MACS3批量运行教程：样本对照表、独立日志与资源限制

完整操作教程 · 参考固定 · 往返恢复验收

以明确处理/对照清单和有限并发运行MACS3，逐样本保留日志、退出码与峰文件验收。

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。

MACS3批量峰识别首先要保证每个处理样本关联了正确对照，再控制并发。把所有BAM一次传给-t会合并处理，并不等于逐样本运行。本篇限定人类双端TF ChIP-seq的窄峰示例；ATAC插入位点、单端或宽峰任务需要另定协议，不能只换文件名。

先建样本与对照清单

samples.tsv使用sample、treatment、control三列表头，每行一个独立分析。路径指向已验收的BAM，同一参考组装、染色体命名和过滤规则均应一致。对照可以按实验设计复用，但要保留对应关系；生物学重复不要为了节省任务数随意合并。

检查项	固定内容	验收记录
关联	一个样本ID对应处理和对照	ID唯一、路径存在、设计合适
协议	BAMPE、hs、q=0.01、去重策略	记录实际版本和完整参数
资源	两个并发实例，每例预算2CPU	总RAM、I/O与作业额度
输出	每例日志、XLS和narrowPeak	退出码、峰数及质量复核

用固定队列运行，错误也留记录

下例依赖支持--threads预算的MACS3 3.1接口，运行前保存macs3 --version及callpeak --help，并确认参数存在。旧版不应猜测同名选项。代码保存为run_peaks.py，在新目录运行；两例并发只是演示，应按实测单例RAM和获配资源降低或增加。

```
import csv,os,re,subprocess
from pathlib import Path
from concurrent.futures import ThreadPoolExecutor
with open("samples.tsv",encoding="utf-8-sig",newline="") as f:
    reader=csv.DictReader(f,delimiter="    ")
    assert reader.fieldnames==["sample","treatment","control"]
    rows=list(reader)
assert rows and len({r["sample"] for r in rows})==len(rows)
for r in rows:
    assert re.fullmatch(r"[A-Za-z0-9][A-Za-z0-9_.-]*",r["sample"])
    for field in ("treatment","control"):
        q=Path(r[field]).resolve(strict=True)
        assert q.is_file() and q.stat().st_size>0
        r[field]=str(q)
help_text=subprocess.run(["macs3","callpeak","--help"],
    check=True,capture_output=True,text=True).stdout
assert "--threads" in help_text,"■■■■MACS3■■"
root=Path("macs-batch-001");root.mkdir()
env=os.environ.copy()
env.update(OMP_NUM_THREADS="1",OPENBLAS_NUM_THREADS="1",MKL_NUM_THREADS="1")
def one(r):
    name=r["sample"]; out=root/name;out.mkdir()
    cmd=["macs3","callpeak","-t",r["treatment"],"-c",r["control"],
         "-f","BAMPE","-g","hs","-q","0.01","--keep-dup","auto",
         "--threads","2","-n",name,"--outdir",str(out)]
    with (out/"run.log").open("w") as log:
        rc=subprocess.run(cmd,stdout=log,stderr=subprocess.STDOUT,env=env).returncode
    peak=out/(name+"_peaks.narrowPeak"); xls=out/(name+"_peaks.xls")
```

```
    valid=peak.is_file() and xls.is_file() and xls.stat().st_size>0
    status=rc if rc else (0 if valid else 1)
    return [name,status,peak.stat().st_size if peak.is_file() else "missing"]
with ThreadPoolExecutor(max_workers=2) as pool:
    result=list(pool.map(one,rows))
with (root/"status.tsv").open("w",newline="") as f:
    w=csv.writer(f,delimiter="    ");w.writerow(["sample","status","peak_bytes"]);
w.writerows(result)
raise SystemExit(1 if any(x[1]!=0 for x in result) else 0)
```

一个样本失败仍记录其他已完成任务，最终汇总返回非零，重跑时另建目录并仅选择确认需要重做的行。空narrowPeak可能代表无峰，不自动当缺文件；它仍需结合XLS、日志与实验质量判断。队列只验收基本产物，BAM损坏、参考不匹配或对照错误应在正式批量前排除。

