

Genome Wide Proteomics of ERBB2 and EGFR and Other Oncogenic Pathways in Inflammatory Breast Cancer

Emma Yue Zhang,[†] Massimo Cristofanilli,^{‡,§} Fredrika Robertson,^{||} James M. Reuben,^{§,||} Zhaomei Mu,[‡] Ronald C. Beavis,[⊥] Hogune Im,[∇] Michael Snyder,[∇] Matan Hofree,[#] Trey Ideker,[#] Gilbert S. Omenn,[¶] Susan Fanayan,[○] Seul-Ki Jeong,[◆] Young-ki Paik,[◆] Anna Fan Zhang,[†] Shiao-Lin Wu,[†] and William S. Hancock^{*,†}

[†]Barnett Institute and Department of Chemistry and Chemical Biology, Northeastern University, 360 Huntington Avenue, Boston, Massachusetts 02115, United States

[‡]Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, Pennsylvania 19107, United States

[§]The Morgan Welch Inflammatory Breast Cancer Research Program and Clinic, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States

^{||}Departments of Experimental Therapeutics and Hematopathology, The University of Texas, MD Anderson Cancer Center, Houston, Texas 77030, United States

[⊥]Department of Biochemistry and Medical Genetics, George and Faye Yee Centre for Healthcare Innovation, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

[∇]Department of Genetics, Stanford University, Stanford, California 94305, United States

[#]Department of Bioengineering, University of California, San Diego, La Jolla, California 92093, United States

[¶]National Center for Integrative Biomedical Informatics, University of Michigan, Ann Arbor, Michigan 48109, United States

[○]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia

[◆]Yonsei Proteome Research Center, Yonsei University, Seoul, Korea


Supporting Information

ABSTRACT: In this study we selected three breast cancer cell lines (SKBR3, SUM149 and SUM190) with different oncogene expression levels involved in ERBB2 and EGFR signaling pathways as a model system for the evaluation of selective integration of subsets of transcriptomic and proteomic data. We assessed the oncogene status with reads per kilobase per million mapped reads (RPKM) values for ERBB2 (14.4, 400, and 300 for SUM149, SUM190, and SKBR3, respectively) and for EGFR (60.1, not detected, and 1.4 for the same 3 cell lines). We then used RNA-Seq data to identify those oncogenes with significant transcript levels in these cell lines (total 31) and interrogated the corresponding proteomics data sets for proteins with significant interaction values with these oncogenes. The number of observed interactors for each oncogene showed a significant range, e.g., 4.2% (JAK1) to 27.3% (MYC). The percentage is measured as a fraction of the total protein interactions in a given data set vs total interactors for that oncogene in STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, version 9.0) and I2D (Interologous Interaction Database, version 1.95). This approach allowed us to focus on 4 main oncogenes, ERBB2, EGFR, MYC, and GRB2, for pathway analysis. We used bioinformatics sites GeneGo, PathwayCommons and NCI receptor signaling networks to identify pathways that contained the four main oncogenes and had good coverage in the transcriptomic and proteomic data sets as well as a significant number of oncogene interactors. The four pathways identified were ERBB signaling, EGFR1 signaling, integrin outside-in signaling, and validated targets of C-MYC transcriptional activation. The greater dynamic range of the RNA-Seq values allowed the use of transcript ratios to correlate observed protein values with the relative levels of the ERBB2 and EGFR transcripts in each of the four pathways. This provided us with potential proteomic signatures for the SUM149 and 190 cell lines, growth factor receptor-bound protein 7 (GRB7), Crk-like protein (CRKL) and Catenin delta-1 (CTNND1) for ERBB signaling; caveolin 1 (CAV1), plectin (PLEC) for EGFR signaling; filamin A (FLNA) and

continued...

ErbB receptor signaling network with RNA-Seq ratios

Gene	RPKM-SUM149	RPKM-SUM190	RATIO (149/190) ^a	ERBB2 interact	Novoseek tumor hits
EGFR	60.1	0.6	5.2	+	6585
AREG (AR)	73.7	4.9	3.7		121
EREG (EPR)	9.1	0.6	2.7	+	27
TGFA	3.9	0.5	1.7	+	926
BTC	0.5	1.0	-0.4	+	13
EGF	ND	ND	--	+	1513
HBEGF	1.5	4.7	-1.2	+	122
ERBB3	6.6	23.1	-1.7	+	169
ERBB4	0.0	3.0	-2.0	+	103
ERBB2	14.4	399.7	-4.7		5807

More 190  More 149

Received: February 19, 2013

Published: May 7, 2013

actinin alpha1 (ACTN1) (associated with high levels of EGFR transcript) for integrin signalings; branched chain amino-acid transaminase 1 (BCAT1), carbamoyl-phosphate synthetase (CAD), nucleolin (NCL) (high levels of EGFR transcript); transferrin receptor (TFRC), metadherin (MTDH) (high levels of ERBB2 transcript) for MYC signaling; S100-A2 protein (S100A2), caveolin 1 (CAV1), Serpin B5 (SERPINB5), stratifin (SFN), PYD and CARD domain containing (PYCARD), and EPH receptor A2 (EPHA2) for PI3K signaling, p53 subpathway. Future studies of inflammatory breast cancer (IBC), from which the cell lines were derived, will be used to explore the significance of these observations.

KEYWORDS: ERBB2, EGFR, Inflammatory breast cancer, Chromosome-centric Human Proteome Project,

■ INTRODUCTION

Breast cancer is a major health problem with over 40 000 deaths each year in the United States. We have previously studied proteomics and glycoproteomics in samples collected from breast cancer patients^{2–4} as potential markers for the early detection of breast cancer. As an extension of these studies, we report in this manuscript on a study of protein expression as measured by both RNA-Seq¹ and proteomics of two cell lines established from primary inflammatory breast cancer (IBC) tumors,⁵ namely, SUM149 and SUM190, which are ER (–) and PR (–), as well as the well-studied cell line SKBR3, which is known to express high levels of ERBB2 and is ER (–) and PR (–).

EGFR and ERBB2 are members of the epidermal growth factor receptor (EGFR) family, one of 20 subfamilies of human receptor tyrosine kinases (RTK).⁶ The EGF family is one of the best studied growth factor receptor systems, often overexpressed in human tumors.^{7–9} Several small molecule inhibitors and protein drugs have been developed to modulate disorders in the EGFR family.^{10,11} Moreover, determination of ERBB2 status by immunohistochemistry (IHC) or fluorescent in situ hybridization (FISH) has been recommended by the American Society of Clinical Oncology (ASCO) as a marker for diagnosis and evaluation in primary invasive breast cancer.¹² Initially we will describe the analysis of the RNA-Seq data to determine the presence or absence of oncogenes typically associated with breast cancer as well as the levels of the target oncogenes ERBB2 and EGFR. These studies demonstrated the importance of EGFR and ERBB family members in the cell lines, as well as other oncogenes such as TP53, CRKL, EZR and MYC. We then explored different approaches to integrate the proteomic information with the transcriptome data and compared the proteomic levels as measured by spectral count with the transcript level as well as interaction values of the observed proteins with the panel of oncogenes. These comparisons highlighted the 4 oncogenes, namely, EGFR, ERBB2, MYC and GRB2, and allowed the identification of protein-based subpathways of interest for the different cell lines.

■ MATERIALS AND METHODS

Cell lines, Cell Lysis, and In-Gel Digestion

Cell Lines SKBR3, SUM149 and SUM190. The human breast cancer cell lines SKBR3 (ER/PR–, HER2+, metastatic pleural effusion), was obtained from the American Type Culture Collection (Manassas, VA) and maintained in culture with DMEM/F-12 medium supplemented with 10% FBS (Tissue Culture Biologicals, Seal Beach, CA) and 1% of Antibiotic-Antimycotic 100X (Gibco, Carlsbad, CA).

SUM149 and SUM190 cells were obtained from Dr. Stephen Ethier (Kramanos Institute, MI, USA) and are commercially available (Asterand, Detroit, MI). SUM149 cells are ER/PR–, HER2– (triple receptors negative), and the SUM190 cells are ER/PR–, HER2+. Both human IBC cell lines were maintained in culture with Ham's/F-12 medium supplemented with 10% FBS

(Tissue Culture Biologicals, Seal Beach, CA), 5 µg/mL of insulin, 1 µg/mL of hydrocortisone and 1% of Antibiotic-Antimycotic 100X (Gibco, Carlsbad, CA).

Twenty microliters of lysis buffer (2% SDS in 50 mM NH₄CO₃) was added to 10 µL of cell lysate. Cells were solubilized by sonication using 20 s bursts, followed by cooling on ice for 20 s, in a process that was repeated for 10 times. The entire extract was concentrated down to 15 µL in a speed vacuum and loaded onto a gel (SDS-PAGE, 4–12% gradient) to separate proteins by molecular weight. After staining with Coomassie blue, each gel lane was cut into five individual slices as shown in Figure S1 (Supporting Information).

Each slice was further minced into smaller pieces (approximately 0.5 mm²). The gel slices were washed with 600 mL of water for 15 min and centrifuged, supernatant was removed, and 50% ACN was added (1 mL), followed by shaking until no visible Coomassie stain remained. Proteins were then reduced with dithiothreitol (DTT) by adding 250 µL of 10 mM DTT in 0.1 M NH₄CO₃ and incubated for 30 min at 56 °C. Samples were subsequently alkylated at room temperature and in the dark for 80 min with 250 µL of 55 mM iodoacetamide (IAA) in 0.1 M NH₄CO₃. Trypsin digestion reagent (200 µL; 10 ng/mL of trypsin in 50 mM NH₄CO₃, pH 8.0) was added, and samples were incubated for 30 min at 4 °C. The trypsin concentration was based upon an estimate of approximately 0.1–0.5 mg of protein per gel slice and adjusted as necessary. The solution was then replaced with 50 mM NH₄CO₃ to cover the gel pieces (50 µL) and incubated overnight at 37 °C to elute peptides from the gel. Following this step, supernatant was removed and stored. Gel pieces were further extracted with 5% formic acid (30 µL) and acetonitrile (ACN, 400 µL) at 37 °C for 10 min and then twice with 5% formic acid (30 µL) and ACN (200 µL). The formic acid solution containing tryptic peptides was combined with the previous supernatant and concentrated to 5–10 µL. The concentrated solution (trypsin-digested peptides) was subjected to LC–MS analysis.

