Ensembl ID 换成基因名后怎么少了几百行?

把基因原始 ID、去版本后的查找键和基因符号分开保存，区分无法映射、一对多、同名与同一稳定 ID 的多个版本，并为每种处置验算分子总数。



运行教学示例

Python 3.10 或以上，标准库即可运行。

```
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

counts.tsv 简化的原始计数表，一行一个带版本的基因 ID，其余列为样本或条码。

mapping.tsv 本次参考注释中的稳定 ID 与 symbol 对应关系，允许多条候选。

核对输出，再换自己的数据

mapping_audit.tsv 为每个原始计数行记录查找键、候选符号及歧义类型。

safe_counts.tsv / review_counts.tsv 分别输出唯一无冲突映射与待查的原始行。

mass_balance.tsv 按输入列复算保留与待查计数之和。

用示例验算

1 7 个示例基因中，2 行可安全使用唯一符号。

2 两个不同 ID 都映射到 SHARED，两个版本行使用同一查找键，另一个 ID 对到两个符号。

3 示例 S1 的总计数为 39，已确认行 12，待查行 27。

换用真实数据

换用本物种、同一基因组与注释释放版本对应的基因映射。保留原始行名，不接受跨物种表直接匹配。对大稀疏矩阵可先只审查行名，再用稳定基因 ID 将结果连接回行元数据。

示例使用 Ensembl 形式的人工 ID 与教学符号。安全策略将候选冲突隔离，待查行在独立表中保留全部计数。把同名实体合并需要与特征定义和下游问题相符的依据。

方法资料

Ensembl 稳定 ID 与版本规则

10x 特征矩阵中 ID 与名称的格式