Molecular Markers in Ductal Carcinoma *in Situ* of the Breast

Dale Porter,[1,6] Jaana Lahti-Domenici,[1] Aparna Keshaviah,[2] Young Kyung Bae,[8] Pedram Argani,[8] Jeffrey Marks,[10] Andrea Richardson,[3,6] Amiel Cooper,[4] Robert Strausberg,[9] Gregory J. Riggins,[10] Stuart Schnitt,[5] Edward Gabrielson,[8] Rebecca Gelman,[2,7] and Kornelia Polyak[1,6]

[1]Departments of Medical Oncology and [2]Biostatistics, Dana-Farber Cancer Institute, Boston, MA; [3]Department of Pathology, Brigham and Women's Hospital, Boston, MA; [4]Department of Pathology, Faulkner and Brigham and Women's Hospital, Boston, MA; [5]Department of Pathology, Beth-Israel Deaconess Medical Center, Boston, MA; [6]Harvard Medical School and [7]Harvard School of Public Health, Boston, MA; [8]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD; [9]National Cancer Institute, Bethesda, MD; and [10]Department of Pathology, Duke University Medical Center, Durham, NC

Abstract

Gene expression patterns in ductal carcinoma *in situ* (DCIS), and in invasive, and metastatic breast tumors were determined using serial analysis of gene expression (SAGE). We used mRNA *in situ* hybridization to examine gene expression at the cellular level and immunohistochemistry on tissue microarrays to determine association between gene expression patterns and histopathologic characteristics of the tumors. We found that that the most dramatic transcriptome change occurs at the normal to DCIS transition, while there is no clear universal ''*in situ*'' or ''invasive'' tumor molecular signature. From the 16,430 transcripts analyzed, we identified only 5 and 11 that were preferentially up-regulated in DCIS and invasive tumors, respectively. The majority of invasive cancer specific SAGE tags correspond to novel genes. The genes we identified may define biologically and clinically meaningful subgroups of DCIS with a high risk of progression to invasive disease.

Introduction

Ductal carcinoma *in situ* (DCIS) of the breast includes a heterogeneous group of preinvasive breast tumors with a wide range of malignant potential. Some DCIS, if untreated, will rapidly progress to invasive cancer, while others will change very little in 5–20 years. DCIS now represents a significant (up to 25%) fraction of newly diagnosed breast cancer cases, yet the clinical management of DCIS patients is still controversial. The major challenge is to reliably determine the risk of local recurrence and progression to invasive disease, and to treat patients accordingly. Understanding the pathobiology of DCIS may provide insight into their initiation and progression, but very little is known about DCIS at the molecular level. Current classification of DCIS is based on cytological and architectural features. Some studies have demonstrated a correlation between these features, particularly nuclear grade and the presence of comedo necrosis, and clinical behavior (1–3). In general, high nuclear grade and comedo DCIS are associated with high rate of local recurrence, while low-grade non-comedo tumors have low recurrence rates (4). However, classification based on histological features can be determined only with moderate consistency and is complicated by intratumoral heterogeneity. Moreover, none of these features is able to predict the risk of progression to invasive disease accurately. Therefore, the molecular characterization of DCIS with the aim of identifying biologically and potentially clinically meaningful subgroups is of utmost importance.

Comprehensive gene expression profiling has proven useful for the molecular classification of invasive breast tumors, but no such extensive study has been performed in DCIS largely due to technical difficulties with obtaining DCIS samples suitable for RNA-based analysis (5, 6). Our laboratory uses serial analysis of gene expression (SAGE) to identify genes implicated in breast tumorigenesis (7–9). SAGE requires a relatively small amount of tissue (50,000 cells or less) for the generation of comprehensive expression profiles without the requirement for RNA/cDNA amplification steps; thus, it is particularly well suited for the analysis of small, preinvasive tumors (10). Although SAGE allows the quantitative measurement of the mRNA levels of thousands of genes simultaneously in one specimen, to establish how frequently an emerging candidate gene is differentially expressed, it is necessary to examine hundreds of breast specimens. The recently developed tissue microarray technology allows rapid profiling of hundreds of specimens on one slide and is therefore ideally suited for the further evaluation of candidate genes emerging from SAGE analysis (11, 12).

Here we report that based on SAGE analyses of breast tumors of different histopathologic stages, we identified several transcripts that belong to one of the following groups: up- or down-regulated in breast cancer regardless of stage; preferentially up-regulated in DCIS; or in invasive carcinomas. The expression of 14 genes was confirmed at the cellular level by mRNA *in situ* hybridization using a panel of frozen DCIS and invasive breast tumors. Tissue microarrays composed of

Received 12/2/02; revised 2/10/03; accepted 2/13/03.
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.
Grant support: National Cancer Institute Cancer Genome Anatomy Project and Specialized Program in Research Excellence in Breast Cancer at Dana-Farber/Harvard Cancer Center (CA89393) and Johns Hopkins University (CA88843); Department of Defense Breast Cancer Center of Excellence Grants.
Requests for reprints: Kornelia Polyak, Dana-Farber Cancer Institute, 44 Binney St. D740C, Boston, MA 02115. Phone: (617) 632-2106; Fax: (617) 632-4005. E-mail: Kornelia_Polyak@dfci.harvard.edu

primary breast cancers of different pathological stages were used to explore the potential clinical usefulness of 10 genes by investigating their relationship to histopathologic features of the tumors and patient outcome.

Results

Normal and Cancerous Breast Transcriptomes

Genes differentially expressed between normal and cancerous breast tissues were identified using SAGE. Confirming our previous report using a limited set of SAGE libraries (8), we found that the most dramatic difference in gene expression patterns occurs at the normal to *in situ* carcinoma transition and it involves the uniform down-regulation of 34 genes (Table 1). Because many of these genes encode secreted proteins and genes related to epithelial cell differentiation, loss of the differentiated epithelial phenotype and abnormal autocrine/paracrine interactions appear to play an essential role in the initiation of breast tumorigenesis. We also identified 144 genes up-regulated in a fraction of *in situ* and invasive or metastatic tumors (Table 2). Nearly one-fourth of these SAGE tags currently have no database match, indicating that many transcripts specifically expressed in certain breast carcinomas remain to be identified. To delineate overall similarities and differences among samples, the 19 SAGE libraries were analyzed by hierarchical clustering (Fig. 1A). A dendogram created using this program revealed that the two normal samples (N1 and N2) and primary invasive tumors and lymph node metastases (I1 and LN1, and I2 and LN2) derived from the same patients were most similar to each other. *In situ* and invasive tumors and metastases did not form distinct clusters, suggesting that in this tumor set, there is not a pronounced and common "*in situ*," "invasive," or "metastasis" signature. Correlating with this observation using clustering and other statistical analyses, we were not able to identify any gene that was universally and specifically up- or down-regulated in DCIS or invasive tumors (Fig. 1A). This result confirms previous studies performed in invasive breast carcinomas and highlights that DCIS tumors are just as heterogeneous at the molecular level as their invasive counterparts (6).

To analyze the relationships among DCIS tumors in more detail, we performed hierarchical clustering using the eight DCIS libraries (Fig. 1B). Genes expressed by non-epithelial cells apparently play a predominant role in defining the relatedness of samples, because the BerEP4 purified (D2, D3, D6, and D7) and unpurified (D1, D4, D5, and T18) tumors formed two distinct clusters. Tumors also appeared to cluster according to their histological grade with high-grade (D3, D6, D4, and D5) and intermediate-grade (D2, D7) DCIS showing highest similarity to each other. However, T18, an intermediate-grade, non-comedo DCIS, showed highest similarity to D1, a high-grade comedo DCIS, suggesting that despite the histological features, this DCIS appears to have the molecular signature of a high-grade comedo DCIS.

Putative Molecular Markers in DCIS

To determine if there are genes that are statistically significantly more likely to be expressed in DCIS or invasive tumors, we performed various statistical tests (see "Materials and Methods"). On the basis of these analyses, we found that the expression of CD74 and a SAGE tag (CTGGGCGCCC) with no database match were significantly more abundant in invasive or metastatic tumors than in DCIS ($P = 0.02$ and $P = 0.05$, respectively, Table 3). The expression of MGC2328, DCD, and eight other genes was also more likely to occur in invasive/metastatic tumors than in DCIS, but none of these reached statistical significance (Table 3). Similarly, the expression of S100A7 and keratin 19 (KRT19) was more frequent and at higher levels in DCIS than in invasive/metastatic tumors, but this was also only marginally statistically significant. In a second statistical analysis, receiver operating characteristic (ROC) curve analysis was used to choose the "best" cutoff for values that result in the most samples being correctly classified as DCIS or invasive, weighing both kinds of misclassification equally (Table 3). Tags that do not include 0.50 in the CI might be potentially useful for the differential diagnosis of *in situ* and invasive carcinomas. This includes all the tags that had $P \leq 0.13$ using the higher of two normal cutoffs as well as three other high in DCIS tags and three other high in invasive tags (Table 3). Using the best cutoff values, several of the SAGE tags correctly classified most of the DCIS and invasive SAGE libraries. For example, KRT19 classified 75% of the DCIS and 0% of the invasive libraries as DCIS, while MGC23280 identified 78% of the invasive cancer and 0% of the DCIS libraries as invasive. Thus, MGC23280 had 78% sensitivity and 100% specificity to correctly categorize breast tumors as DCIS or invasive/metastatic in this data set.

Next we selected 26 genes for further validation studies that appeared to be the most highly differentially expressed between normal and DCIS samples or between an intermediate (D2) and a high-grade (D1) DCIS at $P \leq 0.001$ using the SAGE 2000 software (Table 4). We hypothesized that genes most highly differentially expressed between normal and DCIS tissue or two different types of DCIS tumors could be used as molecular markers for defining biologically and potentially clinically meaningful subgroups of DCIS. This hypothesis was supported by the fact that clustering analysis of the eight DCIS libraries using only these 26 genes gave a nearly identical dendogram to that of using over 500 genes (Fig. 1C).

