Hormones and Cancer
https://doi.org/10.1007/s12672-019-00372-3

ORIGINAL PAPER

A Computational Statistics Approach to Evaluate Blood Biomarkers for Breast Cancer Risk Stratification

Kaan Oktay[1] · Ashlie Santaliz-Casiano[2] · Meera Patel[3] · Natascia Marino[3,4] · Anna Maria V. Storniolo[3,4] · Hamdi Torun[5] · Burak Acar[1] · Zeynep Madak Erdogan[2,6,7,8,9,10]

Received: 26 September 2019 / Accepted: 25 November 2019

Abstract

Breast cancer is the second leading cause of cancer mortality among women. Mammography and tumor biopsy followed by histopathological analysis are the current methods to diagnose breast cancer. Mammography does not detect all breast tumor subtypes, especially those that arise in younger women or women with dense breast tissue, and are more aggressive. There is an urgent need to find circulating prognostic molecules and liquid biopsy methods for breast cancer diagnosis and reducing the mortality rate. In this study, we systematically evaluated metabolites and proteins in blood to develop a pipeline to identify potential circulating biomarkers for breast cancer risk. Our aim is to identify a group of molecules to be used in the design of portable and low-cost biomarker detection devices. We obtained plasma samples from women who are cancer free (healthy) and women who were cancer free at the time of blood collection but developed breast cancer later (susceptible). We extracted potential prognostic biomarkers for breast cancer risk from plasma metabolomics and proteomics data using statistical and discriminative power analyses. We pre-processed the data to ensure the quality of subsequent analyses, and used two main feature selection methods to determine the importance of each molecule. After further feature elimination based on pairwise dependencies, we measured the performance of logistic regression classifier on the

Kaan Oktay and Ashlie Santaliz-Casiano contributed equally to this work.

✉ Zeynep Madak Erdogan
zmadake2@illinois.edu

Kaan Oktay
kaan.oktay@boun.edu.tr

Ashlie Santaliz-Casiano
ashlies2@illinois.edu

Meera Patel
patel88253@gmail.com

Natascia Marino
marinon@iu.edu

Anna Maria V. Storniolo
astornio@iu.edu

Hamdi Torun
hamdi.torun@northumbria.ac.uk

Burak Acar
acarbu@boun.edu.tr

[1] VAVlab, Electrical & Electronics Engineering Department, Bogazici University, Istanbul, Turkey

[2] Division of Nutritional Sciences, University of Illinois Urbana-Champaign, Urbana, IL, USA

[3] Susan G. Komen Tissue Bank at the IU Simon Cancer Center, Indianapolis, IN, USA

[4] Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA

[5] Faculty of Engineering and Environment, University of Northumbria, Newcastle upon Tyne, UK

[6] Department of Food Sciences and Human Nutrition, University of Illinois Urbana-Champaign, Urbana, IL, USA

[7] National Center for Supercomputing Applications, University of Illinois Urbana-Champaign, Urbana, IL, USA

[8] Cancer Center at Illinois, University of Illinois Urbana-Champaign, Urbana, IL, USA

[9] Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, Urbana, IL, USA

[10] Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, Urbana, IL, USA

Published online: 19 December 2019

remaining molecules and compared their biological relevance. We identified six signatures that predicted breast cancer risk with different specificity and selectivity. The best performing signature had 13 factors. We validated the difference in level of one of the biomarkers, SCF/KITLG, in plasma from healthy and susceptible individuals. These biomarkers will be used to develop low-cost liquid biopsy methods toward early identification of breast cancer risk and hence decreased mortality. Our findings provide the knowledge basis needed to proceed in this direction.

Keywords Liquid biopsy · Breast cancer risk · Circulating biomarker · Machine learning · Feature selection

Abbreviations

ERα	Estrogen receptor alpha
PgR	Progesterone receptor
HER2	Human epidermal growth factor receptor
(BRCA1 and 2)	Breast cancer susceptibility 1 and 2
CTCs	Circulating tumor cells
ctDNA	Circulating tumor DNA
IRB	Institutional review board
UIUC	University of Illinois, Urbana-Champaign
GC/MS	Gas chromatography–mass spectrometry
AUC	Area under curve
ROC	Receiver operator characteristic
pCC	Pearson's correlation coefficient
LR	Logistic regression
ELISA	Enzyme-linked immunosorbent assay
SCF/KITLG	Stem cell factor/KIT ligand
HRP	Horseradish peroxidase
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
BMI	Body mass index

Background

Breast cancer is the second leading cause of death among adult women. According to World Health Organization, there is a sharp rise in overall number of breast cancer incidences worldwide due to changes in lifestyle, reproductive factors, and increased life expectancy [1]. Fifty eight percent of all breast cancer–related deaths occur in middle- and low-income countries. While survival rates for breast cancer are around 80% in developed countries, this rate decreases to 60% in middle-income and to 40% in low-income countries due to lack of early detection programs leading to diagnoses in late stages, where 80% of these tumors are incurable [2, 3]. In the middle- and low-income countries, mammography and other expensive and technologically complicated methods are unattainable due to high costs and shortage of trained personnel [4, 5]. Moreover, mammograms are more likely to detect ER-positive breast cancer [6] and are not recommended for younger women. In addition, diagnosis at an earlier stage using conventional procedures is not prognostic for all race groups, for example, the probability of an African-American woman with small-sized tumors presenting with metastasis is higher than that of a Caucasian woman [7]. Thus, there is a critical need for affordable, portable, and accurate means of detecting breast cancer risk before the tumors arise. Development of such technologies has the potential to expedite the solution for the growing health problem to prevent increasing death and disability among women especially in low- and middle-income countries.

Currently, a handful of biomarkers are used in the clinic for breast cancer diagnosis. These biomarkers are proteins overexpressed in certain subtypes of breast tumors and help clinicians plan treatment. Up to date, a limited number of breast cancer biomarkers demonstrated clinical utility, including estrogen receptor alpha (ERα), progesterone receptor (PgR) [8], and human epidermal growth factor receptor 2 (HER2) to predict effectiveness of systemic therapy and the Oncotype DX-21 gene score to predict benefits of chemotherapy [9–11]. Studies evaluating other predictive biomarkers are in progress for breast cancer susceptibility genes (BRCA1 and BRCA2) circulating tumor cells (CTCs), HER2 (+), TOP2A (in subjects with HER2 overexpression), and HER2 (when negative in tumors but positive in the CTCs) [12]. Circulating tumor DNA (ctDNA) is increasingly used in the clinic, particularly for advanced solid tumors [13–15]. However, clinical utility and validity of ctDNA assays in early stage cancers is not as clear [15]. Further, we still lack reliable biomarkers to detect breast cancer risk before the tumors arise. Lack of such biomarkers hinders establishment of reliable screening or prevention programs.

To address this critical need, we systematically evaluated metabolites and proteins in plasma to identify potential biomarkers for breast cancer risk that can be utilized to develop minimally invasive, affordable, portable, and accurate screening devices. In this study, our focus is on liquid biopsy samples from plasma that have the potential to provide simple and minimally invasive information for diagnostic decisions. We developed an efficient pipeline to analyze liquid biopsy samples, to detect blood biomarkers, and to identify the risk for breast cancer before tumors arise. This pipeline paves the way toward developing the aforementioned screening devices to be used in the field by basic-level healthcare workers in low-resource environments.

Methods

Patients and Plasma Samples

All studies were approved by the Indiana University Institutional Review Board (IRB protocol number 1011003097). All research was carried out in compliance with the Helsinki Declaration. Donors provided broad written consent for the use of their specimens in research. The written consent document informed the donor that the donated specimens and medical data would be used for the general purpose of helping to determine how breast cancer develops. It was explained in the written consent that the exact laboratory experiments were unknown at the time of donation, and that proposals for use of the specimens would be reviewed and approved by a panel of independent researchers before specimens and/or data were released for research purposes. Hematoxylin and eosin–stained sections of the FFPE tissue of the identified donors were reviewed by a pathologist to confirm the absence of histological abnormalities. In order to exclude or control confounding variables such as age, racial and ethnic background, and menopausal status, the subjects in the two cohorts, susceptible and healthy controls, were matched by selection of the comparison group (healthy controls) with respect to the distribution of the aforementioned confounders in susceptible group.

Plasma Preparation

Blood was drawn into the Plasma Separator tube (Vacutainer Venous Blood Collection Tubes; SST* Plasma Separation Tube, Fisher Scientific cat. #0268396) and gently mixed by inverting the tube five times. Forty-five minutes (±10 min) after the blood had been drawn, the Plasma Separator Tube was placed into a minicentrifuge (Eppendorf centrifuge 5702) and centrifuged at 1200 rcf for 10 min at room temperature. A repeater pipette was used to aliquot 600 μl of the plasma into each of five cryogenic vials. Samples were stored at − 80 °C until use.

OLINK Protein Biomarker and Whole Metabolite Profiling Assays

All the samples from human studies were handled and analyzed in accordance with UIUC IRB protocol #06741 and as previously described [16]. Ten microliters of plasma samples from Komen Tissue Bank was submitted to OLINK biosciences for cancer and inflammation biomarker analysis. In total, 50 μl of plasma samples was submitted to the Metabolomics Center at UIUC. GC/MS whole metabolite profiling was performed to detect and quantify the metabolites by using gas chromatography–mass spectrometry (GC/MS) analysis. Metabolites were extracted from 50 μl of plasma according to Agilent Inc. application notes. Hentriacontanoic acid was added to each sample as the internal standard prior to derivatization. Metabolite profiles were acquired using an Agilent GC/MS system (Agilent 7890 gas chromatograph, an Agilent 5975 MSD, and an HP 7683B autosampler). The spectra of all chromatogram peaks were evaluated using the AMDIS 2.71 and a custom-built database with 460 unique metabolites. All known artificial peaks were identified and removed prior to data analysis. To allow the comparison between samples, all data were normalized to the internal standard in each chromatogram.

Statistical Analysis

Preprocessing of Measurements

We normalized all individuals' plasma data in each dataset with respect to the healthy individuals' data in the respective dataset to factor out potential differences in data acquisition. More specifically, we performed the following procedure separately for both datasets. For each molecule in a dataset, we subtracted the mean measurement of that molecule in healthy individuals from all individuals' measurements and divided this difference by the standard deviation of that molecule's measurements in healthy individuals. Thus, we converted each single measurement to a z-score which describes the deviation of that measurement from the mean of healthy individuals', in terms of the standard deviation among healthy individuals. As the final step, we merged two datasets, which were normalized with respect to their own healthy individuals, and obtained a dataset with 49 susceptible and 47 healthy individuals.

