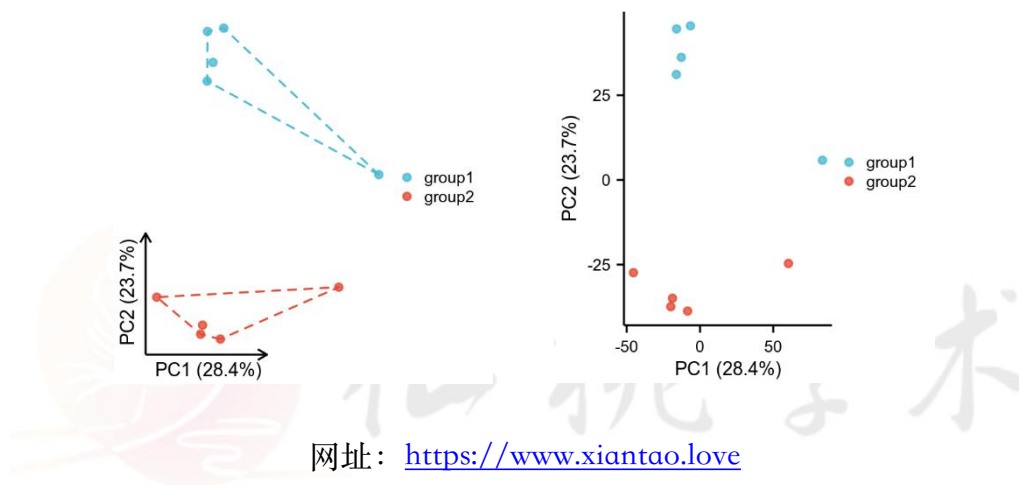


差异表达 - PCA 图



更新时间: 2023.02.28

目录

基本概念	3
应用场景	3
主要结果	4
数据格式	5
参数说明	6
数据处理	6
点	7
外圈	8
标注	9
标题	9
图注(Legend)	10
风格	11
图片	11
结果说明	13
主要结果	13
补充结果	14
方法学	14
如何引用	15
常见问题	16

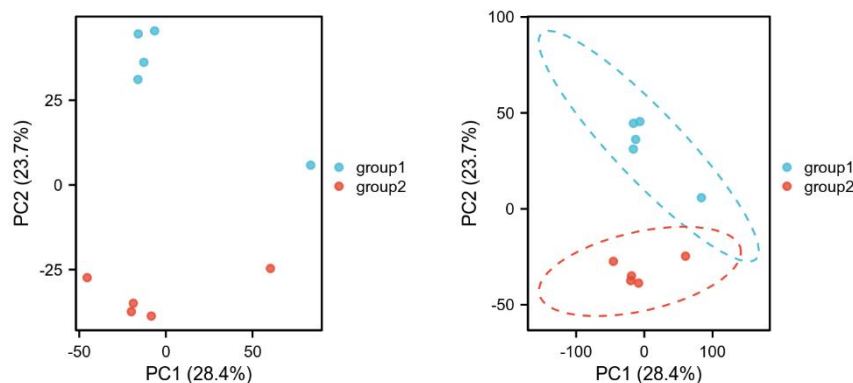
基本概念

- PCA（主成分分析）：**数据降维**的方法。从高维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

- 可以用于查看数据特征情况，具体有但不限于以下场景：
 - 高通量数据中样本之间聚类分布情况。
 - ...

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 28.4%的数据的特征差异，其中图中 (y 轴标题) PC2 能体现 23.7%的数据的特征差异，故整个 PCA 图能体现数据接近一半的差异。
(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组，这部分来自上传数据中的 #注释头部内容，具体可见数据格式说明。
- 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)

数据格式

	A	B	C	D	E	F	G	H
1	#group	group1	group1	group1	group1	group1	group2	group2
2	Gene.Symbol	GSM831759	GSM831760	GSM831761	GSM831762	GSM831763	GSM831846	GSM831847
3	EEF1A1	14.629414	14.6186636	14.6854056	14.5026081	14.5688883	14.6287792	14.6567724
4	RPL41	14.4363619	14.4330003	14.4432824	14.4370109	14.4197727	14.467098	14.4809999
5	TG	14.2929907	14.2520247	14.4814395	13.5969026	14.5549623	14.5664211	14.5841419
6	RPL37A	14.2777219	14.2239444	14.235647	14.2009891	14.2149069	14.2260096	14.2404146
7	RPS4X	14.1429545	14.3008664	14.3029239	14.0351491	14.0905598	14.1927383	14.1735178
8	ND4	14.2252866	14.0511525	14.1893861	14.2098896	14.1550673	14.119611	13.9041902
9	RPS14	13.9267186	13.9305106	14.0226636	13.9913862	14.0405961	14.0795947	14.0226327
10	UBC	14.1675819	14.0454482	14.062307	13.9123953	13.9786911	14.4317457	14.1871612
11	UBB	14.1229222	14.0869451	14.0424096	13.7438065	14.170523	14.1181767	13.9711698
12	HUWE1	14.0853574	14.0918922	13.9576396	14.0948174	14.0615839	13.9418584	14.0415618
13	RPS18	13.8651533	13.9590089	13.7244157	13.7897666	14.0422688	13.8212073	14.0854229
14	RPL39	13.8593892	13.885389	13.8337786	14.0786627	14.1691781	13.9162947	13.7244003
15	COX1	13.8554848	14.4313677	13.6290888	13.4360858	13.6424797	13.7054224	13.4435966
16	RPL23A	13.864491	13.9474918	14.0083676	13.9549066	13.8926654	14.0085861	14.1194867

