

Ontology/本体

——GO (Gene Ontology, 基因本体); HPO (Human Phenotype Ontology, 人类表型本体) 和 PTO (Plant Trait Ontology, 植物性状本体)

Jingbo Xia

Huazhong Agricultural University

xiajingbo.math@gmail.com

2025-10-30/31

Gene Ontology I

Basic knowledge of GO

The Gene Ontology ¹: “a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing”.

- ① Molecular function —elemental activity or task
- ② Biological process —broad objective or goal
- ③ Cellular component —location or complex



[Jingbo] 讨论: GO 中概念的三个分类的设置是否合理?

Gene Ontology II

Basic knowledge of GO

- 1 Molecular function —elemental activity or task
- 2 Biological process —broad objective or goal
- 3 Cellular component —location or complex



[Jingbo] 讨论: GO 中概念的三个分类的设置是否合理?



[Jingbo] 以下概念应该属于三个类别中的哪一个?

nuclease activity,
DNA binding,
microtubule motor activity,
mitosis cell cycle,
signal transduction,
lipid metabolic process,
nucleus,
ribosome

3. GO Enrichment (富集)



[Jingbo] 讨论：何为富集分析？

我们通常是从 GO 了解本体及其注释的，那么何为富集呢？富集分析中的“富集”，通常指一个被选定的基因群体，是否较为“显著”地整体与一个对应表型产生关联性？

传统情况下，GO 富集分析中，首先用或干或湿的实验手段获得一个基因集合，然后用超几何检验等检验方法判断该集合是否与某 Gene 本体的条目“显著相关”。

在课程讲授过程中，我们介绍了 GO 和 HPO 两个实体，一个关注基因分子层面，一个关注人类疾病表型层面。这两个生物本体虽然关注对象的宏观、微观层次不同，但是围绕它们已经均发展了较为充分的（基因-本体条目关联性）背景数据，这为富集分析提供了数据基础。

R 安装和环境设置

R 语言安装步骤

- 1 访问 R 项目官网: <https://cran.r-project.org/>
- 2 根据你的操作系统点击对应的下载链接 (如"Download R for Windows")
- 3 对于 Windows 用户: 选择"install R for the first time"
- 4 下载最新版本的安装程序 (例如 R-4.5.1)
- 5 运行安装程序, 按照向导完成安装

注意事项

- 安装过程中接受默认设置通常是最安全的选择
- 建议在安装新版本前卸载旧版本
- 确保下载与你的系统架构匹配的版本 (32 位或 64 位)

R 安装和环境设置

RStudio 安装步骤

- ❶ 访问 RStudio 官网: <https://www.rstudio.com/products/rstudio/download/#download>
- ❷ 选择免费的 RStudio Desktop 版本并点击下载
- ❸ 运行下载的安装程序 (例如 RStudio-2025.05.1-513)
- ❹ 遵循安装向导的指示完成安装

重要提示 RStudio 是一个集成开发环境 (IDE), 它需要依赖 R 语言。因此, **必须先安装 R**, 然后再安装 RStudio。

操作步骤

- ① 从开始菜单或桌面快捷方式打开 RStudio
- ② 在 RStudio 底部的控制台 (Console) 中, 输入以下测试命令: `print("Hello World!")`
- ③ 按回车键执行
 - 如果控制台输出 `[1] "Hello World!"`, 则表示 R 和 RStudio 环境配置正常。
 - 如果 RStudio 启动失败或报错, 请检查 R 是否已正确安装。

R 代码示例: GO enrichment analysis



[Jingbo] 讨论: 如何获取 GO 富集分析的网络资源?

