Mlsp : A bioinformatics tool for predicting molecular subtypes and prognosis in patients with breast cancer

Zhu, Jie

2022

Zhu , J , Kong , W , Huang , L , Wang , S , Bi , S , Wang , Y , Shan , P & Zhu , S 2022 , ' Mlsp : A bioinformatics tool for predicting molecular subtypes and prognosis in patients with breast cancer ' , Computational and Structural Biotechnology Journal , vol. 20 , pp. 6412-6426 . https://doi.org/10.1016/j.csbj.2022.11.017

http://hdl.handle.net/10138/355364
https://doi.org/10.1016/j.csbj.2022.11.017

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

MLSP: A bioinformatics tool for predicting molecular subtypes and prognosis in patients with breast cancer

Jie Zhu [a,b,1], Weikaixin Kong [b,c,1], Liting Huang [a], Shixin Wang [a], Suzhen Bi [a], Yin Wang [a,*], Peipei Shan [a,*], Sujie Zhu [a,1,*]

[a] *Institute of Translational Medicine, The Affiliated Hospital of Qingdao University, College of Medicine, Qingdao University, Qingdao, China*
[b] *Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Finland*
[c] *Institute Sanqu Technology (Hangzhou) Co., Ltd., Hangzhou, China*

ARTICLE INFO

Article history:
Received 30 April 2022
Received in revised form 18 October 2022
Accepted 7 November 2022
Available online 11 November 2022

Keywords:
Breast cancer
Subtypes
RGP score
Web tool

ABSTRACT

The molecular landscape in breast cancer is characterized by large biological heterogeneity and variable clinical outcomes. Here, we performed an integrative multi-omics analysis of patients diagnosed with breast cancer. Using transcriptomic analysis, we identified three subtypes (cluster A, cluster B and cluster C) of breast cancer with distinct prognosis, clinical features, and genomic alterations: Cluster A was associated with higher genomic instability, immune suppression and worst prognosis outcome; cluster B was associated with high activation of immune-pathway, increased mutations and middle prognosis outcome; cluster C was linked to Luminal A subtype patients, moderate immune cell infiltration and best prognosis outcome. Combination of the three newly identified clusters with PAM50 subtypes, we proposed potential new precision strategies for 15 subtypes using L1000 database. Then, we developed a robust gene pair (RGP) score for prognosis outcome prediction of patients with breast cancer. The RGP score is based on a novel gene-pairing approach to eliminate batch effects caused by differences in heterogeneous patient cohorts and transcriptomic data distributions, and it was validated in ten cohorts of patients with breast cancer. Finally, we developed a user-friendly web-tool (https://sujiezhulab.shinyapps.io/BRCA/) to predict subtype, treatment strategies and prognosis states for patients with breast cancer.

1. Introduction

Breast cancer is the most common malignancy in women all over the world. In 2018, there were 270,000 newly diagnosed cases and more than 40,000 new deaths from breast cancer in the United States [1,2]. In the past decades, mortality of breast cancer has decreased mainly due to early detection and more effective systemic treatments [3]. However, most patients with breast cancer are often only diagnosed at a middle or late disease stage [4].

Breast cancer patients are currently stratified using the histologic classification and AJCC staging system [5,6]. However, given that breast cancer is a very heterogeneous disease, showing large differences in aggressiveness and response to therapy, patients with similar clinical features may have very different prognoses [7]. Therefore, it is necessary to consider many other factors to facilitate and improve patients surveillance and treatment. In clinical practice, the immunohistochemical markers estrogen (ER), progesterone (PR), and human epidermal growth factor receptor2 (HER2) are used to guide diagnosis and treatment decisions [8]. Earlier studies based on gene expression profiling suggest that breast cancer can be stratified into five main molecular subtypes: luminal A, luminal B, HER2-enriched, basal-like, and normal-like [7,9]. The molecular classification of breast tumors based on gene expression patterns has also been successfully translated into tests to support clinical decisions [8].

The tumor microenvironment has also been linked with prognosis in breast cancer. High infiltration of tumor-infiltrating lymphocytes (TILs) is associated with prognosis in many cancer types, including breast cancer [10–17]. Furthermore, numerous studies reported that TILs also play an important role in cancer initiation and progression [18–20]. A better understanding of the immune activity of TILs in breast cancer would provide clinicians with more accurate information on their patients' prognoses.

[*] Corresponding authors.
E-mail addresses: wangyin@qdu.edu.cn (Y. Wang), shanpeipei@qdu.edu.cn (P. Shan), zhusujie@bjmu.edu.cn (S. Zhu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2022.11.017

Increased genomic instability (GI) is also a relevant feature of breast tumors, both at somatic DNA copy-number alterations (SCNA) levels and point mutation [21,22]. These two processes (TILs and GI) play an important role in activating oncogenes or inactivating tumor suppressors genes [19,23], therefore potentially also helping elucidate breast cancer biology.

However, the development and application of a prognostic gene signature have limitations, such as inconsistent data formats and batch effects between different profiling platforms. In addition, data sets from various patient cohorts might also have been processed with different data methods, leading to suboptimal signatures. Therefore, it is necessary to design a more robust prognostic model that can eliminate the influence of batch effects.

Currently, most of the available prognostic models have been based on single pathway genes [24–27]; however, there is a wide appreciation that cancer development involves multiple signaling pathways, including cell cycle, immune, and metabolic pathways. In this study, we identified three clinically relevant subtypes of breast cancer based on multiple cancer-related signaling. To better understand three subtypes of breast cancer, integrative multi-omics analysis is used to explain the biological processes contributing to breast cancer aggressiveness, recurrence, and progression in The Cancer Genome Atlas (TCGA). The molecular characterization of breast tumors can both help the development of new therapies and the selection of patients for the appropriate therapies. Combined with PAM50 subtypes, we identified 15 subtypes of breast cancer and predicted medication guidance for distinct subtypes using L1000 platform [28] and mixed machine learning methods, namely univariate COX regression and unsupervised clustering. Then we wanted to establish a robust prognosis model for breast cancer patients. Based on the gene pairing approach, we propose a novel analytic approach to design a robust prognostic model, which we have validated in ten external test sets. Finally, we designed a machine learning guided molecular subtypes and prognosis (MLSP) prediction platform to help aid clinical decisions from diagnosis to treatment of breast cancer patients.

2. Methods

2.1. Breast cancer datasets

We performed a bioinformatic analysis on publicly available transcriptomic and genomic data from the 11 breast cancer cohorts (Supplementary Table 1). The TCGA is hosted by the NCI's Genomic Data Commons (GDC) https://portal.gdc.cancer.gov/ and contains RNA-seq, copy number, mutation and clinical data. These data were downloaded from TCGA GDC, after combing the clinical information with expression data, we obtained expression daya of 1,900 patients and 113 control sample (Supplementary Tables 1-4).The METABRIC cohort was accessed via the cbioportal website (https://www.cbioportal.org/). Data from other cohorts (GSE20685, GSE21653, GSE17705, GSE11121, GSE7390, GSE20711, GSE1456, GSE31448, GSE4922) were obtained from the GEO database (http://www.ncbi.nlm.nih.gov/gds).

We used ANOVA test and Chi-square test to do clinical information analysis across these cohorts. As for missing values in expression data, we deleted the genes with any missing value across these 11 cohorts. And we changed all expression data in these 11 cohorts to log2(expression +1) format. Then, we used TCGA data for clustering, development of the signature, and training of the prognosis model. The METABRIC cohort samples, which contained both complete normalized RNA microarray profiling on the Illumina HT-12 v3 array [29] and clinical data (1981 fresh-frozen primary breast cancer samples), were used for constructing a nomogram.

2.2. Consensus clustering

We retrieved the queries including "breast cancer", "prognosis signature" and "breast cancer prognosis model" in PubMed, after we manually pick the articles containing the genes with potential prognostic value, and there are 34 studies left containing 183 genes (Supplementary Table 5) [30–63], noting that all of them are published within two years. The key reason why we chose papers within two years is that the accuracy of prognostic models have improved in the past two years since the researchers tend to verify the prediction model in many cohorts instead of just using one cohort, so the genes in these papers are relatively more accurate, which enabled our next analysis. Furthermore, since the studies related to prognosis of BRCA are too much, to avoid the curse of dimensionality, we focused on the studies within two years. Then, using univariate survival analysis in the TCGA-BRCA cohort, 65 genes were identified to be associated with prognosis ($p < 0.05$, Supplementary Table 6). Based on the expression of 65 genes, the unsupervised k-means consensus clustering was performed to divide the patients into different clusters, with 50 repetitions. Notably, 80 % of subsampling will be repeated for each time, and k-varying was seting from 2 to 10 clusters. The optimal number of clusters was determined according to the cumulative distribution function (CDF), which contained the information about the corresponding empirical cumulative distribution. Finally, we identified three cluster ($k = 3$), namely Cluster A, B and C. During this process, the R packages Consensus ClusterPlus was used to perform clustering analysis [64]. To verify the importance of these 65 genes, this method was also used to perform clustering in the Pan cancer datasets (33 kinds of cancers in TCGA database), since that proved these genes are closely related to the outcomes of cancer patients to some degree if the patients can be divided into different clusters with different survival in more than one type of cancer based on these genes.

2.3. Genomic instability and somatic copy-number alterations (SCNA) analysis

The score of genomic instability (GI) and somatic copy-number alterations (SCNAs) in the TCGA dataset were obtain from the previous study [65]. After matching the other clinical information, such as the age, gender, stage,.et, there were 979 patients left, noting that these 979 patients contained all the information, including age, gender, stage, T/M/N, status, Cluster, focal-level SCNA score, chromosome-level SCNA score, arm-level SCNA score, Chromosome/arm score, overall score and GI, as the summary data of these information above was shown in Supplementary Table 7.

2.4. Gene set variation analysis (GSVA) and functional annotation

To explore the biological mechanisms underlying the behavior of distinct subtypes of breast cancer and the degree of enrichment of the KEGG pathways, we used gene set variation analysis (GSVA) [66] and GSVA R packages. The main reason why we performed the GSVA instead of GSEA is that most GSE methods, such as GSEA, are supervised and population based, in that they compute an entichment score per gene set to seacribe the entire data set, modeled on a phenotype (such as patient vs control). More importantly, GSEA method did not take the gene correlations into account, which might lead to an increases number of false-positive gene sets, thus increasing the number of false-positive gene sets with respect to GSEA. GSVA was developed to these major drawbacks above, and also utilizes density estimates for evaluating sample-wise enrichment, but by omitting phenotypic information, which enables more broader applications with higher statistical power than the other methods[66]. Specifically, the GSVA method includes four steps, including evaluating the gene expression level statistic, ranking order per sample, using the Kolmogorov-Smirnov (KS) like random walk statistic, and using two approaches to turn the KS like random walk stastic into an enriment statistc (ES, also called GSVA score). The gene sets of "c2.cp.kegg.v6.2.symbols" were downloaded from MsigDB datasets to run GSVA analysis. An adjusted P of <0.05 was considered statistically significant. To clarify the differential signaling pathway among three subtypes of BRCA, the intersection pathways among cluster A compared with cluster B, cluster A compared with cluster C and cluster B compared with cluster C were selected, and displayed in a heatmap.

