Genomics Proteomics Bioinformatics 18 (2020) 525–538

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

# Kinase–substrate Edge Biomarkers Provide A More Accurate Prognostic Prediction in ER-negative Breast Cancer

**Yidi Sun [1,2,#], Chen Li [1,#,§], Shichao Pang [3], Qianlan Yao [4], Luonan Chen [1,5,6,*] Yixue Li [4,5,7,8,9,*], Rong Zeng [1,5,*]**

[1] *CAS Key Laboratory of Systems Biology, CAS Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[2] *University of Chinese Academy of Sciences, Shanghai 200031, China*
[3] *Deptartment of Statistics, School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China*
[4] *School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China*
[5] *Department of Life Sciences, ShanghaiTech University, Shanghai 201210, China*
[6] *CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China*
[7] *Bio-Med Big Data Center, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[8] *Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai 200032, China*
[9] *Shanghai Center for Bioinformation Technology, Shanghai Academy of Science & Technology, Shanghai 201203, China*

Received 19 September 2018; revised 27 August 2019; accepted 11 November 2019
Available online 13 January 2021

Handled by Edwin Wang

**KEYWORDS**

ER-negative breast cancer;
Edge biomarkers;
Kinase;
Substrate;
Prognostic prediction

**Abstract** The estrogen receptor (ER)-negative breast cancer subtype is aggressive with few treatment options available. To identify specific prognostic factors for **ER-negative breast cancer**, this study included 705,729 and 1034 breast invasive cancer patients from the Surveillance, Epidemiology, and End Results (SEER) and The Cancer Genome Atlas (TCGA) databases, respectively. To identify key differential **kinase–substrate** node and **edge biomarkers** between ER-negative and ER-positive breast cancer patients, we adopted a network-based method using correlation coefficients between molecular pairs in the kinase regulatory network. Integrated analysis of the clinical and molecular data revealed the significant prognostic power of kinase–substrate node and edge features

* Corresponding authors.
E-mail: lnchen@sibs.ac.cn (Chen L), yxli@sibs.ac.cn (Li Y), zr@sibcb.ac.cn (Zeng R).
# Equal contribution.
§ Current address: Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2019.11.012

for both subtypes of breast cancer. Two promising kinase–substrate edge features, *CSNK1A1–NFATC3* and *SRC–OCLN*, were identified for more accurate **prognostic prediction** in ER-negative breast cancer patients.

## Introduction

Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer mortality among females worldwide [1]. Each year, 25% of all cancer occurrences and 15% of cancer deaths among females are attributed to breast cancer [2]. Based on the presence or absence of estrogen receptor (ER), this heterogeneous disease can be divided into two subtypes. The ER-positive subtype is more common and can be treated by ER modulators, but drug resistance and relapse happen frequently in this subtype [3,4]. The ER-negative subtype is less frequent but is more aggressive and associated with poor prognosis. To date, ER-negative patients have limited effective therapies, and chemotherapy is the mostly used treatment options [5]. Retrospective studies revealed that hormone therapies did not reduce the risk of ER-negative breast cancer [6]. Targeted therapies, such as various kinase inhibitors, offered more hope for the treatment of ER-negative breast cancer with the expression of *HER2* (Human epidermal growth factor receptor 2) [7]. A recent study reported that *PTEN* (Phosphatase and tensin homolog) loss in African American females was significantly correlated with the occurrence of ER-negative breast ductal cancer [8]. Inherited mutations in *PALB2* (Partner and localizer of *BRCA2*) and *FANCM* (Fanconi anemia complementation group M) are also found to be connected with the absence of ER [9]. These studies indicate the importance of exploring the genetic characteristics of breast cancer for the development of targeted therapies for ER-negative breast cancer. Several studies have identified differential gene expression patterns in the ER-negative subtype but are limited by relatively small-scale observational studies or particular geographic regions [10–12]. The Cancer Genome Atlas (TCGA) project has generated a substantial amount of data from a large number of patient samples [13,14], and the clinical utility of these genomic data has been assessed in several cancer types [15,16]. Studies based on large databases that seek to identify the potential clinical utility of molecular profiles for ER-negative breast cancer treatment are needed.

Previous studies have reported the important roles of kinases in the development of various diseases, and numerous kinases are involved in promoting cell proliferation and cancer [17]. Kinases typically contain a serine, threonine or tyrosine residue to catalyze their substrates [18]. Considering the shared conservative secondary structure element, kinases are favourable spots for targeted drugs, such as imatinib and sorafenib [19], which are effective kinase inhibitors for chronic myeloid leukaemia treatment. Although kinases are involved in many key signaling pathways in breast cancer through phosphorylating downstream substrates [12,20–22], the interactions between kinases and substrates are also critical in biological pathways and cell signaling related to other diseases [23–25]. In other words, compared with single kinase or substrate molecule, the kinase–substrate network or edge interactions constituted by these molecules are considered to be more credible and permanent for characterizing complex diseases. To date, various advances have been achieved in discovering network-based biomarkers thanks to the vast accessibility of omics data [26]. A recent study reported that the prediction of multiple phenotypes has been improved based on the pathway modules constructed from the biological network [27]. Another study identified a significant association between a module enriched with cell death genes and ovarian cancer survival based on the co-expression network approach [28]. Molecular network-based markers are generally represented as the correlation coefficient between a pair of molecules, but false positives are highly likely to occur given numerous indirect associations detected by this method [29]. To solve this problem and considering the important role of kinase network in breast cancer, we mainly focused on the kinase–substrate interaction network in this study.

The global aim of this study was to identify differential prognostic kinase–substrate network biomarkers between ER-positive and ER-negative subtypes as potential drug targets for the treatment of breast cancer patients. We first analyzed the clinicopathological features and survival probabilities of ER-positive and ER-negative subtypes from the Surveillance, Epidemiology, and End Results (SEER) and TCGA databases to identify clinical prognostic factors. In addition, we selected key differential kinase–substrate node and edge features between the two breast cancer subtypes and integrated these selected features with clinical characteristics for prognostic prediction based on TCGA data. Moreover, we explored the possibility of kinase–substrate biomarkers for prognostic prediction in ER-positive and ER-negative breast invasive carcinoma.

## Results

### Clinicopathological characteristics of ER-positive and ER-negative patients

We included 705,729 SEER and 1034 TCGA breast cancer patients in this study. More than 70% of patients were ER-positive in both databases, whereas ER-negative cases accounted for 21.3% and 22.7% of patients in SEER and TCGA, respectively. As shown in **Table 1**, in SEER, ER-negative patients were diagnosed at younger ages and later disease stages with a larger proportion of African American patients compared with ER-positive patients (Chi-squared test; $P < 0.001$). In addition, ER-negative patients had a larger proportion of poorly differentiated tumors and larger tumor sizes (Chi-squared test; $P < 0.001$). In TCGA, significant differences also existed between the ER-negative and ER-positive groups, in terms of age (Chi-squared test; $P = 0.007$), race (Chi-squared test; $P < 0.001$), and lymph node status (Chi-squared test; $P < 0.001$). No significant difference in tumor stages was observed between the two subtypes, but a larger proportion of stage II patients were found for the ER-negative subtype compared with that for the ER-positive group in TCGA (62.6% *vs* 54.5%), which was consistent with that in SEER (Table 1).

**Table 1 Clinicopathological characteristics of ER± and ER⁻ breast cancer patients**

| | SEER | | | TCGA | | |
|---|---|---|---|---|---|---|
| | **ER⁺ (%)** | **ER⁻ (%)** | ***P* value** | **ER⁺ (%)** | **ER⁻ (%)** | ***P* value** |
| **Total** | 555,151 (78.7) | 150,578 (21.3) | | 796 (77.2) | 238 (22.7) | |
| **Age** | | | < 0.001 | | | 0.007 |
| < 50 years | 125,227 (22.6) | 49,352 (32.8) | | 199 (25.0) | 77 (32.4) | |
| 50–69 years | 266,436 (48.0) | 71,315 (47.4) | | 414 (52.0) | 126 (52.9) | |
| ≥ 70 years | 163,488 (29.4) | 29,911 (19.8) | | 183 (23.0) | 35 (14.7) | |
| **Race** | | | < 0.001 | | | < 0.001 |
| Caucasian | 464,037 (83.6) | 113,860 (75.6) | | 568 (71.4) | 138 (58.0) | |
| African American | 44,967 (8.1) | 24,149 (16.0) | | 107 (13.4) | 69 (29.0) | |
| American Indian / Alaska Native | 2719 (0.5) | 888 (0.6) | | 0 (0.0) | 1 (0.4) | |
| Asian or Pacific Islander | 40,675 (7.3) | 11,085 (7.4) | | 38 (4.8) | 19 (8.0) | |
| Unknown | 2753 (0.5) | 596 (0.4) | | 83 (10.4) | 11 (4.6) | |
| **AJCC stage** | | | < 0.001 | | | 0.168 |
| I | 268,288 (48.3) | 49,879 (33.1) | | 142 (17.8) | 35 (14.7) | |
| II | 176,015 (31.7) | 57,820 (38.4) | | 434 (54.5) | 149 (62.6) | |
| III | 64,363 (11.6) | 26,992 (17.9) | | 190 (23.9) | 46 (19.3) | |
| IV | 20,878 (3.8) | 7844 (5.2) | | 14 (1.8) | 3 (1.3) | |
| Unknown | 25,607 (4.6) | 8043 (5.3) | | 16 (2.0) | 5 (2.1) | |
| **Lymph node status** | | | < 0.001 | | | < 0.001 |
| Positive | 322,321 (58.1) | 81,391 (54.1) | | 365 (45.9) | 82 (34.5) | |
| Negative | 167,893 (30.2) | 52,123 (34.6) | | 304 (38.2) | 125 (52.5) | |
| Unknown | 64,937 (11.7) | 17,064 (11.3) | | 127 (16.0) | 31 (13.0) | |
| **Tumor grade** | | | < 0.001 | | | |
| I (well differentiated) | 126,423 (22.8) | 4281 (2.8) | | | | |
| II (moderately differentiated) | 243,684 (43.9) | 27,524 (18.3) | | | | |
| III (poorly differentiated) | 128,559 (23.2) | 100,654 (66.8) | | | | |
| IV (undifferentiated) | 5529 (1.0) | 4843 (3.2) | | | | |
| Unknown | 50,956 (9.2) | 13,276 (8.8) | | | | |
| **Tumor size** | | | < 0.001 | | | |
| < 2.0 cm | 313,903 (56.5) | 57,648 (38.3) | | | | |
| 2–10 cm | 211,102 (38.0) | 78,912 (52.4) | | | | |
| Other or unknown | 30,146 (5.4) | 14,018 (9.3) | | | | |

