Breast Cancer Res Treat (2015) 153:445–454
DOI 10.1007/s10549-015-3554-5

Candidate early detection protein biomarkers for ER+/PR+ invasive ductal breast carcinoma identified using pre-clinical plasma from the WHI observational study

Matthew F. Buas[1] · Jung-hyun Rho[1] · Xiaoyu Chai[1] · Yuzheng Zhang[1] · Paul D. Lampe[1] · Christopher I. Li[1,2]

Received: 31 May 2015 / Accepted: 21 August 2015 / Published online: 30 August 2015

Abstract Estrogen receptor (ER)-positive/progesterone receptor (PR)-positive invasive ductal carcinoma accounts for ~45 % of invasive breast cancer (BC) diagnoses in the U.S. Despite reductions in BC mortality attributable to mammography screening and adjuvant hormonal therapy, an important challenge remains the development of clinically useful blood-based biomarkers for risk assessment and early detection. The objective of this study was to identify novel protein markers for ER+/PR+ ductal BC. A nested case–control study was conducted within the Women's Health Initiative observational study. Pre-clinical plasma specimens, collected up to 12.5 months before diagnosis from 121 cases and 121 matched controls, were equally divided into training and testing sets and interrogated using a customized antibody array targeting >2000 proteins. Statistically significant differences ($P < 0.05$) in matched case versus control signals were observed for 39 candidates in both training and testing sets, and four markers (CSF2, RYBP, TFRC, ITGB4) remained significant after Bonferroni correction ($P < 2.03 \times 10^{-5}$). A multivariate modeling procedure based on elastic net regression with Monte Carlo cross-validation achieved an estimated AUC of 0.75 (SD 0.06). Most candidates did not overlap with those described previously for triple-negative BC, suggesting sub-type specificity. Gene set enrichment analyses identified two GO gene sets as upregulated in cases—microtubule cytoskeleton and response to hormone stimulus ($P < 0.05$, $q < 0.25$). This study has identified a pool of novel candidate plasma protein biomarkers for ER+/PR+ ductal BC using pre-diagnostic biospecimens. Further validation studies are needed to confirm these candidates and assess their potential clinical utility for BC risk assessment/early detection.

Keywords ER+ breast cancer · Blood biomarkers · Proteomics · Early detection · Antibody array

Introduction

In recent decades, mammography has become a widely implemented screening tool for breast cancer (BC) among women aged 40 and older. Meta-analyses of randomized clinical trials indicate that routine mammographic screening reduces BC mortality by 30 % among women 50–69 years of age [1–4], although the magnitude of this mortality benefit remains controversial [5]. Under current practice, the effectiveness of population-based screening mammography is compromised by at least two important factors: the absence of rationally targeted screening, and mammography's limited sensitivity and specificity [6–8]. Deployment of this imperfect screening tool based almost exclusively on age and sex results in millions of healthy women being frequently imaged, many women without cancer needlessly sent for breast biopsies, and other women with cancer falsely declared cancer free. Novel clinical methods are needed to enhance the selection of women

Electronic supplementary material The online version of this article (doi:10.1007/s10549-015-3554-5) contains supplementary material, which is available to authorized users.

✉ Matthew F. Buas
mbuas@fhcrc.org

✉ Christopher I. Li
cili@fredhutch.org

[1] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

[2] Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA 98109, USA

who initially undergo screening, and to improve the accuracy of imaging among those who do [9–11]. Blood-based biomarkers with high sensitivity and specificity could be used in conjunction with radiographic approaches to increase early detection of BC while reducing false positives and unnecessary clinical follow-up [12].

Many past studies have sought to identify circulating biomarkers for BC, and multiple components of blood have been mined for diagnostic potential—autoantibodies, proteins, miRNAs, cell-free DNA, and circulating tumor cells [13–20]. Despite advances in molecular classification of BC sub-types and in our understanding of disease pathogenesis, no clinically useful blood-based biomarkers for early diagnosis have yet been identified. Recent studies based on mass spectrometry analysis of pre-clinical blood specimens have suggested that plasma proteome alterations may be detectable prior to BC diagnosis [19].

Breast cancer represents a heterogeneous disease, with multiple sub-types defined by distinct histological and molecular characteristics, epidemiologic risk profiles, and clinical features [21–24]. Infiltrating ductal carcinoma (70–80 % of invasive lesions) and infiltrating lobular carcinoma (5–10 %) comprise the two most common histological types of invasive BC. In contrast to ductal BC, the rarer lobular form is associated with older age, larger and better differentiated dispersed tumors, ER positivity, absence of E-cadherin, and increased early (≤10 years) survival [25]. Defined by joint ER/PR status and histology, the most common sub-type in the U.S. is ER+/PR+ invasive ductal carcinoma (IDC). While 5-year survival rates for ER+/PR+ IDC are higher than those observed for the less-common but more lethal triple-negative (ER−/PR−/HER2−) BC, ER+/PR+ cancer remains a significant cause of morbidity and mortality among women overall, particularly when diagnosed at late stages [26]. Given the molecular heterogeneity of BC, it seems plausible that plasma proteomic profiles associated with disease onset, if they exist, may exhibit sub-type specific patterns [12]. We employed a custom-designed antibody-array platform [27] to interrogate pre-diagnostic plasma from 121 BC cases (all ER+/PR+ IDC), and 121 matched controls enrolled in the Women's Health Initiative (WHI) observational study, with the aim of identifying protein biomarkers that are differentially abundant in the plasma of cases relative to controls.

Materials and methods

Study design

We conducted a nested case–control study of ER+/PR+ IDC within the WHI observational study, a prospective cohort of 93,676 post-menopausal women enrolled from 1993–1998 in the U.S. [28, 29]. Our study included 121 ER+/PR+ ductal BC cases and 121 controls without a prior history of any type of cancer, individually matched 1:1 to cases on age at enrollment (±3 years), race/ethnicity (white, non-white), body mass index (BMI, ≤25, 25–29, or 30+), hormone replacement therapy use (estrogen only, estrogen plus progestin, or non-user), and year of blood draw. Cases with an available study blood specimen drawn within 12.5 months prior to their BC diagnosis were eligible for inclusion. ER, PR, and HER2 status were extracted from medical records and centrally adjudicated by WHI staff. The 121 matched pairs were randomly assigned to a training set ($n = 60$ pairs) and a testing set ($n = 61$ pairs). The use of human samples was approved by the Institutional Review Board of the Fred Hutchinson Cancer Research Center.

