

蛋白质组学数据分析

一、实验目的

通过蛋白质组学技术对JRC48耐药性人膀胱移行细胞癌T24-JC48细胞株(低/中/高耐药组)与T24细胞株(对照组)进行全蛋白表达谱分析筛选显著差异表达蛋白,揭示耐药相关功能通路及关键蛋白标志物。

二、实验流程

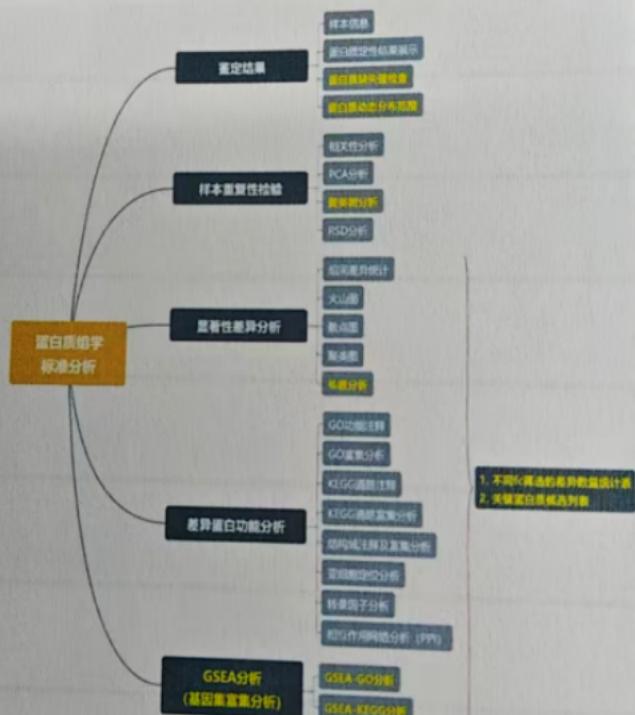


图 1 蛋白质组学信息分析技术流程

2.1 样本制备

- 1) 细胞裂解提取总蛋白, BCA法定量, SDS-PAGE验证蛋白完整性。
- 2) 酶切段肽后质控(肽段量 $\geq 200\text{ng}$), 采用Tims TOF Pro质谱仪进行PDA检测。

2.2 质谱数据采集和分析

- 1) 质谱原始数据经 Max Quant 软件进行数据库搜索 (Homo sapiens UniProt)
- 2) Label-tree 定量, 筛选组内 $\geq 50\%$ 非空值的数据进行差异分析。
- 3) 显著性差异标准: Fold change > 1.2 (上调) 且 P value < 0.05 .

2.3 生物信息学分析

- 1) 差异蛋白筛选: 火山图、散点图、聚类热图。
- 2) 功能注释和富集分析: GO、KEGG、结构域、亚细胞定位。
- 3) GSEA 分析: 基于基因集富集评估相关功能模块。
- 4) 关键蛋白筛选: 整合差异分析中, 折数、P 值、通路富集、PPI 网络等维度。

三、实验结果

3.1 显著性差异分析 表在定量结果的显著性差异分析中, 我们首先筛选样本组内重复实验数据至少有一半为非空值的数据进行差异比较分析, 筛选表达差异倍数大于 1.2 倍 (上调) 且 P value (t test/significance) 小于 0.05 筛选标准的蛋白质视为差异表达蛋白质。

表 1 差异基因统计结果

比较组	上调蛋白数	下调蛋白数	总差异蛋白数
high vs control	1229	1327	2556
low vs control	424	519	943
moderate vs control	894	1049	1943
high VS moderate	1094	1081	2175
high VS low	1260	1266	2526
moderate VS low	723	820	1543

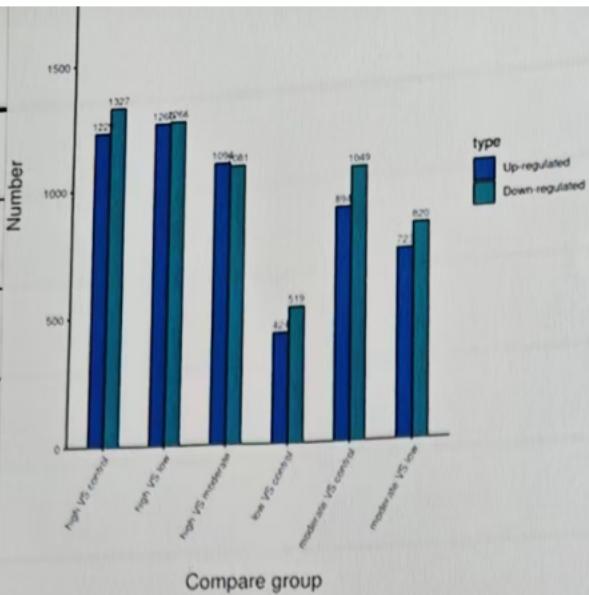


图2 差异蛋白统计柱状图

火山图(Volcano Plot)能够直观展示对比组间和分布差异情况。我们绘制了不标注基因名称和标准注上下调基因名称(剔除筛选以下图中 P value 从小到大排序的 Top 10 个基因)的火山图,结果如下。图中的点,蓝色为上调蛋白,青色为下调蛋白,灰黑色为无差异。

