Integrated Proteomic, Transcriptomic, and Biological Network Analysis of Breast Carcinoma Reveals Molecular Features of Tumorigenesis and Clinical Relapse*

Marcin Imielski‡, Sangwon Cha§, Tomas Rejtar§, Elizabeth A. Richardson‡, Barry L. Karger§¶, and Dennis C. Sgroi‡

Gene and protein expression changes observed with tumorigenesis are often interpreted independently of each other and out of context of biological networks. To address these limitations, this study examined several approaches to integrate transcriptomic and proteomic data with known protein-protein and signaling interactions in estrogen receptor positive (ER+) breast cancer tumors. An approach that built networks from differentially expressed proteins and identified among them networks enriched in differentially expressed genes yielded the greatest success. This method identified a set of genes and proteins linking pathways of cellular stress response, cancer metabolism, and tumor microenvironment. The proposed network underscores several biologically intriguing events not previously studied in the context of ER+ breast cancer, including the overexpression of p38 mitogen-activated protein kinase and the overexpression of poly(ADP-ribose) polymerase 1. A gene-based expression signature biomarker built from this network was significantly predictive of clinical relapse in multiple independent cohorts of ER+ breast cancer patients, even after correcting for standard clinicopathological variables. The results of this study demonstrate the utility and power of an integrated quantitative proteomic, transcriptomic, and network analysis approach to discover robust and clinically meaningful molecular changes in tumors.

Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.014910, 1–15, 2012.

Breast cancer is a complex disease driven by multiple genomic, gene regulatory, proteomic, and metabolomic changes. In the post-genome era, the study of breast cancer has been partly driven by high throughput technologies that allow global profiling of “-omic” alterations in tumor tissue. These technologies enable unbiased searches for molecular events that drive tumor biology and inform prognosis and therapy.

The most widely applied of these technologies, transcriptomic profiling, has yielded multiple gene expression signatures associated with breast tumorigenesis, tumor subtype, and metastatic potential (1–3). Array comparative genomic hybridization (4) and single-nucleotide polymorphism microarrays (5) have been used to identify key somatic amplification and deletion events in breast cancer and guide molecular pathological classification. Ongoing whole genome and whole exome next generation sequencing studies are providing unbiased glimpses into the mutational landscape of breast cancer (6–8). A rapidly developing -omic technology that has shown promising application in breast cancer is MS-based proteomic analysis (9–16).

An important challenge, common to all -omic analyses of solid tumors, is the presence of nonmalignant or stromal cell admixture. Varying levels of nonmalignant cells can confound normal to tumor comparisons, distorting the effect of amplification/overexpression and deletion/underexpression of genes and proteins in the cancer cell population (17). This problem can be addressed by laser capture microdissection (LCM), a technology that allows for the molecular analysis of highly enriched populations of distinct cell types within a complex tissue sample (18, 19). LCM, however, yields a small number of cells (e.g. numbering in the ten thousands), the molecular interrogation of which requires highly sensitive an-

Want to cite this article? Please look on the last page for the proper citation format.
analytical methods. We have overcome these barriers in previous studies, producing both proteomic and transcriptomic characterization of breast cancer using LCM-enriched tumor and normal samples (14, 20).

Once the technical challenges to experimental data acquisition are overcome, -omic profiling data pose both opportunities and difficulties in their interpretation. Although these profiles can offer a comprehensive picture of a disease state relative to healthy tissue, their analysis often yields long lists of molecular changes that are difficult to interpret. This difficulty is compounded when the profiling data are obtained from multiple platforms (e.g. genomic, transcriptomic, and proteomic) that require additional integration of complex molecular signals. Novel bioinformatics techniques are required to overcome these challenges and realize the potential of -omic technologies to shed light on fundamental disease biology.

Having obtained in two previously published studies both proteomic and transcriptomic profiles of estrogen receptor positive (ER+) invasive breast carcinoma, a subtype that constitutes ~70% of all invasive human breast cancers diagnosed worldwide, we sought to integrate and interpret these data in a clinically and biologically meaningful manner (14, 20). We first investigated whether proteomic and transcriptomic profiles obtained from independent tumor versus normal purified breast epithelium sample sets yielded a concordant picture of breast cancer disease biology. We then examined three approaches for integrating these profiles with annotated biological networks to discover sets of genes and proteins associated with breast tumorigenesis (network discovery phase; Fig. 1, a–c). Finally, we investigated the ability of these gene sets to predict relapse in independent breast cancer clinical cohorts (prognostic validation phase; Fig. 1d).

In the network discovery phase of our analysis, we explored three related approaches to identify sets of functionally linked molecules with significantly different gene and protein expression in highly enriched malignant breast epithelium (MBE) from invasive breast carcinoma (IBC) versus normal breast epithelium (NBE). The first approach built networks from proteins that were significantly differentially expressed at both the transcriptomic and proteomic level (Fig. 1a). The second, "gene-centric" approach built networks from differentially expressed genes and identified those networks that were significantly enriched in differentially expressed proteins (Fig. 1b). The third, "protein-centric" approach built networks from differentially expressed proteins and scored these networks for their enrichment in differentially expressed genes (Fig. 1c).

**Fig. 1.** Panels a–c depict the three approaches we took to integrate our proteomic and transcriptomic differential expression data sets with annotated biological networks. a, Approach 1 applied network analysis on the list of significantly differentially expressed genes whose corresponding proteins were also differentially expressed in NBE and MBE. b and c, Approaches 2 (gene-centric NSEA) and 3 (protein-centric NSEA) built networks in either the transcriptomic or proteomic data set, respectively, and chose among these the subset enriched in differentially expressed proteins or genes, respectively. Black and white arrows in the histological tissue sections denote NBE and IBC epithelium, respectively. d, shows the clinical validation approach we used to test gene expression clinical classifiers based on networks generated by Approaches 1–3 using independent training and validation cohorts of ER+ breast cancer patients. The numbers alongside each test and training data set indicate the number of ER+ patients used in the analysis from each in each cohort.
In the prognostic validation phase of our analysis (Fig. 1d), the gene sets underlying significantly altered networks were subsequently trained and tested as prognostic classifiers in large sets of publicly available and clinically annotated ER+ breast cancer gene expression data. After permutation analysis, only the “protein-centric” integration approach produced a classifier that was significantly predictive of relapse in these independent cohorts. This classifier was predictive despite the fact that these clinical expression data were generated from unpurified, stromally contaminated tumor tissue collected from clinically heterogeneous breast cancer cases.

In summary, this study provides motivation and outlines a framework for the integrated analysis of protein and gene expression profiling data. The results suggest that the analysis of global molecular changes in small sample sets of microdissected primary tumor cells can uncover alterations relevant to large independent cohorts of complex tumor tissue samples. Finally, this study uncovers a novel molecular signature that yields potential insight into the biology of breast cancer tumorigenesis and relapse.

