Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Circulating Tumor DNA Mutation Profiling Reveals Prognostic Signatures in Breast Cancer

Jie LI

lijie@kmmu.edu.cn

Kunming Medical University

kangnan Chen

Lilan Yu

Yan Mou

Bo Yang

Kexin Fan

Yajuan Chen

Jie Xu

Ting Zhang

Dan Yang

Rong Yu

Yining Jin

Lixian Yang

Xingru Chen

Chaohan Xu

Ceshi Chen

Yang Chen

First people's Hospital of Yunnan Province https://orcid.org/0000-0002-3547-5773

Sheng Huang

Fudan University Shanghai Cancer Center;Shanghai Medical College, Fudan University

Article

Keywords: Circulating tumor DNA, breast cancer, whole-genome sequencing, prognostic biomarkers, liquid biopsy

Posted Date: August 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10492982/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** conflict of interest to disclose.

Circulating Tumor DNA Mutation Profiling Reveals Prognostic Signatures in Breast Cancer

Kangnan Chen[†1,2], Lilan Yu[†1], Yan Mou[†1], Bo Yang[†1], Kexin Fan[†1], Yajuan Chen[1], Jie Xu[5], Ting Zhang[5], Dan Yang[1,4], Rong Yu[7] ,Yining Jin[1], Lixian Yang[5], Xingru Chen[5], Chaohan Xu[*2] , Ceshi Chen[*1,5], Yang Chen[*3,5], Sheng Huang[*1,5], Jie Li[*1]

[1]Yunnan Key Laboratory of Breast Cancer Precision Medicine, Academy of Biomedical Engineering, Kunming Medical University, Kunming, 650500, China.

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

[3]Department of Pathology, The First People's Hospital of Yunnan Province, Kunming 650032, Yunnan Province, China.

[4]Ningxia Medical University Laboratory, Department of Oncology Surgery, Tumor Hospital, Affiliated General Hospital of Ningxia Medical University, Ningxia Medical University, Yinchuan, 750004, China.

[5]Yunnan Key Laboratory of Breast Cancer Precision Medicine, Yunnan Cancer Hospital, The Third Affiliated Hospital of Kunming Medical University, Peking University Cancer Hospital Yunnan, Kunming, 650118, China.

[6]Department of Pathology, The Affiliated Hospital of Kunming University of Science and Technology, Kunming 650032, Yunnan, China.

[7]Medical School, Kunming University of Science and Technology, Kunming 650500, Yunnan, China.

† These authors contributed equally.

∗ Corresponding author: Jie Li (lijie@kmmu.edu.cn), Sheng Huang (sammer312@126.com), Yang Chen (cy_koasde@163.com), Ceshi Chen (chenc@kmmu.edu.cn), Chaohan Xu (chaohanxu@hrbmu.edu.cn).

ABSTRACT

Circulating tumor DNA (ctDNA) mutation profiling provides a minimally invasive strategy for identifying prognostic biomarkers in breast cancer. Whole-genome sequencing (WGS) data from 349 patients with breast cancer were analyzed to characterize plasma ctDNA mutation profiles, with 1,124 healthy donors used for background variant filtering. ctDNA profiling identified 10,380 plasma-derived variants involving 5,930 genes across the breast cancer cohort, comprising 8 mutation classes, predominantly represented by missense mutations, frameshift deletions, and nonsense mutations. Recurrently mutated genes included *MUC12*, *HRNR*, *OBSCN*, *TTN*, *EPPK1*, *HMCN2*, *RYR1*, *LRP1*, and *MUC16*. Subtype-associated mutation analysis identified representative subtype-enriched mutated genes, including *TTN* and *OBSCN* in luminal A, *MUC12* and *AMZ1* in luminal B, *ARHGEF17* and *DNHD1* in HER2-enriched disease, and *LARGE2* and *COL7A1* in triple-negative breast cancer.Tissue-level mutation profiles and transcriptomic survival cohorts were used to prioritize and validate ctDNA-derived prognostic biomarkers. A prognostic model integrating *NFKBIA*, *PDCD1*, *ARID1B*, *CFB*, *JAK1*, *MUC4*, and *FCGBP* stratified breast cancer patients into distinct survival-risk groups and was strongly associated with overall survival (HR = 2.31, 95% CI: 1.64–3.26, $P < 0.0001$). Mutation-based classification of these seven genes further revealed distinct survival patterns across genetically defined subgroups ($P = 0.011$). In conclusion, ctDNA mutation profiling demonstrates the utility of plasma-derived genomic features and a seven-gene prognostic score in risk stratification of 1,385 breast cancer patients within 7 datasets, which may support future liquid biopsy-based biomarker discovery and prognostic assessment.

Key words: Circulating tumor DNA; breast cancer; whole-genome sequencing; prognostic biomarkers; liquid biopsy

INTRODUCTION

Breast cancer is the most commonly diagnosed malignancy among women and contributes substantially to cancer-related mortality worldwide. Despite substantial advances in surgery, chemotherapy, endocrine therapy, anti-HER2 therapy, immunotherapy, and targeted therapy, patient outcomes remain highly heterogeneous across disease stages and molecular subtypes[1]. Some patients with apparently favorable clinicopathological features eventually develop recurrence or distant metastasis, whereas others with high-risk features may achieve long-term disease control[2]. Current risk stratification mainly relies on tumor stage, lymph node status, hormone receptor status, HER2 status, Ki-67 index, histological grade, and molecular subtype, but these factors do not fully capture tumor-derived molecular heterogeneity or dynamic disease evolution[3]. ctDNA has shown clinical potential for monitoring tumor burden, treatment response, minimal residual disease, recurrence, and prognosis, and may help guide individualized treatment and surveillance[4,5]. However, its application in breast cancer has largely focused on targeted panels or predefined hotspot mutations, while the value of WGS-based ctDNA mutation profiling beyond predefined targeted panels for prognostic biomarker discovery remains insufficiently explored. Therefore, more precise and minimally invasive approaches are needed to identify individual prognostic risk and to support future treatment escalation, de-escalation, and surveillance strategies in breast cancer[6].

Tissue biopsy remains the standard approach for molecular characterization of breast cancer. However, tissue-based profiling is limited by procedural invasiveness, spatial heterogeneity, sampling bias, and the difficulty of repeated sampling during disease evolution. A single tissue specimen usually represents one anatomical site at one time point and may not fully reflect systemic tumor burden or temporal molecular changes. These limitations are particularly relevant when the goal is to monitor evolving disease biology or to identify molecular features associated with recurrence and survival. Thus, non-invasive approaches that can complement tissue profiling are of increasing clinical interest[7].

ctDNA has emerged as a promising liquid biopsy analyte for non-invasive assessment of breast cancer. Serial ctDNA monitoring can detect molecular relapse before clinical recurrence[12]. Garcia-Murillas et al. reported ctDNA detection before metastatic relapse in 12 of 15 relapsed patients, with a median lead time of 7.9 months[13]. A subsequent personalized surveillance study detected ctDNA before clinical or radiological relapse in 16 of 18 relapsed patients, with 89% sensitivity, a median lead time of 8.9 months, and up to 24 months before

overt recurrence[14].

In residual TNBC after neoadjuvant chemotherapy, postoperative ctDNA positivity was associated with inferior distant disease-free survival (HR = 2.99) and worse disease-free and overall survival[15]. In advanced breast cancer, longitudinal ctDNA allele fraction dynamics in the BEECH trial were used to monitor treatment response, while plasmaMATCH profiling of 800 patients revealed subclonal resistance and pathway alterations involving *ESR1*, *PIK3CA*, *HER2*, and *MAPK*[16]. However, most breast cancer ctDNA studies still rely on targeted panels, hotspot mutations, digital PCR, or personalized mutation tracking[17]. Although these strategies are highly sensitive for selected alterations, they may miss broader plasma-derived mutation patterns, mutation burden, and genomic heterogeneity beyond the assay design. Thus, the prognostic value of broader ctDNA mutation profiling in breast cancer remains insufficiently explored.

In this study, we performed plasma ctDNA mutation profiling in 349 breast cancer patients. A panel of normals constructed from 1,124 healthy donors was incorporated to reduce recurrent technical artifacts and background variants, enabling systematic characterization of mutation burden, variant allele frequency patterns, recurrent alterations, subtype-associated genomic features, and broad mutational spectra. Among these patients, 323 with available pathology information were used to assess the relationship between ctDNA-derived features and clinical characteristics. By incorporating tissue-level mutation profiles, transcriptomic survival information, and independent validation cohorts, we further prioritized ctDNA-derived candidate biomarkers and evaluated their prognostic relevance. This study aimed to determine whether ctDNA mutation profiling can provide a liquid biopsy-based framework for prognostic biomarker discovery and risk stratification in breast cancer.

RESULTS

Patient clinicopathologic characteristics and ctDNA-derived features

A total of 349 female patients with breast cancer underwent plasma ctDNA profiling (Table 1, Fig. 1A). AS ctDNA represents the tumor-derived fraction of total plasma cfDNA, peripheral blood samples were first processed for plasma preparation, cfDNA isolation, and cfDNA sequencing. Tumor-derived mutations were then identified through variant calling and

mutation filtering, thereby generating ctDNA mutation profiles for downstream analysis. Among these patients, 323 with available pathological information were included in subsequent pathological and clinical association analyses.

Among patients with available stage information, most patients had stage II disease (41.8%), followed by stage I (28.3%), stage III (21.1%), and stage IV disease (8.9%) (Fig. 1B). Most annotated patients were younger than 60 years (77.6%), whereas 72 patients were 60 years or older (22.4%) (Fig. 1C). Among patients with molecular subtype information, luminal A was the predominant subtype (70.1%), followed by luminal B (15.5%), triple-negative breast cancer (9.8%), and HER2-enriched breast cancer (4.6%) (Fig. 1D). Receptor profiling showed a predominance of ER-positive (81.3%) and PR-positive tumors (82.0%), while HER2 status was equivocal in 105 patients (44.5%), negative in 94 patients (39.8%), and positive in 37 patients (15.7%) (Fig. 1E–G).

Pathological tumor size and clinical nodal status further supported the clinical heterogeneity of the annotated cohort. T1 and T2 disease were predominant, with 88 and 111 cases, respectively, together accounting for 85.0% of patients with pT annotation. T3 and T4 disease were less common, with 16 and 19 cases, respectively (Fig. S2A). Clinical nodal status included 155 node-negative and 116 node-positive cases, indicating nodal involvement in 42.8% of patients with available nodal information (Fig. S2B).

CfDNA concentration showed broadly overlapping distributions across healthy donors, breast cancer stages, and molecular subtypes (Fig. 1H, Fig. S2E). A modest difference was observed between healthy donors and stage IV disease (P = 0.061; Fig. 1H), while the distributions across other stage comparisons remained largely overlapping. Mutation count showed stage-related variation, with a significant difference between stage I and stage II disease (P = 0.044; Fig. S2C). Mean VAF remained broadly comparable across molecular subtypes (Fig. 1I, Fig. S2D).

Subtype composition across clinical stages showed that luminal A remained the dominant subtype in stage I–III disease, whereas triple-negative breast cancer increased to 36% in stage IV samples (Fig. 1J). These patterns further illustrated the clinical and molecular heterogeneity of the cohort and provided a basis for interpreting ctDNA-derived features in relation to disease stage, receptor status, and molecular subtype[18] (Table 1).

Plasma ctDNA Mutation Landscape Reveals Genomic Heterogeneity in Breast Cancer

To further characterize plasma-derived genomic alterations, ctDNA sequencing was performed in 349 female breast cancer patients, followed by a standardized downstream variant-processing workflow (Fig. S1). This workflow included read mapping, duplicate marking, base recalibration, somatic SNV and indel calling, variant filtering, genotype refinement, and functional annotation. A panel of normals constructed from 1,124 healthy donors was incorporated to reduce recurrent technical artifacts and background variants. After these filtering and annotation steps, analysis-ready variants were generated for mutation landscape analysis. In total, 10,380 variants involving 5,930 genes were detected in plasma ctDNA samples from 323 patients, with a median of 7 variants per sample. The oncoplot of the top altered genes showed a highly heterogeneous mutation pattern across patients, with marked interpatient differences in tumor mutation burden and pathology information. *MUC12*, *HRNR*, *OBSN*, *TTN*, *EPPK1*, *HMCN2*, *RYR1*, *LRP1*, *MUC16* were the most frequently altered genes (Fig. 2A).

Subtype-associated mutation patterns were observed across breast cancer molecular subtypes. Representative recurrent alterations included *TTN*, *OBSCN*, *HRNR*, *TET3*, and *MUC12* in luminal A tumors; *AMZ1*, *DNAH1*, *FBLIM1*, *ELP1* and *C17orf113* in luminal B tumors; *ARHGEF17*, *DNHD1*, *GCN1*, *HDLBP* and *MAPK8IP3* in HER2-enriched tumors, and *LARGE2*, *COL7A1*, *MAP1A*, *AHNAK2*, and *ATP1A4* in triple-negative breast cancer. These genes showed variable mutation frequencies and median TMB levels, supporting subtype-related genomic heterogeneity in plasma ctDNA profiles (Fig. 2B). Missense mutation was the predominant variant class (12,492 variants, 77.8%), followed by frameshift deletion (2,293 variants, 14.3%) and nonsense mutation (955 variants, 5.9%) (Fig. 2C). SNV class analysis showed that C>T and T>C substitutions represented the major substitution categories, and the 96-channel trinucleotide spectrum further illustrated the nucleotide-context distribution of plasma-derived ctDNA alterations (Fig. 2C–D).

ctDNA mutation profiling generated a plasma-based mutation landscape that captured recurrent alterations, mutation burden heterogeneity, subtype-associated genomic patterns, broad mutation classes, and trinucleotide mutational spectra[19]. The recurrently altered genes were enriched for functional categories related to mucin biology, extracellular matrix organization, and structural or cytoskeletal components, extending the ctDNA mutation landscape beyond canonical breast cancer driver alterations[20]. Collectively, ctDNA mutation profiling characterized plasma-derived genomic heterogeneity beyond the scope of fixed hotspot-based assays and provided a basis for downstream biomarker prioritization.

Plasma-detectable candidates linked to immune and stromal remodeling

To prioritize biologically interpretable candidate biomarkers from the plasma ctDNA mutation landscape, ctDNA-detectable altered genes were integrated with TCGA-BRCA tissue-level mutation data and a curated tumor microenvironment gene set[21-24] (Fig. 3A). The schematic illustrates the tumor microenvironmental context, including immune cells and cancer-associated fibroblasts, whereas the Venn diagram summarizes the overlap among the three evidence layers. This integrative screening identified 47 genes shared by ctDNA-detectable alterations, TCGA-BRCA altered genes, and curated TME genes. In addition, 36 genes overlapped between ctDNA-detectable alterations and the curated TME gene set, supporting the presence of plasma-detectable alterations associated with tumor microenvironmental biology (Fig. 3A).

