功能聚类 - GOKEGG 联合 FC 分析

ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	zscore
BP	GO:0140014	mitotic nuclear division	31/197	293/18800	2.61e-22	7.82e-19	6.71e-19	-0.53882
BP	GO:0000280	nuclear division	35/197	446/18800	8.47e-21	1.27e-17	1.09e-17	-1.1832
BP	GO:0000070	mitotic sister chromatid se...	24/197	171/18800	2.26e-20	1.83e-17	1.57e-17	-0.8165
CC	GO:0005819	spindle	26/203	402/19594	9.72e-14	2.88e-11	2.53e-11	-1.1767
CC	GO:0072686	mitotic spindle	17/203	160/19594	8.73e-13	1.29e-10	1.14e-10	-0.72761
CC	GO:0000775	chromosome, centromeric regi...	19/203	227/19594	2.81e-12	2.77e-10	2.44e-10	-1.1471
MF	GO:0008017	microtubule binding	19/192	272/18410	7.14e-11	3.28e-08	3e-08	-2.0647
MF	GO:0015631	tubulin binding	19/192	376/18410	1.57e-08	3.6e-06	3.3e-06	-2.0647
MF	GO:0003777	microtubule motor activity	8/192	67/18410	4.66e-07	6.74e-05	6.18e-05	-0.70711
KEGG	hsa04110	Cell cycle	11/95	126/8164	1.99e-07	4.18e-05	3.98e-05	2.1106
KEGG	hsa04114	Oocyte meiosis	10/95	131/8164	2.55e-06	0.0003	0.0003	0.63246
KEGG	hsa04218	Cellular senescence	10/95	156/8164	1.22e-05	0.0009	0.0008	1.8974

网址：https://www.xiantao.love

更新时间：2023.02.13

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。
- 富集分析联合 logFC：就是在富集分析的基础上，利用提供的分子的 logFC，计算每个条目对应的 zscore，初步判断对应的条目是正调节（zscore 为正）还是负调节（zscore 为负）。zscore 计算方法见下：

$$zscore = \frac{(Up - Down)}{\sqrt{Counts}}$$

  - 其中，这里的 Up Down 代表对应条目分子的 logFC 为正以及为负分别对应数量，Counts 代表条目对应的分子总数（这里不是指 Z-score 标准化，是 GOplot 包所使用的概念和提供的方法）

（注意：相对于 GOKEGG 富集分析模块，这个模块只是在同样的富集方法的基础上，另外再计算了每个条目对应的 zscore 值）

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。

- 如果是单个分子，可以考量做单基因差异分析【在足够样本量的疾病组中，按照某个分子的表达分成高低表达组，模拟过表达或者敲减的效果，分析高低表达组两组的差异，从而获得差异分子列表】或者单基因相关性分析【在足够样本量的疾病组中，分析指定分子和其他分子的相关性，设定一个相关性阈值（0.2,0.3 ...或者 p 值），从而获得有显著相关的分子列表】，获得分子列表后再做 GOKEGG 分析，反向推倒这个分子可能涉及的功能或者通路。

