Identification and Validation of Immune-Related Gene Prognostic Signature for breast cancer

Bing Zhou
The Third Hospital of Nanchang
Linyi Zhang
Nanchang University
Qinyu Wang
The Third Hospital of Nanchang
Changqin Pu
Nanchang University
Sixuan Guo
Medical College of Nanchang University
Shuhui Lai
Medical College of Nanchang University
Heming Zhang
Nanchang University
Wenwei Li
The Third Hospital of Nanchang
Zhibing Zhou
The Third Hospital of Nanchang
Yuexia Chen
The Third Hospital of Nanchang
Yao Zhou
The Third Hospital of Nanchang
Liping Zeng (356700210001@email.ncu.edu.cn)
Nanchang University

Article

Keywords: breast cancer, immune-relevant genes, prognosis, overall survival

Posted Date: May 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-386385/v2

Abstract

Background

Although the outcome of breast cancer patients has been improved by advances in early detection, diagnosis and treatment, prognostic assessment still faces challenges due to the heterogeneity of the disease. The accumulated data indicate that there is a clear correlation between the tumor immune microenvironment and clinical outcomes.

Methods

We screened prognostic immune-relevant gene pairs (IRGPs) via univariate Cox regression analysis in the Cancer Genome Atlas (TCGA) cohort. Then, TCGA cohort were further divided into a training set (n = 755) and internal validation sets (n = 320). Least absolute shrinkage and selection operator (LASSO) Cox regression analysis was used to constituted the IRGP prognostic signature in training set. And the IRGP prognostic signature was validated in the internal TCGA validation cohort and external validation cohorts GSE20685 (n = 327), GSE42568 (n = 104), and GSE20711 (n = 88) from Gene Expression Omnibus (GEO) database. We estimated the immune cell infiltration in tumor microenvironment via CIBERSORT and ESTIMATE. We used the TIDE algorithm and stemness indices to evaluate the potential of IRGP signature as an indicator of response to immunotherapy. We used gene set enrichment analysis (GSEA) to elucidate the biological functions of the IRGP prognostic signature.

Results

We generated a IRGP prognostic signature consisting of 16 IRGPs. Subsequently, the 16 IRGPs grouped breast cancer patients into high- and low-risk groups. Survival analysis indicated that the IRGP signature possessed an independent prognostic value. The low-IRGP group exhibited a higher level of immune cell infiltration, higher expression of immune checkpoint molecules, lower tumor stemness indices, and was much more sensitive to immunotherapy. The functional enrichment analysis indicated that low IRGP value was correlated with biological processes related to immune.

Conclusion

The 16-IRGP prognostic signature was developed to provide new insights for the identification of high-risk breast cancer and the evaluation of the possibility of immunotherapy in personalized breast cancer treatment.

1. Introduction

Breast cancer is the most frequently diagnosed cancer in women and ranks second among causes for cancer related death in women [1]. Previous research account for the high mortality of breast cancer, and the metastasis of cancer cell to vital organs is explored as principle cause [2]. In the United States, the number of newly diagnosed cases exceeds 268,000 each year, resulting in more than 41,000 deaths and accounting for an estimated 15% of cancer-related deaths in women [3]. Accurate prediction of prognosis has been proved to greatly decrease the mortality of breast cancer patients [4], and TNM staging correlation has been identified as the core prognostic factor of breast cancer, with higher staging predicting worse prognosis [5, 6]. However, due to the high heterogeneity of breast cancer, different stages of TNM cannot accurately predict the prognosis of patients with similar clinical features. Therefore, prognostic prediction models constructed based on other prognostic factors are urgently needed.

Based on the immune-related genes, like immune checkpoints, immunotherapy is revolutionizing the therapeutic strategies of breast cancer. Recently, several immune checkpoint inhibitors (ICIs) were proved to relieve tumor-mediated immune-cell suppression for breast cancer patients, including programmed cell death-1 (PD-1/PDCD1), programmed cell death ligand-1 (PD-L1/CD274), cytotoxic T-lymphocyte antigen-4 (CTLA-4), T-cell immunoglobulin and ITIM domain (TIGIT), VISTA/VSIR and B7-H3 (CD276)[7, 8]. Moreover, the most clinically developed immune checkpoint proteins are PDCD1, CD274, and CTLA-4 [8]. Nevertheless, only one-third of patients respond to ICI in most cancer types [9]. Thus, it is necessary to investigate predictive biomarkers to indicate ICI responsiveness and prognosis. Since the construction of immune gene pairs allows direct comparison of the expression of two genes, without the need for comparison or correction between samples from various sources, we can obtained prediction results with high accuracy [10]. However, a breast cancer prognostic model based on immune gene pairs was still lacked, which could effectively guide the clinical treatment of breast cancer patients.

In this study, we established a 16-IRGP signature to predict the individual prognostic characteristics of breast cancer by univariate Cox regression analysis and LASSO model. In order to validate of the signature, we investigated its accuracy and efficiency in determining the prognosis of breast cancer patients with KM curves, ROC curve analysis, univariate and multivariate Cox regression analysis. The findings of this study showed and proved that the 16-IRGP signature have the potential to apply in the clinical prognosis of breast cancer patients, and informing the immunotherapy.

2. Materials And Methods

2.1 Transcriptome data acquisition and pre-processing

The breast cancer transcriptome data and related clinical characteristics were downloaded from The Cancer Genome Atlas (TCGA) database [11]. In total 1075 primary breast cancer samples with overall survival information and gene expression profiles were extracted from TCGA. Then, all selected patients were further randomly separated into training and internal validation groups (7:3). The models are identified and evaluated by “caret” package with its

“createDataPartition” function [12]. In addition, the GSE20685 (n = 327)[13], GSE42568 (n = 104) [14]and GSE20711 (n = 88) [15]datasets were downloaded from the Gene Expression Omnibus (GEO) database [16] (http://www.ncbi.nlm.nih.gov/geo/) as external validation cohorts. The GEO data were preprocessed as follows: (1) Only primary breast cancer tissue samples with overall survival information were retained. (2) the overall survival time was converted from year/ day to month. All selected clinical characteristics of breast cancer patients were listed in Table 1. A single-cell RNA sequencing (scRNA-seq) dataset GSE169246 [17] was used to detect the expression of gene panel in breast cancer tumor microenvironment. The cell type annotation was also obtained from the article attachment. The breast cancer tissue before treatment was extracted in our study. All sequencing data and related clinical information were publicly available. The data acquisition and pre-processing procedure were carried out in accordance with the publication guidelines provided by TCGA and GEO database.

2.2 Prognostic immune-relevant gene pairs identification

A total of 2483 immune-relevant genes (IRGs) were obtained from immport database (https://www.immport.org/home) (accessed on September 17, 2021). IRGs with relatively high variation (median absolute deviation > 0.5) were extracted from TCGA and GEO datasets. The common IRGs were used to construct immune-relevant gene pairs (IRGPs). Each IRGP was calculated by comparing the expression levels of genes in a particular sample, and if IRG 1 > IRG 2, then IRGP equals to 1, otherwise IRGP equals to 0. Besides, some IRGPs with score of 1 or 0 in over 80% of the samples were excluded to avoid the biases. The prognostic IRGPs were identified via univariate Cox regression analysis in the TCGA cohort, where IRGPs with P-value < 0.001 were selected for further analysis.

2.3 The IRGP prognostic signature construction and validation

TCGA samples were randomly separated into training (755 patients) and internal validation (320 patients) cohorts based on tumor stage. Least absolute shrinkage and selection operator (LASSO) is a biased estimation tool for complex collinearity data, which can select variables and estimate parameters at the same time, and solve the problem of multicollinearity in regression analysis [18]. Thus, LASSO Cox regression analysis was applied to reduce dimensionality of IRGPs by R package glmnet. IRGPs represented by optimal values of the penalty parameter λ, which were determined by ten-fold cross-validations, constituted the IRGP prognostic signature in this study. On the basis of IRGP prognostic signature, the risk score for each breast cancer sample was calculated according to the following formula: risk score = $\text{expressionpair}_1 \times \beta\text{pair}_1 + \text{expressionpair}_2 \times \beta\text{pair}_2 + \cdots + \text{expressionpair}_x \times \beta\text{pair}_x$, (x = the number of IRGPs; β = coefficient value for each IRGP). All breast cancer samples were classified into high- and low-risk groups based on the optimal cut-off of risk score determined by “cutp” function of survMisc package in R [19]. The cut point was chosen by hazard function with the maximal sensitivity and specificity for survival rate, based on the best P level less than 0.05 [20]. Kaplan-Meier (KM) survival curves were applied to calculate the overall survival (OS), disease-free survival (DFS) and distant metastasis-free survival (DMFS) differences between two groups, and the statistical significance was obtained by log-rank test. Survival predictive accuracy of prognostic models was assessed based on a time-dependent receiver operating characteristic curve (ROC) analysis. Then, the multivariate Cox regression model was used to evaluate whether the prognostic value of IRGP prognostic signature was independent of clinical characteristics. Furthermore, the comparison between IRGP prognostic signature and clinical features was performed by forest plots to determine the effectiveness of the prognostic value.

2.4 Clinical utility of IRGP prognostic signature

A composite nomogram was developed based on the IRGP prognostic signature and available clinical factors to predict the OS of breast cancer patients with the rms package in R software [21]. Calculate the concordance index (C-index) to evaluate the discriminative ability of the nomogram. Besides, a calibration curve was drawn to compare the predicted probability and actual probability of OS. Each component of the nomogram is given points, and their sum represents the total points of a patient has obtained.

2.5 Acquisition of somatic mutation data

A total of 4 subtypes of data files from the TCGA database were selected, from which the “Masked Somatic Mutation” data was processed by VarScan software. Somatic mutation data was processed into Mutation Annotation Format (MAF) file and visualized by “maftools” [22] R package, which can provide multiple analyzing models. There was no ethical conflict to declare because all the data in this research were from public databases.

2.6 Estimation of immune infiltration

Immune infiltration assessment was performed by cibersort method to quantify the absolute abundance of 22 immune cell populations in heterogeneous tissues from transcriptome data [23]. The cibersort method was applied by CIBERSORT package to convert mRNA data into the levels of infiltrating immune cells. Furthermore, the "ESTIMATE" R package [24] was used to estimate stromal scores, immune scores, and tumor purity.

2.7 Immunotherapeutic response prediction

Tumour Immune Dysfunction and Exclusion (TIDE) algorithm was used to estimate the ICI response of breast cancer patients [25]. A higher TIDE score is associated with worse ICI response and worse survival rate.

Cancer stem cells (CSCs) play a crucial role in the development, recurrence, metastasis, and resistance of cancers [26, 27]. Malta et al. used one-class logistic regression (OCLR) machine learning algorithm to calculate the stemness index, with mRNA expression-based stemness index (mRNAsi) and epigenetic regulation-based stemness index (EREG-mRNAsi)[28]. The stemness index value range from 0 to 1, and higher value is associated with a reduced PD-L1 expression and lower sensitivity of immune treatments.

