StringTie多样本转录本合并：清单、参考与资源试跑

完整操作教程 · 参考固定 · 往返恢复验收

用样本表检查GTF清单，建立共同转录本目录，再按统一注释重估各样本并区分结构与表达。

运行边界：merge只建立共同转录本结构；表达重估和差异分析需另行完成。

StringTie合并多个样本，输入是各样本已经验证的转录本GTF，不是把BAM或表达矩阵拼在一起。merge模式建立共同的转录本结构集合，不能替代各样本表达量估计。下面从样本清单生成文件列表，运行一次可追溯合并，再说明怎样把共同注释用于后续定量。

先确认这些GTF可以一起使用

所有样本应对应同一基因组组装，染色体名称、组装软件模式和主要处理条件有记录。不要把不同物种、不同参考坐标或未经核对的短读长与长读长结果混合。每份GTF关联真实样本与原BAM；样本ID唯一，不要求不同样本的局部MSTRG编号全局唯一。

阶段	输入	输出及用途
逐样本组装	对应样本的已验证比对	各样本转录本GTF
merge	GTF清单及可选参考注释	共同结构目录，不是合并计数
重新估计	各样本原BAM与共同GTF	统一结构下的样本表达结果
下游分析	符合目标工具要求的数据	另行完成质量与统计检查

从样本表生成不漏项的列表

在Linux项目目录准备samples-gtf.tsv，表头为sample_id和gtf，两列用Tab分隔，gtf为绝对路径。本例路径不含空白字符。下列标准库脚本检查空表、重复样本、重复路径与空文件，然后新建merge-list.txt；不会覆盖旧列表，也不证明GTF生物学内容正确。

```
import csv
from pathlib import Path

with open("samples-gtf.tsv", encoding="utf-8-sig", newline="") as f:
    rows = list(csv.DictReader(f, delimiter="\t"))
if not rows:
    raise ValueError("■■■■■■■")
ids = [r["sample_id"].strip() for r in rows]
if any(not x for x in ids) or len(ids) != len(set(ids)):
    raise ValueError("■■ID■■■■■")
paths = []
for row in rows:
    raw = row["gtf"]
    p = Path(raw)
    if any(c.isspace() for c in raw) or not p.is_absolute():
        raise ValueError("■■■■■■■■■■■")
    p = p.resolve(strict=True)
    if not p.is_file() or p.stat().st_size == 0:
        raise ValueError("GTF■■■■■")
    paths.append(str(p))
if len(paths) != len(set(paths)):
    raise ValueError("■■GTF■■■■■")
with open("merge-list.txt", "x", encoding="utf-8", newline="\n") as f:
    f.write("\n".join(paths) + "\n")
print("listed samples", len(paths))
```

在新目录运行一次合并

需要已安装兼容StringTie和GNU time，guide.gtf是与输入组装相匹配的已验证参考注释。没有可靠参考时应按研究设计另选无参考合并，不用无关物种注释凑齐参数。先固定列表与过滤设置，不为减少资源随意提高长度、覆盖、FPKM或TPM阈值。

```
set -euo pipefail
guide=/project/reference/guide.gtf
test -s "$guide" && test -s merge-list.txt || exit 2
mkdir stringtie-merge-run1
stringtie --version > stringtie-merge-run1/version.txt
cp merge-list.txt stringtie-merge-run1/inputs.txt
sha256sum "$guide" > stringtie-merge-run1/guide.sha256
/usr/bin/time -v -o stringtie-merge-run1/resources.txt \
  stringtie --merge -G "$guide" \
    -o stringtie-merge-run1/merged.gtf merge-list.txt \
  > stringtie-merge-run1/stdout.log 2> stringtie-merge-run1/stderr.log
test -s stringtie-merge-run1/merged.gtf
```

文件非空只是基本检查。核对日志、有效GTF记录、染色体范围和转录本结构，确认没有意外漏掉样本或参考。合并可重新分配ID，要保存最终GTF与清单版本；不能仅凭原样本相同的MSTRG编号，把两个局部转录本视为同一个全局对象。

