Laboratory Investigation (2008) 88, 591–601

The fibromatosis signature defines a robust stromal response in breast carcinoma

Andrew H Beck[1], Inigo Espinosa[1], C Blake Gilks[2], Matt van de Rijn[1] and Robert B West[1,3]

Breast cancer is a heterogeneous disease, and the influence of stromal gene and protein expression patterns on the biological and clinical heterogeneity of the disease is poorly understood. We previously demonstrated that evaluation of the gene expression patterns of two soft-tissue tumors (desmoid-type fibromatosis (DTF) and solitary fibrous tumor) could be used to identify distinct stromal reaction patterns in breast carcinoma. In the current study, we examined four additional data sets obtained from four different institutions and containing gene expression data from a total of 561 breast cancer patients. We identified a core set of 66 DTF-associated genes that were consistently coordinately expressed in a subset of 25–35% of breast cancers. Breast carcinomas defined by high levels of coordinated expression of DTF core genes tend to be lower grade, express estrogen receptor, and show significantly longer survival across the four data sets. Using multiple tissue microarrays of archival breast cancer specimens obtained from a total of 745 patients, we demonstrated that a subset of breast cancers show coordinate expression of DTF core proteins by stromal cells in the tumor microenvironment. We evaluated the protein expression of a single DTF core protein (SPARC) on a tissue microarray with clinical outcome data and demonstrated that breast cancers with strong stromal protein expression of SPARC show a trend for increased survival. Our data demonstrate that the DTF core gene set is a robust descriptor of a distinct stromal response that is associated with improved clinical outcome in breast cancer patients.

Laboratory Investigation (2008) **88,** 591–601; doi:10.1038/labinvest.2008.31; published online 14 April 2008

KEYWORDS: breast cancer; gene expression profiling; stroma; stromal signature; tumor–stroma interactions; tumor microenvironment

Epithelial carcinogenesis results, in part, from acquisition of genetic mutations in the epithelium, leading to initiation and progression of carcinomas. In the past decade, important epithelial cancer genes, and the pathways they control, have been identified, permitting development of therapies targeted at specific altered pathways.[1,2]

Carcinoma cells live in a complex microenvironment that includes a variety of non-epithelial cell types, including fibroblasts, smooth-muscle cells, cells composing the vasculature, and inflammatory cells, as well as diffusible growth factors and cytokines.[3–5] Carcinoma-associated stromal cells adopt an altered phenotype, characterized by increased proliferative activity and enhanced secretion of extracellular matrix proteins, serine proteases, matrix metalloproteinases, and growth factors.[6] The altered phenotype of the carcinoma-associated stromal cell is thought to allow it to function as a 'coconspirator' in cancer initiation and progression, and heterotypic signaling between epithelial tumor cells and stromal cells profoundly influences many steps of tumor progression.[3,7,8]

The expression profiles of carcinoma-associated stromal cells are only partially known, and it is likely that several currently unrecognized subtypes exist.[9,10] Most studies of tumor stroma consider it as a relatively uniform entity, and the contribution of inter-patient stromal variability to the biological and clinical heterogeneity of breast cancer is only beginning to be recognized and understood.[11–16]

We hypothesize that within a particular group of tumors (eg, breast carcinoma) there exist distinct types of stromal reaction patterns that affect tumor growth in different ways. We propose that fibroblastic soft-tissue tumors, which are thought to be clonal outgrowths derived from distinct subtypes or precursors of fibroblasts, and which yield relatively pure and consistent gene expression profiles, can function as discovery tools for identifying distinct fibroblastic reaction patterns.[17]

[1]Department of Pathology, Stanford University Medical Center, Stanford, CA, USA; [2]Genetic Pathology Evaluation Centre, Vancouver Hospital and Health Sciences Centre and BC Cancer Agency, Vancouver, BC, Canada and [3]Pathology and Laboratory Service, Palo Alto Veterans Affairs Health Care System, Palo Alto, CA, USA
Correspondence: Dr RB West, MD, PhD, Department of Pathology, Stanford University Medical Center, 300 Pasteur Drive, Stanford, CA 94305, USA.
E-mail: rbwest@stanford.edu

Received 21 January 2008; revised 22 February 2008; accepted 22 February 2008

In a proof-of-principle study, we used gene expression profiles from two different fibroblastic soft-tissue tumors, desmoid-type fibromatosis (DTF) and solitary fibrous tumor (SFT), to identify new subtypes of tumor stroma.[15] In comparing the SFT and DTF gene expression profiles, we found a remarkably large number of differentially expressed genes. One of the more striking differences was in the variation of genes involved in the fibrotic response and basement membrane synthesis. Many genes highly expressed in DTF are also expressed during scar formation, such as type-I and type-III collagen, and profibrotic signaling proteins such as connective tissue growth factor (CTGF) and transforming growth factor-β (TGF-β). In contrast, genes highly expressed in SFT include those typically found in epithelial-support stroma, including a number of genes associated with basement membrane function. We examined the expression patterns of DTF- and SFT-associated genes in a large publically available breast cancer data set from the Netherlands Cancer Institute (NKI),[18] and showed that the subset of breast cancers with high levels of coordinated expression of DTF-associated genes demonstrated significantly better outcome, and tumors identified by elevated levels of expression of SFT-associated genes had a much worse outcome.[15]

Due to the high dimensionality of gene array data, associations between elevated levels of expression for a group of genes with outcome should be interpreted with caution.[19] In fact, results from a number of expression profiling papers could not be reproduced on separate data sets,[20] and there is consensus in the biomedical community that demonstrating reproducibility of gene-expression signatures in independent data sets is essential for establishing their robustness and validity.[21,22]

In the current paper, we sought to establish the validity and robustness of the DTF expression signature in breast carcinoma. First, we evaluated the expression of DTF-associated genes in four independent breast cancer data sets and show that in each data set a subset of breast cancers demonstrate increased coordinate expression of DTF genes and these breast cancers show improved survival. Second, we identified the core subset of DTF genes that are consistently and coordinately expressed in breast cancer across a total of five data sets. Third, using breast cancer tissue microarrays (TMAs), we showed that these DTF core proteins tend to be coordinately expressed in breast cancer stroma. Fourth, we performed functional gene-set analysis using a variety of computational techniques to demonstrate that the DTF core gene set is highly enriched for genes known to function in the extracellular matrix, and that the DTF core gene set encodes proteins that are predicted to operate in common functional modules in the regulation of the extracellular matrix. Through integrated use of gene expression analysis, breast cancer TMAs, and functional gene set analysis, we have characterized a distinct and robust stromal response that is consistently seen in a subset of breast carcinomas and correlates with improved survival.

MATERIALS AND METHODS

The DTF/SFT Gene Set

The DTF/SFT gene set consists of 786 gene spots that were significantly differentially expressed between DTF ($n = 10$ cases) and SFT ($n = 13$ cases), as described by West *et al.*[15] The list contains 493 gene spots that showed increased expression in DTF and 293 gene spots that showed increased expression in SFT (false discovery rate = 0.13%).

Breast Cancer Data Sets

We searched for publicly available data sets containing gene expression data from tumor samples of invasive breast carcinoma with clinical follow-up documenting at least one of the following: disease-free survival, disease-specific survival, or overall survival. Data sets were excluded if clinical data were not available. Following application of these inclusion and exclusion criteria, we identified a total of four data sets (the Perreard data set,[23] GSE1379,[24] GSE1456,[25] and GSE3494[26]), which contain gene expression data on a total of 561 cases with clinical follow-up, in addition to the NKI data set[18] previously analyzed. In the current study, we used the NKI data set to contribute to development of the DTF core gene signature and to correlate the DTF core gene signature with clinicopathological features not evaluated in our initial publication. Detailed information on the five data sets can be found in the Supplementary Information.

