BIOF3 组学数据分析

表观组学实践手册

BioF3 表观组学专栏导出版

导出日期：2026年5月12日

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。

BIOF3 组学数据分析

表观组学实践手册目录

BioF3 表观组学专栏导出版

01 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。

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）

参考资源

- ENCODE Portal
- MACS2 文档
- ChIPseeker Bioconductor 页面
- deepTools 文档
- Bismark 手册
- methylKit 用户指南

02 实验类型与数据格式

表观组学实验类型多，但分析思路可以归为两大类：**开放区域检测**（ATAC-seq、DNase-seq）和**蛋白-DNA 结合检测**（ChIP-seq）。两者的分析流程几乎一样，区别在 peak calling 参数和质量指标。

三种主要实验

实验	测什么	peak 类型	典型 QC 指标
ATAC-seq	染色质开放区域	narrow	TSS enrichment、fragment size 分布
ChIP-seq	TF 结合位点 / 组蛋白修饰	narrow (TF) / broad (histone)	FRiP、IDR
WGBS/RRBS	DNA 甲基化	不做 peak calling	覆盖度、转化率

本专栏前 4 个模块聚焦 ATAC-seq 和 ChIP-seq（它们共享 peak-based 分析框架）。甲基化后续单独开。

从 FASTQ 到 peak 文件

不管是 ATAC 还是 ChIP，从原始数据到可分析的 peak 文件，标准流程是：

```
FASTQ → fastp (trim) → Bowtie2/BWA (align) → samtools (sort/filter)
     → Picard (dedup) → MACS2 (peak calling) → narrowPeak / broadPeak
```

BioF3 的表观组教程**从 peak 文件开始**。如果你需要从 FASTQ 跑起，参考 nf-core/chipseq 或 nf-core/atacseq 流水线。

peak 文件格式

MACS2 输出的 `.narrowPeak` 是 BED6+4 格式：

```
chr1  9356548  9356648  peak_1  100  .  5.0  10.5  7.2  50
```

列	含义
1-3	染色体、起始、终止
4	peak 名
5	score
6	strand（通常 `.`）
7	fold enrichment
8	-log10(pvalue)
9	-log10(qvalue)
10	summit 相对于 start 的偏移

R 里用 `ChIPseeker::readPeakFile()` 或 `rtracklayer::import()` 读入，得到 `GRanges` 对象。

下一步

- 02 Peak 注释与多样本比较

参考资源

- MACS2 文档
- nf-core/chipseq
- nf-core/atacseq
- ENCODE 实验标准

03 Peak 注释与多样本比较

拿到 peak 文件之后，第一个问题是"这些 peak 落在基因组的什么位置"。ChIPseeker 把 peak 注释到最近的基因、标注它在启动子 / 内含子 / 基因间等区域，一步出图。

本章用 ChIPseeker 自带的 AR（雄激素受体）ChIP-seq 数据演示：3 个剂量（0M / 1nM / 100nM）的 peak 文件，看剂量增加时 peak 数量、位置分布和关联基因如何变化。

核心流程

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

peak <- readPeakFile("peaks.narrowPeak")
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene

anno <- annotatePeak(peak, TxDb = txdb, tssRegion = c(-3000, 3000))
plotAnnoPie(anno)
plotDistToTSS(anno)
```

`tssRegion = c(-3000, 3000)` 定义"启动子"的范围：TSS 上下游 3kb 以内的 peak 算启动子区域。

真实示例

配套脚本 `epi02_chipseeker_sci.R` 在 ChIPseeker 内置的 AR ChIP-seq 数据上跑完整流程：

