

**Predicting the Clinical Status of Human Breast Cancer using
Gene Expression Profiles**

Mike West¹, Carrie Blanchette², Holly Dressman³, Erich Huang³, Seiichi Ishida³,
Rainer Spang¹, Harry Zuzan¹, Jeffrey R. Marks², and Joseph R. Nevins^{3,4}

¹Institute of Statistics and Decision Sciences

Duke University

²Department of Surgery

³Department of Genetics

Duke University Medical Center

⁴Howard Hughes Medical Institute

Durham, NC 27710

Abstract

The practice of tumor diagnosis depends largely on visual interpretation of gross pathological and histological specimens together with limited biochemical data. These visual features, as well as immunohistochemical staining patterns, are reflective of the genes expressed within the tumor cell. By measuring gene expression directly, there is the potential for refining the diagnosis and classification of neoplastic tissues based on thousands of parameters where previously only a few existed. To do this, we have developed Bayesian statistical regression models that provide predictive capability based on gene expression data derived from DNA microarray analysis of a series of primary breast cancer samples. These patterns have the capacity to discriminate breast tumors on the basis of estrogen receptor (ER) status, and also on the basic categorized lymph node status. Most importantly, we assess the utility and validity of such models in predicting status of tumors in cross-validation determinations. The practical value of such approaches relies critically on the ability to not only 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 in the ability to apply these methodologies to actual tumor diagnosis.

Introduction

The delineation of phenotypes in biological systems is critical to the understanding of diverse phenomena including the development of human disease. For example, the understanding of human cancer depends heavily on the recognition of characteristics of tumor cells, including the phenotypic characterization of the development of tumors from the initial stages of the disease to the advanced, aggressive forms. In many instances, however, the currently used methods for phenotypic characterization are limited and do not have the ability to discern subtle differences that may be of importance to developing a better understanding of the tumor and the development of therapeutic strategies for the treatment of the disease. Breast cancer provides an example where further molecular characterization is needed to improve diagnosis and therapeutic strategies. Numerous studies have correlated genetic alterations with clinical outcome including a strong correlation between amplification of the ERBB2 receptor gene (Her-2) and poor clinical outcome^{1,2}. In addition, overexpression of erbB-2 is a strong predictor of response to adriamycin based therapy³. Nevertheless, such correlations are limited in number and often do not adequately define tumor subtypes.

The inability to define a subclass of tumor type that may be refractory to standard therapies restricts the development of new therapeutic strategies that may be more successful in these cases. Access to complex gene expression information generated by DNA microarray analysis provides the potential for the identification of molecular phenotypes of tumors that can be correlated with the clinical outcome of the therapy. Although the analysis of gene expression does not directly measure the genetic alterations known to define cancer cells, it certainly does represent an indirect measure of these alterations since, in most instances, the genetic 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 prospect of creating much more precise determinations of tumor cell phenotypes. Indeed, several reports have described the use of DNA microarrays to define gene expression patterns that characterize human tumors⁴⁻⁶, including for the purpose of identifying distinct classes of tumors that might aid in clinical decisions and treatment strategies⁷⁻¹⁰.

We have used DNA microarray expression data from a series of primary breast cancer samples to discriminate and predict the estrogen receptor status of these tumors as well as the lymph node status of the patient at the time the tumor was surgically removed. In both cases, statistical methodologies have revealed patterns that can usefully discriminate clinical status, and identify clusters of genes that underlie the discrimination. Much more importantly, the methods have been tested in out-of-sample predictions of clinical status to assess the potential utility of such information in therapeutic decision-making. Our analysis of ER status illustrates high predictive accuracy in such out-of-sample validation studies, and the analysis of nodal status further illustrates both discrimination and prediction, now in a context where the inherent heterogeneity is much more substantial. The latter analysis also emphasizes the need for honest assessment of uncertainties about out-of-sample predictions. As such, this work goes well beyond analyses that simply attempt to organize expression data into two groups without formal, quantitative assessment of the relevance of the discrimination in an out-of-sample, case-by-case predictive context, together with associated assessments of uncertainties about the predictions. Further, we emphasize that the practice of screening to select discriminatory subsets of genes that are then used for within-

sample cross-validation studies is inherently flawed since it leads to overly confident conclusions about the extent of accuracy of discrimination in a true predictive context, and will generally prove to be less reliable in future classification of new cases than expected, as a result. Screening for gene subsets is indeed critical, but simply must be couched in a predictive context to have practical validity. Our analyses of both ER and nodal status outcomes exemplify this point quite clearly.

We discuss and exemplify these concepts in both ER and nodal status studies, where we have identified subsets of discriminatory genes and elucidated the honest uncertainties about predictive validity that are due to inherent heterogeneity in expression patterns of tumors relative to these clinical outcome categories. We also question, again with examples, the current practice of others in selecting discriminatory subsets of genes based on tumor samples that are then used for cross-validation studies, but without repeating the gene subset selection procedure for each cross-validation analysis. Finally, we describe the utility of our analysis approach in aiding the identification of specific transcripts that contribute most significantly to a classification, facilitating correlation of underlying biology to a clinical phenotype.

Results

Measurement of gene expression in primary breast cancer

Our purpose in these studies is to develop methodologies that permit the classification of breast cancer subtypes that relate to clinical phenotypes, including response to therapy, and to use these classifications to predict status of an unknown tumor sample. We chose tumor samples from the Duke Breast Cancer SPORE tissue resource that contains frozen breast tumor specimens together with all pertinent clinical and pathologic information. The collection of samples includes mostly Stage II cancers and above (Table 1). All cancer samples have the same histology (invasive ductal carcinoma) and each is between 1.5 and 5 cm in their largest dimension. The tumor samples were chosen to include roughly an equal representation of hormone receptor–positive versus hormone receptor–negative cancers. All tissues were screened for tumor content and cases that contained less than 60% tumor cells were excluded. We have chosen to analyze the bulk tumor samples without attempt to separate tumor cells from other contributing cell types. We reasoned that even with the heterogeneity of solid tumors such as breast cancer, it would still be possible to identify gene expression patterns even if the overall magnitude of expression varied between samples.

RNA was prepared from each of the frozen tumor samples and utilized for hybridization to high density oligonucleotide DNA microarrays. We made use of Affymetrix GeneChip DNA arrays containing approximately 7100 human gene sequences and ESTs. RNA from each of the samples was converted to target following established procedures as described in Methods and then used to hybridize to the GeneChip arrays. The hybridized chips were then processed and analyzed as described in Methods. For the purpose of this study, we have taken a modified version of the Affymetrix average

difference (AD) as a summary measure of expression. Further analyses using both this and the Affymetrix average log ratio measure together in our statistical model provide similar results. The Affymetrix GeneChip software computes the summary AD measure for each gene; our analysis as reported here utilizes the transformed value $\log_2(\max(1, AD))$; i.e., adopting a log scale and truncated so that genes estimated as unexpressed by AD are set at zero in the data for analysis. Affymetrix control sequences were removed prior to analysis.

Classification of tumor samples based on ER status

As an initial step in the development of methodologies for tumor diagnosis by gene expression profiling, we analyzed the series of primary breast cancer samples to evaluate the ER status. Our approach develops probit regression models that predict outcome status as a binary indicator. For any tumor, write $p(\mathbf{x})$ for the probability of ER+ versus ER- status conditional on the vector \mathbf{x} of measured expression levels for all genes on the array. A standard probit regression sets $p(\mathbf{x})=F(\mathbf{b}'\mathbf{x})$ where F is the standard normal distribution function of the linear function of the elements of \mathbf{x} , namely $\mathbf{b}'\mathbf{x}$ where \mathbf{b} is a regression coefficient column vector and \mathbf{b}' is its row vector transpose. The standard computational machinery of Bayesian statistical modeling allows, in principle, for model fitting to observed data — a set of observed expression profiles and the corresponding ER status — and for out-of-sample validation based on predicting the values of $p(\mathbf{x})$ for new tumors not used in the model fitting. The critical issue addressed in this work is that the number of regression parameters in \mathbf{b} — one per gene — is very much larger than the number of available microarray samples — typically several tens or low hundreds at the very best. Hence there is an implicit need for massive dimension reduction, but with concern that gene-specific information be recovered from the analysis.

