# inferCNV大数据任务教程：核对输入、分阶段试跑与内存监控

完整操作教程 · 参考固定 · 往返恢复验收

为10x表达型CNV推断建立明确输入与方法基线，保留阶段对象和日志，分别记录预处理、HMM及绘图的资源需求。

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。

inferCNV根据RNA表达沿染色体的位置模式推断相对CNV信号，不能直接替代DNA检测。本文针对已完成细胞质控、具有合适正常参考的10x Gene Expression计数，建立一次可复核的完整任务。先核对细胞注释与基因位置，再观察错误发生在哪个阶段，避免把参考不合适解释成服务器不足。

## 先验收三类必需输入

| 输入 | 格式 | 验收要点 |
| --- | --- | --- |
| counts.rds | 基因乘细胞dgCMatrix原始计数 | 唯一基因和细胞名称；非负整数 |
| annotations.tsv | 无表头的细胞、群体两列 | 恰好覆盖计数列；正常群体可信 |
| gene_order.tsv | 无表头的基因、染色体、起点、终点 | 与表达使用相同ID及基因组组装 |

另准备refs.txt，每行一个经确认的正常参考群体名。基因位置由对应组装的注释生成，保留来源与坐标约定；不把另一版本的BED坐标直接改名使用。下例允许表达矩阵有未进入位置表的基因，但会打印交集数量，应解释丢弃范围，而非看到程序运行就忽略。

```
library(infercnv)
library(Matrix)
out <- commandArgs(TRUE)
stopifnot(length(out)==1,!file.exists(out),dir.create(out,recursive=TRUE))
z <- readRDS("counts.rds")
ann <- read.delim("annotations.tsv",header=FALSE,stringsAsFactors=FALSE)
pos <- read.delim("gene_order.tsv",header=FALSE,stringsAsFactors=FALSE)
refs <- readLines("refs.txt"); refs <- refs[nzchar(refs)]
stopifnot(is(z,"dgCMatrix"),ncol(ann)==2,ncol(pos)==4,
 !is.null(rownames(z)),!is.null(colnames(z)),!anyNA(dimnames(z)[[1]]),
 !anyNA(dimnames(z)[[2]]),!anyDuplicated(rownames(z)),!anyDuplicated(colnames(z)),
 all(is.finite(z@x)),all(z@x>=0),all(z@x==floor(z@x)),
 !anyNA(ann),!anyNA(pos),!anyDuplicated(ann[[1]]),
 !anyDuplicated(pos[[1]]),setequal(ann[[1]],colnames(z)),
 all(refs %in% ann[[2]]),length(refs)>0,
 all(is.finite(pos[[3]])),all(is.finite(pos[[4]])),
 all(pos[[3]]>=0),all(pos[[4]]>=pos[[3]]))
cat("position_overlap=",sum(rownames(z) %in% pos[[1]]),"of",nrow(z),"\n")
obj <- CreateInfercnvObject(raw_counts_matrix=z,
 annotations_file="annotations.tsv",gene_order_file="gene_order.tsv",
 ref_group_names=refs,delim="\t",max_cells_per_group=NULL)
saveRDS(obj,file.path(out,"created-object.rds"))
writeLines(capture.output(sessionInfo()),file.path(out,"session.txt"))
ans <- infercnv::run(obj,cutoff=0.1,out_dir=out,cluster_by_groups=TRUE,
 denoise=TRUE,HMM=TRUE,num_threads=1,resume_mode=FALSE,save_rds=TRUE,
 save_final_rds=TRUE)
stopifnot(file.exists(file.path(out,"run.final.infercnv_obj")))
saveRDS(ans,file.path(out,"returned-object.rds"))
```

保存为cnv.R，先在方法已确认的输入上运行。cutoff=0.1来自官方10x示例，denoise及HMM也属于这里明确选择的方法；其他建库不能照搬。创建函数还有计数过滤与默认染色体排除，应检查创建日志和对象维度，确认保留范围符合项目方案。

```
/usr/bin/time -v -o cnv.time.txt Rscript cnv.R cnv-attempt1 > cnv.log 2>&1
```

## 按阶段判断失败，而不是只看热图

用日志标记预处理、平滑、聚类、HMM与绘图的开始和结束，把失败前最后阶段写入任务记录。GNU Time提供整进程峰值；要定位阶段峰值，还需在调度器或监控工具中按时间采样。线程数不是硬内存额度，稀疏输入也不保证全部中间步骤保持稀疏。

官方run接口提供阶段RDS与resume_mode。代码首先关闭恢复，以免旧对象掩盖完整耗时。真正重启时应保留同一目录、输入、版本和方法参数，按该版本恢复规则开启resume_mode并确认日志实际复用了哪一步；不要把另一参数方案的中间文件搬来混用。

