Cicero共可及性实操：输入检查、区域任务与资源试跑

完整操作教程 · 参考固定 · 往返恢复验收

核对Monocle3对象、峰坐标与聚合入口，明确Cicero区域拆分和模型验收的边界。

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。

Cicero用相似细胞聚合后的可及性数据估计峰间共可及性。大量细胞的任务既包含聚合，也包含局部基因组窗口模型；两者的内存压力不同。下面固定Monocle3兼容分支，先核输入，再跑一份有完整日志的任务，不把共可及性直接解释成已经证实的增强子—靶基因作用。

先确认对象类型和参考坐标

准备完成质控和合适LSI/UMAP的cell_data_set，counts为峰×细胞非负稀疏计数。Monocle2的CellDataSet与Monocle3对象不能混用。记录cicero、monocle3、R版本及源码修订；安装能导入只证明环境入口，输入峰集和降维仍需验收。

阶段	检查内容	资源记录
输入	峰ID、细胞ID、UMAP与参考长度表	峰数、细胞数、坐标匹配
聚合	邻近细胞聚合k与随机种子	聚合后列数、时间与RAM
距离估计	窗口、抽样数量、物种假设	有效抽样窗口与估计警告
局部模型	峰密度、窗口重叠与组装	失败窗口、连接数及磁盘

从明确的已有对象跑通

把以下保存为cicero_probe.R。示例限峰名为chr名称_起点_终点且染色体名不含下划线，chrom.sizes为同一组装的无表头两列表。若项目含其他命名，应按官方格式准备映射，不能用简单字符串截断碰巧通过。示例采用哺乳动物默认窗口500kb与100个抽样窗口，其他物种先重新论证参数。

```
library(cicero)
library(monocle3)
library(SingleCellExperiment)
cds <- readRDS("atac-monocle3-ready.rds")
stopifnot(is(cds,"cell_data_set"),ncol(cds)>50,
          !anyDuplicated(rownames(cds)),!anyDuplicated(colnames(cds)))
x <- counts(cds)
stopifnot(inherits(x,"dgCMatrix"),all(is.finite(x@x)),
          all(x@x>=0),all(x@x==floor(x@x)))
coords <- reducedDims(cds)$UMAP
stopifnot(is.matrix(coords),ncol(coords)>=2,
          identical(rownames(coords),colnames(cds)),all(is.finite(coords)))
ids <- rownames(cds)
stopifnot(length(ids)>0,all(grepl("^[^_]+_[0-9]+_[0-9]+$",ids)))
parts <- do.call(rbind,strsplit(ids,"_",fixed=TRUE))
genome <- read.delim("chrom.sizes",header=FALSE,colClasses=c("character","numeric"))
stopifnot(ncol(genome)==2,!anyDuplicated(genome[[1]]),
          all(is.finite(genome[[2]])),all(genome[[2]]>0))
idx <- match(parts[,1],genome[[1]])
start <- as.numeric(parts[,2]); end <- as.numeric(parts[,3])
stopifnot(!anyNA(idx),all(start>=0),all(end>start),
          all(end<=genome[[2]][idx]))
rd <- as.data.frame(rowData(cds))
if (all(c("chr","bp1","bp2") %in% names(rd))) {
  old_start <- suppressWarnings(as.numeric(as.character(rd$bp1)))
  old_end <- suppressWarnings(as.numeric(as.character(rd$bp2)))
  stopifnot(identical(as.character(rd$chr),parts[,1]),
            all(is.finite(old_start)),all(is.finite(old_end)),
            all(old_start==start),all(old_end==end))
}
```

```
stopifnot(dir.create("cicero-probe-001"))
writeLines(capture.output(sessionInfo()),"cicero-probe-001/session.txt")
set.seed(2026)
agg <- make_cicero_cds(cds,reduced_coordinates=coords,k=50)
saveRDS(agg,"cicero-probe-001/aggregated.rds")
con <- run_cicero(agg,genome,window=500000,sample_num=100)
stopifnot(all(c("Peak1","Peak2","coaccess") %in% names(con)))
write.csv(con,"cicero-probe-001/connections.csv",row.names=FALSE)
```

