RESEARCH ARTICLE

Coupling proteomics and transcriptomics in the quest of subtype-specific proteins in breast cancer

Maria P. Pavlou[1,2], Apostolos Dimitromanolakis[3] and Eleftherios P. Diamandis[1,2,3]

[1] Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada
[2] Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, ON, Canada
[3] Department of Clinical Biochemistry, University Health Network, Toronto, ON, Canada

Breast-cancer subtypes present with distinct clinical characteristics. Therefore, characterization of subtype-specific proteins may augment the development of targeted therapies and prognostic biomarkers. To address this issue, MS-based secretome analysis of eight breast cancer cell lines, corresponding to the three main breast cancer subtypes was performed. More than 5200 non-redundant proteins were identified with 23, four, and four proteins identified uniquely in basal, HER2-neu-amplified, and luminal breast cancer cells, respectively. An in silico mRNA analysis using publicly available breast cancer tissue microarray data was carried out as a preliminary verification step. In particular, the expression profiles of 15 out of 28 proteins included in the microarray (from a total of 31 in our subtype-specific signature) showed significant correlation with estrogen receptor (ER) expression. A MS-based analysis of breast cancer tissues was undertaken to verify the results at the proteome level. Eighteen out of 31 proteins were quantified in the proteomes of ER-positive and ER-negative breast cancer tissues. Survival analysis using microarray data was performed to examine the prognostic potential of these selected candidates. Three proteins correlated with ER status at both mRNA and protein levels: ABAT, PDZK1, and PTX3, with the former showing significant prognostic potential.

Received: November 20, 2012
Revised: December 19, 2012
Accepted: January 7, 2013

Keywords:
Breast cancer / Cell biology / Gene expression profiling / MS / Prognosis / Subtype-specific proteins

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Advances in screening and improved treatment options contributed to the decline of breast cancer mortality in the western world during the last decade. However, breast cancer remains the most frequently diagnosed cancer and the second cancer-related cause of death in women, highlighting the numerous unmet clinical needs [1].

Molecular profiling of breast cancer tissues has shown that breast cancer is not a single entity but rather a collection of diseases developed at the same anatomical site. Using gene expression analysis of breast tumor tissues, Perou et al. identified at least four molecular subtypes of breast cancer (luminal, HER2-neu-amplified, basal-like, and normal-like) [2]; a classification that was independently reproduced [3, 4]. One of the implications of the molecular taxonomy is that different subtypes are associated with distinct clinical characteristics such as prognosis and response to therapy [5, 6]. More specifically, patients with luminal cancers

Correspondence: Dr. Eleftherios P. Diamandis, Mount Sinai Hospital, Joseph & Wolf Lebovic Ctr., 60 Murray St [Box 32]; Flr 6–Rm L6–201, Toronto, ON, M5T 3L9, Canada
E-mail: ediamandis@mtsinai.on.ca
Fax: +416-619-5521

Abbreviations: ABAT, 4-aminobutyrate aminotransferase; **ACTB**, beta actin; **DFS**, disease free survival; **ER**, estrogen receptor; **ERBB2**, v-erb-b2 erythroblastic leukemia viral oncogene homolog 2; **ESR1**, estrogen receptor 1; **FDR**, false discovery rate; **NAF**, nipple aspirate fluid; **PCSK9**, proprotein convertase subtilisin/kexin type 9; **PSMG1**, proteasome assembly chaperone 1; **SCOC**, short coiled-coil protein; **SIL1**, endoplasmic reticulum chaperone SIL1 homolog; **SRM**, selected reaction monitoring

Colour Online: See the article online to view Figs. 1–3 in colour.

(ER-positive) show better prognosis compared to patients with HER2-neu-amplified and basal-like breast tumors [5]. Additionally, luminal and HER2-neu overexpressing cancer patients benefit from targeted therapies—endocrine therapies and trastuzumab, respectively—whereas patients with basal-like breast cancer are left with the option of conventional cytotoxic therapies [7].

During the last decade it has been shown that molecular classification of breast cancer holds promise for the development of novel prognostic tools and treatment targets. However, proteins are the mediators of all biological processes and the molecular targets of the majority of drugs. Moreover, the proteome integrates the cellular genetic information and environmental influences [8]. Hence, studying the proteome in combination with transcriptomic studies may augment the development of targeted therapies and the identification of disease-relevant protein networks.

Due to the distinct clinical characteristics of each subtype, we reasoned that subtype-specific proteins may have significant potential as prognostic biomarkers and/or therapeutic targets for breast cancer. Toward our goal, a "bottom-up" proteomics approach and a 2D LC-MS/MS platform on a linear ion trap–orbitrap were utilized to analyze the conditioned media of eight breast cancer cell lines. The cell lines were selected to represent the three main breast cancer subtypes: luminal, basal, and HER2-neu-amplified. Proteins common to all cell lines of the same subtype but not present in the other two subtypes were identified and proposed as subtype-specific proteomic panels. An in silico mRNA analysis using publicly available data from four independent experiments was performed as a preliminary verification step to confirm the subtype-specificity of our proteins. Since most of the identified proteins have not been connected to breast biology or breast cancer before, we sought to examine if these proteins are expressed in breast cancer tissues at the proteome level using a mass spectrometric approach. The prognostic potential of the top candidate proteins was examined using publicly available microarray data.

2 Materials and methods

2.1 Breast tumor-derived cell lines

Eight cell lines were obtained from American Type Culture Collection (ATCC, Manassas, VA, USA) and were grown using the recommended conditions: MCF-7, HCC-1428, BT-438, HCC-38, HCC-1143, MDA-MB-231, SK-BR-3, and HCC-202. (Supporting Information Table 1).

2.2 Cell culture

The seeding density and incubation periods were optimized as previously described [9] to maximize protein secretion and minimize cell death. Optimum conditions are summarized in Supporting Information Table 2. See Supporting Information Materials and Methods for detailed cell culture conditions. Total protein concentration was determined using a Coomasie (Bradford) total protein assay (12) and a volume corresponding to 1 mg of total protein was processed as follows. The experiment was performed in duplicate for each cell line.

2.3 Conditioned media sample preparation

The samples were dialyzed, lyophilized, and processed for trypsin digestion. For a detailed protocol see Supporting Information Materials and Methods. Trypsin digests were lyophilized to dryness.

2.4 Human tissue samples

Sixteen breast cancer tissues, eight ER-positive and eight ER-negative, from patients with primary breast cancer were processed as previously described [10] and stored as cytosolic extracts in liquid nitrogen. Selection of samples was based on the availability of stored cytosolic extracts that remained after routine ER analyses. The samples were not subjected to freezing-thaw cycles prior to analysis.

2.5 Breast cancer cytosol sample preparation

Equal amounts of total protein from each sample, as determined by the Bradford total protein assay (Pierce, USA), were pooled to create two pools (total protein concentration 250 μg): one containing ER-positive and a second containing ER-negative samples. The samples were diluted four-fold with 50 mM ammonium bicarbonate buffer solution and each pool was then divided in two technical replicates that were processed individually. Proteins were denatured with 0.05% RapiGest (Waters) at 60°C and the disulfide bonds were reduced with DTT (final concentration, 10 mM; Sigma). Following reduction, the samples were alkylated with iodoacetamide in room temperature for 45 min in the dark (final concentration, 20 mM; Sigma). The samples were then trypsin-digested overnight at 37°C (trypsin: protein ratio of 1:50; Promega, sequencing-grade modified porcine trypsin). RapiGest was cleaved with 1% TFA and samples were centrifuged at 453 × *g* for 20 min.

