定位SCTransform内存峰值并检查实现与参数

完整操作教程 · 参考固定 · 往返恢复验收

区分SCTransform拟合、残差和多层整理阶段，检查参数真实作用，用单层明确输入验证后再判断内存缺口。

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。

SCTransform拟合后会在SCT assay中保存校正counts、data和Pearson residuals。residuals通常放在非稀疏的scale.data里，特征数和细胞数都会影响内存。估算时先问：下游到底需要哪些residuals？不要只凭“用了SCT”就推断一定要全基因残差或必须换大内存。

先用矩阵维度估算载荷下限

一个双精度稠密矩阵的数值载荷约为“特征数 × 细胞数 × 8字节”。例如2万个基因、10万个细胞，单个全基因矩阵约14.9 GiB；这还不包含counts、SCT模型、中间副本、R和并行worker。此估算只用于识别数量级，不能当成完整峰值。

```
genes <- 20000
cells <- 100000
residual_gib <- genes * cells * 8 / 1024^3
print(residual_gib)
# ■■■■■■■■■■■■
```

选项/内容	作用	使用前要确认
return.only.var.genes	仅在scale.data保留可变基因residuals	下游是否需要全基因residuals
conserve.memory	避免构建完整residual矩阵，降低内存	会更慢，并强制仅返回可变基因
ncells	控制拟合阶段抽样细胞数	它不等于最终保留的细胞数
vst.flavor	选择sctransform模型版本	记录版本和方法，不能为省资源随意切换

在副本上记录对象和会话

先保存原对象，再检查counts维度、layer和当前assay。用接近正式数据的一份子集跑完SCTransform及计划内的PCA/邻接步骤；记录输出特征数、scale.data维度、对象大小和系统/作业峰值。object.size只能近似描述R对象，不代表函数执行时的临时峰值。

```
library(Seurat)
stopifnot(inherits(obj, "Seurat"), "RNA" %in% Assays(obj))
dim(obj[["RNA"]])
Layers(obj[["RNA"]])
format(object.size(obj), units = "GiB")
gc()
sessionInfo()
```

资源优化要服从分析目标

如果下游只需可变特征用于PCA，默认仅保留这些residuals可能已符合流程；若模型或下游检验明确需要全基因residuals，先查方法要求再比较资源。不要为降低峰值而改变基因筛选、回归变量或拟合版本，却仍把结果称作相同分析。每种参数配置要单独保存日志、对象和方法说明，并核对输出特征集合。

用成功标准决定是否调整容量

记录SCTransform开始和结束时间、退出码、失败步骤、worker数、作业额度及峰值来源。确认SCT assay存在、必要的counts/data/scale.data维度合理，PCA等后续步骤能使用预期数据。若失败发生在建立residuals阶段，比较全基因和仅可变基因的目标差别；若只在并行时失败，先做单worker对照。只有目标分析保持不变且代表性数据仍超出可用额度时，才据实测峰值规划更高内存。

把内存记录和对象结构对应起来

建议每个检查点同时记录细胞数、基因数、SCT scale.data的行列数、是否保留RNA assay、运行worker数和调度器峰值。若峰值接近估算的稠密矩阵载荷，可能还有模型及副本未计入；若高峰出现在保存或合并对象时，应单独测该阶段。清理对象前确认它是否是原始counts或分析复核所需检查点。rm和gc能帮助回收不再引用的对象，但不能把算法本身必须生成的矩阵变小，也不能保证操作系统立即回收所有常驻内存。

保留可重复的资源记录

每次试跑用唯一目录保存输入摘要、参数、版本、开始结束时间和日志；同时记录申请内存、实际峰值、峰值来源及任务退出码。若平台只提供主进程RSS，就在记录中注明这一点，不把它写成作业组峰值。同一输入重复测量时保留每次结果和节点负载，能帮助判断测量波动是否来自共享环境。

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

Seurat SCTransform reference · 来源资料

Seurat SCTransform vignette · 来源资料

R object.size documentation · 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。