Thirteen recurrently altered candidates, including *MUC12*, *MUC16*, *LRP1*, *MUC5B*, *KMT2D*, *MUC5AC*, *LAMA5*, *ADAMTSL3*, *FCGBP*, *LAMA1*, *MUC17*, *HSPG2*, and *CREBBP*, were visualized in the ctDNA cohort (Fig. 3B). Their mutation frequencies and median TMB values varied across molecular subtypes. In ER-positive–related luminal subtypes, *LRP1*, *MUC5AC*, *KMT2D*, *FCGBP*, and *ADAMTSL3* showed detectable mutation signals in luminal A tumors, whereas *MUC16*, *MUC4*, and *LAMA5* were detected in luminal B tumors. *MUC5B* showed the highest subtype-associated frequency in HER2-enriched tumors, while *MUC16* and *LAMA1* showed relatively higher frequencies in triple-negative breast cancer (Fig. 3C). Functional enrichment analysis linked these ctDNA-prioritized candidates to mononuclear cell differentiation, lymphocyte differentiation, regulation of hematopoiesis, extracellular structure organization, extracellular matrix organization, T cell differentiation, endoderm development, and collagen metabolic processes (Fig. 3D).

Tissue-level mutation analysis validates ctDNA-prioritized candidate genes

Somatic mutation data from a large TCGA-BRCA tissue cohort were used as an orthogonal reference to evaluate whether ctDNA-prioritized candidates reflected broader breast cancer genomic features[25]. To evaluate whether ctDNA-prioritized candidate genes were detectable at the solid tumor tissue level, we examined their mutation status in the TCGA-BRCA tissue cohort. In the overall TCGA-BRCA cohort, somatic alterations were detected in 919 of 986

tissue samples, providing a global tissue-level mutation background for comparison with plasma-derived mutation signals (Fig. S3A). Multiple ctDNA-detectable candidate genes, including *MUC16*, *MUC17*, *ARID1A*, *KMT2D*, *MUC5B*, *LRP1*, and *FCGBP*, were also recurrently altered in breast cancer tissues and were displayed together with pathology information, including gender, pathological T stage, pathological N stage, pathological M status, disease stage, and sample type (Fig. 4A). Their tissue-level mutation classes were broadly consistent with those observed in plasma ctDNA, mainly including missense, frameshift, and nonsense mutations. Notably, the most frequently mutated genes in TCGA-BRCA tissues differed from those identified in plasma ctDNA, suggesting that plasma ctDNA mutation profiling captures a tissue-detectable but partially distinct mutation spectrum.

Trinucleotide-context analysis showed that C>T-related substitution contexts contributed prominently to the overall mutation spectrum (Fig. S3B). Variant classification showed that missense mutations represented the predominant variant type in the tissue cohort, followed by nonsense mutations and frameshift insertions (Fig. S3C). SNV classification further identified C>T as the most frequent single-nucleotide substitution class (Fig. S3D).

The mutation spectrum of these tissue-detectable ctDNA candidate genes was dominated by missense mutations, with additional nonsense mutations, frameshift insertions, and multi-hit events observed in a subset of samples. Together, these findings indicate that several ctDNA-prioritized genes were not restricted to the plasma cohort but were also detectable in breast cancer tissues, supporting their tumor relevance while highlighting the complementary nature of plasma-derived mutation profiling.

Stage-stratified analysis showed gene-specific heterogeneity in mutation frequency and median TMB. Across disease stages, several candidate genes displayed higher median TMB signals in specific clinical strata, with *ARID1A* and *KMT2D* showing prominent TMB-related signals in advanced-stage disease (Fig. 4B). Similar patterns were observed across pathological tumor stages, where *DNMT3A* and *SETD2* showed the strongest median TMB signals in T4 disease, indicating that selected candidate gene alterations were associated with higher mutation burden in invasive tumors (Fig. 4C). Across pathological nodal stages, *FCGBP* showed relatively higher median TMB signals in N0 groups, whereas other candidate genes showed lower or more variable TMB levels (Fig. 4D). Tissue-level mutation analysis supported the relevance of ctDNA-prioritized candidate genes and provided a genomic basis for subsequent prognostic evaluation.

ctDNA-prioritized candidates stratify breast cancer survival risk

To evaluate whether ctDNA-prioritized candidate genes carried prognostic information at the expression level, we analyzed TCGA-BRCA FPKM-normalized bulk RNA-seq profiles from primary tumor samples with matched overall survival annotations. After matching expression profiles with overall survival time and event status, 1,084 primary tumor samples were included in the univariate Cox regression analysis. This analysis identified multiple survival-associated genes among the ctDNA-prioritized candidates, with *ARID1B* and *MUC4* showing risk-associated effects, whereas *NFKBIA*, *PDCD1*, *CFB*, *JAK1*, and *FCGBP* showed protective-associated trends (Fig. 5A). Consistently, single-gene Kaplan-Meier analysis showed that expression-based stratification of several candidate genes was associated with overall survival. High *ARID1B* expression showed a trend toward poorer survival ($P = 0.071$; HR = 1.35, 95% CI: 0.97 – 1.86; Fig. S4A), whereas high expression of *CFB* ($P = 0.025$; HR = 0.69, 95% CI: 0.50 – 0.96; Fig. S4B), *NFKBIA* ($P = 0.0043$; HR = 0.63, 95% CI: 0.45 – 0.87; Fig. S4C), *PDCD1* ($P = 0.019$; HR = 0.68, 95% CI: 0.49 – 0.94; Fig. S4D), and *JAK1* ($P = 0.013$; HR = 0.66, 95% CI: 0.48 – 0.92; Fig. S4E) was significantly associated with improved overall survival.

Candidate genes with $P < 0.1$ in univariate Cox regression were further refined using LASSO Cox regression, resulting in a prognostic model integrating *NFKBIA*, *PDCD1*, *ARID1B*, *CFB*, *JAK1*, *MUC4*, and *FCGBP* (Fig. S5A-B). Based on the Cox coefficient-weighted risk score derived from the selected genes, patients were stratified into high-risk and low-risk groups. Univariate Cox regression analysis demonstrated that high-risk classification was significantly associated with increased overall survival risk. Older age and advanced pathological stage, especially stage III and stage IV, were also associated with worse survival, whereas gender and stage II were not significant prognostic factors (Fig. S6). Multivariable Cox regression incorporating the ctDNA-derived risk group and clinicopathological variables was performed to evaluate whether the risk classification independently predicted patient prognosis. After adjustment for age, gender, and pathological stage, the high-risk group remained significantly associated with poorer overall survival, indicating that the ctDNA-derived risk group provided prognostic information independent of conventional clinicopathological factors (Fig. 5B). Advanced pathological stage also retained prognostic relevance, with stage III and stage IV showing increased survival risk.

Using the Cox coefficient-weighted risk score, patients were divided into high-risk and low-

risk groups by the median cutoff. The high-risk group had significantly worse overall survival than the low-risk group ($P < 0.0001$; HR = 2.31, 95% CI: 1.64–3.26; Fig. 5C), indicating that the expression-based model effectively stratified breast cancer survival risk.

The potential clinical relevance of this risk classification was also explored through treatment-associated survival analysis within the high-risk group. Patients with documented treatment records for paclitaxel, cyclophosphamide, tamoxifen, or anastrozole showed significantly better overall survival than those without corresponding drug records. The corresponding HRs were 2.17 for paclitaxel (95% CI 1.23–3.85, $P = 0.0059$), 2.63 for cyclophosphamide (95% CI 1.69–4.17, $P < 0.0001$), 2.00 for tamoxifen (95% CI 1.14–3.57, $P = 0.014$), and 2.78 for anastrozole (95% CI 1.47–5.56, $P = 0.0012$) (Fig. S7A–D). These findings indicate that treatment-associated survival differences remained detectable among patients classified as high risk by the seven-gene model.

We next examined the mutation-defined prognostic patterns of the seven ctDNA-prioritized model genes. Patients without mutations in these genes were assigned to the wildtype reference group, whereas mutation-positive patients were classified into G1, G2, and G3. G1 included patients with mutations restricted to protective genes, G2 included patients with mutations restricted to risk genes, and G3 included patients with mutations affecting both protective and risk genes. Kaplan–Meier analysis showed significant overall survival differences among these four groups ($P = 0.011$; Fig. 5D). Pairwise comparisons using the wildtype group as the reference showed that G1 was associated with significantly poorer overall survival ($P = 0.009$), whereas G2 showed no significant difference from the wildtype group ($P = 0.817$). G3 showed a trend toward better survival, although the difference did not reach statistical significance ($P = 0.080$). After excluding wildtype patients, survival separation remained evident among G1, G2, and G3 ($P = 0.026$; Fig. S5C). In pairwise comparisons, G1 differed significantly from G3 (HR = 12.13, $P = 0.010$), whereas the comparisons between G2 and G1, and between G2 and G3, showed weaker separation ($P = 0.085$ and $P = 0.087$). We further focused on G1 patients, whose mutations were restricted to protective genes. Within this subset, patients were stratified according to mutation type. Further stratification by mutation type showed significant differences in overall survival across groups ($P = 2 \times 10^{-4}$; Fig. 5E). In particular, frame-shift insertion mutations were associated with increased survival risk compared with the wildtype reference group (HR = 26.19, $P = 0.009$), suggesting that disruptive mutations in favorable-direction genes may attenuate their protective prognostic effect. In comparison, the same mutation-type stratification analysis in G2 patients did not reveal a significant overall survival difference (*P*

= 0.29; Fig. S5D).These results indicate that ctDNA-prioritized candidate genes can be translated into complementary expression-based and mutation-based frameworks for breast cancer survival risk stratification.

External validation confirms the prognostic value of the ctDNA-prioritized risk model

Independent breast cancer expression cohorts with available survival information were used to validate the same prognostic model based on *NFKBIA*, *PDCD1*, *ARID1B*, *CFB*, *JAK1*, *MUC4*, and *FCGBP*. In total, seven GEO breast cancer expression datasets comprising 1,385 patients were included, namely GSE4922 (n = 289), GSE6532 (n = 87), GSE7390 (n = 198), GSE11121 (n = 200), GSE19783 (n = 115), GSE31448 (n = 357), and GSE45255 (n = 139) (Table 2). For each cohort, patients were classified into high-risk and low-risk groups using the cohort-specific median risk score to account for differences in platform, sample size, endpoint definition, and clinical composition.

In GSE4922, 289 breast cancer patients with disease-free survival (DFS) information were evaluated. Patients classified into the high-risk group showed significantly poorer disease-free survival than those in the low-risk group (HR = 1.56, P = 0.042; Fig. 6A), indicating that the seven-gene risk score retained prognostic value in this independent cohort. In GSE7390, which included 198 patients with overall survival (OS) information, the high-risk group also exhibited significantly worse survival outcome, with an HR of 2.21 (P = 0.0037; Fig. 6B). This result supported the ability of the model to stratify overall mortality risk beyond the TCGA-BRCA discovery setting.

The adverse prognostic direction of the risk model was further confirmed in cohorts with endpoints related to distant metastasis and breast cancer-specific survival. In GSE11121, including 200 patients with distant metastasis-free survival (DMFS) information, high-risk patients had a significantly increased risk of distant metastatic events compared with low-risk patients (HR = 1.91, P = 0.033; Fig. 6C). Similarly, in GSE19783, which included 115 patients with breast cancer-specific survival (BCSS) information, the high-risk group showed markedly worse outcome than the low-risk group (HR = 3.53, P = 0.0011; Fig. 6D). Additional validation was performed in GSE31448 and GSE45255. In GSE31448, the largest validation cohort with 357 patients, the high-risk group showed a consistent trend toward poorer DFS compared with the low-risk group (HR = 1.48, P = 0.08; Fig. 6E). Although this association did not reach conventional statistical significance, the direction of effect was

consistent with the other validation cohorts. In GSE45255, which included 139 patients with disease-specific survival (DSS) information, high-risk patients also showed an adverse survival trend (HR = 2.54, *P* = 0.067; Fig. 6F). In GSE6532, which included 87 patients with relapse-free survival (RFS) information, high-risk patients showed significantly poorer survival than low-risk patients (HR = 2.60, *P* = 0.012; Fig. S5E). These findings indicate that the seven-gene model consistently stratified patients toward adverse survival outcomes across independent breast cancer cohorts, including endpoints related to OS, DFS, DMFS, BCSS, DSS and RFS.

Overall, patients classified as high risk consistently exhibited poorer survival outcomes across the external validation cohorts. The concordant adverse-risk direction across multiple endpoints, including disease-free survival, overall survival, distant metastasis-free survival, breast cancer-specific survival, and disease-specific survival, supports the robustness of the ctDNA-prioritized prognostic model in heterogeneous breast cancer populations. These findings suggest that ctDNA-prioritized candidate genes, when integrated with transcriptomic survival modeling, can provide a reproducible framework for identifying patients at increased prognostic risk and may inform future liquid biopsy-guided risk stratification strategies in breast cancer(Table 2).

Discussion

Several studies have demonstrated the clinical value of ctDNA in breast cancer, particularly for treatment monitoring, minimal residual disease detection, recurrence prediction, and therapy-guided molecular profiling[26-28]. However, many previous studies have focused on targeted panels, predefined hotspot mutations, digital PCR assays, or longitudinal tracking of selected variants. The broader value of plasma ctDNA mutation profiling calling from WGS for discovering prognostic biomarkers and linking plasma-derived genomic signals with clinically relevant breast cancer biology remains less fully characterized[29]. In this study, we performed plasma ctDNA mutation profiling in 349 female breast cancer patients and integrated ctDNA-derived alterations with pathology information, with 1,124 healthy donors included for background variant filtering, TCGA-BRCA tissue mutation data, transcriptomic survival information, and independent validation cohorts. This strategy showed that ctDNA profiling can capture heterogeneous plasma-derived genomic alterations, nominate biologically interpretable candidate genes, and support a prognostic model with reproducible

adverse-risk direction across breast cancer cohorts[30].

The clinicopathologic characteristics and ctDNA-derived features of the cohort provided an important context for interpreting plasma mutation profiles in breast cancer. The cohort included patients across different disease stages and molecular subtypes, with luminal A as the predominant subtype. Pre-library cfDNA concentration broadly overlapped across healthy controls, disease stages, and molecular subtypes, whereas mutation count and VAF reflected additional dimensions of ctDNA-derived variation. These observations suggest that cfDNA input quantity alone is insufficient to represent tumor stage or subtype, and that mutation-based ctDNA features should be interpreted together with clinical stage, receptor status, and molecular subtype[31,32].

