癌症与分子诊断研究，2020 年，第 9 卷，第 3 篇，第 1-14 页

研究报告
Research Report

联合 TCGA 和 GEO 数据库筛选乳腺癌易感基因

王建[1]　邵荣金[1]　龚伟达[1]　吕旭[1]　李金平[2*]
1 宜兴市肿瘤医院, 宜兴, 214200; 2 宁夏医科大学总院肿瘤医院, 银川, 750004
* 通信作者, ayou423x1@hotmail.com

摘　要　本研究是利用公共基因芯片数据库筛选乳腺癌的预后基因，预测和探索这些基因在乳腺癌进展中的可能机制和临床价值。首先，我们筛选了公共基因芯片数据库(gene expression omnibus, EO)GSE22820 和癌症基因组图谱(the cancer genome atlas,TCGA)乳腺癌数据库的重叠差异表达基因，联合 R 语言分析乳腺癌组织与癌旁正常组织差异表达的基因；其次，基于 STRING 数据库及 Cytoscape 软件构建蛋白质相互作用网络图，分析并识别了中枢基因和前三个模块；之后进行了更多的功能分析，包括基因本体(gene ontology, GO)和京都基因与基因组百科全书(kyoto encyclopedia of genes and genomes, KEGG)通路分析以及基因集富集分析 (gene set enrichment analysis, GSEA)，以研究这些基因的作用以及潜在的潜在机制；最后再进行了 Kaplan-Meier 分析和 Cox 比例风险分析，以阐明这些基因的诊断和预后效果。相关数据分析表明 15 个基因的表达水平与生存预后相关，高表达基因患者的总生存时间短于低表达患者(p<0.05)；Cox 比例风险分析表明这 3 个基因 UBE2T、ERCC6L 和 RAD51 是预后生存的独立因素 (p<0.05)；GSEA 分析表明在 UBE2T、ERCC6L 和 RAD51 基因中细胞周期、基础转录因子和卵母细胞减数分裂明显富集。最终，我们得出结论，这三种基因标志物的高表达是乳腺癌预后不良因素，可作为预测乳腺癌患者转移和预后的有效生物标志物。
关键词　生物信息学, 生物标志物, 乳腺癌, 预后标志物

The Identification of New Biomarkers for Breast Cancer: a Study Based on TCGA and GEO Datasets

Wang Jian[1]　Shao Rongjin[1]　Gong Weida[1]　Lv Xu[1]　Li Jinping[2*]
1 Cancer Hospital of Yixing City, Yixin, 214200; 2 Cancer Hospital of Ningxia Medical University, Yinchuan, 750004
* Corresponding author, ayou423x1@hotmail.com
DOI: 10.5376/cmdr.cn.2020.09.0003

Abstract　The purpose of this study was to forecast and explore the possible mechanism and clinical value of genetic markers in the evolution of breast cancer with a merged database to screen the prognostic genes of breast cancer. First,we screened the overlapped differentially expressed genes (DEGs) of GSE22820 and TCGA breast caner datasets by R language. Second, subsequent protein-protein interactions network analysis recognized the hub genes and top three modules among these DEGs in Cytoscape software. Then more functional analysis including Gene Ontology and KEGG pathway analysis and gene set enrichment analysis were processed to investigate the role of these genes and potential underlying mechanisms in BC. And finally Kaplan-Meier analysis and Cox hazard ratio analysis were performed to elucidate the diagnostic and prognostic effects of these genes. Analysis of relevant data shows that the expression levels of fifteen genes were interrelated with survival prognosis, and the overall survival time of the patients with high expression of the gene was shorter than those with low expression (p<0.05).

收稿日期：2020 年 2 月 6 日；接受日期：2020 年 3 月 29 日；发表日期：2020 年 5 月 25 日
引用格式: 王建, 邵荣金, 龚伟达, 吕旭, 李金平, 2020, 联合 TCGA 和 GEO 数据库筛选乳腺癌易感基因, 癌症与分子诊断研究, 9(3): 1-14 (doi: 10.5376/cmdr.cn.2020.09.0003) (Wang J., Shao R.J., Gong W.D., LV X., and Li J.P., 2020, The identification of new biomarkers for breast cancer: a study based on TCGA and GEO datasets, Aizheng Yu Fenzi Zhenduan Yanjiu (Cancer and Molecular Diagnosis Research), 9(3): 1-14 (doi: 10.5376/cmdr.cn.2020.09.0003))

But the Cox proportion hazard ratio analysis that the 3 genes were Significance, UBE2T, ERCC6L, and RAD51 could be considered independent factors for prognosis survival(p<0.05). Gene set enrichment analysis showed that the cell cycle, basic transcription factors and oocyte meiosis were significantly enriched in UBE2T, ERCC6L and RAD51 genes. Finally, we got a Conclusion that The high expression of three genetic markers is a poor prognostic factor for breast cancer and can be used as an effective biomarker to predict metastasis and prognosis of breast cancer patients.

Keywords　Bioinformatics, Biomarker, Breast cancer, Prognostic markers

乳腺癌是全世界女性中最常见的癌症类型之一，也是与癌症相关的死亡的主要原因(DeSantis et al., 2017; Torre et al., 2017)。尽管随着技术的发展，乳腺癌的治疗取得了巨大进展，但仍然是女性恶性肿瘤死亡的常见原因(Clegg et al., 2009)。在 2018 年诊断出近 210 万新病例，占女性所有癌症的 25%，导致全球每年超过 600 000 例死亡(Bray et al., 2018)。在美国，它是与癌症相关的死亡的第二大最常见原因(Rugo et al., 2016)。因此，早期诊断乳腺癌是提高患者生存率的关键。

近年来，关于癌基因、抑癌基因，各种异源蛋白和肿瘤抗原的研究很多，针对特定肿瘤标志物的药物的开发和应用也取得了一定进展 (Baselga et al., 2017)。越来越多的证据表明，分子靶向治疗是癌症治疗的有前途的研究方向(Koshiba, 2016)。因此，迫切需要揭示乳腺癌发生和发展的详细分子机制。

随着微阵列技术的广泛应用，公共数据库用户可以获得大量数据。基于整合生物信息学方法，发现了有效、新颖和可靠的分子标记物。癌症基因组图谱(the cancer genome atlas, TCGA) 是最广泛测序结果的数据库，可为研究人员提供有关肿瘤分期、患者年龄、生存、转移、性别和相应临床数据的全面癌症基因组数据集。公共基因芯片数据库(Gene expression omnibus, GEO) 是美国国家生物技术信息中心(National center of biotechnology information, NCBI)中全面的基因表达数据库，该数据库是世界上最大的基因芯片数据库之一。

我们从 TCGA 和 GEO 数据库中寻找可靠的生物标志物，发现 UBE2T、ERCC6L、RAD51 可能是乳腺癌的生物标志物，它们均与乳腺癌患者的预后有关。我们的研究可能是未来乳腺癌治疗的新的诊断标志物和潜在治疗靶标。

1 结果与分析

1.1 差异表达基因的筛选

通过从 TCGA 下载乳腺癌数据集获得了基因表达谱(1 066 例肿瘤组织和 112 个正常组织)。使用 R×64 3.6.0 软件 Limma 软件包分析了 TCGA 乳腺癌数据集。总共鉴定出 2 775 个差异表达的 mRNA (1 113 个上调，1662 个下调) (图 1)。随后，使用火山图分析直接鉴定 mRNA 的差异基因(图 2)。筛选来自 GEO 的 GSE22820 微阵列数据集，筛选标准 p<0.05 和 |FC|>2，其中包括 639 个上调基因和 476 个下调基因，以作进一步研究。最后，我们得到了两个数据集的重叠差异表达基因，包括 182 个上调基因和 267 个下调基因(图 3)。

