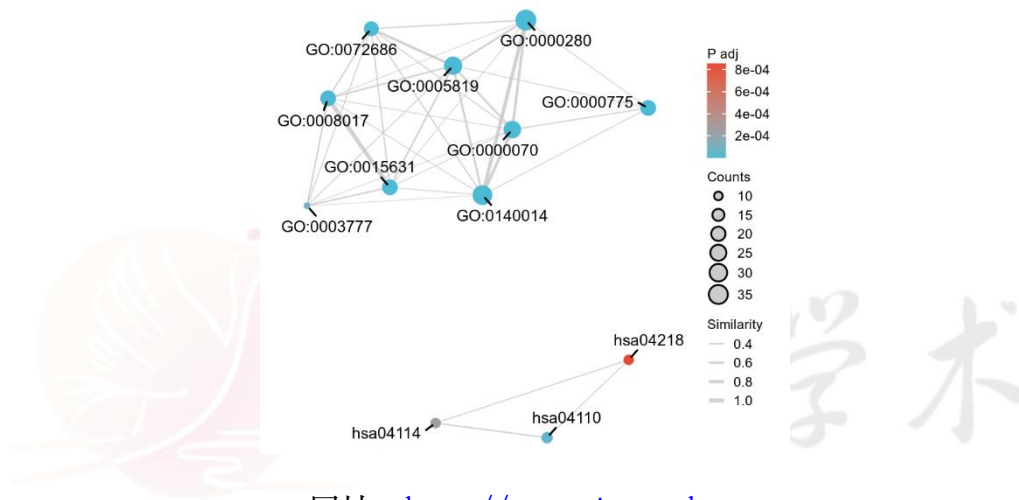


功能聚类 - GOKEGG-EMAP 图



网址: <https://www.xiantao.love>



更新时间: 2023.10.19

目录

基本概念	3
应用场景	3
主要结果	5
云端数据	6
参数说明	7
ID 列表	7
阈值	8
节点映射	9
样式	9
连线	10
点	11
标注	12
标题	13
图注	14
风格	14
图片	15
结果说明	16
主要结果	16
补充结果	17
方法学	18
如何引用	19
常见问题	20

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。
- Jaccard 相似性指数：Jaccard similarity index, 定义为 A 与 B 交集的大小与 A 与 B 并集的大小的比值。本模块即使用 JC 方法，计算 term 与 term 之间交集基因的占比，从而获得一个类目与另一个类目之间的成对相似性。

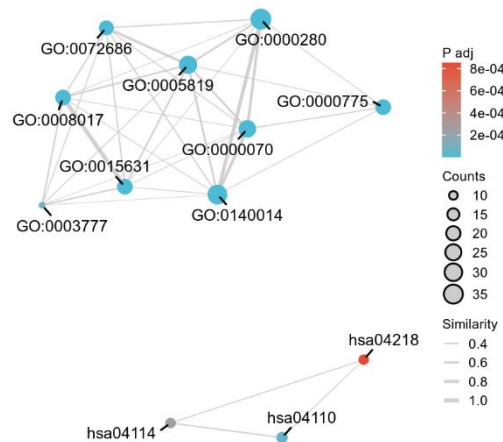
应用场景

本模块为 GO|KEGG 富集分析后结果的可视化展示。EMAP 图的形式，对于所选择的类目 (ID)，先计算类目之间的成对相似性，再以网络图对类目间的相似性进行可视化，更直观的展示功能模块之间的联系。

注意：模块需要先进行 GO|KEGG 富集分析 并 保存结果后，此处的云端数据才会有结果记录，然后才能进行可视化的操作。主要基本参数中有 **ID 列表** 输入框，可以将对应云端数据记录的富集分析表格中的感兴趣的 ID 列复制到此处，进行可视化。



主要结果



通过 EMAP 图的形式展示 GOKEGG 富集分析结果。

- 例图中，节点代表类目，连线代表条目和条目之间的关系，**线条粗细代表了条目之间的成对相似性(Similarity)**，线越粗相似性更高，反之更低。
- 节点的颜色为所选择的 **颜色映射(点)** 内容。(如上图，对应校正后 p 值)
- 节点的大小为所选择的 **大小映射(点)** 内容。(如上图，对应包含 ID 的数量)
- 节点的形状为所选择的 **形状映射(点)** 内容。(默认 无)

默认展示各类别的 top 几个结果 (默认 满足校正后 p 值<0.05)，分面是对应的数据库或者分类。可以挑选在满足阈值下的 top 的类目，或者一些感兴趣的类目。

