

# Integrative profiling of multi-modal plasma cfRNA signatures enables detection and prognostic risk stratification in breast cancer

Jun Wang<sup>a,1</sup>, Liu Yang<sup>a,1</sup>, Kai Fang<sup>b,c,1</sup>, Hongxian Wang<sup>d</sup>, Jianxia Chen<sup>e</sup>, Binbin Tan<sup>f</sup>, Lirong Shu<sup>b</sup>, Jinfeng Ye<sup>b</sup>, Jiayu Wu<sup>a</sup>, Xinyang Lin<sup>a</sup>, Deming Gou<sup>a,\*</sup>, Yun Wang<sup>b,c,\*</sup>

<sup>a</sup> Shenzhen Key Laboratory of Microbial Genetic Engineering, Vascular Disease Research Center, College of Life Sciences and Oceanography, Guangdong Provincial Key Laboratory of Regional Immunity and Disease, Carson International Cancer Center, School of Medicine, Shenzhen University, Shenzhen 518060, China

<sup>b</sup> Longhua Innovation Institute for Biotechnology, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen 518055, China

<sup>c</sup> Center for Clinical Research and Translation, Southern University of Science and Technology Hospital, Shenzhen 518055, China

<sup>d</sup> Thyroid and Breast Department, Huazhong University of Science and Technology Union Shenzhen Hospital, Shenzhen 518052, China

<sup>e</sup> Clinical Laboratory, Longgang Central Hospital of Shenzhen, Shenzhen 518116, China

<sup>f</sup> Guangdong Provincial Key Laboratory of Regional Immunity and Diseases, Department of Pharmacology and International Cancer Center, School of Medicine, Shenzhen University, Shenzhen 518060, China

## ARTICLE INFO

### Keywords:

Breast cancer  
Cell-free RNA  
Diagnosis  
Prognosis  
Tumor microenvironment

## ABSTRACT

Breast cancer requires non-invasive biomarkers for accurate detection and risk stratification. We comprehensively profiled plasma cell-free RNA (cfRNA) from 41 patients with malignant and 42 with benign breast lesions using SLiPiR-seq. Multiple cfRNA subtypes displayed distinct expression patterns, and machine-learning models were developed with repeated stratified four-fold cross-validation. The integrated cfRNA model achieved a mean AUC of 0.795, while the cf-miRNA model performed best (AUC: 0.814) and was externally validated in an independent cohort (AUC: 0.867). A three-gene tissue expression signature comprising DLST, DOCK4, and EGFL7 further stratified patients by overall survival in TCGA-BRCA. High-risk tumors showed increased PIK3CA mutations, PI3K-AKT pathway activation, and altered immune features. Single-cell analysis revealed distinct localization of the three genes across epithelial cells, tumor-associated macrophages, cancer-associated fibroblasts, and T-cell populations. Collectively, multi-modal plasma cfRNA profiling may enable non-invasive breast cancer detection and provide candidate markers for prognostic risk stratification and future precision oncology applications.

## 1. Introduction

Breast cancer is one of the most prevalent malignancies worldwide and a leading threat to women's health [1,2]. Despite significant progress in early detection and therapeutic intervention, breast cancer continues to pose a substantial clinical challenge due to its inherent molecular heterogeneity, and the diverse biological mechanisms driving

its initiation and progression [3,4]. Current diagnostic approaches primarily rely on invasive biopsies, mammography, and immunohistochemical (IHC) evaluation of biomarkers including estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) [5,6]. These methods, however, are associated with invasiveness, limited sensitivity at early-stage of tumors, and an inability to dynamically monitor disease progression or therapeutic

**Abbreviations:** cfRNA, cell-free RNA; TNBC, triple-negative breast cancer; ER, estrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; BEN, benign breast lesions; MAL, malignant breast lesions; RF, Random Forest; LR, Logistic Regression; GB, Gradient Boosting; HGB, Histogram-based Gradient Boosting; AUC, The area under the receiver operating characteristic curve; ROC, Receiver operating characteristic; KM, Kaplan-Meier; TME, tumor microenvironment; GO, Gene Ontology; BP, Biological Processes; CC, Cellular Components; MF, Molecular Functions; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, Gene Set Enrichment Analysis; DERNA, Differentially expressed RNAs; FDR, False Discovery Rate; log2FC, log2FoldChange; DEGs, Differentially Expressed Genes; HR, Hazard ratio; CI, confidence intervals; SNVs, single nucleotide variants; VAF, Tumor variant allele frequency; TAMs, tumor-associated macrophages; TCA, tricarboxylic acid.

\* Corresponding authors.

E-mail addresses: [dmgou@szu.edu.cn](mailto:dmgou@szu.edu.cn) (D. Gou), [wangyun@sustech-hospital.cn](mailto:wangyun@sustech-hospital.cn) (Y. Wang).

<sup>1</sup> Co-first authors: Jun Wang, Liu Yang and Kai Fang contributed equally to this article.

<https://doi.org/10.1016/j.ygeno.2026.111324>

Received 14 February 2026; Received in revised form 17 August 2026; Accepted 21 September 2026

Available online 22 September 2026

0888-7543/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

responses [7,8].

Early detection is pivotal for improving outcomes of breast cancer, as localized breast cancer exhibits a 5-year survival rate exceeding 90%, compared to 27% for metastatic cases [9,10]. Yet conventional imaging techniques, such as mammography and ultrasound, frequently miss small tumors or morphologically ambiguous lesions [11–13]. In addition, conventional molecular subtyping based on ER/PR/HER2 expression, combined with the Ki-67 proliferation index, provides insight into tumor behavior but falls short in offering a detailed prognosis for certain aggressive subtypes, such as TNBC patients who currently has limited targeted therapeutic options [14,15]. Consequently, there is a critical need to develop novel biomarkers that can capture the dynamic biological processes during tumor progression.

Liquid biopsy, which analyze circulating biomarkers in blood, has recently emerged as a promising non-invasive approach for cancer detection and monitoring. Cell-free DNAs (cfDNAs) has been extensively explored for mutation detection, minimal residual disease (MRD) monitoring, and treatment response assessment [16–18]. However, the clinical utility of cfDNA is constrained by its relatively short half-life (16–160 min) and often low abundance, which may limit its sensitivity for early detection and continuous monitoring [19,20]. In contrast, cell-free RNAs (cfRNAs) encompassing miRNAs, mRNAs, lncRNAs, tRNAs, and other species exhibits remarkable stability in plasma due to its encapsulation within extracellular vesicles (EVs) or association with RNA-binding proteins [21–23]. Unlike cfDNA, cfRNA can enter circulation via active secretion mechanisms, such as exosome-mediated signaling; circulating RNA originates from a mixture of cell types, including tumor cells, immune cells, and other host cells [24,25]. Furthermore, cfRNA may reflect the real-time transcriptional activity of its cellular origins, including the tumor and its microenvironment, offering a rich source of diagnostic and prognostic insights. Several studies have demonstrated that specific cfRNA signatures can distinguish malignant from benign conditions, as well as predict clinical outcomes. For instance, circulating miRNA panels (e.g., miR-21, miR-106b, miR-181a, miR-484 and miR-1260b) have achieved an AUC of 0.908 in distinguishing malignant from benign breast lesions [25]. While tumor-derived mRNAs enable non-invasive detection of cancer [26], non-coding RNAs (ncRNAs), such as lncRNAs and tRNAs, also contribute to oncogenesis by modulating immune responses and signaling pathways, underscoring their biomarker potential [27,28].

Despite these advances, most of the investigations focused on a single class of cfRNAs, potentially overlooking the synergistic insights that could be gained by analyzing the entire cfRNA landscape. In this study, we sought to address this gap by conducting comprehensive profiling of the plasma cfRNA transcriptome in patients with breast cancer. To this end, we applied SLiPiR-seq, a cfRNA library preparation method developed by our group that is independent of RNA terminal modifications and enables the simultaneous capture of both small and long cfRNA species (cf-miRNA, cf-tRNA, cf-ysRNA, cf-rsRNA, cf-piRNA, cf-mRNA, and cf-lncRNA) from as little as 200  $\mu$ L of plasma [29]. By simultaneously assessing a wide array of cfRNAs, including circulating miRNAs, mRNAs, lncRNAs, tRNAs, ysRNAs, rsRNAs, and piRNAs, we aimed to develop a diagnostic model using machine learning, identify potential prognostic biomarkers linked to tumor microenvironment remodeling, and establish a framework for molecular subtyping to refine risk stratification. By integrating multi-omics analysis, we also intended to decipher the interplay among cfRNA dynamics, immune evasion mechanisms, and signatures of somatic mutations. As protein-coding genes possess well-annotated biological functions amenable to tissue-level follow-up in public cohorts, cf-mRNAs was further explored for its potential prognostic relevance in tumor tissue.

2. Materials and methods

2.1. Clinical cohort and sample collection

Blood samples from patients with malignant ( $n = 41$ ) and benign ( $n = 42$ ) breast lesions were obtained at the time of diagnosis from Huazhong University of Science and Technology Union Shenzhen Hospital. All blood samples were collected at the time of first diagnosis, before any treatment (surgery, chemotherapy, radiotherapy, or endocrine therapy) and without any pretreatment. Whole blood samples were collected by experienced nurses using anticoagulant tubes. Fresh peripheral blood was gently inverted several times to ensure proper mixing and immediately placed on ice. Samples were balanced and centrifuged (4  $^{\circ}$ C, 1000 rpm, 5 min) to separate components. A fresh 50 mL centrifuge tube was prepared for each sample and labeled. 10 mL of lymphocyte separation medium was added to each tube. After centrifugation, 4–5 mL of the upper yellow layer (plasma) was carefully aspirated. The plasma was aliquoted into pre-labeled 1.5 mL EP tubes (0.5–1 mL per tube). Tubes were immediately placed on ice for short-term storage, then transferred to a  $-80^{\circ}$ C freezer for long-term preservation. All patients in this study signed an informed consent document approved by the Institutional Review Board before entering any study. Clinical characteristics for all participants in this study were listed in Table 1 and Table S1.

