Salmon自建参考索引：核对序列与索引设置，再测内存峰值

完整操作教程 · 参考固定 · 往返恢复验收

核对转录本与decoy角色、FASTA ID和k参数，测建库峰值，再用代表性reads验索引可用性。

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。

Salmon给新物种建索引时内存不足，先核对索引包含什么。只有转录本，与把整个基因组作为decoy加入的索引，规模不同；序列重复、错误的ID或把全基因组误当成定量目标，也不能靠增加RAM解决。本文以bulk RNA-seq的decoy-aware参考为例，先验证输入，再测一次建库峰值。

明确转录本与decoy的角色

准备来源、组装和注释版本相符的transcripts.fa与genome.fa，均为已经核对的未压缩核酸FASTA。转录本是待定量对象，decoy用于帮助识别来自相似基因组区域的错误映射。新物种没有可靠注释时，要先确认转录本生成方法；不能拿近缘物种参考直接当作同一物种的等价替代。

需要核对	记录内容	常见误判
转录本	注释、序列ID、数量与总长度	把基因组所有contig当转录本
decoy策略	部分decoy或整基因组及来源	两种策略被当成相同参考
索引参数	Salmon版本、k、完整命令	为省内存改k却不复核结果
资源	任务额度、峰值、工作盘	把FASTA压缩大小当RAM

先检查FASTA标识，再组成参考

下例采用整基因组decoy策略，可能比部分decoy需要更多建库内存。两个FASTA的首个空白前ID须唯一且不相互冲突；发生冲突时回查来源与命名规则，不随意改名破坏注释对应。awk只检查头部标识，不是完整FASTA质控。所有输出放进新的工作目录。

```
set -euo pipefail
transcripts=/project/reference/transcripts.fa
genome=/project/reference/genome.fa
mkdir salmon-index-run1
cd salmon-index-run1
salmon --version > version.txt
salmon index --help > index-help.txt 2>&1
sha256sum "$transcripts" "$genome" > inputs.sha256
awk '/^>/ {
  id=substr($1,2); n[FILENAME]++
  if(id=="" || seen[id]++) {print "invalid or duplicate ID: " id > "/dev/stderr"; bad=1}
} END {
  if(n[ARGV[1]]<1 || n[ARGV[2]]<1) bad=1
  exit bad
}' "$transcripts" "$genome"
awk '/^>/ {print substr($1,2)}' "$genome" > decoys.txt
{ cat "$transcripts"; printf '\n'; cat "$genome"; printf '\n'; } > gentrome.fa
/usr/bin/time -v -o index.resources.txt salmon index \
  -t gentrome.fa --decoys decoys.txt -i index -k 31 -p 2 \
  > index.stdout.log 2> index.stderr.log
```

需要预先安装兼容Salmon和GNU time。2线程仅为演示，应在获配范围内；k=31也不是所有读长的通用设置。当前官方说明较短reads可考虑更小k，但它影响有效匹配与敏感度，应按方法另建索引比较，不能当作不改变结果的内存开关。

区分输入错误与真实容量缺口

若在解析序列或decoy清单时退出，先核对FASTA格式、ID、清单与合并文件的对应关系。若只有Killed，按同一时间核对调度器或系统记录，区分OOM、时限和取消。容器或作业额度可能小于主机总RAM，free显示很多内存不代表当前任务可用。

整基因组decoy的构建压力确实较高时，可按官方流程评估部分decoy方案；这会改变参考构成，需要重新验证适用性与定量结果。不要在已有索引中删文件来“压缩”，也不要为建库成功随意删掉部分染色体或重复序列。

用代表性reads检查索引可用性

索引命令成功后，仍在salmon-index-run1目录中，用该物种、配对正确且来源明确的小批reads执行quant，按项目确认文库类型；自动推断也要查看日志核对。示例用于读入与结果链检查，不能凭小输入的映射率断言参考足够或生物学结论正确。

```
set -euo pipefail
salmon quant -i index -l A \
  -1 /data/pilot_R1.fastq.gz -2 /data/pilot_R2.fastq.gz \
  --validateMappings -p 2 -o pilot-quant \
  > pilot.stdout.log 2> pilot.stderr.log
head -n 3 pilot-quant/quant.sf
```