Data Analysis

For all data sets, the expression data were downloaded and imported into the dChip 2006 software (http://biosun1.harvard.edu/complab/dchip/). Expression data were standardized gene-wise by subtracting the mean and dividing by the standard deviation of the expression values for each gene. Unsupervised hierarchical clustering was performed with the Cluster 3.0 software (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv), using uncentered Pearson correlation as the distance metric and average linkage clustering. The resulting heatmap and dendrogram were visualized with Java Treeview (http://jtreeview.sourceforge.net/).

Determination of DTF Core Gene and DTF-Like Case Clusters

To create an objective and reproducible rule to define the DTF core gene cluster and DTF-like case cluster in each data set, we defined the DTF core gene cluster as the cluster of genes in each data set that was composed of greater than 50 genes showing greater than 25% correlated expression. We defined the DTF-like case cluster as the cluster of cases showing high levels of expression of the DTF core gene cluster and exhibiting greater than 10% correlated expression. After applying these rules, we defined the 'DTF core gene set' as being composed of genes that are present on the microarray platform in at least two data sets and present in the DTF core gene cluster in either all data sets (if gene was present on platform in only 2–3 of data sets) or absent in a maximum of one data set (if gene was present on platform in 4–5 data sets).

Analysis of Clinicopathological Variables

In the data sets examined, the measured survival outcomes included disease-free survival (Perreard, GSE1379, GSE1456, combined $n = 294$), disease-specific survival (GSE1456, GSE3494, combined $n = 395$), and overall survival (Perreard, GSE1456, combined $n = 234$). The Kaplan–Meier estimate was used to compute survival curves, and log-rank *P*-value was computed to assess statistical significance. Cox proportional hazard analysis was performed to calculate hazard ratios. For association analysis, Pearson χ^2-test was used for nominal variables and the Kendall's tau-b for ordinal variables. To compare ordinal or ratio variables in two independent groups, either Mann–Whitney *U*-test or Student's *t*-test was performed. Statistical computation was performed using SPSS 15.0 for Windows.

Evaluation of DTF Core Protein Localization in Breast Cancer Tumor Microenvironment

To determine the patterns of DTF core protein expression in the breast cancer tumor microenvironment, we performed immunohistochemistry on three breast cancer TMAs (TA108, TA221, and the Vancouver General Hospital TMA) containing samples from a total of 745 cases of breast cancer. TA108 contains samples from 24 cases of breast carcinoma, as described by West *et al.*[15] The Vancouver General Hospital TMA contains samples from a cohort of 438 sequential cases of invasive breast carcinoma with median follow-up of 15.4 years, as described by Makretsov *et al.*[27] TA221 contains samples from 283 breast carcinomas obtained from Stanford University Medical Center. Immunohistochemical studies were performed for three DTF core proteins, SPARC, CSPG2, and AEBP1. To select DTF markers for evaluation by immunohistochemistry on breast cancer TMAs, we identified genes that were most consistently present in the DTF core cluster and showed the highest degree of correlated expression over all the data sets. Out of this list of candidates, we identified SPARC, CSPG2, and AEBP1 as three markers with commercially available antibodies that performed well in immunohistochemistry on formalin-fixed paraffin embedded tissue. Detailed information on the antibodies used, staining procedure, and scoring technique can be found in the Supplementary Information. The institutional review board approved these studies.

Functional Gene-Set Analysis

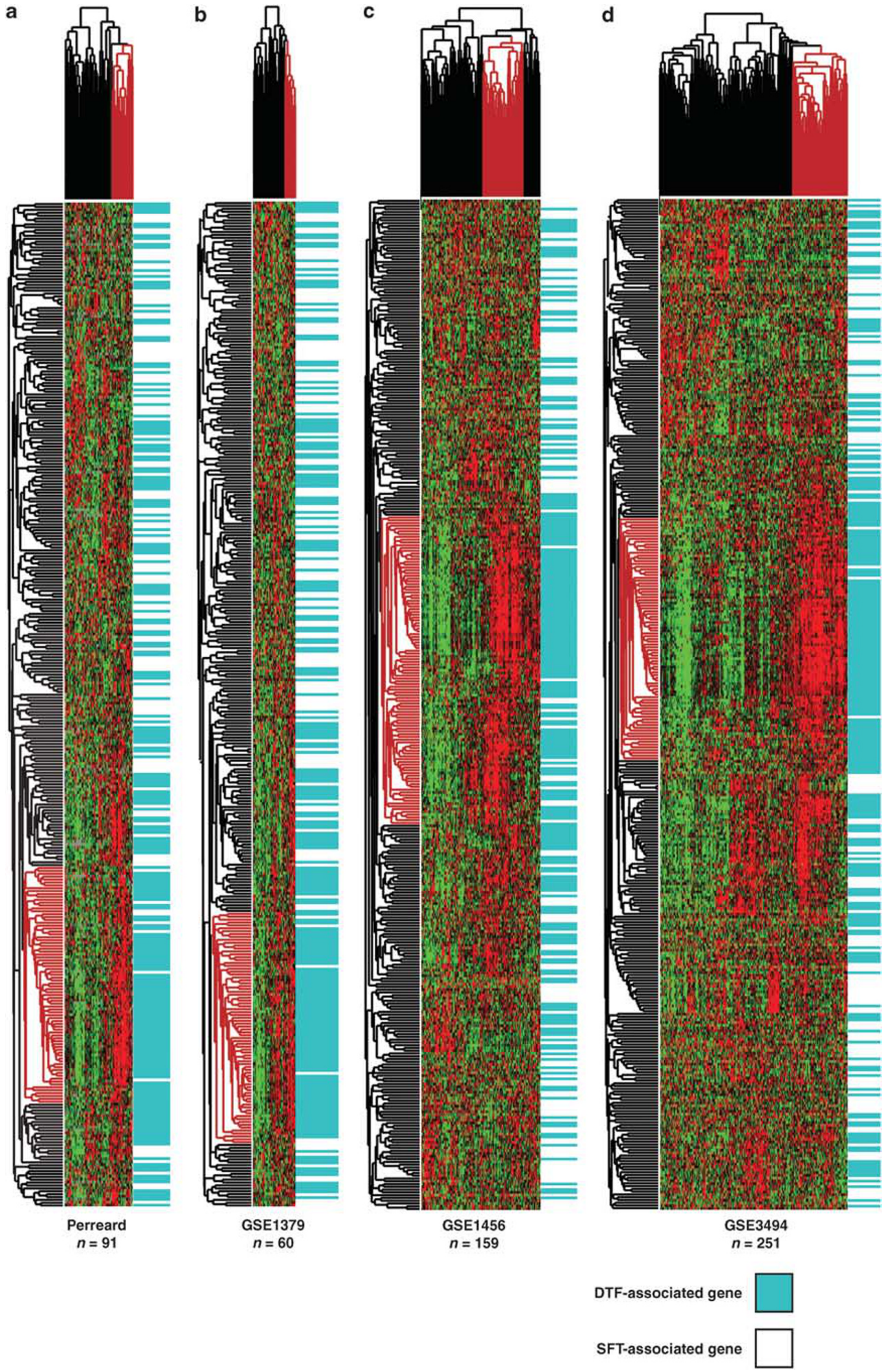
To determine the functional significance of gene sets, we used DAVID (Database for Annotation, Visualization, and Integrated Discovery), which integrates gene and protein annotation information from several databases.[28] To evaluate the properties of the protein–protein interaction (PPI) networks encoded by the DTF non-core and DTF core gene sets, we used STRING 7.0 resource, which is a searchable database of known and predicted PPIs.[29] To generate PPI networks, gene sets were uploaded into STRING and the following active prediction methods were employed: neighborhood, coexpression, gene fusion, experiments, co-occurrence, database, and text mining, with a medium confidence score (0.400). The Cytoscape software platform[30] was used to visualize the PPI networks, and the Cytoscape plug-in Network Analyzer 2.5 was used to evaluate the topological characteristics of the networks (http://med.bioinf.mpi-inf.mpg.de/netanalyzer/).

