Breast Cancer Classification Based on Proteotypes Obtained by SWATH Mass Spectrometry

Graphical Abstract

Highlights
- Proteotyping of 96 breast tumors by SWATH mass spectrometry
- Three key proteins for breast tumor classification
- Varying degrees of heterogeneity within conventional breast cancer subtypes
- Generally modest correlation between protein and transcript levels in tumor tissue

Authors
Pavel Bouchal, Olga T. Schubert, Jakub Faktor, ..., Eva Budinska, Rudolf Nenutil, Ruedi Aebersold

Correspondence
bouchal@chemi.muni.cz (P.B.), aebersold@imsb.biol.ethz.ch (R.A.)

In Brief
Bouchal et al. explore and confirm the suitability of SWATH-MS for proteotyping of human tumor samples at relatively high throughput. Results indicate that proteotype-based classification resolves more variability than is apparent from conventional subtyping and potentially improves current classification.
Breast Cancer Classification Based on Proteotypes Obtained by SWATH Mass Spectrometry

Pavel Bouchal,1,2,10* Olga T. Schubert,3,4 Jakub Faktor,2 Lenka Capkova,1 Hana Imrichova,1,5 Karolina Zoufalova,1 Vendula Paralova,1 Roman Hrstka,1 Yansheng Liu,1 Holger Alexander Ebhardt,3,7 Eva Budinska,2,8 Rudolf Nenutil,2 and Ruedi Aebersold1,9,*

1Department of Biochemistry, Faculty of Science, Masaryk University, Brno, Czech Republic
2Regional Centre for Applied Molecular Oncology, Masaryk Memorial Cancer Institute, Brno, Czech Republic
3Department of Biology, Institute of Molecular Systems Biology, Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland
4Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA
5Center for Human Genetics, University of Leuven, Leuven, Belgium
6Department of Pharmacology, Yale Cancer Biology Institute, Yale University School of Medicine, West Haven, CT, USA
7Systems Biology Ireland, University College Dublin, Dublin, Ireland
8Research Centre for Toxic Compounds in the Environment, Faculty of Science, Masaryk University, Brno, Czech Republic
9Faculty of Science, University of Zurich, Zurich, Switzerland
10Lead Contact

*Correspondence: bouchal@chemi.muni.cz (P.B.), aebersold@imsb.biol.ethz.ch (R.A.)

https://doi.org/10.1016/j.celrep.2019.06.046

SUMMARY

Accurate classification of breast tumors is vital for patient management decisions and enables more precise cancer treatment. Here, we present a quantitative proteotyping approach based on sequential windowed acquisition of all theoretical fragment ion spectra (SWATH) mass spectrometry and establish key proteins for breast tumor classification. The study is based on 96 tissue samples representing five conventional breast cancer subtypes. SWATH proteotype patterns largely recapitulate these subtypes; however, they also reveal varying heterogeneity within the conventional subtypes, with triple negative tumors being the most heterogeneous. Proteins that contribute most strongly to the proteotype-based classification include INPP4B, CDK1, and ERBB2 and are associated with estrogen receptor (ER) status, tumor grade status, and HER2 status. Although these three key proteins exhibit high levels of correlation with transcript levels (R > 0.67), general correlation did not exceed R = 0.29, indicating the value of protein-level measurements of disease-regulated genes. Overall, this study highlights how cancer tissue proteotyping can lead to more accurate patient stratification.

INTRODUCTION

Despite the progress achieved in early cancer diagnosis and therapy, many patients develop fatal disease. This also applies to breast cancer, even though it is one of the best characterized malignant diseases. Breast cancer is currently classified into five intrinsic subtypes, typically using immunohistological markers (estrogen receptor [ER], progesterone receptor [PR], HER2 gene, and/or ERBB2 protein status), tumor grade, and/or proliferation. We will refer to these subtypes as “conventional subtypes”; they have been defined as follows: luminal A (ER+, HER2−, low proliferation), luminal B HER2+ (ER+, HER2+, high proliferation), luminal B HER2+ (ER−, HER2+, high proliferation), HER2 enriched (ER−, HER2+, high proliferation), and triple negative (ER−, PR−, HER2−, high proliferation; Brouckaert et al., 2013; Lam et al., 2014; Parise and Caggiano, 2014). This classification guides decisions for the adjuvant therapy, which, however, fails in a substantial proportion of cases due to cancer recurrence, therapy resistance, and/or metastasis (Parise and Caggiano, 2014). The development of advanced, generalized disease despite the therapy guided by the tumor classification into the subtypes described above indicates that the current classification scheme may not fully capture the genetic and molecular status of the cancer and that a refined classification system might better predict which patient groups respond best to the range of available therapies.

Nowadays, the search for better tumor classifiers significantly concentrates on the application of omics approaches, which are able to analyze thousands of gene sequences, gene transcripts, or proteins in a single experiment. The biochemical effector molecules in cells are proteins, and their direct measurement is, therefore, in principle preferable over the inference of protein quantities from transcript measurements (expression arrays and RNA sequencing). However, the commonly used proteomic approaches based on mass spectrometry analysis in data-dependent acquisition (DDA) mode are often hampered by limited consistency and quantitative accuracy and are therefore less suitable for application to clinical cohorts of significant size. In contrast, targeted proteomic technologies overcome some of these limitations and provide improved quantification precision and reproducibility (Pernikárová and Bouchal, 2015). Kennedy and colleagues (Kennedy et al., 2014) recently demonstrated the ability of the targeted proteomic technique selected or multiple reaction monitoring (S/MRM) to quantify...
319 breast-cancer-associated proteins with high inter-laboratory reproducibility. The data discriminated basal versus luminal breast cancer phenotypes and largely correlated with ER levels in 30 cell lines.

To increase the number of proteins reproducibly quantified across samples, in the present study, we use a highly multiplexed mode of targeted proteomics, sequential windowed acquisition of all theoretical fragment ion spectra-mass spectrometry (SWATH-MS), a next-generation proteomics approach developed by Gillet and colleagues (Gillet et al., 2012). For the targeted analysis of the acquired data, we built a comprehensive breast-cancer-specific SWATH assay library. We applied the SWATH-MS technique to obtain digital proteome maps (or „proteotypes“) for a set of 96 breast tumor lysates (Data S1) and classified them into five proteotype-based subtypes using a conditional reference tree algorithm (Hothom et al., 2006). The algorithm found three key proteins that are highly effective for group separation; the agreement between our proteotype-based subtypes and the conventional subtypes is 84%. The triple-negative subtype showed the highest degree of heterogeneity of protein expression. In addition to allowing a more refined classification of breast cancer subtypes, the obtained SWATH-MS data allowed us to compare protein and transcript levels of over 2,700 genes. Although the correlation of protein and transcript levels was low for most differentially expressed genes, it was strong for the three classifying proteins. In conclusion, this study describes the application of the SWATH-MS technique to generate large-scale quantitative proteomics profiles of breast cancer tissues for tumor classification. Discrepancies between the conventional tumor subtypes and our proteotype-based subtypes indicate patients that could potentially benefit from different treatment strategies.

RESULTS

Generation of an Assay Library for Quantifying Breast-Cancer-Associated Proteins by SWATH-MS

To extract quantitative protein information from SWATH-MS datasets acquired from breast cancer patient tissue samples in a targeted manner, we generated an extensive spectral library based on samples of all classical breast cancer subtypes described above and fractionated pools thereof. From this spectral library, we obtained reference spectra for 28,233 proteotypic peptides and their modified variants (false discovery rate [FDR] < 0.01), representing 2,842 proteins across all pooled samples of each of the five subtypes. For each subtype, lymph-node-negative and lymph-node-positive samples were pooled separately, generating ten sample pools in total. Using the SWATH assay library described above, we were able to extract quantitative data for 25,278 proteotypic peptides and their modified variants, representing 2,842 proteins across all individual samples. These 2,842 consistently quantified proteins cover the majority of molecular processes known to be involved in breast cancer (Figure S1).

Comparison of Prototype-Based Subtypes and Conventional Subtypes of Breast Cancer

Using the thus generated proteotypes for 96 samples, we first asked to what extent tumor classification based on proteotypes correlated with the conventional subtype classification. We performed unsupervised hierarchical clustering on the proteotypes of the pooled samples. Figure 1A shows that pools of lymph-node-positive and negative samples of each subtype clustered closely together, indicating high reproducibility of our measurements. Moreover, clustering revealed proteotype similarity between less aggressive luminal A and luminal B subtypes, whereas the more aggressive HER2 and triple-negative subtypes formed a separate cluster. The luminal B HER2+ group was more similar to the cluster with high aggressiveness, in agreement with its worse therapy response (Figure 1A).

