Heliyon 10 (2024) e28586

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Differential whole-genome doubling based signatures for improvement on clinical outcomes and drug response in patients with breast cancer

Yingli Lv [*,1], Guotao Feng [1], Lei Yang, Xiaoliang Wu, Chengyi Wang, Aokun Ye, Shuyuan wang, Chaohan Xu [**], Hongbo Shi [***]

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang, 150081, China

ARTICLE INFO

Keywords:
Whole-genome doubling
WGD-Related genes
Prognosis genes
Drug response

ABSTRACT

Whole genome doublings (WGD), a hallmark of human cancer, is pervasive in breast cancer patients. However, the molecular mechanism of the complete impact of WGD on survival and treatment response in breast cancer remains unclear. To address this, we performed a comprehensive and systematic analysis of WGD, aiming to identify distinct genetic alterations linked to WGD and highlight its improvement on clinical outcomes and treatment response for breast cancer. A linear regression model along with weighted gene co-expression network analysis (WGCNA) was applied on The Cancer Genome Atlas (TCGA) dataset to identify critical genes related to WGD. Further Cox regression models with random selection were used to optimize the most useful prognostic markers in the TCGA dataset. The clinical implication of the risk model was further assessed through prognostic impact evaluation, tumor stratification, functional analysis, genomic feature difference analysis, drug response analysis, and multiple independent datasets for validation. Our findings revealed a high aneuploidy burden, chromosomal instability (CIN), copy number variation (CNV), and mutation burden in breast tumors exhibiting WGD events. Moreover, 247 key genes associated with WGD were identified from the distinct genomic patterns in the TCGA dataset. A risk model consisting of 22 genes was optimized from the key genes. High-risk breast cancer patients were more prone to WGD and exhibited greater genomic diversity compared to low-risk patients. Some oncogenic signaling pathways were enriched in the high-risk group, while primary immune deficiency pathways were enriched in the low-risk group. We also identified a risk gene, ANLN (anillin), which displayed a strong positive correlation with two crucial WGD genes, KIF18A and CCNE2. Tumors with high expression of ANLN were more prone to WGD events and displayed worse clinical survival outcomes. Furthermore, the expression levels of these risk genes were significantly associated with the sensitivities of BRCA cell lines to multiple drugs, providing valuable insights for targeted therapies. These findings will be helpful for further improvement on clinical outcomes and contribution to drug development in breast cancer.

[*] Corresponding author.
[**] Corresponding author.
[***] Corresponding author.
E-mail addresses: lyu.hrb.bio@hotmail.com (Y. Lv), chaohanxu@hrbmu.edu.cn (C. Xu), shihongbo@ems.hrbmu.edu.cn (H. Shi).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e28586
Received 12 August 2023; Received in revised form 20 March 2024; Accepted 20 March 2024
Available online 26 March 2024

1. Introduction

Whole-genome doubling (WGD), also known as tetraploidization, refers to the duplication of an entire genome through doubling DNA contents in a cell. It is a regulated program in plants and fungi, and can occur in specific tissues as part of terminal differentiation in human beings [1]. Majority of human diploid cells undergo cell division using the normal cell cycle mechanism. However, an error in the normal process can lead diploid cell transitions to a tetraploid state [2], and contribute to a multitude of malignant phenotypes [3]. The known mechanisms of WGD generation include mitotic slippage, cytokinesis failure, cell-cell fusion, and endoreplication [1,4]. Persistent telomere dysfunction, a common event in human tumorigenesis [5], was also shown to induce tetraploidization [6].

WGD has been a common genomic event in cancer, leading to genome instability [7]and promoting resistance to a broad spectrum of chemotherapeutic drugs [8]. This event occurs in approximately 40% of solid tumors, with varying frequencies across different cancer types [9]. WGD has long been recognized as closely associated with tumorigenesis [3,10–12] and has been identified as a key contributor to genome instability [13,14], making it a potential therapeutic target. WGD has shown the ability to identify primary tumors with poor-prognosis, offering valuable insights for the design of adjuvant trials targeting specific high-risk patient populations [9].

The TCGA project revealed that approximately 40% of breast cancer (BRCA) patients have undergone WGD and patients with a high frequency of WGD experience poor prognoses with high mortality rates [9]. These mechanics along with intratumoral heterogeneity adds to the complexity of breast cancer pathogenesis [15,16]. The treatment of breast cancer is a long-term process, and patients can develop drug resistance over time, making it challenging to improve long-term survival rates [17]. Previous studies on molecular research in breast cancer have made significant contributions to clinical diagnosis and prognosis [18]. The heterogeneity of breast cancer subtypes has also been explained through various genomic events including large-scale somatic calls [19]. Moreover, anti-cancer peptides (ACPs), a new type of anti-cancer predictors, demonstrate the ability to inhibit the proliferation or migration of tumor cells while exhibiting a reduced propensity for inducing drug resistance [20]. All peptides were analyzed for ACPs utilizing the web-based prediction servers such as AntiCP [21], ACPP [22],iACP [23], iACP-GAEnsC [24], ACPred [25], cACP-DeepGram [26], etc. In the treatment of breast cancer, two ACPs were found: NRC-03 and NRC-07, which can be used alone or in combination with conventional chemotherapy drugs to treat breast cancer [27]. These findings bring new hope for the treatment of breast cancer. Recently, Taylor et al. applied a pan-cancer analysis and found a correlation between WGD and poor survival in cancer patients with advanced-stage disease, even with the presence of metastasis [9]. Ganem et al. identified that WGD tumors overexpress genes important for cellular proliferation, mitotic spindle formation, and DNA repair [28]. McGranahan et al. explored the evolutionary importance of WGD in cancer, and showed how this can be exploited to identify novel cancer genes [14]. Chia-Hsin Wu et al. discovered the distribution patterns of WGD in triple-negative BRCA patients in the Taiwanese population (BCTW) and provided insights into the impact of WGD on the timing of tumor initiation and tumor maintenance events in BRCA subtypes [19]. However, Taylor's analysis exclusively anticipated the worse overall survival across various cancer types due to genomic doubling. Ganem et al. identified only a substantial pool of overexpressed genes that were affected by WGD in both pan-cancer and 21 specific cancer types. McGranahan et al. applied the fundamental principles of Darwinian evolution to study tumor development. While Ganem et al.'s research outlined the subtype specificity of WGD and CIN in BCTW. Collectively, there is still limited understanding regarding the exploration and elucidation of the role of WGD and its effects on other relevant molecular alterations in breast cancer prognosis and treatment response. The molecular characteristics of breast cancer driven by large-scale chromosomal abnormalities have not been thoroughly analyzed in terms of their impact on the development, prognosis, and treatment of breast cancer [19]. The understanding of how WGD contributes to shaping the patterns of genome characteristics for improving the prognosis of human breast cancers remains scarce.

This study presents a comprehensive bioinformatics analysis that integrates genomic data of breast cancer patients with WGD status to investigate the differential features following the WGD event and explore the fundamental mechanisms underlying clinical prognostic differences. We performed comprehensive screening to identify crucial genes associated with WGD status in breast cancer. By combining univariate/multivariate Cox regression models with random model approaches, we successfully identified a panel of breast cancer risk genes linked to WGD in both training and validation datasets. Furthermore, we conducted an exploratory analysis to investigate the functional mechanisms, distributions of clinical-pathological features, and variations in genomic features among the different breast cancer risk groups affected by WGD. Our findings contribute to a deeper understanding of the genomic characteristics associated with WGD in BRCA and provide potential diagnostic tools for the benefit of BRCA patients. More importantly, our study highlights the value of evaluating WGD status in accelerating the prognostic evaluation of cancer patients. This will also provide a basis for promoting WGD as a diagnostic factor to guide clinical physicians. The workflow is depicted in Supplementary Fig. 1.

2. Materials and methods

2.1. Cohort datasets and preprocessing

The gene expression and clinical information data of human primary breast cancer (BRCA) were obtained from TCGA (https://portal.gdc.cance r.gov). The RNA-seqV2 data was used for expression analysis, while the copy number ratios for each gene were derived from the BRCA segmentation file after processing it through GISTIC2.0 (http://api.gdc.cancer.gov/data/00a32f7a-c85f-4f86-850d-be53973cbc4d). Tumor purity, ploidy, and WGD calls for BRCA samples were downloaded from TCGA Pan-Cancer Atlas (https://gdc.cancer.gov/about-data/publications/pancanatlas).To ensure sample integrity, repeat sequencing samples were filtered based on maximum ploidy, and when the ploidy was the same, the sample with the maximum purity was selected. For expression, genes with zero expression values in more than 80% of samples were removed, and the remaining expression values were log2-transformed after adding a pseudo-count of 0.01. This led to 987 BRCA samples containing 16292 genes for subsequent analysis. Additionally, eight BRCA cohorts from gene expression omnibus (GEO) datasets (GSE20711, GSE20713, GSE22249, GSE24450, GSE37751, GSE39004, GSE45255, and GSE16228), which included overall survival times, as well as the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset, were collected. Each BRCA cohort was processed and analyzed individually as discussed above using the R package limma (version 3.56.1).

2.2. Genome features and special pattern analysis associated with WGD status

2.2.1. Mutational burden

We first divided the tumors into two groups based WGD status as WGD+ (4N) and WGD- (2N). We obtained the tumor mutation burden (TMB) data from the published work of Vé steinn Thorsson et al. [29]. In their study, TMB was defined as the number of non-synonymous mutations per megabase (log10-transformed). We divided the TMB of each sample by its ploidy and marked it as ploidy-corrected TMB. Further we analyzed differential patterns among WGD+ and WGD-tumor based on the TMB and ploidy-corrected TMB by using Wilcoxon rank-sum test.

2.2.2. Aneuploidy score profiling

To examine the differences in aneuploidy between the WGD+ and WGD- BRCA samples, we compared the aneuploidy scores both at the arm-level and chromosome-level. Aneuploidy scores were calculated using the established methods [30,31] based on copy number segmentation data which was retrieved from TCGA (http://api.gdc.cancer.gov/data/00a32f7a-c85f-4f86-850d-be53973cbc4d). Furthermore, we depicted the distribution patterns of the aneuploidy scores for each chromosome arm and chromosome separately (see Supplementary method).

2.2.3. Chromosome instability score

The absolute copy number data for BRCA was obtained from the TCGA PanCan Atlas, which was generated using ABSOLUTE based on tumor purity and ploidy estimates [32]. To assess chromosomal instability (CIN), we quantified the ratio of genomic events involving gain and loss for each chromosome in two WGD status groups. The sum of CIN scores for all chromosomes was defined as the CIN of each tumor sample. We compared the CIN scores between WGD+ and WGD-samples for each chromosome and for each sample in BRCA from the TCGA data (see Supplementary method).

2.2.4. Copy number variation score

Copy number variation (CNV) data from the PanCan Atlas was utilized to investigate CNVs in both WGD+ and WGD-samples of BRCA. The CNV for each chromosome was quantified by counting the occurrences of gain, loss, and both events separately. The CNV score for each tumor sample was defined as the sum of CNVs across all chromosomes. We log10-transformed the CNV scores and compared the three-level CNVs (gain, loss, and both events) for each chromosome and for each sample in both WGD+ and WGD-samples of TCGA-BRCA tumors (see Supplementary method).

2.3. Identification of WGD-related crucial genes

We used linear model and weighted gene co-expression network analysis (WGCNA) to identify genes associated with WGD in BRCA using gene-expression profiles. First a linear model was applied to each gene in BRCA with expression as a function of WGD status, tumor purity, and CN_Local. The ABSOLUTE estimated tumor purity, and 'CN_Local' is the log2-transformed copy number ratio for that gene in each tumor estimated by GISTIC2.0. A total of 8808 genes were selected. Further, we conducted WGCNA analysis on the top 25% of all genes with the highest expression variation. In the WGCNA procedure, an appropriate soft threshold β was calculated to meet the criteria for a scale-free network. The weighted adjacency matrix was then transformed into a topological overlap matrix (TOM), from which the corresponding dissimilarity (1-TOM) was generated. The dynamic tree cutting approach was employed to identify gene modules. To recognize gene modules significantly correlated with WGD status, the module with the highest significant positive and negative correlations was selected for further analysis. Genes within the significant modules that exhibited both high gene significance (GS) and module membership (MM) were defined as candidate WGD-related genes as reported in a previous study [33]. Finally, the overlapping genes derived from the linear model and the candidate WGD-related gene set obtained from WGCNA were identified as WGD-related crucial genes (Supplementary Fig. 1).

2.4. Determination of risk genes linked to WGD status

In the TCGA-BRCA sample set, we performed a random split of WGD+ and WGD-samples, allocating 70% of each group to the training set and 30% to the internal validation set. We repeated the random split process 100 times to ensure the accuracy and robustness of the prognostic model training.

For each random split of the training set, univariable Cox proportional hazards regression analysis was used to select gene sets associated with patient survival from the WGD-related crucial gene set. FDR correction was applied to the P values to control for multiple hypothesis testing. Additionally, we utilized the least absolute shrinkage and selection operator (LASSO)-Cox regression model, implemented using the "glmnet" R package (version 4.1–7), to select useful prognostic genes based on the molecular profile of the aforementioned gene set. Genes with nonzero coefficients were selected using the LASSO Cox regression model with the optimal lambda (minimized lambda) value determined by 10-fold cross-validation. Subsequently, we constructed a prognostic risk score for each patient using the LASSO Cox regression model, using the following formula:

$$Risk\ Score = \sum_{i=1}^{n} Beta(i)GeneExp(i)$$

Here, Beta(i) represents the i-th regression coefficient derived from the LASSO Cox regression model, and GeneExp(i) denotes the expression level of the i-th prognostic gene. We utilized the risk scores to predict the prognostic outcomes of the patients in the training set and evaluated the model's performance using the concordance index (C-index). This process was repeated 100 times, generating 100 C-index values. The model with the highest C-index was selected to predict the risk score in the corresponding internal validation set. Additionally, we calculated the time-dependent area under the curve (AUC) of the receiver operating characteristic (ROC) to assess the model's predictive performance. Finally, the genes optimized through this modeling approach were considered as the risk genes associated with WGD status in BRCA.

2.5. Risk groups of prognostic genes and survival analysis

The optimized model was used to assign patients into high and low-risk groups according to the median value of the risk scores determined by the "survminer" package (version 0.4.9). Associations between the risk scores and clinicopathological features (age, clinical stage, and subtype) were examined using Kruskal-Wallis analysis and visualized using Sankey diagrams, which effectively illustrated the correlations between the risk score and different survival outcomes.

The prognostic genes related with WGD in BRCA were subsequently evaluated in the internal validation set and the entire TCGA-BRCA tumors. Kaplan-Meier analysis curves (KM) were generated using the R package "survival" (version 3.5–5) to compare the trends in overall survival (OS) or progression-free survival (PFI) between the high and low-risk groups. The results of univariate and multivariate Cox analyses were visualized as a forest plot, while the distribution of risk scores and patient survival status were depicted through ranked dot and scatter plots. Additionally, ROC curve analysis was conducted to assess the specificity and sensitivity of the risk genes for three, five, seven, and ten-year survival, utilizing R package "survivalROC" (version 1.0.3.1). The AUC values were calculated to designate the ROC performance. KM curves and ROC curves were also generated for eight independent GEO validation cohorts and METABRIC to evaluate the clinical impact of the risk genes.

