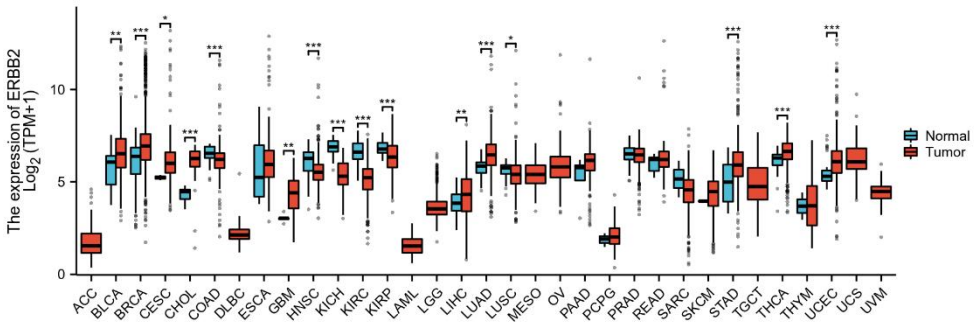
系列模块 - [泛癌] 分组比较



网址：https://www.xiantao.love

更新时间：2023.03.10

目录

基本概念

- 泛癌：泛癌分析旨在研究在不同肿瘤类型中发现的基因组和细胞变化之间的相似性和差异。本模块的 TCGA 的 RNA 表达数据直接来自 TCGA 数据库整理（https://portal.gdc.cancer.gov/），包括 TPM/FPKM/RPM 三种标准化形式；GTEx 的 RNA 表达数据来源 XENA 数据库（https://xenabrowser.net/datapages/?host=https%3A%2F%2Ftoil.xenahubs.net）中经过 Toil 流程统一处理的数据，包括 TPM/FPKM 两种标准化形式。
- 分组比较：针对同一个分子（基因），在泛癌疾病数据中，进行肿瘤组和正常（癌旁）组之间的比较分析。
- 统计方法：
  - T test，亦称 student t 检验（Student's t test），主要用于两组之间的比较，两组需要满足 正态性 和 方差齐性 的要求。
  - Weltch`t test，又称不等方差检验，即当两组仅满足正态而不满足方差齐性的要求时，可以选择用该方法进行两组的比较。
  - Wilcoxon rank sum test,也叫 Mann-Whitney U test（曼-惠特尼 U 检验），或者 Wilcoxon-Mann-Whitney test。秩和检验是一个非参的假设检验方法，一般用于两组不满足正态性的情况。