2.5. Single sample gene expression pathway analysis

To further investigate gene programs enriched by each TCGA breast samples, we employed a single sample gene set enrichment analysis (ssGSEA) method from the GSVA R package [66]. This method uses the gene expression data and the gene set as input, noting the phenotype information is not necessary, and then ranks the inputed gene based on gene expression, finally calculates the Enrichment score (ES) based on the input data. The gene set for marking infiltration of immune cell type in the tumor microenvironment, obtained as described before [67–69], which contained various human immune cell subtypes, such as activated CD8 T cell, activated dendritic cell, macrophage, natural killer T cell, and regulatory T cell and so on. The one-way ANOVA test was used to compare the infiltration of immune cell among distinct three subtypes. The enrichment scores calculated by ssGSEA analysis were utilized to represent the relative abundance of each infiltrating cell in the tumor microenvironment in each sample.

2.6. PD1/CTLA4 response prediction and PAM50 subtyping.

To predict the immunotherapy response of patients with distinct subtypes of breast cancer, we downloaded the immunotherapy prediction information from TCIA database(https://tcia.at/home) database (Supplementary Table 8), which provides results of comprehensive immunogenomic analyses of next generation sequencing data (NGS) data for 20 solid cancers from the TCGA and other data sources. The immunophenoscore (IPS) can be used to predict the response to the immunotherapy agents PD1 and CTLA4. The information of PAM50 subtyping was downloaded from TCIA (Supplementary Table 10), and we combined the PAM50 subtyping with our subtypes identified in this studies to obtain more refined subtypes of BRCA as well as closing to the clinic (Supplementary Table 10), as the PAM50 subtyping is commonly used in clinic.

2.7. Tumor mutation burden (TMB) analysis

The tumor mutation burden (TMB) has been shown to be associated with the efficacy of immunotherapy in multiple cancer types. Therefore, we estimated the TMB value of breast cancer samples by calculating the number of gene mutations per million bases, which was calculated using the TMB function in "maftools" packages in R software, and the result of TMB value was shown in Supplementary Table 9. Mutation data were downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov/) and intersected with samples from the expression dataset. Then, we obtained samples from 974 patients with breast cancer for which both expression and mutation data were available. For these patients, we used the "maftools" packages in R to plot waterfall charts and mutation gene cloud charts, and identified differentially mutated genes (DMGs) between different subtypes of breast cancer. The one-way ANOVA test was used to compare the TMB values of distinct three subtypes. Given the significant survival differences between cluster A and cluster C, we further explored the DMGs between these two clusters and identified three genes related to survival. The significance criteria for determining DMGs, as well as for the survival and gene mutation data was set as a p value of<0.05.

2.8. Gene feature selection of each subtype of BRCA and drug analysis

To more precisely define subtypes of breast cancer, we combined 5 kinds of PAM50 subtypes with our new three clusters and identified 15 subtypes. To identify the gene features of each specific subtype, the empirical Bayesian approach of limma R package was applied. Briefly, we determined differentially expressed genes (DEGs) between each of the 15 subtypes and normal breast tissues [70]. The significance criteria for determining DEGs was set as an adjusted P value <0.001 and $|\log_2 FC| > 1.5$. Then, the protein–protein interaction (PPI) network was constructed for these DEGs by Cytoscape based on the STRING database, with the corresponding topologies were shown in Supplementary Table 11, including the number of nodes, edges and average degree etc. Next, the Degree method (Node connect degree) in Plug-in CytoHubba were used to select the key genes in PPI (Supplementary Table 12). These genes were input of L1000 [28] (https://clue.io/), a tool used to screen drugs that can reverse gene expression from disease state to healthy state. And these drugs were regarded as effective drugs for special disease. In our research, drugs with CMap connectivity (tau) score [28] of < -0.9 were selected and included in our recommendation list if the drug was already approved by the Food and Drug Administration (FDA) for treatment of breast cancer (Supplementary Table 13).

2.9. Prognostic model building

To construct a robust model for predicting the prognosis of patients with breast cancer, we performed a gene-paired method to eliminate batch effect using 65 prognosis related genes. In our previous study, we have verified the robustness of gene-paired method, and it is an effective way to reduce the batch effect without changing the distribution of raw data, so this method can work both for RNAseq and Microarray data [71]. If the expression of gene A > expression of gene B, then the feature "Gene A| Gene B" was marked as 1; otherwise it is marked as 0. In addition, if the expression level of gene A in all of the samples was higher than that of gene B, then the feature Gene A|Gene B was marked as 1 in all of the samples. Such features do not contain classification information, as it only contain the 0 and 1 information, and, therefore, gene pairs whose frequency of the "1" label in the training set was<0.1 or greater than 0.9 were deleted. Based on the gene pair method, we identified 891 gene-pairs as features, which were further reduced to 34 gene-pairs using the univariate Cox regression ($p < 0.001$) and LASSO regression. Then, multivariate Cox regression was used to construct a Robust Gene-Paired (RGP) signature containing 16 gene-pairs (Supplementary Table 14). In this process, we used "glmnet", "survival", and "survminer" R packages.

2.10. Prognostic model validation

We first performed Kaplan-Meier survival analysis (Log-rank test) based on the RGP signature in the TCGA-BRCA cohort (training set) and the other 10 test sets. The different types of survival time in these cohorts were used in this process (Supplementary

Table 1), including overall survival time (OS), disease-free survival time (DFS), relapse-free survival time (RFS), and distant metastasis-free survival time (DMFS). To further examine the robustness of the RGP signature, we compared its performance against other six signatures used as baseline models, namely the autophagy related gene signature (ARG) [50], ferroptosis related gene signature (FRG)[41], macrophage marker gene signature (MMGS) [49], nuclear receptors related gene signature (NR) [55], DNA repair related gene signature (DRG) [60], and hypoxia related gene signature (HRG) [42], by analyzing the area under the receiver operating characteristic (ROC-AUC) curves for predicting patient survival. The 3-, 5-, and 7-year AUC-ROC values were then used to compare the prognostic ability of the models using a paired *t*-test (Table 1).

2.11. Nomogram construction and validation

To construct a nomogram, we divided the METABRIC cohort, which contains various types of clinical information, into a training and a test set. To explore whether the RGP score is an independent risk factor among other clinical features, we conducted a univariate and multivariate independent prognostic analysis using the METABRIC training set. Then, we selected the significant factors in both univariate and multivariate independent prognostic analysis to establish a nomogram using the training set. The nomogram performance was examined in the training and test set by analysis of the AUC-ROCs and calibration curves.

2.12. Design of a web app

To use a nomogram based on our clustering results and RGP score, and provide clinical guidance to help inform patient's prognosis and therapeutics, we designed a web app (https://sujiezhu-lab.shinyapps.io/BRCA/). As the input, we used the expression of 65 genes and 3 types of clinical information (age, PAM50 and Nottingham prognostic index), with the output being treatment-related information, including the subtypes breast cancer patients belonged to, potential drugs, and the patient's survival rate.

2.13. Pan-cancer analysis of overall survival

To explore the value of the RGP score to other types of cancers, data from 33 kinds of cancer from the TCGA were used for pan-cancer analysis. The TCGA pan-cancer data were downloaded from the UCSC Xena database (https://xena.ucsc.edu/), and included 10,071 cancer patients with complete expression profiles, mutation data, overall survival (OS) and survival status. These expression data were used to calculate the RGP score and perform Cox prognostic analysis to calculate the hazard ratio. Kaplan-Meier survival analysis (Log-rank test) was also used to evaluate the prognosis value of the RGP score in the pan-cancer datasets.

3. Results

3.1. Multiple-Pathway-Based stratification of breast cancer

Breast cancer is a heterogeneous disease characterized by distinct transcriptional patterns, biology, and immune composition [72–74]. Here, we selected 183 genes from distinct signaling pathways based on a literature search related to breast cancer prognosis [30–63] (Fig. 1, Supplementary Fig. 1, Supplementary Table 5). We leveraged data from the TCGA cohort, which included 1,090 patients and offered the most comprehensive clinical annotation (Supplementary Table 2). Next, we identified 65 genes with prognostic value by univariate Cox regression ($p < 0.05$) which we analyzed further (Supplementary Table 6). In addition, the sample screening process was shown in Supplementary Fig. 2. In collected cohorts, the clinical information exhibited significant difference in age, OS time, DFS time, PAM50 subtyes, radiation therapy and pharmaceutical therapy (Supplementary Figs. 3 and 4, ANOVA test and Chi-square test, $p < 0.001$). The models trained and evaluated based on different clinical patterns would be robust and have a wide application range.

Several studies have previously identified between two to four distinct transcriptional subtypes of breast cancer based on analysis of single pathways [45,52,75–77]. For more refined clustering analysis, we performed unsupervised k-means consensus clustering of the TCGA-BRCA cohort using multiple key tumor processes (Supplementary Table 6). The optimal number k of clusters was determined to be three based on the area under the curve of the

Table1
Comparison of the RGP score with other signatures.

	RGP Signature	ARG Signature	FRG Signature	MMGS Signature	NR Signature	DRG Signature	HRG Signature
METABRIC	0.590 ± 0.018	0.594 ± 0.023	0.509 ± 0.006*	0.604 ± 0.024#	0.520 ± 0.002*	0.528 ± 0.005*	0.601 ± 0.012
p value		0.234	0.005	0.042	0.018	0.029	0.067
GSE20685	0.650 ± 0.010	0.628 ± 0.016	0.634 ± 0.035	0.648 ± 0.007	0.532 ± 0.017*	0.599 ± 0.033	0.569 ± 0.008*
p value		0.132	0.284	0.397	0.010	0.106	0.009
GSE21653	0.698 ± 0.026	0.551 ± 0.019*	0.579 ± 0.026*	0.639 ± 0.002*	0.506 ± 0.018*	0.630 ± 0.017*	0.646 ± 0.002
p value		0.001	0.039	0.040	0.008	0.037	0.061
GSE17705	0.687 ± 0.020	0.621 ± 0.057	0.571 ± 0.012*	0.530 ± 0.021*	0.486 ± 0.058*	0.588 ± 0.027*	0.602 ± 0.021*
p value		0.083	0.004	0.001	0.030	0.043	0.034
GSE11121	0.703 ± 0.008	0.573 ± 0.022*	0.645 ± 0.029*	0.446 ± 0.016*	0.483 ± 0.038*	0.677 ± 0.075	0.544 ± 0.026*
p value		0.003	0.040	0.002	0.010	0.324	0.003
GSE7390	0.616 ± 0.040	0.600 ± 0.017	0.618 ± 0.027	0.647 ± 0.007	0.467 ± 0.023*	0.463 ± 0.025*	0.587 ± 0.057
p value		0.244	0.487	0.193	0.012	0.009	0.103
GSE20711	0.687 ± 0.072	0.529 ± 0.033*	0.402 ± 0.066*	0.634 ± 0.017	0.538 ± 0.029	0.608 ± 0.020	0.586 ± 0.038
p value		0.027	0.049	0.158	0.076	0.083	0.151
GSE1456	0.573 ± 0.057	0.498 ± 0.028	0.678 ± 0.007	0.535 ± 0.061	0.556 ± 0.048	0.564 ± 0.045	0.627 ± 0.059
p value		0.166	0.073	0.224	0.350	0.454	0.077
GSE31448	0.677 ± 0.026	0.560 ± 0.014*	0.567 ± 0.027*	0.620 ± 0.005*	0.496 ± 0.016*	0.620 ± 0.014	0.622 ± 0.003
p value		0.004	0.047	0.033	0.010	0.060	0.056
GSE4922	0.555 ± 0.018	0.602 ± 0.030#	0.543 ± 0.007	0.518 ± 0.001*	0.547 ± 0.015	0.484 ± 0.005*	0.592 ± 0.008
p value		0.025	0.256	0.044	0.380	0.008	0.071

Footnote: The standard deviation was calculated over the 3-, 5-, and 7-year ROC-AUC values. p values were obtained by paired *t*-test comparing the RGP Signature and other six signatures in 3-,5- and 7-year three time points. *means ROC-AUCs of the RGP signature are significantly higher than values of this labeled signature; #means ROC-AUCs of RGP signature are significantly lower than values of this labeleed signature significantly.

