

表观组学实践教程

导出日期：2026年6月27日

表观组学实践教程

表观组学关注的不是基因本身的序列，而是基因表达能不能被"打开"：染色质开放程度、转录因子结合位置、组蛋白修饰、DNA 甲基化。同一个基因组在不同细胞里呈现不同的表观图谱，这些图谱决定了细胞状态和对环境的反应。

本专栏会围绕最常用的三类实验展开：ATAC-seq（染色质开放区域）、ChIP-seq（蛋白-DNA 结合）、WGBS/RRBS（DNA 甲基化）。思路和 bulk RNA-seq 类似：先搞清楚每步产物是什么，再决定用什么工具把它们跑出来。

表观组要回答的核心问题

转录组告诉你"这个基因表达了多少"，表观组告诉你"为什么它能表达 / 不能表达"：

问题	转录组	表观组
哪些基因被打开	看 mRNA 量	看启动子开放 + TF 结合 + 组蛋白活化标记
这个 TF 在哪里发挥作用	找下游差异基因（间接）	ChIP-seq 直接看 TF 结合位点
这个增强子调控哪个基因	看不到	ATAC + Hi-C / 4C 配套
这个表型为什么稳定遗传	看不到	甲基化（可遗传）
治疗为什么没起效	看不到响应基因变化	看染色质是否压缩、TF 还能否结合

表观组适合的项目：发育、分化、肿瘤进展（耐药机制）、神经精神疾病、细胞身份建立。**不适合：**单纯比较两个稳态条件的差异（直接做转录组就行，表观组只是补充证据）。

一个项目大致是什么样子

ATAC-seq 和 ChIP-seq 的分析主线非常像。从一批样本的 FASTQ 出发，到"差异开放区域"或"差异结合位点"表，大致要经过：

步骤	典型产物	常用工具
接头剪切与质控	清洗后的 FASTQ	fastp、FastQC、MultiQC
比对到参考基因组	sorted.bam	Bowtie2、BWA-MEM
过滤线粒体 / 重复	清洁 BAM	samtools、Picard
peak calling	narrowPeak / broadPeak	MACS2、MACS3
peak 注释	基因附近 peak 映射表	ChIPseeker、HOMER
motif 分析	显著富集的 motif	HOMER、MEME Suite
差异分析	差异 peak 列表	DiffBind、DESeq2 on peak counts
与表达整合	peak ↔ 基因关联	自定义脚本 + ggplot2

WGBS/RRBS 的流程不同：比对要用 bisulfite-aware 工具（Bismark、BWA-Meth），输出是每个 CpG 的甲基化率，差异分析常用 methylKit 或 DSS。

常见工具栈

下面是 BioF3 例子里会优先使用的组合：

阶段	工具	说明
ATAC/ChIP 比对	Bowtie2、BWA-MEM	两者都行，Bowtie2 对 ATAC 友好
BAM 处理	samtools、Picard	去重、过滤 MAPQ、去线粒体
peak calling	MACS2	ATAC 和 ChIP 都支持，参数略不同
peak 注释	ChIPseeker	R 包，输出表格和图
motif	HOMER、MEME	HOMER 一条命令出 motif 报告
差异 peak	DiffBind、DESeq2	样本数少时 DiffBind 更便捷
可视化	deepTools、IGV、pygenometracks	覆盖度曲线、heatmap、基因浏览器图
甲基化	Bismark、methylKit	标准 WGBS/RRBS 流水线

推荐公开数据集

表观组教学比较依赖公开数据，下面几份是常用参照：

数据集	类型	适合	入口
ENCODE K562 ATAC-seq	ATAC-seq, 细胞系	ATAC 流程练习	ENCODE
ENCODE H3K27ac ChIP-seq	ChIP-seq + input	ChIP 流程、peak 注释	ENCODE
10x Genomics PBMC scATAC 10k	scATAC	单细胞方向的过渡	10x Genomics
TCGA / GDC 甲基化	450K / EPIC 芯片	差异甲基化入门	GDC

