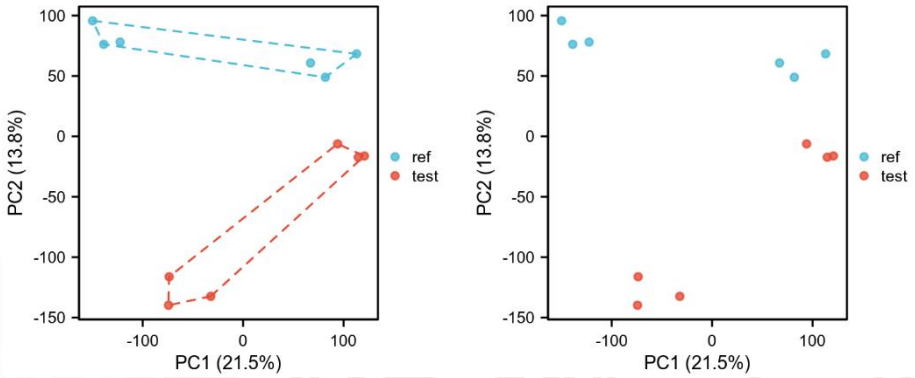
数据集工具 – [数据集] PCA 图



网址：https://www.xiantao.love

更新时间：2023.03.13

目录

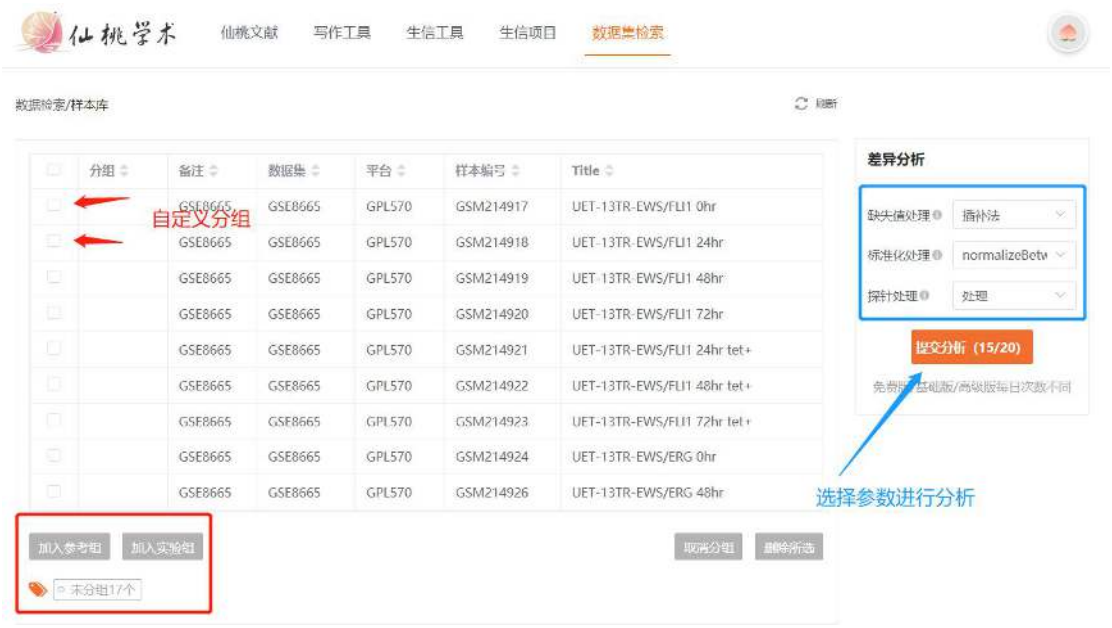
基本概念

- 数据集差异分析：从数据集检索模块中，针对特定 GEO 数据集的数据，进行芯片差异分析的过程，类似 GEO2R。
- PCA（主成分分析）：数据降维的方法。从高纬数据中提取数据的特征向量（成分），转换为低维数据并且用二维或者三维的图来展示这些特征。从特征向量中提取最能体现数据特征（差异）的 2 个特征向量（成分）用于可视化，这就是 PCA 图。