Confirming Gene Expression by Messenger RNA in Situ Hybridization

mRNA *in situ* hybridization determines gene expression at the cellular level, which is particularly useful in solid tumors heterogeneous in their cellular composition. We used 18 frozen DCIS and invasive breast cancer samples for this purpose. Whenever possible, tumors were selected to include normal, DCIS, and invasive components on the same slide to obtain expression data in these three stages of breast tumorigenesis. Examples of *in situ* hybridization results are depicted in Fig. 2A. Interestingly, we found that the expression of several genes up-regulated in DCIS was localized mostly or exclusively to non-epithelial cells. Specifically, CTGF (connective tissue growth factor) and RGS5 (regulator of G protein signaling) were highly expressed in DCIS myoepithelial cells and stromal fibroblasts, although in certain tumors, we detected expression in DCIS epithelial cells as well (Fig. 2A). Cumulative scores for *in situ*

Table 1. Genes Universally Down-Regulated in Breast Cancer Irrespective of Pathological Stage

Tag sequence	Unigene	Gene	Normal			*in Situ*									Invasive							Metastatic			
			N1	N2	Ave	D1	D2	D3	D4	D5	D6	D7	T18	Ave	I1	I2	I3	I4	I5	I6	Ave	LN1	LN2	MET	Ave
Secreted proteins																									
AAATATCCAG	624	interleukin 8[a]	15	5	10	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TGGAAGCACT	624	interleukin 8[a]	368	352	360	8	39	12	1	0	94	15	0	21	2	0	1	0	0	0	1	0	0	0	0
AAGCTCGCCG	62492	secretoglobin, family 3A, member 1 (HIN-1)	125	44	85	0	0	0	3	0	9	0	0	1	0	0	0	0	0	0	0	0	0	4	1
TTGAAACTTT	789	CXCL1 (GRO1)[a]	394	453	423	11	12	14	1	0	61	1	4	13	0	0	1	0	1	0	0	0	0	2	1
TTGCAGGCTC	789	CXCL1 (GRO1)[a]	13	40	26	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ATAATAAAAG	89690	GRO3	24	205	114	4	0	6	4	4	2	0	5	3	7	5	3	8	4	8	6	6	7	11	8
TTGGTTTTTG	164021	small inducible cytokine subfamily B (Cys-X-Cys), member 6	56	16	36	0	3	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	4	1
GAGGGTTTAG	75498	small inducible cytokine subfamily A (Cys-Cys), member 20	44	30	37	2	0	0	0	0	2	2	0	1	0	0	1	0	0	0	0	0	0	0	0
GTACTAGTGT	303649	small inducible cytokine A2	33	12	22	2	0	3	1	0	2	1	0	1	2	3	3	0	1	4	2	0	0	2	1
GCCTTAACAA	239138	pre-B-cell colony-enhancing factor	45	30	38	11	15	0	7	6	17	9	2	8	7	4	5	4	1	4	4	4	3	7	5
GCCTTGGGTG	2250	leukemia inhibitory factor	64	135	99	0	3	8	1	0	4	10	0	3	0	0	1	0	0	4	1	0	0	0	0
Cell surface proteins/receptors																									
ACCAAATTAA	51233	tumor necrosis factor receptor superfamily, member 10b	31	35	33	11	0	0	1	2	6	13	2	4	4	8	1	3	7	12	6	6	7	7	7
AGAAAGATGT	78225	annexin A1	83	77	80	11	3	15	12	10	9	4	23	11	4	16	19	3	7	16	11	6	0	20	9
TGACTGGCAG	278573	CD59 antigen p18-20	49	33	41	15	9	11	0	4	6	9	4	7	4	1	14	11	1	0	5	0	3	5	3
GTCCGAGTGC	374348	ESTs, highly similar to A42926 L6 surface protein	134	96	115	11	33	11	1	2	23	13	4	12	2	0	0	8	0	8	3	2	3	5	3
Cell growth and survival																									
GCTTGCAAAA	372783	superoxide dismutase 2	210	121	166	6	12	5	3	0	10	3	0	5	4	0	1	1	1	4	2	6	3	7	5
ACCAGGCCAC	101382	tumor necrosis factor α-induced protein 2	24	23	23	0	0	0	9	0	7	7	0	3	0	1	1	0	10	0	2	2	0	4	2
TTTGAAATGA	28491	spermidine/spermine N1-acetyltransferase	129	133	131	13	45	37	29	6	20	55	5	26	4	12	40	11	13	20	17	4	4	7	5
CTTGCAAACC	127799	baculoviral IAP repeat-containing 3	16	26	21	0	6	2	1	0	1	2	0	2	2	1	1	0	1	4	2	0	1	4	2
CCATTGAAAC	75517	laminin, β 3	20	21	20	2	3	2	1	0	2	0	7	2	0	0	5	1	1	0	1	0	1	2	1
CCCGAGGCAG	155223	stanniocalcin 2	62	23	43	4	6	0	0	2	4	4	2	3	0	4	6	3	4	0	3	0	1	2	1
CTGGCCCTCG	348024	v-*ral* simian leukemia viral oncogene homologue B	296	145	220	55	117	9	0	31	12	74	69	46	2	1	0	0	1	0	1	2	3	2	2
GACACGAACA	25829	RAS, dexamethasone-induced 1	45	30	38	6	0	8	4	0	2	2	9	4	9	3	1	7	0	0	3	2	4	11	6
GCTGCCCTTG	272897	tubulin, α 3	103	75	89	13	30	3	10	8	18	32	2	15	11	9	13	15	12	20	13	6	12	16	11
Differentiation																									
CGAATGTCCT	335952	keratin 6B	53	49	51	0	0	17	0	0	4	0	0	3	0	0	0	1	0	0	0	0	0	2	1
CTCACTTTTT	76722	CCAAT/enhancer binding protein (C/EBP), delta	154	112	133	38	45	11	16	33	22	22	12	25	7	4	12	17	0	0	6	4	6	23	11
Unknown function																									
AGAATTTAGG	105094	ESTs	13	26	19	2	0	0	0	0	0	0	2	0	0	1	3	0	1	0	1	2	0	0	1
AGTCAAAAAT	NA	No reliable match	13	14	13	0	0	0	0	0	1	4	0	1	0	0	0	0	1	0	0	0	0	0	0
ATTAGTGTTG	23740	KIAA1598 protein	15	7	11	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	4	0	0	1
CTTTGGAAAT	6820	*Homo sapiens* cDNA FLJ32718 fis	16	54	35	4	0	3	1	0	4	5	0	2	0	0	0	0	0	8	1	2	0	9	4
GCAACTTAGA	NA	No reliable match	29	21	25	6	3	0	1	0	2	1	7	3	0	0	4	3	0	0	1	0	0	0	0
GGGACGAGTG	NA	No reliable match	250	460	355	48	493	34	29	53	89	51	49	106	25	9	8	117	3	32	32	16	19	88	41
GGGTTTGTTT	75969	proline rich 2	38	44	41	4	0	3	4	4	20	8	0	5	2	1	6	11	1	8	5	2	1	14	6
GTCTTAAAGT	177781	*Homo sapiens*, clone IMAGE:4711494, mRNA	100	58	79	0	0	3	1	0	21	8	0	4	2	0	5	4	1	8	4	4	1	2	2

Note: Ave, average number of SAGE tags/histological stage.

[a]From interleukin 8 and GRO1, two independent SAGE tags were derived and both were down-regulated in tumors.

Table 2. Genes Up-Regulated in Breast Cancer

Tag	Unigene	Gene	Normal			*in Situ*									Invasive							Metastatic			
			N1	N2	Ave	D1	D2	D3	D4	D5	D6	D7	T18	Ave	I1	I2	I3	I4	I5	T15	Ave	LN1	LN2	MET	Ave
Secreted proteins and ECM related																									
ATGTCTTTTC	1516	insulin-like growth factor binding protein 4	4	5	4	17	36	6	32	59	9	9	4	21	13	29	33	7	19	24	21	8	29	2	13
CATATCATTA	119206	insulin-like growth factor binding protein 7	0	0	0	11	6	6	63	39	4	3	42	22	49	63	59	59	28	80	57	55	12	18	28
CTCCACCCGA	352107	trefoil factor 3 (intestinal)	34	7	21	511	854	17	26	451	31	38	261	274	369	124	15	0	94	16	103	285	244	2	177
ACGTTAAAGA	350570	dermcidin (IBC-1)	0	0	0	0	0	0	1	0	0	0	0	0	177	101	3	0	0	12	49	199	0	0	66
ATTTTCTAAA	91011	anterior gradient 2 homologue	4	7	5	13	75	2	39	2	7	5	0	18	13	17	3	0	12	0	7	2	54	0	19
AGTGGTGGCT	230	fibromodulin	0	0	0	17	0	2	22	0	0	2	34	9	34	36	3	1	70	12	26	22	6	25	18
ATCTTGTTAC	287820	fibronectin 1	0	0	0	4	0	5	7	14	0	2	2	4	2	4	15	4	21	12	10	2	1	0	1
TTATGTTTAA	79914	lumican	0	0	0	2	3	2	28	4	1	1	11	6	0	20	21	1	25	20	14	16	6	11	11
CTCATCTGCT	82109	syndecan 1	0	0	0	0	3	2	25	14	20	2	11	9	4	5	10	36	10	0	11	10	1	9	7
ACATTCCAAG	245188	tissue inhibitor of metalloproteinase 3	0	2	1	13	24	0	12	12	2	7	9	10	7	3	9	1	15	4	6	6	9	7	7
CCAGAGAGTG	180884	carboxypeptidase B1 (tissue)	0	0	0	0	9	0	0	0	0	21	0	4	107	115	0	1	0	0	37	0	354	2	119
TTTGGTTTTC	179573	collagen, type I, α 2	0	0	0	231	0	8	175	53	4	3	12	61	92	90	159	11	158	40	92	138	70	48	85
ACCAAAAACC	172928	collagen, type I, α 1	2	5	3	282	3	8	108	41	22	8	85	70	92	71	83	3	185	189	104	153	34	57	81
TGGAAATGAC	172928	collagen, type I, α 1	2	2	2	191	0	8	260	80	9	0	11	70	184	91	218	23	254	40	135	252	87	39	126
TTTGTTTTTA	3622	procollagen-proline, 2-oxoglutarate 4-dioxygenase	0	0	0	0	3	2	3	2	1	4	2	2	7	7	27	4	21	4	11	2	18	0	7
TGGCCCCAGG	268571	apolipoprotein C-I	2	2	2	8	0	3	44	47	1	3	19	16	87	58	22	8	45	92	52	81	28	32	47
CGACCCCACG	169401	apolipoprotein E	5	2	4	13	0	15	16	33	4	2	65	18	29	37	14	3	54	173	52	31	28	32	31
AACACAGCCT	170250	complement component 4A	5	5	5	25	3	0	52	4	1	5	110	25	29	17	51	0	160	84	57	4	46	7	19
GAATTTCCCA	2253	complement component 2	0	0	0	17	0	0	1	2	0	0	19	5	2	7	1	6	1	8	4	6	1	7	5
CAAACTAACC	153261	immunoglobulin heavy constant μ	0	0	0	11	0	2	50	0	1	0	28	11	172	70	40	1	0	0	47	320	13	193	176
GAAATAAAGC	300697	immunoglobulin heavy constant γ 3	0	0	0	55	0	129	459	10	1	0	247	113	721	665	53	43	0	2442	654	1445	109	770	775
AAACCCCAAT	181125	immunoglobulin λ joining 3	0	0	0	15	0	17	102	4	1	1	44	23	163	87	78	3	0	241	95	258	10	38	102
Cell surface proteins/receptors																									
AAGCACAAAA	9963	TYRO protein tyrosine kinase binding protein	0	0	0	2	0	0	13	12	0	0	0	3	20	12	8	3	16	12	12	14	7	23	15
TGGTTTGCGT	6459	putative G-protein coupled receptor GPCR41	4	7	5	29	36	5	36	45	13	23	12	25	27	25	5	72	12	8	25	24	37	16	25
TACAATAAAC	9071	progesterone receptor membrane component 2	0	0	0	4	9	0	17	18	1	5	0	7	9	5	14	6	18	8	10	20	16	9	15
AGGAAGGAAC	323910	v-*erb*-b2	0	0	0	8	9	11	157	43	110	24	81	55	60	42	13	11	6	96	38	104	12	4	40
ACATTCTTTT	82226	glycoprotein (transmembrane) nmb	2	0	1	4	0	2	7	8	1	0	5	3	4	9	13	18	9	36	15	10	6	25	14
CACCCTGTAC	25450	solute carrier family 29	0	0	0	0	0	2	3	8	0	0	44	7	4	1	5	157	9	20	33	2	9	4	5
GTTCACATTA	84298	CD74 antigen	7	33	20	29	6	25	188	70	6	13	28	46	159	208	226	32	428	474	254	203	72	72	115
CAAGCAGGAC	179516	integral type I protein	2	0	1	17	15	0	38	6	2	4	64	18	29	15	12	30	13	44	24	14	28	16	19
TGCTGCCTGT	118110	bone marrow stromal cell antigen 2	4	9	6	13	57	2	38	14	12	85	57	35	22	41	22	10	21	153	45	6	78	41	42
CCCATCATCC	306122	glycoprotein, synaptic 2	0	0	0	0	6	0	7	16	1	10	16	7	4	8	17	1	15	4	8	2	6	7	5
GCAGTGGCCT	184276	solute carrier family 9	5	7	6	19	96	8	13	53	13	25	9	30	45	32	6	7	19	12	20	31	32	13	25
Cell cycle and apoptosis																									
AAAGTCTAGA	82932	cyclin D1	7	2	5	19	63	6	42	39	29	17	4	27	56	114	36	3	53	12	46	20	140	2	54
CTGGCGCCGA	183180	APC11 anaphase promoting complex subunit 11	4	2	3	11	42	2	7	29	2	2	12	13	22	17	19	11	15	28	19	26	28	20	24
Protein synthesis, transport, and degradation																									
TTTCAGAGAG	75975	signal recognition particle 9 kDa	13	9	11	86	18	23	92	64	10	34	25	44	51	71	83	48	89	24	61	53	60	41	51
TTCTTGCTTA	169895	ubiquitin-conjugating enzyme E2L 6	0	0	0	0	6	3	7	12	2	7	11	6	9	12	14	6	6	36	14	4	25	5	11
GAGAGTGGGG	252259	ribosomal protein S3	0	0	0	6	0	0	0	0	0	0	14	3	18	4	0	0	0	12	6	10	25	0	12