```
Rscript scripts/epigenomics/epi02_chipseeker_sci.R
```

每张图看什么

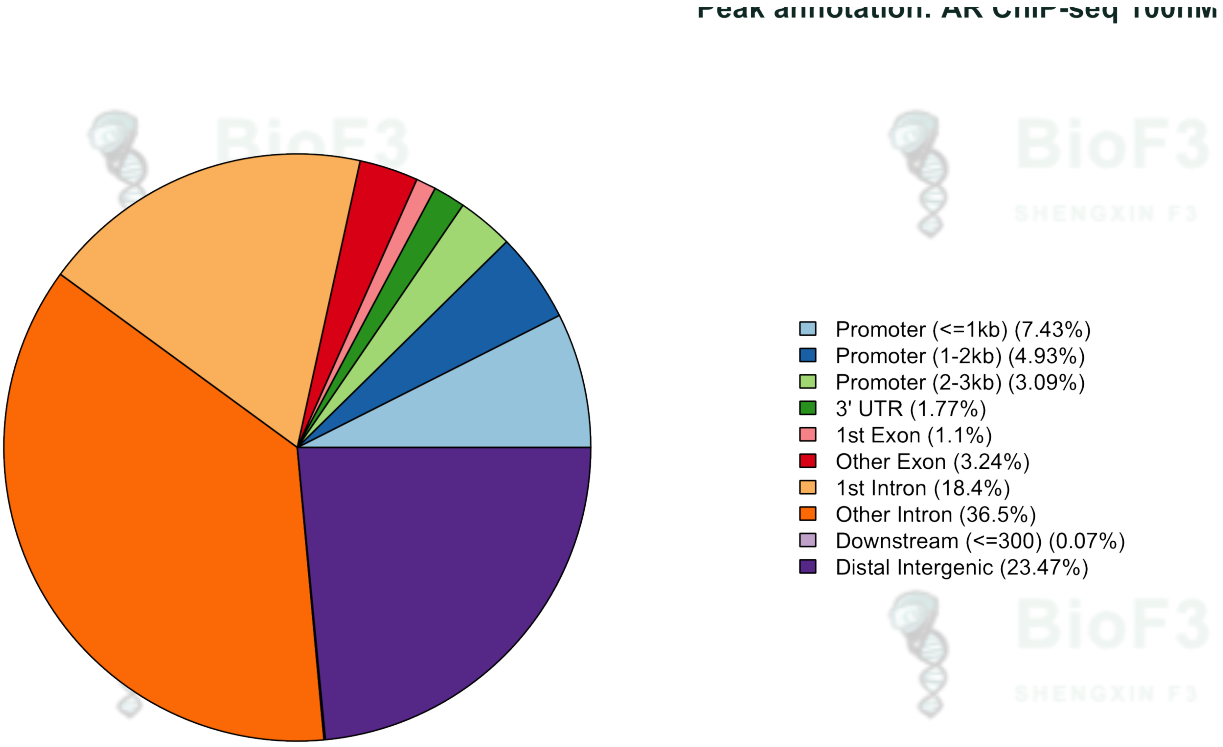


图 1：AR 100nM 的 peak 落在哪些基因组区域。启动子占比越高，说明这个 TF 越倾向于结合在基因起始位点附近。AR 是经典的启动子 + 增强子结合 TF，所以启动子和远端基因间区域都有不少。

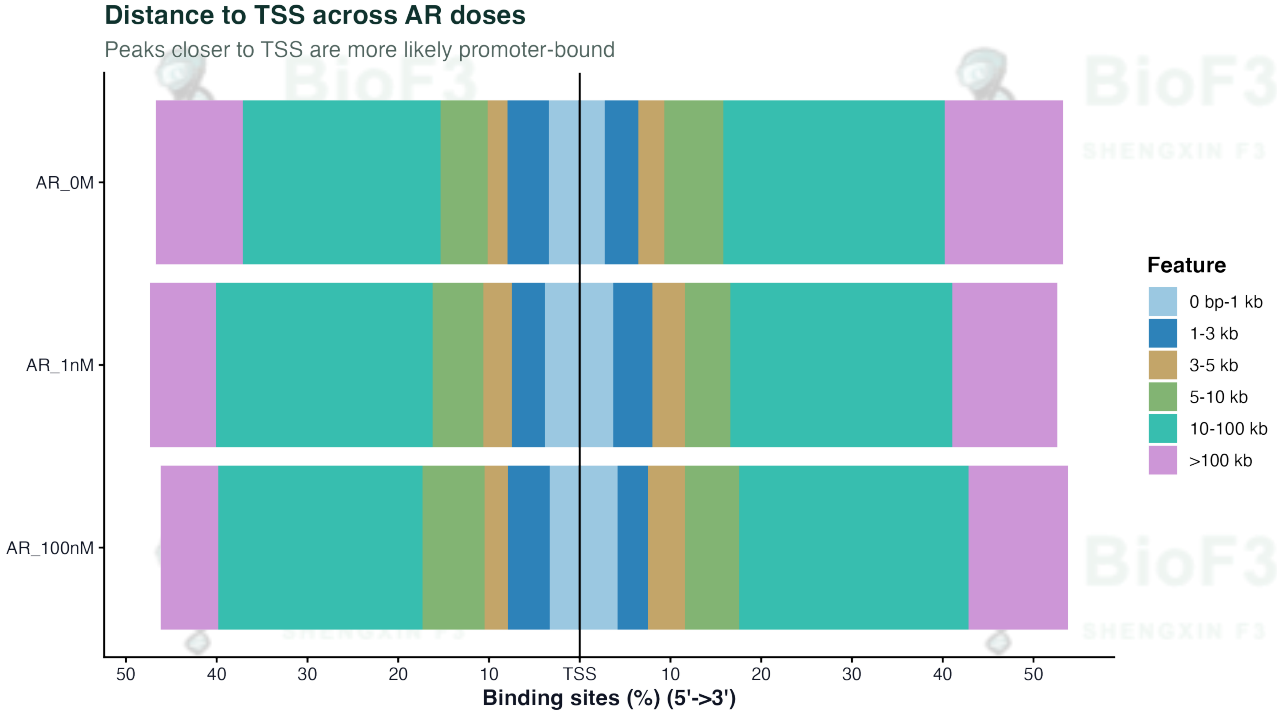


图 2：三个剂量的 peak 到最近 TSS 的距离分布。剂量越高，靠近 TSS 的 peak 比例越大 —— 说明高剂量下 AR 更多地占据启动子。

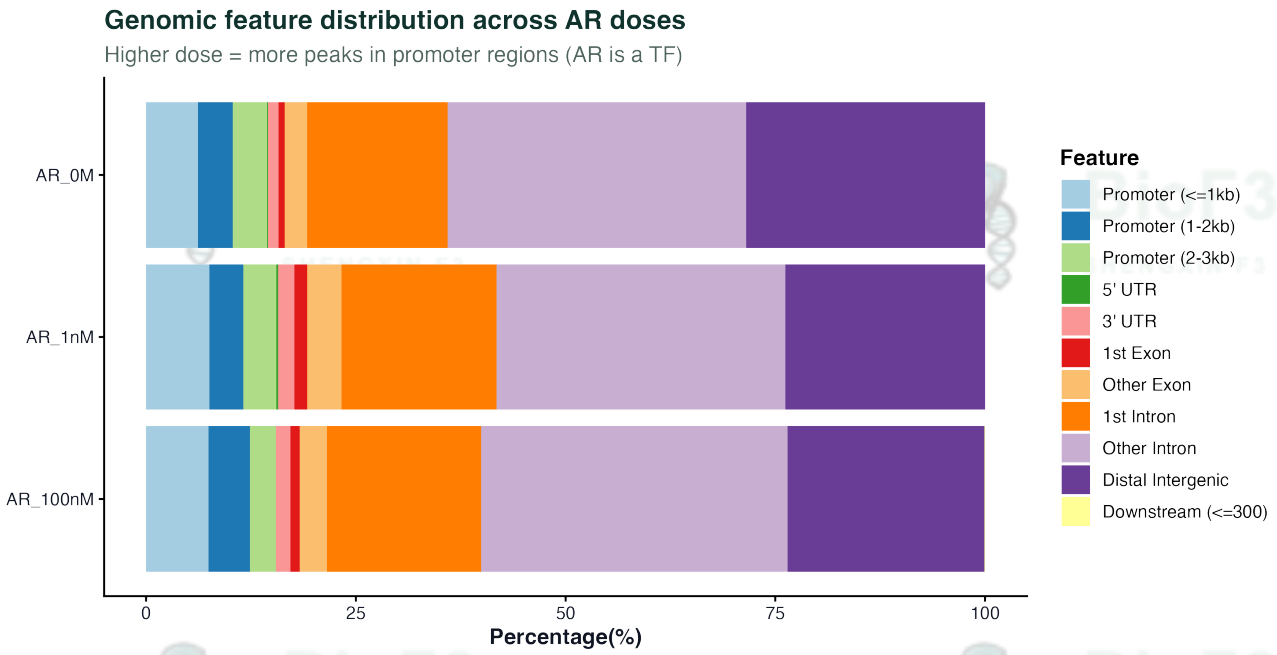


图 3：三个剂量的基因组区域分布对比。堆叠条形图一眼看出"启动子占比随剂量增加"。

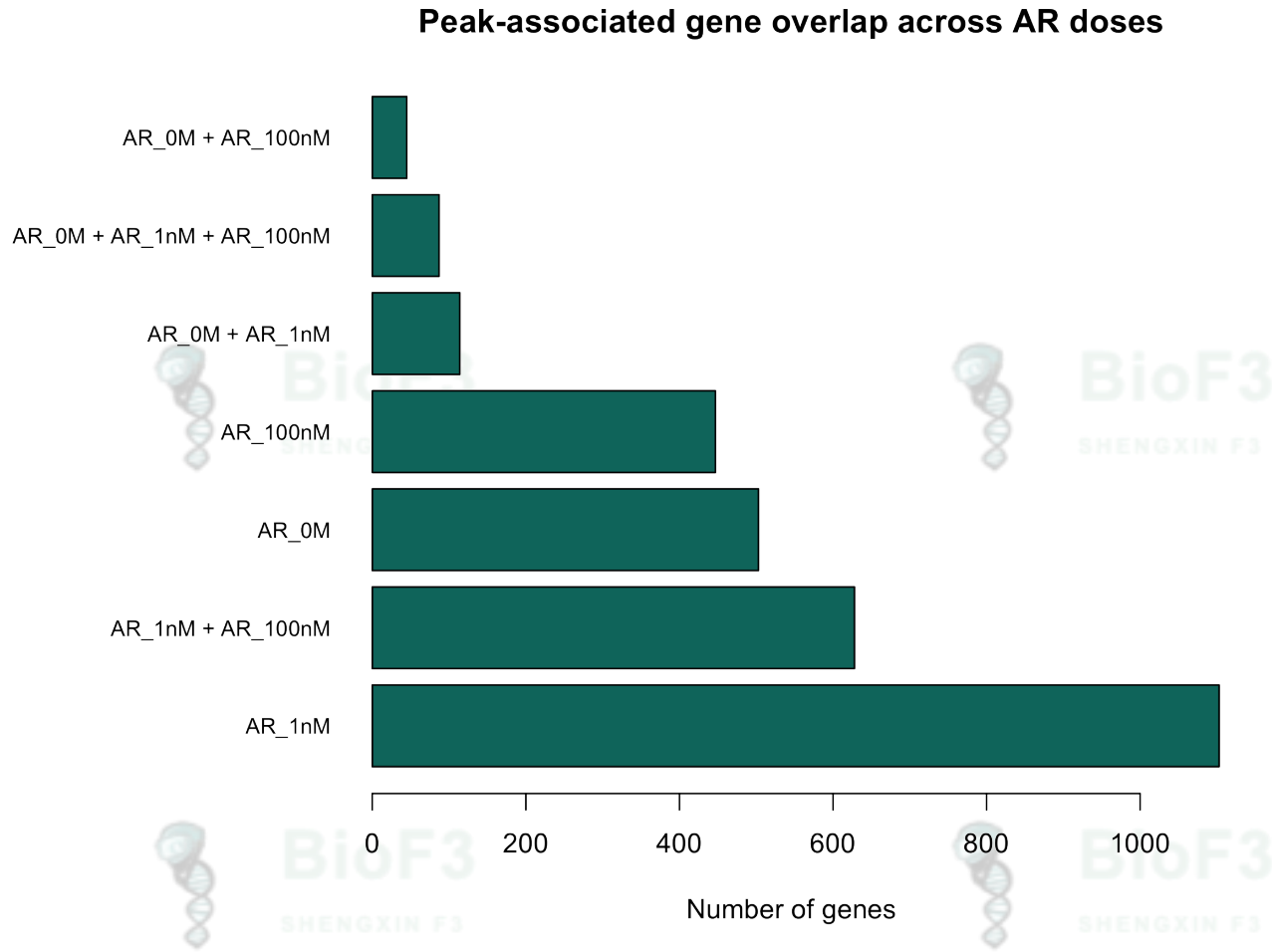


图 4：三个剂量的 peak 关联基因重叠。"三个剂量都有"的基因是 AR 的核心靶基因；"只在 100nM 出现"的是高剂量特异的。

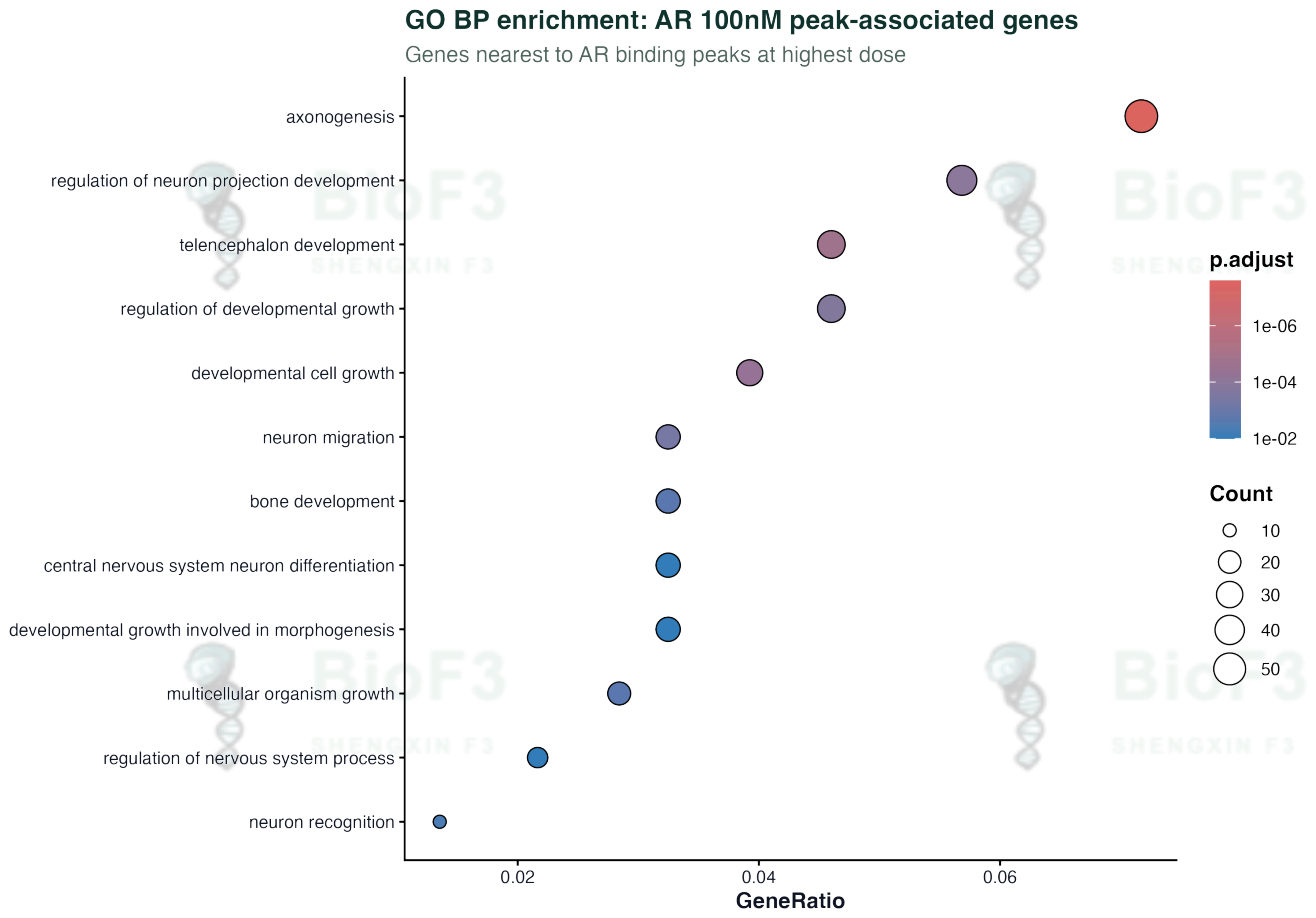


图 5：AR 100nM peak 关联基因的 GO BP 富集。应该能看到雄激素响应、细胞增殖调控等通路。

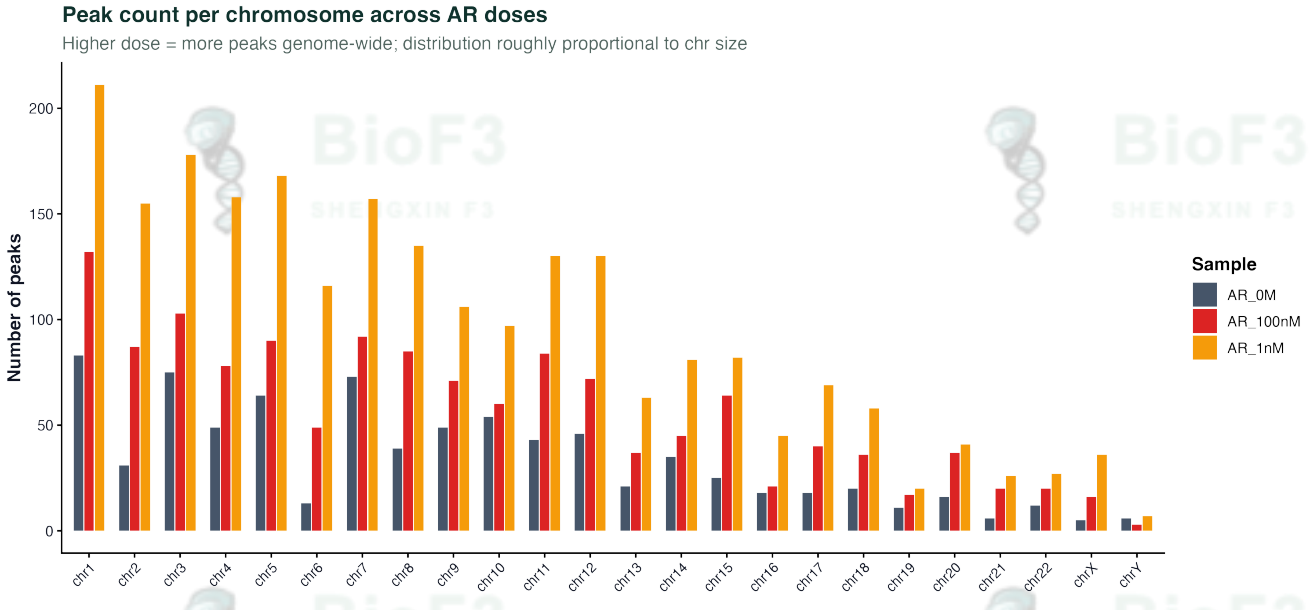


图 6：每条染色体上的 peak 数量。分布大致和染色体大小成正比，但某些染色体可能因为基因密度高而偏多。

