

摘 要

基于转录组及 DNA 甲基化组数据对乳腺癌病理机制的研究及预后靶标的预测

目的：据国家癌症中心最新数据统计报道，乳腺癌超越肺癌成为中国女性发病率最高的肿瘤，在全球乳腺癌发病率中高居首位。中国乳腺癌 5 年生存率仍与发达地区存在差距，死亡率也逐年升高。乳腺癌仍是危害女性健康的最主要肿瘤，为中国卫生健康事业带来巨大的挑战。为此，本研究以 DNA 甲基化测序、转录组测序现代生物学测序技术获得的数据资源为基础，从多组学水平探究乳腺癌发生过程中的病理机制，挖掘乳腺癌预后标志物，为乳腺癌的治疗和诊断提供新的靶点和方向。

方法：1. 通过收集 TCGA 数据库中乳腺癌组织样本和癌旁组织样本的临床特征资料及转录组数据，预处理后筛选出差异表达基因 (DEGs)。以 DEGs 为背景数据，利用 pearson 相关性算法进行差异表达基因聚类分析，构建加权基因共表达网络，识别出模块基因。进一步研究模块基因与临床特征 Event、Age、M、N、T 之间相关性分析，选择高度相关的模块进行 GO 富集分析，并利用 cytoscape 软件搭建 PPI (蛋白互作) 网络，识别 hub 基因。最后，通过生存分析和 ROC 曲线验证 hub 基因。2. 通过收集 TCGA 数据库中乳腺癌组织样本和癌组织样本的临床特征数据及 DNA 甲基化芯片数据，进行样本质控分析后筛选出差异甲基化位点及对应的差异甲基化基因。接下来通过 pearson 相关性算法分析差异甲基化位点数据，构建加权共甲基化网络及识别共甲基化模块。然后应用 spearman 相关性分析，计算临床特征 Event、Age、M、N、T 与模块基因之间的相关性，选择与其高度相关的模块进行 GO 富集分析，并识别出其中关键甲基化位点。最后应用生存分析验证甲基化位点。

结果：1. 转录组数据共有 1217 个样本，包含 1072 个癌组织样本，99 个癌旁组织样本，60488 个基因，从中筛选出 14174 个 DEG (设置 $\log_2\text{FoldChange} > 0$, $\text{padj} < 0.05$)，包括上调基因 2804 个、下调基因 2052 个。在加权基因共表达网络分析中聚类识别出 6 个模块，其中，与临床特征 Event 高度相关的基因模块为 blue 模块；与临床特征 Age 高度相关的基因模块为 turquoise 模块；与临床特

征 M 高度相关的基因模块为 yellow 模块和 brown 模块；与临床特征 N 高度相关的基因模块为 yellow 模块；与临床特征 T 高度相关的基因模块为 blue 模块。与临床特征 Event 高度相关的 blue 模块主要参与细胞分裂及细胞周期的调控等；与临床特征 M、N 高度相关的 yellow 模块主要参与细胞外基质的调控、微丝形成的调控、骨骼系统的发育等；与临床特征 Age、T 高度相关的 turquoise 模块主要涉及肌肉系统的活动及发育，离子通道激活调控等生物过程。根据生存分析和 ROC 曲线结果发现九个关键基因（CDK1、COL11A1、COL10A1、MMP13、CENPF、COMP、TPX2、NCAPG、AURKA）与乳腺癌预后及诊断密切相关。2. 乳腺癌的甲基化数据共有 1888 个样本，包含 1785 个癌组织样本，98 个癌旁组织样本，筛选出 221086 个差异甲基化位点，对应 20622 个差异甲基化基因。在差异甲基化共表达网络分析里识别出 9 个模块，与临床特征 Event 高度相关的基因模块为 blue 模块；与临床特征 Age 高度相关的基因模块为 turquoise 模块；与临床特征 M 高度相关的基因模块为 yellow 模块；与临床特征 N 高度相关的基因模块为 turquoise 模块；与临床特征 T 高度相关的基因模块为 blue 模块。富集分析结果显示 blue 模块在生物过程中主要参与胚胎器官的发育、泌尿生殖系统的发育等，在细胞组成中参与细胞微丝、肌动蛋白细胞骨架形成等，在分子功能中参与 DNA 转录、蛋白质丝氨酸/苏氨酸/酪氨酸激酶信号转导通路调控等生物学功能。K-M 生存分析显示该模块中 TECR (cg23690893)、MAD1L1 (cg05434287)、ASAP1 (cg07799299)、SLC12A4 (cg01817009) 这四个甲基化位点甲基化状态与肿瘤的预后生存密切相关。