Next, we systematically correlated the quantitative proteotypes of the 96 individually measured breast cancer samples and ordered the resulting correlation coefficients according to the classical tumor subtypes (Figure 1B). Spearman correlation of proteomic profiles across the entire dataset was high (R > 0.97). The highest intra-group correlation of proteotypes was within luminal A subtype (R = 0.9900). Very high correlation was also observed within the luminal B subtype (both HER2-, R = 0.9866, and HER2+, R = 0.9878) and between luminal A and luminal B subtypes (R = 0.9865; Figure 1B). Furthermore, we found a high correlation of some samples of the HER2-enriched subtype with some (mostly lymph-node-positive) luminal B HER2+ samples, indicating that a higher degree of similarity in HER2+ tumors of luminal B and HER2-enriched subtypes were apparent from the proteotype. The group of triple-negative tumors exhibited slightly lower inter- (R < 0.9852) and particularly intra-group (R = 0.9840) correlation, potentially indicating tumor heterogeneity not captured by the conventional tumor classification. In summary, we found that clustering by proteotypes closely recapitulates conventional tumor subtyping, but we also found that some of these subtypes are more heterogeneous (triple-negative tumors) than others (Figure 1B).
Pathways and Proteins Associated with Key Breast Cancer Characteristics

Having a large, high-quality proteomic dataset at hand, we were interested in identifying pathways and proteins that are important for breast cancer biology and progression. We first identified proteins that are differentially expressed in tumors of different ER status, tumor grade, HER2 status, or lymph node status (Data S3). We then used gene set enrichment analysis (GSEA) to find pathways that are enriched among the most differentially abundant proteins in these comparisons (Figure 2). Among these, there were several pathways known to be associated with the particular phenotype, for example, an enrichment of the nuclear factor xB (NF-xB) pathway in ER+ tumors, in agreement with its role in proliferation and metastasis of luminal tumors (Azim et al., 2015; Bouchal et al., 2015; Pratt et al., 2009). The list of pathways enriched in high-grade tumors was led by the MCM pathway, which includes pro-proliferation proteins of the MCM family regulating cyclin-dependent kinases and DNA replication (Shetty et al., 2005; Wojnar et al., 2010). In HER2+ tumors, we found an enrichment of proteins belonging to the VEGF pathway, namely seven upregulated subunits of eukaryotic translation initiation factors 2 and 2B, which are known to be regulated by HER2 (Sequeira et al., 2009). In lymph-node-positive tumors, we found members of the CARM1 and regulation of the ER pathway (CARM ER) to be enriched, potentially indicating an involvement of chromatin-remodeling factors in breast cancer progression and metastasis (Wang et al., 2014). All these and further enriched pathways shown in Figure 2 could be highly relevant for breast cancer biology and warrant further investigation as potential targets of breast cancer therapy.

Selection of Discriminant Proteins for Improved Classification of Breast Cancer Subtypes

To examine the potential of proteotyping for breast cancer classification, we next constructed a decision tree to classify the 96 tumors into the five conventional subtypes based on their proteotypes. We started by selecting the most differentially abundant proteins (log2FC > 1.5; FDR-adj. p < 0.05) from the following comparisons: ER+ versus ER-/C0 (8 proteins); grade 3 versus grade 1 (2 proteins); HER2+ versus HER2-/C0 (2 proteins); luminal B versus luminal A (3 proteins); luminal B HER2+ versus luminal A (3 proteins); HER2-enriched versus luminal A (7 proteins); triple-negative versus luminal A (5 proteins); and
HER2-enriched versus luminal B (2 proteins). This procedure resulted in a list of 22 key proteins (partially overlapping among different comparisons). In a next step, we applied a recursive partitioning algorithm for continuous data in a conditional inference framework (Hothorn et al., 2006). The algorithm automatically selected discriminant proteins from the protein list and provided their quantitative thresholds as well as the structure of the decision tree. The algorithm generated a decision tree with three key nodes (Figure 3A), representing three key proteins: type II inositol 3,4-bisphosphate 4-phosphatase (INPP4B); cyclin-dependent kinase 1 (CDK1); and receptor tyrosine-protein kinase erbB-2 (ERBB2). Importantly, the differential expression of the selected proteins reflects key clinical parameters defining breast cancer subtypes: ER status (INPP4B; Figure 3B); tumor grade (CDK1; Figure 3C); and HER2 status (ERBB2; Figure 3D). Furthermore, we found that the proteotype-based decision tree assigned 84% of the tumors into their diagnosed conventional subtypes (Figure 3A).

Validation of the Three Key Proteins Selected by the Decision Tree

We next asked whether the changes in protein levels of the three key proteins from the decision tree, INPP4B, CDK1, and ERBB2, have general discriminative potential and biological validity beyond our 96-patient dataset. Analysis of a published proteomic dataset of 60 human tumor cell lines (http://proteomics.wzw.tum.de/nci60) confirmed high levels of INPP4B protein in ER+ breast cancer cell lines (MCF-7 and T47D), and no INPP4B protein was found in ER−/C0 breast cancer cell lines (MDA-MB-231, MDA-MB-468, BT549, and HS 578T), supporting the link between INPP4B and ER status. CDK1 and ERBB2 proteins were not covered in this reference dataset. We furthermore compared our protein-level data with gene expression data in five published microarray datasets (883 patients; Figure S3; Haibe-Kains et al., 2012) and a published RNA sequencing data-set (1,078 patients) by The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov). This analysis confirmed the connection of INPP4B with ER status, CDK1 with tumor grade, and ERBB2 with HER2 status (Figures 4 and S4). Furthermore, we found that gene expression of INPP4B, CDK1, and ERBB2 was statistically significantly connected with patient survival in the same manner as the commonly used reference genes ESR1 (for ER status) and MKI67 (for tumor grade or proliferation; Figure S5).

Higher Level of ERBB2 in ER+/HER2− versus ER+/HER2− Tumors

An interesting feature of our decision tree is that the algorithm decided between two HER2− subtypes based on ERBB2
Figure 3. Classification of Breast Cancer Patients Based on Protein Levels in Tumor Tissue

(A) Decision tree classification. The top panel shows the decision tree generated from 22 proteins selected from proteotypes of 96 patients (see Data S1A and S1B for details). The bar plots (bottom part) show the number of patients, classified by the protein-based decision tree, that coincide with the conventional subtype classification.

(legend continued on next page)
protein levels: whereas lower levels of ERBB2 protein seem to be associated with ER+/HER2+ grade 3 tumors, higher levels were found in ER-/HER2+ grade 3 tumors (Figure 3A).

To test whether this observation is of general validity, we manually validated the SWATH-MS-based protein quantification and performed independent analyses at both protein and transcript level. Transcript-level analysis of the same 96 tumor samples described in this study (Bouchal et al., 2015), transcript-level analysis in four additional datasets of a total of 116 tumor samples (Haibe-Kains et al., 2012), as well as immunohistochemistry (IHC) in an independent tumor collection of 78 patients (described in STAR Methods) all confirmed a statistically significantly increased level of ERBB2 in ER-/HER2+ versus ER+/HER2+ tumors (Figure 5). This observation supports the notion that proteotypes potentially reveal finer graded classification than provided by conventional subtyping.

Analysis of Global Correlation between Proteins and Transcripts

To see how our protein-level data correlate with transcript-level data globally, we compared our comprehensive SWATH-MS dataset against the five microarray datasets of 883 patients mentioned above (Haibe-Kains et al., 2012; see Data S4 for details). We performed 475,755 individual comparisons of overlaps of differentially abundant proteins (FDR-adj. p < 0.05) versus their cognate transcripts (with the same trend) for 2,782 matching transcript-protein pairs between patient groups with different subtype, ER status, HER2 status, tumor grade, and lymph node status (Data S4). Overall, 6% of protein-level observations and 7%–15% of transcript-level observations (depending on the set of patients) exhibited statistically significant changes (Data S4B). Of these, 13%–28% of differentially abundant proteins also showed a statistically significant change with the same direction on the transcript level. From the reverse perspective, 9%–18% of significantly regulated transcripts showed a significant change with the same trend also on protein level. The global correlation coefficients for fold changes between transcripts and proteins ranged from R = 0.17 to R = 0.29, depending on the dataset (Figure 6A). In contrast, the correlation for the three key proteins from the decision tree was very high, with correlation coefficients from R = 0.67 to R = 0.81 (Figure 6B). A decision tree constructed from the five independent transcriptomics datasets using expression data for 1,036 genes resulted in a tree with three nodes and similar structure (Figure S3). Taken together, although high correlation of protein and transcript levels was observed for the key proteins INPP4B, CDK1, and ERBB2, correlation and overlap of differentially expressed proteins and transcripts on a global scale was rather low, indicating the importance of protein-level measurements to study breast cancer biology.

DISCUSSION

High-Throughput Proteotyping by SWATH-MS as a Next-Generation Approach for Cancer Classification

The currently used classification of breast cancer tissues primarily relies on semiquantitative IHC, which is based on manual evaluation of antibody-stained tissue sections by a pathologist. Transcript-level approaches have been used for expression profiling of breast-cancer-associated genes and classification; however, as confirmed by our data, gene expression does not generally reflect levels of proteins. Protein-level quantification, although technically more difficult, is hence expected to provide the most relevant information. In this study, we employed a recently established massively parallel targeted proteomics technique, SWATH-MS, for the classification of human breast cancer tissues. The technique generally requires no more than 1 or 2 μg of total peptide sample and is capable of analyzing tissue samples obtained by needle biopsy (Guo et al., 2015). Moreover, it has good quantitative accuracy with high specificity due to targeted MS/MS data extraction (Gillet et al., 2012), low cost per run, and relatively high sample throughput, enabling the analysis of 10–24 samples per day. The hereby established proteotypes mostly recapitulated the five conventional subtypes, confirming the general applicability of proteotyping for the identification of cancer subtypes. The inconsistencies between the proteotype-based and conventional classification might reflect further breast cancer subtypes (Prat et al., 2015), which could, for example, arise from additional genetic mutations. This is well illustrated by the TP53 mutation status in our 96 tumor samples: although 50% of tumors with more aggressive subtypes (triple-negative, HER2-enriched, and luminal B HER2+) had mutations in TP53, less aggressive luminal B and luminal A subtypes included only 12.5% and 0.0% of TP53-mutated tumors, respectively (Data S1B). Proper classification of such additional mutational heterogeneity could help to improve diagnostics and treatment of breast cancer.

