交互网络 - 批量相关性

id	correlation_pearson	pvalue_pearson	padj_pearson	correlation_spearman	pvalue_spearman
COL17A1	-0.18826	0.0620	0.3283	-0.032233	0.7511
CAPG	-0.1874	0.0633	0.3329	-0.11866	0.2416
SH2D2A	-0.18469	0.0672	0.3431	-0.12887	0.2033
LAMC2	-0.1832	0.0695	0.3476	-0.091058	0.3694
CALCR	-0.1731	0.0866	0.3872	-0.11574	0.2540
SLC4A8	-0.17207	0.0886	0.3872	-0.17341	0.0861
CD38	-0.17122	0.0902	0.3872	-0.08679	0.3924
CALCRL	-0.17057	0.0914	0.3872	-0.11192	0.2700
ARNTL2	-0.16427	0.1042	0.4149	-0.10836	0.2852
LCP2	-0.16219	0.1087	0.4264	-0.09368	0.3558
TMSB10	-0.16005	0.1136	0.4360	-0.21022	0.0369
MCUB	-0.15987	0.1140	0.4360	-0.062684	0.5370

网址：https://www.xiantao.love

更新时间：2023.03.08

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

批量相关性常用来展示主要变量与其他变量之间的相关性

分析过程

上传数据 ⟹ 数据处理(清洗)⟹分析

- 数据格式：（具体数据格式要求可以看后面过程的“数据格式”部分）
  - 数据第 1 列表示变量/分子/基因，分类类型，（或者说每一行表示一个变量/分子/基因）
  - 数据第 2 列直至以后表示不同的样本，数值类型

	A	B	C	D	E	F	G	H	I	J	K
1	id	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
2	TSPAN6	33.5806	11.48	35.6261	21.7267	8.742	11.1755	16.9678	26.8519	25.5501	24.3983
3	TNMD	0	0.0242	0.118	0	0	0.0614	0.1078	0.0286	0.0303	0
4	DPM1	109.0146	62.1787	157.3143	175.5211	108.883	74.5302	102.9359	173.0159	166.7452	122.0475
5	SCYL3	9.1668	6.8375	8.6556	4.8179	7.5524	4.295	7.7129	7.7216	7.967	4.2038
6	C1orf112	7.1977	2.4819	8.8316	4.5954	5.3899	2.4152	7.9592	6.0859	14.7341	7.0515
7	FGR	7.1953	3.1641	12.1902	4.8928	2.4174	9.2104	4.7657	3.7351	4.8047	1.7554
8	CFH	11.5801	13.3693	12.0354	4.869	5.5554	17.9946	25.0097	11.3613	15.1796	2.7128
9	FUCA2	69.5308	38.947	153.3975	90.1033	86.8354	39.0433	81.341	75.1984	79.0756	110.9578
10	GCLC	31.0188	27.7368	28.3933	15.2642	75.7259	39.4648	20.9048	18.0893	10.451	14.5119
11	NFYA	24.7061	22.9516	29.8181	22.6624	65.2757	60.7372	21.146	134.9353	42.803	40.0468
12	STPG1	7.4737	3.8602	3.1725	1.3379	3.1129	1.2946	3.3432	2.9436	1.6094	2.5577
13	NIPAL3	22.1396	11.7316	13.6762	7.3527	25.1217	7.4846	19.86	12.9585	7.7514	11.035
14	LAS1L	13.9111	14.6492	11.1092	9.7298	5.502	6.6342	7.04	8.2308	11.6021	12.5214
15	ENPP4	68.1308	43.1565	60.3924	21.7591	67.9239	41.2312	36.3941	39.7186	33.3614	17.6327
16	SEMA3F	89.4541	52.549	28.7326	14.2327	13.8334	75.6039	34.9556	95.3505	45.989	12.6683

- 数据处理
  - 对数据中除了第 1 列外所有非数值类型的数据进行处理
    - 第 1 列作为变量/分子/基因，不能有空的值，或者是无法识别的字符
    - 除了第 1 列外所有列/样本都需要纯数值类型的数据
    - 不能有非数值，特殊值(特殊符号等)，并且每一个变量/分子(每一行)不能都是一个值
- 分析：
  - 相关性分析：将处理(清洗)后的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：可以在主要参数[主分子]设置主要变量，如果没有则默认上传数据第 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（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
COL17A1	-0.18826	0.0620	0.3283	-0.032233	0.7511
CAPG	-0.1874	0.0633	0.3329	-0.11866	0.2416
SH2D2A	-0.18469	0.0672	0.3431	-0.12887	0.2033
LAMC2	-0.1832	0.0695	0.3476	-0.091058	0.3694
CALCR	-0.1731	0.0866	0.3872	-0.11574	0.2540
SLC4A8	-0.17207	0.0886	0.3872	-0.17341	0.0861
CD38	-0.17122	0.0902	0.3872	-0.08679	0.3924
CALCRL	-0.17057	0.0914	0.3872	-0.11192	0.2700
ARNTL2	-0.16427	0.1042	0.4149	-0.10836	0.2852
LCP2	-0.16219	0.1087	0.4264	-0.09368	0.3558
TMSB10	-0.16005	0.1136	0.4360	-0.21022	0.0369
MCUB	-0.15987	0.1140	0.4360	-0.062684	0.5370