结论：1. 从转录组水平及 DNA 甲基化水平上分析，本研究发现细胞外基质的调控、微丝及肌动蛋白骨架形成的调控、以及骨骼系统的发育，可能共同影响肿瘤的微环境促进了乳腺癌的复发和转移。同时细胞分裂、细胞周期的调控及细胞内信号通路的激活等生物过程也与乳腺癌的侵袭、增殖密切相关。

2. CDK1、COL11A1、COL10A1、MMP13、CENPF、COMP、TPX2、NCAPG、AURKA 在乳腺癌的发生、转移上发挥重要作用，并与乳腺癌的预后及诊断密切相关。TECR (cg23690893)、MAD1L1 (cg05434287)、ASAP1 (cg07799299)、SLC12A4 (cg01817009) 四个基因的甲基化位点与乳腺癌预后密切相关，上述基因可能成为乳腺癌新的生物标志物。

关键词：基因共表达网络，乳腺癌，差异基因，差异甲基化位点，多组学

Abstract

Research on the Pathological Mechanism of breast cancer and Prediction of Prognostic Targets Based on Transcriptome and DNA Methylation Group Data

Objective: According to the latest statistics from the National Cancer Center, breast cancer has surpassed lung cancer to become the tumor with the highest incidence rate among Chinese women and the highest incidence rate of breast cancer in the world. The 5-year survival rate of breast cancer in China still lags behind that in developed regions, and the mortality rate is increasing year by year. Breast cancer is still the most important tumor endangering women's health, which brings huge challenges to China's health care. Therefore, based on the data resources obtained by DNA methylation sequencing and transcriptome sequencing modern biological sequencing technology, this study explores the pathological mechanism in the process of breast cancer from a multi group level, excavates prognostic markers of breast cancer, and provides some new therapeutic targets and directions for breast cancer.

Methods: 1. By collecting the clinical characteristics and transcriptome data of breast cancer tissue samples and paracancerous tissue samples in the TCGA database, the differential expression genes (DEGs) were screened after pretreatment. With DEGs as background data, pearson correlation algorithm is used to cluster differentially expressed genes, construct weighted gene co-expression network, and identify module genes. Further study the correlation analysis between module genes and clinical features Event, Age, M, N, T, select highly correlated modules for GO enrichment analysis, and use cytoscape software to build PPI (protein interaction) network to identify hub genes. Finally, the hub gene was verified by survival analysis. 2. By collecting the clinical characteristic data and DNA methylation chip data of breast cancer

tissue samples and cancer tissue samples in the TCGA database, we screened out the differential methylation sites and corresponding differential methylation genes after sample quality control analysis. Next, pearson correlation algorithm is used to analyze the data of differential methylation sites, build a weighted co-methylation network and identify the co-methylation module. Then use spearman correlation analysis to calculate the correlation between clinical features Event, Age, M, N, T and module genes, select the modules highly related to them for GO enrichment analysis, and identify the key methylation sites. Finally, the methylation sites were verified by survival analysis.