Singular–value decomposition (SVD) analysis of expression array data sets is extremely valuable in exploratory analysis ¹¹ as well as a key element of our formal statistical models. Write \mathbf{X} for the matrix whose columns are the vectors \mathbf{x} of expression for all arrays, including cases to be predicted for validation as well as cases to be used in model fitting. The SVD maps expression levels in \mathbf{X} to a smaller matrix \mathbf{F} whose columns \mathbf{f} are values of singular factors. With a total of n arrays there are at most n such factors. These "supergene" factors underlie expression patterns in that the level of expression of each gene is a linear combination of the values of the supergene factors; each gene has a "weight" on each factor in this linear combination. The regression probability $p(\mathbf{x}) = \mathbf{F}(\mathbf{b}'\mathbf{x})$ for an array has the equivalent form $\mathbf{F}(\mathbf{g}'\mathbf{f})$ where \mathbf{f} is the vector of supergene factors for that array, and \mathbf{g} is a regression coefficient vector. With n expression arrays to be analyzed or predicted, \mathbf{f} has only n elements, so represents a massive reduction in dimension from the several thousands in \mathbf{x} , and our Bayesian analysis methods ¹² provide for model fitting and prediction. Most importantly, our theory shows that the Bayesian SVD regression framework allows direct inversion to infer the parameters \mathbf{b} from \mathbf{g} . This is critical as it provides inferences about which genes are important in defining $p(\mathbf{x})$, and how subsets of genes interact.

The initial analysis involved a set of 49 tumors, each classified as either ER+ or ER– based on immunohistochemistry (IHC) at the time of diagnosis. These samples were made available sequentially; we received a first set of 43, followed by the second set of 6. To explore out–of–sample predictive validity of the statistical model, we treated the second set of 6 tumors as a validation sample, i.e., 'new' cases to be predicted. Of the 43 training cases, two cases have expression array images at generally very low levels of intensity, reflecting problems in hybridization, and were removed from the study, leaving 41 training

cases. Our initial statistical analysis raised questions about the original ER classification of a few tumors, so the full set of tissues was subjected to a protein immunoblotting assay for estrogen receptor. The immunoblotting test results conflict with the initial classification in 3 of the training cases and 2 of the 6 validation cases. As a result, these 3 training tumors are treated as of unknown ER status, and are combined with the 6 validation cases to be predicted, assuming that the agreement of the two biological methods of ER status assessment implies a certain determination. Hence, the validation sample has a total of 9 tumors, and of the 38 training cases we have 18 ER+ and 20 ER- as determined by immunohistochemistry.

Using the ER outcomes of only the 38 training arrays, we first implemented a simple screen to identify the 100 genes maximally correlated with outcome. This screening strategy aims to reduce noise contributed by irrelevant or unexpressed genes by an initial selection process, and the choice of the number 100 was determined by repeat experimentation. This screen computed sample correlation coefficients between genes and ER+/ER- binary outcomes, and selected those genes giving the 100 largest absolute values of this correlation. The binary regression model was then fitted to this set of 100 selected genes, using the resulting SVD factors based on these 100 genes. Figure 1A shows that two of the implied supergene factors in the vector f together provide a good discrimination between the ER+ and ER- cases. That this discrimination is related to many genes, not just a few of the 100, is clear from an examination of inferences on the b vector, indicating many significant values (not illustrated here).

Figure 1B depicts the estimates of classification probabilities $p(x)$ together with associated 90% probability intervals illustrating the degree of uncertainty. This figure must be interpreted carefully since it shows fitted classification probabilities for each of the 38

training cases, but does not predict the probability of a sample fitting into a classification. This analysis does provide a useful visual assessment of how clearly the samples are discriminated.

The 'top 100' genes can be ordered by the magnitude of the estimated \mathbf{b} vector to provide one initial set of genes ordered by just how much they contribute to this discrimination analysis of the training data. The genes that contribute to the scoring are listed in Table 2 along with the estimate regression parameters for the top 100 genes in this discrimination. In addition, the expression levels of the genes is depicted in Figure 2, ordered as a function of estimated regression coefficients. The group of genes includes many that directly or indirectly function in the estrogen receptor pathway, including the estrogen receptor gene itself as well as a number of known targets for ER. Perhaps more importantly, the group also includes genes that are not regulated by ER but instead are known to function in concert with ER such as HNF3 α and androgen receptor, suggesting that the discrimination is not just similar expression patterns but also functional relationships. In addition, it is also clear that some of the genes that make significant contributions to the discrimination do so inversely with ER+ status (exhibit a negative regression value); many of these encode proteins that are in fact known to have inverse relationships with ER function. For instance, the expression of maspin and GST-Pi are inversely related to ER status. Interestingly, maspin has also been shown to be induced by the estrogen antagonist tamoxifen.

Figure 3 illustrates the formal and honest predictive view, and the real accuracy of the statistical methods developed here, in the predictions for the 9 cases in the validation sample based on the factor score derived from the analysis of expression of the 100 genes. Some of these cases are evidently quite surely predicted as of either ER+ or ER- status, but

those in the central region are uncertain, and the probability intervals reflect this uncertainty. Tumor samples 45 and 46 were initially determined to be ER– at the time of diagnosis by immunohistochemistry; subsequent analysis by immunoblotting indicates an ER+ status. This change in ER status could reflect an initial borderline reading at the time of diagnosis that was more clearly positive by immunoblot assay or it could reflect tumor heterogeneity that influenced the assay based on sampling differences. On the basis of the statistical analysis and prediction, it is clear that the expression profiles are much more consistent with the immunoblotting results. Tumor samples 14, 31, and 33 were initially determined to be ER+ by immunohistochemistry but subsequent analysis by immunoblotting indicates an ER– status. Again, this difference could reflect tumor heterogeneity. In these cases, the statistical analysis indicates an expression profile consistent with the initial determination of a positive ER status for tumor 31, and the subsequent immunoblotting result of negative ER status for tumor 33; for tumor 14, the expression profile yields an uncertain prediction.

Cross validation analysis of ER status

Figure 4A depicts the results of an initial cross-validation analysis. Using the set of 100 genes selected from the full training sample study, the regression model was repeatedly refitted to the training data, each time removing the ER status of one of the tumors and then estimating the classification probability for that tumor. The figure displays estimates of classification probabilities $p(\mathbf{x})$ together with associated 90% probability intervals illustrating the degree of uncertainty, and provides a useful visual assessment of how clearly the samples are discriminated, together with an indication of uncertainties in the discrimination. Some of the uncertainty comes from the small sample of tumors, but much

arises as a result of the heterogeneity in the expression data with respect to the ER classification. This is a standard, 'one-at-a-time' cross-validation prediction analysis; the status of each tumor in the training sample is predicted based on the remaining training cases. Quite critically, these analyses are each based on the pre-screened subset of 100 genes that are most discriminatory for the full training sample. We note, and stress, that this parallels precisely the format of analysis adopted by other recent studies in the use of an overall initial screen to select a small subset of genes (though, of course, our statistical modeling framework is quite different and novel).

Honest prediction of ER status in the cross validation analysis

The discriminatory classification of the tumor samples as achieved in Figure 1, 3, and 4A, as in similar recent studies, is useful. However, the major practical interest and potential clinical value of such statistical analyses lies in the ability to predict the status of new cases based on a gene expression profile; that is, to provide a rational, theoretically well-founded estimate of the probability of ER status for any new case, and accompanied by a truly honest assessment of uncertainty. Such uncertainties may be high due to limited information and population heterogeneity, and it is critical that this uncertainty be reported and communicated to clinical researchers and clinicians along with point estimates of outcome probabilities. Statistical analysis must address the need for such an honest assessment head-on.

