Published in final edited form as:

Cancer Res. 2018 October 15; 78(20): 6001–6010. doi:10.1158/0008-5472.CAN-18-1079.

Clinical Proteomics of Breast Cancer Reveals a Novel Layer of Breast Cancer Classification

Gali Yanovich¹, Hadar Agmon¹, Michal Harel¹, Amir Sonnenblick^{2,3}, Tamar Peretz⁴, and Tamar Geiger¹

¹Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

²Oncology Division, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel

³Oncology Department, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

⁴Sharett Institute of Oncology, Hadassah-Hebrew University Medical Center, Jerusalem, Israel

Abstract

Breast cancer classification has been the focus of numerous worldwide efforts, analyzing the molecular basis of breast cancer subtypes and aiming to associate them with clinical outcome and improve the current diagnostic routine. Genomic and transcriptomic profiles of breast cancer have been well established, however the proteomic contribution to these profiles has yet to be elucidated. In this work, we utilized mass spectrometry (MS)-based proteomic analysis on more than 130 clinical breast samples to demonstrate inter-tumor heterogeneity across three breast cancer subtypes and healthy tissue. Unsupervised analysis identified four proteomic clusters, among them one that represents a novel luminal subtype characterized by increased PI3K signaling. This subtype was further validated using an independent protein-based dataset, but not in two independent transcriptome cohorts. These results demonstrate the importance of deep proteomic analysis, which may affect cancer treatment decision making.

Introduction

Breast cancer is the most prevalent cancer type in women and one of the most common death causes in women worldwide. Tumors are routinely classified based on the expression pattern of three receptors: estrogen receptor (ER), progesterone receptor (PR), and HER2 (1–3). This classification provides the basis for treatment decision making. The luminal, ER positive subtype, is the most prevalent one, with a more favorable prognosis (1). As such, endocrine treatments such as tamoxifen, efficiently target this subtype (4). HER2+ tumors are likely to respond to Trastuzumab (Herceptin), an inhibitory antibody targeting this receptor (1,4). The triple-negative (TN) subtype, which suffers from the absence of known drug targets, has less favorable prognosis (5). The molecular mechanisms underlying its

Corresponding author: Prof. Tamar Geiger, Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, 69978 Tel Aviv, Israel, geiger@tauex.tau.ac.il.

Conflict of interest: The authors declare no potential conflicts of interest.

tumorigenicity are less clear and therefore it is often treated using non-specific chemotherapy (4–6).

Beyond the routine clinical classification, genomic and transcriptomic studies defined four “intrinsic” breast cancer subtypes based on their gene expression patterns: luminal A, luminal B, basal and HER2+ (3,7). The Cancer Genome Atlas Network (2012) further associated the gene expression profiles of each subtype with DNA copy number variations and mutation profiles. For example, basal tumors presented a high frequency of TP53 mutations, whereas ER+ tumors (both luminal A and B) showed a high frequency of PIK3CA mutations. The Metabric consortium analyzed 2000 breast tumors and based on the integration of mRNA and DNA copy-number analysis divided the samples into 10 different subgroups (2). Other groups studied triple-negative tumors more deeply, further dividing them to more specific subgroups (8,9). However, classification ambiguities still exist, and while multiple studies have been focusing on molecular characterization of breast cancer profiles, the clinical routine still relies primarily on immunostaining of ER, PR and HER2 as a basis for classification (10).

In contrast to the genomic studies, protein expression levels in breast cancer and their association with classification have not yet been thoroughly studied. Previously, the TCGA performed reversed-phase protein arrays (RPPA) in order to evaluate how genetic alterations affect the protein level. However, this method is limited to hundreds of pre-selected proteins and several phosphorylation sites, and does not provide a system-wide view of the proteome. Analyzing the cancer proteome in an unbiased manner, two recent mass-spectrometry (MS)-based proteomics studies analyzed dozens of tumors and evaluated the relevance of gene-expression levels to the proteome level. In the first study we analyzed a cohort of 40 breast tumors derived from different subtypes, quantified the proteins using the super-SILAC technology (11), and reached the depth of more than 10,000 proteins (12). The second study, which was conducted by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (13), analyzed MS-based proteomics and phosphoproteomics data from a cohort containing 77 samples, used iTRAQ quantification and reached the depth of 12,553 proteins and 33,239 phosphorylation sites. They further integrated the proteomic with the genomic profiles of the same cohort from the TCGA, and associated the mutational landscape with the corresponding proteomic and phosphoproteomic profiles. Both studies showed limited concordance between the overall protein and mRNA profiles. This was evident both in the 19-protein signature identified by Tyanova et al., in which a third was found to be regulated exclusively on the proteomic level; and also in the global protein-mRNA correlation analysis of the CPTAC, which resulted in a median Pearson correlation of 0.39.

The larger scope of the CPTAC study enabled them to perform unsupervised classification of the tumors and identify novel proteomic and phosphoproteomic subgroups. Compared to the four expected mRNA-based subtypes, they identified only three proteomic groups, out of which two corresponded to the known classification (a basal cluster and a luminal cluster), while the third one was a new “stromal-enriched” cluster. This cluster was characterized by samples originating from all subtypes, and thus may be associated with stroma-rich regions within the tumors.

