合并两个 10x 样本后出现一串同名细胞，这些条码该怎么处理?

把文库号与细胞条码当作两列来连接表达与注释，找出跨库同名、同库重复、缺少注释和多余记录，并检查连接前后的计数。



运行教学示例

Python 3.10 或以上，只使用标准库。

```
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

cells.tsv 每行一个实际细胞，含 library、barcode、total_umi 与 marker_umi。

labels.tsv 每行一个文库内条码的注释，含 library、barcode、celltype。

核对输出，再换自己的数据

annotated.tsv 保持表达表原顺序和所有条码的带注释结果。

unmatched.tsv 分开记下无注释细胞及只在注释表出现的条码。

audit.json 单条码连接的假想行数、双键连接的实际行数与计数守恒检查。

用示例验算

1 8 个输入细胞连接以后仍为 8 个。

2 单用条码左连接会得到 12 行，双键连接为 8 行。

3 L2 / BB 在表达表中缺少注释，L1 / ZZ 仅在注释表中。

换用真实数据

从独立捕获的文库保留来源键。不要直接删除条码末尾的 GEM 组信息。若有合并工具创建的前缀，从原始映射表恢复文库和条码，再用双列键连接。

本例的数据粒度是每次捕获中的细胞。若注释文件为每细胞多条记录，先定义要合并的是多套注释还是多个标签版本；多条 V(D)J contig 等一对多信息应另建表。

方法资料

10x 输出矩阵的条码文件

Seurat 对象与细胞名称