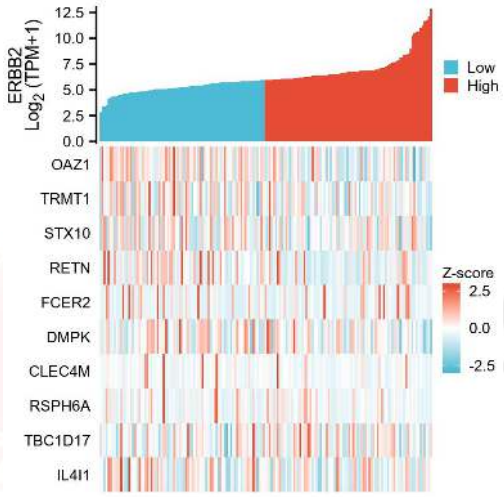
基础绘图 - [云]共表达热图



网址：https://www.xiantao.love

更新时间：2023.02.17

目录

基本概念

- 热图：热图是一个以颜色变化来显示数据情况的矩阵
- 共表达热图：上部分用柱/点/线的形式来展示主要分子的值，下部分用热图的形式来展示其他分子的值，观察其他分子随主要分子的变化趋势
- 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

基于云端数据 分析和可视化 单个基因和其他多个基因之间表达的相关趋势情况

分析过程

云端数据 → 相关性分析 → 可视化

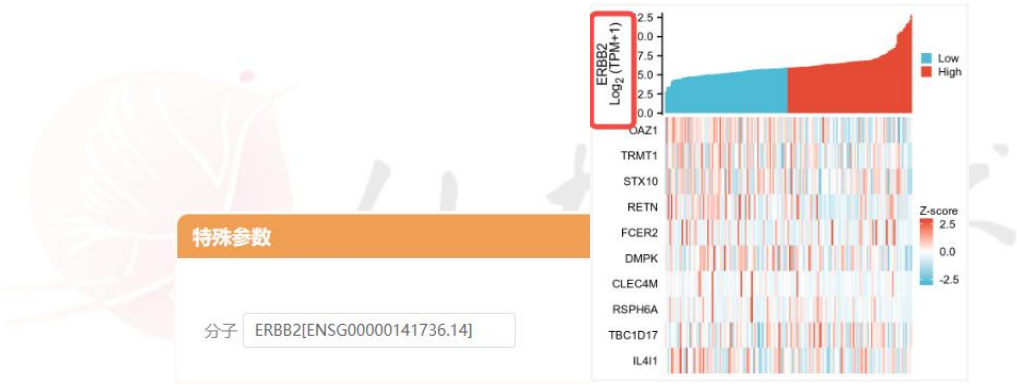
- 云端数据：提供预清洗好的云端数据，不同平台的云端数据集的分子可能会有不同。 注意：选择了不同的平台，搜索出来的分子可能是不一样的
  - 第 1 列为样本/样本编号

