frontiers in Genetics

ORIGINAL RESEARCH
published: 21 October 2021
doi: 10.3389/fgene.2021.689715

Identification of Pathway-Based Biomarkers with Crosstalk Analysis for Overall Survival Risk Prediction in Breast Cancer

Xiaohua Liu[1†], *Lili Su*[2†], *Jingcong Li*[1] *and Guoping Ou*[1*]

[1]State Key Laboratory of Oncology in South China, Sun Yat-sen University Cancer Center, Guangzhou, China, [2]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China

Recently, many studies have investigated the role of gene-signature on the prognostic assessment of breast cancer (BC), however, the tumor heterogeneity and sequencing noise have limited the clinical usage of these models. Pathway-based approaches are more stable to the perturbation of certain gene expression. In this study, we constructed a prognostic classifier based on survival-related pathway crosstalk analysis. We estimated pathway's deregulation scores (PDSs) for samples collected from public databases to select survival-related pathways. After pathway crosstalk analysis, we conducted K-means clustering analysis to cluster the patients into G1 and G2 subgroups. The survival outcome of the G2 subgroup was significantly worse than the G1 subgroup. Internal and external dataset exhibits high consistency with the training dataset. Significant differences were found between G2 and G1 subgroups on pathway activity, gene mutation, immune cell infiltration levels, and in particular immune cells/pathway's activities were significantly negatively associated with BC patient's outcomes. In conclusion, we established a novel classifier reflecting the overall survival risk of BC and successfully validated its clinical usage on multiple BC datasets, which could offer clinicians inspiration in formulating the clinical treatment plan.

Keywords: breast cancer, deep-learning, pathway's deregulation scores, prognosis, classifier

OPEN ACCESS

Edited by:
Wan Zhu,
Stanford University, United States

Reviewed by:
Ning Wang,
Arcus Biosciences, United States
Salima Akter,
Kyung Hee University, South Korea

***Correspondence:**
Guoping Ou
ouguop@sysucc.org.cn

†These authors share co-first authorship

Specialty section:
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

Received: *01 April 2021*
Accepted: *28 September 2021*
Published: *21 October 2021*

Citation:
Liu X, Su L, Li J and Ou G (2021) Identification of Pathway-Based Biomarkers with Crosstalk Analysis for Overall Survival Risk Prediction in Breast Cancer. Front. Genet. 12:689715. doi: 10.3389/fgene.2021.689715

INTRODUCTION

As a highly metastatic and invasive malignant tumor with high incidence, breast cancer (BC) seriously threatens women's health and quality of life (Veronesi et al., 2005; Siegel et al., 2019; Rüschoff et al., 2020). BC occupied a quarter of all malignant tumors, which has received numerous clinical attention worldwide (Ferlay et al., 2015). At present, the primary treatment options for BC are chemotherapy, surgery, and radiotherapy (Shi et al., 2019). However, BCs tend to exhibit drug resistance and high recurrence rates on account of heterogeneity, making the therapeutic effects and prognosis of the disease unsatisfactory (Natarajan et al., 2012). Screening biomarkers for BC has a significant effect on reducing mortality, early diagnosis, and the improvement of prognosis in BC.

With the development of RNA-Seq high-throughput sequencing technology, various gene expression profiles of BC have been accumulated. Plenty of excellent models have been constructed to decode BC, the majority of them were built based on a single gene list. For example, van de Vijver et al. (2002) established a 70-gene prognosis profile to classify 295 BC patients, which is a powerful predictor for monitoring the prognosis of young BC patients. Tekpli

et al. (2019) identified clinically relevant immune clusters by integrating 15 BC cohorts, and discovered that patients with pro-tumorigenic immune infiltration were associated with poor prognosis. PAM50 (Parker et al., 2009) gene signature is a well-known molecular subtyping signature for BC, which could classify the BC into five molecular intrinsic subtypes: Normal-like, Basal-like, HER2+, Luminal A, and Luminal B. These efforts have helped us gain a deeper understanding of BC. Nevertheless, studies have found that due to the tumor heterogeneity and sequencing noise, gene-based signatures were highly unstable and the identified biomarkers were dramatically affected by the selection of training datasets (Michiels et al., 2005; Domany, 2014). In recent years many researchers indicated pathways could be helpful to extract more stable and interpretable features for risk prediction. Efforts have been made to decode cancer at levels of predefined pathways available in biological databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016), Reactome (Fabregat et al., 2018), and Gene-Set Enrichment Analysis (GSEA) (He et al., 2018). However, most existing pathways are general rather than disease-specific, and disease progression can only be partially affected by them. For pathway pairs with many common genes, we call it crosstalk. Taking the impact of overlapping genes on the pathway activity score (PAS) quantification of the two pathways into consideration can help identify disease-related features. Although it is intuitively believed that pathways will influence each other, especially when genes are shared, the existence of this phenomenon has not been studied in PAS estimation. And few studies have explored the PAS in cancer with crosstalk accommodated among well-established pathways to identify cancer-specific sub-pathways that could be used to predict the prognosis of cancer patients. Therefore, subtyping patients based on PAS and pathway crosstalk analysis is essential to promote personalized medicine.

In this study, we constructed a novel classifier reflecting the overall survival risk of BC based on survival-related pathway crosstalk analysis. We calculated the PAS for each pathway obtained from KEGG and GO resources based on the expression matrix. And then investigated the influence of crosstalk between these selected pathways on different cohorts to select the most critical 100 sub-pathways among all cohorts. we further conducted a K-means clustering analysis to cluster the patients into G1 (moderate) and G2 (aggressive) subgroups. Internal and external dataset exhibits high consistency with the training dataset.

MATERIALS AND METHODS

Data Source

We collected BC gene expression profiles from TCGA and GEO datasets, and the dataset with less than 20 samples or without overall survival information was excluded from our selection. TCGA mRNA expression data (level 3) and clinical features were downloaded from the UCSC Xena webserver (https://xenabrowser.net/datapages), while GSE16446, GSE42568, GSE7390, GSE20711, GSE1456A, GSE1456B and GSE20685 microarray data and relevant clinical information were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/gds/). After removing normal and non-survival information samples, we finally obtained 1,090 (TCGA), 107 (GSE16446), 104 (GSE42568), 198 (GSE7390), 88 (GSE20711), 159 (GSE1456A), 159 (GSE1456B) and 327 (GSE20685) BC samples for each dataset.

Pathway Activity Score

The pathway activity score (PAS) for each dataset was calculated based on the method proposed by Bhandari et al. (Bhandari et al., 2019). We downloaded all pathways from the gene ontology (GO) database (http://geneontology.org/) and generated a new mRNA expression matrix that contains only genes that exist in it for each pathway. After that, for each gene, based on its expression level, we classified the tumors into two subgroups, the samples in the higher group were scored +1, while the others were scored −1. Finally, we averaged all gene scores in this pathway as the pathway activity score for each tumor sample. A higher PAS indicates a higher pathway activity in the sample, and otherwise, a lower score means lower activity in the sample.