在独立进程用GNU Time记录完整墙钟与主机RAM，日志中保留距离参数估计、模型失败及警告。峰名的数值边界检查不等于已证明坐标体系正确，仍要抽查原始峰文件与参考；不要为消除报错无依据地给所有起点加一或删掉chr前缀。

```
set -euo pipefail
/usr/bin/time -v -o cicero-resources.txt \
  Rscript cicero_probe.R > cicero-probe.log 2>&1
```

区域拆分先处理统计边界

若rowData已同时含chr、bp1、bp2，Cicero会保留这些坐标而非总从峰名重建，因此代码也检查已有元数据与峰ID是否一致。发现冲突先回到原峰文件与参考确认来源，另存经核对的输入，不能只让峰名通过校验后继续使用旧元数据。

run_cicero先估计距离惩罚，再生成重叠窗口模型并组装连接。随意对每一小段独立重估后拼表，会同时改变惩罚和边界，不能称为与整条染色体等价。需要分区域时，按官方分步接口保存既定距离参数、窗口及组装规则，并记录跨边界连接如何处理；不把重复峰对简单相加。

试跑应覆盖不同峰密度，不能只取很稀疏区域估算整套任务。减小窗口、减少估计抽样或删除密集峰都会改变推断条件，应单列方法变化。局部模型峰数过多时，先检查异常峰集及模型日志，再判断容量；降低并发可以减少多任务叠加，却不减少单个必要模型的工作集。

任务表至少逐项写出输入RDS校验值、参考长度表、区域范围、聚合k、窗口、距离参数来源、随机种子与输出目录。失败重跑也创建新目录，先定位报错阶段，避免把旧聚合对象和另一轮降维混用。大文件搬迁前后核对校验值，不能仅凭文件名相同复用。

检查结果后再扩大任务

核对Peak1/Peak2都属于输入峰集、分数非有限值和模型失败比例，保留原始结果，不把NA自动改成零。无连接可能来自数据覆盖、窗口或模型失败，不能只看CSV存在。最终连接需结合距离、基因注释、重复与独立证据解释，不能当作物理接触或因果调控的直接证明。

峰坐标、基因组和降维必须一致

Cicero的峰区间、chrom.sizes、UMAP或其他低维坐标必须来自同一批细胞和同一基因组版本。BED坐标常采用0-based half-open，而峰名常写成1-based闭区间或不同约定；在生成峰名和chrom.sizes时记录具体转换规则。若染色体使用chr1与1两种命名，应通过明确映射统一，不要仅靠删掉前缀猜测。

逐行验证峰起点小于终点、终点不超过染色体长度，峰ID唯一且无格式歧义。UMAP行名必须严格匹配细胞列名，样本及过滤后细胞要与ATAC矩阵一致。更换参考组装或重做peak calling后，旧chrom.sizes、降维和Cicero连接不能混用。

细胞聚合与局部窗口模型

Cicero会依据低维空间相近细胞形成聚合，再在局部基因组窗口中估计共可及性。聚合数量、窗口长度和抽样窗口数都会影响运行时间与结果。抽样窗口少适合验证入口与资源，不足以支撑全基因组稳定结论。哺乳动物教程默认参数不适用于所有物种或基因组结构，需根据组装和数据密度确认。

聚合后检查每个伪细胞包含的原始细胞数、不同样本的构成，以及是否有某些样本完全缺失。若聚合空间主要由处理组或批次驱动，估计出的可及性关系可能反映组成差异。必要时按有科学依据的组别分别建立输入并进行比较，但不能在看过目标连接后选择性拆分。

共可及性不等于调控验证

Cicero连接表示在所用细胞集合和算法设置下，两个区域的可及性存在统计关联。它可帮助提出候选调控关系，但不能单独证明峰是某基因增强子、连接方向或真实因果作用。将连接与距离、基因表达、motif、染色质互作及独立实验信息结合，报告多条证据的来源。

筛选高分连接时说明阈值和多重比较处理，保留未筛选原始结果和所有参数。若研究比较不同条件，确保每组使用可比的峰集、细胞筛选和窗口设置，并检查连接稳定性。不能把某个阈值以上的边直接写成“已验证靶基因”。

资源排查及窗口失败

聚合阶段耗时或内存升高时查看细胞数、邻居图和低维矩阵规模；局部模型阶段则与峰密度、窗口重叠及抽样量有关。将日志按阶段保存，统计失败窗口及其坐标，避免只看到脚本结束码。大多数窗口成功而少数失败时，检查对应区域的峰坐标、稀疏度和组装边界。