In the current study, in order to explore inter-tumor heterogeneity and examine proteomics-based classification, we integrated three datasets from both published and unpublished resources, all generated using the same approach. The entire cohort consisted of 212 formalin-fixed paraffin-embedded (FFPE) samples originating from different breast cancer subtypes, as well as healthy tissue. Out of these, 131 samples were selected for analysis, based on the data quality and sample annotation. Unsupervised analysis of the integrated cohort enabled us to identify a novel luminal breast cancer subtype. Comparing our findings to independent published data, we found that this classification is reproduced only on the proteomic level. This proteomic cohort enabled us to both support and challenge the known breast cancer classification, and thus to facilitate the analysis of similarities and differences between distinct omics levels.

Materials and Methods

Sample Preparation and Super-SILAC mix

The proteomic data integrated three datasets, including two published ones (12,14) and a novel dataset, Dataset-III, which was generated using the same techniques. For Dataset-III, 70 Formalin-fixed, paraffin-embedded (FFPE) samples were obtained from the Department of Oncology, Hadassah Medical Center, Jerusalem, Israel, upon approval of the IRB committee and in accordance with the Declaration of Helsinki ethical guidelines. Samples were pathologically defined based on the IHC staining of ER, PR and HER2. Samples were macro-dissected from tissue slices. Each sample was enriched to include cellular areas and exclude stromal components. We used the super-SILAC standard to serve as a reference sample for relative quantification and for normalization between the three datasets (11). The standard was prepared as previously described (11). Briefly, four breast cancer cell lines of different origin, stage and receptor status (HCC1599, MCF7, HCC1937 and HCC2218) and normal mammary epithelial cells (HMEC) were SILAC labeled with heavy lysine and arginine and their lysates were combined at equal ratios (protein amount). Super-SILAC was generated from cell lines within 6 months from their purchase from the ATCC or ECACC, or within 6 months from their authentication at the Genomics Core Facility of BioRap Technologies and the Rappaport Research Institute in the Technion, Israel. All tissues and cells were solubilized in buffer containing 4% SDS, 0.1 M dithiothreitol in 0.1 M Tris-HCl pH 7.6. FFPE samples were incubated for 1h at 95°C followed by sonication. Prior to protein digestion, proteins from each clinical sample were combined with equal protein amounts of the super-SILAC mix. The joint lysates were then trypsin-digested following the FASP protocol, on top of 30 kDa cutoff Amicon filters (15). To increase analytical depth, the resulting peptides were fractionated using pH-based strong anion exchange chromatography (SAX) in a StageTip format (15). Each sample was separated into six fractions using buffers of different pH values (11, 8, 6, 5, 4, 3) followed by desalting and concentration on C18 StageTips (16). In 25 cases, when the protein amount was insufficient, we performed in-solution trypsin digestion and/or did not fractionate the samples. Prior to MS analysis, peptides were eluted from the StageTips with 80% acetonitrile, vacuum-concentrated and diluted in MS loading buffer (2% acetonitrile, 0.1% formic acid).

Liquid Chromatography – Mass Spectrometry (LC-MS) Analysis

Peptides were separated by reverse-phase chromatography using the nano- ultra high performance liquid chromatography system (UHPLC) (Easy-nLC1000, Thermo Fisher Scientific), coupled to the Q-Exactive or Q-Exactive Plus mass spectrometers (Thermo Fisher Scientific). Peptides were separated with 220-240 min linear gradients of water/ acetonitrile. MS resolution was 70,000 with an automatic gain control (AGC) target of 3e6, and MS/MS resolution was 17,500 with AGC target of 1e5. MS analysis was performed using a top10 method in which every high resolution MS scan is followed by fragmentation of the 10 most abundant peaks by higher-energy collisional dissociation (HCD).

Protein Identification and Quantification

MS raw files of all three datasets (Dataset-I (12), Dataset-II (14), and Dataset-III) were jointly analyzed by MaxQuant (17,18) version 1.5.3.16 and the integrated Andromeda search engine (19). MS/MS spectra were referenced to the Uniprot human proteome (<http://www.uniprot.org/>) released September 2015. False discovery rate (FDR) of 0.01 was used on both the peptide and protein levels using the target-decoy approach (20). The search included methionine oxidation and protein N-terminal acetylation as variable modifications, and carbamidomethyl cysteine as a fixed modification. SILAC was employed for protein quantification, defined as two MS1 labels reflecting the SILAC heavy labels (Lysine8 and Arginine10). Maximum number of allowed missed peptide cleavages was set to 2, and the minimum peptide length was set to 7 amino-acids. An initial search followed by mass recalibration was performed with a mass tolerance of 20 ppm.

Data pre-processing and Statistics

Analyses were performed using the Perseus software (21) and R environment (where indicated). Initial dataset consisted of 212 samples from all three datasets, and these were then filtered based on their origin and data quality. After exclusion of samples as described in the Results section, the analyzed cohort consisted of 131 samples, including 109 high quality tumor samples and 22 high quality healthy samples. The excluded samples were 30 lymph node samples (healthy and metastases), one male patient, samples lacking clinical/ pathological annotation and samples with less than 3000 quantified proteins.

