BIOF3 组学数据分析

蛋白质组学实践手册

BioF3 蛋白质组学专栏导出版

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

蛋白质组学实践手册目录

BioF3 蛋白质组学专栏导出版

01 蛋白质组学实践教程

转录组告诉你哪些基因被转录，蛋白质组告诉你哪些蛋白真的被翻译、它们的丰度如何、以及是否被修饰。质谱是目前最常用的蛋白组测量手段：样本酶解后肽段上机，仪器给出质荷比和强度，数据库比对后得到蛋白身份和定量值。

做过转录组分析的读者上手蛋白组往往会遇到两个反差：数据维度小（几千个蛋白 vs 几万个基因）、缺失值多（某些蛋白在部分样本根本没测到）。因此思路上 bulk RNA-seq 的很多经验可以复用，但预处理和统计模型要调整。

本专栏假设原始质谱数据已经由 MaxQuant、FragPipe、DIA-NN、Spectronaut 等工具处理好，起点是一份带强度或 LFQ 值的蛋白表。

一个项目大致是什么样子

给你 20 个样本的 MaxQuant `proteinGroups.txt` 或 DIA-NN `report.pg_matrix.tsv`，从这里到一张能写进论文的差异蛋白结果，大致要走：

步骤	典型产物	常用工具
结果表清洗	去除污染、反向库、仅 1 肽识别的蛋白	R，dplyr
缺失值评估	每个样本缺失率分布	naniar、VIM
缺失值插补	完整的强度矩阵	MinProb、KNN、DEP
归一化	按中位数 / 中位线对齐	DEP、limma 的 normalizeBetweenArrays
批次效应检查	PCA、样本相关性	ggplot2、ComplexHeatmap
差异蛋白分析	差异蛋白表	limma、DEP、MSstats
功能富集	GO、Reactome、pathway 结果	clusterProfiler、ReactomePA
相互作用与网络	蛋白-蛋白互作图	STRING、Cytoscape

MaxQuant 和 FragPipe 是 DDA 的两条主流选择，DIA-NN 和 Spectronaut 面向 DIA 实验。处理好后的结果表在字段名上有差异，但后续的差异分析和富集分析可以共用。

常见工具栈

差异分析这一段，`limma` 仍然是最稳的选择，因为它对样本量小、噪音大、存在技术重复的实验设计非常友好。`DEP` 在 limma 之上包了一层适合蛋白组的数据流程，适合初学者。

阶段	工具	运行环境
原始数据处理	MaxQuant、FragPipe、DIA-NN、Spectronaut	GUI / bash
数据读取	MSnbase、QFeatures、DEP	R
缺失值处理	DEP、imputeLCMD	R
差异分析	limma、DEP、MSstats	R
功能注释	clusterProfiler、ReactomePA	R
蛋白互作	STRING 网页、STRINGdb、Cytoscape	R / 网页
可视化	ggplot2、ComplexHeatmap、EnhancedVolcano	R

推荐公开数据集

蛋白组的公开数据比转录组少，但以下几个入口能覆盖大多数教学需求：

数据集	类型	适合	入口
PRIDE 数据集	各种质谱项目	找到文献配套原始/结果	PRIDE
MassIVE	各种质谱项目	北美研究为主	MassIVE
DEP 包 `UbiLength`	8 样本泛素化实验	差异分析最小入门数据	DEP Bioconductor
CPTAC	癌症蛋白组 + 表型	多组学整合练习	CPTAC

`DEP` 包自带的 `UbiLength` 数据是 HeLa 细胞不同处理时间的泛素化组实验，8 个样本、大约 2700 个蛋白，规模合适教学，下面的最小例子也基于它。

最小可跑的例子

这段代码基于 `DEP` 包自带的真实数据。安装好包后能直接运行完成："读入 → 过滤 → 归一化 → 插补 → 差异分析 → 可视化" 一条链路：

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

library(DEP)

# 自带真实数据：UbiLength MaxQuant 结果
data <- UbiLength
experimental_design <- UbiLength_ExpDesign

# 清洗：去除污染蛋白、反向库匹配、仅 1 肽鉴定的蛋白
data_unique <- make_unique(data, "Gene.names", "Protein.IDs", delim = ";")
data_se <- make_se(data_unique, grep("LFQ.", colnames(data_unique)), experimental_design)

# 过滤：每组至少两个样本里被测到
data_filt <- filter_missval(data_se, thr = 0)

# 归一化 + 缺失值插补
data_norm <- normalize_vsn(data_filt)
data_imp  <- impute(data_norm, fun = "MinProb", q = 0.01)

# 差异分析（limma 后端）
data_diff <- test_diff(data_imp, type = "control", control = "Ctrl")
dep <- add_rejections(data_diff, alpha = 0.05, lfc = 1)

# 火山图
plot_volcano(dep, contrast = "Ubi6_vs_Ctrl", label_size = 2, add_names = TRUE)
```

跑完后得到的是一张火山图，横轴是 log2 fold change，纵轴是 `-log10(p)`。被标注出来的蛋白就是这组对比下显著差异的候选，可以再去做富集分析或文献查证。

真实项目里有几件事比这段复杂：样本分组更多、协变量要纳入模型、DIA 数据的插补策略不同、要输出 MSstats 报告等等，但起步的框架就是这样。

专栏模块规划

模块	主题	状态
01	质谱结果表格式与实验设计	已上线
02	DEP 差异蛋白分析	已上线
03	功能富集与 Reactome 通路	已上线
04	可视化：火山图、热图与蛋白互作	已上线
05	limma 直接建模（不用 DEP 包装）	规划中
06	DIA 数据的处理差异	规划中
07	STRING 与蛋白互作网络	规划中
08	转录组与蛋白组联合分析	规划中

目前 01-04 已上线，每个模块带一份可直接跑的 R 脚本。

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
- bulk RNA-seq 实践教程（差异分析思路有很多共通点）

参考资源

- DEP Bioconductor 文档
- limma 用户手册
- MaxQuant 官方文档
- FragPipe
- DIA-NN
- PRIDE Archive
- STRING 数据库

02 质谱结果表与实验设计

这一章的目标是让读者在动手做差异蛋白分析之前，先搞清楚"手上拿到的是什么格式的表、每一列代表什么、实验设计怎么对应到分析里"。

质谱数据处理的起点

蛋白质组学的原始数据是质谱仪产出的 `.raw` / `.mzML` 文件。这些文件需要经过数据库搜索（DDA）或者谱图库匹配（DIA）才能得到蛋白鉴定和定量结果。常用的处理工具：

工具	适用	输出
MaxQuant	DDA	`proteinGroups.txt` 、 `evidence.txt`
FragPipe	DDA	`combined_protein.tsv`
DIA-NN	DIA	`report.pg_matrix.tsv`
Spectronaut	DIA	自定义导出表

BioF3 的蛋白组教程**从这些工具的输出表开始**，不涉及原始质谱数据处理。

MaxQuant proteinGroups.txt 的关键列

MaxQuant 是目前最常用的 DDA 处理工具。它的 `proteinGroups.txt` 是一张宽表，每行一个蛋白组（protein group），关键列：

列名	含义
`Protein IDs`	UniProt accession，多个用分号分隔
`Gene names`	基因名（用于可视化和富集）
`LFQ intensity XXX`	每个样本的 Label-Free Quantification 强度
`iBAQ XXX`	另一种定量方式（绝对丰度估计）
`Razor + unique peptides`	用于该蛋白定量的肽段数
`Reverse`	是否匹配到反向库（应过滤掉）
`Potential contaminant`	是否是常见污染物（应过滤掉）
`Only identified by site`	是否仅通过修饰位点鉴定（通常过滤掉）

读入后第一件事：过滤掉 `Reverse == "+"` 和 `Potential contaminant == "+"` 的行。这些不是真正的样本蛋白。

实验设计表

和 bulk RNA-seq 一样，蛋白组分析也需要一张样本表把"哪个列对应哪个条件"说清楚。DEP 包要求的格式：

```
label      condition  replicate
Sample1    Control    1
Sample2    Control    2
Sample3    Control    3
Sample4    Treatment  1
Sample5    Treatment  2
Sample6    Treatment  3
```

- `label` 要和 `proteinGroups.txt` 里 `LFQ intensity` 列名的后缀对得上
- `condition` 是分组变量（后面做差异分析的对比就基于它）
- `replicate` 是生物学重复编号

缺失值：蛋白组的核心难题

和转录组最大的区别：蛋白组的缺失值非常多。一个蛋白在某些样本里完全没有信号（LFQ = 0 或 NA），原因可能是：

- **MNAR（Missing Not At Random）**：蛋白丰度太低，低于检测限 → 这种缺失和真实丰度有关
- **MCAR（Missing Completely At Random）**：随机的技术波动 → 和丰度无关

两种缺失的处理策略不同：MNAR 通常用"从分布左尾采样"（MinProb）来插补；MCAR 可以用 KNN 或均值插补。DEP 默认用 MinProb，适合大多数蛋白组场景。

从 proteinGroups.txt 到 DEP 对象

DEP 包把上面这些步骤封装成几个函数：

```r
library(DEP)

