Breast Cancer Research

RESEARCH

Open Access

DNA methylation-predicted plasma protein levels and breast cancer risk

Jacob K. Kresovich[1,2,3*], Brett M. Reid[1], Katie M. O'Brien[3], Zongli Xu[4], Doratha A. Byrd[1], Clarice R. Weinberg[4], Dale P. Sandler[3] and Jack A. Taylor[3,5]

Abstract

Background Blood DNA methylation (DNAm) profiles have been used to show that changes in circulating leukocyte composition occur during breast cancer development, suggesting that peripheral immune system alterations are markers of breast cancer risk. Blood DNAm profiles have recently been used to predict plasma protein concentrations ("Protein EpiScores"), but their associations with breast cancer risk have not been examined in detail.

Methods Whole blood DNAm profiles were obtained for a case-cohort sample of participants in the Sister Study and used to calculate 109 Protein EpiScores. Of the 4,479 women included, 2,151 (48%) were diagnosed with breast cancer within 15 years of their baseline blood draw (median time to diagnosis: 8.6 years; 1,673 invasive cancer and 478 ductal carcinomas in situ). Protein EpiScores associations with breast cancer incidence were estimated using weighted Cox regression models, overall and stratified by time and participant characteristics.

Results Protein EpiScores for RARRES2, IGFBP4, and CCL21 were positively associated with invasive breast cancer risk (hazard ratios from 1.17 to 1.24), while those for F7, SELL, CXCL9, CD48, and IL19 were inversely associated (hazard ratios from 0.82 to 0.86) (all FDR < 0.10). Eight immune response-related Protein EpiScores (CXCL9, CD48, FCGR3B, CXCL11, CCL21, CRTAM, VCAM1, GZMA) were associated with invasive cancers diagnosed within five years of enrollment. Protein EpiScore associations were consistently stronger for estrogen receptor-negative tumors.

Conclusions Several Protein EpiScores, including many related to immune response, were associated with breast cancer risk, highlighting novel changes to the peripheral immune system that occur during breast cancer development.

Keywords Breast Cancer, DNA methylation, Proteomics, Epidemiology, Prospective cohort

*Correspondence:
Jacob K. Kresovich
jacob.kresovich@moffitt.org
[1]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612, USA
[2]Department of Breast Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612, USA
[3]Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, Durham, NC 27709, USA
[4]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, Durham, NC 27709, USA
[5]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, Durham, NC 27709, USA

Introduction

Blood DNA methylation (DNAm) at individual cytosine-phosphate-guanine (CpG) sites and combinations of CpGs, used as DNAm-based predictors of biological age or leukocyte composition, have been investigated as markers of breast cancer risk [1–11] and may improve breast cancer risk prediction [12].

DNAm-based predictors of individual proteins (e.g., C-reactive Protein) and physiological traits (e.g., smoking status) may have better predictive performance for disease incidence than direct measurements [13–21]. For example, DNAm-based predictors of smoking history outperform self-reported information in predicting lung cancer incidence and death [19, 20]. Paired plasma proteomic and leukocyte DNAm data from thousands of individuals have recently been used to derive a library of DNAm-based predictors of plasma proteins, called "Protein EpiScores" [22]. Protein EpiScores were developed using elastic net regularization to identify sets of CpGs throughout the genome where DNAm levels correlated with rank-based, inverse normalized, plasma protein concentrations, adjusted for age, sex, and known protein quantitative loci (pQTL) [23–25]. The number of CpGs used in any individual Protein EpiScore ranges from one to 395 (mean 96 CpGs), with some CpGs appearing in multiple scores (e.g., cg05574921 from *AHRR*) [22]. Protein EpiScores were considered validated if, in external testing datasets, their Pearson correlation coefficient with their directly measured protein was >0.10. In total, 109 Protein EpiScores met this threshold, explaining between 1% and 58% of the variance in their corresponding proteins [22]. In initial studies, Protein EpiScores were found to be associated with the risk of lung cancer [22], colorectal cancer [22], cognitive decline [26], cardiometabolic conditions [22], and cardiovascular disease [22, 27].

Protein EpiScores are an emerging class of DNAm-based metrics that may help quantify previously unmeasured aspects of breast cancer risk and provide insights into breast cancer development. In this study, we used existing whole blood DNAm data from a racially diverse case-cohort sample of 4,479 women enrolled in the Sister Study to examine the relationship between Protein EpiScores and breast cancer incidence (Fig. 1).

Methods

Study population

The Sister Study enrolled a prospective cohort of 50,884 women from the United States, who were aged 35 to 74,

