

Gene Expression Profiling for Prediction of Clinical Characteristics of Breast Cancer

ERICH HUANG,* MIKE WEST,[†] AND JOSEPH R. NEVINS*^{‡||}

**Department of Molecular Genetics and Microbiology; [†]Institute of Statistics and Decision Sciences; [‡]Howard Hughes Medical Institute; ^{||}Duke University Medical Center; Duke University, Durham, North Carolina 27710*

ABSTRACT

We have applied techniques of gene expression analysis to the analysis of human breast cancer by identifying metagene models with the capacity to discriminate breast tumors based on estrogen receptor (ER) status as well as the propensity for lymph node metastasis. We assess the utility and validity of these models in predicting status of tumors in cross-validation determinations. The practical value of such approaches relies on the ability not only to assess relative probabilities of clinical outcomes for future samples but also to provide an honest assessment of the uncertainties associated with such predictive classifications, based on the selection of gene subsets for each validation analysis. This latter point is of critical importance to the ability of applying these methodologies to clinical assessment of tumor phenotype. It is also clear from ER predictions that these analyses identify genes known to be involved in ER function but also identify new candidate genes involved in ER function. We believe these gene expression phenotypes have the potential to characterize the complex genetic alterations that typify the neoplastic state in a way that truly reflects the complexity of the regulatory pathways that are affected.

I. Introduction

Breast cancer is a prime example of a disease where further molecular characterization is needed to improve diagnostic and therapeutic strategies. Numerous studies have correlated genetic alterations with clinical outcome, including a strong correlation between the amplification of the *erbB-2* receptor gene (*Her-2*) and poor clinical outcome (Tandon *et al.*, 1989; Ciocca *et al.*, 1992). In addition, overexpression of *erbB-2* is a strong predictor of response to adriamycin-based therapy (Muss *et al.*, 1994). Nevertheless, such correlations are few and often do not adequately define tumor subtypes from the viewpoint of substantially impacting therapeutic decisions. The inability to define a subclass of tumor type that may be refractory to standard therapies restricts the development of new, more-efficacious therapeutic strategies.

The analysis of gene expression represents an indirect measure of the genetic alterations in tumors since, in most instances, these alterations affect gene

regulatory pathways. Given the tremendous complexity that can be scored by measuring gene expression with DNA microarrays, together with the absence of bias in assumptions as to what type of pathway might be affected in a particular tumor, the analysis of gene expression profiles offers the potential to impact clinical decision making based on more-precise determinations of tumor cell phenotypes. It is critical that such analyses characterize the inherent variability and the resulting uncertainty about the predicted clinical status of tumors with out-of-sample predictions, in order to properly assess the potential utility of such information in therapeutic decision making. This has been the focus of much of our work (West *et al.*, 2001; West, 2002).

II. Using Genome-scale Gene Expression Analysis to Study Cellular Phenotypes

Nucleic acid arrays, and the genome-scale expression data they capture, represent a major advance in the biological sciences, not only for the efficient, high-throughput data collection they afford but also for the fresh analytic methodologies and new avenues for interpreting biology demanded by data sets of such novelty, mass, and complexity. Whatever the specific platform, microarrays depend on the same principle: that complementary nucleic acid sequences will hybridize preferentially to sequences mounted on a substrate. As such, microarrays are no more than refinements of a technology that has existed and been employed ubiquitously since its description by Southern (1975). The power of microarrays depends on two significant advances: that the genomes of several organisms have been characterized or substantially characterized and the ability to precisely fix many thousands of sequences reproducibly on a substrate. The synthesis of these genomic and technical revolutions brings about an unprecedented capability to assess and quantify a significant portion, if not virtually all, mRNA sequences present in a tissue at one time. However, the sheer quantity of such data forces the biologist to face the challenge of interpreting hundreds of thousands of data points in a statistically robust and responsible manner that can be conveyed readily to the scientific public.

In its current nascent state, analysis of gene expression data is hampered by the fact that the number of expressed genes being assayed is generally orders of magnitude greater than the number of experimental samples. A typical experiment will involve at most a couple of hundred samples, while microarrays typically assay tens of thousands of sequences. This results in algorithms or models that are highly susceptible to “overfitting,” where random noise in a data set is mistaken for substantive biological structure. Until the point is reached where sample numbers begin to match the number of genes represented on an array, it is exceedingly important that investigators first genuinely understand the data-analysis methodologies they employ and how effectively these methods

distinguish noise from valid structure. Out-of-sample predictive evaluation of models, using cross-validation or bootstrap techniques, is fundamental to validate analyses and their interpretations. Traditional analytical methodologies sufficient for small data sets involving a handful of replicates and a few thousand data points are insufficient to critically evaluate genome-scale data. A further danger is assuming that traditional modes of thinking about biology are sufficient for explaining gene expression data. Such data already are demonstrating that assumptions about the linearity and independence of biological pathways belie the complexity and richness seen in genome-scale data.

An important innovation in analyzing and interpreting gene expression data has been the use of unsupervised learning procedures to identify structure in microarray data sets. A variety of methodologies — hierarchical clustering, K-means clustering, and self-organizing maps — identify genes that share similarity in their patterns of expression and group them taxonomically according to that similarity. Again, patterns of expression are determined at a single-gene level but now similarity metrics allows one to group genes into coregulated clusters.

Alternatively, one can approach genome-scale expression data as a complex molecular phenotype and seek methods that treat such data as a composite of both up- and downregulated genes *in toto* rather than as ontologically isolated genes. This approach also seeks to define structure in expression data but differs from clustering methods by directly relating structure to phenotype, rather than simply describing that structure. We term these alternative integrated data structures of coordinately up- and downregulated genes “metagenes.” We apply them to models that test the association of genes with phenotypic states and provide a predictive capability that facilitates rigorous cross-validation of conclusions generated from microarray experiments.

