功能聚类 - GOKEGG 分组

Group	ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Up1	BP	GO:0010948	negative regulation of cell c...	8/40	301/18800	1.938983e-07	3.098495e-04	2.049199e-04
Up1	CC	GO:0098687	chromosomal region	7/40	366/19594	8.213998e-06	9.281817e-04	6.311809e-04
Up1	MF	GO:0035173	histone kinase activity	3/38	16/18410	4.459956e-06	6.415481e-04	4.536198e-04
Up1	KEGG	hsa04110	Cell cycle	6/29	126/8164	4.259266e-06	4.387044e-04	3.990260e-04
Down1	BP	GO:0010965	regulation of mitotic sister c...	5/52	65/18800	9.684173e-07	4.698010e-04	3.896998e-04
Down1	CC	GO:0072686	mitotic spindle	4/53	160/19594	9.184963e-04	4.949544e-02	4.113194e-02
Down1	MF	GO:0032395	MHC class II receptor activity	2/52	10/18410	3.470636e-04	6.455383e-02	4.822357e-02
Down1	KEGG	hsa05330	Allograft rejection	3/27	38/8164	2.519592e-04	1.421750e-02	1.125632e-02
Up2	BP	GO:0000212	meiotic spindle organization	3/37	15/18800	3.141279e-06	2.824010e-03	2.169136e-03
Up2	CC	GO:0000775	chromosome, centromeric r...	6/43	227/19594	9.642806e-06	6.484437e-04	4.603395e-04
Up2	MF	GO:0008009	chemokine activity	4/37	49/18410	2.741683e-06	2.193347e-04	1.500711e-04
Up2	KEGG	hsa04061	Viral protein interaction wit...	4/15	100/8164	2.608345e-05	1.851925e-03	1.647376e-03

网址：https://www.xiantao.love

更新时间：2023.11.16

目录

基本概念

- 富集分析：简单而言，就是取一部分有功能注释的分子与所有有功能注释的分子去比较（超几何分布检验），确定这一部分分子中都涉及了哪些功能作用。注意：单独几个分子做富集分析意义并不大。
- GO（Gene Ontology，基因本体）数据库：把基因的功能分成了三类：生物过程（biological process, BP）、细胞组分（cellular component, CC）、分子功能（molecular function, MF）。利用 GO 数据库，可以得到目标基因在 CC, MF 和 BP 三个层面上有什么关联。
- KEGG（Kyoto Encyclopedia of Genes and Genomes）数据库：一种通路数据库，收集了很多通路相关的数据库。通路数据库还包括 wikipathway，reactome 等。
- 超几何分布检验：超几何分布（hypergeometric）是统计学上一种离散概率分布。它描述了在 N 个物件中指定 M 个种类的物件，不放回的抽取 n 个，成功抽中指定类型物件的个数（k）的事件。

（注意：相对于 GOKEGG 富集分析模块，本模块是在同样的富集方法的基础上，增加同时分析多组的功能，比较不同分组的富集结果可以揭示不同分组之间的功能共性和差异）

应用场景

如果手上有一堆分子列表，想要看这一堆分子中都涉及哪个方面的功能和通路。注意：单独几个分子做富集分析是没有意义的，单独几个分子直接去查对应分子的功能注释即可，无须做富集分析。

另外，GO 库和 KEGG 库中的有注释的分子一般都是编码分子，如果手上有一堆非编码如 miRNA 或者 lncRNA 或者 circRNA 是没办法直接做富集分析的。一般这种会先找对应的靶功能分子，通过对靶分子富集分析来反向推断设计的功能和通路。