MATERIALS AND METHODS

Proteomic Specimen Acquisition and Processing—Proteomic data used in the network discovery phase were previously published (14). Briefly, NBE and MBE samples were obtained by LCM from each of nine different mammoplasty reduction specimens from women with a history of breast cancer and from each of nine different ER+ breast cancer specimens. Sample label permutation analysis of SpI was performed to identify differentially abundant proteins, using a nominal two-tailed test, as opposed to the use of limma (which was used in the original analysis of Ref. 25) to determine significance. Further details of sample acquisition and data processing for proteomic analysis can be found in Ref. 14.

Gene Expression Data—For the network discovery phase of our analysis, we used gene expression data that was previously published (20). Briefly, these data represent seven pairs of patient-matched LCM-purified MBE samples from IBC and NBE samples that were assayed on the Affymetrix U133XXP platform (see Table III for clinical characteristics); these samples are a subset of the patients reported in Ref. 20, corresponding to MBE samples and patient-matched NBE samples from the estrogen receptor positive (ER+) IBC specimens in the cohort. 13 of these 14 samples were obtained from patients not assayed in the proteomic analysis (the tumor sample from patient 153 was assayed in both transcriptomic and proteomic analyses). Additional details of sample acquisition and expression measurement can be found in Ref. 20. We renormalized the data in Ref. 20 from raw CEL files as described below.

For the prognostic validation phase of our analysis, we used five clinically annotated gene expression data sets (22–26). These data sets were culled from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), NCI caArray (https://array.nci.nih.gov/caarray/home.action), and the NKI website. Gene Expression Omnibus data sets used in the analysis were obtained via identifiers GSE2990 (22), GSE4922 (23), and GSE2034 (24) from the Gene Expression Omnibus website (http://www.ncbi.nlm.nih.gov/geo/). Expression data for the NKI295 breast cancer gene expression study (25) was obtained via the NKI website (http://bioinformatics.nki.nl/data.php). Expression data for Ref. 26 was obtained from the NCI caArray website (https://cabig.nci.nih.gov/tools/caArray) via identifier gray-00248. Only arrays that were directly collected and generated in the cited publications (22–26) were used, and no duplicate samples were noted to be present from both careful analysis of the data pedigree in the source publications and clustering analysis of expression profiles (not shown).

For each gene expression data set, the data were acquired as CEL files and reprocessed using the affy package in MATLAB (Mathworks Inc., Natick, MA) and the Bioconductor affy package (http://www.bioconductor.org/) in R using custom chip definition files obtained from the BrainArray website (http://brainarray.mbnli.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp) (27). Probe sets achieving fewer than 20% MAS5 present calls in any single data set (via the R function mas5calls) were removed from subsequent analyses. For the MBE versus NBE comparison of the data in Ref. 20, differential expression was determined using a t test with a q value threshold of 0.05 (nominal p < 0.0089) (Note that the q value for a gene is the false discovery rate that is implied by setting a threshold on the original analysis of Ref. 25 was assayed on an Agilent oligonucleotide array platform. All of the patient samples in these published clinical data sets were obtained from treatment-naive post-surgical bulk tumor (i.e. nonmicrodissected) samples, although a subset of the patients in all cohorts with the exception of (24) underwent post-operative (i.e. adjuvant) systemic therapy. The study identifiers, cohort of origin, array platform, and clinical covariate values of all clinical samples analyzed in the prognostic validation stage of our study are provided as supplemental data.

For each gene expression data set, the data were acquired as CEL files and reprocessed using the affy package in MATLAB (Mathworks Inc., Natick, MA) and the Bioconductor affy package (http://www.bioconductor.org/) in R using custom chip definition files obtained from the BrainArray website (http://brainarray.mbnli.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp) (27). Probe sets achieving fewer than 20% MAS5 present calls in any single data set (via the R function mas5calls) were removed from subsequent analyses. For the MBE versus NBE comparison of the data in Ref. 20, differential expression was determined using a t test with a q value threshold of 0.05 (nominal p < 0.0089) (Note that the q value for a gene is the false discovery rate that is implied by setting a significance threshold equal to the p value of that gene (28).) The choice of a t test, as opposed to the use of limma (which was used in the original analysis of Ref. 20 to quantify differential expression), was chosen to yield a simple test statistic to facilitate downstream set-based enrichment analyses.

Gene-Protein Set Concordance Enrichment Analysis—Gene set enrichment analysis (GSEA) was adapted to evaluate the concordance of differential gene and protein profiles of breast cancer data (29). Restricting the analysis to 1236 genes/proteins identified and quantitated on both proteomic and transcriptomic platforms, a gene/protein “concordance score” was computed by multiplying the t statistic (from the gene expression analysis) with SpI (from the proteomic analysis). This score reflected the magnitude and direction of the concordance (i.e. agreement) of observed gene and protein expression changes, with a positive score indicating concordance (gene up, protein up or gene down, protein down) and a negative score indicating discordance (gene down, protein up or gene up, protein down) of a gene/protein pair. The above genes/proteins were mapped to the MSigDB collection v.2.5 (http://www.broadinstitute.org/gsea/msigdb/index.jsp), and sets with fewer than 15 genes/proteins were removed. For each gene/protein set, an enrichment score (ES) was computed as a weighted running sum statistic over the concordance...
score. The ES was then recomputed for each gene set across 20,000 random shufflings of “tumor” (MBE) and “normal” (NBE) labels of samples in both proteomic and transcriptomic data sets and rescaled to yield a distribution of normalized enrichment scores (NESs). As in Ref. 29, the histogram of NES across all genes/protein sets and permutations was used to derive a background distribution for concordance enrichment, from which p and q values were computed. All of the computations were performed in MATLAB (Mathworks Inc.).

Network Set Enrichment Analysis (NSEA)—NSEA was developed to uncover biological pathways with both altered protein and gene expression in the tumor samples. Two versions of this approach were implemented (gene-centric and protein-centric NSEA), both employing Ingenuity Pathway Analysis (v8.8, content version 3204, http://www.ingenuity.com/) and a GSEA-like statistic to quantify enrichment (29). Gene-centric NSEA uses IPA to uncover biological networks significantly enriched in differentially expressed genes and chooses among these a subset of networks that is significantly enriched in differentially expressed proteins. A protein-centric version of NSEA employs IPA to uncover biological networks significantly enriched in differentially expressed proteins and chooses among these a subset of networks that is significantly enriched in differentially expressed genes. NSEA is closely related to GSEA, using Ingenuity network-derived gene and protein lists in place of manually curated gene sets and unsigned phenotype correlation statistics (Pearson $r^2$, $|S|p$) instead of the signal-to-noise or Pearson $r$ statistics used in GSEA (29). The NSEA algorithm is described in more detail below and is also depicted in a “road map” shown in supplemental Fig. 1. To explain gene-centric and protein-centric NSEA compactly, parentheses are used to denote the “reverse” analysis, i.e., genes to proteins (proteins to genes).

Significantly altered genes (proteins) from the transcriptomic (proteomic) analysis of LCM-enriched MBE versus NBE were inputted into IPA with standard settings (using all molecules, all direct/indirect relationships, network size of 35) to obtain an “observed” collection of enriched biological networks. A background collection of 50 (500) additional gene (protein) lists was obtained through sample label permutation of the transcriptomic (proteomic) data set, choosing genes (proteins) with $p < 0.0089$ ($|S|p > 0.556$) from each permutation. These permuted gene (protein) lists were also inputted into the Ingenuity Pathway Analysis tool to obtain a background collection of 500–1000 networks enriched in differentially expressed genes (proteins) from phenotype label-permuted data sets. The range of 500–1000 was chosen to be on the order of the size of the standard pathway collections used for GSEA (e.g., MSigDB).