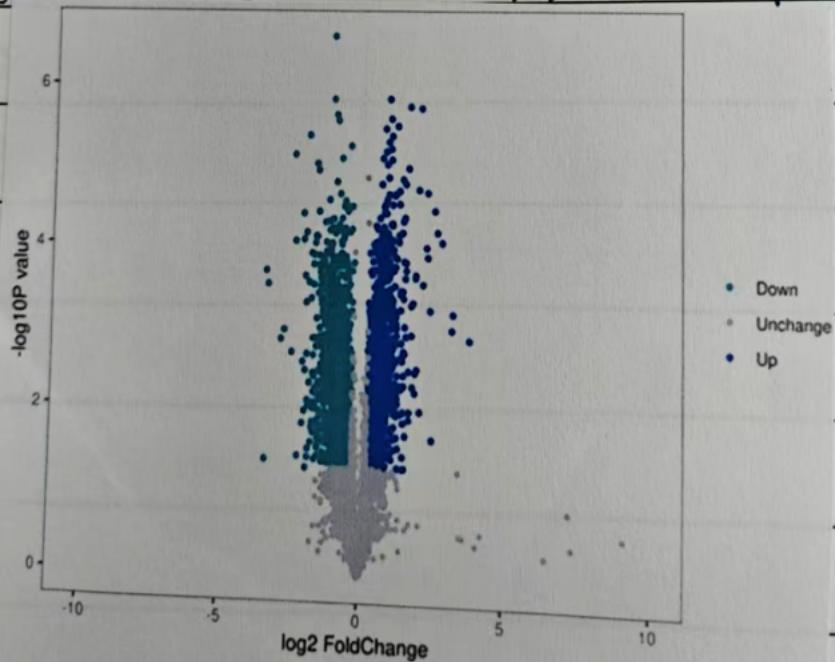


图3 差异蛋白统计火山图

聚类分析是一种常用的数据分析方法,其目的是在相似性的基础上对数据进行分组、归类。聚类分组的结果中,组内的数据

模式相似性较高，而组间的数据模式相似性较低。

在聚类分析过程中，聚类算法对样本和变量两个维度进行分类。对样本的聚类结果可以检验所筛选的目标蛋白质的合理性，即这些目标蛋白质表达量的变化可否代表生物学处理对样本造成的影响。结果如下图所示，显示高耐药组与对照组差异蛋白表达模式显著分离。

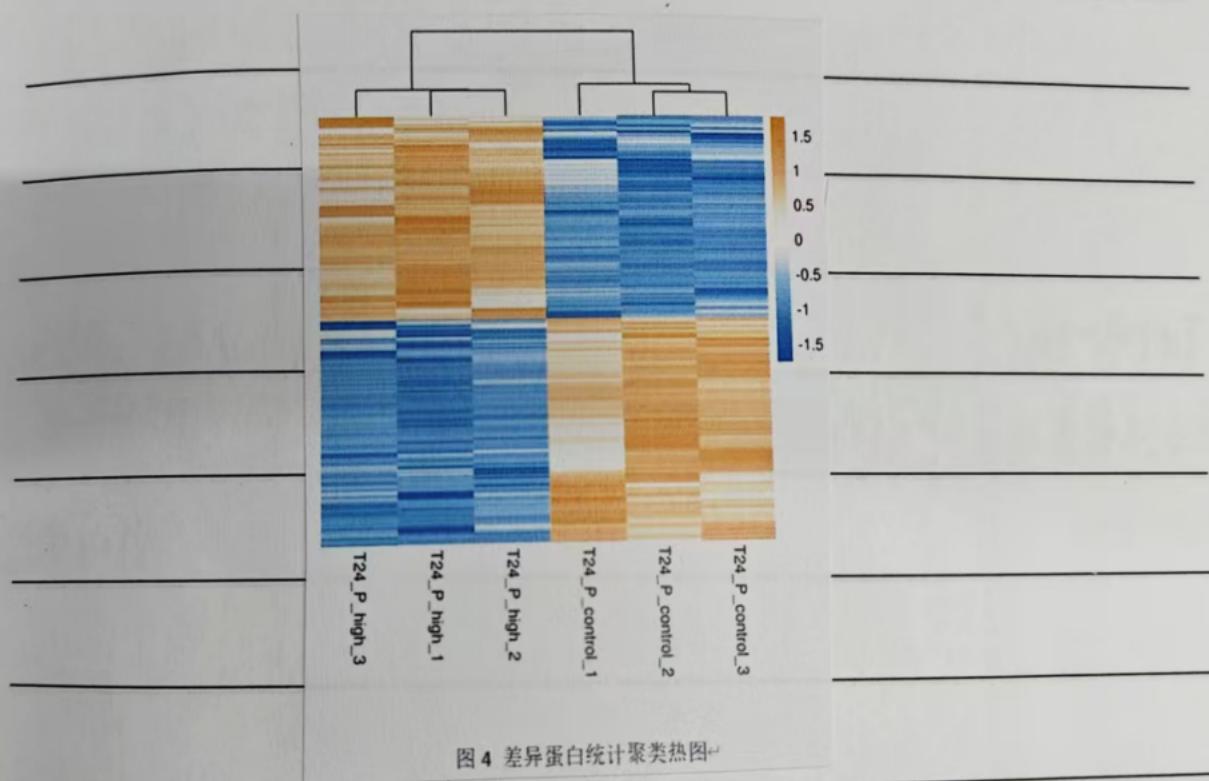


图 4 差异蛋白统计聚类热图

3.2 差异蛋白功能分析。

对于蛋白质组学筛选出的差异结果，我们想知道这些差异蛋白质究竟有怎样的功能。Gene Ontology (GO) 即基因本体论，是一个重要的生物信息学分析方法和工具，用于表述基因和基因产物的各种属性。GO

