Juicer大型数据流程：输入清单、步骤依赖与计算资源实测

完整操作教程 · 参考固定 · 往返恢复验收

以固定CPU脚本和QC范围建立Juicer基线，记录阶段前件、完整工作盘与后处理边界。

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。

Juicer大任务应把比对、排序去重、统计、接触图与后续特征分析分开测量。CPU脚本、集群脚本和juicer_tools版本也不能混用。本篇固定已核对的CPU脚本入口，用QC模式跑一份可验收基线，保留全部阶段依赖，再规划正式大规模任务。

固定部署、参考和实验范围

本次源码修订为177eb610397e4207fc56db1df169b2d08d06d43a，CPU脚本声明2.0。部署目录应有scripts/juicer.sh及配套common工具，BWA参考索引与FASTA相邻，chrom.sizes、限制酶位置文件来自同一组装。下例限定MboI方案，不适合不经核对就换成其他酶或Micro-C。

项目	填写或核对	验收边界
输入	一个文库的已质控配对FASTQ	R1/R2及样本、文库身份
CPU脚本	兼容BWA、samtools、Java和juicer_tools	不是只有juicer.sh一个文件
比对线程	-t控制BWA相关额度	不等于所有阶段总线程
建图线程	-T控制hic创建相关额度	Java内存及临时盘另核
QC模式	--qc限制最低分辨率并跳过后处理	不当作正式全部特征分析耗时

独立目录运行一份QC基线

这个固定修订在建图前硬编码_JAVA_OPTIONS为-Xmx150000m -Xms150000m，QC也执行此段；4线程或QC不会自动降低Java堆。运行前由维护者核对获配RAM和实际堆，若需调整，另存修改后的脚本、记录新SHA并重新验收，不能把QC理解为低内存模式。

以下需要已验收的完整CPU部署及GNU Time；路径替换后，先检查安装脚本帮助包含本例参数。脚本只引用已核对输入，创建本次独立fastq目录；同一文库的多lane合并规则应先定义，不将不同生物学样本放进同一个目录自动混合。

```
#!/usr/bin/env bash
set -euo pipefail
juicer=/opt/juicer-cpu2
ref=/refs/genome.fa
sizes=/refs/chrom.sizes
sites=/refs/genome_MboI.txt
r1=/data/S1_R1.fastq.gz
r2=/data/S1_R2.fastq.gz
for file in "$juicer/scripts/juicer.sh" "$ref" "$sizes" "$sites" "$r1" "$r2"; do
  test -s "$file"
done
test "$r1" != "$r2"
gzip -t "$r1" "$r2"
mkdir juicer-qc-001
root="$PWD/juicer-qc-001"
mkdir "$root/fastq"
ln -s "$r1" "$root/fastq/S1_R1.fastq.gz"
ln -s "$r2" "$root/fastq/S1_R2.fastq.gz"
bash "$juicer/scripts/juicer.sh" -h > "$root/help.txt" 2>&1
sha256sum "$juicer/scripts/juicer.sh" "$r1" "$r2" "$ref" "$sizes" "$sites" \
  > "$root/inputs.sha256"
/usr/bin/time -v -o "$root/resources.txt" \
```

```
  bash "$juicer/scripts/juicer.sh" -D "$juicer" -d "$root" \
  -z "$ref" -p "$sizes" -s MboI -y "$sites" -b GATCGATC \
  -t 4 -T 4 -i S1 -k library1 --qc > "$root/run.log" 2>&1
test -s "$root/aligned/merged_dedup.bam"
samtools quickcheck -v "$root/aligned/merged_dedup.bam"
test -s "$root/aligned/inter.hic"
test -s "$root/aligned/inter_30.hic"
```

QC模式在该修订中建图到1000bp并跳过后续注释，不代表已经完成更高分辨率或环等特征分析。BAM基础检查及hic非空也不证明矩阵可用，应实际读取接触图，检查统计和已知区域；脚本内部日志的第一条实质错误不能被最后一个文件掩盖。

恢复前核对每一步的前件

阶段恢复参数需按安装版帮助选择：chimeric要求比对完成，merge要求嵌合处理完成，dedup要求已合并排序，final要求去重完成而hic尚未生成。postproc另有已有hic和后处理资源要求，不能把这些名字当任意断点。输入、脚本或方法变更后，不无说明复用另一轮的中间文件。