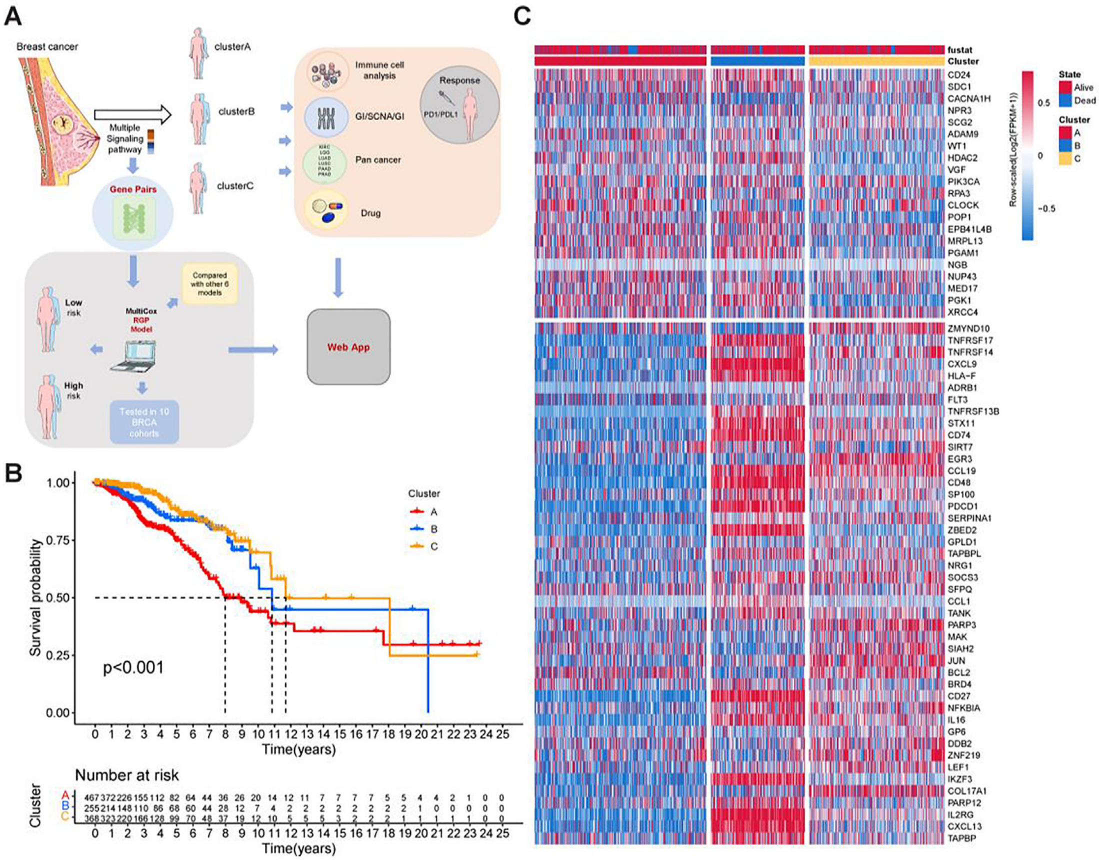


Fig. 1. A. Work design. B. Kaplan-Meier for the survival of clusters A, B and C identified in the TCGA-BRCA cohort. C. Expression heat map of 65 prognosis related genes among cluster A, B and C. The upper part (CD24-XRCC4) shows unfavorable genes with HR > 1; the lower part (ZMYND10-TAPBP) shows favorable genes with HR < 1.

consensus distribution function (CDF, Supplementary Fig. 5), including 467 patients in cluster A, 255 patients in cluster B, and 368 patients in cluster C (Fig. 1B). We next examined whether different subtypes were associated with different prognoses and found that patients in cluster C had a higher survival rate (Fig. 1B). There was significant distinction existed on the 65 genes transcriptional profile among three different subtypes (Fig. 1C). Cluster A was characterized by high expression of unfavorable prognosis genes and lower favorable prognosis genes; cluster B exhibited high expression of favorable prognosis genes; cluster C exhibited significant lower expression of unfavorable prognosis genes.

To elucidate the transcriptional pathways driving subtypes in breast cancer, we performed pathway single-sample gene set variation analysis (GSVA) to identify differentially expressed gene sets across subtypes (Fig. 2A). Cluster A and cluster B displayed enrichment in pathways associated with cell cycle, mismatch repair, and DNA replication. Interestingly, cluster B exhibited enrichment in pathways associated with immune cells (cell adhesion, inflammation, T cell signaling and NK cell signaling) and immune function (antigen presentation and interferon-gamma response). Cluster C was enriched in stromal and carcinogenic activation pathways such as MAPK, MTOR, and ERBB signaling pathway. All subtypes, with exception of cluster A, showed enrichment in genes involved in the apoptosis signaling pathway, which is likely related to higher survival rates in these clusters.

Cancer genomic instability (GI), somatic copy-number alterations (SCNA) played an important role in increasing the adaptive potential of the tumors and poor prognosis [78]. The SCNA score as a representation of the level of SCNA occurring in a tumor was calculated at three different levels: focal, arm, and chromosome level, and the overall score calculated from the sum of all three levels [65,79,80]. We found that tumors from cluster A showed remarkably high levels of SCNA and GI (Fig. 2A, Supplementary Table 7). Altogether these findings highlight the diversity of underlying biology and genomic alterations.

3.2. Breast cancer subtypes based on Multiple-Pathway analysis exhibit distinct immune features

Our findings based on GSVA analyses revealed three subtypes of breast cancer with distinct levels of immune activation (Fig. 2A). To further examine the characteristics of the tumor microenvironment (TME) in terms of immune cell infiltration, we performed ssGSEA on these three subtypes (Fig. 2B). The three subtypes showed significantly distinct characteristics of cell infiltration in the tumor microenvironment. Cluster A exhibited an immune-desert phenotype which was characterized by the suppression of immunity. Cluster B exhibited an immune-excluded phenotype which was characterized by activation of different types of immune cells, including stromal immune cells, such as immunosuppressive MDSCs. Cluster C exhibited an immune-inflamed phenotype which was characterized by moderate immune activation (Fig. 2B) and is associated with higher rates of survival (Fig. 1B). It is noted that there are higher TIL level in Cluster B and Cluster C, emphasizing that the patients with higher TIL tend to better clinical outcomes, which is coincident with previous studies (Ref: TIL related papers).

Patients with higher expression of immune checkpoint markers (such as PDL1) might be more responsive to immunotherapies

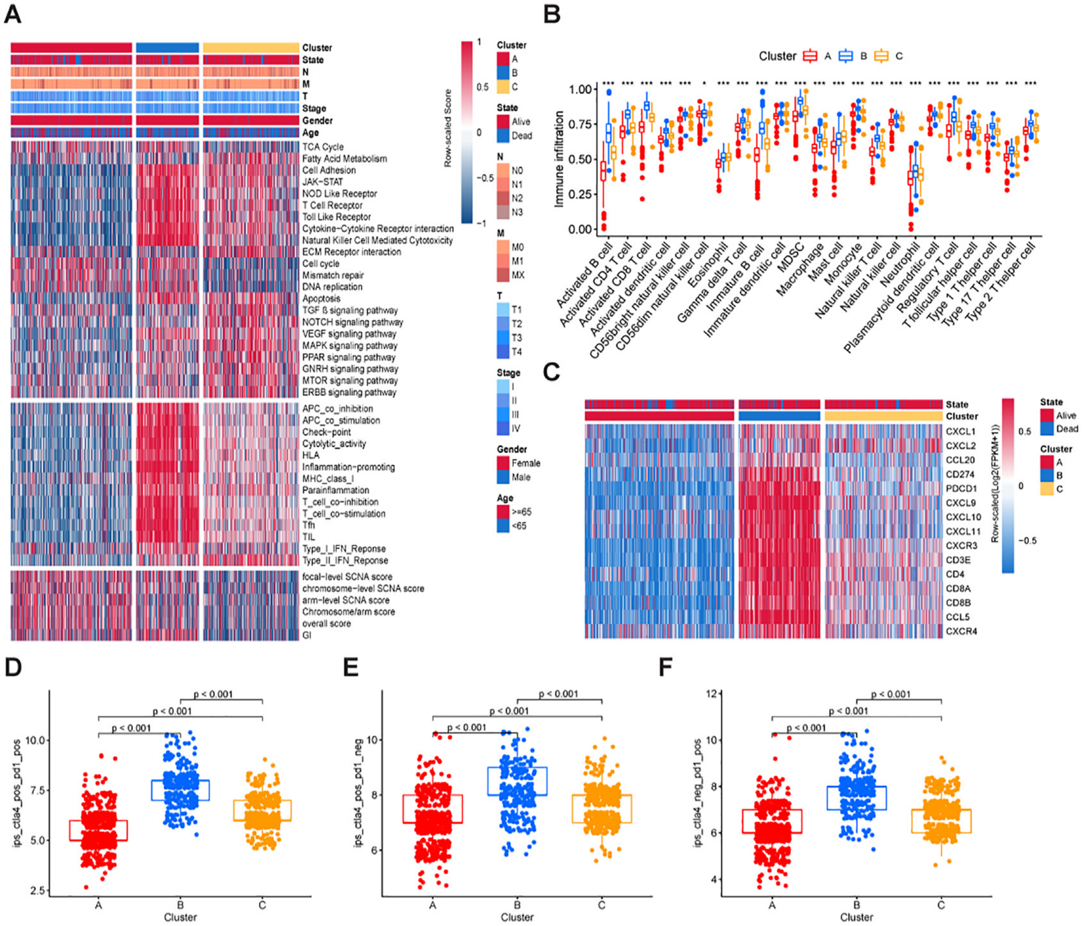


Fig. 2. A. The biology behavior and genomic alterations features among three BRCA subtypes. The upper part of the heat map represents the key pathways derived from GSVA analysis using expression data; the color represents the GSVA score after scaled. The middle part represents the key immune processes derived from the ssGSEA analysis using expression data; the color represents the ssGSEA score after scale. The lower part represents overall, focal-level and chromosome/arm-level SCNA score which was downloaded from the previous study with using mutation data and copy-number alteration. B. ssGSEA analysis of three subtypes (Avona test). * $p < 0.05$, **$p < 0.01$, ***$p < 0.001$. C. Expression of hot and cold tumor marker genes in the three subtypes. D. Immunotherapy prediction downloaded from TICA database. The y-axis values represent mixed score of sensitivity to CTLA4 immunotherapy and sensitivity to PD1 immunotherapy. E. Immunotherapy prediction downloaded from TICA database. The y-axis values represent mixed score of sensitivity to CTLA4 immunotherapy and insensitivity to PD1 immunotherapy. F. Immunotherapy prediction downloaded from TICA database. The y-axis values represent mixed score of insensitivity to CTLA4 immunotherapy and sensitivity to PD1 immunotherapy.

