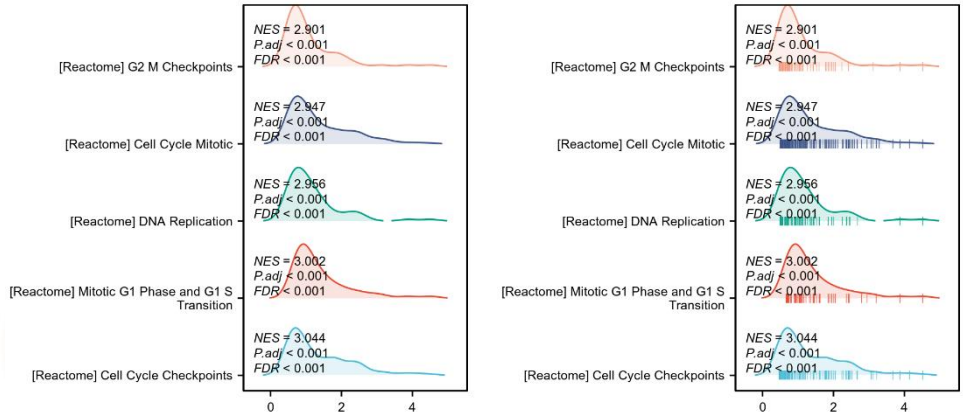
功能聚类 - GSEA 山峦图



网址：https://www.xiantao.love

更新时间：2023.02.08

目录

基本概念

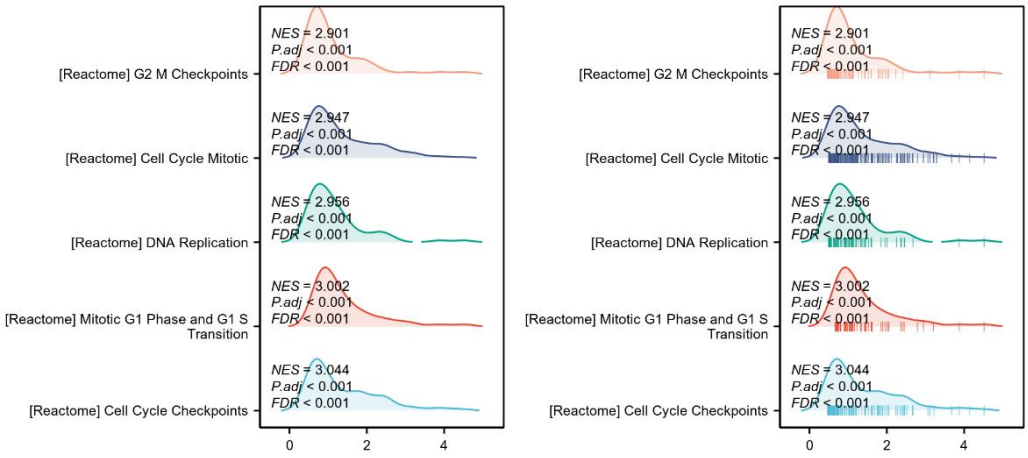
- 基因集富集分析（Gene Set Enrichment Analysis，GSEA）：用一个预先定义的基因集中的基因来评估在与表型相关度排序的基因表中的分布趋势，从而判断其对表型的贡献。这个与表型相关度排序可以是 logFC 值。
- 数据集来自（https://www.gsea-msigdb.org/gsea/msigdb/index.jsp）MSigDB 数据库，如果想要了解数据集的选择以及细节，可以到 MSigDB 数据库进一步了解。

应用场景

想要知道进行了差异分析的两组别有什么功能和通路的差别，并且手上已经有大部分的功能分子以及对应的值，这个值可以是 logFC。可以用这个 logFC 作为分子的排序，从而来评估在预先定义的基因集中是否显著富集。

