

## An Unsupervised Approach to Identify Molecular Phenotypic Components Influencing Breast Cancer Features

Florin M. Selaru,<sup>1</sup> Jing Yin,<sup>1</sup> Andreea Olaru,<sup>1</sup> Yuriko Mori,<sup>1</sup> Yan Xu,<sup>1</sup> Steven H. Epstein,<sup>1</sup> Fumiaki Sato,<sup>2</sup> Elena Deacu,<sup>1</sup> Suna Wang,<sup>1</sup> Anca Sterian,<sup>1</sup> Amy Fulton,<sup>2</sup> John M. Abraham,<sup>1</sup> David Shibata,<sup>3</sup> Claudia Baquet,<sup>4</sup> Sanford A. Stass,<sup>2</sup> and Stephen J. Meltzer<sup>1,2</sup>

<sup>1</sup>Department of Medicine, Division of Gastroenterology, and <sup>2</sup>Departments of Pathology, <sup>3</sup>Surgery, <sup>4</sup>Epidemiology and Preventive Medicine, and University of Maryland Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, Maryland

### Abstract

To discover a biological basis for clinical subgroupings within breast cancers, we applied principal components (PCs) analysis to cDNA microarray data from 36 breast cancers. We correlated the resulting PCs with clinical features. The 35 PCs discovered were ranked in order of their impact on gene expression patterns. Interestingly, PC 7 identified a unique subgroup consisting of estrogen receptor (ER); (+) African-American patients. This group exhibited global molecular phenotypes significantly different from both ER (–) African-American women and ER (+) or ER (–) Caucasian women ( $P < 0.001$ ). Additional significant PCs included PC 4, correlating with lymph node metastasis ( $P = 0.04$ ), and PC 10, with tumor stage (stage 2 versus stage 3;  $P = 0.007$ ). These results provide a molecular phenotypic basis for the existence of a biologically unique subgroup comprising ER (+) breast cancers from African-American patients. Moreover, these findings illustrate the potential of PCs analysis to detect molecular phenotypic bases for relevant clinical or biological features of human tumors in general.

### Introduction

Among women living in the United States, the three most commonly diagnosed cancers in 2003 will be cancers of the breast, lung, and colon (1). Breast cancer alone is expected to account for 32% (211,300) of all new cancer cases and 39,800 deaths in this country (1). Numerous clinical factors influence the prognosis of this disease, including tumor stage at time of diagnosis, age (2), histological grade, hormone receptor status (3), tumor size, and lymph node (LN) status (4). High estrogen receptor (ER) mRNA levels have been associated with absent or minimal necrosis, as well as vascular invasion (3). It has also been suggested that there is differential expression of ER isoforms between African-American (AA) and Caucasian (C) patients (5). In tumors of AA women, the protective ER $\beta$  isoform was decreased significantly relative to matched normal tissue (5). It is believed that these and other biological differences may contribute to the higher mortality and lower survival rates observed in AA breast cancer patients (6–8). However, comprehensive molecular approaches to understanding this specific problem (for example, applying gene microarray data) have not yet been reported. Therefore, in the current study, we analyzed global gene expression data from breast cancers using an unsupervised bioinformatics approach, principal components analysis (PCA).

We used this approach because whereas supervised analyses of gene expression data assign cases to predefined clinical groups, un-

supervised strategies reveal natural groupings of patients. For example, one of the first and most widely used unsupervised techniques has been hierarchical clustering (9, 10). Clustering reveals groupings in data resulting from the superimposition of numerous biological characteristics or dimensions. Thus, cluster analysis may fail to distinguish among subtle categories of disease or between relevant and irrelevant genetic data (11). PCA delineates key dimensions, or components, within a multidimensional gene dataset to explain clinical differences, such as tumor aggressiveness (12). In contrast to hierarchical clustering, PCA reveals multiple layers of meaning designated components within the complex genetic dataset. These components are independent, allowing PCA to mine the data in a layer-oriented fashion, isolating each layer in turn from the next (13). Moreover, PCA measures the fraction of variance contributed by each component to variance within the entire dataset (13). Finally, PCA suggests explanations for a given component by ranking all of the genes in order of relative influence on that component. The current findings provide global molecular phenotypic evidence for the existence of an ER-positive AA breast tumor biological subgroup. This study also illustrates the potential of PCA to detect molecular phenotypic profiles of other clinical features, including age, ethnic origin, tumor size, progesterone receptor status, and LN metastasis, as well as to identify other previously uncharacterized but biologically important tumor parameters. The application of PCA to cDNA microarray data offers an analytic means to identify and explain important biological aspects of malignancy.

