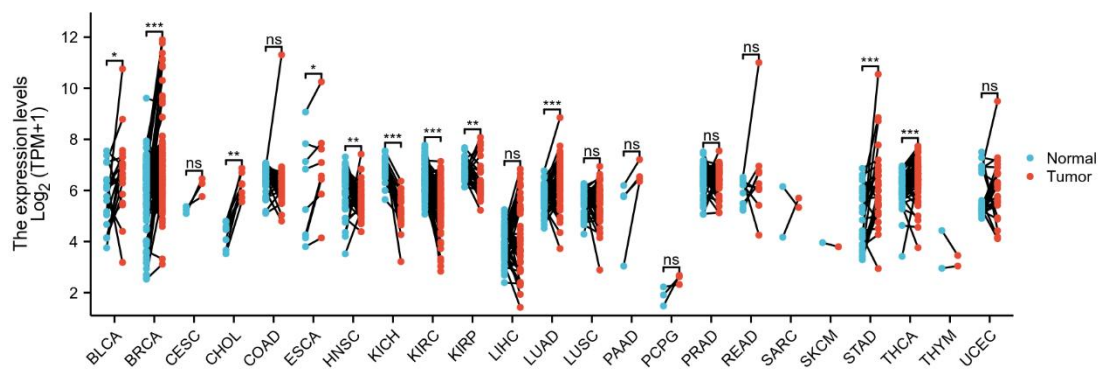


系列模块 - [泛癌] 配对图



网址: <https://www.xiantao.love>



更新时间: 2023.03.13

目录

基本概念	3
应用场景	3
分析流程	4
主要结果	5
云端数据	6
参数说明	7
特殊参数	7
统计分析	7
间距设置	8
点	9
连线	10
箱	10
标题	11
图注(Legend)	12
坐标轴	12
风格	13
图片	14
结果说明	15
主要结果	15
补充结果	17
方法学	19
如何引用	20
常见问题	21

基本概念

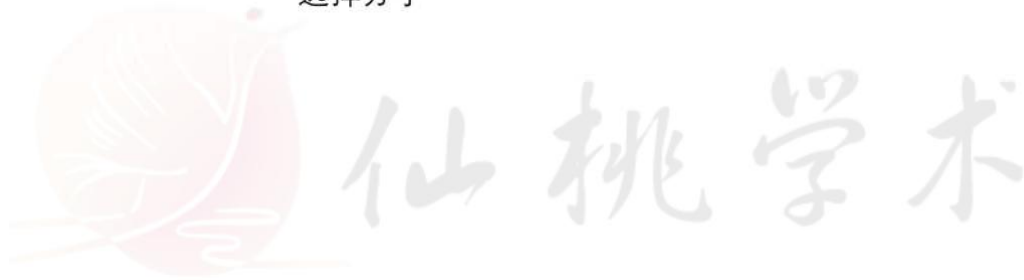
- 泛癌：泛癌分析旨在研究在不同肿瘤类型中发现的基因组和细胞变化之间的相似性和差异。本模块的 TCGA 的 RNA 表达数据直接来自 TCGA 数据库整理 (<https://portal.gdc.cancer.gov/>)，包括 TPM/FPKM/RPM 三种标准化形式。
- 配对图：将有配对关系的样本进行可视化的一种方式。
- 统计方法：统计要求每组样本都要满足 3 个样本以上，并且每组样本的方差不能为 0，如果不满足条件，就不会进行统计分析。
 - **配对样本 T 检验**：用于检验配对类型的参数检验方法。适用条件：连续变量、配对关系、差值服从或者近似服从正态性。
 - **Wilcoxon signed rank test**：符号秩和检验，用于检验配对类型的非参数检验方法，适用于不满足配对样本 T 检验的研究，即不满足正态性！