All data analyses were performed on the logarithmized (log2) ratios of the unlabeled samples towards the same heavy labeled super-SILAC mix. For all pairwise comparisons we filtered the data to retain only proteins with valid quantification values in at least 50% of the samples. Missing values were imputed by drawing from a normal distribution with a width of 0.3 of the overall data distribution and a down-shift of 1.5 standard deviations. The initial filtration of 50% valid quantification values per protein, ensures that all of them are seen in the heavy channel, and missing values represent lowly expressed proteins.

Batch effects originating from the dataset integration-related variability were corrected for using principal component analysis (PCA) and component subtraction. We found that the second component of a PCA separates between datasets (9% of the variation), and therefore subtracted this component before integrated downstream analysis. Student's t-tests were performed using a permutation-based FDR cutoff of 0.05, except for the tumor-healthy

analysis where we used a cutoff of 0.1. Enrichment analyses were performed using Fisher exact test with a Benjamini-Hochberg FDR (22) of 0.02. 2D Annotation enrichment test was performed in Perseus software as previously described (23), based on gene sets provided by the Molecular Signature Database (MSigDB) (24). Venn diagrams were plotted using Biovenn web tool (25).

Clustering and Network Analysis

For the consensus clustering analysis, ratios across samples and across proteins were z-score normalized followed by their clustering using the consensus cluster algorithm (26). The algorithm performs multiple iterations of hierarchical clustering, while in each iteration only a subset of the data is considered and the range of the number of output clusters is determined by the user. This results in a separate consensus matrix for each number of clusters, describing the frequency of two samples clustering together out of the total number of iterations. We used R environment (version 3.2.3) and the ConsensusClusterPlus package (27) to run this analysis (parameters used: max. Number of clusters – 10; number of iterations – 1000; subsampling fraction - 0.8; clustering algorithm – hierarchical; distance matrix - Pearson correlation).

The significantly upregulated proteins in each consensus cluster were applied to the String database (<http://string-db.org>) in order to derive the protein-protein interaction information within each group. Next, we performed a betweenness centrality analysis of each consensus cluster network, in which a node is considered central if it appears in many shortest paths between two other nodes. The top 10% central nodes from each network were used for further analysis. Centrality analysis was performed using Cytoscape and the CytoNCA plugin (28).

TCGA Reverse phase protein array (RPPA) data

The TCGA RPPA data matrix was downloaded from The Cancer Proteome Atlas (TCPA) download portal (<http://tcpaportal.org/tcpa/download.html>).

Data Availability

The entire dataset of 212 samples is available through the PRIDE repository (Dataset-I accession: PXD002619; Dataset-II accession: PXD000815; Dataset-III accession: PXD009766).

Results

Integration of three breast cancer datasets

We generated an integrated breast cancer proteomic dataset by combining three datasets from published and unpublished resources, all assembled in our laboratory, and all analyze macrodissected tumors. The first cohort (Dataset-I) was obtained from Tyanova et al. (12), and consisted of 40 breast tumor samples representative of invasive ductal carcinoma (IDC) of the three main subtypes (ER+, HER2+ and TN). The second cohort (Dataset-II), from Pozniak et al. (14), consisted of 88 samples including luminal IDC, matching adjacent healthy tissue or matched lymph node metastases. For the third dataset (Dataset-III), 70

Formalin-fixed, paraffin-embedded (FFPE)-blocks were pathologically defined based on the expression of ER, PR and HER2. Samples were analyzed by MS-based proteomics using super-SILAC as an internal standard. The three datasets were integrated, generating a cohort of 212 FFPE samples in total, originating from the different subtypes (131 ER+, 30 TN, 15 HER2 and 9 with no annotation) and from adjacent healthy tissue (27 samples).

Computational analysis of the integrated cohort using MaxQuant software resulted in the identification of 10,819 proteins in total. Next, we filtered out tumor samples with less than 3000 identifications and excluded samples of metastatic origin or of missing sample annotation. Altogether, the final cohort included 131 samples, including 22 healthy controls (for a detailed list of the samples before and after filtration see Supplementary Table S1; see Supplementary Fig. S1A and S1B for summary of the sample processing workflow).

Accurate comparison of samples was possible since all datasets were quantified relative to the same super-SILAC reference sample, in which each sample is quantified against a heavy-labelled mix of healthy and cancerous breast cell lines (11). To remove minor batch effects associated with the source of tumor samples and the sample preparation, we identified and removed the discriminating component using principal component analysis (PCA, see Methods and Supplementary Fig. S1C). The entire cohort of 131 samples was then analyzed using supervised approaches, to validate the relevance to the known breast cancer subtypes; and using unsupervised approaches to uncover novel breast cancer classification.

Supervised comparison between breast cancer subtypes

Aiming to validate whether the known subtype characteristics were evident in the proteomics data, analysis of the changes of known subtype markers (29,30) (ERBB2/HER2, MKI67, KRT18, EGFR, CD44 and FOXA1) between the healthy tissues and each subtype reproduced most of the expected subtype patterns (Fig. 1A). For example, HER2 was higher in HER2 samples compared to ER+ and TN; as expected, KRT18 and FOXA1 were significantly lower in TN tumors; however, EGFR and CD44 were only slightly higher in TN than the other subtypes. Interestingly, comparison to the healthy tissue showed that some of these cancer markers are not exclusively expressed in the transformed tissue.

