Mixscape扰动分析实操：输入核对、对照设置与任务批处理

完整操作教程 · 参考固定 · 往返恢复验收

明确gRNA标签、非靶向对照与重复关系，计算扰动签名并保存Mixscape完整对象和分类验收表。

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。

Mixscape用于区分携带靶向gRNA的细胞中是否存在可检测的转录扰动。开始批处理前，要先把gRNA分配、靶基因和非靶向对照对应好。gRNA被检测到不代表编辑一定成功，Mixscape的KO标签也不是直接测出的基因组敲除证据。

先限定本轮比较范围

本例是一组已经选定、适合比较的细胞群，输入为完成RNA常规质控、归一化、变异基因、缩放与PCA的Seurat对象。分析混合细胞类型或不同处理条件前，应先设计比较与对照；不能让细胞类型组成差异直接替代扰动效应。

对象内容	本例约定	错误时先查
RNA	原始计数和可用归一化data	多个未合并层或过期assay接口
gene	靶基因标签，非靶向统一写NT	guide编号与靶基因映射混淆
replicate	真实生物学重复标识	不要把随机切块命名为重复
PCA	至少三十维，来源与RNA预处理一致	缺失降维不能临时用行号替代

固定对照池后计算扰动签名

当前Seurat 6.0官方API仍在这两个函数中使用slot参数，不能机械改成layer。保存对象前核对安装版与官方流程兼容。本例每个重复至少二十一个NT细胞，是为二十邻居设置的保守前置条件，不是统计效能保证；对照质量和数量仍需按设计评估。

保存为mixscape.R，输入文件是mixscape-input.rds，输出目录必须新建。CalcPerturbSig按replicate分别寻找对照邻居；RunMixscape的split.by则用于处理条件或细胞类型特异效应，意义不同，本例不额外设置。研究若需要该分层，应先确定每层有效对照和细胞数。

```
library(Seurat)
obj <- readRDS("mixscape-input.rds")
stopifnot(inherits(obj, "Seurat"), "RNA" %in% Assays(obj),
          "pca" %in% Reductions(obj), !"PRTB" %in% Assays(obj))
meta <- obj[[]]
stopifnot(all(c("gene", "replicate") %in% colnames(meta)),
          identical(rownames(meta), colnames(obj)),
          !anyDuplicated(colnames(obj)),
          !anyNA(meta[, c("gene", "replicate")]),
          all(nzchar(as.character(meta$gene))),
          all(nzchar(as.character(meta$replicate))),
          any(meta$gene == "NT"), any(meta$gene != "NT"),
          !any(grepl("^mixscape_class", colnames(meta))),
          ncol(Embeddings(obj, "pca")) >= 30)
n_nt <- tapply(meta$gene == "NT", meta$replicate, sum)
stopifnot(all(n_nt >= 21), length(VariableFeatures(obj[["RNA"]])) > 0)
out <- "mixscape-run01"
stopifnot(!file.exists(out), dir.create(out))
set.seed(101)
obj <- CalcPerturbSig(obj, assay="RNA", slot="data", gd.class="gene",
  nt.cell.class="NT", split.by="replicate", reduction="pca", ndims=30,
  num.neighbors=20, new.assay.name="PRTB")
DefaultAssay(obj) <- "PRTB"
VariableFeatures(obj) <- VariableFeatures(obj[["RNA"]])
```

```
obj <- ScaleData(obj, do.scale=FALSE, do.center=TRUE)
obj <- RunMixscape(obj, assay="PRTB", slot="scale.data", labels="gene",
  nt.class.name="NT", min.de.genes=5, min.cells=5, iter.num=10,
  de.assay="RNA", verbose=FALSE, prtb.type="KO")
stopifnot("mixscape_class.global" %in% colnames(obj[[]]),
          identical(colnames(obj), rownames(meta)))
saveRDS(obj, file.path(out, "mixscape.rds"))
write.csv(obj[[]], file.path(out, "metadata.csv"))
write.csv(as.data.frame(table(obj$replicate, obj$gene,
                              obj$mixscape_class.global)),
          file.path(out, "class-counts.csv"), row.names=FALSE)
writeLines(capture.output(sessionInfo()), file.path(out, "session.txt"))
```

