BIOF3 组学数据分析

数据与环境准备

导出日期：2026年5月12日

数据与环境准备

BioF3 的每一章都配有可直接运行的 R 脚本。这一篇统一说明：每个脚本要什么数据、数据放在哪、怎么一次性装好所需的 R 包。

数据目录约定

所有脚本默认去**同一个数据根目录**里找数据，避免每个项目再重新下载：

```
~/biof3-data/
├── pbmc3k/              # 单细胞 module01~07 用
├── pbmc5k-citeseq/      # 单细胞 module07（CITE-seq 真实示例）用
└── pbmc10k-scatac/      # 单细胞 module10（scATAC-seq 真实示例）用
```

脚本里读这个路径的方式：

```
data_root <- Sys.getenv("BIOF3_DATA_DIR",
                        file.path(path.expand("~"), "biof3-data"))
pbmc_dir  <- file.path(data_root, "pbmc3k")
```

怎么换位置：

```
# 临时：当次运行覆盖
BIOF3_DATA_DIR=/mnt/shared/biof3-data Rscript scripts/single-cell/sc03_qc_cluster_sci.R

# 长期：写进 shell 配置
echo 'export BIOF3_DATA_DIR=/mnt/shared/biof3-data' >> ~/.zshrc
```

数据清单

需要下载的数据

这三份数据由 10x Genomics 提供，前两份可以在笔记本上跑，scATAC 数据大一点、磁盘占用需要留意：

数据集	用在哪	体积	说明
PBMC 3k	单细胞 01 ~ 07、06 配套脚本、07 配套脚本	~35 MB	10x Genomics 第一代示例，贯穿 scRNA-seq 主线
PBMC 5k CITE-seq	单细胞 07 CITE-seq 真实示例	~75 MB	RNA + 32 个抗体的 TotalSeq-B panel
PBMC 10k scATAC v2	单细胞 10 scATAC-seq 真实示例	~200 MB	peak matrix + singlecell.csv（不含 fragments）

脚本会在首次运行时自动下载，之后缓存。如果想先一次性把三份都下好、之后离线跑，用下面的一键准备脚本。

用 Bioconductor 内置数据的模块

下面这些模块**不需要手工下载任何数据**，装好对应的 Bioconductor 包就能跑：

模块	用的数据包
单细胞 08 TCR/BCR	`scRepertoire`（自带 `contig_list` 和 `scRep_example`）
单细胞 09 空间转录组	`SeuratData::stxBrain`（首次 `InstallData` 一次，本地缓存）
bulk 02 DESeq2 / 03 富集 / 04 可视化 / 07 多工具对比	`airway`
bulk 05 LRT 时间序列	`fission`
bulk 06 批次效应	`bladderbatch` + `airway`

一键准备脚本

配套脚本 `biof3_prepare_data.R` 一次性完成：

1. 检查 R 版本（推荐 R ≥ 4.3）
2. 从 CRAN + Bioconductor 装所有脚本用到的 R 包
3. 下载 PBMC 3k / 5k / 10k 三份数据到 `~/biof3-data/`
4. 加载 SeuratData 的 `stxBrain`（空间转录组）

```
Rscript scripts/biof3_prepare_data.R
```

默认会把数据放在 `~/biof3-data/`。换目录：

```
BIOF3_DATA_DIR=/mnt/shared/biof3-data Rscript scripts/biof3_prepare_data.R
```

脚本里每一步都是幂等的：装过的包、下过的数据不会重复下载。

biof3_prepare_data.R
12 KB

下载一键准备脚本 ↗

R 包清单

按脚本粗分：

单细胞主线（module01~07）：`Seurat`、`SeuratObject`、`Matrix`、`ggplot2`、`dplyr`、`patchwork`、`RColorBrewer`

单细胞专题：

- module05 轨迹：`slingshot`、`RColorBrewer`、`viridis`
- module06 通讯：`CellChat`（GitHub `jinworks/CellChat`）、`patchwork`
- module07 CITE-seq：`Seurat` v5 及以上（用到 `CreateAssay5Object`）
- module08 TCR/BCR：`scRepertoire`
- module09 空间：`SeuratData`、`stxBrain.SeuratData`