Selection of Survival-Related Sub-Pathways

Based on PASs and survival information, we calculated the log-rank *p*-value for each pathway by regression analysis. The pathways were then ranked based on the log-rank *p*-value. To minimize the false positive rate, we used the common significant pathways of these three large breast cancer cohorts, instead of using any single data set. The combined rank of each pathway was determined by the sure independence screening (SIS) method. We further selected the top n pathways for pathway crosstalk analysis. The threshold n was set to 100, which is much bigger than N/log(N), where N is the sample size of each cohort.

Different pathways often share some of the same genes, which can lead to crosstalk in the prognostic associations of different pathways. Considering the influence of overlapping genes on the PAS quantification of the two pathways can help identify cancer-related features. We further identified the crosstalk among the 100 selected survival-related pathways to define sub-pathways related to survival. The crosstalk between two pathways with at least three genes in common could be classified into three types. The overlapped genes between pathway A and pathway B could be defined as $P_A \cap P_B$, while the unique genes that specifically exist in pathway A or pathway B were defined as $P_A - (P_A \cap P_B)$ and $P_B - (P_A \cap P_B)$. Based on this classifier, each pathway pair could generate three sub-pathways.

To make sure each sub-pathway contains enough genes for further analysis, we obtained sub-pathways that consist of at least three genes. The Cox-PH model was used to calculate the survival risk *p*-value based on the recalculated PAS for each sub-pathway. After Bonferroni correction, we identified critical survival-related pathways and sub-pathways (FDR *p*-value < 0.01) for each dataset, and the overlapped pathways were finally adopted for further modeling.

Model Construction and Evaluation

With the pathways generated above, we constructed a pathway activity matrix with the row names are sub-pathways and the column names are sample IDs for each dataset. We performed consensus clustering with the pathway features acquired above to classify the TCGA samples and obtained the best cluster number as 2 based on three metrics, including C-index, Brier score, and log-rank *p*-value to redefine the samples as G1 and G2 subgroups. To predict these two subgroups for other datasets, we used several machine learning methods, including SVM, Adaboost, and Gaussian, to build a prediction model and obtained the best performance based on the pathway activity matrix. The robustness of the model was evaluated in the TCGA testing dataset and several external individual GEO datasets. We further built a classification model using several machine learning methods, including SVM, Adaboost, and Gaussian, based on these labels. We used the grid search to slightly turn the hyperparameters of the classifier. In the cross-validation procedure, TCGA samples were divided into training and testing datasets in a 4:1 ratio, and the training dataset was used to perform 10-fold cross-validation. To predict the GEO dataset, we used all TCGA samples to build the classification model.

We then compared three metrics, including C-index, Brier score, and log-rank *p*-value, to evaluate the model's performance. These metrics can quantify the proportion of patient pairs in a cohort whose risk prediction is highly consistent with survival outcomes. Usually, a higher C-index indicates more precise prediction performance, and 1 means perfect prediction, while 0.5 means the prediction performance is similar to random prediction. To calculate the C-index, we built a Cox-PH model based on the clustering labels and the patient's survival information in the TCGA training dataset and predicted the survival rate in the testing dataset, which was calculated by the R "survcomp" package. Brier score reflects the mean difference between observed and predicted survival after a certain period in survival analysis, and a lower score means good performance. Log-rank *p*-value was calculated by the R "survival" package to show the survival difference between the two groups, and a lower score means a more significant survival difference.

Survival Analysis

The log-rank test compares the survival difference of two groups at each observed event time was performed by R "survival" package. Kaplan-Meier analysis was applied to obtain a survival-curve plot of BRCA subtypes. Multivariate Cox regression analysis determined the independent role of this newly established predictor. Besides, we adopted Fisher's exact test to compare the census gene mutation differences between G2 and G1 subgroups in the TCGA cohort. We also compared the distributions of G2 and G1 in different clinicopathological features, such as tumor stage, new tumor event, and sex, by using Fisher exact tests.

We used the "DESeq2" R package (Love et al., 2014) to real the differential expressed genes between G2 and G1 subgroup; the significant DEGs were identified as $|\text{LogFoldChange}| > 1$ and false discovery rate (FDR) < 0.05.

Gene Set Enrichment Analysis

GSEA analysis was used to compare the pathway activity difference between G2 and G1, in which the R "ClusterProfier" package was performed. We adopted Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways to perform GSEA analysis and the top 20 significant pathways were displayed.

Mutation Analysis

The R "maftools" package was utilized to analyze and visualize the mutation data. The mutation data were compared between one group and the other groups using the chi-square test. A *p*-value of less than 0.05 was considered significant.

RESULTS

Identification of Overall Survival Risk Subtypes in BC

We obtained seven BC datasets (TCGA, GSE1456A, GSE1456B, GSE7390, GSE16446, GSE20685, GSE20711, and GSE42568) gene expression profiles and available clinical survival information from the TCGA and GEO databases. After calculating the PAS for each pathway obtained from KEGG and GO resources and selecting the survival-related pathways, we investigated the influence of crosstalk between these selected pathways on different cohorts, and then the most critical 100 sub-pathways among all cohorts were identified (**Supplementary Table S1**). K-means clustering analysis was used to divide the TCGA patients into two subgroups, defined as group 1 (G1, moderate) and group 2 (G2, aggressive) (**Supplementary Figure S1A**, and **Supplementary Table S2**). Notably, patients from the G2 subgroup show significantly worse clinical outcomes (overall survival and relapse-free survival) compared to the G1 subgroup (**Supplementary Figures S1B,C**; $p = 0.0053$ and $p = 0.0031$, respectively; log-rank test). We further built a classifier based on the TCGA training dataset with the label defined by k-means clustering analysis (*Materials and Methods*).

Evaluation of the Performance

To evaluate the robustness of OS risk prediction of the model, we tested the model performance on the TCGA testing dataset and several external GEO datasets, including GSE1456A, GSE7390, GSE16446, GSE20685, GSE20711, and GSE42568. As shown in **Figure 1** and **Table 1**, the model was stable and exhibited excellent classification capability, indicated by C-index and log-rank *p*-values between G2 and G1. The TCGA test cohort generated a high C-index (0.661), low Brier score (0.179), and significant average log-rank *p*-value ($p = 0.00123$) on survival difference. In different datasets, our classifier can significantly divide the samples into a good prognostic group and a poor prognostic group, which suggested that the newly developed classifier is able to universally predict the overall survival outcomes for patients with BC.

