

## **Classification of Breast Cancer Using Genetic Algorithms and Tissue Microarrays**

Marisa Dolled-Filhart, Lisa Rydén, Melissa Cregger, Karin Jirström, Malini Harigopal  
Robert L. Camp, and David L. Rimm

**Abstract Purpose:** A multitude of breast cancer mRNA profiling studies has stratified breast cancer and defined gene sets that correlate with outcome. However, the number of genes used to predict patient outcome or define tumor subtypes by RNA expression studies is variable, nonoverlapping, and generally requires specialized technologies that are beyond those used in the routine pathology laboratory. It would be ideal if the familiarity and streamlined nature of immunohistochemistry could be combined with the rigorously quantitative and highly specific properties of nucleic acid – based analysis to predict patient outcome.

**Experimental Design:** We have used AQUA-based objective quantitative analysis of tissue microarrays toward the goal of discovery of a minimal number of markers with maximal prognostic or predictive value that can be applied to the conventional formalin-fixed, paraffin-embedded tissue section.

**Results:** The minimal discovered multiplexed set of tissue biomarkers was GATA3, NAT1, and estrogen receptor. Genetic algorithms were then applied after division of our cohort into a training set of 223 breast cancer patients to discover a prospectively applicable solution that can define a subset of patients with 5-year survival of 96%. This algorithm was then validated on an internal validation set ( $n = 223$ , 5-year survival = 95.8%) and further validated on an independent cohort from Sweden, which showed 5-year survival of 92.7% ( $n = 149$ ).

**Conclusions:** With further validation, this test has both the familiarity and specificity for widespread use in management of breast cancer. More generally, this work illustrates the potential for multiplexed biomarker discovery on the tissue microarray platform.

A multitude of breast cancer mRNA profiling studies has stratified breast cancer and defined gene sets that correlate with outcome (1). These studies have resulted in plans for prospective application of nucleic acid–based tests to select patients that do not need further therapy after their primary resection (2, 3). However, the number of genes used to predict patient outcome or define tumor subtypes by RNA

expression studies is variable, nonoverlapping, and generally requires specialized technologies. Immunohistochemical studies can be done with many fewer markers (4, 5), but suffer from the inherent flaw of subjective analysis and variable reproducibility. Thus, it would be ideal if the familiarity and streamlined nature of immunohistochemistry could be combined with the rigorously quantitative and highly specific properties of nucleic acid–based analysis to predict patient outcome. The AQUA method has the potential to combine the best properties of each of these assays. It is similar to immunohistochemistry in its format and slide preparation, but it allows rigorously quantitative and reproducible assessment of protein expression. These results require a specialized slide-reading platform that is commercially available from HistoRx (New Haven, CT).

Tissue microarrays provide a method for high-throughput protein expression analysis of large cohorts of cancer patients on a single slide (6) that results in standardization of many variables and capacity for embedded controls (reviewed in ref. 7). This technology has been slow to be used for discovery because the analysis of these arrays has generally been subjective, thus invalidating many of the algorithms used for discovery in nucleic acid array experiments. The recently developed automated quantitative analysis technology allows for immunofluorescent quantification reflecting the biological continuum of protein expression (8). Studies from our laboratory have shown that AQUA scores are highly correlated

**Authors' Affiliation:** Department of Pathology, Yale University School of Medicine, New Haven, Connecticut

Received 6/7/06; revised 7/27/06; accepted 8/15/06.

**Grant support:** Patrick and Catherine Weldon Donaghue Foundation for Medical Research, NIH, National Cancer Institute grant R21 CA100825, and U.S. Army grant DAMD-17-02-0463 (D.L. Rimm); grants from Stig and Ragna Gorthon Foundation and the Swedish Cancer Society (L. Rydén); U.S. Army Breast Cancer Research grant DAMD17-03-1-0349 (M. Dolled-Filhart); NIH grant K0-8 ES11571 (R.L. Camp); and the Breast Cancer Alliance of Greenwich, CT (D.L. Rimm and R.L. Camp).

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.

**Note:** Supplementary data for this article are available at Clinical Cancer Research Online (<http://clincancerres.aacrjournals.org/>).

Dr. Rimm is a founder, stockholder and consultant to HistoRx, the exclusive licensee of the Yale owned AQUA patent.

**Requests for reprints:** David L. Rimm, Department of Pathology, Yale University School of Medicine, 310 Cedar Street, P.O. Box 208023, New Haven, CT 06520-8023. Phone: 203-737-4204; Fax: 203-737-5089; E-mail: david.rimm@yale.edu.

© 2006 American Association for Cancer Research.

doi:10.1158/1078-0432.CCR-06-1383

**Table 1.** Antibody staining conditions, sources, and references of microarray studies

Marker	Antibody source	Species	Dilution	Time	References*
ACADSB	Gift from Gerry Vockley (53)	Rabbit polyclonal	1:5,000	1 h	(10, 11, 43, 54)
AGR2	Gift from Devon Thompson (55)	Rabbit polyclonal	1:1,000	1 h	(43, 56–58)
BCAS1	Gift from Kevin Brown (59)	Rabbit polyclonal	1:500	1 h	(43, 60)
BCL2	DAKO, clone 124	Mouse monoclonal	1:400	1 h	(2, 10, 43, 54, 61, 62)
CA12	Gift of William Sly (63)	Rabbit polyclonal	1:2,000	30 min	(43, 64–66)
CAV1	Transduction Labs, clone 2297	Mouse monoclonal	1:100	O/N	(11, 67, 68)
CDH3	BD Transduction Labs clone 56	Mouse monoclonal	1:200	O/N	(10, 11, 58, 69)
COX6C	Molecular Probes clone 3G5	Mouse monoclonal	1:100	O/N	(11, 43, 56, 62, 65, 69)
CTSD	DAKO	Rabbit polyclonal	1:1,000	1 h	(43, 56, 61, 70, 71)
ESR1	DAKO clone 1D5	Mouse monoclonal	1:50	1 h	(10, 11, 43, 54, 61, 62, 69, 72)
GATA3	Santa Cruz, clone HG3-31	Mouse monoclonal	1:100	1 h	(10, 11, 43, 57, 61, 62, 66, 68, 69, 73)
GGH	Gift from Thomas J. Ryan (74)	Rabbit polyclonal	1:400	1 h	(11, 43)
GLUL	BD Transduction Labs clone 6	Mouse monoclonal	1:1,000	O/N	(43, 62, 65)
GSTP1	DAKO clone 353-10	Mouse monoclonal	1:50	1 h	(61, 62, 68, 75, 76)
HER2	DAKO	Rabbit polyclonal	1:8,000	1 h	(10, 11, 58, 61, 75)
HSP27	Neomarkers clone Ab-1 G3.1	Mouse monoclonal	1:50	30 min	(43, 58)
IGFBP2	Santa Cruz	Goat polyclonal	1:2,000	30 min	(11, 69, 72)
IGFBP4	Austral Biologicals	Mouse monoclonal	1:50	1 h	(43, 56, 62, 64, 71, 77, 78)
IGFBP5	Austral Biologicals	Mouse monoclonal	1:100	1 h	(11, 43, 56, 68, 72, 77, 78)
IRAK1	Santa Cruz	Rabbit polyclonal	1:100	1 h	(54, 66)
KRT7	DAKO clone TL 12/30	Mouse monoclonal	1:50	1 h	(10, 11, 54, 69, 79)
KRT8	DAKO clone 25BH11	Mouse monoclonal	1:100	1 h	(11, 62, 80)
KRT18	DAKO clone DC10	Mouse monoclonal	1:50	1 h	(10, 43, 62, 68, 79)
MUC1	Novocastra	Mouse monoclonal	1:100	O/N	(11, 54, 56, 57, 61, 75, 80)
MYC	DAKO clone 1D5	Mouse monoclonal	1:200	1 h	(11, 54, 56, 61, 64)
NAT1	Gift from Edith Sim (81)	Rabbit polyclonal	1:1,000	1 h	(10, 11, 43, 62, 82)
PFK	Gift from George Dunaway (83)	Rabbit polyclonal	1:2,000	1 h	(43, 54, 56, 66, 69)
SERPINA3	DAKO	Rabbit polyclonal	1:3,200	10 min	(11, 56, 62, 65, 76–78)
SLC7A5	Serotec	Rabbit polyclonal	1:50	1 h	(54, 56, 67, 69, 77, 78, 82)
SLC9A3R1	Gift from Vijaya Ramesh (84)	Rabbit polyclonal	1:50	O/N	(43, 54, 56, 57, 69)
TFF1	DAKO clone BC04	Mouse monoclonal	1:5,000	O/N	(56, 57, 62, 67, 80, 82)
TFF3	Gift from Daniel Podolsky (85)	Rabbit polyclonal	1:500	1 h	(10, 11, 43, 54, 62, 69, 79, 80)
THBS1	Neomarkers clone A6.1 Ab4	Mouse monoclonal	1:50	O/N	(61, 68, 69, 71, 75)
TIMP3	Oncogene Research 136-13H4	Mouse monoclonal	1:50	O/N	(11, 54, 62, 69, 82)
XBP1	Santa Cruz	Rabbit polyclonal	1:200	O/N	(10, 11, 43, 54, 57, 61, 62, 64–66)