应用场景

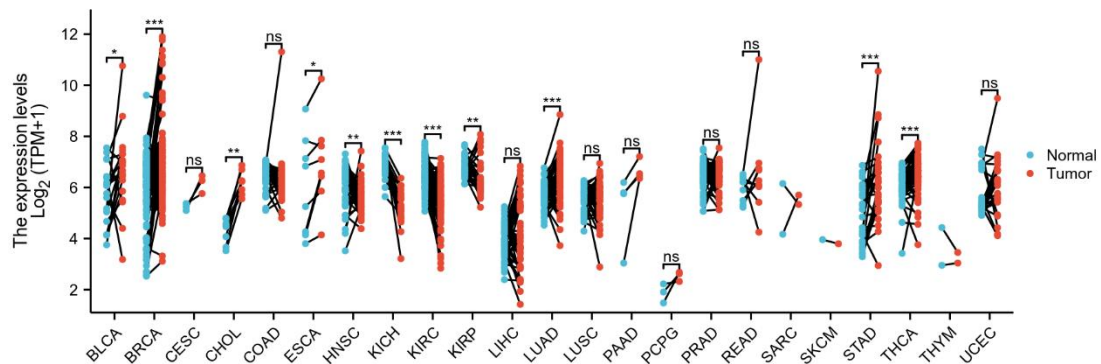
基于公共数据（**云端数据**）直接分析分子在不同泛癌数据中的差异，进行配对的肿瘤和癌旁分组间的差别分析。一般绘制点线图进行直观比较。

注意：配对信息为肿瘤-癌旁/正常，无配对信息的数据将无法进行分析和可视化！

分析流程



主要结果



- 横坐标，为公共数据（**云端数据**）中的肿瘤疾病类型，纵坐标为所选分子在不同肿瘤数据中的表达量，默认 \log_2 化处理，即 $\log_2(\text{value}+1)$ 。
- 每个连线代表一个配对样本，即所选公共数据中的癌旁（Normal）-vs-肿瘤（Tumor）样本。**注意，无配对关系的数据将无法进行分析。**
- 一般线的趋势方向越一致，并且越倾斜，两组的差异越明显。如果线的趋势不明显，此时**可以通过添加箱式图添加整体的中位数情况**，可能会更加直观地显示出两组的差异情况，具体的情况需要查看统计描述以及统计检验的结果。
- 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）。

云端数据

云端数据 ×

疾病 请选择

数据过滤: 无

数据格式: log2(value+1)

	疾病系统	疾病名	疾病英文	来源	获取时间	数据集	平台	Wo
<input checked="" type="checkbox"/>	泛疾病	泛癌	Pan-cancer	TCGA	202208	TCGA-ALL	RNAseq	STA
<input type="checkbox"/>	泛疾病	泛癌	Pan-cancer	TCGA	202208	TCGA-ALL	RNAseq	STA
<input type="checkbox"/>	泛疾病	泛癌	Pan-cancer	TCGA	202208	TCGA-ALL	miRNA-seq	BC

选择平台数据

① 只有合适这个模块的云端数据才会展示

确认

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



特殊参数

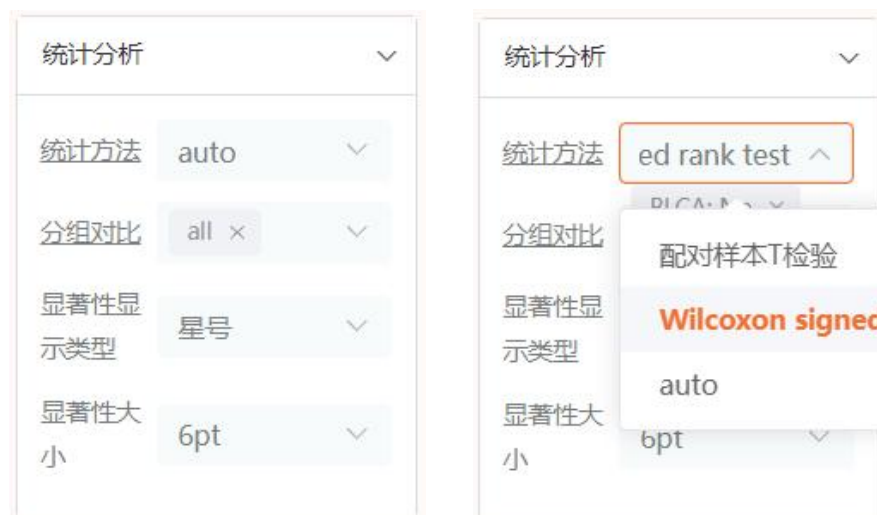
分子

主要参数

- OAZ1[ENSG00000104904.12]
- TRMT1[ENSG00000104907.12]
- STX10[ENSG00000104915.15]
- RETN[ENSG00000104918.8]

