Molecular Oncology

FEBS PRESS science publishing by scientists

High-throughput proteomics of breast cancer interstitial fluid: identification of tumor subtype-specific serologically relevant biomarkers

Thilde Terkelsen[1], Maria Pernemalm[2], Pavel Gromov[3], Anna-Lise Børresen-Dale[4], Anders Krogh[5,6], Vilde D. Haakensen[4], Janne Lethiö[2], Elena Papaleo[1,7] and Irina Gromova[3]

1 Computational Biology Laboratory, Danish Cancer Society Research Center, Copenhagen, Denmark
2 Cancer Proteomics Mass Spectrometry, Science for Life Laboratory, Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden
3 Breast Cancer Biology Group, Genome Integrity Unit, Danish Cancer Society Research Center, Copenhagen, Denmark
4 Department of Cancer Genetics, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Norway
5 Department of Computer Science, University of Copenhagen, Denmark
6 Department of Biology, University of Copenhagen, Denmark
7 Translational Disease System Biology, Faculty of Health and Medical Sciences, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark

Keywords
breast cancer; interstitial fluid; proteome; serological markers; subtype; tumor infiltrating lymphocytes

Correspondence
I. Gromova, Breast Cancer Biology Group, Genome Integrity Unit, Danish Cancer Society Research Center, Copenhagen 2100, Denmark
Tel: +45 35257390
E-mail: iig@cancer.dk
and
E. Papaleo, Computational Biology Laboratory, Danish Cancer Society Research Center, Strandboulevarden 49, 2100, Copenhagen, Denmark
Tel: +45 35257711
E-mail: elenap@cancer.dk

Thilde Terkelsen and Maria Pernemalm contributed equally to the work as first authors
Elena Papaleo and Irina Gromova contributed equally to the work as last senior authors

(Received 22 January 2020, revised 13 August 2020, accepted 9 November 2020, available online 4 January 2021)

Despite significant advancements in breast cancer (BC) research, clinicians lack robust serological protein markers for accurate diagnostics and tumor stratification. Tumor interstitial fluid (TIF) accumulates aberrantly externalized proteins within the local tumor space, which can potentially gain access to the circulatory system. As such, TIF may represent a valuable starting point for identifying relevant tumor-specific serological biomarkers. The aim of the study was to perform comprehensive proteomic profiling of TIF to identify proteins associated with BC tumor status and subtype. A liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis of 35 TIFs of three main subtypes: luminal (19), Her2 (4), and triple-negative (TNBC) (12) resulted in the identification of > 8800 proteins. Unsupervised hierarchical clustering segregated the TIF proteome into two major clusters, luminal and TNBC/Her2 subgroups. High-grade tumors enriched with tumor infiltrating lymphocytes (TILs) were also stratified from low-grade tumors. A consensus analysis approach, including differential abundance analysis, selection operator regression, and random forest returned a minimal set of 24 proteins associated with BC subtypes, receptor status, and TIL scoring. Among them, a panel of 10 proteins, AGR3, BCAM, CELSR1, MIEN1, NAT1, PIP4K2B, SEC23B, THTPA, TMEM51, and ULBP2, was found to stratify the tumor subtype-specific TIFs. In particular, upregulation of BCAM and CELSR1 differentiates luminal subtypes, while upregulation of MIEN1 differentiates Her2 subtypes. Immunohistochemistry analysis showed a direct correlation between protein abundance in TIFs and intratumor expression levels for all 10 proteins. Sensitivity and specificity were estimated for this protein panel by using an independent, comprehensive breast tumor proteome dataset. The results of this analysis

Abbreviations
AUC, area under the curve; BC, breast cancer; DAA, differential abundance analysis; ER, estrogen receptor; FIF, fat interstitial fluid; Her2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; LASSO, least absolute shrinkage and selection operator; MPs, microparticles; NIF, normal interstitial fluid; PgR, progesterone receptor; PPI, protein-protein interaction; RF, random forest; TIF, tumor interstitial fluid; TILs, tumor infiltrating lymphocytes; TNBC, triple-negative breast cancer.

doi:10.1002/1878-0261.12850

strongly support our data, with eight of the proteins potentially representing biomarkers for stratification of BC subtypes. Five of the most representative proteomics databases currently available were also used to estimate the potential for these selected proteins to serve as putative serological markers.

1. Introduction

Breast cancer (BC) is the most prevalent form of cancer among women worldwide, with 2.1 million new cases registered in 2018 [1]. The three main BC subtypes, namely luminal, Her2, and triple negative (TNBC), have been defined based on the expression of estrogen receptor (ER), progesterone receptor (PgR), and epidermal growth factor receptor, ErbB2/Her2 [2,3]. BC is a remarkably heterogeneous disease and molecular profiling has revealed a high level of diversity even within the same tumor subtype. However, this diversity represents a major challenge for tumor stratification, accurate patient diagnosis, and targeted treatment [4,5]. While studies of the transcriptome and genome of BC have been conducted, the differential protein composition of breast tumors, in particular the secreted/extracellular protein complement [6,7], has not been thoroughly investigated.

A large body of research has established that tumors represent complex systems in which numerous cell types, including inflammatory, immune, smooth muscle, and adipocyte cells, mediate varied interactions to ensure tumor survival and development [5,8]. These signaling-mediated, multidirectional interactions inside the tumor-stroma milieu are facilitated via the tumor interstitial fluid (TIF). As part of the tumor microenvironment, the TIF permeates the interstitial tumor space and forms an interface between circulating bodily and intracellular fluids. TIF also serves as a transport medium for nutrients, discarded cellular waste, and as a storage space for signaling substances which are synthesized locally or which are brought to organs through the circulation [9,10]. Molecular complements of tumor-proximal fluids accumulate within the interstitial tumor space via classical endoplasmic reticulum/Golgi pathways [11], via noncanonical protein secretion [12], or through the shedding of membrane vesicles (i.e., exosomes) from intracellular compartments [13,14]. Genesis, turnover, and drainage of TIF depend on many different factors, including tumor type, grade, and stage, as well as the composition of the tumor microenvironment. All of these factors are implicated in the regulation of tumor ecosystems and, therefore, are predicted to have a profound influence on the neoplastic progression of lesions. In recent years, increasing attention has been directed toward analyses of cancer secretomes. Accordingly, proximal lesion sampling, in combination with -omics profiling of TIF is currently considered a promising approach for gaining a greater understanding of the signaling events which underlie BC biology. Furthermore, it is hypothesized that deep proteomic analyses of TIF can lead to the identification of novel protein markers for breast tumor stratification and could form the basis for the development of new blood-based disease diagnostics.

Blood is the most commonly analyzed clinical biospecimen, and it is considered a promising resource for the screening and diagnosis of various pathologies, including cancer. Obtaining a blood sample can also be minimally invasive for a patient [15]. However, despite tremendous efforts, no robust BC-associated protein biomarkers in blood have been implemented into clinical practice, mainly due to difficulties in monitoring tumor heterogeneity and the very high dilution factor of a potential biomarker in blood [15]. Over the past decade, several research labs, including our own, have explored the hypothesis that biomolecules which are aberrantly externalized by breast tumor cells and stromal cells into the tumor interstitium are present at higher, detectable levels within the local tumor space [7,9]. It has also been predicted that cancer-related alterations specific to tumor development can be more prominent in TIFs than in nonmalignant interstitial fluids harvested from the same patient. Therefore, we hypothesize that biomarker signatures, which can be identified from the breast tumor secretome, can be used to establish a tumor-specific, noninvasive blood-based test for monitoring breast malignancy [7,9].

In recent years, we have conducted a number of extensive studies to establish standard operative procedures for the collection and analysis of biomolecular complements in proximal fluids recovered from tumorigenic and normal breast tissues [9,16]. We have also performed several types of quantitative -omics profiling studies with the aim of characterizing levels of cytokines, micro-RNAs, and N-glycans in breast TIF and corresponding serum [17–20]. In addition, we have explored the proteome of proximal breast fluids with gel-based proteomics coupled with mass spectrometry and immunohistochemistry (IHC). In the course of these studies, we have generated several representative proteome datasets, which contain complementary information regarding the secretome of breast tumor lesions, normal mammary glands, and a number of benign breast lesions [6]. We have identified a set of 26 proteins, which are upregulated in breast tumors as compared to normal and benign counterparts, and the expression levels of nine of these proteins were validated in an independent set of 70 malignant breast carcinomas of various grades of atypia [6]. Two-dimensional gel/MALDI-TOF-based proteomics has also been applied by our group to mammary adipose tissues and corresponding interstitial fluids with the aim of investigating the role of adipocytes and related molecular circuitry in the breast tumor microenvironment [21]. However, since gel-based proteomics mainly detect proteins present at moderate to high abundance [22], a more extensive characterization of the breast tumor secretome requires more sensitive tools. The latest developments in quantitative LC-MS/MS, in combination with advanced computational algorithms and bioinformatics, can provide much better proteome coverage, as well as more robust protein identification. Therefore, multiple high-throughput BC proteomics studies have emerged over the last decade [23–25]. However, most studies conducted so far have focused on profiling tumor tissues [26–28] or serum samples [29,30]. To the best of our knowledge, only one pilot study conducted by Raso *et al.* [31] applied tandem mass tags (TMT) quantitative mass spectrometry combined with the MudPIT technique to breast TIF samples, which were isolated from three patients (two patients with infiltrating ductal carcinomas and one patient with a phyllodes tumor) [31]. In the latter study, the authors identified ~ 1700 proteins and demonstrated that this approach could be used to discriminate between normal and tumoral interstitial fluid samples. However, the number of proteins identified was rather limited. Moreover, important tumor characteristics such as subtype, grade, stage, and impact of the tumor microenvironment were not taken into account in this analysis.

In this study, we carried out a detailed quantitative high-throughput LC-MS/MS profiling of the protein complement of interstitial fluid samples, recovered from breast tumors of three main subtypes: luminal, Her2, and TNBC as well as from nonmalignant counterparts obtained from women with untreated BC, who underwent mastectomy at the Copenhagen University Hospital. The aim of this study was to identify a panel of proteins that are externalized from breast tumor components into the local interstitial site and to identify TIF proteins associated with BC tumor status and subtype. Up-to-date bioinformatics methods were applied to a database containing over 8800 proteins to conduct a comprehensive, system-wide, and quantitative characterization of breast tumor secretomes. It is anticipated that this work will lead to the discovery of novel putative serological protein markers to improve our ability to detect and stratify breast malignancies.

2. Methods

2.1. Collection and handling of clinical samples

Fresh tissue samples were collected from patients defined as high-risk according to the guidelines of the Danish Breast Cooperative Group (DBCG, www.dbcg.dk accessed 22.10.2009). Patients had undergone a mastectomy between 2003 and 2012, and samples were collected as part of the Danish Center for Translational Breast Cancer Research program at Copenhagen University Hospital, Denmark. More details on the criteria used to define high-risk cancer patients are reported in our previous publications [17]. Normal samples were collected from nonmalignant areas located at least 5 cm from the tumors. All of the patients presented a unifocal tumor, and none of the patients had a history of breast surgery or had received preoperative treatment (naive samples). Registered clinicopathological data for the patients were retrieved from the Department of Pathology, Rigshospitalet, Copenhagen University Hospital. This study was conducted in compliance with the Helsinki II Declaration and written informed consent was obtained from all participants. The procedures of this study were approved by the Copenhagen and Frederiksberg regional division of the Danish National Committee on Biomedical Research Ethics (KF 01-069/03).

At the time of collection, tissue specimens were divided into two pieces. One piece was stored at −80 °C and subsequently prepared as a formalin-fixed paraffin-embedded (FFPE) sample to undergo histological characterization, tumor subtyping, tumor infiltrating lymphocyte (TIL) scoring, and IHC analysis (see below). The second biopsy piece was placed in PBS at 4 °C within 30–45 min of surgical excision and then was subjected to interstitial fluid recovery (see below).

2.2. IHC of tissue biopsies: histological assessment and tumor subtyping

FFPE blocks prepared from two or three different parts of a tissue specimen were subjected to IHC analysis as previously described [6]. Then, tissue morphology, tumor cell composition, and tumor-stroma percentages were evaluated as previously described [17,32]. The BC subtype of each tissue sample was determined based on ER, PgR, Her2, and Ki67 status, in accordance with St. Gallen International Breast Cancer Guidelines [33]. Three major BC subtypes were identified: luminal, Her2, and TNBC. Due to the small number of samples available, the luminal type tissues analyzed in the present study included both luminal A and B subtypes as a merged category. The cutoffs used for ER, PgR, Her2, and Ki67 for tumor stratification were previously described [19]. The antibodies used in this study (including vendor, origin, dilution, and scoring criteria) are summarized in Table S1. Two researchers (IIG, PSG) blindly reviewed all IHC staining. For each staining, a positive control slide was included in accordance with the manufacturer's instructions. For a negative control, slides were incubated with PBS instead of primary antibody.

Information regarding all of the patients included in this study [i.e., patient age, tumor size, grade, receptor status, stratification of tumor subtype, and proportion of immuno-infiltrate within corresponding biopsies (see below)] is summarized in Table S2. A commercially available tissue microarray (TMA) containing normal tissues from 33 human organs was used (Pantomics, Inc., San Francisco, CA, USA).

2.3. Evaluating and scoring TILs within tumor samples

We examined the most prominent components of the immune microenvironment (i.e., TILs) in the corresponding tumor biopsies used for breast TIF recovery. The number of lymphoid cells present was evaluated with hematoxylin and eosin staining as previously described [17], and with IHC staining with antibodies raised against CD45 (clone 2B11+PD7/26, DAKO) (Table S1). The proportion of TILs in the tissue sections were evaluated in accordance with recommendations of the International TILs Working Group 2014 [32]. Total leukocytes were scored as: 1+ (> 10%), 2+ (10–50%), or 3+ (> 50%). For immune cell population, the expression results were classified as low (neg and 1+) or high (2+ and 3+) (see details in Ref. [17]).

2.4. Interstitial fluid recovery

Tumor interstitial fluid was extracted from fresh breast tumor specimens as previously described [9]. Briefly, 0.1–0.3 g clean tissue was cut into small pieces (~ 1 mm^3 each). After the tissue pieces were washed twice in cold PBS to remove blood and cell debris, they were incubated in PBS at 37 °C in a humidified CO_2 incubator. After 1 h, the samples were centrifuged at 200 ***g*** and 4000 ***g*** for 2 min and 20 min, respectively, both at 4 °C. The supernatants were aspirated, and total protein concentrations were determined with the Bradford assay [34]. The same procedure was used to recover interstitial fluids from lesions enriched with both nonmalignant epithelial and adipose cells, which were dissected approximately 5 cm from a tumor margin. Corresponding normal interstitial fluid (NIF) and fat interstitial fluid (FIF) samples were prepared from 20 and 12 corresponding dissected tissue specimens and pooled for further analysis. To ensure minimum contamination by structural proteins that may originate from cell or tissue lysis, TIF, NIF, and FIF samples and corresponding tissue biopsies were originally subjected to comparative 2D-gel electrophoresis in combination with MS analysis, as previously described [9,16]. The protein component of breast TIF was found to be greatly depleted of structural and nuclear proteins. Quantitation of the ratios of several proteins known to be externalized from tumor tissue to three cytokeratins (CK14, 18, and 19) in both TIF samples and in corresponding whole tumor lysates yielded values that differed by a factor of 10 or more confirming that the release of nonspecific proteins due to cell death is not a significant contributor to TIF.

2.5. LC-MS/MS proteomic experiments

2.5.1. Sample preparation and TMT labeling

Samples were applied to 5 kD cutoff filters (Agilent Technologies, Santa Clara, CA, USA) to perform buffer exchange. Then, 5× the sample volume of 50 mM Hepes buffer (pH 7.6) was added to each sample. The filters were then centrifuged for 20 min at 2000 ***g*** and the flow through was discarded. This step was repeated three times to ensure that a complete exchange was achieved. Protein concentrations of the collected samples were subsequently determined with a DC Protein assay (Bio-Rad, Hercules, CA, USA). The volume of each sample containing 30 µg protein was adjusted to 120 µL with the addition of 50 mM HEPES (pH 7.6). The samples were subsequently denatured at 99 °C for

5 min. Reduction and alkylation were performed by adding 13 μL of 100 mM dithiothreitol and 20 μL of 100 mM iodoacetamide to each sample. Tryptic digestion was performed overnight at 37 °C (trypsin:sample ratio, 1 : 60), followed by TMT labeling, according to the manufacturer's instructions (Thermo Scientific, Waltham, MA, USA). After digestion, 5 μL of each sample (TIF, pooled NIF and FIF) was taken off and run on a short gradient LC-MS/MS for quality control. Pooled NIF and FIF samples were then dissolved in 15 μL of mobile phase A (95% water, 0.1% formic acid) and 1 μL and subjected to LC-MS/MS analysis by using a hybrid Q-Exactive mass spectrometer (Thermo Scientific) as described in 2.5.3. To create an internal standard to link the four TMT sets, a pooled internal standard from TIFs was prepared by taking 4 μg from each sample. TIF samples for in-depth analysis were subjected to TMT labeling according to the manufacturer's instructions (Thermo Scientific). The four TMT-labeled sets were then desalted and cleaned up by applying them to Strata SCX cartridges, according to the manufacturer's instructions (Phenomenex, Torrance, CA, USA), followed by lyophilization. The samples were stored at −20 °C until further analyzed.

2.5.2. Peptide isoelectric focusing and extraction (HiRIEF)

After clean up, the TMT-labeled samples underwent isoelectric focusing (IEF) on four 24-cm, 3.7–4.9 immobilized pH gradient (IPG) strips (GE Healthcare, Uppsala, Sweden). Briefly, samples were rehydrated in 8 M urea with bromophenol blue and 1% Pharmalyte (GE Healthcare), loaded onto IPG strips, and separated, according to previously published protocols [35]. The IPG strips were subsequently subjected to passive elution with MilliQ water into 72 fractions by using an in-house robot. The obtained fractions were dried with a SpeedVac and stored at −20 °C.