Molecule Ranking, Elimination, and Performance Assessment

A two-stage procedure is applied to identify the molecule sets with high discriminative power between the healthy and the susceptible groups. The first stage involves ranking all molecules with respect to their individual discriminative powers (importance ranking). The second stage involves molecule elimination (selection) based on their interdependencies.

To independently assess each of 181 molecules, we used two different methods. In the first method, we applied Student's t test to test the null hypothesis that the measurements in the two groups come from the same distribution. All molecules were ranked based on the corresponding p-values to get a short-list of the top-ranking 20 molecules with the lowest p values, discarding the others from further processing. In the second method, we applied the random forest algorithm to assess the discriminative power of each of the 181 molecules individually by using the mean decrease impurity (Gini importance), which is defined as the mean decrease in node impurity over all the trees in the forest. This time, all molecules were ranked based on their

Table 1 The p values of the paired t test analysis for each pair of molecules among the top 20 molecules ranked by applying Student's t test to all 181 molecules

	SCF	MAD HOMOLOG 5	FGF-5	FASLG	MMP-10	PPY	XPNPEP2	FGF-21	CXL17	MCP-3	ESM-1	HK11	TRAIL	FGF-BP1	EN-RAGE	C15:0	TNFB	CTSV	ADA	CD160
SCF	0	0.6963	0.0004	0.0006	0.3286	0.0017	0.361	0.003	0.2659	0.189	0.0003	0.1552	0.1309	0.1303	0.1191	0.1324	0.0029	0.0045	0.001	0.0015
MAD HOMOLOG 5	0.6963	0	0.0013	0.0014	0.6643	0.0038	0.5633	0.0026	0.4489	0.2985	0.0006	0.3257	0.5298	0.2913	0.354	0.2454	0.008	0.0062	0.0083	0.007
FGF-5	0.0004	0.0013	0	0.6618	0.002	0.7183	0.0018	0.6703	0.0017	0.0027	0.6315	0.0064	0.0033	0.01	0.0027	0.0091	0.4087	0.4174	0.6858	0.4709
FASLG	0.0006	0.0014	0.6618	0	0.0018	0.9554	0.0015	0.9738	0.0034	0.002	0.9137	0.0015	0.0039	0.0112	0.005	0.0067	0.6151	0.548	0.9595	0.6497
MMP-10	0.3286	0.6643	0.002	0.0018	0	0.0033	0.9115	0.0006	0.8086	0.542	0.0073	0.5825	0.8731	0.5308	0.6119	0.417	0.0071	0.0154	0.0169	0.0086
PPY	0.0017	0.0038	0.7183	0.9554	0.0033	0	0.0054	0.9356	0.0018	0.0055	0.8852	0.0063	0.0096	0.0166	0.0124	0.0109	0.6315	0.6427	0.93	0.6801
XPNPEP2	0.361	0.5633	0.0018	0.0015	0.9115	0.0054	0	0.0034	0.9172	0.66	0.0125	0.6878	0.9708	0.6209	0.6971	0.4563	0.0108	0.0055	0.0188	0.0214
FGF-21	0.003	0.0026	0.6703	0.9738	0.0006	0.9356	0.0034	0	0.0014	0.0029	0.9489	0.006	0.0063	0.013	0.0134	0.0096	0.7003	0.6644	0.9805	0.7499
CXL17	0.2659	0.4489	0.0017	0.0034	0.8086	0.0018	0.9172	0.0014	0	0.7384	0.0038	0.7438	0.9536	0.6714	0.7648	0.5413	0.0172	0.026	0.019	0.0126
MCP-3	0.189	0.2985	0.0027	0.002	0.542	0.0055	0.66	0.0029	0.7384	0	0.0089	0.9712	0.7017	0.9032	0.9911	0.788	0.0183	0.017	0.0211	0.0241
ESM-1	0.0003	0.0006	0.6315	0.9137	0.0073	0.8852	0.0125	0.9489	0.0038	0.0089	0	0.0032	0.0061	0.013	0.0112	0.0289	0.7453	0.7257	0.9715	0.7768
HK11	0.1552	0.3257	0.0064	0.0015	0.5825	0.0063	0.6878	0.006	0.7438	0.9712	0.0032	0	0.7392	0.8734	0.9815	0.7421	0.0202	0.0234	0.037	0.0087
TRAIL	0.1309	0.5298	0.0033	0.0039	0.8731	0.0096	0.9708	0.0063	0.9536	0.7017	0.0061	0.7392	0	0.6469	0.6816	0.5747	0.0092	0.0274	0.0071	0.0078
FGF-BP1	0.1303	0.2913	0.01	0.0112	0.5308	0.0166	0.6209	0.013	0.6714	0.9032	0.013	0.8734	0.6469	0	0.9125	0.8767	0.0416	0.0442	0.0429	0.0441
EN-RAGE	0.1191	0.354	0.0027	0.005	0.6119	0.0124	0.6971	0.0134	0.7648	0.9911	0.0112	0.9815	0.6816	0.9125	0	0.8043	0.0374	0.025	0.0115	0.0327
C15:0	0.1324	0.2454	0.0091	0.0067	0.417	0.0109	0.4563	0.0096	0.5413	0.788	0.0289	0.7421	0.5747	0.8767	0.8043	0	0.0409	0.0464	0.0725	0.0502
TNFB	0.0029	0.008	0.4087	0.6151	0.0071	0.6315	0.0108	0.7003	0.0172	0.0183	0.7453	0.0202	0.0092	0.0416	0.0374	0.0409	0	0.969	0.7488	0.9447
CTSV	0.0045	0.0062	0.4174	0.548	0.0154	0.6427	0.0055	0.6644	0.026	0.017	0.7257	0.0234	0.0274	0.0442	0.025	0.0464	0.969	0	0.7132	0.9266
ADA	0.001	0.0083	0.6858	0.9595	0.0169	0.93	0.0188	0.9805	0.019	0.0211	0.9715	0.037	0.0071	0.0429	0.0115	0.0725	0.7488	0.7132	0	0.8146
CD160	0.0015	0.007	0.4709	0.6497	0.0086	0.6801	0.0214	0.7499	0.0126	0.0241	0.7768	0.0087	0.0078	0.0441	0.0327	0.0502	0.9447	0.9266	0.8146	0

The paired t test assesses the pairwise dependencies of the most discriminative 20 molecules. Pairs with $p > 0.05$ show strong dependency within that pair

Table 2 The p values of the paired t-test analysis for each pair of molecules among the top 20 molecules ranked by applying the random forest to all 181 molecules

	SCF	MAD HOMOLOG 5	PPY	FASLG	FGF-5	CXCL1	MMP-10	XPNPEP2	ESM-1	PHOSPHORIC ACID	PD-L1	EPHA2	FLT3L	4E-BP1	TRAIL	MCP-1	TLR3	CD27	FGF-BP1	HK14
SCF	0	0.6963	0.0017	0.0006	0.0004	0.0023	0.3286	0.361	0.0003	0.0859	0.0333	0.0073	0.0056	0.1307	0.1309	0.0148	0.0432	0.0116	0.1303	0.1215
MAD HOMOLOG 5	0.6963	0	0.0038	0.0014	0.0013	0.0153	0.6643	0.5633	0.0006	0.1622	0.146	0.0305	0.0442	0.2792	0.5298	0.0506	0.0586	0.0252	0.2913	0.2387
PPY	0.0017	0.0038	0	0.9554	0.7183	0.4656	0.0033	0.0054	0.8852	0.0413	0.0721	0.1935	0.2713	0.0397	0.0096	0.1833	0.2008	0.2943	0.0166	0.0205
FASLG	0.0006	0.0014	0.9554	0	0.6618	0.4225	0.0018	0.0015	0.9137	0.0376	0.0732	0.142	0.2782	0.0277	0.0039	0.166	0.1646	0.2553	0.0112	0.0051
FGF-5	0.0004	0.0013	0.7183	0.6618	0	0.3008	0.002	0.0018	0.6315	0.0209	0.0378	0.1085	0.1423	0.0145	0.0033	0.0857	0.0983	0.1787	0.01	0.0147
CXCL1	0.0023	0.0153	0.4656	0.4225	0.3008	0	0.0115	0.0376	0.5516	0.2144	0.2567	0.5558	0.6492	0.1072	0.0076	0.4841	0.5033	0.691	0.0583	0.0429
MMP-10	0.3286	0.6643	0.0033	0.0018	0.002	0.0115	0	0.9115	0.0073	0.3345	0.1225	0.0486	0.0419	0.5293	0.8731	0.0542	0.1309	0.0226	0.5308	0.4234
XPNPEP2	0.361	0.5633	0.0054	0.0015	0.0018	0.0376	0.9115	0	0.0125	0.3469	0.2814	0.1044	0.0921	0.5733	0.9708	0.1341	0.1636	0.0719	0.6209	0.519
ESM-1	0.0003	0.0006	0.8852	0.9137	0.6315	0.5516	0.0073	0.0125	0	0.0593	0.114	0.1401	0.3336	0.0335	0.0061	0.2512	0.2426	0.3007	0.013	0.0129
PHOSPHORIC ACID	0.0859	0.1622	0.0413	0.0376	0.0209	0.2144	0.3345	0.3469	0.0593	0	0.829	0.4717	0.446	0.7561	0.422	0.517	0.5868	0.3887	0.6436	0.762
PD-L1	0.0333	0.146	0.0721	0.0732	0.0378	0.2567	0.1225	0.2814	0.114	0.829	0	0.5517	0.4038	0.5749	0.1934	0.5838	0.7439	0.415	0.5413	0.547
EPHA2	0.0073	0.0305	0.1935	0.142	0.1085	0.5558	0.0486	0.1044	0.1401	0.4717	0.5517	0	0.8913	0.3318	0.0648	0.9749	0.9147	0.7529	0.2438	0.2666
FLT3L	0.0056	0.0442	0.2713	0.2782	0.1423	0.6492	0.0419	0.0921	0.3336	0.446	0.4038	0.8913	0	0.2581	0.0333	0.8507	0.8279	0.94	0.2181	0.2617
4E-BP1	0.1307	0.2792	0.0397	0.0277	0.0145	0.1072	0.5293	0.5733	0.0335	0.7561	0.5749	0.3318	0.2581	0	0.5541	0.1918	0.4236	0.2596	0.924	0.9861
TRAIL	0.1309	0.5298	0.0096	0.0039	0.0033	0.0076	0.8731	0.9708	0.0061	0.422	0.1934	0.0648	0.0333	0.5541	0	0.066	0.1962	0.0647	0.6469	0.5674
MCP-1	0.0148	0.0506	0.1833	0.166	0.0857	0.4841	0.0542	0.1341	0.2512	0.517	0.5838	0.9749	0.8507	0.1918	0.066	0	0.9421	0.8181	0.2741	0.3109
TLR3	0.0432	0.0586	0.2008	0.1646	0.0983	0.5033	0.1309	0.1636	0.2426	0.5868	0.7439	0.9147	0.8279	0.4236	0.1962	0.9421	0	0.7566	0.3418	0.3587
CD27	0.0116	0.0252	0.2943	0.2553	0.1787	0.691	0.0226	0.0719	0.3007	0.3887	0.415	0.7529	0.94	0.2596	0.0647	0.8181	0.7566	0	0.1532	0.1312
FGF-BP1	0.1303	0.2913	0.0166	0.0112	0.01	0.0583	0.5308	0.6209	0.013	0.6436	0.5413	0.2438	0.2181	0.924	0.6469	0.2741	0.3418	0.1532	0	0.8822
HK14	0.1215	0.2387	0.0205	0.0051	0.0147	0.0429	0.4234	0.519	0.0129	0.762	0.547	0.2666	0.2617	0.9861	0.5674	0.3109	0.3587	0.1312	0.8822	0

