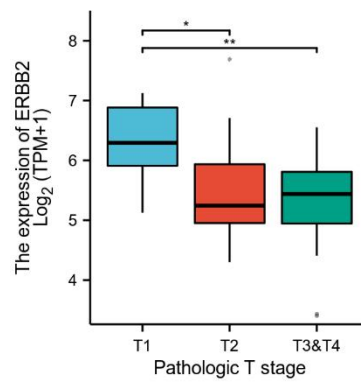


临床意义 - [云] 临床意义分组



网址: <https://www.xiantao love>



更新时间: 2023.03.13

目录

基本概念	3
应用场景	4
分析流程	4
主要结果	5
云端数据	9
参数说明	10
特殊参数	10
统计分析	11
间距设置	12
点	13
箱/柱	14
小提琴	15
误差线	16
标题	17
图注(Legend)	17
坐标轴	18
风格	19
图片	20
结果说明	21
主要结果	21
补充结果	21
方法学	24
如何引用	25
常见问题	26

基本概念

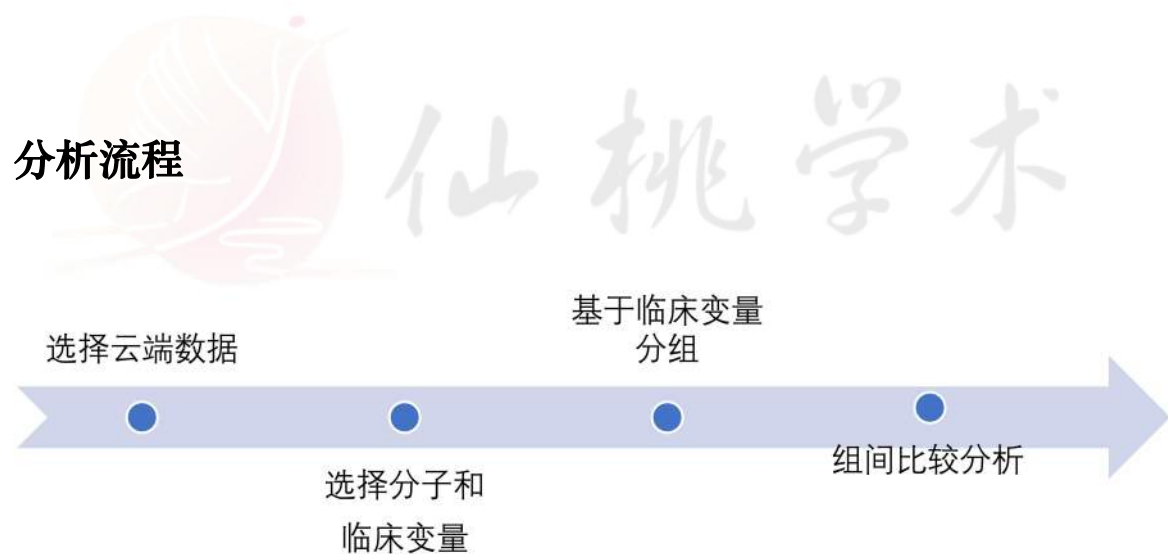
- 分组比较：两组或者多组单个维度或者两个维度的比较。
- 统计方法：
 - **T test**, 亦称 student t 检验 (Student's t test), 主要用于两组之间的比较, 两组需要满足正态性和方差齐性的要求。
 - **Welch's test**, 又称不等方差检验, 即当 两组仅满足正态而不满足方差齐性的要求时, 可以选择用该方法进行两组的比较
 - **Wilcoxon rank sum test**, 也叫 **Mann-Whitney U test** (曼-惠特尼 U 检验), 或者 Wilcoxon-Mann-Whitney test。秩和检验是一个非参的假设检验方法, 一般用于 两组不满足正态性的情况。
 - **One-way ANOVA**, 单因素方差分析是指对单因素试验结果进行分析, 检验因素对试验结果有无显著性影响的方法。单因素方差分析是两个样本平均数比较的引伸, 它是用来检验多个平均数之间的差异, 从而确定因素对试验结果有无显著性影响的一种统计方法。 需要满足正态性和方差齐性的要求。
 - **Welch one-way ANOVA**: 一种特殊的方差分析方法, 当数据 仅满足正态而不满足方差齐性的要求时, 建议选用该方法进行组间比较。
 - **Kruskal-Wallis test**: 又叫克鲁斯卡沃利斯测试, 非参数检验方法。检测是利用多个样本的秩和来推断各样本分别代表的总体的位置有无差别。一般用于 多组样本不满足正态性的情况。
 - **Two-way ANOVA**, 双因素方差分析 (Double factor variance analysis) 有两种类型: 一个是无交互作用的双因素方差分析, 它假定因素 A 和因素 B 的效应之间是相互独立的, 不存在相互关系; 另一个是有交互作用的双因素方差分析, 它假定因素 A 和因素 B 的结合会产生出一种新的效应。当存在有多个因素并怀疑可能存在 交互作用时, 可以选择。