Network collections derived from observed and permuted transcriptomic (proteomic) data were tested for proteomic (transcriptomic) enrichment against the LCM-enriched malignant breast epithelium versus normal epithelium protein (gene) expression data set using a weighted Kolmogorov-Smirnov-like ES for each network (29). In this step, $|S|p$ (Pearson $r^2$) values were computed for all proteins (transcripts) in the breast protein (gene) expression data set of correlation with NBE versus MBE phenotypic status. For each observed or background network, an ES was computed as the maximum of a weighted running sum statistic of ranked $|S|p$ ($r^2$) values corresponding to proteins (genes) encoding nodes in the given network. An unsigned score, i.e., $|S|p$ ($r^2$), was chosen as correlation metric to capture networks enriched in both up-regulated and down-regulated proteins (genes). This score was then recomputed across 20,000 permutations of tumor (MBE) and normal (NBE) sample labels in the protein (gene) expression data set to yield a background ES distribution for each set. ESs were Z-transformed across all phenotype label permutations for each network set to yield NESs. The histogram of NES across all network sets and phenotype label permutations was used to derive a null distribution for network set enrichment. Each observed network set was superimposed on this null distribution to obtain a one-tailed permutation $p$ value for proteomic (transcriptomic) enrichment. $p$ values were Bonferroni-corrected by a factor corresponding to the number of observed networks. Networks with $p < 0.05$ following the Bonferroni correction were determined to be significantly enriched in the proteomic (transcriptomic) data set. NSEA was restricted to gene (protein) sets that had more than five genes (proteins) measured in the transcriptomic (proteomic) data set, because small sets may distort the background distribution of NESs, as discussed in Ref. 29. All of the computations were performed in MATLAB (v7.9.0; Mathworks Inc.; http://www.mathworks.com).

Prognostic Classifier Validation—Networks enriched in significantly altered genes and proteins were tested for their ability to prognostically stratify the patients in data sets of Refs. 22–26 on the basis of gene expression. Each network was mapped to a set of genes on each expression array via the Entrez Gene ID and Hugo Symbols provided with the Ingenuity output file and the gene mappings bundled with the Affymetrix chip definition files (22–24, 26) or UniGene identifiers provided with the text output of (25). A support vector machine classifier with a linear kernel was trained on 5-year distant metastasis-free survival in ER+ patients from Ref. 26 and tested in ER+ patients from Refs. 23–26. Briefly, this classifier consisted of a threshold on a weighted sum of gene expression values; a patient whose score fell above or below the threshold was, respectively, assigned a “good” or “poor” prognostic category label by the classifier. The support vector machine algorithm determined the optimal weights and threshold for the classifier during the training step. Z-transformed gene expression data were used as input into the classifier in both training and test sets.

The efficacy of the classifier was then validated against disease-free (DFS) and relapse-free survival (RFS) measures (DFS in Refs. 23, 25 and 26 and RFS in Ref. 24). DFS, RFS, and distant metastasis-free survival are closely related survival measures that refer to the time elapsed between diagnosis and the reappearance of breast cancer (including a second primary tumor), local or distant breast cancer metastasis, or distant metastasis, respectively, in a patient. The estimated hazard ratios of high risk versus low risk groups in each cohort and in meta-analysis across cohorts were determined using Cox regression; the hazard ratio is a standard metric used in clinical trials to quantify the effect of a grouping variable (e.g., risk stratification) on the rate of events (e.g., disease relapse) in a group of patients over the period of the study. For cohorts having additional clinical data (i.e., lymph node status, ER status, age, tumor size, grade, and treatment), the independent prognostic ability of each classifier subject to these additional covariates using multivariate Cox regression was also determined. We also computed receiver operator characteristic curves for each classifier and each test-cohort by varying the classifier threshold from a value in which all patients received a “Poor” to a value in which all patients received a “Good” prognosis label. Classifier labels were compared with the 5-year survival measure (DFS in (23), RFS in (24), DFS in (25), DFS in (26)) to compute the sensitivity and specificity of the test for each value of the classifier threshold. Receiver operator characteristic curves were computed by plotting sensitivity versus 1-specificity, and the areas under the curve (AUCs) were computed and reported. AUCs are measures of the efficacy of a binary classifier and vary between 0.5 and 1. Permutation testing was applied to compare classifiers obtained from observed (i.e., network-derived) gene lists with random gene sets. For each observed gene list, random gene lists of the same size were chosen from the list of genes assayed on the training set microarray (22). Classifiers were trained on these random gene lists, and hazard ratios in the test cohorts were computed using Cox regression ($\pm$ clinical covariates). A background distribution of hazard ratios from these 1000 random gene lists was computed, and the
observed hazard ratios (either in meta-analysis or individual cohorts and either univariate or multivariate) were compared against the appropriate background distribution. All of the computations were performed in MATLAB (Mathworks Inc.).

RESULTS

Concordance of Differential Gene and Protein Expression in ER+ Invasive Breast Carcinoma—Global protein and gene expression were examined in two cohorts of laser capture microdissected ER+ breast cancer samples (Tables I–III). As noted above, both data sets have been previously published (14, 20); the proteomic data set was derived from LCM-enriched MBE from ER+ breast cancer patients and NBE from mammoplasty specimens of non-breast cancer patients (Tables I and II), whereas the transcriptomic data set consisted of patient-matched MBE and patient-matched NBE (Table III). Quality control filtering (see “Materials and Methods”) of gene expression data from seven ER+ IBC and patient-matched adjacent normal tissue yielded 9804 genes, of which 966 were differentially expressed between purified NBE and IBC cells (q < 0.05, Student’s t test). (Note that these results differ from those presented in Ref. 20 because we are analyzing only ER+ IBC and patient-matched adjacent normal tissue yield 9804 genes, of which 966 were differentially expressed between purified NBE and IBC cells (q < 0.05, Student’s t test).) Analysis of protein expression in nine ER+ MBE tumors and (non-patient matched) NBE samples yielded 1623 proteins, of which 121 were differentially expressed ([SpI] > 0.556).

A total of 1236 genes-protein pairs were detected on both transcriptomic and proteomic platforms in the two cohorts. The normal-tumor differential expression of these genes/proteins as measured by the \( t \) statistic and SpI, respectively, is plotted in Fig. 2a. Global concordance of differential gene and protein expression was modest but highly significant (\( r^2 = 0.19, p < 1 \times 10^{-12} \)). Moreover, there was a significant overlap between genes/proteins determined to be significantly differentially expressed on each platform (\( p < 5 \times 10^{-8}; \) Fisher’s exact test). Examining the direction of differential expression (i.e. up versus down in tumor) in 210 genes/proteins called significant in either platform showed striking concordance (\( p < 1 \times 10^{-12}; \) Fisher’s exact test; Table IV). Restricting the comparison with 28 genes/proteins called significant on both platforms showed 100% concordance in the direction of expression change (\( p = 1.21 \times 10^{-7}; \) Fisher’s exact test; Table IV). These 28 genes/proteins are shown in supplemental Table 1.

