10x 交付里有 raw 和 filtered 两个矩阵，分别能回答什么问题?

对齐 10x 的未过滤与已过滤稀疏矩阵，从原始 UMI 排名追到被保留的条码，再导出低计数背景候选和需要复查的高计数条码。



运行教学示例

Python 3.10 或以上，使用 numpy、scipy 和 matplotlib。解压后可用 pip 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

raw/ 与 filtered/ 各含标准 10x 的 matrix.mtx、features.tsv、barcodes.tsv，可使用 .gz。

settings.json background_max_umi 是教学中挑选低计数候选的上限，只针对本次文库。

核对输出，再换自己的数据

rank.tsv / rank.png 在未过滤计数上排名，并标记原有细胞识别保留的条码。

background.tsv / review.tsv 分别列出低计数候选和未保留但超过该区间的条码。

background_profile.tsv 对候选背景液滴求和后的基因分子数及相对组成。

用示例验算

1 8 个有计数的原始条码，filtered 保留 3 个。

2 上限为 5 时，教学背景有 3 个条码，总 UMI 为 9。

3 D4 有 16 个 UMI，未进 filtered，进入复查清单。

换用真实数据

使用同一份文库和同一次上游运行的两套输出。先只提取 Gene Expression 特征，再按 ID 和条码对齐已有过滤矩阵。用原始条码总量、液滴排名曲线和实验背景共同选择候选背景范围，适当改动阈值观察背景谱是否稳定。

此练习重建排名并复核原先的识别结果，未把每个未过滤条码都判为空液滴。需要正式进行空液滴统计时可使用 DropletUtils 的方法和多重检验结果。

方法资料

Cell Ranger MEX 矩阵定义

DropletUtils 实例