A decisive point in developing any analytical method for genome-scale data is drawing the distinction between using methods to *describe* the data and developing predictive *models* from the data. Data-mining techniques (e.g., unsupervised or machine learning procedures) are exceedingly powerful in identifying and describing features in large, complex data sets. Their weakness is that they only describe the data in hand and do not formally address whether features of that data are generalizable and applicable to the real world in a predictive sense. Assuming a particular data set is a reasonable sample of reality, it is quite likely that features identified by a technique such as hierarchical clustering can be thought to be representative of genuine biological processes. Still, there is no proper methodology for directly establishing or testing this link. In contrast, by interpreting a data set through developing models at the outset, the model can be used to immediately predict new samples or, more importantly, the analysis can be stressed and tested to formally ascertain whether it properly predicts outcome. While this may be a subtle point, it is exceedingly important. A data set is a sample of a real-world distribution. There always is the possibility

that the sample is biased due to chance and the wisdom extracted from the data is relevant to itself and irrelevant to the real world.

In cases where the number of samples in a data set are far fewer than the predictor variables, the only feasible means to correct for overfitting is to test a model derived from the data as rigorously as possible. By performing such corroboration, one can test the possibility that features of the data thought to be important are generated by chance. Validation procedures (e.g., testing against an independent data set, out-of-sample cross-validation) are exceedingly important. Validation represents the thin line between mistakenly accepting adventitious or confounding structure in data as being biologically relevant and truly defining structure that corresponds to biological function. Most data-mining techniques are sensitive enough to find any and all structure present. It is the responsibility of the scientist to ascertain which structure is genuinely useful.

To demonstrate the importance of this issue, we generated two data sets for analysis. The first was a “real” data set comprising replicate samples from cells with ectopically expressed E2F genes (E2F1, E2F2, and E2F3) versus a Control set. The second was a “mock” data set in which a random number generator was used to create a synthetic gene expression data set with identical but arbitrary class assignments: A, B, C, and Control. The only stipulation for the mock set was that it possess the same mean and variance as the real set. Each data set was screened for the top 100 most-correlated genes and analyzed using metagene modeling techniques (West *et al.*, 2001; Spang *et al.*, 2002). Simply stated, metagenes are singular factors (or principal components) that are derived using singular-value decomposition methods, a standard method of data decomposition that isolates unrelated linear combinations of genes that each measure aspects of the patterns of variation and covariation in the full set of data. A three-dimensional metagene plot is created for the resulting metagenes, one for each experiment (Figure 1). It is evident that metagenes can “find” structure in the data that unambiguously separate the E2F1-, E2F2-, and E2F3-expressing cells from each other and from the Control. Strikingly, a similar robust separation was achieved with the mock data set of randomly generated values. Clearly, the massive complexity of this data set allows one to find structure simply by chance. In order to generate a predictive model for E2F1 versus Control, we sought out structure specific to the E2F1 phenotype, as can be seen in the fitted classification based on combined metagene scores. The technique successfully identifies structure that can distinguish E2F1 samples from Control and does so with an estimate of probability of accurate classification. Once again, a similar analysis for the mock data set generates a classification that is as clean as for the real data. For all intents and purposes, the fitted model for the mock data appears to be as valid as the fitted model for the genuine data set.

Only under the stringent conditions of out-of-sample cross-validation does it become obvious that a model generated from the randomly generated data set

fails to deliver a genuine, generalizable predictive model for distinguishing experimentals from controls. To assess this, analysis was repeated a number of times, each time removing one of the samples from the data set, then predicting its class based on the remaining data. This hold-one-out cross-validation approach is the most-searching assessment of predictive value of a model and has been stressed as a cornerstone of our work to date. As evident in Figure 1, the mock classification model is no more accurate than a coin flip — even after many repetitions, the classification of samples was merely stochastic, while the model generated from the real data set robustly predicts E2F1 status versus Control. This example illustrates how all techniques for exploring and modeling data can be sensitive to the fallacies of overfitting. Representations of data — such as multidimensional scaling, principal component plots, or clustering schemes — can demonstrate misleadingly unambiguous class separation. However, unless that structure is built into a model and tested, it cannot be assessed whether that structure is genuine or generated by chance. This issue is important because data sets of the complexity associated with gene expression data inevitably will possess random structure that can be taken erroneously as biologically relevant.

III. Application of Gene Expression Analysis to Breast Cancer

Expression data have opened a realm of possibilities for understanding the process of neoplastic disease at the molecular level. Relying on the assumption that the complement of transcripts in a cell represents fundamental characteristics of phenotype, many groups have sought to create essentially a “molecular pathology” of neoplastic disease using gene expression data. This in the hope that data sets comprising many thousands of molecular features, as opposed to the few available previously, will allow for more-precise diagnosis, prognosis, and prediction of treatment response. Highly promising preliminary studies of leukemias, lymphomas, and solid neoplasms demonstrate that gene expression data can highlight differences between otherwise histologically identical diseases, point toward new prognostic methodologies, and even emphasize patterns idiosyncratic of specific individuals.

We have investigated the potential of metagene analysis using DNA microarray data for a better understanding of breast cancer, with the goal of harnessing gene expression data to identify novel prognostic or predictive methodologies to enhance clinical decision making. Often, traditional methods of phenotypic characterization are limited and are not able to discern subtle differences that may be important for developing a better understanding of the tumor and advancing therapeutic strategies for disease treatment. We have taken a two-pronged approach in creating a novel statistical method that provides robust probabilistic prediction and classification of tumors based on gene expression data and also permits formal assessment of the uncertainties inherent in