Finally, the pathway concordance of gene and protein differential expression was investigated using the MSigDB database and gene-protein set concordance enrichment analysis, analogous to GSEA. Briefly, the 1236 genes/proteins detected by both proteomic and transcriptomic platforms were mapped onto 1015 gene/protein sets of 15 members or more. Gene/protein sets were assigned an enrichment score permuting this enrichment score for original and phenotype label data sets yielded observed and null distributions of the NESs and \( p \) values for enrichment across gene/protein sets (see “Materials and Methods” for additional details). The vast majority of gene/protein sets (927 of 1015) demonstrated positive enrichment (NES \( > 0 \)) (tumor versus normal), with 310 showing significant positive enrichment (\( q < 0.25 \)) (Fig. 2b). These results suggest a significant degree of pathway level concordance in the differential gene/protein expression in invasive ER+ breast cancer. More broadly, both the single gene/protein and pathway level results strongly suggest that
proteomic and transcriptomic profiling approaches reflect consistent, although complementary, features of the same biological process. This concordance is notable in that it emerges from the analysis of nearly independent (31 of 32 samples) breast cancer sample sets with multiple sources of biological (tumor-normal, tumor-tumor, and gene-protein) and experimental (microarray and LC-MS) variation.

**Network Integration of Transcriptomic and Proteomic Results**—Three distinct strategies were examined to integrate transcriptomic and proteomic data with known networks to gain insight into the biology of invasive ER+/H11001 breast tumorigenesis (Fig. 1a). The first and simplest approach built IPA networks directly from genes/proteins that were significantly altered in both proteomic and transcriptomic data sets. The second approach (gene-centric NSEA) built IPA networks from significantly altered genes and then chose the subset of those networks significantly enriched in differentially expressed proteins (Fig. 1b). The third approach (protein-centric NSEA) built IPA networks from significantly altered proteins and then chose the subset of those networks significantly enriched in differentially expressed genes (Fig. 1c). Networks emerging from each approach were evaluated for their ability to prognostically stratify patients across multiple clinical breast cancer gene expression data sets (Fig. 1d). Underlying this strategy was the hypothesis that a network of genes/proteins contributing both to tumorigenesis and clinical relapse would reveal potentially important features of breast cancer biology.

**Approach 1: Build Networks from Both Differentially Expressed Genes and Proteins**—A list of 28 genes/proteins with significant expression changes in both transcriptomic (q < 0.05) and proteomic data sets (|SpI| > 0.556) (supplemental Table 1) was subjected to IPA core analysis with standard settings (see “Materials and Methods”), yielding two significant networks containing 15 and 13 focus molecules. Each of the two networks was examined for their ability to predict relapse using ER+ breast cancer patients from five independent cohorts (22–26) and their corresponding gene expression data sets. For each network, a support vector machine classifier was trained in one cohort (22) to predict 5-year survival (22). Significant survival time differences between the resulting proteomic and transcriptomic profiling approaches reflect consistent, although complementary, features of the same biological process. This concordance is notable in that it emerges from the analysis of nearly independent (31 of 32 samples) breast cancer sample sets with multiple sources of biological (tumor-normal, tumor-tumor, and gene-protein) and experimental (microarray and LC-MS) variation.

**Table III**

**Patient-matched MBE and NBE samples employed in the transcriptomic analysis**

Each row of the table represents a patient-matched tumor normal pair. The table was adapted from Ref. 14. N/A, not available; ND, not determined.

| Patient number | Age (years) | Grade | Estrogen receptor | Progesterone receptor | Her-2 | Size | Nodal status | Tumor type |
|----------------|-------------|-------|-------------------|-----------------------|-------|------|-------------|------------|
| 44             | 28          | III   | Positive          | Positive              |       | 1    | Negative    | Ductal     |
| 45             | 36          | I     | Positive          | Positive              |       | N/A  | Negative    | Ductal     |
| 79             | 54          | I     | Positive          | Positive              | 2.1   |      | Positive    | Ductal     |
| 102            | 55          | I     | Positive          | Negative              |       | 5.2  | Positive    | Ductal     |
| 121            | 45          | II    | Positive          | Positive              | 1.5   |      | Positive    | Ductal     |
| 148            | 42          | II    | Positive          | Negative              | 1.9   |      | Positive    | Ductal     |
| 153            | 46          | I     | Positive          | ND                    | N/A   |      | Positive    | Ductal     |

**Fig. 2.** Concordance of differential gene and protein expression in NBE and MBE epithelial samples. 

**a**, scatter plot of differential protein expression (SpI) and differential gene expression (t-statistic) for 1236 genes/proteins measured on both transcriptomic and proteomic platforms. Each point represents a gene/protein. Points are colored according to whether the corresponding gene/protein is significant in none, one, or both of the genomic and proteomic data sets. 

**b**, gene-protein set concordance analysis results for 1015 MsigDB gene/protein sets plotted as a histogram of observed normalized concordance enrichment score (black) against the background distribution obtained through shufflings of tumor and normal sample labels (red). NES scores for gene-protein sets showing significant concordance (q < 0.25, NES > 1.36) are highlighted in blue. See “Materials and Methods” for details of the analysis.
significant gene/protein enriched networks were tested for work 1; see supplemental Fig. 2 for network 2). These two significant networks. Protein-centric NSEA found that two pressed proteins were inputted into IPA, yielding seven because of this lack of observed proteomic concordance.

tomically derived networks as clinical relapse predictors for 28 hypotheses). We did not test any of the transcrip-
tic NSEA (see “Materials and Methods”) was applied to de-
gene list into IPA yielded 45 significant networks. Gene-cen-
matical derived networks were included in the enrichment analysis because of inadequate power,
i.e. less than five proteins in the network were represented in the proteomic data set. The remaining 28 networks contained be-
5 and 15 detected proteins, but none of these networks showed a significant enrichment when compared with the back-
ground NES distribution (p < 0.05, following Bonferonni correction for 28 hypotheses). We did not test any of the transcrip-
tomically derived networks as clinical relapse predictors because of this lack of observed proteomic concordance.

Approach 2: Gene-centric NSEA—666 of 9804 genes with significant differential gene expression between NBE and MBE (q < 0.05, Student’s t test) were identified. Input of this gene list into IPA yielded 45 significant networks. Gene-centric NSEA (see “Materials and Methods”) was applied to determine enrichment of these 45 networks in the proteomic data set. Interestingly, none of the transcriptomically derived networks yielded significant enrichment in differentially expressed proteins. Specifically, 17 of the 45 networks were not included in the enrichment analysis because of inadequate power, i.e. less than five proteins in the network were represented in the proteomic data set. The remaining 28 networks contained between 5 and 15 detected proteins, but none of these networks showed a significant enrichment when compared with the background NES distribution (p < 0.05, following Bonferonni correction for 28 hypotheses). We did not test any of the transcriptomically derived networks as clinical relapse predictors because of this lack of observed proteomic concordance.