*Note*: AJCC, American Joint Committee on Cancer; ER, estrogen receptor; SEER, the Surveillance, Epidemiology, and End Results; TCGA, The Cancer Genome Atlas. Chi-squared test was used for *P* value calculation.

### Clinical prognostic factors in ER-positive and ER-negative subtypes

Patients with ER-negative breast cancer subtype exhibited a significantly lower 5-year overall survival probability in both SEER (Log-rank test; $P < 0.001$) and TCGA (Log-rank test; $P = 0.018$) datasets (**Figure 1**). Taking tumor stages and age groups into consideration, the ER-negative patients exhibited lower or a tendency of lower 5-year survival rates than ER-positive patients in all stages and age groups in both datasets, except the ER-negative patients aged $\geq 70$ years in TCGA which showed a tendency of a higher 5-year survival rate (Figures S1 and S2; Table S1). However, the 10-year survival rate of ER-negative patients aged $\geq 70$ years in TCGA was considerably reduced compared with the ER-positive patients (Figure S2). In addition, the most significant survival difference between patients with these two cancer subtypes was found for stage III ($P < 0.001$ in both SEER and TCGA, Table S1). Not surprisingly, ER-negative patients exhibited a significantly lower 10-year survival probability in the younger group of patients less than 50 years old (SEER, $P < 0.001$; TCGA, $P = 0.049$, Figure S2). For patients with positive lymph node status, absence of ER is an effective indicator of poor prognosis (SEER, $P < 0.001$; TCGA, $P = 0.007$, Figure S3).

Univariate Cox proportional hazard analysis was conducted on all the clinical factors to explore their effects on overall survival (Table S2). ER-negative patients exhibited worse survival probabilities in both the SEER and TCGA datasets [SEER, $P < 0.001$, hazard ratio (HR) = 1.372, 95% confidence interval (CI) = 1.358–1.386; TCGA, $P = 0.021$, HR = 1.642, 95% CI = 1.078–2.501]. Multivariate Cox regression survival analysis adjusted for age, race, AJCC stage, lymph node status, tumor grade, and tumor size consistently exhibited a strong correlation of the ER-negative subtype with a poor survival probability in SEER dataset ($P < 0.001$, HR = 1.356, 95% CI = 1.337–1.376). The same phenomenon also occurred in TCGA dataset ($P = 0.002$, HR = 2.170, 95% CI = 1.330–3.541) after excluding other covariates (Table S3).

### Differential kinase–substrate features between two subtypes

The kinases included in this study comprised 470 genes annotated in the UniProtKB/Swiss-Prot database [30]. Experimentally validated substrates of these kinases from PhosphositePlus [31] were also incorporated. Kinase–substrate edge features were constructed based on the method described previously [24]. The kinase–substrate node features were transformed into kinase–substrate edge features according to the correlation of each kinase–substrate pair (see Method). We subsequently conducted feature selection of these node and edge features between ER-positive and ER-negative subtypes. By using 100 times of Monte Carlo cross validation, the selected kinase–substrate node and edge features were integrated with clinical characteristics for prognostic prediction (**Figure 2**). The clinical characteristics reported here included age, race, tumor stage, and lymph node status, all of which exhibited significant differences between ER-positive and ER-negative subtypes (Table 1).

By using the least absolute shrinkage and selection operator (LASSO) [32], a tatol of 46 differential kinase–substrate node and edge features between ER-positive and ER-negative subtypes were identified from the molecular dataset (**Figure 3**A). More than half of the selected features were upregulated in the ER-negative subtype, while *ESR1* and *MAPK3* were highly expressed in ER-positive subtype, which was consistent with the previous report [13]. Five-fold cross validation was performed during the process to tune the value of lambda in LASSO, and the performance was evaluated by the area under the curve (AUC), which was 0.908 (Figure S4, see Method). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was conducted on the selected node and edge features, and most of them were highly enriched in cell cycle and cancer-related pathways (Figure 3B). Moreover,

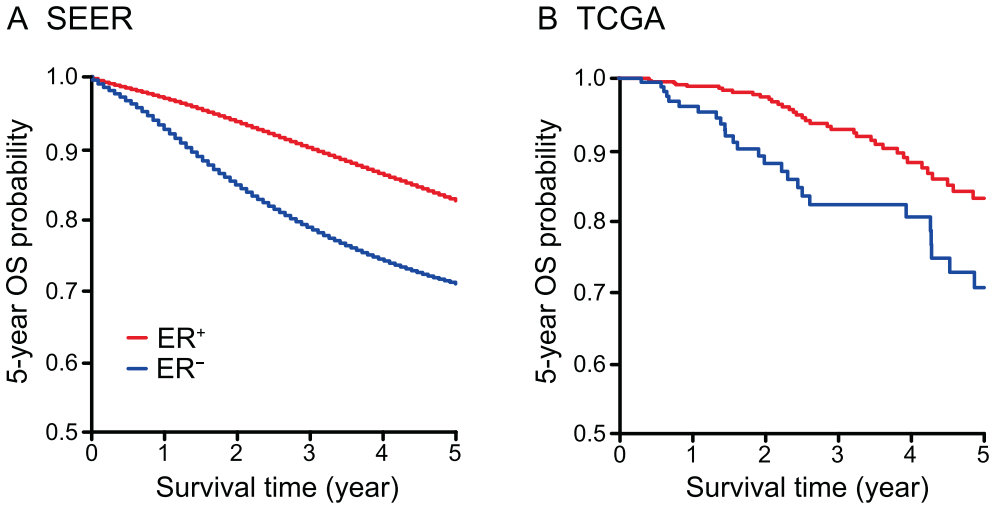



**Figure 1 OS probability of ER-positive and ER-negative patients**

**A.** The 5-year OS probability of ER-positive and ER-negative breast cancer patients in SEER database. Log-rank test, $P < 0.001$. **B.** The 5-year OS probability of ER-positive and ER-negative breast cancer patients in TCGA database. Log-rank test, $P = 0.02$. ER, estrogen receptor. OS, overall survival; SEER, the Surveillance, Epidemiology, and End Results; TCGA, The Cancer Genome Atlas.