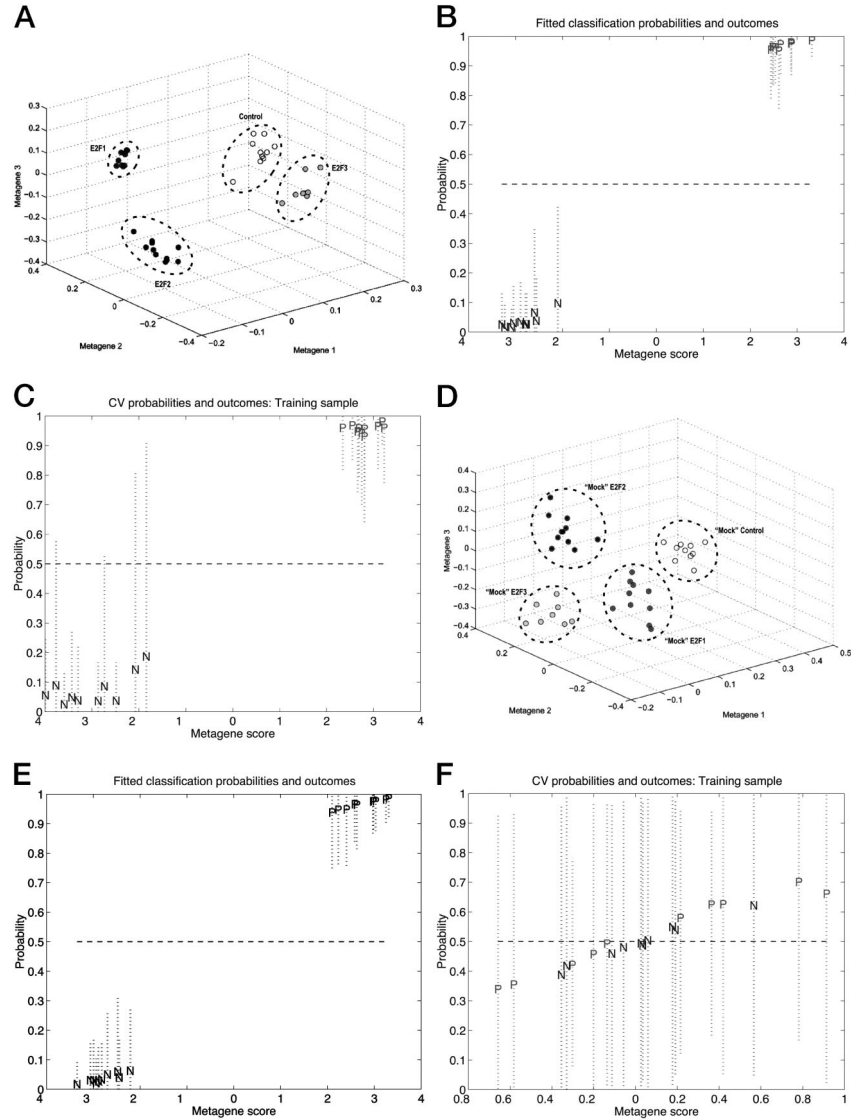


FIG. 1. An example of overfitting of data in microarray analysis. The top set of panels (A-C) represents analysis of data from a "real" gene expression data set derived from the deregulated expression of E2F proteins. Quiescent mouse embryo fibroblasts were infected with recombinant adenoviruses expressing either the E2F1, E2F2, or E2F3 protein. RNA was prepared 18 hours after infection and used to generate hybridization probes. These probes then were applied to Affymetrix Mu11K GeneChips for analysis. Panel A depicts a plot of the analysis of the data based on three

any predictive model. Such an approach is critical in an arena where clinicians must gauge their certainty of a tumor's phenotypic properties against the potential morbidities of specific interventions.

IV. Modeling Breast Cancer Based on Metagene Analysis

Metagene-based methodology for interpreting breast cancer, or any type of expression data, can be summarized in the following manner: each gene in a microarray experiment may be thought of as representing one dimension. Thus, an array representing 10,000 individual sequences defines a gene space of 10,000 dimensions. Each individual tumor sample that is hybridized to an array represents a point plotted in that 10,000-dimensional space. Therefore, 200 individual points plotted in 10,000 gene dimensions would represent a group of experiments encompassing data from 200 tumors. The data "cloud" of 200 points plotted in high-dimensional space possesses a certain structure that can be related to biological features of the data. The essence of metagene definition is drawing multiple regression lines through this cloud of data that successively whittle down the cloud's structure. Each regression line represents a metagene, a composite summarizing the impact of many genes that quantitatively weights the "pull" that each gene dimension exerts on that particular regression line. The operation of deriving metagenes is equivalent to a singular value decomposition (SVD), a standard matrix factorization in linear algebra. Many genomics visualization packages use SVD for principal component plots of experiments. The difference in the approach we describe here is that metagenes are utilized as composite-weighted proxies for multiple genes and fed into a binary regression model for classifying experiments by their gene expression patterns. Thus, each metagene is treated as if it were a single gene, although each actually summarizes the inputs of many genes, thereby simplifying the process of interpreting complex, high-dimensional data by aggregating many dimensions into fewer. In many cases when applying this approach, we can render a question of 100 dimensions into five to 10 without "losing" data. A further advantage is that aggregating genes into metagenes is an additional method of grouping genes that

(Figure 1 caption, continued)

metagenes that provide discrimination of the samples expressing the individual E2F proteins from control. Panel B depicts a fitted classification using the top 100 genes from a combined metagene to classify the E2F1 and control samples, with estimated probabilities for the classification. Panel C depicts a one-at-a-time, out-of-sample cross-validation for the E2F1 classification, demonstrating the ability of the E2F1 metagene score to predict the status of samples treated as unknown. The bottom set of panels (D-F) represent an identical analysis but using a "mock" data set generated with random numbers and with a dimension (number of data points, mean, and variance) the same as the real data set. It is evident from this analysis that patterns can be identified in this mock data that separate the samples equally well as in the real data set but then fail upon cross-validation predictions.

share some biological property. In our models of estrogen receptor (ER) status, many of the most-influential genes in metagene analysis are known ER targets.