The paired t test assesses the pairwise dependencies of the most discriminative 20 molecules. Pairs with $p > 0.05$ show strong dependency within that pair

Table 3 Pearson's correlation coefficient between each pair of 20 most important molecules ranked by their Student's *t*-test *p* values

	SCF	MAD HOMOLOG 5	FGF-5	FASLG	MMP-10	PPY	XPNPEP2	FGF-21	CXL17	MCP-3	ESM-1	HK11	TRAIL	FGF-BP1	EN-RAGE	C15:0	TNFB	CTSV	ADA	CD160
SCF	1	0.136	0.129	0.076	0.318	−0.088	0.03	−0.189	0.215	0.03	0.265	0.261	0.66	0.228	0.415	0	0.073	−0.024	0.312	0.223
MAD HOMOLOG 5	0.136	1	0.057	0.064	−0.048	−0.119	0.148	0.021	0.323	0.217	0.305	0.208	0.244	0.118	0.132	0.017	0.012	0.083	0.089	0.082
FGF-5	0.129	0.057	1	0.21	0.047	0.15	0.123	0.104	0.151	0.116	0.067	−0.048	0.113	−0.073	0.233	−0.027	0.184	0.049	0.219	0.161
FASLG	0.076	0.064	0.21	1	0.079	0.169	0.179	0.192	0.037	0.165	0.149	0.246	0.127	−0.068	0.154	0.075	0.285	0.384	0.12	0.446
MMP-10	0.318	−0.048	0.047	0.079	1	−0.002	0.156	0.32	0.323	0.281	−0.087	0.264	0.294	0.076	0.143	0.097	0.159	−0.042	0.043	0.158
PPY	−0.088	−0.119	0.15	0.169	−0.002	1	−0.047	0.085	0.199	0.012	−0.006	0.022	−0.065	−0.147	−0.034	0.003	0.089	−0.175	0.035	0.173
XPNPEP2	0.03	0.148	0.123	0.179	0.156	−0.047	1	0.11	0.168	0.15	−0.169	0.188	−0.03	0.004	0.118	0.227	0.138	0.281	0.083	−0.018
FGF-21	−0.189	0.021	0.104	0.192	0.32	0.085	0.11	1	0.276	0.2	−0.124	0.092	0.101	0.003	0.005	0.103	0.016	0.038	0.065	0.097
CXL17	0.215	0.323	0.151	0.037	0.323	0.199	0.168	0.276	1	0.129	0.164	0.32	0.198	0.092	0.141	0.131	0.042	−0.096	0.119	0.17
MCP-3	0.03	0.217	0.116	0.165	0.281	0.012	0.15	0.2	0.129	1	0.02	0.068	0.214	0.112	0.319	−0.112	0.11	0.133	0.202	0.079
ESM-1	0.265	0.305	0.067	0.149	−0.087	−0.006	−0.169	−0.124	0.164	0.02	1	0.269	0.16	0.072	0.117	−0.187	0.075	−0.026	0.233	0.299
HK11	0.261	0.208	−0.048	0.246	0.264	0.022	0.188	0.092	0.32	0.068	0.269	1	0.176	0.187	0.047	0.13	0.112	0.071	0.007	0.352
TRAIL	0.66	0.244	0.113	0.127	0.294	−0.065	−0.03	0.101	0.198	0.214	0.16	0.176	1	0.134	0.435	−0.066	0.28	−0.009	0.379	0.333
FGF-BP1	0.228	0.118	−0.073	−0.068	0.076	−0.147	0.004	0.003	0.092	0.112	0.072	0.187	0.134	1	−0.055	0.008	−0.017	−0.034	0.062	0.006
EN-RAGE	0.415	0.132	0.233	0.154	0.143	−0.034	0.118	0.005	0.141	0.319	0.117	0.047	0.435	−0.055	1	−0.136	0.019	0.165	0.41	0.107
C15:0	0	0.017	−0.027	0.075	0.097	0.003	0.227	0.103	0.131	−0.112	−0.187	0.13	−0.066	0.008	−0.136	1	0.035	−0.012	−0.162	−0.001
TNFB	0.073	0.012	0.184	0.285	0.159	0.089	0.138	0.016	0.042	0.11	0.075	0.112	0.28	−0.017	0.019	0.035	1	0.048	0.204	0.275
CTSV	−0.024	0.083	0.049	0.384	−0.042	−0.175	0.281	0.038	−0.096	0.133	−0.026	0.071	−0.009	−0.034	0.165	−0.012	0.048	1	0.25	−0.087
ADA	0.312	0.089	0.219	0.12	0.043	0.035	0.083	0.065	0.119	0.202	0.233	0.007	0.379	0.062	0.41	−0.162	0.204	0.25	1	0.054
CD160	0.223	0.082	0.161	0.446	0.158	0.173	−0.018	0.097	0.17	0.079	0.299	0.352	0.333	0.006	0.107	−0.001	0.275	−0.087	0.054	1

Pairs with pCC > 0.5 show strong dependency within that pair

Table 4 Pearson's correlation coefficient between each pair of 20 most important molecules ranked by their importance calculated by random forest

	SCF	MAD HOMOLOG 5	PPY	FASLG	FGF-5	CXCL1	MMP-10	XPNPEP2	ESM-1	PHOSPHORIC ACID	PD-L1	EPHA2	FLT3L	4E-BP1	TRAIL	MCP-1	TLR3	CD27	FGF-BP1	HK14
SCF	1	0.136	−0.088	0.076	0.129	0.227	0.318	0.03	0.265	0.05	0.363	0.382	0.395	0.297	0.66	0.277	−0.011	0.131	0.228	0.159
MAD HOMOLOG 5	0.136	1	−0.119	0.064	0.057	−0.029	−0.048	0.148	0.305	0.067	0.061	0.271	0.105	0.231	0.244	0.152	0.15	0.133	0.118	0.157
PPY	−0.088	−0.119	1	0.169	0.15	−0.038	−0.002	−0.047	−0.006	−0.016	0.235	0.148	0.102	−0.014	−0.065	0.184	−0.017	−0.011	−0.147	−0.072
FASLG	0.076	0.064	0.169	1	0.21	0.19	0.079	0.179	0.149	−0.011	0.229	0.347	0.098	0.107	0.127	0.262	0.149	0.159	−0.068	0.246
FGF-5	0.129	0.057	0.15	0.21	1	0.044	0.047	0.123	0.067	0.09	0.28	0.218	0.253	0.18	0.113	0.319	0.178	0.081	−0.073	−0.068
CXCL1	0.227	−0.029	−0.038	0.19	0.044	1	0.195	−0.094	0.01	−0.158	0.317	0.197	0.263	0.202	0.439	0.402	0.126	0.235	0.066	0.319
MMP-10	0.318	−0.048	−0.002	0.079	0.047	0.195	1	0.156	−0.087	−0.111	0.499	0.352	0.348	0.101	0.294	0.369	0.047	0.343	0.076	0.209
XPNPEP2	0.03	0.148	−0.047	0.179	0.123	−0.094	0.156	1	−0.169	0.063	0.119	0.151	0.145	0.2	−0.03	0.071	0.02	0.048	0.004	0.121
ESM-1	0.265	0.305	−0.006	0.149	0.067	0.01	−0.087	−0.169	1	−0.015	0.166	0.489	0.117	0.179	0.16	0.136	0.029	0.23	0.072	0.196
PHOSPHORIC ACID	0.05	0.067	−0.016	−0.011	0.09	−0.158	−0.111	0.063	−0.015	1	−0.046	0.073	−0.066	0.082	−0.058	−0.015	−0.138	−0.166	−0.009	−0.196
PD-L1	0.363	0.061	0.235	0.229	0.28	0.317	0.499	0.119	0.166	−0.046	1	0.525	0.635	0.316	0.484	0.519	0.083	0.452	−0.053	0.296
EPHA2	0.382	0.271	0.148	0.347	0.218	0.197	0.352	0.151	0.489	0.073	0.525	1	0.358	0.165	0.431	0.368	0.17	0.642	0.064	0.195
FLT3L	0.395	0.105	0.102	0.098	0.253	0.263	0.348	0.145	0.117	−0.066	0.635	0.358	1	0.27	0.534	0.507	0.128	0.159	0.043	0.075
4E-BP1	0.297	0.231	−0.014	0.107	0.18	0.202	0.101	0.2	0.179	0.082	0.316	0.165	0.27	1	0.422	0.579	0.036	0.054	0.067	0.12
TRAIL	0.66	0.244	−0.065	0.127	0.113	0.439	0.294	−0.03	0.16	−0.058	0.484	0.431	0.534	0.422	1	0.466	0.01	0.217	0.134	0.13
MCP-1	0.277	0.152	0.184	0.262	0.319	0.402	0.369	0.071	0.136	−0.015	0.519	0.368	0.507	0.579	0.466	1	0.042	0.155	0.03	0.121
TLR3	−0.011	0.15	−0.017	0.149	0.178	0.126	0.047	0.02	0.029	−0.138	0.083	0.17	0.128	0.036	0.01	0.042	1	0.138	−0.125	0.083
CD27	0.131	0.133	−0.011	0.159	0.081	0.235	0.343	0.048	0.23	−0.166	0.452	0.642	0.159	0.054	0.217	0.155	0.138	1	0.05	0.316
FGF-BP1	0.228	0.118	−0.147	−0.068	−0.073	0.066	0.076	0.004	0.072	−0.009	−0.053	0.064	0.043	0.067	0.134	0.03	−0.125	0.05	1	0.239
HK14	0.159	0.157	−0.072	0.246	−0.068	0.319	0.209	0.121	0.196	−0.196	0.296	0.195	0.075	0.12	0.13	0.121	0.083	0.316	0.239	1