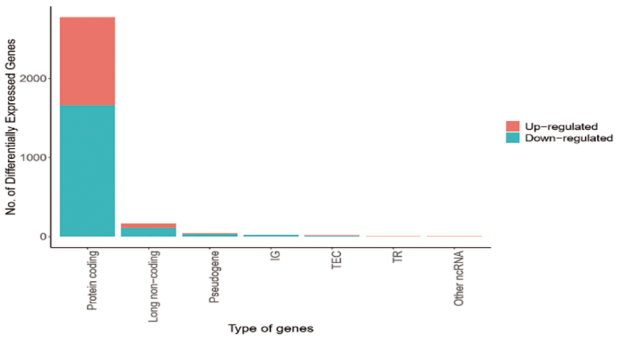

图 1 TCGA 乳腺癌数据库中不同表达的基因

Figure 1 Differentially expressed genes from TCGA breast caner

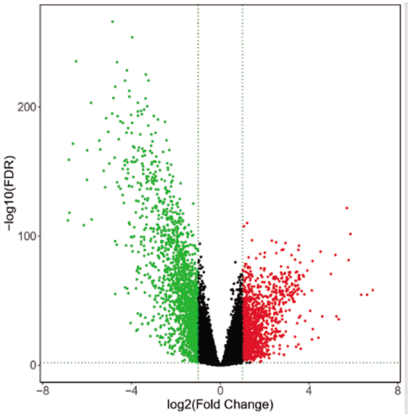

图 2 TCGA 乳腺癌数据库中差异表达基因的火山图

注: p<0.05 和 |FC|>2 被用作截止标准

Figure 2 Volcano plot of differentially expressed genes from TCGA breast cancer

Note: p<0.05 and |FC|>2 were used as the cut off criteria

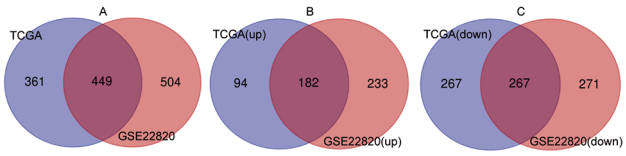

图 3 TCGA 和 GEO 数据库差异表达基因的筛选

注: A: TCGA 乳腺癌和 GSE22820 数据集的差异表达基因; B: 上调基因; C: 下调基因

Figure 3 Screening of differentially expressed genes from TCGA and GEO databases

Note: A.: Differentially expressed genes in TCGA breast cancer and GSE22820 databases; B: Upregulated genes; C: Downregulated genes

1.2 差异表达基因功能富集及通路分析

通过功能富集分析，我们发现这些差异表达基因主要富含蛋白质细胞外基质的细胞成分和细胞外基质成分，包括细胞 - 细胞连接(cell-cell junction)，基底膜 (basement membrane)，基底膜和纺锤体(basement membrane and spindle)。关于生物过程(biological process, BP)，差异表达基因富含细胞外结构组织(extracellular structure organization)，细胞外基质组织(extracellular matrix organization)，泌尿生殖系统发育(urogenital system development)，肾脏系统发育(renal system development)，肾脏发育(kidney development)，骨化(ossification)和姐妹染色单体分离(sister chromatid segregation)。至于分子功能，糖胺聚糖结合(glycosaminoglycan binding) 和跨膜受体蛋白激酶活性 (transmembrane receptor protein kinase activity)包含在前十个富集类别中，p<0.05 (图 4) (表 1)。

在进行 KEGG 通路分析时，我们发现这些差异表达基因主要富集于黏着斑 (focal adhesion)，ECM-受体相互作用(ECM-receptor interaction)，PI3K-Akt 信号通路，轴突导向(Axon guidance)，细胞周期(Cell cycle)，脂肪细胞中脂肪降解调节(Regulation of lipolysis in adipocytes)，Rap1 信号通路，人乳头瘤病毒感染(Human papillomavirus infection)，MAPK 信号通路和 PPAR 信号通路(图 5; 表 2)。

1.3 蛋白质相互作用网络的构建和关键基因的筛选

通过将重叠的差异基因放入在线工具 STRING 中，我们获得了这些基因的蛋白质相互作用网络图(Protein-protein interaction,PPI)。随后，将这些 PPI 网络图导入 Cytoscape 软件中，以进一步体现这些基因编码蛋白质之间的相互作用(图 6)，并且我们在 PPI 网络中检测顶层关键基因。具有最高连通度的前 20 个基因被视为顶层关键基因(表 7)。此外，我们使用 Cytoscape 中的 MCODE 应用程序提取了 PPI 网络中最重要的三个模块(图 8)。

1.4 预后生存 Kaplan–Meier 法分析

为了验证这些顶层关键基因和前三个模块，进行了 R×64 3.6.0 软件中的 Kaplan-Meier 分预后生存析。结果显示：SLC27A6、CXCL2、CX3CL1、SAA1、NTRK2、ACAN、SDC1、KRT14、EXO1、MND1、RAD-54L、MCM10、UBE2T、ERCC6L 和 RAD51 的表达水平较高的患者的总生存期较差，而其他患者则无明显变化(图 9)。

1.5 临床病理特征的 Cox 风险比例分析

随后，我们进行了 Cox 风险比例分析，以确认这些关键基因的预后价值。分析结果表明：UBE2T (HR, 1.01; p=0.039)，ERCC6L (HR，1.16; p=0.008)和 RAD51 (HR, 1.07; p=0.031)的表达状态与患者的总

