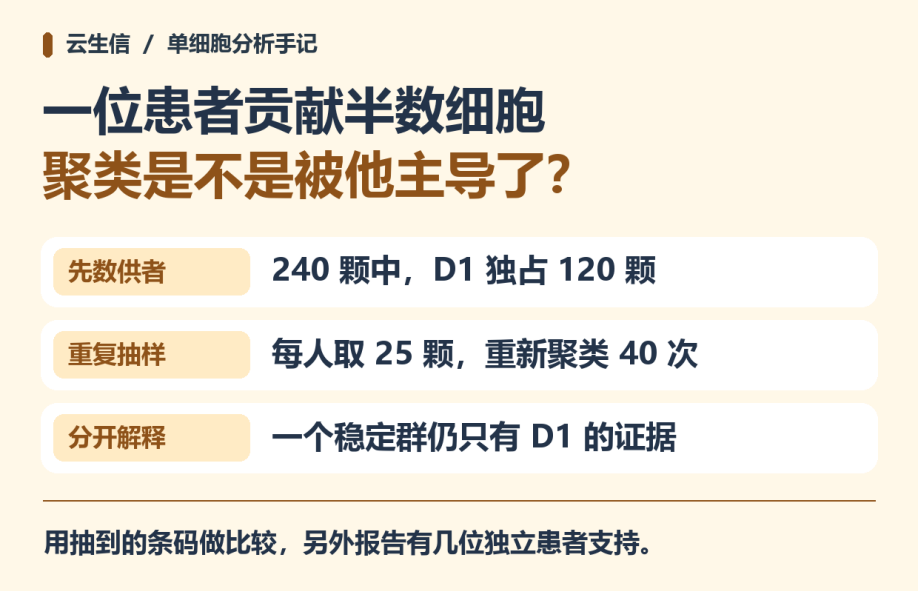
一位患者贡献了半数细胞，怎样看出聚类是否被他主导?

对同一组六维坐标做四十次供者等额抽样，重新聚类、按共同条码匹配分区并核对标记，区分抽样可重复与跨供者重现。



运行教学示例

Python 3.10 或以上，需 numpy、scipy、scikit-learn、matplotlib、threadpoolctl。解压后安装 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

pcs.tsv、cells.tsv、markers.tsv 逐细胞六维坐标、供者以及八个标记的原始计数和全转录组总量。

settings.json 与 --per-donor 控制重复数、配额、聚类个数、随机种子和匹配门槛。另附每人五颗的实跑验算表。

核对输出，再换自己的数据

full_composition.tsv、stability.tsv 原簇供者构成、参考标记、被抽中与达到匹配阈值的次数。

sampled_memberships.tsv、matches.tsv、splits.tsv 每轮抽样条码、原新簇匹配及拆合明细。图为各簇四十轮的 Jaccard。

用示例验算

1 默认每位取 25 颗，共 40 轮，每轮 100 颗；四个簇均通过 40 次。

2 第三群全量 39 颗均来自 D1，反复抽样仍匹配，覆盖的独立供者仍为一位。

3 改为 --per-donor 5，第三群有 33 轮被抽到、31 轮达到匹配阈值；其中 30 轮保留两个相同标记。

换用真实数据

固定输入的特征空间做第一组敏感性检查，以相同条码匹配分区。替换个人项目的真实聚类方法时保持其距离、参数和种子，另记录多供者存在情况。

本例采用 K-means 并固定四群，其目的是检查样本权重对现有空间内划分的影响。不能把多次抽样当成多位患者。

方法资料

Bioconductor 的聚类稳定性与分区比较

scikit-learn 的 K-means 实现