In order to compare the risk prediction capabilities of our predictor with some other clinical information, we performed univariate cox regression analysis for each clinical information

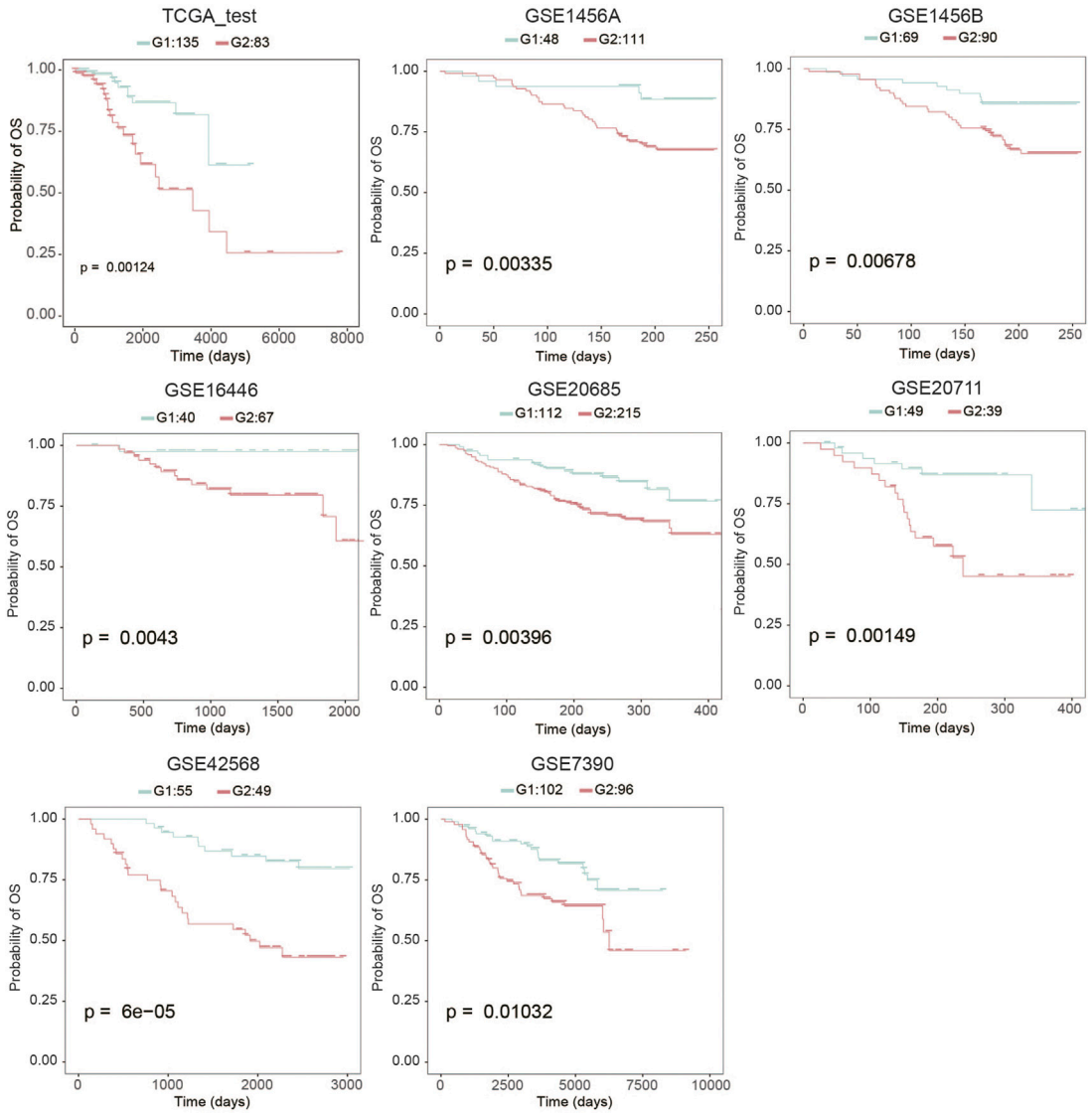

FIGURE 1 | Kaplan-Meier analysis to determine the survival differences between group 2 (G2) and group 1 (G1).

TABLE 1 | Cross-validation based performance robustness of classifier on TCGA training and testing cohorts.

Cohorts	Omics type	Samples	C-index	Brier score	Log-rank *p*
TCGA_test	RNA-Seq	218	0.661	0.179	1.23E-03
GSE1456A	Microarray	159	0.600	0.122	3.35E-03
GSE1456B	Microarray	159	0.599	0.121	6.78E-03
GSE16446	Microarray	107	0.646	0.112	4.30E-03
GSE20685	Microarray	327	0.579	0.153	3.96E-03
GSE20711	Microarray	88	0.656	0.189	1.49E-03
GSE42568	Microarray	104	0.679	0.195	6.27E-05
GSE7390	Microarray	198	0.597	0.180	1.03E-02

TCGA, The Cancer Genome Atlas.

(including age, tumor stage, and PAM50 subtyping) in the TCGA dataset as well as the external validation datasets. As shown in **Figure 2**, our classifier has a more general prognostic ability than other clinical information ($p < 0.05$ in all datasets). We further introduced several published transcriptomic-based predictors as previous study (Lee et al., 2021), including the proliferation index

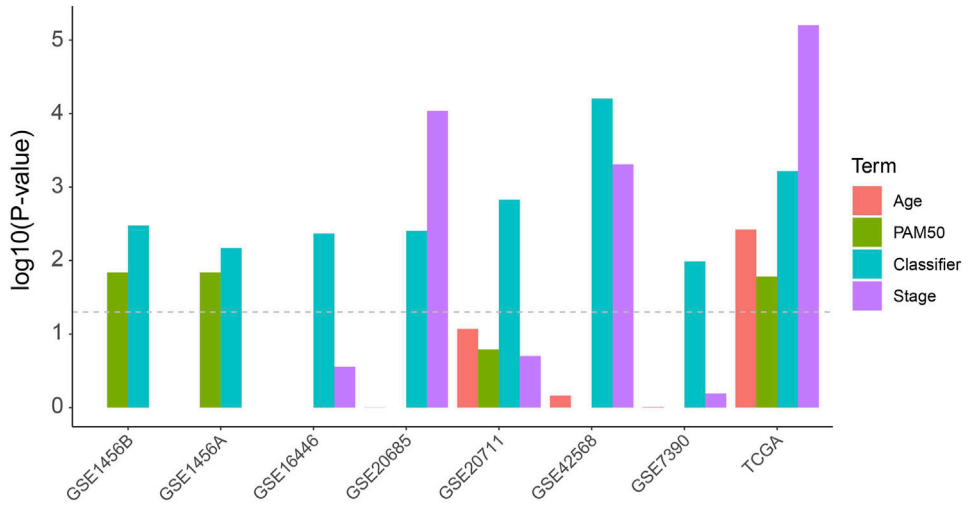

FIGURE 2 | Univariate Cox analysis of the classifier as well as regular clinical classification (Age, PAM50 and tumor stage) in TCGA and other extermal validation cohorts.

