交互网络 - 单基因相关性筛选[云]

id	correlation_pearson	pvalue_pearson	padj_pearson	correlation_spearman	pvalue_spearman
ENSG00000227028.6	-0.4332	5.36e-05	0.0058	-0.42563	8.7e-05
ENSG00000229276.1	-0.43311	5.38e-05	0.0058	-0.37459	0.0006
ENSG00000286457.1	-0.43054	6.02e-05	0.0061	-0.43552	5.74e-05
ENSG00000238140.1	-0.42797	6.74e-05	0.0065	-0.40287	0.0002
ENSG00000203506.5	-0.42433	7.88e-05	0.0071	-0.44131	3.73e-05
ENSG00000228171.1	-0.42382	8.06e-05	0.0072	-0.39175	0.0003
ENSG00000279319.1	-0.42282	8.41e-05	0.0073	-0.40508	0.0002
ENSG00000280378.1	-0.42088	9.14e-05	0.0077	-0.30835	0.0051
ENSG00000266274.2	-0.41944	9.71e-05	0.0080	-0.35842	0.0011
ENSG00000229331.1	-0.41793	0.0001	0.0083	-0.36299	0.0009
ENSG00000089486.17	-0.41735	0.0001	0.0083	-0.38708	0.0004
ENSG00000162552.15	-0.41652	0.0001	0.0085	-0.26341	0.0177

网址：https://www.xiantao.love

更新时间：2023.03.09

目录

基本概念

- 批量相关性：将一个主要变量/分子/基因，分别与其他批量的变量/分子/基因进行两两间相关性分析
- 涉及的统计方法：
  - Pearson 相关：参数相关性检验，衡量两组之间是否存在线性关系
  - Spearman 相关：非参数相关性检验，通过秩次来判断两组是否存在相关性。如果不懂具体的选择条件，可以选择该方法
- 注意：相关不等于因果，也就是两者是可能不存在直接的关系

应用场景

单基因相关性筛选常用来展示主要分子/基因/变量与其他分子/基因/变量之间的相关性

分析过程

云端数据 ⟶ 分析

- 云端数据：提供预清洗好的云端数据，不同平台的云端数据集的分子可能会有不同。 注意：选择了不同的平台，搜索出来的分子可能是不一样的
- 分析：
  - 相关性分析：将云端数据进行相关性分析
    - 主要变量与其它变量之间分别进行两两间相关性分析
      - 主要变量：可以在特殊参数[分子]设置主要变量，如果没有则默认云端数据第 1 个变量/分子/基因（数据第 1 行）作为主要变量

      - 其他变量：将所有的变量/分子/基因包括主要变量作为其它变量（云端数据所有行）
    - 相关性分析表
      - 包含不同方法（Pearson、Spearman）计算的相关性系数值与统计学 p 值等

id	correlation_pearson	pvalue_pearson	padj_pearson	correlation_spearman	pvalue_spearman
ENSG00000227028.6	-0.4332	5.36e-05	0.0058	-0.42563	8.7e-05
ENSG00000229276.1	-0.43311	5.38e-05	0.0058	-0.37459	0.0006
ENSG00000286457.1	-0.43054	6.02e-05	0.0061	-0.43552	5.74e-05
ENSG00000238140.1	-0.42797	6.74e-05	0.0065	-0.40287	0.0002
ENSG00000203506.5	-0.42433	7.88e-05	0.0071	-0.44131	3.73e-05
ENSG00000228171.1	-0.42382	8.06e-05	0.0072	-0.39175	0.0003
ENSG00000279319.1	-0.42282	8.41e-05	0.0073	-0.40508	0.0002
ENSG00000280378.1	-0.42088	9.14e-05	0.0077	-0.30835	0.0051
ENSG00000266274.2	-0.41944	9.71e-05	0.0080	-0.35842	0.0011
ENSG00000229331.1	-0.41793	0.0001	0.0083	-0.36299	0.0009
ENSG00000089486.17	-0.41735	0.0001	0.0083	-0.38708	0.0004
ENSG00000162552.15	-0.41652	0.0001	0.0085	-0.26341	0.0177