With this goal, it is quite clear that the accuracy in discrimination apparent in Figure 4A is illusory from a predictive viewpoint. The pre-screening strategy underlying that figure biases the analysis towards a clearer discrimination, and is simply not a reliable indicator of how the analysis will perform in real application. For a true, honest predictive

analysis and assessment, the screening to select a subset of 100 genes must be performed repeatedly: removing a tumor from the training set to predict it based on the rest, we must then screen to a reduced subset of 100 most predictive genes by applying the screening criterion (here, simple correlation with ER status) to only the remaining training samples, NOT including the one held out. This honest analysis mirrors the real-life circumstances that will be faced in using such models and methods to predict future outcomes. In the cross-validation study, holding out each of the 38 training tumors one at a time then leads to a set of 38 different samples of 100 screened genes, one for each case. Certainly these sets of genes are highly overlapping, containing many genes in common, but also showing some variety as we move between hold-out cases. This variation in screened subsets reflects sample variability and inherent heterogeneity in expression profiles of breast tumors, and is, in contrast, inappropriately ignored by the earlier analysis using a single, overall screen as is common in the literature.

Figure 4B illustrates the cross-validatory predictions resulting from this formal and honest predictive analysis. Note that the uncertainty intervals tend to be fairly wide for tumors whose predicted probabilities are in the central region, 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. Note also the case of tumor 16; in this predictive sense, its expression profile is more in accord with those of the ER+ cases than with those sharing its designated ER- status. This case has a low level of expression of the estrogen receptor gene, based on the microarray data, consistent with its ER- determination, but with relatively elevated levels of other genes in the top group, such as a marginally elevated level of pS2. The two additional highly uncertain cases, numbered 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. For instance, tumors 7 and 8, for which the hybridization levels are low and problematic, exhibit a generally reduced level of all genes in the group (Figure 2). The ER- cases 16, 40, and 43, that are most borderline in the discrimination, also 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, clearly exhibits a pattern consistent with an ER+ state.

With these exceptions, the predictive accuracy of the analysis is very high. In particular, it is important to recognize that in this analysis of 38 tumors, 34 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 have an ability to predict ER status of unknown samples, demonstrating the validity of the link between expression and clinical phenotype. Note further, to reaffirm the earlier discussion, the clear differences between this display and that of Figure 1B, and the extent to which the clean classification in Figure 4A is shown to be less reliable than is suggested when compared with the more relevant and appropriate results in Figure 4B. In particular, the latter highlights the increased uncertainties about cases 16, 40, 43 in the middle ground.

Classification of breast cancer based on lymph node status

The analysis of ER status through gene expression measures demonstrates the power to predict status of samples with associated assessments of uncertainty about the predictions. To extend this analysis to a practical and clinically relevant example, we have explored the methodology in analysis of the reported lymph node status of the same set of tumors. The

determination of the extent of lymph node involvement in primary breast cancer is the single most important risk factor in disease outcome¹³. The potential power in making this determination at 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.

Analysis of the nodal status of this set of breast tumors forms an initial study in comparison of primary cancers that have not spread beyond the breast to ones that had metastasized to the axillary lymph nodes at the time of diagnosis. In the total of 47 training cases (ignoring the 2 of the initial 49 that had very low hybridization levels) we identified tumors as "reported negative" for cases where no positive lymph nodes were discovered, and "reported positive" for tumors having at least 3 identifiably positive nodes. The 13 cases with only 1 or 2 positive nodes were treated as uncertain, comprising validation samples to be predicted from analysis of the remaining 34 training cases. Of these 34, 12 are in the "reported positive" (1) class, and 22 in the "reported negative" (0) class. Gene expression measurements were analyzed using the combined SVD and regression approach, as in the ER study.

Figure 5A shows that the analysis generates a factor structure underlying the discrimination according to reported lymph node status. The fitted probabilities from the binary regression model analysis, together with estimated uncertainties, are shown in Figure 5B. This analysis again provides a good classification based on lymph node status, quite comparable to that for the ER discrimination.

Table 3 provides a list of the 100 genes with highest estimated regression coefficients in this discrimination, and Figure 6 depicts the expression levels of these genes grouped according to lymph node status. It is clear that the reported node negatives and

positives share gene expression patterns beyond those that directly relate to reported nodal status, and that variation in expression patterns as we move across tumors is high; these two facts together undoubtedly obscure the subtler patterns that relate to reported nodal status.

Figure 7 provides the display of predictions for the 13 validation cases having just one or two reported positive nodes. Here the predictions vary widely. Three cases have probabilities likely below 0.1, indicating that their expression profiles are more consistent with the reported negatives in the training set than they are with the reported positives. A few cases are clustered around 0.5 with wide intervals, indicating the high uncertainty about true status; their expression profiles are as similar to those of the reported negatives in the training set as they are to the positives. This reflects a high degree of inherent heterogeneity of profiles, and may also be an indication of subclustering; we are placing tumors into one of two classes, whereas some may simply look different to representatives of each class. Finally, five cases have high probabilities of true positive status, consistent with their reported status. This analysis has, in effect, provided a refined clustering of these validation cases into these three groups based on the formal, out-of-sample predictive analysis.

Cross validation and honest predictions for lymph node status

The cross-validation probabilities from the binary regression model analysis, together with estimated uncertainties, are shown in Figure 8A. As in the ER study, this predictive analysis uses the overall screened subset of 100 genes most correlated with nodal status in the training sample, and so is mainly of interest to demonstrate the very clear discriminatory ability of this subset of genes, and hence underscore the potential for underlying biological interpretation. This analysis again provides a good classification based on lymph node status, quite comparable to that for the ER discrimination.

Figure 8B illustrates the more appropriate cross-validation analysis that adopts a screen to select potentially different genes for each hold-out case, repeating the gene selection process and model fitting with each tumor removed from the sample for analysis and then predicted based on the rest. This should be compared to the "gold standard" of Figure xx. The screened subsets of 100 most discriminatory genes vary more widely than that seen in the ER analysis as we move across tumors, again reflecting higher levels of natural variation in gene expression patterns with respect to nodal status. The resulting graph shows high uncertainties in some predictions, in terms of wide intervals around the estimated probabilities. All of the reportedly positive cases have estimated probabilities appropriately above 0.5, though some are close to that boundary with 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, these could be cases that have not yet metastasized but are poised to do so.

This analysis of nodal status provides a very clear illustration of the importance of the use of out-of-sample prediction in gauging the validity of the classification. Whereas gene expression profiles can provide a very clean separation based on nodal status, as illustrated in Figure 8A, the validation measurements clearly reveal the uncertainty in these predictions, 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, it remains true that the analysis does identify patterns that have predictive capability. In this example of 38 tumors, 10 were accurately predicted as node negative and

7 as node positive. Of the remaining 17 tumors, 10 are predicted appropriately but with considerable uncertainty. Clearly, it is the analysis of those tumors in the uncertain region that must be the focus of further studies.

Discussion

Recent studies of breast cancer ¹⁰, leukemia ⁷, and lymphoma ⁸ have shown that the analysis of patterns of gene expression has the capacity to classify tumors as well as to define tumor sub-types. The analyses presented here further demonstrate that clinically-relevant phenotypes can be determined for primary breast tumor samples through the analysis of gene expression. We go further in developing analyses that possess predictive power in a formal probabilistic sense, allowing new clinical samples of unknown status to be evaluated in a formal probabilistic framework. Such predictive capability brings gene expression analysis a 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 to demonstrate the validity of their classification for leukemia and breast cancer studies ^{7,14}. However, we have point out through the examples of ER and nodal status, the fundamental need to select smaller subsets of genes with value in predictive discrimination must be recognized in such cross-validation studies. The approach that uses an initial, overall screen to select one single subset of genes based on all the training data will lead to illusions of greater predictive accuracy and validity than will be experienced in practice.

