

论著·基础研究

基于生存分析的生物信息学方法筛选乳腺癌枢纽基因和关键通路

陈思*, 刘春良*, 赵倩, 孙海鹏, 刘云霞

上海交通大学基础医学院病理生理学系, 细胞分化与凋亡教育部重点实验室, 上海 200025

[摘要] **目的**·利用生物信息学分析, 将基因表达数据与临床生存分析相结合, 以筛选乳腺癌的枢纽基因和关键通路。**方法**·从 GEO 数据库 (Gene Expression Omnibus) 下载了 3 个乳腺癌基因表达数据集, 筛选出乳腺癌中的差异表达基因。Kaplan-Meier plotter 数据库进一步筛选出与总体生存期相关的差异表达基因, 并对这些基因进行 GO (Gene Ontology) 功能分析和 KEGG (Kyoto Encyclopedia of Genes and Genomes) 通路分析。通过构建蛋白-蛋白相互作用网络筛选乳腺癌枢纽基因。利用 Oncomine 数据库和人类蛋白质图谱数据库来验证枢纽基因的表达。实时荧光定量 PCR 检测人乳腺癌细胞 MDA-MB-231 和人正常乳腺上皮细胞 MCF-10A 中枢纽基因的表达情况。**结果**·筛选到 262 个差异表达基因与乳腺癌患者的总体生存期显著相关。GO 功能分析和 KEGG 通路分析结果显示, 这些基因与细胞核分裂、细胞分裂和染色体分离相关, 并且主要富集在细胞周期、FoxO 信号通路和卵母细胞减数分裂等通路上。蛋白质相互作用网络构建确定了 10 个枢纽基因。经数据库验证, 它们在乳腺癌中均高表达; 实时荧光定量 PCR 结果显示, 10 个枢纽基因中有 8 个在乳腺癌细胞中高表达。**结论**·通过基于生存分析的生物信息学方法筛选出了参与乳腺癌发生发展的关键基因和通路, 这些基因和通路主要与细胞周期调控和细胞分裂相关。

[关键词] 乳腺癌; 生物信息学; 枢纽基因; 生存分析; 生物学标志物

[DOI] 10.3969/j.issn.1674-8115.2020.03.004 **[中图分类号]** R737.9 **[文献标志码]** A

Identification of hub genes and key pathways in breast cancer by survival-based bioinformatics analysis

CHEN Si*, LIU Chun-liang*, ZHAO Qian, SUN Hai-peng, LIU Yun-xia

Key Laboratory of Cell Differentiation and Apoptosis of National Ministry of Education, Department of Pathophysiology, Shanghai Jiao Tong University College of Basic Medical Sciences, Shanghai 200025, China

[Abstract] **Objective**·To identify hub genes and key pathways in breast cancer by bioinformatics analysis that integrated gene expression data with clinical survival analysis. **Methods**·Three gene expression profilings downloaded from Gene Expression Omnibus (GEO) were used to identify differentially expressed genes (DEGs) in breast cancer. Kaplan-Meier plotter was used to identify the DEGs that were significantly associated with overall survival in breast cancer. Gene Ontology (GO) function analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were performed. Next, hub genes were identified from the protein-protein interaction (PPI) network. Oncomine and the Human Protein Atlas (HPA) database were used to validate the expression of the hub genes. The expressions of hub genes in MDA-MB-231 cells and MCF-10A cells were detected by quantitative real-time PCR (qPCR). **Results**·Among the DEGs, 262 genes were significantly correlated with overall survival of breast cancer patients. The results of GO functional analysis and KEGG pathway analysis showed that these genes were associated with nuclear division, cell division and chromosome segregation, and were mainly enriched on the pathways such as cell cycle, FoxO signaling pathway and oocyte meiosis. PPI network construction identified ten hub genes. They were all highly expressed in breast cancer, which were validated by the databases. The results of qPCR showed that 8 out of 10 hub genes were highly expressed in breast cancer cells. **Conclusion**·The hub genes and key pathways involved in the development of breast cancer are identified by survival-based bioinformatics analysis, which are mainly associated with cell cycle regulation and cell division.

[Key words] breast cancer; bioinformatics; hub gene; survival analysis; biomarker

乳腺癌是最为常见的恶性肿瘤之一, 也是女性中发病率最高的恶性肿瘤^[1]。根据雌激素受体 (estrogen receptor, ER)、孕激素受体 (progesterone receptor, PR) 和人类表皮生长因子受体 2 (human epidermal growth factor receptor-2, HER2) 的表达情况, 可以将乳腺癌分为 4 种亚型 (Luminal

A、Luminal B、HER2 阳性和三阴乳腺癌)。根据不同乳腺癌亚型的生物学特征和临床病理分期制定相应的个体化治疗策略, 可以使乳腺癌 5 年生存率达 90% 以上^[2]。然而, 乳腺癌的复发和转移仍旧是一大难题。除此之外, 一些乳腺癌亚型, 例如三阴乳腺癌, 由于缺乏有效的治疗靶点,

[基金项目] 国家自然科学基金 (81570717)。

[作者简介] 陈思 (1994—), 女, 硕士生; 电子信箱: leoway122@163.com。刘春良 (1993—), 男, 硕士生; 电子信箱: liuchunliang@sjtu.edu.cn。* 为共同第一作者。

[通信作者] 刘云霞, 电子信箱: jxsdklyx@126.com。

[Funding Information] National Natural Science Foundation of China (81570717)。

[Corresponding Author] LIU Yun-xia, E-mail: jxsdklyx@126.com。



一直以来是临床治疗的一个瓶颈。因此,明确乳腺癌发生发展相关的关键基因和通路,有助于认识乳腺癌潜在的发病机制,或将为临床寻找更多的诊断和治疗靶点提供参考。

自基因芯片技术和高通量测序技术问世以来,生物信息学迅速发展,目前已发现了许多疾病的生物学标志物^[3-4]。从信息丰富的公共数据库如 GEO 数据库 (Gene Expression Omnibus, 基因表达汇编) 和 TCGA 数据库 (The Cancer Genome Atlas, 癌症基因组图谱) 中可获得基因表达数据,通过生物信息学方法,对数据库中的基因表达数据进行聚类分析、统计分析、通路分析和可视化作图等,能够预测基因的功能以及基因间的相互作用,了解疾病基因层面的发病机制,发现潜在的生物学标志物,从而为疾病的分子靶向药物研发和精准治疗提供理论依据。本研究将乳腺癌基因表达数据与临床生存分析相结合,以筛选枢纽基因和关键信号通路;基于生存期筛选出来的基因可能更具有临床意义,或能为乳腺癌的诊断和治疗提供新的思路。