2.2. cfRNA extraction and sequencing library construction

Following established protocols [29], cfRNA extraction and library construction were performed with slight modifications. In brief, 200  $\mu$ L of plasma was processed using the Apostle MiniMax™ High Efficiency cfRNA Isolation Kit (Apostle), and the cfRNA was eluted in 10  $\mu$ L of RNase-free water for subsequent library preparation. For the simultaneous polyadenylation and reverse transcription (RT) reactions, a 10  $\mu$ L reaction was assembled in 0.2 mL PCR tubes containing 6.75  $\mu$ L of the extracted cfRNA, 1.25  $\mu$ L of 4 $\times$  polyA/RT buffer, 1  $\mu$ L of a 5 nM RT primer, and 1  $\mu$ L of a premixed solution of poly-A polymerase (Enzymatics) and M-MLV reverse transcriptase (Fapon Biotech). The mixture was incubated at 37  $^{\circ}$ C for 30 min, then supplemented with 1  $\mu$ L of Exonuclease I (New England BioLabs) to remove residual primers, followed by an additional incubation at 37  $^{\circ}$ C for 30 min and a subsequent 80  $^{\circ}$ C incubation for 20 min. After a 5-min denaturation at 95  $^{\circ}$ C to generate single-stranded cDNA, the 11  $\mu$ L RT product was transferred to a ligation reaction. This reaction contained 1  $\mu$ L T4 DNA ligase, 1  $\mu$ L of

Table 1  
Clinical information of the participants in this study.

|         | BEN cohort  | MAL cohort                                                                                                |
|---------|-------------|-----------------------------------------------------------------------------------------------------------|
| n       | 42          | 41                                                                                                        |
| Age     | 22–53       | 24–71                                                                                                     |
| Gender  | F (42,100%) | F (41,100%)                                                                                               |
| T Stage | N/A         | T1 (19, 46.3%)<br>T2 (12, 29.2%)<br>T3 (4, 9.8%)<br>T4 (2, 4.9%)<br>N/A (4, 9.8%)                         |
| N Stage | N/A         | N0 (23, 56%)<br>N1 (9, 22%)<br>N2 (2, 4.9%)<br>N3 (3, 7.3%)<br>N/A (4, 9.8%)                              |
| M Stage | N/A         | M0 (36, 89%)<br>M1 (1, 2.4%)<br>N/A (4, 9.8%)                                                             |
| Stage   | N/A         | Stage I (16, 39%)<br>Stage II (11, 26.8%)<br>Stage III (10, 24.4%)<br>Stage IV (1, 2.4%)<br>N/A (3, 7.3%) |

10× T4 DNA Ligase buffer (New England BioLabs), 2 µL of 50% polyethylene glycol 4000 (Thermo), 2 µL of 10 mM ATP (Enzymatics), 1 µL of a 10 µM pre-mixed synthetic adapter, and 2 µL of double-distilled water. Ligation was carried out at 20 °C for 1 h, followed by a 10-min incubation at 65 °C. Subsequently, 1 µL of USER enzyme (New England BioLabs) was added and the mixture was incubated at 37 °C for 15 min. Amplification of the final libraries was performed in a 50 µL reaction using 2× KAPA HiFi HotStart ReadyMix (Roche Diagnostics) under the following cycling conditions: an initial denaturation at 95 °C for 3 min, then 16 cycles of 98 °C for 20 s, 60 °C for 30 s, 72 °C for 30 s, and a final extension at 72 °C for 1 min. Purification was achieved using AMPure XP beads (Beckman Coulter) at 1.8× volume, and libraries were sequenced on an Illumina NovaSeq 6000 (HaploX, Shenzhen) to yield at least 20 million paired-end reads per sample. Pre-analytical variables were strictly controlled: plasma was isolated within 3 h of blood collection, hemolyzed samples were discarded, and all 200 µL aliquots were subjected to only one freeze-thaw cycle before use. A fixed-volume input (200 µL) was adopted to circumvent quantification-induced bias, and the post-amplification library concentrations are listed in Supplementary Table S1.

### 2.3. Raw reads processing and cfRNA annotation

Raw sequencing reads were processed through a bioinformatics pipeline adapted from previously established methods to generate expression matrices for each cfRNA sub-species [29]. Initially, raw reads underwent quality control using Cutadapt (v2.10) and Trimmomatic (v0.32) to remove adapter sequences, low-quality reads, and any reads shorter than 15 nucleotides. The remaining high-quality reads were sequentially aligned to reference databases for cf-rsRNA, cf-ysRNA, cf-tRNA, cf-miRNA, and cf-piRNA using a custom Python script as previously described. Reads not mapping to these RNA references were then aligned to the GRCh38 (hg38) reference genome using Bowtie2 [29]. mRNA and lncRNA expression levels were annotated using FeatureCounts with the GENCODE annotation (release 41 [30]).

### 2.4. Calculation of coverage and fragment length distribution

Coverage analysis for RNY and rRNA was performed by mapping reads to their corresponding parent RNAs (RNY1, RNY2, RNY4, RNY5; 28S rRNA, 18S rRNA, 5S rRNA, 5.8S rRNA, mt12s rRNA, mt16s rRNA). Coverage calculations were based on reads mapped across the entire length of these RNA sequences. Fragment length distributions were determined by analyzing the size and depth of aligned reads across different RNA sub-species using a custom Python script as described previously [29].

### 2.5. Differential expression analysis

Raw count matrices were filtered to retain genes expressed in over 75% of the samples. The filtered count data were then analyzed for differential expression using the DESeq2 R package [31], which models count data using a negative binomial distribution. Differentially expressed cfRNAs (DE-cfRNAs) were identified based on a significance threshold of  $p$ -value  $< 0.05$ , determined using the Wald test. Visualization of expression patterns was performed using volcano plots generated with the ggplot2 R package and clustering heatmaps produced with the pheatmap R package [32].

### 2.6. Development and evaluation of breast cancer diagnostic models

To avoid data leakage, all data-dependent steps were performed strictly within each training fold. Briefly, cfRNA expression matrices were normalized to counts per million (CPM). Within each training fold, differentially expressed cfRNAs (DE-cfRNAs) were identified by the Mann-Whitney  $U$  test ( $p$ -value  $< 0.05$ ) with  $|\log_2\text{FoldChange}(\log_2\text{FC})|$

$> 1$  for cf-ysRNA and cf-rsRNA and  $|\log_2\text{FC}| > 0.60$  for the other cfRNA sub-species, and the top 10 features of each cfRNA type were selected; ratio features constructed from negatively correlated gene pairs (Spearman  $r < -0.35$ , up to eight pairs per type) and per-fold mean-centering batch correction were additionally applied. Features were scaled using the QuantileTransformer before model fitting. Random Forest (RF), Logistic Regression (LR) and HistGradientBoosting (HGB) models were trained with the scikit-learn library in Python under strong regularization (RF: maximum depth of 3, minimum samples per leaf of 8, 500 estimators, balanced class weights; LR:  $C = 0.01$ , L2 penalty, balanced class weights; HGB: maximum depth of 2, learning rate of 0.03, L2 regularization of 10, balanced class weights) [33]. Each model underwent 100 iterations of training, employing stratified shuffle split cross-validation to maintain class distribution. Model hyper-parameters were optimized as follows. RF and GB: maximum depth of 3, minimum samples required to split an internal node set to 10, minimum samples required at a leaf node set to 5, 500 estimators, and balanced class weights. LR: up to 1000 iterations with balanced class weights. Models were evaluated by repeated stratified 4-fold cross-validation (20 repeats; 80 independent evaluations) on the 83 study samples. For each cfRNA type and for the integrated all-RNA feature set, the best-performing model was selected based on the mean cross-validated area under the receiver operating characteristic curve (AUC), and out-of-fold (OOF) predictions were collected for ROC visualization. For external validation, the cf-miRNA models trained on the present cohort were applied to an independent cohort (GSE270497;  $n = 180$ ) after quantile transformation of both datasets. The top 10 features for each cfRNA type were visualized using boxplots generated with the matplotlib and seaborn libraries in Python [34,35].

### 2.7. Survival analysis

Prognostic relevance of the top cf-mRNAs contributing to the diagnostic models was evaluated using clinical data from the TCGA-BRCA dataset [36]. Univariate Cox proportional hazards regression analysis was performed to identify genes associated with overall survival. Patients were stratified by cumulative gene expression of the three signature genes (DLST, DOCK4, EGFL7): patients with  $\geq 2$  of the three genes expressed above the cohort median were defined as high-risk, and those with  $\leq 1$  gene above the cohort median were defined as low-risk. Kaplan-Meier (KM) survival curves were plotted to visualize associations between identified RNA signature and patient survival outcomes. Analysis was conducted using the “survival” and “survminer” R packages, with survival curves generated using the ggSurv function from the survminer package.

### 2.8. Immune infiltration analysis

The tumor microenvironment (TME) of breast cancer was characterized using multiple computational methods. The relative abundance of various immune cell types in tumor samples was estimated using the CIBERSORT algorithm [37]. For this analysis, the LM22 signature matrix was employed, encompassing 22 immune cell types. The ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data) algorithm was applied to assess stromal and immune cell content within tumor tissues, which generates three scores: stromal score, immune score, and estimate score.

### 2.9. Genomic mutation analysis in risk groups

Somatic mutation profiles in high- and low-risk patients were obtained from the TCGA-BRCA dataset and analyzed using the maftools R package to visualize mutation landscapes and assess the potential impact of mutations on gene function and patient prognosis [38].

## 2.10. Single-cell analysis of the prognostic RNA signature

Single-cell RNA sequencing data from the GEO dataset GSE161529 were analyzed to investigate the expression of the prognostic RNA signature (DLST, DOCK4, and EGFL7) and their roles in the TME. Cell annotation was performed using established marker genes. Analysis was conducted using the Seurat5 R package for data normalization, clustering, and visualization, with cell-cell trajectory analysis performed using the CytoTRACE and Monocle R packages [39,40], and cell-cell communication analysis was conducted using the CellChat R package [41].

## 2.11. Gene function and pathway enrichment analysis

The R package clusterProfiler was employed for Gene Ontology (GO) analysis, assessing enrichment in Biological Processes (BP), Cellular Components (CC), and Molecular Functions (MF). Additionally, this package was used for Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis and Gene Set Enrichment Analysis (GSEA) to explore associated signaling pathways [42,43]. For the analysis of DE-cf-mRNA, GO terms with a minimum of two enriched genes were selected for subsequent further analysis and visualization using the R package clusterProfiler.

For the GSEA analysis, DERNA's were utilized as input. Subsequently, the terms for each ontology in GO and KEGG were selected for visualization using R package ggplot2.

## 3. Results

### 3.1. Comprehensive cfRNA profiling revealed distinct expression patterns in malignant and benign breast lesions

To comprehensively investigate the cfRNA landscape in breast cancer, plasma samples were collected from 41 patients with stage I-IV malignant breast cancer and 42 age-matched patients with benign breast lesions between 2018 and 2019. Plasma cfRNAs were profiled using the previously established SLiPiR-seq method [29]. Multiple RNA sub-species were identified in both malignant (MAL) and benign (BEN) breast lesion cohorts, including cf-miRNA, cf-mRNA, cf-lncRNA, cf-tRNA, cf-ysRNA, cf-piRNA, and cf-rsRNA. Notably, cf-tRNA and cf-ysRNA represented the most abundant fractions, accounting for approximately 45.30% and 26.50% of the mapped cfRNA respectively (Fig. 1A and Table S2).

Principal component analysis (PCA) of both the overall cfRNA transcriptome and individual cfRNA sub-species revealed clustering of MAL and BEN samples (Fig. 1B and S1A). Correlation analysis demonstrated a high overall concordance in cfRNA expression between the MAL and BEN cohorts ( $R = 0.98$ , Fig. 1C and S1B), a pattern that was consistently observed across individual cfRNA sub-species. These findings indicate robust intra-group reproducibility with no detectable batch effects, suggesting the inter-group differences reflect intrinsic biological variations between MAL and BEN samples.

To investigate co-expression patterns of tRNAs, we categorized cf-tRNA species based on their parent tRNAs. Interestingly, cf-tRNA derived from phenylalanine (tRNA-Phe), glycine (tRNA-Gly), methionine (tRNA-Met), tryptophan (tRNA-Trp), and arginine (tRNA-Arg) exhibited mutually antagonistic expression profiles (Fig. S1C), suggesting potential competition during biogenesis. Further analysis of parent RNA abundance for cf-ysRNA, cf-tRNA, and cf-rsRNA revealed that tRNA-Phe was significantly reduced in MAL compared to BEN cohort, while tRNA-Gly and tRNA-Met showed elevated trends (Fig. 1D). No significant differences were observed for other cfRNA sub-species.

Moreover, mapping of cf-ysRNA and cf-rsRNA reads to their parent RNAs demonstrated conserved fragmentation patterns in both MAL and BEN cohorts (Fig. 1E, S1D, and Table S3). Fragment length distribution analysis of cfRNA sub-species revealed distinct size ranges: cf-mRNA and

cf-lncRNA ranged from 15 to 150 bp, cf-miRNA peaked at 22 bp, and cf-tRNA centered at 17 bp. Despite these differences, fragment length profiles were consistent between MAL and BEN cohorts (Fig. 1F and Table S4), reinforcing high intra-group homogeneity and supporting the reproducibility of the initial analysis.