In order to generate a classification model, we first define a biological or clinical question, then agnostically seek the metagenes that best answer the question, dispensing with noninformative or adventitious metagenes. With most questions, multiple metagenes are required. The logistic regression problem becomes a relatively straightforward binary regression involving multiple metagene predictors that, in turn, each represent multiple genes. As a proof-of-principle undertaking, we elected to model ER status because it signifies a characteristic that is both prognostic and predictive. It also is independently verifiable by both immunohistochemistry (IHC) and protein immunoblotting and many aspects of ER biology function are understood. The advantage of the considerable body of knowledge related to the estrogen hormonal axis is that the roles of genes identified as being important in the model are more likely interpretable in the context of previous knowledge. To construct this classification, we focused on developing a model that could distinguish between tumors evaluated by IHC as having no ER present (ER⁻) versus those with any receptor present (ER⁺).

V. Modeling ER Status in Breast Cancer

For this study, we used 49 tumors collected in the Breast Cancer Program at Duke University Medical Center. Tumors were classified as ER⁺ or ER⁻ at time of diagnosis by IHC. These findings later were confirmed by protein immunoblotting for ER. Conflicts between these assays were identified in five cases. Consequently, these five cases — and an additional four randomly selected tumors—were separated out and treated as validation samples for testing the predictive model developed from the remaining training cases. Of the latter, two samples were rejected due to failed hybridization, leaving a training set of 18 ER⁺ and 20 ER⁻ cases, as determined by both IHC and immunoblotting. The five tumors with contradictory tests for ER raise concerns about heterogeneity in the tumor sample; hence, these were treated as of equivocal status and scrutinized on the basis of expression-based predictions of status employing our metagene model.

For the 38 training arrays of unequivocal ER status (as confirmed by both IHC and immunoblot), we first sought to reduce noise contributed by genes irrelevant to an ER predictive model by computing sample correlation coefficients between individual genes and ER⁺/ER⁻ binary outcomes. After repeated empirical experimentation, we found that selecting about 100 genes with the largest absolute correlation coefficient values minimized noise, while providing enough relevant information for metagene models. In the absence of screening, results were comparable to what we found with a filtered subset but included all variation in the entire data set, which subsequently was integrated into the

singular-factor metagenes. This contributed adventitious influences to the metagenes and clouded the discriminatory ability of metagene models. In the ER analysis using all genes, results are broadly similar to those reported here. However, due to the much higher level of noise influencing the analysis, all predictive probabilities have much higher associated uncertainties and one or two tumors are much less well classified. Screening to a smaller, relevant, discriminatory subset of genes is guaranteed to reduce such unwanted noise, with cleaner and more-accurate results.

With a gene expression space defined by 100 genes culled from genes possessing the highest absolute correlation coefficient for ER status, we derived metagenes, of which the first (Figure 2A) provides strong discrimination between ER+ and ER- cases. By inferences on the gene regression vector in the binary regression, we found that this particular metagene displayed many significant values (not illustrated here). Though it appears that a single metagene can serve in a discriminatory model, in higher-dimensional space, many metagenes make subtle contributions. Thus, our classification model integrates several metagenes that are coalesced into a metagene score and binary regression model developed

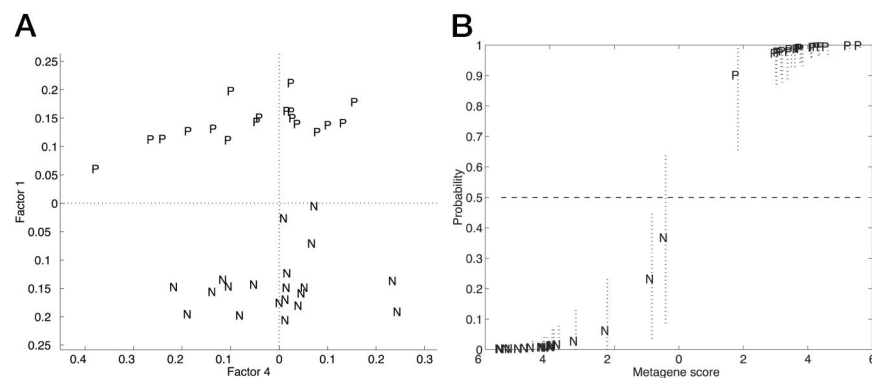


FIG. 2. Factor analysis for ER+/ER- comparison. (A) Pairwise factor analysis for the discrimination of breast tumors based on estrogen receptor (ER) status. Individual tumor samples are depicted in a scatter plot on two dominant factors underlying 100 genes selected in pure discrimination of the training cases. Each tumor is indicated as either ER+ (P) or ER- (N). Only the tumors in the training set are plotted. Factor 1 is clearly discriminatory. (Factor 4 is chosen purely for display purposes.) (B) Fitted classification probabilities for training cases from the factor regression analysis. The values on the horizontal axis are estimates of the overall factor score in the regression. The corresponding values on the vertical axis are fitted/estimated classification probabilities, with corresponding 90% probability intervals marked as dashed lines to indicate uncertainty about these estimated values. Coding is as described in panel A. [Reprinted with permission from West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR 2001 Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 98:11462–11467. Copyright 2001 National Academy of Sciences, U.S.A.]

from these inputs. Figure 2B is a graphical depiction of estimated classification probabilities for each of the training cases, with vertical bars indicating 90% probability intervals reflecting degree of uncertainty. This illustration of our fitted model, looking at each of the tumors in the developmental subset, actually represents in-sample discrimination between ER+ and ER− rather than prediction. Only by performing out-of-sample procedures do we genuinely test the efficacy of the model in classifying tumors by ER status.

Genes can be ordered by the absolute values of the estimated regression vector to provide an assessment of their relevance in the discrimination. Figure 3 depicts expression levels of the genes, with each row representing an individual gene, ordered from top to bottom according to the absolute values of the estimated regression coefficients. The group of genes includes some that function in the ER pathway, including the ER gene itself as well as a number of known ER targets (Table I). Several others contribute to the discrimination inversely with ER+ status (negative coefficients); some of these encode proteins that are known to have inverse relationships with ER function (e.g., maspin, glutathione-S-transferase (GST)-Pi). Also included are genes that are not regulated by ER but are known to function in concert with ER, such as hepatocyte nuclear factor 3 alpha (HNF3 α) and androgen receptor. Although the model is not designed to discover regulatory mechanisms, these factor models may generate clues about relationships between genes that do relate to underlying functional pathways.