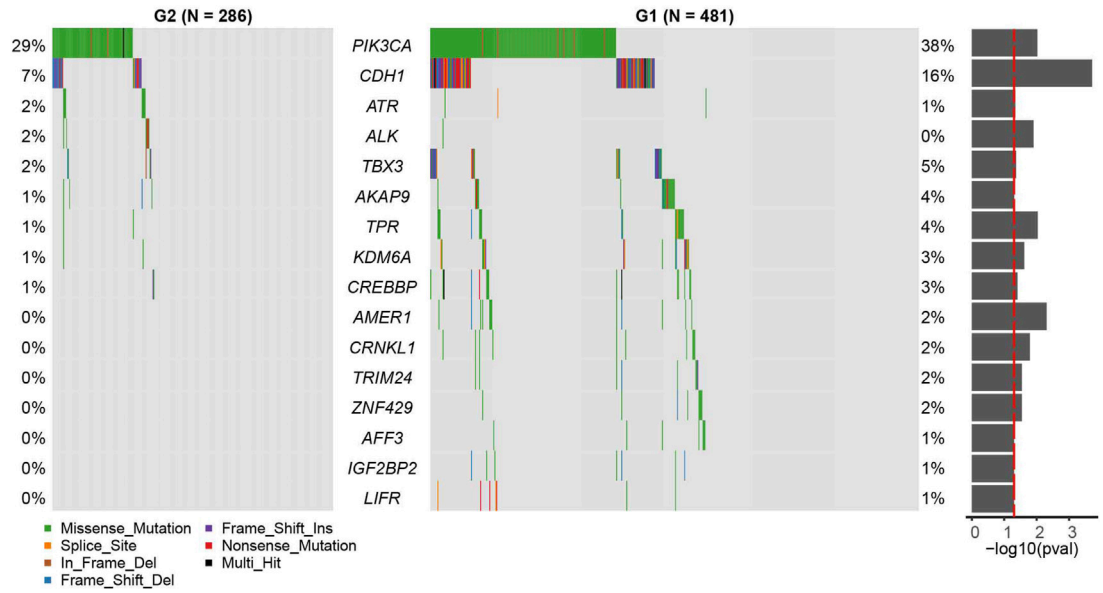

FIGURE 3 | Census mutation landscape between group 2 (G2) and group 1 (G1). Only cancer census genes in the COSMIC database are shown in the plot. The significance of the difference in gene mutation frequency between the two groups is shown in the barplot on the right (fisher's exact test).

(Whitfield et al., 2006), interferon-γ (IFNγ) signature score (Ayers et al., 2017) as well as cytolytic activity score (Rooney et al., 2015), and performed a multivariate Cox regression analysis with age, tumor stage, and our classifier (**Supplementary Figure S2**). In this analysis, the proliferation index and the IFNγ signature score were estimated as ssGSEA score (Yi et al., 2020a) of each gene signature, respectively, and the cytolytic activity score was calculated as the mean expression level of *GZMA* and *PRF1* (Rooney et al., 2015). As shown in **Supplementary Figure S2**, the proliferation index and IFNy signature score show higher predictive power (hizard ratios were 2.13 and 0.27, respectively), but also have larger confidence intervals and *p*-values, which suggesting that they cannot be used as independent prognostic factors of BRCA. Reassuringly, our classifier had a more stable hizard ratioa near statistically significant *p*-value. In addition, we also test the risk prediction performance in different subgroups of age and tumor stage (**Supplementary Figure S3**). This result suggests

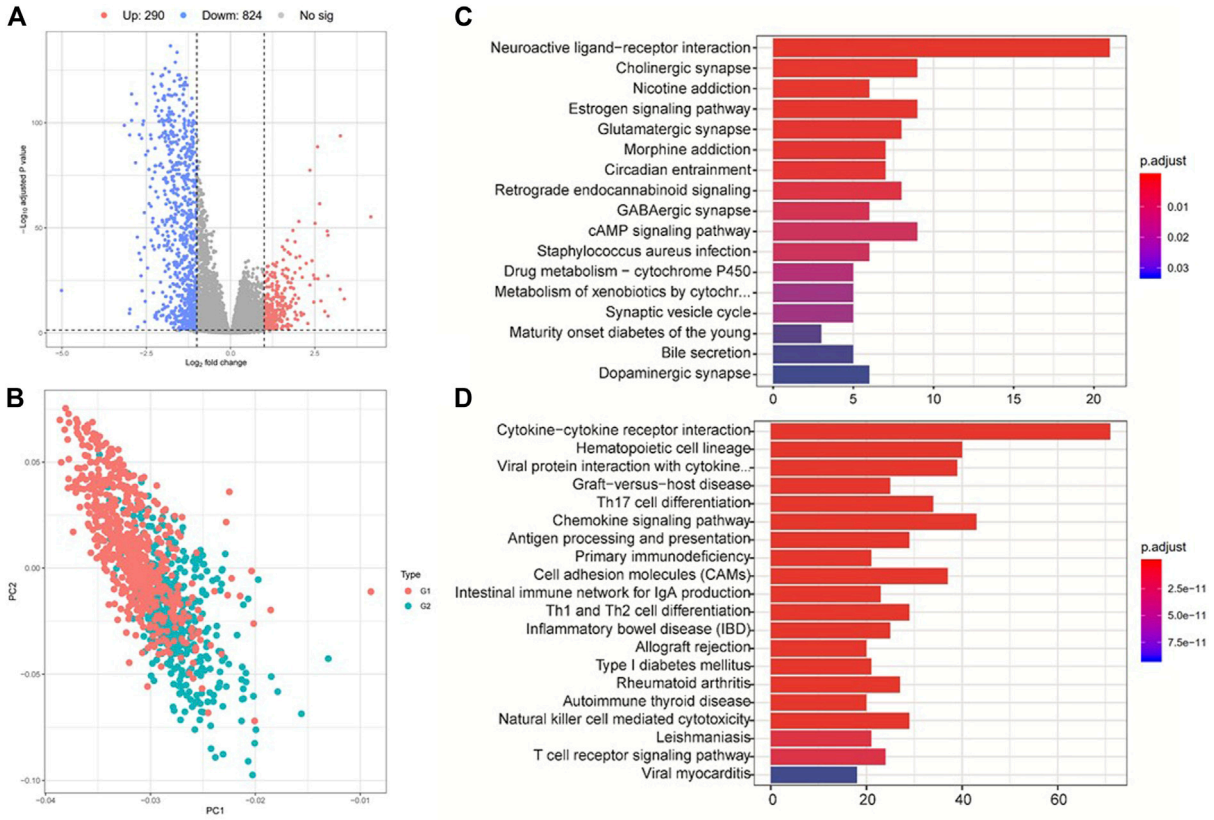

FIGURE 4 | Analysis of differentially expressed genes and their corresponding pathways. **(A)** Gene expression differences between group 2 (G2) and group 1 (G1). **(B)** PCA analysis shows the clustering results of group 2 (G2) and group 1 (G1). **(C)** and **(D)** KEGG pathway enrichment analysis for the up- (up panel) and down-regulated (down panel) genes between G2 and G1 subgroups.

that our classifier can be used essentially for the typing of different clinical subgroups, although in the low-age group and low-level group of TCGA the *p*-values did not reach significance.