更改window、sample_num或抽样范围都会改变建模范围或稳定性，必须作为分析参数记录。若系统OOM，检查R对象副本、内存限额和并发数；若是某些窗口拟合警告，增加RAM可能无效。先在固定小样本上定位，再用代表规模数据评估作业额度与临时盘。

交付文件与审计

保存Monocle3/Cicero版本、输入对象散列、chrom.sizes来源、峰集和坐标转换规则、降维方法、窗口设置、随机种子、日志和连接表。将基因组组装名称写进文件名或manifest，不要仅依赖“hg38”这种未说明来源的简称。输出big对象、连接表及筛选后表之间应有清晰关系。

交付时附至少一个连接示例，展示两个峰的坐标、Cicero得分、关联基因假设和外部支持证据。高分连接没有外部证据时仍属候选关系。受限项目的峰或细胞元数据也可能敏感，共享前检查样本ID和访问权限。

物种和参考长度的专项核对

小鼠、斑马鱼或非模式物种需要使用与研究一致的参考组装和染色体长度，不能直接使用哺乳动物默认参数。线粒体、未定位contig和性染色体的命名也要与peak和chrom.sizes一致。对照组装官网或项目构建记录核验总长度与序列名称，保留版本来源和散列。

若峰名包含下划线、特殊字符或未定位contig，不能用简单split规则解析。应从原始区间表提供独立的chr/start/end列，再生成Cicero支持的输入对象。坐标转换后抽取几个峰与原BED对照，确保没有起点偏移或start/end颠倒。

小样本与全量任务的差别

局部试跑用于确认接口和错误处理，不代表所有基因组窗口均可完成。按固定规则抽取窗口时，记录随机种子和窗口坐标；全量运行需另存完整窗口清单及失败重试状态。扩展前测一个高峰密度区域和一个低密度区域，了解资源上界及是否存在特定窗口拟合失败。

如果局部跑通但全量失败，优先比对失败窗口密度、坐标范围、聚合后细胞构成和临时磁盘，而非简单重复整批。处理有科学依据的异常窗口时，将筛除清单和原因保留，并报告覆盖范围改变。不能只挑成功窗口绘图后隐去未完成区域。

窗口尺度和稳定性

窗口长度会改变进入同一局部模型的基因组范围，抽样窗口数影响估计覆盖。固定参数后可评估合理设置下主要连接是否稳定，预先定义共同峰对比例或分数排序相关性。不要从大量参数组合中只挑支持假设的设置。

跨样本连接比较需要共同峰集或明确的峰ID映射。不同组装的峰坐标不能只按字符串匹配。保存每次连接的窗口参数、抽样种子和过滤步骤，区分生物差异与峰集差异。

交付证据链

交付表列出候选峰对坐标、基因组组装、共可及性分数、是否跨样本稳定及外部证据。没有配套表达或实验验证时写作“候选连接”。附输入散列和软件版本，确保其他分析者可以从结果回查原始数据。

Cicero输出可以与基因组浏览器轨迹和基因注释叠加，抽查连接是否落在正确染色体区段。检查注释版本、TSS定义和链方向后再讨论近端/远端关系。近距离峰对也不一定调控同一基因，远距离候选更需要独立证据支持。

资源记录应区分聚合、距离估计、局部窗口模型和结果合并。申请资源时列出细胞数、峰数、抽样窗口数和最慢窗口，而不只给文件大小。服务器能够运行流程不代表参数适合该物种或研究问题。

对于距离较远的峰对，建议同时核对基因表达、启动子邻近信息或独立染色质证据，并标明证据来自同一队列还是外部数据。若所有支持都来自当前共可及性计算，就仍是待验证假设。

把连接表与chrom.sizes和基因注释版本一并归档，区域名称应能复原到染色体起止坐标。离开原始组装和坐标转换规则，峰对得分无法被其他人正确定位或复核。

报告窗口估计时记录抽样范围、有效窗口数和警告数量。某些区域模型未能收敛时，保留坐标及日志并判断其是否与峰密度或输入坐标有关；不要静默删除失败窗口后将结果称为全基因组完整分析。

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

Cicero Monocle3官方教程 · 来源资料

Cicero聚合与窗口模型源码 · 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。