Advantages of SWATH-MS to Classify Breast Cancer Tumors

Several studies used proteomics approaches to classify breast cancer tissues (Lam et al., 2014), applying a range of methods, from surface-enhanced laser desorption-ionization time-of-flight (SELDI-TOF) MS (Bouchal et al., 2013; Brozkova et al., 2008) and stable isotope labelling with amino acids in cell culture (SILAC)-liquid chromatography-tandem mass spectrometry (LC-MS/MS) (Waldemarson et al., 2016) in breast cancer tumor samples to MS1-based, label-free quantification of secreted proteins in a cell-line panel (Pavlo et al., 2013). These studies confirm the utility of protein expression profiling for the identification of novel molecular markers to classify breast cancer. We previously analyzed the tumor samples of the 96 patients described in the...
present study using an iTRAQ-2DLC-MS/MS approach in an attempt to identify metastasis-associated proteins in low-grade breast cancer (Bouchal et al., 2015). In that study, we quantified 6% more proteins than in the current study (see also Figure S1); however, there, we were limited by significantly lower sample throughput, only allowing the analysis of pooled and not individual samples in a reasonable time, resulting in inferior statistical power. Compared to the iTRAQ method used earlier by us (Bouchal et al., 2015) and the Clinical Proteomic Tumor Analysis Consortium (Mertins et al., 2016), SWATH-MS has a better quantitative accuracy by avoiding the flattening of peptide ratios due to the use of the same iTRAQ reporter ions for quantification of co-isolated precursors. A recent study using SuperSILAC for the proteomic profiling of 40 breast cancer tissues (Tyanova et al., 2016) identified 10,138 endogenous proteins in total, but only a fraction of this number (2,588 proteins) was quantified across all samples (Figure S6). The study found a 19-protein signature discriminative for medium- and high-grade breast cancer subtypes, of which we consistently quantified 14 proteins in our SWATH-MS dataset of 96 patients. The abundance ranks of these 19 proteins in the two independent datasets (their 40 patients and our 96 patients) were highly similar (Table S1). Compared to the SuperSILAC approach, advantages of SWATH-MS are the lower cost and convenience of the label-free quantification but most importantly the consistent quantification of proteins across large sample sets (Figure S6). One of the gold-standard methods to profile proteins in clinical tissue samples is selected or multiple reaction monitoring (S/MRM). Of 319 breast-cancer-associated proteins quantified by S/MRM by Kennedy and colleagues (Kennedy et al., 2014), our SWATH-MS data cover 305 (96%). Similarly,
9 of 10 proteins associated with breast cancer biology (represented by 16 of 17 peptides) were quantified by S/MRM in the same set of tumors (Procházková et al., 2017) as in our current SWATH-MS dataset with high level of correlation (Spearman correlation coefficients 0.439–0.880 and p values 1.1 × 10⁻⁵ to 2.2 × 10⁻¹⁶; Data S5). This comparison well validates our SWATH-MS data using an independent method on individual tumor level. A strong correlation between SWATH-MS and S/MRM was demonstrated already in the first SWATH-MS publication (Gillet et al., 2012) and confirmed in other independent studies (Kockmann et al., 2016; Liu et al., 2013; Nakamura et al., 2016; Schmidlin et al., 2016). These studies include our recent comparison of S/MRM, pseudo-SRM/MRMMR, and SWATH-MS analytical parameters in selected samples from the same breast cancer tissue collection (Faktor et al., 2017). Based on the above data, it has been well demonstrated both experimentally in our breast tumor sample set and in the literature that SWATH-MS provides data highly correlated with S/MRM. In summary, our SWATH-MS-based strategy provided an advantageous combination of sample throughput, quantitative precision (Vowinckel et al., 2013), and proteome coverage in a large sample set. Applying the latest technical developments (e.g., ion mobility MS or faster Orbitrap-based instruments) may further improve the quantitative depth of SWATH-MS or similar data independent acquisition-based studies.

Biological Relevance of the Key Proteins Selected by the Decision Tree

The three key proteins identified by our decision tree are strongly associated with important clinical parameters, namely ER status (INPP4B), tumor grade (CDK1), and HER2 status (ERBB2; Figures 3B–3D). The receptor tyrosine protein kinase ERBB2 is the protein product of the HER-2/NEU gene and is routinely being used for breast cancer classification into HER2⁺ or HER2⁻ phenotypes. It also is the target of anti-HER2 therapy via US

Figure 5. Expression of ERBB2 Protein and Transcript in ER⁻/HER2⁻ versus ER⁺/HER2⁺ Breast Cancer Tissues

(A) Intensity of ERBB2 protein in SWATH-MS proteomics data (16 patients of grade 3; Data S1B).
(B) Immunohistochemistry for ERBB2 in an independent set of patients (78 patients of grade 2+3; Data S1C).
(C) Transcript-level analysis for ERBB2 in the same patients shown in (A) (16 patients of grade 3; Bouchal et al., 2015).
(D) Transcript-level analysis in four independent sets of grade 3 patients (DFHCC [n = 27], IRB [n = 23], PNC [n = 24], and SUPERTAM_HGU133PLUS_2 [n = 42]; Data S4D). Boxes are extended from the 25th to the 75th percentile, with a line at the median. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range (IQR) from the box. The individual points represent outliers or extreme values.
Food and Drug Administration (FDA)-approved humanized monoclonal antibody trastuzumab. The strong association of the ERBB2 protein with HER2 status in our dataset internally validates our proteomics data and design of the study. Of note, higher levels of ERBB2 protein observed here in ER+/HER2- versus ER+/HER2+ tumors are also consistent with the better response to therapy of ER+/HER2- versus ER+/HER2+ tumors (Bhargava et al., 2011).

INPP4B is known to dephosphorylate phosphatidylinositol 3,4-bisphosphate in the PI3K pathway, which co-activates cell growth and movement via Akt kinases (Malek et al., 2017). Hence, it serves as a tumor suppressor, and our earlier observations that it is significantly associated with ER+ tumors (Fedele et al., 2010) suggest that it should be explored as a candidate therapeutic target for ER+ breast cancer. The second of our key proteins, mitotic kinase CDK1, is known to accelerate critical processes required for mitosis (Enserink and Kolodner, 2010) and correlates with tumor grade (Chae et al., 2011). Moreover, inhibitors of the family members CDK4/6 have been FDA-approved for the treatment of metastatic breast cancer in a first-line setting (Bilgin et al., 2017). In conclusion, although this is a pilot discovery study and follow-ups with larger patient cohorts are required to further train and validate our classifier, our findings suggest that both INPP4B and CDK1 are promising alternative targets for anti-cancer therapy, as they exhibit a similar level of association with ER status and tumor grade as ERBB2 with HER2 status, which is already successfully targeted to treat HER2+ breast cancer patients. We would like to note that subsequent validation studies with S/MRM can now be set up easily as the information required for the acquisition methods can be obtained directly from the SWATH assay library. A small panel of validated protein biomarkers could be subsequently implemented as part of an IHC panel or assessed with other techniques used in the clinic.

Molecular Features Available in Proteotype and Not in Conventional Breast Cancer Subtypes

Although there was a high concordance (84%) between classification based on proteotypes and conventional subtypes, some samples with identical conventional subtype showed distinct proteotypes. We find such proteotype heterogeneity, for example, in triple-negative tumors, a genetically heterogeneous group that can indeed be sub-divided further. For example, Lehmann et al. (2011) suggested six subtypes based on gene expression profiling: basal-like 1; basal-like 2; immunomodulatory; mesenchymal; mesenchymal-stem-like; and luminal androgen receptor subtype; others suggested a similar division (Palma et al., 2015). We also found that some HER2-enriched tumors were more similar to luminal B HER2+ tumors than reflected in current subtyping; this is evident also in data from
Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **LEAD CONTACT AND MATERIALS AvAILABILITY**
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
  - Study design
  - Clinical tissue samples

**Global Correlation and Overlaps of Protein and Transcript Level Expression**

Abundance of our three key proteins INPP4B, CDK1, and ERBB2 across tumors of different ER status, tumor grade, and HER2 status correlated well with their respective transcript levels. However, when looking at all differentially expressed proteins and transcripts in our dataset, the overlap and correlation of fold changes was modest. Although it has been shown that protein levels are chiefly determined by transcript levels, particularly in steady state (Schwanhäusser et al., 2011), and that fold changes of transcript and protein levels between different human cell lines can show correlations as high as $R = 0.63$ (Lundberg et al., 2010), our comparisons of transcript and protein data suggest that this correlation is relatively low in human breast cancer tissues ($R = 0.29$). In general, the limited correlation between protein and transcript levels provides a substantial reason to focus on the analysis of proteins instead of transcripts, as these represent the true molecular effectors in cells.

**Conclusions**

This study explored and confirmed the suitability of SWATH-MS for proteotyping of human tumor samples at relatively high throughput. Although larger patient cohorts are needed for validation of the classifier, our results indicate that proteotype-based classification resolves more breast cancer subtypes than apparent from conventional subtyping and potentially improves current classification. Furthermore, the potential of data-independent approaches, such as SWATH-MS, for tissue classification is not limited to breast cancer but valid for other diseases and clinical specimens. Although we are not yet at a point to make clinical decisions based on proteotype data, our study may motivate further the research that in turn may result in more adequate treatment and better clinical outcomes. The breast cancer SWATH assay library and the high-quality proteomics dataset of 96 breast tumors will provide a valuable resource for future protein marker studies.

**STAR★METHODS**

Brozkova and colleagues (Brozkova et al., 2008). All these data indicate that proteotypes have the potential of enabling finer stratification of a patient population than conventional subtyping. Current clinical practice shows that treatment based on conventional subtypes is far from optimal with respect to patient response, and proteotyping can potentially provide a more accurate picture of the actual molecular state of a cancerous tissue and could thereby enable more precise or even personalized treatment.

