BMC Biology

RESEARCH

Open Access

Multimodal analysis of cell-free DNA enhances differentiation of early-stage breast cancer from benign lesions and healthy individuals

Thi Tuong Vi Van[1], Trung Hieu Tran[1], Thi Hue Hanh Nguyen[1], Van Thien Chi Nguyen[1], Dac Ho Vo[1], Giang Thi Huong Nguyen[1], Trong Hieu Nguyen[1], Kim Sang To[2], Anh Luan Nguyen[2], Cao Hong An Tran[2], Thanh Xuan Jasmine[3], Thi Loan Vo[3], Thi Huong Thoang Nai[3], Thuy Trang Tran[3], My Hoang Truong[3], Ngan Chau Tran[3], Thi Loc Le[4], Thi Hong Nhung Nguyen[4], Ngoc Hieu Tu[5], Thanh Son Tran[5], Bao Toan Le[6], Van Phong Tang[6], Pham Thanh Nhan Nguyen[7], Khac Tien Nguyen[8], Van Chien Ho[8], Xuan Vinh Nguyen[1], Nhu Nhat Tan Doan[1], Thi Trang Tran[1], Thi Minh Thu Tran[1], Vu Uyen Tran[1], Minh Phong Le[1], Thi Luyen Vu[1], Ba Linh Tieu[1], Huu Tam Phuc Nguyen[1], Luu Hong Dang Nguyen[1], Ngoc Minh Phan[1], Thi Van Phan[1], Thi Thanh Thuy Do[1], Thi Huyen Dao[1], Hung Sang Tang[1], Duy Sinh Nguyen[1], Hoa Giang[1], Minh Duy Phan[1], Hoai-Nghia Nguyen[1], Duc Hieu Vo[2*] and Le Son Tran[1*]

Abstract

Background Breast cancer (BC) remains the second leading cause of cancer-related mortality among women worldwide. Liquid biopsy based on circulating tumor DNA (ctDNA) offers a promising noninvasive approach for early detection; however, differentiating malignant tumors from benign abnormalities remains a significant challenge.

Results Here, we developed a multimodal approach to analyze cfDNA methylation and fragmentomic patterns in 273 BC patients, 108 individuals with benign breast conditions, and 134 healthy controls. Genome-wide analyses revealed distinct cfDNA copy number alterations and cytosine-enriched cleavage sites in BC patients. Targeted sequencing further revealed unique methylation patterns, including hypermethylation in *GPR126*, *KLF3*, and *TLR10* and hypomethylation in *TOP1* and *MAFB*. Our machine-learning model achieved an AUC of 0.90, with 93.6% specificity and 62.1–66.3% sensitivity for stage I–II cancers. In symptomatic populations, sensitivities were 50.0%, 68.2%, and 64.7% for BI-RADS categories 3, 4, and 5, respectively, with 96.1% specificity.

Conclusions These findings underscore the potential of cfDNA biomarkers to enhance BC detection and reduce the rate of unnecessary biopsies.

Keywords Breast cancer, CfDNA, Benign abnormalities, Methylation and fragmentomic

*Correspondence:
Duc Hieu Vo
hieuvobvub@gmail.com
Le Son Tran
leson1808@gmail.com
Full list of author information is available at the end of the article

Background

Breast cancer (BC) is the second most frequent cancer globally, accounting for 11.6% of all cancer cases in 2022 [1]. It is the most common malignancy among women, with about 2.3 million new cases and 660,000 deaths reported in the same year [1]. Projections indicate that by 2040, the global burden of BC will surpass 3 million new cases and 1 million deaths annually [2]. Early detection is critical for improving prognosis, survival rates, and overall patient outcomes by facilitating timely clinical interventions [3]. Therefore, advancing and optimizing effective screening and diagnostic techniques for BC is essential to improving survival outcomes.

Mammography and ultrasonography are the most widely employed methods for BC detection, with mammography recognized as the gold standard for early detection [4]. Despite its proven effectiveness, mammography has several limitations. These include reduced tumor visibility in dense fibroglandular tissue and the risks associated with repeated radiation exposure [5, 6]. Additionally, the sensitivity of mammography is approximately 86.9%, leaving certain BC types undetected [7]. Some breast tumors, such as benign lumps, integrate into the natural breast architecture, further complicating detection. In young women and those with dense breast tissue, mammography often yields a high false-positive rate, leading to unnecessary follow-up procedures, including biopsies, as recommended by clinical guidelines [8]. Among benign breast conditions, fibroadenomas are the most common, followed by fibrocystic disease [9]. Fibroadenomas present imaging characteristics that overlap with a variety of other benign and malignant breast lesions, contributing to their significant representation among breast biopsies [10].

Magnetic resonance imaging (MRI) has been proposed as a promising alternative to mammography, particularly because it eliminates the risks of radiation exposure [11]. However, its application in clinical practice is limited by factors such as low specificity, complex image interpretation, and the absence of standardized diagnostic guidelines. As a result, breast MRI is generally reserved for women at high risk of BC [12]. The Breast Imaging Reporting and Data System (BI-RADS) categories are frequently utilized alongside imaging modalities to evaluate malignancy risk. These categories provide a spectrum of risk, ranging from >0 to $\leq 2\%$ for BI-RADS 3, >2 to $<95\%$ for BI-RADS 4, and $\geq 95\%$ for BI-RADS 5 [13]. Despite their utility, distinguishing benign breast lesions from malignant tumors remains a significant challenge in clinical practice, underscoring the need for improved diagnostic tools and techniques.

The limitations of current BC diagnostic methods underscore the need for non-invasive approaches capable of accurately diagnosing BC and differentiating between malignant and benign tumors. In this context, liquid biopsy has emerged as a promising diagnostic method, offering higher sensitivity and specificity for early BC detection [14]. Liquid biopsy relies on the analysis of epigenetic and genetic alterations in circulating tumor cells (CTCs), circulating tumor DNA (ctDNA), cell-free DNA (cfDNA), mRNAs, microRNAs (miRNAs), and proteins, which can be detected in peripheral blood or other body fluids using advanced technologies [15]. Several studies have demonstrated the potential of ctDNA as a biomarker for identifying specific mutations associated with BC. However, the sensitivity of mutation-based ctDNA detection methods remains limited due to the lack of universally prevalent mutations in BC [16]. Recent advancements in DNA methylation analysis within liquid biopsies offer a promising alternative. Methylation markers are particularly advantageous due to their stability, relevance across diverse cancer types, and ability to provide comprehensive insights into cancer biology, including early detection [17, 18]. Despite these advantages, most methylation-based studies have not included benign lesions, and the overall performance in BC detection remains suboptimal. High rates of false positives and false negatives persist, largely due to benign conditions that can release DNA with methylation or mutation patterns resembling those of cancer, thereby reducing specificity [19, 20].

To address these challenges, we previously developed the SPOT-MAS (Screening for the Presence of Tumor by DNA Methylation and Size) multimodal assay to analyze genome-wide cfDNA methylomics and fragmentomics, with the goal of enhancing BC detection rates [21, 22]. However, our previous study did not fully address the differentiation of BC from benign lesions, which limited its applicability for general population screening, where patients with benign lesions are common [21, 22].

In the current study, we expanded the feature space by incorporating previously unexplored features, including specific 21-mer end motifs around cfDNA cleavage sites, as well as methylation patterns and copy number aberrations across 450 targeted genomic regions. Importantly, we conducted pairwise comparisons between 273 non-metastatic BC patients and 108 patients with benign lesions, as well as 134 healthy individuals, to identify multiple BC-specific signatures. These signatures were then used to build a classifier model to improve the differentiation of early-stage BC from both healthy individuals and patients with benign lesions (Fig. 1). To further demonstrate the utility of our model, we assessed its performance in detecting BC using an external validation cohort of patients diagnosed with BI-RADS 3–5 lesions through standard-of-care (SOC) imaging tests (Fig. 1).

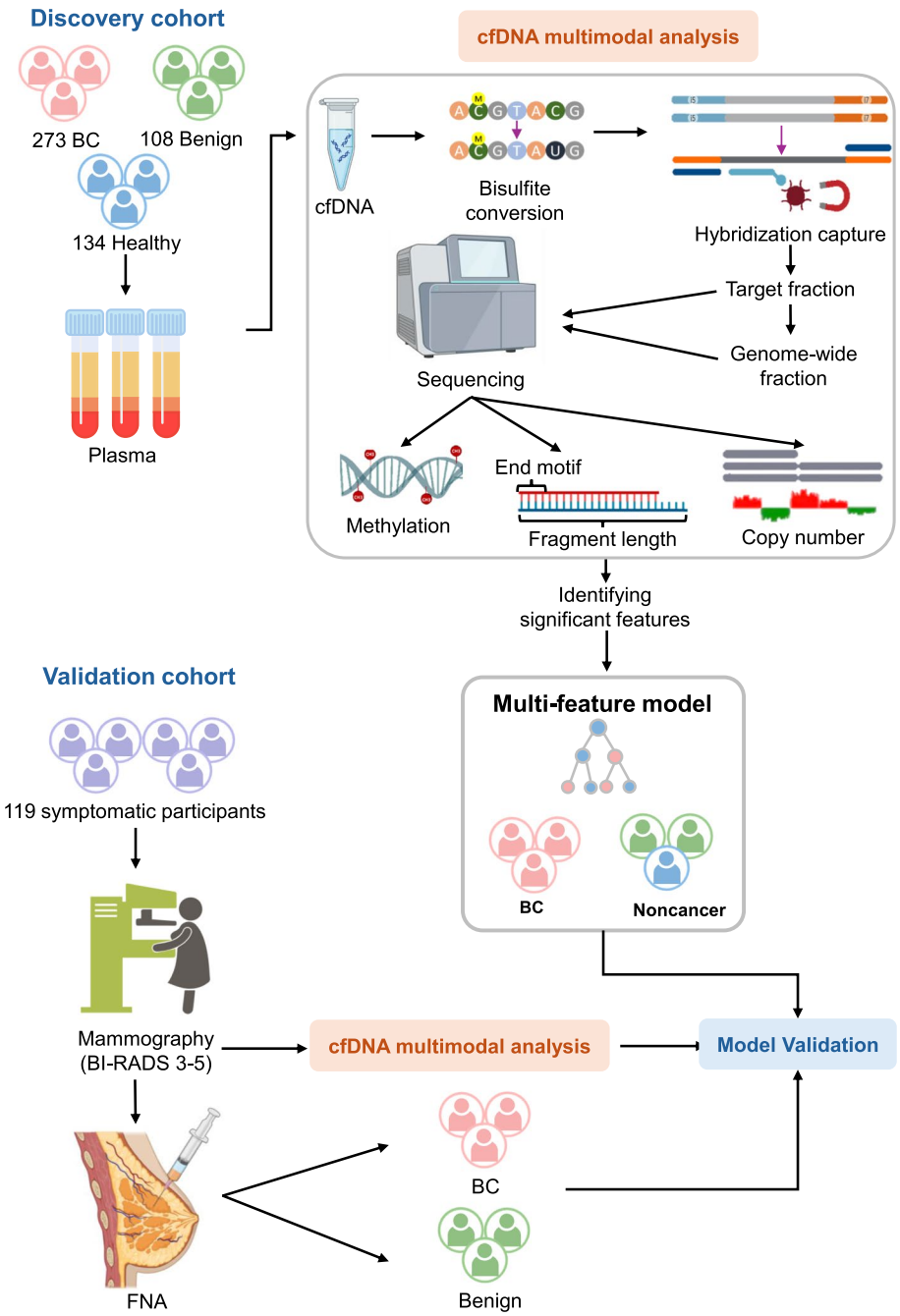

Fig. 1 Overview of study design. The plasma samples in the discovery cohort underwent the SPOT-MAS assay. Features of cfDNA (Methylation, End motif, Fragment length, Copy number) were collected from both the target and the genome-wide fraction after sequencing. A multi-feature model was developed using these features to distinguish BC patients from non-cancer (benign and healthy) individuals. To validate this model, we recruited 119 symptomatic participants who had been diagnosed with BI-RADS 3-5 by mammography. The participants in the validation cohort also underwent the SPOT-MAS assay, which was similar to those in the discovery cohort and confirmed by fine needle aspiration (FNA) afterward

Results

Clinical characteristics of BC patients, individuals with benign breast disease and healthy participants in the discovery cohort

We recruited 273 BC patients, 108 patients with benign breast lumps, and 134 healthy subjects as a discovery cohort to profile cfDNA signatures capable of differentiating BC patients from non-cancer individuals (Table 1). All BC patients were diagnosed following abnormal imaging results (mammography or ultrasound), with

Table 1 Summary of the clinical features of participants in discovery cohort

Clinical Information		Discovery cohort (BC: 273; Benign: 108; Healthy: 134)									
		Train set					Test set				
		BC	Noncancer		*p*-value(Mann-Whitney U test)		BC	Noncancer		*p*-value(Mann-Whitney U test)	
		BC*N*=143	Benign*N*=52	Healthy*N*=65			BC*N*=130	Benign*N*=56	Healthy*N*=69		
		N(%)	N(%)	N(%)	BC vs. Benign	BC vs. Healthy	N(%)	N(%)	N(%)	BC vs. Benign	BC vs. Healthy
Age	Median	49.0	39.5	49.0	4.47E-06	1.00E+00	53.5	41.5	49.0	5.69E-05	7.95E-02
	Min	28.0	22.0	25.0			24.0	20.0	22.0		
	Max	75.0	68.0	67.0			77.0	70.0	70.0		
Stage	I, IA, IB	34(23.8)					29 (22.3)				
	II, IIA, IIB	77(53.8)					83 (63.8)				
	III, IIIA	21(14.7)					15 (11.5)				
	NA	11(7.7)					3 (2.3)				
Subtypes	Luminal A	27(18.9)					22 (16.9)				
	Luminal B	43(30.1)					46 (35.4)				
	Luminal B-HER2	27(18.9)					25 (19.2)				
	HER2	22(15.4)					21 (16.2)				
	TNBC	14(9.8)					11(8.5)				
	NA	10(7.0)					5(3.8)				
Subtypes_Benign	Fibroadenomas		36(69.2)					38(67.9)			
	Fibrocystic changes		16(30.8)					18(32.1)			

malignancy confirmed through subsequent tissue biopsy. Patients in the benign group were confirmed to have non-malignant breast conditions, such as fibroadenomas or fibrocystic changes. Both benign patients and healthy participants underwent mammography as part of their annual health check-ups and were monitored for 12 months to confirm their cancer-free status. The clinical characteristics of participants in the BC, benign, and control groups are summarized in Table 1 and detailed in Additional file 1:Table S1.