2.5.3. LC-MS/MS

For each LC-MS analysis of a HiRIEF fraction, the auto sampler (Ultimate 3000 RSLC System; Thermo Scientific Dionex) dispensed 15 μL of mobile phase A (95% water, 5% dimethyl sulfoxide, 0.1% formic acid) into the corresponding well of a 96-well V-bottom polystyrene microtiter plate (Corning, New York, USA). After mixing the samples added to the plate by aspirating/dispensing a 10-μL volume 10 times, a 7-μL aliquot was injected onto a C18 guard desalting column (Acclaim Pepmap 100, 75 μm × 2 cm, NanoViper, Thermo Scientific). After 5 min with the loading pump at a flow rate of 5 μL·min^{-1}, the 10-port valve switched to analysis mode with the NC pump providing a flow rate of 250 nL·min^{-1} through the guard column. The curved gradient (curve 6 in CHROMELEON software, ThermoFisher Scientific, Waltham, MA, USA) was subsequently applied with 3% mobile phase B (95% acetonitrile, 5% water, 0.1% formic acid) increased to 45% mobile phase B over 50 min, followed by a wash with 99% mobile phase B and re-equilibration. The total LC-MS run time was 74 min. A nano EASY-Spray column (Pepmap RSLC, C18, 2 μm bead size, 100 Å, 75 μm internal diameter, 50 cm length; ThermoFisher Scientific) was used on the nano electrospray ionization (NSI) EASY-Spray source (ThermoFisher Scientific) at 60 °C. Online LC-MS was performed by using a hybrid Q-Exactive mass spectrometer (Thermo Scientific). FTMS master scans with 70 000 resolution and a mass range of 300–1700 *m*/*z* were followed by data-dependent MS/MS at 35 000 resolution for the top five ions by using higher energy collision dissociation (HCD) at 30% normalized collision energy. Precursors were isolated with a 2 *m*/*z* window. Automatic gain control targets were 1e6 for MS1 and 1e5 for MS2. Maximum injection times were 100 ms for MS1 and 450 ms for MS2. The entire duty cycle lasted ~ 2.5 s. Dynamic exclusion was used with 60 s duration. Precursors with an unassigned charge state or a charge state of 1 were excluded. An underfill ratio of 1% was used. MS/MS data were searched by using Sequest HT of the PROTEOME DISCOVERER 1.4 software platform (Thermo Scientific) against the UniProt protein sequence database (140407) with a 1% peptide false discovery rate (FDR) cutoff. A precursor mass tolerance of 10 p.p.m. and product mass tolerances of 0.02 Da were used. Additional settings were as follows: trypsin with 1 missed cleavage; IAA on cysteine, TMT on lysine, N-terminal as fixed modification, oxidation of methionine, and phosphorylation of serine, threonine, or tyrosine as variable modifications. Quantitation of TMT 10-plex reporter ions was performed by Proteome Discoverer (Thermo Scientific) on HCD-FTMS tandem mass spectra by using an integration window tolerance of 20 p.p.m. FDR rate was estimated by using percolator (part of PD 1.4). The mass spectrometry proteomics data obtained have been deposited into the ProteomeXchange Consortium2 via the PRIDE partner repository with the dataset identifier, PXD001686.

2.6. Normalization of samples, data filtering, and batch corrections

We quantified peptides with samples by using a pooled internal standard. Sample ratios were corrected based on mean protein abundance. Normalization was performed in Sequest (Proteome Discoverer User Guide, Software Version 2.2, XCALI-97808, June 2017; Thermo Fisher), after which the data were log2 transformed for bioinformatics analyses. The proteomics data were filtered to remove proteins for which more than 12 samples had missing values (which is the size of the TNBC group) in order to improve the statistical power of the analyses. Missing value imputation with least local squares was performed to infer the remaining missing values before analysis [36]. Since the LC-MS/MS experiments were performed with samples split into four different pools, we explored potential batch effects with a multidimensional scaling (MDS) analysis using Euclidean distance. For visualization purposes, we performed batch correction by using the Combat function [37] implemented in the R-package, SVA, to remove technical pool variation. For differential abundance analysis (DAA), LC-MS/MS pools were used as covariates within the design matrix. We carried out Least Absolute Shrinkage and Selection Operator (LASSO) and random forest (RF) analyses using both batch-corrected and non-batch-corrected data, see Fig. S1 for a comparison of group-wise variances and clustering of samples before and after correction for batch.

2.7. Hierarchical clustering

Hierarchical clustering was applied to batch-corrected data according to Ward's clustering method [38] and algorithm ward.D2. This initial analysis was performed to identify covariates which might contribute to patient stratification and guide the design of DAA (see Results for additional details).

2.8. Evaluation of relevant hits

To evaluate the validity of protein candidates identified as potential serological biomarkers, we investigated their presence or absence in the following relevant publicly available databases and datasets:

-Human Plasma PeptideAtlas [39] https://www.hupo.org/plasma-proteome-project. We downloaded a total of 3529 plasma proteins from https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildInfo?_subtab = 2.

-ExoCarta database [40] http://www.exocarta.org. At the time of our download, this database included entries for 9769 proteins secreted via the exosomal pathway (derived from 286 studies).

-Microparticles from human plasma [41]. This dataset includes 2357 proteins derived from twelve samples.

-A comprehensive dataset of proteins secreted from eleven breast cancer cell lines. This dataset includes 3386 entries and was downloaded from [42].

-MS/MS-based dataset of breast TIF proteins derived from six samples from three patients. Approximately 1000 proteins are available [31,43].

2.9. SignalP and Phobius

FASTA sequences of 6763 proteins (see above) were queried based on UniProt ID, with the R-package, protr [44]. Sequences of 6582 of these entries could be retrieved and were included in our analyses. The remaining proteins (n = 181) were not included due to redundancy or discontinuation of their UniProt ID. Signal peptides were predicted from fasta by using SignalP V. 4.1 [45] (http://www.cbs.dtu.dk/services/SignalP/) and Phobius [46] (http://phobius.sbc.su.se/). For Signal P analysis, the default cutoff of the mean signal peptide score (S_{mean}) was > 0.3 to define a signal.

2.10. Differential abundance analysis

The statistical software, LIMMA (linear models for microarray data) implemented in R, is powerful for small sample sizes due to shrinkage of feature-specific variances [47]. A number of studies have demonstrated the versatility of this software for the analysis of different -omics data, including proteomics data [48,49]. Here, DAA used a corrected P-value (FDR ≤ 0.05) as the cutoff for significance, as well as log-fold change (logFC) ≥ 1 (±5%) or ≤ -1 (±5%) for up- or down-regulated proteins, respectively. In our DAA, the following groups were used for comparisons: (a) all pairwise subtype combinations, (b) hormone receptor status (ER, PgR, Her2), (c) high TIL status (2+ and 3+) versus low TIL status (neg and 1+), and (d) high-grade tumors (3) versus low-grade tumors (1 and 2). Based on the hierarchical clustering, we also included information on LC-MS/MS pools as contrasts in the design matrices for DAA.

2.11. LASSO regression

We performed LASSO regression with tenfold cross-validation, as implemented in the R-package, GLMNET [50]. We used contrasts from the differential expression analysis which yielded significant hits, including (a) BC subtype, (b) TIL status, and (c) hormone receptor status of ER and PgR as a response. The full set of proteins retained after filtering were used as input for the LASSO regression. We ran each LASSO model 10 times with 10 different random seeds and extracted the overlap of selected proteins across runs to a consensus set. For the models with BC subtype, estrogen receptor status, and progesterone status, we split the dataset into a training set and test set, which was used to estimate the model accuracy—See Table S3. We did not split the dataset for LASSO regression with Her2 receptor, degree of TILs and tumor grade, as this resulted in large cross-validation errors. The higher error rates observed for these models were in part related to a highly unbalanced number of samples within each clinicopathological group, especially for the Her2 receptor, in combination with large variances within these homogeneous groups. As such, we can only provide cross-validation errors for these models, see Table S3. N.B we are aware that not splitting the datasets is bias and will produce over-fitted models; however, here it is important to note that we do not use the results of LASSO regression as a stand-alone method for selecting proteins, but as a way of filtering results from differential abundance analysis.

2.12. Random forest

We performed RF with the contrasts from the DAA, which yielded significant hits, including (a) breast cancer subtype; (b) TIL status, and (c) hormone receptor status of ER and PgR as a response. All proteins retained after filtering were used as input for the RF models. For the datasets which converged, defined by a class error ≤ 25%, variable selection was performed by re-running the RF 10 times with different random seeds. The results of each run were overlapped to obtain a consensus of protein classifiers. The R-packages, random Forest and varSelRF [51,52], were used to conduct RF classification. Similarly to the LASSO regression, we split the datasets into training and test sets for models with BC subtype, estrogen receptor status, and progesterone receptor status. However, for the models with Her2, TILs, and tumor grade, we used all data for training due to the large out-of-bag (OOB) errors associated with these. Table S3 contains OBB errors and accuracies for the six RF models.

2.13. Protein-protein interaction networks

All human protein-protein interaction pairs from the STRING database V11.0 [53] (https://string-db.org/cgi/download.pl?sessionId=sm2jwqiNPyyz) were downloaded and used for analysis. In order to get the most comprehensive networks all protein-protein interactions (PPIs) with a score (support/confidence of interaction) above the lower 25th quantile of all scores were kept for analysis. Networks were created for sets of differentially abundant proteins from contrasts with BC subtypes (luminal, Her2, and TNBC). PPIs were retained in a given network if both proteins in the pair were significantly differentially abundant in the same contrast. Networks were visualized using CYTOSCAPE V3.8.0 [54], nodes were colored according to logFC, edges according to the directionality of node pair, and edge width in accordance with score (confidence) of interaction.

2.14. Comparison of quantitative proteomics with IHC scoring

Adjacent heatmaps of protein abundances from high-throughput MS/MS were generated, along with IHC scores obtained from paired tumor tissues (see IHC of tissue biopsies: histological assessment and tumor subtyping). While the latter values were discrete (i.e., 0–3), protein abundances from MS/MS data were continuous. The level and expression pattern for the set of selected proteins that revealed subtype-specific differential TIF abundancy quantified by LC-MS/MS were blindly inspected and compared across the two plots (i.e., IHC vs. LC-MS/MS). IHC scores of the proteins from luminal versus TNBC sample comparisons were evaluated with Fisher's exact test. Her2 subtypes were not included in the test due to the low number of available samples. Significance was defined as a *P*-value < 0.05, and no correction for multiple testing was needed since only 10 tests were performed.

2.15. Protein biomarker sensitivity and specificity

Validation of subtype-specific expression of the TIF proteins identified in this study was performed using the area under the curve (AUC) of receiver operating characteristic (ROC) curves on the independent dataset published [26]. The data from the Tyanova *et al.* [26] publication were generated using the super-SILAC mass spectrometry technique, and the dataset acquired for ROC analysis, contained normalized H/L ratios between the standard and the tissue. Normalization of data had been performed using the MAXQUANT software, see [26] for specifics. Any missing values were imputed by using the llsImpute function of the R-package, pcaMethods [36], with k neighbors values of 4–7 yielding consistent results. Data were log2 transformed to push the protein abundance toward a normal distribution. After imputation of the missing values, and transformation, pROC [55] and nnet [56] were used to generated AUCs with 95% confidence intervals for: (a) each protein individually for relevant pairwise comparisons, (b) proteins combined for the relevant pairwise comparisons, and (c) proteins combined for all three groups together (multinomial model). For the latter setup, we split the data into a training set (which included 2/3 of the data) and a test set (which included 1/3 of the data). A few different seeds were selected as the starting point for splitting in order to evaluate variability among the samples within the dataset and AUC stability. The advantage of this score is that it is independent of threshold selection. Note that AUC varies between 0 and 1 and that an AUC score of 1 means perfect biomarker classification of a subtype.

2.16. Data and script availability

All the scripts, code, and documentation to reproduce our bioinformatic and biostatistical analyses are reported in the GitHub repository https://github.com/ELELAB/Proteomics-TIF. The repository also contains the data and the outputs of the analyses. Moreover, we have published the Cancer BioMarker Prediction Pipeline (CAMPP) [19], a pipeline which may be used to perform most of the analyses.

3. Results and Discussion

3.1. An overview of the TIF proteome

To obtain an initial, yet comprehensive, characterization of the TIF proteome, we performed high-throughput LC-MS/MS quantitative proteome profiling of TIF samples recovered from 35 tumor specimens originating from BC patients. Our aims were to elucidate whether the composition of these TIF proteins: (I) can be used for patient subgroup stratification, (II) is dependent on the composition of the tumor microenvironment,which is known to play an essential role in tumor development, and (III) can prove useful for a putative noninvasive BC test. The experimental and computational workflow for this study is summarized in Fig. 1.

A total of 8855 proteins were identified. At a 1% peptide FDR, this coverage represents approximately six orders of magnitude of dynamic range. After normalization, filtering, and batch correction, 6763 proteins were retained, and these were included in our downstream analysis (see Fig. 1). To the best of our knowledge, this is the most comprehensive dataset of proteins externalized from breast tumors. It is worth noting that the number of proteins comprising the breast TIF dataset is less than the number reported in the largest breast tumor proteomic dataset available to date, which includes 10 135 proteins identified by high-throughput LC-MS/MS screening of whole breast tumor tissue samples [26].

In parallel with the quantitative profiling of proteins externalized from tumor masses, we examined the protein composition of interstitial fluids recovered from far-distant tumor lesions containing a high proportion of nonmalignant mammary epithelium or adipose tissue (i.e., NIF and FIF samples, respectively) (see Methods). Protein spectra from NIF/FIF samples were analyzed with lower analytical depth and proteome coverage, thereby yielding the most abundant proteins externalized. We considered this to represent a baseline of normalcy. Within the pooled NIF and FIF samples, 318 proteins and 391 proteins were detected, respectively (Fig. 2A). A total of 155 proteins were common to all of the TIF, NIF, and FIF samples, and this subset represents approximately 50% of all the NIF and FIF proteins identified. In addition, 53 proteins are shared between the TIF and FIF samples, while 25 proteins are shared between the TIF and NIF samples. Meanwhile, 260 proteins in the FIF and/or NIF samples were not identified in any tumor fluids (Fig. 2A, Table S4).

To gain further insight into the secretion potential of proteins identified in our TIF samples, we compared our dataset to secreted protein entries in the five most representative and relevant protein datasets and databases currently available:

-The Human Plasma PeptideAtlas database [39] is the most comprehensive resource of proteins present in human blood, independent of origin and type of secretion. A comparative analysis between TIF and plasma protein complements has the potential to identify secreted proteins that enter the blood circulation, and thus, may assist in prioritizing candidates for further studies. In total, the concatenation of datasets from this database yielded 3529 proteins for analysis.

-The ExoCarta database [40] is the largest database of exosomal proteins, containing more than 40 000 protein entries (9769 proteins). Exosomes are small membranous vesicles (30–150 nm in diameter), which are released by a variety of cells into the extracellular environment. Exosomes represent a nonclassical, vesicle-mediated secretory pathway for the transport and exchange of a variety of biomolecules between cells as a means of communication. Thus, a comparison of TIF with ExoCarta enabled us to identify proteins which are most likely externalized into breast TIF through exosome-associated secretion pathways.

-A dataset of proteins associated with circulating human plasma microparticles (MPs) [41]. Plasma and other bodily fluids contain membranous MPs, which

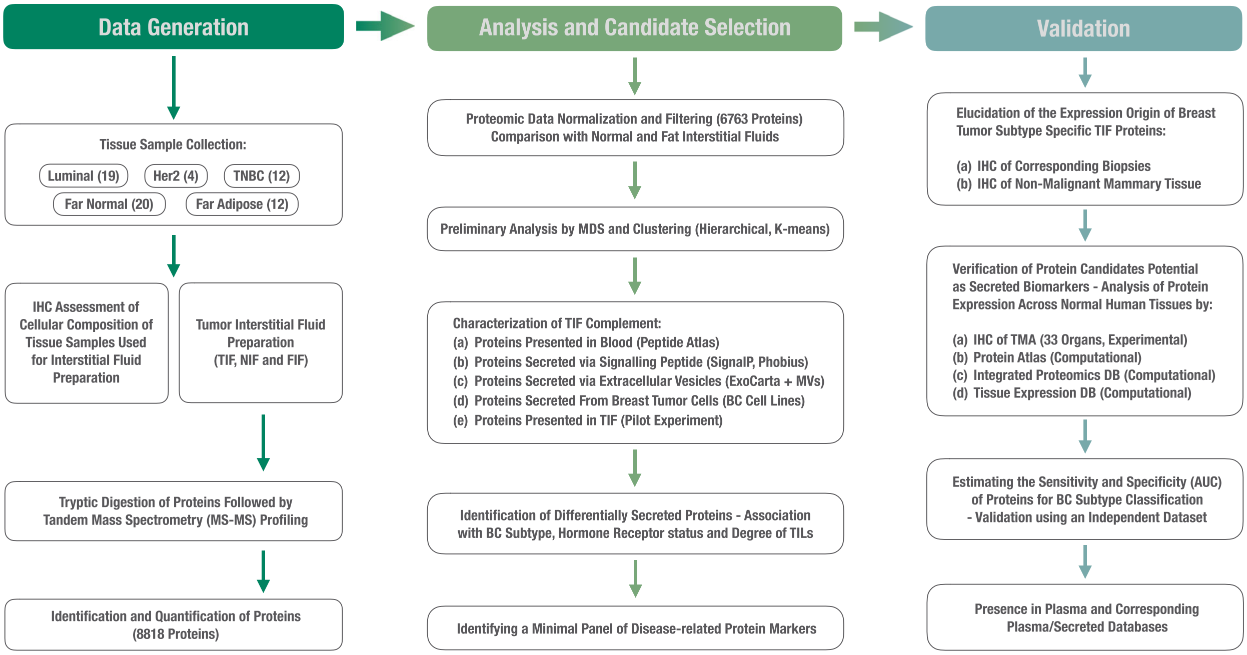

Data Generation
Analysis and Candidate Selection
Validation
Tissue Sample Collection:
Luminal (19)
Her2 (4)
TNBC (12)
Far Normal (20)
Far Adipose (12)
IHC Assessment of Cellular Composition of Tissue Samples Used for Interstitial Fluid Preparation
Tumor Interstitial Fluid Preparation (TIF, NIF and FIF)
Tryptic Digestion of Proteins Followed by Tandem Mass Spectrometry (MS-MS) Profiling
Identification and Quantification of Proteins (8818 Proteins)
Proteomic Data Normalization and Filtering (6763 Proteins) Comparison with Normal and Fat Interstitial Fluids
Preliminary Analysis by MDS and Clustering (Hierarchical, K-means)
Characterization of TIF Complement:
(a) Proteins Presented in Blood (Peptide Atlas)
(b) Proteins Secreted via Signalling Peptide (SignalP, Phobius)
(c) Proteins Secreted via Extracellular Vesicles (ExoCarta + MVs)
(d) Proteins Secreted From Breast Tumor Cells (BC Cell Lines)
(e) Proteins Presented in TIF (Pilot Experiment)
Identification of Differentially Secreted Proteins - Association with BC Subtype, Hormone Receptor status and Degree of TILs
Identifying a Minimal Panel of Disease-related Protein Markers
Elucidation of the Expression Origin of Breast Tumor Subtype Specific TIF Proteins:
(a) IHC of Corresponding Biopsies
(b) IHC of Non-Malignant Mammary Tissue
Verification of Protein Candidates Potential as Secreted Biomarkers - Analysis of Protein Expression Across Normal Human Tissues by:
(a) IHC of TMA (33 Organs, Experimental)
(b) Protein Atlas (Computational)
(c) Integrated Proteomics DB (Computational)
(d) Tissue Expression DB (Computational)
Estimating the Sensitivity and Specificity (AUC) of Proteins for BC Subtype Classification - Validation using an Independent Dataset
Presence in Plasma and Corresponding Plasma/Secreted Databases

Fig. 1. A flow chart of the experimental and computational workflow for this study. Number of tumor interstitial samples used for analysis, sample curation, and data filtering are summarized. Methodological steps in the bioinformatic/biostatistical analyses performed, along with validation (both experimental and in relation to available literature) of candidate proteins, are also presented.

are thought to be derived from various cell types, including BC cells [57]. MPs differ from cellular exosomes in size and cellular origin, with the latter originating from intracellular multi-vesicular bodies. The importance of MPs as mediators of cellular signaling is supported by recent data which demonstrate that MPs serve as vectors in the intercellular transfer of functional proteins and nucleic acids, and also in drug sequestration [58]. Moreover, an important role for MPs in facilitating evasion of cancer cell immune surveillance has been demonstrated [59]. In total, 2357 proteins associated with MPs are currently available for analysis.