主要结果

	A	B	C	D	E	F	G	H	I	J	K
1	ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count	zscore
2	BP	GO:0140014	mitotic nucle	31/197	293/18800	2.609E-22	7.823E-19	6.71E-19	BMP4/CCNB	31	-0.538816
3	BP	GO:0000280	nuclear divis	35/197	446/18800	8.469E-21	1.269E-17	1.089E-17	BMP4/CCNB	35	-1.183216
4	BP	GO:0000070	mitotic siste	24/197	171/18800	2.262E-20	1.826E-17	1.566E-17	CCNB1/CDC2	24	-0.816497
5	BP	GO:0048285	organelle fis	36/197	493/18800	2.436E-20	1.826E-17	1.566E-17	BMP4/CCNB	36	-1
6	BP	GO:0000819	sister chrom	25/197	205/18800	1.177E-19	7.056E-17	6.053E-17	CCNB1/CDC2	25	-1
7	BP	GO:0007059	chromosome	28/197	348/18800	4.748E-17	2.373E-14	2.035E-14	CCNB1/CDC2	28	-1.511858
8	BP	GO:1902850	microtubule	20/197	151/18800	1.216E-16	5.208E-14	4.467E-14	CCNB1/CDC2	20	0
9	BP	GO:0098813	nuclear chro	25/197	287/18800	3.993E-16	1.496E-13	1.283E-13	CCNB1/CDC2	25	-1
10	BP	GO:0007052	mitotic spinc	17/197	124/18800	1.429E-14	4.76E-12	4.083E-12	CCNB1/CDC2	17	0.2425356
11	BP	GO:0007051	spindle orga	19/197	188/18800	1.123E-13	3.367E-11	2.888E-11	CCNB1/CDC2	19	-0.229416
12	BP	GO:0051304	chromosome	14/197	97/18800	1.579E-12	4.303E-10	3.691E-10	CCNB1/CENF	14	0.5345225
13	BP	GO:0010965	regulation of	12/197	65/18800	3.103E-12	7.752E-10	6.649E-10	CCNB1/CENF	12	1.1547005

- 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。
- zscore：条目对应分子中表达增高分子 和 表达降低分子 之差 和 条目分子开根号 相除的结果，如果为正，说明对应的条目可能是正调节，如果为负，对应条目可能是负调节；绝对值越大，说明高表达分子和低表达分子的数量差相对比较大，说明调节程度可能更高。

（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
1	id	logFC
2	NAT1	2.921874291
3	ADH1B	2.104560046
4	BIRC5	2.879420509
5	AQP9	2.918582581
6	BCL2A1	2.11695989
7	BMP4	2.334144218
8	C7	2.202394322
9	CA12	2.735250901
10	CACNA1D	2.422592568
11	CAMP	2.638625193
12	CCNA2	2.373657878
13	CCNB1	2.428388797

数据要求提供 2 列：

- 第 1 列除了列名外，下面的可以是分子名、Ensembl 编号、Entrez ID
  - 分子为差异分析有显著的分子列表（一般几十到几百或者上千不等），不是所有的分子！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。
  - 分子至少是 10 个以上，10 个以下可能无法进行富集分析。
- 第 2 列为分子对应的 logFC 值（这个值来自于差异分析后的结果中，如果是相关性分析，则为相关系数）。