Laboratory methods

Plasma samples were interrogated using a customized antibody array populated with 3290 full-length antibodies to 2036 distinct targets, encompassing several classes of proteins involved in diverse signaling pathways. Of the 3290 antibodies, 292 were included as standards targeting CA19-9 ($n = 96$), CEACAM1 ($n = 48$), CSF3 ($n = 48$), CXCL12 ($n = 48$), or IL1β ($n = 48$), while four antibodies on the array served as technical controls targeting Cy3 ($n = 2$) or Cy3/Cy5 ($n = 2$). These standards and controls included replicates of two independent antibodies for CA19-9 and replicates of a single antibody for the other targets. Protocols for array fabrication, sample treatment and labeling, incubation of plasma with arrays, and array scanning have been previously described [30–32]. Briefly, for this study, 200 μg case and control plasma protein were labeled with Cy5 and separately incubated for 90 min with Cy3-labeled 200 μg reference plasma protein (a common pool of plasma composed of samples collected from seven women aged 45–72 years was used as a reference for all samples) in 80 μl total volume (kept from drying using LifterSlips, Fisher Scientific, Pittsburgh, PA). Samples from a given matched pair (case and control) were incubated on the same day. After washing, slides were scanned in a GenePix 4000B microarray scanner and data extracted using GenePix Pro 6.0 software (Molecular Devices, Sunnyvale, CA, USA).

Statistical analyses

Genepix results that raw data files were initially processed using the R package *limma* [33, 34]. For each antibody, fold change of signal (red channel) compared to reference (green channel), the M value, was calculated as $\log_2(R_c/$

G_c), where R_c is red corrected, and G_c is green corrected (using the normexp background correction method). After exclusion of technical control spots ($n = 4$ antibodies) and all spots with GenePix flags ≤ -50, antibody features on each array were filtered based on the coefficient of variation (CV) for observed M values across triplicate spots. Technical sources of variation were normalized by loess procedures developed for microarrays, including within-array print-tip loess and between-arrays reference channel quartile normalization. Following normalization, triplicate spots were summarized using their median. Antibody features with missing M values in > 30 % of the 121 matched pairs (after CV filtering) were excluded from the analysis ($n = 819$ of 3286). All statistical analyses were performed using M values. Intra-array reproducibility was evaluated by calculating the CV across triplicate spots of each antibody feature (Figure S1). Inter-array reproducibility was examined by calculating the Pearson pair-wise correlation coefficient (ρ) between arrays probed with an identical plasma sample (Figure S2).

Paired t tests were conducted for each antibody feature using the 121 BC cases and 121 matched controls. Antibodies for which the mean difference in M values between paired cases and controls differed significantly from zero (at $P < 0.05$) in both the training and testing sets were selected as candidate markers. After exclusion of the four control antibodies, Bonferroni correction for multiple comparisons was conducted ($\alpha = 0.05/2467 = 2.03 \times 10^{-5}$) using the pooled dataset (training and testing), and Benjamini–Hochberg false discovery rate q values were generated. Performance metrics (AUC/sensitivity/specificity) were calculated using covariate-adjusted marker values.

Multivariate classification models were constructed using regularized logistic regression with elastic net penalty (R package: glmnet) [35]. After exclusion of $n = 7$ subjects due to missing covariate values, 235 subjects were available for analysis. Missing M values for the 2467 antibody features included in the main analysis were imputed via the k-nearest neighbor method (R package: knn). Equal numbers of cases and controls were randomly allocated to a training set (75 %: $n = 175$), used for variable selection and model selection, or a testing set (25 %: $n = 60$), used for (preliminary) model validation. With the mixing parameter (α) set to 0.5, fivefold cross-validation was conducted within the training set only to select the optimal value of the penalty parameter (λ). A model was generated using the complete training set data and used to predict class values for subjects in the testing set. Monte Carlo cross-validation (MCCV) was conducted using 100 different training (and associated testing) sets randomly selected from the total study sample. Mean estimated area under the receiver-operating characteristic (ROC) curve (AUC) was calculated across the 100 testing sets. A composite average ROC curve was generated (R package: ROCR) [36].

Gene set enrichment analyses (GSEA) were based on the Kyoto encyclopedia of genes and genomes (KEGG) and gene ontology (GO) gene sets, available from the Molecular signatures database (MSigDB). After exclusion of antibodies specific for carbohydrate or non-human protein targets, 2396 of 2467 antibodies included in the main analysis were available for GSEA (corresponding to proteins encoded by 1693 unique genes). Of the 186 KEGG gene sets obtained from MSigDB, 128 sets contained at least five genes coding for proteins targeted by our arrays (Table S1A). Of the 1454 available GO gene sets, 822 sets contained at least five genes coding for proteins targeted by our arrays (Table S1B). Wilcoxon rank-sum tests were conducted on the test statistics obtained from paired t tests (comparing M values of matched cases and controls). A null distribution for the Wilcoxon test was generated from analysis of 1000 permuted datasets in which the case/control status of subjects in each matched pair was randomly maintained or reversed. The (one-sided) P value for each pathway was calculated as the fraction of all permuted datasets in which the sum of the ranks of the paired t test statistics for markers in the gene set was more extreme than the observed sum of ranks for these markers' test statistics in the real dataset. All statistical analyses were conducted using Stata v13.1 (College Station, TX) or R v3.03.