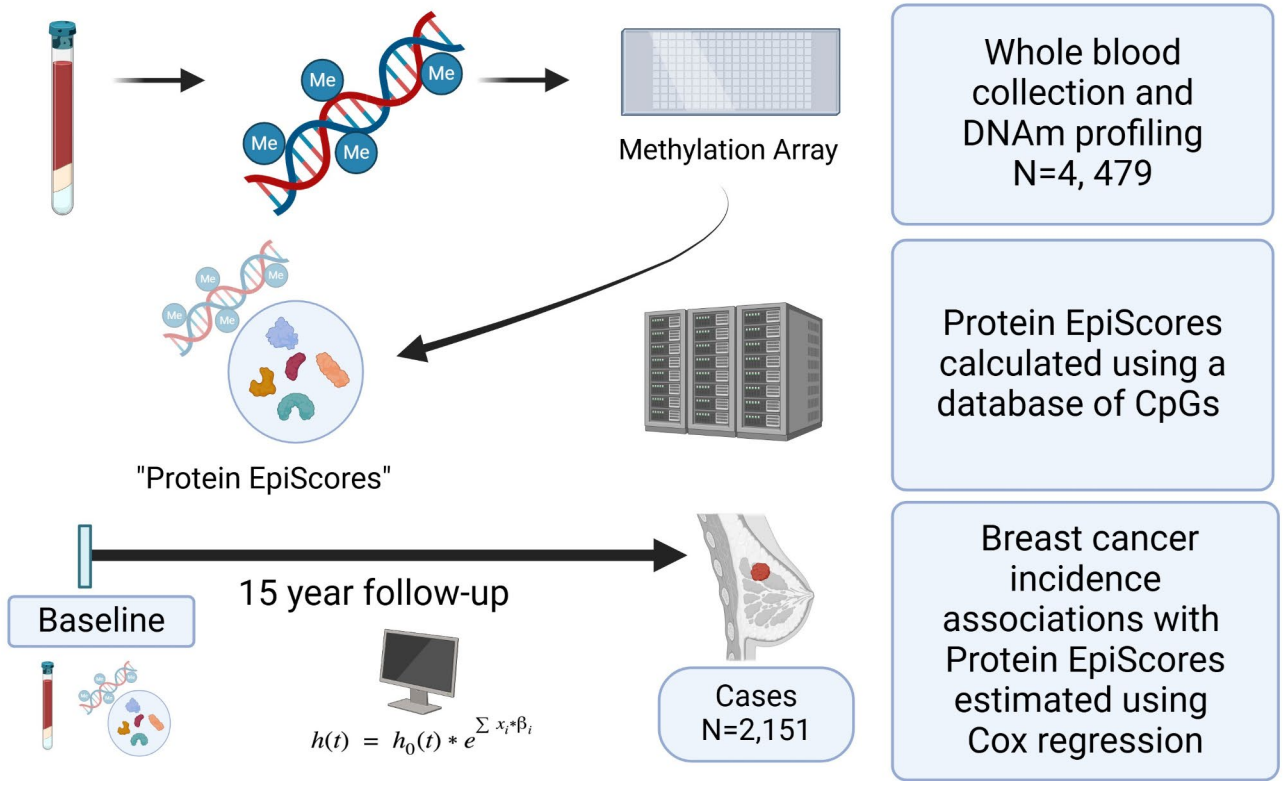

Fig. 1 Schematic of study design. Whole blood samples were collected from Sister Study participants at enrollment when all were breast cancer-free and assayed for genome-wide DNA methylation (DNAm) profiles using either the HumanMethylation450 or MethylationEPIC BeadChips. DNAm data were used to calculate circulating concentrations of 109 plasma proteins ("Protein EpiScores") using a pre-identified set of CpGs. Breast cancer incidence associations with Protein EpiScores were estimated using weighted Cox regression models overall and stratified by time and participant characteristics

between July 2003 and March 2009 and was designed to identify novel environmental and biological factors associated with breast cancer incidence and survival [28]. By design, enrollment was restricted to women who were breast cancer-free, but had a first-degree family history of breast cancer [29]. Upon enrollment, participants completed a detailed computer-assisted telephone interview to provide information on demographics, lifestyle, and health. During an in-person home visit, trained medical examiners collected body measurements and whole blood samples following standardized procedures. Participants are recontacted annually to self-report major changes in health, with response rates over 85%. Every three years, a more comprehensive questionnaire gathers updates on lifestyle, environmental exposure, and health. Written informed consent from all participants was collected at the home visit, and the study is overseen by the Institutional Review Board of the National Institutes of Health. Research activities were performed in accordance with the Declaration of Helsinki. More information and procedures for accessing Sister Study data can be found at: https://sisterstudy.niehs.nih.gov/English/coll-data.htm.

Whole blood DNA methylation assessment and quality control

Two case-cohort samples of women were selected for DNAm profiling in this study. In 2014, blood DNA samples from 2,878 self-identified non-Hispanic White women were assayed using the Infinium HumanMethylation450 BeadChip, including 1,633 women diagnosed with breast cancer during study follow-up. In 2019, blood DNA samples from 2,166 self-identified Black (Hispanic or non-Hispanic, $n = 736$) and non-Hispanic White women ($n = 1{,}430$) were assayed using the Infinium MethylationEPIC BeadChip, including 999 women (Black, $n = 265$; White, $n = 734$) diagnosed with breast cancer after study enrollment [30]. This second case-cohort sample was intentionally enriched for Black women and those diagnosed with estrogen receptor (ER) negative breast cancer. 541 women had DNAm profiled on both arrays; for these women, only DNAm data from the HumanMethylation450 BeadChip were used for analysis.

Genomic DNA was extracted from whole blood aliquots using an automated system (AutoPure, Gentra Systems) in the NIEHS Molecular Genetics Core Facility or using DNAQuik at BioServe Biotechnologies LTD (Beltsville, MD). Extracted DNA was bisulfite-converted using the EZ DNA Methylation Kit (Zymo Research, Orange County, CA). Samples were tested for complete bisulfite conversion, and converted DNA was analyzed using Illumina's Infinium BeadChip protocols with high-throughput robotics to minimize batch effects. Methylation analysis was conducted at the National Institutes of Health Center for Inherited Disease Research (Baltimore, MD) for the 2014 case-cohort and at the National Cancer Institute (Rockville, MD) for the 2019 case-cohort.

DNAm data was preprocessed using the *ENmix* software pipeline, which included background correction, dye-bias correction, inter-array normalization, and probe-type bias correlation [31–33]. Samples were excluded if they did not meet quality control measures, including bisulfate intensity < 5,000, more than 5% of probes with low-quality methylation values (detection $P > 0.000001$, < 3 beads, or values outside 3 times the interquartile range), or if they were outliers for their methylation beta value distributions. In total, 4,483 samples passed quality control. Four participants were missing age at end of follow-up, resulting in a DNAm analytic sample of 4,479 participants (Supplemental Fig. 1).

Protein EpiScore derivation and calculation in the Sister Study

Blood DNAm data were used to calculate circulating levels of 109 Protein EpiScores using the 'methscore' function of *ENmix* [34]. Additional details on the Protein EpiScore development, including component CpGs (and weights) and generalized functions of the target proteins, can be found in the original report [22]. The target proteins were classified into 39 unique groups, with some proteins assigned to multiple groups. The most common functions were immune response (49 proteins, 33%), metabolic (11 proteins, 7%), cell adhesion (11 proteins, 7%), growth factors (7 proteins, 5%) and vascular (6 proteins, 4%).

Breast cancer incidence and characteristics

Incident breast cancers and dates of diagnosis of Sister Study participants are self-reported. Women who report a breast cancer diagnosis are asked to provide a personal copy of their pathology report and are re-contacted six months later to obtain permission to contact their healthcare providers for medical records. Women self-report information on tumor characteristics, which included invasiveness (invasive vs. ductal carcinoma in situ [DCIS]) and ER status. Medical record and pathology reports, when available, are abstracted for information on tumor characteristics and treatments. Because agreement between self-reported and abstracted information is high (e.g., positive predictive values > 99% for invasiveness and ER positivity) [35], self-reported information is used when medical records are not available.