# 1. 读入 MaxQuant 结果
data <- read.delim("proteinGroups.txt")

# 2. 过滤污染和反向库
data <- data[data$Reverse != "+", ]
data <- data[data$Potential.contaminant != "+", ]

# 3. 确保蛋白名唯一
data_unique <- make_unique(data, "Gene.names", "Protein.IDs", delim = ";")

# 4. 构建 SummarizedExperiment
lfq_cols <- grep("LFQ.intensity.", colnames(data_unique))
data_se <- make_se(data_unique, lfq_cols, experimental_design)
```

`make_se` 之后得到的是一个标准的 Bioconductor `SummarizedExperiment` 对象，后续的过滤、归一化、插补、差异分析都在这个对象上操作。

下一步

- 02 DEP 差异蛋白分析
- 03 功能富集与 Reactome 通路

参考资源

- MaxQuant 官方文档
- DEP Bioconductor 文档
- Perseus 教程（MaxQuant 配套可视化）
- DIA-NN 文档

03 DEP 差异蛋白分析

DEP（Differential Enrichment analysis of Proteomics data）把蛋白组差异分析的完整流程封装成一条管线：过滤 → 归一化 → 插补 → limma 差异检验 → 多重校正。底层用的是 limma，所以统计学上和 bulk RNA-seq 的差异分析是同一套框架。

本章用 DEP 自带的 `UbiLength` 数据走一遍完整流程。数据是 HeLa 细胞在不同泛素链长度处理下的蛋白组（4 个条件 × 3 个重复 = 12 个样本，约 3000 个蛋白）。

完整流程

```
library(DEP)

data(UbiLength)
data(UbiLength_ExpDesign)

# 1. 确保蛋白名唯一
data_unique <- make_unique(UbiLength, "Gene.names", "Protein.IDs", delim = ";")

# 2. 构建 SummarizedExperiment
lfq_cols <- grep("LFQ.intensity.", colnames(data_unique))
data_se <- make_se(data_unique, lfq_cols, UbiLength_ExpDesign)

# 3. 过滤：每组至少 2 个样本有值
data_filt <- filter_missval(data_se, thr = 0)

# 4. VSN 归一化
data_norm <- normalize_vsn(data_filt)

# 5. 缺失值插补（MinProb：从左尾采样）
data_imp <- impute(data_norm, fun = "MinProb", q = 0.01)

# 6. limma 差异检验
data_diff <- test_diff(data_imp, type = "control", control = "Ctrl")

# 7. 标记显著蛋白
dep <- add_rejections(data_diff, alpha = 0.05, lfc = 1)
```

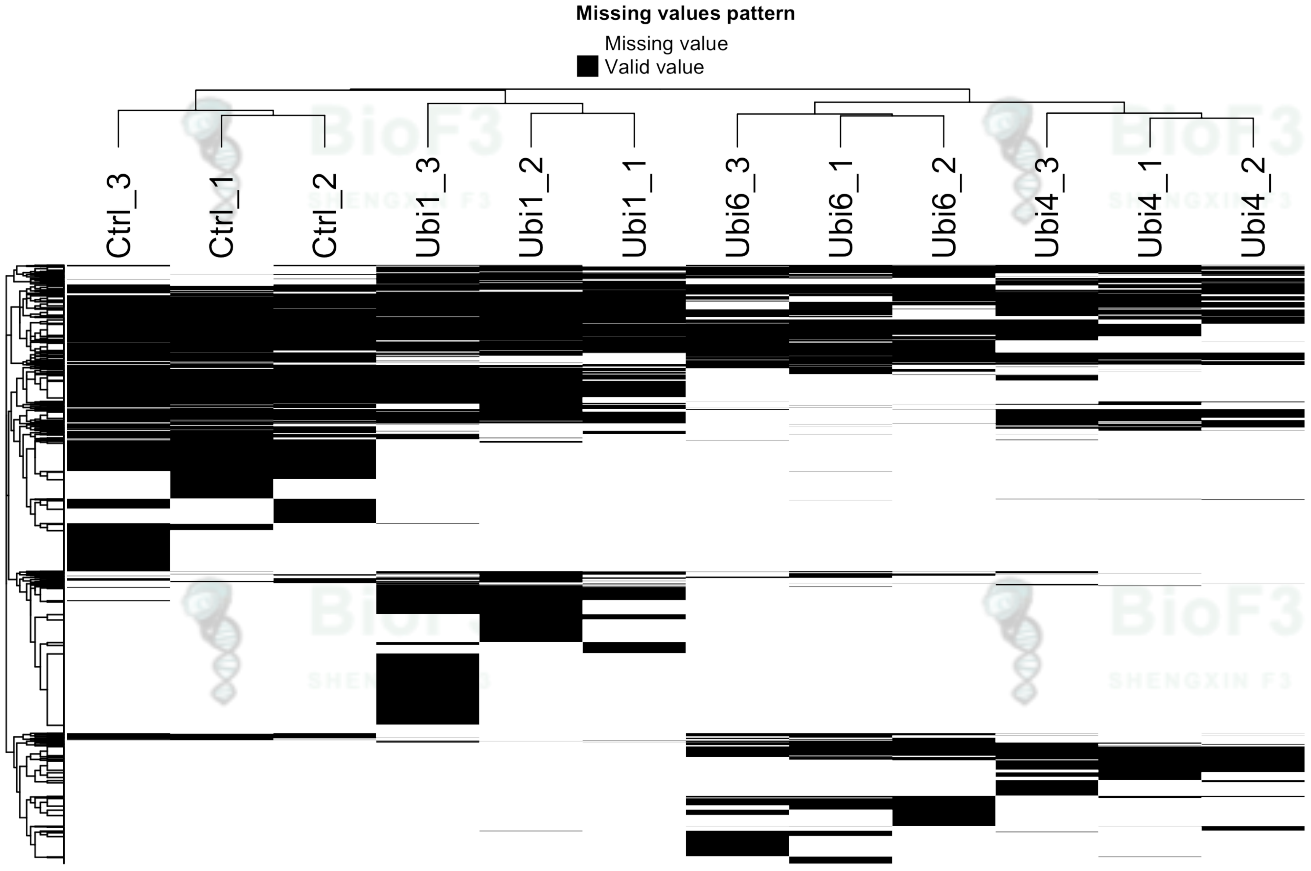
`test_diff(type = "control", control = "Ctrl")` 会自动生成所有"XX vs Ctrl"的对比。如果要做两两比较，用 `type = "all"` 。

真实示例：UbiLength 上的 DEP 分析

配套脚本 `prot02_dep_sci.R` 把上面的流程完整跑了一遍，输出 6 张图：