验收分类与生物学解释

metadata.csv保留原标签和分类信息，class-counts.csv按重复、靶基因及NT、NP、KO计数。检查所有原细胞是否保留、标签是否缺失、各组是否有足够细胞，并查看扰动分数和后验概率分布。min.cells以下的靶类别可能全判NP，不应解释为编辑无效的确定结论。

对重点靶基因核对预期通路、独立蛋白或其他实验指标，并比较不同gRNA和重复的一致性。NP表示方法未检出相应转录扰动，也可能受细胞状态、信号弱、统计信息不足影响。仅提高KO比例不是调参目标，min.de.genes、迭代设置或分组变化都需作为分析条件记录。

批量运行与失败分流

以有完整对照上下文的独立实验集为任务，不随机拆成几块细胞后直接拼分类，因为邻居池和模型会改变。参数探索也用独立目录，保存输入校验、随机种子、版本和日志，避免覆盖上一轮。使用Rscript mixscape.R启动，资源记录应覆盖签名、缩放与分类全过程。

若报找不到data或scale.data，先查assay、层及预处理顺序；找不到NT或邻居不足，先查元数据和重复内对照。若进程被终止，再查作业错误与实际内存限额。PRTB及中心化矩阵可能产生较大内存载荷，先用科学上有效的小设计核查，不能用随意减少对照伪装成等价低内存运行。

汇报时同时列出每组总细胞数、有效对照数和分类计数，避免只报百分比。小组的比例波动不能与大组等同解释，异常组回查输入和实验记录后再决定是否进入后续比较。

批次清单与对照设计

在运行Mixscape前，将每个细胞的guide、靶基因、非靶向状态、donor、replicate、处理条件、细胞类型及测序批次放入可审计的元数据表。guide到gene的映射需来自实验设计并核对多gRNA靶向情况，不能只依赖列名推测。不同生物学重复必须是真实独立实验单位，随机将细胞拆分后命名为replicate会造成伪重复，改变对照邻居及统计解释。

NT对照应与扰动细胞处于可比的实验流程、细胞类型和批次。每个replicate的NT数量门槛只是邻居计算的输入守卫，不保证对照能代表所有目标细胞。运行前查看NT数量、各guide细胞数、细胞类型构成和RNA质控，若某个靶点只在单一批次或单一细胞群出现，需谨慎解释并考虑分层方案。

扰动签名和分类步骤分别检查

CalcPerturbSig先依据指定对照邻居为扰动细胞生成签名；replicate分层用于限制对照匹配范围。检查PCA确实来自用于比较的RNA处理流程，维度足够且细胞顺序与对象元数据一致。若改变ndims或邻居数，应在方法上说明并评估签名稳定性，不要为了让所有扰动都能分类而任意增加维度或减少邻居。

RunMixscape再利用扰动签名进行分类。检查PRTB assay是否按预期创建、中心化是否使用了正确变量基因、分类元数据列是否包含每个原始细胞且没有丢失ID。KO/NP等标签是基于表达扰动模式推断的类别，不是DNA编辑状态的直接检测。可将分类与guide检测、表达变化、编辑效率和独立验证并列分析。

条件和细胞类型分层的选择

split.by在CalcPerturbSig和RunMixscape中承担的作用应分别查阅当前Seurat版本文档，不要认为两个函数中的同名参数做同一件事。若研究问题关注细胞类型特异响应，应先保证每个层级都有足够NT和扰动细胞，再按层级估算。小层级样本太少时结果可能不稳定，应报告未分类或未分析，不将多个条件无依据合并。

分层越细，细胞量和计算次数越多，但并行运行并不一定更快；多个Seurat对象副本会增加内存。先在一个分层上建立可复现基线，然后逐层运行并记录细胞数、线程、内存和分类比例。细胞随机拆块会让不同块拥有不同NT邻居池，因此不能把此类拆分当成无影响的资源优化。

异常结果排查与科学解释

若几乎所有细胞都被分到同一类，检查NT定义、guide映射、基因表达矩阵、PRTB签名和迭代分类参数；若某replicate报错，先查看该重复的NT数量和PCA数据是否有效。某些基因扰动未产生可检测转录变化是可能的，不应为了获得预期KO比例而反复调阈值。保留所有参数试验并说明选择理由。