(continued on next page)

Table 2. (*continued*)

Tag	Unigene	Gene	Normal			*in Situ*									Invasive							Metastatic			
			N1	N2	Ave	D1	D2	D3	D4	D5	D6	D7	T18	Ave	I1	I2	I3	I4	I5	T15	Ave	LN1	LN2	MET	Ave
Transcription, chromatin, other nuclear proteins																									
TGAGCAAGCC	27801	zinc finger protein 278	0	0	0	6	0	2	1	2	1	0	7	2	18	11	3	0	9	4	7	14	16	2	11
CCTGTACCCC	32317	high-mobility group 20B	0	0	0	2	3	3	3	8	4	6	25	7	7	7	8	7	6	12	8	2	7	0	3
CCTTTCACAC	278589	general transcription factor II, i	4	2	3	13	15	5	22	59	1	13	14	18	27	24	31	47	37	8	29	16	35	9	20
CACCAGCATT	75847	CREBBP/EP300 inhibitory protein 1	4	0	2	19	15	3	22	18	0	7	30	14	27	15	15	0	9	0	11	22	21	2	15
TTTTGTAATT	75890	membrane-bound transcription factor protease	0	0	0	0	3	3	4	0	1	3	14	4	4	9	8	0	7	4	5	2	16	9	9
GTGCAGGGAG	79414	prostate epithelium-specific Ets transcription factor	2	0	1	8	21	0	57	33	11	13	110	32	56	54	28	3	32	24	33	59	41	2	34
ATGACTCAAG	239752	nuclear receptor subfamily 2	0	0	0	15	9	3	19	39	7	16	5	14	27	21	24	29	23	8	22	18	48	11	26
ATTGTTTATG	181163	high-mobility group nucleosomal binding domain 2	2	9	6	13	18	3	55	55	4	21	14	23	60	53	60	43	47	20	47	51	34	9	31
AAGGATGCCA	169946	GATA binding protein 3	4	0	2	55	9	0	1	14	9	24	9	15	13	7	17	0	26	16	13	8	38	0	15
CTTGTAATCC	183253	nucleolar RNA-associated protein	9	2	6	4	72	78	22	55	7	80	4	40	27	21	14	19	7	104	32	4	62	7	24
TAGTTTGTGG	78934	mutS homologue 2	0	0	0	8	9	5	4	8	0	0	4	5	13	12	12	15	4	0	9	37	10	11	19
Signal transduction																									
CGGTCTTATG	75842	dual-specificity phosphorylation regulated kinase 1A	0	0	0	2	0	0	15	27	4	0	5	7	7	11	18	21	7	8	12	4	3	2	3
TGAAAAGCTT	2384	tumor protein D52	2	2	2	19	21	5	26	47	5	15	2	17	49	44	22	69	19	28	38	18	109	25	50
TTAAGAGGGA	178137	transducer of ERBB2, 1	0	0	0	11	3	8	13	16	0	1	2	7	18	19	28	47	12	4	21	29	12	2	14
TATTTCACCG	138860	Rho GTPase activating protein 1	2	0	1	2	6	3	25	20	5	1	5	8	27	22	12	8	15	0	14	20	9	11	13
GTCTTTCTTG	151536	RAB13, member RAS oncogene family	2	2	2	13	0	2	12	20	0	6	4	7	11	19	32	37	25	8	22	22	9	13	14
CCAGGGGAGA	278613	IFN, α-inducible protein 27	0	0	0	4	36	3	4	90	5	176	2	40	0	21	5	1	3	104	23	2	31	77	37
GAGCAGCGCC	112408	S100 calcium binding protein A7 (psoriasin 1)	18	0	9	1018	3	3	373	16	1	2	890	288	0	0	0	1	0	20	4	0	0	0	0
GCTCTGCTTG	112408	S100 calcium binding protein A7 (psoriasin 1)	2	0	1	76	0	0	20	0	0	0	55	19	0	0	0	0	0	0	0	0	0	0	0
CGCCGACGAT	265827	IFN, α-inducible protein (IFI-6-16)	4	0	2	17	644	3	90	418	18	366	4	195	130	171	5	63	12	161	90	14	526	181	240
GTGTGTTTGT	118787	transforming growth factor, β induced, 68 kDa	0	0	0	8	0	2	10	6	1	0	4	4	13	11	21	8	22	44	20	24	10	9	14
CCAATAAAGT	101850	retinol binding protein 1, cellular	2	0	1	0	3	0	0	2	6	11	7	4	49	28	6	8	0	0	15	102	32	21	52
GTCTAGAATC	92384	vitamin A responsive; cytoskeleton related	0	0	0	21	6	0	25	6	1	4	32	12	16	7	21	11	15	24	15	20	10	5	12
ATCCGCGAGG	180142	calmodulin-like skin protein	0	0	0	0	0	3	22	0	20	0	0	6	47	25	0	52	19	0	24	20	0	0	7
GATTTTGCAC	274479	nucleoside diphosphate kinase 7	0	0	0	19	6	0	7	0	6	1	16	7	9	1	4	1	6	0	4	2	18	2	7
Metabolism																									
ACCTTGTGCC	878	sorbitol dehydrogenase	0	2	1	4	18	0	20	4	1	3	9	7	22	26	1	6	110	4	28	4	95	0	33
TGCCGTTTTG	2006	glutathione *S*-transferase M3 (brain)	0	2	1	0	48	0	1	20	7	25	2	13	9	12	3	4	19	8	9	4	13	7	8
CCGTGCTCAT	9857	dicarbonyl/L-xylulose reductase	11	7	9	2	51	8	20	18	4	5	67	22	99	56	21	7	12	56	42	77	34	7	39
GTTTCTATCA	12540	lysophospholipase I	0	2	1	6	15	0	25	49	1	7	0	13	25	12	26	45	19	8	22	12	38	2	17
CAAATAAAAT	71465	squalene epoxidase	2	2	2	0	24	2	19	55	4	0	5	14	9	8	3	40	13	12	14	4	6	39	16
GGAACTTTTA	43857	similar to glucosamine-6-sulfatases	0	2	1	17	36	3	7	6	4	14	25	14	9	8	26	0	60	0	17	10	10	5	8
TTACCTTTTT	79222	galactosidase, β 1	0	0	0	4	3	0	10	14	0	2	2	4	2	4	8	18	6	16	9	18	3	5	9
TTGGGGAAAC	81029	biliverdin reductase A	4	5	4	4	24	0	22	27	1	9	7	12	43	19	8	3	18	32	20	22	29	11	21
TGATCTCCAA	83190	fatty acid synthase	16	5	10	53	63	6	201	182	31	47	5	74	168	33	105	17	314	4	107	254	46	21	107
TTTGGTGTTT	83190	fatty acid synthase	5	0	3	8	24	2	57	27	5	28	21	21	36	41	62	14	57	12	37	28	10	4	14
TTAACCCCTC	78224	ribonuclease, RNase A family, 1 (pancreatic)	2	0	1	25	0	6	20	10	1	1	5	9	31	57	13	6	0	32	23	18	46	9	24
GCTTTGATGA	89649	epoxide hydrolase 1, microsomal (xenobiotic)	0	2	1	0	6	2	52	20	2	9	12	13	16	29	13	6	29	40	22	29	6	14	17
TACAGTATGT	170171	glutamate-ammonia ligase	0	5	2	13	12	3	36	82	4	24	228	50	4	19	87	26	56	56	41	4	16	0	7

(*continued on next page*)

Table 2. (*continued*)

