HiC-Pro部署实操：酶切与参考配置、样例运行及工作盘规划

完整操作教程 · 参考固定 · 往返恢复验收

先核酶切和成套参考，再以独立样本目录运行HiC-Pro并检查有效连接、矩阵与资源。

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。

HiC-Pro从原始双端读段到接触矩阵，最容易踩坑的是酶切、连接位点和参考不一致。CPU再多也不能修复这些配置。本篇按已核对官方配置和独立运行入口，准备一个样本并保留完整阶段记录，先让结果可验收，再扩大数据量。

先准备成套参考与实验设置

记录物种、组装、文库方案、限制酶及接头处理。Bowtie2索引、染色体长度与酶切片段BED应来自同一参考，染色体命名一致；连接位点不是把酶名直接写进配置。DNase或Micro-C等方案不能照搬限制酶模板，需按适用流程另核过滤设置。

项目	填写或核对	验收边界
N_CPU / SORT_RAM	获配CPU与排序相关设置	不是整个作业RAM硬上限
PAIR1_EXT / PAIR2_EXT	与样本R1/R2文件名匹配	配对读段身份和顺序已验收
BOWTIE2_IDX_PATH / REFERENCE_GENOME	索引目录和实际前缀	整套索引文件保持完整
GENOME_SIZE / GENOME_FRAGMENT	长度表与适用酶切片段	参考组装与区间范围一致
LIGATION_SITE / BIN_SIZE	实验连接位点与目标分辨率	方法设置不能因机器变化省略

从安装版附带的config-hicpro.txt另存config-hicpro.checked.txt，按项目填写并复核，不能只沿用示例的hg19和HindIII默认值。当前核对源码修订为57a0bf2cb1d7ebeb44b4f11e78058b504260c5c9，手册中有历史版本示例，因此参数以实际安装帮助和对应配置为准。

按一个样本一个目录组织读段

以下需要已正确安装且兼容的HiC-Pro及依赖、GNU Time，以及已经质控并验证配对的FASTQ。脚本将本次输入映射到独立rawdata/S1目录；原始读段只通过软链接引用，不按目录名猜测哪些文件属于同一个生物学样本。

```
#!/usr/bin/env bash
set -euo pipefail
r1=/data/S1_R1.fastq.gz
r2=/data/S1_R2.fastq.gz
config="$PWD/config-hicpro.checked.txt"
test "$r1" != "$r2"
test -s "$r1"
test -s "$r2"
test -s "$config"
gzip -t "$r1" "$r2"
mkdir hicpro-pilot-001
root="$PWD/hicpro-pilot-001"
mkdir -p "$root/rawdata/S1"
ln -s "$r1" "$root/rawdata/S1/S1_R1.fastq.gz"
ln -s "$r2" "$root/rawdata/S1/S1_R2.fastq.gz"
cp "$config" "$root/config.txt"
HiC-Pro -h > "$root/help.txt" 2>&1
```

```
sha256sum "$r1" "$r2" "$root/config.txt" > "$root/inputs.sha256"
/usr/bin/time -v -o "$root/resources.txt" \
  HiC-Pro -i "$root/rawdata" -o "$root/results" -c "$root/config.txt" \
  > "$root/run.log" 2>&1
shopt -s nullglob
pairs=("$root/results/hic_results/data/S1/"*allValidPairs)
test "${#pairs[@]}" -eq 1
test -s "${pairs[0]}"
```

gzip校验只检查压缩完整性，不证明R1/R2配对或样本身份。脚本检查有效配对文件只是最低产物门槛，仍要核对所选分辨率矩阵和归一化结果。输入和配置校验值与参考来源、版本清单共同保存，不能只保留一份最终热图。

按阶段解释失败与资源

分别观察全局与连接位点相关比对、配对过滤、样本合并去重、建图和ICE归一化。比对率异常先查参考与读段；无效连接比例异常先查实验、片段和过滤；主机Killed再结合平台或系统内存记录判断，不把所有失败统称为算力不足。

