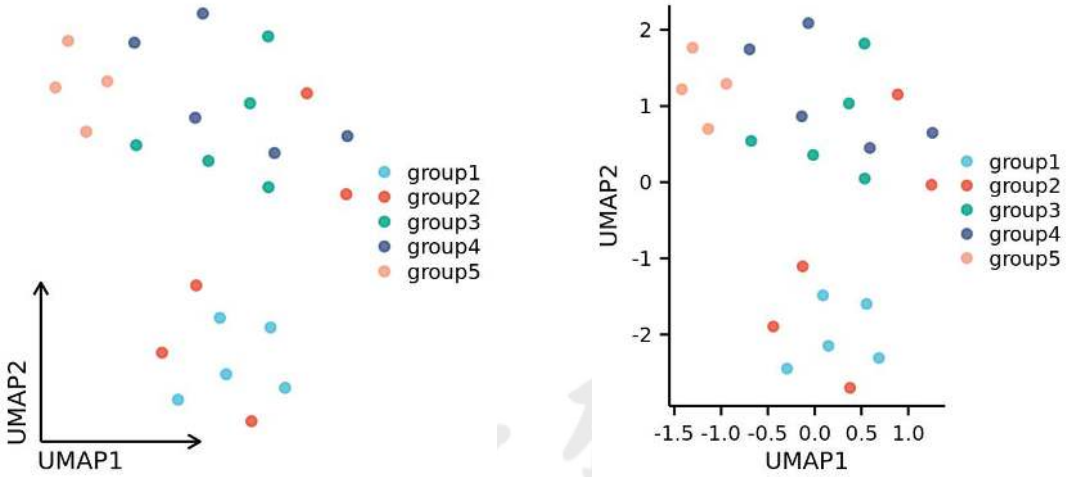
差异表达 - UMAP 图



网址：https://www.xiantao.love

更新时间：2023.09.07

目录

基本概念

- UMAP：一种降维技术，假设数据样本均匀（Uniform）分布在拓扑空间（Manifold）中，可以从这些有限数据样本中近似（Approximation）距离关系映射（Projection）到低维空间。UMAP 图，就是根据 UMAP 降维分析的结果绘制点图。