较为丰富的网络资源

- <https://geneontology.org/docs/go-enrichment-analysis/>
- <https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/goenrichment/tutorial.html>
- <http://bioinformatics.sdstate.edu/go/>
- 我们提供了一段 BioC 的基础代码

R 代码示例: GO enrichment analysis

Packages Installation

```
1 install.packages('devtools', repos = 'https://mirrors.tuna.tsinghua.edu.cn/CRAN')
2 install.packages('BiocManager', repos = 'https://mirrors.tuna.tsinghua.edu.cn/CRAN')
3 library('devtools')
4 BiocManager::install(version = "3.10")
5 Needed=c("bit", "formatR", "hms", "triebeard", "tweenr", "polyclip", "RcppEigen", "
  RcppArmadillo", "zlibbioc", "bit64", "blob", "plogr", "lambda.r", "futile.options",
  "progress", "urltools", "gridGraphics", "ggforce", "ggrepel", "viridis", "tidygraph",
  "graphlayouts", "bitops", "XVector", "IRanges", "RSQLite", "futile.logger", "snow",
  "data.table", "gridExtra", "fastmatch", "cowplot", "europepmc", "ggplotify", "
  ggraph", "ggridges", "igraph", "dplyr", "tidyselect", "RCurl", "Biostrings", "
  AnnotationDbi", "BiocParallel", "DO.db", "fgsea", "GOSemSim", "qvalue", "S4Vectors",
  "BiocGenerics", "graph", "Biobase", "GO.db", "SparseM", "matrixStats", "DBI", "
  enrichplot", "rvcheck", "tidyr", "org.Hs.eg.db", "KEGGgraph", "XML", "Rgraphviz", "
  png", "KEGGREST")
6 install.packages(Needed, repos = 'https://mirrors.tuna.tsinghua.edu.cn/CRAN')
7 BiocManager::install(c("DOSE", "topGO", "clusterProfiler", "pathview"))
```

R 代码示例: GO enrichment analysis

Packages Installation

DOSE package (Disease Ontology Semantic and Enrichment analysis) ³

ClusterProfiler ⁴

```
1 library(DOSE)
2 library(org.Hs.eg.db)
3 library(topGO)
4 library(clusterProfiler)
5 library(pathview)
6
7 # Human gene types
8 keytypes(org.Hs.eg.db)
9 #[1] "ACCNUM" "ALIAS" "ENSEMBL" "ENSEMBLPROT" "ENSEMBLTRANS" "ENTREZID"
10 #[7] "ENZYME" "EVIDENCE" "EVIDENCEALL" "GENENAME" "GO" "GOALL" "IPI"
11 #[16] "MAP" "OMIM" "ONTOLOGY" "ONTOLOGYALL" "PATH" "PFAM" "PMID"
12 #[23] "PROSITE" "REFSEQ" "SYMBOL" "UCSCKG" "UNIGENE" "UNIPROT"
```

³<https://bioc.ism.ac.jp/packages/3.6/bioc/manuals/DOSE/man/DOSE.pdf>

⁴Guangchuang Yu, Li-Gen Wang, Yanyan Han and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology. 2012, 16(5):284-287.

<https://www.bioconductor.org/packages/devel/bioc/manuals/clusterProfiler/man/clusterProfiler.pdf>  

R 代码示例: GO enrichment analysis

Gene names importing

```
1 # Input several genes
2 MyGeneSet <- c("AASDH", "ABCB11", "ADAM12", "ADAMTS16", "ADAMTS18")
3 MyGeneIDSet = bitr(MyGeneSet,
4                     fromType="SYMBOL", # input format is "symbol"
5                     toType="ENTREZID", # output format is "ENTERZID"
6                     OrgDb="org.Hs.eg.db") #Use human data
7
8 #> MyGeneIDSet
9 #   SYMBOL   ENTREZID
10 #1   AASDH    132949
11 #2  ABCB11     8647
12 #3  ADAM12     8038
13 #4 ADAMTS16    170690
14 #5 ADAMTS18    170692
```

R 代码示例: GO enrichment analysis

Gene names importing

```
1 # Import genes from file
2 ## Create a file "GeneNames" which contains genes symbol for each line.
3 setwd("/Users/zhs1/RTest/")
4 MyGeneSet2 <- read.table("GeneNames", header=FALSE)
5 MyGeneSet2$V1 <- as.character(MyGeneSet2$V1)
6
7 MyGeneIDSet2 = bitr(MyGeneSet2$V1, fromType="SYMBOL", toType=c("ENSEMBL", "ENTREZID", "GO
  "), OrgDb="org.Hs.eg.db")
8 head(MyGeneIDSet2, 6)
9 #> MyGeneIDSet2
10 # SYMBOL      ENSEMBL ENTREZID
11 #1      AASDH  ENSG00000157426   132949
12 #2     ABCB11  ENSG00000073734     8647
13 #3     ABCB11  ENSG00000276582     8647
14 #4     ADAM12  ENSG00000148848     8038
15 #5  ADAMTS16  ENSG00000145536   170690
16 #6  ADAMTS18  ENSG00000140873   170692
```

