Research Paper

Immunoglobulin superfamily genes are novel prognostic biomarkers for breast cancer

Yue Li[1,*], Maoni Guo[2,*], Zhenkun Fu[3,*], Peng Wang[2], Yan Zhang[2], Yue Gao[2], Ming Yue[2], Shangwei Ning[2], Dianjun Li[3]

[1] Department of Medical Oncology, Harbin Medical University Cancer Hospital, Harbin, 150081, China

[2] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China

[3] Department of Immunology, Harbin Medical University, Harbin, 150081, China

[*] These authors contributed equally to this work

Correspondence to: Dianjun Li, **email:** dianjunli@163.com
Shangwei Ning, **email:** ningsw@ems.hrbmu.edu.cn

Keywords: *immunoglobulin superfamily, breast cancer, network module, topology feature, prognostic biomarker*

Received: *July 18, 2016* ***Accepted:*** *November 22, 2016* ***Published:*** *November 29, 2016*

ABSTRACT

Breast cancer progression is associated with dysregulated expression of the immunoglobulin superfamily (IgSF) genes that are involved in cell-cell recognition, binding and adhesion. Despite widespread evidence that many IgSF genes could serve as effective biomarkers, this potential has not been realized because the studies have focused mostly on individual genes and not the entire network. To gain a global perspective of the IgSF-related biomarkers, we constructed an IgSF-directed neighbor network (IDNN) and an IgSF-directed driver network (IDDN) by integrating multiple levels of data, including IgSF genes, breast cancer driver genes, protein-protein interaction (PPI) networks and gene expression profiling data. Our study shows that IgSF genes in the PPI network have important topological features related to cancer. Most IgSF genes are either cancer driver genes themselves or associated with them. We also identified a 21-gene IgSF network module with enriched mutations that are associated with overall survival based on 450 breast cancer patient samples extracted from The Cancer Genome Atlas (TCGA) and multiple independent microarray validation datasets. These results highlight the potential of IgSF genes as novel diagnostic, prognostic and therapeutic targets for breast cancer.

INTRODUCTION

Breast cancer is the leading cause of cancer death among women worldwide. In Chinese women, breast cancer is the most prevalent form of cancer with more than 1.6 million people diagnosed and 1.2 million people dying every year. The most common type of breast cancer is invasive ductal carcinoma (IDC) that can spread from the ducts or the lobules to the surrounding tissue. Prognostic biomarkers are useful to choose the appropriate treatment for IDC, and they significantly affect the process of cancer therapy [1–3]. Studies have shown that the genetic diversity in breast cancer impacts response to treatment and patient outcomes. This is exemplified by the estrogen receptor negative (ER−) and positive (ER+) subtypes that have different prognostic gene signatures and responses to treatment [4]. Therefore, there is scope to identify novel signatures that can enhance predicting the prognostic and clinical behavior.

Gene expression of many IgSF members is altered in breast cancer, and hence, they are promising candidates as prognostic biomarkers. ALCAM (CD166) is a potential breast cancer biomarker and a therapeutic target due to its role in induction of programmed cell death, apoptosis and autophagy in breast cancer [5]. Down regulation of CXCR4 inhibits cell migration in breast cancer cells [6]. The expression of MUC18 (CD146) promotes the progression of human breast cancer cells by increasing their motility, invasiveness and tumorigenesis [7]. L1CAM is potentially an early diagnostic biomarker in breast cancer progression as it promotes cell adhesion and migration *in vitro* [8]. Although these findings demonstrate the important role of IgSF members in breast cancer progression and metastasis, these studies focus on one or a few IgSF members analyzed in either cell lines or in limited patient samples and therefore do not present a global perspective of the entire immunoglobulin superfamily.

Current advances in cancer biology and genomic methods have generated large-scale gene expression profiling datasets (such as TCGA) and other OMICs and provide an opportunity to study the entire network of IgSF genes as well as identify novel biomarkers for breast cancer. Previously, Li and others used cancer gene microarray and network data to develop a network-based method for cancer prognostic biomarker identification [9]. Similarly, Chuang and others used a protein-network-based approach using data from protein interaction databases to identify markers as sub networks [10]. These studies demonstrate that integrating gene expression and protein-protein interaction data can improve prediction performance in network biomarker identification.

Our aim was to study the role of the IgSF network in breast cancer and test its diagnostic, prognostic and therapeutic potential. Towards this goal we constructed an IgSF-directed neighbor network (IDNN) and an IgSF-directed driver network (IDDN) to address the role of the IgSF network in breast cancer. We identified IgSF genes in the PPI network with hub topological features connected to breast cancer. We also identified a 21-gene module from the IDDN network that was associated with the overall survival of breast cancer patients. This module included several key IgSF and breast cancer driver genes with enriched mutations, demonstrating the functional significance of the IgSF genes to predict breast cancer. Our findings highlight the novel role of the IgSF-directed network in breast cancer. It also highlights their potential for biomarker-guided development of preclinical and clinical therapeutic modalities.

RESULTS

IgSF genes play a crucial role in breast cancer

We constructed an IgSF-directed neighbor network (IDNN) that had 1050 nodes including 283 IgSF genes and their 767 neighbors from the PPI network (Figure 1A, Supplementary Table S1). We found 33 IgSF genes (KIT, PDGFRB, KDR, FGFR1, CD28, NTRK1, etc.) that were themselves breast cancer driver genes as well as 250 IgSF genes (including PTPN11, LCK, GRB2, ABL1, APP, etc. Supplementary Table S2) that associated with the breast cancer driver genes. The top 10 IgSF genes (KIT, PDGFRB, KDR, FGFR3, FGFR1, CD4, NTRK1, IGSF21, HSPG2, and PDGFRA) had a direct connection in the network, indicating that IgSF genes play hub roles in the sub-networks (Figure 1B). We also found a greater degree of genes (mean = 16.76) that were both IgSF and breast cancer driver genes in the IDNN suggesting a intricate link between the IgSF genes and breast cancer (Figure 1C, Supplementary Figure S1). Also, enrichment of IgSF cancer driver genes in specific modules suggests that the IgSF driver genes may play consequential roles in special biological modules (Supplementary Figure S2, Supplementary Table S3).

Notably, six of the top 10 IgSF genes in the PPI network were well-known breast cancer driver genes, including KIT, whose high expression occurs infrequently in breast cancer [11]. In the triple negative breast cancer, immunohistochemical expression of C-kit and mutations of PDGFRA are frequent suggesting that they are good candidates for molecular targeted therapy [12]. Our findings also suggest that CD4^{+} follicular helper T cells (Tfh) may be prognostic indicators as they are found in the breast tumors [13]. Furthermore, down-regulation of KDR expression induces apoptosis in breast cancer cells [14]. Also, FGFR1 activity is required for the survival of an FGFR1-amplified breast cancer cell line [15]. Among the IgSF neighbor genes, ABL1 and PDGFR are well known breast cancer driver genes that promote acquired resistance to aromatase inhibitor (AI) therapy in ER+ breast cancers [16]. The activation of the PDGFR and ABL1 pathways is associated with long-term estrogen deprivation in MCF7 breast cancer cells and decreased anti-proliferative response to AI treatment in primary ER-positive breast carcinomas.

IgSF genes directly interact with breast cancer driver genes

To further explore the relationship between IgSF and breast cancer driver genes, we constructed a network called IgSF-directed driver network (IDDN). This included IgSF genes and breast cancer driver genes that were extracted from the IDNN (Figure 2A, Supplementary Table S4). The IDDN contained 253 genes, of which 103 were IgSF and 121 were breast cancer driver genes. Among these, 29 of the IgSF genes were also breast cancer genes. Compared to the nodes in IDNN, the nodes in the IDDN had higher degrees, betweenness centrality and closeness centrality (avg. 3.440 vs. 3.220 for degrees, p = 0.001, Supplementary Figure S3A; avg. 0.020 vs. 0.012 for betweenness centrality, p = 2.74e-07, Supplementary Figure S3B; avg. 0.242 vs. 0.213 for closeness centrality, p = 6.78e-09, Supplementary Figure S3C; Wilcoxon rank sum test). This suggested that the IDDN obtained from IDNN was closer in structure and played a crucial role in the biogenesis of breast cancer.