```
Rscript scripts/proteomics/prot02_dep_sci.R
```

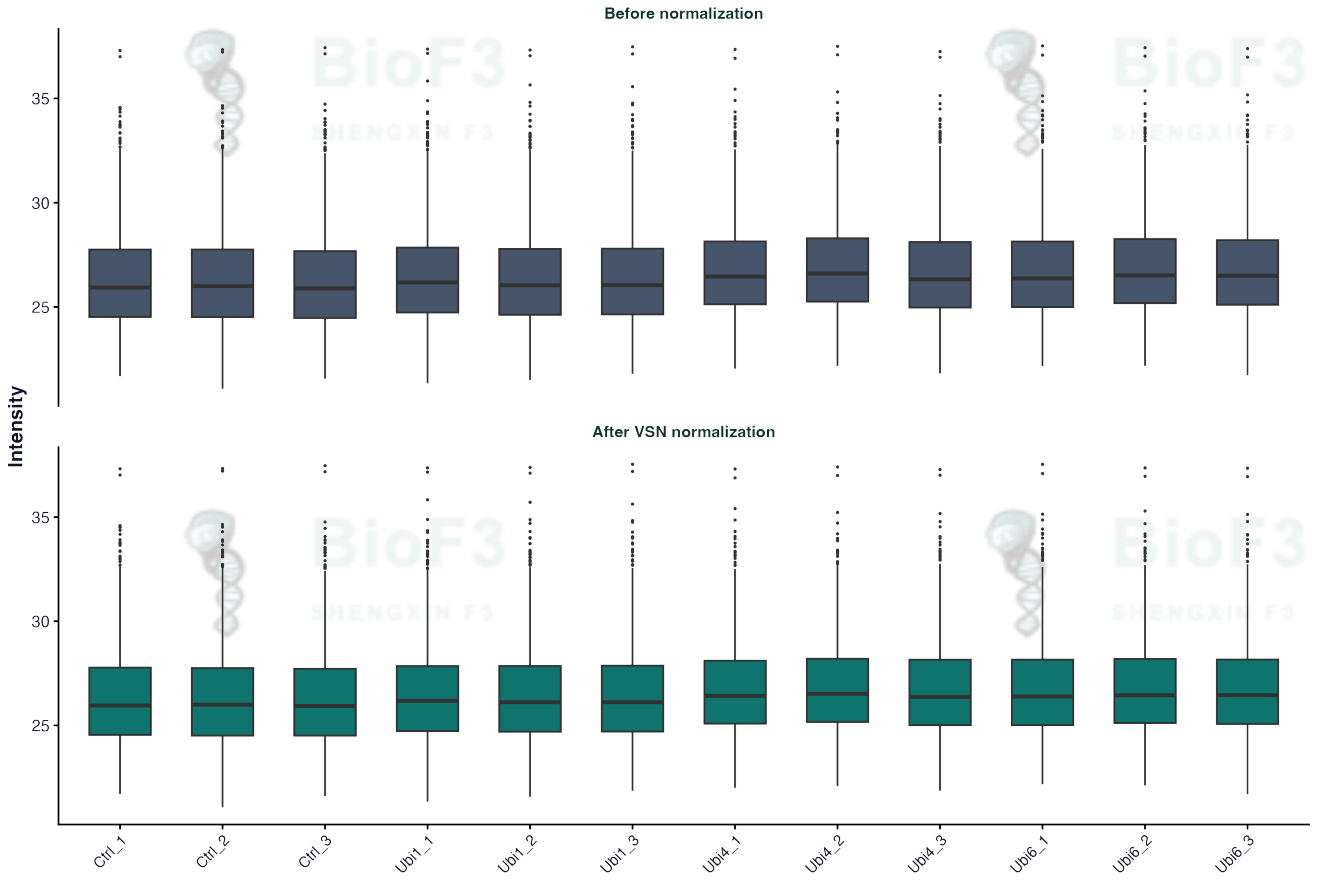
图 1：缺失值热图



每行是一个蛋白，每列是一个样本。黑色 = 有值，灰色 = 缺失。蛋白组的缺失不是随机的 —— 低丰度蛋白在所有样本里都容易缺失（整行灰色），这就是 MNAR 的典型模式。

这张图在 QC 阶段最重要的用途：**看有没有某个样本缺失率异常高**（整列灰色比别的多很多）。如果有，要考虑是不是该样本的质谱跑坏了。

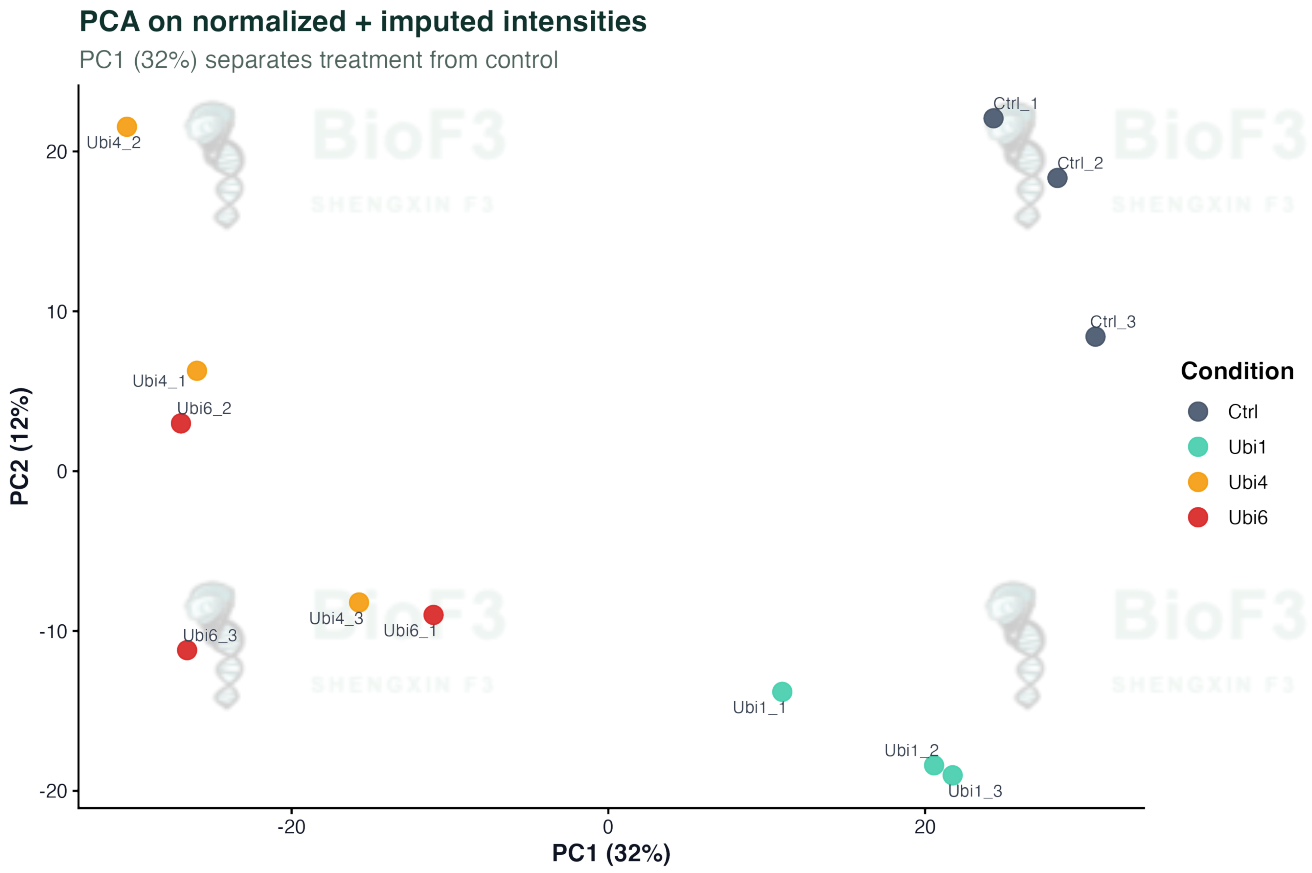
图 2：归一化前后的强度分布



上面是归一化前，下面是 VSN 归一化后。归一化的目标是让所有样本的中位线对齐、方差稳定。如果归一化前某些样本的 box 明显偏高或偏低，说明上样量或质谱响应有系统偏差。

VSN（Variance Stabilizing Normalization）和 bulk RNA-seq 里的 `vst` 思路一样：让高强度和低强度蛋白的方差都差不多，后续 limma 的 t 检验才不会被少数高丰度蛋白主导。

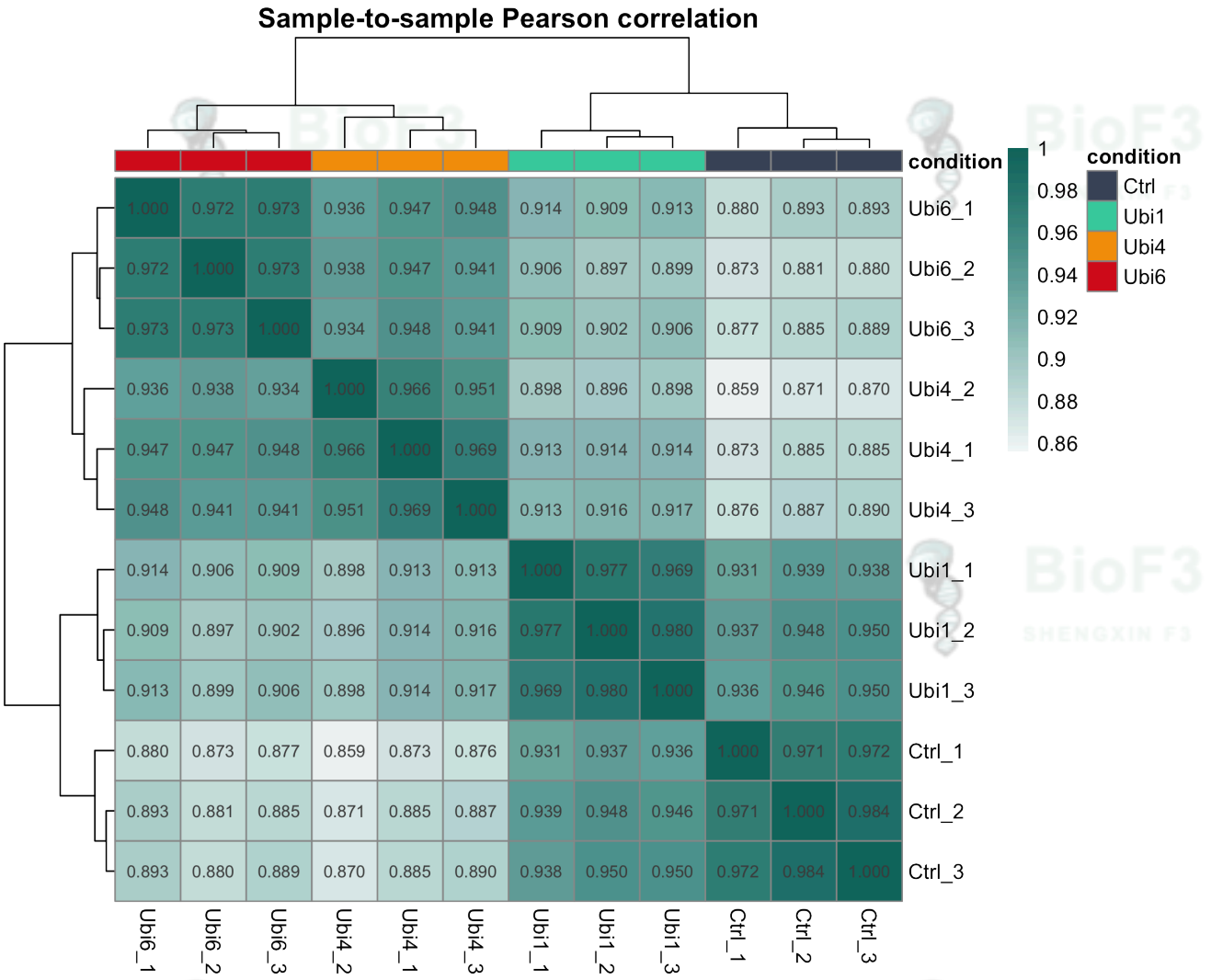
图 3：PCA



归一化 + 插补之后做 PCA。颜色是条件（Ctrl / Ubi1 / Ubi4 / Ubi6）。PC1 应该把处理组和对照组分开。如果同一条件的重复散得很开，说明技术变异大或者某个重复有问题。