应用场景

- 将高维的 scRNA-seq 数据降到二维或三维，并可视化单细胞细胞表达和细胞类型之间的关系。
- 高通量数据中样本之间聚类分布情况。
- 其他…

分析过程

上传数据 → 数据处理（清洗） → umap 降维分析 → 可视化

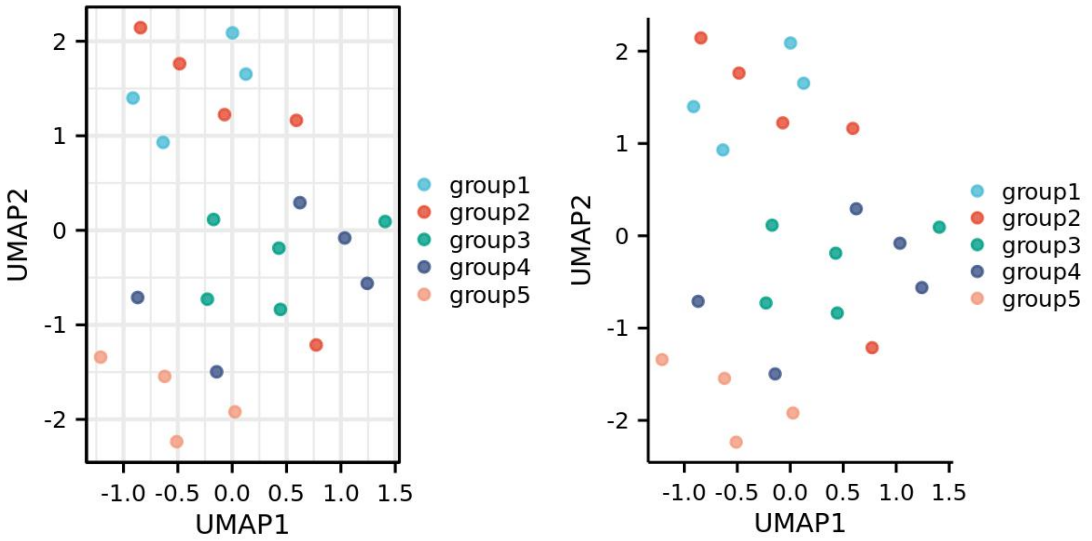
- 数据格式：（具体数据格式要求可以看后面过程的“数据格式”部分）

- 数据第 1 行必须是以#开头的注释信息
- 数据第 2 行必须是与注释信息对应的样本名，必须不能重复或者为空
- 数据的行名不是必须提供的
- 必须提供第 1 行和第 2 行（分组注释信息）；数据至少提供 5 行 4 列（数值部分）

	A	B	C	D	E	F	G	H	I
1	#group	group1	group1	group1	group1	group1	group2	group2	group2
2	Gene.Symbol	GSM831759	GSM831760	GSM831761	GSM831762	GSM831763	GSM831846	GSM831847	GSM83184
3	EEF1A1	392	229	619	421	377	375	419	338
4	RPL41	359	396	349	376	290	353	368	324
5	TG	352	397	267	577	291	365	331	241
6	RPL37A	303	334	313	454	366	360	326	427
7	RPS4X	298	272	228	235	269	314	271	253
8	ND4	290	271	263	356	296	324	347	598
9	RPS14	274	217	200	277	196	181	177	172
10	UBC	268	243	212	279	258	262	321	242
11	UBB	255	243	174	242	278	228	310	222
12	HUWE1	221	393	239	193	228	237	209	265
13	RPS18	221	253	207	145	196	226	184	222
14	RPL39	221	207	321	242	351	343	305	364
15	COX1	219	225	198	328	188	181	201	133
16	RPL23A	217	222	238	171	241	189	216	191

- 数据处理：对每一列数值类型的数据及其他列数据进行相应处理
  - 数值类型数据只能是纯数值类型数据，不能非数值、不规则的值等
  - ……
- 可视化：数据清洗后，使用 umap 包降维分析，再用 ggplot2 包进行可视化

结果解读



典型 UMAP 图以点图形式展示。

- x 轴和 y 轴分别是 umap 降维分析的二维坐标位置，并没有实际的意义，只是提供坐标位置来可视化降维结果。
- 点与点（样本与样本）间的距离能够体现样本相似度，相似度越大的点会聚得越近，相似度低的点会分得更开。
- 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。

数据格式

	A	B	C	D	E	F	G	H	I
1	#group	group1	group1	group1	group1	group1	group2	group2	group2
2	Gene.Symbol	GSM831759	GSM831760	GSM831761	GSM831762	GSM831763	GSM831846	GSM831847	GSM83184
3	EEF1A1	392	229	619	421	377	375	419	338
4	RPL41	359	396	349	376	290	353	368	324
5	TG	352	397	267	577	291	365	331	241
6	RPL37A	303	334	313	454	366	360	326	427
7	RPS4X	298	272	228	235	269	314	271	253
8	ND4	290	271	263	356	296	324	347	598
9	RPS14	274	217	200	277	196	181	177	172
10	UBC	268	243	212	279	258	262	321	242
11	UBB	255	243	174	242	278	228	310	222
12	HUWE1	221	393	239	193	228	237	209	265
13	RPS18	221	253	207	145	196	226	184	222
14	RPL39	221	207	321	242	351	343	305	364
15	COX1	219	225	198	328	188	181	201	133
16	RPL23A	217	222	238	171	241	189	216	191

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。
- 数据最多支持 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

- 转化：对数据进行 log 转换，可以选无、log2+1、log2、log10。
- 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。