- 相关性统计：对相关性分析所得的结果进行筛选，得到最终筛选出来的变量
  - 筛选方法：筛选相关性一些常见阈值(|Cor|大于0.3或者0.5或者0.7)下同时满足pvalue<0.05的数量，也可以根据需要下载差异分析结果用excel表进行过滤

相关性统计

此表对应相关性统计方法: pearson

筛选相关性一些常见阈值(|Cor|大于0.3或者0.5或者0.7)下同时满足pvalue<0.05的数量, 也可以根据需要下载差异分析结果用excel表进行过滤

筛选条件	筛选后的数量
|Cor|>0.3 & pvalue<0.05	4057
|Cor|>0.5 & pvalue<0.05	102
|Cor|>0.7 & pvalue<0.05	9

此处是结合p值<0.05，是否结合校正后p值可根据需要下载分析结果后用excel进行筛选

结果解读

id	correlation_pearson	pvalue_pearson	padj_pearson	correlation_spearman	pvalue_spearman
ENSG00000227028.6	-0.4332	5.36e-05	0.0058	-0.42563	8.7e-05
ENSG00000229276.1	-0.43311	5.38e-05	0.0058	-0.37459	0.0006
ENSG00000286457.1	-0.43054	6.02e-05	0.0061	-0.43552	5.74e-05
ENSG00000238140.1	-0.42797	6.74e-05	0.0065	-0.40287	0.0002
ENSG00000203506.5	-0.42433	7.88e-05	0.0071	-0.44131	3.73e-05
ENSG00000228171.1	-0.42382	8.06e-05	0.0072	-0.39175	0.0003
ENSG00000279319.1	-0.42282	8.41e-05	0.0073	-0.40508	0.0002
ENSG00000280378.1	-0.42088	9.14e-05	0.0077	-0.30835	0.0051
ENSG00000266274.2	-0.41944	9.71e-05	0.0080	-0.35842	0.0011
ENSG00000229331.1	-0.41793	0.0001	0.0083	-0.36299	0.0009
ENSG00000089486.17	-0.41735	0.0001	0.0083	-0.38708	0.0004
ENSG00000162552.15	-0.41652	0.0001	0.0085	-0.26341	0.0177

	A	B	C	D	E	F	G	H	I
1	id	correlation_pearsor	pvalue_pearson	padj_pearson	correlation_spearman	pvalue_spearman	padj_spearman	gene_name	gene_type
2	ENSG00000000003.15	0.150461487	0.180000749	0.499340584	0.262240289	0.018272247	0.143395398	TSPAN6	protein_coding
3	ENSG00000000005.6	0.025310666	0.822530636	0.933252071	-0.013311786	0.906107946	0.970258082	TNMD	protein_coding
4	ENSG00000000419.13	0.090679176	0.420768641	0.727041792	0.083242999	0.459238832	0.75706446	DPM1	protein_coding
5	ENSG00000000457.14	0.328377185	0.002763175	0.056411107	0.376513098	0.000579936	0.01557441	SCYL3	protein_coding
6	ENSG00000000460.17	0.3987385	0.00022689	0.013089593	0.398102981	0.000260594	0.009072737	C1orf112	protein_coding
7	ENSG00000000938.13	-0.186601281	0.095316906	0.367772341	-0.168947606	0.131450364	0.431445016	FGR	protein_coding
8	ENSG00000000971.16	-0.256779227	0.020665582	0.169846806	-0.24302168	0.029047919	0.188176566	CFH	protein_coding
9	ENSG00000001036.14	0.168691233	0.132212289	0.433120562	0.097493225	0.385812085	0.703423565	FUCA2	protein_coding
10	ENSG00000001084.13	0.232271638	0.03692593	0.22845234	0.214543812	0.054597191	0.271240967	GCLC	protein_coding
11	ENSG00000001167.14	0.004258503	0.969902171	0.990154626	0.059214092	0.59882248	0.838876485	NFYA	protein_coding
12	ENSG00000001460.18	0.27994543	0.011368041	0.124438902	0.359485095	0.001051862	0.023044272	STPG1	protein_coding
13	ENSG00000001461.17	0.211344743	0.05822597	0.287496065	0.350790425	0.001409054	0.027841351	NIPAL3	protein_coding
14	ENSG00000001497.18	0.314208156	0.004281398	0.071525852	0.283514002	0.010541851	0.102571008	LAS1L	protein_coding
15	ENSG00000001561.7	0.188749095	0.091498848	0.36009231	0.255194219	0.021735574	0.158485446	ENPP4	protein_coding
16	ENSG00000001617.12	0.076223222	0.498830264	0.778674316	0.118947606	0.289582814	0.619681872	SEMA3F	protein_coding