- 第 2 列直至以后为每一个分子/变量/基因

	A	B	C	D	E	F	G	H	I	J	K	L
1	sample_id	ERBB2	TSPAN6	TNMD	DPM1	SCYL3	C1orf112	FGR	CFH	FUCA2	GCLC	NFYA
2	TCGA-IC-A6RF-01A-13R-	7.08817977	4.69674502	0.07450544	6.53239645	2.86161797	1.99696715	2.8350954	3.23922999	4.70561304	5.08202979	4.7061603
3	TCGA-IG-A3I8-01A-11R-A	5.3260591	5.88374299	0.20489192	7.96714555	3.24558693	3.58100168	1.75915583	6.39147217	4.77161624	4.48920252	5.629065137
4	TCGA-IG-A3QL-01A-11R-	5.81655619	6.31655126	0.03562391	7.28418651	3.03331727	3.55392567	1.28220255	3.5081624	4.87725345	6.8538385	5.083898379
5	TCGA-IG-A3YA-01A-11R-	5.08085802	5.27261299	0.06598583	6.50076223	2.95311621	2.77865015	4.84587639	4.9947924	4.48481484	5.81758233	4.771753554
6	TCGA-IG-A3YB-01A-11R-	5.45138587	3.4316497	0.03308817	6.31215689	2.7086949	1.27572226	3.1257668	3.28356624	3.07584065	5.39794668	4.520969049
7	TCGA-IG-A3YC-01A-11R-	5.3026371	5.12894497	0	6.6581572	3.1793518	2.24613442	4.55404851	5.59561107	5.16225033	3.80140702	4.043248252
8	TCGA-IG-A4P3-01A-11R-	6.24438667	4.48605143	0.04166353	6.61200741	3.20992119	2.97731656	3.39746073	5.72198231	5.03449746	5.67039566	4.923943062
9	TCGA-IG-A50L-01A-11R-	4.93491274	4.37970413	0.13264264	7.5146754	2.9465063	3.25698044	3.28160924	5.6856795	5.82190151	6.31654402	4.369647943
10	TCGA-IG-A51D-01A-11R-	5.84043554	4.15387828	0	7.56184238	3.1498284	4.19008541	3.83395266	4.80444083	5.49328866	6.13930781	5.132630305
11	TCGA-IG-A5B8-01A-11R-	3.43170318	4.09498763	0.0713508	7.05423312	2.53229202	2.3663362	1.96731619	2.54248094	5.48471174	3.18778358	4.256897432
12	TCGA-IG-A5S3-01A-11R-	5.85764018	5.67664823	0.16104922	7.64181189	4.02986805	4.16513211	5.11039665	4.74797274	4.02811824	3.82280152	5.807231257
13	TCGA-IG-A625-01A-11R-	5.40525041	5.57617745	0.26555692	6.99499939	3.48284828	2.84020136	2.79141819	3.74550627	5.26456646	8.50862672	4.776946192
14	TCGA-IG-A6QS-01A-12R-	4.42344371	3.89572916	0	7.21838209	2.56467082	2.30462785	2.62709317	4.07847321	5.22945718	4.93418622	4.173447286
15	TCGA-IG-A8O2-01A-11R-	5.97101006	6.22004549	0.04124298	7.49762833	3.05595915	2.84641308	2.938681	5.55269983	5.17872659	6.57394857	4.860212958