2.6 Strong cation-exchange LC

The tryptic peptides were subjected to strong cation exchange chromatography to reduce sample complexity. Refer to Supporting Information Materials and Methods for a detailed protocol.

2.7 MS

Mass spectrometric analysis was performed as previously described [9] and is also described in Supporting Information Materials and Methods.

Instrument performance was monitored using BSA as quality control. More specifically, 10 fmol of digested BSA were injected in the mass spectrometer and five peptides (m/z: 582.316, 722.817, 653.360, 733.282, 740.401) were monitored in terms of retention time, peak shape, area under the curve, and mascot score to evaluate proper and consistent instrument performance.

2.8 Data analysis

2.8.1 Cell lines

Data analysis was performed as previously described [9] and is also described in Supporting Information Materials and Methods. Protein Center Professional Edition (v. 3.5.2.1; Proxeon Bioinformatics, Odense, Denmark) was utilized for obtaining genome ontology information including cellular localization, function, process annotations, and further analysis.

2.8.2 Breast cancer cytosols

Data analysis was performed by MaxQuant software (v 1.1.1.25) [11]. MS/MS spectra were searched against the human international protein index database (v3.68, 87,061 sequences) by Andromeda search engine supplemented with frequently observed contaminants and concatenated with reversed copies of all sequences. Enzyme specificity was set to trypsin and a maximum of two missed cleavages was allowed. Carbamidomethylcysteine was set as fixed and *N*-acetylation and methionine oxidation were set as variable modifications. The initial maximum allowed mass deviation was set to 7 ppm for monoisotopic precursor ions and 0.5 Da for MS/MS peaks. The required minimum peptide length was six amino acids. The false discovery rate (FDR) at the peptide level and protein level was set to 1%.

Label-free quantification was performed also in MaxQuant by extracting isotope patterns for each peptide in each run and matching to each other across runs using peptide identifications, very high mass accuracy, and retention time. In order to identify differentially expressed proteins between ER-positive and ER-negative samples, the average intensity of the two technical replicates for ER-positive and ER-negative sets was used and the $\log_2$ ratio ER-positive/ER-negative was computed. The ratio was then transformed to z-scores by using the median of $\log_2$ ratio and a robust estimate of the SD based on interquartile range. Outliers were identified by computing a p-value for each z-score and then adjusting it for FDR based on Benjamini and Holberg correction. A cutoff of 0.0025 was used to identify significantly differentially expressed proteins.

2.9 ELISAs

ELISA for proprotein convertase subtilisin/kexin type 9 (PCSK9) was purchased commercially and performed according to the manufacturer's instructions (R&D systems, catalog # DCP900). Kallikreins 5 and 6 were measured in conditioned media using in-house developed ELISA assays, as described previously [12].

2.10 Cell line mRNA expression microarray data

Cell line mRNA expression data were obtained from NCBI GEO [13], experiment GSE12790 [14], performed on the Affymetrix HGU133Plus2 platform. From all raw files, eight cell lines that corresponded to our cell line work were selected. The CEL files were imported into *R* and normalized using gcRMA [15] in the Bioconductor [16] environment. Expression values for the eight cell lines were obtained as $\log_2$ transformed intensities. Probes that showed little variation (i.e. having a maximum over minimum ratio of less than two between different cell lines) were excluded.

From the normalized expression measurements, probes that matched the genes of interest were selected using the current Affymetrix annotation file. The $\log_2$ intensity values were used for constructing a heatmap on which genes were ordered based on which cell lines they were identified, in our proteomic data analysis. The cell lines were grouped according to the reported subtype. For the color coding of the heatmap, we centered each row on the median expression and used the centered value as a color scale. Thus, the qualitative difference in color is proportional to the $\log_2$ expression ratio between each two-cell lines.

2.11 Tissue mRNA expression and survival data

The NCBI GEO was queried for datasets with breast cancer tissue microarray data that were performed on Affymetrix arrays (HG-U133A or HG-U133Plus2), one of the most common platforms with wide gene coverage. The datasets A: GSE7390 [17], B: GSE2034 [18], C: GSE21653 [19], and D: GSE4922 [20] were selected for further analysis.

Raw CEL files for the four datasets were obtained from NCBI GEO repository. Normalization and quality control metrics were computed in R 2.14/Bioconductor 2.8 and simpleaffy [21] packages. Quality control metrics were evaluated (average background, RNA degradation, scale factors, percent present) by simpleaffy. Samples that showed high 5′ to 3′ ratio for control genes (>2.5 for ACTB, >2 for GAPDH) or flagged as outliers in other metrics were excluded. After quality control, expression data were normalized using gcRMA [15]. In

total 924 samples passed, quality control criteria and were used for further analysis.

All probes matching to the genes identified by proteomic analysis were selected. To ensure consistency of interpretation among datasets, only probes that were common between all experiments were included, effectively selecting only probes on the U133A chip. In total, 46 probes matching to 28 genes identified by proteomics were used in downstream analysis.

Evaluation of the probe correlation with ER expression was based on estrogen receptor 1 (ESR1) probe 205225_at, which exhibited excellent correlation with ER status based on clinical data (ANOVA test $p < 10^{-30}$ in all datasets). Pearson's product moment correlation coefficient and significance of the correlation was evaluated for all genes, independently on each dataset.

2.12 SRM assay development for ABAT

2.12.1 Peptide selection

Global proteome machine proteomics database (http://mrm.thegpm.org) was used to select top nine peptides for 4-aminobutyrate aminotransferase (ABAT) protein. Peptides were then confirmed in selected reaction monitoring (SRM) atlas (http://www.srmatlas.org) or in our LC-MS/MS identification data. Fully tryptic and doubly charged peptides with 7–20 aminoacids were chosen. From the nine initial candidate peptides those with methionine and N-terminal cysteine residues were excluded resulting in four candidate peptides. Peptides were also analyzed with the Basic Local Alignment Search Tool to ensure that they were unique to the protein of interest. In silico protein digestion and peptide fragmentation were performed with Pinpoint software (Thermo Scientific).

2.12.2 SRM conditions

In the first step of method development, four peptides and 27 transitions were included in one SRM method. For method optimization, digested samples of breast cancer cytosols used previously in our analysis were loaded onto a 2 cm C18 column with 15 μm inner diameter and were eluted to a resolving 5 cm analytical C18 column (inner diameter 75 μm) for separation. This setup was online coupled to a triple-quadrupole mass spectrometer (TSQ Vantage, Thermo Scientific) using a nano-ESI source (Proxeon Biosystems). Details regarding LC and MS methods can be found in our previous study [22]. Parameters of SRM method were as follows: predicted collision energy values, 0.002 m/z scan width, 20 ms scan time, 0.2 resolution at the first quadrupole, 0.7 resolution at the third quadrupole, 1.5 mTorr pressure at the collision cell, tuned tube lens values, 7 V skimmer offset. Retention times, relative intensities of peptides, most intense and selective transitions per peptide were recorded at that step.