### 3.2. Development of a cfRNA-based diagnostic model for breast cancer prediction

To investigate whether cfRNA profiles could distinguish MAL from BEN cases, we first identified differentially expressed cfRNAs (DE-cfRNAs) between cohorts. Using stringent criteria (expression in  $\geq 75\%$  of samples,  $|\log_2FC| \geq 1$ , and  $p$ -value  $< 0.05$ ), we detected 978 DE-cfRNAs (Fig. 2A–B and Table S5). As anticipated, cf-rsRNA, cf-tRNA, and cf-ysRNA constituted the largest proportions of DE-cfRNAs, reflecting their abundance and dysregulation in malignancy (Fig. 2B).

Gene Set Enrichment Analysis (GSEA) of the DE-cf-mRNAs identified several pathways distinguishing MAL and BEN patient groups, including *cytoplasmic translation* (false discovery rate (FDR) =  $7.801 \times 10^{-08}$ ), *cell morphogenesis involved in neuron differentiation* (FDR = 0.021), and *ribosome biogenesis* (FDR =  $9.778 \times 10^{-05}$ ) (Fig. 2C).

To construct the diagnostic models while preventing information leakage, all feature-selection steps were performed strictly within each training fold, and models were evaluated by repeated stratified 4-fold cross-validation (20 repeats; 80 independent evaluations; details in Section 2.6). Under this protocol, cf-miRNA achieved the highest mean AUC ( $0.814 \pm 0.088$ , Random Forest), cf-ysRNA the lowest ( $0.700 \pm 0.122$ ), and the remaining cfRNA species ranged from 0.756 to 0.789; the model integrating all cfRNA species achieved a mean AUC of  $0.795 \pm 0.090$  (Fig. 2D–E and S2A). To further assess generalizability, the cf-miRNA models were applied to an independent external cohort (GSE270497;  $n = 180$ , 120 screening-positive and 60 negative), in which the logistic-regression model achieved an AUC of 0.867 (95% CI: 0.81–0.92) (Fig. 2F).

Feature importance analysis revealed that the top contributors exhibited marked expression differences between MAL and BEN samples (Fig. S2B and Table S6). Notably, cf-ysRNA and cf-tRNA dominated the discriminative signature in the overall DE-cfRNA diagnostic model, aligning with their abundance in the plasma cfRNA. Collectively, these findings underscore the utility of cfRNA-based biomarkers in differentiating malignant from benign breast lesions. The integrated diagnostic model, combining multi-class cfRNA signatures, demonstrates promising clinical potential for non-invasive detection of breast cancer.

### 3.3. Tissue-level prognostic relevance of cfRNA-derived biomarkers in breast cancer

To evaluate whether diagnostic cfRNA markers harbor prognostic relevance, we investigated the association between the top cf-mRNA contributors from our diagnostic model and clinical outcomes in breast cancer. Univariate Cox regression analysis of the TCGA-BRCA dataset ( $n = 1194$ ) identified high expression (FPKM  $> 1$ ) of DLST, DOCK4, and EGFL7 as potential prognostic markers for reduced overall survival (Fig. 3A), with each gene showing a statistically significant hazard ratio: DLST (HR = 1.522, 95% CI: 1.050–2.207,  $p = 0.027$ ), DOCK4 (HR = 1.407, 95% CI: 1.107–1.788,  $p = 0.005$ ), and EGFL7 (HR = 1.196, 95% CI: 1.053–1.360,  $p = 0.006$ ). To evaluate the tissue-level prognostic relevance of the signature, we stratified TCGA-BRCA patients into high-risk ( $n = 560$ ,  $\geq 2$  of the three genes expressed above cohort median) and low-risk ( $n = 634$ ,  $\leq 1$  gene expressed above cohort median) groups. Kaplan-Meier analysis revealed markedly poorer survival in the high-risk cohort (median OS: 8.386 years [95% CI: 7.6–10.8] vs. 17.676 years [95% CI: 10.6–NA], log-rank  $p$ -value = 0.005) with 5-year survival rates of 74.7% vs. 83.0% (Fig. 3B).

To delineate biological mechanisms linking these cfRNA-derived signatures to tumor progression, differential expression analysis (|

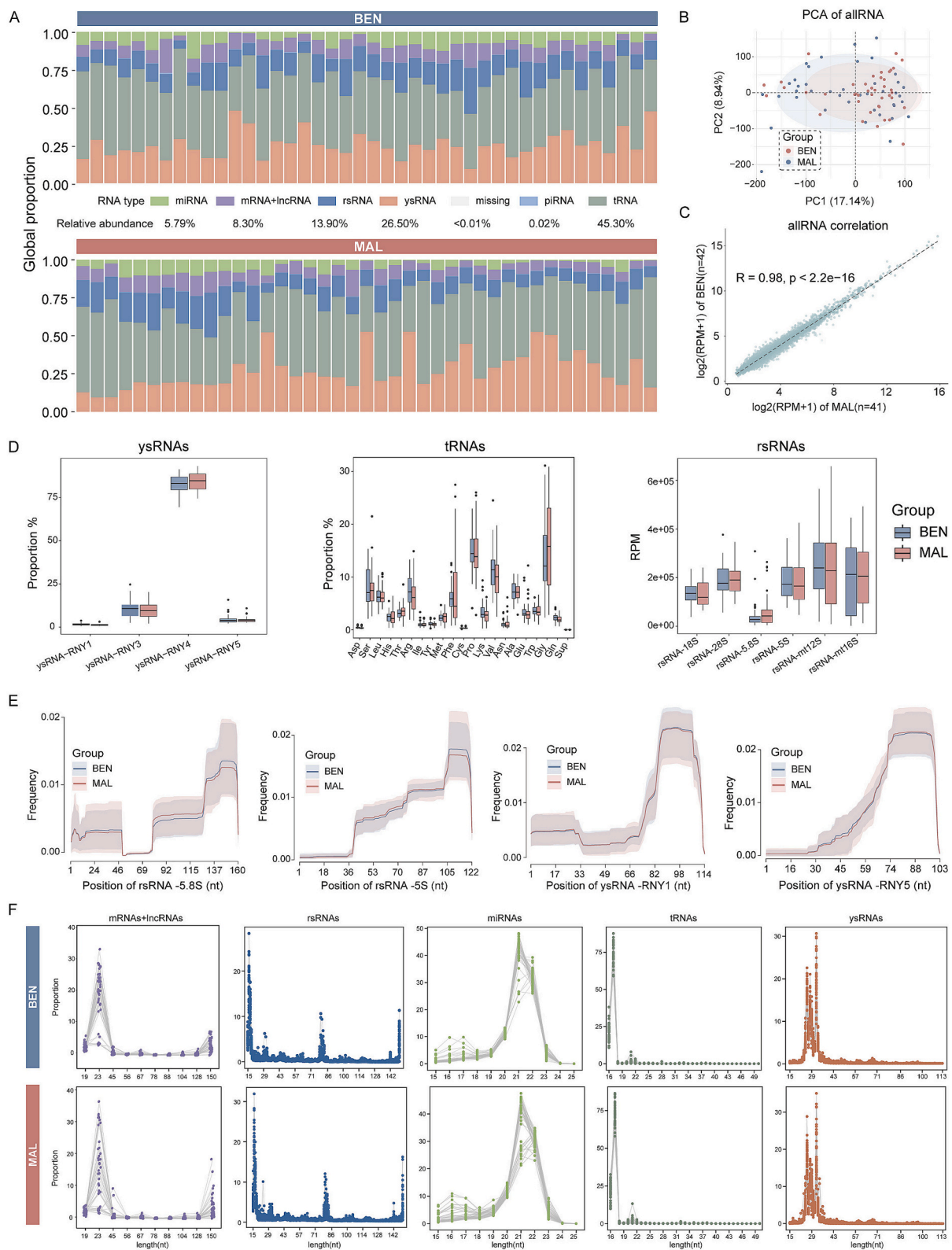


**Fig. 1. Comparative profiling of plasma cfRNAs in MAL and BEN breast lesion cohorts.** **A.** Composition of plasma cfRNA sub-species, including cf-miRNA, cf-mRNA/cf-lncRNA, cf-rsRNA, cf-ysRNA, cf-piRNA, and cf-tRNA, were shown as proportional distributions for BEN ( $n = 42$ ) and MAL ( $n = 41$ ) cohorts. **B.** Principal component analysis (PCA) of the overall cfRNA transcriptome, with PC1 (17.14% variance) and PC2 (8.4% variance) illustrating the clustering of BEN and MAL samples. **C.** Correlation of  $\log_2$ -transformed RPM (reads per million) value between BEN and MAL cohorts; each point represents the expression level of a transcript at the cohort level. **D.** Subtype-specific proportions of cf-ysRNA, cf-tRNA, cf-rsRNA in both BEN and MAL cohorts. **E.** Genome browser visualization of read coverage profiles of cf-rsRNA mapped to human ribosomal RNA (rRNA), and cf-ysRNA mapped to human Y RNA loci. Solid lines denote mean RPM values, shaded regions represent standard deviation (SD). **F.** Density plots showing fragment length distribution for each cfRNA sub-species between groups.