1 材料与方法

1.1 数据采集

为了获取乳腺癌的基因表达数据集,本研究在 GEO 数据库下载了 3 个乳腺癌数据集 (<http://www.ncbi.nlm.nih.gov/geo/>), 分别为 GSE54002、GSE29431 和 GSE61304, 这 3 个数据集都是基于 GPL570 平台。GSE54002 包含 417 例乳腺癌样本和 16 例正常样本。GSE29431 包含了 54 例乳腺癌样本和 12 例正常样本; 54 例乳腺癌样本中有 15 例 HER2 免疫组织化学评分为 3+, 且伴有 HER2 基因扩增; 26 例评分为 2+, 其中 13 例伴有 HER2 基因扩增, 13 例不伴有 HER2 基因扩增; 13 例评分为 0/1+, 且不伴有 HER2 基因扩增。GSE61304 包含了 58 例乳腺癌样本和 4 例正常样本; 58 例乳腺癌样本中 18 例为 ER⁺PR⁺, 19 例为 ER⁺PR⁻, 4 例为 ER⁻PR⁻, 1 例为 ER⁻PR⁺, 其他 16 例样本未说明。

1.2 筛选差异表达基因

在 Rstudio 软件中 (版本 3.4.0) 加载并下载 Bioconductor 网站上软件包来分析上述 3 个乳腺癌数据集。首先使用 affy 包导入 CEL 文件, 使用 simpleaffy 包评估微阵列数据质量^[5], gcrma 包中的 RMA 算法预处理原始数据^[6], genefilter 包过滤非特异性结合的探针和数据质量低的探针, limma 包进行差异基因表达的统计学验证^[7]。基因表达变化倍数的对数值的绝对值 ($|\log_2 FC|$) > 1 且 $P < 0.05$ 认定为差异表达基因。最后, 为了提高差异表达基因的稳健性, 使用 Funrich 软件 (版本 3.1.3) 获得 3 个数据集中均

上调或下调的基因, 用于下一步的分析。

1.3 筛选与总体生存期相关的差异表达基因

为了分析差异表达基因对乳腺癌患者总体生存期的影响, 在 Kaplan-Meier plotter 数据库 (www.kmplot.com) 中根据基因的中位表达值将患者样本分为高表达组和低表达组。使用默认参数, 计算每个基因高表达组和低表达组的中位生存期; 若 log-rank $P < 0.05$, 则该基因被视为与总体生存期相关的差异表达基因。

1.4 分析与总体生存期相关的差异表达基因的功能

基因本体数据库 (Gene Ontology, GO) 富集分析和京都基因与基因组百科全书 (Kyoto Encyclopedia of Genes and Genome, KEGG) 通路富集分析被广泛用于识别基因的功能和通路。本研究使用 Rstudio 软件中 clustProfiler 包对与乳腺癌患者总体生存期相关的差异表达基因进行 GO 分析和 KEGG 分析, $P < 0.05$ 认为具有统计学意义^[8]。

1.5 蛋白-蛋白相互作用网络构建及筛选枢纽基因

蛋白与蛋白相互作用在调节生物学过程中起着至关重要的作用。这种关系可以通过蛋白-蛋白相互作用 (protein-protein interaction, PPI) 网络表示, 每个节点代表一个蛋白, 边代表蛋白质之间的相互作用。紧密相连的区域可以作为富集功能群。STRING 数据库 (<https://string-db.org/>) 包含了丰富的蛋白质之间相互作用的信息^[9]。为了评估与总体生存期相关的差异表达基因之间的相互关系, 将差异表达基因列表导入 STRING 数据库, 并设定信度为 0.4; 接着将 PPI 表格数据导入 Cytoscape 软件中构建 PPI 网络, 使用软件中的插入式分子复合物检测 (MCODE 评分 > 4 分, 节点数 > 5 个) 筛选出 PPI 网络中的枢纽模块, 最后通过 CytoHubba 插件计算网络中每一个基因的最大团中心性 (maximal clique centrality, MCC) 分数, 将得分前 10 的基因作为枢纽基因^[10-12]。

1.6 枢纽基因验证

1.6.1 数据库验证枢纽基因表达 利用 Oncomine 数据库 (<https://www.oncomine.org/resource/main.html/>) 和人类蛋白质图谱 (Human Protein Atlas, HPA) 数据库 (<https://www.proteinatlas.org/>) 对枢纽基因在乳腺癌肿瘤组织和正常组织间的 mRNA 和蛋白水平的表达进行验证^[13]。

1.6.2 RNA 抽提及实时荧光定量 PCR 人乳腺癌细胞 MDA-MB-231 和人正常乳腺上皮细胞 MCF-10A (购自中国科学院干细胞库) 的总 RNA 抽提按照 Invitrogen 公司 TRIzol 试剂盒提供的方法。反转录以 1 000 ng RNA 为模

板, 按 TaKaRa 公司 M-MLV Reverse Transcription Kit 试剂盒说明配制反转录反应液, 进行反转录反应。实时荧光定量 PCR (quantitative real-time PCR, qPCR) 使用 ABI 公司 7500 Real-Time PCR System 试剂盒。每份采用 10 μ L 体系, cDNA 模板 1 μ L, 每份样品做 3 个复孔, 求平均值; 以 18S rRNA 为内参, 引物序列见表 1。

表 1 qPCR 引物序列
Tab 1 Primer sequences for qPCR

Gene	Forward sequence (5' \rightarrow 3')	Reverse sequence (5' \rightarrow 3')
<i>BIRC5</i>	AGGACCACGCATCTCTACAT	AAGTCTGGCTCGTTCTCAGTG
<i>NDC80</i>	CCTCTCCATGCAGGAGTTAAGA	GGTCTCGGGTCTTGATTTTCT
<i>MAD2L1</i>	GTTCTTCTCATTCCGCATCAACA	GAGTCCGTATTCTGCACTCG
<i>CENPF</i>	AAAGAAACAGACGGAACAACTG	CCAAGCAAAGACCGAGAACT
<i>CCNB1</i>	CTTGCAAGTAAATGATGTGGATG	GTGACTTCCCGACCCAGTAG
<i>BUB1</i>	ACAATCAACGGAGAAAGCATGA	CTCCACCACCTGATGCAACT
<i>BUB1B</i>	AAATGACCCTCTGGATGTTTGG	GCATAAACGCCCTAATTTAAGCC
<i>KIF2C</i>	ACTCTAGGACTTGCGATGATTGCC	TGGGTGTCAAACCAAACAGA
<i>CDC20</i>	ACGGTTTTGATGTAGAGGAAGC	GATACGGTCTGGCAGGGAAG
<i>CDC48</i>	CTTCGCCCTTGGAGGAAACAA	GGTGCTGAATAGCTTCTGCTG

Note: *BIRC5*—baculoviral IAP repeat containing 5; *NDC80*—NDC80 kinetochore complex component; *MAD2L1*—mitotic arrest deficient 2 like 1; *CENPF*—centromere protein F; *CCNB1*—cyclin B1; *BUB1*—mitotic checkpoint serine/threonine kinase; *BUB1B*—BUB1 mitotic checkpoint serine/threonine kinase B; *KIF2C*—kinesin family member 2C; *CDC20*—cell division cycle 20; *CDC48*—cell division cycle associated 8.