To explore the biological functions of these genes, we conducted pathway analysis using the Subpathway Miner [17]. The T cell receptor signaling pathway emerged as the most significant pathway from this analysis with several of the IgSF genes located in key positions, such as CBL, LCK, ZAP70, FYN, CD28, PD-1, CTLA4, ICOS, and CD4/8 (Figure 2B). Since the expression of CBL can inhibit LCK and ZAP70 gene expression, CBL gene may play a crucial role in the pathway [18].

We then performed a GO enrichment analysis using IgSF and breast cancer driver genes. Some common GO terms that we found associated with breast cancer included cell surface receptor-linked signal transduction, immune response, regulation of cell proliferation, regulation of cell activation, regulation of T cell activation, regulation of T cell differentiation, and negative regulation of the immune system process. Significance of this analysis is exemplified by the fact that mutations of BTLA, CD28, CD4, and CD8A genes in the GO term cell surface receptor-linked signal transduction contribute to sporadic breast cancer risk. More importantly, we found overlapping GO terms between the IgSF genes and the breast cancer driver genes, including phosphate metabolic process, protein amino acid phosphorylation, enzyme linked receptor protein signaling pathway, transmembrane receptor protein tyrosine kinase signaling pathway, cell surface receptor-linked signal transduction, and regulation of cell proliferation (Figure 2C). This indicated a close connection between IgSF genes and breast cancer. Additionally, the degree of the common genes of IgSF and the breast cancer driver genes derived from the IDDN was also greatest (Figure 2D).

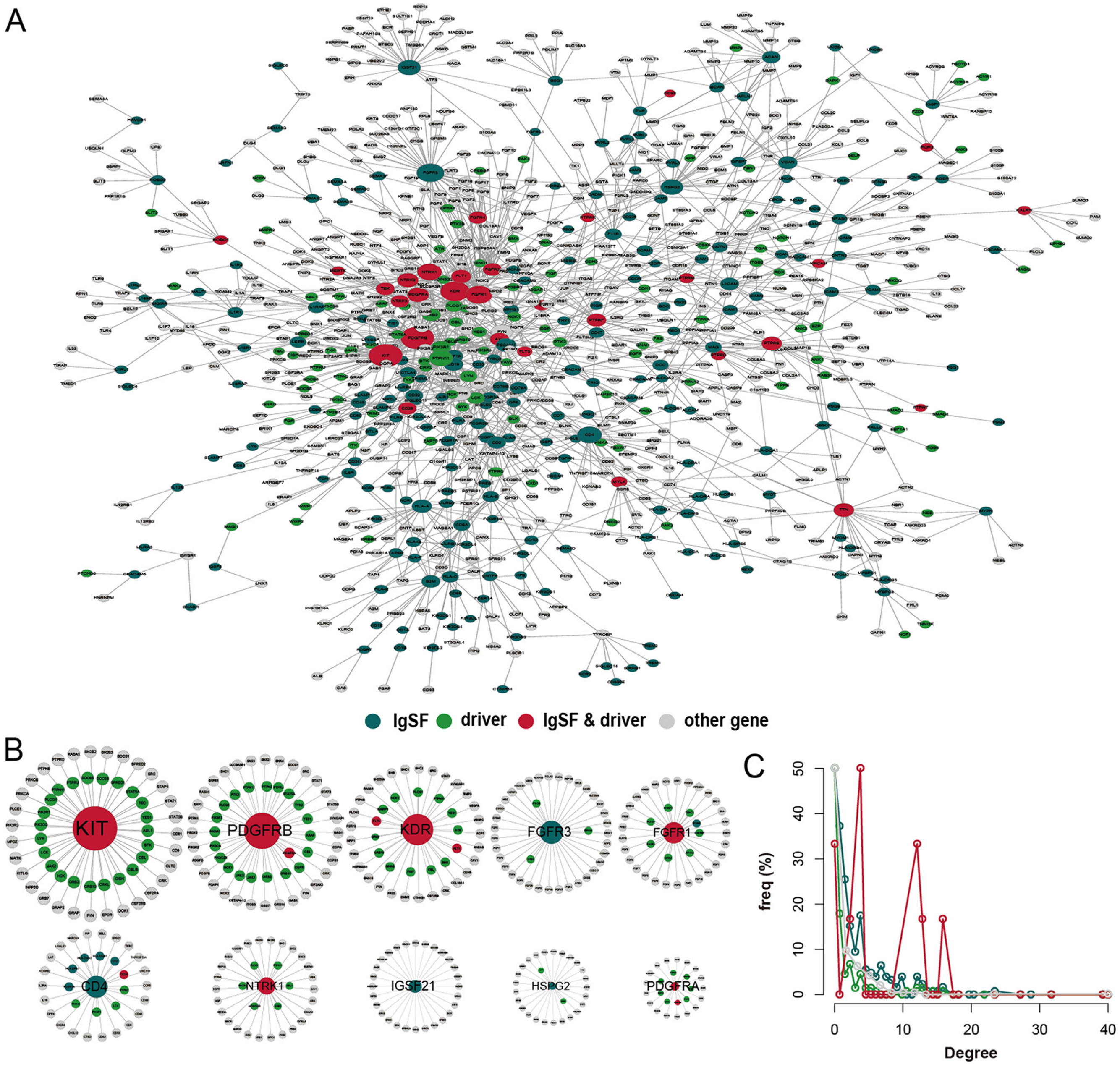

Figure 1: The properties of the IgSF-directed neighbor network (IDNN). (**A**) The global IDNN. (**B**) Top 10 IgSF genes ranked by gene degree (KIT, PDGFRB, KDR, FGFR3, FGFR1, CD4, NTRK1, IGSF21, HSPG2, and PDGFRA). (**C**) The degree distribution of the nodes in IDNN.

IgSF genes associate with driver genes in functional modules

To understand the communication between the IgSF and the breast cancer driver genes, we performed a module analysis in the IDDN. We found six significant IgSF-associated modules in which IgSF and breast cancer driver genes were closely associated (Figure 3). We conducted a GO terms enrichment analysis of the module genes using DAVID [19] and identified the biological process terms that enriched the module genes using a cutoff of FDR < 0.05 (Supplementary Table S5). The top-ranked GO terms included immune response, cell adhesion, biological adhesion, regulation of alpha-beta T cell proliferation, transmembrane receptor protein tyrosine kinase signaling pathway, and T cell activation. The GO term cell adhesion enriched by the second and sixth module genes is a very important biological process of the IgSF genes (VCAM1, ICAM1, F11R, ITGAL, EZR, ICAM4, ICAM5, ICAM2, ICAM3, ITGB5, ITGB2, and CD226) that participate in the immune response. Among these, the expression of members of the ICAM family regulates tumorigenesis and are potential diagnostic biomarkers and therapeutic targets for breast cancer [20]. Moreover, both the first and fourth modules enriched the common GO term phosphate metabolic process, which is differentially regulated in breast cancer [21].

