TOBIAS多样本流程：参考、motif库与分阶段资源试跑

完整操作教程 · 参考固定 · 往返恢复验收

分阶段运行TOBIAS校正、足迹打分与motif预测，保存对应关系和验收记录。

运行边界：示例代码需按安装版本和真实项目输入复核；未提供真实生物数据，语法或小样本检查不代表科学结果和性能验收。

TOBIAS足迹流程包含Tn5偏好校正、信号打分和motif结合预测。多样本任务应统一参考与分析峰集，逐阶段验收，不能把普通覆盖度轨迹直接当已经做过偏好校正的足迹。本篇从一个已验收ATAC样本跑通入口，再按相同协议扩展清单。

先核对参考、峰集和原始处理

准备坐标排序且索引有效的ATAC BAM、同一组装FASTA及索引、共识峰BED、合适黑名单和motif库。确认染色体命名与坐标范围一致，保留重复处理、比对过滤和片段策略。ATACorrect会按其方法处理cutsite，不能不加说明地将已位移的输入再做同类位移。

阶段	固定材料	验收内容
ATACorrect	BAM、FASTA、峰集与黑名单	校正轨迹、报告和日志
ScoreBigwig	同一校正轨迹与既定区域	足迹bigWig可读且范围正确
BINDetect	打分轨迹、FASTA与motif库	motif扫描及结合预测文件
多样本	固定峰集、motif版本和参数	样本—信号—条件的对应表

按三个阶段记录资源

以下需要兼容TOBIAS和GNU Time；路径替换为已验收材料。先确认安装版TOBIAS ScoreBigwig --help可用，当前源码使用这个名称替代早期FootprintScores。4核仅为试跑额度，独立目录和日志便于定位失败阶段，不自动覆盖旧结果。

```
#!/usr/bin/env bash
set -euo pipefail
bam=/data/A.bam
genome=/refs/genome.fa
peaks=/refs/consensus.bed
blacklist=/refs/blacklist.bed
motifs=/refs/motifs.jaspar
for file in "$bam" "$genome" "$peaks" "$blacklist" "$motifs"; do
  test -s "$file"
done
mkdir tobias-pilot-001
root="$PWD/tobias-pilot-001"
TOBIAS --version > "$root/version.txt" 2>&1
/usr/bin/time -v -o "$root/atacorrect.time.txt" \
  TOBIAS ATACorrect --bam "$bam" --genome "$genome" --peaks "$peaks" \
  --blacklist "$blacklist" --prefix A --outdir "$root/correction" --cores 4 \
  > "$root/atacorrect.log" 2>&1
test -s "$root/correction/A_corrected.bw"
/usr/bin/time -v -o "$root/score.time.txt" \
  TOBIAS ScoreBigwig --signal "$root/correction/A_corrected.bw" \
  --regions "$peaks" --output "$root/A_footprints.bw" --cores 4 \
  > "$root/score.log" 2>&1
test -s "$root/A_footprints.bw"
/usr/bin/time -v -o "$root/bindetect.time.txt" \
  TOBIAS BINDetect --signals "$root/A_footprints.bw" --motifs "$motifs" \
  --genome "$genome" --peaks "$peaks" --cond_names A \
  --outdir "$root/bindetect" --cores 4 > "$root/bindetect.log" 2>&1
test -s "$root/bindetect/bindetect_results.txt"
```

校正轨迹可含负值，不要为便于画图把负值统一裁成零再送入下一步。检查bigWig染色体与长度、峰内信号及报告；文件非空只是最低检查。若日志提示参考缺失或坐标越界，先修材料，不用减少核数来掩盖输入问题。

多条件保持文件顺序清楚

样本清单至少记录BAM校验值、参考组装、峰集版本、打分输出和条件名称。第一次计算后抽查一个已知开放区域及一个低信号区域，比较校正前后和足迹分数，确认变化能够解释。motif库越大，扫描结果和输出文件越多，应按实际选择记录，不能用压缩库文件大小直接推算运行RAM。

每个样本或既定条件完成相同校正和打分，再按实验方案进入联合BINDetect。signals列表与cond_names位置逐一对应，不能依赖通配符碰巧排序；有生物学重复时先明确重复如何进入方法与后续评价，不把简单合并重复当作天然正确。

