用 Seurat sketch 分析 130 万个小鼠脑细胞，怎样检查有没有漏掉稀有群?

在 1,306,127 个公开小鼠脑细胞上实际运行两套五万细胞的 Seurat sketch 并投回全集，另用真实子集做轻量对照，查清为何抽中更多却可能漏报整群。



运行教学示例

轻量核对仅需 Python、numpy 和 matplotlib。完整复算另需公开 HDF5、h5py、scipy、R、Seurat 5、Matrix；流式处理全集用 BPCells。

```
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

all_marker_audit.tsv.gz 公开原图谱全部 1,306,127 个条码与七个基因的原始计数、文库后缀和等概率子集标记。

子集投影及 full_projections/ 全集投影 四份实跑的逐条码 Seurat 投影，附运行环境、提取及重跑代码。全集两份各含 1,306,127 行。

核对输出，再换自己的数据

full_atlas.tsv、capture_suffix.tsv 全集三标记同阳性分母、拆分文库后缀的计数，以及真实子集覆盖。

method_comparison.tsv、full_method_comparison.tsv 等 分别给出子集与全集的直接命中、未抽中条码的找回及多簇敏感性统计，另附每簇的分子分母。

用示例验算

1 全集有 8,559 个 C1qa 阳性，C1qa、C1qb、Tyrobp 三者均检出为 6,348 个。真实子集 16,000 颗中有 70 个候选。

2 在全集各抽五万颗，直接命中数为 247 与 2,944；在真实子集各抽 1,500 颗时为 4 与 40。

3 全集杠杆路线的最纯簇仅找回 18/3,404。把训练阳性至少十颗且过半的两簇合计，找回 3,148/3,404。

换用真实数据

先用独立面板标记一批候选并记录分母，再同预算比较抽样方法。只用抽中的条码指定要追查的群，把投影评价限定到没被抽中的细胞。

这是对原始标记规则的覆盖核查，不能直接等同巨噬细胞鉴定准确率。多簇规则属于解释本例的敏感性分析，需独立数据复核。

方法资料

GSE93421 原始小鼠脑数据

Seurat 百万级 sketch 教程