VI. Cross-validation Analysis of ER Status and Honest Prediction

A major practical interest and potential clinical value of such statistical analyses lie in the ability to predict clinical conditions based on gene expression profiles of the primary tumor. In the pilot study of ER status, the prediction of status is based on gene expression patterns. The goal is to develop a rational, theoretically well-founded estimate of the probability of ER status for any new case, accompanied by a realistic assessment of uncertainty. Because such uncertainties may be high, due to limited information and population heterogeneity, it is critical that this uncertainty be reported and communicated to clinical researchers and clinicians along with point estimates of outcome probabilities.

Using the set of 100 genes selected from the full training sample study, the regression model was refitted repeatedly to the training data, each time removing the ER status of one of the tumors and then estimating the classification probability for that tumor. This is a standard, “one-at-a-time” cross-validation analysis; the status of each tumor in the training sample is predicted based on the remaining cases. For a true predictive assessment, gene screening and selection must be performed separately in each “hold-one-out” analysis, mirroring the real-life circumstances that will be faced in using such models and methods to predict future outcomes. In each of the 38 analyses, this leads to a different subset

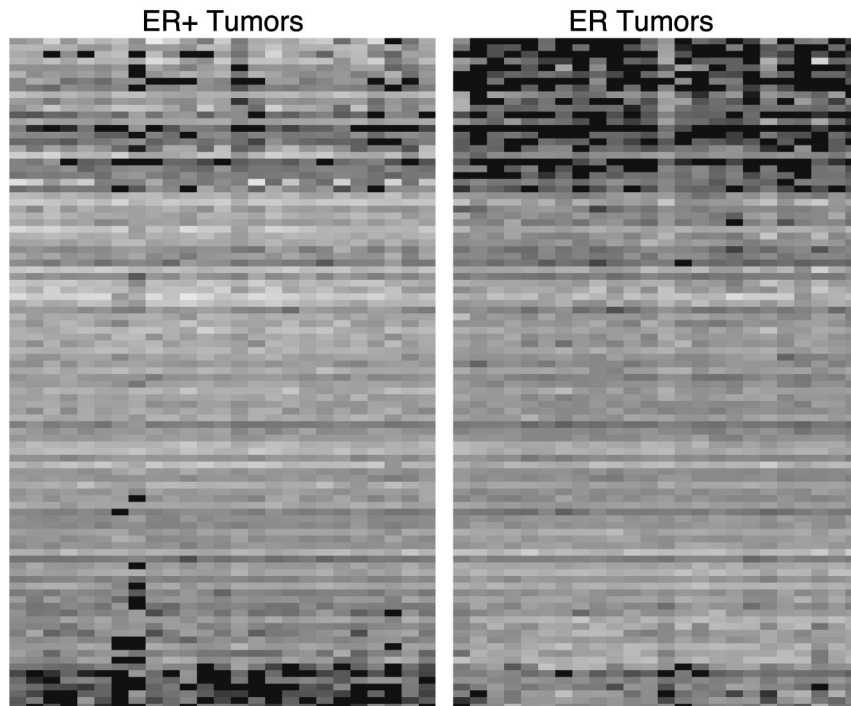


FIG. 3. Expression levels of top 100 genes providing pure discrimination of ER status. Expression levels are depicted by color coding, with black representing the lowest level and white as the highest level of expression, with shades of gray in between. Each column in the figure represents all 100 genes from an individual tumor sample, which are grouped according to determined ER status. Each row represents an individual gene, ordered from top to bottom according to regression coefficients. [Reprinted with permission from West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR 2001 Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 98:11462–11467. Copyright 2001 National Academy of Sciences, U.S.A.]

of 100 screened genes. These subsets are highly overlapping but reveal additional genes on a case-by-case basis, reflecting sample variability and inherent heterogeneity in expression profiles. Figure 4 illustrates the results: uncertainty intervals tend to be fairly wide for tumors whose predicted probabilities are in the central region (i.e., nearer 0.5 than 0 or 1). This reflects the ambiguity discovered in the expression profiles of these cases relative to the 100 genes found to be most discriminatory among the other 37 cases. These “uncertain” cases are of obvious special interest for further study. Case 16 clearly has an expression profile more in accord with those of the ER+ cases than with those sharing its designated ER– status. This case has a low level of ER gene expression, consistent with its

TABLE I
Genes That Contribute to Discrimination of Estrogen Receptor Status

Rank	Weight	Unigene cluster	Estrogen relation
1	0.08	Trefoil factor 1 (pS2)	Estrogen induced
2	0.079	Estrogen receptor (ER) 1	ER
3	0.067	Cytochrome P450, subfamily IIB	
4	0.064	Trefoil factor 3	Estrogen induced
5	0.061	(Insulin-like growth factor)	Estrogen induced
6	0.057	Human clone 23948 mRNA sequence	
7	0.056	Microtubule-associated protein tau	Estrogen induced
8	0.055	Hepsin	
9	0.048	GATA-binding protein 3	Coexpressed with ER
10	0.047	v-myb Avian myeloblastosis viral oncogene homolog	Estrogen induced
11	-0.043	Serine proteinase inhibitor, clade B, member 5 (Maspin)	Induced by tamoxifen; inverse with ER
12	0.041	N-acetyltransferase 1	
13	-0.041	S100 calcium binding protein A9	
14	-0.041	Retinoic acid receptor responder 1	
15	-0.039	Small inducible cytokine subfamily D, member 1	
16	0.039	Hepatocyte nuclear factor 3 Alpha	Synergistic with ER
17	0.038	37 kDa Leucine-rich repeat protein	
18	0.038	(Androgen receptor)	Physical interaction with ER
19	-0.038	Cathepsin C	
20	0.037	Inositol polyphosphate-4-phosphatase, type II, 105 kD	
21	0.036	Purinergic receptor P2X, ligand-gated ion channel, 4	Estrogen biosynthesis
22	-0.036	KIAA0125 gene product	
23	0.036	(Neuropeptide Y receptor Y1)	
24	0.035	Meis (mouse) homolog 3	
25	0.035	LIV-1 protein	Estrogen induced

[Genes are listed according to the discriminatory ranking, with gene 1 having the greatest weight in the discrimination. Negative values indicate an inverse correlation with ER+ status (and thus a positive correlation with ER- status). Reprinted with permission from West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR 2001 Proc Natl Acad Sci 98:11462. Copyright 2001 National Academy of Sciences, U.S.A.]

