



Integrated Multi-Omics Analysis Model to Identify Biomarkers Associated With Prognosis of Breast Cancer

Yeye Fan^{1†}, Chunyu Kao^{2†}, Fu Yang², Fei Wang^{3,4}, Gengshen Yin^{3,4}, Yongjiu Wang^{3,4}, Yong He^{1,2}, Jiadong Ji^{2*} and Liyuan Liu^{1,3,4*}

¹ School of Mathematics, Shandong University, Jinan, China, ² Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan, China, ³ Department of Breast Surgery, The Second Hospital, Cheeoo College of Medicine, Shandong University, Jinan, China, ⁴ Institute of Translational Medicine of Breast Disease Prevention and Treatment, Shandong University, Jinan, China

OPEN ACCESS

Edited by:

Francesco Schettini,
Institut de Recerca Biomèdica August
Pi i Sunyer (IDIBAPS), Spain

Reviewed by:

Xiaoyong Fu,
Baylor College of Medicine,
United States
Abedalrman Alkhateeb,
University of Windsor, Canada

*Correspondence:

Jiadong Ji,
jiadong@sdu.edu.cn
Liyuan Liu,
liuliyuan@sdu.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Breast Cancer,
a section of the journal
Frontiers in Oncology

Received: 19 March 2022

Accepted: 12 May 2022

Published: 10 June 2022

Citation:

Fan Y, Kao C, Yang F, Wang F, Yin G,
Wang Y, He Y, Ji J and Liu L (2022)
Integrated Multi-Omics Analysis Model
to Identify Biomarkers Associated
With Prognosis of Breast Cancer.
Front. Oncol. 12:899900.
doi: 10.3389/fonc.2022.899900

Background: With the rapid development and wide application of high-throughput sequencing technology, biomedical research has entered the era of large-scale omics data. We aim to identify genes associated with breast cancer prognosis by integrating multi-omics data.

Method: Gene-gene interactions were taken into account, and we applied two differential network methods JDINAC and LGCDG to identify differential genes. The patients were divided into case and control groups according to their survival time. The TCGA and METABRIC database were used as the training and validation set respectively.

Result: In the TCGA dataset, C11orf1, OLA1, RPL31, SPDL1 and IL33 were identified to be associated with prognosis of breast cancer. In the METABRIC database, ZNF273, ZBTB37, TRIM52, TSGA10, ZNF727, TRAF2, TSPAN17, USP28 and ZNF519 were identified as hub genes. In addition, RPL31, TMEM163 and ZNF273 were screened out in both datasets. GO enrichment analysis shows that most of these hub genes were involved in zinc ion binding.

Conclusion: In this study, a total of 15 hub genes associated with long-term survival of breast cancer were identified, which can promote understanding of the molecular mechanism of breast cancer and provide new insight into clinical research and treatment.

Keywords: multi-omics, survival prediction, differential network, breast cancer, prognosis

INTRODUCTION

Gene variation and expression play an important role in the development of cancer. Breast cancer ranks as the greatest killer among women's cancers (1). Therefore, research on genes related to long-term survival in breast cancer is of great significance for medical workers, in order to enable the development of targeted drugs and formulation of reasonable plans.

The development of high-throughput sequencing technology provides a unique opportunity for the prognostic prediction of breast cancer (2). Most of the early breast cancer studies were conducted based on single omics data such as gene expression (3, 4). For example, 70 genes related to the survival of breast cancer patients were identified by feature screening in 295 samples of breast cancer gene expression data using multivariate analysis (5). However, the development of cancer is a multiplex, multi-factorial process, involving a variety of molecular-level biological mechanisms; it is difficult for single omics analysis to elucidate the biological process of breast cancer development (6). The integration of multiple omics data is conducive to comprehending the mechanism of disease occurrence and development and can inject new blood into biological research (7). Many studies have found that integrating multiple omics data can improve clinical classification performance (8–10). Zeng integrated radiological and genomic data to predict the survival of clear-cell renal carcinoma using multiple machine learning classifiers such as logistic regression and support vector machine methods, and found that multi-omics models were more accurate than single-omics models (11).

The occurrence and development of cancer are often related to interactions between multiple genes. The heterogeneity of genomic data and the characteristics of interaction analysis result in limitations for traditional statistical methods in the application of the whole genome (12). Differential network estimation has become an important tool for exploring biological mechanisms, and the interaction patterns can provide opportunities for screening important biomarkers in disease research, which has a wide range of biological and clinical research significance (13–15). Kim used a graph-based data-fusion approach to treat multiple omics data as different nodes in a heterogeneous network for the clinical postoperative prediction of the stage, grade, and survival of ovarian cancer patients (16). Gatto constructed genome-scale metabolic-network models for 13 cancers based on the cancer genome atlas (TCGA) dataset and found that different cancers showed similar metabolic networks (17). However, most differential network models are based on single omics data, and few studies have combined multiple omics data for differential network analysis.

In this study, two advanced differential network methods for continuous and discrete data were combined to identify the differential genes and interaction networks related to breast cancer. Gene expression profiles, somatic mutations, and copy

number variations (CNVs) were collected from the cancer genome atlas (TCGA) and molecular taxonomy of breast cancer international consortium (METABRIC). By integrating genomic and transcriptomic data, we screened prognostic markers and constructed gene-interaction networks related to the long-term survival of breast cancer patients. Functional enrichment analysis was used to identify the important biological processes associated with breast cancer. The current study provides insights into the molecular mechanisms underlying breast cancer prognosis and will support the development of clinical trials and breast cancer research.

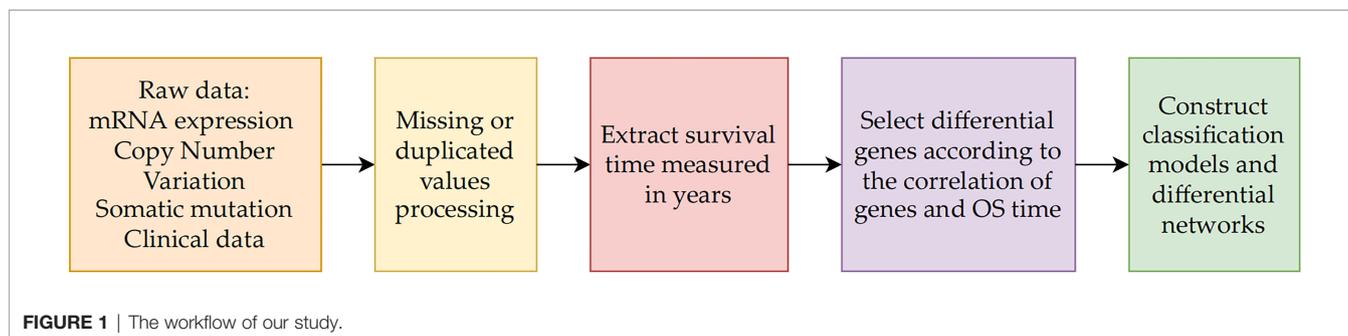
MATERIALS AND METHODS

Data Sources and Preprocessing

Omics data for mRNA gene expression, CNVs, and mutations were integrated into our study. The TCGA database was used as the training set (18). The gene expression profiles with the HTSeq-FPKM format of BRCA samples and mutation profiles were obtained directly from the data portal of TCGA (<https://portal.gdc.cancer.gov/>). We used R to convert RNAseq data from fragments per kilobase million (FPKM) format to transcripts per million (TPM) format. CNV profiles and survival data were downloaded from <http://xena.ucsc.edu/>, and the validation set METABRIC database was downloaded from <http://www.cbioportal.org/> (19–22).

