引文格式:乔娇,安美文,焦雄. 基于多组学数据的乳腺癌阶段特异性基因研究[J]. 太原理工大学学报,2026,57(5):1121-1129.

QIAO Jiao, AN Meiwen, JIAO Xiong. Stage-specific genetic studies of breast cancer based on multi-omics data [J]. Journal of Taiyuan University of Technology, 2026, 57(5): 1121-1129.

基于多组学数据的乳腺癌阶段特异性基因研究

乔 娇,安美文,焦 雄

(太原理工大学 人工智能学院,山西 晋中)

摘 要:**【目的】**精确识别癌症不同阶段的关键基因并了解其生物学功能对于深入理解癌症的发病机制至关重要。提出一种新的计算框架,提升疾病阶段相关基因识别的准确性。**【方法】**整合分析乳腺癌的基因表达、DNA甲基化和拷贝数变异样本对数据,并结合蛋白质-蛋白质相互作用(PPI)网络筛选关键基因,构建不同阶段的基因调控网络。采用加权基因共表达网络分析(WGCNA)方法,识别阶段特异性网络模块,并通过富集分析、生存分析和网络拓扑分析进行验证。**【结果】**研究鉴定出与癌症各分期相关的阶段特异性候选疾病基因。共获得7个阶段特异性基因模块,并筛选出乳腺癌三个分期的29、20和25个关键基因,其中,89.7%、75%和72%的基因与乳腺癌的发展显著相关,并验证了该计算框架在识别癌症进展相关基因方面的有效性。

关键词:乳腺癌;多组学;阶段特异性;基因调控网络;WGCNA;模块保存

中图分类号:R736.3 **文献标识码**:A

DOI:10.16355/j.tyut.1007-9432.20240687 **文章编号**:1007-9432(2026)05-1121-09

Stage-specific Genetic Studies of Breast Cancer Based on Multi-omics Data

QIAO Jiao, AN Meiwen, JIAO Xiong

(*College of Artificial Intelligence, Taiyuan University of Technology, Jinzhong, Shanxi, China*)

Abstract:**【Purposes】** Precisely identifying key genes at different stages of cancer and understanding their biological functions are essential for a deeper understanding of cancer pathogenesis. A new computational framework is proposed to improve the accuracy of identifying genes associated with disease stages. **【Methods】** Through gene expression integrat and analysis, DNA methylation, copying of multiple variant breast cancer sample pair data, and combining protein-protein interaction (PPI) networks to screen key genes, gene regulatory networks at different stages were constructed. Weighted gene co-expression network analysis (WGCNA) method was used to identify stage-specific network modules, and they were validated through enrichment analysis, survival analysis, and network topology analysis. **【Results】** Disease stage-specific candidate genes associated with various stages of cancer are identified. A total of seven stage-specific gene modules are obtained, and 29, 20, and 25 key genes in three stages of breast cancer are screened, among which 89.7%, 75%, and 72% of the genes are significantly associated with the development of breast cancer, and the effectiveness of the

收稿日期:2024-09-05;**修回日期**:2024-10-03

基金项目:山西省基础研究计划项目(202203021221063)

第一作者:乔娇,硕士研究生,(E-mail)17835266773@163.com

通信作者:焦雄,教授,博士生导师,主要从事生物信息学和影像基因组学方面的研究,(E-mial)jiaoxiong@tyut.edu.cn

computational framework in identifying genes related to cancer progression is verified.

Keywords: breast cancer; multi-omics; stage specificity; gene regulatory network; WGCNA; module preservation

癌症的发展通常源于关键基因的功能失调，这些基因负责调控细胞增殖、基因组稳定性、血管生成、侵袭和转移等生物过程[1]。这些关键基因的突变和异常表达可能促进肿瘤的形成和扩散。例如，原癌基因和抑癌基因的突变导致细胞过度增殖和存活。此外，其他基因如PD-L1，IDO1，TGF-β和IL-10等的异常表达也可能导致细胞凋亡减少、免疫逃逸以及肿瘤微环境的重塑，从而进一步促进癌症的发展。精确识别与癌症各阶段相关的驱动基因对于早期诊断、预后评估和个体化治疗具有重要意义。高通量测序技术的发展，如癌症基因组图谱计划（TCGA）[2]和国际癌症基因组联盟（ICGC）[3]等数据库提供了丰富的组学数据，涵盖基因组学、蛋白质组学、转录组学以及代谢组学等领域。这些数据的多样性使我们能够更全面地分析癌症相关基因。

目前，针对癌症分期特异性基因的研究方法主要采用单一组学数据分析和多组学数据整合分析两种方式。单组学数据分析的研究进展如下：Fu等人[4]对基因表达谱数据进行WGCNA分析和富集分析，鉴定出了49个与乳腺癌病理分期相关的枢纽基因；Wang[5]利用单因素方差分析和分层聚类方法分析RNA测序（RNA-seq）数据，识别出3 576个与胃癌分期相关的基因；Kanathezath等人[6]通过基因表达谱分析，从乳腺癌癌组织中发现13个I期、12个II期和42个III期特异性基因，这些基因可能作为潜在治疗靶点。