CPU预算与并发数分别计算

当前接口的--threads是单次调用预算，包含主执行参与者；它不代替外层队列和调度器配额。两例各2CPU不能都按整节点额度启动，RAM与读盘压力也会叠加。保留版本和帮助信息，不能把旧版默认行为当作新版统一规则；设置预算更高也不保证所有阶段并行。

先用两个具有代表性的处理/对照组合测完整墙钟、内存和工作盘，再比较串行与双任务重叠。CPU空闲但读盘等待高，优先检查共享存储；Killed先看作业内存和系统记录。不要为了让任务跑完而改变q值、宽窄峰类型、重复处理或对照集合，再按原方法报告。

结果按实验含义验收

资源试跑表同时写清处理与对照的输入大小、完整耗时、最大RAM和成功样本数。首次从单例开始，再测两个重叠任务；不把两个样本耗时平均后忽略其中一例失败。若多个样本共享同一对照，缓存可能影响重复读取，单列首次和复用情形，不能据此承诺整批都保持同样速度。

抽查峰坐标、参考范围、对照背景及已知阳性/阴性区域，结合测序质量、富集和重复一致性判断。峰多不等于质量高，q阈值也不能代替实验重复。保留清单、完整命令、日志和原始峰文件，后续共识峰或重复一致性分析另有规则，不能把不同样本峰表简单求并集当最终结论。

先检查样本配对和BAM质量

样本清单中每个sample都应唯一，treatment和control应来自同一实验设计并适合目标比较。批量运行前由熟悉实验的人员确认IP、input/IgG对照、抗体、重复和处理组关系；程序只能检查路径和文件存在，不能判断实验对照是否合理。若一个对照用于多个处理样本，应在台账明确记录，避免误把它当作独立重复。

MACS3按BAMPE处理双端片段时，确保输入是与格式匹配的成对比对结果，并确认排序、重复处理、参考组装和染色体命名一致。运行前检查BAM头、索引、mapping统计和片段长度分布；samtools quickcheck或idxstats只能做部分检查。若处理与对照的参考或过滤规则不同，应先解决输入可比性。

参数会改变峰定义

本例中的BAMPE、有效基因组大小、q阈值和keep-dup共同决定峰识别结果。hs适用于特定人类参考背景，非人类、不同染色体集合或特殊实验设计应重新确定有效基因组大小。q值阈值与重复策略不能为了得到更多或更少峰而随意调整，应记录选择依据并按实验协议固定。

MACS3处理合并输入与逐样本独立运行回答的问题不同。把多份BAM给一个命令可能将样本合并，而不是分别得到每份文库峰。若目标是生物学重复，保留逐重复结果及质控，再采用合适的重复一致性或差异结合分析流程；不能把每份narrowPeak简单拼成一张表称作差异分析。

结果质量检查

对每个样本核对退出码、narrowPeak和xls是否存在、峰数、信号分布、峰宽、FRiP或其他适合实验的质量指标。没有统一适用于所有抗体和细胞类型的峰数阈值，应结合历史批次和实验设计。抽查峰是否落在预期区域、背景是否异常高，并与处理/对照信号或已知阳性区域比较。

并发批次中一个任务失败不能让总批次被错误标为成功。汇总每个sample的结束码、日志末尾、输出大小和质控状态；缺少任一必要产物都应标记未完成。对于失败项保留原目录并在新attempt目录重跑，避免被已有的部分文件误导。

并发和资源如何安排

示例两任务并发且每任务2线程时，线程预算是示意值，实际CPU和内存要按单任务实测。MACS3线程、Python运行时和I/O会共同影响资源；多个任务同时读取大型BAM时，磁盘吞吐可能成为瓶颈。先串行测一个代表样本，再测低并发，记录总墙钟时间、每任务峰值RAM、临时盘和读写等待。

若增加并发后每个任务变慢，通常应先检查磁盘共享、缓存、内存争用和线程过量，而不是继续加线程。样本数量多时用受控队列限制同时运行数，并确保对照文件在任务期间只读。不要用同一个输出目录重跑任务，也不要将所有后台进程的资源需求只按CPU核心数估算。

