功能聚类 – GSVA 自定义基因集

IDs	s1	s2	s3	s4	s5
gs1	-0.192655151	-0.0860944286	-0.1013756794	-0.260717481	0.145957458
gs2	0.148577606	-0.0204790697	0.0168391070	0.114559924	-0.112953383
gs3	0.078764819	-0.2105647580	-0.1478470438	0.024141368	0.164408619
gs4	-0.216089322	-0.3965959406	-0.0244867979	0.123615377	0.333483397
gs5	-0.048289052	0.0414135075	-0.0694476395	0.010819867	0.065202501
gs6	0.037988818	-0.1819210076	-0.1111575263	0.275762114	-0.058125316
gs7	-0.279400679	-0.3016609984	-0.1763779665	-0.167907396	-0.147600401
gs8	0.223400377	0.1340883058	0.0905533832	0.053868679	-0.076058782
gs9	-0.098771793	0.0551448150	-0.0937835062	0.081060991	-0.087523918
gs10	0.049315000	0.0002629903	-0.0981344049	-0.122054051	-0.105298167
gs11	-0.095474673	-0.1071615776	0.1354543509	-0.081089320	0.037202338
gs12	-0.093517051	0.0069120972	0.1259126580	0.145737077	-0.075640843

网址：https://www.xiantao.love

更新时间：2023.11.09

目录

基本概念

- 基因集变异分析(Gene set variation analysis, GSVA)：是一种非参数、无监督的算法，从单个基因作为特征的表达矩阵，转化为特定基因集作为特征的表达矩阵的过程。

应用场景

与 GSEA 不同，GSVA 不需要预先对样本进行分组、或者对分子/基因排序，该算法默认处理表达量数据（log 化）或者原始 count 计数数据，从而计算每个样本中特定基因集的富集分数。

如果研究对象非功能基因或者是希望自定义特定集合，可以使用自定义基因集进行 GSVA 分析。

主要结果

	A	B	C	D	E	F	G
1	IDs	s1	s2	s3	s4	s5	s6
2	gs1	-0.19265515	-0.08609443	-0.10137568	-0.26071748	0.14595746	-0.00388346
3	gs2	0.14857761	-0.02047907	0.01683911	0.11455992	-0.11295338	-0.07945878
4	gs3	0.07876482	-0.21056476	-0.14784704	0.02414137	0.16440862	0.26612282
5	gs4	-0.21608932	-0.39659594	-0.0244868	0.12361538	0.3334834	0.56707211
6	gs5	-0.04828905	0.04141351	-0.06944764	0.01081987	0.0652025	-0.08705512
7	gs6	0.03798882	-0.18192101	-0.11115753	0.27576211	-0.05812532	-0.02287783
8	gs7	-0.27940068	-0.301661	-0.17637797	-0.1679074	-0.1476004	0.23509886
9	gs8	0.22340038	0.13408831	0.09055338	0.05386868	-0.07605878	-0.08125558
10	gs9	-0.09877179	0.05514481	-0.09378351	0.08106099	-0.08752392	0.1285223
11	gs10	0.049315	0.00026299	-0.0981344	-0.12205405	-0.10529817	-0.25895853
12	gs11	-0.09547467	-0.10716158	0.13545435	-0.08108932	0.03720234	0.0486299
13	gs12	-0.09351705	0.0069121	0.12591266	0.14573708	-0.07564084	-0.10040315

- 表格中的行名（第一列）代表基因集名称，列（从第二列开始）代表样本，数值代表每个样本在对应基因集的富集得分。

结果包含所有满足条件（基因集内定义的分子数目>2 且分子存在于表达量数据矩阵中）的基因集的样本富集得分。可以作为组间比较（limma）的输入数据，从而评估不同基因集在不同组间的富集程度(后续会开发 limma 差异分析模块)。