2.6. Construction and validation of the nomogram model

The prognostic criteria was established through univariate and multivariate COX proportional hazard regression analysis, incorporating clinical features (age, stage, subtype), risk scores, and WGD status. To facilitate clinical diagnosis of BRCA patients, a predicted nomogram was constructed using the significant clinical features for patients in training, internal validation, and entire TCGA cohort by the "rms" (version 6.7-0) program. The hazard ratio (HR), 95% confidence interval (CI), and P-values obtained by log-rank test of the risk genes were presented.

Each variable in the nomogram scoring system was assigned a score, and the total score was calculated by summing the scores from all factors in each sample. Calibration curves were utilized to evaluate the consistency between the nomogram predictions and clinical observations for three, five, seven, and ten-year OS and PFI. The nomogram's predictive performance for three, five, seven, and ten-year OS survival was evaluated using ROC curves. Further it was compared with the ROC curves of other clinical variables in the training, internal validation, and entire TCGA cohort. The C-index was calculated to determine the nomogram's predictive potential. Furthermore, decision curve analysis (DCA) was employed to compare the clinical benefits of different models by R package "DCA" (version 2.0).

2.7. Functional enrichment analysis

We performed functional enrichment analysis for 247 genes associated with WGD status and differentially expressed genes (DEGs) related with risk groups in BRCA patients. The R package "DESeq2″ was used to filter DEGs between high and low-risk patient groups, with genes having an adjusted P value < 0.05 and $|\text{logFC}| \geq 0.5$ considered statistically significant and shown in a Volcano plot. Functional categories and pathways enriched in the high and low-risk groups were identified and visualized using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses performed with the R package "clusterProfiler" (version 4.8.1). The P-value was adjusted by the Benjamini and Hochberg method. Additionally, Gene Set Enrichment Analysis (GSEA) was conducted using the R package "GSVA" (version 1.48.0) and "h.all.v7.2.symbols.gmt" was set as the reference database. Pathways with a normalized $P < 0.05$ and a false-discovery rate (FDR) $q < 0.25$ were considered statistically significant, and the top ten enriched pathways were selected by ranking of normalized enrichment scores (NESs). Pathway enrichment using GO and KEGG was also performed to visualize the function of 247 genes.

2.8. Analysis of drug sensitivity

Cancer Cell Line Encyclopedia (CCLE) encompasses omics data in ~1000 cancer cell lines, while the Cancer Therapeutics Response Portal (CTRP, https://portals.broadinstitute.org/) provides information regarding 481 small-molecule probes and drugs across 860 extensively characterized cancer cell lines. We integrated both resources to compile 40 BRCA cell lines, including the expression levels of 21 out of 22 risk genes as well as the sensitivity of these cell lines to 500 drugs (Area Under Curve [34]).

The BRCA cell lines were split into high and low-risk groups according to the median risk score of 40 cell lines using the LASSO Cox regression model constructed for BRCA TCGA samples. Subsequently, we calculated Pearson correlations between the expression levels of the 21 risk genes and the sensitivity of the 500 drugs in these cell lines. Correlations with a *P*-value <0.05 were considered significant connections between the high and low-risk groups. To assess the risk genes' usefulness in clinical therapy, the differences in AUC values of compounds were compared between the two groups by the Wilcoxon rank-sum test.

2.9. Statistical analysis

Data analysis and visualization were conducted using R software (version 4.3.0) with the necessary packages. The Wilcoxon rank-sum test or Kruskal-Wallis test was applied to analyze continuous variables, while the chi-square test was utilized to analyze the proportions of patients with WGD in the high and low-risk groups, and the proportions of patients with mutations in the two risk groups. Pearson correlation analysis was used to assess the relationships between risk scores and expression levels of differentially expressed genes, as well as to correlate genomic characteristics with risk scores. All P-values were two-sided, and statistical significance was determined at a threshold of $P < 0.05$.

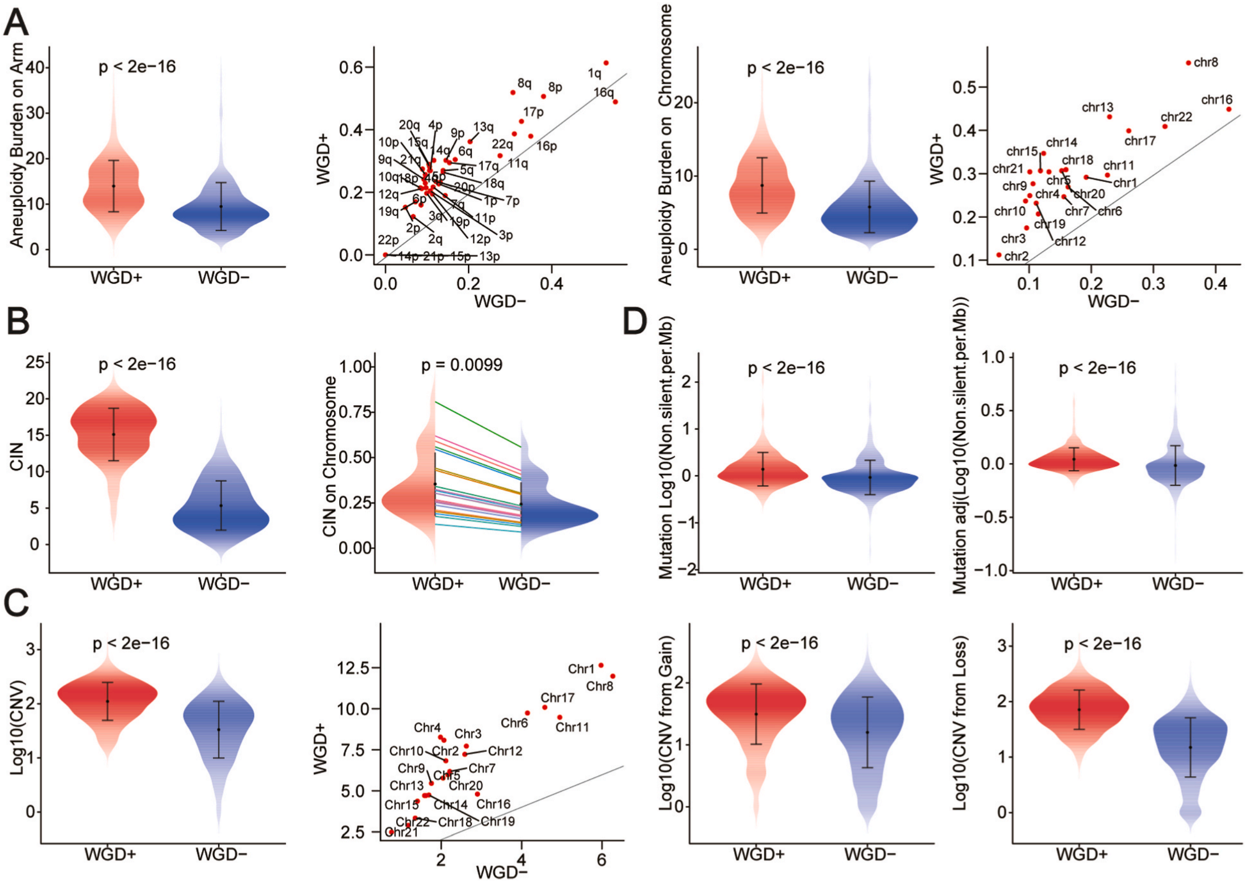


Fig. 1. Comparision analysis of genomic features in WGD+ and WGD-tumors from TCGA breast cancers. (A) The variation of aneuploidy burden on arm/chromosome level were shown in violin plots; The average variation of aneuploidy burden across tumors on arm/chromosome level were showed in scatter plots. (B) The variation of CIN (left); The average variation of CIN on chromosome level across tumors (right), the lines connect the same chromosomes in two groups. (C) The variation of CNV log 10 transformed (the first plot); Variation of Log 10 transformed CNV from chromosome gain/loss events (the third and the fourth plots); The average variation in log 10 transformed CNV on chromosome across tumors (the second plot). (D) Total mutation burden (left) and corrected mutation burden by ploidy (right). The gray lines in all scatter plots represent the equal values between the two groups. Statistic test: Two-sided Wilcoxon rank-sum test.

3. Results

3.1. Genomic features associated with WGD in breast cancer

Chromosomal instability (CIN), copy number variation (CNV), and aneuploidy are commonly inherent features of unstable genome that are linked to tumorigenesis. The higher rates of somatic CNVs, CIN, and aneuploidy [9,35–37] are associated with whole genome doubling (WGD) events. To investigate the genomic differences between WGD+ and WGD-breast cancer (BRCA) samples, we obtained WGD status from TCGA for roughly 1000 primary tumors. Consistent with the previous estimates [9], the frequency of WGD in these tumors was found to be 44%. The tumors were divided into two groups based on their WGD status where 401 were positive (WGD+) and 586 were negative (WGD-) cases. We quantified CIN, CNV, and aneuploidy levels for each BRCA sample and explored the distinct patterns of genomic changes between WGD+ and WGD-tumors.

We observed that WGD + tumors exhibited a significantly higher aneuploidy burden than WGD-tumors ($P < 2e\text{-}16$), regardless of whether the ploidy abnormality resulted from a chromosome arm or entire chromosome (Fig. 1A). Furthermore, WGD + tumors displayed markedly a higher CIN compared to WGD-tumors ($P < 2e\text{-}16$). To delve deeper into the CIN patterns within each tumor sample and individual chromosomes, we compared the CIN between WGD+ and WGD-tumors. CIN at the chromosome level reflects copy number amplification, deletion, or both events. Notably, we found that each chromosome in WGD + tumors demonstrated elevated rates of CIN compared to near-diploid breast tumors (Fig. 1B). Additionally, we compared the CNV events based on chromosome gain, loss and both events in the two groups (Fig. 1C, Supplementary Fig. 2). The CNV (log10-transformed) was significantly higher in WGD + tumors ($P < 2e\text{-}16$) both at tumor and chromosome levels. Our findings suggest a strong association of WGD events with a higher frequency of genomic variations.

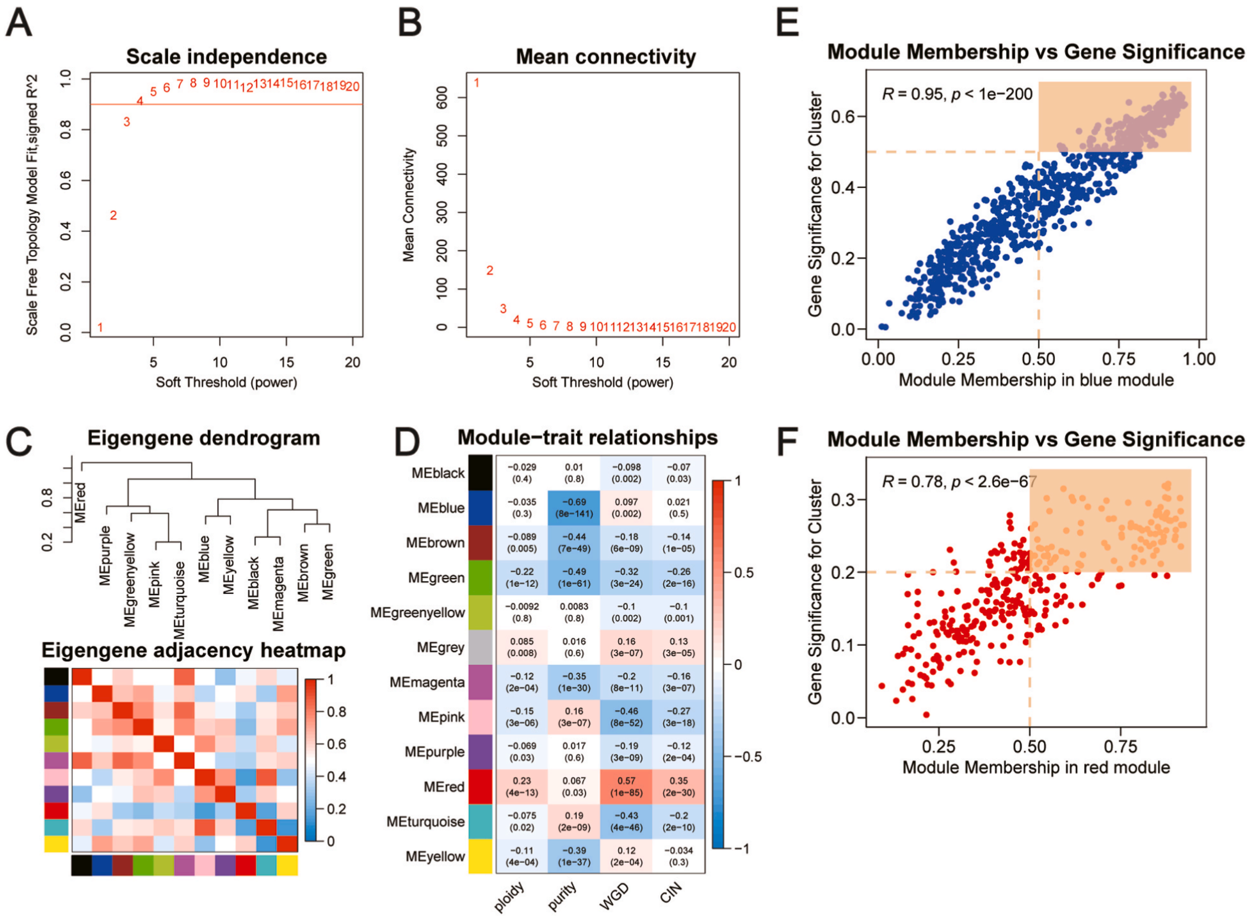


Fig. 2. Identification of WGD-related crucial genes via WGCNA procedure. (A and B) Analysis of network topology for various soft-threshold powers. (A) The impact of soft-threshold power on the scale-free topology fit index; (B) The impact of soft-threshold power on the mean connectivity. (C) Eigengene dendrogram and eigengene adjacency plot. (D) Correlation analysis between module eigengenes and WGD-related traits in TCGA breast cancer cohort. (E and F) The high correlation between GS and MM in the (E) blue module and the (F) red module. Dots within the two rectangle were defined as WGD-related crucial genes, with both high GS and MM. Statistic test: Pearson's correlation coefficient, two-sided unpaired *t*-test.

Along with somatic copy number variation events, we also investigated the association of mutations with WGD cases. The Wilcoxon rank-sum test revealed that WGD + tumors had a significantly higher tumor mutation burden (TMB) than WGD-tumors ($P < 2e\text{-}16$), irrespective of ploidy corrected or non-corrected TMB (Fig. 1D). These findings implied that WGD + breast tumors exhibited extensive genome abnormalities and may have implications for clinical diagnosis and prognosis of these patients. Given the poor survival outcome of the patients with WGD + tumors [9], we aimed to systematically explore and validate whether genes affected by WGD could function as potential biomarkers for the clinical diagnosis of BRCA patients.