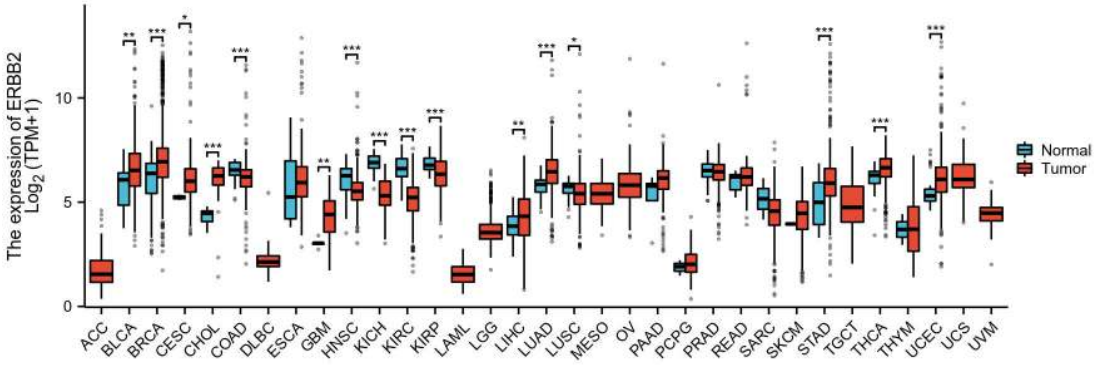
应用场景

基于公共数据（云端数据）直接分析分子在不同泛癌数据中的差异，进行肿瘤和癌旁分组间的比较分析，分析其是否有统计学差异。

一般绘制箱式图进行直观比较，本模块支持点图、箱式图、小提琴图及各自组合的可视化形式。

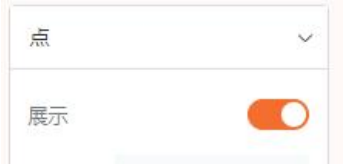
分析流程

主要结果



可视化形式：默认-箱式图

- 横坐标，为公共数据（云端数据）中的肿瘤疾病类型，纵坐标为所选分子（如上图 ERBB2）在不同肿瘤数据中的表达量，默认 log 化处理，即 $log_2(value+1)$。
- 箱子（或其他形式）代表肿瘤和正常分组，如上图将样本分为 Normal、Tumor 两组。
- 默认情况下，模块会根据数据的情况，如正态性和方差齐性自动选择合适的统计方法进行统计分析（具体方法见基本概念中的统计方法）。
- 可视化形式：需要选择对应参数中【展示】



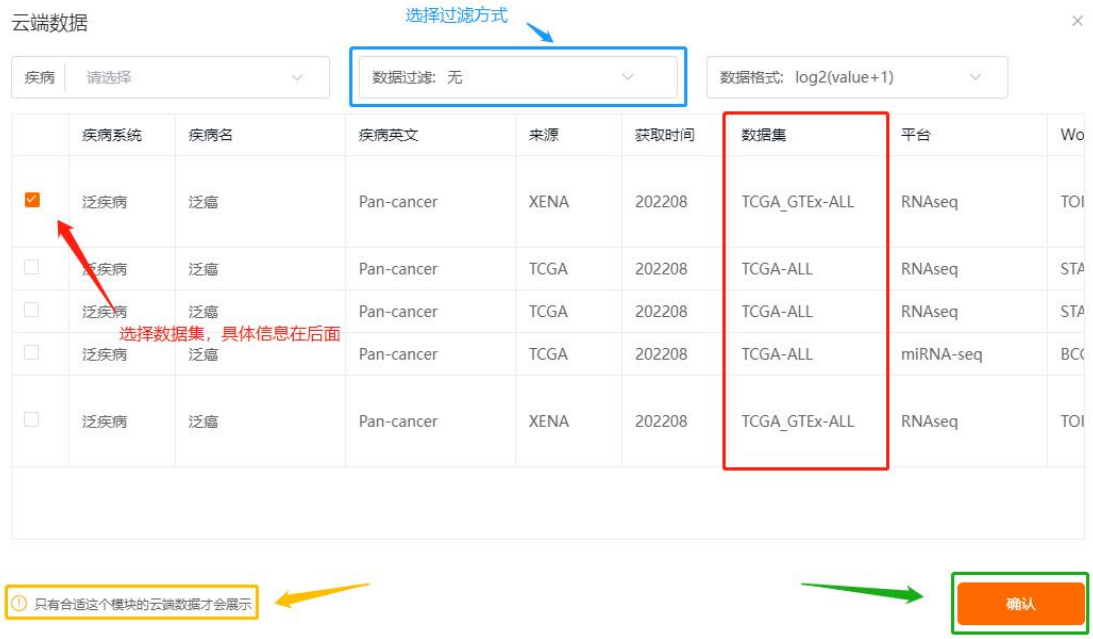
  - 点图：将分组内所有的值用点的位置来进行表示，同时还会另外加上误差线以表征组内的变异情况。点图能够直接看到分组内各样本的分组情况。

- 箱式图/柱状图
  - 箱式图：常见分组比较图之一，箱子中间的横向代表中位数，箱子的上下边代表上四分位（75 百分位数）和下四分位（25 百分位数）。一般而言，箱子的上方和下方的线，如果分组内不存在离群值(Q1-1.5*IQR or Q3+1.5*IQR，下四分位-1.5 倍四分位距)，那么线的最远位置就为最小值或者最大值。箱子的上方或者下方的点代表离群值的点。
  - 柱状图：常见分组比较图之一，柱状图高度一般代表每组的均值情况，同时附带有误差线，表征组内变异的程度。

