Journal of proteome research

Article

pubs.acs.org/jpr

Integrating Meta-Analysis of Microarray Data and Targeted Proteomics for Biomarker Identification: Application in Breast Cancer

Maria P. Pavlou,[†,‡] Apostolos Dimitromanolakis,[‡] Eduardo Martinez-Morillo,[§] Marcel Smid,[∥] John A. Foekens,[∥] and Eleftherios P. Diamandis*[,†,‡,§]

[†]Department of Laboratory Medicine and Pathobiology, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada

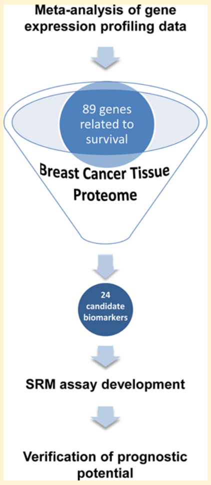
[‡]Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, 60 Murray Street, Toronto, ON M5T 3L9, Canada

[§]Lunenfeld-Tanenbaum Research Institute, Joseph and Wolf Lebovic Health Complex, Mount Sinai Hospital, 60 Murray Street, Toronto, ON M5T 3L9, Canada

[∥]Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Groene Hilledijk 301, 3075 EA Rotterdam, The Netherlands

Supporting Information

ABSTRACT: The development of signature biomarkers has gained considerable attention in the past decade. Although the most well-known examples of biomarker panels stem from gene expression studies, proteomic panels are becoming more relevant, with the advent of targeted mass spectrometry-based methodologies. At the same time, the development of multigene prognostic classifiers for early stage breast cancer patients has resulted in a wealth of publicly available gene expression data from thousands of breast cancer specimens. In the present study, we integrated transcriptome and proteome-based platforms to identify genes and proteins related to patient survival. Candidate biomarker proteins have been identified in a previously generated breast cancer tissue extract proteome. A mass-spectrometry-based assay was then developed for the simultaneous quantification of these 20 proteins in breast cancer tissue extracts. We quantified the relative expression levels of the 20 potential biomarkers in a cohort of 96 tissue samples from patients with early stage breast cancer. We identified two proteins, KPNA2 and CDK1, which showed potential to discriminate between estrogen receptor positive patients of high and low risk of disease recurrence. The role of these proteins in breast cancer prognosis warrants further investigation.


Meta-analysis of gene expression profiling data
89 genes related to survival
Breast Cancer Tissue Proteome
24 candidate biomarkers
SRM assay development
Verification of prognostic potential

KEYWORDS: *selected reaction monitoring, gene expression data meta-analysis, breast cancer prognosis, mass spectrometry, targeted proteomics*

INTRODUCTION

The development of biomarker panels, rather than single biomarkers, has emerged as an attractive approach after recognizing the biological heterogeneity of human disease and the multiple molecular pathways involved during disease progression. Although the most well-known examples of biomarker panels stem from gene expression studies,[1] similar panels could be developed at the proteome level. Targeted mass-spectrometry-based methodologies (such as selected reaction monitoring, SRM) provide an effective platform for evaluation of proteomic signatures. As implied by the name, targeted mass spectrometric approaches require a priori knowledge of the analytes to be detected and enable relative or absolute quantification of multiple peptides, and therefore proteins, in a biological sample.[2] Small sample amount requirements, multiplexing capability, high selectivity, and cost- and time-efficient development of assays are the major advantages of targeted mass-spectrometry-based assays.

Breast cancer is a major health issue, affecting annually ∼1.4 million women worldwide.[3] The advent of high-throughput platforms for gene expression analysis, such as microarrays, has led to studies that revolutionized the way breast cancer is perceived. Microarray studies undertaken the past decade gave rise to a molecular classification system and numerous prognostic multigene classifiers for breast cancer.[4] A "by-product" originating from these studies is an unprecedented access to gene expression data from hundreds to thousands of breast cancer specimens deposited in public data repositories (e.g., the National Centre for Biotechnology Information (NCBI) gene expression omnibus (GEO),[5] ArrayExpress[6]). Given that the samples analyzed belong to well-annotated cohorts with long follow-up, the significance of these publicly available data is high. Taking also into consideration the fact

Received: January 17, 2014
Published: May 5, 2014

that sample availability is one of the main obstacles in biomarker development,[7] the publicly available gene expression data are a valuable resource for the scientific community.

Despite the undeniable contribution of gene expression in the field of breast oncology, proteins are the mediators of all biological processes and the molecular targets of the majority of drugs. Additionally, they are more dynamic than nucleic acids and may be more reflective of cellular physiology because they integrate the cellular genetic information with the effect of the immediate environment.[8] More specifically, in the breast cancer field, proteomic technologies have been recently applied in the quest of novel biomarkers.[9,10] Furthermore, quantitative protein-based assays are cost-effective and easy to use; therefore, they are considered the gold standard in the clinic.

In the present study, we integrated transcriptome and proteome-based platforms to identify genes related to patient survival that have also been identified at the protein level in a breast cancer tissue extract proteome previously generated by our group,[11] postulating that these proteins may have prognostic potential. Given that the debate on assessing prognosis is particularly intense in the management of breast cancer patients with localized, estrogen receptor (ER)-positive tumors, where the option of targeted therapies exists, we focused on ER-positive patients only. We sought to identify genes that are related to 5-year disease-free survival (DFS) of ER-positive patients using gene expression profiling data from four independent cohorts. Then, we compared the generated list to our breast cancer tissue extract proteome and selected only genes that have been identified at the protein level. Next, we developed an SRM method for the relative quantification of the selected proteins in breast cancer tissue samples. Finally, the relative levels of candidate biomarkers were measured using the developed SRM method in a selected cohort composed of 96 patients with primary, lymph-node-negative breast cancer.

■ MATERIALS AND METHODS

Tissue mRNA Expression Microarray Data Analysis

NCBI GEO was queried for data sets with breast cancer tissue microarray data. The data sets A: GSE7390,[12] B: GSE2034,[13] C: GSE21653,[14] and D: GSE4922[15] were selected to identify genes related to 5-year DFS (in the case of GSE2034 data set the end point was distant metastasis-free survival). The four studies were performed on a common microarray platform (Affymetrix Human Genome 133A chip − a high coverage chip that includes 22 283 probes mapping to 12 688 unique genes), and description of clinical outcome along with censoring status were also available. Microarray data were normalized using gcRMA algorithm and quality-controlled according to Affymetrix guidelines, as previously described.[11] Because many probes in any microarray experiment do not show significant variation (resulting from genes not expressed in a specific tissue or nonspecific binding of probe sets), a limit on the mean interquartile range (IQR) of each of the probes across the four data sets was imposed. Probes showing less than two-fold ratio across the patient median were excluded. This exercise resulted in a set of 3124 probes that showed highly variable expression among patients. This filtering also limited the impact of multiple testing penalty. Patients were then divided into two groups: patients with DFS lower and higher than 5 years. A *t* test was performed to identify genes differentially expressed between the two groups ($p < 0.05$) in at least three out of four data sets.

Breast Cancer Tissue Samples and Preparation for SRM Analysis

To evaluate the relative expression levels of potential biomarkers with SRM, we selected 96 breast cancer tissues. The selection was based on ER status and on whether the patients had developed a distant metastasis within 5 years (poor prognosis) or remained free of disease for more than 7 years (good prognosis). The selection was such that of the included 48 ER-positive patients 24 had a good prognosis and 24 had a poor prognosis. A detailed description of the cohort can be found in Table 1. Breast cancer tissues were processed as previously described[16] and remained frozen at −80 °C until assayed. The total protein concentration of all samples was adjusted to 1 mg/mL.

Table 1. Demographics of the Patients for the 96 Breast Cancer Tissue Samples Analyzed in the Present Study

		ER status	
		ER-negative	ER-positive
total number		48	48
tumor stage	T1	15	32
	T2	29	13
	T3	2	2
	Tx	2	1
menopausal status	pre-	22	24
	post-	26	24
age	≤40	11	1
	>40−≤55	16	30
	>55−≤70	17	11
	>70	4	6
age mean		53.7	55
age median		52.5	53

Sample volume corresponding to 30 μg of total protein was diluted four times with 50 mM ammonium bicarbonate (Fisher Scientific) buffer solution, and proteins were denatured with 0.05% RapiGest (Waters) at 60 °C. The disulfide bonds were reduced with dithiothreitol (final concentration, 10 mM; Sigma-Aldrich) before being subjected to alkylation with iodoacetamide at room temperature for 45 min in the dark (final concentration, 20 mM; Sigma-Aldrich). Samples were then digested with sequencing-grade-modified trypsin (trypsin: protein ratio of 1:30; Promega, sequencing-grade modified porcine trypsin) overnight at 37 °C. RapiGest (Waters) was cleaved with 1% trifluoroacetic acid (TFA, Fisher Scientific), samples were centrifuged at 453 g for 10 min, and supernatant was carefully collected to avoid pellet contamination. In experiments where isotope-labeled peptides were utilized, they were spiked into the samples after protein digestion and prior to RapiGest precipitation by TFA (Fisher Scientific). Volume corresponding to 15 μg of peptides was purified and extracted using ZipTip C18 pipet tips (Millipore) and was eluted using 4.5 μL of mobile phase B (55% acetonitrile (ACN), 0.1% formic acid, Fisher Scientific). Fifty-six μL of mobile phase A (0.1% formic acid, Fisher Scientific) was added to each sample to yield three injections of 18 μL. During the verification experiment, all samples were processed at the same time (96-well plate) in a randomized and blinded manner.