- 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



统计分析

统计方法 auto

分组对比 all x

显著性显示类型 星号

显著性大小 6pt

统计分析

统计方法 ed rank test

分组对比

显著性显示类型 Wilcoxon signed

显著性大小 6pt

- **统计方法**：统计方法默认为 auto（自动选择），当第一次点击确认分析后，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信息，默认为 all（全部都标注）。当第一次点击确认分析后，会自动替换成对应上传数据的分组！**此处暂时无作用**。
- **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。

间距设置

间距设置

组间距离

- 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



点

填充色 ▼ ▼

描边色 ▼ ▼

样式 圆形 × ▼

大小 1

不透明度 1

- **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，第一色卡控制癌旁（Normal）分组，第二色卡控制肿瘤（Tumor）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，第一色卡控制癌旁（Normal）分组，第二色卡控制肿瘤（Tumor）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。
- **不透明度**：点的透明度。0 为完全透明，1 为完全不透明。

连线

连线
 ▼

颜色
 ▼

类型
 实线
▼

粗细
 0.75pt
▼

- 颜色：点之间连线的颜色，默认黑色。不受配色方案全局性影响。
- 类型：连线的类型，可选 实线、虚线。
- 粗细：连线的粗细，默认为 0.75pt。



箱

箱
 ▼

展示
 ☐

填充色
 ▼
▼

描边色
 ▼
▼

描边粗细
 0.75pt
▼

不透明度
 1

箱子宽度
 0.6

- 展示：可选是否展示。

- **填充色**：箱子的填充色颜色选项，有多少个分组会提取多少个颜色，第一色卡控制癌旁（Normal）分组，第二色控制肿瘤（Tumor）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**：箱子描边的粗细，默认为 0.75pt。
- **不透明度**：箱子的透明度。0 为完全透明，1 为完全不透明
- **箱子宽度**：箱子的宽度控制，默认 0.6。

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)

图注

是否展示
☒

图注标题
图注标题内容

图注位置
默认

- 展示：是否展示图注
- 图注标题：可以添加图注标题
- 图注位置：可选右、上，默认为右。

坐标轴

坐标轴

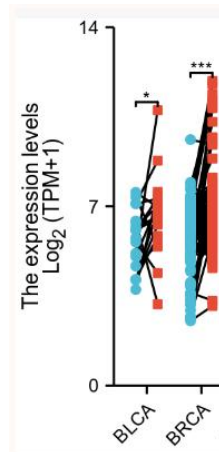
x轴分组名
, +空格隔开

x轴标注旋
转
0

y轴范围+刻度
()包裹,内容用',' +

- **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：用于修改 y 轴范围以及刻度，如果需要分割，需要用小括号(英文输入法)隔开，数值间需要用逗号隔开，例如(1,1,2,5,5)。如果调整过大可能会无作用。
- 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

