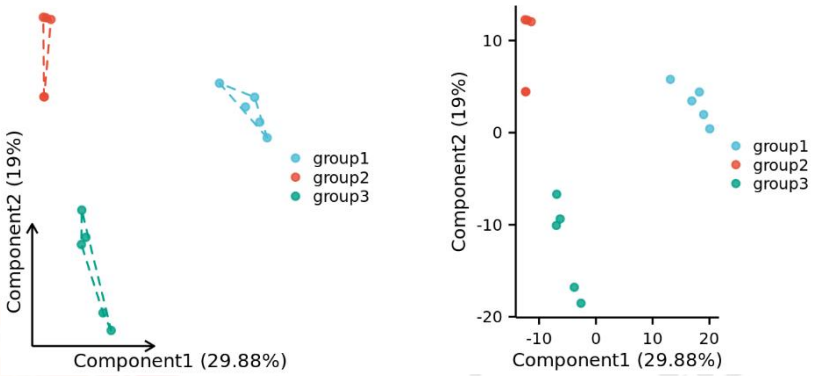
差异表达 - PLS-DA 图



网址：https://www.xiantao.love

更新时间：2023.09.11

目录

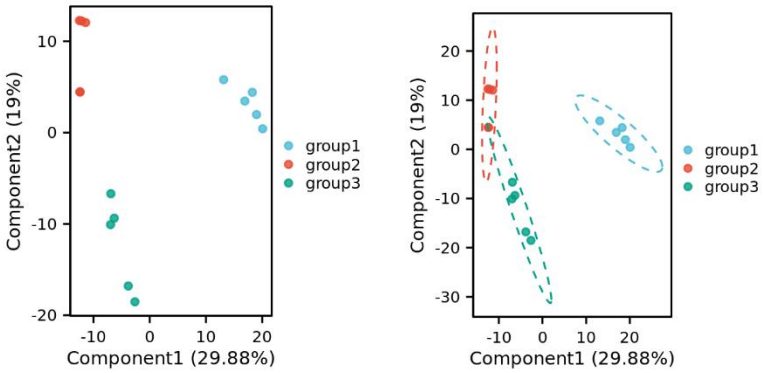
基本概念

- PLS-DA（偏最小二乘法判别分析）：PLS 是“有监督”模式的偏最小二乘法分析，DA 是判别分析，PLS-DA 用偏最小二乘回归的方法，在对数据“降维”的同时，建立了回归模型，并对回归结果进行判别分析。

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：
  - 可用于微生物多样性分析，组间差异较小或者组内样本个数差异较大的情况。
  - …

主要结果



典型的 PLS-DA 图以点图形式展示。

- 横坐标（Component1）表示第一个组分，纵坐标（Component2）表示第二个组分，括号内代表组分解释比例。
- 图中每个点代表每个样本在 Component1 和 Component2 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

#group	group1	group1	group1	group1	group1	group2	group2	group2	group2	group2
OTU_ID	CK1	CK2	CK3	CK4	CK5	T1	T2	T3	T4	T5
1	17	22	0	3	2	23	10	1	8	1
2	143	137	223	186	222	69	56	5	14	9
3	272	214	8	8	6	20	16	80	58	7
4	62	59	6	1	3	0	0	0	0	0
5	13	12	0	0	1	7	7	0	1	0
6	28	34	66	57	86	0	0	0	0	0
7	208	257	122	45	119	45	33	3	2	5
8	73	62	23	4	29	22	20	1	0	1
9	198	210	386	242	322	191	346	214	277	322
10	106	125	203	171	200	142	107	56	70	81
11	1361	1287	280	239	295	245	183	73	99	43
12	122	142	371	507	473	22	14	1	1	1
13	73	58	23	18	29	19	15	5	6	5
14	2	6	59	47	58	0	2	6	1	41
15	59	54	10	7	19	1	3	0	0	0
16	47	59	74	106	110	21	10	3	5	23
17	598	532	127	89	94	44	46	44	42	24
18	121	91	29	15	26	1	0	0	0	2

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，每行的分组最少 2 个最多 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。
  - 主体的第一列为变量名（示例是微生物多样性测序数据中的 OTU id，也可以是其他组学的数据）。
  - 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 500 列，5000 行。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。）

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，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 4 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。
- 线条类型：外圈的线条类型，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。
- 线条粗细：外圈的线条粗细，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。
- 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

