·论　　著·　　DOI：10.3969/j.issn.1672-9455.2022.23.016

基于 GEO 及 TCGA 数据库建立乳腺癌他莫昔芬耐药相关预后模型

郑　洁[1]，开震天[2]，夏超然[2]，罗　鹏[2]，刘　晖[1]，王建东[1]，王　凤[1]，胡俊艳[1△]

1.上海中医药大学附属曙光医院乳腺外科，上海 201203；2.上海鼎晶生物医药科技股份有限公司，上海 201321

摘　要：目的　通过生物信息学的方法，构建并验证预测雌激素受体阳性乳腺癌他莫昔芬治疗预后的长链非编码 RNA（lncRNA）模型。**方法**　利用基因表达综合（GEO）数据库提取雌激素受体阳性乳腺癌芯片测序数据，并从中下载他莫昔芬耐药及敏感细胞系的 lncRNA 表达数据集，通过单因素和多因素 Cox 回归分析构建预后模型。根据模型计算风险比（*HR*），分析高危组、低危组的临床病理参数与预后的关系，并用验证集样本对训练集的结果进行验证。**结果**　在 GEO 数据库和癌症基因组图谱（TCGA）数据库对他莫昔芬敏感与他莫昔芬耐药乳腺癌患者的差异 lncRNA 进行比对筛选，最终获得 116 个差异 lncRNA。对差异 lncRNA 进行单因素 Cox 回归分析，计算每个 lncRNA 与乳腺癌患者总生存率的 *HR* 及 *P* 值，最终得到 8 个相关的 lncRNA（$P<0.05$）。通过多因素 Cox 回归分析最终建立由 6 个 lncRNA 组成的预后模型。按照训练集中 *HR* 的中位数将训练集及验证集中的患者分为高危组、低危组，Log-rank 检验结果发现，高危组与低危组在训练集和验证集中的生存率差异均有统计学意义（$P=7\times10^{-7}$、0.008）。**结论**　乳腺癌他莫昔芬耐药相关预后模型 6 个 lncRNA 对雌激素受体阳性乳腺癌患者预后具有一定价值，可为逆转他莫昔芬耐药、治疗癌症的研究提供依据。

关键词：乳腺癌；　雌激素受体；　他莫昔芬耐药；　预后模型

中图法分类号：R737.9　　**文献标志码：**A　　**文章编号：**1672-9455(2022)23-3227-04

Establishment of a prognostic model of tamoxifen resistance in breast cancer based on GEO and TCGA databases

ZHENG Jie[1]，*KAI Zhentian*[2]，*XIA Chaoran*[2]，*LUO Peng*[2]，*LIU Hui*[1]，*WANG Jiandong*[1]，*WANG Feng*[1]，*HU Junyan*[1△]

1. *Department of Breast Surgery，Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine，Shanghai* 201203，*China*；

2. *Shanghai Topgen Biomedical Technology Co.，Ltd，Shanghai* 201321，*China*

Abstract：Objective　To construct and verify a long non-coding RNA (lncRNA) model for predicting the prognosis of estrogen receptor positive breast cancer treated with tamoxifen by bioinformatics method. **Methods**　Gene Expression Omnibus (GEO) database was used to extract the chip sequencing data of estrogen receptor positive breast cancer，from which lncRNA expression data sets of tamoxifen-resistant and sensitive cell lines were downloaded. The prognostic models were constructed by univariate and multivariate Cox regression analysis. Hazard ratio (*HR*) was calculated according to the model，and the relationship between clinicopathological parameters and prognosis was analyzed in high-risk group and low-risk group. The relevant data sets in the cancer genome atlas database were randomly divided into training set and validation set，and the results of training set samples were verified by validation set. **Results**　Differential lncRNAs of tamoxifen-sensitive and tamoxifen-resistant breast cancer patients were compared and screened in GEO database and the cancer genome atlas (TCGA) database，and finally 116 lncRNAs were obtained. Subsequently，univariate Cox regression analysis was performed to calculate the *HR* and *P* values of each lncRNA and the overall survival rate of breast cancer patients，and finally 8 significantly related lncRNAs were obtained ($P<0.05$). Multivariate Cox regression analysis was used to establish a prognostic model consisting of 6 lncRNAs. According to the median *HR* in the training set，patients in the training set and validation set were divided into high-risk group and low-risk group. Log-rank test results showed that there were statistically significant differences in the survival rates of high-risk group and the low-risk group in the training set and validation set ($P=7\times10^{-7}$，0.008).