2.12.3 Sample preparation

Twenty breast cytosolic extracts (10 ER-positive and 10 ER-negative) were processed for trypsin digestion individually as follows. Total protein for each sample was measured by the Bradford total protein assay (Pierce), and the volume was adjusted to extract equal amounts of total protein (30 μg) from the individual samples. Samples were diluted four times with 50 mM ammonium bicarbonate buffer solution and proteins were denatured with 0.05% Rapigest (Waters) at 60°C, and the disulfide bonds were reduced with DTT (final concentration, 10 mM; Sigma) before being subject to alkylation with iodoacetamide in room temperature for 45min in the dark (final concentration, 20 mM; Sigma). 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. Sixty femtomoles of heavy 13C6, 15N2 L Lysine-labeled peptide (LSEPAELTDAVK*) of KLK3 protein was added as an internal standard for microextraction. Rapigest was cleaved with 1% TFA, and samples were centrifuged at 453 g for 10 min and supernatant was carefully collected to avoid pellet contamination. Volume corresponding to 5 μg of peptides were purified and extracted using ZipTip C_{18} pipette tips (Millipore) twice for each sample, and were eluted using 2 μL of mobile phase B (55% ACN, 0.1% formic acid). Eighteen μL of mobile phase A (0.1% formic acid) was added to each sample to yield one injection of 18 μL.

2.12.4 Protein quantification by SRM

Six transitions of best performing peptide (IDIPSFDWPIAPFPR) were used for the quantification of ABAT protein in breast cancer cytosolic extracts (Supporting Information Table 3). Housekeeping proteins beta actin (ACTB) and GAPDH were selected to serve as relative internal standards (SRM methods developed previously [22]). The final SRM method targeted 45 transitions of 5 peptides (one peptide for ABAT, two peptides for GAPDH, one peptide for ACTB and heavy labelled peptide of KLK3 as internal standard, Supporting Information Table 3) during a 60 min LC gradient. Scan time was set to 30 msec and was calculated to ensure the measurement of at least 15–20 points per LC peak. Peptides were separated by 60-min C18 RP LC (EASY-nLC, Proxeon) and analyzed by a triple-quadrupole mass spectrometer (TSQ Vantage, Thermo Scientific) using a nano-ESI source, as previously described [22]. Analytical nano-LC column performed well several days before and after the analysis, so stability of SRM signal was not compromised. Reproducibility of SRM signal was ensured by running two quality control solution of 0.25 fmol/mL BSA every ten runs. Raw files recorded for each sample were analyzed using Pinpoint software, and CSV

files with peptide areas were extracted (Supporting Information file 1). It should be noted that three out of 40 injected samples were excluded for further analysis due to inadequate microextraction based on the signal of the quality control peptide. The quality control peptide was not used in further data analysis.

2.12.5 Data analysis

Normalization among samples was performed by calculating the ratio of ABAT peptide extracted area over the sum of GAPDH and ACTB peptides extracted area scaled by a factor of 10^4. For samples with two technical replicates the average normalized peak area was used for further analysis. The results were analyzed using the GraphPAD Prism (GraphPAD Software).

3 Results

3.1 Delineation of breast cancer cell lines secretome

The conditioned media of eight breast cancer cell lines representing the main breast cancer subtypes (ER-positive, basal and HER2-neu-amplified) were analyzed using a bottom-up tandem mass-spectrometric approach. Two biological replicates were analyzed for each cell line, yielding between 1492 and 2828 proteins in each of the eight cell lines, with FDR of less than 1% (Supporting Information Table 4). The reproducibility between biological replicates, defined as the overlap of identified proteins in each replicate, was over 70% in all cell lines analyzed; this reproducibility is considered acceptable in experiments of similar nature [9, 23]. More than 60% of the proteins were identified with two or more peptides (Supporting Information Table 4).

In the absence of cell death, conditioned media is expected to contain proteins secreted or shed from the cellular plasma membrane. Thus, to verify that proteomic analysis of the conditioned media resulted in enrichment for the subproteome of interest (secreted or shed proteins), cellular localization annotation of the identified proteins was performed using the Gene Ontology consortium database as provided in ProteinCenter. Our datasets were enriched for proteins annotated as "extracellular," "cell surface," and "membrane" (Supporting Information Table 4). Although cell growth conditions were optimized (Supporting Information Table 2), uncontrolled cell death can account for the large proportion of proteins annotated as cytosolic. Alternatively, proteins annotated as "cytosolic" could reach the extracellular space through exosome secretion [24].

A summary of proteins per cell line along with international protein index identifiers, gene names, number of identified peptides, cellular localization annotation, and signal peptide information can be found in Supporting Information file 2. An extensive peptide report for the eight cell lines can be found in Supporting Information file 3.

Overall, this experiment resulted in the identification of 5222 non-redundant proteins in the secretomes of eight breast cancer cell lines. To verify the biological relevance of proteins identified using our in vitro system, the generated non-redundant list of proteins was compared against our previously studied nipple aspirate fluid (NAF) proteome [25]. A total of 553 out of 863 (64%) previously identified NAF proteins were present in the compiled breast cancer cell line proteome. It is worth mentioning that approximately 50% of proteins identified in the NAF proteome were annotated as cytosolic, further supporting the notion that cytosolic proteins could reach the extracellular space through diverse pathways including exosome secretion [24].

The MS-based identification of KLK5 and KLK6 (proteins previously studied in the context of breast cancer [26]) and PCSK9 a protein never connected to breast cancer before) was verified using immunosorbent assays. KLK5 was expressed in HCC-1143 and HCC-38 cell lines at concentrations 42 μg/L and 30μg/L, respectively. KLK6 was expressed by BT-483 (0.5μg/L), MCF-7 (0.3 μg/L), HCC-1143 (7 μg/L), and HCC-38 (5 μg/L). Finally, PCSK9 was expressed only by the basal cell lines MDA-MB-231, HCC-38, HCC-1143 at the level of 10 μg/L, 5 μg/L, and 2,5 μg/L, respectively.

3.2 Identification of subtype-specific proteins

Next, we sought to determine breast cancer subtype-specific secretome signatures in our data. To achieve so, the identified proteomes were qualitatively compared to select proteins common among the cell lines of the same subtype, yet unique to each subtype. ProteinCenter was utilized for the comparisons among the eight cell lines and results were manually verified. To increase stringency, only proteins present in both biological replicates, with two peptides identified in at least one replicate were selected. Additional filtering was added for proteins with multiple isoforms, whereby they were excluded from further analysis to avoid gene name promiscuity. This step-wise selection of proteins is depicted in Supporting Information Fig. 1. By using these criteria, we managed to identify 23 basal, four ER-positive and four HER2-neu amplified specific proteins, as shown in Table 1. Notably, v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 (ERBB2) was one of the proteins uniquely identified in the HER2-neu-amplified subtype.