预先定义的基因集来自 MSigDB 数据库（https://www.gsea-msigdb.org/gsea/msigdb/index.jsp），这些预先定义的基因集中的分子基本为功能基因为主，如果手上只有非功能基因（比如 miRNA、lncRNA、circRNA），那么将由于缺少基因集而无法进行 GSEA 分析。

主要结果



通过山峦图展示各基因集的 GSEA 富集分析结果。

- 纵坐标为基因集名称，横坐标为对应基因集 core_enrichment 中基因对应数值的分布情况（GSEA 分析模块示例数据中的第二列）。
- 图中山峦的颜色（填充色和描边色）为所选择的 颜色映射 内容。（如上图，对应基因集 ID）

一般只要满足阈值（p.adj<0.05 & qvalue<0.25），就关注基因集的名字（最前面是对应的数据库或者分类）即可。可以挑选在满足阈值下的 NES top 的分子，或者一些感兴趣的分子。

云端数据

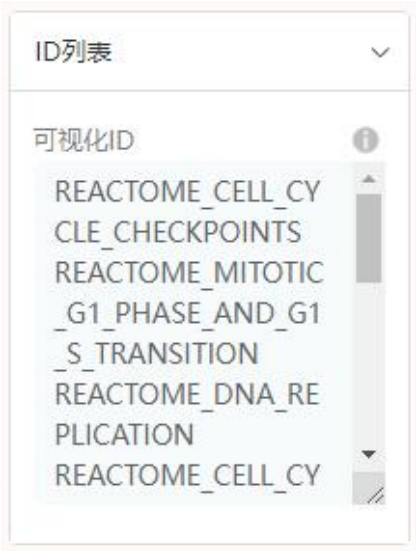
这里的云端数据与历史记录汇总 GSEA 富集分析模块的数据记录是保持一致的，可以在历史记录中找到相应的数据记录。

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：输入想要可视化的基因集 ID，默认为对应云端数据结果中每个类目的前 5 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 10 个基因集。

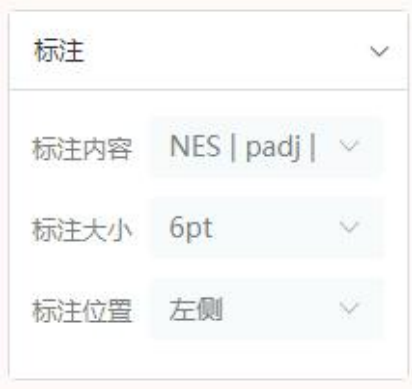
样式

- ID 换行：ID 名称过长时，可以根据需要选择换行模式。可选择 全名(自动换行)、一行 20 长度、一行 30 长度、一行 40 长度、一行 50 长度、一行 60 长度、一行 70 长度、一行 80 长度、不换行。
- ID 前缀是否去除：默认不去除。
- 样式：可选择 山峦图、峦图-数据分布竖线。
- 颜色映射：主要影响山峦的取色范围，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 ID、富集分数（enrichmentScore）、NES、p 值（pvalue）、校正后 p 值（p.adj）、FDR（q 值，qvalue）和不映射。

山峦

- 填充色：山峦的填充色颜色选项，取决于 颜色映射 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 ID）时，有多少个分类会提取多少个颜色，最多支持修改 10 个颜色。受配色方案全局性修改。
- 描边色：山峦的描边色颜色选项，取决于 颜色映射 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 ID）时，有多少个分类会提取多少个颜色，最多支持修改 10 个颜色。受配色方案全局性修改。
- 描边粗细：山峦描边的粗细，默认为 0.75pt。
- 不透明度：山峦的透明度。0 为完全透明，1 为完全不透明。
- 宽度：山峦的间隔宽度。

标注



- 标注内容：在图中标注 GSEA 富集结果中基因集对应的统计量，可选择 NES | padj | FDR、NES | padj、NES | pvalue、NES 和 不标注。
- 标注大小：标注的字体大小，默认 6pt。
- 标注位置：对应上面 标注内容 参数的展示位置，可选择 左侧、右侧。