LTQ-FT MS

The in-gel digested peptides were analyzed by online LC using a linear IT coupled to a Fourier transfer mass spectrometer (LTQ-FT MS, Thermo Electron, San Jose, CA) with a Dionex nano-LC instrument (Ultimate 3000, Sunnyvale, CA) and a 75 mm i.d. × 15 cm C-18 capillary column packed with Magic C18 (3 mm, 200 Å pore size) (Michrom Bioresources, Auburn, CA). The LTQ-FT mass spectrometer was operated in the data-dependent mode to switch automatically between MS and MS/MS acquisition. Survey full-scan MS spectra with two microscans (m/z 400–2000) were acquired in the Fourier transform ion cyclotron resonance cell with a mass resolution of 100 000 at m/z 400 (after accumulation to a target value of 2×10^6 ions in the linear IT), followed by ten sequential LTQ-MS/MS scans throughout the 90 min separation. The analytical separation was carried out using a three-step linear gradient, starting from 2% B to 40% B in 40 min (A: water with 0.1% formic acid; B: ACN with

0.1% formic acid), increased to 60% B in 10 min, and then to 80% B in 5 min. The column flow rate was maintained at 200 nL/min.

Protein Identification

Peptide sequences were identified using Thermo Proteome Discoverer 1.3 from a human database SP.human.S6.5 with full trypsin specificity and up to three internal missed cleavages. The tolerance was 50 ppm for precursor ions and 0.8 Da for product ions. Dynamic modifications were deamidation of asparagine, and static modification was carbamidomethylation for cysteine. Peptides were identified with Xcorr scores above the following thresholds: ≥ 3.8 for 3+ and higher charge state ions, ≥ 2.2 for 2+ ions, and ≥ 1.9 for 1+ ions.

We used the spectral count approach to measure relative abundance of protein samples as reported by Choi et al.¹³ We have selected several housekeeping proteins, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), b-actin (ACTB), b-tubulins (2A, 2B, 2C, 3 and 5),¹⁴ which are ubiquitously expressed in a wide range of tissues and cell types,¹⁵ as internal standards for relative quantification in order to minimize variations in the amount of samples loaded on the 1D SDS-PAGE gel. These proteins met the required criteria of high abundance and consistent ratios across the 3 cell lines, as measured by peptide counts and extracted ions in the same gel section between the different cell lines.

The protein list also was submitted to the Gene A La Cart (provided by www.genecards.com, uploaded to Gene A La Cart for analysis in August, 2011) to acquire data for bioinformatics analysis, including gene symbols and other genomic information.

RNA-Seq Measurement

Strand-specific RNA-Seq libraries were prepared and sequenced on a lane of the Illumina HiSeq 2000 instrument per sample to obtain transcript data.¹⁶ All RNA-Seq data are available at Short Read Archive (SRS366582, SRS366583, SRS366584, SRS366609, SRS366610, SRS366611).

From total RNA, strand-specific RNA-Seq libraries were prepared according to Illumina TruSeq standard procedures and sequenced at both ends (paired-end RNA-sequencing) on Illumina HiSeq 2000. Tophat embedded with Bowtie was used to align the sequence reads to human genome (hg19). Using Cufflinks, the alignments were assembled into gene transcripts (NCBI build 37.2), and their relative abundances (RPKM) were calculated.

RESULTS AND DISCUSSION

We have previously studied on the role of two driver oncogenes, EGFR, ERBB2, in epithelial cancers^{17,18} and have investigated the changes in their glycosylation patterns.^{2–4,19} To further expand on our previous observations, we have performed a comparative study to explore the total lysate proteome of a well-established epithelial breast cancer cell line, SKBR3, which overexpresses ERBB2 and two primary cell lines (SUM149 and SUM190) isolated from patients with inflammatory breast cancer.⁵ We have employed a traditional proteomic analysis of the data and compared these results with an alternative format, namely genome-wide proteomics using the chromosome format (C-HPP²⁰), which is being developed as part of the HUPO human proteome initiative. One benefit of such approach is the facile integration of proteomic and transcriptomics data as well as allowing for the identification of genomic regions in which a driver oncogene may effect gene transcription of adjacent genes.

Analysis of Cell Lines SKBR3, SUM149, and SUM190

Each cell line was analyzed in triplicate, and relative quantitation was achieved with spectral counts using a correction factor based on housekeeping proteins. With the availability of a deep measurement of the transcriptome, by RNA-Seq (100 million reads), it is common to measure 10 000–11 000 transcripts in a cell line study. In contrast, a proteomic study comparable to what is reported here will sample only approximately 10% of the expressed set of proteins. While the transcriptome can enhance the proteomic measurement, the opposite is also true as a medium level protein study can be used to explore the major phenotypic patterns observed in a study of disease versus normal cell lines and patient tissue. In addition, there are examples of a protein being identified in the absence of a measurable transcript level.²¹

In the proteomic analysis we used a conservative protocol for identifying proteins in replicate analysis, which included high protein confidence and high peptide rank (Proteome Discoverer) and with a FDR of less than 1%. We identified a total of 1444, 1396, and 964 proteins (numbers of proteins with 2 or more peptides) in the SKBR3, SUM149 and 190 cell line samples, respectively (numbers of proteins with 2 or more peptides were 1071, 1134, and 686 for SKBR3, SUM149 and SUM190, respectively). In addition, selected proteins identified by one single peptide were further analyzed using additional criteria such as high mass accuracy, fragmentation spectra and observation of the corresponding transcript (see Table 1). In the cell line studies a comparison of the SKBR3 with SUM190, SKBR3 with SUM149, and SUM190 with SUM149 proteome contents identified 751, 934, and 695 common proteins, respectively.

Characterization of EGFR and ERBB2

EGFR was identified in SUM149 and SKBR3 cell lysates, while ERBB2 was identified in SKBR3 and SUM190 cell lysate preparations, consistent with IHC results in previous studies.^{5,22} As shown in Table S1 (Supporting Information), EGFR and ERBB2 were identified with 11 and 13 peptides for cell lines SUM149 and SKBR3, respectively. This table employs data from GPMDB (Global Proteome Machine database)²³ to assess the quality of peptides observed for the two proteins. The peptides observed in our study have been frequently reported in the literature, e.g., rank 1–5 and 1–4 for the most frequently observed, as well as other peptides for EGFR and ERBB2, respectively. The MS/MS data for a diagnostic peptide for EGFR and ERBB2 is shown in Figure S2 (Supporting Information). Both EGFR and ERBB2 were detected with good sequence coverage (15.5 and 15.8%), although peptides derived from the N-terminal domain of ERBB2 were not observed. The identification of ERBB2 was confirmed by immunoprecipitation with the monoclonal antibody trastuzumab (Herceptin) and subsequent analysis on 1D SDS-PAGE and detected at an approximate molecular weight (MW) of 110 000 (theoretical 138 kD, data not shown).

Protein Observations with RNA-Seq Data and Expressed in a Genome Wide Format (Chromosomes)

Besides proteomic analysis, we have also discovered potential proteins of interest by comparing proteomics data with the corresponding transcriptomic data in a chromosome format (see Tables S3 and S4, Supporting Information, for the RNA-Seq results for SKBR3, SUM149 and SUM190). We collected the genomic information from the Gene A La Cart tool provided by www.genecards.org. In doing so, UniProt accession numbers for

Table 1. List of Oncogenes Associated with Breast Cancer with Associated Proteomic and Transcriptomic Data

Gene Symbol ^a	SKBR3 ^b	SUM190 ^b	SUM149 ^b	SKBR3 ^c	SUM190 ^c	SUM149 ^c
ACOT8	ND ^d	ND	ND	34.4	2.2	9.4
APC	ND	ND	ND	ND	3.4	1.2
BRCA 1	ND	ND	ND	1.0	28.3	1.9
CDKN2A	ND	ND	ND	6.4	3.3	ND
CEACAM6	ND	ND	ND	ND	16.2	ND
CRKL	ND	9	ND	3.3	56.3	12.3
DEK	5	ND	ND	32.4	12.5	26.0
EGFR ^e	12	ND	22	1.4	ND	60.1
ERBB2 ^e	46	19	ND	300.1	399.7	14.4
ERBB2IP	ND	ND	ND	1.1	15.2	8.3
ERBB3	ND	ND	ND	6.6	23.1	6.6
ERBB4	ND	ND	ND	ND	3.0	ND
EZR	92	40	46	54.0	10.9	26.4
GRB2 ^e	ND	ND	ND	9.0	7.7	10.0
GRB7	2 ^f	2 ^f	ND	30.3	57.5	5.8
JAK1	ND	ND	ND	1.8	12.0	7.9
KRAS	ND	ND	2	1.5	8.5	6.9
MET	ND	ND	ND	ND	5.6	9.3
MLH1	ND	ND	ND	3.5	3.6	6.3
MSH2	2 ^f	ND	3	2.8	2.9	10.6
MTOR	ND	ND	ND	1.7	7.2	16.2
MYC ^e	ND	ND	ND	19.0	9.6	79.1
PIK3R1	ND	ND	ND	ND	16.3	3.3
PIK3R2	ND	ND	ND	7.4	1.7	16.0
PIK3R3	ND	ND	ND	4.5	28.1	1.5
PPP1R1B	ND	ND	ND	1.3	ND	ND
PTEN ^e	ND	ND	ND	0.8	6.1	0.8
RB1	ND	ND	ND	1.2	2.0	1.7
RHOC	11	ND	16	53.6	24.7	47.8
SRC	ND	ND	ND	6.2	ND	7.5
STAT1	ND	ND	1 ^f	1.5	9.7	11.3
TOP2A	58	ND	17	50.2	5.3	25.9
TP53	ND	ND	ND	14.8	1.3	40.6

^aGene symbols are from Genecards. ^bSpectral counts. ^cRPKM values. ^dND = not detected. ^eOncogenes used for pathway analysis are highlighted by box. ^fIdentifications of single peptide proteins are shown in Figure S3 (Supporting Information).

their corresponding proteins were first extracted from the search result files in Proteome Discoverer, prior to submission to Gene a la Cart as identifiers to retrieve their genomic information, including gene symbols, genomic locations (chromosome number, base pair location of gene start and end, and gene size), and Ensemble cytobands. This would allow the protein list to be organized by their locations on different chromosomes. The resulting data sets for the three cell lines are shown in Table S3 (Supporting Information).