Results: 1. There are 1217 samples of transcriptome data, including 1072 cancer tissue samples, 99 para-cancer tissue samples, 60488 genes, and 14174 DEG (set $\log_2\text{FoldChange} > 0$, $\text{padj} < 0.05$), including 2804 up-regulated genes and 2052 down-regulated genes. In the weighted gene co-expression network analysis, six modules were identified by clustering, among which the gene module highly related to clinical feature Event was blue module; The gene module highly related to clinical feature Age is turquoise module; The gene modules highly related to clinical characteristics M are yellow module and brown module; The gene module highly related to clinical feature N is yellow module; The gene module highly related to clinical feature T is blue module. The blue module highly related to the clinical characteristics of Event is mainly involved in the regulation of cell division and cell cycle; The yellow module, which is highly related to clinical characteristics M and N, is mainly involved in the regulation of extracellular matrix, the regulation of microfilament formation, and the development of skeletal system; The turquoise module, which is highly related to clinical characteristics Age and T, mainly involves biological processes such as the activity and development of muscle system, activation and regulation of ion channels. According to the survival score and ROC curve, nine key genes (CDK1、COL11A1、

COL10A1、MMP13、CENPF、COMP、TPX2、NCAPG、AURKA) are closely related to the prognosis of breast cancer. 2. There were 1888 samples of methylation data of breast cancer, including 1785 cancer tissue samples and 98 adjacent tissue samples. 221086 differential methylation sites were screened, corresponding to 20622 differential methylation genes. In the analysis of differential methylation co-expression network, 9 modules were identified, and the gene module highly related to clinical feature Event was blue module; The gene module highly related to clinical feature Age is turquoise module; The gene module highly related to clinical feature M is yellow module; The gene module highly related to clinical feature N is turquoise module; The gene module highly related to clinical feature T is blue module. The enrichment analysis results showed that the blue module was mainly involved in the development of embryonic organs, the development of urogenital system, the formation of microfilament and actin cytoskeleton in cell composition, and the regulation of DNA transcription, protein serine /threonine /tyrosine kinase signal transduction pathway in molecular function. K-M survival analysis showed that the methylation status of four methylation sites in the module, TECR (cg23690893), MAD1L1 (cg05434287), ASAP1 (cg07799299), SLC12A4 (cg01817009), was closely related to the prognosis and survival of the tumor.

Conclusion: 1. From the transcriptome level and DNA methylation level, this study found that the regulation of extracellular matrix, the regulation of microfilament and actin skeleton formation, and the development of skeletal system may jointly affect the tumor microenvironment to promote the recurrence and metastasis of breast cancer. At the same time, biological processes such as cell division, cell cycle regulation and activation of intracellular signaling pathways are also closely related to the invasion and proliferation of breast cancer. 2. CDK1, COL11A1, COL10A1, MMP13, CENPF, COMP, TPX2, NCAPG, AURKA play an important role in the occurrence and metastasis of breast cancer, and are closely related to the

prognosis and diagnosis of breast cancer. The methylation sites of four genes TECR (cg23690893), MAD1L1 (cg05434287), ASAP1 (cg07799299) and SLC12A4 (cg01817009) are closely related to the prognosis of breast cancer, and these genes may become new target markers of breast cancer.

Keywords: Gene co expression network, breast cancer, differential gene, differential methylation site, multiomics

目录

第 1 章	前言	1
第 2 章	文献综述	2
第 3 章	材料与方法.....	6
3.1	数据来源	6
3.2	应用工具	6
3.3	基于乳腺癌转录组数据集分析流程.....	6
3.3.1	数据下载及预处理	7
3.3.2	筛选差异表达基因 (DEGs)	7
3.3.3	构建加权基因共表达网络	7
3.3.4	将基因模块与临床特征关联分析.....	8
3.3.5	模块基因的 GO 功能富集分析	8
3.3.6	识别 hub 基因	8
3.3.7	验证 hub 基因	9
3.4	基于乳腺癌 DNA 甲基化芯片数据分析流程	9
3.4.1	数据下载及样本质控分析	9
3.4.2	差异甲基化位点的识别	9
3.4.3	共表达网络的构建与模块基因的识别	10
3.4.4	模块基因与临床数据相关性分析.....	10
3.4.5	模块基因富集分析.....	10
3.4.6	识别并验证关键甲基化位点	10
第 4 章	结果	11
4.1	乳腺癌转录组数据结果	11
4.1.1	数据预处理	11
4.1.2	差异表达基因分析.....	11
4.1.2	构建加权基因共表达网络	11
4.1.3	基因模块与临床特征关联分析.....	12
4.1.4	基因模块的富集分析.....	13
4.1.5	hub 基因的识别	14
4.2	乳腺癌甲基化芯片数据分析结果	16
4.2.1	样本质控结果	16
4.2.2	差异甲基化基因的识别	17
4.2.3	加权基因调控网络的构建和基因模块的鉴定	17
4.2.4	基因模块与临床特征关联分析.....	18
4.2.5	基因模块的富集分析.....	19
4.2.6	关键甲基化位点的验证	20
第 5 章	讨论	21
5.1	基于转录组数据分析结果讨论.....	21
5.2	基于 DNA 甲基化数据分析结果讨论	23

