chromVAR大型任务排查：峰矩阵、背景集合与资源测量

完整操作教程 · 参考固定 · 往返恢复验收

保留完整细胞与峰集及固定背景，按motif块计算chromVAR，分别定位背景与打分内存。

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。

chromVAR输入稀疏，不代表输出和背景抽样都很小。大型任务应分别检查峰×细胞计数、峰×motif匹配、背景索引和motif×细胞得分。下面保留全部既定细胞与峰集、背景和期望，按motif块写出结果，减少一次保留全部motif得分的压力。

先把三种行顺序对齐

准备已完成质量过滤的SummarizedExperiment，counts为非负原始片段稀疏矩阵，rowRanges对应峰坐标。按正确组装运行addGCBias后，rowData的bias保存GC偏好；motif匹配矩阵必须按同一峰顺序。不得用归一化表达、TF-IDF或另一峰集的motif结果替代原始计数。

阶段	固定材料	验收内容
计数	峰×细胞、唯一ID和非空列	稀疏类型、深度与非零项
motif	峰×motif的逻辑匹配	峰ID顺序与motif标签
背景	同峰集的GC和平均可及性	背景索引、抽样数与种子
打分	固定全体期望与背景	输出列顺序、异常值及内存

固定背景后按motif块写出

将下列保存为chromvar_chunks.R，在已验证的chromVAR环境执行。示例要求obj-gc.rds已正确添加GC偏好，motif-matches.rds为逻辑稀疏矩阵。每块100个motif只是演示，先按细胞量与单块峰值调整。无匹配峰的motif需记录并解释；本例要求每个保留motif至少匹配一个峰。

```
library(chromVAR)
library(SummarizedExperiment)
library(BiocParallel)
register(SerialParam())
obj <- readRDS("obj-gc.rds")
motif <- readRDS("motif-matches.rds")
x <- assays(obj)$counts
stopifnot(inherits(x,"dgCMatrix"),nrow(x)>0,ncol(x)>0,
          !is.null(rownames(obj)),!is.null(colnames(obj)),
          !anyDuplicated(rownames(obj)),!anyDuplicated(colnames(obj)),
          all(is.finite(x@x)),all(x@x>=0),all(x@x==floor(x@x)),
          all(Matrix::colSums(x)>0),all(Matrix::rowSums(x)>0))
stopifnot(inherits(motif,"lgCMatrix"),ncol(motif)>0,!anyNA(motif@x),
          identical(rownames(motif),rownames(obj)),
          !is.null(colnames(motif)),!anyDuplicated(colnames(motif)),
          all(Matrix::colSums(motif)>0))
stopifnot("bias" %in% names(rowData(obj)),
          all(is.finite(rowData(obj)$bias)),
          all(rowData(obj)$bias>=0 & rowData(obj)$bias<=1))
stopifnot(dir.create("chromvar-chunks-001"))
set.seed(2026)
bg <- getBackgroundPeaks(obj,niterations=50)
expect <- computeExpectations(obj)
saveRDS(list(background=bg,expectation=expect),
        "chromvar-chunks-001/background-expectation.rds")
writeLines(capture.output(sessionInfo()),"chromvar-chunks-001/session.txt")
for (start in seq.int(1L,ncol(motif),by=100L)) {
  ix <- seq.int(start,min(start+99L,ncol(motif)))
  dev <- computeDeviations(obj,annotations=motif[,ix,drop=FALSE],
                           background_peaks=bg,expectation=expect)
  stopifnot(nrow(dev)==length(ix),
            identical(colnames(dev),colnames(obj)))
  cat("motif chunk",start,"nonfinite",sum(!is.finite(deviationScores(dev))),"\n")
```

```
  saveRDS(list(motif_ids=colnames(motif)[ix],cell_ids=colnames(obj),result=dev),
          sprintf("chromvar-chunks-001/chunk-%08d.rds",start))
  rm(dev);gc()
}
```

当前核心要求每个峰在送入计算的细胞集合中都有计数，因此不直接拆细胞，避免出现块内全零峰。先查峰、细胞与GC输入，不能不记录就删除零深度列。代码要求这些质量处理已完成。背景计算仍使用全部既定细胞的平均可及性；每块重新生成背景或期望，会改变打分基线，不能直接合并成同一分析。

