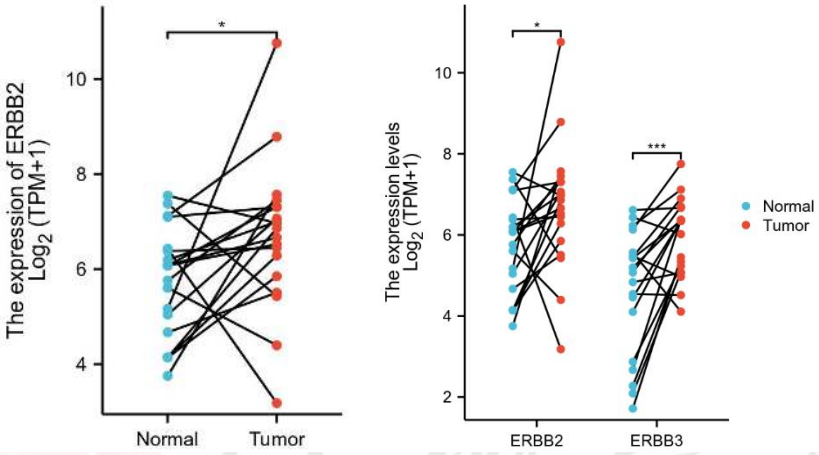
表达差异 - [云]配对样本



网址：https://www.xiantao.love

更新时间：2023.02.24

目录

基本概念

- 配对图：将有配对关系的样本进行可视化的一种方式。

  配对样本能有效控制其他混杂因素的影响。配对有这几种类型：① 前后配对；② 同一受试对象，两个不同部位配对；③ 条件配对：根据某一个特征相同，比如年龄、性别相同的患者配成 1 对；④ 同一份样本两种检测方法配对。

- 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。

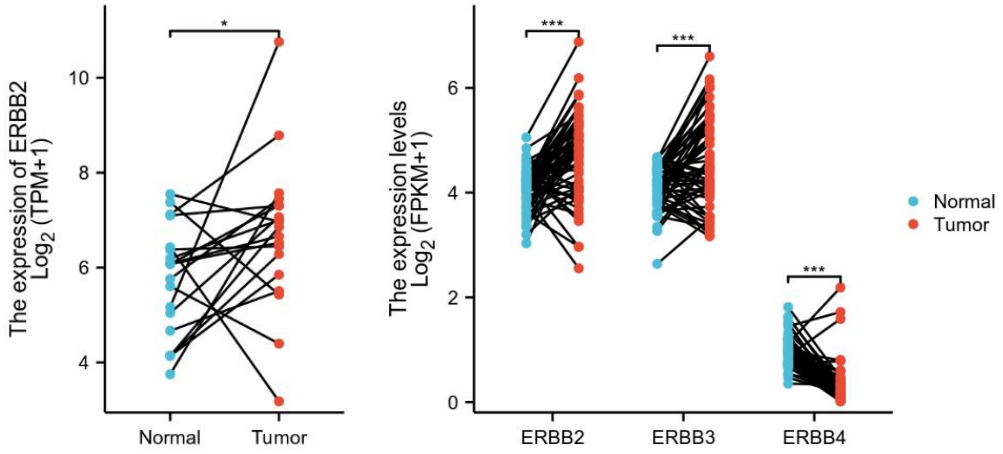
  - 配对样本 T 检验：用于检验配对类型的参数检验方法。适用条件：连续变量、配对关系、差值服从或者近似服从正态性。
  - Wilcoxon signed rank test：符号秩和检验，用于检验配对类型的非参数检验方法，适用于不满足配对样本 T 检验的研究，即不满足正态性！