### Materials and Methods

#### Surgical Specimens and cDNA Microarray Preparation

Thirty-six breast cancers, surgically resected between 1994 and 1999, were obtained from the University of Maryland Greenebaum Cancer Center. The clinical and molecular features of these tumors are depicted in Table 1. In dichotomizing patient age, a cutoff value of 50 was determined *a priori* based on the average age of onset of menopause in our patient population. Genomic DNA and total RNA were extracted from fresh-frozen specimens. Amplified RNA (aRNA) was amplified from 20–50  $\mu\text{g}$  of total RNA using a T7-based protocol (14). Labeling was performed on 3–6  $\mu\text{g}$  of aRNA by incorporating Cy3- or Cy5-labeled dCTP using random primers and Superscript reverse transcriptase (15). A universal reference probe was prepared from an equimolar mixture containing aRNAs from eight human cancer cell lines (10, 11, 15, 16). cDNA microarray slides containing 8064 human cDNA clones were prepared according to a protocol described previously (15). The Lawrence Livermore Laboratory cDNA library was used as a clone source (Invitrogen, Carlsbad, CA). All 8064 of the clones were independently sequence-verified and checked for correct annotation in our laboratory (10, 11, 15, 16). Microarrays were cohybridized to Cy5-labeled specimen aRNA and Cy3-labeled universal reference probe aRNA at 65°C overnight. After hybridization, each slide was scanned using a GenePix 4000A dual-laser slide scanning system (Axon Instruments, Union City, CA).

Received 10/12/03; revised 12/12/03; accepted 12/23/03.

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.

**Note:** F. M. Selaru, J. Yin, and A. Olaru contributed equally to this work.

**Requests for reprints:** Stephen J. Meltzer, 655 West Baltimore Street, BRB 8-009, Baltimore, MD 21201. Phone: (410) 706-3375; Fax: (410) 706-1099; E-mail: smeltzer@medicine.umaryland.edu.

Table 1 *Clinical and molecular data on breast cancer patients*

ID	Date of procedure	Age	Size	LN <sup>a</sup>	ER	PR	Race	Stage
BC 1	1995	59	NK	Yes	NK	NK	AA	3
BC 2	1996	36	2.1	No	(-)	(-)	AA	3
BC 3	1996	32	3.6	0/14	(-)	(-)	C	2
BC 4	1996	59	2.5	0/31	(+)	(-)	C	2
BC 5	1996	55	6.5	0/16	NK	NK	C	3
BC 6	1996	48	2.2	0/5	(-)	(-)	AA	2
BC 7	1996	45	10	Yes	(-)	(-)	C	2
BC 8	1998	45	3	1/9	(-)	(-)	AA	2
BC 9	1998	37	2.6	1/8	(-)	(-)	AA	2
BC 10	1998	65	2.5	No	(+)	(+)	C	3
BC 11	1998	54	3	4/6	(+)	(+)	C	3
BC 12	1998	73	7.2	NK	(-)	(-)	C	3
BC 13	1998	40	2.4	4/26	(-)	(+)	C	3
BC 14	1998	56	3.7	11/12	(+)	(+)	AA	3
BC 15	1999	46	2.4	3/20	(-)	(-)	AA	2
BC 16	1999	35	2.1	3/6	(+)	(+)	C	2
BC 17	1999	47	3.8	No	(+)	(+)	AA	2
BC 18	1999	45	2	0/4	(-)	(-)	C	2
BC 19	1996	58	7	3/12	(+)	(+)	C	3
BC 20	1996	53	7	29/29	(+)	(+)	C	3
BC 21	1994	44	NK	1/13	(-)	(+)	AA	2
BC 22	1994	39	10	Yes	(-)	(-)	AA	3
BC 23	1995	47	2.5	0/12	(+)	(-)	C	2
BC 24	1995	38	6	NK	NK	NK	C	3
BC 25	1994	78	3.6	0/14	(+)	(-)	C	2
BC 26	1995	67	3	1/12	(-)	(-)	C	2
BC 27	1995	41	6	0/7	(+)	(-)	C	3
BC 28	1994	81	6.8	No	(+)	(-)	AA	3
BC 29	1994	91	5.8	NK	(+)	(+)	AA	3
BC 30	1995	92	5	3/3	(+)	(-)	AA	2
BC 31	1995	30	9	0/19	(-)	(-)	AA	3
BC 32	1995	47	4.5	1/26	(-)	(-)	AA	2
BC 33	1995	44	3.5	2/25	(+)	(-)	AA	2
BC 34	1994	48	3	0/13	(-)	(-)	AA	2
BC 35	1994	64	10	Yes	NK	NK	AA	3
BC 36	1995	60	2.5	11/15	(+)	(-)	C	2

<sup>a</sup> LN, lymph node; ER, estrogen receptor; PR, progesterone receptor; C, Caucasian; AA, African-American; NK, not known.

