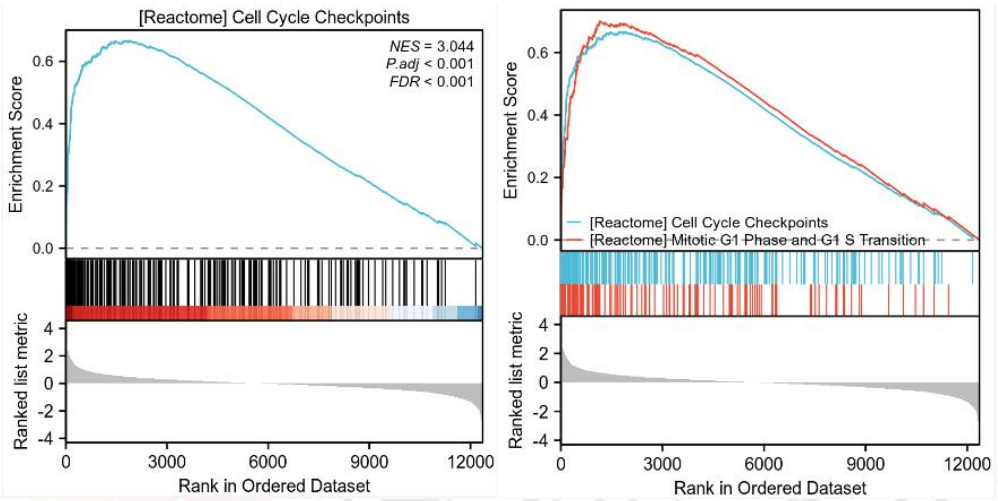
# 功能聚类 – GSEA 经典可视化




网址：https://www.xiantao.love

更新时间：2023.02.04

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。

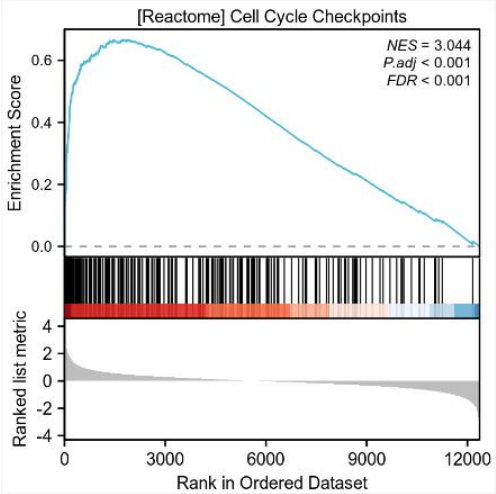
# 应用场景

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

一个基因集

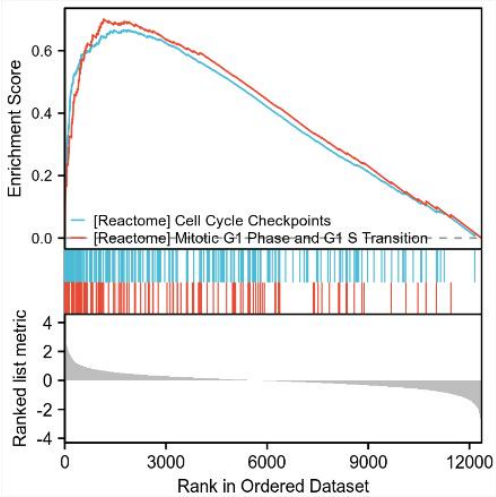



典型结果图由上、中、下三个部分组成：

- 上：为富集评分(ES)的情况，如果 NES 为正（如上图），则峰出现在左侧（头部富集）（高表达组富集），基因集中核心分子主要集中在左侧高表达组中；如果 NES 为负，则尾部会出现谷（尾部富集）（低表达组富集），基因集中核心分子主要集中在右侧低表达组中。
- 中：每一根竖线代表基因集中一个分子，上传数据的分子根据给定的值进行排序，排序后单独提取当前基因集中的定义的分子，分子在整个排序中的位置情况即为中间部分的所示。
- 下：把上传数据分子给定的值进行归一化后的值进行可视化。下部分的结果可以不用怎么关注。

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。

## 多个基因集




展示多个基因集（最多 6 个）：

- 上：选择展示多少个基因集，则绘制多少条曲线。不同颜色代表不同基因集，图注默认出现在左下角。
- 中：选择展示多少个基因集，则展示多少行。不同颜色代表不同基因集。
- 下：把上传数据分子给定的值进行归一化后的值进行可视化。下部分的结果保持不变。