2.8 Functional enrichment analysis

Gene set enrichment analysis (GSEA) was conducted to investigate the association between potential biological phenotypes and IRGP prognostic signature. The gene sets in the "Molecular Signatures Database" of "c5.go.v7.2.symbols.gmt" and "c2.cp.kegg.v7.2.symbols.gmt" were downloaded as the reference gene set. The significance threshold was set at adjusted $P < 0.05$.

2.9 Statistical analysis

We conducted Cox regression analysis and Kaplan-Meier curves analysis with the log-rank test by survival packages. The continuous variables conforming to normal distribution were analyzed via unpaired Student's t-test, and the statistical significance of differences between non-normally distributed variables were estimated using Wilcoxon test. For comparisons of three or more groups, Kruskal-Wallis and one-way ANOVA tests were utilized as non-parametric and parametric methods, respectively. Pearson correlation was used to explore the correlation between the IRGP prognostic signature and estimate score or immune checkpoint molecules. All statistical analyses were performed in R studio (version 4.0.3), and $P < 0.05$ was considered statistically significant.

3. Results

3.1 Prognostic immune-relevant gene pairs identification

The workflow chart for data collection and analysis was shown in Fig. 1. We obtained 212 common high variable IRGs overlapped between TCGA cohort, GSE20685 cohort, GSE42568 cohort, and GSE20711 cohort (Fig. 2A). Then we acquired 3075 common IRGPs after removing IRGPs with scores of 0 or 1 in over 80% of the samples in all cohorts (Fig. 2B). Univariate Cox regression analysis was performed for the 3075 IRGPs, of which 35 IRGPs showed significant prognostic potential (P-value < 0.001). Based on the cutting off value (HR = 1), we identified 14 risky IRGPs and 21 protective IRGPs (Table S1).

3.2 Construction and validation of the IRGP prognostic signature

The IRGPs in the risk model were selected by applying LASSO Cox regression analysis. All samples in TCGA cohort were regrouped into training and internal validation groups randomly for prognostic analyses. No significant difference was observed when comparing patient characteristics between the two groups (Table S2). The IRGP prognostic signature constructed by 16 IRGPs consisting of 28 unique IRGs was generated through the LASSO model (Fig. S1A-B). Most of these 28 IRGs encoded molecules related to antimicrobials, cytokines, and cytokine receptors (Table 2).

Furthermore, we uploaded the 28 unique IRGs list to Metascape (https://metascape.org/gp/index.html) to annotate biological function of these genes. The GO enrichment results shown that these genes were correlated with cancer immune-related biological functions, including cell chemotaxis, leukocyte migration, myeloid leukocyte migration and neutrophil migration (Fig. S2A, Table S3). Similarly, the KEGG pathway enrichment results indicated that these genes were associated with cytokine-cytokine receptor interaction, chemokine signaling pathway and several classical signal pathways (Fig. S2B).

The 16 IRGPs grouped breast cancer patients into high- and low-risk groups based on the optimal cut-off point determined by "cutp" function from survMisc package. Kaplan-Meier curves indicated that patients in low-risk group have significantly longer survival time than that in high-risk group in training cohort (P-value = 6.12E-13, HR = 3.90, 95%CI = 2.65–5.74; Fig. 3A), internal validation cohort (P-value = 8.15E-6, HR = 3.70, 95%CI = 1.74–7.89; Fig. 3B), GSE20685 validation cohort (P-value = 1.83E-9, HR = 3.44, 95%CI = 2.02–5.86; Fig. 3C), GSE42568 validation cohort (P-value = 1.80E-3, HR = 4.51, 95%CI = 2.27–8.96; Fig. 3D), and GSE20711 validation cohort (P-value = 2.72E-2, HR = 2.41, 95%CI = 1.09–5.33; Fig. 3E). As shown in Fig. 3F, the IRGP prognostic signature possessed a high specificity and sensitivity in the training cohort, with the area under the curve (AUC) of 0.849, 0.827, and 0.781 for 1-, 3- and 5-year OS prediction. The AUC of the IRGP prognostic signature for 1-, 3- and 5-year OS prediction were 0.688, 0.796 and 0.714 in internal validation cohort (Fig. 3G), 0.584, 0.752 and 0.748 in GSE20685 validation cohort (Fig. 3H), 0.538, 0.620 and 0.689 in GSE42568 validation cohort (Fig. 3I), and 0.920, 0.542 and 0.657 in GSE20711 validation cohort (Fig. 3J). In addition, the patients in the high-risk group had a shorter disease-free survival (DFS) and distant metastasis-free survival (DMFS) time compared with low-risk group in training cohort, internal validation cohort, GSE20685, GSE42568 validation cohort, and GSE20711 validation cohort (Fig. S3A-E). And the prognostic value of IRGP signature for DFS and DMFS prediction was also evaluated in both training and validation cohorts (Fig. S3F-J). The above analyses presented that the IRGP prognostic signature could has a good predictive value in OS, DFS and DMFS prediction.

Univariate Cox regression analysis showed a statistical significance for IRGP prognostic signature in training cohort (P-value = 2.63E-11, HR = 4.04, 95%CI = 2.68–6.09), internal validation cohort (P-value = 3.28E-5, HR = 3.80, 95%CI = 2.03–7.15), GSE20685 validation cohort (P-value = 1.63E-8, HR = 3.50, 95%CI = 2.27–5.40), GSE42568 validation cohort (P-value = 4.45E-3, HR = 4.54, 95%CI = 1.60-12.88), and GSE20711 validation cohort (P-value = 3.23E-2, HR = 2.47, 95%CI = 1.08–5.66) (Table 3). In order to explore the independence of IRGP prognostic signature in survival prediction, a multivariate Cox regression analysis was performed in the TCGA training cohort and internal validation cohort, including IRGP prognostic signature, age, M stage, N stage, T stage, tumor stage, fraction genome altered, immune subtype, and breast cancer subtypes. Whereas only age, M stage, N stage and T stage were available for GSE20685 validation cohort and age as well as tumor grade were available for GSE42568 and GSE20711 validation cohort, we integrated IRGP prognostic signature and these clinical features in multivariate cox regression analysis. The prognostic values of IRGP prognostic signature was significant compared with other clinical characteristics in training cohort (P-value = 2.02E-6, HR = 3.53, 95%CI = 2.10–5.94), internal validation cohort (P-value = 7.21E-05, HR = 8.88, 95%CI = 3.02–26.10), GSE20685 validation cohort (P-value = 3.40E-6, HR = 2.98, 95%CI = 1.88–4.72), GSE42568 validation cohort (P-value = 1.22E-2, HR = 3.93, 95%CI = 1.35–11.48), and GSE20711 validation cohort (P-value = 3.68E-2, HR = 2.47, 95%CI = 1.06–5.78) (Table 3). It indicated that IRGP prognostic signature was a strong independent risk factor. The predictive power of the IRGP prognostic signature was further tested in various subgroups stratified by TNM stage, age, fraction genome altered, ER status, HER2 status, PR status, immune subtype, and breast cancer subtype in the TCGA entire cohort. The forest plot shown that higher risk score was correlated with worse prognosis in almost all subgroups (Fig. 4). Collectively, the results indicated that the predictive ability of the IRGP prognostic signature was independent of other clinical parameters and was predictive of OS of breast cancer patients.

3.3 Increasing IRGP risk score is associated with clinical characteristics of breast cancer patients

The relationship between IRGP prognostic signature and clinical characteristics of breast cancer patients was further investigated in the entire TCGA cohorts. In terms of clinical features, increasing IRGP risk score was correlated with more advanced tumor stage (P = 3.10E-03; Fig. 5A), M stage (P = 1.20E-03; Fig. 5B), N stage (P = 9.50E-02, without statistical significance; Fig. 5C), T stage (P = 1.40E-05; Fig. 5D), and patients who had died (P-value = 1.10E-11; Fig. 5E) or progression (P-value = 2.50E-3; Fig. 5F) due to the disease. Furthermore, age influenced the value of IRGP prognostic signature (P-value = 8.20E-6; Fig. 5G). In addition, it is explored that more genomic changes (P-value < 2.20E-16; Fig. 5H) and triple negative breast cancer (TNBC) patients (P-value = 1.20E-8; Fig. 5I) relative to higher values of IRGP prognostic signature. In terms of molecular characteristics, patients in molecular subtypes C4 (P-value < 2.20 E-16; Fig. 5J) and HER2-enriched as well as Basal-like (P-value < 2.20 E-16; Fig. 5K) exhibited significantly higher values of IRGP prognostic signature than others. Above all, increasing IRGP risk score is associated with worse clinical status or molecular subtype.

3.4 Establishment of a nomogram predicting OS in breast cancer patients

In order to develop a clinically relevant quantitative method to predict the mortality rate of patients, we constructed a nomogram integrating IRGP prognostic signature derived score and clinical information to predict survival probability of breast cancer patients in TCGA entire cohort (Fig. 6A). The concordance index (C-index) of the nomogram was 0.820 in TCGA cohort, which indicating a good discriminatory ability of the nomogram. The calibration plots showed that the derived nomograms performed well compared to the performance of an ideal model at 1-year, 3-years and 5-years (Fig. 6B-D). Decision curve analysis shows that the clinical utility of nomograms greatly exceeds TNM staging system (Fig. 6E). The nomogram analysis indicated that the IRGP prognostic signature is the most important factor to predict survival probability of breast cancer patients, which has the greatest contribution to survival with largest regression coefficient.

3.5 Potential of IRGP prognostic signature as an indicator of response to immunotherapy

It has been reported that the infiltration of immune cells is associated with the prognosis of breast cancer patients. The expression levels of genes in IRGP signature in differential immune cell types were investigated in scRNA-seq dataset GSE169246. The results showed that CMTM8, CXCL13, TNFRSF12A, ICAM2, PIK3R3, IL7R, PDGFD were higher expressed in T cells (Fig. S4). The immune cell infiltration analysis shown that high-risk group patients was obviously infiltrated by M0 macrophages and M2 macrophages, while low-risk group patients showed an obvious increase trend in infiltration by naive B cell, memory resting CD4 T cells and CD8 T cells (Fig. 7A). Then, we estimated the tumor microenvironment (TME) in the two risk groups. We found that the patients in high-risk group had a lower stromal score (Fig. 7B) and lower immune score (Fig. 7C). And increasing IRGP risk score was significantly positively correlated with tumor purity (Fig. 7D).