Statistical analysis

Because the study population comprised two case-cohort samples, all models were weighted based on the participants' inverse probability of selection for DNAm

profiling. As a result, the association estimates reported here are generalizable to the full sample of Black and non-Hispanic White women in the Sister Study [36]. Sample characteristics are described using weighted means and standard deviations (SD) or weighted proportions overall and stratified by breast cancer status.

Breast cancer risk associations with the Protein EpiScores, per 1-SD increase, were estimated using case-cohort Cox regression models with robust standard errors and reported as hazard ratios (HRs) with 95% confidence intervals (CIs) and 2-sided P-values [36]. In all models, chronological age was used as the primary timescale [37] and left truncation was determined by age at blood draw. Follow-up ended at age of the breast cancer event, end of study follow-up (September 30, 2019), loss-to-follow-up, or death. Associations were examined for invasive breast cancer, where a DCIS diagnosis was treated as a censoring event, and for all breast cancer events combined (DCIS and invasive). To exclude the influence of occult breast cancer, associations were examined excluding the first two years of follow-up. Additional analyses examined Protein EpiScore associations with short-term breast cancer risk by estimating associations for breast cancers diagnosed within five years of the baseline blood draw.

Primary Protein EpiScore associations were estimated in models adjusted for methylation platform (HumanMethylation450, MethylationEPIC), age (years), and self-reported race (White, Black). Sensitivity analyses were conducted by additionally adjusting for breast cancer risk factors (smoking history, physical activity, body mass index, menopausal status, and an interaction term between body mass index and menopausal status) and leukocyte proportions (granulocytes, CD8 + T, CD4 + T, B cells, natural killers, monocytes), as estimated by the top-performing deconvolution algorithm [38]. Because leukocyte proportions sum to one, monocyte proportions were excluded from the models to avoid overfitting. Effect modification was examined by self-reported race, and menopausal status, and examiner-measured body mass index at baseline.

Joint Cox regression models were performed to investigate Protein EpiScore associations with invasive breast cancer risk, stratified by ER status (positive, negative). Women were censored if they were diagnosed with the alternative subtype of interest, had missing ER information, or were diagnosed with DCIS. To test for statistical interaction, the joint model was parameterized to allow for direct comparison of subtype-specific associations using a Wald test [39]. In the primary analyses of breast cancer risk, statistical significance was determined using a Benjamini-Hochberg False Discovery Rate (FDR) < 0.10 [40]. To explore potential effect modification, significant statistical interaction was defined at FDR < 0.15. All analyses were performed using SAS (version 9.4, Cary, NC) and R (version 4.1.0, R Foundation for Statistical Computing, Vienna, Austria).

Results

Sample population

A total of 4,479 women were followed for up to 15 years (median follow-up, 8.6 years). The participants had a weighted mean age of 56 years at enrollment. Most were non-Hispanic White (84%), had at least some college education (85%), and were postmenopausal (69%) (Table 1). Of the 2,500 women randomly selected into the case-cohort ("random subcohort"), 174 (7%) were diagnosed with breast cancer after study enrollment. An additional 1,977 women were sampled because they were diagnosed with breast cancer after study enrollment ("case sample"). Of the 2,151 women diagnosed with incident breast cancer, 1,673 were diagnosed with invasive breast cancer (1,483 non-Hispanic White, 190 Black) and 478 were diagnosed with DCIS (409 non-Hispanic White, 69 Black). Among the invasive breast cancer cases, 1,145 were ER positive (1,040 non-Hispanic White, 105 Black) and 205 were ER negative (167 non-Hispanic White, 38 Black) (Supplemental Fig. 1). Compared to women who remained breast cancer-free, those with breast cancer were slightly older (57 years vs. 56 years), reported greater alcohol use at baseline (4.5 vs. 4.1 drinks/week), were more likely to be non-Hispanic White (88% vs. 80%), and be postmenopausal (72% vs. 66%) (Table 1).

Protein EpiScore associations with breast cancer incidence

In weighted Cox regression models adjusted for age, race, and DNAm platform, eight of the 109 Protein EpiScores were significantly associated with the incidence of invasive breast cancer (Fig. 2). The strongest associations were observed for F7 and RARRES2 EpiScores: F7 was inversely associated with invasive breast cancer (HR = 0.82, 95% CI: 0.74, 0.91, $P = 3.0 \times 10^{-4}$, FDR = 0.03), while RARRES2 was positively associated (HR: 1.24, 95% CI: 1.10, 1.40, $P = 6.0 \times 10^{-4}$, FDR = 0.03). Among the eight Protein EpiScores associated with invasive breast cancer risk, five are related to immune response, with four of these (SELL, CXCL9, CD48, and IL19) showing inverse associations. Pearson correlation coefficients between the eight Protein EpiScores ranged from −0.40 (RARRES2 and IL19) to 0.85 (CD48 and CXCL9) (Supplemental Fig. 2). In models adjusted for breast cancer risk factors, although HRs for RARRES2 and CCL21 were slightly attenuated, the HRs for CXCL9, CD48, F7, IGFBP4, SELL, and IL19 were essentially unchanged; despite the similar strengths of associations, no associations reached statistical significance (all FDR > 0.10, Supplemental Table 1). In models adjusted for leukocyte composition, associations with RARRES2, F7, SELL, IGFBP4, CCL21, and

Table 1 Weighted characteristics of sister study subsample with DNAm data (*N* = 4,479)

Characteristic	Overall *N* = 4,479	Breast cancer-free *N* = 2,328	Breast Cancer *N* = 2,151
Age, mean yrs.	55.7	55.5	57.4
Physical activity, mean METs/week	51.1	51.6	47.3
Smoking history, mean pack-years	2.4	2.3	3.2
Alcohol use, mean drinks/week	4.2	4.1	4.5
Parity, mean live births	1.9	1.9	1.9
Body mass index, mean kg/m2	27.8	27.7	28.2
Self-reported Race			
Non-Hispanic White	84%	80%	88%
Black (Hispanic or non-Hispanic)	16%	20%	12%
Educational Attainment			
High school or less	14%	15%	14%
Some college	33%	34%	32%
College graduate or more	52%	51%	54%
Smoking status			
Never	54%	56%	53%
Former	37%	35%	40%
Current	8%	9%	7%
Menopausal status			
Premenopausal	31%	34%	28%
Postmenopausal	69%	66%	72%
Tumor invasiveness			
Invasive			78%
DCIS			22%
Invasive tumor ER status			
Positive			69%
Negative			12%
Missing			19%

CXCL9 remained statistically significant (all FDR < 0.10) (Supplemental Table 2). Protein EpiScore associations for IL19 and CD48 were also similar, but associations did not reach statistical significance (FDR = 0.11 for both). In a combined analysis of DCIS and invasive breast cancers, Protein EpiScore associations were attenuated compared to those for invasive breast cancer alone, and none reached statistical significance (Supplemental Table 3).