## Data Preprocessing

We included in this analysis only clones yielding expression information in at least 97% of the tumors (*i.e.*, clones lacking information for, at most, one tumor). This minimal-information threshold was surpassed by 7513 of 8064 printed clones. Data points representing gene expression ratios were log-transformed. We then normalized data to exclude intensity-dependent bias. In this fashion, local distortions in signal and background intensity within different regions of a slide were overcome. We based this procedure on the assumption that Cy5:Cy3 ratios should not depend on spot intensity. This type of data distortion was removed by a robust scatter-plot smoothing method (17). Using SigmaPlot version 5 (SPSS, San Rafael, CA), we calculated the LOWESS fitting curve using a fitting parameter of 40%. Data for each slide was normalized so that the mean in-slide expression value was 0 and the SD was 1.

## Data Analysis

**Clustering and Derivation of Principal Components (PCs).** Data imported from GenePix were manipulated and clustered using average linkage clustering with centered correlation (9). The second step in the analysis involved PCA. All of the PCA calculations were performed in MatLab (MathWorks, Inc., Natick, MA). The data, filtered as described above, were input into MatLab and normalized so that for each specimen, mean gene expression equaled 0 and SD equaled 1. Because the independent dimensions in PCA typically equal the number of specimens minus 1, and because there were 36 breast cancers in this study, 35 independent components were derived. The relative contribution of each component to the total data variance was calculated, and components were ranked in decreasing order of their relative contribution to this variance.

**Associations between PCs and Clinicomolecular Data.** Beginning with the first-ranked component, attempts were made to correlate each component with known clinical data. Various statistical techniques were considered for this task, among them multivariate fitting models. For its simplicity and the

ready availability of validation techniques, one-way ANOVA was used. A literature search was conducted and subgroups of breast cancers reported previously as having distinct molecular, biological, or clinical features were used in our correlative analyses. Thus, a *P* was calculated for the one-way ANOVA test of the association between each component and the following tumor characteristics: year of surgery, age, ethnic origin, LN status, ER status, progesterone receptor status, location (left *versus* right breast), stage, size, and histological type (intraductal *versus* lobular). One-way ANOVA calculations were performed in Statistica (StatSoft, Tulsa, OK).

**Validation of Associations by Permutation Testing.** It is important to note that PCs are not observed data, but are rather summary statistics, which capture the main variance in data. Therefore, it is predictable that low *P*s will be obtained when performing one-way ANOVA analyses of associations with these components. To ensure that these *P*s are significant, they need to be adjusted. One method to adjust *P*s in this fashion is based on permutation testing. We performed permutation-based confirmatory analyses for each observed association between a PC and a clinical feature. First, the *P* of the one-way ANOVA test for association between the PC and the clinical feature was calculated (see above); next, identification tags for the patients were randomly shuffled, and the *P*s of the one-way ANOVA test for the association between each component and the clinical feature were recalculated. This calculation resulted in 35 different *P*s, 1 for the association between each component and the given clinical or molecular feature. Among these 35 *P*s, the highest *P* was recorded. Next, step 2 was repeated 999 times, yielding a total of 999 "highest" *P*s. Finally, the *P* calculated at step 1 was ranked among the 999 values obtained by random permutations. The rank thus obtained constituted the "corrected" *P*, representing the probability of having obtained the original *P* by chance (18).

**Gene Loading Values in PCs.** The loading value was the number assigned by PCA to represent the influence, within a particular component, of a given gene relative to other genes. Thus, the greater the relative impact of a given gene on a particular component, the more extreme its positive or negative loading value in that component. When the loading value for a given gene was close to 0 for a given component, the gene exerted a minimal influence on that component.

## Results

**Cluster Analysis.** As a first unsupervised bioinformatics strategy, we used hierarchical agglomerative clustering. Clustering failed to recognize any clinical tumor characteristics, being influenced mainly by the age of the surgical specimen (data not shown). This result suggested that additional unsupervised approaches would be necessary if meaningful groups within the microarray data were to be discovered.

**PCA.** PCA extracted 35 PCs. To discover the biological significance of these PCs, they were correlated with clinical or biological features using one-way ANOVA (13). These associations revealed a significant impact of several features on global molecular phenotype (Table 2A). The following PCs were found to be associated with only one clinical characteristic: PC 1, 4, 5, 6, 9, 10, 12, 23, and 25. PC 7 was associated with LN metastatic status ( $P = 0.03$ , one-way ANOVA), ER status ( $P = 0.009$ , one-way ANOVA), and with a subgroup consisting of ER (+) AA patients ( $P < 0.0001$ , one-way ANOVA). The association between PC 7 and the ER (+) AA patients was significantly stronger than were the other two associations. Thus, the fundamental factor impacting PC 7 was the biological difference between AA patients with ER (+) tumors and all of the remaining breast cancers, rather than ER status or LN metastasis alone.