Then, we investigate the relationship between IRGP prognostic signature and ICI genes, including PDCD1, CD274, CTLA-4, TIGIT, VSIR and CD276. As shown in Fig. S5, the IRGP prognostic signature was markedly negatively related with PDCD1, CTLA-4, TIGIT, VSIR (Fig. S5A-D), and positively correlated with CD276 (Fig. S5E). The expression of CD274 was negatively related with IRGP prognostic signature but without statistical significance (Fig. S5F). Moreover, patients in low-risk group tend to express high expression level of PDCD1 (P-value = 2.30E-15; Fig. 8A), CTLA-4 (P-value = 2.50E-6; Fig. 8B), CD274 (P-value = 5.80E-4; Fig. 8C) compared with patients in high-risk group. In addition, there were statistical significance between the risk group and expression level of TIGIT, VSIR, CD276 (Fig. S5G-I). This result indicated that IRGP prognostic signature was associated with the expression levels of ICI genes (PDCD1, CTLA-4, TIGIT, VSIR and CD276).

Furthermore, we used the TIDE score and stemness indices to explore whether the IRGP prognostic signature could reflect the immunotherapeutic benefit in breast cancer patients. Interestingly, we found that the TIDE score was significantly lower in low-risk group and low-risk patients may be more likely to respond to immunotherapy compared with high-risk patients (Fig. 8D-E). As excepted, we observed that the breast cancer patients in low-risk group had lower stemness indices value than high-risk group (Fig. 8F-G). Collectively, these results indicated that low-risk patients were more likely to respond to immunotherapy.

3.6 Differential somatic mutation burden landscape between two risk score levels

We visualized the results of 390 high-risk and 565 low-risk breast cancer patients using the " maftools" package, based on mutation data with VCF format. The waterfall plot shows the mutation information for each gene in each sample, and the different types of mutations are annotated by different colors at the bottom (Fig. 9A-B). The result exhibited the top 10 mutated genes with ranked percentages, where TP53 shown the highest mutation rate in high risk breast cancer patients (39%) compared with low risk breast cancer patients (19%). Besides, the number of altered bases in each sample of high risk breast cancer patients was higher than low risk breast cancer patients (Fig. 9C-D).

3.7 Identification of IRGP prognostic signature related biological pathways and processes

GSEA was performed to elucidate the biological functions of the IRGP prognostic signature. According to the GO analysis results (Fig. 10A, Table S4), these genes highly expressed in the low-IRGP risk score group were associated with immune-related gene set, including T cell receptor complex, B cell receptor signaling pathway, phagocytosis recognition, and immunoglobulin receptor binding. While the high-IRGP risk score group related genes showed significant enrichment in DNA transcription-related gene set, such as the mRNA cis splicing via spliceosome, pre mRNA binding, mRNA splice site selection, and mRNA 5’ splice site recognition. The KEGG pathway results also shown that pathways involved in immune response were activated in low-IRGP risk score group, including T cell receptor signaling pathway, natural killer cell mediated cytotoxicity, and antigen processing and presentation (Fig. 10B).

4. Discussion

Immunity is closely linked to tumors. For example, Li B et al inferred the abundance of 6 immune cell types (B cells, CD4 T cells, CD8 T cells, neutrophils, macrophages, and dendritic cells), and discovered significant associations between cell abundance and prognosis in 23 cancers[29]. In addition to prolonging the survival of patients, CD8 T cells may also play an important role in preventing tumor recurrence, such as in melanoma, colorectal cancer, and cervical

cancer [29]. In addition, increasing evidence points to the importance of biomarkers (especially genes) in determining cancer outcomes, which provides new opportunities for integrating this information into treatment algorithms [10]. Many previous studies have shown that immune-related genes can be used as prognostic indicators for breast cancer[30–32]. However, researches were often limited by the singularity of genes and differences between samples. Therefore, new methods are urgently needed to improve the accuracy of breast cancer prediction.

Immune-related gene pairs were widely used in tumor analysis, and great progress has been made in many cancers, such as melanoma, ovarian cancer and pancreatic cancer [33–35]. However, more research on breast cancer was required. Our study focused on predicting breast cancer survival using the 16-IRGP signature combined with clinical information. We hoped that our findings would offer a new insight on clinical decision-making and prognostic monitoring in breast cancer patients.

The aim of this study was to use bioinformatics tools and databases to demonstrate the predictive prognostic potential of 16-IRGP signature in breast cancer patients. The 16-IRGP signature could predict breast cancer patient prognosis including OS, DFS and DMFS. The higher IRGP value was correlated with worse oncological outcomes. And the IRGP prognostic signature was an independent risk factor for OS prediction of breast cancer patients. Nomogram analysis shows that the clinical effectiveness of the IRGP prognostic signature is significantly higher than that of the TNM stage.

GSEA showed that low-IRGP related genes were related to immune related gene sets, including T cell receptor complex, B cell receptor signaling pathway, phagocytosis recognition, and immunoglobulin receptor binding, while high-IRGP related genes were significantly enriched in the DNA transcription-related gene set, including mRNA cis splicing via spliceosome, pre mRNA binding, mRNA splice site selection, and mRNA 5’ splice site recognition. At the same time, we also analyzed the somatic mutations in patients with breast cancer, and we found that TP53 is highest mutation rate in high risk breast cancer patients (39%) compared with low risk breast cancer patients (19%). The TP53 protein is a DNA-bound transcription factor that has the potential to bind to hundreds of different promoter elements in the genome to regulate gene expression[36]. Tumors with TP53 mutations are usually characterized by poor differentiation, increased invasiveness and high metastatic potential, which are associated with poor prognosis [37, 38]. Combined GSEA results with somatic mutation analysis, the mutations of TP53 affect the prognosis of patients with breast cancer to a great extent. Therefore, we can determine the reliability of our functional enrichment results [36, 39] and our IRGPs may play a role in the prognosis of breast cancer.

At present, several immune-related therapies for tumors have achieved good results in clinical trials. For instance, by blocking macrophage chemokines (such as CXCL12) and preventing macrophages from entering tumors, the development and proliferation of cancer cells can be inhibited[40]. Although the focus of this research is not on the mechanism of immune cells, it still provides strong evidence that tumor-related immune genes may become potential targets for cancer treatment. Our research focuses on immune-related genes and uses strict standard-level screening to obtain genes that may be prognostic targets for breast cancer [41]. And our results demonstrated that low-risk group patients were obviously infiltrated by naive B cell, memory resting CD4 T cells and CD8 T cells. The IRGP prognostic signature may have a predictive ability for breast cancer immunotherapy.

The current study has some limitations. First, due to the retrospective nature of this study, the patient population was heterogeneous. Secondly, prospective clinical trials are needed to validate our findings. In conclusion, the IRGPS gene map is a powerful tool for predicting breast cancer survival and guiding treatment.

Abbreviations

ICIs: immune checkpoint inhibitors

PD-1/PDCD1: programmed cell death-1

PD-L1/CD274: programmed cell death ligand-1

CTLA-4: cytotoxic T-lymphocyte antigen-4

TIGIT: T-cell immunoglobulin and ITIM domain ()

scRNA-seq: A single-cell RNA sequencing

IRGs: immune-relevant genes

IRGPs: immune-relevant gene pairs

LASSO: Least absolute shrinkage and selection operator

KM: Kaplan-Meier

OS: overall survival

DFS: disease-free survival

DMFS: distant metastasis-free survival

C-index: concordance index

MAF: Mutation Annotation Format

TIDE: Tumour Immune Dysfunction and Exclusion

CSCs: Cancer stem cells

OCLR: one-class logistic regression

GSEA: Gene set enrichment analysis

AUC: area under the curve

TME: the tumor microenvironment

Declarations

Ethics approval and consent to participate

Not applicable.

Data availability statement

All data generated or analysed during this study are included in this published article (and its Supplementary Information files). All sequencing data and related clinical information were publicly available.

Competing interests

The authors declare no conflicts of interest.

Funding

No funding was received.

Authors' contributions

All authors contributed substantially to the preparation of this manuscript. B Zhou, S. X. Guo, S. H. Lai, L. P. Zeng and Y Zhou were responsible for protocol design. B Zhou, L. Y. Zhang, S. X. Guo, C. Q. Pu, S. H. Lai and L. P, Zeng were responsible for data acquisition. Q. Y. Wang, W. W. Li, Z. B. Zhou, Y. X. Chen and Y. Zhou were responsible for data analysis. All authors interpreted the data. L. Y. Zhang, S. X. Guo, C. Q. Pu, H. M. Zhang and L. P. Zeng wrote the manuscript. All authors revised and finalized the manuscript.

Acknowledgements

Not applicable.

Patient consent for publication

Not applicable.

References

1. Libson S, Lippman M: **A review of clinical aspects of breast cancer**. Int Rev Psychiatry 2014, **26**(1):4–15.
2. Maughan KL, Lutterbie MA, Ham PS: **Treatment of breast cancer**. Am Fam Physician 2010, **81**(11):1339–1346.
3. Liang Y, Zhang H, Song X, Yang Q: **Metastatic heterogeneity of breast cancer: Molecular mechanism and potential therapeutic targets**. Semin Cancer Biol 2020, **60**:14–27.
4. Li G, Hu J, Hu G: **Biomarker Studies in Early Detection and Prognosis of Breast Cancer**. Adv Exp Med Biol 2017, **1026**:27–39.
5. Edge SB, Compton CC: **The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM**. Ann Surg Oncol 2010, **17**(6):1471–1474.
6. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, Meyer L, Gress DM, Byrd DR, Winchester DP: **The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging**. CA Cancer J Clin 2017, **67**(2):93–99.
7. Workman CJ, Dugger KJ, Vignali DA: **Cutting edge: molecular analysis of the negative regulatory function of lymphocyte activation gene-3**. J Immunol 2002, **169**(10):5392–5395.
8. Gaynor N, Crown J, Collins DM: **Immune checkpoint inhibitors: Key trials and an emerging role in breast cancer**. Semin Cancer Biol 2022, **79**:44–57.
9. Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A: **Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy**. Cell 2017, **168**(4):707–723.
10. Li B, Cui Y, Diehn M, Li R: **Development and Validation of an Individualized Immune Prognostic Signature in Early-Stage Nonsquamous Non-Small Cell Lung Cancer**. JAMA Oncol 2017, **3**(11):1529–1537.