A global subtype comparison resulted in hundreds of significant proteins upregulated in each subtype (Fig. 1B; Student's t-test FDR<0.1 and Supplementary Tables S2A-C). Enrichment analyses identified subtype-specific processes e.g. galactose metabolism in HER2 tumors and oxidative phosphorylation in ER+ tumors – corresponding to the well-characterized metabolic activity of breast cancer subtypes (31,32). Upregulated immune activity, demonstrated by Type I interferon and antigen presentation processes, was significantly enriched in TN tumors, as previously described (8,33) (Fig. 1C, see Supplementary Tables S2D-F for the full list). To obtain a more general functional and subtype-related view of the samples, we used lower stringency criteria for subtype specificities, (one-sided t-test, p-value<0.05) and constructed KEGG pathway proteomaps (34) (Fig. 1D). In agreement with previous findings (35–37), these maps show the subtype-specificity of signaling pathways. For example, MAPK signaling was prominent in ER+ tumors, whereas ERBB2 signaling activity was prominent, as expected, in HER2 tumors. Other than subtype-specific processes, we identified significantly enriched processes in the group of proteins that are common to all

subtypes when compared to healthy. These include actin cytoskeleton organization, extracellular vesicular exosome and type I interferon signaling (Fig. 1B and Supplementary Tables S2G and H). Taken together, these results indicate that proteomic data capture many of the well-established subtype characteristics, both globally and with regard to specific markers.

Unsupervised analysis of the cancer proteomes identifies four tumor clusters

In the second step of the analysis we aimed to examine unsupervised classification of the tumors. To eliminate masking of subtype differences, we excluded the 22 healthy controls, and classified the 109 tumor samples using the consensus clustering algorithm (26). Optimization of the number of clusters in the dataset resulted in four clusters that best separated the samples. This was evaluated quantitatively using the area under the cumulative distribution curve, which indicated that generating four clusters results in more than 15% improvement in accurate grouping of the data, compared to three clusters; but more than four clusters adds less than 10%. (Supplementary Fig. S2A and S2B). A visual inspection of the consensus matrix also provided support of a clear cluster separation (Fig. 2A). Examination of the distribution of the classical subtypes between the consensus clusters showed partial concordance with the known classification; most TN samples belonged to consensus cluster (CC) 4; HER2 samples spread across clusters; and ER+ samples (LumA or LumB) were divided into three clusters (Fig. 2B). These analyses, as well as t-SNE analysis (Fig. 2C), revealed that although CC1, CC2 and CC3 are all associated with ER+ samples, CC2 and CC3 are closer to the TN-enriched cluster, CC4; and CC1 is relatively segregated from them. Notably, running the consensus clustering algorithm with $k=3$, the ER+ CC2 was co-clustered with the ER- CC4, thus indicating their higher global proximity (Supplementary Fig. S2C). To examine whether contaminating proteins of the extracellular compartment affect this classification, we repeated the analysis and omitted all extra-cellular matrix (ECM) proteins from the dataset using Gene Ontology Cellular Component annotation. Re-analyzing with consensus clustering resulted in three clusters, reproducing CC1 and CC3, and creating a combined cluster of CC2 and CC4. This suggests that ECM proteins do not play a key role in the separation of the luminal samples (for the sample-wise comparison see Supplementary Fig. S2D).

Next, we explored the characteristics of the newly identified proteomic groups and their relation to the known subtype markers. Plotting six selected subtype markers (as mentioned above) showed the subtype-expected pattern according to the CC distribution and the known classification (Fig. 3). For example, KRT18 was significantly higher in all ER+ clusters compared to the TN-enriched CC4; and the TN markers EGFR and CD44 were significantly higher in CC4 when compared to CC3 and CC1, respectively. Examination of the LumB marker MKI67 suggested that CC1 represents this subtype. This was further supported by an enrichment analysis that identified gene sets of RB1-regulated and MYC-related targets - with which LumB subtype is known to be associated (1) - as significantly enriched within CC1 samples when compared to either CC2 or CC3 (Supplementary Fig. S3A). This analysis also showed that the majority of significantly enriched processes are higher in CC1 in both comparisons. Very few processes are higher in CC2 and CC3, and the common ones

include mostly interferon related pathways (Supplementary Fig. S3B; for the full lists see Supplementary Table S3A).

Having associated CC4 with TN subtype and CC1 with LumB subtype, CC2 and CC3 remained as two possible LumA clusters. In order to characterize them, we first analyzed each one of the two LumA clusters compared to the TN CC4. We speculated that the comparison to the distinct cluster will highlight the specific luminal features of each one of the clusters. First, to ensure that both clusters are indeed luminal, in agreement with their pathological annotation, we performed a 2D-enrichment analysis looking for gene sets from the Molecular Signatures Database (MSigDB) that behave similarly in both CC2 and CC3. We found gene sets such as YANG_BREAST_CANCER_ESR1_UP and SMID_BREAST_CANCER_BASAL_DN to be significantly enriched in CC2 and CC3 when compared to CC4, while gene sets such as SMID_BREAST_CANCER_BASAL_UP were significantly de-enriched. In addition, both CC2 and CC3 were enriched for bone metastasis gene sets (Fig. 4A and Supplementary Table S3B), recapitulating the known ER+ metastatic pattern (38,39).

