功能聚类 - GSEA 分析

ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue
REACTOME_CELL_CYC...	REACTOME_CELL_CYCLE_CH...	237	0.667	3.044	1e-10	7.58e-09	6.02e-09
REACTOME_MITOTIC_...	REACTOME_MITOTIC_G1_P...	142	0.701	3.002	1e-10	7.58e-09	6.02e-09
REACTOME_DNA_REPL...	REACTOME_DNA_REPLICATI...	137	0.696	2.956	1e-10	7.58e-09	6.02e-09
REACTOME_CELL_CYC...	REACTOME_CELL_CYCLE_MI...	458	0.607	2.947	1e-10	7.58e-09	6.02e-09
REACTOME_G2_M_CHE...	REACTOME_G2_M_CHECKP...	134	0.687	2.901	1e-10	7.58e-09	6.02e-09
REACTOME_SYNTHESI...	REACTOME_SYNTHESIS_OF_...	110	0.711	2.901	1e-10	7.58e-09	6.02e-09
REACTOME_MITOTIC_...	REACTOME_MITOTIC_META...	201	0.650	2.887	1e-10	7.58e-09	6.02e-09
WP_RETINOBLASTOMA...	WP_RETINOBLASTOMA_GE...	84	0.730	2.814	1e-10	7.58e-09	6.02e-09
REACTOME_S_PHASE	REACTOME_S_PHASE	145	0.655	2.799	1e-10	7.58e-09	6.02e-09
REACTOME_MITOTIC_...	REACTOME_MITOTIC_SPIN...	92	0.702	2.772	1e-10	7.58e-09	6.02e-09

网址：https://www.xiantao.love

更新时间：2023.02.03

目录

基本概念 3
应用场景 3
主要结果 4
数据格式 7
参数说明 9
    基因集 9
    分析参数 10
结果说明 11
    主要结果 11
    方法学 13
如何引用 14
常见问题 15

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。

应用场景

手上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

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	Description	setSize	enrichmentS	NES	pvalue	p.adjust	qvalue	rank	leading_edg	core_enrichment	
2	REACTOME_	REACTOME_	237	0.66674999	3.04379824	1E-10	7.5813E-09	6.0164E-09	1898	tags=43%, lis	CDC45/CDCA8/MCM10/CD	
3	REACTOME_	REACTOME_	142	0.70095021	3.00209513	1E-10	7.5813E-09	6.0164E-09	1151	tags=41%, lis	CDC45/MCM10/MYBL2/TOI	
4	REACTOME_	REACTOME_	137	0.69582082	2.95608573	1E-10	7.5813E-09	6.0164E-09	1763	tags=49%, lis	CDC45/MCM10/UBE2C/UBE	
5	REACTOME_	REACTOME_	458	0.6073098	2.94715166	1E-10	7.5813E-09	6.0164E-09	1763	tags=36%, lis	CDC45/CDCA8/MCM10/CD	
6	REACTOME_	REACTOME_	134	0.68661939	2.90126992	1E-10	7.5813E-09	6.0164E-09	1898	tags=49%, lis	CDC45/MCM10/CCNB2/CDI	
7	REACTOME_	REACTOME_	110	0.71134747	2.90065947	1E-10	7.5813E-09	6.0164E-09	1898	tags=54%, lis	CDC45/UBE2C/UBE2S/CCN	
8	REACTOME_	REACTOME_	201	0.65023395	2.88715897	1E-10	7.5813E-09	6.0164E-09	1850	tags=40%, lis	CDCA8/CDC20/CENPE/CCN	
9	WP_RETINOI	WP_RETINOI	84	0.73015214	2.81434613	1E-10	7.5813E-09	6.0164E-09	1329	tags=51%, lis	CDC45/CCNB2/TOP2A/RRM	
10	REACTOME_	REACTOME_	145	0.65531327	2.79876078	1E-10	7.5813E-09	6.0164E-09	1763	tags=46%, lis	CDC45/UBE2C/UBE2S/CCN	
11	REACTOME_	REACTOME_	92	0.70154933	2.77241955	1E-10	7.5813E-09	6.0164E-09	449	tags=27%, lis	CDCA8/CDC20/CENPE/NDC	
12	REACTOME_	REACTOME_	111	0.68070952	2.75987156	1E-10	7.5813E-09	6.0164E-09	1898	tags=46%, lis	CDC45/MCM10/UBE2C/UBE	
13	REACTOME_	REACTOME_	162	0.64291611	2.75628681	1E-10	7.5813E-09	6.0164E-09	1762	tags=38%, lis	CDCA8/CDC20/CENPE/NDC	
14	NABA_CORE	NABA_CORE	201	-0.66811865	-2.75251153	1E-10	7.5813E-09	6.0164E-09	1537	tags=50%, lis	MFAP3/IGFBP7/THSD4/VW	
15	KEGG_CELL_	KEGG_CELL_	114	0.66878969	2.72918483	1E-10	7.5813E-09	6.0164E-09	1230	tags=40%, lis	CDC45/CDC20/CCNB2/CCN	