第 6 章	结论	25
参考文献		26
简介		32
致谢		33

英文缩写词

英文缩写	英文全称	中文
WGCNA	Weighted Gene Co-Expression Network Analysis	加权基因共表达网络分析
M	Metastasis	转移
N	Node	淋巴结播散情况
T	Tumor	原发肿瘤的范围和大小
HER-2	Human epidermal growth factor receptor 2	人表皮生长因子受体 2
BRCA1	breast cancer susceptibility gene	乳腺癌易感基因 1
CDO1	Cysteine Dioxygenase Type 1	半胱氨酸双加氧酶 1
DMP	Differential Methylation Probe	差异甲基化位点
DMR	Differential Methylation Region	差异甲基化区域
TCGA	the Cancer Genome Atlas database	癌症基因组图谱数据库
STRING	Search Tool for the Retrieval of Interacting Genes	STRING 数据库
PPI	Protein-Protein Interaction Networks	蛋白互作网络
DEG	differential expressed gene	差异基因
ME	Module eigengenes	模块基因
GO	Gene Ontology	基因本体论
PCA	Principal Component Analysis	主成分分析
ROC 曲线	receiver operating characteristic curve	受试者工作特征曲线
AUC	area under the curve	曲线下面积
GS	Genetic significance	基因显著性
MM	Module membership	模块隶属度
CDK1	Cyclin Dependent Kinase 1	细胞周期蛋白依赖性激酶 1
COL11A1	Collagen Type XI Alpha 1 Chain	XI 型胶原 α 1 链

COL10A1	Collagen type X alpha 1	X 型胶原 α 1 链
MMP13	Matrix Metalloproteinase 13	基质金属蛋白酶 13
COMP	Cartilage oligomeric matrix protein	软骨寡聚基质蛋白
TPX2	targeting protein for Xenopus kinesin-like protein 2	非洲爪蟾驱动蛋白样蛋白 2 的靶向蛋白
NCAPG	Non-SMC Condensin I complex subunit G	非 SMC 凝蛋白 I 复合物亚基 G
AURKA	Aurora kinase A	极光激酶 A
CENPF	Centromere protein F	着丝粒蛋白 F
TECR	Trans-2,3-Enoyl-CoA Reductase	反-2,3-酰基-CoA 还原酶
MAD1L1	Mitotic Arrest Deficient 1 Like 1	有丝分裂阻滞缺陷 1 样 1
SLC12A4	Solute Carrier Family 12 Member 4	溶质载体家族 12 成员 4
ASAP1	ArfGAP With SH3 Domain, Ankyrin Repeat And PH Domain 1	Arf 核糖激化因子 GTP 酶活化蛋白 1

第1章 前言

乳腺癌是全球女性中最常见的肿瘤，也是癌症相关死亡的第二大原因^[1]。据最新统计报道，中国女性乳腺癌发病率逐年上升。与过去相比，随着乳腺癌的早期诊断的不断提高及治疗手段不断更新，癌症的死亡率有所下降^[2]，但晚期侵袭性乳腺癌的预后仍然不佳。所以乳腺癌治疗方案的制定亟待采取精准医疗模式，旨在根据患者生活方式、基因检测结果、肿瘤分期等因素制定个体化治疗方案，从而提高患者的生存率及生存质量。因此，寻找乳腺癌与早期诊断、病理机制或预后相关的靶基因至关重要。