主要结果

	A	B	C	D	E	F	G	H	I	J	K
1	Group	ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
2	Up1	BP	GO:0010948	negative reg	8/40	301/18800	1.939E-07	0.00030985	0.00020492	AURKB/BMP	8
3	Up1	BP	GO:0070098	chemokine-r	5/40	89/18800	1.2255E-06	0.00060063	0.00039723	ACKR1/CX3C	5
4	Up1	BP	GO:0045786	negative reg	8/40	387/18800	1.2964E-06	0.00060063	0.00039723	AURKB/BMP	8
5	Up1	BP	GO:1990868	response to	5/40	97/18800	1.8793E-06	0.00060063	0.00039723	ACKR1/CX3C	5
6	Up1	BP	GO:1990869	cellular resp	5/40	97/18800	1.8793E-06	0.00060063	0.00039723	ACKR1/CX3C	5
7	Up1	BP	GO:0016572	histone pho	4/40	45/18800	2.4575E-06	0.00065452	0.00043287	AURKB/CCN	4
8	Up1	BP	GO:0045839	negative reg	4/40	48/18800	3.1947E-06	0.0007293	0.00048232	AURKB/BMP	4
9	Up1	BP	GO:0051784	negative reg	4/40	55/18800	5.5398E-06	0.00110658	0.00073184	AURKB/BMP	4
10	Up1	BP	GO:0045930	negative reg	6/40	234/18800	9.3948E-06	0.00165228	0.00109274	AURKB/BMP	6
11	Up1	BP	GO:0051783	regulation o	5/40	139/18800	1.0984E-05	0.00165228	0.00109274	AURKB/BMP	5
12	Up1	BP	GO:0000086	G2/M transit	5/40	140/18800	1.1374E-05	0.00165228	0.00109274	AURKB/CCN	5
13	Up1	BP	GO:1901988	negative reg	6/40	255/18800	1.531E-05	0.00203875	0.00134833	AURKB/CCNI	6

- Group：对应输入的分组信息（上传文件的列名）
- ONTOLOGY：类目，包括 BP、CC、MF、KEGG
- ID：对应的功能或者通路的 ID 编号，由数据库给定。
- Description：对应的功能或者通路的名字，详细信息。
- GeneRatio：输入的分子（经过 ID 转换后）与对应 ID 条目内分子的交集总数 / 输入的分子（经过 ID 转换后）与库内（BP、CC、MF 和 KEGG 都是分开的注释库）总的有功能注释的分子的交集总数。
- BgRatio：对应 ID 条目内分子总数 / 库内（BP、CC、MF 和 KEGG 都是分开的注释库）总的有功能注释的分子的交集总数。
- pvalue：超几何分布检验统计的 p 值。
- p.adjust：通过 p 值校正方法得到的校正后的 p 值。
- qvalue：通过 p 值校正方法得到的校正后的 q 值，代表错误率。
- geneID：输入的分子（经过 ID 转换后）与对应 ID 条目内分子的交集的具体的分子 ID。
- Count：输入的分子（经过 ID 转换后）与对应 ID 条目内分子的交集总数。

（一般在文章里面常见设定 p.adj < 0.05 为显著富集的结果。结果可以展示 top 几的条目，也可以挑一些结果来放到文章里面或者进行可视化。）

数据格式

	A	B	C	D
1	Up1	Down1	Up2	Down2
2	ABCA8	FABP4	ABLIM3	IL1R2
3	ACKR1	FMO5	ADAMDEC1	INAVA
4	ADH1B	FOXA1	ADIPOQ	IQCH
5	AGR2	FOXM1	ADIRF	KCNE4
6	ANXA9	GABRP	AK5	KCNK15
7	AQP9	GAMT	APOBEC3B	KIAA0754
8	AURKB	GATA3	ASPM	KIF18A
9	BCL2A1	GPR19	ASPN	KIF18B
10	BIRC5	GRP	AURKA	KIF20A
11	BMP4	GZMB	C1orf21	KIF23
12	C7	HLA-DQA1	C3orf18	KIF4A
13	CA12	HLA-DQA2	CALML5	LAMP3

数据要求：

- 至少 1 列数据，第 1 行为列名，将作为分组信息进行数据整理。每 1 列代表 1 个分组
- 每 1 列为不同分组的分子列表，可以是分子名、Ensembl 编号、Entrez ID，不支持非编码基因(lncRNA/miRNA 等)。
- 分子至少是 10 个以上，10 个以下无法进行富集分析。

提醒：分子列表不是把所有的分子都放入，一般是差异分子列表（数目大概是在几十到几百不等），一次性放入上千或者上万个分子得到的富集分析结果没有实际的参考价值。

参数说明

（说明：标注了颜色的为常用参数。）

类别

- 物种：物种选择，可以选择人源(Homo sapiens)、小鼠(Mus musculus)、大鼠(Rattus norvegicus)。

富集参数

- 条目：可选 GO+KEGG、GO、KEGG、GO:BP、GO:CC、GO:MF 等，默认全部 GO+KEGG。
- p 值校正方法：默认为 BH 法，一般不需要改动。如果有需要也可以进行相应修改。