PC 7 identified a unique subgroup of ER (+) AA patients. Initially, we observed an association between PC 7 and ER status *per se* ( $P 0.009$ , one-way ANOVA). Upon additional testing, we noticed that PC 7 had an even stronger association with a subgroup of ER (+) AA patients. PC7 grouped all of the ER (+) AA breast cancers into one category and all of the remaining tumors, regardless of ethnic origin or ER status, into a second category ( $P < 0.0001$ , one-way ANOVA; Fig. 1). The statistical

Table 2 Components and genes correlating with clinical features of breast cancer patients  
Table 2A Correlations between PCs<sup>a</sup> and clinical and molecular features

	Date	LN	Race	Histologic type (ductal vs. lobular)	ER (+) AA vs. the others	Age (≤/ >50)	Stage (2 or 3)	Size (≤5/ >5 cm)	PR status
PC 1	2 × 10 <sup>-7</sup>								
PC 4		0.048							
PC 5			0.003						
PC 6				0.017					
PC 7					0.0002				
PC 9						0.003			
PC 10							0.007		
PC 23								0.01	
PC 25									0.036

Table 2B Genes with highest loading values in PC 7

Loading	Accession no.	Gene name
-0.0613	NM_006103	WAP four-disulfide core domain 2 (WFDC2), mRNA
-0.0579	NM_004374	Cytochrome c oxidase subunit VIc (COX6C), nuclear gene encoding mitochondrial protein, mRNA
-0.053	NM_006088	Tubulin, β, 2 (TUBB2), mRNA
-0.0527	NM_001871	Carboxypeptidase B1 (tissue) (CPB1), mRNA
-0.0518	XM_048179	Interferon-induced protein with tetratricopeptide repeats 1 (IFIT1), mRNA
-0.0511	NM_003467	Chemokine (C-X-C motif), receptor 4 (fusin) (CXCR4), mRNA
-0.05	NM_003191	Threonyl-tRNA synthetase (TARS), mRNA
-0.0498	XM_027680	Interferon-stimulated protein, 15 kDa (ISG15), mRNA
-0.0486	AF071020	Clone 19 MHC class I antigen (HLA-G) mRNA, partial cds
-0.0476	AF054838	Tetraspan TM4SF (TSPAN-1) mRNA, complete cds
-0.0473	XM_006121	Cathepsin D (lysosomal aspartyl protease) (CTSD), mRNA
-0.0453	AA504348	Topoisomerase (DNA) II α (170kD)
-0.045	AA504348	Topoisomerase (DNA) II α (170kD)
-0.0448	A1460128	<i>N</i> -acetyltransferase 2 (arylamine <i>N</i> -acetyltransferase)
-0.0447	S58545	Acrosomal serine protease inhibitor mRNA, complete cds
-0.0437	D87995	PACE4A-II, complete cds
-0.0435	AA598508	ae35a02.s1 Gessler Wilms tumor cDNA clone IMAGE:897770 3' similar to gb:M68867 RETINOIC ACID-BINDING PROTEIN II, CELLULAR (); mRNA sequence
-0.0418	AA700604	Sorbitol dehydrogenase
-0.0408	AF029082	14-3-3 δ protein mRNA, complete cds
0.0496	AF151898	CGI-140 protein mRNA, complete cds
0.0525	X03212	mRNA fragment for mesothelial type II keratin K7
0.0536	XM_032969	Trefoil factor 3 (intestinal) (TFF3), mRNA
0.054	T61948	FBJ murine osteosarcoma viral oncogene homolog B
0.058	XM_012318	Early growth response 1 (EGR1), mRNA
0.0606	NM_001885	Crystallin α B (CRYAB), mRNA
0.0608	NM_003064	Secretory leukocyte protease inhibitor (antileukoproteinase) (SLPI), mRNA
0.0621	H73727	Ribosomal protein S14
0.0679	NM_003064	Secretory leukocyte protease inhibitor (antileukoproteinase) (SLPI), mRNA
0.0682	NM_002343	Lactotransferrin (LTF), mRNA

<sup>a</sup> PC, principal component; LN, lymph node; ER, estrogen receptor; AA, African-American; PR, progesterone receptor.

significance of this relationship was maintained after permutation testing ( $P = 0.006$ ; see "Materials and Methods," above).

PC 4 correlated with LN metastatic status ( $P = 0.04$ ), PC 5 with ethnic origin (*C versus AA*;  $P = 0.002$ ), PC 6 with histological subtype (strictly intraductal *versus* lobular/mixed;  $P = 0.01$ ), PC 9 with age (*< versus > 50 years old*;  $P = 0.003$ ), PC 10 with tumor stage (stage 2 *versus* stage 3;  $P = 0.007$ ), PC 23 with tumor size (*< versus > 5 cm*;  $P = 0.009$ ), and PC 25 with progesterone receptor status ( $P = 0.03$ ). We were not able to find an association between PC 1 and any of the clinical, biological, or molecular features that were tested. It was not due to a batch effect. This finding may have been due to a fundamental biological subgrouping in breast cancers, previously unidentified, with a very strong impact on global molecular phenotypes.

