·临床研究·

WGCNA联合LASSO-COX分析筛选并构建乳腺癌患者5基因预后预测模型

陈彩萍　钱燕芳

[摘要] **目的**　基于癌症基因组图谱数据库(TCGA)筛选乳腺癌预后相关的关键基因作为生物标志物,并构建预后预测模型。**方法**　从TCGA数据库收集乳腺癌和正常样本的基因表达图谱,利用limma算法筛选乳腺癌样本组和正常组之间的差异表达基因(DGEs),采用WGCNA联合LASSO-COX分析DGEs获得预后相关的关键基因,再由关键基因构建预后模型评估患者风险,并在基因表达综合数据库(GEO)乳腺癌数据集中进行验证。最后,通过GSEA方法分析高低风险组患者涉及的关键信号通路。**结果**　通过差异基因分析获得1 000个DGEs;WGCNA联合LASSO-COX分析DGEs,获得5个乳腺癌预后关键基因:FBXL19、HAGHL、PHKG2、PKMYT1和TXNDC17,由这些基因构建预后预测模型并计算患者风险评分;ROC分析表明该模型具有良好的预测性能并在GEO数据库得到验证;生存分析显示高风险评分与患者不良预后相关;GSEA分析表明p53信号通路富集于高风险评分组。**结论**　由FBXL19、HAGHL、PHKG2、PKMYT1和TXNDC17组成的预后预测模型可用于乳腺癌患者预后预测,为乳腺癌患者基因靶向治疗提供参考。

[关键词] 乳腺癌;　预后;　WGCNA;　LASSO-COX

Construction of prognosis model based on five genes that screened by WGCNA and LASSO-COX for breast cancer patients CHEN Caiping, QIAN Yanfang. Department of Ultrasound, The First People's Hospital of Xiaoshan District, Hangzhou 311200, China.

[Abstract] Objective　To identify prognostic signatures to predict the prognosis of breast cancer based on a series of comprehensive analysis of gene expression data. **Methods**　The RNA-sequencing expression data and corresponding breast cancer patient clinical data were collected from the Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases. Firstly, the differentially expressed genes (DEGs) related to prognosis between tumor tissues and normal tissues were ascertained by performing R package "limma" based on the TCGA database. Second, DEGs were used to construct a polygenic risk scoring model by the weighted gene co-expression network analysis (WGCNA) and the least absolute shrinkage and selection operator (LASSO)-COX regression model. Third, we performed the survival analysis to investigate the risk score values in the TCGA cohort. Simultaneously, the GEO cohort was used to validate the model. Lastly, the Gene Set Enrichment Analysis (GSEA) was performed based on gene expression profile and risk score grouping to enrich relevant pathways and molecular mechanisms. **Results**　A total of 1 000 DEGs were identified. A prognostic signature comprising 5 genes including FBXL19, HAGHL, PHKG2, PKMYT1, and TXNDC17 was developed to divide patients into high-risk and low-risk groups, and its prognostic prediction was great both in training and validation cohorts. The high-risk group generally had a poorer prognosis. **Conclusion**　The 5 genes including FBXL19, HAGHL, PHKG2, PKMYT1, and TXNDC17 risk model can predict the prognosis of breast cancer patients, which is a reference for the gene-targeted therapy of breast cancer.