Using data from The Cancer Genome Atlas (TCGA), we identified post-menopausal female subjects with Stage I/II ER+/PR+ infiltrating ductal breast carcinoma. Those with available tumor RNA expression data were selected ($n = 130$). Paired normal tissue had been profiled for $n = 10$ of these women. RNA expression data was also included from an additional 51 normal tissue specimens isolated from BC cases not satisfying our inclusion criteria. Technical details pertaining to RNA profiling have been published previously [37, 38]. Microarray data were loess normalized, and the ratio of the Cy5 channel (sample) and Cy3 channel (reference) was $\log_2$-transformed to generate gene expression values ($E = \log_2$ Cy5/Cy3). Welch's t tests were conducted to compare mean RNA expression levels in the $n = 130$ tumor samples and $n = 61$ normal samples. Marker fold changes were calculated as the ratio of geometric mean signals in tumor samples versus normals: $2^\wedge\mu_{E,\text{tumors}}/2^\wedge\mu_{E,\text{normals}}$.

Results

Cases and controls in our training and testing sets were well-balanced with respect to matching covariates: age, race/ethnicity, BMI, and HRT use (Table 1). None of the

Table 1 Subject characteristics of 121 matched case–control pairs randomly assigned to either a training set or testing set

	Training set				**Testing set**			
	Cases ($n = 60$)		Controls ($n = 60$)		Cases ($n = 61$)		Controls ($n = 61$)	
	n	%	n	%	n	%	n	%
Age								
50–60	9	15.0	7	11.7	7	11.5	7	11.5
60–70	21	35.0	27	45.0	26	42.6	28	45.9
70–80	30	50.0	26	43.3	28	45.9	26	42.6
Race/ethnicity[a]								
White	53	89.8	53	88.3	53	86.9	53	86.9
Non-white	6	10.2	7	11.7	8	13.1	8	13.1
Body mass index (kg/m2)								
<25.0 (normal)	26	43.3	20	33.3	25	41.0	21	34.4
25–29.9 (overweight)	17	28.3	20	33.3	22	36.1	21	34.4
30.0 (obesity)	17	28.3	20	33.3	14	23.0	19	31.1
Current use of HRT[a]								
Non user	20	34.5	20	34.5	21	35.0	21	35.0
Estrogen + progestin	13	22.4	13	22.4	13	21.7	13	21.7
Unopposed estrogen	25	43.1	25	43.1	26	43.3	26	43.3

[a] Numbers may not add to total subjects due to missing data

observed differences in the distributions of these variables between cases and controls, in either the training or testing set, reached statistical significance when assessed by χ^2 tests (data not shown). The antibody-array profiling platform exhibited high levels of both intra-array and inter-array reproducibility (Figures S1/S2), consistent with past performance [30, 39]. When comparing cases to their matched controls, statistically significant ($P < 0.05$) mean differences in signal intensity were observed for 328 of 2467 antibodies (13.3 %) in the training set. Of these 328 candidates, 39 were validated ($P < 0.05$) in the testing set (Table S2). In the pooled analysis (training and testing sets combined), geometric mean within-pair fold changes of (R_c/G_c) for the 39 identified candidates, comparing cases to matched controls, ranged in magnitude from 0.76–1.11 (Table 2). Four markers remained significant after Bonferroni correction for multiple comparisons ($P < 0.05/2467 = 2.03 \times 10^{-5}$): Colony stimulating factor 2 (CSF2), transferrin receptor (TFRC), RING1 and YY1 binding protein (RYBP), and integrin β4 (ITGB4) (Figure S3). Based on Western immunoblots, the printed array antibodies for these top candidates recognized antigens of the expected molecular weight and observed intensity changes across different plasma samples appeared consistent with array-based M values (Figure S4).

Nineteen additional candidates were captured when the false discovery rate was set to $q < 0.05$. TFRC and BRCA2 were each identified as nominal hits by two different antibodies. AUCs based on covariate-adjusted marker values ranged from 0.58 (CTGF) to 0.71 (CSF2), with maximum sensitivity of 33 % at 95 % specificity (RYBP), and maximum specificity of 30 % at 95 % sensitivity (TFRC). Ten of the 39 candidates had a sensitivity of ≥20 % at 95 % specificity, and three had a specificity of ≥20 % at 95 % sensitivity. Among the top-performing markers, some exhibited superior performance at high fixed levels of sensitivity, relative to high fixed levels of specificity, while others exhibited skewed performance in the opposite direction, as illustrated by the asymmetrically shaped ROC curves for TFRC and RYBP, respectively (Fig. 1). Exploratory analyses suggested that several markers identified may exhibit stronger associations (lower P values and similar or larger fold changes) closer to the time of BC diagnosis: TFRC, ITGB4, SNX5, CD2, NTRK1, CSF3R, BRCA1, and RASGRF2, among others (Figure S5). Conclusive interpretation was limited by reduced study power.

We next determined whether data from TCGA might provide further support for any of the ~40 candidate markers identified. Since protein expression data were not available, we evaluated RNA profiles for the corresponding genes in 130 Stage I/II ER+/PR+/ductal breast tumors and 61 normal tissue samples (Table S3). 25 of the 37 transcripts assessed exhibited statistically significant differences in mean abundance when comparing tumor and normal tissue samples, including three of our top five

Table 2 Top-ranked antibodies by paired *t* test, comparing cases versus matched controls ($P < 0.05$ in both training and testing sets)

