同一种细胞在 10x 和 Smart-seq2 里检出率差很多，两批数据怎么比?

拿两份不同计数单位的表达矩阵，只在共同基因、相同供者和可信细胞大类中核对读深、检测比例和平台内部的标记排序。



运行教学示例

Python 3.10 或以上，需 numpy、scipy 和 matplotlib。先在解压目录运行 python -m pip install -r 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

10x/、Smart-seq2/、assays.tsv 两套有单独基因表的原始计数，使用 UMI 和分配到基因的读段两类单位。

cells.tsv、settings.json 每颗细胞的供者、细胞大类、比对读段数及要比较的类别和标记。

核对输出，再换自己的数据

feature_coverage.tsv、cell_depths.tsv 基因可测集合、原计数单位、比对读段及共同集合检出数。

detections.tsv、paired_detection.tsv、marker_ranks.tsv 各供者每个大类的检出比例、对齐后的平台差、各平台内的标记效应排序。另附配对图。

用示例验算

1 两份矩阵各 38 个特征，交集 37 个；X_ONLY 和 PLATE_ONLY 记录为各自平台特有。

2 36 颗 T 细胞中 LOW_MARK 的检出分别为 6/36 和 22/36；D1、D2、D3 的检出率差依次为 5/12、5/12、6/12。

3 两个平台上 T 对 B 的前三个标记均为 CD3D、TRAC、IL7R，排序基于各自原单位归一化后的结果。

换用真实数据

确认两端的物种、基因 ID 和计数类型，明确总比对读数从哪个样本级或细胞级文件获取。限制到相同供者来源与可信大类，在共同基因上报告缺失情况。

两个平台采集的是同一位供者的不同细胞。检出率会受读深、建库、过滤和组成影响，演示数值不能用于推断平台普遍优劣。

方法资料

Ding 等比较多种单细胞测序方法的原始研究

Smart-seq2 原始方案

Cell Ranger 的基因表达计数