RSEM多样本定量实操：组织参考、输入与任务资源

完整操作教程 · 参考固定 · 往返恢复验收

固定RSEM参考、文库方向和样本清单，保存独立结果与日志，再按真实STAR任务峰值规划批次。

运行边界：定量结果依赖参考、链特异性和输入样本对应关系；演示参数需按实际软件版本核对。

RSEM多样本定量，先统一参考与输入路线，再安排批次。转录本结果能否合并，取决于注释和ID是否一致，不是把几个results文件按行拼起来。下面以已质控的短读长双端bulk RNA-seq为例，使用RSEM调用STAR完成比对与定量；已有比对文件的路线需另查输入约束。

准备相互匹配的参考和样本清单

基因组FASTA与GTF应对应同一组装，染色体名称一致，GTF中基因和转录本ID可追溯。记录RSEM、STAR与参考版本，保留输入校验值。没有可靠注释时，应先确定转录本来源和基因映射，不把别的物种索引改名后直接使用。

输入	本例要求	容易混淆
参考	同一组装FASTA与GTF	基因组名相同但注释发布不同
样本表	唯一sample_id、R1、R2路径	lane被当成独立生物重复
文库方向	确认none、forward或reverse	未知方向直接当非链特异
输出	每样本独立目录与前缀	多个任务覆盖同名临时文件

只准备一次参考，单独记录建库资源

需要已安装兼容RSEM、STAR和GNU time。将示例路径换成真实、已验证参考，输出根目录必须不存在。线程2只是试跑档位；STAR索引构建和以后定量的峰值分别测量，不能以一个数字代替全部阶段。

```
set -euo pipefail
genome=/project/reference/genome.fa
gtf=/project/reference/annotation.gtf
test -s "$genome" && test -s "$gtf" || exit 2
case "$genome$gtf" in *","*|*" "*) exit 2;; esac
mkdir rsem-batch1
mkdir rsem-batch1/reference rsem-batch1/samples
rsem-calculate-expression --version > rsem-batch1/version.txt
STAR --version > rsem-batch1/star-version.txt
sha256sum "$genome" "$gtf" > rsem-batch1/reference.sha256
/usr/bin/time -v -o rsem-batch1/reference.time.txt \
  rsem-prepare-reference --gtf "$gtf" --star -p 2 \
    "$genome" rsem-batch1/reference/ref \
  > rsem-batch1/reference.stdout.log 2> rsem-batch1/reference.stderr.log
```

按清单顺序运行，先验一个样本

准备无表头、Tab分隔的samples.tsv，每行是sample_id、R1绝对路径、R2绝对路径，ID仅用字母数字、下划线或连字符；本例路径避免空格、逗号、Tab或换行，尤其逗号在RSEM中可能表示文件列表。先只放一个代表样本，通过后再增加行。仍从创建rsem-batch1时的项目目录执行下面代码。

```
set -euo pipefail
: "${STRAND:?■■■■■none■forward■reverse}"
case "$STRAND" in none|forward|reverse) :;; *) exit 2;; esac
test -s samples.tsv
while IFS=$'\t' read -r sample r1 r2 || [[ -n "${sample:-}" ]]; do
  r2=${r2%$'\r'}
  [[ "$sample" =~ ^[A-Za-z0-9_-]+$ ]] || exit 2
  [[ "$r1" != "$r2" && -s "$r1" && -s "$r2" ]] || exit 2
  [[ "$r1" = /* && "$r2" = /* ]] || exit 2
  case "$r1$r2" in *","*|*" "*) exit 2;; esac
  gzip -t "$r1" "$r2"
```

```
  out="rsem-batch1/samples/$sample"
  mkdir "$out"
  /usr/bin/time -v -o "$out/resources.txt" \
    rsem-calculate-expression --paired-end --star --star-gzipped-read-file \
      --strandedness "$STRAND" -p 2 "$r1" "$r2" \
      rsem-batch1/reference/ref "$out/$sample" \
    > "$out/stdout.log" 2> "$out/stderr.log"
  test -s "$out/$sample.genes.results"
  test -s "$out/$sample.isoforms.results"
done < samples.tsv
```

本例限定gzip FASTQ，压缩完整性不代替配对与序列质控。当前源码中STAR压缩输入参数为star-gzipped-read-file，旧网页示例与本机帮助不一致时，先核对安装版本对应接口，不盲目复制。已有SAM/BAM需满足转录本坐标、同一read比对聚集、mate相邻及SEQ/QUAL等条件，不能直接把基因组坐标排序BAM送入另一条路线。