注释分为3大类：生物进程(Biological Process), 细胞组成(Cellular Component)和分子功能(Molecular Function)(Ashburner et al., 2000)，从不同角度阐释蛋白的生物学作用。因此，通过GO功能注释，可以帮助我们了解功能基因。我们采用Blast2Go(<http://www.blast2go.com/>)软件(Götz et al., 2008; Hung et al., 2014)对所有差异蛋白质进行GO功能。同时，在GO二级功能注释层级上对差异蛋白数目进行统计，结果如图所示。

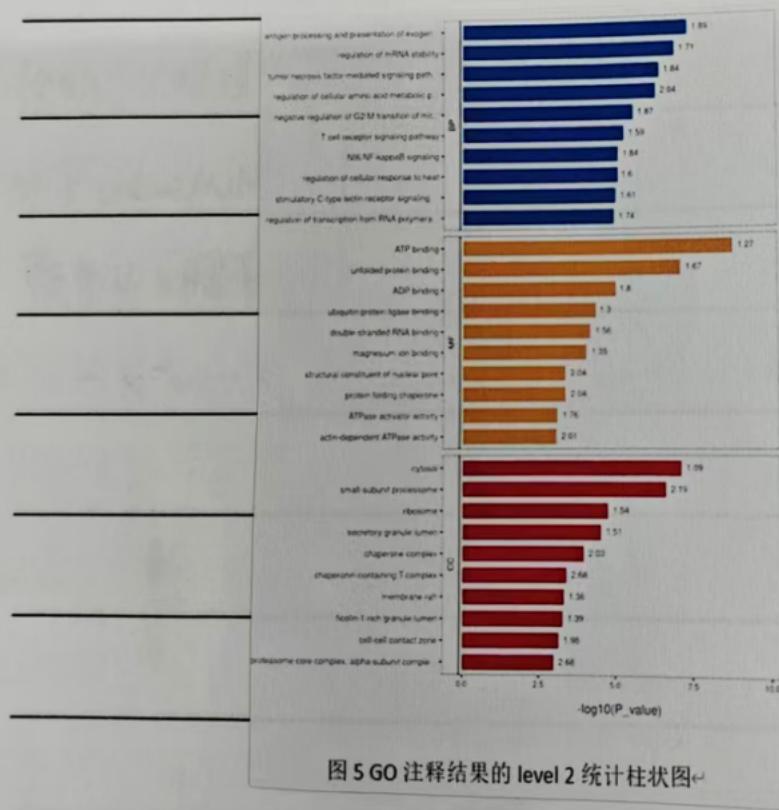


图 5 GO 注释结果的 level 2 统计柱状图

以气泡图展示三大分类里富集前10的分类结果：

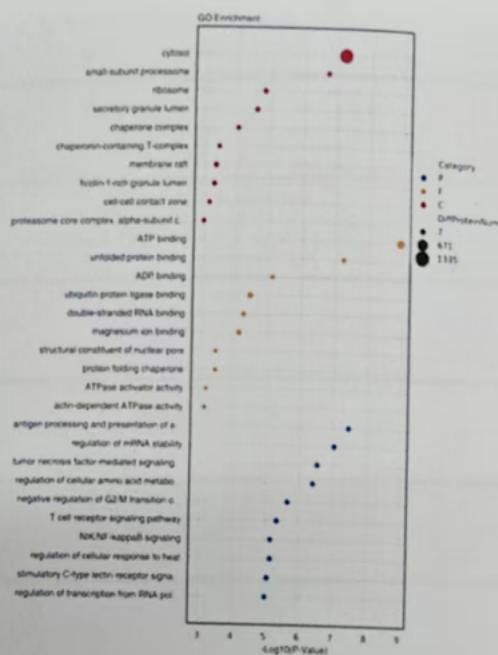


图 6 GO term 的富集统计气泡图 (top 10) ↵

根据 Fold change, 差异蛋白可分为上调和下调和下调分类。为进一步了解上调、下调差异蛋白的功能, 我们绘制了上下调分开展示的GO富集柱状图, 结果如图所示。