作者简介：郑洁，女，主治医师，主要从事乳腺癌方面的研究。　△ **通信作者**，E-mail：julianahu2021@163.com。

Conclusion The 6 lncRNAs have certain prognostic value in patients with estrogen receptor positive breast cancer,which might provide a basis for the research of reversing tamoxifen resistance and treating cancer.

Key words:breast cancer; estrogen receptor; tamoxifen-resistant; prognostic model

乳腺癌是女性最常见的恶性肿瘤。根据最新全球癌症数据统计,乳腺癌已上升为全球发病率第一的恶性肿瘤,其中约有 75%的乳腺癌患者为雌激素受体阳性的乳腺癌[1],内分泌治疗是该类乳腺癌的重要治疗策略,包括选择性雌激素受体调节剂、芳香化酶抑制剂、孕激素、卵巢功能抑制剂等。他莫昔芬可使雌激素受体阳性乳腺癌患者的 5 年复发风险降至 33.2%,5 年病死率降至 25.6%[2],但仍有 40%左右的内分泌治疗患者出现原发或继发他莫昔芬耐药,导致复发转移,影响预后。因此,临床亟待找到可以早期预估雌激素受体阳性乳腺癌患者耐药风险的靶点。

长链非编码 RNA(lncRNA)是一类长度超过 200 个核苷酸的非编码 RNA。近年研究发现,lncRNA 具有多种重要的功能,可影响基因转录调控、转录后修饰、表观遗传修饰等多种生理、病理过程,其转录或功能的异常可促进或抑制肿瘤的转移和耐药性产生[3-4]。多项研究发现,lncRNA 的表达差异与乳腺癌的发生、发展、预后及治疗耐药密切相关[5-7]。

近年来,基于高通量平台的微阵列已成为筛选癌症发生过程中重要的遗传或表观遗传学改变的有效工具,并且利用该技术去寻找癌症诊断和预后的潜在生物标志物具有广阔前景。本研究利用基因表达综合(GEO)数据库提取雌激素受体阳性乳腺癌芯片测序数据,并筛选雌激素受体阳性乳腺癌患者群体中出现他莫昔芬耐药的 lncRNA,从而分析乳腺癌他莫昔芬耐药的分子机制及治疗靶点。

1 材料与方法

1.1 数据下载及处理 使用 R 包 GEOquery 从 GEO 数据库中下载他莫昔芬耐药及敏感细胞系的 lncRNA 表达数据集,编号为 GSE159981,用于挖掘他莫昔芬耐药相关的差异 lncRNA。TANRIC(https://ibl.mdanderson.org/tanric/design/basic/main.html)数据库是一个涵盖了 20 种癌症 lncRNA 的数据库,其中数据来源包括癌症基因组图谱(TCGA)、癌症细胞系百科全书(CCLE)及大量的独立数据集,可用于探索 lncRNA 在各种癌症中的功能及临床相关性。本研究从其中下载 TCGA 乳腺癌患者的基因表达数据集,共 837 例,并从 UCSC Xena (http://xena.ucsc.edu/)数据库中获取 TCGA 乳腺癌患者及所对应的生存数据,用于建立预后模型。

1.2 生物信息分析

1.2.1 差异 lncRNA 分析 使用 R 包 GEOquery 读取 GSE159981 数据集中 GPL 20115 平台对应的注释文件。将其中标记为 lncRNA 的探针提取出来,并统一使用 lncRNA ID 进行注释。随后,使用 R 包 limma 对他莫昔芬敏感组 MCF-7 与他莫昔芬耐药组 LCC-2 中的 lncRNA 进行 t 检验。然后,按照错误发现率(FDR)矫正 P 值<0.05 且$|\log_2 FC|>1.5$(FC 为差异倍数)的标准筛选其中的差异 lncRNA。最后,使用 R 包 ggplot2 绘制差异 lncRNA 的火山图。