The plasma mutation landscape further demonstrated that ctDNA profiling can capture broad genomic heterogeneity in breast cancer[33,34]. In 323 analyzable ctDNA samples, we identified 10,380 variants involving 5,930 genes, with a median of 7 variants per sample. The most frequently altered genes included *MUC12*, *HRNR*, *OBSCN*, *TTN*, *EPPK1*, *HMCN2*, *RYR1*, *LRP1*, and *MUC16*. Missense mutation was the predominant variant class, followed by frameshift deletion and nonsense mutation, while C>T and T>C substitutions represented major SNV categories. Together with the 96-channel trinucleotide spectrum, these results provided a broad view of plasma-derived mutation classes and nucleotide-context patterns. Importantly, recurrent alterations involved mucin-related genes, extracellular matrix-related genes, and structural or cytoskeletal genes, which provided the basis for subsequent prioritization of biologically interpretable candidate genes. By integrating ctDNA-detectable alterations with tissue-level mutation data and transcriptomic survival information, we further narrowed these candidates to a prognostic gene set comprising *NFKBIA*, *PDCD1*, *ARID1B*, *CFB*, *JAK1*, *MUC4*, and *FCGBP*. This seven-gene model linked the plasma mutation landscape to survival-risk stratification and supported the overall framework of using ctDNA profiling to nominate candidate biomarkers with potential clinical relevance.

This study also has limitations. The ctDNA cohort was primarily designed for plasma mutation profiling and did not include matched longitudinal plasma samples or long-term survival follow-up for each patient. Therefore, it is still premature to determine whether the identified ctDNA-prioritized markers can directly monitor recurrence or individual survival risk[35]. In addition, the prognostic model was supported by public transcriptomic cohorts with survival annotations rather than survival data from the ctDNA cohort itself. Future prospective studies with serial plasma sampling and clinical follow-up are needed to evaluate the potential of these markers for non-invasive recurrence surveillance and prognostic

monitoring in breast cancer[36,37].

In conclusion, our study demonstrates the utility of plasma ctDNA mutation profiling for capturing tumor-derived genomic heterogeneity and prioritizing prognostically relevant candidate genes in breast cancer. By integrating ctDNA-detectable alterations with tissue-level genomic evidence and transcriptomic survival modeling, we identified a ctDNA-prioritized prognostic model that stratified patients with distinct survival risks and showed consistent adverse-risk direction across independent breast cancer cohorts. Although prospective longitudinal validation is still needed, these findings highlight the potential of ctDNA-informed biomarkers for non-invasive prognostic assessment and provide a framework for future studies exploring recurrence surveillance, survival-risk monitoring, and individualized risk stratification in breast cancer management[38].

Methods

Sample collection

Breast cancer plasma samples used in this study were collected from patients recruited from multiple hospitals in China, including Peking University Cancer Hospital, Yunnan Cancer Hospital, the First People Hospital Of Yunnan Province, Beijing Tsinghua Changgung Hospital, the Cancer Hospital of the Chinese Academy of Medical Sciences. Peripheral blood samples were collected from breast cancer patients for plasma cfDNA analysis and downstream ctDNA mutation profiling[39]. All samples and associated clinical information were de-identified before analysis.This study was reviewed and approved by the Medical Ethics Committee of Kunming Medical University (approval number: KMMU2025MEC328). In total, 349 breast cancer plasma samples were included for ctDNA mutation profiling. Among them, 323 samples with available pathology information were retained for subsequent clinicopathologic and molecular association analyses. 1,124 healthy donor samples were used as a background reference for variant filtering.

Sample preparation and processing

cfDNA Extraction. Plasma was isolated from peripheral blood using a two-step centrifugation protocol. Briefly, peripheral blood samples were first centrifuged at 2,000 rpm

for 10 min to separate plasma from cellular components. The plasma supernatant was carefully transferred into EP tubes and subjected to a second centrifugation at 14,000 rpm for 10 min at 4°C to remove residual cells and debris. The clarified plasma supernatant was transferred into fresh EP tubes and stored at −80°C until extraction. Cell-free DNA (cfDNA) was extracted from stored plasma samples according to the manufacturer's protocol. The concentration and quality of cfDNA were assessed before downstream library construction.

Library construction and sequencing. cfDNA libraries were constructed using a cfDNA library preparation workflow according to the manufacturer's protocol. Because cfDNA is naturally fragmented, no additional fragmentation step was performed. Briefly, purified cfDNA was subjected to end preparation in a 60 μL reaction containing 7 μL End Prep Buffer 2, 3 μL End Prep Enzyme 2, and 50 μL cfDNA template with ddH2O. The reaction was performed at 20°C for 30 min and 65°C for 30 min, followed by holding at 4°C. Adapter ligation was then performed in a 110 μL reaction containing 60 μL end-prepared product, 30 μL Rapid Ligation Buffer 5, 10 μL Rapid DNA Ligase 2, 5 μL adapter, and 5 μL ddH2O, and incubated at 20°C for 15 min. Adapter-ligated products were purified using 88 μL VAHTS DNA Clean Beads, washed twice with 80% ethanol, and eluted with 22.5 μL 10 mM Tris-HCl.

Library amplification was performed in a 50 μL PCR reaction containing 20 μL purified adapter-ligated product, 5 μL PCR primers, and 25 μL VAHTS HiFi Amplification Mix 2. PCR was performed as follows: 98°C for 45 s; 15 cycles of 98°C for 15 s, 60°C for 30 s, and 72°C for 30 s; 72°C for 1 min; and hold at 4°C. Amplified libraries were purified using 50 μL VAHTS DNA Clean Beads, washed twice with 80% ethanol, and eluted with 22.5 μL ddH2O. Library concentration and fragment size distribution were assessed before sequencing. Qualified libraries were pooled and sequenced on a high-throughput next-generation sequencing platform with paired-end reads.

Sequencing data analysis and variant calling

Preprocessing and alignment. Raw paired-end sequencing reads were trimmed using Trim Galore to remove terminal adapters and low-quality sequences with a quality cutoff of 20, Phred33 encoding, adapter stringency of 3, and a minimum read length of 20 bp. Clean reads were aligned to the human reference genome GRCh38 using the BWA-MEM algorithm

implemented in Burrows-Wheeler Aligner (BWA, version 0.7.19). Read group information was added during alignment. The resulting alignments were converted to BAM format, sorted, and indexed using SAMtools (version 1.9). PCR duplicates were marked and removed using Picard MarkDuplicates. Base-quality score recalibration was then performed before downstream variant calling. Reads with mapping quality ≥30 were retained for subsequent analyses.

ctDNA variant calling. Somatic single-nucleotide variants (SNVs) and small insertions and deletions (InDels) were detected from plasma cfDNA sequencing data using GATK Mutect2 from GATK(version 4.2.5.0). To reduce recurrent technical artifacts and background signals, an unmatched panel of normals constructed from 1,124 healthy donors was incorporated into the variant-calling workflow. A germline resource and population common-variant database were also used to reduce likely germline variants and common polymorphisms. The allele-frequency parameter for variants absent from the germline resource was set to 0.00003125.

Contamination estimation and variant filtering. For each cfDNA sample, pileup summaries were generated at common germline variant sites using GATK GetPileupSummaries. Sample-level contamination was estimated using GATK CalculateContamination. Raw Mutect2 calls were then filtered using GATK FilterMutectCalls by incorporating contamination estimates and standard Mutect2 filtering criteria. Candidate variants were further evaluated using supporting read evidence to reduce PCR- and sequencing-derived errors. Variants were retained when they passed Mutect2 filtering, were supported by consensus read evidence, were not recurrently observed in the unmatched panel of normals, and were not annotated as common germline polymorphisms. Variants passing these filters were considered putative ctDNA-derived somatic variants.

Variant annotation and downstream analysis. Filtered VCF files were converted into ANNOVAR-compatible input format and annotated using ANNOVAR with the hg38 genome build. Annotation protocols included refGene, cytoBand, exac03, and dbnsfp30a. Per-sample annotated variant tables were aggregated into a unified mutation table. After sample-level and variant-level quality control, analysis-ready variants were retained for downstream mutation profiling, including mutation count, variant allele frequency, variant classification, recurrently altered gene identification, mutation spectrum analysis, and integration with pathology

information.

Definition of mutation-defined patient groups

Candidate genes were first evaluated using univariate Cox proportional hazards regression analysis based on gene expression and overall survival data. Hazard ratios (HRs) and 95% confidence intervals were calculated for each gene. The prognostic direction of each gene was defined according to its HR value: genes with HR > 1 were considered risk-associated genes, indicating that higher expression was associated with increased mortality risk, whereas genes with HR < 1 were considered protective-associated genes, indicating that higher expression was associated with reduced mortality risk. Candidate genes with potential prognostic relevance were then incorporated into a LASSO-penalized Cox model for feature selection and risk-score construction.

For mutation-based stratification, patients were grouped according to the mutation status of risk-associated and protective-associated genes among the genes retained in the final LASSO-penalized Cox model. Patients carrying mutations only in protective-associated genes were assigned to G1, patients carrying mutations only in risk-associated genes were assigned to G2, and patients carrying mutations in both risk-associated and protective-associated genes were assigned to G3. Patients without mutations in any of the retained model genes were classified as the wildtype group.

Survival analysis and prognostic model construction

To evaluate the prognostic relevance of ctDNA-detectable candidate genes in breast cancer, survival information and corresponding gene expression profiles were integrated for Cox regression-based survival modeling. Candidate genes identified from the ctDNA mutation analysis were first subjected to univariate Cox proportional hazards regression analysis. Genes with HR > 1 were considered potential risk-associated genes, whereas genes with HR < 1 were considered potential protective genes. Genes showing survival-associated signals were further evaluated using least absolute shrinkage and selection operator Cox regression to construct a parsimonious prognostic model.

The final LASSO-penalized Cox model retained seven genes, including *ARID1B*, *JAK1*,

MUC4, *CFB*, *PDCD1*, *NFKBIA*, and *FCGBP*. The model coefficients were fixed and subsequently applied to independent GEO validation cohorts. For each GEO validation cohort, expression values of the seven retained genes were standardized using z-score transformation within the corresponding dataset. A LASSO Cox-derived risk score was then calculated for each patient according to the following formula:

$$Risk\ score = \sum_{i=1}^{n} coefficient_i \times standardized_expression_i$$

where the coefficients were defined as follows: *ARID1B* = 1.16870213, *JAK1* = −0.31956337, *MUC4* = 0.31647725, *CFB* = −0.24106792, *PDCD1* = −0.22657677, *NFKBIA* = −0.22434813, and *FCGBP* = −0.07262415. Patients were classified into high-risk and low-risk groups using the median risk score within each GEO cohort.

Kaplan–Meier survival analysis and log-rank tests were used to compare survival differences between the high- and low-risk groups. Hazard ratios and 95% confidence intervals were calculated using Cox proportional hazards regression. The reproducibility of the prognostic model was assessed by evaluating whether the high-risk group consistently showed poorer survival outcomes across independent GEO validation cohorts.

Statistical analysis

Categorical variables were summarized as counts and percentages, and continuous variables were summarized as medians with interquartile ranges. The Kruskal–Wallis test was used to compare differences in continuous or ordinal variables across groups. Spearman's rank correlation analysis was used to evaluate correlations between variables.

Survival time was defined as the time from diagnosis to death, the occurrence of the recorded endpoint event, or last follow-up. Survival analyses were performed using the Kaplan–Meier method, and survival differences between groups were assessed using the log-rank test. Hazard ratios and 95% confidence intervals were estimated using Cox proportional hazards regression. Prognostic model construction was performed using Cox regression and LASSO Cox regression where appropriate.

All analyses and visualizations were performed using R software version 4.5.1. The main R packages included survival (v3.8.3), survminer (v0.5.2), ggplot2 (v4.0.0), pheatmap (v1.0.13), pROC, glmnet (v5.0), maftools (v2.24.0), ComplexHeatmap (v2.24.1), and VennDiagram (v1.7.3). All statistical tests were two-sided, and $P < 0.05$ was considered statistically significant.

Data availability

All human plasma cfDNA sequencing data generated in this study have been deposited in the Genome Sequence Archive for Human (GSA-Human: HRA018400) , Chinese Academy of Sciences, and are available under controlled access [https://ngdc.cncb.ac.cn/gsa-human/s/rS8L22sM]. The cfDNA-derived VCF data generated in this study have been deposited in the Genome Variation Map (GVM: GVM001462) at NGDC and are available at https://bigd.big.ac.cn/gvm/getProjectDetail?Project=GVM001462. The human plasma cfDNA sequencing data generated from healthy donors and used as normal control datasets in this study have been deposited in the Genome Sequence Archive for Human (GSA-Human) under accession numbers HRA019452 and HRA019254. The HRA019452 dataset is available under controlled access at https://ngdc.cncb.ac.cn/gsa-human/s/26Qy2Wp4, while the HRA019254 dataset, which was also used as an additional healthy-donor reference, is available under controlled access at https://ngdc.cncb.ac.cn/gsa-human/s/qg08o94P. The independent validation datasets GSE4922, GSE6532, GSE7390, GSE11121, GSE19783, GSE31448, and GSE45255 were obtained from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/). Access to individual-level human genomic data can be requested through the corresponding Data Access Committee and may be granted for academic research use upon approval. To protect the privacy and confidentiality of our patients, identifiable clinical data are not made publicly available in a repository or in the supplementary material of the article. The remaining data supporting the findings of this study are available within the article and its supplementary materials.

Code availability

The code used for the analysis is available from the GitHub repository at https://github.com/Leelab-Kmmu/ctDNA-Mutation-Profiling-Reveals-Prognostic-Biomarkers-in-Breast-Cancer. The code used for cfDNA variant calling analysis is available

at https://github.com/Leelab-Kmmu/cfDNA_Mutect2_generator.

Ethics approval

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and was approved by the Medical Ethics Committee of Kunming Medical University (Approval No. KMMU2025MEC328). Written informed consent was obtained from all participants prior to sample collection. All samples and associated data were de-identified before sequencing and analysis.

Funding Declaration

This work was supported in part by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (No. 2025ZD0545500, 2025ZD0545502), the National Natural Science Foundation of China (No. 82404091, 82560572), the Yunnan Provincial Major Science and Technology Special Plan Projects (202501CF070023), the Health Commission Foundation of Yunnan Province (2023-KHRCBZ-B15), Yunnan Province Applied Basic Research Program Kunming Medical University Joint Project (202401AY070001-120).

Competing interests

The authors have no competing interests to declare.