标题

- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 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(Legend)

- 是否展示：是否展示图注（颜色映射内容为数值型时）
- 图注标题：可以添加图注标题
- 图注位置：可选择 默认、右、上。

风格

- 外框：是否添加外框
- 网格：是否添加网格
- 文字大小：针对图中所有文字整体的大小控制

图片

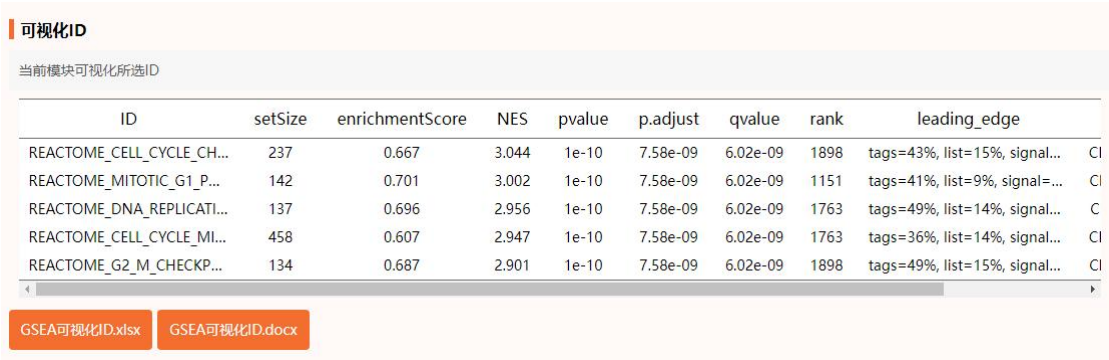
- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

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

ID	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	rank	leading_edge	
REACTOME_CELL_CYCLE_CH...	237	0.667	3.044	1e-10	7.58e-09	6.02e-09	1898	tags=43%, list=15%, signal...	Cl
REACTOME_MITOTIC_G1_P...	142	0.701	3.002	1e-10	7.58e-09	6.02e-09	1151	tags=41%, list=9%, signal=...	Cl
REACTOME_DNA_REPLICATI...	137	0.696	2.956	1e-10	7.58e-09	6.02e-09	1763	tags=49%, list=14%, signal...	C
REACTOME_CELL_CYCLE_MI...	458	0.607	2.947	1e-10	7.58e-09	6.02e-09	1763	tags=36%, list=14%, signal...	Cl
REACTOME_G2_M_CHECKP...	134	0.687	2.901	1e-10	7.58e-09	6.02e-09	1898	tags=49%, list=15%, signal...	Cl

GSEA可视化ID.xlsx GSEA可视化ID.docx

此表格提供当前可视化的 GSEA 富集分析结果，提供 Excel、Docx 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 包（用于可视化）

基因集数据库：MSigDB Collections

（https://www.gsea-msigdb.org/gsea/msigdb/index.jsp）

处理过程：使用 ggplot2 包对富集分析结果进行可视化。

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）。

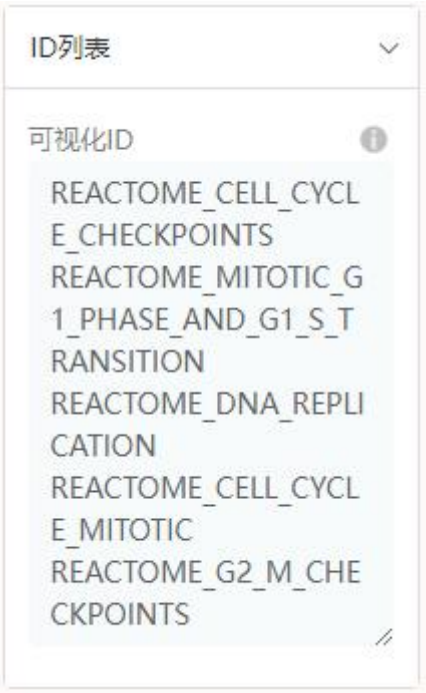
方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

常见问题

1. 可视化结果能否更换别的 ID?