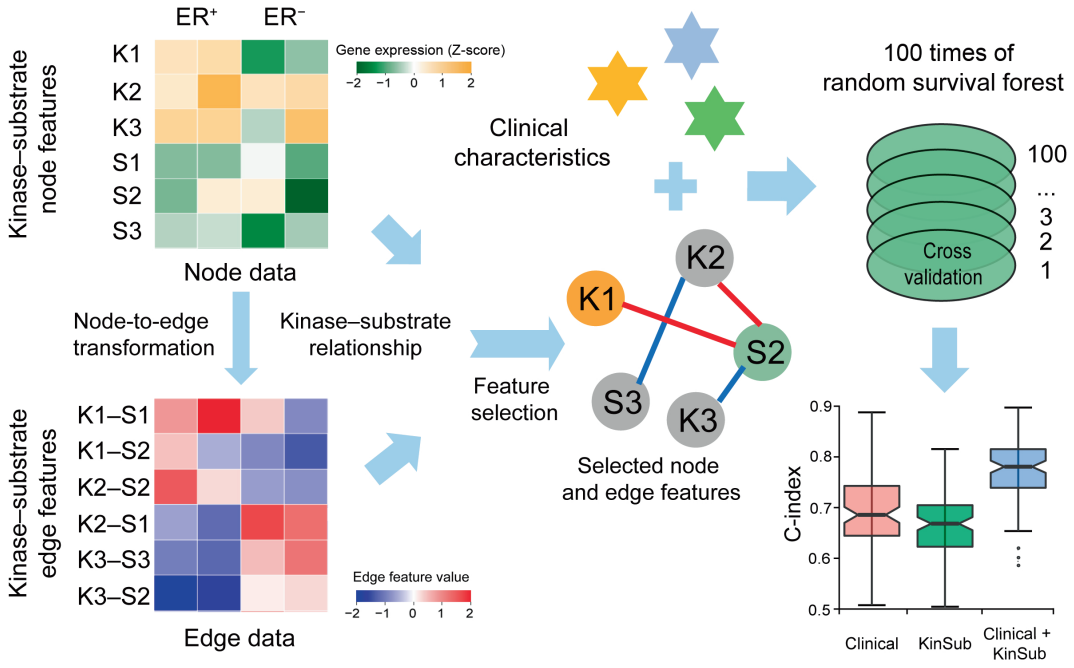



**Figure 2 Workflow of the kinase–substrate biomarker detection process**

Gene expression levels (Z-score transformed) are presented in a green-yellow color gradient with green and yellow for low and high expression, respectively. Features in the node data matrix are transformed into edge data based on correlation between each pair of kinase and substrate, generating the edge dataset in a blue-red color gradient with blue and red for low and high correlation, respectively. "K" and "S" depict kinase and substrate, respectively, and numbers 1–3 indicate different kinases and substrates, *e.g.*, "K1" means "Kinase1" and "S1" means "Substrate1". The yellow and green colored "K1" and "S2" indicate the differentially expressed kinase and substrate, respectively. The red and blue lines delineate the positive and negative correlations, respectively, between the pair of kinase and substrate. Node and edge features are subsequently selected using the LASSO regression algorithm, and 100 times of Monte Carlo cross validation are performed to identify the prognostic value of these selected features by integrating clinical characteristics for prognostic prediction. Clinical: clinical variables only; KinSub, kinase–substrate node and edge features only; Clinical + KinSub, clinical plus kinase–substrate node and edge features.

by analyzing the associated drugs of these features, we found that they could be highly enriched in existing drugs, such as glutathione and genistein (Figure 3C).

### Kinase–substrate node and edge features improve prognostic prediction

To assess whether these kinase and substrate features can provide additional prognostic power compared with clinical variables, we built predictive models by integrating clinical variables with expression values of the selected node and edge features. Concordance index (C-index) was used to measure the predictive power of node and edge features together with clinical variables, and a C-index greater than 0.5 indicates prediction accuracy other than random guess [33]. We applied 100 times of five-fold and two-fold cross validation for each model and calculated 100 C-indexes for each group of predictive variables (**Figure 4**, Figure S5). Notably, models integrating clinical variables with either node features ("Clinical + Node" model) or edge features ("Clinical + Edge" model) significantly increased the predictive accuracy compared with the model based exclusively on clinical variables ("Clinical" model) (C-index 0.744 *vs.* 0.683, $P = 3.46 \times 10^{-5}$; C-index 0.708 *vs.* 0.683, $P = 0.021$; Figure 4B). The final model integrating clinical variables with all the kinase–substrate node and edge features ("Clinical + KinSub" model) demonstrated the highest prediction power (C-index 0.781 *vs.* 0.683, $P = 5.89 \times 10^{-14}$; Figure 4B). Moreover, we retrieved the expression of a 50-gene qPCR assay (PAM50) gene set from our dataset and built a prognostic model by integrating clinical variables with "PAM50" ("Clinical + PAM50" model) using the same method. PAM50 gene signatures are widely used intrinsic subtype markers in breast cancer with independent prognostic values [34], but surprisingly, the "Clinical + PAM50" model performed no better than the "Clinical + KinSub" model in our analysis (C-index 0.743 *vs.* 0.781, $P$ value $= 2.73 \times 10^{-4}$; Figure S6), implying the better prognostic potential of kinase–substrate node and edge features in breast cancer.

### Kinase–substrate biomarkers exhibit subtype-specific prognostic power

To identify subtype-specific biomarkers for prognostic prediction, we performed univariate survival analysis for each of the 46 molecular features in both ER-positive and ER-negative subtypes (**Table 2**). Four node features *SAT1* ($P = 0.027$, HR = 0.539, 95% CI = 0.312–0.931), *GMPS* ($P = 0.011$, HR = 1.895, 95% CI 1.158–3.101), *PHKG2* ($P = 0.005$, HR = 0.491, 95% CI = 0.297–0.811), *CCNE1* ($P = 0.016$, HR = 1.833, 95% CI = 1.122–2.995) and one edge feature

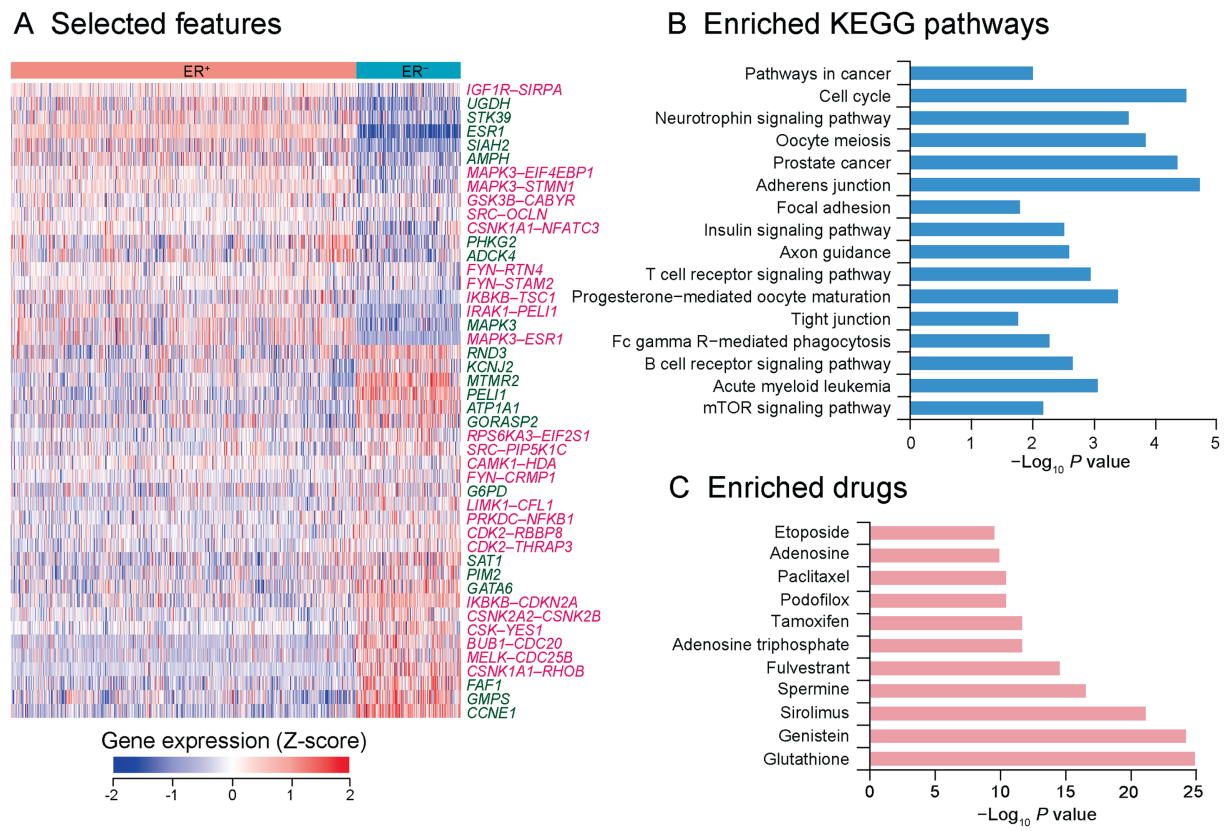



**Figure 3 Features selected from LASSO regression**

**A.** Heatmap demonstrating the Z-score transformed expression levels of the 21 differential node features and 25 differential edge features between ER-positive and ER-negative subtypes. Blue and red represent low and high expression, respectively. Green and pink texts indicate node and edge features, respectively. **B.** KEGG-enriched pathways of the 46 selected node and edge features. **C.** Drugs enriched for the 46 selected node and edge features. KEGG, Kyoto Encyclopedia of Genes and Genomes.

*BUB1–CDC20* ($P$ = 0.013, HR = 1.867, 95% CI = 1.139–3.063) demonstrated significantly prognostic power in ER-positive subtype. Two edge features *CSNK1A1–NFATC3* ($P$ = 0.043, HR = 2.048, 95% CI = 1.021–4.107) and *SRC–OCLN* ($P$ = 0.048, HR = 2.04, 95% CI 1.006–4.134) showed significantly prognostic power in the ER-negative group. To exclude the influence of clinical covariates, a multivariate Cox model was constructed and the prognostic values of these candidate biomarkers were validated (Table S4).

To further assess the relationship of these candidate biomarkers with clinical outcome of ER-positive and ER-negative patients, Kaplan–Meier curves were constructed using Log-rank test to stratify the patients into high- and low-risk groups according to the expression levels of node features or correlation values of edge features (median split) (**Figure 5**, Figure S7). Poor survival was observed in the high-risk groups of ER-negative patients stratified by *CSNK1A1–NFATC3* (Figure 5C) and *SRC–OCLN* (Figure 5F). Moreover, Kaplan–Meier curves were also plotted based on the expression of *CSNK1A1*, *NFATC3*, *SRC*, and *OCLN* by median split (Figure 5A, B, D, and E), but the expression of these kinases and substrates did not demonstrate prognostic values in ER-negative patients, which supports the significant power of the correlations between kinases and substrates in clinical practice. Independent datasets from Gene Expression Omnibus (GEO), including GSE42568 (HG-U133A Plus2 platform) [35], GSE22055 (HG-U133A platform) [36], and ten other datasets with available survival information and ER statuses, were used to determine whether the identified node and edge biomarkers could provide prognostic information for ER-positive and ER-negative patients. We observed that these potential biomarkers could also stratify the survival of high- and low-risk groups in these independent datasets, suggesting that the prognostic power of these biomarkers is stable and reliable in practice (Figure 5G–J, Figure S7; Table S5).