The specific steps of our study were as follows: (1) The original data were obtained. (2) Gene expressions with $\geq 5\%$ missing values were deleted, and those with $< 5\%$ missing values were interpolated with the median. The expression values of genes with repeated sample IDs were replaced by the mean values. (3) The survival time (“OS.time” in TCGA and “OS_MONTHS” in META) was extracted, and the units were uniformly converted into years. (4) The coefficients of the distance correlations between genes and the survival time were calculated. (5) We selected the common differential genes in three omics of the TCGA dataset to train the classification model and construct the gene-interaction network. Finally, 966 patients in TCGA and 1,866 patients in METABRIC were used in our study. The workflow of our study is shown in **Figure 1**.

We considered four survival time categories: 1-year, 3-year, 5-year, and 10-year. The classification performance was measured using the receiver operating characteristic (ROC) curve, Kaplan-



Meier curve, area under the ROC curve (AUC), and accuracy. The high-risk group and low-risk group in Kaplan-Meier curves were truncated by the cutoff calculated by ROC curves. The differential networks were drawn to identify genes associated with breast cancer prognosis.

DC-SIS Variable Selection

The dimension of biological data is too large, and they contain many genes with little significance. To effectively utilize the data and reduce the cost of machine learning, variable selection is required. We used the sure independence screening procedure based on distance correlation (DC-SIS) method to select differential genes using the “energy” package (v1.7-8) in R (23).

The DC-SIS method measures the correlation between two random vectors according to their distance correlation coefficients (24). The distance covariance of two random vectors u and v is defined as

$$dcov(u, v) = \int_{R^{d_u+d_v}} \|\varphi_{u,v}(t, s) - \varphi_u(t)\varphi_v(s)\|^2 \omega(t, s) dt ds,$$

in which d_u and d_v are the dimensions of u and v , respectively; $\varphi_u(t)$ and $\varphi_v(s)$ are their respective eigenfunctions; $\varphi_{u,v}(t, s)$ is their joint eigenfunction; and

$$\omega(t, s) = \left\{ c_{d_u} c_{d_v} \|t\|^{1+d_u} \|s\|^{1+d_v} \right\}^{-1},$$

$$c_d = \pi^{(1+d)/2} / \Gamma\{(1+d)/2\}.$$

The distance correlation of u and v is obtained by dividing their distance covariance by the product of their distance standard deviations, which is

$$dcor(u, v) = \frac{dcov(u, v)}{\sqrt{dcov(u, u)dcov(v, v)}}.$$

Classification and Differential Network Analysis

Since gene expression data are continuous variables, whereas CNVs and mutations are discrete variables, we applied two advanced differential network estimation methods to infer the interaction networks and differential genes. In this paper, the differential network consists of difference edges that filtered out from at least two omics of data. We used Cytoscape (25) to plot the differential network. A flow chart of the study is shown in **Figure 2**. Among them, we used TCGA as the training set and METABRIC as the verification set. The omics data mRNA, CNV and somatic mutation in TCGA and mRNA and CNV in METABRIC were used to construct the differential network respectively.

For continuous-variable mRNA data, we referred to the joint density-based non-parametric differential interaction network analysis and classification (JDINAC) method to measure the interaction between the two variables, and then used L_1 -Penalized logistic regression to build the prediction model and screen the differential genes (26). A total of 50% of the data were used to fit the joint density function, and 50% were used to fit the regression model; the number of data splits was

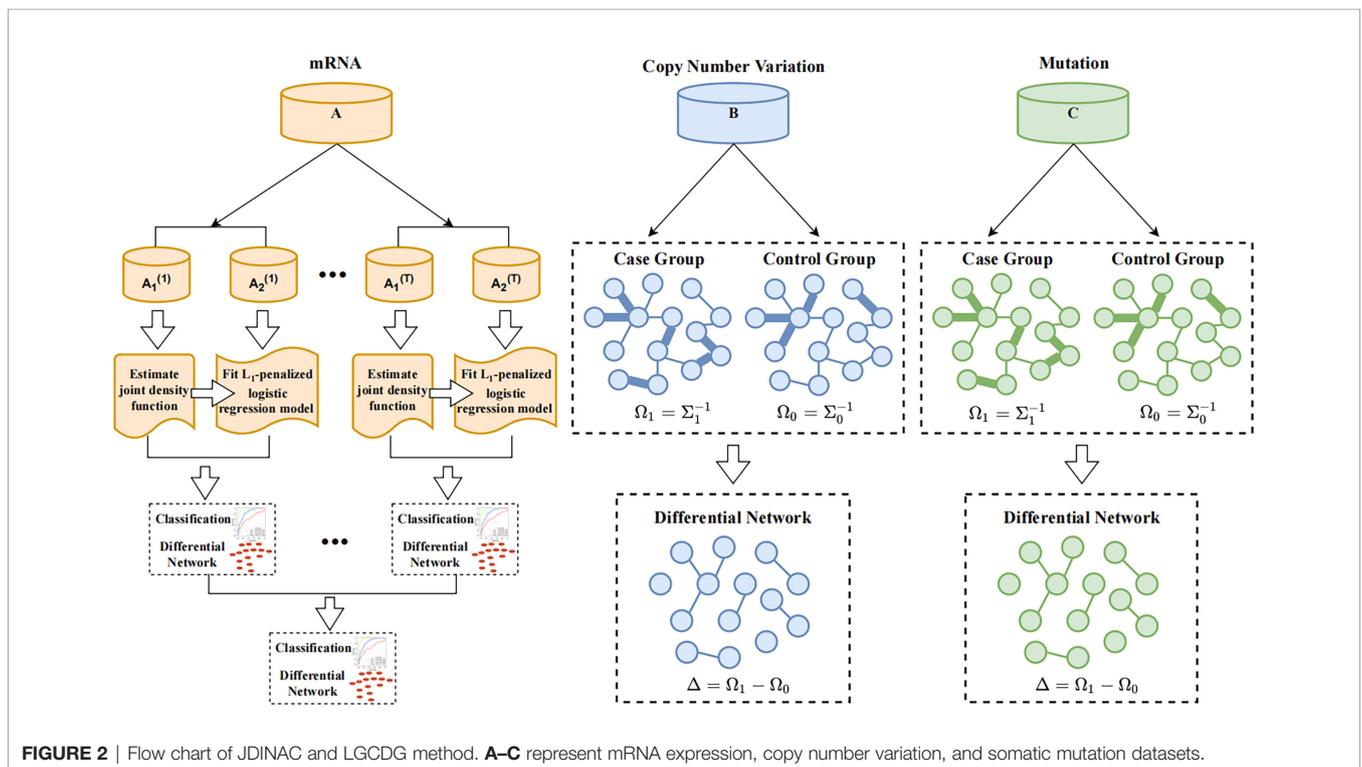


FIGURE 2 | Flow chart of JDINAC and LGCDG method. **A–C** represent mRNA expression, copy number variation, and somatic mutation datasets.