11. Balbin OA, Malik R, Dhanasekaran SM, Prensner JR, Cao X, Wu YM, Robinson D, Wang R, Chen G, Beer DG *et al*: **The landscape of antisense gene expression in human cancers**. Genome Res 2015, **25**(7):1068–1079.
12. Kuhn M: **Caret**: **Classification and regression training**. 2013.
13. Kao KJ, Chang KM, Hsu HC, Huang AT: **Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization**. BMC Cancer 2011, **11**:143.
14. Clarke C, Madden SF, Doolan P, Aherne ST, Joyce H, O'Driscoll L, Gallagher WM, Hennessy BT, Moriarty M, Crown J *et al*: **Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis**. Carcinogenesis 2013, **34**(10):2300–2308.
15. Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, Michiels S, Volkmar M, Deplus R, Luciani J *et al*: **DNA methylation profiling reveals a predominant immune component in breast cancers**. EMBO Mol Med 2011, **3**(12):726–741.
16. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M *et al*: **NCBI GEO: archive for functional genomics data sets–update**. Nucleic Acids Res 2013, **41**(Database issue):D991-995.
17. Zhang Y, Chen H, Mo H, Hu X, Gao R, Zhao Y, Liu B, Niu L, Sun X, Yu X *et al*: **Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer**. Cancer Cell 2021, **39**(12):1578–1593.e1578.
18. Duan J, Soussen C, Brie D, Idier J, Wan M, Wang YP: **Generalized LASSO with under-determined regularization matrices**. Signal Processing 2016, **127**:239–246.
19. Xu Q, Yang Q, Zhou Y, Yang B, Jiang R, Ai Z, Teng Y: **A long noncoding RNAs signature to improve survival prediction in endometrioid endometrial cancer**. J Cell Biochem 2018.
20. Contal C, O'Quigley J: **An application of changepoint methods in studying the effect of age on survival in breast cancer**. Computational Statistics & Data Analysis 1999, **30**(3):253–270.
21. Jr F: **rms**: **Regression Modeling Strategies**. 2015.
22. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP: **Maftools: efficient and comprehensive analysis of somatic variants in cancer**. Genome Res 2018, **28**(11):1747–1756.
23. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA: **Robust enumeration of cell subsets from tissue expression profiles**. Nat Methods 2015, **12**(5):453–457.
24. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA *et al*: **Inferring tumour purity and stromal and immune cell admixture from expression data**. Nat Commun 2013, **4**:2612.
25. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, Li Z, Traugh N, Bu X, Li B *et al*: **Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response**. Nat Med 2018, **24**(10):1550–1558.
26. Dawood S, Austin L, Cristofanilli M: **Cancer stem cells: implications for cancer therapy**. Oncology (Williston Park) 2014, **28**(12):1101–1107, 1110.
27. Nassar D, Blanpain C: **Cancer Stem Cells: Basic Concepts and Therapeutic Implications**. Annu Rev Pathol 2016, **11**:47–76.
28. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O *et al*: **Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation**. Cell 2018, **173**(2):338–354.e315.
29. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S *et al*: **Comprehensive analyses of tumor immunity: implications for cancer immunotherapy**. Genome Biol 2016, **17**(1):174.
30. Xie P, Ma Y, Yu S, An R, He J, Zhang H: **Development of an Immune-Related Prognostic Signature in Breast Cancer**. Front Genet 2019, **10**:1390.
31. Xu H, Wang G, Zhu L, Liu H, Li B: **Eight immune-related genes predict survival outcomes and immune characteristics in breast cancer**. Aging (Albany NY) 2020, **12**(16):16491–16513.
32. Chen L, Dong Y, Pan Y, Zhang Y, Liu P, Wang J, Chen C, Lu J, Yu Y, Deng R: **Identification and development of an independent immune-related genes prognostic model for breast cancer**. BMC Cancer 2021, **21**(1):329.
33. Huang RZ, Mao M, Zheng J, Liang HQ, Liu FL, Zhou GY, Huang YQ, Zeng FY, Li X: **Development of an immune-related gene pairs index for the prognosis analysis of metastatic melanoma**. Sci Rep 2021, **11**(1):1253.
34. Zhang B, Nie X, Miao X, Wang S, Li J, Wang S: **Development and verification of an immune-related gene pairs prognostic signature in ovarian cancer**. J Cell Mol Med 2021, **25**(6):2918–2930.
35. Li Y, Yao P, Zhao K, Ye Z, Zhang H, Cao J, Zhang S, Xing C: **Individualized prognostic signature for pancreatic carcinoma validated by integrating immune-related gene pairs (IRGPs)**. Bioengineered 2021, **12**(1):88–95.
36. Silwal-Pandit L, Langerod A, Borresen-Dale AL: **TP53 Mutations in Breast and Ovarian Cancer**. Cold Spring Harb Perspect Med 2017, **7**(1):a026252.
37. Olivier M, Hollstein M, Hainaut P: **TP53 mutations in human cancers: origins, consequences, and clinical use**. Cold Spring Harb Perspect Biol 2010, **2**(1):a001008.
38. Silwal-Pandit L, Vollan HK, Chin SF, Rueda OM, McKinney S, Osako T, Quigley DA, Kristensen VN, Aparicio S, Børresen-Dale AL *et al*: **TP53 mutation spectrum in breast cancer is subtype specific and has distinct prognostic relevance**. Clin Cancer Res 2014, **20**(13):3569–3580.
39. Mosele F, Stefanovska B, Lusque A, Tran Dien A, Garberis I, Droin N, Le Tourneau C, Sablin MP, Lacroix L, Enrico D *et al*: **Outcome and molecular landscape of patients with PIK3CA-mutated metastatic breast cancer**. Ann Oncol 2020, **31**(3):377–386.
40. Gholamin S, Mitra SS, Feroze AH, Liu J, Kahn SA, Zhang M, Esparza R, Richard C, Ramaswamy V, Remke M *et al*: **Disrupting the CD47-SIRPalpha anti-phagocytic axis by a humanized anti-CD47 antibody is an efficacious treatment for malignant pediatric brain tumors**. Sci Transl Med 2017, **9**(381):eaaf2968.

41. Li J, Liu C, Chen Y, Gao C, Wang M, Ma X, Zhang W, Zhuang J, Yao Y, Sun C: **Tumor Characterization in Breast Cancer Identifies Immune-Relevant Gene Signatures Associated With Prognosis**. Front Genet 2019, **10**:1119.

Tables

Table 1
Clinical characteristics for breast cancer patients.

Variables	Training cohort (755 patients)	Internal validation cohort (320 patients)	GSE20685 validation cohort (327 patients)	GSE42568 validation cohort (104 patients)	GSE20711 validation cohort (88 patients)
OS times (days) (mean, range)	1222 (1-8481)	1293 (7-8432)	2880 (146–5146)	1898 (138–3026)	2525 (318–5139)
DFS times (days) (mean, range)	1103 (1-7892)	1189 (7-8432)	1546 (73-4088)	1633 (84-3026)	2050 (0-5139)
PFS times (days) (mean, range)	/	/	1058 (0-3869)	/	/
OS status (n, %)					
Alive	644 (85.30)	280 (87.50)	244 (74.62)	69 (66.35)	63 (71.59)
Dead	111 (14.70)	40 (12.50)	83 (25.38)	35 (33.65)	25 (28.41)
DFS status (n, %)					
Non-relapse	605 (80.13)	266 (83.13)	0 (0)	56 (53.85)	49 (55.68)
Relapse	85 (11.26)	28 (8.75)	25 (7.65)	48 (46.15)	39 (44.32)
Unknown	65 (8.61)	26 (8.12)	302 (92.35)	/	/
PFS status (n, %)					
Non-progress	/	/	0 (0)	/	
Progress	/	/	83 (25.38)	/	
Unknown	/	/	244 (74.62)	/	
Age (n, %)					
<=50 years	231 (30.60)	96 (30.00)	209 (63.91)	27 (25.96)	27 (30.68)
> 50 years	524 (69.40)	224 (70.00)	118 (36.09)	77 (74.04)	61 (69.32)
Fraction genome altered (n, %)					
<=0.25	359 (47.55)	164 (51.25)	/	/	/
> 0.25	383 (50.73)	151 (47.19)	/	/	/
Unknown	13 (1.72)	5 (1.56)	/	/	/
T stage (n, %)					
T1	197 (26.09)	81 (25.31)	101 (30.89)	/	/
T2	439 (58.15)	184 (57.50)	188 (57.49)	/	/
T3	90 (11.92)	43 (13.44)	26 (7.95)	/	/
T4	28 (3.71)	10 (3.12)	12 (3.67)	/	/
Unknown	1 (0.13)	2 (0.63)	/	/	/
N stage (n, %)					
N0	356 (47.15)	148 (46.25)	137 (41.90)	/	/
N1	243 (32.19)	113 (35.31)	87 (26.61)	/	/
N2	81 (1073)	39 (12.19)	63 (19.26)	/	/
N3	58 (7.68)	17 (5.31)	40 (12.23)	/	/
Unknown	17 (2.25)	3 (0.94)	/	/	/
M stage (n, %)					
M0	621 (82.25)	272 (85.00)	319 (97.55)	/	/
M1	15 (1.99)	7 (2.19)	8 (2.45)	/	/
Unknown	119 (15.76)	41 (12.81)	/	/	/
Tumor stage (n, %)					

Variables	Training cohort (755 patients)	Internal validation cohort (320 patients)	GSE20685 validation cohort (327 patients)	GSE42568 validation cohort (104 patients)	GSE20711 validation cohort (88 patients)
Stage I	126 (16.69)	54 (16.88)	/	/	/
Stage II	427 (56.56)	182 (56.88)	/	/	/
Stage III	171 (22.65)	72 (22.50)	/	/	/
Stage IV	14 (1.85)	6 (1.87)	/	/	/
Unknown	17 (2.25)	6 (1.87)	/	/	/
Tumor grade (n, %)					
G1	/	/	/	11 (10.58)	13 (14.77)
G2	/	/	/	40 (38.46)	5 (5.68)
G3	/	/	/	53 (50.96)	70 (79.55)
ER status (n, %)					
Negative	169 (22.38)	68 (21.25)	/	34 (32.70)	45 (51.14)
Positive	551 (72.98)	240 (75.00)	/	67 (64.42)	42 (47.73)
Unknown	35 (4.64)	12 (3.75)	/	3 (2.88)	1 (1.13)
PR status (n, %)					
Negative	235 (31.13)	104 (32.50)	/	/	/
Positive	482 (63.84)	204 (63.75)	/	/	/
Unknown	38 (5.03)	12 (3.75)	/	/	/
HER2 status (n, %)					
Negative	390 (51.66)	163 (50.94)	/	/	62 (70.45)
Positive	110 (14.57)	50 (15.63)	/	/	26 (29.55)
Unknown	255 (33.77)	107 (33.44)	/	/	0 (0)
TNBC status (n, %)					
NO	593 (78.54)	253 (79.07)	/	/	/
YES	81 (10.73)	34 (10.62)	/	/	/
Unknown	81 (10.73)	33 (10.31)	/	/	/
BRCA subtype (n, %)					
Basal-like	121 (16.03)	49 (15.31)	/	/	27 (30.68)
HER2-enriched	43 (5.70)	27 (8.44)	/	/	26 (29.55)
Luminal B	352 (46.62)	150 (46.88)	/	/	22 (25.00)
Luminal A	128 (16.95)	57 (17.81)	/	/	13 (14.77)
Normal-like	101 (13.38)	35 (10.94)	/	/	/
Unknown	10 (1.32)	2 (0.62)	/	/	/
Immune subtype (n, %)					
C1	258 (34.17)	107 (33.44)	/	/	/
C2	274 (36.29)	109 (34.06)	/	/	/
C3	130 (17.22)	60 (18.75)	/	/	/
C4	59 (7.82)	28 (8.75)	/	/	/
C6	24 (3.18)	14 (4.38)	/	/	/
Unknown	10 (1.32)	2 (0.62)	/	/	/

Abbreviation: OS, overall survival; DFS, disease-free survival; PFS, progression-free survival; TNBC, triple negative breast cancer.