应用场景

本模块为 数据集检索 – 差异分析 后结果的可视化展示。可以用于查看数据特征情况，查看数据集表达谱中样本间差异的情况。

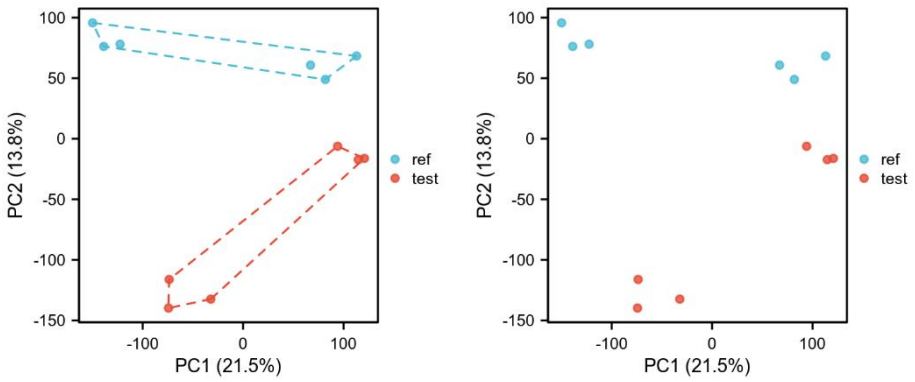
注意：模块需要先进行 数据集检索 – 差异分析 后，此处的云端数据才会有结果记录，然后才能进行可视化的操作。


仙桃学术
仙桃文献
写作工具
生信工具
生信项目
数据集检索
数据检索/样本库
刷新
分组
备注
数据集
平台
样本编号
Title
GSE8665 GSE8665 GPL570 GSM214917 UET-13TR-EWS/FLI1 0hr
GSE8665 GSE8665 GPL570 GSM214918 UET-13TR-EWS/FLI1 24hr
GSE8665 GSE8665 GPL570 GSM214919 UET-13TR-EWS/FLI1 48hr
GSE8665 GSE8665 GPL570 GSM214920 UET-13TR-EWS/FLI1 72hr
GSE8665 GSE8665 GPL570 GSM214921 UET-13TR-EWS/FLI1 24hr tet+
GSE8665 GSE8665 GPL570 GSM214922 UET-13TR-EWS/FLI1 48hr tet+
GSE8665 GSE8665 GPL570 GSM214923 UET-13TR-EWS/FLI1 72hr tet+
GSE8665 GSE8665 GPL570 GSM214924 UET-13TR-EWS/ERG 0hr
GSE8665 GSE8665 GPL570 GSM214926 UET-13TR-EWS/ERG 48hr
自定义分组
差异分析
缺失值处理
插补法
标准化处理
normalizeBetw
探针处理
处理
提交分析（15/20）
选择参数进行分析
加入参考组
加入实验组
未分组17个
取消分组
删除所选


分析记录
时间最新在最上方记录
历史记录中超过30天的记录会自动清理!
刷新
ID 名称 模块 状态 类型 时间 操作
1 芯片-差异分析 完成 表格 2023-03-12 20:35:56 更名 | 删除 | 下载 | 查看
2 芯片-差异分析 完成 表格 2023-03-12 20:14:16 更名 | 删除 | 下载 | 查看
3 芯片-差异分析 完成 表格 2023-03-12 20:13:12 更名 | 删除 | 下载 | 查看
4 芯片-差异分析 完成 表格 2023-03-12 20:10:10 更名 | 删除 | 下载 | 查看

分析流程

主要结果



典型的 PCA 图以点图形式展示。

- x 轴和 y 轴分别代表 主成分 1（PC1）和主成分 2（PC2），其中图中（x 轴标题）PC1 能体现 21.5% 的数据的特征差异，其中图中（y 轴标题）PC2 能体现 13.8% 的数据的特征差异，故整个 PCA 图能体现数据还不到一半的差异。（因为数据是高维数据，前两个主成分未必就能体现绝大部分的差异，具体数据具体分析）。
- 图中每个点代表每个样本在主成分 1 和主成分 2 中对应的映射位置信息，单个样本的数值大小不能体现单个样本说明特征情况，需要整体来看。点与点（样本与样本）间的距离情况能体现样本间的差异。
- 图中不同的颜色表征不同样本所属的分组，即在差异分析阶段，自定义的参考组（默认 ref）和实验组（默认 test）。

