BIOF3 组学数据分析

01 实践数据集与数据获取

导出日期：2026年5月12日

01 实践数据集与数据获取

BioF3 的教程需要配套真实数据，不能只依赖手造数据和示意代码。本页整理适合教学复现的公开测试数据集，并给出建议使用场景、下载命令和数据规模。

所有数据都来自公开资源，建议下载到项目外的本地数据目录，例如 `~/biof3-data`，不要直接提交到网站仓库。

```
mkdir -p ~/biof3-data
cd ~/biof3-data
```

推荐数据集

数据集	适用章节	类型	大小	用途
PBMC 3k	02-06	scRNA-seq	约 7.4 MB	入门、质控、聚类、注释、差异分析
5k PBMC CITE-seq	07	RNA + ADT	约 37 MB	多模态分析、WNN、蛋白标志物
PBMC scATAC 10k	10	scATAC-seq	约 162 MB	染色质可及性、LSI、peak matrix
Visium Breast Cancer	09	空间转录组	约 74 MB	空间表达、组织切片可视化

PBMC 3k

PBMC 3k 是 10x Genomics 公开的外周血单个核细胞数据，也是 Seurat 和 Scanpy 入门教程常用数据。它体积小、下载快，适合本教程前半部分的大多数练习。

适用章节：

- 02 原始数据处理与 Cell Ranger
- 03 质量控制、聚类与细胞类型注释
- 05 轨迹推断与拟时序分析
- 06 细胞-细胞通讯分析

底部下载资源区会直接从 10x Genomics 原始地址下载。也可以使用命令行下载：

```
mkdir -p ~/biof3-data/pbmc3k
cd ~/biof3-data/pbmc3k

curl -L -O https://cf.10xgenomics.com/samples/cell-exp/1.1.0/pbmc3k/pbmc3k_filtered_gene_bc_matrices
tar -xzf pbmc3k_filtered_gene_bc_matrices.tar.gz
```

Seurat 读取示例：

```
library(Seurat)

data_dir <- "~/biof3-data/pbmc3k/filtered_gene_bc_matrices/hg19"
counts <- Read10X(data.dir = data_dir)
pbmc <- CreateSeuratObject(counts = counts, project = "PBMC3K")
pbmc
```

PBMC 3k 真实图表脚本

本节配套脚本会自动下载 PBMC 3k 数据，读取 10x Genomics 的真实 `matrix.mtx` 、 `genes.tsv` 和 `barcodes.tsv` ，并生成本教程使用的图表。

运行方式：

```
Rscript scripts/single-cell/sc01_data_sci.R
```

默认数据目录：

```
~/biof3-data/pbmc3k
```

默认图片输出目录：

```
static/img/tutorial/single-cell/module01/
```

脚本生成的图表分为两类：

图片	来源	说明
01-gene-expression-bar.png	PBMC 3k 真实矩阵	总 UMI counts 最高的基因
02-cell-counts-distribution.png	PBMC 3k 真实矩阵	每个细胞的 total UMI counts 分布
03-gene-mean-distribution.png	PBMC 3k 真实矩阵	每个基因的平均表达分布
04-expression-matrix-heatmap.png	PBMC 3k 真实矩阵	真实表达矩阵子集热图
05-qc-scatter.png	PBMC 3k 真实矩阵	total counts、detected genes 和 mitochondrial percentage
06-database-comparison.png	概念图	公共数据源使用场景对比
07-workflow.png	概念图	公开数据可追溯分析流程
08-qc-combined.png	PBMC 3k 真实矩阵	PBMC 3k 组合 QC 指标