1.2.2 预后模型建立与分析 使用 R 包 biomaRt 注释 TCGA 基因表达谱中的基因名,并将筛选出来的差异 lncRNA 与 TCGA 的基因表达谱中包含的 lncRNA 取交集。随后,对上述交集部分的 lncRNA 进行单因素 Cox 回归分析,计算每个 lncRNA 与乳腺癌总生存率的风险值及 P 值,以 $P<0.05$ 为标准筛选出与预后显著相关的他莫昔芬耐药 lncRNA,表达量完全一致的 lncRNA 中仅保留一个。为保证模型的稳定性,在训练集中采用多因素 Cox 回归分析建立预后模型。将所建模型计算得到的风险值按中位数将患者分为高危组和低危组,使用 R 包中的 survival 和 survminer 绘制两组患者的 Kaplan-Meier 曲线并采用 Log-rank 检验比较两组患者的生存差异。

1.2.3 预后模型的验证 利用所建立的模型计算出 TCGA 乳腺癌验证集中的风险比(HR),按照相同阈值将患者划分为高危组及低危组,绘制两组患者的生存曲线并用 Log-rank 检验两组患者的生存差异。

2 结 果

2.1 差异 lncRNA 的筛选 以 FDR<0.05 及$|\log_2 FC|>1.5$ 为筛选标准,在他莫昔芬敏感组与他莫昔芬耐药组中找到差异 lncRNA 共 416 个,其中上调表达的 lncRNA 200 个,下调表达的 lncRNA 216 个。FDR 排序前十位的差异 lncRNA 和对应 P 值见表 1。

2.2 预后模型建立 将找出的差异 lncRNA 与 TCGA 基因表达谱中重叠的部分进行比对,最终获得 116 个 lncRNA。对 116 个差异表达的 lncRNA 进行单因素 Cox 回归分析,以计算各 lncRNA 与乳腺癌患者总生存率的 HR 与 P 值,得到 8 个显著相关的 lncRNA($P<0.05$),见表 2。随后,将 TCGA 随机分为训练集($n=470$)与验证集($n=157$),并在训练集中利用多因素 Cox 回归对上述 8 个 lncRNA 进行多因素 Cox 回归分析,最终确立 6 个 lncRNA(ENSG00000230440、ENSG00000231128、ENSG00000232986、ENSG00000249346、ENSG00000253898、ENSG00000258412)的预后模型。按照所建模型计算得到的 HR 的中位数进行区分,将患者分为高危组和低

危组，并进行生存分析。Log-rank 检验结果发现，高危组与低危组在训练集和验证集中的生存率差异均有统计学意义（$P=7\times10^{-7}$、0.008）。

表 1　他莫昔芬耐药相关的差异 lncRNA（FDR 排序前十位）

lncRNA ID	$\log_2$FC	平均表达水平	P	FDR
NR_109925.1	5.939 366	−0.514 31	1.55×10^{-9}	6.32×10^{-5}
ENST00000400946.2	4.172 953	−4.877 06	3.26×10^{-9}	6.67×10^{-5}
TCONS_00003745	4.181 320	−4.784 87	5.17×10^{-9}	7.06×10^{-5}
ENST00000437492.1	3.897 739	−0.091 67	1.06×10^{-8}	8.00×10^{-5}
ENST00000549807.1	5.560 677	−4.099 42	1.19×10^{-8}	8.00×10^{-5}
ENST00000562361.1	3.570 117	−5.139 26	1.19×10^{-8}	8.00×10^{-5}
ENST00000519983.1	3.324 306	−5.197 50	1.74×10^{-8}	8.00×10^{-5}
TCONS_00022864	−4.796 710	−2.507 10	1.85×10^{-8}	8.00×10^{-5}
uc002flc.1	−4.125 460	−4.792 64	1.87×10^{-8}	8.00×10^{-5}
RNA33552|snoRNA_scaRNA_148_104	3.535 041	−3.267 42	1.95×10^{-8}	8.00×10^{-5}

表 2　8 个 lncRNA 的单因素分析结果

ensembl_gene_ID	P	HR（95%CI）
ENSG00000224257	0.031	$1.1e^{18}$（44～$2.6e^{34}$）
ENSG00000230440	0.002	530（11～25 000）
ENSG00000231128	0.004	$1.1e^{-7}$（$2e^{-12}$～0.006）
ENSG00000232986	0.018	3.7（1.2～11.0）
ENSG00000249346	0.021	0.85（0.75～0.98）
ENSG00000253898	0.006	1（1～1）
ENSG00000258412	0.005	$6.6e^{-5}$（$7.8e^{-8}$～0.056）
ENSG00000259583	0.048	0.098（0.009 9～0.980 0）