1.7 统计学分析

应用 OriginPro 2017C 和 Adobe Illustraor CS6 软件进行统计学分析和作图, qPCR 数值用 $\bar{x} \pm s$ 表示, 组间比较采用独立样本 *t* 检验。 $P < 0.05$ 认为差异具有统计学意义。

2 结果

2.1 筛选差异表达基因

在 $|\log_2FC| > 1$ 且 $P < 0.05$ 的筛选条件下, 从 GSE54002

中得到差异表达基因 3 389 个, 其中上调基因 1 561 个, 下调基因 1 828 个; GSE29431 中得到差异表达基因 3 660 个, 其中上调基因 1 097 个, 下调基因 2 563 个; GSE61304 中得到差异表达基因 1 828 个, 其中上调基因 821 个, 下调基因 1 007 个。然后用 Funrich 软件的 Vene 图筛选得到了 211 个共同上调和 374 个共同下调的差异表达基因, 如图 1 所示。



Notes: A. 211 differentially expressed genes up-regulated in the three datasets. B. 374 differentially expressed genes down-regulated in the three datasets.

图 1 GSE29431、GSE54002 和 GSE61304 中筛选得到的差异表达基因的 Venn 图

Fig 1 Venn diagrams of the differentially expressed genes screened in GSE29431, GSE54002 and GSE61304

2.2 乳腺癌中与总体生存期相关的差异表达基因

为了筛选与总体生存期相关的基因, 使用 Kaplan-Meier plotter 数据库对这 585 个差异表达基因进行总体生存分析, 计算每个基因的高表达组和低表达组的中位生存期和

log-rank *P* 值。结果显示, 有 262 个基因的高表达组和低表达组的总体生存期之间的差异有统计学意义。表 2 列出了与生存相关最显著的前 10 个差异表达基因。这 262 个与总体生存期相关的差异表达基因将用于下一步的基因功能分析。

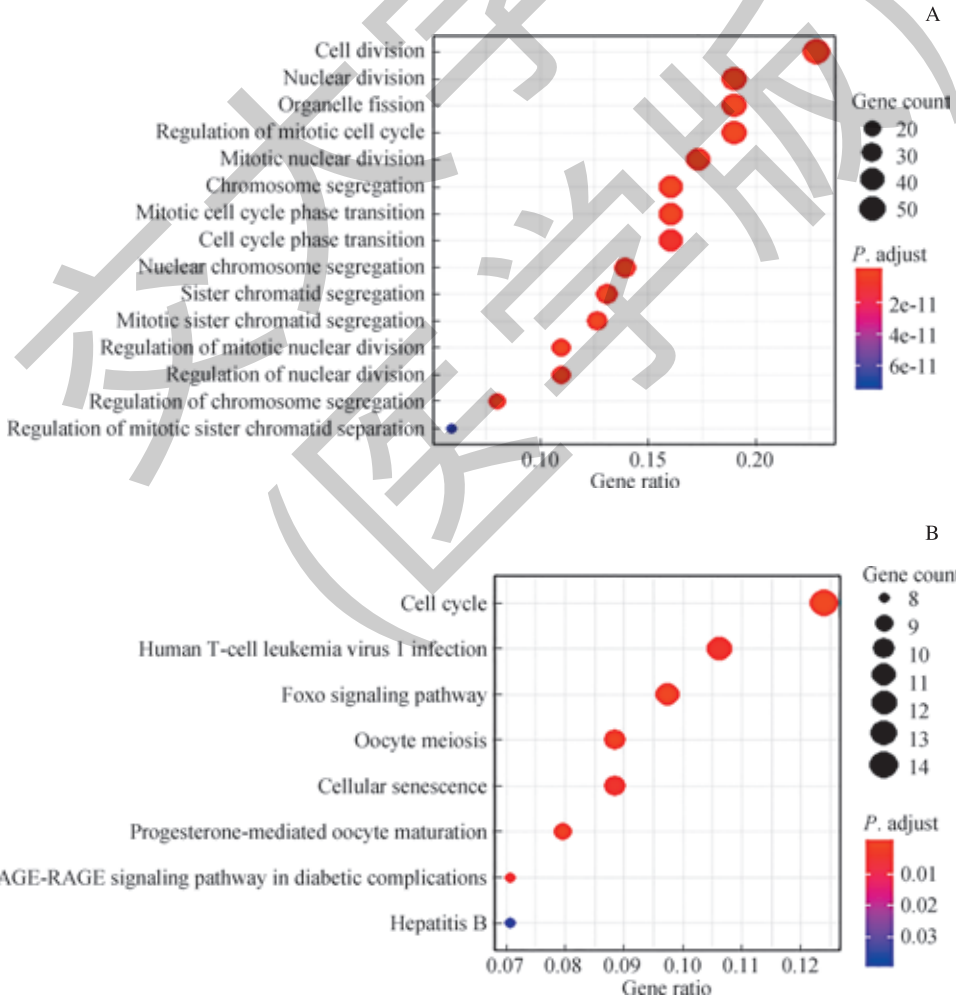
表 2 与总体生存期相关最显著的前 10 个差异表达基因
Tab 2 List of the top 10 overall survival-related differentially expressed genes

Gene	Median survival/month		Log-rank <i>P</i> value
	Low expression cohort	High expression cohort	
<i>CCNB1</i>	151.04	63.00	5.2×10^{-11}
<i>MELK</i>	151.04	63.25	9.2×10^{-11}
<i>CAP-G</i>	143.74	63.06	9.9×10^{-11}
<i>MAD2L1</i>	151.04	63.83	1.8×10^{-10}
<i>CDC20</i>	143.74	63.25	2.1×10^{-10}
<i>NUSAP1</i>	169.20	65.64	4.8×10^{-10}
<i>MYBL2</i>	169.20	64.80	5.5×10^{-10}
<i>ASE1/PRC1</i>	169.20	63.52	5.9×10^{-10}
<i>CDCA8</i>	169.20	64.80	6.1×10^{-10}
<i>SPAG5</i>	143.74	63.06	6.2×10^{-10}

Note: *MELK*—maternal embryonic leucine zipper kinase; *CAP-G*—condensin subunit CAP-G; *NUSAP1*—nucleolar and spindle associated protein 1; *MYBL2*—MYB proto-oncogene like 2; *ASE1/PRC1*—anaphase spindle elongation/protein regulator of cytokinesis 1; *SPAG5*—sperm associated antigen 5.