\*This includes references in which the marker was identified in a breast cancer microarray expression/profiling study.

to protein content as measured by ELISA (9).<sup>1</sup> Thus, the combination of tissue microarrays and AQUA provides follow-up and extension of technology for cDNA microarray-based target discovery by allowing an *in situ* protein assay of markers of interest on large cohorts of tumors with the inclusion of spatial subcellular localization information and multiplexed analysis.

The oldest and most valuable prognostic and predictive marker in breast cancer is the estrogen receptor (ER). Estrogen not only regulates differentiation and growth in the mammary gland, but it is important in breast cancer development and progression. A multitude of studies have investigated the relationship between gene expression in ER-positive and ER-negative breast cancers to generate a list of genes that are able to discriminate between the two classes. There is overlap between ER status discriminating genes and estrogen-responsive genes, but the sets are substantially different. These studies, and the results of the subtype analyses (10–12), were the basis for our selection of 35 ER and estrogen-related markers as a starting

point for the tissue microarray (TMA) marker discovery studies toward classification and prognostication of breast cancer (further justification of marker selection is discussed in the Materials and Methods section and in Table 1).

## Materials and Methods

**Tissue microarray construction.** Formalin-fixed, paraffin-embedded breast cancer tumors from the archives of the Yale University Department of Pathology were used in the construction of the tissue microarrays. Briefly, representative areas of invasive ductal carcinoma were selected from whole tissue H&E sections by pathologists (R.L.C., D.L.R.). Cores sized 0.6 mm in diameter were spaced 0.8 mm apart in a grid layout using a manual Tissue Microarrayer (Beecher Instruments, Silver Spring, MD). The resulting tissue microarray blocks were cut to 5- $\mu$ m sections with a microtome, placed on slides with an adhesive tape-transfer method (Instrumedics, Inc., Hackensack, NJ), and UV cross-linked.

**Cohort information.** The Yale breast cancer cohort consists of 688 samples of invasive ductal carcinoma selected from the Yale University Department of Pathology archives as available from 1961 to 1983. The mean follow-up time of this cohort is 12.8 years and the mean age of diagnosis is 58.1 years. The median follow-up time is 8.9 years and the median age of diagnosis is 58.0 years. Of the 656 patients with follow-up time, 328 were censored at 20 years and 276 were uncensored at

<sup>1</sup> Dolled-Filhart M, McCabe A, Giltneane J, Cregger M, Camp RL, Rimm DL. Quantitative *in situ* analysis of  $\beta$ -catenin expression in breast cancer shows decreased expression is associated with poor outcome. *Cancer Res* 2006;66:5487–94.

20 years. Of the 328 censored patients, their median follow-up was 21.4 years, with the minimum at 4.2 months. This cohort contains approximately half node-positive specimens and half node-negative specimens, which have been used previously in other studies from our laboratory (9, 13–18). Complete clinical treatment information was not available for this cohort; however, some limited information was available for the node-positive patients. Most node-negative patients were treated with local radiation, and none received Herceptin. About 15% of the node-positive patients were treated with chemotherapy (primarily Adriamycin, cytoxan, and 5-fluorouracil), and subsequently ~27% received tamoxifen (post-1978). The node-negative patients were largely treated only by surgical resection.

The Swedish breast cancer cohort consists of 564 premenopausal patients with invasive breast cancer, Unio Internationale Contra Cancrum stage II, enrolled in a randomized prospective clinical trial from 1986 to 1991 at two centers in Sweden (19). The patients were randomized to receive adjuvant tamoxifen ( $n = 276$ ) for 2 years or no adjuvant treatment ( $n = 288$ ) after the primary surgical treatment. Locoregional radiotherapy was delivered to all node-positive patients. Less than 2% of all of the patients received other adjuvant treatment. The median follow-up time was 13.9 years. The arms were well balanced, both having a median age of 45 years old. Between 27% and 30% of each arm showed no nodal metastasis at time of diagnosis.

The analysis of the 35 markers was all done on a 250-case cohort (9) of half node-positive and half node-negative specimens; the associations with clinical and pathologic variables with patient outcome are displayed in Table S1A. There were 105 classified as ER negative (44.9%) and 129 cases classified as ER positive (55.1%) for those tumors in which the clinical ER status (as determined by immunohistochemical analysis) were available. Previous analysis from our laboratory has shown a high correlation between our AQUA scores and pathologist-designated ER status (8). A larger cohort, used for further analysis of ER, GATA3, and NAT1, consists of a total of 675 tumors, including cores from the 250-case tissue microarray tumors, with the additional samples consisting of 193 ER-negative cases (48.3%) and 207 ER-positive cases (51.8%) for those cases with available ER status information. The associations with clinical and pathologic variables with patient outcome for the entire cohort are displayed in Supplementary Table S1B. This entire cohort contains approximately half node-positive specimens and half node-negative specimens, and has also been used previously in other studies (8, 13, 14, 16, 18, 20, 21) and is described in detail in Supplementary Data. The cohort used for validation is from the control arm of a study from a randomized trial of tamoxifen conducted in Sweden from 1986 to 1991 (19). Details of this cohort are described in Supplementary Data. In all cohorts, tissue microarray spots without sufficient breast tumor epithelium were excluded from the analysis.

**Description of selection of genes.** We mined the published results (and where possible, the full data sets) of 30 breast cancer RNA expression profiling studies to identify an overall list of candidate biomarkers to screen using tissue microarrays and AQUA analysis. The list of biomarkers were prioritized based on a combination of the following information: (a) appearance in multiple independent breast cancer profiling studies (references listed in Table 1); (b) classification strength/ranking of biomarkers in each study; (c) mapping to regions of breast cancer genetic alterations, amplification, or deletions based on a breast cancer cytogenetic meta-analysis of 15 papers that studied an overall total of 644 breast cancer tumors and/or cell lines (22–36); and (d) whether an antibody was available (either commercially or from individual laboratories) with validation in the literature by Western blotting and/or successful immunohistochemical analyses. This list was then further focused to only those who appeared in estrogen-related studies and that were successfully used for immunofluorescence, resulting in 35 markers successfully screened on tissue microarrays by AQUA in this study.

**Immunofluorescence staining.** The tissue microarrays were deparaffinized by two xylene rinses of 30 minutes each followed by two rinses

with 100% ethanol for 1 minute each and a rinse in water. Antigen retrieval was done by boiling the slides in a pressure cooker in a sodium citrate buffer at a pH of 6.0. After rinsing briefly in  $1 \times$  TBS, a 30-minute incubation with 2.5% hydrogen peroxide/methanol was used to block endogenous peroxidases. To reduce nonspecific background staining, slides were incubated with 0.3% bovine serum albumin/ $1 \times$  TBS for 1 hour at room temperature, followed by a series of 2-minute rinses in  $1 \times$  TBS,  $1 \times$  TBS/0.01% Triton,  $1 \times$  TBS (TBS washes). Slides were incubated overnight at 4°C with either a monoclonal mouse anti-cytokeratin antibody (clone AE1/AE3; DAKO, Carpinteria, CA; 1:100) when using a rabbit or goat target antibody, or a rabbit anti-cytokeratin antibody (1:100; DAKO) when using a mouse target antibody. The sources, dilutions, and incubation times of the target antibodies are listed in Supplementary Table S1. Antibodies were selected that had either been generated and validated by individual laboratories, or were well-characterized antibodies frequently used for immunohistochemistry in the literature. Slides were washed in  $1 \times$  TBS rinses and incubated with secondary antibodies for 1 hour at room temperature as follows: Alexa 488 goat anti-mouse or Alexa 488 goat anti-rabbit (1:100, Molecular Probes, Eugene, OR) for detecting cytokeratin, and species-specific horseradish peroxidase with a dextran-polymer backbone (Envision, DAKO) along with 6-diamidino-2-phenylindole (1:100, DAKO) for visualization of nuclei. For goat primary antibody detection, slides were incubated for 1 hour with biotinylated anti-goat (1:200, Vector, Burlingame, CA) and Cy-2-donkey anti-mouse (1:50, The Jackson Laboratory, Bar Harbor, ME). This was followed by a 1-hour incubation with streptavidin horseradish peroxidase (1:200, Perkin-Elmer, Wellesley, MA). All slides were washed with TBS rinses followed by a 10-minute incubation with Cy-5 tyramide (1:50 dilution in Amplification Diluent, Perkin-Elmer). The slides were mounted in 0.6% *n*-propyl gallate (an antifade mounting medium) and coverslipped. Examples of immunofluorescent staining are shown in Fig. 1.