3.2. *Identification of crucial genes derived from WGD status*

We conducted WGCNA analysis on the top 25% genes with the highest expression variation, utilizing a soft threshold (β) of 0.9 for co-expression network construction (Fig. 2A and B). Along with, we applied linear models to the gene expression data and identified 8808 genes associated with WGD + tumors. Further we identified 11 modules represented by distinct colors, each containing no less than 30 genes. The eigengene, which is the first principal component of gene expression within a module, was considered as the representative of the module. The heatmap displayed the eigengene adjacency of the modules (Fig. 2C). We also identified the association between modules and the tumor genome traits, such as ploidy, purity, WGD status, and CIN. The red module exhibited the highest significant positive correlation (Corr = 0.57, P = 1e-85), while the blue module displayed the highest significant negative correlation (Corr = −0.69, P = 8e-141) in the module-trait relationship (Fig. 2D). In addition, the red module containing 323 genes had a high correlation coefficient between gene significance (GS) and module membership (MM) of 0.78 (P < 2.6e-67), and the blue module with 777 genes had a correlation coefficient of 0.9 (P < 1e-200), indicating the quality of gene module construction (Fig. 2E

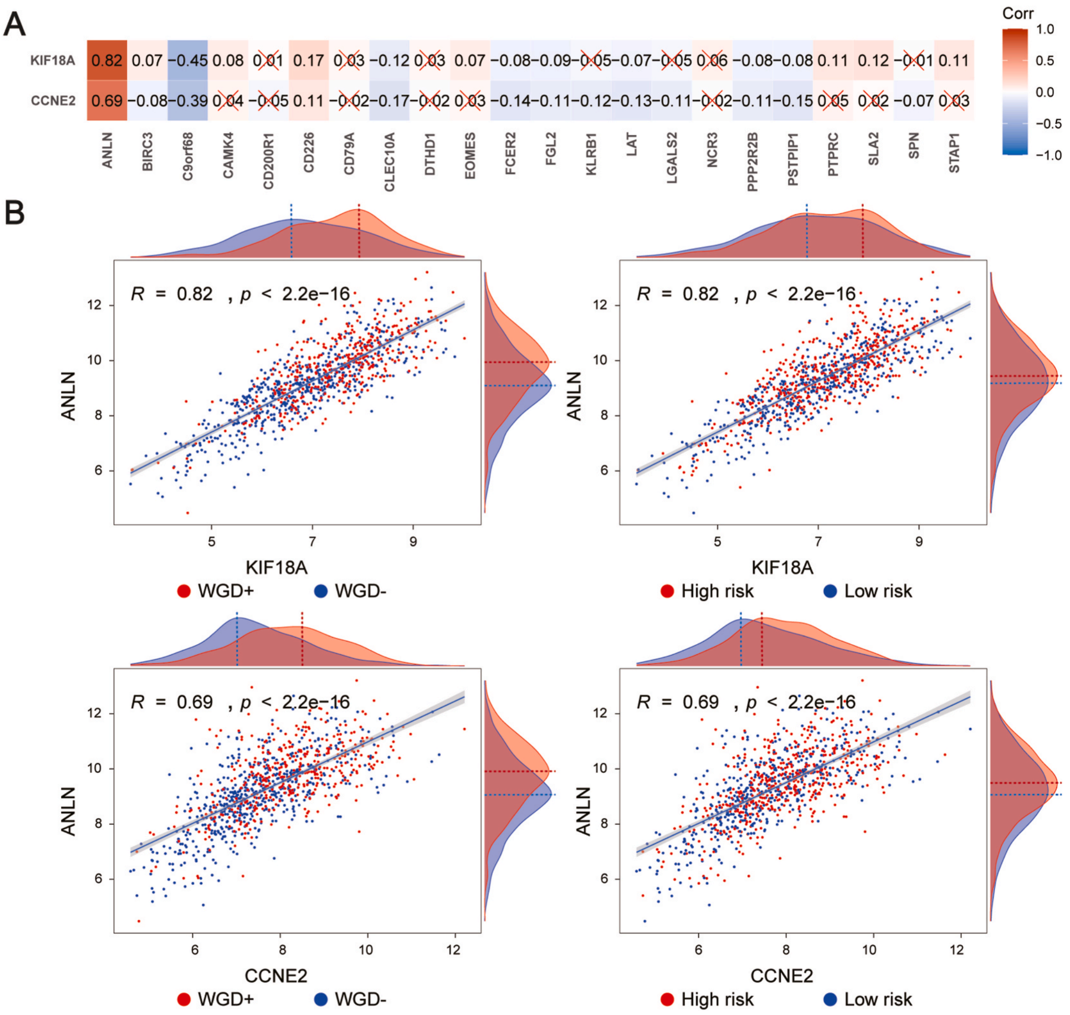


Fig. 3. Correlation analysis of 22 risk genes with WGD crucial genes. (A) Correlation heatmap between risk genes and KIF18A and CCNE2. The pearson correlation coefficient is shown in each little box. The cross indicates a non-significant correlation coefficient ($P \geq 0.05$). (B) Correlations between ANLN with KIF18A and CCNE2 grouped by WGD status and risk score. The density represents the distribution of the expression of corresponding genes among tumors with different groups. Dot lines point to the peak of the density.

and F). To identify the hub genes associated with WGD patterns in the two significant modules, we considered 76/323 genes with GS > 0.2 and MM > 0.6 in the red module and 233/777 genes with GS > 0.5 and MM > 0.5 in the blue module as candidate WGD genes. These 309 candidate genes were mainly selected based on the significant association marked in the upper right quarter of the two modules (Fig. 2E and F). We combined 8808 genes derived from linear models with the 309 genes, 247 overlapping genes were identified and marked as WGD-related crucial genes.

3.3. Identification of prognostic genes associated with WGD status for BRCA patients

We generated a prognostic model in order to identify the impact of 247 WGD-related crucial genes on the survival of BRCA patients (for TCGA data, details in methods section). The data was divided into training and test cohorts using random sampling with 100 iterations. This resulted in 100 prognostic univariate Cox regression models along with LASSO regression models for the TCGA BRCA training cohort. For each model, a 100 C-index was along generated. The model with the highest C-index (C-index = 0.73) from 100 iterations was selected as the optimal prognostic model for the training cohort. Univariable Cox regression analysis identified a set of 122 genes out of 247 WGD-related crucial genes that were significantly correlated with OS (*P*-value <0.001) after being adjusted for multiple testing by the BH procedure. While LASSO Cox regression model identified 22 genes out of 247 as the most useful risk markers associated with WGD status in terms of prognosis in BRCA patients. For the LASSO model, we depicted the resultant change trajectory of each independent variable and the confidence intervals under each lambda (Supplementary Fig. 3). The 22 risk genes were related to WGD status with minimized lambda (0.008959). Among them, 21 genes have been recorded in the DisGeNET [38] database (v7.0), excluding C9orf68, and 8 risk genes (ANLN [39], BIRC3 [40], CD200R1 [41], CD226 [42], CLEC10A [43], PPP2R2B [44], PTPRC [45], SPN [46]) are already reported for BRCA. We also identified 8 other genes (CD79A [47], EOMES [48], CAMK4 [49], FGL2 [50], KLRB1 [51], LGALS2 [52], SLA2 [53], STAP1 [54] that have been reported in previous BRCA studies. In addition, three cancer genes (BIRC3, CD79A, and PTPRC) overlapped with the Cancer Gene Census in COSMIC (v97) [55], further supporting their relevance in the context of cancer, majorly BRCA. Supplementary Table 1 summarized the relevance of risk genes and diseases in particular with BRCA.

3.4. Potential correlation of risk genes with KIF18A and CCNE2 in BRCA

Previous studies have shown that the depletion of KIF18A(Kinesin Family Member 18A) leads to significantly prolonged mitoses in tetraploid cells, potentially contributing to tumor development and progression [28]. Likewise, CCNE2 (Cyclin E2) has been associated with high genome ploidy in breast cancers, and its overexpression promotes aberrant mitosis, suggesting its role as a driver of genome doubling in cancer [56]. Notably, both KIF18A and CCNE2 have been found to be overexpressed in human BRCA and are associated with worse prognosis for cancer patients [56,57]. Therefore, we tried to investigate the association of 22 risk genes in WGD with KIF18A and CCNE2 in the TCGA BRCA data set. Our analysis revealed that a total of 15 risk genes (ANLN, BIRC3, C9orf68, CAMK4, CD226, CLEC10A, EOMES, FCER2, FGL2, LAT, PPP2R2B, PSTPIP1, PTPRC, SLA2, STAP1) exhibited a significant correlation with KIF18A expression, while 13 genes (ANLN, BIRC3, C9orf68, CD226, CLEC10A, FCER2, FGL2, KLBR1, LAT, LGALS2, PPP2R2B, PSTPIP1, SPN) showed the significant association with CCNE2 expression (Fig. 3A). Remarkably, ANLN displayed a highly positive correlation with both KIF18A (Cor = 0.82, P < 2.2e-16) and CCNE2 (Cor = 0.69, P < 2.2e-16) expression. Meanwhile, tumors exhibiting high expression of ANLN or CCNE2 were found to be more prone to WGD events and displayed a higher clinical risk score (Fig. 3B). In contrast, the remaining genes did not exhibit the same trend. ANLN has been reported to play a crucial role in cell cycle progression in primary BRCA [58]. It exhibits high nuclear expression in breast tumor cells and is significantly associated with high histological grade, elevated proliferation rate, and poor prognosis [59]. Experimental knockdown of ANLN remarkably inhibited cell proliferation and migration as well as cell invasion, arrested the cells in G2/M phase, and induced apoptosis in BRCA cells [39]. These findings led us to speculate that ANLN may contribute to the occurrence of WGD; however, further validation is required.

Additionally, we conducted an exhaustive literature review to elucidate the predicted correlations. The PI3K-AKT signaling pathway is frequently dysregulated in cancer, with hyperactivation observed in approximately 50% of breast cancers [60,61]. Studies indicate that the overexpression of KIF18A can promote the activation of the PI3K-AKT signaling pathway [62], subsequently influencing the nuclear localization and stability of ANLN, a protein positively regulated by the PI3K-AKT pathway [63–66]. This is in concordance with a positive correlation between the expression of KIF18A and ANLN. Furthermore, numerous studies have reported the overexpression of ANLN as a potential biomarker of lung cancer progression [67,68]. The elevated expression of CCNE2 is also significantly correlated with advanced tumors and worse OS in lung cancer [69,70]. Berberine (BBR) treatment have inhibited the activity of PI3K/AKT pathway by suppressing CCNE2 expression, demonstrating the inhibitory impact on the progression of NSCLC [71]. This leads to a positive correlation between the expression of CCNE2 and ANLN, which is positively regulated by the PI3K/Akt pathway. However, the biological regulatory mechanism underlying the positive correlation between them in BRCA needs further validation. We also found that reduced PTEN and PPP2R2B expression was associated with activated AKT/mTOR and PDK1/MYC pathways [72], which provided a potential foundation for validating the inverse correlation between PPP2R2B and CCNE2, given that CCNE2 is highly expressed in the activated PI3K/AKT pathway.

3.5. The predictive power of the risk genes

To assess the forecasting performance of the 22 risk genes, we conducted a clinical stratified analysis using TCGA BRCA training cohort, focusing on three clinicopathological features: stage, age, subtype, and WGD status (Supplementary Fig. 4). We identified no significant differences between high and low-risk groups in terms of OS for the patients with the normal and basal subtype. On the contrary remaining 9 clinical groups revealed that a high-risk score was associated with a poor prognosis compared to a low-risk score.

In addition, we investigated the relationship between these clinicopathological features and the risk genes across the training, test, and entire TCGA cohort. Using the same risk score formula as implied in the training set, the patients were stratified into high and low-risk groups in both the test and entire cohort. Kruskal-Wallis test revealed a distinct distribution of risk scores among BRCA patients with varying stages, age groups, and cancer subtypes. Patients with older age, and the luminal B subtype exhibited significantly higher risk scores (Fig. 4A). Further, in all three cohorts, we observed striking disparities between the high and low-risk groups in terms of patient distribution related to age and subtypes. The Sankey association diagram illustrated the flow from the two risk subgroups to different clinical outcomes, age, and clinical subtypes (Fig. 4B).

3.6. *The protective power of the optimal model*

A prognostic model was generated based on the linear combination of 22 risk-associated genes. In the TCGA training set, patients were categorized as high or low risk based on the median risk score. The heatmap visually depicted the expression patterns of the 22 risk-associated genes in the high and low-risk groups (Fig. 5A). The expression levels of 21 genes decreased with the increasing risk scores, with ANLN being an exception. The expression of ANLN significantly increased with the increase in risk scores (Fig. 5B) ($R = 0.16$, $P = 9.859e\text{-}07$). Further we identified the hazard ratios (HR), 95% CI, and the respective *P*-values by log-rank tests for the 22 genes as shown in a forest plot (Fig. 5C). Of the 22 genes, 21 were identified as protective factors with $HR < 1$, whereas ANLN emerged as a risk factor with $HR > 1$. To assess the clinical implications of the risk score, we performed a Kaplan-Meier survival analysis comparing the two risk groups within the training cohort. More death events were observed in the high-risk group, indicating that low-risk patients experienced better clinical outcomes than high-risk patients (Fig. 5D).