In order to further characterize the differences between the LumA clusters, we examined the proteins significantly upregulated in each cluster. Student's t-test resulted in hundreds of significantly different proteins between CC2 and CC3, out of which 539 were higher in CC2 and 127 were higher in CC3 (for the list of significant proteins see Supplementary Table S3C). Interestingly, 265 proteins were commonly and significantly upregulated in both CC2 (luminal) and CC4 (TN) when compared to CC3. To identify the key proteins in this comparison, we plotted the protein-protein interaction networks of the significantly higher proteins in CC2 and CC4 (compared to CC3), and performed a betweenness centrality analysis of each network. Selecting the 10% top central proteins in each group showed 21 overlapping proteins between the CC2 and the CC4 networks. Among these proteins, we detected signaling proteins such as NFKB1, MTOR and STAT3 (Fig. 4B and C). Collectively, these results indicate that CC2 shows high expression of key TN signaling markers, but simultaneously still maintains an intrinsic-ER+ signature.

Clinical relevance of the new LumA cluster (CC2)

In order to further validate our results on an independent sample cohort and based on a distinct technological platform, we reanalyzed the reversed phase protein array (RPPA) data provided by the TCGA (1). These data included examination of 217 antibodies on 848 tumor samples, of those we selected only IDC samples (n=621), which correspond to the histological subtype included in our study. Interestingly, similar to the MS-based data, k-means clustering of the TCGA RPPA data identified two LumA clusters. Defining k-means clustering with four clusters resulted in one HER2 cluster, one basal cluster, and two clusters enriched for LumA samples, one of which was also enriched for LumB samples (Supplementary Fig. S4A). Running the same algorithm with optional five clusters gave similar results, but separated the combined LumA/LumB cluster into two different clusters (Fig. 5A). Comparison of the two LumA RPPA clusters also reproduced the differences in specific proteins (Fig. 5B and Supplementary Table S4), including significant differences in PI3K-related proteins and their phosphorylated forms. Interestingly, the cluster with

upregulated PI3K-related proteins was also characterized by increased ER signaling – demonstrated by significant upregulation of both non-phosphorylated and phosphorylated ER. Phosphorylation of ER on Ser118 is known to be mediated by AKT signaling and was shown to increase ER transcriptional activity (40,41). Additionally, ER-involved MTOR/PI3K activity is a known resistance mechanism to endocrine therapy such as tamoxifen (40,42,43). To further compare the RPPA results to the original MS data, we examined the connection between PI3K and ER on the original dataset. Since the ER protein was below signal detection level in many of the samples, we searched for ER downstream proteins to serve as a proxy for ER expression. We gathered a list of 1346 genes that are annotated as ER targets in MSigDB, out of which 411 were present in our data. In agreement with the RPPA results, also the proteomics results show significantly higher predicted ER activity in CC2 (p-value = 7.71E-11, see Supplementary Fig. S4B).

LumA subtypes distinction enabled specifically by proteomic data

Previous studies highlighted major differences in unsupervised classification based on proteomic compared to the genomic approaches (1–3,7). We therefore sought to examine whether the proteomic signature of differentially expressed proteins in CC2 and CC3 also separates the LumA samples based on their RNA profiles. In contrast to the separation based on the protein levels (Fig. 5C), hierarchical clustering of the RNA data of the same genes showed no clear differentiation between the intrinsic subtypes, even when sorted according to subtype (Fig. 5D). We applied the same analysis to an independent gene expression dataset, from the Metabric study (2012). Sorting the samples according to the 10 integrative clusters identified in the Metabric analysis and plotting the gene expression of the same signature yielded results similar to the TCGA RNA analysis, and did not manage to differentiate clearly between the clusters (Fig. 5E). This analysis indicates that the distinction between LumA sub-subtypes is potentially unique to protein-based data.

Discussion

We integrated three proteomic datasets of breast cancer samples originating from the three main clinical subtypes of IDC and from healthy tissue. Supervised analysis of the tumor samples validated many of the previous findings regarding their specific subtype-related pathways and functionalities, including immune activity in TN tumors and elevated MAPK signaling in ER+ tumors. The size of the cohort, despite being significantly smaller than current genomic studies, is still larger than the microarray studies that initially identified the breast cancer intrinsic subtypes (3,7). Additionally, the careful tumor dissection enabled focusing on the cancer cells rather than stroma, which affected past classifications. Unsupervised tumor classification using the consensus clustering algorithm separated the data into four clusters, consisting of one TN-enriched group and three ER+ enriched groups, while HER2 samples were distributed across groups.