**AQUA analysis.** AQUA software linked to a fluorescence microscopy system allowed for quantification of the protein of interest within the tumor region of each tissue microarray core, as described in detail by Camp et al. (8). Image acquisition begins with a low-resolution ( $64 \times 64$  pixels) image capture with a  $\times 4$  objective. The rows and columns of the tissue microarray are defined to form a grid on which the histospots are placed. Monochromatic, high-resolution ( $1,024 \times 1,024$  pixels), in-focus and out-of-focus images are then captured for each relevant wavelength for each histospot with a  $\times 10$  objective. The pan-cytokeratin antibody separates the epithelial breast cancer tumor component by binary masking from the surrounding stroma, and other fluorescent tags designate subcellular compartments (i.e., 6-diamidino-2-phenylindole for nuclei). Coalescence of cytokeratin staining was used to define the nonnuclear compartment for analysis of cytoplasmic and/or membranous staining. The images of the markers were captured at the Cy-5 wavelength because it is outside the range of tissue autofluorescence. Two algorithms, rapid exponential subtraction algorithm and pixel-based locale assignment for compartmentalization of expression, were then used for AQUA analysis. RESA improves subcellular compartment assignment by taking into account overlapping of cells due to the thickness of the tissue microarray section on the slide by subtracting the in-focus information from the out-of-focus information. The second algorithm, pixel-based locale assignment for compartmentalization of expression, assigns target staining in Cy-5 to the predefined subcellular compartments and quantifies its intensity within each compartment to give a quantitative measurement of the expression level for each histospot. The resulting AQUA scores are the measurements of the biomarker pixel intensity within a compartment divided by the total area of tumor (to normalize for differences in tumor area in each spot), and were used as their raw values in the statistical analyses described below, unless otherwise noted. Tumor samples with more than one AQUA score for a given case were averaged.

**Unsupervised hierarchical clustering.** Z-score transformation (37) was used to normalize between experiments to perform clustering

analyses with this formula:  $[(\text{AQUA score}) - (\text{mean of AQUA scores on tissue microarray } X)] / \text{SD}$ , where  $X$  is a given tissue microarray experiment assessed for a single marker. This has previously been used for analysis of AQUA scores (38).<sup>1</sup> CLUSTER (39) was used to perform unsupervised average linkage hierarchical clustering on unweighted Z-score-transformed AQUA data and visualized with TREEVIEW (39).

**Genetic algorithm for generation of a multiplex marker assay.** A genetic algorithm was used to develop a multiplex marker assay for separation of prognostic groups based on expression of ER, GATA3, and NAT. A training set ( $n = 223$ ) and validation set ( $n = 223$ ) were randomly assigned to generate two groupings with equivalent baseline survival curves for all patients with values for all three markers and both censor and follow-up information. Data for each marker was normalized to control for differences in exposure time based on linear regression of redundant cases.

The genetic algorithm was used to determine the optimal cutpoints for the three selected biomarkers (40, 41). We generated software to produce random "solutions," each containing three algorithms (one each for ER, GATA3, and NAT1), where each algorithm represents an equation that is a true/false statement. For each marker, a value is either greater than or less than or equal to a randomly selected cutpoint. The solutions consisted of three if/then statements that can be calculated as

true or false. Each tumor was then classified on the basis of the number of true conditions where they are given one point for each criteria that is true, or zero points when false. Thus, each tumor has a final genetic algorithm solution score between 0 and 3. These scores are used to group the tumors as follows; group A (3 points), group B (2 points), group C (1 point), and group D (zero points). Then, the prognostic value of a solution was determined by assessing the risk ratio for disease-specific death of the D group (high risk) compared with the A group (low risk). The process is then repeated with new random cutpoint values for each marker and the new solution is compared with the previous solution in an iterative manner—selecting the solution with the highest risk ratio and discarding other solutions until the model converged onto a maximum predictive value (over 20 million iterations). As a further limitation, solutions were only evaluated if they resulted in at least 10% of the training set being assigned into the D group and at least 10% assigned into the A group (to prevent finding groups of less than 20 patients each in the low and high risk groups). The genetic algorithm resulted in a model with the best solution (i.e., highest relative risk) on the training set for stratifying 0-scoring tumors from 3-scoring tumors as follows: ER > 4.6, GATA3 > 33.8, NAT1 > 18.6. The tumors in the training set were assigned into one of the four groups: high risk (score of 0, group D), intermediate risk 1 (score = 1, group C), intermediate risk 2 (score = 2, group B), or low risk (score = 3, group A). This model was applied to the validation set.

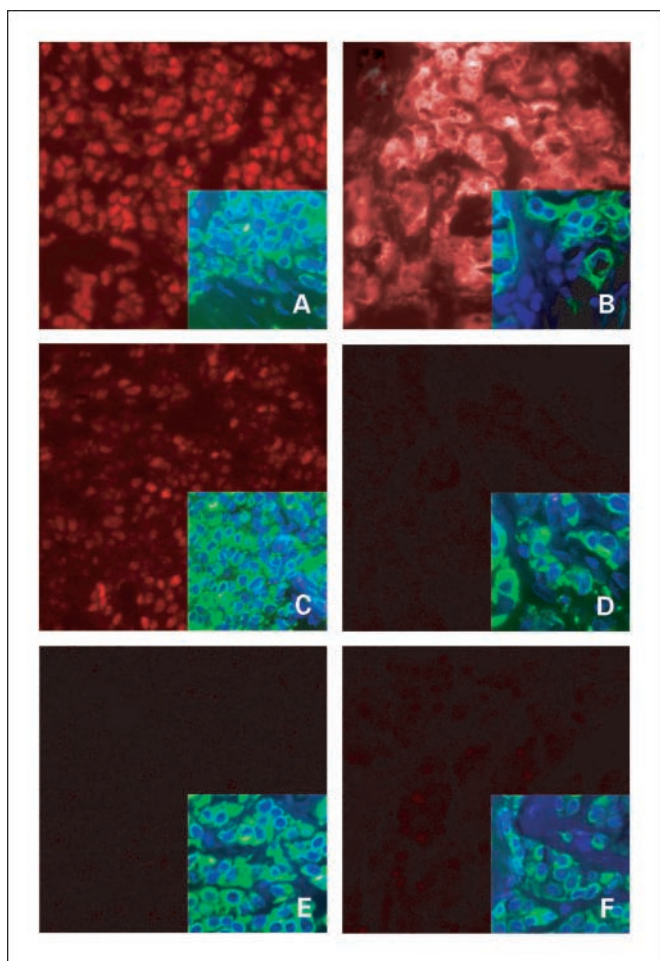
**Statistical analyses.** Statview 5.0.1 (SAS Institute, Inc., Cary, NC) was used for the Cox proportional hazards analysis. Clusters from the unsupervised hierarchical clustering analyses and groups from the multiplex marker assay were assessed for relationship to patient prognosis by Kaplan-Meier survival curves with their significance analyzed by the Mantel-Cox log-rank test with Statview 5.0.1 or SPSS v12 (SPSS, Inc., Chicago, IL).