验证分块，而不是假定等价

正式扩量前，在能完整处理的小集合上固定同一bg和expect，分别做完整与分块计算，核对细胞顺序、motif顺序、得分与缺失位置。异常值可能涉及背景方差或无效集合，应结合输入和方法解释，不把NA/Inf批量替换为零。每块显式保留motif和细胞ID，按原motif清单拒绝遗漏与重复，同时核对各块细胞顺序完全相同。

偏差和Z分数是不同输出，Z分数不能直接当成该TF表达量或实验结合证据。相近motif可能共享峰集合，细胞组比较仍需合适的样本设计。画图颜色较深不自动意味着统计显著，也不能把单细胞数当生物学重复数。

定位仍然超出的那一步

motif分块仍保留完整计数，并未减少一个motif处理全部细胞时的背景数组。若单个motif或原始计数本身已超出容量，这个分块方案也不够；应按实际失败阶段调整并验证可用实现，不自行删峰重编号后沿用旧背景。

输出规模也可先作演算：1000个motif×10万个细胞仅一份双精度数值主体就约0.8GB，偏差与Z分数两份还需更多空间，未含对象开销。稀疏峰计数即便很小，得分也可能稠密；按需要读取motif输出块，避免立即全部合并又回到原来的内存峰值。

用独立R进程及GNU Time记录完整墙钟与主机RAM，同时看背景生成和单块打分日志。若在getBackgroundPeaks前后失败，缩小motif块无法解决背景工作集；若多个任务同时打分，先收紧外层并发与BiocParallel。对象大小不包括全部内部临时量，释放变量也不保证操作系统立刻回收所有内存。

峰过滤、motif范围、背景抽样次数和细胞集合都属于方法设置，改变后另存协议并重新验收。仅将计数转成稀疏格式不能消除得分矩阵开销，直接as.matrix会使问题恶化。保留输入、背景、期望和输出块的共存空间；后续全量合并、降维或热图还要单独测量。

为什么背景峰应固定

chromVAR用GC含量与平均可及性相近的背景峰来构建对照期望。若每个motif块分别重新抽背景，背景期望会因抽样和参数变化而不同，块间结果就不再来自同一基线。示例先生成一次bg并固定随机种子，再按motif分块计算，便于控制结果对照；正式流程要记录背景矩阵和峰顺序散列。

背景匹配依赖与计数相同的峰集及GC偏好信息。若更换peak set、过滤峰或重新生成bias，旧bg就不再对应当前对象。空匹配motif应作为明确状态处理，不应静默删除后忘记原始motif清单。背景集合的抽样数和随机性会影响期望及方差，需在资源和统计解释中说明。

分块计算和内存仍要全局验收

按motif分块可以避免一次保留全部motif×细胞得分，但峰×细胞counts、GC bias、background indices和当前块结果仍在内存。把输出块保存后释放临时对象，记录每块motif ID、输出维度和完成状态；最后按原始motif顺序合并，检查每个motif恰好出现一次、没有缺项和重复。

全量背景生成本身也可能是资源热点。分别测读取对象、计算背景、计算期望和每块deviation阶段，不要只测分块循环。改变niterations或块大小可能改变计算时间与结果稳定性；先在固定对象上验证不同块大小只改变分片方式而不改变总体输出，保存比较证据。

GC偏好和motif映射

addGCBias需要与peak coordinates匹配的正确参考序列。bias字段应有限且位于预期范围，峰ID顺序与counts、motif匹配矩阵一致。若峰坐标版本不匹配，GC比例会错误；程序仍可能生成数值，但背景匹配失去意义。抽查几个峰坐标和GC结果，并记录参考序列版本。

motif矩阵行是峰、列是motif/注释；检查峰顺序、motif名和非零匹配数。无匹配峰的motif不能产生有意义的偏差估计。计算完成后将motif ID映射到TF时记录数据库版本及多对多关系，避免把同一基序的家族成员压成唯一因子。

输出解释与样本设计

chromVAR deviations相对背景期望描述motif集合可及性的偏离，不是直接的TF表达量或结合强度。Z-score等输出还依赖背景集合和期望计算。要解释样本或细胞状态差异，应结合实验重复、细胞组成、测序深度和其他证据，不能把每个细胞当作独立实验重复。

多样本任务要保留sample列、批次、条件和真实重复单位。可视化按预先定义的分组展示，并报告每motif的结果覆盖。若某motif只匹配少数峰或某样本覆盖很低，应标记结果限制而非与充分覆盖的motif同等解释。

