

功能聚类 – GOKEGG 分析

ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
BP	GO:0140014	mitotic nuclear division	31/197	293/18800	2.61e-22	7.82e-19	6.71e-19
BP	GO:0000280	nuclear division	35/197	446/18800	8.47e-21	1.27e-17	1.09e-17
BP	GO:0000070	mitotic sister chromatid se...	24/197	171/18800	2.26e-20	1.83e-17	1.57e-17
CC	GO:0005819	spindle	26/203	402/19594	9.72e-14	2.88e-11	2.53e-11
CC	GO:0072686	mitotic spindle	17/203	160/19594	8.73e-13	1.29e-10	1.14e-10
CC	GO:0000775	chromosome, centromeric r...	19/203	227/19594	2.81e-12	2.77e-10	2.44e-10
MF	GO:0008017	microtubule binding	19/192	272/18410	7.14e-11	3.28e-08	3e-08
MF	GO:0015631	tubulin binding	19/192	376/18410	1.57e-08	3.6e-06	3.3e-06
MF	GO:0003777	microtubule motor activity	8/192	67/18410	4.66e-07	6.74e-05	6.18e-05
KEGG	hsa04110	Cell cycle	11/95	126/8164	1.99e-07	4.18e-05	3.98e-05
KEGG	hsa04114	Oocyte meiosis	10/95	131/8164	2.55e-06	0.0003	0.0003
KEGG	hsa04218	Cellular senescence	10/95	156/8164	1.22e-05	0.0009	0.0008

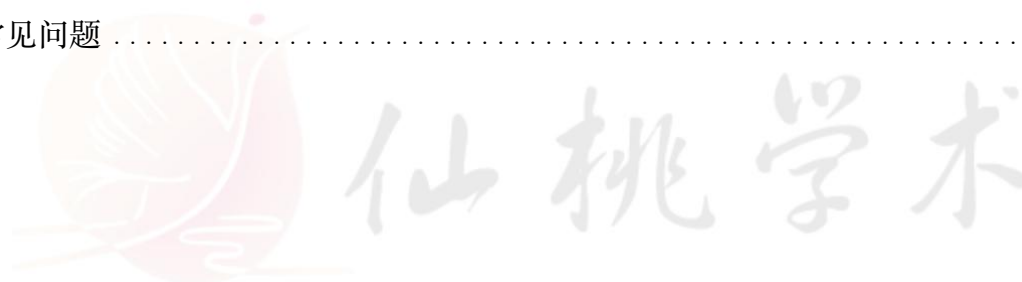
网址: <https://www.xiantao.love>



更新时间: 2023.02.09

目录

基本概念	3
应用场景	4
主要结果	5
数据格式	6
参数说明	7
分子列表	7
富集参数	8
结果说明	9
主要结果	9
补充结果	10
方法学	11
如何引用	12
常见问题	13



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。

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
1	ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
2	BP	GO:0140014	mitotic nucle	31/197	293/18800	2.6093E-22	7.8228E-19	6.7101E-19	BMP4/CCNB1	31
3	BP	GO:0000280	nuclear divisi	35/197	446/18800	8.4685E-21	1.2694E-17	1.0889E-17	BMP4/CCNB1	35
4	BP	GO:0000070	mitotic sister	24/197	171/18800	2.2624E-20	1.8261E-17	1.5664E-17	CCNB1/CDC20	24
5	BP	GO:0048285	organelle fissi	36/197	493/18800	2.4365E-20	1.8261E-17	1.5664E-17	BMP4/CCNB1	36
6	BP	GO:0000819	sister chroma	25/197	205/18800	1.1768E-19	7.0563E-17	6.0527E-17	CCNB1/CDC20	25
7	BP	GO:0007059	chromosome	28/197	348/18800	4.7484E-17	2.3726E-14	2.0351E-14	CCNB1/CDC20	28
8	BP	GO:1902850	microtubule c	20/197	151/18800	1.2161E-16	5.2083E-14	4.4675E-14	CCNB1/CDC20	20
9	BP	GO:0098813	nuclear chron	25/197	287/18800	3.9927E-16	1.4963E-13	1.2834E-13	CCNB1/CDC20	25
10	BP	GO:0007052	mitotic spindl	17/197	124/18800	1.4288E-14	4.7596E-12	4.0826E-12	CCNB1/CDC20	17
11	BP	GO:0007051	spindle organ	19/197	188/18800	1.123E-13	3.3666E-11	2.8878E-11	CCNB1/CDC20	19