结果说明

主要结果


主要结果 补充结果 方法学 保存结果 下载整份报告

GOKEGG分组

GOKEGG分组: 拿一堆有功能注释的分子与所有有功能注释的分子去比较（超几何分布检验），确定那一堆中都涉及了哪些功能作用或者通路。比较不同分组的富集结果可以揭示不同分组之间的功能共性和差异。

过程: 输入的分子列表 →(内部)→ 转成Entrez ID →(内部)→ 根据分组在GOKEGG的库进行超几何分析 → 获得各分组的分析结果

Group	ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Up1	BP	GO:0010948	negative regulation of cell c...	8/40	301/18800	1.938983e-07	3.098495e-04	2.049199e-04
Up1	CC	GO:0098687	chromosomal region	7/40	366/19594	8.213998e-06	9.281817e-04	6.311809e-04
Up1	MF	GO:0035173	histone kinase activity	3/38	16/18410	4.459956e-06	6.415481e-04	4.536198e-04
Up1	KEGG	hsa04110	Cell cycle	6/29	126/8164	4.259266e-06	4.387044e-04	3.990260e-04
Down1	BP	GO:0010965	regulation of mitotic sister c...	5/52	65/18800	9.684173e-07	4.698010e-04	3.896998e-04
Down1	CC	GO:0072686	mitotic spindle	4/53	160/19594	9.184963e-04	4.949544e-02	4.113194e-02
Down1	MF	GO:0032395	MHC class II receptor activity	2/52	10/18410	3.470636e-04	6.455383e-02	4.822357e-02
Down1	KEGG	hsa05330	Allograft rejection	3/27	38/8164	2.519592e-04	1.421750e-02	1.125632e-02
Up2	BP	GO:0000212	meiotic spindle organization	3/37	15/18800	3.141279e-06	2.824010e-03	2.169136e-03
Up2	CC	GO:0000775	chromosome, centromeric r...	6/43	227/19594	9.642806e-06	6.484437e-04	4.603395e-04
Up2	MF	GO:0008009	chemokine activity	4/37	49/18410	2.741683e-06	2.193347e-04	1.500711e-04
Up2	KEGG	hsa04061	Viral protein interaction wit...	4/15	100/8164	2.608345e-05	1.851925e-03	1.647376e-03

GOKEGG分组.xlsx GOKEGG分组.docx

主要结果格式为表格结果，提供 Excel、Docx 格式下载。

注意：页面仅展示各条目类型的前 n 个结果，Word 三线表同页面的情况。所有的富集结果需要下载 Excel 表来进行查看。

如果需要富集结果进行可视化，请先保存结果，保存成功后再到[GOKEGG 分组]对应的可视化模块直接进行可视化。如果删除了数据记录，将无法进行可视化。

补充结果

ID转换情况

输入的分子列表会先转成Entrez id后再进行GOKEGG富集分析。（备注：这里转化id用到的R包是%s，一般转换的总数和比例不要太少即可。）

Up1		
输入ID总数	成功转化的ID总数	转换比例(%)
41	41	100
Down1		
输入ID总数	成功转化的ID总数	转换比例(%)
53	53	100
Up2		
输入ID总数	成功转化的ID总数	转换比例(%)
45	45	100
Down2		
输入ID总数	成功转化的ID总数	转换比例(%)
70	69	98.6

ID转换情况.xlsx

此表格提供各分组 ID 转换情况，上传的分子都会换成 Entrez ID，只要这个转换比例不要过低（<10%）影响到富集分析即可，提供 Excel 格式下载。

GOKEGG分组富集情况

GOKEGG分组富集分析显著性cut-off值一般设置为 校正后p值<0.05

Up1				
阈值条件	BP	CC	MF	KEGG
p.adj<0.1	254	35	19	8
p.adj<0.05	168	17	17	3
Down1				
阈值条件	BP	CC	MF	KEGG
p.adj<0.1	92	10	11	21
p.adj<0.05	57	3		9
Up2				
阈值条件	BP	CC	MF	KEGG
p.adj<0.1	85	17	10	6
p.adj<0.05	56	15	9	4
Down2				
阈值条件	BP	CC	MF	KEGG
p.adj<0.1	51	23	10	
p.adj<0.05	35	22	9	