数据格式

表达量

	A	B	C	D	E	F	G
1	IDs	s1	s2	s3	s4	s5	s6
2	g1	-0.59808	1.636706	-0.40887	0.178157	1.486084	0.614145
3	g2	0.379568	1.304403	-1.29235	-1.17344	0.890876	1.621795
4	g3	1.921686	-1.0319	0.458201	1.387484	-0.4117	-1.34292
5	g4	-0.29105	0.429697	0.678384	-1.47682	1.05833	-1.21538
6	g5	1.862092	1.22481	0.491899	0.711793	0.981387	0.697307
7	g6	-2.29404	-1.06946	-0.52327	0.059466	0.482005	0.951184
8	g7	-0.80498	-0.51614	0.904459	-0.35037	-1.57431	-0.69293
9	g8	-0.42672	0.749551	0.048787	-1.27132	-0.21031	0.887832
10	g9	-0.85795	0.852707	0.357472	-0.35265	0.057828	0.444108
11	g10	0.248116	-1.44779	-1.02924	-1.05662	1.486841	-1.39108
12	g11	-1.39174	-1.11407	-0.5052	0.887208	1.110899	-0.74754
13	g12	1.156562	-0.0013	-0.70316	0.022295	0.760419	-1.01562

数据要求：

- 数据至少有 2 列以上，至少需要 100 行数据。第一行为样本编号，第一列为分子 ID（基因名，需要跟自定义基因集数据的分子能匹配上），不能含有缺失、重复及特殊字符。数值部分为不同分子在各样本中的表达量，功能基因可以是 log(value)或 raw count。支持 RNAseq 数据或者是芯片数据。
- 最多支持 500 列，60000 行。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。

自定义基因集

	A	B	C	D	E	F	G	H
1	gs1	gs2	gs3	gs4	gs5	gs6	gs7	gs8
2	g9397	g3131	g5254	g5403	g7607	g5384	g1444	g4957
3	g7429	g9553	g9596	g8936	g2756	g7801	g9482	g9250
4	g1852	g3890	g1463	g5121	g4768	g1196	g7871	g411
5	g4608	g975	g5210	g4993	g4090	g582	g7776	g1014
6	g4861	g891	g1119	g6095	g3348	g3335	g5459	g2092
7	g4729	g1807	g8958	g7116	g5567	g5449	g9127	g5231
8	g9372	g242	g7726	g7017	g4137	g8951	g6798	g2312
9	g6251	g5173	g1372	g1533	g437	g4143	g641	g8179
10	g5860	g3968	g1925	g1843	g2603	g4242	g5435	g1862
11	g6942	g9499	g7265	g5060	g474	g9215	g5054	g5327
12	g1786	g2030	g7186	g9400	g4519	g5386	g774	g5339
13	g6295	g9078	g9291		g7482	g7257	g8847	g9680
14	g1751	g6007	g6745		g573	g2064	g3282	g8155
15	g7777	g1863	g5393		g815	g6166	g2203	g9989
16	g536	g8570	g2183		g2048	g4397	g3803	g2238
17		g1123	g3378		g4899	g9914		g4379
18		g1426	g9895		g9694	g7514		g1674
19		g82	g480		g6927	g9677		g6767
20		g5255	g8900		g1328	g629		g7893

数据要求：

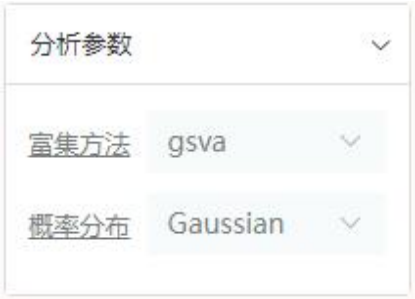
- 数据至少有 1 列以上，每一列至少需要 2 行数据。
- 第一行为自定义基因集名称，每一列为对应基因集的分子 ID 信息。数据应为字符类型，并且要保证基因集中的分子 ID 与表达量数据中的第一列分子 ID 至少 2 个以上匹配，否则无法进行富集分析。
- 最多支持 100 列，10000 行。若验证数据时返回报错，需要在上传数据内进行相应的调整，然后再上传数据。

