环境 RNA 与 HBB

从示例数据跑到第一张结果表

准备同一文库的三组液滴，逐条码计算 HBB 占比，再比较共同条码的检出率。教学数据特意留了一条只有原始计数的 T 细胞，方便观察分母为何改变。



准备 R 环境

仅使用 R 自带的软件包。 示例已在 R 4.3.3 运行。

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

```
d <- read.delim("example/droplets.tsv")
t <- subset(d, library_id == "L1" & group == "T")
mean(t$raw_hbb > 0)
matched <- subset(t, !is.na(adjusted_hbb))
c(before = mean(matched$raw_hbb > 0),
  after  = mean(matched$adjusted_hbb > 0))
```

在 L1 的 8 个候选 T 细胞中，原始 HBB 检出 5 个。若只保留在校正矩阵也找得到的 7 个条码，原始检出是 4 个，校正后为 1 个。比较前后用的是同样的 7 个条码。运行下一段，把阳性参照也摆在旁边。

用第二项结果核对

```
s <- read.delim("results/library_group.tsv")
s[s$library_id == "L1",
  c("group", "n_raw", "n_matched",
    "matched_raw_detection", "matched_adjusted_detection")]
```

示例中 L1 的 5 个红系细胞在前后都能检出 HBB，空液滴只有原始计数。脚本汇总已有的计数，未在这些教学数据上运行去污染模型。对真实文库，先从 SoupX 或 CellBender 的输出按条码提取校正值。

应能复现的三个观察

- L1 / T 的 n_raw 为 8，n_matched 为 7。
- L1 / T 配对检出率从 4/7 变为 1/7。
- L1 的红系参照配对检出率前后均为 1。

空液滴反映背景组分，血红蛋白占比不能直接换算为每个 T 细胞的污染率。校正后很小的正数也会被本练习计入“>0”，换模型或阈值时请在结果注明口径。

换成自己的数据

以同一捕获批次的未过滤计数、细胞注释和校正结果为起点。用 HBB 所在的基因行提取分子数，用该表达层全部基因的列和作总数。将校正矩阵按文库与条码左连接到原始表。先在同一批次里确定背景和红系参照，再按相同字段追加其他文库。

先对齐输入字段

droplets.tsv	一行一个文库内条码。library_id 与 barcode 联合作键，group 取 empty、erythroid 或 T。
raw_total / raw_hbb	原始总 UMI 和 HBB 的原始 UMI。总数以同一表达层的全部基因为分母，保留 HBB 为零的条码。
adjusted_total / adjusted_hbb	校正表达层中同一条码的总量与 HBB。两列一起填值或一起留空。脚本允许非整数校正值。

运行前准备 raw_counts 与 corrected_counts（基因 × 条码），以及有 barcode、group 两列的 groups 标签表。只核查原始数据时可将 corrected_counts 设为 NULL。合并多份文库后再写出 droplets.tsv。

```
source("prepare.R", encoding="UTF-8")
out <- make_hbb_table(raw_counts, corrected_counts, groups, "L1")
write.table(out, "droplets.tsv", sep="\t", quote=FALSE, row.names=FALSE)
```

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

droplet_metrics.tsv	逐条码的原始与校正占比。
library_group.tsv	每个文库、每组的原始与配对分母，以及相应检出率。
hbb-by-library.pdf	按文库分页的原始 HBB 占比箱线图。

方法资料 / SoupX 作者教程

把 HBB 检出门槛从 1 个 UMI 提到 2 个，会发生什么?

单个 HBB 分子就算阳性，和至少两个分子才算阳性，你对 T 细胞的判断会差多少？在两种计数层都找得到的条码上改阈值，才能把检出门槛和细胞名单变化分开。

```
Rscript advanced.R example results 1,2,3
```

最后一个参数是以英文逗号隔开的正整数。每个阈值分别生成一行，空液滴也保留原始覆盖数。校正值若有小数，直接用填写的阈值与原值比较。

L1 • T 细胞 • 相同的 7 个条码

检出门槛	校正前	校正后	失去检出
1 UMI	4 / 7	1 / 7	3 个
2 UMI	2 / 7	1 / 7	1 个
3 UMI	1 / 7	0 / 7	1 个

模拟文库 L1 的 T 细胞有 7 个共同条码。门槛为 1 时，校正前后各为 4/7 和 1/7，有 3 个条码从检出变为未检出。门槛为 2 时，各为 2/7 和 1/7，只差 1 个条码。两次计算的分母都是 7，八个原始条码的覆盖数另外记录。

从结果表复算

```
x <- read.delim("results/threshold-sensitivity.tsv")
subset(x, library_id == "L1" & group == "T",
       c(min_umi, n_paired, paired_before, paired_after,
         lost, gained, paired_delta))
```

n_paired 是前后都可比较的条码数。lost 和 gained 分别是向下和向上越过阈值的条码数。paired_delta 是前后检出比例之差，按校正后减校正前计算。

再单独选择红系细胞，在这三档门槛下，共同条码的前后检出都应为 5/5。若自己的数据中阳性参照也同步大幅减少，就要回查校正强度。在没有校正值的空液滴中，配对差记作 NA。