Approach 3: Protein-centric NSEA—121 differentially expressed proteins were inputted into IPA, yielding seven significant networks. Protein-centric NSEA found that two of the seven protein-derived networks showed significant enrichment in the gene expression data set (p < 0.05, after Bonferonni correction for seven hypotheses) (Fig. 3 for network 1; see supplemental Fig. 2 for network 2). These two significant gene/protein enriched networks were tested for their ability to predict relapse across independent clinical gene-expression cohorts. As in Approach 1, an support vector machine classifier on network genes was trained on 5-year relapse-free survival in one ER+ breast cancer cohort (22), and then the performance of this network was tested across four independent test cohorts of ER+ patients applying Cox regression on DFS (23, 25, 26) or RFS (24). The classifier performance (quantified as hazard ratios) was compared between the network genes and classifiers created from 1000 random gene sets to yield a stringent permutation p value for predicting relapse versus relapse-free survival.

Although both of the above networks showed nominal significance in meta-analysis across the four test cohorts, only one network (Network 1; see Fig. 3) was significant following permutation (p = 0.004, permutation test). Stated differently, Network 1 outperformed 99.6% of random gene sets as a prognostic clinical classifier, yielding a nominal p value of 6.5 × 10−11 and a hazard ratio of 2.35 in meta-analysis across the four cohorts. Network 1 was also nominally significant in all four individual test cohorts (p = 0.002 in Ref. 23, p = 0.0003 in Ref. 24, p = 0.0004 in Ref. 25, and p = 0.002 in Ref. 26; Cox regression) (Table V). Kaplan-Meier curves illustrating the performance of the Network 1-based classifier on the four test cohorts and in meta-analysis are shown in Fig. 4. Receiver operator characteristic curves were also computed for this classifier across each of the cohorts (AUC = 0.61 in Ref. 23, AUC = 0.64 in Ref. 24, AUC = 0.67 in Ref. 25, and AUC = 0.66 in Ref. 26) and are provided in supplemental Fig. 3. Genes contributing to the Network 1-based classifier were: ACY, ASAHI1, BLVRA, COL14A1, CRABP2, DCN, FTH1, KRT15, MVP, MYH11, MYLK, NME1, and PARP1.

Multivariate Cox regression analysis revealed that the Network 1-based predictor was independent of histological grade,

### TABLE IV
Concordance in up versus down expression of genes/proteins detected and called significant on either or both proteomic and transcriptomic platform

Both show highly significant concordance (p < 1 × 10−12 and p = 1.21 × 10−7, respectively, Fisher’s exact test).

| Genes/proteins called significant on either proteomic or transcriptomic platform (210 in total) | Up | Down |
|---|---|---|
| Genes/proteins called significant on both proteomic and transcriptomic platforms (28 in total) | Up | Down |

| Network | P (Cox) | HR (Cox) | P (perm) | HR (perm) | P (Cox) | HR (Cox) | P (perm) | HR (perm) | P (Cox) | HR (Cox) | P (perm) | HR (perm) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.0E-04 | 2.18 | 0.13 | 0.01 | 4.20 | 0.10 | 0.17 | 1.36 | 0.67 | 0.22 | 1.34 | 0.46 | 8.3E-05 | 1.66 | 0.28 |
| 2 | 1.3E-03 | 2.17 | 0.13 | 0.37 | 1.64 | 0.66 | 0.21 | 1.38 | 0.66 | 0.94 | 0.98 | 0.95 | 0.01 | 1.45 | 0.66 |

### TABLE V
Cox regression results for two gene-protein networks arising from Approach 1

| Network | Ishina | Chin | Wang | Van de Vijver | Meta-analysis |
|---|---|---|---|---|---|
| | P (Cox) | HR (Cox) | P (perm) | HR (perm) | P (Cox) | HR (Cox) | P (perm) | HR (perm) | P (Cox) | HR (Cox) | P (perm) | HR (perm) |
| 1 | 8.0E-04 | 2.18 | 0.13 | 0.01 | 4.20 | 0.10 | 0.17 | 1.36 | 0.67 | 0.22 | 1.34 | 0.46 | 8.3E-05 | 1.66 | 0.28 |
| 2 | 1.3E-03 | 2.17 | 0.13 | 0.37 | 1.64 | 0.66 | 0.21 | 1.38 | 0.66 | 0.94 | 0.98 | 0.95 | 0.01 | 1.45 | 0.66 |
lymph node status, age at diagnosis, and tumor size in the data sets from Refs. 24 (p = 0.002, Cox regression) and 25 (p = 0.0026, Cox regression) and showed a trend for independence in the Ref. 23 data set (p = 0.06, Cox regression); clinical covariate data (beyond ER status) was not available for the data set in Ref. 24. Applying multivariate Cox-regression in a meta-analysis across the three cohorts yielded a robust p value of $5.7 \times 10^{-6}$ and a hazard ratio of 2.14 after correcting for the four clinical variables. Importantly, this result remained significant after comparison with 1000 random gene sets (p = 0.01, permutation). Multivariate Cox regression results for the Network 1 based classifier are shown in Table VII. These results were also unchanged when including systemic adjuvant chemotherapy as a fifth covariate (supplemental Table 2).

In summary, protein-centric NSEA yielded a network that was robustly informative of breast cancer relapse. Further-

Fig. 3. The single gene and protein network generated by Approach 3 (Network 1) that demonstrated significance as a clinical relapse predictor following a stringent permutation based test. Nodes correspond to proteins in the Ingenuity Pathway Database. a and b show the network with protein and gene expression overlaid, respectively. In each figure, red and green coloring represents proteins/genes that are up- and down-regulated, respectively, in the laser capture microdissected cancer cells relative to normal epithelium. Nodes in a and b are annotated with the spectral index and log fold change for that protein and gene, respectively.
more, this network-based expression biomarker was prognostically significant even after correcting for commonly used clinicopathological variables and comparison against 1000 random gene sets.

**DISCUSSION**

**Methodological Considerations**—We examined three approaches for integrating differential proteomic and transcriptomic results with annotated biological networks for ER+ invasive breast cancer-derived data sets. Approach 1 built a network from intersecting differentially expressed genes and proteins identified in the proteomic and transcriptomic analyses. Although simple, the approach ignored much of the transcriptomic data set, the majority of which mapped to proteins that were undetected on the proteomic platform. Furthermore, this approach only took into account the most significant changes in both data sets, which is likely to be conservative and insensitive. Approach 2 (gene-centric NSEA) built networks from significantly differentially expressed genes and evaluated those for enrichment with significantly altered proteins. This approach failed to yield any altered networks. There are several likely explanations for this result. First, the comparatively small number of proteins detected on the proteomic platform (versus gene expression) resulted in a minimal overlap between transcriptomically derived networks and proteins detected in the proteomic data set. Second, transcriptomic analysis yielded a large number of networks, many of which are likely driven by “biological noise” (i.e. false positives).
multivariate Cox regression was performed correcting for lymph node status, age, tumor size, and grade in each cohort and in meta-analysis. Age, tumor size, and grade covariate data were not available for the data in Ref. 24.