Tag	Unigene	Gene	Normal			*in Situ*									Invasive								Metastatic			
			N1	N2	Ave	D1	D2	D3	D4	D5	D6	D7	T18	Ave	I1	I2	I3	I4	I5	T15	Ave	LN1	LN2	MET	Ave
TGGGGTTCTT	272499	dehydrogenase/reductase (SDR family) member 2	2	2	2	0	0	2	0	113	0	84	0	25	7	13	10	0	0	0	5	0	32	0	11
TTACTTCCCC	184641	fatty acid desaturase 2	2	0	1	2	0	0	138	29	9	2	0	22	29	19	10	32	43	4	23	53	4	4	20
AAGAATCTGA	183435	NADH dehydrogenase	0	0	0	15	0	3	31	31	1	3	0	10	34	20	14	17	35	0	20	71	46	2	39
GTCCCTGCCT	279837	glutathione *S*-transferase M2	0	5	2	4	18	0	10	53	1	6	5	12	4	13	22		47	0	16	4	12	11	9
AATATGTGGG	351875	cytochrome *c* oxidase subunit VIc	11	5	8	38	707	6	19	219	2	112	23	141	325	337	77	30	185	24	163	28	1250	14	431
GGAGCTCTGT	227750	NADH dehydrogenase 1 β subcomplex, 4	4	5	4	11	39	5	17	27	5	21	14	17	18	11	30	22	29	16	21	16	31	9	19
GAAGGAGATA	171889	choline phosphotransferase 1	0	0	0	4	3	0	0	10	0	1	0	2	9	15	14	34	4	4	13	2	23	2	9
TCAGACTTTT	334305	diacylglycerol *O*-acyltransferase homologue 2	0	0	0	11	0	0	15	0	2	0	28	7	2	22	1	17	0	4	8	2	0	30	11
TCTTGTAACT	256549	nucleotide binding protein 2	0	0	0	0	12	0	9	4	5	4	2		11	13	4	1	4	48	14	22	12	2	12
ESTs																										
TGATGAGTGT	356209	ESTs	0	0	0	2	0	0	1	6	0	3	0	2	2	0	6	6	7	0	4	2	0	0	1
CTGCAACCTA	374393	ESTs	2	0	1	11	6	2	13	8	4	8	9	7	2	7	8	4	7	12	7	12	16	16	15
TGAGTGGTTT	29672	ESTs	0	0	0	4	0	0	3	14	0	0	2	3	4	3	10	12	6	8	7	2	6	5	4
CACTGTGTTG	350475	EST clone IMAGE:4430514	4	0	2	2	3	0	4	2	1	3	18	4	9	7	12	12	7	12	10	6	21	5	11
TTAAGAAGTT	275360	ESTs	7	0	4	15	0	3	63	0	0	0	2	10	2	1	55	0	18	0	13	14	6	0	7
GCGACAGTAA	170853	ESTs	0	0	0	4	0	0	6	16	0	5	16	6	9	8	9	3	15	20	11	2	1	4	2
TCAACTTGAA	99244	ESTs	0	0	0	21	3	3	7	4	12	0	0	6	16	19	9	3	10	0	9	28	40	16	28
TTTCTGGAGG	129943	KIAA0545 protein	2	0	1	15	3	3	4	12	6	1	2	6	16	12	12	6	7	4	9	20	6	13	13
GGGGCTGGAG	301685	KIAA0620 protein	0	0	0	11	6	5	13	29	6	6	4	10	2	9	14	6	7	16	9	8	13	18	13
GTCTCATTTC	90419	KIAA0882 protein	4	0	2	8	3	2	4	23	1	33	0	9	0	13	14	3	21	0	8	0	29	0	10
ACCGCCTGTG	79625	chromosome 20 open reading frame 149	2	5	3	4	36	2	1	80	4	121	19	33	4	7	13	19	21	12	13	6	6	9	7
GAAGAACAGA	29341	chromosome 20 open reading frame 81	0	0	0	13	3	3	4	16	0	2	2	5	4	9	14	8	6	0	7	6	15	7	9
TCGTAACGAG	11197	chromosome 20 open reading frame 92	4	2	3	11	0	0	15	8	4	3	23	8	25	8	18	19	4	12	14	22	10	16	16
GTGATGGGGC	62620	chromosome 6 open reading frame 1	2	0	1	2	12	0	13	2	0	4	11	5	16	3	6	6	13	0	7	20	10	9	13
GAGAGAAAAT	181444	hypothetical protein LOC51235	0	2	1	40	9	0	10	6	7	7	21	13	4	8	9	11	18	0	8	6	10	27	14
GCCCACATCC	84753	hypothetical protein FLJ12442	4	0	2	0	0	3	4	0	4	1	26	5	63	26	1	12	6	48	26	49	1	11	20
GTATTTAACT	209065	hypothetical protein FLJ14225	0	0	0	17	6	3	28	12	6	8	9	11	9	16	15	6	16	0	10	20	10	18	16
GGCTGGTCTC	324844	hypothetical protein IMAGE3455200	2	2	2	6	6	5	6	12	2	3	11	6	18	7	10	18	12	16	13	6	18	20	14
AACACTTCTC	333526	hypothetical protein MGC14832	4	0	2	2	6	0	25	8	1	2	4	6	27	19	4	0	9	4	10	18	6	4	9
AATAAAGAGA	28149	hypothetical protein BC010626	0	2	1	0	3	0	6	23	0	1	60	12	7	4	21	0	31	0	10	6	0	2	3
GAGAAACATT	267245	hypothetical protein FLJ14803	0	2	1	17	0	0	4	8	1	2	2	4	7	5	14	12	13	4	9	14	12	5	10
TTTGGTCTTT	109773	hypothetical protein FLJ20625	0	0	0	8	0	3	6	10	4	4	4	5	20	28	12	15	15	24	19	10	10	0	7
TGTGGTGGTG	83422	MLN51 protein	5	2	4	6	3	2	55	39	7	7	4	15	87	25	18	22	13	36	34	92	18	5	38
GAAAGATGCT	334370	brain expressed, X-linked 1	2	0	1	6	48	0	1	0	1	1	0	7	29	37	1	1	1	0	12	0	162	2	54
TAGCAGACCC	349196	myeloid/lymphoid or mixed-lineage leukemia	0	0	0	0	3	3	1	4	2	7	12	4	13	13	12	7	4	20	12	18	1	0	6
No database match																										
AACGCTGCGA	NA	No reliable match	7	5	6	36	24	0	4	35	1	10	0	14	31	60	23	1	19	0	22	29	101	23	51
AATGGATGAA	NA	No reliable match	0	0	0	38	0	0	3	2	1	0	44	11	2	0	0	0	0	60	10	4	1	0	2
ACATCGTAGT	NA	No reliable match	0	0	0	0	15	0	3	31	0	2	2	7	13	20	4	4	10	4	9	0	60	0	20
ACCCGCCGGG	NA	No reliable match	11	7	9	103	18	3	4	0	1	6	166	38	20	8	0	1	4	193	38	31	23	0	18
AGTGCAGGGA	NA	No reliable match	0	0	0	2	0	2	15	2	0	0	37	7	38	9	23	1	1	48	20	26	0	7	11
ATCAAGAATC	NA	No reliable match	2	0	1	2	3	3	9	8	0	3	9	5	18	13	15	4	16	72	23	22	13	13	16
ATGTGGCACA	NA	No reliable match	4	2	3	2	24	0	20	31	1	9	34	15	18	16	12	44	23	8	20	14	15	9	12
CAAACCTTTA	NA	No reliable match	0	0	0	11	6	0	16	25	1	5	0	8	16	16	13	23	13	8	15	33	15	34	27
CAATGCTGCC	NA	No reliable match	11	12	11	53	12	3	23	33	9	3	64	25	580	145	18	18	26	44	139	588	28	11	209

(continued on next page)

Table 2. (*continued*)

Tag	Unigene	Gene	Normal			*in Situ*									Invasive							Metastatic			
			N1	N2	Ave	D1	D2	D3	D4	D5	D6	D7	T18	Ave	I1	I2	I3	I4	I5	T15	Ave	LN1	LN2	MET	Ave
CAGCTTAATT	NA	No reliable match	4	2	3	4	3	0	25	20	0	1	2	7	36	20	0	0	4	4	11	90	6	5	34
CCGACGGGCG	NA	No reliable match	4	2	3	67	3	0	3	0	1	4	87	21	7	0	0	0	0	181	31	4	7	0	4
CCTTTGAACA	NA	No reliable match	2	0	1	4	6	5	0	10	2	3	14	6	9	13	5	12	6	16	10	2	4	4	3
CCTTTGCCCT	NA	No reliable match	0	0	0	0	9	2	73	16	1	14	5	15	27	26	19	0	9	0	14	28	9	0	12
CGGTTTAATT	NA	No reliable match	2	0	1	23	0	0	12	10	1	3	53	13	13	9	26	3	25	16	15	20	0	0	7
CTTTATTCCA	NA	No reliable match	0	0	0	19	0	2	48	2	0	0	5	9	25	22	31	4	16	0	16	18	15	5	13
GAAGTCGGAA	NA	No reliable match	4	0	2	48	0	2	3	2	27	3	2	11	20	3	4	12	4	0	7	18	9	7	11
GATCTCGCAA	NA	No reliable match	4	7	5	44	21	0	31	25	7	1	0	16	40	13	12	22	16	4	18	47	38	64	50
GCACCTCCTA	NA	No reliable match	2	0	1	8	9	2	7	12	4	1	2	6	13	12	6	11	10	0	9	12	6	7	8
GCCGTGAGCA	NA	No reliable match	2	0	1	17	12	0	6	8	2	1	5	6	25	17	1	6	13	0	10	12	31	20	21
GGAAAGTGAC	NA	No reliable match	0	0	0	2	6	2	4	10	0	5	7	5	11	22	12	6	26	0	13	12	23	9	15
GGACCTTTAT	NA	No reliable match	2	0	1	23	3	0	1	23	1	0	37	11	2	1	1	0	1	0	1	4	3	0	2
GGCAGACAAT	NA	No reliable match	0	0	0	13	0	0	12	14	1	2	7	6	16	5	1	15	7	0	7	18	12	13	14
GGCAGCACAA	NA	No reliable match	0	5	2	23	18	0	16	27	20	12	5	15	49	11	5	12	6	4	15	35	25	29	30
GGTAGCTGCT	NA	No reliable match	0	0	0	6	3	0	3	20	0	6	14	7	7	4	4	4	3	0	4	2	1	4	2
GGTAGTTTTA	NA	No reliable match	13	0	6	59	21	3	32	41	2	13	18	24	18	28	39	0	59	16	26	18	79	0	32
GGTCAGTCGG	NA	No reliable match	5	5	5	76	15	2	0	0	39	3	102	30	25	3	1	7	1	80	20	18	13	2	11
GTAATCCTGC	NA	No reliable match	4	2	3	34	6	12	0	4	187	28	51	40	22	17	6	25	1	52	21	24	7	7	13
GTAGTTACTG	NA	No reliable match	2	2	2	8	120	0	1	25	0	21	4	22	38	33	13	7	19	0	18	8	172	4	61
TCACAGTGCC	NA	No reliable match	2	2	2	15	3	2	13	39	1	7	14	12	29	5	42	28	21	8	22	20	6	13	13
TCTGGTTTGT	NA	No reliable match	2	2	2	6	12	3	10	33	5	2	7	10	29	16	4	50	3	12	19	41	6	7	18
TGAAGCAGTA	NA	No reliable match	4	2	3	99	3	2	36	27	9	5	25	26	74	46	122	57	85	12	66	57	40	25	41
TGTCATAGTT	NA	No reliable match	0	0	0	0	15	0	9	55	0	3	9	11	34	42	9	4	34	4	21	6	197	0	68
TTACGATGAA	NA	No reliable match	2	0	1	0	6	0	3	18	1	1	0	4	51	41	4	1	7	0	18	73	9	2	28
TTCGGTTGGT	NA	No reliable match	2	0	1	101	3	0	55	16	0	0	7	23	58	40	40	1	60	4	34	55	22	11	29

Note: Ave, average number of SAGE tags/histological stage.

hybridization were used for hierarchical clustering analysis and statistical tests. A dendogram of the 18 different tumors and 5 normal breast tissues determined that using 14 genes, we were able to differentiate between normal and cancer samples and group the tumors into subclasses (Fig. 2B). Confirming our SAGE results, we were not able to differentiate DCIS and invasive tumors based on gene expression profiles. Surprisingly, in the majority of cases within the same tissue sample, the *in situ* and invasive areas of the tumor did not show the highest similarity to each other (Fig. 2B). Although this result could be due to the use of mRNA *in situ* hybridization and the selected genes, it may suggest that gene expression profiles are not necessarily maintained during tumor progression. Fisher's exact test revealed significant positive correlation between the expression of TFF3 and IFI-6-16 ($P = 0.01$), LOC51235, and BEX1 ($P = 0.05$), while inverse correlation was found between the expression of S100A7 and RGS5Tu (tumor cell specific expression) ($P = 0.04$), S100A7 and TFF3 ($P = 0.04$), and CTGF and TM4SF1 ($P = 0.01$). No statistically significant associations were found between the expression of the genes and histopathologic features of the tumors.