After excluding the first two years of follow-up, Protein EpiScores for F7, RARRES2, and IGFBP4 remained statistically significantly associated with the incidence of invasive breast cancer at FDR < 0.10 (Supplemental Fig. 2). The other five Protein EpiScores showed similar strengths of association but were not statistically significant. In an analysis of invasive breast cancer events occurring within the first five years of follow-up, IGFBP4, CXCL9, CD48, and CCL21 were significantly associated (Supplemental Fig. 3). Statistically significant associations were also observed for FCGR3B, CXCL11, CRTAM, VCAM1, LGALS3BP, SMPD1, and GZMA. Notably, eight of the eleven Protein EpiScores associated with invasive breast cancer incidence occurring in the first five years of follow-up are related to immune response (CXCL9, CD48, FCGR3B, CXCL11, CCL21, CRTAM, VCAM1, and GZMA), with seven of these showing inverse associations.

Protein EpiScore associations stratified by participant characteristics

In analyses stratified by participant characteristics, significant interactions by race were observed for six Protein EpiScores: HGF, LYZ, CLEC11A, VEGFA, F7, and LFT. The inverse association between the F7 Protein EpiScore and invasive breast cancer was significant only among non-Hispanic White women (White, HR: 0.78, 95% CI: 0.66, 0.87, $P = 1.0 \times 10^{-4}$; Black, HR: 0.97, 95% CI: 0.86, 1.10, $P = 0.69$; Interaction FDR = 0.12) (Supplemental Table 4). Although the Protein EpiScores for HGF, LYZ, CLEC11A, VEGFA, and LFT were not significantly associated with invasive breast cancer risk in the primary analysis, they were all positively associated in Black women and inversely associated in White women (all interaction FDR < 0.15). Only two of these Protein EpiScores are related to immune response (LYZ and LTF). No significant invasive breast cancer interactions were observed between any Protein EpiScores and body mass index or menopause status (Supplemental Tables 5 and 6).

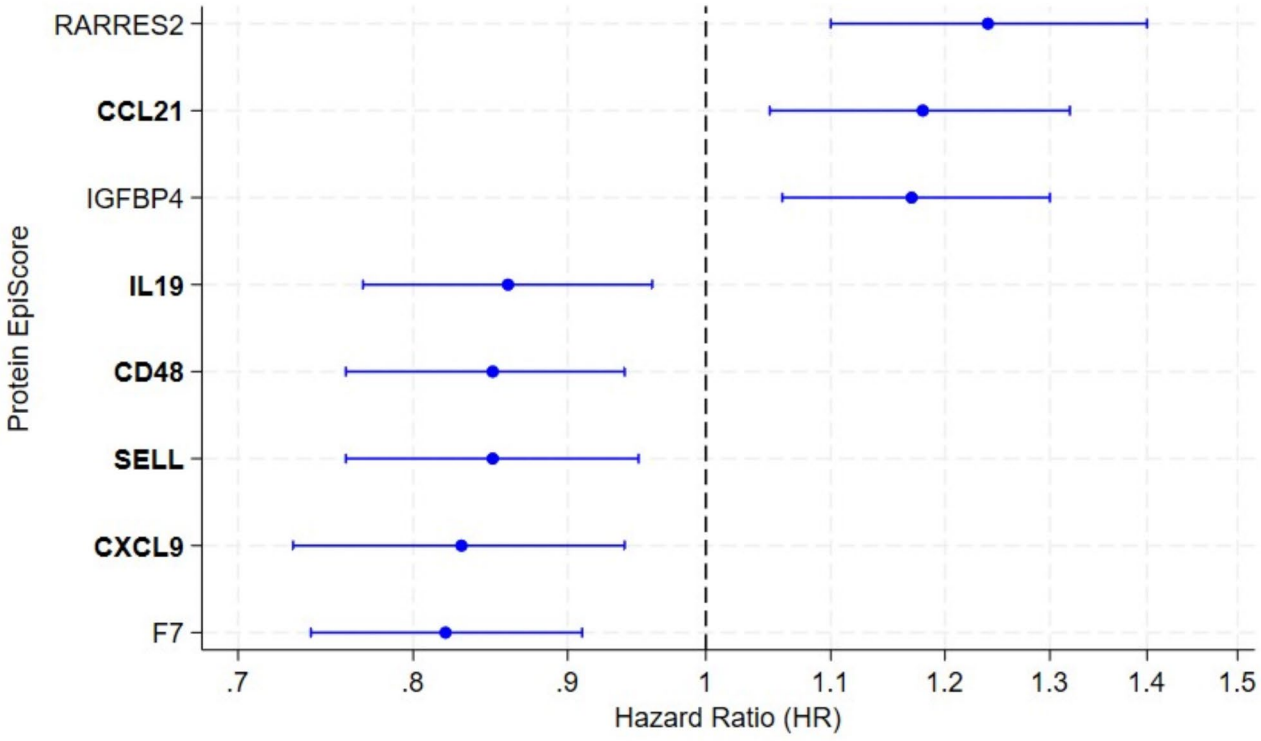

Fig. 2 Protein EpiScore associations with invasive breast cancer incidence. Hazard Ratios and 95% confidence intervals from weighted Cox regression models displaying associations for the Protein EpiScores significantly associated with invasive breast cancer incidence (FDR < 0.10). Models treat age as the time-scale and are therefore adjusted for age. The models additionally include DNAm platform (HumanMethylation450, MethylationEPIC) and self-reported race (non-Hispanic White, Black) as covariates. Bolded Protein EpiScores represent those that are designed to predict plasma proteins with 'immune response' functions