Table 2

Genes of IRGP prognostic signature.

Gene pairs	Coef	Gene1	Full Name	Immune Processes	Gene2	Full Name	Immune Processes
CCL2|CXCL9	0.2149	CCL2	C-C motif chemokine ligand 2	Antimicrobials	CXCL9	C-X-C motif chemokine ligand 9	Chemokines
CCL5|TNFRSF12A	-0.4298	CCL5	C-C motif chemokine ligand 5	Antimicrobials	TNFRSF12A	TNF receptor superfamily member 12A	Cytokine_Receptors
CCL8|ICAM2	0.5647	CCL8	C-C motif chemokine ligand 8	Antimicrobials	ICAM2	intercellular adhesion molecule 2	NaturalKiller_Cell_Cytotoxicit
CMTM8|RAC2	0.0581	CMTM8	CKLF like MARVEL transmembrane domain containing 8	Cytokines	RAC2	Rac family small GTPase 2	BCRSignalingPathway
CXCL11|IGLV1-44	0.1136	CXCL11	C-X-C motif chemokine ligand 11	Antimicrobials	IGLV1-44	immunoglobulin lambda variable 1–44	BCRSignalingPathway
CXCL13|PIK3R3	-0.1407	CXCL13	C-X-C motif chemokine ligand 13	Chemokines	PIK3R3	phosphoinositide-3-kinase regulatory subunit 3	NaturalKiller_Cell_Cytotoxicit
CXCL14|HMOX1	-0.0966	CXCL14	C-X-C motif chemokine ligand 14	Chemokines	HMOX1	heme oxygenase 1	Antimicrobials
CXCL14|PDGFRB	-0.1959	CXCL14	C-X-C motif chemokine ligand 14	Chemokines	PDGFRB	platelet derived growth factor receptor beta	Antimicrobials
GHR|IL7R	0.2607	GHR	growth hormone receptor	Cytokine_Receptors	IL7R	interleukin 7 receptor	Antimicrobials
IGF1R|TNFRSF21	-0.4738	IGF1R	insulin like growth factor 1 receptor	Cytokine_Receptors	TNFRSF21	TNF receptor superfamily member 21	Cytokine_Receptors
IGLV3-25|PDGFD	-0.3061	IGLV3-25	immunoglobulin lambda variable 3–25	BCRSignalingPathway	PDGFD	platelet derived growth factor D	Cytokines
INHBA|STC2	0.1536	INHBA	inhibin subunit beta A	Cytokines	STC2	stanniocalcin 2	Cytokines
INHBB|STC2	0.0299	INHBB	inhibin subunit beta B	Cytokines	STC2	stanniocalcin 2	Cytokines
MAPT|TNFRSF21	-0.0109	MAPT	microtubule associated protein tau	Antimicrobials	TNFRSF21	TNF receptor superfamily member 21	Cytokine_Receptors
SAA1|SECTM1	-0.1199	SAA1	serum amyloid A1	Chemokines	SECTM1	secreted and transmembrane 1	Cytokines
SEMA4B|STC2	0.1775	SEMA4B	semaphorin 4B	Chemokines	STC2	stanniocalcin 2	Cytokines

Abbreviation: IRGP, immune-relevant gene pair.

Table 3
Univariate and multivariate survival analyses of the IRGP prognostic signature and clinical variables.

	Univariate analysis			Multivariate analysis		
Variables	HR	95% CI	P-value	HR	95% CI	P-value
Training cohort (755 patients)						
Risk group (vs. Low risk)	4.04	2.68–6.09	2.63E-11	3.53	2.10–5.94	2.02E-06
Age (vs. <=50)	1.78	1.17–2.71	7.54E-03	1.54	0.93–2.53	9.31E-02
M1 stage (vs. M stage0)	5.02	2.67–9.44	5.75E-07	16.30	2.35–112.81	4.69E-03
N stage (vs. N stage0)						
N1 stage	1.45	0.91–2.31	1.14E-01	0.85	0.43–1.68	6.39E-01
N2 stage	2.68	1.51–4.74	7.54E-04	0.92	0.28–3.02	8.93E-01
N3 stage	2.75	1.35–5.59	5.16E-03	0.74	0.21–2.6	6.34E-01
T stage (vs. T stage1)						
T2 stage	1.21	0.76–1.93	4.12E-01	0.94	0.41–2.16	8.91E-01
T3 stage	1.41	0.77–2.56	2.64E-01	0.64	0.21–1.95	4.36E-01
T4 stage	2.53	1.21–5.3	1.36E-02	0.62	0.17–2.19	4.54E-01
Tumor stage (vs. Stage I)						
Stage II	1.42	0.77–2.63	2.67E-01	1.55	0.52–4.63	4.34E-01
Stage III	2.60	1.36–4.96	3.74E-03	4.71	0.93–23.92	6.14E-02
Stage IV	9.27	4.04–21.23	1.43E-07	NA	NA	NA
Fraction genome altered (vs. <=0.25)	1.38	0.95–2.02	9.49E-02	1.16	0.67–2.01	5.89E-01
Immune subtype (vs. C3)						
C1	0.95	0.54–1.68	8.70E-01	0.66	0.29–1.47	3.09E-01
C2	1.04	0.60–1.81	8.83E-01	0.68	0.31–1.46	3.19E-01
C4	2.19	1.03–4.67	4.16E-02	1.02	0.39–2.69	9.68E-01
C6	0.83	0.24–2.83	7.65E-01	1.03	0.22–4.8	9.69E-01
BRCA subtype (vs. Luminal A)						
Normal-like	2.81	1.78–4.43	9.43E-06	2.94	1.67–5.18	1.83E-04
Luminal B	1.34	0.73–2.49	3.46E-01	0.66	0.29–1.5	3.23E-01
HER2-enriched	1.94	0.86–4.36	1.09E-01	1.23	0.44–3.48	6.94E-01
Basal-like	1.11	0.59–2.08	7.50E-01	1.25	0.57–2.74	5.69E-01
Internal validation cohort (320 patients)						
Risk group (vs. Low risk)	3.80	2.03–7.15	3.28E-05	8.88	3.02–26.10	7.21E-05
Age (vs. <=50)	1.14	0.59–2.23	6.90E-01	0.83	0.35–1.98	6.73E-01
M1 stage (vs. M stage0)	5.18	2.06–13.04	4.75E-04	54.24	1.99–1474.97	1.78E-02
N stage (vs. N stage0)						
N1 stage	3.20	1.5–6.84	2.67E-03	3.14	0.95–10.37	6.05E-02
N2 stage	2.65	0.82–8.57	1.05E-01	0.51	0.06–4.25	5.35E-01
N3 stage	13.79	4.42–43.07	6.26E-06	8.88	1.18–66.8	3.39E-02
T stage (vs. T stage1)						
T2 stage	1.62	0.65–4.04	2.98E-01	0.43	0.06–3.25	4.16E-01
T3 stage	2.24	0.75–6.7	1.50E-01	0.56	0.05–5.94	6.27E-01
T4 stage	19.98	5.85–68.3	1.79E-06	0.37	0.03–4.43	4.31E-01
Tumor stage (vs. Stage I)						

	Univariate analysis			Multivariate analysis		
Stage II	2.34	0.69–7.96	1.73E-01	2.33	0.19–28.6	5.10E-01
Stage III	5.10	1.43–18.19	1.20E-02	13.05	0.55–310.72	1.12E-01
Stage IV	38.77	8.78–171.24	1.39E-06	NA	NA	NA
Fraction genome altered (vs. <=0.25)	1.81	0.95–3.47	7.14E-02	4.78	1.57–14.5	5.80E-03
Immune subtype (vs. C3)						
C1	0.81	0.34–1.90	6.28E-01	0.28	0.08–1.01	5.10E-02
C2	0.85	0.37–1.97	7.05E-01	0.22	0.06–0.79	2.03E-02
C4	0.87	0.23–3.27	8.31E-01	0.37	0.06–2.5	3.10E-01
C6	0.50	0.06–3.98	5.14E-01	0.09	0.01–1.52	9.41E-02
BRCA subtype (vs. Luminal A)						
Normal-like	4.57	1.78–11.73	1.57E-03	8.40	2.52–28	5.32E-04
Luminal B	2.25	0.91–5.58	8.02E-02	1.93	0.55–6.8	3.09E-01
HER2-enriched	2.50	0.80–7.85	1.15E-01	0.27	0.04–1.86	1.84E-01
Basal-like	1.67	0.67–4.13	2.68E-01	1.48	0.41–5.31	5.49E-01
GSE20685 validation cohort (327 patients)						
Risk group (vs. Low risk)	3.50	2.27–5.40	1.63E-08	2.98	1.88–4.72	3.40E-06
Age (vs. <=50)	0.83	0.52–1.31	4.18E-01	0.84	0.52–1.37	4.89E-01
M1 stage (vs. M stage0)	5.20	2.39–11.33	3.22E-05	1.21	0.38–3.85	7.51E-01
N stage (vs. N stage0)						
N1 stage	2.40	1.24–4.66	9.62E-03	2.76	1.39–5.46	3.59E-03
N2 stage	5.10	2.74–9.48	2.77E-07	4.37	2.2–8.67	2.48E-05
N3 stage	5.10	2.55–10.23	4.38E-06	4.39	1.98–9.72	2.70E-04
T stage (vs. T stage1)						
T2 stage	1.14	0.66–1.94	6.42E-01	0.65	0.36–1.15	1.39E-01
T3 stage	4.80	2.44–9.43	5.29E-06	1.41	0.63–3.17	4.05E-01
T4 stage	4.43	1.95–10.08	3.89E-04	1.30	0.39–4.26	6.70E-01
GSE42568 validation cohort (104 patients)						
Risk group (vs. Low risk)	4.54	1.60–12.88	4.45E-03	3.93	1.35–11.48	1.22E-02
Age (vs. <=50)	0.91	0.44–1.89	7.93E-01	0.90	0.43–1.88	7.79E-01
Tumor grade (vs. Grade 1)						
Grade 2	2.09	0.26–16.68	4.88E-01	1.08	0.13–9.08	9.43E-01
Grade 3	7.37	1.00–54.39	5.03E-02	3.76	0.48–29.32	2.06E-01
GSE20711 validation cohort (88 patients)						
Risk group (vs. Low risk)	2.47	1.08–5.66	3.23E-02	2.47	1.06–5.78	3.68E-02
Age (vs. <=50)	2.24	0.84–5.98	1.09E-01	2.31	0.86–6.22	9.61E-02
Tumor grade (vs. Grade 1)				0.80		
Grade 2	0.97	0.09–10.95	9.81E-01	2.07	0.07–9.50	8.62E-01
Grade 3	2.35	0.55–10	2.48E-01	2.47	0.48–8.98	3.30E-01
Abbreviation: IRGP, immune-relevant gene pair.						

