分步骤记录整合耗时并核对算法与对象大小

完整操作教程 · 参考固定 · 往返恢复验收

把Seurat整合拆成归一化、缩放、PCA与整合阶段记录，区分方法条件、对象体积和实际内存峰值。

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。

Seurat v5把多个样本放在一个对象的不同layer中，再由IntegrateLayers调用CCA、RPCA、Harmony、FastMNN或scVI等方法。整合前的ScaleData和整合方法会产生中间矩阵；分块可影响计算速度与内存，但选择算法和特征不能只按省内存决定。

先确定整合前要准备什么

记录样本、批次和生物学条件，检查RNA counts/data层是否对应正确样本，并确定使用何种归一化、可变特征和整合方法。整合解决的是跨数据集的对齐问题，不代表所有批次差异都应删除。比较方法时先设定要保留的生物差异和评估标准，不能只以图上混合得更好作为正确性证据。

步骤	记录内容	验收点
输入	细胞数、样本、layer、版本	样本标识和细胞数正确
ScaleData	assay、features、block.size	输出特征符合下游计划
整合	method、reduction、dims及参数	产生预期降维且无警告
资源	峰值口径、worker、时限	代表性任务稳定完成

分块设置不是算法等价开关

ScaleData的features默认使用变量特征；block.size控制单次处理的特征数，调大可能更快，但会增加内存。先按既定研究流程选择features和回归变量，再固定版本试跑。下面只演示显式记录参数，具体assay和特征列表应从项目对象检查后填入。

```
features.use <- VariableFeatures(obj)
length(features.use)
obj <- ScaleData(
  obj, assay = "RNA", features = features.use,
  block.size = 500, verbose = TRUE
)
dim(obj[["RNA"]][["scale.data"]])
```

不同block.size通常是计算实现参数，但仍要按所用Seurat版本核对行为，并确认输出维度、缺失特征和下游PCA可用。若为了减内存改变features、回归变量或归一化方法，分析输入已变，应在方法中明确说明。

按阶段测资源

在接近正式规模的一份输入上，分开记录NormalizeData、FindVariableFeatures、ScaleData、RunPCA和IntegrateLayers的时间与峰值。GNU Time的主进程RSS未必覆盖全部worker；若并行运行，保存作业平台的聚合统计。每轮用独立副本和输出目录，记录Seurat版本、样本数、features数、worker和退出码。

验收整合结果和数据用途

检查降维结果名称、维度、细胞数与样本组成，按样本及生物学条件分别查看嵌入图和关键标记。若做差异表达，确认使用未整合的表达层或合适的样本级统计输入；整合降维不应自动替代原始表达矩阵用于所有检验。记录JoinLayers是否执行以及理由，保留未整合对象以便复核。只有结果满足预设生物学目标、资源峰值在可用额度内，才扩大到全部数据。

从小批次判断工作盘与重算成本

整合流程除内存外还要记录临时目录、对象保存路径和输入是否需要重复读取。试跑结束后检查目录剩余文件、是否能从保存点继续，以及重新运行一个阶段需要多久。不要在不同候选配置间复用被覆盖的输出目录；保留输入对象、参数文件和关键日志，后续才能区分数据结构改变与资源配置改变。若研究目标允许按样本分批预处理，先确认分批结果如何合并以及哪些步骤仍需全体细胞一起计算，不能把“分批跑完”当作整合等价。

运行前还要检查安装版本是否支持所选整合方法、其所需包是否在当前R库中，以及对象的输入reduction是否已生成。配置错误时先用小型官方或项目样例确认方法入口和参数名，不要同时升级多个包、改变输入并切换算法；否则报错和结果差异难以归因。将用于比较的RDS副本和原始对象分开保存。

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 v5 integration vignette · 来源资料

Seurat IntegrateLayers reference · 来源资料

Seurat ScaleData reference · 来源资料

GNU Time memory resources · 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。