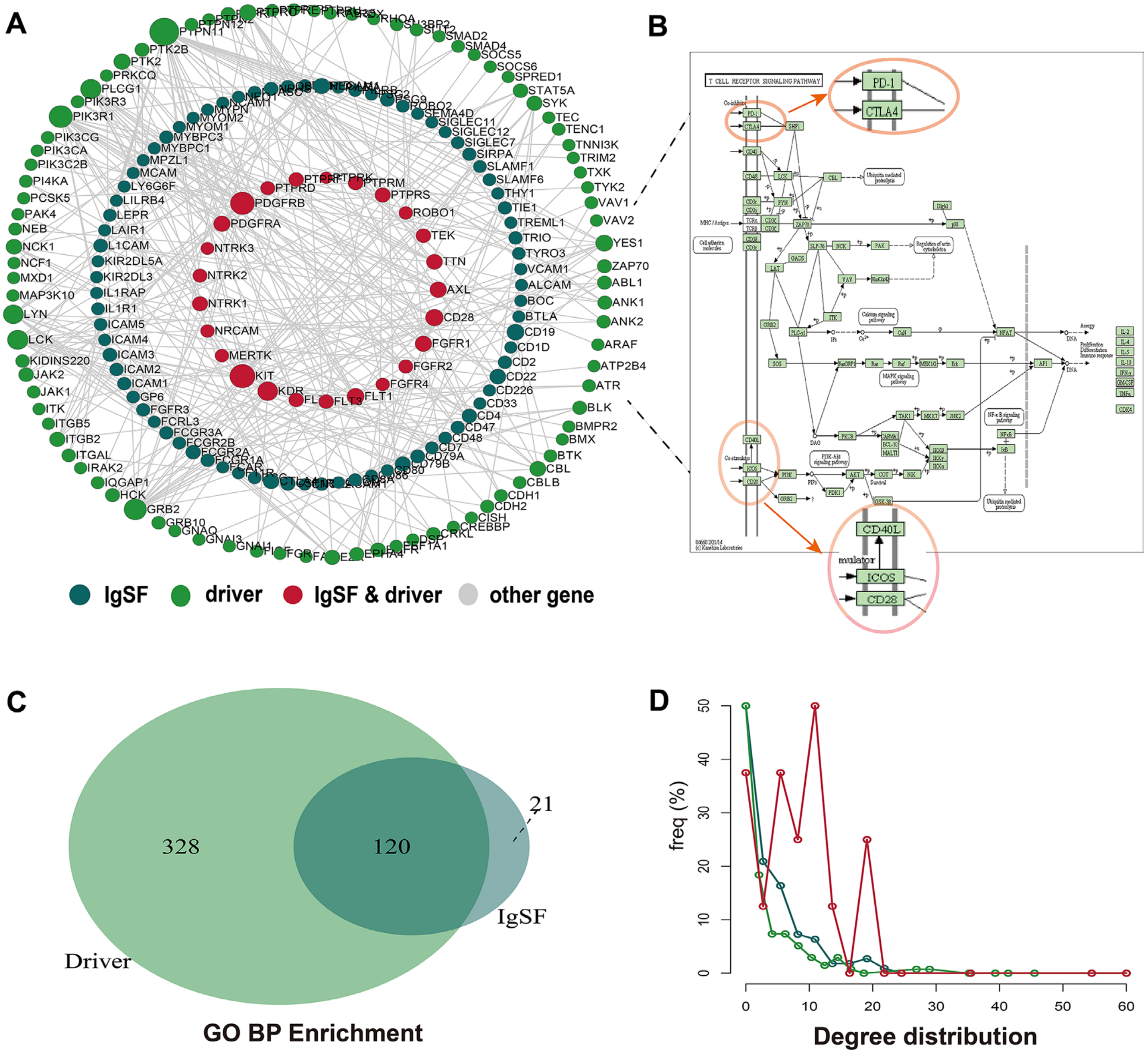

Figure 2: The properties of the IgSF-directed driver network (IDDN). (**A**) The global IDDN and the size of the sphere represent the degree of the gene. (**B**) The functional sub-pathway of the T cell receptor signaling pathway (**C**) The overlapping GO terms between IgSF and breast cancer driver genes. (**D**) The degree distribution of nodes in IDDN.

IgSF-related modules are enriched with cancer mutations

Since genetic mutations can cause cancer, we analyzed if IgSF genes in our modules are mutated. Previously, the IgSF genes in the module that includes well-known breast cancer genes have been shown to be mutated. Some SNPs in six immunological genes, BTLA, ITGAL, CTLA4, ICOS, PDCD1, and VTCN1 were reported as breast cancer risk mutations in previous studies [22–27] (Supplementary Table S6). We performed gene mutation enrichment analysis and evaluated if these module genes were enriched in the top 15% of mutated genes in breast cancer. We identified three module genes that overlapped with the top 15% mutated genes in breast cancer (*p* = 0.001 for module 1, Supplementary Figure S4A; *p* = 0.001 for module 4, Supplementary Figure S4B; *p* = 1.28e-07 for module 5, Supplementary Figure S4C; Hypergeometric Test).

Next, we identified the somatic mutations of the genes in the IDDN using TCGA breast cancer somatic mutation data and determined if they overlapped with the top 15% mutated genes in the IDDN using hypergeometric test. We found that the second module with 21 genes was enriched in mutated genes (Figure 4A, $P < 0.001$). Four of the six common genes (PTPN11, TRIM2, FGR and ZAP70) were IgSF genes (Figure 4B). PTPN11, a HER2-inhibition up-regulated PTP (protein tyrosine and dual-specificity phosphatase) transduces positive signals and is an oncogene [28]. TRIM2 is a diagnostically significant and conserved element of the SOX10 signature in BBC (breast basal-like carcinomas) cell lines [29]. FGR and ZAP70 have diagnostic and therapeutic potential due to their relationship with breast cancer development and progression [30, 31].

IgSF-mutated module has prognostic potential

In the TCGA gene expression data, only 20 genes (BTLA, CD33, FCRL3, FGR, LAIR1, LEPR, LILRB4, MPZL1, PILRA, PILRB, PTPN11, SIGLEC11, SIGLEC12, SIGLEC7, SIRPA, SLAMF1, SLAMF6, TREML1, TRIM2, and ZAP70) had expression values. We created a risk-score formula according to the expression of