- ID：基因集的名字，以下划线作为分隔，最前面代表来自哪个数据库，比如 KEGG_xxxxx，就说明来自 KEGG 的基因集。
- Description：基因集的名字
- setSize：基因集中定义的分子数量
- enrichmentScore：富集得分。ES 反应基因集中的基因（S）在排序列表基因（L）的两端富集的程度。
  - 以 FoldChange（FC）值作为排序基因列表（L）对应数值为例，对于一个基因集（S）：

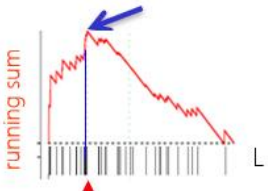
    计算方式是，从基因列表 L 的第一个基因开始扫描，计算一个累计统计值。当遇到一个在 S 里面的基因，则增加统计值；遇到一个不在 S 里面的基因，则降低统计值。每一步统计值增加或减少的幅度与基因的表达变化程度是相关的。统计值连成线，最高峰定义为对应基因集（如例子中的 S）的富集得分 ES。正值 ES 表示基因集在列表的顶部富集，负值 ES 表示基因集在列表的底部富集。

Start with ranked list (L) of genes that are in (Hit) or not in (Miss) a gene set (S), using fold change (FC) as example metric

Ranked List(L)	FC		Contribution to running sum for ES	Hits +|FC| / Σ	Misses -1/(N-N_H)	Running sum for ES
———	15	Hit	+0.15	+0.15		0.15
———	12	Hit	+0.12	+0.12		0.27
———	10	Miss	-0.01		-0.01	0.269
———	9	Hit	+0.09	+0.09		0.359
———	8	Hit	+0.08	+0.08		0.439
———	6	Miss	-0.01		-0.01	0.438
.....	.....	.....	.....	.....	.....	.....

Hits: Genes ∈ S +|FC| / Σ
Misses: Genes ∉ S -1/(N-N_H)

Σ 表达矩阵L中基因的FC之和(e.g., 100)
N 表达矩阵L的基因总数(e.g., 1020)
N_H 某一 基因集对应的基因数(e.g., 20)



ES(S) = value of maximum deviation from 0 of the running sum

- NES（normalize enrichment score）：校正后归一化的富集得分。富集评分的标准化考虑了基因集个数和大小。
- pvalue: 统计检验的 p 值，也称为 NOM p-val。通过基于表型而不改变基因之间关系的排列检验（permutation test）计算观察到的富集得分(ES)出现的可能性。若样品量少，也可基于基因集做排列检验（permutation test），计算 p-value。
- p.adjust: 通过 p 值校正方法得到的校正后的 p 值。(一般这个要满足 <0.05)
- qvalue: 通过 p 值校正方法得到的校正后的 q 值，也称为 FDR。
- rank: 当 ES 值最大时，对应基因在排序好的基因列表 L 中的位置。
- Leading-edge subset，对富集得分贡献最大的基因成员，即核心基因集，也是对 ES 影响较大的基因；该处有 3 个统计值，tags 表示核心基因集占该基因集 S 中基因总数的百分比；list 表示核心基因集占基因列表 L 中基因总数的百分比；signal，将前两项统计数据结合在一起计算出的富集信号强度。
- core_enrichment: 核心富集的分子，即对应的基因集中核心的分子。