云端数据

云端数据

	记录名称	来源模块	时间	补充说明
<input checked="" type="checkbox"/>	GOKEGG	GOKEGG富集分析 @1.0	2023-02-06 16:14:19	数据记录可以在历史记录中找到

这里的云端数据与历史记录汇总 GOKEGG 富集分析模块的数据记录是保持一致的，可以在历史记录中找到相应的数据记录。

根据需要可视化的项目 选择好对应的云端数据记录。默认使用最近生成的分析记录。



参数说明

(说明：标注了颜色的为常用参数。)

ID 列表

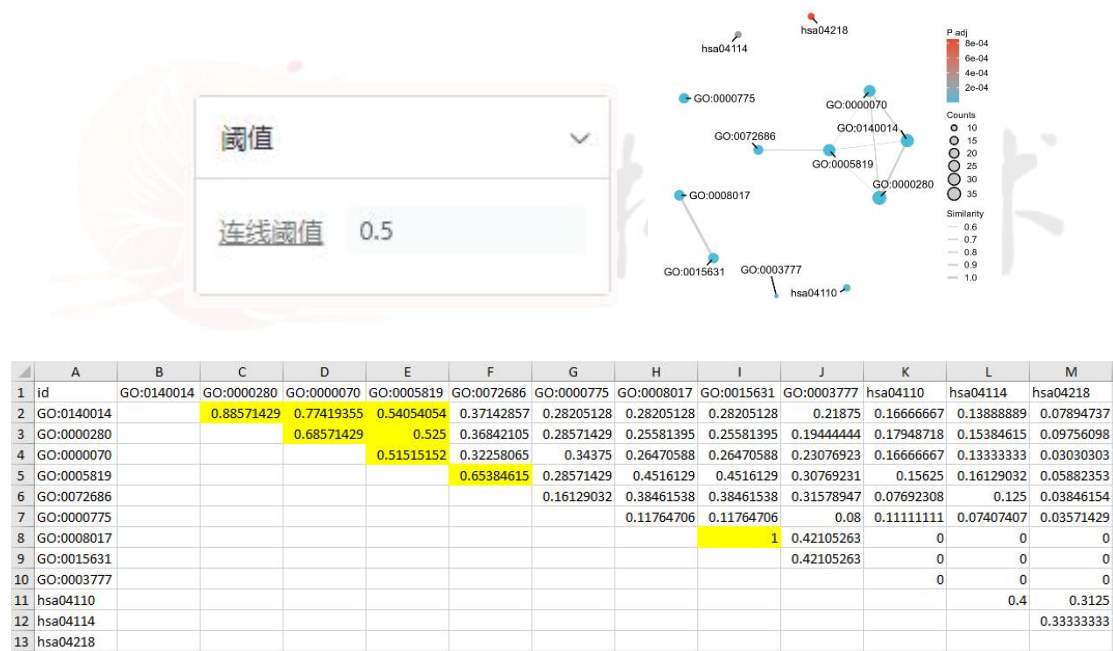


- 可视化 ID: 输入想要可视化的功能或者通路的条目 ID，默认为对应云端数据结果中每个类目前几个条目，可以根据需要进行输入修改。注意：输入的 ID 来自所选云端数据记录的结果，需要先在历史记录中找到对应的记录，下载 excel 结果，复制想要展示的 ID 到这个输入框中，一行代表一个。最多支持 1 张图同时绘制 30 个类目，至少绘制 2 个类目。

阈值



- **连线阈值**：可以设置连线之间的最小阈值，默认 0.2，节点之间的连线代表条目之间的成对相似性(Similarity)，可通过下载相似性表格查阅。该设置将影响连线的显示，如下：



(注意，当设置的阈值超过数据范围，图形不进行改变)

节点映射

节点映射

颜色映射 校正后p值(p)

大小映射 包含ID的数量

形状映射 无

- 颜色映射：主要影响点的颜色范围，注意映射内容的数值类型，数值型数据为渐变色，分类型数据为单个颜色。可选择 p 值(pvalue)、校正后 p 值(padj)、q 值(qvalue)(错误率)、类别(Ontology)、无。
- 大小映射：主要影响点的大小，可选择 包含 ID 的数量、无。
- 形状映射：主要影响点的形状，可选择 类别(Ontology)、无。

样式

样式

布局样式 发散(kk)

连线样式 直线

- 布局样式：网络图的布局形式。可选择 发散(kk)、发散(fr)、发散(nicely)、环形(circle)、球体(sphere)、环绕(star)、点阵(grid)、随机(randomly)。
- 连线样式：节点之间的连线样式。可选择 直线、曲线。