- A comprehensive dataset of secreted proteins from 11 BC cell lines with different origins that are representative of different stages of BC development [42]. This is the most comprehensive dataset of proteins detected in conditioned media of BC cells, and it represents the major BC subtypes. This dataset encompasses 3386 proteins.
- The only high-throughput LC-MS/MS-based pilot study of breast TIF protein composition published to date [31]. This dataset contains approximately 1000 proteins and derived from an analysis of breast TIF samples from three healthy individuals and three patients with tumors.

The results from comparing these datasets are presented in Fig. 2B. Overlaps (i.e., intersections) between the proteins found in the different datasets are indicated with vertical bars. In total, 4830 out of the 6763 proteins (71.4%) detected in the TIF samples were included in at least one of the datasets used for comparison. The main overlap observed involved exosomal proteins present in the ExoCarta database, with 3203 proteins found to be present in TIF and in exosomes. This result indicates that many of the TIF proteins (~ 50%) are likely externalized through exosomal signaling pathways. In addition, 2567 proteins were present in TIF and secreted from BC cell lines, while 2230 proteins and 1599 proteins were found to be shared between TIF and plasma or plasma MPs, respectively. Taken together, these results highlight the secretory nature of the TIF proteome complement. Compared with the TIF dataset published previously by Raso *et al.* [31], 775 proteins are common to the present breast cancer TIF dataset. This overlap represents approximately 84% of the 924 proteins identified by this group. These results emphasize the validity of our experimental framework and they also indicate that high compliance exists between the data obtained in both of these studies on TIF.

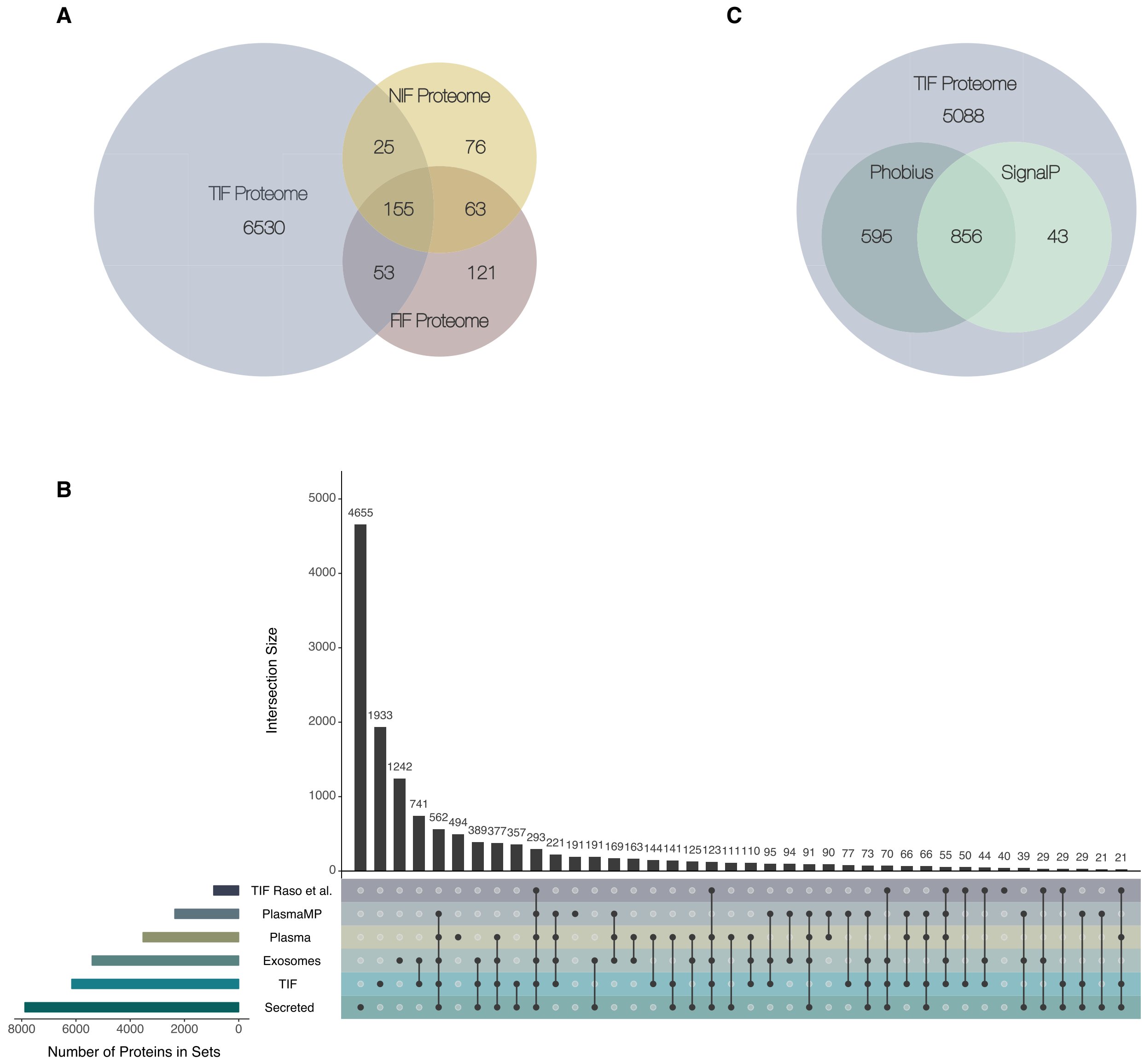


Fig. 2. Overall characterization of the TIF dataset. Both a comparison of secretome-related datasets and an evaluation of the secretion pathway were performed. (A) Venn diagram of the overlap between proteins identified within TIF samples and within pooled NIF and FIF samples. (B) An Upset plot illustrates the overlap of the final set of 6763 TIF proteins, 6066 unique gene symbols, included in further analyses with: (a) human plasma—PeptideAtlas database proteins [39]; (b) human exosomal proteins from the ExoCarta database (http://www.exocarta.org; [40] (c) a dataset of proteins associated with circulating human plasma microparticles [41]; (d) a secretome of 11 BC cell lines of different origins [42]; and (e) an interstitial fluid BC protein complement published by Raso *et al.* [31]. Colors denote the individual protein datasets. Horizontal bars indicate the size of each dataset; vertical bars indicate the number of proteins shared between all combinations of the six sets. (C) Venn diagram showing how many TIF proteins are predicted by SignalP [45,103] and/or Phobius [46] to encompass a signal peptide.

Additionally, an important observation is that our TIF protein complement contains 1933 unique entities which do not overlap with any of the five datasets used for comparison (Fig. 2B). These proteins were potentially identified due to the depth of our analysis. A subset of these proteins may be specific to breast TIF and may originate from malignant cells and/or cells present in the tumor microenvironment. However, we cannot exclude the possibility that some of these proteins are presented in TIF as a result of partial cellular or tissue lysis during sample preparation.

After comparing our dataset to these relevant databases and datasets, we used the prediction tools, SignalP [45] and Phobius [46], to segregate classically secreted proteins within the TIF proteome from proteins externalized via nonclassical pathway(s). With the use of these two tools, the presence of signal peptides within TIF proteins was predicted. While these signal peptides may indicate which proteins are targeted for the secretory pathway, they may not necessarily drive secretion. Among the TIF proteins, 6582 out of 6763 had FASTA sequences available for analysis. According to SignalP analysis, a total of 899 TIF proteins are predicted to contain a signaling peptide. In comparison, Phobius predicted that 1451 proteins contain a signaling peptide. The overlap between the two predictors was quite high, with 856 commonly predicted secreted proteins present in the TIF proteome (Fig. 2C). However, this set of proteins only accounted for 10% of the TIF proteome (Fig. 2C), suggesting that a significant proportion of TIF proteins undergo nonclassical secretion (i.e., though membrane pore formations or via specialized secretory autophagosomes [60]).

3.2. TIF proteins distinguish low-grade versus high-grade tumors and are associated with different levels of TILs

To evaluate the potential for TIF proteome profiles to be segregated according to BC subtype, we stratified available tumor samples into three major subtype-specific groups: luminal, Her2, and TNBC. Corresponding plots with multidimensional scaling (MDS) revealed considerable segregation of TIF proteomes across all three tumor subtypes (Fig. 3). A slight overlap of two samples, a Her2- and a TBNC-subtype with the TBNC and luminal groups, respectively, was observed (Fig. 3). However, one Her2 sample appeared to be a very clear outlier (Fig. 3, indicated with an arrow). A morphological analysis of the corresponding tumor biopsy revealed the presence of apocrine dysplasia(s) with multiple cyst structures scattered within the tumor mass. Apocrine dysplasia tumors have a high secretion potential [6,61], and this characteristic is consistent with the unique TIF profile of this particular Her2 tumor sample. Therefore, we removed this sample from subsequent analyses.

To gain insight into which clinical and morphological covariates may have an impact on the externalized protein patterns within the interstitial space, unsupervised hierarchical clustering was performed (see Hierarchical clustering). As shown in the dendrogram in Fig. 4, both hormone receptor status and level of TILs within the corresponding biopsies were able to stratify the TIF proteomes of the BC patients examined. Hierarchical clustering revealed two main clusters of TIF samples, Cluster 1 and Cluster 2. Cluster 1 almost exclusively encompassed ER+ luminal samples (93% of samples within the cluster), the majority of which were positive for PgR, and from lower grade tumors approximately, 71% of samples. Cluster 1 samples were more often characterized by a low level of immune cell infiltration (65% within-cluster and 69% across clusters), although this pattern was more subtle. In contrast, Cluster 2 mainly consisted of ER^-/PgR^- samples, 70%, and 75%, respectively, originating from the TNBC subtype (55% of samples and 92% of all TNBCs across clusters) and Her2 specimens—15% within-cluster and 100% of Her2 samples. Most of the samples within Cluster 2 were enriched in TILs and originated from high-grade tumors, 80% and 75%, respectively (Fig. 4). These findings are in agreement with the results of our recent publications on cytokine and N-glycan profiling of breast tumor interstitial fluids, which demonstrate that high-grade Her2, and especially TNBC, tumors exhibit a high level of TILs [17,19]. Meanwhile, neither the percentage of malignant cells in the tumor samples, nor patient age, stratified the TIF proteomes. It should be noted that while

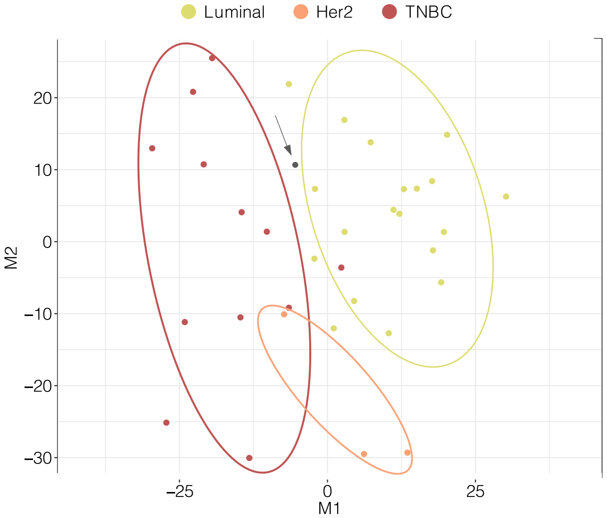


Fig. 3. Multidimensional scaling plot of 35 BC TIF samples according to subtype based on protein abundances. The *x*-axis and *y*-axis denote multidimensional scaling components 1 (M1) and 2 (M2), respectively, which best retain the distance relationship (squared Euclidian) between the samples in two-dimensional space. The single gray dot indicated with an arrow represents a Her2 outlier sample (apocrine dysplasia(s) with multiple cyst structures), which was excluded from further analysis.

there was a propensity toward the clustering of samples with similar clinicopathological characteristics. Figure 4 also highlighted sample heterogeneity, a well-established issue within the field of breast cancer research. The distribution of different clinicopathological groups within clusters and across clusters may be found in Table S5.

3.3. Identification of differentially abundant (DA) proteins associated with BC subtypes, hormone receptor status, and degree of TILs

Next, we performed differential abundant analysis (DAA) to identify which secreted proteins are able to discriminate between TIFs originating from the three major BC subtypes, tumors of different grades, and tumors with varying degrees of infiltrating lymphocytes (Fig. 4). Specifically, DAA was applied to the following group comparisons: (a) all pairwise subtype combinations (i.e., luminal vs. Her2, luminal vs. TNBC, and Her2 vs. TNBC); (b) ER^+ versus ER^-, (c) PgR^+ versus PgR^-, (d) high Her2 ($3^+/2^+$) versus low Her2 (1+/0), (e) high ($3^+/2^+$) versus low ($1^+/0$) TIL status, and (f) high-grade tumors (i.e., GR3) versus low-grade tumors (GR2/1).

From these six comparisons, a total of 174 DA proteins (FDR < 0.05 and logFC > 1 or < −1) were identified. Among these, 151 proteins were associated with BC subtypes, 64 proteins were associated with ER/PgR/Her2 status, and 15 proteins were associated with TIL scoring. Four of these proteins, ADIRF, S100A9 (both in luminal vs. TNBC), HSPB1, and POSTN (TIL associated), were found in the NIF/FIF background datasets. Despite the observed partitioning of samples based on tumor grade in Fig. 4, we did not identify any DA proteins when we compared TIF samples from high- versus low-grade tumors. This result may be due to the almost total confounding of tumor grade with TIL status. Thus, when we corrected for TILs as a confounder in the statistical analysis, we lost

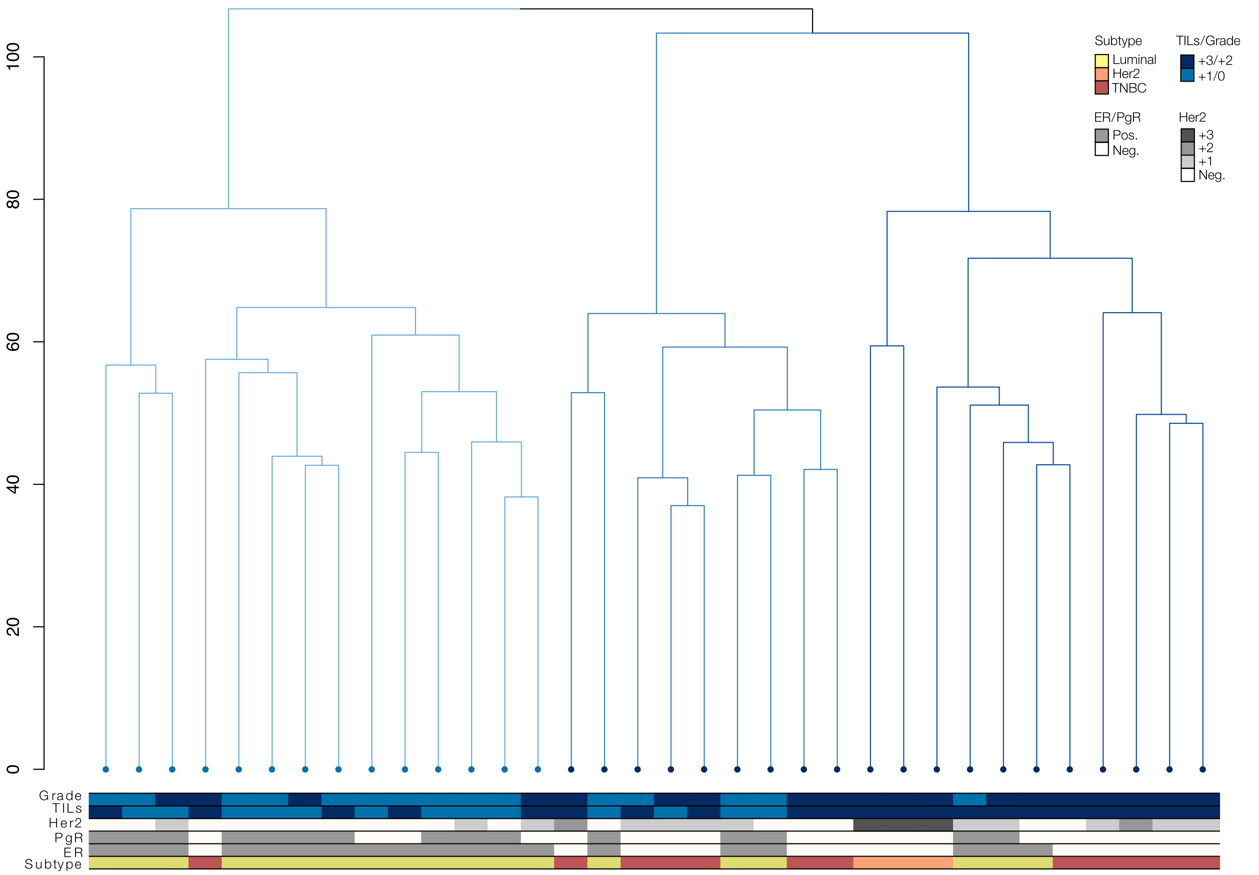


Fig. 4. Dendrogram clustering of 34 BC samples based on TIF protein abundances. Coloring of bars as gray (positive) or white (negative) labels ER/PgR/Her2 status of the samples. Subtype, TIL grade, ER/PgR, and Her2 status are colored as indicated. TIL samples were stratified according to recommendations of the International TILs Working Group 2014 [32] and as described in Methods.

most of the biological variance between tumor grades. The complete list of DA genes used for each comparison, as well as the directionality in the comparison (i.e., up- vs. downregulation), are reported in Table S6. Individual set-wise list of DAA results, including test-statistics, *P*-values and logFCs, may be found in Table S7. Additionally, set-wise heatmaps of differentially abundant proteins from each comparison are presented in Fig. S2.

Among the 151 proteins which exhibited differential abundance in the TIF samples from three BC subtypes, we identified 60 unique DA proteins when we compared Her2 and luminal samples, 66 unique DA proteins when we compared luminal and TNBC samples, and only eight DA proteins when we compared Her2 and TNBC samples. The majority of DA proteins were unique to one subtype, although a few overlapped across pairs of subtypes (Fig. 5, black vertical bars across subtype-wise comparison). Five of the proteins (BCAM, COPS9, DNJC12, TCEAL3, and ZSCAN18) revealed higher abundances in the TIF samples with luminal origin, compared to both the Her2-enriched and TNBC TIF samples. Meanwhile, there were 11 proteins (BCAS1, CDK12, CRYM, ERBB2, FDFT1, GRB7, HMGCS1, IDI1, MIEN1, SRCIN1, and VPS13B), which were enriched in the Her2 TIF samples compared to the luminal and TNBC TIF samples. Conversely, two proteins (LRMDA/C10orf11, TMEM51) exhibited depleted abundances in the Her2 TIF samples compared to the luminal and TNBC TIF samples, while one protein (PADI2) exhibited low abundance in the luminal TIF samples compared to both the Her2 and TNBC TIF samples. Finally, the 9 + 4 proteins identified as DA in the comparisons of high versus low Her2 and PgR^+ versus PgR^- TIF samples, respectively, were redundant, with those identified in the subtype contrasts. Similarly, the majority of the 51 proteins found to be DA between the ER^+ versus ER^- TIF samples were also DA between the luminal versus Her2 and/or TNBC TIF samples.