Metabolic Labeling of Breast Cancer Cell Lines

Two cell lines (SK-BR-3 and MDA-MB-231) were purchased by the American Tissue Culture Collection (ATCC) and metabolically labeled as follows. Stable isotope labeling with amino acids in cell culture (SILAC) media was prepared from customized RPMI-1640 media devoid in two essential amino acids: L-arginine and L-lysine (AthenaES). Heavy amino acids, L-arginine-6 (^{13}C) and L-lysine-8 (^{13}C and ^{15}N) (Cambridge Isotope Laboratories), were supplemented to the medium at a concentration of 87 and 54 mg/L, respectively, for the "heavy" medium. For the control medium ("light"), amino acids L-arginine and L-lysine (Sigma-Aldrich) were supplemented at a final concentration of 84 and 52 mg/L each. Both heavy and light media were supplemented with L-proline (Sigma-Aldrich) at a concentration of 150 mg/L. All amino acids were reconstituted in phosphate-buffered saline (PBS, Gibco) and were filtered through a 0.22 μm filter to obtain a sterile solution (Millipore). Additionally, 10% of dialyzed fetal bovine serum (FBS, Gibco) was added to both heavy and light media. A minimum of five doubling times was ensured to achieve high-efficiency (>97%) labeling.

Sample Preparation of Cell Lines for SRM Analysis

Cells were washed twice with PBS (Gibco), detached using trypsin (Gibco), and centrifuged at 290*g* for 10 min, and supernatants were discarded. Cell pellets were kept at −80 °C until they were further processed. Cell lysis and protein digestion was performed as previously described,[17] and equimolar amounts of "heavy" and "light" cell lysates were mixed. Volume corresponding to 15 μg of peptides was purified and extracted using ZipTip C18 pipet tips (Millipore) and were eluted using 4.5 μL of mobile phase B (55% ACN, 0.1% formic acid, Fischer Scientific). Fifty-six μL of mobile phase A (0.1% formic acid, Fisher Scientific) was added to each sample to yield three injections of 18 μL.

Peptide Selection for SRM Method Development

Between three and five doubly charged proteotypic (PTP) peptides (length of 8 to 20 amino acids) per protein were initially selected. Peptides were selected from the peptide spectral library of the breast tissue extract proteome previously generated in-house. Peptides with N-terminus glutamine (Q), cysteine (C), or asparagine (N) were excluded. Also peptides containing histidine (H) in the middle of the sequence were avoided, if possible. In cases that the discovery data did not render sufficient numbers of peptides (at least three), Global Proteome Machine (GPM) database (http://gpmdb.thegpm.org/) was searched. Peptide uniqueness was confirmed by searching against the Basic Local Alignment Search Tool (BLAST; http://blast.ncbi.nlm.nih.gov/). In silico digestion, fragmentation, and prediction of collision energy were performed using Skyline software.[18]

Liquid Chromatography (LC)/Mass Spectrometry Conditions

Samples were loaded onto a 2 cm trap column (C18, 5 μm) with an inner diameter of 150 μm, and the peptides were eluted onto a resolving 5 cm analytical column (C18, 3 μm) with an inner diameter of 75 and 8 μm tip (New Objetive). The LC setup, EASY-nLC 1000 (Thermo Fisher), was coupled online to a triple-quadrupole mass spectrometer (TSQ Vantage, Thermo Fisher) using a nanoelectrospray ionization source (nano-ESI, Thermo Fisher). A three-step 60 min gradient with an injection volume of 18 μL was used. Buffer A contained 0.1% formic acid in water, and buffer B contained 0.1% formic acid in acetonitrile (Fisher Scientific). Peptides were analyzed by SRM assays with the following parameters: predicted collision energy values, 0.2 Da fhwm at the first quadrupole, 0.7 Da fhwm at the third quadrupole, 1.5 mTorr pressure at the collision cell, tuned tube lens values, and 7 V skimmer offset.

Identification of Optimum Peptides for SRM Method Development

Peptide identification was confirmed in four ways: (1) by observing the coelution of, at least, six transitions per peptide; (2) prediction of retention times (RTs) using SRRCalc 3.0, 300 Å (Skyline software, version 1.4); (3) comparing the observed fragmentation pattern of these peptides (SRM methods) with the fragmentation pattern displayed in our in-house breast cancer tissue extract proteome (discovery data); and (4) by observing the coelution of transitions originating from the "heavy" and "light" peptides, as also described by Liu et al.[19] For RT prediction, a 0.2 mg/mL bovine serum albumin (BSA, Sigma) solution with 10 isotope-labeled standard peptides (SpikeTides TQL, JPT Peptide Technologies) was used. A multiplex SRM assay with 28 peptides (18 peptides from BSA and 10 isotope-labeled peptides) was ran in a 60 min gradient, and the measured RTs were utilized to predict the RT and 95% confidence intervals (CIs) of target peptides using Skyline software (SRRCalc 3.0).

Selection of Transitions for SRM Method Development

Three transitions per peptide were selected based on two main criteria: relative intensity (according to the results in breast cancer tissues) and presence of interferences. Transitions with the highest intensity were preferred. Presence of interferences was predicted by using the SRM collider software, version 1.4 (www.srmcollider.org). SRM collider predicts unique ion signatures (UISs) for each peptide. The search parameters utilized were: SSRCalc window: 10 arbitrary units; Q1 mass window: 0.2 Th; Q3 mass window: 0.7 Th; low and high mass threshold for transitions: 300 and 1500 Th, respectively; genome: Human Peptide Atlas; consider isotopes up to 3 amu; one missed cleavage; find UIS up to order 3; and finally, charge check, modifications, and all background ion series were selected.

Optimization of the Amount of Spiked-in Isotope-Labeled Peptides

Lyophilized peptides (JPT Peptide Technologies) were reconstituted in 100 μL of 20% ACN (Fisher Scientific) in 0.1 M ammonium bicarbonate (Fisher Scientific) and divided in three aliquots to ensure no repeating freeze−thaw cycles. Equal volumes of the heavy peptides were mixed to create a master stock solution. A pool of breast cancer cytosols was digested as previously described. Before precipitation of RapiGest (Waters) with TFA (Fisher Scientific), isotope-labeled peptides were added in the matrix, and serial dilutions covering three orders of magnitude were prepared and analyzed with our method. Scan time and time windows were adjusted to ensure the measurement of at least 15−20 points per LC peak.

Data Analysis

The raw files were uploaded to Pinpoint software, version 1.0 (Thermo Fisher), which was used for quantification of the area under the curve (AUC). The ratio AUC_{light}/AUC_{heavy} was multiplied by the amount of isotope-labeled peptide added in the sample to estimate the relative amount of each native peptide (expressed in fmoles per injection). These values were used for further analysis.

Analytical Range and Limit of Quantification of SRM Assays

For the study of linearity, a pool of breast cancer tissues that expressed most of the investigated proteins was prepared. The pool was digested and then separated in 13 parts. Isotope-labeled peptides (250 fmoles of each peptide per injection) were added in the first part, and this solution was sequentially diluted (1:2) to generate 13 points of calibration (250, 125, 62.5, 31.25, 15.62, 7.81, 3.9, 1.95, 0.98, 0.48, 0.24, 0.12, and 0.06 fmoles per injection). The standard solutions were analyzed in triplicate (except 250 that was analyzed in singleton) and in order from lowest to highest concentration. The limit of quantification (LOQ) was estimated as the concentration with a coefficient of variation (CV) lower than 20% and within the linear range.

Statistical Analysis

Protein levels were univariately associated with the relapse status of the patients using the nonparametric Mann−Whitney U-test. All statistical tests were performed using Stata v11 (StataCorp, College Station, Texas), and two-sided p value <0.05 was considered significant.

■ RESULTS

Identification of Genes Related to Disease Free Survival Using Publicly Available Microarray Data

The gene selection procedure was based on microarray profiles of breast cancer tissue samples from 607 ER-positive patients across four different studies summarized in Table 2. Although these studies included both ER-positive and ER-negative patients, only ER-positive patients were used in the analysis. The breast cancer patients of all studies were untreated at the time of sampling (surgery), and three out of four studies (experiments A−C) included only lymph-node-negative patients. Following microarray data meta-analysis, 89 genes were found to be differentially expressed (p < 0.05) between patients with DFS lower or higher than 5 years in at least three out of four data sets. Out of these 89 genes, 76 were overexpressed in the group of patients with poor prognosis (DFS < 5 years), whereas 13 of these were overexpressed in the group of patients with favorable prognosis (DFS > 5 years). On the basis of this selection procedure, the statistically expected number of false-positive findings is fewer than 3 out of the 89 genes (two tail; p < 0.05). All genes, except for one, exhibited the same direction of effect (favorable/unfavorable) in all data sets, even for the cases where the effect itself was not statistically significant, as it can be seen in Supplementary Table 1 in the Supporting Information.