Contributions

J.L., S.H., Yan.C., C.C., and C.X. conceived and designed the study and supervised the project. L.Yu, Y.M., J.X., T.Z., L.Yang, and X.C. collected and curated the data. K.C., B.Y., K.F., Yaj.C., D.Y., R.Y., and Y.J. performed the experiments and analyzed the corresponding results. K.C., L.Yu, Y.M., B.Y., K.F., Y.J., and X.C. performed the bioinformatics and statistical analyses. K.C., L.Yu, Y.M., Yaj.C., D.Y., R.Y., L.Yang, and X.C. contributed to data validation and visualization. J.L., S.H., Yan.C., C.C., and C.X. provided experimental resources, technical guidance, and project supervision. K.C., L.Yu, Y.M., B.Y., K.F., and J.L. drafted the initial manuscript. All authors reviewed, revised, and approved the final

manuscript.

Reference

1 Biomarkers for Systemic Therapy in Metastatic Breast Cancer: ASCO Guideline Update. *Journal of Clinical Oncology* (2022). https://doi.org/10.1200/JCO.22.01063

2 Identification of a five-lncRNA signature for predicting the risk of tumor recurrence in patients with breast cancer. *International Journal of Cancer* (2018). https://doi.org/10.1002/ijc.31573

3 Systematical Identification of Breast Cancer-Related Circular RNA Modules for Deciphering circRNA Functions Based on the Non-Negative Matrix Factorization Algorithm. *International Journal of Molecular Sciences* (2019). https://doi.org/10.3390/ijms20040919

4 Circulating tumor DNA in neoadjuvant-treated breast cancer reflects response and survival. *Annals of Oncology* (2021). https://doi.org/10.1016/j.annonc.2020.11.007

5 Results of the c-TRAK TN trial: a clinical trial utilising ctDNA mutation tracking to detect molecular residual disease and trigger intervention in patients with moderate- and high-risk early-stage triple-negative breast cancer. *Annals of Oncology* (2023). https://doi.org/10.1016/j.annonc.2022.11.005

6 Perspective on new cell-free DNA technologies for early cancer detection. *Cancer Biology & Medicine* (2023). https://doi.org/10.20892/j.issn.2095-3941.2023.0159

7 Liquid Biopsy: From Discovery to Clinical Application. *Cancer Discovery* (2021). https://doi.org/10.1158/2159-8290.CD-20-1311

8 Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer* (2017). https://doi.org/10.1038/nrc.2017.7

9 Novel technologies in cfDNA analysis and potential utility in clinic. *Chinese Journal of Cancer Research* (2021). https://doi.org/10.21147/j.issn.1000-9604.2021.06.07

10 Use of ctDNA in early breast cancer: analytical validity and clinical potential. *NPJ Breast Cancer* (2024). https://doi.org/10.1038/s41523-024-00653-3

11 Analysis of circulating tumor DNA to monitor metastatic breast cancer. *New England Journal of Medicine* (2013). https://doi.org/10.1056/NEJMoa1213261

12 Assessment of Molecular Relapse Detection in Early-Stage Breast Cancer. *JAMA Oncology* (2019). https://doi.org/10.1001/jamaoncol.2019.1838

13 Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Science Translational Medicine* (2015). https://doi.org/10.1126/scitranslmed.aab0021

14 Personalized Detection of Circulating Tumor DNA Antedates Breast Cancer Metastatic Recurrence. *Clinical Cancer Research* (2019). https://doi.org/10.1158/1078-0432.CCR-18-3663

15 Clinical significance and biology of circulating tumor DNA in high-risk early-stage HER2-negative breast cancer receiving neoadjuvant chemotherapy. *Cancer Cell* (2023). https://doi.org/10.1016/j.ccell.2023.04.008

16 Circulating tumour DNA analysis to direct therapy in advanced breast cancer (plasmaMATCH): a multicentre, multicohort, phase 2a, platform trial. *The Lancet Oncology* (2020). https://doi.org/10.1016/S1470-2045(20)30444-7

17 Whole genome sequencing-powered ctDNA sequencing for breast cancer detection. *Annals of Oncology* (2025). https://doi.org/10.1016/j.annonc.2025.01.021

18 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* (2012). https://doi.org/10.1038/nature10983
19 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* (2016). https://doi.org/10.1038/nature17676
20 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020). https://doi.org/10.1038/s41586-020-1943-3
21 Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discovery* **12**, 31–46 (2022). https://doi.org/10.1158/2159-8290.CD-21-1059
22 Sahai, E., Astsaturov, I. & Cukierman, E. A framework for advancing our understanding of cancer-associated fibroblasts. *Nature Reviews Cancer* **20**, 174–186 (2020). https://doi.org/10.1038/s41568-019-0238-1
23 Nagarsheth, N., Wicha, M. S. & Zou, W. Chemokines in the cancer microenvironment and their relevance in cancer immunotherapy. *Nature Reviews Immunology* **17**, 559–572 (2017). https://doi.org/10.1038/nri.2017.49
24 Roumenina, L. T., Daugan, M. V. & Petitprez, F. Context-dependent roles of complement in cancer. *Nature Reviews Cancer* **19**, 698–715 (2019). https://doi.org/10.1038/s41568-019-0210-0
25 The Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* (2012). https://doi.org/10.1038/nature11412
26 Dynamic ctDNA tracking stratifies relapse risk for triple negative breast cancer patients receiving neoadjuvant chemotherapy. *Nature Communications* (2025). https://doi.org/10.1038/s41467-025-57988-z
27 Molecular Residual Disease in Breast Cancer: Detection and Therapeutic Interception. *Clinical Cancer Research* (2023). https://doi.org/10.1158/1078-0432.CCR-23-0757
28 Circulating tumor DNA for predicting recurrence in patients with operable breast cancer: a systematic review and meta-analysis. *ESMO Open* (2024). https://doi.org/10.1016/j.esmoop.2024.102390
29 Advances in liquid biopsy: From exploration to practical application. *Cancer Cell* (2025). https://doi.org/10.1016/j.ccell.2024.11.009
30 ctDNA detectability and representativeness in seven body liquids from patients with metastatic breast cancer. *Nature Communications* (2025). https://doi.org/10.1038/s41467-025-65838-1
31 Cell-free DNA-associated multi-feature applications in cancer diagnosis and treatment. *Clinical and Translational Discovery* (2024). https://doi.org/10.1002/ctd2.280
32 Detection of Circulating Tumor DNA Using a Tissue-Free Epigenomic Assay Is a Highly Prognostic Biomarker in Early-Stage Triple-Negative Breast Cancer. *Clinical Cancer Research* (2025). https://doi.org/10.1158/1078-0432.CCR-24-3145
33 Uncovering chromatin accessibility and cancer diagnostic potential via cell-free DNA utilization. *Science Bulletin* (2024). https://doi.org/10.1016/j.scib.2024.04.013
34 Genomic landscape of circulating tumor DNA in HER2-low metastatic breast cancer. *Signal Transduction and Targeted Therapy* (2024). https://doi.org/10.1038/s41392-024-02047-0
35 Detecting Liquid Remnants of Solid Tumors: Circulating Tumor DNA Minimal Residual Disease. *Cancer Discovery* (2021). https://doi.org/10.1158/2159-8290.CD-21-0634
36 Ultrasensitive Detection and Monitoring of Circulating Tumor DNA Using Structural Variants in Early-Stage Breast Cancer. *Clinical Cancer Research* (2025). https://doi.org/10.1158/1078-0432.CCR-24-3472

37 Serial Postoperative Circulating Tumor DNA Assessment Has Strong Prognostic Value During Long-Term Follow-Up in Patients With Breast Cancer. *JCO Precision Oncology* (2024). https://doi.org/10.1200/PO.23.00456
38 Liquid biopsy enters the clinic - implementation issues and future challenges. *Nature Reviews Clinical Oncology* (2021). https://doi.org/10.1038/s41571-020-00457-x
39 Integrated analysis toolkit for dissecting whole-genome-wide features of cell-free DNA. *Clinical and Translational Medicine* (2023). https://doi.org/10.1002/ctm2.1212

Figure 1

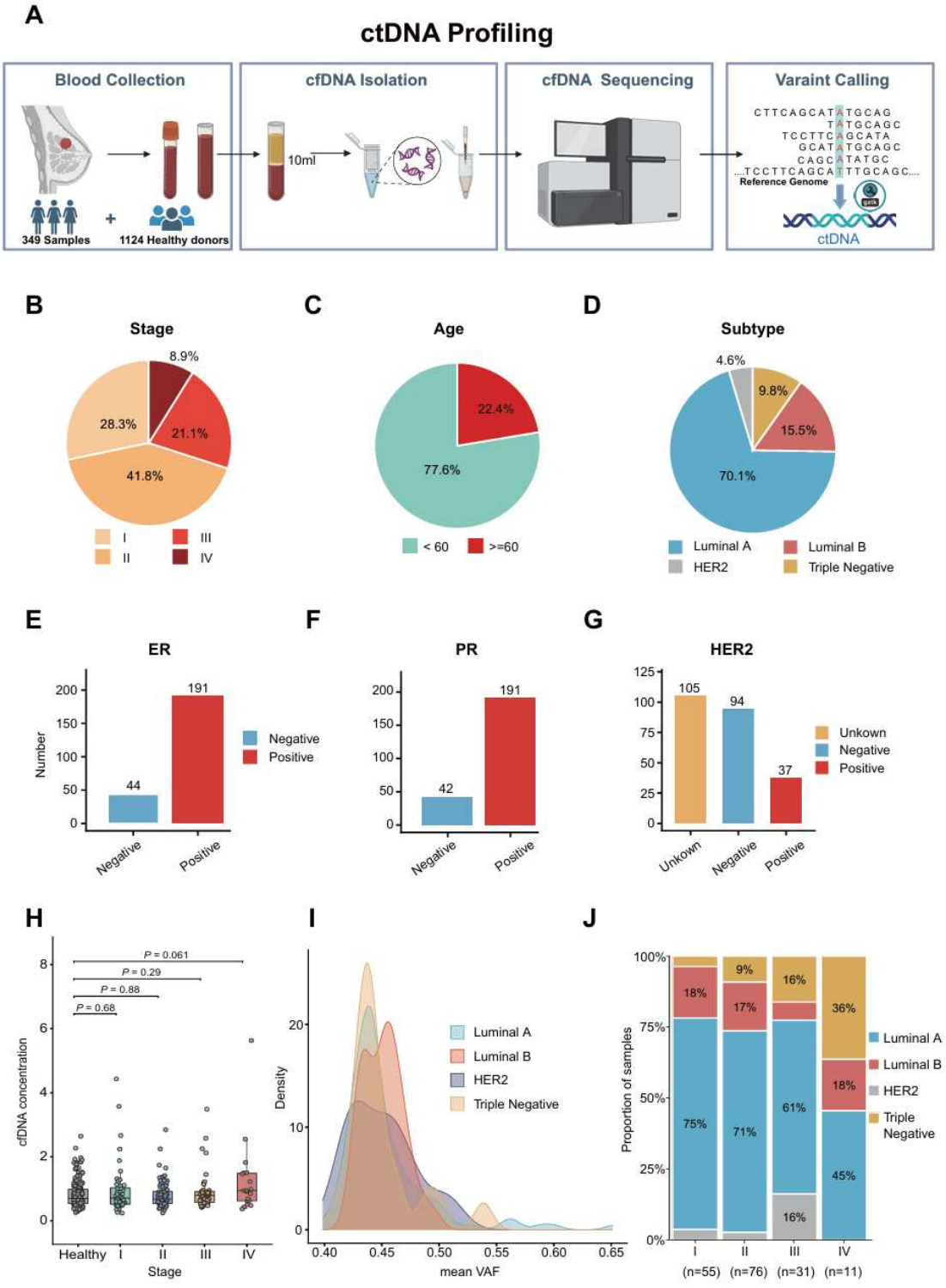

A
ctDNA Profiling
Blood Collection
cfDNA Isolation
cfDNA Sequencing
Varaint Calling
10ml
349 Samples
1124 Healthy donors
Reference Genome
ctDNA
B
Stage
8.9%
21.1%
41.8%
28.3%
I
II
III
IV
C
Age
22.4%
77.6%
< 60
>=60
D
Subtype
4.6%
9.8%
15.5%
70.1%
Luminal A
Luminal B
HER2
Triple Negative
E
ER
Number
191
44
Negative
Positive
F
PR
191
42
Negative
Positive
G
HER2
105
94
37
Unkown
Negative
Positive
H
P = 0.061
P = 0.29
P = 0.88
P = 0.68
cfDNA concentration
Healthy
I
II
III
IV
Stage
I
Density
Luminal A
Luminal B
HER2
Triple Negative
mean VAF
J
Proportion of samples
18%
75%
9%
17%
71%
16%
61%
16%
36%
18%
45%
(n=55)
(n=76)
(n=31)
(n=11)

Fig. 1 Study overview and baseline clinicopathologic and ctDNA-related characteristics of the breast cancer cohort.

(A) Workflow of WGS-based plasma ctDNA mutation profiling in breast cancer. Peripheral blood samples from 349 breast cancer patients and 1,124 healthy donors were processed for plasma separation, cfDNA isolation, whole-genome sequencing, and variant calling. Plasma-derived tumor-associated variants were used to generate ctDNA mutation profiles for downstream analyses. **(B)** Distribution of clinical stage among patients with available stage information. Stage II disease was the most common subgroup (41.8%), followed by stage I (28.3%), stage III (21.1%), and stage IV disease (8.9%). **(C)** Age distribution of the annotated cohort, showing that most patients were younger than 60 years (<60 years, 77.6%; ≥60 years, 22.4%). **(D)** Distribution of molecular subtypes, with luminal A representing the predominant subtype (70.1%), followed by luminal B (15.5%), triple-negative breast cancer (9.8%), and HER2-enriched tumors (4.6%). **(E-G)** Bar plots showing receptor status in the annotated cohort. ER-positive and PR-positive tumors were predominant, with 191 ER-positive and 191 PR-positive cases, whereas HER2 status included 105 equivocal, 94 negative, and 37 positive cases. **(H)** Box plot comparing pre-library cfDNA concentration between healthy controls and breast cancer patients stratified by clinical stage. Selected pairwise P values are shown, including the comparison between healthy controls and stage IV disease. **(I)** Density distribution of mean variant allele frequency (VAF) across molecular subtypes. **(J)** Stacked bar plot showing molecular subtype composition across clinical stages. Luminal A remained the dominant subtype in stage I–III disease, whereas triple-negative breast cancer accounted for a larger proportion of stage IV samples. Sample numbers for each stage group are shown below the bars.

