BANKSY大型数据实操：比较lazy与物化PCA并记录资源

用同一份SingleCellExperiment分别测物化与lazy PCA，先核验归一化矩阵、坐标、细胞顺序和版本；记录RSS与输出差异，不把省内存误解为免资源。

BANKSY将表达特征与空间邻域信息结合后做降维。大型数据评估要分清两种路径：物化路径会先构造邻域特征矩阵，再做PCA；lazy路径在PCA迭代中按需计算，不物化完整BANKSY矩阵。当前文档的lazy PCA只支持M=0（不使用AGF），并仍需读取表达矩阵、构图和保存PC；它不是“零内存”，也不代表聚类或后续整合同样省资源。

先固定可比输入

项目	检查内容	影响
表达	同一normcounts、特征顺序和稀疏格式	避免尺度差异污染比较
位置	x/y有限值，样本字段准确	切片不可串图
参数	同k_geom、lambda、npcs、seed	分离实现差别与设计差别
路径	物化use_agf=FALSE；lazy限M=0	仅比较共同支持的设置

示例假设normcounts是dgCMatrix；若数据由BPCells或DelayedArray后端管理，不要为通过类型检查而强制转成内存稀疏矩阵，应按对应后端核验lazy支持。输入用归一化表达但不要把两套对象分别归一化后直接比较；确认列名唯一、坐标和细胞列顺序一致，并记录特征数、细胞数、稀疏度和样本数。lambda控制空间信息权重，k_geom定义邻域尺度，应按目标是细胞分型还是空间域、以及平台几何结构选取。示例值只作资源探测，不是推荐参数。正式分析应以参数敏感性或组织特定方案为依据，并记录选择理由。Visium网格的邻居数有明确环层解释，其他技术不能照搬。

分进程做资源探测

下面假设banksy-input.rds为SingleCellExperiment，含normcounts和colData的x/y。物化路径先运行computeBanksy再PCA；lazy路径不必先运行computeBanksy。每次在新R进程和独立输出目录运行，避免前一轮对象留在内存中。近期版本才有lazy和pca_backend接口，先锁定Banksy版本并检查函数参数。安装环境应保存锁文件和版本。

```
library(Banksy)
library(SingleCellExperiment)
args <- commandArgs(TRUE)
stopifnot(length(args)==2, args[1] %in% c("materialized","lazy"))
mode <- args[1]; out <- args[2]
se <- readRDS("banksy-input.rds")
stopifnot(inherits(se,"SingleCellExperiment"),
          "normcounts" %in% assayNames(se),
          all(c("x","y") %in% names(colData(se))),
          !anyDuplicated(colnames(se)), length(reducedDimNames(se))==0)
x <- assay(se,"normcounts")
stopifnot(inherits(x,"dgCMatrix"), all(is.finite(x@x)), all(x@x>=0),
          all(is.finite(as.matrix(colData(se)[,c("x","y")]))),
          all(c("lazy","pca_backend") %in% names(formals(runBanksyPCA))))
dir.create(out, showWarnings=FALSE)
stopifnot(length(list.files(out))==0)
if(mode=="materialized") {
  se <- computeBanksy(se, assay_name="normcounts",
    coord_names=c("x","y"), compute_agf=FALSE, k_geom=15,
    chunk_size=100, parallel=FALSE)
  saveRDS(se,file.path(out,"neighborhood.rds"))
  se <- runBanksyPCA(se, assay_name="normcounts", use_agf=FALSE,
    lambda=0.2, npcs=20, seed=101, lazy=FALSE)
} else {
  se <- runBanksyPCA(se, assay_name="normcounts",
    coord_names=c("x","y"), k_geom=15, spatial_mode="kNN_median",
    lambda=0.2, npcs=20, seed=101, lazy=TRUE, pca_backend="cpp")
}
pc <- reducedDimNames(se)
stopifnot(length(pc)>0, all(is.finite(reducedDim(se,pc[1]))))
saveRDS(se,file.path(out,"banksy-pca.rds"))
writeLines(capture.output(sessionInfo()),file.path(out,"session.txt"))
write.csv(data.frame(cell=colnames(se)),file.path(out,"cells.csv"),row.names=FALSE)
```

在Linux服务器上可分别用/usr/bin/time -v Rscript probe.R materialized materialized-run和lazy lazy-run包住命令，记录Maximum resident set size及耗时；两次应在同一节点顺序运行，避免其他任务并发干扰。同时保存日志、sessionInfo和输出维度；RSS包含读入、建图、PCA和写盘阶段，定位峰值时要结合日志时间点。物化阶段单独保存邻域矩阵，观察它的文件大小与PCA阶段峰值；lazy路径没有该中间矩阵，不应把两者峰值拆开后误算总任务资源。

验收结果并解释差异

比较两次的细胞条码顺序、PC维度、有限值、耗时和RSS。PCA方向可因符号翻转而不同，不能逐元素要求相等；可比较PC子空间、邻近结构或固定下游流程的聚类稳定性。若目标分析需要AGF，当前lazy路径不适用，不能为了省内存直接删去该生物学特征。若物化路径在chunk计算阶段报向量长度错误，核对细胞数、特征数和chunk_size；computeBanksy的parallel在Windows不支持。

lazy降低了完整特征矩阵的物化开销，但输入矩阵、邻居索引、PCA工作区、输出PC和后续图聚类仍占内存与CPU。记录软件版本、n_cells、n_features、稀疏度、样本数、k_geom、lambda、npcs、AGF设置、backend、峰值RSS和时间。用代表样本做试跑后，再测完整数据；只有实测显示设备达不到目标周期时，才按这些记录评估计算资源。

参考资料

BANKSY lazy PCA官方接口和限制 官方文档
BANKSY物化邻域矩阵接口 官方文档
BANKSY参数选择与lambda、k_geom说明 官方文档