分析参数

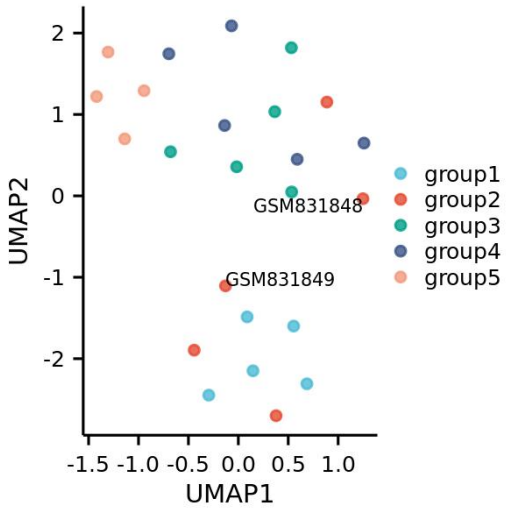
- 最近邻个数：通过设置最近邻的超参数，构造高维数据近似最近邻的数量，有效地控制最终降维结果中局部和全局结构之间的平衡。较小的值能够在降维过程中更关注局部结构，而较大的值会更关注全局结构失去细节，可以通过调试结果选择最适合的值。这里使用 umap 包的默认值 15。注意最近邻个数不能大于样本个数，示例数据中样本数 24 个，则最近邻个数不能调整为 25。
- 距离算法：可选欧氏距离(euclidean，即 N 维空间中两点之间的直线距离)和曼哈顿距离(manhattan，即两个点在标准坐标系上的在坐标轴上投影的距离总和)，默认欧氏距离(euclidean)。
- 种子号：设置种子号可以保证数据降维的坐标位置可重复，默认为 2023，此参数请输入非零整数。调整该参数可改变数据降维的坐标位置。

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。当多选时，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，默认为 1。
- 不透明度：点的明度。0 为完全透明，1 为完全不透明。

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（样本编号）。注意样本编号需要与上传数据的样本信息保持一致！



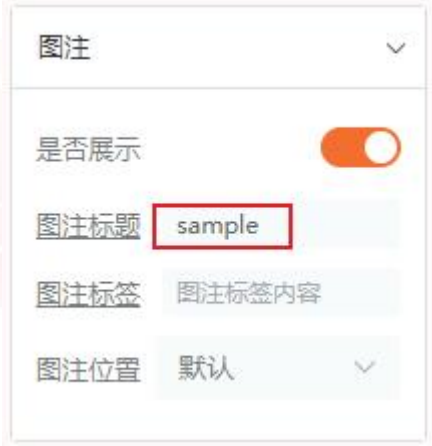
- 标注大小：控制图中标注内容的文字大小，默认为 5pt。

标题文本

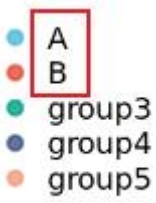
- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本
- 补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如{{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如[[2]]