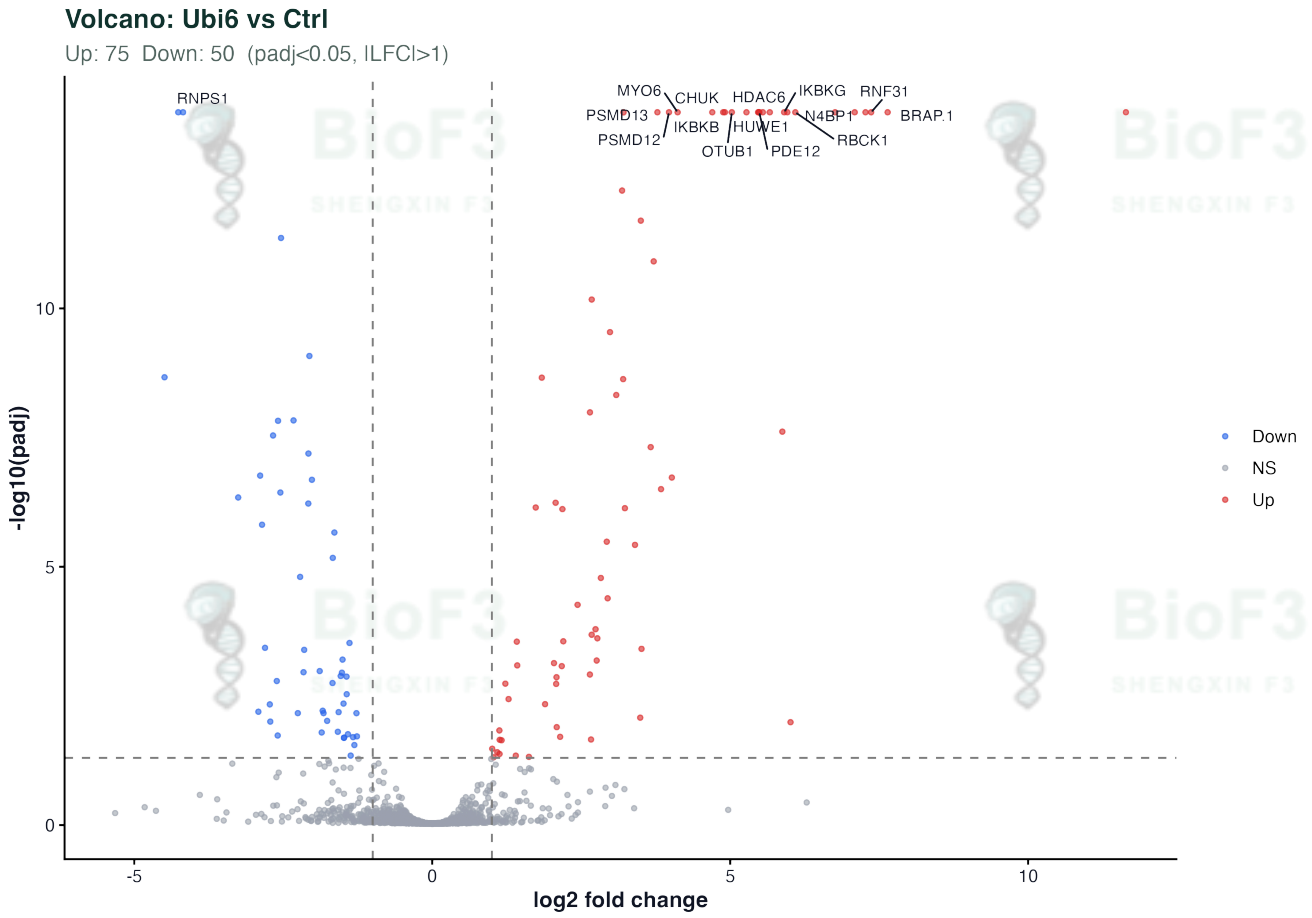
UbiLength 里 Ubi6（最长泛素链）和 Ctrl 分得最远，Ubi1 和 Ctrl 最近 —— 符合生物学预期：泛素链越长，蛋白组变化越大。

图 4：样本相关性热图



样本两两之间的 Pearson 相关。同一条件的样本之间相关性应该最高（对角线附近的深色块）。如果某个样本和同组的相关性反而低于和别组的，要回头看 QC。

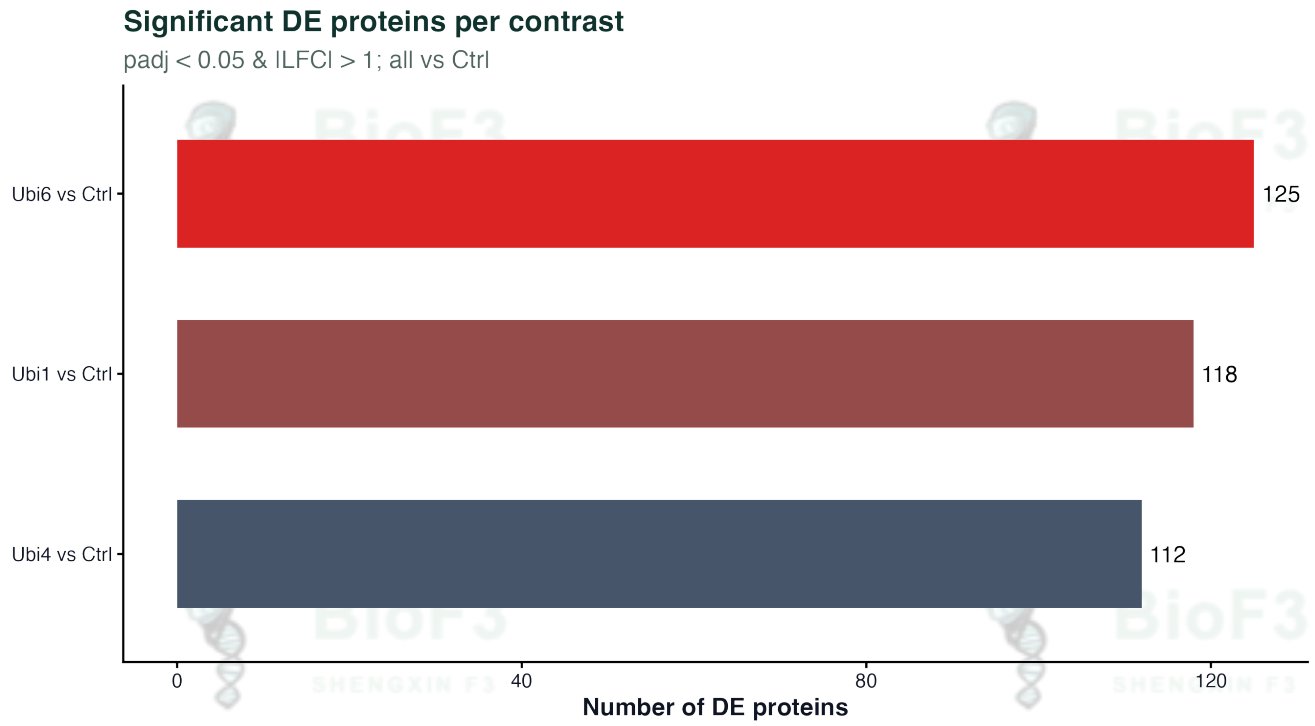
图 5：火山图（Ubi6 vs Ctrl）



横轴 log2 fold change，纵轴 `-log10(padj)`。红色 = 上调（Ubi6 里更高），蓝色 = 下调。标注了 padj 最小的 15 个蛋白。

和 RNA-seq 火山图的读法完全一样。蛋白组的特点是：点数少（几千 vs 几万）、效应大小通常比转录组小（蛋白翻译后调控会缓冲 mRNA 变化）。

图 6：每个对比的显著蛋白数



三个对比（Ubi1/Ubi4/Ubi6 vs Ctrl）各自有多少显著差异蛋白。Ubi6 最多、Ubi1 最少，和 PCA 的分离程度一致。

套到自己数据上

脚本的前 10 行把 `UbiLength` 换成自己的 `proteinGroups.txt` 就行：

