转基因实验做完了，为什么表达矩阵里连 GFP 这一行都没有?

从转基因报告序列出发，同步核对基因组 FASTA、GTF 外显子、链方向、转录本锚序列与最终特征列表，定位 GFP 缺行发生在哪一步。



运行教学示例

Python 3.10 或以上，只需标准库。可在普通电脑上直接运行两个教学例子。

```
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

reference.fa、genes.gtf、features.tsv 对应同一参考和运行的 FASTA、外显子注释及表达矩阵特征轴。均支持 .gz。

reporter.json 目标 gene_id、contig 名和按成熟转录本方向填写的唯一性候选锚序列。

核对输出，再换自己的数据

transcripts.tsv 报告每条目标转录本的链方向、外显子数、拼接长度与锚序列位置。

audit.json 汇总 contig 长度、目标转录本数、矩阵行状态和需要追查的具体问题。

用示例验算

1 正例中的 REP_TX1 有两个外显子，拼接后 26 个碱基，指定锚序列在该转录本中仅出现一次。

2 正例的特征表存在 REPORTER_GFP，audit.json 的 flags 为空。

3 用 broken 目录再跑，得到 check_annotation 和 matrix_row_absent 两个标记。

换用真实数据

用实际载体的转录方向和末端结构挑选锚序列，替换为真实参考序列、GTF 和目标样本的特征表。返回该次比对记录查看锚附近覆盖，并核查相似序列可能造成的多重比对。

教学程序核验指定转录本的结构、局部序列和矩阵入口，不能证明在宿主全基因组唯一、实际表达或测序足够。建立正式 Cell Ranger 索引和重新计数需要相应软件、完整参考与 FASTQ。

方法资料

Cell Ranger 自定义基因表达参考

10x 的特征条码矩阵