■ 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	40
|Cor|>0.5 & pvalue<0.05	1
|Cor|>0.7 & pvalue<0.05	1

结果解读

id	correlation_pearson	pvalue_pearson	padj_pearson	correlation_spearman	pvalue_spearman
COL17A1	-0.18826	0.0620	0.3283	-0.032233	0.7511
CAPG	-0.1874	0.0633	0.3329	-0.11866	0.2416
SH2D2A	-0.18469	0.0672	0.3431	-0.12887	0.2033
LAMC2	-0.1832	0.0695	0.3476	-0.091058	0.3694
CALCR	-0.1731	0.0866	0.3872	-0.11574	0.2540
SLC4A8	-0.17207	0.0886	0.3872	-0.17341	0.0861
CD38	-0.17122	0.0902	0.3872	-0.08679	0.3924
CALCRL	-0.17057	0.0914	0.3872	-0.11192	0.2700
ARNTL2	-0.16427	0.1042	0.4149	-0.10836	0.2852
LCP2	-0.16219	0.1087	0.4264	-0.09368	0.3558
TMSB10	-0.16005	0.1136	0.4360	-0.21022	0.0369
MCUB	-0.15987	0.1140	0.4360	-0.062684	0.5370

	A	B	C	D	E	F	G
1	id	correlation_pearson	pvalue_pearson	padj_pearson	correlation_spearman	pvalue_spearman	padj_spearman
2	TSPAN6	1	0	0	1	0	0
3	TNMD	0.238890544	0.017250199	0.174244436	0.284651711	0.004295507	0.049173076
4	DPM1	0.271035715	0.006657437	0.101650987	0.284291899	0.00447475	0.049173076
5	SCYL3	0.225082927	0.025095495	0.218221699	0.281756339	0.004856245	0.05131148
6	C1orf112	0.219767855	0.028838518	0.227907043	0.306716141	0.002103551	0.033752132
7	FGR	0.005348113	0.958100335	0.97714448	0.037068646	0.715217705	0.834111457
8	CFH	-0.003999832	0.968657155	0.981415557	0.085949289	0.396984024	0.599673753
9	FUCA2	0.030685457	0.763025729	0.914898956	-0.069251701	0.495162644	0.675528846
10	GCLC	0.211292265	0.035782027	0.251512482	0.132096475	0.192122747	0.394811565
11	NFYA	0.132896898	0.189743306	0.540579219	0.211168831	0.036069792	0.147917855
12	STPG1	0.030135012	0.767154622	0.91491941	0.13598021	0.179290477	0.376660665
13	NIPAL3	0.080489885	0.428372775	0.72360266	0.220217131	0.028504801	0.13356957
14	LAS1L	0.265927415	0.007804222	0.107699817	0.226171923	0.024577355	0.122275396
15	ENPP4	0.002196919	0.982781812	0.988714097	0.064007421	0.52844775	0.694443939
16	SEMA3F	0.219626504	0.028944194	0.227907043	0.16533086	0.101922131	0.266811862

- id：表示变量/分子/基因，对应上传数据第 1 列
- 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 1 表示变量 TSPAN6 与自身做相关性分析得到的相关性系数为 1（这里的 TSPAN6 作为主要变量）

	A	B	C
1	id	correlation_pearson	pvalue_pearson
2	TSPAN6	1	0
3	TNMD	0.238890544	0.017250199
4	DPM1	0.271035715	0.006657437
5	SCYL3	0.225082927	0.025095495
6	C1orf112	0.219767855	0.028838518