**METHOD DETAILS**

- Tissue quality control via RNA integrity measurement
- Proteomics sample preparation
- LC-MS analyses for spectral library generation
- LC-MS analyses in SWATH-MS mode
- TP53 sequencing
- ERBB2 immunohistochemistry
- Validation of SWATH-MS quantitation through selected reaction monitoring

**QUANTIFICATION AND STATISTICAL ANALYSIS**

- SWATH-MS assay library generation
- SWATH-MS data processing in OpenSWATH
- Statistical analysis
- Relative quantification with MSstats and differential protein expression analysis between subtypes and related clinical-pathological variables
- KEGG pathway analysis
- Gene set enrichment analysis
- Correlation analysis of breast cancer tissue proteomes
- Construction of the decision tree
- Analysis of ERBB2 gene expression in the same sample set
- Analysis of gene expression in independent microarray and RNA-Seq sets of samples
- Analysis of patient survival
- Statistical analysis of the IHC data
- Correlation analysis of SWATH-MS and SRM quantitation

**DATA AND CODE AVAILABILITY**

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.06.046.

**ACKNOWLEDGMENTS**

We thank the women who provided their tissue for this research and all of the clinically related staff involved in their treatment. We are grateful to Dr. Hannes Röst and Dr. George Rosenberger for their help with SWATH-MS data analysis, to Dr. Ben Collins for the script to generate Data S3C, to Anna Pacinkova for preparing Figure 4D based upon data generated by the TCGA Research Network (http://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga?redirect=true), to Anna Najbrtova for data processing in Data S5A–S5C, to Dr. Josef Planeta for nano-LC column preparation, and to Dr. Philip J. Coates for critical reading of the manuscript. This work was supported by the Czech Science Foundation (project no. 17-09567S). Publication fee was covered by Grant Agency of Masaryk University. J.F. was supported by MEYS - NPS I - LO1413, R.N. by MH CZ - DRO (MCC1; L00209805), and E.B. by the CETOCOEN PLUS project and the RECETOX Research Infrastructure (LM2015051). R.A. was supported by the Swiss National Science Foundation (SNSF; 3100A-166435) and European Research Council grant 20140AdG 670821.
datasets; K.Z. performed KEGG pathway analyses; V.P. performed SWATH-MS data analyses in Skyline software; R.H. analyzed p53 status of tumors; Y.L. significantly contributed to SWATH-MS measurements and to manuscript preparation; H.A.E. contributed to SWATH-MS measurements and to manuscript preparation; E.B. constructed the decision tree and co-supervised all data analyses; R.N. designed and selected the set of tissues and contributed to data interpretation; and R.A. approved the joint study, provided computational and instrument capacity, and wrote and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 26, 2017
Revised: March 6, 2019
Accepted: June 12, 2019
Published: July 16, 2019

REFERENCES

Azim, H.A., Jr., Peccatori, F.A., Brohée, S., Branstetter, D., Loi, S., Viale, G., Piccart, M., Douglou, W.C., Pruneri, G., and Sotiriou, C. (2015). RANK-ligand (RANKL) expression in young breast cancer patients and during pregnancy. Breast Cancer Res. 17, 24.

Bhargava, R., Dabbs, D.J., Beriwal, S., Yildiz, I.A., Badve, P., Soran, A., Johnson, R.R., Brufsky, A.M., Lembersky, B.C., McGuire, K.P., and Ahrendt, G.M. (2011). Semiquantitative hormone receptor level influences response to trastuzumab-containing neoadjuvant chemotherapy in HER2-positive breast cancer patients. Breast Cancer Res. Treat. 129–016717.

Bourchulda, E., Boillot, A., Galland, M., Boissin, A., Canard, M., Perennou, D., et al. (2015). Expression of the indoleamine 2,3-dioxygenase gene in breast cancer: correlation with CD8+ tumor-infiltrating lymphocytes. Int. J. Cancer 137, 2183–2192.

Buchala, A., Kremser, G., Gross, S., Mehl, F., Zangerl, P., et al. (2015). Targeted proteomic analysis of breast cancer tissues by SWATH-MS: prospective digital proteome maps. Nat. Med. 21, 407–413.

Györfy, B., Lanczky, A., Eklund, A.C., Denkert, C., Budczies, J., Li, Q., and Szallasi, Z. (2010). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res. Treat. 123, 725–731.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A.C., Bonetti, G., Quackenbush, J., and Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. J. Natl. Cancer Inst. 104, 311–325.

Hoehn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. 15, 651–674.

Keller, A., Nesvizhskii, A.I., Kokker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392.

Koh, H., Knauf, P., Koh, C.C., Gillet, L.C., Wolski, W.E., Röst, H.L., Rosenberger, G., Collins, B.C., Blum, L.C., Gillesen, S., et al. (2015). Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat. Med. 21, 407–413.

Kockmann, T., Trachsel, C., Panse, C., Wältner, A., Selevsek, N., Grossmann, J., Wolski, W.E., and Schlappach, R. (2016). Targeted proteomics coming of age - SRM, PRM and DIA performance evaluated from a core facility perspective. Proteomics 16, 2183–2192.

Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., King, N., Stein, S.E., and Aebersold, R. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7, 655–667.

Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., Stein, S.E., and Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. Nat. Methods 5, 873–875.

Lam, S.W., Jimenez, C.R., and Boven, E. (2014). Breast cancer classification by proteomic technologies: current state of knowledge. Cancer Treat. Rev. 40, 129–138.
Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. 121, 2750–2767.

Liu, Y., Hüttenhain, R., Surinova, S., Gilliet, L.C., Mouritsen, J., Brunner, R., Navarro, P., and Aebersold, R. (2013). Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. Proteomics 13, 1247–1256.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Alégäs, C., Lundberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. Mol. Syst. Biol. 6, 450.

Malek, M., Kielkowska, A., Chessa, T., Anderson, K.E., Barneda, D., Pir, P., Nakaniishi, H., Eguchi, S., Koizumi, A., Sasaki, J., et al. (2017). PTEN regulates PI(3,4)P2 signaling downstream of class I PI3K. Mol. Cell 64, 488–498.

Mertins, P., Mani, D.R., Ruggles, K.V., Gilliet, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al. NCI CPTAC (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 55–62.

Nakamura, K., Hirayama-Kurogi, M., Ito, S., Kuno, T., Yoneyama, T., Obuchi, W., Terakasai, T., and Ohtsuki, S. (2016). Large-scale multiplex absolute protein quantification of drug-metabolizing enzymes and transporters in human intesti- ne, liver, and kidney microsomes by SWATH-MS: comparison with MRM/ SRM and HR/PRM/PRM. Proteomics 16, 2106–2116.

Palma, G., Fraisci, G., Chirico, A., Esposito, E., Siani, C., Saturnino, C., Arna, C., Ciiberto, G., Giordano, A., and D’Auito, M. (2015). Triple negative breast cancer: looking for the missing link between biology and treatments. Oncotarget 6, 26560–26574.

Parise, C.A., and Caggiano, V. (2014). Breast cancer survival defined by the ERVRHER2 subtypes and a surrogate classification according to tumor grade and immunohistochemical biomarkers. J. Cancer Epidemiol. 2014, 489251.

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

Perniková, V., and Bouchal, P. (2015). Targeted proteomics of solid cancers: from quantification of known biomarkers towards the digital proteome maps. Expert Rev. Proteomics 12, 651–667.

Planetas, J., Karásek, P., and Vojrosta, J. (2003). Development of packed capil- lary columns using carbon dioxide slurries. J. Sep. Sci. 26, 525–530.

Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L., Diez, M., Viladot, M., Arance, A., and Muñoz, M. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. Breast 24 (Suppl 2), S26–S35.

Pratt, M.A., Tibbo, E., Robertson, S.J., Jansson, D., Hurst, K., Perez-Irurtxueta, C., Lau, R., and Niu, M.Y. (2009). The canonical NF-kappaB pathway is required for formation of luminal mammary neoplasias and is activated in the mammary progenitor population. Oncogene 28, 2710–2722.

Procházková, I., Lenců, J., Fucíková, A., Dresler, J., Čapková, L., Hrstka, R., Nenutil, R., and Bouchal, P. (2017). Targeted proteomics driven verification of biomarker candidates associated with breast cancer aggressiveness. Biochim. Biophys. Acta. Proteins Proteomics 1865, 488–498.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a bioc- conductor package for differential expression analysis of digital gene expres- sion data. Bioinformatics 26, 139–140.

Rosenberger, G., Ludwig, C., Röst, H.L., Aebersold, R., and Malmström, L. (2014). aLFQ: an R-package for estimating absolute protein quantities from label-free LC-MS/MS proteomics data. Bioinformatics 30, 2511–2513.

Röst, H.L., Rosenberger, G., Navarro, P., Gilliet, L., Miladinović, S.M., Schubert, O.T., Wolseki, W., Collins, B.C., Malmström, J., Malmström, L., and Aebersold, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. 32, 219–223.

Röst, H.L., Liu, Y., D’Agostino, G., Zanella, M., Navarro, P., Rosenberger, G., Collins, B.C., Gilliet, L., Testa, G., Malmström, L., and Aebersold, R. (2016). TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat. Methods 13, 777–783.

Schmidlin, T., Garrigues, L., Lane, C.S., Mulder, T.C., van Doorn, S., Post, H., de Graaf, E.L., Lemeer, S., Heck, A.J., and Alkaear, A.F. (2016). Assessment of SRM, MRM(3), and DIA for the targeted analysis of phosphorylation dynamics in non-small cell lung cancer. Proteomics 16, 2193–2205.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature 473, 337–342.