Pairs with pCC > 0.5 show strong dependency within that pair

Table 5 Ranking of molecules identified by initial Student's *t* test for each consequent feature elimination method

Rank	Student's *t* test	Manual selection by LR	Correlation analysis	Paired *t* test
1	SCF	SCF	SCF	SCF
2	MAD HOMOLOG 5	MAD HOMOLOG 5	MAD HOMOLOG 5	FGF-5
3	FGF-5	FGF-5	FGF-5	
4	FASLG	FASLG	FASLG	
5	MMP-10	MMP-10	PPY	
6	PPY	XPNPEP2	XPNPEP2	
7	XPNPEP2	FGF-21	FGF-21	
8	FGF-21	CXL17	MCP-3	
9	CXL17	MCP-3	FGF-BP1	
10	MCP-3	ESM-1	C15:0	
11	ESM-1	TNFB		
12	HK11	CTSV		
13	TRAIL	CD160		
14	FGF-BP1			
15	EN-RAGE			
16	C15:0			
17	TNFB			
18	CTSV			
19	ADA			
20	CD160			

The "Student's t test" column lists the top-ranking most discriminative 20 molecules among 181 molecules. The "Manual selection by LR," "Paired *t* test," and "Correlation analysis" columns list the molecules selected from these 20 top molecules by applying the iterative molecule elimination procedures using manual selection by LR based on classification performance, paired *t* test, and correlation analysis, respectively, as described in "Statistical Analysis" section

Gini importance values to get the top-ranking 20 molecules with the highest importance values. No further threshold was applied to these top-ranking molecules at this stage for both methods, as the low-ranking molecules in these lists may potentially have significant marginal contribution to a subset of molecules when used together.

To generate an optimum subset of the top 20 molecules identified by Student's *t* test or random forest, we used the following iterative procedure. We initialized a "selected molecules" list (S-list) with the top-ranking molecule and an "unselected molecules" list (U-list) with the remaining 19 ranked molecules. We iteratively assessed the individual molecules in the U-list with respect to the molecules set represented by the S-list and added the ones that have a positive contribution to the S-list while discarding the others. Three different approaches are applied to assess whether a molecule has a positive contribution to the S-list: (1) Manual selection: Logistic Regression (LR) classifiers, to identify healthy and susceptible groups, are trained and tested iteratively by using the *selected* molecules (S-list) and the top-ranking *unselected* molecule (U-list) as the features. The classifier performance is assessed using the selected molecules' AUC (area under curve) of ROC (receiver operator characteristic) curves. After each iteration, if the AUC is increased, the top-ranking *unselected* molecule is added to the S-list, otherwise discarded. The iterations stop when the U-list is exhausted. (2) Paired *t* test: The inter-molecule dependencies, as measured by the paired *t* test, is used to select the molecules from the U-list to be added to the S-list. We first computed the paired *t* test *p* values for each pair of molecules among the aforementioned top-ranking 20 molecules with the null hypothesis being that both come from the same distribution. Using these *p* values, we iteratively discarded the molecules from the U-list that have a *p* value larger than 0.05 when tested with anyone of the molecules from the S-list and moved the *unselected* molecule from U-list to S-list with the lowest maximum *p* value (< 0.05) when tested with the *selected* molecules. The iterations stop when the U-list is exhausted. (3) Correlation analysis: The second approach described above is repeated by replacing the null hypothesis testing with the correlation analysis as measured by Pearson's correlation coefficient (pCC). We used 0.5 as the pCC threshold.

Finally, we performed LR classification (4-fold cross-validation with 500 iterations) using the top-ranking N molecules in each list, where N runs from one to the length of the corresponding list. Of note, use of LR for performance assessment of classification at this last step is distinct from the earlier use of LR for manual selection of the molecules.

Table 6 Ranking of molecules identified by initial random forest method for each consequent feature elimination method

Rank	Random forest	Manual selection by LR	Correlation analysis	Paired *t* test
1	SCF	SCF	SCF	SCF
2	MAD HOMOLOG 5	MAD HOMOLOG 5	MAD HOMOLOG 5	PPY
3	PPY	PPY	PPY	
4	FASLG	FASLG	FASLG	
5	FGF-5	FGF-5	FGF-5	
6	CXCL1	MMP-10	CXCL1	
7	MMP-10	XPNPEP2	XPNPEP2	
8	XPNPEP2	ESM-1	PHOSPHORIC ACID	
9	ESM-1	FLT3L	TLR3	
10	PHOSPHORIC ACID	HK14	CD27	
11	PD-L1		FGF-BP1	
12	EPHA2			
13	FLT3L			
14	4E-BP1			
15	TRAIL			
16	MCP-1			
17	TLR3			
18	CD27			
19	FGF-BP1			
20	HK14			

The "Random forest" column lists the top-ranking most discriminative 20 molecules among 181 molecules. The "Manual selection by LR," "Paired *t* test," and "Correlation analysis" columns list the molecules selected from these 20 top molecules by applying the iterative molecule elimination procedures using manual selection by LR based on classification performance, paired *t* test, and correlation analysis, respectively, as described in "Statistical Analysis" section

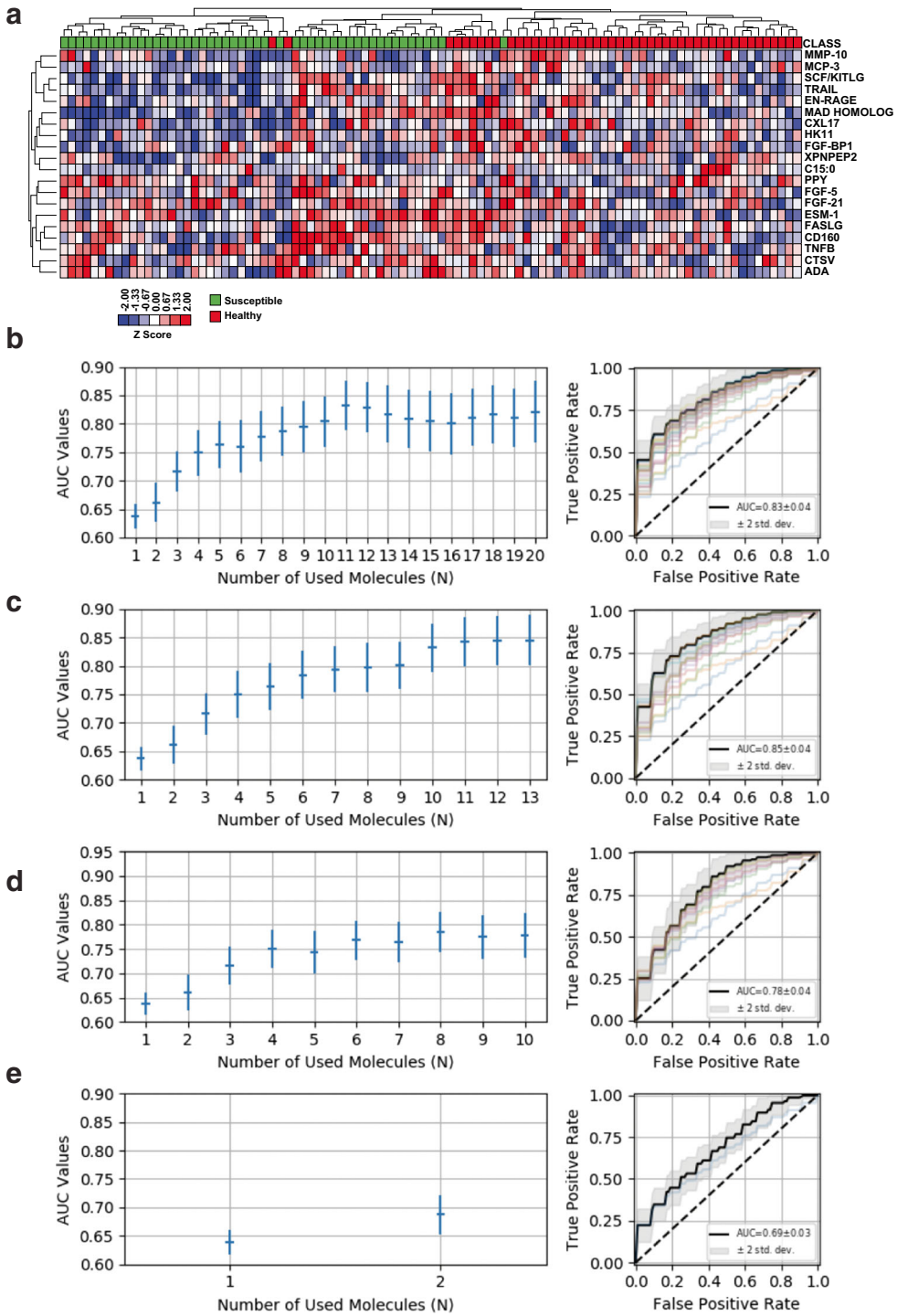
SCF/KITLG Quantification Using Enzyme-Linked Immunosorbent Assay (ELISA)

Plasma samples from both groups were collected and stored at − 80 °C until the time of assay. We used an ELISA kit for SCF/KITLG (Sigma, catalog no. RAB0330). Samples were diluted 2-fold per suggestion from the manufacturer. For the SCF/KITLG antibody, concentrate was diluted 100-fold with 1× diluent buffer. To prepare the HRP–streptavidin concentrate, the vial was spun and diluted 400 times with 1× diluent buffer. A 50 ng/ml stock solution was used to make the standard curve: 2000 pg/ml, 666.7 pg/ml, 222.2 pg/ml, 74.07 pg/ml, 24.69 pg/ml, 8.23 pg/ml, and 2.74 pg/ml for SCF/KITLG. The human SCF/KITLG antibody-precoated ELISA wells were filled with 100 μl of either serially diluted standard protein or plasma samples. After 2.5 h of incubation with gentle shaking at room temperature, 100 μl of 1× SCF/KITLG biotinylated detection antibody was added to the wells. After 1-h incubation with shaking at room temperature, the solution was discarded and the wells were washed four times using 300 μl wash buffer solution. Final wash was aspirated and plates were inverted to remove any remaining buffer. Then, 100 μl of prepared HRP–streptavidin solution was added to each well and incubated for 45 min at room temperature with gentle shaking. The solution was discarded and washed four times as described previously. Then 100 μl of ELISA colorimetric TMB reagent was added to each well and incubated for 30 min at room temperature in the dark. After this, 50 μl of stop solution was added to each well. Immediately after color development, the OD values were measured at 450 nm using Cytation 5 Cell Imaging Multi-Mode Reader (Biotek) and SCF/KITLG concentrations were calculated from specific calibration curves prepared with known standard solutions. Diluent buffer served as blank and the OD of these wells was subtracted from the values.

