# PLINK群体结构分析：LD剪枝、PCA结果核对与资源规划

在已完成基因型质控的数据上分步进行LD剪枝和PCA，按FID/IID核对样本，并用代表性运行记录内存、线程和耗时。

PLINK 群体结构分析不只是运行一次 PCA。输入样本、低频位点、亲缘关系和测序批次都会影响主成分；大队列还要提前估算内存与运行时间。本文以 PLINK 2 的 PGEN 文件组（同前缀的 .pgen/.pvar/.psam）为例，把 LD 剪枝和 PCA 分成两次运行，保留输入清单、日志及资源记录。开始前应完成基因型质控，并确认研究设计需要纳入哪些样本。

## 先固定输入与分析范围

先核对 PLINK 版本、PGEN 三件套是否来自同一次转换，以及样本表中的 FID/IID 是否唯一。记录参考基因组、变异过滤规则、样本筛选清单和文件校验值。PCA 前需处理极低 MAF 位点；LD 剪枝可减少少数高连锁区域对主成分的支配。具体阈值受物种、队列和研究目的影响，以下窗口和阈值仅是命令示例，不能直接当成通用最佳值。

| 设置/检查 | 作用 | 注意事项 |
| --- | --- | --- |
| --indep-pairwise | 生成低 LD 位点清单 | 千碱基窗口时步长设为 1；阈值需结合队列判断 |
| --extract | 让后续 PCA 真正使用剪枝位点 | .prune.in 本身只是位点 ID 清单 |
| --pca 10 | 计算指定数量的主成分 | PC 数量不能超过样本和变异提供的信息 |
| --threads | 限制计算线程 | 按任务配额设置，避免与其他任务争抢 |

## 剪枝和 PCA 分开运行

先为本次尝试新建空目录，记录二进制版本及输入 SHA256。下面的 500kb、步长 1、$r^2$=0.2、10 个 PC 和 4 线程是演示值；根据数据和分配资源修改。PLINK 2 的千碱基窗口要求步长为 1。剪枝命令会写出 prune.prune.in 与 prune.prune.out，但不会自动改变本次运行后续分析使用的变异，因此第二次 PCA 命令要显式加 --extract。

```
set -euo pipefail
mkdir run1
plink2 --version > run1/version.txt
sha256sum /data/qc.pgen /data/qc.pvar /data/qc.psam > run1/input.sha256
/usr/bin/time -v -o run1/prune.time.txt \
  plink2 --pfile /data/qc --indep-pairwise 500kb 1 0.2 \
  --threads 4 --out run1/prune > run1/prune.log 2>&1
test -s run1/prune.prune.in
/usr/bin/time -v -o run1/pca.time.txt \
  plink2 --pfile /data/qc --extract run1/prune.prune.in \
  --pca 10 --threads 4 --out run1/pca > run1/pca.log 2>&1
test -s run1/pca.eigenvec && test -s run1/pca.eigenval
head -n 3 run1/pca.eigenvec
```

先看两份日志中的样本数、变异数、过滤结果和警告，再打开 .eigenvec、.eigenval；不要只凭文件非空就认定结果正确。输出中的样本键应与原始 FID/IID（无 FID 时按 PLINK 规则处理）对应。与表型或协变量合并时按键连接并检查匹配率，不能因行数相同就横向拼接。绘图前检查重复样本、缺失 PC、离群点和各批次分布。

## 解释主成分并规划资源

PLINK 2 默认提取 10 个 PC；标准算法与 approx 随机近似算法计算路径不同。官方文档指出，大于约 5,000 个样本时可考虑近似方法，超过 50,000 个样本时通常很有必要；近似算法会对缺失基因型做均值填补。若使用近似法，应记录版本、PC 数、随机种子、线程与内存设置，并做结果稳定性检查。对更小队列，可先用标准算法建立基线，不要为了省资源贸然更换方法。

用 /usr/bin/time -v 分别记录剪枝和 PCA 的墙钟时间、峰值常驻内存、退出码及输出大小。PLINK 2 默认尝试为主工作区预留约一半系统 RAM；共享机器或容器应确认作业实际可用内存后，再用 --memory 调整，给操作系统和其他进程留余量。线程也要限制在调度器分配范围内。以代表性样本和变异规模先跑一次；不要按总位点数简单线性外推，因为样本量、PC 数和算法类型都会改变内存需求。

主成分图分开不同群体并不自动证明祖源差异。检查 PC 是否同时对应测序中心、芯片批次、亲缘关系、样本重复或极端缺失；若主要信号来自技术因素，应先处理或在解释中明确。做关联分析时，PC 数量要依研究设计与模型诊断选择并记录依据。任何过滤阈值变化都应生成独立结果目录，并一并保留输入哈希、样本键、剪枝清单、日志和资源记录，便于复核。

## 参考资料

PLINK 2 官方 LD 剪枝参数与输出 官方文档
PLINK 2 官方 PCA 与群体结构说明 官方文档
PLINK 2 官方内存、线程与随机种子说明 官方文档