100. The mean of the predicted values is taken as the final prediction probability.

JDINAC is a nonparametric kernel method that considers gene-gene interaction, which is characterized by estimating the conditional joint density of gene pairs (26). If (x_i, x_j) denotes one gene pair, the response variable is $y = \{0,1\}$; patients with short survival time were labeled 1, and those with long survival time were labeled 0. For example, in the case of 1-year classification, samples with survival time less than or equal to 1 year were labeled 1, and those with survival time greater than 1 year were labeled 0. $f_{ij}(x_i, x_j)$ and $g_{ij}(x_i, x_j)$ represent the class conditional densities for class 1 and class 0, respectively, where $f_{ij}(x_i, x_j) = P((x_i, x_j) | y = 1)$ and $g_{ij}(x_i, x_j) = P((x_i, x_j) | y = 0)$. The log ratio of the two-dimensional class conditional density $\ln \frac{f_{ij}(x_i, x_j)}{g_{ij}(x_i, x_j)}$ was used as the classification predictor variable $\ln \frac{f_{ij}(x_i, x_j)}{g_{ij}(x_i, x_j)} > 0$ indicates that the gene pair is more closely related in class 1, whereas $\ln \frac{f_{ij}(x_i, x_j)}{g_{ij}(x_i, x_j)} < 0$ indicates that there is stronger dependency between genes in class 0. Based on the L_1 -penalized logistic regression model, the prediction accuracy is improved in the multivariate classifier, and the logistic model can be

$$\text{logit}(P) = \beta_0 + \sum_{i=1}^p \sum_{j=i+1}^p \beta_{ij} \ln \frac{f_{ij}(x_i, x_j)}{g_{ij}(x_i, x_j)}, \sum_{i=1}^p \sum_{j=i+1}^p |\beta_{ij}| < c, \quad c > 0.$$

To explore the differential networks of the discrete-variable CNV and mutation data, we applied the latent Gaussian copula differential graphical (LGCDG) model, which defines the differential network as the difference between the precision matrices of the short-term (labeled 1) and long-term (labeled 0) survival groups (27). We transferred the CNV data into binary variables; specifically, the non-zero elements were encoded as 1, indicating that the copy number is out of the normal range.

LGCDG assumes that the 0/1 binary data $D = (D_1, D_2, \dots, D_p)^T \in \{0,1\}^p$ satisfies the latent Gaussian copula model (LGCM), that is, the binary data are generated by discretizing a latent continuous variable at some unknown cutoff. In this assumption, the continuous variable follows a non-paranormal distribution, which is $X \sim \text{NPN}(\mathbf{0}, \Sigma, f)$, and the binary variable can be $D_j = I(X_j > C_j)$; then, $D \sim \text{LGCM}(\Sigma, \Lambda)$, where $\Lambda_j = f_j(C_j)$.

We assume $D^1 \sim \text{LGCM}(\Sigma^1, \Lambda^1)$ and $D^0 \sim \text{LGCM}(\Sigma^0, \Lambda^0)$ are the binary data from the case and control group, respectively. The differential network is defined as the difference between the two precision matrices, denoted by $\Delta = (\Sigma^1)^{-1} - (\Sigma^0)^{-1}$. The estimator of Δ can be obtained by solving the following optimization problem:

$$\arg \min |\Delta|_1, \text{ subject to } \left| (\hat{b}^1 \otimes \hat{b}^0) \text{Vec}(\Delta) - \text{Vec}(\hat{b}^1 - \hat{b}^0) \right|_\infty \leq \lambda_n,$$

where \hat{b}^1 and \hat{b}^0 are the Kendall's tau rank-based correlation matrix estimators for Σ^1 and Σ^0 .

The underlying differential network of binary data can be inferred through the LGCDG method, which provides a deeper understanding of the unknown mechanism than that among the observed binary variables.

GO Function Enrichment

Gene ontology (GO) enrichment analysis was performed to better understand the biological functions of the differential genes selected by DC-SIS method, and the “clusterProfiler” package (v4.2.1) and “org.hs.eg.db” package (v3.14.0) in R were used (28). With reference to the whole human genome, significant functional categories and the biological functions of the differential genes were identified.

RESULTS

DC-SIS Variable Selection

The coefficients of the distance correlations between genes and survival time were calculated; a total of 140 common genes for three omics were screened (listed in **Table S1**). The differential genes were considered as the genes highly expressed and mutated in breast cancer and were used in the following study.

Classification and Differential Network in TCGA

We compared the JDINAC model with classical binary classifiers logistic regression and random forest using 5-fold cross-validation;

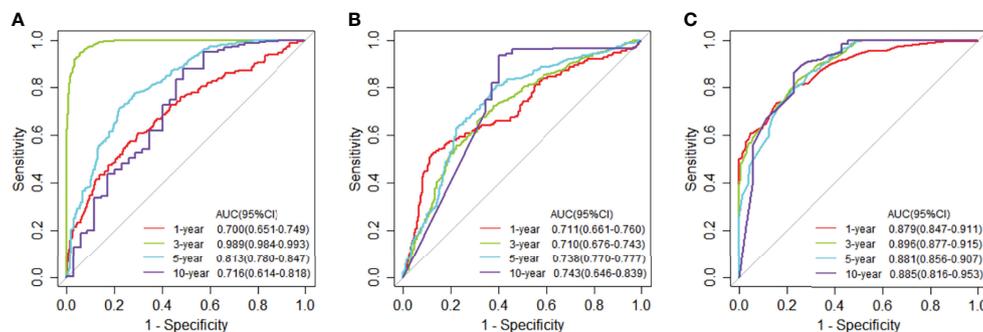
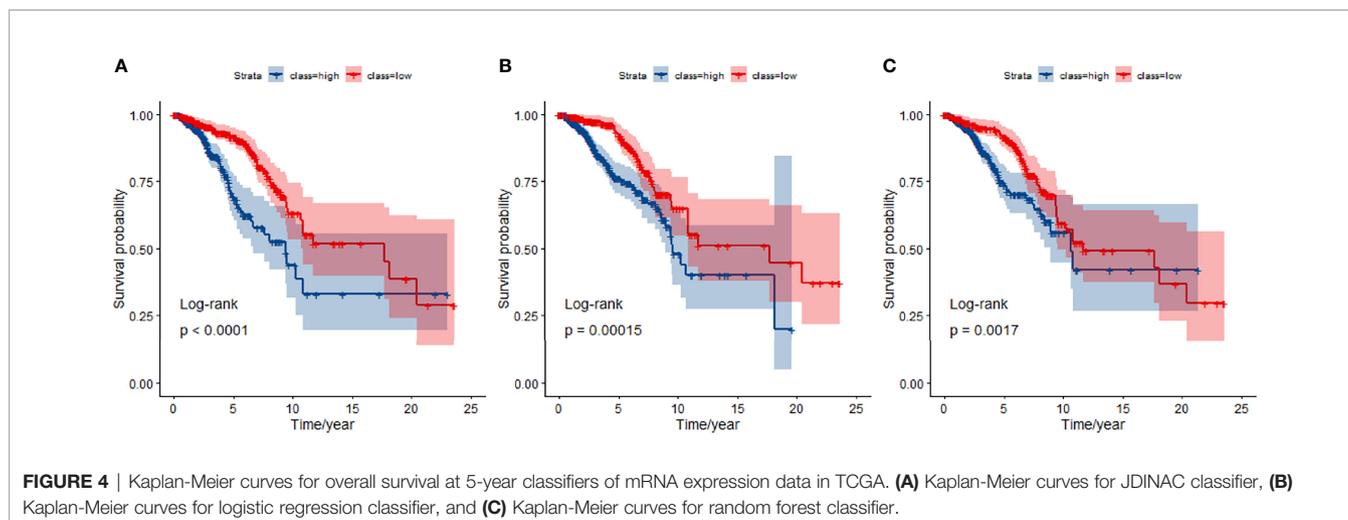