差异结合与生物学结论

单个样本callpeak产生候选峰集合，不等于组间差异结合显著性。组间比较应采用合适的统一峰集、计数矩阵和统计模型，考虑生物学重复及实验设计。不同样本峰数差异可能来自深度、抗体效率、背景和参数，不能单凭峰数量得出条件效应。

报告中区分峰识别、峰注释和差异检验。保留每个样本独立峰文件、对照配对、质控指标、软件版本和完整命令。若仅进行候选峰探索，应明确为探索性分析，并将后续差异测试和实验验证作为独立步骤。

重复样本的峰一致性

先分别评估生物学重复的峰数、信号强度、片段长度和富集指标，再决定是否进入合并或差异模型。峰集一致性可按预先定义的重叠规则比较，但相邻峰合并或阈值改变会影响重叠率。保留每个重复的独立峰文件和质控信息，避免将重复先合并后失去可评估的个体差异。

若同一抗体不同重复峰数差异很大，先查测序深度、抗体批次、对照选择、重复率和实验质量。MACS3参数可以影响候选峰，但不能用调阈值把质量差的重复变得可比。差异结合需要样本级计数和模型，技术重复的处理方式应在方案中说明。

输入核对的可复现步骤

记录每个处理/对照BAM的SHA256、参考版本、索引生成时间、比对软件和过滤规则；执行前检查样本表中的路径解析到预期文件且没有同一文件被误用成多个样本。脚本可检测路径唯一性，但如果实验设计确实共享对照，需显式允许并在台账记录，不能默默假设共享是错误。

在正式任务开始前对一个处理-对照组合人工核对BAM header、染色体集合及基本比对统计。宏观QC通过之后再执行批量队列。命令、版本帮助和运行日志放入每个样本目录，方便在MACS3更新后知道哪些参数是当时接受的。

并发队列失败恢复

批处理函数应收集每个样本的退出码和必需产物完整性，并在所有Future完成后生成汇总。某个子进程失败后，检查是否有同名部分narrowPeak文件，再决定是否清理或保留；重试使用新的attempt目录。队列达到并发上限后仍应保留待运行状态，不能将尚未启动任务计作成功。

系统中断时查看作业管理器记录和OOM事件，区分进程异常、节点回收、磁盘满和用户取消。记录失败时间与最近完成步骤。共享机器上不要通过杀掉其他用户进程腾资源；降低自己的并发或申请新的作业额度，再按台账恢复。

重复样本的峰一致性

分别评估生物学重复的峰数、信号、片段长度和富集指标，再决定进入合并或差异模型。峰集重叠会受峰合并规则影响，需预先定义。保留各重复的独立文件与质控信息，避免先合并而无法评估个体差异。

若重复峰数差异很大，先查测序深度、抗体批次、对照选择、重复率和实验质量。不能靠调阈值把质量差的重复变得可比。差异结合需要样本级计数和模型，技术重复的处理方式应在方案中说明。

输入和失败复核

记录处理/对照BAM散列、参考版本、索引生成信息、比对软件与过滤规则。执行前人工核对一对代表样本的header、染色体集合及基本比对统计。某个样本异常退出后保留部分产物，使用新attempt目录重跑；不能让旧narrowPeak冒充当前输入的结果。

资源预算

资源计划包含样本数、BAM规模、线程、并发上限、临时空间和重试余量。处理与对照同时从共享盘读取时，并发可能造成I/O争用。先测最大样本和计划并发，记录峰值RAM、读写等待和总时间；若并发后变慢，应降低队列并发检查共享盘，而不是继续增加线程。

批次汇总需区分峰识别成功和样本质控通过。若程序产生narrowPeak但实验QC未达标，状态仍应是“计算完成、质控待定”，不能标为可用于差异分析。

对照BAM复用应在样本清单明确标识，资源估算按真实并行读取量计算。任务结束后设置轨迹目录为只读，防止后续重跑覆盖原始峰调用结果。

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

MACS3 callpeak格式与参数 · 来源资料

MACS3 CPU预算说明 · 来源资料

MACS3发布说明 · 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。