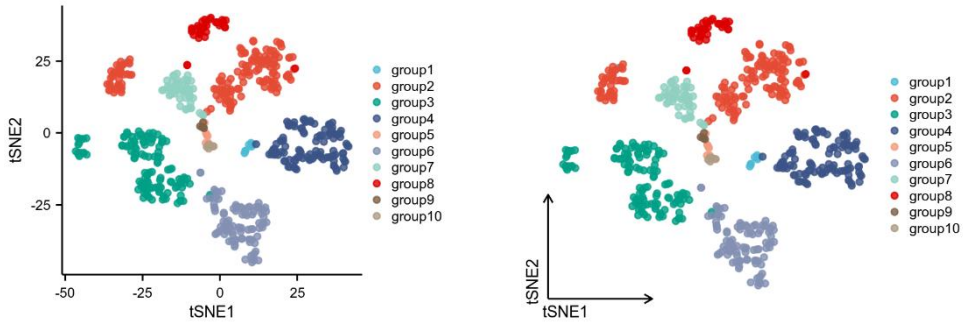
差异表达 - tSNE 图



网址：https://www.xiantao.love

更新时间：2023.08.18

目录

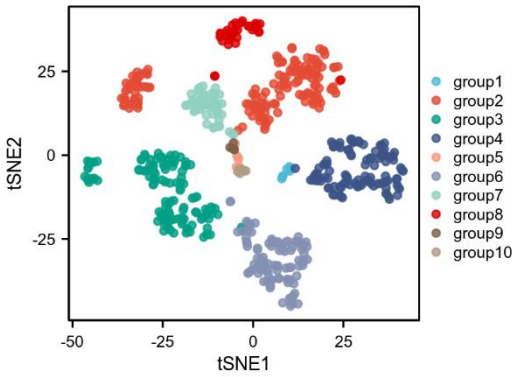
基本概念

- t-SNE 分析，全称为 t-distributed Stochastic Neighbor Embedding，t 分布-随机近邻嵌入，是一种可以把高维数据降到二维或三维的降维技术，适用于高维数据集的可视化。

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：
  - 高通量数据中展示样本的整体分群情况。
  - 在单细胞转录组的数据分析中，t-SNE 应用的更为广泛。
  - …

主要结果



t-SNE 图是以点图形式展示：

- 图中每个点代表一个样本，x 轴和 y 轴分别代表 样本在二维空间的两个主要坐标，但坐标轴的大小没有实际意义。（与 PCA 图不同，该分析无"主成分"的说法）。
- t-SNE 降维后，较大相似度的样本（点），t 分布在低维空间中的距离稍小一点，即相似的样本能够聚集在一起；而对于低相似度的样本（点），t 分布在低维空间中的距离需要更远，即差异大的样本能够有效地分开。但是要注意，t-SNE 中距离本身是没有意义，都是概率分布问题。
- 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	J	K	L	M
1	#group	group1	group1	group1	group1	group1	group1	group1	group1	group1	group2	group2	group2
2	Gene	cell1	cell2	cell3	cell4	cell5	cell6	cell7	cell8	cell9	cell10	cell11	cell12
3	PLN	-0.18545	-0.18545	-0.18545	-0.18545	-0.18545	-0.18545	-0.18545	-0.18545	-0.18545	-0.18545	-0.18545	-0.18545
4	ATP1A2	-0.18567	-0.18567	-0.18567	-0.18567	-0.18567	-0.18567	-0.18567	-0.18567	-0.18567	-0.18567	-0.18567	-0.18567
5	DLK1	-0.20579	-0.20579	-0.20579	-0.20579	-0.20579	-0.20579	-0.20579	-0.20579	-0.20579	-0.20579	-0.20579	-0.20579
6	MYH11	-0.28861	-0.28861	-0.28861	-0.28861	-0.28861	-0.28861	-0.28861	-0.28861	-0.28861	-0.28861	-0.28861	-0.28861
7	CCDC17	-0.1754	-0.1754	-0.1754	-0.1754	-0.1754	-0.1754	-0.1754	-0.1754	-0.1754	-0.1754	-0.1754	-0.1754
8	BEX2	-0.22878	-0.22878	-0.22878	-0.22878	-0.22878	-0.22878	-0.22878	-0.22878	-0.22878	-0.22878	-0.22878	-0.22878
9	PNRC2	-0.22425	1.706128	-0.22425	-0.22425	2.274861	-0.22425	-0.22425	-0.22425	-0.22425	2.164991	-0.22425	-0.22425
10	C1orf194	-0.15879	-0.15879	-0.15879	-0.15879	-0.15879	-0.15879	-0.15879	-0.15879	-0.15879	-0.15879	-0.15879	-0.15879
11	MT1A	-0.26342	-0.26342	-0.26342	-0.26342	-0.26342	-0.26342	-0.26342	-0.26342	-0.26342	-0.26342	-0.26342	-0.26342
12	TFPI2	-0.23622	-0.23622	-0.23622	-0.23622	-0.23622	-0.23622	-0.23622	-0.23622	-0.23622	-0.23622	-0.23622	-0.23622
13	ERP27	-0.24419	-0.24419	-0.24419	-0.24419	-0.24419	-0.24419	-0.24419	-0.24419	-0.24419	-0.24419	-0.24419	-0.24419
14	CD79A	-0.14851	-0.14851	-0.14851	-0.14851	-0.14851	-0.14851	4.062934	-0.14851	-0.14851	-0.14851	-0.14851	-0.14851
15	PRKCH	-0.24751	1.272529	5.596326	3.505189	1.720369	5.203066	1.806454	-0.24751	-0.24751	-0.24751	-0.24751	-0.24751
16	PDK4	-0.42698	-0.42698	1.491204	-0.42698	-0.42698	-0.42698	-0.42698	-0.42698	-0.42698	-0.42698	-0.42698	-0.42698