- 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

	A	B	C	D	E	F	G	H	I	J	K
1	id	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10
2	TSPAN6	33.5806	11.48	35.6261	21.7267	8.742	11.1755	16.9678	26.8519	25.5501	24.3983
3	TNMD	0	0.0242	0.118	0	0	0.0614	0.1078	0.0286	0.0303	0
4	DPM1	109.0146	62.1787	157.3143	175.5211	108.883	74.5302	102.9359	173.0159	166.7452	122.0475
5	SCYL3	9.1668	6.8375	8.6556	4.8179	7.5524	4.295	7.7129	7.7216	7.967	4.2038
6	C1orf112	7.1977	2.4819	8.8316	4.5954	5.3899	2.4152	7.9592	6.0859	14.7341	7.0515
7	FGR	7.1953	3.1641	12.1902	4.8928	2.4174	9.2104	4.7657	3.7351	4.8047	1.7554
8	CFH	11.5801	13.3693	12.0354	4.869	5.5554	17.9946	25.0097	11.3613	15.1796	2.7128
9	FUCA2	69.5308	38.947	153.3975	90.1033	86.8354	39.0433	81.341	75.1984	79.0756	110.9578
10	GCLC	31.0188	27.7368	28.3933	15.2642	75.7259	39.4648	20.9048	18.0893	10.451	14.5119
11	NFYA	24.7061	22.9516	29.8181	22.6624	65.2757	60.7372	21.146	134.9353	42.803	40.0468
12	STPG1	7.4737	3.8602	3.1725	1.3379	3.1129	1.2946	3.3432	2.9436	1.6094	2.5577
13	NIPAL3	22.1396	11.7316	13.6762	7.3527	25.1217	7.4846	19.86	12.9585	7.7514	11.035
14	LAS1L	13.9111	14.6492	11.1092	9.7298	5.502	6.6342	7.04	8.2308	11.6021	12.5214
15	ENPP4	68.1308	43.1565	60.3924	21.7591	67.9239	41.2312	36.3941	39.7186	33.3614	17.6327
16	SEMA3F	89.4541	52.549	28.7326	14.2327	13.8334	75.6039	34.9556	95.3505	45.989	12.6683

数据要求：

- 数据至少 5 列以上，每列至少 2 个观测（即至少 2 行数据），最多支持 1300 列和 70000 行数据
  - 数据第 1 列表示变量/分子/基因，分类类型，（或者说每一行表示一个变量/分子/基因）
  - 数据第 2 列直至以后表示不同的样本，数值类型
  - 第 1 列作为变量/分子/基因，不能有空的值，或者是无法识别的字符
  - 除了第 1 列外所有列/样本都需要纯数值类型的数据，不能有非数值，特殊值(特殊符号等)，并且每一个变量/分子(每一行)不能都是一个值
- 列名（样本名）不能重复

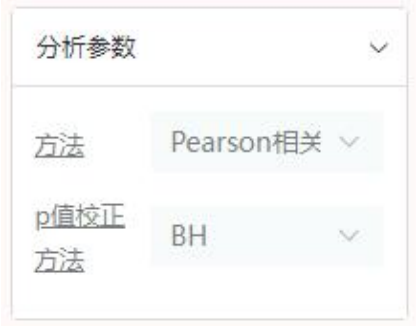
参数说明

（说明：标注了颜色的为常用参数。）

主分子

- 主分子 ID：可以手动输入主要分子（主要变量），与上传数据第 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上传数据第 1 列第 1 个变量/分子/基因作为主要变量进行后续分析

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

批量相关性分析

相关性筛选: 基于所选数据把 TSPAN6 和其他所有分子(ID)的数据进行<批量相关性分析>。
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	padj_spe
PHKA1	0.39488	5.24e-05	0.0075	0.3695	0.0002	0.01
TMEM161A	0.38908	6.89e-05	0.0086	0.26069	0.0093	0.06
TRAPPC6A	0.37264	0.0001	0.0162	0.244	0.0149	0.09
PHF7	0.36681	0.0002	0.0188	0.3048	0.0022	0.03
TRMT11	0.35887	0.0003	0.0241	0.14753	0.1449	0.33
IL20RA	0.35458	0.0003	0.0255	0.44263	5.66e-06	0.00
ZFP64	0.35194	0.0004	0.0255	0.35309	0.0004	0.01
AIFM2	0.35171	0.0004	0.0255	0.3585	0.0003	0.01
USP28	0.34941	0.0004	0.0262	0.34994	0.0004	0.01
FECH	0.34514	0.0005	0.0292	0.39259	6.68e-05	0.00
TNPO3	0.34104	0.0006	0.0307	0.38037	0.0001	0.00
ARFGEF1	0.34097	0.0006	0.0307	0.44891	4.05e-06	0.00

- 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 几十或者几百或者是
  - 看一些感兴趣分析的相关性高低情况 或者是
  - 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	40
|Cor|>0.5 & pvalue<0.05	1
|Cor|>0.7 & pvalue<0.05	1

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