3.4. Identifying a minimal subset of disease-related proteins: A consensus approach (DA analysis, LASSO regression, and random forest)

Our DAA returned a relatively large number of proteins. Therefore, we further employed two independent approaches to pinpoint the most prominent set of protein candidates, which would have the potential to discriminate different subgroups of breast tumors. These analyses included random forest (RF) classification and least absolute shrinkage and selection operator regression (LASSO) with leave-one-out k-fold cross-validation (see Methods).

For LASSO regression with BC subtypes (without Her2), estrogen, and progesterone receptors, we split the sets into training and test sets; however, for regression analysis with Her2 status, degree of TILs and tumor grade we kept all samples for training, as these models returned large cross-validation (CV) errors even when all samples were included in the model. The larger CV errors observed for these regression models (~ 25%) were somewhat attributed to an unequal distribution of classes (subgroups), especially for Her2 status, a high degree of sample heterogeneity within these clinicopathological groups, in addition to our small sample size. Table S3 contains the variables returned from each LASSO, ordered according to the variables weight in the model, in addition to cross-validation errors and accuracies with confidence intervals for regressions where data could be split into training and test sets. Average cross-validation errors for models with subtype, ER, and PgR ranged from 7% to 10%, while accuracies estimated from the test sets were ~ 0.89 (CI: 0.5–0.99) for all three.

Similarly to regression analysis, random forest returned large class errors (> 25%) and poor convergence for models with Her2 receptor status, degree of TILs, and tumor grade. In contrast, convergence for RFs with BC subtype (without Her2), ER, and PgR status was okay (~ 15% misclassified), with accuracies of 94%, 89%, and 78%, respectively. See Table S3 for out-of-bag errors, class errors, and accuracies. Generally, RF and LASSO with Her2, TILs, and grade as outcome, had poor overlap of selected variables, and those which were identified by both approaches were weighted quite differently. This observation is supported by the large OOB and CV errors associated with these models.

We derived a list of candidates for each comparison (i.e., BC subtypes, ER/PgR/Her2 status, and TIL level) where proteins were identified by at least two out of the three methods applied (Fig. 6 and Table S8). Both RF classification and LASSO regression returned five of the original 15 DA proteins detected in the TIF samples with high versus low TILs, here among: COL5A3, HSPB1, GPC1, MAPT, and SPATA18 (Table S8). HSPB1 was excluded from the final list of proteins because it was detected in NIF/FIF samples (Table S4). RF and LASSO regression also returned protein GPC1, which had significant adjusted *P*-values in the differential expression analysis but fell short of the logFC cutoff (logFCs −0.86 and 0.44, respectively). Accordingly, GPC1 was also excluded from the minimal consensus subset of proteins. Interestingly, all

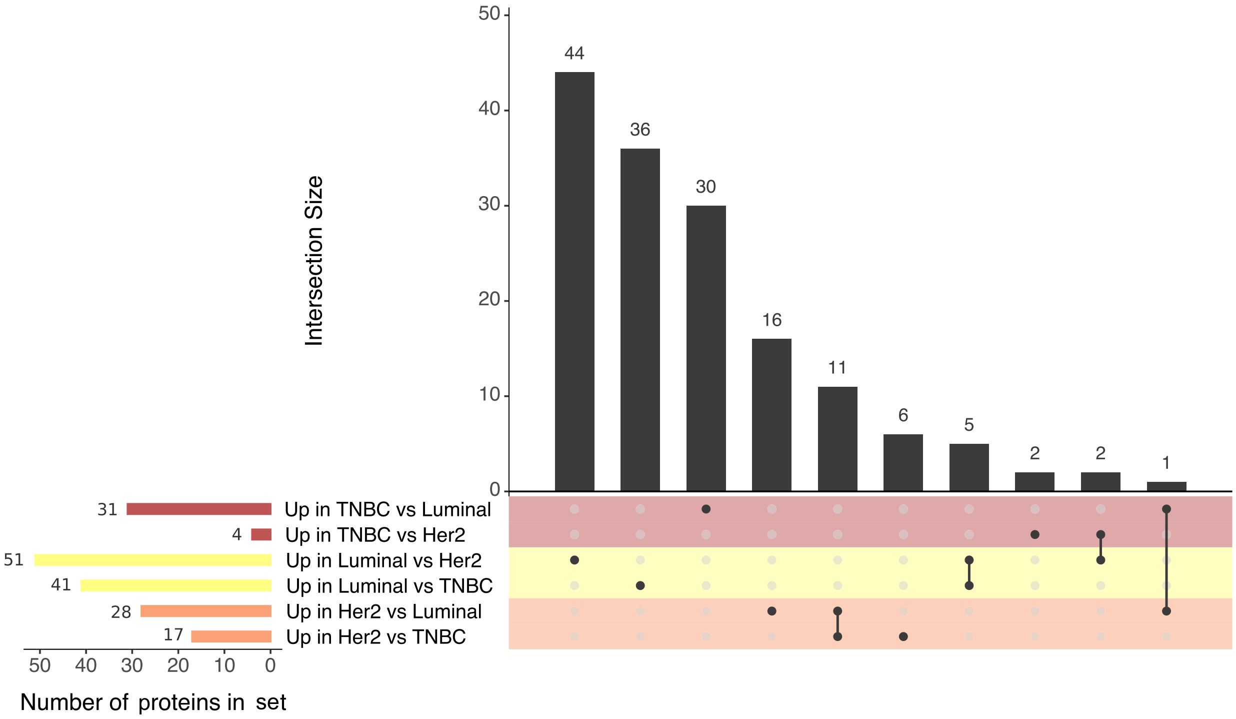


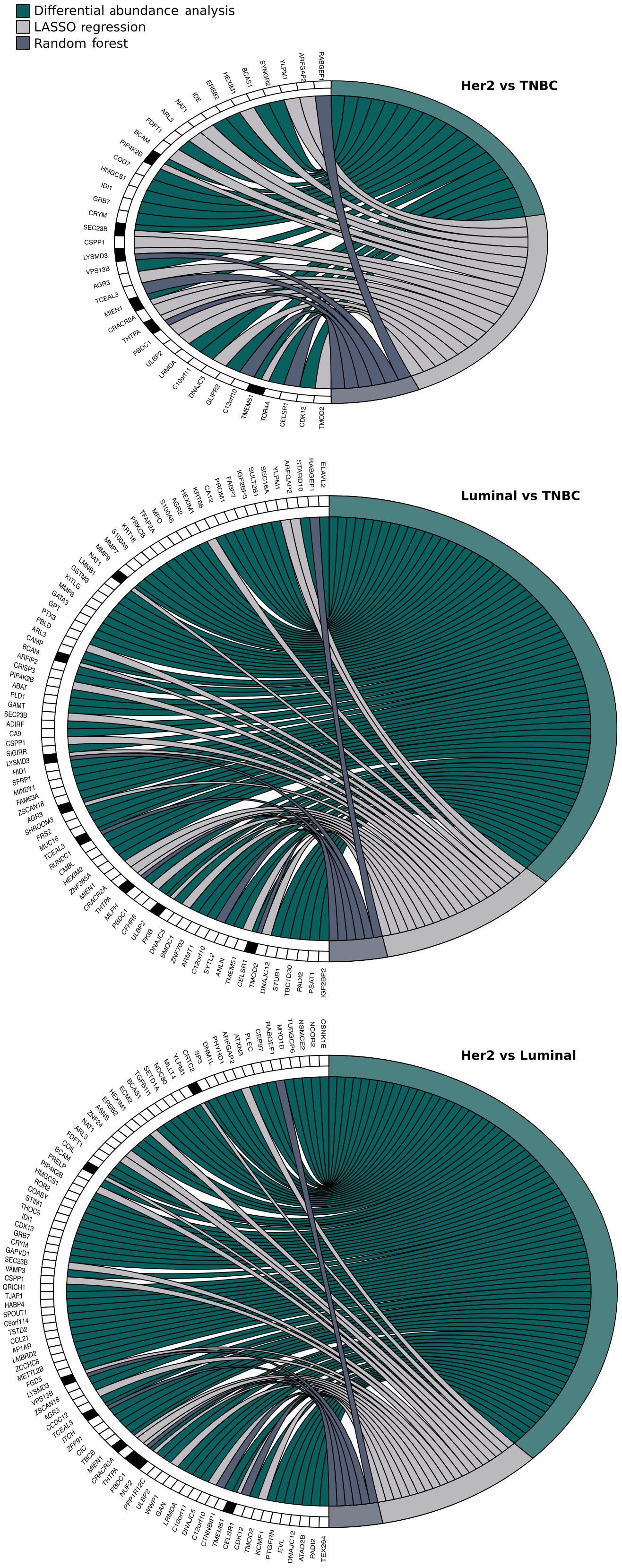
Fig. 5. Upset plot showing the overlap of TIF proteins which exhibited DA according to BC subtype. Horizontal bars denote the size of each dataset; vertical bars indicate the number of proteins shared between all combinations of the six sets. Colors are used to indicate comparisons made with TNBC (red), Her2 (orange), or luminal (yellow) subtype samples.

three TIL-associated proteins exhibited an inverse directionality of correlation between their protein abundance in TIF samples and the number of TILs observed in matched tumors. Decreased abundances of COL5A3, MAPT, and SPATA18 were detected in TIF samples derived from tumor specimens with a high level of leukocyte infiltrate (+3/+2). These results imply that metabolites produced by white blood cells that penetrate tumors may suppress the production and/or secretion of these proteins by neoplastic cells. Interestingly, recent evidence suggests that one of these identified proteins, SPATA18, plays an important role in suppressing the progression of breast and colorectal cancers in a hypoxic tumor microenvironment [62–64]. Meanwhile, lower expression of MAPT in TNBC tumors (which are often enriched with TILs), and not in other BC subtypes, has been observed [65]. Darlix *et al.* [66] (and references within) have also recently demonstrated a prognostic value for the serum level of MAPT in metastatic BC patients, as well as its correlation with brain metastases).

When hormone receptor status was compared with the results of LASSO regression and RF classification, 10 proteins were identified. Six proteins were associated with estrogen status (CELSR1, SEC23B, THTPA, TCEAL3, ZNF703, ZSCAN18), two proteins were related to progesterone status (BCAM, COMP), and three proteins were related to Her2 status (ERBB2, SP3, ZNF24) (Table S8).

Out of the 24 proteins identified as a minimal subset of disease-related proteins (Table S8), we selected 10 proteins, namely AGR3, BCAM, CELSR1, MIEN1, NAT1, PIP4K2B, SEC23B, THTPA, TMEM51, and ULBP2, which represent potential candidates for segregating TIFs from different BC subtypes. The selection was based on the commercial availability of antibodies that met the criterion for their sensitivity and specificity (Methods: IHC of tissue biopsies: histological assessment and tumor sub-

Fig. 6. Consensus analysis of DA proteins according to BC subtype. Circle plots represent the consensus observed across DAA (green), LASSO (light gray), and RF (dark gray) analyses. Genes identified with each approach are noted on the left-hand side of each plot. The black boxes indicate genes identified with more than one method.


🟩 Differential abundance analysis
▨ LASSO regression
■ Random forest
Her2 vs TNBC
RABGEF1
ARFGAP2
YLPM1
SYNGR2
BCAS1
HEXIM1
ERBB2
IDE
NAT1
ARL3
FDFT1
BCAM
PIP4K2B
COG7
HMGCS1
IDI1
GRB7
CRYM
SEC23B
CSPP1
LYSMD3
VPS13B
AGR3
TCEAL3
MIEN1
CRACR2A
THTPA
PBDC1
ULBP2
LRMDA
C10orf11
DNAJC5
GLIPR2
C12orf10
TMEM51
TOR4A
CELSR1
CDK12
TMOD2
Luminal vs TNBC
ELAVL2
RABGEF1
STARD10
ARFGAP2
YLPM1
SEC16A
SULT2B1
IGF2BP3
FABP7
PROM1
CA12
KRT86
HEXIM1
AGR2
S100A8
MPO
TFAP2A
PRKCB
KRT18
S100A9
MMP7
MMP9
NAT1
LMNB1
GSTM3
KITLG
MMP8
GATA3
GPT
PTX3
PBLD
ARL3
CAMP
BCAM
ARFIP2
CRISP3
PIP4K2B
ABAT
PLD1
GAMT
SEC23B
ADIRF
CA9
CSPP1
SIGIRR
LYSMD3
HID1
SFRP1
MINDY1
FAM63A
ZSCAN18
AGR3
SHROOM3
FRS2
MUC16
TCEAL3
RUNDC1
CMBL
HEXIM2
ZNF385A
MIEN1
CRACR2A
THTPA
MLPH
PBDC1
CFHR5
ULBP2
PKIB
DNAJC5
SMOC1
ZNF703
ARMT1
C12orf10
SYTL2
ANLN
TMEM51
CELSR1
TMOD2
DNAJC12
STUB1
TBC1D30
PADI2
PSAT1
IGF2BP2
Her2 vs Luminal
CSNK1E
NCOR2
NSMCE2
TUBGCP6
MYO1B
RABGEF1
CEP97
PLEC
ATXN3
ARFGAP2
PHYHD1
DNM1L
SP3
CRTC2
YLPM1
MLLT4
NDC80
SETD1A
TGFB1I1
BCAS1
ECM2
HEXIM1
ERBB2
ASNS
ZNF24
NAT1
ARL3
FDFT1
COIL
BCAM
PRELP
PIP4K2B
HMGCS1
ROR2
COASY
STIM1
THOC5
IDI1
CDK13
GRB7
CRYM
GAPVD1
SEC23B
VAMP3
CSPP1
QRICH1
TJAP1
HABP4
SPOUT1
C9orf114
TSTD2
CCL21
AP1AR
LMBRD2
ZCCHC8
METTL2B
FGD5
LYSMD3
VPS13B
ZSCAN18
AGR3
CCDC12
TCEAL3
ITCH
ZFP91
CIC
TBCB
MIEN1
CRACR2A
THTPA
PBDC1
NUF2
PPP1R12C
ULBP2
WWP1
GAN
LRMDA
C10orf11
DNAJC5
C12orf10
CTNNBIP1
TMEM51
CELSR1
CDK12
TMOD2
KCMF1
PTGFRN
EVL
DNAJC12
ATAD2B
PADI2
TEX264

typing, Table S1) required for further confirmatory studies by IHC (see TIF subtype-specific protein signatures: Expression origin and externalization pattern below).

Two of these proteins were identified in contrast with estrogen status (CELSR1, THTPA), while one was associated with progesterone status (BCAM). Based on available datasets and literature, we evaluated whether these proteins are present in human plasma, in exosomes, and/or among proteins secreted by breast cancer cell lines of different origins. In addition, we checked whether these proteins may be externalized through classical pathways based on the predicted presence of secretory signal peptides (SignalP or Phobious). Finally, to ensure that these candidate proteins originate from malignant cells and not from the normal epithelium and/or adipose cells, we investigated their presence within pooled NIF and FIF samples. None of the NIF/FIF proteins were included in the consensus set, thereby confirming the potential of this subset to serve as BC-specific biomarkers (Table S8).

To estimate the potential significance of DA TIF proteins associated with BC subtypes, we compared them to the gene-set of the well-established BC classifier, PAM50. The PAM50 signature is one of the most powerful predictive and prognostic classifiers [67,68] currently implemented in the clinic [approved by the US FDA in September 2013]. The PAM50 signature is characterized by expression levels of 50 transcripts, including mostly hormone receptors, proliferation-related genes, and genes exhibiting myoepithelial and basal features. When we compared the DA proteins associated with BC subtypes (a total of 151 proteins, see TIF proteins distinguish low-grade versus high-grade tumors and are associated with different levels of TILs) to the PAM50 classifier set of genes, eight proteins (MLPH, ANLN, ERBB2, GRB7, NAT1, SFRP1, NUF2/CDCA1, and NDC80/KNTC2) were identified (Table S9). We subsequently compared the abundance directionality (up or down) of these eight TIF proteins with intratumor mRNA levels for pairwise subtype contrasts [69] and observed full concordance between mRNA expression levels and levels of the corresponding proteins (Table S9). Furthermore, only NAT1 from a minimal subset of TIF candidates (n = 14) identified by using a consensus approach (i.e., DAA, elastic-net regression, and machine learning) matched the corresponding transcripts in PAM50. Tyanova *et al.* [26] previously reported a higher number of proteins (41) from BC tumor subtypes that matched the 50 transcripts of PAM50; yet only 21 had quantitative data available from more than 70 samples for which PAM50 genes enabled partial segregation of classical subtypes at the protein level. In particular, the authors highlighted four well-described proteins for differentiating breast cancer subtypes, namely Her2, Grb7, FOXA1, and MLPH, which were clearly selected in the PAM50 and proteomic signatures. It should be noted that two of these proteins, Grb7 and MLPH, are present in our 8-protein set, which overlaps with PAM50.

3.5. Protein-protein interaction networks

To assess whether the interplay between differentially abundant proteins from contrasts with BC subtype, we constructed protein-protein interaction (PPI) networks using the STRING database [53] and visualized with CYTOSCAPE [54] (see Methods and Table S10). The results of network analysis with DA proteins from comparisons with BC subtypes are shown in Fig. 7. Specifically, we were interested in whether protein candidates selected using the consensus approach (DAA, LASSO, and RF) were highly interconnected within the PPIs, indicating a regulatory role or potentially a driver role, or if they were leaf nodes.

The network of DA TIF proteins from the Her2 vs TNBC comparison was small and almost completely redundant with the Her2 vs luminal network—all nodes in the Her2 vs TNBC network were upregulated in Her2, see Fig. 7. All hub nodes are marked in Fig. 7 with a star. Hub nodes from the Her2 vs luminal network included ERBB2, ATAD2B both upregulated, and NCOR2, SETD1A, downregulated. ERBB2 and ATAD2B, were not hub proteins in the same subnetwork but in each their own, connecting 30 and 16 proteins, respectively. High levels of ATAD2(B) are known to be associated with increased cell survival, tumor cell migration, and a poor prognosis in patients with breast cancer, supported by multiple studies [70,71], and in accordance with this, TIFs from Her2 samples had a greater abundance of this protein as compared to luminal samples.

The hub protein SETD1A, which was upregulated in luminal compared to Her2 samples, is a component of the histone methyltransferase (HMT) complex. SETD1A has been shown to be involved in the regulation of mitotic gene expression, and the knockdown of this gene leads to cellular senescence [72]. SETD1A is amplified in 7–24% of breast cancers and was found to promote survival and migration of ER-positive breast cancers, specifically [73].