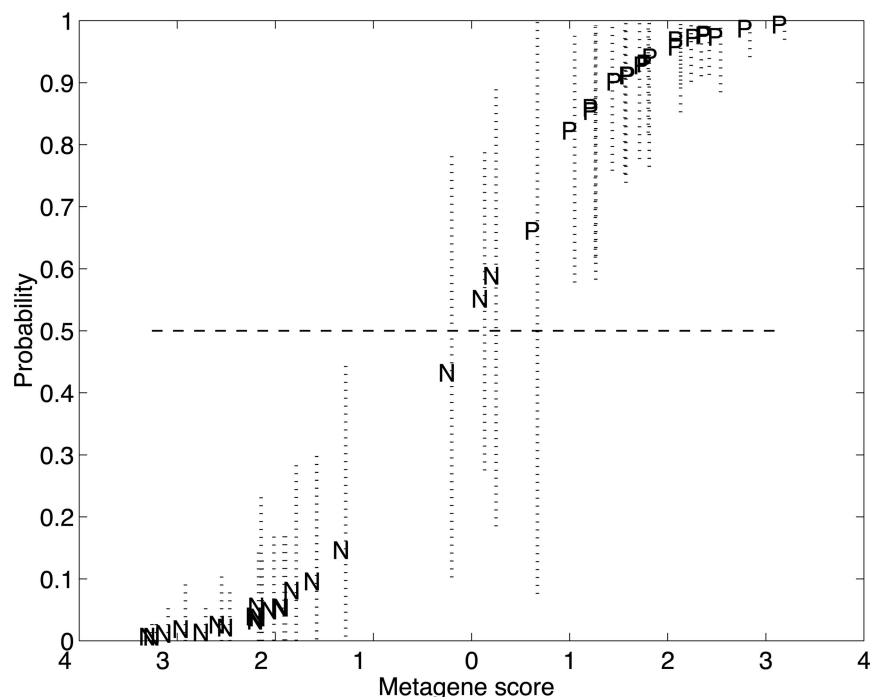


FIG. 4. Out-of-sample cross-validation predictions of ER status. One-at-a-time cross-validation predictions of classification probabilities for training cases in the ER study. In this instance, each case is predicted based only on the ER status of the remaining training tumors, with the subset of 100 genes reselected in each case. The figure presents the resulting honest uncertainties about the extent of true predictive accuracy in a practical setting, reflecting inherent variability due to heterogeneity of expression profiles. [Reprinted with permission from West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR 2001 Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 98:11462–11467. Copyright 2001 National Academy of Sciences, U.S.A.]

ER⁻ determination, but with relatively elevated levels of other genes in the top group, such as a marginally elevated level of ps2. Cases 40 and 43 share similar expression characteristics to tumor 16, exhibiting elevated levels of several known estrogen-regulated genes. In some cases, the discrepancy in clinical classification versus molecular classification is evident from the expression data. The ER⁻ cases (16, 40, and 43) that are most borderline exhibit patterns that lie somewhere between the ER⁺ and ER⁻, as does the ER⁺ case of tumor 11. Tumor 31, whose laboratory ER status determinations were conflicting, strongly exhibits a pattern consistent with an ER⁺ state.

With these exceptions, the predictive accuracy of the analysis is very high. In particular, 34 of the 38 tumor samples are predicted accurately with a high degree of confidence. Thus, not only do these expression patterns derived from regression analysis have the capacity to classify on the basis of ER status, they also have an ability to honestly predict ER status of unknown samples, demonstrating the validity of the link between expression and clinical phenotype. Note, again, the clear differences between this display and that of Figure 2B and the extent to which the true prediction in Figure 4 highlights the increased uncertainties about cases 16, 40, and 43 in the middle ground.

The validation procedures we have performed are essential for testing whether the metagene structure genuinely reflects, in a predictive manner, biological characteristics related to ER status. In the absence of such testing, the identified distinctions between ER+ and ER- tumor samples are, at best, highly promising but, at worst, the product of chance.

VII. Models That Predict Lymph Node Metastasis in Breast Cancer

The analysis of ER status demonstrates the power to predict status of samples with associated assessments of predictive uncertainties. A second analysis concerns the clinically important issue of metastatic spread of the tumor. Determination of the extent of lymph node involvement in primary breast cancer is the single most important risk factor in disease outcome (Shek and Godolphin, 1988). Here, the analysis compares primary cancers that have not spread beyond the breast to ones that have metastasized to the axillary lymph nodes at the time of diagnosis. The potential power in making this determination from the primary cancer is significant in those instances where a positive lymph node might be missed or where a tumor is poised to metastasize to the lymph node but has not yet done so.

