从 CELLxGENE 拼肺组织图谱，怎样发现被收录两次的细胞?

下载两个真实肺图谱集合的逐细胞元数据，按主记录字段核算筛选前后的数量，追溯混有主记录的亚群文件和跨技术重复使用的供者。



运行教学示例

离线核对需 Python 3.10 或以上和 matplotlib。完整重取公开元数据的脚本另需 Linux 和 tiledbsoma，详见《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

records.tsv.gz、datasets.tsv 两个肺图谱集合实际的逐条目主记录标记、供者、技术，以及研究标题和 DOI。

query.json、release_lung_summary.tsv 固定版本、查询列、组织筛选、数据集条件与该版本全肺总表。

核对输出，再换自己的数据

dataset_counts.tsv、selection_ids.tsv 逐数据集的保留/排除数，每个 Census 行号的筛选决定。

donor_sources.tsv、donor_review.tsv、scope_counts.tsv 追溯同一研究的供者标记与技术来源，分开报告两个集合和整个发布版。

用示例验算

1 两个集合共 222,418 行，保留主记录 141,592 行，排除 80,826 行。

2 胎肺成纤维与平滑肌文件中 39,028 条为非主记录，但该文件也有 176 条主记录。

3 成人肺同一研究的供者 2 同时出现在两种技术；不按供者编号删除细胞。

换用真实数据

从细胞表先查看主记录标记，再按研究、供者、技术追溯重复来源。保存版本、表达式与原始行号；个体分析需另整理供者的真实对应。

Census 的行号标识一个版本中的观察记录。两个集合内的非主记录并不能直接逐一对应到本切片某一行；精确配对需原始条码、文库和来源。

方法资料

Census 肺组织官方实例

Census 字段与数据结构