```
data <- read.delim("proteinGroups.txt")
data <- data[data$Reverse != "+", ]
data <- data[data$Potential.contaminant != "+", ]
data_unique <- make_unique(data, "Gene.names", "Protein.IDs", delim = ";")
```

`UbiLength_ExpDesign` 换成自己的样本表（label / condition / replicate 三列）。

几个常见调整：

- **DIA 数据**：DIA-NN 输出的 `report.pg_matrix.tsv` 已经是宽表，列名就是样本名，直接 `make_se` 即可
- **插补策略**：如果缺失率 > 50%，MinProb 可能不够好，试 `fun = "knn"` 或 `fun = "mixed"`
- **多因子设计**：`test_diff(type = "manual", test = ...)` 可以传自定义 contrast

下载资源

prot02_dep_sci.R
8 KB

下载 DEP 差异蛋白完整脚本 ↗

下一步

- 03 功能富集与 Reactome 通路
- 04 可视化与蛋白互作网络

参考资源

- DEP Bioconductor 文档
- limma 用户手册
- Zhang et al. 2017, UbiLength 原始研究

04 功能富集与 Reactome 通路

差异蛋白列表本身只是中间产物。真正有用的是"这些蛋白对应哪些通路"。本章用 02 DEP 差异分析 的结果，走一遍 GO + Reactome + GSEA。

思路和 bulk RNA-seq 03 富集 完全一致，只是输入从基因换成了蛋白（通过 gene symbol 映射到 Entrez ID）。

真实示例

配套脚本 `prot03_enrichment_sci.R` 在 UbiLength 的 Ubi6 vs Ctrl 差异蛋白上跑 GO BP、Reactome、GSEA：