To evaluate the reliability of the optimized trained model, comprehensive tests were conducted using both internal TCGA test set

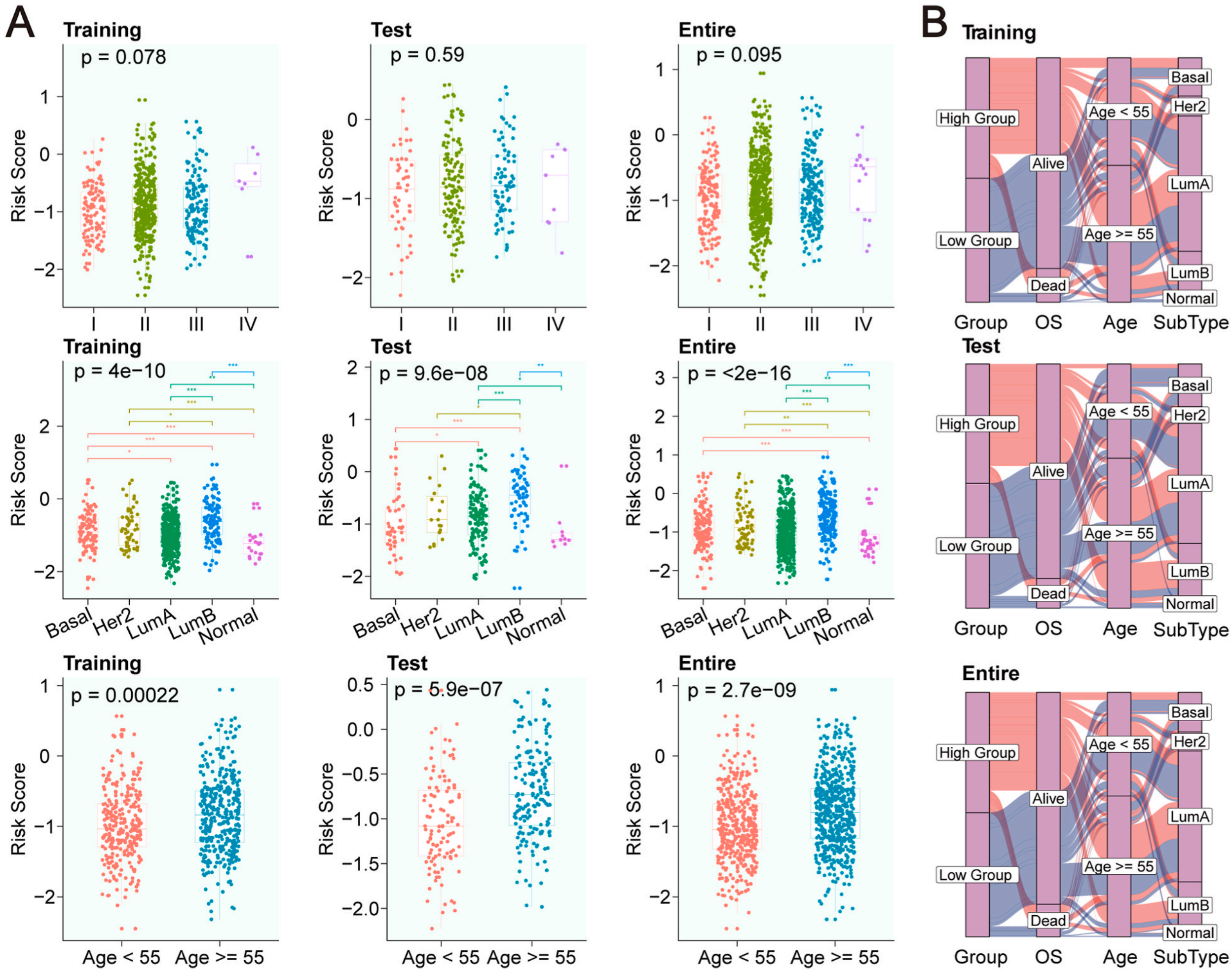


Fig. 4. Associations between clinicopathological features and the risk genes in the training, testing, and entire cohorts. (A) Variation pattern of risk score in different clinicopathological groups (stage, age and subtype). (B) Sanky plots of risk score and significant clinicopathological features (age and subtype).

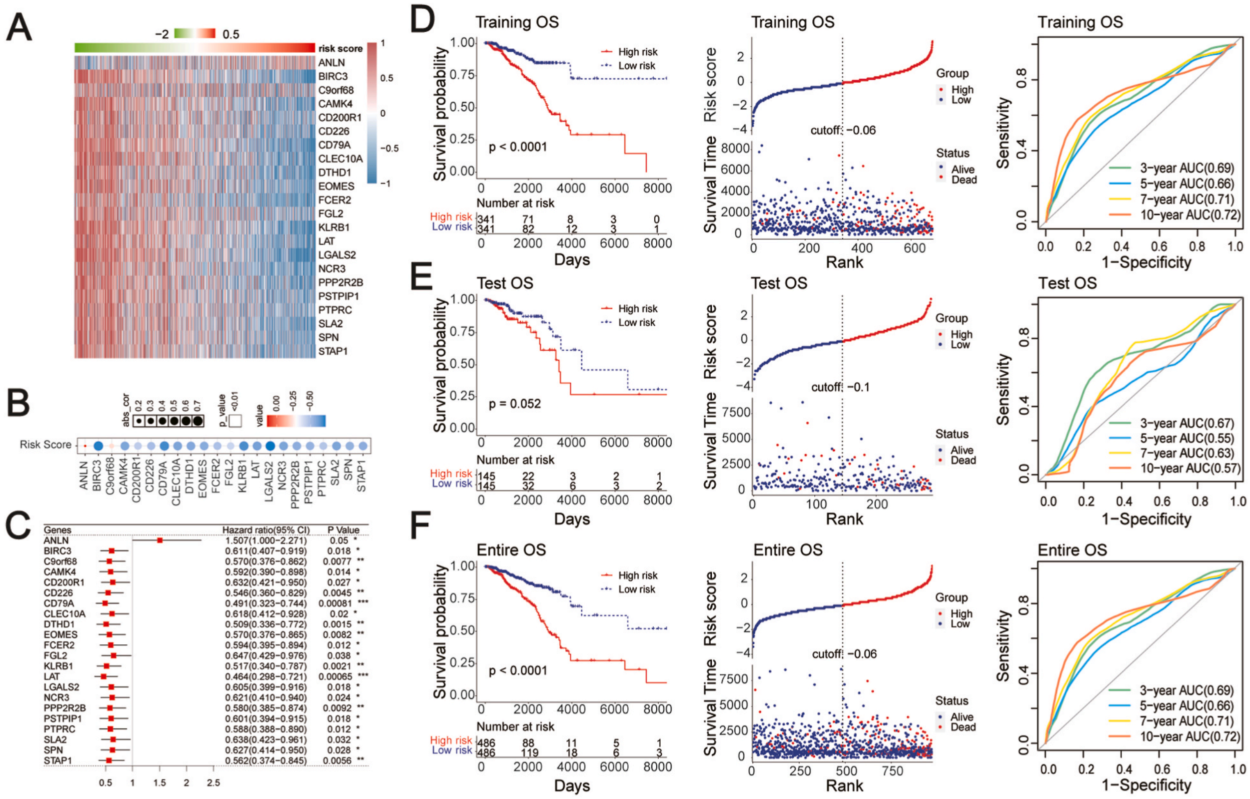


Fig. 5. Determination of risk gene signatures associated with WGD status in TCGA breast cancer patients. (A) The expression pattern of 22 genes constructing the risk score. (B) Correlations between 22 genes and the risk score. (C) Forest plot of the prognostic ability of the 22 WGD-related risk genes. (D–F) The Kaplan–Meier curves of OS according to the risk score in (D) training cohort (log-rank test: $P < 0.0001$), (E) testing cohort (log-rank test: $P = 0.052$) and the (F) entire cohort (log-rank test: $P < 0.0001$). The risk score distribution and patient survival status are depicted in ranked dot and scatter plots in the middle from D to F. Time-dependent ROC analysis for predicting OS at 3-, 5-,7- and 10 year are on the right from D to F.

(validation) and entire TCGA cohort. In the test cohort, we identified 20 risk genes that exhibited a significant differential expression distribution between the two risk groups. Similarly, in the training and entire datasets, 22 risk genes displayed the differential patterns (Supplementary Fig. 5). Consistent with the findings in the training cohort, patients stratified into the high-risk group exhibited significantly worse OS than those in the low-risk group, as confirmed by the log-rank test in both the test cohort ($P = 0.052$) and the entire cohort ($P < 0.0001$) (Fig. 5E and F).

The ROC analysis indicated that the risk genes are powerful in predicting patient OS in the training cohort, with AUC values of 0.69, 0.66, 0.71, and 0.72 for three, five, seven, and ten-year OS, respectively (Fig. 5D). Similar trends were observed in the test and entire cohort, reaffirming the accuracy of the prediction model (Fig. 5E and F). Furthermore, the model's predictive ability was also assessed using PFI in the three cohorts which led to consistent results (Supplementary Fig. 6). To confirm the robustness of our findings, we validated the model's performance on eight independent GEO validation data sets and METABRIC datasets. Remarkably, the high-risk group consistently demonstrated significantly poor OS and the AUC values obtained were consistent with those observed in the TCGA dataset for all independent validation data sets (Fig. 6A and B and Supplementary Fig.7).

3.7. Nomogram establishment and validation with clinical features

To enhance the clinical applicability of the identified WGD-related risk genes and to provide visualized risk prediction, we developed a quantitative analysis algorithm. The method aimed at predicting the expected survival of individuals with BRCA using patients from the training, test, and entire cohort. By univariate and multivariate Cox regression analyses, we selected three significant clinical factors (age, stage, and risk score) in the training cohort. The HR and P-value of these factors were illustrated in Fig. 7A and B. The three factors were further incorporated to calculate the individual sample's summary score and the total score across the three cohorts, based on three, five, seven and ten-year survival probabilities in each cohort (Fig. 7C and Supplementary Fig. 8). Notably, the risk score contributed the most to risk points compared to age and stage information in the training and entire cohort.

The calibration curves of three, five, seven, and ten-year OS in both the training and entire cohorts demonstrated near-optimal performance. Moreover, the calibration curves of three and five-year OS in the test cohort also exhibited a good fit, indicating

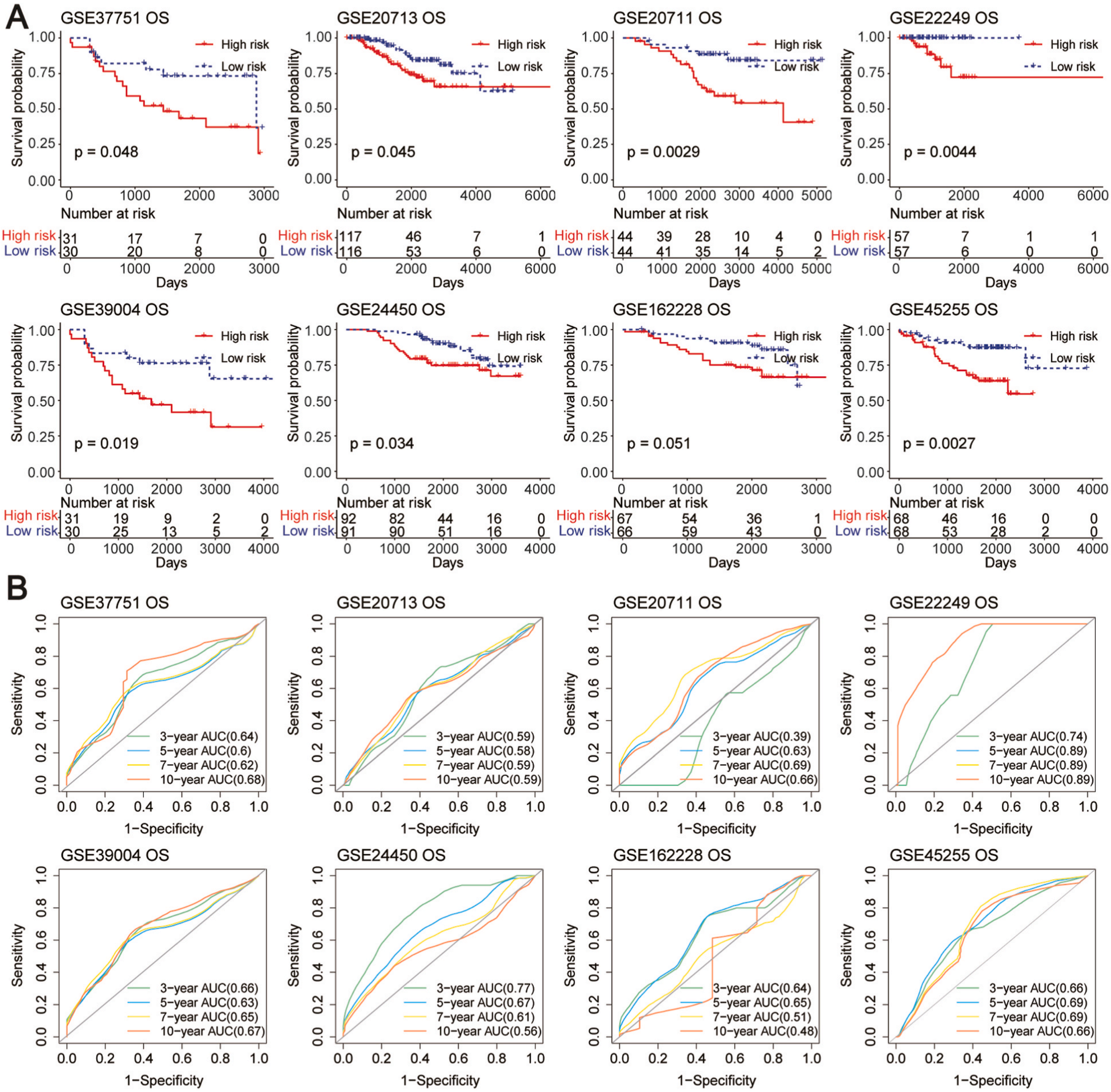


Fig. 6. The risk model's ability for prediction of OS for breast cancer patients in external cohorts. (A) Kaplan-Meier curves of OS and (B) the time-dependent ROC analysis for predicting OS at 3-, 5-, 7- and 10-year were showed for each cohort.

consistency between the actual measured prognostic value and the value projected by the nomogram (Fig. 7D and Supplementary Fig. 8). The decision curves further demonstrated that the nomogram had a higher net benefit in predicting OS probability compared to either the treat-all-patients scheme or the treat-none scheme, except for the results of seven and ten-year in the test cohort (Fig. 7E and Supplementary Fig. 8). The prognostic performance of the nomogram for PFI in the training, test and entire cohorts were also assessed. Supplementary Fig. 9 illustrated a similar trend of PFI with OS in the three cohorts, with the risk score emerging as the most important factor affecting patient survival, followed by tumor stage and age. The calibration curves of three, five, seven, and ten-year PFI in the three cohorts performed similarly to those of OS, indicating the reliability of the nomogram's predictions.

The ROC analysis was further used to evaluate the accuracy of the nomogram. The predicted AUC values of the three, five, and ten-year OS nomograms in the three cohorts were comparable to or higher than those of the risk score or the clinical features (age and stage) individually (Fig. 7F). This indicated that the nomogram outperformed the other predictors for predicting BRCA patient survival. Additionally, the C-indices of these models were 0.8, 0.77, and 0.79 in the three cohorts respectively, demonstrating the robust predictive power of the nomogram for TCGA-BRCA. Overall, the nomogram based on WGD-related risk gene-based risk scores provided valuable guidance for clinical diagnosis and survival prediction in BRCA patients.