Protein EpiScore associations by tumor ER status

In the analysis stratified by ER status, 40 Protein EpiScores showed significant interactions across tumor subtypes (interaction FDR < 0.15; Supplemental Table 7). Among these, five Protein EpiScores—CXCL9, IL19, CD48, CCL21, and F7—were significantly associated with invasive breast cancer risk in the primary analysis, including four related to immune response. Notably, nearly all of the Protein EpiScores, including all five identified in the primary analysis, displayed stronger associations in women diagnosed with ER-negative tumors (Fig. 3). The most pronounced heterogeneity for Protein EpiScores identified in the primary analysis were for CXCL9 (ER-positive, HR: 0.87, 95% CI: 0.76, 0.99, $P = 0.04$; ER-negative, HR: 0.44, 95% CI: 0.34, 0.58, $P = 3.2 \times 10^{-9}$; Interaction FDR = 0.0003) and IL19 (ER-positive, HR: 0.93, 95% CI: 0.83, 1.04, $P = 0.21$; ER-negative, HR: 0.56, 95% CI: 0.43, 0.73, $P = 1.6 \times 10^{-5}$; Interaction FDR = 0.004) (Fig. 4, Supplemental Table 7).

Discussion

In this prospective case-cohort study of 4,479 women, including 2,151 who were diagnosed with breast cancer after the enrollment blood draw, we identified eight Protein EpiScores significantly associated with invasive breast cancer risk. Five of these were designed to predict proteins involved in immune response, with four displaying inverse associations. In the analysis of short-term invasive breast cancer risk, eight Protein EpiScores related to immune response were significantly associated, with seven displaying inverse associations. There was considerable association heterogeneity by breast cancer subtype, with stronger associations for women diagnosed with ER-negative tumors. These findings highlight novel changes in peripheral immunity that may occur during breast cancer development.

Various Protein EpiScores have been reported as risk markers for cancer [22], cardiometabolic conditions [22], cognitive decline [26], and cardiovascular disease [22, 27]. However, interpreting Protein EpiScore associations is complex. Protein EpiScores are only modestly correlated with their corresponding protein targets, with Pearson correlations ranging from 0.10 (STC1) to 0.73 (MST1) [22]. These imperfect correlations stem partly from the decision to train the Protein EpiScore models on plasma protein concentrations after adjustment for age, sex, and pQTLs—factors that strongly influence circulating protein levels [23–25, 41, 42]. Additionally, Protein EpiScores were developed using genome-wide DNAm data from leukocytes, even though only some of

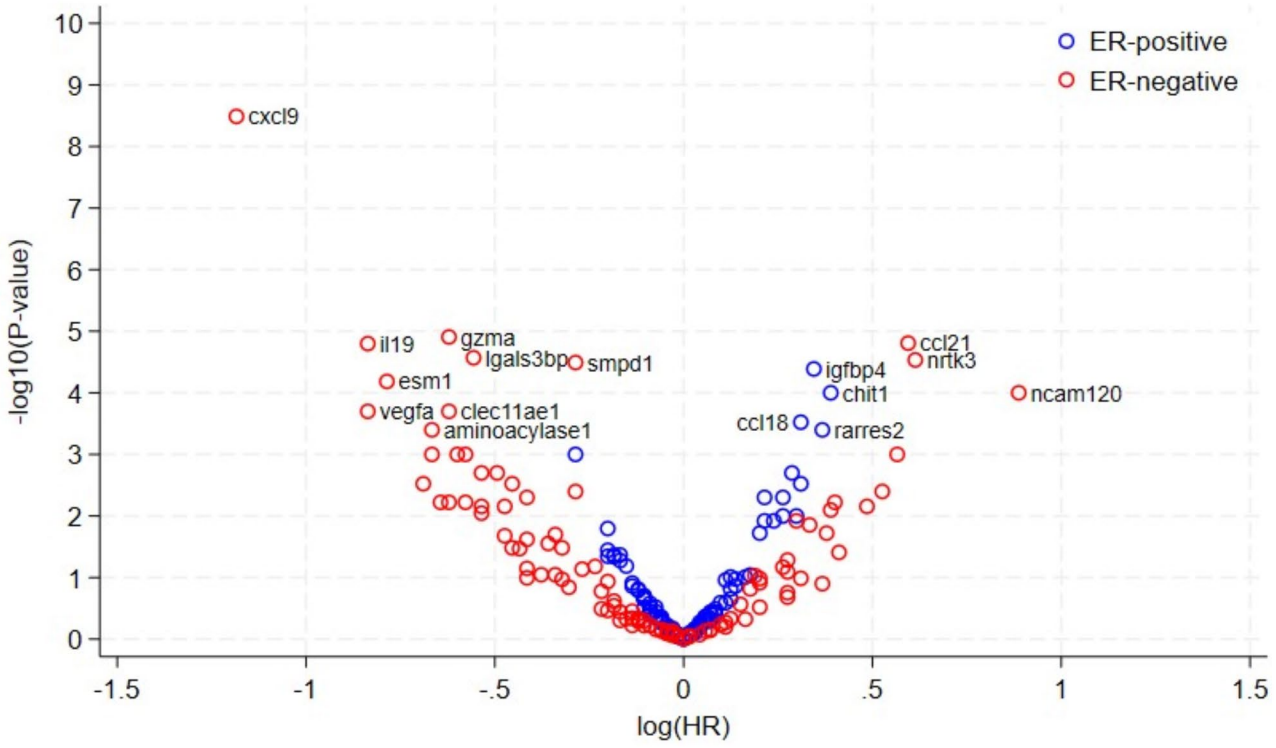

Fig. 3 Protein EpiScore associations with invasive breast cancer incidence, stratified by tumor ER status. Volcano plot depicting the log(hazard ratios) and -$\log_{10}$(P-values) for all 109 Protein EpiScores and invasive breast cancer incidence, stratified by tumor ER status. Protein EpiScore names displayed for those significantly associated with invasive breast cancer incidence at $P < 0.001$. Results from weighted Cox regression models treating age as the time scale and adjusted for DNAm platform and race

the targeted proteins are principally produced by leukocytes. Consequently, the CpGs used in the Protein EpiScores do not necessarily map to the genes encoding the target proteins; for example, only one of the 26 CpGs in the F7 Protein EpiScore is located in the *F7* gene [22]. Compared to plasma protein levels, which are known to vary over short periods [43, 44], leukocyte DNAm profiles are relatively more stable over time [45]. Protein EpiScores may, therefore, represent a time-integrated assessment of leukocyte transcriptional responses to different plasma protein concentrations rather than a direct measurement of those proteins. In contrast, Protein EpiScores for immune response proteins may reflect the long-term transcriptional state of leukocytes. Protein EpiScores are known to replicate certain well-established protein-disease associations, suggesting that they may complement direct protein measurements in studying disease development [46–50].