[Key words] breast cancer;　prognostic;　WGCNA;　LASSO-COX

DOI:10.13558/j.cnki.issn1672-3686.2023.004.013

作者单位:311200　浙江杭州,杭州市萧山区第一人民医院超声科

2020年乳腺癌已经成为全球第一大癌症,是女性发病率最高的恶性肿瘤,在女性肿瘤患者中,占

比高达30%[1]，严重威胁女性的生命健康。乳腺癌具有很高的异质性，然而目前临床上是根据肿瘤的TNM分期，激素表达水平以及人表皮生长因子2（human epidermal growth factor 2，Her-2）水平制定治疗方案和进行预后预测[2]。预后预测不准确会对轻微患者过度治疗造成二次伤害，对重度患者疏忽治疗会引起复发，严重影响她们的生活质量和生存时间[3]。因此，寻找精准分子标记物来预测乳腺癌患者的预后十分迫切。目前用于诊断、治疗、预测患者预后的乳腺癌的分子标记物主要分为以下五类：①肿瘤细胞生长类：Ki67抗体，拓扑异构酶Ⅱ，细胞周期素D1；②激素水平类：雌激素受体，孕激素受体；③预后类：HER2；④肿瘤发生类：AKT，mTOR，p53，PK13，BRCA1；⑤调解通路类：调解ERK信号通路的EGFL7[4]，作用于p53信号通路的NRF2[5]。然而，多数研究集中于单基因诊断、预后预测模型的构建，预测效果具有局限性。因此，本次研究通过多种生物信息学方法构建一种新的多基因综合风险评分模型用于预测乳腺癌患者的预后。

1 资料与方法

1.1 数据获取与处理 从癌症基因组图谱（the cancer genome atlas program，TCGA）数据库，获取1 098例乳腺癌肿瘤样本和113例正常样本的mRNA表达数据及962例乳腺癌患者临床信息作为训练集，从基因表达综合数据库（gene expression omnibus，GEO）获取一个具有238例乳腺癌基因表达数据集（GSE103091）作为验证集。

1.2 差异基因筛选 首先通过训练集筛选乳腺癌样本组和正常组之间的差异表达基因（differential expressed genes，DEGs）。应用limma R软件包对RNA-seq表达数据行归一化处理，并以$P<0.01$和$|\log_2 FC|>1$为阈值，筛选具有统计学意义的DEGs。

1.3 WGCNA与LASSO-COX分析 先计算每个基因的中位数绝对离差，剔除值小的前50%的基因，利用Good Samples Genes方法去除离群的基因和样本，再通过计算尺度独立性和平均连通性以确定表征基因符合无尺度分布的软阈值，*Pearson*法选定与生存状态相关基因模块。LASSO分析通过引入惩罚系数（λ）将冗余变量的系数压缩为0，最后剩余系数非零的变量为最终变量。本研究中使用5折交叉验证确定最优惩罚系数，得出有效基因再将这些有效基因进行多因素COX回归分析，计算每个基因的回归系数，构建风险评分方程。

1.4 预后模型建立与验证 根据上述LASSO-COX回归分析的结果，构建基于基因表达和回归系数的模型方程，计算每个样本的风险评分。公式如下：风险评分=$\beta1\times$mRNA 1EXP+$\beta2\times$mRNA 2EXP+.....+$\beta n\times$mRNA nEXP[6,7]。β为相应mRNA的多因素回归系数，mRNA EXP为相应mRNA的表达量。根据风险评分的最优截断值将患者分为高、低风险两组，利用Kaplan-Meier法进行生存分析，Log-rank检验进行组间比较；利用R软件绘制模型的绘制受试者工作特征（receiver operating characteristic，ROC）曲线并计算曲线下面积（area under curve，AUC），评估模型预测效能且在GSE103091数据集中进行验证。

1.5 基因集富集分析（gene set enrichment analysis，GSEA） 利用GSEA研究风险评分与京都基因和基因组百科全书（kyotoencyclopdia of genes and genomes，KEGG）通路间的相关性。采用GSEA软件，根据风险评分的最优截断值将患者分为高、低风险两组，并根据c2.cp.kegg.v7.4.symbols.gmt子集合，评估相关途径和分子机制。GSEA分析参数设置：基因集范围为[5,5000]，1000次重抽样。$P<0.05$，FDR<0.25作为显著差异性。