连线



连线

颜色

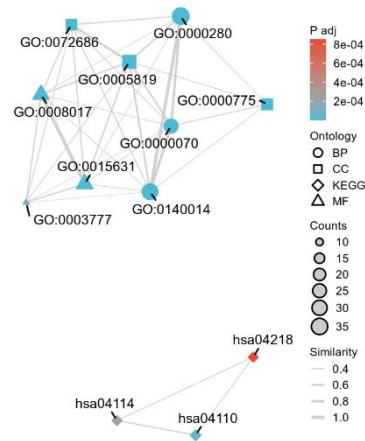
类型 实线

不透明度 0.8

- 颜色：条目节点和条目节点之间连线的颜色，最多支持修改 1 个颜色。默认灰色，不受配色方案全局性修改。
- 类型：连线的类型，可选择 实线、虚线。
- 不透明度：连线的透明度。0 为完全透明，1 为完全不透明。

点

- **填充色**：点的填充色颜色选项，取决于 [颜色映射\(点\)](#) 参数所选择的内容，展示数值型内容时，修改第一和第二色卡作为数值从小到大的渐变色；展示分类型内容（如 类别）时，有多少个功能类别会提取多少个颜色，最多支持修改 4 个颜色。受配色方案全局性修改。
- **描边色**：点的描边色颜色选项，取决于 [颜色映射\(点\)](#) 参数所选择的内容，展示数值型内容时，修改第一和第二色卡作为数值从小到大的渐变色；展示分类型内容（如 类别）时，有多少个功能类别会提取多少个颜色，最多支持修改 4 个颜色。受配色方案全局性修改。
- **样式**：点的样式类型，取决于 [形状映射\(点\)](#) 参数所选择的内容，可选择 圆形、正方形、菱形、三角形、倒三角，多选后不同的分组中点的类型也会有不同。



- 大小比例：点的相对大小，取决于 **大小映射(点)** 参数所选择的内容。
- 不透明度：点的透明度。0 为完全透明，1 为完全不透明。

标注

- 类型选择：选择所标注的变量内容，可选项 不标注、标注全部、标注类目ID、标注类目名、标注下面特定分子、标注下面特定分子。
- 特定变量：可以输入想要标注的变量，1 行 1 个，与 **类型选择** 参数中的部分选项关联使用。



- 标注大小：标注内容的字体大小。

标题



- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- 补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]。

