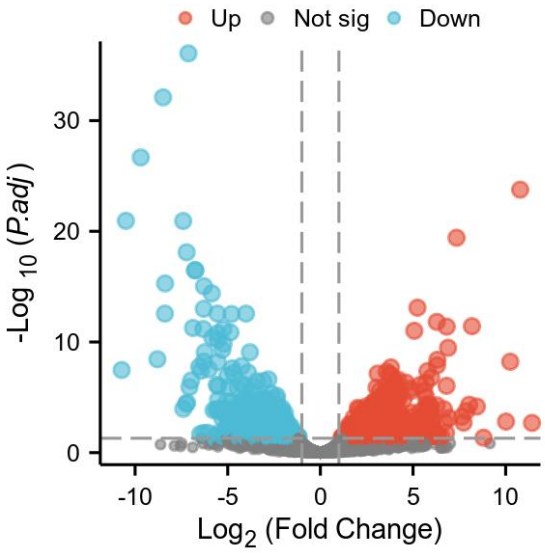
表达差异 - 火山图



网址：https://www.xiantao.love

更新时间：2023.02.19

目录

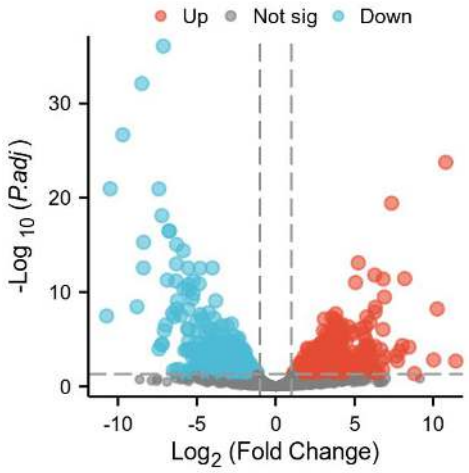
基本概念

火山图：散点图的一种，它将统计测试中的统计显著性量度（如 p-value）和变化幅度相结合，从而能够帮助快速直观地识别那些变化幅度较大且具有统计学意义的数据点（基因等）

一些问题：

- Fold change ：差异倍数，简单来说就是基因在一组样品中的表达值的均值除以其在另一组样品中的表达值的均值。所以火山图只适合展示两组样品之间的比较
- 为什么要做 Log 2 转换？ 两个数相除获得的结果（fold change）要么大于 1，要么小于 1，要么等于 1。对于基因差异，简单说，大于 1 表示上调（可以描述为上调多少倍），小于 1 表示下调（可以描述为下调为原来的多少分之多少）。大于 1 可以到多大呢？多大都有可能。小于 1 可以到多小呢？最小到 0。用原始的 fold change 描述上调方便，描述下调不方便。绘制到图中时，上调占的空间多，下调占的空间少，展示起来不方便。所以一般会做 Log 2 转换。默认我们都会用两倍差异 （fold change == 2 | 0.5）做为一个筛选标准。Log 2 转换的优势就体现出来了，上调的基因转换后 Log 2 (fold change) 都大于等于 1，下调的基因转换后 Log 2 (fold change)都小于等于-1。无论是展示还是描述是不是都更方便。

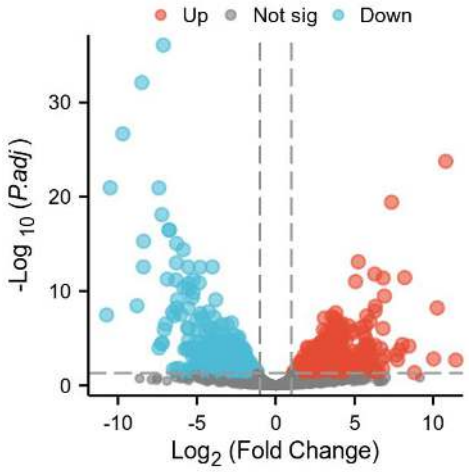
- 图形构成



应用场景

常用于转录组差异分析的结果可视化，也能应用于基因组，蛋白质组，代谢组等统计数据

结果解读



- 图中横坐标代表 logFC，纵坐标代表 p 值或者校正后 p 值。图中每个点表示一个检测到的基因。红色和蓝色分别表示上调和下调基因，灰色表示无差异基因。
- 整体来看，基因有上调就有下调，图整体是以 X=0 的垂线左右对称的。点越偏离中心，表示差异倍数越大。如果数据中大部分点都是上调或下调，成偏态分布时，需考虑标准化步骤没有处理好，或数据存在批次效应，导致数据存在系统偏差。
- 图的左上角和右上角是差异基因集中的地方，也是我们关注的重点。

数据格式

	A	B	C
1	gene	logFC	padj
2	TINAG	11.42176924	0.002045602
3	CPLX2	10.78346635	1.72926E-24
4	INS-IGF2	-10.72502805	3.40332E-08
5	CLEC4M	-10.4939252	1.13013E-21
6	HOXA13	10.24743564	6.05726E-09
7	PAGE4	10.02038069	0.001547695
8	CLEC1B	-9.694371799	2.12659E-27
9	CHP2	9.146296978	0.147290991
10	RTBDN	8.80808934	0.044341667
11	CFTR	-8.793154523	3.58571E-09
12	SPRR1B	-8.625761728	0.181574636
13	CXCL14	-8.488664898	7.93508E-33
14	TM4SF20	8.446470341	6.82178E-05
15	PZP	-8.391045162	2.72023E-13
16	MARCO	-8.37961331	5.15118E-16
17	EEF1A2	8.180360625	3.72551E-12
18	LHFPL4	8.031405565	4.66021E-05
19	PNCK	7.87480692	0.000175449
20	SPRR2D	-7.87463873	0.238833257