Immunohistochemical Analysis of Tissue Microarrays and Clinicopathologic Associations

Next we analyzed the expression of 10 genes by immunohistochemistry using tissue microarrays composed of tumors of different pathological stages. In total, we analyzed 769 tumor samples [634 primary invasive tumors, 53 metastases (distant and lymph node), and 82 DCIS-71 of which were pure DCIS] obtained from eight different cohorts (tissue microarrays), but not all genes were analyzed in all data sets. An example of immunohistochemical staining of a DCIS using multiple genes is depicted in Fig. 2C. Cumulative scores for immunohistochemical staining were used for statistical analyses to determine associations between the expression of the genes and histopathologic features of the tumors or between different genes. In addition, we also analyzed S100A7 expression in relation to clinical outcome (overall survival and distant metastasis free survival) in two of the patient cohorts. Confirming our SAGE results, the expression of DCD was limited to a subset of invasive breast carcinomas with only 1 out of 70 DCIS tumors showing detectable DCD expression (Fig. 2C and data not shown). The expression of CTGF, TFF3, and SPARC in the stroma was statistically significantly related to pathological stage with TFF3 and SPARC less likely to be expressed in DCIS than in invasive or metastatic tumors (Table 5). Statistically significant association between S100A7 expression and estrogen receptor (ER) negativity, high histological grade, and more than four positive lymph nodes was demonstrated in logistic regression models in primary invasive tumors (Table 5). Because all these tumor characteristics are known to correlate with poor prognosis, it is likely that S100A7 identifies a clinically meaningful subgroup of tumors. Kaplan-Meier analysis demonstrated decreased overall survival for patients with S1007 positive tumors, but this did not reach statistical significance ($P = 0.41$) possibly due to relatively short patient follow-up data and insufficient sample size (data not shown). The expression of fatty acid synthase (FASN) was higher in ER or HER2 positive high-grade tumors, while the expression of SPARC (osteonectin) inversely correlated with high histological grade and TNM stage 3 (Table 5). The fraction of breast tumors that expressed the cytokines CXCL1 (GRO1), CXCL2 (GRO2), and IL-8 was, as

expected, very low, because these genes were highly expressed in normal mammary epithelium based on SAGE and immunohistochemistry (data not shown), but GRO1 and IL-8 were frequently co-expressed in the same tumors ($P = 0.006$). Using Fisher's exact test, the expression of S100A7 was associated with a larger likelihood of expression of FASN ($P = 9.95 \times 10^{-6}$) and TFF3 ($P = 0.002$), and a lower likelihood of expression of CTGF ($P = 0.005$); the expression of FASN was associated with that of TFF3 ($P = 3.5 \times 10^{-6}$) and SPARC in the tumor ($P = 6.6 \times 10^{-4}$) and stromal ($P = 0.01$) cells, TFF3 expression was associated with that of NFIKBA ($P = 0.03$) and SPARC-Tumor ($P = 4 \times 10^{-5}$), while the expression of OST in tumor cells correlated with that of stromal cells ($P = 0.03$).

Discussion

The goal of our study was to identify genes that may be used as markers for the molecular classification of DCIS tumors into biologically and potentially clinically meaningful groups. To find these genes, we quantitatively compared transcripts present in normal mammary epithelial cells, and different types of DCIS, and invasive or metastatic breast cancer using SAGE. On the basis of hierarchical clustering analyses of these SAGE libraries, we detected no clear molecular signatures specific for *in situ*, invasive, or metastatic breast tumors. However, using further statistical tests, we identified genes that appear to be preferentially expressed in DCIS or invasive/metastatic breast tumors. The majority of SAGE tags fairly specific for invasive/metastatic tumors correspond to novel transcripts, suggesting that many of the genes specific for a particular cell or cancer type still remain to be identified and that SAGE is a powerful technique for the discovery of such genes. The failure to find differences at the molecular level between primary invasive and metastatic tumors is in agreement with that of another gene expression profiling study performed using DNA microarrays (13). To our knowledge, there are no published gene expression studies analyzing *in situ* and invasive breast carcinomas using DNA microarrays.

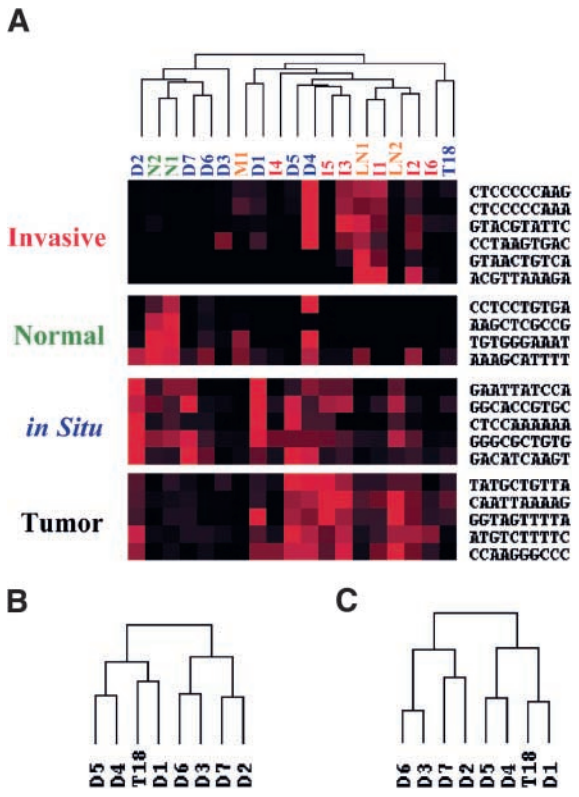

FIGURE 1. Hierarchical clustering analysis of breast SAGE libraries. **A.** Dendogram depicting the relatedness of SAGE libraries generated from normal mammary epithelial cells (*N1* and *N2*, *green*), DCIS (*D1–7* and *T18*, *blue*), primary invasive breast tumors (*I1–6*, *red*), and lymph node (*LN1* and *LN2*, *orange*) and distant lung metastases (*M1*, *orange*). Examples for SAGE tags preferentially present in normal breast tissue, or invasive or *in situ* carcinomas. **B.** Dendogram demonstrating similarities among DCIS tumors using 515 genes for hierarchical clustering. **C.** Dendogram of hierarchical clustering performed using the 26 most highly differentially expressed genes selected for further validation studies (genes listed in Table 4).

Although no genes were absolutely specific for DCIS tumors, two calcium binding proteins of the S100 family, S100A7 (psoriasin) and S100A9, trefoil factor 3 (TFF3), keratin 19 (KRT19), and apolipoprotein D (APOD) were preferentially more abundant in DCIS than in normal or invasive/metastatic cancerous breast tissue. S100A7 has been previously identified as a gene differentially expressed between *in situ* and invasive breast cancer and it has been demonstrated to bind calcium and act as a chemoattractant for T-lymphocytes and neutrophils (14). Correlating with this, S100A7 expression was particularly high in high-grade comedo DCIS that frequently demonstrate strong lymphocytic infiltration. Similarly, S100A9 is known to influence cell migration (15). However, it is likely that both S100A7 and S100A9 have additional intracellular functions such as the regulation of cell growth, differentiation, or survival. Keeping with this, the expression of S100A7 was shown to be dramatically up-regulated in response to apoptosis inducing stimuli and in psoriatic and premalignant keratinocytes that are unable to differentiate (16–18). The high expression of S100A7 in ER negative, poorly differentiated, and lymph node positive invasive breast tumors suggests that its expression may predict bad clinical outcome and high risk of recurrence or progression in DCIS. Kaplan-Meier curve analysis demonstrating a somewhat decreased overall survival of patients with S100A7 positive invasive breast tumors supports this hypothesis, but to conclusively determine if S100A7 could be useful for the prognostication of breast cancer patients requires further studies.

Another gene highly expressed in DCIS is trefoil factor 3, a secreted protein that has been implicated in tumorigenesis, wound healing, and regulation of epithelial integrity (19). Although its role in breast tumorigenesis is unclear, in colon cancer cells, it has been demonstrated to confer apoptosis resistance (20). Keratin 19 and apolipoprotein D were also more abundant in DCIS than in normal epithelial cells and invasive carcinomas. Keratin 19 is the smallest known acidic keratin that is preferentially expressed in breast tumors of ductal origin (21). Apolipoprotein D is a glycoprotein involved in lipid transport that is negatively regulated by estrogen and its expression in invasive breast cancer may correlate with shorter disease free survival (22, 23). Although the differential expression of KRT19, S100A9, TFF3, and APOD between *in situ* and invasive carcinomas has not been analyzed in detail, the differences we detected by SAGE would be impossible to confirm by mRNA *in*

Table 3. Genes "Specific" for *in Situ* and Invasive or Metastatic Breast Cancer SAGE Libraries

Tag sequence	Unigene	Gene	*P*	ROC area × 100	ROC area × 100 95% CI	ROC best cutoff	DCIS % > cutoff	IDC % > cutoff	Normal		*in Situ*								Invasive						Metastatic		
									N1	N2	D1	D2	D3	D4	D5	D6	D7	T18	I1	I2	I3	I4	I5	I6	LN1	LN2	M1
DCIS "specific" genes																											
GAGCAGCGCC	112408	S100A7[a] (psoriasin)	0.29	92	77–100	2.00	88	11	18	0	1018	3	3	373	16	1	2	890	0	0	0	1	0	20	0	0	0
GCTCTGCTTG	112408	S100A7[a] (psoriasin)	0.08	69	51–87	54.70	38	0	2	0	76	0	0	20	0	0	0	55	0	0	0	0	0	0	0	0	0
GGACCTTTAT	352107	TFF3[a] (trefoil factor 3)	0.33	64	35–93	3.00	50	11	2	0	23	3	0	1	23	1	0	37	2	1	1	0	1	0	4	3	0
CTCCACCCGA	352107	TFF3[a] (trefoil factor 3)	1.00	69	42–97	16.80	100	56	34	7	511	854	17	26	451	31	38	261	369	124	15	0	94	16	285	244	2
GTGGCCACGG	112405	S100A9 (calgranulin B)	0.29	85	63–100	4.10	88	22	29	30	200	0	9	238	4	20	15	92	0	1	1	3	0	72	0	0	4
GACATCAAGT	182265	KRT19 (keratin 19)	0.06	83	58–100	58.90	75	0	33	35	59	165	3	118	139	59	153	34	20	40	41	25	31	20	10	34	16
CCCTACCCTG	75736	APOD (apolipoprotein D)	0.21	76	52–100	7.70	100	44	4	58	15	42	8	293	215	9	12	49	2	16	41	3	4	44	0	3	16
Invasive or metastatic breast cancer "specific" genes																											
ACGTTAAAGA	350570	DCD (Dermcidin)	0.13	75	55–95	2.50	0	56	0	0	0	0	0	1	0	0	0	0	177	101	3	0	0	12	199	0	0
CCAGAGAGTG	180884	CPB1 (carboxypeptidase B1)	0.33	67	43–91	1.30	25	56	0	0	0	9	0	0	0	0	21	0	107	115	0	1	0	0	0	354	2
GGAGTAAGGG	5163	MGC23280 (hypothetical protein)	0.06	86	68–100	1.46	0	78	0	0	0	0	0	1	0	0	1	0	22	8	0	3	1	0	22	1	2
CTGGGCGCCC	NA	No reliable match	0.05	80	61–99	12.00	0	56	0	0	0	0	2	0	0	0	0	0	40	25	0	0	0	12	26	1	34
CCAATAAAGT	101850	RBP1 (retinol binding protein)	0.33	78	54–100	6.40	25	78	2	0	0	3	0	0	2	6	11	7	49	28	6	8	0	0	102	32	21
TTTGTTTTTA	131740	FLJ30428 (hypothetical protein)	1.00	84	62–100	4.01	0	78	0	0	0	3	2	3	2	1	4	2	7	7	27	4	21	4	2	18	0
ATCCGCGAGG	180142	CLSP (calmodulin-like skin protein)	0.64	64	38–89	19.00	25	56	0	0	0	0	3	22	0	20	0	0	47	25	0	52	19	0	20	0	0
GACCACACCG	367741	NUDT8 (nudix)	0.64	69	43–96	8.00	0	56	2	2	2	0	0	7	0	7	0	5	27	21	1	0	0	8	33	9	0
CGATATTCCC	37616	MGC14480 (hypothetical protein)	0.33	79	57–100	6.40	25	78	4	2	4	6	0	3	12	1	6	7	36	26	6	4	9	12	31	13	2
AAACCCCAAT	181125	IGL (immunoglobulin λ)	1.00	72	46–97	38.00	25	67	0	0	15	0	17	102	4	1	1	44	163	87	78	3	0	241	258	10	38
GTTCACATTA	84298	CD74 antigen	0.02	93	81–100	31.70	25	100	7	33	29	6	25	188	70	6	13	28	159	208	226	32	428	474	203	72	72

Note: Human Gene Nomenclature Database (www.gene.ucl.ac.uk/nomenclature) symbols are provided if available. *P* is based on using the SAGE tag number which was highest of two normals as cutoff (details in "Materials and Methods"). The first ROC column gives the ROC area, the second the approximate 95% CI, the third the "best" cutoff, while the last two columns show the percentage of DCIS specimens with values greater than or equal to the ROC best cutoff and the percentage of invasive specimens with values greater than or equal to the ROC best cutoff. "Specific" — enriched/more abundant, but not necessarily absolutely specific.