Results

Identification of Circulating Factor Signatures for Future Breast Cancer Risk Assessment

Because we wanted to identify circulating factors that might indicate future breast cancer risk, we utilized plasma samples from a cohort of healthy controls (healthy) and individuals who were clinically healthy at the time of plasma collection but later had a diagnosis of breast cancer (susceptible). We analyzed plasma samples using whole metabolite profiling



and OLINK biomarker analysis for a panel of inflammation and cancer-related proteins. We used two different sample sets, one with 39 susceptible and 36 healthy and the other with 10 susceptible and 11 healthy individuals, which were collected at different times. In the first set, 22 out of 39 susceptible and 23 out of 36 healthy individuals were postmenopausal status and remaining ones were premenopausal. In the second dataset, 7 out of 10 susceptible and 8 out of 11 healthy individuals were postmenopausal status and the remaining ones were premenopausal. Average time to diagnosis was 3.7 years after sample donation (median is 4 years). Data from two datasets were pre-processed separately because they were acquired at different times and were expected to have a variation due to external factors. Plasma levels of 295 different molecules for the first dataset and 339 different molecules for the second dataset were detected for the individuals. Some

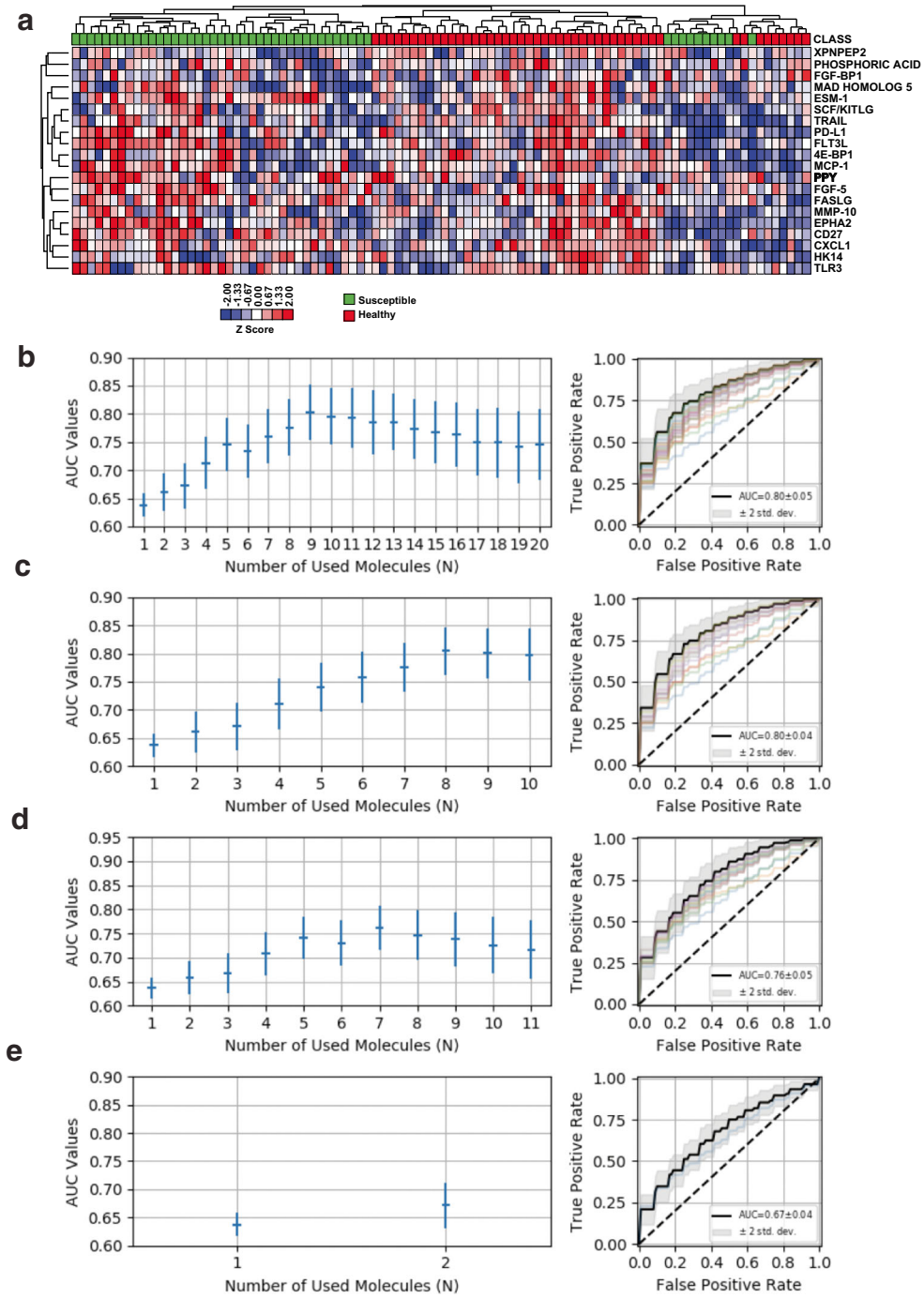
◀ **Fig. 1** Identification and performance assessment of circulating factor signatures for future breast cancer risk assessment using Student's *t* test as initial feature selection method. (**a**) Levels of top 20 molecules identified by Student's *t* test in 47 healthy (red) and 49 susceptible (green) individuals using OLINK analysis. Z-Scores were not log transformed or centered. Unsupervised hierarchical clustering was performed using Cluster 3 software for Z-scores of molecule concentrations with uncentered correlation as similarity metric and average linkage as clustering method. Data are visualized using Java Tree view software. In the lower panel, each column represents an individual and each row represents a molecule, with elevated levels in red, reduced levels in blue, and mean control levels in white. Bar indicates the coloring for Z-scores of molecule concentrations. (**b**) LR classification performances (AUC values) using the top-ranking N (1–20) molecules, ranked by their *p* values in Table 5, and the ROC curves of every AUC value where the bold black line indicates ROC curve of the best-performing (the highest AUC value) molecule set. (**c**) LR classification performances (AUC values) using the top-ranking N (1–13) molecules selected manually by considering the LR classification performance given in (**b**) and the ROC curves of every AUC value where the bold black line indicates ROC curve of the best-performing (the highest AUC value) molecule set. (**d**) LR classification performances (AUC values) using the top-ranking molecules selected from the list of 20 molecules, ranked by Student's *t* test in Table 5, by iterative elimination using pairwise Pearson correlation coefficients of features in Table 3 ($|pCC| = 0.5$ is the significance threshold). The ROC curves of every AUC value where the bold black line indicates ROC curve of the best-performing (the highest AUC value) molecule set. (**e**) LR classification performances (AUC values) using the top-ranking N (1–2) molecules selected from the list of 20 molecules, ranked by Student's *t* test in Table 5, by iterative elimination using paired *t*-test *p* values of features in Table 1 ($p = 0.05$ is the significance threshold). The ROC curves of every AUC value where the bold black line indicates ROC curve of the best-performing (the highest AUC value) molecule set

molecules had missing values (were not detected by metabolomics or OLINK approach) for some individuals, and further, some molecules were not measured for both datasets. All these molecules were excluded from the analysis. Therefore, we analyzed 181 different molecules, consisting of metabolites and proteins, which have plasma level values for every subject in both datasets.

In order to generate an inclusive list of features that would best discriminate between healthy and susceptible individuals, we took a stepwise approach where we first screened all molecules that contribute to increased classifier performance (LR) and then iteratively eliminate the redundant ones for both top-ranking molecule lists obtained by either of the initial molecule selection methods. We initially selected two different groups of 20 molecules (out of 181 molecules) using Student's *t* test and random forest (600 trees) methods to rank all 181 molecules with respect to their (healthy versus susceptible) discriminative power. Top-ranking 20 molecule sets obtained by two different feature selection methods, Student's *t* test and random forest, contain 10 common molecules, highly concentrated in the upper halves of the lists. For example, four out of five top-ranking molecules are common in both datasets. To assess the pairwise dependencies among the most discriminative 20 molecules and further reduce the number of features in our lists, we used the paired *t* test (Table 1—Student's *t* test, Table 2—random forest) or pairwise correlation analysis (Table 3—Student's *t* test, Table 4—random forest). To ensure that all molecules that might positively contribute to classifier performance are included in the signature, we performed logistic regression. Finally, to eliminate redundant molecules, we utilized paired *t* test *p* values ($p > 0.05$) and/or correlation coefficients ($pCC > 0.5$) to discard one of the molecules in that pair. Our approach resulted in six molecule signatures (Table 5—Student's *t* test, Table 6—random forest).