2 结果

2.1 临床特征 本研究共纳入962例乳腺癌患者，其中女性950例（98.75%）、男性12例（1.25%）；年龄≤45岁171例（17.78%）、年龄>45岁791例（82.22%）；死亡116例（12.06%）、生存846例（87.94%）；临床分期为Ⅰ期、Ⅱ期、Ⅲ期、Ⅳ期分别有163例（16.94%）、559例（58.11%）、205例（21.31%）、18例（1.87%），不确定17例（1.77%）；发生远端转移19例（1.97%）、未发生远端转移799例（83.06%），不确定144例（14.97%）；淋巴转移分期为N0、N1、N2、N3分别有461例（47.92%）、314例（32.64%）、109例（11.33%），60例（6.24%），不确定18例（1.87%）；T分期为T1、T2、T3、T4分别有252例（26.20%）、569例（59.15%）、104例（10.81%）、34例（3.53%），不确定3例（0.31%）；死于乳腺癌70例（7.28%），生存或者死于其他疾病为892例（92.72%）。

2.2 DEGs 运用函数limma算法筛选乳腺癌与正常组间的DEGs。通过条件$P<0.05$，$|\log_2 FC|>1$，筛选出1 000个符合条件的DEGs，包含396个上调和604个下调基因。

2.3 WGCNA分析 通过WGCNA的尺度独立性和平均连通性比较发现，基因间联系软阈值为9，设置模块合并阈值为0.2，获取6个基因模块。临床特征与6个基因模块性状相关性分析见表1。

表1 临床特征与基因模块的相关性

临床特征	蓝色		黄色		红色		棕色		绿色		灰色	
	r	*P*	*r*	*P*	*r*	*P*	*r*	*P*	*r*	*P*	*r*	*P*
总体生存状态	0.04	>0.05	−0.02	>0.05	-2.70×10^{-3}	>0.05	−0.03	>0.05	−0.02	>0.05	5.70×10^{-4}	>0.05
总体生存时间	0.03	>0.05	−0.07	<0.05	−0.06	>0.05	0.02	>0.05	0.02	>0.05	−0.08	<0.05
疾病特异性生存状态	0.06	>0.05	−0.03	>0.05	−0.02	>0.05	−0.04	>0.05	−0.05	>0.05	-5.10×10^{-4}	>0.05
疾病特异性生存时间	0.03	>0.05	−0.07	<0.05	−0.06	>0.05	0.02	>0.05	0.02	>0.05	−0.08	<0.05
无复发生存状态	0.05	>0.05	−0.05	>0.05	−0.03	>0.05	−0.02	>0.05	4.60×10^{-4}	>0.05	−0.04	>0.05
无复发生存时间	4.40×10^{-4}	>0.05	−0.07	<0.05	−0.07	<0.05	0.04	>0.05	0.03	>0.05	−0.08	<0.05
T分期	0.08	<0.05	0.05	>0.05	−0.07	<0.05	−0.05	>0.05	−0.06	>0.05	0.10	<0.05
N分期	−0.02	>0.05	0.04	>0.05	0.07	<0.05	0.07	<0.05	0.03	>0.05	0.06	>0.05
M分期	0.03	>0.05	0.07	<0.05	-1.10×10^{-4}	>0.05	−0.06	>0.05	−0.04	>0.05	0.07	<0.05
临床分期	0.05	>0.05	0.06	>0.05	-8.00×10^{-3}	>0.05	−0.02	>0.05	−0.03	>0.05	0.09	<0.05
性别	0.05	>0.05	0.11	<0.05	5.70×10^{-3}	>0.05	−0.04	>0.05	−0.04	>0.05	0.11	<0.05
年龄	−0.14	<0.05	0.05	>0.05	−0.08	<0.05	−0.06	>0.05	0.02	>0.05	−0.01	>0.05