```
Rscript scripts/proteomics/prot03_enrichment_sci.R
```

每张图看什么

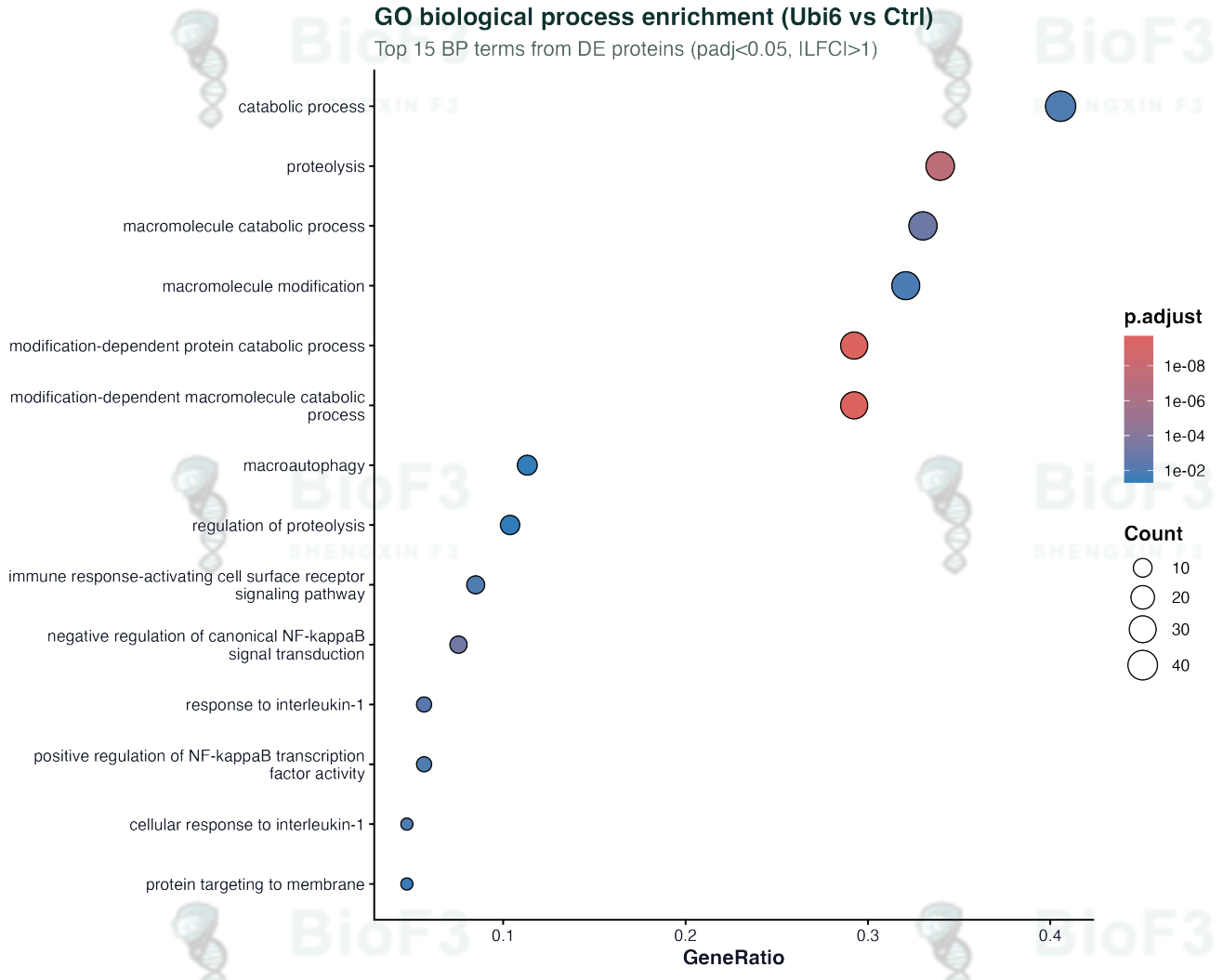


图 1：GO biological process 富集 dotplot。泛素化相关通路（蛋白降解、泛素连接酶活性）排在最前面，符合实验设计。

图 2：Reactome 通路富集。Reactome 的通路粒度比 GO 更细，适合定位到具体的信号级联。

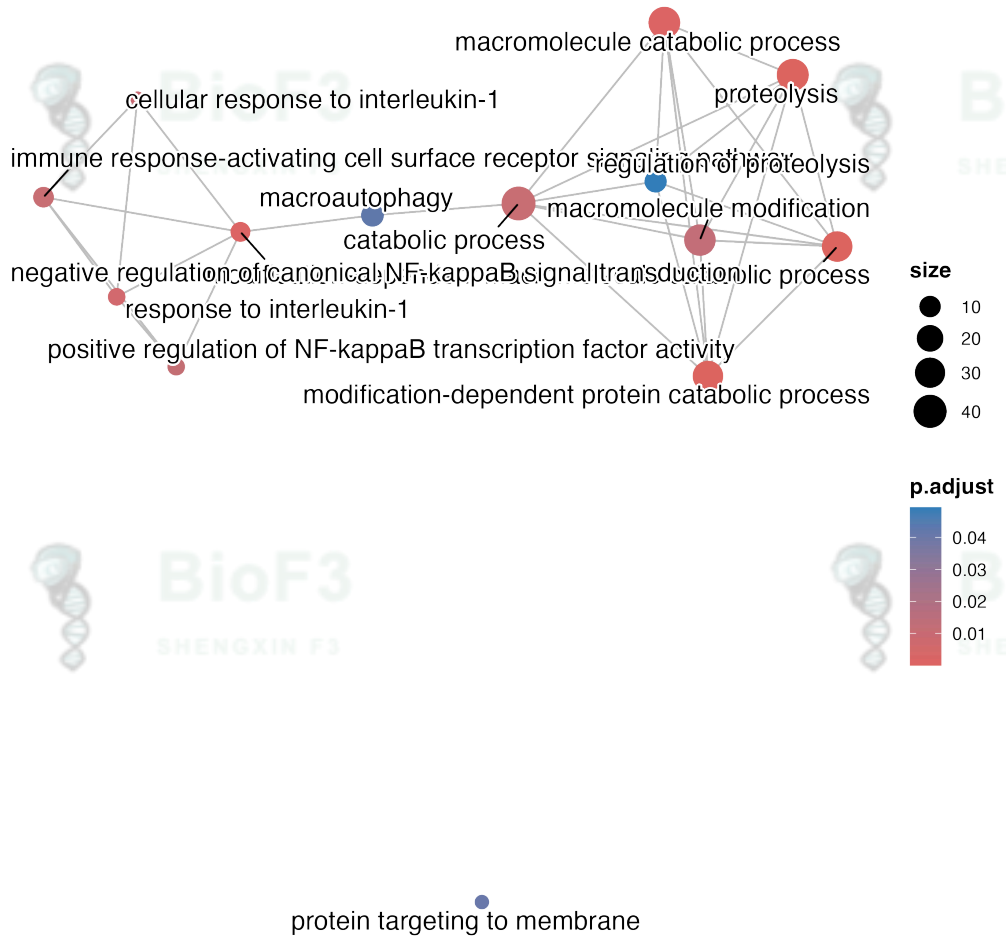


图 3：GO term 相似度网络。共享蛋白的 term 连在一起，帮助去冗余。

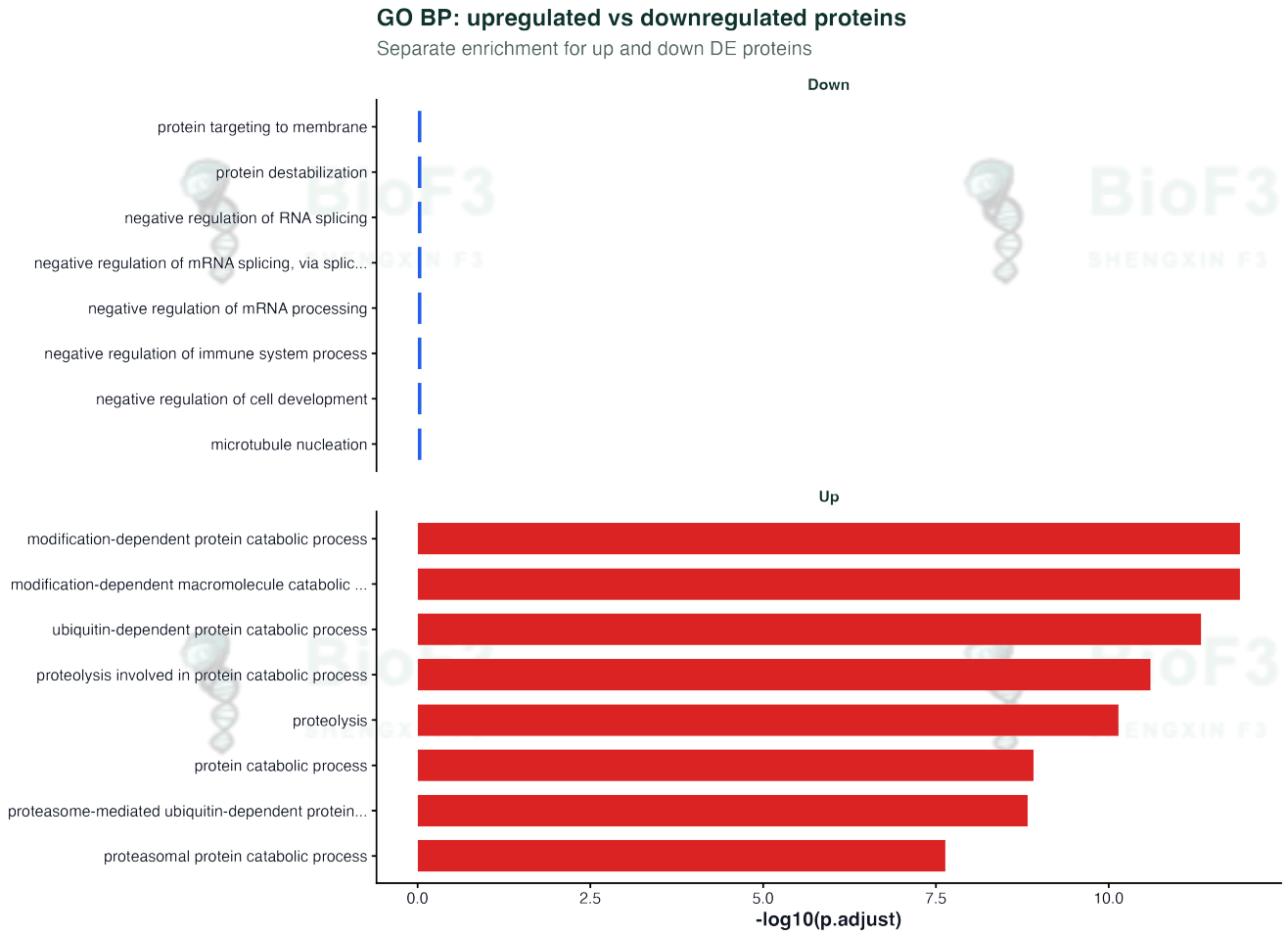


图 4：分别对上调和下调蛋白做 GO 富集。上调蛋白富集在泛素化相关通路，下调蛋白富集在代谢或稳态维持通路。

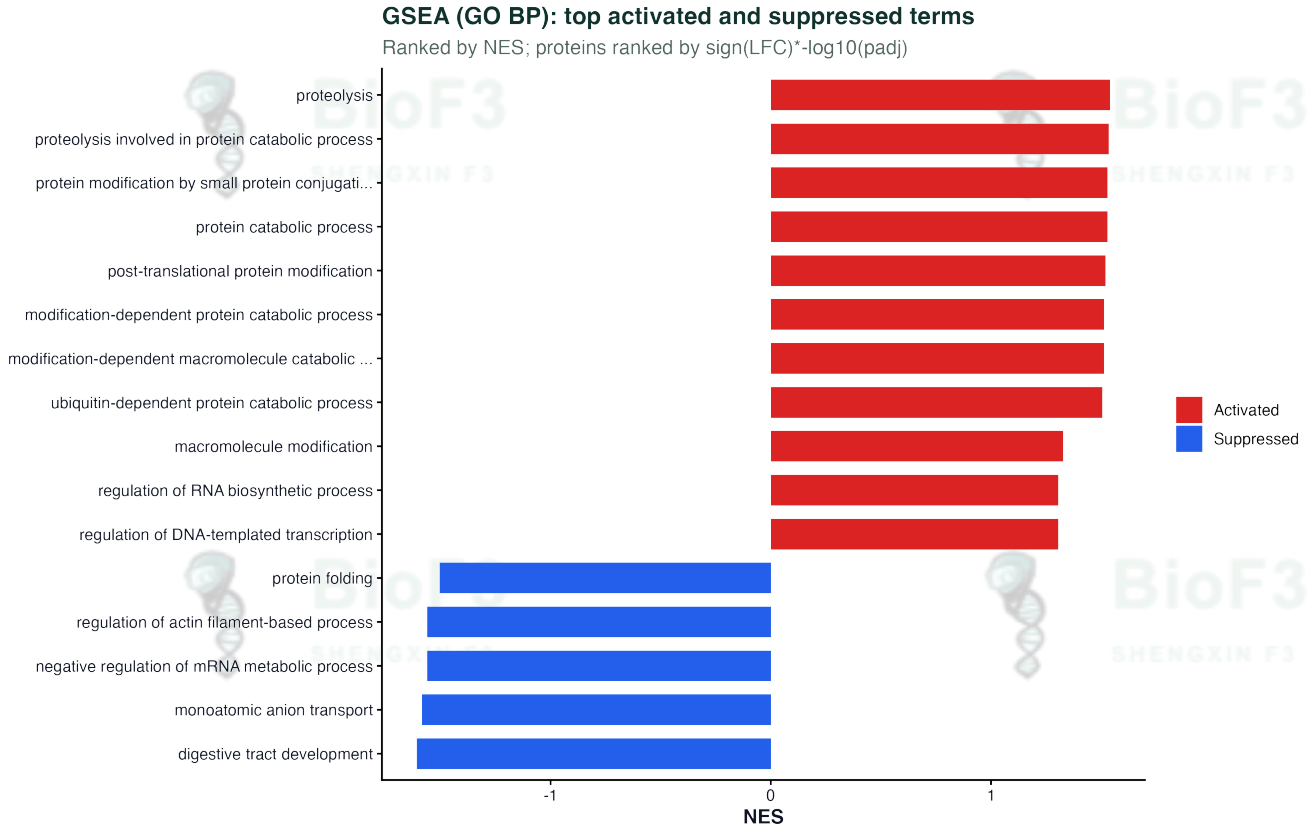


图 5：GSEA 瀑布图。不依赖阈值，直接看整条通路的蛋白是否整体偏向一个方向。

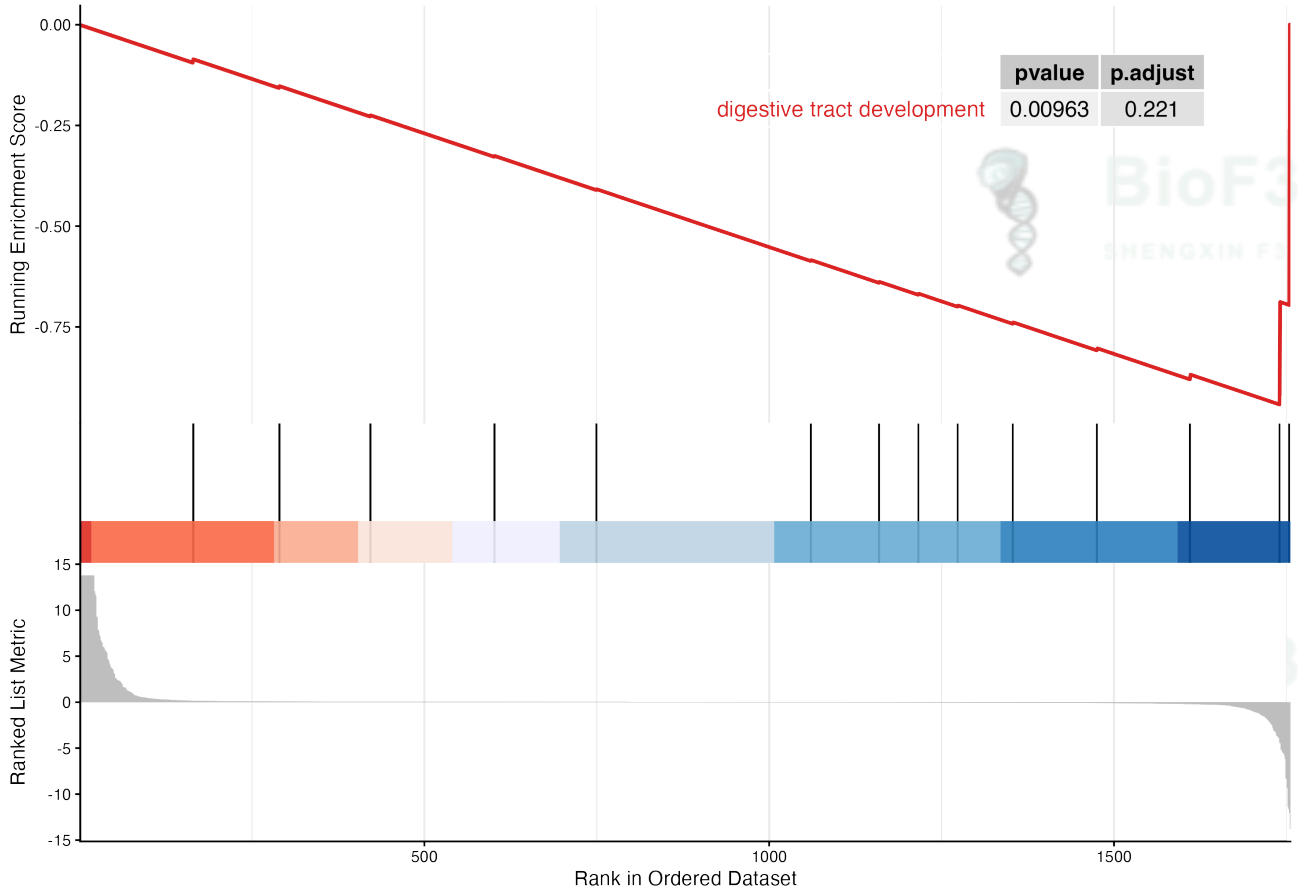


图 6：最强 NES 的 term 的 running-score 图。

下载资源

prot03_enrichment_sci.R
8 KB

下载蛋白组富集分析完整脚本 ↗

参考资源

- clusterProfiler 教程
- ReactomePA
- Reactome 数据库

05 可视化：火山图、热图与蛋白互作

本章把 DEP 差异分析的结果变成能直接用于论文的图。每张图配一个最常见的用途。

配套脚本 `prot04_visualization_sci.R` 在 UbiLength 数据上输出 6 张可视化成品：