RESULTS

In our initial study, we observed that a large cluster of genes, which are highly expressed in the fibroblastic tumor DTF, showed elevated coordinated expression in a subset of breast cancers from the NKI data set. We now identify the core set of DTF genes that are consistently coordinately expressed in DTF-like cases of breast cancer, by evaluating the NKI data set and four additional independent data sets from separate institutions.

Identification of DTF Core Gene Set and DTF-like Breast Cancer Cases

Following unsupervised hierarchical clustering of the breast cancer cases with the DTF and SFT gene sets, in each data set we noted a tight cluster of DTF-associated genes that were coordinately expressed at high levels in a subset of breast cancers, representing the DTF core gene cluster and the DTF-like breast cancer case cluster (Figure 1a–d). To objectively identify the DTF core gene cluster in each data set, we chose the single cluster of genes containing greater than 50 genes and showing greater than 25% correlated expression. In each data set, only a single cluster of genes fulfilled this criterion, and this cluster was highly enriched for DTF genes (86–95% in each data set, 91% overall). The size of the DTF core gene cluster ranged from a minimum of 70 genes (23% of DTF/SFT gene set present on the NKI data set) to a maximum of 111 genes (30% of DTF/SFT gene set present on GSE1456 data set), and overall accounted for 25% of the DTF/SFT gene set present on the arrays. On the basis of these analyses, we defined the 'DTF core gene set' as being composed of the genes that are present in the DTF core gene cluster in all data sets (if the gene was present on the microarray platform in only 2–3 of data sets) or absent in only one data set (if gene was present on the platform in 4–5 data sets). The resulting DTF core gene set contains 66 genes (64 DTF-associated genes and 2 SFT-associated genes) that were consistently coordinately expressed in a subset of breast cancers across five independent data sets (Supplementary Workbook). We defined the DTF-like breast cancer case cluster as the cluster of cases that demonstrated greater than 10% correlation and increased levels of expression of the DTF core gene cluster. This technique identified a consistent subset of 25–35% of the breast cancer cases in each data set (Figure 1a–d). These


a
b
c
d
Perreard
n = 91
GSE1379
n = 60
GSE1456
n = 159
GSE3494
n = 251
DTF-associated gene
SFT-associated gene

findings confirm that across multiple independent breast cancer data sets, a quarter to one-third of breast cancers show high levels of coordinated expression of a consistent core subset of DTF genes.

When evaluating expression of the SFT genes in the breast carcinoma data sets by unsupervised hierarchical clustering, we did not see a cluster of SFT-associated genes that was as distinct as the DTF gene cluster. The SFT genes tended to show smaller clusters of coordinately expressed genes in small separate clusters of breast cancer cases (Figure 1a–d). The heterogeneous pattern of expression of SFT-associated genes in breast carcinoma made it difficult to identify SFT-enriched cases of breast cancer by unsupervised hierarchical clustering in an objective reproducible manner across data sets, as was achievable with the DTF gene signature. Consequently, in the current paper, we have focused exclusively on defining the core genes, protein localization, and biological and prognostic features of the DTF stromal signature.

Clinical Features of DTF-like Breast Carcinomas

The DTF-like cases of breast cancer tended to be lower grade than the non-DTF cases (31% grade 1, 47% grade 2, 22% grade 3 in DTF-like cases *vs* 19% grade 1, 42% grade 2, and 39% grade 3; Mann–Whitney $U = 53434.5$, $Z = -5.0$, $P = 0.0000006$). Patients with DTF-like tumors tended to be younger (52.1 years with DTF-like tumors *vs* 54.9 years; $P = 0.011$). The DTF-like tumors were more likely to express estrogen receptor (89% in DTF-like tumors *vs* 75%; $P = 0.00004$) and showed significantly improved outcomes in disease-free survival (75% at 10 years *vs* 53%; log-rank $P = 0.0004$; HR = 0.372, 95% CI = 0.210–0.657, Wald $P = 0.001$), disease-specific survival (82% at 10 years. *vs* 73%; log-rank $P = 0.049$; HR = 0.607, 95% CI = 0.367–1.00, Wald $P = 0.051$), and overall survival (83% at 8 years *vs* 65%; log-rank $P = 0.014$; HR = 0.432, 95% CI = 0.217–0.858, Wald $P = 0.017$) (Figure 2). The DTF-like tumors showed no significant difference in mean tumor size (22.1 mm for DTF-like tumors *vs* 22.8 mm; $P = 0.477$) or frequency of lymph node metastasis (43.7% for DTF-like tumors *vs* 42.5%; $P = 0.786$).

Molecular Subtype Characteristics of DTF-like Breast Carcinomas

The NKI, Perreard, and GSE1456 data sets have previously been stratified by others into molecular subcategories based on initial gene expression studies by Perou, Sorlie, and coworkers (basal, ERBB2 +, luminal A, luminal B, normal-like).[31,32] The cases were classified as 'N/A' if gene expression profile did not show sufficient correlation with one of the molecular subtypes. Combining information on molecular subtyping from the three different data sets showed the DTF core cases to be significantly less likely to be luminal B (7% in DTF-like cases *vs* 18%, $P = 0.001$) or basal (5% in DTF-like cases *vs* 24%, $P = 0.00000009$), and more likely to be normal-like (18% in DTF-like cases *vs* 5%, $P = 0.000001$) or N/A (35% in DTF-like cases *vs* 18%, $P = 0.00003$). There was no significant difference in the frequency of ERBB2 + or luminal A molecular subtypes between the two groups.

Chang *et al*[33] defined a 'wound response signature' based on gene expression profiling of fibroblasts exposed to serum. On the basis of the expression of 'core serum response genes,' the investigators defined two expression patterns ('quiescent' and 'activated') and demonstrated that breast cancer cases with the 'activated' expression pattern show worse prognosis across multiple different cancer types, including breast.[14,33] In the NKI data set, the DTF-like cases were significantly more likely to show a 'quiescent' pattern of gene expression (80% in DTF-like cases *vs* 47%; $P = 0.00000004$). There is little overlap of the DTF core gene set with the core serum response gene set (only three DTF core cluster genes (LUM, SDC1, and ID3) are present in the unique core serum response gene set (3 of 531 (0.6%)).

Van't Veer *et al* used a supervised analysis of gene expression data to define a '70-gene prognosis signature' that was a significant predictor of prognosis in the NKI breast cancer data set.[18,34] In the NKI data set, the DTF-like cases of breast cancer were significantly more likely to be classified as 'good prognosis' by the 70-gene signature (52% in DTF-like cases *vs* 33% for non-DTF core cases; $P = 0.003$). There is only 1 gene (WISP1) in common between the 70-gene prognosis signature and the DTF core gene list.