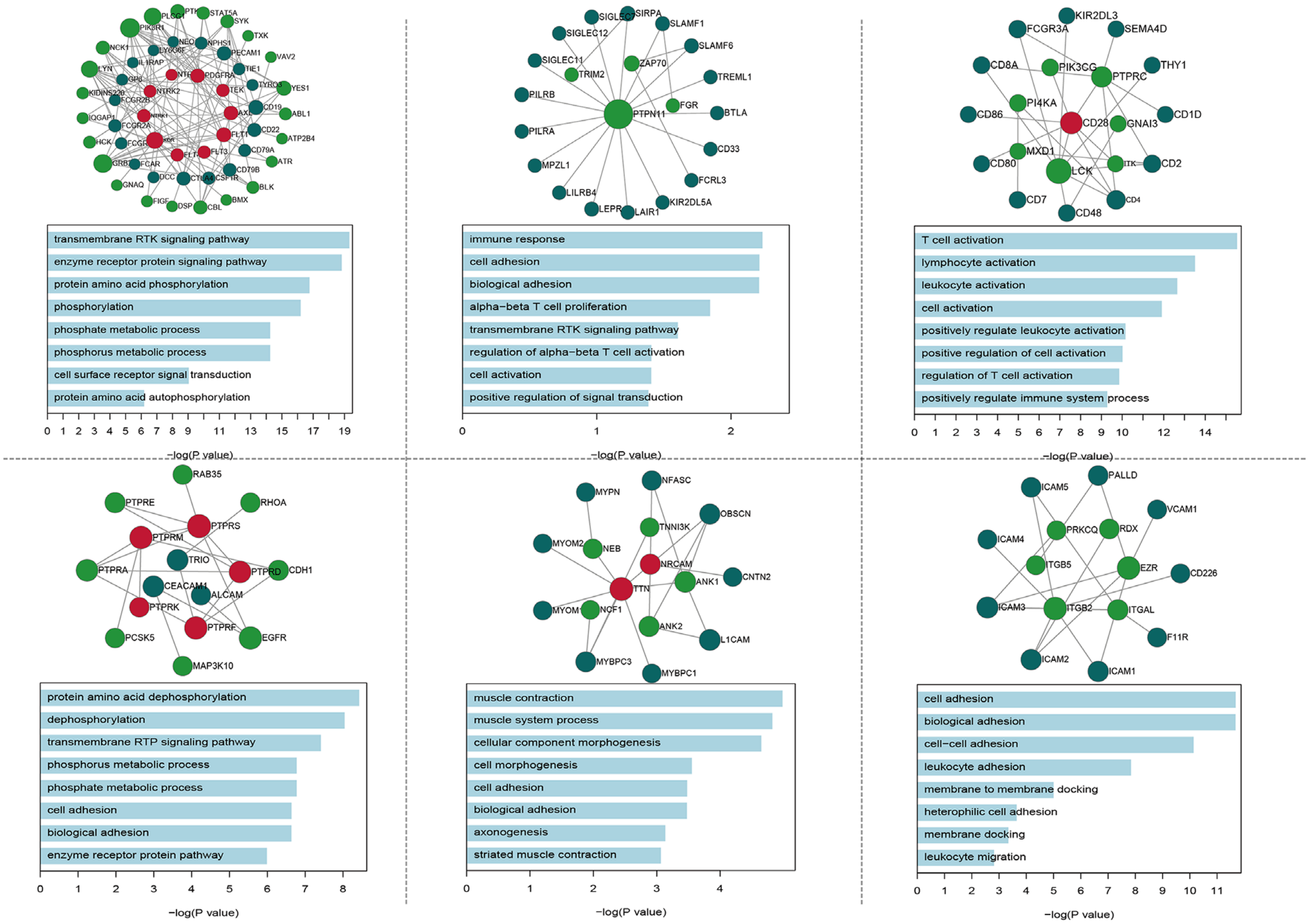

Figure 3: The GO BP terms of six significantly functional modules.

these 20 genes to generate OS (overall survival) prediction (see the Material and Methods section). The Cox regression coefficients of the total samples, ER+ samples and ER– samples are as listed in Supplementary Table S7. Using the median risk score of the test series as the cutoff point, we calculated the risk scores for the 20 genes for each patient and then ranked the patients according to their risk score. The patients grouped into a high-risk (*N* = 225) or a low-risk (*N* = 225) category by using the median risk score of the test series as the cutoff point. Patients in the high-risk group had significantly shorter median OS than those in the low-risk group (Figure 5A, HR = 3.82, *P* = 2.81e-05). In addition, patients with high risk in the ER+ series (HR = 2.78, *P* = 0.00507) as well as the ER– series (HR = 8.77, *P* = 0.0124) had significantly shorter median OS than those in the low-risk group (Figure 5A). Based on the distribution of gene risk score, the survival status and the gene expression signature of the breast cancer patients, patients with high risk scores expressed higher levels in the ten risk genes, whereas patients with low risk scores expressed higher levels in the remaining ten protective genes (Figure 5B–5D).

Validating the prognostic potential of the IgSF module in independent patients

To confirm the prognostic value of the IgSF module, we validated the 21 gene signatures in four independent microarray datasets (Table 1). Using the same risk score formula, we classified patients in GSE4922 into high-risk and low-risk groups using the median score of the test series as the cutoff point. Consistent with our previous findings, patients in the high-risk group had significantly shorter median OS than those in the low-risk group (Figure 6A, all samples *P* = 0.076; Figure 6B, ER+ samples *P* = 0.012; Figure 6C, ER– samples *P* = 0.021). Similarly, we classified the patients in GSE7390 into a high- and a low-risk group (Figure 6D, all patients, *P* = 0.091; Figure 6E, ER+ patients, *P* = 0.018 and Figure 6F, ER– patients, *P* = 0.054) and obtained similar results.

After further adjusting for other markers, the univariate analysis indicated that the IgSF module, as an independent risk factor, was significantly associated with the overall survival of breast cancer patients from TCGA (HR = 2.71, 95% CI: 2.02–3.64, *P* = 2.75e-11), GSE4922 (HR = 1.69, 95% CI: 0.94–3.05, *P* = 0.079), and GSE7390 (HR = 1.58, 95% CI: 0.92–2.73, *P* = 0.094) datasets (Table 2). Further, when multivariate analysis was performed to investigate the independence of the module to other clinical factors, the high- and low-risk groups remained independent of other clinical factors in TCGA patients (HR = 2.60, 95% CI: 1.84–3.67, *P* = 4.32e-8), GSE4922 (HR = 1.86, 95% CI: 1.01–3.43, *P* = 0.045), and GSE7390 (HR = 1.70, 95% CI: 0.97–2.96, *p* = 0.060) datasets (Table 2). Data stratification analysis on TCGA patients also indicated that the module was independent of PR, age and tumor stage. It performed similarly in PR category (log-rank test *p* = 0.005 for PR+ group and log-rank test *p* = 0.002 for PR- group) as well as the age category older or younger than 60 (log-rank test *p* = 0.01 for the older patients and log-rank test *P* = 9.64e-4 for the younger patients) and tumor stage III/IV category (log-rank test *P* = 0.012) (Figure 7A–7E). Similar trend was observed in GSE4922 patients with G3 grade (log-rank

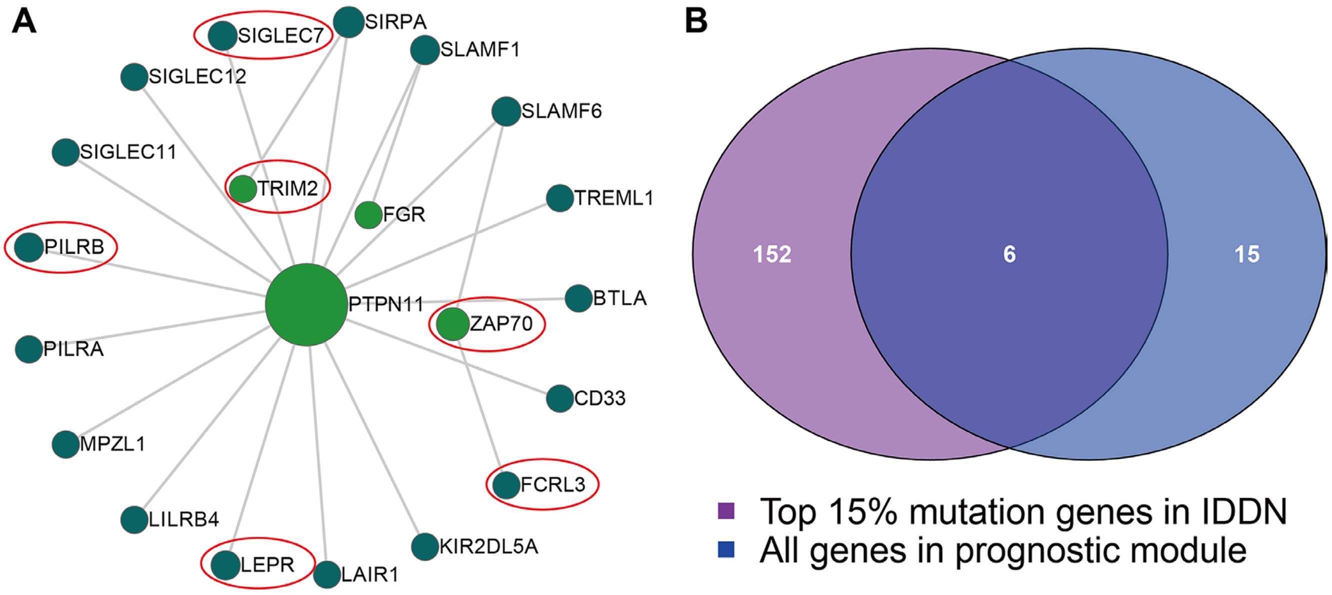

Figure 4: The module with mutated IgSF genes. (**A**) The module includes 21 genes, with the size of the sphere depicting the degree of the gene and the red circles showing the six common mutated genes found in this module. (**B**) A Venn plot between the IgSF and the mutated genes.