	Antibody name	Gene	N^a	FC^b	P^c	q^d	N^e	$AUC^{f,i}$	$Sens^{g,i}$	$Spec^{h,i}$
1	Colony stimulating factor 2 (granulocyte–macrophage)	CSF2	97	1.08	4.00E − 08*	9.86E − 05	207	0.71	0.29	0.13
2	Transferrin receptor	TFRC	95	0.76	3.10E − 06*	0.004	207	0.67	0.06	0.30
3	RING1 and YY1 binding protein	RYBP	102	1.05	8.66E − 06*	0.007	216	0.69	0.33	0.10
4	Integrin, beta 4	ITGB4	120	1.05	1.24E − 05*	0.008	234	0.68	0.15	0.17
5	Matrix metallopeptidase 15 (membrane-inserted)	MMP15	119	1.10	2.03E − 05	0.008	233	0.64	0.17	0.11
6	Sorting nexin 5	SNX5	119	1.06	2.06E − 05	0.008	233	0.68	0.21	0.11
7	IMP (inosine 5′-monophosphate) dehydrogenase 2	IMPDH2	118	1.06	3.43E − 05	0.009	232	0.67	0.18	0.12
8	CD2 molecule	CD2	117	1.06	3.44E − 05	0.009	231	0.66	0.21	0.13
9	Chloride channel, voltage-sensitive 3	CLCN3	119	1.04	3.57E − 05	0.009	233	0.66	0.22	0.09
10	Homeobox D13	HOXD13	118	1.07	3.69E − 05	0.009	232	0.67	0.16	0.13
11	DNA replication and sister chromatid cohesion 1	DSCC1	121	1.04	0.0001	0.024	235	0.63	0.17	0.09
12	G protein-coupled receptor 19	GPR19	120	1.08	0.0001	0.028	234	0.61	0.15	0.12
13	Neurotrophic tyrosine kinase, receptor, type 1	NTRK1	112	1.07	0.0002	0.032	227	0.64	0.13	0.05
14	Breast cancer 2, early onset	BRCA2	121	1.05	0.0002	0.033	235	0.66	0.15	0.10
15	Colony stimulating factor 3 receptor (granulocyte)	CSF3R	121	1.04	0.0003	0.033	235	0.63	0.26	0.05
16	Nephroblastoma overexpressed	NOV	108	1.06	0.0003	0.033	221	0.63	0.19	0.09
17	Plexin D1	PLXND1	115	1.03	0.0003	0.033	229	0.63	0.18	0.15
18	Transferrin receptor	TFRC	116	0.92	0.0004	0.037	230	0.64	0.03	0.17
19	Breast cancer 1, early onset	BRCA1	109	1.06	0.0004	0.037	223	0.63	0.23	0.11
20	Tumor protein p63	TP63	118	1.04	0.0004	0.037	232	0.64	0.14	0.11
21	Angio-associated, migratory cell protein	AAMP	113	0.96	0.0005	0.043	227	0.61	0.06	0.10
22	Connective tissue growth factor	CTGF	112	0.96	0.0006	0.043	227	0.58	0.09	0.20
23	EPH receptor B6	EPHB6	120	1.03	0.0006	0.044	234	0.63	0.15	0.13
24	WD repeat and FYVE domain containing 3	WDFY3	120	1.03	0.0008	0.050	234	0.62	0.22	0.07
25	RAD23 homolog A (S. cerevisiae)	RAD23A	110	1.10	0.0009	0.053	224	0.63	0.15	0.07
26	Epithelial cell adhesion molecule	EPCAM	115	1.04	0.0009	0.054	229	0.60	0.16	0.04
27	Stromal cell derived factor 4	SDF4	119	1.04	0.0010	0.054	233	0.64	0.18	0.14
28	Ras protein-specific guanine nucleotide-releasing factor 2	RASGRF2	118	1.02	0.0013	0.062	232	0.63	0.23	0.03
29	Kallikrein-related peptidase 5	KLK5	120	1.03	0.0013	0.062	234	0.61	0.14	0.06
30	Breast cancer 2, early onset	BRCA2	118	1.05	0.0014	0.063	232	0.62	0.12	0.10
31	CD40 ligand	CD40LG	119	1.04	0.0015	0.063	232	0.60	0.11	0.11
32	Guanine nucleotide binding protein-like 3 (nucleolar)	GNL3	118	1.05	0.0018	0.068	232	0.60	0.12	0.10
33	Thrombospondin 3	THBS3	119	1.02	0.0019	0.070	233	0.63	0.14	0.09
34	Argonaute RISC catalytic component 2	AGO2	118	1.11	0.0020	0.073	232	0.60	0.20	0.10
35	PBX/knotted 1 homeobox 1	PKNOX1	116	1.04	0.0022	0.077	231	0.61	0.19	0.07
36	Microtubule-associated protein, RP/EB family, member 1	MAPRE1	120	1.04	0.0025	0.085	234	0.61	0.09	0.09
37	Glucagon-like peptide 1 receptor	GLP1R	111	0.95	0.0026	0.085	225	0.61	0.05	0.21
38	Corticotropin releasing hormone binding protein	CRHBP	98	1.08	0.0028	0.091	212	0.61	0.10	0.07
39	Zic family member 3	ZIC3	100	1.07	0.0032	0.101	213	0.60	0.11	0.09

[a] Number of case/control pairs with non-missing marker values

[b] Geometric mean within-pair fold change of R_c/G_c, comparing cases to controls: 2^[$M_{\text{case}} - M_{\text{control}}$]

[c] Paired *t* test *P* value

[d] False discovery rate (Benjamini and Hochberg)

[e] Number of subjects with non-missing marker values

[f] Area under the curve (AUC)

[g] Sensitivity at 95 % specificity

[h] Specificity at 95 % sensitivity

[i] Performance metric calculated using marker values adjusted for age, race, BMI, HRT use, hybridization day, *$P < (0.05/2467) = 2.03 \times 10^{-5}$ (Bonferroni threshold)