Figure 2

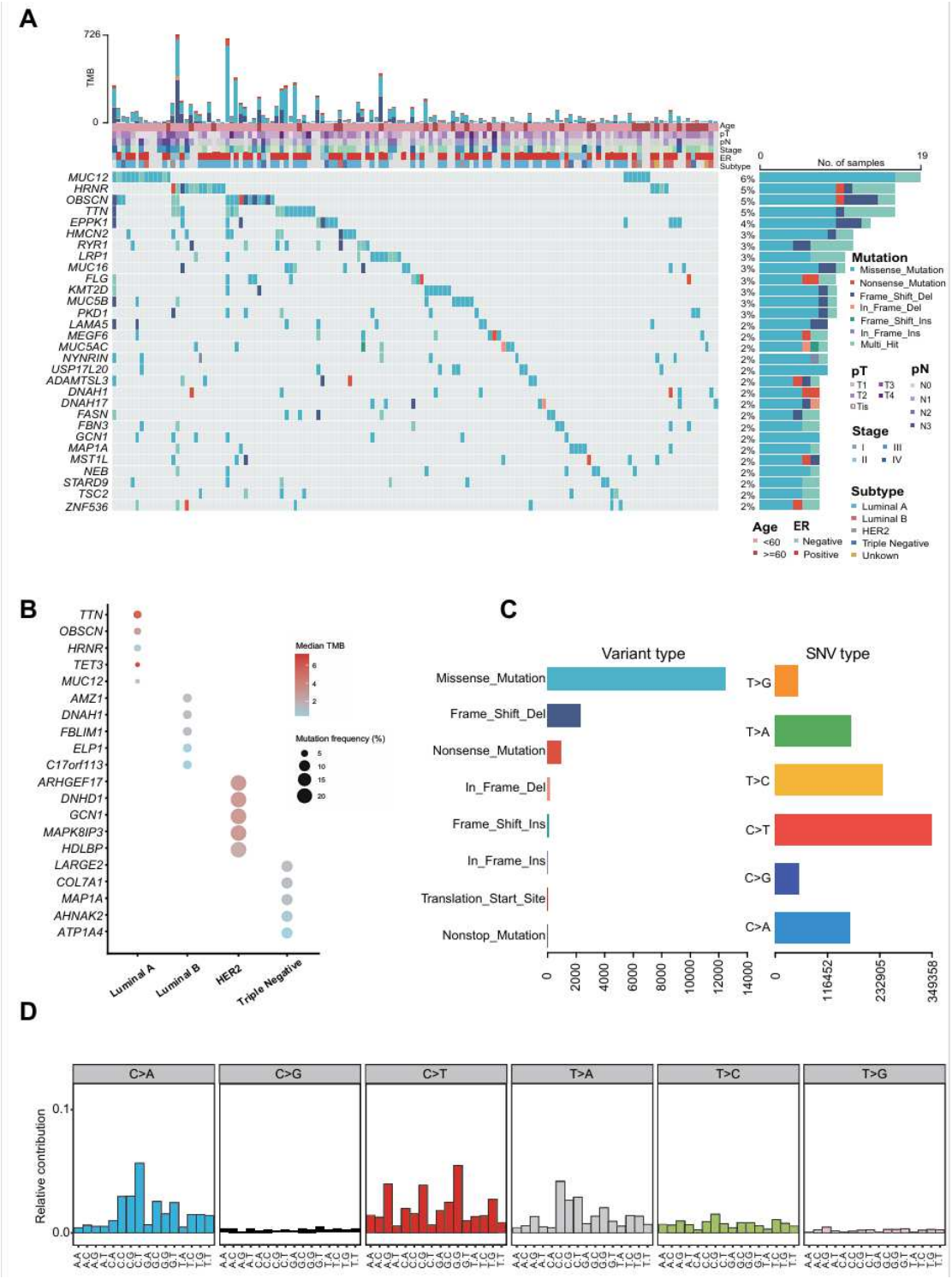


Fig. 2 Plasma ctDNA mutation landscape and mutational spectrum in breast cancer.
(A) Oncoplot showing recurrent ctDNA alterations and clinicopathological annotations across

breast cancer samples. The upper bar plot shows sample-level TMB, and the right bar plot shows the number of mutated samples for each gene. Mutation colors indicate variant classifications, including missense mutation, nonsense mutation, frameshift deletion, in-frame deletion, frameshift insertion, in-frame insertion, and multi-hit events. **(B)** Bubble plot showing subtype-associated altered genes across luminal A, luminal B, HER2-enriched tumors, and triple-negative breast cancer. Dot size indicates mutation frequency, and color indicates median TMB. **(C)** Bar plots showing the distribution of variant classifications and SNV classes. Missense mutation was the predominant variant class, and C>T/T>C substitutions were the major SNV classes. **(D)** Ninety-six-channel trinucleotide mutational spectrum showing the relative contribution of each nucleotide substitution context in breast cancer ctDNA samples.

Figure 3

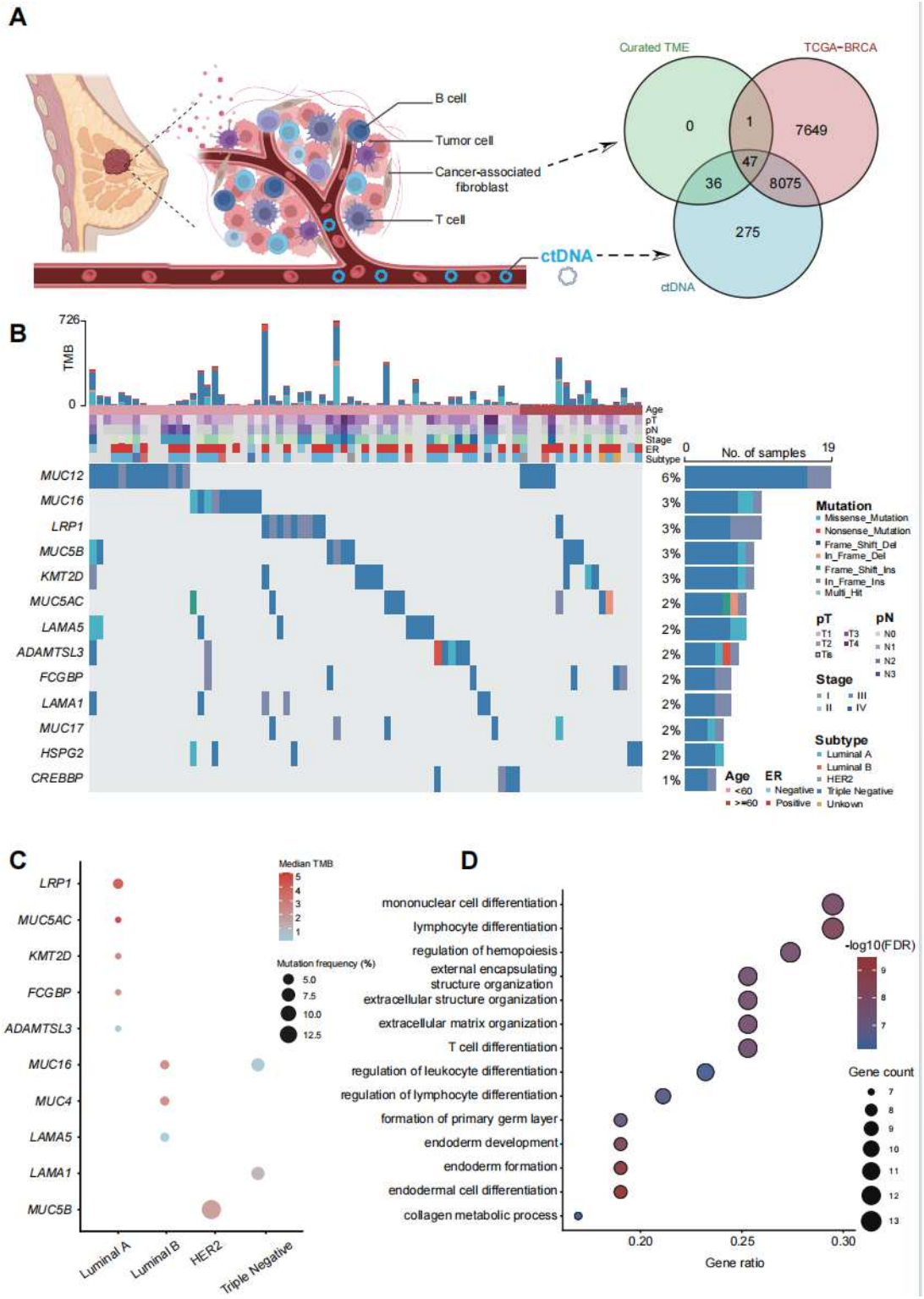


Fig. 3 ctDNA-detectable tumor microenvironment-related candidate genes in breast cancer.

(A) Schematic and Venn diagram showing the integration of ctDNA-detectable altered genes, TCGA-BRCA tissue-level altered genes, and curated tumor microenvironment-related genes. **(B)** Oncoplot showing recurrent alterations of 13 ctDNA-prioritized TME-related candidate genes and clinicopathological annotations across breast cancer samples. **(C)** Bubble plot showing the mutation frequency and median TMB of candidate genes across molecular subtypes. Dot size indicates mutation frequency, and color indicates median TMB. **(D)** Functional enrichment analysis of ctDNA-prioritized TME-related candidates.

Figure 4

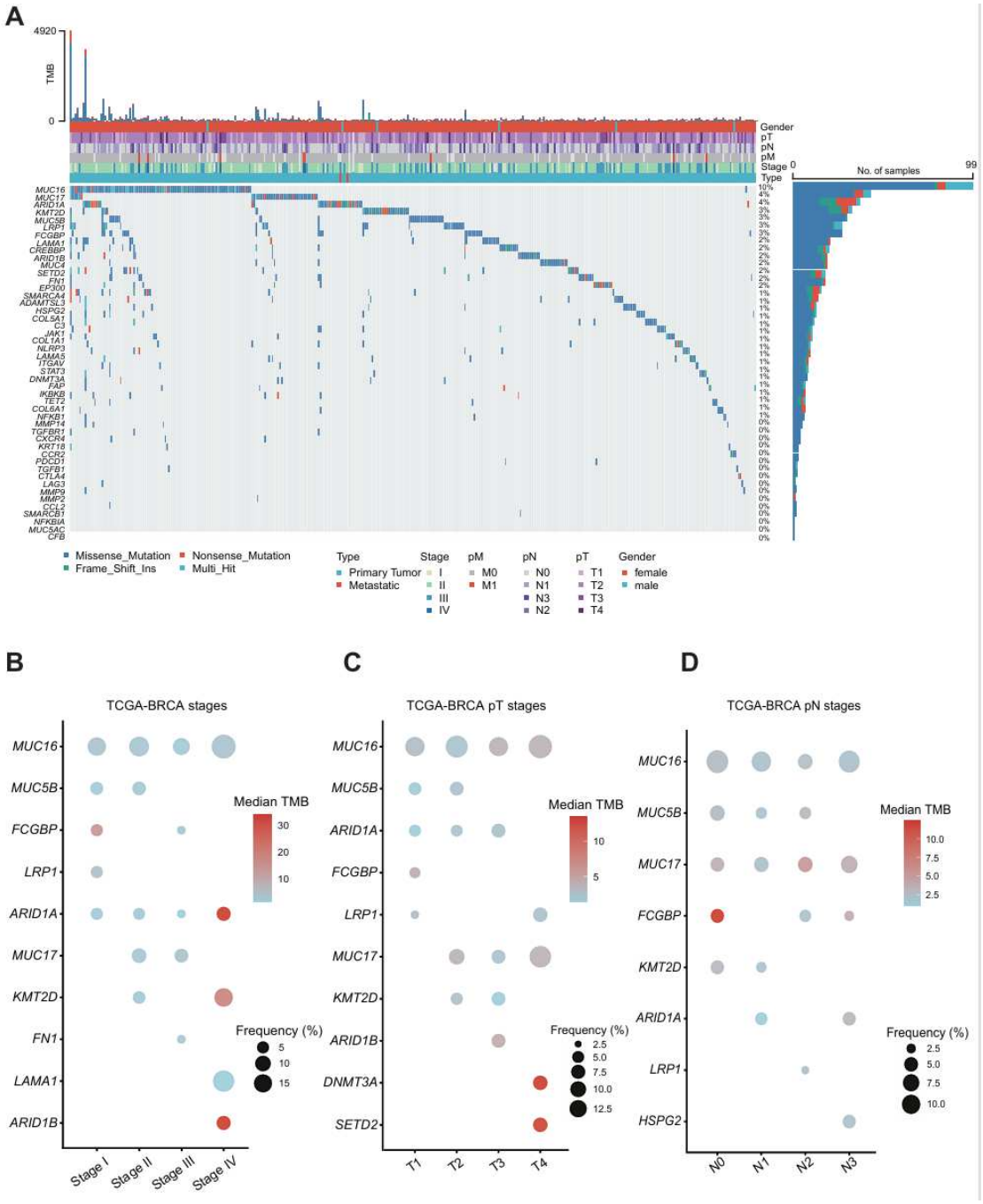


Fig. 4 Tissue-level validation of ctDNA-prioritized candidate genes in TCGA-BRCA. **(A)** Oncoplot showing tissue-level somatic alterations of ctDNA-prioritized candidate genes in the TCGA-BRCA cohort, with clinical annotations including gender, pT stage, pN stage, pM status, disease stage, and sample type. The upper bar plot shows sample-level TMB, and the right bar plot shows the number of mutated samples for each gene. Mutation colors indicate variant classifications, including missense mutation, nonsense mutation, frameshift insertion, and multi-hit events. **(B-D)** Bubble plots showing mutation frequency and median TMB of candidate genes across TCGA-BRCA clinical strata, including overall disease stage

(B), pathological tumor stage **(C)**, and pathological nodal stage **(D)**. Dot size indicates mutation frequency, and color indicates median TMB.

