整合前后的近邻追踪

从示例数据跑到第一张结果表

把整合前后两套低维坐标和原群条码放在一起。先数哪些条码还在，再在共同细胞全集里查询邻居。这个例子故意打乱第二张坐标表的行序，并漏掉一个原群成员。



准备 R 环境

需要 RANN；本练习用 RANN 2.6.1 验证。 示例已在 R 4.3.3 运行。

```
install.packages("RANN", repos="https://cloud.r-project.org")
```

完整解压资料包，在 run.R 所在文件夹打开终端，运行以下命令。

```
Rscript run.R example results
```

若使用 RStudio，将工作目录设为同一文件夹，再在控制台输入下面这一行。

```
source("run.R", encoding="UTF-8")
```

打开 results 文件夹查看输出。expected 文件夹中有示例的参考结果，示例数字均来自教学模拟。

逐步复算示例

在示例已经跑完的同一个资料包目录中打开 R。下面两段代码可依次粘贴，也可以执行包内的“读表练习.R”。

读第一张表

```
r <- read.delim("results/retention.tsv")
target <- subset(r, target)
with(target, table(in_before, in_after))
target$cell_id[!target$in_after]
```

示例中 6 个目标条码有 5 个进入两套坐标，D2_C15 只在整合前出现。先把它记为未匹配，剩余五个再比较近邻，避免把条码缺失解释成细胞状态消失。

用第二项结果核对

```
e <- read.delim("results/neighbour_list.tsv")
one <- subset(e, cell_id == "D1_C01")
a <- one$neighbour_id[one$space == "before"]
b <- one$neighbour_id[one$space == "after"]
length(intersect(a, b)) / length(union(a, b))
```

D1_C01 两套各有 3 个邻居，交集 1 个、并集 5 个，Jaccard 为 0.20。整合后它有 2 个邻居来自别的供者。要解释状态是否仍在，继续对照同条码在原始 RNA 层的表达与供者分布。

应能复现的三个观察

- 目标条码 6 个，其中 5 个在两个空间匹配。
- D2_C15 在 after 中缺失。
- D1_C01 的近邻交并比为 0.20。

脚本在两个表示空间分别按欧氏距离精确查近邻，排除查询细胞自身。这里比较的是重建的近邻名单，Seurat 的 SNN 图还包含共享近邻权重；取距离相同的边界邻居时具体名单也可能改变。

换成自己的数据

Seurat 对象可用 Embeddings(obj, reduction = "pca") 取出未整合坐标，再用对象中实际的整合降维名称提取另一套坐标。把行名保存到 cell_id 列。覆盖所有候选邻居，在 cells.tsv 单独标记目标细胞，然后增加合理的维度和 k。

先对齐输入字段

cells.tsv	cell_id 为带文库前缀的唯一细胞键。donor 为供者，target 用 1/0 标记整合前的目标群。
before.tsv / after.tsv	首列 cell_id，其余列全部是数值维度。前表填未整合 PCA，后表填整合表示，每行对应一个细胞。
settings.tsv	k 是每个目标细胞保留的邻居数，须小于两个空间共有的细胞数。

以下是 Seurat v5 对象的导出示例。obj 应是读者已运行分析的对象，示例的 integrated.rpca 请替换为 Reductions(obj) 中对应本次分析的名称。分开保存的两个对象也可分别提取坐标，只要条码键统一。

```
source("prepare.R", encoding="UTF-8")
write_embedding(SeuratObject::Embeddings(obj, "pca"), "before.tsv")
write_embedding(SeuratObject::Embeddings(obj, "integrated.rpca"), "after.tsv")
```

将输入放入 my-data，补齐其余示例字段和设置，再运行 Rscript run.R my-data my-results。

输出文件怎样读

retention.tsv	核对元数据中的每个条码是否出现在两个空间。
neighbour_metrics.tsv / neighbour_list.tsv	每个目标细胞的同群、跨供者比例、近邻交并比，以及可手查的邻居姓名单。
target-neighbours.pdf	整合前后的同群近邻比例配对图。

方法资料 / Seurat 整合教程

近邻数换成 2、3、5，同一群细胞还聚在一起吗?

整合前后各取三个近邻，可能恰好让某个边界条码显得变化很大。固定共同细胞和两套坐标，逐个更改近邻数，看看目标同伴的比例以及具体同伴的身份各自怎样变化。

```
Rscript advanced.R example results 2,3,5
```

末尾给出想比较的 k，多个值用英文逗号隔开。所有 k 必须小于共同细胞数。算法先在两个空间各取足够长的精确近邻列表，随后截取各个 k，供者汇总使用目标条码的中位数。

D1_C01 • 共同细胞上的精确近邻

k	保留邻居	Jaccard	整合后目标占比
2	1	0.333	1.00
3	1	0.200	1.00
5	3	0.429	0.80

对模拟条码 D1_C01，k 为 2、3、5 时，与整合前相同的邻居分别有 1、1、3 个。以 k 为 3 为例，前后各 3 人，交集 1、并集 5，Jaccard 为 0.20；整合后的 3 个邻居却全属于原目标群。同伴的具体身份换了，细胞状态仍可能相近。

从结果表复算

```
x <- read.delim("results/k-by-cell.tsv")
subset(x, cell_id == "D1_C01",
       c(k, retained_neighbours, jaccard, target_before,
         target_after))
```

retained_neighbours 是同一条码前后近邻的交集大小。target_after 是整合后近邻中属于原目标群的比例。k-by-donor.tsv 另外列出每位供者贡献的目标条码个数。

打开 k-by-donor.tsv，D1 在 k 为 3 时汇总了 3 个目标条码，D2 只有 2 个。先给每一位供者画范围，再决定是否值得追踪同一标记群。两个坐标空间若存在相同距离的并列邻居，边界名次还可能随并列处理改变。