2.3　6 个 lncRNA 预测预后的受试者工作特征（ROC）曲线　使用 R 包 timeROC 分别计算 3 年及 5 年生存率的曲线下面积（AUC），并绘制出相应的 ROC 曲线以评价模型的特异度和灵敏度。在整体数据集中，所构建的 6 个 lncRNA 预后模型的 3 年和 5 年生存率 AUC 分别为 0.75 和 0.68（图 1），均能较好的预测出乳腺癌患者的生存情况。

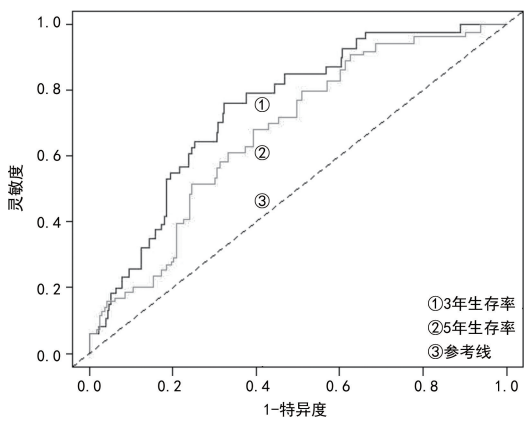

图 1　6 个 lncRNA 预测预后的 ROC 曲线

3　讨　论

他莫昔芬是一种结构与雌激素相似的人工合成的非甾体类抗雌激素药，它通过与雌激素竞争肿瘤细胞雌激素受体，减少雌激素与受体有效结合，阻止雌激素发挥有效作用，从而抑制乳腺癌细胞的增殖[8]。虽然他莫昔芬的使用明显改善了雌激素受体阳性乳腺癌患者的预后，但不可忽视的耐药问题严重影响了他莫昔芬的整体疗效，因此找到特异性的且有治疗意义的内分泌治疗耐药靶点具有重要的临床意义。

在对他莫昔芬耐药的探索中，lncRNA 的作用越来越被人重视，也取得了一些成果，比如有 WU 等[9]发现，lncRNA UCA1 可通过激活 mTOR 信号转导途径对他莫昔芬产生耐药。多项研究表明，lncRNA HOTAIR 可与雌激素受体相互作用，增强其转录活性，从而促进他莫昔芬耐药[10-12]。李均勇等[13]研究发现，lncRNA GAS5 在 MCF-7/他莫昔芬耐药细胞中呈低表达，lncRNA GAS5 过表达后 MCF-7/他莫昔芬耐药细胞增殖活性降低、对他莫昔芬敏感性增强，其机制可能与靶向调控 miR-223-5p，进而抑制下游 PI3K/Akt 通路表达有关。另有研究发现，lncRNA BCAR4 可通过 HER2 信号通路参与乳腺癌的细胞侵袭和他莫昔芬耐药[12]。

本文主要是基于 GEO 及 TCGA 数据库对他莫昔芬在乳腺癌治疗中的耐药机制进行研究，从中筛选出与他莫昔芬耐药相关的 lncRNA，并构建出能够用于评估患者生存状态的 6 个 lncRNA 预后模型。此预后模型提示高风险及低风险患者的生存曲线存在着明显的分离。与高风险评分患者相比，低风险评分患者的生存时间延长、预后较好，表明 lncRNA 在雌激素受体阳性乳腺癌内分泌治疗疗效及预后中可能

起着一定作用。

目前,对于乳腺癌内分泌治疗患者来说,仍然缺少能够有效判断内分泌治疗疗效及预后的工具。本研究中较少的 lncRNA(6 个)便可预测内分泌治疗的效果及预后,为乳腺癌患者的内分泌治疗方案选择提供参考。同时,本研究报道的 6 个 lncRNA 均为现有文献尚鲜见报道与他莫昔芬耐药相关的标志物,可能成为研究乳腺癌他莫昔芬耐药机制及逆转耐药的新靶点。

但本研究存在一定的局限性,由于高通量测序数据具有一定的误差及背景噪声[14],本研究虽然在分析前已对数据进行标准化及批次校正,且通过独立训练集和验证集初步验证了模型的稳定性,但结果仍需进一步结合大量临床标本及预后数据来验证其在临床应用的价值。

综上所述,本研究针对雌激素受体阳性乳腺癌构建了他莫昔芬耐药相关的 6 个 lncRNA 预后模型,并初步显示了该模型预测他莫昔芬耐药风险、预后情况及进一步逆转他莫昔芬耐药、治疗癌症的潜力。

参考文献

[1] MASOUD V, PAGÈS G. Targeted therapies in breast cancer: new challenges to fight against resistance[J]. World J Clin Oncol, 2017, 8(2): 120-134.