We identified tumors as “reported negative” when no positive lymph nodes were discovered and “reported positive” for tumors with at least three identifiably positive nodes, resulting in 12 reported positives [1] and 22 reported negatives [0]. Following screening to select the top 100 most-correlated genes, the metagene analysis was performed as described for ER discrimination. As in the ER study, this first analysis used an overall screened subset of 100 genes and again provided a good classification based on lymph node status, quite comparable to that for ER discrimination. Figure 5 illustrates the practically relevant cross-validation analysis that adopts a screen to select potentially different genes for each held-out case. The screened subsets of 100 most-discriminatory genes vary more widely than that seen in the ER analysis as we move across tumors, reflecting higher levels of natural variation in gene expression patterns with respect to nodal status. All of the reportedly positive cases appropriately have estimated probabilities above 0.5, though some are close to that boundary with

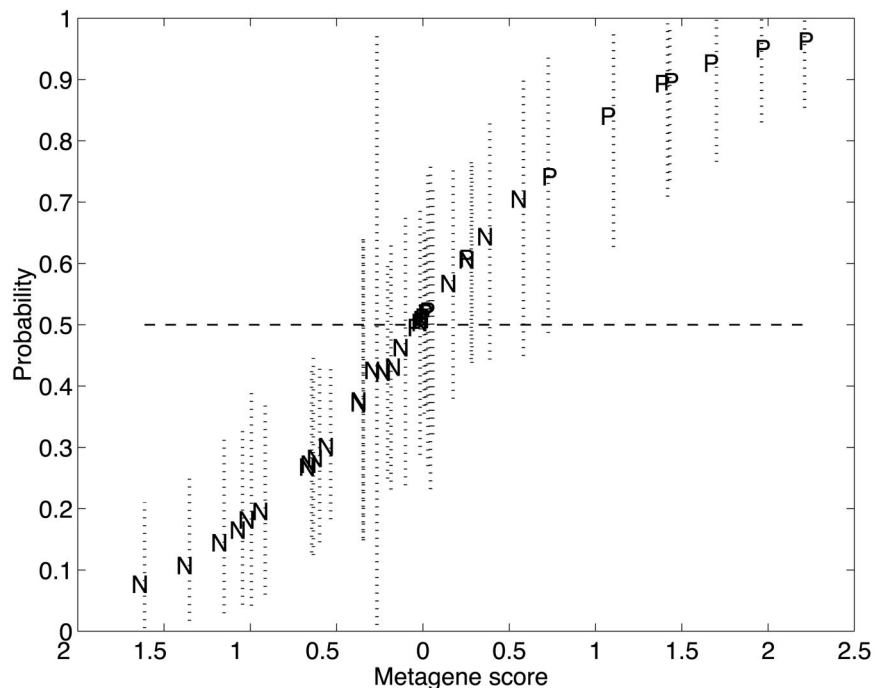


FIG. 5. Analysis for nodal comparisons. One-at-a-time cross-validation predictions in the analysis of lymph node metastasis. Each case is predicted based only on the nodal status of the remaining training tumors, with the subset of 100 genes reselected in each case. As such, the analysis exhibits the resulting uncertainties about the extent of true predictive accuracy in a practical setting, reflecting inherent variability due to heterogeneity of expression profiles. Node-positive tumors are indicated as P and node-negative tumors as N. [Reprinted with permission from West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR 2001 Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 98:11462–11467. Copyright 2001 National Academy of Sciences, U.S.A.]

moderate uncertainty. Perhaps most interesting are the few reportedly negative cases whose predicted probabilities slightly exceed 0.5. Cases like this are of paramount interest, since identifying genomic predictors of the progression from node negative to positive is a major goal from the viewpoint of potential therapeutic implications. These cases could, in principle, represent tumors that have metastasized but were missed in the nodal determination or they could be cases that have not yet metastasized but are poised to do so. Such analysis of nodal status clearly illustrates the importance of honest cross-validation studies of predictions in gauging the validity of the classification. The honest cross-validation predictions reveal realistic levels of uncertainty, likely due to heterogeneity in the profiles and the clinical phenotypes, and stress the importance of

the validation studies to verify the significance of the classification. Nevertheless, the analysis does identify gene expression patterns that have predictive capability. Clearly, it is the analysis of those tumors in the uncertain region that must be the focus of further studies.

VIII. Future Considerations

Numerous studies of leukemia, lymphoma, and breast cancer indicate that analysis of gene expression reveals patterns that can serve to classify tumors and define tumor subtypes. The method we describe here first explicitly poses a clinical question — in this case, whether ER+ and ER− tumors can be distinguished by the mRNA of primary tumor samples and whether axillary lymph node status can be classified by similar methods. Methodologically, this differs significantly from the dominant analytical method for functional genomics, hierarchical clustering. The most important difference is that it is, at the outset, a model-building approach that seeks an outcome, finds the data that best fit the outcome, and critically evaluates whether the outcome and data are robustly linked. In contrast, no question is posed with clustering. Clustering organizes a data set without human intervention solely by a similarity criterion, thus generating taxonomic trees of genes and experiments that appear related. This technique belongs to a class of data-mining tools called unsupervised learning algorithms. Clustering has demonstrated remarkable power for delineating the superorganization of data inherent in gene expression data sets. From the point of view of delineating structure, clustering and metagene-based approaches are quite comparable. It is in actively evaluating the link between a clinical or phenotypic question that the metagene approach has the advantage of providing quantitative relations of genes to outcome rather than to themselves. In purely utilitarian terms, because the process of deriving metagenes directly rotates and recomputes the values in gene expression hyperspace for each gene in the context of overall structure in the data, the rotated expression value can be fed directly into the models linking that data to an outcome. With clustering, no rotation takes place: relations between genes are drawn by similarity metrics. Inevitably, because of the complexity and sheer quantity of gene expression data, much structure is due to chance. Testing whether the composition of a cluster or membership of a particular gene in a cluster are the consequences of random processes becomes very difficult to assess. Because metagenes are rigorously assessed by their ability to successfully answer a clinical question, the relation of structure and phenotype is more rigorous and transparent.

At the same time, a model-based methodology forces the investigator not only to pose a question but also to pose the right question. Seeking a classification that cannot be corroborated in gene expression will result in weak models. Even if qualitative visual assessments by two-dimensional metagene plots,

multidimensional scaling, or clustering show a dramatic separation between experiments and controls, this separation is only quantifiable and testable if one proceeds to the next step and builds a predictive model based on that ostensible separation. In our comparison of a genuine and a synthetic data set, data-mining techniques found unambiguous and seemingly irrefutable gene expression patterns that properly separated two different classes. Even fitted models appeared to probabilistically confirm the validity of structure found in the synthetic set that drew the distinction between classes. Out-of-sample cross-validation, the gold standard stress test for a predictive model, represents the last word on whether structure is real or merely stochastic. It was only through this method that real data could be distinguished from randomly generated data. Similarly, with any gene expression experiment, if a biological question cannot be answered by the data, the weakness of the predictive model built on that assumption will be exposed — and very likely only be exposed — by validation testing.