Table 1: The independent microarray datasets used in this study

Datasets	Platform	Number of patients	Overall type	Number of ER+ patients	Number of ER– patients
GSE4922	HG-U133A	289	OS	211	34
GSE7390	HG-U133A	198	OS&DMFS	134	64

Abbreviations: Note: OS, Overall Survival; DMFS, Distant Metastasis-Free Survival.

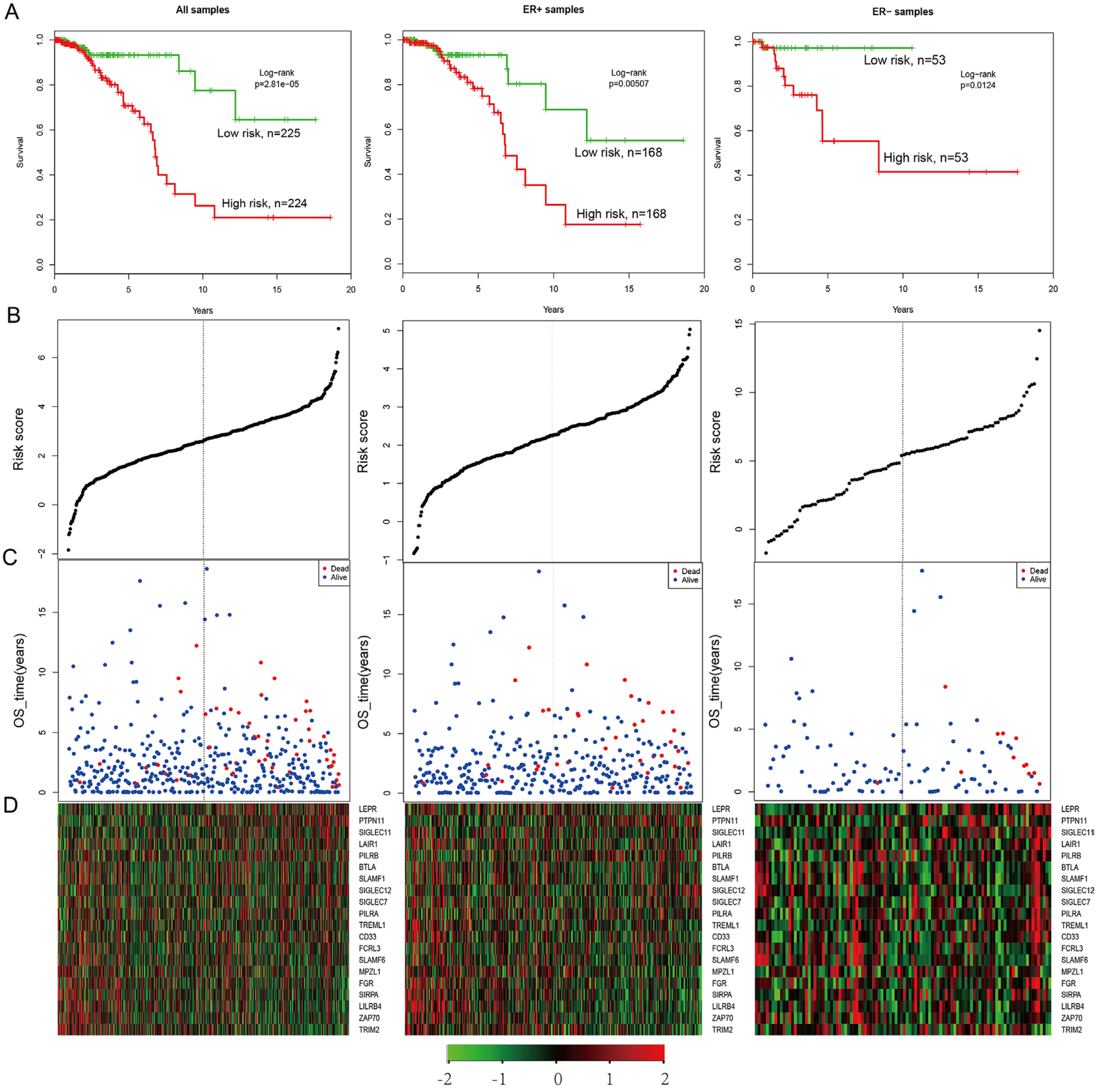

Figure 5: The survival analysis of the IgSF-mutated module. (**A**) The Kaplan–Meier curve for the overall survival of two patient groups with high and low risk scores in the TCGA patient set (n = 450), ER+ samples (N = 337) and ER– samples (N = 106). The difference between the two curves was evaluated by a two-sided log-rank test. (**B**) The gene-based risk score distribution of the 20 genes (**C**) The gene-based patient survival status of the 20 genes. (**D**) The heat map depicting expression profiles of the 20 genes. The black dotted line represents the cutoff value of the risk score derived from the corresponding set that separates patients into high- and low-risk groups.

Table 2: Statistical analysis of the IgSF module gene signature and overall survival of breast cancer patients in the TCGA and GEO cohorts

Variables	Univariable model			Multivariable model		
	HR	95% CI of HR	*P*-value	HR	95% CI of HR	*P*-value
TCGA (*N* = 450)						
Module risk score	2.7183	2.0252–3.6485	< 0.0001	2.6063	1.8499–3.6720	< 0.0001
ER	1.1332	0.5728–2.2418	0.7195	0.6366	0.2439–1.6615	0.3562
PR	0.7344	0.4214–1.2799	0.2761	0.3486	0.1353–0.8984	0.0291
HER2	0.9361	0.5859–1.4954	0.7823	0.7739	0.4528–1.3227	0.3486
Age	1.0275	1.0044–1.0511	0.0194	1.0529	1.0193–1.0876	0.0018
Stage						
II	2.1360	0.7406–6.1609	0.1602	1.6061	0.5178–4.9821	0.4120
III/IV	3.5429	1.1950–10.5034	0.0225	3.5199	1.0733–11.5434	0.0378
GSE4922 (*N* = 289)						
Module risk score	1.6933	0.9400–3.0505	0.0795	1.8644	1.0133–3.4306	0.0453
ER	0.8583	0.4667–1.5785	0.6230	1.1735	0.6184–2.2271	0.6245
Age	0.9971	0.9815–1.0130	0.7223	0.9999	0.9847–1.0152	0.9862
Grade						
G2	1.8232	1.0336–3.2160	0.0381	1.6455	0.9269–2.9211	0.0890
G3	3.1634	1.6929–5.9113	0.0003	3.2902	1.7080–6.3381	0.0004
GSE7390 (*N* = 226)						
Module risk score	1.5873	0.9234–2.7285	0.0946	1.7023	0.9762–2.9686	0.0608
ER	0.4755	0.2806–0.8060	0.0058	0.4239	0.2277–0.7891	0.0068
Age	1.0131	0.9771–1.0505	0.4802	1.0110	0.9752–1.0482	0.5516
Grade						
G2	1.1056	0.4738–2.5799	0.8165	1.0323	0.4392–2.4266	0.9419
G3	1.3785	0.5976–3.1799	0.4516	0.9614	0.3795–2.4353	0.9338

test P = 6.48e-07; Figure 7F). Based on all these data, we conclude that the IgSF-related module is a strong prognostic indicator for breast cancer.