确认quant.sf可读、转录本ID和字段符合预期，并复核库类型、映射与decoy相关日志。当前Salmon1.12分支面向bulk，不将这套流程直接当单细胞alevin教程。最后保存输入校验、decoy策略、版本、峰值与索引目录；建库峰值和以后批量定量的并发资源分别测量，再决定容量与使用周期。

先决定decoy策略和转录本来源

Salmon参考由转录本和decoy组成。decoy用于竞争性识别可能更适合基因组区域的reads；全基因组decoy比部分decoy更全面，但参考体积和建索引资源通常更高。新物种若使用自己组装的转录本，要确认转录本ID、GTF、基因ID和组装版本对应；没有可靠注释时，索引构建无法弥补注释缺失。

FASTA头部ID取首个空白前字段，transcripts和genome不能存在冲突。decoys.txt每行应是gentrome中实际存在的基因组序列ID，且按Salmon要求组织。生成gentrome前后核对记录数、ID唯一、序列长度与来源。仅用awk检查标题不验证FASTA序列完整性、碱基合法性和序列对应关系。

索引参数k和读长的关系

k-mer长度会影响索引规模和可匹配片段要求。官方说明中k为奇数且不超过31，默认31；短读长可评估较小k，但会改变灵敏度和匹配表现。为减少内存而改变k不是纯硬件调整，应建立新索引并用同一批小样本比较mapping rate、有效转录本和定量差异。

同一项目的所有样本应使用同一个锁定索引、decoy列表和Salmon版本。索引目录包含多个文件，不能只复制一个索引文件到新机器。记录完整命令、线程数、临时目录、k值和索引清单散列，并通过`salmon index`完成日志确认索引生成。

建库峰值与读取顺序

整基因组decoy策略需要把transcriptome和genome合并为gentrome，建索引期间原始FASTA、合并文件和索引中间文件可能同时占用磁盘。索引峰值RAM不能从压缩FASTA大小直接估算。先对代表性组装实测峰值RSS与scratch高水位，确认作业额度覆盖后再做完整构建。

若解析阶段失败，检查FASTA格式、重复ID、空序列、序列名以及decoy清单；若系统OOM，查看调度器实际限制和索引阶段资源；若磁盘满，清点gentrome和临时文件。不要在失败后覆盖原始文件或清理日志，保留失败证据并在新目录重跑。

索引完成后的验收

记录index命令退出码、stdout/stderr、索引目录文件列表和散列。用小规模reads运行Salmon quant检查索引能否读取及转录本ID映射，再检查quant.sf中的目标ID、TPM/NumReads列和mapping统计。小样本成功不代表定量准确，但能发现索引损坏、引用ID不匹配和输入格式问题。

如果转录本注释含重复或版本后缀不一致，下游基因级汇总可能出现缺失。使用明确的transcript-to-gene映射并记录版本，不要对ID随意截断造成多个转录本合并。比对率偏低时优先检查参考、读长、物种和污染，而不是立刻修改k值或增加资源。

批量定量和资源规划

索引通常构建一次后供多个样本只读使用。为每个样本独立建立quant输出目录，固定Salmon版本、libType、读段类型和索引散列；保存每样本日志、mapping率、输入fastq散列与退出码。多个样本并发会共享索引读取并占用CPU/内存，先测单样本和低并发。

资源需求拆分为索引构建阶段与样本定量阶段。若只有索引OOM，增加定量线程没有帮助；若quant阶段I/O受限，扩RAM也未必缩短时间。咨询资源时提供FASTA数量和总长度、decoy策略、k、索引峰值、样本读段量、并发及工作盘峰值。

FASTA与注释ID逐项对应

转录本ID应能对应GTF/注释中的transcript_id和gene_id，decoy ID对应基因组FASTA首字段。生成gentrome前比较两个ID集合、重复项和版本后缀，并随机核对若干转录本序列来源。只检查header不能验证序列完整性或注释匹配。