应用场景

临床意义分组比较，是基于公共数据（云端数据）根据所选分子在过滤好的样本中的表达量，直接分析所选临床变量分组之间的差异，分析其是否有统计学差异。

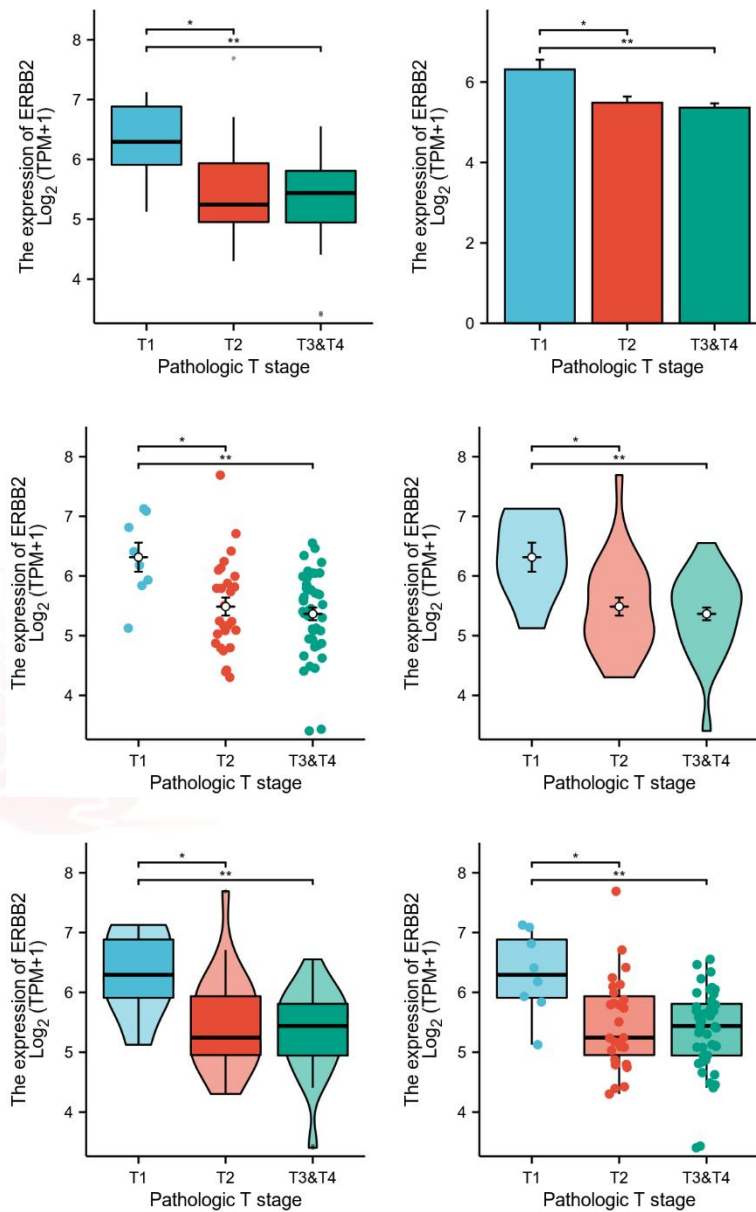
一般绘制箱式图进行直观比较，本模块支持点图、箱式图、柱状图、小提琴图及各自组合的可视化形式。