R 代码示例: GO enrichment analysis

Enrichment Analysis and Visualization

Gene name normalization. ^a

Common Identifiers

Gene

[Ensembl](#) [ENSG00000139618](#)

[Entrez Gene](#) [675](#)

Unigene [Hs.34012](#)

RNA transcript

GenBank [BC026160.1](#)

[RefSeq](#) [NM_000059](#)

Ensembl [ENST00000380152](#)

Protein

Ensembl [ENSP00000369497](#)

[RefSeq](#) [NP_000050.2](#)

[UniProt](#) [BRCA2_HUMAN](#) or

[A1YBP1_HUMAN](#)

IPI [IPI00412408.1](#)

EMBL [AF309413](#)

PDB [1MIU](#)

Species-specific

HUGO HGNC [BRCA2](#)

MGI [MGI:109337](#)

RGD [2219](#)

ZFIN [ZDB-GENE-060510-3](#)

FlyBase [CG9097](#)

WormBase [WBGene00002299](#) or [ZK1067.1](#)

SGD [S000002187](#) or [YDL029W](#)

Annotations

InterPro [IPR015252](#)

OMIM [600185](#)

Pfam [PF09104](#)

Gene Ontology [GO:0000724](#)

SNPs [rs28897757](#)

Experimental Platform

Affymetrix [208368_3p_s_at](#)

Agilent [A_23_P99452](#)

CodeLink [GE60169](#)

Illumina [GI_4502450-S](#)

R 代码示例: GO enrichment analysis

Enrichment Analysis and Visualization

```
1 data(geneList, package="DOSE") #21,126 genes as background genes
2 names(geneList)
3
4 # Enrich all.
5 ego_ALL <- enrichGO(gene = MyGeneIDSet2$ENTREZID,
6                     universe = names(geneList), #Background Gene
7                     OrgDb = org.Hs.eg.db,
8                     #keytype = 'ENSEMBL',
9                     ont = "ALL", # Other options: CC or BP or MF
10                    pAdjustMethod = "BH", # adjust method, "holm", "hochberg", "
11                                     hommel", "bonferroni", "BH", "BY", "fdr", "none"
12                    pvalueCutoff = 1,
13                    qvalueCutoff = 1,
14                    readable = TRUE)
15
16 write.csv(summary(ego_ALL), "ALL-enrich.csv", row.names = FALSE)
17
18 #Enrich CC, BP, or MF
19 ego_MF <- enrichGO(gene = MyGeneIDSet2$ENTREZID, universe = names(geneList), OrgDb = org.
20                    Hs.eg.db, ont = "MF", pAdjustMethod = "BH", pvalueCutoff = 1, qvalueCutoff = 1, readable
21                    = TRUE)
```