Use of RNA-Seq Data to Explore ERBB2 Signaling Pathways

As a first step we generated a list of 33 oncogenes associated with breast cancer from the Sanger, Genecards databases, and literatures,^{24,25} which had either measurable transcript level (RPKM >1) and in some cases proteomic data (see Table 1). The RNA-Seq values showed that the cell line SUM149 had a high level of transcript for EGFR (RPKM = 60) and a relatively low value for ERBB2 (RPKM = 14). Conversely the cell line SUM190 had values of 400 and ND for ERBB2 and EGFR, respectively. The immortalized cell line SKBR3 expressed a high

level of ERBB2 and a low level of EGFR (RPKM = 300 and 1.4, respectively). Other oncogenes with a high level of transcript (RPKM > 40) were TP53, MYC (SUM149); GRB7, CRKL (SUM190) and EZR, TOP2A (SKBR3). As described in a later section we also explored reported interactions between the group of 31 oncogenes and the proteins observed in the SUM149 and 190 proteomic results.

Figure 1A and B compares the ERBB2 signaling pathway in two IBC cell lines, SUM149 (high levels of EGFR transcript) and SUM190 (high levels of ERBB2 transcript) with the ERBB2 pathway derived from the KEGG database. SUM190 presents an interesting situation with high transcript levels of ERBB2 and ERBB3 (RPKM = 400 and 23, respectively) and a low level for ERBB4 (RPKM = 3), without detectable transcript levels of EGFR (ERBB1) and a low RNA-Seq value (4.91) for amphiregulin (AR in Figure 1), one of the EGFR ligands.²⁶ ERBB2 is a special member in the ERBB family in that there has been no ligand discovered for ERBB2 and signaling largely depends on heterodimer formation with either EGFR, ERBB3 or

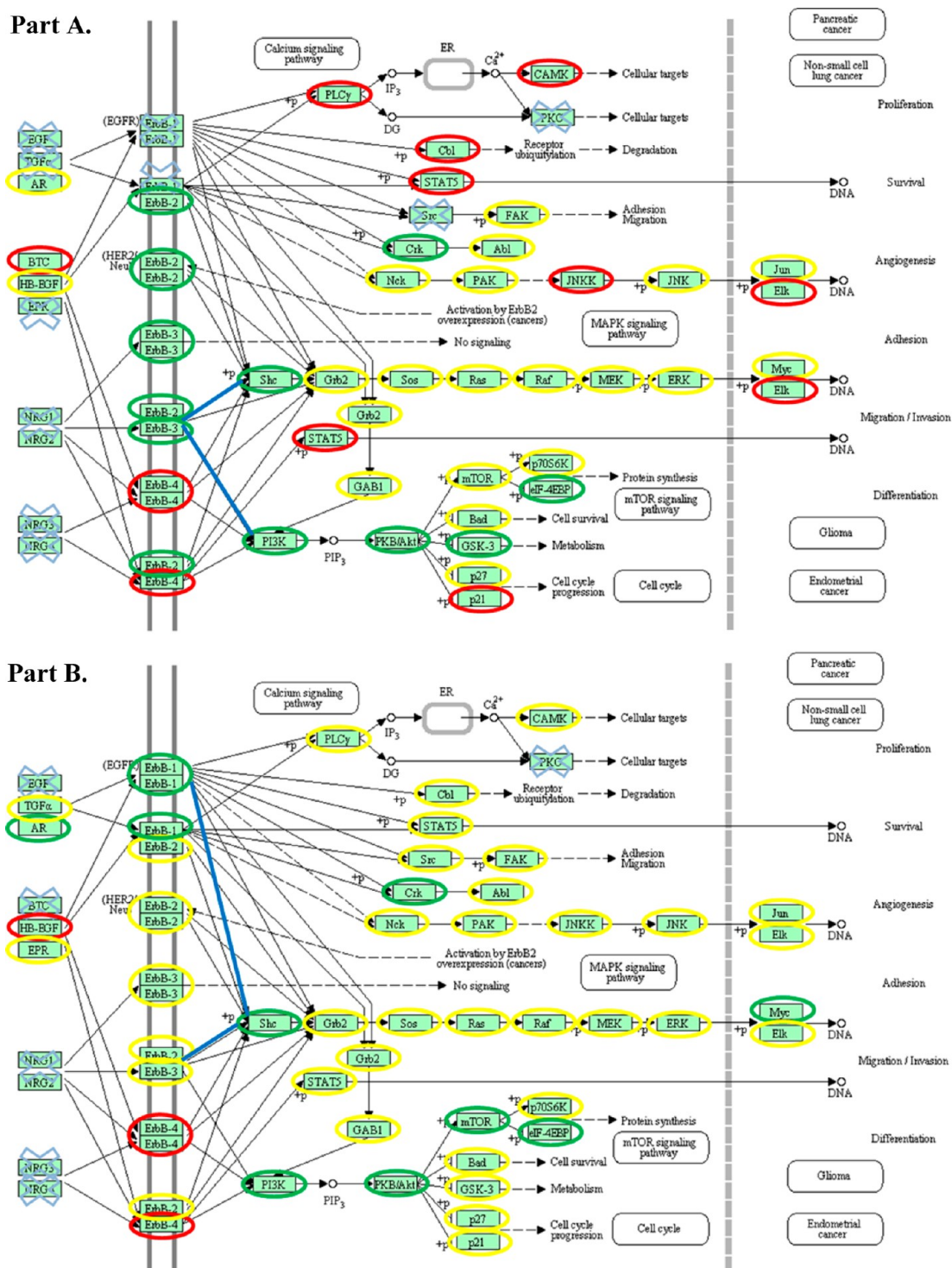


Figure 1. Annotation of KEGG ERBB2 signaling pathways with transcriptomic data, (A) SUM190, (B) SUM149. The pathway was derived from <http://www.genome.jp/kegg/pathway/hsa/hsa04012.html> in January 2012. The RPKM values are shown as follows. Green circle: RPKM value is larger than 15; yellow circle: RPKM value is between 3 and 15; red circle: RPKM value is between 1 and 3; blue cross: transcription value is under detection limit; blue line: potential preferred signaling based on transcript levels.

ERBB4.^{27,28} However, a high level of ERBB3 is found in the SUM190 transcript, and it has been reported that the ERBB2/ERBB3 heterodimer is active in cell proliferation in breast tumor cells (see highlighted blue lines in Figure 1A).²⁹ Conversely, as is shown in Figure 1B (highlighted blue lines) with the observed transcript values in the SUM149 cell line for the EGFR family signaling pathway there are several possibilities for signaling with

involvement of EGFR dimers, ERBB2 heterodimers with EGFR or ERBB3. Since ERBB4 is not detected at either the transcript or protein level, it is presumably not part of the signaling cascade. Thus RNA-Seq studies identified potential differences between the two cell lines and thus set the stage for a proteomic investigation. Another advantage of the RNA-Seq studies was the greater dynamic range than the proteomic measurement; one

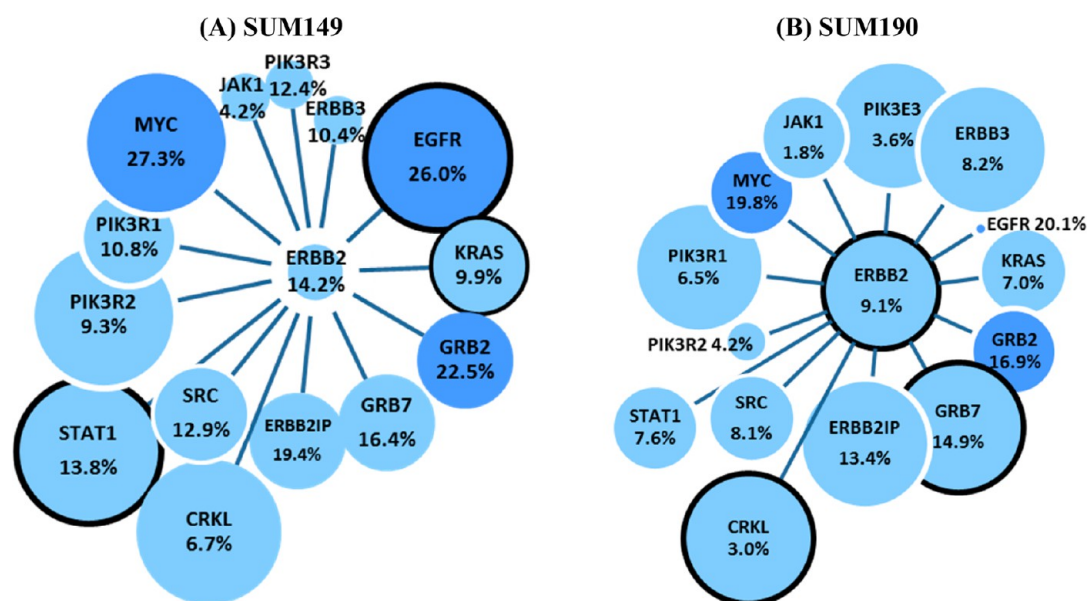


Figure 2. A composite of SUM149 (A) and SUM190 (B) transcriptomic, proteomic, and interaction data for significant oncogenes observed in SUM149 and SUM190. The following notations are used. Line length: Interaction score (shorter line, stronger interaction with ERBB2). Circle size: RPKM value (largest: RPKM > 15, medium: RPKM between 3 and 15, small: RPKM between 1 and 3, spot: RPKM < 1). Black circle: if observed in proteomic experiments. Percentage: percentage of proteins identified in SUM149 or 190 with specific oncogene interactions as listed by STRING or I2D in Genecards.org.

important example was identification of high levels of the transcript for the MYC oncogene in SUM149 and 190 (19 and 10, respectively) in the absence of a proteomics signal. The importance of this oncogene is consistent with the importance of the MEK/ERK pathway in carcinogenesis (see arrow in Figure 1) and is supported by the large number of MYC interactors identified in the proteomics study (see Figure 2 and discussion later).