We next compared the kinase–substrate biomarkers with existing breast cancer prognostic biomarkers from previous studies [12,37]. National Research Council (NRC) gene signatures NRC-1 (33 genes), NRC-2 (46 genes), and NRC-3 (47 genes) were reported to predict disease-free survival of ER-positive patients with high accuracy. NRC-7, NRC-8, and NRC-9 gene sets (39, 25, and 20 genes, respectively) were prognostic signatures for ER-negative subtype. We compared the performance of our newly identified biomarkers with the previously reported signatures by C-indexes based on 100 times of cross validation in the two subtypes, respectively (see Method). For the ER-positive biomarkers, *CCNE1* and the combination of the five biomarkers demonstrated higher prognostic effects than both NRC-1 and NRC-2 (Figure S8A). For the ER-negative biomarkers, *SRC–OCLN*, *CSNK1A1–NFATC3*, and their combination significantly

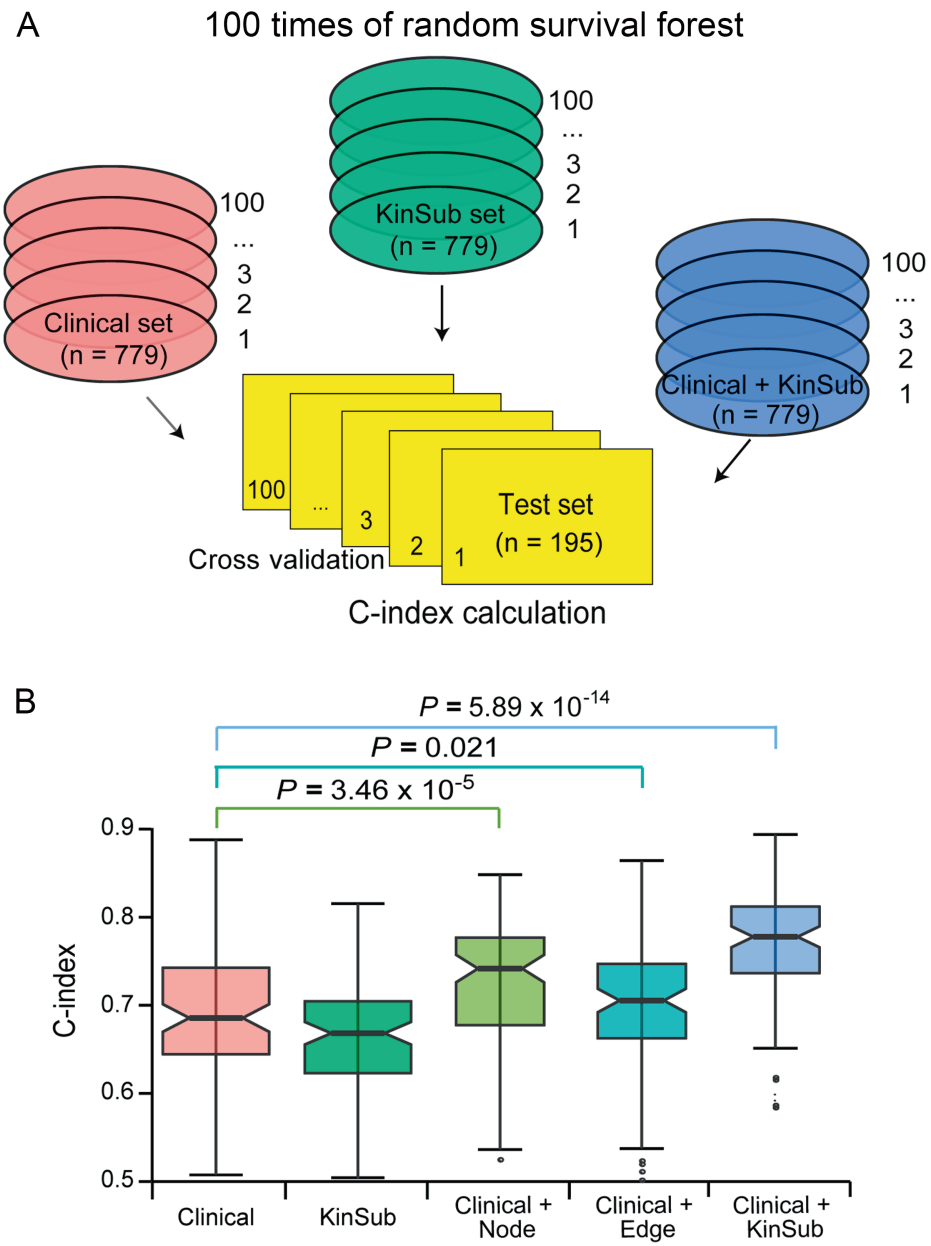



**Figure 4 Random survival forest models trained from clinical variables and kinase–substrate node and edge features**
**A.** 100 times of five-fold cross validation in the clinical, KinSub, and their combined datasets for C-index calculation. **B.** Comparison of C-indexes of 100 times of cross validation results for Clinical, KinSub, Clinical plus KinSub node features (Clinical + Node), Cnical plus KinSub edge features (Clinical + Edge), and Clinical + KinSub. n, the number of samples. Two-sided Wilcoxon rank-sum test was used for significance test.

outperformed NRC-7, NRC-8, NRC-9 and their combination, as well as "S6 kinase" markers, which include genes encoding kinases in the S6 kinase signaling pathway, such as *RPS6KA3*, *SMG-1*, and *RPS6KA1* [12] (Figure S8B). These results suggested that the novel kinase–substrate biomarkers identified in this study had better performance than existing biomarkers, which were gene sets comprising dozens of genes.

To demonstrate the additional prognostic power of clinical characteristics from the previous analysis, we also explored the combinations of these node and edge biomarkers with clinical factors. The C-index and *P* value from the Cox regression model were presented for all possible combinations. We found that the inclusion of age group, AJCC stage, and lymph node status in ER-positive or ER-negative subtype could significantly improve the prognostic power (Table S6). Kaplan–Meier survival curves also demonstrated that combination of clinical factors with these biomarkers achieved increased prognostic power (Figure 5C and F, Figures S7A and S9).

To evaluate the potential for specific targeted drug development for these node and edge biomarkers, we compared the expression of the selected kinases and substrates in ER-positive and ER-negative patients with that in normal individuals in the TCGA database. The expression of most of the selected kinases and substrates was increased in the breast cancer compared with that in normal samples (Figure S10). Specifically, the expression level of SRC was also higher in

**Table 2 Univariate survival analysis of 46 kinase–substrate node and edge biomarkers in breast cancer based on ER statuses**