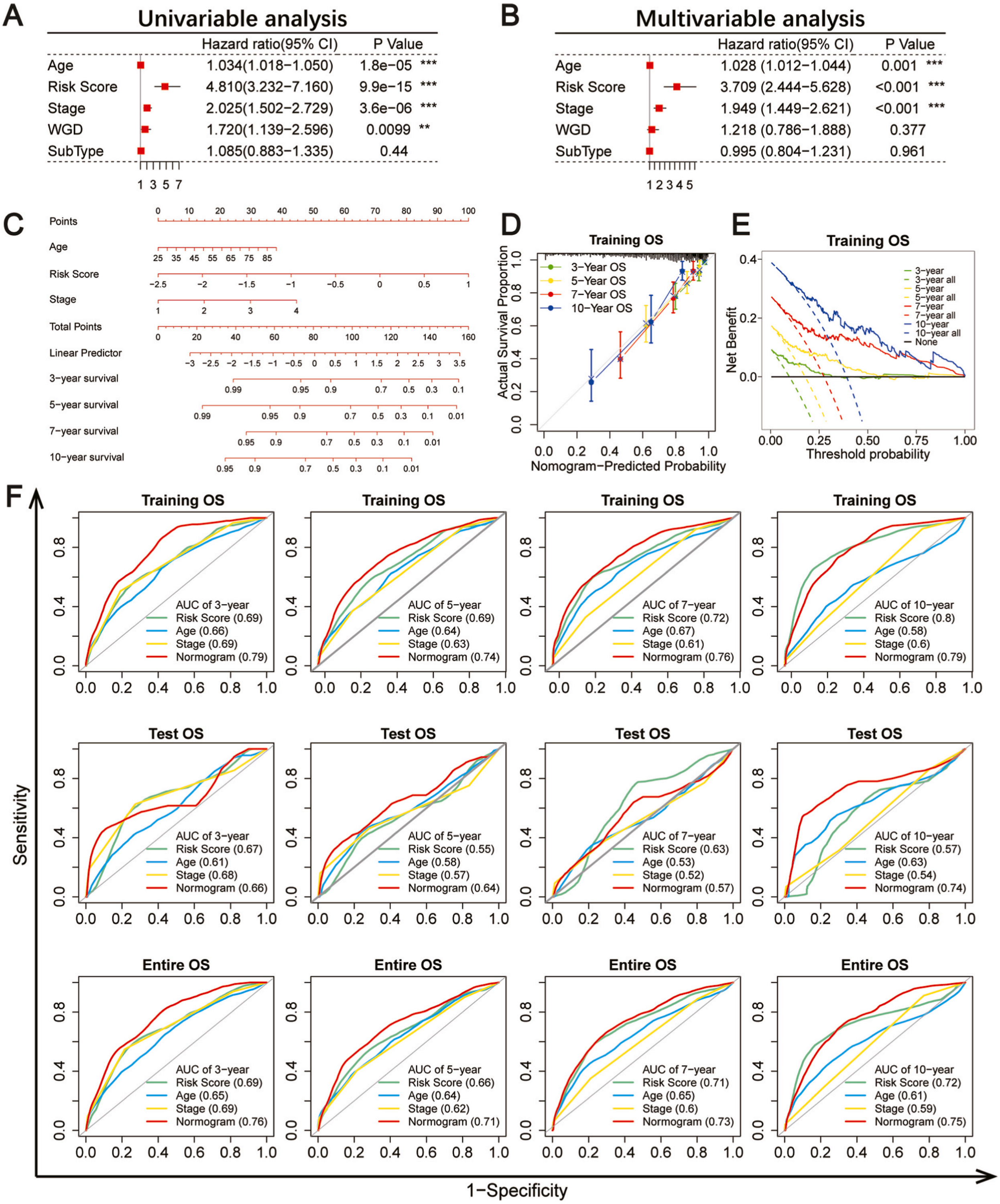


Fig. 7. Establishment and verification of Nomogram for 3-, 5-,7- and 10-year OS in TCGA training, test and entire cohorts. (A and B) Univariate and multivariate Cox regression analysis for determination of significant clinical factors. (C) The nomogram was constructed with the significant clinical factors and risk score incorporated. (D) Calibration plot of the nomogram in terms of agreement between the predicted and observed 3-, 5-,7- and 10-year outcomes. The 95% confidence intervals were represented by the close-ended vertical lines, the nomogram predicted and actual OS were illustrated on the x-axis and the y-axis. The dashed line along the 45-degree line represents the ideal performance of a nomogram. (E) Decision curve analysis of the nomogram for 3-, 5-,7- and 10-year risk. The black line represents the assumption that no patients died at 3-, 5-,7- or 10-year. (F) ROC curves of the nomograms compared with those of other clinical variables with regard to 3-,5-,7- and 10-year OS.

3.8. Genomic feature alteration among risk groups in patients with BRCA

Investigating the differences in genome features between high- and low-risk patients, we conducted a Wilcoxon test to assess aneuploidy burden and CIN in the training, test, and entire cohort. In both the training and entire cohorts, the high-risk group exhibited significantly higher levels of both features. However, only aneuploidy burden showed a similar distribution between the high- and low-risk groups in the test cohort ($P = 0.00022$) (Fig. 8A).

Furthermore, we analyzed the association (Pearson correlation) between aneuploidy burden and CIN with risk scores in the three cohorts and found consistently significant associations across all three cohorts. Moreover, we observed a higher concentration of patients with WGD events in the high-risk tumors (Fig. 8B). To validate this observation, we performed a chi-square test comparing the two groups of risk patients with different WGD status, revealing a higher number of WGD + tumors in the high-risk group compared to the low-risk group (Supplementary Fig. 10A). However, we only observed a significant *P*-value in the training ($P = 0.003$) and entire cohorts ($P = 0.003$), which may be attributed to the limited sample size of the test set. Additionally, we noted a higher concentration of

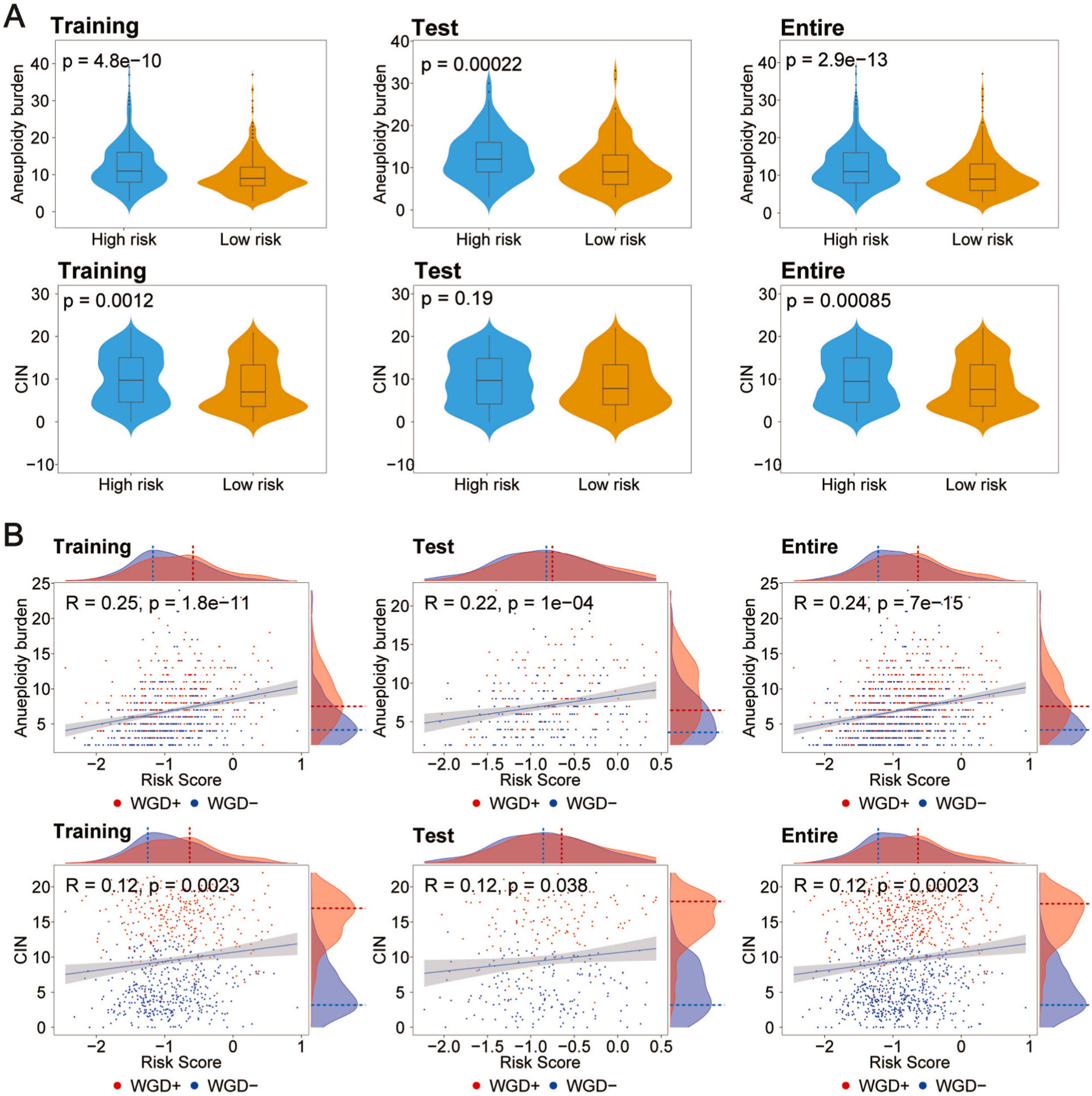


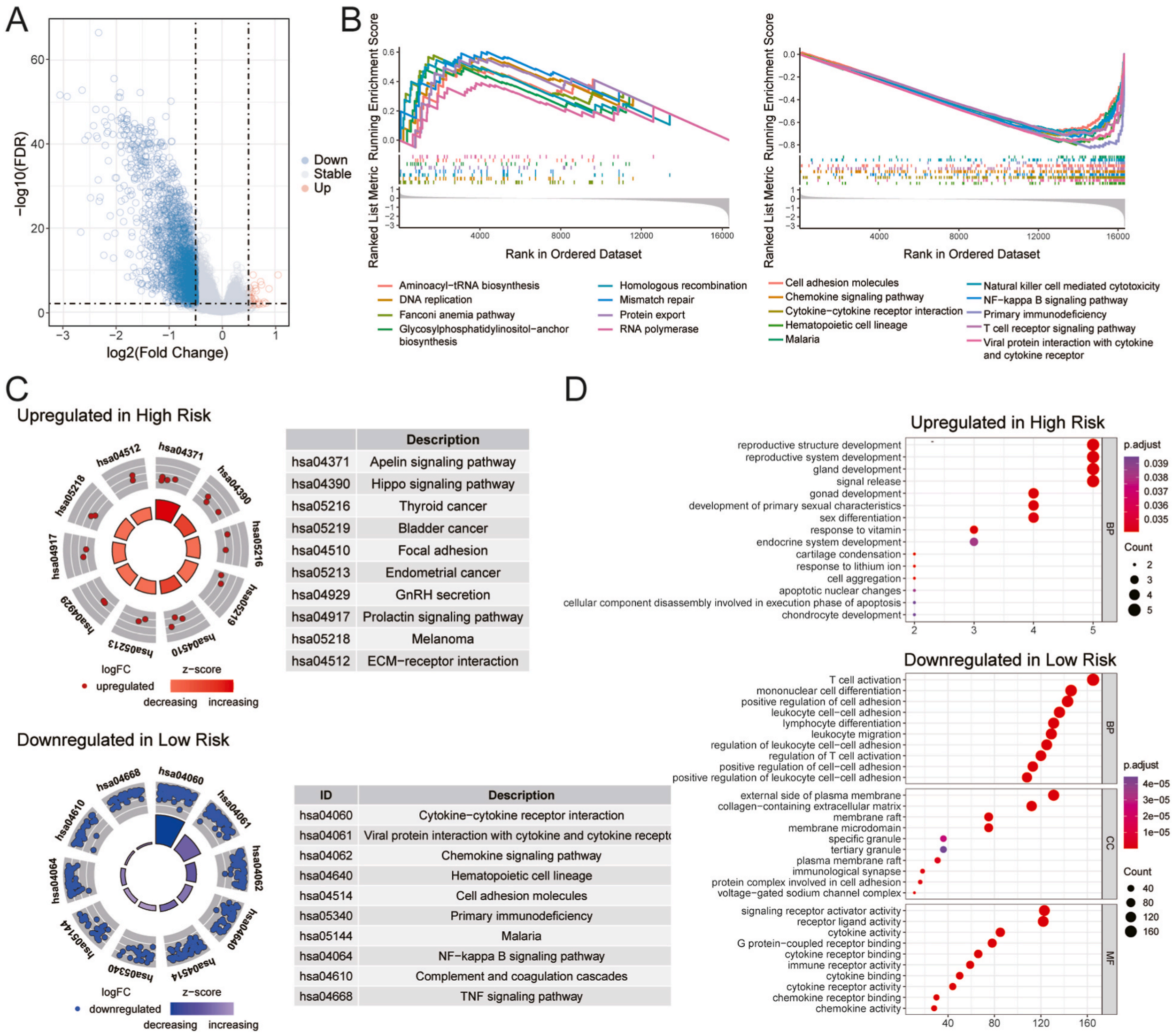
Fig. 8. Genome features analysis between high- and low-risk patients in the training, test and entire cohort. (A) Variation of aneuploidy burden and CIN grouped by risk score. (B) Correlation of aneuploidy burden and CIN with risk score. Dot lines point to the peak of the density.

patients with mutations in the low-risk group (Supplementary Fig. 10B). These findings revealed substantial variations in genomic features between the low- and high-risk groups.

3.9. Functional enrichment analysis of the WGD-related genes

To explore the functional implications of 247 WGD-related crucial genes, we conducted comprehensive GO and KEGG pathway analyses. The GO enrichment analysis revealed that 247 WGD-related genes were predominantly enriched in cellular components closely associated with cell replication, such as chromatin, centromeres, and kinetochores. Furthermore, these candidate genes were also enriched in biological processes related to immune responses, including T cell activity, lymphocyte differentiation, T cell differentiation, and immune response regulation (Supplementary Fig.11). KEGG pathway analysis unveiled the involvement of candidate genes in crucial biological pathways, such as cytokine receptor interactions, T cell receptor signaling pathway, Th17/Th1/Th2 cell differentiation, primary immune deficiency, and cell cycle. We further explored the underlying mechanisms that contributed to the different outcomes stratified by the 22 risk genes by performing KEGG pathway, GSEA, and GO analyses.

Moreover, the differential expression analysis of the TCGA training cohort identified 2128 down-regulated genes (DEGs) and 33 up-regulated genes (DEGs) between the high- and low-risk groups (Fig. 9A). The KEGG pathway analysis of the DEGs highlighted significant alterations in pathways associated with cytokine-cytokine receptor interactions, the chemokine signaling pathway, and the



	Description
hsa04371	Apelin signaling pathway
hsa04390	Hippo signaling pathway
hsa05216	Thyroid cancer
hsa05219	Bladder cancer
hsa04510	Focal adhesion
hsa05213	Endometrial cancer
hsa04929	GnRH secretion
hsa04917	Prolactin signaling pathway
hsa05218	Melanoma
hsa04512	ECM−receptor interaction

ID	Description
hsa04060	Cytokine−cytokine receptor interaction
hsa04061	Viral protein interaction with cytokine and cytokine recepto
hsa04062	Chemokine signaling pathway
hsa04640	Hematopoietic cell lineage
hsa04514	Cell adhesion molecules
hsa05340	Primary immunodeficiency
hsa05144	Malaria
hsa04064	NF−kappa B signaling pathway
hsa04610	Complement and coagulation cascades
hsa04668	TNF signaling pathway

Fig. 9. Functional analysis of DEGs based on risk genes related with WGD status between high- and low-risk breast cancer patients in TCGA training cohort. (A) Volcano map of DEGs between the high- and low-risk groups. (B) GSEA enrichment plots of the high- and low-risk groups. (C) KEGG pathways enriched in the high- and low-risk groups. (D) The 15 most significantly enriched GO terms in high-risk group and three groups of top 10 significantly enriched GO terms in low-risk group.

