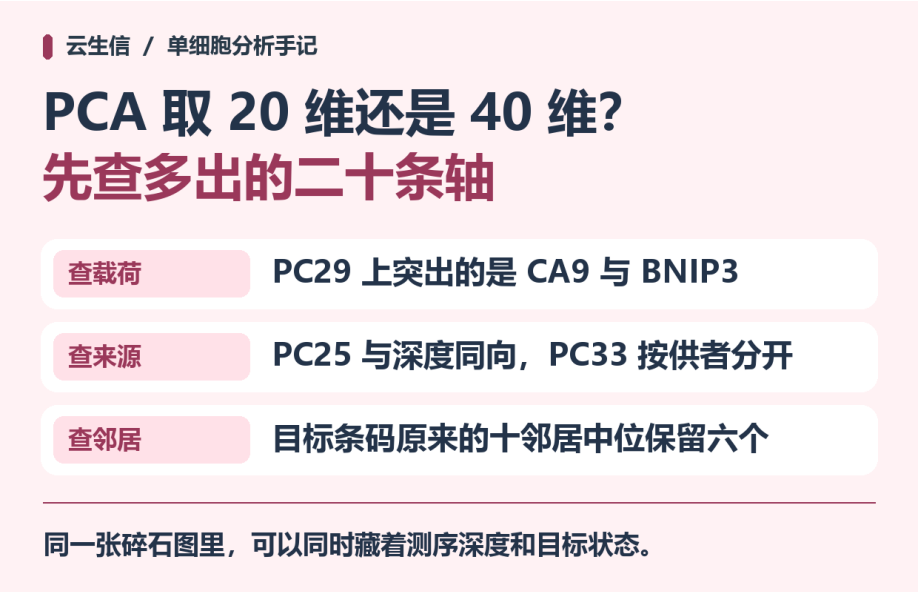
PCA 用前 20 个还是 40 个主成分，怎么知道多出来的维度有什么?

在同一套 PCA 得分上分别使用前二十维和前四十维，从新增轴的载荷、供者来源和深度关系追到目标细胞近邻。



运行教学示例

Python 3.10 或以上，需 numpy、scipy、scikit-learn 和 matplotlib。先在独立环境运行 python -m pip install -r requirements.txt。

```
python run.py --input example --out results
```

在解压后的资料包根目录运行。example 是教学输入，results 保存本次输出，expected 保存可供核对的参考结果。

输入字段

log_expression.tsv 已变换的教学表达表，一行一颗细胞，一列一个基因。构造的矩阵用于揭示指定的轴，不是真实组织数据。

cells.tsv、settings.json 供者、原始总 UMI、独立目标标签以及两档维数、邻居数和每轴展示的基因数。

核对输出，再换自己的数据

axis_audit.tsv、loadings.tsv 每个主成分的方差占比、深度与目标相关和供者解释比例；绝对载荷最高的基因。

cell_neighbors.tsv、edges.tsv、target_summary.tsv 条码的十个邻居重合数、两套逐边名单及各供者目标细胞的本类邻居比例。

component_audit.png 全四十轴的方差曲线和新增二十轴与已知协变量的关系。

用示例验算

1 PC25 的深度秩相关绝对值接近 1，PC29 的两个最大平方载荷各为 0.5，PC33 的供者解释比例约为 0.996。

2 20 个目标细胞，每颗在两个图中有十个邻居，重合数的中位数为六。

3 目标细胞发出的同目标邻接记录由 17/200 到 32/200，第三位供者两次均为 9/50。

换用真实数据

在项目同一表达尺度和特征集上拟合一次至少四十维的 PCA，保存细胞得分和基因载荷。用独立的目标标注与单独保存的原始深度检查新增轴。

两档数字取决于固定特征、归一化、缩放与度量，本例未将每个基因缩放为单位方差。供者关联是描述性统计，不具有因果含义。

方法资料

Bioconductor 主成分选择与方差

Seurat 从主成分到图聚类