Table 2. Demographics of the Four Microarray Experiments Used for Identifying Genes Related to Disease Free Survival (DFS)[a]

ID	experiment	number of ER+ samples	survival data	5 year DFS TRUE/FALSE
A	GSE7390	130	125	93/32
B[b]	GSE2034	144	138	97/41
C	GSE21 653	122	91	64/27
D	GSE4922	211	196	58/138
	total	**607**	**550**	

[a] Breast cancer patients of all studies were untreated at time of sampling and three out of four studies (A−C) included only lymph-node-negative patients. Although these studies included both estrogen receptor (ER)-positive and ER-negative patients, only ER-positive patients for each study were used for our analysis. Patients were divided into two groups: patients with DFS lower (FALSE) and higher than 5 years (TRUE). [b] End point in the GSE2034 data set was distant metastasis-free survival (MFS).

The list of the 89 selected genes was uploaded and analyzed in the DAVID Functional Annotation Tool. This tool performs a gene ontology (GO)-term enrichment analysis to highlight the most relevant GO terms associated with a given gene list. The top enriched GO categories associated with the studied genes are summarized in Supplementary Table 2 in the Supporting Information and mainly include GO terms connected to cell proliferation such as M phase, cell division, and mitotic cell cycle. This finding underscores that high levels of tumor cell proliferation play a central role in breast cancer prognosis of ER-positive patients.

Selecting Potential Prognostic Biomarkers by Integrating Transcriptomic and Proteomic Information

The proteomic analysis of breast cancer tissues previously performed by our group[11] provided us with a comprehensive database of breast-cancer-related proteins that can be identified (and potentially be quantified) by mass spectrometry. Initially, we sought to examine whether the 89 genes that were identified to discriminate between good and poor prognosis breast cancer patients at the mRNA level were present in the breast cancer tissue proteome. Twenty out of 89 genes were identified by mass spectrometry in the breast cancer tissue proteome with at least two peptides: 14 related to poor prognosis and 6 related to favorable prognosis, summarized in Table 3.

Identification of Proteotypic Peptides for SRM Method Development

The final list of proteins for SRM method development included the 20 identified candidate biomarkers, two proteins (ABAT and PTX3) previously identified by our group as potential biomarkers,[11] two proteins (MARCSL1 and DDX1) previously reported in the literature as breast cancer prognostic markers[20,21] and the two biomarkers used in the clinic (ESR1 and ERBB2) − a total of 26 proteins.

The SRM method development was a multistep process with various rounds of optimization. In the first step, 97 peptides from 26 proteins and 580 transitions were monitored over the complete duration of the gradient (60 min) in a pool of breast cancer tissue samples. To ensure that at least 15 points were measured per LC peak, the scan time was set to 0.03 s, and no more than 60 transitions were included per method, resulting in 10 methods to be run. The coelution of all selected transitions (at least 6 in most cases) would indicate the presence of the peptide of interest. This process was repeated three times with three different pools of tissues, and the peptide yield out of this exercise is summarized in Supplementary Table 3 in the Supporting Information.

At the end of step one of method development, not all proteins were represented by at least one PTP, possibly due to sample complexity. For this reason, a more homogeneous and less complex system was selected − cancer cell lines. The cell-line proteome was searched to identify in which cell line(s) the proteins were identified with the greatest abundance. Two cell lines, SK-BR-3 and MDA-MB-231, were found to express almost all of the proteins in relatively high amounts (based on the spectral counts) and were selected for SRM method development.

Table 3. 20 Candidate Prognostic Biomarkers along with the *p* Value and the Coefficient of the End-Point Analysis in the Four Gene Expression Data Sets Used in the Present Study[a]

	experiment A		experiment B		experiment C		experiment D	
gene name	*p* value	coefficient	*p* value	coefficient	*p* value	coefficient	*p* value	coefficient
CDK1	0.02	−0.68	0.06	−0.42	<0.01	−1.19	<0.01	−0.77
CTTN	0.02	−0.60	0.22	−0.32	0.01	−1.01	<0.01	−0.67
CIAPIN1	0.39	−0.11	0.04	−0.38	0.04	−0.30	0.01	−0.35
FEN1	0.01	−0.40	0.07	−0.25	0.01	−0.66	,0.01	−0.35
HN1	0.02	−0.51	0.16	−0.26	0.01	−0.66	<0.01	−0.73
KPNA2	0.02	−0.43	0.01	−0.46	0.23	−0.26	<0.01	−0.37
LMNB1	0.02	−0.29	0.56	−0.09	<0.01	−0.92	0.01	−0.53
LRRC59	<0.01	−0.62	0.02	−0.42	0.01	−0.48	0.20	−0.17
MCM2	0.02	−0.42	0.24	−0.20	0.01	−0.60	<0.01	−0.56
NOL3	0.02	−0.34	<0.01	−0.58	0.01	−0.50	0.04	−0.25
PAICS	0.01	−0.38	0.15	−0.20	0.02	−0.44	0.02	−0.37
PNP	0.01	−0.54	0.01	−0.42	0.05	−0.44	0.04	−0.25
RRM2	0.05	−0.64	0.03	−0.68	<0.01	−1.34	<0.01	−1.20
TXNRD1	<0.01	−0.50	0.03	−0.31	0.28	−0.21	0.05	−0.21
CD74	0.06	0.38	0.02	0.36	0.04	0.51	<0.01	0.35
ALDH2	0.02	0.47	0.04	0.39	0.02	0.60	0.18	0.19
FAM129A	0.08	0.53	0.03	0.61	0.01	0.75	<0.01	0.63
HLA-DPA1	0.03	0.51	0.20	0.28	0.04	0.66	0.01	0.41
KCTD12	0.57	0.11	0.02	0.39	0.01	0.73	<0.01	0.54
SH3BGRL	0.02	0.48	0.20	0.22	0.03	0.50	0.01	0.27

[a] Difference in means of $\log_2$ expression between patients with DFS greater than 5 years (positive value) or lower than 5 years (negative value).

The use of cell lines offers another advantage to method development: the relatively cost-efficient generation of isotope-labeled peptides. Light- and heavy-labeled cells were lysed following our optimized protocol,[17] mixed in a total protein ratio of 1:1, and processed for trypsin digestion. The SRM methods were modified to monitor both light and heavy peptides. Given that chromatographic properties of heavy and light peptides are identical, both of them elute at the same time (and thus have the same RT). Also, $\mathrm{AUC}_{\mathrm{light}}/\mathrm{AUC}_{\mathrm{heavy}}$ should be ∼1. Additionally, the order of transition intensities should be the same for light and heavy peptides. However, because of the difference in the mass (+6 Da for arginine (R) and +8 Da for lysine (K)), the *m*/*z* of heavy peptides is shifted by six and eight units for peptides containing an arginine and lysine, respectively. A graphic representation of these properties can be seen in Figure 1A,B. Following this approach, we were able to confirm the identity of 10 peptides (corresponding to eight proteins) previously found in tissue samples by our SRM method. Additionally, we were able to detect 12 peptides (corresponding to 10 proteins) not previously observed in tissues (Supplementary Table 3 in the Supporting Information).

Notably, not all peptides identified by SRM in tissues were found also in the cell lines. An additional step to increase our confidence that the detected peaks in tissues correspond to the peptides of interest is predicting the RT and comparing it to the experimental RT. The RT of 25 peptides corresponding to 20 proteins was within the CI predicted by SSRCalc 3.0 (Supplementary Table 3 in the Supporting Information); also see Figure 1C. Notably, there were two peptides for which coelution of at least six transitions was observed, but the experimental RTs were different than the predicted ones, flagging these peptides as false-positives. These two peptides (corresponding to proteins CD74, CDK1) were excluded from further analysis.

For peptides confirmed in the two previous approaches, an extra confirmation step was performed that included the comparison of data obtained through previous shotgun proteomic experiments[11] and current SRM analyses. For this purpose, the MS/MS spectra obtained for the peptides of interest were retrieved, and the order of transition intensities was recorded and then compared with the order of transition intensities from the SRM approach. Because of the similar way of peptide fragmentation, the transition intensity order should be similar in the two experiments. A graphic representation of these fragmentation patterns is depicted in Figure 1D.

Through the procedure previously described, 46 peptides that corresponded to 21 proteins were detected. For proteins with multiple peptides, several criteria to select the final peptide(s) were applied. First, Universal Protein Resource (Uniprot) was utilized to investigate whether the selected peptides carry post-translational modifications (www.uniprot.org), which may affect *m*/*z* of the peptides and may differ among individuals, thus introducing variation in our analysis. The peptides that could possibly carry a modification were discarded. Second, among peptides from the same protein, the ones with highest signal intensities were preferred. Finally, peptides with minimally overlapping RTs (when possible) were selected for the final assay. The proteins and peptides of the final SRM method are summarized in Table 4.

Upon peptide selection, three transitions per peptide were retained for the final assay. SRM collider was used to identify UIS of the selected peptides, and the UIS that contained the most intense transitions was preferred. All peptides and corresponding transitions were scheduled in a single multiplex scheduled SRM method within 5 min (±2.5 min) intervals during a 60 min LC gradient. Scan times were optimized for each peptide in the final SRM method to ensure the measurement of 15−20 points per LC peak per transition.