FIGURE 3 | Time-dependent receiver operating characteristic (ROC) curves at 1, 3, 5, and 10 years of mRNA expression data in TCGA. **(A)** The ROC curves for JDINAC classifier. **(B)** The ROC curves for logistic regression classifier. **(C)** The ROC curves for random forest classifier.



the ROC curves and 5-year Kaplan-Meier curves of mRNA gene expression data are shown in **Figures 3, 4**. The other Kaplan-Meier curves in TCGA dataset were shown in **Figure S1**. The classification performance of the three classifiers was measured in terms of the AUC, specificity, sensitivity, and accuracy (**Table S2**). The results show that our model has better classification performance than logistic regression model, and it can achieve comparable performance to the random forest method. In addition, the AUCs of JDINAC are all above 0.7 and even reached 0.989 in the 3-year classification category, which is sufficient to prove the efficient classification performance of JDINAC. The Kaplan-Meier curves also show that our method has better classification ability than the other two models.

The interaction networks of genes were performed by combining JDINAC and LGCDG, in which the three omics data were integrated. The differential network is composed of the common edges screened out from omics data. Genes are represented by nodes, the interactions between genes are represented by edges between nodes, and genes with at least three edges are regarded as hub genes. The hub genes were identified under four taxonomic conditions, of which C11orf1, OLA1, RPL31, SPDL1, and IL33 were identified in at least two interaction networks, and all of these genes were found in 5- or 10-year interaction networks (**Figure 5**).

Classification and Differential Network in METABRIC

We used the selected 140 differential genes in TCGA to evaluate the performance in METABRIC. Mutation data were not included in the model due to insufficient sample size. The JDINAC classification performance of the mRNA expression data was compared with the logistic regression and random forest methods using 5-fold cross-validation, and the ROC and Kaplan-Meier curves are shown in **Figures 6, 7**. The other Kaplan-Meier curves in METABRIC dataset were shown in **Figure S2**. The AUCs, specificities, sensitivities, and accuracies of the three classifiers are listed in **Table S3**. The classification performance of JDINAC is as good as that of random forest

method, and is better than that of logistic regression. The AUCs of JDINAC were all above 0.8, which indicates that the JDINAC method showed excellent classification performance.

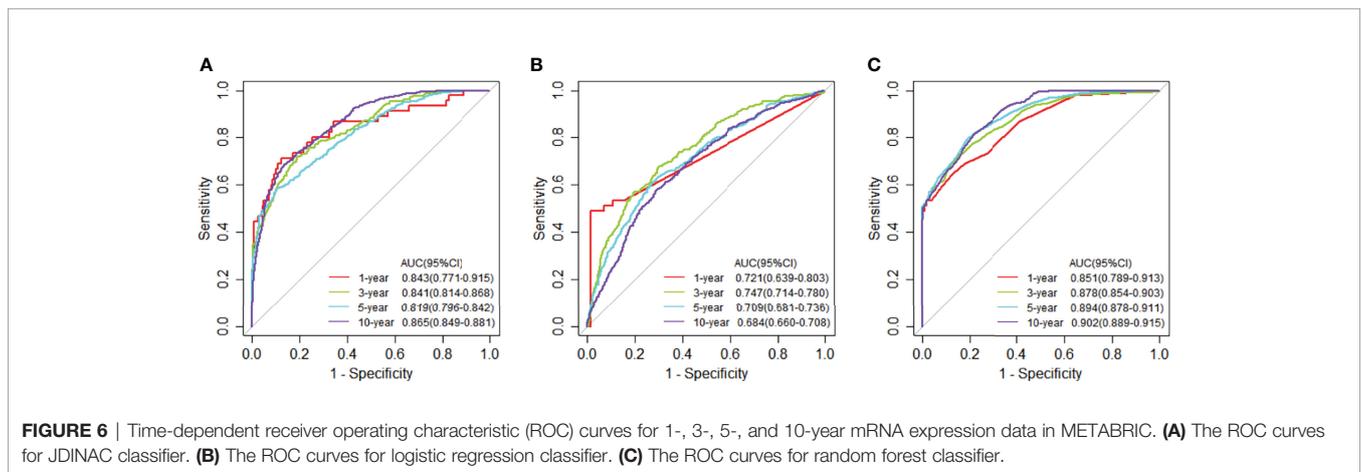
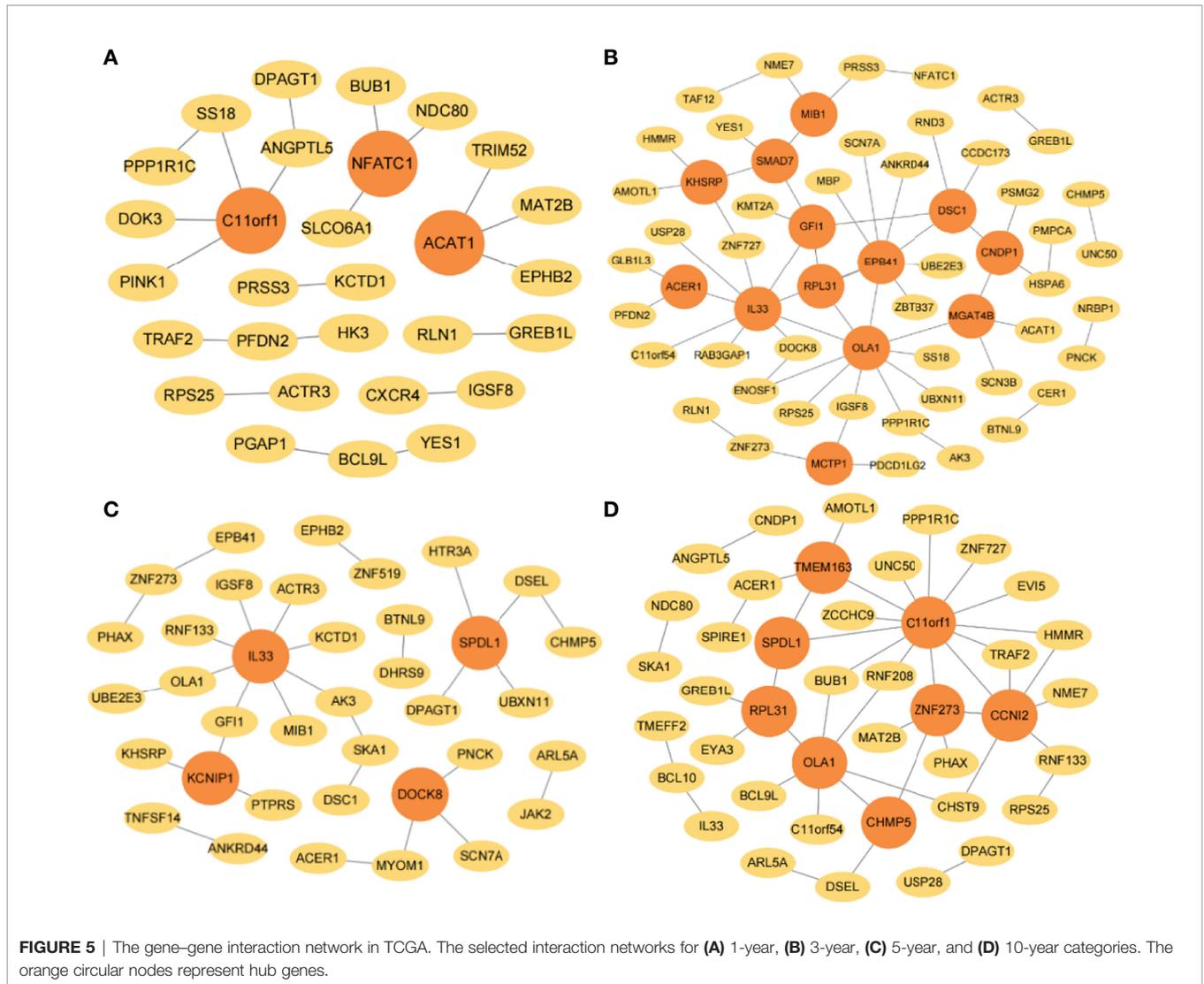
The gene-interaction networks for four classification categories were determined by combining JDINAC and LGCDG methods, and the identified hub genes are marked by orange circles (**Figure 8**). The results show that ZNF273, ZBTB37, TRIM52, TSGA10, ZNF727, TRAF2, TSPAN17, USP28, and ZNF519 were identified in at least two interaction networks, in which ZNF273, ZBTB37, and ZNF727 are related to 5-year or 10-year survival in breast cancer. Additionally, it is interesting that RPL31, TMEM163, and ZNF273 were selected as hub genes in both the TCGA and METABRIC databases, this is a finding that cannot be ignored.