## Results

We obtained antibodies to 35 markers that were either well-characterized commercial antibodies or well-studied antibodies from individual laboratories (Table 1). We then assessed a 250-case breast cancer cohort by AQUA to quantify the expression of each marker in the epithelial tumor regions of the tissue microarray histospots. The resulting scores are measurements of the biomarker pixel intensity within a compartment divided by the total area of epithelium to account for differences in epithelial area in each spot (see ref. 8 for complete explanation of AQUA technology). The generation of continuous data for each marker allowed us to objectively assess protein expression and to use the data in rigorous algorithmic analyses.

To further analyze each marker, we used the survival and disease-specific death information for the cohort to investigate the relationship of each marker with prognosis by univariate Cox proportional hazards (Table 2). Six of the 35 markers were significant when assessed by their continuous protein expression levels for relationship with 5-year disease-specific survival, including BCL2 ( $P = 0.0400$ ), COX6C ( $P = 0.0049$ ), ER ( $P = 0.0115$ ), GATA3 ( $P = 0.0011$ ), HER2 ( $P = 0.0048$ ), and NAT1 ( $P = 0.0206$ ). Examples of immunofluorescence staining are shown in Fig. 1 for high protein expression of ER, NAT1, and GATA3 (Fig. 1A, B, and C, respectively), and for very low expression of ER, NAT1, and GATA3 (Fig. 1C, D, and E, respectively).

Although the individual prognostic value of markers is important, the classification strength from array studies derives from the power of multiplexing. We began by using the hierarchical clustering and visualization of a heat map to provide a context for exploration of the biological coexpression.



**Fig. 1.** Immunofluorescent staining of breast cancer tissue microarray tumors. Images of tissue microarray breast tumor cores are shown at  $\times 40$  magnification, with the marker staining (red), and cytokeratin staining (green) and 6-diamidino-2-phenylindole staining (blue) in the inset to indicate the presence of tumor in that area in which we show marker expression. ER staining as an example of high expression (A) and low expression (D). High (B) and low (E) expression of NAT1, and high (C) and low expression of GATA3 (F).

**Table 2.** Cox univariate analysis of breast cancer samples with continuous variables on a 250-case cohort (5-year disease-free survival analysis)

Marker	Cox P	Relative risk and 95% confidence interval*
ACADSB cytoplasmic	0.3607	0.988 (0.961-1.104)
AGR2 cytoplasmic	0.8775	0.999 (0.991-1.008)
BCAS1 cytoplasmic	0.2940	1.010 (0.991-1.029)
<b>BCL2 cytoplasmic</b>	<b>0.0400</b>	<b>0.975 (0.952-0.999)</b>
CA12 membranous	0.0621	1.011 (0.999-1.023)
CAV1 membranous	0.7030	0.996 (0.977-1.016)
CDH3 membranous	0.1272	1.012 (0.997-1.027)
<b>COX6C cytoplasmic</b>	<b>0.0049</b>	<b>1.029 (1.009-1.050)</b>
CTSD cytoplasmic	0.7265	0.997 (0.983-1.012)
<b>ER nuclear</b>	<b>0.0115</b>	<b>0.983 (0.969-0.996)</b>
<b>GATA3 nuclear</b>	<b>0.0011</b>	<b>0.949 (0.920-0.979)</b>
GGH cytoplasmic	0.7840	0.998 (0.980-1.015)
GLUL cytoplasmic	0.1192	0.993 (0.983-1.002)
GSTP1 nuclear	0.3406	1.005 (0.995-1.015)
<b>HER2 membranous</b>	<b>0.0048</b>	<b>1.004 (1.001-1.006)</b>
HSP27 cytoplasmic	0.0638	0.990 (0.980-1.001)
IGFBP2 cytoplasmic	0.8729	0.999 (0.986-1.012)
IGFBP4 cytoplasmic	0.2655	0.984 (0.956-1.013)
IGFBP5 cytoplasmic	0.1227	1.027 (0.993-1.063)
IRAK1 cytoplasmic	0.1376	1.012 (0.996-1.028)
KRT18 cytoplasmic	0.1910	0.993 (0.982-1.004)
KRT7 cytoplasmic	0.3000	1.004 (0.997-1.011)
KRT8 cytoplasmic	0.8989	0.999 (0.989-1.010)
MUC1 cytoplasmic	0.6845	0.998 (0.985-1.008)
MYC nuclear	0.7628	1.004 (0.976-1.034)
<b>NAT1 cytoplasmic</b>	<b>0.0206</b>	<b>0.986 (0.974-0.988)</b>
PFK cytoplasmic	0.0500	1.026 (1.000-1.052)
SERPINA3 cytoplasmic	0.9325	1.001 (0.987-1.015)
SLC7A5 membranous	0.4607	0.992 (0.970-1.014)
SLC9A3R1 cytoplasmic	0.1597	0.979 (0.951-1.008)
TFF1 cytoplasmic	0.8775	0.998 (0.969-1.027)
TFF3 cytoplasmic	0.3938	0.989 (0.965-1.014)
THBS1 cytoplasmic	0.9793	0.999 (0.950-1.051)
TIMP3 cytoplasmic	0.0675	1.010 (0.999-1.021)
XBP1 cytoplasmic	0.7146	1.004 (0.985-1.023)

NOTE: Bold indicates *P* values <0.05; associated relative risks are also in bold type.

\*Relative risk calculations are based on the continuous measurements.

Unsupervised average linkage clustering was applied to this set of markers as shown in the heat map in Fig. 2A. The cases shown include those with data for at least 80% of the markers to allow inclusion of a larger number of samples due to missing data. The total number of tumors clustered is 151 due to missing values. ER is in a small cluster with two other markers, GATA3 and NAT1. The larger cluster containing the GATA3, NAT1, and ER grouping also includes BCL2, HSP27, SLC9A3R1, and IGFBP4. When looking at the results of linear regression analysis for these markers individually with ER protein expression levels, they had statistically significant direct relationships (i.e., higher ER levels corresponding to higher marker levels) as follows: BCL2 ( $R = 0.423$ ,  $P < 0.0001$ ), GATA3 ( $R = 0.497$ ,  $P < 0.0001$ ), HSP27 ( $R = 0.327$ ,  $P < 0.0001$ ), IGFBP4 ( $R = 0.203$ ,  $P = 0.0055$ ), NAT1 ( $R = 0.487$ ,  $P < 0.0001$ ), and SLC9A3R1 ( $R = 0.187$ ,  $P = 0.0182$ ). There was a noticeable grouping of markers (CDH3, KRT7, GGH, and HER2) away from the rest of the markers in the heat map that generally have indirect relationships with ER levels (i.e., higher ER levels correspond to