[a]From two transcripts (S100A7 and TFF3), two independent SAGE tags were derived and both found to be specific for DCIS.

Table 4. Genes Selected for Messenger RNA *in Situ* Hybridization and Immunohistochemical Analyses

Tag sequence	Unigene	Gene	Normal			*in Situ*									Invasive							Metastatic				Method	Cell[a]
			N1	N2	Ave	D1	D2	D3	D4	D5	D6	D7	T18	Ave	I1	I2	I3	I4	I5	I6	Ave	LN1	LN2	M1	Ave		
Normal specific																											
AAGCTCGCCG	62492	SCGB3A1 (HIN-1, High in Normal-1)	125	44	85	0	0	0	3	0	9	0	0	1	0	0	0	0	0	0	0	0	0	4	1	ISH	LE
GTCCGAGTGC	351316	TM4SF1 (transmembrane 4 superfamily member 1)	134	96	115	11	33	11	1	2	23	13	4	12	2	0	0	8	0	8	3	2	3	5	3	ISH	LE
GACTGCGCGT	10086	FN14 (Type I transmembrane protein Fn14)	40	26	33	0	36	6	3	4	22	32	4	13	0	3	0	1	1	8	2	0	0	0	0	ND	LE
TTGAAGCTTT	75765	CXCL2 (GRO2, growth-related protein 2)	122	247	184	2	3	15	0	0	29	5	0	7	0	1	4	0	0	0	1	0	0	0	0	IH	LE
TTGAAACTTT	789	CXCL1 (GRO1, growth-related protein 1)	394	453	423	11	12	14	1	0	61	1	4	13	0	0	1	0	1	0	0	0	0	2	1	IH	LE
TGGAAGCACT	624	IL-8 (interleukin-8)	368	352	360	8	39	12	1	0	94	15	0	21	2	0	1	0	0	0	1	0	0	0	0	IH	LE
TAACAGCCAG	81328	NFKBIA (NFKB inhibitor α)	136	152	144	6	39	23	4	2	28	125	19	31	4	7	8	7	9	4	6	2	10	20	11	IH	LE
Tumor specific																											
CAATTAAAAG	149923	XBP1 (X-box binding protein)	80	58	69	147	196	29	366	322	27	97	214	175	244	247	535	18	531	129	284	199	599	7	268	ISH	LE
TTTGGTGTTT	83190	FASN (fatty acid synthase)	5	0	3	8	24	2	57	27	5	28	21	21	36	41	62	14	57	12	37	28	10	4	14	IH	LE
TGATCTCCAA	83190	FASN (fatty acid synthase)	16	5	10	53	63	6	201	182	31	47	5	74	168	33	105	17	314	4	107	254	46	21	107	IH	LE
CTCCACCCGA	82961	TFF3 (trefoil factor 3)	34	7	21	511	854	17	26	451	31	38	261	274	369	124	15	0	94	16	103	285	244	2	177	ISH + IH	LE + F
Intermediate-grade DCIS specific																											
CGCCGACGAT	265827	IFI-6-16 (IFN α-inducible protein)	4	0	2	17	644	3	90	418	18	366	4	195	130	171	5	63	12	161	90	14	526	181	240	ISH	LE + F
TTTGGGCCTA	17409	CRIP1 (cysteine-rich protein 1)	33	5	19	21	66	29	22	33	49	223	4	56	7	49	37	0	35	4	22	2	60	7	23	ISH	LE
AATCTGCGCC	833	ISG15 (IFN-stimulated protein, 15 kDa)	0	0	0	2	48	2	3	20	1	42	2	15	9	5	1	0	1	28	8	4	29	16	16	ISH	LE
CCAGGGGAGA	278613	IFI27 (IFN α-inducible protein)	0	0	0	4	36	3	4	90	5	176	2	40	0	21	5	1	3	104	23	2	31	77	37	ISH	LE
GAAAGATGCT	334370	BEX1 (brain expressed, X-linked 1)	2	0	1	6	48	0	1	0	1	1	0	7	29	37	1	1	1	0	12	0	162	2	54	ISH	ME
CAGACTTTTT	293884	LOC150678 (helicase/primase protein)	7	5	6	4	54	5	1	4	0	31	5	13	2	9	4	1	4	0	4	0	4	4	3	ISH	LE
CTGGCGCCGA	183180	NAPC11 (anaphase promoting complex subunit 11)	4	2	3	11	42	2	7	29	2	2	12	13	22	17	19	11	15	28	19	26	28	20	24	ND	NA
TGAGCTACCC	72222	FER1L4 (Fer-1-like 4)	0	0	0	0	33	0	0	6	0	0	11	6	2	0	0	1	0	4	1	0	0	0	0	ND	NA
High-grade DCIS specific																											
GAGCAGCGCC	112408	S100A7 (psoriasin)	18	0	9	1018	3	3	373	16	1	2	890	288	0	0	0	1	0	20	4	0	0	0	0	ISH + IH	LE
TTTGCACCTT	75511	CTGF (connective tissue growth factor)	0	0	0	141	6	18	63	18	9	6	41	38	9	42	43	66	19	16	32	10	7	48	22	ISH + IH	ME + F
TATGAGGGTA	24950	RGS5 (regulator of G-protein signaling 5)	0	0	0	40	0	0	1	0	0	6	46	12	4	0	1	0	0	8	2	0	1	4	2	ISH	ME
GAAGTTATAA	137476	PEG10 (paternally expressed 10)	0	7	4	44	3	0	6	0	33	1	16	13	0	4	0	4	1	0	2	8	0	0	3	ISH	LE
ATGTGAAGAG	111779	SPARC (osteonectin)	4	0	2	118	3	6	79	39	22	6	12	36	112	97	185	47	194	96	122	163	32	129	108	IH	LE + F
GAGAGAAAAT	181444	LOC51235 (hypothetical protein)	0	2	1	40	9	0	10	6	7	7	21	13	4	8	9	11	18	0	8	6	10	27	14	ND	NA
CTCCCCCAAA	293441	SNC73 (immunoglobulin heavy μ chain)	2	14	8	78	0	20	605	37	1	0	11	94	159	86	186	0	6	12	75	140	19	109	89	ISH	LY

Note: Gene names and Unigene identification numbers are provided when available. ISH, *in situ* hybridization; IH, immunohistochemistry; ND, not determined; NA, not applicable. LE, luminal epithelial cells; ME, myoepithelial cells; F, fibroblasts; LY, leukocytes.

[a]Cell refers to cell type expressing the gene based on mRNA *in situ* hybridization or immunohistochemistry.

situ hybridization or immunohistochemistry due to the semi-quantitative nature of these latter techniques. Because we believe that the most likely clinical application of molecular markers would be their evaluation by immunohistochemistry, these four genes have limited potential as molecular markers.

In addition to selecting DCIS and invasive/metastatic cancer specific genes, we also identified 26 transcripts most highly differentially expressed between normal and DCIS or between different types of DCIS and demonstrated that these genes are able to classify DCIS as accurately as over 500 genes could (Fig. 1, B and C). The expression of 15 and 10 of these genes was confirmed by mRNA *in situ* hybridization and immunohistochemistry, respectively (Fig. 2 and Table 5), and the expression of some of them appeared to correlate with the histopathologic features of the tumors suggesting that they may identify subgroups of DCIS.

The majority of genes that are preferentially expressed in invasive/metastatic carcinomas correspond to uncharacterized genes. Four transcripts (MGC23280, MGC14480, FLJ30428, and NUDT8) encode hypothetical proteins with no known functions, while one of the SAGE tags currently has no ESTs match (Table 3). Dermcidin (DCD) was recently identified as a small secreted protein highly expressed in sweat glands of the skin, but its role in breast cancer is unclear (24). Calmodulin-like skin protein is a calcium-binding protein with a putative role in keratinocyte differentiation (25). Similarly, cellular retinol-binding protein is a protein carrier of retinol, a compound essential for normal epithelial cell differentiation (26). The high expression of immunoglobulin λ and CD74 antigen may reflect leukocytic infiltration of the tumors. However, CD74, the invariant (γ) chain of the MHC class II antigen, was also highly abundant in SAGE libraries generated from purified epithelial cells; thus, it is likely to be expressed by the cancer cells themselves although the functional relevance of this is unclear. We hypothesized that these "invasive cancer specific" genes may be used for the identification of DCIS tumors with the highest risk for recurrence and progression. As a first step toward testing this hypothesis, we confirmed by immunohistochemical analysis that the expression of DCD is very rarely detected in DCIS (data not shown). However, testing this hypothesis will require the examination of hundreds of DCIS tumors with long-term clinical follow-up data.

