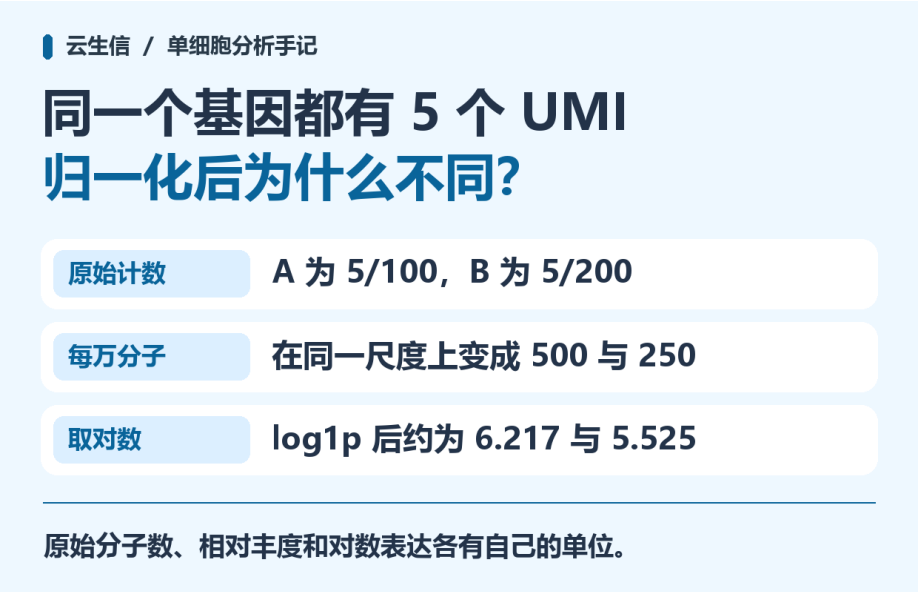
明明是同一个细胞，原始计数和归一化表达为什么差这么多?

从两个细胞各有五个原始 UMI 的例子出发，逐步计算文库大小归一化、自然对数和反向验算，再用同一张矩阵检查先平均还是先取对数为何不同。



运行教学示例

Python 3.10 或以上，需 numpy 和 scipy。运行前用 pip 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

counts/ 原始、非负整数的 Gene Expression MEX，包含特征表、条码表和稀疏计数。

settings.json、inspect.tsv 归一化目标总量、均值示例基因，以及希望逐项手算的条码与基因 ID。

核对输出，再换自己的数据

handcheck.tsv、cell_depth.tsv、means.tsv 手算表、归一化前后的列和及平均与取对数先后顺序的比较。

normalized.mtx、log1p.mtx、轴文件 两份只含有效条码的稀疏结果及各自的基因和细胞顺序。

用示例验算

1 A 与 B 的 G 都有 5 个 UMI，但深度分别为 100、200，归一化后变成 500、250。

2 A 的 R 为 40、B 的 R 为 80；两者各占本细胞的 40%，归一化和对数表达相同。

3 三个有效细胞的归一化列和均为一万；零深度条码 Z 单列，未生成表达值。

换用真实数据

从自己的 RNA 原始计数层导出 MEX，确认计算分母包含哪些特征，选择实际对象中同一个条码、基因和归一化比例，逐项核对表达层。

本练习仅实现固定目标总量再取自然对数这一种方法。校正残差、按行标准化与其他模型得到的值有不同定义，不可套用这里的反算公式。

方法资料

Seurat NormalizeData 的 LogNormalize 定义

Scanpy 按总量归一化说明