- 右图中给样本不同组增加了置信椭圆的圈（如果分组内样本差异过大，可能会没办法圈住样本的椭圆的圈）

云端数据

云端数据

	记录名称	来源模块	时间	补充说明
☑		芯片-差异分析 @1.0	2023-03-12 20:35:56	数据记录可以在历史记录中找到

这里的云端数据与历史记录汇总 数据集检索工具样本库中【差异分析】 的数据记录是保持一致的，可以在历史记录中找到相应的数据记录。

根据需要可视化的项目 选择好对应的云端数据记录。默认使用最近生成的分析记录。

参数说明

（说明：标注了颜色的为常用参数。）

数据处理

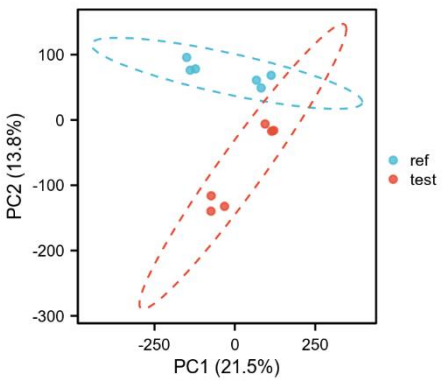
- 归一化：对特征进行归一化可以有效减少特征之间数量级过大的问题，可以选 对行(变量)归一化、无。

点

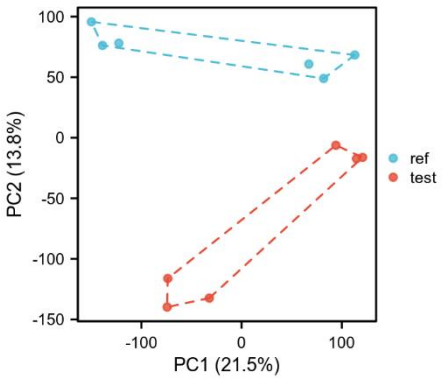
- 填充色：点的填充色颜色选项，有多少个分组会提取多少个颜色，第一色卡控制参考组（默认 ref）分组，第二色卡控制实验组（默认 test）分组，最多支持修改 2 个颜色。受配色方案全局性修改。
- 描边色：点的描边色颜色选项，有多少个分组会提取多少个颜色，第一色卡控制参考组（默认 ref）分组，第二色卡控制实验组（默认 test）分组，最多支持修改 2 个颜色。受配色方案全局性修改。
- 样式：点的样式类型，可选择 圆形、正方形、菱形、三角形、倒三角，默认为圆形。多选，多选后不同的分组/分类中的点的类型也会有相应变化，循环取该参数值。
- 大小：点的大小。
- 不透明度：点的透明度。0 为完全透明，1 为完全不透明。

外圈

- 展示：是否需要圈住分组的不同分类。
- 样式：外圈的样式类型，可选择 连线、椭圆，默认为连线。单选，选择类型后所有圈的样式都统一改变。
  - 椭圆，即置信椭圆。（注意，不是所有的分类都能有圈的，如果分类内含有极端的样本，可能没有办法有圈，另外样本多少也会影响是否能有圈，如单个分组内少于 3 个样本则无法添加）