The inability to identify a distinct group of HER2 samples in proteomics data might stem from the diverse phenotype of these tumors beyond HER2 expression. This result was previously shown by the CPTAC consortium (13) using similar techniques and a smaller cohort, and also by the Metabric study (2), which shows that HER2 is high in several

integrative clusters based on copy number aberrations and mRNA expression. Additionally, a mass cytometry-based proteomic analysis of breast cancer also suggested that HER2 expressing cells do not represent a single phenotype, and are instead characterized by variable expression of other proteins (44). The RPPA results did separate the HER2 subtype, presumably due to the pre-selection of a large number of HER2 related proteins in the arrays. In the current work, we consistently assessed HER2 and its associated proteins in an unbiased manner. They were indeed high in the HER2 clinical subtype, but showed variable expression across consensus clusters. While the importance of HER2 is not questionable given the efficacy of HER2-directed therapies, our results suggest that this pathway affects only a subset of the proteome and does not represent a distinct subtype when examining the global proteome profiles. Thus, it is useful as a therapeutic marker rather than a representative of a defined tumor subtype. Additional classification ambiguities between studies may originate from the pre-selection of features (genes/proteins) based on their standard deviation, as done by previous studies (2,13). Furthermore, major discrepancies have been identified between all 'omics'-based classifications (including ours), which are based on averaging of large tumor regions, as compared to IHC approaches. Apart from potentially overlooking the effect of tumor microenvironment components (e.g. extracellular matrix, immune infiltrates, and fibroblasts), “averaging” the cancer cells masks the intra-tumor heterogeneity that is well documented in breast cancer (45,46). This intra-tumor heterogeneity is likely to underlie some of the observed inconsistencies between classification using the clinical IHC approaches and the 'omics' approaches.

Comprehensive investigation of the three ER+ clusters identified one LumB cluster and two LumA clusters. One of the two LumA clusters was characterized with increased activity of TN-associated signaling proteins, thus suggesting a novel, sub-subtype of LumA tumors. Our classification was validated using an independent dataset obtained from TCGA, which also assisted in the characterization of this novel subtype as potentially more aggressive and therapy resistant. The mechanism underlying this resistance relies on the coordinated interaction between ER itself and growth factor signaling proteins. ER can be phosphorylated on several sites by MAPK as well as PI3K downstream kinases, such as AKT, P70S6K1 and MAPK14. Phosphorylation leads to increased transcriptional activity of ER, which in turn upregulates the transcription of genes belonging to those pathways (47,48). One of the common phosphorylation sites is Serine118, which alongside other sites (such as Serine167), has been implicated in resistance of ER+ tumors to tamoxifen (42,49). Resistance is suggested to be due to reduced affinity to the drug, accompanied by increased binding of estrogen (40,42). The signaling proteins which constitute the profile of the proteomic CC2, together with upregulation of both phosphorylated and non-phosphorylated ER, reflect this therapy-resistant phenotype. Thus, the new LumA cluster potentially represents a group of more aggressive tumors, more likely to be resistant to treatment.

The identification of a novel LumA subtype has several implications. Our findings suggest that classifying luminal breast tumors based on protein expression may set the ground for a more comprehensive clinical evaluation of breast tumors, and thus benefit treatment decision making. This work may therefore represent a starting point for a directed study of response to hormonal treatment. A focused study of this sort will include patient survival information as well as clinical data regarding patient course of treatment and response. These could

increase the understanding and interpretation of newly identified tumor profiles, and may provide insights on their biological nature and clinical implications.

The work presented here shows the added value of cancer proteomics, both independently as a robust methodology and as a complementary layer of biological information to genomics and transcriptomics. The integration of multiple samples yielded high-coverage and high-accuracy data, shedding light on the complex diversity of breast cancer, and altogether demonstrating the importance of a deep, systemic and multi-level point of view in breast cancer research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the members of the Geiger laboratory for fruitful discussions and technical assistance. We thank Dana Silverbush and Dvir Netanel for insightful bioinformatics assistance. We specifically thank Diana Nemichenitzer for assisting in sample preparation of Dataset III. This work was supported by the Israel Cancer Research Fund; G. Yanovich, M. Harel and T. Geiger were supported by the European Research Council (ERC; PROTEOMICAN 639534).

References

1. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012; 490(7418):61–70. [PubMed: 23000897]
2. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012; 486(7403):346–52. [PubMed: 22522925]
3. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001; 98(19):10869–74. [PubMed: 11553815]
4. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Ann Oncol*. 2013; 24(9):2206–23. [PubMed: 23917950]
5. Tan DS, Marchio C, Jones RL, Savage K, Smith IE, Dowsett M, et al. Triple negative breast cancer: molecular profiling and prognostic impact in adjuvant anthracycline-treated patients. *Breast Cancer Res Treat*. 2008; 111(1):27–44. [PubMed: 17922188]
6. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin Cancer Res*. 2007; 13(15 Pt 1):4429–34. [PubMed: 17671126]
7. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature*. 2000; 406(6797):747–52. [PubMed: 10963602]
8. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SA, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res*. 2015; 21(7):1688–98. [PubMed: 25208879]
9. Craig DW, O'Shaughnessy JA, Kiefer JA, Aldrich J, Sinari S, Moses TM, et al. Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. *Mol Cancer Ther*. 2013; 12(1):104–16. [PubMed: 23171949]
10. Yap TA, Gerlinger M, Futreal PA, Pusztai L, Swanton C. Intratumor heterogeneity: seeing the wood for the trees. *Sci Transl Med*. 2012; 4(127):127ps10.

11. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature Methods*. 2010; 7(5):383–5. [PubMed: 20364148]
12. Tyanova S, Albrechtsen R, Kronqvist P, Cox J, Mann M, Geiger T. Proteomic maps of breast cancer subtypes. *Nature Communications*. 2016; 7 10259.
13. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*. 2016; 534(7605):55–62. [PubMed: 27251275]
14. Pozniak Y, Balint-Lahat N, Rudolph JD, Lindskog C, Katzir R, Avivi C, et al. System-wide Clinical Proteomics of Breast Cancer Reveals Global Remodeling of Tissue Homeostasis. *Cell Systems*. 2016; 2(3):172–84. [PubMed: 27135363]
15. Wisniewski JR, Ostasiewicz P, Mann M. High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. *J Proteome Res*. 2011; 10(7):3040–9. [PubMed: 21526778]
16. Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc*. 2007; 2(8):1896–906. [PubMed: 17703201]
17. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008; 26(12):1367–72. [PubMed: 19029910]
18. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc*. 2016; 11(12):2301–19. [PubMed: 27809316]
19. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res*. 2011; 10(4):1794–805. [PubMed: 21254760]
20. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*. 2007; 4(3):207–14. [PubMed: 17327847]
21. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*. 2016; 13(9):731–40. [PubMed: 27348712]
22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. 1995; 57(1):289–300.
23. Cox J, Mann M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics*. 2012; 13(Suppl 16):S12.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005; 102(43):15545–50. [PubMed: 16199517]
25. Hulsen T, de Vlieg J, Alkema W. BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*. 2008; 9:488. [PubMed: 18925949]
26. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*. 2003; 52:91–118.
27. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010; 26(12):1572–3. [PubMed: 20427518]
28. Tang Y, Li M, Wang J, Pan Y, Wu FX. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *Biosystems*. 2015; 127:67–72. [PubMed: 25451770]
29. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009; 27(8):1160–7. [PubMed: 19204204]
30. Park SY, Lee HE, Li H, Shipitsin M, Gelman R, Polyak K. Heterogeneity for stem cell-related markers according to tumor subtype and histologic stage in breast cancer. *Clin Cancer Res*. 2010; 16(3):876–87. [PubMed: 20103682]

31. Choi J, Kim DH, Jung WH, Koo JS. Metabolic interaction between cancer cells and stromal cells according to breast cancer molecular subtype. *Breast Cancer Res.* 2013; 15(5):R78. [PubMed: 24020991]
32. Koo HR, Park JS, Kang KW, Cho N, Chang JM, Bae MS, et al. 18F-FDG uptake in breast cancer correlates with immunohistochemically defined subtypes. *Eur Radiol.* 2014; 24(3):610–8. [PubMed: 24097303]
33. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest.* 2011; 121(7):2750–67. [PubMed: 21633166]
34. Liebermeister W, Noor E, Flamholz A, Davidi D, Bernhardt J, Milo R. Visual account of protein investment in cellular functions. *Proc Natl Acad Sci U S A.* 2014; 111(23):8488–93. [PubMed: 24889604]
35. Yang F, Foekens JA, Yu J, Sieuwerts AM, Timmermans M, Klijn JG, et al. Laser microdissection and microarray analysis of breast tumors reveal ER-alpha related genes and pathways. *Oncogene.* 2006; 25(9):1413–9. [PubMed: 16261164]
36. Biswas DK, Shi Q, Baily S, Strickland I, Ghosh S, Pardee AB, et al. NF-kappa B activation in human breast cancer specimens and its role in cell proliferation and apoptosis. *Proc Natl Acad Sci U S A.* 2004; 101(27):10137–42. [PubMed: 15220474]
37. Nakshatri H, Bhat-Nakshatri P, Martin DA, Goulet RJ Jr, Sledge GW Jr. Constitutive activation of NF-kappaB during progression of breast cancer to hormone-independent growth. *Mol Cell Biol.* 1997; 17(7):3629–39. [PubMed: 9199297]
38. Smid M, Wang Y, Zhang Y, Sieuwerts AM, Yu J, Klijn JG, et al. Subtypes of breast cancer show preferential site of relapse. *Cancer Res.* 2008; 68(9):3108–14. [PubMed: 18451135]
39. Kennecke H, Yerushalmi R, Woods R, Cheang MC, Voduc D, Speers CH, et al. Metastatic behavior of breast cancer subtypes. *J Clin Oncol.* 2010; 28(20):3271–7. [PubMed: 20498394]
40. Hare SH, Harvey AJ. mTOR function and therapeutic targeting in breast cancer. *Am J Cancer Res.* 2017; 7(3):383–404. [PubMed: 28400999]
41. Paplomata E, O'Regan R. The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. *Ther Adv Med Oncol.* 2014; 6(4):154–66. [PubMed: 25057302]
42. Viedma-Rodriguez R, Baiza-Gutman L, Salamanca-Gomez F, Diaz-Zaragoza M, Martinez-Hernandez G, Ruiz Esparza-Garrido R, et al. Mechanisms associated with resistance to tamoxifen in estrogen receptor-positive breast cancer (review). *Oncol Rep.* 2014; 32(1):3–15. [PubMed: 24841429]
43. Boulay A, Rudloff J, Ye J, Zumstein-Mecker S, O'Reilly T, Evans DB, et al. Dual inhibition of mTOR and estrogen receptor signaling in vitro induces cell death in models of breast cancer. *Clin Cancer Res.* 2005; 11(14):5319–28. [PubMed: 16033851]
44. Giesen C, Wang HA, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods.* 2014; 11(4):417–22. [PubMed: 24584193]
45. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer.* 2012; 12(5):323–34. [PubMed: 22513401]
46. Polyak K. Breast cancer: origins and evolution. *J Clin Invest.* 2007; 117(11):3155–63. [PubMed: 17975657]
47. Gutierrez MC, Detre S, Johnston S, Mohsin SK, Shou J, Allred DC, et al. Molecular changes in tamoxifen-resistant breast cancer: relationship between estrogen receptor, HER-2, and p38 mitogen-activated protein kinase. *J Clin Oncol.* 2005; 23(11):2469–76. [PubMed: 15753463]
48. Miller TW, Balko JM, Arteaga CL. Phosphatidylinositol 3-kinase and antiestrogen resistance in breast cancer. *J Clin Oncol.* 2011; 29(33):4452–61. [PubMed: 22010023]
49. Likhite VS, Stossi F, Kim K, Katzenellenbogen BS, Katzenellenbogen JA. Kinase-specific phosphorylation of the estrogen receptor changes receptor interactions with ligand, deoxyribonucleic acid, and coregulators associated with alterations in estrogen and tamoxifen activity. *Mol Endocrinol.* 2006; 20(12):3120–32. [PubMed: 16945990]