DISCUSSION

Elucidating the functional significance and molecular mechanism of IgSF members in breast cancer may provide new opportunities for the early detection and treatment of cancer. A variety of molecular biomarkers in breast cancer have been previously identified. Van de Vijver and others categorized breast cancer based on the gene expression profile of 70 genes and found that their classification was better than standard systems that were based on clinical and histological criteria [32] . Pawitan and others developed a 64-gene signature to predict the response of breast cancer patients to therapy [33]. Venet and others showed that many random gene expression signatures are significantly associated with breast cancer survival, although the underlying mechanism of the biomarkers was unclear implying the need for more effective biomarkers [34]. We performed a global analysis of the IgSF genes in breast cancer and developed a network-based strategy to identify the IgSF-related PPI network and modules. Our study shows that the IgSF network genes closely interact with the cancer driver genes in the mutations enriched module and can serve as a survival-associated biomarker for breast cancer. Also, the IgSF genes participate in the dysregulation of breast cancer driver genes. These results provide a novel understanding of the role of IgSF network in breast cancer progression. Our study also highlighted the importance of immune-related IgSF genes. Immune genes play a key role in cell-cell communication, and dysfunctional immune response cause various diseases in humans including cancers [9]. In breast cancer, the immune genes may modulate the communication between malignant cells and normal cells. We had postulated that the IgSF-related modules mediate breast cancer progression by regulating cancer metastasis. However, functional enrichment analysis showed that the IgSF-related modules were involved in a variety of biological events, including immune response, cell adhesion, biological adhesion and regulation of alpha-beta T cell proliferation. This suggests that in the early stage of breast cancer, IgSF-related modules restrict cell proliferation by regulating signal transduction and the immune response.

In summary, our data shows that the IgSF-related module that is of prognostic importance. Our data also reveal novel functional insights for dysregulated IgSF-related members and modules in the etiology of breast cancer.

MATERIALS AND METHODS

Human IgSF gene dataset

The IgSF gene data was downloaded from the HUGO Gene Nomenclature Committee website (HGNC, http://www.genenames.org/), constituting 478 IgSF genes from 651 records, including 42 C1-set, 40 C2-set, 245 Immunoglobulin-like, 161 I-set and 163 V-set domain containing records.

Breast cancer driver gene dataset

We downloaded cancer associated genes from the Catalogue of Somatic Mutations in Cancer (COSMIC v70; Aug 2014), which is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer. Firstly, we obtained the breast cancer associated genes from the Cancer Gene Census by searching the keyword "breast" in the tumor types. In addition, we searched cosmic mutation data with the keyword "breast" in the primary site, "y" if the entire genome/exome is sequenced and "CANCER" if the mutations affected the tumor generation in pathology. In summary, we obtained 1,307 unique genes acting as the breast cancer driver genes for subsequent analyses.

Human protein-protein interaction data

The protein-protein interaction (PPI) data was downloaded from the Human Protein Reference Database (HPRD Release9, http://www.hprd.org/) [35]. It contained more than 42,000 manually curated interactions between 9,826 human genes.

Analysis of breast cancer gene expression and somatic mutations in clinical datasets

In this study, we focused only on invasive ductal carcinoma (IDC) patients with breast cancer. The breast cancer gene expression, somatic mutation data and the corresponding clinical data was downloaded from The

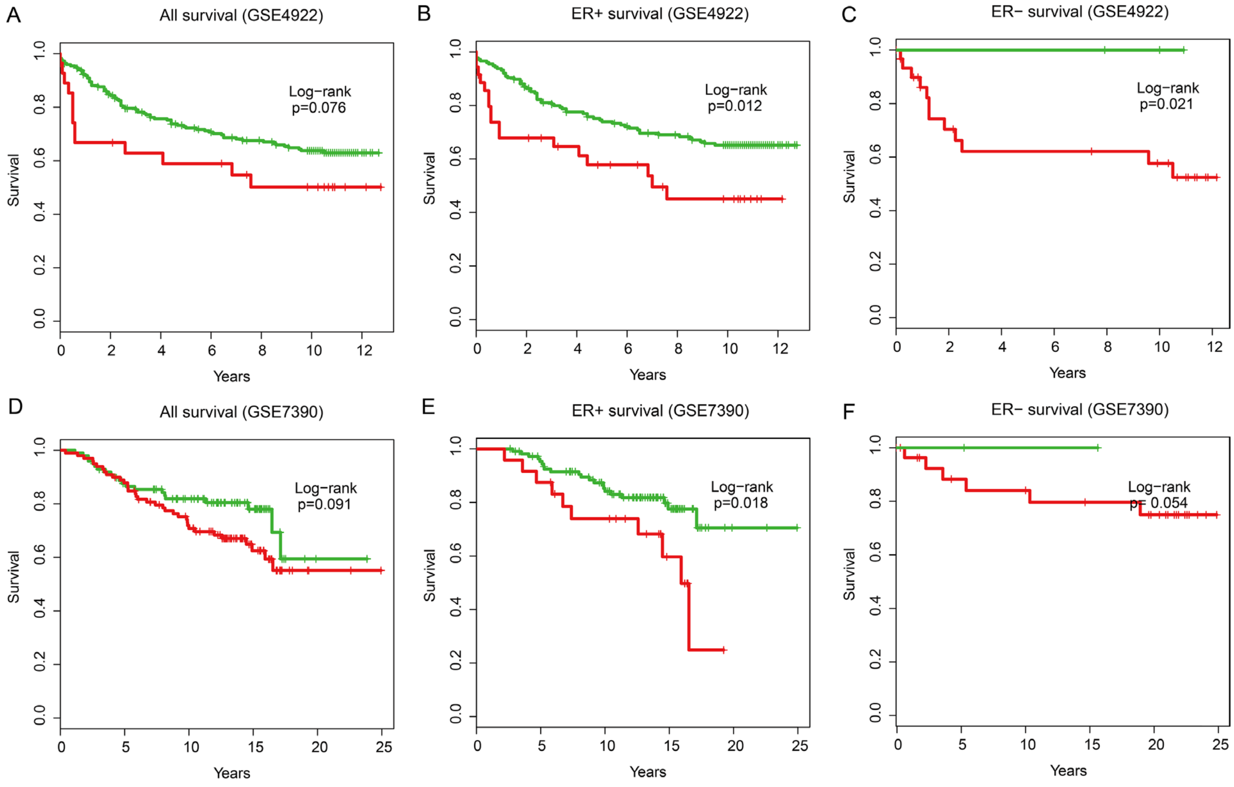

Figure 6: The overall survival analysis in independent validation datasets. Kaplan–Meier survival curves were plotted for (**A**) GSE4922 all patients (*n* = 289), (**B**) GSE4922 ER+ patients (*n* = 211), (**C**) GSE4922 ER– patients (*n* = 34), (**D**) GSE7390 all patients (*n* = 198), (**E**) GSE7390 ER+ patients (*n* = 134) and (**F**) GSE7390 ER– patients (*n* = 64).

Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). The gene expression profiling from 591 breast cancer patients was generated by the UNC__AgilentG4502A_07_3 microarray platform. The samples that contained only expression profiling without clinical data were removed from the analysis. We generated 450 IDC samples that had both expression profiling and clinical data, involving 17,814 genes. Also, the patients were further stratified into the estrogen receptor negative (ER−) and positive (ER+) groups.

Validation of microarray datasets

The four independent microarray breast cancer datasets [36–39] that were used in this study were obtained from the Gene Expression Omnibus (GEO) database. All the datasets were produced by the HGU133A or HGU133-PLUS2 platform. The datasets were chosen based on the criteria of no less than 100 samples and the availability of clinical outcome data. Raw microarray datasets were normalized using Robust Multichip Average [40]. The three main steps were as follows: background correction, quantile-normalization and log2-transformation. All the probes were mapped based on their EntrezGeneID. When multiple probes were mapped to the same gene ID, the mean value was used to represent the expression value of the single gene. To account for differences in systematic measurement between different datasets, each dataset was standardized independently by transforming the expression of each gene into a mean of 0 and a standard deviation of 1.