Inversely to SETD1A, a high level of the hub protein NCOR2 (also up in luminal vs Her2), may be associated with increased metastasis-free-survival in

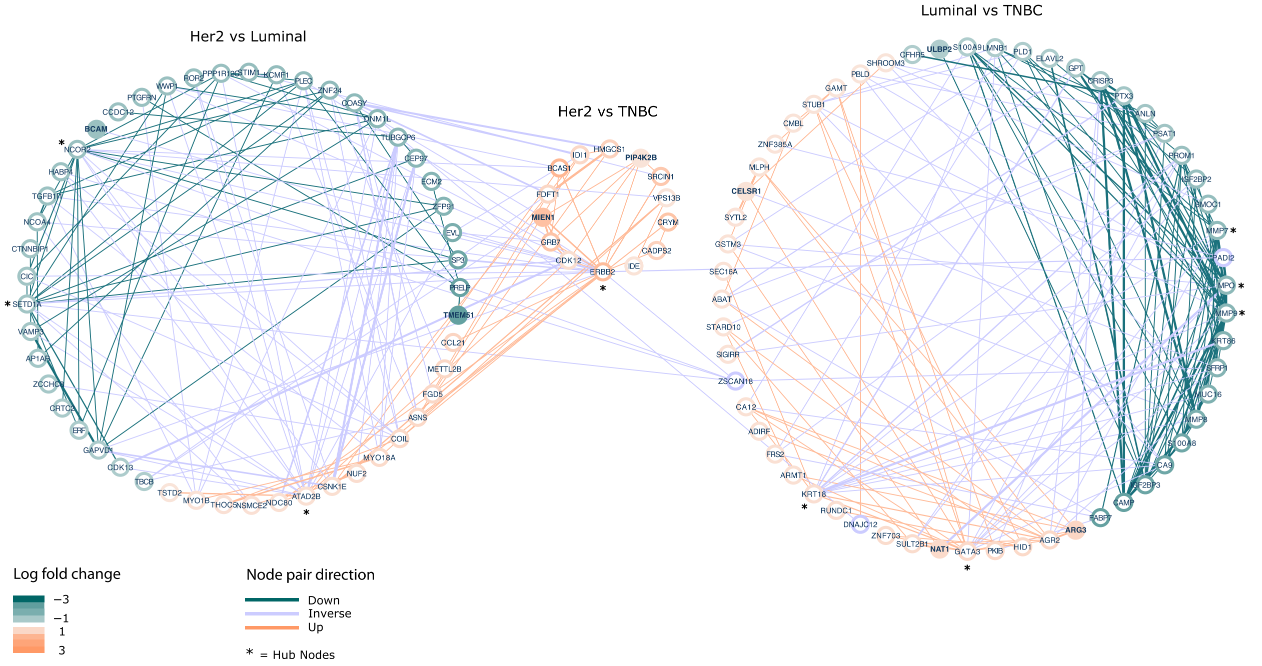


Fig. 7. Protein-protein interaction networks based on DA proteins from the comparison of BC subtypes (luminal, Her2, and TNBC). The plot contains three networks; (I) Her2 vs luminal, (II) Her2 vs TNBC and (III) luminal vs TNBC. Nodes (proteins) are colored according to logFC: green < −1 and orange > 1. Edges are colored based on directionality of node pair: green = both nodes down, orange = both nodes up, purple = inverse directionality of nodes. The width of edges denote the node pair interaction score (support) from STRING, and ranges from 0.25 to 1.0. Purple nodes are a part of more than one network and have opposite directionality in the networks.

BC patients with ER-positive tumors [74]. NCOR2 is an established tumor suppressor gene in prostate cancer and a hallmark of this cancer type (COSMIC database) [75].

Hub proteins from the luminal vs TNBC network included GATA3 and KRT18 (upregulated), both of which are well-known markers of luminal breast cancer [76], as well as MMP7, MMP9, and MPO (downregulated). Matrix metalloproteinases MMP7 and MMP9 are thought to be drivers of tumor cell invasion and metastasis in patients with TNBC, and have therefore been proposed as therapeutic targets for drug treatment of this more aggressive type of BC [77,78]. In accordance, the abundance of hub proteins MMP7 and MMP9 were low in luminal samples compared to TNBC.

Three proteins ZSCAN18, PADI2, and DNAJC12 connected the Her2 vs luminal network with the luminal vs TNBC network. ZSCAN18, which had a high abundance in Luminal vs Her2 and TNBC samples, was identified by the consensus method as one of the best candidates for ER+ vs ER− classification (Table S8). The role of ZSCAN18 in breast cancer is not well-studied; however, this gene is proposed to be a strong methylation marker for colorectal, gastric, and pancreatic cancers [79]. TIFs from patients with Luminal ER+ breast tumors had a high abundance of both ZSCAN18 as well as DNAJC12. In agreement with this observation, the expression level of the DNAJC12 gene is known to be significantly positively correlated with estrogen receptor-positive status, and may be regulated by estrogen itself via response elements in the genes promoter [80]. Inversely to ZSCAN18 and DNAJC12, the protein PADI2 was depleted in TIFs from luminal samples in comparison to Her2 tumors. Expression of the PADI2 gene, a member of the peptidyl arginine deiminase family, has been strongly linked to the amplification of Her2 (ERBB2). The inhibition of PADI2 gene expression results in a decrease in the level of cell cycle genes p21 and Ki67 [81], as well as other genes associated with aggressive breast cancer phenotypes, here among ACSL4 and BIRC3 [82]. PAD12 has been proposed as a biomarker for Her2 tumors and a potential therapeutic target for BC treatment.

Out of the 10 proteins selected as secreted candidate biomarkers for separation of BC subtypes, eight were retained in the PPI networks, while two proteins, THPTA and SEC23B, did not have any interactions annotated in STRING. Three proteins, BCAM, TMEM51, and ULBP2 were leaf nodes in their respective networks. The single interaction partner of BCAM (Her2 vs Luminal network) was ERBB2, although evidence for this interaction (score) was low, while TMEM51 was a part of a small subnetwork including PLEC, PTGFRN, and ECM2 within the larger network with hub proteins NCOR2 and SETD1A. ULBP2, included in the Luminal vs TNBC network, interacted with CAMP, which was tightly connected to the hub nodes, matrix metalloproteinases MMP7 and MMP9.

CELSR1 and NAT1, both from the Luminal vs TNBC network, had three interaction partners each (average number of interactions per node), CELRS1 was connected to SFRP1 and SHROOM3 in addition to hub node GATA3, while NAT1 interacted with GSTM3, SULT2B1, and hub node MPO.

The most interconnected of the 10 protein candidates, with 6–7 edges each, were MIEN1, PIP4K2B, and ARG3. The latter of these, ARG3 was a part of the subnetwork with hub nodes GATA3 and KRT18, along with its homolog ARG2. MIEN1 and PIP4K2B were included in the highly interconnected subnetwork with ERBB2 as the hub and interacted with each other as well a hand full of proteins, all of which were upregulated in TIFs from Her2 tumors.

3.6. TIF subtype-specific protein signatures: Expression origin and externalization pattern

To elucidate whether selected protein candidates with subtype-specific patterns originate from malignant cells, normal cells, or TILs, we performed an extensive IHC analysis of 10 proteins, including anterior gradient protein 3 (AGR3), lutheran/basal cell-adhesion molecule (BCAM), Cadherin EGF LAG seven-pass G-type receptor 1 (CELSR1), membrane-anchored protein C35 (MIEN1), *N*-acetyltransferase 1 (NAT1), phosphatidylinositol-5-phosphate (PtdIns5P)-4-kinase (PIP4K2B), Sec23 Homolog A (SEC23B), Thiamine triphosphatase (THTPA), Transmembrane protein (TMEM51), and UL16-binding protein 2 (ULBP2) (Table 1). These proteins were selected for analysis based on the availability of highly specific antibodies and their quality and specificity exhibited in a series of control experiments (Table S1). Protein expression was analyzed for both proximal and distant samples, which were collected with TIF recovery and from paired normal lesions (Fig. 8) based on stratification criteria described in Methods (see IHC of tissue biopsies: histological assessment and tumor subtyping and Table S1). Representative examples of high (3+) versus low (0–1+) expression levels within tumor samples, as well as within nonmalignant areas, are shown in Fig. 8 (panel A).

A key advantage of IHC analysis is that it provides visualization of spatial tissue architecture, including inter- and intracellular expression context. Upon the first examination of the IHC images obtained, we observed that most of the proteins exhibited expression patterns in both the cytoplasm and membrane, which is consistent with available literature. PIP4K2B exhibited strong nucleic positivity in several samples in addition to classical cytoplasmic and apical membrane staining, and this result is also consistent with previously published data [83]. It has been hypothesized that differential intracellular localization of PIP4K2B may be associated with the cellular functions of particular PI5P4K isoforms in different tumor subtypes [84]. With the exception of PI5P4K, all of the proteins exhibited significantly higher expression levels in malignant cells compared with normal cells (see the inset panels within the corresponding IHC panels of Fig. 8 and data presented in Table 1). The expression of PIP4K3B in normal mammary epithelium is consistent with data published by Keune *et al*. [85]. Interestingly, we also observed that the expression levels of all 10 proteins were significantly lower in the TILs located inside corresponding tumor lesions, regardless of subtype. In addition, positivity was not detected for any of the 10 proteins in adipose cells, nor in infiltrated or distant lesions (results not shown). Thus, collectively, our IHC data are in agreement with our comparative MS-based analysis [see Identifying a minimal subset of disease-related proteins: A consensus approach (DA analysis, LASSO regression, and random forest)], which did not detect any of these proteins in the NIF or FIF pooled samples. Taken together, these results clearly indicate that the abundance of our selected 10-protein panel in the tumor interstitium is predominantly due to externalization of these proteins from neoplastic cells rather than from nonmalignant ductal epithelium, adipose cells, or the immuno-complement of the tumor microenvironment.

Next, we compared the abundance of these 10 proteins across tumor tissues from different BC subtypes in order to validate a presumed correlation between intratumor expression levels and abundance within TIF. Due to the small sample size of the Her2 group, we were only able to apply Fisher's exact test of IHC score distribution to the luminal versus TNBC samples. Significant *P*-values were obtained for all of the proteins except MIEN1 and TMEM51. This is in full accordance with the differential abundance/expression of MIEN1 and TMEM51 associated with the Her2 subtype, and these samples were not included in the

Table 1. Detailed information for the top 10 candidate proteins found to be DA in TIF samples obtained from luminal, Her2, and TNBC BC subtypes. Parentheses around arrows denote a protein was borderline significant (i.e., corrected *P*-value < 0.05, and log-fold change close to, but not quite reaching, the cutoff of −|+ 1) (a) Primary UniProt nomenclature, (b) Primary SwissProt nomenclature, (c) see Table S6 for details, (d) Arrows denote abundance directionality (up or down) in TIF samples, (e) see Table S7 and Fig. 7 for details, (f) see Fig. 7 for details, (g) see Fig. 7 for details, (h) see Table S8 for details, (↓)—borderline significant. Additional references in table are [104–146].

Gene symbol (a)	Protein name (b)	Molecular function/ Biological process	Differential abundancy in TIF (c)			Differential expression in tumor biopsy (e)			Correlation TIF/Tumor (f)	Expression in normal breast tissue (close vicinity) (g)	Far-distant	Presence in pooled NIF/FIF	Presence in exosome database [40]	Presence in Plasma database [39,118]	Secreted (BC cell lines) [42]	Presence in serum (PubMed)	Expression in 33 normal human tissues (h)	Expression in BC biopsies (PubMed)	Expression in other cancers (i)
			Luminal	Her2	TNBC	Luminal	Her2	TNBC											
AGR3	Anterior gradient protein 3	Dystroglycan binding; negative regulation of cell death	↑ (d)	Not significant	↓	High (3+)	Medium/ Low	Low (1+ – 0)	Yes	Low/Neg	Low/Neg	No	Yes	Yes	Yes	Yes [88]	Neg but medium in bladder, stomach, small intestine, colon and rectum	mRNA [139]; protein [88,89,109,146]	Liver (protein [119,129]), ovarian (mRNA [144]; protein [141]), prostate (mRNA [122])
BCAM	Lutheran/basal cell-adhesion molecule, Lu/ Bcam/CD239	Transmembrane signaling; cell adhesion and migration	↑	↓	↓	High (3+)	Medium/ Low	Low (1+ – 0)	Yes	Low/Neg	Low/Neg	No	Yes	Yes	Yes	Yes [90]	Neg but medium in bladder, fallopian tube, esophagus, heart, kidney, lung, parathyroid, prostate and thyroid	Protein [90,91]	Colon (protein [134]), hepatocellular carcinoma (plasma [130]; protein [113]), ovarian (protein [110]), pancreatic cancer (serum [115]), skin (protein [104,106])
CELSR1	Cadherin EGF LAG seven-pass G-type receptor 1	Transmembrane signaling; cell adhesion and migration	↑	(↓)	↓	High (3+)	Low (1+ – 0)	Low (1+ – 0)	Yes	Low/Neg	Low/Neg	No	No	No	Yes	No Data	Neg in all	Gene copy [99,121]	Hepatocellular carcinoma (DNA methylation [120]), glioma (mRNA [142]), lymphocytic leukemia (protein [127])
MIEN1	membrane-anchored protein C35, C17orf37	Regulation of cell migration and apoptosis	↓	↑	↓	Medium (2+)	High (3+)	Low (1+ – 0)	Yes	Low/Neg	Low/Neg	No	Yes	Yes	Yes	No Data	Neg in all	Protein [76,93,94,111]	Ovarian (mRNA [128]), oral (mRNA & protein [132])
NAT1	*N*-acetyltransferase 1	Drugs and carcinogens metabolizing enzyme	↑	Not significant	↓	High (3+)	Low (1+ – 0)	Low (1+- 0)	Yes	Low/Neg	Low/Neg	No	Yes	No	No	No Data	Neg in all but medium in bladder	mRNA [25,105,107,108]; protein [92,112,126]	No data
PIP4K2B	PhosphatidyliNositol-5-phosphate (PtdIns5P)-4-kinase	Stress-regulated lipid kinase; cell surface receptor	Not significant	↑	↓	Medium (2+)	High (3+)	Low (1+ – 0)	Yes	Comparable	Comparable	No	Yes	Yes	No	No Data	Neg or very low in all	mRNA & protein [85,125]	Lung adenocarcinoma (mRNA [138]), myeloid

Table 1. (Continued).

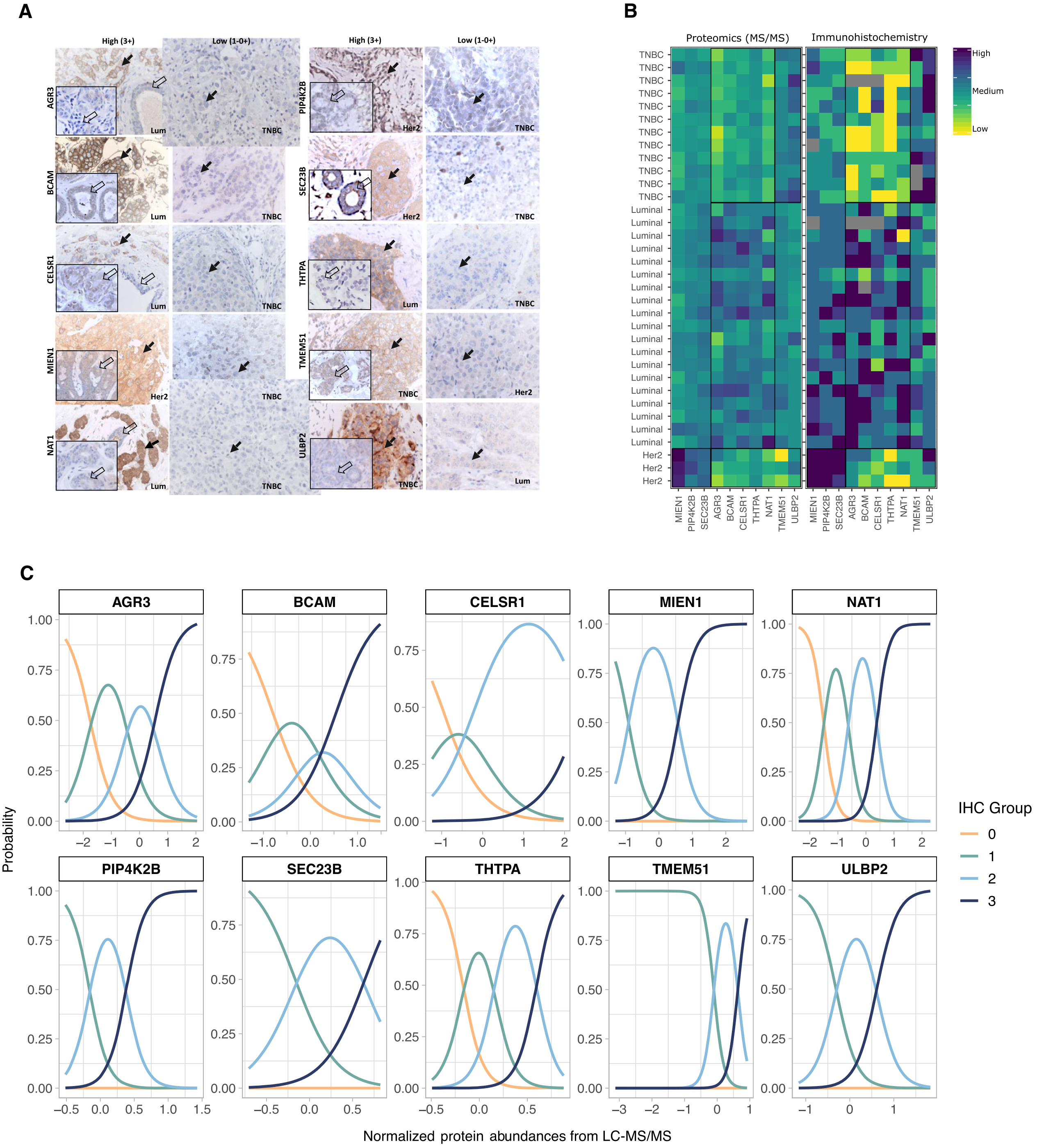
Gene symbol (a)	Protein name (b)	Molecular function/ Biological process	Differential abundancy in TIF (c)			Differential expression in tumor biopsy (e)			Correlation TIF/Tumor (f)	Expression in normal breast tissue (close vicinity) (g)	Far-distant	Presence in pooled NIF/FIF	Presence in exosome database [40]	Presence in Plasma database [39,118]	Secreted (BC cell lines) [42]	Presence in serum (PubMed)	Expression in 33 normal human tissues (h)	Expression in BC biopsies (PubMed)	Expression in other cancers (i)
			Luminal	Her2	TNBC	Luminal	Her2	TNBC											
		signaling pathway																	leukemia (mRNA [145])
SEC23B	Sec23 Homolog A	GTPase activator activity; cellular transport	Not significant	↑	↓	Medium (2+)	High (3+)	Low (1+ – 0)	Yes	Low/Neg	Low/Neg	No	Yes	No	Yes	No Data	Neg in all but medium in eye	mRNA & protein [133]	Thyroid & endometrial (protein [133])
THTPA	Thiamine triphosphatase	Hydrolase activity; metabolic and energy processes	↑	Not significant	↓	High (3+)	Low (1+ – 0)	Low (1+ – 0)	Yes	Low/Neg	Low/Neg	No	No	No	No	No Data	Neg or very low in all	No data	No data
TMEM51	Transmembrane protein	Unknown function	↑	↓	↑	Medium (2+)	Low (1+ – 0)	High (3+)	Yes	Low	Low	No	Yes	No	No	No Data	Neg in all but medium in bladder, kidney and pancreas	No data	Pancreatic (mRNA [136])
ULBP2	UL16-binding protein 2	Cell surface glycoprotein; natural killer cell activation; immune response	↓	Not significant	↑	Low (1+ – 0)	Low (1+ – 0)	High (3+)	Yes	Low/Neg	Low/Neg	No	No	No	Yes	Yes [124,131]	Neg in all but medium in heart	mRNA [143]; protein [96]	Colon (mRNA [140]), lung (plasma [123]), ovarian (mRNA [116]; protein [114], pancreatic (protein [135,137]; plasma [117])

test (Table 1 and Table S11). These results also support the TIF-MS data in terms of differential expression/abundance levels of particular proteins detected with the paired tumor comparison of BC subtypes. To visualize the correlation between TIF and intratumor protein abundance, we generated tile plots of (a) batch-corrected, MS-based protein abundances of the top 10 candidates for discriminating BC subtypes and (b) IHC scores of the same 10 proteins (Table S11). The patterns of protein abundance/expression for the two plots (Fig. 8, panel B) were highly comparable, and importantly, a strong association was observed for Her2 samples, which were not subjected to Fisher's exact test. To better assess the correlation observed between the two tile plots, we performed logistic regression using the discrete IHC scores from solid tissues as the response and the normalized TIF protein abundances from LC-MS/MS as the predictor. Results are visualized in Fig. 8 (panel C). The plot shows the probability of a given IHC score in response to the normalized protein abundance in TIF. All logistic regression models (one for each of the 10 proteins of interest) had an overall significant *P*-value, indicating that TIF protein level was indeed predictive of IHC score, see Table S12. However, as was evident from the group-specific *P*-values in Table S12, TIF protein abundances were not found to have a significant effect on all levels of IHC for all proteins. In the case of ARG3, NAT1, and THTPA, there was a good correlation between the level of protein in TIF and all IHC scores, supported by the distinct probability peaks in Fig. 8 (panel C). In contrast, proteins BCAM and CELSR1 displayed less clear-cut patterns, specifically with respect to the intermediate IHC score of 1–2, in accordance with accompanying *P*-values.