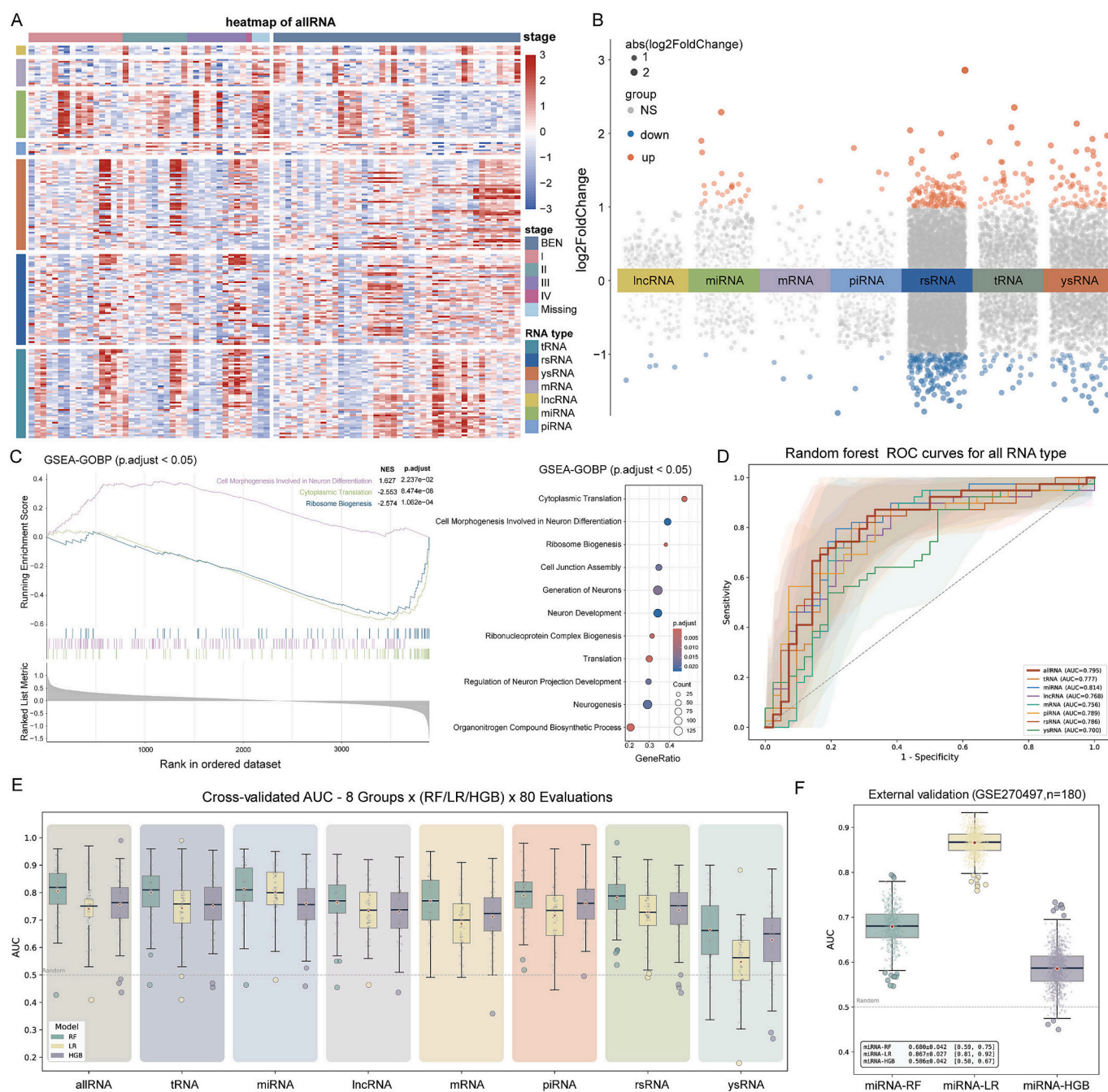


**Fig. 2. Evaluation of plasma cfRNA signature in breast cancer detection.** **A.** Unsupervised clustering heatmap of differentially expressed cfRNAs (DE-cfRNAs) between malignant (MAL,  $n = 41$ ) and benign (BEN,  $n = 42$ ) cohorts. MAL samples are annotated by clinical stage (I–IV). Expression levels are z-score normalized. **B.** Volcano plot of DE-cfRNAs ( $|\log_2FC| \geq 1$ ,  $p$ -value  $< 0.05$ ). Dots colored red (upregulated in MAL) and blue (downregulated in MAL) met significance thresholds; gray dots indicate non-significant (NS) cf-RNAs. **C.** Gene Set Enrichment Analysis (GSEA) of differential cf-mRNAs between MAL and BEN. Left panel: running enrichment score trajectory across ranked gene list (x-axis: rank position, y-axis: enrichment score). Right panel: bubble plot of significant pathways (x-axis: GeneRatio, y-axis: pathway terms; bubble color: adjusted  $p$ -value, size: enriched gene count). **D.** Receiver operating characteristic (ROC) curves of the best-performing models for individual cfRNA sub-species and the integrated allRNAs model (out-of-fold predictions from 80 independent evaluations). The dashed line represents the performance expected by random chance. **E.** Distribution of AUC values across 80 independent cross-validated evaluations for RF, LR, and HGB models. **F.** External validation of the cf-miRNA models in an independent cohort (GSE270497;  $n = 180$ , 120 screening-positive vs 60 negative controls). Boxplots show bootstrap distributions of AUCs (1000 resamples) for the Random Forest (RF), Logistic Regression (LR), and HistGradientBoosting (HGB) models trained on the present cohort. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$|\log_2FC| \geq 1$ ,  $FDR < 0.05$ ) between risk groups was performed in this study (Fig. S2C). Gene Ontology (GO) enrichment of 149 DEGs in high-risk tumors highlighted processes including extracellular matrix organization ( $FDR = 1.421397e-06$ ), angiogenesis ( $FDR = 5.083e-03$ ), and cytokine activity ( $FDR = 0.014$ ) (Fig. 3C). KEGG pathway analysis

further implicated high-risk tumors in ECM-receptor interaction ( $FDR = 0.018$ ), AMPK signaling ( $FDR = 0.018$ ), and PI3K-Akt activation ( $FDR = 0.039$ ) (Fig. 3D).

Collectively, these analyses show that the DLST/DOCK4/EGFL7 signature is associated with extracellular matrix remodeling,

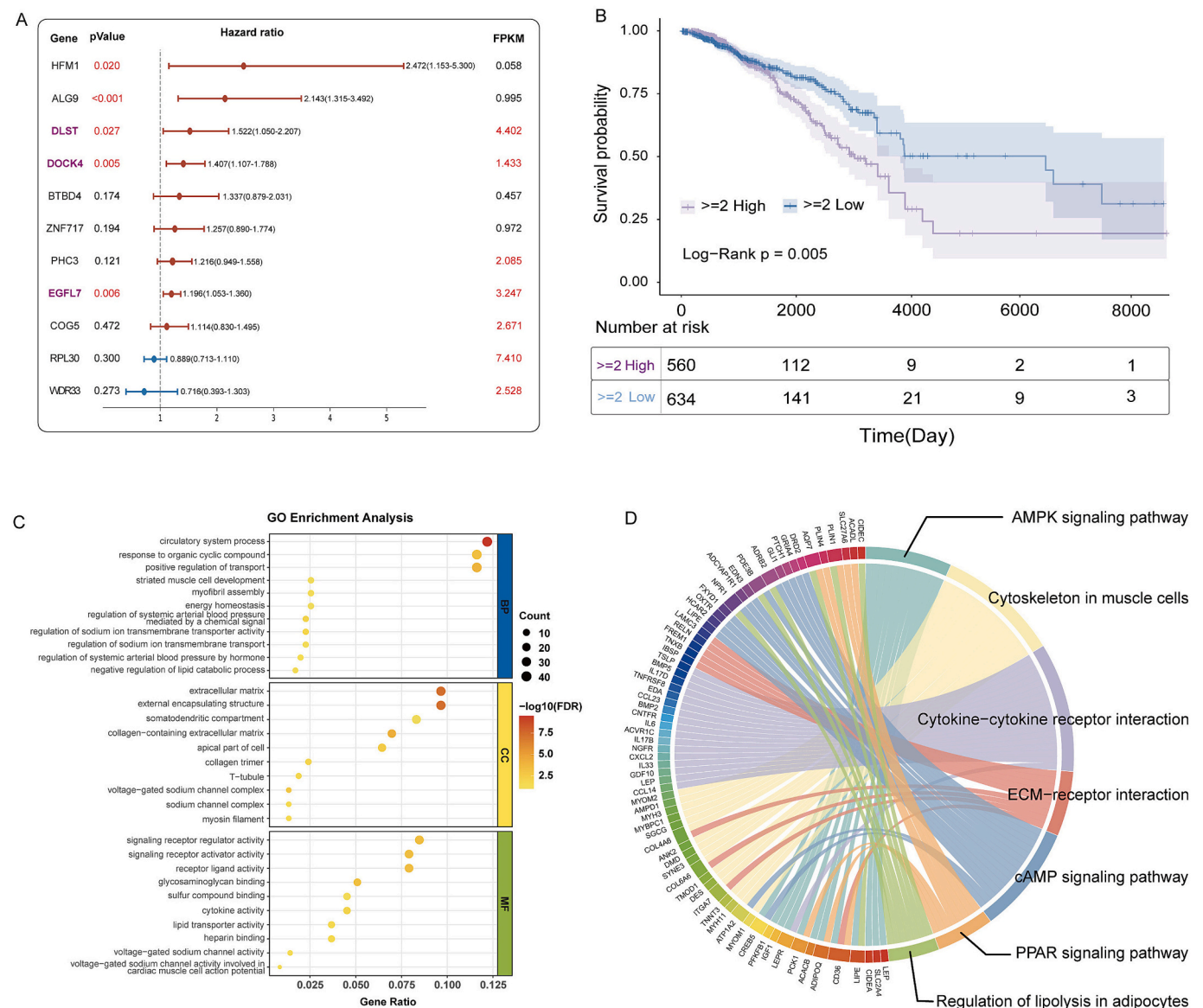


**Fig. 3. Construction and validation of candidate cf-mRNAs as prognostic biomarkers.** **A.** Univariate cox proportional hazards regression analysis of 18 non-redundant cf-mRNAs (top 10 contributors from three cf-mRNA training models) in the TCGA-BRCA dataset. Hazard ratio (HR) with 95% confidence intervals (CI) were shown. Genes with HR > 1 (red) represented mortality risk factors, while HR < 1 (blue) indicated protective factors. Significance was determined by the Wald test. **B.** Kaplan-Meier survival curves stratified by cumulative risk score (high-risk:  $\geq 2$  upregulated risk genes; low-risk:  $\leq 1$  upregulated risk genes). Log-rank test demonstrated significant survival difference ( $p$ -value = 0.005). **C.** Bubble plot showed gene ontology (GO) enrichment analysis of DEGs between risk groups. Top enriched terms span Biological Processes (BP), Cellular Components (CC), and Molecular Functions (MF). **D.** KEGG pathway analysis visualized using a chord plot, showing significant pathway (FDR < 0.05) and associated DEGs. Outer arcs represent pathways, with connecting ribbons indicating linked genes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

angiogenesis- and metabolism-related programs in high-risk tumors.

3.4. Exploratory analysis of the association between risk groups and immune features

As an exploratory analysis, we compared computationally inferred immune infiltration profiles between high- and low-risk patients using bulk tumor tissue transcriptomes from TCGA-BRCA. Although the overall immune landscape across 22 immune cell subsets was broadly similar, 12 immune subsets exhibited significant differences. Notably, the low-risk group showed elevated infiltration of M0 and M1 macrophages (Fig. 4A–B), indicative of enhanced antitumor and anti-infective immune activity. In contrast, the high-risk group was characterized by increased infiltration of resting CD4+ memory T cells, plasma cells, and

activated mast cells, a profile associated with immunosuppression and poor clinical outcomes (Fig. 4B).

Correlation analysis between the cf-mRNA risk signature and immune cell infiltration revealed significant negative correlations with M0 macrophages, M1 macrophages, and follicular helper T cells (Fig. 4C). These computational findings suggest that elevated expression of the risk-associated RNA signatures may be associated with reduced anti-tumor immune activity. Tumor microenvironment (TME) characterization further revealed that the low-risk group had significantly higher tumor purity but lower immune and stromal scores compared to the high-risk group (Fig. 4D).

Collectively, these exploratory analyses suggest potential immunomodulatory roles of the mRNA signature within the TME, providing hypotheses for future functional validation.