Network construction and analysis of topological features

We constructed a human PPI network based on the HPRD data. Then, we distinguished all the IgSF genes in this network using HGNC IgSF family data. Finally, using IgSF genes and their direct interacting genes in the network (referred to as IgSF neighbor genes), we constructed a sub-network of the human PPI network named the IgSF-directed neighbor network (IDNN). Furthermore, we extracted all the IgSF genes and the breast cancer driver genes to construct an IgSF-directed driver network (IDDN) for subsequent analysis. The Cytoscape software was used for the construction of

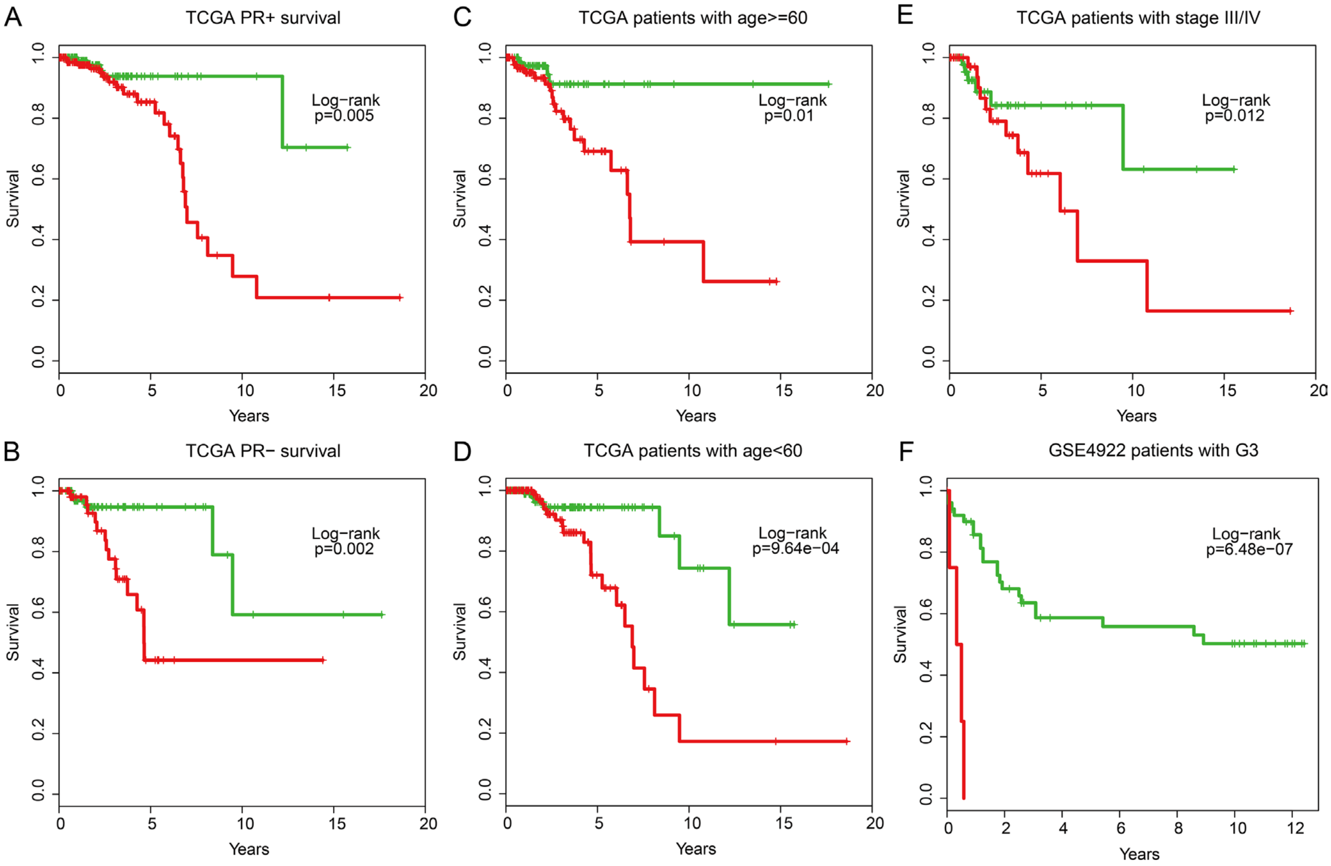

Figure 7: Stratification analyses of the IgSF mutated module with available PR, age, tumor stage and grade information for all patients. (**A**) Kaplan-Meier survival curves for TCGA patients with PR+, (**B**) Kaplan-Meier survival curves for TCGA patients with PR–. (**C**) Kaplan-Meier survival curves for TCGA patients with age > = 60, (**D**) Kaplan-Meier survival curves for TCGA patients with age < 60. (**E**) Kaplan-Meier survival curves for TCGA patients with stage III/IV. (**F**) Kaplan-Meier survival curves for GSE4922 patients with G3 grade. *P*-values were calculated using the log-rank test.

networks. The topological properties of the IgSF genes were analyzed in both IDNN and IDDN.

Identifying functional modules in the IgSF networks

Based on the IDDN constructed by the IgSF and breast cancer driver genes, we identified all the network modules using GraphWeb [41], a web server for identifying the network-based biomarkers that best represent the property of the network. The GraphWeb web server was made of three component processes: (i) Network datasets (to input human protein-protein interaction pairs; IDDN in this study), (ii) Network algorithm (we used the betweenness centrality clustering method and the default values were set) and (iii) Network settings (including default edge settings, node setting and module settings with less than 3 nodes and insignificant modules hidden).

Statistical analysis

Hypergeometric test was used to explore the overlap between the genes in the IgSF-related modules and the top mutated genes in the sub-network for breast cancer. We also studied the enrichment of breast cancer driver genes in IgSF by using this method. Univariate and multivariate analyses were performed using Cox proportional hazards regression model to determine whether the IgSF-related prognostic module was independent of other clinical variables, and adjusted for ER, PR, HER2, age, stage and grade. Hazard ratio (HR) and 95% confidence intervals (CI) were estimated by Cox proportional hazards regression model. To verify if the modules we identified are associated with patient survival, we determined the regression coefficient of every gene in the module related to patient survival using the gene expression data. The classifier was built as a linear combination of the gene expression values of select immune-related genes with the standardized Cox regression coefficient as the weight. A risk score formula for each patient was established by including the expression values of each selected gene, weighed by their estimated regression coefficients in the multivariate Cox regression analysis [42]. Finally, the patients were divided into high- and low-risk groups using the median of the risk score as the threshold. The patients with high-risk scores were classified as poor outcomes. Kaplan-Meier survival plots and log-rank tests by R package "survival" were used to assess the differences in overall survival (OS) time between the high- and low-risk patients. Bioinformatic analysis was performed with R 3.0.0 statistical software.

Functional enrichment analysis

Functional enrichment analysis at the GO and KEGG levels were performed using DAVID bioinformatics resources (http://david.abcc.ncifcrf.gov/, version 6.7) [43]. The DAVID enrichment analysis was limited to KEGG pathways and GO-FAT biological process (BP) terms with the whole human genome as background. Functional categories with a *p*-value of < 0.05 were considered statistically significant.

CONFLICTS OF INTEREST

No potential conflicts of interest were disclosed.

FUNDING

This work was supported by the National Natural Science Foundation of China [31401090], the Postdoctoral Science Foundation of China [2015M571432, 2016T90308], the Postdoctoral Foundation of Heilongjiang Province [LBH-Z14148], Harbin Special Funds for Innovative Talents of Science and Technology Research Project [RC2016QN003028], and Yu Weihan Outstanding Youth Training Fund of Harbin Medical University.