- id：表示变量/分子/基因的编号（1 个变量对应 1 个编号，比如：特殊参数[分子]所选择的分子为 ERBB2[ENSG00000141736.14]，其中 ERBB2 就表示变量名，ENSG00000141736.14 就表示编号，后面将用编号作为变量来说明），对应云端数据第 1 列所有变量/分子/基因

- correlation_pearson：表示主要变量与其他变量通过 Pearson 统计方法计算得到的相关性系数，比如：这里的 0.150461487 表示变量 ENSG00000000003.15 与主要变量 ENSG00000141736.14 做相关性分析得到的相关性系数为 0.150461487

	A	B	C
1	id	correlation_pearson	pvalue_pearson
2	ENSG00000000003.15	0.150461487	0.180000749
3	ENSG00000000005.6	0.025310666	0.822530636
4	ENSG00000000419.13	0.090679176	0.420768641
5	ENSG00000000457.14	0.328377185	0.002763175
6	ENSG00000000460.17	0.3987385	0.00022689

- pvalue_pearson：表示主要变量与其他变量通过 Pearson 统计方法计算得到的统计学 p 值
- padj_pearson：表示主要变量与其他变量通过 Pearson 统计方法计算出来的统计学 p 值再经过 p 值校正方法(BH)校正后得到的 p 值
- correlation_spearman：表示主要变量与其他变量通过 Spearman 统计方法计算得到的相关性系数
- pvalue_spearman：表示主要变量与其他变量通过 Spearman 统计方法计算得到的统计学 p 值
- padj_spearman：表示主要变量与其他变量通过 Spearman 统计方法计算得到的统计学 p 值再经过 p 值校正方法(BH)校正后得到的 p 值

数据格式

提供预清洗好的云端数据，不同平台的云端数据集的分子可能会有不同。 注意：选择了不同的平台，搜索出来的分子可能是不一样的

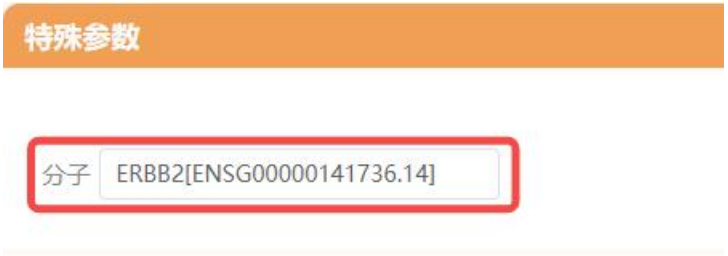
（该样本数据：如下：）

参数说明

（说明：标注了颜色的为常用参数。）

特殊参数

分子



- 分子：选择主要分子（主要变量），也可以手动输入，并与选择的云端数据第 1 列变量/分子/基因进行匹配，匹配成功则作为主要分子进行后续分析，匹配不成功则会默认把选择的云端数据第 1 列第 1 个变量/分子/基因作为主要变量进行后续分析

主要参数

分析参数

- 方法：可以选择主要变量与其他变量间进行相关性分析的方法，两种相关性方法都可以选择，分析结果均会返回两种结果，补充结果统计部分将会当前选择的方法为准
  - Pearson 相关: Pearson 为参数检验方法，数据需要满足双正态
  - Spearman 相关: Spearman(默认)为非参数检验方法，数据可以不需要满足正态性
- p 值矫正方法：可以选择进行 p 值矫正的方法，默认为 BH，还可以选择 bonferroni、BY、fdr、holm、hochberg、hommel

