Original Article Open Access

Archives Of Breast Cancer

DOI: 10.32768/abc.5842169374-058

An Integrated Circulating MicroRNA–EPCAM–Protein Liquid Biopsy Signature for Early Breast Cancer Detection in Iraqi Women: A Multilayer Molecular Framework

Adam Neama Al-Ebadi*[a], Rasha Abd Ali Al-Khalidi[b,c]

[a] Biotechnology Department, College of Science, University of Baghdad, Baghdad, Iraq
[b] Department of Biotechnology, College of Science, University of Baghdad, Baghdad, Iraq
[c] Biomedical Application Department, Artificial Intelligence, University of Baghdad, Baghdad, Iraq

* **Corresponding Author:** Adam Neama Al-Ebadi. Email: adam.n@sc.uobaghdad.edu.iq

ARTICLE INFO

Received:
06 May 2026
Revised:
05 July 2026
Accepted:
28 July 2026

Keywords:
breast neoplasms, microRNAs, liquid biopsy, epithelial cell adhesion molecule, biomarkers, tumor, CA-15-3 antigen, early detection of cancer, logistic models

ABSTRACT

Background: Early breast cancer is characterized by subtle molecular alterations that may escape detection by conventional single-marker assays. Liquid biopsy offers a noninvasive approach to capturing this complexity, yet most circulating biomarker studies remain limited to a single molecular layer. This study aimed to develop and evaluate an integrated circulating signature combining microRNAs (miRs), epithelial cell adhesion molecule (EPCAM)–associated epithelial dissemination signals, and tumor-associated serum proteins for early breast cancer detection in Iraqi women.

Methods: This case-control study included 75 women (25 patients with early breast cancer and 50 healthy controls) recruited at the National Hospital for Oncology and Hematology, Najaf, Iraq, between February and August 2025. Three computationally prioritized circulating miRs (hsa-miR-4734, hsa-miR-5787, hsa-miR-4532) identified from the public serum miR dataset GSE73002 were quantified by SYBR Green quantitative reverse transcription-polymerase chain reaction (qRT-PCR) using hsa-miR-16 as a reference. EPCAM expression was assessed by qRT-PCR with GAPDH normalization in a subset of 13 participants. Serum cancer antigen 15-3 (CA15-3), carcinoembryonic antigen (CEA), and mammaglobin-1 (MGB1) were measured by sandwich enzyme-linked immunosorbent assay (ELISA). An integrated multimarker classifier was constructed using leave-one-out cross-validated logistic regression.

Results: Levels of hsa-miR-4734 (4.2-fold; $P < .001$; area under the receiver operating characteristic curve (AUC), 0.75), hsa-miR-5787 (4.0-fold; $P = .01$; AUC, 0.68), and hsa-miR-4532 (5.0-fold; $P = .09$; AUC, 0.62) were upregulated in patients. EPCAM expression was also elevated in a 13-participant subset (7 patients, 6 controls; $P = .04$; AUC, 0.86; 95% confidence interval (CI), 0.50–0.97). The integrated 5-marker classifier achieved a leave-one-out cross-validated AUC of 0.91 (95% CI, 0.77–0.96), sensitivity of 84%, and specificity of 90%.

Conclusion: An integrated circulating signature encompassing posttranscriptional regulation, epithelial shedding, and tumor-associated protein release may support noninvasive early detection of breast cancer. Diagnostic performance estimates should be considered exploratory pending external prospective validation in age-matched cohorts.

© 2026 The Author(s). This open-access article is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License, which permits sharing and adapting the material for any purpose except commercial purposes, provided appropriate credit is given, a link to the license is provided, and any changes are indicated.

INTRODUCTION

Breast cancer is the most common cancer among women and the leading cause of cancer-related death in women worldwide. While incidence and mortality are generally increasing, the extent of this trend varies significantly across low- and middle-income countries, including Iraq.[1] Despite significant progress in screening and adjuvant therapy, prognosis remains largely determined by disease stage at diagnosis, which continues to be delayed in a substantial proportion of cases until locoregional or systemic metastatic spread has occurred.[2] Incipient tumors induce molecular changes that are often

subtle, heterogeneous, and scattered across multiple regulatory layers; this limitation of conventional single-analyte diagnostic methods motivates the search for biologically integrated biomarker frameworks.[3]

In recent years, liquid biopsy has emerged as a game-changer for cancer diagnostics with its ability to noninvasively interrogate tumor-derived content from peripheral blood.[4,5] In addition to their proximal involvement in posttranscriptional regulation of oncogenes and tumor suppressors, circulating microRNAs (miRs) exhibit distinctive properties favoring their use in early diagnosis. They are highly abundant in plasma and serum and exhibit exceptional stability, attributed to their packaging within extracellular vesicles or binding to Argonaute-lipoprotein complexes, which protects them from RNase degradation in the circulation.[6,7] Research investigating this regulatory facet in Iraqi cohorts has begun to outline this dimension; for example, miR-10b was persistently upregulated across stages of breast cancer in patients recruited through Baghdad oncology centers, which supports its diagnostic and prognostic utility.[8]

However, current diagnostic and prognostic applications of circulating miRs have been limited by their reliance on single markers; this approach overlooks the holistic, multilayered view of early tumor biology. Mounting evidence now suggests that early breast cancer is more appropriately viewed as a system-wide disruption rather than simple disturbance of an individual molecule. Multimarker panels that combine miRs with traditional protein biomarkers have consistently outperformed single analytes in published cohorts.[9–12] In parallel, work in the same Iraqi clinical context has examined relationships between serum and tissue markers in patients with breast cancer, supporting evaluation of complementary molecular layers.[13]

In addition to regulatory and antigenic markers, the cellular dimension of liquid biopsy introduces an additional layer. EPCAM is expressed at very low levels in normal epithelial tissues but upregulated on circulating tumor cells (CTCs) and shed in CTC clusters during early dissemination. EPCAM-based immunocapture has been demonstrated to reliably detect CTCs across nearly all breast cancer subtypes, and the presence of high EPCAM-expressing CTC subpopulations has emerged as a negative prognostic indicator.[14–16] Notably, in standard PCR-based assessments, we can detect EPCAM transcripts in both healthy and pathological samples; the expression ratio captures epithelial dissemination and CTC-associated activity, rather than a binary score.

Informed by these theoretical and empirical perspectives, we here incorporate 3 molecular dimensions of liquid biopsy—circulating miRs, the EPCAM-associated CTC layer, and tumor-associated serum proteins—into a unified diagnostic framework. A bioinformatics analysis of the GSE73002 serum microRNA dataset prioritized 3 microRNAs (hsa-miR-4734, hsa-miR-5787, and hsa-miR-4532) that were subsequently validated by qRT-PCR in an Iraqi case-control cohort. EPCAM expression was profiled on CTCs as a measure of epithelial dissemination, while ELISA determined serum levels of CA15-3, CEA, and MGB1 to capture the protein layer. The aim was not simply to validate the differential expression of each analyte independently, but to test the hypothesis that early breast cancer creates a coherent molecular signature across multiple biological layers and that integrating these complementary signals may improve overall diagnostic characterization.

METHODS

Study design and participants

This case-control study enrolled 75 female participants: 25 patients with pathologically confirmed early-onset breast cancer (aged ≤ 50 years) and 50 healthy adult controls. Controls were selected based on the absence of cancer and chronic inflammatory disease; no formal matching was performed. Ages ranged from 13 to 60 years overall, with a mean (standard deviation (SD)) age of 46.7 (8.8) years for patients and 32.3 (12.4) years for controls. Patients were recruited from the National Hospital for Oncology and Hematology, Najaf, Iraq, between February and August 2025. The recruitment window was defined by the project timeline and clinical availability; no seasonal selection of biomarker expression was expected, although the 6-month window did not preclude systematic variation in referral patterns that could influence cohort composition. Inclusion criteria for patients included histopathologically confirmed primary breast carcinoma, no history of antitumor treatment (chemotherapy, radiotherapy, or surgery) prior to sample acquisition, and complete clinical records. Patients with hematological diseases, autoimmune conditions, or any second primary malignancy were excluded. Healthy controls had no personal history of cancer and no chronic inflammatory or autoimmune disease at the time of recruitment. Place of residence (urban or rural) was recorded for all participants based on self-report at the time of enrollment.