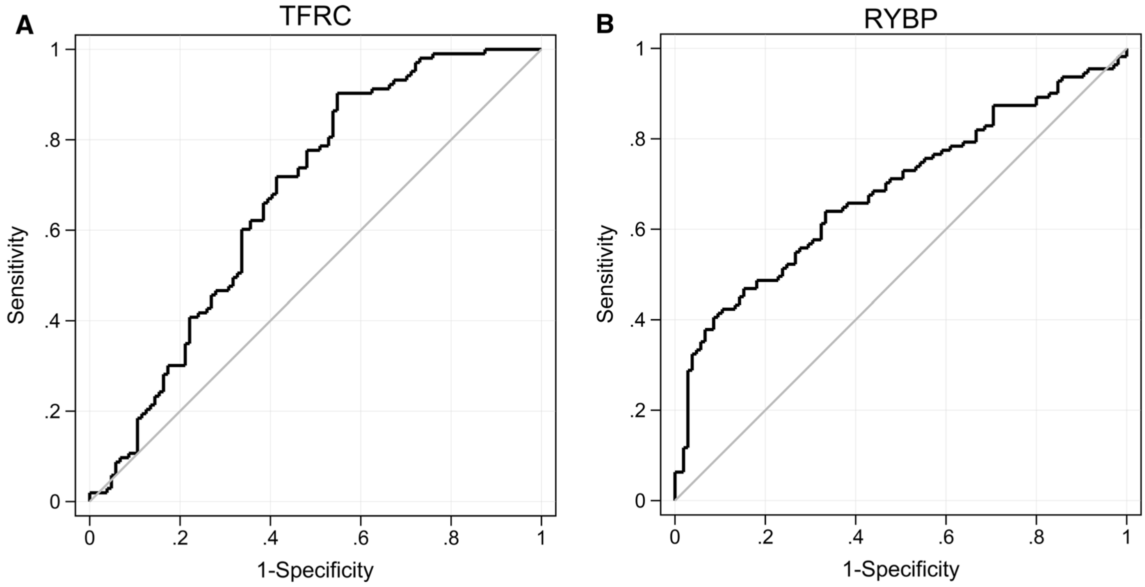

Fig. 1 Receiver-operating characteristic (ROC) curves for TFRC (**a**) and RYBP (**b**), based on covariate-adjusted marker values in the pooled sample set

ranked candidates (TFRC, RYBP, ITGB4). 14 of 25 showed changes directionally concordant with the plasma protein alterations identified in our BC cases versus controls.

To assess the utility of combining multiple candidates into a composite marker panel, we conducted multivariate modeling using regularized logistic regression with elastic net penalty [40]. The mean AUC of models built using the 39 candidates listed in Table 2 and assessed across 100 rounds of MCCV, was 0.75 (SD = 0.06), with estimated sensitivity of ~30 % at 95 % specificity (Fig. 2a). Parsimonious models built with the top four markers reaching Bonferroni significance (CSF2, TFRC, RYBP, ITGB4) achieved comparable classification accuracy (Fig. 2b).

Gene set enrichment analyses were conducted to assess whether plasma levels of proteins encoded by genes in

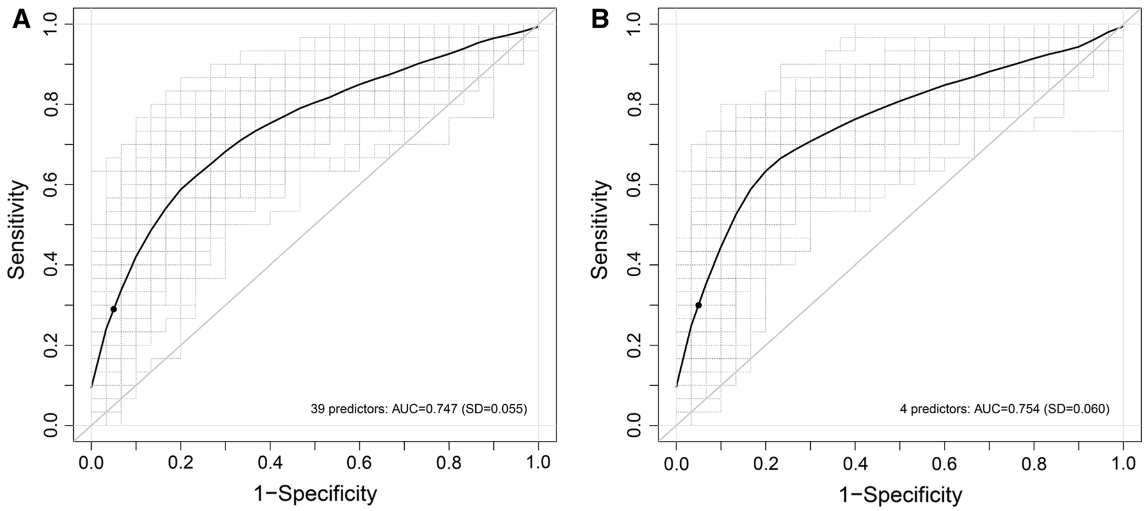

Fig. 2 Receiver-operating characteristic (ROC) curves for multivariate classifiers based on **a** all 39 candidate protein markers identified (Table 2), or **b** the top four candidate protein markers reaching Bonferroni significance (CSF2, TFRC, RYBP, ITGB4). Study participants were randomly divided into a training set (75 %: $n = 175$) or testing set (25 %: $n = 60$). A multivariate classifier was derived in the training set using the indicated markers and evaluated in the testing set. After 100 rounds of Monte Carlo cross-validation (MCCV), a composite average ROC curve was generated (*solid black*) from the individual ROC curves (*dotted lines*) derived in each testing set. Mean area under the ROC curve (AUC) was calculated across all MCCV iterations. Estimated sensitivity at 95 % specificity indicated by *black circles*

Table 3 Top gene sets identified as differentially abundant in cases versus matched controls, using training, testing, or pooled sample sets

S no.	GO gene set		P^a			q^b
			Training	Testing	Pooled	
1	Microtubule cytoskeleton	↑	0.001	0.024	<0.001	0.21
2	Response to hormone stimulus	↑	0.003	0.046	<0.001	0.21
3	Golgi apparatus	↑	0.045	0.042	0.004	0.50
4	Ruffle	↓	0.007	0.036	0.004	0.50
5	Tubulin binding	↑	0.035	0.035	0.010	0.50
6	Monocarboxylic acid metabolic process	↑	0.033	0.016	0.011	0.50
7	Negative regulation of cellular biosynthetic process	↑	0.041	0.022	0.014	0.50
8	Organic acid metabolic process	↑	0.033	0.020	0.014	0.50
9	Carboxylic acid metabolic process	↑	0.033	0.020	0.014	0.50
10	Negative regulation of biosynthetic process	↑	0.041	0.022	0.014	0.50
11	Embryo implantation	↓	0.029	0.042	0.031	0.50
12	Enzyme activator activity	↑	0.045	0.037	0.036	0.50

Arrows denote upregulation (↑) or downregulation (↓) of indicated gene set in cases relative to matched controls, based on results from the pooled sample set

[a] *P* value derived from Wilcoxon rank-sum test, comparing an observed sum of ranks (for a given pathway) in the actual dataset to a null distribution generated from 1000 permuted datasets

[b] False discovery rate (Benjamini and Hochberg), pooled sample set. 2396 antibody features were used in these analyses. Each included marker had non-missing values for at least 85 case: control pairs (≥70 %)

defined KEGG/GO gene sets were significantly, concordantly altered in cases relative to matched controls. Twelve of the GO gene sets (and none of the KEGG sets) reached significance in both the training and testing sets. (Table 3 and Figure S6). Two gene sets, microtubule cytoskeleton and response to hormone stimulus, remained significant in the pooled analysis when allowing for a false discovery rate of $q < 0.25$. Both appeared to be upregulated in cases versus controls.