[81]. Therefore, we next examined the expression of immune checkpoint genes (Supplementary Fig. 6). Of the three clusters, tumors from cluster A showed the lowest expression of checkpoint genes, and those from cluster B showed the highest expression of immune checkpoint markers. Tumors can be categorized into hot or cold according to their response to immunotherapy[82]. To investigate which subtype might be more responsive to immunotherapy, we examined the expression of genes associated with a hot and cold phenotype in the three subtypes of breast cancer (Fig. 2C) and found that cluster A displayed low expression of genes, while cluster B was characterized by low expression of cold tumor related genes and high expression of hot tumor related genes, Tumors in cluster C exhibited moderate expression of both hot and cold-associated genes.

These findings show that tumors in cluster B exhibit higher levels of immune cell and immune checkpoint genes, and of genes associated with a hot tumor phenotype, suggesting that cluster B tumors might be good candidates for immunotherapy. To test this hypothesis, we obtained the prediction of immunotherapy effect among three subtypes of breast cancer (Fig. 2D-F) from TICA database. Consistently with our results, tumors from cluster B were more responsive to PD1/CTLA4 therapies, whereas those from cluster A showed were less responsive.

3.3. Subtypes based on Multiple-Pathway analysis show distinct tumor mutation burden

Gene alterations acquired during tumor evolution can be predictive of poor outcomes in patients with breast cancer, as well as resistance to therapy [83]. To better understand the role of genetic alterations in the three subtypes of breast cancer, we further analyzed the distribution differences of somatic mutations. Cluster B subtype presented with a more extensive tumor mutation burden than cluster A and cluster C (Supplementary Table 9), with an average mutation rate of top 10 mutations at 17 % (Supplementary Fig. 7B) versus 13.9 % (Supplementary Fig. 7A) and 14.3 % (Supplementary Fig. 7C), for cluster A and C, respectively. These findings suggest again that tumors in cluster B might be more responsive to immunotherapy, given that tumor mutation burden is associated with immunotherapy response [84]. When selecting subtype-specific mutated genes, we found that tumors in cluster A were characterized with high mutation rates of TP53 and GATA3

(Supplementary Figure 8A), while those in cluster B were characterized by mutations of TP53, PIK3CA and TTN (Supplementary Figure 8B), and those in cluster C were characterized by mutations of PIK3CA and CDH1 (Supplementary Figure 8C). Tumors in cluster C exhibited the lowest mutation rate of TP53, and also had the best prognosis outcomes of the three subtypes.

Mutational processes help drive tumor evolution and genetic complexity. Therefore, we next assessed tumor mutational burden in the three distinct subtypes of breast cancer. Compared with cluster A and cluster C (Fig. 3A), tumor mutation burden was higher in cluster B. Mounting evidence demonstrated that patients with a high tumor mutation burden presented more durable clinical responses to anti-PD-1/PD-L1 immunotherapy, suggesting again that tumors cluster B might be more responsive to immune checkpoint blockade therapies. Pre-clinical studies and clinical trials revealed an association between higher somatic tumor mutation burden, enhanced response with durable clinical benefits to immune checkpoint blockade therapies, and increased long-term survival. Considering that tumors in cluster A and cluster C showed the worst and best prognosis, respectively, and that gene mutation

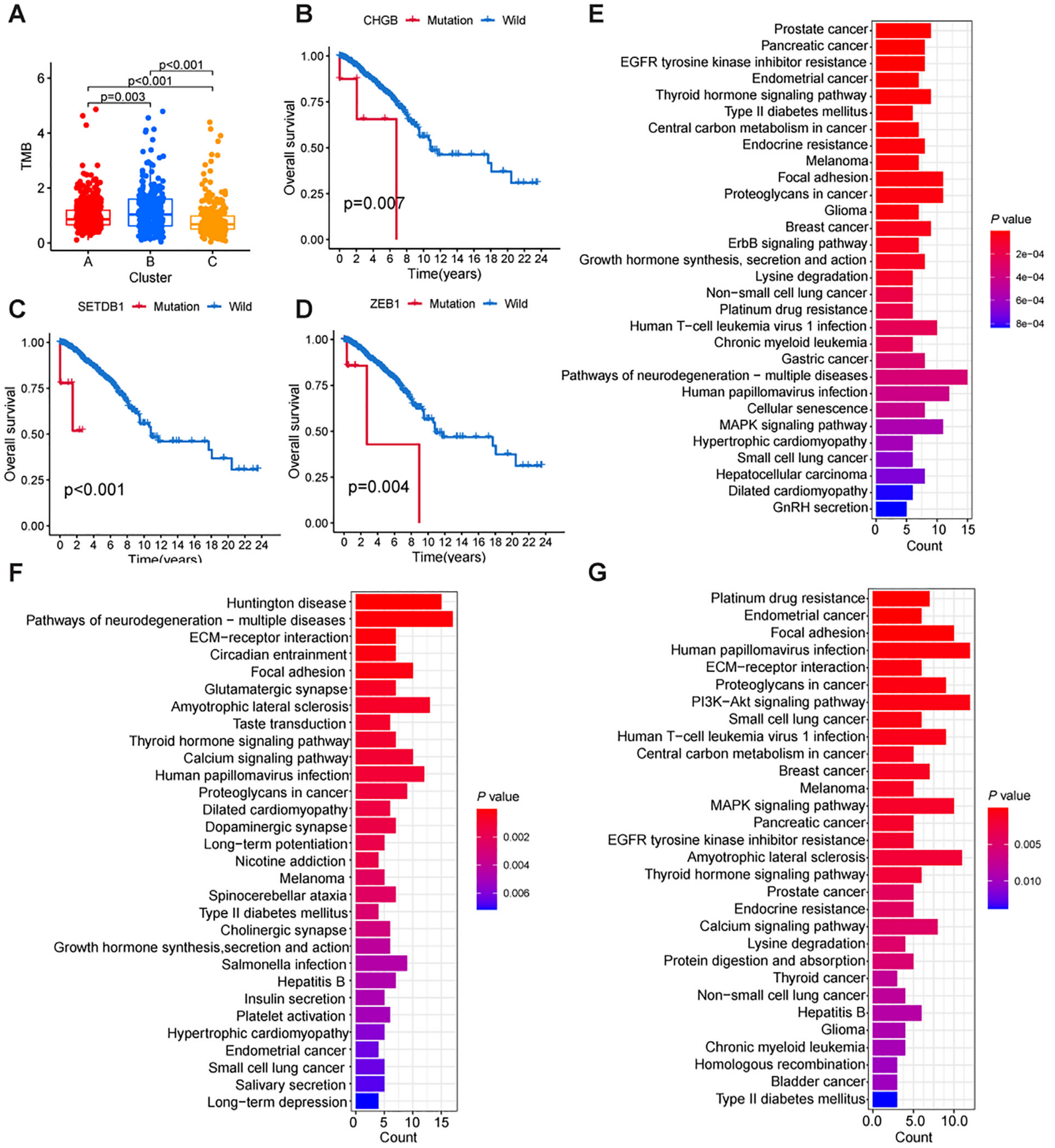


Fig. 3. Mutation analysis results in TCGA-BRCA cohort. A. Comparison of tumor mutation burden among the three subtypes by Wilcoxon test. B. Kaplan-Meier survival curves of the CHGB-mutation and the CHGB-wild groups (Log-rank test). C. Kaplan-Meier survival curves of the SETDB1-mutation and SETDB1-wild groups (Log-rank test). D. Kaplan-Meier survival curves of the ZEB1-mutation and ZEB1-wild groups (Log-rank test). E. GO analysis showing the top 100 most frequent mutations in the A subtype. F. GO analysis showing the top 100 most frequent mutations in the B subtype. G. GO analysis showing the top 100 most frequent mutations in the C subtype.

is related to survival, we analyzed the different mutation rates between cluster A and cluster C. We identified three genes related to survival (CHGB, SETDB1 and ZEB1, Fig. 3B-D), which were all more often mutated in tumors in cluster A.

Mutations can lead to substantial biological changes during tumorigenesis. Therefore we examined the pathways associated with the most prominent gene mutations in the distinct subtypes of breast cancer. We selected the genes which were more often mutated (in the top 200) in the three distinct subtypes and performed a KEGG analysis for these genes (Fig. 3E-G). Tumors in cluster A were enriched for mutations in Erbb, MAPK, PI3K-Akt, and EGFR tyrosine kinase inhibitor resistance pathways (Fig. 3E). Tumors in cluster B exhibited abnormal ECM-receptor interaction (Fig. 3F), while tumors in cluster C exhibited mutations in ECM-receptor interaction, PI3K-Akt, EGFR tyrosine kinase inhibitor resistance, and MAPK signaling pathways (Fig. 3G).

3.4. Drug screening in Multiple-Pathway-Based and PAM50-Based subtypes

According to the PAM50 subtyping system (Supplementary Table 10), breast cancer patients can be subdivided into Luminal A, Luminal B, Her2, Basal, and Normal subtypes. Prognostic analysis for the PAM50 subtypes revealed that patients who belong to the Normal subtype have better survival rates initially and but these quickly decrease at the 10th year (Fig. 4A), similarly to tumors in the Luminal A (LumA) subtype, whereas tumors in the Basal subtype showed an opposite survival trend to those in the Normal sub-

