从 GEO 下载了 12 个单细胞样本，怎么查清它们来自几位患者?

把 GEO 文库记录、论文样本表和文件名单对齐，做出按患者、部位与时间点可追溯的清单，再数清一次比较真正有几个人。



运行教学示例

Python 3.10 或以上，使用标准库，无需安装额外软件包。

```
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

libraries.tsv 每行一个文库，保存 GEO 样本键、文库键和标本键。

specimens.tsv 每行一个实际取材标本，记录供者、器官与时间点。

files.tsv 一行一个输入文件，记录所属文库。允许同一文库对应多个文件。

核对输出，再换自己的数据

manifest.tsv 逐文库列出供者、标本、时间、文件数与核对状态。

donors.tsv 逐供者清点可用标本，标记基线与随访是否齐全。

issues.tsv / summary.json 列出缺失映射与没有文件的文库，并给出可复核的计数。

用示例验算

1 12 份模拟文库中，11 份能追溯到标本和供者。

2 可确认的独立供者为 5 位，只有 P01 与 P04 有可用的前后配对。

3 L12 的标本待查，L10 有来源但没有列入输入文件。

换用真实数据

从 GEO 样本记录中提取作者使用的文库名，再从论文补充表恢复同一人的不同标本。把每个文件分配给真正生成它的文库。保留原始拼写与来源文件，确认映射后再填统一编码。真实数据里的患者键请使用研究内去标识编号。

缺失患者字段的文库无法用文件名补成一个新供者。练习按 baseline 与 followup 识别时间配对；真实研究还应检查器官、条件和实验设计是否允许比较。

方法资料

NCBI GEO 数据结构

NCBI GEO 数据下载说明