Figures

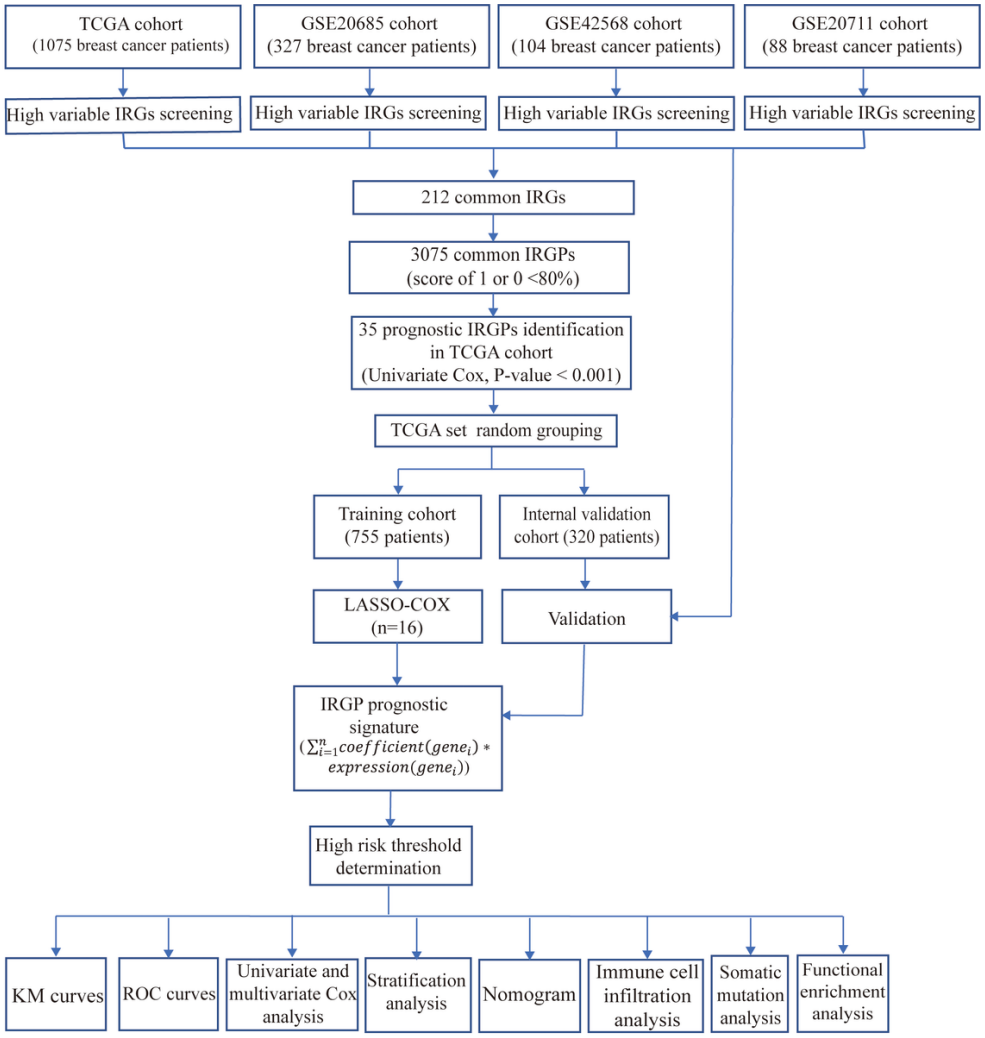

Figure 1

Workflow chart for data collection and analysis. IRGs, immune-relevant genes; IRGPs, immune-relevant gene pairs; ROC, receiver operating characteristic; KM, Kaplan-Meier.

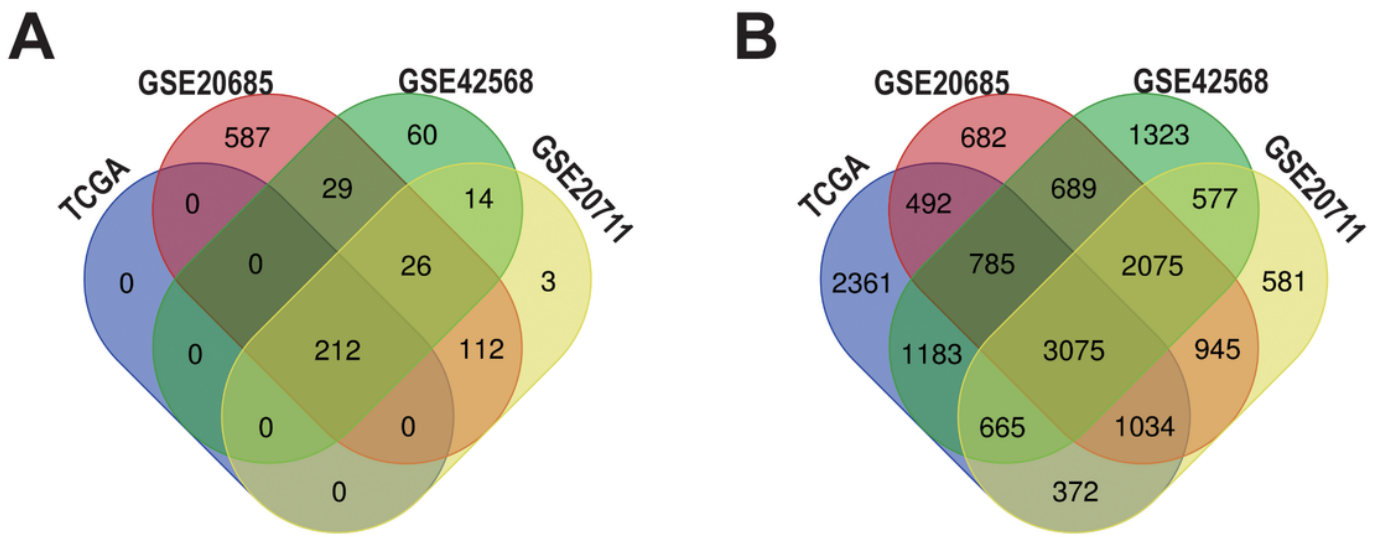

Figure 2

Candidate IRGPs identification. Venn diagram for high variable IRGs (A) and IRGPs with scores of 0 or 1 < 80% (B) in TCGA cohort, GSE20685 cohort, GSE42568 cohort, and GSE20711 cohort. IRGs, immune-relevant genes; IRGPs, immune-relevant gene pairs.

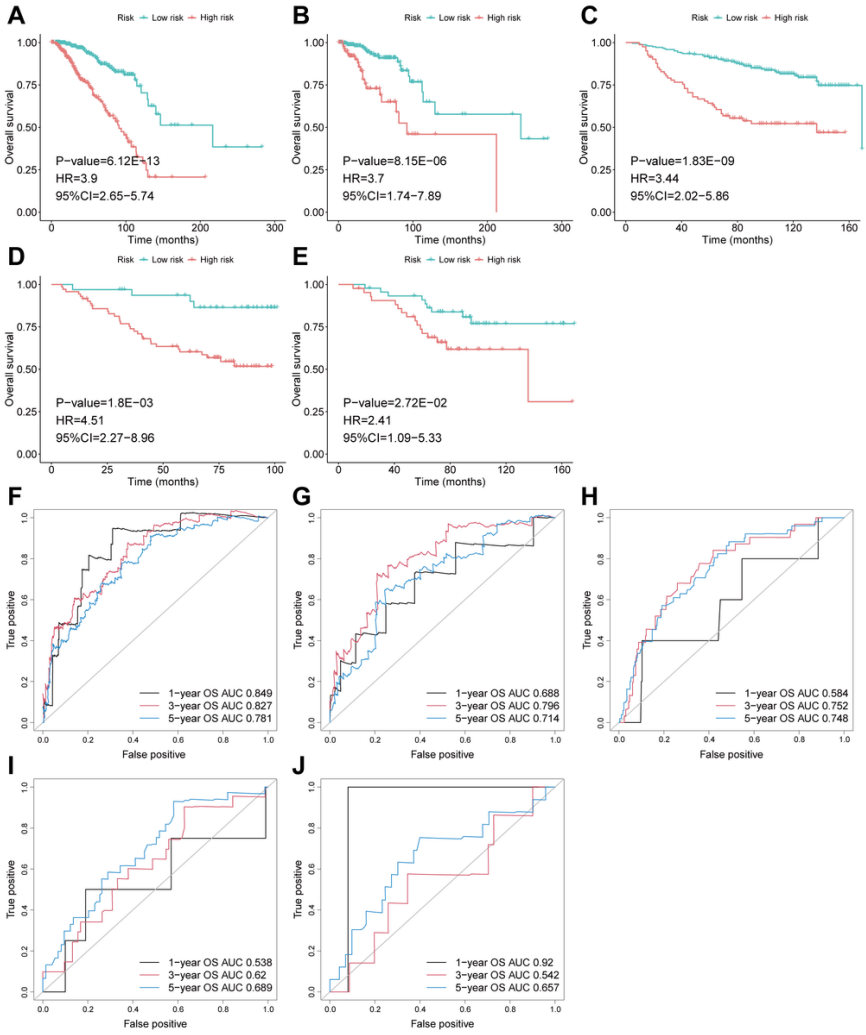

Figure 3

Construction and validation of the IRGP prognostic signature. (A) Kaplan-Meier curves for overall survival between 406 low-risk and 349 high-risk of breast cancer patients in TCGA training cohort. (B) Kaplan-Meier curves for overall survival between 236 low-risk and 84 high-risk of breast cancer patients in TCGA internal validation cohort. (C) Kaplan-Meier curves for overall survival between 245 low-risk and 82 high-risk of breast cancer patients in GSE20685 validation cohort. (D) Kaplan-Meier curves for overall survival between 33 low-risk and 71 high-risk of breast cancer patients in GSE42568 validation cohort. (E) Kaplan-Meier curves for overall survival between 45 low-risk and 43 high-risk of breast cancer patients in GSE20711 validation cohort. (F–J) ROC curves analysis for overall survival prediction according to IRGP prognostic signature in TCGA training cohort (F), TCGA internal validation cohort (G), GSE20685 validation cohort (H), GSE42568 validation cohort (I), and GSE20711 validation cohort (J). IRGP, immune-relevant gene pair; ROC, receiver operating characteristic.