保存每阶段开始/结束、退出状态、输入输出及校验；发现失败先确认原进程已结束，再定位可复用前件。不要仅创建一个同名空文件让流程跳过。大任务拆分也要保持读段配对与去重范围，不能分别去重后简单拼结果就称为全量等价。

把资源观察放回完整交付

GNU Time的最大RSS不等于所有并发子进程内存的总和；多阶段、多进程任务还应看作业或控制组监控，并记录采样口径。对照试跑固定同一QC范围，不能拿另一组更低分辨率的数据直接比较速度。

工作盘应包括FASTQ、splits中的中间比对、排序临时量、去重BAM、建图临时目录、两个hic和待保存旧结果。最终hic体积不能倒推峰值。测单任务及小并发的RAM、I/O和墙钟，BWA线程更多也可能受读写或后段Java计算限制，不给每个实例都申请整节点额度。

正式流程关闭QC限制或加入后处理时，重新核对Java堆、juicer_tools和具体算法的CPU/GPU支持，再测对应阶段。不能将QC基线时间当全部正式分析报价。结合重复率、有效互作、距离分布和测序深度验收目标分辨率，保留配置与科学结果检查记录后再扩展队列。

Juicer参数和Java堆必须与可用内存匹配

固定修订脚本中的Java堆设置需要在运行前人工核对。`-Xmx150000m -Xms150000m`意味着Java堆初始与最大值均约150GB；若节点或作业实际额度低于此值，JVM可能无法启动或触发系统OOM。QC模式、少量线程或测试数据都不会自动降低这个堆设置。确认该脚本版本后，按获配额度修改副本并记录diff和散列。

`-t`和`-T`分别控制相关阶段的线程额度，不代表整条工作流只有这些线程或内存需求。比对、排序、去重、Juicer_tools建图和Java堆会在不同时段使用资源。通过配置或脚本本身确认每个阶段的线程和内存逻辑，再以实际日志和调度器记录验证。

输入参考和酶切文件的核对

参考FASTA、BWA索引、chrom.sizes和酶切位点文件必须来自同一组装、同一染色体命名和MboI方案。校验FASTA索引和BWA索引文件齐全，读取随机区域确认索引可用；核对sizes长度与FASTA字典相同。酶切位点文件应由准确的参考及酶切规则生成，不能用另一版染色体长度文件替代。

FASTQ应来自同一文库的正确R1/R2，检查配对记录和gzip完整性。多lane归并要依照实验记录；不同文库需分别建立任务目录。Juicer CPU脚本、集群脚本和juicer_tools版本之间参数可能不同，保存脚本提交、帮助输出及工具版本。

QC模式输出如何解释

QC模式用于评估部分流程和质量，可能限制矩阵分辨率或跳过某些后续步骤。QC可验证部分输入、比对和建图路径，但不能替代正式高分辨率接触矩阵及下游特征分析。记录QC运行实际完成到哪一阶段、生成哪些文件、哪些步骤明确未执行。

检查比对统计、重复、valid pairs、矩阵分辨率和联系图基本结构。不同实验方案的质量阈值不同，不能套用一个固定比例做所有样本的通过标准。若QC结果不理想，先检查参考、酶切、FASTQ质量和建库，而不是直接申请更大服务器。

阶段依赖和恢复规则

Juicer多个阶段依赖前序生成的中间目录和文件。检查每一步完成标志、日志和文件大小，使用工具支持的resume/重新启动机制。不同脚本或版本产生的中间输出不一定兼容；不要把另一次运行的merged_nodups或hic文件复制进当前目录伪装为已完成。

任务被调度器终止时，查看OOM、时间限额、磁盘满和节点回收记录。确认Java和子进程均结束后，保留失败目录与日志，在新的attempt中恢复。改过Java堆或线程后重新计算配置散列，并重新验证一份代表样本。

存储、并发与完成时间

FASTQ、临时SAM/BAM、去重pairs和`.hic`文件会在不同阶段同时占用空间。按代表样本测量工作目录高水位，并为并发样本和失败重试留余量。将输入放在只读参考目录，输出按sample/library/run ID分隔；确认快速scratch是否会在作业结束时自动清除。