To further explore the difference between EGFR and ERBB2 signaling in SUM190 and 149 transcriptome, we used the ratio of the RPKM values to interrogate the NCI ErbB receptor signaling network and visualized the data by assigning different colors based on the ratio values. First, EGFR and ERBB2 are the most differentially expressed genes in this network. As can be seen in Table 2 increased levels of EGFR transcript are associated with increased levels of the ligands amphiregulin (AR), epiregulin (EPR) and transforming growth factor, alpha (TGFA) for SUM149 vs 190, while the transcript levels for ERBB2 and associated receptors/ligands HBEGF, ERBB3 and 4 are increased in SUM190 vs 149. Amphiregulin is identified as a ligand of EGFR²⁶ and acts as an effective mitogen for epithelial cells.³⁰ Epiregulin is another EGFR ligand that binds directly to EGFR and regulates tyrosine phosphorylation of EGFR.³¹ On the other hand, ERBB3 and ERBB4 are reported to be part of ERBB2 heterodimer in ERBB2 signaling, and ERBB3 has been reported to be necessary for tumor cell proliferation in breast cancer.²⁹

Proteomic Analysis of SKBR3, SUM149, and 190 Cell Lines

With the importance of ERBB2 and EGFR signaling indicated by the RNA-Seq data, we then examined the correlation between our proteomics data, transcript levels and chromosome location. In Table S2 (Supporting Information) we have ranked the 20 most abundant proteins as measured by spectral count in the SKBR3 cell line (highest number of protein observations) and compared these values with the corresponding RNA-Seq levels as well as the proteomic values for SUM149 and 190 cell lines. As has been reported elsewhere⁹ there is a general correlation

Table 2. ErbB Receptor Signaling Network^a with RNA-Seq Ratios (SUM149 vs SUM190)^b

Gene	RPKM-SUM149	RPKM-SUM190	RATIO (149/190) ^b	ERBB2 interact	Novoseek tumor hits
EGFR	60.1	0.6	5.2	+	6585
AREG (AR)	73.7	4.9	3.7		121
EREG (EPR)	9.1	0.6	2.7	+	27
TGFA	3.9	0.5	1.7	+	926
BTC	0.5	1.0	-0.4	+	13
EGF	ND	ND	--	+	1513
HBEGF	1.5	4.7	-1.2	+	122
ERBB3	6.6	23.1	-1.7	+	169
ERBB4	0.0	3.0	-2.0	+	103
ERBB2	14.4	399.7	-4.7		5807

More 190 → → → → More 149

^aGene set in this pathway is retrieved from the following link: http://pid.nci.nih.gov/search/pathway_landing.shtml?pathway_id=erbb_network_pathway&pathway_name=ErbB%20receptor%20signaling%20network&source=NCI-Nature%20curated&what=graphic&jpg=on&page=1. All pathways from Nature Cancer Institute were released on October 12, 2011.

^bRPKM values are used to show the expression differences in two IBC cell lines, and the values in the ratio column are calculated as follows: Ratio (149/190) = $\log_2 \left(\frac{(\text{RPKM SUM149} + 1)}{(\text{RPKM SUM190} + 1)} \right)$. By adding 1 to RPKM values artificially, the ratio could still be calculated even if RPKM value is 0.

between the levels of a transcript and the corresponding proteins, although relative differences in transcript and protein stability as well as temporal events can result in exceptions to this rule. The genes TUBB, ACTB and GAPDH, which were selected as housekeeping proteins for normalization of the proteomic data, were indeed observed at high levels (spectral count rank 17, 8, and 7, respectively, for SKBR3), and these genes were also observed with high transcript values (RPKM of 209, 1391, and 2966, respectively). Conversely, the genes HIST1H4A, EPPK1, ENO3 and FLNA offer examples of poor correlation with a rank of 15, 9, 10, 11 in the proteomic data and a RPKM of only 3, 4, 2, and 6, respectively. While the selection of 20 examples in Table S2 (Supporting Information) as a representative protein set is arbitrary, it is of interest to note that 11 of the 20 proteins are

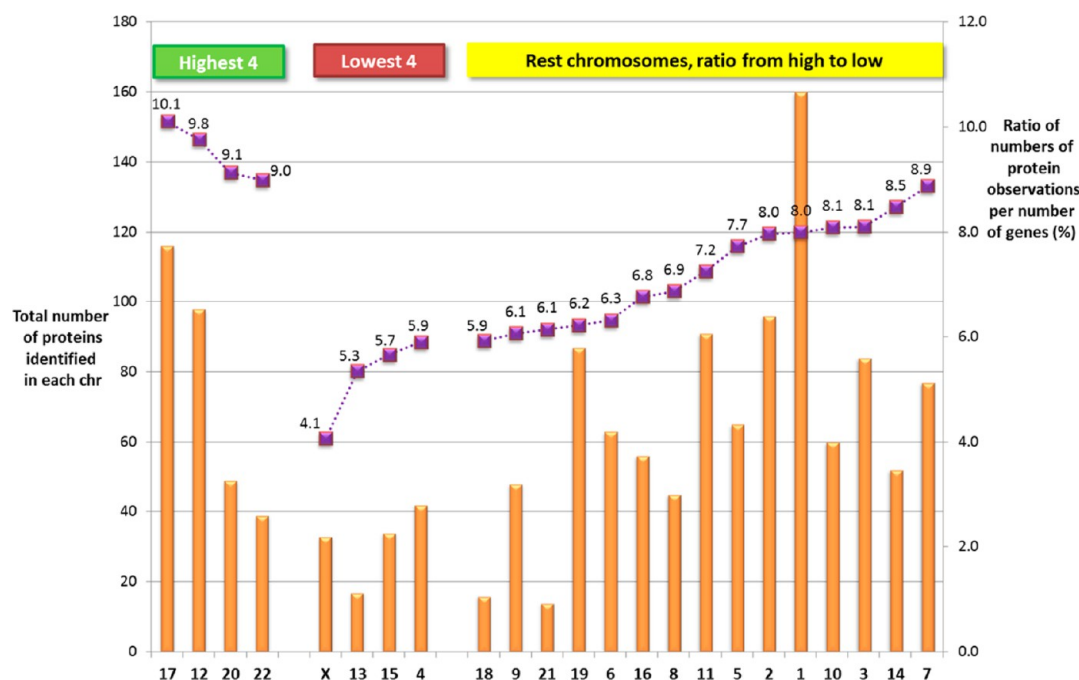


Figure 3. Ratio of number of protein observations per number of genes for each chromosome. Solid bar: total number of proteins identified in each chromosome for proteomic analysis in SKBR3. Squares: ratio of proteins identified in proteomic experiments relative to total gene numbers for each chromosome (as a percentage).

located on just 3 chromosomes: 6, 12, and 17. The possible significance of this observation will be discussed in the next section. One of the proteins coded by the gene MYH9 is a known oncogene,³² and such a high level of expression is of potential interest.

It has been reported there is a relationship between levels of gene expression and gene density in a chromosome region.⁹ Figure 3 shows the number of proteins identified in the SKBR3 study reported for each chromosome together with the % observed (number of protein observations divided by the number of protein coding genes on the chromosome). It is not surprising that the highest number of protein observations occurs for chromosome 1 and the lowest for chromosome 13 (largest chromosome and chromosome with lowest number of protein coding gene density, respectively). The highest % values were observed for genes 17, 12, 20, and 22, and while there is some correlation with reported gene densities on each chromosome (order of gene density is 19, 17, 20 and 22, high to low) it is relevant to note that chromosome 12 had 5 of the 20 most abundant proteins in Table S2 (Supporting Information), followed by chromosome 17 (3). Another factor is that chromosome 17 contains the highly expressed oncogene ERBB2 that can amplify a set of genes collocated near this oncogene (9).

Comparison of Proteomic Observations between Cell Lines

One of the challenges of studies with cancer cell lines compared with patient derived tumor samples is the lack of a suitable control samples. We chose the levels of ERBB2 as the comparator and compared the relative abundance of proteins in the two ERBB2 expressing cell lines (SUM190 and SKBR3, RPKM = 400 and 300) with SUM149 (RPKM = 14) in terms of unique proteins and for proteins with a 10-fold higher expression (see Tables S3, Supporting Information). Examples of proteins observed with this approach include the RAS associated proteins that are commonly activated in tumors in which ERBB2 is over-expressed.^{31,33} RAS-related proteins were preferentially observed


in SUM190 and SKBR3 (ERBB2+) in that of the 24 different types of RAS-related proteins identified, SUM190 and SKBR3 accounted for 15 and 20, respectively, while only 6 were shared by all three cell lines. In addition, there are 5 RAS proteins with relative abundance 2-fold higher in SUM190 and SKBR3 compared to SUM149. Another example is cathepsin D which was elevated 6X and 10X more in SUM190 and SKBR3 compared to SUM149 and has previously been associated with Her2 amplification²³ and is a marker of breast cancer marker.¹² While this type of data analysis did detect some proteins with cancer associations it did not lead to pathway discoveries similar to that observed with the RNA-Seq analysis, and thus we explored alternative approaches.