Figure5

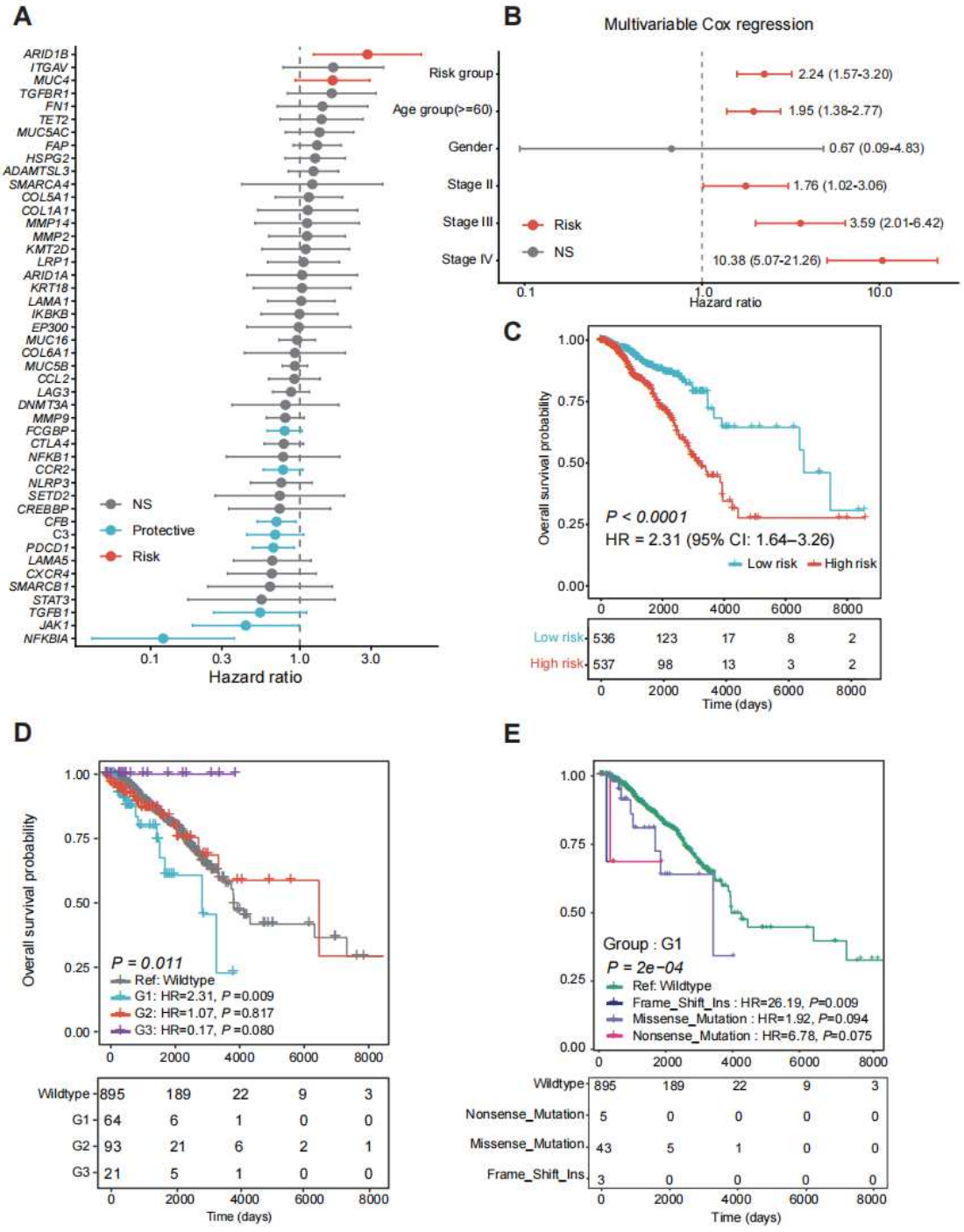


Fig. 5 Prognostic value of ctDNA-prioritized candidate genes in breast cancer.
(A) Univariate Cox regression analysis of ctDNA-prioritized candidate genes using TCGA-BRCA FPKM-normalized bulk RNA-seq data with matched overall survival annotations. *ARID1B* and *MUC4* showed risk-associated effects, whereas *NFKBIA*, *JAK1*, *PDCD1*, *CFB*, and *FCGBP* showed protective-associated trends. Red indicates risk-associated genes, blue indicates protective-associated genes, and gray indicates non-significant genes. **(B)**

Multivariable Cox regression analysis incorporating the seven-gene risk group and clinicopathological variables, including age, gender, and pathological stage. The high-risk group remained independently associated with increased overall survival risk after adjustment for clinical covariates. Advanced pathological stage, particularly stage III and stage IV, was also associated with increased survival risk. **(C)** Kaplan–Meier survival analysis of high- and low-risk groups stratified by the seven-gene risk score. High-risk patients showed significantly worse overall survival than low-risk patients (HR = 2.31, 95% CI: 1.64–3.26, $P < 0.0001$). **(D)** Survival analysis of mutation-defined patient groups based on the seven model genes. Patients were classified as wildtype, G1, G2, or G3 according to mutation status: G1 included patients with mutations only in genes with HR < 1, G2 included patients with mutations only in genes with HR > 1, and G3 included patients with mutations in both HR < 1 and HR > 1 genes. Overall survival differed significantly among these groups ($P = 0.011$). **(E)** Mutation-type-specific survival analysis in G1 patients, showing significant overall survival differences across mutation categories in genes with HR < 1 ($P = 2 \times 10^{-4}$).

Figure 6

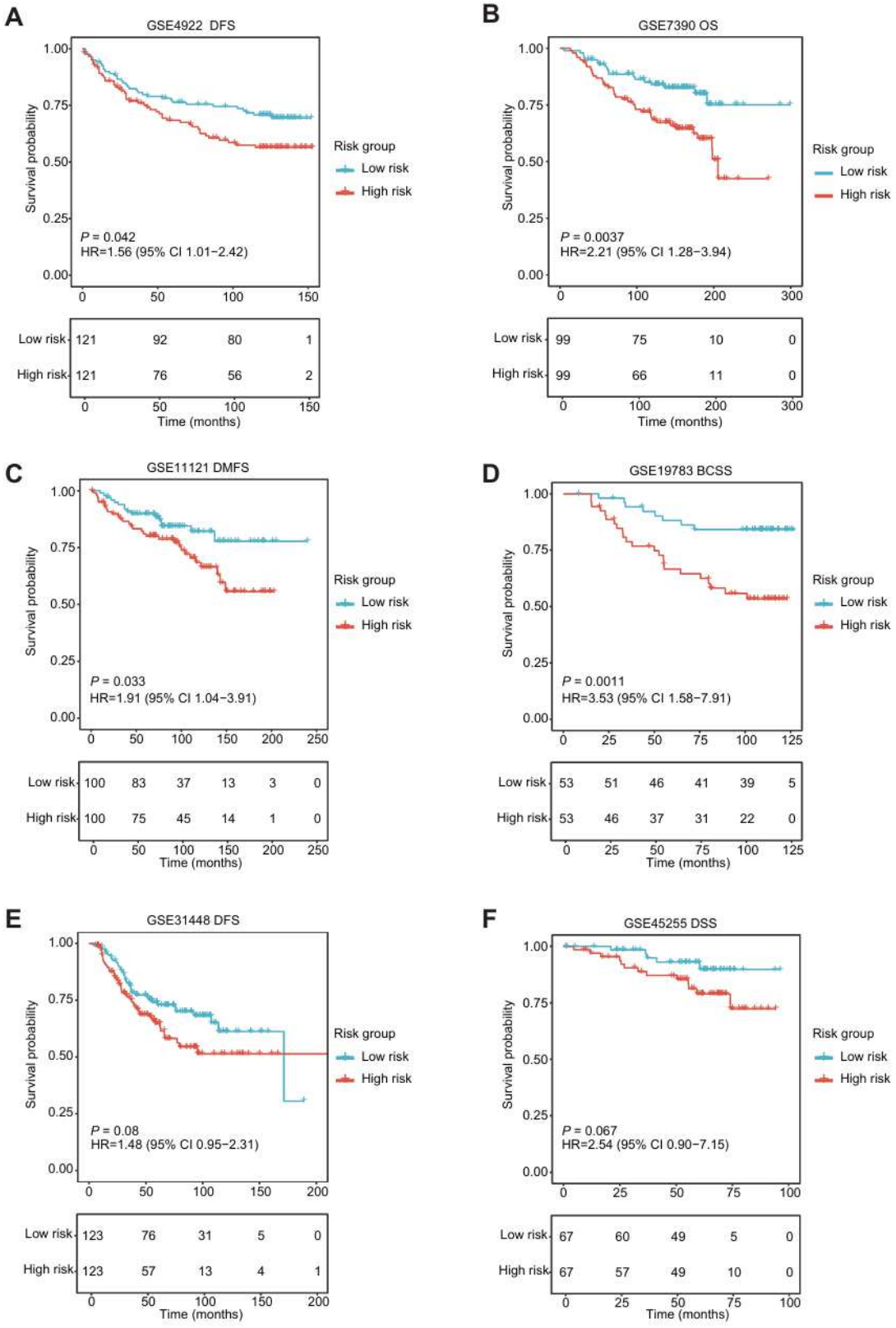


Fig. 6 External validation of the seven-gene prognostic model in independent breast cancer expression cohorts.

(A-F) Kaplan–Meier survival curves showing the prognostic performance of the seven-gene risk model in six independent GEO breast cancer cohorts. Patients in each cohort were

stratified into high- and low-risk groups using the cohort-specific median risk score calculated from *NFKBIA*, *PDCD1*, *ARID1B*, *CFB*, *JAK1*, *MUC4*, and *FCGBP* expression. **(A)** Disease-free survival in GSE4922 (HR = 1.56, $P = 0.042$). **(B)** Overall survival in GSE7390 (HR = 2.21, $P = 0.0037$). **(C)** Distant metastasis-free survival in GSE11121 (HR = 1.88, $P = 0.033$). **(D)** Breast cancer-specific survival in GSE19783 (HR = 3.40, $P = 0.0011$). **(E)** Disease-free survival in GSE31448 (HR = 1.48, $P = 0.08$). **(F)** Disease-specific survival in GSE45255 (HR = 2.40, $P = 0.067$). *P* values were calculated using the log-rank test. DFS, disease-free survival; OS, overall survival; DMFS, distant metastasis-free survival; BCSS, breast cancer-specific survival; DSS, disease-specific survival.

Table 1

Table 1 | Pathology information

Clinicopathologic characteristics		*N (%)*
Age (years)	<60	250 (77.6%)
	>=60	72 (22.4%)
Pathological tumor size (pT)	T1	88 (37.6%)
	T2	111 (47.4%)
	T3	16 (6.8%)
	T4	19 (8.1%)
Clinical nodal status	Negative	155 (57.2%)
	Positive	116 (42.8%)
Stage	I	67 (28.3%)
	II	99 (41.8%)
	III	50 (21.1%)
	IV	21 (8.9%)
ER status	Negative	44 (18.7%)
	Positive	191 (81.3%)
PR status	Negative	42 (18.0%)
	Positive	191 (82.0%)
HER2 status	Negative	94 (39.8%)
	Unkown	105 (44.5%)
	Positive	37 (15.7%)
Ki67	Low	12 (5.2%)
	High	221 (94.8%)
Metastasis status	M0	252 (92.3%)
	M1	21 (7.7%)
Molecular subtype	Luminal A	136 (70.1%)
	Luminal B	30 (15.5%)
	HER2	9 (4.6%)
	Triple Negative	19 (9.8%)

Table 1 Pathology information of the breast cancer ctDNA cohort.
Clinicopathologic characteristics of 323 breast cancer patients with available clinical annotations. Data are shown as number and percentage, n (%). ER, estrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; pT, pathological tumor size; M0, no distant metastasis; M1, distant metastasis.

Table 2

Table 2 | GEO expression datasets pathology information

Characteristic	Category	GSE4922 (n=289)	GSE6532 (n=87)	GSE7390 (n=198)	GSE11121 (n=200)	GSE19783 (n=115)	GSE31448 (n=357)	GSE45255 (n=139)
Age (years)	Age <60	103 (35.6%)	33 (37.9%)	195 (98.5%)	–	–	232 (65.0%)	91 (65.5%)
	Age >=60	146 (50.5%)	54 (62.1%)	3 (1.5%)	–	–	119 (33.3%)	48 (34.5%)
Metastasis status	M0	160 (55.4%)	59 (67.8%)	136 (68.7%)	154 (77.0%)	–	167 (46.8%)	104 (74.8%)
	M1	89 (30.8%)	28 (32.2%)	62 (31.3%)	46 (23.0%)	–	83 (23.2%)	32 (23.0%)
Grade	Grade 1	68 (23.5%)	17 (19.5%)	30 (15.2%)	29 (14.5%)	–	45 (12.6%)	17 (12.2%)
	Grade 2	126 (43.6%)	37 (42.5%)	83 (41.9%)	136 (68.0%)	–	89 (24.9%)	52 (37.4%)
	Grade 3	55 (19.0%)	16 (18.4%)	83 (41.9%)	35 (17.5%)	–	123 (34.5%)	67 (48.2%)
ER status	Negative	34 (11.8%)	–	64 (32.3%)	–	43 (37.4%)	161 (45.1%)	48 (34.5%)
	Positive	211 (73.0%)	87 (100.0%)	134 (67.7%)	–	70 (60.9%)	188 (52.7%)	89 (64.0%)
PR status	Negative	–	21 (24.1%)	–	–	–	153 (42.9%)	60 (43.2%)
	Positive	–	64 (73.6%)	–	–	–	167 (46.8%)	72 (51.8%)
HER2 status	Negative	–	–	–	–	82 (71.3%)	–	69 (49.6%)
	Positive	–	–	–	–	19 (16.5%)	1 (0.3%)	62 (44.6%)
Tumor stage	T1	126 (43.6%)	43 (49.4%)	102 (51.5%)	112 (56.0%)	–	67 (18.8%)	43 (30.9%)
	T2	117 (40.5%)	42 (48.3%)	96 (48.5%)	85 (42.5%)	–	141 (39.5%)	87 (62.6%)
	T3	6 (2.1%)	2 (2.3%)	–	3 (1.5%)	–	70 (19.6%)	9 (6.5%)
	T4	–	–	–	–	–	–	–
Lymph node stage	N0	159 (55.0%)	29 (33.3%)	198 (100.0%)	200 (100.0%)	–	137 (38.4%)	94 (67.6%)
	N1	81 (28.0%)	58 (66.7%)	–	–	–	150 (42.0%)	45 (32.4%)
	N2	–	–	–	–	–	–	–
	N3	–	–	–	–	–	–	–
Subtype	Luminal A	–	–	–	–	45 (39.1%)	90 (25.2%)	–
	Luminal B	–	–	–	–	16 (13.9%)	49 (13.7%)	–
	HER2	–	–	–	–	19 (16.5%)	26 (7.3%)	–
	TNBC	–	–	–	–	16 (13.9%)	98 (27.5%)	–

Table 2 Pathology information of independent GEO breast cancer expression cohorts. Clinicopathologic characteristics of seven independent GEO breast cancer expression datasets used for external validation or cohort characterization, including GSE4922 (n = 289), GSE6532 (n = 87), GSE7390 (n = 198), GSE11121 (n = 200), GSE19783 (n = 115), GSE31448 (n = 357), and GSE45255 (n = 139). Data are shown as number and percentage, n (%). ER, estrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; TNBC, triple-negative breast cancer; M0, no distant metastasis or no distant event; M1, distant metastasis or distant event.