➢ 相关性分析

- 将云端数据进行相关性分析
  - 主要分子（可以在参数部分选择）与其他分子之间，如下：



    - 此处的分子就表示在数据中的主要分子

  - 相关性分析表

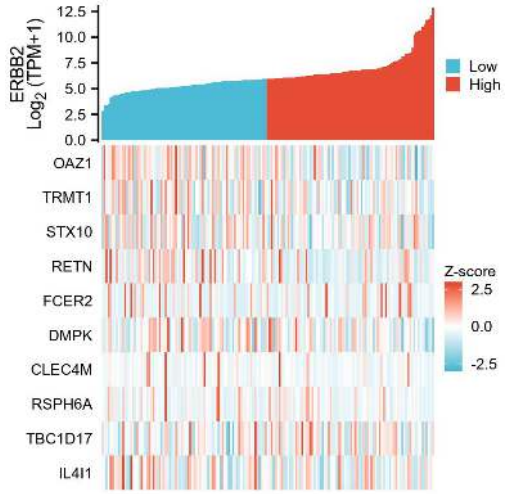
相关性分析

提供Pearson和Spearman统计方法的结果

主变量	次变量	自由度(df)	统计量-Pearson	相关系数-Pearson	p值-Pearson	统计量-Spearman	相关系数-Spearman	p
ERBB2	OAZ1	161	-3.7074	-0.280457	0.0003	9.75e+05	-0.350893	
ERBB2	TRMT1	161	-2.08243	-0.161952	0.0389	8.44e+05	-0.169294	
ERBB2	STX10	161	-0.968776	-0.0761287	0.3341	7.656e+05	-0.0607983	
ERBB2	RETN	161	-2.06727	-0.160803	0.0403	8.612e+05	-0.193218	
ERBB2	FCER2	161	-0.853311	-0.0670987	0.3948	7.194e+05	0.00326035	
ERBB2	DMPK	161	-3.29184	-0.25112	0.0012	8.359e+05	-0.158107	
ERBB2	CLEC4M	161	-1.19814	-0.0940084	0.2326	8.087e+05	-0.120386	
ERBB2	RSPH6A	161	-1.51854	-0.11883	0.1308	7.602e+05	-0.0531982	
ERBB2	TBC1D17	161	-0.381143	-0.0300247	0.7036	6.802e+05	0.0576477	
ERBB2	IL4I1	161	-1.42929	-0.111936	0.1549	8.139e+05	-0.127596	

➢ 将分析后得到的结果（相关性系数与 p 值）进行后续的相关性热图可视化

结果解读



对于上图（柱状图）：

- 横坐标表示各个样本；纵坐标表示主要分子在各样本之间的表达值
- 图形是按照主要分子在各个样本间的表达值从低到高进行排序，再以表达值的中位数为分界线划分为高低两个表达组

对于下图（热图）：

- 列表示各个样本；行表示除了主要分子之外的其他分子
- 每一个方块表示其他分子进行 z-score 转换之后在各样本间的表达值，颜色的深浅表示值的绝对值大小

补充：

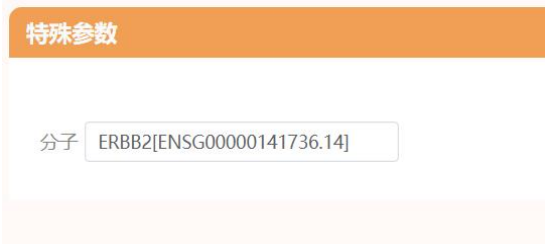
- zscore 转换是绘制热图中常用的一种对数据进行转换的方法（每个分子在单个样本中的表达值减去其在所有样本中的表达均值后，再除以标准差）
- zscore 转换可减少不同分子表达值差异过大而影响整个热图的可视化效果并且保留了单个分子在样本间的差异情况

云端数据

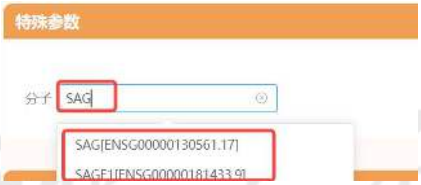
提供预清洗好的云端数据，不同平台的云端数据集的分子可能会有不同。 注意：选择了不同的平台，搜索出来的分子可能是不一样的

参数说明-特殊参数

分子



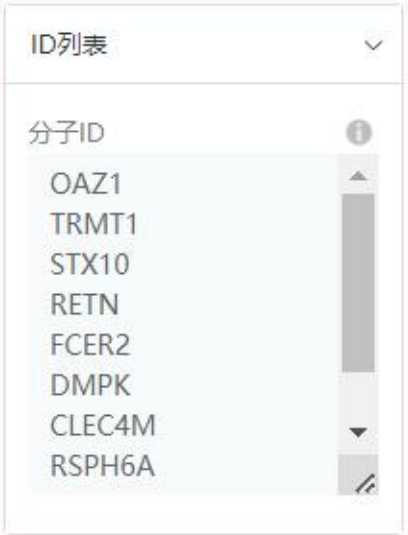
- 分子：根据云端数据所得结果，点击中间输入框后（无须删除），直接键盘输入，即可搜索对应云端数据的分子。 （如果已经是在输入框里面的分子，是没有办法再搜索得到的），如下：