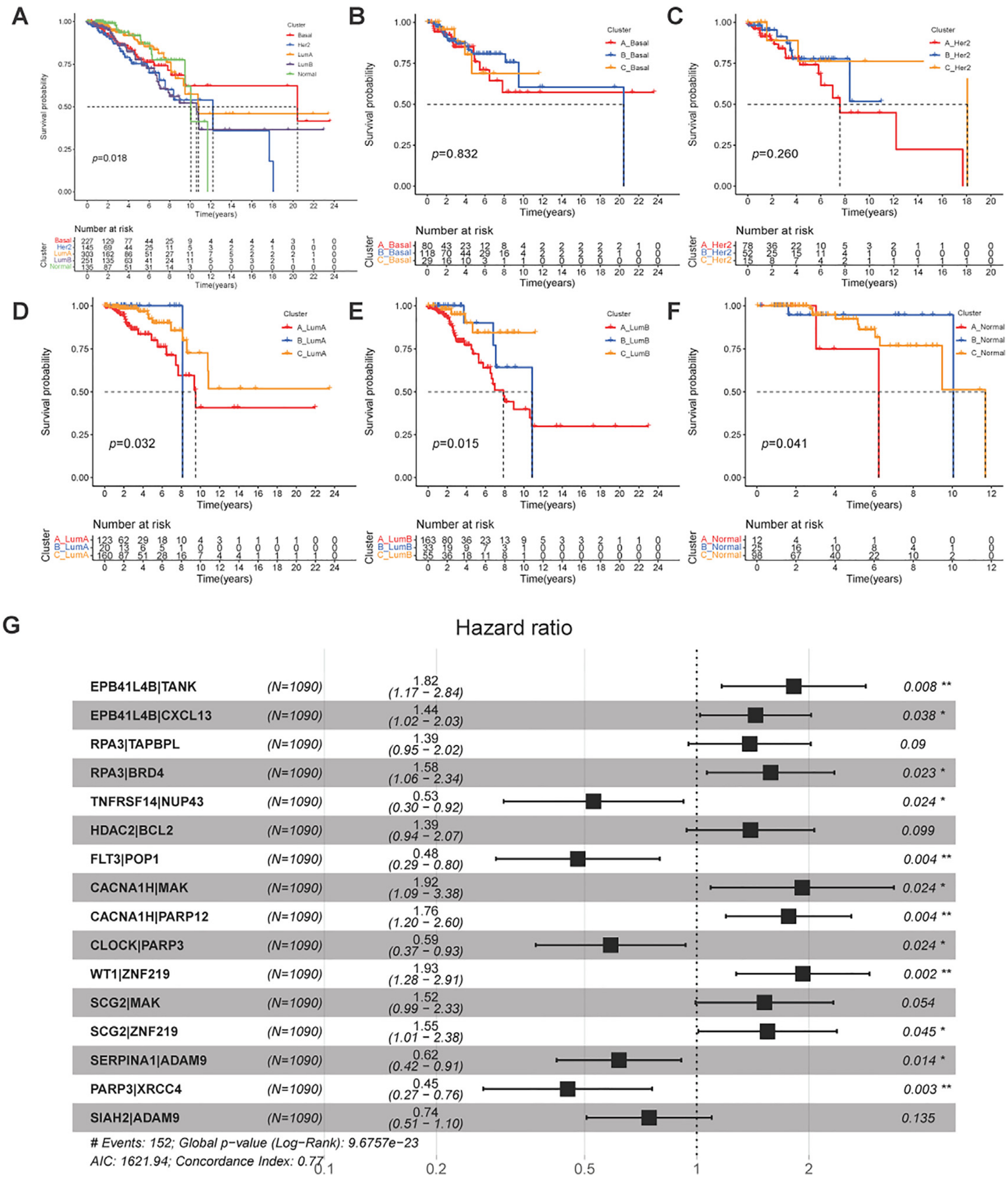


Fig. 4. A. Kaplan-Meier survival curves of 5 different PAM50 types in the TCGA-BRCA cohort (Log-rank test); the number of alive patients is shown in the box below labeled as "Number at risk". B-F. Kaplan-Meier survival curves of different cancer types in the TCGA-BRCA cohort (Log-rank test); the number of alive patients is shown in the box below labeled as "Number at risk". B. Basal type. C. Her2 type. D. LumA type. E. LumB type. F. Normal type. G. The coefficients of RGP signature.

type. Tumors in the Her2 and Luminal B (LumB) subtypes had the worst survival rates before the 10th year. In analyzing the three subtypes of breast cancer identified in our study (Supplementary Figure 9), we found that tumors in cluster A were often from the LumB subtype and less often from the Normal subtype, those in cluster B were often from the Basal subtype, and those in cluster C more often from the LumA subtype and less often from the Basal and Her2 subtypes. Based on the PAM50 subtyping system and our three subtypes, breast cancer patients can be subdivided into 15 clusters, namely cluster A_Luminal A (AA); cluster B_Luminal A (BA); cluster C_Luminal A (CA); cluster A_Luminal B (AB); cluster B_Luminal B (BB); cluster C_Luminal B (CB); cluster A_Basal (ABa); cluster B_Basal (BBa); cluster C_Basal (CBa); cluster A_Her2 (AH); cluster B_Her2 (BH); cluster C_Her2 (CH); cluster A_Normal (AN); cluster B_Normal (BN); and cluster C_Normal (CN). Interestingly, the Normal, LumA and LumB subtypes in cluster A (AN, AA, AB) had worse prognoses than those subtypes in cluster B and cluster C (Fig. 4D-F, $p < 0.05$), confirming the prognostic value of our 3 subtypes for patients with breast cancer.

Our results showed that PAM50 subtyping (LumA, LumB, Basal, Her2 and Normal subtype) was associated with distinct survival among our identified three subtypes of breast cancer. We then further investigated potential drug treatments for the 15 subtypes of cancer (AA, BA, CA, AB, BB, CB, ABa, BBa, CBa, AH, BH, CH, AN, BN, and CN) using CMPA. Connectivity Map (CMAP) is a computational biology screening strategy previously shown to be effective in advancing anti-obesity therapies based on gene signature-based drug screening [85]. Promising candidate drugs should revert the gene signature of the disease of interest compared with controls. Furthermore, the CMAP version2 named L1000 expanded the scope of screening. First, we identified DEGs by comparing transcriptional changes between normal breast samples and the distinct 15 subtypes of breast cancer samples. Then, hub genes were screened using Protein-Protein network (PPI) analysis (Supplementary Tables 11 and 12). The hub genes were used as input to predict drugs that can reverse the expression data from a subtype state to a control state. The drugs with CMap connectivity (tau) score [28] <−0.9 were selected and regarded as effective drugs for this specific breast cancer subtype. Considering the drugs can be used clinically, we then focused on the FDA-approved drugs for breast cancer and the drugs information was listed in Supplementary Table 13. Different drugs were identified for each subtype of breast cancer, further suggesting that each subtype of breast cancer has its unique genomic feature.

3.5. Construction of a prognostic robust gene pair (RGP) score for breast cancer

We next explored whether our identified 65 genes with the potential prognostic value to predict prognosis in the pan cancer datasets. Indeed, our analysis indicates that these genes are associated with the patient's survival ($p < 0.05$, Log-rank test, Supplementary Figure 10).

The 65 prognostic genes were then used for gene pairing (Supplementary Fig. 1), resulting in a total of 2,080 (65*64/2) gene-pairs, 891 of which have a frequency of "gene A > gene B expression" between 10 % and 90 % in the TCGA-BRCA cohort (training set), and were therefore considered to provide sufficient information to be included in the model (Supplementary Fig. 1). By univariate Cox regression ($p < 0.001$) and LASSO regression (Supplementary Figure 11), we then selected 34 gene-pairs. Finally, using multivariate Cox regression, we identified 16 gene-pairs that were associated with survival differences (Supplementary Table 14); these gene pairs include 10 risk factors (HR greater than 1, Fig. 4G) and 6 protective factors (HR < 1, Fig. 4G). The RGP score was calculated as below:

$$\begin{aligned}\mathrm{Sum} = {} & 0.599 \times \mathrm{EPB41L4B|TANK} + 0.363 \times \mathrm{EPB41L4B|CXCL13} \\ & + 0.327 \times \mathrm{RPA3|TAPBPL} + 0.456 \times \mathrm{RPA3|BRD4} \\ & - 0.638 \times \mathrm{TNFRSF14|NUP43} + 0.332 \times \mathrm{HDAC2|BCL2} \\ & - 0.735 \times \mathrm{FLT3|POP1} + 0.652 \times \mathrm{CACNA1H|MAK} \\ & + 0.567 \times \mathrm{CACNA1H|PARP12} - 0.533 \times \mathrm{CLOCK|PARP3} \\ & + 0.656 \times \mathrm{WT1|ZNF219} + 0.418 \times \mathrm{SCG2|MAK} \\ & + 0.437 \times \mathrm{SCG2|ZNF219} - 0.481 \times \mathrm{SERPINA1|ADAM9} \\ & - 0.798 \times \mathrm{PARP3|XRCC4} - 0.295 \times \mathrm{SIAH2|ADAM9} \qquad (1)\end{aligned}$$

$$\mathrm{RGP\ score} = e^{Sum} \qquad (2)$$

The detailed coefficients of the RGP score are shown in Supplementary Table 5. To investigate the prognostic performance of the 16 gene-pair RGP score, we applied it to ten independent breast cancer patient cohorts, i.e cohorts that had not been used in the construction of the RGP score (Supplementary Table 1). Differences in survival time, including overall survival time (OS), disease-free survival time (DFS), relapse-free survival time (RFS), and distant metastasis-free survival time (DMFS) were observed among the different cohorts. We used the median value of each breast cancer cohort as the cutoff to distinguish high from low-risk groups. Notably, despite the heterogeneity of the disease and cohort differences in terms of patient characteristics, follow-up times and transcriptomic platforms, the unified RGP score held value in distinguishing patient survival status of the patients across breast cancer cohorts ($p < 0.05$, Log-rank test, Fig. 5). Indeed, the high risk group was associated with a worse prognosis in all of the cohorts (Fig. 5).

3.6. Evaluation of the RGP score

To confirm the robustness of the paired RGP score, we compared its performance against six other existing prognostic signatures in breast cancer (ARG; FRG; MMGS; NR; DRG; and HRG signatures) [41,42,49,50,55,60] using AUC-ROC analyses (Supplementary Figure 12 and Supplementary Figure 13). The *t*-paired test was used to perform the comparison. These results (Table 1) showed that paired RGP score provided a robust and overall the most accurate prognostic risk score for breast cancer patients (Table 1).

We further performed pan-cancer analysis using survival information of 33 types of cancers in TCGA, including 2 hematological cancers and 31 solid tumors. The RGP score showed significant prognostic value in 12 different types of cancer ($p < 0.05$, Wald's test, Fig. 6A) using Cox regression. Kaplan-Meier survival analysis indicated that the RGP score had significant prognostic survival value in seven types of cancer ($p < 0.05$, Log-rank test, Fig. 5A and Fig. 6B-G).

3.7. Construction and validation of a nomogram based on the RGP score

To design the RGP score easy to use, a nomogram including different types of clinical information was constructed in the METABRIC. The METABRIC cohort was randomly divided into a training (n = 1,132, Supplementary Fig. 1) and a testing set (n = 567, Supplementary Fig. 1). We first performed the univariate independent analysis in the training set based on the RGP score and other clinical information, namely PAM50 (containing five factors), Nottingham Prognostic Index (NPI), and age (Fig. 7A). We retrieved six significant factors (Fig. 7A) which were then used to perform multivariate independent prognostic analysis in the training set to eliminate correlations among factors. Then we got four significant factors (Fig. 7B) which we used to construct a nomogram (Fig. 7C).

To evaluate the nomogram, we predicted the survival state of patients 3-,5- and 7- time points and drew related ROC and calibra-