Sequeira, S.J., Wen, H.C., Avivar-Valderas, A., Farias, E.F., and Aguirre-Ghiso, J.A. (2009). Inhibition of eIF2alpha dephosphorylation inhibits ErbB2-induced deregulation of mammalian acinar morphogenesis. BMC Cell Biol. 10, 64.

Shetty, A., Loddo, M., Fanshawe, T., Prevost, A.T., Sainsbury, R., Williams, G.H., and Stoebber, K. (2005). DNA replication licensing and cell cycle kinetics of normal and neoplastic breast. Br. J. Cancer 93, 1295–1300.

Shteynberg, D., Deutsch, E.W., Lam, H., Eng, J.K., Sun, Z., Tasman, N., Mendozra, L., Moritz, R.L., Aebersold, R., and Nesvizhskii, A.I. (2011). Prophet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol. Cell. Proteomics 10, M111.007690.

Tyanova, S., Albrechtsen, R., Kronqvist, P., Cox, J., Mann, M., and Geiger, T. (2016). Proteomic maps of breast cancer subtypes. Nat. Commun. 7, 18259.

Vowinckel, J., Capuano, F., Campbell, K., Deery, M.J., Lilley, K.S., and Ralser, M. (2013). The beauty of being (label)-free: sample preparation methods for SWATH-MS and next-generation targeted proteomics. F1000Res. 2, 272.

Waldemarson, S., Kurbasic, E., Krogh, M., Cifani, P., Berggård, T., Borg, Å., and James, P. (2016). Proteomic analysis of breast tumors confirms the mRNA intrinsic molecular subtypes using different classifiers: a large-scale analysis of fresh frozen tissue samples. Breast Cancer Res. 18, 69.

Wang, L., Zhao, Z., Meyer, M.B., Saha, S., Yu, M., Guo, A., Wisinski, K.B., Huang, W., Cai, W., Pike, J.W., et al. (2014). CARM1 methylates chromatin remodeling factor BAF155 to enhance tumor progression and metastasis. Cancer Cell 25, 21–36.

Wiśniewski, J.R., Ostatiewicz, P., and Mann, M. (2011). High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. J. Proteome Res. 10, 3040–3049.

Wojnar, A., Kobierzynski, C., Krollica, A., Pula, B., Podhorska-Okolow, M., and Dziegieł, P. (2010). Correlation of Ki-67 and MCM-2 proliferative marker expression with grade of histological malignancy (G) in ductal breast cancers. Folia Histochem. Cytobiol. 49, 442–446.
# STAR METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| **Antibodies**      |        |            |
| anti-HER2 antibody   | Leica Biosystems Novocastra | Cat# NCL-L-CBE-356; RRID:AB_563769 |
| anti-HER2 antibody   | Leica Biosystems Novocastra | Cat# NCL-c-erbB-2-316; RRID:AB_563764 |
| antibody diluent     | Agilent/DakoCytomation | S0809 |
| biotinylated anti-mouse secondary antibody | Vector-Elite, Vector Laboratories | BA-9200; RRID:AB_2336171 |
| biotinylated anti-rabbit secondary antibody | Vector-Elite, Vector Laboratories | BA-1000; RRID:AB_2313606 |
| peroxidase ABC reagent | Vector-Elite, Vector Laboratories | PK-6100; RRID:AB_2336819 |
| **Chemicals, Peptides, and Recombinant Proteins** |        |            |
| trypsin             | SCIEX  | Cat# 4352157 |
| guanidine hydrochloride | Sigma-Aldrich | Cat# G4505 |
| Triton X-100        | Sigma-Aldrich | Cat# T9284 |
| tris(2-carboxyethyl)phosphine | Sigma-Aldrich | Cat# C4706 |
| S-methyl methanethiosulfonate | Sigma-Aldrich | Cat# 64306 |
| urea                | Sigma-Aldrich | Cat# U6504 |
| ammonium formate    | Sigma-Aldrich | Cat# 70221 |
| **Critical Commercial Assays** |        |            |
| iRT-kit             | Biognoys | Cat# Ki-3002 |
| RC-DC Protein Assay | Bio-Rad | Cat# 5000122 |
| RNA 6000 Nano Kit   | Agilent | Cat# 5067-1511 |
| **Deposited Data**  |        |            |
| SWATH assay library | this paper | http://www.SWATHAtlas.org Human breast cancer library 2017 |
| mass spectrometry SWATH data | this paper | https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/Pass_View?identifier=PASS00864 |
| mass spectrometry spectral library data | this paper | https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/Pass_View?identifier=PASS00857 |
| microarray mRNA breast cancer data | (Haibe-Kains et al., 2012) | DFHCC, DFHCC2, IRB, PNC and SUPERTAM_HGU133PLUS_2 |
| RNA-Seq TCGA breast cancer data | (Cancer Genome Atlas, 2012; Ciriello et al., 2015) | https://portal.gdc.cancer.gov/projects/TCGA-BRCA |
| survival analysis data | (Györffy et al., 2010) | http://kmplot.com/analysis/index.php?p=service&cancer=breast |
| **Oligonucleotides** |        |            |
| p53 sequencing sense primer | EastPort | 5’ TCCCTCCCATGTGCTCAAGACTG 3’ |
| p53 sequencing antisense primer | EastPort | 5’ GGAGCCCGGGACAAACGAATGG 3’ |
| **Software and Algorithms** |        |            |
| MSstats (version 2.1.3) | (Choi et al., 2014) | http://bioconductor.org/packages/release/bioc/html/MSstats.html |
| GSEA | Broad Institute, Inc. | http://software.broadinstitute.org/gsea/login.jsp |
| ctree() | (Hothorn et al., 2006) | https://cran.r-project.org/web/packages/partly/index.html |
| limma | (Ritchie et al., 2015) | http://bioconductor.org/packages/release/bioc/html/limma.html |
| edgeR | (Robinson et al., 2010; McCarthy et al., 2012) | https://bioconductor.org/packages/release/bioc/html/edgeR.html |
| OpenMS tool | | https://www.openms.de/ |

(Continued on next page)
LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Pavel Bouchal (e-mail: bouchal@chemi.muni.cz).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Study design
The objective of the study was to compare classification of breast cancer tissues based on proteotypes obtained using a novel next generation proteomics approach, SWATH-MS, with clinically used subtypes classified by immunohistological markers and grade. To avoid lymph node status as confounding factor in tumor classification into subtypes, we decided for the same representation of lymph node positive and lymph node negative tumors in the sample set. A secondary aim was to compare SWATH-MS data with previous measurement using discovery proteomics via data dependent analysis (DDA) in the same sample set. To this end we designed a retrospective pilot discovery study on a cohort of well characterized breast tumor samples available from Masaryk Memorial Cancer Institute (MMCI) (Bouchal et al., 2015). The samples were analyzed by SWATH-MS and the findings confirmed both by immunochemical validation and by meta-analysis of corresponding mRNA levels in independent publicly available sets of patients.

Clinical tissue samples
Informed patient consent forms and the use of collected tissues for targeted proteomics analysis were approved by the Ethics committee of the Masaryk Memorial Cancer Institute (MMCI). Breast cancer tissue samples were frozen in liquid nitrogen within 20 min of surgical removal and stored at −180°C in the tissue bank at MMCI. A set of 96 preoperatively untreated women’s breast carcinomas of 11-20 mm maximum diameter (pT1c) was selected. The set consisted of 48 ER+, PR+, HER2-, grade 1 tumors (luminal A (LA) subtype); 16 ER+, PR−/+, HER2-, grade 3 tumors (luminal B (LB) subtype); 8 ER+, PR−/-, HER2+, grade 3 tumors (luminal B-like HER2 positive (LBH) subtype); and 16 ER−, PR−, HER2-, grade 3 tumors (triple negative (TN) subtype). Half of the tumors in each group was lymph node positive and half was lymph node negative at the time of diagnosis. Full details are available in Data S1A and S1B. The cases were reviewed by involved pathologist (Rudolf Nenutil) before entering the study. The tumors were classified and reviewed using FFPE blocks, taken in parallel with the native deeply frozen samples used for proteomics. The samples with very low cellularity of invasive tumor component (e.g., below 20%), and/or dominant in situ component and/or apparent clonal morphological heterogeneity were not used. As the dataset attempted to represent the main phenotypes, the cases were of variable malignancy and different cellularity. On average, the low grade tumors are inherently of lower cellularity compared to high grade ones. Based on the results, additional independent set of 78 grade 2 and 3 breast tumors was used for IHC validation of ERBB2 protein levels in HER2+, ER+ (n = 41) versus HER2+, ER− (n = 37) tumors (Data S1C). Sample sets used for meta-analysis of mRNA levels are described in Statistical Analysis section.