```
Rscript scripts/proteomics/prot04_visualization_sci.R
```

每张图看什么

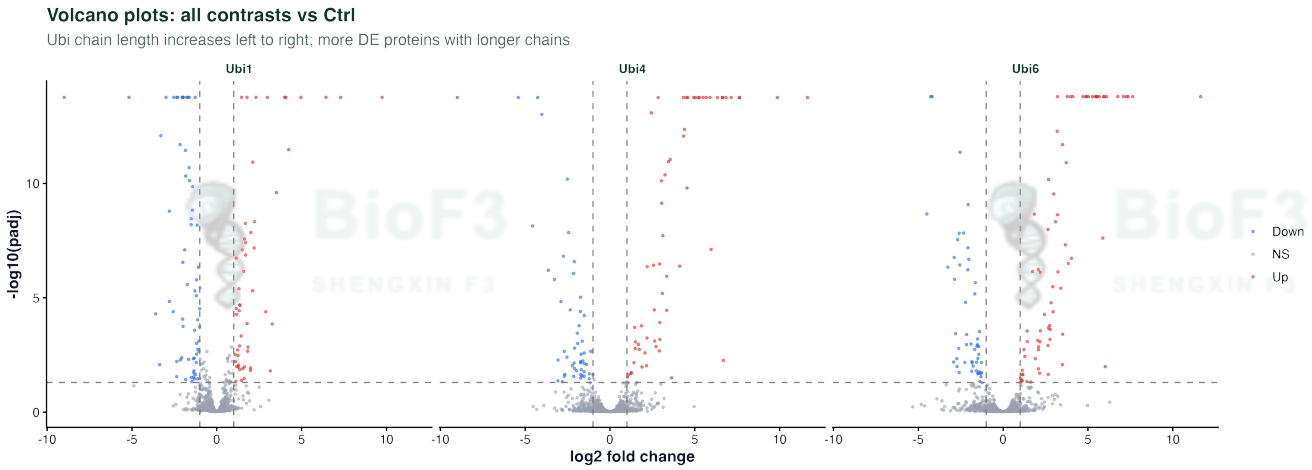


图 1：三个对比（Ubi1/Ubi4/Ubi6 vs Ctrl）的火山图并排。泛素链越长，显著蛋白越多、fold change 越大。一张图同时展示剂量-效应关系。

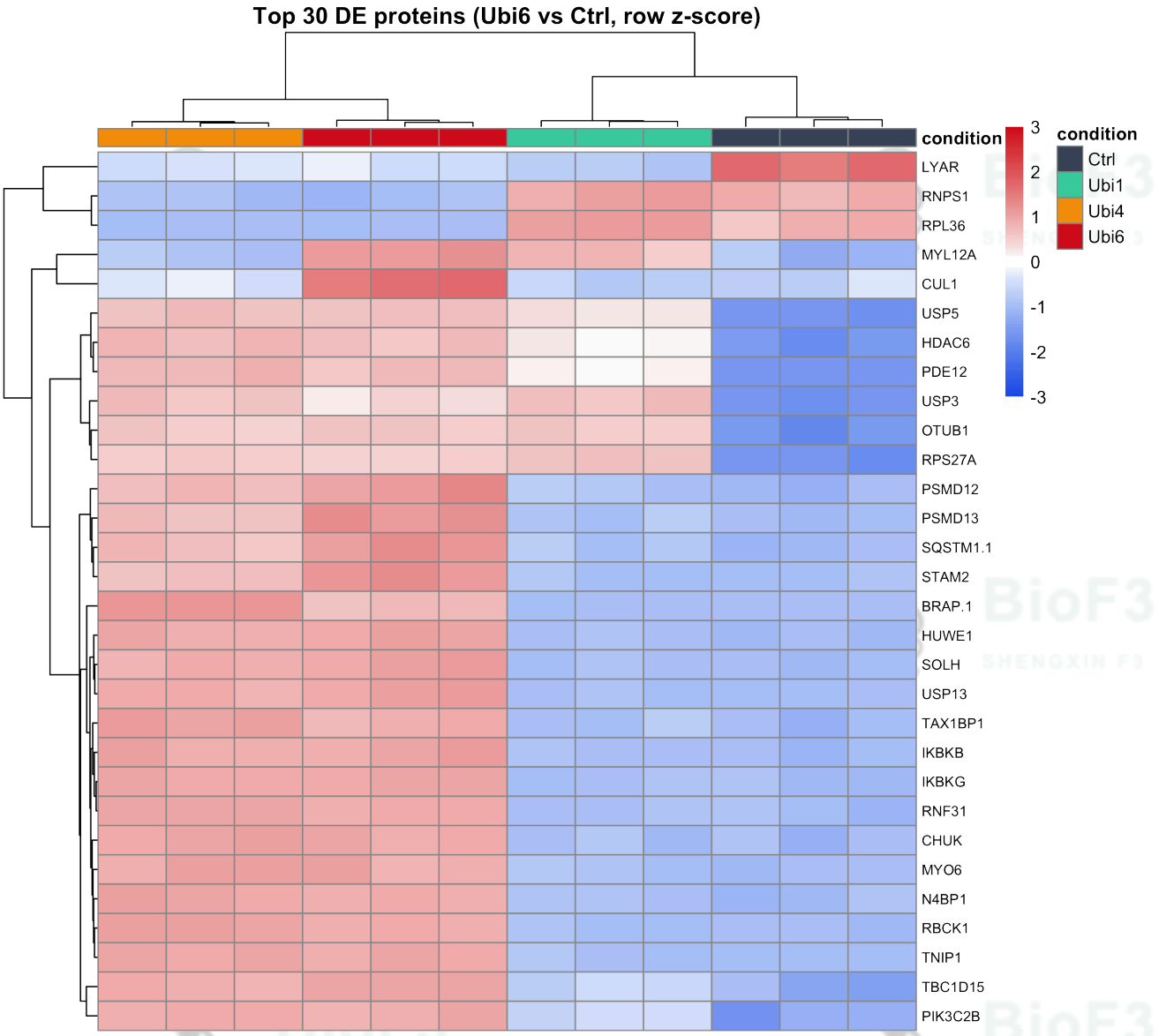


图 2：Ubi6 vs Ctrl 最显著的 30 个蛋白的 z-score 热图。列按条件排序，行聚类。可以直接看到哪些蛋白在 Ubi6 里一致上调 / 下调。

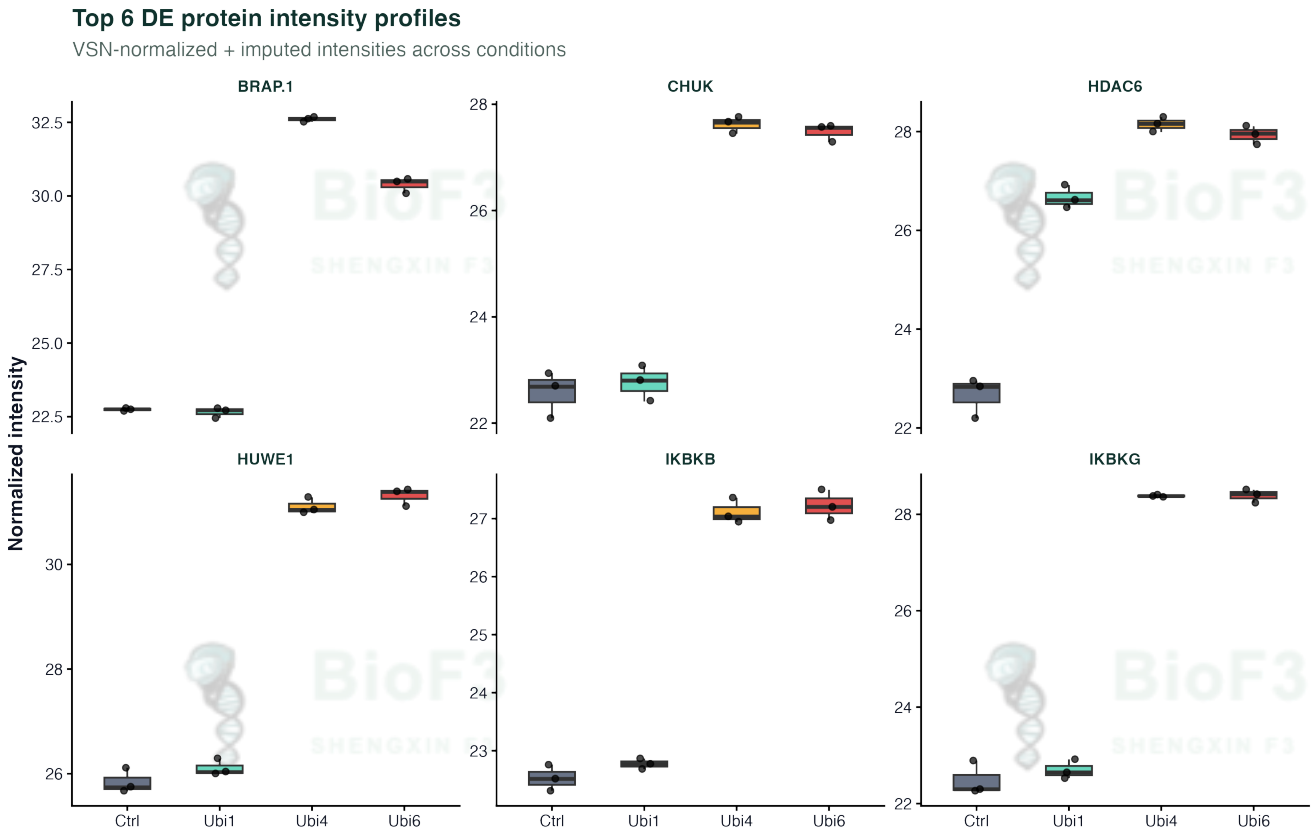


图 3：Top 6 差异蛋白在 4 个条件下的归一化强度分布。boxplot + jitter 能看到每个重复的值，确认差异不是被个别 outlier 驱动。