任务清单和资源记录

记录Bioconductor、chromVAR、MotifDb/JASPAR来源、对象散列、峰与细胞数量、背景抽样参数、seed、块大小、线程、峰值RSS和磁盘。结果目录按输入版本和attempt唯一命名，避免多个分析者覆盖共享块文件。

资源申请提供GC参考、motif数、峰数、细胞数和固定背景矩阵大小。若瓶颈是背景峰生成或大对象读取，增加motif分块并不一定解决；若瓶颈是得分矩阵，再评估块大小。提交前确认所有块齐全并与manifest一致。

背景抽样的检查和敏感性

固定背景矩阵后检查维度、列数、索引范围和峰顺序，确保每个背景索引指向当前obj中的峰。随机种子应在生成背景前设置，并记入manifest。若改变背景迭代数或匹配峰规则，重新计算期望并标记为新方法版本；不应让一部分motif使用旧背景、另一部分使用新背景。

对关键motif可评估背景设置变化时分数排序是否稳定，重点观察峰匹配数较少或GC偏差明显的集合。背景集用于建立统计参照，不是生物学阴性对照。无法匹配足够背景峰时应记录原因并谨慎解释，不能随意放宽条件而不披露。

块结果合并与失败恢复

为每个motif块写一个状态文件，包含块编号、motif ID、输入散列、开始结束时间、退出码和输出散列。合并时按原始motif顺序检查完整性，失败块单独重跑。不要通过复制上次运行的同名输出跳过未确认完成的块。

跨块合并后抽查若干motif，比较整块运行和分块运行在同一背景及期望下的结果。若差异超出预先设定的数值容差，检查分块时是否错误重算期望、改变峰顺序或改变随机背景。只有完整检查后才能把合并对象标为通过。

错误信息与资源路径

若背景峰生成失败，先检查rowData bias、counts非零、峰注释与参考是否匹配；若某motif得分缺失，检查匹配峰数、块输出和对象列名。系统OOM时看作业内存限制与峰值记录，R错误则读完整调用栈。两类问题的处置不同，不要只扩大内存后重复同一输入。

每次试跑记录输入对象散列、背景矩阵散列、峰和细胞数量、motif块边界及软件版本。发生中断后确认旧R进程结束，保留已完成块的清单并只重跑失败块。若背景或输入发生变化，则所有块应一起重算，不能新旧混合。

可复核的单motif示例

报告中挑选一个代表motif，列出数据库ID、匹配峰数量、deviation在各样本或细胞群的分布及背景设定。该例用于说明如何读表，不应只展示最显著项。检查motif名称映射到TF家族的歧义，并在文字中区分motif activity proxy和TF实际表达/占位。

背景期望和deviation输出的对象结构依版本而异。按安装版本确认函数返回值和assay名称，合并块时保留细胞顺序、motif标签和背景参数。升级后不要用旧脚本假设的列名自动合并，否则可能产生尺寸匹配但标签错位的矩阵。

在同一固定输入上抽取少量motif完整计算，与分块路径结果核对数值和维度。该检查验证分块流程的一致性，不验证motif注释或生物学解释。

若一部分motif块失败，不要用上一次运行中没有manifest的文件填补。重新计算失败块前核对输入对象和背景矩阵散列一致，再把新旧块合并并验证每个motif恰好一次。若对象或背景已变化，所有块都应重新计算。

建议为结果保存元数据表，明确每列对应细胞、每行对应motif，以及deviation值和Z-score分别来自哪个assay。热图色阶可影响视觉判断，需保存绘图参数并说明是否进行中心化、缩放或裁剪。

计算期望与deviation时保持同一峰集合和细胞顺序。每块输出文件应保存其motif ID向量，合并后与原始motif清单逐项比较。发现重复或缺失ID时停止合并并追查块边界。

对样本级统计使用生物学样本作为重复，单细胞分数可用于展示分布但不能替代重复设计。跨条件的推断需采用适合数据结构的模型。

在可复现实验中把背景矩阵和随机种子作为关键中间证据保存，文件较大时可保存生成脚本、参数和散列，并按项目存储策略保留实际对象。

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

chromVAR偏差和背景使用说明 · 来源资料

chromVAR期望计算接口 · 来源资料

chromVAR当前打分实现 · 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。