然而，单一组学数据不足以破译不同阶段下癌症进展中基因组的动态复杂性，可能会遗漏一些与癌症驱动因素突变相关的潜在分子。因此，通过多组学研究，分析多个层次和尺度的高维数据集，可以进一步揭示细胞及其环境的复杂性，并识别出在不同数据集中变化趋势一致的分子或临床特征。基于多组学数据的研究有以下工作：Fan等人[7]通过整合乳腺癌基因表达数据和DNA甲基化数据，构建基因调控网络，识别三个阶段的20、12和22个关键基因；Yue等人[8]对肺腺癌（LUAD）患者的多组学数据进行了分析，涵盖了下一代测序（NGS）panel、全基因组测序（WGS）、RNA-seq和染色质可及性测序（ATAC-seq）等数据集，发现晚期样本在免疫相关基因通路中的突变增多；Yao等人[9]对尿路上皮膀胱癌（UC）患者在整个疾病过程中对不同分期和分级样本进行多组学分析，他们的蛋白质基因组整合分析揭示了两种亚型的分子特征和临床预后。分支状尿路上皮癌（PUC）亚型与癌基因突变有关，原位癌（CIS）则与抑癌基因突变相关，这些研究为深入理解癌症机制提供了重要信息。

此外，一些研究者尝试结合基因组学、蛋白质组学和代谢组学数据来揭示癌症分期特异性基因的作用机制。尽管利用多组学数据的研究已经取得了一定的进展，但仍存在不足。例如，Fan等人[4]的研究中，整合多组学数据的方式过于简单，只关注了各阶段样本对的差异基因和高低甲基化基因的交集基因来构建基因调控网络，而忽略了其他可能与表型相关的基因。Yue等人的研究虽然整合了早期和晚期LUAD患者的多组学数据，但仅限于免疫相关基因通路的分析，未能深入研究其他影响肺癌进展和发展的基因调控机制。此外，Yao等人虽然使用全面的多组学分析对UC患者在整个疾病过程中的不同阶段进行了研究，但是没有明确找到不同阶段之间的特异性。

为此，我们提出了一种新的计算框架，拟采用综合方法整合配对样本的多组学数据，以深化对疾病进展机制的分子层面理解，并精准识别不同阶段关键治疗靶点。

1 材料与方法

我们探索乳腺癌不同阶段的关键特异性疾病基因及其功能，技术路线如图1所示。研究主要包括5个步骤，并额外添加模块保存分析步骤，以揭示在乳腺癌发展过程中起重要作用的生物学通路。

1.1 材料

从TCGA（38.0-August 31—2023）[10]下载乳腺浸润性癌（BRCA）基因表达谱、DNA甲基化、拷贝数变异3级数据和临床数据，数据详情见表1。为降低样本间批次效应的干扰，选取乳腺癌各分期的配对样本数据进行分析，得到Stage I、Stage II、Stage III的配对样本交集数量分别为9对、32对、15对。由于Stage IV中仅发现两个样本对交集，样本量不足，缺乏说服力，因此不将Stage IV纳入本研究的分