# 云端数据

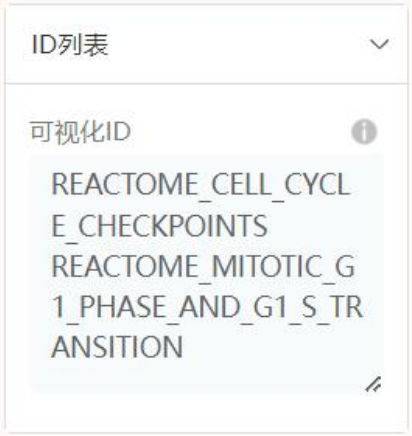

这里的云端数据与历史记录汇总 GSEA 富集分析模块的数据记录是保持一致的，可以在历史记录中找到相应的数据记录。

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 2 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 张图绘制同时绘制 6 个基因集。

# 样式


- 可视化：可以根据需要是否展示 3 个部分的内容，可选择 上中下、上中、上。
- 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。
- 标注内容：只有当输入的可视化 ID 只有 1 个时候才会生效(在图中标注对应的统计量)，可选择 NES | padj | FDR、NES | padj、NES | pvalue、NES。
- 标注位置：对应上面 标注内容 参数的展示位置，可选择 右上、右下、左上、左下、无。

## 线


- 颜色：线条颜色，有多少个基因集取多少个颜色，最多支持 6 个。受配色方案全局性修改。
- 线条类型：可选择 实线、虚线。
- 线条粗细：线的粗细，默认为 0.75pt。
- 不透明度：线条的透明度。0 为完全透明，1 为完全不透明。

## 标题


- 大标题：大标题文本
- x 轴标题：x 轴标题文本

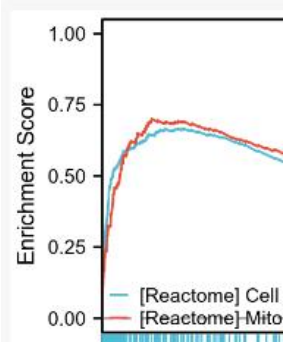
- y 轴标题：y 轴标题文本
- 补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]。

## 图注(Legend)

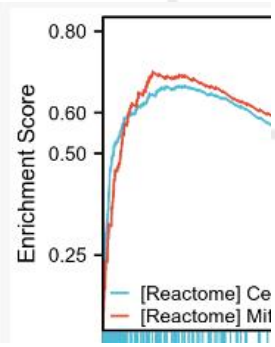

- 是否展示：是否展示图注（可视化 2 个及以上基因集时有效）
- 图注标题：可以添加图注标题
- 图注位置：可选择 默认、右、上、右上、右下、左上、左下。

## 坐标轴

- 主要图对应的 y 轴范围+刻度:(注意:范围的修改如果超过原本值范围的 20%会失效)
  - 如果只是想要修改范围，可以只输入两个范围值，比如 0,1




  - 如果同时想要修改范围+刻度，可以输入比如：0.1,0.25,0.5,0.6,0.8,0.8。注意，此时最大和最小值会被当做范围值，不会作为刻度，如果需要刻度，需要类似于 0.8 那样同时写两次。