全基因组decoy策略将transcriptome和genome合并，decoys.txt每行一个基因组序列首字段且这些记录须在gentrome中存在。序列名冲突应回查参考来源和命名空间，不要临时改名后忘记更新GTF映射。

decoy策略和k值会改变索引

decoy用于让相似基因组区域与转录本竞争，减少不适当分配。全基因组decoy更全面但索引构建通常更耗内存和磁盘；部分decoy也需依官方流程准备。记录策略、来源和索引目录散列，不将两种策略的结果视为相同。

Salmon索引的k为奇数且不超过31，默认31。较短读段可评估较小k，但k影响有效匹配和灵敏度，不是单纯的硬件旋钮。若要比较，使用同一reads重新定量并查看mapping、转录本和基因级差异，建立新索引版本。

建索引时测量峰值

输入FASTA、合并gentrome和索引中间文件可能同时占用工作盘；RAM峰值不能从压缩FASTA大小简单推算。用代表性参考实测RSS、墙钟和scratch高水位，按作业实际额度规划。检查器或awk脚本也会读取输入，需考虑预检本身的I/O。

重复ID、decoy不匹配或格式错误先修输入；只有调度器确认OOM且输入已通过检查，才评估更大额度。磁盘满则规划临时空间，不要覆盖唯一的参考文件。失败保留stderr和资源记录，在新索引目录重跑。

完成后的索引与quant验收

保存索引命令、Salmon版本、日志、文件清单和散列。用一份小型reads运行quant验证索引可读取、转录本ID能对应并查看mapping统计；小样本成功不等于定量准确，但可发现损坏索引和ID不匹配。

所有样本固定同一个索引、libType、读段布局及Salmon参数，每个样本独立输出并保存FASTQ散列与日志。检查quant.sf中的ID、NumReads、TPM和有效长度。跨样本差异推断需适当统计模型，不以TPM直接代替统计检验。

资源需求与错误分流

索引构建与样本quant是不同资源任务。全基因组decoy主要增加建库负载；quant另受FASTQ规模、索引读取、线程和并发影响。分别测单任务峰值和低并发吞吐，不用一个阶段的资源推算另一个阶段。

mapping率异常低时先查物种、组装、读长、R1/R2、libType和污染；索引OOM查看cgroup/作业额度；读取失败查目录完整性。每次改变转录本版本、decoy策略或k都更新索引版本号并绑定相应quant结果。

索引生成后将gentrome.fa、decoys.txt和参考注释文件的散列写入索引manifest。迁移索引目录时复核目录内文件数量和散列；不要仅凭文件夹可打开就认定复制完整。索引更新与旧版本并存，便于定量结果回溯。

对每个样本在quant前核对FASTQ顺序、样本名、读段数和索引版本。若使用外显子捕获或非标准读段类型，确认Salmon模式适用。多样本并行时给每个样本独立目录，避免日志和quant文件互相覆盖。

正式差异分析前查看样本mapping统计、异常转录本数量和样本间表达分布，发现离群样本时回查实验和读段质量。建索引成功只是参考准备完成，不能替代样本级定量QC。

由于全基因组decoy索引可能较大，索引长期存放在只读目录并按版本管理。每个quant结果在manifest中写明其索引散列，避免更新索引后误用旧样本结果进行比较。

建索引和quant两种任务均保留stderr、资源记录和结束码。目录应标明reference版本、decoy策略和k值，让后续使用者不必从日志猜测设置。

若参考注释后来修订，应重新核对ID映射并评估是否重建索引；仅文件名保持不变不能说明FASTA内容相同。

索引构建之后不要直接删除输入FASTA和映射文件；如按项目策略移除，应至少归档校验值、版本来源和重建说明。这样才能追溯索引对应的参考。

读取清单应包括FASTQ校验值、读段布局和索引散列。这样mapping率异常时能判断是输入样本改变，还是索引版本和参数变动造成。

生成索引时还要记录运行机器、作业限额和临时目录，便于解释不同环境下的资源峰值。迁移后如果重建索引，作为新版本保存并更新quant manifest，不能沿用旧散列。

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

Salmon1.12官方：decoy参考、索引及quant接口 · 来源资料

Salmon官方推荐参考准备教程 · 来源资料

GNU time统计范围 · 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。