共识峰范围、blacklist、motif版本或打分参数改变都会影响结果，不能同时更换后只归因于服务器。motif相似的转录因子可能难以区分，预测结合不等于该TF已经在实验中直接验证。单条件入口也没有形成可靠的组间变化结论。

按真实瓶颈安排批量

分别记录每阶段墙钟、RAM峰值、工作盘及输出数，motif扫描与逐碱基轨迹处理的开销不相同。先顺序跑一例，再测试两例重叠；每例4核不代表可以无限叠加。高I/O等待先检查共享盘和输入读取，主机Killed要结合系统或调度器内存记录判断。

排期时保留原始BAM、未校正与校正轨迹、足迹信号和BINDetect输出的共存空间，以及失败现场与待回传文件。不要只计算最终汇总表。小区域适合验证参数，不能用其耗时保证全峰集；正式扩量后重新核对峰密度、motif数量、有效样本和完整交付时间。

三阶段各自检查，不要只看最终motif表

ATACorrect输出后先核对校正bigWig存在、染色体长度、文件可读性和报告中的峰/片段统计。ScoreBigwig输入必须确实来自同一批次ATACorrect，而不是普通coverage或未经校正轨迹。BINDetect再使用正确的footprint signal、FASTA、峰集和motif库。每一步使用单独目录，保存命令、退出码、日志与资源时间，才能定位文件格式或方法阶段的错误。

多样本BINDetect比较需要样本与条件清单、相同参考和可比的信号处理协议。比较的对象是footprint score和motif匹配条件下的结合预测，不等于直接测量TF蛋白结合。样本峰集不同、motif库版本不同或参考不一致时，结果差异不能简单解释为条件特异结合。

检查motif库与参考序列

motif库需记录来源、版本、物种及文件散列；检查motif ID唯一、序列格式有效，motif扫描阈值与研究问题相符。不同motif库可能包含重复或相近基序，预测TF家族时要保留motif ID和候选TF映射，不要把一个基序预测当作唯一TF身份。

FASTA、BAM、峰集和黑名单的组装必须一致。抽取若干峰检查染色体名称、范围和FASTA序列可提取性，核对索引与参考文件版本。峰集外的足迹预测没有同样输入含义；ATACorrect未完成或校正异常时，应先修正，而不是跳到BINDetect继续。

资源和文件句柄排查

记录ATACorrect、ScoreBigwig、BINDetect的峰值RSS、CPU时间、输入大小、输出大小和打开文件数。BINDetect处理大量motif时可能同时打开许多文件；遇到Too many open files应检查用户文件句柄限制和流程并行方式，而不是只加RAM。共享服务器应先确认单任务资源额度，再确定每个样本并发。

若ATACorrect读取BAM慢，查看索引、磁盘吞吐和FASTA访问；ScoreBigwig慢则检查信号轨迹与区域量；BINDetect慢则进一步查看motif数、扫描阈值和临时输出。更改cores只影响计算调度，不解决组装不符、缺少索引或motif库格式错误。

足迹结果的科学解释

足迹图中的局部信号形状会受Tn5偏好、测序深度、峰可及性和motif位置共同影响。BINDetect提供预测性结合与差异线索，不直接证明某TF发生了占位。报告候选motif时结合TF表达、基因调控证据、外部ChIP/CUT&Tag数据或实验验证，并说明TF家族基序的歧义。

跨样本比较时保留每样本ATACorrect校正轨迹、footprint score、条件映射和分析窗口。对低深度样本的空信号要区分技术覆盖不足与真实无占位。不要仅按一个差异p值或火山图排名就得出调控机制结论。

复现和资源咨询清单

清单至少包含TOBIAS版本、BAM/FASTA/peak/motif散列、黑名单版本、运行命令、cores、峰值内存、临时盘、样本条件映射及每阶段状态。服务器资源评估应基于代表性样本和最慢阶段，注明同一节点是否会并发运行多个样本。

请求环境支持时提供工具版本和错误阶段，说明参考与motif库体量、样本数、峰数及输出保留要求。若问题是缺少正确motif库或参考数据，增加CPU/RAM并不会补齐输入，也不会提升预测可信度。

常见报错怎样定位

校正bigWig缺失时先查ATACorrect日志、BAM读取、FASTA索引、峰集/黑名单坐标和输出权限。ScoreBigwig报regions或signal错误时检查输入是不是校正轨迹及染色体字典是否一致。BINDetect结果不完整则检查motif库、峰集、信号文件列表和打开文件限制。按阶段分开日志可以避免把所有失败都归因于内存。