此表格提供在一些阈值条件下 各分组 各个类目的数目，一般富集分析有意义的定义是 p.adj<0.05。

方法学

所有分析和可视化均在 R 4.2.1 中进行

涉及的 R 包：clusterProfiler 包（用于富集分析), org.Hs.eg.db 包（用于 ID 转换）

处理过程：

⑴ 对输入的分子列表进行 ID 转换

⑵ 根据分组，用 clusterProfiler 包进行分组富集分析。

如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分（Acknowledge）致谢仙桃学术（www.xiantao.love）。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

常见问题

1. GOKEGG 需要输入的内容?

答:

一般是差异分子列表，或者其他来源的几十到上百个分子组成的列表。

2. 为什么才富集了这么一些?

答:

页面仅仅展示了前几个的结果,所有的富集结果需要下载 excel 表格来进行查看。一般大于 4 个分组,页面仅展示各分类中的 1 个,具体还是要下载所有结果来看。

3. 分子列表怎么来?

答:

分子列表一般是两组间进行差异分析后按照阈值过滤后得到的分子列表,当然也可以是其他方式得到的分子列表（比如批量相关性分析，或者预测靶分子等），只要能有一堆功能基因分子列表，即可做富集分析。本模块为多组富集分析，即可前提进行多个分组的差异分析，如同一项研究中，存在多个实验处理，treat1-vs-control、treat2-vs-control、treat3-vs-control 分别进行差异分析，分别筛选显著差异分子后，按照本模块数据格式整理数据后，即可进行分组富集分析。

4. 富集分析结果不好（结果很少或者只有其中的一类），怎么办?

答:

可以试试别的富集分析的数据库，比如 metascape 等。

5. 已经输入了很多分子，但是富集结果不好，是什么问题?

答：

① 首先要关注 ID 转换情况，如果补充结果中 ID 转换比例很低，这个是会影响到富集的结果的。

② 其次是要注意分子类型是否是编码基因，如果很多都是比如 miRNA 或者 lncRNA，这些是没有功能注释的分子，这些分子都是没办法进行富集分析的。

如果在功能基因中混有一些这些分子，是不需要手动剔除的，一般是不怎么会影响结果的。

6. 结果的排序规则是什么？

答：

结果是 各分组 按照校正后的 p 值进行排序的。

7. 我用别的数据库（比如 DAVID）做的结果为什么跟工具做的不一样？

答：

主要由于不同的注释库的差异导致的，工具是利用 R 中的 org.Hs.eg.db 包作为注释库以及 ID 转换的。统计学检验的方法应该都是类似的。所以出现了不同的结果也是很常见的。

8. 如何进行可视化？

答：

在 GOKEGG 分组 分析模块完成后，点击保存结果，此时数据记录会保存到历史记录中，同时下载对应的结果文件，然后到 GOKEGG 分组 可视化模块中，选择对应的数据记录，即可进行可视化。想要修改可视化的条目，可以从结果表格中复制 ID 到 可视化 ID 参数中。

9. 如何进行 KEGG 通路分析?

答:

在富集参数中的 **条目** 参数中，选择 KEGG 即可。

10. 我已经选了 GO+KEGG，为什么结果里面只有 BP 或者 CC 或者 MF 或者 KEGG 中的一种，其他的都没有？为什么有一些类（BP、CC、MF、KEGG）只有一条或者少数几条结果?