| Gene | $ER^+$ | | $ER^-$ | |
|---|---|---|---|---|
| | HR (95% CI) | *P* value | HR (95% CI) | *P* value |
| *ATP1A1* | 1.016 (0.621−1.663) | 0.950 | 2.046 (0.909−4.605) | 0.084 |
| *ESR1* | 1.663 (0.944−2.93) | 0.078 | 0.596 (0.141−2.528) | 0.483 |
| *SAT1* | 0.539 (0.312−0.931) | 0.027 | 1.086 (0.524−2.251) | 0.825 |
| *UGDH* | 1.13 (0.683−1.871) | 0.634 | 0.751 (0.227−2.487) | 0.640 |
| *GORASP2* | 1.428 (0.872−2.339) | 0.157 | 0.732 (0.364−1.469) | 0.380 |
| *SIAH2* | 0.781 (0.475−1.285) | 0.331 | 0.987 (0.344−2.835) | 0.981 |
| *MAPK3* | 1.022 (0.62−1.685) | 0.931 | 1.604 (0.769−3.348) | 0.208 |
| *GMPS* | 1.895 (1.158−3.101) | 0.011 | 1.692 (0.515−5.564) | 0.386 |
| *G6PD* | 1.07 (0.653−1.754) | 0.789 | 1.294 (0.62−2.702) | 0.492 |
| *RND3* | 0.775 (0.473−1.271) | 0.313 | 1.066 (0.478−2.378) | 0.876 |
| *FAF1* | 0.79 (0.473−1.319) | 0.368 | 2.045 (0.779−5.367) | 0.146 |
| *STK39* | 0.732 (0.446−1.202) | 0.218 | 0.906 (0.372−2.209) | 0.828 |
| *MTMR2* | 1.574 (0.963−2.572) | 0.070 | 1.026 (0.394−2.671) | 0.959 |
| *PHKG2* | 0.491 (0.297−0.811) | 0.005 | 0.74 (0.35−1.566) | 0.432 |
| *PELI1* | 0.689 (0.416−1.143) | 0.149 | 0.904 (0.37−2.208) | 0.824 |
| *ADCK4* | 0.804 (0.491−1.319) | 0.388 | 0.822 (0.405−1.668) | 0.587 |
| *PIM2* | 0.667 (0.398−1.119) | 0.125 | 0.733 (0.358−1.502) | 0.396 |
| *AMPH* | 0.728 (0.445−1.191) | 0.206 | 0.748 (0.259−2.161) | 0.592 |
| *CCNE1* | 1.833 (1.122−2.995) | 0.016 | 4.444 (0.602−32.809) | 0.144 |
| *GATA6* | 0.734 (0.444−1.216) | 0.230 | 1.968 (0.806−4.805) | 0.137 |
| *KCNJ2* | 1.053 (0.64−1.733) | 0.838 | 0.641 (0.318−1.289) | 0.212 |
| *BUB1−CDC20* | 1.867 (1.139−3.063) | 0.013 | 0.708 (0.289−1.739) | 0.452 |
| *CAMK1−HDAC9* | 0.973 (0.595−1.591) | 0.914 | 1.189 (0.592−2.387) | 0.627 |
| *CDK2−RBBP8* | 0.806 (0.491−1.322) | 0.392 | 0.962 (0.444−2.088) | 0.923 |
| *CDK2−THRAP3* | 1.239 (0.752−2.04) | 0.401 | 1.087 (0.523−2.26) | 0.823 |
| *CSK−YES1* | 0.64 (0.382−1.072) | 0.090 | 0.994 (0.469−2.104) | 0.987 |
| *CSNK1A1−NFATC3* | 0.964 (0.586−1.585) | 0.884 | 2.048 (1.021−4.107) | 0.043 |
| *CSNK1A1−RHOB* | 0.902 (0.541−1.505) | 0.693 | 0.486 (0.228−1.035) | 0.061 |
| *CSNK2A2−CSNK2B* | 1.214 (0.739−1.997) | 0.444 | 1.078 (0.482−2.408) | 0.855 |
| *FYN−CRMP1* | 1.101 (0.674−1.799) | 0.700 | 0.879 (0.439−1.759) | 0.715 |
| *FYN−RTN4* | 0.619 (0.374−1.023) | 0.061 | 1.228 (0.6−2.514) | 0.574 |
| *FYN−STAM2* | 1.004 (0.615−1.641) | 0.986 | 1.247 (0.609−2.554) | 0.546 |
| *GSK3B−CABYR* | 1.208 (0.739−1.974) | 0.451 | 1.05 (0.514−2.144) | 0.894 |
| *IGF1R−SIRPA* | 0.656 (0.401−1.075) | 0.094 | 0.936 (0.441−1.988) | 0.864 |
| *IKBKB−CDKN2A* | 1.412 (0.845−2.362) | 0.188 | 2.484 (0.865−7.133) | 0.091 |
| *IKBKB−TSC1* | 0.684 (0.418−1.118) | 0.130 | 0.832 (0.358−1.931) | 0.668 |
| *IRAK1−PELI1* | 1.45 (0.846−2.484) | 0.177 | 0.378 (0.114−1.249) | 0.111 |
| *LIMK1−CFL1* | 0.848 (0.518−1.389) | 0.513 | 1.562 (0.736−3.315) | 0.246 |
| *MAPK3−EIF4EBP1* | 0.658 (0.403−1.075) | 0.095 | 1.016 (0.488−2.117) | 0.966 |
| *MAPK3−ESR1* | 1.03 (0.62−1.712) | 0.909 | 0.71 (0.215−2.343) | 0.574 |
| *MAPK3−STMN1* | 1.008 (0.616−1.652) | 0.973 | 0.97 (0.445−2.116) | 0.939 |
| *MELK−CDC25B* | 0.976 (0.596−1.598) | 0.924 | 0.757 (0.358−1.603) | 0.468 |
| *PRKDC−NFKB1* | 1.036 (0.634−1.694) | 0.888 | 0.863 (0.421−1.768) | 0.687 |
| *RPS6KA3−EIF2S1* | 1.13 (0.69−1.85) | 0.628 | 0.84 (0.413−1.709) | 0.631 |
| *SRC−OCLN* | 1.424 (0.86−2.358) | 0.170 | 2.04 (1.006−4.134) | 0.048 |
| *SRC−PIP5K1C* | 1.331 (0.811−2.184) | 0.257 | 0.883 (0.421−1.849) | 0.741 |

*Note*: CI, confidence interval; HR, hazard ratio. Wald test was used for *P* value calculation.

the ER-negative group than that in the ER-positive subtype (Figure S10). Considering the poor survival of high-risk groups of ER-negative patients stratified by *SRC–OCLN*, the edge biomarkers could serve as potential drug targets; however, further studies are needed.

## Discussion

This study analyzed 706,763 first primary invasive breast cancer patients with available ER status from two independent databases (SEER and TCGA). This study was the first to integrate clinical factors with kinase–substrate node and edge biomarkers for prognostic prediction between ER-positive and ER-negative breast cancer subtypes in large datasets. In addition, we identified prominent kinase–substrate node and edge biomarkers in both subtypes, and these signatures might be potential biomarkers to distinguish ER-positive and ER-negative breast cancer.

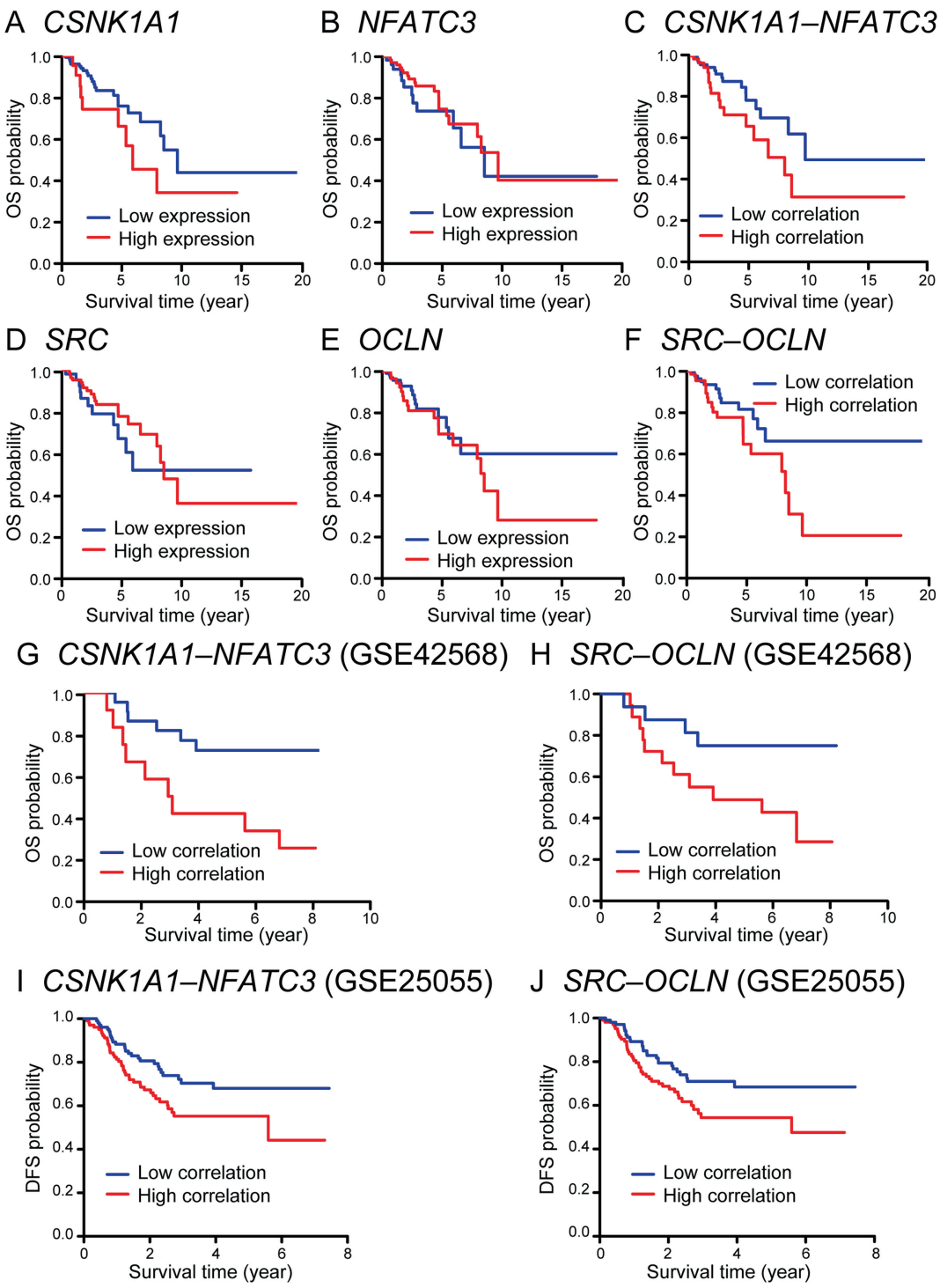
The data presented here confirmed that the ER-negative subtype exhibited poorer survival with an increased occurrence rate among younger ages and patients of African American compared with ER-positive patients. This finding was consistent with the phenomenon observed in the California dataset [38–40]. In addition, we found that features including later disease stages and larger tumor sizes were associated with the ER-negative subtype. These observations were reasonable given the lower survival rate of ER-negative patients. Interestingly, in both datasets, patients in the 50–69-year-old group exhibit the largest proportion of ER-negative breast cancer occurrence (Table 1), and the survival probability of this group was similar in SEER or even higher in TCGA compared with patients less than 50 years old (Figure S2). This finding might be related to hormone levels in the body. Carey et al. [38] reported the increased prevalence of the more aggressive subtype in premenopausal women compared with postmenopausal women in the Carolina breast cancer cohort. Further studies are needed to better characterize the influence of hormone level on the occurrence of breast cancer subtypes.