研究表明，表观遗传变异在乳腺癌的早期发生及转移中非常普遍，DNA 甲基化不仅被证实参与了乳腺癌的早期发展，也与不良预后密切相关^[3]。因此，DNA 甲基化作为指导早期乳腺癌诊断、分组及判断预后的分子标志物具有很高的潜力及广泛的应用范围。在后基因组时代，高通量实验技术（微列针、RNA 测序等）的高速发展，产生了大量多组学数据，推动了数据挖掘的研究。基于转录组及 DNA 甲基化组数据分析有助于了解乳腺癌的发生及发展、预测肿瘤转移及判断预后。加权基因共表达网络分析（WGCNA）是一种可以分析多样本基因表达数据的生物信息学方法^[4]。WGCNA 方法是分析表型数据和转录组数据相关关系的重要工具，它可以利用基因相关关系的权重搭建的矩阵进行分层聚类来构建共表达模块，并且提供动态分支切割，识别表型与模块间的生物学关联等功能。目前在癌症数据挖掘中，WGCNA 方法成为应用最广泛的共表达分析方法。

在此，我们基于乳腺的转录组、DNA 甲基化组及临床数据，提出了一种加权网络分析方法研究了乳腺癌发生、发展中的病理机制及和预后相关的基因，包括差异表达基因及差异甲基化基因的识别、加权网络的构建、模块下游分析、关键甲基化位点及 hub 基因的验证，为乳腺癌的预后及病理诊断探索新的发展方向。

第2章 文献综述

2.1 乳腺癌流行病学及进展

近年来，乳腺癌的发病率逐年攀升，伴有年轻化趋势^[1]。据国家癌症中心最新数据统计报道，乳腺癌超越肺癌成为中国女性发病率最高的肿瘤，在全球乳腺癌发病率中高居首位^[5]。在中国，乳腺癌较其他癌种预后较好，但5年生存率仍与发达地区存在差距，可能原因为早诊率低及晚期病例诊疗存在不规范等。乳腺癌仍是危害女性健康的最主要肿瘤，为中国卫生健康事业带来巨大的挑战。

20世纪70年代，学者们逐渐发现雌激素受体（ER）是高复发率及短生存期的独立预后指标，并且是转移性肿瘤对内分泌治疗反应的良好预测指标^[6]。雌激素受体（ER）和孕激素受体（PR）检测在晚期乳腺癌患者中普及开来，osborne等人发现含有高定量ER含量和PR阳性的患者对内分泌治疗应答率高^[7]，内分泌治疗成为受体阳性患者的标准治疗。1984年，Weinberg^[8]及其同事首次发现了人表皮生长因子受体2（HER-2），该基因编码跨膜酪氨酸激酶受体蛋白^[9]，参与Ras/丝裂原激活蛋白激酶途径，磷脂酶C途径，Janus激酶/信号转导，PI3K/Akt途径等多个主要信号传导途径^[10]，最终影响细胞增殖，运动和黏附。在30%左右的乳腺癌患者中检测到HER-2基因扩增或蛋白质过表达，该类患者通常被发现为ER及PR阴性，且淋巴结转移阳性^[8]。在乳腺癌中，HER-2基因扩增被研究发现与肿瘤转移，增殖及加速血管生成相关^[10]。以曲妥珠单抗为首的HER-2靶向药物已经成为HER-2阳性患者的最重要的选择^[11]。目前根据全基因组表达和DNA拷贝数分析，人们将乳腺癌按照雌激素受体（ER）、孕激素受体（PR）、HER-2和Ki67的表达情况主要分为4型，分别是：Luminal A/B型、HER-2阳性型、三阴型。这一分类使得乳腺癌患者避免了不必要的辅助治疗，加上癌症筛查的普及，乳腺癌的死亡率在过去15年内已经显著下降^[2]。然而，这些分类仍不能完全解释乳腺癌病理和临床异质性，仍有相对一部分患者死于晚期乳腺癌。乳腺癌是一类高度异质性疾病，主要表现在分子分型相同或分期相同的乳腺癌也可能具有不同的临床行为，对于同一种治疗方案可能存在不同的敏感性及预后，所以乳腺癌治疗方案的制定亟待采取精准医疗模式，旨在根据患者生活方式、基因检测结果、肿瘤分期等因素制定个体化治疗方案，

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/656211023132010054>