Our study develops gene expression analysis in a context exhibiting a considerable degree of biological heterogeneity, much greater than that encountered in the study of AML/ALL leukemia samples, for example, and that involve and address quite subtle aspects of tumor phenotype. Importantly, our methods 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 clinical data from new patients. Because prognostic and predictive factors are couched in probabilistic language, clinicians can make judgments based on unbiased assessments of the uncertainties in a classification.

Although it is relatively straightforward to assay ER status by immunohistochemistry, and we have viewed the ER determination as a test case, it nevertheless is true that the discrimination using gene expression profiles has potential practical clinical value. For instance, the assay of ER status by immunohistochemistry is not perfect and can produce erroneous results. In addition, such assays would not score alterations that disable the ER pathway. Thus, if the clinically significant determination is the status of the pathway, not just 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. Moreover, 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, such as HNF3 α and androgen receptor, points to the potential power of the analysis in identifying functionally significant relationships.

The presence of metastatic breast cancer in axillary lymph nodes is the most significant factor in overall survival ¹³. Although the determination of lymph node status is relatively routine, it is true that selectivity in the process of identifying nodes for examination induces biases that suggest some reported negatives may indeed be truly positive ^{15,16}. Perhaps of more significance is the patient with truly negative lymph nodes but with a primary tumor that is poised to metastasize. Although more data is needed to determine the precision of the predictive capability for lymph node status, it is possible that

a gene expression profile could predict metastatic potential even in the clear absence of reportedly positive lymph nodes. That is, the analysis would predict the imminent potential of the behavior of the tumor and thus allow more aggressive and appropriate therapy to be instituted. Our analysis explicitly identifies cases that are reportedly node negative but whose gene expression patterns share much in common with node positives, and which therefore suggest that this objective may indeed be achievable. In parallel, we encounter the reciprocal situation – cases with just one or two nodes reported positive but whose expression patterns share higher similarity to those of true node negatives than true node positives in the sample. These cases are almost equally important and potentially informative about expression patterns that characterize the early or initial stages of metastasis.

Finally, the derivation of discriminating factors, that are based on the expression of genes in the samples, not only provides the potential for diagnostic discrimination but also provides information about the nature of the gene expression patterns that define the differences in the samples. As seen in the ER discrimination, this includes genes known to lie in the ER pathway as well as genes that may function together with ER. As such, one could view these discriminating patterns to reflect not just co-regulated genes but also genes functioning coordinately in a pathway. Although the nature of the genes providing discrimination in lymph node status is more difficult to interpret, there are several genes in the top 100 that are known to influence metastasis. For instance, previous work has implicated PAF in lymph node metastasis^{17,18}, Maspin as a suppressor of tumor growth and metastasis¹⁹⁻²², and S100A1 in modulating metastatic potential²³. One would hope that as the process of gene annotation goes forward, particularly with respect to function, it will be possible to not only use these analyses for discrimination but also for developing an

understanding of the pathways that define the difference between tumors in a clinically meaningful way.

Acknowledgements

This work was supported by the Duke SPORE in Breast Cancer (CA 68438), the Early Detection Research Network (CA 84955), and pilot project funds from the Duke Comprehensive Cancer Center (CA xxx). JRN is an Investigator of the Howard Hughes Medical Institute. RS and HZ were partially supported as postdoctoral fellows of the National Institute of Statistical Sciences.

Experimental Procedures

Breast tumor samples. Primary breast tumors from the Duke Breast Cancer SPORE frozen tissue bank were selected for this study on the basis of several criteria. Tumors were either positive for both the estrogen and progesterone receptors or negative for both receptors. Each tumor was diagnosed as invasive ductal carcinoma and was between 1.5 and 5 cm in maximal dimension. Each case had a diagnostic axillary lymph node dissection performed. Each potential tumor was examined by H&E staining and only those that were >60% tumor (on a per cell basis), with few infiltrating lymphocytes or necrotic tissue, were carried on for RNA extraction. The final collection of tumors consisted of 13 ER+ LN+ tumors, 12 ER– LN+ tumors, 12 ER+ LN– tumors, and 12 ER– LN– tumors (see Table 1).

RNA preparation. Approximately 30 mg of frozen breast tumor tissue was added to a chilled BioPulverizer H tube (Bio101). Lysis buffer from the Qiagen RNeasy Mini kit was added and the tissue was homogenized for 20 seconds in a Mini–Beadbeater (Biospec Products). Tubes were spun briefly to pellet the garnet mixture and reduce foam. The lysate was transferred to a new 1.5 ml tube using a syringe and 21 ga. needle, followed by passage through the needle 10 times to shear genomic DNA. Total RNA was extracted using the Qiagen RNeasy Mini kit. Two extractions were performed for each tumor and the total RNA was pooled at the end of the RNeasy protocol, followed by a precipitation step to reduce volume. Quality of the RNA was checked by visualization of the 28S:18S ribosomal RNA ratio on a 1% agarose gel. Concentration of total RNA was determined by spectrophotometry.

Affymetrix GeneChip analysis. The targets for Affymetrix DNA microarray analysis were prepared according to the manufacturers instructions. All assays used the human HuGeneFL GeneChip microarray. Arrays were hybridized with the targets at 45° C for 16 hr and then washed and stained using the GeneChip Fluidics station according to the manufacturers instructions. DNA chips were scanned with the GeneChip scanner and signals obtained by the scanning were processed by GeneChip Expression Analysis algorithm (version 3.2) (Affymetrix).

Statistical methods. Our analysis uses standard binary regression models combined with singular factor decompositions (SVDs) and with stochastic regularization using Bayesian analysis. In the ER status study as an example, probit regression models have $p(x)$ for the probability of ER+ versus ER- status conditional on the vector x of measured expression levels for all genes on the array. A standard probit regression sets $p(x)=F(b'x)$ where F is the standard normal distribution function of the linear function of the elements of x , namely $b'x$ where b is a regression coefficient column vector and b' is its row vector transpose. A singular factor decomposition of the matrix of expression vectors maps this to the equivalent form $F(g'f)$ where f is the vector of supergene factors for that array, and g is a regression coefficient vector. Standard Bayesian analysis methods^{24,25} provide for model fitting and prediction, and we use standard prior distributions for g that induce an unbiased analysis which providing stochastic regularization to overcome the inherent instabilities in the resulting likelihood function in this model with n binary observations and n predictors (independent variables) in f . Some elements of g are set to zero to remove factors with low levels of variation, representing only irrelevant noise in the expression data; analysis reported selects five non-zero elements. For these, prior distributions on elements of g are

independent Student T distributions with 2 degrees of freedom, centered at zero, so represent very vague priors and an unbiased initial position with respect to what they predict about classification probabilities. Most importantly, our theory shows that the Bayesian SVD regression framework allows direct inversion to infer the parameters b from g . This is critical as it provides inferences about which genes are important in defining $p(x)$, and how subsets of genes interact. Model fitting uses standard iterative Markov chain Monte Carlo (MCMC) simulation methods of Bayesian analysis^{41,42} to impute sets of simulated parameter values whose distributions are summarized to produce point and interval estimates of model parameters g, b as well as of probabilities of ER status in both the fitted model and for future, validation samples. Multiple repeat analyses varying aspects of the model prior distributions, including the number of non-zero elements of g selected, and control parameters for the MCMC analysis have been carried out to verify the stability of the reported results.