2.3 乳腺癌中与总体生存期相关的差异表达基因的功能分析

为进一步了解这些基因的功能, 利用 Rstudio 软件中的 clusterProfiler 包对得到的生存期相关的差异表达基因进行功能分析, $P<0.05$ 认为具有统计学意义。GO 功能分析结果显示, 这些基因的功能主要与细胞分裂 (cell division)、核分裂 (nuclear division)、细胞器分裂 (organelle fission)、细胞周期的调控 (regulation of mitotic cell cycle)、负向调控细胞有丝分裂 (mitotic nuclear division) 以及染色体分离 (chromosome segregation) 等生物学过程有关 (图 2A)。KEGG 通路分析显示这些基因参与细胞周期 (cell cycle)、人 T 细胞白血病病毒 1 感染 (human T-cell leukemia virus 1 infection)、FoxO 信号通路 (FoxO signaling pathway)、卵母细胞减数分裂 (oocyte meiosis)、细胞衰老 (cellular senescence) 以及孕酮介导的卵母细胞成熟 (progesterone-mediated oocyte maturation) 等信号通路 (图 2B)。



Note: A. GO function analysis. B. KEGG pathway analysis. Sorted by number of genes.

图 2 与总体生存期相关的差异表达基因 GO 功能分析和 KEGG 通路分析
Fig 2 GO function analysis and KEGG pathway analysis of the overall survival-related differentially expressed genes

将与总体生存期相关的差异表达基因列表上传至 STRING, 并设定信度 0.4 作为判断相互作用是否有意义的标准, 构建了 PPI 网络 (图 3)。Cytoscape 软件根据 MCODE 评分排序筛选出了 3 个枢纽模块 (图 4), 并对枢纽模块中的基因进行 GO 功能分析, 结果显示模块 1 和模

块 2 的基因主要富集在细胞分裂 (cell division)、细胞核分裂 (nuclear division) 和细胞器分裂 (organelle fission) 等生物学过程, 模块 3 的基因主要集中在血小板脱颗粒 (platelet degranulation)、负向调控线粒体细胞色素 c 的释放 (negative regulation of release of cytochrome c from mitochondria) 等生物学过程 (表 3)。

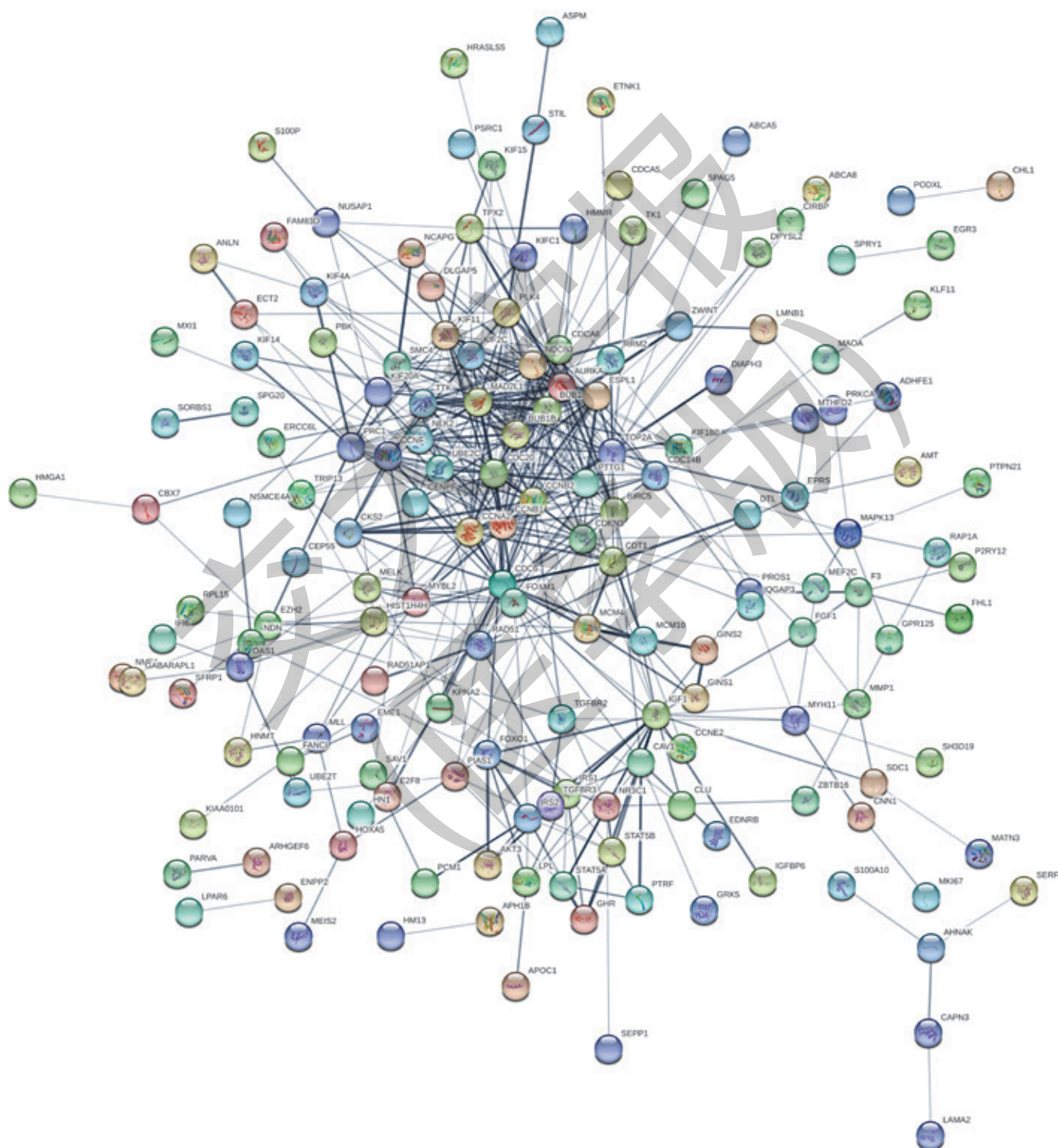
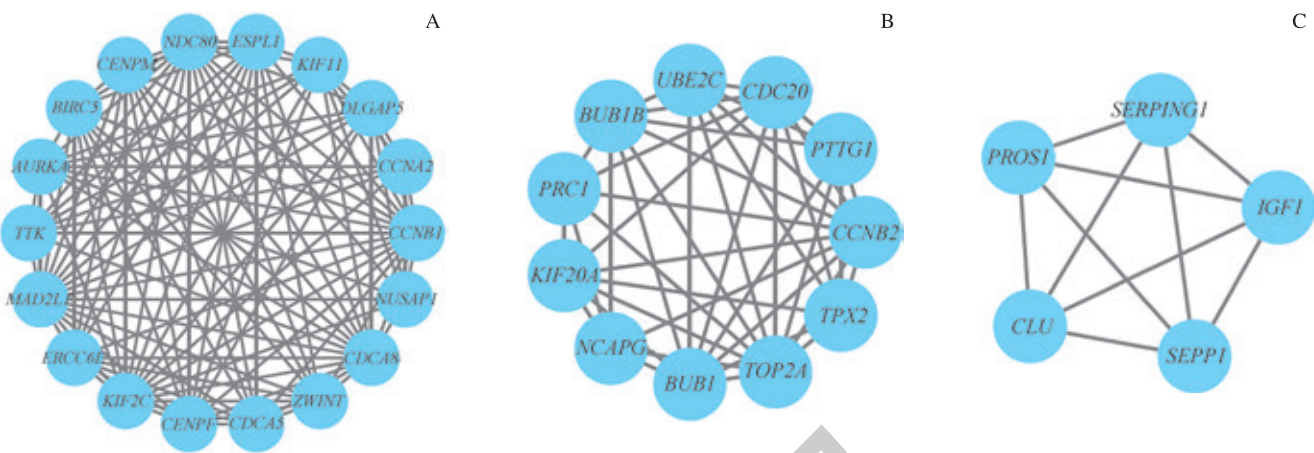