检查结果含义，再扩大并发

逐样本查看genes.results与isoforms.results，核对ID集合、字段、数值和比对日志，异常缺失与低映射先解释。expected_count是模型期望计数，可能为非整数；TPM、FPKM与计数含义不同，后续按分析工具支持的输入方式导入，不直接按行号拼表或把TPM当原始计数。

输出文件非空只是基础检查。保留每个样本的原始日志、参考版本及成功或失败状态，重复ID会因目录已存在而停止，应回查清单，不给失败任务补一个完成标记。比较不同批次前确认注释、模型和文库设置一致。

先顺序测代表样本，再验证计划并发。官方STAR路线中，各任务会有自己的基因组内存副本，磁盘上一份索引不代表RAM只算一份。另记临时BAM、结果和输入共存空间；合理方法仍有资源缺口时，用实测任务组峰值与整批窗口选择环境。

先把技术重复与生物学样本分开

样本清单中的sample_id应对应实验设计里的生物学样本，而不是测序lane或FASTQ文件。一个样本有多个lane时，要按测序方案决定合并文件还是把文件列表按RSEM当前接口传递；不要把lane当成独立生物重复输入差异分析。清单宜另存一份样本元数据表，包含分组、批次、文库类型、建库日期和预期配对关系，主分析只用稳定的sample_id关联，不从文件名猜分组。

跑任务前检查每个R1/R2都存在、压缩文件可完整读取、路径唯一，且清单没有重复样本或配错mate。抽查FASTQ标题中的read标记和条数；gzip -t只验证压缩流完整，不能证明read配对、质量或污染合格。若一组样本的lane需合并，保留原lane到样本的映射和哈希，合并后再核对记录数与顺序。对于少数单端样本，不要直接混进此双端命令，应按RSEM单端输入规则另行建模并记录。

链特异性应来自实验方案而非猜测

STRAND值应根据试剂盒、文库构建记录或实验负责人确认。RSEM支持none、forward、reverse等设置，但不同软件对“正向/反向”的描述可能以不同read或链定义为参照，不能仅凭“stranded”一词映射。先查对应建库类型的官方说明，并在少量已知样本上观察比对日志、转录本覆盖方向或已知表达基因是否符合预期。方向不清时先标记未确认，不用多跑几个选项后挑最好看的结果。

文库方向设置错误可能使比对率或有效定量发生系统变化，输出文件仍可能生成。因此，判断成功需同时看日志中的输入读段数、有效比对比例、估计率和异常警告，并与质控报告交叉核对。一个样本出现显著低映射时先查参考物种、接头污染、read长度、参考ID及样本身份，不通过调整方向强行“救回”所有样本。

理解RSEM与STAR的中间文件和并发成本

参考构建只做一次，但需要的磁盘和内存应按实际FASTA/GTF与STAR索引单独测量。定量阶段启用STAR时，每个并发作业会有自己的基因组内存副本；因此索引在盘上只存一份，不意味着同时跑多个样本只占一份RAM。用一个样本测定量峰值后，逐步增加并发，记录每作业内存、整组峰值、CPU使用、工作盘读写和任务等待时间。

RSEM临时目录默认在任务结束后删除；如果需要保留中间文件，应明确设置临时目录和保留选项，估算输入、临时BAM、结果及日志同时存在的峰值空间。临时路径应位于足够容量的目标文件系统，而不是依赖默认系统盘。保留中间文件会增加空间并可能包含较大的未压缩比对数据；只在排错或项目审计需要时开启，完成后按批准流程归档或清理。

线程参数不是整机资源需求。多进程并发时，线程总数、STAR的内存副本、文件系统吞吐和调度器限制要一起考虑。线程从2增到8不一定按比例缩短墙钟，尤其当磁盘读写或共享节点竞争成为瓶颈。测试配置时固定参考、样本、版本和参数，每次只改变一个因素，并保存`/usr/bin/time -v`或调度器报告，避免把不同条件下的运行时间直接比较。

检查genes.results和isoforms.results的字段

RSEM会输出gene与isoform层结果表。打开表头确认ID列、长度、effective_length、expected_count、TPM、FPKM及可选的后验估计字段，不能假定不同选项运行出的列完全一样。expected_count来自模型估计，可能不是整数；TPM/FPKM是归一化表达度量，不能把它们当作原始整数计数直接输入要求count的差异检验。选择下游工具时依据其统计模型与数据格式文档，而不是只看文件名里写着results。