The study was approved by the Ethics Committee of the University of Baghdad, College of Science, Department of Biotechnology (Approval No. 3763E/24/10/2024), and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants. Sample size was determined a priori using the Hanley-McNeil method for receiver operating characteristic (ROC)–based power estimation, targeting 80% power to detect an AUC of 0.75 or greater against a null of 0.50 (2-sided $\alpha = .05$) for the primary single-biomarker candidate (CA15-3), which required a minimum of 21 cases and 42 controls; the enrolled 25 patients and 50 controls met this threshold for the single-biomarker analysis. This calculation does not extend to the 5-predictor multivariable model. For that model, the events-per-variable ratio (EPV = 5) falls below the conventional 10-EPV standard, and no separate a priori sample size justification is available. Leave-one-out cross-validation (LOOCV)

was employed as the recommended small-sample estimator of out-of-sample performance; however, LOOCV reduces rather than eliminates overfitting risk, and multivariable discrimination metrics should be interpreted as exploratory pending adequately powered external validation.

Bioinformatic prioritization of candidate microRNAs

Candidate circulating microRNAs were selected through differential expression analysis of the publicly available serum miR dataset GSE73002 (GPL18941 platform). Differential expression analysis identified 47 microRNAs meeting joint significance thresholds ($|\log_2$ fold change$| \geq 1$; Benjamini-Hochberg-adjusted $P < .05$), of which 31 were upregulated and 16 downregulated in breast cancer serum relative to healthy controls. Secondary prioritization applied 3 additional filters in sequence: (1) biological relevance—target gene enrichment in cancer-relevant Kyoto Encyclopedia of Genes and Genomes (KEGG)/Reactome pathways (PI3K-Akt, epithelial-mesenchymal transition (EMT), apoptosis); (2) serum expression detectability—exclusion of candidates with mean intensity below the platform detection threshold in the validation set; and (3) PCR feasibility—availability of validated stem-loop primer sequences. Three miRs (hsa-miR-4734, hsa-miR-5787, hsa-miR-4532) were the highest-ranked candidates satisfying all 3 secondary filters and were advanced to experimental validation.

Sample collection and RNA extraction

Peripheral blood samples were collected from each participant using 2 tubes: a 5 mL ethylenediaminetetraacetic acid (EDTA)–anticoagulated tube for RNA extraction and gene expression analyses, and a separate clot activator tube for serum biomarker measurements. Samples were processed within 2 hours of collection. Total RNA, including the small RNA fraction, was extracted using TRIzol reagent (Invitrogen) according to the manufacturer's protocol. RNA concentration and purity were assessed on a NanoDrop spectrophotometer; samples with an A_{260}/A_{280} ratio between 1.8 and 2.1 were considered acceptable. RNA integrity was verified by agarose gel electrophoresis.

Plasma samples obtained from the EDTA tubes were visually inspected for hemolysis prior to RNA extraction. Samples displaying macroscopic pink or red discoloration indicative of erythrocyte lysis were excluded from downstream analysis. We acknowledge that spectrophotometric quantification of the hemolysis index or assessment by the miR-23a/miR-451a ratio would provide a more sensitive control; the absence of this formal assay represents a preanalytical limitation of the current study.

Reverse transcription and quantitative PCR

MicroRNA quantification used a stem-loop reverse transcription strategy. For each target miR (hsa-miR-4734, hsa-miR-5787, hsa-miR-4532) and the endogenous reference hsa-miR-16, miR-specific stem-loop primers were used to generate cDNA. Quantitative PCR was performed on a Bio-Rad CFX96 system with SYBR Green chemistry according to the following cycling conditions: 95 °C for 3 minutes; 40 cycles at 95 °C for 15 seconds, followed by annealing (60 °C, 30 seconds), and extension (72 °C, 30 seconds); this cycle was followed by melt-curve analysis to verify amplicon specificity. Oligo(dT) primers were used to reverse-transcribe EPCAM mRNA, and cycling conditions were identical to those used for GAPDH quantification as the housekeeping reference. Each reaction was conducted in technical duplicates, and no-template controls were included on each plate. Relative expression was calculated using the comparative $2^{-\Delta\Delta C_t}$ method in which the mean ΔCt of the control group is used as a calibrator. The hsa-miR-16 was selected as the endogenous normalizer on the basis of its established use in serum microRNA studies and its application in the GSE73002 source dataset, thereby ensuring methodological consistency across bioinformatic discovery and experimental validation phases. We acknowledge that formal stability validation using NormFinder, geNorm, or exogenous spike-in approaches (eg, cel-miR-39) was not performed; this represents a limitation, and future studies should include multireference stability assessment before adopting a single normalizer.

Serum protein biomarker quantification

Serum obtained from the serum-separator (gel) tubes was centrifuged (2000g, 10 minutes) and stored at −80 °C until analysis. The CA15-3 was quantified using a sandwich ELISA kit (ab277214; Abcam; sensitivity: 0.5 U/mL; intra-assay coefficient of variation (CV) < 10%); CEA using a commercial sandwich ELISA (E-EL-H0109; Elabscience; sensitivity: 0.09 ng/mL; intra-assay CV < 10%); and MGB1 (SCGB2A2) using a sandwich ELISA (SEF536Hu; Cloud-Clone Corp; sensitivity: 12.4 pg/mL; intra-assay CV < 10%). Manufacturer-declared cross-reactivity data confirmed no significant interference from closely related antigens. All samples were measured in duplicate, and intra- and interassay CVs were maintained below 10%. Standard curves were constructed by 4-parameter logistic regression, and unknown sample concentrations were derived from the calibrated curves.

Statistical analysis

All analyses were performed in R version 4.3, and Python version 3.12. Normality of continuous variables was assessed by the Shapiro-Wilk test. Group differences were tested by the Welch *t* test or Mann-Whitney U test. Diagnostic performance was evaluated by ROC curve analysis with AUCs and 95% CIs computed by the DeLong method. Optimal diagnostic cutoffs for each individual biomarker and for the integrated model were determined by maximizing the Youden index (sensitivity + specificity − 1) on the ROC curve. An integrated multimarker classifier was constructed using ridge (L2-penalized) logistic regression, with the regularization

parameter (λ) selected via internal cross-validation to minimize deviance while controlling for the limited EPV ratio, with performance estimated by LOOCV. Pairwise correlations among biomarkers are reported as uncorrected *P* values, given their descriptive, hypothesis-generating intent; correction for 21 simultaneous tests (eg, a Bonferroni threshold of $P < .002$) would reduce the number of statistically significant correlations and should be applied in confirmatory analyses. Reporting of this observational case-control study adheres to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist.

Age was not included as an additional covariate in the penalized multivariable model, as doing so would further reduce the EPV ratio (from 5 to ≈ 4.2) in an already EPV-constrained model; the association between age and each individual biomarker was instead assessed post hoc (see Discussion), and age-matched or propensity-adjusted controls are identified as a requirement for future confirmatory studies. To assess potential model optimism and overfitting, 2 additional internal validation procedures were performed. First, bootstrap optimism correction was conducted using 1000 bootstrap resamples according to Harrell's framework, with optimism estimated as the average difference between apparent and bootstrap test performance and subsequently subtracted from the apparent AUC. Second, nested cross-validation was implemented using an outer LOOCV loop and an inner training loop for model fitting, providing an unbiased estimate of model generalizability. These analyses were performed for the integrated liquid biopsy model that combines circulating miRs and protein biomarkers.