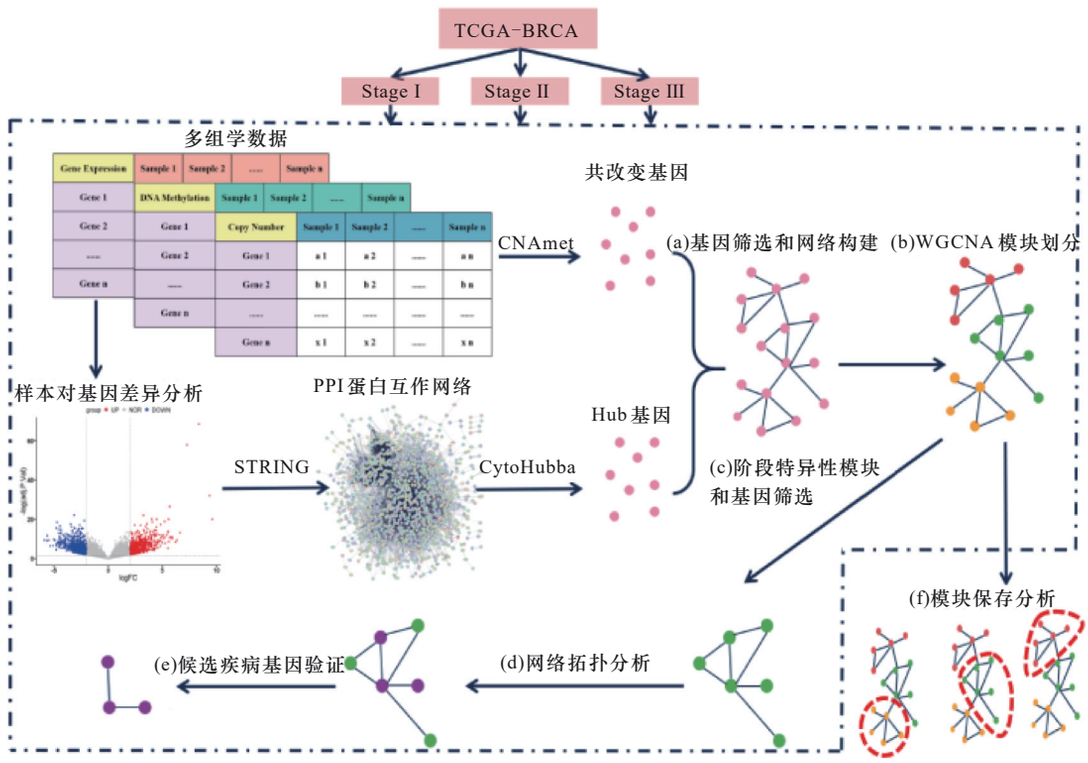


图 1　总体研究技术路线

Fig. 1　Overall research technical route

析范围。

表 1　TCGA-BRCA 数据集中各数据的整体分布情况

Table 1　Overall distributions of of data in TCGA-BRCA dataset

数据类型	检测平台	样本量/个	数据量/个
Gene expression	-	1 231	基因数：60 660
DNA methylation	Illumina human methylation 450	895	CpG 位点：486 427
Copy number	Affymetrix snp 6. 0	2 229	CNV：463 441
Clinical data	-	1 097	-

1.2　方法

1.2.1　多组学数据预处理

对于基因表达谱，下载原始的 Count 数据，提取蛋白编码的 mRNA 进行后续研究，选择表达量最大值法去除重复的基因名，得到 19 938 个基因。

对于 DNA 甲基化，使用连续的 β 值数据，范围从 0（无甲基化）到 1（完全甲基化）。首先，去除两个重复样本以及 5 个非原发肿瘤和正常实体组织样本。然后，计算每个 CpG 位点的缺失率，并筛选出缺失率低于 10% 的位点。对于剩余的缺失数据，我们采用平均值插补法进行填充。将甲基化 β 值矩阵转化为 M 值矩阵，使用 Lumi 包对其进行标准化，根据数据分布，将第一个十分位数的数据定义为低甲基化，第十个十分位数的数据定义为高甲基化，以此实现矩阵的二值化处理。

对于拷贝数变异，我们获取了 Level3 的拷贝数片段数据，并使用 GISTIC 2. 0 识别细胞拷贝数变化。通过将置信水平设置为 0. 90，确定了具有显著拷贝数变化的区域，并获得了样本中基因的离散二值化矩阵。

1.2.2　基因筛选

本研究采用了两种互补的策略来完成基因筛选，目标是全面地从多个维度理解基因在疾病中的功能。多组学数据中改变基因的筛选：使用 CNAmet 包[11]整合各阶段肿瘤样本的基因表达谱、DNA 甲基化和拷贝数变异数据，优先排序潜在的癌症驱动基因，旨在鉴定表达量上调、低甲基化和扩增的基因；或者对应的表达量下调、高甲基化和缺失的基因。拷贝数扩增会导致基因表达量的上调，而这种上调可能伴随着低甲基化的状态。同样，拷贝数缺失可能导致基因表达量的下调，而这种下调可能伴随着高甲基化的状态。通过设定 P 值小于 0. 05 的严格筛选阈值，我们成功识别了两种表达模式下的目标基因，这些基因更准确地反映了 3 个组学数据之间的相互作用关系。

蛋白质相互作用网络的枢纽基因：对基因表达谱，我们使用 R 中的 DESeq2 包[12]筛选样本对之间

的差异表达基因，筛选阈值为|logFC|大于2和adj. P值小于0.05，分别得到Stage I，Stage II，Stage III的1 364个，1 368个和1 351个差异基因。接着，利用Search Tool for the Retrieval of interaction genes（STRING，https://string-db.org/）数据库[13]构建差异表达基因（DEGs）的PPI网络，平均置信度为0.4认为相互作用是显著的。为了预测和探索PPI网络中的关键基因，首先利用CytoHubba插件[14]提供的12种拓扑分析方法来评估网络中节点的重要性，其中包括4种基于局部的方法：度中心性（Degree）、最大邻域分量（Maximum Neighborhood Component）、最大邻域分量密度（Density of Maximum Neighborhood Component）、最大团中心性（Maximum Clique Centrality），以及8种基于全局的方法：紧密中心性（Closeness）、介数中心性（Betweenness）、离心率（Eccentricity）、瓶颈中心性（Bottleneck）、聚类系数（Clustering Coefficient）、生物网络中心性（Biological Network Centrality）、邻域组件中心性（Neighborhood Component Centrality）、最大邻域组件中心性（Maximum Neighborhood Component Centrality）和边缘渗透分量（Edge Percolated Component）。对每个节点的核心度进行排名，以确定其在各个拓扑

分析方法中的相对重要性。接下来，采用Robust Rank Aggreg（RRA）[15]排序聚合算法对这12个排序结果进行综合评价，识别在所有分析结果中重要性排名靠前的基因。为每个基因分配一个显著性评分，评分小于0.05的基因被认为是PPI网络中最显著的枢纽基因。

1.2.3 构建阶段特异性基因调控网络

利用筛选得到的目标基因和具有相关性的转录因子构建不同阶段的基因调控网络。从基因调控网络数据库（Gene Regulatory Network database，GRNdb）[16]下载人类TCGA-BRCA的转录因子靶基因对，过滤出靶基因包含在筛选基因中的TF-靶基因对。根据TF-target的表达水平计算每个靶基因对的皮尔森相关系数（Pearson Correlation Coefficient，PCC），选择相关性大于0.5，P值小于0.05的基因，构建阶段特异性基因调控网络。