Our analysis strategy reduces noise contributed by irrelevant or unexpressed genes by an initial screening process to select a subset of genes maximally correlated with outcome. This screening strategy aims to reduce noise contributed by irrelevant or unexpressed genes by an initial selection process, and the choice of the number 100 was determined by repeat experimentation. This screen computes sample correlation coefficients between genes and binary outcomes, and selects those genes giving the 100 largest absolute values of this correlation. The binary regression model is then fitted to this set of 100 selected genes, using the resulting SVD factors based on these 100 genes. In out-of-sample prediction, either in the validation study of new cases or in the one-at-a-time analysis, expression data for all tumors, including those to be predicted, is included in the expression matrix for SVD analysis, even though the binary outcomes are unknown for

these cases. This is important in that the SVD analysis is then informed by these cases as well as those in the training set, so better identifying the underlying factor structure.

Predictive computations are trivial, since they are based on including the unknown binary outcomes as latent/hidden variables to be iteratively simulated in the Bayesian MCMC analysis, another standard technique^{8,41,42}. We stress the importance of such analysis in providing an honest assessment of uncertainty about predictions, as well a rational, theoretically well-founded estimate of the probability of ER status for any new case.

We further stress that the selection of gene subsets for prediction is repeated for each one-at-a-time validation analysis. This is fundamentally important from a practical viewpoint, since these gene subsets can vary depending on the specific training sample and to do otherwise would not adequately represent the real-life circumstances that will arise in future use of such analyses. This contrasts with other studies that involve an initial overall screen to select a gene subset based on all the samples, and then use only that selected subset in one-at-a-time predictions, and that are potentially misleading as a result. Our empirical results validate this view, clearly indicating the extent of additional uncertainties in one-at-a-time predictions compared to the fitted model predictions, and therefore elucidating the extent of inherent heterogeneity and some of its implications in predictive analysis that will ultimately inform therapeutic decisions.

Figure Legends

Figure 1. Factor analysis for ER+/ER– comparison

- A.** Pairwise factor analysis. Breast tumors depicted in a scatter plot on two dominant factors underlying 100 genes selected in pure discrimination of the training cases. Each tumor is indicated by a simple index number (see Table 1) and is color coded with red indicating ER+ cases and blue indicating ER– cases. Only the tumors in the training set are plotted.
- 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. Color coding is as described in panel A.

Figure 2. Expression levels of top 100 genes providing pure discrimination of ER status.

Figure 3. Predictive probabilities for ER status of each tumor in the validation sample.

The analysis was based on the selected subset of 100 genes in the full training sample analysis. The format color coding follows the same scheme described in Figure 1.

Figure 4. Out-of-sample cross validation predictions of ER status.

- C.** One-at-a-time cross-validation predictions of 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 estimated classification probabilities with corresponding 90% probability intervals marked as dashed lines to indicate uncertainty about these estimated values. The analysis and predictions for each tumor are based on the screened subset of 100 most discriminatory genes to parallel current practice in expression studies by other groups.
- D.** One-at-a-time cross-validation predictions of classification probabilities for training cases in the ER study, in a format similar to that of panel A. The difference here is that the predictive analysis is honest in that 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 exhibits the resulting honest uncertainties about the extent of true predictive accuracy in a practical setting, reflecting inherent variability due to heterogeneity of expression profiles.

Figure 5. Factor analysis for nodal status comparison

- A.** Pairwise factor analysis. Breast tumors depicted in a scatter plot on two dominant factors underlying 100 genes selected in pure discrimination of the training cases. Each tumor is indicated by a simple index number (see Table 1) and is color coded with red indicating node positive cases with at least 3 identified positive nodes and blue indicating lymph node negative cases. Only the tumors in the training set are plotted.
- 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. Color coding is as described in panel A.

Figure 6. Expression levels of top 100 genes providing pure discrimination of nodal status.

Figure 7. Predictive probabilities for nodal status of each tumor in the validation sample.

The analysis was based on the selected subset of 100 genes in the full training sample analysis. The format color coding follows the same scheme described in Figure 5.

Figure 8. Out-of-sample cross validation predictions of nodal status.

- A.** One-at-a-time cross-validation predictions of 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 estimated classification probabilities with corresponding 90% probability intervals marked as dashed lines to indicate uncertainty about these estimated values. The analysis and predictions for each tumor are based on the screened subset of 100 most discriminatory genes to parallel current practice in expression studies by other groups.
- B.** One-at-a-time cross-validation predictions of classification probabilities for training cases in the nodal study, in a format similar to that of panel A. The difference here is that the predictive analysis is honest in that 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. The figure exhibits the resulting honest uncertainties about the extent of true predictive accuracy in a practical setting, reflecting inherent variability due to heterogeneity of expression profiles.

Table 1. Breast tumor samples used for gene expression analysis

Tumor No.	Hist. Grade	Nuclear Grade	IPASTG	ER Status	LN Status
1	3	3	4	Pos	Pos
2	2	2	2B	Pos	Pos
3	3	2	2B	Pos	Pos
4	2	2	2B	Pos	Pos
5	3	2	2B	Pos	Pos
6	3	3	2B	Pos	Pos
7	1	1	2A	Pos	Neg
8	2	2	2A	Pos	Neg
9	2	2	2A	Pos	Neg
10	1	1	1	Pos	Neg
11	2	2	2A	Pos	Neg
12	2	2	2A	Pos	Neg
13	3	3	2B	Pos	Pos
14	3	3	3B	Neg ¹	Pos
15	2	2	2B	Pos	Pos
16	3	3	2B	Neg	Pos
17	3	3	4	Neg	Pos
18	3	3	2A	Neg	Pos
19	3	3	4	Neg	Pos
20	2	2	2A	Neg	Neg
21	3	3	2A	Neg	Neg
22	3	2	2A	Neg	Neg
23	3	3	2B	Neg	Neg
24	3	3	2A	Neg	Neg
25	3	3	2A	Neg	Neg
26	3	3	3B	Neg	Pos
27	3	3	2A	Neg	Neg
28	3	3	4	Pos	Pos
29	3	2	2A	Pos	Pos
30	3	2	2B	Pos	Pos
31	3	3	2B	Neg ¹	Pos
32	2	2	3B	Pos	Neg
33	3	3	2A	Neg ¹	Neg
34	3	2	3B	Pos	Neg
35	3	2	2A	Pos	Neg
36	3	3	2B	Neg	Pos
37	2	2	2A	Neg	Pos
38	3	3	2B	Neg	Pos
39	3	3	1	Neg	Neg
40	3	3	2A	Neg	Neg
41	3	2	2A	Neg	Neg
42	3	3	2A	Neg	Neg
43	2	2	2A	Neg	Neg
44	3	3	2A	Neg	Pos
45	3	3	2B	Pos ²	Pos
46	3	3	2B	Pos ²	Pos
47	2	2	2A	Pos	Neg
48	2	2	2A	Pos	Neg
49	3	3	2B	Neg	Pos