Mapping of Oncogene Interactions with Proteomic Observations

With the use of interaction scores provided by Genecards (String, I2D) we recorded the values for interactions between the proteins identified in the proteomic studies of the two IBC cell lines and 21 oncogenes listed in Table 1. The large data set is given in Tables S3 and S4 (Supporting Information), and a summary is given in Figure 2 with the proteomic and transcriptomic experimental data as well as number of interacting proteins. First, Figure 2 shows oncogenes that are known to interact with ERBB2, and the oncogenes that show a high degree of interaction (EGFR, ERBB3, ERBB2IP, GRB2, GRB7, KRAS) are denoted by a shorter line. A relatively high RNA-Seq measurement is shown by the size of the circle, e.g., ERBB2, GRB7 and MYC, and those oncogenes with a proteomics value are shown with a black outline, e.g., ERBB2, GRB7, CRKL, TOP2A (see Table 1 for numerical values). For each oncogene, the number of interactions with proteins observed in the proteomic studies of either SUM149 or 190 is given in the circle as a percentage of the total oncogene interactions. As shown in Figure 2 the top 3 oncogenes with the greatest number of interactions with observed proteins are MYC, GRB2 and EGFR with 268, 235, and 143 interactions,

Table 3. EGFR1 Signaling from NCI

Gene	SKBR3 ^a	SUM190 ^a	SUM149 ^a	SKBR3 ^b	SUM190 ^b	SUM149 ^b	Ratio 149/190 ^c	EGFR interact	Novoseek tumor hits
KRT17 ^d	66	73	27	1.6	0.2	197.8	7.3	+	24
EGFR ^d	12		22	1.4	0.6	60.1	5.2	+	6585
CAV1 ^d			4	0.4	0.2	33.6	4.9	+	434
PLEC ^d	332	16	91	15.6	7.1	69.7	3.1	+	1
SRC ^e				6.2	0.1	7.5	3.0	+	397
PIK3R2 ^e				7.4	1.7	16.0	2.7	+	2
HTT			1	0.5	1.9	9.0	1.8	+	4
MTA2	3		2	19.5	11.5	37.6	1.6		5
AP2A1	8			9.4	4.3	14.9	1.6	+	0
HAT1 ^d			3	3.0	6.5	12.5	0.8		0
STAT5B	2			3.4	1.4	3.2	0.8	+	92
EPS15			1	2.8	4.1	7.8	0.8	+	2
KRT7 ^d	380	634	50	654.7	338.9	575.6	0.8	+	466
YWHAB	46	70	77	70.2	18.3	29.1	0.6	+	0
CRK	2		1	6.9	14.7	18.2	0.3	+	19
RALB	3		1	11.2	6.6	8.1	0.3		9
EEF1A1 ^d	162		250	903.9	309.6	357.8	0.2	+	20
STAT1			1	1.5	9.7	11.3	0.2	+	121
KRT8 ^d	655	286	28	2025.0	60.0	61.3	0.0	+	116
KRT18 ^d	334	131	21	1470.0	71.3	64.1	-0.2	+	158
KRAS			2	1.5	8.5	6.9	-0.3	+	615
RAB5A	8			5.4	6.7	5.2	-0.3	+	0
STAT3	35			23.4	13.8	10.6	-0.3	+	693
MAP2K1 ^d		2		2.4	10.3	7.8	-0.4	+	45
PEBP1 ^d	26	45	12	104.3	43.1	30.3	-0.5		22
JAK1 ^e				1.8	12.0	7.9	-0.6	+	9
RAC1	15	14	12	207.9	23.0	12.0	-0.9	+	88
ARF4 ^d	11	35		17.2	24.9	12.2	-1.0	+	0
CDC42 ^d		9	20	23.0	37.7	17.5	-1.1	+	43
CTNND1 ^d		28	7	5.5	68.9	19.3	-1.8	+	83
SH3BGR1		2	2	0.7	18.6	3.7	-2.1	+	0
CRKL ^d		9		3.3	56.3	12.3	-2.1	+	4
GRB7	2	2		30.3	57.5	5.8	-3.1	+	22
PIK3R3 ^e				4.5	28.1	1.5	-3.5	+	0

More 190  More 149

^aSpectral counts. ^bRPKM values. ^cRPKM values are used to show the expression differences in two IBC cell lines, and the values in the ratio column are calculated as follows: Ratio (149/190) = $\log_2 \left(\frac{(\text{RPKM SUM149} + 1)/(\text{RPKM SUM190} + 1)}{1} \right)$. ^dProteins with higher expression in SUM149 are highlighted in yellow and in blue those with higher levels in SUM190. ^eKnown oncogene.

respectively, for SUM149. The basis for this approach has been used by others in the development of bioinformatic processes for prioritizing cancer associated genes with gene expression data combined with protein–protein interaction network information,³⁴ as well as the observation that proteomic data when combined with genomic information can add further discrimination to pathway analysis.³⁵ Thus in our approach we have combined mapping of oncogenes with RNA-Seq levels and identification of interacting proteins in the proteomic data set, and we will now use this data to search for additional pathways of interest in breast cancer.

Identification of Pathways That Contain ERBB2, EGFR, GRB2 and MYC Interactors

As an example of our process we describe the selection process for ERBB2 interactors. From the proteomic data set 35 proteins were found to be interacted with ERBB2 on the basis of I2D and STRING databases. We then selected a subset of 14 proteins according to levels of protein expression (spectral counts) and RNA-Seq values in the two IBC cell lines, SUM149, 190 and the model cell line SKBR3 (see Table S3, Supporting Information). The process was repeated for the 3 other oncogenes with greatest number of interactors with the proteomic data set (Figure 2), namely, EGFR, GRB2 and MYC that resulted in 172, 289, and 336 strong interactors with significant RNA-Seq or proteomics levels, respectively.

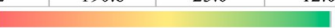
Table S3 (Supporting Information) also lists the chromosome locations of the interacting proteins, and it is noteworthy that

many of the genes in these pathways are located on cytoband 17q12, which is the site of the ERBB2 amplicon.^{21,36} Of this group of chromosome 17 genes, ERBB2, GRB7, STAT3 and KRT17 are located in the same chromosome region (17q12 to q21.2) and have the following Novoseek tumor associations based on literature text-mining (Genecards): 5807, 22, 693, and 24. The next stage in our process was to select disease relevant pathways based on our integration of transcriptomic, proteomic and interaction data. Our goal was to find at least one pathway for each of the 4 oncogenes that were well represented by the proteins listed in Table S3 (Supporting Information) and we used Cytoscape and Pathway Commons in this search. The pathways that we have selected are ERBB2, MYC, and PI3K signaling pathways from NCI Pathway Interaction Database, EGFR from the Cancer Cell Map and Integrin Signaling (GRB2) from GeneGo.

In Table 3 we have listed all the proteins identified in EGFR1 signaling pathway as well as oncogenes (including those only observed with significant levels of transcript) in order of the ratio of SUM149/190 RNA-Seq values. This approach allows us to take advantage of the much greater dynamic range for RNA-Seq vs proteomics to compare differences between the two cell lines. We then compared these ratios with the proteomics data obtained for these two cell lines. The control cell line SKBR3 expresses high levels of ERBB2 transcript (300) and lower levels of EGFR (1.4) and shows proteomic values that are mostly intermediate between SUM149 and 190. In Table 3 we highlighted in yellow the proteins with higher expression in SUM149 and in

Table 4. Integrin Outside-In Signaling^a

Gene	SKBR3 ^b	SUM190 ^b	SUM149 ^b	SKBR3 ^c	SUM190 ^c	SUM149 ^c	Ratio 149/190 ^d	GRB2 interact	Novoseek tomor hits
FN1 ^e	0	0	1	1.3	0.1	111.9	6.6	+	159
FLNA ^e	420	5	524	7.0	12.6	414.3	4.9	+	9
ACTN1 ^e	75	54	154	8.9	8.6	81.9	3.1	+	5
<i>SRC</i> ^f				7.2	0.1	7.5	3.0	+	397
ACTN4 ^e	158	52	103	73.7	11.4	71.2	2.5	+	14
ACTB	499	545	476	1083.8	196.1	559.5	1.5	+	3
VCL ^e	100	5	17	8.6	3.1	8.9	1.3		28
ITGB1 ^e	0	0	7	7.6	12.9	32.1	1.3	+	71
TLN1 ^e	72	0	46	5.3	5.5	12.9	1.1		0
CTNNA1	0	12	12	8.4	21.9	49.0	1.1	+	1681
ITGA2 ^e	0	0	6	0.2	2.4	3.7	0.4		28
<i>GRB2</i> ^f				10.2	7.7	10.0	0.3	+	35
VTN	2	0	0	2.5	0.0	0.1	0.1		102
ITGAV	2	0	0	1.3	7.1	7.9	0.1		23
ACTR3	28	16	17	15.6	20.4	20.8	0.0	+	5
MAP2K1 ^e	0	2	0	3.1	10.3	7.8	-0.4	+	45
RAC1	15	14	12	190.8	23.0	12.0	-0.9	+	88

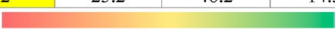
More 190  More 149

^aThis pathway was retrieved from GeneGo in January, 2012. ^bSpectral counts. ^cRPKM values. ^dRPKM values are used to show the expression differences in two IBC cell lines, and the values in the ratio column are calculated as follows: Ratio (149/190) = $\log_2((\text{RPKM SUM149} + 1)/(\text{RPKM SUM190} + 1))$.

^eProteins with higher expression in SUM149 are highlighted in yellow and in blue those with higher levels in SUM190. ^fKnown oncogene.