Statement of significance

Findings utilize extensive proteomics to identify a novel luminal breast cancer subtype, highlighting the added value of clinical proteomics in breast cancer to identify unique features not observable by genomic approaches.

Fig. 1. Supervised analysis of breast cancer subtypes and healthy tissue.

A) Strip plots show the protein expression distribution of selected subtype markers. For significance test of each pairwise comparison we used Student's t-test, $\text{FDR} < 0.05$. **B)** Venn diagram of the numbers of significantly upregulated proteins in each subtype when compared to healthy, and processes enriched within the commonly upregulated proteins (Fisher Exact test, $\text{FDR} < 0.02$). **C)** Selected biological processes (from gene ontology - GO biological processes and molecular functions) that were enriched in the different subtypes when compared to healthy. q-values, after Benjamini-Hochberg FDR correction are

indicated next to each process in each subtype. * indicates $q\text{-value} < 0.02$. **D)** Proteomaps of the significantly different proteins show that each subtype is characterized with specific metabolic and signaling pathways. Each polygon represents a KEGG pathway and its size corresponds to the t-test difference of the proteins it contains.

Fig. 2. The proteomic clusters identified using consensus clustering.

A) Heatmap representing the consensus matrix of clustered samples. Blue color represents a higher consensus between each pair of samples. Visual inspection of the consensus matrix identifies four different groups. **B)** Alluvial plot shows the distribution of clinically defined subtypes across consensus clusters. Consensus clusters 1, 2 and 3 included mostly ER+ samples, while consensus cluster 4 included mostly TN ones. HER2 samples were spread and not clustered together. **C)** T-distributed stochastic neighbor embedding (t-SNE) analysis of 109 tumor samples, based on the ANOVA-significant proteins segregates the consensus

clusters. Consensus cluster 1 is more easily distinguished, while 2, 3 and 4 appear relatively closer on the t-SNE axes.

Fig. 3. Selected subtype markers across consensus clusters.

Selected subtype markers exhibit the subtype-expected pattern (for each pairwise comparison we used Student's t-test, FDR<0.05). MKI67 expression was significantly higher in CC1, suggesting its association with the LumB subtype; and the luminal KRT18 was significantly higher in all ER+ clusters (CC1-3) compared to the TN CC4.

Fig. 4. Comparison between the LumA consensus clusters.

A) 2D enrichment analysis of subtype-specific gene sets revealed common enrichment of ER+ related gene sets as well as ER+ metastasis related gene sets in both CC2 and CC3. Barcode plots show proteins ranked according to fold-change of CC2 or CC3 relative to CC4, where each vertical line represents a single protein. High rank indicates higher expression in either CC2 or CC3. **B)** Network centrality analysis of CC2 and CC4 upregulated proteins resulted in 21 overlapping proteins among the top 10% central in each group. TN-related signaling proteins are represented as dash-lined nodes. **C)** Density and rug-plots of proteins whose significantly high expression is common to CC2 and CC4, but not to CC3.

Fig. 5. Independent validation of the proteomic classification.

A) K-means clustering of TCGA RPPA data, based on 621 invasive ductal carcinoma (IDC) tumor samples, reproduces the proteomic classification. Two LumA-enriched clusters are identified when data is divided into 5 clusters, while the other three are each enriched (Fisher exact test, $FDR < 0.02$) for the known intrinsic subtypes (LumB, Basal and HER2). **B)** Comparison of k-means clusters 2 and 3, which correspond to the proteomic CC2 and CC3, reveals a similar expression pattern of signaling proteins such as MTOR, AKT and MAPK14. Alongside these, ER itself (both phosphorylated and non-phosphorylated) is

upregulated in k-means cluster 2 (Student's t-test, $FDR < 0.05$). **C**) Plotting the 666 proteins distinguishing CC2 and CC3 creates a clear differentiation between the samples in our proteomic data. Clustering of the mRNA levels of same genes from the TCGA RNA-sequencing data (**D**) or the Metabric data (**E**) shows no clear separation of subtypes or integrative clusters.