BAM转CRAM归档教程：固定参考并验证可恢复性

以坐标排序BAM为例，固定FASTA与samtools版本，显式编码CRAM、索引并恢复回BAM；核对记录和关键标签后再批量迁移。

把BAM转成CRAM可以节省存储空间，但“文件能打开”并不等于长期可恢复。解码依赖与CRAM匹配的参考序列，某些标签也可能在读取时重新计算。下面先用单个已验收的坐标排序BAM演练编码、解码和比对记录核对，通过后再按样本批量操作。原BAM在验收完成前应保留。

归档前准备	记录内容	验收要点
输入BAM	样本、排序状态、索引和版本	quickcheck通过
参考FASTA	来源、组装、校验值和索引	序列名、长度与BAM字典相符
格式选项	CRAM版本和MD/NM策略	下游兼容且标签可解释

固定参考并往返转换

先确认参考FASTA及.fai来自同一文件，并放到恢复环境可访问的位置。示例显式选择CRAM 3.0，并保存MD/NM；若项目标签要求不同，应先在样本副本上验证。不要忽略参考校验错误来强行解码。

```
set -euo pipefail
bam=/data/S1.sorted.bam
ref=/refs/genome.fa
test -s "$bam"; test -s "$ref"; test -s "$ref.fai"
samtools quickcheck -v "$bam"
mkdir cram-S1-001
root="$PWD/cram-S1-001"
samtools --version > "$root/version.txt"
sha256sum "$bam" "$ref" "$ref.fai" > "$root/inputs.sha256"
/usr/bin/time -v -o "$root/encode.resources.txt" \
  samtools view -@ 2 -C -T "$ref" \
  --output-fmt-option version=3.0 \
  --output-fmt-option store_md=1 --output-fmt-option store_nm=1 \
  -o "$root/archive.cram" "$bam" 2> "$root/encode.log"
samtools index -@ 2 "$root/archive.cram"
samtools quickcheck -v "$root/archive.cram"
/usr/bin/time -v -o "$root/decode.resources.txt" \
  samtools view -@ 2 -b -T "$ref" --input-fmt-option decode_md=0 \
  -o "$root/restored.bam" "$root/archive.cram" 2> "$root/decode.log"
samtools quickcheck -v "$root/restored.bam"
samtools view "$bam" | cut -f1-11 | sha256sum > "$root/original.core.sha256"
samtools view "$root/restored.bam" | cut -f1-11 | sha256sum > "$root/restored.core.sha256"
cmp "$root/original.core.sha256" "$root/restored.core.sha256"
```

quickcheck快速检查文件头和末尾完整性，不会扫描文件中间每个数据块；因此再做完整解码很关键。最后比较的SAM前11列包含读名、标志、坐标、CIGAR、序列和质量值，且要求记录顺序一致。哈希一致说明这些列一致，不代表所有辅助标签、头信息及下游工具兼容。

单独核对标签和随机访问

根据项目要求列出重要标签，例如RG、CB、UB、MM、ML、OQ或自定义字段，并比较其名称、类型和值；不能因为辅助标签顺序变化就判成内容不同，也不能忽略标签缺失或被重新生成。示例的store_md/store_nm保留策略尤其要和decode_md设置配套。记录任何预期中的头部PG变化。

再把CRAM、索引和FASTA复制到隔离的恢复目录，显式指定参考进行完整读取，并抽查若干基因组区间随机访问。参考应核对序列校验值，而不只看“hg38”名称。还要比较BAM与恢复文件头中的@SQ序列名和长度、@RG样本信息及读段数量；如果下游流程依赖特定辅助标签，挑选覆盖这些标签类型的记录做核对。核心字段哈希、quickcheck和随机读取验证的是不同问题，需都保留结果。

测空间后再扩批

CRAM压缩率取决于测序、参考和标签，不要按BAM大小乘固定比例估容量。先记录一份代表样本的BAM、CRAM、索引、解码副本、参考和日志大小，并观察运行期间工作盘高水位；转换时这些文件可能同时占空间。按实际挂载点、inode和配额检查可用容量。

一份样本通过往返、标签及随机访问验收后，才开始分批转换。每批保存样本ID到CRAM、索引和参考指纹的映射，并把通过、失败、待解释分开登记；抽查不能替代逐文件完整解码。遇到不一致时先暂停该样本，不覆盖原BAM，并保留差异记录供复核。每个样本使用独立目录，失败时保留命令、退出码和日志；重试用新目录，避免旧文件混入。最终交付清单包括输入与参考校验值、samtools版本、编码选项、CRAM索引、解码测试和未解决差异。只有代表性实测显示CPU、内存或I/O不足时，再调整并发与资源申请。

参考资料

Samtools view官方手册：CRAM转换与标签选项 官方文档
Samtools quickcheck官方手册：检查范围与限制 官方文档