套到自己数据上

把 `getSampleFiles()` 换成自己的 `.narrowPeak` 文件路径即可。注意：

- 参考基因组版本要和 peak 文件一致（hg19 / hg38 / mm10）
- 对应的 TxDb 包：`TxDb.Hsapiens.UCSC.hg38.knownGene`（hg38）或 `TxDb.Mmusculus.UCSC.mm10.knownGene`（小鼠）

- 如果是 broadPeak（组蛋白修饰），`annotatePeak` 的参数不用改，但解读时"启动子"的含义要注意

下载资源

epi02_chipseeker_sci.R
8 KB

下载 ChIPseeker peak 注释完整脚本 ↗

参考资源

- ChIPseeker Bioconductor 文档
- Yu et al. 2015, ChIPseeker 论文
- TxDb 包列表

04 DiffBind 差异结合分析

ChIP-seq 的差异分析和 RNA-seq 思路一样：把每个 peak 在每个样本里的 reads 数当作"表达量"，用 DESeq2 或 edgeR 做统计检验。DiffBind 把这条流程封装成几个函数，从 peak 文件 + BAM 到差异结合位点一步到位。

本章用 DiffBind 自带的 tamoxifen 数据演示：11 个乳腺癌细胞系的 ER（雌激素受体）ChIP-seq，分为 Responsive（对他莫昔芬敏感）和 Resistant（耐药）两组。

真实示例

配套脚本 `epi03_diffbind_sci.R` 在 tamoxifen 数据上跑完整的差异结合分析：