RESULTS

Cohort characteristics

Demographic and clinical characteristics of the study cohort are summarized in Table 1. The 25 patients with breast cancer were significantly older than the healthy controls (mean (SD) age, 46.7 (8.8) vs 32.3 (12.4) years; $P < .001$), reflecting the age distribution of breast cancer presentation in this clinical setting. A family history of breast cancer was reported by 32% of the patients compared with 14% of controls. Tumor staging at presentation was predominantly T2 (68%), with T1 in 20% and locally advanced T3/T4 disease in 12%; nodal involvement (N1–N3) was documented in 40% of the patients. Hormone receptor profiling was evaluable in 20 of 25 patients: 10 (50%) were estrogen receptor/progesterone receptor (ER/PR)-positive and human epidermal growth factor receptor 2 (HER2)-negative, 6 (30%) were HER2-positive, and 4 (20%) were triple-negative; receptor status was unavailable for the remaining 5 patients.

Bioinformatic prioritization of circulating microRNAs

Differential expression analysis of GSE73002 identified a panel of upregulated and downregulated circulating microRNAs in breast cancer serum, with hsa-miR-4734, hsa-miR-5787, and hsa-miR-4532 emerging as highly ranked candidates after applying joint thresholds of $|\log_2 \text{fold change}| \geq 1$ and adjusted $P < .05$ (Figure 1, A). These 3 microRNAs met the secondary filters for pathway relevance, serum stability, and PCR feasibility (Figure 1, B). They were, therefore, advanced to experimental validation in the Iraqi cohort.

Coordinated upregulation of circulating microRNAs in early breast cancer

All 3 prioritized microRNAs were upregulated in patients compared with controls (Figure 2, A–C; Table 2). The hsa-miR-4734 showed a 4.2-fold mean increase (mean (SD), 8.26 (9.65) vs 1.97 (1.77); $P < .001$, Mann-Whitney U test), hsa-miR-5787 a 4.0-fold increase (6.21 (9.35) vs 1.55 (1.71); $P = .01$), and hsa-miR-4532 a 5.0-fold increase that approached but did not reach statistical significance (6.89 (10.74) vs 1.39 (0.93); $P = .09$).

EPCAM expression as an epithelial dissemination signal

EPCAM mRNA was quantifiable in both groups, consistent with its constitutive low-level expression in normal epithelial tissue. Importantly, the degree of expression varied significantly between groups: the mean fold change for patient samples was 7.37 (4.17) vs 2.74 (3.07) in controls (Welch $t = 2.37$; $P = .04$), corresponding to approximately 2.7-fold relative upregulation (Figure 2, D). This differential pattern further reinforces the view that elevated EPCAM mRNA detected in the blood of patients with cancer reflects a biological process linked to pathogenic epithelial shedding and active CTCs, rather than simply increased epithelial levels observed as part of basal expression differences in healthy donors. The EPCAM results should be interpreted with particular caution: the 13-participant analytical subset (7 patients, 6 controls) is insufficient for stable AUC estimation, as evidenced by the wide 95% CI (0.50–0.97), and the contribution of this layer to the integrated classifier remains the most provisional element of the proposed framework. Independent replication in a larger cohort with uniform biological material availability is essential before EPCAM can be considered a confirmed component of the diagnostic panel.

Serum protein biomarkers

Of the 3 protein analytes (Figure 2, E–G), CA15-3 was the strongest single discriminator: mean patient concentration (26.6 U/mL) was 4.9-fold higher than the control mean (5.47 U/mL; $P < .001$, Mann-Whitney U test). The MGB1 (SCGB2A2) serum levels were significantly lower in patients than in controls (mean log(ng/mL) 3.73 (0.22) vs 3.92 (0.26); Welch *t* test $P = .002$). The CEA level was mildly elevated in some cases yet not significantly different across the cohort ($P = .51$).

Table 1. Demographic and Clinical Characteristics of Patients and Controls

Characteristic	Patients (n = 25)	Controls (n = 50)	*P* value[a]
Age, mean (SD), y	46.7 (8.8)	32.3 (12.4)	< .001
Family history of breast cancer, No. (%)	8 (32)	7 (14)	.13
Residence, No. (%)			
Urban	14 (56)	50 (100)	NA
Rural	11 (44)	0	NA
Tumor stage at diagnosis, No. (%)			
T1	5 (20)	NA	NA
T2	17 (68)	NA	NA
T3/T4	3 (12)	NA	NA
Nodal involvement (N1–N3), No. (%)	10 (40)	NA	NA
Hormone receptor status, No. (%)			
ER/PR+HER2−	10 (50)	NA	NA
HER2+	6 (30)	NA	NA
Triple-negative	4 (20)	NA	NA

Abbreviations: ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; NA, not applicable; PR, progesterone receptor.
Hormone receptor percentages are based on 20 evaluable patients; receptor status was unavailable for 5 patients.
[a] Continuous variables compared by Welch *t* test; categorical variables by χ^2 test.

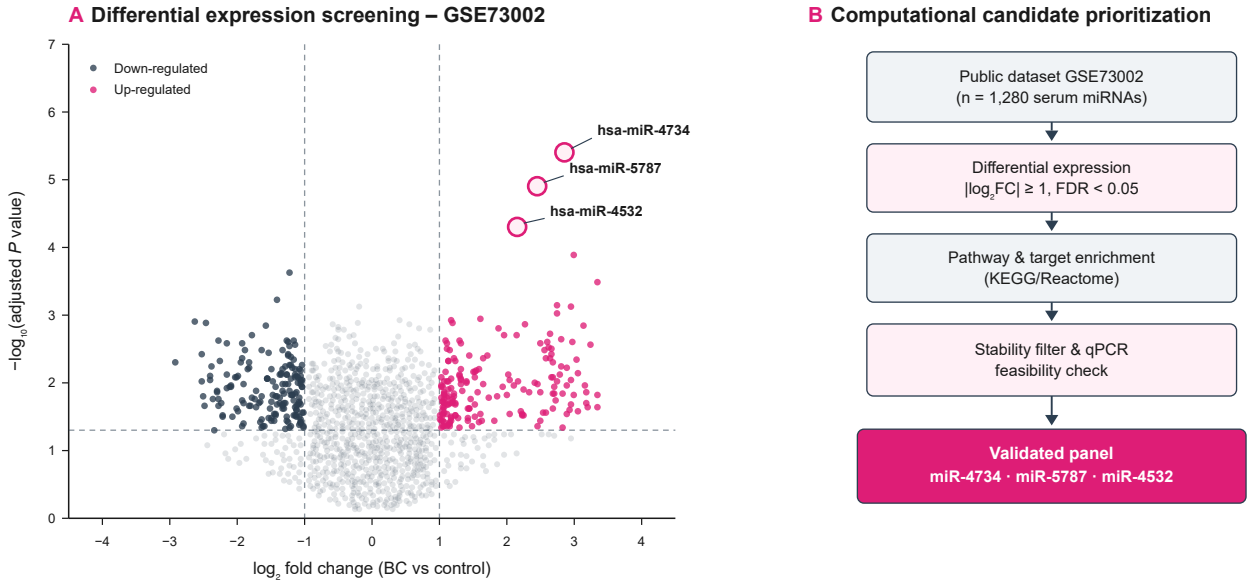


Figure 1. Bioinformatic Prioritization of Candidate Circulating MicroRNAs
A, Volcano plot of differential expression in the public serum microRNA dataset GSE73002 (1280 microRNAs profiled). Dashed lines indicate the significance thresholds (|$\log_2$ fold change| ≥ 1; adjusted P < .05). Red points indicate upregulated and blue points downregulated microRNAs; the 3 prioritized candidates, hsa-miR-4734, hsa-miR-5787, and hsa-miR-4532, are highlighted in gold. B, Stepwise computational prioritization from global microRNA profiling to the final 3-microRNA panel selected for experimental validation.