表格：不同基因的差异分析结果

- 3 列数据: ID 列 | logFC 列 | p 值或者校正后 p 值(p.adj)列（注意列名）. 数据会过滤掉 logFC 或者是 p 值列有缺失的数据

参数说明

（说明：标注了颜色的为常用参数。）

阈值

- logFC 阈值：可以控制图中划分是否有显著差异表达对应的 logFC 阈值
- p 值阈值：可以控制图中划分是否有显著差异表达对应的 p 值阈值
- 对称 x 轴长度：默认是强制 x 轴对称

标注

- 标注 id（第一列）：可以标注想要标注的基因
- 标注大小：默认是 5pt
- 标记差异数量：可以在图中展示差异基因的数量

点

- 填充色：点的填充色
- 描边色：点的描边色
- 样式：可以选择圆形、三角形、正方形等形状
- 大小：点的大小
- 不透明度：0 为完全透明，1 为完全不透明。

标题

- 是否显示：是否显示误差线
- 大标题：大标题文本
- X 轴标题：X 轴标题
- Y 轴标题：Y 轴标题

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

- 展示：是否展示图注
- 图注标题：输入图注的标题内容
- 图注位置：可选右、上，默认为右

风格

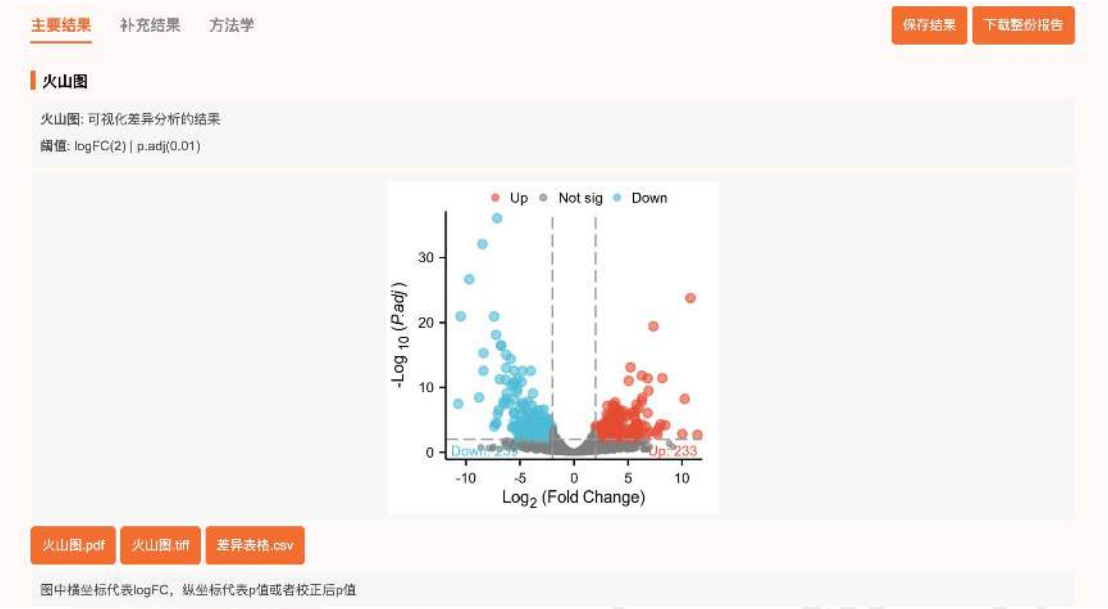
- 边框：是否显示主图边框
- 网格：是否添加网格
- xy 颠倒：x 轴和 y 轴是否转换
- 文字大小：图中的文字部分的大小（包括标签文字和刻度数），默认是 7pt

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

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，差异表格下载。

补充结果

1. 统计


主要结果 补充结果 方法学 保存结果 下载整份报告

差异统计

统计一些常见阈值(|logFC|大于2或者1或者是0.58(0.58换算过来就是1.5倍))下的差异分子数量

筛选条件	筛选后的数量
|LogFC|>2 & p.adj<0.05	700
|LogFC|>1.5 & p.adj<0.05	881
|LogFC|>1 & p.adj<0.05	942
|LogFC|>0.58 & p.adj<0.05	942

统计一些常见阈值(|logFC|大于 2 或者 1 或者是 0.58(0.58 换算过来就是 1.5 倍))下的差异分子数量

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[3.3.6]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. 为什么为什么做-Log 10 转换呢?

答：因为假阳性率 FDR（false discovery rate）值是 0-1 之间，数值越小越是统计显著，也越是我们关注的。-Log 10 (adjusted P-value)转换后正好是反了多来，数值越大越显著，而且以 10 为底很容易换算回去。

2. 什么是 adjusted P-value?

答：做差异基因检测时，要对成千上万的基因分别做差异统计检验。统计学家认为做这么多次的检验，本身就会引入假阳性结果，需要做一个多重假设检验校正。这个校正怎么做呢？最简单粗暴的方法是每一次统计检验获得的 P-value 都乘以总的统计检验的次数获得 adjusted P-value (这就是 Bonferroni correction）。但这样操作太严苛了，很容易降低统计检出力，找不到有差异的基因。后续又有统计学家提出相对不这么严苛的计算方法，如 holm, hochberg, hommel, BH, BY, fdr 等。BH 是我们比较常用的一个校正方法，获得的值是假阳性率 FDR（false discovery rate）。FDR 筛选时就可以不用遵循 0.05 这个标准了。我们可以设置 FDR<0.05 表示我们容许数据中存在至多 5%假阳性率；FDR<0.1 表示我们对假阳性率的容忍度至多是 10%。当然如果说我们设置 FDR<0.5，即数据中最多可能有一半是假阳性就说不过去了。