The association between Protein EpiScores and breast cancer risk was previously examined in the Generation Scotland Cohort, which included only 131 breast cancer events and found no statistically significant associations [22]. Our study, which includes more than 10 times as many invasive breast cancer cases, identified significant associations with eight Protein EpiScores in models adjusted for age, race, and methylation platform. Specifically, we observed positive associations with Protein EpiScores for IGFBP4 and RARRES2 (chemerin), both of which have been reported to be higher in treatment-naïve invasive breast cancer patients than cancer-free women [51, 52]. The other six Protein EpiScores we identified are for proteins not yet investigated in epidemiologic studies of breast cancer. Whether direct measurements of these other proteins are associated with breast cancer risk will need to be examined in future studies.

Protein EpiScores may also provide insights into breast cancer development. In this study, five of the identified Protein EpiScores are related to immune response; four of which may enhance the immune system's ability to detect and destroy malignant cells. Specifically, we found inverse associations between breast cancer and Protein EpiScores for CD48, SELL (L-selectin), CXCL9, and IL19. CD48 and SELL are adhesion molecules that may protect against breast cancer by enhancing leukocytes' ability to bind to malignant cells [53, 54]. CXCL9 and IL19 are signaling molecules that regulate anti-tumor immune cell subsets, such as CD4 + helper and CD8 + cytotoxic T-cells [55, 56], offering additional protection against breast cancer. Conversely, we observed a positive association between breast cancer risk and the Protein EpiScore

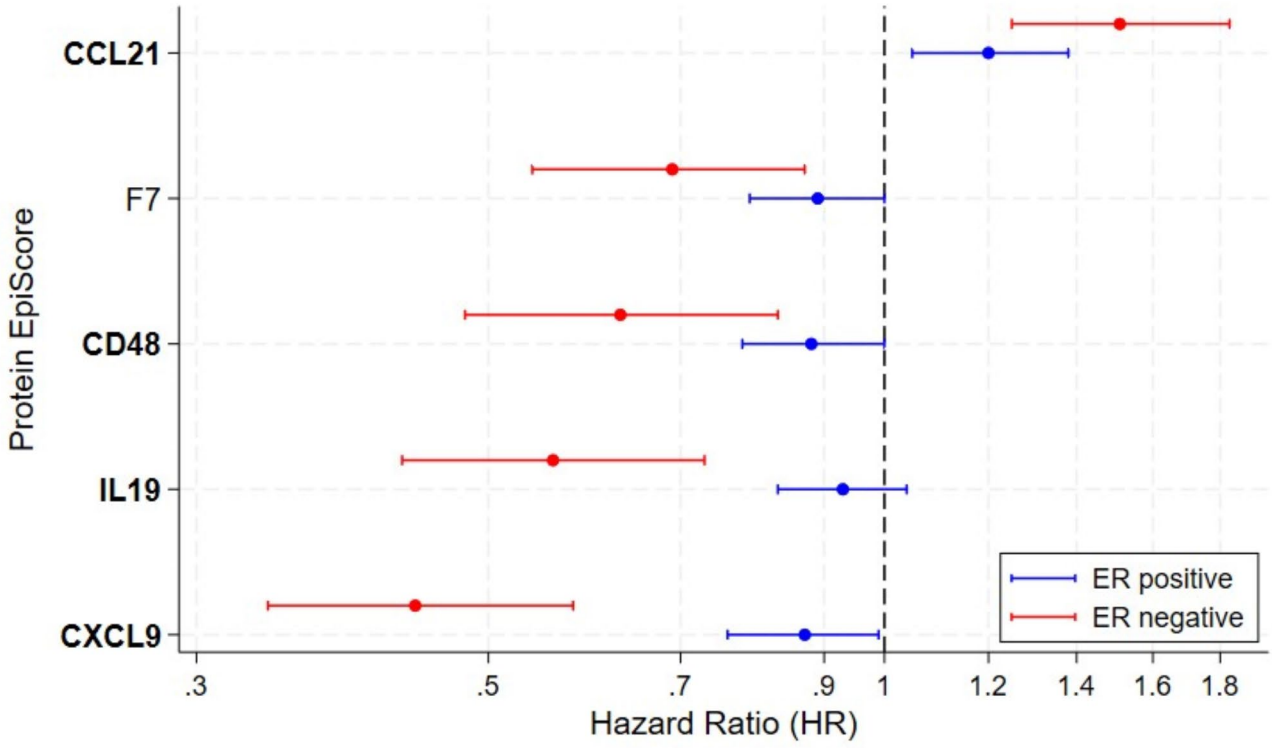

Fig. 4 Protein EpiScores identified in the primary analysis of invasive breast cancer risk that showed significant heterogeneity by tumor ER status. Hazard ratios and 95% confidence intervals displayed only for Protein EpiScores that were significantly associated with invasive breast cancer incidence in the primary analysis and significantly modified by tumor ER status (interaction FDR < 0.15). Models treat age as the time-scale and are therefore adjusted for age. The models additionally include DNAm platform (HumanMethylation450, MethylationEPIC) and self-reported race (non-Hispanic White, Black) as covariates. Bolded Protein EpiScores are those that were designed to predict plasma proteins related to 'immune response.'

for CCL21, a signaling molecule that may reduce the immune system's response to malignant cells by activating immunosuppressive leukocyte subsets (e.g., T regulatory cells) [57]. Notably, in the set of Protein EpiScores associated with short-term breast cancer risk, a majority were related to immune response and displayed inverse associations. There was also considerable variation in Protein EpiScore associations by tumor subtype, with Protein EpiScores consistently showing stronger associations with ER-negative breast cancers. This could reflect the higher immunogenicity of ER-negative breast cancers, which tend to elicit stronger immune responses [58].