这里得到的表格即说明（假设是由两组分析后得到的 logFC 作为分子的值）对应的基因集在两组内有差异，当 ES 或者 NES 为正时，说明该基因集在高表达组（头部）富集，；当 ES 或者 NES 为负时，说明该基因集在低表达组（尾部）富集。

结果这里一般只需要关注满足阈值（p.adj<0.05 & qvalue<0.25）的基因集的名字（最前面是对应的数据库或者分类）。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。

数据格式

	A	B
1	id	value
2	MMP1	4.572612682
3	CDC45	4.514593715
4	NMU	4.418217981
5	CDCA8	4.144075182
6	MCM10	3.876258009
7	CDC20	3.677857006
8	S100A9	3.501962572
9	FOXM1	3.291811805
10	KIF23	3.286223276
11	CENPE	3.219760698
12	KRT16	3.213434697
13	MYBL2	3.208372764
14	MELK	3.197736904
15	CCNB2	3.125238899
16	S100A8	3.086835528

数据要求提供 2 列：

- 第一列除了列名外，下面的可以是分子名、Ensembl 编号、Entrez ID、Symbol ID。
- 第二列为分子对应的数值，这个可以是 logFC 值。
  - 注意：这里的数据不需要对分子过滤，分子越多越好。如果过滤了分子，则可能会富集不到结果。
  - 如果分子存在有重复的话，会只保留第一次出现的分子以及对应的值
- GSEA 一般要求是输入所有的分子和对应的值，因为 GSEA 会对所有分子进行排序，输入分子越少则对结果可能影响较大，甚至可能富集不到结果，参考基因集中的分子中越多分子没有给定值，则该参考基因集富集不出来的概率更高。
- 最少要求 200 行，最多支持 70000 列。若验证数据时返回报错，需要在上传数据内进行相应的调整，然后再上传数据。

这里为任务式模块，提交任务后需要到**历史记录**中刷新并等待任务完成，（分析时间大概在 几分钟 左右，如果任务执行时间过长，刷新后任然在执行阶段，建议删除后重新提交。）

参数说明

（说明：标注了颜色的为常用参数。）

基因集

- 数据集：数据集主要来自 MSigDB 数据库（https://www.gsea-msigdb.org/gsea/msigdb/index.jsp），具体数据集的介绍可以在 MSigDB 数据库查看相关介绍。
- 物种：物种选择，可以选人源、大鼠、小鼠。

分析参数

- 种子：设置种子号。由于 GSEA 会进行重复随机计算，需要设置种子号保证每次输入的结果都是一致的，不同种子号产生的结果都有可能会有一定的差别。
- 计算次数：设置计算的次数，默认无，可以选 1000、5000、100000。提高计算次数能够增加 GSEA 富集结果的稳定性。（可能会有效降低校正后的 p 值。）
- 每个基因集至少含有基因数：一般设置为 10，一般不需要更改。
- 每个基因集至多含有基因数：一般设置为 500，一般不需要更改。
- 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

下载整份报告

GSEA分析

GSEA: (MSigDB数据库)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和相关性。
过程: 分子以及对应的值(例: logFC) → 分子转成Entrez ID → 与所选的基因集合中进行GSEA分析 → 获得分析结果

ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	rank
REACTOME_CELL_CYC...	REACTOME_CELL_CYCLE_CH...	237	0.667	3.044	1e-10	7.58e-09	6.02e-09	1898
REACTOME_MITOTIC_...	REACTOME_MITOTIC_G1_P...	142	0.701	3.002	1e-10	7.58e-09	6.02e-09	1151
REACTOME_DNA_REPL...	REACTOME_DNA_REPLICATI...	137	0.696	2.956	1e-10	7.58e-09	6.02e-09	1763
REACTOME_CELL_CYC...	REACTOME_CELL_CYCLE_MI...	458	0.607	2.947	1e-10	7.58e-09	6.02e-09	1763
REACTOME_G2_M_CHE...	REACTOME_G2_M_CHECKP...	134	0.687	2.901	1e-10	7.58e-09	6.02e-09	1898
REACTOME_SYNTHESI...	REACTOME_SYNTHESIS_OF_...	110	0.711	2.901	1e-10	7.58e-09	6.02e-09	1898
REACTOME_MITOTIC_...	REACTOME_MITOTIC_META...	201	0.650	2.887	1e-10	7.58e-09	6.02e-09	1850
WP_RETINOBLASTOMA...	WP_RETINOBLASTOMA_GE...	84	0.730	2.814	1e-10	7.58e-09	6.02e-09	1329
REACTOME_S_PHASE	REACTOME_S_PHASE	145	0.655	2.799	1e-10	7.58e-09	6.02e-09	1763
REACTOME_MITOTIC_...	REACTOME_MITOTIC_SPIN...	92	0.702	2.772	1e-10	7.58e-09	6.02e-09	449

GSEA.xlsx GSEA.docx

此表格提供 GSEA 富集分析结果（页面只展示 Top10），提供 Excel、Docx 格式下载。

补充结果

GSEA富集情况统计

显著性一般看校正后p值<0.05同时FDR(qvalue)<0.25

条件	个数
FDR(qvalue)<0.25 & p.adjust<0.05	397

此表格提供 GSEA 富集分析结果中显著性满足阈值（p.adj<0.05 & qvalue<0.25）的基因集个数。

这里为任务式模块，提交任务后需要到**历史记录**中刷新并等待任务完成，（分析时间大概在 几分钟 左右，如果任务执行时间过长，刷新后任然在执行阶段，建议删除后重新提交。）任务完成后，提供 Excel 、Docx 格式及完整报告下载。

注意：GSEA 的可视化需要到 可视化模块中进行。如果删除了数据记录，将无法进行可视化。

如果下载的 excel 为空，说明没有富集出来结果。**具体原因可以看 常见问题。**

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)、msigdbr 包(参考基因集来源)

基因集数据库：MSigDB Collections

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

处理过程：

（1） 方法：Gene Set Enrichment Analysis (GSEA)分析

（2） 基因集数据库：MSigDB Collections (数据库超链接)(有各个基因集的详细介绍)

（3） 过程：对输入的数据中的分子进行 ID 转换后，用 clusterProfiler 包进行 GSEA 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。