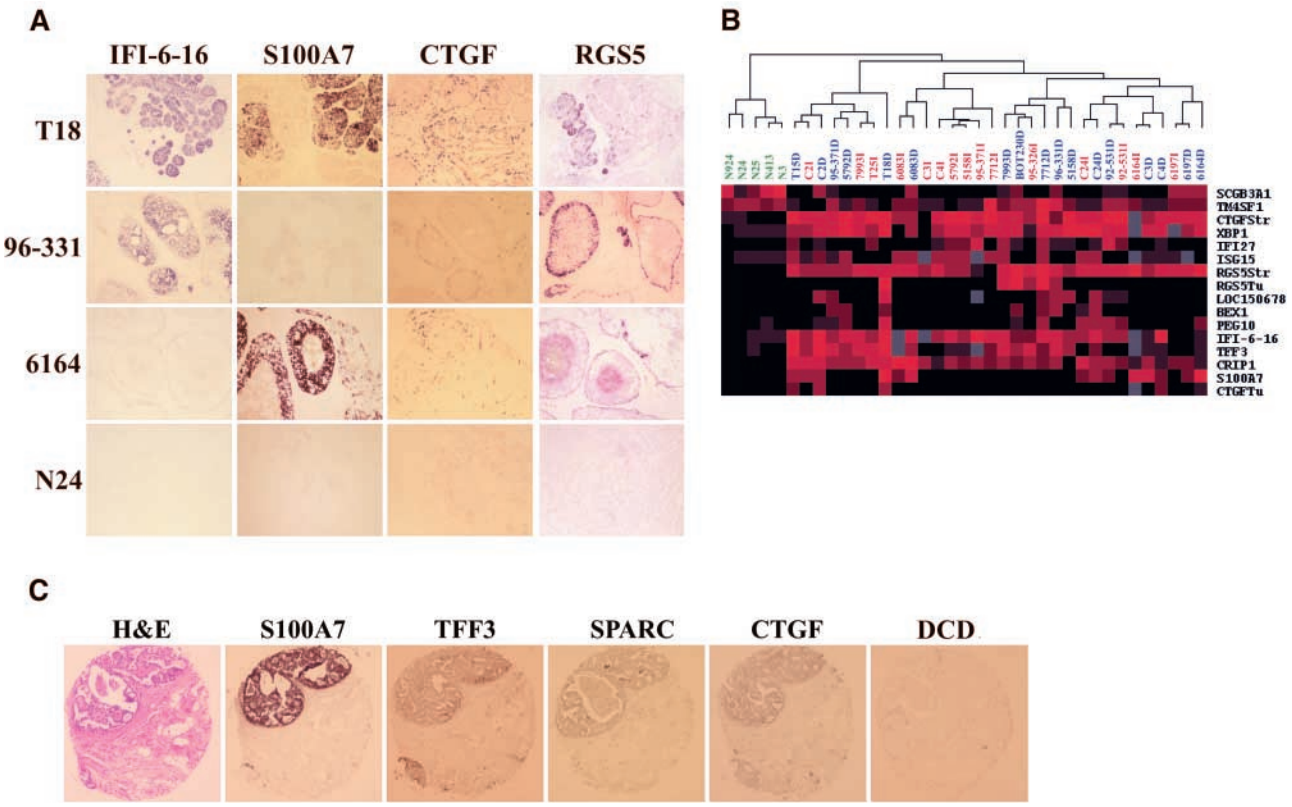

FIGURE 2. Analysis of gene expression in DCIS tumors using mRNA *in situ* hybridization and immunohistochemistry. **A.** Digitonin labeled anti-sense riboprobes corresponding to the indicated genes were hybridized to frozen DCIS tumors as described in "Materials and Methods" to give a dark purple precipitate that corresponds to the presence of the appropriate mRNA. Hybridization with the sense probes gave no signal (data not shown). Strong expression of IFI-6-16 and S100A7 is detected in a subset of DCIS (*T18*, *96-331*, and *6164*), but not normal (*N24*), mammary epithelial cells, while the expression of CTGF and RGS5 is mostly seen in DCIS stromal fibroblasts and myoepithelial cells, respectively. **B.** Dendogram depicting hierarchical clustering of 5 normal, 18 DCIS, and invasive breast tumors based on mRNA *in situ* hybridization results obtained using the indicated 14 genes. Numbers correspond to coded specimen identifiers, "N" denotes normal, "D" DCIS, and "I" invasive breast tissue. Coloring scheme as in Fig. 1. Tu and Str stand for expression in tumor and stromal cells, respectively. **C.** A representative DCIS sample from a tissue microarray composed of DCIS tumors stained with hematoxylin–eosin, or with antibodies against the indicated genes. Immunohistochemistry was performed as described in "Materials and Methods"; black/dark brown precipitate indicates the presence of S100A7, TFF3, SPARC, or CTGF protein. No DCD expression is seen in DCIS.

Table 5. Relationships Between Gene Expression and Histopathological Features of Tumors

	DCIS	Invasive	Metastasis	P[a]	DCIS	Invasive						
					Age ≤50	ER	HER2	Grade 1	Grade 3	Stage 3	Tumor size	4 pos LN
S100A7	23 (37.5)	245 (43.5)	16 (31.4)	0.08	$P = 0.03$	$P = 0.03$[b]	NS	NS	$P < 0.0001$	NS	NS	$P = 0.0008$
FASN	28 (38.9)	126 (51.0)	21 (50.0)	0.2	NS	$P = 0.02$	$P = 0.002$	$P = 0.03$[b]	NS	NS	NS	NS
TFF3	36 (52.2)	196 (77.2)	31 (75.6)	0.0003	NS	$P = 0.02$	NS	NS	NS	NS	NS	NS
CTGF	21 (30.0)	88 (34.7)	5 (12.2)	0.01	NS	NS	NS	NS	NS	NS	NS	NS
SPARC-Tumor	27 (39.1)	136 (50.4)	21 (50.0)	0.25	NS	NS	NS	NS	$P = 0.01$[b]	$P = 0.02$[b]	NS	NS
SPARC-Stroma	63 (87.5)	248 (91.2)	42 (100.0)	0.04	NS	NS	NS	NS	NS	$P = 0.002$[b]	$P = 0.03$	NS
CXCL1 (GRO1)	ND	11 (15.9)	ND	NA	NA	NS	NS	NS	NS	NS	NS	NS
CXCL2 (GRO2)	ND	2 (3.1)	ND	NA	NA	NS	NS	NS	NS	NS	NS	NS
IL-8	ND	5 (7.5)	ND	NA	NA	NS	NS	NS	NS	NS	NS	NS
NFKBIA	ND	46 (93.9)	ND	NA	NA	NS	NS	NS	NS	NS	NS	$P = 0.04$[a]
CCND1	ND	3 (10.7)	ND	NA	NA	NS	NS	NS	NS	NS	NS	NS
CD45	ND	28 (96.6)	ND	NA	NA	NS	NS	NS	NS	NS	NS	NS

Note: Numbers reflect the actual numbers of tumor specimens that were positive for the indicated gene, while the percentage of positive tumors is indicated in brackets. Only data for which there was at least one statistically significant association are listed in the table; see "Materials and Methods" for details.
[a]P is Fisher's exact test P for association between gene expression and tumor category (DCIS, Invasive, or Metastasis), or for the association between NFKBIA expression of positive lymph nodes. All other values of P are likelihood ratio (LR) test P.
[b]Denotes P for inverse correlation.

Although no study identical to ours has been performed using DNA microarrays, several of the genes that were overexpressed in tumors (like trefoil factor 3, X-box binding protein, collagen type I, fibronectin, etc.) were also found to be up-regulated in breast cancer by other groups (5, 6, 27).

In summary, we have identified several genes that could potentially be used for the molecular differential diagnosis of DCIS and invasive breast cancer and for the classification of DCIS tumors. Determining the clinical usefulness of these genes requires further studies.

Materials and Methods

Tissue Samples and Tissue Microarrays

Fresh tissue specimens obtained from the Brigham and Women's Hospital, Massachusetts General Hospital, and Faulkner Hospital (all Boston, MA), Duke University (Durham, NC), University Hospital Zagreb (Zagreb, Croatia), and the National Disease Research Interchange were snap frozen on dry ice and stored at −80°C until use. All human tissue was collected following NIH guidelines and using protocols approved by the Institutional Review Boards. Tumors with significant DCIS component were identified based on pathology report and confirmed by microscopic examination of hematoxylin-eosin-stained frozen sections. Tumors used for SAGE analysis D1, D3, D4, D5, and D6 were high-grade, comedo DCIS, while D2, D7, and T18 were intermediate-grade with no necrosis. Tumors used for mRNA *in situ* hybridization and immunohistochemistry included DCIS cases of all three (low, intermediate, and high grade) histological types. Most of the cases used for *in situ* hybridization and immunohistochemistry were DCIS with concurrent invasive carcinoma and pure DCIS, respectively, while tumors D3 and D6 used for SAGE were pure DCIS. The larger representation of frozen/fresh DCIS cases with concurrent invasive disease was due to logistic issues, because it is extremely difficult to obtain frozen pure DCIS specimens, especially ones with long-term clinical follow-up data. For *in situ* hybridization, 5-μm-thick frozen sections were mounted on silylated slides (CEL Associates Inc., Pearland, TX), air-dried, and stored at −80°C until use. Tissue microarrays were (*a*) obtained from commercial sources; Imgenex, San Diego, CA (49 invasive breast tumor) and Ambion, Austin, TX (92 primary invasive tumors and 41 distant metastasis); (*b*) provided by Cooperative Breast Cancer Tissue Resource (40 normal breast tissue, 20 DCIS-10 pure and 10 with concurrent invasive tumor, 192 primary invasive breast tumors); and (*c*) generated at Johns Hopkins University (299 invasive breast tumors and 10 distant metastasis) and at Beth Israel Deaconess Medical Center (30 invasive breast cancer, and 70 pure DCIS of different histological grades and matched normal breast tissue) following published protocols (11). With the exception of the Imgenex and the DCIS arrays (1 mm punches), all other tissue microarrays (TMAs) had 0.6 mm punches and contained at least two punches/tumor to control for tumor and immunohistochemical staining heterogeneity.

Generation and Analysis of SAGE Libraries

SAGE libraries were generated and analyzed essentially as previously described as part of the National Cancer Institute Cancer Gene Anatomy Project (8, 9, 28, 29). SAGE libraries generated from normal (N1 and N2), two DCIS (D1 and D2), two invasive tumors (I1 and I2), and corresponding lymph node metastases (LN1 and LN2) were previously reported (8). Some of the DCIS tumors were pure DCIS (D3 and D6), while others were derived from patients with concurrent invasive breast carcinomas. Epithelial cells from normal breast tissue (N1 and N2) and tumors (D2, D3, D6, and D7) were purified using BerEP4-coated magnetic beads (Dynal, Oslo, Norway), while other tumors were macroscopically dissected based on adjacent hematoxylin-eosin-stained slides. Approximately 50,000 SAGE tags were obtained from each library. For further analyses, libraries were normalized to the library with the highest tag number (89,541 total tags). Hierarchical clustering was applied to data using the Cluster program developed by Eisen *et al.* (30). Differentially expressed genes were identified based on

statistical analysis of comparisons of groups of normal (two cases), DCIS (eight samples), and invasive or metastatic (nine samples) breast cancer SAGE libraries using the SAGE2000 software (7). Similarly, for the identification of genes specific for DCIS or invasive breast cancer, the eight DCIS samples were treated as a group and the nine invasive or metastatic patients were treated as another group. First we used the SAGE tag numbers highest in two normal libraries (N1 and N2) as the cutoff and evaluated tag numbers in the DCIS and invasive libraries above this ''normal'' value using a two-sided Fisher's exact test without multiple comparisons (see Table 3). In a second test, ROC curve analysis was used to choose the best cutoff for values. A ROC area of 0.50 is no better than chance and a ROC area of 1.00 is the best possible. The bound 2 in the first row means that values greater than or equal to 2 are counted as predicting DCIS.

Messenger RNA in Situ Hybridization

To generate templates for *in vitro* transcription reactions, 300- to 500-bp fragments derived from the 3′ untranslated region of the selected genes were PCR amplified and subcloned into pZERO 1.0 (Invitrogen, Carlsbad, CA). This pZERO 1.0 vector contains a multiple cloning site surrounded by SP6 and T7 RNA polymerase promoters; therefore, the same plasmid was used for the generation of sense and anti-sense riboprobes for mRNA *in situ* hybridizations. Digitonin-labeled sense and anti-sense riboprobes were generated and mRNA *in situ* hybridization was performed as previously described (31). The hybridized sections were observed with a NIKON microscope, images were obtained using a SPOT CCD camera, and processed with Adobe Photoshop. Hybridizations were considered successful if the sense probe gave no significant signal. The intensity and distribution of the hybridization signal were scored (0–3 for intensity and 0–3 for distribution using a scoring scheme described below for immunohistochemistry) independently by three investigators, and scores in Fig. 2 reflect a consensus of the three independent summary scores.