Participants in the discovery cohort were randomly divided into a training set and a test set (Table 1). The training set included 143 BC patients, 52 benign patients, and 65 healthy individuals. In this set, both BC patients and healthy participants had a median age of 49 years (p-value = 1.00, Mann–Whitney U test). By contrast, the benign group had a median age of 39.5 years (range 22–68), which was significantly younger than the BC group (p-value ≤ 0.05, Mann–Whitney U test). Most BC patients in the training set were diagnosed at early stages, with 23.8% at stage I, 53.8% at stage II, and 14.7% at stage IIIA. Staging information was unavailable for 7.7% of patients, though all were confirmed by specialists to have non-metastatic tumors. Regarding BC subtypes, 18.9% were luminal A, 30.1% luminal B, 18.9% luminal B-HER2, 15.4% HER2, and 9.8% triple-negative breast cancer (TNBC). In the benign group, 69.2% were diagnosed with fibroadenomas and 30.8% with fibrocystic changes.

The test set consisted of 130 BC patients, 56 benign patients, and 69 healthy individuals (Table 1). The median ages of participants in the BC, benign, and healthy groups in the test set were comparable to those in the training set. Similar to the training set, most BC patients in the test set presented with early-stage tumors, with 22.3% at stage I and 63.8% at stage II. The distributions of BC subtypes and benign conditions in the test set were consistent with those in the training set (Table 1).

Shallow genome-wide sequencing reveals distinct CNA and motif end signatures in cfDNA of BC patients compared to benign and healthy individuals

Our prior research using shallow genome-wide sequencing revealed distinct plasma CNA and EM signatures in BC patients compared to healthy individuals [21]. However, a comparative analysis between BC and benign patients has not been previously conducted. To identify ctDNA signatures capable of distinguishing BC from both healthy and benign individuals, we examined overlapping significant CNA and EM features across two pairwise comparisons: BC versus benign and BC versus healthy.

To investigate the CNA patterns in cfDNA sequencing reads from the whole-genome fraction were aligned to the human reference genome, segmented into 2691 bins of 1 Mb each. The DNA copy number was calculated for each bin. In BC patients compared to benign patients, we identified 863 bins with a significant gain and 592 bins with a significant loss across 22 chromosomes in BC patients (p-value ≤ 0.05, Mann–Whitney U test, Benjamini–Hochberg correction, Fig. 2A, C). When comparing CNA values between BC patients and healthy individuals, we identified only 5 bins with a significant gain and 4 bins with a significant loss in BC patients (p-value ≤ 0.05, Mann–Whitney U test, Benjamini–Hochberg correction, Fig. 2B, D). Notably, 5 bins overlapped between the two pairwise comparisons (Fig. 2E). Among these, two bins with a significant gain were located on chromosomes 7 and 16 (Fig. 2F), while three bins with a significant loss were observed on chromosomes 4, 5, and 12 in BC patients (Fig. 2G).

We next compared the frequencies of all 256 possible 4-mer EMs in cfDNA fragments from BC, benign, and healthy individuals. We identified 18 EMs that were differentially enriched in BC cfDNA compared to benign or healthy individuals (Fig. 3A). Of these, 15 EMs displayed consistent enrichment patterns across both pairwise comparisons (p-value ≤ 0.05, Mann–Whitney U test, Benjamini–Hochberg correction, Fig. 3A). Among the differentially enriched EMs, 8 (highlighted in red, Fig. 3A) were more frequently observed in BC patients than in benign or healthy individuals, while 7 (highlighted in blue, Fig. 3A) were less frequently observed in BC patients compared to these groups. Interestingly, all EMs significantly enriched in BC tended to start with cytosine (C), including CGCC, CCCC, CGCT, CGAC, CGCA, CCCA, CCAC, and CCAG (Fig. 3A, B). In contrast, 5 out of 7 EMs with reduced frequencies in BC began with guanine (G), including GTGT, GTTA, GTTC, GTTG, and GTTT, while the remaining 2 EMs began with cytosine (C) and thymine (T): CTTG and TTTG (Fig. 3A, C).

Beyond the 4-mer EMs, a previous study [23] demonstrated that nucleotide motifs in regions 10 bp downstream and upstream of cfDNA cleavage sites are consistently conserved in healthy controls, while tumor-derived cfDNA fragments exhibit aberrant changes in these motifs. Building on this foundation, we investigated the nucleotide composition (A/T/G/C) within a ± 10 bp interval surrounding the 5′ end of each cfDNA fragment (a total of 21 positions, with the fragment start site indexed as position 0) from BC, benign, and healthy subjects. To distinguish this feature from the 4-mer EM, we designated it as Motif 21 (ME21). Our analysis identified 8 significant ME21s shared by both pairwise comparisons (p-value ≤ 0.05, Mann–Whitney U test, Benjamini–Hochberg correction, Fig. 3D). Among these, 2 ME21s

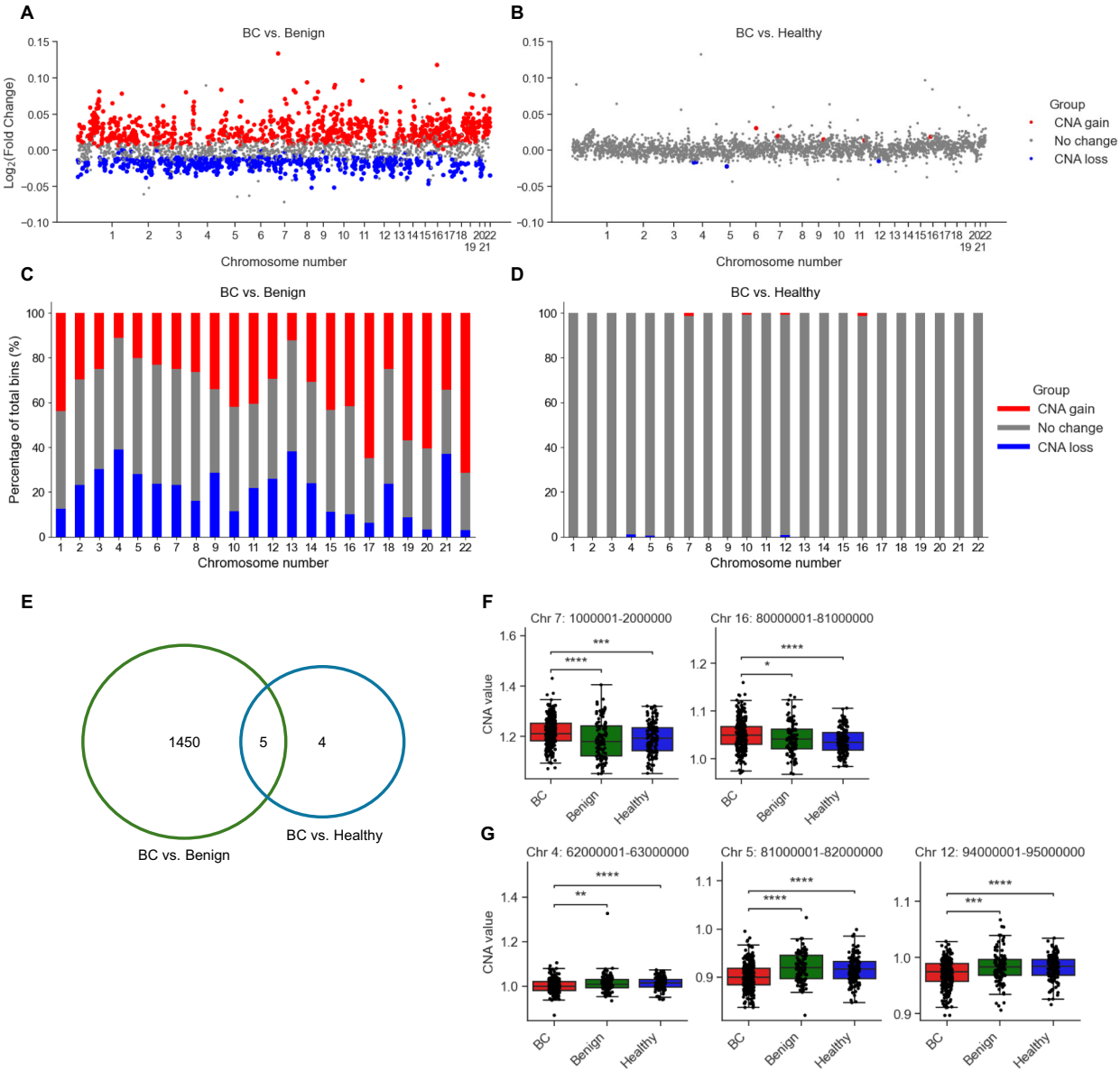

Fig. 2 Analysis of genome-wide copy number aberration (CNA) in cfDNA. Scatter plot shows log2 fold change of DNA copy number in each bin across 22 chromosomes of 273 BC patients versus 108 benign patients (**A**) or 134 healthy subjects (**B**) in the discovery cohort. Each dot represents a bin identified as gain (red), loss (blue), or no change (grey) in the copy number. Proportions of different CNA bins in each chromosome for BC versus benign patients (**C**) and BC versus healthy individuals (**D**). **E** Venn diagram indicates the number of significant bins overlapped between two pair-wise comparisons (BC versus benign and BC versus healthy). Boxplots showing 2 gain bins (**F**) and 3 loss bins (**G**) of copy number in BC patients compared to benign or healthy individuals

were located at position − 1, where we observed significant enrichment of thymine (T) and a reduced frequency of Guanine (G) (Fig. 3D,E). Another two ME21s were located at position − 2, where we identified a notable increase in cytosine (C) and a decrease in guanine (G) (Fig. 3D,E). The remaining 4 significant ME21s exhibited an increased proportion of cytosine (C) at positions − 3, 0, 1, and 2 (Fig. 3D,E).

Altogether, our genome-wide sequencing analysis reveals significant changes in CNA and ME that differentiate BC patients from both healthy individuals and those with benign lesions. These findings underscore their potential utility as biomarkers for differential diagnosis.