METHOD DETAILS

Tissue quality control via RNA integrity measurement
After homogenization in a MM301 mechanical homogenizer (Retsch, Germany) using a metal ball for 2 × 2 min at 25 s⁻¹ in 600 μL of RLT buffer (QIAGEN, Germany) with 1% β-mercaptoethanol, total RNA was isolated using RNeasy Mini Kit (QIAGEN, Germany) following the manufacturer’s protocol. RNA was eluted with 30 μL of RNase-free water, quantified at 260 nm using NanoDrop ND-1000 (Thermo Fisher Scientific, USA) and quality checked by measurement of RNA integrity number (RIN) on Agilent 2100 Bioanalyzer using RNA 6000 Nano Kit (both Agilent Technologies, USA). Samples which did not pass the criterion of RNA quality...
Proteomics sample preparation
Frozen breast cancer tissue (approx. 20 mm³) was homogenized in 150 μL lysis buffer (6 M guanidine hydrochloride; 0.1 M Na-phosphate buffer, pH 6.6; 1% Triton X-100) in a MM301 mechanic homogenizer (Retsch, Germany) using a metal ball for 2 × 2 min at 20 s⁻¹, needle-sonicated (Bandelin 2200 Ultrasonic homogenizer, Bandelin, Germany; 30 × 0.1 s pulses at 50 W) and kept on ice for 1 h. After 14,000 × g centrifugation at 4 °C for 20 min, protein concentration was measured in the supernatant using RC-DC assay (Bio-Rad, USA). An aliquot of the lysate containing 60 μg total protein mass was digested using a filter aided sample preparation protocol (Wisniewski et al., 2011) with modifications. Briefly, aliquots of the lysate were digested with 200 μL 8 M urea in 0.5 M triethylammonium bicarbonate (TEAB) pH 8.5 on Vivacon 500 filter device, cut-off 10K (Sartorius Stedim Biotech, Germany). The device was centrifuged at 14,000 × g at 20 °C for 20 min (all of the following centrifugation steps were performed applying the same conditions). Subsequently, 100 μL 5 mM tris(2-carboxyethyl)phosphine in 8 M urea, 0.5 M TEAB, pH 8.5 was added to the filter, proteins were reduced at 37 °C for 60 min at 600 rpm and centrifuged. Next, 100 μL 10 mM S-methyl methanethiosulfonate in 8 M urea and 0.5 M TEAB, pH 8.5 were added to the filter, cysteine groups of peptides were alkylated at 20 °C for 10 min and centrifuged. The resulting concentrate was diluted with 100 μL 8 M urea in 0.5 M TEAB, pH 8.5 and concentrated again. This step was repeated twice. The concentrate was subjected to proteolytic digestion by adding 100 μL 0.5 M TEAB, pH 8.5 containing trypsin (TPCK treated, SCIEX, USA) reconstituted in water (trypsin to protein weight ratio 1:30) and by incubating at 37 °C for 16 h. The digests were collected by centrifugation into clean tubes, dried in a vacuum concentrator and C18 desalted as previously described (Bouchal et al., 2009) using 0.1% trifluoroacetic acid as an ion pairing reagent. Eleven retention time anchor peptides (commercial iRT peptide solution, Biognosys, Switzerland) (Escher et al., 2012) were added into each sample at a ratio of 1:40 v/v. For SWATH-MS analysis, equal amounts of samples (estimated to be 1.33 μg protein) were injected in single technical replicates.

LC-MS analyses for spectral library generation
As an input for generating the SWATH-MS assay library, the following samples were prepared: (i) 10 pooled samples (each pooled from 4-8 patients) of 5 the breast cancer subtypes mentioned above. Each subtype group involved two pools of tumors (lymph node positive and lymph node negative cases separately); (ii) pool of aliquots of all samples in the sample set (400 μg in total) fractionated using HILIC chromatography as follows: HILIC Kinetex column (Phenomenex, USA, 2.6 μm, 150 × 2.1 mm, 100 A) was run in an Agilent Infinity 1260 LC system (Agilent, USA). Mobile phase (A) was composed of 100% acetonitrile (Merck, Germany), mobile phase (B) of water (Milliq, Millipore) and mobile phase (C) of 50 mM ammonium formate (pH 3.2). 20 μL mobile phase (B) were added to the sample which was then sonicated in an ultrasonic bath for 2 min. Then, 20 μL mobile phase (A) and 5 μL mobile phase (C) were added. After a further 2 min of sonication, the sample was centrifuged at 16,000 × g at 20 °C for 20 min. The sample injection volume was 40 μL and the separation method was set as follows: 5 min isocratic 0% B, 7 min gradient to 20% B, 23 min gradient to 34% B, 5 min gradient to 50% B, 5 min isocratic 50% B, 0.5 min gradient to 0% B and for 4.5 min isocratic 0% B; mobile phase C was kept at 10% all the time. The flow rate was 0.2 mL/min, column temperature was set to 30 °C and the UV signal was monitored at 280 nm. Fractions were collected every 1 min, some neighboring fractions with lower signal intensity were subsequently pooled to generate a final set of 20 fractions with similar peptide content. These were vacuum-dried and stored at −80 °C.

MS/MS datasets for spectral library generation were acquired on a TripleTOF 5600+ mass spectrometer (SCIEX, Canada) interfaced to an Eksigent Expert nanoLC 400 system (SCIEX, Canada). Prior to separation, the peptides were concentrated on a C18 PepMap100 pre-column (Thermo Fisher Scientific, USA; particle size 5 μm, 100 Å, 300 μm x 5 mm). After 10 min washing with a solvent consisting of 2% acetonitrile and 0.05% (v/v) trifluoroacetic acid, the peptides were eluted from a capillary column (75 μm × 250 mm, X-Bridge BEH C18 130 Å, particle size 2.5 μm, Waters, USA, prepared as described in (Planeta et al., 2003)) using 2% mobile phase B for 10 min (mobile phase A was composed of 0.1% (v/v) formic acid in water, mobile phase B of 0.1% (v/v) formic acid in acetonitrile) followed by gradient elution from 2% to 40% mobile phase B in the next 120 min at a flow rate of 300 nL/min. Output of the separation column was directly coupled to nano-electrospray source. MS1 spectra were collected in the range of 400-1250 m/z for 250 ms. The 20 most intense precursors with charge states of 2 to 5 that exceeded 50 counts per second were selected for fragmentation, rolling collision energy was used for fragmentation and MS2 spectra were collected in the range of 200-1600 m/z for 100 ms. The precursor ions were dynamically excluded from reselection for 12 s.

LC-MS analyses in SWATH-MS mode
SWATH-MS datasets of the individual patients were acquired on a TripleTOF 5600+ mass spectrometer (SCIEX, Canada); the same chromatographic system, settings, and gradient conditions as described above for spectral library generation were used. Using an isolation width of 9.7 m/z (containing 1 m/z for the window overlap), a set of 69 overlapping SWATH mass windows was constructed covering the precursor mass range of 400-1000 m/z. The effective isolation windows can be considered as 400.5-408.2 (first nar-rower window), 408.2-416.9, 416.9-425.6 etc. SWATH MS2 spectra were collected from 360 to 1460 m/z. The collision energy was optimized for each window according to the calculation for a charge 2+ ion centered upon the window with a spread of 15 eV. An accumulation time (dwell time) of 50 ms was used for all fragment ion scans in high-sensitivity mode, and for each SWATH cycle a survey scan was also acquired for 50 ms, resulting in a duty cycle of 3.5 s and a typical LC peak width of ~30 s.
Compared to the above conditions, for the analysis of pooled samples (see previous paragraph for pooling scheme) the parameters were changes as follows: (i) chromatographic separation of peptides was performed on 20-cm emitter (75 μm inner diameter, #PF360-75-10-N-5, New Objective, USA) packed in-house with C18 resin (Magic C18 AQ 3 μm diameter, 200 Å pore size, Michrom BioResource, USA); (ii) a linear gradient from 2%–30% solvent B (98% ACN/0.1% FA) was run over 120 min at a flow rate of 300 nl/min; (iii) because of the increased sample complexity due to the pooling strategy, a set of 64 SWATH windows (containing 1 m/z for the window overlap) with variable width optimized for human samples was used to cover the precursor mass range of 400-1200 m/z (Collins et al., 2017).

TP53 sequencing
Total cellular RNA was extracted using TRI Reagent (MRC). TP53 mRNA from tumor tissue was amplified using the SuperScript™ III One Step RT-PCR System with Platinum® Taq High Fidelity (Invitrogen, USA), sense primer: 5′ TCCCCTCCCCCATGTGCTCAAGACTG 3′ and antisense primer: 5′ GGAGCCCGGGGACAAAGCAAATGG 3′. PCR products were purified by MinElute™ PCR Purification Kit (Qiagen, Germany) and sequenced using the ABI PRISM BigDye® Terminator v 3.1 Cycle Sequencing Kit on an ABI 3130 genetic analyzer (Applied Biosystems, USA).

ERBB2 immunohistochemistry
After removal of paraffin wax with xylene and rehydration, endogenous peroxidase activity was blocked with 3% hydrogen peroxide in phosphate buffered saline (PBS) pH 7.5, for 15 min. No antigen retrieval was performed. After three PBS washes, nonspecific binding was blocked with 5% non-fat dried milk in PBS for 15 min. The cocktail of anti-HER-2 primary antibodies was diluted in antibody diluent (DakoCytomation) to 1:500 for Novocastra NCL-c-erbB-2-316, and 1:1,000 for Novocastra NCL-L-CBE-356 (both Leica Biosystems, Germany) and applied overnight at 4°C. Reactive sites were identified with biotinylated anti-mouse and anti-rabbit secondary antibodies and peroxidase ABC reagents (Vector-Elite, Vector Laboratories, USA) according to the manufacturer’s instructions and peroxidase activity was visualized with DAB+ reagents (DakoCytomation). Sections were washed in distilled water and counterstained with Gill’s hematoxylin, dehydrated, cleared, and mounted. Membrane staining of tumor cells was evaluated as 0, 1+, 2+, 3+ according to the HercepTest™ Interpretation Manual (DakoCytomation).