The GSE3494 data set contains information pertaining to the p53 mutation status of 251 cases of breast carcinoma. From this data set, Miller *et al*[26] defined a p53 expression signature and used diagonal linear discriminant analysis to classify cases according to the signature. Whereas DTF core cases showed no significant correlation with p53 mutation status, there was a negative correlation with the p53 mutation gene signature (19% of DTF-like cases enriched with signature *vs* 33%; $P = 0.033$), suggesting that high expression of the DTF core gene set is inversely correlated with a gene expression pattern seen in the setting of p53 mutations.

◀

Figure 1 Unsupervised hierarchical clustering of breast carcinomas with DTF- and SFT-associated genes. (**a**) Perreard ($n = 91$); (**b**) GSE1379 ($n = 60$); (**c**) GSE1456 ($n = 159$); and (**d**) GSE3494 ($n = 251$). Uncentered Pearson correlation was used as the distance metric with average linkage for the unsupervised hierarchical clustering. Within the heatmap, red represents high expression, black represents median expression, and green represents low expression. The sidebar on the right of each heatmap indicates whether a gene is associated with DTF (blue) or SFT (white). The red-highlighted region of the dendrogram to the left of the heatmaps indicates the DTF core gene cluster within each data set, and the red-highlighted region of the dendrogram above the heatmaps indicates the DTF-like breast cancer case cluster within each data set. The heatmap dendrograms of the gene expression of the DTF and SFT genes in the NKI data set was published previously.[6]

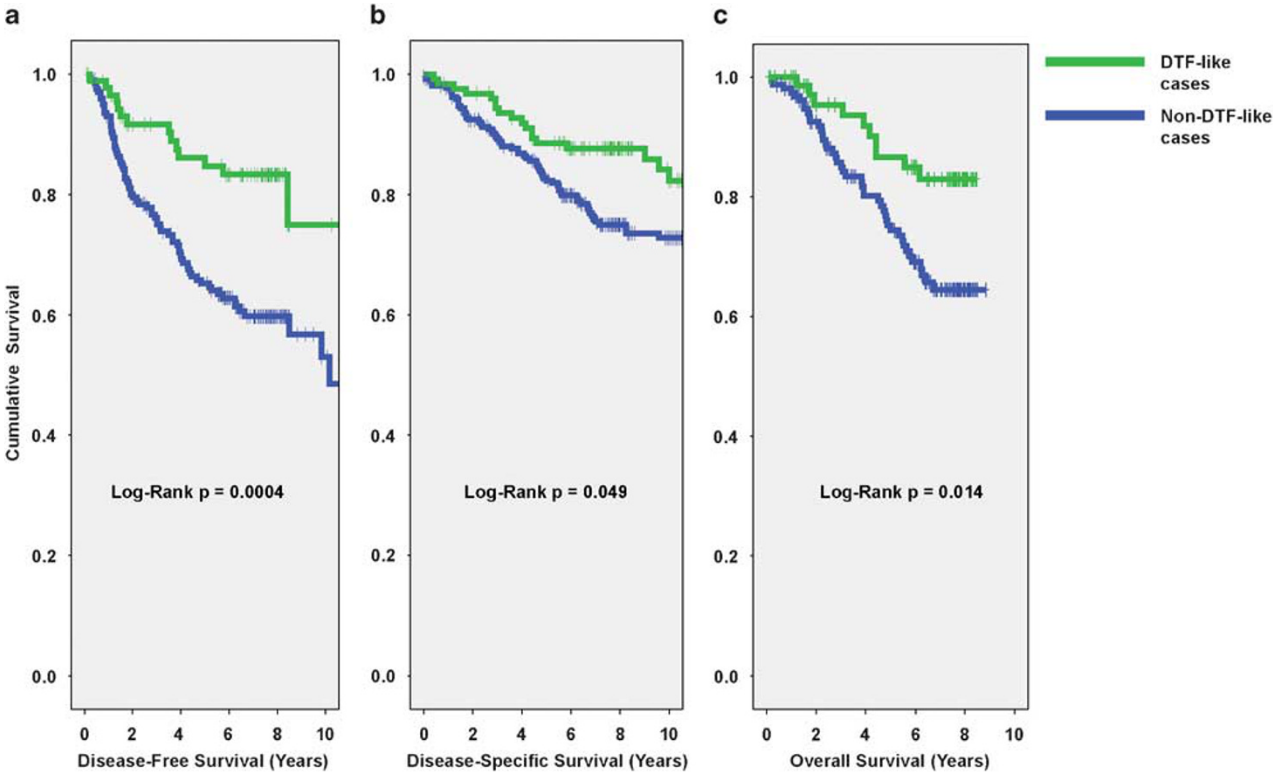


Figure 2 Kaplan–Meier survival curves for breast cancer cases stratified by expression of DTF-associated genes. In each plot, the cases are stratified according to their presence in the DTF-like cluster of breast cancer cases. The Kaplan–Meier survival curves plot (**a**) disease-free survival, (**b**) disease-specific survival, and (**c**) overall survival. Disease-free survival data were available for 294, disease-specific survival data for 395, and overall survival data were available for 234 patients. The Kaplan–Meier survival curves are compared by Cox–Mantel log-rank test to assess statistical significance. The *x*-axis indicates time from diagnosis and the unit of measurement is years. The *y*-axis indicates cumulative probability of survival.

Localization of DTF Core Proteins in the Breast Cancer Tumor Microenvironment: Coordinated Expression of DTF Core Proteins

To investigate the cellular localization of DTF core proteins in the breast cancer tumor microenvironment, we evaluated the protein expression of a selection of DTF core proteins on multiple breast cancer TMAs ($n = 745$). We first evaluated the stromal expression of three DTF core proteins (SPARC, CSPG2, AEBP1) by immunohistochemistry on TA108, which is a TMA that contains representative 2 mm cores from 24 breast cancers. Integrating these immunohistochemistry results with those from RNA *in situ* hybridization of a DTF core gene measured in our prior study (POSTN (OSF2)) demonstrated that 9 of 24 cases of breast carcinoma showed stromal expression of at least three of the four DTF core markers (Figure 3a). The DTF core proteins showed distinct patterns of expression within the breast cancer tumor microenvironment, with the CSPG2 (versican) labeling protein in the extracellular compartment throughout the tumor stroma (Figure 3b). This finding is consistent with CSPG2's role as a large extracellular matrix proteoglycan involved in cell adhesion, migration, proliferation, and ECM assembly.[35] In contrast to CSPG2's pan-stromal extracellular pattern of protein expression, SPARC expression was localized primarily to the cytoplasm of peri-tumoral fibroblasts and endothelial cells (Figure 3b). This finding is consistent with SPARC's known function as a matricellular protein expressed at high levels in many types of cancers by cells associated with tumor stroma and vasculature.[36] AEBP1, which is a transcriptional repressor whose expression is abolished during differentiation of pre-adipocytes into mature adipocytes,[37] showed staining of a combination of epithelium and peritumoral stroma (Figure 3b).