For a molecule to be considered a serological cancer-specific biomarker, it should be secreted predominantly from malignant tissues, not normal tissues. Therefore, we also analyzed expression levels of the selected candidate proteins in normal human tissues, using a TMA of normal human tissues deriving from 33 different organs (Pantomics, Inc., San Francisco, CA, USA), which is recommended by the FDA in its guidelines for testing cross-reactivity. The results obtained are summarized in Table S13. We supplemented these data with publicly available information regarding the expression patterns of our candidates in normal tissues from: (a) the Protein Atlas (https://www.proteinatlas.org/), (b) the Integrated Proteomics Database (https://www.proteomicsdb.org/proteomicsdb/#overview), and (c) the Tissues Expression Database (https://tissues.jensenlab.org/Search). The latter is often used as a reference for protein expression in literature. We observed good consensus between the data available in these databases and our IHC results. For example, CELSR1, MIEN1, NAT1, PIP4K2B, SEC23B, THPPA, and ULBP1 were only detected at background levels in almost all of the normal tissues analyzed. Meanwhile, AGR3, BCAM, and TMEM51 exhibited noticeable expression levels in a number of normal tissues.

It is interesting to note that the minimal TIF protein set proposed in our study does not overlap with the protein signature of breast cancer subtypes described in the comprehensive proteome dataset of BC tissue biopsies published by Tyanova *et al.* [26]. This discrepancy may be due to significant differences in the methods used to obtain protein lysates for subsequent MS/MS analysis in these two studies. In contrast to our work in which fresh tissue was used, Tyanova *et al.* extracted proteins from FFPE tissue blocks with deparaffinization in xylene and ethanol. It cannot be ruled out that such differences in procedure may have affected the protein profiles of the samples examined.

Fig. 8. Differential intratumor and TIF abundance of 10 proteins which discriminate between BC subtypes. Intracellular expression levels of selected proteins were estimated by IHC across tumor biopsies used for TIF recovery, panel A. The expression level of each protein was considered positive if at least 10% of the tumor cells have intensities of expression scored as (0–1+), medium (2+), or high (3+), in accordance with previously described criteria (Keune *et al.* [85]). Representative examples of high (3+) versus low (0–1+) expression levels for AGR3, BCAM, CELSR1, MIEN1, NAT1, PIP4K2B, SEC23B, THTPA, TMEM51, and ULBP2 proteins are presented. BC subtypes are specified within each panel. Far-normal areas of each tumor biopsy used for TIF recovery were analyzed in parallel and representative staining is shown (insertions within corresponding IHC panels). Black arrows within representative IHC images of the tumor biopsies indicate either positive (left panels) or negative (right panels) malignant cells. Transparent arrows indicate normal-like mammary ducts located in close proximity of tumor cells. Magnification of images shown is 40×. Correlation between TIF and intracellular abundancy is summarized in panel B with tile plots containing: (a) batch-corrected MS/MS-based protein abundances of the top 10 candidates for discriminating breast cancer subtypes and (b) IHC scores of the same 10 proteins. Protein names are indicated on the *x*-axis. Sample ID, along with assigned BC subtype, are denoted on the *y*-axis. Coloring indicates low abundance/low IHC score (yellow), ranging to high abundance/high IHC score (blue). Black squares highlight which of the three subtypes the samples belonging to. Panel C contains logistic regression plots, one for each of the 10 proteins of interest, showing the probability of a given IHC score (*y*-axis) in relation to normalized TIF protein abundance (*x*-axis). Colors denote IHC scores of 0 (0–0.5), 1 (1–1.5), 2 (2–2.5), and 3.



In addition, different statistical algorithms used in the two studies may have resulted in differences in the corresponding hits.

Therefore, to further validate the breast cancer subtype-specific expression pattern of the 10 TIF proteins and to evaluate their potential clinical applicability as candidates for classification of BC subtype, we examined their sensitivity and specificity on Tyanova's proteome dataset. An AUC analysis was performed for eight of the 10 proteins (i.e., AGR3, BCAM, CELSR1, MIEN1, NAT1, PIP4K2B, SEC23B, and THTPA) because TMEM51 and ULBP2 had almost exclusively missing values, and thus, could not be included in the analysis. In light of more recent demonstrations that use of multiple markers can significantly increase the specificity and sensitivity of disease classification compared to the use of single biomarkers [86,87], we estimated AUCs for both individual proteins and various protein combinations. Briefly, AUCs were estimated for: (a) individual AUCs for proteins DA within pairwise subtype contrasts, (b) combined AUC for multiple proteins associated with pairwise subtype contrasts, and (c) combined AUC for all proteins with all three BC subtypes included. The results obtained are summarized in Table 2.

The AUCs of the individual proteins ranged from 0.74 to 0.91, with SEC23B and PIPI4K2B having the lowest and highest specificity/sensitivity values, respectively, for classification of Her2 versus TNBC subtypes. Strikingly, MIEN1 had AUC = 1.0, which we partly attribute to the abundance of this protein being completely distinct between Her2 samples and TNBC samples. However, since MIEN1 was also one of the proteins which had the largest number of missing values (~ 35%), it was difficult to determine an exact AUC for this protein. Moreover, although the specificity/sensitivity of MIEN1 is likely to be high, it is doubtful that it would reach 1.0 in a larger dataset. For pairwise subtype classification by using a combination of markers, the AUCs were good, with values > 0.9 obtained for all three models (Table 2). In the combined model with luminal versus TNBC and Her2 versus TNBC, we observed a redundancy of markers. Additionally, we observed that use of more than two markers did not increase AUC in any noticeable manner. This is shown in Table 2 where a dual-protein combination which maximized AUC is shown first, and AUCs associated with the remaining proteins which are specific for various subtype comparisons (denoted in parentheses) are below. Lastly, the combined AUC estimate for classification in the multinomial set-up with all three subtypes in one model ranged from 0.85 to 0.91 (Table 2). The difference in AUC depended on how the dataset was split into test and training sets (see Methods) and essentially reflected how a small dataset with a large variance is highly sensitive to division. Overall, the AUC scores strongly support our observation that eight of the proteins selected (AGR3, BCAM, CELSR1, MIEN1, NAT1, PIP4K2B, SEC23B, and THTPA) exhibit potential as biomarkers for stratification of BC subtypes. However, their possible validity as serological markers for breast malignancy should be investigated, using an independent serum proteome dataset from BC patients.

Unfortunately, we were not able to verify the presence of the identified protein panel directly in matched blood samples since the latter samples had been consumed in previous studies [17,18]. We also could not find any publicly available database of plasma proteins classified by BC subtype and containing information about potential serological BC subtype-related biomarkers. Therefore, we indirectly evaluated the validity of our 10-protein set as a potential serological signature by curating relevant databases (Fig. 2). The information used included: (a) a breast TIF MS-dataset acquired from a pilot experiment [31], (b) the most updated versions of plasma and exosome databases [39,40], (c) a secretome dataset derived from BC cell lines [42], and (d) the presence of these proteins in human serum according to available literature. The results of these curations, in combination with relevant protein characteristics (i.e., differential abundance in TIF and correlation with expression levels in matched tumors), are summarized in Table 1. A PubMed search revealed that six out of the 10 subtype-specific proteins identified in our study have previously been characterized as differentially expressed in breast tumors at the protein and/or mRNA levels, compared to nonmalignant counterparts. These include AGR3 [88,89], BCAM [90,91], NAT1 [92], MIEN1 [93–95], PIP4K2B [85], and ULBP2 [96]. It should be noted, however, that with a few exceptions, most of these studies analyzed protein/mRNA levels in tumors without regard to subtyping, intratumor context, or secretion status. It has been reported that AGR3 is associated with less aggressive breast tumors and better BC patient outcome [89]. However, a PubMed search of thiamine triphosphatase (THTPA), transmembrane protein 51 (TMEM51), and cadherin EGF LAG seven-pass G-type receptor 1 (CELSR1), did not return any relevant literature about their expression in breast tumors. Thus, to the best of our knowledge, the present results appear to provide a first indication of the value of these three proteins for tumor subtype stratification.

The expression profile of all 10 protein candidates, along with relative abundances in relation to the three main tumor subtypes, are schematically presented in Fig. 9. Six of the ten proteins (AGR3, BCAM, CELSR1, NAT1, THTPA, and TMEM51) were found to be significantly upregulated in the tumor/TIF of the luminal subtype samples. In particular, BCAM and CELSR1 are significantly elevated in the luminal subtype compared to the TNBC and Her2 subtypes (Fig. 9). Thus, these two proteins may represent a specific biomarker signature which can discriminate luminal subtype breast tumors.

BCAM, also known as CD239, is a plasma membrane glycoprotein and a receptor for the extracellular matrix protein, laminin [97]. Expression of CD239/

Table 2. Validation of BC subtype-specific expression profiles of selected protein biomarkers. AUC was estimated by using protein abundances quantified from 40 BC specimens (subtypes: luminal, Her2, and TNBC) by Tyanova *et al.* [26]. AUCs were estimated from: (a) models with individual markers used as classifiers of BC subtypes (pairwise), (b) a generalized linear model with additive combinations of markers in BC subtypes (pairwise), and (c) a multinomial log-linear model fit via a neural network (Venables and Ripley [56]) by using a train/test set-up with additive combinations of markers in BC subtypes (all). Parentheses around a protein name indicate that the corresponding AUC only increased minimally, or not at all when this protein was included in the model. A period indicates that the marker was not relevant for a given pairwise comparison. CI = 95% confidence interval was relevant/possible.

(a) Individual AUC scores with 95% confidence intervals for each protein associated with a pairwise subtype contrast

	AGR3	BCAM	CELSR1	MIEN1	NAT1	PIP4K2B	SEC23B	THTPA
Luminal vs TNBC	0.85 (CI: 0.67–1)	0.80 (CI: 0.59–1)	0.77 (CI: 0.56–0.98)	.	0.84 (CI: 0.65–1)	.	.	0.77 (CI: 0.56–0.97)
Luminal vs Her2	.	0.87 (CI: 0.72–1.0)	.	0.83 (CI: 0.67 –0.1.0)	.	.	.	.
Her2 vs TNBC	.	.	.	1.0 (CI: NA)	.	0.91 (CI: 0.8 –1.0)	0.74 (CI: 0.55–0.93)	
.								

(b) AUC score with 95% confidence interval for combined proteins associated with a pairwise subtype contrast

Luminal vs TNBC	**ARG3** + **BCAM** (+ **CELSR1** + **NAT1** + **THTPA**) 0.92 (CI: 0.8–1.0)	
Luminal vs Her2	**BCAM** + **MIEN1** 0.94 (CI: 0.87–1.0)	
Her2 vs TNBC	**PIP4K2B** (+ **SEC23B**) 0.91 (CI: 0.8–1.0)	**MIEN1** + **PIP4K2B** (+ **SEC23B**) 1.0 (CI: NA)

(c) AUC scores (min. observed – max. observed) for the combination of all markers and three subtypes together. AUCs from a train/test multinomial log-linear model, fit via neural network (cite)

Luminal vs Her2 vs TNBC	**ARG3** + **BCAM** + **MIEN1** + **PIP4K2B** (+ **CELSR1** + **NAT1** + **SEC23B** + **THTPA**) *0.85–0.91 (CI:NA)*

BCAM is increased in invasive ductal carcinomas [91], and has also been found to be elevated level in a subset of BC tissues, particularly Her2-negative tumors [91]. The present results are in agreement with these results. Furthermore, it has been hypothesized that BCAM represents a promising antigen for antibody-drug conjugate-based BC therapy [91]. BCAM is a secreted protein, and the significantly higher serum level of BCAM, determined by ELISA, has been reported in BC patients compared to normal individuals [90]. However, in the latter study, the tumors analyzed were not stratified, according to subtype.

Another member of our luminal-specific signature is CELSR1, a protein shown to have a key role in epithelial planar cell polarity [98]. In general, very little is known about the potential role of this protein in carcinogenesis, and particularly in BC progression. A recently published study [99] demonstrated that CELSR1 is commonly amplified in pure, yet not mixed, ductal carcinoma *in situ* (DCIS) and is associated with invasion. Amplification of the 22q arm of chromosome 13, the position of *CELSR1*, is also frequently observed in DCIS [99]. In our IHC analysis, CELSR1 positivity was mainly associated with the cytoplasmic compartment, as expected for primary breast carcinomas. Moreover, CELSR1 positivity strongly correlated with less aggressive luminal type tumors. In contrast, TNBC tumors and normal ducts distant from the tumor site were almost exclusively negative for CELSR1 (Fig. 8, panel A).

Three proteins, MIEN1, PIP412B, SEC23B, were upregulated in tumor/TIF samples of the Her2-enriched subtype compared to the levels detected in the luminal (MIEN1) and TNBC (MIEN1, PIP412B, and SEC23B) subtypes (Fig. 9). However, only the expression of MIEN1 was found to be specific to Her2 tumors (Fig. 9). MIEN1, migration and invasion enhancer 1, is a membrane-anchored protein, which is highly expressed in various types of cancer. It was recently reported that expression of MIEN1 in human BC tissue is higher than in adjacent noncancerous breast tissue [93], which is consistent with our data. Notably, a direct correlation between upregulation of MIEN1 and upregulation of neighboring genes,

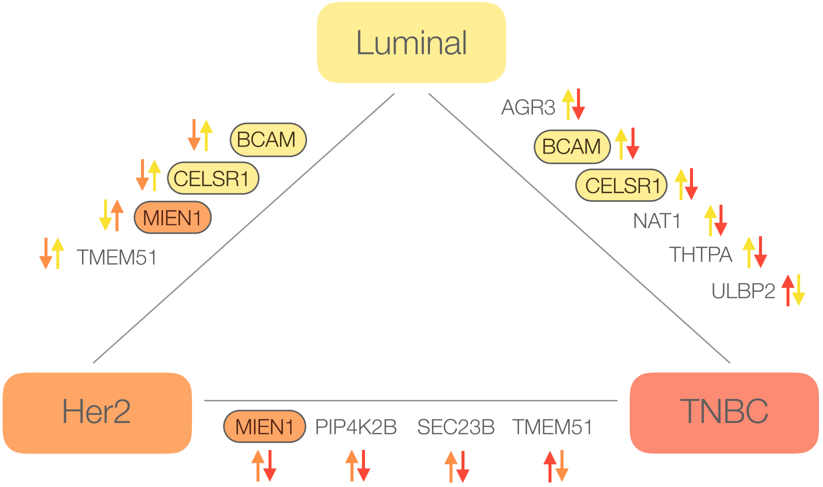


Fig. 9. A schematic which represents the regulation pattern of 10 protein candidates in tumor and matched TIF samples according to luminal (yellow), Her2 (orange), and TNBC (red) tumor subtypes. Comparative directionality of protein abundance between subtypes is represented as arrows in corresponding colors. The proteins that are specifically upregulated in one subtype compared to the other two subtypes are marked with an oval filled with the corresponding color.

ERBB2 and GRB7, was recently shown in a variety of cancers, including BC [93,100,101]. Meanwhile, *in vitro*, it has been demonstrated that overexpression of MIEN1 may promote cell dissemination and invasion in breast cancer by regulating cytoskeletal-focal adhesion dynamics [102]. There is no information currently available regarding MIEN1 in human plasma. However, with curation of the Human Plasma PeptideAtlas [39], it has been confirmed that MIEN1 is present in circulation (Table 1). Thus, MIEN1 is a component of the breast tumor secretome in Her2-positive patients, and targeting MIEN1 in the bloodstream may represent a promising approach to prevent breast tumor metastasis, especially for Her2-enriched cancers.

In summary, our studies have led to the identification of three proteins, which have the potential to specifically discriminate between BC subtypes, particularly luminal (BCAM and CELSR1) and Her2 (MIEN1) enriched subtypes. The expression of TMEM51 was also found to be specific to the Her2 subtype (Fig. 9), although its downregulation in Her2 samples diminishes its value as a potential biomarker. Six additional proteins which we identified also exhibited expression levels relative to the BC subtypes examined, which manifested as pairwise differences.

4. Conclusions

Overall, by characterizing breast TIF proteome with high-throughput LC-MS/MS and bioinformatics analyses, we generated a database containing over 8800 proteins externalized from breast tumors into the tumor microenvironment,which represents the most comprehensive BC secretome dataset published to date. To maximize the probability of finding protein signature(s) associated with BC subtypes, ER/PgR/Her2 status and scoring of TILs we used a consensus bioinformatics approach including DAA, LASSO, and RF that led to the identification a minimal panel of 24 proteins, 10 of which namely, AGR3, BCAM, CELSR1, MIEN1, NAT1, PIP4K2B, SEC23B, THTPA, TMEM51, and ULBP2 were analyzed by IHC on matched tumor tissue samples, confirming their potential to stratify the BC tumor subtype-specific TIFs. In particular, increased abundancy of BCAM and CELSR1 in TIF differentiates luminal, while upregulation of MIEN1 differentiates Her2 subtypes. The sensitivity and specificity were estimated for this 10-protein panel in an independent, comprehensive breast tumor proteome dataset [26] using the AUC scores and the results strongly support our evidence that eight of the proteins (AGR3, BCAM, CELSR1, MIEN1, NAT1, PIP4K2B, SEC23B, and THTPA) might serve as biomarkers for stratification of luminal, Her 2 and TNBC tumor subtypes. Curation of the most relevant and current datasets of secreted and plasma proteins hypothesized the potential of identified proteins to serve as tumor-specific biomarkers for plasma screening. Further studies are warranted to confirm the validity of these deregulated proteins as classifiers for particular breast tumor subtypes and to evaluate their value as potential serological biomarkers. We believe that the results presented in our study provide a system-wide, quantitative baseline map and data resource of the breast interstitial fluid proteome, which extends the existing human tissue proteome databases.