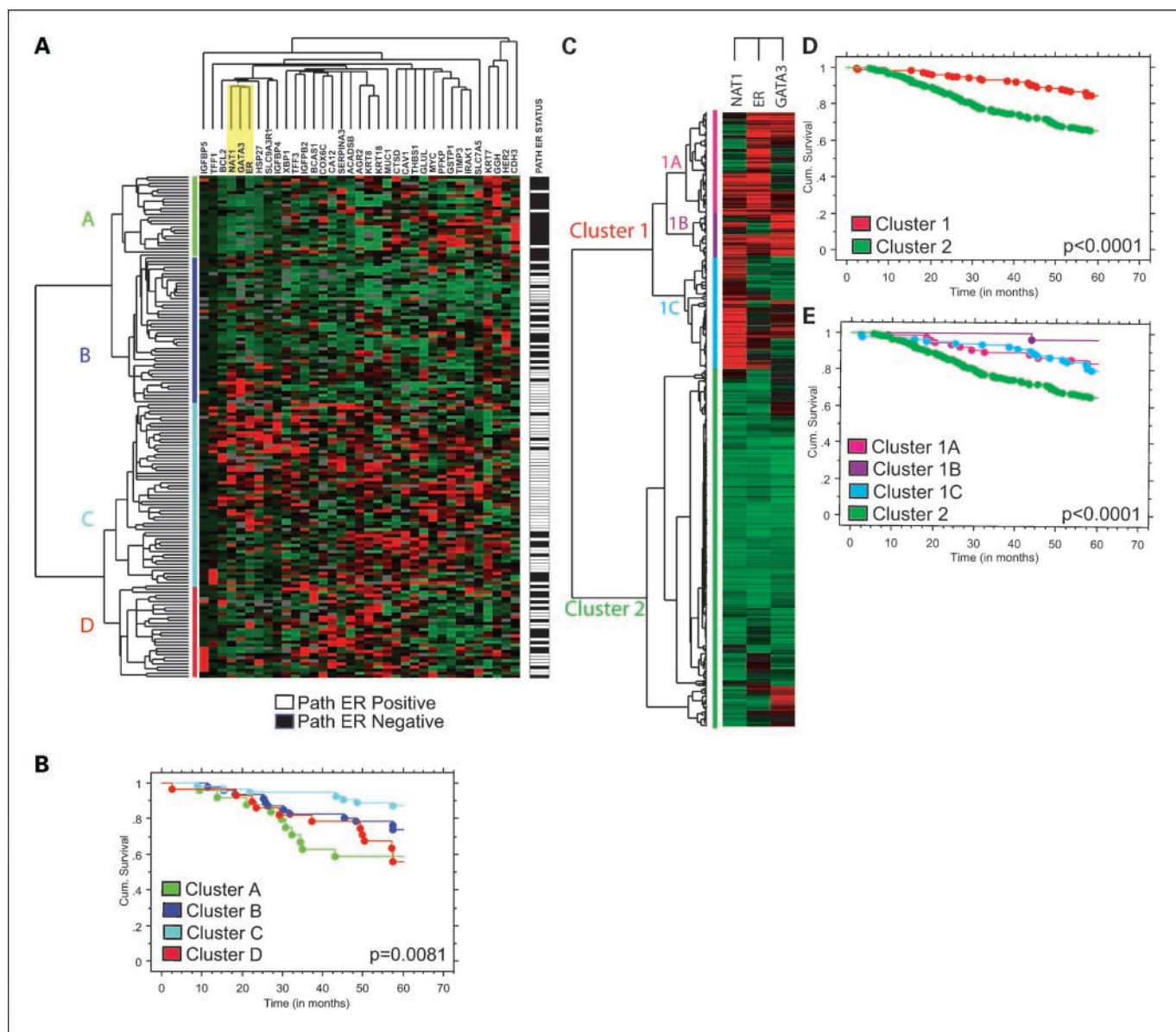
lower levels). Also to be noted is the close relationship in the clustering analysis of keratin 8 and 18, which are coexpressed (42).

The cluster with the highest ER levels, cluster C, had the best prognosis. Cluster A, with the lowest ER levels, had a poor prognosis (87.5% and 60.0% 5-year survival, respectively; Fig. 2A and B). When comparing the cluster groupings with traditional immunohistochemically determined ER status (>10% nuclei staining), there were differences in ER status positivity between the groupings. Cluster C consists of 81% ER-positive tumors, cluster B with 56% ER-positive tumors, cluster D with 53% ER-positive tumors, and cluster A with 12% ER-positive tumors. This analysis shows that multiplex expression analysis can reveal prognostic classes that are substantially different than that obtained with current immunohistochemical methods.

As the goal of this discovery was to find a minimal number of markers with maximal prognostic value, we focused on the small grouping of two markers, GATA3 and NAT1, with highest correlation with ER protein expression (highlighted in Fig. 2A). As shown in the univariate Cox proportional hazards analyses in Table 2, higher levels of ER, GATA3, and NAT1 are all individually related to better patient prognosis.

To more rigorously evaluate these three markers, AQUA analysis for ER, GATA3, and NAT1 was done on an expanded set of breast cancer tumors on tissue microarrays (cohort details in Supplementary Table S1B). In this cohort, pairwise evaluation of these markers by Cox multivariate analysis methods shows that each maintains independent predictive value. The results of unsupervised hierarchical clustering of ER, GATA3, and NAT1 for all of the breast cancer tumors (with data for all three markers) defined two main clusters, cluster 1 with different subsets of coexpression of NAT1, GATA3, and ER, and cluster 2 with predominantly low levels of the three markers (Fig. 2C). These two clusters had significant differences in prognosis as shown in Fig. 2D with 84.6% 5-year survival in cluster 1 (log-rank  $P < 0.0001$ ) versus 65.9% in cluster 2. Although cluster 2 shows remarkably homogeneity in expression of the three markers, cluster 1 illustrates that although having overall higher levels ER, NAT1, and GATA3, there are groupings of marker coexpression that subclassify patients with good outcome (Fig. 2E). Specifically, one cluster (1B) shows a 97% 5-year survival compared with another (1C) with only 80%. Although the unsupervised clustering can define classes that represent dramatic differences in outcome, the method does not lend itself to prospective assignment of class, nor does it take advantage of outcome-based information in class discovery. We sought another method for analysis of this three marker set for potential prospective use.

Genetic algorithm methods have been used extensively in other fields (for review, see ref. 40) to discover a set of mathematical functions that best define properties of a population and more recently have been used on gene expression data (41). They are less popular in array analysis due to the large number of genes and, hence, lower likelihood of convergence. But the limited number of protein biomarkers abstracted from the clustering analysis lends itself well to the genetic algorithm approach. Iterative analysis of algorithms (fitness tests) on a training set results in convergence to optimal predictive values that can then be tested on a validation set. Thus, to optimally select patient subsets with the best prognosis by a multiplex assay, we divided our entire cohort into two randomly



**Fig. 2.** Unsupervised hierarchical clustering of 35 ER and related markers and subset clustering of NAT1, GATA3, and ER. The heat maps shown have the overall similarity ordering of the breast cancer tumors on the vertical axis, and the ordering of antibody immunoreactivities (AQUA scores normalized by Z-score transformation as described in Materials and Methods) on the horizontal axis. Black, indicates protein expression level equal to the mean; green, indicates protein expression level below the mean; red, indicates protein expression level above the mean. The branch lengths and pattern demonstrate the relatedness of the tumors on the vertical axis and the antibody staining on the horizontal axis. *A*, heat map of 161 tumors (with 80% filtering) for 35 markers related to estrogen responsiveness and ER status. Gray, indicates missing values. The yellow-highlighted section, indicates the markers that clustered in a small group with ER (GATA3 and NAT1). *B*, Kaplan-Meier survival curve analysis of clusters *A*, *B*, *C*, and *D* as designated in (*A*). *C*, heat map of 502 tumors with values for NAT1, GATA3, and ER protein expression. *D*, Kaplan-Meier survival curve analysis of the two major cluster groups (cluster 1,  $n = 210$ , and cluster 2,  $n = 292$ ) as shown in (*C*) at 5-year disease-specific survival. *E*, this Kaplan-Meier curve separates out clusters 1A ( $n = 84$ ), 1B ( $n = 34$ ), 1C ( $n = 92$ ), and 2 ( $n = 292$ ), from the heat map in (*C*) including all patients for which we had censor and survival data ( $n = 446$ ).

assigned, training and validation sets with equal baseline survival. We used a genetic algorithm to define a series of risk groups for disease-specific death at 5 years. Each algorithm in this case represents an equation that is a true/false statement, such that the marker value is either greater than or less than or equal to a randomly selected cutpoint. A "solution" thereby consists of three if/then statements that can be calculated as true or false. Each tumor is then assessed one point for each criterion that it meets when true, or no points when false. Tumors can thereby receive a 0, 1, 2, or a 3 cumulative score. These scores are referred to as A (3 points), B (2 points),

C (1 point), and D (zero points). The predictive value of a solution was determined by assessing the risk ratio of the D group (high risk) compared with the A group (low risk). The solution that converged after over 10,000,000 generations (iterations) with the best fit (highest relative risk between good and poor prognosis groups) was selected for validation. The solution was  $ER > 4.6$ ,  $GATA3 > 33.8$ , and  $NAT1 > 18.6$ , where the units are normalized AQUA scores.

The application of this solution to the training set showed that 27 patients were assigned to the good prognosis (low risk), with one event resulting in 96.3% survival at 5 years, and

23 patients assigned to the poor prognosis group (high risk) with eight events (34.8% survival at 5 years). We were also able to assign patients into two intermediate risk groups, with 78.6% and 62.3% 5-year disease-specific survival. The Kaplan-Meier survival curves of the four groups for the training set are shown in Fig. 3A (log-rank  $P < 0.0001$ ). We then applied this algorithm to the internal validation set, which resulted in the Kaplan-Meier survival curves shown in Fig. 3B (log-rank  $P = 0.0073$ ). In the validation set, the 24 patients in the low-risk group had only one event (95.8% 5-year survival), and the 24 patients in the high-risk group had 58.3% 5-year survival, showing again that our model was able to stratify patients into different prognostic groupings on the basis of a prospectively assignable score using only three protein expression measurements; this grouping remains independent in a multivariate Cox model, including age at diagnosis; nuclear grade; tumor size; nodal status; clinical immunohistochemical ER status; and the individual AQUA scores for ER, GATA3, and NAT1 (Table 3).