3.3 In silico verification of the proposed subtype-specific protein panels using publicly available microarray data

Our list of 31 proteins (Table 1) was identified using a discovery-based mass spectrometric approach. However, for most of these proteins there are no commercially available quantitative methods at present. Consequently, we opted to

Table 1. Summary of subtype-specific proteomic panels

Basal			
VIM	FBN1	LAMB3	TGM2
GSTP1	LOXL2	AHNAK2	PTX3
PCSK9	COL4A2	MFI2	PTRF
AKR1B1	DCTD	MT2A	TRAP1
AKR1C3	LMAN1	DNAJB4	PSMG1
ICAM2	BAG2	SIL1	
HER2-neu-amplified			
VAMP8	VTCN1	SCOC	ERBB2
ER-positive			
PDZK1	CLSTN2	ABAT	SEMA4C

SIL1, endoplasmic reticulum chaperone SIL1 homolog; FBN1, fibrillin 1; PCSK9, proprotein convertase subtilisin/kexin type 9; COL4A2, collagen type IV alpha 2; DCTD, dCMP deaminase; PSMG1, proteasome assembly chaperone 1; short coiled-coil protein; ERBB2, v-erb-b2 erythroblastic leukemia viral oncogene homolog 2.

preliminarily examine the relationship of these candidates with breast cancer subtypes by using an in silico approach, based on transcriptomic data. We first studied the correlation between mRNA and protein levels of the genes of interest identified through our analysis. Towards this aim, we performed a meta-analysis of available mRNA expression data focusing on the cell lines and genes of interest [14]. The qualitative concordance between microarray and protein expression data was high (Fig. 1, Supporting Information Table 5). Twenty of 23 basal-specific proteins (except TRAP1, endoplasmic reticulum chaperone SIL1 homolog and proteasome assembly chaperone 1) showed higher expression in basal cell lines by microarray analysis, with ratios ranging from 2.16 to 369, when compared to ER-positive/HER2-neu-amplified cell lines. Among ER-positive-specific proteins, all four (ABAT, CLSTN2, PDZK1, and SEMA4C) were validated using the microarray data, with PDZK1 having a ratio of 640 in comparison to the mean value of the other cell lines. From the three proteins identified by our proteomic work in HER2-amplified cell lines (short coiled-coil protein (SCOC), VAMP8, and VTCN1), none exhibited a ratio of 2, however VTCN1 was uniquely expressed in the two HER2-amplified cell lines as well as HCC-1428. Similar results were obtained when analysis was expanded in 51 breast cancer cell lines (Supporting Information Fig. 2). In total, 24 of our 30 candidates showed microarray expression patterns consistent with the proteomic data.

The good qualitative concordance between mRNA and protein expression levels in the cell lines used in the study, encouraged us to examine the subtype specificity of our panels in breast cancer tissue samples. We performed an in silico mRNA expression analysis using publicly available data from four independent experiments containing a total of 1039 patients with primary breast cancer [17–20]. A common microarray platform, non-overlapping patient cohorts and documented clinical information were prerequisites for dataset selection. Patient characteristics are summarized in Table 2. All genes except two (PCSK9, SCOC) had at least one probe on the selected microarray platform and were qualified for further analysis.

ER correlation was evaluated, separately on each of the four datasets. Pearson correlation coefficients (r) for the probes that showed significant correlation with ER status ($p < 0.05$ in at least three datasets) are shown in Fig. 2. Among the ER-positive-specific genes, ABAT, CLSTN2, and PDZK1 exhibited a positive correlation with ER-status, reaching a significance level of $p = 10^{-45}$ for ABAT (Supporting Information Table 6) and all three genes had a consistent pattern among the four datasets (Fig. 2). SEM4AC did not show a significant correlation with ESR1 expression. Among basal-type-specific proteins, 13 showed a consistent negative correlation with ESR1 at highly significant levels while three (dCMP deaminase, fibrillin 1, and endoplasmic reticulum chaperone SIL1 homolog) showed a positive correlation. Correlation of the proposed HER2-neu-specific proteins (VAMP8, VTCN1, and SCOC) with ERBB2 failed to reveal any consistent and statistically significant association.

3.4 MS-based verification of the proposed subtype-specific proteins in breast cancer tissues

Given that most proteins found to be subtype-specific have not been studied in the context of breast cancer before, we sought to verify their expression in breast cancer tissues. Due to the correlation of a subset of these proteins with ESR1 expression, we were interested in examining this correlation at the proteome level. Toward our aim, we performed an extensive proteomic analysis of 16 breast cancer cytosol samples; eight ER-positive and eight ER-negative samples. The samples of each type were pooled to obtain sufficient amount of sample for MS-based analysis. Approximately 3300 and 3500 proteins were identified in the two technical replicates of ER-negative and ER-positive samples, respectively (Supporting Information File 4). The false discovery rate was 1% and almost 70% of proteins were identified with at least two peptides. Notably, ER protein was identified in the pool of ER-positive samples but was absent from the ER-negative pool, as expected. A similar pattern was observed for progesterone receptor—an estrogen-regulated protein. Additionally, KLK3, previously reported to be identified in breast cancer cytocols by immunoassay was also identified (38). In summary, a total of 4124 nonredundant proteins were identified by the analysis of 16 breast cancer tissue pooled samples.

Eighteen out of 30 proposed subtype-specific proteins based on the cell lines analysis were also identified in the breast cancer cytosol proteome (Supporting Information Table 7). Label-free MS-based quantification using extracted ion current was utilized for relative quantification of the identified proteins between ER-positive and ER-negative samples. In accordance with the cell line work, protein PTX3 was found to be significantly (p-value < 0.0001)