1.2.4 阶段特异性模块

我们利用WGCNA[17]识别基因调控网络中表达模式相似的模块。首先对样本进行聚类，去除簇之间距离大于10 000的离群样本；再构建邻接矩阵，选择通过pickSoftThreshold函数计算得出最优软阈值β；构建拓扑重叠矩阵（TOM）进行模块检测；构建共表达网络对基因进行聚类；采用Dynamic Tree Cut方法[18]进行分支切割，生成不同的网络模块。设置最小模块的基因数为30，相似性模块的合并高度为0.25。

1.2.5 网络拓扑分析

随着癌症的发展，某些基因可能会在正常状态下发生突变或表达过度，这些基因会成为基因调控网络中的关键基因。使用Cytoscape软件[19]（https://cytoscape.org/）中的Analyze network功能进行网络拓扑分析，选择介数中心度、紧密中心度和度分布这三个指标[20]作为分析依据筛选关键基因。通过与相关性分析相结合，最终确定了疾病分期特异性模块中的候选基因。

1.2.6 候选疾病基因验证

在Metascape[21]平台上，对各阶段的候选疾病基因进行了联合富集分析，以验证富集通路的特异性；接下来，利用Kaplan Meier Plotter在线数据库[22]中的2 976个样本数据，对这些基因进行了Kaplan-Meier生存分析，根据基因表达量的中位数将样本划分成高低两组，以验证基因的高低表达与乳腺癌患者生存期之间的相关性。当$P<0.05$时，认为基因的表达水平与乳腺癌患者的生存期存在显著的相关性；最后，利用人类在线孟德尔遗传数据库（Online Mendelian Inheritance in Man，OMIM），人类癌症体细胞突变数据库（Cancer Somatic Mutations Catalog，COSMIC）以及HuGE Navigator数据库[23]对基因与乳腺癌的关联进行进一步检查，同时利用PubMed研究了其他基因与乳腺癌的关系。

1.2.7 模块保存分析

癌症是一个逐步发展的连续过程，我们利用R包NetRep[24]工具对乳腺癌3个阶段的基因调控网络模块进行保存性分析，以识别跨阶段显著的模块。基于置换法和模块保存度量，以综合P值小于0.001作为测试网络模块保留的标准，筛选出在不同阶段间保持一致性的模块，并进行GO和KEGG富集分析，以发现潜在的治疗靶点和诊断标志物，深化对乳腺癌进展的理解。

2 结果

2.1 基因筛选与网络构建

通过CNAmet方法从多组学数据中筛选出的协同变化基因，结合CytoHubba算法和RRA算法从PPI网络中识别的枢纽基因，计算基因表达数据的

Pearson相关系数，筛选出靶基因对用来构建基因调控网络。各阶段筛选所得基因与基因调控网络详情如表2所示。

表2　各阶段筛选所得基因与基因调控网络信息

Table 2　Genes obtained from screening at each stage and gene regulatory network information

阶段	PPI Hub基因数/个	CNAmet共改变基因数/个	TF靶基因数/个	基因共表达网络节点/个	基因共表达网络边数/条
Stage I	476	49	5 567	717	2 621
Stage II	525	435	9 368	995	3 401
Stage III	494	156	6 431	801	2 627

2.2　阶段特异性模块

利用WGCNA将阶段特异性基因调控网络划分为具有独特表达模式的模块，如图2所示。每个模块中存在的基因数量在表3中详细列出。此外，将只存在于一个阶段而不存在于其他阶段的基因定义为阶段特异性基因。同样，确定了阶段特异性基因主要存在的模块为阶段特异性模块。发现Stage I的141个特异性基因主要集中在S1_blue和S1_turquoise模块，Stage II的368个特异性基因主要集中在S2_blue和S2_turquoise模块，Stage III的193个基因主要集中在S3_turquoise、S3_blue和S3_green模块，我们选择这7个模块作为要重点研究的特异性模块。