For external validation of these findings, we repeated this assay on a cohort of patients from a randomized clinical trial of tamoxifen conducted in Sweden from 1986 to 1991 (19). A tissue microarray from these tumors were analyzed for ER, GATA3, and NAT1, normalized for exposure times to the Yale cohort, and then grouped as defined by the genetic algorithm above. The data from 149 patients from the control arm of this study show a 92.7% 5-year survival in group A compared with 58.8% 5-year survival in the high risk (group D). Like the Yale cohort, this group of patients was treated only with surgery. Unlike the Yale cohort, this group was limited to premenopausal patients with low stage (I or II) breast cancer representing a more challenging cohort for classification.

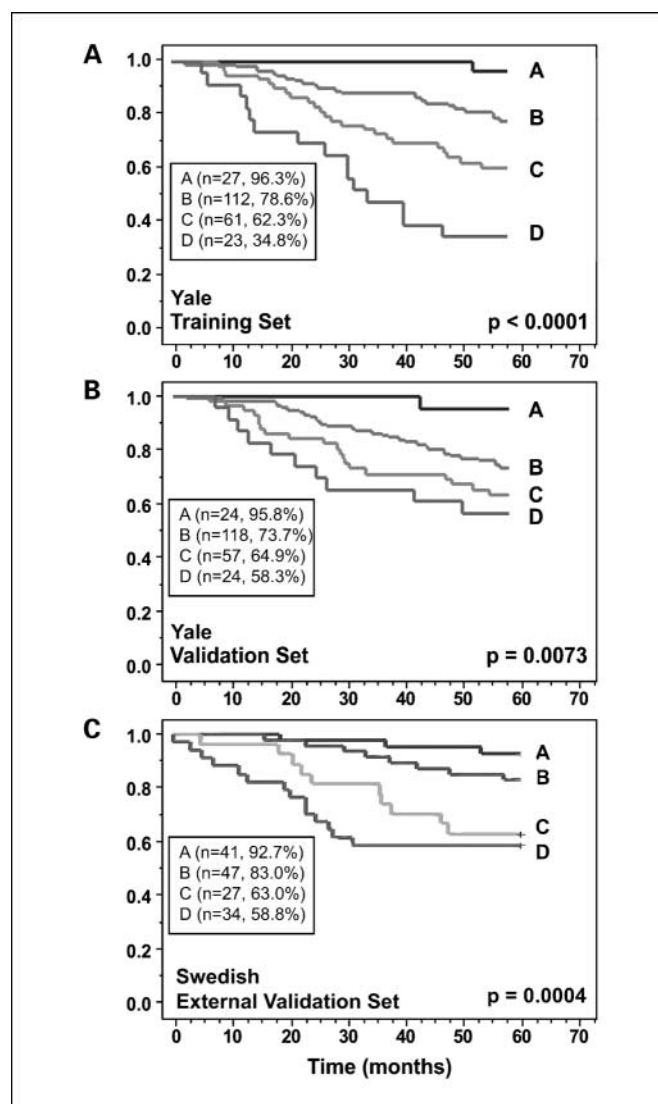
## Discussion

The ability to multiplex markers allows for greater complexity in the assessment of multiple biomarkers that can contribute to predicting patient outcome. Several breast cancer studies have used such a strategy, such as the ratio between *HOXB13* and *IL17BR* gene expression to predict response to adjuvant tamoxifen therapy (3) or the use of the 21 reverse transcription-PCR based test described by Paik et al. (2). There are several commercially available tests based on gene prognosis signatures developed from expression profiling studies (e.g., refs. 2, 43, 44), which could be valuable for subsets of breast cancer patients. This study extends the multiplexing concept to protein expression analysis by quantitative extension of the familiar immunohistochemistry platform.

Several recent breast cancer tissue microarray studies have used hierarchical clustering to immunoprofile breast cancer cohorts with semiquantitative scoring methods to identify prognostic subgroups. These publications include classification by binarized ER, HER2, COX-2, p53, and vascular endothelial growth factor expression (45); analysis of breast cancer expression of neuroendocrine markers (46); definition of the basal subtype (5); confirmation of cDNA expression data (47); and several others focused on prognosis prediction (48–50). The most recent work using this strategy is by Ring et al. (51) and, like this work, it results in patient stratification on the basis of outcome. Some statisticians argue that this type of clustering is invalid because the clustering is done on ordinal,

rather than continuous, data. This study, as well as another recent AQUA-based study (from our laboratory) used hierarchical clustering on continuous data to identify a subset of patients with poor prognosis (52).

Multiplexing, as described in this work, is a combination of true multiplexing and serial section multiplexing. True multiplexing is when multiple markers are measured on the same tissue spot. Although we are multiplexing by measuring keratin, 6-diamidino-2-phenylindole, and our target all on the same tissue spot, we are only quantitatively measuring one marker (the target) per tissue section. We are able to do true multiplexing with multiple targets, but currently only two targets can be assessed per tissue spot. We clustered our data



**Fig. 3.** Kaplan-Meier survival curve analysis of low-, intermediate-, and high-risk patient groups. The scores of patients for the training set – generated model, equating a single point for each criteria met of ER  $< 4.6$ , GATA3  $< 33.8$ , and NAT1  $< 18.6$ , were stratified into four groups of high risk (group D, 0 points), intermediate risk 1 (group C, 1 point), intermediate risk 2 (group B, 2 points), and low risk (group A, 3 points). **A**, Kaplan-Meier survival curves of the training set tumors. **B**, Kaplan-Meier survival curves of the validation set tumors. **C**, Kaplan-Meier survival curves of the external validation set from the Swedish cohort. In each frame, the inset shows each patient grouping by letter (A–D) with the number of patients in the group and the 5-year disease-specific survival.

**Table 3.** Multivariate analysis of multiplex model with clinical factors (5-year disease-specific survival)

Variable	Relative risk (95% confidence interval)	P
Model		<b>0.0084</b>
Group A	1.00	
Group B	5.43 (1.09-26.97)	<b>0.0385</b>
Group C	8.98 (1.64-49.15)	<b>0.0114</b>
Group D	13.99 (2.37-82.51)	<b>0.0036</b>
Age (y)		
Age >50	1.00	
Age ≤50	1.23 (0.79-1.93)	0.3594
Nuclear grade		0.6636
1	1.00	
2	1.09 (0.61-1.97)	0.7642
3	1.28 (0.70-2.34)	0.4308
Tumor size (cm)		<b>&lt;0.0001</b>
≤2	1.00	
>2 < 5	1.94 (1.23-3.08)	<b>0.0047</b>
≥5	2.91 (1.78-4.73)	<b>&lt;0.0001</b>
Nodal status		
Node Negative	1.00	
Node Positive	2.18 (1.41-3.36)	<b>0.0004</b>
Clinical ER immunohistochemical status		
ER positive	1.00	
ER negative	1.20 (0.71-2.03)	0.5062
AQUA ER nuclear	0.99 (0.97-1.01)	0.1766
AQUA GATA3 nuclear	1.01 (0.97-1.04)	0.7340
AQUA NAT1 cytoplasmic	1.00 (0.99-1.01)	0.9623

NOTE: Bold indicates P values <0.05; associated relative risks are also in bold type.

based on a serial section multiplexing. Because the data is quantitative and continuous, we believe it is statistically valid to do this type of clustering analyses.

This analysis was restricted only to genes with available antibodies and could thus be confounded by this selection. However, the ultimate goal of this type of effort is to discover the smallest set of biomarkers with the maximum prognostic value. Toward the same goal, other groups have come up with

similarly small, but unique sets of antibodies that are highly prognostic (51). We anticipate that future studies with other antibodies may suggest the addition of new markers or even the removal of some of the three markers identified in this study. Thus, it is possible, or even likely, that the ultimate most prognostic algorithm will extend or otherwise change the three-marker set defined in this work. Furthermore, without any antibodies available or generated, semiquantitative analysis of *in situ* hybridization results could provide additional information on highly interesting targets as supplement to this type of study.