| Network | P (Cox) | HR (Cox) | P (perm) | P (Cox) | HR (Cox) | P (perm) | P (Cox) | HR (Cox) | P (perm) | P (Cox) | HR (Cox) | P (perm) | P (Cox) | HR (Cox) | P (perm) |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Network 1 | 0.060  | 1.60  | 0.31  | 0.0020  | 7.54  | 0.026  | 0.0026  | 2.37  | 0.041  | 5.7E-06  | 2.14  | 0.014  |

Guidelines for Evaluating Clinical Relapse Signatures—Part of the above analysis involved evaluation of significantly altered gene/protein networks for their ability to predict breast cancer relapse. An important caveat of evaluating gene sets and gene signatures for relapse is that it is quite possible to train a prognostically informative classifier (i.e. a biomarker(s) that predicts clinical outcome) from a random gene set. As has been previously observed in the gene expression literature, such a “random classifier” will even perform well on multiple validation data sets (37, 38). This baseline predictive ability may arise from the “footprint” of proliferative index or tumor grade on all genes in the genome (37–39). As a result, a clinical classifier built from a random gene set will encode this generic molecular feature and be somewhat predictive of prognosis, without necessarily reflecting novel tumor biology. Indeed, this caveat has been observed with many of the network-derived and random gene sets of the present study.
In prognostic validation phase of our study, this challenge was overcome through permutation, i.e. by ranking the prognostic ability of an observed gene set relative to a collection of randomly chosen gene sets. A gene set that performs no better than random on the test set is likely not encoding an interesting biological principle. For example, such was the case with both networks obtained by Approach 1. However, with Approach 3, a network was found that robustly surpassed this stringent permutation-based significance criterion. Furthermore, this network-based biomarker signature was significant even after correcting for commonly used clinicopathological variables used in determining breast cancer prognosis. As a result, this network very likely encodes a biologically significant feature of breast cancer molecular biology.

**Biological and Clinical Implications of Discovered Network**—Protein and gene expression changes superimposed on the prognostically significant network obtained via proteincentric NSEA (Fig. 3) suggest an overlap of three interrelated biological processes: 1) activation of stress response signaling, 2) altered small molecule metabolism, and 3) modulation of cytoskeletal and extracellular matrix components. These three themes comprise a host of interesting molecules and fuel many intriguing conjectures, a subset of which is discussed below.

Central to this network are the two canonical stress-activated protein kinases, JNK and p38 MAPK. These kinases mediate cellular response to environmental and genotoxic stresses, in particular, reactive oxygen species formation, while also playing important roles in inflammation, differentiation, survival, and cell migration (40). The impact of these kinases in cancer is complex- and context-specific. Thus, although certain cells employ the signaling pathways involving these kinases to inhibit growth and promote apoptosis, other cells employ JNK and p38MAPK to facilitate proliferation, survival, and invasion through activation of Ap1. This dual function mirrors the dual role of oxidative stress in cancer, in which tumor progression is promoted by inducing genomic instability, whereas apoptosis and other forms of cell death are simultaneously enhanced. Although phosphorylation activity of these two stress-activated protein kinases has not been measured, mild overexpression of JNK at the gene level and of p38 MAPK on both the protein and gene levels was observed in this study, as shown in Fig. 3.

In addition, consistent with JNK activation, dramatic down-regulation of ferritin heavy chain at the protein level was observed. Ferritin is the major iron storage protein in human cells, and its down-regulation causes reactive oxygen species accumulation and JNK activation through the increase of intracellular labile iron pools (41, 42). Conversely, JNK activation has also been shown to cause ferritin degradation, release of labile iron pools, and reactive oxygen species formation (43). Down-regulation of ferritin commonly occurs between early and late stages of esophageal carcinoma and is an essential event in transforming growth factor-β-induced epithelial to mesenchymal transition in cultured cancer cells (42). Interestingly, ferritin heavy chain is not found to be down-regulated at the transcript level. However, ferritin heavy chain is classically known to be post-transcriptionally regulated via a pathway involving iron response proteins IRP1 and IRP2. The analysis did not determine post-translational or transcription modifications, but either of these factors could be a possible explanation for this discordance (44).

Striking proteomic and transcriptomic up-regulation of heat shock protein HSBP1 (also known as Hsp27) was also observed. This finding is consistent with previous observations in which HSBP1 was found to be overexpressed in more than 50% of breast cancers and other epithelial tumors; heat shock protein has also been associated with multidrug resistance to chemotherapy (45–47). Similar to JNK and p38 MAPK, heat shock proteins are activated by cellular stress. In addition, HSBP1 is associated with the activation of JNK and inhibition of p38 MAPK in hematopoietic cells (48, 49) and is a substrate of the kinase MAPKAP-K2, which is directly phosphorylated by p38 MAPK (50).

HSBP1 prevents the cleavage of PARP1 (51), which was one of the most significantly altered proteins in the proteomic data set. PARP1 cleavage is a hallmark of programmed cell death in both normal and cancer cells (52). The enzyme mediates a key response to genotoxic stress by sensing double- and single-stranded breaks in DNA and rapidly recruiting a plethora of DNA repair proteins to the lesion site. PARP1 is also linked intricately to cancer cell metabolism, because its enzymatic activity causes significant depletion of cellular NAD+ and ATP, leading to a novel form of metabolic cell death termed parthananos (53). Most importantly, PARP1 has emerged as a centerpiece of breast cancer therapeutics, because PARP1 inhibitors have shown efficacy in the treatment of “triple negative” (i.e. ER, PR, and Her-2-Neu negative) breast cancers (52). Furthermore, recent overexpression and RNA interference knockdown experiments have implicated p38 MAPK as a major modulator of sensitivity to PARP1 inhibitors in breast cancer cell lines, providing an intriguing link between these two nodes in Network 1 (54, 55). PARP1 is significantly up-regulated across both the proteomic and gene expression data sets, which suggests an intriguing novel role for this molecule in the stress response of ER+ tumors.

Small molecule metabolism enzymes acid ceramidase, ATP citrate lyase (ACLY), and nucleotide diphosphate kinase 1 (NME1) were concordantly up-regulated in both the proteomic and transcriptomic data sets. Acid ceramidase, an enzyme that mediates the conversion of ceramide to sphingosine, links small molecule metabolism and stress response pathways in the network in Fig. 3. Ceramide accumulation provides a key molecular signal to coordinate the cellular response to serum deprivation, ionizing radiation, inflammatory cytokine stimulation, and other stresses (56). Notably, ceramide activates JNK and also causes programmed cell death through PARP1 cleavage (57, 58). Ceramide-mediated apo-
ptosis can be prevented through the overexpression of acid ceramidase, which metabolizes intracellular ceramide to sphingosine (59).

ACLY is a central metabolic enzyme that is essential for acetyl-CoA synthesis from citrate (60). The enzyme shuttles carbon moieties derived from glycolysis into de novo lipid synthesis, providing an important link between the Warburg effect and cancer anabolism (61). Pharmacological and genetic inhibition of ACLY suppresses tumor growth in multiple cell lines and xenografts (61). In addition to its established role in lipogenesis, it was shown that acetyl-CoA products derived from ACLY serve as the primary substrate for chromatin modifying histone acetylases, suggesting that ACLY may be involved in the epigenetic regulation of the differentiation state of tumor cells (62). Intriguingly, ACLY has been shown to undergo phosphorylation at a catalytic histidine residue by another protein in Network 1, the enzyme NME1 (63).