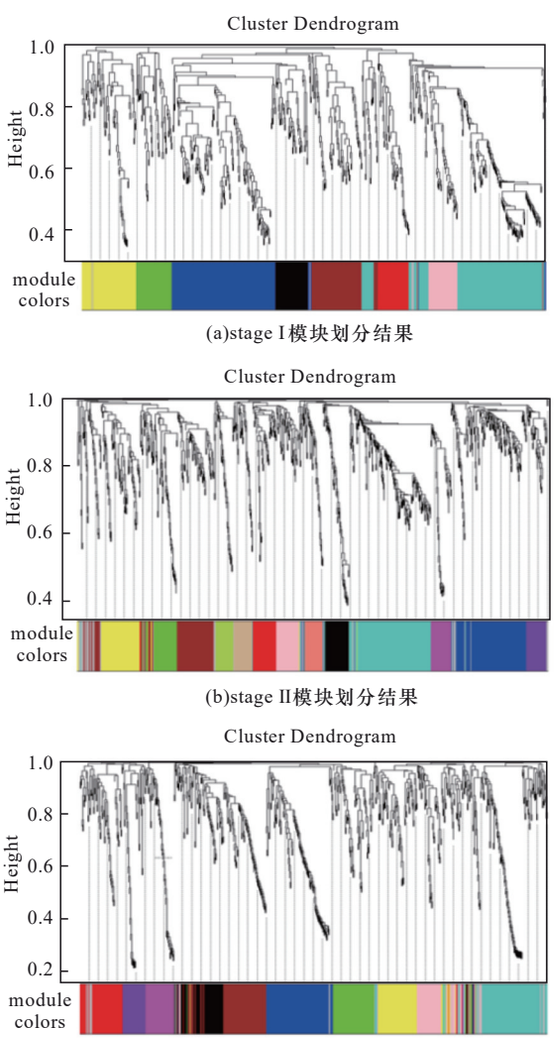
2.3　网络拓扑分析

在进行网络拓扑分析时，我们根据特定的阈值标准进行筛选：介数中心性大于0；紧凑中心性大于0.5，度分布占核心基因的前10%。在选定的7个阶段特异性模块中，进一步通过计算基因的相关性矩阵，筛选出相关性系数超过0.8且P值小于0.05的基因。综合这两种筛选策略，我们交叉鉴定出了分别对应I期、II期和III期的29、20和25个候选疾病基因，详细信息如表4所示。

2.4　候选疾病基因验证

在Metascape平台上对不同阶段的候选疾病基因进行联合富集分析，结果如图3所示。结果揭示了乳腺癌不同阶段特异性疾病基因所涉及的主要通路，表明这些基因在各个阶段所扮演的角色不同，为识别阶段特异性疾病基因在各自阶段中的重要性和有效性提供了有力验证。Stage I的基因主要与细胞周期、胚胎和组织发育、细胞分化以及肿瘤发生相关。Stage II的基因富集于细胞周期和胚胎发育。而Stage III的基因则涉及转录调控和发育生物学。特别地，有丝分裂细胞周期作为一个共有的富集途径，在三个阶段的候选疾病基因中均有显著富集，这表明细胞周期的调控机制在乳腺癌的连续发展阶段中扮演着贯穿始终的角色。

生存分析结果如补充图2所示，除Stage I中的4个基因DTL（P=0.08）、GATA4（P=0.47）、HOXA9（P=0.33）和UHRF1（P=0.15），Stage II中的3个基因ATF2（P=0.11）、DTL（P=0.08）、SPTY2D1（P=0.076），以及Stage III中的两个基因CAV1（P=0.08），EGR3（P=0.52）之外，其余基因的表达水平与乳腺癌患者的生存期存在显著的相关性。这些发现为乳腺癌的基因治疗提供了潜在靶点，并有助于开发新的预后生物标志物。



(a)stage I模块划分结果

(b)stage II模块划分结果

(c)stage III模块划分结果

图2　各阶段WGCNA模块划分结果

Fig. 2　Results of WGCNA module division in each stage

经过OMIM、COSMIC、HuGE Navigator数据库验证和PubMed搜索对乳腺癌3个阶段的候选基因进行进一步的验证，分别有26个、15个和18个枢纽基因与乳腺癌治疗和预后显著相关，验证率达到

表3 各分期模块中特异性基因的分布情况

Table 3 Distributions of specific genes in each stage module

StageI 模块颜色	对应模块的基因数量/个	阶段特异性基因数量/个	StageII 模块颜色	对应模块的基因数量/个	阶段特异性基因数量/个	StageIII 模块颜色	对应模块的基因数量/个	阶段特异性基因数量/个
black	49	14	black	48	21	black	62	19
blue	156	28	blue	143	78	blue	111	23
brown	74	19	brown	106	24	brown	85	16
green	57	18	green	53	6	green	71	22
grey	3	0	greenyellow	36	20	grey	32	12
pink	41	9	grey	52	16	magenta	49	19
red	49	19	magenta	43	25	pink	60	10
turquoise	190	21	pink	47	25	purple	39	13
yellow	73	13	purple	43	18	red	68	15
			red	51	27	turquoise	115	36
			salmon	34	17	yellow	78	8
			tan	36	24			
			turquoise	170	37			
			yellow	85	30			

表4 各分期候选疾病基因

Table 4 Candidate disease genes at each stage

阶段	候选疾病基因
Stage I	CD80#、CDC25A#、CXCL10#、CXCL11#、DTL*、E2F1#、E2F2#、E2F7*、E2F8*、EMX1、ESPL1#、GATA4*、HOXA4*、HOXA7*、HOXA9*、KIF14*、LEF1、MKI67#、MLXIPL#、MYBL2*、NEK2#、PPARG、RXRG#、TCF21#、TFF3*、TP63#、TWIST2*、UHRF1*、WT1#
Stage II	AFTPH*、ARHGAP11A、ATAD2#、ATF2*、CDC25A#、CEBPA#、DTL*、E2F1#、E2F7*、E2F8*、FBXO30、KIF14*、LMNB1、MYBL2#、NEK2#、PAPOLA*、RACGAP1#、SPTY2D1、SUZ12、TTK#
Stage III	ATF3#、CAV1*、CCNA2#、CCNE2*、CENPU*、E2F8*、EGR1*、EGR2*、EGR3*、FOSB*、FOXC1*、GBX2、HOXA3、HOXA5*、HOXB1、HOXB6、IRF2BP1、MLXIPL#、NAA30、NEIL3#、RAB5A*、SKA1、TP63#、TPX2#、VEZF1*