答:

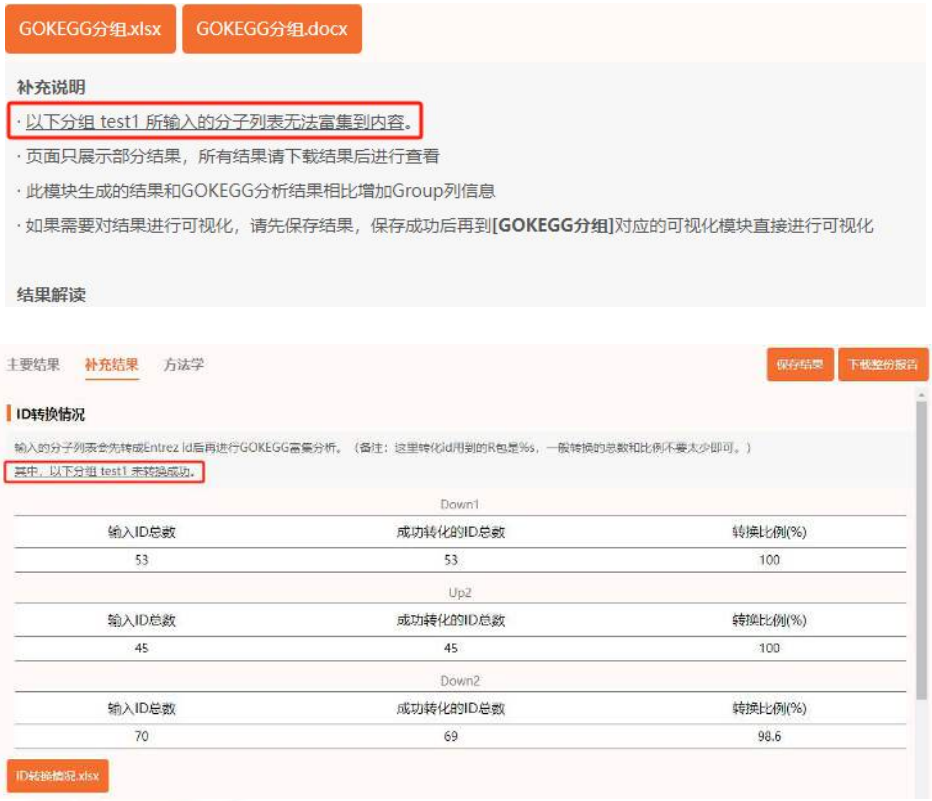
最终的表格只保留了满足较宽的阈值（p<0.1 以及 qvalue<0.2）的结果，而不满足这一较宽阈值下的条目都会被过滤，如果整个类（BP、CC、MF、KEGG）都不满足这个阈值，那么最终的表格中就会缺少这个类。如果富集的结果不是很理想，可以尝试别的富集分析的数据库，比如 metascape 等。

11. 为什么分组少了?

答:

本模块是在 GOKEGG 富集分析 模块的基础上，增加同时分析多组的功能，分析也是按照分组（即按照上传文件中的列）进行富集分析。缺少分组时，可以先看主要结果-补充说明部分。结合问题 5 中的说明，可以了解到，一般是对应分组（列）中的分子无法进行 ID 转换等情况，从而导致无法进行富集分析。

	A	B	C	D
1	test1	Down1	Up2	Down2
2	a	FABP4	ABLIM3	IL1R2
3	b	FMO5	ADAMDEC1	INAVA
4	c	FOXA1	ADIPOQ	IQCH
5	d	FOXM1	ADIRF	KCNE4
6		GABRP	AK5	KCNK15
7		GAMT	APOBEC3B	KIAA0754


GOKEGG分组.xlsx
GOKEGG分组.docx
补充说明
· 以下分组 test1 所输入的分子列表无法富集到内容。
· 页面只展示部分结果，所有结果请下载结果后进行查看
· 此模块生成的结果和GOKEGG分析结果相比增加Group列信息
· 如果需要对结果进行可视化，请先保存结果，保存成功后再到[GOKEGG分组]对应的可视化模块直接进行可视化
结果解读
主要结果
补充结果
方法学
保存结果
下载整份报告
ID转换情况
输入的分子列表会先转成Entrez Id后再进行GOKEGG富集分析。（备注：这里转化id用到的R包是%s，一般转换的总数和比例不要太少即可。）
其中，以下分组 test1 未转换成功。
Down1
输入ID总数
成功转化的ID总数
转换比例(%)
53
53
100
Up2
输入ID总数
成功转化的ID总数
转换比例(%)
45
45
100
Down2
输入ID总数
成功转化的ID总数
转换比例(%)
70
69
98.6
ID转换情况.xlsx