Association Between Different Survival Subtypes and Genomic Feature

We found that the mutation rates of *PI3KCA* and *CDH1* were significantly higher in the G1 group than G2 group (*PI3KCA*: OR = 0.655, 95%CI: 0.471–0.907, $p = 0.00951$; *CDH1*: OR = 0.389, 95%CI: 0.220–0.660, $p = 0.000195$, **Figure 3**, and **Supplementary Table S3**, Fisher's exact test). Other significant differentially mutated genes between the two groups, including *ATR*, *ALK*, *TBX3*, *AKAP9*, *TPR*, *KDM6A*, *CREBBP*, *AMER1*, *CRNKL1*, *TRIM24*, *ZNF429*, *AFF3*, *IGF2BP2*, and *LIFR* (**Supplementary Table S3**, all $p < 0.05$). No significant tumor mutation burden (TMB) level difference was found between G1 and G2 subgroups (**Supplementary Figure S4**). *PI3KCA* and *CDH1* are two frequently mutated genes in many cancers, including breast cancer, gastric cancer, colorectal carcinoma, and head and neck squamous cell carcinoma (Hansford et al., 2015; Millis et al., 2016; Zhang et al., 2017; An et al., 2018). However, the association of PIK3CA mutation and prognosis has not been clarified clearly, PI3KCA mutation can be associated with a better prognosis (Barbareschi et al., 2007; Maruyama et al., 2007; Pérez-Tenorio et al., 2007; Kalinsky et al., 2009) or a worse prognosis (Li et al., 2006; Lerma et al., 2008). In some studies, PIK3CA mutation even has no obvious relationship with the prognosis (Saal et al., 2005; Lai et al., 2008; Stemke-Hale et al., 2008; Michelucci et al., 2009; Loi et al., 2010; Boyault et al., 2012). A similar phenomenon was found for CDH1 mutation as well (Corso et al., 2018).

We then performed differential expression analysis between G2 and G1 subgroups, and identified 290 upregulated and 824 downregulated genes ($|\log_2 \text{fold change}| > 1$ and FDR *p*-value > 0.05) (**Figures 4A,B**, and **Supplementary Table S4**) based on the TCGA cohort. KEGG pathway analysis indicated that these upregulated genes were mostly enriched in neuroactive ligand–receptor interaction, cholinergic synapse and estrogen signaling pathways (**Figure 4C**). The downregulated genes were mostly enriched in immune-related pathways, such as cytokine–cytokine receptor interaction, hematopoietic cell lineage, graft–versus–host disease, and Th17 cell differentiation (**Figure 4D**). These results prompted that the G1 subgroup might be immune activated subtype, which could be associated with its better overall survival. We also tested the correlations between the two survival subtypes (G2 and G1) and

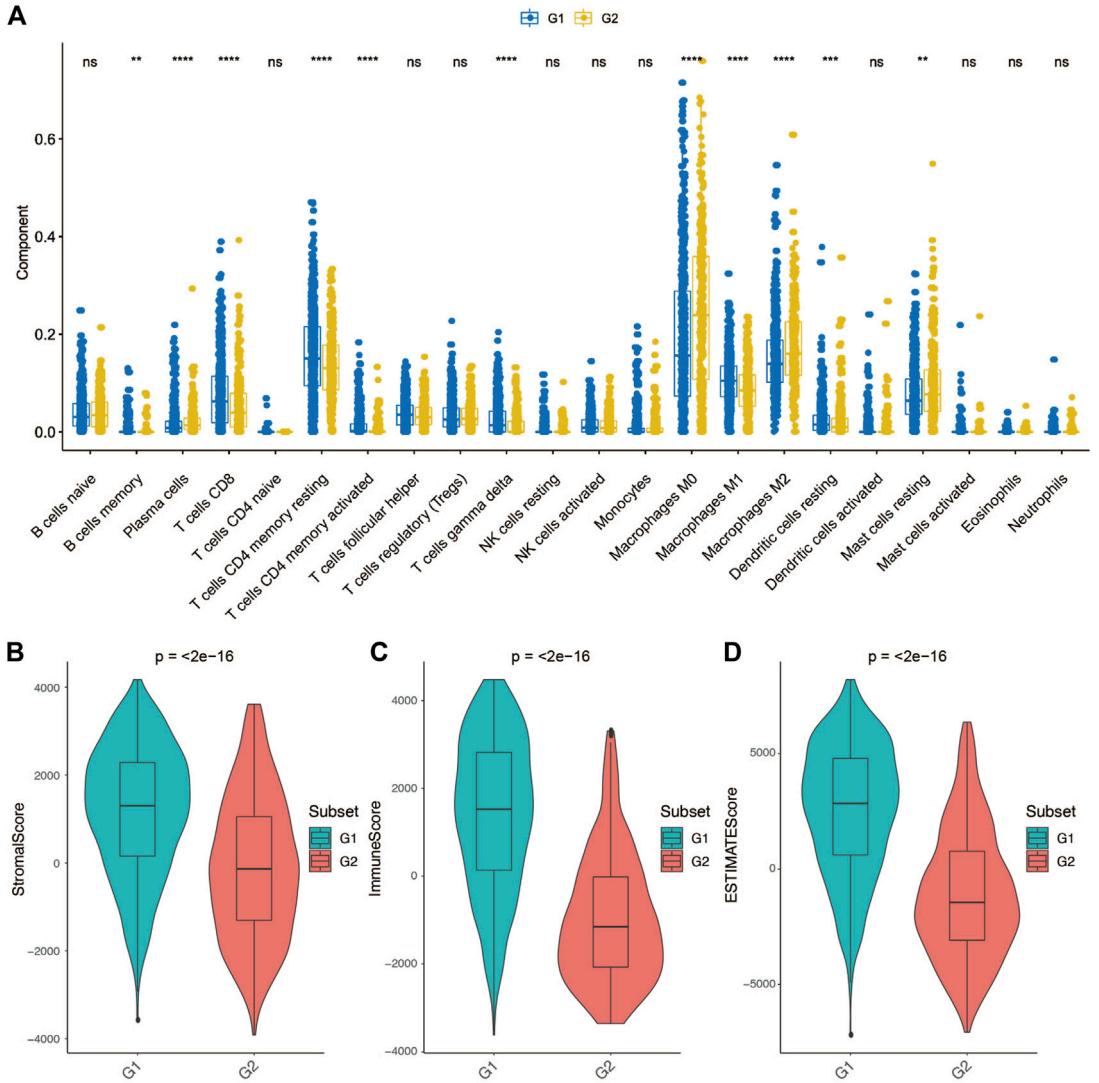

FIGURE 5 | Tumor microenvironment differences between group 2 (G2) and group 1 (G1) subgroups. **(A)** Comparison of each immune cell component between G1 and G2. **(B)** Comparison of stroma score between G1 and G2. **(C)** Comparison of immune score between G1 and G2. **(D)** Comparison of ESTIMATE score between G1 and G2.

clinicopathological characteristics from the TCGA cohort and found that no significant differences were revealed in age, sex, tumor stage, metastasis coded, estrogen receptor status, progesterone receptor status, and histological type subgroups, instead of PAM50 subtype (**Supplementary Figure S5**, **Supplementary Table S5**).