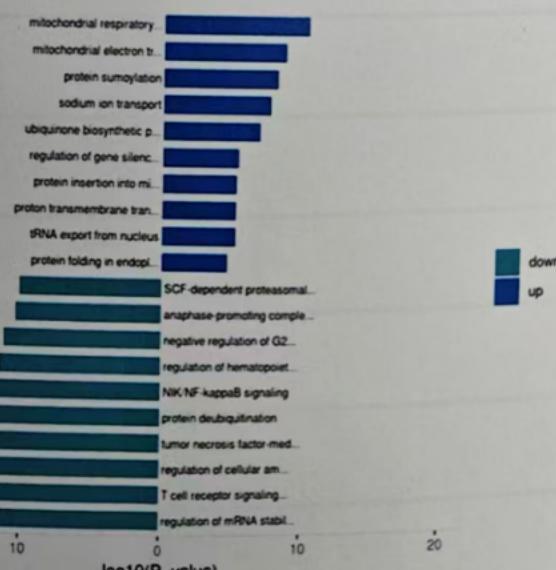


图 7 上下调蛋白质的 GO 功能富集柱状图 (BP) ↵

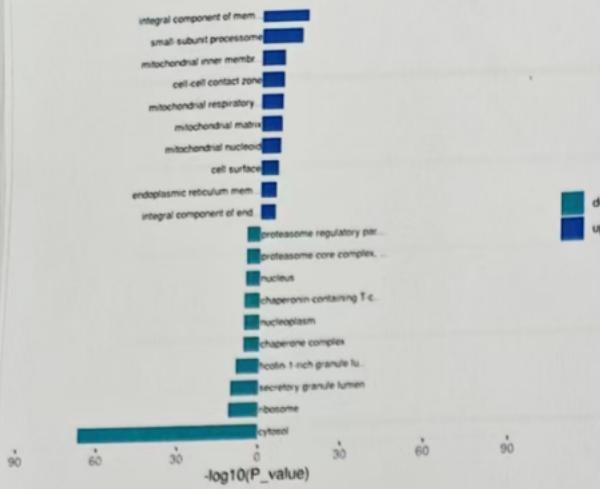


图 8 上下调蛋白质的 GO 功能富集柱状图 (CC)

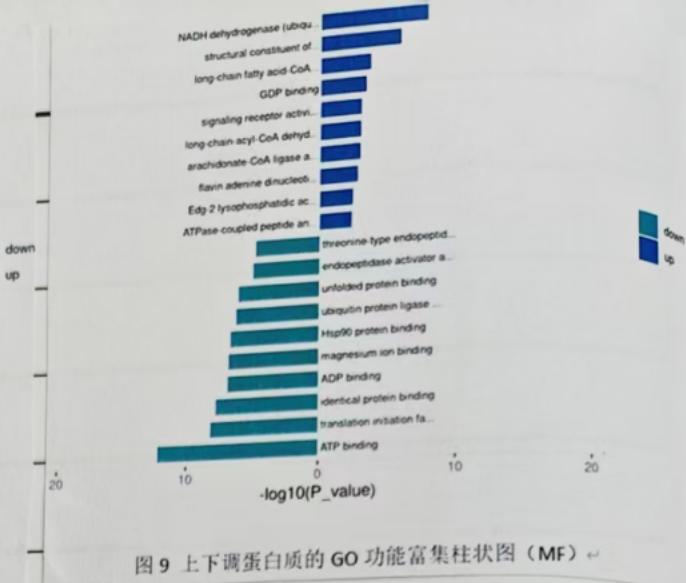


图 9 上下调蛋白质的 GO 功能富集柱状图 (MF)

3.3 差异蛋白通路分析

KEGG (Kyoto Encyclopedia of Genes and Genomes) 是常用于通路研究的数据
库之一，它是由研究人员阅读海量文献后以特定的图形语言描述代谢
途径以及各途径之间的相互关系 (Kanehisa et al., 2012)。KEGG 数据库相
关资料参见：<http://www.kegg.jp/>。通过对显著性差异表达的蛋白质进行 KEGG
通路注释，从而显示蛋白质从胞表而到胞核一系列变化过程。我们根据 KEGG
注释结果，以通路中的差异蛋白质数量为依据进行排序，展示 Top20 的 KEGG
通路注释结果，如图所示。

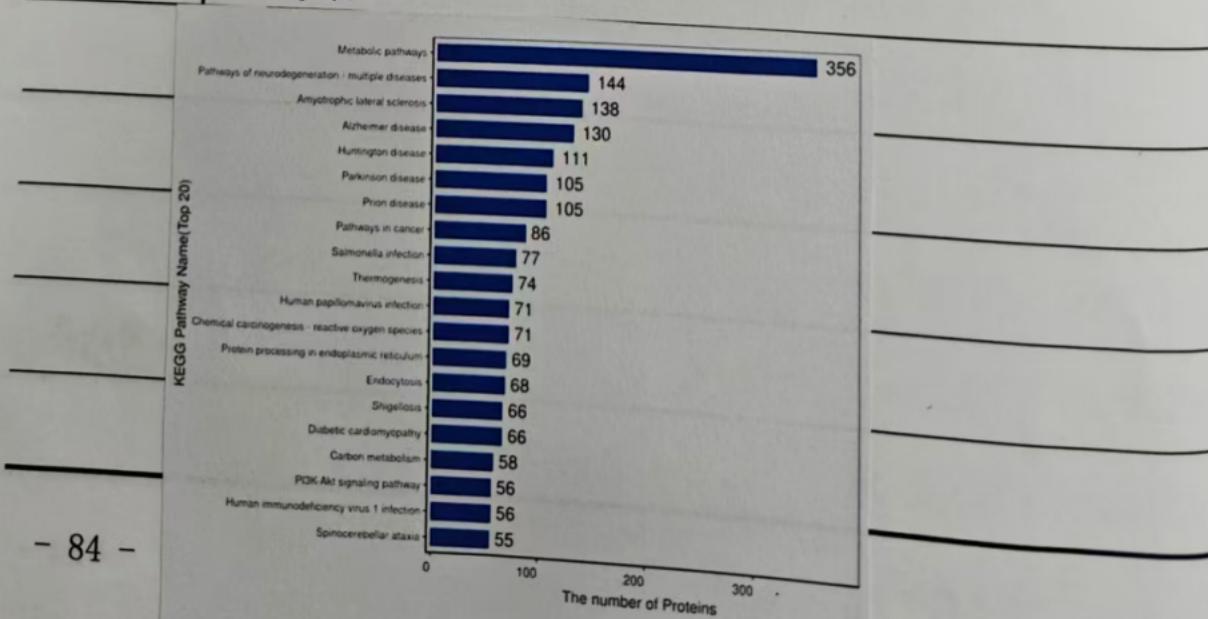


图 10 KEGG 通路注释结果柱状图 (top 20)