## 验收与容量结论

最终对象存在只是计算验收。还需核对参考组、保留基因和细胞数，查看参考是否出现系统性偏移、观测群体模式及HMM相关警告，并结合独立证据解释。若先做代表性小范围试跑，记录明确选择规则；小数据峰值只能用于阶段诊断，不能按细胞数简单线性保证全量容量。

HMM、去噪、窗口、群体组织与细胞抽样均可能改变结果，不应为了内存私自调整后仍称为同一分析。方法和范围固定后若必要步骤确有RAM缺口，再安排更大内存与足够工作盘；把多个患者并发时，预算应覆盖各自阶段峰值重叠和阶段RDS占盘。

## 补齐本次运行记录

建议结果表逐行填写输入细胞数、参考群体与细胞数、位置交集基因数、创建后保留范围、各阶段退出状态和最终对象路径。参数改变或参考群体变更时另建任务目录；对照两次信号变化时，先确认比较的是同一细胞与基因范围。

## 参考细胞定义了 inferCNV 的基线

inferCNV 将表达模式沿染色体排序和平滑后，与参考细胞表达基线比较以突出大尺度拷贝数样信号。reference cell 群体的选择会影响中心化结果。优先根据实验设计与细胞注释选用可信的正常细胞，确认样本来源、批次和细胞类型适合比较；若使用外部参考或不提供参考组，必须说明基线构建假设。把肿瘤细胞误放进 reference 会压低真实改变，正常细胞混入观察组也会影响图形解释。

输入至少要保证 raw counts 矩阵、细胞注释与 gene order 文件之间 ID 完全匹配。核对 cell barcode 去重、注释类别、基因符号版本、染色体命名和基因组坐标排序。gene order 中缺失或重复基因会改变连续平滑；先统计交集和丢弃项，再决定是否运行。

## cutoff 依测序技术与输入尺度决定

官方指南示例中 Smart-seq 与 10x 数据采用不同 cutoff 参考值，说明阈值与测序技术和表达尺度相关。不能将一套项目的 cutoff 照搬到另一套数据。结合表达分布、每细胞测序深度、过滤后基因数和参考信号选择阈值，并记录阈值改变后保留的基因数量。cutoff 过高可能丢失低表达但有用的基因，过低会增加噪声和内存负担。

inferCNV 期望输入细胞已做基本质量控制。先过滤低质量细胞、核对每个细胞 counts 范围和注释，再比较不同 cutoff 的代表结果。避免在没有记录的情况下同时改变 cutoff、参考组和 denoise 参数，否则很难判断图形变化由何种因素造成。

## HMM 输出不是 DNA 测序结果

HMM=TRUE 会根据表达推断状态段，输出是 RNA-seq 表达推断的 CNV 候选，不等于 DNA 层面的拷贝数实测。状态数、平滑、参考基线、细胞分群与模型参数会改变结果。区分原始 inferCNV 热图、denoise 后信号和 HMM 状态输出，报告使用哪一种作为下游结果，不要混称为同一种预测。

对重要扩增或缺失事件，结合 DNA 测序、FISH、已知细胞系或其他独立证据验证。细胞周期、细胞类型表达差异和染色体区域基因密度也可能造成类似连续信号。观察到一条染色体颜色变化不应直接写成已确认拷贝数改变。

## 大矩阵的内存与阶段运行

内存需求与保留的基因数、细胞数、矩阵副本、平滑与聚类步骤、HMM 和绘图共同相关。R 中稠密矩阵可能远大于输入稀疏矩阵，转换时还会产生临时副本。先用代表子集和真实参数跑通完整阶段，记录 `object.size` 只能帮助了解对象本身，作业峰值仍需由调度器或系统监控测量。

降低细胞或基因数会改变推断结果，不应仅为了装进内存而无记录地下采样。先查明是矩阵读入、聚类、HMM 还是绘图阶段占用峰值；根据真实瓶颈选择分批、减少并发、调整支持的参数或增加内存。每次资源调整都保存分析参数并比较科学输出。

## 图形与输出的复核清单

运行后检查日志、对象保存、热图、基因排序与细胞分组。确认参考组与观察组标签、染色体顺序和基因数符合预期；图形范围、颜色和阈值会影响视觉印象，应配套输出数据和参数。不同样本的热图若采用独立色标，颜色强度不一定可直接横向比较。

交付保存 counts 来源、细胞过滤、注释文件、gene order、参考群、cutoff、denoise/HMM 参数、R 包版本、运行日志、峰值内存和结果目录清单。把 CNV 候选标注为推断结果，并注明需要独立验证。数据和细胞注释若含敏感信息，存储路径与共享权限应符合单位要求。