We then performed GSEA analysis to compare the G2 and G1 subgroups, aiming to identify critical pathways that displayed different activities between the G1 and G2 subgroups (**Supplementary Figure S6**, **Supplementary Tables S6–S8**). Hallmark pathway enrichment analysis showed that immune-related pathways including the inflammatory response, allograft rejection, interferon-gamma response and TNFA-signaling *via* NFκB were enriched in the G1 subgroup, while metabolic-related pathways such as oxidative phosphorylation signaling were activated in the G2 subgroup (**Supplementary Table S6**). Pathway enrichment analysis indicated that the differences between these two groups were concentrated in the KEGG pathways of "Graft *vs.* host disease", "primary immunodeficiency", and "allograft rejection" (**Supplementary Table S7**) and Reactome pathways related to co-stimulation by the CD28 family, generation of second messenger molecules, and cytokine signaling in the immune system (**Supplementary Table S8**). Previous studies have proved that metabolic pathway activities like oxidative phosphorylation signaling were negatively correlated with immune infiltration

and contributed to a worse prognosis in TNBC (Gong et al., 2021), which is consistent with our results.

Comparison of Tumor Microenvironment Between G2 and G1

We further employed the CIBERSORT algorithm to investigate the distributions of infiltrated immune cells between the G2 and G1 subgroups (**Supplementary Figure S7**). The result revealed that significant differences were obtained between two groups in CD8+T cells, CD4+T memory cells (resting), CD4+T memory (activated), γδ T cells, Macrophages M0, Macrophages M1, Macrophages M2, Dendritic cells (resting), and Mast cells (resting) (**Figure 5A**). Among macrophages, Macrophages M1 accounts for a higher proportion of the G1 subgroup, while the G2 subgroup consists of a higher proportion of Macrophages M2. Macrophages M2 was found to be dominant in BC and associated with poor clinical outcomes of BC (Bao et al., 2021), which could be the reason that the G2 subgroup patients have a worse overall survival.

Considering that the tumor tissue has tumor cells, stromal cells and immune cells, we measured stromal score and immune score based on specific gene expression signature to represent the level of immune infiltration and stroma infiltration of each tumor following the previous reported method-ESTIMATE (Yoshihara et al., 2013). Also, an ESTIMATE score also calculated which reflects the overall level of both immune infiltration and stromal infiltration. As shown in **Figures 5B–D**, G1 presented a higher stromal score, immune score and ESTIMATE score compared with G2. These results consistent with the previous definition that the G1 subgroup might be immune activated subtype, there was abundant crosstalk in the tumor microenvironment of this type of tumor, which could benefit from immunotherapy.

DISCUSSION

In the era of personalized medicine, there is an urgent need for a molecular marker-based approach to predict the prognostic outcomes of cancer patients accurately. Previous studies have reported many gene-based signatures to subtype BC (van de Vijver et al., 2002; Pu et al., 2020) (Tekpli et al., 2019) (Parker et al., 2009). Here, we constructed an overall survival risk model to classify samples into two subgroups. Internal and external datasets validation exhibits high consistency with the training dataset. Significant differences were found between the G2 and G1 subgroups including pathway activity, gene mutation, immune cell infiltration levels. In particular, immune cells/pathway's activities were significantly negatively associated with BC patient's outcomes.

In order to test whether our classifier is applicable to all ages and tumor grades, we performed prognostic association analysis for different clinical subgroups. For a data set with sufficient samples (more than 20 samples for each subgroup), our classifier can basically distinguish patients with different overall survival periods, although the high-age group and the low-stage group of TCGA have not reached statistical significance. Although the *p*-value of the high-age group of TCGA does not reach statistical significance (0.076), a clear trend can still be seen. However, it is challenging to explain why our classifier is unable to distinguish OS in the low-stage samples of TCGA with prognostic significance, though it performed well in the other two verification sets (**Supplementary Figures S3M,O**).

We found a significant mutation rate difference of PI3KCA and CDH1 gene between the G1 and G2 subgroups. It is not yet clear whether PIK3CA mutation is associated with clinical outcome, PI3KCA mutation can be associated with a better prognosis (Barbareschi et al., 2007; Maruyama et al., 2007; Pérez-Tenorio et al., 2007; Kalinsky et al., 2009) or a worse prognosis (Li et al., 2006; Lerma et al., 2008). In some studies, PIK3CA mutation even has no obvious relationship with the prognosis (Saal et al., 2005; Lai et al., 2008; Stemke-Hale et al., 2008; Michelucci et al., 2009; Loi et al., 2010; Boyault et al., 2012). A similar phenomenon was found for CDH1 mutation as well (Corso et al., 2018). Our results suggested that mutations of PI3KCA are positively associated with a favorable prognosis, but future studies are needed to investigate the potential mechanisms.

We also found that the G1 subgroup displayed significant higher level of immune infiltration, stromal infiltration level than the G2 subgroup. As reported, Th1 and cytotoxic types of memory T cells and CD8+ T cells can predict better prognosis in diverse cancers (Wei et al., 2018; Yi et al., 2020b; St. Paul and Ohashi, 2020). Several studies showed that the existence of mature antigen-presenting dendritic cells (DCs) could infiltrate colon cancer and theoretically increase immune response, which are correlated with improved survival as well (Schwaab et al., 2001). Other immune cells such as macrophages always produce plenty of factors influencing tumor cell's survival and growth, chemotaxis, cell invasion, angiogenesis, or repress T cell responses (Pagès et al., 2010). Therefore, a high rate of tumor-associated macrophages typically serves as a poor prognostic factor. It is valuable to predict the efficacy of specific therapies, especially immunotherapy. For example, Peng et al. recently developed a computational method named TIDE to accurately predict immunotherapy outcomes of melanoma (Jiang et al., 2018). The level of immune infiltration was significantly associated with the efficacy of immunotherapy (Galon and Bruni, 2019). The significant immunological differences between G1 and G2 suggest that our classifier may be predictive of immunotherapy efficacy. We will collect relevant data resources for more in-depth study in our future work.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: TCGA and Gene Expression Omnibus (GSE1456A, GSE16446, GSE20685, GSE20711, GSE42568, and GSE7390).

AUTHOR CONTRIBUTIONS

XL and GO designed this experiment. XL and LS performed the research and analyzed the data. XL and JL wrote the manuscript, and XL, LS, JL, and GO revised the manuscript. All these authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.689715/full#supplementary-material

Supplementary Figure S1 | The establishment and preliminary validation of the classifier. **(A)** Heatmap shows the clustering of two subclassification groups. **(B)** Kaplan-Meier plot shows the OS difference between group 2 (G2) and group 1(G1). **(C)** Kaplan-Meier plot shows the RFS difference between group 2 (G2) and group 1(G1).

Supplementary Figure S2 | Multivariate Cox regression analysis of prognostic factors for overall survival in TCGA patients. Age, tumor stage, proliferation score, IFNg score, cytolytic activity as well as our classifier were taken into consideration in this analysis.

Supplementary Figure S3 | KM curves show the prognostic performance of the classifier in different age level and tumor stage level of all included cohorts. Only cohorts with available clinical information are shown in the plot.

Supplementary Figure S4 | Comparison of tumor mutation burden between the G1 and G2 subgroups.

Supplementary Figure S5 | Sankey plot shows the distribution of group 2 (G2) and group 1 (G1) in different clinicopathological and genetic subgroups.

