BIOF3 组学数据分析

01 多组学项目设计与样本匹配

导出日期：2026年5月12日

01 多组学项目设计与样本匹配

多组学整合的第一步不是跑模型，而是确认"哪些样本在所有组学层都有数据"。样本 ID 不一致、批次不对齐、缺失模式不清楚——这些问题不解决，后面的分析全是空中楼阁。

样本 ID 的现实

不同组学平台对同一个样本的命名方式几乎不会一样。TCGA 用 barcode（`TCGA-A1-A0SK-01A-11R-A084-07`），蛋白组可能只保留前 12 位，甲基化数据用 Sentrix ID。你需要一张映射表把它们统一起来。

```
library(dplyr)

# 假设三张表各有自己的 ID 列
rna_samples   <- colnames(rna_mat)
meth_samples  <- colnames(meth_mat)
prot_samples  <- colnames(prot_mat)

# 用映射表统一
id_map <- read.csv("sample_id_mapping.csv")
# 列：sample_id, rna_id, meth_id, prot_id

# 找到三层都有的样本
common <- id_map %>%
  filter(rna_id %in% rna_samples,
         meth_id %in% meth_samples,
         prot_id %in% prot_samples)
cat("三层共有样本数:", nrow(common), "\n")
```

缺失模式可视化

在决定用哪些样本之前，先画一张缺失模式图。UpSet plot 比 Venn 图更适合三层以上的情况：

```
library(UpSetR)

sample_list <- list(

  RNA        = id_map$sample_id[!is.na(id_map$rna_id)],
  Methylation = id_map$sample_id[!is.na(id_map$meth_id)],
  Protein    = id_map$sample_id[!is.na(id_map$prot_id)]
)
upset(fromList(sample_list), order.by = "freq")
```

如果某一层缺失比例很高，需要考虑是否把它排除在整合之外，或者用支持缺失值的方法（如 MOFA2）。

MultiAssayExperiment 容器

Bioconductor 的 `MultiAssayExperiment`（MAE）是多组学数据的标准容器。它把多层数据、样本映射和临床信息绑在一起，后续分析函数可以直接操作。

```
library(MultiAssayExperiment)

# 构建 ExperimentList
exp_list <- ExperimentList(
  RNA   = SummarizedExperiment(assays = list(counts = rna_mat)),
  Meth  = SummarizedExperiment(assays = list(beta = meth_mat)),
  Prot  = SummarizedExperiment(assays = list(abundance = prot_mat))
)

# 样本映射表：每行说明某个 primary sample 在某层对应哪个 colname
smap <- listToMap(list(
  RNA  = data.frame(primary = common$sample_id, colname = common$rna_id),
  Meth = data.frame(primary = common$sample_id, colname = common$meth_id),
  Prot = data.frame(primary = common$sample_id, colname = common$prot_id)
))

# 临床信息
col_data <- DataFrame(row.names = common$sample_id,
                      subtype = common$subtype,
                      stage   = common$stage)

mae <- MultiAssayExperiment(experiments = exp_list,
                            colData     = col_data,
                            sampleMap   = smap)
mae
```

设计阶段的几个决策

1. **样本量**：多组学整合对样本量要求不低。MOFA2 官方建议至少 15 个样本；SNF 在 30 个以下效果不稳定。如果某一层只有 10 个样本，考虑是否值得纳入。
2. **批次效应**：不同组学的批次效应需要分别处理。RNA-seq 用 ComBat-seq，甲基化用 ComBat，蛋白组用 limma removeBatchEffect。不要在合并矩阵之后再做批次校正。
3. **配对 vs 非配对**：如果不是所有样本都有全部组学数据，需要决定是只用完全配对的子集，还是用能处理缺失的方法。MOFA2 支持部分缺失，SNF 不支持。

参考资源

- MultiAssayExperiment Bioconductor
- MultiAssayExperiment 使用手册
- UpSetR 包
- TCGA barcode 说明

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。