图注



- 是否展示：是否展示图注
- 图注位置：默认位于右侧。

风格



- 文字大小：针对图中所有文字整体的大小控制

图片

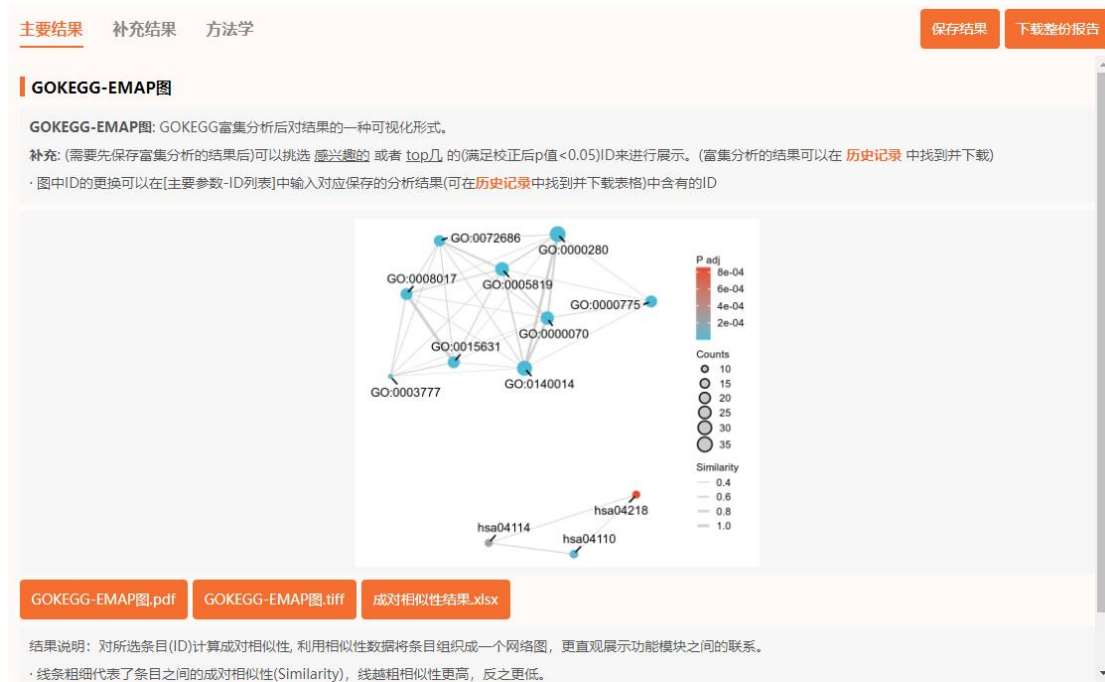
图片	▼
宽度 (cm)	7
高度 (cm)	6
字体	Arial ▼

- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体



结果说明

主要结果



主要结果格式为图片格式, 提供 PDF、TIFF 格式下载, 结果报告可以下载包括 pdf 以及说明文本的内容。

补充结果

可视化ID

当前模块可视化所选ID

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/CDC20/...
BP	GO:0000280	nuclear division	35/197	446/18800	8.47e-21	1.27e-17	1.09e-17	BMP4/CCNB1/CDC20/...
BP	GO:0000070	mitotic sister chromatid se...	24/197	171/18800	2.26e-20	1.83e-17	1.57e-17	CCNB1/CDC20/CENPE...
CC	GO:0005819	spindle	26/203	402/19594	9.72e-14	2.88e-11	2.53e-11	BIRC5/CCNB1/CDK1/...
CC	GO:0072686	mitotic spindle	17/203	160/19594	8.73e-13	1.29e-10	1.14e-10	CDK1/CENPE/KIF11/...
CC	GO:0000775	chromosome, centromeric r...	19/203	227/19594	2.81e-12	2.77e-10	2.44e-10	BIRC5/CCNB1/CENPE...
MF	GO:0008017	microtubule binding	19/192	272/18410	7.14e-11	3.28e-08	3e-08	BIRC5/CENPE/KIF11...
MF	GO:0015631	tubulin binding	19/192	376/18410	1.57e-08	3.6e-06	3.3e-06	BIRC5/CENPE/KIF11...
MF	GO:0003777	microtubule motor activity	8/192	67/18410	4.66e-07	6.74e-05	6.18e-05	CENPE/KIF11/KIFC1...
KEGG	hsa04110	Cell cycle	11/95	126/8164	1.99e-07	4.18e-05	3.98e-05	CCNA2/CCNB1/CDK1/...
KEGG	hsa04114	Oocyte meiosis	10/95	131/8164	2.55e-06	0.0003	0.0003	CCNB1/CDK1/CDC20/...
KEGG	hsa04218	Cellular senescence	10/95	156/8164	1.22e-05	0.0009	0.0008	CACNA1D/CCNA2/CCN...

GOKEGG可视化ID.xlsx

GOKEGG可视化ID.docx

此表格提供当前可视化的 GOKEGG 富集分析结果，提供 Excel、Docx 格式下载。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	GO:0140014	GO:0000280	GO:0000070	GO:0005819	GO:0072686	GO:0000775	GO:0008017	GO:0015631	GO:0003777	hsa04110	hsa04114	hsa04218
2	GO:0140014		0.8857143	0.7741935	0.5405405	0.3714286	0.2820513	0.2820513	0.2820513	0.21875	0.1666667	0.1388889	0.0789474
3	GO:0000280			0.6857143	0.525	0.3684211	0.2857143	0.255814	0.255814	0.1944444	0.1794872	0.1538462	0.097561
4	GO:0000070				0.5151515	0.3225806	0.34375	0.2647059	0.2647059	0.2307692	0.1666667	0.1333333	0.030303
5	GO:0005819					0.6538462	0.2857143	0.4516129	0.4516129	0.3076923	0.15625	0.1612903	0.0588235
6	GO:0072686						0.1612903	0.3846154	0.3846154	0.3157895	0.0769231	0.125	0.0384615
7	GO:0000775							0.1176471	0.1176471	0.08	0.1111111	0.0740741	0.0357143
8	GO:0008017									1	0.4210526	0	0
9	GO:0015631									0.4210526		0	0
10	GO:0003777										0	0	0
11	hsa04110											0.4	0.3125
12	hsa04114												0.3333333
13	hsa04218												

同时提供类目之间的成对相似性结果 Excel 格式下载。

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: ggplot2、igraph、ggraph 包（用于可视化）

处理过程:

- 1) 通过 Jaccard 的相似性指数(JC)计算富集项 (term) 的成对相似性;
- 2) 使用 ggplot2 包、igraph 包、ggraph 包对相似性结果进行 EMAP 图可视化。



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. 可视化结果能否更换别的 ID? 可视化结果的 ID 从哪里获得? 为什么某个条目 (BP、CC、MF、KEGG) 只有 3 个?