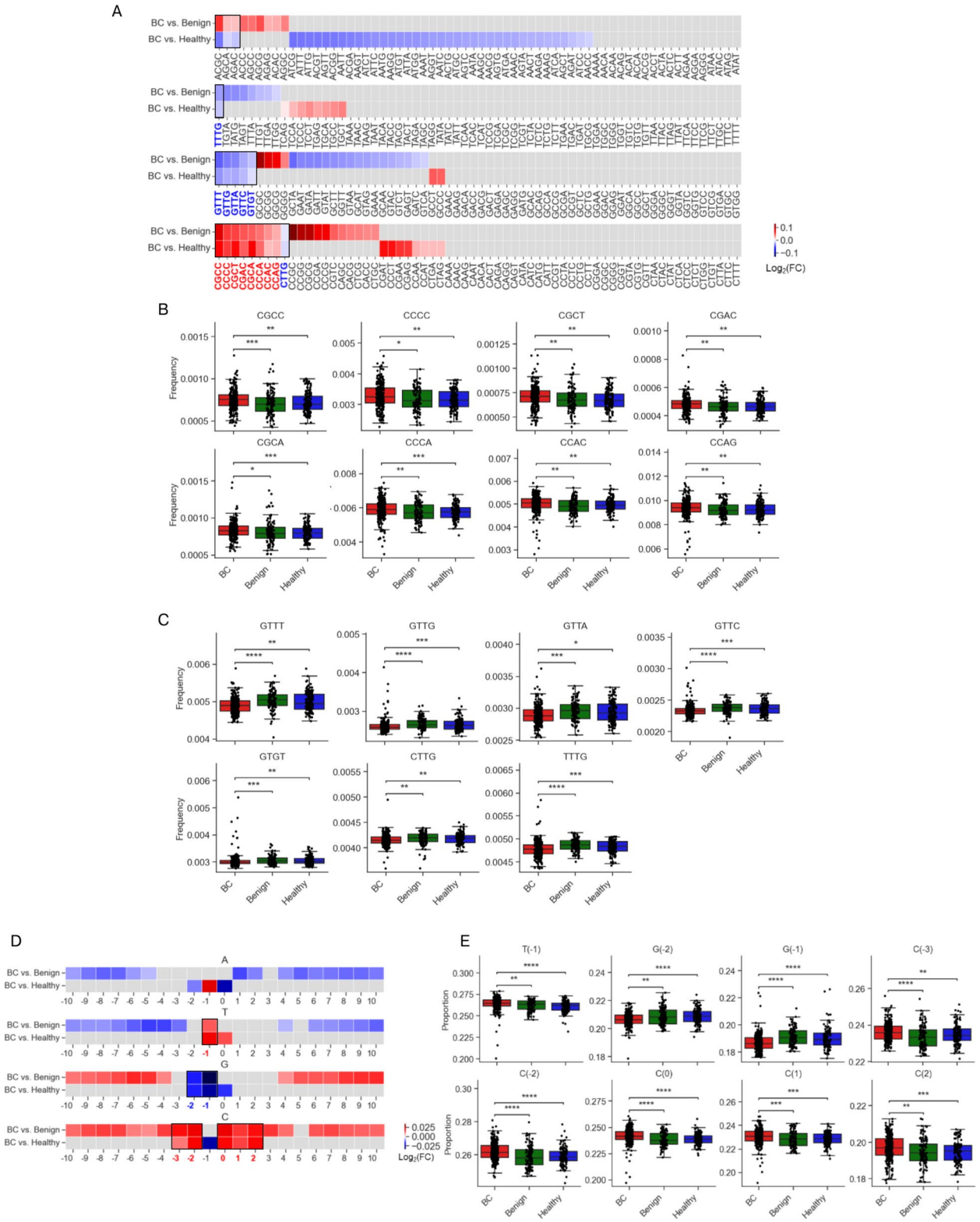

Fig. 3 Distinct end motif patterns of plasma cfDNA in BC, benign and healthy individuals. **A** Heatmap shows log2 fold change of 256 4-mer end motifs in BC patients compared to benign or healthy subjects (end motifs were highlighted in red for higher frequency and blue for lower frequency in BC patients). Boxplots showing 8 increased EMs (**B**) and 7 decreased EMs (**C**) of copy number in BC patients compared to benign patients or healthy controls. **D** Heatmap indicates log2 fold change of motif 21 between BC versus benign patients and BC versus healthy subjects. **E** Boxplots show 8 significant ME21s in BC patients compared to benign or healthy individuals

Identification of methylation changes and CNA in target regions for distinguishing BC patients from benign and healthy individuals

DNA methylation alterations are critical epigenetic modifications that regulate the expression of cancer-associated genes, playing a crucial role in cancer carcinogenesis [24]. To explore this, we profiled methylation changes across 450 target regions selected for their importance in the transcriptional regulation of cancer-associated genes. Among these regions, we identified three DMRs, *HIVEP2_GPR126*, *KLF3_TLR10*, and *MAFB_TOP1*, that were consistent across two pairwise comparisons in the discovery cohort (p-value ≤ 0.05, Mann–Whitney U test, Benjamini–Hochberg correction, Fig. 4A).

BC patients exhibited significant hypermethylation in the *HIVEP2_GPR126* region, which are associated with *HIVEP2* and *GPR126*, and in the *KLF3_TLR10* region, which regulates *KLF3* and *TLR10* ($Log_2FC > 0$, Fig. 4A,B). Conversely, hypomethylation was observed in the *MAFB_TOP1* region, which are associated with *MAFB* and *TOP1* ($Log_2FC < 0$; Fig. 4A,B). Notably, genes such as *GPR126*, *KLF3*, *TLR10*, and *TOP1* are well-documented for their roles in BC cell differentiation [25–28]. Thus, our methylation analysis identified three significant DMRs

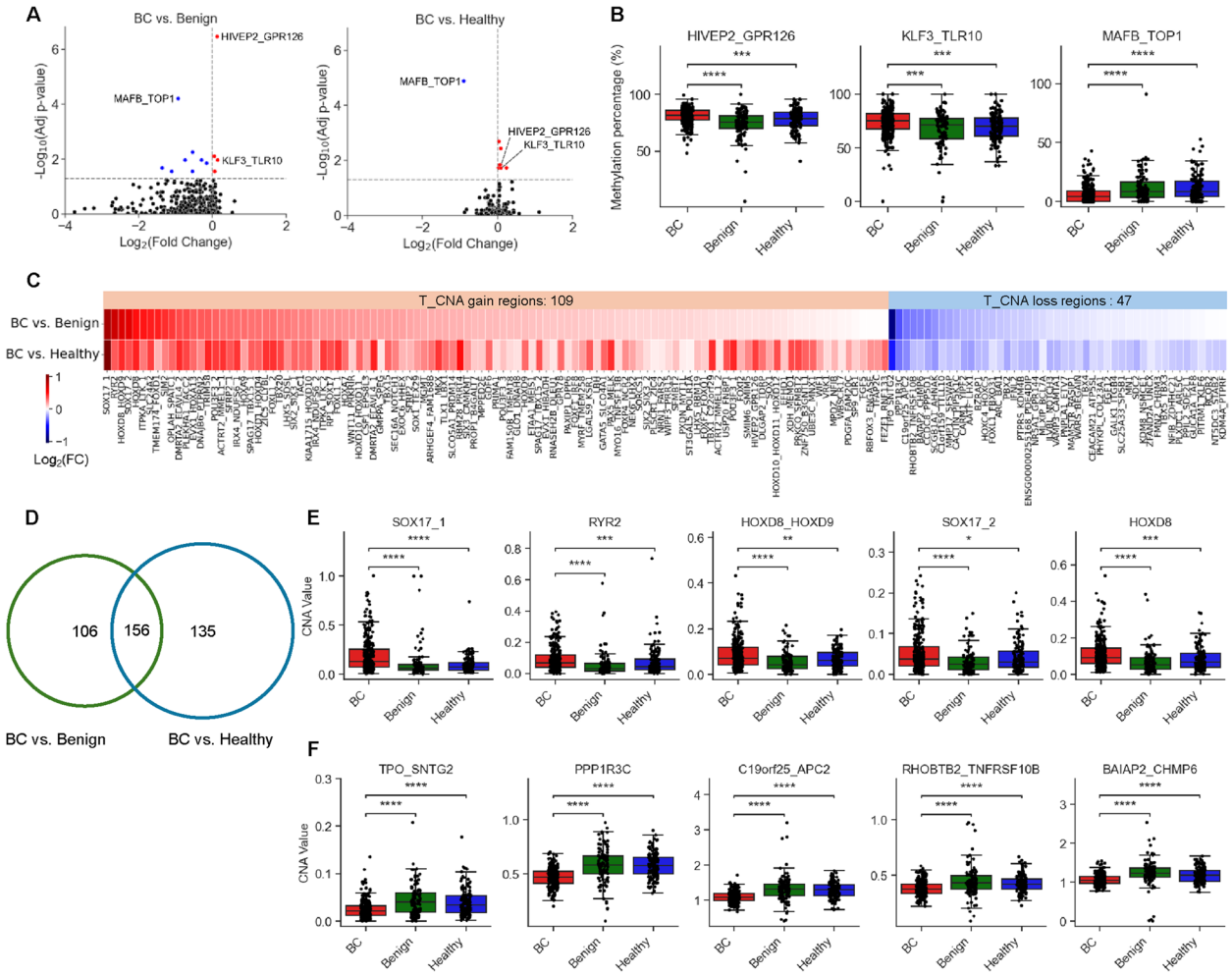

Fig. 4 Analysis of targeted methylation and copy number in plasma cfDNA. **A** Volcano plot shows log2 fold change and significance methylation (−log10 Benjamini- Hochberg adjusted p-value from Mann-Whitney U test) of 450 target regions when comparing BC patients to benign or healthy controls in the discovery cohort. There are 3 DMRs (adjusted p-value< 0.05) and overlapped across these two pairwise comparisons, color-coded by genomic locations (highlighted in red for log2FC>0 and blue for log2FC<0). **B** Boxplots showing 3 regions (HIVEP2_GPR126, KLF3_TLR10, and MAFB_TOP1) in BC patients, benign patients, and healthy participants. **C** Heatmap shows log2 fold change of 156 regions with significant copy number in BC patients compared to benign or healthy individuals. The number of significant regions in each pairwise comparison and overlapped regions was indicated via Venn diagram (**D**). Boxplots showing the top 5 with significantly raised CNA values (**E**) and the top 5 with significantly reduced CNA values (**F**) of BC patients compared to benign or healthy individuals

associated with the transcriptional regulation of BC-related genes, highlighting their potential as biomarkers to distinguish BC from benign or healthy individuals.

DNA methylation changes are often associated with CNAs, as both can reflect underlying changes in gene regulation [29]. To complement the methylation analysis, we assessed CNAs in the same 450 target regions among BC, benign, and healthy individuals. This localized analysis of CNAs, distinct from genome-wide CNA studies, was termed as "T_CNA." Of these 450 regions, 156 (34.7%) exhibited significant CNAs across the two pairwise comparisons (p-value ≤ 0.05, Mann–Whitney U test, Benjamini–Hochberg correction, Fig. 4C,D). Of these, 109/156 regions (69.9%) displayed a significant T_CNA gain ($Log_2FC > 0$, Fig. 4C), while 47/156 regions (30.1%) exhibited a significant T_CNA loss ($Log_2FC < 0$, Fig. 4C) in BC patients compared to benign or healthy controls. The top 5 regions with a significant T_CNA gain in BC patients included *SOX17_1*, *RYR2*, *HOXD8_HOXD9*, *SOX17_2*, and *HOXD8* (Fig. 4E). In contrast, regions with a significant T_CNA loss included *TPO_SNTG2*, *PPP1R3C*, *C19orf25_APC2*, *RHOBTB2_TNFRSF10B*, and *BAIAP2_CHMP6* (Fig. 4F).

Our study identifies three key DMRs and several regions with significant T_CNA alterations, underscoring their potential utility as biomarkers for BC diagnosis. Alterations in methylation and copy number at key regions may provide valuable insights into the regulatory mechanisms underlying BC progression.

A multimodal analysis combining signatures from both the target and genome-wide fraction to enhance the accuracy of BC detection

The identification of multiple significant signatures from both the targeted and genome-wide fractions motivated the development of a classification model to discriminate BC patients from individuals with benign conditions and healthy controls.

To construct this model, we divided the discovery cohort dataset into training and test sets. Dimensionality reduction was applied to the training dataset to remove highly correlated features using the Kendall rank correlation coefficient ($r \geq 0.7$). Subsequently, we performed tenfold cross-validation to identify stable features, employing the SelectKBest method to retain the top 500 features based on their recurrence across iterations (Fig. 5A).

We assessed the performance of five machine learning algorithms, including LR, SVM, DT, RF, and XGB, using the HyperClassifierSearch function with tenfold cross-validation and default parameters. Among these, XGB outperformed the other algorithms (Additional file 1:Fig. S1). After selecting XGB as the optimal algorithm, we fine-tuned its hyperparameters using tenfold cross-validation to optimize model performance. The finalized model was evaluated on the test set, achieving outstanding performance metrics.

In the training set, our multimodal model achieved area under the ROC curve (AUC) values of 0.97 (95% CI 0.95–0.99), 0.95 (95% CI 0.92–0.98), and 0.98 (95% CI 0.96–0.99) for discriminating BC patients from non-cancer individuals (Fig. 5B), benign lesion cases (Fig. 5C), and healthy controls (Fig. 5D), respectively. Validation on the test set demonstrated robust performance for pairwise classifications, with AUC values of 0.90 (95% CI 0.87–0.94), 0.88 (95% CI 0.83–0.94), and 0.92 (95% CI 0.88–0.96) for the same classifications (Fig. 5B–D).

To further enhance clinical utility, we implemented a cutoff value of 0.88 to ensure a specificity of at least 95%, thereby minimizing false-positive rates. Under this threshold, the model achieved a sensitivity of 80.4% for detecting BC patients, with specificities of 97.4%, 96.2%, and 98.5% for differentiating BC patients from benign cases, and healthy controls, respectively (Fig. 5E). When applied to the test set, the model maintained high specificities of 93.6%, 91.1%, and 95.7% for the same classifications, achieving an overall sensitivity of 66.2% for detecting BC patients (Fig. 5E).