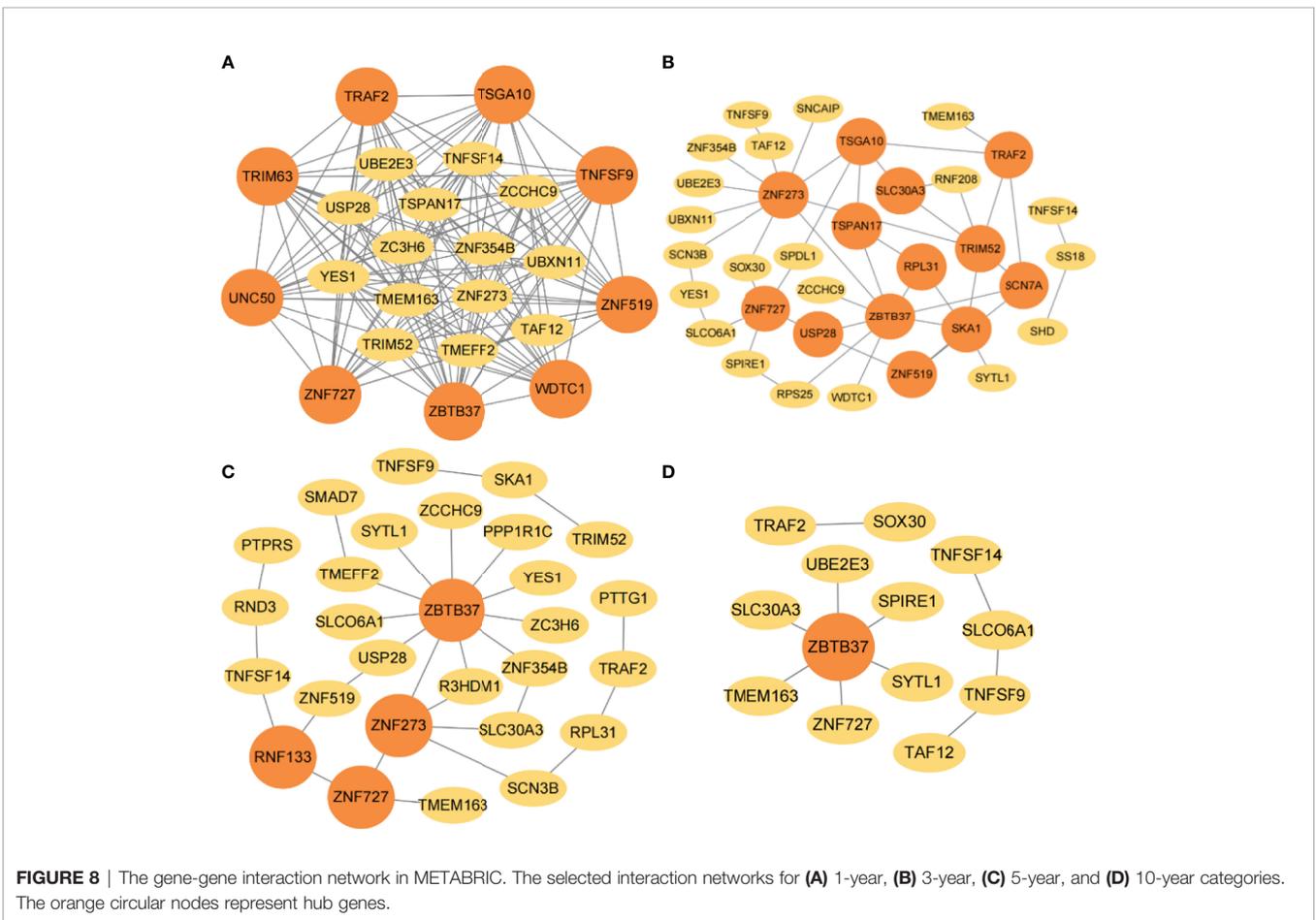
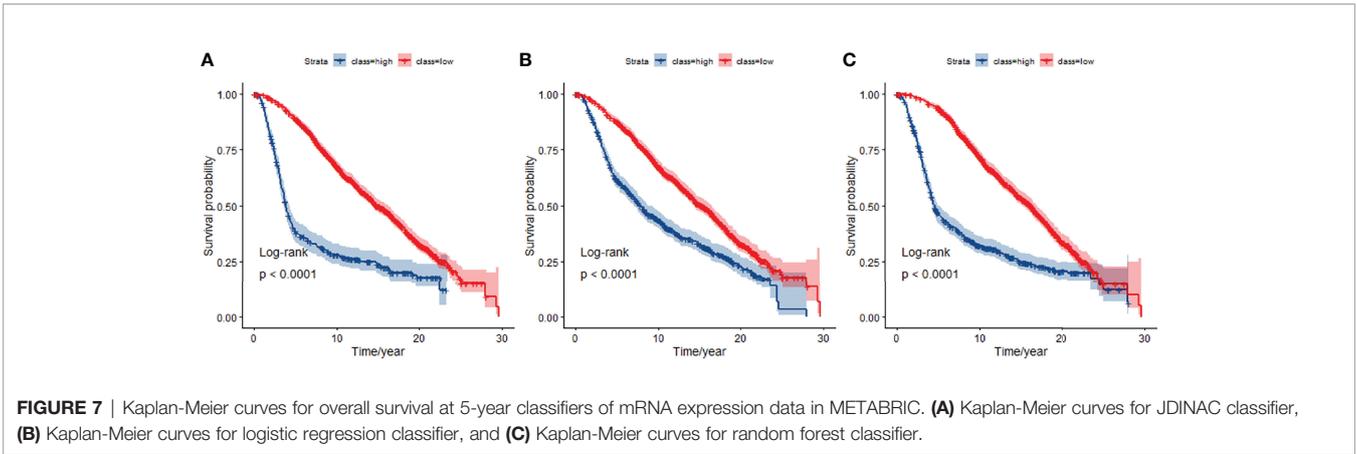
GO Function Enrichment

We performed GO enrichment analysis to assess which functional categories of genes were most connected to the prognosis of breast cancer. Enrichment analysis revealed that these differential genes were significantly enriched in 34 GO terms, mainly associated with regulation of cell-cell adhesion, positive regulation of cell activation, and positive regulation of leukocyte activation (**Figure 9**). Combining these results with TCGA and METABRIC data, a total of 15 genes related to the prognosis of breast cancer were screened out. The GO terms enriched by the 15 genes show that the metastasis and prognosis of breast cancer are closely related to zinc-ion binding (**Table S4**), which means that genes related to zinc-ion binding have significant reference value in the study of breast cancer prognosis. Among them, ZBTB37, ZNF273, ZNF519, ZNF727 and IL33 are all involved in the biological process of transcription DNA-templated, which can be used in targeted gene therapy.