**Relative Contributions of Individual Genes to Each Component.** The individual contributions of genes to particular components were assessed by examining gene loading values. Loading values represent association coefficients between genes and components (13). They range from  $-1$  to  $+1$ , with these extremes representing perfect negative or positive associations, respectively. Genes with high loading values in a given PC are associated with high weights in the equation determining total PC output value (13). In the current study, genes were ordered according to their loading values, representing their degree of influence on each component. Because component 7 showed the most significant  $P$  after permutation testing,

attention was focused on this component. Genes with extreme positive or negative loading values in this component are displayed in Table 2B (PC 7). ER (+) AA specimens received negative values on PC 7, whereas the remaining specimens received positive values on PC 7. Thus, genes with negative loading values in PC 7 correlate with ER (+) AA specimens, whereas genes with positive loading values on PC 7 correlate with the remaining specimens. Genes with negative loading values are important in defining the biology of ER (+) AA cancers; genes with positive loading values are important in defining the biology of non-ER (+) AA cancers. The more extreme the loading value (positive or negative), the more important the gene in defining the biology of these cancer subgroups. Thus, genes with the most extreme loading values (positive or negative) are logical targets for future in-depth studies. In table 2B, proteases included NM\_006103 (carboxypeptidase B1), 17472831 (cathepsin D), and NM\_003064 (secretory leukocyte protease). Interestingly, NM\_003064, which was printed in two separate locations on the microarray, appears twice on this list of genes with highest loading values, supporting the validity of our microarray data as well as the PCA method we used. Also of interest, a number of the most highly ranked genes have putative relationships to tumorigenicity, including *topoisomerase II α*, *IFN-stimulated protein*, *FBJ murine osteosarcoma viral oncogene homologue*, *early growth response 1*. Wilms tumor homologue, trefoil factor 3, cathepsin D, and 14-3-3-σ.

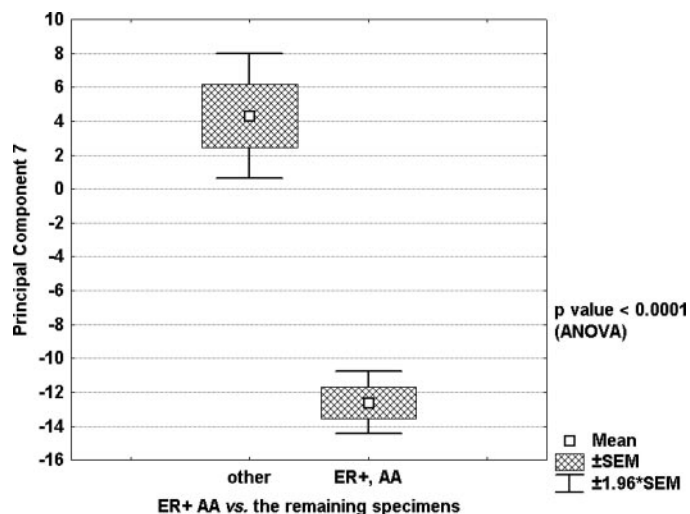


Fig. 1. Graphic display of association of PC 7 with race and estrogen receptor (ER) status. X axis displays two categories: the first category includes ER (+) African-Americans (AA), whereas the second category includes the rest of the specimens, i.e., ER (-) AA, ER (-) C and ER (+) C patients. Y axis shows the output value for each of these four groups in PC 7;  $P$ s were obtained by one-way ANOVA; bars,  $\pm$ SE. The ER (+) AA tumors (bottom rightmost group) tended to have much lower output values in PC 7. This result demonstrates a clear separation between the 6 ER (+) AA tumors and the remaining cancers. Within the other category, there are 9 ER (-) AA patients, 11 ER (+) patients, and 6 ER (-) C patients. The  $\square$  represent the mean, the  $\boxplus$ ,  $\pm$ SE, and bars  $\pm 1.96 * SE$ .

## Discussion

In the current study, we applied an unsupervised bioinformatics approach to identify naturally occurring groups within a breast cancer population. This approach, PCA, reduces complex data to a smaller number of components (12, 19). It is ideally suited to cDNA microarray analysis, which generates thousands of data points from a single experiment. In our study, PC 4 correlated significantly with LN status, component 6 with histological grade, component 9 with age, component 10 with stage, and component 23 with size. The association between PC 1 and year of surgery is not of clinical significance *per se*. This association suggests that these microarray data may have been influenced to a significant degree by factors such as surgical technique, tissue collection, and preservation.