Immunohistochemistry

The expression of the indicated genes in primary breast tumors was analyzed by the use of immunohistochemistry to eight tissue microarrays that contained evaluatable paraffin-embedded specimens derived from 80 DCIS, 675 primary invasive breast cancer, and 33 distant metastases. Antigen Retrieval Citra solution (Research Genetics, San Ramon, CA) and boiling in microwave (5 min at high power) was used to enhance staining. Isotype control antiserum was used as negative control. A standard indirect immunoperoxidase protocol with 3,3′-diaminobenzidine as a chromogen was used for the visualization of antibody binding (ABC-Elite; Vector Laboratories, Burlingame, CA). Primary antibodies used were as follows: mouse monoclonal anti-psoriasin antibody (17); affinity-purified rabbit polyclonal anti-CTGF (a generous gift of Dr. D. Brigstock, Childrens' Research Institute, Columbus, OH); affinity-purified rabbit polyclonal anti-TFF3 (a kind gift of Prof. Hoffman, Universitaetsklinikum, Magdeburg, Germany); mouse monoclonal anti-IL-8, GRO-1, and GRO-2 were purchased from R&D Systems (Minneapolis, MN); anti-SPARC antibody was from Hematologic Technologies (Essex Junction, VT); and anti-FASN antibody from Transduction Labs. (San Diego, CA). Antibodies were used at 1:100 dilution in PBS containing 10% heat-inactivated goat serum. Antibody staining was subjectively scored by three investigators independently in a scale of 0–3 for intensity (0 = no staining, 1 = faint signal, 2 = moderate, and 3 = intense staining) and 0–3 for extent (0 = no, 1 = $\leq$30%, 2 = 30–70%, and 3 = 70% positive cells) of staining. Cumulative scores were calculated based on the average scores assigned by the three independent observers. For statistical analyses, a cumulative score at or above 3 was considered positive. Relationships between the expression of genes determined by mRNA *in situ* hybridization or immunohistochemistry were analyzed by Fisher's exact test without correction for multiple comparisons.

Statistical Analyses of Clinical Correlates

The relationship of gene expression to clinicopathologic parameters and the association between the expression of different genes determined by immunohistochemistry were analyzed by the following statistical methods. The eight separate tissue microarray data sets and one combined data set were analyzed for association of gene positivity and prognostic factors using a logistic regression model (with gene positivity as the outcome), and a forward, or step-up, selection procedure to determine the best fitting model. Clinicopathologic factors analyzed were: expression of the estrogen and progesterone receptors and HER2 by immunohistochemistry, histological grade, TNM stage, tumor size, number of positive lymph nodes, patient age, and overall and distant metastasis free survival. If all patients or no patients with a particular level of a covariate had gene positivity, then the logistic regression did not converge and a significance level was obtained using Fisher's exact test. If, however, there remained some patients with and without gene positivity after deleting patients with the particular level of the covariate, then a step-up logistic regression was done on them. The significance of the variables in the logistic regression models was tested using likelihood ratio tests. The cutoff used for entry into the model was $\alpha = 0.05$. In addition to the analyses described above, Kaplan-Meier curves were generated and Cox models were run for two data sets that contained survival information. Calculated times to distant failure and times to survival were used and were based on the failure/death and accession dates.

Acknowledgments

Special thanks to the National Cancer Institute Cooperative Breast Cancer Tissue Resource for breast tissue microarrays, the National Disease Research Interchange for frozen DCIS specimens, Dr. D. Brigstock (Childrens' Research Institute, Colombus, OH) for anti-CTGF, and Prof. Hoffman (Universitaetsklinikum, Magdeburg, Germany) for anti-TFF3 antibodies provided for this study.

References

1. Walker, R. A., Jones, J. L., Chappell, S., Walsh, T., and Shaw, J. A. Molecular pathology of breast cancer and its application to clinical management. Cancer Metastasis Rev., *16:* 5–27, 1997.

2. Gupta, S. K., Douglas-Jones, A. G., Fenn, N., Morgan, J. M., and Mansel, R. E. The clinical behavior of breast carcinoma is probably determined at the preinvasive stage (ductal carcinoma *in situ*). Cancer, *80:* 1740–1745, 1997.

3. Badve, S., A'Hern, R. P., Ward, A. M., Millis, R. R., Pinder, S. E., Ellis, I. O., Gusterson, B. A., and Sloane, J. P. Prediction of local recurrence of ductal carcinoma *in situ* of the breast using five histological classifications: a comparative study with long follow-up. Hum. Pathol., *29:* 915–923, 1998.

4. Silverstein, M. (ed.). Ductal Carcinoma *in Situ* of the Breast. Baltimore, MD: Williams & Wilkins, 1997.

5. Luzzi, V., Holtschlag, V., and Watson, M. A. Expression profiling of ductal carcinoma *in situ* by laser capture microdissection and high-density oligonucleotide arrays. Am. J. Pathol., *158:* 2005–2010, 2001.

6. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. Molecular portraits of human breast tumours. Nature, *406:* 747–752, 2000.

7. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. Science, *270:* 484–487, 1995.

8. Porter, D. A., Krop, I. E., Nasser, S., Sgroi, D., Kaelin, C. M., Marks, J. R., Riggins, G., and Polyak, K. A SAGE (serial analysis of gene expression) view of breast tumor progression. Cancer Res., *61:* 5697–5702, 2001.

9. Krop, I. E., Sgroi, D., Porter, D. A., Lunetta, K. L., LeVangie, R., Seth, P., Kaelin, C. M., Rhei, E., Bosenberg, M., Schnitt, S., Marks, J. R., Pagon, Z., Belina, D., Razumovic, J., and Polyak, K. HIN-1, a putative cytokine highly expressed in normal but not cancerous mammary epithelial cells. Proc. Natl. Acad. Sci. USA, *98:* 9796–9801, 2001.

10. St Croix, B., Rago, C., Velculescu, V., Traverso, G., Romans, K. E., Montgomery, E., Lal, A., Riggins, G. J., Lengauer, C., Vogelstein, B., and Kinzler, K. W. Genes expressed in human tumor endothelium. Science, *289:* 1197–1202, 2000.

11. Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G., and Kallioniemi, O. P. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat. Med., *4:* 844–847, 1998.

12. Moch, H., Schraml, P., Bubendorf, L., Mirlacher, M., Kononen, J., Gasser, T., Mihatsch, M. J., Kallioniemi, O. P., and Sauter, G. High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma [see comments]. Am. J. Pathol., *154:* 981–986, 1999.

13. Ramaswamy, S., Ross, K. N., Lander, E. S., and Golub, T. R. A molecular signature of metastasis in primary solid tumors. Nat. Genet., *33:* 49–54, 2003.

14. Watson, P. H., Leygue, E. R., and Murphy, L. C. Psoriasin (S100A7). Int. J. Biochem. Cell Biol. *30:* 567–571, 1998.

15. Donato, R. Functional roles of S100 proteins, calcium-binding proteins of the EF- hand type. Biochim. Biophys. Acta, *1450:* 191–231, 1999.

16. Madsen, P., Rasmussen, H. H., Leffers, H., Honore, B., Dejgaard, K., Olsen, E., Kiil, J., Walbum, E., Andersen, A. H., Basse, B., Lauridsen, J. B., Ratz, G. P., Celis, A., Vandekerckhove, J., and Celis, J. E. Molecular cloning, occurrence, and expression of a novel partially secreted protein "psoriasin" that is highly up-regulated in psoriatic skin. J. Invest. Dermatol., *97:* 701–712, 1991.

17. Enerback, C., Porter, D. A., Seth, P., Sgroi, D., Gaudet, J., Weremowicz, S., Morton, C. C., Schnitt, S., Pitts, R. L., Stampl, J., Barnhart, K., and Polyak, K. Psoriasin expression in mammary epithelial cells *in vitro* and *in vivo*. Cancer Res., *62:* 43–47, 2002.

18. van Ruissen, F., Jansen, B. J., de Jongh, G. J., van Vlijmen-Willems, I. M., and Schalkwijk, J. Differential gene expression in premalignant human epidermis revealed by cluster analysis of serial analysis of gene expression (SAGE) libraries. FASEB J., *16:* 246–248, 2002.

19. Sands, B. E. and Podolsky, D. K. The trefoil peptide family. Annu. Rev. Physiol., *58:* 253–273, 1996.

20. Taupin, D. R., Kinoshita, K., and Podolsky, D. K. Intestinal trefoil factor confers colonic epithelial resistance to apoptosis. Proc. Natl. Acad. Sci. USA, *97:* 799–804, 2000.

21. Bratthauer, G. L., Moinfar, F., Stamatakos, M. D., Mezzetti, T. P., Shekitka, K. M., Man, Y. G., and Tavassoli, F. A. Combined E-cadherin and high molecular weight cytokeratin immunoprofile differentiates lobular, ductal, and hybrid mammary intraepithelial neoplasias. Hum. Pathol., *33:* 620–627, 2002.

22. Diez-Itza, I., Vizoso, F., Merino, A. M., Sanchez, L. M., Tolivia, J., Fernandez, J., Ruibal, A., and Lopez-Otin, C. Expression and prognostic significance of apolipoprotein D in breast cancer. Am. J. Pathol., *144:* 310–320, 1994.

23. Harding, C., Osundeko, O., Tetlow, L., Faragher, E. B., Howell, A., and Bundred, N. J. Hormonally-regulated proteins in breast secretions are markers of target organ sensitivity. Br. J. Cancer, *82:* 354–360, 2000.

24. Schittek, B., Hipfel, R., Sauer, B., Bauer, J., Kalbacher, H., Stevanovic, S., Schirle, M., Schroeder, K., Blin, N., Meier, F., Rassner, G., and Garbe, C. Dermcidin: a novel human antibiotic peptide secreted by sweat glands. Nat. Immunol., *2:* 1133–1137, 2001.

25. Mehul, B., Bernard, D., Simonetti, L., Bernard, M. A., and Schmidt, R. Identification and cloning of a new calmodulin-like protein from human epidermis. J. Biol. Chem., *275:* 12841–12847, 2000.

26. Colantuoni, V., Cortese, R., Nilsson, M., Lundvall, J., Bavik, C. O., Eriksson, U., Peterson, P. A., and Sundelin, J. Cloning and sequencing of a full length cDNA corresponding to human cellular retinol-binding protein. Biochem. Biophys. Res. Commun., *130:* 431–439, 1985.

27. Sgroi, D. C., Teng, S., Robinson, G., LeVangie, R., Hudson, J. R., and Elkahloun, A. G. *In vivo* gene expression profile analysis of human breast cancer progression. Cancer Res., *59:* 5656–5661, 1999.

28. Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L., and Riggins, G. J. A public database for gene expression in human cancers. Cancer Res., *59:* 5403–5407, 1999.

29. Boon, K., Osorio, E. C., Greenhut, S. F., Schaefer, C. F., Shoemaker, J., Polyak, K., Morin, P. J., Buetow, K. H., Strausberg, R. L., De Souza, S. J., and Riggins, G. J. An anatomy of normal and malignant gene expression. Proc. Natl. Acad. Sci. USA, *99:* 11287–11292, 2002.

30. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, *95:* 14863–14868, 1998.

31. Qian, Y., Fritzsch, B., Shirasawa, S., Chen, C. L., Choi, Y., and Ma, Q. Formation of brainstem (nor)adrenergic centers and first-order relay visceral sensory neurons is dependent on homeodomain protein Rnx/Tlx3. Genes Dev., *15:* 2533–2545, 2001.