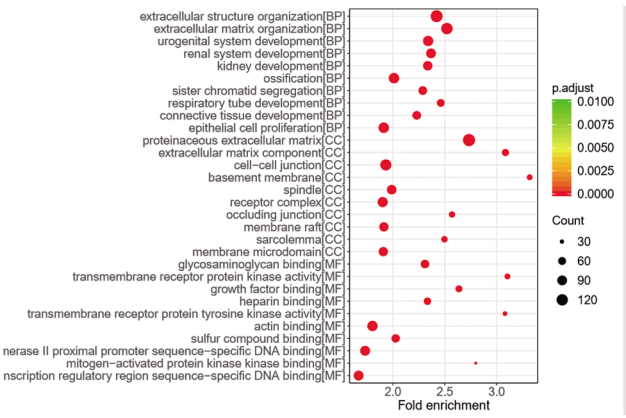

图 4 差异表达基因功能分析

Figure 4 Gene ontology analysis of differentially expressed genes

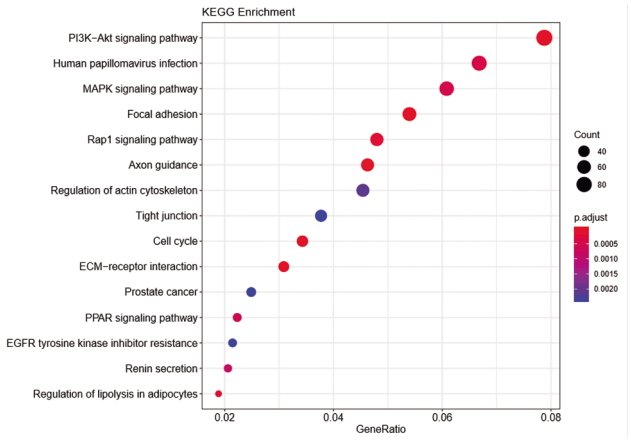

图 5 差异表达基因 KEGG 通路分析

Figure 5 KEGG pathway analysis of differentially expressed genes

表 1 乳腺癌差异表达基因功能分析

Table 1 Gene ontology analysis of DEGs in breast cancer

类别 Category	术语 Term	计数 Count	*p* 值 *p*-value	q 值 q-value
生物过程 Biological processes				
GOTERM_BP_FAT	GO:0043062~ 细胞外结构组织 GO:0043062~extracellular structure organization	138	2.70E-24	1.21E-20
GOTERM_BP_FAT	GO:0030198~ 细胞外基质组织 GO:0030198~extracellular matrix organization	124	1.18E-23	2.64E-20
GOTERM_BP_FAT	GO:0001655~ 泌尿生殖系统发育 GO:0001655~urogenital system development	101	5.76E-17	8.57E-14
GOTERM_BP_FAT	GO:0072001~ 肾系统发育 GO:0072001~renal system development	91	8.93E-16	9.97E-13
GOTERM_BP_FAT	GO:0001822~ 肾脏发育 GO:0001822~kidney development	85	1.87E-14	1.67E-11
GOTERM_BP_FAT	GO:0001503~ 骨化 GO:0001503~ossification	104	1.16E-12	8.66E-10
GOTERM_BP_FAT	GO:0000819~ 姐妹染色单体分离 GO:0000819~sister chromatid segregation	73	3.96E-12	2.53E-09
GOTERM_BP_FAT	GO:0030323~ 呼吸管发育 GO:0030323~respiratory tube development	59	1.70E-11	9.49E-09
GOTERM_BP_FAT	GO:0061448~ 结缔组织发育 GO:0061448~connective tissue development	72	2.12E-11	1.05E-08
GOTERM_BP_FAT	GO:0050673~ 上皮细胞增殖 GO:0050673~epithelial cell proliferation	105	2.47E-11	1.10E-08
细胞成分 Cellular components				
GOTERM_CC_FAT	GO:0005578~ 蛋白质细胞外基质 GO:0005578~proteinaceous extracellular matrix	143	3.25E-31	1.81E-28
GOTERM_CC_FAT	GO:0044420~ 细胞外基质成分 GO:0044420~extracellular matrix component	52	1.19E-14	3.31E-12
GOTERM_CC_FAT	GO:0005911~ 细胞间连接 GO:0005911~cell-cell junction	118	7.11E-13	1.32E-10
GOTERM_CC_FAT	GO:0005604~ 基底膜 GO:0005604~basement membrane	38	1.55E-12	2.16E-10
GOTERM_CC_FAT	GO:0005819~ 纺锤体 GO:0005819~spindle	87	1.54E-10	1.71E-08
GOTERM_CC_FAT	GO:0043235~ 复合体受体 GO:0043235~receptor complex	97	2.01E-10	1.87E-08
GOTERM_CC_FAT	GO:0070160~ 闭合连接 GO:0070160~occluding junction	43	2.11E-09	1.67E-07
GOTERM_CC_FAT	GO:0045121~ 膜筏 GO:0045121~membrane raft	82	3.65E-09	2.29E-07
GOTERM_CC_FAT	GO:0042383~ 肌膜炎 GO:0042383~sarcolemma	44	3.81E-09	2.29E-07
GOTERM_CC_FAT	GO:0098857~ 膜微区 GO:0098857~membrane microdomain	82	4.23E-09	2.29E-07

续表 1

Continuing table 1

类别 Category	术语 Term	计数 Count	*p* 值 *p*-value	q 值 q-value
分子功能 Molecular functional				
GOTERM_MF_FAT	GO:0005539~ 糖胺聚糖结合 GO:0005539~glycosaminoglycan binding	71	4.77E-12	4.53E-09
GOTERM_MF_FAT	GO:0019199~ 跨膜受体蛋白激酶活性 GO:0019199~transmembrane receptor protein kinase activity	39	1.84E-11	8.72E-09
GOTERM_MF_FAT	GO:0019838~ 生长因子结合 GO:0019838~growth factor binding	49	5.71E-11	1.81E-08
GOTERM_MF_FAT	GO:0008201~ 肝素结合 GO:0008201~heparin binding	55	7.75E-10	1.84E-07
GOTERM_MF_FAT	GO:0004714~ 跨膜受体蛋白酪氨酸激酶活性 GO:0004714~transmembrane receptor protein tyrosine kinase activity	32	1.45E-09	2.46E-07
GOTERM_MF_FAT	GO:0003779~ 肌动蛋白结合 GO:0003779~actin binding	102	1.56E-09	2.46E-07
GOTERM_MF_FAT	GO:1901681~ 硫化合物结合 GO:1901681~sulfur compound binding	68	6.35E-09	8.61E-07
GOTERM_MF_FAT	GO:0000982~ 转录因子活性，RNA 聚合酶 II 近端启动子序列特异性 DNA 结合 GO:0000982~transcription factor activity, RNA polymerase II proximal promoter sequence specific DNA binding	94	5.37E-08	6.37E-06
GOTERM_MF_FAT	GO:0031434~ 有丝分裂原激活的蛋白激酶激酶结合 GO:0031434~mitogen-activated protein kinase kinase binding	28	1.70E-07	1.79E-05
	GO:0001228~ 转录激活子活性，RNA 聚合酶 II 转录调控区序列特异性 DNA 结合 GO:0001228~transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific DNA binding	94	3.28E-07	3.12E-05

注: BP: 生物过程; CC: 细胞成分; GO: 基因本体; MF: 分子功能

Note: BP: Biological process; CC: Cellular component; GO: Gene ontology; MF: Molecular function.

体生存率相关。为了确保这三个关键基因在乳腺癌组织中的表达水平，我们提取了数据并进行分析并绘制了图表。在肿瘤组织中 UBE2T，ERCC6L 和RAD51 的表达增加，正常组织被下调。单变量分析表明，除性别与乳腺癌患者的总体生存率没有显着相关性，其他都有相关性。之后我们分别对每个基因进行了多变量分析后，除年龄之外，所有变量均与乳腺癌患者的总体生存率无显着相关性(图 10)。但是，单因素及多因素分析都表达诊断三个关键基因有统计学意义，最后，我们得出 UBE2T、ERCC6L 和 RAD51 表达水平可被视为预后生存的独立因素(表 3)。

1.6 差异基因功能富集分析

使用 GSEA 分析，我们发现在 UBE2T、ERCC6L 和 RAD51 基因中细胞周期(cell cycle)、基础转录因子 (basal transcription factors) 和卵母细胞减数分裂(oocyte meiosis)明显富集。此外，p53 信号传导途径、DNA 复制(DNA replication)、蛋白酶体(proteasome)、嘌呤代谢 (purine metabolism) 也在三个基因中高表达。另外还发现了下调的 UBE2T 基因在 MARK 信号通路中富集(图 11)。

2 讨论

乳腺癌是世界上最常见的恶性肿瘤之一(Ghoncheh et al., 2016)。由于早期临床症状较少，因此绝大多数乳腺癌患者诊断明确时都是晚期。所以，探讨乳腺癌的发病机制和发展过程，寻找有效的肿瘤标志物具有重要意义(Park et al., 2010, Januskeviciene and

表 2 差异表达基因 KEGG 通路分析

Table 2 KEGG pathway analysis of differentially expressed genes in breast cancer

类别 Category	术语 Term	P 值 P value	P 值矫正 p. adjust	计数 Count
KEGG 通路 KEGG pathway	hsa04510~ 局灶性粘连 hsa04510~Focal adhesion	1.08E-09	1.70E-07	63
	hsa04512~ ECM- 受体相互作用 hsa04512~ECM-receptor interaction	1.09E-09	1.70E-07	36
	hsa04151~PI3K-Akt 信号通路 hsa04151~PI3K-Akt signaling pathway	1.72E-08	1.79E-06	92
	hsa04360~ 轴突指导 hsa04360~Axon guidance	1.59E-07	1.24E-05	54
	hsa04110~ 细胞循环 hsa04110~Cell cycle	7.13E-07	4.45E-05	40
	hsa04923~ 调节脂肪细胞中的脂肪分解 hsa04923~Regulation of lipolysis in adipocytes	4.79E-06	0.000223	22
	hsa04015~Rap1 信号通路 hsa04015~Rap1 signaling pathway	5.00E-06	0.000223	56
	hsa05165~ 人乳头瘤病毒感染 hsa05165~Human papillomavirus infection	1.15E-05	0.000448	78
	hsa04010~MAPK 信号通路 hsa04010~MAPK signaling pathway	1.49E-05	0.000517	71
	hsa03320~PPAR 信号通路 hsa03320~PPAR signaling pathway	1.95E-05	0.000609	26

