把同一份文库的 reads 减半，哪些细胞类型和标记基因最先消失?

从两份配对 FASTQ、微型参考序列和条码表出发，按对抽样、定位并去重，分开统计消失的条码与保留下来却读不到的标记。



运行教学示例

Python 3.10 或以上，先安装 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

R1.fastq.gz、R2.fastq.gz、transcripts.fa 完整的教学原始读段、四条虚拟转录本和精确匹配参考。

barcodes.tsv、make_example.py 六十个预期条码及其已知类型，可改低表达与抽样条件。

核对输出，再换自己的数据

read_depth.tsv、type_summary.tsv 原始读对、保留读对、去重分子与按类型的两类损失。

barcode_audit.tsv、retained_barcodes.png 每个条码的全量与降深计数、标记与入选状态。

用示例验算

1 全量 4,286 对读段产生 2,551 个不同基因-条码-UMI。

2 种子 98 保留 2,178 对，得到 1,683 个分子。

3 种子 98 有一颗保留的 T 条码丢失原有标记；稀有组五颗原入选条码均掉出。

换用真实数据

对真实文库按同一决策抽读对，固定参考、化学版本、比对与细胞调用，保留全量和降深两张条码名单。

教学计数器使用无错短序列的唯一精确匹配和固定 UMI 线，不代替真实参考上的比对、纠错和细胞调用软件。

方法资料

10x 对原始 FASTQ 输入的要求

seqtk 作者的成对抽样用法