Slingshot多任务实操：固定输入、拆分任务与恢复失败批次

完整操作教程 · 参考固定 · 往返恢复验收

按独立分析任务固定降维、分群和起点，用单任务入口保存拟时序、版本及失败证据，避免随意拆细胞改变轨迹。

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。

Slingshot同时跑多个样本或方案，先决定哪些分析在生物学上应独立。它依据降维坐标和分群推断谱系及曲线，不能为省内存随机把同一轨迹的细胞切块，再将各块拟时序拼成全局结果。多任务管理应围绕已确定的分析设计。

为每个任务冻结输入合同

每行记录任务ID、输入校验值、细胞范围、降维维度、分群版本、起点依据和种子。不同样本单独建轨迹与整合后共建轨迹回答的问题不同；独立运行得到的拟时序尺度也不能直接相加或当作同一时间轴比较。

材料	本例内容	验收
rd	细胞×降维维度数值矩阵	行名唯一，全部有限
cluster	与rd同顺序的命名分群向量	名称与细胞逐项一致
start	有生物学依据的起点簇	存在于本任务分群
任务记录	软件、参数、种子和输出路径	失败与成功分开登记

给一个任务准备完整入口

将可信输入存成含rd、cluster、start的R列表RDS，保存下例为fit_one.R。前提是兼容R与slingshot已安装。approx_points=150是曲线近似分辨率的明确设置，按当前官方建议范围使用；它不限制拟时序只有150种取值，也不能代替输入质量验证。

```
args <- commandArgs(TRUE)
stopifnot(length(args) == 3)
input <- args[1]; out <- args[2]; seed <- as.integer(args[3])
stopifnot(!is.na(seed))
x <- readRDS(input)
stopifnot(is.matrix(x$rd), is.numeric(x$rd), all(dim(x$rd) > 0), all(is.finite(x$rd)))
ids <- rownames(x$rd)
stopifnot(length(ids) == nrow(x$rd), !anyNA(ids), all(nzchar(ids)), !anyDuplicated(ids))
stopifnot(identical(names(x$cluster), ids), !anyNA(x$cluster))
stopifnot(length(unique(x$cluster)) >= 2, length(x$start) == 1,
          !is.na(x$start), x$start %in% x$cluster)
stopifnot(!dir.exists(out), dir.create(out, recursive=TRUE))
library(slingshot)
set.seed(seed)
fit <- slingshot(x$rd, clusterLabels=x$cluster,
                 start.clus=x$start, approx_points=150)
pt <- slingPseudotime(fit)
stopifnot(nrow(pt) == nrow(x$rd))
saveRDS(fit, file.path(out, "slingshot.rds"))
write.csv(data.frame(cell=ids, pt, check.names=FALSE),
          file.path(out, "pseudotime.csv"), row.names=FALSE)
writeLines(capture.output(sessionInfo()), file.path(out, "session.txt"))
writeLines(as.character(seed), file.path(out, "seed.txt"))
```

在项目根目录运行下面命令，输出目录和相邻日志都使用本次唯一名称。GNU time只记录这次进程的资源观察，不把一个小任务的峰值外推到全部谱系或正式细胞量。

```
set -eu
/usr/bin/time -v -o pilot-A.time.txt \
  Rscript --vanilla fit_one.R inputs/pilot-A.rds runs/pilot-A 101 \
```

```
> pilot-A.stdout.log 2> pilot-A.stderr.log
```

先验收谱系与细胞对应

核对任务退出、输出对象、细胞行数和谱系列数，再绘制簇连接与曲线检查起点和分支。未分配到某谱系的细胞可能有NA，不能一律补零；也不因曲线平滑就认定细胞真实发育方向已经证明。起点与分群改变需另立方法记录。

结果应保存完整对象及输入对应，不能只交一列无细胞ID的拟时序。缺维度、同顺序检查失败或非有限坐标时先修输入；只有明确进入曲线拟合后才讨论资源。不要用删掉异常细胞的方式让任务完成，却不记录筛选依据。