Table 5. Validated Targets of C-MYC Transcriptional Activation (A Subpathway of c-MYC Pathway)

Gene	SKBR3 ^a	SUM190 ^a	SUM149 ^a	SKBR3 ^b	SUM190 ^b	SUM149 ^b	Ratio 149/190 ^c	MYC interact	Novoseek tomor hits
RCC1	2			0.1	0.0	26.8	4.8	+	0
HMGA1 ^d			1	35.6	3.5	119.0	4.7	+	99
BCAT1 ^d			4	0.0	0.0	11.1	3.6	+	0
CAD ^d	43		18	2.5	3.2	22.0	2.5		2
RUVBL2 ^d	2	2	4	73.0	10.4	53.9	2.3	+	0
NCL ^d	60	45	142	118.1	14.9	70.0	2.2	+	21
GAPDH	522	495	539	2966.0	205.2	800.7	2.0	+	88
BAX ^d	2		2	12.8	2.8	13.3	1.9	+	848
RUVBL1	2	7	10	13.2	6.3	26.3	1.9	+	0
CCNB1 ^d			1	17.3	8.8	32.3	1.8	+	138
PFKM	5			7.0	5.5	19.3	1.6	+	0
EIF4A1	49	96	77	128.0	21.6	61.8	1.5	+	2
ENO1	758	1073	661	481.8	169.7	464.3	1.4	+	49
TRRAP ^d	2		1	1.2	3.3	10.7	1.4	+	1
LDHA ^d	179	115	215	213.6	19.2	48.2	1.3	+	27
TK1	6	5	3	120.2	11.0	28.3	1.3	+	67
SLC2A1	5			22.1	7.7	19.7	1.3	+	451
HUWE1 ^d	15		37	3.3	16.1	36.0	1.1	+	6
HSPD1	216	117	107	73.6	14.5	31.8	1.1	+	85
PRDX3	31	7	7	23.6	14.6	31.3	1.1	+	1
ACTL6A ^d			4	12.1	12.6	24.1	0.9	+	0
EIF4G1 ^d	25	12	45	35.8	37.8	60.9	0.7	+	16
NME1	29	23	20	0.1	26.4	36.0	0.4		732
PTMA ^d	6		3	550.7	77.5	88.0	0.2		18
EIF2S1	6	9	6	5.9	13.1	12.9	0.0	+	14
DDX18 ^d	2	2	4	6.9	6.7	6.5	0.0	+	5
HSPA4	26	31	36	8.0	23.0	21.6	-0.1	+	385
HSP90-AA1	260	326	287	254.6	58.8	42.3	-0.5	+	485
EIF4E	3			7.5	3.1	1.8	-0.5	+	181
MTDH ^d		5		13.6	18.4	9.7	-0.9	+	25
TFRC ^d	41	21	12	23.2	48.2	14.3	-1.7	+	0

More 190  More 149


^aSpectral counts. ^bRPKM values. ^cRPKM values are used to show the expression differences in two IBC cell lines, and the values in the ratio column are calculated as followed: Ratio (149/190) = $\log_2((\text{RPKM SUM149} + 1)/(\text{RPKM SUM190} + 1))$. ^dProteins with higher expression in SUM149 are highlighted in yellow and in blue those with higher levels in SUM190.

blue those with higher levels in SUM190. In general there was good agreement between RNA-Seq and proteomic values, e.g., CAV1, PLEC for higher ratios of EGFR vs ERBB2 and GRB7, CRKL and CTNND1 for higher ratios of ERBB2 vs EGFR. CRKL has been shown to associate with lamellipodia formation in breast carcinoma,³⁷ and coactivation of CRKL and estrogen

receptor alpha has been shown to be a promoter of tumorigenesis.³⁸ These observations are supported by literature reports, such as CTNND1 was genomically correlated to breast cancer and cell proliferation in ERBB2 positive breast cancer cell lines.^{39–41} The overexpression of caveolin-1 (CAV1) is frequently related to breast cancer⁴² and has been reported to be

Table 6. p53 Pathway (A Subpathway of Class I PI3K Signaling Events Mediated by Akt)

Gene	SKBR3 ^a	SUM190 ^a	SUM149 ^a	SKBR3 ^b	SUM190 ^b	SUM149 ^b	Ratio 149/190 ^c	PTEN interact	Novoseek tomor hits
S100A2 ^d			14	1.6	0.4	53.8	5.3		66
EGFR ^d	12		22	1.4	0.6	60.1	5.2		6585
CAV1 ^d			4	0.4	0.2	33.6	4.9	+	434
TP53 ^e				14.8	1.3	40.6	4.2	+	24003
SERP-INB5 ^d			14	0.0	0.4	23.5	4.2		482
SFN ^d	6	23	88	15.2	2.8	64.2	4.1		25
PRMT1 ^d			2	58.8	7.4	94.2	3.5		10
NDRG1	18			50.6	0.9	19.7	3.5		43
PYCARD ^d			3	0.1	0.0	9.5	3.4		21
GPX1 ^d			1	0.1	13.5	132.2	3.2		12
EPHA2 ^d			4	2.1	6.0	51.4	2.9		208
PRKCD	3			39.6	0.1	5.3	2.5		19
CSE1L ^d	66	21	95	19.9	10.1	59.8	2.5		7
BAX ^d	2		2	12.8	2.8	13.3	1.9		848
TRIM28 ^d	12	9	5	42.8	18.1	69.7	1.9		1
HTT ^d			1	0.5	1.9	9.0	1.8		4
SMAR-CA4 ^d	5		2	21.7	8.4	31.4	1.8		61
CCNB1 ^d			1	17.3	8.8	32.3	1.8		138
MSH2 ^d	2		3	2.8	2.9	10.6	1.6		338
PCNA	25	16	20	47.2	22.6	63.8	1.5		2016
TRRAP ^d	2		1	1.2	3.3	10.7	1.4		1
BID ^d			2	2.5	1.4	5.4	1.4		24
PRMT5 ^d			3	6.8	6.9	19.5	1.4		5
HUWE1 ^d	15		37	3.3	16.1	36.0	1.1		6
CTSD ^d	52	31	5	346.5	34.9	68.6	1.0		430
MLH1 ^e				3.5	3.6	6.3	0.7		1123
MET ^e				0.8	5.6	9.3	0.7		379
RPL5	34	35	29	658.1	98.9	145.3	0.6		0
ATR ^d			1	0.6	3.8	5.3	0.4		10
NEDD8 ^d	9		11	100.7	15.1	19.7	0.4		4
USP7	2			8.3	9.2	8.9	0.0		13
RB1 ^e				1.2	2.0	1.7	-0.1		1484
PPP2CA ^d		5		6.4	22.0	15.1	-0.5		0
DDX5	6	5	8	32.2	40.2	19.8	-1.0		6
APC ^e				0.2	3.4	1.2	-1.0		573
PTEN ^e				0.8	6.1	0.8	-1.9		2252

More 190  More 149

^aSpectral counts. ^bRPKM values. ^cRPKM values are used to show the expression differences in two IBC cell lines, and the values in the ratio column are calculated as followed: Ratio (149/190) = $\log_2((\text{RPKM SUM149} + 1)/(\text{RPKM SUM190} + 1))$. ^dProteins with higher expression in SUM149 are highlighted in yellow and in blue those with higher levels in SUM190. ^eKnown oncogene.

associated with EGFR activation.⁴³ Interestingly, the overexpression of both CAV1 and CAV2 has been discovered in triple negative (TN) invasive breast cancer.⁴⁴ In our study, SUM149 is only the TN cell line, and CAV1 is only identified by proteomics in this cell line and a RPKM value (33.6) that is much higher than for two ERBB2+ cell lines, i.e., SUM190 (0.2) and SKBR3 (0.4). At the other extreme of Table 3, higher levels of ERBB2 transcript are associated with the proteomic measurement of GRB7, growth factor receptor-bound protein 7, which is part of the ERBB2 amplicon in breast cancer.⁴⁵ In addition most of the proteins in Table 3 have been reported to interact with EGFR (30/34) and had literature associations with cancer (27/34).

Table 4 shows a similar analysis of the Integrin outside-in signaling pathway, which was selected as an example of the oncogene GRB2, and shows an elevation of filamin A (FN1), actinin alpha1 (ACTN1) in both the transcriptome and proteome of SUM149 vs 190 cell lines. Of interest, Filamin A phosphorylation has been shown to mediate the effects of caveolin-1 on cancer cell migration.⁴⁶ For the c-MYC pathway (Table 5) the higher ratios of EGFR transcript were associated with increased proteomic levels of branched chain amino-acid transaminase 1 (BCAT1), cytosolic, carbamoyl-phosphate synthetase 2 (CAD) and nucleolin (NCL), while higher ERBB2 ratios are associated with transferrin receptor (TFRC) and metadherin (MTDH).

Examples of the significance of these proteins include the observation that nucleolin colocalizes with BRCA1 in breast carcinoma tissue,⁴⁷ and metadherin is a valuable marker of breast cancer progression, and high expression may play a role in tumorigenesis of breast cancer.^{48,49} As was observed for the EGFR pathway, most of the proteins in Table 4 (GRB2) and Table 5 (MYC) contained a significant number of interactors (12/17 and 28/31) and literature associations with cancer (16/17 and 24/31) respectively.

A similar analysis of the p53 pathway is shown in Table 6. This pathway is a subpathway of Class I PI3K signaling events mediated by Akt and was selected as an example of the oncogene PTEN (phosphatase and tensin homologue). Tumor suppressor PTEN has been observed to be deleted in TN breast cancer, which shown related to resistance of EGFR targeting therapy.⁵⁰ In our data set, SUM149, which is a classic TN breast cancer, has a very low level transcript expression of PTEN (0.8), compared to SUM190 (6.1). Interestingly, another tumor suppressor, SERPINB5 (Serpine B5), which has been reported to be negatively correlated with both ER and PGR genes in a quantitative DNA analysis,⁵¹ was only observed in SUM149 (proteomics and transcriptomics), which is the only TN cell line in the study. Likewise, amplification of S100A2 (Protein S100-A2) was observed in both proteomics and transcriptomics experiments.

This protein, as one of S100 families, has been reported to be upregulated in mRNA expression in ER-negative breast cancer patients and potentially promote cancer metastasis.⁵² SFN (14-3-3 protein sigma), which acts as p53-regulated inhibitor of G2/M progression, has been reported to be silencing due to DNA hypermethylation in breast cancer.^{53,54} A similar silencing due to methylation for PYCARD (or TMS1) has been observed in breast cancer cells.⁵⁵ However, both overexpression of SFN and PYCARD in transcript and proteomic level was detected in SUM149, which could provide a potential diagnostic marker for TN breast cancer. Similarly, EPH2, which overexpresses in more than 60% of breast cancer patients,⁵⁶ has been listed as potential clinical target in TN breast cancer.⁴⁴ Expression of EPH2 has been observed to be stimulated by the activation of EGFR.⁵⁷ This is consistent with the EPH2 expression in our experiment, in which EPH2 was only identified in SUM149 (proteomics) and greatly amplified in transcriptomic level.