To better understand the contribution of cfDNA signatures in our breast cancer detection model, we applied SHAP analysis to identify the top 20 most impactful features [30]. Among these, the proportion of Guanine (G) at position 0 of cfDNA fragments emerged as the most influential feature (Additional file 1:Fig.S3). Additionally, the proportions of Adenine (A), Guanine (G), and Thymine (T) at position − 1 were also ranked among the top 20 features. These nucleotide patterns at cleavage sites—particularly the enrichment of A and T and depletion of G—may reflect the activity of specific endonucleases and chromatin structural changes [31].

This multimodal approach effectively integrates targeted and genome-wide signatures, demonstrating strong potential for distinguishing BC patients from non-cancer, benign, and healthy individuals. The high specificity and competitive sensitivity suggest its promise for clinical applications.

The multimodal assay enabled effective detection of BC patients at early stages and with heterogeneous molecular subtypes

Detecting early-stage BC is particularly challenging due to the low levels of ctDNA present in the bloodstream [32]. The model demonstrated robust performance in detecting stage I, II, and III tumors, with an AUC of 0.96 (95% CI 0.92–0.99), 0.98 (95% CI 0.96–0.99), and 0.98 (95% CI 0.97–1.00), respectively, on the training set

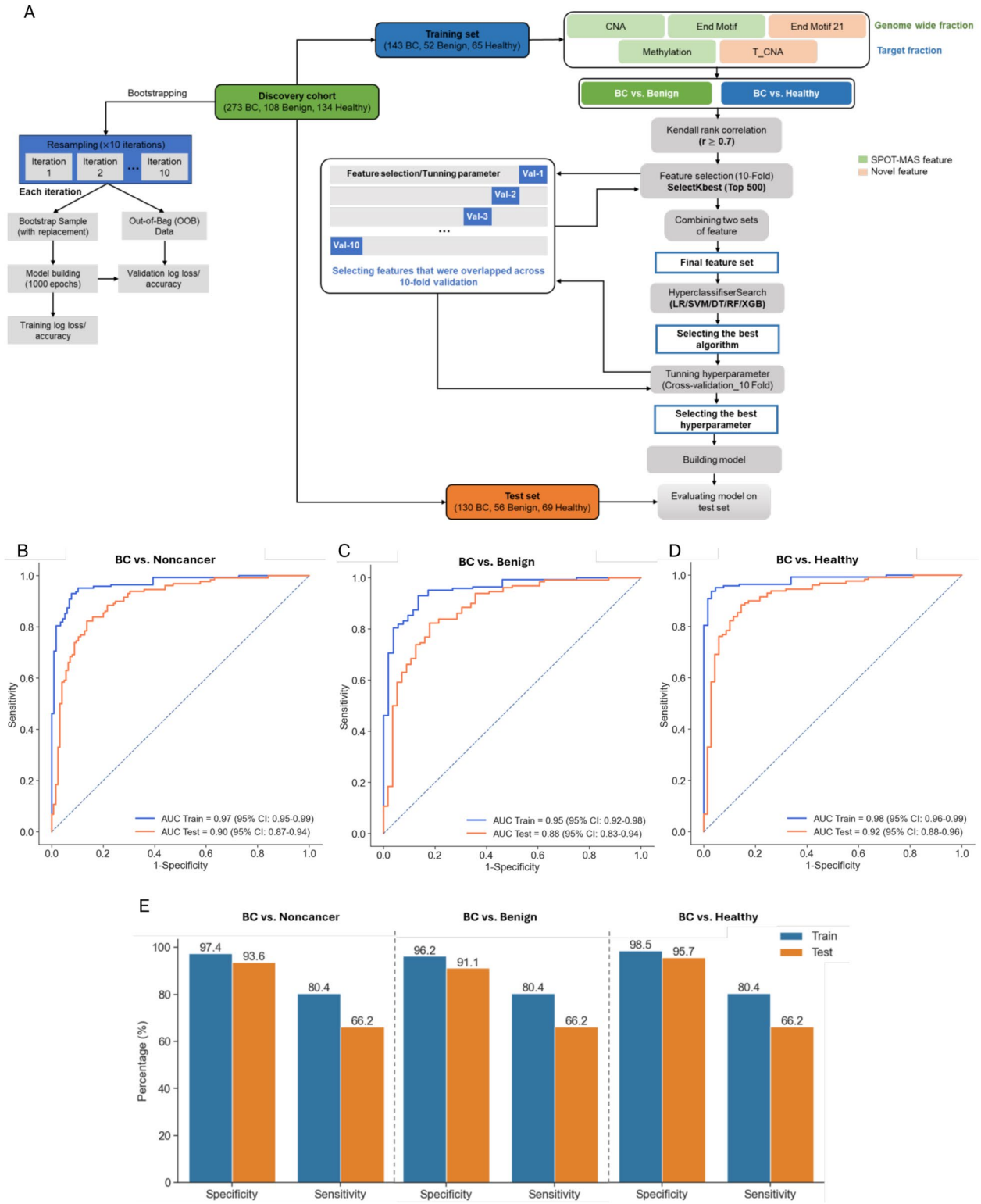

Fig. 5 Model construction and evaluation of multi feature model. **A** Model construction workflow. Receiver Operating Characteristic (ROC) curves showing the performances of the multi-feature models with the classification BC versus noncancer (**B**), BC versus benign (**C**), and BC versus healthy (**D**). **E** Bar plot displaying sensitivity and specificity values in the train and the test set

(Fig. 6A). Validation on the test set showed AUCs of 0.92 (95% CI 0.87–0.97) for stage I, 0.90 (95% CI 0.85–0.94) for stages II, and 0.90 (95% CI 0.83–0.97) for stages III (Fig. 6B). Sensitivities increased incrementally with disease stage, detecting BC patients in stages I, II, and III with sensitivities of 62.1%, 66.3%, and 66.7%, respectively (Fig. 6C).

BC is a highly heterogeneous disease, comprising diverse subtypes with distinct genetic, molecular, and clinical characteristics [33]. To assess the ability of model to identify these subtypes, we evaluated its performance across five molecular subtype groups: Luminal A, Luminal B, Luminal B-HER2, HER2-enriched, and TNBC. The model exhibited excellent performance in the training set, achieving AUC values ranging from 0.95 to 0.99 for all subtypes (Fig. 6D). In the test set, the model maintained strong performance, with AUC values ranging from 0.91 to 0.92 (Fig. 6E). The Luminal B-HER2 subtype demonstrated the lowest test set performance, with an AUC of 0.86 (95% CI 0.76–0.94) (Fig. 6E). Sensitivity analysis revealed that the model effectively differentiated between molecular subtypes in the test set. The highest sensitivity was observed for Luminal B (69.6%), followed by TNBC (63.6%), Luminal A (63.6%), HER2-enriched (61.9%), and Luminal B-HER2 (60.0%) (Fig. 6F).

Our multimodal assay achieved high accuracy in detecting early-stage BC and demonstrated strong performance across heterogeneous molecular subtypes. These results highlight the potential of this approach as a robust diagnostic tool for early detection and molecular classification of BC.

The multimodal assay demonstrated consistently high performance in a validation cohort of participants with breast lesions suspected of malignancy

To further validate the performance of the multimodal model in a diagnostic setting, we recruited 119 participants presenting lesions suspected of BC, classified as BI-RADS 3 to 5 following mammography. These participants, forming the validation cohort, were referred for diagnostic testing via FNA to confirm lesion types. Blood samples for the multimodal assay were collected prior to the FNA procedure. The assay results were subsequently

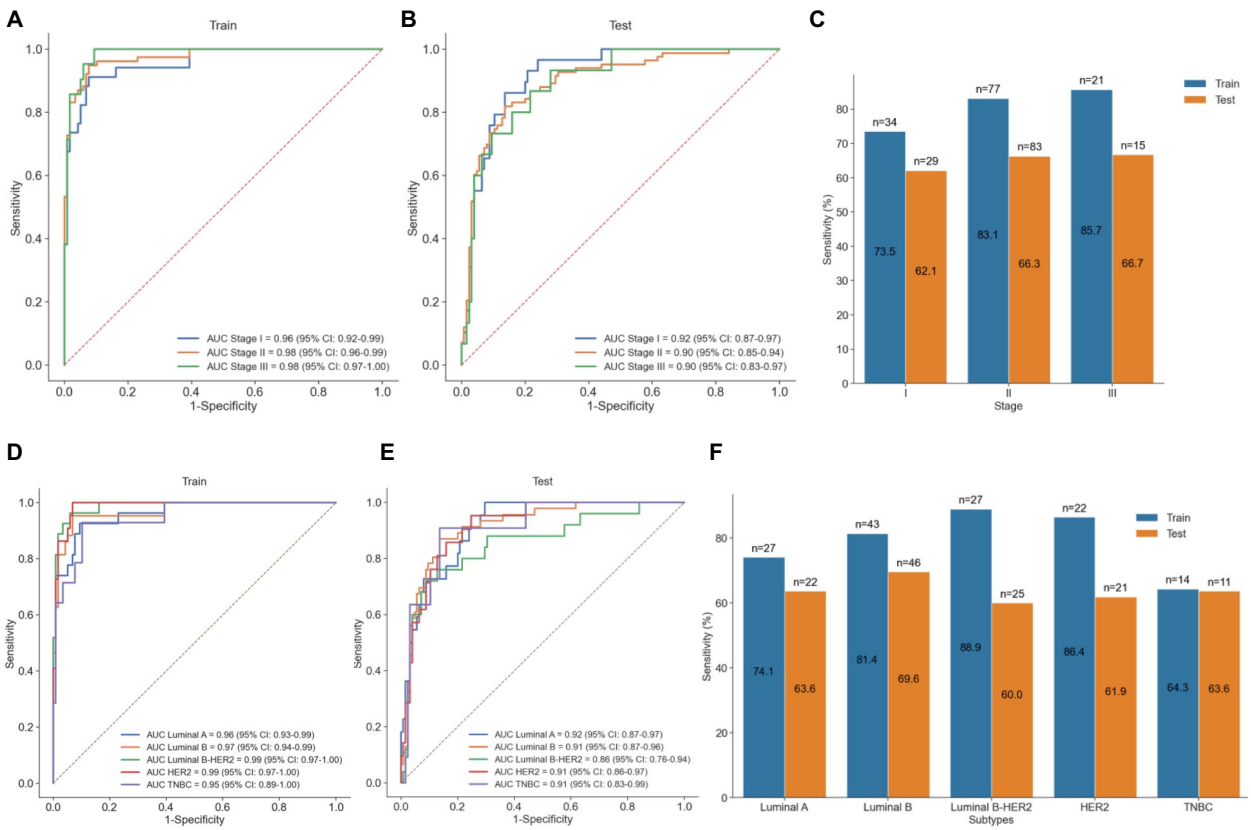

Fig. 6 Performance of the model in BC detection at different stages or with different subtypes. ROC curves showing the performance of the model for BC stage I to III in the train set (**A**) and the test set (**B**). **C** Bar plot showing the sensitivities of the model according to stages I–III. ROC curves showing the performance of the model on different BC subtypes in the train set (**D**) and the test set (**E**). **F** Bar plot showing the sensitivities of the model according to subtypes of BC patients

compared with FNA outcomes and tissue biopsy findings from participants who underwent surgery.