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 5 列以上，至少需要 6 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，因为这里为单细胞测序数据，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，不允许样本（某一列）数据完全一样。
  - 样本数量与【混乱度参数】有关，样本数不应少于（3 * 混乱度 + 1）的范围，如，样本数仅有 4，混乱度参数应设置为 1。当样本数过少时，应对应调整混乱度参数范围。
- 最多支持 600 列，10000 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

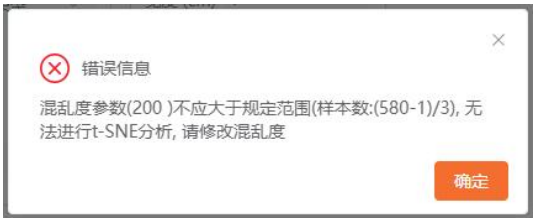
（说明：标注了颜色的为常用参数。）

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。

分析参数

- 种子号：设置种子数可以保证统计检验 p 值结果可重复，默认为 2023，此参数请输入非零整数。
- 混乱度：表示 t-SNE 在优化过程中考虑邻近点的多少，默认 5（因此，要求起码 16 个样本），对于大数据量应该使用较高的混乱度，一般设置 5-50 之间最优。注意，该值不应大于三分之一的样本数，样本数 ≥（3 * 混乱度 + 1），当样本数太少时，应对应调整参数范围。