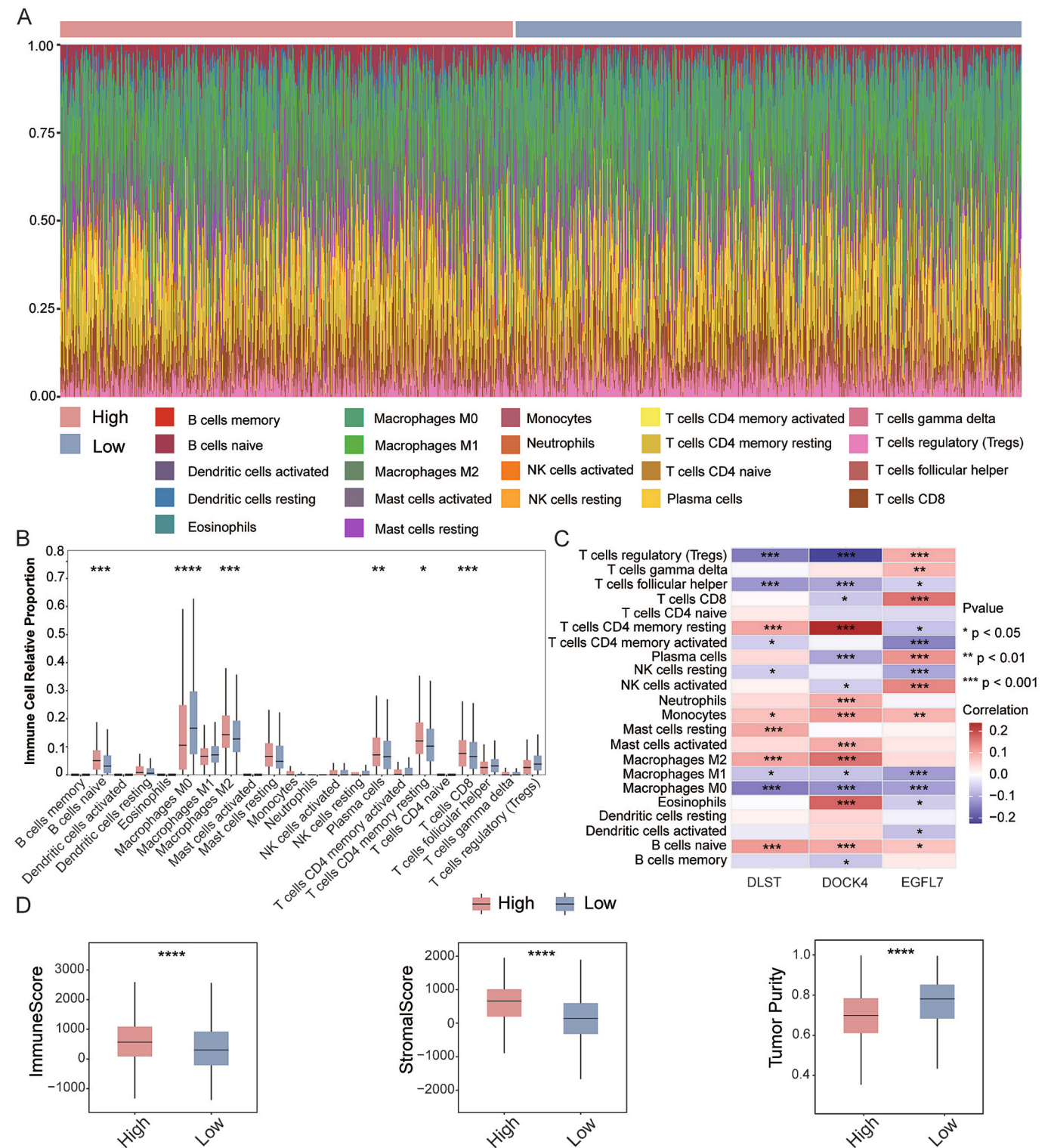


**Fig. 4. Differential immune landscape and tumor microenvironment features stratified by the cf-mRNA risk signature.** A. Immune cell composition of 22 functionally distinct subpopulations in high-risk and low-risk groups. B. Differential abundance of immune cell subtypes between risk groups. Statistical significance was assessed using a two-sided Wilcoxon rank-sum test (\* $p$ -value  $< 0.05$ ; \*\* $p$ -value  $< 0.01$ ; \*\*\* $p$ -value  $< 0.001$ ; \*\*\*\* $p$ -value  $< 0.0001$ ). C. Correlation analysis between expression of risk-associated gene signatures (DLST, DOCK4, EGFL7) and levels of immune cell infiltration. D. Immune score, stromal score, and tumor purity estimated by the ESTIMATE algorithm in the high-risk and low-risk groups.

3.5. Genomic mutation landscape and pathway activation differences between risk groups

To further delineate the genomic determinants underlying breast

cancer progression in distinct risk groups, we analyzed somatic mutation profiles and their functional implications. Comparative analysis revealed significant differences in mutation patterns between high- and low-risk groups, with distinct molecular pathways associated with

disease progression.

Analysis of mutation types demonstrated that missense mutations, predominantly single nucleotide variants (SNVs), were the most frequent alteration in both groups (Fig. S3). Among the top 10 mutated genes, PIK3CA emerged as the most frequently altered oncogene in the high-risk group (42%), followed by TP53 (23%) (Fig. 5A–B). Notably, the high-risk group exhibited a significantly higher prevalence of C > T transitions ( $p$ -value = 0.013), a hallmark of DNA methylation-associated mutagenesis resulting from spontaneous deamination of methylated cytosines (Fig. 5C).

Tumor variant allele frequency (VAF) analysis further distinguished the risk groups, with the high-risk group showing elevated VAFs in key oncogenes (Fig. 5D). Specifically, PIK3CA missense mutations were enriched in the high-risk group (41.94% vs. 28.71% in low-risk;  $p$ -value = 0.027), with clustering in functional domains critical for PI3K signaling, including the kinase and helical domains (Fig. 5E). This domain-specific mutation pattern suggests enhanced activation of PI3K-driven oncogenic pathways.

To validate this hypothesis, we conducted GSEA, which revealed significant activation of the PI3K-AKT signaling pathway in the high-risk group (FDR =  $2.917 \times 10^{-5}$ , Fig. 5F). These findings implicate PIK3CA mutations as potential drivers of pathway dysregulation in high-risk tumors, suggesting their prognostic relevance and potential utility for risk stratification.

### 3.6. Single-cell resolution of prognostic RNA signature expression in the breast cancer microenvironment

As a further exploratory analysis, we examined the expression of the prognostic cf-mRNA signature (DLST, DOCK4, EGFL7) in a single-cell RNA sequencing (scRNA-seq) dataset from an integrated cohort of estrogen receptor (ER)-positive, human epidermal growth factor receptor 2 (HER2)-positive, and triple-negative breast cancer (TNBC) patients (GSE161529,  $n = 26,196$  high-quality cells). Unsupervised clustering resolved six major cell populations: epithelial cells, tumor-associated macrophages (TAMs), B cells, cancer-associated fibroblasts (CAFs), T cells, and T/NK cells (Fig. 6A). Cell identities were confirmed by canonical markers (e.g., CAFs: COL1A1, POSTN; TAMs: FPR3, C1QC; epithelial cells: LY6D, FDCSP).

Expression mapping revealed distinct spatial localization of the prognostic RNA signature within the TME (Fig. 6B–C). DLST was predominantly expressed in epithelial cells and T cells, while DOCK4 localized to TAMs and EGFL7 to CAFs. This cell type-specific expression pattern indicates compartmentalized functions of the three genes within the tumor microenvironment, which warrant further functional characterization.

To characterize functional consequences of dysregulated prognostic RNA signature, we stratified cells into high- and low-expression subgroups (Fig. 6D and Table S7). Differential expression analysis identified 44 upregulated and 6 downregulated genes in the high-expression group (FDR < 0.05;  $|\log_2 \text{FC}| > 1$ ). Pathway enrichment revealed significant activation of ribosome biogenesis and MAPK signaling (Fig. 6E), aligning with prior results of elevated translational activity and MAPK-driven proliferation in MAL patients. Collectively, these exploratory results suggest that the prognostic RNA signature may exert pleiotropic effects across TME compartments, potentially modulating protein synthesis, stromal activation, and proliferative signaling.

Comprehensive subtyping of cancer-associated fibroblasts (CAFs) was performed, with definitive marker genes identified for each subpopulation (Fig. S4A). The CytoTRACE algorithm was employed to predict the differentiation trajectory of CAF subtypes during breast cancer progression (Fig. S4B and S4C). Our analysis suggested that DLST, DOCK4, and EGFL7 may be associated with the phenotypic transition from epithelial-like fibroblasts to tumor endothelial-associated fibroblasts during breast cancer progression. This differentiation pattern was further characterized using Monocle pseudotemporal

analysis (Fig. S4D).

## 4. Discussion

The molecular heterogeneity of breast cancer and the limitations of conventional diagnostic tools underscore the requirement of non-invasive biomarkers. In this study, we established a novel, non-invasive approach for both diagnosing breast cancer and stratifying patients into prognostically distinct subgroups. By profiling plasma cfRNA in 41 MAL and 42 BEN breast lesion patients, we uncovered striking differences in cfRNA composition between groups. Malignant tumors exhibited dysregulation of cfRNA, with tRNA-Phe significantly depleted and tRNA-Gly/Met elevated in MAL cohorts. These subtype-specific cfRNA alterations enabled the development of machine learning models integrating multiple cfRNA sub-species, which achieved exceptional diagnostic accuracy in distinguishing MAL from BEN lesions (mean AUC of 0.795 for the integrated model and 0.814 for the cf-miRNA model, with external validation AUC of 0.867). These results support the potential of cfRNA profiling as a non-invasive tool for breast cancer detection.

Furthermore, we identified potential prognostic markers composed of DLST, DOCK4, and EGFL7, whose intra-tumoral expression levels stratify patients into high- and low-risk groups based on survival outcomes. High-risk tumors were characterized by a higher prevalence of PIK3CA mutations, significant activation of the PI3K-AKT signaling pathway, and an immunosuppressive tumor microenvironment. Single-cell RNA sequencing, as an exploratory analysis, provided additional insights, revealing that these genes are compartmentalized within the tumor microenvironment: DLST predominantly in epithelial cells and T cells, DOCK4 in tumor-associated macrophages, and EGFL7 in cancer-associated fibroblasts. Collectively, these findings highlight the diagnostic utility of cfRNA profiling and the potential prognostic relevance of the cf-mRNA-derived signature in tumor tissue.

The machine learning models in this study integrated multiple cfRNA sub-species. Although direct comparisons across studies are constrained by differences in cohort composition and study design, the external validation performance of our cf-miRNA diagnostic model (AUC = 0.867) remained robust and was comparable to previously reported panels, including a 4-up-regulated miRNA panel (AUC  $\approx$  0.930) and a 5-down-regulated miRNA panel (AUC  $\approx$  0.920) [44], as well as a cf-mRNA panel based on EVs protein and mRNA markers (AUC  $\approx$  0.923) [45]. These findings support the generalizability of cfRNA-based diagnostic signatures. Moreover, the dysregulation of cf-tRNA, cf-ysRNA, and cf-miRNA in MAL cohorts aligns with emerging evidence that non-coding RNAs are actively secreted by tumors via extracellular vesicles, reflecting dynamic transcriptional activity [46,47]. Specifically, the depletion of tRNA-Phe and elevation of tRNA-Gly/Met in malignant plasma may mirror tumor metabolic reprogramming, as tRNA fragmentation is closely linked to gene regulation and protein translation [48,49]. Our findings corroborate the hypothesis that multi-RNA integration captures broader molecular signatures of malignancy, offering a superior diagnostic tool to single-analyte approaches.

GSEA of the differentially expressed cf-mRNAs revealed enrichment of cytoplasmic translation (FDR =  $7.8 \times 10^{-8}$ ), ribosome biogenesis (FDR =  $9.8 \times 10^{-5}$ ) and cell morphogenesis involved in neuron differentiation (FDR = 0.021) in malignant relative to benign samples. Although these results derive from circulating transcripts and must be interpreted with caution, the enrichment of translation- and ribosome-related programs is consistent with the elevated protein synthesis demand of proliferating tumor cells, while the neuron-differentiation morphogenesis term may reflect cytoskeletal gene programs shared between neural morphogenesis and tumor cell remodeling rather than bona fide neuronal differentiation. Whether these pathway perturbations causally contribute to breast tumorigenesis remains to be established in future functional studies.