1 Initially scored as ER+ by immunohistochemistry but confirmed as ER- by immunoblot assay

2 Initially scored as ER- by immunohistochemistry but then ER+ by immunoblot assay

Table 2. Genes that contribute to discrimination of ER status

ER Rank	ER Weight	Acc. #	Unigene Cluster	Symbol	Estrogen Relation	Ref
1	0.08	x52003	Trefoil Factor 1 (pS2)	TFF1	Estrogen induced	26,27
2	0.079	x03635	Estrogen Receptor 1	ESR1	ER	
3	0.067	m29874	Cytochrome P450, Subfamily IIB	CYP2B		
4	0.064	l08044	Trefoil Factor 3	TFF3	Estrogen induced	28
5	0.061	s37730	(Insulin-Like Growth Factor)	IGFBP2	Estrogen induced	29,30
6	0.057	u79293	Human Clone 23948 mRNA Sequence	N/A		
7	0.056	j03778	Microtubule-Assoc'd Protein Tau	MAPT	Estrogen induced	31
8	0.055	x07732	Hepsin	HPN		
9	0.048	x58072	GATA-binding protein 3	GATA3	Co-expressed with ER	32-34
10	0.047	u22376	v-myb Avian Myeloblastosis Viral Oncogene Homolog	MYB	Estrogen induced	35,36
11	-0.043	u04313	Serine Proteinase Inhibitor, Clade B, Member 5 (Maspin)	SERPINB5	Induced by tamoxifen; inverse with ER	37, 38 ₁
12	0.041	x17059	N-Acetyltransferase 1	NAT1		
13	-0.041	m26311	S100 Calcium Binding Protein A9	S100A9		
14	-0.041	u27185	Retinoic Acid Receptor Responder 1	RARRES1		
15	-0.039	u84487	Small Inducible Cytokine Subfamily D, Member 1	SCYD1		
16	0.039	u39840	Hepatocyte Nuclear Factor 3 Alpha	HNF3A	Synergistic with ER	39
17	0.038	u32907	37 kDa Leucine-Rich Repeat (LRR) Protein	P37NB		
18	0.038	m35851	(Androgen Receptor)	AR	Physical interaction with ER	40
19	-0.038	x87212	Cathepsin C	CTSC		
20	0.037	u96922	Inositol Polyphosphate-4-Phosphatase, Type II, 105 kD	INPP4B		
21	0.036	af000234	Purinergic Receptor P2X, Ligand-Gated Ion Channel, 4	P2RX4	Estrogen biosynthesis	41,42
22	-0.036	d50915	KIAA0125 Gene Product	KIAA0125		
23	0.036	l07615	(Neuropeptide Y Receptor Y1)	NPY1R		
24	0.035	u68385	Meis (Mouse) Homolog 3	MEIS3		
25	0.035	u41060	LIV-1 Protein	LIV-1	Estrogen induced	43
26	0.034	hg3548-hf3749	(CCAAT Displacement Protein)	CUTL1		
27	0.032	d38437	Postmeiotic Segregation Increased 2-Like 3	PMS2L3		

28	-0.031	x04470	Secretory Leukocyte Protease Inhibitor		SLPI	
29	0.029	m81057	Carboxypeptidase B1		CPB1	
30	0.027	u95740	KIAA0430 Gene Product		KIAA0430	
31	-0.027	m24485	Glutathione S-Transferase Pi		GSTP1	Inverse relation with ER 44
32	0.025	x55037	GATA-binding protein 3		GATA3	Estrogen induced 32-34
33	0.023	m31627	X-Box Binding Protein 1		XBP1	
34	-0.023	x13794	Lactate Dehydrogenase B		LDHB	
35	0.022	z29083	5T4 Oncofetal Trophoblast Glycoprotein		5T4	
36	0.022	u21931	(Fructose-1,6-Biphosphatase)		FBP1	
37	0.021	m23263	Androgen Receptor		AR	Physical interaction with ER 40
38	0.021	u09770	Cysteine-Rich Protein 1		CRIP1	
39	0.021	x13238	Cytochrome C Oxidase Subunit Vic		COX6C	
40	-0.02	u03057	Singed (Drosophila)-Like		SNL	
41	-0.019	u05340	CDC20		CDC20	
42	-0.018	x01630	Argininosuccinate Synthetase		ASS	
43	-0.018	s45630	Crystallin, Alpha B		CRYAB	
44	-0.017	hg3494- ht3688	(Nuclear Factor NF-IL6)			
45	0.016	x83425	Lutheran Blood Group		LU	
46	0.016	m14745	B-Cell CLL/Lymphoma 2		BCL2	
47	-0.015	x87241	FAT Tumor Suppressor		FAT	
48	0.015	m99701	Transcription Elongation Factor A (SII)-Like 1		TCEAL1	
49	0.015	hg3125- ht3301	(Estrogen Receptor)		ESR1	ER
50	0.015	x12876	Keratin 18		KRT18	
51	0.015	u96113	(Nedd-4-Like Ubiquitin Protein Ligase)		WWP1	
52	0.015	m27891	(Cystatin C)		CST3	
53	-0.014	d63880	Chromosome Condensation-Related SMC-Associated Protein 1		KIAA0159	
54	0.014	x68733	Serine Proteinase Inhibitor, Clade A, Member 3		SERPINA3	
55	0.014	z23090	Heat Shock 27 kD Protein 1		HSPB1	Estrogen induced 45,46
56	0.014	x16665	Homeo Box B2		HOXB2	
57	0.014	m62403	Insulin-Like Growth Factor Binding Protein 4		IGFBP4	
58	-0.014	u85193	Nuclear Factor I/B		NFIB	

59	0.013	z48633	H. sapiens mRNA for RetroTransposon	N/A	
60	0.013	d50840	UDP-Glucose Ceramide Glucosyltransferase	UGCG	
61	0.013	l15702	B-Factor	BF	
62	0.013	d78134	Cold Inducible RNA-Binding Protein	CIRBP	
63	0.013	u72649	BTG Family, Member 2	BTG2	47 Regulate ER-mediated activation
64	0.012	x75861	Testis Enhanced Gene Transcript	TEGT	
65	0.012	u52100	Epithelial Membrane Protein 2	EMP2	
66	0.011	x53002	Integrin, Beta 5	ITGB5	
67	0.011	u29656	Non-Metastatic Cells 3, Protein Expressed In	NME3	
68	0.011	x59834	Glutamate-Ammonia Ligase	GLUL	
69	0.011	l40401	Peroxisomal Long-Chain Acyl-CoA Thioesterase, Putative Protein	ZAP128	
70	-0.011	M32313	Steroid-5-Alpha-Reductase, Alpha Polypeptide 1	SRD5A1	
71	0.011	X57351	Interferon Induced Transmembrane Protein 2	IFITM2	
72	-0.011	D38550	Human mRNA for KIAA0075 Gene		
73	-0.01	D90209	Activating Transcription Factor 4	ATF4	
74	0.01	U68142	RAB2, Member of RAS Oncogene Family-Like	RAB2L	
75	0.01	S69272	Serine Proteinase Inhibitor, Clade B, Member 6	SERPINB6	
76	-0.0099	M69066	Moesin	MSN	48 Inverse relation with ER
77	0.0091	hg1515- ht1515	(Transcription Factor BTF3B)	BTF3	
78	0.0091	U67963	Lysophospholipase-Like	HU-K5	
79	-0.0091	U65785	Oxygen Regulated Protein (150 kD)	ORP150	
80	0.0089	U09196	Polymerase (DNA-Directed), Delta 4	POLD4	
81	0.0088	U11791	Cyclin H	CCNH	
82	0.0086	M29877	Fucosidase, Alpha-L-1, Tissue	FUCA1	
83	-0.0083	U96131	Thyroid Hormone Receptor Interactor 13	TRIP13	
84	0.0082	U14603	Protein Tyrosine Phosphatase Type IVA, Member 2	PTP4A2	
85	-0.0079	M14328	Endolase 1	ENO1	
86	0.0078	X74929	Keratin 8	KRT8	
87	0.0078	X71973	Glutathione Peroxidase 4	GPX4	
88	0.0077	M74715	Iduronidase, Alpha-L	IDUA	
89	0.0076	U14394	Tissue Inhibitor of Metalloproteinase 3	TIMP3	

90	0.0075	U30827	Splicing Factor, Arginine/Serine-Rich 5	SFRS5
91	0.0074	U79262	Deoxyhypusine Synthase	DHPS
92	0.0071	U94831	Transmembrane 9 Superfamily Member 1	TM9SF1
93	-0.007	I20298	Core-Binding Factor, Beta Subunit	CBFB
94	0.0069	U72066	Retinoblastoma-Binding Protein 8	RBBP8
95	-0.0068	I16862	G Protein-Coupled Receptor Kinase 6	GPRK6
96	-0.0066	M96982	U2(RNU2) Small Nuclear RNA Auxiliary Factor 1	U2AF1
97	-0.0066	U09564	SFRS Protein Kinase 1	SRPK1
98	0.0061	hg2238-h2321	(Nuclear Mitotic Apparatus Protein 1)	NUMA1
99	-0.0055	D86978	KIAA0225 Protein	KIAA0225
100	0.0039	X91648	H. sapiens mRNA for Pur Alpha Extended 3' Untranslated Region	N/A

Legend to Table 2. 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). In those instances where a gene appears more than once in the list, the DNA microarray contained more than one representation of that gene.

