转录组里明明有 T 细胞，为什么有些条码在 V(D)J 表里找不到?

用两份会复用短条码的模拟文库，重现 V(D)J 结果全都对不上和对得太多两种错误。随后沿着原始 contig、过滤表、生产性链和 αβ 配对逐层统计检出率。



运行教学示例

Python 3.10 或以上，pandas、numpy 和 matplotlib，见 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

gex_cells.tsv、all_contigs.tsv、filtered_contigs.tsv 两个模拟捕获文库的 240 个 RNA 条码，以及过滤前后的 TCR contig。

make_example.py 固定随机种子及可编辑的生成规则，含一个不在 RNA 细胞中的原始 contig。

核对输出，再换自己的数据

barcode_funnel.tsv、stratified.tsv、stage_totals.tsv 每个条码和每个文库类型从原始 contig 到生产性双链的追踪。

join_audit.tsv、false_matches_without_library.tsv、detection_funnel.png 三种连接方法、41 个串库错配条码及分层检出率图。

用示例验算

1 直接凭完整编号得到 0，去前缀忽略文库得到 150，使用文库和条码得到 109 个过滤表匹配。

2 RNA 中有 133 个 T 细胞，117 个有任一原始 contig、107 个进入过滤表、105 个至少有一条生产性 α/β 链，59 个两条均有。

3 模拟 NK 样条码中 2 个有生产性链但无双链；需要核对表达标签和技术因素。

换用真实数据

从测序交付单建立 GEX 与同一捕获的 V(D)J 文库对照；检查后缀、原过滤规则和物种链类型，再输出 RNA 细胞内的条件检出率。

这里只模拟了教学所需的几列和过滤过程，各版本 Cell Ranger 的真实过滤规则以原输出为准；有双链也不等于配对唯一或抗原特异性已知。

方法资料

10x Cell Ranger V(D)J 注释字段

AIRR 受体重排字段定义