共同结构确定后，重新估计各样本

各样本用自己的基因组坐标排序BAM对共同注释重新估计，避免把不同样本原有的局部目录直接拼成表达矩阵。下例只示范一份已经确认fr-firststrand的短读长文库，因此使用--rf；其他建库方向按实际修改，并保留项目既定参数。BAM还需满足相应剪接标签与输入要求。

```
set -euo pipefail
mkdir stringtie-reestimate-S01
stringtie /project/bam/S01.sorted.bam \
  -G stringtie-merge-run1/merged.gtf -e --rf -p 2 \
  -o stringtie-reestimate-S01/S01.gtf \
  -A stringtie-reestimate-S01/gene-abundance.tsv \
  > stringtie-reestimate-S01/stdout.log 2> stringtie-reestimate-S01/stderr.log
```

2线程为演示设置，在获配范围内使用。按原样本清单逐一登记重估结果，确认字段、ID和需要的表达范围；TPM、FPKM不是原始计数，进入差异分析前使用对应工具支持的整理方式。合并完成本身不产生差异结论。

资源试跑分开记录merge与逐样本重估：前者受目录规模和结构复杂度影响，后者还要读BAM。先测代表任务与工作盘共存，再验证正式集合；不要拿GTF文件数量线性推内存，也不把增加CPU当作读取或注释错误的解决方式。

样本GTF的来源与参考坐标要逐份对应

每个GTF都应能追溯到sample_id、原始比对文件、基因组组装、StringTie版本及组装参数。合并前检查所有文件使用同一染色体命名和坐标系统，确认GTF行数非零、九列格式完整、属性字段可解析，并抽查转录本与exon记录的ID关联。若个别样本使用了另一版参考或不同染色体前缀，不能只改文件名或批量替换chr字符串后直接合并；需确认坐标和注释确实可比。

局部转录本ID可能基于样本或比对文件名生成。跨样本看到相同MSTRG编号并不代表是同一转录本，不同编号也不必然代表结构不同。StringTie的merge步骤会构建非冗余的共同转录本集合，输出统一GTF供后续重估；它不会把所有样本的TPM、FPKM或count列合并起来。把表达矩阵直接拼接不是merge模式。

理解参考注释与过滤阈值的作用

提供-G时，guide GTF/GFF会参与共同集合构建；它应与比对使用的组装一致。若没有可信参考注释，应按分析目标使用无参考合并路线，不能把不匹配的注释只为保留已知基因名称而硬塞进去。merge的-m、-c、-F、-T和-f等阈值会影响哪些转录本进入共同集合。默认值也是分析决策的一部分，先记录版本默认设置，不为减少文件大小或缩短运行时间而随意提高门槛。

若需要调整过滤，先明确依据：目标是保留低表达候选、只关注较高覆盖转录本，还是为了特定实验设计。用代表样本和小批次比较GTF的转录本数、外显子结构、参考匹配比例及已知控制基因，记录阈值改变带来的增减。过严阈值可能丢失低表达但真实的转录本，过宽则可能保留噪声。没有普遍适用的一组阈值，参数要结合设计与下游目的说明。

样本清单检查后再运行merge

Python脚本确保sample_id唯一、GTF路径为存在的非空文件、路径无空白且没有重复。进一步还应检查行首是否误带空格、表头是否准确、文件是否为预期UTF-8/Tab格式，以及每个样本的GTF是否来自同一分析批次。保存清单原件与生成的merge-list.txt，并对两者计算校验和。清单生成成功只证明路径可读，不证明文件所用参考和参数兼容。

merge命令在新的尝试目录运行，记录StringTie版本、guide哈希、输入列表哈希、标准输出、错误输出与资源测量。若任务失败，不在已存在的输出上继续覆盖；用新目录重试并记录变更。对于超大样本集合，可先以代表子集测试格式和参数，再扩大到完整集合，但子集的GTF只能用于接口验证，不应拿来作为最终全样本共同注释。

合并结果的结构验收

merged.gtf非空后先检查GTF九列、feature类型、start/end是否合法、strand取值及属性中的gene_id、transcript_id。检查每个transcript是否有对应exon记录、exon顺序与坐标合理、没有超出参考contig长度。若提供参考注释，抽查reference_id等对应关系；未匹配到参考不自动意味着错误，也可能是新结构，但需要回到输入读段和注释版本核验。

用gffcompare或项目认可的工具可汇总匹配与新结构类别；应按工具版本和参数记录结果，并把“新转录本候选”与“已确认新基因”区分开。结构注释与表达量不是同一个证据。合并结果中没有TPM/FPKM/count数值是预期行为，不要因缺少这些列就把表达表从原GTF强行拼入。最终GTF是后续重估的结构参考，需作为一个有版本号的独立资产保存。