## 风格

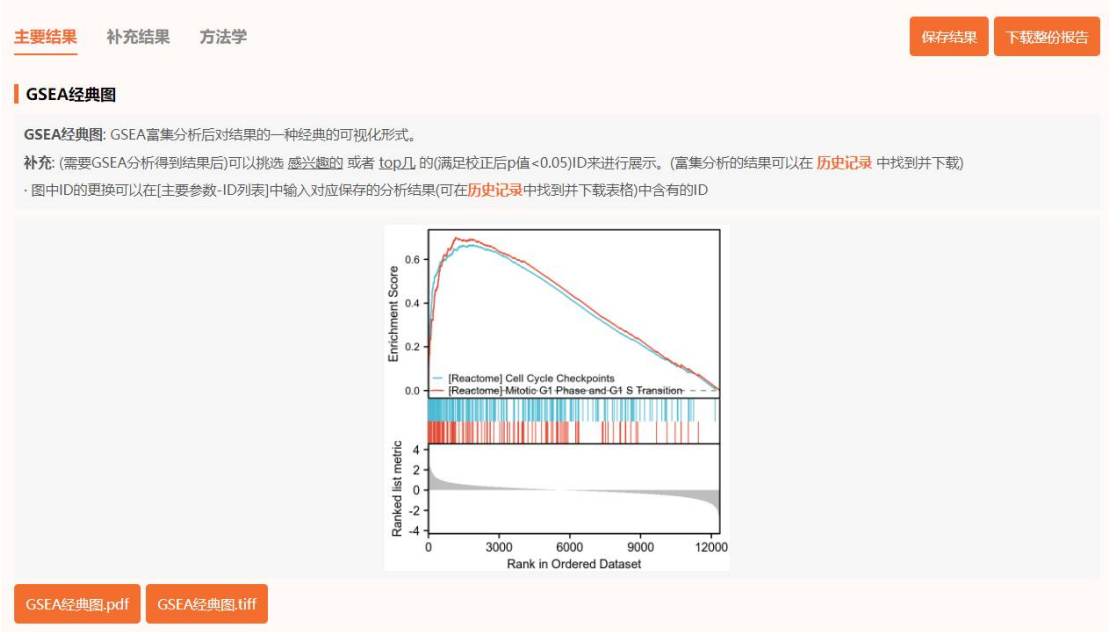

- 外框：是否添加外框
- 网格：是否添加网格
- 文字大小：针对图中所有文字整体的大小控制

## 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

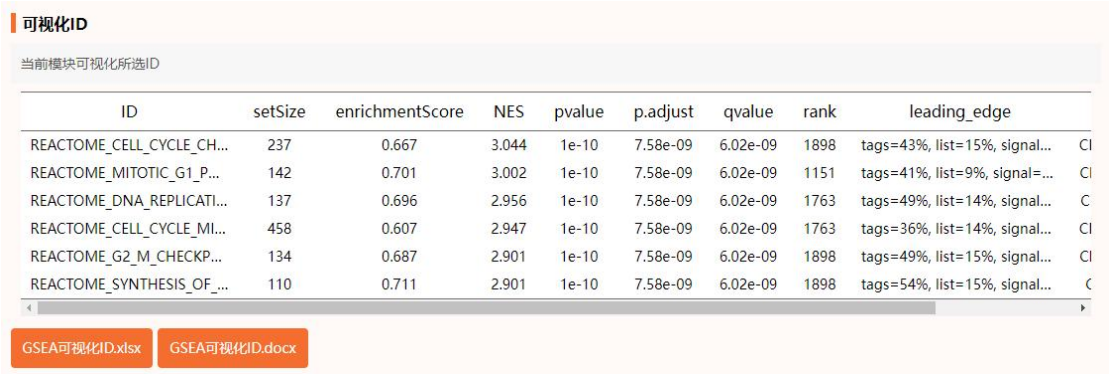
# 结果说明

## 主要结果




主要结果格式为图片格式，提供 PDF 和 TIFF 格式格式下载，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 |
| REACTOME_SYNTHESIS_OF_... | 110 | 0.711 | 2.901 | 1e-10 | 7.58e-09 | 6.02e-09 | 1898 | tags=54%, list=15%, signal... | C |

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

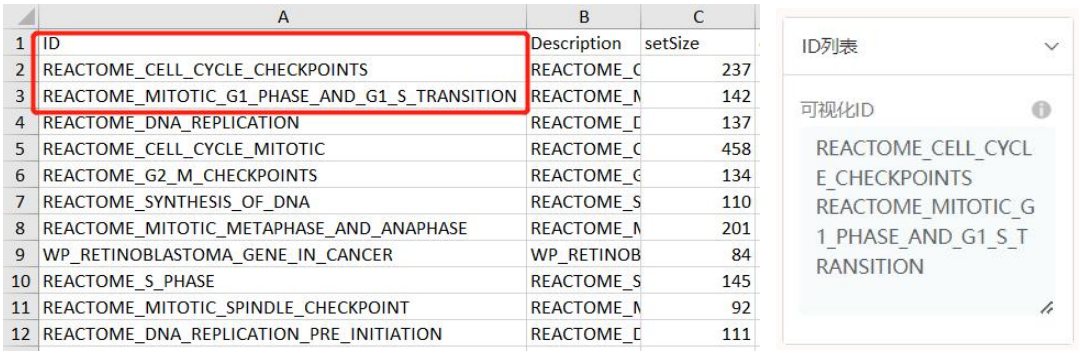
## 1. 可视化结果能否更换别的 ID?