NME1 is a small molecule metabolism enzyme that was up-regulated in both the transcriptomic and proteomic data sets. NME1 catalyzes synthesis of non-ATP nucleotide diphosphates. NME1 overexpression is also associated with the inhibition of JNK and p38 MAPK, as well as broad inhibition of the proliferative kinases ERK, Akt, and PI3K (64, 65). NME1 is also known as a “metastasis suppressor,” because loss or underexpression of the protein is associated with high metastatic potential in breast cancer and other tumor types (65). This property may underlie its prognostic value in the network-based clinical predictor reported in this study.

The periphery of the proposed network contains multiple down-regulated extracellular matrix and cytoskeletal proteins that include decorin and type XIV collagen (COL14A1). Decorin is an extracellular matrix proteoglycan whose overexpression causes inhibition of the oncosgenic Erbb2 protein in breast carcinoma cell lines (66). Decorin is degraded by matrix metalloproteases, which are endopeptidases that cleave multiple components of the extracellular matrix and are ubiquitously up-regulated across human cancers, including breast cancers (67, 68). Matrix metalloproteinases are transcriptionally activated by JNK and p38 MAPK but are generally regulated through post-translational cleavage from an inactivezymogen state (40, 67). Furthermore, decorin binds to type XIV collagen, and this interaction is postulated to promote the organized connection of collagen fibrils in the extracellular matrix (69). In vitro-based data demonstrate that decorin suppresses breast metastases (70–72), whereas direct clinical data show that human breast tumors with reduced decorin expression are associated with poor outcome (73).

Taken together, Network 1 highlights an intriguing interplay of stress response, cancer metabolism, and tumor microenvironment pathways. The dual property of this network as being 1) proteomically and transcriptomically altered in the breast cancer versus normal epithelium analysis and 2) prognostically informative across a large clinical collection of breast cancer patients underscores the potentially important role of these pathways in the biology of tumorgenesis and metastasis in ER+ breast cancers.

The functional role of this network in breast cancer is not established by our study. The patterns discussed above represent hypotheses that would require significant additional experimentation to functionally validate. Because many of the central nodes in this network are involved in phosphosignaling (JNK and p38 MAPK), targeted and global kinase profiling of MBE and NBE would critically test these hypotheses. Additional functional insight could be gained through knockdown and overexpression experiments in human breast cancer cell lines or genetically engineered mouse models combined with histological and molecular phenotypic profiling following each perturbation. Such experiments are outside the scope of our present study but represent a potentially fruitful direction of future experimental efforts.

Conclusion—This study has demonstrated the power of an integrated quantitative proteomic, transcriptomic, and network analysis approach to discover robust and potentially meaningful clinical molecular changes in breast cancer. The work began by demonstrating a basic concordance between proteomic and transcriptomic expression changes measured in two studies of ER+ invasive ductal carcinoma patients. Leveraging this concordance, three novel methods were implanted to integrate proteomic and transcriptomic expression changes with biologically annotated networks, leading to the discovery of novel molecular features in breast cancer biology. The most successful approach constructed networks from significantly altered proteins and then chose among these networks significantly enriched in differentially expressed genes. Strikingly, it was found that a network enriched in differentially expressed genes and proteins in cancerous versus normal breast epithelium was informative of breast cancer relapse. This network linked pathways in cellular stress response, cancer metabolism, and tumor microenvironment through experimentally annotated signaling and binding interactions. The present results suggest physiologically important intersections of diverse cellular pathways that mediate both tumorgenesis and clinical aggressiveness in breast cancer. The findings have the potential to open new avenues of research into the molecular mechanisms of breast tumorgenesis and the development of novel therapeutics strategies for human breast cancer.

Acknowledgments—We thank the Enterprise Research IS group at Partners Healthcare for in-depth support and access to the HPC facilities. This work is Contribution 962 from the Barnett Institute.

* This work was supported by National Institutes of Health Grants R01GM15847 (to B. L. K.) and R01CA112021 (to D. C. S.) and funds from the Susan G. Komen Breast Cancer Foundation (to B. L. K. and D. C. S.), the NCI SPORE in breast cancer at Massachusetts General Hospital (to D. C. S.), and the Avon Foundation (to D. C. S.). The costs
of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

REFERENCES

1. Blièche, I., and Lidereau, R. (2011) Genome-based and transcriptome-based molecular classification of breast cancer. Curr. Opin. Oncol. 23, 93–99

2. Bombonati, A., and Sgroi, D. C. (2011) The molecular pathology of breast cancer progression. J. Pathol. 223, 307–317

3. Zhang, X. H., Wang, Q., Gerald, W., Hudis, C. A., Norton, L., Smid, M., Foekens, J. A., and Massague, J. (2009) Latent bone metastasis in breast cancer tied to Src-dependent survival signals. Cancer Cell 16, 67–78

4. Easton, D. F., Pooley, K. A. D., Arndt, A. M., Pharoah, P. D., Thompson, D., Easton, D. F., Pooley, K. A., Dunning, A. M., PHORO, P. D., Thompson, D., Easton, D. F., Pooley, K. A., Dunning, A. M., and Small, I. L. (2011) The molecular pathology of breast cancer: Understanding the molecular basis of histologic grade peritumoral interface zone. J. Mammary Gland Biol. Neoplasia 8, 340–348

5. Horlings, H. M., Lai, C., Nuyten, D. S., Halfwerk, H., Kristel, P., van Beers, J. M. D., Thompson, D. M., Jr, Ivanovich, J. L., Goodfellow, P. J., Perou, O. C., and Aparicio, S. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. J. Pathol. 216, 93–99

6. Bateman, N. W., Sun, M., Bhargava, R., Hood, B. L., Darfler, M. M., Bateman, N. W., Sun, M., Bhargava, R., Hood, B. L., Darfler, M. M., and Aparicio, S. (2003) A statistical integrated network analysis of breast carcinoma. J. Proteome Res. 6, 681–690

7. Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., Wallis, J. W., Harris, C. C., McLeinan, M. D., Fulton, R. S., Fulton, L. L., Abbott, R. M., and Pietenpol, J. A. (2010) Identification of differentially expressed proteins in triple-negative breast carcinomas using DIGE and mass spectrometry. J. Mass Spectrom. 45, 1197–1207

8. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

9. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

10. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

11. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

12. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

13. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

14. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

15. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

16. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

17. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

18. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

19. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

20. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

21. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

22. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

23. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

24. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

25. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

26. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

27. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

28. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

29. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

30. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

31. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

32. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

33. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

34. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

35. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

36. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

37. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

38. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

39. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

40. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

41. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

42. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

43. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

44. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

45. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.

46. To whom correspondence may be addressed. E-mail: b.karger@neu.edu.

47. To whom correspondence may be addressed. E-mail: dsgroi@partners.org.
24. Wagner, Y., Klijn, J. G., Zhang, Y., Siewerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Foekens, J. A. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679.

