肿瘤上皮细胞里有哪些成套变化的基因？用 cNMF 拆开表达程序

把一个上皮细胞当成多种表达程序的组合，在三种 K 下各做十二次真实 cNMF 拟合，通过稳定性、重建误差、标记和逐供者使用度判断应解释哪些程序。



运行教学示例

复核用 Python 3.10 或以上与 requirements.txt。重算官方 cNMF 建议另建 Python 3.12 环境，按 model-requirements.txt 安装。

```
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

epithelial_counts.tsv、cells.tsv 360 颗教学上皮细胞的 84 个原始基因计数、四位供者及用于事后核验的模拟组成。

fit_cnmf.py、example/fit/ cNMF 1.7.1 的初始分解、共识代码及三种 K 的程序、使用度和模型指标。

核对输出，再换自己的数据

k_summary.tsv、program_identity.tsv 重建误差与稳定性，前三个程序的标记及其和模拟使用度的相关。

donor_usage.tsv、k4_marker_coverage.tsv、stability_donors.png 各供者程序使用度、第四成分的标记审查及比较图。

用示例验算

1 K=2、3、4 的重建误差约 9,025、5,076、4,724；K=4 的稳定性明显下降。

2 K=3 的各分量排名前八的基因均命中相应预设面板八项，相关均大于 0.92。

3 K=4 的第 4 分量前八项均为模拟背景；K=3 的三种优势程序在四位供者中都有。

换用真实数据

对已确认上皮谱系的原始非负表达，在一组合理 K 下重复拟合，再从稳定性、误差、基因和供者四方面决定程序是否值得解释。

本例为教学模拟，知道生成权重。真实肿瘤中发现的程序需独立样本或实验验证，单个成分不能证明恶性身份或调控机制。

方法资料

Kotliar 等 cNMF 原始论文

作者代码与运行示例