- 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。

数据格式

	A	B	C
1	id		
2	NAT1		
3	ADH1B		
4	BIRC5		
5	AQP9		
6	BCL2A1		
7	BMP4		
8	C7		
9	CA12		
10	CACNA1D		
11	CAMP		
12	CCNA2		
13	CCNB1		
14	CDK1		
15	CDC20		
16	CENPE		

数据要求提供 1 列：

- 第一列除了列名外，下面的可以是分子名、Ensembl 编号、Entrez ID
- 分子至少是 10 个以上，10 个以下无法进行富集分析。
- 上传数据后，分子会出现在主要参数的 分子列表 的参数 分子 ID 中。



也可以直接将分子列表复制黏贴到上面的输入框内，每一行代表一个分子。

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。)

分子列表

分子列表

分子ID

- NAT1
- ADH1B
- BIRC5
- AQP9
- BCL2A1
- BMP4
- C7
- CA12

物种 人源(Homo)

- **分子 ID**：待分析的分子列表，可以直接从表格中复制分子到这里，类型支持 基因名 | Ensembl | Entrez ID 三类，输入的 ID 会自动转换成 Entrez ID 后进行富集分析(标准流程)。每一行代表一个分子。至少要 10 个分子以上才能进行富集分析。
- 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

富集参数

▼

条目

全部GO+KEGG

▼

p值校正方法

BH

▼

- **条目**：可选 GO、KEGG、GO:BP、GO:CC、GO:MF 等。
- **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富集分析

GOKEGG: 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。

过程: 输入的分子列表 → (内部) → 转成Entrez ID → (内部) → 在GOKEGG的库进行超几何分析 → 获得分析结果

ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID
BP	GO:0140014	mitotic nuclear division	31/197	293/18800	2.61e-22	7.82e-19	6.71e-19	BMP4/CCNB1/CDC2
BP	GO:0000280	nuclear division	35/197	446/18800	8.47e-21	1.27e-17	1.09e-17	BMP4/CCNB1/CDC2
BP	GO:0000070	mitotic sister chromatid se...	24/197	171/18800	2.26e-20	1.83e-17	1.57e-17	CCNB1/CDC20/CEN
BP	GO:0048285	organelle fission	36/197	493/18800	2.44e-20	1.83e-17	1.57e-17	BMP4/CCNB1/CDC2
BP	GO:0000819	sister chromatid segregation	25/197	205/18800	1.18e-19	7.06e-17	6.05e-17	CCNB1/CDC20/CEN
CC	GO:0005819	spindle	26/203	402/19594	9.72e-14	2.88e-11	2.53e-11	BIRC5/CCNB1/CDK
CC	GO:0072686	mitotic spindle	17/203	160/19594	8.73e-13	1.29e-10	1.14e-10	CDK1/CENPE/KIF1
CC	GO:0000775	chromosome, centromeric r...	19/203	227/19594	2.81e-12	2.77e-10	2.44e-10	BIRC5/CCNB1/CEN
CC	GO:0098687	chromosomal region	23/203	366/19594	5.15e-12	3.81e-10	3.35e-10	BIRC5/CCNB1/CDK
CC	GO:0000779	condensed chromosome, ce...	16/203	156/19594	7.23e-12	4.28e-10	3.76e-10	BIRC5/CCNB1/CEN
MF	GO:0008017	microtubule binding	19/192	272/18410	7.14e-11	3.28e-08	3e-08	BIRC5/CENPE/KIF1
MF	GO:0015631	tubulin binding	19/192	376/18410	1.57e-08	3.6e-06	3.3e-06	BIRC5/CENPE/KIF1

[GOKEGG.xlsx](#)
[GOKEGG.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 5 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]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

输入分子列表会先转成Entrez id后再进行GOKEGG富集分析

输入ID总数	成功转化的ID总数	转换比例(%)
209	208	99.5

ID转换情况.xlsx

一共输入了209个ID，其中，成功转成208个Entrez ID，转换的百分比为99.5。（备注：这里转化id用到的R包是org.Hs.eg.db，一般转换的总数和比例不要太少即可。）

此表格提供 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富集情况

GOKEGG富集分析显著性cut-off值一般设置为 校正后p值<0.05

阈值条件	BP	CC	MF	KEGG
p.adj<0.1	258	35	23	11
p.adj<0.05	188	26	21	10

此表格提供在一些阈值条件下各个类目的数目，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:

(1) 对输入的分子列表进行 ID 转换后, 用 clusterProfiler 包进行富集分析。



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 5 的结果, 所有的富集结果需要下载 excel 表格来进行查看。

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。

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 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 到分子 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。