With these considerations in mind, the analyses presented here demonstrate that clinically relevant phenotypes can be determined for primary breast tumor samples through the analysis of gene expression. We also develop predictive analyses that bring gene expression analysis to real-world clinical applicability, facilitating the use of complex gene expression patterns as discrete prognostic or predictive factors. Similar studies have utilized gene expression profiles in out-of-sample cross-validation studies. Most approaches use some form of initial gene screening to select discriminatory subsets. We have stressed and illustrated the practical importance of repeating such gene-screening exercises within each cross-validation, in order to adequately assess realistic uncertainties about predictions and avoid misleading confidence in predictive accuracy and validity.

Classifications of leukemias and lymphomas that have been achieved in recent analyses of gene expression patterns represent a significant step in the development of methodologies to phenotype tumors (Golub *et al.*, 1999; Alizadeh *et al.*, 2000). The analysis of breast cancer phenotypes likely represents a context of considerably more biological heterogeneity, reflecting subtle aspects of tumor phenotype. As such, the fact that the cross-validation predictions reveal tumors with an uncertain classification, particularly for the lymph node analysis, is not unexpected. Indeed, it would be surprising to find that such an analysis would yield two cleanly separated groups. In this context, it is critical to develop methods that not only validate classifications with out-of-sample cross-validation methods but also provide appropriate and adequate assessments of the inherent uncertainties found with such predictions. The predictive or prognostic capacity demonstrated here is particularly relevant because clinical decision making depends on a rational, theoretically well-founded model for assessing data from new patients. Because such prognostic and predictive factors are couched in probabilistic language, clinicians can make judgments based on unbiased assessments of the uncertainties in a classification.

The assay of ER status by IHC is far from perfect and can produce erroneous results, as highlighted by our study. In addition, such assays would not score alterations that disable the ER pathway as opposed to the receptor itself. Thus, if the clinically significant determination is the functional status of the pathway, not the status of ER itself, then measurements of gene expression profiles that reflect activity of the pathway could provide an important advance in understanding the behavior of breast cancers. The finding that the group of genes that contribute most weight to the discrimination include not only ER and ER pathway genes but also genes that encode proteins that synergize with ER (e.g., HNF3 α , androgen receptor) points to the potential power of the analysis in identifying functionally significant relationships.

An additional important benefit of these analyses is the potential for identifying gene pathways underlying an observed phenotype. A key point is the capacity to identify not just highly expressed genes but also those whose expression pattern highly correlates with the phenotype, regardless of level of expression. Perhaps most important is the fact that these analyses identify not only genes expected to be involved in the phenotype (i.e., ER-regulated genes), thus validating the process, but also genes for which a connection is not immediately clear. It is the identification of this latter group of genes that represents a major focus of these studies: the use of expression analysis to identify genes that highly correlate with the observed phenotype, thus providing additional insight into the underlying biological pathways.

Finally, we note that the presence of metastatic breast cancer in axillary lymph nodes is the most-significant factor in overall survival (Shek and Godolphin, 1988). Although the determination of lymph node status is relatively routine, the surgical procedure is highly invasive. Selectivity in the process of identifying nodes for examination induces biases that suggest some reported negatives may, indeed, be truly positive (Kjaergaard *et al.*, 1985; Hill *et al.*, 1999). Furthermore, the ability to accurately predict axillary lymph node status based on an expression profile of the primary tumor may obviate the need for axillary lymph node dissection and the significant morbidity associated with this procedure. Perhaps of more significance is the patient with truly negative lymph nodes but a primary tumor that is poised to metastasize. Many more data are needed to determine the precision of the predictive capability for lymph node status but it is clearly possible that a gene expression profile could predict metastatic potential, even in the absence of reportedly positive nodes.

REFERENCES

- Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani T, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM** 2000 Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503–511
- Ciocca DR, Fujimura FK, Tandon AK, Clark GM, Mark C, Lee Chen GJ, Pounds GJ, Vendely P, Owens MA, Pandian MR** 1992 Correlation of HER-2/neu amplification with expression and with other prognostic factors in 1103 breast cancers. *J Natl Cancer Inst* 84:1279–1282
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES** 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537
- Hill AD, Tran KN, Akhurst T, Yeung H, Yeh SD, Rosen PP, Borgen PI, Cody HS** 1999 Lessons learned from 500 cases of lymphatic mapping for breast cancer. *Ann Surg* 231:148–149
- Kjaergaard J, Blichert-Toft M, Andersen JA, Rank F, Pedersen BV** 1985 Probability of false negative nodal staging in conjunction with partial axillary dissection in breast cancer. *Br J Surg* 72:365–367
- Muss HB, Thor AD, Berry DA, Kute T, Liu ET, Koerner F, Cirrincione CT, Budman DR, Wood WC, Barcos M** 1994 c-erbB-2 expression and response to adjuvant therapy in women with node-positive early breast cancer. *N Engl J Med* 331:211
- Shek LL, Godolphin W** 1988 Model for breast cancer survival: relative prognostic roles of axillary nodal status, TNM stage, estrogen receptor concentration, and tumor necrosis. *Cancer Res* 48:5565–5569
- Southern E** 1975 Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98:503–517
- Spang R, Zuzan H, West M, Nevins JR, Blanchette C, Marks J** 2003 Prediction and uncertainty in the analysis of gene expression profiles. *Silico Biol*, in press
- Tandon AK, Clark GM, Chamness GC, Ullrich A, McGuire WL** 1989 HER-2/neu oncogene protein and prognosis in breast cancer. *J Clin Oncol* 7:1120–1128
- West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR** 2001 Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 98:11462–11467
- West M** 2003 Bayesian factor regression models in the “large p, small n” paradigm. *Bayesian Statistics*, in press