Supplementary Figure S6 | Gene Set Enrichment Analysis (GSEA) analysis shows the differences between group 2 (G2) and group 1 (G1) at Hallmark, (Kyoto Encyclopedia of Genes and Genomes) KEGG, and Reactome pathway aspects.

Supplementary Figure S7 | CIBERSORT analysis shows the distributions of 22 tumor infiltrated immune cells in each breast cancer sample.

REFERENCES

An, Y., Adams, J. R., Hollern, D. P., Zhao, A., Chang, S. G., Gams, M. S., et al. (2018). Cdh1 and Pik3ca Mutations Cooperate to Induce Immune-Related Invasive Lobular Carcinoma of the Breast. *Cel Rep.* 25 (3), 702–714. doi:10.1016/j.celrep.2018.09.056

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., et al. (2017). IFN-γ-related mRNA Profile Predicts Clinical Response to PD-1 Blockade. *J. Clin. Invest.* 127 (8), 2930–2940. doi:10.1172/jci91190

Bao, X., Shi, R., Zhao, T., Wang, Y., Anastasov, N., Rosemann, M., et al. (2021). Integrated Analysis of Single-Cell RNA-Seq and Bulk RNA-Seq Unravels Tumour Heterogeneity Plus M2-like Tumour-Associated Macrophage Infiltration and Aggressiveness in TNBC. *Cancer Immunol. Immunother.* 70 (1), 189–202. doi:10.1007/s00262-020-02669-7

Barbareschi, M., Buttitta, F., Felicioni, L., Cotrupi, S., Barassi, F., Del Grammastro, M., et al. (2007). Different Prognostic Roles of Mutations in the Helical and Kinase Domains of the PIK3CA Gene in Breast Carcinomas. *Clin. Cancer Res.* 13 (20), 6064–6069. doi:10.1158/1078-0432.Ccr-07-0266

Bhandari, V., Hoey, C., Liu, L. Y., Lalonde, E., Ray, J., Livingstone, J., et al. (2019). Molecular Landmarks of Tumor Hypoxia across Cancer Types. *Nat. Genet.* 51 (2), 308–318. doi:10.1038/s41588-018-0318-2

Boyault, S., Drouet, Y., Navarro, C., Bachelot, T., Lasset, C., Treilleux, I., et al. (2012). Mutational Characterization of Individual Breast Tumors: TP53 and PI3K Pathway Genes Are Frequently and Distinctively Mutated in Different Subtypes. *Breast Cancer Res. Treat.* 132 (1), 29–39. doi:10.1007/s10549-011-1518-y

Corso, G., Veronesi, P., Sacchini, V., and Galimberti, V. (2018). Prognosis and Outcome in CDH1-Mutant Lobular Breast Cancer. *Eur. J. Cancer Prev.* 27 (3), 237–238. doi:10.1097/cej.0000000000000405

Domany, E. (2014). Using High-Throughput Transcriptomic Data for Prognosis: a Critical Overview and Perspectives. *Cancer Res.* 74 (17), 4612–4621. doi:10.1158/0008-5472.Can-13-3338

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 46 (D1), D649–d655. doi:10.1093/nar/gkx1132

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012. *Int. J. Cancer* 136 (5), E359–E386. doi:10.1002/ijc.29210

Galon, J., and Bruni, D. (2019). Approaches to Treat Immune Hot, Altered and Cold Tumours with Combination Immunotherapies. *Nat. Rev. Drug Discov.* 18 (3), 197–218. doi:10.1038/s41573-018-0007-y

Gong, Y., Ji, P., Yang, Y.-S., Xie, S., Yu, T.-J., Xiao, Y., et al. (2021). Metabolic-Pathway-Based Subtyping of Triple-Negative Breast Cancer Reveals Potential Therapeutic Targets. *Cel Metab.* 33 (1), 51–64. doi:10.1016/j.cmet.2020.10.012

Hansford, S., Kaurah, P., Li-Chang, H., Woo, M., Senz, J., Pinheiro, H., et al. (2015). Hereditary Diffuse Gastric Cancer Syndrome. *JAMA Oncol.* 1 (1), 23–32. doi:10.1001/jamaoncol.2014.168

He, W., Chen, L., Yuan, K., Zhou, Q., Peng, L., and Han, Y. (2018). Gene Set Enrichment Analysis and Meta-Analysis to Identify Six Key Genes Regulating and Controlling the Prognosis of Esophageal Squamous Cell Carcinoma. *J. Thorac. Dis.* 10 (10), 5714–5726. doi:10.21037/jtd.2018.09.55

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T Cell Dysfunction and Exclusion Predict Cancer Immunotherapy Response. *Nat. Med.* 24 (10), 1550–1558. doi:10.1038/s41591-018-0136-1

Kalinsky, K., Jacks, L. M., Heguy, A., Patil, S., Drobnjak, M., Bhanot, U. K., et al. (2009). PIK3CA Mutation Associates with Improved Outcome in Breast Cancer. *Clin. Cancer Res.* 15 (16), 5049–5059. doi:10.1158/1078-0432.Ccr-09-0632

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res.* 44 (D1), D457–D462. doi:10.1093/nar/gkv1070

Lai, Y.-L., Mau, B.-L., Cheng, W.-H., Chen, H.-M., Chiu, H.-H., and Tzen, C.-Y. (2008). PIK3CA Exon 20 Mutation Is Independently Associated with a Poor Prognosis in Breast Cancer Patients. *Ann. Surg. Oncol.* 15 (4), 1064–1069. doi:10.1245/s10434-007-9751-7

Lee, J. S., Nair, N. U., Dinstag, G., Chapman, L., Chung, Y., Wang, K., et al. (2021). Synthetic Lethality-Mediated Precision Oncology via the Tumor Transcriptome. *Cell* 184 (9), 2487–2502. doi:10.1016/j.cell.2021.03.030

Lerma, E., Catasus, L., Gallardo, A., Peiro, G., Alonso, C., Aranda, I., et al. (2008). Exon 20 PIK3CA Mutations Decreases Survival in Aggressive (HER-2 Positive) Breast Carcinomas. *Virchows Arch.* 453 (2), 133–139. doi:10.1007/s00428-008-0643-4

Li, S. Y., Rong, M., Grieu, F., and Iacopetta, B. (2006). PIK3CA Mutations in Breast Cancer Are Associated with Poor Outcome. *Breast Cancer Res. Treat.* 96 (1), 91–95. doi:10.1007/s10549-005-9048-0

Loi, S., Haibe-Kains, B., Majjaj, S., Lallemand, F., Durbecq, V., Larsimont, D., et al. (2010). PIK3CA Mutations Associated with Gene Signature of Low mTORC1 Signaling and Better Outcomes in Estrogen Receptor-Positive Breast Cancer. *Proc. Natl. Acad. Sci.* 107 (22), 10208–10213. doi:10.1073/pnas.0907011107