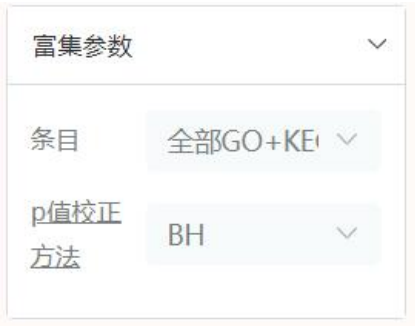
参数说明

（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)。

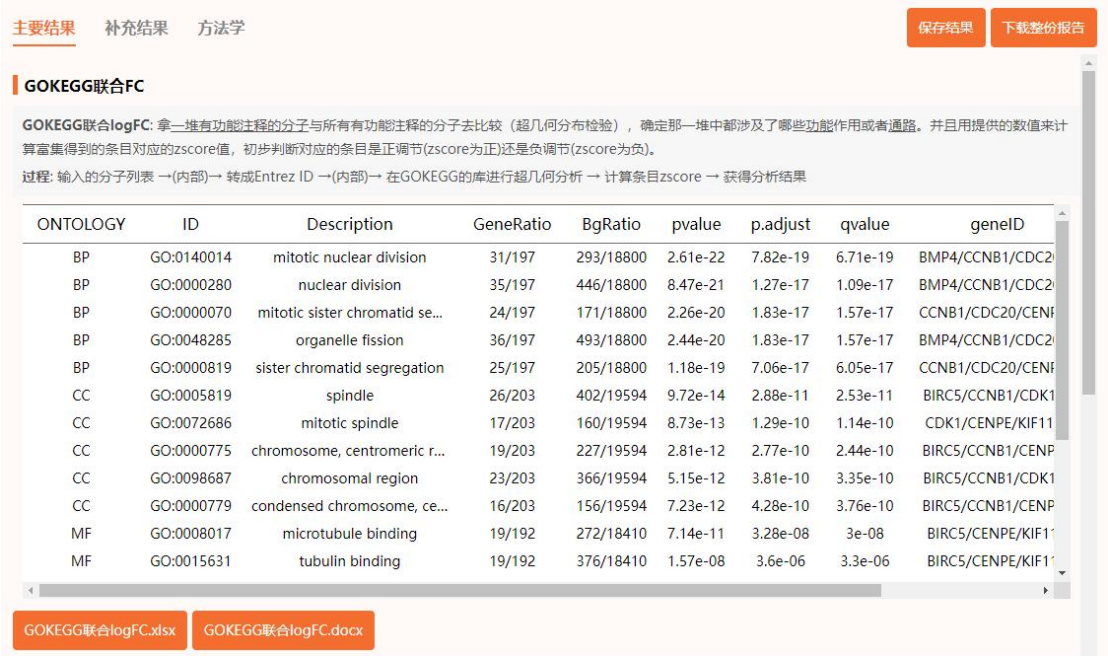
富集参数



- 条目：条目所属，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。

如果需要将 GOKEGG 富集（联合 logFC）结果进行可视化，请先保存结果，保存成功后再到【GOKEGG 联合 FC】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

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），GOplot 包（用于计算 zscore）

处理过程：

⑴ 对输入的分子列表进行 ID 转换后，用 clusterProfiler 包进行富集分析，利用提供的分子的数值通过 GOplot 包计算每个富集条目对应的 zscore 值。

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 联合 FC 模块需要输入的内容?

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。差异分子列表不是进行差异分析后的所有分子，而是经过 logFC 和校正后 p 值过滤后的有（显著）差异的分子。

2. 结果中 zscore 是什么，这个值能说明什么?

答:

zscore 的计算方法来自 GOplot 包，计算方法见下:

$$zscore = \frac{(Up - Down)}{\sqrt{Counts}}$$

其中，这里的 Up Down 代表对应条目分子的 logFC 为正以及为负分别对应数量，Counts 代表条目对应的分子总数。

如果 zscore 为正，说明对应的条目可能是正调节，如果为负，对应条目可能是负调节；绝对值越大，说明高表达分子和低表达分子的数量差相对比较大，说明调节程度可能更高。

注意，GOplot 提供的计算 zscore 方法是没有考虑分子在对应的条目里面是对这个条目正调节还是负调节的，也就存在如果有低表达的负调节的分子，在 zscore 里面是记为 down，但是因为负负得正，应该是对这个条目正调节，记为正才合理。GOplot 就只是提供了这个计算方法，而且 GOKEGG 库里面也并没有记录每个条目每个分子是对这个条目是正还是负调节的数据信息，尚且都还达不到这个粒度，所以这个 zscore 仅仅只能作为一种可能性参考。

3. 这个模块和 GOKEGG 富集分析模块的差别是什么?

答:

这个模块和 GOKEGG 富集分析用的是一样的富集分析的内容（富集出来的结果是一致的），只是在这个基础上，还会另外再计算每个条目对应的 zscore 值（GOplot 包提供）。

这个模块的分析结果会有对应专属的可视化模块（因为有 zscore 值，会有更多的可视化模块），GOKEGG 富集分析是对应另外的可视化模块。

4. 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。

5. 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:

具体原因可以见下一个问题。如果还是没有解决，可以试试别的富集分析的数据库，比如 metascape 等。

6. 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。

7. 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。

8. 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。

9. 如何进行可视化?

答:

在 【GOKEGG 联合 FC】 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 联合 FC】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。

10. 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。

11. 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。