```
Rscript scripts/epigenomics/epi03_diffbind_sci.R
```

每张图看什么

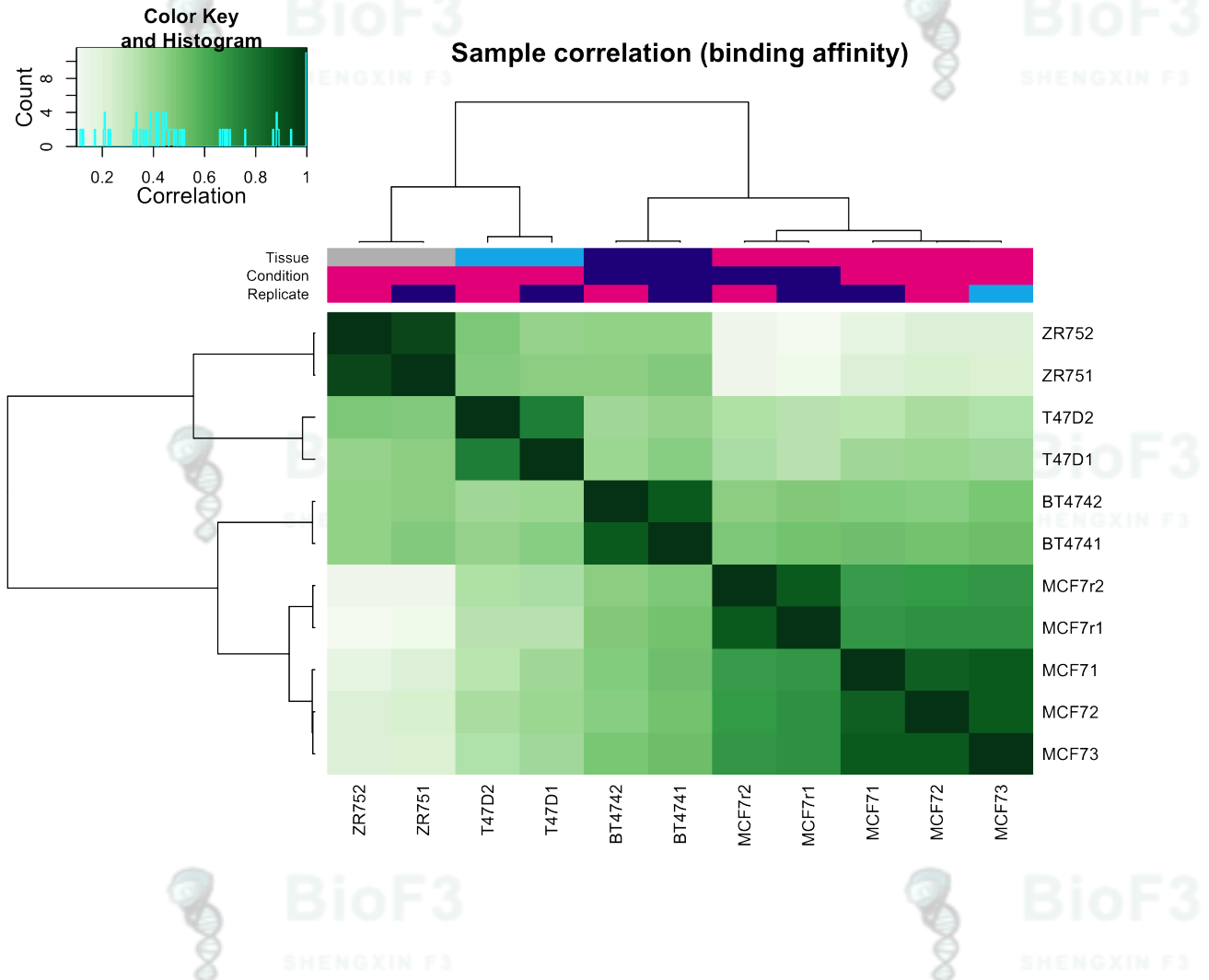


图 1：样本间 binding affinity 的相关性热图。同一条件的样本应该聚在一起。

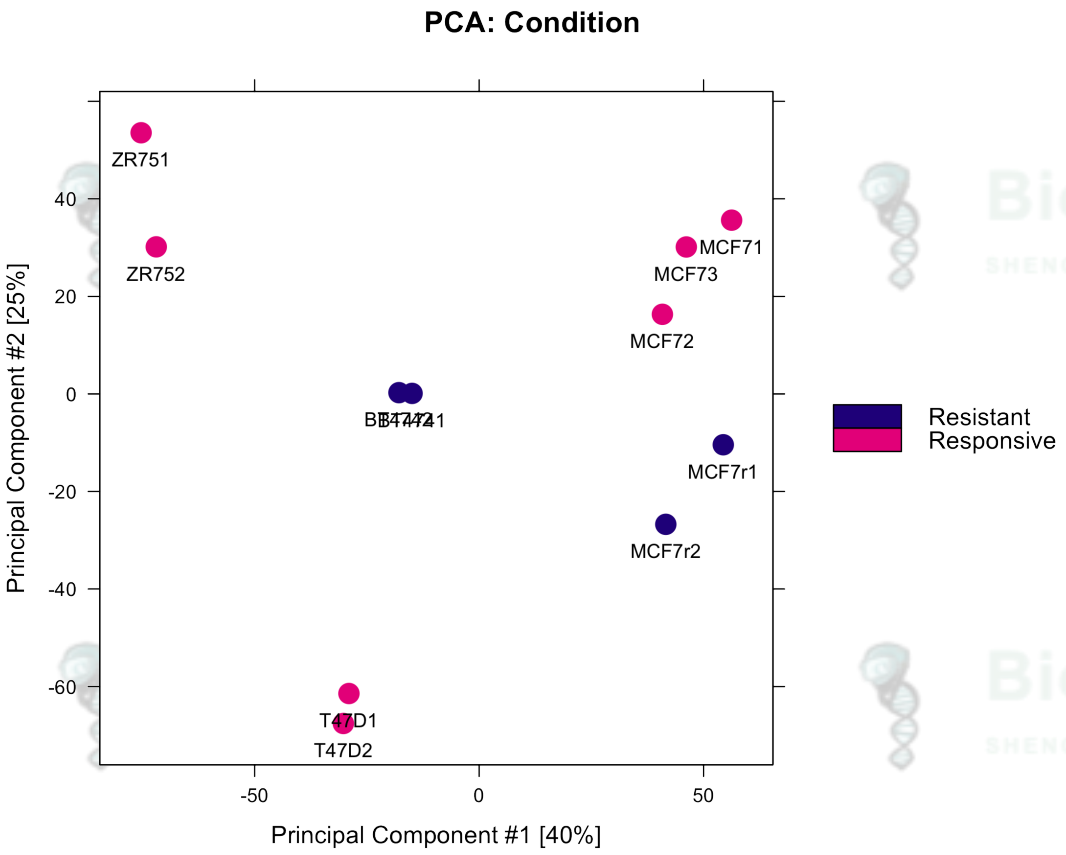


图 2：PCA 按条件着色。Responsive 和 Resistant 在 PC1 上分开。

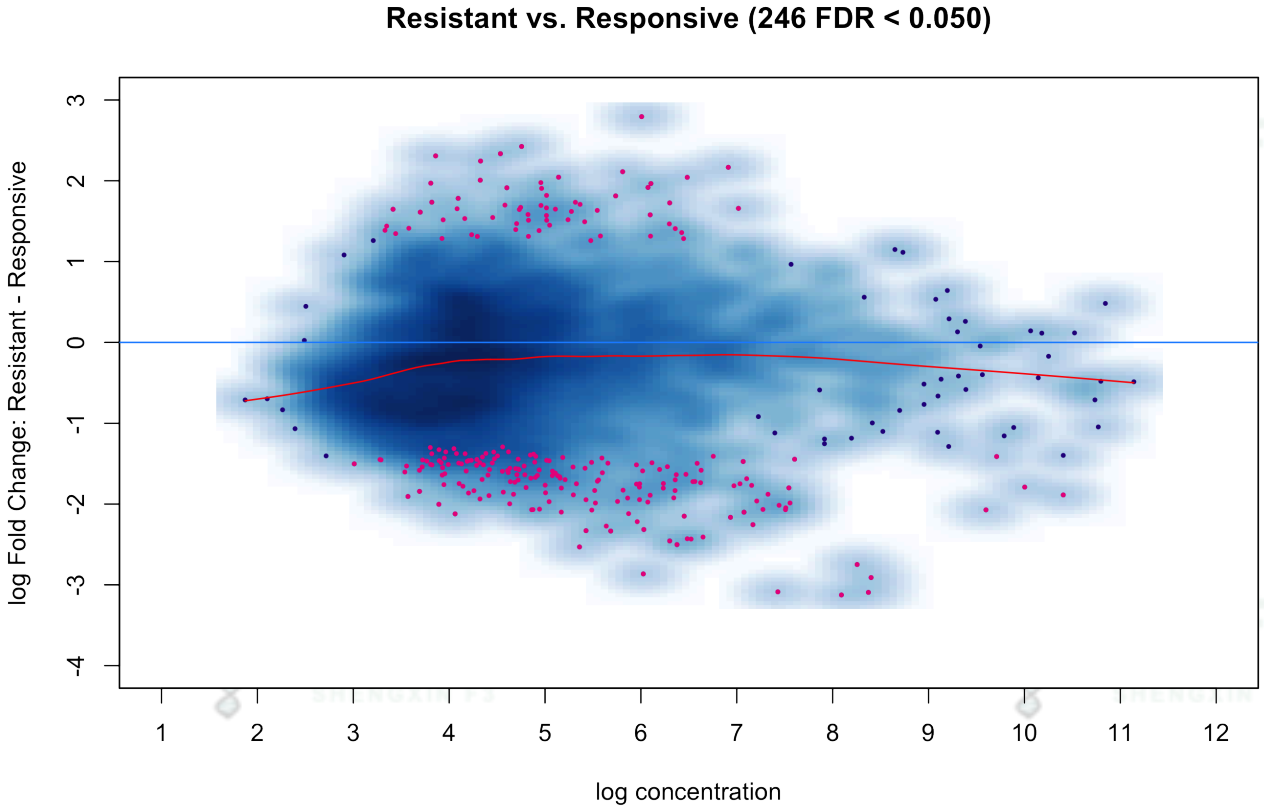


图 3：MA plot。红点是 FDR < 0.05 的差异结合位点。