Assessment of Classification Performances of Molecule Signatures Using Machine-Learning Approach

In order to test the classification performance of each molecule signature, we performed LR classification using the molecules indicated in Tables 5 and 6. Of note, use of LR for performance assessment of classification at the last step is distinct from the earlier use of LR for manual selection of the molecules. Our top 20 feature list generated by Student's *t* test contained MMP-10, MCP-3, SCF/KITLG, TRAIL, ENRAGE, MAD HOMOLOG 5 (SMAD5), CXL17, HK11, FGF-BP1, XPNPEP2, C15:0 (pentadecanoic acid), PPY, FGF-5, FGF-21, ESM-1, FASLG, CD160, TNFB, CTSV, and ADA (Fig. 1a). Unsupervised clustering of the data using this list of molecules separated healthy and susceptible individuals; only two healthy individuals were classified with susceptible individuals and only one individual was classified together with healthy individuals (Fig. 1a). This list without any further feature elimination achieved AUC value of 0.83 (Fig. 1b). Reduction of feature number to 13 using manual selection increased AUC value to 0.85 ± 0.04 (Fig. 1c). Further reduction of feature using correlation analysis (Fig. 1d; $AUC = 0.78 \pm 0.04$) or paired *t* test (Fig. 1e; $AUC = 0.69 \pm 0.03$). On the other hand, AUC values achieved by molecule signatures using random forest had lower performance (Fig. 2). This list contained XPNPEP2, phosphoric acid, FGF-BP1, MAD HOMOLOG 5, ESM-1, SCF/KITLG, TRAIL, PD-L1, FLT3L, 4E-BP1, MCP-1, PPY, FGF-5, FASLG, MMP-10, EPHA2, CD27, CXCL1, HK14, and TLR3 (Fig. 2a). Unsupervised clustering of the data using this list of molecules was less successful in separating healthy and susceptible individuals; 10 susceptible individuals were classified with healthy individuals (Fig. 2a). Using all 20 factors achieved AUC of 0.80 ± 0.05 (Fig. 2b). Reducing the molecule number to 10 using manual selection (Fig. 2c, $AUC = 0.80 \pm 0.04$), to 11 using correlation analysis (Fig. 2d, $AUC = 0.76 \pm 0.05$), or to 2 using paired *t* test (Fig. 2e, $AUC = 0.67 \pm 0.04$) did not improve the AUC values. To sum up, initial feature selection using Student's *t* test followed by manual selection using LR gave us the best performing list of 13



circulating molecules from plasma for differentiating between healthy and susceptible individuals.

Biological Relevance of Biomarkers

Our best-performing list contained SCF/KITLG, MMP-10, MAD HOMOLOG5, CXL17, MCP-3, FGF05, FASLG, CD160, TNFB, ESM-1, FGF-21, XPNPEP2, and CTSV (Fig. 3a). In order to increase our understanding of molecules in the best-performing molecule list. Unsupervised clustering of the data using this list of molecules separated healthy and susceptible individuals accurately (Fig. 3a). To delve further into direction of change in the plasma levels of identified molecules, we compared the level of individual molecules in healthy versus susceptible individuals. Six of the 13 molecules, including SCF/KITLG, MAD HOMOLOG 5, FASLG, MMP-10, XPNPEP2, and CXL17, were statistically significantly different between the two groups (Fig. 3b). Since our aim was to identify the molecules that have marginal but significant contribution to the classification task when used

◀ **Fig. 2** Identification and performance assessment of circulating factor signatures for future breast cancer risk assessment using random forest as initial feature selection method. (**a**) Levels of top 20 molecules identified by random forest method in 47 healthy (red) and 49 susceptible (green) individuals using OLINK analysis. Z-Scores were not log transformed or centered. Unsupervised hierarchical clustering was performed using Cluster 3 software for Z-scores of molecule concentrations with uncentered correlation as similarity metric and average linkage as clustering method. Data are visualized using Java Tree view software. In the lower panel, each column represents an individual and each row represents a molecule, with elevated levels in red, reduced levels in blue, and mean control levels in white. Bar indicates the coloring for Z-scores of molecule concentrations. (**b**) LR classification performances (AUC values) using the top-ranking N (1–20) molecules, ranked by their feature importance values (computed by random forest) in Table 6, and the ROC curves of every AUC value where the bold black line indicates ROC curve of the best-performing (the highest AUC value) molecule set. (**c**) LR classification performances (AUC values) using the top-ranking molecules selected manually by considering the LR classification performance given in (**b**) and the ROC curves of every AUC value where the bold black line indicates ROC curve of the best-performing (the highest AUC value) molecule set. (**d**) LR classification performances (AUC values) using the top-ranking N (1–11) molecules selected from the list of 20 molecules, ranked by random forest in Table 6, by iterative elimination using pairwise Pearson correlation coefficients of features in Table 4 ($|pCC| = 0.5$ is the significance threshold). The ROC curves of every AUC value where the bold black line indicates ROC curve of the best-performing (the highest AUC value) molecule set. (**e**) LR classification performances (AUC values) using the top-ranking N (1–2) molecules selected from the list of 20 molecules, ranked by random forest in Table 6, by iterative elimination using paired t-test p values of features in Table 2 ($p = 0.05$ is the significance threshold). The ROC curves of every AUC value where the bold black line indicates ROC curve of the best-performing (the highest AUC value) molecule set

together with other molecules, even if they have weak discriminative power on their own, we still included these molecules with poor t-test performance individually, p value > 0.05, or low random forest importance in the final lists. We were particularly interested in SCF/KITLG as this molecule was the top molecule identified in both feature selection methods (Tables 5 and 6). Overall, SCF/KITLG levels were lower in individuals with increased breast cancer risk (Fig. 3c). We also validated our finding from OLINK analysis using another independent method, ELISA analysis, and verified that the level of this protein is lower in susceptible individuals (Fig. 3b).

Discussion

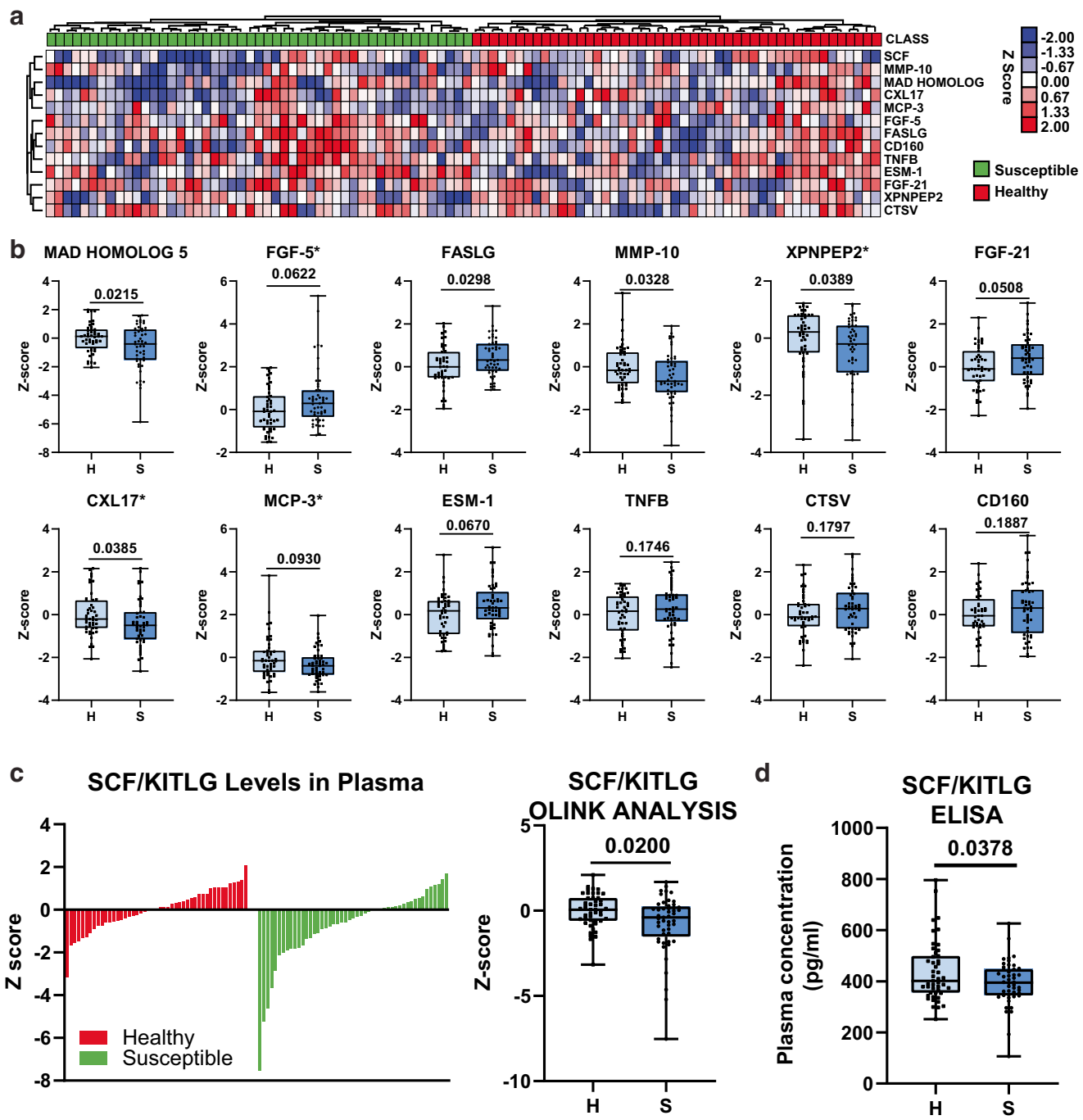
In this study, we developed a pipeline to identify plasma biomarkers of breast cancer risk using a combination of classical statistics methods and machine-learning approaches, and independently validated one of the identified biomarkers, SCF/KITLG. By iterative feature selection, elimination, and performance testing, we generated a molecular signature of plasma biomarkers that can discriminate between healthy and breast cancer–susceptible individuals. Because of our approach, some of the molecules in this signature had weak discriminative power on their own, yet they contributed significantly to the discriminative power of the signature.

A biomarker is a biomolecule such as DNA, RNA, proteins, hormones, and chemical modifications that can be measured to describe that an abnormal or a normal process is taking place within the organism [12]. A cancer biomarker can arise due to changes in the DNA (mutations), rearrangements, deletions (missing copies), or amplifications. Biomarkers might affect various hallmarks of cancer including cell cycle, cell death, or immunological properties of the tumor and indicate the risk of developing cancer, its progression, and response to therapy [8, 9, 12].