We examined the expression patterns of two DTF core proteins, SPARC and CSPG2, on a larger TMA containing 0.6 mm cores from a total of 283 breast carcinomas. For SPARC, of the interpretable cases 54 of 257 (21%) cases showed strong stromal staining, 115 of 257 (45%) showed weak stromal staining, and 88 of 257 (34%) showed no stromal staining. For CSPG2, of the interpretable cases 60 of 256 (23%) cases showed strong stromal staining, 108 of 256 (42%) showed weak stromal staining, and 88 of 256 (34%) showed no stromal staining. The SPARC and CSPG2 staining patterns showed significant degree of correlation (Kendall's tau-b = 0.350, $P = 2.76e-013$). SPARC and CSPG2 showed strongly discordant staining (strong *vs* negative) in only 11 of the 243 cases (4.5%) in which both stains were interpretable. Of the 243 cases with interpretable cores for both SPARC and

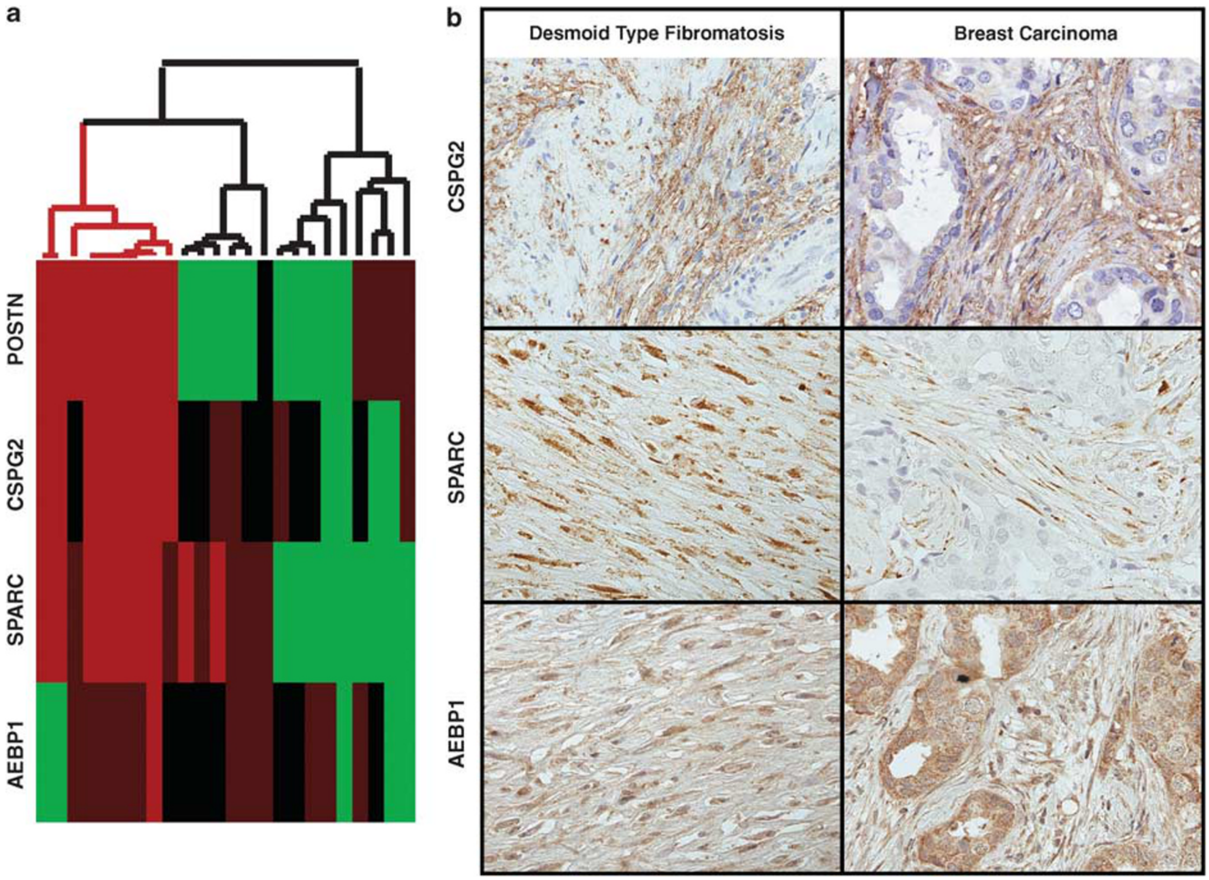


Figure 3 Coordinate expression of DTF core proteins in DTF and in the breast cancer tumor microenvironment. (**a**) Unsupervised hierarchical clustering of 24 breast carcinomas based on TMA staining with markers from the DTF core gene set: CSPG2, SPARC, AEBP1 (immunohistochemistry), and POSTN (*in situ* hybridization). Bright red represents strong expression, dull red weak expression, green no expression, and black represents no data. Uncentered Pearson correlation was used as the distance metric with average linkage for unsupervised hierarchical clustering. The cluster of cases showing increased levels of coordinated expression of DTF core markers is highlighted in red on the dendrogram above the heatmap. (**b**) Examples of DTF core protein expression (CSPG2, SPARC, AEBP1) measured by immunohistochemistry in DTF (left column) and breast carcinoma (right column). Magnification, × 400.

CSPG2, 24 of 243 cases (10%) showed coordinate strong stromal staining of the two proteins.

SPARC Protein Expression and Disease-Specific Survival

We evaluated the stromal expression of SPARC on the Vancouver General Hospital TMA, which contains representative cores from 438 cases of invasive breast cancer with median follow-up of 15.4 years, as described in Makretsov *et al.*[27] Of the 364 interpretable cases of unilateral invasive breast carcinoma, 80 (22%) showed strong diffuse stromal SPARC staining, 151 (42%) showed weak SPARC stromal staining, and 133 (37%) showed no staining. The 80 cases showing strong SPARC stromal staining showed a trend for increased disease-specific survival in univariate analysis (75% survival at 20 years *vs* 60% in SPARC weak or negative cases; log-rank $P = 0.085$; HR = 0.652, 95% CI = 0.399–1.065; Wald $P = 0.087$) (Figure 4). A Cox multivariate analysis incorporating lymph node status, tumor size, and SPARC staining showed no significant association of SPARC with disease-specific survival in the multivariate model (SPARC multivariate HR = 0.673, 95% CI = 0.38–1.19; $P = 0.175$).

The trend for improved disease-specific survival with strong SPARC expression in the univariate analysis is similar to that seen in the gene expression data for the DTF-like cases of breast cancer (82 *vs* 73% at 10 years; log-rank $P = 0.049$; HR = 0.607, 95% CI = 0.367–1.00, Wald $P = 0.051$), although statistical power is less in the TMA data, due, in part, to smaller sample size.