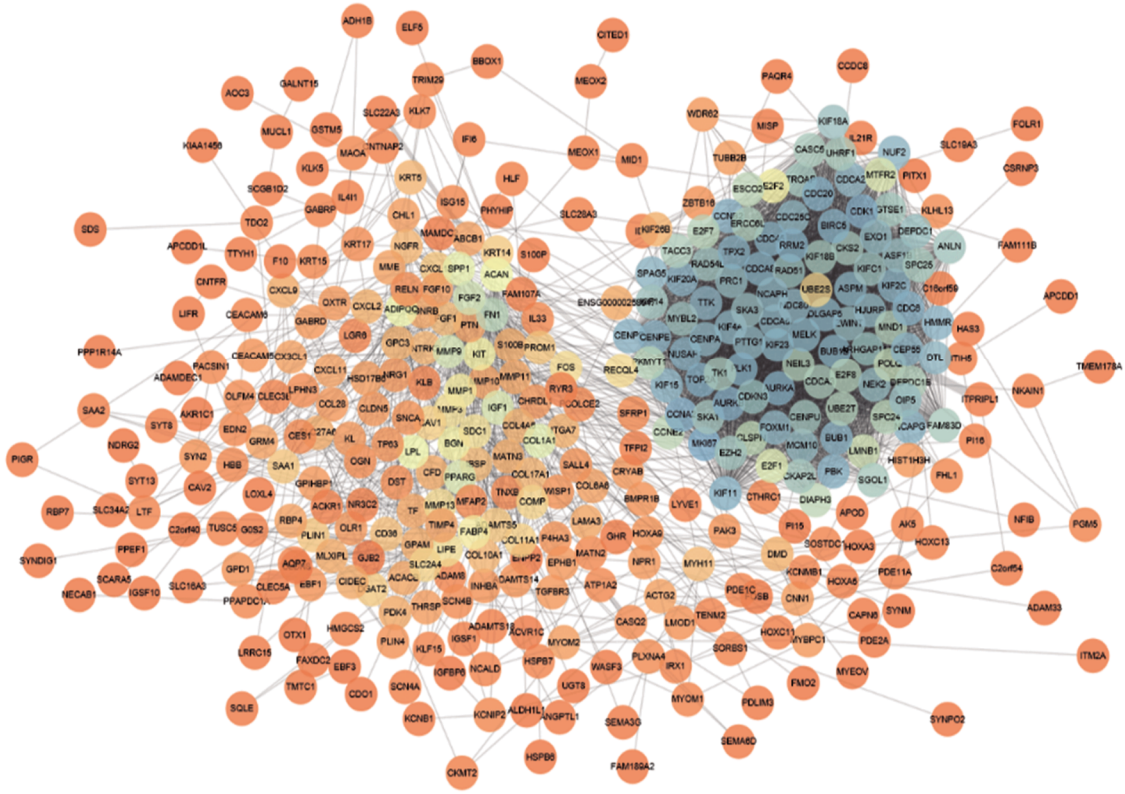

图 6 重叠差异表达基因的蛋白质互作网络分析

Figure 6 PPI network analysis of differentially expressed genes in breast cancer

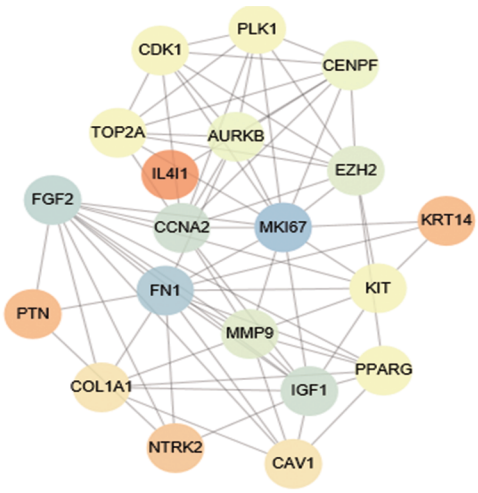

表 7 蛋白质互作网络中顶层前 20 个基因

Figure 7 Top 20 genes in PPI network

Petrikaite, 2019)。

这项研究我们一共发现 499 个乳腺癌差异表达基因，并且，我们获得了这些基因的蛋白质互作网络图。随后，我们筛选了网络图中的前 20 个基因和前三个模块。我们对这些关键基因的整体生存分析以及这些基因在正常组织和肿瘤组织中的表达分析。结果表明这有 15 个基因的表达水平与生存预后相关，基因低表达的患者的总生存时间长于基因高表达的患者。但是 COX 风险比例分析提示只有 UBE2T、ERCC6L 和 RAD51 这 3 个基因被认为是乳腺癌的独立预后指标。这些研究可能为乳腺癌提供新的诊断方法和治疗目标，从而可以改善乳腺癌患者的预后。

泛素结合酶 E2T (ubiquitin-conjugating enzyme E2T, UBE2T)是泛素结合酶 E2 家族的成员。早期的研究发现 UBE2T 可以通过调节范可尼贫血互补群基因 D2 (fanconi anemia complementation group D2, FANCD2) 单泛素化水平来影响细胞 DNA 损伤修复程序并触发肿瘤发生(Machida et al., 2006)。UBE2T 在多种肿瘤中均上调，例如在肺癌、前列腺癌、膀胱癌、鼻咽癌、乳腺癌和骨肉瘤中，UBE2T 的上调可调