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 2 个 ID（根据 p 值最小排序），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 6 个。



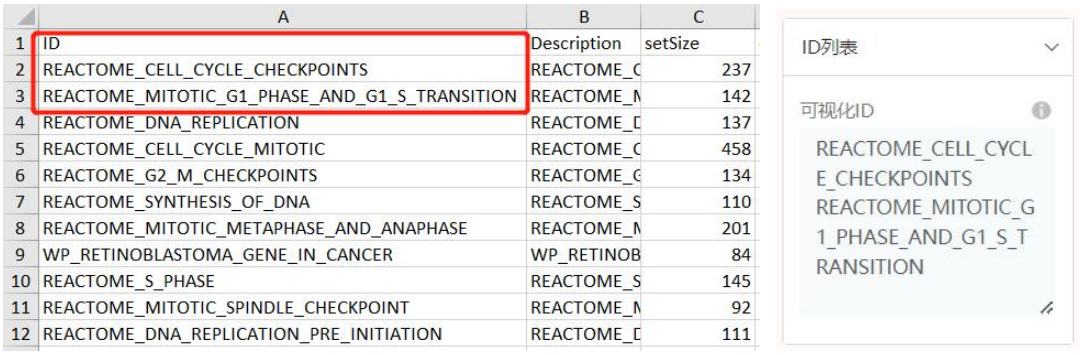
| | A | B | C |
|---|---|---|---|
| 1 | ID | Description | setSize |
| 2 | REACTOME_CELL_CYCLE_CHECKPOINTS | REACTOME_C | 237 |
| 3 | REACTOME_MITOTIC_G1_PHASE_AND_G1_S_TRANSITION | REACTOME_M | 142 |
| 4 | REACTOME_DNA_REPLICATION | REACTOME_D | 137 |
| 5 | REACTOME_CELL_CYCLE_MITOTIC | REACTOME_C | 458 |
| 6 | REACTOME_G2_M_CHECKPOINTS | REACTOME_C | 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. 为什么只能可视化这么一两条数据？如何修改？（为什么别人的 GSEA 那么多的图，而工具只有 1 个？）**

答：

因为工具是可以个性化修改的，所以只可视化一个图（可以是 1-6 个基因集）。首先需要先下载对应云端数据的记录（可以在历史记录中找到对应的），下载好表格后，可以从表格中找到 ID 列，再把想要可视化的 ID 列输入到右侧的基本参数内有基因集 ID 的输入框，即可对输入的 ID 进行可视化。一张图最多可以可视化 6 个 ID。ID 输入框内默认是提取了前 2 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_C | 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 |



**3. 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。

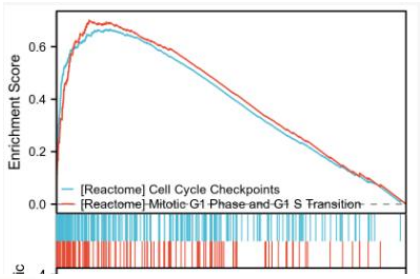
**4. 如何图中标注的文本的位置？固定位置？**

答：

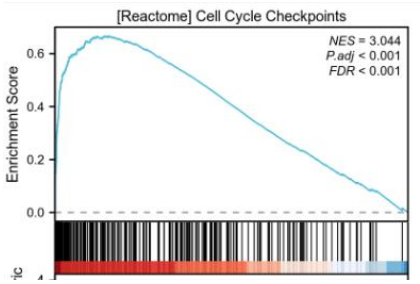
当基本参数的 ID 内只有 1 个时，此时是不会绘制 legend，只会有标注(统计值)的文本。标注文本的位置由 标注位置 参数控制，可以（根据 NES 的正负调整）修改为右上或者左下。当基本参数的 ID 输入 2 个以及以上时，此时是不会标注(统计值)文本，有 legend 代表基因集信息。

**5. 为什么图中没有统计学数值的标注？如何标注？**

答：




当输入了 2 个以及以上的基因集 ID 时，是不会有统计学标注的，只有输入 1 个基因集 ID 时，图中才会有，主要看样式中的 标注内容 和 标注位置 参数。




**6. 标题太长了，如何修改？图注太长了，如何修改？**

答：

当标题太长时，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，同时影响 legend 所代表的基因集信息。