Biological Relationships in the DTF Core Gene Set

The DTF core gene set is highly enriched for genes encoding proteins involved in diverse aspects of extracellular matrix structure and function, including collagens, proteins involved in collagen binding, proteins involved in calcium ion binding, proteins involved in cell–cell adhesion, and proteins involved in cell-surface receptor-linked signal transduction regulation (Supplementary Workbook). KEGG pathways enriched in the DTF core gene set include extracellular matrix–receptor interaction, focal adhesion, and cell commu-

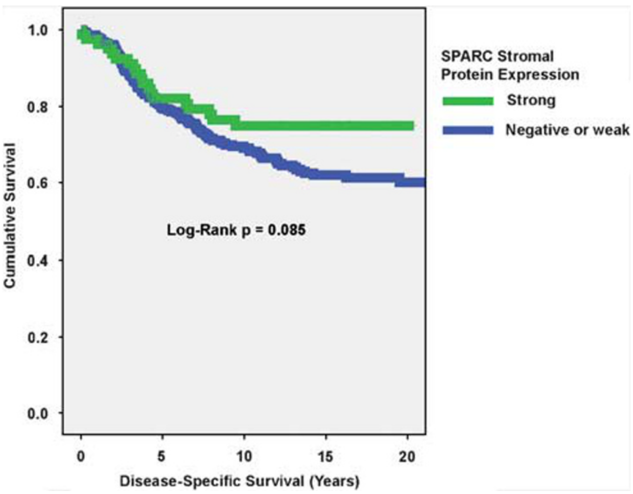


Figure 4 Kaplan–Meier survival curves for breast cancer cases stratified by SPARC protein expression. Survival curves display the disease-specific survival in breast cancer cases with strong stromal SPARC protein expression ($n=80$) as compared with cases with weak or no staining ($n=284$) measured by immunohistochemistry. The Kaplan–Meier survival curves are compared using Cox–Mantel log-rank test to assess statistical significance. The *x*-axis indicates time from diagnosis and the unit of measurement is years. The *y*-axis indicates cumulative probability of disease-specific survival.

nication. The DTF core gene set contains CTGF, which is known to be expressed by stromal cells and serves multiple functions, including interaction with integrin receptors and growth factors such as TGF-β.[38] The DTF core gene list contains several members of the TGF-β signaling-family (THBS2, ID3, INHBA, FBN1), and the list contains genes involved in Wnt signaling (WISP1, WNT2, DKK3, SDC1). Dysregulation of Wnt signaling and increased levels of stabilized β-catenin have been shown by others to play central roles in the pathogenesis of DTF,[39–42] and the Wnt system has been shown to play a potential role in both fibroblastic differentiation[43] and in mammary gland development and breast carcinogenesis.[44] These functional annotation data support the hypothesis that the DTF core gene list encodes stromal proteins involved in common functional modules in a subset of breast cancers.

We constructed PPI networks for the 46 DTF core proteins and 107 DTF non-core proteins with available PPI data from the STRING database.[29] In these networks, a link is formed between two proteins if they share a predicted functional interaction. This analysis shows that the network created by the DTF core proteins contains more links per node (3.4 *vs* 1.8, $P=0.002$) than the network created by the DTF non-core proteins (Figure 5). The DTF core network contains eight proteins (SDC1, BGN, FN1, COL3A, COL1A2, SPARC, COL1A1, CTGF) with greater than seven links to other proteins, whereas the non-DTF core network contains no proteins with greater than seven links (Figure 5). These findings support the hypothesis that the DTF core genes encode proteins showing a higher degree of biological interrelatedness and a greater likelihood of being involved in a common biological pathway as compared with the non-core DTF genes.

DISCUSSION

Breast cancer is a clinically and biologically heterogeneous disease.[31,32] Breast carcinogenesis results from the acquisition of genetic mutations in the epithelium and the attainment of a tumor micro-environment that provides key growth factors and physical interactions to facilitate tumor initiation and progression.[45,46] Despite the knowledge that the stroma in breast cancer adopts certain characteristic features (such as increased numbers of fibroblasts, increased capillary density, increased type-I collagen and fibrin deposition, and increased fibroblastic expression of α-smooth-muscle actin and stroma-derived factor-1), the contribution of inter-patient stromal variability to the biological and clinical heterogeneity of breast cancer remains poorly understood.[6,7]

Major hurdles in the study of cancer stroma involve the cellular complexity of the tumor microenvironment, both in modeling the microenvironment and isolating pure populations of stromal cell types.[4,6,17] A novel approach developed in our laboratory is the use of gene expression profiles of soft-tissue tumors to investigate different stromal response types in carcinomas. In a proof-of-principle study, we used gene expression profiles from two fibroblastic soft-tissue tumors (DTF and SFT) to identify distinct stromal reaction patterns in breast cancer.[15] Using a single breast cancer gene expression data set, we demonstrated that a subset of breast cancers show increased coordinated expression of DTF-associated genes and that these breast cancers demonstrate significantly improved survival.

Due to high dimensionality of gene array data, associations between elevated levels of expression for a group of genes with outcome should be interpreted with caution.[19] In fact, results from a number of expression profiling studies could not be reproduced on separate data sets.[20]

In the current study, we sought to confirm the robustness and validity of the DTF gene-expression signature in breast cancer and to identify the core subset of DTF genes that are consistently and coordinately expressed in breast cancer. In this study, we demonstrate in four additional independent data sets that the DTF-like stromal response occurs in between 25–35% of invasive breast cancers. We show that over the four data sets this stromal response is associated with lower tumor grade, increased expression of estrogen receptor, and improved survival. In addition, we show that the DTF core signature is significantly associated with multiple independent gene-expression signatures that confer improved prognosis (70-gene ‘good prognosis’ signature, ‘quiescent’ core serum response signature) and negatively correlates with others conferring poor prognosis (basal and luminal B molecular subtypes, p53 mutation signature).