风格

边框

网格

xy颠倒

文字大小 7pt

- 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

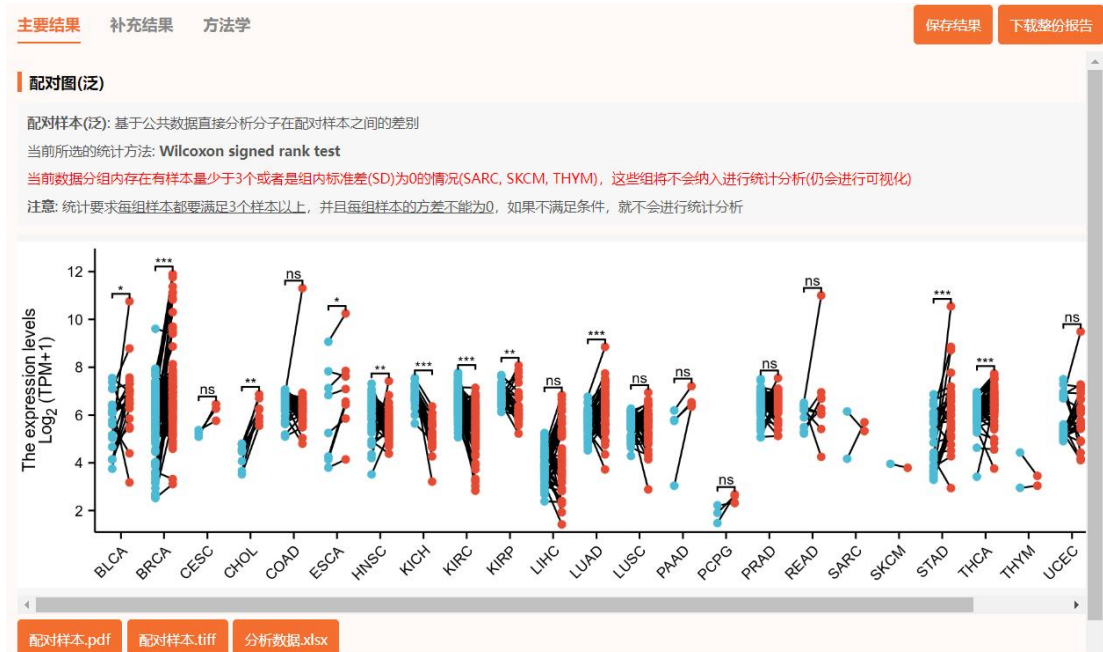


图片	
宽度 (cm)	17
高度 (cm)	5
字体	Arial

- 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。

配对样本(泛): 基于公共数据直接分析分子在配对样本之间的差别
当前所选的统计方法: **Wilcoxon signed rank test**
当前数据分组内存在有样本量少于3个或者是组内标准差(SD)为0的情况(SARC, SKCM, THYM), 这些组将不会纳入进行统计分析(仍会进行可视化)
注意: 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	id	group	x	ERBB2
2	TCGA-22-4593	Tumor	LUSC	6.381710787
3	TCGA-22-4593	Normal	LUSC	5.643001861
4	TCGA-22-4609	Tumor	LUSC	5.255104126
5	TCGA-22-4609	Normal	LUSC	5.751311067
6	TCGA-22-5471	Tumor	LUSC	5.80013371
7	TCGA-22-5471	Normal	LUSC	6.266835254
8	TCGA-22-5472	Tumor	LUSC	6.514669875
9	TCGA-22-5472	Normal	LUSC	6.027605266
10	TCGA-22-5478	Tumor	LUSC	5.985475383
11	TCGA-22-5478	Normal	LUSC	6.008045347
12	TCGA-22-5481	Tumor	LUSC	2.886159913
13	TCGA-22-5481	Normal	LUSC	5.777961446



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)
BLCA	Normal	19	3.7513	7.5472	6.0726	1.5477	4.8555	6.4033	5.7255	1.1719
BLCA	Tumor	19	3.1795	10.758	6.8598	1.2404	6.0662	7.3066	6.7019	1.5798
BRCA	Normal	113	2.5265	9.6093	6.3835	1.441	5.4213	6.8623	5.9514	1.366
BRCA	Tumor	113	3.1105	11.898	6.9148	1.3386	6.1308	7.4694	7.0037	1.5879
CESC	Normal	3	5.0947	5.3666	5.2369	0.13594	5.1658	5.3018	5.2328	0.13598
CESC	Tumor	3	5.7592	6.4568	6.2672	0.34884	6.0132	6.362	6.1611	0.36074
CHOL	Normal	8	3.5203	4.7991	4.4967	0.65884	3.9677	4.6265	4.2993	0.4927
CHOL	Tumor	8	5.5539	6.868	6.236	0.50301	5.8683	6.3713	6.1868	0.45165
COAD	Normal	41	5.1001	7.0693	6.5428	0.60947	6.2952	6.9047	6.489	0.51071
COAD	Tumor	41	4.8006	11.307	6.2912	0.69994	5.8708	6.5707	6.3256	0.93403
ESCA	Normal	8	3.7975	9.0664	6.04	3.0706	4.234	7.3046	6.0387	1.9485
ESCA	Tumor	8	4.1412	10.251	6.8246	1.3823	6.29	7.6723	6.9739	1.7593