Table 3. Genes that contribute to discrimination of lymph node status

Ran k	Weig ht	Acc#	Unigene Cluster	Symbol
1	-0.072	x58079	S100 Calcium-Binding Protein A1	S100A1
2	-0.061	x84707	Melanoma Inhibitory Activity	MIA
3	-0.055	u39817	Bloom Syndrome	BLM
4	-0.055	m99435	Transducin-Like Enhancer of Split 1, Homolog of Drosophila E(sp1)	TLE1
5	0.054	u32674	G Protein-Coupled Receptor 9	GPR9
6	-0.052	s69115	Natural Killer Cell Group 7 Sequence (NKG7)	NKG7
7	-0.051	u27185	Retinoic Acid Receptor Responder (Tazarotene Induced) 1 (RARRES1)	RARRES1
8	-0.051	u90065	Potassium Channel, Subfamily K, Member 1	KCNK1
9	-0.049	l22454	Nuclear Respiratory Factor 1	NRF1
10	-0.048	s74445	Cellular Retinoic Acid-Binding Protein 1 (CRABP1)	CRABP1
11	-0.048	d85425	Nuclear Transcription Factor Y, Gamma	NFYC
12	-0.047	l06419	Procollagen-Lysine, 2-Oxoglutarate 5-Dioxygenase (Lysine Hydroxylase, Ehlers-Danlos Syndrome Type VI)	PLOD
13	-0.046	u04313	Serine (or Cysteine) Proteinase Inhibitor, Clade B (Ovalbumin), Member 5	SERPINB5, MASPIN
14	-0.043	d87071	KIAA0233 Gene Product	KIAA0233
15	0.042	x02158	Erythropoietin	EPO
16	-0.042	u02609	Transducin (Beta)-Like 3	TBL3
17	-0.042	u90908	Hypothetical Protein From Clones 23549 and 23762	LOC58504
18	-0.042	m84820	Retinoid X Receptor, Beta	RXRβ
19	-0.041	d25217	KIAA0027 protein	
20	-0.041	ab000897	n/a	
21	0.041	s77361	n/a	
22	-0.04	j02960	Adrenergic, Beta-2-, Receptor, Surface	ADRB2
23	-0.04	d50923	KIAA0133 Gene Product	KIAA0133
24	-0.039	m13955	Keratin 7	KRT7
25	-0.039	s76942	Dopamine Receptor D4	DRD4
26	-0.038	d28235	Prostaglandin-Endoperoxide Synthase 2 (Cyclooxygenase 2)	PTGS2
27	-0.038	u41344	n/a	
28	-0.038	m60047	Heparin-Binding Growth Factor Binding Protein	HBP17
29	0.038	d63390	Platelet-Activating Factor Acetylhydrolase, Isoform IB, Beta Subunit	PAFAH1B2
30	-0.038	m95740	Iduronidase, Alpha-L-	IDUA
31	-0.038	l07597	Ribosomal Protein S6 Kinase, 90kD, Polypeptide 1	RPS6KA1
32	-0.038	d31889	Proteasome 26S Subunit, Non-ATPase, 5	PSMD5

33	0.037	hg3998-- ht4268	Splicing Factor 3a, Subunit 1, 120 kD	SF3A1
34	-0.037	x85237		
35	0.037	hg3548-- ht3749		
36	-0.037	m97347	Glucosaminyl (N-Acetyl) Transferase 1, Core 2 (Beta-1,6-N-Acetylglucosaminyltransferase)	GCNT1
37	0.036	u06454	Protein Kinase, AMP-Activated, Alpha 2 Catalytic Subunit	PRKAA2
38	-0.036	d13645	KIAA0020 gene product	KIAA0020
39	-0.036	l02547	Cleavage Stimulation Factor, 3' pre=RNA, Subunit 1, 50 kD	CSTF1
40	0.035	m15517	n/a	
41	-0.035	u84540	n/a	
42	0.035	m11437	n/a	
43	0.034	u37673	Adaptor-Related Protein Complex 3, Beta 2 Subunit	AP3B2
44	0.034	m77144	Hydroxy-Delta-5-Steroid Dehydrogenase, 3 Beta- and Steroid Delta-Isomerase 2	HSD3B2
45	0.034	u57093	RAB27B, Member RAS Oncogene Family	RAB27B
46	0.033	x79781	RAB35, Member RAS Oncogene Family	RAB35
47	-0.033	d25215	Hect Domain and RLD 3	HERC3
48	0.033	l11695	Transforming Growth Factor, Beta Receptor 1	TGFBFR1
49	-0.033	af000560	Homo sapiens TTF-I interacting peptide 20 mRNA, partial cds	
50	-0.032	u29589	Cholinergic Receptor, Muscarinic 3	CHRM3
51	0.031	y09980	Homeobox D3	HOXD3
52	-0.031	x14690	Pre-Alpha (Globulin) Inhibitor, H3 Polypeptide	ITI13
53	-0.029	z14978	ARP1 (Actin-Related Protein 1, Yeast) Homolog A (Centractin Alpha)	ACTR1A
54	-0.028	x07618	Cytochrome p450, Subfamily IID, Polypeptide 7a (Pseudogene)	CYP2D7AP
55	-0.028	x66503	Adenylsuccinate Synthase	ADSS
56	0.028	x91196	n/a	
57	0.027	hg855-- ht855		
58	-0.025	x05276	Tropomyosin 4	TPM4
59	-0.024	u91985	DNA Fragmentation Factor, 45 kD, Alpha Polypeptide	DFFA
60	-0.024	z25535	Nucleoporin 153 kD	NUP153
61	-0.024	m34539	FK506-Binding Protein 1A	FKBP1A
62	-0.024	hg2999-- ht4756		
63	-0.024	hg3638-- ht3849		
64	0.023	u04840	Neuro-Oncological Ventral Antigen 1	NOVA1
65	-0.022	y08319	Kinesin Heavy Chain Member 2	KIF2
66	-0.021	u50534	Putative Gene Product	13CDNA73
67	-0.02	u72761	Karyopherin (Importin) Beta 3	KPNB3

68	0.02	m99701	Transcription Elongation Factor A (SII)-Like 1	TCEAL1
69	0.018	x07834	Superoxide Dismutase 2, Mitochondrial	SOD2
70	-0.017	m86737	Structure Specific Recognition Protein 1	SSRP1
71	0.016	u38276	Sema Domain, Ig Domain, Short Basic Domain, Secreted, (Semaphorin) 3F	SEMA3F
72	0.015	x66922	Inositol(Myo)-1(or 4)-Monophosphatase 1	IMPA1
73	-0.014	u49835	Chitinase 3-Like 2	CHI3L2
74	-0.014	l20010	Host Cell Factor C1	HCFC1
75	0.014	d83018	Nef-Like 2	NELL2
76	-0.014	l40379	Thyroid Hormone Receptor Interactor 10	TRIP10
77	-0.013	d13988	GDP Dissociation Inhibitor 2	GDI2
78	-0.013	u58658	Human Unknown Protein mRNA within p53 Intron 1	
79	0.012	s82198	Chymotrypsin C (Caldecrin)	CTRC
80	0.012	x92518	High-Mobility Group Protein Isoform I-C	HMGIC
81	0.012	m33684	n/a	
82	0.012	s69272	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 6	SERPIN B6
83	-0.011	hg1612- ht1612		
84	0.011	l18877	Melanoma Antigen, Family A, 12	MAGEA12
85	-0.011	u28686	RNA Binding Motif Protein 3	RBM3
86	-0.009	y00264	Amyloid Beta Precursor Protein	APP
87	-0.009	m76378	n/a	
88	0.009	x56681	JunD Proto-Oncogene	JUND
89	0.0087	u49869	n/a	
90	0.0086	u64444	Ubiquitin Fusion Degradation 1-Like	UFD1L
91	-0.008	z50781	Delta Sleep Inducing Peptide, Immunoreactor	DSIP1
92	0.0083	x51521	Villin 2 (Ezrin)	VIL2
93	-0.007	z22536	Activin A Receptor, Type IB	ACVR1B
94	0.0063	u70671	Ataxin 2 Related Protein	A2LP
95	0.0062	x15341	Cytochrome C Oxidase Subunit Via Polypeptide 1	COX6A1
96	-0.006	u02493	Non-POU-Domain-Containing, Octamer-Binding	NONO
97	-0.005	d31884	KIAA0063 Gene Product	KIAA0063
98	-0.002	l34075	FK506 Binding Protein 12-Rapamycin Associated Protein 1	FRAP1
99	-0.001	x15525	Acid Phosphatase 2, Lysosomal	ACP2

100 -0.000
9

m69013 Guanine Nucleotide Binding Protein, Alpha 11

GNA11

Legend to Table 3. 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 positive lymph node status (and thus a positive correlation with a node–negative status).