由表1可见，黄色模块与临床性状的关联性最高，确定其为关键模块。

2.4 LASSO-COX分析 根据生存时间、状态和基因表达数据，使用LASSO-COX对关键模块包含的36个基因进一步筛选，设置λ值为0.0095，最终获得5个关键预后基因，见封三图1。

由封三图1A可见，五个关键基因为：FBXL19、HAGHL、PHKG2、PKMYT1、TXNDC17。由封三图1B可见，使用交叉验证建立模型，结果一致。

2.5 预后模型 根据得到的回归系数及基因表达值，按照如下公式构建综合风险评分模型：风险评估=0.015×FBXL19+0.055×HAGHL−0.244×PHKG2+0.011×PKMYT1+0.019×TXNDC17。基于5个基因风险评分的生存曲线及ROC曲线见封三图2。

由封三图2A、B可见，该风险评分显著影响患者的预后，且高风险评分组的患者生存率较低。利用GSE103091数据集对模型加以验证，生存分析表明高风险组和低风险组的预后差异具有统计学意义，且低风险值有利于预后。

由封三图2C、D可见，基于5个基因风险评分的ROC曲线的AUC为0.66（95%*CI* 0.59~0.73），表明该模型具有良好的预测性能。验证集AUC为0.60（95%*CI* 0.46~0.75），表明模型预测性能良好。

2.6 GSEA分析见封三图3

由封三图3可见，GSEA分析表明p53信号通路与风险评分密切相关，该通路富集于高风险组。

3 讨论

本研究得到5个预后关键基因：FBXL19、HAGHL、PHKG2、PKMYT1、TXNDC17，根据这些基因构建预后模型且计算患者的风险评分。该模型表明低风险评分有利于患者预后。ROC分析表明该模型具有良好的预测性能，且在GSE103091数据集得以验证。最后通过GSEA分析，发现p53信号通路富集于高风险组。

FBXL19可通过调节食管癌进展中起重要作用的Rac3蛋白的降解和泛素化，抑制TGF-β1通路诱导的钙粘蛋白下调从而抑制癌细胞转移[8]。它可能通过调节不同通路影响乳腺癌细胞的增殖与生长，需深入研究。Im等[9]研究表明HAGHL与骨吸收通路相关，是患癌女性儿童化疗后预测骨折风险的重要生物标志。PHKG2与cAMP依赖型蛋白激酶A激活、糖原代谢通路相关，其突变会导致人患糖原贮积病Ⅸc型疾病，增加肝纤维化和肝硬化风险[10]。在乳头状甲状腺癌中，该基因有过度甲基化特点[11]；在乳腺癌中，除非致瘤性乳腺上皮细胞外，在其他乳腺细胞系中高表达[12]，且其突变可破坏其他乳腺癌重要相关基因的功能[13]。PKMYT1属于丝氨酸/苏氨酸蛋白激酶家族，通过细胞周期蛋白依赖性激酶1的磷酸化和失活的方式，抑制高尔基体和内质网组装，阻滞卵母细胞从G2期进入M期[14]。

Liu 等[15]研究表明PKMYT1的表达与乳腺癌的雌激素、孕激素水平相关，且高表达不利于预后，该研究GSEA分析还发现高风险评分组激素相关通路的基因表达水平较高。TXNDC17属于硫氧还蛋白家族[16]，其在肺癌中过表达可增加Atg5和Beclin1的表达，激活细胞自噬，促进癌细胞增殖[17]，在卵巢癌中同样高表达，与患者较差预后相关[18]。p53通路调节多种基因的表达和信号通路，包括细胞凋亡，分化，基因修复和抑制血管生成。研究表明，乳腺癌的发生与p53异常有关。乳腺癌的生长需要血管提供大量的营养物质，而p53可通过影响促血管生成的血管内皮生长因子的表达，抑制肿瘤恶化。本研究中，GSEA分析表明高风险组患者p53信号通路被抑制，且预后较差，可能是因为高风险组患者p53处于异常状态，其活性变化影响到对血管内皮生长因子的调控，使得肿瘤血管生成过程不受抑制，进而促进肿瘤发展。

综上所述，本文通过生物信息学方法筛选出可为潜在预后标志物的5个基因，并构建了具有良好预测性能的预后模型。这对乳腺癌的预后预测与基因靶向治疗有重要意义，其中潜在调控机制还需体外和体内实验进一步验证。

参考文献

1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020[J]. CA Cancer J Clin, 2020, 70(1): 7-30.