Diagnostic performance and integrated multimarker model

Receiver operating characteristic analysis revealed differential discriminatory power across the markers (Figure 3, A; Table 3). The CA15-3 was the strongest single discriminator (AUC, 0.92; 95% CI, 0.83–0.97; sensitivity, 76%; specificity, 96%), followed by EPCAM (AUC, 0.86; 95% CI, 0.50–0.97), hsa-miR-4734 (AUC, 0.75; 95% CI, 0.60–0.86), MGB1 (AUC, 0.73; 95% CI, 0.59–0.83), and hsa-miR-5787 (AUC, 0.68; 95% CI, 0.51–0.81). The hsa-miR-4532 (AUC, 0.62) and CEA (AUC, 0.55) showed only modest individual discrimination, consistent with their nonsignificant t and Mann-Whitney U test results.

Because age differed between cases and controls, an exploratory post hoc Spearman correlation analysis was performed to assess the relationship between age and the measured biomarkers. None of the 3 circulating miRs showed significant correlations with age (all |ρ| < 0.20, all P > .15). The CA15-3 demonstrated a weak positive correlation with age (ρ = 0.24, P = .04), whereas MGB1, CEA, and EPCAM showed no statistically significant associations (all P > .05). These findings suggest that age-related effects cannot be excluded but are unlikely to fully account for the observed biomarker differences between the groups.

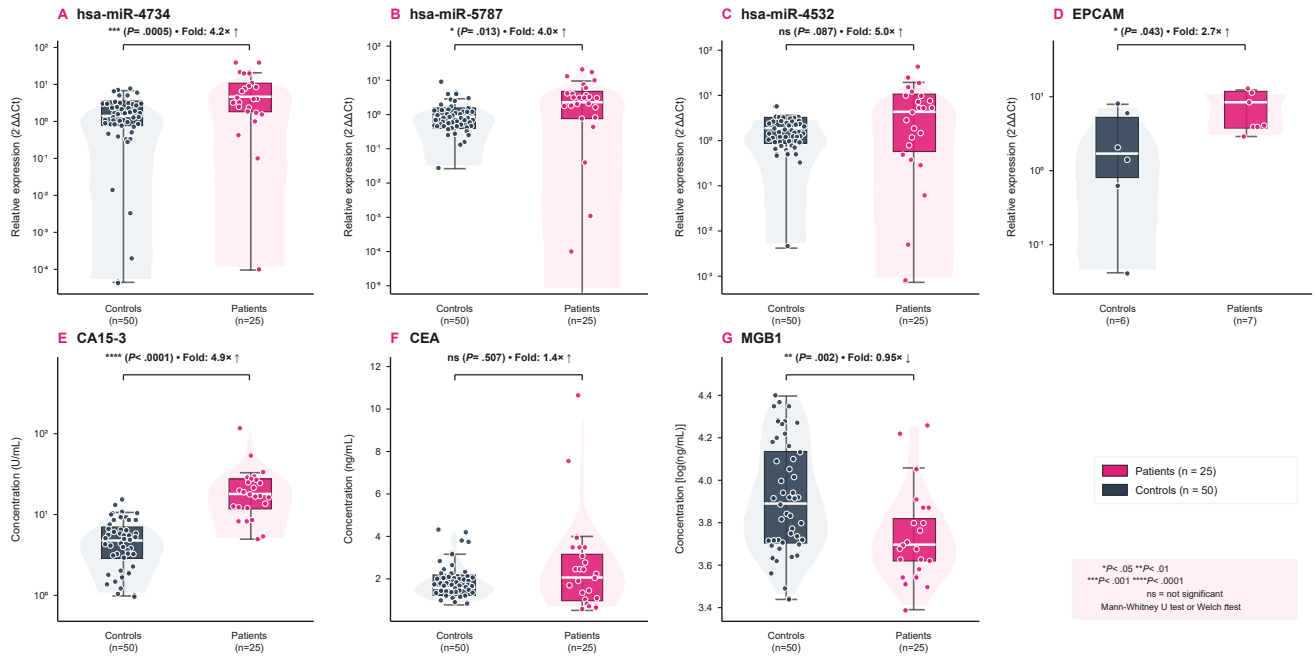


Figure 2. Distribution of Circulating Biomarkers in Patients With Early Breast Cancer vs Healthy Controls
A–C, Quantitative reverse transcription-polymerase chain reaction–derived relative expression ($2^{-\Delta\Delta C_t}$) of hsa-miR-4734, hsa-miR-5787, and hsa-miR-4532. D, Relative EPCAM expression in the 13-participant subset (7 patients, 6 controls). E–G, Serum concentrations of CA15-3, CEA, and mammaglobin-1 (MGB1), respectively. Boxes indicate interquartile ranges with medians; whiskers extend to 1.5 × IQR, and dots represent individual participants. Significance bars indicate *P* values from the Mann-Whitney U or Welch *t* test, as appropriate; fold changes are ratios of group means. **P* < .05; ***P* < .01; ****P* < .001; *****P* < .0001; ns, not significant. For MGB1, a fold change below 1 denotes lower expression in patients.

Table 2. Comparison of Circulating Biomarker Levels Between Patients and Controls

Biomarker	Patients, mean (SD)	Controls, mean (SD)	Fold change[a]	Statistical test	*P* value
hsa-miR-4734	8.26 (9.65)	1.97 (1.77)	4.2	Mann-Whitney U	< .001
hsa-miR-5787	6.21 (9.35)	1.55 (1.71)	4.0	Mann-Whitney U	.01
hsa-miR-4532	6.89 (10.74)	1.39 (0.93)	5.0	Mann-Whitney U	.09
EPCAM (relative expression)[b]	7.37 (4.17)	2.74 (3.07)	2.7	Welch *t* test	.04
CA15-3, U/mL	26.6 (29.2)	5.47 (3.35)	4.9	Mann-Whitney U	< .001
CEA, ng/mL	2.55 (2.26)	1.87 (0.83)	1.4	Mann-Whitney U	.51
MGB1, log(ng/mL)	3.73 (0.22)	3.92 (0.26)	.95	Welch *t* test	.002

Abbreviations: CA15-3, cancer antigen 15-3; CEA, carcinoembryonic antigen; EPCAM, epithelial cell adhesion molecule; MGB1, mammaglobin-1.
[a] Fold change calculated as the ratio of group means; for MGB1, < 1 reflects lower expression in patients.
[b] EPCAM data are derived from a 13-participant subset (7 patients, 6 controls) due to differences in the availability of biological material.

Integrating the 3 miRs with CA15-3 and MGB1 in an LOOCV logistic regression model yielded high discriminatory performance, with an AUC of 0.91 (95% CI, 0.77–0.96), sensitivity of 84%, and specificity of 90% at the optimal Youden cutoff (Figure 3, B). The integrated panel achieved discrimination comparable to that of the strongest individual marker (CA15-3; AUC, 0.92), while providing higher sensitivity (84% vs 76%) at a slightly lower specificity (90% vs 96%), illustrating the complementary value of cross-layer molecular integration.

Internal validation and model stability

Internal validation analyses demonstrated good model stability. The apparent model AUC was 0.95. Bootstrap optimism correction using 1000 resamples yielded an optimism-corrected AUC of 0.92. Nested cross-validation produced an AUC of 0.90, suggesting that the integrated biomarker panel retained substantial discriminative performance when evaluated under more conservative validation procedures. The relatively small decline in performance across validation approaches indicates that severe overfitting is unlikely despite the limited sample size (Table 4).

