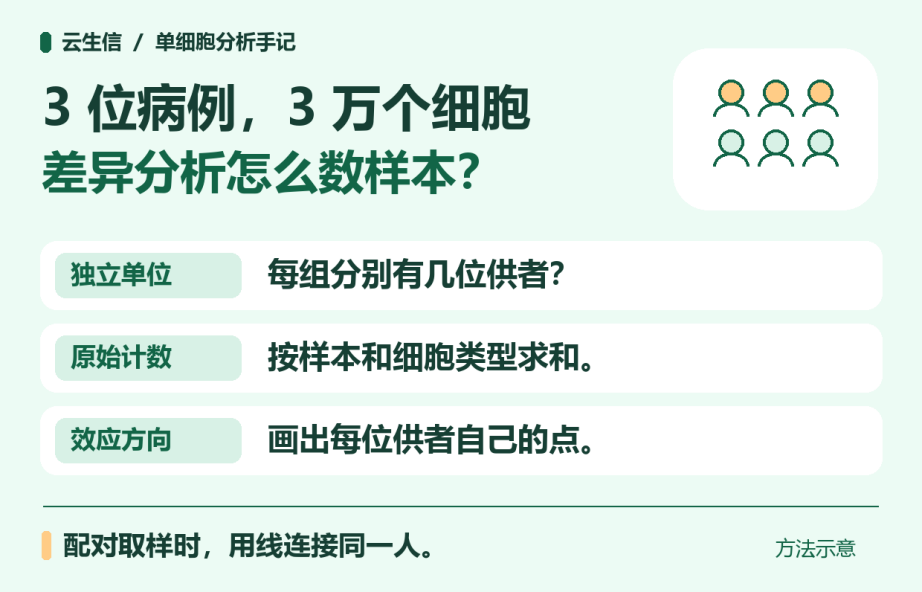
按供者汇总与差异检验

从示例数据跑到第一张结果表

运行从稀疏原始计数、细胞标签、样本覆盖到 edgeR 准似然检验的完整流程。文章开头用三位病例解释独立样本量。下载练习另用 4 位虚构供者的前后配对，便于亲手检查设计矩阵。



准备 R 环境

需要 Matrix 和 Bioconductor 的 edgeR。实测版本分别为 1.6-5 和 3.38.4。 示例已在 R 4.3.3 运行。

```
if (!requireNamespace("BiocManager", quietly=TRUE))
    install.packages("BiocManager", repos="https://cloud.r-project.org")
BiocManager::install(c("Matrix", "edgeR"), ask=FALSE, update=FALSE)
```

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

```
cov <- read.delim("results/sample_coverage.tsv")
with(subset(cov, use), table(donor, condition))
design <- read.delim("results/design.tsv", check.names = FALSE)
design[, c("sample_id", "conditionafter")]
```

模拟数据有 192 个细胞，每份样本 18 个单核细胞、6 个 T 细胞。筛单核细胞后 144 个细胞汇入 8 列，每位供者各有一前一后。这里可估计 4 对供者的条件差异。

用第二项结果核对

```
de <- read.delim("results/differential.tsv")
subset(de, gene_id == "ISG15", c(gene_id, logFC, FDR))
patients <- read.delim("results/donor_expression.tsv")
patients[order(patients$donor, patients$condition),
         c("donor", "condition", "n_cells", "logCPM")]
```

在给定的模拟计数中，ISG15 被设定为处理后上调，实跑的 log□ 倍数变化约为 +1.31。配对图可查看各人方向。若要分析独立病例和对照，在设置表改成 independent，并为两组提供不同供者；有重复供者或配对缺失时脚本会指出设计问题。

应能复现的三个观察

- 单核细胞 144 个，共 8 份样本和 4 对供者。
- 矩阵 120 个基因，汇总前后逐基因计数守恒。
- ISG15 的 logFC 约 +1.31，方向为 after 相对 before。

示例的最少细胞数只是练习设置。重复很少时估计会很不稳定。此脚本演示两个条件的简单独立或配对比较；多批次协变量、交互或复杂重复采样需要重新建模。

换成自己的数据

从原始 RNA counts 层导出稀疏矩阵，按实际行列顺序写出基因、条码清单。每个 cell_id 需在各文库间唯一。先把目标细胞类型和样本设计填入设置，再检查覆盖表；处于两组比较之外的样本不进入本次分析。

先对齐输入字段

counts.mtx / genes.tsv / barcodes.tsv	基因 × 细胞的非负整数 UMI 稀疏矩阵。后两表首列 gene_id 和 cell_id 的顺序分别对应矩阵的行与列。
cells.tsv	每个 cell_id 对应 sample_id、donor、condition 和 cell_type。供者可重复出现于不同条件，样本编号须唯一指向供者与条件。
settings.tsv	指定 cell_type、设计 paired 或 independent、reference、comparison、min_cells 与候选 plot_gene。

raw_counts 是带有唯一基因、细胞行列名的原始计数矩阵，cell_metadata 包含输入表中的五个字段。导出后根据研究设计修改 my-data/settings.tsv。

```
source("prepare.R", encoding="UTF-8")
write_count_inputs(raw_counts, cell_metadata, "my-data")
file.copy("example/settings.tsv", "my-data/settings.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

sample_coverage.tsv / pseudobulk_counts.tsv	逐样本细胞覆盖与逐基因聚合后的原始计数，可手算核对。
design.tsv / differential.tsv	实际设计矩阵和 edgeR 的效应量、原始 P 值与 FDR。
donor_expression.tsv / donor-expression.pdf	目标基因的标准化样本值和供者散点图。

方法资料 / edgeR 官方手册

把 4 位供者依次拿掉一位，ISG15 的方向还稳吗?

完整队列得到一个很小的 FDR 时，可以问得更具体：把任意一位供者的一对样本同时拿掉，估计的处理效应是否改变方向？每次从原始 pseudobulk 计数重做表达过滤、归一化、离散度估计与准似然检验。

```
Rscript advanced.R example results ISG15
```

末尾参数替换为目标基因 ID。该练习用于成对设计，至少要有三位完整供者，才能保证每次留一后仍有两对样本。每次重算都记录被保留的基因数，目标基因若被过滤则保留 filtered 状态。

ISG15 · 成对条件 · 模拟计数

去掉供者	配对数	log□ 倍数变化	FDR
保留全部	4	+1.313	0.000072
D1	3	+1.162	0.0726
D3	3	+1.094	0.000029

模拟队列全样本拟合得到 log□ 倍数变化 +1.313。留一后，ISG15 的四个估计约为 +1.162、+1.405、+1.094、+1.492。拿掉 D1 时 FDR 约为 0.073，效应仍为正向。读表时应分别报告方向是否一致、效应变动多少，以及统计精度如何变化。

从结果表复算

```
x <- read.delim("results/leave-one-donor.tsv")
x[, c("omitted", "pairs", "status", "logFC", "FDR")]
range(x$logFC[x$omitted != "(none)"], na.rm = TRUE)
```

omitted 表示本次去掉的整位供者，(none) 对应全队列。pairs 是留下的完整配对数，logFC 是 comparison 相对 reference 的方向，FDR 来自当次拟合通过过滤的基因集合。

将基因改成 MX1，再看哪一位供者对该基因影响较大。留一拟合共享大部分样本，四个结果互相依赖。不要挑其中最小的 FDR 当最后结果，也不能将四次拟合解释为来自四个独立队列的验证。