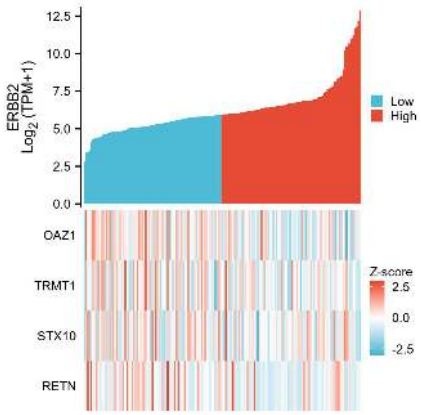
参数说明-主要参数

（说明：标注了颜色的为常用参数。）

ID 列表

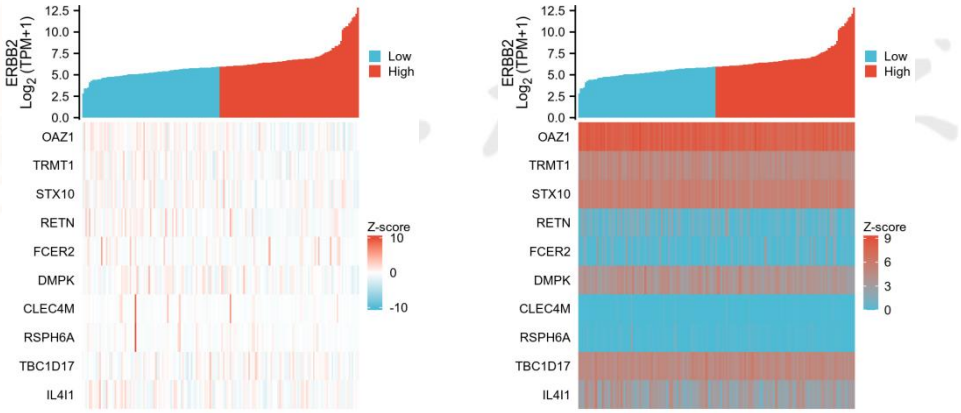


- 分子 ID：这部分输入的分子匹配数据并对应到热图部分的分子列表（一行一个 ID，可以是分子名，也可以是分子 ID，最多支持 20 个）
  - 这部分分子可以来自「单基因差异分析」或者「单基因相关性筛选」两个模块筛选后再进行选择，建议是结合两者一起来看，如果想要热图结果好看一些，建议是从「单基因相关性筛选」模块中挑相关性高的分子进行可视化（因为相关趋势更加明显）



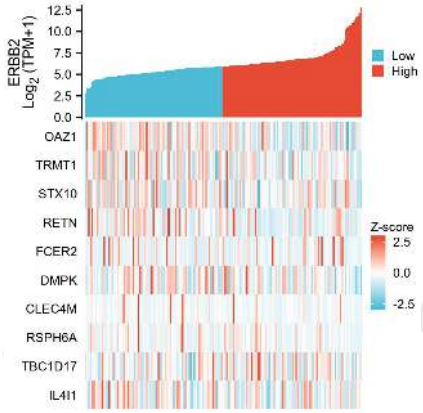
数据处理

- 数据处理：可以选择是否对其他分子进行归一化处理（为了更好展示热图中不同分子之间的差异，一般都需要对热图中的分子进行归一化处理）
  - 默认选择 归一化(处理极值：归一化后大于 3 的值固定为 3,小于-3 固定为 -3)
  - 还可以选择 归一化 或 不归一化，如下：左侧为归一化，右侧为不归一化

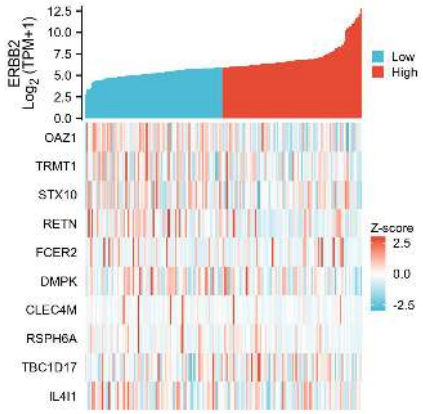


统计

- 统计方法：可以选择对主要分子和其他分子进行相关性分析处理时的方法
  - Spearman：Spearman(默认)为非参数检验方法，数据可以不需要满足正态性