Previously, Kazarian et al. studied pre-diagnostic samples from the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Serum samples were taken from 239 women who were diagnosed with invasive ductal carcinoma in breast, months to years after sample donation [17]. These patients were post-menopausal women with ages ranging from 50 to 74, who were healthy cases at the moment of recruitment but later developed breast cancer. Hence, this group studied the ability of several serum markers to detect breast cancer cases before these patients were diagnosed. They studied CA 15-3, RANTES/CCL5, OPN, PAI-1, SLP1, HSP90A, IGFBP3, APOC1, and PAPPA. They concluded that only three out of the nine serum markers (CA 15-3, PAI1, and HSP90A) were potential prognostic biomarkers [17]. Those analyses were performed using a limited panel of proteins. However, in our analysis, we characterized more than 300 proteins and metabolites in plasma and used a final list of 181 molecules to generate our signatures.

One potential marker of interest we identified is SCF/KITLG protein. KITLG protein is expressed in 53% of breast cancer cell lines [18]. SCF/KITLG was shown to have a proliferative role in BCK4 cells, and when it is reduced, it decreased estrogen-induced proliferation [19]. We identified a lower level of this biomarker in plasma from women with breast cancer risk. Since at the time of the blood draw the women did not have tumors, it is not possible for us to infer the level of this protein in the tumor. Whether SCF plays a role in the induction of breast tumors or lower plasma levels of this protein contributes to the tumor biology needs to be determined.

Several of the molecules in our signature were also implicated in cancer biology. For example, MAD HOMOLOG5/SMAD5 plays a role in breast cancer cell stemness and resistance to chemotherapy [20, 21]. FGF-5, FASLG, CTSV, and ESM-1 expression is associated with lower survival and worst outcomes [22–27]. MMP-10 affects angiogenesis and apoptosis [28, 29]; XPNPEP2 is overexpressed in cervical cancer patients and increases motility and invasiveness of tumors [30]. FGF-21, TNFB contributes to metastatic potential of



breast cancer cells [31, 32]. CXL17 [33], MCP-3 [34], and CD160 [35] play a role in recruitment of immune cells. TNFB/LTA polymorphisms increased the cancer risk in various populations [36, 37]. All these studies focused on the tumors or patients that already have cancers. The impact of proteins in our signature on breast cancer risk and initiation remains to be established. Direction of differences in the plasma levels of these proteins between healthy and susceptible individuals might be different from what is reported in already established tumors and might indicate a different role for these proteins at early stages of tumor development.

More recently, liquid biopsy methods supported with machine-learning approaches have been used for the detection of different cancer types [15, 38, 39]. For example, Cohen et al. recently demonstrated the capability of detecting eight different cancer types including breast cancer using circulating tumor DNA (ctDNA) and protein biomarkers [40]. They reported remarkable sensitivity values > 95% for ovarian and liver cancers. However, the reported sensitivity for breast cancer is rather low at 33%. The novelty of our study is identifying circulating molecules that are associated with future cancer risk and developing a pipeline to utilize these markers in

◀ **Fig. 3** Validation of biomarker identification. (**a**) Levels of 13 molecules identified by Student's *t* test followed by manual selection in 47 healthy (red) and 49 susceptible (green) individuals using OLINK analysis. Z-Scores were not log transformed or centered. Unsupervised hierarchical clustering was performed using Cluster 3 software for Z-scores of molecule concentrations with uncentered correlation as similarity metric and average linkage as clustering method. Data are visualized using Java Tree view software. In the lower panel, each column represents an individual and each row represents a molecule, with elevated levels in red, reduced levels in blue, and mean control levels in white. Bar indicates the coloring for Z-scores of molecule concentrations. (**b**) Changes in the levels of 12 of 13 signature molecules in 47 healthy and 49 susceptible individuals. Anderson–Darling and Kolmogorov–Smirnov tests for normality were used. If the dataset did not pass the normality test, non-parametric Mann–Whitney *U* test was used to assess if level of a molecule is statistically significantly different in plasma from healthy versus susceptible individuals (molecules with *). Otherwise, unpaired *t* test was used to assess if level of a molecule is statistically significantly different in plasma from healthy versus susceptible individuals. All data points are plotted. *P* values are indicted on the graphs. (**c**) Level of SCF/KITLG in 47 healthy and 49 susceptible individuals. Anderson–Darling and Kolmogorov–Smirnov tests for normality were used. Non-parametric Mann–Whitney *U* test was used to assess if level of a molecule is statistically significantly different in plasma from healthy versus susceptible individuals. All data points are plotted (as histogram on the left side and as box–whiskers graph on right side). *P* values are indicted on the graph. (**d**) Results from (**c**) are independently validated using ELISA using all the samples. Level of identified biomarkers in human plasma samples were compared using unpaired *t* test. *P* value is reported on the graph. Values from all the samples are presented

generation of biosensors based on our previous work to detect breast cancer risk [41].

We used a combination of various statistical analysis methods to identify biomarkers. Although Student's *t* test and/or random forest gives some information about the ability of a biomarker to discriminate between healthy and susceptible patients, it alone is not sufficient. To identify the biomarkers with high classification performance, we applied logistic regression. Area under curve (AUC) of receiver operating characteristic (ROC) curves resulting from the classification operations on these biomarkers is commonly used as an indicator for the discriminative capacity of a single molecule or a set of molecules. Previously, logistic regression was performed on predictors consisting of serum levels of several molecules, but authors did not report any confidence interval for that AUC value and did not split the data into training and test sets [42]. In another study, authors used Student' *t* test and its non-parametric equivalence (Mann–Whitney *U* test) to find potential biomarkers, but the lower bound of their reported confidence intervals was dramatically low, suggesting that those biomarkers were not robust, and they also did not split the training and test sets [43]. Several other studies have also used these methods to identify potential biomarkers but have not utilized a training–set split for their datasets [17, 44–46]. Training (model building) and testing on the same dataset is not an ideal practice in machine learning as the model is likely to over-fit to the data. This approach results in artificially high predictive rates, in other words, *low generalizability*, which refers to poor applicability of the model to unseen data. The cross-validation that we employed in this study is a common approach to circumvent the problem of overfitting.

Conclusion

We identified biomarkers of breast cancer risk using metabolomics and protein profiling in plasma samples from healthy and susceptible individuals. Future studies are required to validate these markers in bigger data sets, to determine their role in breast tumorigenesis, develop liquid biopsy/biosensor-based approaches, and move this information to clinic for early identification of breast cancer risk. In addition, further molecular studies in cell lines and animal models are required to show conclusively whether or not each or a combination of these markers can be utilized as indicators of breast cancer risk without having observable effects on breast cancer cells or can have other roles at the earlier stages of carcinogenesis. Overall, our analysis offers novel plasma biomarkers for further validation and functional characterization.

Author Contributions K.O., B.A., H.T., and Z.M.E. designed the study. K.O. performed the biomarker identification analysis. M.P., N.M., and A.M.V.S. provided the samples. A.S.-C. prepared the samples for analysis and performed the experiments. K.O., B.A., H.T., A.S.-C., and Z.M.E. wrote the manuscript. All authors have read and approved the manuscript.

Funding This work was supported by grants from the University of Illinois, ACES Future interdisciplinary research explorations grant (Z.M.E., data collection), Office of International Programs—Conrad Award (Z.M.E., data collection), National Institute of Food and Agriculture, U.S. Department of Agriculture, award ILLU-698-909 (to Z.M.E., data collection and analysis), UIUC Graduate College ASPIRE fellowship (A.S.-C.), Boğaziçi University research funds grant no. 12360 (to B.A. and H.T., data analysis) and TÜBİTAK 2210-A National Scholarship Programme for MSc Students (K.O.).

Data Availability All the data will be available from Komen Tissue Bank website upon acceptance of manuscript.

Compliance with Ethics Standards KTB studies were approved by the Indiana University Institutional Review Board (IRB protocol nos. 1011003097 and 1607623663). All research was carried out in compliance with the Helsinki Declaration. Donors provided broad written consent for the use of their specimens in research. The consent document informed the donor that the donated specimens and medical data would be used for the general purpose of helping to determine how breast cancer develops and the exact laboratory experiments were unknown at the time of donation, and that proposals for use of the specimens would be reviewed and approved by a panel of independent researchers before specimens and/or data were released for research purposes.

Conflict of Interest Z.M.E. has investigator-initiated grant from Karyopharm Therapeutics and is a co-inventor on several patents entitled "Novel Compounds which Activate Estrogen Receptors and Compositions and Methods of Using the Same." Z.M.E. was a PI on an investigator-initiated grant from Corteva Agrisciences and Pfizer Inc.

