检查检验方法、细胞分组与重复计算

完整操作教程 · 参考固定 · 往返恢复验收

Seurat差异表达运行很慢，先确认要比较什么，再检查哪一步耗时。细胞类型标记寻找与处理组差异推断有不同的统计目标，不能为了速度把每个细胞都当作独立生物学重复。本文聚焦已经具备合适对象的标记比较，给出从分组检查、方法核对到单比较计时的步骤；涉及处理条件效应时，应先确定样本层面的分析设计。 先写出比较

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差异表达运行很慢，先确认要比较什么，再检查哪一步耗时。细胞类型标记寻找与处理组差异推断有不同的统计目标，不能为了速度把每个细胞都当作独立生物学重复。本文聚焦已经具备合适对象的标记比较，给出从分组检查、方法核对到单比较计时的步骤；涉及处理条件效应时，应先确定样本层面的分析设计。

先写出比较单位与目标

Seurat当前官方参考提醒，处理组之间的差异应考虑跨样本生物学变异，例如使用适用的pseudobulk路径。若只是查找某类细胞的标记，可以在明确适用范围内进行细胞级比较，但不能直接把这一输出解释为受试者层面的处理效应。若设计尚不清楚，先解决分组与重复问题，再讨论加CPU。

问题	操作前应核对	对耗时判断的影响
标记比较	目标细胞身份和对照身份	比较范围决定参与的细胞与特征
处理组效应	独立样本、供者和批次设计	可能需要不同统计路径
表达层	RNA、SCT或其他assay来源	不能用不适配层替代
检验方法	版本、输入及支持参数	不能只按速度选择
重复计算	同一比较是否已有相同参数结果	避免无意义重跑

核对对象和身份，避免算错范围

下面假设object.rds是已经正确预处理的Seurat对象，RNA assay有适用的单一标准化表达层，且metadata里存在celltype列。A、B只是待替换的身份标签。多layer对象、SCT对象和其他版本应先按对应文档完成准备；不要因为报错就盲目把所有层合并，或改用另一个assay绕过问题。

```
library(Seurat)
obj <- readRDS("object.rds")
stopifnot("celltype" %in% colnames(obj[[]]))
Idents(obj) <- "celltype"
print(table(Idents(obj)))
stopifnot(all(c("A", "B") %in% levels(Idents(obj))))
print(SeuratObject::Layers(obj[["RNA"]]))
sessionInfo()
```

表格里的细胞数要与预期一致。检查是否混入不该比较的样本、是否有空身份，尤其不要把“所有其余细胞”误认为一个特定对照。身份与样本设计固定后，记录参与的细胞范围和特征范围，后续试跑才能比较相同任务。

单独计时一个已确认的比较

先运行一个必要比较，不立即启动所有细胞群两两组合。下例明确使用Wilcoxon路径和稀疏输入，不代表它适用于所有研究问题，也没有把阈值作为单纯的提速旋钮。具体阈值和特征过滤应在分析计划中确认，再在版本记录中保存。

```
elapsed <- system.time({
  markers <- FindMarkers(
    obj, ident.1 = "A", ident.2 = "B",
    assay = "RNA", slot = "data",
    test.use = "wilcox", densify = FALSE
```

```
  )
})
print(elapsed)
print(dim(markers))
write.csv(markers, "markers_A_vs_B.csv")
saveRDS(list(timing = elapsed, version = sessionInfo()),
        "markers_A_vs_B_run.rds")
```

若程序只使用少量CPU，不意味着它异常；并行支持取决于检验实现、包版本和运行设置。先看时间花在对象加载、准备、检验还是写文件，不能用全机CPU占用低直接推出需要更多核心。对大对象，额外worker还可能复制数据并提高总内存。

优化前后保持同一分析问题

可以检查是否重复读取同一大对象、是否重复做不变的预处理，或者是否无必要地把所有基因转成稠密矩阵。若调整max.cells.per.ident、过滤阈值或特征集合，参与检验的数据就可能变化，应作为另一分析方案记录，而不宣称只是硬件加速且结果必然等价。

验收包括比较身份与输入正确、输出列和结果数量合理、日志无未解释错误，并保留与原方案的差异说明。只看脚本结束或CSV存在不够；异常空结果也应核对过滤、表达和分组，而不是先认为CPU不足。保存任务的版本记录，后续软件更新也不能直接把新结果混入旧比较。

什么时候评估服务器资源

在方法、数据层和比较范围确认后，若必要比较数量很大，现有机器无法在项目时间内完成，或并行时内存成为限制，再评估CPU与内存组合。可以向云生信提供检验方法、样本和细胞规模、比较清单、单任务耗时及峰值，沟通自行运行分析的租赁环境。方法选择、统计解释和科研结论不默认包含在租金内。

参考资料

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 FindMarkers官方参考 · 来源资料

GNU time手册（Debian发布） · 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。