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。）

分析参数



- 富集方法：估计每个样本基因集富集分数的方法，默认 gsva，可选 ssgsea。只对一个样本进行富集分析时，应选择 ssgsea 方法。当样本>1 且选择 gsva 方法时，数据中不能存在行全为同一个数值，即行的方差不能为 0。
- 概率分布：累积分布函数中非参数估计样本间表达水平的方法，仅对 gsva 方法起作用（微阵列或 CPM、RPKM、TPM 测序数据使用 Gaussian，测序 count 数据使用 Poisson）。

结果说明

主要结果

GSVA-自定义基因集

GSVA分析: 基因集变异分析(Gene set variation analysis, GSVA), 从单个基因作为特征的表达矩阵，转化为特定基因集作为特征的表达矩阵的过程。
· 页面只展示前5的结果，所有结果请下载结果后进行查看。

IDs	s1	s2	s3	s4	s5
gs1	-0.192655151	-0.0860944286	-0.1013756794	-0.260717481	0.145957458
gs2	0.148577606	-0.0204790697	0.0168391070	0.114559924	-0.112953383
gs3	0.078764819	-0.2105647580	-0.1478470438	0.024141368	0.164408619
gs4	-0.216089322	-0.3965959406	-0.0244867979	0.123615377	0.333483397
gs5	-0.048289052	0.0414135075	-0.0694476395	0.010819867	0.065202501
gs6	0.037988818	-0.1819210076	-0.1111575263	0.275762114	-0.058125316
gs7	-0.279400679	-0.3016609984	-0.1763779665	-0.167907396	-0.147600401
gs8	0.223400377	0.1340883058	0.0905533832	0.053868679	-0.076058782
gs9	-0.098771793	0.0551448150	-0.0937835062	0.081060991	-0.087523918
gs10	0.049315000	0.0002629903	-0.0981344049	-0.122054051	-0.105298167
gs11	-0.095474673	-0.1071615776	0.1354543509	-0.081089320	0.037202338
gs12	-0.093517051	0.0069120972	0.1259126580	0.145737077	-0.075640843
gs13	0.002162300	-0.3645511747	0.0575838479	-0.281510814	0.173432586

富集得分表.xlsx

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 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：GSVA 包

处理过程：

首先根据表达谱矩阵基因的累计密度分布对基因进行排序，对每一个基因集进行类似 K-S 检验的秩统计量计算，将表达矩阵转换成基因集富集打分(ES)矩阵，获得每个基因集对应每个样本的 GSVA 富集打分。

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. 可以使用 Count 数据进行 GSVA 分析吗?

答：可以。

GSVA 不需要预先对样本进行分组、或者对分子/基因排序，该算法默认处理表达量数据（log 化）或者原始 count 计数数据，从而计算每个样本中特定基因集的富集分数。

(1) 按照【数据格式】要求，整理原始 count 计数数据即可。第一行为样本编号，第一列为分子 ID（基因名，需要跟自定义基因集数据的分子能匹配上）。如下图所示：

	A	B	C	D	E	F	G	H	I
1	id	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8
2	TSPAN6	3420	1220	2135	1754	1829	2635	3067	1024
3	TNMD	1	0	1	2	0	6	1	3
4	DPM1	2602	900	6094	1392	2556	1452	3223	1106
5	SCYL3	671	479	1142	1025	1920	828	1617	621
6	C1orf112	443	70	868	199	1146	225	374	109
7	FGR	3855	149	904	522	181	318	1252	365
8	CFH	9476	1484	5260	7614	2238	2658	8874	3405
9	FUCA2	3495	1223	8052	2085	5487	1499	3065	1129
10	GCLC	1495	1599	2982	3437	4364	2239	4132	1747
11	NFYA	1494	501	3435	1123	4968	1240	2859	1026
12	STPG1	276	179	830	339	693	373	748	284
13	NIPAL3	1786	1872	5751	3895	5379	3641	3746	2571
14	LAS1L	2451	949	2668	1433	2879	1447	2674	1410
15	ENPP4	1301	1339	1297	2662	2644	4734	3057	4692
16	SEMA3F	5210	948	7200	2857	2585	1942	2772	1615

(2) gsva 方法进行富集分析时，需要选择【概率分布】参数，而针对原始 count 数据，算法建议使用 "Poisson"。