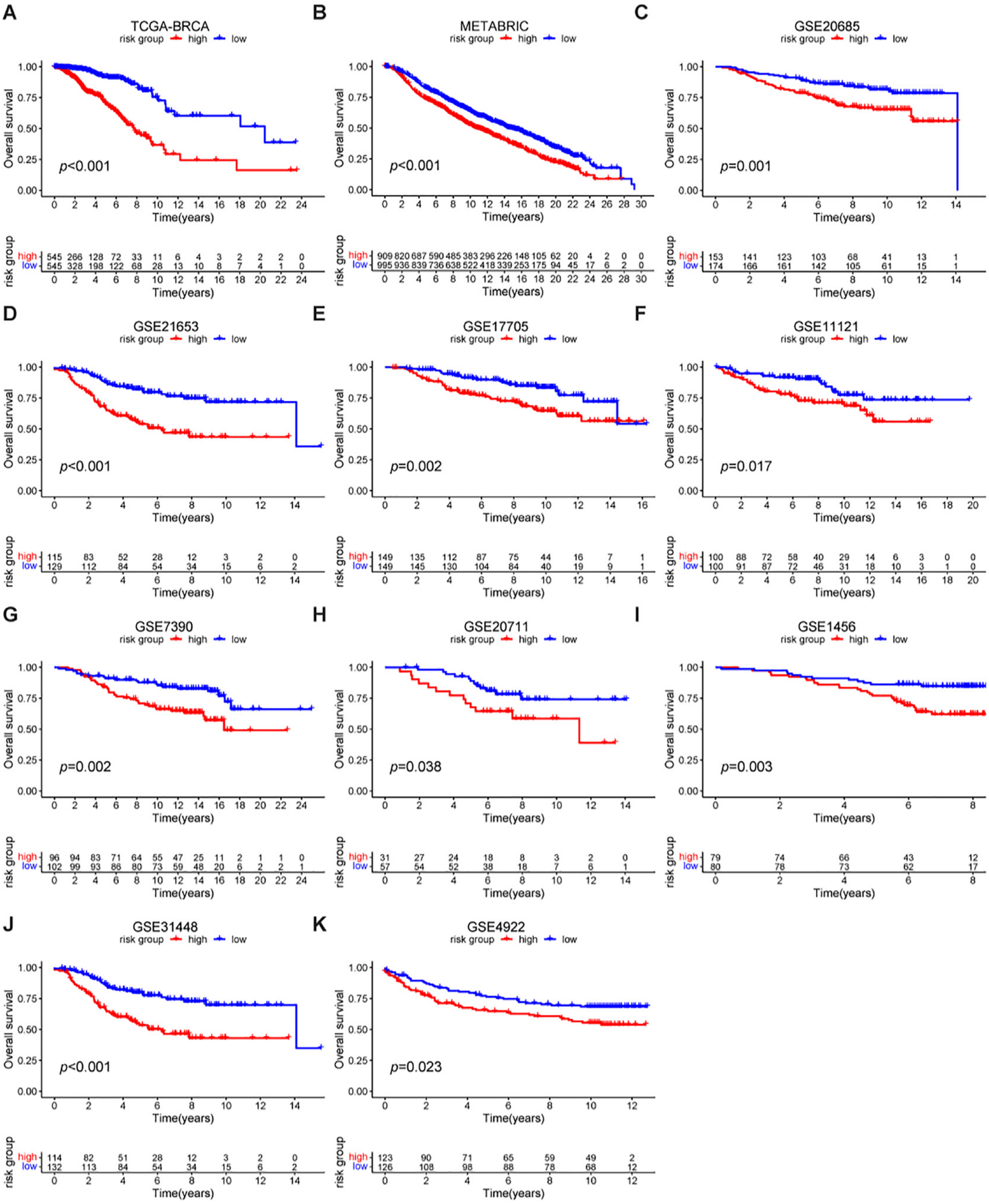


Fig. 5. Kaplan-Meier survival curves of the RGP signature in training and 10 test sets (Log-rank test); the number of alive patients is shown in the box below labeled as "Number at risk". A. TCGA-BRCA cohort (training set). B. METABRIC cohort. C. GSE20685. D.GSE21653. E. GSE17705. F. GSE11121. G. GSE7390. H. GSE20711. I. GSE1456. J. GSE31448. K. GSE4922.

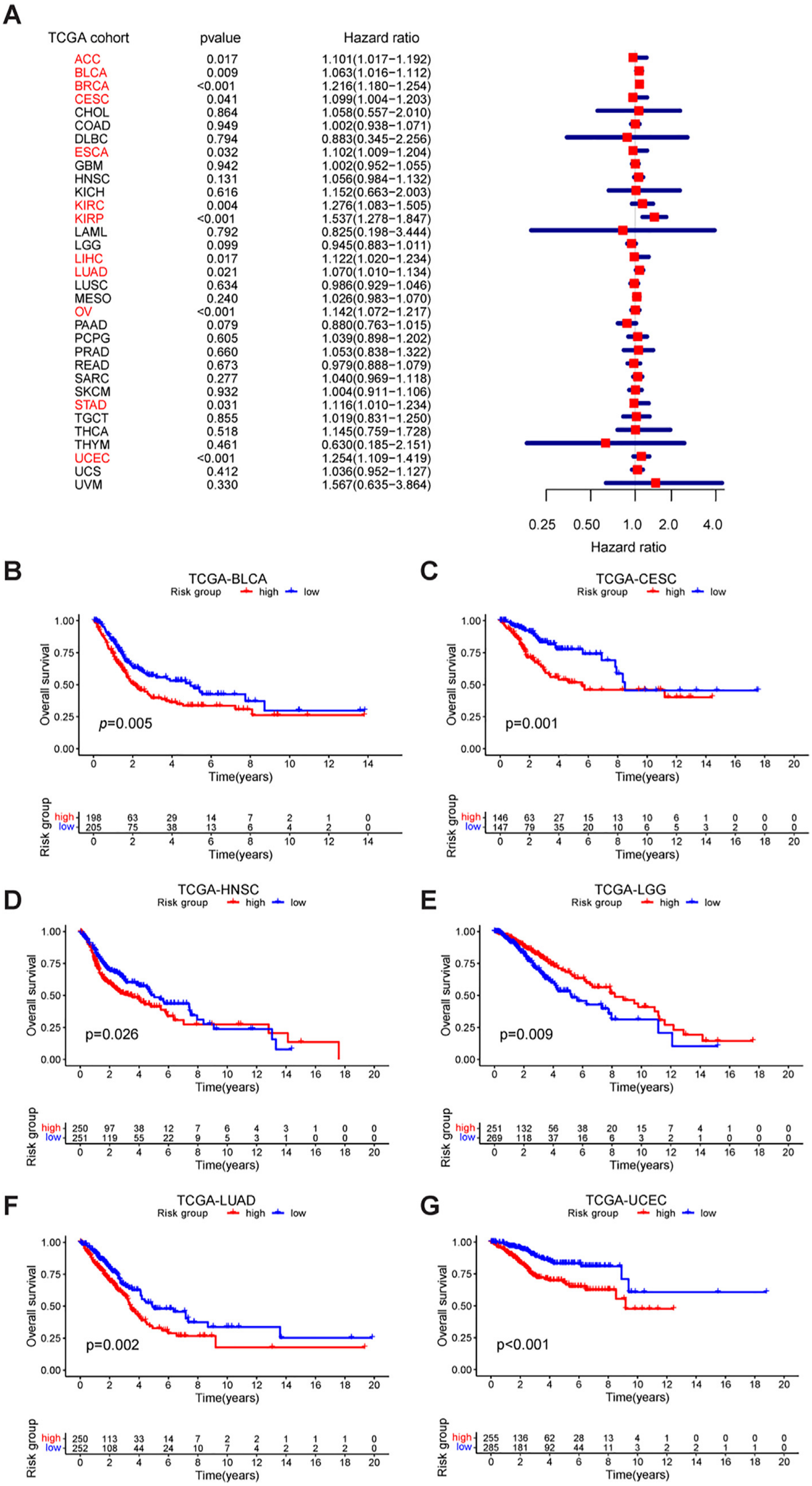
tion curves in both the training and test sets (Fig. 7D-G). The average ROC-AUCs were greater than 0.6. When only considering ROC corves, there was some overfitting in the training set, however, when considering calibration curves, we achieved high accuracy in both sets suggesting the robustness of the nomogram.

Our clustering prediction, recommended drugs and nomogram are available using a web-app (https://sujiezhulab.shinyapps.io/BRCA/).

4. Discussion

Despite recent improvement in the outcome of patients with breast cancer, most patients with advanced disease are at a higher risk of relapse and distant metastasis, and therefore high mortality [86,87]. Traditional classification systems and current prognostic prediction markers fail to accurately reflect the biological heterogeneity and clinical complexity of breast cancer. Here, we identify-three novel molecular subtypes of breast cancer that reflect disease heterogeneity and can help guide the clinical management of the disease. Our web-tool presents a tool with predictive value for different subtypes of breast cancer and patient treatment; implementation of the RGP score and nomogram provides an easy-to-use risk score for breast cancer patients.

The three subtypes we identified exhibited significantly different types of tumor microenvironment cell infiltrations. Cluster A was characterized by the suppression of immunity; cluster B was



TCGA cohort	pvalue	Hazard ratio
ACC	0.017	1.101(1.017−1.192)
BLCA	0.009	1.063(1.016−1.112)
BRCA	<0.001	1.216(1.180−1.254)
CESC	0.041	1.099(1.004−1.203)
CHOL	0.864	1.058(0.557−2.010)
COAD	0.949	1.002(0.938−1.071)
DLBC	0.794	0.883(0.345−2.256)
ESCA	0.032	1.102(1.009−1.204)
GBM	0.942	1.002(0.952−1.055)
HNSC	0.131	1.056(0.984−1.132)
KICH	0.616	1.152(0.663−2.003)
KIRC	0.004	1.276(1.083−1.505)
KIRP	<0.001	1.537(1.278−1.847)
LAML	0.792	0.825(0.198−3.444)
LGG	0.099	0.945(0.883−1.011)
LIHC	0.017	1.122(1.020−1.234)
LUAD	0.021	1.070(1.010−1.134)
LUSC	0.634	0.986(0.929−1.046)
MESO	0.240	1.026(0.983−1.070)
OV	<0.001	1.142(1.072−1.217)
PAAD	0.079	0.880(0.763−1.015)
PCPG	0.605	1.039(0.898−1.202)
PRAD	0.660	1.053(0.838−1.322)
READ	0.673	0.979(0.888−1.079)
SARC	0.277	1.040(0.969−1.118)
SKCM	0.932	1.004(0.911−1.106)
STAD	0.031	1.116(1.010−1.234)
TGCT	0.855	1.019(0.831−1.250)
THCA	0.518	1.145(0.759−1.728)
THYM	0.461	0.630(0.185−2.151)
UCEC	<0.001	1.254(1.109−1.419)
UCS	0.412	1.036(0.952−1.127)
UVM	0.330	1.567(0.635−3.864)

Fig. 6. Pan-cancer analysis of the RGP score. A. Cox regression analysis of the RGP score across 33 cancer types. The red color indicates significant results. B-G. Kaplan-Meier survival curves of the RGP signature in 6 types of cancers (Log-rank test). B. TCGA-BLCA cohort. C.TCGA-CESC cohort. D. TCGA-HNSC cohort. E. TCGA-LGG cohort. F. TCGA-LUAD cohort. G. TCGA-UCEC cohort. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