图 3 与总体生存期相关的差异表达基因的 PPI 网络
Fig 3 PPI network of overall survival-related differentially expressed genes



Note: A. Module 1; MCODE score was 13.2. B. Module 2; MCODE score was 8.6. C. Module 3; MCODE score was 5.0. *ESPL1*—extra spindle pole bodies like 1; *KIF11*—kinesin family member 11; *DLGAP5*—DLG associated protein 5; *CCNA2*—cyclin A2; *ZWINT*—ZW10 interacting kinetochore protein; *CDCA5*—cell division cycle associated 5; *ERCC6L*—ERCC excision repair 6 like, spindle assembly checkpoint helicase; *TTK*—Thr/Tyr kinase; *AURKA*—aurora kinase A; *CENPM*—centromere protein M; *UBE2C*—ubiquitin-conjugating enzyme E2 C; *PTTG1*—PTTG1 regulator of sister chromatid separation, securin; *CCNB2*—cyclin B2; *TPX2*—TPX2 microtubule nucleation factor; *TOP2A*—DNA topoisomerase II alpha; *NCAPG*—non-SMC condensin I complex subunit G; *KIF20A*—kinesin family member 20A; *PRC1*—protein regulator of cytokinesis 1; *IGF1*—insulin like growth factor 1; *SEPP1*—selenoprotein P; *CLU*—clusterin; *PROS1*—protein S; *SERPING1*—serpin family G member 1.

图 4 MCODE 评分排序筛选出的 3 个枢纽模块
Fig 4 Three modules selected by MCODE scoring sorting

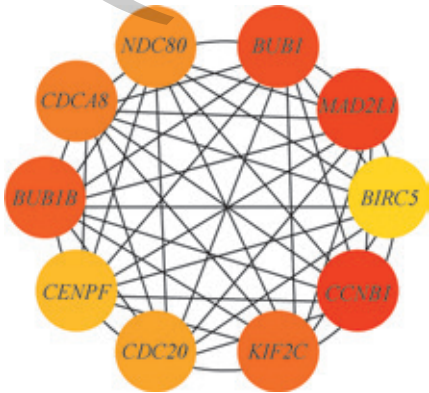
表 3 3 个模块基因的 GO 功能分析
Tab 3 GO functional analysis of the differentially expressed genes in three modules

Module	GO ID	Description	P. adjust value
Module 1	GO:0140014	Mitotic nuclear division	1.82×10^{-18}
	GO:0000280	Nuclear division	7.46×10^{-17}
	GO:0048285	Organelle fission	1.87×10^{-16}
Module 2	GO:0051301	Cell division	2.94×10^{-16}
	GO:0000280	Nuclear division	2.35×10^{-15}
	GO:0048285	Organelle fission	6.07×10^{-15}
Module 3	GO:0002576	Platelet degranulation	9.82×10^{-7}
	GO:0090201	Negative regulation of release of cytochrome c from mitochondria	6.97×10^{-6}
	GO:0006958	Complement activation, classical pathway	3.55×10^{-5}

2.5 乳腺癌中枢纽基因的鉴定

枢纽基因是一类在生物学过程中发挥至关重要作用的基因，在相关通路中其他非枢纽基因的调控往往要受到这类基因的影响，因此枢纽基因有可能成为乳腺癌的生物学

标志物和治疗靶标。使用 Cytoscape 中的插件 Cytohubba，通过 MCC 法得到了得分前 10 的枢纽基因，分别是 *NDC80*、*BUB1*、*CDCA8*、*BUB1B*、*BIRC5*、*CCNB1*、*KIF2C*、*CENPF*、*MAD2L1* 和 *CDC20*（图 5）。



Note: The darker the color, the higher the MCC score.
图 5 通过 MCC 法得到的 10 个枢纽基因及其相互作用
Fig 5 Ten hub genes and their interactions by MCC

2.6 枢纽基因的验证

通过 Oncomine 数据库和 HPA 数据库对筛选得到的 10 个枢纽基因进行表达验证。Oncomine 数据库结果显示, *BIRC5*、*CDC20*、*NDC80*、*CENPF*、*MAD2L1*、*CDCA8*、*KIF2C*、*BUB1*、*CCNB1* 和 *BUB1B* 的 mRNA 水平在乳腺癌组织中明

显上调 (图 6)。HPA 数据库免疫组织化学检测结果显示, 肿瘤组织中 *CDCA8*、*BIRC5*、*CDC20*、*CENPF*、*MAD2L1* 和 *CCNB1* 的蛋白表达水平高于正常乳腺组织, 另外 4 个基因 *NDC80*、*KIF2C*、*BUB1*、*BUB1B* 未被 HPA 数据库收录。以上结果提示, 筛选得到的枢纽基因具有较好的稳健性。