REFERENCES

1. Doppler H, Bastea L, Borges S, Geiger X, Storz P. The phosphorylation status of VASP at serine 322 can be predictive for aggressiveness of invasive ductal carcinoma. Oncotarget. 2015; 6:29740–29752. doi: 10.18632/oncotarget.4965.
2. Zhu MZ, Yu XF, He XM, Feng WL, Fan JH, Li J, Xu F, Tang ZH, Zhang BN, Qiao YL, Zheng S, Yang HJ. Clinicopathological features of invasive lobular carcinoma of the breast: A nationwide multicenter study in China. J Cancer Res Ther. 2015; 11:C89–94.
3. Nascimento JC, Ferreira Sde A, Vasconcelos JL, Silva-Filho JL, Barbosa BT, Bezerra MF, Rocha CR, Beltrao EI. Fut3 role in breast invasive ductal carcinoma: Investigating its gene promoter and protein expression. Exp Mol Pathol. 2015; 99:409–415.
4. Cizeron-Clairac G, Lallemand F, Vacher S, Lidereau R, Bieche I, Callens C. MiR-190b, the highest up-regulated miRNA in ERalpha-positive compared to ERalpha-negative breast tumors, a new biomarker in breast cancers? BMC Cancer. 2015; 15:499.
5. Jezierska A, Matysiak W, Motyl T. ALCAM/CD166 protects breast cancer cells against apoptosis and autophagy. Med Sci Monit. 2006; 12:BR263–273.
6. Chen Y, Stamatoyannopoulos G, Song CZ. Down-regulation of CXCR4 by inducible small interfering RNA inhibits breast cancer cell invasion *in vitro*. Cancer Res. 2003; 63:4801–4804.
7. Zeng GF, Cai SX, Wu GJ. Up-regulation of METCAM/MUC18 promotes motility, invasion, and tumorigenesis of human breast cancer cells. BMC Cancer. 2011; 11:113.

8. Li Y, Galileo DS. Soluble L1CAM promotes breast cancer cell adhesion and migration *in vitro*, but not invasion. Cancer Cell Int. 2010; 10:34.

9. Li J, Lenferink AE, Deng Y, Collins C, Cui Q, Purisima EO, O'Connor-McCourt MD, Wang E. Identification of high-quality cancer prognostic markers and metastasis network modules. Nat Commun. 2010; 1:34.

10. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007; 3:140.

11. Simon R, Panussis S, Maurer R, Spichtin H, Glatz K, Tapia C, Mirlacher M, Rufle A, Torhorst J, Sauter G. KIT (CD117)-positive breast cancers are infrequent and lack KIT gene mutations. Clin Cancer Res. 2004; 10:178–183.

12. Zhu Y, Wang Y, Guan B, Rao Q, Wang J, Ma H, Zhang Z, Zhou X. C-kit and PDGFRA gene mutations in triple negative breast cancer. Int J Clin Exp Pathol. 2014; 7:4280–4285.

13. Gu-Trantien C, Loi S, Garaud S, Equeter C, Libin M, de Wind A, Ravoet M, Le Buanec H, Sibille C, Manfouo-Foutsop G, Veys I, Haibe-Kains B, Singhal SK, et al. CD4(+) follicular helper T cell infiltration predicts breast cancer survival. J Clin Invest. 2013; 123:2873–2892.

14. Zhang X, Ge YL, Zhang SP, Yan P, Tian RH. Downregulation of KDR expression induces apoptosis in breast cancer cells. Cell Mol Biol Lett. 2014; 19:527–541.

15. Elbauomy Elsheikh S, Green AR, Lambros MB, Turner NC, Grainge MJ, Powe D, Ellis IO, Reis-Filho JS. FGFR1 amplification in breast carcinomas: a chromogenic *in situ* hybridisation analysis. Breast Cancer Res. 2007; 9:R23.

16. Greuber EK, Smith-Pearson P, Wang J, Pendergast AM. Role of ABL family kinases in cancer: from leukaemia to solid tumours. Nat Rev Cancer. 2013; 13:559–571.

17. Li C, Li X, Miao Y, Wang Q, Jiang W, Xu C, Li J, Han J, Zhang F, Gong B, Xu L. SubpathwayMiner: a software package for flexible identification of pathways. Nucleic Acids Res. 2009; 37:e131.

18. Thien CB, Langdon WY. c-Cbl: a regulator of T cell receptor-mediated signalling. Immunol Cell Biol. 1998; 76:473–482.

19. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003; 4:P3.

20. Guo P, Yang J, Jia D, Moses MA, Auguste DT. ICAM-1-Targeted, Lcn2 siRNA-Encapsulating Liposomes are Potent Anti-angiogenic Agents for Triple Negative Breast Cancer. Theranostics. 2016; 6:1–13.

21. Schramm G, Surmann EM, Wiesberg S, Oswald M, Reinelt G, Eils R, Konig R. Analyzing the regulation of metabolic pathways in human breast cancer. BMC Med Genomics. 2010; 3:39.

22. Wang L, Li D, Fu Z, Li H, Jiang W. Association of CTLA-4 gene polymorphisms with sporadic breast cancer in Chinese Han population. BMC Cancer. 2007; 7:173.

23. Zhang J, Zhang M, Jiang W, Wang L, Fu Z, Li D, Pang D. B7-H4 gene polymorphisms are associated with sporadic breast cancer in a Chinese Han population. BMC Cancer. 2009; 9:394.

24. Fu Z, Li D, Jiang W, Wang L, Zhang J, Xu F, Pang D. Association of BTLA gene polymorphisms with the risk of malignant breast cancer in Chinese women of Heilongjiang Province. Breast Cancer Res Treat. 2010; 120:195–202.

25. Fu Z, Jiao M, Zhang M, Xu F, Yuan W, Pang D, Li D. LFA-1 gene polymorphisms are associated with the sporadic infiltrative duct breast carcinoma in Chinese Han women of Heilongjiang Province. Breast Cancer Res Treat. 2011; 127:265–271.

26. Hua Z, Li D, Xiang G, Xu F, Jie G, Fu Z, Jie Z, Da P. PD-1 polymorphisms are associated with sporadic breast cancer in Chinese Han population of Northeast China. Breast Cancer Res Treat. 2011; 129:195–201.

27. Xu F, Li D, Zhang Q, Fu Z, Zhang J, Yuan W, Chen S, Pang D. ICOS gene polymorphisms are associated with sporadic breast cancer: a case-control study. BMC Cancer. 2011; 11:392.

28. Lucci MA, Orlandi R, Triulzi T, Tagliabue E, Balsari A, Villa-Moruzzi E. Expression profile of tyrosine phosphatases in HER2 breast cancer cells and tumors. Cell Oncol. 2010; 32:361–372.

29. Ivanov SV, Panaccione A, Nonaka D, Prasad ML, Boyd KL, Brown B, Guo Y, Sewell A, Yarbrough WG. Diagnostic SOX10 gene signatures in salivary adenoid cystic and breast basal-like carcinomas. Br J Cancer. 2013; 109:444–451.

30. Montero JC, Seoane S, Ocana A, Pandiella A. Inhibition of SRC family kinases and receptor tyrosine kinases by dasatinib: possible combinations in solid tumors. Clin Cancer Res. 2011; 17:5546–5552.

31. Wu T, Wang X, Li J, Song X, Wang Y, Zhang L, Li Z, Tian J. Identification of Personalized Chemoresistance Genes in Subtypes of Basal-Like Breast Cancer Based on Functional Differences Using Pathway Analysis. PLoS One. 2015; 10:e0131183.

32. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002; 347:1999–2009.

33. Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res. 2005; 7:R953–964.

34. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011; 7:e1002240.

35. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN,

Deshpande N, Suresh S, Rashmi BP, Shanker K, Padma N, et al. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. 2004; 32:D497–501.

36. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci USA. 2005; 102:13550–13555.

37. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005; 365:671–679.

38. Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, Lindahl T, Pawitan Y, Hall P, Nordgren H, Wong JE, Liu ET, Bergh J, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. 2006; 66:10292–10301.

39. Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, Viale G, Delorenzi M, Zhang Y, d'Assignies MS, Bergh J, Lidereau R, Ellis P, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res. 2007; 13:3207–3214.

40. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003; 4:249–264.

41. Reimand J, Tooming L, Peterson H, Adler P, Vilo J. GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. Nucleic Acids Res. 2008; 36:W452–459.

42. Alizadeh AA, Gentles AJ, Alencar AJ, Liu CL, Kohrt HE, Houot R, Goldstein MJ, Zhao S, Natkunam Y, Advani RH, Gascoyne RD, Briones J, Tibshirani RJ, et al. Prediction of survival in diffuse large B-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. Blood. 2011; 118:1350–1358.

43. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37:1–13.