按细胞数、基因数、对象结构和步骤估算内存

完整操作教程 · 参考固定 · 往返恢复验收

“128G能分析多少个细胞”没有统一答案。同样十万个细胞，只有稀疏计数矩阵的任务，与同时保留多个assay、全基因缩放结果和整合中间对象的任务，内存峰值可以很不一样。先把对象结构和要运行的步骤列出来，才能判断当前机器能否完成；直接按细胞数购买规格，容易漏掉真正占内存的环节。

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。

“128G能分析多少个细胞”没有统一答案。同样十万个细胞，只有稀疏计数矩阵的任务，与同时保留多个assay、全基因缩放结果和整合中间对象的任务，内存峰值可以很不一样。先把对象结构和要运行的步骤列出来，才能判断当前机器能否完成；直接按细胞数购买规格，容易漏掉真正占内存的环节。

先区分四种大小

FASTQ文件大小描述原始读段，不等于进入R之后的表达矩阵大小；RDS是序列化文件，也不能当作进程运行内存。对象在读入、转换和计算时还会产生临时数据。对于Seurat等复合对象，应分别检查计数、标准化、缩放和降维结果，不要只看文件管理器里的一个数字。

需要记录的量	它能回答什么	它不能代替什么
细胞数与基因数	输入维度与可能的计算规模	不能单独推出峰值
非零元素数	稀疏计数的实际密度	不能说明后续是否转成稠密矩阵
对象大小	当前R对象的近似占用	不能代表函数执行时的峰值
进程峰值与可用额度	某次运行真正消耗多少	不能脱离版本和步骤套给其他任务

用一个矩阵演算理解稠密化风险

以下只是容量算术示例，不是Seurat的配置承诺。假设20,000个基因、100,000个细胞，若把一个矩阵存成每元素8字节的双精度稠密矩阵，仅数值区就需要约14.9 GiB。若已有计数、又生成标准化或缩放矩阵、再产生临时副本，总需求会继续增加。实际对象还有索引、名称和其他内容。

```
genes <- 20000
cells <- 100000
dense_gib <- genes * cells * 8 / 1024^3
print(dense_gib)
# ■■■■■■■■■■■■■■■■■■■
# ■■■■■■■■■■ as.matrix(■■■■■■■)
```

稀疏表示的好处取决于非零比例和具体操作是否保留稀疏结构。不能把“输入是稀疏矩阵”理解成所有后续步骤都会低内存。Seurat官方大规模教程展示了磁盘后端和sketch路径，但采用这些路径前，需要确认它们覆盖你的分析目标及所用版本。

在读入和每个大步骤后记录对象

下面假设当前已经有一个名为obj的Seurat对象。先运行诊断而不是删除数据。object.size给的是近似对象大小；外部指针、共享内容和磁盘后端的文件数据未必完整反映在这个值中。测量应和进程监控一起使用。

```
stopifnot(exists("obj"))
print(dim(obj))
print(SeuratObject::Assays(obj))
print(format(object.size(obj), units = "GiB"))
print(gc())
sessionInfo()
# ■■■■■■■■■■■■■■■■■
```

■■■■■■■■■■■■■■■■■■■

确认哪些结果仍然需要：旧的整合中间对象是否只是临时变量，缩放是否确实要覆盖所有基因，多个版本的对象是否在同一个R会话里同时保留。处理前先保存可恢复的检查点，避免为了省内存丢掉后续复查需要的原始计数。不要在不清楚对象结构时批量清空assay或layer。

把“128G够不够”改成可验收的问题

选择接近正式输入的数据，记录最吃内存的步骤、是否有并行worker、同机是否还运行其他任务，以及容器或账号的可用额度。机器标称128G，不代表某个会话可以独占全部内存。若已经频繁换页或被限额终止，先查进程、调度和容器限制，不能用提高R里的某个导出上限来替代物理容量。

验收至少要求目标步骤成功结束、输出维度与必要内容正确、日志无内存终止，并且能在约定并发下复现。合理的数据表示和计算范围仍超过可用内存时，再评估容量。保留细胞量、对象结构、卡住步骤和实测峰值，作为后续资源比较的依据。

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

R对象大小估计 · 来源资料

Seurat大规模数据与sketch教程 · 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。