- 小提琴图：形状类似小提琴，同一水平线上分布的样本越多，则越宽，否则就越窄。小提琴图能有效展示分组内的样本情况的分布。

组合图：点图和箱式图

云端数据



本模块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。

注意查看当前数据参数选中的云端数据。

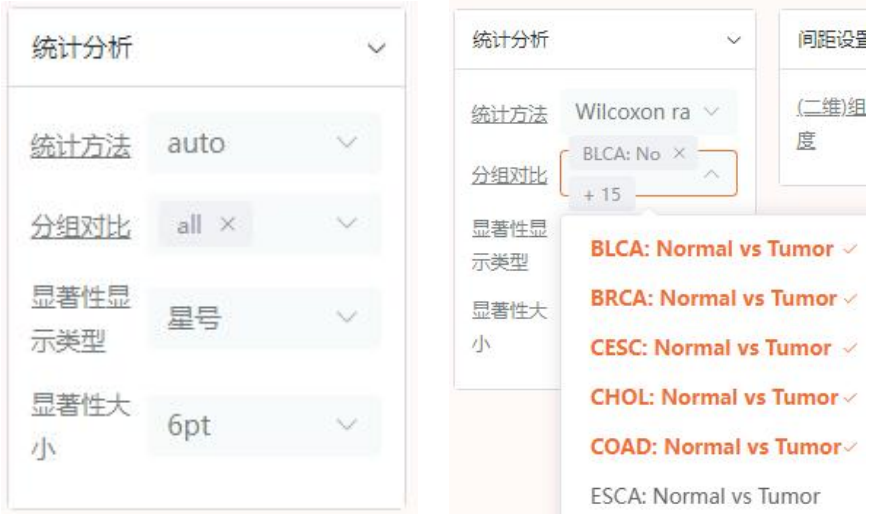
参数说明

（说明：标注了颜色的为常用参数。）

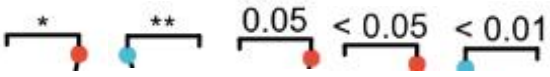
特殊参数

- 分子：下拉框将列出对应所选数据集分子，可以输入关键字搜索分子，基因 symbol 或 Ensembl ID，只能选单个分析。

统计分析

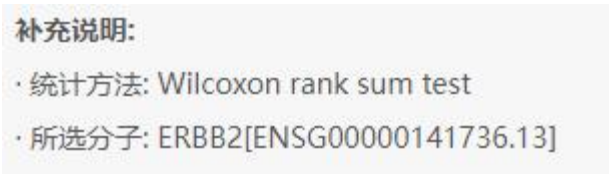


- 统计方法：统计方法默认为 auto（自动选择），当点击确认进行分析后，会自动替换成适合于对应公共数据的统计方法，之后可以自行选择和修改别的统计方法。统计方法的选择依据可以参考“基本概念”中统计方法的说明。
- 分组对比：统计学差异标注的分组，默认为 all（全部都标注）。当点击确认进行分析后，会自动替换成对应数据的分组。之后可以自行选择想要保留和去掉的比较。（如果分组不满足>3 个观测以及标准差>0 的情况，则这些组将不会纳入进行统计分析，如参数中灰色字体的分组，但仍会进行可视化。）。
- 显著性显示类型：影响分组比较中显著性标注，默认为星号。可选择星号或者 p 值以及其他形式，可以选 星号、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<)、p = 科学计数、p = 数值(小于 0.05 自动<)、p = 数值(小于 0.001 自动<)、无。



- 显著性大小：可以修改显著性标注的大小。

参数使用情况：



间距设置

- 组间距离：两组之间的宽度，只有在二维数据(含 legend)的时候才会有效果。主要控制单个分子两组之间的距离。

点

- 展示：可选是否展示。可组合图形。
- 填充色：点的填充色颜色选项，有多少个分组会提取多少个颜色，第一色卡控制低表达分组，第二色控制高表达分组，最多支持修改 2 个颜色。受配色方案全局性修改。
- 描边色：点的描边色颜色选项，有多少个分组会提取多少个颜色，第一色卡控制低表达分组，第二色控制高表达分组，最多支持修改 2 个颜色。受配色方案全局性修改。
- 样式：点的样式类型，可选择 圆形、正方形、菱形、三角形、倒三角。可以多选，多选后不同的分组中点的类型也会有不同。
- 大小：点的大小。
- 透明度：点的透明度。0 为完全透明，1 为完全不透明。
- 分布宽度：图中的点会在一个水平线上随机分布，此处影响点能随机水平移动的范围。