CONCLUSION

In view of the importance of EGFR/ERBB2 heterodimer signaling in breast cancer, it is of interest to explore the transcriptomic and proteomic analysis of two primary cell lines isolated from inflammatory breast cancer patients, one (SUM149) that expresses high levels of EGFR transcript with much lower levels of ERBB2 (1/4), while the other expresses very high levels of ERBB2 transcript (SUM190) and no detectable EGFR transcript. As a control we used a SKBR3 cell line that expressed high levels of ERBB2 transcript and low levels of EGFR. Analysis of the transcript levels indicated that the most likely signaling pathway for SUM190 involved the ERBB2/ERBB3 heterodimer, while SUM149 had several possibilities with involvement of EGFR dimers, ERBB2 heterodimers with EGFR and ERBB2 or ERBB3. We then explored the proteome of the two cell lines in terms of correlations between the transcriptome and proteomic measurements, identification of a panel of 21 oncogenes expressed in the two cell lines, interaction analysis of the observed proteins with this panel of oncogenes and selection of relevant cancer pathways. The analysis resulted in 4 pathways in addition to ERBB2 signaling (EGFR, integrin, MYC signaling, and PI3K signaling, see Tables 4–6) that contained many of the oncogene interacting proteins. In general there was reasonable agreement between the RNA-Seq and proteomic values shown in these tables except for some house-keeping proteins (see Discussion section). We list here those proteins that were correlated with higher levels of EGFR or ERBB2 transcript, respectively. **EGFR signaling:** caveolin 1 (CAV1), plectin (PLEC) (EGFR); growth factor receptor-bound protein 7 (GRB7), Crk-like protein (CRKL) and Catenin delta-1 (CTNND1) (ERBB2). **Integrin signaling:** filamin A (FLNA) and actinin alpha1 (ACTN1) (EGFR). **MYC signaling:** branched chain amino-acid transaminase 1 (BCAT1), carbamoyl-phosphate synthetase (CAD), nucleolin (NCL) (EGFR); transferrin receptor (TFRC), metadherin (MTDH) (ERBB2). **p53 signaling:** S100-A2 protein (S100A2), caveolin 1 (CAV1), Serpin B5 (SERPINB5), stratifin (SFN), PYD and CARD domain containing (PYCARD), and EPH receptor A2 (EPHA2) (EGFR). While the depth of the proteomic analysis was limited partly because of technical issues with analysis of the primary cell lines, this study was designed to use proteomics to identify higher level protein expressions that correlated with the transcriptome study. In this study we have demonstrated that one of the goals of the chromosome-centric human proteome project (C-HPP), which is to integrate RNA-Seq with proteomics measurement, is

of value. We plan in a future study to explore the potential of the proteins identified in this study as markers of ERBB2 and EGFR signaling as well as activation of the oncogenes MYC and GRB2 in a study of breast cancer tumor samples.

ASSOCIATED CONTENT

Supporting Information

Supplemental figures and tables. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: wi.hancock@neu.edu. Tel: 617-373-4881. Fax: 617-373-8795.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by following research grants: (Korea) The World Class University program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10086-0 (W.S.H. and Y.-K.P.)), National Project for the Personalized Genomic Medicine A111218-11-CP01 (to Y.-K.P.) from the Korean Ministry of Health and Welfare; (USA) The National Institutes of Health Grants, U01-CA128427 to W.S.H., U54DA021519, UL1 RR024986, RM-08-029, and U54ES017885 to G.S.O.; NIH grant (M.P.S. and H.L.); Texas State Rider for the Morgan Welch Inflammatory Breast Cancer Program and the G. Morris Dorrance Jr. Chair in Medical Oncology (M.C., Z.M.).

REFERENCES

- (1) Mortazavi, A.; Williams, B. A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **2008**, *5* (7), 621–8.
- (2) Yang, Z.; Harris, L. E.; Palmer-Toy, D. E.; Hancock, W. S. Multilectin affinity chromatography for characterization of multiple glycoprotein biomarker candidates in serum from breast cancer patients. *Clin. Chem.* **2006**, *52* (10), 1897–905.
- (3) Zeng, Z.; Hincapie, M.; Haab, B. B.; Hanash, S.; Pitteri, S. J.; Kluck, S.; Hogan, J. M.; Kennedy, J.; Hancock, W. S. The development of an integrated platform to identify breast cancer glycoproteome changes in human serum. *J. Chromatogr., A* **2010**, *1217* (19), 3307–15.
- (4) Zeng, Z.; Hincapie, M.; Pitteri, S. J.; Hanash, S.; Schalkwijk, J.; Hogan, J. M.; Wang, H.; Hancock, W. S. A proteomics platform combining depletion, multi-lectin affinity chromatography (M-LAC), and isoelectric focusing to study the breast cancer proteome. *Anal. Chem.* **2011**, *83* (12), 4845–54.
- (5) Forozan, F.; Veldman, R.; Ammerman, C.; Parsa, N.; Kallioniemi, A.; Kallioniemi, O.-P.; Ethier, S. Molecular cytogenetic analysis of 11 new breast cancer cell lines. *Br. J. Cancer* **1999**, *81* (8), 7.
- (6) Liang, Z.; Zeng, X.; Gao, J.; Wu, S.; Wang, P.; Shi, X.; Zhang, J.; Li, T. Analysis of EGFR, HER2, and TOP2A gene status and chromosomal polysomy in gastric adenocarcinoma from Chinese patients. *BMC Cancer* **2008**, *8* (363), 12.
- (7) Mendelsohn, J.; Baselga, J. The EGF receptor family as targets for cancer therapy. *Oncogene* **2000**, *19* (56), 16.
- (8) Ross, J. S.; Fletcher, J. A. The HER-2/neu oncogene: prognostic factor, predictive factor and target for therapy. *Semin. Cancer Biol.* **1999**, *9* (2), 125–38.
- (9) Kauraniemi, P.; Hautaniemi, S.; Autio, R.; Astola, J.; Monni, O.; Elkhouloun, A.; Kallioniemi, A. Effects of Herceptin treatment on global gene expression patterns in HER2-amplified and nonamplified breast cancer cell lines. *Oncogene* **2004**, *23* (4), 1010–3.