分析流程



主要结果

单个分子的可视化

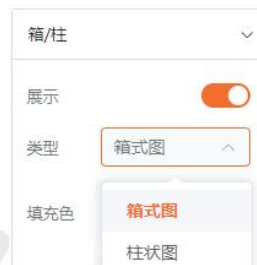


可视化形式：默认-箱式图

- 横坐标，临床变量的分组信息，根据公共数据（云端数据）的临床变量中分类，可自定义肿瘤样本分组，根据需求组合分类，从而形成不同的分组信息。
(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)

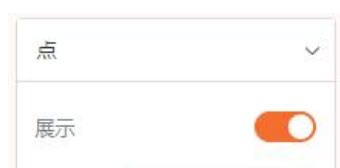


- 纵坐标，为所选分子在不同肿瘤样本中的表达量，默认 \log_2 化处理，即 $\log_2(\text{value}+1)$ 。
- 默认情况下，模块会根据数据的情况，如正态性和方差齐性自动选择合适的统计方法进行统计分析（具体方法见基本概念中的统计方法）。
- 可视化形式：需要选择对应参数中【展示】



■ 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。



- 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。

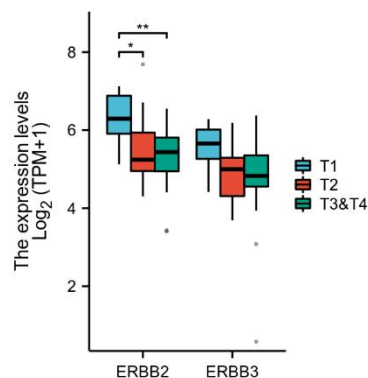


- 小提琴图：形状类似小提琴，同一水平线上分布的样本越多，则越宽，否则就越窄。小提琴图能有效展示分组内的样本情况的分布。



组合图：点图和箱式图

(多选) 多个分子的可视化



上图为，选择多个分子（基因）时的可视化形式。

- 横坐标为分子，纵坐标为分子的表达量数据，默认 \log 化处理，即 $\log_2(\text{value}+1)$ 。
- 箱子（或其他形式）代表临床变量分组，如上图按照临床变量 (Pathologic_T_stage) 在云端数据集中的分类将肿瘤样本分为 T1、T2、T3&T4 三组。



云端数据

数据参数
重置参数

云端数据

食管鳞癌 / TCGA / TCGA-ESCC / RNAseq / STAR / TPM @过滤:去除正常 @处理:log2(value+1)

疾病名/来源/数据集/平台/分析流程/数据格式
@数据处理方式

云端数据 选择疾病
过滤数据: 默认 去除正常 + 去除无临床信息
×

疾病 请选择 v

数据过滤: 去除正常 + 去除无临床信息 v

数据格式: log2(value+1) v

	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"/>	食管	食管鳞癌	Esophagus squamous cell carcinoma	TCGA	202208	TCGA-ESCC	RNAseq	STA
<input type="checkbox"/>	食管	食管癌	Esophageal carcinoma	TCGA	202208	TCGA-ESCA	RNAseq	STA
<input type="checkbox"/>	食管	食管癌	Esophageal carcinoma	TCGA	202208	TCGA-ESCA	RNAseq	STA
<input type="checkbox"/>	食管	食管腺癌	Esophagus adenocarcinoma	TCGA	202208	TCGA-ESAD	RNAseq	STA

共 115 条 上一页 1 2 3 4 5 6 ... 12 下一页

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和临床变量可能会有不同。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，[可选多个分析](#)。



- 临床变量：下拉框将列出对应所选数据集的临床变量，选中变量后，右侧可选关联的分类信息，如 Pathologic_T_stage 对应 T1-T4 分类。



- **分组**：在变量对应的分类中自定义比较分组。   加减号修改分组，一个框内的分类组可以合成一个组，如 T3 和 T4 分类作为一组等等。[注意](#),

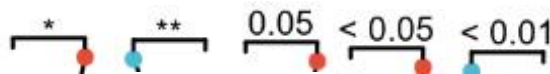
选择的分类（单个组）中样本数需要大于等于 3，且分组数大于等于 2，才能进行统计分析。根据具体情况可以自由选择参考组的分组。

统计分析



- **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，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，则可能不会出现在此处。）
- 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：