Acknowledgements

The authors wish to thank Maria Grønvig Nielsen and Lisbeth Lisbeth Schade Hansen for expert technical assistance, Dr Vera Timmermans Wielenga and Dr

Maj-Lis Møller Talman (Department of Pathology, Center of Diagnostic Investigations, Copenhagen University Hospital, Copenhagen, Denmark) for their invaluable support. This project is supported by grants from The Eurocan Platform, which has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 260791. The Danish Cancer Society, the 'Race Against Breast Cancer' foundation, the John and Birthe Meyer Foundation, the Danish National Research Foundation (Centre of Excellence: CARD, DNRF125), the Innovation Fund Denmark grant 2016-2019 5189-0052B, and The Regional Health Authorities in South-East of Norway. RS acknowledges funding from the Science Foundation Ireland Starting Investigator Research grant (SFI SIRG) under Grant Number 13/SIRG/2164. The calculations described in this paper were performed using the DeiC National Life Science Supercomputer Computerome at DTU.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

IIG, EP, TT, and MP conceived and designed the project. MP acquired the data. IIG, EP, TT, MP, and PG analyzed and interpreted the data. IIG, EP, and TT wrote the manuscript. PSG and IIG collected the material and participated in the evaluation of data. All of the authors read and revised the manuscript critically and approved the final manuscript.

Peer Review

The peer review history for this article is available at https://publons.com/publon/10.1002/1878-0261.12850.

References

1 Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA & Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394–424.

2 Prat A, Parker JS, Fan C & Perou CM (2012) PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. *Breast Cancer Res Treat* **135**, 301–306.

3 Anderson WF, Rosenberg PS, Prat A, Perou CM & Sherman ME (2014) How many etiological subtypes of breast cancer: two, three, four, or more? *J Natl Cancer Inst* **106**, dju165.

4 Martelotto LG, Ng CK, Piscuoglio S, Weigelt B & Reis-Filho JS (2014) Breast cancer intra-tumor heterogeneity. *Breast Cancer Res* **16**, 210.

5 Beca F & Polyak K (2016) Intratumor heterogeneity in breast cancer. *Adv Exp Med Biol* **882**, 169–189.

6 Gromov P, Gromova I, Bunkenborg J, Cabezon T, Moreira JMA, Timmermans-Wielenga V, Roepstorff P, Rank F & Celis JE (2010) Up-regulated proteins in the fluid bathing the tumour cell microenvironment as potential serological markers for early detection of cancer of the breast. *Mol Oncol* **4**, 65–89.

7 Papaleo E, Gromova I & Gromov P (2017) Gaining insights into cancer biology through exploration of the cancer secretome using proteomic and bioinformatic tools. *Expert Rev Proteomics* **14**, 1021–1035.

8 Place AE, Jin Huh S & Polyak K (2011) The microenvironment in breast cancer progression: biology and implications for treatment. *Breast Cancer Res* **13**, 227.

9 Gromov P, Gromova I, Olsen CJ, Timmermans-Wielenga V, Talman ML, Serizawa RR & Moreira JMA (2013) Tumor interstitial fluid — a treasure trove of cancer biomarkers. *Biochim Biophys Acta (BBA)* **1834**, 2259–2270.

10 Wagner M & Wiig H (2015) Tumor interstitial fluid formation, characterization, and clinical implications. *Front Oncol* **5**, 115.

11 Viotti C (2016) ER to Golgi-dependent protein secretion: the conventional pathway. *Methods Mol Biol* **1459**, 3–29.

12 Nickel W (2003) The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. *Eur J Biochem* **270**, 2109–2119.

13 Koga K, Matsumoto K, Akiyoshi T, Kubo M, Yamanaka N, Tasaki A, Nakashima H, Nakamura M, Kuroki S, Tanaka M, *et al.* (2005) Purification, characterization and biological significance of tumor-derived exosomes. *Anticancer Res* **25**, 3703–3707.

14 Simpson RJ, Lim JW, Moritz RL & Mathivanan S (2009) Exosomes: proteomic insights and diagnostic potential. *Expert Rev Proteomics* **6**, 267–283.

15 Whittaker K, Burgess R, Jones V, Yang Y, Zhou W, Luo S, Wilson J & Huang RP (2019) Quantitative proteomic analyses in blood: a window to human health and disease. *J Leukoc Biol* **106**, 759–775.

16 Gromov P & Gromova I (2016) Characterization of the tumor secretome from tumor interstitial fluid (TIF). *Methods Mol Biol* **1459**, 231–247.

17 Espinoza JA, Jabeen S, Batra R, Papaleo E, Haakensen V, Timmermans Wielenga V, Møller Talman ML, Brunner N, Børresen-Dale AL, Gromov P *et al.* (2016) Cytokine profiling of tumor interstitial fluid of the breast and its relationship with lymphocyte

infiltration and clinicopathological characteristics. *Oncoimmunology* **5**, e1248015.

18 Halvorsen AR, Helland Å, Gromov P, Wielenga VT, Talman MM, Brunner N, Sandhu V, Børresen-Dale AL, Gromova I & Haakensen VD (2017) Profiling of microRNAs in tumor interstitial fluid of breast tumors - a novel resource to identify biomarkers for prognostic classification and detection of cancer. *Mol Oncol* **11**, 220–234.

19 Terkelsen T, Haakensen VD, Saldova R, Gromov P, Hansen MK, Stockmann H, Lingjaerde OC, Børresen-Dale AL, Papaleo E, Helland Å *et al.* (2018) N-glycan signatures identified in tumor interstitial fluid and serum of breast cancer patients: association with tumor biology and clinical outcome. *Mol Oncol* **12**, 972–990.

20 Jabeen S, Espinoza JA, Torland LA, Zucknick M, Kumar S, Haakensen VD, Lüders T, Engebraaten O, Børresen-Dale AL, Kyte JA *et al.* (2019) Noninvasive profiling of serum cytokines in breast cancer patients and clinicopathological characteristics. *Oncoimmunology* **8**, e1537691.

21 Celis JE, Moreira JMA, Cabezón T, Gromov P, Friis E, Rank F & Gromova I (2005) Identification of extracellular and intracellular signaling components of the mammary adipose tissue and its interstitial fluid in high risk breast cancer patients. *Mol Cell Proteomics* **4**, 492–522.

22 Celis JE & Gromov P (2003) Proteomics in translational cancer research: toward an integrated approach. *Cancer Cell* **3**, 9–15.

23 Lam SW, Jimenez CR & Boven E (2014) Breast cancer classification by proteomic technologies: current state of knowledge. *Cancer Treat Rev* **40**, 129–138.

24 Mardamshina M & Geiger T (2017) Next-generation proteomics and its application to clinical breast cancer research. *Am J Pathol* **187**, 2175–2184.

25 Zhang B, Whiteaker JR, Hoofnagle AN, Baird GS, Rodland KD & Paulovich AG (2019) Clinical potential of mass spectrometry-based proteogenomics. *Nat Rev Clin Oncol* **16**, 256–268.

26 Tyanova S, Albrechtsen R, Kronqvist P, Cox J, Mann M & Geiger T (2016) Proteomic maps of breast cancer subtypes. *Nat Commun* **7**, 10259.

27 Bouchal P, Schubert OT, Faktor J, Capkova L, Imrichova H, Zoufalova K, Paralova V, Hrstka R, Liu Y, Ebhardt HA *et al.* (2019) Breast cancer classification based on proteotypes obtained by SWATH mass spectrometry. *Cell Rep* **28**, 832–843, e7.

28 Johansson HJ, Socciarelli F, Vacanti NM, Haugen MH, Zhu Y, Siavelis I, Fernandez-Woodbridge A, Aure MR, Sennblad B, Vesterlund M *et al.* (2019) Breast cancer quantitative proteome and proteogenomic landscape. *Nat Commun* **10**, 1600.

29 Gajbhiye A, Dabhi R, Taunk K, Jagadeeshaprasad MG, RoyChoudhury S, Mane A, Bayatigeri S, Chaudhury K, Santra MK & Rapole S (2017) Multipronged quantitative proteomics reveals serum proteome alterations in breast cancer intrinsic subtypes. *J Proteomics* **163**, 1–13.

30 Zeidan B, Manousopoulou A, Garay-Baquero DJ, White CH, Larkin SET, Potter KN, Roumeliotis TI, Papachristou EK, Copson E, Cutress RI *et al.* (2018) Increased circulating resistin levels in early-onset breast cancer patients of normal body mass index correlate with lymph node negative involvement and longer disease free survival: a multi-center POSH cohort serum proteomics study. *Breast Cancer Res* **20**, 19.

31 Raso C, Cosentino C, Gaspari M, Malara N, Han X, McClatchy D, Park SK, Renne M, Vadalà N, Prati U *et al.* (2012) Characterization of breast cancer interstitial fluids by TmT labeling, LTQ-Orbitrap Velos mass spectrometry, and pathway analysis. *J Proteome Res* **11**, 3199–3210.

32 Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, Wienert S, Van den Eynden G, Baehner FL, Penault-Llorca F *et al.* (2015) The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann Oncol* **26**, 259–271.

33 Esposito A, Criscitiello C & Curigliano G (2015) Highlights from the 14(th) St Gallen International Breast Cancer Conference 2015 in Vienna: dealing with classification, prognostication, and prediction refinement to personalize the treatment of patients with early breast cancer. *Ecancermedicalscience* **9**, 518.

34 Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**, 248–254.

35 Branca MM, Orre M, Johansson J, Granholm V, Huss M, Pérez-Bercoff Å, Forshed J, Käll L & Lehtiö J (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat Methods* **11**, 59–62.

36 Stacklies W, Redestig H, Scholz M, Walther D & Selbig J (2007) pcaMethods – a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23**, 1164–1167.

37 Leek JT, Johnson WE, Parker HS, Jaffe AE & Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883.

38 Murtagh FaL P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J Classif* **31**, 274–295.

39 Schwenk JM, Omenn GS, Sun Z, Campbell DS, Baker MS, Overall CM, Aebersold R, Moritz RL & Deutsch EW (2017) The human plasma proteome draft of 2017: building on the human plasma PeptideAtlas from mass

spectrometry and complementary assays. *J Proteome Res* **16**, 4299–4310.

40 Keerthikumar S, Chisanga D, Ariyaratne D, Al Saffar H, Anand S, Zhao K, Samuel M, Pathan M, Jois M, Chilamkurti N *et al.* (2016) ExoCarta: a web-based compendium of exosomal cargo. *J Mol Biol* **428**, 688–692.

41 Braga-Lagache S, Buchs N, Iacovache MI, Zuber B, Jackson CB & Heller M (2016) Robust label-free, quantitative profiling of circulating plasma microparticle (MP) associated proteins. *Mol Cell Proteomics* **15**, 3640–3652.

42 Boersema PJ, Geiger T, Wisniewski JR & Mann M (2013) Quantification of the N-glycosylated secretome by super-SILAC during breast cancer progression and in human blood samples. *Mol Cell Proteomics* **12**, 158–171.

43 Celis JE, Gromov P, Cabezón T, Moreira JMA, Ambartsumian N, Sandelin K, Rank F & Gromova I (2004) Proteomic characterization of the interstitial fluid perfusing the breast tumor microenvironment. *Mol Cell Proteomics* **3**, 327–344.

44 Xiao N, Cao DS, Zhu MF & Xu QS (2015) protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **31**, 1857–1859.

45 Nielsen H (2017) Predicting secretory proteins with SignalP. *Methods Mol Biol* **1611**, 59–73.

46 Kall L, Krogh A & Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, 1027–1036.

47 Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W & Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47.

48 Kammers K, Cole RN, Tiengwe C & Ruczinski I (2015) Detecting significant changes in protein abundance. *EuPA Open Proteom* **7**, 11–19.

49 van Ooijen MP, Jong VL, Eijkemans MJC, Heck AJR, Andeweg AC, Binai NA & van den Ham HJ (2018) Identification of differentially expressed peptides in high-throughput proteomics data. *Brief Bioinform* **19**, 971–981.

50 Friedman J, Hastie T & Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**, 1–22.

51 Diaz-Uriarte R & Alvarez de Andres S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3.

52 Diaz-Uriarte R (2007) GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics* **8**, 328.

53 Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47** (D1), D607–D613.

54 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B & Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504.

55 Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC & Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77.

56 Venables WN and Ripley BD (2002) Modern Applied Statistics. Springer, New York, NY.

57 Jaiswal R, Luk F, Dalla PV, Grau GE & Bebawy M (2013) Breast cancer-derived microparticles display tissue selectivity in the transfer of resistance proteins to cells. *PLoS One* **8**, e61515.

58 Jaiswal R, Raymond Grau GE & Bebawy M (2014) Cellular communication via microparticles: role in transfer of multidrug resistance in cancer. *Future Oncol* **10**, 655–669.

59 Robbins PD & Morelli AE (2014) Regulation of immune responses by extracellular vesicles. *Nat Rev Immunol* **14**, 195–208.

60 Rabouille C (2017) Pathways of unconventional protein secretion. *Trends Cell Biol* **27**, 230–240.

61 Celis JE, Gromov P, Moreira JMA, Cabezón T, Friis E, Vejborg IMM, Proess G, Rank F & Gromova I (2006) Apocrine cysts of the breast. *Mol Cell Proteomics* **5**, 462–483.

62 Tsuneki M, Nakamura Y, Kinjo T, Nakanishi R & Arakawa H (2015) Mieap suppresses murine intestinal tumor via its mitochondrial quality control. *Sci Rep* **5**, 12472.

63 Kamino H, Nakamura Y, Tsuneki M, Sano H, Miyamoto Y, Kitamura N, Futamura M, Kanai Y, Taniguchi H, Shida D *et al.* (2016) Mieap-regulated mitochondrial quality control is frequently inactivated in human colorectal cancer. *Oncogenesis* **4**, e181.

64 Gaowa S, Futamura M, Tsuneki M, Kamino H, Tajima JY, Mori R, Arakawa H & Yoshida K (2018) Possible role of p53/Mieap-regulated mitochondrial quality control as a tumor suppressor in human breast cancer. *Cancer Sci* **109**, 3910–3920.

65 Wang D, Li J, Cai F, Xu Z, Li L, Zhu H, Liu W, Xu Q, Cao J, Sun J *et al.* (2019) Overexpression of MAPT-AS1 is associated with better patient survival in breast cancer. *Biochem Cell Biol* **97**, 158–164.

66 Darlix A, Hirtz C, Thezenas S, Maceski A, Gabelle A, Lopez-Crapez E, De Forges H, Firmin N, Guiu S, Jacot W *et al.* (2019) The prognostic value of the Tau protein serum level in metastatic breast cancer patients

and its correlation with brain metastases. *BMC Cancer* **19**, 1–13.

67 Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–1167.

68 Liu MC, Pitcher BN, Mardis ER, Davies SR, Friedman PN, Snider JE, Vickery TL, Reed JP, DeSchryver K, Singh B *et al.* (2016) PAM50 gene signatures and breast cancer prognosis with adjuvant anthracycline- and taxane-based chemotherapy: correlative analysis of C9741 (Alliance). NPJ. *Breast Cancer* **2**, 1–8.

69 Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, Díez M, Viladot M, Arance A & Muñoz M (2015) Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast* **24**, S26–S35.

70 Caron C, Lestrat C, Marsal S, Escoffier E, Curtet S, Virolle V, Barbry P, Debernardi A, Brambilla C, Brambilla E *et al.* (2010) Functional characterization of ATAD2 as a new cancer/testis factor and a predictor of poor prognosis in breast and lung cancers. *Oncogene* **29**, 5171–5181.

71 Kalashnikova EV, Revenko AS, Gemo AT, Andrews NP, Tepper CG, Zou JX, Cardiff RD, Borowsky AD & Chen HW (2010) ANCCA/ATAD2 overexpression identifies breast cancer patients with poor prognosis, acting to drive proliferation and survival of triple-negative cells through control of B-Myb and EZH2. *Cancer Res* **70**, 9402–9412.

72 Tajima K, Matsuda S, Yae T, Drapkin BJ, Morris R, Boukhali M, Niederhoffer K, Comaills V, Dubash T, Nieman L *et al.* (2019) SETD1A protects from senescence through regulation of the mitotic gene expression program. *Nat Commun* **10**, 2854.

73 Jin ML, Kim YW, Jin HL, Kang H, Lee EK, Stallcup MR & Jeong KW (2018) Aberrant expression of SETD1A promotes survival and migration of estrogen receptor α-positive breast cancer cells. *Int J Cancer* **143**, 2871–2883.

74 van Agthoven T, Sieuwerts AM, Veldscholte J, Meijer-van Gelder ME, Smid M, Brinkman A, den Dekker AT, Leroy IM, van Ijcken WF, Sleijfer S *et al.* (2009) CITED2 and NCOR2 in anti-oestrogen resistance and progression of breast cancer. *Br J Cancer* **101**, 1824–1832.

75 Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res* **47** (D1), D941–D947.

76 Zhao SG, Chang SL, Erho N, Yu M, Lehrer J, Alshalalfa M, Speers C, Cooperberg MR, Kim W, Ryan CJ *et al.* (2017) Associations of luminal and basal subtyping of prostate cancer with prognosis and response to androgen deprivation therapy. *JAMA Oncol* **3**, 1663–1672.

77 Merdad A, Karim S, Schulten HJ, Dallol A, Buhmeida A, Al-Thubaity F *et al.* (2014) Expression of matrix metalloproteinases (MMPs) in primary human breast cancer: MMP-9 as a potential biomarker for cancer invasion and metastasis. *Anticancer Res* **34**, 1355–1366.

78 Radisky ES & Radisky DC (2015) Matrix metalloproteinases as breast cancer drivers and therapeutic targets. *Front Biosci (Landmark Ed)* **20**, 1144–1163.

79 Marie Vedeld H, Andresen K, Andrassy Eilertsen I, Nesbakken A, Seruca R, Gladhaug IP, Thiis-Evensen E, Rognum TO, Muri Boberg K & Lind GE (2015) The novel colorectal cancer biomarkers CDO1, ZSCAN18 and ZNF331 are frequently methylated across gastrointestinal cancers. *Int J Cancer* **136**, 844–853.

80 De Bessa SA, Salaorni S, Patrao DF, Neto MM, Brentani MM & Nagai MA (2006) JDP1 (DNAJC12/Hsp40) expression in breast cancer and its association with estrogen receptor status. *Int J Mol Med* **17**, 363–367.

81 McElwee JL, Mohanan S, Griffith OL, Breuer HC, Anguish LJ, Cherrington BD, Palmer AM, Howe LR, Subramanian V, Causey CP *et al.* (2012) Identification of PADI2 as a potential breast cancer biomarker and therapeutic target. *BMC Cancer* **12**, 500.

82 Wang H, Xu B, Zhang X, Zheng Y, Zhao Y & Chang X (2016) PADI2 gene confers susceptibility to breast cancer and plays tumorigenic role via ACSL4, BINC3 and CA9 signaling. *Cancer Cell Int* **16**, 61.