Cross-layer correlation structure

To examine whether the multiple molecular layers reflect coordinated or independent biological signals, we computed pairwise Pearson correlations among all measured biomarkers within the patient group

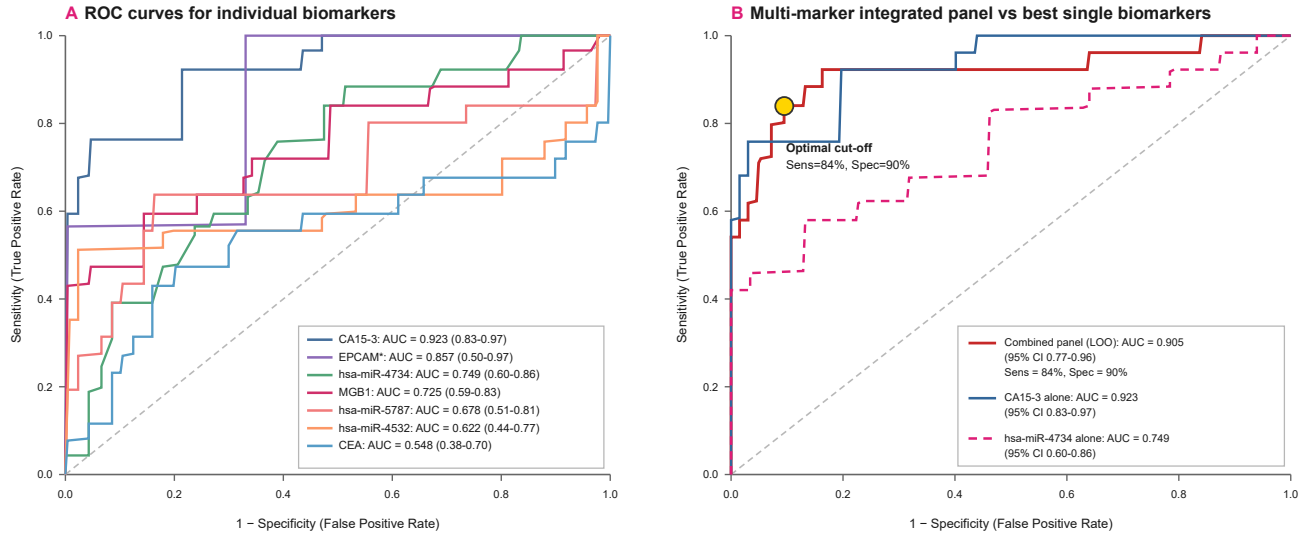


Figure 3. Diagnostic Discriminatory Performance of Individual and Integrated Circulating Biomarkers
A, Receiver operating characteristic (ROC) curves for the individual biomarkers, with 95% CIs estimated using the DeLong method. B, ROC curve of the integrated 5-marker logistic regression model (hsa-miR-4734, hsa-miR-5787, hsa-miR-4532, CA15-3, and MGB1) evaluated by leave-one-out cross-validation and compared with CA15-3 and hsa-miR-4734 alone. The gold marker indicates the optimal Youden cutoff (sensitivity, 84%; specificity, 90%). EPCAM was evaluated in a 13-participant subset (7 patients, 6 controls), and its AUC should not be directly compared with estimates from the full cohort.

Table 3. Diagnostic Performance of Individual Biomarkers and the Integrated Multimarker Panel

Biomarker or model	AUC	95% CI	Cutoff	Sensitivity, %	Specificity, %
hsa-miR-4734	.75	0.60–0.86	FC ≥ 1.45	60.0	85.7
hsa-miR-5787	.68	0.51–0.81	FC ≥ 1.50	64.0	84.0
hsa-miR-4532	.62	0.44–0.77	FC ≥ 2.45	52.0	98.0
EPCAM[a]	.86	0.50–0.97	FC ≥ 2.85	100.0	66.7
CA15-3	.92	0.83–0.97	9.7 U/mL	76.0	96.0
CEA	.55	0.38–0.70	NA	NA	NA
MGB1	.73	0.59–0.83	3.81	76.0	60.0
Integrated model (LOOCV)	.91	0.77–0.96	π = 0.36	84.0	89.8

Abbreviations: AUC, area under the receiver operating characteristic curve; CA15-3, cancer antigen 15-3; CEA, carcinoembryonic antigen; EPCAM, epithelial cell adhesion molecule; FC, fold change; LOOCV, leave-one-out cross-validation; MGB1, mammaglobin-1; NA, not applicable.
Note: The integrated model is a logistic regression on log-transformed and standardized values of the 3 miRs, CA15-3, and MGB1.
For the integrated model, π = 0.36 denotes the predicted-probability decision threshold selected by the Youden index from the leave-one-out cross-validated logistic regression; predicted probabilities of at least 0.36 were classified as case-positive.
[a] Derived from a 13-participant subset (7 patients, 6 controls).

Table 4. Internal Validation Metrics for the Integrated 5-Marker Classifier

Method	AUC
Apparent AUC	0.95
Bootstrap corrected	0.92
Nested CV	0.90

Abbreviations: AUC, area under the receiver operating characteristic curve; CV, cross-validation.

(Figure 4). Notable positive correlations were observed between hsa-miR-5787 and EPCAM (r = 0.64), suggesting concordant regulatory and epithelial dissemination dynamics, and between CA15-3 and CEA (r = 0.46), consistent with shared mucinous antigen biology. A biologically meaningful negative correlation was identified between hsa-miR-4532 and MGB1 (r = −0.42, P < .05), congruent with the posttranscriptional repression of mammaglobin-related transcripts by this microRNA. The CA15-3 was inversely correlated with EPCAM (r = −0.47), a pattern consistent with the known compartmental distinction between secreted soluble antigens and cell-surface markers retained on circulating epithelial cells. Together, these correlations argue against redundancy among the markers and reinforce the rationale for multilayer integration. These correlation P values are uncorrected for multiple comparisons and