注：# OMIM，COSMIC，HuGE Navigator验证的基因；*PubMed验证的基因。

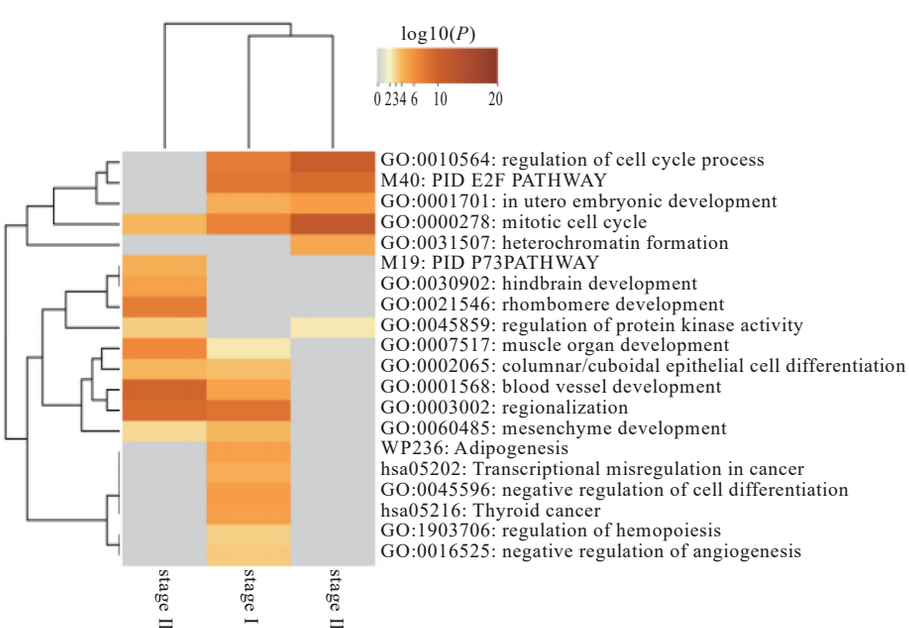


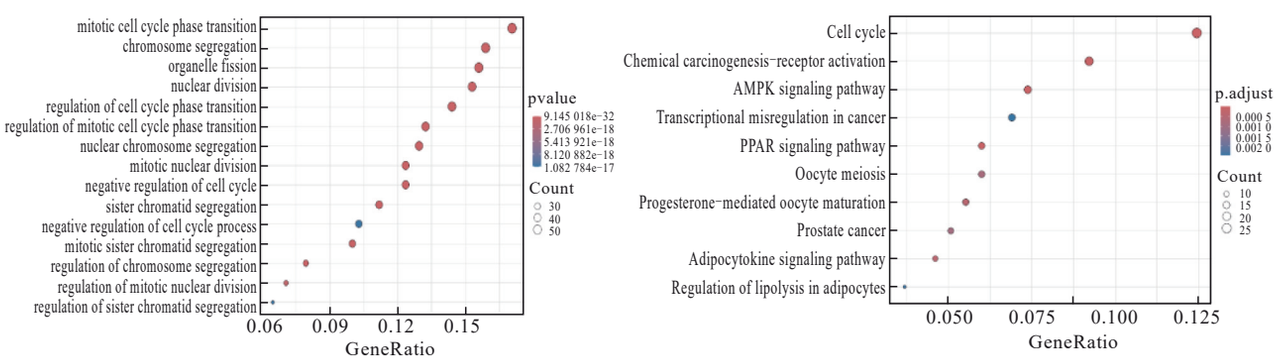
图3 阶段特异性模块的联合富集分析

Fig. 3 Joint enrichment analysis of stage-specific modules

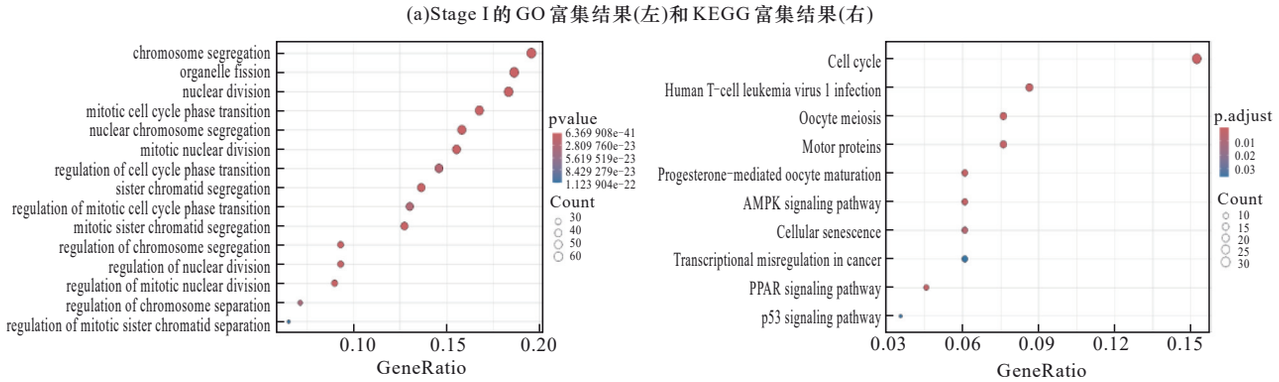
89.7%、75% 和 72%，这些枢纽基因可能作为生物标志物，有助于区分乳腺癌患者的不同病理阶段。其中 CD80 是免疫球蛋白；CDC25A 是一种磷酸酶；CXCL10 和 CXCL11 是趋化因子；E2F1 和 E2F2 是转录激活因子；ESPL1 是膜蛋白；MKI67 是一种 359 kD 的核蛋白；MLXIPL 是一种酪氨酸激酶；RXRG 是一种核受体；WT1 是一种肿瘤抑制基因；ATAD2 是一种 ATP 酶；NEK2 是丝氨酸/苏氨酸蛋白激酶；RACGAP1 是酶激活蛋白；TTK 是一种酪氨酸激酶；CCNA2 是一种细胞周期蛋白；NEIL3 是一种 DNA 修复酶；CEBPA、TCF21、TP63、MYBL2、ATF3 和 TPX2 是转录因子。