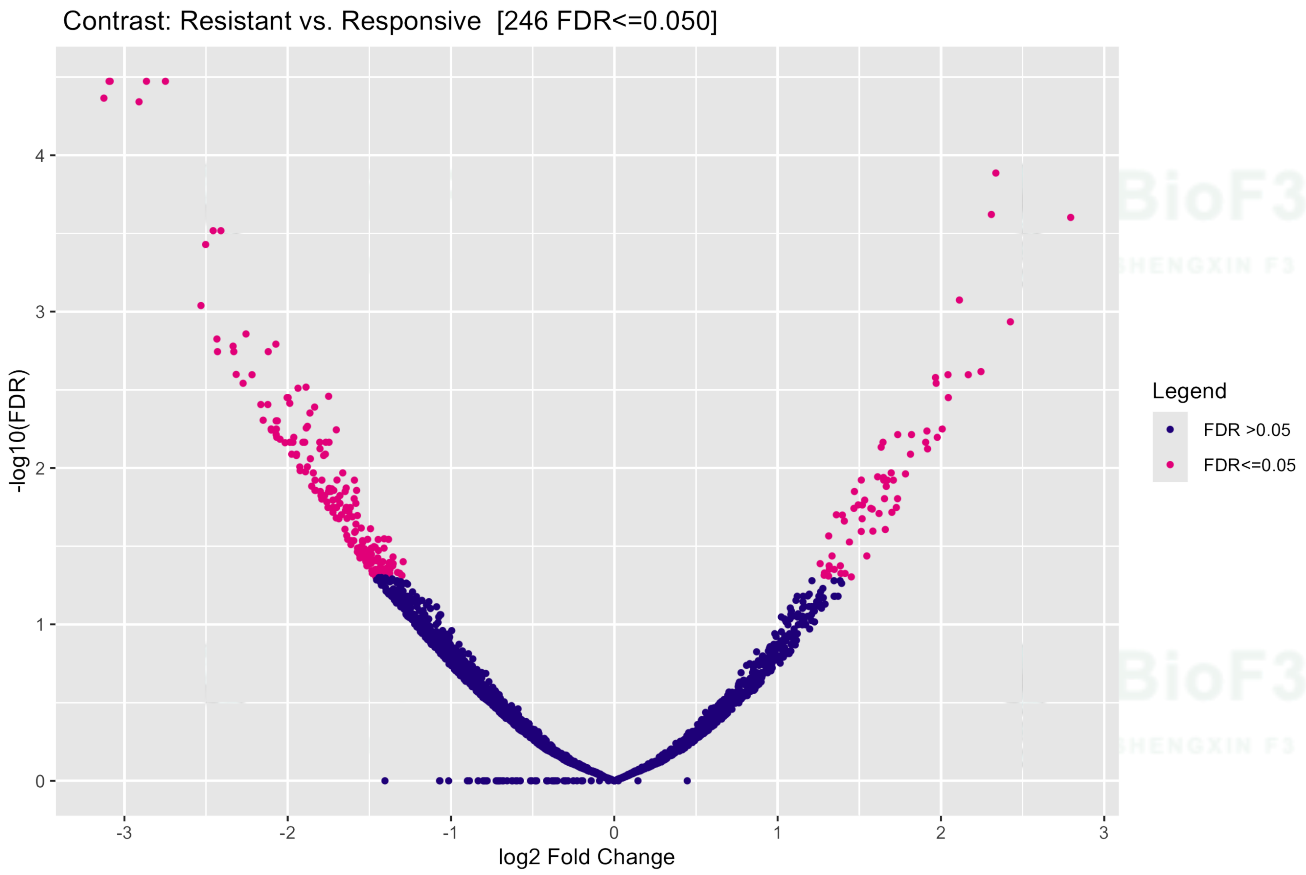


图 4：火山图。横轴 log2FC，纵轴 -log10(FDR)。

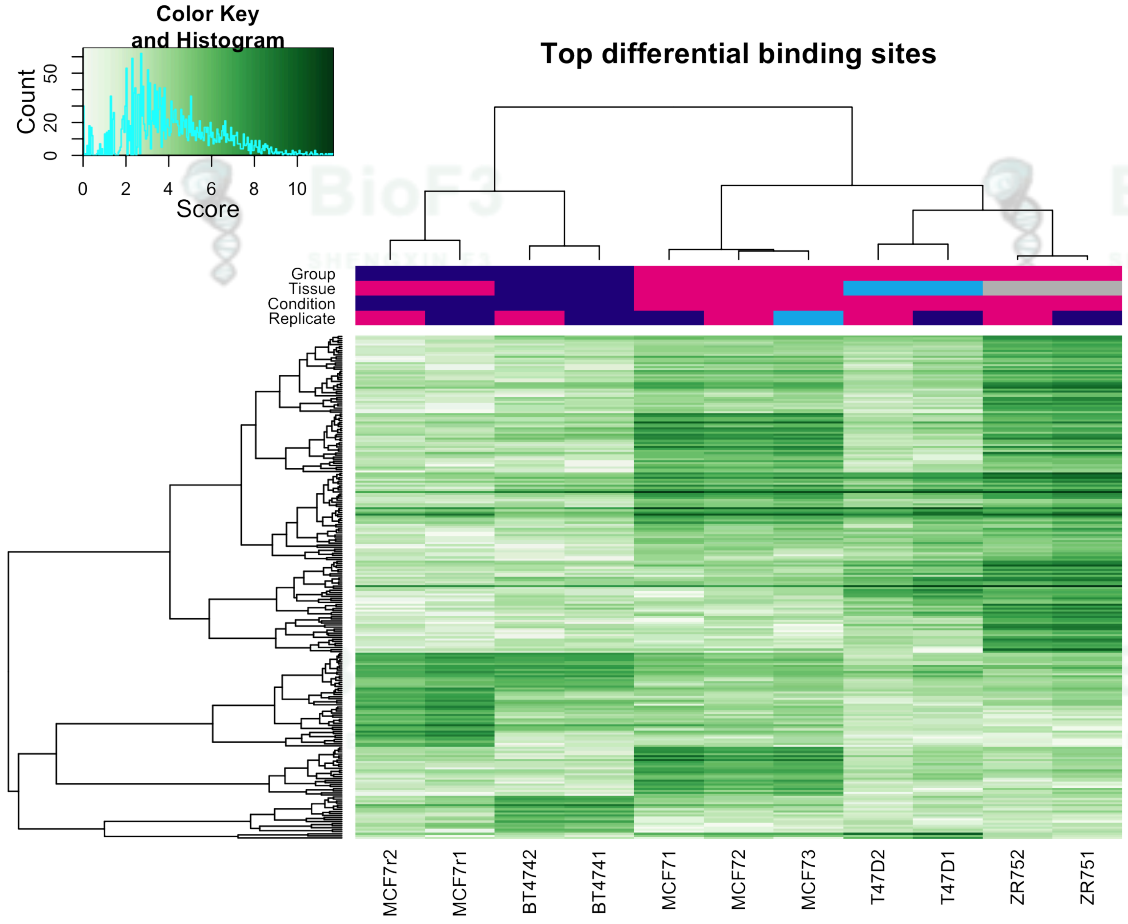
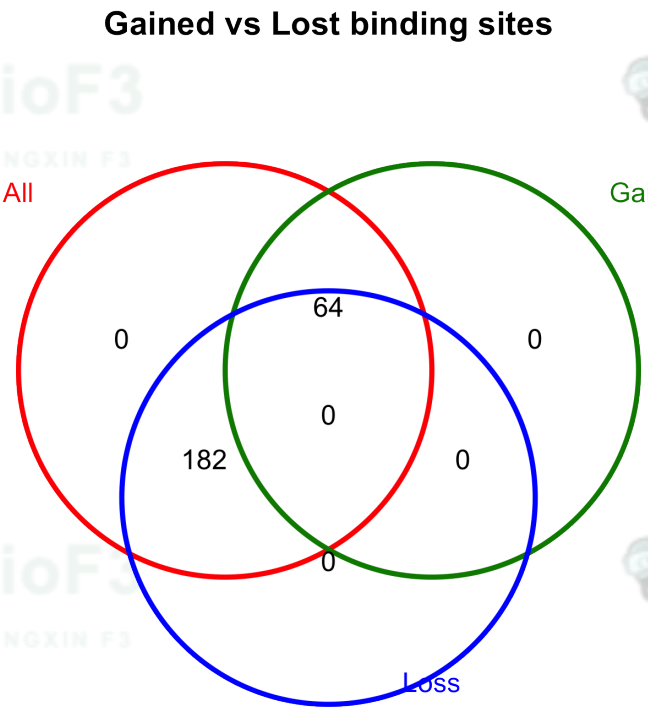


图 5：差异结合位点的 binding affinity 热图。每行一个 peak，每列一个样本。



Resistant vs. Responsive:DB:DESeq2

图 6：差异位点按方向分：Gained（Responsive 里更强）vs Lost（Resistant 里更强）。

核心代码