提示

后续章节的图表也会逐步按这个标准整理：优先使用真实公开数据；如果是概念图，会明确标注为概念图，不和真实分析结果混在一起。

5k PBMC CITE-seq

5k PBMC CITE-seq 数据同时包含 RNA 表达矩阵和抗体衍生标签（ADT）矩阵，适合讲解多模态单细胞分析。

适用章节：

- 07 多模态单细胞分析

底部下载资源区会直接从 10x Genomics 原始地址下载。也可以使用命令行下载：

```
mkdir -p ~/biof3-data/pbmc5k-citeseq
cd ~/biof3-data/pbmc5k-citeseq

curl -L -O https://cf.10xgenomics.com/samples/cell-exp/3.1.0/5k_pbmc_protein_v3_nextgem/5k_pbmc_prot
tar -xzf 5k_pbmc_protein_v3_nextgem_filtered_feature_bc_matrix.tar.gz
```

Seurat 读取示例：

```
library(Seurat)

data_dir <- "~/biof3-data/pbmc5k-citeseq/filtered_feature_bc_matrix"
counts <- Read10X(data.dir = data_dir)

pbmc <- CreateSeuratObject(counts = counts$`Gene Expression`, project = "PBMC5K_CITE")
pbmc[["ADT"]] <- CreateAssayObject(counts = counts$`Antibody Capture`)
pbmc
```

Visium Breast Cancer

Visium Breast Cancer 是 10x Genomics 的空间转录组数据，适合练习组织切片上的空间表达可视化和空间邻域分析。

适用章节：

- 09 空间转录组学

该数据体积较大，当前不放入网站仓库。底部下载资源区会直接从 10x Genomics 原始地址下载。也可以使用命令行下载：

```
mkdir -p ~/biof3-data/visium-breast-cancer
cd ~/biof3-data/visium-breast-cancer

curl -L -O https://cf.10xgenomics.com/samples/spatial-exp/1.1.0/V1_Breast_Cancer_Block_A_Section_1/V
tar -xzf V1_Breast_Cancer_Block_A_Section_1_filtered_feature_bc_matrix.tar.gz
```

说明

完整空间分析通常还需要组织切片图片和 `spatial/` 坐标文件。后续补强09 时，会把矩阵、图片、坐标和 Seurat/Scanpy 读取流程整理成完整示例。

PBMC scATAC 10k

PBMC scATAC 10k 是 10x Genomics 的单细胞 ATAC-seq 数据，适合练习 peak matrix、TF-IDF、SVD/LSI 和染色质可及性分析。

适用章节：

- 10 scATAC-seq

该数据体积较大，当前不放入网站仓库。底部下载资源区会直接从 10x Genomics 原始地址下载。也可以使用命令行下载：

```
mkdir -p ~/biof3-data/pbmc10k-scatac
cd ~/biof3-data/pbmc10k-scatac

curl -L -O https://cf.10xgenomics.com/samples/cell-atac/2.1.0/10k_pbmc_ATACv2_nextgem_Chromium_Contr
```

使用建议

1. 初学者先下载 PBMC 3k，优先完成01-03。
2. 做多模态分析时再下载 5k PBMC CITE-seq。
3. 空间转录组和 scATAC 数据体积更大，建议在理解标准 scRNA-seq 流程后再使用。
4. 网站仓库只保存教程、脚本和小型示例，不保存大型原始数据。

下载资源

module01_complete_sci.R
14 KB
下载图表生成脚本 ↗

pbmc3k_filtered_gene_bc_matrices.tar.gz
7.3 MB
下载 PBMC 3k 数据 ↗

5k_pbmc_protein_v3_nextgem_filtered_feature_bc_matrix.tar.gz
37 MB
下载 5k PBMC CITE-seq 数据 ↗

V1_Breast_Cancer_Block_A_Section_1_filtered_feature_bc_matrix.tar.gz
74 MB
下载 Visium Breast Cancer 表达矩阵 ↗

10k_pbmc_ATACv2_nextgem_Chromium_Controller_filtered_peak_bc_matrix.h5
162 MB
下载 PBMC scATAC 10k 数据 ↗

数据来源

- 10x Genomics Datasets
- Seurat PBMC 3k Guided Tutorial
- Scanpy PBMC 3k Preprocessing Tutorial

扫码关注微信公众号【生信F3】

获取文章完整内容，分享生物信息学最新知识。