Validation of SWATH-MS quantitation through selected reaction monitoring
Aliquots of the same tryptic digests were used for (i) SWATH-MS (as described above) and for (ii) selected reaction monitoring (SRM) with mTRAQ labeling (as described previously (Procházková et al., 2017)). Shortly, aliquots for SRM were labeled by mTRAQ prior to the analysis. For mTRAQ labeling, two aliquots from each sample corresponding to 10 μg of digested protein were processed. One sample group was labeled with mTRAQ-Δ0 and the other sample group was labeled with mTRAQ-Δ8 label. Samples labeled with mTRAQ-Δ8 were pooled together to create global internal standard. This pool was then divided into 96 aliquots that were added to each mTRAQ-Δ0 labeled sample and measured in SRM as described previously (Procházková et al., 2017).

QUANTIFICATION AND STATISTICAL ANALYSIS
SWATH-MS assay library generation
Raw data files (wiff) were centroided and converted into mzML format using the SCIEX converter (beta release 111102) and subsequently converted into mzXML using openMS (version 1.9.0, Feb 10 2012, Revision 9534). The converted data files were searched using the search engines XTandem (k-score, version 2011.12.01.1) and Comet (version 2013.02, revision 2) against all human proteins annotated in UniProt/SwissProt (2014_04) and the sequences of 11 iRT peptides (iRT-kit, Biognosys). The searched database also contained a decoy protein sequence (reversed protein sequence) for each database protein. Only fully tryptic peptides with up to two missed cleavages were allowed for the database search. The tolerated mass errors were 15 ppm on MS1 level and 0.1 Da on MS2 level. Methylthiolation of cysteines was defined as a fixed modification and methionine oxidation as a variable modification. The search results were processed with PeptideProphet (Keller et al., 2002) and iProphet (Shteynberg et al., 2011) as part of the TPP 4.6.0 (Deutsch et al., 2010). The SWATH-MS assay library was constructed from the iProphet results with an iProphet cut-off of 0.8360 which corresponds to 1% FDR on peptide level. The raw and consensus spectral libraries were built with SpectraST (version 4.0) (Lam et al., 2007; Lam et al., 2008) using the -cICID_QTOF option for high resolution and high mass accuracy. Retention times were normalized and converted to iRT space using spectratrust2spectrat_iRT.py (imsproteomicstools R356). The 6 most intense y and b fragment ion of charge state 1, 2 and 3 were extracted from the consensus spectral library using spectratrust2tsv.py (imsproteomicstools). Neutral losses –17 (NH3), –18 (H2O) and –64 (CH3SO2) typical for oxidized methionine were also included if they were among the 6 most intense fragment ions. Fragment ions falling into the SWATH window of the precursor were excluded as the resulting signals are often highly interfered. The library was converted into TraML format using the OpenMS tool ConvertTSToTraML (version 1.10.0). Decoy transition groups were generated based on shuffled sequences (decays similar to targets were excluded) by the OpenMS tool OpenSwathDecoyGenerator (version 1.10.0) and appended to the final SWATH library in TraML format.
SWATH-MS data processing in OpenSWATH
The SWATH-MS data was analyzed using OpenSWATH (Röst et al., 2014) with the following parameters: Chromatograms were extracted with 0.05 Th around the expected mass of the fragment ions and with an extraction window of ± 5 min around the expected retention time (see Data S3C for justification). The best models to separate true from false positives (per run) were determined by pyProphet with 10 cross-validations. The runs were subsequently aligned with a target FDR of 0.01 for aligned features (Röst et al., 2016). Background signals were extracted for features that could not be confidently identified (Röst et al., 2016). To reduce the size of the output data and remove low-quality features, two filtering steps were introduced: (i) keep only the 10 most intense peptide features per protein and (ii) of these, keep only features that were identified with an FDR < 0.01 in at least four samples over all runs, corresponding to the smallest tumor group in the dataset defined by a combination of subtype and lymph node status.

Statistical analysis
All statistical tests were two-tailed and the results were considered statistically significant at alpha = 0.05 or FDR = 0.05, if not stated otherwise. Definition of error bars in all figures: Boxes are extended from the 25th to the 75th percentile, with a line at the median. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range (IQR) from the box. The individual points represent outliers or extreme values.

Relative quantification with MSstats and differential protein expression analysis between subtypes and related clinical-pathological variables
We used the R (version 3.0.3) package MSstats 2.1.3 (Choi et al., 2014) for relative quantification of protein levels among the five different breast cancer conventional subtypes and related clinical-pathological variables (ER, grade, HER2, lymph node status). Before MSstats and correlation analysis, the OpenSWATH output was further reduced to contain up to five peptide features per protein and the intensities were log2 transformed and median-equalized. The differences in protein expression between conventional subtypes and related clinical-pathological variables were compared pairwise using mixed effect models as implemented in the groupComparison function of MSstats, with expanded scope of biological and restricted scope of technical replication. Resulting p values were corrected for multiple hypotheses testing by the Benjamini-Hochberg method.

KEGG pathway analysis
The list of 4,443 proteins in the SWATH-MS library of assays (Data S2A) and the list of SWATH-MS 2,842 quantified proteins (Data S3) were inserted in Kyoto Encyclopedia of Genes and Genomes (KEGG) Mapper (https://www.kegg.jp/kegg/tool/map_pathway2.html), searched against hsa (Homo sapiens) database the subset of proteins related to Pathways in cancer (hsa05200) was displayed.

Gene set enrichment analysis
Gene set enrichment analysis (GSEA) in GSEA Java desktop application (http://software.broadinstitute.org/gsea/login.jsp) was conducted using the pre-ranked list (according to protein fold changes between ER+’/ER-, tumor grade 3/grade 1, HER2+/HER2-, lymph node positive/negative patient groups) of 2,842 proteins quantified by SWATH-MS to find pathways enriched in ER+, high grade, HER2+, and lymph node positive phenotypes separately, with a priori defined pathways from BioCarta (https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways). We used default settings, except that we decreased the minimal size of a gene set to 1 and we did not use any normalization method to normalize the enrichment scores across analyzed gene sets.

Correlation analysis of breast cancer tissue proteomes
For the correlation analysis of the pooled samples, label-free quantification was conducted using the R package aLFQ (1.3.2) (Roenneberger et al., 2014). The method ProteinInference with default parameters (summing the three most intense transitions per peptide and averaging the two most intense peptides per protein) but without consensus feature selection was used to compute a protein intensity for all 1,832 proteins for which at least one peptide has been quantified by OpenSWATH (only including proteotypic peptides). Hierarchical clustering with Spearman’s correlation-based distance matrix and average linkage algorithm was performed in Perseus 1.5.1.6 software (https://www.maxquant.org) on log2 transformed, Z-score normalized (on both samples and proteins according to median) protein abundance values, including only proteins quantified in all pools. For correlation analysis of individual samples, we selected all 2,842 proteins for which proteotypic peptides were quantified by OpenSWATH and performed Spearman’s correlation among samples based on log2 protein intensities.

Construction of the decision tree
We used a conditional reference tree algorithm for automated selection of the most discriminative variables (proteins) between conventional subtypes and for the generation of a decision tree, using ctrees() function of R package party (Hothorn et al., 2006) with the control parameters set to default, except for the minimum sum of weights in a node in order to be considered for splitting (mssplit = 50) and the minimum sum of weights in a terminal node (minbucket = 3). The analysis was based on the set of 22 proteins with significantly different abundances between different conventional subtypes and related clinical-pathological characteristics, as determined by MSstats and presented in the Results part. The selected proteins were further validated in gene expression datasets and through immunohistochemistry. The decision tree was constructed also on gene expression data (see section g below).
Analysis of ERBB2 gene expression in the same sample set

The data were extracted from our previously published dataset (Bouchal et al., 2015).

Analysis of gene expression in independent microarray and RNA-Seq sets of samples

Publicly available gene expression datasets DFHCC, DFHCC2, IRB, PNC and SUPERTAM_HGU133PLUS_2 (all platform Affymetrix Human Genome U133A, 937 samples in total) were downloaded in the log2 normalized form (Haibe-Kains et al., 2012) and used in order to confirm at the transcriptome level the hypotheses derived from our analysis of proteomic SWATH-MS data. For this purpose, subsets of 883 samples with available information on gene expression, ER status, tumor grade, HER2 status or lymph node status were used, based on the type of comparison (see Figure S2).

First, we performed analysis of differential expression between conventional subtypes (pairwise) and between the categories of clinical-pathological variables using moderated t-statistics (method implemented in the R package Limma of R 3.0.2), on the set of 6,895 probesets representing 2,782 genes with corresponding products (proteins) measured also by SWATH-MS in our experiment. This means that out of 2,842 proteins measured by SWATH-MS, 97.9% had corresponding genes in the gene expression datasets. P values were adjusted for multiple hypothesis testing by Benjamini-Hochberg FDR correction; see Data S4 for details. In this analysis we also validated INPP4B, CDK1 and ERBB2 (the proteins selected in the proteotype classification tree) as differentially regulated between ER+ versus ER- tumors, high versus low grade tumors, and HER2+ versus HER2- tumors, respectively.

Second, we correlated log2 protein fold-changes (log2FC) as obtained from pairwise group comparison with the respective transcript log2FCs from the same comparisons in the five transcriptomic datasets. The same analysis was performed also for a subset of ERBB2, INPP4B and CDK1 protein-transcript pairs.

Third, a decision tree based on gene transcript levels was constructed in order to classify the samples into the five conventional subtypes and thus to compare the resulting model in terms of performance to the model (decision tree) based on proteotypes. For this purpose, the same procedure as described above (“Construction of the Decision Tree” section) was applied on 1036 most variable (top 5%) probesets representing unique gene symbols and a set of 474 samples.