We identified several key distinct kinase–substrate node and edge features between ER-positive and ER-negative subtypes. By analyzing the associated drugs of these features, we found that they could be highly enriched in existing drugs, such as glutathione and genistein. Glutathione is involved in numerous biological processes, including but not limited to cell development, differentiation, antioxidation, and immune response modulation. Therefore, disorders in glutathione metabolism could lead to the development and progression of numerous human diseases, including cancer [41–43]. In fact, as a drug the glutathione has been used in the treatment of lung cancer and liver cancer [44–48], implying that it could potentially be used in the treatment of ER-negative invasive breast cancer.

Integrated analysis of molecular biomarkers with clinical prognostic factors in breast cancer patients demonstrated the utility of the inclusion of kinase–substrate node and edge biomarkers for prognostic prediction. Neither of the clinical plus nodes (“Clinical + Node” model) and clinical plus edges (“Clinical + Edge” model) exhibited increased prognostic power compared with the integrated prediction model (“Clinical + KinSub” model), suggesting that both the expression of kinases or substrates (nodes) and the correlation between kinases and their targeted substrates (edges) play important roles in the regulation of networks in our body. Moreover, the “Clinical + KinSub” model achieved better performance than the model built by the widely used “PAM50” gene set. The result confirmed the importance of kinase–substrate network in our body. In addition, the kinase–substrate node and edge biomarkers identified in both ER-positive and ER-negative subtypes outperformed the existing markers of breast cancer in prognosis. Considering the feasibility and convenience of kinases to be drug targets, our analysis provides a prominent kinase–substrate set for the drug development, which will give more available intervention on patients with ER-negative breast cancer in the future.

The improved prognostic value of the kinase–substrate edge biomarkers validated the utility of our method for identifying the correlation between genes instead of DEGs as functional drivers. These edge markers are generally missed by traditional methods [24,49]. The disruption of the correlations between kinases and substrates can potentially improve the clinical outcome of breast cancer patients, which enlarges the scope of drug target development. The results from 12 independent GEO datasets also confirmed the effectiveness of these biomarkers (Figure 5, Figure S7; Table S5). Besides, in the two different subtypes, node features contributed more to the prognostic probability of ER-positive patients, whereas ER-negative subtype mostly relied on the edge biomarkers (Table 2). This phenomenon may underpin different regulation networks in different types of diseases. Previous studies also reported different results. Specifically, edge biomarkers were considered to be more reliable for subtyping in one study [24], whereas Speers et al. [12] demonstrated that kinases alone were effective in ER-negative breast cancer subtyping. Given the different cohorts and analyzing methods used in these studies, a more comprehensive dataset with broader networks in addition to kinase–substrate networks is needed in the future to testify the usefulness of these two types of biomarkers.

Particularly, two kinase–substrate pairs, *CSNK1A1–NFATC3* and *SRC–OCLN*, demonstrated strong correlations with clinical outcome in the ER-negative subtype. *NFATC3* is one member of the NFAT (nuclear factor of activated T cells) transcription factor gene family, which play important roles in T cell activation [50]. The activation of calcineurin-NFAT pathway was observed in triple-negative breast cancer and substantially contributed to the tumorigenesis and metastasis of mammary tumor cell lines [51–53]. In our study, the high correlation between *CSNK1A1* and *NFATC3* is associated with poor survival in the ER-negative group, which suggests that pharmacological inhibition of *NFATC3* by targeting *CSNK1A1* could be of therapeutic interest for breast cancer patients. The tight junction structure is one of the inevitable barrier for cancer cells to enable metastasis, whereas OCLN (Occludin) is one of the early identified tight junction proteins [54–56]. Slight association between reduced *OCLN* expression and poor overall survival was observed in a cohort with 10-year follow-up of breast cancer patients, and studies conducted on human cell lines demonstrated that OCLN phosphorylation by SRC attenuates its assembly at the tight junctions [56–59]. Given that the high correlation group between *SRC* and *OCLN* had worse survival performance in ER-negative subtype, the association between *SRC* and *OCLN* represents a potential “driver” of cell proliferation in ER-negative breast cancer.


A CSNK1A1
B NFATC3
C CSNK1A1–NFATC3
D SRC
E OCLN
F SRC–OCLN
G CSNK1A1–NFATC3 (GSE42568)
H SRC–OCLN (GSE42568)
I CSNK1A1–NFATC3 (GSE25055)
J SRC–OCLN (GSE25055)
OS probability
DFS probability
Survival time (year)
Low expression
High expression
Low correlation
High correlation

In conclusion, our study depicts a model for the identification of promising molecular biomarkers with utility in clinical prognosis. This population-based research suggests distinct clinicopathological characteristics between ER-positive and ER-negative breast cancer patients. Prognostic clinical factors and kinase–substrate node and edge features were identified based on the comparison of these two subtypes. Compared with using the clinical variables only ("Clinical" model), incorporating kinase–substrate node and edge features greatly improves the predictive accuracy, indicating the advantages of kinases and substrates as well as their regulation in clinical diagnosis. Furthermore, our analyses also provide promising kinase–substrate node and edge biomarkers for clinically relevant refinement of prognostic assessment in the ER-positive and ER-negative subtypes, and these biomarkers also serve as candidate drug targets for the treatment of breast invasive cancer in the future. In addition, this work can be applied to the analyses of network biomarkers [24,49,60–63] and dynamic network biomarkers [25,64–66] for disease diagnosis and disease prediction, respectively.

## Method

### Clinical database

The SEER 1973 to 2012 database (http://seer.cancer.gov/about/overview.html) represented approximately 28% of the US population. We analyzed breast cancer survival in all women diagnosed with first primary breast cancer with ER status (available from 1990). Characteristics, including age at diagnosis, race, AJCC stage, lymph node status, tumor grade, and tumor size, were examined for each patient. Survival information included vital status, cause of death, and survival time. All characteristics studied and information regarding ER status were based on standard coding rules of SEER records. After excluding patients without survival information, the final study cohort of SEER was reduced to 705,729 from an initial dataset of 705,740.

Clinical data of breast invasive carcinoma from TCGA were used for this study (http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/BRCA/20160128). The clinicopathological information for each patient included age at diagnosis, race, AJCC stage, and lymph node status. ER status was determined according to the current clinical guideline jointly issued by the American Society of Clinical Oncology (ASCO) and the College of American Pathology [67]. After excluding male patients and cases lacking information on ER status, the final study cohort was reduced to 1034 from an initial dataset of 1097.

### Gene expression dataset

Analysis of kinase–substrate features was performed on gene expression data (RNAseqV2) of TCGA Breast Invasive Carcinoma (BRCA) (http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/BRCA/20160128). Upper quartile normalized RNA-seq by Expectation-Maximization (RSEM) data were $\log_2$ transformed, and the data were then Z-score centred on the gene level [68].

Gene expression data were matched with clinical data using a TCGA barcode for each patient, excluding cases lacking clinical records or expression information, which resulted in a cohort of 1017 breast invasive cancer patients. In total, 470 of 521 known human kinases and 552 experimentally validated substrates of these kinases were identified in the expression dataset and characterized as node features [30].

All the GEO datasets were obtained from the GEO website (GEO: GSE42568, GSE22055, GSE10893, GSE2034, GSE21653, GSE22133, GSE22219, GSE48408, GSE4922, GSE53031, GSE6532, and GSE7390), and are publicly accessible at https://www.ncbi.nlm.nih.gov/geo.

### Kinase–substrate edge construction

Kinase–substrate node features were transformed into kinase–substrate edge features based on the correlation of each kinase–substrate pair, performed according to the method previously described [24]. The transformation is described below.

$$\begin{matrix} kinase, u \\ substrate, v \end{matrix} \begin{pmatrix} x_{u,j,k} \\ x_{v,j,k} \end{pmatrix} -> edge< u - v >_k \left( \frac{x_{u,j,k} - \mu_{u,k}}{\sigma_{u,k}} . \frac{x_{v,j,k} - \mu_{v,k}}{\sigma_{v,k}} \right)$$

where $x_{u,j,k}$ represents the original value of $u^{th}$ kinase in $j^{th}$ sample from $k^{th}$ class, $x_{v,j,k}$ represents the original value of $v^{th}$ substrate in $j^{th}$ sample from $k^{th}$ class, and k was set to 1 or 2 to represent the ER-positive or ER-negative subtype. In addition, $\mu_{u,k} = \frac{1}{n_k}\sum_{j=1}^{n_k}(x_{u,j,k} - \mu_{u,k})$ and $\mu_{v,k} = \frac{1}{n_k}\sum_{j=1}^{n_k}(x_{v,j,k} - \mu_{v,k})$ are sample means of kinase u and substrate v, and $\sigma_{u,k} = \sqrt{\frac{1}{n_k}\sum_{j=1}^{n_k}(x_{u,j,k} - \mu_{u,k})^2}$ and $\sigma_{v,k} = \sqrt{\frac{1}{n_k}\sum_{j=1}^{n_k}(x_{v,j,k} - \mu_{v,k})^2}$ are the corresponding uncorrected sample standard deviation. After edge transformation, the final expression dataset consisted of 1022 kinase–substrate node features and 2606 edge features.

