用磁盘后端与sketch策略搭建大规模分析流程

完整操作教程 · 参考固定 · 往返恢复验收

使用BPCells保存完整counts，在明确的sketch assay上探索并投影到全量，同时验证后端目录与资源边界。

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。

面对几十万细胞，首先要区分“保留完整表达数据”与“每一步都在完整数据上计算”。Seurat的磁盘矩阵与sketch流程提供一种可选路径：完整counts留在磁盘，选定细胞用于部分探索，再把学到的表示或标签投影到全量数据。它不是把大任务自动变成完全无内存需求的操作。

下面演示当前公开工作流，要求Seurat、SeuratObject和BPCells兼容，保存辅助函数还需要fs等依赖。例子以已经确认是RNA基因表达计数的10x HDF5为输入；含ADT、其他模态或不同特征命名的文件应先按模态核对，不能直接把所有行当作RNA基因。

先建立磁盘counts和Assay5对象

使用新的工作目录，保留原HDF5不动。BPCells负责磁盘矩阵的读写，Seurat保存对象结构和路径。创建之后核对assay类、特征和细胞名，不调用as.matrix，也不根据一个很小的RDS大小判断完整数据只占很少空间。

```
library(Seurat)
library(SeuratObject)
library(BPCells)
dir.create("analysis_bundle", showWarnings=FALSE)
bp_dir <- "analysis_bundle/counts_bp"
if (dir.exists(bp_dir)) stop("[redacted]")
mat <- open_matrix_10x_hdf5("filtered_feature_bc_matrix.h5")
write_matrix_dir(mat=mat, dir=bp_dir)
counts <- open_matrix_dir(bp_dir)
obj <- CreateSeuratObject(counts=counts)
stopifnot(inherits(obj[["RNA"]], "Assay5"),
          identical(Layers(obj[["RNA"]]), "counts"))
print(list(dim=dim(obj[["RNA"]]), class=class(obj[["RNA"]]$counts)))
writeLines(capture.output(sessionInfo()), "analysis_bundle/session.txt")
```

建立sketch，并只在明确的assay上计算

下面的细胞数、维度和分辨率仅为演示参数，要用目标数据检验。示例单counts层意味着ncells对应本层抽取量；多层时当前接口允许按层指定数量，单个整数也可能按每层解释，不能误以为总数永远等于该整数。

```
DefaultAssay(obj) <- "RNA"
obj <- NormalizeData(obj)
obj <- FindVariableFeatures(obj)
obj <- SketchData(obj, assay="RNA", ncells=min(10000L, ncol(obj)),
                  method="LeverageScore", sketched.assay="sketch", seed=2026L)
DefaultAssay(obj) <- "sketch"
obj <- FindVariableFeatures(obj)
obj <- ScaleData(obj, features=VariableFeatures(obj))
npcs <- min(30L, length(VariableFeatures(obj))-1L, ncol(obj[["sketch"]])-1L)
stopifnot(npcs>=2L, ncol(obj[["sketch"]])>50L)
obj <- RunPCA(obj, npcs=npcs)
obj <- FindNeighbors(obj, dims=seq_len(npcs))
obj <- FindClusters(obj, resolution=0.5)
obj <- RunUMAP(obj, dims=seq_len(npcs), return.model=TRUE)
```

对象或步骤	主要作用	必须核对的边界
RNA磁盘counts	保留全部原始计数访问	后端目录是数据的一部分
sketch assay	支持一部分探索和建模	不是完整数据的同义替身
PCA与UMAP模型	在抽取细胞上学习表示	参数与稀有群体表现需检查
全量投影	扩展所选表示或标签	不等于所有步骤全量重算

对象或步骤	主要作用	必须核对的边界
保存与迁移	保留对象与后端依赖	不能只复制一个RDS文件

需要全量标签时，再明确投影步骤

用与sketch一致的维度和UMAP模型投影，并分别命名全量结果。得到全量标签后，检查细胞覆盖和已知标记，特别关注稀有群体是否被过度合并。sketch有自己的抽样目标，标签投影也有近似性，不能凭图形好看保证每个稀有群体都被正确发现。

```
obj <- ProjectData(obj, assay="RNA", sketched.assay="sketch",
                   sketched.reduction="pca", full.reduction="pca.full",
                   umap.model="umap", dims=seq_len(npcs),
                   normalization.method="LogNormalize",
                   refdata=list(cluster_full="seurat_clusters"))
DefaultAssay(obj) <- "RNA"
stopifnot(nrow(Embeddings(obj, "pca.full"))==ncol(obj[["RNA"]]))
print(table(is.na(obj$cluster_full)))
SaveSeuratRds(obj, file="analysis_bundle/analysis.rds",
              move=FALSE, relative=TRUE)
check <- LoadSeuratRds("analysis_bundle/analysis.rds")
print(dim(check[["RNA"]]))
```

保存、资源和下一步验收

此处后端原本建在bundle内，保存时显式不移动已有数据，并使用相对路径。将完整目录复制到新位置后，还应在新R会话重新载入、抽查少量值和必要降维。不要在验证前清理原始输入或后端目录；对象能被读取但外部路径失效，后续访问仍会失败。

将技术验证与细胞质量判断分开

HDF5文件能够被打开，只说明读取路径可用。仍需按实验情况确认低质量细胞、空液滴及特征命名等问题，并保存质量筛选依据。不能为了让对象更小就随意删掉难处理的样本，也不能把没有做质量检查的示例流程直接当作最终科学结果。

在全量投影前，先在sketch中查看与研究问题相关的标记和样本覆盖。如果某个关键群体根本没有合适的参考细胞，投影不会自动提供新的证据。对需要深入分析的群体，应返回原始细胞集合建立适用的后续方案，并说明该方案的计算范围。

大规模绘图也可能产生额外时间与内存需求。先用适当范围核查表达和标签，再决定是否输出全量图；绘图慢、后端路径失效和模型计算不足，应分别排查。最终交接时，保留目录清单、版本与关键检查结果，才能让新的服务器验证同一套对象，而不是只收到一个可以打开外壳的RDS。目标账号还应能读取全部后端目录，权限验收也属于交接检查的一部分。

分别测量写盘、归一化、sketch、投影和保存的内存与I/O峰值。后续算法如果不支持该矩阵类或会物化数据，仍可能需要较大内存。确认必要步骤和资源缺口后，再评估适合的大内存CPU、工作盘与租期，服务器租赁的环境部署和数据保留范围需逐项确认。

参考与版本核对

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：BPCells磁盘矩阵与保存 · 来源资料

Seurat：sketch与全量投影工作流 · 来源资料

Seurat：SketchData参数 · 来源资料

Seurat：ProjectData投影参考 · 来源资料
SeuratObject：保存与载入磁盘后端对象 · 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。