During the SRM development phase, 60 min LC gradients were used. In the interest of reducing machine run time, the

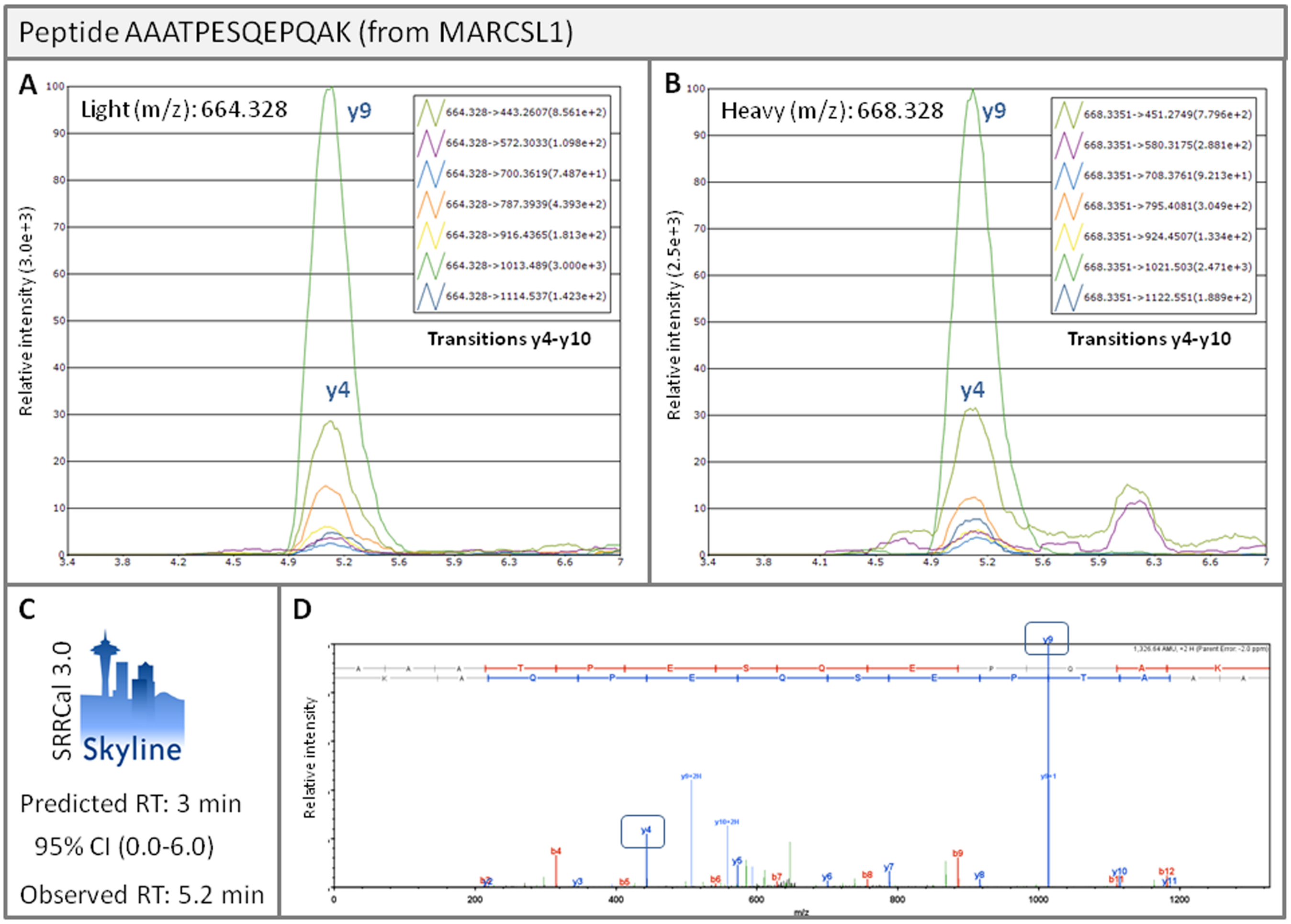


Figure 1. Identification of proteotypic peptides for SRM method development: a representative example. Co-elution of seven transitions originating from the endogenous (A) and the spiked-in isotope-labeled (B) peptide. Both "light" and "heavy" transitions elute at the same RT and with the same order of transition intensities. The relative intensity of "light" and "heavy" is almost equal. (C) Predicted retention time (RT) and 95% confidence interval (CI) for peptide AAATPESQEPQAK, according to SRRCalc. 3.0, and observed RT in a 60 min gradient. (D) MS/MS spectrum of the doubly charged peptide AAATPESQEPQAK (m/z = 664.3), acquired in an LTQ Orbitrap XL, and identification of b and y ions in Scaffold software.

method was modified to a 30 min gradient. Comparison of signal intensities for peptides monitored in the 30 versus the 60 min method (Supplementary Table 4 in the Supporting Information) did not reveal considerable differences. At the same time, sample analysis time was reduced to 57 min compared with 90 min previously. Therefore, the final SRM method was run in a 30 min gradient.

Optimization of the Amount of Spiked-in Isotope-Labeled Peptides

Upon selection of the final peptides to be included in the assay, the corresponding isotopically labeled peptides in unpurified form were purchased. An estimated concentration of 25 nM was provided by the manufacturer and was used for further calculations. On the basis of these estimated concentrations and using a pool of breast cancer cytosols as matrix, serial dilutions of the "heavy" peptides covering three orders of magnitude (2−1000 fmoles per injection) were prepared and analyzed with the developed SRM method. As expected, all spiked-in isotope-labeled peptides were detected. However, not all the peptides were present in the matrix. The SRM assays (for the 14 detected proteins) showed good linearity (coefficient of determination, R^2 > 0.99 for all proteins except ABAT that exhibited R^2 = 0.985) in the entire concentration range (2−1000 fmol/injection) (Supplementary Table 5 in the Supporting Information).

The purpose of adding isotope-labeled peptides in the samples was to control for variations during sample preparation and mass spectrometric analysis of samples. These peptides were used for normalization of signal intensities. It is preferable that the spiked-in amounts of "heavy" peptides are close to the levels of the endogenous counterparts. This exercise allowed us to determine what amount of heavy peptides should be added to the samples to obtain a heavy-to-light ratio close to one. On the basis of the relative abundances of heavy and light peptides, the optimum amount of heavy peptides was determined. For peptides that were not present in the matrix, the relative abundance compared with other peptides was taken into consideration. The optimum amount for each peptide is summarized in Supplementary Table 6 in the Supporting Information.

Calibration Curves and Limit of Quantification

Calibration curves with 13 points ranging from 0.06 to 250 fmol/injection were generated to define the LOQ of the SRM assays. Analysis of the second to last point (125fmol/injection) failed due to technical problems; therefore, that point was excluded from further analysis. Additionally, the endogenous

Table 4. Proteins, Peptides, and Transitions of the Developed SRM Method

protein	peptide	peptide m/z	transition m/z	ion type
ABAT	IDIPSFDWPIAPFPR	885.964	1098.573	y9
			983.546	y8
			797.467	y7
ALDH2	ANNSTYGLAAAVFTK	764.394	1040.578	y10
			877.514	y9
			395.229	y3
ALDH2	ELGEYGLQAYTEVK	800.399	1171.599	y10
			1008.536	y9
			710.372	y6
CDK1	SPEVLLGSAR	514.790	844.489	y8
			503.294	y5
			390.210	y4
CTTN	SAVGFDYQGK	536.259	814.373	y7
			757.352	y6
			610.283	y5
DDX1	ELAEQTLNNIK	636.843	959.516	y8
			830.473	y7
			488.283	y4
ESR1	YLENEPSGYTVR	714.343	1151.532	y10
			1022.490	y9
			779.404	y7
FAM129A	VLTSEDEYNLLSDR	827.402	1124.522	y9
			717.389	y6
			377.178	y3
FEN1	LIADVAPSAIR	563.335	713.430	y7
			614.362	y6
			543.325	y5
KCTD12	SGYITIGYR	515.272	609.335	y5
			395.204	y3
			338.182	y2
KPNA2	ASLSLIEK	430.758	789.472	y7
			589.356	y5
			389.239	y3
LMNB1	IQELEDLLAK	586.332	930.514	y8
			801.472	y7
			688.388	y6
MARCKSL1	AAATPESQEPQAK	664.328	1013.490	y9
			787.394	y7
			443.261	y4
MCM2	VAVGELTDEDVK	637.827	1005.473	y9
			706.325	y6
			490.251	y4
NOL3	LLLLVQGK	442.302	657.429	y6
			544.345	y5
			431.261	y4
PAICS	EVYELLDSPGK	625.319	1021.520	y9
			858.457	y8
			503.246	y5
PNP	ANHEEVLAAGK	569.796	687.404	y7
			558.361	y6
			459.293	y5
PTX3	LTSALDELLQATR	715.896	945.500	y8
			588.346	y5
			475.262	y4
RRM2	IEQEFLTEALPVK	758.917	870.529	y8
			656.398	y6
			343.234	y3
SH3BGRL	GDYDAFFEAR	595.759	740.373	y6
			669.335	y5
			522.267	y4
TXNRD1	IGLETVGVK	458.279	802.467	y8
			745.445	y7
			632.361	y6

levels of FEN1 were very low, and thus the generation of a calibration curve was not possible. The SRM assays showed good linearity (R^2 > 0.99) in both the entire concentration range, from 250 fmoles to the LOQ, and in the lower range (five lower concentration standards). The LOQ, defined as the concentration for which monitored transitions could be clearly detected, was within the linear range and showed a CV lower or equal to 20%. Table 5 summarizes the LOQs for the SRM assays.