所有样本使用同一参考与注释，才能按稳定的gene_id或transcript_id对齐矩阵。先验证ID是否重复、集合是否异常变化、行数是否合理，再构建矩阵。样本间缺少某些转录本行时，应依据工具输出规则和统一ID表处理，不可按行号合并。若注释版本改变导致ID集合改变，应视为不同的定量批次，不要把其结果静默拼在同一矩阵。

定量结果是估计，不等于生物学结论。检查RSEM日志、样本相关性、分组与批次设计、异常样本和潜在混杂；差异分析还需恰当的重复数、设计矩阵和统计流程。若从RSEM输出构建表达矩阵，先明确用expected_count还是TPM等数据，并按下游方法要求做过滤与变换。多种表达单位不得在同一矩阵里混用。

失败时按阶段定位而不是重跑整批

现象	先检查	处理动作
参考构建失败	FASTA/GTF组装、染色体名称、磁盘和日志	修正参考并创建全新索引目录
STAR参数报错	本机RSEM版本、STAR版本和对应帮助	先对一个样本确认压缩参数接口
样本读段数异常	FASTQ完整性、R1/R2清单和样本身份	隔离该样本，不继续合并矩阵
低有效比对	物种、参考注释、方向、read长度及污染	查明原因后再决定是否重跑
任务被杀或磁盘满	cgroup/调度器日志、临时盘和并发	先保留日志，降低并发后单样本重试

每个样本独立目录和状态表可让失败重跑不覆盖已成功结果。开始前确认输出路径不存在；再次运行前把旧结果移入有记录的隔离目录，而不是直接删除整个批次。工作清单记录sample_id、输入哈希、参考哈希、软件版本、STRAND、线程、状态和输出路径。最终样本数与设计表一致，结果字段完整，且下游能够读取，才把批次标记为完成。

用代表性样本估算整批运行窗口

选择一个数据量接近中位数和一个偏大的样本分别试跑，记录读段数、压缩字节、参考大小、墙钟、CPU时间、峰值内存和临时盘峰值。单样本耗时不能简单乘样本数再除以并发数就当作完工承诺，因为样本复杂度、队列等待、索引、磁盘争用和失败重试都会改变总时长。可以先将线性估算作为排期基线，再用真实小批次测量修正。

并发测试从1开始，再增加到2或计划上限；若总耗时没有改善，查看CPU利用、读写和内存曲线后再决定。多个STAR任务可能因重复占用基因组内存导致互相竞争。此时减少并发可能比增加线程或内存更有效。记录“每样本资源”和“同时任务资源”两个口径，并把排队与人工检查时间单独列出。

当已有校内集群额度足够且软件环境可复现时，先用现有资源完成代表样本试跑。如果长期任务因队列、内存或工作盘受限，再基于日志讨论新增环境。咨询服务器资源时准备RSEM/STAR版本、参考组装、样本数与read规模、计划并发、单样本峰值及项目周期；不需要提交完整FASTQ就能先判断工作盘、内存和计算并发的主要需求。

从多个样本生成矩阵时保留结果来源

若用RSEM配套脚本从多个genes.results或isoforms.results生成矩阵，先确认输入全部处于同一层级，并按项目样本表的固定顺序列出文件。矩阵生成只负责按RSEM结果组织字段，不会检查分组设计、重复是否独立或批次效应，也不自动替代差异分析工具的输入审核。保存文件清单和命令，检查矩阵列名与样本表完全对应，防止路径顺序错位。

RSEM配套矩阵可用于其文档支持的后续分析场景；切换到DESeq2、edgeR等工具时，需按其官方指南使用适合的计数定义和模型。不要因为列名叫counts就跳过分布、整数要求、有效长度或不确定性处理的核查。表达单位与工具兼容性是分析设计的一部分，应在正式多样本批次开始前确定。

配套练习包

本PDF对应的ZIP包含本文代码块、主题练习数据与样本清单模板、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

RSEM官方参考与定量工作流 · 官方资料

RSEM参考构建接口 · 官方资料

RSEM定量输入与结果字段 · 官方资料

RSEM当前STAR压缩输入参数定义 · 官方资料

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。