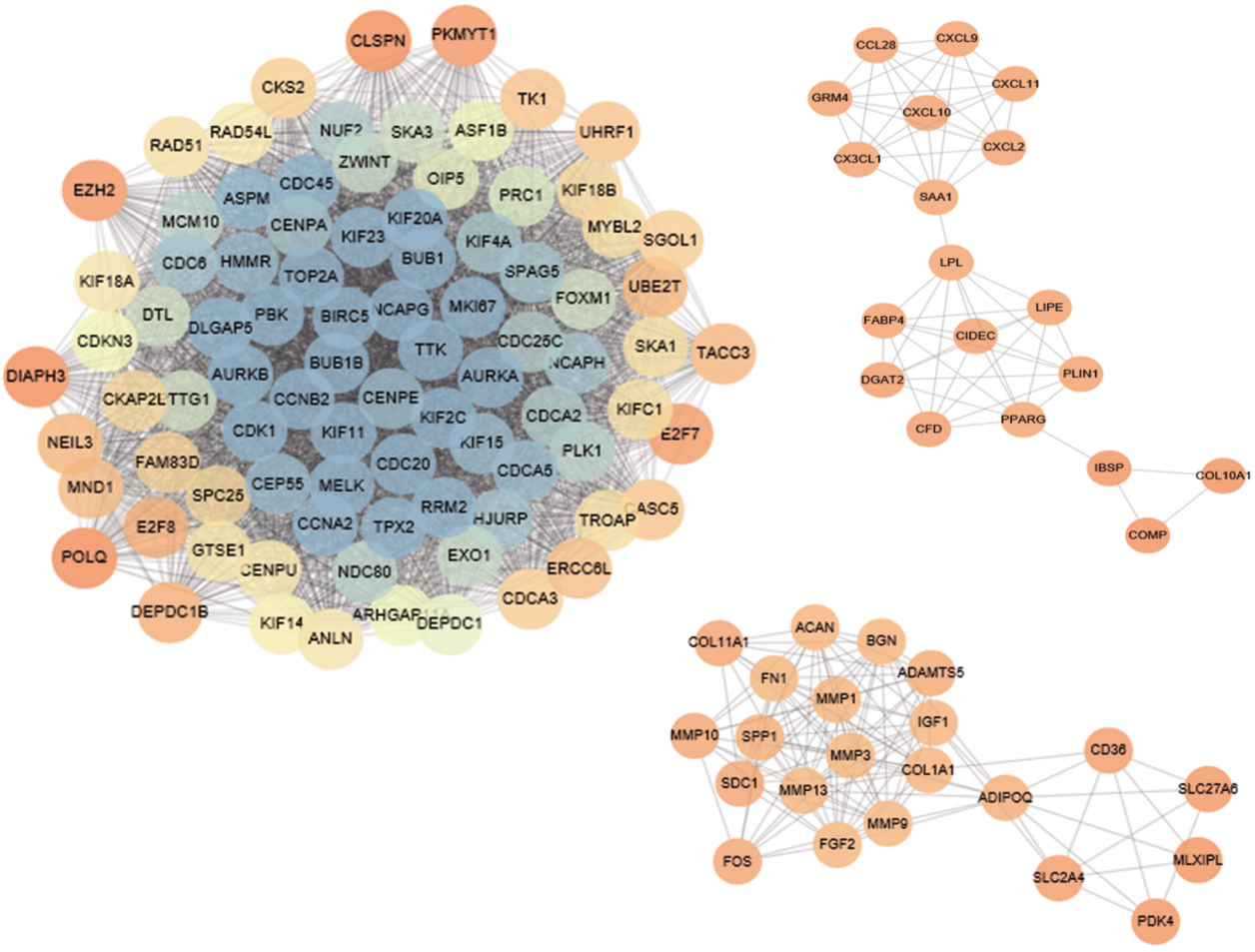

图 8 蛋白质互作网络图中最重要的 3 个模块

Figure 8 The most important first three modules in PPI network

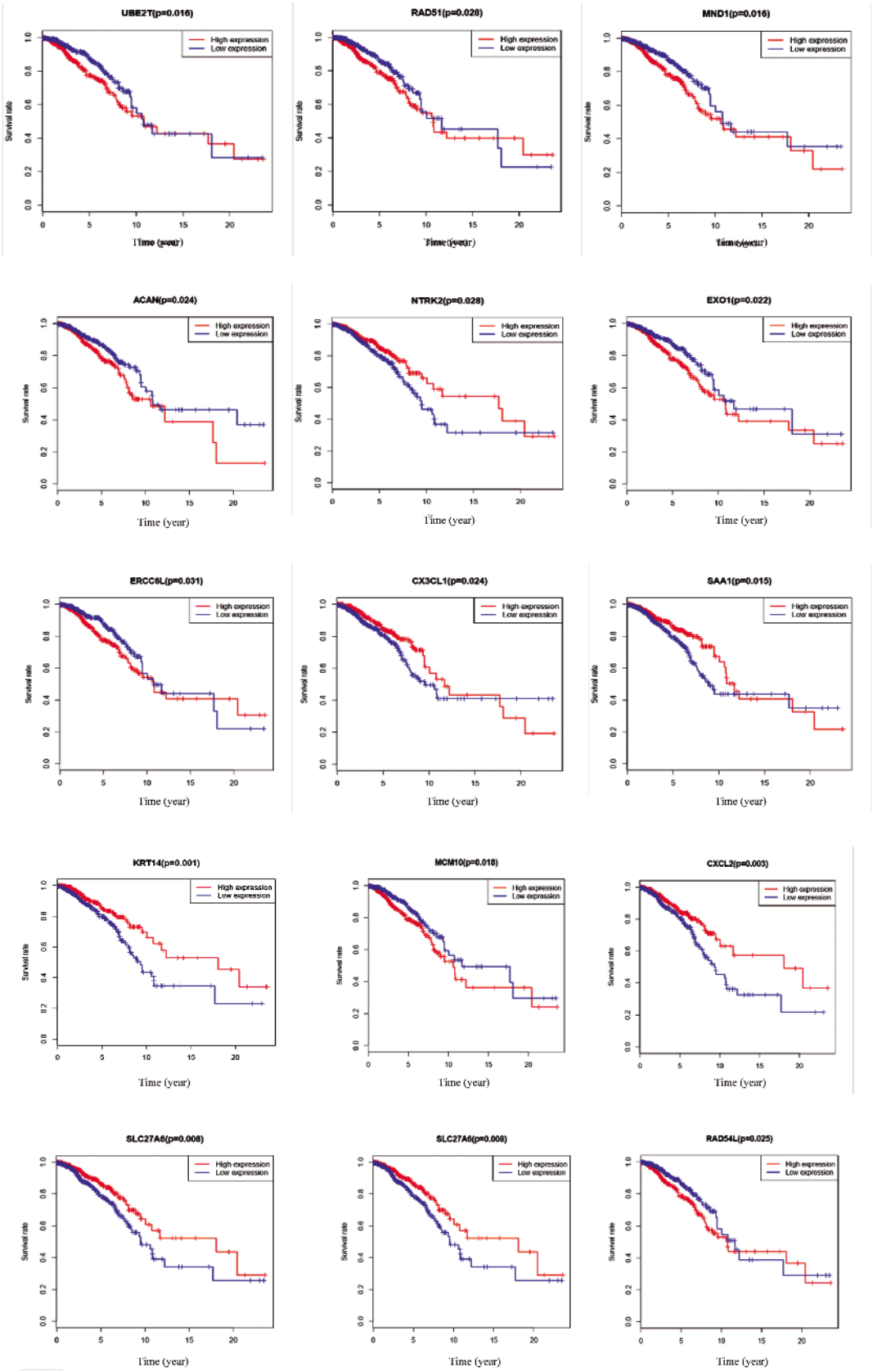

图 9 15 个差异表达基因的 Kaplan-Meier 分析

注: p<0.05 有统计学意义

Figure 9 Kaplan-Meier analysis of 15 differentially expressed genes

Note: p<0.05 was considered as statistically significant

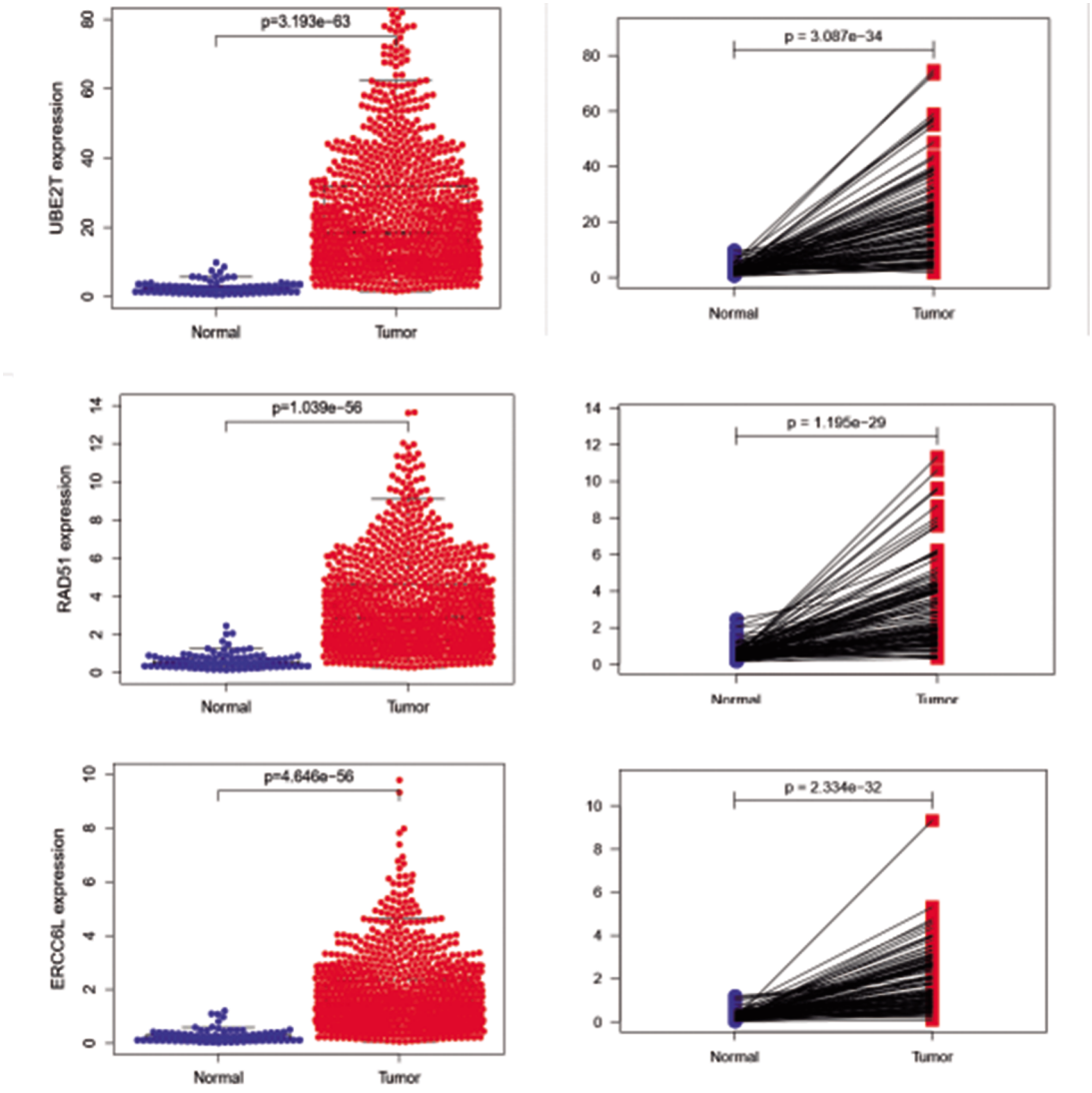

图 10 *UBE2T*, *ERCC6L* 和 *RAD51* 在癌组织和正常组织的表达分析

Figure 10 Expression levels of *UBE2T*, *ERCC6L* and *RAD51* in normal tissues and tumor tissues