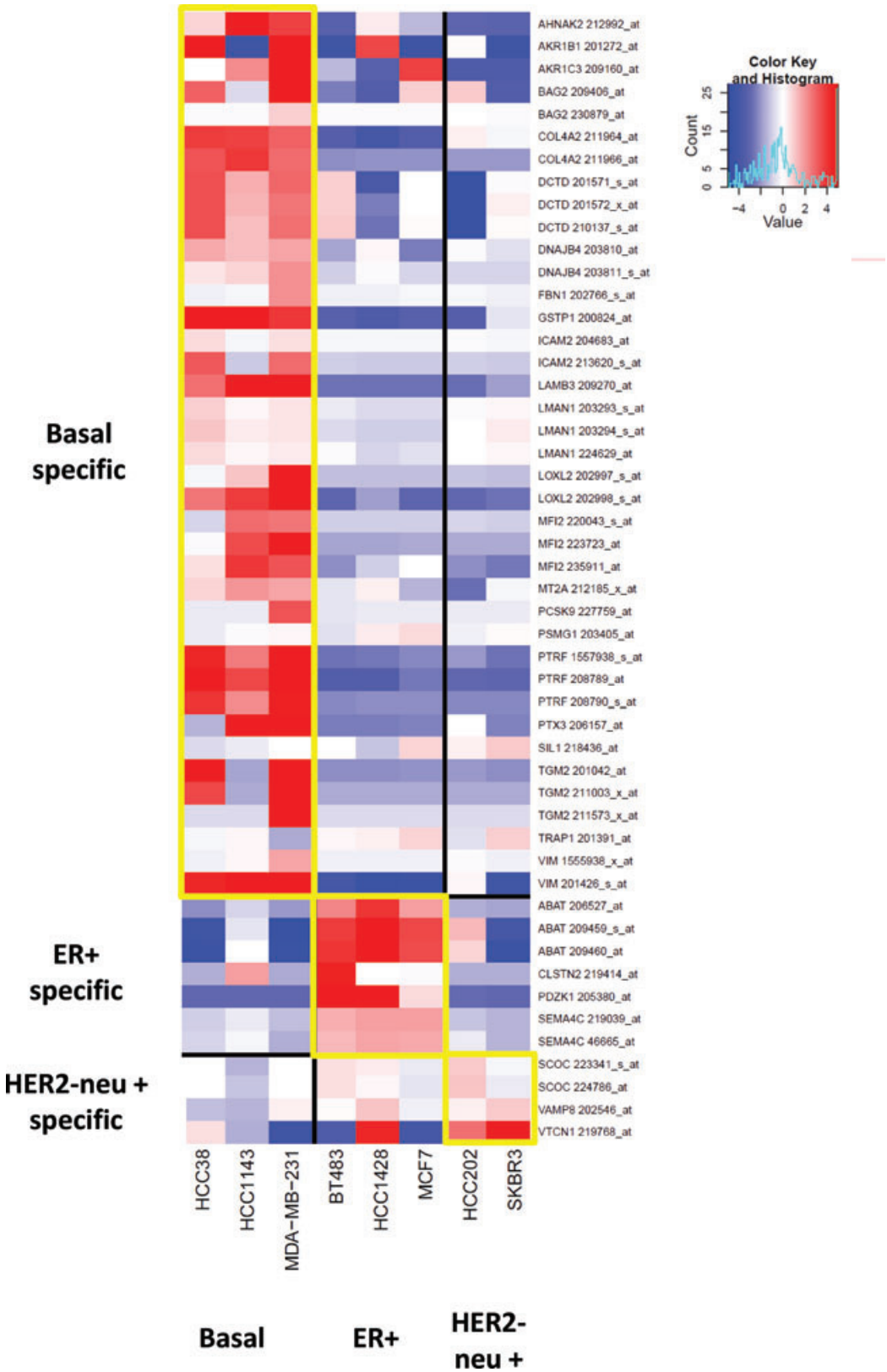


Figure 1. mRNA expression of the selected genes in eight cell lines displayed as a heat map. Inside the cells are $\log_2$ expression values. Red corresponds to higher than mean expression whereas blue to lower. The color of each cell is relative to the mean value of each row and can be used to judge over/under expression.

Table 2. Demographics of the four microarray experiments used for verification.

ID	Experiment	Number of samples	Failed Q.C.	Survival data	ER-positive	ER-negative	Two-year DFS TRUE/ FALSE	Five-year DFS TRUE/ FALSE	References
A	GSE7390	198	6	192	130	62	168/23	131/51	[17]
B	GSE2034	286	81	205	144	61	174/30	131/63	[18]
C	GSE21653	266	26	226	132	105	180/26	94/60	[19]
D	GSE4922	289	2	245	211	34	197/42	158/69	[20]
	Total	1039	115	868	617	262			

Summary demographics are shown, along with disease free survival for the 2-year/5-year endpoints. Patients who were censored and the status could not be validated were excluded from endpoint analysis. DFS, disease-free survival.

underexpressed in ER-positive breast cancer tissues with 60-fold difference between ER-positive and ER-negative. In the case of proteins specific to ER, ABAT showed statistically significant 14-fold overexpression (p-value = 0.002) in ER-positive breast cancer tissues. Additionally, PDZK1 was identified only in the ER-positive samples.

The overexpression of ABAT in ER-positive breast cancer cytosolic extracts was also validated in an independent set of twenty individual samples (ten ER-positive and ten ER-negative) using a targeted mass spectrometric approach. Each trypsin-digested sample was microextracted and analyzed twice and microextraction efficiency was assessed using a heavy labeled peptide that was spiked-in as quality control. Three out of 40 injected samples were excluded due to inadequate microextraction. For samples having two technological replicates median CV was 12.1%, also depicted in Supporting Information Table 8. For comparison of the ER-positive and ER-negative group the average normalized expression of the two technical replicates of each sample (when applicable) was used. The mean normalized expression of ABAT in ER-negative ($n = 10$) and ER-positive ($n = 10$) samples was 2.8 (SEM = 0.37) and 15.4 (SEM = 3.2), respectively (Fig. 3, panel A). The difference between the two groups was tested using independent samples t-test and was found to be statistically different ($n = 20$, df = 18, p-value = 0.001).

3.5 Exploring the prognostic potential of ABAT, PDZK1, and PTX3

Breast cancer subtypes show distinct clinical outcomes with ER-positive breast cancer patients having better prognosis in comparison to basal or HER2-neu-amplified tumor-carrying patients [5]. Based on this observation, we hypothesized that proteins expressed uniquely by cancer cells representing those subtypes may have prognostic potential. Examining the expression of the three genes at a two-year endpoint for disease-free survival (DFS), ABAT showed consistently higher expression in patients with no recurrence and reached significance in each of the four datasets independently ($p = 4.27 \times 10^{-5}$, 0.027, 0.015, and 6.14×10^{-4}), (also depicted in Fig. 3). The expression levels of ABAT were on average 2.3-fold higher in patients with DFS of more than two years. Additionally, ABAT expression remained significantly changed at the five-year endpoint in all four datasets, with consistent direction of effect and a mean ratio of 1.5. Survival analysis using the online tool Gene expression-based Outcome for Breast cancer Online [27] revealed that patients with high expression of ABAT have slightly longer relapse-free survival compared to those with low expression ($p = 0.036$, Supporting Information Fig. 3). When survival analysis was performed in subgroups of patients, it was shown that patients with ER-positive disease and high ABAT expression have slightly better prognosis than those with low expression ($p = 0.037$, Supporting Information Fig. 3). Moreover, tamoxifen-treated breast cancer patients with high expression of the ABAT gene have better prognosis than those with low expression (Fig. 3) and a similar pattern was observed for breast cancer patients with grade II tumors (Fig. 3). The associations remained significant in a multivariate analysis using ER status and grade as covariates (Supporting Information Fig. 4).

PDZK1 was found to have consistently higher expression in patients with no recurrence at a two-year endpoint but the association reached significance in two out of the four datasets. PTX3 was not found to be related to DFS at the mRNA level.

4 Discussion

Given the distinct clinical characteristics of each subtype, subtype-specific proteins may be useful as prognostic biomarkers or therapeutic targets especially in the case of triple-negative breast cancer disease that lacks targeted therapies. The present study provides an insight of the value of breast cancer cell secretomics for identifying subtype-specific breast cancer proteins. Given that intracellular and cell surface proteins have been previously studied in the quest for novel subtype-specific proteins [28, 29], the current study focuses on proteins secreted or shed by breast cancer cells. To our knowledge, cancer cell secretomes have not been explored in the field of breast cancer subclassification, although a large number of secreted proteins have been shown to be implicated in various steps of cancer development and progression [30]. The use of established cancer cell lines for