## 组织与细胞注释会改变参考比较

观察细胞与参考细胞的细胞类型和批次构成应在分析前核对。参考群不应只是“看起来正常”的标签；需要结合组织来源、病理状态和表达质量判断。不同正常细胞类型基线差异较大时，选择单一参考或多个参考组都会影响中心化结果，建议依据实验设计比较合理的参考方案，并解释取舍。

如果没有合适正常细胞，使用整体平均信号作基线会引入不同假设，尤其当观察群中多数细胞都可能有相同方向改变时。此时需明确模型假设，避免把相对基线信号当作绝对拷贝数。尽量寻找匹配正常样本或独立 DNA 层面证据。

## R 对象和临时文件的空间需求

inferCNV 会生成多个中间对象、热图和 HMM 文件。输出目录可能包含大型 RDS、矩阵与图形；运行前确认输出盘可用空间、inode 与配额，设置每个项目单独目录。若使用并行计算，进程副本会增加 RAM 消耗；多线程不代表内存占用只计算一次。记录 R 会话、包版本、线程数和作业级峰值。

失败后查看日志最后完成的步骤和对象文件状态，再判断是否支持从中间对象恢复。不要将上一次运行残留目录直接作为新参数运行目录，否则新旧输出可能混杂。保留完整结果后再清理可重建临时文件，并记录删除策略。

## 把推断结果与细胞标注一起交付

热图和 HMM 状态应附细胞分组、参考组定义、基因排序版本和主要参数，避免图像被单独分享后无法解释。图中染色体位置、色标范围、去噪设置和聚类顺序都会影响视觉判断。对于候选异常区域，导出对应基因列表和细胞范围供后续检查。

若样本有已知肿瘤纯度、组织学或 DNA 变异信息，结合这些资料做交叉验证。inferCNV 适合从单细胞表达观察大尺度拷贝数样模式，不应单独作为临床诊断或精细断点定位证据。

## 比较参数时固定输入对象

cutoff、denoise、HMM、聚类方式和参考组应逐项比较。一次只改变少数参数，并在同一输入对象、软件版本和随机条件下保存运行目录。不同运行的热图颜色范围与图形分辨率要统一后再比较，否则视觉差别可能来自绘图设置。

批量样本分析按样本保存运行参数和退出状态，汇总哪些对象成功、哪些因输入注释或资源不足失败。不要只统计成功样本后忽略失败样本的系统性特征；失败本身可能与细胞数或数据质量相关。

## 把推断结论限制在表达数据分辨率

inferCNV 对沿染色体连续变化的表达信号进行平滑，适合识别大尺度 CNV 样模式，不适合精确定位断点或代替 DNA 测序。基因密度、表达噪声和基因组注释都会影响信号边界。将图中区段作为候选区域，并用更高分辨率的数据验证。

若同一事件在不同参考群、cutoff 或平滑参数下不稳定，应将它标记为不确定并避免强结论。保存敏感性比较的参数与图形，说明结论在哪些设置下保持一致。

## 确认对象序列化后可以恢复

inferCNV 中间 RDS 保存后，使用同一包版本在独立 R 会话尝试读取并检查对象维度、细胞注释与 gene order。仅有文件存在或大小合理，不足以证明对象完整。若输出目录迁移，更新记录中的绝对路径，并在新位置重新打开关键对象后再清理旧副本。

## 按项目计划保留中间结果

输出目录可能含热图、矩阵、对象和运行日志。先确认哪些是可重新生成的缓存、哪些是唯一审核证据，再按团队归档规则清理。数据敏感时限制中间表达矩阵的访问范围。

## 把参考细胞选择写进结果解释

参考细胞不是一个普通的运行参数，它决定了表达信号相对于什么基线被比较。建议在方法和结果中说明每个参考组的细胞类型、样本来源、纳入数量、排除标准以及选择理由；若有多个候选正常群体，可做敏感性分析，观察主要 CNV 样式是否稳定。恶性细胞标注不确定、参考群体混入肿瘤细胞，或者不同样本的细胞组成差异很大，都可能改变热图的相对模式。对少量候选染色体片段，应检查基因位置和连续性，确认不是低表达、测序深度、细胞周期或批次造成的偏移。把 inferCNV 结果用于细胞注释时，最好与病理信息、突变或拷贝数实验、已知标记等独立证据交叉核对；如果没有独立证据，应把它描述为支持性线索，而不是确诊。

## 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


inferCNV官方运行教程 · 来源资料

inferCNV run参数及恢复实现 · 来源资料

输入对象与过滤参数 · 来源资料

inferCNV 官方运行指南 · 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。