```
library(DiffBind)
data(tamoxifen_counts)

# 设置对比
tamoxifen <- dba.contrast(tamoxifen, categories = DBA_CONDITION)

# 差异分析（DESeq2 后端）
tamoxifen <- dba.analyze(tamoxifen, method = DBA_DESEQ2)

# 提取结果
db_report <- dba.report(tamoxifen)
```

下载资源

epi03_diffbind_sci.R
6 KB

下载 DiffBind 差异结合完整脚本 ↗

参考资源

- DiffBind vignette
- Ross-Innes et al. 2012, tamoxifen 数据源

05 Peak 可视化与多样本比较

本章把 ChIPseeker + DiffBind 的结果变成能直接用于论文的图。

配套脚本 `epi04_visualization_sci.R` 输出 6 张图：

```
Rscript scripts/epigenomics/epi04_visualization_sci.R
```

每张图看什么

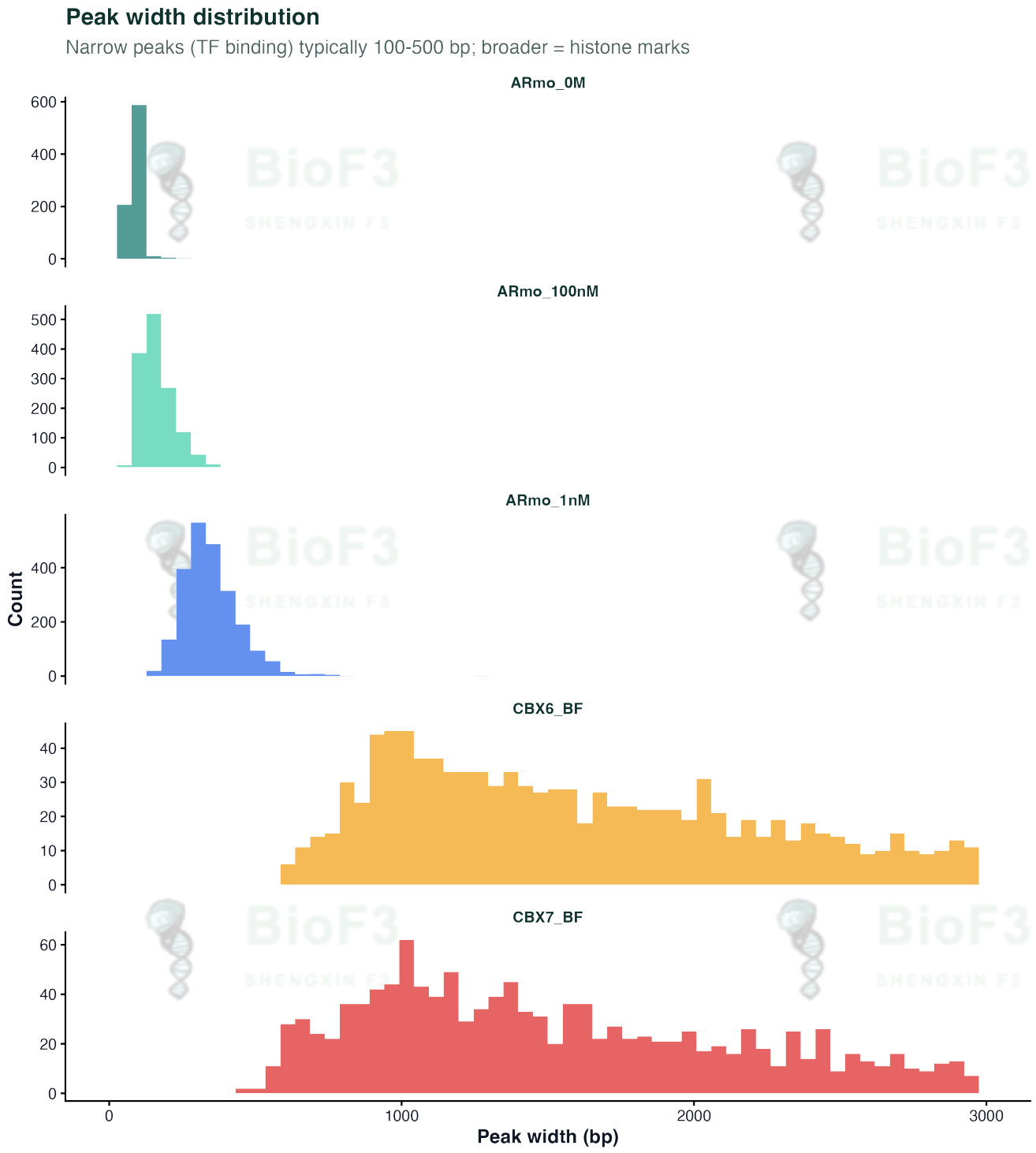


图 1：每个样本的 peak 宽度分布。TF ChIP-seq 的 narrow peak 通常 100-500bp；组蛋白修饰的 broad peak 可以到几 kb。

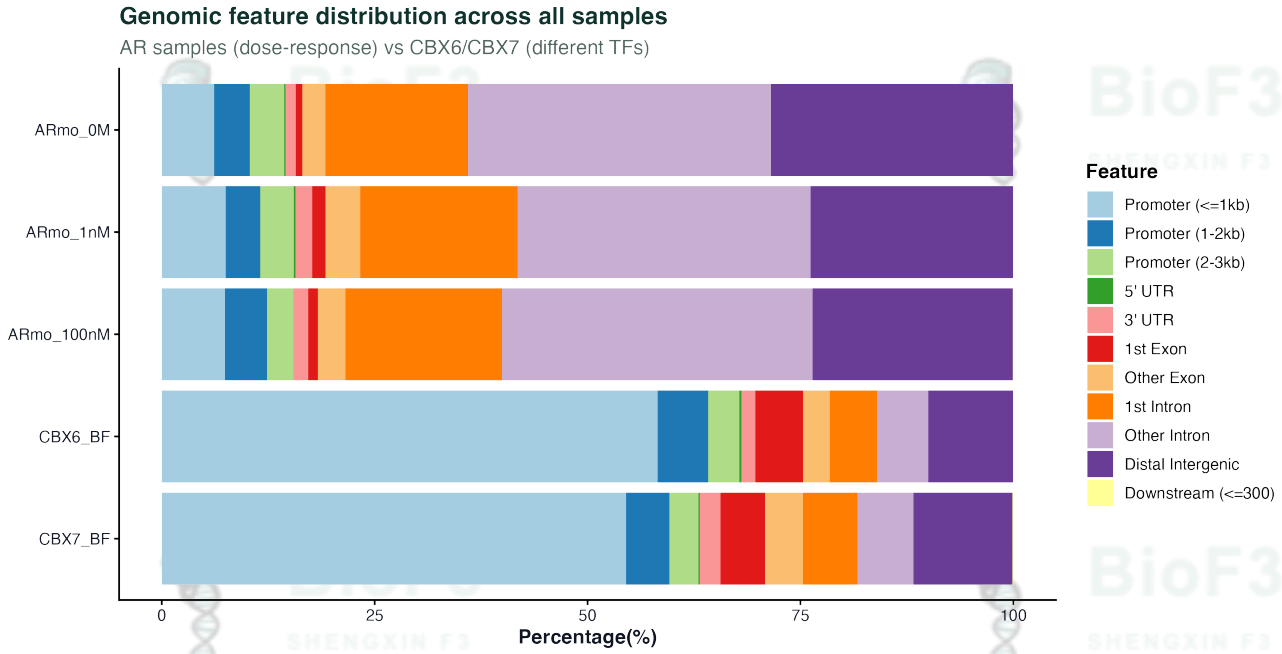


图 2：5 个样本的基因组区域分布对比。AR（TF）和 CBX6/CBX7（chromatin reader）的分布模式不同。

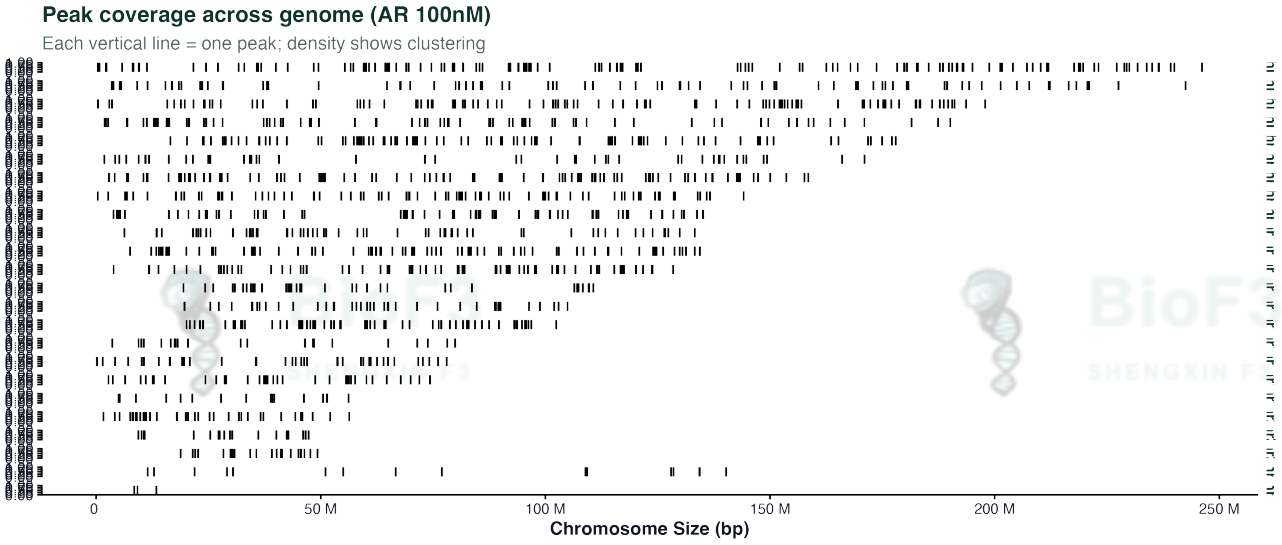