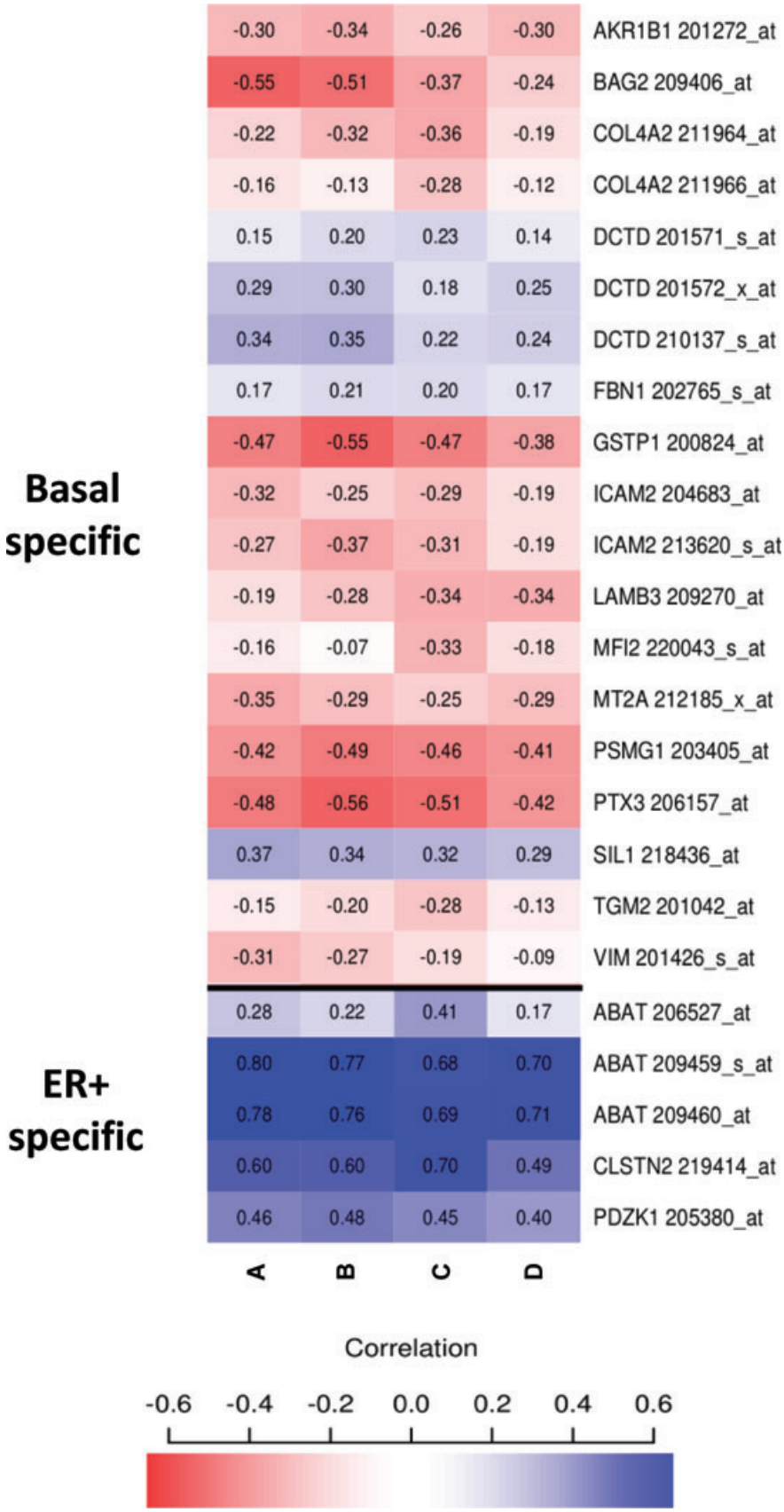
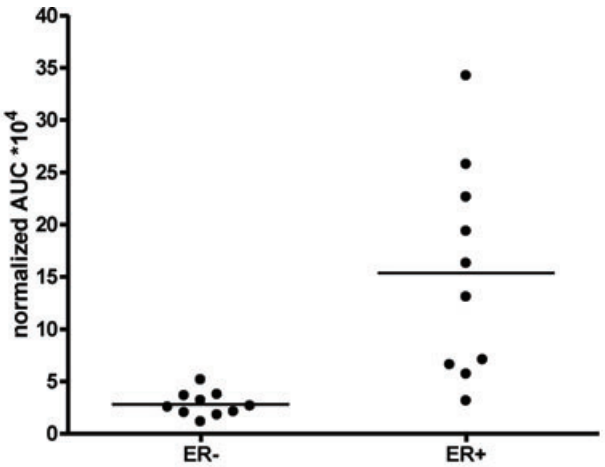


Figure 2. Correlation of each probe with ESR1 gene expression based on tissue microarray data in four gene-profiling experiments with breast cancer tissues. Pearson correlation coefficient (*r*) is shown inside the cells. Red corresponds to negative and blue to positive correlation.

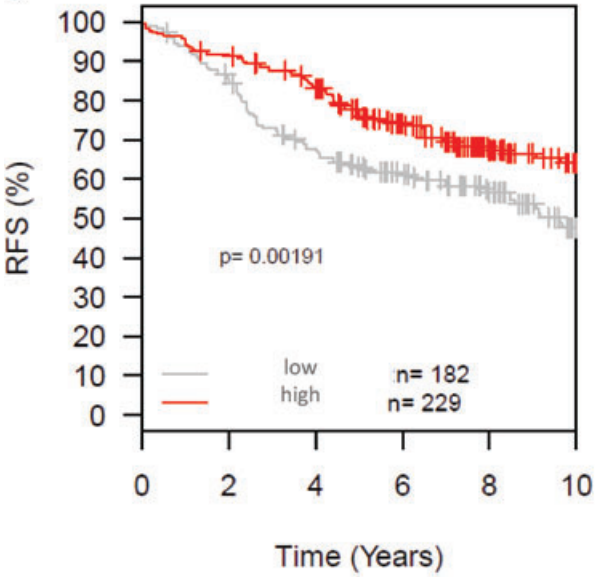
A



B

	Dataset A	Dataset B	Dataset C	Dataset D	
Gene	**Ratio**	**Ratio**	**Ratio**	**Ratio**	**Mean Ratio**
ABAT	3.2 (p=0.00004)	2 (p=0.027)	1.9 (p=0.015)	2.1 (p=0.006)	2.3

C



D

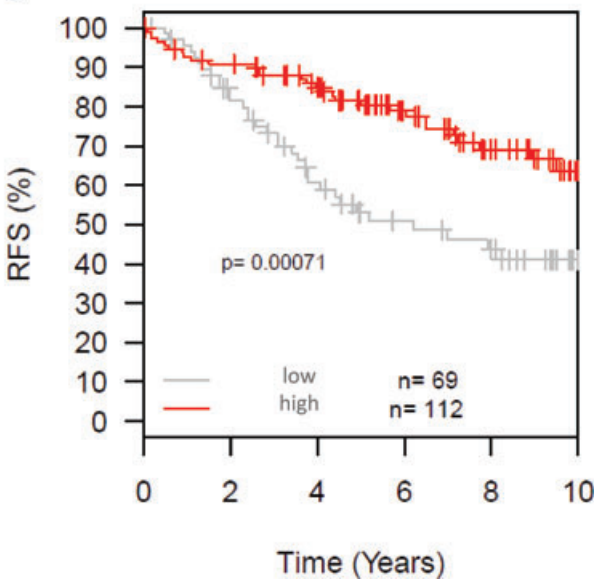


Figure 3. Overexpression of ABAT in ER-positive breast cancer tissues and survival analysis for ABAT using publicly available microarray data and GOBO. (A) Levels of ABAT measured with a selected reaction monitoring (assay were found to be significantly upregulated (*t*-test, *p*-value = 0.001) in ER-positive breast cancer cytosolic extracts when compared to ER-negative samples. The *y*-axis depicts the normalized area of the peak (area under the curve, AUC) multiplied by 10^4 (for details refer to Section 2). (B) Summary of the results from the Cox proportional hazards model used to evaluate the significance of probe expression levels with relapse-free survival for ABAT. A positive ratio indicates good prognostic potential with higher expression in patients showing no 2-year disease recurrence. (C) Kaplan–Meier analysis for (i) tamoxifen-treated breast cancer patients and (ii) patients with grade II breast cancer using the online tool GOBO. Patients with high ABAT expression have better prognosis when compared to patients with low expression. Relapse-free survival was used as end-point.

biomarker discovery overcomes the issues of cellular heterogeneity and contamination by stromal proteins but raises concerns as to whether these cells truly recapitulate the cancer in vivo. To compensate for this potential limitation, we selected multiple cell lines representing the major breast cancer subtypes. The significant overlap between the cell line secretomes and the NAF proteome, along with the identification of known breast cancer biomarkers (e.g. ERBB2) indicates that, despite the known limitations, established cancer cell lines could be a useful source for biomarker discovery.