Fourth, preprocessed Level 3 RNA-seq data (Cancer Genome Atlas, 2012; Cirriello et al., 2015) were downloaded from the TCGA data portal (https://portal.gdc.cancer.gov). Filtering and normalization was performed using edgeR package (Robinson et al., 2010; McCarthy et al., 2012), Limma (Ritchie et al., 2015) “RemoveBatchEffect” function was executed on log2 transformed Count Per Million (CPM) data. Batch corrected log2CPM values were then used in order to validate the hypotheses derived from our analysis of proteomic SWATH-MS data also on transcript level. Subsets of 1078 samples with available information on gene expression, ER status, or HER2 status were used to perform analysis of differential expression between the categories of clinical-pathological variable: 791 ER+ versus 237 ER- patients were compared in term of INPP4B expression, and 161 HER2+ versus 554 HER2- patients were compared in term of ERBB2 expression using Wilcoxon rank sum test. As the information on tumor grade was unavailable for the dataset, we performed Spearman correlation of CDK1 expression with expression data on commonly used proliferation marker MKI67.

Analysis of patient survival

Survival analysis was performed using Kaplan-Meier Plotter (http://kmplot.com) for relapse-free survival (RFS) involving a microarray dataset from 3951 breast cancer tissues (2018 database version) (Gyorffy et al., 2010). Each gene was represented by user-defined probe set, Affymetrix IDs were as follows: 205376_at (INPP4B), 203213_at (CDK1), referenced as CDC2 in kmplot database, 216836_s_at (ERBB2), for reference genes 205225_at (ESR1) and 212023_s_at (MKI67). The population was split into high and low expression groups based on the incidence: (i) upper tertile for INPP4B and ESR1 based on approximate proportion of ER+ and ER- tumors, (ii) median for CDK1 and MKI67 genes based on approximate proportion of high and low grade tumors, and (iii) lower quartile for ERBB2 based on approximate proportion of HER2+ and HER2- tumors in the breast cancer population. 120 months follow up threshold was applied.

Statistical analysis of the IHC data

Associations between ERBB2 staining intensity and ER status were assessed by Fisher’s exact test in R 3.0.2.

Correlation analysis of SWATH-MS and SRM quantitation

The SRM data were extracted from Supplementary file 3A of the previous publication (Procházková et al., 2017) and the data from three technical replicates (injections) were averaged. To compare SWATH-MS and SRM quantitation, comparison based on intensity ratios was applied as follows: Intensity ratios in SWATH-MS dataset were calculated as a ratio between peptide intensity of endogenous peptide in the sample (I_D) and average intensity of the particular peptide across the dataset (I_mean). Intensity ratios in SRM dataset were calculated as a ratio between peptide intensity of endogenous peptide in the sample (I_D) and intensity of global internal standard (I_int). Data analysis and correlation was performed in stringr (version 1.3.0.) and ggplot2 (version 3.0.0.) packages from cran.r-project.org, in R version 3.4.4.
DATA AND CODE AVAILABILITY

All MS/MS data files in wiff and mzXML format as well as intermediary and final files of the library building workflow are available at https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/PASS_View?identifier=PASS00857, the accession number in PeptideAtlas is PASS00857. Breast cancer SWATH assay library has been made available through the SWATHAtlas database (http://www.SWATHAtlas.org). All SWATH-MS raw data and OpenSWATH output files are available at https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/PASS_View?identifier=PASS00864, the accession number in PeptideAtlas is PASS00864. OpenSWATH related software is available on http://www.openswath.org/en/latest/.
Supplemental Information

Breast Cancer Classification Based on Proteotypes Obtained by SWATH Mass Spectrometry

Pavel Bouchal, Olga T. Schubert, Jakub Faktor, Lenka Capkova, Hana Imrichova, Karolina Zoufalova, Vendula Paralova, Roman Hrstka, Yansheng Liu, Holger Alexander Ebhardt, Eva Budinska, Rudolf Nenutil, and Ruedi Aebersold
Figure S1. Overlap of cancer-related proteins identified by SWATH-MS and iTRAQ. Related to Figures 1 and 2. The overlap in coverage of cancer-related proteins between SWATH-MS and iTRAQ is shown on KEGG’s “Pathways in Cancer” map. The same set of 96 breast cancer tissues was analysed. Yellow boxes correspond to proteins quantified by iTRAQ only (Bouchal et al., 2015), red boxes correspond to proteins quantified by SWATH-MS only, and green boxes correspond to proteins present in both datasets. The FDR cut-off for both methods is 1%. The total number of proteins in the SWATH-MS dataset is 2,842 and the total number of proteins in the iTRAQ dataset is 3,007. Detailed information for the pathway map is available on the KEGG website at www.genome.jp/kegg/.
Figure S2. Overview of samples from independent transcriptomics datasets DFHCC, DFHCC2, IRB, PNC and SUPERTAM_HGUPLUS_2 and how they were used for comparisons with the proteomic data and to build a decision tree. Related to Figures 4-6. Data set source: (Haibe-Kains et al., 2012).
Figure S3. Decision tree based on gene expression data. Related to Figures 3-4. (A). Five selected breast cancer gene expression datasets (DFHCC, DFHCC2, IRB, PNC, SUPERTAM_HGU133PLUS2) (Haibe-Kains et al., 2012) were used to construct the decision tree using the same recursive partitioning algorithm as for the proteomic data. See STAR Methods for further details on datasets and analysis. (B) Data table for the generation of the decision tree shown in A. Breast cancer subtypes: LA, luminal A; LB, luminal B; LBH, luminal B HER2+, HER, HER2 enriched; TN, triple negative.
Figure S4. Independent validation of INPP4B, CDK1 and ERBB2 association with ER status, tumour grade, and HER2 status (full version). Related to Figure 4. Five independent transcriptomics datasets of 937 patients (DFHCC (n=115), DFHCC2 (n=84), IRB (n=129), PNC (n=92) and SUPERTAM_HGU133PLUS_2 (n=517)) (Haibe-Kains et al., 2012), see STAR Methods and Data file S4D for details) were analysed for gene expression of INPP4B, CDK1, and ERBB2. For each of the three genes, transcript levels were significantly different (p<0.05).
depending on ER status (for INPP4B), tumour grade (for CDK1), or HER2 status (for ERBB2). All available Affymetrix probes per gene are shown here; data for the most variable Affymetrix probe are shown in Figure 4.
Figure S5. Relapse-free survival (RFS) in breast cancer patients with high vs. low expression of the three key genes INPP4B, CDK1, ERBB2. Related to Figures 3-4. A microarray dataset of 3955 breast cancer patients available through the KMplot tool (www.kmplot.com, 2018 database version) was used to generate Kaplan-Meier plots. Selected cut-offs for high vs. low expression correspond to representation in breast cancer patient population. We found that INPP4B expression was significantly connected with RFS in the same manner as the commonly used reference gene ESR1 (for ER status), and that CDK1 expression was significantly connected with RFS in the same manner as the commonly used reference gene MKI67 (for tumour proliferation/grade).
Figure S6. Comparison of the number of proteins consistently quantified across samples by SWATH-MS and SuperSILAC. Related to Figures 1 and 3. Plots show the number of proteins consistently quantified across samples as a function of samples analysed. Shown on the left is the SWATH-MS dataset of 96 patients and on the right the SuperSILAC dataset of 40 patients (Tyanova et al., 2016). Note that the comparison is based only on light peptide variants originating from tissue; since the labeled (“heavy”) internal standard peptides used in the SuperSILAC study were prepared from cell lines and spiked into the tissue sample, they are not relevant for the number of tissue proteins quantified. „Requantification“ refers to a software tool that is run as part of the SWATH-MS data analysis pipeline for peptides not detectable above an FDR of 0.01. In these cases, the tool infers the peak boundaries from the closest neighboring run and quantifies the background fragment ion signal within those boundaries. These imputed values serve as upper bounds of the analyte signal for the run in question (Rost et al., 2016).
| Gene symbol | UniProt Acc. | Abundance rank |
|-------------|--------------|----------------|
| HER2        | P04626       | 997            |
|             |              | 197            |
| GRB7        | Q14451       | 959            |
|             |              | 1,080          |
| MAPK3       | P27361       | 1,302          |
|             |              | 1,775          |
| EEF1G       | P26641       | 168            |
|             |              | 156            |
| AGR2        | O95994       | 291            |
|             |              | 212            |
| NDUFAB1     | O14561       | 2,149          |
|             |              | 2,338          |
| MIPEP       | Q99797       | -              |
|             |              | 2,820          |
| MLPH        | Q9BV36       | 2,396          |
|             |              | 2,322          |
| MCM5        | P33992       | 1,481          |
|             |              | 1,486          |
| HID1        | Q8IV36       | 1,275          |
|             |              | 2,567          |
| STMN1       | P16949       | 460            |
|             |              | 1,123          |
| CMBL        | Q96DG6       | 1,456          |
|             |              | 1,124          |
| GLS         | O94925       | 2,536          |
|             |              | 2,031          |
| FOXA1       | B7ZAP5       | -              |
|             |              | 3,574          |
| C9orf114    | Q5T280       | -              |
|             |              | 4,924          |
| RCL1        | Q9Y2P8       | -              |
|             |              | 2,871          |
| ECM1        | Q16610       | 1,323          |
|             |              | 2,552          |
| CAPN7       | Q9Y6W3       | -              |
|             |              | 3,998          |
| ENO1        | P06733       | 26             |
|             |              | 28             |

Tab S1: Coverage and abundance ranks of 19-protein signature identified by SuperSILAC compared to the abundance rank of the same proteins in our SWATH-MS dataset. Related to Figure 1. Abundance ranks of 19 signature proteins identified by Tyanova and colleagues (Tyanova et al., 2016) presented in their Fig. 6A were compared to their abundance ranks in our study. Of the 19 key proteins, 14 were quantified in our study with a similar abundance rank.