BIOF3 组学数据分析

公共数据库与数据检索

导出日期：2026年5月11日

公共数据库与数据检索

组学分析的第二个问题是：从哪里找到可信的数据，并确认它是否适合自己的问题？

本章介绍常用公共数据库 GEO、SRA、CELLxGENE、Human Cell Atlas 和 Expression Atlas 分别适合什么数据，以及如何根据 accession 编号追踪数据来源。

AI 工具如何辅助数据检索，详见 AI 辅助编程与智能体工具。

学习目标

完成本章后，你应该能够：

- 知道不同数据库分别适合解决什么问题
- 看懂常见 accession 编号的层级
- 用合适的关键词检索公共数据
- 判断一个数据集是否适合自己的研究问题

公共数据库怎么选

不同数据库解决不同问题。先判断你需要的是原始测序数据、处理后的表达矩阵，还是可交互浏览的注释数据。

需求	优先看哪里
找论文配套表达矩阵	GEO
下载原始 FASTQ	SRA / ENA
浏览单细胞注释数据	CELLxGENE
找人类细胞图谱项目数据	HCA Data Portal
查基因在组织/细胞中的表达	Expression Atlas

常见 accession 编号

公开数据通常通过 accession 编号追踪。

编号	常见含义	示例
GSE	GEO Series，一个研究或数据集	GSE12345
GSM	GEO Sample，一个样本	GSM123456
SRP	SRA Study	SRP123456
SRS	SRA Sample	SRS123456
SRX	SRA Experiment	SRX123456
SRR	SRA Run，常用于下载 reads	SRR1234567

分析前要确认编号层级。很多新手拿到 `GSE` 后直接找 FASTQ，会发现真正下载 reads 需要进一步找到对应的 `SRR` 。

主要数据库

GEO

GEO 是 NCBI 维护的功能基因组学数据库，常见于论文数据提交。它可以包含表达矩阵、样本信息、平台信息和补充文件。

网址：https://www.ncbi.nlm.nih.gov/geo/

适合：

- 查找论文配套数据
- 下载处理后的表达矩阵
- 查看样本分组和元数据
- 追踪到 SRA 原始数据

检索建议：

```
关键词 + 物种 + 技术 + 组织/疾病
```

例如：

```
single cell RNA-seq human liver fibrosis
```

SRA

SRA 是 NCBI 的原始测序数据归档库。需要 FASTQ 时通常会用到它。

网址：https://www.ncbi.nlm.nih.gov/sra/

下载常用 SRA Toolkit：

```
conda install -c bioconda sra-tools

prefetch SRR1234567
fasterq-dump SRR1234567 --split-files -O data/raw/
```

注意：

- FASTQ 文件可能很大
- 下载前确认磁盘空间

- 批量下载前先测试一个 run
- 记录 SRR 列表和下载日期

CELLxGENE

CELLxGENE Discover 提供许多可浏览的单细胞数据集，通常可以在线查看 UMAP、细胞类型和基因表达。

网址：https://cellxgene.cziscience.com/

适合：

- 快速浏览单细胞数据
- 查看细胞类型注释
- 寻找可下载的 h5ad 数据
- 比较公开数据中的基因表达模式

使用建议：

- 先在线检查数据是否符合研究问题
- 下载前确认样本、组织、物种和处理流程
- 注意数据是否已经标准化或整合

Human Cell Atlas

Human Cell Atlas 关注人类细胞参考图谱。HCA Data Portal 提供社区生成的多组学开放数据。

网址：https://data.humancellatlas.org/

适合：

- 查找人类组织和器官图谱
- 获取大型参考数据
- 了解细胞类型注释和 atlas 项目
- 作为数据整合和注释参考

Expression Atlas / Single Cell Expression Atlas

EMBL-EBI 的 Expression Atlas 和 Single Cell Expression Atlas 提供基因表达查询和单细胞表达浏览。

网址：https://www.ebi.ac.uk/gxa/

适合：

- 查询某个基因在不同组织或细胞中的表达
- 浏览经过整理的表达数据
- 做初步假设生成
- 辅助解释 marker gene

下一步

继续学习：

- R 数据整理与 ggplot2 可视化
- 单细胞实践 01：实践数据集与数据获取

参考资源

- GEO：https://www.ncbi.nlm.nih.gov/geo/
- SRA：https://www.ncbi.nlm.nih.gov/sra/

- CELLxGENE：https://cellxgene.cziscience.com/
- Human Cell Atlas Data Portal：https://data.humancellatlas.org/
- Expression Atlas：https://www.ebi.ac.uk/gxa/

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。