Our findings from the cell line secretomes were supported by in silico verification, using publicly available gene expression data. The wealth of generated microarray data from patient samples, accompanied with clinical variables, is an attractive resource for validation studies and meta analyses [31]. Although mRNA levels can account for 40% of the variability at the protein level [32], a good concordance between mRNA and protein levels for the genes of interest in the eight breast cancer cell lines was observed. Thus, utilization of microarray data from breast cancer tissues as an in silico verification step could be informative in a biomarker discovery pipeline, as the one described here.

Not all the proteins identified during the secretomics approach as subtype-specific were verified during subsequent analyses. Two proteins, PCSK9 and SCOC did not have probes in the Affymetrix arrays and 12 proteins were not identified in the breast cancer tissue proteomes. The development of multiplex quantitative MS-based assays for the targeted quantification of those proteins [33, 34] in breast cancer tissues will be the focus of our subsequent investigations.

The identified proteins could be utilized in a variety of ways. In the case of ER-positive specific proteins, a protein that is specifically expressed/secreted by ER-positive breast cancer tumors could be used as a surrogate marker for the ER status of metastatic breast cancer. Due to discordances between the primary and metastatic site, re-biopsy and re-assessment of the ER status has been recommended [35]. However, the procedure is invasive and could also be challenging especially in the case of bone metastasis [36]. Therefore, a blood-based test for assessing ER status in metastatic breast cancer could be highly beneficial. This type of biomarker should not only be specific ER-positive disease but also be absent or in low concentrations in the plasma of normal individuals. Based on the human plasma proteome reference set with estimated concentrations that can be found in PeptideAtlas [37], PDZK1 is almost undetectable in normal plasma whereas ABAT has an estimated concentration of 0.5 ng/mL. The usefulness of these proteins toward that direction warrants further investigation. PDZ domain containing 1 (PDZK1) protein has been mostly described as an estrogen-regulated protein in the context of breast cancer [38]. However, PDZK1 has been reported to be over-expressed in human carcinomas and interact with multidrug resistance-associated protein 2 (ABCC2), suggesting that it could play a role in the cellular mechanisms associated with drug resistance [39]. Furthermore, overexpression of PDZK1 has been found to associated with drug resistance in multiple myeloma [40]. Given that ER-positive breast cancer tumors are less sensitive to chemotherapy [41], the role of PDZK1 in chemotherapeutic resistance should be investigated. Finally, PTX3 is a soluble pattern recognition receptor found to be inducible by inflammation. It has been shown that PTX3 overexpressing breast cancer cells inhibit angiogenesis in vitro and decreases tumor volume in vivo. However, very recently, PTX3 was reported to be highly expressed in breast cancer tissues from patients classified as high risk based on the results of OncotypeDx [42]. Although controversial, further studies are required for elucidating the role of PTX3 in breast cancer biology.

High expression of ABAT was shown to be associated with better prognosis of breast cancer patients, especially in the case of tamoxifen-treated patients and patients with grade II disease. Notably, ABAT has never been studied in the context of breast cancer before. This protein is responsible for catabolism of gamma-aminobutyric acid (GABA), the most abundant neurotransmitter of the CNS, into succinic semialdehyde [43]. Interestingly, it has been previously suggested that genes related to GABA synthesis may be regulated by estrogen in the nervous system [43]. The role of ABAT in breast cancer biology is not yet clear, thus it warrants further investigation.

Similarly, VAMP8 that is associated with the HER2-neu subtype, has been found to be regulated by the HER2 oncogene [44]. Finally, numerous studies have demonstrated the association of LOXL2 expression with highly invasive properties, metastatic potential and basal-like phenotype of breast cancer tumors [45–47]. Collectively, all these observations underscore the validity of our findings and may render our discovery-based strategy (patho) physiologically relevant and concrete.

In summary, we performed an extensive proteomic analysis of eight breast cancer cell lines, generating a database of approximately 5200 breast cancer-related proteins. Using bioinformatics, we were able to generate subtype-specific proteomic panels. Our in silico verification, utilizing publicly available microarray datasets along with mass spectrometric analysis of breast cancer tissues confirmed the existence of three subtype-specific proteins with one of the candidates showing significant prognostic potential.

The authors would like to thank Dr J. Foekens for providing the breast cancer cytosolic extracts, Dr. E. Martinez-Morillo and G.S. Karagiannis for helpful comments and discussions. M.P.P. is a recipient of the Ontario Graduate Scholarship (OGS).

The authors have declared no conflict of interest.

5 References

[1] Maxmen, A., The hard facts. *Nature* 2012, *485*, S50–S51.

[2] Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M. et al., Molecular portraits of human breast tumours. *Nature* 2000, *406*, 747–752.

[3] Bertucci, F., Finetti, P., Cervera, N., Maraninchi, D. et al., Gene expression profiling and clinical outcome in breast cancer. *OMICS* 2006, *10*, 429–443.

[4] Sorlie, T., Tibshirani, R., Parker, J., Hastie, T. et al., Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci.* U.S.A. 2003, *100*, 8418–8423.

[5] Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T. et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* U.S.A. 2001, *98*, 10869–10874.

[6] Rouzier, R., Perou, C. M., Symmans, W. F., Ibrahim, N., et al., Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin. Cancer Res.* 2005, *11*, 5678–5685.

[7] Hudis, C. A., Gianni, L., Triple-negative breast cancer: an unmet medical need. *Oncologist* 2011, *16*, 1–11.

[8] Banks, R. E., Dunn, M. J., Hochstrasser, D. F., Sanchez, J. C., et al., Proteomics: new perspectives, new biomedical opportunities. *Lancet* 2000, *356*, 1749–1756.

[9] Makawita, S., Smith, C., Batruch, I., Zheng, Y., et al., Integrated proteomic profiling of cell line conditioned media and pancreatic juice for the identification of pancreatic cancer biomarkers. *Mol. Cell Proteomics* 2011, *10*, M111.

[10] 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*, 1790–1796.

[11] Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., et al., Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 2011, *10*, 1794–1805.

[12] Shaw, J. L., Diamandis, E. P., Distribution of 15 human kallikreins in tissues and biological fluids. *Clin. Chem.* 2007, *53*, 1423–1432.

[13] Edgar, R., Domrachev, M., Lash, A. E., Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002, *30*, 207–210.