按任务恢复，再增加并发

先逐任务建立成功基线，记录墙钟、RAM和结果大小。需要并发时只对已独立定义的任务安排外层进程，同时核对底层数学库线程和作业额度；不同任务资源差异较大时按实测分组，不只平均分配任务数。

失败后确认原进程已结束，保留日志，用新尝试目录重跑对应任务。输入、参数和版本未变且内容验收通过的任务才复用；目录存在不等于成功，也不从残留曲线文件任意续算。最后按任务清单汇总成功、失败及未验收项，不把最后完成的一项当作全批完成。

任务清单至少写明任务名、输入文件、根簇、种子、输出位置与验收状态。根簇应来自已知生物学或实验信息；改换根簇会影响轨迹方向，不能仅为了让拟时序与预期一致而反复挑选却不记录。

运行前先做一次小样本闭环

不要一上来用完整对象试内存。先从真实任务中选取一个有代表性的样本，保留它的真实分群结构和候选起点；小样本用于检查接口、日志和输出格式，不用于替代完整数据得出生物学结论。创建输入后立即记录R版本、Slingshot版本、降维方法、输入文件SHA256和细胞数。每次改了降维、聚类或根簇，都应该生成新的运行编号，避免旧目录里的图和表被误认为新结果。

Slingshot输入的行顺序必须同时对应坐标矩阵和cluster向量。建议在写入RDS之前，以细胞ID显式对齐，而不是只检查两个对象行数相同。坐标列代表哪些维度、是否已缩放、是否由同一个对象导出，也要写进清单。若坐标包含NA/Inf、簇标签缺失、重复细胞名，应该在运行前停止；不要等拟合结束后再猜测异常来自算法还是输入。

开始小样本前，先用项目记录解释根簇为什么代表候选起点，例如采样时间、已知标记或实验设计。只有“图上看起来最早”不足以单独证明起点。若没有可靠起点，可先按方法允许的无根分析探索拓扑，再把方向解释留到有证据之后；不要为了得到符合预期的箭头反复换start.clus。

读懂谱系输出并形成可复核的结果

运行结束后逐项检查曲线数、各曲线包含的细胞数、拟时序矩阵维度以及每个谱系中NA的比例。不同谱系覆盖细胞不同，NA可能表示未分配到该谱系；它不是零时刻，也不应无条件删除。将拟时序与细胞ID写成整洁表时，保留每个谱系独立列，并额外记录是否纳入该谱系，便于下游作图时避免把未分配细胞放在起点。

图至少要能看出原始簇、最小生成树连接、平滑曲线和起点标记。把曲线叠在同一套降维坐标上，检查是否穿过明显不相关区域、是否将互不连通簇硬连起来、分支方向是否受少数细胞支配。可做合理的参数敏感性检查，但每个候选方法的结果应并列呈现，不能只挑一条最符合假设的曲线。

拟时序是由输入坐标、聚类和模型产生的相对排序，不是实验时间的直接测量。不同样本分别拟合时，数值范围、曲线密度和分支结构不能直接相加；若要比较样本，先定义共同参考、匹配谱系和统计方案。报告中应写明比较对象和方法假设，避免把“排序靠后”写成细胞一定经历更长时间。

为多任务设置明确的完成状态

为每个任务使用独立目录，并设置planned、running、failed、validated等状态。只有进程退出码为零且必要对象、表格、日志、谱系数与细胞对应全部通过检查，才标记validated。对于中断任务，先确认调度器和子进程均已退出，保存错误末尾及系统资源记录，再决定重跑。Slingshot不是可以任意按细胞块拆开后拼接成一条全局轨迹的逐行工具；若内存不足，先排查对象副本、输入维度、内存限制和并发，再评估有科学依据的分析设计调整。