应用场景

配对样本图，是基于公共数据（云端数据）直接分析分子在配对的样本之间的差别。可同时展示同一个配对组中的多个分子的表达情况，一般绘制点线图。

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！

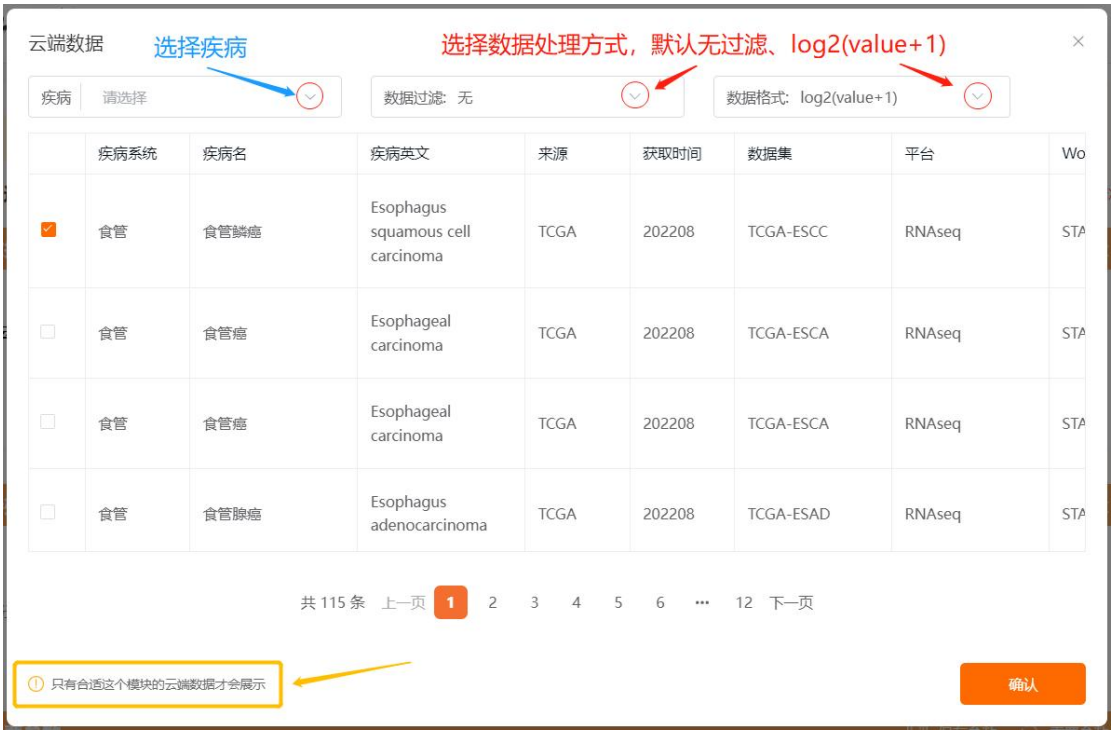
主要结果



左图：单个指标/分子的情况；右图：多个指标/分子的情况

- 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。
- 最多 1 个图 15 个分子，更多的分子建议分成多个图展示。

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

- 特殊参数：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，支持多个分子，最多 15 个分子！

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

- 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

- 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。
- 透明度：点的透明度。0 为完全透明，1 为完全不透明。

连线

- 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

- 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]]

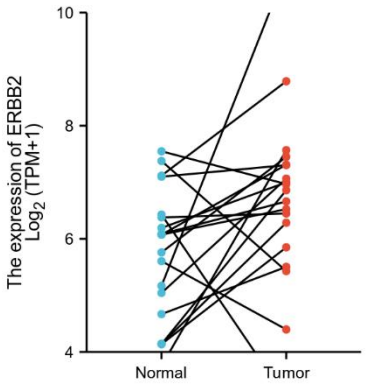
图注(Legend)

- 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。

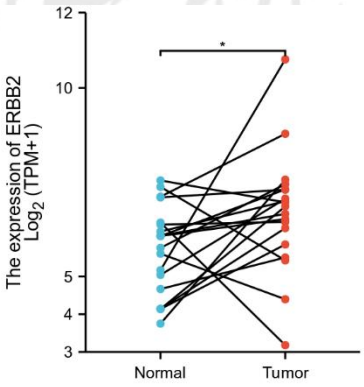
坐标轴

- 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。
  - 如果只是想要修改范围，可以只输入两个范围值，比如 4,10：



  - 如果同时想要修改范围+刻度，可以输入比如：3,3,4,5,10,12,12。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 3 和 12 那样同时写两次：



风格

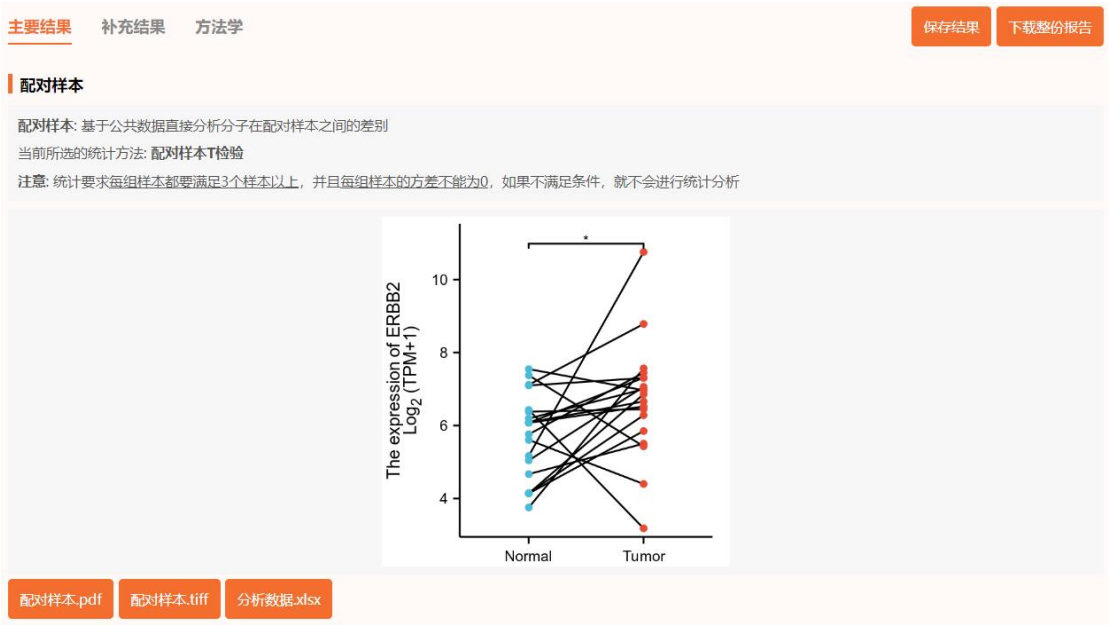
- 外框：是否添加外框
- 网格：是否添加网格
- 是否颠倒 XY 轴：可以颠倒 xy 轴
- 文字大小：针对图中所有文字整体的大小控制