[2] EBCT E. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials[J]. Lancet, 2005, 365(9472): 1687-1717.

[3] WANG Y E, XU Z Y, JIANG J F, et al. Endogenous miRNA sponge lncRNA-RoR regulates Oct4, nanog, and Sox2 in human embryonic stem cell self-renewal[J]. Dev Cell, 2013, 25(1): 69-80.

[4] XU S, WANG P, YOU Z, et al. The long non-coding RNA EPB41L4A-AS2 inhibits tumor proliferation and is associated with favorable prognoses in breast cancer and other solid tumors[J]. Oncotarget, 2016, 7(15): 20704-20717.

[5] KIM J, PIAO H L, KIM B J, et al. Long noncoding RNA MALAT1 suppresses breast cancer metastasis[J]. Nat Genet, 2018, 50(12): 1705.

[6] LIU L, ZHU Y, LIU A M, et al. Long noncoding RNA LINC00511 involves in breast cancer recurrence and radioresistance by regulating STXBP4 expression via miR-185[J]. Eur Rev Med Pharmacol Sci, 2019, 23(17): 7457-7468.

[7] TOMAR D, YADAV A S, KUMAR D, et al. Non-coding RNAs as potential therapeutic targets in breast cancer[J]. Biochim Biophys Acta Gene Regul Mech, 2020, 1863(4): 194378.

[8] DING Y L, WANG H X, YU T F. Tamoxifen clinical application: research advances[J]. J Int Pharm Res, 2016, 43(2): 275-279.

[9] WU C, LUO J. Long non-coding RNA (lncRNA) urothelial carcinoma-associated 1 (UCA1) enhances tamoxifen resistance in breast cancer cells via inhibiting mTOR signaling pathway[J]. Med Sci Monit, 2016, 22: 3860-3867.

[10] XUE X, YANG Y A, ZHANG A, et al. LncRNA HOTAIR enhances ER signaling and confers tamoxifen resistance in breast cancer[J]. Oncogene, 2016, 35(21): 2746-2755.

[11] HAYES E L, LEWIS-WAMBI J S. Mechanisms of endocrine resistance in breast cancer: an overview of the proposed roles of noncoding RNA[J]. Breast Cancer Res, 2015, 17: 40.

[12] GODINHO M F, SIEUWERTS A M, LOOK M P, et al. Relevance of BCAR4 in tamoxifen resistance and tumour aggressiveness of human breast cancer[J]. Br J Cancer, 2010, 103(8): 1284-1291.

[13] 李均勇,徐璞,陈顺勤,等. LncRNA GAS5 调控 miR-223-5p 逆转乳腺癌他莫昔芬耐药性分析[J]. 中国药师,2021,24(8):468-472.

[14] LAEHNEMANN D, BORKHARDT A, MCHARDY A C. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction[J]. Brief Bioinform, 2016, 17(1): 154-179.

(收稿日期:2021-12-17 修回日期:2022-09-19)

(上接第 3226 页)

the first trimester of pregnancy to predict hypertensive disorders of pregnancy[J]. J Human Hypertens, 2022, 36(2): 184-191.

[16] PAPASTEFANOU I, WRIGHT D, SYNGELAKI A, et al. Competing-risks model for prediction of small-for-gestational-age neonate from maternal characteristics and serum pregnancy-associated plasma protein-A at 11 — 13 weeks′ gestation[J]. Ultrasound Obstet Gynecol, 2020, 56(4): 541-548.

[17] KARATOPRAK E, TOSUN O. Effects of valproic acid and levetiracetam monotherapy on carotid intima-media and epicardial adipose tissue thickness in non-obese children with epilepsy[J]. Brain Develop, 2020, 42(2): 165-170.

[18] LU W H, ZHANG W Q, SUN F et al. Correlation between occupational stress and coronary heart disease in northwestern China: a case study of Xinjiang[J]. Biomed Res Int, 2021, 2021: 8127873.

(收稿日期:2022-05-10 修回日期:2022-09-15)