补充说明：

- 统计方法: One-way ANOVA
- 所选分子: ERBB2[ENSG00000141736.14]

间距设置

间距设置
▼

(二维)组内总宽度

- 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，最多支持修改 6 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，最多支持修改 6 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。
- 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**：箱子的填充色颜色选项，有多少个分组会提取多少个颜色，最多支持修改 6 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 6 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 0.8。

小提琴

小提琴

展示 ☐

填充色

描边色

描边粗细 0.75pt

不透明度 0.5

宽度 0.8

宽度校正 1

- 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，最多支持修改 6 个颜色。受配色方案全局性修改。
- 描边色：小提琴的描边色颜色选项，有多少个分组会提取多少个颜色，默认黑色，最多支持修改 6 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**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

标题 ▼

大标题

大标题内容

x轴标题

x轴标题内容

y轴标题

y轴标题内容

- 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 默认、右、上，默认为右。

坐标轴

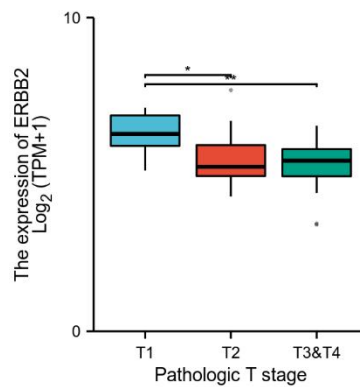
坐标轴

x轴分组名 ,+空格隔开

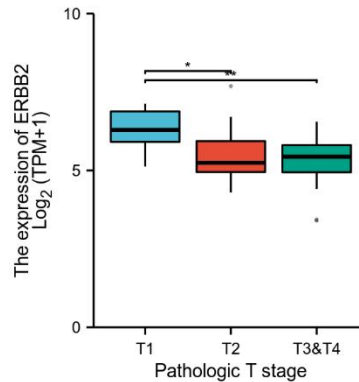
x轴标注旋
转 0

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**：（注意：范围的修改如果调整过大会失效）
 - 如果只是想要修改范围，可以只输入两个范围值，比如 0,0,10,10



- 如果同时想要修改范围+刻度，可以输入比如：0,0,0.5,1,1 。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 0,1 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

图片

▼

宽度 (cm)

5

高度 (cm)