- (10) Wu, S.-L.; Kim, J.; Bandle, R. W.; Liotta, L.; Petricoin, E.; Karger, B. L. Dynamic profiling of the post-translational modifications and interaction partners of epidermal growth factor receptor signaling after stimulation by epidermal growth factor using extended range proteomic analysis (ERPA). *Mol. Cell. Proteomics* **2006**, *5* (9), 18.
- (11) Ahn, E. R.; Vogel, C. L. Dual HER2-targeted approaches in HER2-positive breast cancer. *Breast Cancer Res. Treat.* **2012**, *131* (2), 371–83.
- (12) Harris, L.; Fritsche, H.; Menne, R.; Norton, L.; Ravdin, P.; Sheila, T.; Somerfield, M. R.; Hayes, D. F.; Bast, R. C., Jr American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J. Clin. Oncol.* **2007**, *25* (33), 5287–5312.
- (13) Choi, H.; Fermin, D.; Nesvizhskii, A. I. Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* **2008**, *7* (12), 2373–85.
- (14) Ferguson, R. E.; Carroll, H. P.; Harris, A.; Maher, E. R.; Selby, P. J.; Banks, R. E. Housekeeping proteins: A preliminary study illustrating some limitations as useful references in protein expression studies. *Proteomics* **2005**, *5*, 6.
- (15) Zhang, L.; Li, W.-H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* **2004**, *21* (2), 4.
- (16) Chen, R.; Mias, G. I.; Li-Pook-Than, J.; Jiang, L.; Lam, H. Y.; Miriami, E.; Karczewski, K. J.; Hariharan, M.; Dewey, F. E.; Cheng, Y.; Clark, M. J.; Im, H.; Habegger, L.; Balasubramanian, S.; O'Huallachain, M.; Dudley, J. T.; Hillenmeyer, S.; Haraksingh, R.; Sharon, D.; Euskirchen, G.; Lacroute, P.; Bettinger, K.; Boyle, A. P.; Kasowski, M.; Grubert, F.; Seki, S.; Garcia, M.; Whirl-Carrillo, M.; Gallardo, M.; Blasco, M. A.; Greenberg, P. L.; Snyder, P.; Klein, T. E.; Altman, R. B.; Butte, A. J.; Ashley, E. A.; Gerstein, M.; Nadeau, K. C.; Tang, H.; Snyder, M. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **2012**, *148* (6), 1293–307.
- (17) Orazine, C. I.; Hincapie, M.; Hancock, W. S.; Hattersley, M.; Hanke, J. H. A proteomic analysis of the plasma glycoproteins of a MCF-7 mouse xenograft: a model system for the detection of tumor markers. *J. Proteome Res.* **2008**, *7* (4), 1542–54.
- (18) Wu, S. L.; Hancock, W. S.; Goodrich, G. G.; Kunitake, S. T. An approach to the proteomic analysis of a breast cancer cell line (SKBR-3). *Proteomics* **2003**, *3* (6), 1037–46.
- (19) Wu, S. L.; Taylor, A. D.; Lu, Q.; Hanash, S. M.; Im, H.; Snyder, M.; Hancock, W. S. Identification of potential glycan cancer markers with sialic acid attached to sialic acid and up-regulated fucosylated galactose structures in epidermal growth factor receptor secreted from A431 cell line. *Mol. Cell. Proteomics* **2013**, *12* (5), 1239–49.
- (20) Paik, Y. K.; Hancock, W. S. Uniting ENCODE with genome-wide proteomics. *Nat. Biotechnol.* **2012**, *30* (11), 1065–7.
- (21) Greenbaum, D.; Colangelo, C.; Williams, K.; Gerstei, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **2003**, *4* (9), 117.
- (22) Lev, D.; Kim, L.; Melnikova, V.; Ruiz, M.; Ananthaswamy, H.; Price, J. Dual blockade of EGFR and ERK1/2 phosphorylation potentiates growth inhibition of breast cancer cells. *Br. J. Cancer* **2004**, *91* (4), 8.
- (23) Zhang, C.-C.; Rogalski, J. C.; Evans, D. M.; Klockenbusch, C.; Beavis, R. C.; Kast, J. In silico protein interaction analysis using the global proteome machine database. *J. Proteome Res.* **2011**, *10* (2), 13.
- (24) Bignell, G. R.; Santarius, T.; Pole, J. C.; Butler, A. P.; Perry, J.; Pleasance, E.; Greenman, C.; Menzies, A.; Taylor, S.; Edkins, S.; Campbell, P.; Quail, M.; Plumb, B.; Matthews, L.; McLay, K.; Edwards, P. A.; Rogers, J.; Wooster, R.; Futreal, P. A.; Stratton, M. R. Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.* **2007**, *17* (9), 1296–303.
- (25) Ethier, S. P. Identifying and validating causal genetic alterations in human breast cancer. *Breast Cancer Res. Treat.* **2003**, *78* (3), 285–7.
- (26) Johnson, G. R.; Kannan, B.; Shoyab, M.; Stromberg, K. Amphiregulin induces tyrosine phosphorylation of the epidermal growth factor receptor and p185erbB2. Evidence that amphiregulin acts exclusively through the epidermal growth factor receptor at the surface of human epithelial cells. *J. Biol. Chem.* **1993**, *268* (4), 2924–31.
- (27) Franklin, M. C.; Carey, K. D.; Vajdos, F. F.; Leahy, D. J.; de Vos, A. M.; Sliwkowski, M. X. Insights into ErbB signaling from the structure of the ErbB2-pertuzumab complex. *Cancer Cell* **2004**, *5* (4), 317–28.
- (28) Hudelist, G.; Singer, C. F.; Manavi, M.; Pischinger, K.; Kubista, E.; Czerwenka, K. Co-expression of ErbB-family members in human breast cancer: Her-2/neu is the preferred dimerization candidate in nodal-positive tumors. *Breast Cancer Res. Treat.* **2003**, *80* (3), 353–61.
- (29) Holbro, T.; Beerli, R. R.; Maurer, F.; Koziczak, M.; Barbas, C. F., 3rd; Hynes, N. E. The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (15), 8933–8.
- (30) Eckstein, N.; Servan, K.; Girard, L.; Cai, D.; von Jonquieres, G.; Jaehde, U.; Kassack, M. U.; Gazdar, A. F.; Minna, J. D.; Royer, H. D. Epidermal growth factor receptor pathway analysis identifies amphiregulin as a key factor for cisplatin resistance of human breast cancer cells. *J. Biol. Chem.* **2008**, *283* (2), 739–50.
- (31) Komurasaki, T.; Toyoda, H.; Uchida, D.; Morimoto, S. Epiregulin binds to epidermal growth factor receptor and ErbB-4 and induces tyrosine phosphorylation of epidermal growth factor receptor, ErbB-2, ErbB-3 and ErbB-4. *Oncogene* **1997**, *15* (23), 2841–8.
- (32) Futreal, P. A.; Coin, L.; Marshall, M.; Down, T.; Hubbard, T.; Wooster, R.; Rahman, N.; Stratton, M. R. A census of human cancer genes. *Nat. Rev. Cancer* **2004**, *4*, 7.
- (33) Downward, J. Targeting RAS signalling pathways in cancer therapy. *Nat. Rev. Cancer* **2003**, *3* (1), 12.
- (34) Wu, C.; Zhu, J.; Zhang, X. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinf.* **2012**, *13*, 182.
- (35) Lessner, D. J.; Li, L.; Li, Q.; Rejtar, T.; Andreev, V. P.; Reichlen, M.; Hill, K.; Moran, J. J.; Karger, B. L.; Ferry, J. G. An unconventional pathway for reduction of CO₂ to methane in CO-grown Methanosarcina acetivorans revealed by proteomics. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (47), 17921–6.
- (36) Kauraniemi, P.; Barlund, M.; Monni, O.; Kallioniemi, A. New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. *Cancer Res.* **2001**, *61*, 6.
- (37) Lamorte, L.; Royal, I.; Naujokas, M.; Park, M. Crk adapter proteins promote an epithelial-mesenchymal-like transition and are required for HGF-mediated cell spreading and breakdown of epithelial adherens junctions. *Mol. Biol. Cell* **2002**, *13* (5), 1449–61.
- (38) Padmanabhan, R. A.; Nirmala, L.; Murali, M.; Laloraya, M. CrkL is a co-activator of estrogen receptor α that enhances tumorigenic potential in cancer. *Mol. Endocrinol.* **2011**, *25* (9), 1499–512.
- (39) Paredes, J.; Correia, A. L.; Ribeiro, A. S.; Milanezi, F.; Cameselle-Teijeiro, J.; Schmitt, F. C. Breast carcinomas that co-express E- and P-cadherin are associated with p120-catenin cytoplasmic localisation and poor patient survival. *J. Clin. Pathol.* **2008**, *61* (7), 856–62.
- (40) Paredes, J.; Correia, A. L.; Ribeiro, A. S.; Schmitt, F. Expression of p120-catenin isoforms correlates with genomic and transcriptional phenotype of breast cancer cell lines. *Cell. Oncol.* **2007**, *29* (6), 467–76.
- (41) Saijo, Y.; Perlaky, L.; Valdez, B. C.; Busch, R. K.; Henning, D.; Zhang, W. W.; Busch, H. The effect of antisense p120 construct on p120 expression and cell proliferation in human breast cancer MCF-7 cells. *Cancer Lett.* **1993**, *68* (2–3), 95–104.
- (42) Pinilla, S. M.; Honrado, E.; Hardisson, D.; Benitez, J.; Palacios, J. Caveolin-1 expression is associated with a basal-like phenotype in sporadic and hereditary breast cancer. *Breast Cancer Res. Treat.* **2006**, *99* (1), 85–90.
- (43) Agelaki, S.; Spiliotaki, M.; Markomanolaki, H.; Kallergi, G.; Mavroudis, D.; Georgoulas, V.; Stournaras, C. Caveolin-1 regulates EGFR signaling in MCF-7 breast cancer cells and enhances gefitinib-induced tumor cell inhibition. *Cancer Biol. Ther.* **2009**, *8* (15), 1470–7.
- (44) Tan, D. S.; Marchio, C.; Jones, R. L.; Savage, K.; Smith, I. E.; Dowsett, M.; Reis-Filho, J. S. Triple negative breast cancer: molecular profiling and prognostic impact in adjuvant anthracycline-treated patients. *Breast Cancer Res. Treat.* **2008**, *111* (1), 27–44.
- (45) Stein, D.; Wu, J.; Fuqua, S. A.; Roonprapunt, C.; Yajnik, V.; D'Eustachio, P.; Moskow, J. J.; Buchberg, A. M.; Osborne, C. K.; Margolis, B. The SH2 domain protein GRB-7 is co-amplified,

overexpressed and in a tight complex with HER2 in breast cancer. *EMBO J.* **1994**, *13* (6), 1331–40.

(46) Ravid, D.; Chuderland, D.; Landsman, L.; Lavie, Y.; Reich, R.; Liscovitch, M. Filamin A is a novel caveolin-1-dependent target in IGF-I-stimulated cancer cell migration. *Exp. Cell Res.* **2008**, *314* (15), 2762–73.

(47) Tulchin, N.; Chambon, M.; Juan, G.; Dikman, S.; Strauchen, J.; Ornstein, L.; Billack, B.; Woods, N. T.; Monteiro, A. N. BRCA1 protein and nucleolin colocalize in breast carcinoma tissue and cancer cell lines. *Am. J. Pathol.* **2010**, *176* (3), 1203–14.

(48) Li, J.; Yang, L.; Song, L.; Xiong, H.; Wang, L.; Yan, X.; Yuan, J.; Wu, J.; Li, M. Astrocyte elevated gene-1 is a proliferation promoter in breast cancer via suppressing transcriptional factor FOXO1. *Oncogene* **2009**, *28* (36), 3188–96.

(49) Li, J.; Zhang, N.; Song, L. B.; Liao, W. T.; Jiang, L. L.; Gong, L. Y.; Wu, J.; Yuan, J.; Zhang, H. Z.; Zeng, M. S.; Li, M. Astrocyte elevated gene-1 is a novel prognostic marker for breast cancer progression and overall patient survival. *Clin. Cancer Res.* **2008**, *14* (11), 3319–26.

(50) Marty, B.; Maire, V.; Gravier, E.; Rigall, G.; Vincent-Salomon, A.; Kappler, M.; Lebigot, I.; Djelti, F.; Tourdès, A.; Gestraud, P. Frequent PTEN genomic alterations and activated phosphatidylinositol 3-kinase pathway in basal-like breast cancer cells. *Breast Cancer Res.* **2008**, *10* (6), R101.

(51) Cheol Kim, D.; Thorat, M. A.; Lee, M. R.; Cho, S. H.; Vasiljević, N.; Scibior-Bentkowska, D.; Wu, K.; Ahmad, A. S.; Duffy, S.; Cuzick, J. M. Quantitative DNA methylation and recurrence of breast cancer: A study of 30 candidate genes. *Cancer Biomarkers* **2011**, *11* (2), 75–88.

(52) McKiernan, E.; McDermott, E. W.; Evoy, D.; Crown, J.; Duffy, M. J. The role of S100 genes in breast cancer progression. *Tumor Biol.* **2011**, *32* (3), 441–50.

(53) Ferguson, A. T.; Evron, E.; Umbricht, C. B.; Pandita, T. K.; Chan, T. A.; Hermeking, H.; Marks, J. R.; Lambers, A. R.; Futreal, P. A.; Stampfer, M. R.; Sukumar, S. High frequency of hypermethylation at the 14-3-3 sigma locus leads to gene silencing in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (11), 6049–54.

(54) Lodygin, D.; Hermeking, H. Epigenetic silencing of 14-3-3sigma in cancer. *Semin. Cancer Biol.* **2006**, *16* (3), 214–24.

(55) Levine, J. J.; Stimson-Crider, K. M.; Vertino, P. M. Effects of methylation on expression of TMS1/ASC in human breast cancer cells. *Oncogene* **2003**, *22* (22), 3475–88.

(56) Vaught, D.; Brantley-Sieders, D. M.; Chen, J. Eph receptors in breast cancer: roles in tumor promotion and tumor suppression. *Breast Cancer Res.* **2008**, *10* (6), 217.

(57) Larsen, A. B.; Pedersen, M. W.; Stockhausen, M. T.; Grandal, M. V.; van Deurs, B.; Poulsen, H. S. Activation of the EGFR gene target EphA2 inhibits epidermal growth factor-induced cancer cell motility. *Mol. Cancer Res.* **2007**, *5* (3), 283–93.