Perhaps the most interesting finding in the current study was the significant separation observed between ER (+) AA tumors and the remaining breast cancers (PC 7). This finding supports an impact of ER status on gene expression profiles. Although it was known that ER status influences gene expression (20) and that breast cancers from AA patients exhibit unique biological features (7, 8, 21, 22), the current data now suggest that ER status has a different impact on AA breast cancer patients than on Caucasian breast cancer patients. The current data identify this ER (+) AA subgroup based on an unsupervised analysis of global gene expression data and reveal genes associated with these spontaneously revealed categories.

In addition to identifying natural subgroupings among breast cancer patients, PCA identified genes influencing these groupings. Genes displaying high loading values suggested possible pathways or molecular bases underlying each PC (13). Although other methods, such as significance analysis of microarrays (18), may identify differentially expressed genes among predefined clinical subgroups, we have shown that PCA identifies genes relevant to natural subgroups of breast cancer patients. However, these genes still require individual verification, either by significance analysis of microarrays or using quantitative reverse transcriptase-PCR.

Many of the genes identified by PCA already had been linked to other human cancers or to breast cancers in general and ER status in

particular. Among the genes relevant to PC 7 [the ER (+) AA component] was *Homo sapiens cathepsin D*, which is overexpressed in aggressive human breast cancers (25) and induced by estrogens in hormone-responsive breast cancer cells (26). High cathepsin D concentrations in primary breast cancers correlate with an increased risk of metastasis and are particularly useful in orienting LN-negative tumors to adjuvant therapy (25). *Topoisomerase 2  $\alpha$* , which was also related to PC 7, is associated with mammalian cell proliferation (27), and its overexpression is linked to cellular dedifferentiation and a biologically aggressive breast cancer phenotype (27). Another gene linked to PC 7, *14-3-3- $\sigma$* , interacts with cyclin-dependent kinases and controls the rate of entry of cells into mitosis (28). The protein product of this gene has been implicated in the neoplastic transformation of breast epithelial cells by virtue of its role as a tumor suppressor; as such, it may constitute a robust biomarker with clinical utility (28). Additional studies have implicated *14-3-3- $\sigma$*  as a target of methylation in breast cancers (29). It has been suggested that hypermethylation and loss of expression of *14-3-3- $\sigma$*  occurs at an early stage in the progression to invasive breast cancer (29).

The relationship between breast cancer behavior and clinical or molecular factors has been explored in numerous previous studies (21, 30–35). There are studies reporting that AA women tend to develop highly aggressive breast cancers (5, 36), with a higher mortality rate relative to their Caucasian counterparts (7, 37, 38). The basis of this disparity has not been found. Several investigators have linked this difference to coexisting variables, including socioeconomic status and limited access to health care (39), or to stage of disease at diagnosis (40). However, biological bases for this disparity have also been suggested. For example, *cyclin D* overexpression is more prevalent in non-C breast cancers (7), and variations in estrogen-mediated signaling due to differences in ER isoforms may account for differences in breast tumor behavior (5). In fact, ER status *per se* has been correlated with ethnic origin (22, 41). The current study shows that based on global molecular phenotyping, ER (+) AA breast cancers constitute a distinct biological subgroup ( $P < 0.0001$ , one-way ANOVA). This finding, which suggests that ER status has disparate effects on AA and C patients, was obtained by an unsupervised bioinformatics approach (PCA) and additionally validated by permutation analysis ( $P = 0.006$ , one-way ANOVA). The unsupervised fashion by which this finding was derived supports the conclusion that this is a natural, biologically meaningful group of patients. The other associations found between PCs and clinical features were less strong than the distinction between ER (+) AA and the remaining breast cancers (i.e., at permutation analyses the other  $P$ s were not statistically significant). Nonetheless, these associations should be additionally investigated, as they may indicate a natural biologic grouping among breast cancer patients.

It is a commonly held hope that molecular biology, by offering a more objective and scientifically based view of tumors and other human diseases, will eventually supplement our current clinicopathologic classification schemes. cDNA microarrays hold immense promise in this regard, considering the large statistical power of the data contained within them. The current study suggests that PCA can reveal multiple levels of meaning within microarray data by viewing these data from multiple viewpoints. By accomplishing this task, PCA increases the likelihood that cDNA microarrays will become part of our molecular taxonomic armamentarium. Moreover, the current study suggests that PCA can simplify complex microarray data, both by identifying and providing insight into biological and clinical categories. In some cases, PCs themselves may have more biological or clinical significance than our current clinical categories do. For example, a given component may identify cancers with a poor outcome or precancerous lesions with a higher risk of neoplastic progression. Ultimately, large-scale prospective and other clinical correlative stud-

ies are needed that apply PCA to predict the clinical behavior and natural history of human malignant disease.