2 Vuong D, Simpson PT, Green B, et al. Molecular classification of breast cancer[J]. Virchows Arch, 2014, 465(1): 1-14.

3 李伟华，张广凤，马骊骊，等. 基于TCGA数据库代谢相关基因构建乳腺癌预后模型[J]. 实用肿瘤学杂志，2022，36(1)：37-43.

4 姜青明，周文文，肖觉. 乳腺癌EGFL7及β-catenin蛋白检测的临床意义[J]. 临床与病理杂志，2018，38(3)：472-479.

5 丁牧遥，张倩，袁胜涛，等. NRF2信号通路在乳腺癌中的研究进展[J]. 中国细胞生物学学报，2021，43(10)：2085-2092.

6 Yu L, Xiang L, Feng J, et al. mRNA-21 and mRNA-223 expression signature as a predictor for lymph node metastasis, distant metastasis and survival in kidney renal clear cell carcinoma[J]. J Cancer, 2018, 9(20): 3651-3659.

7 Liu Q, Diao R, Feng G, et al. Risk score based on three mRNA expression predicts the survival of bladder cancer [J]. Oncotarget, 2017, 8(37): 61583-61591.

8 Dong S, Zhao J, Wei J, et al. F-box protein complex FBXL19 regulates TGFbeta1-induced E-cadherin down-regulation by mediating Rac3 ubiquitination and degradation[J]. Mol Cancer, 2014, 13: 76.

9 Im C, Li N, Moon W, et al. Genome-wide Association Studies Reveal Novel Locus With Sex -/Therapy-Specific Fracture Risk Effects in Childhood Cancer Survivors[J]. J Bone Miner Res, 2021, 36(4): 685-695.

10 Shao Y, Li T, Jiang M, et al. A very rare case report of glycogen storage disease type IXc with novel PHKG2 variants[J]. BMC Pediatr, 2022, 22(1): 267.

11 Kikuchi Y, Tsuji E, Yagi K, et al. Aberrantly methylated genes in human papillary thyroid cancer and their association with BRAF/RAS mutation[J]. Front Genet, 2013, 4: 271.

12 Fu S, Cheng J, Wei C, et al. Development of diagnostic SCAR markers for genomic DNA amplifications in breast carcinoma by DNA cloning of high-GC RAMP-PCR fragments[J]. Oncotarget, 2017, 8(27): 43866-43877.

13 Xiao F, Kim YC, Snyder C, et al. Genome instability in blood cells of a BRCA1+ breast cancer family[J]. BMC Cancer, 2014, 14: 342.

14 Mueller PR, Coleman TR, Kumagai A, et al. Myt1: A membrane-associated inhibitory kinase that phosphorylates Cdc2 on both threonine-14 and tyrosine-15[J]. Science, 1995, 270(5233): 86-90.

15 Liu Y, Qi J, Dou Z, et al. Systematic expression analysis of WEE family kinases reveals the importance of PKMYT1 in breast carcinogenesis[J]. Cell Prolif, 2020, 53(2): e12741.

16 Jeong W, Yoon HW, Lee SR, et al. Identification and characterization of TRP14, a thioredoxin-related protein of 14 kDa. New insights into the specificity of thioredoxin function[J]. J Biol Chem, 2004, 279(5): 3142-3150.

17 闫红江，任红欣，高伟年，等. TRP14促进非小细胞肺癌细胞增殖移动的作用机制研究[J]. 循证医学，2021，21(5)：294-300.

18 Zhang SF, Wang XY, Fu ZQ, et al. TXNDC17 promotes paclitaxel resistance via inducing autophagy in ovarian cancer[J]. Autophagy, 2015, 11(2): 225-238.

（收稿日期 2022-07-08）

（本文编辑 葛芳君）