KEGG通路富集分析方法与GO富集分析相似，即以KEGG通路为单位，以所有定性蛋白质为背景，通过Fisher精确检验(Fisher's Exact Test)，来分析计算各个通路蛋白质富集度的显著性水平，从而确定受到显著影响的代谢和信号转导途径。对差异蛋白质进行KEGG富集分析，以圈图、柱状图和气泡图的形式来进行结果展示。

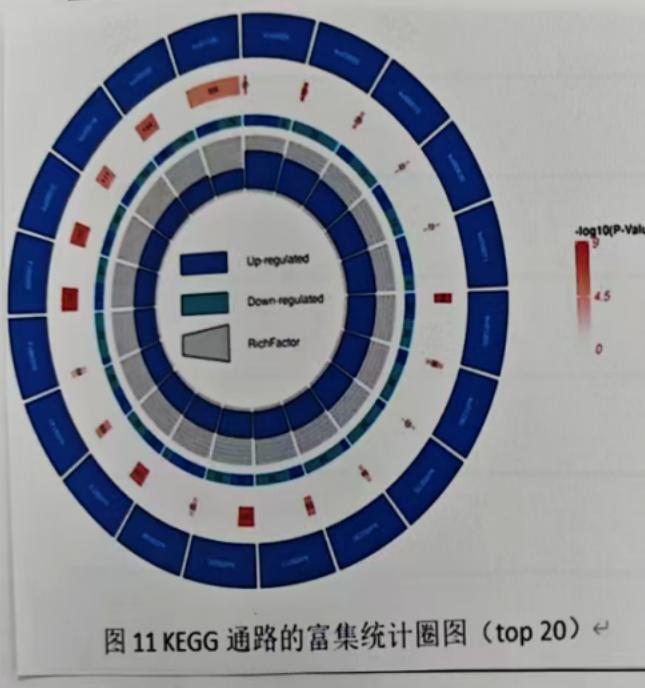


图 11 KEGG 通路的富集统计圈图 (top 20) ↪

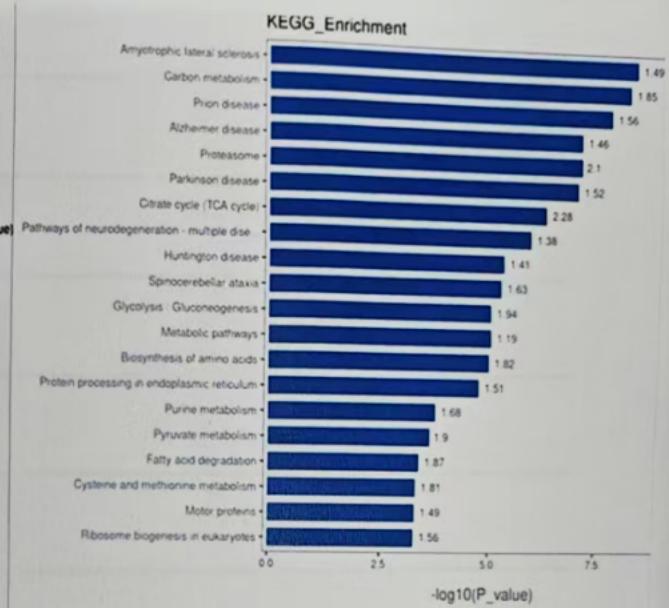


图 12 KEGG 通路的富集统计柱状图 (top 20) ↪

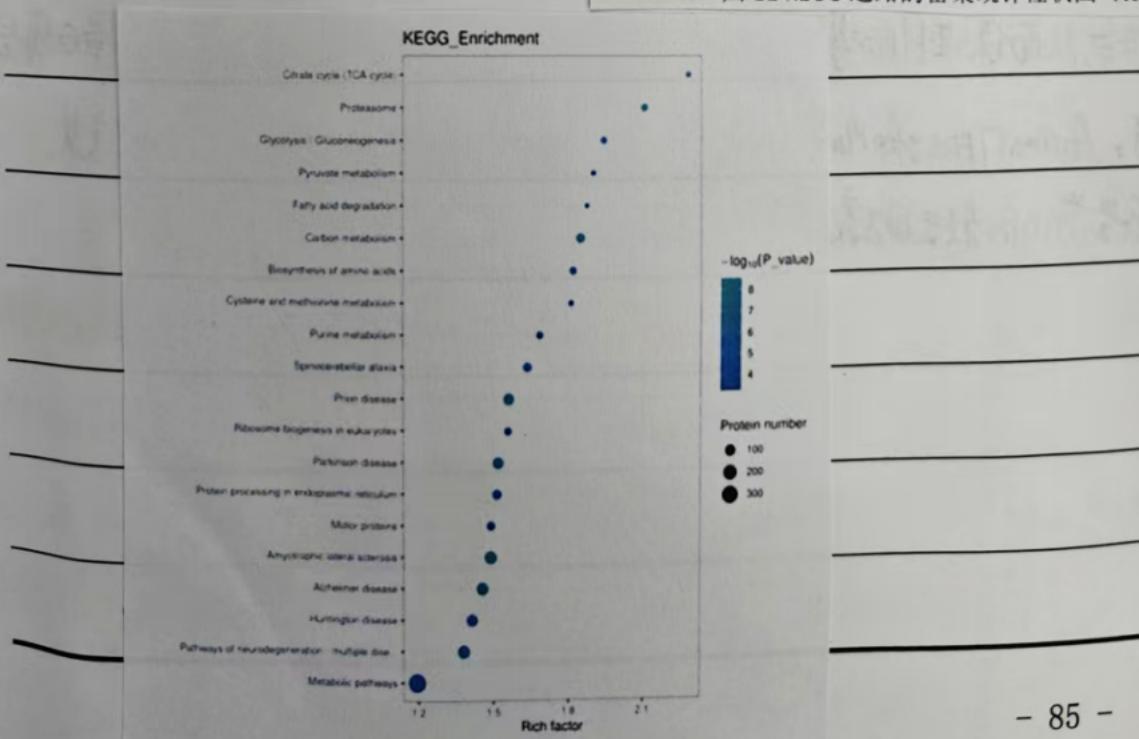