节细胞的增殖、侵袭、转移、细胞周期和细胞凋亡(Alpi et al., 2016)。Ueki 等 (Ueki et al., 2009) 发现 UBE2T 与 BRCA1 / BARD1 (BRCA1-associated RING domain1)在乳腺癌细胞中共同表达，并能抑制乳腺癌中 BRCA1 基因的表达，从而促进乳腺癌细胞的侵袭和转移。沉默 UBE2T 表达可以上调 BRCA1，推测出 UBE2T 或其抑制剂可能是乳腺癌治疗的靶标。从浸润前导管内癌进展到浸润性导管癌的跨物种基因组学分析表明，这一过程伴随 UBE2T 基因改变(Colak et al., 2013)。perez 等(Perez-Pena et al., 2017)发现在基底样和管腔型乳腺癌患者中 UBE2T 过度表达与发病结果相关，该基因在约 12%的乳腺肿瘤中被扩增。

切除修复交叉互补 6 (excision repair cross-complementation group 6, ERCC6) 基因是与 SWI/SNF 复合物相关的三磷酸腺苷酶家族成员之一，与许多疾病相关 (Xu et al., 2014 , Liu et al., 2016)。Nielsen 等(Nielsen et al., 2015) 鉴定到 ERCC6L 与拓扑异构酶 II 在有丝分裂中协同作用，促进姐妹染色单体分离。研究表明，ERCC6L 在多种类型的人类实体瘤中高度表达，因此，它被认为是癌症治疗的潜在目标(Santamaria et al., 2007)。Liu 等(Liu et al., 2013)揭示了 DNA 修复基因 ERCC6rs1917799 多态性与中国人群的胃癌风险有关。 ERCC6 多态性还与口腔癌、肺癌、膀胱癌和结肠直肠癌的易感性有关(Chang et al., 2009 , Ma et al., 2009 , Ramaniuk et al., 2014)。PU 等(Pu et al., 2017)证实，ERCC6L 高表达与乳腺癌和肾癌的较差的临床存活率显着相关。最近的研究还观察到，ERCC6L 在 91.51%的乳腺癌患者中高表达(Liu et al., 2018)。这些发现表明，ERCC6L 可能是参与

表 3 *UBE2T*, *ERCC6L* 和 *RAD51* 的 Cox 风险比例分析

Table 3 Cox risk ratio analysis of *UBE2T*, *ERCC6L* and *RAD51*

变量 Variable	单因素分析 Univariate analysis			多因素分析 Multivariate analysis		
	风险比 Hazard ratio	95%置信区间 95%Confidence intervals	*p* 值 *p* value	风险比 Hazard ratio	95%置信区间 95%Confidence intervals	*p* 值 *p* value
UBE2T mRNA						
年龄(年) Age(year)	1.03	1.02-1.05	0.000	1.04	1.02-1.05	0.000
性别(男 / 女) Gender (male/female)	0.89	0.12-6.36	0.905	0.45	0.06-3.31	0.432
分期(IV/III/II/I) Stage (IV/III/II/I)	2.11	1.66-2.69	0.000	1.59	0.94-2.70	0.085
T 分期(T4/T3/T2/T1) T classification (T4/T3/T2/T1)	1.46	1.17-1.82	0.000	0.97	0.71-1.32	0.837
N 分期(N3/N2/N1/N0) N classification (N3/N2/N1/N0)	1.70	1.41-2.06	0.000	1.25	0.92-1.69	0.148
M 分期(M1/M0) M classification (M1/M0)	6.52	3.65-11.65	0.000	1.53	0.66-3.55	0.327
UBE2T(高表达 / 低表达) UBE2T (high/low expression)	1.01	1.00-1.01	0.039	1.21	1.02-1.43	0.031
ERCC6L mRNA						
年龄(年) Age(year)	1.03	1.02-1.05	0.000	1.04	1.02-1.05	0.000
性别(男 / 女) Gender (male/female)	0.89	0.12-6.36	0.905	0.51	0.07-3.75	0.511
分期(IV/III/II/I) Stage (IV/III/II/I)	2.11	1.66-2.69	0.000	1.61	0.95-2.77	0.078
T 分期(T4/T3/T2/T1) T classification (T4/T3/T2/T1)	1.46	1.17-1.82	0.000	0.96	0.70-1.31	0.805
N 分期(N3/N2/N1/N0) N classification (N3/N2/N1/N0)	1.70	1.41-2.06	0.000	1.23	0.91-1.66	0.185
M 分期(M1/M0) M classification (M1/M0)	6.52	3.65-11.65	0.000	1.68	0.72-3.94	0.233
ERCC6L(高表达 / 低表达) ERCC6L (high/low expression)	1.16	1.04-1.29	0.008	1.54	1.14-2.01	0.002
RAD51 mRNA						
年龄(年) Age(year)	1.03	1.02-1.05	0.000	1.04	1.02-1.05	0.000
性别(男 / 女) Gender (male/female)	0.89	0.12-6.36	0.905	0.44	0.06-3.25	0.421
分期(IV/III/II/I) Stage (IV/III/II/I)	2.11	1.66-2.69	0.000	1.65	0.97-2.82	0.065
T 分期(T4/T3/T2/T1) T classification (T4/T3/T2/T1)	1.46	1.17-1.82	0.000	0.95	0.70-1.30	0.761
N 分期(N3/N2/N1/N0) N classification (N3/N2/N1/N0)	1.70	1.41-2.06	0.000	1.23	0.91-1.66	0.186
M 分期(M1/M0) M classification (M1/M0)	6.52	3.65-11.65	0.000	1.65	0.70-3.86	0.249
RAD51(高表达 / 低表达) RAD51 (high/low expression)	1.07	1.00-1.14	0.031	1.39	1.09-1.77	0.008