DISCUSSION

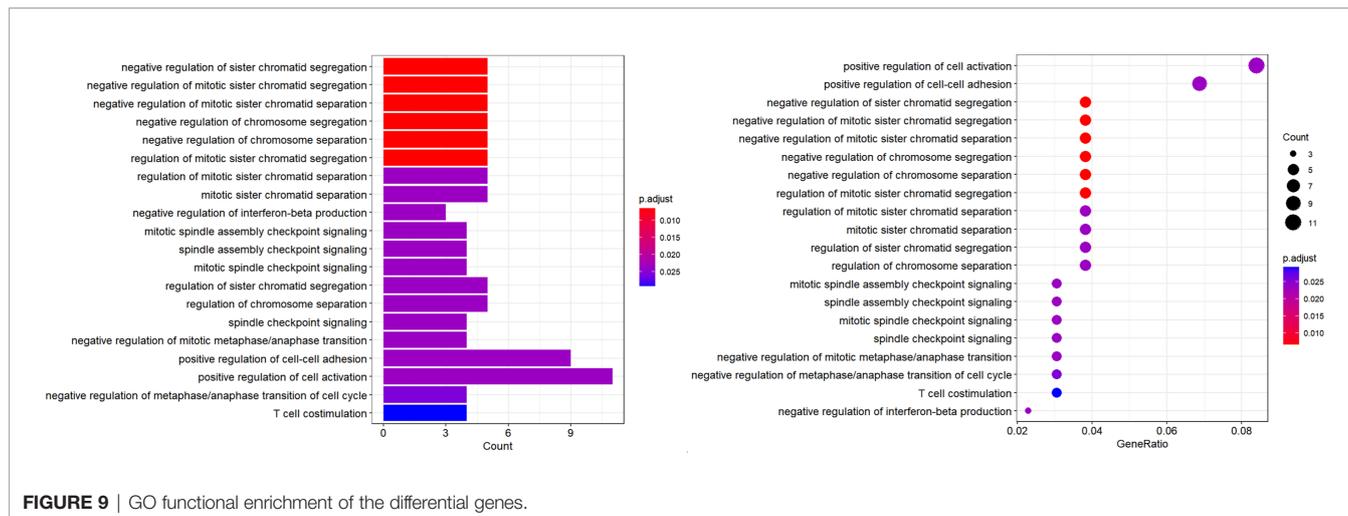
The incidence of breast cancer ranks the highest among malignant tumors in females (1). With the improvement of medical technology, the mortality rate for breast cancer has decreased significantly. However, drug resistance, recurrence, and





metastasis remain poorly addressed, resulting in low long-term survival (29). To improve the efficacy of treatment for breast cancer patients, in-depth research on potential prognostic molecular markers related to long-term survival is of great significance. In this study, we utilized multi-omics data from TCGA and METABRIC to construct gene-gene interaction networks and identify differential genes, which can provide an important basis for the clinical diagnosis of and medical research on breast cancer.

In order to avoid information redundancy, the pre-screening of differential genes is essential. We usually screen differential genes by calculating the correlation of variables according to a certain principle, such as the p-value, Pearson’s correlation coefficient, and Kendall’s tau correlation coefficient. However, the omics data include continuous and discrete data, and the traditional screening criteria often assume that the variables obey certain distributions and tend to ignore the sample information.



Compared with traditional statistical methods, DC-SIS can deal with multiple response variables, regardless of whether the response variables are continuous, discrete, or classified. It ensures that all important variables can be selected in a sufficient sample size. In addition, it does not make any model assumptions about responses and predictors, thus making model misrepresentation unlikely.

In recent years, many survival prediction models have been developed to identify prognostic biomarkers. Researchers usually use Kaplan–Meier and time-dependent ROC curves to measure predictive performance (30, 31). However, these studies are only based on the probabilities calculated using a single prediction model, which lacks discernibility in long-term survival. Zhou et al. used high-dimensional embedding and residual neural network method to extract hub genes by analyzing multi-omics data of breast cancer, but only analyzed the hub genes of each omics, lacking comprehensive consideration of multiple omics (32). In this study, we divided patients according to survival time, and constructed gene-interaction networks. Then, we focused on the differential genes associated with 5-year and 10-year survival, which makes more sense for the long-term survival of breast cancer patients.

Public sequencing platforms such as TCGA and GEO provide abundant omics data for biological researchers and facilitate molecular mechanism and clinical research. However, these datasets are highly heterogeneous, which poses significant challenges for existing approaches of data integration. There are many studies using multiple omics data and data-integration methods to analyze the survival of breast cancer patients. However, studies that combine gene interactions with multi omics are rare. Most of the studies on gene interaction focus on single omics, while the studies on multiple omics data often consider the impact of a single gene and ignore the gene interaction (33–36). We overcame the limitations of omics-data heterogeneity and applied interaction-network methods that are more suitable for multiple data types to identify hub genes.

In this study, we identified genes associated with breast cancer prognosis. Interestingly, most of the screened genes are involved in

protein binding and zinc-ion binding. The results indicate that the tumorigenesis and development of breast cancer are closely related to zinc-ion binding, which is consistent with the findings in previous studies (37–39). Many studies have found that zinc is significantly correlated with the carcinogenesis of various tissues and cells in the body, and a change in the zinc content in the human body is closely related to the occurrence and development of tumors (40–42). In addition, zinc deficiency can cause immune dysfunction, which can enhance the inflammatory effects of interleukin, inhibit the effects of interleukin on lymphocytes, and promote apoptosis, angiogenesis, and metastasis. Zinc is often involved in gene expression, the maintenance of protein and nucleic acid structure, intracellular molecular transport, and immune functions performed by zinc-finger proteins (43–45). Studies have shown that zinc lipoprotein is involved in cancer-related biological processes, can inhibit the proliferation and invasion of cancer cells, and has a protective effect on the occurrence of prostate cancer (46–48).

There are endless studies about breast cancer prognosis, and our method has several advantages compared with other methods. Firstly, we used multiple omics data for gene expression, copy number variations, and somatic mutations, making full use of multiple levels of biological information to make the study more complete. Secondly, the interaction between genes was taken into account. The correlation between genes was incorporated into the predictive variables to preliminarily explore the biological mechanisms of complex diseases. Thirdly, we classified different survival periods, mainly focusing on the genes related to long-term survival. Genes associated with 5-year and 10-year survival were identified, and their biological functions were analyzed. Finally, we solved the problem of data heterogeneity. Appropriate differential network approaches were used to estimate gene differential networks and identify hub genes.

