BIOF3 组学数据分析

08 整合结果的生物学解读与报告

导出日期：2026年5月12日

08 整合结果的生物学解读与报告

跑完 MOFA、SNF 或机器学习模型之后，你手上有一堆因子、模块、聚类标签和特征列表。这些数字本身不是结论——需要把它们翻译成生物学语言，并以可复现的方式写进论文。

从因子到通路

MOFA 因子的权重列表本质上是一个 ranked gene list，可以直接送进 GSEA：

```
library(MOFA2)
library(fgsea)
library(msigdbr)

# 获取 Factor 1 在 mRNA 层的权重
weights <- get_weights(model, views = "mRNA", factors = 1, as.data.frame = TRUE)
gene_ranks <- setNames(weights$value, weights$feature)
gene_ranks <- sort(gene_ranks, decreasing = TRUE)

# Hallmark 基因集
hallmark <- msigdbr(species = "Homo sapiens", category = "H")
pathways <- split(hallmark$gene_symbol, hallmark$gs_name)

fgsea_res <- fgsea(pathways, gene_ranks, minSize = 15, maxSize = 500)
sig_pathways <- fgsea_res[padj < 0.05][order(NES, decreasing = TRUE)]
head(sig_pathways[, .(pathway, NES, padj)], 10)
```

如果 Factor 1 富集在 cell cycle 和 DNA repair 通路，而且它在 Mutations 层也有高方差贡献，你可以说"Factor 1 捕捉了与基因组不稳定性相关的跨组学变异"。

从聚类到亚型特征

SNF 或 consensus clustering 给出的聚类标签需要进一步表征：

```r
# 每个亚型的差异特征
library(limma)

design <- model.matrix(~ 0 + factor(clusters))
colnames(design) <- paste0("C", 1:ncol(design))

fit <- lmFit(rna_final, design)
contrast_mat <- makeContrasts(C1 - C2, C1 - C3, C2 - C3, levels = design)
fit2 <- contrasts.fit(fit, contrast_mat)
fit2 <- eBayes(fit2)

# C1 vs C2 的 top 差异基因
topTable(fit2, coef = 1, number = 20)
```

把每个亚型的 marker 基因列出来，再做 GO/KEGG 富集，就能给亚型一个生物学标签（比如"免疫活跃型""代谢重编程型"）。

多层证据汇总

好的多组学分析不是把每层结果分别报告，而是展示"多层证据指向同一个结论"。一个常见的汇总方式：

```r
# 某个基因在三层的一致性
gene <- "TP53"

# RNA 层：在亚型间差异表达
rna_p <- t.test(rna_final[gene, ] ~ clusters)$p.value

# 甲基化层：promoter 甲基化差异
meth_p <- t.test(meth_final[gene, ] ~ clusters)$p.value

# 蛋白层：蛋白丰度差异
prot_p <- t.test(prot_final[gene, ] ~ clusters)$p.value

cat(gene, "across layers:\n")
cat("  RNA p-value:", format(rna_p, digits = 3), "\n")
cat("  Methylation p-value:", format(meth_p, digits = 3), "\n")
cat("  Protein p-value:", format(prot_p, digits = 3), "\n")
```

写 Methods 段

多组学论文的 Methods 段需要覆盖以下内容。下面是一个模板：

```
Multi-omics integration was performed using MOFA2 (v1.8.0).
Input data included transcriptomics (N genes after filtering),
DNA methylation (N CpG sites, promoter-level aggregation),
and proteomics (N proteins). Each layer was independently
preprocessed: RNA-seq counts were variance-stabilized using
DESeq2; methylation beta values were Noob-normalized with minfi;
protein intensities were log2-transformed and batch-corrected
with limma::removeBatchEffect. Features were mapped to gene
symbols using biomaRt. The model was trained with K=15 factors,
slow convergence mode, and random seed 42. Downstream enrichment
analysis used fgsea with MSigDB Hallmark gene sets.
```

关键原则：

- 写清楚每层的预处理方法和工具版本
- 说明特征映射策略（probe → gene 用了什么注释）
- 报告整合方法的参数（因子数、迭代次数、收敛标准）
- 提供代码仓库链接

可视化建议

多组学论文常见的图：

图类型	用途	工具
方差解释热图	展示因子在各层的贡献	MOFA2 内置
多层 heatmap	同一批样本在不同组学的模式	ComplexHeatmap
Sankey / alluvial 图	展示亚型在不同方法间的对应	ggalluvial
网络图	融合网络结构	igraph、Cytoscape
Forest plot	多组学 Cox 回归系数	forestplot

```r
library(ComplexHeatmap)

# 多层 heatmap：同一批样本，上面是 RNA，中间是甲基化，下面是蛋白
ht1 <- Heatmap(rna_final[top_genes, order(clusters)], name = "RNA",
               show_column_names = FALSE, cluster_columns = FALSE)
ht2 <- Heatmap(meth_final[top_genes, order(clusters)], name = "Meth",
               show_column_names = FALSE, cluster_columns = FALSE)
ht3 <- Heatmap(prot_final[top_genes, order(clusters)], name = "Prot",
               show_column_names = FALSE, cluster_columns = FALSE)

ht_list <- ht1 %v% ht2 %v% ht3
draw(ht_list, column_title = "Multi-omics heatmap by subtype")
```

可复现性清单

发表前检查：

- 原始数据是否上传到公共数据库（GEO、PRIDE、GDC）

- 分析代码是否在 GitHub/Zenodo 上公开
- R session info 是否记录（`sessionInfo()`）
- 随机种子是否固定
- 中间结果（模型文件、聚类标签）是否保存

参考资源

- fgsea Bioconductor
- msigdbr 基因集
- ComplexHeatmap 完整手册
- ggalluvial 包
- 多组学论文写作指南 (Subramanian et al. 2020)