# 按参考基因组、索引和并发数检查内存需求

分别核对STAR索引构建、私有索引载入和BAM排序的内存来源，用真实版本、参考与阶段日志估计需求。

STAR参考索引占用多少磁盘，不等于建索引或比对时需要多少内存。建索引时的FASTA大小、索引参数和线程会影响峰值；比对时若使用默认私有索引，多份并发任务可能各自载入一份基因组。先记录真实索引、STAR版本和内存设置，再按阶段测量，避免只按FASTQ或索引目录大小估算服务器。

## 先记录索引由什么生成

保留STAR完整版本、FASTA/GTF组装和来源、建索引命令、参数文件及日志。确认任务实际指向的genomeDir与预期一致，并记录索引目录大小和校验值。磁盘占用可以帮助发现误用物种或索引版本，但它不是运行时RSS。若参考基因组、注释或STAR版本变了，旧有内存测量也不一定适用。

## 区分索引构建参数与比对内存

STAR维护者参数表中，limitGenomeGenerateRAM是索引构建可用RAM的上限设置，不等于自动申请的内存；实际作业额度仍由主机、Slurm或容器控制。参数表的默认值约31,000,000,000字节。小基因组和大量contig时，genomeSAindexNbases、genomeChrBinNbits等值也可能需按官方说明调整。不能为绕过OOM就随意改索引参数，否则索引结构和比对表现会变；先确认参考输入正确，再对照适用版本和报错调整。

| 阶段 | 记录什么 | 常见误判 |
| --- | --- | --- |
| 生成索引 | FASTA大小、参数、内存额度、峰值 | 把磁盘索引大小当作RAM |
| 单样本比对 | 索引载入方式、线程、作业RSS | 仅按FASTQ压缩大小估算 |
| 多样本并发 | 同时运行数、总内存、I/O等待 | 只看单任务峰值 |

## 先测一个真实流程，再看并发

挑一份代表性FASTQ，固定索引、线程、输出参数和临时目录，先完成一次比对并记录作业峰值及结果。用/usr/bin/time -v可以补充墙钟时间和进程资源统计；在集群上还应结合作业记账，因为单进程记录未必覆盖所有子任务。核对日志完成标记、输出文件和下游验收结果，失败提前退出的低峰值不能用于配置决策。

```
/usr/bin/time -v -o star.resources.txt \
  STAR --runMode alignReads \
  --genomeDir /ref/star_index \
  --readFilesIn sample_R1.fastq.gz sample_R2.fastq.gz \
  --runThreadN 8 \
  --outFileNamePrefix pilot/sample_ \
  2>&1 | tee star.log
echo "exit=$?"
```

这是参数记录格式示意，正式命令要加入与FASTQ压缩格式匹配的读取选项及项目要求的输出。若参考索引特别大，检查基因组和注释是否符合实验设计；不要为了省内存盲目删掉序列或注释。

## 评估共享索引和总内存

STAR默认NoSharedMemory，每个任务使用私有索引副本；LoadAndKeep等模式可使用共享内存，但需要节点支持、正确预载索引并由管理员约定清理。不能把该模式当成所有容器或队列都可用。先按私有模式建立单任务基线，再逐步增加一项同类任务，观察总内存峰值、共享内存限制、读盘等待和完成时间。若要启用共享索引，需确认所有进程使用完全相同的索引版本并遵循集群管理规则。

为选型保留：索引校验值、STAR版本、构建与比对参数、单任务峰值、并发总峰值、临时盘增长及成功输出证据。并发配置依赖具体节点和共享内存设置；这些示例不是任何数据集的固定规格，最终以代表性任务的实测和集群配额为准。

## 参考资料

STAR参数说明（维护者仓库） · 官方资料

STAR README（维护者仓库） · 官方资料

GNU time资源统计手册 · 官方资料