GenomeScope基因组特征估计：k-mer计数与拟合验收

以短读长WGS为例，先用Meryl计数并检查频数直方图，再用GenomeScope拟合；区分工具输出、模型假设和可解释结论。

GenomeScope根据短读段中的k-mer频数分布估计基因组大小、杂合度和重复比例。结果依赖输入质量、计数方式、k值和倍性假设；不能把一个模型点估计直接当作真实基因组大小。以下以单个二倍体短读长WGS样本为例，把k-mer计数与模型拟合拆开留档。

阶段	固定内容	检查结果
计数	同一读段集合、k、Meryl版本	日志、资源峰值、计数库
直方图	每个频次对应的k-mer数	两列整数、频次递增
拟合	GenomeScope版本、k、倍性	日志、曲线、参数与不确定性

先检查输入并完成计数

示例要求已完成接头处理和基本质控的双端FASTQ，并使用Meryl与GNU Time。k=21、4线程和16 GB内存是演示参数，需根据读长、覆盖和服务器额度调整；Meryl的memory参数是计数阶段上限，不代表全流程总内存承诺。先在空目录运行一份材料，保留原始计数库。

```
set -euo pipefail
r1=/data/material_R1.fastq.gz
r2=/data/material_R2.fastq.gz
test -s "$r1"; test -s "$r2"
gzip -t "$r1" "$r2"
mkdir kmer-profile-001
cd kmer-profile-001
meryl --version > meryl-version.txt 2>&1
/usr/bin/time -v -o count.resources.txt \
  meryl k=21 threads=4 memory=16 count "$r1" "$r2" output reads.meryl \
  > count.log 2>&1
test -d reads.meryl
meryl histogram reads.meryl > reads.hist 2> histogram.log
test -s reads.hist
```

核对直方图两列含义

Meryl直方图每行是出现频次和具有该频次的不同k-mer数量。下面的标准库脚本检查两列整数、频次递增并汇总不同k-mer数与观测次数；它只检查文件结构，不判断峰是否生物学合理，也不能证明输入身份。

```
rows = distinct = observations = previous = 0
with open("reads.hist") as handle:
    for line in handle:
        if not line.strip(): continue
        fields = line.split()
        if len(fields) != 2 or not all(x.isdigit() for x in fields):
            raise ValueError("expected two integer columns")
        multiplicity, count = map(int, fields)
        if multiplicity <= previous: raise ValueError("frequency not increasing")
        previous = multiplicity; rows += 1; distinct += count
        observations += multiplicity * count
if not rows or not distinct: raise ValueError("empty histogram")
print("bins", rows, "distinct", distinct, "observations", observations)
```

第二列求和是不同k-mer总数；频次乘以该行数量后再求和，才是按直方图计算的总出现次数。不要把两者混为一谈。canonical计数、读段过滤和工具版本也会影响分布，应在分析记录中写明并保持批次一致。若要比较不同k值，应从同一FASTQ重新计数并分别拟合，不能把不同k生成的直方图混在一次模型运行中。

再拟合并判断模型是否适用

只有输入与直方图检查通过后才运行GenomeScope。若样本测序深度不足或质量报告显示明显污染、GC偏倚，先判断是否满足模型输入前提，再决定是否继续估计。二倍体设置应来自样本生物学；倍性未知时先澄清假设，不要为了接近期望结果反复改ploidy或截断低频区。输出目录生成成功只说明程序完成，仍要检查观测曲线与拟合、拟合范围和日志。

```
set -euo pipefail
/usr/bin/time -v -o model.resources.txt \
  genomescope.R -i reads.hist -o profile -k 21 -p 2 > model.log 2>&1
test -d profile
```

低覆盖、错误尾、污染、GC偏倚或重复复杂度都可能使峰难以区分。若拟合明显偏离主体分布、参数碰到边界或结果对合理k值敏感，应把结论标为探索性并说明限制，保留不同参数的结果，而不是只报告最符合预期的一次。基因组大小、杂合度和重复比例是模型估计，不是组装质量或参考基因组的替代指标。

按实测安排批次资源

计数通常比拟合更消耗时间、内存和工作盘。记录FASTQ散列、过滤流程、Meryl/GenomeScope版本、k、倍性、完整日志、直方图、拟合图、峰值RSS和磁盘峰值。先测最大样本，再逐步测试并发；计数库可能远大于最终直方图，多个样本会叠加内存和I/O。失败时分别排查FASTQ损坏、计数资源不足、直方图格式和模型不收敛；不要把所有失败都归为机器不够用。

参考资料

Meryl官方项目与命令文档 官方文档
GenomeScope 2.0官方代码与用法 官方文档