**Figure 5 Prognostic values of two edge biomarkers in ER-negative patients**

**A.** Kaplan–Meier curves of high and low expression groups stratified by *CSNK1A1* for ER-negative breast cancer patients in TCGA. Log-rank test, $P = 0.111$. **B.** Kaplan–Meier curves of high and low expression groups stratified by *NFATC3* for ER-negative breast cancer patients in TCGA. Log-rank test, $P = 0.418$. **C.** Kaplan–Meier curves of high and low correlation groups stratified by *CSNK1A1–NFATC3* for ER-negative breast cancer patients in TCGA. Log-rank test, $P = 0.039$. **D.** Kaplan–Meier curves of high and low expression groups stratified by *SRC* for ER-negative breast cancer patients in TCGA. Log-rank test, $P = 0.476$. **E.** Kaplan–Meier curves of high and low expression groups stratified by *OCLN* for ER-negative breast cancer patients in TCGA. Log-rank test, $P = 0.378$. **F.** Kaplan–Meier curves of high and low correlation groups stratified by *SRC−OCLN* for ER-negative breast cancer patients in TCGA. Log-rank test, $P = 0.043$. **G.** Kaplan–Meier curves of high and low correlation groups stratified by *CSNK1A1–NFATC3* in the GSE42568 dataset. Log-rank test, $P = 0.01$. **H.** Kaplan–Meier curves of high and low correlation groups stratified by *SRC–OCLN* in the GSE42568 dataset. Log-rank test, $P = 0.036$. **I.** Kaplan–Meier curves of high and low correlation groups stratified by *CSNK1A1–NFATC3* in the GSE25055 dataset. Log-rank test, $P = 0.022$. **J.** Kaplan–Meier curves of high and low correlation groups stratified by *SRC–OCLN* in the GSE25055 dataset. Log-rank test, $P = 0.033$. DFS, disease-free survival.

### Feature selection and performance comparison

We first selected the node and edge features using the Student's *t*-test with a *P* value cut-off of 0.05 between ER-positive and ER-negative subtypes, which reduced the dataset to 2275 features. We used these expression data as the explanatory variables and two subtypes as the response variables to build a binary classifier (family = "binomial") for feature selection by five-fold cross validation. For each of the five iterations, 80% of the data were used for training by LASSO using the R package "glmnet" [69], and the prediction was conducted on the remaining 20% of the data. The prediction results from the five-fold cross validation were combined, and the AUC was calculated by the R package "ROCR". Ultimately, we trained the entire dataset with LASSO, and all the 46 features with nonzero coefficients were retained for subsequent analysis.

Clinical data, including ER status, age, AJCC stage, and lymph node status, were combined with molecular data that included the 46 selected features for model training by random survival forest (RSF) using the R package "randomForestSRC" [70]. For each dataset, we used two criteria to randomly split the samples into two parts. One method used 80% as the training set and the remaining 20% as test set. The other method divided the entire set in half: 50% served as the training set and 50% served as the test set. The model built based on the training set was then applied to the test set for prediction, and the C-index was calculated using the R package "survcomp". For each dataset, the procedure was iterated 100 times; thus, 100 C-indexes were obtained. Wilcoxon rank-sum test was then used to compare the results from different datasets. Furthermore, each of the 46 node and edge features was evaluated for prognostic values in ER-positive and ER-negative subtypes using univariate Cox regression. For features demonstrating significant prognostic power, multivariate survival analysis was also conducted to exclude the influence of covariates.

For the performance comparison of the identified kinase–substrate biomarkers in ER-positive and ER-negative subtypes with existing biomarkers, we built RSF models for each biomarker and their combinations by five-fold cross validation. By iterating 100 times of the procedure, we compared the C-indexes between kinase–substrate biomarkers and existing biomarkers by Wilcoxon rank-sum test.

### Functional analysis

We performed KEGG pathway and drug association enrichment analysis for the 46 kinase–substrate features using WebGestalt (http://www.webgestalt.org/). Hypergeometric test was used for enrichment evaluation analysis, and the Benjamini & Hochberg method was used for *P* value adjustment. *P* values less than 0.05 were considered significant.

### Statistical analysis

R version 3.2.2 (http://www.R-project.org/) was used to perform all the statistical analyses in this work. The relationships of ER-positive and ER-negative groups with clinicopathological characteristics were analyzed using the Chi-square ($\chi 2$) test. Survival curves were generated using the Kaplan–Meier method, and the Log-rank test was applied to calculate differences between the curves. HRs and their 95% CI were estimated for each multivariate and univariate survival analyses using Cox proportional hazards models. All tests conducted were two-sided, and significant differences were noted by *P* values less than 0.05.

## CRediT author statement

**Yidi Sun:** Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing - original draft. **Chen Li:** Conceptualization, Writing - review & editing. **Shichao Pang:** Writing - review & editing. **Qianlan Yao:** Writing - review & editing. **Luonan Chen:** Supervision, Writing - review & editing. **Yixue Li:** Supervision, Writing - review & editing. **Rong Zeng:** Conceptualization, Supervision, Writing - review & editing. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing financial interests.

## Acknowledgments

We thank Yuan Lu (Fudan University, China) and Jie Ping (Vanderbilt University Medical Center, USA) for critical reading of the manuscript. This work was supported by the National Key R&D Program of China (Grant No. 2017YFA0505500), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA12010000), the National Program on Key Basic Research Project of China (Grant Nos. 2014CBA02000 and 2014CB910500), and the National Natural Science Foundation of China (Grant Nos. 91029301, 30700397, 91529303, and 31771476). The authors gratefully acknowledge the support of the SANOFI-SIBS Distinguish Young Scientist Award Scholarship Program.

## Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2019.11.012.

## ORCID

0000-0002-4191-2917 (Yidi Sun)
0000-0002-6367-0998 (Chen Li)
0000-0002-4111-2864 (Shichao Pang)
0000-0003-0737-4147 (Qianlan Yao)
0000-0002-3960-0068 (Luonan Chen)
0000-0002-1198-7176 (Yixue Li)
0000-0003-0685-4333 (Rong Zeng)

## References

[1] Maajani K, Jalali A, Alipour S, Khodadost M, Tohidinik HR, Yazdani K. The global and regional survival rate of women with breast cancer: a systematic review and meta-analysis. Clin Breast Cancer 2019;19:165–77.