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

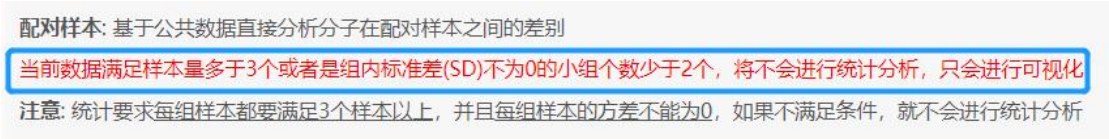
结果说明

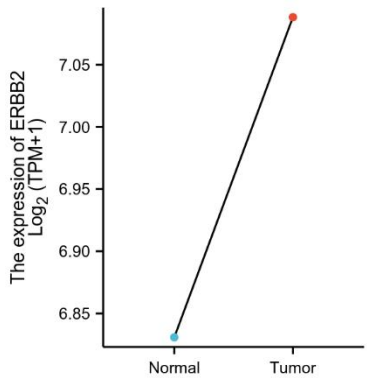
主要结果



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及分组的表达量数据，提供 EXCEL 格式下载：

	A	B	C
1	pair_info	status	ERBB2
2	TCGA-BL-A13J	Tumor	6.283282861
3	TCGA-BL-A13J	Normal	4.145889361
4	TCGA-BT-A20N	Tumor	5.849068511
5	TCGA-BT-A20N	Normal	4.137601886
6	TCGA-BT-A20Q	Tumor	6.859774543
7	TCGA-BT-A20Q	Normal	4.138093596
8	TCGA-BT-A20R	Tumor	7.566767216
9	TCGA-BT-A20R	Normal	3.751335171
10	TCGA-BT-A20U	Tumor	6.660378183
11	TCGA-BT-A20U	Normal	6.084848772
12	TCGA-BT-A20W	Tumor	7.309896482
13	TCGA-BT-A20W	Normal	6.072601052

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(S
Normal	19	3.7513	7.5472	6.0726	1.5477	4.8555	6.4033	5.7255	1.1719	0.26884
Tumor	19	3.1795	10.758	6.8598	1.2404	6.0662	7.3066	6.7019	1.5798	0.36244

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(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	离群值	异常值
Tumor	3.17952697396995,...	

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

自由度(df)	统计量	p值
19	0.97758	0.9107

正态性检验结果显示，各组配对样本<差值>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。

配对样本T检验

应用条件: 各组内两两配对样本差值满足正态性检验

组别I	组别J	自由度(df)	统计量t	差值(J-I)	置信区间(95%CI)	p值
Normal	Tumor	18	2.129	0.97639	0.012873 - 1.9399	0.0473

p值满足<0.05时，可认为两组存在统计学上差异

此表格为 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. 为什么配对和非配对的数据结果之间会有差异？一个有意义，一个没有意义？

答：

结果有差异一般有这么几种可能：① 样本量不同，可能非配对的样本量更大，增加了很多别的样本；② 统计方法不同，本身配对和非配对的统计方法就是不同的，出来的结果不同也是有可能的

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，难免是有可能会出现 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和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，分析时候可能会有剔除样本的情况，比如，有一些样本是只有正常而没有配对的，有一些是只有疾病而没有正常的，所以在非配对中看到的正常样本数未必就等于在配对中的正常数 !具体需要看说明文本中对于数据的处理情况的说明。

4. 为什么有一些分组没有标注显著性？

答:

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。

5. 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。

6. 为什么在云端数据中找不到一些数据集，但是在一些别的模块能找到?

答:

如果对应的数据集不满足对应模块的要求,是不会在对应模块的云端数据中出现的。

7. 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。