Figure S1

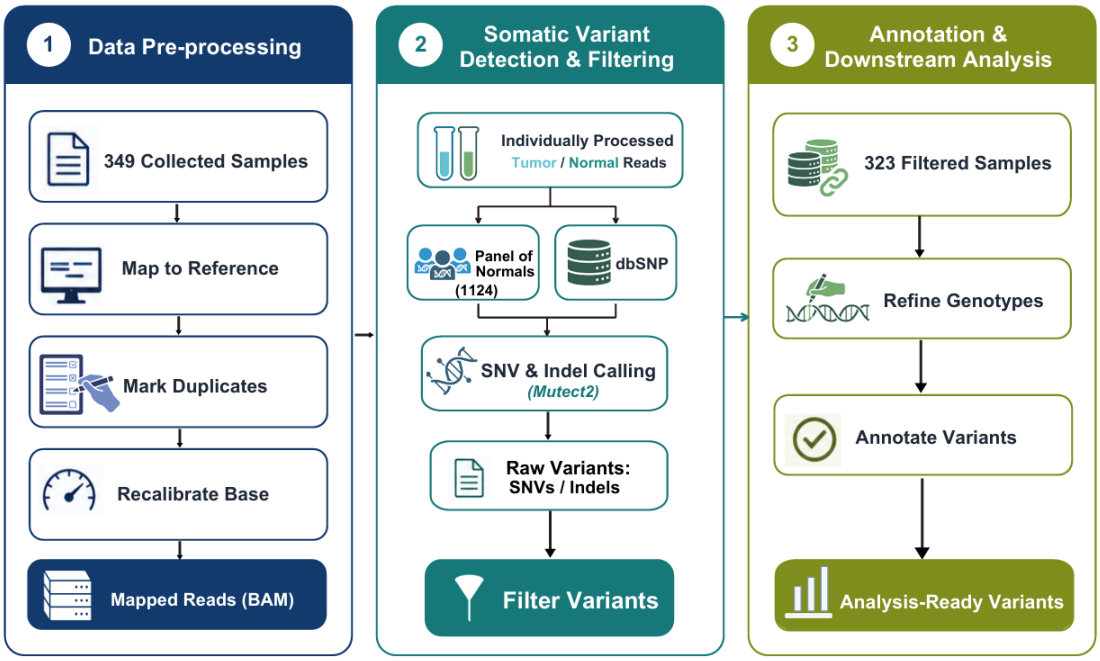


Fig. S1 Variant calling workflow for plasma ctDNA mutation profiling.
Flowchart showing the variant-processing framework used for plasma ctDNA sequencing data. The workflow included three main modules: data preprocessing, somatic variant detection and filtering, and annotation with downstream analysis. A total of 349 collected samples were mapped to the reference genome, followed by duplicate marking and base recalibration to generate BAM files. Somatic SNVs and InDels were called using Mutect2 with a panel of normals constructed from 1,124 normal samples and dbSNP-assisted filtering. After variant filtering, genotype refinement, and functional annotation, 323 filtered samples were retained for downstream mutation analysis. SNVs, somatic single-nucleotide variants; InDels, small insertions and deletions; BAM, binary alignment map; PON, panel of normals.

Figure S2

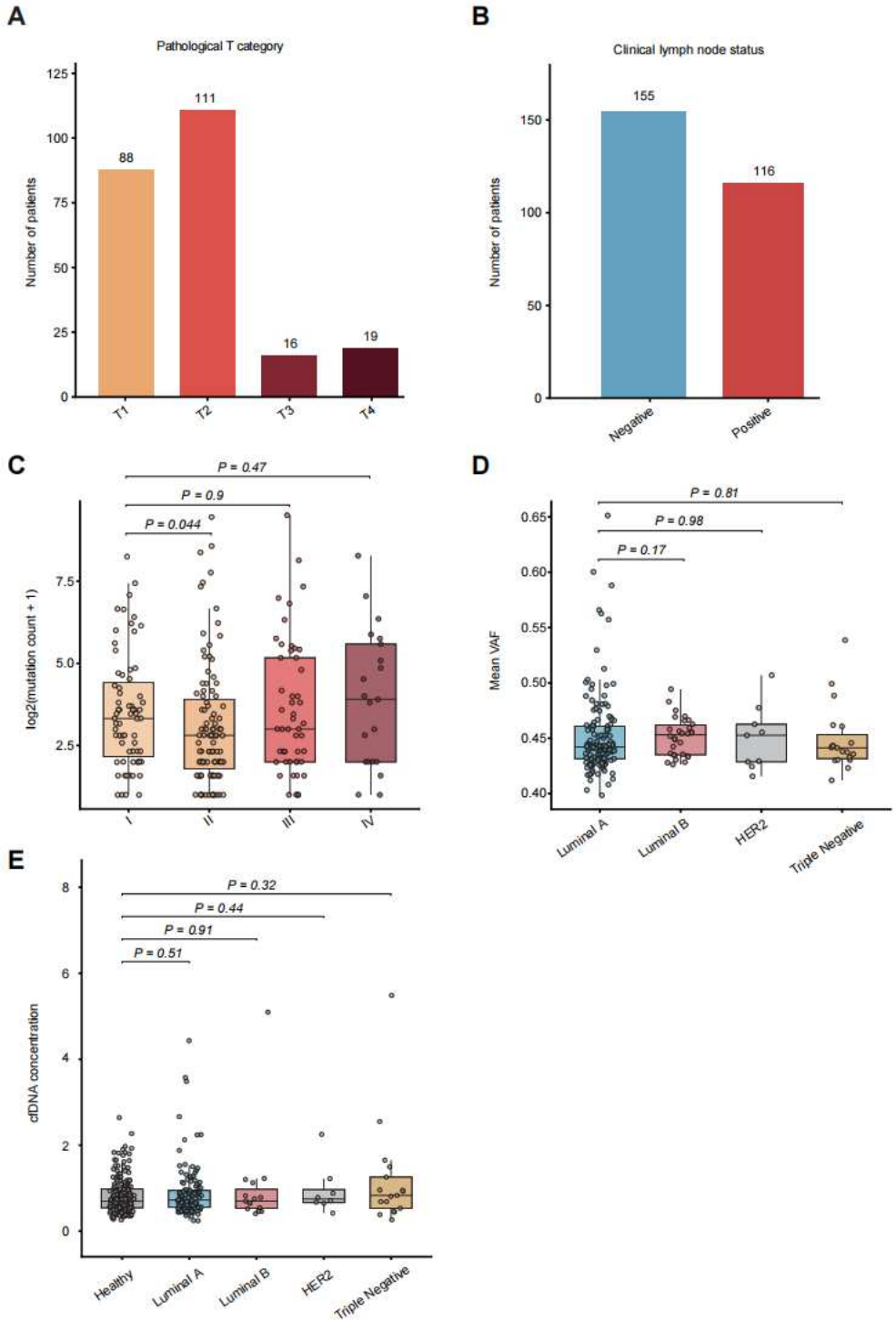


Fig. S2 Additional clinicopathologic and ctDNA-related characteristics of the breast cancer cohort.

(A) Bar plot showing the distribution of pathological tumor size in the annotated cohort, including T1 (n = 88), T2 (n = 111), T3 (n = 16), and T4 (n = 19) tumors. **(B)** Bar plot showing clinical nodal status, with 155 node-negative and 116 node-positive patients. **(C)** Comparison of log2-transformed mutation count across clinical stages. A significant difference was observed between stage I and stage II disease (P = 0.044), whereas other

selected comparisons were not significant. **(D)** Comparison of mean VAF across molecular subtypes, including luminal A, luminal B, HER2-enriched, and triple-negative breast cancer; selected pairwise comparisons were not significant. **(E)** Comparison of pre-library cfDNA concentration between healthy controls and breast cancer molecular subtypes. No significant differences were observed in selected comparisons. P values are shown above the plots. VAF, variant allele frequency.

Figure S3

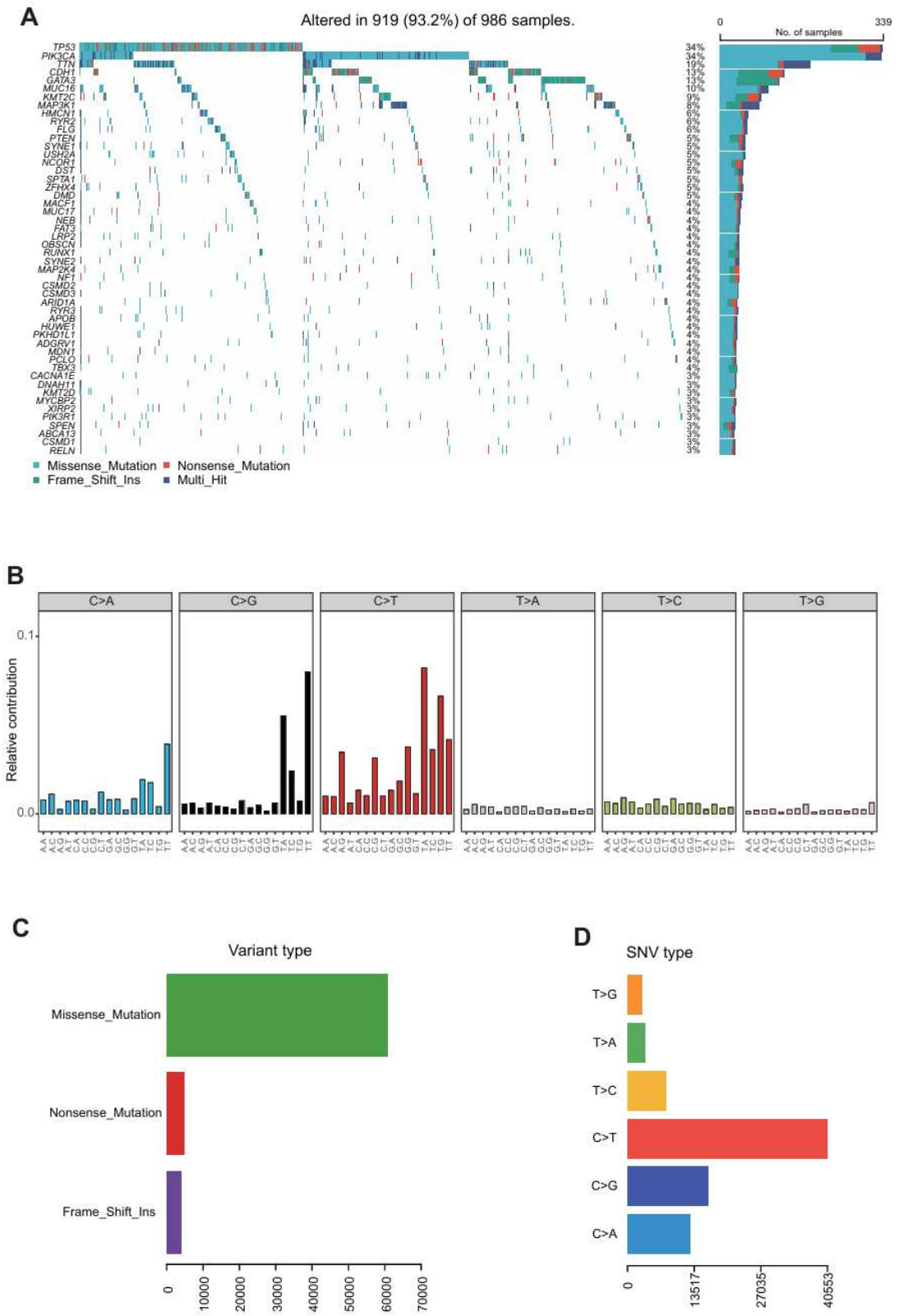


Fig. S3 TCGA-BRCA tissue mutation landscape and mutational spectrum.
(A) Oncoplot showing the top altered genes in the TCGA-BRCA tissue mutation cohort. Somatic alterations were detected in 919 of 986 samples, with *TP53*, *PIK3CA*, *TTN*, *CDH1*,

GATA3, *MUC16*, and *KMT2C* among the most frequently altered genes. Mutation colors indicate variant classifications, including missense mutation, nonsense mutation, frameshift insertion, and multi-hit events. **(B)** Ninety-six-channel trinucleotide mutational spectrum showing the relative contribution of each nucleotide substitution context in TCGA-BRCA tissue samples. **(C)** Bar plot showing the distribution of variant classifications, with missense mutation as the predominant mutation type, followed by nonsense mutation and frameshift insertion. **(D)** Bar plot showing the distribution of SNV classes, with C>T representing the major substitution class.

Figure S4

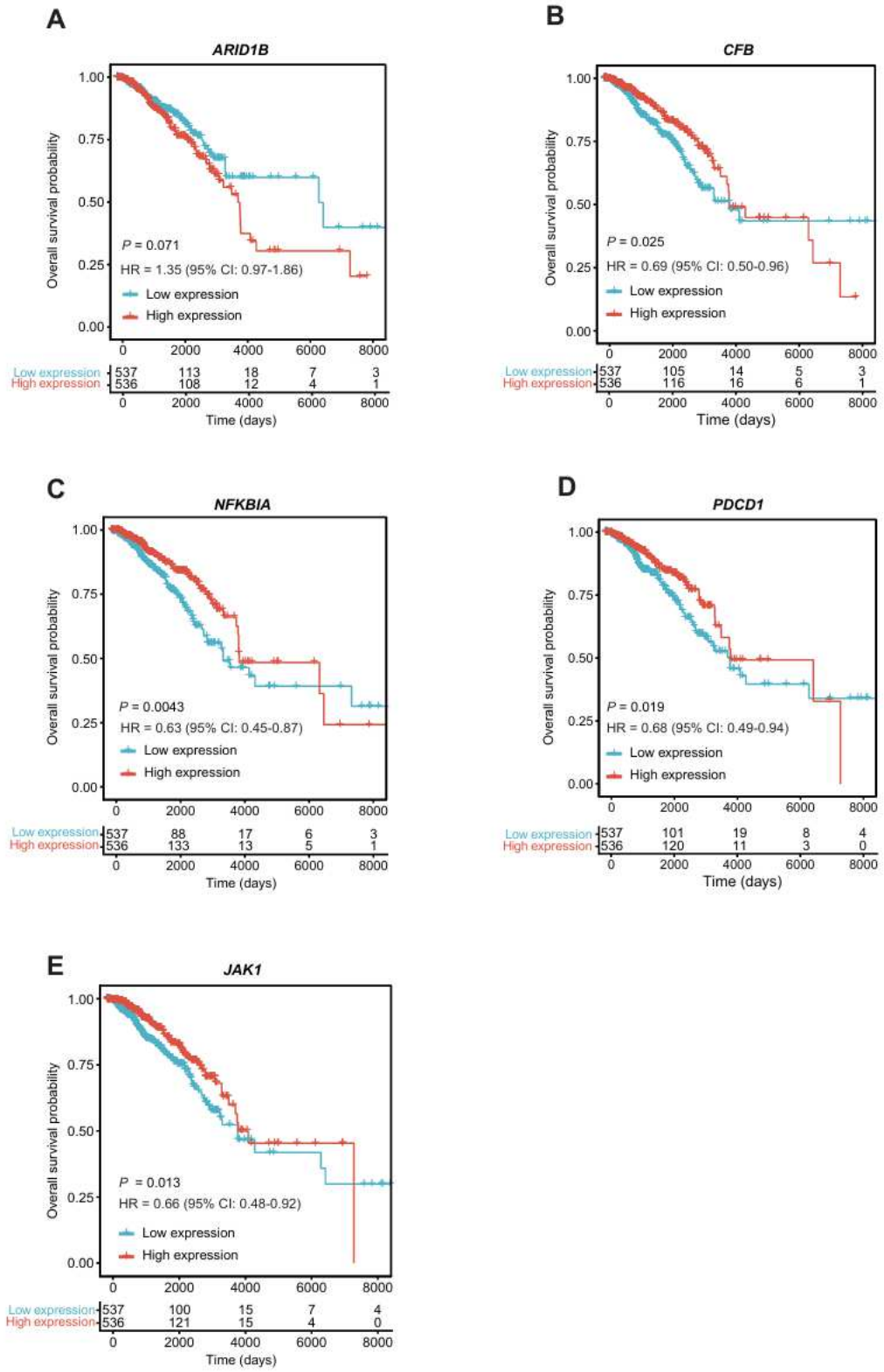


Fig. S4 Kaplan–Meier survival analysis of individual prognostic genes in TCGA-BRCA. Kaplan–Meier curves showing overall survival differences between high- and low-expression

groups for selected ctDNA-prioritized prognostic genes in the TCGA-BRCA cohort. Patients were stratified by the median expression level of each gene, with 537 patients in the low-expression group and 536 patients in the high-expression group. **(A)** High *ARID1B* expression showed a trend toward poorer overall survival compared with low *ARID1B* expression (P = 0.071). **(B)** High CFB expression was associated with significantly better overall survival (P = 0.025). **(C)** High *NFKBIA* expression was associated with improved overall survival (P = 0.0043). **(D)** High *PDCD1* expression was associated with better overall survival (P = 0.019). **(E)** High *JAK1* expression was associated with improved overall survival (P = 0.013). P values were calculated using the log-rank test. OS, overall survival.