Measurement of Relative Amounts of 20 Proteins in 96 Breast Cancer Tissue Samples

Using the developed SRM assays, the relative amounts of 20 proteins in 96 breast cancer tissues were measured. All samples were analyzed in duplicate; however, 10 out of 192 injections failed due to technical problems. It the end, 86 samples were assayed in duplicate and 10 samples in singleton.

As expected, all spiked-in isotope-labeled peptides were detected. Native peptides from proteins PNP, RRM2, NOL3, DDX1, and TXNRD1 were not detected in the majority of samples, indicating low abundance of these proteins in this set of samples. It is worth mentioning that even in cases where the transitions of monitored peptides were detected the signal-to-noise ratio was low, so they were not taken into further consideration. The native peptide of ESR1 was detected but with low sensitivity, allowing only the qualitative and not quantitative assessment of ER (presence or absence). All of the other native peptides were clearly detected. The majority of $\mathrm{AUC}_{\mathrm{light}}/\mathrm{AUC}_{\mathrm{heavy}}$ for all peptides was between 0.1 and 1, indicating that the spiked-in amount of isotope-labeled peptides was optimized. For samples analyzed in duplicate, the average CV was 4%, ranging from 1 to 26%. The estimated amount of each native peptide along with the CVs is depicted in Supplementary File 1 in the Supporting Information.

Given that the SRM assay included two peptides for ALDH2, we sought to examine the extent of their correlation. The signal intensities for both peptides were similar (10^5) except for eight samples that the second peptide showed a 10-fold decrease in signal intensity. As can be seen in Figure 2, the relative amounts of the two peptides were significantly correlated (R^2 = 0.87). Notably, only one of the two peptides had the isotope-labeled counterpart; therefore, raw AUC values were compared without any normalization.

Association of PTX3 and ABAT with ER Status

In a previous study by our group, protein ABAT was identified as specific to ER-positive breast cancer tumors, whereas PTX3 was proposed as an ER-negative subtype-specific protein.[11] In the present study, the relative amounts of ABAT and PTX3 in a set of samples that contained both ER-positive and ER-negative cases were measured to investigate whether this finding could be independently reproduced. Indeed, PTX3 was shown to be significantly associated (Mann−Whitney U-test, p < 0.0001) to ER-negative samples (Figure 3A). ABAT was found to be

Table 5. Analytical Characteristics for the Quantification of 14 Endogenous Peptides from 14 Proteins[a]

protein	peptide	min/max	LOQ	CV	analytical range	R^2
ABAT	IDIPSFDWPIAPFPR	0.8−52	0.49	1%	250−0.49	0.999
					62.5−0.49	0.998
ALDH2	ANNSTYGLAAAVFTK	1−52	0.98	20%	250−0.98	0.994
					62.5−0.98	0.993
CDK1	SPEVLLGSAR	0.1−13	0.06	8%	250−0.06	0.998
					15.6−0.06	0.995
CTTN	SAVGFDYQGK	0.6-31	0.49	10%	250−0.49	0.996
					31.25−0.49	0.998
FAM129A	VLTSEDEYNLLSDR	0.9−37	0.49	2%	250−0.49	0.997
					31.25−0.49	0.989
KCTD12	SGYITIGYR	3−28	1.95	5%	250−1.95	0.996
					31.25−1.95	0.994
KPNA2	ASLSLIEK	0.4−18	0.24	8%	250−0.24	0.996
					31.25−0.24	0.998
LMNB1	IQELEDLLAK	0.1−4	0.12	13%	250−0.12	0.996
					7.81−0.12	0.982
MARCSL1	AAATPESQEPQAK	2.6−101	0.49	10%	250−0.49	0.993
MCM2	VAVGELTDEDVK	0.2−27	0.24	5%	250−0.24	0.996
					31−0.24	0.998
MCM6	ESEDFIVEQYK	0.08−8	0.06	20%	250−0.06	0.999
					7.8−0.06	0.995
PAICS	EVYELLDSPGK	1.3−26	0.49	7%	250−0.49	0.998
					31.25−0.49	0.998
PTX3	LTSALDELLQATR	0.1−25	0.06	16%	250−0.06	0.998
					31.25−0.06	0.994
SH3BGRL	GDYDAFFEAR	3−185	0.12	14%	250−0.12	0.998

[a] min/max, the minimum and maximum amount of the corresponding protein quantified in the 96 clinical samples; LOQ, limit of quantification (fmoles per injection); CV, coefficient of variation (triplicates) at LOQ; R^2, coefficient of determination.

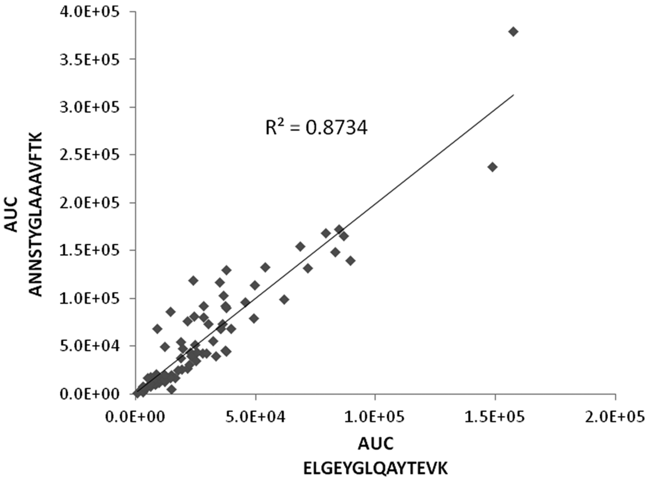


Figure 2. Correlation of the relative amounts (shown as AUC) of two peptides (ANNSTYGLAAAVFTK and ELGEYGLQAYTEVK) originating from the same protein (ALDH2).

significantly (Mann−Whitney U-test, $p < 0.0001$) associated with ER-positive samples (Figure 3B). These results provide strong confirmation of our previous study.

Given that this study was performed in a blinded fashion, we sought to investigate whether we could predict the ER status of the samples based on the ABAT and ESR1 levels measured by SRM. The ER status was correctly assigned in 80 out of 96 assayed samples, a result highly significant as shown by a chi-square test ($p < 0.0001$).

Association of Candidate Biomarker Levels with Clinical Outcome

Analysis of results was performed independently for ER-positive and ER-negative patients. In ER-negative patients, the average levels of all proteins between good and poor prognosis patients did not show a significant difference. This finding was expected given that the candidate biomarkers were chosen to have prognostic potential in ER-positive patients. In the subgroup of ER-positive patients, the majority of proteins were not found to be differentially expressed between patients with good and poor prognosis. However, two proteins, KPNA2 and CDK1, were found to be overexpressed (Mann−Whitney U-test, $p < 0.05$) in the poor prognosis patient group by approximately two-fold (Figure 4). The direction of effect was in accordance with our proposal that CDK1 and KPNA2 are biomarkers of poor prognosis. Notably, the levels of those two proteins did not differ significantly in the subgroup of ER-negative patients (Supplementary Figure 1 in the Supporting Information).

■ DISCUSSION

In the present study, we integrated transcriptome- and proteome-based platforms for identifying potential prognostic biomarkers for stratification of ER-positive breast cancer patients into groups of low and high risk for disease recurrence. The selection of our candidate prognostic biomarkers was a two-step process. First, we identified genes that are related to 5-year DFS of ER-positive patients by performing meta-analysis of gene expression profiling data from four independent studies. Then, we compared the generated gene list to a previously generated breast cancer tissue proteome and selected only

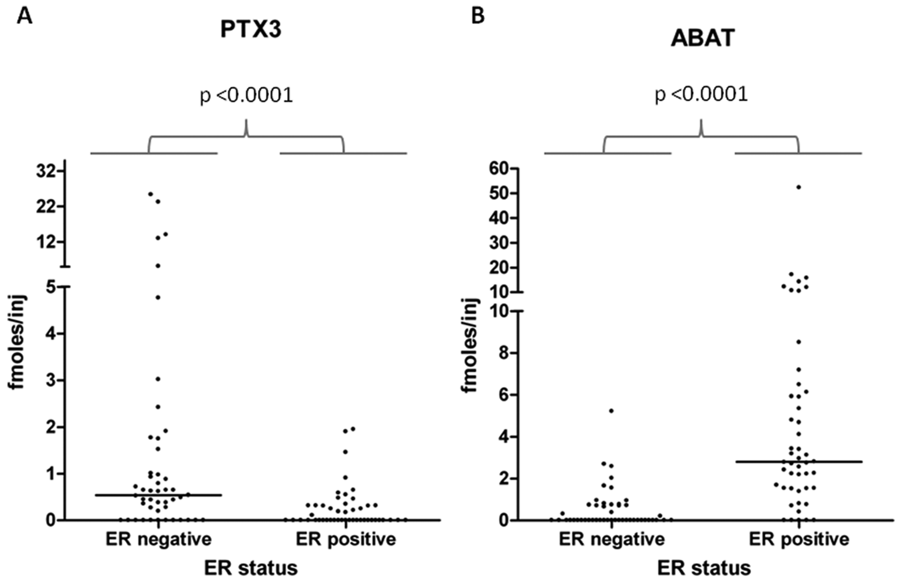


Figure 3. Association of PTX3 (A) and ABAT (B) expression levels with ER status. Lines define median levels of each protein. Median levels of PTX3 in ER-negative and ER-positive patients were 0.54 and 0 fmoles per injection, respectively. Median levels of ABAT in ER-negative and ER-positive patients were 0 and 2.8 fmoles per injection, respectively. The statistical analysis (Mann−Whitney U test) was performed by comparing the samples from ER-negative patients (n = 48) versus the samples from ER-positive patients (n = 48).