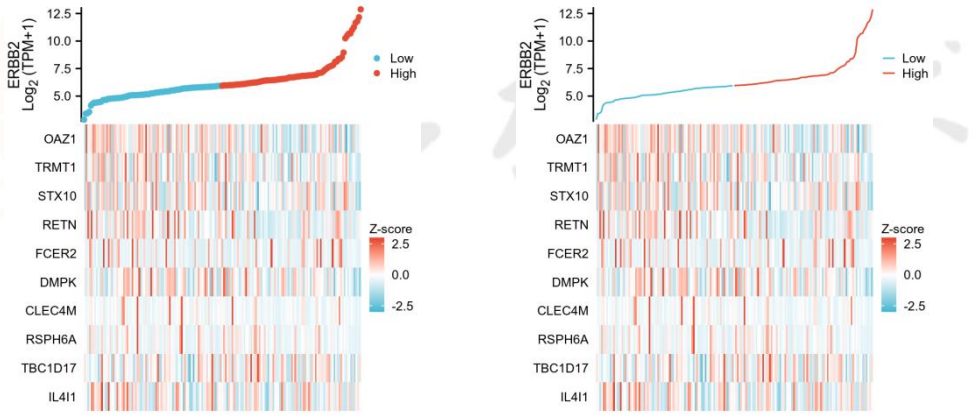


  - Preason：Pearson 为参数检验方法，数据需要满足双正态

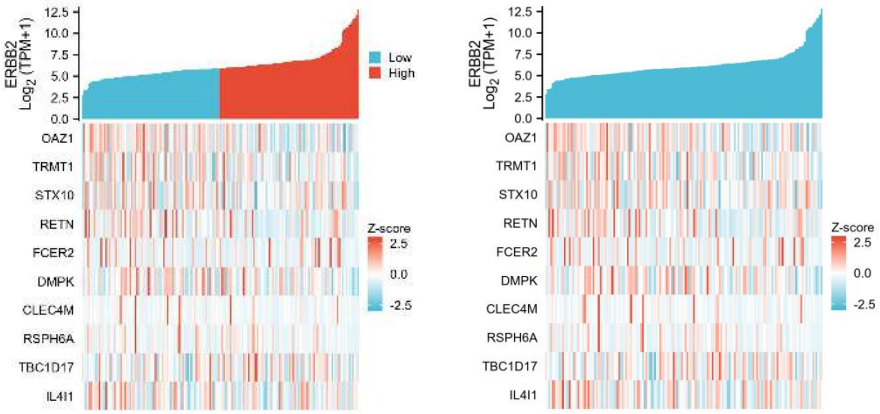


上部分

- 样式：可以选择上部分图形的样式，默认为柱状图，还可以选择点图或者折线图（如下：左侧为点图，右侧为折线图）



- 分组：可以选择是否对主要分子进行分组，默认为以 中位数分组，分为高低两个表达组，还可以选择不分组，如下：（左侧为中位数分组（默认），右侧为不分组）



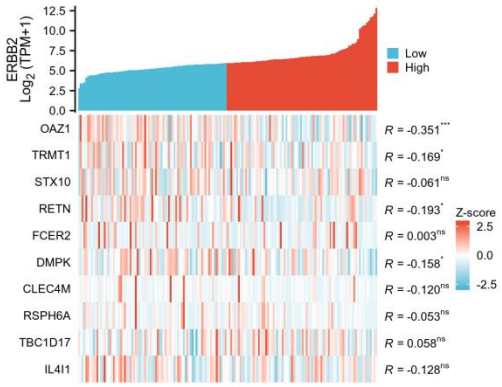
- 颜色：可以修改主要分子样本对应柱子/点/折线的颜色
- 大小：可以修改主要分子对应柱子的宽度/点的大小/折线的粗细
- 不透明度：可以修改主要分子样本对应高低表达组柱子/点/折线的不透明度，1 表示完全不透明，0 表示完全透明