如果峰集中有大量不存在的染色体或motif扫描没有命中，检查物种、组装和命名；不要用忽略警告的方式让工具继续。文件非空仍需核对内容和染色体长度。失败阶段、版本、输入散列和退出码都应记录到每样本状态表。

重复运行与流程改动

重跑失败阶段前确认旧进程结束，输出目录新建并保留旧日志。若改变黑名单、峰集、motif库或ATACorrect参数，后续ScoreBigwig和BINDetect也应使用一致的新版本输入，不能混用旧校正轨迹。生成一个新的run ID并将完整链路关联，便于识别每个预测来自哪些输入。

不同版本之间可能改变默认阈值和输出列。升级后在小样本上比较关键轨迹和预测表结构，并核对官方帮助。旧结果只有在输入及参数匹配时才可复用，不要按同名文件自动跳过整个阶段。

样本级状态表

为每个样本记录输入文件散列、参考组装、共识峰集、黑名单、motif库、运行版本、每阶段状态和人工复核。多样本分析使用同一份样本条件映射表，避免信号轨迹与条件标签顺序错位。失败样本保留失败阶段和最后成功文件，不能只统计最终成功的motif数。

完成结果前检查三个阶段的输出文件相互对应：校正轨迹来自该样本BAM，footprint score来自相同区域和轨迹，BINDetect配置引用正确的motif资源。任何一阶段重跑后更新依赖记录，确保下游文件不是旧版本缓存。

比较预测与外部证据

如果一个TF被预测为差异结合，核对其motif是否在多个TF间共享，结合TF在相关细胞中的表达、靶基因变化和实验条件解释。不同motif库版本可能改变匹配结果，方法说明应保留数据库版本和阈值。把预测结果描述为候选结合变化，不直接写成已验证的转录因子占位。

网络文件系统上大量motif文件同时读取时，检查文件句柄限制、inode和随机读取延迟。若BINDetect提示打开文件数不足，先复核系统限制及工具官方建议，再按合规方式调整用户级限制或串行化任务。不要通过减少motif库来绕过错误，除非研究设计明确允许并记录影响。

在最终清单中将每个样本的footprint轨迹与对应BAM、峰集和条件一一关联。对照与处理样本需具有可解释的实验配对；若样本缺失或质量不达标，统计流程应单独标记，而不是用一个轨迹替代。

TOBIAS不同阶段可能由不同机器或容器运行，迁移中间文件时记录来源、版本和散列，确认BigWig染色体字典、峰集及参考一致。ScoreBigwig的输入文件必须在清单中指向当前ATACorrect结果；如果前一步重新跑过，旧footprint轨迹不能自动保留为新结果。

样本级汇总除成功状态外，还记录每阶段峰值内存、墙钟和输出体积。若ATACorrect通过但BINDetect失败，应单独重试后者并保持前两阶段文件散列不变。日志里记录命令行中可能包含敏感路径，公开练习包应使用占位符。

每个样本完成后核对BINDetect条件名称与样本清单一致，尤其是文件顺序和组别拼写。条件名称误配会让输出仍然存在，却产生错误比较。将条件映射表作为正式输入保存并计算散列。

足迹与motif文件较多时，输出目录应按sample、stage和run_id分层，避免不同样本同名文件覆盖。归档前抽查文件头、染色体字典和motif ID，确认各阶段的样本对应关系。

对照峰和黑名单文件也要与组装版本绑定。资源清单中保存下载来源、日期、文件散列和染色体命名；若参考库更新，必须重新确认校正和扫描的输入一致。

资源测量可在同一节点记录系统内存、作业容器限额和进程峰值，三者口径不同。交付时注明使用哪一种监控数据，避免把主机总内存误认为单个任务实际可用额度。

配套练习包

本PDF对应的ZIP包含本文代码块、主题练习模板或合成数据、README和SHA256校验文件。请从同一教程页面下载练习包；先核对校验值，再在独立目录复现。

参考资料

TOBIAS ATACorrect官方说明 · 来源资料

ScoreBigwig参数与旧命令替代关系 · 来源资料

BINDetect多条件与结果说明 · 来源资料

操作建议：先在合成小样本上验证流程，再依据项目参考、标签合同、存储策略与获配资源验收真实数据。