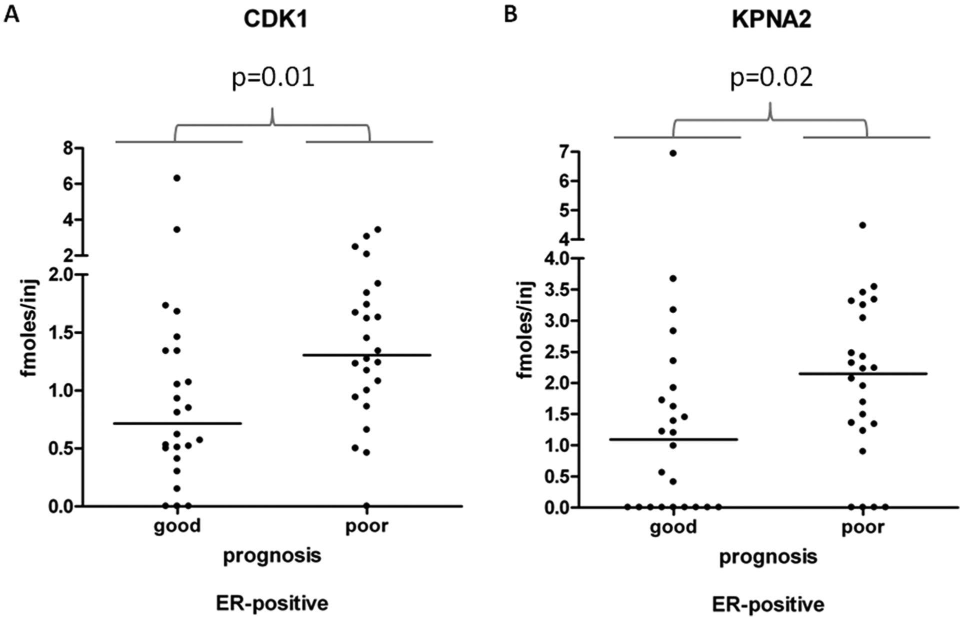


Figure 4. Association of CDK1 and KPNA2 expression levels with clinical outcome in ER-positive patients. CDK1 (A) and KPNA2 (B) expression levels were found increased in ER-positive patients with poor prognosis compared with ER-positive patients with favorable prognosis. The lines in scatter plots define median levels of each protein. Median levels of CDK1 in poor and good prognosis ER-positive patients were 1.3 and 0.7 fmoles per injection, respectively. Median levels of KPNA2 in poor and good prognosis ER-positive patients were 2.2 and 1.1 fmoles per injection, respectively. The statistical analysis (Mann−Whitney U test) was performed by comparing the samples from ER-positive patients with poor (n = 24) versus good prognosis (n = 24).

genes that have been identified at the protein level. Following candidate identification, we developed a multiplex mass-spectrometry-based assay for the simultaneous quantification of the 26 selected proteins in breast cancer tissues. We were able to develop a single multiplex SRM assay for the quantification of 21 peptides (corresponding to 20 proteins) in breast cancer tissues. Finally, the prognostic potential of the candidate biomarkers for which the SRM method was successfully developed was investigated. The relative amounts of 20 proteins were measured by SRM in a cohort of 96 samples from patients with early-stage primary breast cancer. Two proteins, KPNA2 and CDK1, were found to discriminate between patients with favorable and poor prognosis.

SRM has recently emerged as a promising technology for mass spectrometry-based quantification of targeted proteins in clinical specimens. A drawback of current SRM technology is relatively low sensitivity. Without sample prefractionation, SRM measurements have been limited to only moderately abundant proteins in human plasma, present in the low microgram per milliliter range.[22] This limitation may explain why we were not able to develop an SRM method for 6 out of the 26 candidate proteins. Some approaches to increase assay sensitivity include sample fractionation, depletion of high abundance proteins, or affinity purification of low abundant proteins using antibodies (immuno-SRM) or aptamers. Additionally, several efforts to circumvent sensitivity limitations have been undertaken recently. The online sample fractionation using cation exchange chromatography,[23] isoelectric focusing,[24] and high-resolution reverse-phase chromatography has been described and demonstrated sensitivities of pg/mL.[25] However, these experimental setups significantly decrease method throughput and increase development time. In an attempt to increase method throughput, Whiteaker et al. demonstrated the feasibility of multiplexing immuno-SRM assays for the simultaneous quantification of up to 47 peptides.[26] Nevertheless, good-quality purification reagents are of paramount significance. In general, these strategies decrease the throughput and increase variability due to extensive sample handling, and in the case of immuno-SRM, a high-quality antibody is required. For these reasons, in the present study, we decided to employ minimum sample manipulation and focus on medium to high abundance proteins. Some reasons for our inability to identify the protein counterparts of 69 out of 89 genes could include low protein abundance, protein degradation during sample preparation, or protein insolubility.

In an SRM experiment, the proteotypic peptides act as surrogates for the quantification of the corresponding protein. It is recommended that two peptides per protein and at least three transitions per peptide should be monitored;[2] however, this may not be always feasible. First, it should be noted that not all possible peptides originating from a protein can be detected by the mass spectrometer due to factors such as poor ionization efficiency. Additionally, a set of heuristics exists regarding peptide selection for SRM. In general, peptide length should be between 8 and 20 amino acids, accommodating the m/z range of the quadrupole analyzer (typically 50 to 1500 m/z). Also, peptides shorter than eight amino acids (very small m/z) are prone to suffer from more interferences. For sensitive analysis, it is important to monitor the predominant charge state of a peptide. Because of our focus on doubly charged peptides, peptides containing histidine (positively charged amino acid) in the middle of the sequence were avoided, if possible. Furthermore, peptides with N-terminus glutamine, cysteine, or asparagine should be excluded because these residues are susceptible to chemical modifications that will alter the peptide mass. Ideally, peptides should not contain any post-translational modifications that again will affect the mass of the peptide and thus the m/z. Given that usually only part of native peptides will be modified, there is no straightforward way to calculate the total amount of all modified forms. When all of these rules of the thumb are taken into consideration the number of candidate peptides is reduced. It has been previously reported that the failure rate of omitting peptides during peptide selection was close to 75%.[27] In the present study, we were able to identify two (or more) peptides for approximately half of the studied proteins. However, the final method contains one peptide per protein (except for ALDH2) to maximize assay sensitivity. In the case of ALDH2, five eligible peptides were identified, and the top two performing were included in the final assay.

An integral part of our approach was the utilization of publicly available microarray data for identification of genes related to 5-year DFS. Following the publication of the first studies describing gene profiling of breast cancer tissues for the development of prognostic signatures, numerous studies have reported the meta-analysis of publicly available gene expression data in the quest of novel multigene classifiers.[28−35] One of the challenges in building classification models is overfitting, which results in nonreproducible findings.[36] Overfitting occurs when multivariable models demonstrate discrimination between two conditions by chance, and a model is prone to overfitting when the number of parameters tested is large and the number of samples is small. To overcome this challenge, and unlike previous studies, we identified individual genes that are associated with survival by calculating the differential expression between patients with poor or good prognosis, at the gene-by-gene level, in four independent patient cohorts. It should be noted that the OncotypeDX gene selection model[37] was developed based on 447 tissue samples (fewer than our 607 samples) and included both ER positive and negative tumors. Paik et al.[37] also used a less strict criterion for gene selection ($p < 0.1$ in three out of three studies or $p < 0.05$ in two out of three studies). Our premise is that by combining genes that individually show prognostic potential, we could develop a powerful multiparametric prognostic signature. Our approach should be less prone to overfitting bias, given that we use multiple independent data sets and we refrain from applying multiple testing.

Assessing the prognostic utility of a candidate biomarker requires careful selection of samples to be included in the study. First, the samples should originate from patients that have been monitored for an extended period of time after their disease diagnosis. Second, given that a prognostic marker should provide information about the natural history of the disease independent of a specific therapy, the patients should not receive any type of adjuvant therapy that will affect the disease course. The cohort analyzed in the present study included tissue specimens from patients with early stage (lymph-node-negative) disease who, at the time of diagnosis, were considered as low-risk for disease recurrence and were treated by local treatment only (surgical removal of the tumor, with or without radiotherapy). This very strictly selected set of pilot samples allows for the investigation of the true prognostic potential of the candidate biomarkers free of potential confounding effects of systemic therapy. Given that in our analysis we focused only on ER-positive patients, we would expect the candidate markers not to show prognostic utility in ER-negative patients. The inclusion of ER-negative patients in our cohort confirmed our original assumption. It should be noted that this cohort was selected for this preliminary study; an independent study with larger number of samples should be performed to verify the findings.