Figure S5

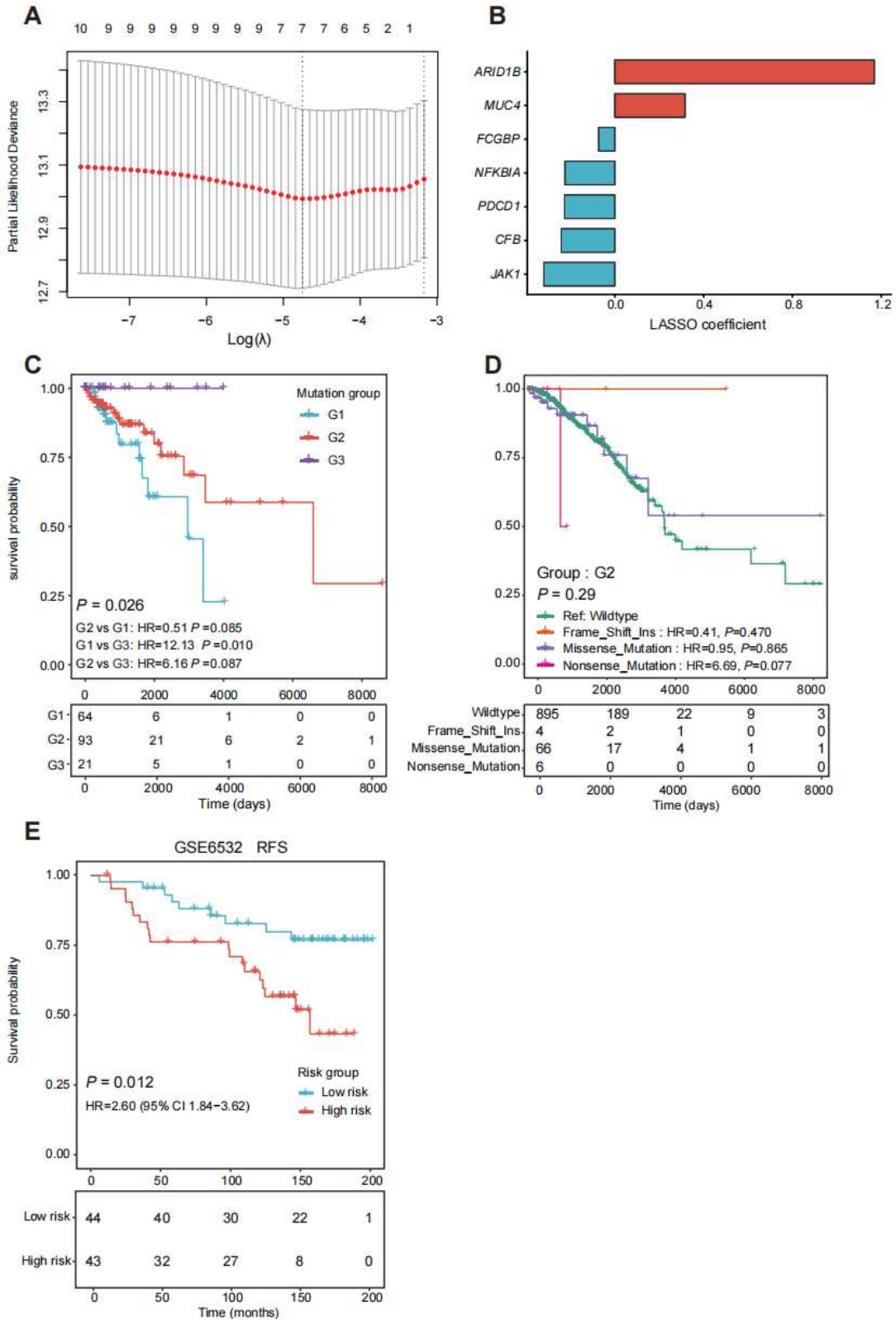


Fig. S5 Additional model construction and mutation-based survival analyses.
(A) Cross-validation curve for LASSO-penalized Cox model showing partial likelihood deviance across different values of log(λ). The dotted vertical lines indicate the selected λ

values. **(B)** Coefficient plots of the seven-gene prognostic model. LASSO coefficients are shown for *ARID1B*, *MUC4*, *FCGBP*, *PDCD1*, *CFB*, *JAK1*, and *NFKBIA*. Red bars indicate risk-associated genes, and blue bars indicate protective-associated genes. **(C)** Kaplan–Meier survival analysis among mutation-positive patients stratified into G1, G2, and G3 groups. G1 included patients with mutations only in genes with HR < 1, G2 included patients with mutations only in genes with HR > 1, and G3 included patients with mutations in both HR < 1 and HR > 1 genes. Overall survival differed significantly among the three mutation-positive groups (P = 0.026). **(D)** Mutation-type-specific survival analysis in G2 patients. No significant overall survival difference was observed across mutation-type groups in genes with HR > 1 (P = 0.29). P values were calculated using the log-rank test. **(E)** Relapse-free survival(RFS) analysis in the GSE6532 cohort. High-risk patients showed significantly poorer relapse-free survival than low-risk patients (HR = 2.60, P = 0.012).

Figure S6

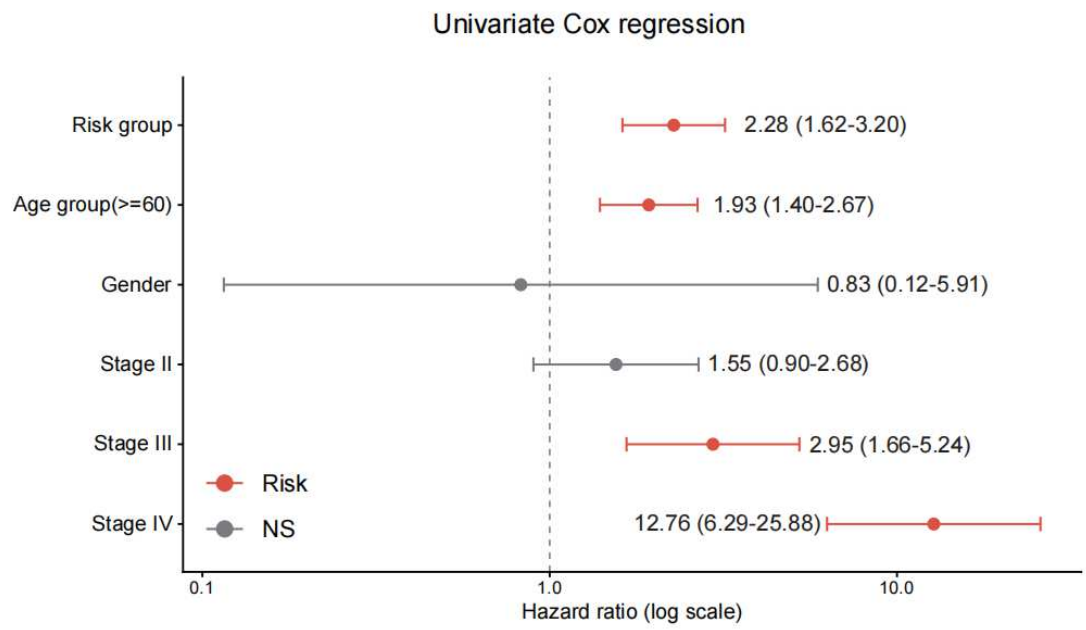


Fig. S6 Univariate Cox regression analysis of the risk score and clinicopathological variables.

Univariate Cox regression analysis evaluating the prognostic effects of the seven-gene risk score and clinicopathological variables in breast cancer patients with matched overall survival annotations. Red indicates risk-associated variables, and gray indicates non-significant variables. Hazard ratios are shown on a log scale. The high-risk group and age ≥60 years were significantly associated with an increased risk of death. Compared with stage I disease, stages III and IV were significantly associated with increased mortality risk, whereas stage II and gender were not statistically significant. Red indicates variables significantly associated with increased risk, and gray indicates non-significant variables. Hazard ratios are shown on a logarithmic scale.

Figure S7

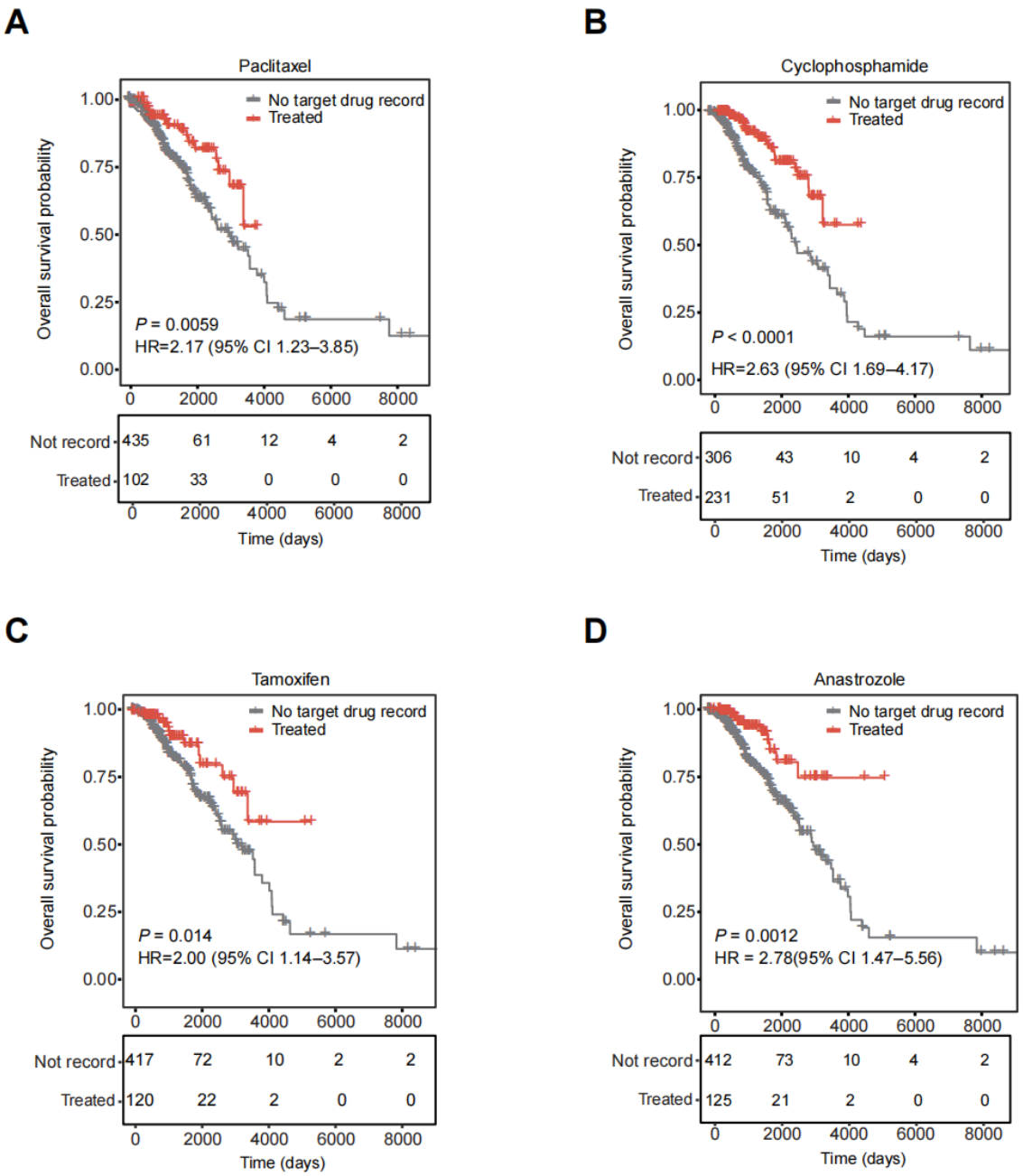


Fig. S7 Treatment-associated survival analysis in the high-risk group.

Kaplan–Meier survival analyses evaluating the associations between documented drug treatment and overall survival among high-risk breast cancer patients. Patients with treatment records for paclitaxel, cyclophosphamide, tamoxifen, or anastrozole showed significantly better overall survival than those without corresponding drug records. (A) Paclitaxel treatment was associated with improved survival (HR = 2.17, 95% CI 1.23–3.85, P = 0.0059) . (B) Similar associations were observed for cyclophosphamide (HR = 2.63, 95% CI 1.69–4.17, P < 0.0001). (C) tamoxifen (HR = 2.00, 95% CI 1.14–3.57, P = 0.014) and (D) anastrozole (HR = 2.78, 95% CI 1.47–5.56, P = 0.0012).

Supplementary Files

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

- Table1pathologyinformation.csv
- Table2GEOpathologyinformation.csv