图 3：AR 100nM 的 peak 在全基因组上的覆盖密度。某些染色体区域 peak 特别密集，可能对应 super-enhancer 或基因密集区。

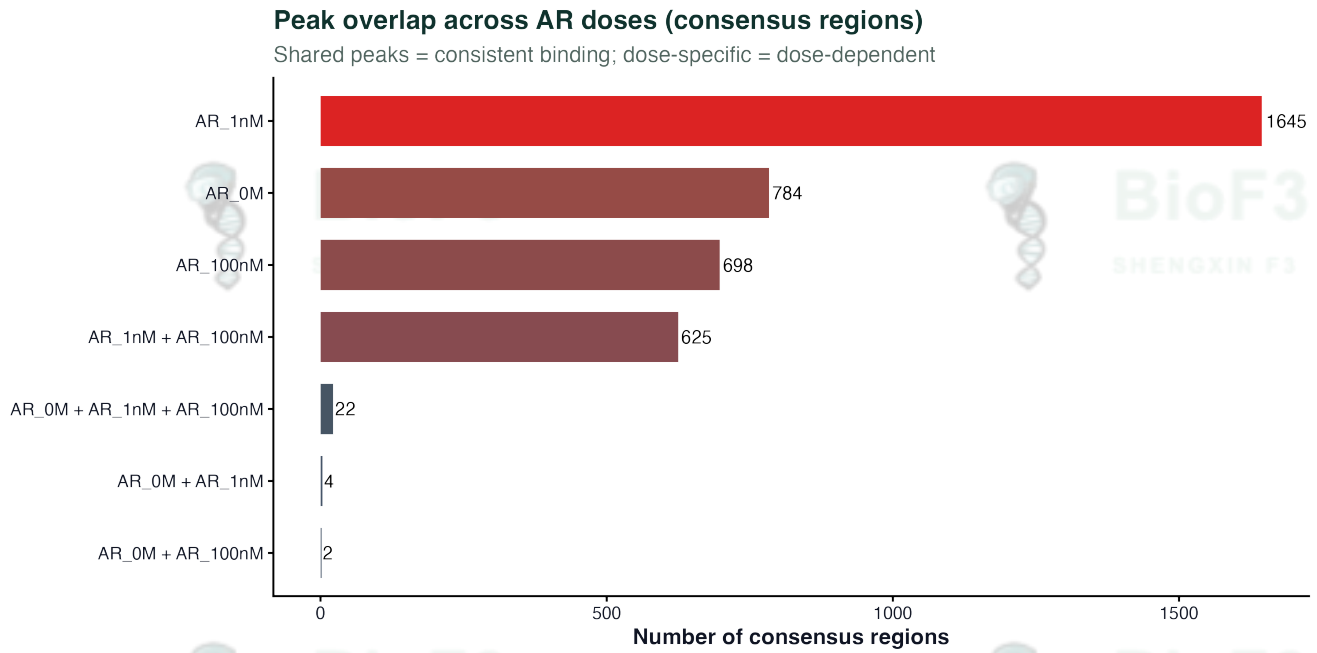


图 4：AR 三个剂量的 consensus peak 重叠。"三个剂量都有"的是核心结合位点；"只在 100nM 出现"的是剂量依赖的新增位点。

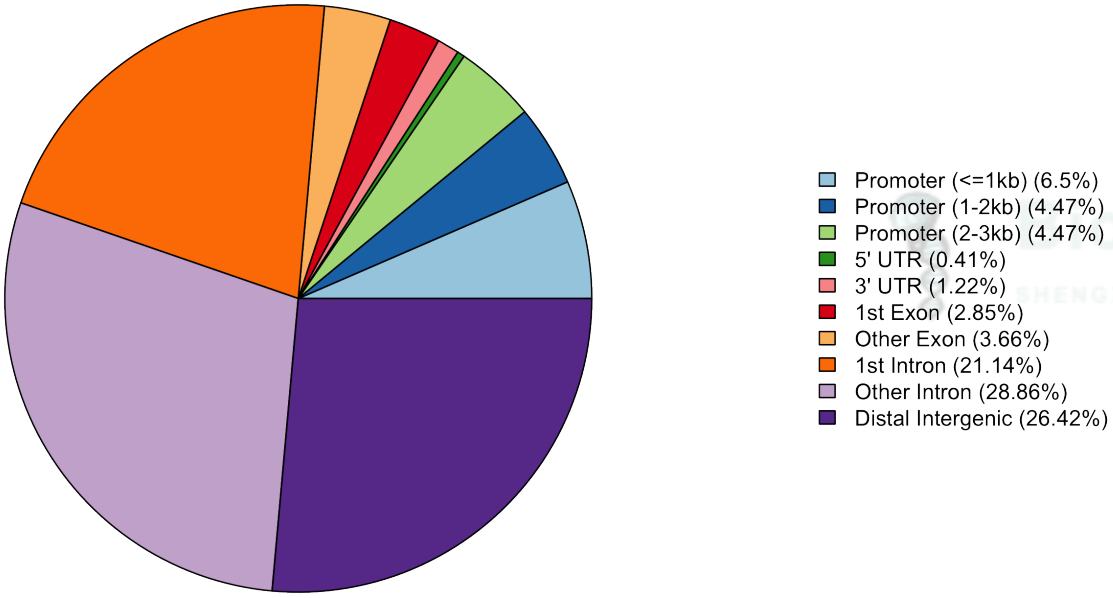


图 5：DiffBind 差异结合位点的基因组区域注释。差异位点主要落在哪里（启动子 vs 增强子）决定了它们的功能解读方向。

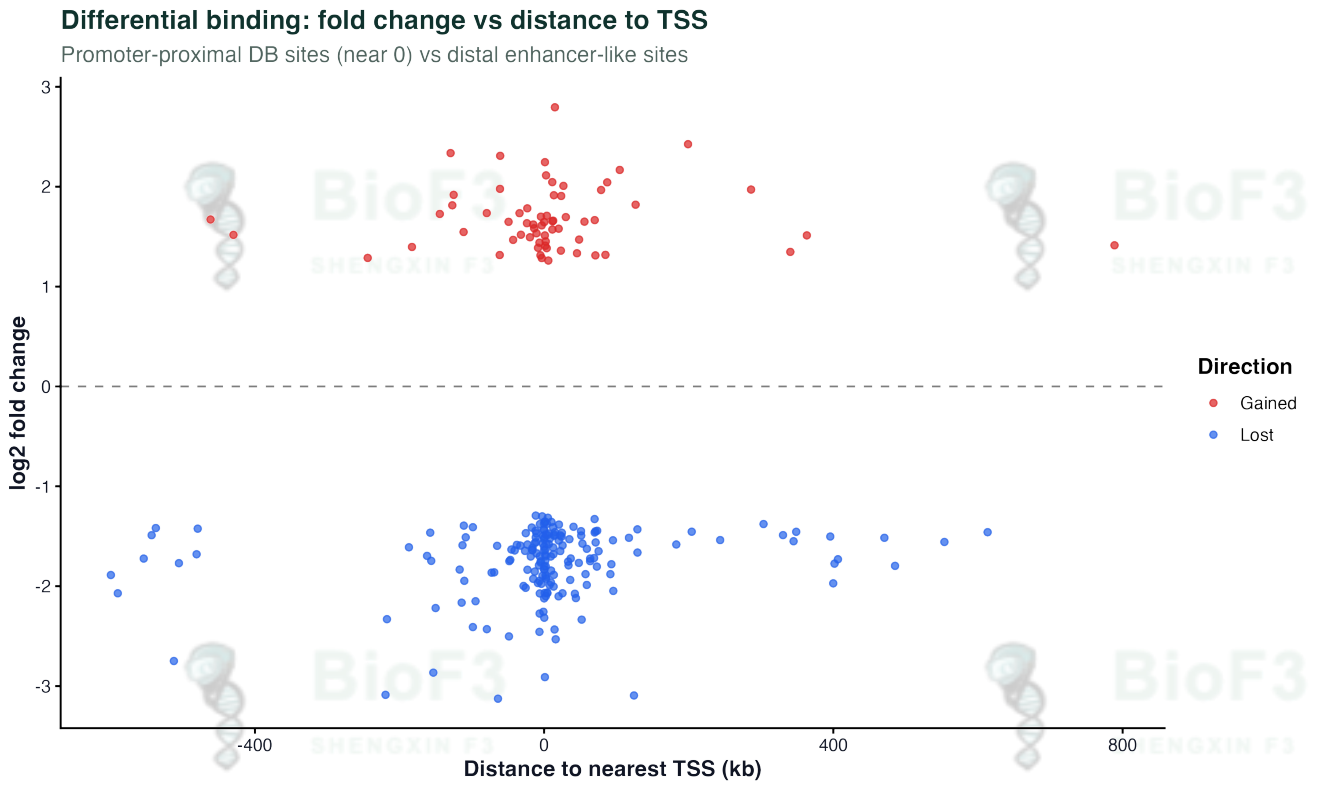


图 6：差异结合位点的 fold change vs 到最近 TSS 的距离。靠近 TSS 的位点（启动子区域）和远离 TSS 的位点（增强子区域）可能有不同的 fold change 分布。

下载资源

epi04_visualization_sci.R
8 KB

下载表观组可视化完整脚本 ↗

参考资源

- ChIPseeker 文档
- DiffBind 文档
- deepTools（Python 端覆盖度可视化）

06 Motif 富集与 HOMER

Peak 本身只是"这里有信号"，motif 分析回答的是"什么转录因子可能在这里结合"。思路是：从 peak 序列里统计过表达的短序列模式（motif），和已知 TF motif 数据库比对。

常用工具

工具	特点
HOMER	一条命令出完整报告（已知 + de novo motif），最常用
MEME Suite	学术标准，de novo 发现能力强
motifmatchr + JASPAR	R 里做 motif scanning，适合和 DiffBind 结果联动

HOMER 典型用法