R 代码示例: GO enrichment analysis

Enrichment Analysis and Visualization

ALL-enrich.csv									
使用“Microsoft Excel”打开									
ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
BP	GO:0015721	bile acid and bile salt transport	1/5	21/11512	0.00908927496913425	0.328504361565643	0.315984231633014	ABCB11	1
BP	GO:0006699	bile acid biosynthetic process	1/5	28/11512	0.0121043014992035	0.328504361565643	0.315984231633014	ABCB11	1
BP	GO:0007520	myoblast fusion	1/5	34/11512	0.0146827646305175	0.328504361565643	0.315984231633014	ADAM12	1
BP	GO:0008206	bile acid metabolic process	1/5	37/11512	0.0159699756334489	0.328504361565643	0.315984231633014	ABCB11	1
BP	GO:0000768	syncytium formation by plasma membrane fusion	1/5	45/11512	0.0193959625197148	0.328504361565643	0.315984231633014	ADAM12	1
BP	GO:0140253	cell-cell fusion	1/5	45/11512	0.0193959625197148	0.328504361565643	0.315984231633014	ADAM12	1
BP	GO:0006949	syncytium formation	1/5	46/11512	0.0198235390599957	0.328504361565643	0.315984231633014	ADAM12	1
BP	GO:0014902	myotube differentiation	1/5	80/11512	0.0342726443795521	0.496953343503506	0.478013197925332	ADAM12	1
BP	GO:0035725	sodium ion transmembrane transport	1/5	104/11512	0.0443690257863316	0.514680699121447	0.495064919300121	ABCB11	1
BP	GO:0046717	acid secretion	1/5	104/11512	0.0443690257863316	0.514680699121447	0.495064919300121	ABCB11	1
BP	GO:0015718	monocarboxylic acid transport	1/5	127/11512	0.0539653122592263	0.532684970100477	0.512383002093018	ABCB11	1
BP	GO:0045766	positive regulation of angiogenesis	1/5	143/11512	0.0605954022949936	0.532684970100477	0.512383002093018	ADAM12	1
BP	GO:1904018	positive regulation of vasculature development	1/5	160/11512	0.0675990893085697	0.532684970100477	0.512383002093018	ADAM12	1
BP	GO:0006814	sodium ion transport	1/5	161/11512	0.0680097662424541	0.532684970100477	0.512383002093018	ABCB11	1
BP	GO:0006694	steroid biosynthetic process	1/5	164/11512	0.0692409288818571	0.532684970100477	0.512383002093018	ABCB11	1
BP	GO:0015850	organic hydroxy compound transport	1/5	196/11512	0.0822925585478718	0.532684970100477	0.512383002093018	ABCB11	1
BP	GO:0051146	striated muscle cell differentiation	1/5	224/11512	0.0935922485962097	0.532684970100477	0.512383002093018	ADAM12	1
BP	GO:1901617	organic hydroxy compound biosynthetic process	1/5	231/11512	0.0963997041193237	0.532684970100477	0.512383002093018	ABCB11	1

图 11: Data format of GO enrichment result

ONTOLOGY ID Description GeneRatio BgRatio pvalue p.adjust qvalue geneID Count

BP GO:0015721 bile acid and bile salt transport 1/5 21/11512 0.0090893 0.3285043 0.3159842 ABCB11 1

R 代码示例: GO enrichment analysis

Enrichment Analysis and Visualization

```
1 barplot(ego_MF, showCategory=20, title="EnrichmentGO_MF")
```

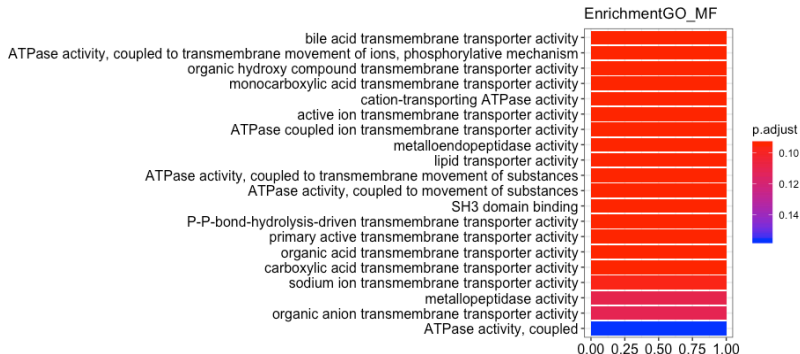


图 12: Bar plot of GO enrichment result

R 代码示例: GO enrichment analysis

Enrichment Analysis and Visualization

```
plotGOgraph(ego_MF)
```

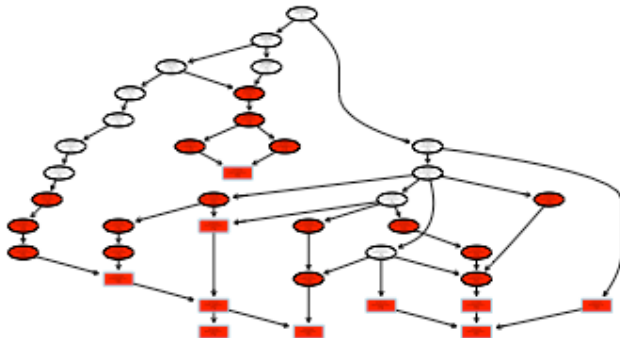


图 13: Hierarchical tree of enriched GO terms