答:

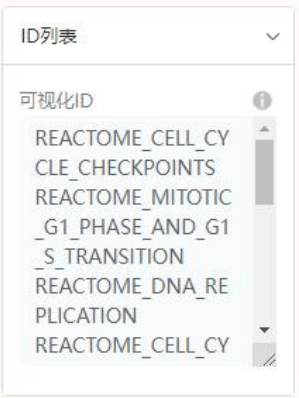
在“ID 列表”选项卡中，有基因集 ID 的输入框:



选项框内**默认**选择对应云端记录结果中前 5 个 ID，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 10 个。

	A	B	C
1	ID	Description	setSize
2	REACTOME_CELL_CYCLE_CHECKPOINTS	REACTOME_C	237
3	REACTOME_MITOTIC_G1_PHASE_AND_G1_S_TRANSITION	REACTOME_M	142
4	REACTOME_DNA_REPLICATION	REACTOME_D	137
5	REACTOME_CELL_CYCLE_MITOTIC	REACTOME_C	458
6	REACTOME_G2_M_CHECKPOINTS	REACTOME_G	134
7	REACTOME_SYNTHESIS_OF_DNA	REACTOME_S	110
8	REACTOME_MITOTIC_METAPHASE_AND_ANAPHASE	REACTOME_M	201
9	WP_RETINOBLASTOMA_GENE_IN_CANCER	WP_RETINOB	84
10	REACTOME_S_PHASE	REACTOME_S	145
11	REACTOME_MITOTIC_SPINDLE_CHECKPOINT	REACTOME_M	92
12	REACTOME_DNA_REPLICATION_PRE_INITIATION	REACTOME_D	111



2. 要选择哪些 ID 来进行可视化？每个 ID 是什么含义？

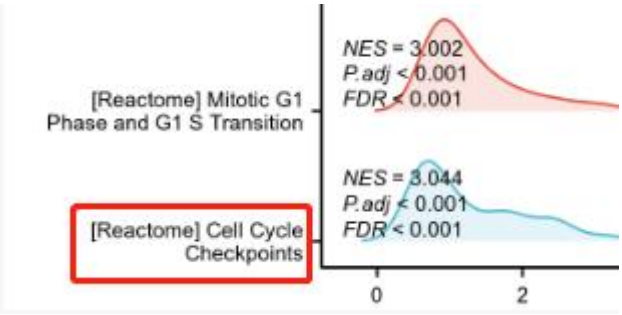
答：

在满足阈值（p.adj<0.05 & qvalue<0.25）下，可以是 TOP 几，也可以是自己感兴趣的想要展示的条目。具体数据集可以通过 MSigDB 数据库（https://www.gsea-msigdb.org/gsea/msigdb/index.jsp）进行了解。

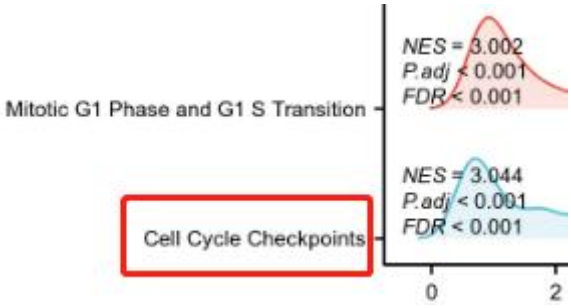
3. 基因集名称太长了，如何修改？

答：

当基因集名称太长时，可以在样式中的 ID 换行 或 ID 前缀是否去除 参数中进行换行和修改。



上图：ID 换行 - 一行 20 长度



上图：ID 前缀是否去除 - 去除