BIOF3 组学数据分析

多组学整合实践教程

导出日期：2026年6月27日

多组学整合实践教程

单一组学常常回答不了完整问题。转录组看到 mRNA 变化，却不知道蛋白层面有没有跟着改变；表观组解释了染色质开放，却不直接告诉你哪些基因表达了；基因组变异落在了某个启动子上，也需要表达和修饰数据才能说明它的效应。多组学整合的目标不是把所有数据塞进一个矩阵，而是让每一层提供一段证据，共同支撑一个生物学判断。

本专栏讲如何把多层数据放进同一个分析框架：从样本匹配、特征对齐，到常用整合方法（相关、网络、因子模型、机器学习）的适用场景和取舍。

一个项目大致是什么样子

假设你手上有一个队列：50 个样本，每个样本都有转录组、甲基化和蛋白质组数据，外加临床表型（亚型、分期、生存时间）。从这里到"一个能解释某个亚型的多组学特征"，大致要走：

步骤	典型产物	常用工具
样本对齐	三层数据都能匹配到同一批样本 ID	R、dplyr
各层预处理	归一化、特征筛选后的矩阵	各组学专属工具
特征对应	甲基化 ↔ 基因、蛋白 ↔ 基因的映射	biomaRt、org.Hs.eg.db
相关性探索	cis 相关、跨层相关图	ggplot2、ComplexHeatmap
整合建模	多组学因子 / 共识聚类 / 分类模型	MOFA2、mixOmics、SNF
模块解读	每个因子的生物学含义	GSEA、文献查证
与表型关联	因子/模块 vs 表型	线性模型、生存分析
交付	图、表、方法段	R Markdown、Quarto

在"整合建模"这一步，常见做法分三类：

- **早期整合**（early integration）：把各层特征拼成一个大矩阵，后续按单组学思路处理。简单，但容易被维度高的那一层主导。
- **中期整合**（intermediate integration）：在模型层面同时考虑多组学，典型代表是 MOFA 和 mixOmics。
- **晚期整合**（late integration）：各层独立分析，最后在结果层面汇总，比如合并 p 值或做共识聚类。

没有谁最好。样本少、特征差异大时 MOFA 更稳；样本充足、特征维度可比时早期整合也能跑通；目标是临床预测时机器学习整合（Random Forest、神经网络）更常见。

常见工具栈

阶段	工具	说明
数据载体	MultiAssayExperiment、SummarizedExperiment	R / Bioconductor 标准容器
相关性与网络	WGCNA、igraph	模块识别和拓扑分析
多组学因子	MOFA2、mixOmics	最常用的中期整合
相似性网络	SNFtool	适合亚型发现
机器学习	caret、tidymodels、scikit-learn	分类/回归/预测模型
生存分析	survival、survminer	Cox 回归、KM 曲线
可视化	ComplexHeatmap、ggplot2、pheatmap	多层 heatmap、associations 图

MOFA2 是个值得单独提的工具。它用概率矩阵分解识别"跨组学共同变异"的潜因子，对缺失值友好，对样本量小也能用，现在几乎是多组学整合教学的标配。

推荐公开数据集

多组学公开数据集数量不多，但下面几个足够做教学练习：

数据集	内容	适合	入口
TCGA 泛癌	转录组 + 甲基化 + 变异 + 临床	整合分析、亚型发现	GDC Portal
CPTAC	蛋白组 + 转录组 + 基因组	蛋白-RNA 整合	CPTAC
MOFA2 包示例（CLL）	表达 + 甲基化 + 突变 + 药物响应	MOFA 入门，200 个样本	MOFA2
TCGA + xCell 免疫组分	癌症转录组 + 免疫细胞组成推断	跨层相关分析	GDC

MOFA2 的 CLL 数据集特别适合教学：4 个组学层、200 个样本，包里 load 进来就能跑。

最小可跑的例子

下面用 MOFA2 自带的 CLL 数据做一次最简单的整合分析，从加载数据到看出潜因子的生物学含义，不到 15 行代码：

```r
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
BiocManager::install("MOFA2")

library(MOFA2)

# 载入示例数据：慢淋（CLL）多组学
file <- system.file("extdata", "CLL_data.hdf5", package = "MOFA2")
model <- load_model(file)

# 先看整体：每个组学在模型里的方差贡献
plot_variance_explained(model, x = "view", y = "factor")

# 看某个因子在各层数据上解释了多少方差
plot_variance_explained(model, max_r2 = 10)

# 某个因子上权重最高的特征（基因/甲基化位点等）
plot_top_weights(model, view = "mRNA", factor = 1, nfeatures = 10)

# 把因子得分投射到样本元信息上（比如亚型）
metadata <- samples_metadata(model)
head(metadata)
```

`plot_variance_explained` 会输出一张热图，每行是潜因子、每列是组学层。有些因子只在一层里有高贡献（组学特异），有些跨多层都有贡献（跨组学共同变异），后者通常更有生物学意义。

理解了这个例子，再去读 MOFA 论文或 CPTAC 的整合分析会容易很多。

专栏模块规划

模块	主题	状态
01	多组学项目设计与样本匹配	已上线
02	各组学数据清洗与特征对应	已上线
03	跨层相关性探索	已上线
04	WGCNA 与共表达模块	已上线
05	MOFA2 因子分析实战	已上线
06	SNF 相似性网络与亚型识别	已上线
07	机器学习预测模型	已上线
08	整合结果的生物学解读与报告	已上线

所有 8 个模块已上线。Module 05 带可跑脚本和真实数据图。

推荐前置知识

- bulk RNA-seq 实践教程
- 表观组学实践教程
- 蛋白质组学实践教程

- R 数据整理与 ggplot2 可视化

多组学分析的每一层预处理都涉及该组学的专属知识，建议先熟悉至少两个单组学的流程再来做整合。

常见坑

坑 1：期望整合一定比单组学好

很多人认为"数据越多结论越好"，但如果某一层数据质量差或与研究问题无关，强行整合反而引入噪声。跑完整合模型后一定要和单组学结果做对照实验，看是否真的有提升。没有提升本身也是有价值的结论。

坑 2：把相关当因果

跨层相关性分析发现甲基化和表达负相关，不代表甲基化"导致"了沉默——两者可能同时受上游信号（如转录因子结合）驱动。需要因果推断时应考虑 Mendelian randomization 或 mediation analysis，而不是仅凭 Pearson r 下结论。

坑 3：忽略各层数据量级差异导致主导效应

早期整合（拼接矩阵）时，如果 RNA-seq 有 20000 个基因而蛋白组只有 3000 个，模型会被转录组主导。解决方案是在拼接前做特征数平衡（每层取 top N）或用中期整合方法（MOFA2、mixOmics）让各层权重自动学习。

坑 4：样本量不够就上复杂模型

MOFA2 建议至少 15 个配对样本，SNF 在 30 以下不稳定。10 个样本就想跑 multi-omics factor analysis，结果的因子不可复现。样本少时先做简单的 cis 相关和单基因多层验证，不要强上降维模型。

下一步

接着深入：

- 01 多组学项目设计与样本匹配 — 从实验设计角度理解整合的起点
- 05 MOFA2 因子分析实战 — 动手跑最常用的中期整合方法

横向延伸：

- 蛋白质组学实践教程 — 整合中最常搭配的第二层组学
- 表观组学实践教程 — 甲基化数据的预处理和分析基础

参考资源

- MOFA2 官方教程
- mixOmics 文档
- SNFtool 说明
- MultiAssayExperiment Bioconductor
- TCGA GDC Portal
- CPTAC