箱/柱

- 展示：可选是否展示。可组合图形。
- 填充色：箱子/柱子的填充色颜色选项，有多少个分组会提取多少个颜色，第一色卡控制低表达分组，第二色控制高表达分组，最多支持修改 2 个颜色。受配色方案全局性修改。
- 描边色：箱子/柱子的描边色颜色选项，有多少个分组会提取多少个颜色，默认黑色，最多支持修改 2 个颜色。不受配色方案全局性影响。
- 描边粗细：箱子/柱子描边的粗细，默认为 0.75pt。
- 不透明度：箱子/柱子的透明度。0 为完全透明，1 为完全不透明
- 宽度：箱子/柱子的宽度控制，默认 0.6。

小提琴

- 展示：可选是否展示。可组合图形。
- 填充色：小提琴的填充色颜色选项，有多少个分组会提取多少个颜色，第一色卡控制低表达分组，第二色控制高表达分组，最多支持修改 2 个颜色。受配色方案全局性修改。
- 描边色：小提琴的的描边色颜色选项，有多少个分组会提取多少个颜色，默认黑色，最多支持修改 2 个颜色。不受配色方案全局性影响。
- 描边粗细：小提琴描边的粗细，默认为 0.75pt。
- 不透明度：小提琴的透明度。0 为完全透明，1 为完全不透明。
- 宽度：小提琴的宽度。
- 宽度校正：用于提高小提琴中较窄位置的宽度和整体宽度。

误差线

误差线只有在没有箱式图时才会显示（箱式图本身自带类似误差线）。

- 展示：可选是否展示。
- 样式：可选 上、上下。
- 类型：可选均值±标准差、均值±标准误、中位数~上下四分位，建议选择均值±标准差。
- 颜色：误差线颜色，默认为纯黑，不受配色方案全局性影响。
- 描边粗细：误差线粗细，默认为 0.75pt
- 宽度：误差线的宽度。

标题

- 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]]

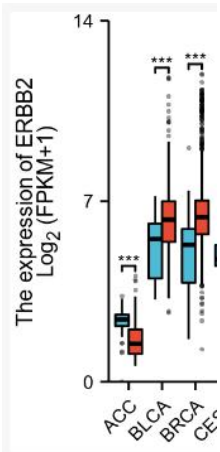
图注(Legend)

- 展示：是否展示图注
- 图注标题：可以添加图注标题

- 图注位置：可选 默认、右、上，默认为右。

坐标轴

- X 轴分组名：支持直接修改 x 轴各个分组的名字，每个名字之间需要用英文输入法的逗号隔开，比如 group1, group2。这里支持换行，需要换行的位置可以插入\n
- X 轴标注旋转：支持对 x 轴文字进行旋转。适合于 x 轴文字过长的时候。
- Y 轴范围+刻度：（注意：范围的修改如果调整过大会失效）
  - 如果同时想要修改范围+刻度，可以输入比如：0,0,7,14,14 。注意，此时最大和最小值会被当做范围值，不会作为刻度，如果需要刻度，需要类似于 0,14 那样同时写两次



风格

- 外框：是否添加外框
- 网格：是否添加网格
- 是否颠倒 XY 轴：可以颠倒 xy 轴
- 文字大小：针对图中所有文字整体的大小控制

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

主要结果格式为图片格式，提供 PDF、TIFF 格式下载，结果报告可以下载包括 pdf 以及说明文本的内容。

- 如果数据可以进行统计分析，将会进行统计分析。统计分析默认是根据数据情况选择合适的统计方法。统计要求每组样本都要满足 3 个样本以上，并且每组样本的方差不能为 0，如果不满足条件，就不会进行统计分析。
- 此外，还提供公共数据中分子在不同样本的表达量，及样本分组和肿瘤类型信息，提供 EXCEL 格式下载：