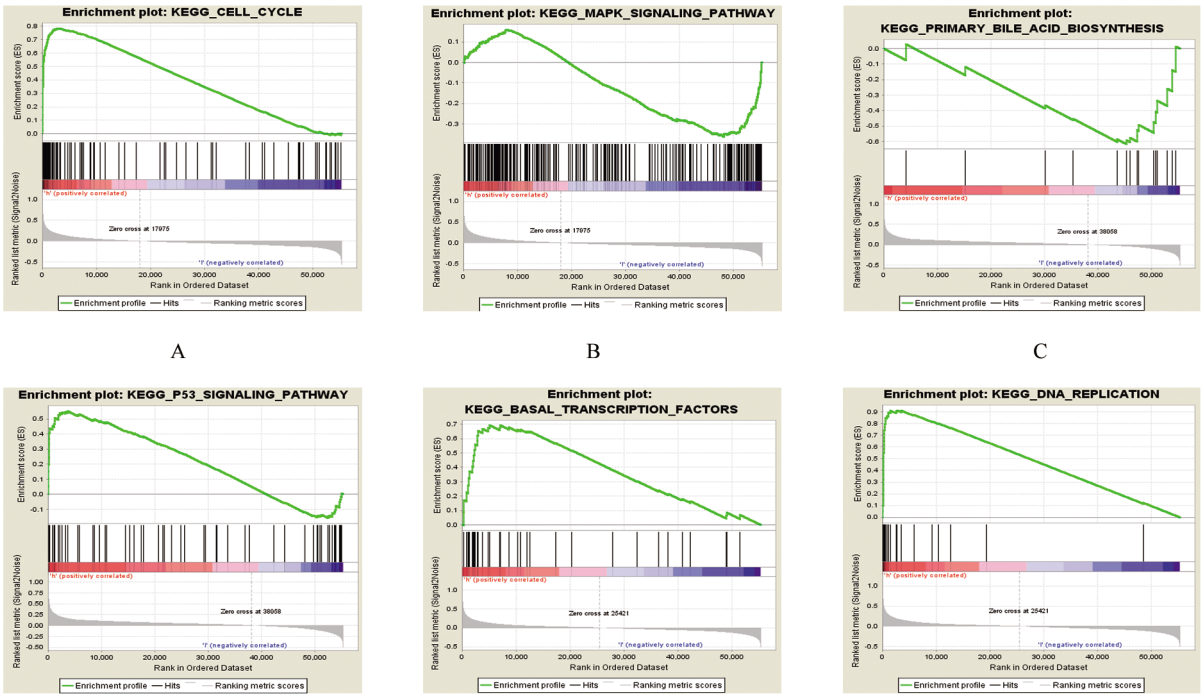

图 11 *UBE2T*, *ERCC6L* 和 *RAD51* 的基因功能富集分析

注: A,B: UBE2T; C,D: ERCC6L; E,F: RAD51; p<0.05 and FDR<0.05 有统计学意义

Figure 11 Gene set enrichment analysis of *UBE2T*, *ERCC6L* and *RAD51*

Note: A,B: UBE2T; C,D: ERCC6L; E,F: RAD51; p<0.05 and FDR<0.05 were considered significantly enriched

肿瘤进展的致癌基因，可能被认为是有效诊断和开发乳腺癌疗法的潜在靶标。

重组蛋白 A (Recombination protein A, RAD51)与大肠杆菌 RecA 的同源物，是减数分裂和有丝分裂重组以及修复双链 DNA 断裂所必需的 (Chen et al., 1998)。近年来，聚腺苷酸二磷酸核糖聚合酶抑制剂[Poly (ADP-ribose) polymerase inhibitors,PARP]已证明在治疗具有同源重组 DNA 修复缺陷的恶性肿瘤中具有临床实用性。最初，PARP 被批准用于治疗乳腺癌易感基因 (breast cancer susceptibility genes, BRCA)缺乏的乳腺癌和卵巢癌中(Stewart et al., 2018)。Toh 等(Toh et al., 2019)的研究显示 BRCA1 异二聚体和 RAD51 共同定位，与它们位于 BRCA1 和 RAD51 远端结合域位置是一致的。在临床上，BARD1 和 BRCA1 细胞株的致病变异均富含更具侵袭性的乳腺癌表型，例如三阴性乳腺癌(triple-negative breast cancer, TNBC)，与更高的复发、进展和死亡率相关(Couch et al., 2015; Mani et al., 2019)。Jia 等(Jia et al., 2019)表明，RAD51 的高表达是 ER 阳性乳腺癌患者行新辅助内分泌治疗的不利指标，包括使用芳香化酶抑制剂治疗疗效和生存期方面都很差。

总之，目前我们研究表明这三种基因被认为是乳腺癌潜在肿瘤标志物，具有作为预后指标，对乳腺癌患者的诊断以及治疗具有较高的价值。我们以生物信息学分析的方式检测了这些基因潜在的机制，所以，以上潜在功能基因与乳腺癌的相关性及相关机制的研究仍需在临床样本中进一步的验证。

3 材料与方法

3.1 数据来源

微阵列数据集 GSE22820 从 GEO 数据库(https://www.ncbi.nlm.nih.gov/geo/)获得。该阵列测定了 176 例原发性乳腺癌样本和 10 例癌旁正常组织。TCGA 乳腺癌数据集以及临床数据(包括 112 例正常组织和 1066 例癌组织) 可从癌症基因组图谱数据库下载(https://portal.gdc.cancer.gov/)。

3.2 差异表达基因筛选

采用 R 语言 Limma 数据包 (Ritchie et al., 2015)对乳腺癌组织与癌旁正常组织进行差异表达基因筛选，筛选标准为 p<0.05，|FC(fold change)|>2；随后，使用在线工具 venny2.1.0 (http://http://bioinformatics.

psb.ugent.be/webtools/Venn/)来识别两个基因表达微阵列中的重叠差异表达基因。随后，分别测量上调和下调的基因。

3.3 基因功能富集和注释

基于前述所得的差异表达基因，使用 R 语言中软件包库进行了注释，依据基因本体(gene ontology, GO)数据库对差异表达基因进行生物学功能注释。同时利用京都基因与基因组百科全书 (Kyoto encyclopedia of genes and genomes, KEGG) 通路数据库进行差异基因信号通路的富集(Huang et al., 2007)。$p<0.05$ 被认为具有统计学意义。

3.4 蛋白质相互作用网络构建

STRING 是一种在线工具(https://string-db.org/)，可以构建已知的蛋白质与蛋白质互作网络并整合这些信息(Szklarczyk et al., 2017)。在获得蛋白质相互作用网络图后导入 Cytoscape 软件(Cytoscape_v3.6.1)中进行可视化显示，利用软件中 cytoHubba 插件筛选网络内顶部前 20 个基因(Shannon et al., 2003)，并通过 MCODE 插件筛选出前三个模块。

3.5 基因富集分析

基因富集分析(gene set enrichment analysis,GSEA) 可以确定预先定义的一组基因在两个生物学状态之间是否有统计学上差异的一种方法(Subramanian et al., 2007)。为了研究 UBE2T、ERCC6L 和 RAD51 在乳腺癌中的作用进行了 GSEA 分析，当正常 $p<0.05$ 和 FDR<0.25 时，则认为该基因集显着富集。

3.6 统计学分析

通过 R×64 3.6.0 软件分析了从 GEO 和 TCGA 数据集下载的重叠差异表达基因临床信息。Cox 风险比例模型进行单变量和多变量分析。Kaplan-Meier (K-M)方法用于生存分析。所有结果均以平均值±标准差表示，$p<0.05$ 被认为揭示了统计学上的显着差异。

作者贡献

王建、李金平是本研究设计的执行人；王建、邵荣金及龚伟达完成数据分析，论文初稿的写作；吕旭参与本研究的设计，研究结果分析；李金平是项目的构思者及负责人，指导本研究的设计、数据分析，论文写作与修改。全体作者都阅读并同意最终的文本。

致谢

本研究由宜兴市科技创新(社会发展类)专项资金项目(2019SF23)资助。

参考文献

Alpi A.F., Chaugule V., and Walden H., 2016, Mechanism and disease association of E2-conjugating enzymes: lessons from UBE2T and UBE2L3, Biochem J., 473: 3401-3419

Baselga J., Coleman R.E., Cortes J., and Janni W., 2017, Advances in the management of HER2-positive early breast cancer, Crit. Rev. Oncol. Hematol., 119: 113-122

Bray F., Ferlay J., Soerjomataram I., Siegel R.L., Torre L.A., and Jemal A., 2018, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin., 68: 394-424

Chang C.H., Chiu C.F., Wang H.C., Wu H.C., Tsai R.Y., Tsai C. W., Wang R.F., Wang C.H., Tsou Y.A., and Bau D.T., 2009, Significant association of ERCC6 single nucleotide polymorphisms with bladder cancer susceptibility in Taiwan, Anticancer Res., 29: 5121-5124

Chen P.L., Chen C.F., Chen Y., Xiao J., Sharp Z.D., and Lee W. H., 1998, The BRC repeats in BRCA2 are critical for RAD51 binding and resistance to methyl methanesulfonate treatment, Proc. Natl. Acad. Sci. USA., 95: 5287-5292

Clegg L.X., Reichman M.E., Miller B.A., Hankey B.F., Singh G. K., Lin Y.D., Goodman M.T., Lynch C.F., Schwartz S.M., Chen V.W., Bernstein L., Gomez S.L., Graff J.J., Lin C.C., Johnson N.J., and Edwards B.K., 2009, Impact of socioeconomic status on cancer incidence and stage at diagnosis: selected findings from the surveillance, epidemiology, and end results: National longitudinal mortality study, Cancer Causes Control, 20: 417-435

Colak D., Nofal A., Albakheet A., Nirmal M., Jeprel H., Eldali A., Al-Tweigeri T., Tulbah A., Ajarim D., Malik O.A., Inan M.S., Kaya N., Park B.H., and Bin Amer S.M., 2013, Age-specific gene expression signatures for breast tumors and cross-species conserved potential cancer progression markers in young women, PLoS One, 8: e63204

