用代表性样本比较整合方案和CPU内存组合

完整操作教程 · 参考固定 · 往返恢复验收

将方法筛选与服务器比较分开，在同一准备对象上验收CCA/RPCA结果，再固定方法测资源。

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。

比较Seurat整合方法或服务器配置时，先固定输入对象、样本分组、归一化、特征、随机设置和下游评估流程。CCA、RPCA、Harmony、FastMNN和scVI的假设、依赖与校正行为不同；不能因某一种运行更快或混合更明显，就认定它适合当前研究。

制定小而可复核的比较

选覆盖主要批次和生物学条件的代表性数据；尽量保留样本不平衡、稀有细胞等正式分析关注的结构。先跑未整合基线，再跑候选方法，记录每种方法的输入、版本、参数、运行时间、作业峰值、错误和降维输出。资源比较应使用同一硬件额度和存储位置；若某个方法依赖GPU或额外包，也要单独记录安装与设备条件。

比较方面	需要查看	防止的误判
运行条件	版本、输入、features、dims、依赖	参数变化被误当算法差异
资源	墙钟、峰值、worker、I/O	只拿主R进程RSS代表总量
整合质量	批次混合及已知生物标签	过度校正也会造成混合
下游结果	标记、稀有群、样本差异	只凭UMAP视觉判断

统一资源记录和错误处理

每次运行使用新目录，保存sessionInfo、控制台日志、输出校验和作业统计。失败后先定位步骤：ScaleData失败时检查features、block.size和内存；IntegrateLayers失败时检查方法依赖、输入layer、降维和worker日志。不要因为一个方法失败就只增加内存，先用固定子集确认接口和输入，再看是否触及资源额度。

```
sessionInfo()
Layers(obj[["RNA"]])
length(VariableFeatures(obj))
dim(obj[["RNA"]][["scale.data"]])
gc()
```

形成方法选择结论

汇总表应同时列出质量观察和资源证据：哪些已知细胞状态保留、批次差异是否降低、样本内结构是否稳定、峰值内存及运行时间是多少。不同方法结果不必完全相同；需要研究者依据实验设计判断取舍，并留存未整合对象。正式数据、样本结构或软件版本变化后，重新验证关键步骤。只有在结果可解释、运行可复现且资源需求清楚时，再据此规划服务器配置和批次。

同时检查过校正和未校正信号

比较候选方法时，除看批次标签是否混合，还要按细胞类型、条件和已知标记查看生物信号是否保留。对照未整合与整合后的降维，检查稀有群是否被合并、已知差异是否减弱、不同样本中共同类型是否能对应。若输入批次与实验条件完全重合，算法无法从数据本身可靠地区分技术差异和生物效应，结论要限定在设计能支持的范围内。

把评估图、定量指标、参数和人工判断依据存进同一结果目录。必要时请实验设计熟悉者复核，而不是根据一个整合分数选算法。方法选择涉及生物解释，资源较低但过度校正的方案不应仅因能在现有机器上运行就成为默认路线。

保存可复核的比较材料

为每种候选方法保存独立对象、运行脚本、sessionInfo、参数清单、降维结果和评估图。注明随机种子及是否使用并行，因为采样和worker顺序可能影响结果。对资源不足的方案，记录失败阶段、退出原因及是否可从中间结果恢复；不要只留一张最终UMAP图或一组耗时数字。这样合作者能复核方法取舍，也能在软件更新后重做关键比较。

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。