Among the 119 patients, 43 (36.1%) were confirmed to have BC based on FNA results. Clinical characteristics of the cohort are detailed in Additional file 1:Table S2 and Table S3. Of the 63 patients with BI-RADS 3, 4 (6.3%) were diagnosed with malignant lesions, while 59 (93.7%) had benign lesions, including 46 fibroadenomas, 3 cases of fibrocystic changes, 6 cysts, 1 atypical hyperplasia, and 3 unavailable subtype information cases (Additional file 1:Table S2 and Fig. 7A). In contrast, higher proportions of patients with BI-RADS 4 (59.5%) and BI-RADS 5 (89.5%) were confirmed to have malignant lesions. Among the 43 BC patients, 36 (83.7%) were diagnosed at early stages (I–II), while 7 (16.3%) were diagnosed at late stages (III) (Additional file 1:Table S2 and Fig. 7A). Evaluation of the multimodal model on this cohort demonstrated an overall sensitivity of 65.1%, with sensitivity of 50.0% for BI-RADS 3, 68.2% for BI-RADS 4, and 64.7% for BI-RADS 5 patients in identifying BC (Fig. 7B). Notably, the model achieved high specificity of 96.1% in differentiating benign individuals from BC patients, including 100% specificity for BI-RADS 5 classification. Among BI-RADS 4 patients, one benign case (a fibroadenoma) was misclassified as malignant, but the model still maintained a specificity of 93.3% for this category (Fig. 7B). Similarly, in the BI-RADS 3 group, two benign cases (a cyst and a fibroadenoma) were misclassified as malignant, yielding a specificity of 96.6% (Fig. 7B). Stratification of BC patients by stage revealed that the model detected BC with sensitivities of 53.3% for stage I, 76.2% for stage II, and 57.1% for stage III (Fig. 7C). Moreover, the model exhibited high sensitivity for certain molecular subtypes, achieving the highest sensitivity (80%) for

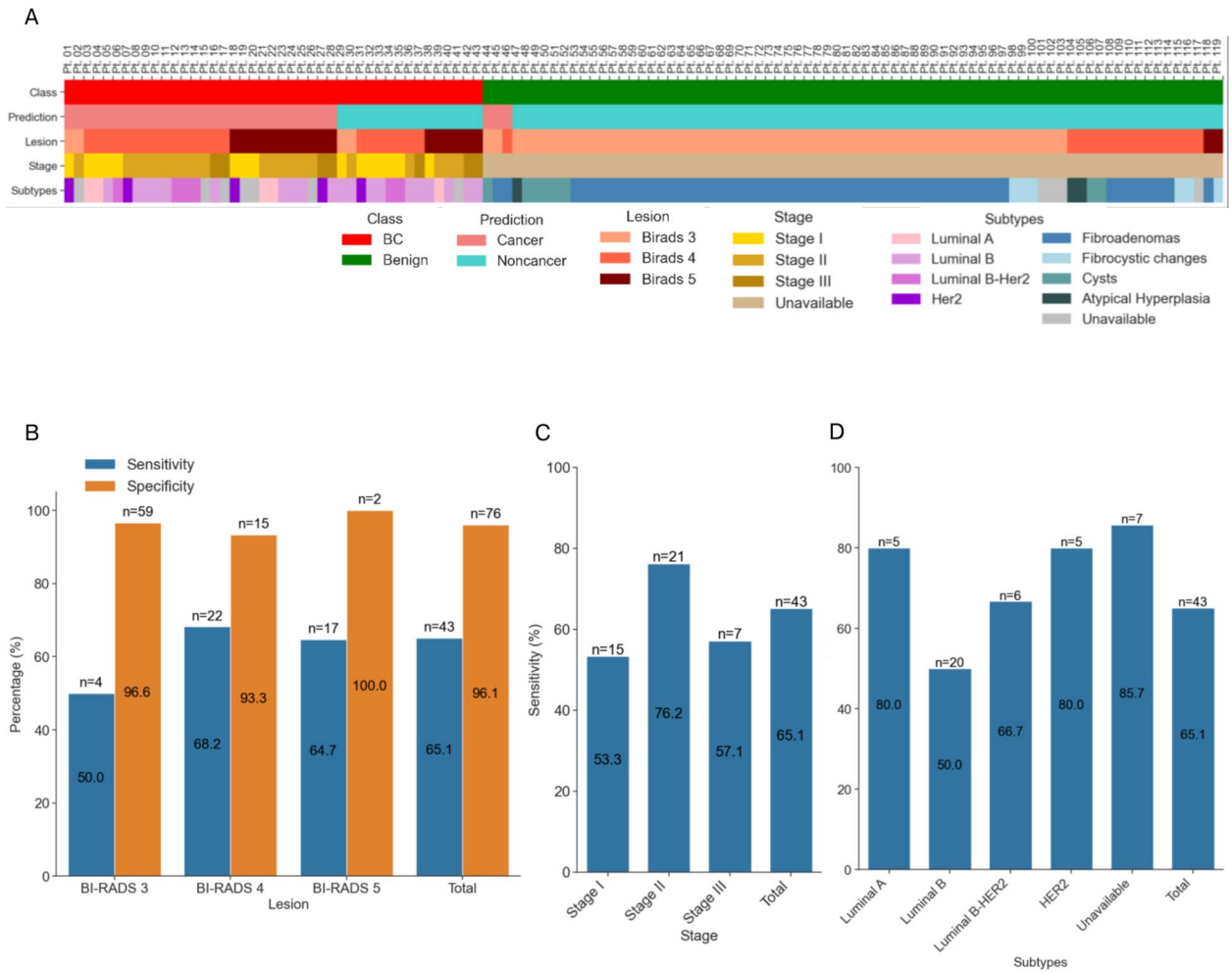

Fig. 7 Performance of the model in the validation cohort. **A** Heatmap shows the demographic details of 119 symptomatic participants in the validation cohort. **B** Bar plot showing sensitivities and specificities of the model according to BI-RADS lesions. **C** Bar plot indicates sensitivities of BC patients in different stages. **D** Bar plot displaying sensitivities of BC patients in different subtypes

both Luminal A and HER2-enriched subtypes. Detection sensitivity for the Luminal B-HER2 and Luminal B subtypes was slightly lower, at 66.7% and 50%, respectively (Fig. 7D).

These findings demonstrate that the multimodal assay effectively differentiates benign from malignant lesions in a diagnostic setting, achieving high specificity for differentiating benign lesions from early-stage tumors. This highlights its potential as a robust companion tool complementing SOC screening tests in clinical practice.

Discussion

Recent studies have highlighted liquid biopsy as a promising approach for early BC detection. However, distinguishing benign lesions from malignant tumors remains a significant challenge [34–36]. In this study, our multimodal analysis of cfDNA characteristics demonstrated robust performance in differentiating BC patients not only from healthy controls but also from individuals with benign conditions. This distinction has the potential to reduce false-positive rates, thereby minimizing unnecessary biopsies in clinical practice. Furthermore, our protocol integrates both targeted enrichment and shallow genome-wide sequencing (0.5X) from a single 10 mL blood sample and library preparation (Fig. 1). This "all-in-one" approach allows for the simultaneous profiling of multiple cfDNA signatures while keeping testing costs low compared to other assays [22], thus supporting its potential implementation in routine clinical practice to differentiate breast cancer from benign and healthy cases.

BC is a heterogeneous disease, characterized by variations in the genomic, epigenomic, transcriptomic, and proteomic profiles of cancer cells [37]. Additionally, at early stages, the concentration of ctDNA in the bloodstream is extremely low, posing a challenge for detection [38]. To address these complexities, we previously analyzed multiple features of cfDNA, including methylation and fragmentomic patterns, to identify a broader spectrum of BC-specific signatures. In this study, we expanded on this approach by integrating three genome-wide signatures—CNA, EM, and ME21, and two targeted signatures—TMD and targeted CNA (T_CNA).

Aberrant DNA copy number changes in breast cancer primarily arise from genomic instability during DNA replication—a hallmark of cancer. These alterations, involving the gain or loss of DNA segments, often confer growth or survival advantages to tumor cells [39]. Our statistical analysis identified significant CNAs across several chromosomes, consistent with previously reported findings. For instance, gain of chromosome 16 has been associated with high nuclear grade, larger tumor size, advanced stage, increased expression of topoisomerase IIα, and poorer overall survival in breast cancer [40]. Furthermore, luminal B-like and HER2-overexpressing subtypes frequently exhibit gains in 1q, 8q, 16p, and 17q—features characteristic of the "firestorm" CNA profile [41]. In our genome-wide fraction, CNA analysis revealed significant alterations across all 22 chromosomes when comparing BC patients with benign conditions, with more pronounced CNA signatures than those observed between BC and healthy subjects. Benign breast disease may represent an intermediate stage in the progression from normal tissue to cancer [42], which could explain the complex and variable patterns seen in this group. Furthermore, inflammation in benign breast tissue may contribute to genomic instability, reflecting changes in the tumor microenvironment [43]. Similarly, altered fragmentation patterns in cfDNA likely result from tumor-specific chromatin accessibility and nucleosome positioning, influenced by the cancer cell's epigenetic state and apoptosis dynamics [44]. During apoptosis, DNA in cancer cells and normal cells cleaves differently, resulting in variations in the DNA patterns at the end of each fragment. Previous studies have demonstrated that 4-mer end motifs can be leveraged for detecting ctDNA in the plasma of patients with hepatocellular carcinoma and other cancers [22, 45]. In this study, we identified distinct end motif patterns distinguishing BC from healthy controls as well as from patients with benign patients. Notably, all enriched end motifs in BC began with "C," consistent with findings from our previous multi-cancer early detection study, which included BC [22]. Similarly, Jin et al. identified three 4-mer end motifs (CCCA, CCTG, and CCAG) as preferred end motifs in hepatocellular carcinoma [46]. Serpas et al. further suggested that *DNASE1L3* may contribute to the generation of the "CCCA" end motif [47]. In a related study, Zhu et al. [48] employed an experimental model using the HL60 human leukemia cell line undergoing apoptosis to investigate DNA fragmentation patterns via shallow whole-genome sequencing. Their results revealed that C-end motif preferences (e.g., fragments beginning with cytosine) were retained both intracellularly and extracellularly in apoptotic cells, even in the absence of *DNASE1L3*. Interestingly, they also observed a reduced abundance of C-end motifs in cfDNA from cancer patients. One possible explanation is that while *DNASE1L3* may be downregulated in tumors, other nucleases—such as *DNASE1*, *DFFB*, or additional endonucleases—could compensate or preferentially cleave at C-rich regions, contributing to the observed motif enrichment in vivo [49]. Although the HL60 apoptosis model offers valuable mechanistic insights under controlled in vitro conditions, it does not fully capture the biological complexity of cfDNA fragmentation in vivo. In the physiological setting, factors such as tumor

heterogeneity, the diversity of cell death pathways, and dynamic interactions within the tumor microenvironment collectively influence nuclease activity and chromatin accessibility. These findings underscore the need for appropriate in vivo models in future studies to disentangle the relative contributions of specific nucleases and chromatin context to cfDNA fragmentomic patterns in cancer.

Previous studies have also shown that cfDNA fragment start and end sites contain distinct nucleotide motifs that can differentiate tumor-derived fragments [50]. For example, Chandrananda et al. demonstrated that 10-bp motifs upstream and downstream of the cfDNA cleavage site exhibit consistent patterns in cancer-derived cfDNA [23]. Consistently, we identified significant ME21 signatures enriched in BC compared to benign patients or healthy individuals. This finding highlights the non-random fragmentation process of BC-derived ctDNA, underscoring its potential as a marker to discriminate benign lesions from malignant tumors. The proportions of Guanine (G) at position 0, and Adenine (A), Guanine (G), and Thymine (T) at position − 1, were identified as key cfDNA features contributing to model performance (Additional file 1:Fig.S3). These nucleotide patterns at cfDNA cleavage sites—particularly the enrichment of A and T and the depletion of G—are thought to reflect the enzymatic activity of endonucleases and tumor-specific chromatin architecture. Supporting this, Bai et al. [52] demonstrated that cfDNA is initially cleaved intracellularly by *DFFB* and *DNASE1L3*, which preferentially generate A-end and C-end fragments, respectively. Subsequent extracellular cleavage is mediated by *DNASE1L3* and *DNASE1*, with *DNASE1* favoring the generation of T-end fragments [51]. In addition, a previous study reported substantial intratumoral heterogeneity in the expression of DNase family genes across 33 tumor types, which may contribute to the distinct nucleotide profiles observed at cleavage sites in breast cancer-derived ctDNA [52].

Across 450 target regions, we identified three DMRs (*HIVEP2_GPR126*, *KLF3_TLR10*, and *MAFB_TOP1*, Fig. 4A) between BC and both benign patients and healthy subjects. Aberrant DNA methylation in breast cancer often reflects transcriptional silencing of tumor suppressor genes or activation of oncogenic pathways, which is mirrored in the plasma cfDNA methylation landscape [53]. Among the hypermethylated targets in our study, *GPR126*, a member of the adhesion G protein-coupled receptor family, plays a key role in the progression of triple-negative breast cancer [25]. We also observed hypermethylation in regulatory regions associated with *KLF3* and *TLR10*, consistent with previous studies that identified both genes as tumor suppressors [54, 55] and reported their downregulation in BC patients compared to healthy controls [55]. Similarly, decreased expression of several toll-like receptors (*TLRs*), including *TLR10*, has been documented in BC tissues relative to normal tissues [27]. In addition, we observed hypomethylation of genomic regions associated with *TOP1*, aligning with reports of elevated *TOP1* expression in BC patients, as described by Tsyganov et al. [28]. Moreover, we investigated methylation differences between breast tumor and normal tissues using methylation array data from the TCGA-BRCA project. Out of 354,454 CpG sites analyzed, 262,289 were found to be significantly differentially methylated between tumor and normal samples (p-value ≤ 0.05, t-test, Benjamini–Hochberg correction). These CpG sites were annotated to 19,163 genes. When compared to the 584 genes annotated from our 450 targeted regions, 542 genes overlapped with those identified in the TCGA dataset. Notably, all six genes corresponding to the three most statistically significant cfDNA-derived regions (*HIVEP2_GPR126, KLF3_TLR10*, and *MAFB_TOP1*) also overlapped with the TCGA data (Additional file 1:Fig.S4A). In addition, two key genes identified through SHAP analysis—*SNAI1* and *UBE2V1*—were also present among the TCGA-derived differentially methylated genes (Additional file 1:Fig.S4B). These results support the biological relevance of our identified methylation signatures and suggest their potential involvement in breast cancer occurrence and progression.