Couch F.J., Hart S.N., Sharma P., Toland A.E., Wang X., Miron P., Olson J.E., Godwin A.K., Pankratz V.S., Olswold C., Slettedahl S., Hallberg E., Guidugli L., Davila J.I., Beckmann M.W., Janni W., Rack B., Ekici A.B., Slamon D.J., Konstantopoulou I., Fostira F., Vratimos A., Fountzilas G., Pelttari L.M., Tapper W.J., Durcan L., Cross S.S., Pilarski R., Shapiro C.L., Klemp J., Yao S., Garber J., Cox A., Brauch H., Ambrosone C., Nevanlinna H., Yannoukakos D., Slager S.L., Vachon C.M., Eccles D.M., and Fasching P.A.,

2015, Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer, J. Clin. Oncol., 33: 304-311

DeSantis C.E., Ma J., Goding Sauer A., Newman L.A., and Jemal A., 2017, Breast cancer statistics, 2017, racial disparity in mortality by state, CA Cancer J, Clin., 67: 439-448

Ghoncheh M., Pournamdar Z., and Salehiniya H., 2016, Incidence and mortality and epidemiology of breast cancer in the world, Asian Pac. J. Cancer Prev., 17: 43-46

Huang D.W., Sherman B.T., Tan Q., Collins J.R., Alvord W.G., Roayaei J., Stephens R., Baseler M.W., Lane H.C., and Lempicki R.A., 2007, The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists, Genome Biol., 8: R183

Januskeviciene I., and Petrikaite V., 2019, Heterogeneity of breast cancer: The importance of interaction between different tumor cell populations, Life Sci., 239: 117009

Jia Y., Song Y., Dong G., Hao C., Zhao W., Li S., and Tong Z., 2019, Aberrant regulation of RAD51 promotes resistance of neoadjuvant endocrine therapy in ER-positive breast cancer, Sci. Rep., 9: 12939

Koshiba M., 2016, Molecular targeted therapy and laboratory tests, Rinsho. Byori., 64: 709-716

Liu J., Deng N., Xu Q., Sun L., Tu H., Wang Z., Xing C., and Yuan Y., 2016, Polymorphisms of multiple genes involved in NER pathway predict prognosis of gastric cancer, Oncotarget, 7: 48130-48142

Liu J., Sun J., Zhang Q., and Zeng Z., 2018, shRNA knockdown of DNA helicase ERCC6L expression inhibits human breast cancer growth, Mol. Med. Rep., 18: 3490-3496

Liu J.W., He C.Y., Sun L.P., Xu Q., Xing C.Z., and Yuan Y., 2013, The DNA repair gene ERCC6 rs1917799 polymorphism is associated with gastric cancer risk in Chinese, Asian Pac. J. Cancer Prev., 14: 6103-6108

Ma H., Hu Z., Wang H., Jin G., Wang Y., Sun W., Chen D., Tian T., Jin L., Wei Q., Lu D., Huang W., and Shen H., 2009, ERCC6/CSB gene polymorphisms and lung cancer risk, Cancer Lett., 273: 172-176

Machida Y.J., Machida Y., Chen Y., Gurtan A.M., Kupfer G.M., D'Andrea A.D., and Dutta A., 2006, UBE2T is the E2 in the Fanconi anemia pathway and undergoes negative autoregulation, Mol. Cell., 23: 589-596

Mani C., Jonnalagadda S., Lingareddy J., Awasthi S., Gmeiner W.H., and Palle K., 2019, Prexasertib treatment induces homologous recombination deficiency and synergizes with olaparib in triple-negative breast cancer cells, Breast Cancer Res., 21: 104

Nielsen C.F., Huttner D., Bizard A.H., Hirano S., Li T.N., Palmai-Pallag T., Bjerregaard V.A., Liu Y., Nigg E.A., Wang L.H., and Hickson I.D., 2015, PICH promotes sister chromatid disjunction and co-operates with topoisomerase II in mitosis, Nat. Commun., 6: 8962

Park S.Y., Gonen M., Kim H.J., Michor F., and Polyak K., 2010, Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype, J. Clin. Invest., 120: 636-644

Perez-Pena J., Corrales-Sanchez V., Amir E., Pandiella A., and Ocana A., 2017, Ubiquitin-conjugating enzyme E2T (UBE-2T) and denticleless protein homolog (DTL) are linked to poor outcome in breast and lung cancers, Sci. Rep., 7: 17530

Pu S.Y., Yu Q., Wu H., Jiang J.J., Chen X.Q., He Y.H., and Kong Q.P., 2017, ERCC6L, a DNA helicase, is involved in cell proliferation and associated with survival and progress in breast and kidney cancers, Oncotarget., 8: 42116-42124

Ramaniuk V.P., Nikitchenko N.V., Savina N.V., Kuzhir T.D., Rolevich A.I., Krasny S.A., Sushinsky V.E., and Goncharova R.I., 2014, Polymorphism of DNA repair genes OGG1, XRCC1, XPD and ERCC6 in bladder cancer in Belarus, Biomarkers., 19: 509-516

Ritchie M.E., Phipson B., Wu D., Hu Y., Law C.W., Shi W., and Smyth G.K., 2015, limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res., 43: e47

Rugo H.S., Rumble R.B., Macrae E., Barton D.L., Connolly H.K., Dickler M.N., Fallowfield L., Fowble B., Ingle J.N., Jahanzeb M., Johnston S.R., Korde L.A., Khatcheressian J.L., Mehta R.S., Muss H.B., and Burstein H.J., 2016, Endocrine therapy for hormone receptor-positive metastatic breast cancer, american society of clinical oncology guideline, J. Clin. Oncol., 34: 3069-3103

Santamaria A., Neef R., Eberspacher U., Eis K., Husemann M., Mumberg D., Prechtl S., Schulze V., Siemeister G., Wortmann L., Barr F.A., and Nigg E.A., 2007, Use of the novel Plk1 inhibitor ZK-thiazolidinone to elucidate functions of Plk1 in early and late stages of mitosis, Mol. Biol. Cell., 18: 4024-4036

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., and Ideker T., 2003, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res., 13: 2498-2504

Stewart M.D., Zelin E., Dhall A., Walsh T., Upadhyay E., Corn J. E., Chatterjee C., King M.C., and Klevit R.E., 2018, BARD1 is necessary for ubiquitylation of nucleosomal histone H2A and for transcriptional regulation of estrogen metabolism genes, Proc. Natl. Acad. Sci. USA., 115: 1316-1321

Subramanian A., Kuehn H., Gould J., Tamayo P., and Mesirov J. P., 2007, GSEA-P: a desktop application for gene set enrichment analysis, Bioinformatics., 23: 3251-3253

Szklarczyk D., Morris J.H., Cook H., Kuhn M., Wyder S., Simonovic M., Santos A., Doncheva N.T., Roth A., Bork P., Jensen L.J., and Von Mering C., 2017, The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, Nucleic Acids Res., 45: D362-D368

Toh M.R., Chong S.T., Chan S.H., Low C.E., Ishak N.D.B., Lim J.Q., Courtney E., and Ngeow J., 2019, Functional analysis of clinical BARD1 germline variants, Cold Spring Harb. Mol. Case Stud., 5: a004093

Torre L.A., Islami F., Siegel R.L., Ward E.M., and Jemal A., 2017, Global cancer in women: Burden and trends, Cancer Epidemiol Biomarkers Prev., 26: 444-457

Ueki T., Park J.H., Nishidate T., Kijima K., Hirata K., Nakamura Y., and Katagiri T., 2009, Ubiquitination and downregulation of BRCA1 by ubiquitin-conjugating enzyme E2T overexpression in human breast cancer cells, Cancer Res., 69: 8752-8760

Xu Q., Liu J.W., He C.Y., Sun L.P., Gong Y.H., Jing J.J., Xing C.Z., and Yuan Y., 2014, The interaction effects of pri-let-7a-1 rs10739971 with PGC and ERCC6 gene polymorphisms in gastric cancer and atrophic gastritis, PLoS One., 9: e89203