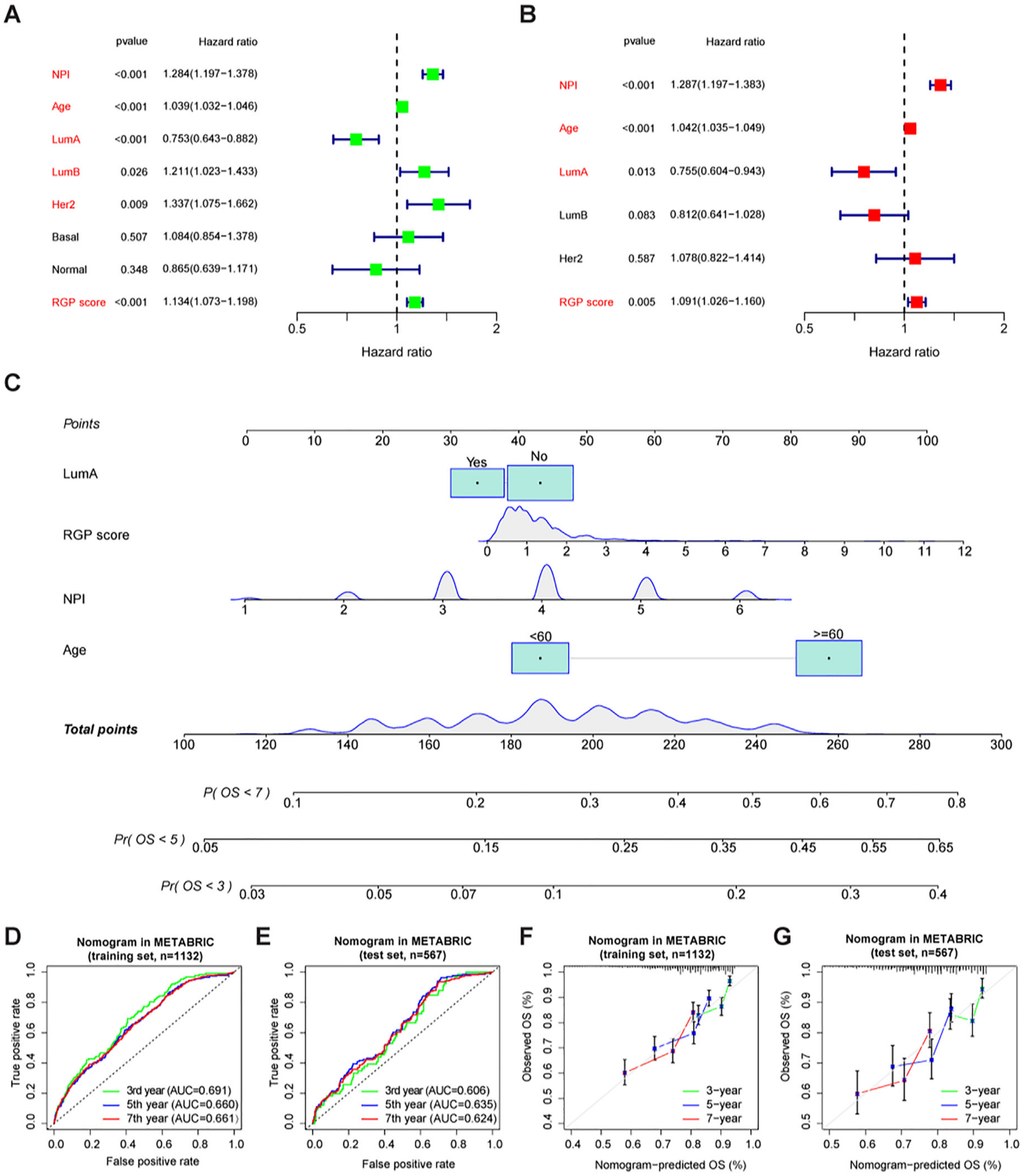


Fig. 7. Nomogram generation and validation in the METABRIC set. A. Univariate independent prognostic analysis. The red color indicates significant results. B. Multivariate independent prognostic analysis. The red color indicates significant results. C. A 4-factor nomogram containing the RGP score. D. ROC curves of the nomogram in the METABRIC training set. E. ROC curves of the nomogram in the METABRIC test set. F. Calibration curves of the nomogram in the METABRIC training set. G. Calibration curves of the nomogram in the METABRIC test set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

characterized by activation of immune cells; cluster C was characterized by activation of adaptive immunity. Although previous studies suggested that the immune-excluded phenotype could be regarded as the non-inflamed tumor and the immune-inflamed phenotype might be regarded as the hot tumor, we found that tumors in cluster B were characterized by low expression of cold tumor marked genes and high expression of hot tumor genes. Furthermore, tumors in cluster B had higher scores in terms of immunotherapy response.

In this study, three novel subtypes had distinct survival in PAM50 subtypes, which indicted that the detailed subtype is necessary for patients to improve treatment and prognosis. Combined with PAM50, BRCA was subdivided into 15 subtypes, which has distinct survival. Given that tumor subtypes based on the PAM50 classifier have distinct prognoses, and respond differently to systemic therapy, we provided a new insight of therapy for 15 subtypes that we constructed. Screening drugs based on gene expression has proven effective, therefore, we have used applied this approach to the 15 subtypes of breast cancer. Taking into account the safety of the drug, we focused on drugs that had been approved by FDA for patients with breast cancer. This list of drugs included targeted drugs and chemotherapy agents, serving as a reasonable reference for clinicians.

Although several prognostic models of breast cancer have been previously constructed, they have not been sufficiently robust to have clinical value. Therefore, we established a web tool based on RGP score. This web-tool included the precision subtypes, treatment and prognosis for patients with breast cancer. To the best of our knowledge, this is the first web app for breast cancer that simultaneously considers two key questions including treatment and prognosis. Using 10 independent breast cancer cohorts, and show it can predict the prognosis in more robust and accurate ways than other existing signatures. Our RGP score also effectively eliminated batch effects between different patient cohorts and data sets, and was shown to be an independent prognostic factor in patients with breast cancer, unexpectedly also holding prognostic value in the pan-cancer datasets.

Some genes involved in the RGP score are related to breast cancer, such as *HDAC2* and *PARP3*, and the former codes for a protein belonging to the histone deacetylase family. This protein forms transcriptional repressor complexes by associating with many other proteins, which plays an important role in transcription as well as drug resistance (). Notably, HDAC inhibitors are potential regents to overcome chemotherapeutic resistance [88]. Furthermore, recent study showed that HDAC2 promotes IFNγ-induced PDL1 expression, and targeting the HDAC2 may enhance antitumor immunity in triple-negative breast cancer (TNBC) [89]. *PAPR3* codes the protein that is a member of PAPR family, which is required for the DNA repair, regulation of apoptosis, and maintenance of genomic stability [90]. It is shown that PARP3 knockdown reduces the survival of BRCA1-deficient TNBC cells, and PAPR3 inhibition is a promising strategy for BRCA1-deficient tumors [91].

Our work also has some limitations. Cancer is a very heterogeneous disease in molecular terms, and cancer development involves multiple signaling pathways rather than single pathway genes. Our findings provide new opportunities for identifying breast cancer subtypes, which can potentially be generalizable to other cancers but it still need further investigation and validation. Although L1000 has shown a strong predictive value of drugs on gene expression, these need to be further validated before they can be used in a clinical setting. Another limitation is that all the analyses were based on public datasets. Future studies are needed to examine the accuracy of our predictions for potential drugs for the distinct subtypes and the RGP score to predict survival.

5. Conclusion

In conclusion, our study leveraged multiple omics data to identify distinct molecular breast cancer subtypes with unique behavior and clinical traits. Novel treatment strategies that are subtype-specific are recommended for patients with breast cancer. A robust prognosis model, RGP score, was developed too. And our web-tool can help promote personalized therapi.

Funding

This study was supported by the National Natural Science Foundation of China (Grant No. 81903539, Sujie Zhu holding), (Grant No. 31671447, Yin Wang holding).

CRediT authorship contribution statement

Jie Zhu: Conceptualization, Formal analysis, Methodology, Software, Writing – original draft. **Weikaixin Kong:** Conceptualization, Formal analysis, Methodology, Software, Writing – original draft. **Liting Huang:** Formal analysis. **Shixin Wang:** Formal analysis. **Suzhen Bi:** Investigation. **Yin Wang:** Supervision, Writing – review & editing. **Peipei Shan:** Supervision, Writing – review & editing. **Sujie Zhu:** Supervision, Funding acquisition, Writing – review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.017.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68:7–30.

[2] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69:7–34.

[3] Curigliano G, Burstein HJ, Winer EP, Gnant M, Dubsky P, Loibl S, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. Ann Oncol. 2017; 28: 1700-12.

[4] Fan L, Strasser-Weippl K, Li JJ, St Louis J, Finkelstein DM, Yu KD, et al. Breast cancer in China. Lancet Oncol 2014;15:e279–89.

[5] Ali HR, Rueda OM, Chin SF, Curtis C, Dunning MJ, Aparicio SA, et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biol 2014;15:431.

[6] Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 2015;163:506–19.

[7] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.

[8] Zaha DC. Significance of immunohistochemistry in breast cancer. World J Clin Oncol 2014;5:382–92.

[9] Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001;98:10869–74.

[10] Erdag G, Schaefer JT, Smolkin ME, Deacon DH, Shea SM, Dengel LT, et al. Immunotype and immunohistologic characteristics of tumor-infiltrating immune cells are associated with clinical outcome in metastatic melanoma. Cancer Res 2012;72:1070–80.

[11] Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pages C, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science 2006;313:1960–4.

[12] Yang C, Lee H, Jove V, Deng J, Zhang W, Liu X, et al. Prognostic significance of B-cells and pSTAT3 in patients with ovarian cancer. PLoS One 2013;8:e54029.

[13] Zhang L, Conejo-Garcia JR, Katsaros D, Gimotty PA, Massobrio M, Regnani G, et al. Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. N Engl J Med 2003;348:203–13.

[14] Yang L, Wang S, Zhang Q, Pan Y, Lv Y, Chen X, et al. Clinical significance of the immune microenvironment in ovarian cancer patients. Mol Omics 2018;14:341–51.

[15] Yang L, Lv Y, Wang S, Zhang Q, Pan Y, Su D, et al. Identifying FL11 subtype by characterizing tumor immune microenvironment in prostate adenocarcinoma via Chou's 5-steps rule. Genomics 2020;112:1500–15.

[16] Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. Lancet Oncol 2018;19:40–50.

[17] Jin YW, Hu P. Tumor-infiltrating CD8 T cells predict clinical breast cancer outcomes in young women. Cancers 2020;12.

[18] Condamine T, Ramachandran I, Youn JI, Gabrilovich DI. Regulation of tumor metastasis by myeloid-derived suppressor cells. Annu Rev Med 2015;66:97–110.

[19] Fassler DJ, Torre-Healy LA, Gupta R, Hamilton AM, Kobayashi S, Van Alsten SC, et al. Spatial characterization of tumor-infiltrating lymphocytes and breast cancer progression. Cancers 2022;14.

[20] Sheu BC, Kuo WH, Chen RJ, Huang SC, Chang KJ, Chow SN. Clinical significance of tumor-infiltrating lymphocytes in neoplastic progression and lymph node metastasis of human breast cancer. Breast (Edinburgh, Scotland) 2008;17:604–10.

[21] Duijf PHG, Nanayakkara D, Nones K, Srihari S, Kalimutho M, Khanna KK. Mechanisms of Genomic Instability in Breast Cancer. Trends Mol Med 2019;25:595–611.

[22] Kalimutho M, Nones K, Srihari S, Duijf PHG, Waddell N, Khanna KK. Patterns of genomic instability in breast cancer. Trends Pharmacol Sci 2019;40:198–211.

[23] Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet 2013;45:1134–40.

[24] Wang Z, Bao A, Liu S, Dai F, Gong Y, Cheng Y. A Pyroptosis-related gene signature predicts prognosis and immune microenvironment for breast cancer based on computational biology techniques. Front Genet 2022;13:801056.

[25] Zhang K, Ping L, Du T, Liang G, Huang Y, Li Z, et al. A Ferroptosis-Related lncRNAs signature predicts prognosis and immune microenvironment for breast cancer. Front Mol Biosci 2021;8:678877.

[26] Zhou L, Rueda M, Alkhateeb A. Classification of breast cancer nottingham prognostic index using high-dimensional embedding and residual neural network. Cancers 2022;14.

[27] Tabl AA, Alkhateeb A, Pham HQ, Rueda L, ElMaraghy W, Ngom A. A novel approach for identifying relevant genes for breast cancer survivability on specific therapies. Evol Bioinform Online 2018;14.

[28] Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell 2017;171:1437–5217.

[29] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486:346–52.

[30] Yuan J, Duan F, Zhai W, Song C, Wang L, Xia W, et al. An aging-related gene signature-based model for risk stratification and prognosis prediction in breast cancer. Int J Womens Health 2021;13:1053–64.

[31] Guo L, Jing Y. Construction and identification of a novel 5-gene signature for predicting the prognosis in breast cancer. Front Med (Lausanne) 2021;8:669931.