## References

- Jemal, A., Murray, T., Samuels, A., Ghafoor, A., Ward, E., and Thun, M. J. Cancer statistics, 2003. *CA Cancer J. Clin.*, 53: 5–26, 2003.
- Joslyn, S. A., and West, M. M. Racial differences in breast carcinoma survival. *Cancer (Phila.)*, 88: 114–123, 2000.
- Nagai, M. A., Marques, L. A., Yamamoto, L., Fujiyama, C. T., and Brentani, M. M. Estrogen and progesterone receptor mRNA levels in primary breast cancer: association with patient survival and other clinical and tumor features. *Int. J. Cancer*, 59: 351–356, 1994.
- Silva, O. E., and Zurrada, S. *Breast Cancer: A Practical Guide*. Amsterdam: Elsevier Science 71, 2000.
- Poola, I., Clarke, R., DeWitty, R., and Leffall, L. D. Functionally active estrogen receptor isoform profiles in the breast tumors of African American women are different from the profiles in breast tumors of Caucasian women. *Cancer (Phila.)*, 94: 615–623, 2002.
- Rose, D. P., and Royak-Schaler, R. Tumor biology and prognosis in black breast cancer patients: a review. *Cancer Detect. Prev.*, 25: 16–31, 2001.
- Joe, A. K., Arber, N., Bose, S., Heitjan, D., Zhang, Y., Weinstein, I. B., and Hibshoosh, H. Cyclin D1 overexpression is more prevalent in non-Caucasian breast cancer. *Anticancer Res.*, 21: 3535–3539, 2001.
- Simon, M. S., and Severson, R. K. Racial differences in breast cancer survival: the interaction of socioeconomic status and tumor biology. *Am. J. Obstet. Gynecol.*, 176: S233–239, 1997.
- 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.
- Selaru, F. M., Zou, T., Xu, Y., Shustova, V., Yin, J., Mori, Y., Sato, F., Wang, S., Oлару, A., Shibata, D., Greenwald, B. D., Krasna, M. J., Abraham, J. M., and Meltzer, S. J. Global gene expression profiling in Barrett's esophagus and esophageal cancer: a comparative analysis using cDNA microarrays. *Oncogene*, 21: 475–478, 2002.
- Xu, Y., Selaru, F. M., Yin, J., Zou, T. T., Shustova, V., Mori, Y., Sato, F., Liu, T. C., Oлару, A., Wang, S., Kimos, M. C., Perry, K., Desai, K., Greenwald, B. D., Krasna, M. J., Shibata, D., Abraham, J. M., and Meltzer, S. J. Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. *Cancer Res.*, 62: 3493–3497, 2002.
- Raychaudhuri, S., Stuart, J. M., and Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, 455–66, 2000.
- Kachigan, S. Multivariate statistical analysis: a conceptual introduction, pp. 236–261. New York: Radius Press, 1991.
- Luo, L., Salunga, R. C., Guo, H., Bittner, A., Joy, K. C., Galindo, J. E., Xiao, H., Rogers, K. E., Wan, J. S., Jackson, M. R., and Erlander, M. G. Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nat. Med.*, 5: 117–122, 1999.
- Selaru, F. M., Xu, Y., Yin, J., Zou, T., Liu, T. C., Mori, Y., Abraham, J. M., Sato, F., Wang, S., Twigg, C., Oлару, A., Shustova, V., Leytin, A., Hytiroglou, P., Shibata, D., Harpaz, N., and Meltzer, S. J. Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. *Gastroenterology*, 122: 606–613, 2002.
- Zou, T., Selaru, F. M., Xu, Y., Shustova, V., Yin, J., Mori, Y., Shibata, D., Sato, F., Wang, S., Oлару, A., Deacu, E., Liu, T. C., Abraham, J. M., and Meltzer, S. J. Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. *Oncogene*, 21: 4855–4862, 2002.
- Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Stat. Assoc.*, 74: 829–836, 1979.
- Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98: 5116–5121, 2001.
- Landgrebe, J., Wurst, W., and Welzl, G. Permutation-validated principal components analysis of microarray data. *Genome Biol.*, 3: 19.1–19.11, 2002.
- Grubberger, S., Ringner, M., Chen, Y., Panavally, S., Saal, L. H., Borg, A., Ferno, M., Peterson, C., and Meltzer, P. S. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.*, 61: 5979–5984, 2001.
- Bernstein, L., Teal, C. R., Joslyn, S., and Wilson, J. Ethnicity-related variation in breast cancer risk factors. *Cancer (Phila.)*, 97: 222–229, 2003.
- Lyman, G. H., Kuderer, N. M., Lyman, S. L., Cox, C. E., Reintgen, D., and Baekey, P. Importance of race on breast cancer survival. *Ann. Surg. Oncol.*, 4: 80–87, 1997.
- van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, 347: 1999–2009, 2002.