5

字体

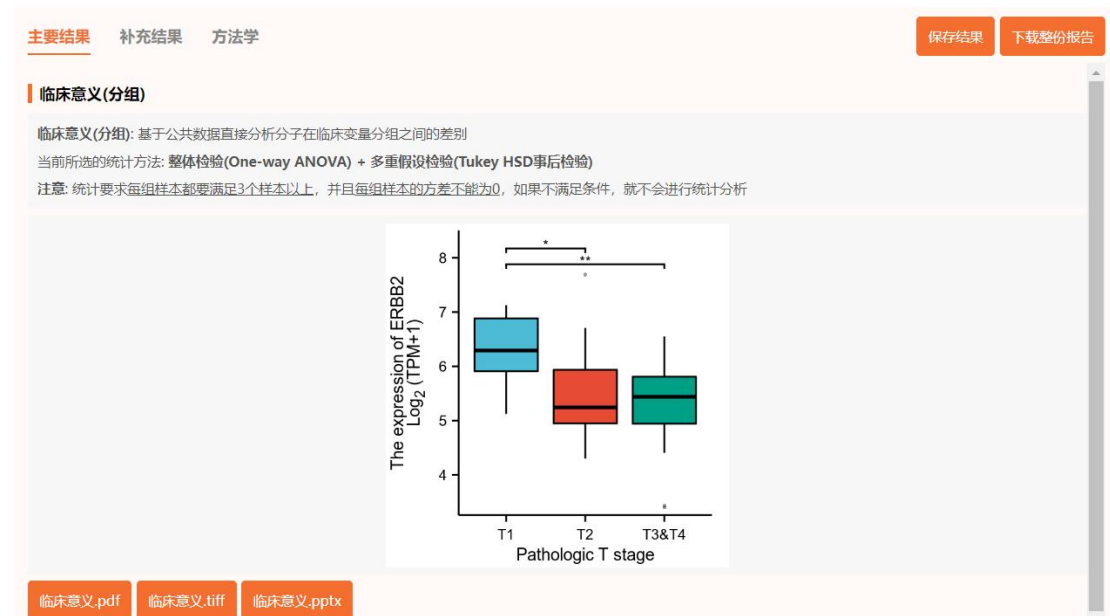
Arial

▼

- 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。

补充结果

统计描述

各个组常见「统计描述指标」

组别	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)	标准误(SE)
T1	8	5.1254	7.1264	6.2935	0.97343	5.9092	6.8826	6.3142	0.68861	0.24346
T2	27	4.3016	7.6902	5.2439	0.98525	4.9512	5.9365	5.4874	0.77983	0.15008
T3&T4	44	3.4034	6.5524	5.4386	0.86507	4.9441	5.8092	5.3648	0.69884	0.10535

统计描述.xlsx

此表格提供统计描述的结果，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	离群值	异常值
T2	7.69016214286973	
T3&T4	3.43170318148885,...	

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	自由度(df)	统计量	p值
T1	8	0.94677	0.6787
T3&T4	44	0.95084	0.0591
T2	27	0.95249	0.2463

正态性检验结果显示，观测变量在各组内接近正态分布($P > 0.05$)，建议选择用 参数检验的方法

此表格为正态性检验的结果。

方差齐性检验

检验方法: Levene's test

· Base on Mean

自由度1(df1)	自由度2(df2)	统计量	p值
2	76	0.3415	0.7118

方差齐性检验显示，各组观测变量的方差相等($P > 0.05$)

此表格为方差齐性检验的结果。

One-way ANOVA					
比较的组	分子自由度(DFn)	分母自由度(DFd)	统计量	p值	η^2
组内比较	2	76	5.7865	0.0046	0.13215

多重假设检验(Tukey HSD事后检验)				
分组I	分组J	估计值(J-I)	置信区间(95%CI)	校正后p值
T1	T3&T4	-0.94938	-1.617 - -0.28173	0.0031
T1	T2	-0.82677	-1.526 - -0.12753	0.0164
T3&T4	T2		-	NA

此表格为比较组之间统计检验的结果。

(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: 对主变量进行分组后, 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. 为什么工具出来的结果跟别的同样数据集的数据库的不一样?

答:

① **表达数据可能不同**: 工具用的表达数据是直接对应数据库最新下载。另外, 数据过滤和处理的情况也会有影响, 工具所有经过的处理都在方法学中进行了描述。总之, **不同的数据**、**数据格式**、**数据处理**和**样本例数**情况, 均会影响结果。

② 临床变量可能会不同: 不同时间版本的。

③ **统计方法**: 统计方法都是成熟的, 不存在方法学上的不同导致的。

所以, 如果只是单纯想要拿一些结果来充实自己的研究, 那么可以只放满足自己想要的趋势的数据。

2. 在别的数据库上看到一个分子的趋势 跟 工具做出来的不一样?

答:

即便是同一个分子同一个疾病, 不同数据集得到的结果都可能会有差别, 甚至是存在趋势相反的情况。不同数据集之间可能存在有很多混杂因素, 并不能完全做到控制好所有的变量和情况, 难免是有可能出现 趋势不同或者相反的情况的。所以, 如果只是单纯想要拿一些结果来充实自己的研究, 那么可以只放满足自己想要的趋势的数据。

3. 在云端数据框内看到的例数、选择临床变量的数目以及分析时候的例数不同, 这个是什么情况?

答:

云端数据的例数一般是对应组学所有的例数, 选择临床变量的选项的数目为没有去除重复样本的变量数量, 分析时候可能会有剔除样本(数据信息中可选去除重复), 具体需要看说明文本中对于数据的处理情况的说明。

有一些云端数据是存在有一些临床样本检测了多次的情况，去除重复检测的样本，能够降低同一份临床数据被同时纳入而影响结果。虽然存在有重复检测，但是一般这些重复检测的样本的数量很少。同样，也有一些云端数据对应的临床数据是只有临床数据，而没有对应的平台（组学）的检测的，一般这些没有检测的数据都是会被剔除的。另外，也存在有进行了对应平台（组学）检测，但是临床信息缺失的情况。（具体的数量对应和过滤的细节，可以看下一个问题）。

4. 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。

5. 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。