整套资源记录包括输入、拆分/比对中间文件、排序临时文件、有效配对、多个分辨率矩阵和待回传副本。压缩FASTQ大小不是工作盘峰值。N_CPU只描述相关并行额度，多样本叠加还要核对总CPU、RAM和I/O；先单样本基线，再测试允许的小并发。

验收后再决定恢复与扩量

检查总读段、唯一比对、有效连接、重复、cis/trans及距离分布，结合实验类型解释，不用某一个固定比例包打天下。矩阵能画图不等于分辨率有足够支持；更细bin会增加计算和稀疏程度，需要测序深度与分析目的共同支撑。

分步恢复时，mapping、proc_hic、merge_persample、build_contact_maps与ice_norm接受的输入阶段不同，先读安装版帮助及对应输出说明，不把原FASTQ目录传给每一步。保留失败日志和中间文件，确认旧进程结束后再按当前流程恢复；未经验收不要批量删除可追溯材料。

先从实验方案确认酶切参数

Hi-C分析配置中的restriction enzyme、ligation site、GENOME_FRAGMENT和参考长度必须来自实际建库方案与同一基因组组装。示例配置中的HindIII或hg19不能直接用于其他实验。多酶或酶切位点特殊时，依据试剂盒和项目实验记录核实连接位点及切片段文件，不能只替换配置中的酶名称。

FASTQ配对、lane和文库是不同层级。先建立样本表，列出R1/R2、lane、library_id、sample_id和是否属于同一生物样本。只有同一文库的多个lane才按既定规则汇总；不同样本不能因文件名相似而放入同一输入目录。保存文件散列和gzip完整性检查。

配置文件逐项审阅

从安装版本提供的config模板复制出项目专用副本，逐项填写Bowtie索引前缀、参考长度、酶切位点、比对质量过滤、重复/片段处理、线程、排序RAM和临时目录。不要把排序内存参数解释为整个作业RAM上限，HiC-Pro各阶段还会调用不同程序并有其他内存需求。

对配置变更做diff并由第二人复核，尤其是参考索引前缀、GENOME_SIZE、GENOME_FRAGMENT、LIGATION_SITE和BIN_SIZE。软件默认值可能随版本变化，保存模板来源、提交版本和最终配置散列。若没有匹配的参考索引或片段文件，先准备并验证，不要在分析脚本里临时猜测文件路径。

逐阶段看结果和质量

HiC-Pro会产生比对、过滤、去重、valid pairs、矩阵等阶段结果。检查各阶段读段数是否有合理递减，未比对、无效连接、重复和有效pairs的比例，并使用配套统计报告判断文库质量。不同文库质量差异可能来自建库而非服务器；不能仅用最终矩阵是否存在作为通过标准。

matrix构建要核对分辨率、矩阵对称性、染色体长度和valid pairs支持。小分辨率矩阵可用于检查流程，但不代表高分辨率交互足够稳定。比较样本时确保使用一致的有效pairs过滤与归一化方法。HiC-Pro本身完成并不代表下游差异互作分析已完成。

磁盘与临时空间规划

原始FASTQ、未压缩中间文件、BAM、valid pairs和不同分辨率矩阵可能同时存在。开始前估算工作目录峰值占用，确认inode、配额和临时目录容量。软链接只节省原始文件副本空间，但中间结果仍会写入输出目录；共享存储上多个样本并发可能造成I/O拥塞。

清理中间文件前核对哪些文件可重建、哪些是最终审计所需。保留统计报告、配置、日志和最终矩阵的同时，应按机构策略清理临时FASTQ副本或大BAM。若只保存矩阵而删除必要的valid-pairs或参数记录，后续可能无法复现特定分辨率。

故障分流和恢复

输入识别或配对错误时检查命名规则、软链接目标和样本目录；比对阶段失败先查索引、参考、内存和bowtie2日志；排序/去重阶段失败查临时盘和排序RAM；矩阵阶段失败则看valid pairs、分辨率和工具版本。每阶段保存退出码和日志末尾，防止只看到总脚本失败而无法定位。