In the target fraction, we observed several regions with significant copy number changes in BC compared to both benign patients and healthy subjects. Previous studies have reported that associations between somatic CNA and DNA methylation can be either positive or negative. In contrast, associations between CNA and gene expression may occur independently, as DNA methylation does not strongly mediate this relationship [29, 54]. In our study, we identified regions with T_CNA alterations localized in non-differentially methylated regions of BC. Prior research indicates that the copy number variant of *SOX17* influences both methylation and gene expression in BC [55].

Our multimodal approach demonstrated that combining distinct cfDNA features enhances sensitivity for BC detection while maintaining high specificity in distinguishing benign and healthy classifications. Although the model performed well on the training set, with a sensitivity of 80.4% (Fig. 5E), sensitivity declined to 66.2% on the test set (Fig. 5E). In this study, our multi-omics analysis resulted in a high-dimensional feature space, which could increase the risk of overfitting during model construction [56]. To mitigate this, we implemented a train-test split combined with tenfold cross-validation within the training set (Fig. 5A). This strategy enabled robust

hyperparameter tuning and stable feature selection across folds, thereby reducing the risk of overfitting to any specific subset. Notably, the test set was completely withheld during model development and used only once for final evaluation, ensuring an unbiased assessment of the model's generalization performance. This approach allows for efficient use of limited data while maintaining the integrity of performance evaluation. Previous studies have emphasized the effectiveness of cross-validation in mitigating overfitting and highlighted the importance of independent test set evaluation to avoid optimistic bias in performance estimation [57, 58]. To evaluate model stability and assess the risk of overfitting, we performed bootstrap resampling across 1000 training epochs on the discovery cohort (Fig. 5A). For each of 10 bootstrap iterations, an XGBoost classifier was trained on a resampled dataset drawn with replacement from the discovery cohort. Model performance was monitored across 10 independently bootstrapped models. Beyond 200 epochs, log-loss remained consistently low (Additional file 1:Fig. S5A), and classification accuracy remained high (Additional file 1:Fig.S5B), indicating stable model behavior. Notably, both log-loss and accuracy stabilized early in the training process, suggesting efficient convergence and obviating the need for extended training beyond 200 epochs (Additional file 1:Fig.S5). The narrow and stable performance gap between training and validation sets further supports good model generalizability and indicates an absence of significant overfitting (Additional file 1:Fig.S5B). Despite these challenges, our performance in screening BC was highly competitive, particularly in the context of benign lesion inclusion. Comparatively, the Galleri test, which utilizes a panel of > 100,000 informative methylation loci via targeted whole-genome bisulfite sequencing of plasma DNA to detect > 50 cancer types, achieved a specificity of 99.3%. However, BC exhibited the lowest sensitivity, with detection rates of less than 25% for stage I and approximately 50% for stage II cases in case–control studies [59]. Compared to our MCED test (SPOT-MAS) [22], which utilizes features selected through comparisons across multiple cancer types, the multi-feature model presented in this study demonstrated superior performance (66.2% vs. 49.3%). These findings support our hypothesis that focusing on BC-specific features for training a machine-learning model enhances its performance for detecting a particular cancer type.

In this study, we demonstrated that the performance of the multi-feature model was not subtype dependent. Breast cancer is a biologically heterogeneous disease comprising several molecular subtypes, including Luminal A, Luminal B, HER2-enriched, and triple-negative breast cancer (TNBC), each characterized by distinct molecular alterations, biological behaviors, and clinical outcomes [60]. Approximately 20% of Luminal B tumors are HER2-positive, a group classified as Luminal B–HER2 ($ER^+/HER2^+$)[60]. In our study, the lower sensitivity observed for this subtype in test set (60%) may reflect subtype-specific differences in ctDNA shedding dynamics and epigenetic features. Less aggressive or more indolent subtypes, such as Luminal A and Luminal B ($ER^+/HER2^-$), tend to release lower levels of ctDNA into circulation. In contrast, more aggressive subtypes, including HER2-enriched and TNBC, are associated with higher ctDNA release due to increased tumor proliferation and necrosis [61]. Although Luminal B–HER2 tumors are more proliferative than Luminal A [62], they may still shed less ctDNA than HER2-enriched non-luminal or TNBC subtypes, which exhibit more aggressive tumor biology and higher ctDNA abundance [63, 64]. This biological heterogeneity likely contributes to differences in detection sensitivity across subtypes. Our multimodal approach, which integrates a range of ctDNA-based features, may help mitigate this variability by capturing a broader spectrum of molecular signatures unique to each subtype. Notably, TNBC exhibited a high sensitivity of 63.6% in the test set (Fig. 6F), aligning with previous studies that emphasize the utility of ctDNA in diagnosing TNBC across early and advanced stages [56].

To further validate our multimodal model, we evaluated its performance in detecting BC within a high-risk cohort of participants with breast lesions by mammography and recommended for FNA for diagnostic confirmation. The malignancy rates for BIRADS-3, BIRADS-4, and BIRADS-5 were 6.3%, 59.5%, and 89.5%, respectively (Additional file 1:Table S2), consistent with previous reports [65, 66]. These malignancy rates suggest that the high-risk cohort reflects real-world clinical scenarios. The model accurately identified 50.0% (2/4) of BC cases in the BIRADS-3 group, 68.2% (15/22) in BIRADS-4, and 64.7% (11/17) in BIRADS-5 (Fig. 7B). Overall, the model achieved a sensitivity of 65.1% (Fig. 7B) while maintaining a high specificity of 96.1% (Fig. 7B) in distinguishing BC from benign lesions across BIRADS categories. This performance is consistent with that observed in the original test set, which demonstrated a sensitivity of 66.2% (p-value = 1.0, Fisher's exact test) and a specificity of 91.1% (p-value = 0.28, Fisher's exact test) (Fig. 5E). These results indicate that the model maintains robust and consistent performance when prospectively validated in an independent external cohort. Moreover, these results highlight the potential of the multimodal model, trained on BC-specific signatures, to enhance differentiation between malignant and benign lesions and reduce the need for invasive diagnostic procedures in high-risk populations.

While sensitivity for stage I breast cancer remains moderate (53–62%), this performance surpasses that of several published cfDNA-based assays and reflects inherent biological challenges related to minimal ctDNA shedding in early disease [32, 67]. Despite these challenges, our test achieved a sensitivity of 62–66.3% for stage I–II breast cancers (Fig. 6C and Additional file 1:Table S4)—substantially higher than other reported ctDNA-based assays, such as Galleri [59] (25–50%), CancerSEEK (33%), DELFI (57%) [68], and SPOT-MAS MCED (49.3%) [22]. A nucleosome profiling-based cfDNA assay developed by Han et al. evaluated 173 malignant, 158 benign, and 102 healthy samples, reporting a sensitivity of 70.8% but a relatively low specificity of 76.5% in distinguishing breast cancer from benign lesions [69]. In contrast, our assay demonstrated a unique strength in differentiating breast cancers from benign breast conditions, achieving a high specificity of 96.1% (Fig. 7B and Additional file 1:Table S4) in a validation cohort representative of real-world high-risk populations. This highlights its potential utility in clinical settings where benign breast lesions are common and often lead to diagnostic uncertainty. Recently, Guardant Reveal™ has been approved for the detection of minimal residual disease (MRD) and recurrence monitoring in early-stage cancers, although it is not primarily intended for early detection or population-level screening. In exploratory analyses involving breast cancer, the assay demonstrated a sensitivity of 79% for detecting distant recurrence but only 13% for localized disease, while maintaining a specificity of 100% [70]. These findings highlight the persistent challenge of detecting ctDNA shed by early-stage breast tumors.

Despite these promising results, our study has several limitations. First, while the multimodal model showed strong classification performance, the validation cohort sample size was limited. Larger cohorts are needed for further validation before applying the model in real-world screening settings. Second, 16.3% of breast cancer patients were without subtype information (7/43) in the validation cohort (Additional file 1:Fig.S2 and Table S2). These patients underwent initial imaging at study hospitals but chose different facilities for subsequent tests and surgeries, limiting access to complete histological data. In addition, the limited sample size for each molecular subtype remains a constraint, particularly for Luminal B–HER2, HER2-enriched, and triple-negative breast cancer (TNBC). As reported by Clarke et al. [71], the distribution of breast cancer subtypes varies by age and population. Among Asian women, $HR^+/HER2^-$ tumors (Luminal A and Luminal B) represent approximately 55% of cases, $HR^+/HER2^+$ (Luminal B–HER2) account for ~ 12%, HER2-enriched ($HR^-/HER2^+$) for ~ 8%, and TNBC for ~ 9%[71]. Consequently, acquiring sufficient numbers of early-stage samples from these less common subtypes poses a significant challenge for subtype-specific model development and validation. Future validation of our model in more diverse populations and clinical contexts, and evaluation using real-world data, is required to confirm the robustness of our assay. Currently, we are conducting a prospective study, K-ACCELERATE (NCT06391749), which aims to recruit 1000 participants presenting with symptoms or signs suggestive of the five most common cancer types, including breast cancer. This study will provide an opportunity to assess the diagnostic performance of our model in a high-risk diagnostic population and evaluate its sensitivity across molecular subtypes. Third, while our protocol, which combines targeted and shallow genome-wide sequencing (Fig. 1), enables low testing costs, it does not capture the full spectrum of molecular signatures associated with breast cancer. We are actively incorporating deep sequencing to comprehensively profile the methylome and fragmentome, with the goal of integrating additional features into our framework. This expanded feature set is expected to further improve detection sensitivity and enhance biological interpretability. We believe this iterative strategy will enable the progressive refinement of our assay's performance while maintaining its clinical feasibility. Finally, while our assay leverages shallow whole-genome sequencing and an integrated library preparation protocol to reduce per-sample costs, we acknowledge that a comprehensive cost-effectiveness analysis—including direct comparison with existing modalities such as mammography in terms of cost-per-test, infrastructure requirements, and workflow integration—is needed to fully assess clinical scalability. Future prospective studies are warranted to evaluate these parameters in real-world settings.

Conclusions

Our findings underscore the potential of multimodal plasma cfDNA analysis to identify novel biomarkers for accurately distinguishing BC patients from benign lesions and healthy individuals. This capability could reduce false-positive rates and unnecessary biopsies, offering significant clinical utility for high-risk populations.

Methods

Patient enrollment

This study included a discovery cohort of 515 female participants, consisting of 273 BC patients, 108 individuals with benign breast disease, and 134 healthy controls. BC diagnoses were confirmed through abnormal mammography findings followed by tissue biopsy. Disease staging was determined using the TNM (Tumor, Node, Metastasis) classification system, based on guidelines from the American Joint Committee on Cancer (Version

VIII) and the International Union for Cancer Control. To emphasize the importance of early cancer detection, only patients with non-metastatic BC (stages I–IIIA) were included in the study.

Participants with benign breast disease and healthy controls underwent routine mammography as part of their annual health check-ups. Diagnoses of benign breast disease were confirmed at the time of inclusion and were characterized as non-malignant lumps, including fibroadenomas and fibrocystic changes. These individuals were monitored for 12 months to confirm their cancer-free status. Participants for the discovery cohort were recruited between October 2021 and December 2023 from Ho Chi Minh City Oncology Hospital and the Medical Genetics Institute in Ho Chi Minh City, Vietnam.

An external validation cohort was also recruited, comprising 119 female participants identified with BI-RADS 3–5 lesions through mammography. All participants underwent fine needle aspiration (FNA) for diagnostic confirmation. This validation cohort was recruited between April 2024 and April 2025 from Thai Nguyen National General Hospital, Can Tho Oncology Hospital, Nghe An Oncology Hospital, Da Nang Oncology Hospital and Buon Ma Thuot Medical University Hospital, Vietnam.

The study complied with the ethical principles set forth in the Declaration of Helsinki. The research protocol was approved by the Ethics Committees of all participating institutions, with ethics review number 294/BVUB-HĐĐĐ for the discovery cohort and 460/HĐĐĐ-ĐHYD for the validation cohort. Written informed consent was obtained from all participants before sample collection. Importantly, all samples were collected prior to the initiation of any therapeutic interventions to ensure the integrity of the analysis.

Study design

Plasma samples were isolated from 273 BC patients, 108 individuals with benign breast disease, and 134 healthy controls in the discovery cohort. We employed the SPOT-MAS assay, as described in our previous publication [22], to simultaneously analyze multiple cfDNA signatures from both targeted and genome-wide fractions. To enhance BC detection, we extended the analysis of 21-mer motif ends (ME21), targeted methylation density (TMD) and copy number aberration (T_CNA) signatures in targeted regions (Fig. 1).