建议把每次任务的sessionInfo、调用参数、输入散列、输出散列和绘图脚本一并归档。重跑同一任务时先对比输入与软件版本；若输入变化，旧输出即使可以读取，也不应被无说明地复用。若版本升级，应在小样本上重新跑并对比分支、拟时序和运行日志，再切换正式版本。这样才能区分“计算恢复”与“分析定义发生变化”。

变更参数时做成对照，而不是覆盖旧结果

当结果与预期不符时，先写下具体问题：是分支数不对、某个谱系拟时序缺失，还是曲线走向异常。一次只改变一个有依据的因素，例如候选起点或降维表示，并保持其余输入和参数一致。为每个版本保留参数文件、对象散列与输出，不覆盖基线目录。这样才能判断变化来自哪个因素，而不是在多个同时修改后把差异归因于服务器资源。

同一输入重复运行的结果若有差异，检查随机数种子、上游随机算法、软件版本、线程库和输入排序。种子固定可以帮助复现部分随机过程，但不保证不同R版本或并行环境逐位一致。对关键图形和数值结果，定义可以解释的比较指标，例如谱系数、共享细胞比例和拟时序排序相关性，并预先说明差异达到什么程度需要人工复核。

交付时附上方法与局限

给合作者的文件不应只有CSV。至少应附带拟合对象、细胞ID、谱系列定义、根簇依据、运行参数、版本信息和一张可解释的轨迹图。README中列出哪些细胞没有进入某谱系、是否存在被排除的簇，以及任务失败或未验收的项目。数据量较大时，可单独交付轻量结果和原始对象存放位置，但应保持结果与输入的唯一对应。

撰写结论时使用“在当前降维、分群和起点假设下，模型推断出……”这样的表达，避免把拟时序直接写成真实发育时间或因果机制。若轨迹分支与已知实验条件一致，也需要描述支持证据来自哪里，以及尚未验证的替代解释。方法边界说清楚，比单独报告一个漂亮的拟时序图更能让后续分析者判断是否可复用。

一个可直接复制的任务台账

可用CSV维护任务表，字段包括task_id、sample_id、input_sha256、embedding_key、cluster_key、start_cluster、seed、slingshot_version、status、output_dir、reviewer和review_note。每行对应一个预先定义的轨迹分析。运行脚本前核对task_id唯一、输入散列非空、输出目录不同；批次结束后统计validated、failed和pending各多少项。把台账与代码放在同一版本控制提交中，能让其他人知道这次轨迹究竟用了哪份输入。

一个示例状态变化是：planned表示输入尚未进入计算；running表示任务已启动且进程ID与日志已记录；failed表示命令失败或输出守卫未通过；validated表示代码退出、对象与细胞对应、图形检查及人工复核均完成。状态不能只按文件夹是否存在自动判断。共享目录中残留的旧文件尤其容易造成误判，因此建议新尝试使用不可复用的编号。

常见报错的第一步

如果提示起点簇不存在，先比较字符类型和分群因子水平：标签看起来相同但一个是数字、一个是字符时，显式统一再检查。若提示坐标非有限，定位具体细胞并回看上游归一化和降维，不要直接将其替换为零。若谱系曲线数与预期不同，应先检查分群拓扑、候选起点和数据覆盖，不能通过增加内存来解决拓扑定义问题。

如果拟时序文件有行但下游join后出现NA，通常要核对细胞ID是否在导出时被改写、是否加了样本前缀，以及CSV是否把ID列自动读成因子。保留稳定、唯一的原始cell ID，并在合并时显式匹配而非按行号对齐。每次整理ID规则后做一次输入和输出ID集合对比，避免静默丢掉细胞。

每次复核结束后由另一位组内分析者按输入清单重读方法选择、起点证据和输出细胞覆盖，记录发现的问题及处理人。若暂时没有第二位分析者，可将审核状态保留为待复核，不要把脚本运行成功等同于结论已经独立确认。

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

Slingshot官方vignette：输入、曲线与approx_points · 来源资料
Slingshot官方项目及源码 · 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。