This study is not without limitations. First, we lacked direct measurements of plasma proteins in the Sister Study, so we could not compare association estimates between Protein EpiScores and directly measured proteins. Second, although we observed statistical heterogeneity by ER status, our sample size was insufficient to examine more finely stratified associations, such as those among women diagnosed with triple-negative breast cancer. Third, Protein EpiScores were derived in populations of European ancestry and their translation to individuals with different ancestries has not yet been examined in detail. Lastly, because the Sister Study enrolled only women with a first-degree family history of breast cancer, the findings may not be generalizable to women without a family history. Despite these limitations, this study benefits from a large study population, the prospective case-cohort study design, and the exploration of a novel class of DNAm-based metrics.

In summary, we find that Protein EpiScores are significantly associated with invasive breast cancer risk, with most related to immune response. Interestingly, the associations appeared stronger for short-term breast cancer risk and for women diagnosed with ER-negative breast cancer. This study represents the first large-scale investigation of Protein EpiScores and breast cancer risk. Findings can help prioritize protein targets for future research and offer insights into the leukocyte transcriptional programs related to breast cancer development.

Acknowledgements
Not applicable.

Author contributions
J.K.K. and J.A.T. conceptualized the research question. B.M.R. performed the data analysis. J.K.K. and J.A.T. drafted the manuscript. All authors aided in the interpretation of the study findings, provided additional edits, and approved of the manuscript.

Funding
This research was supported by the Miles for Moffitt (Center for Women's Oncology) pilot study mechanism and the Intramural Research Program of the National Institutes of Health (National Institute of Environmental Health Sciences Z01-ES049033, Z01-ES049032, Z01-ES044005).

Data availability
A limited dataset for replication purposes can be requested via the Sister Study website: https://sisterstudy.niehs.nih.gov/English/coll-data.htm.

Declarations

Ethical approval and consent to participate
Written informed consent from all participants was collected at the home visit, and the study is overseen by the Institutional Review Board of the National Institutes of Health. Research activities were performed in accordance with the Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 4 December 2024 / Accepted: 16 March 2025
Published online: 26 March 2025

References

1. Guan Z, Yu H, Cuk K, et al. Whole-Blood DNA Methylation Markers in Early Detection of Breast Cancer: A Systematic Literature Review. Cancer Epidemiol Biomarkers Prev. 2019;28(3):496–505.
2. Xu Z, Sandler DP, Taylor JA. Blood DNA methylation and breast cancer: A prospective case-cohort analysis in the Sister Study. J Natl Cancer Inst. 2019. https://doi.org/10.1093/jnci/djz065.
3. Xu Z, Bolick SC, DeRoo LA, et al. Epigenome-wide association study of breast cancer using prospectively collected sister study samples. J Natl Cancer Inst. 2013;105(10):694–700.
4. Johansson A, Flanagan JM. Epigenome-wide association studies for breast cancer risk and risk factors. Trends Cancer Res. 2017;12:19–28.
5. Tang Q, Holland-Letz T, Slynko A, et al. DNA methylation array analysis identifies breast cancer associated RPTOR, MGRN1 and RAPSN hypomethylation in peripheral blood DNA. Oncotarget. 2016;7(39):64191–202.
6. van Veldhoven K, Polidoro S, Baglietto L, et al. Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. Clin Epigenetics. 2015;7:67.
7. Severi G, Southey MC, English DR, et al. Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer. Breast Cancer Res Treat. 2014;148(3):665–73.
8. Brennan K, Garcia-Closas M, Orr N, et al. Intragenic ATM methylation in peripheral blood DNA as a biomarker of breast cancer risk. Cancer Res. 2012;72(9):2304–13.
9. Kresovich JK, O'Brien KM, Xu Z, et al. Prediagnostic Immune Cell Profiles and Breast Cancer. JAMA Netw Open. 2020;3(1):e1919536.
10. Kresovich JK, Xu Z, O'Brien KM, et al. Methylation-based biological age and breast cancer risk. J Natl Cancer Inst. 2019;111(10):1051–8.
11. Kresovich JK, Xu Z, O'Brien KM, et al. Epigenetic mortality predictors and incidence of breast cancer. Aging. 2019;11(24):11975–87.
12. Kresovich JK, Xu Z, O'Brien KM, et al. Blood DNA methylation profiles improve breast cancer prediction. Mol Oncol. 2021. https://doi.org/10.1002/1878-0261.13087.
13. Lu AT, Quach A, Wilson JG, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging. 2019;11(2):303–27.
14. Stevenson AJ, Gadd DA, Hillary RF, et al. Creating and Validating a DNA Methylation-Based Proxy for Interleukin-6. J Gerontol Biol Sci Med Sci. 2021;76(12):2284–92.
15. Corley J, Cox SR, Harris SE, et al. Epigenetic signatures of smoking associate with cognitive function, brain structure, and mental and physical health outcomes in the Lothian Birth Cohort 1936. Transl Psychiatry. 2019;9(1):248.
16. Hamilton OKL, Zhang Q, McRae AF, et al. An epigenetic score for BMI based on DNA methylation correlates with poor physical health and major disease in the Lothian Birth Cohort. Int J Obes (Lond). 2019;43(9):1795–802.
17. Verschoor CP, Vlasschaert C, Rauh MJ, et al. A DNA methylation based measure outperforms circulating CRP as a marker of chronic inflammation and partly reflects the monocytic response to long-term inflammatory exposure: A Canadian Longitudinal Study on Aging analysis. Aging Cell. 2023;22(7):e13863.
18. Bernabeu E, McCartney DL, Gadd DA, et al. Refining epigenetic prediction of chronological and biological age. Genome Med. 2023;15(1):12.
19. Zhang Y, Elgizouli M, Schottker B, et al. Smoking-associated DNA methylation markers predict lung cancer incidence. Clin Epigenetics. 2016;8:127.
20. McCartney DL, Hillary RF, Stevenson AJ, et al. Epigenetic prediction of complex traits and death. Genome Biol. 2018;19(1):136.
21. Lu AT, Seeboth A, Tsai PC, et al. DNA methylation-based estimator of telomere length. Aging. 2019;11(16):5895–923.
22. Gadd DA, Hillary RF, McCartney DL et al. Epigenetic scores for the circulating proteome as tools for disease prediction. Elife 2022;11.
23. Suhre K, Arnold M, Bhagwat AM, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017;8:14357.
24. Hillary RF, Trejo-Banos D, Kousathanas A, et al. Multi-method genome- and epigenome-wide studies of inflammatory protein levels in healthy older adults. Genome Med. 2020;12(1):60.
25. Hillary RF, McCartney DL, Harris SE, et al. Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. Nat Commun. 2019;10(1):3160.
26. Smith HM, Moodie JE, Monterrubio-Gomez K, et al. Epigenetic scores of blood-based proteins as biomarkers of general cognitive function and brain health. Clin Epigenetics. 2024;16(1):46.
27. Chybowska AD, Gadd DA, Cheng Y, et al. Epigenetic Contributions to Clinical Risk Prediction of Cardiovascular Disease. Circ Genom Precis Med. 2024;17(1):e004265.
28. Sandler DP, Hodgson ME, Deming-Halverson SL, et al. The Sister Study Cohort: Baseline Methods and Participant Characteristics. Environ Health Perspect. 2017;125(12):127003.
29. Weinberg CR, Shore DL, Umbach DM, et al. Using risk-based sampling to enrich cohorts for endpoints, genes, and exposures. Am J Epidemiol. 2007;166(4):447–55.
30. Kresovich JK, Sandler DP, Taylor JA. Methylation-Based Biological Age and Hypertension Prevalence and Incidence. Hypertension. 2023;80(6):1213–22.
31. Xu Z, Niu L, Li L, et al. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Res. 2016;44(3):e20.
32. Xu Z, Langie SA, De Boever P, et al. RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. BMC Genomics. 2017;18(1):4.
33. Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. Bioinformatics. 2016;32(17):2659–63.
34. Xu Z, Niu L, Kresovich JK et al. methscore: a comprehensive R function for DNA methylation-based health predictors. Bioinformatics 2024;40(5).
35. D'Aloisio AA, Nichols HB, Hodgson ME, et al. Validity of self-reported breast cancer characteristics in a nationwide cohort of women with a family history of breast cancer. BMC Cancer. 2017;17(1):692.
36. O'Brien KM, Lawrence KG, Keil AP. The Case for Case-Cohort: An Applied Epidemiologist's Guide to Reframing Case-Cohort Studies to Improve Usability and Flexibility. Epidemiology. 2022;33(3):354–61.
37. Thiebaut AC, Benichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. Stat Med. 2004;23(24):3803–20.
38. Salas LA, Zhang Z, Koestler DC, et al. Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. Nat Commun. 2022;13(1):761.
39. Xue X, Kim MY, Gaudet MM, et al. A comparison of the polytomous logistic regression and joint cox proportional hazards models for evaluating multiple disease subtypes in prospective cohort studies. Cancer Epidemiol Biomarkers Prev. 2013;22(2):275–85.
40. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Royal Stat Soc Ser B (Methodological). 1995;57:289–300.
41. Oh HS, Rutledge J, Nachun D, et al. Organ aging signatures in the plasma proteome track health and disease. Nature. 2023;624(7990):164–72.
42. Tanaka T, Biancotto A, Moaddel R, et al. Plasma proteomic signature of age in healthy humans. Aging Cell. 2018;17(5):e12799.