The panel of candidate prognostic biomarkers contained both proteins that have been previously connected to breast cancer (including prognosis) and others that (to our knowledge) have not been studied in the context of breast cancer before. Proteins such as cyclin-dependent kinase 1 (CDK1), karyopherin 2 (KPNA2), and minichromosome maintenance protein 2 (MCM2) are involved in cell proliferation, a tumor characteristic that is tightly connected to prognosis. Cortactin (CTTN) has been the focus of numerous studies in breast and other cancer types. The gene encoding cortactin is located in the 11q13 region that is amplified in up to 15% breast cancer cases and produces a cytoplasmic protein that is a key regulator of actin polymerization. Because of its role in actin polymerization, CTTN has been shown to play a critical role in various actin-mediated processes such as cell invasion and migration, adhesion, and receptor-mediated endocytosis.[38] In breast cancer, overexpression of CTTN has been reported in tumors with and without 11q13 amplification, and its role in tumor invasion and metastasis has been documented.[39,40] Thus, the potential of CTTN as a prognostic biomarker in breast cancer warrants further investigation. CD antigen CD74 has also been previously studied in the context of breast cancer and was found to be associated with increased invasion and metastasis.[41,42] However, both studies report an association of CD74 with triple-negative phenotype, and hence the prognostic value in ER-positive disease remains to be investigated.

Flap endonuclease 1 (FEN1), a nuclease known for its critical roles in Okazaki fragment maturation, DNA repair, and apoptosis-induced DNA fragmentation; thioredoxin reductase 1 (TXNRD1), a key player in oxidative stress control; and nucleolar protein 3 (NOL3), an antiapoptotic protein have been found to be overexpressed in breast cancer in previous studies.[43−45] Proteins FAM129A, SH3BGRL, ALDH2, LMNB1, KCTD12, and PAICS have not been connected to breast cancer previously. Interestingly, high levels of potassium channel tetramerization domain-containing 12 (KCTD12) have been associated with higher percentage of 5-year recurrence-free survival rate in patients with gastrointestinal stromal tumors.[46] Notably, in our approach, KCTD12 was identified as a candidate biomarker of favorable prognosis.

The majority of investigated proteins did not show potential in discriminating between patients with different prognosis. Although mRNA data strongly supported the prognostic potential of those candidates, this was not mirrored in our verification study at the protein level. Possibly, the prognostic utility of these genes could not be observed at the protein level due to protein instability, high turnover, or degradation. Another possible explanation may be related to the microarray data sets used in our analysis. Although the majority of the patients in all four studies were lymph-node-negative, they probably received a variety of adjuvant therapies that may have altered the natural course of disease. Finally, it could be possible that these markers do not show prognostic potential individually but they may perform better in a panel. However, this multiparametric approach will require analysis of significantly larger number of samples to avoid overfitting.

Recently, criticism over the use of the cutoff value of 0.05 for significance (similar to what was used in the present study) has emerged.[47,48] However, apart from the p value (which implies statistical significance), certain findings of this study demonstrate biological significance. As discussed in detail, many proteins identified as candidate markers have been connected to breast cancer (including prognosis) previously. This biologic background supports our notion to further investigate the prognostic potential of these proteins. Additionally, none of the proteins investigated in the present study showed any significant association with prognosis in the ER-negative subgroup of patients. Given that in our initial selection we focused only in ER-positive patients, this finding was anticipated. The absence of statistical significant results (even using p value <0.05) in the ER-negative patient group is encouraging. Nevertheless, this is a preliminary study, and an independent study with larger number of samples should be performed to verify the findings.

Two proteins previously reported as prognostic biomarkers for breast cancer, DDX1 and MARCSL1, were included in our verification. The protein levels of DDX1 were assessed by immunohistochemistry in a study of 113 tumor samples, and cytoplasmic localization of DDX1 was found to correlate with increased risk of recurrence in breast cancer, independently of other prognostic markers such as ER and grade.[21] In our study, although we were able to develop an SRM method for DDX1, the protein levels in the cohort of samples analyzed were below the level of detection. Levels of MARCSL1 were successfully measured in our cohort. Jonsdottir et al. evaluated the expression of this protein by immunohistochemistry in a cohort of 305 operable lymph-node-negative breast cancer patients. High expression of MARCKSL1 was correlated with an increased risk for metastasis and a worse prognosis.[20] However, this association was not observed in our study. This could be attributed to the definition of “high” MARCSL1 expression in the study by Jonsdottir et al. The authors, using the optimal cutoff value from the ROC-analysis to stratify patients in good or bad prognosis, identified 28 out of 305 patients with high levels of MARCSL1 (IHC score >7) and significantly worse outcome. In the present study, comparisons were performed based on the median values.

Two proteins were found to be overexpressed in ER-positive patients with poor prognosis when compared with favorable prognosis patients: karyopherin alpha 2 (KPNA2) and CDK1. The karyopherin family includes more than 20 members that participate in several nuclear transport pathways into and out of the nucleus. Nuclear import of proteins via the classical pathway is mediated by heterodimers of members from the karyopherin beta and karyopherin alpha families. KPNA2 is one of seven described members of the karyopherin alpha family. KPNA2 is highly expressed in multiple cancer types, and its aberrant expression is often associated with adverse patient outcomes. The first connection between breast cancer and KPNA2 was provided by Dahl et al. by performing gene expression profiling of laser-microdissected cancer and corresponding benign breast tissues.[49] The authors found that KPNA2 mRNA levels were up-regulated (fold change >2) in 32% of analyzed tumor/normal pairs. Additionally, immunohistochemical assessment of KPNA2 in a cohort of 272 breast cancer patients showed negative correlation between KPNA2 expression in the primary tumor and overall survival in lymph-node-positive but not node-negative patients. The same group went on to evaluate KPNA2 expression in invasive breast cancer and matched ductal carcinoma in situ in 83 clinicopathologically characterized cases.[50] Nuclear KPNA2 staining was significantly correlated with higher tumor stage, grade, and lymph node status. Consistent with their previous results, survival analysis revealed that patients with KPNA2-positive invasive breast carcinomas had significantly shorter DFS. Notably, the authors report an association between KPNA2 expression and ER-negative disease. Additionally, in an independent study, KPNA2 was shown to predict poor survival in patients with advanced (lymph node-positive) breast cancer.[51] We are the first to report the prognostic potential of KPNA2 in early stage breast cancer and particularly in the subset of ER-positive patients.

■ CONCLUSIONS

In summary, by integrating transcriptomic and proteomic information, we identified 20 proteins as potential prognostic biomarkers in the subset of ER-positive breast cancer patients. We were able to develop an SRM method for monitoring simultaneously the relative levels of 20 candidate biomarkers in breast cancer tissues. The prognostic potential of the candidate biomarkers was preliminarily investigated in a cohort of 96 breast cancer patients with primary early-stage disease. Two proteins were identified that show potential to discriminate between ER-positive patients of high and low risk of disease recurrence. The role of these proteins in breast cancer prognosis warrants further investigation.

■ ASSOCIATED CONTENT

Supporting Information

Association of CDK1 and KPNA2 expression levels with clinical outcome in ER-negative patients. Summary of the 89 genes identified in the present study along with the *p* value and the coefficient of the endpoint analysis in the four independent experiments. Gene ontology (GO)-term enrichment analysis of the 89 genes identified in the present study as potential prognostic biomarkers. Proteotypic peptides identified during the SRM method development. Signal intensities for peptides monitored in the 30 versus the 60 min SRM method. Coefficients of determination for 14 isotope-labelled peptides monitored over three orders of magnitude (2-1000 fmoles/injection). Optimum amount of spiked-in isotope-labelled peptides. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: 416-586-8443. Fax: 416-619-5521. E-mail: ediamandis@mtsinai.on.ca.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Rakha, E. A.; Reis-Filho, J. S.; Ellis, I. O. Combinatorial biomarker expression in breast cancer. *Breast Cancer Res. Treat.* **2010**, *120* (2), 293−308.

(2) Lange, V.; Picotti, P.; Domon, B.; Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **2008**, *4*, 222.

(3) Jemal, A.; Bray, F.; Center, M. M.; Ferlay, J.; Ward, E.; Forman, D. Global cancer statistics. *Ca-Cancer J. Clin* **2011**, *61* (2), 69−90.

(4) Reis-Filho, J. S.; Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* **2011**, *378* (9805), 1812−1823.

(5) Edgar, R.; Domrachev, M.; Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30* (1), 207−210.

(6) Parkinson, H.; Sarkans, U.; Kolesnikov, N.; et al. ArrayExpress update−an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* **2011**, *39* (Database issue), D1002−D1004.

(7) Pavlou, M. P.; Diamandis, E. P.; Blasutig, I. M. The long journey of cancer biomarkers from the bench to the clinic. *Clin. Chem.* **2013**, *59* (1), 147−157.

(8) Banks, R. E.; Dunn, M. J.; Hochstrasser, D. F.; et al. Proteomics: new perspectives, new biomedical opportunities. *Lancet* **2000**, *356* (9243), 1749−1756.

(9) Umar, A.; Kang, H.; Timmermans, A. M.; et al. Identification of a putative protein profile associated with tamoxifen therapy resistance in breast cancer. *Mol. Cell. Proteomics* **2009**, *8* (6), 1278−1294.