并发数要以每任务堆、排序RAM、CPU和存储带宽共同核算。Java固定大堆时不应贸然并发多个样本。先串行完成基线，再按获配资源调整每阶段设置和低并发试跑。报告全流程墙钟时间与各阶段耗时，不把QC小跑时间直接外推为正式分析时长。

结果交付和方法界限

交付内容包含输入散列、Juicer脚本和juicer_tools版本、Java版本与堆配置、参考及酶切文件版本、完整命令、日志、资源记录、QC状态和输出文件散列。注明正式分辨率是否达到，哪些后续分析未执行。文件存在和`.hic`可读只说明格式和流程的一部分，不等于互作结果已完成科学解释。

接触图中的互作信号受测序深度、归一化、分辨率和建库质量影响。跨条件比较需采用适当的重复设计与统计方法，不能仅比较浏览器中颜色深浅。给报告明确区分流程QC、矩阵构建和差异互作推断，避免把QC图当作结论。

实际运行前审阅固定脚本

下载或检出固定Juicer CPU版本后，对脚本中的Java选项、线程、临时目录和路径执行代码审阅并计算散列。若修改-Xmx/-Xms，确认Java参数格式和实际单位；提交前运行帮助和一个小规模QC，并由运维确认作业额度。不要编辑系统共享安装目录，使用项目副本并记录修改。

即使堆额度满足，其他阶段也需要额外RAM供BWA、排序、系统缓存和Java外部内存使用。作业请求应高于Java堆并留出余量，但实际余量由测量决定。观察调度器内存峰值和OOM日志，而不是只看Juicer脚本设置。

QC和正式运行之间的差别

QC运行可验证部分读段处理与矩阵构建，但其分辨率、过滤或跳过步骤可能与正式分析不同。对照输出时列出QC模式打开了哪些条件、限制了哪些阶段、生成了哪些分辨率。正式任务应以研究目标要求的完整参数另行运行并记录，不把QC结束作为全量流程完成。

从QC切换到正式运行时新建目录或使用明确定义的输出隔离，避免低分辨率QC文件混入正式hic目录。整理每个文库QC状态、正式运行状态和后续特征分析状态，按阶段汇报覆盖。

Java堆调整的决策流程

先从调度器确认作业实际可用内存，再查看Juicer脚本的Java堆配置和其他并发步骤的额外内存需求。如果固定堆大于额度，按经过维护者审核的副本降低堆或申请合适额度，并用代表样本验证是否满足算法需求。不能把-Xmx降得很低后宣称配置可运行；堆不足可能只会把错误改成Java heap space。

每次改Java参数都保存修改前后diff、脚本SHA256、作业资源请求及日志。对不同大小FASTQ测试启动阶段、建图阶段和结束阶段内存。提交集群脚本时确保请求内存大于堆并留有操作系统与原生库余量，具体值按峰值测量决定。

交付和空间管理

Juicer输出常包含多个中间文件和接触图；按项目规定保留必需阶段产物、日志和最终`.hic`文件，并记录被清理的可重建项。清理前检查下游是否还依赖merged_nodups或其他中间文件。输出归档时做大小与散列核对，防止大文件复制未完成。

如果需要跨样本比较，确保酶切方案、参考、有效pairs过滤和归一化流程一致。对不同深度样本采用适当分析方法，浏览器图只能做QC和探索。报告流程状态、资源峰值及未执行步骤，避免把矩阵构建写成差异互作分析结论。

Juicer脚本运行前可用grep或人工检查核对脚本内Java options及所有资源相关设置，但要保存复核后的实际脚本，不能只保存原始版本链接。运维人员应知道用户实际调用的是哪个路径和版本。

若大堆导致JVM启动失败，保留完整stderr和调度器记录；若降低堆后运行到中途才报heap space，说明该规模的算法需求超过当前设置，应重新评估任务规模、方法和额度。不要通过清空日志或重复覆盖来掩盖首次失败。

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

固定修订Juicer CPU脚本及参数 · 来源资料

Juicer官方运行与恢复说明 · 来源资料

Juicer官方部署说明 · 来源资料

GNU Time资源统计口径 · 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。