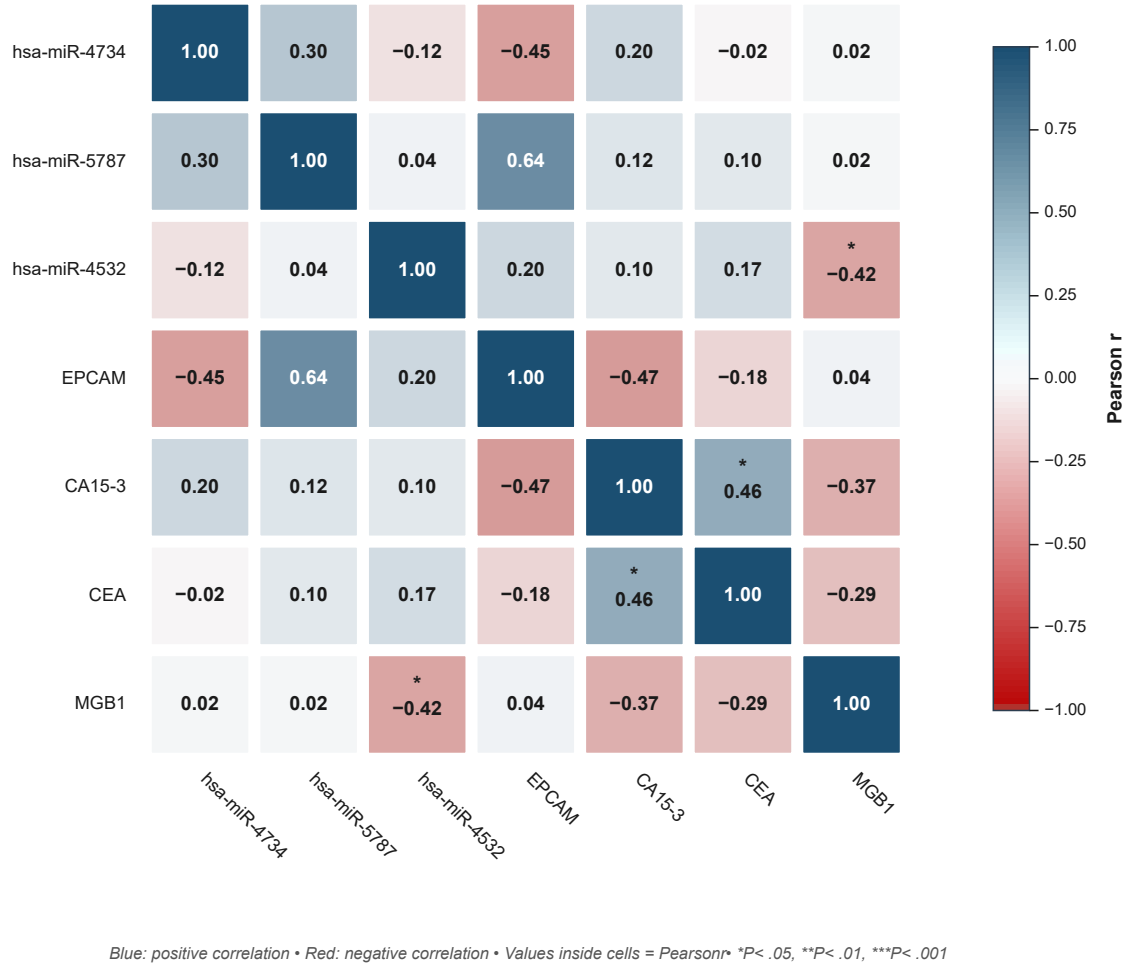


Figure 4. Correlation Matrix of Circulating Biomarkers in Patients With Early Breast Cancer
Heat map of pairwise Pearson correlation coefficients (*r*) among circulating microRNAs, EPCAM expression, and serum protein biomarkers in patients with early breast cancer (n = 25). Correlations were calculated from log-transformed biomarker values; blue indicates positive and red negative correlations. Asterisks indicate statistical significance (*$P < .05$; **$P < .01$; ***$P < .001$). Notable associations include positive hsa-miR-5787–EPCAM and CA15-3–CEA correlations and inverse hsa-miR-4532–MGB1 and CA15-3–EPCAM correlations.

should be interpreted as descriptive and hypothesis-generating rather than confirmatory; applying a Bonferroni correction for the 21 pairwise tests (α = .002) would reduce the number of statistically significant associations.

Integrated multilayer biomarker framework

The collective evidence from differential expression, ROC analysis, the multimarker model, and the cross-layer correlation structure supports a consistent 3-layer biomarker framework, as shown in Figure 5. Layer 1 (posttranscriptional regulation): Coordinated upregulation of hsa-miR-4734, hsa-miR-5787, and hsa-miR-4532 in serum captures systemic regulatory perturbation. Layer 2 (epithelial dissemination) is founded on the surrogate marker of EPCAM upregulation for early CTC activity. Layer 3 (tumor-associated antigens) incorporates the secreted protein dimension through a combined signature involving CA15-3 increase and MGB1 loss. The flow of these orthogonal molecular channels into a single integrated diagnostic readout is indicated by vertical arrows in Figure 5, with robust performance metrics (AUC, 0.91; sensitivity, 84%; specificity, 90%) underscoring the additional value of multilayer profiling relative to the potential clinical impact of any individual biomarker.

DISCUSSION

We developed a framework for a biologically principled, computationally driven liquid biopsy that approaches early breast cancer as an advanced systemic disturbance rather than the disorder of any single circulating molecule, enabling multimolecular-layer capture. Integrating bioinformatically prioritized miRs with EPCAM-associated epithelial dispersion signals and tumor-associated serum proteins in one Iraqi cohort, we provide quantitative evidence that cross-layer integration captures complementary biological information beyond traditional single-analyte testing, with an integrated panel AUC of 0.91, statistically

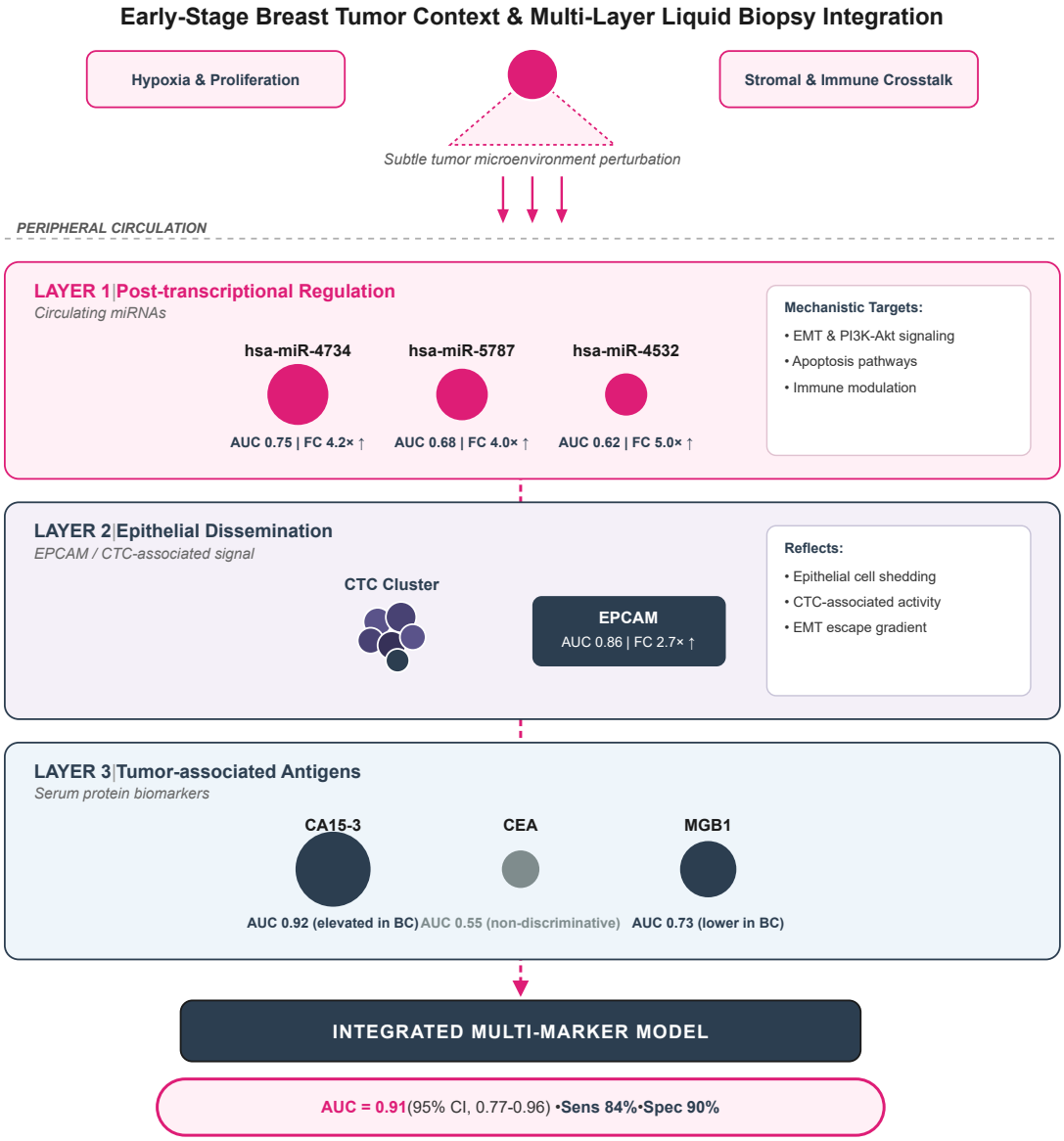


Figure 5. Multi-Layer Circulating Biomarker Framework for Early Breast Cancer Detection
Conceptual schema linking the early breast tumor and microenvironment to 3 complementary molecular layers detectable in peripheral circulation: posttranscriptional regulation (circulating microRNAs), epithelial dissemination (EPCAM/CTC-associated signal), and tumor-associated serum proteins. Bubble size represents individual-marker AUC, and arrows indicate integration of the 3 layers into a single 5-marker classifier, yielding a leave-one-out cross-validated AUC of 0.91 (95% CI, 0.77–0.96), with 84% sensitivity and 90% specificity.