2.5　模块保存分析

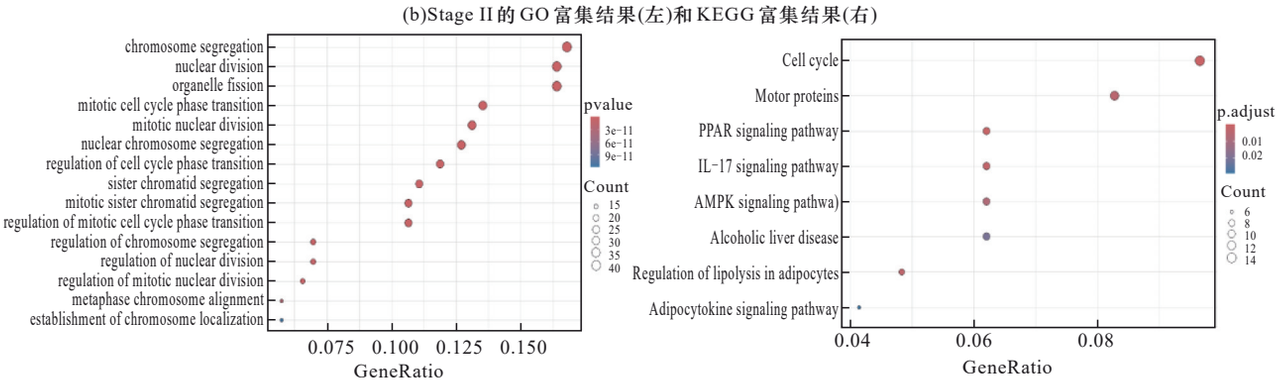
模块保存分析的计算结果如表 5 所示，Stage I 的蓝色模块和绿松石模块，Stage II 的棕色模块、绿色模块和绿松石模块，Stage III 的蓝色模块、粉色模块和黄色模块相互保存。因此，在乳腺癌的发展过程中，可能保留 2 个 I 期模块基因，3 个 II 期模块基因，2 个 III 期模块基因。对这些模块基因进行了 GO 富集和 KEGG 富集分析，并将结果可视化在图 4 中。在第一阶段，主要的 GO 富集通路集中在有丝分裂细胞周期。第二和第三阶段的主要 GO 富集通路则一致指向染色体分离。KEGG 富集分析结果表明，细胞周期在乳腺癌的发病机制和进展过程中扮演着关键角色。这些富集结果进一步验证了我们模态保存分析的准确性和阶段特异性疾病基因的生物学意义。



(a)Stage I 的 GO 富集结果(左)和 KEGG 富集结果(右)



(b)Stage II 的 GO 富集结果(左)和 KEGG 富集结果(右)



(c)Stage III 的 GO 富集结果(左)和 KEGG 富集结果(右)

图 4　各阶段模块保存基因的 GO 和 KEGG 富集分析结果

Fig. 4　Results of GO and KEGG enrichment analyses of the preserved genes in each stage module

表5 各阶段模块保存分析结果

Table 5 Analysis results of module preservation for each stage

I(D) II(T)	I(T) II(D)	II(D)III(T)	II(T)III(D)	I(D) III(T)	I(T) III(D)
Black:0.849150849	Black:0.126873127	Black:0.769230769	Black:0.098901099	Black:0.932067932	Black:0.136863137
Blue:0.000999001	Blue:0.564435564	Blue:0.822177822	Blue:0.000999001	Blue:0.000999001	Blue:0.000999001
Brown:1.000000000	Brown:0.000999001	Brown:0.000999001	Brown:0.394605395	Brown:1.000000000	Brown:0.057942508
Green:0.890109890	Green:0.000999001	Green:0.000999001	Green:0.093906094	Green:0.802197802	Green:0.034965035
Pink:0.527472527	Greenyellow: 0.997002997	Greenyellow: 0.992007992	Magenta: 1.000000000	Pink:0.468531469	Magenta: 0.990009990
Red:1.000000000	Magenta:0.391608392	Magenta:0.994005994	Pink:0.000999001	Red:0.994005994	Pink:0.000999001
Turquoise: 0.000999001	Pink:0.293706294	Pink:0.491508492	Purple:0.999000999	Turquoise: 0.000999001	Purple:1.000000000
Yellow:0.764235764	Purple:0.264735265	Purple:0.987012987	Red:0.712287712	Yellow:0.994005994	Red:0.280719281
	Red:0.978021978	Red:1.000000000	Turquoise: 0.983016983		Turquoise: 1.000000000
	Salmon:0.615384615	Salmon:0.979020979	Yellow:0.000999001		Yellow:0.000999001
	Tan:0.912087912	Tan:0.638361638			
	Turquoise: 0.000999001	Turquoise: 0.000999001			
	Yellow:0.060939061	Yellow:1.000000000			