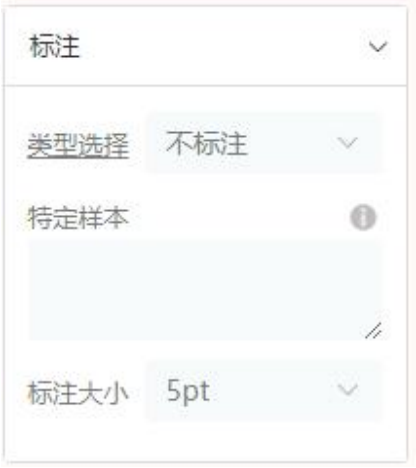


  - 连线，是由各个组最外层的点连接而成，起码两个样本及以上。



- 描边线条类型：外圈的描边样式类型，可选择 实线、虚线，默认为虚线。单选，选择类型后所有圈的描边都统一改变。
- 描边粗细：外圈的描边粗细，默认为 0.75pt。

标注



- 类型选择：是否需要标注样本编号信息。可选择 不标注、标注全部样本、标注下面特定样本，默认为不标注。

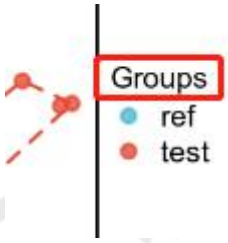
- 特定样本：当上一个参数选择了“标注下面特定样本”时，将根据此参数输入的样本编号在图上进行标注，一行一个。注意样本编号是否与上传数据的样本信息保持一致！
- 标注大小：控制图中需标注的文字大小，默认为 5pt。

标题

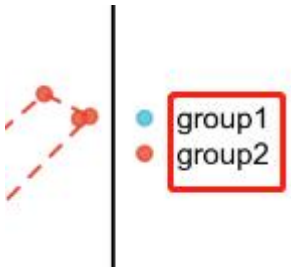
- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本
- 补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]。

图注(Legend)