Discussion

This biomarker discovery study coupled the use of pre-clinical plasma specimens with a powerful antibody-array platform to identify candidate blood-based biomarkers for ER+/PR+ invasive ductal BC. Our arrays included >3000 antibodies directed against >2000 proteins, which encompass many cytokines, adipokines, growth factors, and other secreted or membrane proteins implicated in a range of biological pathways dysregulated in BC. The putative markers identified (37 distinct proteins) include both previously implicated and novel proteins in BC pathogenesis. Half of these candidates are located either in the plasma membrane ($n = 12$) or extracellular space ($n = 8$), with the remainder divided between the nucleus ($n = 10$) or cytoplasm ($n = 7$) (Ingenuity Pathway Analysis). The two largest functional groups are membrane receptors ($n = 9$) and transcription regulators ($n = 7$), with kinases, peptidases, transporters, growth factors, and cytokines accounting for most of the remaining share.

Among four candidates meeting Bonferroni significance, two (TFRC and ITGB4) have been the focus of multiple BC-related studies [41–46]. TFRC (transferrin receptor), which functions in cellular uptake of iron and is expressed on highly proliferating cells, was identified by two independent antibodies on our array. Recent studies have linked elevated TFRC protein expression in ER+ breast tumors to higher clinical grade, increased proliferative activity, and worse prognosis [41, 42]. Membrane-bound TFRC is proteolytically cleaved to generate a soluble form of the receptor, which circulates in the plasma [47]. Interestingly, we observed *lower* levels of TFRC protein in the pre-clinical plasma of BC cases relative to controls. It remains uncertain, however, whether reduced levels of circulating TFRC reflect decreased abundance of cellular TFRC, particularly in breast tissue. In contrast to TFRC, ITGB4 (β4 integrin) was elevated in the plasma of cases relative to controls. ITGB4 belongs to the integrin family of transmembrane receptors, which function in cell adhesion, migration, and proliferation [48]. β4 integrin has been linked to increased breast tumor size, nuclear grade, and BC progression, and is believed to function in part through activation of the MAPK and PI3K/Akt signaling pathways [44–46, 49]. The 37 candidates for ER+/PR+ ductal BC show minimal if any overlap with those identified in our studies of triple-negative BC. While our published report on TN BC [27] used an earlier-generation array with only limited antibody overlap with the current arrays, more recent side-by-side discovery studies for ER+/PR+ ductal and TN BC on the same platform point to distinct patterns of proteomic alterations in the pre-clinical plasma of women who develop these types of cancer (C. Li, unpublished observations). Such differences are compatible with our growing recognition of BC sub-types as disparate diseases on the epidemiologic, clinical, and molecular levels, but larger studies are required for confirmation.

The mechanisms by which pre-clinical BC may lead to changes in the circulating plasma remain largely unknown. Detected alterations, if not due to chance, could represent (i) biological signals originating in the evolving target tissue (breast) or (ii) systemic changes or responses associated with a nascent cancer. In the former scenario, dysregulated neoplastic physiology might yield a modified repertoire of secreted proteins [50], while elevated levels of necrosis/apoptosis could produce increased spillage of cellular components. Alternatively, enhanced systemic inflammation or altered immune profiles [51] could underlie observed changes in the plasma/serum. Future studies using inducible in vivo models of carcinogenesis and paired analysis of blood and tissue may provide further mechanistic insights [52].

Our TCGA analyses indicated that 14 of our top 37 serum protein candidates exhibited significantly altered and concordant changes in RNA expression, when comparing breast tumor versus normal tissue. It remains unknown why only a subset of the markers showed parallel changes in tissue RNA expression, and some showed altered expression in the opposite direction. However, we note that RNA and protein expression levels of a given gene within a particular tissue may show only weak to moderate correlation [53], likely reflecting complex post-transcriptional and post-translational regulatory mechanisms. When considering serum protein and tissue RNA levels, the rate of protein secretion or spillage into the bloodstream represents yet another potential layer of regulation.

If the identified marker candidates are externally validated, an important objective will be to explore their potential clinical utility. Under current practice, age and sex remain the sole factors used on a widespread basis in determining whether or not a screening mammogram is recommended. More refined methods of BC risk assessment that integrate epidemiologic factors with clinical parameters (e.g., plasma protein measurements) could tailor screening recommendations. Three scenarios for using blood-based biomarkers to guide mammography usage relate to identifying subgroups of higher-risk women who are suitable for additional imaging: women aged 50–69 who should be recalled early after a negative mammogram; healthy women aged 75–79 who may benefit from continued mammography; and women aged 50–69 who should be prioritized for imaging in low-resource settings (e.g., rural areas). In these scenarios, marker panels achieving even modest sensitivity (e.g., ~30 %) at high fixed levels of specificity (95 %), such as those derived in this study, would have the potential to capture a significant number of additional women who are most likely to benefit from mammography, without flooding the system with those least likely to need imaging. One important caveat with respect to our modeling results is that performance estimates were derived using the same set of samples used to identify the included biomarker predictor variables. Given the resulting likelihood of some level of optimism bias, studies in external study populations are essential.