对每个原始样本重新估计表达

用merged.gtf作为每份原始BAM的-G输入，按实际文库方向设置--rf、--fr或非链特异选项。原始BAM应对应同一参考且满足StringTie的剪接信息要求。每个样本保存独立输出和日志，核对命令使用的BAM与sample_id映射。不能把另一个样本的BAM重复跑多次并改输出名来伪造完整样本集，也不能仅用合并GTF估计一个全队列表达量。

-e模式限制估计到提供的转录本集合，不再从该样本生成新的结构；这使不同样本在同一转录本目录下评估表达，但并不自动产生差异检验统计量。若后续用Ballgown、prepDE或其他工具，按照该工具对GTF、覆盖表、读长与计数单位的要求准备输入。StringTie或prepDE产生的计数需理解其估算方法和适用统计流程，不能与RSEM原始字段混为同一种计数。

对比重估前后的样本覆盖、表达分布、缺失值和参考转录本数；确认所有样本共同GTF哈希相同。样本之间的文库和批次设计仍需进入下游模型。合并流程帮助建立共同结构坐标，不会消除批次效应、补足生物学重复或替代统计设计。

估算merge与重估的资源并区分I/O

merge阶段主要读取每份GTF并维护待合并结构集合，工作量受样本数、每份GTF行数、注释复杂度和过滤条件影响。逐样本重估阶段则读取BAM和共同GTF，I/O与BAM体积、索引状态、线程及并发相关。分别用小规模代表集合和一个样本估算，不把merge耗时直接当作整批运行耗时，也不以GTF文件字节数直接预测峰值内存。

测量时记录峰值RSS、CPU时间、墙钟、输入输出大小、文件系统位置、同一时段其他任务和运行版本。若GTF合并很快但BAM重估缓慢，瓶颈可能在比对文件读取或共享存储；若内存接近调度上限，先降低并发并检查是否把多个流程同时放在同一节点。针对资源问题再调整作业请求，不用更改结构阈值来掩盖机器容量不足。

故障排查与可交付的复现记录

问题	核验项	处置
输入文件列表为空	样本表表头、分隔符、路径和文件权限	修复清单并重新生成列表
merge报GTF格式错误	九列、坐标、属性引用和文件截断	隔离异常样本，不忽略报错继续合并
输出转录本数骤减	参考版本、默认/显式过滤阈值、输入GTF范围	小批比较参数影响并说明设计依据
样本表达表字段不一致	StringTie版本、共同GTF哈希、模式参数	确认全部样本由同一公共GTF重估
任务超时或被杀	作业资源、输入规模、读盘速度和并发	分别测merge与重估后调整计划

最终复现包建议保留样本表、原GTF清单、每个GTF及guide哈希、StringTie版本、merge参数、merged.gtf、样本重估命令和状态表、资源记录及异常说明。GTF可能包含样本路径或内部标签，分享前按机构规则脱敏。只有所有预期样本都完成、共用同一merged.gtf、字段符合下游要求且结果可重新读取，才宣布这一批处理结束。

若单机或现有集群在经过代表任务测量后，确实无法满足重估并发、内存或工作盘需求，再比较专用计算环境。计算资源不会自动替代参考版本核对、GTF结构审查和差异分析设计。咨询时准备样本数量、GTF行数、BAM大小、StringTie版本、目标并发、项目时限和现有错误日志摘要即可，不必提供完整测序数据。

prepDE与差异分析输入不要混为一谈

StringTie官方流程还提供prepDE，从各样本在-e模式下生成的覆盖信息估算gene/transcript count矩阵。该脚本的估算涉及覆盖、转录本长度和平均read长度；参数和脚本版本应记录，读长不同时不能照搬一个默认值。它生成的矩阵仍需按DESeq2、edgeR等下游工具的要求检查样本列顺序、基因ID、过滤规则和实验设计，不能把TPM、FPKM列改名为count矩阵。

差异分析需要独立生物学重复与合理的模型设计。StringTie merge只统一可比较的转录本结构，不能把技术lane变成重复，也不会自动解决不同批次的实验条件。先由实验设计表建立分组和协变量，再选择适合的计数整理方法和统计模型；教程中的转录本结构流程不替代这一步。

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

StringTie官方：merge与统一转录本定量流程 · 官方资料

Python官方CSV与TSV读写 · 官方资料

GNU time资源口径 · 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。