The potential prognostic markers DLST, DOCK4, and EGFL7 stratify

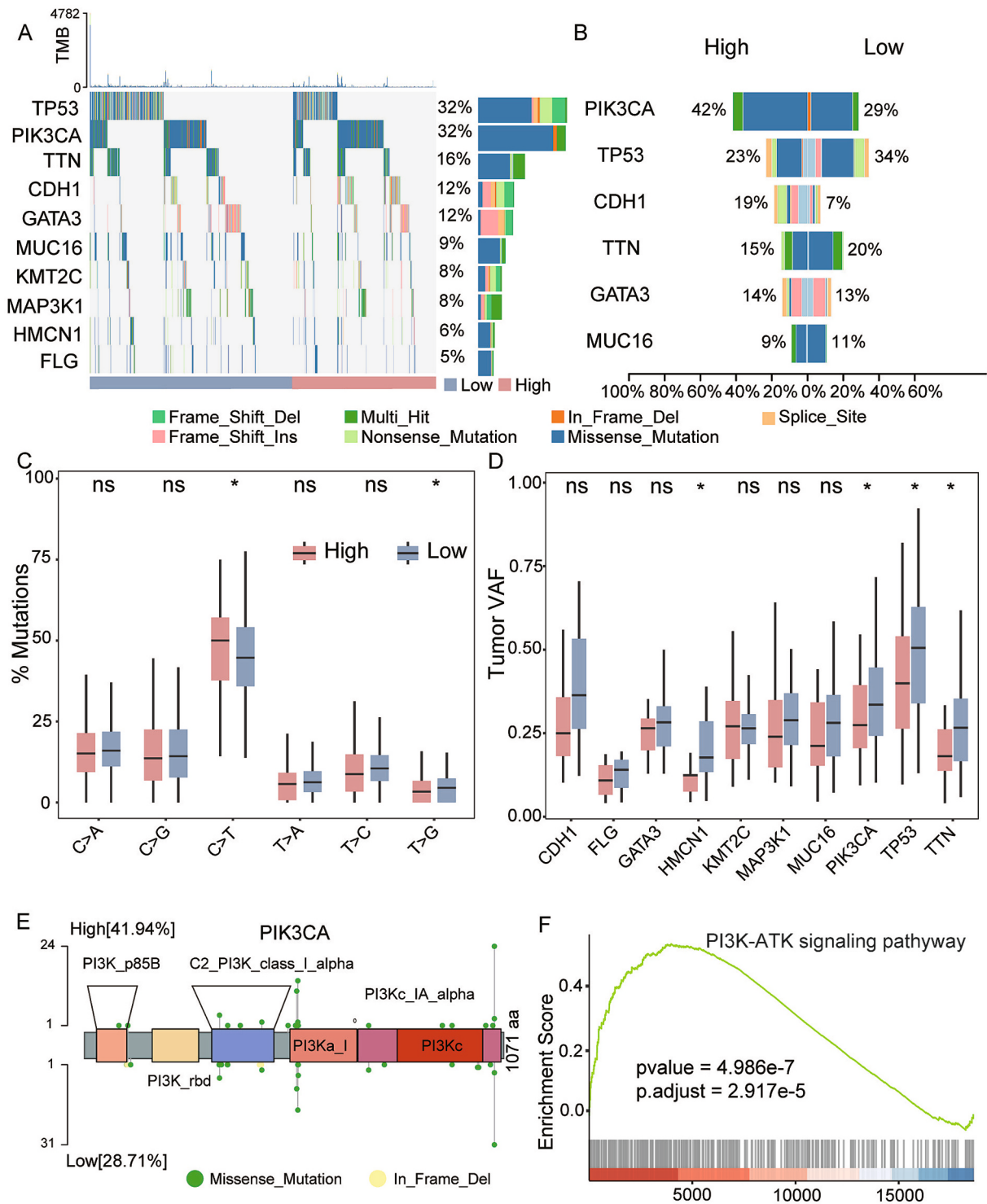


**Fig. 5. Genomic alterations and pathway activation in high- and low-risk groups.** **A.** Waterfall plot illustrates mutation profiles of frequently altered genes (TP53, PIK3CA, etc.), with color-coded mutation types (missense, truncating, etc.) and sample clusters stratified by risk groups. **B.** Mutation frequency comparison of top 10 mutated oncogenes between risk groups. **C.** Proportional distribution of base substitution (e.g., C > A, C > G) between risk groups. The y-axis represents the percentage of mutations. **D.** Tumor variant allele frequency (VAF) distribution for recurrently mutated genes between risk groups. The x-axis represents gene names, and the y-axis denotes tumor VAF. **E.** PIK3CA mutation landscape. Structural domains are annotated, with high-risk tumors harboring a higher proportion of missense mutations compared to low-risk tumors. Green dots denote missense mutations, while yellow dots indicate in-frame deletions. **F.** Enrichment analysis of the PI3K-AKT signaling pathway, with the enrichment score representing pathway enrichment significance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

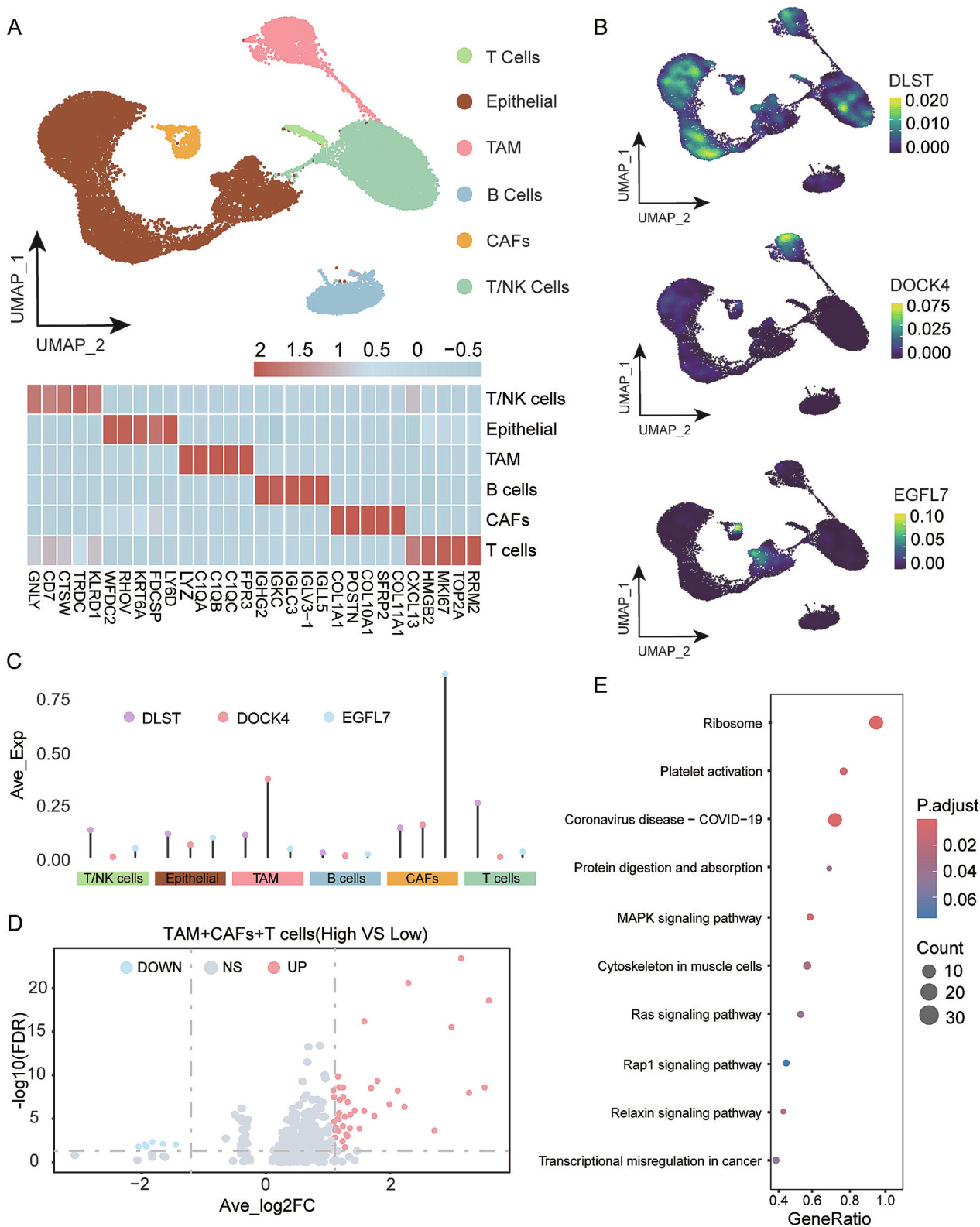


**Fig. 6.** Single-cell transcriptomic characterization of the prognostic RNA signature in breast cancer microenvironment. **A.** Cell clustering and marker gene expression in the scRNA-seq dataset (GSE161529). Top panel: Uniform Manifold Approximation and Projection (UMAP) of scRNA-seq data colored by cell type ( $n = 6$  clusters). Bottom panel: Heatmap of canonical marker genes across clusters. Expression values normalized to Z-score. **B.** UMAP visualization of DLST (yellow), DOCK4 (red), and EGFL7 (blue) expression levels. Color intensity corresponds to log-normalized counts. **C.** Bar plots showing mean expression of DLST, DOCK4, and EGFL7 across cell types. **D.** Volcano plots showing differentially expressed genes (DEGs) between cell populations stratified by high and low expression of the prognostic RNA signature. Genes colored as upregulated (UP:  $\log_2FC > 1.0$ ,  $FDR < 0.05$ ; red), downregulated (DOWN:  $\log_2FC < -1.0$ ,  $FDR < 0.05$ ; blue), or non-significant (NS; gray). **E.** Bubble plot of significantly enriched pathways among DEGs stratified by prognostic RNA signature expression. Dot color corresponds to adjusted  $p$ -values, and dot size reflects the proportion of enriched genes within each pathway. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

breast cancer patients into high- and low-risk groups with significant survival differences in tumor tissue. DLST, a key enzyme in the tricarboxylic acid (TCA) cycle, has been implicated in metabolic adaptation in triple-negative breast cancer (TNBC) [50], while DOCK4, a regulator of Rho GTPases, promotes invasion by modulating cytoskeletal dynamics [51]. EGFL7, an angiogenic factor, is associated with stromal remodeling and poor prognosis in metastatic disease [52,53]. High-risk tumors exhibited a computationally inferred immunosuppressive TME characterized by macrophage polarization (reduced M1, elevated M2-like TAMs) and dysfunctional T-cell activity, consistent with prior studies linking PI3K-AKT activation to immune evasion [54,55]. These findings align with clinical observations that PI3K inhibitors show limited efficacy in tumors with hyperactivated stromal signaling [56,57], emphasizing the need for combinatorial therapies targeting both oncogenic pathways and immune checkpoints. Although descriptive, these pathway associations are consistent with the known functions of DLST (TCA cycle), DOCK4 (cytoskeletal dynamics), and EGFL7 (angiogenesis/stromal remodeling), collectively suggesting that the signature may capture concerted metabolic and stromal reprogramming in aggressive tumors. These links, however, remain to be functionally validated in future studies.

The enrichment of PIK3CA mutations (42.0% in high-risk vs. 28.7% in low-risk) was identified in high-risk tumors. These mutations cluster in functional domains, likely driving constitutive PI3K-AKT pathway activation, which may contribute to the resistance to endocrine therapies [58,59]. cfRNA profiling may offer a non-invasive method to explore such pathway-associated biology, although this application remains to be validated prospectively.

As an exploratory finding, single-cell resolution revealed compartment-specific expression of the prognostic RNA signature. These observations generate hypotheses that require functional validation, for instance, CAF-expressed EGFL7 may contribute to collagen deposition and immune exclusion [60,61], DLST may modulate tumor cell metabolism and T cell function, and DOCK4 may regulate macrophage polarization. These hypotheses align with emerging methodology in precision oncology. Collectively, these exploratory insights lay the foundation for future studies on the translation of cfRNA-based biomarkers into precision oncology.