- 是否展示：是否展示图注
- 图注标题：可以添加图注标题，默认不标注标题，如标注 Groups 时：

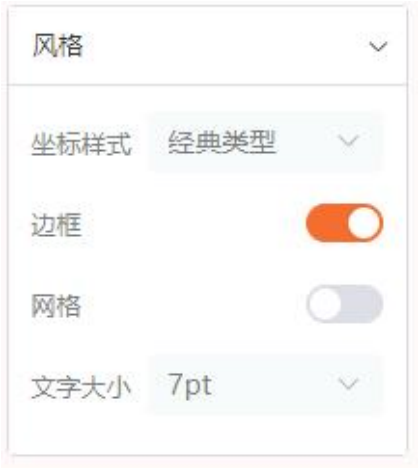


- 图注标签：可以修改图注中分组标签的名字，如果有多个名字要修改，则需要把这些名字以英文逗号的形式合并成一个，类似 group1, group2：

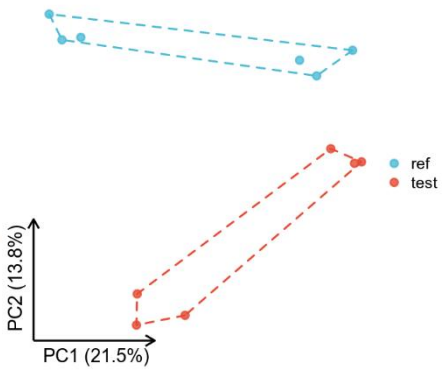


- 图注位置：可选右、上，默认为右。

风格



- 坐标样式：无边框的情况下，坐标轴的样式。可选择 指向类型、经典类型，默认为经典类型。指向类型时，注意需要去除边框，否则无效，如下：



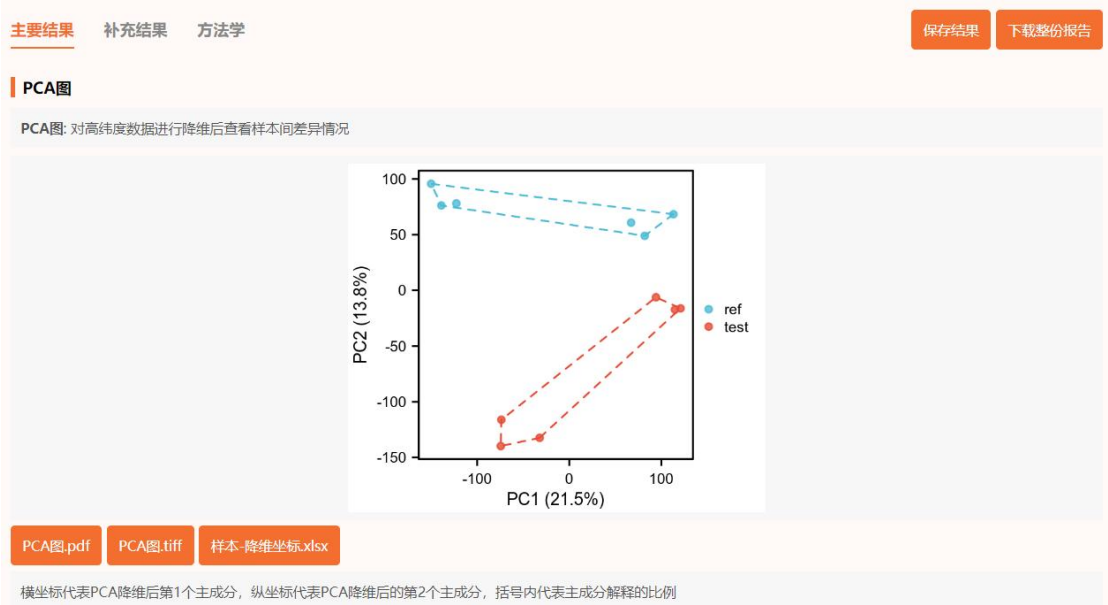
- 边框：是否添加外框
- 网格：是否添加网格
- 文字大小：针对图中所有文字整体的大小控制，默认为 7pt

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
- 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。

- 另外，提供各个样本的降维坐标结果表格 xlsx 下载，含有每个样本对应主成分 1 和主成分 2 的位置信息。

	A	B	C
1	sample	PC1	PC2
2	GSM214918	-122.5087687	78.05869142
3	GSM214919	-138.8452774	76.19981629
4	GSM214920	-149.9155965	95.67133399
5	GSM214926	67.09256718	60.76403204
6	GSM214927	112.9980272	68.29280602
7	GSM214925	81.86165384	48.93515306
8	GSM214921	-73.67186339	-116.1521898
9	GSM214922	-74.31474782	-139.748319
10	GSM214923	-32.19057622	-132.3641181
11	GSM214928	94.08761908	-6.220712565
12	GSM214929	114.6570855	-17.26007075
13	GSM214930	120.7498772	-16.17642262

补充结果

PCA主成分

PCA降维后前10成分对应的解释数据变异情况的比例以及累积比例情况

一般只看主成分1和主成分2解释比例，没有硬性要求要达到多少比例，但是也不能太低

主成分	解释比例(%)	累积比例(%)
PC1	21.5	21.5
PC2	13.8	35.3
PC3	9.8	45.1
PC4	9.2	54.2
PC5	8.0	62.2
PC6	7.2	69.4
PC7	6.6	76.0
PC8	6.5	82.4
PC9	6.0	88.4
PC10	5.8	94.3
PC11	5.7	100.0
PC12	0.0	100.0

此表格为各主成分的解释比例和累积比例，如 PC1 的解释比例为 21.5%，则表示 x 轴的差异可以解释全面分析结果的 21.5%。

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）

处理过程：基于数据集检索模块的差异分析结果，将 PCA 分析结果用 ggplot2 包进行可视化。

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。

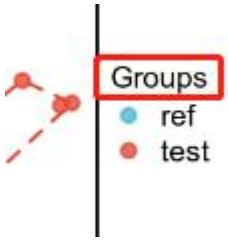
常见问题

1. 如何修改分组名?

答:

可以在 图注 参数中修改【图注标题】、【图注标签】:

- 如修改标题为 Groups:



- 如修改标签为 group1,group2:

- 逗号为英文输入法下的逗号，其他的没办法识别。

2. 能否上传自己的分析数据进行可视化?

答:

自己的差异分析的结果可以上传到表达差异的 PCA 图等模块进行可视化。