83 Tan X, Thapa N, Choi S & Anderson RA (2015) Emerging roles of PtdIns(4,5)P2–beyond the plasma membrane. *J Cell Sci* **128**, 4047–4056.

84 Bultsma Y, Keune WJ & Divecha N (2010) PIP4Kbeta interacts with and modulates nuclear localization of the high-activity PtdIns5P-4-kinase isoform PIP4Kalpha. *Biochem J* **430**, 223–235.

85 Keune WJ, Sims AH, Jones DR, Bultsma Y, Lynch JT, Jirstrom K, Landberg G & Divecha N (2013) Low PIP4K2B expression in human breast tumors correlates with reduced patient survival: a role for PIP4K2B in the regulation of E-cadherin expression. *Cancer Res* **73**, 6913–6925.

86 Russell MR, Graham C, D'Amato A, Gentry-Maharaj A, Ryan A, Kalsi JK, Ainley C, Whetton AD, Menon U, Jacobs I *et al.* (2017) A combined biomarker panel shows improved sensitivity for the early detection of ovarian cancer allowing the identification of the most aggressive type II tumours. *Br J Cancer* **117**, 666–674.

87 Enroth S, Berggrund M, Lycke M, Broberg J, Lundberg M, Assarsson E, Olovsson M, Stålberg K, Sundfeldt K & Gyllensten U (2019) High throughput

proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. *Commun Biol* **2**, 221.

88 Garczyk S, von Stillfried S, Antonopoulos W, Hartmann A, Schrauder MG, Fasching PA, Anzeneder T, Tannapfel A, Ergönenc Y, Knüchel R *et al.* (2015) AGR3 in breast cancer: prognostic impact and suitable serum-based biomarker for early cancer detection. *PLoS One* **10**, e0122106.

89 Hrstka R, Obacz J, Brychtova V, Podhorec J, Fabian P, Vojtesek B & Dobes P (2015) Anterior gradient protein 3 is associated with less aggressive tumors and better outcome of breast cancer patients. *Onco Targets Ther*, 1523.

90 Li L, Chen L, Zhang W, Liao Y, Chen J, Shi Y & Luo S (2017) Serum cytokine profile in patients with breast cancer. *Cytokine* **89**, 173–178.

91 Kikkawa Y, Enomoto-Okawa Y, Fujiyama A, Fukuhara T, Harashima N, Sugawara Y, Negishi Y, Katagiri F, Hozumi K, Nomizu M *et al.* (2018) Internalization of CD239 highly expressed in breast cancer cells: a potential antigen for antibody-drug conjugates. *Sci Rep* **8**, 6612.

92 Savci-Heijink CD, Halfwerk H, Hooijer GKJ, Koster J, Horlings HM, Meijer SL & van de Vijver MJ (2019) Epithelial-to-mesenchymal transition status of primary breast carcinomas and its correlation with metastatic behavior. *Breast Cancer Res Treat* **174**, 649–659.

93 Katz E, Dubois-Marshall S, Sims AH, Faratian D, Li J, Smith ES, Quinn JA, Edward M, Meehan RR, Evans EE *et al.* (2010) A gene on the HER2 amplicon, C35, is an oncogene in breast cancer whose actions are prevented by inhibition of Syk. *Br J Cancer* **103**, 401–410.

94 Yin K, Ba Z, Li C, Xu C, Zhao G, Zhu S & Yan G (2015) Overexpression of C35 in breast carcinomas is associated with tumor progression and lymphnode metastasis. *Biosci Trends* **9**, 386–392.

95 Zhao HB, Zhang XF, Wang HB & Zhang MZ (2017) Migration and invasion enhancer 1 (MIEN1) is overexpressed in breast cancer and is a potential new therapeutic molecular target. *Genet Mol Res* **16**, 1–9.

96 de Kruijf EM, Sajet A, van Nes JG, Putter H, Smit VT, Eagle RA, Jafferji I, Trowsdale J, Liefers GK, van de Velde CJ *et al.* (2012) NKG2D ligand tumor expression and association with clinical outcome in early breast cancer patients: an observational study. *BMC Cancer* **12**, 24.

97 El Nemer W, Colin Y & Le Van KC (2010) Role of Lu/BCAM glycoproteins in red cell diseases. *Transfus Clin Biol* **17**, 143–147.

98 Duncan JS, Stoller ML, Francl AF, Tissir F, Devenport D & Deans MR (2017) Celsr1 coordinates the planar polarity of vestibular hair cells during inner ear development. *Dev Biol* **423**, 126–137.

99 Geradts J, Groth J, Wu Y & Jin G (2016) Validation of an oligo-gene signature for the prognostic stratification of ductal carcinoma in situ (DCIS). *Breast Cancer Res Treat* **157**, 447–459.

100 Benusiglio PR, Pharoah PD, Smith PL, Lesueur F, Conroy D, Luben RN, Dew G, Jordan C, Dunning A, Easton DF *et al.* (2006) HapMap-based study of the 17q21 ERBB2 amplicon in susceptibility to breast cancer. *Br J Cancer* **95**, 1689–1695.

101 Staaf J, Jonsson G, Ringner M, Vallon-Christersson J, Grabau D, Arason A, Gunnarsson H, Agnarsson BA, Malmström PO, Johannsson OT *et al.* (2010) High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer. *Breast Cancer Res* **12**, R25.

102 Kpetemey M, Chaudhary P, Van Treuren T & Vishwanatha JK (2016) MIEN1 drives breast tumor cell migration by regulating cytoskeletal-focal adhesion dynamics. *Oncotarget* **7**, 54913–54924.

103 Petersen TN, Brunak S, von Heijne G & Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**, 785–786.

104 Bernemann TM, Podda M, Wolter M & Boehncke WH (2000) Expression of the basal cell adhesion molecule (B-CAM) in normal and diseased human skin. *J Cutan Pathol* **27**, 108–111.

105 Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al.* (2000) Molecular portraits of human breast tumours. *Nature* **406**, 747–752.

106 Schon M, Klein CE, Hogenkamp V, Kaufmann R, Wienrich BG & Schon MP (2000) Basal-cell adhesion molecule (B-CAM) is induced in epithelial skin tumors and inflammatory epidermis, and is expressed at cell-cell and cell-substrate contact sites. *J Invest Dermatol* **115**, 1047–1053.

107 Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* **98**, 10869–10874.

108 Adam PJ, Berry J, Loader JA, Tyson KL, Craggs G, Smith P, De Belin J, Steers G, Pezzella F, Sachsenmeir KF *et al.* (2003) Arylamine N-acetyltransferase-1 is highly expressed in breast cancers and conveys enhanced growth and resistance to etoposide in vitro. *Mol Cancer Res* **1**, 826–835.

109 Fletcher GC, Patel S, Tyson K, Adam PJ, Schenker M, Loader JA, Daviet L, Legrain P, Parekh R, Harris AL *et al.* (2003) hAG-2 and hAG-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene C4.4a and dystroglycan. *Br J Cancer* **88**, 579–585.

110 Määttä M, Bützow R, Luostarinen J, Petäjäniemi N, Pihlajaniemi T, Salo S, Miyazaki K, Autio-Harmainen H & Virtanen I (2005) Differential expression of laminin isoforms in ovarian epithelial carcinomas suggesting different origin and providing tools for differential diagnosis. *J Histochem Cytochem* **53**, 1293–1300.

111 Evans EE, Henn AD, Jonason A, Paris MJ, Schiffhauer LM, Borrello MA, Smith ES, Sahasrabudhe DM & Zauderer M (2006) C35 (C17orf37) is a novel tumor biomarker abundantly expressed in breast cancer. *Mol Cancer Ther* **5**, 2919–2930.

112 Tozlu S, Girault I, Vacher S, Vendrell J, Andrieu C, Spyratos F, Cohen P, Lidereau R & Bieche I (2006) Identification of novel genes that co-cluster with estrogen receptor alpha in breast tumor biopsy specimens, using a large-scale real-time reverse transcription-PCR approach. *Endocr Relat Cancer*, 1109–1120.

113 Kikkawa Y, Sudo R, Kon J, Mizuguchi T, Nomizu M, Hirata K & Mitaka T (2008) Laminin α5 mediates ectopic adhesion of hepatocellular carcinoma through integrins and/or Lutheran/basal cell adhesion molecule. *Exp Cell Res* **314**, 2579–2590.

114 Li K, Mandai M, Hamanishi J, Matsumura N, Suzuki A, Yagi H, Yamaguchi K, Baba T, Fujii S & Konishi I (2009) Clinical significance of the NKG2D ligands, MICA/B and ULBP2 in ovarian cancer: high expression of ULBP2 is an indicator of poor prognosis. *Cancer Immunol Immunother* **58**, 641–652.

115 Yu KH, Barry CG, Austin D, Busch CM, Sangar V, Rustgi AK & Blair IA (2009) Stable isotope dilution multidimensional liquid chromatography-tandem mass spectrometry for pancreatic cancer serum biomarker discovery. *J Proteome Res* **8**, 1565–1576.

116 McGilvray RW, Eagle RA, Rolland P, Jafferji I, Trowsdale J & Durrant LG (2010) ULBP2 and RAET1E NKG2D ligands are independent predictors of poor prognosis in ovarian cancer patients. *Int J Cancer* **127**, 1412–1420.

117 Chang YT, Wu CC, Shyr YM, Chen TC, Hwang TL, Yeh TS, Chang KP, Liu HP, Liu YL, Tsai MH *et al.* (2011) Secretome-based identification of ULBP2 as a novel serum marker for pancreatic cancer detection. *PLoS One* **6**, e20029.

118 Farrah T, Deutsch EW, Omenn GS, Campbell DS, Sun Z, Bletz JA, Mallick P, Katz JE, Malmström J, Ossola R *et al.* (2011) A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics* **10**, M110 006353.

119 King ER, Tung CS, Tsang YT, Zu Z, Lok GT, Deavers MT, Malpica A, Wolf JK, Lu KH, Birrer MJ *et al.* (2011) The anterior gradient homolog 3 (AGR3) gene is associated with differentiation and survival in ovarian cancer. *Am J Surg Pathol* **35**, 904–912.

120 Ammerpohl O, Pratschke J, Schafmayer C, Haake A, Faber W, von Kampen O *et al.* (2012) Distinct DNA methylation patterns in cirrhotic liver and hepatocellular carcinoma. *Int J Cancer* **130**, 1319–1328.

121 Liao S, Desouki MM, Gaile DP, Shepherd L, Nowak NJ, Conroy J, Barry WT & Geradts J (2012) Differential copy number aberrations in novel candidate genes associated with progression from in situ to invasive ductal carcinoma of the breast. *Genes Chromosomes Cancer* **51**, 1067–1078.

122 Vaarala MH, Hirvikoski P, Kauppila S & Paavonen TK (2012) Identification of androgen-regulated genes in human prostate. *Mol Med Rep* **6**, 466–472.

123 Yamaguchi K, Chikumi H, Shimizu A, Takata M, Kinoshita N, Hashimoto K, Nakamoto M, Matsunaga S, Kurai J, Miyake N *et al.* (2012) Diagnostic and prognostic impact of serum-soluble UL16-binding protein 2 in lung cancer patients. *Cancer Sci* **103**, 1405–1413.

124 Chitadze G, Bhat J, Lettau M, Janssen O & Kabelitz D (2013) Generation of soluble NKG2D ligands: proteolytic cleavage, exosome secretion and functional implications. *Scand J Immunol* **78**, 120–129.

125 Emerling BM, Hurov JB, Poulogiannis G, Tsukazawa KS, Choo-Wing R, Wulf GM, Bell EL, Shim HS, Lamia KA, Rameh LE *et al.* (2013) Depletion of a putatively druggable class of phosphatidylinositol kinases inhibits growth of p53-null tumors. *Cell* **155**, 844–857.

126 Endo Y, Toyama T, Takahashi S, Yoshimoto N, Iwasa M, Asano T, Fujii Y & Yamashita H (2013) miR-1290 and its potential targets are associated with characteristics of estrogen receptor α-positive breast cancer. *Endocr Relat Cancer* **20**, 91–102.

127 Kaucka M, Plevova K, Pavlova S, Janovska P, Mishra A, Verner J, Procházková J, Krejcí P, Kotasková J, Ovesná P *et al.* (2013) The planar cell polarity pathway drives pathogenesis of chronic lymphocytic leukemia by the regulation of B-lymphocyte migration. *Cancer Res* **73**, 1491–1501.

128 Leung TH, Wong SC, Chan KK, Chan DW, Cheung AN & Ngan HY (2013) The interaction between C35 and DeltaNp73 promotes chemo-resistance in ovarian cancer cells. *Br J Cancer* **109**, 965–975.

129 Brychtova V, Zampachova V, Hrstka R, Fabian P, Novak J, Hermanova M & Vojtesek B (2014) Differential expression of anterior gradient protein 3 in intrahepatic cholangiocarcinoma and hepatocellular carcinoma. *Exp Mol Pathol* **96**, 375–381.

130 Kikkawa Y, Miwa T, Tanimizu N, Kadoya Y, Ogawa T, Katagiri F, Hozumi K K, Nomizu M, Mizuguchi

T, Hirata K *et al.* (2014) Soluble Lutheran/basal cell adhesion molecule is detectable in plasma of hepatocellular carcinoma patients and modulates cellular interaction with laminin-511 in vitro. *Exp Cell Res* **328**, 197–206.

131 Zhou YF, Xu LX, Huang LY, Guo F, Zhang F, Hh XY, Yuan YZ & Yao WY (2014) Combined detection of serum UL16-binding protein 2 and macrophage inhibitory cytokine-1 improves early diagnosis and prognostic prediction of pancreatic cancer. *Oncol Lett* **8**, 2096–2102.

132 Rajendiran S, Kpetemey M, Maji S, Gibbs LD, Dasgupta S, Mantsch R, Hare RJ & Vishwanatha JK (2015) MIEN1 promotes oral cancer progression and implicates poor overall survival. *Cancer Biol Ther* **16**, 876–885.

133 Yehia L, Niazi F, Ni Y, Ngeow J, Sankunny M, Liu Z, Wei W, Mester JL, Keri RA, Zhang B *et al.* (2015) Germline heterozygous variants in SEC23B are associated with Cowden syndrome and enriched in apparently sporadic thyroid cancer. *Am J Hum Genet* **97**, 661–676.

134 Bartolini A, Cardaci S, Lamba S, Oddo D, Marchio C, Cassoni P, Amoreo CA, Corti G, Testori A, Bussolino F *et al.* (2016) BCAM and LAMA5 mediate the recognition between tumor cells and the endothelium in the metastatic spreading of KRAS-mutant colorectal cancer. *Clin Cancer Res* **22**, 4923–4933.

135 Chen J, Zhu XX, Xu H, Fang HZ & Zhao JQ (2016) Expression and prognostic significance of unique ULBPs in pancreatic cancer. *Onco Targets Ther* **9**, 5271–5279.

136 Giulietti M, Occhipinti G, Principato G & Piva F (2016) Weighted gene co-expression network analysis reveals key genes involved in pancreatic ductal adenocarcinoma development. *Cell Oncol (Dordr)* **39**, 379–388.

137 Lin X, Huang M, Xie F, Zhou H, Yang J & Huang Q (2016) Gemcitabine inhibits immune escape of pancreatic cancer by down regulating the soluble ULBP2 protein. *Oncotarget* **7**, 70092–70099.

138 Xu H, Ma J, Wu J, Chen L, Sun F, Qu C, Zheng D & Xu S (2016) Gene expression profiling analysis of lung adenocarcinoma. *Braz J Med Biol Res* **49**.

139 Agosto-Arroyo E, Isayeva T, Wei S, Almeida JS & Harada S (2017) Differential gene expression in ductal carcinoma in situ of the breast based on ERBB2 status. *Cancer Control* **24**, 102–110.

140 Demirkol S, Gomceli I, Isbilen M, Dayanc BE, Tez M, Bostanci EB, Turhan N, Akoglu M, Ozyerli E, Durdu S *et al.* (2017) A combined ULBP2 and SEMA5A expression signature as a prognostic and predictive biomarker for colon cancer. *J Cancer* **8**, 1113–1122.

141 Samanta S, Tamura S, Dubeau L, Mhawech-Fauceglia P, Miyagi Y, Kato H, Lieberman R, Buckanovich RJ, Lin YG & Neamati N (2017) Expression of protein disulfide isomerase family members correlates with tumor progression and patient survival in ovarian cancer. *Oncotarget* **8**, 103543–103556.

142 Vastrad B, Vastrad C, Godavarthi A & Chandrashekar R (2017) Molecular mechanisms underlying gliomas and glioblastoma pathogenesis revealed by bioinformatics analysis of microarray data. *Med Oncol* **34**, 182.

143 Abouelghar A, Hasnah R, Taouk G, Saad M & Karam M (2018) Prognostic values of the mRNA expression of natural killer receptor ligands and their association with clinicopathological features in breast cancer patients. *Oncotarget* **9**, 27171–27196.

144 Qiu C, Wang Y, Wang X, Zhang Q, Li Y, Xu Y, Jin C, Bu H, Zheng W, Yang X *et al.* (2018) Combination of TP53 and AGR3 to distinguish ovarian high-grade serous carcinoma from low-grade serous carcinoma. *Int J Oncol* **52**, 2041–2050.

145 Lima K, Coelho-Silva JL, Kinker GS, Pereira-Martins DA, Traina F, Fernandes PACM, Markus RP, Lucena-Araujo AR & Machado-Neto JA (2019) PIP4K2A and PIP4K2C transcript levels are associated with cytogenetic risk and survival outcomes in acute myeloid leukemia. *Cancer Genet* **233-234**, 56–66.

146 Xu Q, Shao Y, Zhang J, Zhang H, Zhao Y, Liu X, Guo Z, Chong W, Gu F & Ma Y (2020) Anterior gradient 3 promotes breast cancer development and chemotherapy response. *Cancer Res Treat* **52**, 218–245.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Figure showing the effects of batch correction on sample clustering and variance of protein abundance.

Fig. S2. Six set-wise heatmaps (A-F) with differentially abundant proteins from DAA comparisons.

Table S1. List of the antibodies used in this study.

Table S2. Clinicopathological characteristics of the breast cancer TIF samples which were examined in this study.

Table S3. Proteins returned from LASSO regression and Random Forest models.

Table S4. An alphabetical list of proteins and their detection profile in tumor, normal, and fat interstitial fluid samples (TIF, NIF, and FIF, respectively).

Table S5. Table with fractions (percentages) of BC samples within clusters belonging to each clinicopathological subgroup.

Table S6. An alphabetical list of 174 (176) differentially abundant TIF proteins according to limma analysis.

Table S7. The results of the differential abundance analysis with limma.

Table S8. Results from differential abundance analysis, LASSO regression, and random forest.

Table S9. Comparison of the expression profiles for eight of the TIF proteins identified in the present study with the PAM50 prognostic signature.

Table S10. Protein-protein interaction networks from analysis using the STRING database.

Table S11. A panel of 10 proteins identified in TIF samples in the present study are segregated according to immunohistochemistry (IHC) scores from paired tumor tissues and BC subtype (Her2, luminal, TNBC).

Table S12. Test-statistics and p-values from ordinal logistic regression with ten protein candidates.

Table S13. Expression profiles for 10 selected proteins in 33 normal human tissues (MNO661; Pantomics, USA) according to immunohistochemistry (IHC) scores.

Supplementary Material