- van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. Gene expression profiling predicts clinical outcome of breast cancer. *Nature (Lond.)*, 415: 530–536, 2002.
- Garcia, M., Augereau, P., Briozzo, P., Capony, F., Cavailles, V., Freiss, G., Maudelonde, T., Montcourrier, P., Vignon, F., and Rochefort, H. Regulation, clinical and biological significance of cathepsin D in breast cancer. *Rev. Esp. Fisiol.*, 46: 39–41, 1990.
- Cavaillès, V., Augereau, P., and Rochefort, H. Cathepsin D gene is controlled by a mixed promoter, and estrogens stimulate only TATA-dependent transcription in breast cancer cells. *Proc. Natl. Acad. Sci. USA*, 90: 203–207, 1993.
- Nakopoulou, L., Lazaris, A. C., Kavantzis, N., Alexandrou, P., Athanassiadou, P., Keramopoulos, A., and Davaris, P. DNA topoisomerase II- $\alpha$  immunoreactivity as a marker of tumor aggressiveness in invasive breast cancer. *Pathobiology*, 68: 137–143, 2000.
- Vercoutter-Edouart, A. S., Lemoine, J., Le Bourhis, X., Louis, H., Boilly, B., Nurcombe, V., Revillion, F., Peyrat, J. P., and Hondermarck, H. Proteomic analysis reveals that 14-3-3 $\sigma$  is down-regulated in human breast cancer cells. *Cancer Res.*, 61: 76–80, 2001.
- Umbricht, C. B., Evron, E., Gabrielson, E., Ferguson, A., Marks, J., and Sukumar, S. Hypermethylation of 14-3-3 $\sigma$  (stratifin) is an early event in breast cancer. *Oncogene*, 20: 3348–3353, 2001.
- Wolff, M. S., Britton, J. A., and Wilson, V. P. Environmental risk factors for breast cancer among African-American women. *Cancer (Phila.)*, 97: 289–310, 2003.
- Olopade, O. I., Fackenthal, J. D., Dunston, G., Tainsky, M. A., Collins, F., and Whitfield-Broome, C. Breast cancer genetics in African Americans. *Cancer (Phila.)*, 97: 236–245, 2003.
- Proceedings of the Summit Meeting on Breast Cancer Among African American Women. Washington, DC, September 8–10, 2002. *Cancer (Phila.)*, 97: 207–341, 2003.
- Monni, O., Hyman, E., Mousse, S., Barlund, M., Kallioniemi, A., and Kallioniemi, O. P. From chromosomal alterations to target genes for therapy: integrating cytogenetic and functional genomic views of the breast cancer genome. *Semin. Cancer Biol.*, 11: 395–401, 2001.
- Widschwendter, M., Berger, J., Hermann, M., Muller, H. M., Amberger, A., Zeschnigk, M., Widschwendter, A., Abendstein, B., Zeimet, A. G., Daxenbichler, G., and Marth, C. Methylation and silencing of the retinoic acid receptor- $\beta$ 2 gene in breast cancer [see comments]. *J. Natl. Cancer Inst.*, 92: 826–832, 2000.
- Martin, K. J., Kritzman, B. M., Price, L. M., Koh, B., Kwan, C. P., Zhang, X., Mackay, A., O'Hare, M. J., Kaelin, C. M., Mutter, G. L., Pardee, A. B., and Sager, R. Linking gene expression patterns to therapeutic groups in breast cancer. *Cancer Res.*, 60: 2232–2238, 2000.
- Jones, B. A., Patterson, E. A., and Calvoceossi, L. Mammography screening in African American women: evaluating the research. *Cancer (Phila.)*, 97: 258–272, 2003.
- Reis, L. A. G., M. P., E., Kosary, C. L., Hankey, B. F., Miller, B. A., Clegg, L. X., Edwards, B. K., and editors SEER Cancer Statistics Review, 1973–1999. National Cancer Institute. Bethesda, MD, 2002.
- Newman, L. A., Mason, J., Cote, D., Vin, Y., Carolin, K., Bouwman, D., and Colditz, G. A. African-American ethnicity, socioeconomic status, and breast cancer survival: a meta-analysis of 14 studies involving over 10,000 African-American and 40,000 White American patients with carcinoma of the breast. *Cancer (Phila.)*, 94: 2844–2854, 2002.
- Franzini, L., Williams, A. F., Franklin, J., Singletary, S. E., and Theriault, R. L. Effects of race and socioeconomic status on survival of 1,332 black, Hispanic, and white women with breast cancer. *Ann. Surg. Oncol.*, 4: 111–118, 1997.
- Wojcik, B. E., Spinks, M. K., and Optenberg, S. A. Breast carcinoma survival analysis for African American and white women in an equal-access health care system. *Cancer (Phila.)*, 82: 1310–1318, 1998.
- Li, C. I., Malone, K. E., and Daling, J. R. Differences in breast cancer hormone receptor status and histology by race and ethnicity among women 50 years of age and older. *Cancer Epidemiol. Biomark. Prev.*, 11: 601–607, 2002.