Azimuth本地部署教程：参考下载、版本匹配与查询样例验证

完整操作教程 · 参考固定 · 往返恢复验收

在独立兼容环境中固定Azimuth与参考材料，使用原始RNA counts的小查询验证本地映射和注释输出。

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。

Azimuth本地映射能避免把查询数据上传到网页，但仍需准备合适参考和兼容R环境。本篇限定RNA查询与官方RunAzimuth入口；ATAC桥接、多模态和新的其他映射路线有不同输入，不能直接复用这段RNA代码。

先固定软件与参考的组合

在新项目环境按官方安装说明准备Azimuth、Seurat及依赖，记录R版本、软件提交或包版本与锁文件。系统编译库由维护者按范围部署，不在正在使用的共享R库中一键更新全部包。旧版网页vignette的示例环境不等于任意新Seurat版本都已经验证兼容。

官方既支持参考名自动下载，也支持本地参考目录。本例选择固定本地目录：从官方Azimuth参考页取得对应材料，保留下载链接、发布信息与发布方校验信息；展开后应有ref.Rds和idx.annoy。只自己计算一个哈希能标识文件，不能独自证明下载来源可信。

检查	需要一致	不宜混用
参考适配	物种、组织、技术与注释层级	不适配参考可能给出貌似精细标签
查询	未标准化RNA counts与唯一基因/细胞ID	不拿scale.data当原始counts
文件	ref.Rds、idx.annoy及同一发布	不从不同目录拼凑参考
环境	R、Azimuth、Seurat和依赖组合	不把安装成功当映射验收

运行一个本地查询样例

下面要求已验证兼容的环境，query.rds是有单一RNA counts层的可信Seurat对象，REFERENCE_DIR是实际参考绝对路径。拆分counts层或旧对象先按所用版本处理并记录，不在这里自动合并或更改表达数据。输出目录必须不存在。

```
library(Seurat)
library(Azimuth)
reference <- Sys.getenv("REFERENCE_DIR")
stopifnot(nzchar(reference))
stopifnot(all(file.exists(file.path(reference, c("ref.Rds", "idx.annoy")))))
query <- readRDS("query.rds")
stopifnot(inherits(query, "Seurat"), "RNA" %in% Assays(query))
stopifnot(!anyDuplicated(colnames(query)), !anyDuplicated(rownames(query[["RNA"]])))
counts <- GetAssayData(query, assay="RNA", layer="counts")
stopifnot(ncol(counts) > 0, nrow(counts) > 0)
stopifnot(!any(grepl("^predicted\\.", colnames(query[[]]))))
stopifnot(!dir.exists("azimuth-pilot"), dir.create("azimuth-pilot"))
mapped <- RunAzimuth(query=query, reference=reference,
                     query.modality="RNA", assay="RNA")
stopifnot(identical(colnames(mapped), colnames(query)))
fields <- grep("^predicted\\.", colnames(mapped[[]]), value=TRUE)
stopifnot(length(fields) > 0)
saveRDS(mapped, "azimuth-pilot/mapped.rds")
write.csv(mapped[[]][, fields, drop=FALSE],
          "azimuth-pilot/predictions.csv")
writeLines(capture.output(sessionInfo()), "azimuth-pilot/session.txt")
```

GetAssayData的layer接口适用于相应SeuratObject版本，已被官方新层访问器取代但仍有兼容入口；安装版若不支持，应先匹配环境，不能把参数错误当查询内存不足。已有预测列会停止，以免混淆旧映射结果与本次输出。

核对输出与映射质量

确认退出成功、全部查询细胞对应、参考降维与注释列存在。不同参考的层级和分数命名不同，查看实际metadata并保存完整对象，不只导出某篇教程中的固定列名。对主要群体和低置信细胞检查标记、分数分布及参考覆盖。

标签很细不代表更可靠，参考没有的状态仍可能被映射到最接近类别。预测分数也不是独立的实验验证；疾病、发育或罕见状态需要结合项目证据，不为让标签比例符合预期而手工替换结果。

分别处理环境、输入与资源失败

参考路径或索引缺失先查完整发布；基因名交集异常先查ID体系和物种；函数或层接口错误先查版本。进入映射后出现确定RAM问题，记录查询规模、参考、阶段与作业峰值，再测代表任务。只用很小查询跑通不证明正式输入能完成。

最后保存参考指纹、查询清单、完整命令与结果验收。扩批保持同一环境和参考，每任务独立输出；升级参考或方法时另建运行，不在共享目录原地替换正在加载的材料。

下载参考时先确认研究对象和软件版本

Azimuth是以参考数据映射查询细胞的流程。实际可用参考取决于组织、参考构建方法、Seurat/Azimuth版本和参考发布状态。先确认参考覆盖的组织、细胞类型层级、物种和参考说明，再决定是否适合问题。参考标签不包含的细胞状态可能被分到相近类别，因此不能仅以返回了label列就断定注释可靠。

参考对象可能包括Seurat对象及索引文件，体积和组成依发布版本而变化。不要把旧教程中的固定文件地址、文件名或下载大小当作永久接口。下载时保存官方页面、版本号、时间和SHA256；解压后确认预期文件均在同一目录且可读。无法取得配套索引时，不要把其他版本的索引随手拼入参考目录。

本地离线运行的目的通常是减少重复下载并固定参考版本，不是免除兼容性检查。建议为每个参考建立不可变版本目录，查询结果写到独立运行目录；升级参考前先保留旧版本并在小批查询上并行比较预测标签、分数和内存。不同版本结果有变化，应作为方法变化记录而非静默覆盖。