比较不同靶点时，报告每个guide的细胞数、replicate分布、分类概率或分数、未分类比例和采用的阈值。低细胞数或失衡对照会影响分类结果。对重要发现开展外部证据核对，如靶基因表达、通路变化或独立实验验证；分类模型本身不证明编辑成功或扰动的因果机制。

结果交付与资源记录

保存原始Seurat对象散列、处理脚本提交、Seurat版本、完整参数、随机种子、分层规则和输出对象。导出每细胞元数据时保留稳定cell ID，并核对对象列顺序。将分类汇总表与样本清单关联，避免只交一列mixscape_class而无法知道每个细胞来自哪个guide、replicate或条件。输出图应标注所用assay和分组，不只展示有利的结果。

记录PCA、扰动签名计算和分类阶段的耗时及峰值内存，并注明作业额度。若读入对象就占用大部分内存，减少分类阶段迭代不会解决输入规模问题；可考虑按预先有科学依据的条件分开处理，同时保留共同分析的边界。对多任务并发前先测单任务资源，避免多个R进程重复加载完整Seurat对象。

把方法选择和服务器建议分开记录

服务器资源记录包含CPU、RAM、作业额度、输入Seurat对象大小、细胞数、基因数、PCA维度和每阶段峰值；方法记录则包含NT定义、重复单位、邻居数、签名分层、分类参数和阈值。这样当分类比例变化时，可以先判断是计算环境变化，还是实验/模型设定变化。调整资源不应静默改变邻居池或随机拆分方式。

如果需要减少内存，可先删除确定不再需要的临时对象或在保持对照设计的前提下按条件顺序执行；任何按细胞拆分都需验证每块仍有代表性NT对照。不要直接把不同随机块叫作replicate，也不要仅为降低输入规模而删除稀有细胞群，除非该排除标准有独立科学依据并已记录。

复现和版本升级

固定Seurat、Mixscape来源、R版本和函数参数，保存sessionInfo与脚本提交。Seurat对象在不同大版本间的assay/layer/slot访问可能变化，因此先检查当前函数签名和官方教程，确认对象中RNA、PRTB和scale.data的位置。升级后用固定小样本重跑，比较扰动签名、分类比例和元数据列；不能只因新版本运行成功就覆盖旧项目结果。

若结果依赖随机过程，记录随机种子并在同版本环境中重复试跑，评估分类结果稳定性。种子固定不代表不同版本、线程库和平台逐位相同。对重要靶点可比较多个合理参数下分类的一致性，并保存选择规则。不能只挑最符合预期KO比例的一次运行。

判断什么时候数据不足

某个guide细胞数极少、某个重复缺少NT或目标基因低表达时，模型输出的NP/KO比例可能不稳定。报告每个guide和replicate的细胞量与对照量，明确哪些组合未达到项目预设标准。不要将低细胞数结果与充足重复结果放在同一证据等级。增加计算资源可以缩短运行时间，但不能制造缺失的实验重复或增加扰动信号。

若分类结果与guide捕获、靶基因表达或独立实验不一致，先检查guide-to-gene映射、RNA处理、对照可比性和模型参数。对每个冲突保留原始预测和复核记录，分析可能技术原因；不要直接手工改写分类标签后将其称为模型结果。若结果仍有歧义，报告为待验证结论。

对照池变化的影响

replicate内NT邻居池大小和组成会影响扰动签名。对每个replicate统计NT细胞数量、细胞类型比例及PCA空间分布；如果NT只覆盖部分状态，其他扰动细胞可能缺少可比邻居。必要时依照预先设计按细胞类型或条件进一步分层，并确保每个层级有足够对照。分层定义应在看结果前确定，避免按分类结果再反向调整对照池。

如果某些guide或条件没有匹配NT，不能借用不相干批次的对照来强行计算。记录未分析的组合及原因，考虑补充实验或采用有依据的统计方案。分类结果可作为扰动响应证据的一部分，但还需检查目标基因表达、通路变化或其他实验读数。缺失的对照不是服务器配置可以补足的。

若多个项目共享Mixscape环境，应将包版本和R库路径固定在项目环境中，避免某个用户升级全局R包影响其他分析。提交作业前确认使用的Rscript可执行文件、项目库和输入对象路径，防止交互式R与批处理脚本调用不同版本。

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官方Mixscape教程 · 来源资料

CalcPerturbSig当前接口 · 来源资料

RunMixscape当前接口 · 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。