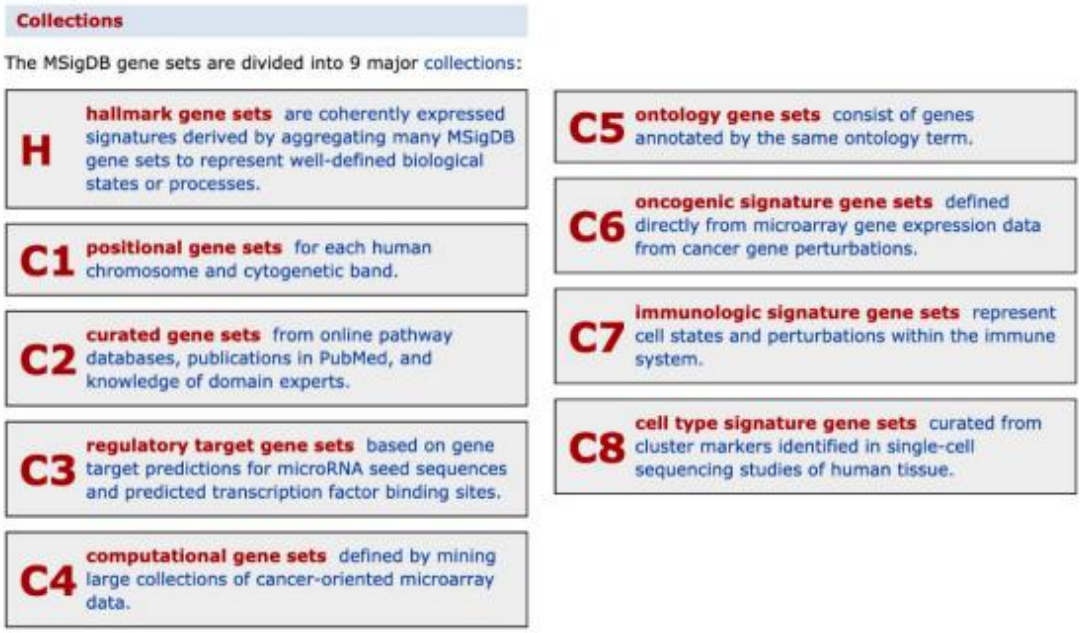
常见问题

1. 数据集选项卡中的每个数据集是什么内容？KEGG 在哪个数据集中？

答：

与 GSEA 分析数据集通用，具体每个数据集都是来源于这个数据库 MSigDB Collections

（https://www.gsea-msigdb.org/gsea/msigdb/index.jsp），数据库内含有每个数据集的介绍：



其中，C2 是通路数据集，包含有 KEGG、Reactome 数据库等的内容。如果是想要分析 KEGG 通路富集，可以选 C2.CP 数据集。如果最终结果中包含有 KEGG 开头的数据集，则说明有 KEGG 通路相关富集。

2. 富集结果不是很好（没有富集出来），有什么可能的原因？下载的文件为空？

答：

下载文件为空，就是说明所选的基因集中都不满足 p.adj < 0.05 & qvalue < 0.25 的阈值，所以就为空文件（只含有列名）。

富集结果不好的原因可能有下：

① 如果定义的每个分子的值本身就是不显著，差别不大，则可能富集出来的结果就是不好的。

② 如果过滤了分子，有一些分子没有出现在表格中，尤其是一些基因集中核心的分子，那么有极大的可能富集不到好的结果。

③ 上传的分子含有较少的功能基因或者无功能基因或者没办法转换 ID 的分子，则可能最终也会富集不到结果。

如果某个基因集结果不好，可以尝试其他的数据集。结果如果还是不理想，那么可能就还是数据的问题或者就是富集不出来，这种工具也没有办法解决。

3. 富集结果的 p.adj 就差一丢丢满足阈值，有没有办法降低以达到阈值?

答：

可以更换别的种子号或者提高计算次数（具体解释见参数说明部分），可能会有不一样的结果。

4. 富集分析的结果很多，如何挑选?

答:

结果这里一般只需要关注满足阈值（p.adj<0.05 & qvalue<0.25）的基因集的名字（最前面是对应的数据库或者分类）。可以挑选在满足阈值下的 NES top 基因集，或者一些感兴趣的基因集。

5. 如何做单基因差异 GSEA?

答:

先进行单基因差异分析（云端数据类型模块在表达差异-单基因差异分析模块），拿到结果后,提取分子列和对应的 logFC,进行 GSEA 分析,即是单基因 GSEA 分析。注意，这里不要按照差异分析结果过滤分子，要把所有的分子都纳入。

6. 如何进行可视化的操作?

答:

提交分析任务完成后，历史记录中会有一条对应的结果记录，可以下载对应的结果表格。同时在 GSEA 可视化模块中可以选择对应的数据记录，可以对数据进行可视化。如果想要更改可视化的基因集，可以从下载的结果表格中复制 ID 列到右边的基本参数中，即可进行想要的基因集的可视化。