25. van de Vijver, M. J., He, Y. D., van’t Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009.

26. Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., et al. (2005) Genome mining of breast cancer: Toward a unified understanding of gene-expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1368–1377.

27. Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Aki, H., Watson, S. J., and Meng, F. (2005) Evolving gene/transcript definitions significantly alter the interpretation of gene-expression data. *Nucleic Acids Res.* **33**, e165.

28. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445.

29. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1544–1549.

30. Bateman, N. W., Sun, M., Hood, B. L., Flint, M. S., and Conrads, T. P. (2010) Defining central themes in breast cancer biology by differential proteinomics: Conserved regulation of cell spreading and focal adhesion kinase. *J. Proteome Res.* **9**, 5311–5324.

31. Zhu, W., Fang, C., Gramatikoff, K., Niemeyer, C. C., and Smith, J. W. (2011) Proteins and an inflammatory network expressed in colon tumors. *J. Proteome Res.* **10**, 2129–2139.

32. Szklarczyk, D., Fraganis, C. D., Kuhn, M., Simonovic, M., Roth, A., Minzge, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. (2011) The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D651–D658.

33. Elkon, R., de la Motte, A., Kruhlak, M., and Lin, M. H. (2005) Identification of gene expression patterns associated with breast cancer in the breast cancer molecular taxonomy. *Cancer Cell* **8**, 529–541.

34. Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002) Discovering biological networks hidden in “omics” data. *Annu. N.Y. Acad. Sci.* **9440–45.

35. Elkon, R., Vesterman, R., Amit, N., Ulitsky, I., Zohar, I., Weisz, M., Mass, G., Dreyfuss, G., and Bork, P. (2006) p38MAPK pathways in breast cancer development. *Neoplasia* **8**, 3168–3177.

36. Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. (2008) A modular hierarchical network model for the analysis of gene expression data. *Bioinformatics* **24**, 1545–1550.

37. Dittrich, M. T., Kuhr, G. L., and Albermann, K. A. (2008) Discovery of conserved gene-expression profiles in human tumors using self-organizing maps. *BMC Bioinformatics* **9**(Suppl 1) S233–S240.

38. Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18**, 1897–1905.

39. Garrido, C., Bruye, J. M., Fromentin, A., Hamman, A., Amigo, A. P., and Solal-Celigny, P. (1998) Identification of p53-independent and apoptosis-induced apoptosis through reduction of PKCα-mediated ROS production. *Oncoogene* **24**, 3715–3725.

40. Rouleau, M., Patel, A., Hendzel, M. J., Kaufmann, S. H., and Poirier, G. G. (2005) PARP inhibition: PARP1 and beyond. *Nat. Rev. Cancer* **10**, 293–301.

41. David, K. K., Andrabu, S. A., Dawson, T. M., and Dawson, V. L. (2009) Parthenos, a messenger of death. *Front. Biosci.* **14**, 1116–1128.

42. Dhanak, D., Hingorani, S. R., Tuveson, D. A., and Thompson, C. B. (2005) ATP-citrate lyase links cellular metabolism and resistance to radiation-induced apoptosis. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7558–1570.

43. Antosiewicz, J., Ziolkowski, W., Kaczor, J. J., and Herman-Antosiewicz, A. (2008) Identifying functional modules in protein-protein interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **36**, 5189–5200.

44. Wellen, K. E., Hatzivassiliou, G., Sachdeva, U. M., Bui, T. V., Cross, J. R., and Thompson, C. B. (2009) ATP-citrate lyase links cellular metabolism with tumor necrosis factor-induc
to histone acetylation. Science 324, 1076–1080
63. Wagner, P. D., Steeg, P. S., and Vu, N. D. (1997) Two-component kinase-like activity of nm23 correlates with its motility-suppressing activity. Proc. Natl. Acad. Sci. U.S.A. 94, 9000–9005
64. Boissan, M., De Wever, O., Lizaraga, F., Wendum, D., Poincloux, R., Chignard, N., Desbois-Mouthon, C., Dufour, S., Nawrocki-Raby, B., Birembaut, P., Bracke, M., Chavrier, P., Gespach, C., and Lacombe, M. L. (2010) Implication of metastasis suppressor NM23-H1 in maintaining adherens junctions and limiting the invasive potential of human cancer cells. Cancer Res. 70, 7710–7722
65. Otsuki, Y., Tanaka, M., Yoshii, S., Kawazoe, N., Nakaya, K., and Sugimura, H. (2001) Tumor metastasis suppressor nm23H1 regulates Rac1 GTPase by interaction with Tiam1. Proc. Natl. Acad. Sci. U.S.A. 98, 4385–4390
66. Santra, M., Eichstetter, I., and Iozzo, R. V. (2000) An anti-oncogenic role for decorin. Down-regulation of ErbB2 leads to growth suppression and cytodifferentiation of mammary carcinoma cells. J. Biol. Chem. 275, 35153–35161
67. Egeblad, M., and Werb, Z. (2002) New functions for the matrix metalloproteinases in cancer progression. Nat. Rev. Cancer 2, 161–174
68. Bissell, M. J., Radisky, D. C., Rizki, A., Weaver, V. M., and Petersen, O. W. (2002) The organizing principle: Microenvironmental influences in the normal and malignant breast. Differentiation 70, 537–546
69. Ehnis, T., Dieterich, W., Bauer, M., Kresse, H., and Schuppan, D. (1997) Localization of a binding site for the proteoglycan decorin on collagen XIV (undulin). J. Biol. Chem. 272, 20414–20419
70. Reed, C. C., Waterhouse, A., Kirby, S., Kay, P., Owens, R. T., McQuillan, D. J., and Iozzo, R. V. (2005) Decorin prevents metastatic spreading of breast cancer. Oncogene 24, 1104–1110
71. Araki, K., Wakahayashi, H., Shintani, K., Morikawa, J., Matsumine, A., Kusuzaki, K., Sudo, A., and Uchida, A. (2009) Decorin suppresses bone metastasis in a breast cancer cell line. Oncology 77, 92–99
72. Goldoni, S., Seidler, D. G., Heath, J., Fassan, M., Baffa, R., Thakur, M. L., Owens, R. T., McQuillan, D. J., and Iozzo, R. V. (2008) An antimetastatic role for decorin in breast cancer. Am. J. Pathol. 173, 844–855
73. Troup, S., Njue, C., Kiewer, E. V., Parisien, M., Roskelley, C., Chakravarti, S., Roughley, P. J., Murphy, L. C., and Watson, P. H. (2003) Reduced expression of the small leucine-rich proteoglycans, lumican, and decorin is associated with poor outcome in node-negative invasive breast cancer. Clin. Cancer Res. 9, 207–214

In order to cite this article properly, please include all of the following information: Imielinski, M., Cha, S., Rejtar, T., Richardson, E. A., Karger, B. L., and Sgroi, D. C. (2012) Integrated Proteomic, Transcriptomic, and Biological Network Analysis of Breast Carcinoma Reveals Molecular Features of Tumorigenesis and Clinical Relapse. Mol. Cell. Proteomics 11(6):M111.014910. DOI: 10.1074/mcp.M111.014910.