(10) Liu, N. Q.; Braakman, R. B.; Stingl, C.; et al. Proteomics pipeline for biomarker discovery of laser capture microdissected breast cancer tissue. *J. Mammary Gland Biol. Neoplasia* **2012**, *17* (2), 155−164.

(11) Pavlou, M. P.; Dimitromanolakis, A.; Diamandis, E. P. Coupling proteomics and transcriptomics in the quest of subtype-specific proteins in breast cancer. *Proteomics* **2013**, *13* (7), 1083−1095.

(12) Desmedt, C.; Piette, F.; Loi, S.; et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* **2007**, *13* (11), 3207−3214.

(13) Wang, Y.; Klijn, J. G.; Zhang, Y.; et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **2005**, *365* (9460), 671−679.

(14) Sabatier, R.; Finetti, P.; Cervera, N.; et al. A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res. Treat.* **2011**, *126* (2), 407−420.

(15) Ivshina, A. V.; George, J.; Senko, O.; et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* **2006**, *66* (21), 10292−10301.

(16) Luo, L. Y.; Diamandis, E. P.; Look, M. P.; Soosaipillai, A. P.; Foekens, J. A. Higher expression of human kallikrein 10 in breast cancer tissue predicts tamoxifen resistance. *Br. J. Cancer* **2002**, *86* (11), 1790−1796.

(17) Drabovich, A. P.; Pavlou, M. P.; Dimitromanolakis, A.; Diamandis, E. P. Quantitative analysis of energy metabolic pathways in MCF-7 breast cancer cells by selected reaction monitoring assay. *Mol. Cell. Proteomics* **2012**, *11*, 422−434.

(18) MacLean, B.; Tomazela, D. M.; Shulman, N.; et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010**, *26* (7), 966−968.

(19) Liu, N. Q.; Dekker, L. J.; Stingl, C.; et al. Quantitative proteomic analysis of microdissected breast cancer tissues: comparison of label-free and SILAC-based quantification with shotgun, directed, and targeted MS approaches. *J. Proteome Res.* **2013**, *12* (10), 4627−4641.

(20) Jonsdottir, K.; Zhang, H.; Jhagroe, D.; et al. The prognostic value of MARCKS-like 1 in lymph node-negative breast cancer. *Breast Cancer Res. Treat.* **2012**, *135* (2), 381−390.

(21) Germain, D. R.; Graham, K.; Glubrecht, D. D.; Hugh, J. C.; Mackey, J. R.; Godbout, R. DEAD box 1: a novel and independent prognostic marker for early recurrence in breast cancer. *Breast Cancer Res. Treat.* **2011**, *127* (1), 53−63.

(22) Kuzyk, M. A.; Smith, D.; Yang, J.; et al. Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. *Mol. Cell. Proteomics* **2009**, *8* (8), 1860−1877.

(23) Krisp, C.; McKay, M. J.; Wolters, D. A.; Molloy, M. P. Multidimensional protein identification technology-selected reaction monitoring improving detection and quantification for protein biomarker studies. *Anal. Chem.* **2012**, *84* (3), 1592−1600.

(24) Rafalko, A.; Dai, S.; Hancock, W. S.; Karger, B. L.; Hincapie, M. Development of a Chip/Chip/SRM platform using digital chip isoelectric focusing and LC-Chip mass spectrometry for enrichment and quantitation of low abundance protein biomarkers in human plasma. *J. Proteome Res.* **2012**, *11* (2), 808−817.

(25) Shi, T.; Fillmore, T. L.; Sun, X.; et al. Antibody-free, targeted mass-spectrometric approach for quantification of proteins at low picogram per milliliter levels in human plasma/serum. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (38), 15395−15400.

(26) Whiteaker, J. R.; Zhao, L.; Zhang, H. Y.; et al. Antibody-based enrichment of peptides on magnetic beads for mass-spectrometry-based quantification of serum biomarkers. *Anal. Biochem.* **2007**, *362* (1), 44−54.

(27) Mortstedt, H.; Karedal, M. H.; Jonsson, B. A.; Lindh, C. H. Screening method using selected reaction monitoring for targeted proteomics studies of nasal lavage fluid. *J. Proteome Res.* **2013**, *12* (1), 234−247.

(28) Ma, S.; Kosorok, M. R. Detection of gene pathways with predictive power for breast cancer prognosis. *BMC Bioinf.* **2010**, *11*, 1.

(29) Yau, C.; Esserman, L.; Moore, D. H.; Waldman, F.; Sninsky, J.; Benz, C. C. A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. *Breast Cancer Res.* **2010**, *12* (5), R85.

(30) Molloy, T. J.; Roepman, P.; Naume, B.; van't Veer, L. J. A prognostic gene expression profile that predicts circulating tumor cell presence in breast cancer patients. *PLoS One* **2012**, *7* (2), e32426.

(31) Rody, A.; Karn, T.; Liedtke, C.; et al. A clinically relevant gene signature in triple negative and basal-like breast cancer. *Breast Cancer Res.* **2011**, *13* (5), R97.

(32) Shen, R.; Ghosh, D.; Chinnaiyan, A. M. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics* **2004**, *5* (1), 94.

(33) Miecznikowski, J. C.; Wang, D.; Liu, S.; Sucheston, L.; Gold, D. Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways. *BMC Cancer* **2010**, *10*, 573.

(34) Wirapati, P.; Sotiriou, C.; Kunkel, S.; et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.* **2008**, *10* (4), R65.

(35) Li, J.; Lenferink, A. E.; Deng, Y.; et al. Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat. Commun.* **2010**, *1*, 34.

(36) Ransohoff, D. F. Rules of evidence for cancer molecular-marker discovery and validation. *Nat. Rev. Cancer* **2004**, *4* (4), 309−314.

(37) Paik, S.; Shak, S.; Tang, G.; et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **2004**, *351* (27), 2817−2826.

(38) Kirkbride, K. C.; Sung, B. H.; Sinha, S.; Weaver, A. M. Cortactin: a multifunctional regulator of cellular invasiveness. *Cell Adhes. Migr.* **2011**, *5* (2), 187−198.

(39) Hui, S.; Choi, J.; Zaidi, S.; et al. Peptide-mediated disruption of calmodulin-cyclin E interactions inhibits proliferation of vascular smooth muscle cells and neointima formation. *Circ. Res.* **2011**, *108* (9), 1053−1062.

(40) Lundgren, K.; Holm, K.; Nordenskjold, B.; Borg, A.; Landberg, G. Gene products of chromosome 11q and their association with CCND1 gene amplification and tamoxifen resistance in premenopausal breast cancer. *Breast Cancer Res.* **2008**, *10* (5), R81.

(41) Greenwood, C.; Metodieva, G.; Al-Janabi, K.; et al. Stat1 and CD74 overexpression is co-dependent and linked to increased invasion and lymph node metastasis in triple-negative breast cancer. *J. Proteomics* **2012**, *75* (10), 3031−3040.

(42) Leth-Larsen, R.; Lund, R.; Hansen, H. V.; et al. Metastasis-related plasma membrane proteins of human breast cancer cells identified by comparative quantitative mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (6), 1436−1449.

(43) Medina-Ramirez, C. M.; Goswami, S.; Smirnova, T.; et al. Apoptosis inhibitor ARC promotes breast tumorigenesis, metastasis, and chemoresistance. *Cancer Res.* **2011**, *71* (24), 7705−7715.

(44) Singh, P.; Yang, M.; Dai, H.; et al. Overexpression and hypomethylation of flap endonuclease 1 gene in breast and other cancers. *Mol. Cancer Res.* **2008**, *6* (11), 1710−1717.

(45) Cadenas, C.; Franckenstein, D.; Schmidt, M.; et al. Role of thioredoxin reductase 1 and thioredoxin interacting protein in prognosis of breast cancer. *Breast Cancer Res.* **2010**, *12* (3), R44.

(46) Hasegawa, T.; Asanuma, H.; Ogino, J.; et al. Use of potassium channel tetramerization domain-containing 12 as a biomarker for diagnosis and prognosis of gastrointestinal stromal tumor. *Hum. Pathol.* **2013**, *44* (7), 1271−1277.

(47) Nuzzo, R. Scientific method: statistical errors. *Nature* **2014**, *506* (7487), 150−152.

(48) Johnson, V. E. Revised standards for statistical evidence. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (48), 19313−19317.

(49) Dahl, E.; Kristiansen, G.; Gottlob, K.; et al. Molecular profiling of laser-microdissected matched tumor and normal breast tissue identifies karyopherin alpha2 as a potential novel prognostic marker in breast cancer. *Clin. Cancer Res.* **2006**, *12* (13), 3950−3960.

(50) Dankof, A.; Fritzsche, F. R.; Dahl, E.; et al. KPNA2 protein expression in invasive breast carcinoma and matched peritumoral ductal carcinoma in situ. *Virchows Arch.* **2007**, *451* (5), 877−881.

(51) Gluz, O.; Wild, P.; Meiler, R.; et al. Nuclear karyopherin alpha2 expression predicts poor survival in patients with advanced breast cancer irrespective of treatment intensity. *Int. J. Cancer* **2008**, *123* (6), 1433−1438.