```
# 安装 HOMER（一次性）
# http://homer.ucsd.edu/homer/introduction/install.html

# 对 narrowPeak 做 motif 富集
findMotifsGenome.pl peaks.narrowPeak hg38 motif_output/ \
  -size 200 -mask -p 8
```

`-size 200` 表示取 peak summit 两侧各 100bp 做分析。输出目录里会有：

- `knownResults.html` — 已知 motif 的富集排名
- `homerResults.html` — de novo 发现的 motif
- 每个 motif 的 logo 图

R 里做 motif scanning

```
library(motifmatchr)
library(TFBSTools)
library(JASPAR2020)

# 获取 JASPAR 的人类 TF motif
pfm_list <- getMatrixSet(JASPAR2020, opts = list(species = "Homo sapiens"))

# 在 peak 序列里扫描 motif
library(BSgenome.Hsapiens.UCSC.hg38)
motif_hits <- matchMotifs(pfm_list, peaks_gr, genome = BSgenome.Hsapiens.UCSC.hg38)

# 统计每个 motif 在 DE peaks vs non-DE peaks 里的富集
```

这种方式的好处是能和 DiffBind 的差异结合结果直接联动：只看"在耐药细胞里 gained 的 peak 富集了什么 motif"。

解读要点

- **已知 motif 富集**：如果你做的是 AR ChIP-seq，排第一的应该是 ARE（androgen response element）。如果不是，说明实验或分析有问题。
- **de novo motif**：HOMER 会尝试从头发现新 motif。如果 de novo 结果和已知 motif 高度相似，说明信号很强。
- **背景选择**：默认用随机基因组区域做背景。如果你的 peak 集中在启动子，用"所有启动子"做背景会更严格。

参考资源

- HOMER 官方文档
- MEME Suite
- JASPAR 数据库
- motifmatchr Bioconductor

07 ATAC-seq 分析要点

ATAC-seq 和 ChIP-seq 共享大部分分析流程（比对 → peak calling → 注释 → 差异），但有几个关键差异需要单独说明。

和 ChIP-seq 的区别

维度	ChIP-seq	ATAC-seq
测什么	特定蛋白结合位点	所有开放染色质区域
需要 input/control	是（IgG 或 input DNA）	通常不需要
fragment size	单一分布	多模态（nucleosome-free + mono/di/tri-nucleosome）
peak calling 参数	`--nomodel` 或默认	`--nomodel --shift -100 --extsize 200`
核心 QC	FRiP、IDR	TSS enrichment、fragment size 分布、FRiP

Fragment size 分布

ATAC-seq 最重要的 QC 图是 fragment size 分布：

```
~100-150 bp  → nucleosome-free fragments（信号最好的部分）
~200 bp      → mono-nucleosome
~400 bp      → di-nucleosome
~600 bp      → tri-nucleosome
```

好的 ATAC-seq 数据应该在 < 150bp 处有一个明显的 peak（nucleosome-free），然后在 ~200bp 处有第二个 peak。如果第一个 peak 不明显，说明 Tn5 转座效率低或者细胞核裂解不充分。

MACS2 参数

```
macs2 callpeak \
  -t sample.bam \
  -f BAMPE \
  --nomodel \
  --shift -100 --extsize 200 \
  -g hs \
  -n sample_atac \
  --keep-dup all \
  -q 0.05
```

关键参数：

- `-f BAMPE`：paired-end 模式，用实际 fragment 长度
- `--nomodel --shift -100 --extsize 200`：不做 fragment size 建模，直接用 Tn5 切割位点两侧 100bp
- `--keep-dup all`：ATAC-seq 的 PCR duplicate 已经在前面用 Picard 去过了

和单细胞 scATAC 的关系

单细胞实践 10 scATAC-seq 里用的 Signac 流程，底层思路和 bulk ATAC 一样（TF-IDF + LSI），只是把"每个样本"换成了"每个细胞"。bulk ATAC 的 peak 可以直接作为 scATAC 的参考 peak set。

推荐流水线

如果不想手动跑每一步，推荐用 nf-core/atacseq：

```
nextflow run nf-core/atacseq \
  --input samplesheet.csv \
  --genome GRCh38 \
  --outdir results/
```

它会自动完成 trim → align → dedup → shift → peak call → QC report 全流程。

参考资源

- ENCODE ATAC-seq pipeline
- nf-core/atacseq
- Buenrostro et al. 2013, ATAC-seq 原始论文
- Yan et al. 2020, ATAC-seq 分析最佳实践

08 DNA 甲基化分析入门

DNA 甲基化（主要是 CpG 位点的 5-甲基胞嘧啶）是最稳定的表观修饰之一。和 ChIP/ATAC 不同，甲基化分析不做 peak calling，而是直接量化每个 CpG 位点的甲基化率（0~100%）。

实验类型

技术	覆盖度	成本	适用
WGBS	全基因组 ~28M CpG	高	全景图、发现新 DMR
RRBS	富集 CpG 岛附近 ~2M CpG	中	启动子甲基化
450K / EPIC 芯片	固定位点 450K~850K	低	大样本量、TCGA 数据

BioF3 这一章聚焦 WGBS/RRBS 的 bisulfite-seq 分析。芯片数据用 `minfi` 包处理，思路不同。

分析流程

```
FASTQ → Bismark (bisulfite-aware alignment) → methylation extraction
     → per-CpG methylation table → DMR calling (methylKit / DSS / dmrseq)
```

Bismark 比对

```
# 建立 bisulfite 索引（一次性）
bismark_genome_preparation --bowtie2 reference/

# 比对
bismark --genome reference/ -1 sample_R1.fq.gz -2 sample_R2.fq.gz

# 去重
deduplicate_bismark sample_pe.bam

# 提取甲基化信息
bismark_methylation_extractor --paired-end --comprehensive --cytosine_report \
  --genome_folder reference/ sample_pe.deduplicated.bam
```

输出的 `CpG_context` 文件每行一个 CpG 位点，包含染色体、位置、甲基化 reads 数、非甲基化 reads 数。

R 里做差异甲基化

```
library(methylKit)

# 读入 Bismark 的 CpG 报告
file_list <- list("sample1.CpG_report.txt", "sample2.CpG_report.txt",
                  "sample3.CpG_report.txt", "sample4.CpG_report.txt")
obj <- methRead(file_list,
                sample.id = list("ctrl1","ctrl2","trt1","trt2"),
                assembly = "hg38",
                treatment = c(0, 0, 1, 1),
                context = "CpG",
                mincov = 10)

# 合并所有样本的 CpG 位点
meth <- unite(obj, destrand = FALSE)

# 差异甲基化位点
diff <- calculateDiffMeth(meth)
diff_25 <- getMethylDiff(diff, difference = 25, qvalue = 0.01)

# 差异甲基化区域（DMR）
# 用 tileMethylCounts 把基因组分成 1kb 窗口再做差异
tiles <- tileMethylCounts(obj, win.size = 1000, step.size = 1000)
meth_tiles <- unite(tiles)
diff_tiles <- calculateDiffMeth(meth_tiles)
```

关键概念

- **甲基化率**：某个 CpG 位点被甲基化的 reads 占总 reads 的比例
- **DMC（Differentially Methylated Cytosine）**：单个 CpG 位点的差异
- **DMR（Differentially Methylated Region）**：连续多个 CpG 位点一起变化的区域
- **覆盖度过滤**：覆盖度 < 10x 的位点噪声太大，通常过滤掉

和表达的关系

启动子区域的高甲基化通常和基因沉默相关；基因体内的甲基化和活跃转录正相关。把 DMR 和 RNA-seq 的差异基因做交叉，能找到"甲基化变化驱动表达变化"的候选基因。

参考资源

- Bismark 手册
- methylKit 教程
- DSS 差异甲基化
- dmrseq
- minfi（芯片数据）

09 表观组与转录组的整合

表观修饰（开放染色质、TF 结合、甲基化）最终要通过影响基因表达来发挥功能。把表观组数据和转录组数据放在一起看，能回答"哪些表观变化真正驱动了表达变化"。

常见整合策略

策略	输入	输出	工具
Peak-gene 关联	差异 peak + 差异基因	重叠基因列表	ChIPseeker + 自定义脚本
相关性分析	peak 信号矩阵 + 表达矩阵	peak-gene 相关性	cor.test / LOLA
调控网络推断	motif + 表达 + peak	TF → target 网络	SCENIC / pySCENIC
多组学因子分析	多层矩阵	共变因子	MOFA2 / mixOmics

Peak-gene 关联（最简单）

```
library(ChIPseeker)

# 差异 peak 注释到最近基因
db_anno <- annotatePeak(db_peaks, TxDb = txdb)
db_genes <- unique(as.data.frame(db_anno)$SYMBOL)

# 差异基因（来自 DESeq2）
de_genes <- res_df$SYMBOL[res_df$padj < 0.05]

# 交集
overlap <- intersect(db_genes, de_genes)
cat("Overlap:", length(overlap), "genes\n")

# Fisher 检验看是否显著富集
fisher.test(matrix(c(
  length(overlap),
  length(setdiff(db_genes, de_genes)),
  length(setdiff(de_genes, db_genes)),
  total_genes - length(union(db_genes, de_genes))
), nrow = 2))
```

如果 overlap 显著大于随机期望，说明表观变化和表达变化确实有关联。

方向一致性检查

更严格的验证：不仅看"有没有重叠"，还看"方向是否一致"：

```r
# 合并 peak FC 和 gene FC
merged <- inner_join(
  data.frame(gene = db_genes_df$SYMBOL, peak_fc = db_genes_df$Fold),
  data.frame(gene = de_df$SYMBOL, rna_fc = de_df$log2FoldChange)
)

# 散点图：peak FC vs RNA FC
ggplot(merged, aes(x = peak_fc, y = rna_fc)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Peak log2FC (ChIP/ATAC)", y = "RNA log2FC")
```

正相关 = 开放区域增加的基因表达也增加（符合预期）。如果是甲基化数据，启动子区域应该是负相关（甲基化增加 → 表达下降）。

SCENIC：从 scATAC + scRNA 推断调控网络

如果有配对的单细胞数据（10x Multiome 或分别测的 scRNA + scATAC），SCENIC+ 能推断出"哪个 TF 通过哪个增强子调控哪个基因"：

```python
# Python (pySCENIC+)
import scenicplus

# 输入：scRNA AnnData + scATAC AnnData + motif 数据库
# 输出：TF → enhancer → gene 的三元组网络
```

这是目前单细胞表观组整合的最前沿方向，计算量大但信息量也最大。

实用建议

1. **先做简单的 peak-gene overlap**，确认方向一致性
2. 如果 overlap 显著且方向一致，再做更复杂的网络推断
3. 多组学整合的结果要用独立实验验证（比如 CRISPRi 敲掉某个增强子看表达是否下降）
4. 不要过度解读"相关性 = 因果性"

参考资源

- SCENIC+
- MOFA2
- LOLA（区域富集）
- Corces et al. 2018, ATAC + RNA 整合