   #注释的最小基因集，默认为10
               maxGSSize = 500,  #注释的最大基因集，默认为500
               qvalueCutoff = 1,  #q值阈值
               readable = TRUE)  #基因ID转换为基因名
```

maxGSSize默认500，这样超过500个基因的基因集在用clusterprofiler分析时就会被去掉；而metascape没有限制基因数

metascape　　clusterProfiler