ENCODE 的好处是样本类型全、质量稳定、每个实验都有对应的 input 或 control，新手跟着跑最不容易出问题。

最小可跑的例子

下面用 R 的 ChIPseeker 包做一次最简单的 peak 注释，它自带一个 Nature Neuroscience 论文发布的真实 narrowPeak 文件。数据很小，跑完只需要几秒：

```
# 一次性安装依赖（如果还没装）
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
BiocManager::install(c("ChIPseeker", "TxDb.Hsapiens.UCSC.hg19.knownGene"))

library(ChIPseeker)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)

# ChIPseeker 自带的示例 peak 文件
peak_files <- getSampleFiles()
peak_files

# 载入其中一个样本的 peak
peak <- readPeakFile(peak_files[[1]])
peak

# 注释到最近的基因
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
anno <- annotatePeak(peak, TxDb = txdb, tssRegion = c(-3000, 3000))

# 看每个 peak 落在什么类型的区域
head(as.data.frame(anno))

# 基因组区域饼图
plotAnnoPie(anno)
```

plotAnnoPie 会画出 peak 主要落在哪些基因组区域（启动子、内含子、基因间等），这也是 ChIP/ATAC 文章里最常见的一张配图。把这个例子读懂，再去跑真实数据的 peak calling、motif 分析就会顺很多。

专栏模块规划

模块	主题	状态
01	实验类型与数据格式	已上线
02	Peak 注释与多样本比较	已上线
03	DiffBind 差异结合分析	已上线
04	Peak 可视化与多样本比较	已上线
05	Motif 富集与 HOMER	已上线
06	ATAC-seq 分析要点	已上线
07	DNA 甲基化分析入门	已上线
08	表观组与转录组的整合	已上线

目前 01-08 全部上线。01-04 带可跑脚本，05-08 为理论 + 代码示例。

推荐前置知识

- [编程基础：R、Python、Bash 学习路径](#)
- [R 数据整理与 ggplot2 可视化](#)
- [10 scATAC-seq 分析](#)（已有的单细胞版 ATAC 分析）

常见误区

新手做表观组项目最容易的几个判断错误：

误区 1：peak 数量等于实验质量

样本 A 出 50000 个 peak，样本 B 只出 30000 个 — 不一定是 A 测得"更深更好"。**peak 数量**受测序深度、过滤阈值、**call peak** 参数严苛程度影响很大。看 FRiP（reads in peaks 比例）、TSS enrichment 这些标准化指标比 peak 数量更可靠。

误区 2：只做 peak 数量比较，不做 peak 强度分析

A 样本和 B 样本 peak 列表 90% 重叠，但同一个 peak 在 A 里可能 reads 数 100，在 B 里只有 10 — **这才是真差异**。用 DiffBind 做基于 reads 强度的差异分析，光比较 peak 集合 overlap 信息量不足。

误区 3：把 ChIP-seq 当 RNA-seq 思路做差异

ChIP / ATAC 不能用全转录组的 DESeq2 直接套。**peak 是按位置定义的**，不同样本的 **peak 不一致** — 要先建 consensus peak set，再在 consensus 上数 reads，DiffBind 把这套封装好了。

误区 4：把启动子 peak 当成调控 peak

很多 ChIP / ATAC 的 peak 落在启动子，但**真正决定细胞身份的往往是远端增强子**（TSS 几十 kb 之外）。注释 peak 时要区分 promoter / enhancer / gene body，不要全归到 "最近基因" 就完事。

误区 5：忽略 input / IgG 对照

ChIP-seq 没有 input 是没法做正经分析的 — peak caller 没办法区分"真信号"和"基因组开放区域的非特异背景"。**项目设计阶段必须每个 condition 至少配一个 input**。ATAC-seq 不强制要 input，但 IgG 对照能区分非特异 Tn5 切割。

参考资源

- [ENCODE Portal](#)
- [MACS2 文档](#)
- [ChIPseeker Bioconductor 页面](#)
- [deepTools 文档](#)
- [Bismark 手册](#)
- [methylKit 用户指南](#)



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。