A limitation of this study is that the predictive value of this test is only valid at 5 years. When this same set of markers was tested for 15 years, the significance is lost. However, this is most likely due to the fact that the genetic algorithm is derived using data for 5-year disease-specific survival. The use of this or any test to exclude patients from chemotherapy, as proposed for the Oncotype Dx test (2), requires assessment of longer-term outcome. Work is under way on this data set to discover a small set of markers that predict low recurrence frequency over a long time period. However, the key value of an assay of this nature is to find the low-risk patients that are highly likely to recur. Table 3 shows that group D patients have a relative risk of nearly 14. Furthermore, the external validation set of young node-negative patients in the external validation cohort shows that group D can define a subset of nearly a quarter of the patient population that has a 40% chance of recurrence. With further validation, this technology has the potential to translate into a useful clinical test. It could be used to select a subset of breast cancer patients with a subclass of tumor that is highly likely to recur and thus patients could be strongly encouraged to undergo adjuvant therapy, even if tumor stage and other clinical variables may suggest otherwise. Further analysis is under way with other markers to increase the classificatory power of this type of assay using the AQUA platform.

### Acknowledgments

We thank Laura Dillon, Thomas D'Aquila, Lori Charette, Matthew Neopolitano, and Jung Kang for their technical assistance in this project.

### References

- van't Veer LJ, Paik S, Hayes DF. Gene expression profiling of breast cancer: a new tumor marker. *J Clin Oncol* 2005;23:1631-5.
- Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;351:2817-26.
- 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;5:607-16.
- Foulkes WD, Brunet JS, Stefansson IM, et al. The prognostic implication of the basal-like (cyclin E high/p27 low/p53+/glomeruloid-microvascular-proliferation+) phenotype of BRCA1-related breast cancer. *Cancer Res* 2004;64:830-5.
- Nielsen TO, Hsu FD, Jensen K, et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res* 2004;10:5367-74.
- Kononen J, Bubendorf L, Kallioniemi A, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998;4:844-7.
- Dolled-Filhart M, Rimm DL. Tissue arrays. 7th ed. In: DeVita VT, Jr., Hellman S, Rosenberg SA, editors. *Cancer: principles and practice of oncology*. Philadelphia: LWW Oncology; 2004. p. 26-34.
- Camp RL, Chung GG, Rimm DL. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nat Med* 2002;8:1323-7.
- McCabe A, Dolled-Filhart M, Camp RL, Rimm DL. Automated quantitative analysis (AQUA) of *in situ* protein expression, antibody concentration, and prognosis. *J Natl Cancer Inst* 2005;97:1808-15.
- Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* 2000;406:747-52.
- Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 2001;98:10869-74.
- Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 2003;100:8418-23.
- Camp RL, Dolled-Filhart M, King BL, Rimm DL. Quantitative analysis of breast cancer tissue microarrays shows that both high and normal levels of HER2 expression are associated with poor outcome. *Cancer Res* 2003;63:1445-8.
- Dolled-Filhart M, Camp RL, Kowalski DP, Smith BL, Rimm DL. Tissue microarray analysis of signal transducers and activators of transcription 3 (Stat3) and phospho-Stat3 (Tyr705) in node-negative breast cancer shows nuclear localization is associated with a better prognosis. *Clin Cancer Res* 2003;9:594-600.
- Kang JY, Dolled-Filhart M, Ocal IT, et al. Tissue microarray analysis of hepatocyte growth factor/Met pathway components reveals a role for Met, matric metalloproteinase, and hepatocyte growth factor activator inhibitor 1 in the progression of node-negative. *Breast Cancer Res* 2003;63:1101-5.