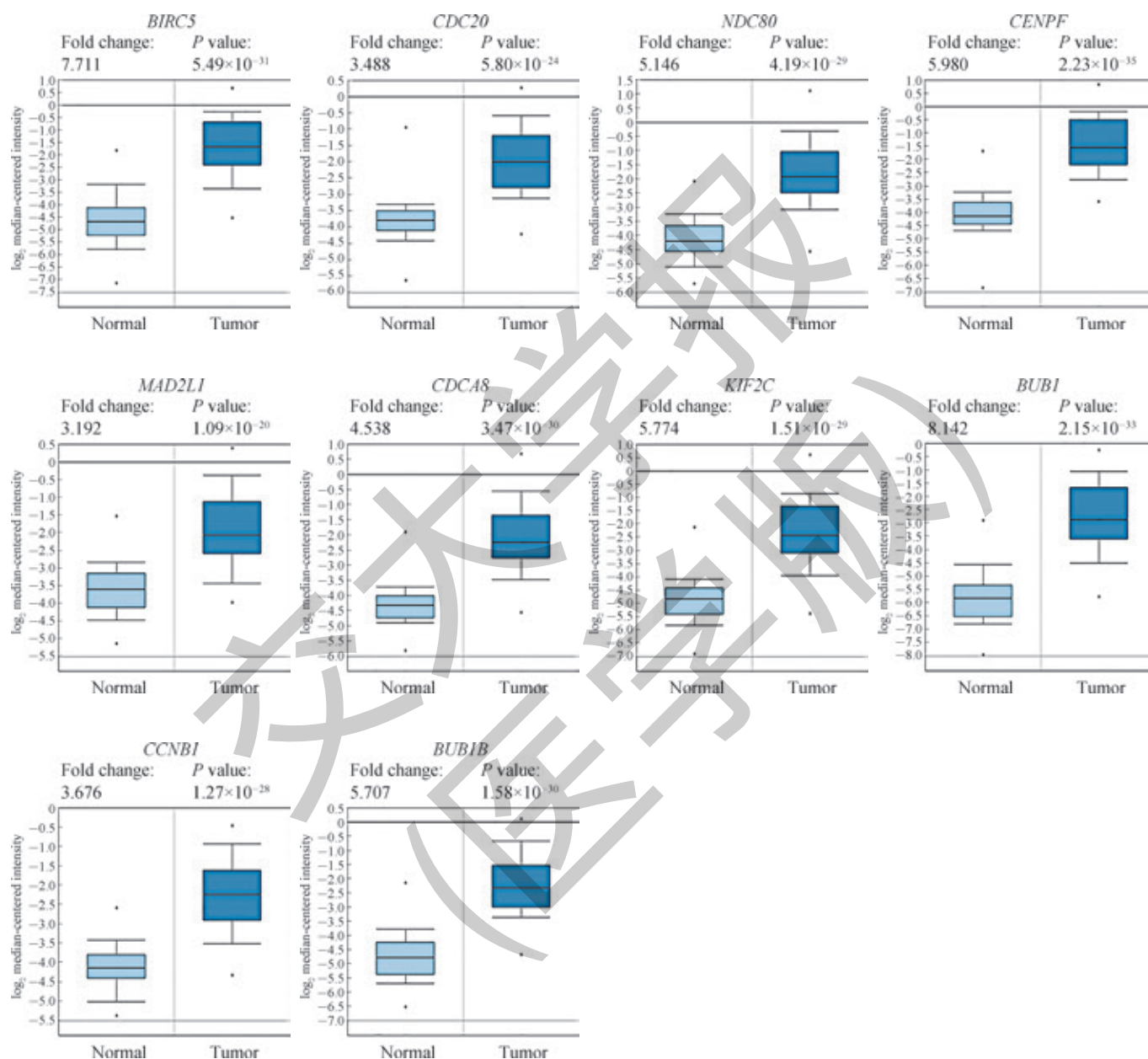
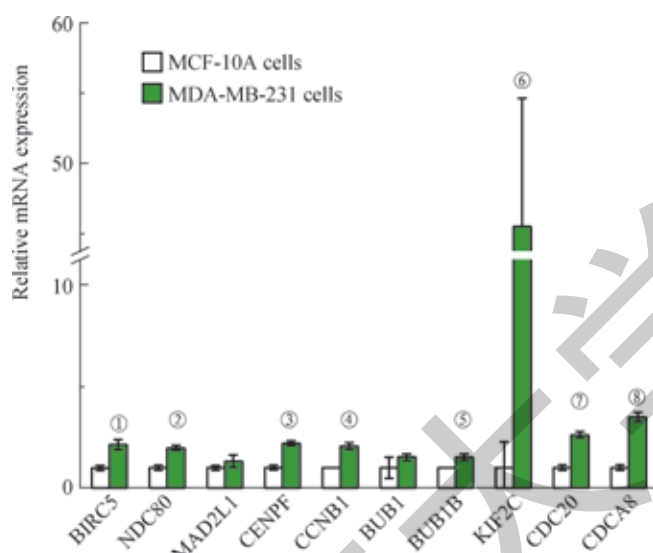


图 6 通过 Oncomine 数据库验证枢纽基因转录水平的表达

Fig 6 Validation of the expression of hub genes at mRNA level using Oncomine database

此外, 通过提取人乳腺癌细胞 MDA-MB-231 和人正常乳腺上皮细胞 MCF-10A 的 RNA 进行反转录和 qPCR, 验证了这些枢纽基因的表达水平。如图 7 结果所示, *BIRC5*、*NDC80*、*MAD2L1*、*CENPF*、*CCNB1*、*BUB1*、*BUB1B*、*KIF2C*、*CDC20* 和 *CDC48* 在乳腺癌细胞中的表达均高于人正常乳腺上皮细胞; 同时, 虽然 2 个枢纽基因 *MAD2L1* 与 *BUB1* 的表达差异无统计学意义, 但乳腺癌细胞中水平仍略高于正常乳腺上皮细胞。以上结果在细胞水平验证了各枢纽基因表达。



Note: ^① $P=0.011$, ^② $P=0.003$, ^③ $P=0.001$, ^④ $P=0.004$, ^⑤ $P=0.012$, ^⑥ $P=0.034$, ^⑦ $P=0.001$, ^⑧ $P=0.001$, compared with MCF-10A cells.