References

1. Siegel RL, Miller KD, Jemal A (2017) Cancer statistics, 2017. CA Cancer J Clin 67(1):7–30
2. Anderson BO, Yip C-H, Smith RA, Shyyan R, Sener SF, Eniu A, Carlson RW, Azavedo E, Harford J (2008) Guideline implementation for breast healthcare in low-income and middle-income countries. Cancer 113(8):2221–2243
3. Coleman MP, Quaresma M, Berrino F, Lutz J-M, De Angelis R, Capocaccia R, Baili P, Rachet B, Gatta G, Hakulinen T et al (2008) Cancer survival in five continents: a worldwide population-based study (CONCORD). The Lancet Oncology 9(8):730–756
4. Li J, Shao Z (2015) Mammography screening in less developed countries. SpringerPlus 4:615
5. da Costa Vieira RA, Biller G, Uemura G, Ruiz CA, Curado MP (2017) Breast cancer screening in developing countries. Clinics 72(4):244–253
6. Dawson SJ, Duffy SW, Blows FM, Driver KE, Provenzano E, LeQuesne J, Greenberg DC, Pharoah P, Caldas C, Wishart GC (2009) Molecular characteristics of screen-detected vs symptomatic breast cancers and their impact on survival. Br J Cancer 101(8): 1338–1344
7. Iqbal J, Ginsburg O, Rochon PA, Sun P, Narod SA (2015) Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. JAMA 313(2): 165–173
8. Hammond MEH, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, Fitzgibbons PL, Francis G, Goldstein NS, Hayes M et al (2010) American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Clin Oncol 28(16):2784–2795
9. Andre F, Ismaila N, Stearns V (2019) Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: ASCO clinical practice guideline update summary. Journal of Oncology Practice 15(9):495–497
10. Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, Bilous M, Ellis IO, Fitzgibbons P, Hanna W et al (2018) Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. J Clin Oncol 36(20):2105–2122
11. Van Poznak C, Somerfield MR, Bast RC, Cristofanilli M, Goetz MP, Gonzalez-Angulo AM, Hicks DG, Hill EG, Liu MC, Lucas W et al (2015) Use of biomarkers to guide decisions on systemic therapy for women with metastatic breast cancer: American Society of Clinical Oncology clinical practice guideline. J Clin Oncol 33(24): 2695–2704
12. Goossens N, Nakagawa S, Sun X, Hoshida Y (2015) Cancer biomarker discovery and validation. Transl Cancer Res 4(3):256–269
13. O'Leary B, Hrebien S, Morden JP, Beaney M, Fribbens C, Huang X, Liu Y, Bartlett CH, Koehler M, Cristofanilli M et al (2018) Early circulating tumor DNA dynamics and clonal selection with palbociclib and fulvestrant for breast cancer. Nat Commun 9(1):896
14. Coombes RC, Page K, Salari R, Hastings RK, Armstrong AC, Ahmed S, Ali S, Cleator SJ, Kenny LM, Stebbing J *et al*: Personalized detection of circulating tumor DNA antedates breast cancer metastatic recurrence. *Clinical Cancer Research* 2019: clincanres.3663.2018
15. Merker JD, Oxnard GR, Compton C, Diehn M, Hurley P, Lazar AJ, Lindeman N, Lockwood CM, Rai AJ, Schilsky RL, Tsimberidou AM, Vasalos P, Billman BL, Oliver TK, Bruinooge SS, Hayes DF, Turner NC (2018) Circulating tumor DNA analysis in patients with cancer: American Society of Clinical Oncology and College of American Pathologists Joint Review. J Clin Oncol 36(16):1631–1641
16. Madak-Erdogan Z, Band S, Zhao YC, Smith BP, Kulkoyluoglu-Cotul E, Zuo Q, Santaliz Casiano A, Wrobel K, Rossi G, Smith RL et al (2019) Free fatty acids rewire cancer metabolism in obesity-associated breast cancer via estrogen receptor and mTOR signaling. Cancer Res
17. Kazarian A, Blyuss O, Metodieva G, Gentry-Maharaj A, Ryan A, Kiseleva EM, Prytomanova OM, Jacobs IJ, Widschwendter M, Menon U, Timms JF (2017) Testing breast cancer serum biomarkers for early detection and prognosis in pre-diagnosis samples. Br J Cancer 116(4):501–508
18. Hines S, Organ C, Kornstein M, Krystal G (1995) Coexpression of the c-kit and stem cell factor genes in breast carcinomas. Cell Growth Differ 6(6):769–779
19. Harrell JC, Shroka TM, Jacobsen BM (2017) Estrogen induces c-kit and an aggressive phenotype in a model of invasive lobular breast cancer. Oncogenesis 6(11):396
20. Opyrchal M, Gil M, Salisbury JL, Goetz MP, Suman V, Degnim A, McCubrey J, Haddad T, Iankov I, Kurokawa CB et al (2017) Molecular targeting of the Aurora-a/SMAD5 oncogenic axis restores chemosensitivity in human breast cancer cells. Oncotarget 8(53):91803–91816
21. Opyrchal M, Iankov I, Ingle JN, Salisbury JL, Galanis E, D'Assoro A (2013) SMAD5 expression and inhibition of the mitotic kinase aurora-A on sensitivity of breast cancer cells to chemotherapy. *Journal of Clinical Oncology* **31**(15_suppl):e13516–e13516
22. Ghassemi S, Vejdovszky K, Sahin E, Ratzinger L, Schelch K, Mohr T, Peter-Vörösmarty B, Brankovic J, Lackner A, Leopoldi A, Meindl D, Pirker C, Hegedus B, Marian B, Holzmann K, Grasl-Kraupp B, Heffeter P, Berger W, Grusch M (2017) FGF5 is expressed in melanoma and enhances malignancy in vitro and in vivo. Oncotarget 8(50):87750–87762
23. Huang Y, Wang H, Yang Y (2018) Expression of fibroblast growth factor 5 (FGF5) and its influence on survival of breast cancer patients. Med Sci Monit 24:3524–3530
24. Ates O, Gedik E, Sunar V, Altundag K (2018) Serum endocan level and its prognostic significance in breast cancer patients. Journal of Oncological Sciences 4(1):15–18
25. Bębenek M, Duś D, Koźlak J (2013) Prognostic value of the Fas/Fas ligand system in breast cancer. Contemp Oncol 17(2):120–122
26. Mor G, Kohen F, Garcia-Velasco J, Nilsen J, Brown W, Song J, Naftolin F (2000) Regulation of Fas ligand expression in breast cancer cells by estrogen: functional differences between estradiol and tamoxifen. J Steroid Biochem Mol Biol 73(5):185–194
27. Toss M, Miligy I, Gorringe K, Mittal K, Aneja R, Ellis I, Green A, Rakha E: Prognostic significance of cathepsin V (CTSV/CTSL2) in breast ductal carcinoma in situ. *Journal of Clinical Pathology* 2019: jclinpath-2019-205939
28. Benson CS, Babu SD, Radhakrishna S, Selvamurugan N, Sankar BR (2013) Expression of matrix metalloproteinases in human breast cancer tissues. Dis Markers 34(6):395–405
29. Zhang G, Miyake M, Lawton A, Goodison S, Rosser CJ (2014) Matrix metalloproteinase-10 promotes tumor progression through regulation of angiogenic and apoptotic pathways in cervical tumors. BMC Cancer 14(1):310
30. Cheng T, Wei R, Jiang G, Zhou Y, Lv M, Dai Y, Yuan Y, Luo D, Ma D, Li F et al (2017) XPNPEP2 is overexpressed in cervical cancer and promotes cervical cancer metastasis. Tumor Biol 39(7): 1010428317717122
31. Knott ME, Ranuncolo SM, Nuñez M, Armanasco E, Puricelli LI, De Lorenzo MS (2015) Abstract 1577: Levels of fibroblast growth factor 21 (FGF21) in serum as diagnostic biomarker in patients with breast cancer. Cancer Res 75(15 Supplement):1577–1577
32. Aukes K, Forsman C, Brady NJ, Astleford K, Blixt N, Sachdev D, Jensen ED, Mansky KC, Schwertfeger KL (2017) Breast cancer

cell-derived fibroblast growth factors enhance osteoclast activity and contribute to the formation of metastatic lesions. PLoS One 12(10):e0185736
33. Matsui A, Yokoo H, Negishi Y, Endo-Takahashi Y, Chun NAL, Kadouchi I, Suzuki R, Maruyama K, Aramaki Y, Semba K, Kobayashi E, Takahashi M, Murakami T (2012) CXCL17 expression by tumor cells recruits CD11b+Gr1 high F4/80− cells and promotes tumor progression. PLoS One 7(8):e44080–e44080
34. Ben-Baruch A, Xu L, Young PR, Bengali K, Oppenheim JJ, Wang JM (1995) Monocyte chemotactic protein-3 (MCP3) interacts with multiple leukocyte receptors: C-C CKR1, a receptor for macrophage inflammatory protein-1α/RANTES, is also a functional receptor for MCP3. J Biol Chem 270(38):22123–22128
35. Sun H, Xu J, Huang Q, Huang M, Li K, Qu K, Wen H, Lin R, Zheng M, Wei H, Xiao W, Sun R, Tian Z, Sun C (2018) Reduced CD160 expression contributes to impaired NK-cell function and poor clinical outcomes in patients with HCC. Cancer Res 78(23): 6581–6593
36. Huang Y, Yu X, Wang L, Zhou S, Sun J, Feng N, Nie S, Wu J, Gao F, Fei B et al (2013) Four genetic polymorphisms of lymphotoxin-alpha gene and cancer risk: a systematic review and meta-analysis. PLoS One 8(12):e82519
37. Kohaar I, Tiwari P, Kumar R, Nasare V, Thakur N, Das C B, Bharadwaj M: Association of single nucleotide polymorphisms (SNPs) in TNF-LTA locus with breast cancer risk in Indian population, vol. 114; 2008
38. Alix-Panabières C, Pantel K (2013) Circulating tumor cells: liquid biopsy of cancer. Clin Chem 59(1):110–118
39. Heitzer E, Ulz P, Geigl JB (2015) Circulating tumor DNA as a liquid biopsy for cancer. Clin Chem 61(1):112–123
40. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, Douville C, Javed AA, Wong F, Mattox A et al (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science (New York, NY)* **359**(6378):926–930
41. Sarangapani K, Torun H, Finkler O, Zhu C, Degertekin L (2010) Membrane-based actuation for high-speed single molecule force spectroscopy studies using AFM. Eur Biophys J 39(8):1219–1227
42. Hwa H-L, Kuo W-H, Chang L-Y, Wang M-Y, Tung T-H, Chang K-J, Hsieh F-J (2008) Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models. J Eval Clin Pract 14(2):275–280
43. Santillán-Benítez JG, Mendieta-Zerón H, Gómez-Oliván LM, Torres-Juárez JJ, González-Bañales JM, Hernández-Peña LV, Ordóñez-Quiroz A (2013) The tetrad BMI, leptin, leptin/adiponectin (L/A) ratio and CA 15-3 are reliable biomarkers of breast cancer. J Clin Lab Anal 27(1):12–20
44. Dalamaga M, Sotiropoulos G, Karmaniolas K, Pelekanos N, Papadavid E, Lekka A (2013) Serum resistin: a biomarker of breast cancer in postmenopausal women? Association with clinicopathological characteristics, tumor markers, inflammatory and metabolic parameters. Clin Biochem 46(7):584–590
45. Assiri AMA, Kamel HFM, Hassanien MFR (2015) Resistin, visfatin, adiponectin, and leptin: risk of breast cancer in pre- and postmenopausal Saudi females and their possible diagnostic and predictive implications as novel biomarkers. Dis Markers 2015: 253519
46. Provatopoulou X, Georgiou GP, Kalogera E, Kalles V, Matiatou MA, Papapanagiotou I, Sagkriotis A, Zografos GC, Gounaris A (2015) Serum irisin levels are lower in patients with breast cancer: association with disease diagnosis and tumor characteristics. BMC Cancer 15:898