	A	B	C	D
1	sample_id	status	tissue	ERBB2
2	GTEX-111CU-0126-SM-5GZWZ	Normal	ACC	2.653026021
3	GTEX-111YS-0126-SM-5987T	Normal	ACC	1.989102637
4	GTEX-1122O-0326-SM-5H124	Normal	ACC	2.333413834
5	GTEX-11DXX-0126-SM-5EGH7	Normal	ACC	2.707048824
6	GTEX-11DXY-1626-SM-5H12L	Normal	ACC	3.066935898
7	GTEX-11DXZ-0226-SM-5EGGZ	Normal	ACC	1.903036568
8	GTEX-11EMC-0526-SM-5EGJN	Normal	ACC	2.841969425
9	GTEX-11EQ9-0126-SM-5986I	Normal	ACC	2.220300283
10	GTEX-11GSP-0326-SM-5A5KW	Normal	ACC	1.831872028
11	GTEX-11I78-1826-SM-5A5M4	Normal	ACC	2.339109276
12	GTEX-11NSD-0226-SM-5A5LR	Normal	ACC	2.80323842
13	GTEX-11P7K-0126-SM-5986E	Normal	ACC	2.735558795

补充结果

统计描述

各个组常见「统计描述指标」

组别1	组别2	数目	最小值	最大值	中位数(Median)	四分位距(IQR)	下四分位	上四分位	均值(Mean)	标准差(SD)
ACC	Normal	128	0	3.3162	2.4141	0.44075	2.1651	2.6058	2.3562	0.43353
ACC	Tumor	77	0.61354	4.106	1.4751	0.95153	1.0841	2.0356	1.6803	0.79308
BLCA	Normal	28	3.1969	7.2035	5.5363	2.1192	4.0107	6.1299	5.2241	1.21
BLCA	Tumor	407	2.6759	11.79	6.2888	1.5698	5.4243	6.9942	6.2889	1.378
BRCA	Normal	292	1.6553	9.0625	5.3125	2.0852	3.84	5.9251	4.9069	1.3547
BRCA	Tumor	1099	1.257	11.931	6.3841	1.3073	5.7215	7.0288	6.5703	1.4752
CESC	Normal	13	4.157	5.9453	4.7506	0.80721	4.4944	5.3016	4.9108	0.56379
CESC	Tumor	306	3.0908	12.809	5.6592	0.99541	5.2006	6.196	5.8174	1.2656
CHOL	Normal	9	3.4383	4.7924	4.391	0.71934	3.9486	4.6679	4.2472	0.50571
CHOL	Tumor	36	1.4751	6.5481	5.8813	1.086	5.1543	6.2403	5.5851	1.0955
COAD	Normal	349	0	6.8487	4.7393	1.7483	4.2403	5.9887	4.9801	1.1058
COAD	Tumor	290	4.0251	9.8527	5.8194	0.67681	5.4549	6.1317	5.7906	0.64638

统计描述.xlsx

此表格提供不同肿瘤疾病中肿瘤和正常分组统计描述的结果，提供 EXCEL 格式下载。

异常值分析

离群值 = Q1(下四分位) - 1.5*IQR(四分位间距) 或者 Q3(上四分位) + 1.5*IQR(四分位间距)
异常值 = Q1(下四分位) - 3.0*IQR(四分位间距) 或者 Q3(上四分位) + 3.0*IQR(四分位间距)

组别1	组别2	离群值	异常值
ACC	Normal	1.1634970132744, ...	0
ACC	Tumor	4.106021850982, 3...	
BLCA	Tumor	10.4868047279133,...	11.7321237236, 11...
BRCA	Normal	9.06249812648734	
BRCA	Tumor	10.9317380232795,...	10.9514280165905,...
CESC	Tumor	11.4596118352206,...	11.4596118352206,...
CHOL	Tumor	1.47509288617573,...	1.47509288617573
COAD	Normal	0, 0, 0, 0	
COAD	Tumor	9.66277918628477,...	9.66277918628477,...
DLBC	Tumor	2.67357809656951,...	
ESCA	Normal	0, 0, 0, 0, 8.924...	
ESCA	Tumor	10.8538789581684,...	11.6810389984502,...
GBM	Normal	4.66102343140225,...	