图 7 qPCR 检测乳腺癌细胞枢纽基因的表达

Fig 7 qPCR analysis of hub genes in breast cancer cells

3 讨论

本研究利用基于生存分析的生物信息学方法, 将基因表达数据与临床生存分析相结合, 筛选乳腺癌中的枢纽基因和关键通路。首先, 我们从 GEO 数据库下载了 3 个乳腺癌组织和癌旁组织的基因表达数据集, 筛选出了 585 个差异表达基因, 其中上调的基因有 211 个, 下调的基因有 374 个。然后用 Kaplan-Meier plotter 数据库筛选出了 262 个与乳腺癌患者总体生存期相关的差异表达基因。接着对这 262 个基因进行了 GO 功能分析, 结果显示这些基因主要与细胞分裂、细胞周期的调控以及染色体分离等生物学过程相关; KEGG 通路分析结果表明, 这些与总体生存期相关的差异表达基因主要富集在细胞周期、FoxO 信号通路和卵母细胞减数分裂等通路上。此外我们发现人类 T 细胞白血病病毒 1 感染信号通路在乳腺癌中失调, 但目前并无关于人类 T 细胞白血病病毒 1 感染信号通路与乳腺癌的报

道; 由于其在乳腺癌中的分子机制仍不明确, 因此需要进一步研究。以上筛选出来的通路可以为乳腺癌发生发展的分子机制研究提供依据。

此外, 我们通过构建 PPI 网络筛选出了 10 个乳腺癌的枢纽基因, 分别为 *BIRC5*、*CDC20*、*NDC80*、*CENPF*、*MAD2L1*、*CDC48*、*KIF2C*、*BUB1*、*CCNB1* 和 *BUB1B*, 经 Oncomine 数据库和 HPA 数据库以及 qPCR 验证, 它们在乳腺癌中均高表达, 并与乳腺癌患者较差的生存期相关。通过查阅文献发现这些基因主要参与了细胞有丝分裂染色体的分离和细胞周期的调控。我们还尝试在 Oncomine 数据库中研究这些枢纽基因的表达是否与乳腺癌分子分型、分期及恶性程度有关; 但由于数据库中数据的一些局限, 无法直接分析枢纽基因在不同分子分型乳腺癌的表达情况, 因此无法确认枢纽基因的表达与分子分型的相关性。但我们发现枢纽基因在 ER 阴性乳腺癌和 PR 阴性乳腺癌的表达量分别高于 ER 阳性乳腺癌和 PR 阳性乳腺癌的表达量; 在 HER2 阳性或阴性的乳腺癌中, 这些基因的表达却没有明显差别 (数据未展示), 提示枢纽基因的表达可能与分子分型有关。通过分析枢纽基因在不同分期的乳腺癌中的表达情况, 我们发现这些基因在 I_B 期的表达量明显低于其他分期的表达量, 而 *BUB1B* 在 III 期的表达量显著高于其他分期 (数据未展示)。 *MAD2L1*、*CDC20*、*CENPF*、*KIF2C*、*CCNB1*、*NDC80*、*BUB1*、*CDC48* 和 *KIF2C* 在各分期的表达量没有明显差异。上述发现说明枢纽基因可能可以作为特定的分子分型和分期治疗的判断依据。

有研究报道, *NDC80*、*MAD2L1*、*CDC20*、*BUB1*、*BIRC5* 和 *CCNB1* 在乳腺癌的发生发展中发挥了重要功能。 *NDC80* 基因编码的 NDC80 蛋白参与构成微管与动粒连接复合体, 为染色体正常分离所必需, 在细胞有丝分裂过程中起着至关重要的作用; *NDC80* 的异常表达会造成染色体的异常分离, 从而使染色体不稳定, 最终导致肿瘤的发生^[14]。多项研究^[15-16]表明 *NDC80* 在多种肿瘤中高表达并参与肿瘤的发生发展。 *NDC80* 抑制剂 TAI-95 可以在体外和体内抑制乳腺癌肿瘤生长, *NDC80* 有望成为乳腺癌的治疗靶点^[17]。 *MAD2L1* 和 *CDC20* 蛋白是纺锤体检查点 (spindle assembly checkpoint, SAC) 复合体的组成成分, SAC 负责确保姐妹染色单体的动粒和纺锤体结合并准确分离到 2 个子细胞中^[18]。研究^[19]显示通过 shRNA 敲除 *MAD2L1* 可以抑制乳腺癌细胞的生长和侵袭能力。 *CDC20* 的高表达与乳腺癌患者较差的预后相关, 且由它介导的 BTG3 相关核蛋白 (BTG3 associated nuclear protein, SMAR1) 降解可以促进乳腺癌细胞迁移和侵袭^[20-21]。

BUB1 是一类丝氨酸 / 苏氨酸蛋白激酶, 在有丝分裂和 DNA 损伤应答中发挥着重要的功能^[22]。Han 等^[23]发现, 敲低 *BUB1* 会降低肿瘤干细胞的潜能, *BUB1* 可能成为针对乳腺癌干细胞的治疗靶点。*BIRC5* 编码的蛋白是凋亡抑制 (inhibitor of apoptosis, IAP) 家族的成员之一, *BIRC5* 具有明显的抑制细胞凋亡的作用。Li 等^[24]认为 *BIRC5* 与乳腺癌患者进展及不良预后相关; 研究^[25-27]发现, *BIRC5* 是 miR-485-5p、ZIC 家族成员 1 (Zic family member 1, ZIC1) 和半乳糖凝集素 1 (galectin-1) 的靶点, 可能在乳腺癌中发挥重要作用。CCNB1 蛋白是调控细胞周期 G₂/M 过渡阶段所必需的; 研究^[28]表明, CCNB1 蛋白与侵袭程度 (肿瘤分级、肿瘤体积和淋巴结状态) 相关, 是乳腺癌的预后因素之一。

目前 *KIF2C*、*CENPF*、*CDCA8* 和 *BUB1B* 这 4 个枢纽基因与乳腺癌相关的研究还比较少, 具体分子机制有待进一步研究。*KIF2C* 和 *CENPF* 蛋白参与有丝分裂的染

色体分离。研究^[29]报道, *KIF2C* 在多种上皮型肿瘤中高表达, 并与肿瘤的分级、分期和预后相关。*CENPF* 基因编码一种与着丝粒-动粒复合体相关的蛋白, 其表达水平在有丝分裂前期增加, 并在后期开始时发生蛋白水解^[30-31]; *CENPF* 的上调与乳腺癌的不良预后相关, 可能是有临床意义的治疗靶点^[32]。*CDCA8* 是细胞周期分裂相关蛋白家族的一员, 此蛋白受细胞周期调控, 对染色质诱导的微管稳定和纺锤体形成是必需的; *CDCA8* 的高表达与乳腺癌患者的低生存率和较差预后相关^[33]。*BUB1B* 基因编码的蛋白是 SAC 复合体的重要成员, *BUB1B* 表达增加是恶性程度高的乳腺癌的特征之一^[34]。