D:验证网络;T:测试网络

3 结论

本文基于多组学数据预测乳腺癌分期特异性疾病基因,旨在识别乳腺癌分期中基因表达、DNA甲基化、拷贝数共同发生改变的基因,并结合PPI网络枢纽基因对TCGA-BRCA项目中的3个不同分期的乳腺癌进行了详尽分析。研究结果显示,通过这种方法筛选出的候选疾病基因大多数与乳腺癌有关联,验证了本文计算框架的有效性。本研究识别的分期特异性疾病基因可对未来的乳腺癌研究提供新的参考信息。然而,本研究存在一定的局限性,我们的计算框架仅基于乳腺癌的多组学数据开展分析,尚未考虑其他类型的癌症,需要进一步在不同疾病数据上开展计算,以提升其普适性。

参考文献:

[1] Zhou Y, Liu Y, Li K, et al. Ican: An integrated co-alteration network to identify ovarian cancer-related genes[J]. Plos One, 2015,10(3):e0116095.

[2] McLendon R, Friedman A, Bigner D, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways[J]. Nature, 2008,455(7216):1061-1068.

[3] Zhang J, Baran J, Cros A, et al. International cancer genome consortium data portal-a one-stop shop for cancer genomics data [J]. Database-the Journal of Biological Databases and Curation, 2011, 2011:bar026.

[4] Fu Y, Zhou Q Z, Zhang X L, et al. Identification of hub genes using co-expression network analysis in breast cancer as a tool to predict different stages[J]. Medical Science Monitor, 2019, 25:8873-8890.

[5] Wang Y. Identifying key stage-specific genes and transcription factors for gastric cancer based on rna-sequencing data[J]. Medicine, 2017, 96(4):e5691.

[6] Kanathezath A, Chembra V, Variyath SKP, et al. Identification of biomarkers and functional modules from genomic data in stage-wise breast cancer[J]. Current Bioinformatics, 2021, 16(5):722-733.

[7] Fan L, Hou J, Qin G. Prediction of disease genes based on stage-specific gene regulatory networks in breast cancer[J]. Frontiers in Genetics, 2021, 12:10.

[8] Yue D S, Liu W R, Gao L W, et al. Integrated multiomics analyses revealing different molecular profiles between early- and late-stage lung adenocarcinoma[J]. Frontiers in Oncology, 2021, 11:12.

[9] Yao Z M, Xu N, Shang G G, et al. Proteogenomics of different urothelial bladder cancer stages reveals distinct molecular features for papillary cancer and carcinoma in situ[J]. Nature Communications, 2023, 14(1):5670.

[10] Wang Z, Jensen M A, Zenklusen J C. A practical guide to the cancer genome atlas (tcga)[J]. Methods in molecular biology, 2016, 1418:111-141.

[11] Louhimo R,Hautaniemi S. Cnamet:An r package for integrating copy number,methylation and expression data[J]. Bioinformatics,2011,27(6):887-888.

[12] Anders S,Huber W. Differential expression analysis for sequence count data[J]. Genome Biology,2010,11(10):R106.

[13] Szklarczyk D,Franceschini A,Wyder S,et al. String v10:Protein-protein interaction networks,integrated over the tree of life[J]. Nucleic Acids Research,2015,43(D1):D447-D452.

[14] Chin C H,Chen S H,Wu H H,et al. Cytohubba:Identifying hub objects and sub-networks from complex interactome[J]. Bmc Systems Biology,2014,8(Suppl 4):S11.

[15] Kolde R,Laur S,Adler P,et al. Robust rank aggregation for gene list integration and meta-analysis[J]. Bioinformatics,2012,28(4):573-580.

[16] Fang L,Li Y,Ma L,et al. Grndb:Decoding the gene regulatory networks in diverse human and mouse conditions[J]. Nucleic Acids Research,2021,49(D1):D97-D103.

[17] Zhang B,Horvath S. A general framework for weighted gene co-expression network analysis[J]. Statistical Applications in Genetics and Molecular Biology,2005,4(1):Article17.

[18] Langfelder P,Zhang B,Horvath S. Defining clusters from a hierarchical cluster tree:The dynamic tree cut package for R[J]. Bioinformatics,2008,24(5):719-720.

[19] Shannon P,Markiel A,Ozier O,et al. Cytoscape:A software environment for integrated models of biomolecular interaction networks[J]. Genome Research,2003,13(11):2498-2504.

[20] Zhang J L,Luo Y. Degree centrality,betweenness centrality,and closeness centrality in social network[C]//2nd International Conference on Modelling,Simulation and Applied Mathematics (MSAM). Bangkok,Thailand,2017,132:300-303.

[21] Zhou Y,Zhou B,Pache L,et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets[J]. Nature Communications,2019,10(1):1523.

[22] Gyorffy B. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer[J]. Computational and Structural Biotechnology Journal,2021,19:4101-4109.

[23] Yu W,Gwinn M,Clyne M,et al. A navigator for human genome epidemiology[J]. Nature Genetics,2008,40(2):124-125.

[24] Ritchie S C,Watts S,Fearnley L G,et al. A scalable permutation approach reveals replication and preservation patterns of network modules in large datasets[J]. Cell Systems,2016,3(1):71-82.

(编辑:贾丽红)