comparable to the strongest individual marker (CA15-3; AUC, 0.92) but with higher sensitivity (84% vs 76%) at a modest reduction in specificity (90% vs 96%) and approaching the performance benchmarks achieved by highly optimized panels in much larger international cohorts.[11,12]

A critical interpretive caveat is that the diagnostic performance metrics reported here are derived from a binary contrast between symptomatic patients with cancer and entirely healthy controls, a design that maximizes intergroup separation but that overstates performance in the clinically relevant setting of population-based early detection, where the target population includes benign breast pathology, equivocal lesions, and premalignant conditions. This spectrum bias is intrinsic to the case-control design and necessitates prospective validation in screening-eligible populations before any translational claims can be supported.

The coordinated upregulation of hsa-miR-4734, hsa-miR-5787, and hsa-miR-4532 supports the increasingly accepted view that early breast cancer is reflected in serum as a structured pattern of posttranscriptional dysregulation rather than as the perturbation of any single locus.[9–11,17] The magnitude of upregulation observed (4- to 5-fold across markers) is consistent with the dynamic range reported for serum miR biomarkers in independent Iraqi cohorts, where miR-10b also exhibited stage-progressive upregulation in patients with breast cancer.[8] Mechanistically, the 3 microRNAs identified here share predicted target enrichment in EMT-related and PI3K-Akt signaling pathways, which links their

elevation in circulation to the canonical proliferative, antiapoptotic, and promigratory programs of early tumor biology. The fact that hsa-miR-4734 was the strongest miR discriminator in our cohort (AUC, 0.75), while hsa-miR-4532 was the weakest (AUC, 0.62), is itself biologically informative: it suggests that not all coordinately upregulated regulators carry equal diagnostic information, and that downstream model integration extracts complementary discriminatory signals even from individually weaker markers. Exploratory correlation analyses suggest limited associations between age and most biomarkers, although residual age confounding cannot be excluded.

EPCAM expression added an orthogonal cellular dimension to the regulatory microRNA layer and was, therefore, evaluated as a surrogate marker of epithelial dissemination and CTC activity. In healthy individuals, EPCAM transcripts are detectable at low basal levels because of routine epithelial turnover, which is why simple band detection in conventional PCR cannot distinguish patients from controls; only quantitative differential expression carries diagnostic meaning, an interpretive principle now established in the CTC literature.[14–16] The 2.7-fold elevation observed in patients (AUC, 0.86) is biologically consistent with low-level early shedding of EPCAM-expressing tumor cells or CTC clusters into peripheral circulation prior to overt metastasis, and the strong positive correlation between EPCAM and hsa-miR-5787 (r = 0.64) suggests a coupled regulatory-dissemination axis that warrants further mechanistic investigation. Conversely, the inverse correlation between EPCAM and CA15-3 (r = −0.47) reflects a compartmental dissociation between cell-surface adhesion molecules retained on circulating tumor cells and shed soluble mucin antigens, a duality that further supports the conceptual rationale for combining these layers rather than substituting one for the other.

Within the protein layer, CA15-3 emerged as the single strongest discriminator (AUC, 0.92; 95% CI, 0.83–0.97). This is somewhat higher than its typical performance in pure early-stage cohorts and likely reflects 2 factors: first, the presence of T2–T4 tumors and 40% nodal involvement in our cohort means that a meaningful subset of patients carry sufficient tumor burden to drive elevated CA15-3 release; and second, CA15-3 demonstrated similarly strong discrimination (mean, 36.15 U/mL in malignant cases) in a recent multiprotein Iraqi cohort study,[12] indicating that this performance is reproducible in Iraqi clinical settings rather than an artifact of single-cohort statistical noise. The inverse pattern observed for MGB1 (significantly lower serum concentrations in patients than in controls) is biologically coherent with prior literature documenting selective MGB1 downregulation in HER2-overexpressing and triple-negative subtypes, both of which are well-represented in our cohort (30% and 20%, respectively) and which are characterized by loss of mammary lineage fidelity. The strong negative correlation between hsa-miR-4532 and MGB1 (r = −0.42) further suggests a direct or indirect posttranscriptional contribution to MGB1 suppression and reinforces the notion that the regulatory and antigenic layers are biologically coupled.[18]

The nonsignificant performance of CEA in our cohort (AUC, 0.55; P = .51) is consistent with its established profile as a marker of metastatic disease and treatment monitoring rather than early-stage detection; in an independent Iraqi multiplex serum-biomarker study, CEA demonstrated high sensitivity and specificity when assessed alongside other serum markers,[12] supporting a stage-dependent rather than population-dependent interpretation. Importantly, in the multimarker integration, even biomarkers with a weaker individual ROC signal contributed to model robustness through complementary error structures: this is the core conceptual advantage of multilayer profiling over single-marker reductionism.

Although no internal validation strategy can fully substitute for external replication, the additional bootstrap optimism correction and nested cross-validation analyses performed in the present study demonstrated only modest reductions in AUC relative to apparent model performance. These findings suggest that severe overfitting is unlikely and support the stability of the integrated biomarker framework, while still emphasizing the need for prospective external validation before clinical application.

Beyond statistical performance, the present framework carries translational implications. In low- and middle-income countries where centralized imaging-based screening programs remain incomplete and high-throughput sequencing technologies are not uniformly available, a 5-marker liquid biopsy panel built on widely accessible qRT-PCR and ELISA platforms represents a realistic technological match. The high specificity of the integrated model (90%) is particularly valuable in this context, because false-positive triage testing imposes downstream cost and anxiety that disproportionately affect resource-limited settings. Coupled with the demonstrated reproducibility of CA15-3 discrimination across independent Iraqi cohorts[12] and the cumulative literature establishing microRNA-based liquid biopsy as a maturing diagnostic modality[4,5,17], the framework presented here provides a foundation for prospective validation in larger multicenter cohorts.

Several limitations should be acknowledged. First, the case-control design introduces spectrum bias: the contrast between symptomatic patients and entirely healthy volunteers is broader than that encountered in a screening setting, and the diagnostic performance metrics reported here should, therefore, be interpreted as upper-bound estimates pending prospective validation in a screening-eligible population. Second, the substantial age disparity between patients and controls (mean, 46.7 vs 32.3 years; Δ = 14.4 years; P < .001) reflects the practical realities of clinical recruitment yet introduces a potential confounder, representing a recognized limitation of this study. Age independently influences circulating microRNA profiles, serum protein concentrations, and epithelial cell turnover rates; accordingly, the biomarker

differences reported here cannot be unequivocally attributed to tumor biology without an age-adjusted analysis. Prospective validation studies must employ propensity scores or frequency-matched controls to isolate the tumor-specific signal from age-related variation. The lower age boundary of 13 years in the control group reflects the institutional minimum age for research participation consent rather than an epidemiologically designed inclusion criterion. The inclusion of adolescent controls alongside adult patients creates a clinically heterogeneous reference population; future studies should apply a minimum control age of 18 years or older and ideally restrict to women within the age range at which breast cancer risk is clinically relevant (≥ 25 years).

Third, the EPCAM analysis was performed in a 13-participant subset, and while descriptive statistics indicated a substantial effect size (AUC CI, 0.50–0.97), the wide CI means that this layer needs further validation.