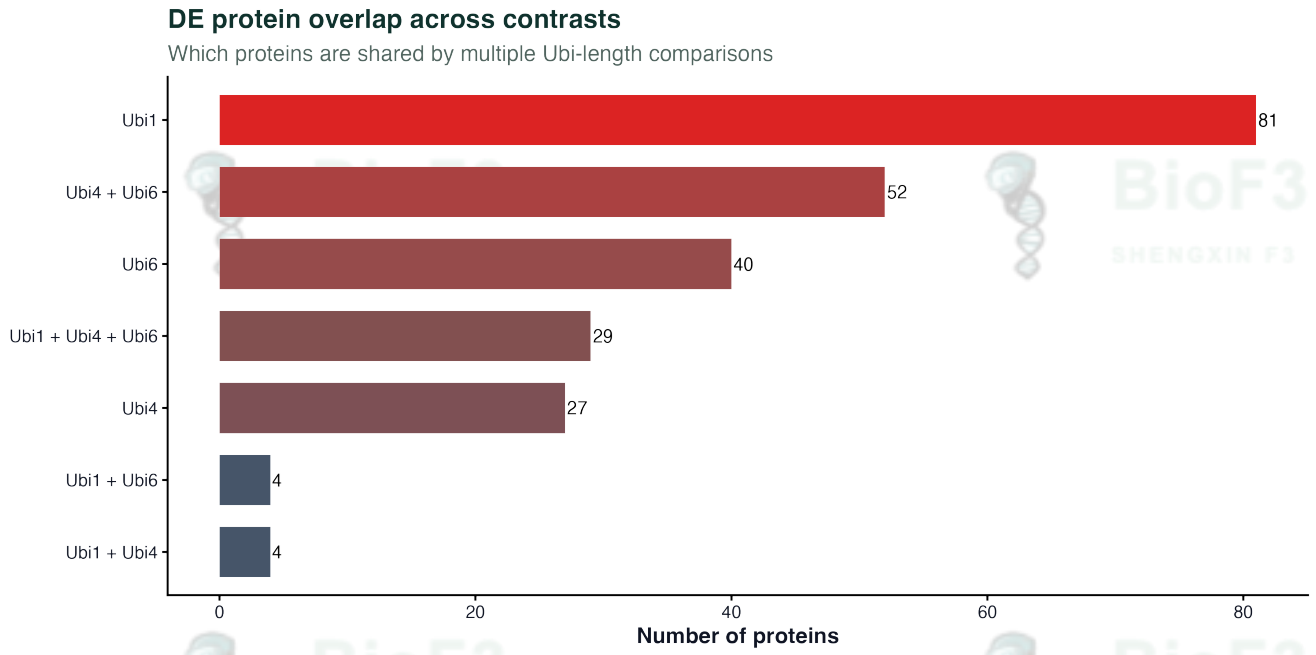


图 4：三个对比的 DE 蛋白集合重叠。"Ubi1 + Ubi4 + Ubi6" 共有的蛋白是最稳健的候选；只在 Ubi6 里出现的可能是链长度特异的效应。

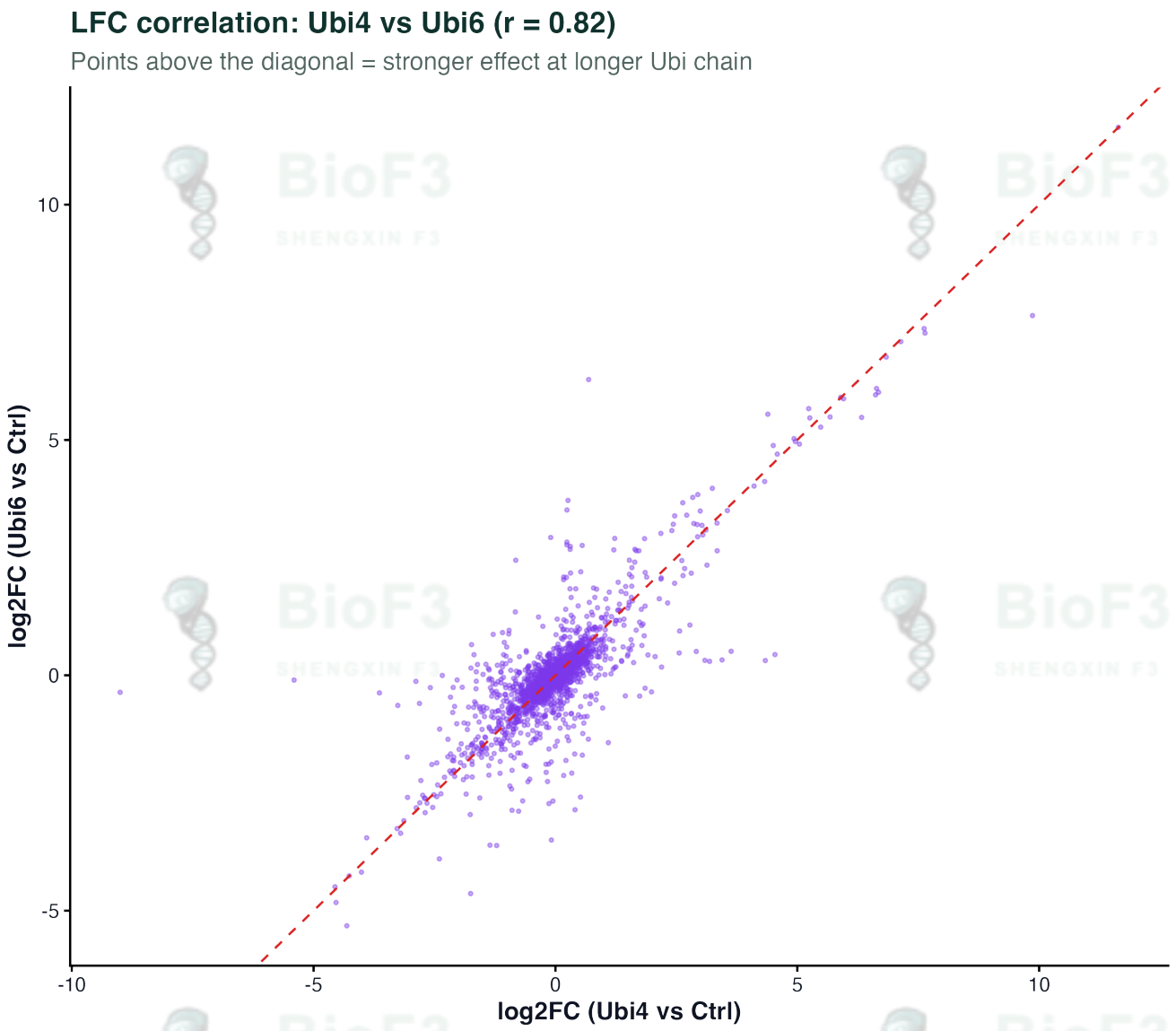


图 5：Ubi4 vs Ubi6 的 log2FC 散点。对角线上方的点 = Ubi6 效应更强。高相关性说明两个条件的蛋白组变化方向一致，只是幅度不同。

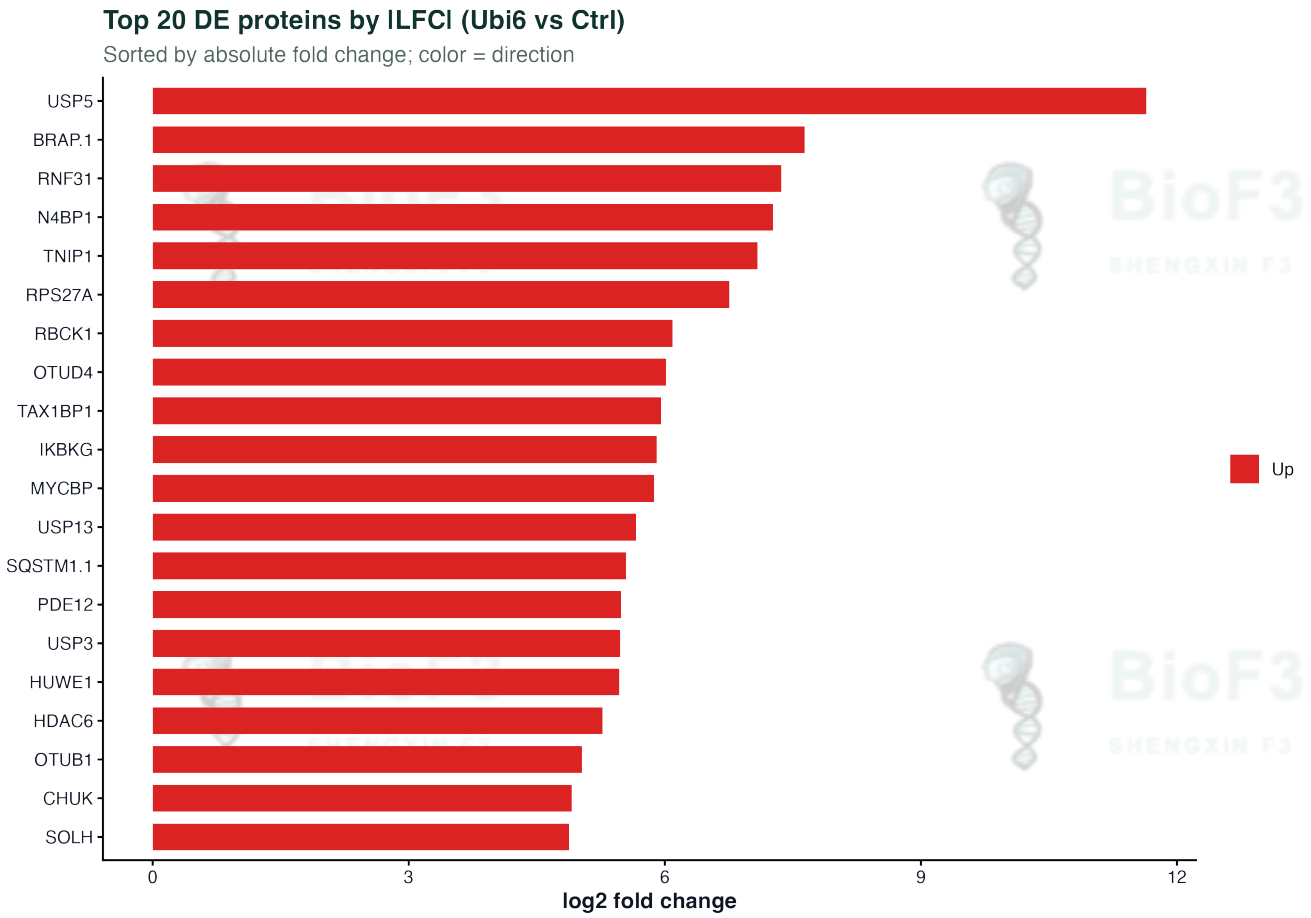


图 6：按 |LFC| 排序的 top 20 差异蛋白条形图。红色上调、蓝色下调。写论文时直接用来列出"效应最大的候选蛋白"。

下载资源

prot04_visualization_sci.R
8 KB

下载蛋白组可视化完整脚本 ↗

参考资源

- pheatmap
- ComplexHeatmap
- ggrepel
- STRING 数据库