- module10 scATAC：`Signac` 、 `EnsDb.Hsapiens.v86` 、 `GenomicRanges`

bulk RNA-seq：`DESeq2` 、 `edgeR` 、 `limma` 、 `airway` 、 `fission` 、 `bladderbatch` 、 `sva` 、 `clusterProfiler` 、 `org.Hs.eg.db` 、 `enrichplot` 、 `DOSE` 、 `ggrepel` 、 `pheatmap` 、 `EnhancedVolcano` 、 `apeglm`

`biof3_prepare_data.R` 会把这些全部装好。如果只跑某几章，按需装对应包即可。

国内网络的注意事项

几个国内环境下常见的坑，解决办法：

1. Bioconductor 主站直连有时慢

`BiocManager::install()` 默认用 Bioconductor 官方源。如果下载很慢，临时手动指定镜像：

```
options(
  repos = c(
    BioCsoft = "https://bioconductor.org/packages/3.22/bioc",
    BioCann  = "https://bioconductor.org/packages/3.22/data/annotation",
    BioCexp  = "https://bioconductor.org/packages/3.22/data/experiment",
    CRAN     = "https://mirrors.tuna.tsinghua.edu.cn/CRAN"
  ),
  timeout = 600
)
install.packages("DESeq2")
```

上面这段写死了 BioC 版本（3.22 对应 R 4.5），**也是 `biof3_prepare_data.R` 内置的 fallback**。装完之后回到默认 `BiocManager` 继续用。

2. CellChat 不在 CRAN / BioC

GitHub 直接装：

```
# CRAN 上装依赖
install.packages(c("sna", "ggnetwork", "collapse"))

# GitHub 装 CellChat
if (!requireNamespace("devtools", quietly = TRUE)) install.packages("devtools")
devtools::install_github("jinworks/CellChat")
```

如果 `devtools::install_github` 连不上，从 `https://codeload.github.com/jinworks/CellChat/tar.gz/HEAD` 下 tarball 手动 `R CMD INSTALL` 也可以。

3. SeuratData 和 msigdbr 的数据服务器在海外

- `SeuratData::InstallData("stxBrain")` 约 136 MB，从 AWS 拉，可能慢
- `msigdbr` 26.x 起从 Zenodo 拉数据，国内直连常超时。如果 bulk03 需要 MSigDB，临时换成 `clusterProfiler` 自带的 GO/KEGG 富集也完全够用

目录结构总览

把上面这些合起来，一个完整的学习环境长这样：

```
~/biof3-data/                    # 数据（脚本共用）
├── pbmc3k/
├── pbmc5k-citeseq/
└── pbmc10k-scatac/

~/BioF3/                         # 代码（git clone 或者直接下脚本）
├── scripts/
│   ├── biof3_prepare_data.R
│   ├── single-cell/
│   │   ├── sc01_data_sci.R
│   │   └── ...
│   ├── bulk-rnaseq/
│   │   ├── bulk02_deseq2_sci.R
│   │   └── ...
└── static/scripts/              # 和 scripts/ 内容一致，供下载使用

~/R/x86_64-apple-darwin20/4.3/   # R 包库（自动管理）
├── Seurat/
├── DESeq2/
└── ...
```

数据根目录和代码根目录完全分开，数据集换机器 / 换项目都能共享。

常见问题

Q：脚本一直在下载同一份数据

A：检查 `~/biof3-data/` 目录是否存在。脚本会用 `file.exists()` 检测是否已下载；如果你换了路径、删了缓存，它就会重新下。

Q：磁盘不够，能不能把数据放在外置硬盘

A：可以，把外置硬盘路径写进 `BIOF3_DATA_DIR` 即可，脚本会跟着去找。

Q：装一个包失败，整个脚本断了

A： `biof3_prepare_data.R` 每装一个包都用 `tryCatch` 包着，失败会打印 warning 但不会中断。装完最后会报告哪些包失败。

参考资源

- 10x Genomics 数据入口
- Bioconductor 镜像列表
- 清华 TUNA CRAN 镜像
- SeuratData 包