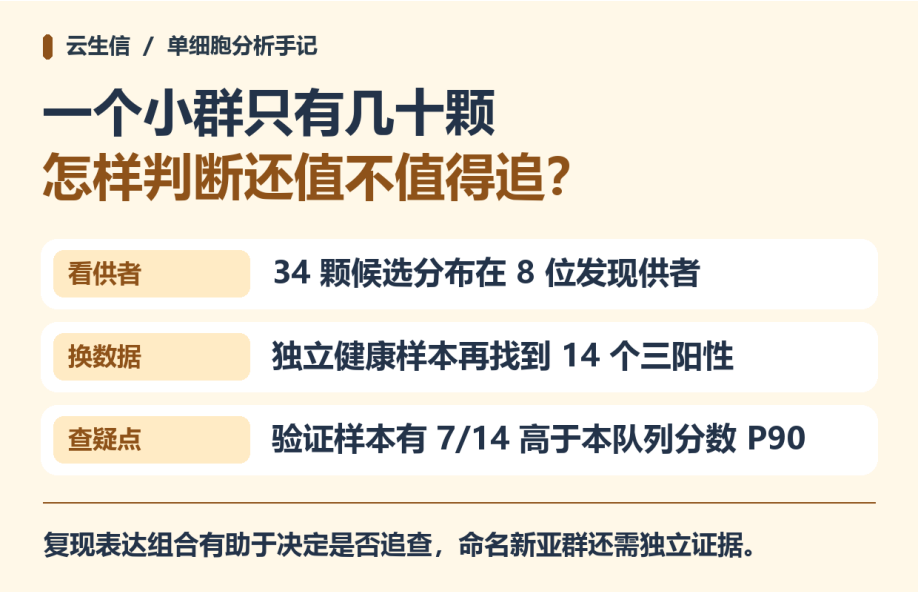
一个新亚群总共只有几十个细胞，凭什么说它值得继续追？

用两个互相独立的公开 PBMC 数据集追查稀有血小板相关候选，逐供者核对三基因计数、双细胞评分与潜在的白细胞混合信号。



运行教学示例

Python 3.10 或以上和 requirements.txt 用于离线核对。重算原始来源与 Scrublet 的版本、命令及校验值见《数据来源》。

```
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

public_cells.tsv、source_hashes.json 两队列 15,015 个真实条码的五项原始标记、供者、原库深度和可用的基因型标注。

kang_scores.tsv、pbmc3k_scores.tsv 分别在两份完整基因矩阵运行 Scrublet 的逐细胞分数，按研究和条码连接。

核对输出，再换自己的数据

candidate_barcodes.tsv、donors.tsv、studies.tsv 所有三标记候选、每位供者的分母、两队列的双细胞复查和混合表达。

threshold_sensitivity.tsv、score_distributions.tsv 每个基因最低计数变化对样本覆盖的影响，另附原始双细胞分布和逐供者图。

用示例验算

1 发现队列 12,315 个条码中 34 个三阳性，分布于 8 位供者；独立队列 2,700 中为 14 个。

2 表达型双细胞分数超过本队列第 90 百分位者分别为 2/34 和 7/14，作为复核首选。

3 每个标记均需至少两个计数时，发现队列剩 7 个、覆盖 5 位，验证队列仍有 11 个。

换用真实数据

先固定候选规则，再对独立研究原样应用，逐供者报告分子分母，连同实验支持与技术疑点回到条码和样本中核查。

三标记定义的是血小板或巨核相关表达组合，黏附、环境 RNA 与双细胞仍可能产生部分信号；独立研究的一个供者不能估计普遍丰度。

方法资料

Kang 等的原始多供者实验

10x 健康供者 PBMC3k