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Maruyama, N., Miyoshi, Y., Taguchi, T., Tamaki, Y., Monden, M., and Noguchi, S. (2007). Clinicopathologic Analysis of Breast Cancers with PIK3CA Mutations in Japanese Women. *Clin. Cancer Res.* 13 (2 Pt 1), 408–414. doi:10.1158/1078-0432.Ccr-06-0267

Michelucci, A., Di Cristofano, C., Lami, A., Collecchi, P., Caligo, A., Decarli, N., et al. (2009). PIK3CA in Breast Carcinoma. *Diagn. Mol. Pathol.* 18 (4), 200–205. doi:10.1097/PDM.0b013e31818e5fa4

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of Cancer Outcome with Microarrays: a Multiple Random Validation Strategy. *The Lancet* 365 (9458), 488–492. doi:10.1016/s0140-6736(05)17866-0

Millis, S. Z., Ikeda, S., Reddy, S., Gatalica, Z., and Kurzrock, R. (2016). Landscape of Phosphatidylinositol-3-Kinase Pathway Alterations across 19 784 Diverse Solid Tumors. *JAMA Oncol.* 2 (12), 1565–1573. doi:10.1001/jamaoncol.2016.0891

Natarajan, K., Xie, Y., Baer, M. R., and Ross, D. D. (2012). Role of Breast Cancer Resistance Protein (BCRP/ABCG2) in Cancer Drug Resistance. *Biochem. Pharmacol.* 83 (8), 1084–1103. doi:10.1016/j.bcp.2012.01.002

Pagès, F., Galon, J., Dieu-Nosjean, M.-C., Tartour, E., Sautès-Fridman, C., and Fridman, W.-H. (2010). Immune Infiltration in Human Tumors: a Prognostic

Factor that Should Not Be Ignored. *Oncogene* 29 (8), 1093–1102. doi:10.1038/onc.2009.416

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Jco* 27 (8), 1160–1167. doi:10.1200/jco.2008.18.1370

Pérez-Tenorio, G., Alkhori, L., Olsson, B., Waltersson, M. A., Nordenskjöld, B., Rutqvist, L. E., et al. (2007). PIK3CA Mutations and PTEN Loss Correlate with Similar Prognostic Factors and Are Not Mutually Exclusive in Breast Cancer. *Clin. Cancer Res.* 13 (12), 3577–3584. doi:10.1158/1078-0432.Ccr-06-1609

Pu, M., Messer, K., Davies, S. R., Vickery, T. L., Pittman, E., Parker, B. A., et al. (2020). Research-based PAM50 Signature and Long-Term Breast Cancer Survival. *Breast Cancer Res. Treat.* 179 (1), 197–206. doi:10.1007/s10549-019-05446-y

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell* 160 (1-2), 48–61. doi:10.1016/j.cell.2014.12.033

Rüschoff, J., Lebeau, A., Sinn, P., Schildhaus, H.-U., Decker, T., Ammann, J., et al. (2020). Statistical Modelling of HER2-Positivity in Breast Cancer: Final Analyses from Two Large, Multicentre, Non-interventional Studies in Germany. *The Breast* 49, 246–253. doi:10.1016/j.breast.2019.12.005

Saal, L. H., Holm, K., Maurer, M., Memeo, L., Su, T., Wang, X., et al. (2005). PIK3CA Mutations Correlate with Hormone Receptors, Node Metastasis, and ERBB2, and Are Mutually Exclusive with PTEN Loss in Human Breast Carcinoma. *Cancer Res.* 65 (7), 2554–2559. doi:10.1158/0008-5472-can-04-3913

Schwaab, T., Weiss, J. E., Schned, A. R., and Barth, R. J., Jr. (2001). Dendritic Cell Infiltration in colon Cancer. *J. Immunother.* 24 (2), 130–137. doi:10.1097/00002371-200103000-00007

Shi, W., Luo, Y., Zhao, D., Huang, H., and Pang, W. (2019). Evaluation of the Benefit of Post-mastectomy R-adiotherapy in P-atients with E-arly-stage B-reast C-ancer: A P-ropensity S-core M-atching S-tudy. *Oncol. Lett.* 17 (6), 4851–4858. doi:10.3892/ol.2019.10197

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer Statistics, 2019. *CA A. Cancer J. Clin.* 69 (1), 7–34. doi:10.3322/caac.21551

Stemke-Hale, K., Gonzalez-Angulo, A. M., Lluch, A., Neve, R. M., Kuo, W.-L., Davies, M., et al. (2008). An Integrative Genomic and Proteomic Analysis of PIK3CA, PTEN, and AKT Mutations in Breast Cancer. *Cancer Res.* 68 (15), 6084–6091. doi:10.1158/0008-5472.Can-07-6854

St. Paul, M., and Ohashi, P. S. (2020). The Roles of CD8+ T Cell Subsets in Antitumor Immunity. *Trends Cel Biol.* 30 (9), 695–704. doi:10.1016/j.tcb.2020.06.003

Tekpli, X., Lien, T., Lien, T., Røssevold, A. H., Nebdal, D., Borgen, E., et al. (2019). An Independent Poor-Prognosis Subtype of Breast Cancer Defined by a Distinct Tumor Immune Microenvironment. *Nat. Commun.* 10 (1), 5499. doi:10.1038/s41467-019-13329-5

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., et al. (2002). A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N. Engl. J. Med.* 347 (25), 1999–2009. doi:10.1056/NEJMoa021967

Veronesi, U., Boyle, P., Goldhirsch, A., Orecchia, R., and Viale, G. (2005). Breast Cancer. *The Lancet* 365 (9472), 1727–1741. doi:10.1016/S0140-6736(05)66546-4

Wei, M., Shen, D., Mulmi Shrestha, S., Liu, J., Zhang, J., and Yin, Y. (2018). The Progress of T Cell Immunity Related to Prognosis in Gastric Cancer. *Biomed. Res. Int.* 2018, 1–6. doi:10.1155/2018/3201940

Whitfield, M. L., George, L. K., Grant, G. D., and Perou, C. M. (2006). Common Markers of Proliferation. *Nat. Rev. Cancer* 6 (2), 99–106. doi:10.1038/nrc1802

Yi, M., Nissley, D. V., McCormick, F., and Stephens, R. M. (2020). ssGSEA Score-Based Ras Dependency Indexes Derived from Gene Expression Data Reveal Potential Ras Addiction Mechanisms with Possible Clinical Implications. *Sci. Rep.* 10 (1), 10258. doi:10.1038/s41598-020-66986-8

Yi, S., Zhang, Y., Xiong, W., Chen, W., Hou, Z., Yang, Y., et al. (2020). Prominent Immune Signatures of T Cells Are Specifically Associated with Indolent B-cell Lymphoproliferative Disorders and Predict Prognosis. *Clin. Transl Immunol.* 9 (1), e01105. doi:10.1002/cti2.1105

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Zhang, Y., Kwok-Shing Ng, P., Kucherlapati, M., Chen, F., Liu, Y., Tsang, Y. H., et al. (2017). A Pan-Cancer Proteogenomic Atlas of PI3K/AKT/mTOR Pathway Alterations. *Cancer Cell* 31 (6), 820–832. doi:10.1016/j.ccell.2017.04.013

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

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