数据要求：

➤ 头部注释行（以#开头）：

- 用于表征每个样本所属的分组，至少需要提供 1 行样本的注释信息，每行的分组最多是 10 个。注意，注释行不能超过 4 行。

➤ 主体部分：

- 数据至少有 4 列以上，至少需要 5 行数据。
- 主体的第一行为样本编号（如图中的第 2 行），这一行不能含有缺失、重复及特殊字符。
- 主体的第一列为基因名（未必需要提供基因名，只要是能表征样本各个维度的情况即可，因为这里为表达谱数据，所以用的是基因名）。
- 主体的其他部分为样本在各个维度对应的数值，不能含有非数值内容。

➤ 最多支持 600 列，70000 行。文件不能大于 120M，若验证数据时返回报错，需要在上传数据内进行相应的调整，然后再上传数据。

参数说明

(说明：标注了颜色的为常用参数。)

数据处理



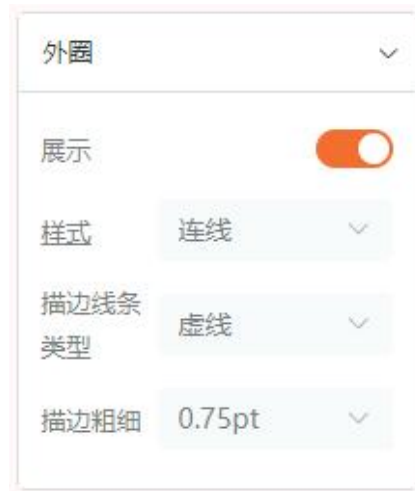
- 转换：对数据进行 log 转换，可以选 无、 \log_2+1 、 \log_2 、 \log_{10} 。
- 归一化：对特征进行归一化可以有效减少特征之间数量级过大的问题，可以选 对行(变量)归一化、无。

点



- **填充色**：点的填充色颜色选项，取决于上传数据中的头部注释行信息，有多少个分组会提取多少个颜色，最多支持修改 10 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，取决于上传数据中的头部注释行信息，有多少个分组会提取多少个颜色，最多支持修改 10 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

标注

类型选择 不标注

特定样本

标注大小 5pt

- 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

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)



- 是否展示: 是否展示图注
- 图注标题: 可以添加图注标题
- 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逗号的形式合并成一个, 类似 A, B
- 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，默认为指向类型。
- 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

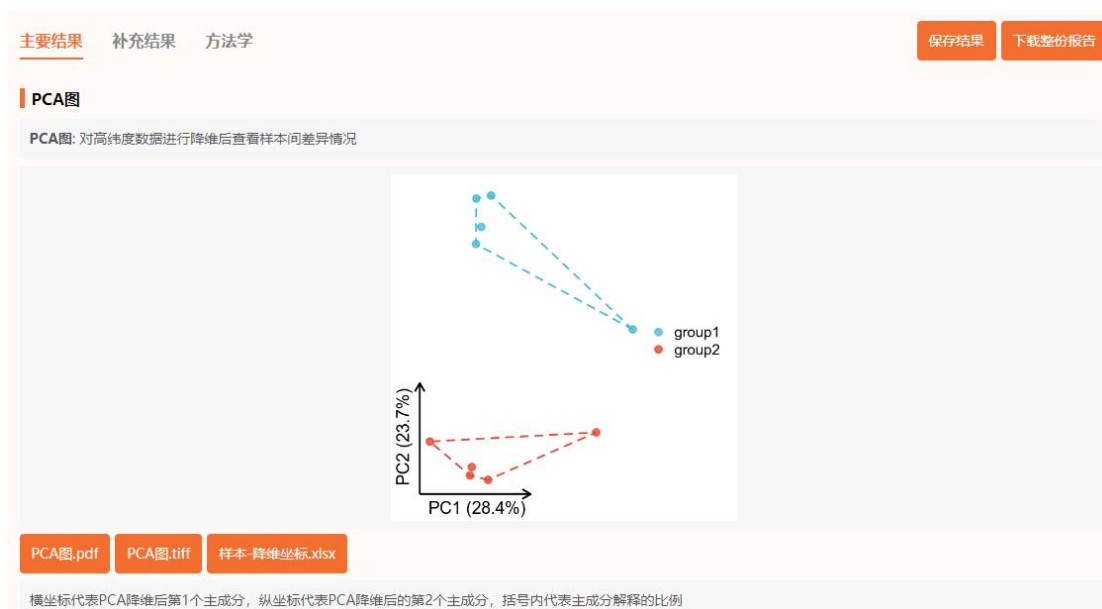


- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体



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。

	A	B	C
1	sample	PC1	PC2
2	GSM831759	-6.494426117	45.47991961
3	GSM831760	-15.96764197	44.58105383
4	GSM831761	-16.12629947	31.12225745
5	GSM831762	83.56836393	5.807391179
6	GSM831763	-12.73430987	36.18489572
7	GSM831846	-8.436819552	-38.72949453
8	GSM831847	-19.94191189	-37.4259903
9	GSM831848	60.34331423	-24.67784692
10	GSM831849	-18.78825249	-34.9586966
11	GSM831850	-45.42201681	-27.38348945

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。

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	28.4	28.4
PC2	23.7	52.1
PC3	13.7	65.8
PC4	9.6	75.4
PC5	7.0	82.4
PC6	6.5	89.0
PC7	5.1	94.1
PC8	3.1	97.2
PC9	2.8	100.0
PC10	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 28.4%，则表示 x 轴的差异可以解释全面分析结果的 28.4%。展示前 10 个成分对应的数据。

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：对数据进行 PCA 分析，分析后结果用 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. RNAseq 数据是否能用 Counts 数据?

答：可以使用。