PI3K-Akt signaling pathway within the low-risk subgroup (Fig. 9C). However, patients with high-risk scores demonstrated convergence towards tumor-related pathways, including the Apelin signaling pathway, Hippo signaling pathway, and ECM-receptor interaction, which have potential associations with BRCA [73] (Fig. 9C). The GO analysis further demonstrated that many biological functions in low-risk patients were primarily associated with immune-related biological processes and molecular functions (Fig. 9D), highlighting the potential involvement of immune mechanisms in determining the divergent outcomes. Additionally, the independent GSEA analysis revealed distinct clustering patterns in gene sets related to aminoacyl-tRNA biosynthesis, DNA replication, GPI-anchor biosynthesis, homologous recombination, mismatch repair, and RNA polymerase among high-risk patients. In contrast, signaling pathways promoting tumor immunity, including the T cell receptor signaling pathway, primary immunodeficiency, and NF-kappa B

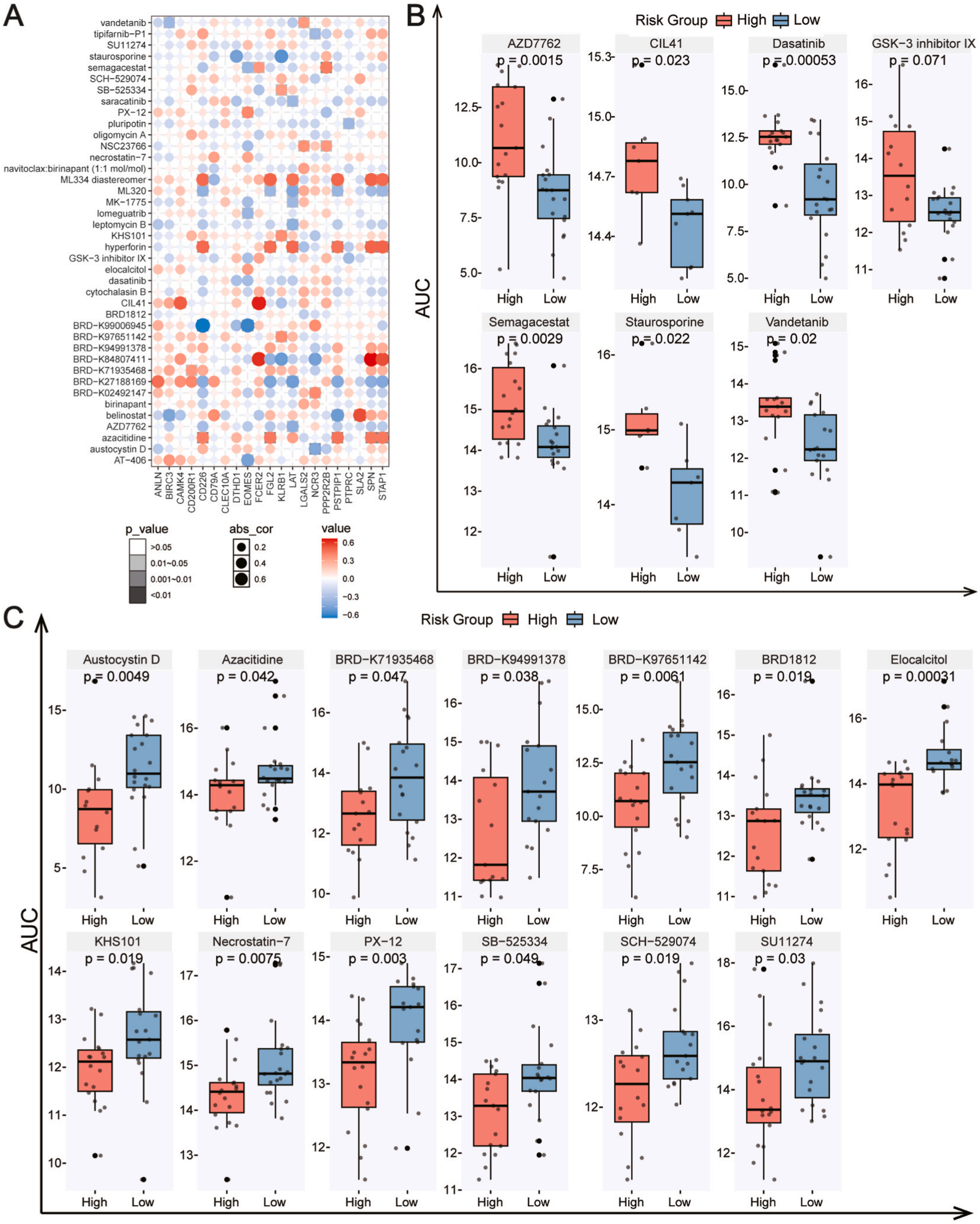


Fig. 10. Drug response of risk genes related with WGD status in breast cancer. (A) Correlation between 22 risk genes expressions and drug sensitivities (AUC). (B and C) Comparison of difference in AUC values of 7 drug components (B) and 13 drug components (C) for treating 40 breast cancer cell lines.

signaling, were more prevalent in low-risk patients (Fig. 9B).

3.10. Analysis of the correlations between WGD-Related risk genes and drug sensitivity

To better understand the impact of the risk genes on drug response, we conducted a comprehensive drug sensitivity analysis based on the expression levels of risk genes across 40 BRCA cell lines obtained from CCLE and the corresponding drug sensitivity data (AUC) from CTRP. Our analysis identified 21 risk genes that exhibited significant correlations with the response to 40 different anticancer drugs (P-value <0.05) (Fig. 10A), indicating the pivotal role of these risk genes in drug response. Moreover, we observed that the AUC values of seven drugs exhibited a significant increase with higher risk scores (Supplementary Fig. 12A). Specifically, the AUC values of these seven drugs in the low-risk group were significantly lower than those in the high-risk group (Fig. 10B). The drugs with this correlation included AZD7762 ($P = 0.0052$), BRD-K84807411 ($P = 0.014$), CIL41 ($P = 0.023$), dasatinib ($P = 0.0061$), staurosporine ($P = 0.022$), semagacestat ($P = 0.0029$), and vandetanib ($P = 0.02$). These results suggest that BRCA cell lines with low-risk genes exhibited an increased sensitivity to these specific drugs. Additionally, the AUC values of 13 drugs, including Austocystin D ($P = 0.027$), azacitidine ($P = 0.042$), BRD-K71935468 ($P = 0.047$), BRD-K94991378 ($P = 0.038$), BRD-K97651142 ($P = 0.011$), BRD1812 ($P = 0.019$), elocalcitol ($P = 0.00031$), KHS101 ($P = 0.019$), necrostatin-7 ($P = 0.0075$), PX−12 ($P = 0.003$), SB−525334 ($P = 0.0049$), SCH−529074 ($P = 0.0019$), and SU11274 ($P = 0.03$)), displayed significantly lower AUC values in the high-risk group compared to the low-risk group, exhibiting a strong negative correlation with the risk score. These findings indicate that the BRCA cell lines with high-risk genes are more sensitive to these 13 specific drugs (Fig. 10C and Supplementary Fig. 12B).

More importantly, our analysis revealed that 8 out of the 20 drugs had been previously reported for the BRCA treatment. Four drugs exhibited significantly lower AUC values in the low-risk group, the remaining exhibited significantly lower AUC values in the high-risk group (Supplementary Table 2). For example, Vandetanib, a multitargeted tyrosine kinase inhibitor with anti-tumor activity in preclinical models of BRCA [74], exhibited significantly higher AUC values in the high-risk group, suggesting that patients in the low-risk group may be more responsive to this drug. Similarly, SU11274, a MET inhibitor, which along with various EGFR inhibitors resulted in synergistic suppression of cell viability and cell survival of MSL subtype TNBC cells [75]. The lower AUC of SU11274 in high-risk group implied its potential efficacy in treating tumors within this subgroup. These results indicate that the risk genes associated with WGD may provide valuable predictive information for drug sensitivity, thereby reducing the risk of adverse drug reactions in patients with BRCA. However, further validation is necessary for these findings and to develop more effective and precise treatment strategies for BRCA.

4. Discussions

In this study, we identified breast cancer-specific genes influenced by WGD using a combination of linear models and gene co-expression network analysis. Furthermore, we investigated a set of WGD-specific genes associated with the clinical survival of BRCA patients. Through comprehensive evaluation and validation using cohorts from TCGA-BRCA, GEO, and METABRIC, we identified 22 genes that exhibited broad and favorable prognostic effects (Supplementary Fig. 13). Among these genes, ANLN stood out as an interesting candidate. ANLN upregulation has been observed in 21 types of cancers and has been associated with poor OS, DFI, and PFI in most cancers [76]. Multiple studies have proposed ANLN as a potential molecular marker for predicting breast cancer diagnosis [59,77]. Additionally, ANLN has been recognized as a crucial regulatory factor involved in various signaling events, particularly those related to the cell cycle and nucleocytoplasmic transport pathways [78].

Meanwhile, WGD is often attributed to potential errors during cell division, such as mitotic slippage and cytokinesis failure [79]. Defects in the G1 checkpoint can lead to replication and give rise to different malignant phenotypes [3], ultimately impacting the cell cycle. We discovered a strong positive correlation between ANLN expression and KIF18A, a cell division driving protein specific to WGD cells [31,80], as well as CCNE2. These findings suggest a potential association between ANLN overexpression and the generation of WGD in BRCA patients.

Subsequently, we identified that three cancer genes (BIRC3, CD79A, and PTPRC) from the risk-genes were reported in the Cancer Gene Census from COSMIC (v97) [55]. BIRC3 belongs to the inhibitor of apoptosis (IAP) family of proteins. Studies have demonstrated that high expression of BIRC3 can lead to increased infiltration of immune cells [81]. Interestingly, we observed a significant upregulation of BIRC3 in the low-risk group, suggesting its potential role in enhancing the inhibition of tumor cell apoptosis and promoting the development of a tumor immune microenvironment. CD79A is a protein primarily expressed in B cells and plays a critical role in activating the immune response. In triple-negative breast cancer, CD79A expression may be associated with an immune response [82]. Notably, CD79A-positive triple-negative breast tumors were found to have a higher density of tumor-infiltrating lymphocytes (TILs) and improved patient survival [47]. PTPRC, also known as CD45, is a protein tyrosine phosphatase expressed on the surface of all nucleated hematopoietic cells. It played a critical role in immune cell signaling and activation. A study revealed that low PTPRC expression was associated with worse OS and DFI in BRCA patients [83]. The study also found a positive correlation between PTPRC expression and immune cell infiltration, indicating that PTPRC may play a role in the immune response to BRCA. In line with these findings, we observed consistent lower expression trends of CD79A and PTPRC in the high-risk group. These studies supported the novel WGD–related risk genes as a potentially measurable prognostic indicator in patients with BRCA, which may provide a theoretical basis for improving the poor prognosis of patients.

In this study, we constructed a predictive clinical model based on a set of 22 identified risk genes. To assess the efficacy of our proposed model, we also employed three distinct WGD-related gene sets, encompassing the entire gene set, the gene set filtered by our linear model, and the gene set filtered by our WGCNA approach. Subsequently, univariate and multivariate COX regression analyses were conducted on these diverse WGD-related gene sets [84,85]. The significant gene sets with the same number of risk genes as our method were selected, and the C-index derived from these gene sets was employed to evaluate the performance of our constructed model. We calculated and compared the C-index values of models generated through various recombination methods on both the TCGA training dataset and the testing dataset. The highest C-index values achieved by these strategies on the TCGA training set and test set were 0.65 and 0.63, respectively, which did not surpass those obtained from our method. The results revealed that our method yielded a superior C-index compared to alternative strategies (Supplementary Table 3). Furthermore, we determined the AUC values for different strategies predicting the three, five, seven, and ten-year OS of patients in eight GEO datasets. The comparative analysis revealed a significant advantage in the indices calculated by our method, providing further validation of the effectiveness of the predictive model we constructed (Supplementary Table. 4).

In this study, we developed a scoring system based on the 22 risk genes to stratify patients into various risk groups, which revealed noteworthy differences in functional mechanisms, genomic feature distributions, and drug responses. These WGD-related risk genes might provide valuable guidance for therapy selection in breast cancer patients with varying risk levels.

In our study, we also observed a variation in the distribution of tetraploid cells among different risk groups, with a higher frequency of tumors exhibiting this genomic event in the high-risk group. However, this trend was not significant in test cohort, potentially due to the limited data available from TCGA-BRCA dataset. When we expanded the sample size to include the entire TCGA-BRCA cohort, the result became significant.

There are several limitations in our study. First, we observed a variation in the distribution of tetraploid cells among different risk groups, with a higher frequency of tumors exhibiting this genomic event in the high-risk group. However, this trend was not significant in the test set, potentially due to small sample size of the test set. The result was further significant for the entire TCGA cohort including both training and test sets. Secondly, we acknowledge that the limited sample size might have influenced the significance of certain findings, indicating constraints in drawing robust conclusions due to the relatively small number of participants. Our study also unveils substantial variations in patient characteristics, such as mutations and clinical features. These variations may introduce confounding factors and restrict the generalizability of the study's findings. Specifically, we note a higher concentration of patients with mutations in the low-risk group, implying a potential influence on the comparison between high-risk and low-risk groups and its impact on our results. Additionally, our study validated the findings across independent GEO validation cohorts and METABRIC datasets, but further experimental validations are required. The absence of experimental validation raises the possibility of overfitting or bias in the model's performance. In short, our study has certain limitations, including the restricted availability of data, small sample size, variations in patient characteristics, potential bias in risk group comparisons, and the absence of experimental validation.

A future direction of our research involves expanding availability of WGD status data. At present, the TCGA project is the primary source providing the comprehensive data required to determine whether a sample has undergone WGD. The limited availability of WGD status data in other datasets presents a great challenge. Further validation in more external datasets would strengthen the confidence in this observation. This is an aspect we intend to address in forthcoming research endeavors.

5. Conclusion

In conclusion, our study has yielded insights into the impact of WGD on BRCA. By considering the high frequency of WGD events in BRCA, we identified 22 key genes and constructed a risk model that holds clinical significance. Further, we uncovered a potential biomarker gene with implications for diagnostic applications. Finally, we identified diagnostic and drug sensitivity-related gene markers associated with tetraploid properties in BRCA patients.

Funding statement

This work was supported by the Heilongjiang Postdoctoral Research Startup Foundation (LBH-Q21152); the National Nature Fund Program (No.32000473); and the Natural Science Foundation of Heilongjiang Province (LH2021F042 and LH2020C100).

Ethics declarations

Review and/or approval by an ethics committee was not needed for this study because the datasets presented in this study are publicly available data, and did not involve animal and human experiments, as well as other data related to human privacy.

Data availability statement

The publicly available datasets are from TCGA, METABRIC and GEO (accession no. GSE20711, GSE20713, GSE22249, GSE24450, GSE37751, GSE39004, GSE45255, and GSE16228) portals. The download addresses and accession number(s) are provided in the main text. The other data supporting the findings of the current study can be found in the Supplementary Materials.