[2] American Cancer Society. Breast cancer facts & figures 2013–2014. Atlanta: American Cancer Society, Inc.; 2013.
[3] Swaby RF, Sharma CGN, Jordan VC. SERMs for the treatment and prevention of breast cancer. Rev Endocr Metab Disord 2007;8:229–39.
[4] Formisano L, Lu Y, Servetto A, Hanker AB, Jansen VM, Bauer JA, et al. Aberrant FGFR signaling mediates resistance to CDK4/6 inhibitors in ER + breast cancer. Nat Commun 2019;10:1373.
[5] Arumugam A, Subramani R, Nandy SB, Terreros D, Dwivedi AK, Saltzstein E, et al. Silencing growth hormone receptor inhibits estrogen receptor negative breast cancer through ATP-binding cassette sub-family G member 2. Exp Mol Med 2019;51:1–13.
[6] Kim S, Ko Y, Lee HJ, Lim JE. Menopausal hormone therapy and the risk of breast cancer by histological type and race: a meta-analysis of randomized controlled trials and cohort studies. Breast Cancer Res Treat 2018;170:667–75.
[7] Gu G, Dustin D, Fuqua SA. Targeted therapy for breast cancer and molecular mechanisms of resistance to treatment. Curr Opin Pharmacol 2016;31:97–103.
[8] Khan F, Esnakula A, Ricks Santi LJ, Zafar R, Kanaan Y, Naab T. Loss of PTEN in high grade advanced stage triple negative breast ductal cancers in African American women. Pathol Res Pract 2018;214:673–8.
[9] Hahnen E, Hauke J, Engel C, Neidhardt G, Rhiem K, Schmutzler RK. Germline mutations in triple-negative breast cancer. Breast Care (Basel) 2017;12:15–9.
[10] Van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
[11] Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98:10869–74.
[12] Speers C, Tsimelzon A, Sexton K, Herrick AM, Gutierrez C, Culhane A, et al. Identification of novel kinase targets for the treatment of estrogen receptor-negative breast cancer. Clin Cancer Res 2009;15:6327–40.
[13] Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumors. Nature 2012;490:61–70.
[14] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) 2015;19:A68–77.
[15] Yuan Y, Allen EMV, Omberg L, Wagle N, Amin Mansour A, Sokolov A, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotechnol 2014;32:644–52.
[16] Shen S, Wang Y, Wang C, Wu YN, Xing Y. SURVIV for survival analysis of mRNA isoform variation. Nat Commun 2016;7:11548.
[17] Zhang J, Yang PL, Gray NS. Targeting cancer with small molecule kinase inhibitors. Nat Rev Cancer 2009;9:28–39.
[18] Manning G, Whyte D, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298:1912–34.
[19] Manley PW, Cowan Jacob SW, Mestan J. Advances in the structural biology, design and clinical development of VEGF-R kinase inhibitors for the treatment of angiogenesis. Biochim Biophys Acta 2004;1697:17–27.
[20] Chen Y, Choong LY, Lin Q, Philp R, Wong CH, Ang BK, et al. Differential expression of novel tyrosine kinase substrates during breast cancer development. Mol Cell Proteomics 2007;6:2072–87.
[21] Weitsman G, Lawler K, Kelleher MT, Barrett JE, Barber PR, Shamil E, et al. Imaging tumor heterogeneity of the consequences of a PKCα-substrate interaction in breast cancer patients. Biochem Soc Trans 2014;42:1498–505.
[22] Hochgräfe F, Zhang L, O'Toole SA, Browne BC, Pinese M, Cubas AP, et al. Tyrosine phosphorylation profiling reveals the signaling network characteristics of basal breast cancer cells. Cancer Res 2010;70:9391–401.
[23] Li QR, Wang ZM, Wewer Albrechtsen NJ, Wang DD, Su ZD, Gao XF, et al. Systems signatures reveal unique remission-path of type 2 diabetes following Roux-en-Y gastric bypass surgery. EBioMedicine 2018;28:234–40.
[24] Zhang W, Zeng T, Liu X, Chen L. Diagnosing phenotypes of single-sample individuals by edge biomarkers. J Mol Cell Biol 2015;7:231–41.
[25] Li M, Li C, Liu WX, Liu C, Cui J, Li Q, et al. Dysfunction of PLA2G6 and CYP2C44-associated network signals imminent carcinogenesis from chronic inflammation to hepatocellular carcinoma. J Mol Cell Biol 2017;9:489–503.
[26] Auffray C, Charron D, Hood L. Predictive, preventive, personalized and participatory medicine: back to the future. Genome Med 2010;2:57.
[27] Haider S, Yao C, Sabine V, Grzadkowski M, Stimper V, Starmans MH, et al. Pathway-based subnetworks enable cross-disease biomarker discovery. Nat Commun 2018;9:4746.
[28] Jin N, Wu H, Miao Z, Huang Y, Hu Y, Bi X, et al. Network-based survival-associated module biomarker and its crosstalk with cell death genes in ovarian cancer. Sci Rep 2015;5:11566.
[29] De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nat Rev Microbiol 2010;8:717–29.
[30] Braconi Quintaje S, Orchard S. The annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot: one small step in manual annotation, one giant leap for full comprehension of genomes. Mol Cell Proteomics 2008;7:1409–19.
[31] Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Murray B, Latham V, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res 2012;40:261–70.
[32] Tibshirani R. Regression shrinkage and selection via the Lasso. J Royal Stat Soc 1996;58:267–88.
[33] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
[34] Liu MC, Pitcher BN, Mardis ER, Davies SR, Friedman PN, Snider JE, et al. PAM50 gene signatures and breast cancer prognosis with adjuvant anthracycline- and taxane-based chemotherapy: correlative analysis of C9741 (Alliance). NPJ Breast Cancer 2016;2:15023–33.
[35] Clarke C, Madden SF, Doolan P, Aherne ST, Joyce H, O'Driscoll L, et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. Carcinogenesis 2013;34:2300–8.
[36] Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA 2011;305:1873–81.
[37] Li J, Lenferink AE, Deng Y, Collins C, Cui Q, Purisima EO, et al. Identification of high-quality cancer prognostic markers and metastasis network modules. Nat Commun 2010;1:34.
[38] Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA 2006;295:2492–502.
[39] Clarke CA, Keegan THM, Yang J, Press DJ, Kurian AW, Patel AH, et al. Age-specific incidence of breast cancer subtypes: understanding the black–white crossover. J Natl Cancer Inst 2012;104:1094–101.
[40] Keegan THM, Derouen MC, Press DJ, Kurian AW, Clarke CA. Occurrence of breast cancer subtypes in adolescent and young adult women. Breast Cancer Res 2012;14:R55.
[41] Ballatori N, Krance SM, Notenboom S, Shi S, Tieu K, Hammond CL. Glutathione dysregulation and the etiology and progression of human diseases. Biol Chem 2009;390:191–214.

[42] Townsend DM, Tew KD, Tapiero H. The importance of glutathione in human disease. Biomed Pharmacother 2003;57: 145–55.
[43] Franco R, Schoneveld OJ, Pappa A, Panayiotidis MI. The central role of glutathione in the pathophysiology of human diseases. Arch Physiol Biochem 2007;113:234–58.
[44] Yang P, Ebbert JO, Sun Z, Weinshilboum RM. Role of the glutathione metabolic pathway in lung cancer treatment and prognosis: a review. J Clin Oncol 2013;24:1761–9.
[45] Rahman I, Macnee W. Oxidative stress and regulation of glutathione in lung inflammation. Eur Respir J 2000;16:534–54.
[46] Estrela M, Ortega A, Obrador E. Glutathione in cancer biology and therapy. Crit Rev Clin Lab Sci 2006;43:143–81.
[47] Purohit V, Abdelmalek MF, Barve S, Benevenga NJ, Halsted CH, Kaplowitz N, et al. Role of *S*-adenosylmethionine, folate, and betaine in the treatment of alcoholic liver disease: summary of a symposium. Am J Clin Nutr 2007;86:14–24.
[48] Yang P, Mandrekar SJ, Hillman SH, Allen Ziegler KL, Sun Z, Wampfler JA, et al. Evaluation of glutathione metabolic genes on outcomes in advanced non-small cell lung cancer patients after initial treatment with platinum-based chemotherapy: an NCCTG-97-24-51 based study. J Thorac Oncol 2009;4:479–85.
[49] Liu X, Wang Y, Ji H, Aihara K, Chen L. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res 2016;44:e164.
[50] Zhu J, Shibasaki F, Price R, Guillemot JC, Yano T, Dotsch V, et al. Intramolecular masking of nuclear import signal on NF-AT4 by casein kinase I and MEKK1. Cell 1998;93:851–61.
[51] Iampietro M, Gravel A, Flamand L. Inhibition of breast cancer cell proliferation through disturbance of the calcineurin/NFAT pathway by human herpesvirus 6B U54 tegument protein. J Virol 2014;88:12910–4.
[52] Yiu GK, Toker A. NFAT induces breast cancer cell invasion by promoting the induction of cyclooxygenase-2. J Biol Chem 2006;281:12210–7.
[53] Quang CT, Leboucher S, Passaro D, Fuhrmann L, Nourieh M, Vincent-Salomon A, et al. The calcineurin/NFAT pathway is activated in diagnostic breast cancer cases and is essential to survival and metastasis of mammary cancer cells. Cell Death Dis 2015;6:e1658.
[54] Martin TA, Watkins G, Mansel RE, Jiang WG. Loss of tight junction plaque molecules in breast cancer tissues is associated with a poor prognosis in patients with breast cancer. Eur J Cancer 2004;40:2717–25.
[55] Tobioka H, Sawada N, Zhong Y, Mori M. Enhanced paracellular barrier function of rat mesothelial cells partially protects against cancer cell penetration. Br J Cancer 1996;74:439–45.
[56] Martin TA, Mansel RE, Jiang WG. Loss of occludin leads to the progression of human breast cancer. Int J Mol Med 2010;26: 723–34.
[57] Elias BC, Suzuki T, Seth A, Giorgianni F, Kale G, Shen L, et al. Phosphorylation of Tyr-398 and Tyr-402 in occludin prevents its interaction with ZO-1 and destabilizes its assembly at the tight junctions. J Biol Chem 2009;284:1559–69.
[58] Takenaga Y, Takagi N, Murotomi K, Tanonaka K, Takeo S. Inhibition of Src activity decreases tyrosine phosphorylation of occludin in brain capillaries and attenuates increase in permeability of the blood-brain barrier after transient focal cerebral ischemia. J Cereb Blood Flow Metab 2009;29:1099–108.
[59] Rao R. Occludin phosphorylation in regulation of epithelial tight junctions. Ann N Y Acad Sci 2009;1165:62–8.
[60] Zhao J, Zhou Y, Zhang X, Chen L. Part mutual information for quantifying direct associations in networks. Proc Natl Acad Sci U S A 2016;113:5130–5.
[61] Zhang X, Liu K, Liu ZP, Duval B, Richer JM, Zhao XM, et al. NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. Bioinformatics 2013;29:106–13.
[62] Zhang X, Zhao J, Hao JK, Zhao XM, Chen L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic Acids Res 2015;43:e31.
[63] Yu X, Zhang J, Sun S, Zhou X, Zeng T, Chen L. Individual-specific edge-network analysis for disease prediction. Nucleic Acids Res 2017;45:e170.
[64] Chen L, Liu R, Liu ZP, Li M, Aihara K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Sci Rep 2012;2:342.
[65] Liu R, Chen P, Aihara K, Chen L. Identifying early-warning signals of critical transitions with strong noise by dynamical network markers. Sci Rep 2015;5:17501.
[66] Yang B, Li M, Tang W, Liu W, Zhang S, Chen L, et al. Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. Nat Commun 2018;9:678.
[67] Hammond ME, Hayes DF, Wolff AC, Mangu PB, Temin S. American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Oncol Pract 2010;6:195–7.
[68] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323.
[69] Ya RA, Series T, Lp S. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
[70] Ishwaran H, Udaya M, Kogalur B. Random survival forests. Ann Appl Stat 2008:841–60.