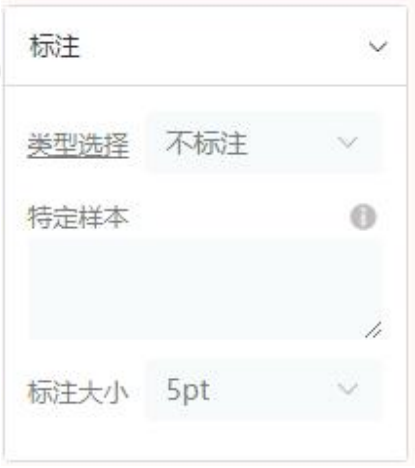


点

- 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。

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

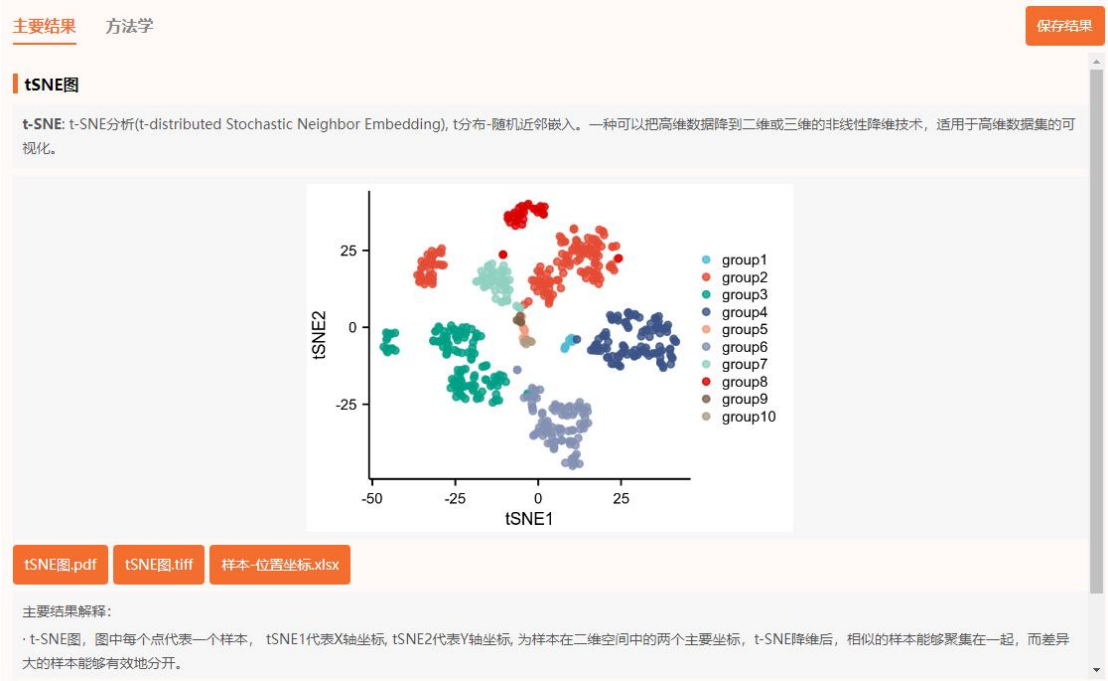
- 坐标样式：无边框的情况下，坐标轴的样式。可选择 指向类型、经典类型，默认为经典类型。
- 边框：是否添加外框
- 网格：是否添加网格
- 文字大小：针对图中所有文字整体的大小控制，默认为 7pt

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。

	A	B	C
1	sample	tSNE1	tSNE2
2	cell1	8.232175814	-6.20236429
3	cell2	10.43569458	-4.111098387
4	cell3	8.202621758	-6.385420656
5	cell4	9.359851758	-4.117550463
6	cell5	7.951275686	-7.033020164
7	cell6	10.16570925	-4.760683054
8	cell7	10.75404948	-3.982512154
9	cell8	-30.318327	24.45528376
10	cell9	10.0601521	-3.464975011
11	cell10	15.24598577	25.12180773

另外，提供各个样本的降维坐标结果表格 xlsx 下载，含有每个样本对应二维空间的 xy 轴位置信息。

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. 与 PCA 有什么区别？一般使用什么样的数据分析？

答：

- PCA 属于线性降维，原理是在多维空间找到两条或者三条最能解释所有点方差的轴，然后将所有点直接投影到这个二维或三维空间。通过高维数据中提取数据的特征向量（主成分，PC 值），并使用解释度的概念来体现数据特征。
- t-SNE 是一种非线性降维算法，原理是将数据点之间的相似度转化为条件概率，原始空间中数据点的相似度由正态分布表示，嵌入空间中数据点的相似度由 t 分布表示。通过原始空间和嵌入空间的联合概率分布的 KL 散度（用于评估两个分布的相似度的指标）来评估嵌入效果的好坏。
  - t-SNE 与 PCA 有很大的区别，在于 PCA 的投影更为直接，而 t-SNE 对点与点的距离进行了 t 分布转化，满足所有点整体上在这个二维或三维空间中的分布情况与多维空间的分布特征一致。
  - t-SNE 图需要从整体上解读所有点的分布特征，而具体比较某两个点之间的距离没有实际意义。在单细胞转录组数据分析中，t-SNE 经常用来展示细胞的整体分群情况，注意，不应该用来具体地比较某几个点的距离！
  - t-SNE 因为随机性强，适用于可视化，不适合添加置信区间。
  - t-SNE 本身受限于参数，不太能用来做分类，只能作为可视化看看样本是不是聚在理想的 cluster。
- PCA 可以使用 counts 或标准化后的数据，tSNE 在单细胞转录组数据分析中一般使用标准化后的数据。