CRediT authorship contribution statement

Yingli Lv: Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Conceptualization. **Guotao Feng:** Writing – review & editing, Validation, Software, Formal analysis, Data curation. **Lei Yang:** Writing – review & editing, Validation. **Xiaoliang Wu:** Validation. **Chengyi Wang:** Validation. **Aokun Ye:** Validation. **Shuyuan wang:** Validation.

Chaohan Xu: Writing – review & editing, Supervision. **Hongbo Shi:** Writing – review & editing, Validation.

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.

Acknowledgments

Not applicable.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e28586.

References

[1] Z. Storchova, C. Kuffer, The consequences of tetraploidy and aneuploidy, J. Cell Sci. 121 (2008) 3859–3866, https://doi.org/10.1242/jcs.039537.

[2] S. Lens, R. Medema, Cytokinesis defects and cancer, Nat. Rev. Cancer 19 (2019) 32–45, https://doi.org/10.1038/s41568-018-0084-6.

[3] T. Fujiwara, M. Bandi, M. Nitta, E. Ivanova, R. Bronson, D. Pellman, Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells, Nature 437 (2005) 1043–1047, https://doi.org/10.1038/nature04217.

[4] S. Gemble, R. Basto, CHRONOCRISIS: when cell cycle Asynchrony generates DNA damage in polyploid cells, Bioessays : news and reviews in molecular, cellular and developmental biology 42 (2020) e2000105, https://doi.org/10.1002/bies.202000105.

[5] F. Barthel, W. Wei, M. Tang, E. Martinez-Ledesma, X. Hu, S. Amin, K. Akdemir, S. Seth, X. Song, Q. Wang, T. Lichtenberg, J. Hu, J. Zhang, S. Zheng, R. Verhaak, Systematic analysis of telomere length and somatic alterations in 31 cancer types, Nat. Genet. 49 (2017) 349–357, https://doi.org/10.1038/ng.3781.

[6] T. Davoli, T. de Lange, Telomere-driven tetraploidization occurs in human cells undergoing crisis and promotes transformation of mouse cells, Cancer Cell 21 (2012) 765–776, https://doi.org/10.1016/j.ccr.2012.03.044.

[7] S. Dewhurst, N. McGranahan, R. Burrell, A. Rowan, E. Grönroos, D. Endesfelder, T. Joshi, D. Mouradov, P. Gibbs, R. Ward, N. Hawkins, Z. Szallasi, O. Sieber, C. Swanton, Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution, Cancer Discov. 4 (2014) 175–185, https://doi.org/10.1158/2159-8290.Cd-13-0285.

[8] A. Kuznetsova, K. Seget, G. Moeller, M. de Pagter, J. de Roos, M. Dürrbaum, C. Kuffer, S. Müller, G. Zaman, W. Kloosterman, Z. Storchová, Chromosomal instability, tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells, Cell Cycle 14 (2015) 2810–2820, https://doi.org/10.1080/15384101.2015.1068482.

[9] C. Bielski, A. Zehir, A. Penson, M. Donoghue, W. Chatila, J. Armenia, M. Chang, A. Schram, P. Jonsson, C. Bandlamudi, P. Razavi, G. Iyer, M. Robson, Z. Stadler, N. Schultz, J. Baselga, D. Solit, D. Hyman, M. Berger, B. Taylor, Genome doubling shapes the evolution and prognosis of advanced cancers, Nat. Genet. 50 (2018) 1189–1195, https://doi.org/10.1038/s41588-018-0165-1.

[10] M. Imkie, M. Davis, D. Persons, M. Cunningham, Biphasic acute myeloid leukemia with near-tetraploidy and immunophenotypic transformation, Archives of pathology & laboratory medicine 128 (2004) 448–451, https://doi.org/10.5858/2004-128-448-bamlwn.

[11] N. Ganem, S. Godinho, D. Pellman, A mechanism linking extra centrosomes to chromosomal instability, Nature 460 (2009) 278–282, https://doi.org/10.1038/nature08136.

[12] A. Olaharski, R. Sotelo, G. Solorza-Luna, M. Gonsebatt, P. Guzman, A. Mohar, D. Eastmond, Tetraploidy and chromosomal instability are early events during cervical carcinogenesis, Carcinogenesis 27 (2006) 337–343, https://doi.org/10.1093/carcin/bgi218.

[13] T. Watkins, E. Lim, M. Petkovic, S. Elizalde, N. Birkbak, G. Wilson, D. Moore, E. Grönroos, A. Rowan, S. Dewhurst, J. Demeulemeester, S. Dentro, S. Horswell, L. Au, K. Haase, M. Escudero, R. Rosenthal, M. Bakir, H. Xu, K. Litchfield, W. Lu, T. Mourikis, M. Dietzen, L. Spain, G. Cresswell, D. Biswas, P. Lamy, I. Nordentoft, K. Harbst, F. Castro-Giner, L. Yates, F. Caramia, F. Jaulin, C. Vicier, I. Tomlinson, P. Brastianos, R. Cho, B. Bastian, L. Dyrskjøt, G. Jönsson, P. Savas, S. Loi, P. Campbell, F. Andre, N. Luscombe, N. Steeghs, V. Tjan-Heijnen, Z. Szallasi, S. Turajlic, M. Jamal-Hanjani, P. Van Loo, S. Bakhoum, R. Schwarz, N. McGranahan, C. Swanton, Pervasive chromosomal instability and karyotype order in tumour evolution, Nature 587 (2020) 126–132, https://doi.org/10.1038/s41586-020-2698-6.

[14] S. López, E. Lim, S. Horswell, K. Haase, A. Huebner, M. Dietzen, T. Mourikis, T. Watkins, A. Rowan, S. Dewhurst, N. Birkbak, G. Wilson, P. Van Loo, M. Jamal-Hanjani, C. Swanton, N. McGranahan, Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution, Nat. Genet. 52 (2020) 283–293, https://doi.org/10.1038/s41588-020-0584-7.

[15] Y. Zou, F. Ye, Y. Kong, X. Hu, X. Deng, J. Xie, C. Song, X. Ou, S. Wu, L. Wu, Y. Xie, W. Tian, Y. Tang, C. Wong, Z. Chen, X. Xie, H. Tang, The single-cell landscape of intratumoral heterogeneity and the immunosuppressive microenvironment in liver and brain metastases of breast cancer, Adv. Sci. 10 (2023) e2203699, https://doi.org/10.1002/advs.202203699.

[16] Y. Liang, H. Zhang, X. Song, Q. Yang, Metastatic heterogeneity of breast cancer: molecular mechanism and potential therapeutic targets, Semin. Cancer Biol. 60 (2020) 14–27, https://doi.org/10.1016/j.semcancer.2019.08.012.

[17] L. Liao, Y. Zhang, L. Deng, C. Chen, X. Ma, L. Andriani, S. Yang, S. Hu, F. Zhang, Z. Shao, D. Li, Protein phosphatase 1 subunit PPP1R14B stabilizes STMN1 to promote progression and paclitaxel resistance in triple-negative breast cancer, Cancer Res. 83 (2023) 471–484, https://doi.org/10.1158/0008-5472.Can-22-2709.

[18] S. Liu, Z. Ye, V. Xue, Q. Sun, H. Li, D. Lu, KIF2C is a prognostic biomarker associated with immune cell infiltration in breast cancer, BMC Cancer 23 (2023) 307, https://doi.org/10.1186/s12885-023-10788-4.

[19] C. Wu, C. Hsieh, Y. Chang, C. Huang, H. Yeh, M. Hou, Y. Chung, S. Tu, K. Chang, A. Chattopadhyay, L. Lai, T. Lu, Y. Li, M. Tsai, E. Chuang, Differential whole-genome doubling and homologous recombination deficiencies across breast cancer subtypes from the Taiwanese population, Commun. Biol. 4 (2021) 1052, https://doi.org/10.1038/s42003-021-02597-x.

[20] M. Kordi, Z. Borzouyi, S. Chitsaz, M. Asmaei, R. Salami, M. Tabarzad, Antimicrobial peptides with anticancer activity: today status, trends and their computational design, Arch. Biochem. Biophys. 733 (2023) 109484, https://doi.org/10.1016/j.abb.2022.109484.

[21] P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, G. Raghava, AntiCP 2.0: an updated model for predicting anticancer peptides, Briefings Bioinf. 22 (2021) bbaa153, https://doi.org/10.1093/bib/bbaa153.

[22] S. Vijayakumar, L. Ptv, ACPP: a web server for prediction and design of anti-cancer peptides, Int. J. Pept. Res. Therapeut. 21 (2015) 99–106, https://doi.org/10.1007/s10989-014-9435-7.

[23] W. Chen, H. Ding, P. Feng, H. Lin, K. Chou, iACP: a sequence-based tool for identifying anticancer peptides, Oncotarget 7 (2016) 16895–16909, https://doi.org/10.18632/oncotarget.7815.

[24] S. Akbar, M. Hayat, M. Iqbal, M. Jan, iACP-GAEnsC: evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space, Artif. Intell. Med. 79 (2017) 62–70, https://doi.org/10.1016/j.artmed.2017.06.008.

[25] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, W. Shoombuatong, ACPred: a computational tool for the prediction and analysis of anticancer peptides, Molecules 24 (2019) 1973, https://doi.org/10.3390/molecules24101973.

[26] S. Akbar, M. Hayat, M. Tahir, S. Khan, F. Alarfaj, cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model, Artif. Intell. Med. 131 (2022) 102349, https://doi.org/10.1016/j.artmed.2022.102349.

[27] A. Hilchie, C. Doucette, D. Pinto, A. Patrzykat, S. Douglas, D. Hoskin, Pleurocidin-family cationic antimicrobial peptides are cytolytic for breast carcinoma cells and prevent growth of tumor xenografts, Breast Cancer Res. 13 (2011) R102, https://doi.org/10.1186/bcr3043.

[28] R. Quinton, A. DiDomizio, M. Vittoria, K. Kotýnková, C. Ticas, S. Patel, Y. Koga, J. Vakhshoorzadeh, N. Hermance, T. Kuroda, N. Parulekar, A. Taylor, A. Manning, J. Campbell, N. Ganem, Whole-genome doubling confers unique genetic vulnerabilities on tumour cells, Nature 590 (2021) 492–497, https://doi.org/10.1038/s41586-020-03133-3.

[29] T. Huang, L. Fu, The immune landscape of esophageal cancer, Cancer Commun. 39 (2019) 79, https://doi.org/10.1186/s40880-019-0427-z.

[30] A. Shukla, T. Nguyen, S. Moka, J. Ellis, J. Grady, H. Oey, A. Cristino, K. Khanna, D. Kroese, L. Krause, E. Dray, J. Fink, P. Duijf, Chromosome arm aneuploidies shape tumour evolution and drug response, Nat. Commun. 11 (2020) 449, https://doi.org/10.1038/s41467-020-14286-0.

[31] Y. Cohen-Sharir, J. McFarland, M. Abdusamad, C. Marquis, S. Bernhard, M. Kazachkova, H. Tang, M. Ippolito, K. Laue, J. Zerbib, H. Malaby, A. Jones, L. Stautmeister, I. Bockaj, R. Wardenaar, N. Lyons, A. Nagaraja, A. Bass, D. Spierings, F. Foijer, R. Beroukhim, S. Santaguida, T. Golub, J. Stumpff, Z. Storchová, U. Ben-David, Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition, Nature 590 (2021) 486–491, https://doi.org/10.1038/s41586-020-03114-6.

[32] S. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. Laird, R. Onofrio, W. Winckler, B. Weir, R. Beroukhim, D. Pellman, D. Levine, E. Lander, M. Meyerson, G. Getz, Absolute quantification of somatic DNA alterations in human cancer, Nat. Biotechnol. 30 (2012) 413–421, https://doi.org/10.1038/nbt.2203.

[33] Z. Liu, L. Liu, S. Weng, C. Guo, Q. Dang, H. Xu, L. Wang, T. Lu, Y. Zhang, Z. Sun, X. Han, Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer, Nat. Commun. 13 (2022) 816, https://doi.org/10.1038/s41467-022-28421-6.

[34] L. Zhang, C. Liu, X. Zhang, C. Wang, D. Liu, Breast cancer prognosis and immunological characteristics are predicted using the m6A/m5C/m1A/m7G-related long noncoding RNA signature, Funct. Integr. Genom. 23 (2023) 117, https://doi.org/10.1007/s10142-023-01026-y.

[35] T. Zack, S. Schumacher, S. Carter, A. Cherniack, G. Saksena, B. Tabak, M. Lawrence, C. Zhsng, J. Wala, C. Mermel, C. Sougnez, S. Gabriel, B. Hernandez, H. Shen, P. Laird, G. Getz, M. Meyerson, R. Beroukhim, Pan-cancer patterns of somatic copy number alteration, Nat. Genet. 45 (2013) 1134–1140, https://doi.org/10.1038/ng.2760.

[36] S. Gemble, R. Wardenaar, K. Keuper, N. Srivastava, M. Nano, A. Macé, A. Tijhuis, S. Bernhard, D. Spierings, A. Simon, O. Goundiam, H. Hochegger, M. Piel, F. Foijer, Z. Storchová, R. Basto, Author Correction: genetic instability from a single S phase after whole-genome duplication, Nature 608 (2022) E27, https://doi.org/10.1038/s41586-022-05099-w.

[37] A. Taylor, J. Shih, G. Ha, G. Gao, X. Zhang, A. Berger, S. Schumacher, C. Wang, H. Hu, J. Liu, A. Lazar, A. Cherniack, R. Beroukhim, M. Meyerson, Genomic and functional approaches to understanding cancer aneuploidy, Cancer Cell 33 (2018) 676–689.e3, https://doi.org/10.1016/j.ccell.2018.03.007.

[38] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, L. Furlong, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic acids research 45 (2017) D833–D839, https://doi.org/10.1093/nar/gkw943.

[39] Z. Wang, S. Hu, X. Li, Z. Liu, D. Han, Y. Wang, L. Wei, G. Zhang, X. Wang, MiR-16-5p suppresses breast cancer proliferation by targeting ANLN, BMC Cancer 21 (2021) 1188, https://doi.org/10.1186/s12885-021-08914-1.

[40] L. Wang, T. Luan, S. Zhou, J. Lin, Y. Yang, W. Liu, X. Tong, W. Jiang, LncRNA HCP5 promotes triple negative breast cancer progression as a ceRNA to regulate BIRC3 by sponging miR-219a-5p, Cancer Med. 8 (2019) 4389–4403, https://doi.org/10.1002/cam4.2335.

[41] N. Erin, A. Podnos, G. Tanriover, Ö. Duymuş, E. Cote, I. Khatri, R. Gorczynski, Bidirectional effect of CD200 on breast cancer development and metastasis, with ultimate outcome determined by tumor aggressiveness and a cancer-induced inflammatory response, Oncogene 34 (2015) 3860–3870, https://doi.org/10.1038/onc.2014.317.