答:

在“ID 列表”选项卡中, 有可视化 ID 的输入框:



选项框内默认选择对应云端记录结果中前几个条目, 可以在此处选择想要可视化的 ID。



注意: 输入的 ID 来自所选云端数据记录的结果, 需要先在历史记录中找到对应的记录, 下载 excel 结果, 复制想要展示的 ID 到这个输入框中, 一行代表一个。最多同时支持 30 个, 至少展示 2 个。

	A	B	C	D
1	ONTOLOGY	ID	Description	GeneRatio
2	BP	GO:0140014	mitotic nucle	31/197
3	BP	GO:0000280	nuclear divis	35/197
4	BP	GO:0000070	mitotic siste	24/197
5	BP	GO:0048285	organelle fise	36/197
6	BP	GO:0000819	sister chrom	25/197
7	BP	GO:0007059	chromosome	28/197
8	BP	GO:1902850	microtubule	20/197
9	BP	GO:0098813	nuclear chro	25/197
10	BP	GO:0007052	mitotic spinc	17/197
11	BP	GO:0007051	spindle orga	19/197

ID列表
可视化ID
GO:0072686
GO:0000775
GO:0008017
GO:0015631
GO:0003777
hsa04110
hsa04114
hsa04218

2. 要选择哪些 ID 来进行可视化?

答:

可以对应云端记录表格中的各个分类的 TOP 几条目,也可以是自己感兴趣的想要展示的条目。

3. 为什么出来的图中少了 KEGG (或者 BP 或者 CC 或者 MF), 明明已经选了 GO+KEGG? (为什么出来的图里面某个分类只有 1 个或者没有?)

答:

GOKEGG 可视化模块仅仅只是对已经完成 GOKEGG 富集分析的数据进行可视化,如果对应保存的数据中就不存在某些类(没有富集出来某些类),可视化是不可能会有这些类的。而在 GOKEGG 富集分析模块中,最终的结果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。

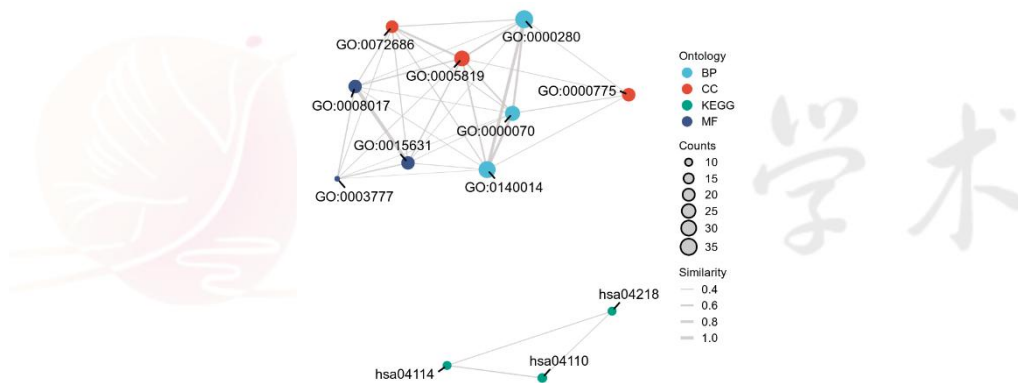
可以先检查 GOKEGG 富集分析的结果,在历史记录中找到保存的记录:



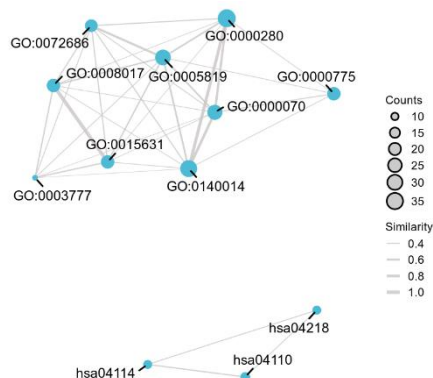
4. 如何修改展示的数据?

答:

可以通过样式中的 **颜色映射(点)** 参数, 下拉框选择 GOKEGG 富集分析结果中不同的结果指标进行展示。



上图: 颜色映射 - 类别(Ontology)



上图: 颜色映射 - 无

5. 能否上传自己的富集数据进行可视化?

答:

由于涉及到类目相关的基因信息，暂时没有对应模块。