[32] Ye M, Li L, Liu D, Wang Q, Zhang Y, Zhang J. Identification and validation of a novel zinc finger protein-related gene-based prognostic model for breast cancer. PeerJ 2021;9:e12276.

[33] Lan Y, Su J, Xue Y, Zeng L, Cheng X, Zeng L. Analysing a novel RNA-binding-protein-related prognostic signature highly expressed in breast cancer. J Healthc Eng 2021;2021:9174055.

[34] Du J, Dong Y, Li Y. Identification and prognostic value exploration of cyclophosphamide (cytoxan)-centered chemotherapy response-associated genes in breast cancer. DNA Cell Biol 2021;40:1356–68.

[35] Liu M, Li Q, Zhao N. Identification of a prognostic chemoresistance-related gene signature associated with immune microenvironment in breast cancer. Bioengineered 2021;12:8419–34.

[36] Jin X, Yan J, Chen C, Chen Y, Huang WK. Integrated analysis of copy number variation, microsatellite instability, and tumor mutation burden identifies an 11-gene signature predicting survival in breast cancer. Front Cell Dev Biol 2021;9:721505.

[37] Peng W, Lin C, Jing S, Su G, Jin X, Di G, et al. A novel seven gene signature-based prognostic model to predict distant metastasis of lymph node-negative triple-negative breast cancer. Front Oncol 2021;11:746763.

[38] He M, Hu C, Deng J, Ji H, Tian W. Identification of a novel glycolysis-related signature to predict the prognosis of patients with breast cancer. World J Surg Oncol 2021;19:294.

[39] Lai YW, Hsu WJ, Lee WY, Chen CH, Tsai YH, Dai JZ, et al. Prognostic value of a glycolytic signature and its regulation by Y-Box-binding protein 1 in triple-negative breast cancer. Cells 2021:10.

[40] Ding S, Sun X, Zhu L, Li Y, Chen W, Shen K. Identification of a novel immune-related prognostic signature associated with tumor microenvironment for breast cancer. Int Immunopharmacol 2021;100:108122.

[41] Liu Q, Ma JY, Wu G. Identification and validation of a ferroptosis-related gene signature predictive of prognosis in breast cancer. Aging (Albany NY) 2021;13:21385–99.

[42] Sun X, Luo H, Han C, Zhang Y, Yan C. Identification of a hypoxia-related molecular classification and hypoxic tumor microenvironment signature for predicting the prognosis of patients with triple-negative breast cancer. Front Oncol 2021;11:700062.

[43] Sun C, Wang S, Zhang Y, Yang F, Zeng T, Meng F, et al. Risk signature of cancer-associated fibroblast-secreted cytokines associates with clinical outcomes of breast cancer. Front Oncol 2021;11:628677.

[44] Yang X, Weng X, Yang Y, Zhang M, Xiu Y, Peng W, et al. A combined hypoxia and immune gene signature for predicting survival and risk stratification in triple-negative breast cancer. Aging (Albany NY) 2021;13:19486–509.

[45] Liu Y, Sun H, Li X, Liu Q, Zhao Y, Li L, et al. Identification of a Three-RNA Binding Proteins (RBPs) signature predicting prognosis for breast cancer. Front Oncol 2021;11:663556.

[46] Yang A, Zhou Y, Kong Y, Wei X, Ye F, Zhang L, et al. Identification and validation of immune-related methylation clusters for predicting immune activity and prognosis in breast cancer. Front Immunol 2021;12:704557.

[47] Qi A, Ju M, Liu Y, Bi J, Wei Q, He M, et al. Development of a novel prognostic signature based on antigen processing and presentation in patients with breast cancer. Pathol Oncol Res 2021;27:600727.

[48] Liu J, Liu Z, Zhou Y, Zeng M, Pan S, Liu H, et al. Identification of a novel transcription factor prognostic index for breast cancer. Front Oncol 2021;11:666505.

[49] Li Y, Zhao X, Liu Q, Liu Y. Bioinformatics reveal macrophages marker genes signature in breast cancer to predict prognosis. Ann Med 2021;53:1019–31.

[50] Ma JY, Liu Q, Liu G, Peng S, Wu G. Identification and validation of a robust autophagy-related molecular model for predicting the prognosis of breast cancer patients. Aging (Albany NY) 2021;13:16684–95.

[51] Zhong S, Lin Z, Chen H, Mao L, Feng J, Zhou S. The m(6)A-related gene signature for predicting the prognosis of breast cancer. PeerJ 2021;9:e11561.

[52] Long M, Hou W, Liu Y, Hu T. A histone acetylation modulator gene signature for classification and prognosis of breast cancer. Curr Oncol 2021;28:928–39.

[53] Zhou X, Zhang FY, Liu Y, Wei DX. A risk prediction model for breast cancer based on immune genes related to early growth response proteins family. Front Mol Biosci 2020;7:616547.

[54] Tan L, He X, Shen G. Identification of a 15-pseudogene based prognostic signature for predicting survival and antitumor immune response in breast cancer. Aging (Albany NY) 2020;13:14499–521.

[55] Wu F, Chen W, Kang X, Jin L, Bai J, Zhang H, et al. A seven-nuclear receptor-based prognostic signature in breast cancer. Clin Transl Oncol 2021;23:1292–303.

[56] Pei J, Li Y, Su T, Zhang Q, He X, Tao D, et al. Identification and validation of an immunological expression-based prognostic signature in breast cancer. Front Genet 2020;11:912.

[57] Wu CC, Ekanem TI, Phan NN, Loan DTT, Hou SY, Lee KH, et al. Gene signatures and prognostic analyses of the Tob/BTG pituitary tumor-transforming gene (PTTG) family in clinical breast cancer patients. Int J Med Sci 2020;17:3112–24.

[58] Tsai HT, Huang CS, Tu CC, Liu CY, Huang CJ, Ho YS, et al. Multi-gene signature of microcalcification and risk prediction among Taiwanese breast cancer. Sci Rep 2020;10:18276.

[59] Tian Z, Tang J, Liao X, Yang Q, Wu Y, Wu G. Identification of a 9-gene prognostic signature for breast cancer. Cancer Med 2020.

[60] Zhang D, Yang S, Li Y, Yao J, Ruan J, Zheng Y, et al. Prediction of overall survival among female patients with breast cancer using a prognostic signature based on 8 DNA repair-related genes. JAMA Netw Open 2020;3:e2014622.

[61] Shi W, Hu D, Lin S, Zhuo R. Five-mRNA signature for the prognosis of breast cancer based on the ceRNA network. Biomed Res Int 2020;2020:9081852.

[62] Zhao Y, Pu C, Liu Z. Exploration the significance of a novel immune-related gene signature in prognosis and immune microenvironment of breast cancer. Front Oncol 2020;10:1211.

[63] Hu W, Li M, Zhang Q, Liu C, Wang X, Li J, et al. Establishment of a novel CNV-related prognostic signature predicting prognosis in patients with breast cancer. J Ovarian Res 2021;14:103.

[64] Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3.

[65] Chatsirisupachai K, Lesluyes T, Paraoan L, Van Loo P, de Magalhaes JP. An integrative analysis of the age-associated multi-omic landscape across cancers. Nat Commun 2021;12:2345.

[66] Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinf 2013;14:7.

[67] Zhang B, Wu Q, Li B, Wang D, Wang L, Zhou YL. m(6)A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer. Mol Cancer 2020;19:53.

[68] Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep 2017;18:248–62.

[69] Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;462:108–12.

[70] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.

[71] Zhu S, Kong W, Zhu J, Huang L, Wang S, Bi S, et al. The genetic algorithm-aided three-stage ensemble learning method identified a robust survival risk score in patients with glioma. Brief Bioinform 2022;23.

[72] Pan JW, Zabidi MMA, Ng PS, Meng MY, Hasan SN, Sandey B, et al. The molecular landscape of Asian breast cancers reveals clinically relevant population-specific differences. Nat Commun 2020;11:6433.

[73] Romero-Cordoba SL, Salido-Guadarrama I, Rebollar-Vega R, Bautista-Pina V, Dominguez-Reyes C, Tenorio-Torres A, et al. Comprehensive omic characterization of breast cancer in Mexican-Hispanic women. Nat Commun 2021;12:2245.

[74] Jindal S, Pennock ND, Sun D, Horton W, Ozaki MK, Narasimhan J, et al. Postpartum breast cancer has a distinct molecular profile that predicts poor outcomes. Nat Commun 2021;12:6341.

[75] Wu J, Zhu Y, Luo M, Li L. Comprehensive analysis of pyroptosis-related genes and tumor microenvironment infiltration characterization in breast cancer. Front Immunol 2021;12:748221.

[76] Zhong X, Li J, Wu X, Wu X, Hu L, Ding B, et al. Identification of N6-methyladenosine-related lncrnas for predicting overall survival and clustering of a potentially novel molecular subtype of breast cancer. Front Oncol 2021;11:742944.

[77] Ye Z, Zou S, Niu Z, Xu Z, Hu Y. A novel risk model based on lipid metabolism-associated genes predicts prognosis and indicates immune microenvironment in breast cancer. Front Cell Dev Biol 2021;9:691676.

[78] McGranahan N, Burrell RA, Endesfelder D, Novelli MR, Swanton C. Cancer chromosomal instability: therapeutic and diagnostic challenges. EMBO Rep 2012;13:528–38.

[79] Yuan J, Hu Z, Mahal BA, Zhao SD, Kensler KH, Pi J, et al. Integrated analysis of genetic ancestry and genomic alterations across cancers. Cancer Cell 2018;34.

[80] Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science 2017;355.

[81] Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer 2012;12:252–64.

[82] Wang H, Li S, Wang Q, Jin Z, Shao W, Gao Y, et al. Tumor immunological phenotype signature-based high-throughput screening for the discovery of combination immunotherapy compounds. Sci Adv 2021:7.

[83] Verret B, Sourisseau T, Stefanovska B, Mosele F, Tran-Dien A, Andre F. The influence of cancer molecular subtypes and treatment on the mutation spectrum in metastatic breast cancers. Cancer Res 2020;80:3062–9.

[84] Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. Mol Cancer Ther 2017;16:2598–608.

[85] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006;313:1929–35.

[86] Britt KL, Cuzick J, Phillips KA. Key steps for effective breast cancer prevention. Nat Rev Cancer 2020;20:417–36.

[87] Baliu-Pique M, Pandiella A, Ocana A. Breast cancer heterogeneity and response to novel therapeutics. Cancers (Basel) 2020:12.

[88] Cao L, Zhao S, Yang Q, Shi Z, Liu J, Pan T, et al. Chidamide combined with doxorubicin induced p53-driven cell cycle arrest and cell apoptosis reverse multidrug resistance of breast cancer. Front Oncol 2021;11:614458.

[89] Xu P, Xiong W, Lin Y, Fan L, Pan H, Li Y. Histone deacetylase 2 knockout suppresses immune escape of triple-negative breast cancer cells via downregulating PD-L1 expression. Cell Death Dis 2021;12:779.

[90] Sharif-Askari B, Amrein L, Aloyz R, Panasci L. PARP3 inhibitors ME0328 and olaparib potentiate vinorelbine sensitization in breast cancer cell lines. Breast Cancer Res Treat 2018;172:23–32.

[91] Beck C, Rodriguez-Vargas JM, Boehler C, Robert I, Heyer V, Hanini N, et al. PARP3, a new therapeutic target to alter Rictor/mTORC2 signaling and tumor progression in BRCA1-associated cancers. Cell Death Differ 2019;26:1615–30.