综上所述, 本研究基于与总生存期相关的差异表达基因, 筛选了乳腺癌中的枢纽基因和关键通路; 基于生存期的枢纽基因可能更具有临床意义, 但这些枢纽基因的功能仍需进一步研究。

参 考 文 献

- [1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019[J]. CA Cancer J Clin, 2019, 69(1): 7-34.
- [2] Akram M, Iqbal M, Daniyal M, et al. Awareness and current knowledge of breast cancer[J]. Biol Res, 2017, 50(1): 33.
- [3] Zhou Z, Cheng Y, Jiang Y, et al. Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis[J]. Int J Biol Sci, 2018, 14(2): 124-136.
- [4] Lu TP, Tsai MH, Lee JM, et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women[J]. Cancer Epidemiol Biomarkers Prev, 2010, 19(10): 2590-2597.
- [5] Gautier L, Cope L, Bolstad BM, et al. affy: analysis of Affymetrix GeneChip data at the probe level[J]. Bioinformatics, 2004, 20(3): 307-315.
- [6] Olson NE. The microarray data analysis process: from raw data to biological significance[J]. NeuroRx, 2006, 3(3): 373-383.
- [7] Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies[J]. Nucleic Acids Res, 2015, 43(7): e47.
- [8] Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters[J]. OMICS, 2012, 16(5): 284-287.
- [9] Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored[J]. Nucleic Acids Res, 2011, 39(Database issue): D561-D568.
- [10] Chin CH, Chen SH, Wu HH, et al. cytoHubba: identifying hub objects and sub-networks for RNA-interactome[J]. BMC Syst Biol, 2014, 8(Suppl 4): S11.
- [11] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks[J]. Genome Res, 2003, 13(11): 2498-2504.
- [12] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks[J]. BMC Bioinformatics, 2003, 4: 2.
- [13] Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome[J]. Science, 2015, 347(6220): 1260419.
- [14] Ju LL, Chen L, Li JH, et al. Effect of NDC80 in human hepatocellular carcinoma[J]. World J Gastroenterol, 2017, 23(20): 3675-3683.
- [15] Meng QC, Wang HC, Song ZL, et al. Overexpression of NDC80 is correlated with prognosis of pancreatic cancer and regulates cell proliferation[J]. Am J Cancer Res, 2015, 5(5): 1730-1740.
- [16] Qu Y, Li J, Cai Q, et al. Hec1/Ndc80 is overexpressed in human gastric cancer and regulates cell growth[J]. J Gastroenterol, 2014, 49(3): 408-418.
- [17] Huang LY, Chang CC, Lee YS, et al. Activity of a novel Hec1-targeted anticancer compound against breast cancer cell lines *in vitro* and *in vivo*[J]. Mol Cancer Ther, 2014, 13(6): 1419-1430.
- [18] Kim Y, Choi JW, Lee JH, et al. Spindle assembly checkpoint MAD2 and CDC20 overexpressions and cell-in-cell formation in gastric cancer and its precursor lesions[J]. Hum Pathol, 2019, 85: 174-183.
- [19] Wang Z, Katsaros D, Shen Y, et al. Biological and clinical significance of *MAD2L1* and *BUB1*, genes frequently appearing in expression signatures for breast cancer prognosis[J]. PLoS One, 2015, 10(8): e0136246.
- [20] Karra H, Repo H, Ahonen I, et al. Cdc20 and securin overexpression predict short-term breast cancer survival[J]. Br J Cancer, 2014, 110(12): 2905-2913.
- [21] Paul D, Ghorai S, Dinesh US, et al. Cdc20 directs proteasome-mediated degradation of the tumor suppressor SMAR1 in higher grades of cancer through the anaphase promoting complex[J]. Cell Death Dis, 2017, 8(6): e2882.
- [22] Yang C, Wang H, Xu Y, et al. The kinetochore protein Bub1 participates in the DNA damage response[J]. DNA Repair (Amst), 2012, 11(2): 185-191.
- [23] Han JY, Han YK, Park GY, et al. Bub1 is required for maintaining cancer stem cells in breast cancer cell lines[J]. Sci Rep, 2015, 5: 15993.
- [24] Li S, Wang L, Meng Y, et al. Increased levels of LAPTM4B, VEGF and survivin are correlated with tumor progression and poor prognosis in breast cancer patients[J]. Oncotarget, 2017, 8(25): 41282-41293.
- [25] Wang M, Cai WR, Meng R, et al. miR-485-5p suppresses breast cancer progression and chemosensitivity by targeting survivin[J]. Biochem Biophys Res Commun, 2018, 501(1): 48-54.
- [26] Nam K, Son SH, Oh S, et al. Binding of galectin-1 to integrin β 1 potentiates drug resistance by promoting survivin expression in breast cancer cells[J]. Oncotarget, 2017, 8(22): 35804-35823.
- [27] Han W, Cao F, Gao XJ, et al. ZIC1 acts a tumor suppressor in breast cancer by targeting survivin[J]. Int J Oncol, 2018, 53(3): 937-948.
- [28] Aaltonen K, Amini RM, Heikkilä P, et al. High cyclin B1 expression is associated with poor survival in breast cancer[J]. Br J Cancer, 2009, 100(7): 1055-1060.
- [29] Gnjatich S, Cao Y, Reichelt U, et al. NY-CO-58/KIF2C is overexpressed in a variety of solid tumors and induces frequent T cell responses in patients with colorectal cancer[J]. Int J Cancer, 2010, 127(2): 381-393.
- [30] Liao H, Winkfein RJ, Mack G, et al. CENP-F is a protein of the nuclear matrix that assembles onto kinetochores at late G2 and is rapidly degraded after mitosis[J]. J Cell Biol, 1995, 130(3): 507-518.
- [31] Rattner JB, Rao A, Fritzler MJ, et al. CENP-F is a 400 kDa kinetochore protein that exhibits a cell-cycle dependent localization[J]. Cell Motil Cytoskeleton, 1993, 26(3): 214-226.
- [32] O'Brien SL, Fagan A, Fox EJ, et al. CENP-F expression is associated with poor prognosis and chromosomal instability in patients with primary breast cancer[J]. Int J Cancer, 2007, 120(7): 1434-1443.
- [33] Phan NN, Wang CY, Li KL, et al. Distinct expression of CDCA3, CDCA5, and CDCA8 leads to shorter relapse free survival in breast cancer patient[J]. Oncotarget, 2018, 9(6): 6977-6992.
- [34] Scintu M, Vitale R, Prencipe M, et al. Genomic instability and increased expression of *BUB1B* and *MAD2L1* genes in ductal breast carcinoma[J]. Cancer Lett, 2007, 254(2): 298-307.

[收稿日期] 2019-06-24

[本文编辑] 瞿麟平