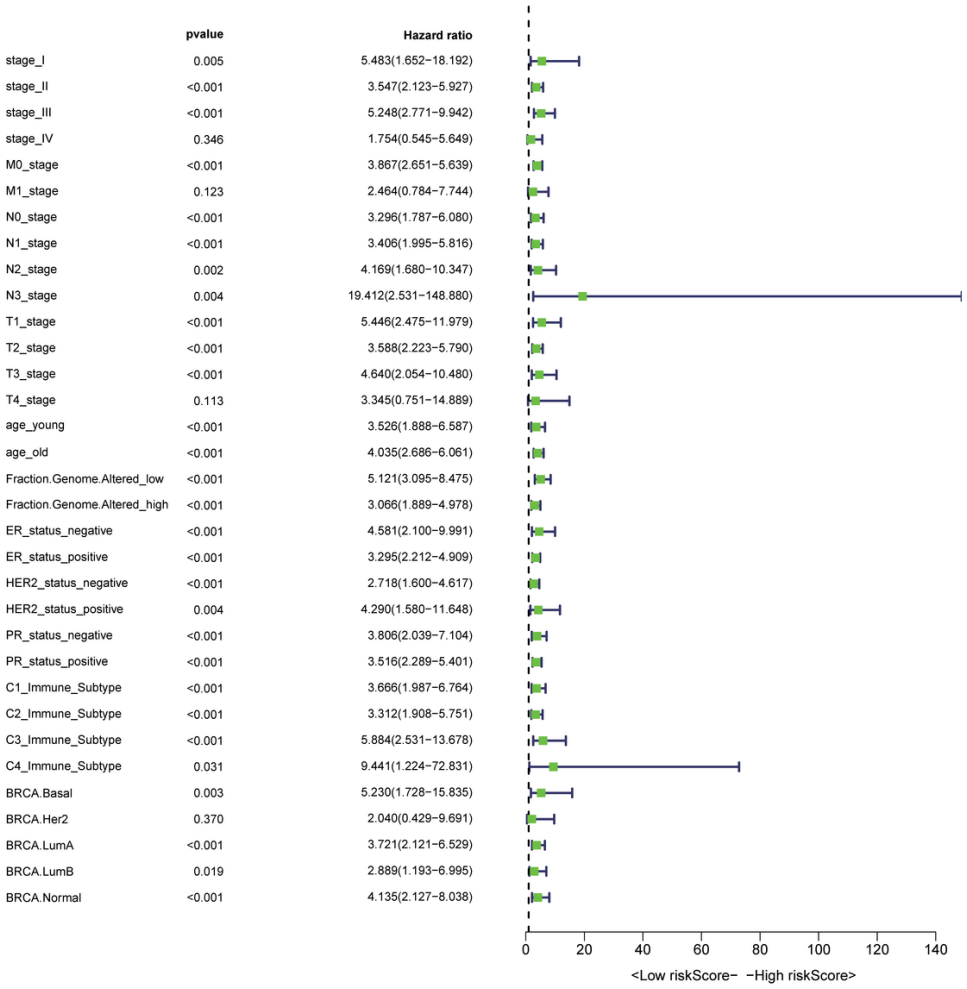

Figure 4

Forest plots of the associations between IRGP prognostic signature and overall survival in various clinical subgroups in TCGA entire cohort. Unadjusted HRs (boxes) and 95% confidence intervals (horizontal lines) are depicted. IRGP, immune-relevant gene pair.

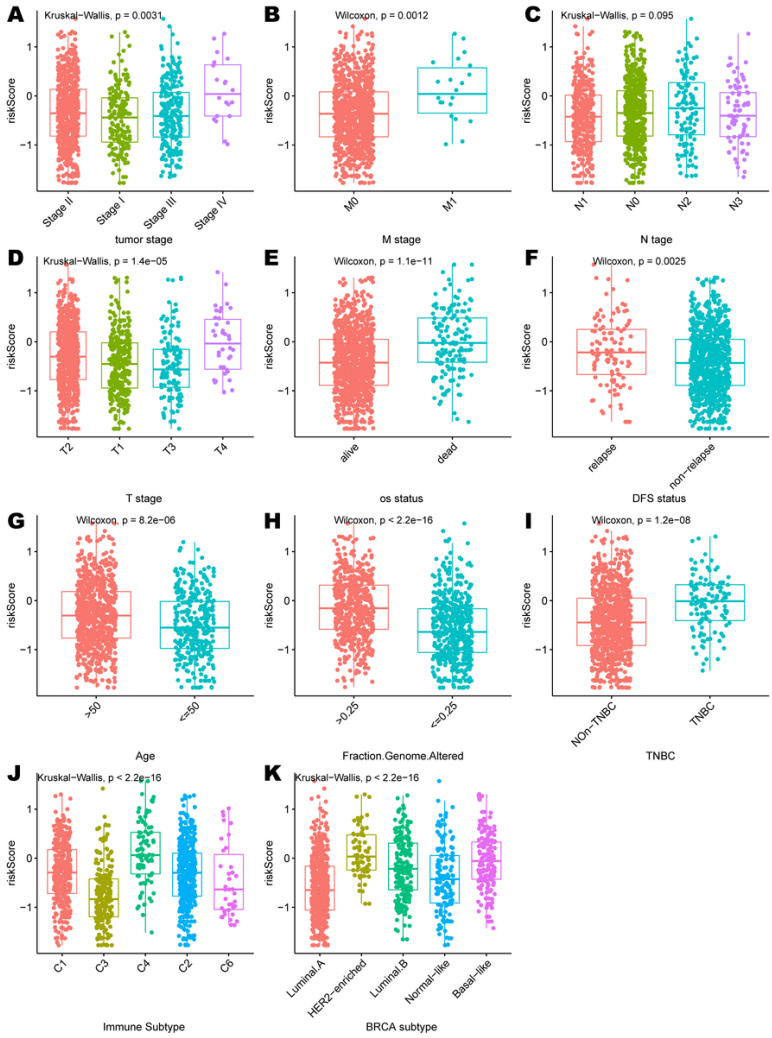

Figure 5

The relationship between IRGP risk score and clinical characteristics of breast cancer patients. The relationship between IRGP risk score and clinical characteristics, including tumor stage (A), M stage (B), N stage (C), T stage (D), died status (E), progression status (F), age (G), fraction genome altered (H), TNBC status (I), immune subtype (J), and breast cancer subtype (K). IRGP, immune-relevant gene pair; TNBC, triple negative breast cancer.

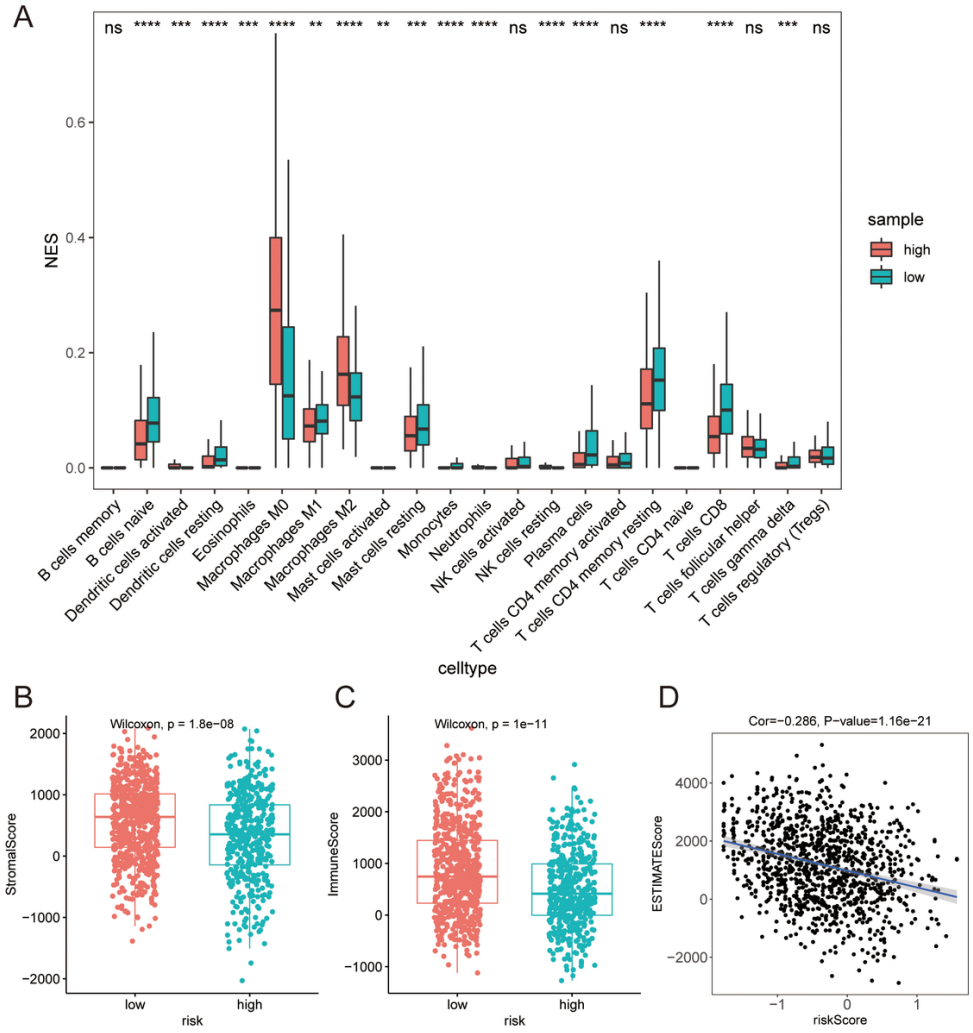

Figure 7

The immune cell infiltration analysis of IRGP prognostic signature. (A) Box plot of the comparison of 22 immune cells between high- and low- risk group with the use of the CIBERSORT analytical tool; (B-C) Box plot of the comparison of stromal score (B) and immune score (C) between high- and low- risk group with the tool of estimate; (D) The negative correlation between IRGP risk score and ESTIMATE score. IRGP, immune-relevant gene pair.