先做一个查询样例并检查表达层

Seurat对象中assay、layer或slot的名称和行为随Seurat主版本变化。运行前检查RNA assay是否含预期counts，且细胞ID唯一、非空；查询数据的物种和基因标识符格式应与参考兼容。不能把已归一化矩阵误当原始counts，也不应为了凑齐基因名而无记录地改名或去除重复基因。

先挑一个较小查询样例运行，把输入细胞ID、输出预测表行名和对象细胞名逐一核对。预测结果至少保存原始查询ID、预测标签、对应层级和分数；将预测写回Seurat对象前先确认行名完全匹配，避免按位置赋值造成标签错位。若命令报缺少reference或索引文件，优先检查目录完整性和版本兼容，而不是重复尝试同一大查询。

Azimuth预测中的标签和分数取决于参考数据及算法。低分或多个标签接近时应保留不确定性，结合标记基因、实验来源和独立证据复核。参考映射用于将查询放到参考框架中理解，不能替代独立聚类、差异分析或实验验证；对于参考之外的新细胞群，合理结果也可能是无法高置信匹配。

评估运行时间、存储和可复现性

先在小查询确认流程后再扩大细胞规模，记录R和相关包版本、Seurat对象大小、参考目录校验值、查询细胞数、开始结束时间及峰值内存。多个进程并发加载同一个大型参考可能重复消耗内存；若分析任务相互独立，可在资源额度明确后安排并发，但应先测单任务峰值并确保磁盘有空间容纳日志和中间文件。

对于网络受限环境，可在联网且获准的节点一次性下载官方参考，再进行校验后迁移到分析节点。迁移前后对文件计算散列，避免传输截断。若参考包含多个配套文件，应把整套目录作为一个版本单元管理，保留下载清单和来源链接。不要把查询对象、参考对象和临时缓存混成不可区分的共享目录。

失败定位按“路径与权限—文件完整性—软件版本—Seurat对象结构—查询规模”顺序进行。一个小样本也失败，通常不该先申请更多内存；只有输入与版本已通过检查且峰值资源确实达到限制，才考虑扩展资源或改变任务拆分方式。改变细胞筛选范围、基因集或参考版本都会改变结果定义，需要重新验证映射稳定性。

参考映射后的质量复核表

每个输出标签最好连同label level、prediction score及查询样本来源一起保存。对照原始聚类、经典标记基因表达和实验条件，查看是否存在大簇被拆成多个低置信标签，或不同生物群被强行归为同一类型。检查预测低分细胞在UMAP上是否形成连片结构；若是，可能是参考覆盖不足，也可能是批次或预处理差异，需要回到数据本身判断。

使用参考的层级标签时，粗层级一致不代表细层级也可靠。可先报告大类，然后只在有足够分数、标记和数量支持时细分。阈值不应从别的组织或项目照搬；若项目设阈值，应记录依据、评估样本和被标为unknown的细胞比例。保留原始预测与人工修订记录，后续可追查每个标签为何被接受或改写。

批量映射时建立稳定的样本清单

批量查询前建立样本清单，至少含样本ID、Seurat对象路径、文件散列、细胞数、物种、建库批次、参考版本、运行目录和状态。路径最好使用无歧义的绝对路径，样本ID避免空格及会影响命令解析的特殊字符。一个样本成功不代表整个批次完成；脚本需要对每个样本分别记录退出状态，并在批次结束后检查清单中是否每项都有匹配结果。

重试失败样本时使用新尝试目录，保留上一轮日志和失败原因。不要为了让批次脚本继续而吞掉错误并在最终报告中只写成功数。汇总时将成功、失败、跳过和人工复核中样本分开统计，并确保预测表中的细胞ID没有在样本之间意外重复。若多个对象会同时加载参考，采用受控队列比同时拉起大量R进程更省资源，也更容易定位异常。

制作可复用的参考清单

建议对每个Azimuth参考维护reference_id、组织、物种、发布版本、官方来源、下载时间、文件相对路径、SHA256、R版本、Seurat版本、Azimuth版本和兼容性测试日期。共享参考目录设为只读，升级另建目录，不覆盖原文件。团队成员提交分析时引用reference_id和散列，就能避免“都用了某器官参考”但实际下载版本不同的情况。

如果参考下载包含索引、元数据和RDS等多种配套文件，应记录各文件散列并检查完整性。压缩包解开后与官方说明的目录结构做比对；不同参考的索引不要混放。磁盘空间计算也应考虑压缩包、解压后文件和运行缓存同时存在的阶段峰值，而非只看下载文件大小。

如何判断资源建议是否必要

先测一份覆盖常见细胞类型的查询，不只选极小数据；记录读入、映射、写出各阶段的时间和峰值。若方法卡在加载参考，缩小查询块未必有帮助；若只在高细胞数查询时失败，则可以评估按样本串行或减小单次查询。涉及参考本身、批次校正或表达层改变时，必须先确认映射定义没有改变。

向服务器服务方咨询前，准备软件版本、参考版本及大小、查询细胞数、对象所用assay/层、实测峰值内存、失败阶段和预计并发数。只说“跑Azimuth很慢”无法判断问题属于下载、兼容、内存还是其他限制。环境配置还需确认R包能否安装、参考是否可存放及项目数据的访问约束。资源应根据这些证据评估，而不是给所有课题组同一配置。

映射失败的样本也应保留在批次台账，写出失败阶段和下一步需要的证据。只有完成质量检查的样本进入后续统计；其余样本不能因为没有输出而从分母中消失。这样可准确报告覆盖率，也便于判断是否需要补充参考或改变实验设计。

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

Azimuth官方安装及本地参考文件 · 来源资料

RunAzimuth官方参数 · 来源资料

Azimuth官方RNA映射教程 · 来源资料

SeuratObject层数据访问接口 · 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。