热图

- 填充色：可以修改对应热图的颜色
- 描边色：可以修改对应热图的描边色
- 描边粗细：可以修改对应热图每个小矩形的描边粗细
- 不透明度：可以修改对应热图的不透明度，1 表示完全不透明，0 表示完全透明

标注

- 内容：可以选择是否对热图右侧进行标注操作，默认为不进行标注，还可以选择进行标注，如下：（相关系数-星号标注）
  - 相关系数
  - 相关系数-星号
  - 星号
  - p 值科学计数法
  - p 值数值(小于 0.05 自动<)
  - p 值数值(小于 0.001 自动<)标注，



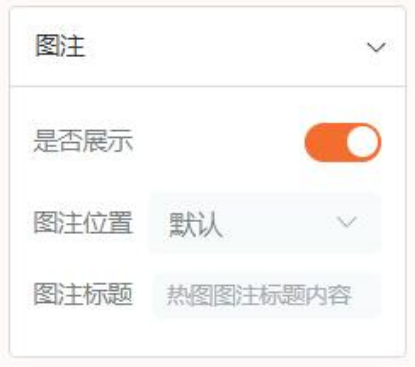
- 标注大小：可以选择并修改标注的大小，默认为 6pt

标题文本

- 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



- 是否展示：可以选择是否展示图注信息
- 图注位置：可以展示图注的位置，默认表示默认为右侧，还可以选择上侧
- 图注标题：可以修改图注标题内容

风格

- 边框：可以选择是否展示边框，默认不展示
- 网格：可以选择是否展示网格，默认不展示
- 文字大小：控制整体文字大小，默认为 7pt

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

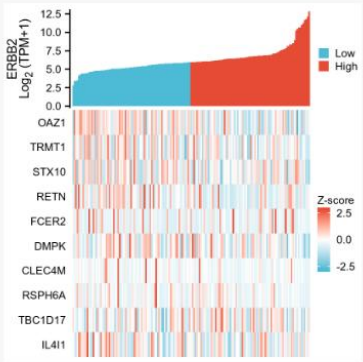
结果说明

主要结果

共表达热图-云

共表达热图: 上部分用柱/点/线的形式来展示主要变量的值, 下部分用热图的形式来展示其他变量的值, 观察其他变量随主要变量的变化趋势

统计方法: spearman



共表达热图.pdf　共表达热图.tiff　分析数据.xlsx

补充: zscore转换

· zscore转换是绘制热图中常用的一种对数据进行转换的方法（每个变量在单个样本中的表达值减去其在所有样本中的表达均值后，再除以标准差）

· zscore转换可减少不同变量表达值差异过大而影响整个热图的可视化效果并且保留了单个变量在样本间的差异情况

补充结果

相关性分析

提供Pearson和Spearman统计方法的结果

主变量	次变量	自由度(df)	统计量-Pearson	相关系数-Pearson	p值-Pearson	统计量-Spearman	相关系数-Spearman	p值-
ERBB2	TSPAN6	161	1.66334	0.129978	0.0982	5.531e+05	0.233664	
ERBB2	TNMD	161	0.565303	0.044508	0.5727	6.661e+05	0.0771218	
ERBB2	DPM1	161	-0.666708	-0.0524715	0.5059	7.871e+05	-0.0905227	
ERBB2	SCYL3	161	4.41664	0.328734	1.83e-05	4.38e+05	0.393084	
ERBB2	C1orf112	161	2.19823	0.170702	0.0294	6.112e+05	0.153233	
ERBB2	FGR	161	-1.55689	-0.121787	0.1215	7.981e+05	-0.105802	
ERBB2	CFH	161	-3.39129	-0.258208	0.0009	9.176e+05	-0.271363	
ERBB2	FUCA2	161	3.42338	0.260486	0.0008	5.564e+05	0.229175	
ERBB2	GCLC	161	-0.852032	-0.0669986	0.3955	7.458e+05	-0.0332962	
ERBB2	NFYA	161	2.06785	0.160848	0.0403	5.074e+05	0.297011	

