BIOF3 组学数据分析

08 表观组与转录组的整合

导出日期：2026年5月12日

08 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