[14] Hoeflich, K. P., O'Brien, C., Boyd, Z., Cavet, G. et al., In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. *Clin. Cancer Res.* 2009, *15*, 4649–4664.

[15] Wu, Z., Irrizari, R. A., Gentleman, R., Murillo, F. M., Spencer, F., Model Based Background Adjustment for Oligonucleotide Expression Arrays. Working Papers edition. Department of Biostatistics, Johns Hopkins University; 2004.

[16] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., et al., Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 2004, *5*, R80.

[17] Desmedt, C., Piette, F., Loi, S., Wang, Y. 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*, 3207–3214.

[18] Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M. et al., Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005, *365*, 671–679.

[19] Sabatier, R., Finetti, P., Cervera, N., Lambaudie, E. et al., A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res. Treat* 2011, *126*, 407–420.

[20] Ivshina, A. V., George, J., Senko, O., Mow, B. et al., Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* 2006, *66*, 10292–10301.

[21] Wilson, C. L., Miller, C. J., Simpleaffy: a BioConductor package for Affymetrix quality control and data analysis. *Bioinformatics* 2005, *21*, 3683–3685.

[22] 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.

[23] Karagiannis, G. S., Petraki, C., Prassas, I., Saraon, P., et al., Proteomic signatures of the desmoplastic invasion front reveal collagen type XII as a marker of myofibroblastic differentiation during colorectal cancer metastasis. *Oncotarget* 2012, *3*, 267–285.

[24] Jang, J. H., Hanash, S., Profiling of the cell surface proteome. *Proteomics* 2003, *3*, 1947–1954.

[25] Pavlou, M. P., Kulasingam, V., Sauter, E. R., Kliethermes, B., Diamandis, E. P., Nipple aspirate fluid proteome of healthy females and patients with breast cancer. *Clin. Chem.* 2010, *56*, 848–855.

[26] Kulasingam, V., Diamandis, E. P., Proteomics analysis of conditioned media from three breast cancer cell lines: a mine for biomarkers and therapeutic targets. *Mol. Cell Proteomics* 2007, *6*, 1997–2011.

[27] Ringner, M., Fredlund, E., Hakkinen, J., Borg, A., Staaf, J., GOBO: gene expression-based outcome for breast cancer online. *PLoS One* 2011, *6*, e17911.

[28] Goncalves, A., Charafe-Jauffret, E., Bertucci, F., Audebert, S. et al., Protein profiling of human breast tumor cells identifies novel biomarkers associated with molecular subtypes. *Mol. Cell. Proteomics* 2008, *7*, 1420–1433.

[29] Dane, K. Y., Gottstein, C., Daugherty, P. S., Cell surface profiling with peptide libraries yields ligand arrays that classify breast tumor subtypes. *Mol. Cancer Ther.* 2009, *8*, 1312–1318.

[30] Karagiannis, G. S., Pavlou, M. P., Diamandis, E. P., Cancer secretomics reveal pathophysiological pathways in cancer molecular oncology. *Mol. Oncol.* 2010, *4*, 496–510.

[31] Abba, M. C., Lacunza, E., Butti, M., Aldaz, C. M., Breast cancer biomarker discovery in the functional genomic age: a systematic review of 42 gene expression signatures. *Biomark Insights* 2010, *5*, 103–118.

[32] Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J. et al., Global quantification of mammalian gene expression control. *Nature* 2011, *473*, 337–342.

[33] Cho, C. K., Drabovich, A. P., Batruch, I., Diamandis, E. P., Verification of a biomarker discovery approach for detection of Down syndrome in amniotic fluid via multiplex selected reaction monitoring (SRM) assay. *J. Proteomics* 2011, *74*, 2052–2059.

[34] Drabovich, A. P., Jarvi, K., Diamandis, E. P., Verification of male infertility biomarkers in seminal plasma by multiplex selected reaction monitoring assay. *Mol. Cell. Proteomics* 2011, *10*, M110.004127.

[35] Amir, E., Clemons, M., Purdie, C. A., Miller, N. et al., Tissue confirmation of disease recurrence in breast cancer patients: pooled analysis of multi-centre, multi-disciplinary prospective studies. *Cancer Treat. Rev.* 2012, *38*, 708–714.

[36] Hilton, J. F., Amir, E., Hopkins, S., Nabavi, M. et al., Acquisition of metastatic tissue from patients with bone metastases from breast cancer. *Breast Cancer Res. Treat.* 2011, *129*, 761–765.

[37] Farrah, T., Deutsch, E. W., Omenn, G. S., Campbell, D. S., et al., A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* 2011, *10*, M110.006353.

[38] Ghosh, M. G., Thompson, D. A., Weigel, R. J., PDZK1 and GREB1 are estrogen-regulated genes expressed in hormone-responsive breast cancer. *Cancer Res.* 2000, *60*, 6367–6375.

[39] Kocher, O., Comella, N., Gilchrist, A., Pal, R. et al., PDZK1, a novel PDZ domain-containing protein up-regulated in carcinomas and mapped to chromosome 1q21, interacts with cMOAT (MRP2), the multidrug resistance-associated protein. *Lab. Invest* 1999, *79*, 1161–1170.

[40] Inoue, J., Otsuki, T., Hirasawa, A., Imoto, I. et al., Overexpression of PDZK1 within the 1q12-q22 amplicon is likely to be associated with drug-resistance phenotype in multiple myeloma. *Am. J. Pathol.* 2004, *165*, 71–81.

[41] Colleoni, M., Bagnardi, V., Rotmensz, N., Gelber, R. D. et al., Increasing steroid hormone receptors expression defines breast cancer subtypes non responsive to preoperative chemotherapy. *Breast Cancer Res. Treat.* 2009, *116*, 359–369.

[42] Muraoka, S., Kume, H., Watanabe, S., Adachi, J. et al., Strategy for SRM-based verification of biomarker candidates discovered by iTRAQ method in limited breast cancer tissue samples. *J. Proteome Res.* 2012, *11*, 4201–4210.

[43] Hudgens, E. D., Ji, L., Carpenter, C. D., Petersen, S. L., The gad2 promoter is a transcriptional target of estrogen receptor (ER)alpha and ER beta: a unifying hypothesis to explain diverse effects of estradiol. *J. Neurosci.* 2009, *29*, 8790–8797.

[44] Bollig-Fischer, A., Dewey, T. G., Ethier, S. P., Oncogene activation induces metabolic transformation resulting in insulin-independence in human breast cancer cells. *PLoS One* 2011, *6*, e17959.

[45] Barker, H. E., Chang, J., Cox, T. R., Lang, G. et al., LOXL2-mediated matrix remodeling in metastasis and mammary gland involution. *Cancer Res.* 2011, *71*, 1561–1572.

[46] Moreno-Bueno, G., Salvador, F., Martin, A., Floristan, A. et al., Lysyl oxidase-like 2 (LOXL2), a new regulator of cell polarity required for metastatic dissemination of basal-like breast carcinomas. *EMBO Mol. Med.* 2011, *3*, 528–544.

[47] Kirschmann, D. A., Seftor, E. A., Fong, S. F., Nieva, D. R. et al., A molecular role for lysyl oxidase in breast cancer invasion. *Cancer Res.* 2002, *62*, 4478–4483.