各组离群值和异常值如上所示，如数据确认非人为记录错误，可不进行处理

此表格为不同肿瘤疾病中肿瘤和正常分组异常值情况表，可以判断数据是否存在异常值。

正态性检验

检验方法: Shapiro-Wilk normality test

组别1	组别2	自由度(df)	统计量	p值
ACC	Normal	128	0.9133	4.99e-07
ACC	Tumor	77	0.8878	5.54e-06
BLCA	Normal	28	0.9409	0.1165
BLCA	Tumor	407	0.96223	9.92e-09
BRCA	Normal	292	0.9449	5.4e-09
BRCA	Tumor	1099	0.89782	2.06e-26
CESC	Normal	13	0.925	0.2928
CESC	Tumor	306	0.77796	4.41e-20
CHOL	Normal	9	0.90144	0.2605
CHOL	Tumor	36	0.72757	7.9e-07
COAD	Normal	349	0.90738	8.59e-14
COAD	Tumor	290	0.89563	3.04e-13
DLBC	Normal	444	0.95163	7.09e-11

正态性检验结果显示，存在有不满足正态分布的分组(P < 0.05)，建议选择用 非参数检验的方法

此表格为不同肿瘤疾病中肿瘤和正常分组正态性检验的结果。

方差齐性检验

检验方法: Levene's test

· Base on Mean

组别	自由度1(df1)	自由度2(df2)	统计量	p值
ACC	1	203	27.092	4.73e-07
BLCA	1	433	0.024116	0.8767
BRCA	1	1389	4.0163	0.0453
CESC	1	317	1.3452	0.2470
CHOL	1	43	1.5812	0.2154
COAD	1	637	106.75	3.06e-23
DLBC	1	489	51.358	2.85e-12
ESCA	1	846	9.3266	0.0023
GBM	1	1321	18.274	2.05e-05
HNSC	1	562	0.019155	0.8900
KICH	1	117	3.2882	0.0723
KIRC	1	629	0.41839	0.5180
KIRP	1	347	7.1593	0.0078

方差齐性检验显示，各组观测变量的方差不相等(P < 0.05)，建议选择用校正方法

此表格为不同肿瘤疾病中肿瘤和正常分组方差齐性检验的结果。

Mann-Whitney U检验(Wilcoxon rank sum test)

组别	组别I	组别J	统计量	差值(J-I)	置信区间(95%CI)	p值
ACC	Normal	Tumor	8010.5	-0.84516	-1.0041 - -0.68171	6.74e-14
BLCA	Normal	Tumor	3260.5	0.958	0.47223 - 1.4872	0.0002
BRCA	Normal	Tumor	6.249e+04	1.3498	1.1766 - 1.5366	5.12e-58
CESC	Normal	Tumor	846	0.79731	0.39454 - 1.1788	0.0005
CHOL	Normal	Tumor	27	1.5415	1.0844 - 1.9526	2.32e-05
COAD	Normal	Tumor	2.93e+04	0.84234	0.65979 - 1.0092	4.65e-20
DLBC	Normal	Tumor	7552	0.43453	0.1665 - 0.70048	0.0018
ESCA	Normal	Tumor	4.495e+04	0.50301	0.31234 - 0.70187	8.93e-08
GBM	Normal	Tumor	2.667e+04	1.3561	1.2097 - 1.5013	2.65e-51
HNSC	Normal	Tumor	1.764e+04	-0.75188	-0.94581 - -0.54549	2.29e-09
KICH	Normal	Tumor	2613	-0.71769	-0.97908 - -0.439	3.89e-06
KIRC	Normal	Tumor	4.694e+04	-1.3324	-1.4723 - -1.1751	3.29e-34
KIRP	Normal	Tumor	8874	-0.03065	-0.23068 - 0.17617	0.7748

此表格为不同肿瘤疾病中肿瘤和正常分组 2 组比较统计检验的结果。

（注意：不同的统计方法会有不一样的统计检验的表格）

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：ggplot2 包（用于可视化）、stats、car（用于统计分析）

处理过程：根据数据格式特征情况选择合适的统计方法进行统计(stats 包以及 car 包)(如果不满足统计要求将不会进行统计分析)，用 ggplot2 包对数据进行可视化。

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. 统计学标注可以用具体 p 值吗?

答:

在“统计分析”选项卡中，【显著性显示类型】参数，里面有显示具体 p 值的选项。另外，需要【分组比对】选择了分组才会显示。

2. 云端数据在哪可以查询?

答:

模块分析后，在方法学标签中，提供了公共数据（云端数据）的具体信息及下载链接。