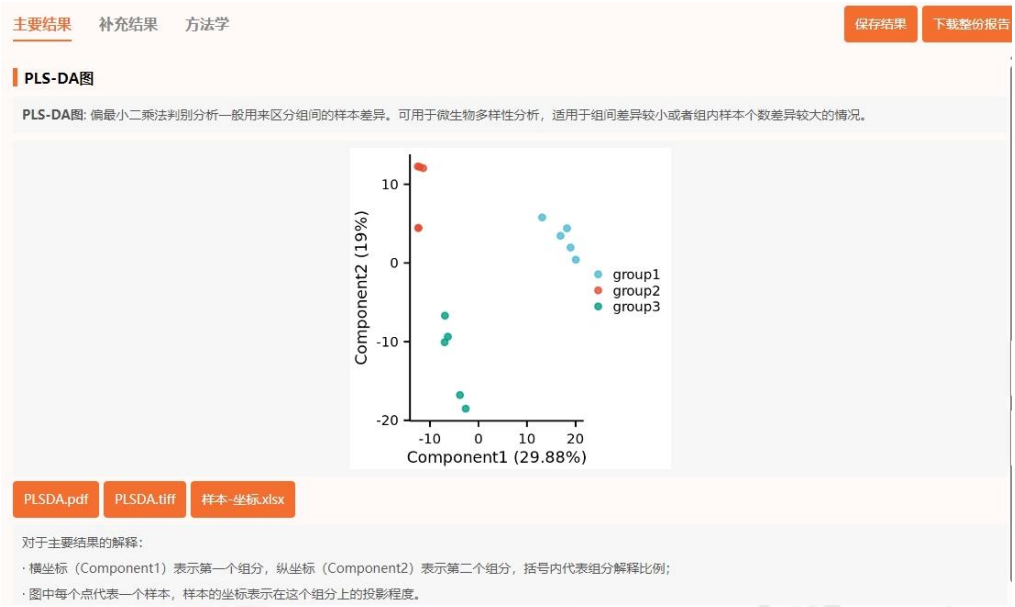
- 坐标样式：无边框的情况下，坐标轴的样式。可选择经典类型、指向类型，默认为经典类型。
- 边框：是否添加外框。
- 网格：是否添加网格。
- 文字大小：针对图中所有文字整体的大小控制，默认为 7pt。

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	comp1	comp2
2	CK1	18.9824593	1.96424719
3	CK2	20.06154	0.41835636
4	CK3	16.8981455	3.44845056
5	CK4	13.1262063	5.78719697
6	CK5	18.2325252	4.41034291
7	T1	-12.3506873	4.45011013
8	T2	-12.4018108	4.4495093
9	T3	-12.0903639	12.2150296
10	T4	-12.4843522	12.2681243
11	T5	-11.3870403	12.0504361

另外，提供各个样本的坐标结果表格 xlsx 下载，含有每个样本对应 Component1 和 Component2 的位置信息补充结果。

补充结果


主要结果 补充结果 方法学 保存结果 下载整份报告

PLS-DA组分

组分对应的解释数据变异情况的比例以及累积比例情况

组分	解释比例(%)	累积比例(%)
comp1	29.883	29.883
comp2	19.002	48.885
comp3	18.543	67.428
comp4	9.6893	77.117
comp5	2.7279	79.845
comp6	3.0175	82.863
comp7	2.5073	85.37
comp8	2.189	87.559
comp9	2.3623	89.921
comp10	2.5366	92.458

注：不满10个样本或去除缺失（和非数值）后的数据不满10个样本的情况，只计算2个组分

此表格为各组分的解释比例和累积比例，如 comp1 的解释比例为 29.883%，则表示 x 轴的差异可以解释全面分析结果的 29.883%。展示前 10 个组分对应的数据。

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）、mixOmics（用于分析）

处理过程：对数据进行 PLS-DA 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. 与 PCA、PCoA、NMDS 等分析的区别?

答：PCA、PCoA、NMDS 都是无监督的降维方法，PLS 是“有监督”模式的偏最小二乘法分析，也就是在分析数据时，已知样本的分组关系，这样可以更好的选择区分各组的特征变量，确定样本之间的关系。DA 是判别分析，PLS-DA 用偏最小二乘回归的方法，在对数据“降维”的同时，建立了回归模型，并对回归结果进行判别分析。

有时我们使用 PCA 或者 PCoA 对微生物群落样本进行分析，得到的结果可能十分混乱，样本分组情况非常不明显，而这时使用 PLS-DA 就会得到比较明显的样本分组聚类结果。

2. 为什么不同于 PCA、PCoA、NMDS 等分析可以一个分组，PLS-DA 至少需要两个分组?

答：这是因为 PLS-DA 是一种有监督学习方法，旨在区分不同组别之间的样本差异。在无监督降维方法（如 PCA、PCoA、NMDS）中，关注的是数据自身的内在结构和变化，对于有监督学习问题，需要根据样本的类别信息进行分类或预测。因此，PLS-DA 需要至少两个分组，以便利用类别标签的信息，通过最大化类别间方差和最小化类别内部方差来找到潜在变量，从而更好地区分不同组别之间的样本差异。

3. 为什么有时候同样的数据和分组在 PLS-DA 得分图上能看到区别，而在 PCA 的得分图上看不到呢?

答：最直观的区别就是 PLS-DA 应用到了标签信息，所以容易找到差异。PCA 和 PLS 的得分图都来自于高维空间向低维空间的投影，但不同点在于，在 PCA 中，投影方向是 X 中方差最大的方向，而 PLS 则是试图将 X 投影到解释 Y 空间方差最大的方向。所以，只有当 X 中方差最大的方向恰好是解释 Y 空间方差最大的

方向时，才能在 PCA 和 PLS 得分图中都能看到差异。而如果两者如果不一致，那么将只能在 PLS 得分图中看到差异。

4. 想要通过 PLS-DA 做差异分析怎么办?

答: mixOmics 包本身并没有提供直接计算 R2、Q2 和 VIP 值的函数，建议用 ropls 这个 R 包做 PLS-DA 分析获取差异分析所用数值。