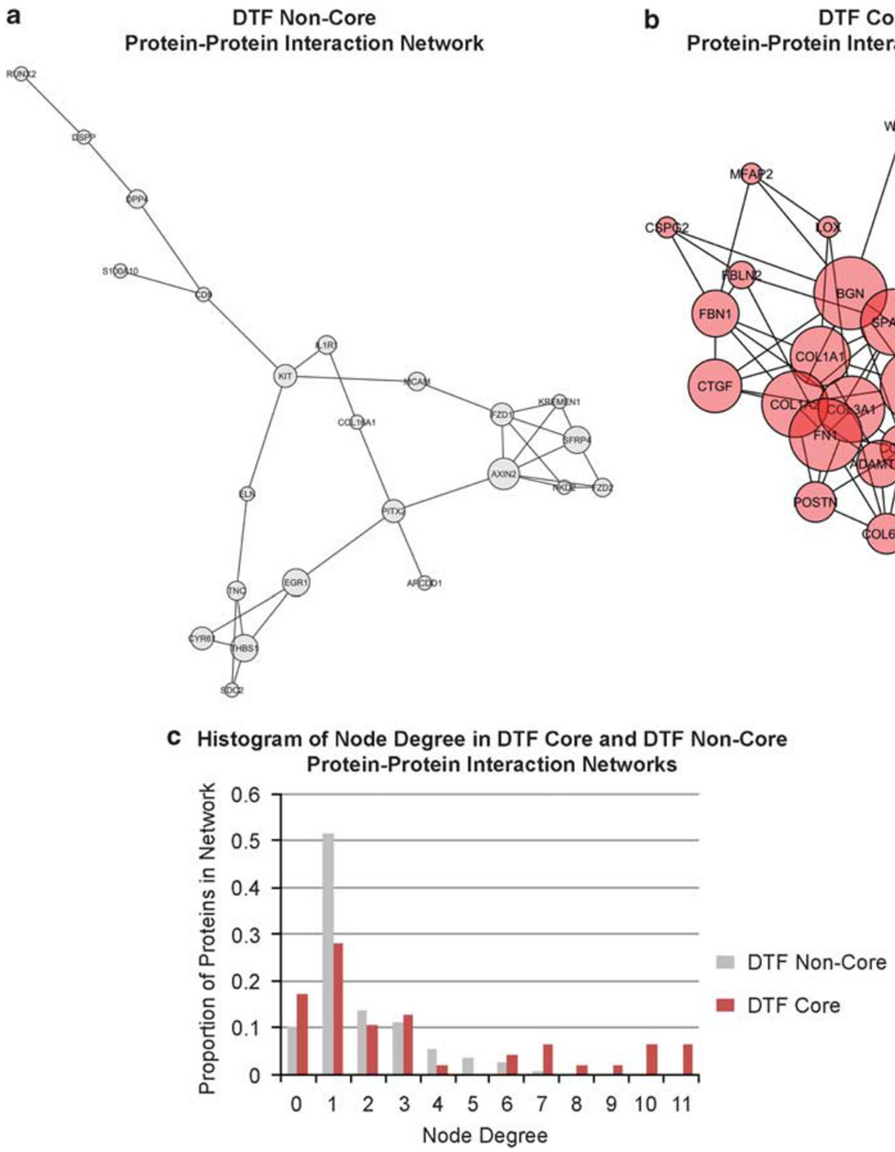


Figure 5 Network properties of the DTF core and DTF non-core protein–protein interaction networks. Each node represents a protein and the links between nodes represent known or predicted functional protein–protein interactions. The degree of a node indicates its number of links to other nodes. The network is visualized using the Cytoscape software platform for visualizing molecular interaction networks with the Cytoscape spring-embedded layout. Only proteins with at least three links to other proteins in the network are displayed. The size of the node correlates with the number of links it has with other proteins. Cytoscape plug-in Network Analyzer 2.5 was used to compute the degree distributions of the two networks. (**a**) The DTF non-core protein–protein interaction network (left, in gray). (**b**) The DTF core protein–protein interaction network (right, in red). (**c**) Histogram of the distribution of links per node in the DTF non-core protein–protein interaction network (gray) and the DTF core protein–protein interaction network (red).

Although breast cancer cases identified by the DTF core signature are significantly correlated with those identified by a variety of other gene-expression signatures, the DTF core signature is unique in that it highlights patterns of gene and protein expression seen primarily in the tumor stroma and not in the carcinoma cells. Consequently, there is very little overlap between the DTF core gene set and the genes identified in other molecular signatures.

The coordinate expression and cellular localization of DTF core proteins was demonstrated on breast cancer TMAs using four genes for which antibody or *in situ* probes are available. Importantly, a further demonstration of the prognostic significance of the DTF gene set was offered by the finding that a single marker (SPARC) showed a trend for improved survival when tested on a breast carcinoma TMA with clinical follow-up. SPARC has been shown to be frequently expressed in the juxtatumoral stroma in breast cancer and tends to be absent from stroma of non-neoplastic breast tissue.[47,48] There have been conflicting reports in the literature as to the prognostic significance of SPARC expression in breast carcinoma, with reports demonstrating that increased SPARC expression by epithelial cells in breast carcinoma is associated with decreased survival[49] and increased propensity for lung metastasis,[50] and other reports showing that endogenous SPARC expression inhibits MDA-MB-231 breast cancer cell metastasis by reducing cancer cell invasive activity and tumor cell-platelet aggregation.[51] Bergamaschi *et al*[13] recently defined extracellular matrix signatures based on evaluating the gene expression profiles of 278 extracellular matrix-related genes derived from the literature, and found SPARC upregulation to be associated with improved outcome in breast cancer. The findings from our study support those of Bergamaschi *et al* and suggest that SPARC is a core member of the DTF-like stromal response seen in breast cancer and associated with improved prognosis.

While stromal gene expression pattern seen in DTF-like cases of breast cancer shares similarities with the gene expression pattern seen in DTF, only a subset of genes highly expressed in DTF consistently show coordinate high levels of expression in DTF-like cases of breast cancer. In the current study, we have used a total of five breast cancer data sets as 'filters' to identify those core DTF genes that define a distinct stromal gene expression pattern in breast carcinoma.

Computational molecular interaction studies can be useful to uncover biological relationships and functional properties of gene sets derived from genome wide studies.[52,53] We have used PPI network analysis to demonstrate that the PPI network created with the DTF core gene set contains significantly increased number of connections per protein as compared with the PPI network created by the non-core DTF proteins (Figure 5). Functional gene set analysis shows that the DTF core gene set is highly enriched for proteins involved in diverse aspects of extracellular matrix structure and function (Supplementary Workbook). These findings suggest that soft-tissue tumors can be used to define relatively large gene sets characteristic of the cell type from which the soft-tissue tumor arises, and carcinoma gene expression data sets can then be employed to identify the core subset of genes that are likely to be involved in common functional modules in regulation of the carcinoma microenvironment.

Several reports posit that a histopathologically defined desmoplastic 'fibrotic focus' correlates with poor prognosis in breast cancer.[54] Our findings suggest that the DTF-like stromal response, which contains profibrotic growth factors (eg, CTGF) and extracellular matrix constituents typically associated with a desmoplastic response (eg, type-I collagens, fibronectin, SPARC), is associated with improved prognosis in breast cancer. Bergamaschi *et al*[13] stratified breast cancers by differences in extracellular matrix gene expression and found no significant difference in stroma morphology between the molecularly defined sub groups. It is conceivable that several distinct molecular pathways result in adoption of a fibrotic appearance to the tumor stroma, and additional research will need to be undertaken to further evaluate the relationship between molecular and morphological changes seen in the breast cancer tumor microenvironment.

Our data demonstrate that a gene set defined by a soft-tissue tumor can be used as a genome-wide search tool to obtain information about gene and protein expression patterns in the microenvironment of breast carcinoma. The elucidation of a distinct stromal gene expression pattern in breast cancer that correlates with molecular and clinicopathologic features of the tumor, advances our knowledge of the biology of the breast cancer tumor microenvironment and provides a valuable resource for future research in this area. This study provides the framework for future studies to evaluate the molecular events that cause certain breast cancers to adopt a DTF-like stroma, and to demonstrate how this particular stromal response type affects tumor cell behavior. It is envisioned that this methodology will be applied to other soft-tissue tumor subtypes to define additional stromal signatures, which can then be evaluated in the tumor microenvironment of carcinomas from a variety of organ systems (eg, prostate, ovary, colon, and pancreas). Ultimately, it is hoped that discovery of genes and proteins playing a concerted role in the tumor stroma will facilitate discovery of therapeutic agents to target and manipulate particular stromal subtypes in the treatment of cancer.[38]

Supplementary Information accompanies the paper on the Laboratory Investigation website (http://www.laboratoryinvestigation.org)

ACKNOWLEDGEMENT

This work was funded in part by an IDEA grant from the California Breast Cancer Research Program (13IB-0076). We thank Maggie Cheang, Chuck Perou, and Chris Fan for helpful statistical support and Torsten Nielsen for valuable comments on the paper.

1. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–799.
2. Sawyers C. Targeted cancer therapy. Nature 2004;432:294–297.
3. Bhowmick NA, Neilson EG, Moses HL. Stromal fibroblasts in cancer initiation and progression. Nature 2004;432:332–337.
4. Bissell MJ, Radisky D. Putting tumours in context. Nat Rev Cancer 2001;1:46–54.
5. Mueller MM, Fusenig NE. Friends or foes—bipolar effects of the tumour stroma in cancer. Nat Rev Cancer 2004;4:839–849.
6. Kalluri R, Zeisberg M. Fibroblasts in cancer. Nat Rev Cancer 2006;6:392–401.
7. Egeblad M, Littlepage LE, Werb Z. The fibroblastic coconspirator in cancer progression. Cold Spring Harb Symp Quant Biol 2005;70:383–388.
8. Elenbaas B, Weinberg RA. Heterotypic signaling between epithelial tumor cells and fibroblasts in carcinoma formation. Exp Cell Res 2001;264:169–184.
9. Chang HY, Chi JT, Dudoit S, *et al*. Diversity, topographic differentiation, and positional memory in human fibroblasts. Proc Natl Acad Sci USA 2002;99:12877–12882.
10. Sugimoto H, Mundel TM, Kieran MW, *et al*. Identification of fibroblast heterogeneity in the tumor microenvironment. Cancer Biol Ther 2006;5:1640–1646.
11. Fukino K, Shen L, Patocs A, *et al*. Genomic instability within tumor stroma and clinicopathological characteristics of sporadic primary invasive breast carcinoma. JAMA 2007;297:2103–2111.
12. Patocs A, Zhang L, Xu Y, *et al*. Breast-cancer stromal cells with TP53 mutations and nodal metastases. N Engl J Med 2007;357:2543–2551.
13. Bergamaschi A, Tagliabue E, Sorlie T, *et al*. Extracellular matrix signature identifies breast cancer subgroups with different clinical outcome. J Pathol 2007;214:357–367.

14. Chang HY, Nuyten DS, Sneddon JB, *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA 2005;102:3738–3743.
15. West RB, Nuyten DS, Subramanian S, *et al.* Determination of stromal signatures in breast carcinoma. PLoS Biol 2005;3:e187.
16. Bacac M, Provero P, Mayran N, *et al.* A mouse stromal response to tumor invasion predicts prostate and breast cancer patient survival. PLoS ONE 2006;1:e32.
17. West RB, van de Rijn M. Experimental approaches to the study of cancer–stroma interactions: recent findings suggest a pivotal role for stroma in carcinogenesis. Lab Invest 2007;87:967–970.
18. van de Vijver MJ, He YD, van't Veer LJ, *et al.* A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
19. Clarke R, Ressom HW, Wang A, *et al.* The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 2008;8:37–49.
20. Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 2003;362:1439–1444.
21. Fan C, Oh DS, Wessels L, *et al.* Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 2006;355:560–569.
22. Allison DB, Cui X, Page GP, *et al.* Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55–65.
23. Perreard L, Fan C, Quackenbush JF, *et al.* Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Res 2006;8:R23.
24. Ma XJ, Wang Z, Ryan PD, *et al.* A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 2004;5:607–616.
25. Pawitan Y, Bjohle J, Amler L, *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 2005;7:R953–R964.
26. Miller LD, Smeds J, George J, *et al.* An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci USA 2005;102:13550–13555.
27. Makretsov NA, Huntsman DG, Nielsen TO, *et al.* Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma. Clin Cancer Res 2004;10:6143–6151.
28. Dennis Jr G, Sherman BT, Hosack DA, *et al.* DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4:P3.
29. von Mering C, Jensen LJ, Kuhn M, *et al.* STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 2007;35:D358–D362.
30. Shannon P, Markiel A, Ozier O, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–2504.
31. Perou CM, Sorlie T, Eisen MB, *et al.* Molecular portraits of human breast tumours. Nature 2000;406:747–752.
32. Sorlie T, Perou CM, Tibshirani R, *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001;98:10869–10874.
33. Chang HY, Sneddon JB, Alizadeh AA, *et al.* Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2004;2:E7.
34. van 't Veer LJ, Dai H, van de Vijver MJ, *et al.* Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–536.
35. Wight TN. Versican: a versatile extracellular matrix proteoglycan in cell biology. Curr Opin Cell Biol 2002;14:617–623.
36. Framson PE, Sage EH. SPARC and tumor growth: where the seed meets the soil? J Cell Biochem 2004;92:679–690.
37. He GP, Muise A, Li AW, *et al.* A eukaryotic transcriptional repressor with carboxypeptidase activity. Nature 1995;378:92–96.
38. Hofmeister V, Schrama D, Becker JC. Anti-cancer therapies targeting the tumor stroma. Cancer Immunol Immunother 2008; 57:1–17.
39. Alman BA, Li C, Pajerski ME, *et al.* Increased beta-catenin protein and somatic APC mutations in sporadic aggressive fibromatoses (desmoid tumors). Am J Pathol 1997;151:329–334.
40. Tejpar S, Nollet F, Li C, *et al.* Predominance of beta-catenin mutations and beta-catenin dysregulation in sporadic aggressive fibromatosis (desmoid tumor). Oncogene 1999;18:6615–6620.
41. Cheon SS, Cheah AY, Turley S, *et al.* Beta-catenin stabilization dysregulates mesenchymal cell proliferation, motility, and invasiveness and causes aggressive fibromatosis and hyperplastic cutaneous wounds. Proc Natl Acad Sci USA 2002;99:6973–6978.
42. Jilong Y, Jian W, Xiaoyan Z, *et al.* Analysis of APC/beta-catenin genes mutations and Wnt signalling pathway in desmoid-type fibromatosis. Pathology 2007;39:319–325.
43. Klapholz-Brown Z, Walmsley GG, Nusse YM, *et al.* Transcriptional program induced by wnt protein in human fibroblasts suggests mechanisms for cell cooperativity in defining tissue microenvironments. PLoS ONE 2007;2:e945.
44. Turashvili G, Bouchal J, Burkadze G, *et al.* Wnt signaling pathway in mammary gland development and carcinogenesis. Pathobiology 2006;73:213–223.
45. Kim JB, Stein R, O'Hare MJ. Tumour–stromal interactions in breast cancer: the role of stroma in tumorigenesis. Tumour Biol 2005;26:173–185.
46. Bissell MJ, Radisky DC, Rizki A, *et al.* The organizing principle: microenvironmental influences in the normal and malignant breast. Differentiation 2002;70:537–546.
47. Barth PJ, Moll R, Ramaswamy A. Stromal remodeling and SPARC (secreted protein acid rich in cysteine) expression in invasive ductal carcinomas of the breast. Virchows Arch 2005; 446:532–536.
48. Iacobuzio-Donahue CA, Ryu B, Hruban RH, *et al.* Exploring the host desmoplastic response to pancreatic carcinoma: gene expression of stromal and neoplastic cells at the site of primary invasion. Am J Pathol 2002;160:91–99.
49. Jones C, Mackay A, Grigoriadis A, *et al.* Expression profiling of purified normal human luminal and myoepithelial breast cells: identification of novel prognostic markers for breast cancer. Cancer Res 2004;64: 3037–3045.
50. Minn AJ, Gupta GP, Siegel PM, *et al.* Genes that mediate breast cancer metastasis to lung. Nature 2005;436:518–524.
51. Koblinski JE, Kaplan-Singer BR, VanOsdol SJ, *et al.* Endogenous osteonectin/SPARC/BM-40 expression inhibits MDA-MB-231 breast cancer cell metastasis. Cancer Res 2005;65:7370–7377.
52. Hu P, Bader G, Wigle DA, *et al.* Computational prediction of cancer-gene function. Nat Rev Cancer 2007;7:23–34.
53. Segal E, Friedman N, Kaminski N, *et al.* From signatures to models: understanding cancer using microarrays. Nat Genet 2005;37(Suppl):S38–S45.
54. Van den Eynden GG, Colpaert CG, Couvelard A, *et al.* A fibrotic focus is a prognostic factor and a surrogate marker for hypoxia and (lymph)angiogenesis in breast cancer: review of the literature and proposal on the criteria of evaluation. Histopathology 2007;51:440–451.