Our antibody-array platform has certain limitations. First, it interrogates only a subset of the complete plasma proteome, dictated by the set of included antibodies. Our pathway analyses were somewhat limited by the absence of data on many proteins belonging to KEGG/GO gene sets. Second, relative to gold standard assays (sandwich ELISA), the antibody array does not incorporate enzymatic amplification or require multiple antibodies to bind a given target at distinct epitopes. Thus, technical sensitivity and specificity are unlikely to match the parameters of targeted assays. Nevertheless, the array platform has shown excellent concordance with ELISA for specific protein targets [27, 30, 31]. Third, some antibodies failed on a sizable number of study subjects, due to low intensity or highly variable signals. The intended use of our array platform is for discovery purposes, however, and any potential downstream clinical assays would likely evaluate a small number markers using ELISA.

Additional studies are needed to replicate our array-based proteomic profiling results on independent sample sets. Confirmed candidates would then be advanced as potential biomarkers for further validation using gold standard assays capable of absolute quantification, when feasible. The potential clinical utility of these results will require further investigation in larger external cohorts. Integration of blood-based biomarker data and clinical/epidemiologic characteristics may ultimately facilitate development of multivariate classification models for accurate BC risk stratification and early diagnosis.

Conflicts of interest

The authors disclose no potential conflicts of interest.

Acknowledgments This work was principally supported by NIH grant U01CA152637 (C.I.L) and NIH training grant T32CA009168 (M.F.B).

References

1. Kerlikowske K, Grady D, Rubin SM et al (1995) Efficacy of screening mammography. A meta-analysis. JAMA 273:149–154
2. Nyström L, Andersson I, Bjurstam N et al (2002) Long-term effects of mammography screening : updated overview of the Swedish randomised trials. Lancet 359:909–919. doi:10.1016/S0140-6736(02)08020-0
3. Humphrey LL, Helfand M, Chan BK, Woolf SH (2002) Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 137:347–360
4. USPSTF (2009) Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 151:716–726. doi:10.7326/0003-4819-151-10-200911170-00008 **W–236**
5. Miller AB, Wall C, Baines CJ (2014) Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. BMJ 348:g366–g366. doi:10.1136/bmj.g366
6. Brawley OW (2012) Risk-based mammography screening: an effort to maximize the benefits and minimize the harms. Ann Intern Med 156:662–663. doi:10.7326/0003-4819-156-9-201205010-00012
7. Pace LE, Keating NL (2014) A systematic assessment of benefits and risks to guide breast cancer screening decisions. JAMA 311:1327–1335. doi:10.1001/jama.2014.1398
8. Onega T, Beaber EF, Sprague BL et al (2014) Breast cancer screening in an era of personalized regimens: a conceptual model and National Cancer Institute initiative for risk-based and preference-based approaches at a population level. Cancer 120:2955–2964. doi:10.1002/cncr.28771
9. Phillips M, Beatty JD, Cataneo RN et al (2014) Rapid point-of-care breath test for biomarkers of breast cancer and abnormal mammograms. PLoS One 9:e90226. doi:10.1371/journal.pone.0090226
10. Lacombe J, Mangé A, Bougnoux A-C et al (2014) A multiparametric serum marker panel as a complementary test to mammography for the diagnosis of node-negative early-stage breast cancer and DCIS in young women. Cancer Epidemiol Biomark Prev 23:1834–1842. doi:10.1158/1055-9965.EPI-14-0267
11. Brooks M (2009) Breast cancer screening and biomarkers. Methods Mol Biol 472:307–321. doi:10.1007/978-1-60327-492-0_13
12. Li CI (2011) Discovery and validation of breast cancer early detection biomarkers in preclinical samples. Horm Cancer 2:125–131. doi:10.1007/s12672-010-0061-3
13. Zaenker P, Ziman MR (2013) Serologic autoantibodies as diagnostic cancer biomarkers–a review. Cancer Epidemiol Biomark Prev 22:2161–2181. doi:10.1158/1055-9965.EPI-13-0621
14. Coronell JAL, Syed P, Sergelen K (2012) The current status of cancer biomarker research using tumour-associated antigens for minimal invasive and early cancer diagnostics. J Proteomics 76:102–115. doi:10.1016/j.jprot.2012.07.022
15. Gong B, Xue J, Yu J et al (2012) Cell-free DNA in blood is a potential diagnostic biomarker of breast cancer. Oncol Lett 3:897–900. doi:10.3892/ol.2012.576
16. Ng EKO, Li R, Shin VY et al (2013) Circulating microRNAs as specific biomarkers for breast cancer detection. PLoS One 8:e53141. doi:10.1371/journal.pone.0053141
17. Cuk K, Zucknick M, Heil J et al (2013) Circulating microRNAs in plasma as early detection markers for breast cancer. Int J Cancer 132:1602–1612. doi:10.1002/ijc.27799
18. Lu H, Ladd J, Feng Z et al (2012) Evaluation of known oncoantibodies, HER2, p53, and cyclin B1, in prediagnostic breast cancer sera. Cancer Prev Res (Phila) 5:1036–1043. doi:10.1158/1940-6207.CAPR-11-0558
19. Pitteri SJ, Amon LM, Busald Buson T et al (2010) Detection of elevated plasma levels of epidermal growth factor receptor before breast cancer diagnosis among hormone therapy users. Cancer Res 70:8598–8606. doi:10.1158/0008-5472.CAN-10-1676
20. Fischer JC, Niederacher D, Topp SA et al (2013) Diagnostic leukapheresis enables reliable detection of circulating tumor cells of nonmetastatic cancer patients. Proc Natl Acad Sci USA 110:16580–16585. doi:10.1073/pnas.1313594110
21. Reis-Filho JS, Pusztai L (2011) Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet 378:1812–1823. doi:10.1016/S0140-6736(11)61539-0
22. Eroles P, Bosch A, Pérez-Fidalgo JA, Lluch A (2012) Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. Cancer Treat Rev 38:698–707. doi:10.1016/j.ctrv.2011.11.005
23. Bao P-P, Shu XO, Gao Y-T et al (2011) Association of hormone-related characteristics and breast cancer risk by estrogen receptor/progesterone receptor status in the shanghai breast cancer study. Am J Epidemiol 174:661–671. doi:10.1093/aje/kwr145
24. Yang XR, Chang-Claude J, Goode EL et al (2011) Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. J Natl Cancer Inst 103:250–263. doi:10.1093/jnci/djq526
25. Pestalozzi BC, Zahrieh D, Mallon E et al (2008) Distinct clinical and prognostic features of infiltrating lobular carcinoma of the breast: combined results of 15 International Breast Cancer Study Group clinical trials. J Clin Oncol 26:3006–3014. doi:10.1200/JCO.2007.14.9336
26. Desantis C, Ma J, Bryan L, Jemal A (2014) Breast cancer statistics, 2013. CA Cancer J Clin 64:52–62. doi:10.3322/caac.21203
27. Li CI, Mirus JE, Zhang Y et al (2012) Discovery and preliminary confirmation of novel early detection biomarkers for triple-negative breast cancer using preclinical plasma samples from the Women's Health Initiative observational study. Breast Cancer Res Treat 135:611–618. doi:10.1007/s10549-012-2204-4
28. Hays J, Hunt JR, Hubbell FA et al (2003) The Women's Health Initiative recruitment methods and results. Ann Epidemiol 13:S18–S77
29. Study, The Women'S Health Initiative (1998) Design of the Women's Health Initiative clinical trial and observational study. Control Clin Trials 19:61–109
30. Loch CM, Ramirez AB, Liu Y et al (2007) Use of high density antibody arrays to validate and discover cancer serum biomarkers. Mol Oncol 1:313–320. doi:10.1016/j.molonc.2007.08.004
31. Ramirez AB, Loch CM, Zhang Y et al (2010) Use of a single-chain antibody library for ovarian cancer biomarker discovery. Mol Cell Proteomics 9:1449–1460. doi:10.1074/mcp.M900496-MCP200
32. Mirus JE, Zhang Y, Li CI et al (2015) Cross-species antibody microarray interrogation identifies a 3-protein panel of plasma biomarkers for early diagnosis of pancreas cancer. Clin Cancer Res 21:1764–1771. doi:10.1158/1078-0432.CCR-13-3474
33. Ritchie ME, Silver J, Oshlack A et al (2007) A comparison of background correction methods for two-colour microarrays. Bioinformatics 23:2700–2707. doi:10.1093/bioinformatics/btm412
34. Smyth GK, Speed T (2003) Normalization of cDNA microarray data. Methods 31:265–273
35. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22

36. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21:3940–3941. doi:10.1093/bioinformatics/bti623
37. The Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474:609–615. doi:10.1038/nature10166
38. The Cancer Genome Atlas Research Network (2012) Comprehensive molecular portraits of human breast tumours. Nature 490:61–70. doi:10.1038/nature11412
39. Rho JH, Mead JR, Wright WS et al (2014) Discovery of sialyl Lewis A and Lewis X modified protein cancer biomarkers using high density antibody arrays. J Proteomics 96:291–299. doi:10.1016/j.jprot.2013.10.030
40. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol) 67:301–320. doi:10.1111/j.1467-9868.2005.00503.x
41. Habashy HO, Powe DG, Staka CM et al (2010) Transferrin receptor (CD71) is a marker of poor prognosis in breast cancer and can predict response to tamoxifen. Breast Cancer Res Treat 119:283–293. doi:10.1007/s10549-009-0345-x
42. Miller LD, Coffman LG, Chou JW et al (2011) An iron regulatory gene signature predicts outcome in breast cancer. Cancer Res 71:6728–6737. doi:10.1158/0008-5472.CAN-11-1870
43. Lu S, Simin K, Khan A, Mercurio AM (2008) Analysis of integrin beta4 expression in human breast cancer: association with basal-like tumors and prognostic significance. Clin Cancer Res 14:1050–1058. doi:10.1158/1078-0432.CCR-07-4116
44. Diaz LK, Cristofanilli M, Zhou X et al (2005) Beta4 integrin subunit gene expression correlates with tumor size and nuclear grade in early breast cancer. Mod Pathol 18:1165–1175. doi:10.1038/modpathol.3800411
45. Lipscomb EA, Simpson KJ, Lyle SR et al (2005) The alpha6beta4 integrin maintains the survival of human breast carcinoma cells in vivo. Cancer Res 65:10970–10976. doi:10.1158/0008-5472.CAN-05-2327
46. Dutta U, Shaw LM (2008) A key tyrosine (Y1494) in the β4 integrin regulates multiple signaling pathways important for tumor development and progression. Cancer Res 68:8779–8787. doi:10.1158/0008-5472.CAN-08-2125
47. Beguin Y, Huebers HA, Josephson B, Finch CA (1988) Transferrin receptors in rat plasma. Proc Natl Acad Sci USA 85:637–640
48. Desgrosellier JS, Cheresh DA (2010) Integrins in cancer: biological implications and therapeutic opportunities. Nat Rev Cancer 10:9–22. doi:10.1038/nrc2748
49. Bon G, Folgiero V, Di Carlo S et al (2007) Involvement of alpha6beta4 integrin in the mechanisms that regulate breast cancer progression. Breast Cancer Res 9:203. doi:10.1186/bcr1651
50. Chen S-T, Pan T-L, Juan H-F et al (2008) Breast tumor microenvironment: proteomics highlights the treatments targeting secretome. J Proteome Res 7:1379–1387. doi:10.1021/pr700745n
51. Anderson KS, Sibani S, Wallstrom G et al (2011) Protein microarray signature of autoantibody biomarkers for the early detection of breast cancer. J Proteome Res 10:85–96. doi:10.1021/pr100686b
52. Hanash SM, Taguchi A (2014) Mouse to human blood-based cancer biomarker discovery strategies. Cold Spring Harb Protoc 2014:144–149. doi:10.1101/pdb.top078808
53. Pascal LE, True LD, Campbell DS et al (2008) Correlation of mRNA and protein levels: cell type-specific gene expression of cluster designation antigens in the prostate. BMC Genom 9:246. doi:10.1186/1471-2164-9-246