16. Kluger HM, Dolled-Filhart M, Rodov S, et al. Macrophage colony-stimulating factor-1 receptor expression is associated with poor outcome in breast cancer by large cohort tissue microarray analysis. *Clin Cancer Res* 2004;10:173–7.
17. Ocal IT, Dolled-Filhart M, DAquila TG, Camp RL, Rimm DL. Tissue microarray-based studies of patients with lymph node negative breast carcinoma show that met expression is associated with worse outcome but is not correlated with epidermal growth factor family receptors. *Cancer* 2003;97:1841–8.
18. Chung GG, Zerkowski MP, Ocal IT, et al.  $\beta$ -Catenin and p53 analyses of a breast carcinoma tissue microarray. *Cancer* 2004;100:2084–92.
19. Ryden L, Jonsson PE, Chebil G, et al. Two years of adjuvant tamoxifen in premenopausal patients with breast cancer: a randomised, controlled trial with long-term follow-up. *Eur J Cancer* 2005;41:256–64.
20. Camp RL, Rimm EB, Rimm DL. Met expression is associated with poor outcome in patients with axillary lymph node negative breast carcinoma. *Cancer* 1999;86:2259–65.
21. Tolgay Ocal I, Dolled-Filhart M, DAquila TG, Camp RL, Rimm DL. Tissue microarray-based studies of patients with lymph node negative breast carcinoma show that met expression is associated with worse outcome but is not correlated with epidermal growth factor family receptors. *Cancer* 2003;97:1841–8.
22. Buerger H, Otterbach F, Simon R, et al. Comparative genomic hybridization of ductal carcinoma *in situ* of the breast-evidence of multiple genetic pathways. *J Pathol* 1999;187:396–402.
23. Forozan F, Mahlamaki EH, Monni O, et al. Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data. *Cancer Res* 2000;60:4519–25.
24. Ried T, Just KE, Holtgreve-Grez H, et al. Comparative genomic hybridization of formalin-fixed, paraffin-embedded breast tumors reveals different patterns of chromosomal gains and losses in fibroadenomas and diploid and aneuploid carcinomas. *Cancer Res* 1995;55:5415–23.
25. Hermsen MA, Baak JP, Meijer GA, et al. Genetic analysis of 53 lymph node-negative breast carcinomas by CGH and relation to clinical, pathological, morphometric, and DNA cytometric prognostic factors. *J Pathol* 1998;186:356–62.
26. Isola JJ, Kallioniemi OP, Chu LW, et al. Genetic aberrations detected by comparative genomic hybridization predict outcome in node-negative breast cancer. *Am J Pathol* 1995;147:905–11.
27. Kuukasjarvi T, Karhu R, Tanner M, et al. Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. *Cancer Res* 1997;57:1597–604.
28. Kytola S, Rummukainen J, Nordgren A, et al. Chromosomal alterations in 15 breast cancer cell lines by comparative genomic hybridization and spectral karyotyping. *Genes Chromosomes Cancer* 2000;28:308–17.
29. Loveday RL, Greenman J, Simcox DL, et al. Genetic changes in breast cancer detected by comparative genomic hybridisation. *Int J Cancer* 2000;86:494–500.
30. Nishizaki T, Chew K, Chu L, et al. Genetic alterations in lobular breast cancer by comparative genomic hybridization. *Int J Cancer* 1997;74:513–7.
31. Pollack JR, Sorlie T, Perou CM, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* 2002;99:12963–8.
32. Roylance R, Gorman P, Harris W, et al. Comparative genomic hybridization of breast tumors stratified by histological grade reveals new insights into the biological progression of breast cancer. *Cancer Res* 1999;59:1433–6.
33. Schwendel A, Richard F, Langreck H, et al. Chromosome alterations in breast carcinomas: frequent involvement of DNA losses including chromosomes 4q and 21q. *Br J Cancer* 1998;78:806–11.
34. Tirkkonen M, Johannsson O, Agnarsson BA, et al. Distinct somatic genetic changes associated with tumor progression in carriers of BRCA1 and BRCA2 germ-line mutations. *Cancer Res* 1997;57:1222–7.
35. Tirkkonen M, Tanner M, Karhu R, et al. Molecular cytogenetics of primary breast cancer by CGH. *Genes Chromosomes Cancer* 1998;21:177–84.
36. Zudaire I, Otero MD, Caballero C, et al. Genomic imbalances detected by comparative genomic hybridization are prognostic markers in invasive ductal breast carcinomas. *Histopathology* 2002;40:547–55.
37. Cheadle C, Vavtver MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. *J Mol Diagn* 2003;5:73–81.
38. Rubin MA, Zerkowski MP, Camp RL, et al. Quantitative determination of expression of the prostate cancer protein  $\alpha$ -methylacyl-CoA racemase using automated quantitative analysis (AQUA): a novel paradigm for automated and continuous biomarker measurements. *Am J Pathol* 2004;164:831–40.
39. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;95:14863–8.
40. Mitchell M. An introduction to genetic algorithms. Cambridge (MA): MIT Press; 1998.
41. Ooi CH, Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics* 2003;19:37–44.
42. Malzahn K, Mitze M, Thoenes M, Moll R. Biological and prognostic significance of stratified epithelial cytokeratins in infiltrating ductal breast carcinomas. *Virchows Arch* 1998;433:119–29.
43. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–6.
44. 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.
45. Zhang DH, Salto-Tellez M, Chiu LL, Shen L, Koay ES. Tissue microarray study for classification of breast tumors. *Life Sci* 2003;73:3189–99.
46. Makretsov N, Gilks CB, Coldman AJ, Hayes M, Huntsman D. Tissue microarray analysis of neuroendocrine differentiation and its prognostic significance in breast cancer. *Hum Pathol* 2003;34:1001–8.
47. Abd El-Rehim DM, Pinder SE, Paish CE, et al. Expression of luminal and basal cytokeratins in human breast carcinoma. *J Pathol* 2004;203:661–71.
48. 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;10:6143–51.
49. Jacquemier J, Ginestier C, Rougemont J, et al. Protein expression profiling identifies subclasses of breast cancer and predicts prognosis. *Cancer Res* 2005;65:767–79.
50. Callagy G, Cattaneo E, Daigo Y, et al. Molecular classification of breast carcinomas using tissue microarrays. *Diagn Mol Pathol* 2003;12:27–34.
51. Ring BZ, Seitz RS, Beck R, et al. Novel prognostic immunohistochemical biomarker panel for estrogen receptor-positive breast cancer. *J Clin Oncol* 2006;24:3039–47.
52. Siddiqui SF, Pawelek J, Handerson T, et al. Coexpression of  $\beta$ 1,6-N-acetylglucosaminyltransferase V glycoprotein substrates defines aggressive breast cancers with poor outcome. *Cancer Epidemiol Biomarkers Prev* 2005;14:2517–23.
53. He M, Burghardt TP, Vockley J. A novel approach to the characterization of substrate specificity in short/branched chain Acyl-CoA dehydrogenase. *J Biol Chem* 2003;278:37974–86.
54. Sotiriou C, Neo SY, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* 2003;100:10393–8.
55. Thompson DA, Weigel RJ, haG-2, the human homologue of the *Xenopus laevis* cement gland gene XAG-2, is coexpressed with estrogen receptor in breast cancer cell lines. *Biochem Biophys Res Commun* 1998;251:111–6.
56. Cicatiello L, Scafoglio C, Altucci L, et al. A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive transcriptome. *J Mol Endocrinol* 2004;32:719–75.
57. Jiang Y, Harlocker SL, Molesh DA, et al. Discovery of differentially expressed genes in human breast cancer using subtracted cDNA libraries and cDNA microarrays. *Oncogene* 2002;21:2270–82.
58. Mackay A, Jones C, Dexter T, et al. cDNA microarray analysis of genes associated with ERBB2 (HER2/neu) overexpression in human mammary luminal epithelial cells. *Oncogene* 2003;22:2680–8.
59. Beardsley DI, Kowbel D, Lataxes TA, et al. Characterization of the novel amplified in breast cancer-1 (NABC1) gene product. *Exp Cell Res* 2003;290:402–13.
60. Clark J, Edwards S, John M, et al. Identification of amplified and expressed genes in breast cancer by comparative hybridization onto microarrays of randomly selected cDNA clones. *Genes Chromosomes Cancer* 2002;34:104–14.
61. Bertucci F, Nasser V, Granjeaud S, et al. Gene expression profiles of poor-prognosis primary breast cancer correlate with survival. *Hum Mol Genet* 2002;11:863–72.
62. West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* 2001;98:11462–7.
63. Watson PH, Chia SK, Wykoff CC, et al. Carbonic anhydrase XII is a marker of good prognosis in invasive breast carcinoma. *Br J Cancer* 2003;88:1065–70.
64. Bouras T, Southey MC, Chang AC, et al. Stanniocalcin 2 is an estrogen-responsive gene coexpressed with the estrogen receptor in human breast cancer. *Cancer Res* 2002;62:1289–95.
65. Iwao K, Matoba R, Ueno N, et al. Molecular classification of primary breast tumors possessing distinct prognostic properties. *Hum Mol Genet* 2002;11:199–206.
66. Pusztai L, Ayers M, Stec J, et al. Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors. *Clin Cancer Res* 2003;9:2406–15.
67. Charpentier AH, Bednarek AK, Daniel RL, et al. Effects of estrogen on global gene expression: identification of novel targets of estrogen action. *Cancer Res* 2000;60:5977–83.
68. Zajchowski DA, Bartholdi MF, Gong Y, et al. Identification of gene expression profiles that predict the aggressive behavior of breast cancer cells. *Cancer Res* 2001;61:5168–78.
69. Gruberger S, Ringner M, Chen Y, et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001;61:5979–84.
70. Oh JJ, Grosshans DR, Wong SG, Slamon DJ. Identification of differentially expressed genes associated with HER-2/neu overexpression in human breast cancer cells. *Nucleic Acids Res* 1999;27:4008–17.
71. Lin CY, Strom A, Vega VB, et al. Discovery of estrogen receptor  $\alpha$  target genes and response elements in breast tumor cells. *Genome Biol* 2004;5:R66.
72. Ahr A, Holtrich U, Solbach C, et al. Molecular classification of breast cancer patients by gene expression profiling. *J Pathol* 2001;195:312–20.
73. Hoch RV, Thompson DA, Baker RJ, Weigel RJ. GATA-3 is expressed in association with estrogen receptor in breast cancer. *Int J Cancer* 1999;84:122–8.
74. Rhee MS, Lindau-Shepard B, Chave KJ, Galivan J, Ryan TJ. Characterization of human cellular  $\gamma$ -glutamyl hydrolase. *Mol Pharmacol* 1998;53:1040–6.
75. Nacht M, Ferguson AT, Zhang W, et al. Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. *Cancer Res* 1999;59:5464–70.

76. Martin KJ, Kritzman BM, Price LM, et al. Linking gene expression patterns to therapeutic groups in breast cancer. *Cancer Res* 2000;60:2232–8.
77. Hayashi S. Prediction of hormone sensitivity by DNA microarray. *Biomed Pharmacother* 2004;58:1–9.
78. Inoue A, Yoshida N, Omoto Y, et al. Development of cDNA microarray for expression profiling of estrogen-responsive genes. *J Mol Endocrinol* 2002;29:175–92.
79. Wilson KS, Roberts H, Leek R, Harris AL, Geradts J. Differential gene expression patterns in HER2/neu-positive and -negative breast cancer cell lines and tissues. *Am J Pathol* 2002;161:1171–85.
80. Bosma AJ, Weigelt B, Lambrechts AC, et al. Detection of circulating breast tumor cells by differential expression of marker genes. *Clin Cancer Res* 2002;8:1871–7.
81. Stanley LA, Coroneos E, Cuff R, et al. Immunohistochemical detection of arylamine *N*-acetyltransferase in normal and neoplastic bladder. *J Histochem Cytochem* 1996;44:1059–67.
82. Itoh T, Karlsberg K, Kijima I, et al. Letrozole-, anastrozole-, and tamoxifen-responsive genes in MCF-7aro cells: a microarray approach. *Mol Cancer Res* 2005;3:203–18.
83. Dunaway GA, Kasten TP, Sebo T, Trapp R. Analysis of the phosphofructokinase subunits and isoenzymes in human tissues. *Biochem J* 1988;251:677–83.
84. Stemmer-Rachamimov AO, Wiederhold T, Nielsen GP, et al. NHE-RF, a merlin-interacting protein, is primarily expressed in luminal epithelia, proliferative endometrium, and estrogen receptor-positive breast carcinomas. *Am J Pathol* 2001;158:57–62.
85. Suemori S, Lynch-Devaney K, Podolsky DK. Identification and characterization of rat intestinal trefoil factor: tissue- and cell-specific member of the trefoil protein family. *Proc Natl Acad Sci U S A* 1991;88:11017–21.
86. Porter D, Lahti-Domenici J, Keshaviah A, et al. Molecular markers in ductal carcinoma *in situ* of the breast. *Mol Cancer Res* 2003;1:362–75.