Distinct cfDNA signatures were profiled from the targeted and genome-wide fractions in two pairwise comparisons: (1) BC versus benign disease, and (2) BC versus healthy subjects. Our analysis prioritized signatures exhibiting consistent trends across both comparisons in BC patients. After identifying BC-specific signatures, we developed a multi-feature model capable of differentiating BC patients from individuals with benign lesions and healthy controls (Fig. 1).

Finally, the predictive model was validated using an independent validation cohort, which had not been included in the training or testing phases. This cohort comprised 119 individuals with BI-RADS 3–5 lesions identified via mammography. Plasma samples from these participants were analyzed using the SPOT-MAS assay, and the model predicted the presence or absence of ctDNA signals. To confirm diagnoses, all participants underwent FNA, which served as the reference standard for determining cancer status. Following FNA, 43 participants were diagnosed with benign breast disease, while 76 participants were confirmed as having BC. The performance of model was evaluated by comparing its predictions with FNA-confirmed diagnoses (Fig. 1).

Isolation of cfDNA

Blood (10 mL) was collected in Cell-Free DNA BCT tubes (Streck, USA) and processed via two-step centrifugation as previously described [22]. The resulting plasma fractions were aliquoted and stored at − 80 °C for long-term preservation. cfDNA was extracted from 1 mL plasma using the MagMAX Cell-Free DNA Isolation Kit (ThermoFisher, USA) and quantified with the QuantiFluor dsDNA System (Promega, USA).

Multimodal analysis of genome-wide and targeted fractions of cfDNA

The cfDNA samples isolated for this study were analyzed using the SPOT-MAS assay, as previously described [22]. The workflow involves three key steps:

Step 1: Bisulfite conversion was performed using the EZ DNA Methylation-Gold Kit (Zymo Research, USA), and libraries were prepared with the xGen™ Methyl-Seq DNA Library Prep Kit (IDT, USA) using Adaptase™ technology.

Step 2: The library is subsequently subjected to hybridization to selectively enrich the target fraction, encompassing 450 cancer-specific regions, as detailed in the design and construction of the capture panel [22]. The non-targeted whole-genome fraction is recovered by collecting the flow-through and re-hybridizing it with probes targeting the adapter sequences of the DNA library. Both the target capture and whole-genome fractions are sequenced to depths of about 52X and 0.55X, respectively, generating 100-bp paired-end reads with a sequencing depth of 20 million reads per fraction. Preprocessing of the data generates five distinct cfDNA feature sets: CNA, EM, ME21, TMD, and T_CNA.

Copy number aberration analysis (CNA)

CNA analysis was performed using the QDNAseq R package [72] with a 1-Mb segmentation strategy. Regions of low mappability and Duke blacklist regions were excluded to ensure data quality. Read counts per bin were calculated using the "binReadCounts" function followed by GC content correction with "estimateCorrection" and "correctBins." To refine the CNA features, we applied "normalizeBins" for normalization and "smoothOutlierBins" to address outliers. This process yielded a final feature vector comprising 2691 bins.

End motif analysis (EM)

Library preparation with Adaptase™ added a random 5′ tail to reverse reads, precluding accurate 5′ end identification. Therefore, EM features were calculated based on the 5′ coordinates of forward reads, as previously described [22].

End motif 21 analysis (ME21)

To define the 21-bp end-motif feature (ME21), we extracted a 21-nucleotide sequence from the 5′ end of each R1 sequencing read. This sequence consisted of 11 bp from within the read and 10 bp extending into adjacent genomic regions. A position weight matrix (PWM) of dimensions 4×21 was constructed, where each row corresponded to a nucleotide (A, T, G, or C) and each column represented a position within the 21-bp sequence [50]. The PWM entries captured the proportional frequency of each nucleotide at each position. The matrix was subsequently flattened into a 1×84 vector, which served as an input feature for further analysis.

Targeted methylation density analysis (TMD)

Paired-end reads were trimmed using Trimmomatic v0.32 (HEADCROP), aligned with Bismark v0.22.3, and processed with Samtools v1.15 and Bedtools v2.28. Methylation calling was performed using Bismark methylation extractor [22]:

$$Methylation\ ratio = \frac{methylated\ cyto\ \sin\ e\ (c)}{methylated\ C\ +\ unmethylated\ C}$$

Methylation fold changes between BC and benign or healthy samples were calculated for each target region. Differentially methylated regions were identified by comparing BC and benign/healthy samples using a Benjamini–Hochberg adjusted p-value threshold of ≤ 0.05 ($-\log_{10}$ adjusted $p \geq 1.301$).

Targeted copy number aberration analysis (T_CNA)

The 450-targeted regions (T_CNA) feature was constructed similarly to the common genome-wide CNA feature. Read counts across these 450 regions were normalized, and copy numbers were estimated using the *QDNAseq* package, with minor modifications to adapt the settings for the targeted regions.

Step 3: The cfDNA features were input into a machine learning algorithm to generate predictive outcomes. The discovery cohort dataset was divided into training and testing sets, with the training set further split into two pairwise comparisons: cancer versus benign and cancer versus healthy, for feature selection. To minimize redundancy, Kendall rank correlation ($r \geq 0.7$) was applied to remove highly correlated features. Cross-validation with 10 folds was used to identify stable features in each pairwise comparison. The *SelectKBest* method was employed to filter the top 500 features in each iteration, based on scores from the"*f_classif*"function. Stable features were selected if they appeared consistently across all 500 feature sets from the iterations. The final feature set was formed by combining the selected features from both pairwise comparisons.

Five machine learning algorithms—Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGB)—were evaluated. To identify the best algorithm, we used the *HyperclassifierSearch* function with tenfold cross-validation and default parameters. After selecting the optimal algorithm, its hyperparameters were fine-tuned, and a classification model was constructed to differentiate BC patients from benign and healthy individuals. To evaluate model stability and assess the risk of overfitting, we performed bootstrap resampling with 1000 training epochs on the discovery cohort. In each of 10 bootstrap iterations, a dataset of equal size was generated by sampling with replacement from the original data. A classifier was trained on each resampled dataset for 1000 boosting rounds. During training, both log loss and accuracy were tracked across epochs using the bootstrap training data and the corresponding out-of-bag (OOB) samples for validation, providing insight into convergence behavior and potential overfitting.

Statistical analysis

The Mann–Whitney U test was employed to evaluate statistically significant differences between BC and benign samples, as well as between BC and healthy samples. DNA methylation data from 1075 breast tumor tissues and 124 adjacent non-tumor tissues were obtained from the TCGA-BRCA project via the TCGA data portal (https://portal.gdc.cancer.gov/). Differentially methylated CpG sites were identified using the Student's t test. To account for multiple comparisons, p-values were adjusted using the Benjamini–Hochberg correction, with

a significance threshold set at $\alpha \leq 0.05$. Sensitivity and specificity values between the test set and the validation cohort were compared using Fisher's exact test, with a significance threshold also set at $\alpha \leq 0.05$.

Receiver operating characteristic (ROC) curves were constructed, and the area under the ROC curve (AUC), along with 95% confidence intervals (CI), was calculated to evaluate the discriminative performance of features. We applied SHapley Additive exPlanations (SHAP) analysis to identify the top 20 most impactful features of the model [30]. All statistical analyses were conducted using Python version 3.11.2 (Python Software Foundation, USA).

Abbreviations

BC	Breast cancer
MRI	Magnetic resonance imaging
BI-RADS	Breast Imaging Reporting and Data System
CTCs	Circulating tumor cells
ctDNA	Circulating tumor DNA
cfDNA	Cell-free DNA
miRNAs	MicroRNAs
SPOT-MAS	Screening for the Presence of Tumor by DNA Methylation and Size
CNA	Copy number aberration
EM	End motif
ME21	Motif 21
AUC	Area under the ROC curve
FNA	Fine needle aspiration
TMD	Targeted methylation density
T_CNA	Targeted copy number aberration
DMRs	Differentially methylated regions
TLRs	Toll-like receptors
MRD	Minimal residual disease
TNM	Tumor, Node, Metastasis
LR	Logistic regression
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
XGB	Extreme Gradient Boosting
OOB	Out-of-bag
SHAP	SHapley Additive exPlanations

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-025-02371-z.

Additional file 1: Table S1-S4 and Figure S1-S5. Table S1. Detailed clinical information of all participants in the discovery cohort. Table S2. Summary of participants' clinical features in the validation cohort. Table S3. Detailed clinical information of all participants in the validation cohort. Table S4. Comparison of model performance. Fig.S1. Comparison of Area Under the Curve (AUC) values obtained from tenfold cross-validation across five machine learning algorithms (Error bars represent ± 1 standard deviation across tenfold cross-validation). Fig. S2. The pie chart shows the percentages according to stages in BC patients (A), subtypes in BC patients (B), and subtypes of benign patients (C). Fig. S3. SHapley Additive exPlanations (SHapley Additive exPlanations (SHAP) analysis reveals the feature importances in the machine learning model.) of the model in training set. Fig. S4. Overlapping genes between 450 target regions and TCGA database. Fig. S5. Log loss and accuracy across bootstrapped models on the discovery cohort.

Acknowledgements
We sincerely thank all participants for their contribution to this study, as well as the clinics and hospitals that provided support for patient consultations and sample collection.

Authors' contributions
KST, ALN, CHAT, TXJ, TLV1, THTN, TTT, MHT, NCT, TLL, THNN, NHT, TST, BTL, VPT, PTNN, KTN, VCH, TBL, HTPN, LHDN, NMP, TVP, TTTD, THD, HST performed patient consultancy and screening. TTVV, THHN, VTCN, DHV, TTT, TMTT, VUT, MPL, TLV2 performed formal analysis. TTVV, THT, THN, XVN, NNTD performed data curation. DSN, HG, MDP, HNN, LST performed the methodology. DSN, HG, MDP, HNN, DHV, LST performed conceptualization. TTVV, THGN, LST performed writing-original draft. TTVV, THGN, LST performed writing-review and editing. TLV1 corresponding to Thi Loan Vo and TLV2 corresponding to Thi Luyen Vu. All authors read and approved the final manuscript.

Funding
This work was supported by Gene Solutions.

Data availability
DNA-seq from this study have been deposited in NCBI SRA as a Bioproject with accession number PRJNA1296750 [73]. All data generated or analyzed during this study are included in this published article, its supplementary information files, and publicly available repositories. The source code for this study is publicly available in the GitHub repository at https://github.com/tuongvi2259/Breast_Cancer_Project and Zenodo at https://doi.org/10.5281/zenodo.16146086[74].

Declarations

Ethics approval and consent to participate
The research protocol was approved by the Ethics Committees of all participating institutions, with ethics review number 294/BVUB-HĐĐĐ for the discovery cohort and 460/HĐĐĐ-ĐHYD for the validation cohort. Informed written consent was obtained from each participant in accordance with the Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors, including LST, DSN, HG, MDP, and HNN, hold equity in Gene Solutions. We confirm that this does not impact on our compliance with the journal's policies regarding data and material sharing.

Author details
[1]Medical Genetics Institute, Ho Chi Minh, Vietnam. [2]Ho Chi Minh City Oncology Hospital, Ho Chi Minh, Vietnam. [3]Medic Medical Center, Ho Chi Minh, Vietnam. [4]Thai Nguyen National General Hospital, Thai Nguyen, Vietnam. [5]Buon Ma, Thuot Medical University Hospital, Buon Ma Thuot, Vietnam. [6]Can Tho Oncology Hospital, Can Tho, Vietnam. [7]Da Nang Oncology Hospital, Da Nang, Vietnam. [8]Nghe An Oncology Hospital, Nghe An, Vietnam.

Received: 20 January 2025 Accepted: 1 August 2025
Published online: 20 August 2025

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63.
2. Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. Breast. 2022;66:15–23.
3. Merino Bonilla JA, Torres Tabanera M, Ros Mendoza LH. Breast cancer in the 21st century: from early detection to new therapies. Radiologia. 2017;59(5):368–79.