恢复时遵循HiC-Pro自身对已完成阶段和输出目录的规则。未确认支持resume前不要手动复制中间文件或用空文件骗过流程检查。若配置变了，使用新的输出目录并完整记录。超时或OOM后，确认残留进程已结束，再用代表样本测试修改后的资源。

对服务器和数据管理的要求

高分辨率Hi-C任务可能需要大量临时存储和较长作业时限。资源沟通时给出FASTQ总量、样本数、目标分辨率、测试阶段峰值、临时盘高水位和并发计划。CPU、RAM、存储吞吐与工作盘空间要分别考虑，不能用单一“核数”概括需求。

Hi-C数据和接触矩阵可能包含敏感样本信息。确认项目目录权限、备份范围、临时文件保留周期和交付访问方式。服务器部署能解决环境隔离和资源执行问题，但酶切方案、建库质量和数据授权仍由研究项目负责。

结果目录和重复样本管理

每个生物学文库使用独立目录和运行ID，保留rawdata软链接来源、FASTQ散列、配置副本和结果统计。样本名称要稳定且不能含会被流程解析成lane或配对标记的歧义字符串。多个lane如需合并，按安装版本规定命名并在manifest注明原始文件到文库的关系。

重复分析应保留每个文库的独立valid pairs和矩阵。合并前核实样本级有效接触数和QC，避免某一个深度高的文库主导联合结果。不同文库实验质量差异要在后续模型中考虑，不能只将其矩阵相加后称作重复验证。

流程检查点和磁盘告警

运行过程中监控工作目录增长、临时目录、可用inode和配额。HiC-Pro多阶段会生成大型中间结果，磁盘写满可能造成看似不相关的后续错误。日志出现文件写入失败或截断时先检查空间，再判断流程代码问题。对共享磁盘预留并发空间，避免多个项目同时写满同一文件系统。

每个阶段可记录输入/输出文件数量、总字节数、完成标志和时间。某阶段输出为空时，查看上游日志及退出码，不要手动创建空文件让下一阶段继续。重跑时尽可能使用工具支持的恢复方式，且保存旧结果用于审计。

质量指标的解释

联系有效率、重复率、比对率和矩阵可用性要结合建库类型、测序深度与实验协议。不同实验室或酶切方案的经验阈值不一定可互用。对异常指标查看原始FASTQ质量、酶切位点统计、连接片段和重复情况，再决定是否继续下游。

矩阵可读且稀疏并不表示有效互作足够。绘制不同分辨率的cis/trans交互、距离衰减和覆盖情况，检查染色体接触结构是否合理。不同样本用同一分辨率比较前，确认每个样本支持的valid-pair数量足够。

配置修改留痕

配置修改后输出差异摘要，包括原值、新值、修改理由和复核人。尤其是bin size、过滤阈值、有效基因组大小和片段参数，会影响输出意义。不要直接修改共享模板；将专用配置放在项目目录并计算散列，后续重跑时确认使用相同文件。

如果不同lane属于同一文库，合并规则应与HiC-Pro识别样本名的方式一致。运行前在临时目录用一对文件验证lane识别，检查日志中样本总数与清单一致，避免流程将lane分别当成生物重复或把不同文库合并。

按基因组分辨率构建矩阵时，记录每种分辨率的valid-pair支持和文件大小。很细的分辨率会造成稀疏矩阵和较大磁盘需求；选择目标分辨率应考虑测序深度与研究问题，不应只因硬盘容量而无说明地降低分辨率。

每批任务的结果状态应区分未启动、运行中、计算完成待QC、QC通过和失败。只有样本级日志、矩阵完整性、主要质控指标和输入配置复核均完成，才标为可交付。程序返回成功码但矩阵不完整或参考不匹配仍属于待处理。

每个样本的统计表应与输入library_id对应，不要在合并多样本结果后丢掉文库标识。样本名规范化时保存原名称与新ID映射。

重新运行后核对结果路径与本次run_id关联，确保报告引用的矩阵不是早期运行残留。对关键输出保存校验值并在归档时复核。

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

HiC-Pro官方输入、分步与质量说明 · 来源资料

同修订官方配置模板 · 来源资料

HiC-Pro结果组织说明 · 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。