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, 注意是配对样本才有, 提供 EXCEL 格式下载。

异常值分析

离群值 = $Q1(\text{下四分位}) - 1.5 * IQR(\text{四分位间距})$ 或者 $Q3(\text{上四分位}) + 1.5 * IQR(\text{四分位间距})$

异常值 = $Q1(\text{下四分位}) - 3.0 * IQR(\text{四分位间距})$ 或者 $Q3(\text{上四分位}) + 3.0 * IQR(\text{四分位间距})$

组别1	组别2	离群值	异常值
BLCA	Tumor	3.17952697396995,...	
BRCA	Normal	2.68826989467837,...	
BRCA	Tumor	10.8351110348054,...	11.7660772710529,...
COAD	Normal	5.10005254667299,...	
COAD	Tumor	11.3068766502765,...	11.3068766502765
ESCA	Tumor	4.14122022953747,...	
HNSC	Normal	3.51675900247186	
HNSC	Tumor	6.83134271687737,...	
KICH	Normal	5.64222791124677	
KICH	Tumor	3.21233587416207	
KIRC	Tumor	2.83430670296291,...	
KIRP	Tumor	8.07721094818306	
LUAD	Normal	4.52162870449255	

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, 注意是配对样本才有, 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	自由度(df)	统计量	p值
BLCA	19	0.97758	0.9107
BRCA	113	0.94154	9.06e-05
CESC	3	0.96789	0.6559
CHOL	8	0.9339	0.5523
COAD	41	0.79879	5.16e-06
ESCA	8	0.88756	0.2221
HNSC	43	0.90956	0.0025
KICH	24	0.94499	0.2106
KIRC	72	0.97602	0.1846
KIRP	32	0.90065	0.0064
LIHC	50	0.98414	0.7341
LUAD	58	0.97446	0.2586
LUSC	49	0.97106	0.2668

正态性检验结果显示, 存在有不满足正态分布的情况($P < 0.05$), 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, 注意是配对样本才有。

Wilcoxon signed rank test

应用条件: 各组内两两配对样本差值满足不满足正态性时

组别	组别I	组别J	统计量V	差值(J-I)	置信区间(95%CI)	p值
BLCA	Normal	Tumor	43	1.0293	0.095191 - 1.8445	0.0361
BRCA	Normal	Tumor	1532	0.89928	0.51284 - 1.3315	1.33e-06
CESC	Normal	Tumor	0	0.95378	0.39254 - 1.3621	0.2500
CHOL	Normal	Tumor	0	1.9022	1.3395 - 2.3861	0.0078
COAD	Normal	Tumor	577	-0.27787	-0.48469 - 0.009316	0.0580
ESCA	Normal	Tumor	1	0.82939	0.19501 - 1.7408	0.0156
HNSC	Normal	Tumor	732	-0.58377	-0.87369 - -0.25189	0.0014
KICH	Normal	Tumor	300	-1.5394	-1.9344 - -1.2081	1.19e-07
KIRC	Normal	Tumor	2452	-1.1311	-1.3924 - -0.87329	1.73e-10
KIRP	Normal	Tumor	411	-0.59487	-0.81815 - -0.23643	0.0050
LIHC	Normal	Tumor	466	0.36041	-0.068135 - 0.82954	0.0988
LUAD	Normal	Tumor	400	0.52685	0.28256 - 0.78355	0.0004
LUSC	Normal	Tumor	656	-0.046535	-0.28114 - 0.18746	0.6715

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, 注意是配对样本才有。

(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. 为什么有一些分组没有标注显著性?

答:

默认只标注满足统计分析要求的分组，当如果样本分组内存在有小于 3 的分组，那么这整个分组都不会进行统计学检验（<3 个样本的分组是没办法进行统计学检验的）。具体每个组的数目可以在 统计描述的表格中找到。

2. TPM、FPKM、RPM 格式的数据有什么区别?

答:

TPM 和 FPKM 是 RNAseq 的一些数据格式,RPM 是 miRNAseq 的数据格式,TPM 是从 FPKM 转换而来，经过了基因组长度的校正。一般建议是用 TPM 用来组间比较，不过也有人用 FPKM 来比的。

3. 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。

4. 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。