Fourth, although bootstrap optimism correction and nested cross-validation demonstrated reasonable model stability, the integrated multimarker model was developed and evaluated within a single cohort. True generalization performance can only be confirmed in an independent external validation cohort.

Menopausal status was not systematically captured for all participants due to incomplete clinical documentation in a subset of controls; its potential independent effect on circulating miR profiles and serum protein concentrations could not be evaluated, representing an additional confounder that future studies must address.

Fifth, the observed interindividual variability, particularly in the miR layer, is biologically meaningful but limits the power to detect subgroup-specific effects in a 25-patient cohort. Finally, this study did not collect detailed dietary, reproductive, or environmental exposure data; the potential influence of these exposures on circulating biomarker levels could not be evaluated.

CONCLUSION

Early breast cancer in Iraqi women is associated with a coordinated, multilayer molecular signature detectable in peripheral blood. Three computationally prioritized circulating microRNAs (hsa-miR-4734, hsa-miR-5787, hsa-miR-4532) were collectively upregulated; EPCAM expression provided an orthogonal epithelial dissemination signal; and serum protein biomarkers contributed CA15-3 elevation and MGB1 reduction as complementary antigenic dimensions. Integrating these layers in a logistic regression model yielded an LOOCV-validated AUC of 0.91 with 84% sensitivity and 90% specificity, achieving performance comparable to the strongest individual marker (CA15-3; AUC, 0.92) while providing higher sensitivity at a modest reduction in specificity. These findings support a biologically grounded liquid biopsy framework in which cross-layer molecular integration rather than single-marker reductionism provides the path forward for noninvasive early detection of breast cancer. Validation in larger, prospectively recruited, age-matched cohorts and in screening-eligible populations is now warranted to translate this framework into clinical practice.

ETHICAL CONSIDERATIONS

All procedures involving human participants were conducted in accordance with the ethical standards of the institutional research committee and the Declaration of Helsinki. The study was approved by the Ethics Committee of the University of Baghdad, College of Science, Department of Biotechnology (reference No. 3763E/24/10/2024). Written informed consent was obtained from all participants prior to enrollment.

DATA AVAILABILITY

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. The public microRNA dataset GSE73002 used for bioinformatic prioritization is available through the NCBI Gene Expression Omnibus.

CONFLICT OF INTERESTS

The authors declare no conflicts of interest.

FUNDING

None.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the management and staff of the National Hospital for Oncology and Hematology, Najaf, Iraq, for their assistance with patient recruitment and sample collection. The authors are deeply indebted to all women who participated in this study.

AI DISCLOSURE

ChatGPT (OpenAI) and Claude (Anthropic) were used solely for language editing, grammar refinement, formatting assistance, and figure preparation support. No AI tool was used for data generation, statistical analysis, result interpretation, or scientific decision-making. The authors take full responsibility for the study design, analyses, interpretation of findings, and conclusions presented in this manuscript.

AUTHOR CONTRIBUTION

ANA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

RAAA: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing.

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. doi: 10.3322/caac.21834
2. Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. *Lancet*. 2021;397(10286):1750-69. doi: 10.1016/s0140-6736(20)32381-3
3. Mihai AM, Ianculescu LM, Suciu N. MiRNAs as potential biomarkers in early breast cancer detection: a systematic review. *J Med Life*. 2024;17(6):549-54. doi: 10.25122/jml-2024-0322
4. 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;23(17):9952. doi: 10.3390/ijms23179952
5. Alba-Bernal A, Lavado-Valenzuela R, Domínguez-Recio ME, Jiménez-Rodriguez B, Queipo-Ortuño MI, Alba E, et al. Challenges and achievements of liquid biopsy technologies employed in early breast cancer. *EBioMedicine*. 2020;62:103100. doi: 10.1016/j.ebiom.2020.103100
6. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan ELA, et al. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A*. 2008;105(30):10513-8. doi: 10.1073/pnas.0804549105
7. Schwarzenbach H, Nishida N, Calin GA, Pantel K. Clinical relevance of circulating cell-free microRNAs in cancer. *Nat Rev Clin Oncol*. 2014;11(3):145-56. doi: 10.1038/nrclinonc.2014.5
8. Abdallah M, Aziz IH, Alsammarraie AZ. Assessment of miRNA-10b expression levels as a potential precursor to metastasis in localized and locally advanced/metastatic breast cancer among Iraqi patients. *Int J Breast Cancer*. 2024;2024:2408355. doi: 10.1155/2024/2408355
9. Su X, Shan Z, Duan S. Harnessing extracellular vesicles using liquid biopsy for cancer diagnosis and monitoring: highlights from AACR Annual Meeting 2024. *J Hematol Oncol*. 2024;17(1):55. doi: 10.1186/s13045-024-01577-y
10. Xiao J, Zhang L, Su R, Zhao B, Dang Y, Zhao C, et al. MicroRNAs in breast cancer—new frontiers in diagnosis, targeted therapy, and prognosis assessment. *Front Oncol*. 2025;15:1529907. doi: 10.3389/fonc.2025.1529907
11. de Miranda FS, Slaibi-Filho J, Calasans dos Santos G, Carmo NT, Kaneto CM, Borin TF, et al. MicroRNA as a promising molecular biomarker in the diagnosis of breast cancer. *Front Mol Biosci*. 2024;11:1337706. doi: 10.3389/fmolb.2024.1337706
12. Sahan K, Sahan ZA, Qazzaz HA. Predictive power of multiplex serum biomarkers in breast cancer: a comprehensive ROC curve analysis. *Iraqi J Cancer Med Genet*. 2025;18(2):51-9. doi: 10.29409/zvqcs091
13. Hussain AM, Ali AH, Mohammed HL. Correlation between serum and tissue markers in breast cancer Iraqi patients. *Baghdad Sci J*. 2022;19(3):0501. doi: 10.21123/bsj.2022.19.3.0501
14. Perelmuter VM, Grigoryeva ES, Savelieva OE, Alifanov VV, Andruhova ES, Zavyalova MV, et al. EpCAM-CD24+ circulating cells associated with poor prognosis in breast cancer patients. *Sci Rep*. 2024;14(1):12245. doi: 10.1038/s41598-024-61516-2
15. Ring A, Mineyev N, Zhu W, Park E, Lomas C, Punj V, et al. EpCAM-based capture detects and recovers circulating tumor cells from all subtypes of breast cancer except claudin-low. *Oncotarget*. 2015;6(42):44623-34. doi: 10.18632/oncotarget.5977
16. de Wit S, Manicone M, Rossi E, Lampignano R, Yang L, Zill B, et al. EpCAM(high) and EpCAM(low) circulating tumor cells in metastatic prostate and breast cancer patients. *Oncotarget*. 2018;9(86):35705-16. doi: 10.18632/oncotarget.26298
17. Foruzandeh Z, Alivand MR, Ghiami-Rad M, Zaefizadeh M, Ghorbian S. Identification and validation of miR-583 and miR-877-5p as biomarkers in patients with breast cancer: an integrated experimental and bioinformatics research. *BMC Res Notes*. 2023;16(1):72. doi: 10.1186/s13104-023-06343-w
18. Abbas RAR. Biochemical profiling of pro-inflammatory (ILs) and hormonal markers (HGH) as predictive biomarkers implications in Iraqi women with breast carcinoma. *J Med Oral Biosci*. 2026;3(2):80-92. doi: 10.58564/jmob.163

How to Cite This Article

Al-Ebadi AN, Al-Khalidi RAA. An Integrated Circulating MicroRNA–EPCAM–Protein Liquid Biopsy Signature for Early Breast Cancer Detection in Iraqi Women: A Multilayer Molecular Framework. Arch Breast Cancer. 2026; 13(4):455-466. doi: 10.32768/abc.5842169374-058
Available from: https://www.archbreastcancer.com/index.php/abc/article/view/1330