相关性.xlsx

相关系数为正，说明两个变量之间存在正相关关系; 相关系数为负，说明两个变量之间存在负相关关系;
相关系数绝对值代表相关程度，0-0.3代表弱或者不相关; 0.3-0.5代表弱相关; 0.5-0.8代表中等程度相关; 0.8-1代表强相关
相关是否有统计学意义还需要结合p值来查看

这里提供（pearson，spearman 两种方法）相关性分析表：可以查看主要分子（变量）与其他分子（变量）之间的相关系数与其对应的检验 p 值等

- 相关系数为正数，说明两个分子（主要分子与其他分子）之间可能存在正相关关系；相关系数为负数，说明两个分子可能存在负相关关系
  - 相关系数绝对值在 0.8-1.0 之间，说明两个分子之间强相关
  - 相关系数绝对值在 0.5-0.8 之间，说明两个分子之间中等程度相关
  - 相关系数绝对值在 0.3-0.5 之间，说明两个分子之间相关程度一般
  - 相关系数绝对值在 0.0-0.3 之间，说明两个分子之间弱相关或者不相关
- 相关是否有统计学意义还需要结合 p 值来查看

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) 对数据中主分子（变量）和次要分子（分子）之间进行相关性分析

(2) 分析结果用ggplot包进行共表达热图可视化

数据：

(1) 数据获取：从TCGA数据库（https://portal.gdc.cancer.gov）下载并整理TCGA-ESCA(食管癌)项目STAR流程的RNAseq数据并提取TPM格式的数据 以及 临床数据

(2) 数据过滤策略：去除正常

(3) 数据处理方法：log2(value+1)

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．选择哪些分子进行可视化？分子列表来自哪里？如何才能让热图更加好看？

答：分子可以来自「单基因差异分析」或者「单基因相关性筛选」两个模块筛选后再 进行选择，建议是结合两者一起来看，可以分别从高和低各自挑选 10 个、15 个 或者 20 个来进行可视化

如果想要热图结果好看一些，建议是从「单基因相关性筛选」模块中挑相关性高的分子进行可视化（因为相关趋势更加明显）

2．选择了单基因差异分析后的分子进行可视化，为什么结果不好看（趋势不明显）

答：单基因差异分析是将主要基因按照表达（连续变量）分成了高低表达组（二分类），分析两组的差异基因，因为是从连续变量变成了二分类，所以从单基因差异分析模块中得到的最显著的分子，未必就是相关性最高的分子。如果是要结果好看，建议结合「单基因相关性筛选」模块的结果一起看，挑一些同时相关性也高的分子

3．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。

4. 图的内容被压缩了，如何处理?

答：由于文字不会被压缩，如果热图部分很长，就可能会导致热图部分重叠。解决方案可以是：

① 增加图片高度；

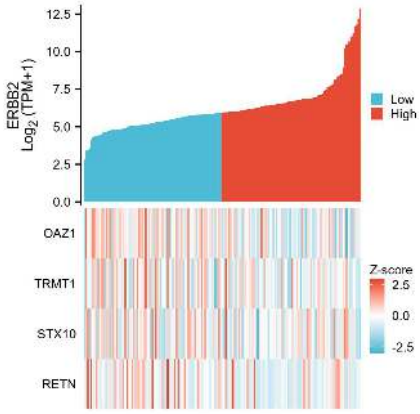
② 减少数据中的列数（分子数量）。

5. 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

6. 为什么在分子输入框内输入了很多的分子，但是出来的图只有几个分子，数目对不上？



答：输入的分子会进行匹配的，只有是正式是分子名才会匹配上，而蛋白或者别名有可能会匹配不上，如果是要精准匹配，建议是输入 ENSG 编号（可以利用 ID 转换工具转换 ID）