图 13 KEGG 通路的富集统计气泡图 (top 20)

根据Fold change, 差异蛋白可分为上调和下调两类。为进一步了解上下调差异蛋白质参与的代谢和信号转导途径, 我们绘制了上下调分开展 KEGG 富集柱状图, 如下图所示。

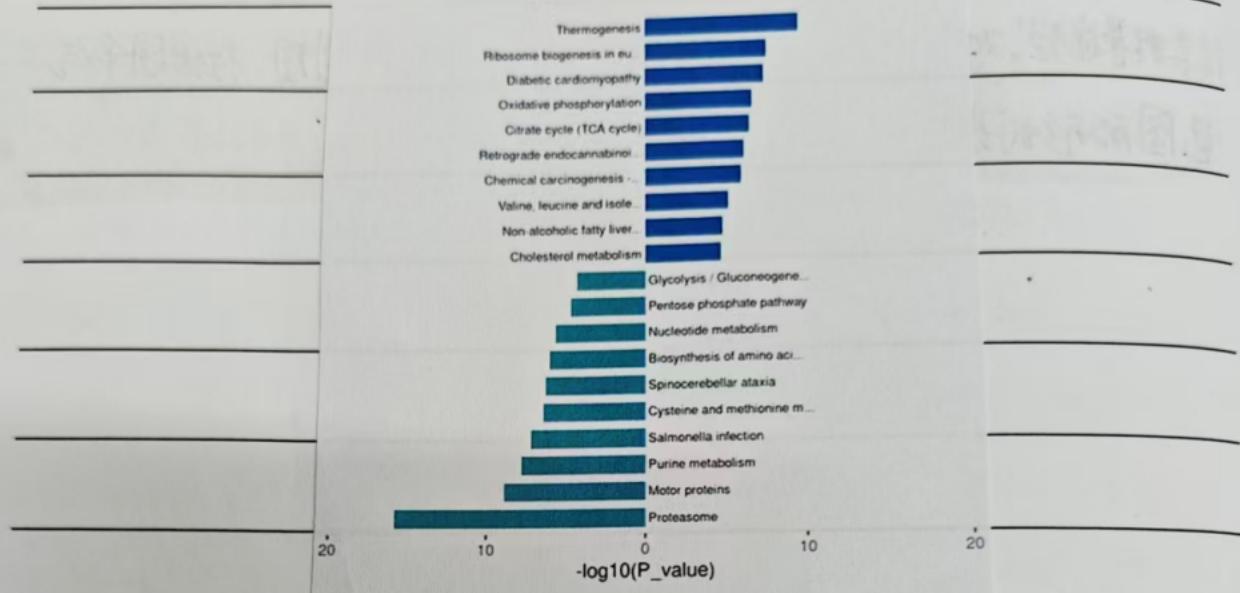


图 14 KEGG 富集柱状图 (上下调分开展示) ↵

3.4 转录因子分析

转录因子(Transcription Factor)是能与基因5'端上游特定序列专一性结合, 从而保证目的基因以特定的强度在特定的时间与空间表达的蛋白质分子。Animal TFDB 和 Plant TFDB 数据库包含动植物转录因子及其家族信息, 数量前10的转录因子家族统计结果如图所示。