结果说明

主要结果

单基因相关性筛选

单基因相关性筛选: 基于所选数据把 ERBB2[ENSG00000141736.14] 和其他所有分子的数据进行<批量相关性分析>。通过此分析可以找到和目标分子相关性高的分子。

页面中仅仅展示正负相关最高各30个的结果，更多的结果需要下载差异分析表格

id	correlation_pearson	pvalue_pearson	padj_pearson	correlation_spearman	pvalue_spearman
ENSG00000141736.14	1	0	0	1	0
ENSG00000131748.16	0.90364	7.99e-31	2.28e-26	0.75488	0
ENSG00000161395.14	0.89926	4.25e-30	8.09e-26	0.80734	0
ENSG00000141741.12	0.88843	1.93e-28	2.76e-24	0.62313	0
ENSG00000141738.14	0.88506	5.86e-28	6.7e-24	0.73821	0
ENSG00000265178.1	0.87627	9.02e-27	8.6e-23	0.75896	2.23e-16
ENSG00000214546.4	0.80291	1.97e-19	1.61e-15	0.69857	4.15e-13
ENSG00000173991.6	0.76143	1.56e-16	1.12e-12	0.56791	5.84e-08
ENSG00000141744.4	0.70187	2.89e-13	1.84e-09	0.54776	2.01e-07
ENSG00000131771.14	0.69188	8.54e-13	4.88e-09	0.65562	0
ENSG00000108344.15	0.68875	1.19e-12	6.18e-09	0.59955	5.84e-09
ENSG00000167258.15	0.68338	2.07e-12	9.89e-09	0.62649	0

- id：变量/分子/基因的编号
- correlation_pearson: Pearson 方法计算得到的相关系数
- pvalue_pearson: Pearson 方法计算得到的 p 值
- padj_pearson: Pearson 方法计算得到的 p 值经过 p 值校正方法(BH)校正后得到的 p 值
- correlation_spearman: Spearman 方法得到的相关系数
- pvalue_spearman: Spearman 方法计算得到的 p 值
- padj_spearman: Spearman 方法计算得到的 p 值经过 p 值校正方法(BH)校正后得到的 p 值
- 结果筛选方法：可根据需要
  - 选 top 几十或者几百
  - 看一些感兴趣分析的相关性高低情况

- 设定相关系数阈值(常见阈值有 0.3, 0.5, 0.7 等)

补充结果

相关性统计

此表对应相关性统计方法: pearson
筛选相关性一些常见阈值(|Cor|大于0.3或者0.5或者0.7)下同时满足pvalue<0.05的数量, 也可以根据需要下载差异分析结果用excel表进行过滤

筛选条件	筛选后的数量
|Cor|>0.3 & pvalue<0.05	4057
|Cor|>0.5 & pvalue<0.05	102
|Cor|>0.7 & pvalue<0.05	9

此处是结合p值<0.05，是否结合校正后p值可根据需要下载分析结果后用excel进行筛选

这里提供相关性统计表：通过相关性系数与 p 值两个的阈值进行筛选，最终得到符合筛选条件的变量/分子/基因

- 筛选相关性一些常见阈值(|Cor|大于 0.3 或者 0.5 或者 0.7)下同时满足 pvalue<0.05 的数量，也可以根据需要下载差异分析结果用 excel 表进行过滤

方法学

统计分析和可视化均在R 4.2.1版本中进行

涉及的R包：ggplot2包（用于可视化）

处理过程：

(1) 提取目标ID（主要变量）数据

(2) 将目标ID（主要变量）和其余所有ID（所有变量）两两进行相关性分析

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. 方法里面的 Spearman 和 Pearson 方法，应该选择哪一个？

答： 两种方法均可以选择。Pearson 会要求数据是满足正态性，Spearman 因为是非参 数的方法，可以不需要满足。可以先选择非参数的 Spearman 相关进行尝试。

2. 相关系数多少为好？

答： 这个没有很统一的标准，可以参考以下：

- 相关系数强弱：
  - 绝对值在 0.8 以上：强相关
  - 绝对值在 0.5-0.8：中等程度相关
  - 绝对值在 0.3-0.5：相关程度一般
  - 绝对值在 0.3 以下：弱或者不相关