[42] W. Tan, M. Liu, L. Wang, Y. Guo, C. Wei, S. Zhang, C. Luo, N. Liu, Novel immune-related genes in the tumor microenvironment with prognostic value in breast cancer, BMC Cancer 21 (2021) 126, https://doi.org/10.1186/s12885-021-07837-1.

[43] J. Wildschutte, D. Ram, R. Subramanian, V. Stevens, J. Coffin, The distribution of insertionally polymorphic endogenous retroviruses in breast cancer patients and cancer-free controls, Retrovirology 11 (2014) 62, https://doi.org/10.1186/s12977-014-0062-3.

[44] Z. Li, Y. Li, X. Wang, Q. Yang, PPP2R2B downregulation is associated with immune evasion and predicts poor clinical outcomes in triple-negative breast cancer, Cancer Cell Int. 21 (2021) 13, https://doi.org/10.1186/s12935-020-01707-9.

[45] P. Li, W. Wang, S. Wang, G. Cao, T. Pan, Y. Huang, H. Wan, W. Zhang, Y. Huang, H. Jin, Z. Wang, PTPRC promoted CD8+ T cell mediated tumor immunity and drug sensitivity in breast cancer: based on pan-cancer analysis and artificial intelligence modeling of immunogenic cell death-based drug sensitivity stratification, Front. Immunol. 14 (2023) 1145481, https://doi.org/10.3389/fimmu.2023.1145481.

[46] A. Carnero, Spinophilin: a new tumor suppressor at 17q21, Curr. Mol. Med. 12 (2012) 528–535, https://doi.org/10.2174/156652412800619987.

[47] R. Harris, A. Cheung, J. Ng, R. Laddach, A. Chenoweth, S. Crescioli, M. Fittall, D. Dominguez-Rodriguez, J. Roberts, D. Levi, F. Liu, E. Alberts, J. Quist, A. Santaolalla, S. Pinder, C. Gillett, N. Hammar, S. Irshad, M. Van Hemelrijck, D. Dunn-Walters, F. Fraternali, J. Spicer, K. Lacy, S. Tsoka, A. Grigoriadis, A. Tutt, S. Karagiannis, Tumor-infiltrating B lymphocyte profiling identifies IgG-biased, clonally expanded prognostic phenotypes in triple-negative breast cancer, Cancer Res. 81 (2021) 4290–4304, https://doi.org/10.1158/0008-5472.Can-20-3773.

[48] S. Wang, A. Beeghly-Fadiel, Q. Cai, H. Cai, X. Guo, L. Shi, J. Wu, F. Ye, Q. Qiu, Y. Zheng, W. Zheng, P. Bao, X. Shu, Gene expression in triple-negative breast cancer in relation to survival, Breast Cancer Res. Treat. 171 (2018) 199–207, https://doi.org/10.1007/s10549-018-4816-9.

[49] F. Zhang, Y. Zhang, T. Hou, F. Ren, X. Liu, R. Zhao, X. Zhang, Screening of genes related to breast cancer prognosis based on the DO-UniBIC method, Am. J. Med. Sci. 364 (2022) 333–342, https://doi.org/10.1016/j.amjms.2022.04.022.

[50] Y. Feng, C. Guo, H. Wang, L. Zhao, W. Wang, T. Wang, Y. Feng, K. Yuan, G. Huang, Fibrinogen-like protein 2 (FGL2) is a novel biomarker for clinical prediction of human breast cancer, Med. Sci. Mon. Int. Med. J. Exp. Clin. Res. : international medical journal of experimental and clinical research 26 (2020) e923531, https://doi.org/10.12659/msm.923531.

[51] J. Zhu, Y. Shen, L. Wang, J. Qiao, Y. Zhao, Q. Wang, A novel 12-gene prognostic signature in breast cancer based on the tumor microenvironment, Ann. Transl. Med. 10 (2022) 143, https://doi.org/10.21037/atm-21-6748.

[52] P. Ji, Y. Gong, M. Jin, H. Wu, L. Guo, Y. Pei, W. Chai, Y. Jiang, Y. Liu, X. Ma, G. Di, X. Hu, Z. Shao, Lgals2In vivo multidimensional CRISPR screens identify as an immunotherapy target in triple-negative breast cancer, Sci. Adv. 8 (2022) eabl8247, https://doi.org/10.1126/sciadv.abl8247.

[53] Y. Meng, T. Huang, X. Chen, Y. Lu, A comprehensive analysis of the expression and regulation network of lymphocyte-specific protein tyrosine kinase in breast cancer, Transl. Cancer Res. 10 (2021) 1519–1536, https://doi.org/10.21037/tcr-21-328.

[54] T. Matsuda, K. Oritani, Possible therapeutic applications of targeting STAP proteins in cancer, Biological & pharmaceutical bulletin 44 (2021) 1810–1818, https://doi.org/10.1248/bpb.b21-00672.

[55] J. Tate, S. Bamford, H. Jubb, Z. Sondka, D. Beare, N. Bindal, H. Boutselakis, C. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. Jupe, C. Kok, K. Noble, L. Ponting, C. Ramshaw, C. Rye, H. Speedy, R. Stefancsik, S. Thompson, S. Wang, S. Ward, P. Campbell, S. Forbes, COSMIC: the catalogue of somatic mutations in cancer, Nucleic acids research 47 (2019) D941–D947, https://doi.org/10.1093/nar/gky1015.

[56] C. Lee, K. Fernandez, S. Alexandrou, C. Sergio, N. Deng, S. Rogers, A. Burgess, C. Caldon, Cyclin E2 promotes whole genome doubling in breast cancer, Cancers 12 (2020) 2268, https://doi.org/10.3390/cancers12082268.

[57] C. Zhang, C. Zhu, H. Chen, L. Li, L. Guo, W. Jiang, S. Lu, Kif18A is involved in human breast carcinogenesis, Carcinogenesis 31 (2010) 1676–1684, https://doi.org/10.1093/carcin/bgq134.

[58] W. Zhou, Z. Wang, N. Shen, W. Pi, W. Jiang, J. Huang, Y. Hu, X. Li, L. Sun, Knockdown of ANLN by lentivirus inhibits cell growth and migration in human breast cancer, Mol. Cell. Biochem. 398 (2015) 11–19, https://doi.org/10.1007/s11010-014-2200-6.

[59] K. Magnusson, G. Gremel, L. Rydén, V. Pontén, M. Uhlén, A. Dimberg, K. Jirström, F. Pontén, ANLN is a prognostic biomarker independent of Ki-67 and essential for cell cycle progression in primary breast cancer, BMC Cancer 16 (2016) 904, https://doi.org/10.1186/s12885-016-2923-8.

[60] K. Zhu, Y. Wu, P. He, Y. Fan, X. Zhong, H. Zheng, T. Luo, PI3K/AKT/mTOR-Targeted therapy for breast cancer, Cells 11 (2022) 2508, https://doi.org/10.3390/cells11162508.

[61] Z. Sun, Q. Jiang, B. Gao, X. Zhang, L. Bu, L. Wang, Y. Lin, W. Xie, J. Li, J. Guo, AKT blocks SIK1-mediated repression of STAT3 to promote breast tumorigenesis, Cancer Res. 83 (2023) 1264–1279, https://doi.org/10.1158/0008-5472.Can-22-3407.

[62] Y. Liu, M. Sun, B. Zhang, W. Zhao, KIF18A improves migration and invasion of colorectal cancer (CRC) cells through inhibiting signaling, Aging 15 (2023) 9182–9192, https://doi.org/10.18632/aging.205027.

[63] C. Suzuki, Y. Daigo, N. Ishikawa, T. Kato, S. Hayama, T. Ito, E. Tsuchiya, Y. Nakamura, ANLN plays a critical role in human lung carcinogenesis through the activation of RHOA and by involvement in the phosphoinositide 3-kinase/AKT pathway, Cancer Res. 65 (2005) 11314–11325, https://doi.org/10.1158/0008-5472.Can-05-1507.

[64] A. Belli, D. Cumberland, Percutaneous atherectomy-early experience in Sheffield, Clin. Radiol. 40 (1989) 122–126, https://doi.org/10.1016/s0009-9260(89)80067-4.

[65] G. Haupert, Regulation of Na+, K+-ATPase by the endogenous sodium transport inhibitor from hypothalamus, Hypertension 10 (1987) I61–I66, https://doi.org/10.1161/01.hyp.10.5_pt_2.i61.

[66] A. Klippel, M. Escobedo, M. Wachowicz, G. Apell, T. Brown, M. Giedlin, W. Kavanaugh, L. Williams, Activation of phosphatidylinositol 3-kinase is sufficient for cell cycle entry and promotes cellular changes characteristic of oncogenic transformation, Molecular and cellular biology 18 (1998) 5699–5711, https://doi.org/10.1128/mcb.18.10.5699.

[67] X. Long, W. Zhou, Y. Wang, S. Liu, Prognostic significance of ANLN in lung adenocarcinoma, Oncol. Lett. 16 (2018) 1835–1840, https://doi.org/10.3892/ol.2018.8858.

[68] L. Sheng, Y. Kang, D. Chen, L. Shi, Knockdown of ANLN inhibits the progression of lung adenocarcinoma via pyroptosis activation, Mol. Med. Rep. 28 (2023) 177, https://doi.org/10.3892/mmr.2023.13064.

[69] X. Xu, L. Xu, H. Huang, J. Li, S. Dong, L. Jin, Z. Ma, L. Li, Identification of hub genes as biomarkers correlated with the proliferation and prognosis in lung cancer: a weighted gene Co-expression network analysis, BioMed Res. Int. 2020 (2020) 3416807, https://doi.org/10.1155/2020/3416807.

[70] P. Gao, H. Wang, J. Yu, J. Zhang, Z. Yang, M. Liu, Y. Niu, X. Wei, W. Wang, H. Li, Y. Wang, G. Sun, miR-3607-3p suppresses non-small cell lung cancer (NSCLC) by targeting TGFBR1 and CCNE2, PLoS Genet. 14 (2018) e1007790, https://doi.org/10.1371/journal.pgen.1007790.

[71] Q. Wang, H. Wu, Q. Wu, S. Zhong, Berberine targets KIF20A and CCNE2 to inhibit the progression of nonsmall cell lung cancer via the PI3K/AKT pathway, Drug Dev. Res. 84 (2023) 907–921, https://doi.org/10.1002/ddr.22061.

[72] X. Qian, Y. Li, Y. Yu, F. Yang, R. Deng, J. Ji, L. Jiao, X. Li, R. Wu, W. Chen, G. Feng, X. Zhu, Inhibition of DNA methyltransferase as a novel therapeutic strategy to overcome acquired resistance to dual PI3K/mTOR inhibitors, Oncotarget 6 (2015) 5134–5146, https://doi.org/10.18632/oncotarget.3016.

[73] Y. Bao, L. Wang, L. Shi, F. Yun, X. Liu, Y. Chen, C. Chen, Y. Ren, Y. Jia, Transcriptome profiling revealed multiple genes and ECM-receptor interaction pathways that may be associated with breast cancer, Cellular & molecular biology letters 24 (2019) 38, https://doi.org/10.1186/s11658-019-0162-0.

[74] A. De Luca, A. D'Alessio, M. Maiello, M. Gallo, S. Bevilacqua, D. Frezzetti, A. Morabito, F. Perrone, N. Normanno, Vandetanib as a potential treatment for breast cancer, Expet Opin. Invest. Drugs 23 (2014) 1295–1303, https://doi.org/10.1517/13543784.2014.942034.

[75] Y. Yi, K. You, E. Bae, S. Kwak, Y. Seong, I. Bae, Dual inhibition of EGFR and MET induces synthetic lethality in triple-negative breast cancer cells through downregulation of ribosomal protein S6, Int. J. Oncol. 47 (2015) 122–132, https://doi.org/10.3892/ijo.2015.2982.

[76] L. Zhang, Y. Wei, Y. He, X. Wang, Z. Huang, L. Sun, J. Chen, Q. Zhu, X. Zhou, Clinical implication and immunological landscape analyses of ANLN in pan-cancer: a new target for cancer research, Cancer Med. 12 (2023) 4907–4920, https://doi.org/10.1002/cam4.5177.

[77] Y. Xiao, Z. Deng, Y. Li, B. Wei, X. Chen, Z. Zhao, Y. Xiu, M. Hu, M. Alahdal, Z. Deng, D. Wang, J. Liu, W. Li, ANLN and UBE2T are prognostic biomarkers associated with immune regulation in breast cancer: a bioinformatics analysis, Cancer Cell Int. 22 (2022) 193, https://doi.org/10.1186/s12935-022-02611-0.

[78] X. Zhang, L. Li, S. Huang, W. Liao, J. Li, Z. Huang, Y. Huang, Y. Lian, Comprehensive analysis of ANLN in human tumors: a prognostic biomarker associated with cancer immunity, Oxid. Med. Cell. Longev. 2022 (2022) 5322929, https://doi.org/10.1155/2022/5322929.

[79] T. Davoli, T. de Lange, The causes and consequences of polyploidy in normal development and cancer, Annu. Rev. Cell Dev. Biol. 27 (2011) 585–610, https://doi.org/10.1146/annurev-cellbio-092910-154234.

[80] C. Marquis, C. Fonseca, K. Queen, L. Wood, S. Vandal, H. Malaby, J. Clayton, J. Stumpff, Chromosomally unstable tumor cells specifically require KIF18A for proliferation, Nat. Commun. 12 (2021) 1213, https://doi.org/10.1038/s41467-021-21447-2.

[81] Y. Zheng, K. Wang, N. Li, Q. Zhang, F. Chen, M. Li, Prognostic and immune implications of a novel pyroptosis-related five-gene signature in breast cancer, Frontiers in surgery 9 (2022) 837848, https://doi.org/10.3389/fsurg.2022.837848.

[82] F. Qi, W. Qin, Y. Zang, Molecular mechanism of triple-negative breast cancer-associated BRCA1 and the identification of signaling pathways, Oncol. Lett. 17 (2019) 2905–2914, https://doi.org/10.3892/ol.2019.9884.

[83] J. Kim, H. Jung, I. Sohn, S. Woo, H. Cho, E. Cho, J. Lee, S. Kim, S. Nam, Y. Park, J. Ahn, Y. Im, Prognostication of a 13-immune-related-gene signature in patients with early triple-negative breast cancer, Breast Cancer Res. Treat. 184 (2020) 325–334, https://doi.org/10.1007/s10549-020-05874-1.

[84] S. Wang, Y. Xiong, Q. Zhang, D. Su, C. Yu, Y. Cao, Y. Pan, Q. Lu, Y. Zuo, L. Yang, Clinical significance and immunogenomic landscape analyses of the immune cell signature based prognostic model for patients with breast cancer, Briefings Bioinf. 22 (2021), https://doi.org/10.1093/bib/bbaa311 bbaa311.

[85] Z. Wang, K. Xing, B. Zhang, Y. Zhang, T. Chai, J. Geng, X. Qin, X. Zhang, C. Xu, Identification of prognostic gene signatures by developing a scRNA-seq-based integration approach to predict recurrence and chemotherapy benefit in stage II-III colorectal cancer, Int. J. Mol. Sci. 23 (2022) 12460, https://doi.org/10.3390/ijms232012460.