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

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

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 status (4). High estrogen receptor (ER) mRNA levels have been associated with absence of 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 ERβ 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, unsupervised 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 μg of total RNA using a T7-based protocol (14). Labeling was performed on 3–6 μg 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).
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 coefficients of data distortion were removed by a robust scatter-plot smoothing method (17). Using SigmaPlot version 5 (SPSS, San Rafael, CA), we calculated 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.

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.

Clustering 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 clinicomolecular features: year of surgery, ethnic origin, 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 $Ps$ will be obtained when performing one-way ANOVA analyses of associations with these components. To ensure that these $Ps$ are significant, they need to be adjusted. One method to adjust $Ps$ 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 $Ps$ of the one-way ANOVA test for the association between each component and the clinical feature were recalculated. This calculation resulted in 35 different $Ps$, 1 for the association between each component and the given clinical or molecular feature. Among these 35 $Ps$, the highest $P$ was recorded. Next, step 2 was repeated 999 times, yielding a total of 999 “highest” $Ps$. 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).

Results

Clustering 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 ($P = 0.009$, one-way ANOVA). Upon additional testing, we noticed that PC 7 had an even stronger association with a subgroup of ER (+) AA patients. PC 7 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
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 \((C \text{ versus } AA; \ P = 0.002)\), PC 6 with histological subtype \((\text{strictly intraductal versus lobular/mixed}; \ P = 0.01)\), PC 9 with age \(< \text{versus} \geq 50 \text{ years old}; \ P = 0.003)\), PC 10 with tumor stage \((\text{stage 2 versus} \text{ stage 3}; \ P = 0.007)\), PC 23 with tumor size \(< \text{versus} > 5 \text{ cm}; \ P = 0.009)\), 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 putative relationships to tumorigenicity, including \(topoisomerase II \alpha\), \(IFN\)-stimulated protein, \(FB\) murine osteosarcoma viral oncogene homologue, \(early\ growth\ response\ 1\). Wilms tumor homologue, trefoil factor 3, cathepsin D, and 14-3-3-\(\sigma\).
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 α*, 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-α, 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-α as a target of methylation in breast cancers (29). It has been suggested that hypermethylation and loss of expression of 14-3-3-α 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 *Ps* 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 supplant 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-

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 first 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; Ps were obtained by one-way ANOVA; bars, ±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 □ represent the mean, the ■, ±SE, and bars ± 1.96 *SE.
ies are needed that apply PCA to predict the clinical behavior and natural history of human malignant disease.

References

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

Florin M. Selaru, Jing Yin, Andreea Olaru, et al.


Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/64/5/1584

Cited articles
This article cites 28 articles, 9 of which you can access for free at:
http://cancerres.aacrjournals.org/content/64/5/1584.full.html#ref-list-1

Citing articles
This article has been cited by 3 HighWire-hosted articles. Access the articles at:
/content/64/5/1584.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.