4. Gerami R, Sadeghi Joni S, Akhondi N, Etemadi A, Fosouli M, Eghbal AF, et al. A literature review on the imaging methods for breast cancer. Int J Physiol Pathophysiol Pharmacol. 2022;14(3):171–6.
5. Gilbert FJ, Tucker L, Young KC. Digital breast tomosynthesis (DBT): a review of the evidence for use as a screening tool. Clin Radiol. 2016;71(2):141–50.
6. Tabár L, Vitak B, Chen TH, Yen AM, Cohen A, Tot T, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology. 2011;260(3):658–63.
7. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DS, Kerlikowske K, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. Radiology. 2017;283(1):49–58.
8. Salem DS, Kamal RM, Mansour SM, Salah LA, Wessam R. Breast imaging in the young: the role of magnetic resonance imaging in breast cancer screening, diagnosis and follow-up. J Thorac Dis. 2013;5 Suppl 1(Suppl 1):S9–18.
9. Boral S, Jagtap SV. Clinicohistopathological study of benign breast lesions in surgically excised specimens in a tertiary care hospital. J Cancer Res Ther. 2023;19(Supplement):S116–20.
10. Klinger K, Bhimani C, Shames J, Sevrukov A. Fibroadenoma: from imaging evaluation to treatment. 2019.
11. Brown AL, Vijapura C, Patel M, De La Cruz A, Wahab R. Breast cancer in dense breasts: detection challenges and supplemental screening opportunities. Radiographics. 2023;43(10): e230024.
12. Jaglan P, Dass R, Duhan M. Breast cancer detection techniques: issues and challenges. Journal of The Institution of Engineers (India): Series B. 2019;100(4):379–86.
13. Merkkola-von Schantz PA, Jahkola TA, Krogerus LA, Hukkinen KS, Kauhanen SM. Should we routinely analyze reduction mammaplasty specimens? J Plast Reconstr Aesthet Surg. 2017;70(2):196–202.
14. Duque G, Manterola C, Otzen T, Arias C, Palacios D, Mora M, et al. Cancer biomarkers in liquid biopsy for early detection of breast cancer: a systematic review. Clin Med Insights Oncol. 2022;16: 11795549221134831.
15. Freitas AJA, Causin RL, Varuzza MB, Calfa S, Hidalgo Filho CMT, Komoto TT, et al. Liquid biopsy as a tool for the diagnosis, treatment, and monitoring of breast cancer. Int J Mol Sci. 2022. https://doi.org/10.3390/ijms23179952.
16. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science. 2018;359(6378):926–30.
17. Manoochehri M, Borhani N, Gerhäuser C, Assenov Y, Schönung M, Hielscher T, et al. DNA methylation biomarkers for noninvasive detection of triple-negative breast cancer using liquid biopsy. Int J Cancer. 2023;152(5):1025–35.
18. Liu J, Zhao H, Huang Y, Xu S, Zhou Y, Zhang W, et al. Genome-wide cell-free DNA methylation analyses improve accuracy of non-invasive diagnostic imaging for early-stage breast cancer. Mol Cancer. 2021;20(1):36.
19. Zhang Y, Cheng L. DNA methylation analysis of peripheral blood mononuclear cells in diagnosing breast cancer from benign breast lesions. J Transl Med. 2024;22(1):1070.
20. Adashek JJ, Kato S, Lippman SM, Kurzrock R. The paradox of cancer genes in non-malignant conditions: implications for precision medicine. Genome Med. 2020;12(1):16.
21. Pham TMQ, Phan TH, Jasmine TX, Tran TTT, Huynh LAK, Vo TL, et al. Multimodal analysis of genome-wide methylation, copy number aberrations, and end motif signatures enhances detection of early-stage breast cancer. Front Oncol. 2023. https://doi.org/10.3389/fonc.2023.1127086.
22. Nguyen VTC, Nguyen TH, Doan NNT, Pham TMQ, Nguyen GTH, Nguyen TD, et al. Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for multi-cancer early detection and localization. Elife. 2023;12: RP89083.
23. Chandrananda D, Thorne NP, Bahlo M. High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. BMC Med Genomics. 2015;8:29.
24. Esteller M. Epigenetics in cancer. N Engl J Med. 2008;358(11):1148–59.
25. An W, Lin H, Ma L, Zhang C, Zheng Y, Cheng Q, et al. Progesterone activates GPR126 to promote breast cancer development via the Gi pathway. Proc Natl Acad Sci U S A. 2022;119(15): e2117004119.
26. Zhu KY, Tian Y, Li YX, Meng QX, Ge J, Cao XC, et al. The functions and prognostic value of Krüppel-like factors in breast cancer. Cancer Cell Int. 2022;22(1): 23.
27. Shi S, Xu C, Fang X, Zhang Y, Li H, Wen W, et al. Expression profile of Toll-like receptors in human breast cancer. Mol Med Rep. 2020;21(2):786–94.
28. Tsyganov MM, Ibragimova MK, Garbukov EY, Tsydenova IA, Gaptulbarova KA, Dolgasheva DS, et al. Predictive and prognostic significance of mRNA expression and DNA copies aberrations of ERCC1, RRM1, TOP1, TOP2A, TUBB3, TYMS, and GSTP1 genes in patients with breast cancer. Diagnostics. 2022. https://doi.org/10.3390/diagnostics12020405.
29. Seal DB, Das V, Goswami S, De RK. Estimating gene expression from DNA methylation and copy number variation: a deep learning regression model for multi-omics integration. Genomics. 2020;112(4):2833–41.
30. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin Transl Sci. 2024;17(11): e70056.
31. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018. https://doi.org/10.1126/science.aav1898.
32. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014;6(224): 224ra24.
33. Fumagalli C, Barberis M. Breast cancer heterogeneity. Diagnostics (Basel). 2021. https://doi.org/10.3390/diagnostics11091555.
34. Matsutani A, Udagawa C, Matsunaga Y, Nakamura S, Zembutsu H. Liquid biopsy for the detection of clinical biomarkers in early breast cancer: new insights and challenges. Pharmacogenomics. 2020;21(5):359–67.
35. Tay TKY, Tan PH. Liquid biopsy in breast cancer: a focused review. Arch Pathol Lab Med. 2021;145(6):678–86.
36. Lone SN, Nisar S, Masoodi T, Singh M, Rizwan A, Hashem S, et al. Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments. Mol Cancer. 2022;21(1):79.
37. Guo L, Kong D, Liu J, Zhan L, Luo L, Zheng W, et al. Breast cancer heterogeneity and its implication in personalized precision therapy. Exp Hematol Oncol. 2023;12(1):3.
38. Garofoli M, Maiorano BA, Bruno G, Giordano G, Falagario UG, Necchi A, et al. Circulating tumor DNA: a new research frontier in urological oncology from localized to metastatic disease. Eur Urol Oncol. 2024;8(3):805-17.
39. Fridlyand J, Snijders AM, Ylstra B, Li H, Olshen A, Segraves R, et al. Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer. 2006;6: 96.
40. Nakopoulou L, Panayotopoulou EG, Giannopoulou I, Tsirmpa I, Katsarou S, Mylona E, et al. Extra copies of chromosomes 16 and X in invasive breast carcinomas are related to aggressive phenotype and poor prognosis. J Clin Pathol. 2007;60(7):808–15.
41. Grifoni A, Alonzi T, Alter G, Noonan DM, Landay AL, Albini A, et al. Impact of aging on immunity in the context of COVID-19, HIV, and tuberculosis. Front Immunol. 2023;14:1146704.
42. Figueroa JD, Gierach GL, Duggan MA, Fan S, Pfeiffer RM, Wang Y, et al. Risk factors for breast cancer development by tumor characteristics among women with benign breast disease. Breast Cancer Res. 2021;23(1):34.
43. Huang A, Cao S, Tang L. The tumor microenvironment and inflammatory breast cancer. J Cancer. 2017;8(10):1884–91.
44. Angeles AK, Janke F, Bauer S, Christopoulos P, Riediger AL, Sültmann H. Liquid biopsies beyond mutation calling: genomic and epigenomic features of cell-free DNA in cancer. Cancers (Basel). 2021. https://doi.org/10.3390/cancers13225615.
45. Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. Cancer Discov. 2020;10(5):664–73.
46. Jin C, Liu X, Zheng W, Su L, Liu Y, Guo X, et al. Characterization of fragment sizes, copy number aberrations and 4-mer end motifs in cell-free DNA of hepatocellular carcinoma for enhanced liquid biopsy-based cancer detection. Mol Oncol. 2021;15(9):2377–89.
47. Serpas L, Chan RWY, Jiang P, Ni M, Sun K, Rashidfarrokhi A, et al. Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. Proc Natl Acad Sci U S A. 2019;116(2):641–9.
48. Zhu D, Wang H, Wu W, Geng S, Zhong G, Li Y, et al. Circulating cell-free DNA fragmentation is a stepwise and conserved process linked to apoptosis. BMC Biol. 2023;21(1):253.

49. Zhu Z, Chen T, Zhang M, Shi X, Yu P, Liu J, et al. Dynamic profiling of cell-free DNA fragmentation uncovers postprandial metabolic and immune alterations. Hum Genomics. 2025;19(1):27.
50. Markus H, Chandrananda D, Moore E, Mouliere F, Morris J, Brenton JD, et al. Refined characterization of circulating tumor DNA through biological feature integration. Sci Rep. 2022;12(1):1928.
51. Zhou Z, Ma ML, Chan RWY, Lam WKJ, Peng W, Gai W, et al. Fragmentation landscape of cell-free DNA revealed by deconvolutional analysis of end motifs. Proc Natl Acad Sci U S A. 2023;120(17): e2220982120.
52. Bai Q, He X, Hu T. Pan-cancer analysis of the deoxyribonuclease gene family. Mol Clin Oncol. 2023;18(3): 19.
53. Szyf M, Pakneshan P, Rabbani SA. DNA methylation and breast cancer. Biochem Pharmacol. 2004;68(6):1187–97.
54. Sun W, Bunn P, Jin C, Little P, Zhabotynsky V, Perou CM, et al. The association between copy number aberration, DNA methylation and gene expression in tumor samples. Nucleic Acids Res. 2018;46(6):3009–18.
55. Gao C, Li H, Liu C, Wu J, Zhou C, Liu L, et al. Determination of genetic and epigenetic modifications-related prognostic biomarkers of breast cancer: genome high-throughput data analysis. J Oncol. 2021;2021: 2021:2143362.
56. Shang M, Chang C, Pei Y, Guan Y, Chang J, Li H. Potential management of circulating tumor DNA as a biomarker in triple-negative breast cancer. J Cancer. 2018;9(24):4627–34.
57. Simon RM, Subramanian J, Li MC, Menezes S. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. Brief Bioinform. 2011;12(3):203–14.
58. Yates LA, Aandahl Z, Richards SA, Brook BW. Cross validation for model selection: a review with examples from ecology. Ecol Monogr. 2023;93(1): e1557.
59. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020;31(6):745–59.
60. Harbeck N. Insights into biology of luminal HER2 vs. enriched HER2 subtypes: therapeutic implications. Breast. 2015;24(Suppl 2):S44–8.
61. Magbanua MJM, Brown Swigart L, Ahmed Z, Sayaman RW, Renner D, Kalashnikova E, et al. Clinical significance and biology of circulating tumor DNA in high-risk early-stage HER2-negative breast cancer receiving neoadjuvant chemotherapy. Cancer Cell. 2023;41(6):1091-102.e4.
62. Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst. 2009;101(10):736–50.
63. Coombes RC, Page K, Salari R, Hastings RK, Armstrong A, Ahmed S, et al. Personalized detection of circulating tumor DNA antedates breast cancer metastatic recurrence. Clin Cancer Res. 2019;25(14):4255–63.
64. Dawson SJ, Tsui DW, Murtaza M, Biggs H, Rueda OM, Chin SF, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med. 2013;368(13):1199–209.
65. Wiratkapun C, Bunyapaiboonsri W, Wibulpolprasert B, Lertsithichai P. Biopsy rate and positive predictive value for breast cancer in BI-RADS category 4 breast lesions. J Med Assoc Thai. 2010;93(7):830–7.
66. Spick C, Bickel H, Polanec SH, Baltzer PA. Breast lesions classified as probably benign (BI-RADS 3) on magnetic resonance imaging: a systematic review and meta-analysis. Eur Radiol. 2018;28(5):1919–28.
67. Sánchez-Martín V, López-López E, Reguero-Paredes D, Godoy-Ortiz A, Domínguez-Recio ME, Jiménez-Rodríguez B, et al. Comparative study of droplet-digital PCR and absolute Q digital PCR for ctDNA detection in early-stage breast cancer patients. Clin Chim Acta. 2024;552: 117673.
68. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570(7761):385–9.
69. Han BW, Cai GX, Liu Q, Yang X, Guo ZW, Huang LM, et al. Noninvasive discrimination of benign and malignant breast lesions using genome-wide nucleosome profiles of plasma cell-free DNA. Clin Chim Acta. 2021;520:95–100.
70. Janni W, Rack B, Friedl TWP, Hartkopf AD, Wiesmüller L, Pfister K, et al. Detection of minimal residual disease and prediction of recurrence in breast cancer using a plasma-only circulating tumor DNA assay. ESMO Open. 2025;10(4): 104296.
71. Clarke CA, Keegan TH, Yang J, Press DJ, Kurian AW, Patel AH, et al. Age-specific incidence of breast cancer subtypes: understanding the black-white crossover. J Natl Cancer Inst. 2012;104(14):1094–101.
72. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 2018. https://doi.org/10.1126/scitranslmed.aat4921.
73. Multimodal analysis of cell-free DNA enhances differentiation of early-stage breast cancer from benign lesions and healthy individuals. NCBI Bioproject Accession: PRJNA1296750. 2025. https://www.ncbi.nlm.nih.gov/search/all/?term=PRJNA1296750.
74. Van TTV. Multimodal analysis of cell-free DNA enhances differentiation of early-stage breast cancer from benign lesions and healthy individuals. Zenodo; 2025.

Publisher's Note