43. Kim CH, Tworoger SS, Stampfer MJ, et al. Stability and reproducibility of proteomic profiles measured with an aptamer-based platform. Sci Rep. 2018;8(1):8382.
44. Haslam DE, Li J, Dillon ST, et al. Stability and reproducibility of proteomic profiles in epidemiological studies: comparing the Olink and SOMAscan platforms. Proteomics. 2022;22(13–14):e2100170.
45. Flanagan JM, Brook MN, Orr N, et al. Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study. Cancer Epidemiol Biomarkers Prev. 2015;24(1):221–9.
46. Ngo D, Benson MD, Long JZ et al. Proteomic profiling reveals biomarkers and pathways in type 2 diabetes risk. JCI Insight 2021;6(5).
47. Gudmundsdottir V, Zaghlool SB, Emilsson V, et al. Circulating Protein Signatures and Causal Candidates for Type 2 Diabetes. Diabetes. 2020;69(8):1843–53.
48. Elhadad MA, Jonasson C, Huth C, et al. Deciphering the Plasma Proteome of Type 2 Diabetes. Diabetes. 2020;69(12):2766–78.
49. Ganz P, Heidecker B, Hveem K, et al. Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. JAMA. 2016;315(23):2532–41.
50. Serban KA, Pratte KA, Bowler RP. Protein Biomarkers for COPD Outcomes. Chest. 2021;159(6):2244–53.
51. Panagiotou G, Papakonstantinou E, Vagionas A, et al. Serum Levels of Activins, Follistatins, and Growth Factors in Neoplasms of the Breast: A Case-Control Study. J Clin Endocrinol Metab. 2019;104(2):349–58.
52. Song Y, Zhu X, Lin Z, et al. The potential value of serum chemerin in patients with breast cancer. Sci Rep. 2021;11(1):6564.
53. McArdel SL, Terhorst C, Sharpe AH. Roles of CD48 in regulating immunity and tolerance. Clin Immunol. 2016;164:10–20.
54. Borsig L. Selectins in cancer immunity. Glycobiology. 2018;28(9):648–55.
55. Liang YK, Deng ZK, Chen MT, et al. CXCL9 Is a Potential Biomarker of Immune Infiltration Associated With Favorable Prognosis in ER-Negative Breast Cancer. Front Oncol. 2021;11:710286.
56. Pabani A, Gainor JF. Facts and Hopes: Immunocytokines for Cancer Immunotherapy. Clin Cancer Res. 2023;29(19):3841–9.
57. Masih M, Agarwal S, Kaur R, et al. Role of chemokines in breast cancer. Cytokine. 2022;155:155909.
58. Loizides S, Constantinidou A. Triple negative breast cancer: Immunogenicity, tumor microenvironment, and immunotherapy. Front Genet. 2022;13:1095839.

Publisher's note