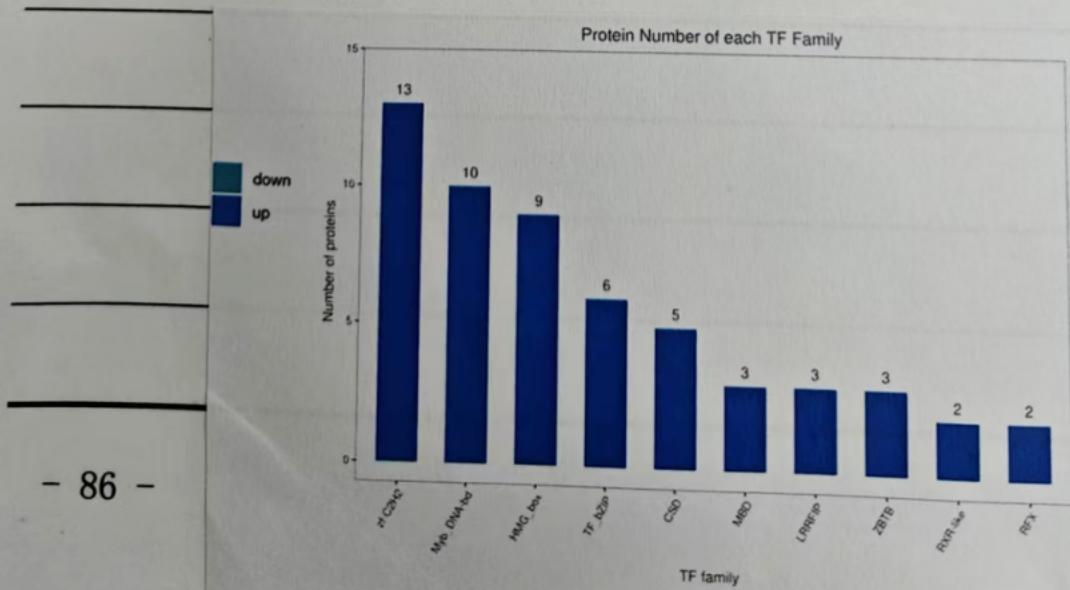


图 15 数量前 10 的 TF 家族统计图 ↵

3.5 GSEA分析

传统的富集分析侧重于比较两组间的基因表达差异，集中关注少数几个显著上调或下调的基因，这种方式存在一定的局限性，比如，(1)由于筛选参数不合理，漏掉部分表达不显著却有着重要生物学意义基因；(2)以及当差异蛋白质数量少的时候，传统富集分析方法得到结果可能会很少，甚至无；(3)难以回答如果传统富集方法富集到的某一通路上，既有上调差异基因，也有下调差异基因，也有下调差异基因，那么这条通路总体的表现形式究竟是怎样？