The results for the selected genes can provide potential targets for the clinical diagnosis of breast cancer. Although we identified potential candidate genes for breast cancer prognosis using bioinformatics approaches, some limitations of this study need to be noted. First, our sample lacked clinical follow-up

information, and the database analysis based on publicly available data is not convincing enough; it needs to be verified by further clinical trials. We also lack the comprehensive consideration of clinical characteristics including age, response to different treatments, and recurrence rate in patients with different molecular subtypes, especially the hormone receptor-positive luminal vs. basal/triple-negative breast cancer (49–51). Second, the interaction between genes is a complex biological process; we only integrated existing differential-network-estimation methods to discover differential genes, which lacks innovation in methodology and comparability with other network-based approaches. Finally, the gene pre-screening process may omit some important characteristics. In the next step, we plan to focus on exploring the relationship between zinc-ion-related genes and breast cancer, and support our research through operational experiments.

CONCLUSIONS

In conclusion, we constructed a breast cancer gene-interaction network and identified genes associated with long-term breast cancer survival. The results show that there is a strong correlation between the prognosis of breast cancer and zinc-ion binding. The screened genes can be used as new prognostic markers of breast cancer, providing a new development direction for clinical research and laying a foundation for subsequent research.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

REFERENCES

- DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast Cancer Statistics, 2019. *CA Cancer J Clin* (2019) 69:438–51. doi: 10.3322/caac.21583
- Boyd SD. Diagnostic Applications of High-Throughput DNA Sequencing. *Annu Rev Pathol* (2013) 8:381–410. doi: 10.1146/annurev-pathol-020712-164026
- Begum F, Ghosh D, Tseng GC, Feingold E. Comprehensive Literature Review and Statistical Considerations for GWAS Meta-Analysis. *Nucleic Acids Res* (2012) 40:3777–84. doi: 10.1093/nar/gkr1255
- LaFramboise T. Single Nucleotide Polymorphism Arrays: A Decade of Biological, Computational and Technological Advances. *Nucleic Acids Res* (2009) 37:4181–93. doi: 10.1093/nar/gkp552
- van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N Engl J Med* (2002) 347:1999–2009. doi: 10.1056/NEJMoa021967
- Eddy S, Mariani LH, Kretzler M. Integrated Multi-Omics Approaches to Improve Classification of Chronic Kidney Disease. *Nat Rev Nephrol* (2020) 16:657–68. doi: 10.1038/s41581-020-0286-5
- Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnol Adv* (2021) 49:107739. doi: 10.1016/j.biotechadv.2021.107739
- Yang X, Regan K, Huang Y, Zhang Q, Li J, Seiwert TY, et al. Single Sample Expression-Anchored Mechanisms Predict Survival in Head and Neck

AUTHOR CONTRIBUTIONS

Conceptualization, CK and LL; Data curation, GY and YW; Formal analysis, YF, CK and FY; Funding acquisition, JJ and LL; Investigation, FY; Methodology, YH and JJ; Project administration, FW, YH, JJ and LL; Resources, GY and YW; Software, YF; Supervision, YH, JJ and LL; Validation, FY; Writing-original draft, YF and CK; Writing-review & editing, FW; Manuscript have been read and approved by all authors.

FUNDING

This work was supported by General Program of China Postdoctoral Science Foundation (2021M691911), General Program of Natural Science Foundation of Shandong Province (ZR2021MH243), National Natural Science Foundation of China (81903410) and the Young Scholars Program of Shandong University.

ACKNOWLEDGMENTS

We thank the TCGA project team and the METABRIC consortium for data access and their participants for providing sequencing data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.899900/full#supplementary-material>

- Cancer. *PLoS Comput Biol* (2012) 8:e1002350. doi: 10.1371/journal.pcbi.1002350
- Yang X, Li J, Wang Y, Li P, Zhao Y, Duan W, et al. Individualized Prediction of Survival by a 10-Long Non-Coding RNA-Based Prognostic Model for Patients With Breast Cancer. *Front Oncol* (2020) 10:515421. doi: 10.3389/fonc.2020.515421
- Yuan L, Zhao J, Sun T, Shen Z. A Machine Learning Framework That Integrates Multi-Omics Data Predicts Cancer-Related lncRNAs. *BMC Bioinf* (2021) 22:332. doi: 10.1186/s12859-021-04256-8
- Zeng H, Chen L, Wang M, Luo Y, Huang Y, Ma X. Integrative Radiogenomics Analysis for Predicting Molecular Features and Survival in Clear Cell Renal Cell Carcinoma. *Aging (Albany NY)* (2021) 13:9960–75. doi: 10.18632/aging.202752
- Ideker T, Krogan NJ. Differential Network Biology. *Mol Syst Biol* (2012) 8:565. doi: 10.1038/msb.2011.99
- Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the Human Interactome Defines Protein Communities and Disease Networks. *Nature* (2017) 545:505–9. doi: 10.1038/nature22366
- Ji J, Yuan Z, Zhang X, Xue F. A Powerful Score-Based Statistical Test for Group Difference in Weighted Biological Networks. *BMC Bioinf* (2016) 17:86. doi: 10.1186/s12859-016-0916-x
- Liu X, Ren X, Deng X, Huo Y, Xie J, Huang H, et al. A Protein Interaction Network for the Analysis of the Neuronal Differentiation of Neural Stem Cells in Response to Titanium Dioxide Nanoparticles. *Biomaterials* (2010) 31:3063–70. doi: 10.1016/j.biomaterials.2009.12.054