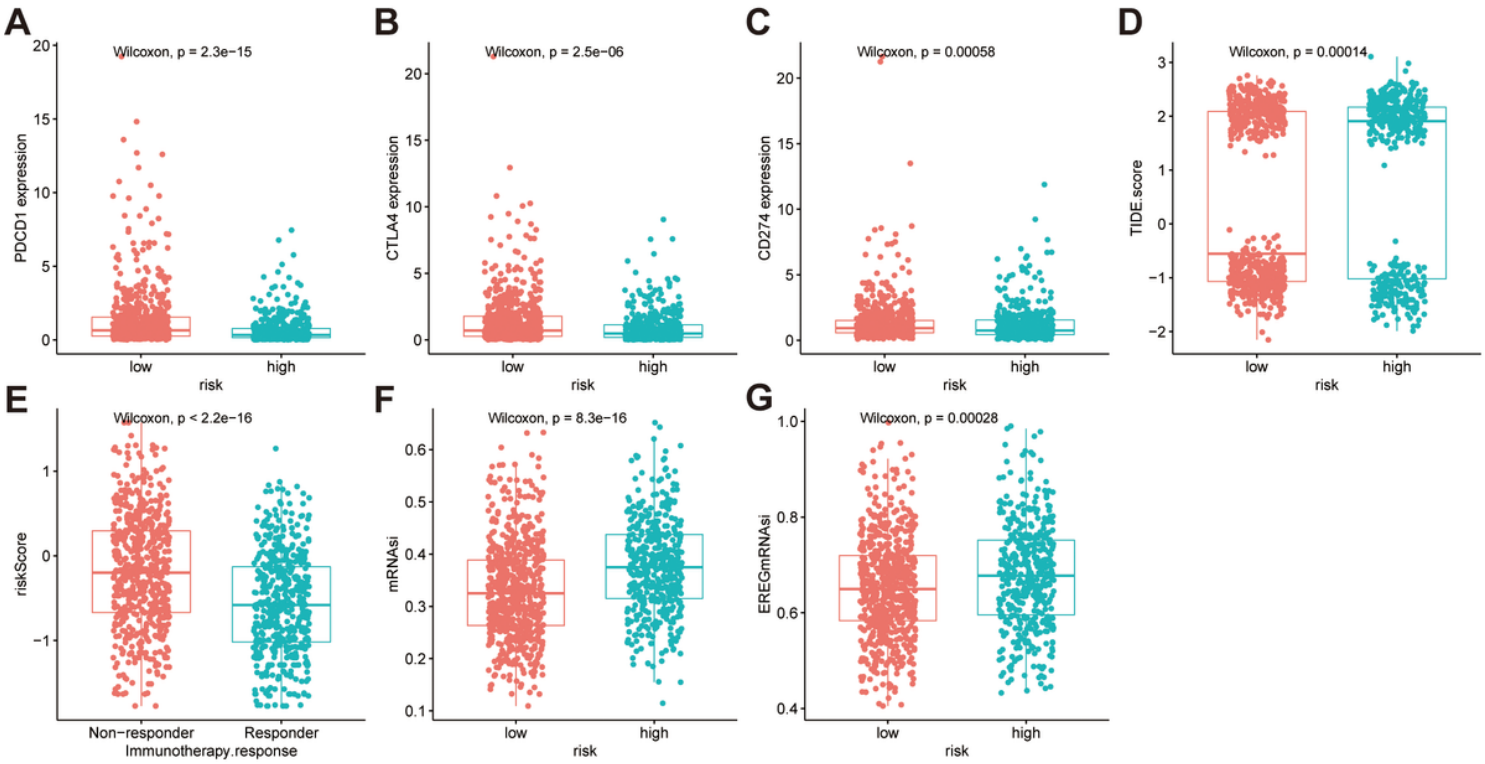

Figure 8

Evaluation of the immunotherapeutic response prediction. (A-C) Box plot of immune checkpoint inhibitors between high- and low- risk group, including PDCD1 (A), CTLA-4 (B), and CD274 (C). (D) Box plot of TIDE score between high- and low-risk group. (E) Correlation between risk score and immunotherapy response. (F) Box plot of mRNAsi between high- and low-risk group. (G) Box plot of EREG-mRNAsi between high- and low-risk group.

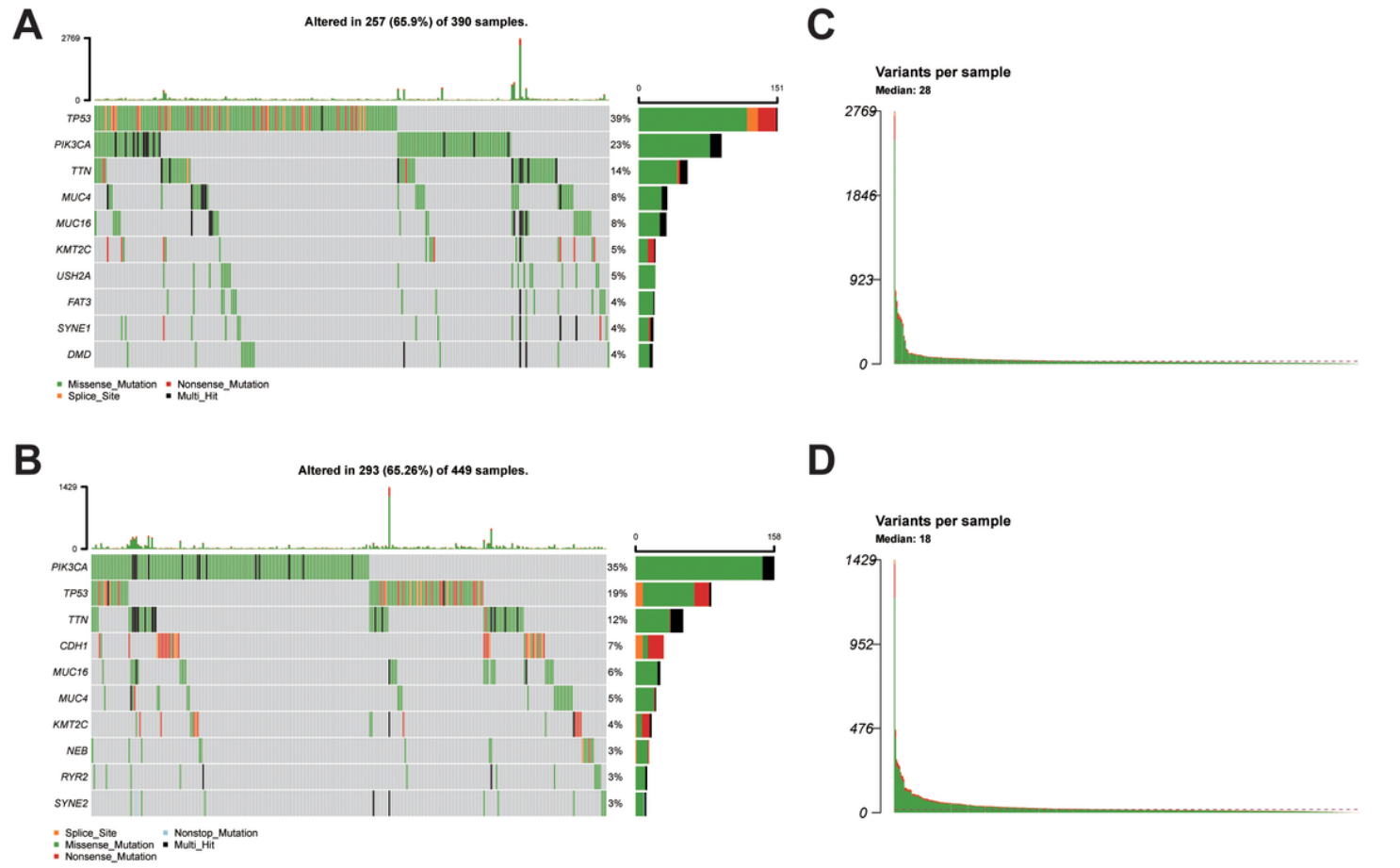

Figure 9

Landscape of mutation profiles in high risk and low risk breast cancer patients. Waterfall plot of mutation information for each gene in each sample in high-risk group (A) and low-risk group (B), in which various colors with annotations at the bottom represented the different mutation types. The barplot above the legend exhibited the mutation burden; Tumor mutation burden in specific samples in high-risk group (C) and low-risk group (D).

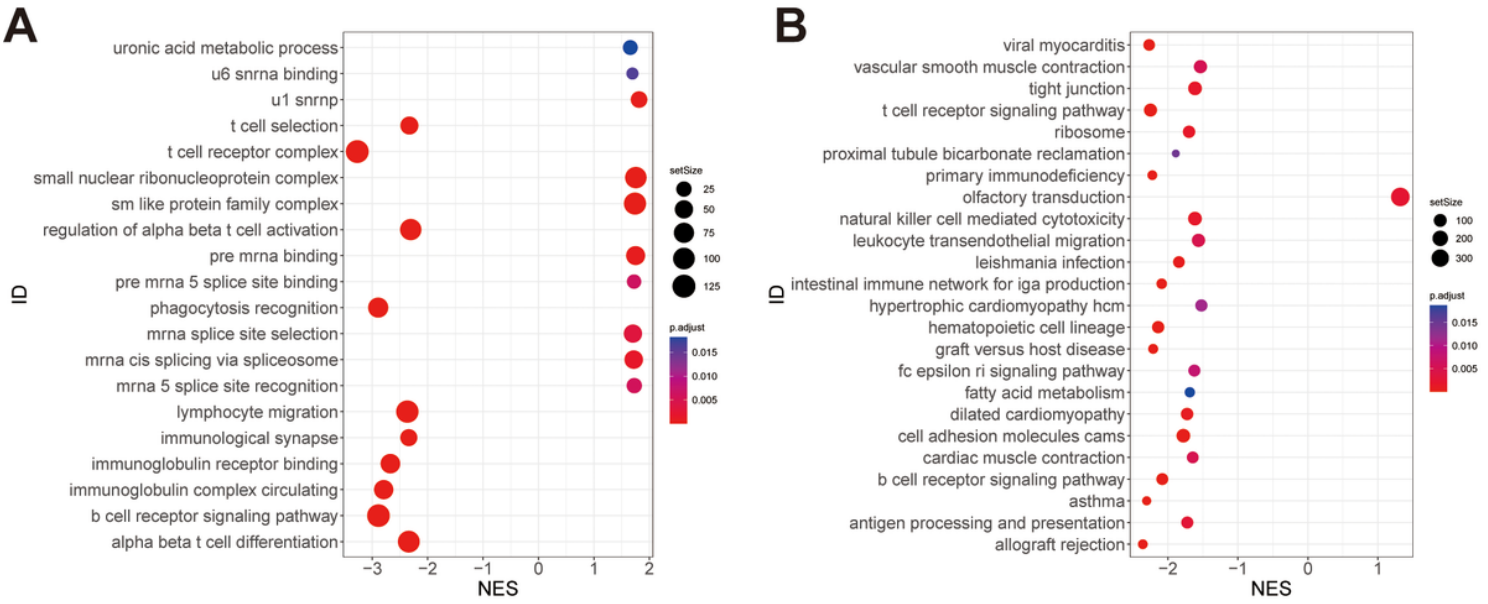

Figure 10

Gene set enrichment analysis results of IRGP prognostic signature. (A) The dot plot of top20 significant GO terms enrichment results. (B) The dot plot of all significant KEGG pathway enrichment results. IRGP, immune-relevant gene pair; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfigures.docx
- FigS1.tif
- FigS2.tif
- FigS4.tif
- FigS5.tif
- TableS1.docx
- TableS3.docx
- rawDatascript.zip