Despite these advances, several limitations must be acknowledged. First, the study was conducted on a moderate-size cohort ( $n = 83$ ), which may affect the generalizability of the signatures; larger, multi-center cohorts are required to validate the robustness of the cfRNA diagnostic model. Second, independent external validation is currently available only for cf-miRNA (GSE270497,  $n = 180$ ); external cohorts covering the other cfRNA sub-species are not yet available, and cross-platform standardization remains a challenge. Third, the prognostic analysis is indirect: the three-gene signature was derived from plasma cf-mRNA, but its survival associations were evaluated in TCGA tumor tissue; prospective plasma-based validation with longitudinal clinical follow-up is required before any prognostic clinical use can be claimed. In summary, while this proof-of-concept study demonstrates the feasibility of multi-modal cfRNA profiling for non-invasive breast cancer detection and risk stratification, its clinical application depends on further prospective validation.

## Ethics declarations

All procedures were performed in compliance with relevant laws and institutional guidelines and have been approved by the appropriate institutional committee(s). The privacy rights of human subjects were strictly observed, and written informed consent was obtained from all participants prior to their inclusion in the study.

## Availability of data and code

The raw sequencing data (FASTQ format) generated from the plasma

samples in this study have been deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human: HRA007921). Access to these data is restricted and they can be obtained upon request from the corresponding author, in compliance with ethical guidelines. The datasets supporting the conclusions of this article are available in the Github repository, [https://github.com/wangjun-bio/miRlab/tree/cfRNA\\_breast\\_cancer](https://github.com/wangjun-bio/miRlab/tree/cfRNA_breast_cancer). The datasets supporting the conclusions of this article are included within the article and its additional file.

**Table 1.** BEN: benign breast lesions; MAL: malignant breast lesions, T Stage: Primary Tumor Size and Invasion; T1: Tumor  $\leq 2$  cm in diameter; T2: Tumor 2–5 cm in diameter; T3: Tumor  $> 5$  cm in diameter; T4: Tumor invades chest wall or skin (e.g., ulcer, satellite nodules), regardless of size, N Stage: N0: No regional lymph node metastasis; N1: 1–3 axillary lymph node metastases, or micrometastasis in internal mammary nodes (detected microscopically); N2: 4–9 axillary lymph node metastases, or clinically evident metastasis in internal mammary nodes (detected by imaging or palpation); N3:  $\geq 10$  axillary lymph node metastases, or supraclavicular/inframammary lymph node metastasis, or axillary lymph nodes fixed to skin/muscles, M Stage: M0: No distant metastasis; M1: Presence of distant metastasis (any organ), Stage: Staging of tumor progression; Stage I: Early Stage, good prognosis; Stage II: Middle Stage, Tumor progression, requires comprehensive treatment; Stage III: Locally Advanced, may require neoadjuvant therapy before surgery; Stage IV: Advanced, Distant metastasis, systemic therapy as main approach, NA: Missing clinical information; N.A.: No patients with this type.

## CRediT authorship contribution statement

**Jun Wang:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Liu Yang:** Writing – original draft, Data curation. **Kai Fang:** Writing – original draft, Validation, Data curation. **Hongxian Wang:** Supervision, Resources. **Jianxia Chen:** Resources, Methodology. **Binbin Tan:** Writing – review & editing. **Lirong Shu:** Resources, Methodology. **Jinfeng Ye:** Validation, Methodology, Formal analysis, Data curation. **Jiayu Wu:** Validation, Methodology, Investigation, Data curation. **Xinyang Lin:** Validation, Methodology, Formal analysis, Data curation. **Deming Gou:** Supervision, Funding acquisition, Conceptualization. **Yun Wang:** Supervision, Project administration, Funding acquisition, Conceptualization.

## Consent for publication

All authors have approved this manuscript and affirm that it has not been previously published nor is it under consideration by another journal. Written informed consent for publication was obtained from all participants.

## Ethics declaration

### Informed consent and patient details

Written informed consent to take part in the study and to publish the article has been obtained from all participants or their legal representatives. The privacy rights of participants have been observed.

### Human and biological material

This study included organ or tissue donors.

This study includes human biological material and consent was obtained by donors, or their next of kin or legal representatives, for use in this study and for publication of the article. The samples used in this research were not sourced from executed prisoners or prisoners of conscience.

## Studies in Human

This study was performed in compliance with relevant laws, regulatory frameworks and guidelines where the research took place.

This study was approved by the Medical Ethics Committee of Union Shenzhen Hospital, Huazhong University of Science and Technology.

(Approval No. 0720001, ky-2020-039-02)

## Funding

This work was supported by National Natural Science Foundation of China (82470056, 82530003); Shenzhen Medical Research Fund (B2302015 and C2404001); Open Project of Respiratory disease Clinical Medical Center of Yunnan Province (2022LCZXKF-HX03 and 2022LCZXKF-HX04); Shenzhen University Medicine-Engineering Interdisciplinary Research Fund (2024YG001); Shenzhen Natural Science Fund (20231120193941001); Intramural Project A (National Project Initiation Program, YN2024A005).

## 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.

## Acknowledgements

We thank the instrument analysis center of Shenzhen University and the public service platform for large-scale instruments and equipment of the College of Life Sciences and Oceanography for their assistance in instruments and equipment. The results in this study are partly based on data generated from The Cancer Genome Atlas managed by the NCI and NHGRI (<http://cancergenome.nih.gov>).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ygeno.2026.111324>.

## Data availability

Data will be made available on request.

## References

- [1] A.N. Giaquinto, H. Sung, L.A. Newman, R.A. Freedman, R.A. Smith, J. Star, et al., Breast cancer statistics 2024, *CA Cancer J. Clin.* 74 (6) (2024) 477–495, <https://doi.org/10.3322/caac.21863>.
- [2] X. Xiong, L.W. Zheng, Y. Ding, Y.F. Chen, Y.W. Cai, L.P. Wang, et al., Breast cancer: pathogenesis and treatments, *Signal Transduct. Target. Ther.* 10 (1) (2025 Feb 19) 49, <https://doi.org/10.1038/s41392-024-02108-4>.
- [3] S. Lobo-Martins, E. Agostinetto, Azambuja E. de, G. Gebhart, Tumour heterogeneity in molecular imaging for breast cancer, *Ann. Oncol.* 35 (11) (2024 Nov 1) 1061–1062, <https://doi.org/10.1016/j.annonc.2024.07.726>.
- [4] C. Denkert, S. Rachakonda, T. Karn, K. Weber, M. Martin, F. Marmé, et al., Dynamics of molecular heterogeneity in high-risk luminal breast cancer—from intrinsic to adaptive subtyping, *Cancer Cell* 43 (2) (2025 Feb 10) 232–247.e4, <https://doi.org/10.1016/j.ccell.2025.01.002>.
- [5] N.H.Y. Li, C.I. Li, Incidence rate trends of breast cancer overall and by molecular subtype by race and ethnicity and age, *JAMA Netw. Open* 8 (1) (2025 Jan 24) e2456142, <https://doi.org/10.1001/jamanetworkopen.2024.56142>.
- [6] C. Liu, L. Sun, N. Niu, P. Hou, G. Chen, H. Wang, et al., Molecular classification of hormone receptor-positive/HER2-positive breast cancer reveals potential neoadjuvant therapeutic strategies, *Signal Transduct. Target. Ther.* 10 (1) (2025 Mar 26) 1–12, <https://doi.org/10.1038/s41392-025-02181-3>.
- [7] I. Meattini, C.E. Coles, T. Tramm, S. Borghesi, D. Krug, A. Montero, et al., Biomarker-directed radiotherapy in breast cancer: a narrative review, *JAMA Oncol.* 11 (3) (2025 Mar 1) 329–339, <https://doi.org/10.1001/jamaoncol.2024.5780>.
- [8] H.J. Burstein, G. Curigliano, B. Thürlimann, W.P. Weber, P. Poortmans, M. Regan, et al., Customizing local and systemic therapies for women with early breast cancer: the st. Gallen international consensus guidelines for treatment of early breast cancer 2021, *Ann. Oncol.* 32 (10) (2021 Oct) 1216–1235, <https://doi.org/10.1016/j.annonc.2021.06.023>.
- [9] J. Kim, A. Harper, V. McCormack, H. Sung, N. Houssami, E. Morgan, et al., Global patterns and trends in breast cancer incidence and mortality across 185 countries, *Nat. Med.* (2025 Feb 24) 1–9, <https://doi.org/10.1038/s41591-025-03502-3>.
- [10] G. Guglielmi, Breast cancer is on the rise: data reveal drastic gap in survival rates, *Nature* 639 (8053) (2025 Feb 24) 16, <https://doi.org/10.1038/d41586-025-00265-2>.
- [11] Y. Zhang, J. Rodriguez, X. Mao, F. Grassmann, J. Tapia, M. Eriksson, et al., Incidence and risk factors of interval and screen-detected breast cancer, *JAMA Oncol.* (2025 Mar 27), <https://doi.org/10.1001/jamaoncol.2025.0167>.
- [12] N. Pashayan, A.C. Antoniou, U. Ivanus, L.J. Esserman, D.F. Easton, D. French, et al., Personalized early detection and prevention of breast cancer: ENVISION consensus statement, *Nat. Rev. Clin. Oncol.* 17 (11) (2020 Nov) 687–705, <https://doi.org/10.1038/s41571-020-0388-9>.
- [13] S.J. Calleja, E.C. Patiño, Diagnostic accuracy of the CxPREVENTIVE breast model for early detection of breast cancer, *J. Clin. Oncol.* 40 (16 suppl) (2022 Jun) e13578, [https://doi.org/10.1200/JCO.2022.40.16\\_suppl.e13578](https://doi.org/10.1200/JCO.2022.40.16_suppl.e13578).
- [14] S. Sueoka, S. Sasada, E. Suzuki, N. Goda, K. Kajitani, A. Emi, et al., Molecular subtyping of breast cancer by dedicated breast PET, *Ann. Oncol.* 29 (2018 Oct 1) viii65, <https://doi.org/10.1093/annonc/mdy270.198>.
- [15] K.M. Turner, S.K. Yeo, T.M. Holm, E. Shaughnessy, J.L. Guan, Heterogeneity within molecular subtypes of breast cancer, *Am. J. Phys. Cell Phys.* 321 (2) (2021 Aug) C343–C354, <https://doi.org/10.1152/ajpcell.00109.2021>.
- [16] W. Li, J.B. Liu, L.K. Hou, F. Yu, J. Zhang, W. Wu, et al., Liquid biopsy in lung cancer: significance in diagnostics, prediction, and treatment monitoring, *Mol. Cancer* 21 (1) (2022 Jan 20) 25, <https://doi.org/10.1186/s12943-022-01505-z>.
- [17] M. Nikanjam, S. Kato, R. Kurzrock, Liquid biopsy: current technology and clinical applications, *J. Hematol. Oncol.* 15 (1) (2022 Sep 12) 131, <https://doi.org/10.1186/s13045-022-01351-y>.
- [18] S. Ma, M. Zhou, Y. Xu, X. Gu, M. Zou, G. Abudushalamu, et al., Clinical application and detection techniques of liquid biopsy in gastric cancer, *Mol. Cancer* 22 (1) (2023 Jan 11) 7, <https://doi.org/10.1186/s12943-023-01715-z>.
- [19] A. Zviran, R.C. Schulman, M. Shah, S.T.K. Hill, S. Deochand, C.C. Khamnei, et al., Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring, *Nat. Med.* 26 (7) (2020 Jul) 1114–1124, <https://doi.org/10.1038/s41591-020-0915-3>.
- [20] M. Ezeogbu, E. Wilkinson, G. Reid, E.J. Rodger, B. Brockway, T. Russell-Camp, et al., Cell-free DNA methylation in the clinical management of lung cancer, *Trends Mol. Med.* 30 (5) (2024 May 1) 499–515, <https://doi.org/10.1016/j.molmed.2024.03.007>.
- [21] Y. Wada, C. Takasu, T. Yoshimoto, D. Ishikawa, H. Kashiara, T. Tokunaga, et al., A liquid biopsy analysis of exosomal miRNA to detect lymph node metastasis in early gastric cancer, *J. Clin. Oncol.* (2025 Feb), [https://doi.org/10.1200/JCO.2025.43.4\\_suppl.469](https://doi.org/10.1200/JCO.2025.43.4_suppl.469).
- [22] C. Baldini, M. Porte, E. Rouleau, M. Touat, Editorial: liquid biopsy in gliomas – are we there yet? *Ann. Oncol.* (2025 Apr 1) <https://doi.org/10.1016/j.annonc.2025.03.017>.
- [23] Z.X. Liu, G. Chen, Z.L. Yu, Advances in subpopulation separation and detection of extracellular vesicles: for liquid biopsy and downstream research, *Theranostics* 15 (3) (2025 Jan 1) 1135–1155, <https://doi.org/10.7150/thno.106459>.
- [24] Q. Zheng, J. Zhao, H. Yu, H. Zong, X. He, Y. Zhao, et al., Tumor-specific transcripts are frequently expressed in hepatocellular carcinoma with clinical implication and potential function, *Hepatology* (Baltimore, Md.) 71 (1) (2020 Jan) 259–274, <https://doi.org/10.1002/hep.30805>.
- [25] Y. Kim, J.Y. Kim, S. Moon, H. Lee, S. Lee, J.Y. Kim, et al., Tumor-derived EV miRNA signatures surpass total EV miRNA in supplementing mammography for precision breast cancer diagnosis, *Theranostics* 14 (17) (2024) 6587–6604, <https://doi.org/10.7150/thno.99245>.
- [26] M.H. Larson, W. Pan, H.J. Kim, R.E. Mauntz, S.M. Stuart, M. Pimentel, et al., A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection, *Nat. Commun.* 12 (1) (2021 Apr 21) 2357, <https://doi.org/10.1038/s41467-021-22444-1>.
- [27] S. Miao, H. Li, X. Song, Y. Liu, G. Wang, C. Kan, et al., tRNA m1A modification regulates cholesterol biosynthesis to promote antitumor immunity of CD8<sup>+</sup> T cells, *J. Exp. Med.* 222 (3) (2025 Mar 3) e20240559, <https://doi.org/10.1084/jem.20240559>.
- [28] Y. Wang, Y. Zhao, W. Guo, G.S. Yadav, C. Bhaskarla, Z. Wang, et al., Genome-wide gain-of-function screening characterized lncRNA regulators for tumor immune response, *Sci. Adv.* 8 (49) (2022 Dec 9) eadd0005, <https://doi.org/10.1126/sciadv.add0005>.
- [29] J. Wang, J. Huang, Y. Hu, Q. Guo, S. Zhang, J. Tian, et al., Terminal modifications independent cell-free RNA sequencing enables sensitive early cancer detection and classification, *Nat. Commun.* 15 (1) (2024 Jan 2) 156, <https://doi.org/10.1038/s41467-023-44461-y>.
- [30] GENCODE - human release 41 [Internet]. <https://www.encodegenes.org/human/release.41.html> accessed 2025 Aug 25.
- [31] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biol.* 15 (12) (2014) 550, <https://doi.org/10.1186/s13059-014-0550-8>.
- [32] R. Kolde, pheatmap: pretty heatmaps [Internet], 2019. <https://cran.r-project.org/web/packages/pheatmap/index.html>.
- [33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in python, *J. Mach. Learn. Res.* 12 (2011) 2825–2830.