16. Kim D, Joung J-G, Sohn K-A, Shin H, Park YR, Ritchie MD, et al. Knowledge Boosting: A Graph-Based Integration Approach With Multi-Omics Data and Genomic Knowledge for Cancer Clinical Outcome Prediction. *J Am Med Inform Assoc* (2015) 22:109–20. doi: 10.1136/amiajnl-2013-002481
17. Gatto F, Ferreira R, Nielsen J. Pan-Cancer Analysis of the Metabolic Reaction Network. *Metab Eng* (2020) 57:51–62. doi: 10.1016/j.ymben.2019.09.006
18. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat Genet* (2013) 45:1113–20. doi: 10.1038/ng.2764
19. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. *Nature* (2012) 486:346–52. doi: 10.1038/nature10983
20. Pereira B, Chin S-F, Rueda OM, Vollan H-KM, Provenzano E, Bardwell HA, et al. The Somatic Mutation Profiles of 2,433 Breast Cancers Refines Their Genomic and Transcriptomic Landscapes. *Nat Commun* (2016) 7:11479. doi: 10.1038/ncomms11479
21. Rueda OM, Sammut S-J, Seoane JA, Chin S-F, Caswell-Jin JL, Callari M, et al. Dynamics of Breast-Cancer Relapse Reveal Late-Recurring ER-Positive Genomic Subgroups. *Nature* (2019) 567:399–404. doi: 10.1038/s41586-019-1007-8
22. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the Cbioportal. *Sci Signal* (2013) 6:pl1. doi: 10.1126/scisignal.2004088
23. Li R, Zhong W, Zhu L. Feature Screening via Distance Correlation Learning. *J Am Stat Assoc* (2012) 107:1129–39. doi: 10.1080/01621459.2012.695654
24. Székely GJ, Rizzo ML, Bakirov NK. Measuring and Testing Dependence by Correlation of Distances. *Ann Stat* (2007) 35:2769–94. doi: 10.1214/009053607000000505
25. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* (2003) 13:2498–504. doi: 10.1101/gr.1239303
26. Ji J, He Di, Feng Y, He Y, Xue F, Xie L. JDINAC: Joint Density-Based Non-Parametric Differential Interaction Network Analysis and Classification Using High-Dimensional Sparse Omics Data. *Bioinformatics* (2017) 33:3080–7. doi: 10.1093/bioinformatics/btx360
27. He Y, Ji J, Xie L, Zhang X, Xue F. A New Insight Into Underlying Disease Mechanism Through Semi-Parametric Latent Differential Network Model. *BMC Bioinf* (2018) 19:493. doi: 10.1186/s12859-018-2461-2
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for the Unification of Biology. *Gene Ontol Consortium Nat Genet* (2000) 25:25–9. doi: 10.1038/75556
29. Waks AG, Winer EP. Breast Cancer Treatment: A Review. *JAMA* (2019) 321:288–300. doi: 10.1001/jama.2018.19323
30. Wang K, Xu J, Zhao L, Liu S, Liu C, Zhang L. Prognostic lncRNA, miRNA, and mRNA Signatures in Papillary Thyroid Carcinoma. *Front Genet* (2020) 11:805. doi: 10.3389/fgene.2020.00805
31. Zhang Y, Li A, Peng C, Wang M. Improve Glioblastoma Multiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning. *IEEE/ACM Trans Comput Biol Bioinform* (2016) 13:825–35. doi: 10.1109/TCBB.2016.2551745
32. Zhou L, Rueda M, Alkhateeb A. Classification of Breast Cancer Nottingham Prognostic Index Using High-Dimensional Embedding and Residual Neural Network. *Cancers (Basel)* (2020) 14:934. doi: 10.3390/cancers14040934
33. Oulas A, Minadakis G, Zachariou M, Sokratous K, Bourdakou MM, Spyrou GM. Systems Bioinformatics: Increasing Precision of Computational Diagnostics and Therapeutics Through Network-Based Approaches. *Brief Bioinform* (2019) 20:806–24. doi: 10.1093/bib/bbx151
34. Rahimi A, Gonen M. Efficient Multitask Multiple Kernel Learning With Application to Cancer Research. *IEEE Trans Cybern* (2021). doi: 10.1109/TCYB.2021.3052357
35. Bucak SS, Rong J, Jain AK. Multiple Kernel Learning for Visual Object Recognition: A Review. *IEEE Trans Pattern Anal Mach Intell* (2014) 36:1354–69. doi: 10.1109/TPAMI.2013.212
36. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: An Ensemble of Deep-Learning and Machine-Learning Models for Prognosis Prediction Using Multi-Omics Data. *Genome Med* (2021) 13:112. doi: 10.1186/s13073-021-00930-x
37. Wang LH, Yang XY, Zhang X, Mihalic K, Fan Y-X, Xiao W, et al. Suppression of Breast Cancer by Chemical Modulation of Vulnerable Zinc Fingers in Estrogen Receptor. *Nat Med* (2004) 10:40–7. doi: 10.1038/nm969
38. Lv J, Xia K, Xu P, Sun E, Ma J, Gao S, et al. miRNA Expression Patterns in Chemoresistant Breast Cancer Tissues. *BioMed Pharmacother* (2014) 68:935–42. doi: 10.1016/j.biopha.2014.09.011
39. Lymburner S, McLeod S, Purtzki M, Roskelley C, Xu Z. Zinc Inhibits Magnesium-Dependent Migration of Human Breast Cancer MDA-MB-231 Cells on Fibronectin. *J Nutr Biochem* (2013) 24:1034–40. doi: 10.1016/j.jnutbio.2012.07.013
40. Chandler P, Kochupurakkal BS, Alam S, Richardson AL, Soybel DI, Kelleher SL. Subtype-Specific Accumulation of Intracellular Zinc Pools is Associated With the Malignant Phenotype in Breast Cancer. *Mol Cancer* (2016) 15:2. doi: 10.1186/s12943-015-0486-y
41. Chakraborty M, Hershinkel M. Zinc Signaling in the Mammary Gland: For Better and for Worse. *Biomedicines* (2021) 9:1204. doi: 10.3390/biomedicines9091204
42. da Cruz RS, Andrade Fd, Carioni VM, Rosim MP, Miranda ML, Fontelles CC, et al. Dietary Zinc Deficiency or Supplementation During Gestation Increases Breast Cancer Susceptibility in Adult Female Mice Offspring Following a J-Shaped Pattern and Through Distinct Mechanisms. *Food Chem Toxicol* (2019) 134:110813. doi: 10.1016/j.fct.2019.110813
43. Wang G, Zheng C. Zinc Finger Proteins in the Host-Virus Interplay: Multifaceted Functions Based on Their Nucleic Acid-Binding Property. *FEMS Microbiol Rev* (2021) 45:1–11. doi: 10.1093/femsre/fuaa059
44. Close P, East P, Dirac-Svejstrup AB, Hartmann H, Heron M, Maslen S, et al. DBIRD Complex Integrates Alternative mRNA Splicing With RNA Polymerase II Transcript Elongation. *Nature* (2012) 484:386–9. doi: 10.1038/nature10925
45. Palermo G, Spinello A, Saha A, Magistrato A. Frontiers of Metal-Coordinating Drug Design. *Expert Opin Drug Discovery* (2021) 16:497–511. doi: 10.1080/17460441.2021.1851188
46. Chen L, Wu X, Xie H, Yao N, Xia Y, Ma G, et al. ZFP57 Suppress Proliferation of Breast Cancer Cells Through Down-Regulation of MEST-Mediated Wnt/ β -Catenin Signalling Pathway. *Cell Death Dis* (2019) 10:169. doi: 10.1038/s41419-019-1335-5
47. Wu X, Zhang X, Yu L, Zhang C, Ye L, Ren D, et al. Zinc Finger Protein 367 Promotes Metastasis by Inhibiting the Hippo Pathway in Breast Cancer. *Oncogene* (2020) 39:2568–82. doi: 10.1038/s41388-020-1166-y
48. Costello LC, Franklin RB. The Clinical Relevance of the Metabolism of Prostate Cancer; Zinc and Tumor Suppression: Connecting the Dots. *Mol Cancer* (2006) 5:17. doi: 10.1186/1476-4598-5-17
49. de Sanctis R, Viganò A, Giuliani A, Gronchi A, de Paoli A, Navarria P, et al. Unsupervised Versus Supervised Identification of Prognostic Factors in Patients With Localized Retroperitoneal Sarcoma: A Data Clustering and Mahalanobis Distance Approach. *BioMed Res Int* (2018) 2018:2786163. doi: 10.1155/2018/2786163
50. Liao G, Jiang Z, Yang Y, Zhang C, Jiang M, Zhu J, et al. Combined Homologous Recombination Repair Deficiency and Immune Activation Analysis for Predicting Intensified Responses of Anthracycline, Cyclophosphamide and Taxane Chemotherapy in Triple-Negative Breast Cancer. *BMC Med* (2021) 19:190. doi: 10.1186/s12916-021-02068-4
51. Pedersen RN, Esen BÖ, Mellekjær L, Christiansen P, Ejlersen B, Lash TL, et al. The Incidence of Breast Cancer Recurrence 10–32 Years After Primary Diagnosis. *J Natl Cancer Inst* (2022) 114:391–9. doi: 10.1093/jnci/djab202

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Fan, Kao, Yang, Wang, Yin, Wang, He, Ji and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.