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，例如：



- 图注标签：可以修改图注中分组标签的名字，如果只修改其中某个，则需要把这些名字以逗号的形式合并成一个，类似 A，B。例如：



- 图注位置：可选择右、上，默认为右

风格

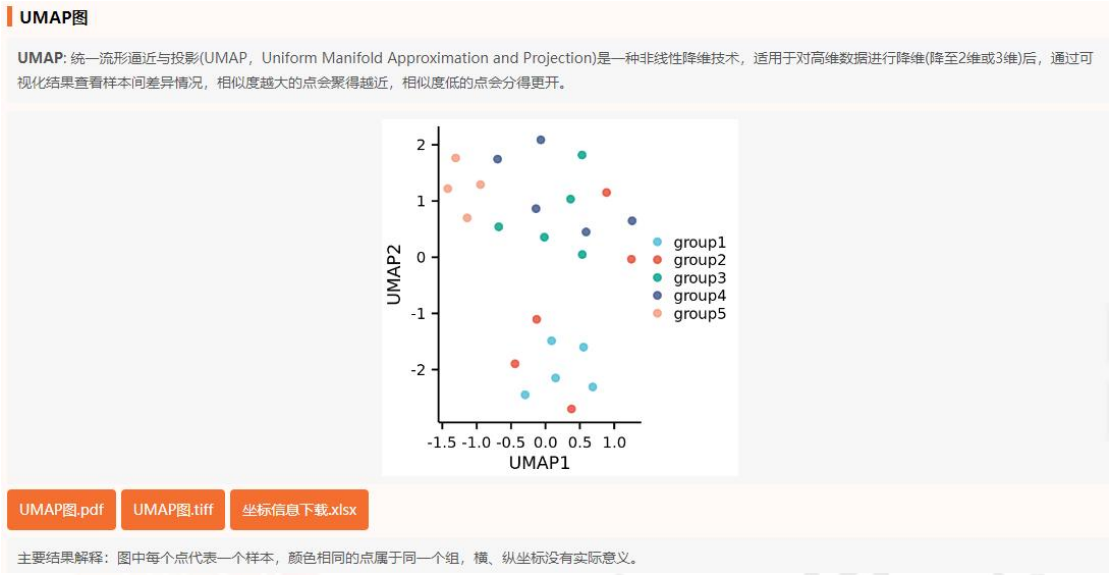
- 坐标样式：不添加边框的情况下，坐标轴的样式。可选择指向类型、经典类型，默认为指向类型。
- 边框：可以选择是否进行添加图形边框的操作
- 网格：可以选择是否进行添加图形内网格的操作
- 文字大小：针对图中所有文字整体的大小控制，默认为 7pt

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	UMAP1	UMAP2
2	GSM831759	0.68715842	-2.30865592
3	GSM831760	-0.29589974	-2.44810169
4	GSM831761	0.55421521	-1.59961765
5	GSM831762	0.14698118	-2.14881772
6	GSM831763	0.08839582	-1.48678247
7	GSM831846	-0.44212745	-1.89543871
8	GSM831847	0.37848135	-2.70063633
9	GSM831848	1.24875779	-0.03502948
10	GSM831849	-0.12828543	-1.10541119
11	GSM831850	0.88724066	1.15213515
12	GSM461850	0.53171838	1.81956619

另外，提供各个样本的降维坐标结果表格 xlsx 下载，含有每个样本对应降维到 2 维坐标的位置信息。

方法学

软件：R（4.2.1)版本

R 包：umap 包（用于分析数据）、ggplot2 包（用于可视化）

处理过程：

(1）将清洗后的数据，通过 umap 包进行 UMAP 分析，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. 可以分析什么数据?

答：可以处理基因表达数据，例如单细胞 RNA 测序（scRNA-seq）数据、批次效应校正后的表达矩阵等

2. 是否需要对数据做归一化?

答：是否需要对数据进行归一化取决于具体的数据类型和分析目的。归一化是将不同尺度或范围的数据映射到统一的区间，以消除数据之间的差异性。例如基因表达数据，通常建议对数据进行归一化，这是因为基因表达量的范围可能会从几个数量级到几万个数量级不等。在这种情况下，如果不进行归一化，则较大数值的特征可能会在降维过程中占据主导地位，而较小数值的特征可能被忽略。

3. 怎么知道最近邻个数是合适的?

可以通过调整最近邻个数，不断观察可视化效果，还可以参考一些文献或经验，常见的最近邻个数取值范围为 5 到 50 之间，但具体取值还需要根据数据集和分析目的进行调整。

4. 距离算法如何选择?

选择距离度量算法时，应根据数据的类型、特性和分析目的来决定。一般来说，在表达谱数据时，计算样本之间的欧氏距离，衡量的是各个样本的表达值在空间

中的差异，而使用曼哈顿距离则不会考虑样本之间的方向，只反映它们的差异大小。