- [34] M.L. Waskom, seaborn: statistical data visualization, *J. Open Source Softw.* 6 (60) (2021) 3021, <https://doi.org/10.21105/joss.03021>.
- [35] J.D. Hunter, Matplotlib: a 2D graphics environment, *Comput Sci Eng.* 9 (3) (2007) 90–95, <https://doi.org/10.1109/MCSE.2007.55>.
- [36] GDC data portal homepage [Internet]. <https://portal.gdc.cancer.gov/> accessed 2025 Aug 25.
- [37] CIBERSORTx [Internet]. <https://cibersortx.stanford.edu/> accessed 2025 Aug 25.
- [38] A. Mayakonda, D.C. Lin, Y. Assenov, C. Plass, H.P. Koeffler, Maftools: efficient and comprehensive analysis of somatic variants in cancer, *Genome Res.* 28 (11) (2018 Nov) 1747–1756, <https://doi.org/10.1101/gr.239244.118>.
- [39] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, et al., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, *Nat. Biotechnol.* 32 (4) (2014 Apr) 381–386, <https://doi.org/10.1038/nbt.2859>.
- [40] M. Kang, J.J.A. Armenteros, G.S. Gulati, R. Gleyzer, S. Avagyan, E.L. Brown, et al., Mapping single-cell developmental potential in health and disease with interpretable deep learning, *Biorxiv Prepr. Serv. Biol.* (2024 Mar 21) 2024.3.19.585637, <https://doi.org/10.1101/2024.03.19.585637>.
- [41] S. Jin, M.V. Plikus, Q. Nie, CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics, *Nat. Protoc.* 20 (1) (2025 Jan) 180–219, <https://doi.org/10.1038/s41596-024-01045-4>.
- [42] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci.* 102 (43) (2005 Oct 25) 15545–15550, <https://doi.org/10.1073/pnas.0506580102>.
- [43] V.K. Mootha, C.M. Lindgren, K.F. Eriksson, A. Subramanian, S. Sihag, J. Lehara, et al., PGC-1 $\alpha$ -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nat. Genet.* 34 (3) (2003 Jul) 267–273, <https://doi.org/10.1038/ng1180>.
- [44] M. Samara, E. Thodou, M. Patoulioti, A. Poultisidi, G.E. Thomopoulou, A. Giakountis, Integrated miRNA signatures: advancing breast cancer diagnosis and prognosis, *Biomolecules* 14 (11) (2024 Oct 24) 1352, <https://doi.org/10.3390/biom14111352>.
- [45] H. Lin, B. Li, J. Guo, X. Mai, H. Yu, W. Pan, et al., Simultaneous detection of membrane protein and mRNA at single extracellular vesicle level by droplet microfluidics for cancer diagnosis, *J. Adv. Res.* 24 (2024 Aug 26), <https://doi.org/10.1016/j.jare.2024.08.026>, 369–2. S2090-1232.
- [46] L. Zhu, J. Li, Y. Gong, Q. Wu, S. Tan, D. Sun, et al., Exosomal tRNA-derived small RNA as a promising biomarker for cancer diagnosis, *Mol. Cancer* 18 (1) (2019 Apr 2) 74, <https://doi.org/10.1186/s12943-019-1000-8>.
- [47] W. Gu, J. Shi, H. Liu, X. Zhang, J.J. Zhou, M. Li, et al., Peripheral blood non-canonical small non-coding RNAs as novel biomarkers in lung cancer, *Mol. Cancer* 19 (1) (2020 Nov 12) 159, <https://doi.org/10.1186/s12943-020-01280-9>.
- [48] M. Duechler, G. Leszczynska, E. Sochacka, B. Nawrot, Nucleoside modifications in the regulation of gene expression: focus on tRNA, *Cell Mol Life Sci CMLS.* 73 (16) (2016 Aug) 3075–3095, <https://doi.org/10.1007/s00018-016-2217-y>.
- [49] M.A. Kay, The role of tRNA derived small RNAs in gene regulation in normal tissues and cancer, *FASEB J.* 34 (S1) (2020) 1, <https://doi.org/10.1096/fasebj.2020.34.s1.00158>.
- [50] N. Shen, S. Korm, T. Karantanos, D. Li, X. Zhang, E. Ritou, et al., DLST-dependence dictates metabolic heterogeneity in TCA-cycle usage among triple-negative breast cancer, *Commun. Biol.* 4 (1) (2021 Nov 16) 1289, <https://doi.org/10.1038/s42003-021-02805-8>.
- [51] N. Hiramoto-Yamaki, S. Takeuchi, S. Ueda, K. Harada, S. Fujimoto, M. Negishi, et al., Ephxin4 and EphA2 mediate cell migration through a RhoG-dependent mechanism, *J. Cell Biol.* 190 (3) (2010 Aug 9) 461–477, <https://doi.org/10.1083/jcb.201005141>.
- [52] G. Hong, V. Kuek, J. Shi, L. Zhou, X. Han, W. He, et al., EGFL7: master regulator of cancer pathogenesis, angiogenesis and an emerging mediator of bone homeostasis, *J. Cell. Physiol.* 233 (11) (2018 Nov) 8526–8537, <https://doi.org/10.1002/jcp.26792>.
- [53] S. Delfortrie, S. Pinte, V. Mattot, C. Samson, G. Villain, B. Caetano, et al., Egfl7 promotes tumor escape from immunity by repressing endothelial cell activation, *Cancer Res.* 71 (23) (2011 Dec 1) 7176–7186, <https://doi.org/10.1158/0008-5472.CAN-11-1301>.
- [54] T. Shi, Y. Zhang, Y. Wang, X. Song, H. Wang, X. Zhou, et al., DKK1 promotes tumor immune evasion and impedes anti-PD-1 treatment by inducing immunosuppressive macrophages in gastric cancer, *Cancer Immunol. Res.* 10 (12) (2022 Dec 2) 1506–1524, <https://doi.org/10.1158/2326-6066.CIR-22-0218>.
- [55] X. Zheng, J. Chen, M. Deng, K. Ning, Y. Peng, Z. Liu, et al., G3BP1 and SLU7 jointly promote immune evasion by downregulating MHC-I via PI3K/akt activation in bladder cancer, *Adv. Sci. Weinh. Baden-Wurt. Ger.* 11 (7) (2024 Feb) e2305922, <https://doi.org/10.1002/adv.202305922>.
- [56] K. Okkenhaug, M. Graupera, B. Vanhaesebroeck, Targeting PI3K in cancer: impact on tumor cells, their protective stroma, angiogenesis, and immunotherapy, *Cancer Discov.* 6 (10) (2016 Oct) 1090–1105, <https://doi.org/10.1158/2159-8290.CD-16-0716>.
- [57] H. Zhang, L. Zhang, Y. He, D. Jiang, J. Sun, Q. Luo, et al., PI3K PROTAC overcomes the lapatinib resistance in PIK3CA-mutant HER2 positive breast cancer, *Cancer Lett.* 598 (2024 Aug 28) 217112, <https://doi.org/10.1016/j.canlet.2024.217112>.
- [58] O. Stål, G. Pérez-Tenorio, L. Akerberg, B. Olsson, B. Nordenskjöld, L. Skoog, et al., Akt kinases in breast cancer and the results of adjuvant therapy, *Breast Cancer Res.* 5 (2) (2003), <https://doi.org/10.1186/bcr569>. R37–44.
- [59] M. Bullock, FOXO factors and breast cancer: outfoxing endocrine resistance, *Endocr. Relat. Cancer* 23 (2) (2016 Feb 1) R113–R130, <https://doi.org/10.1530/ERC-15-0461>.
- [60] C. Fabian, S. Mahajan, M.H.H. Schmidt, EGFL7: an emerging biomarker with great therapeutic potential, *Pharmacol. Ther.* 266 (2025 Feb) 108764, <https://doi.org/10.1016/j.pharmthera.2024.108764>.
- [61] D. Nichol, H. Stuhlmann, EGFL7: a unique angiogenic signaling factor in vascular development and disease, *Blood* 119 (6) (2012 Feb 9) 1345–1352, <https://doi.org/10.1182/blood-2011-10-322446>.