GSEA其基本思想是不需要指定明确的差异基因阈值，而是按照所有基因在两组样本中的差异表达程度进行排序，然后计算预先设定的基因集在顶端或末端的富集程度及其显著性。

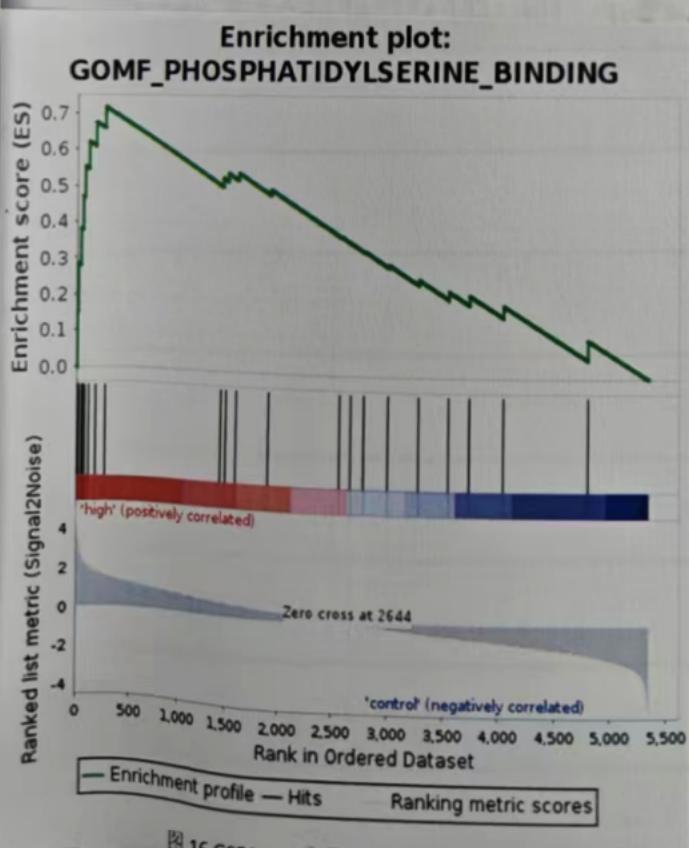


图 16 GSEA-GO 富集得分曲线图



图 17 GSEA-GO 富集统计气泡图

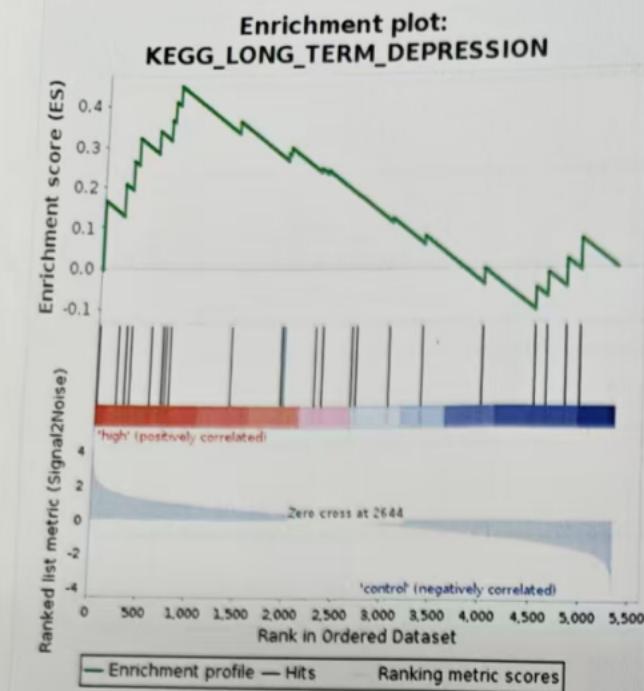


图 18 GSEA-KEGG 富集得分曲线图

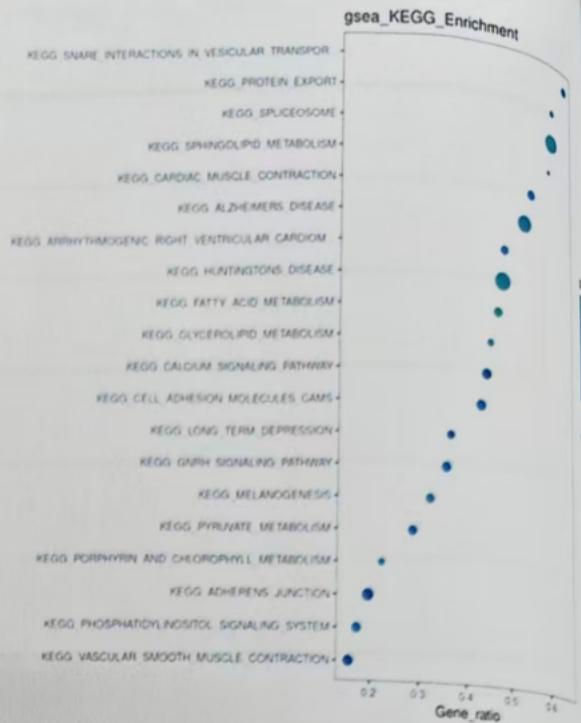


图 19 GSEA-KEGG 富集统计气泡图

8. 关键蛋白质候选列表 基于以下标准筛选出前 20 名核心候选蛋白：

1) 差异显著性：P 值最小 ($P < 0.001$)、Fold change 最大 ($|log_2 FC| > 3$)；2) 功能关联性：参与耐药相关通路；3) 网络中心性：PPI 分析中接壤度 (Degree) > 50 。

表 2 前 20 名核心候选蛋白

Protein IDs ^②	Gene Name ^③	FC ^④	p-value ^⑤	regulation ^⑥
P07204 ⁻³	THBD ⁻³	5.593411 ⁻³	0.004579 ⁻³	UP ⁻³
P05120 ⁻³	SERPINB2 ⁻³	5.383147 ⁻³	3.71E-05 ⁻³	UP ⁻³
O76061 ⁻³	STC2 ⁻³	4.651881 ⁻³	0.000699 ⁻³	UP ⁻³
P34741 ⁻³	SDC2 ⁻³	4.094121 ⁻³	0.000481 ⁻³	UP ⁻³
P26022 ⁻³	PTX3 ⁻³	4.02582 ⁻³	0.001416 ⁻³	UP ⁻³
Q86X29 ⁻³	LSR ⁻³	3.967194 ⁻³	6.92E-05 ⁻³	UP ⁻³
Q8IVT2 ⁻³	MISP ⁻³	3.891363 ⁻³	0.000355 ⁻³	UP ⁻³
P05362 ⁻³	ICAM1 ⁻³	3.452982 ⁻³	0.000328 ⁻³	UP ⁻³
O00622 ⁻³	CCN1 ⁻³	3.328667 ⁻³	0.000932 ⁻³	UP ⁻³
P17275 ⁻³	JUNB ⁻³	3.257302 ⁻³	1.05E-05 ⁻³	UP ⁻³
Q9NX18 ⁻³	SDHAF2 ⁻³	3.234066 ⁻³	0.012645 ⁻³	UP ⁻³
P00749 ⁻³	PLAU ⁻³	3.094785 ⁻³	0.001926 ⁻³	UP ⁻³
P20591 ⁻³	MX1 ⁻³	3.078541 ⁻³	5.48E-05 ⁻³	UP ⁻³

Q01201 ⁻	RELB ⁻	2.950152 ⁻	0.001589 ⁻	UP ⁻
P20592 ⁻	MX2 ⁻	2.844591 ⁻	0.008315 ⁻	UP ⁻
Q13753 ⁻	LAMC2 ⁻	2.832641 ⁻	0.000728 ⁻	UP ⁻
Q9Y4K1 ⁻	CRYBG1 ⁻	2.779093 ⁻	0.002373 ⁻	UP ⁻
Q03405 ⁻	PLAUR ⁻	2.712975 ⁻	0.000375 ⁻	UP ⁻
Q14574 ⁻	DSC3 ⁻	2.641861 ⁻	0.002393 ⁻	UP ⁻
P15407 ⁻	FOSL1 ⁻	2.590803 ⁻	0.000367 ⁻	UP ⁻

三、结果分析.

本研究通过蛋白质组学技术系统解析了T24-RC48耐药细胞的分子特征,发现耐药表型与代谢重编程、蛋白稳态调控及关键信号通路信号异常激活密切相关,筛选出诸多潜在耐药相关蛋白标记物,初步揭示了细胞适应药物压的多层次调控网络,为深入探究RC48耐药机制及开发靶向逆转录策略提供了重要理论依据和实验方向。