Reference List

1. Ciocca,D.R. *et al.* Correlation of HER-2/neu amplification with expression and with other prognostic factors in 1103 breast cancers. *J. Natl. Cancer Inst.* **84**, 1279–1282 (1992).
2. Tandon,A.K., Clark,G.M., Chamness,G.C., Ullrich,A. & McGuire,W.L. HER-2/neu oncogene protein and prognosis in breast cancer. *J. Clin. Oncol.* **7**, 1120–1128 (1989).
3. Muss,H.B. *et al.* c-erbB-2 expression and response to adjuvant therapy in women with node-positive early breast cancer. *N. Engl. J. Med.* **331** , 211 (1994).
4. Alon,U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96**, 6745–6750 (1999).
5. Bittner,M. *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540 (2000).
6. DeRisi,J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* **14**, 457–460 (1996).
7. Golub,T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
8. Alizadeh,A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).

9. Khan,J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
10. Perou,C.M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* **9212**, 9217 (1999).
11. Alter,O., Brown,P.O. & Botstein,D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* **97**, 10101–10106 (2000).
12. West,M., Nevins,J.R., Marks,J.R., Spang,R. & Zuzan,H. Bayesian regression analysis in the "large p, small n" paradigm with application in DNA microarray studies. *Manuscript submitted* (2000).
13. Shek,L.L. & Godolphin,W. 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 (1988).
14. Hedenfalk,I. *et al.* Gene expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344**, 539–548 (2001).
15. Kjaergaard,J., Blichert-Toft,M., Andersen,J.A., Rank,F. & Pedersen,B.V. Probability of false negative nodal staging in conjunction with partial axillary dissection in breast cancer. *Br. J. Surg.* **72**, 365–367 (1985).
16. Hill,A.D. *et al.* Lessons learned from 500 cases of lymphatic mapping for breast cancer. *Ann. Surg.* **231**, 148–149 (1999).

17. Pitton,C. *et al.* Presence of PAF–acether in human breast carcinoma: relation to axillary lymph node metastasis. *J. Natl Cancer Inst* **81**, 1298–1302 (1989).
 18. Im,S.Y. *et al.* Augmentation of tumor metastasis by platelet–activating factor. *Cancer Res* **56**, 2662–2665 (1996).
 19. Zhang,M., Shi,Y., Magit,D., Furth,P.A. & Sager,R. Reduced mammary tumor progression in WAP–TAg/WAP–maspin bitransgenic mice. *Oncogene* **19**, 6053–6058 (2000).
 20. Xiao,G. *et al.* Suppression of breast cancer growth and metastasis by a serpin myoepithelium–derived serine proteinase inhibitor expressed in the mammary myoepithelial cells. *Proc. Nat’l. Acad. Sci. U. S. A.* **96**, 3700–3705 (1999).
 21. Sheng,S. *et al.* Maspin acts at the cell membrane to inhibit invasion and motility of mammary and prostatic cancer cells. *Proc Natl Acad Sci USA* **93**, 11669–11674 (1996).
 22. Zou,Z. *et al.* Maspin, a serpin with tumor–suppressing activity in human mammary epithelial cells. *Science* **263**, 526–529 (1994).
 23. Wang,G., Rudland,P.S., White,M.R. & Barraclough,R. Interaction in vivo and in vitro of the metastasis–inducing S100 protein, S100A4 (p9Ka) with S100A1. *J. Biol. Chem.* **275**, 11141–11146 (2000).
 24. Carlin,J.B. Bayesian Data Analysis. 1996. Chapman & Hall.
- Ref Type: Serial (Book,Monograph)

25. Johnson,V.E. & Albert,J.H. Ordinal Data Modeling. 1999. Springer–Verlag.
Ref Type: Serial (Book,Monograph)
26. Jeltsch,J.M. *et al.* Structure of the human oestrogen–responsive gene pS2. *Nucl. Acids Res.* **15**, 1401–1414 (1987).
27. Berry,M., Nunez,A.M. & Chambon,P. Estrogen–respnsive element of the human pS2 gene is an imperfectly palindromic sequence. *Proc Natl Acad Sci USA* **86**, 1218–1222 (1989).
28. May,F.E. & Westley,B.R. Expression of human intestinal trefoil factor in malignant cells and its regulation by oestrogen in breast cancer cells. *J. Pathol.* **182**, 404–413 (1997).
29. Richmond,R.S., Carlson,C.S., Register,T.C., Shanker,G. & Loeser,R.F. Functional estrogen receptors in adult articular cartilage: estrogen replacement therapy increases chondrocyte synthesis of proteoglycans and insulin–like growth factor binding protein 2. *Arthritis Rheum.* **43**, 2081–2090 (2000).
30. Cardona–Gomez,G.P., Chowen,J.A. & Garcia–Segura,L.M. Estradiol and progesterone regulate the expression of insulin–like growth factor–I receptor and insulin–like growth factor binding protein–2 in the hypothalamus of adult female rats. *J. Neurobiol.* **43**, 269–281 (2000).
31. Matsuno,A. *et al.* Modulation of protein kinases and microtubule–associated proteins and changes in ultrastructure in female rat pituitary cells: effects of estrogen and bromocriptine. *J. Histochem. Cytochem.* **45**, 805–813 (1997).

32. Hoch,R.V., Thompson,D.A., Baker,R.J. & Weigel,R.J. GATA-3 is expressed in association with estrogen receptor in breast cancer. *Int. J. Cancer* **84**, 122-128 (1999).
33. Yang,G.P., Ross,D.T., Kuang,W.W., Brown,P.O. & Weigel,R.J. Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. *Nucl. Acids Res.* **27**, 1517-1523 (1999).
34. Bertucci,F. *et al.* Gene expression profiling of primary breast carcinomas using arrays of candidate genes. *Hum. Mol. Genet.* **9**, 2981-2991 (2000).
35. Jeng,M.H. *et al.* Estrogen receptor expression and function in long-term estrogen-deprived human breast cancer cells. *Endocrinology* **139**, 4164-4174 (1998).
36. Gudas,J.M., Klein,R.C., Oka,M. & Cowan,K.H. Posttranscriptional regulation of the c-myb proto-oncogene in estrogen receptor-positive breast cancer cells. *Clin. Cancer Res.* **1**, 235-243 (1995).
37. Shao,Z.M., Radziszewski,W.J. & Barsky,S.H. Tamoxifen enhances myoepithelial cell suppression of human breast carcinoma progression in vitro by two different effector mechanisms. *Cancer Lett.* **157**, 133-144 (2000).
38. Martin,K.J. *et al.* Linking gene expression patterns to therapeutic groups in breast cancer. *Cancer Res.* **60**, 2232-2238 (2000).
39. Robyr,D., Gegonne,A., Wolffe,A.P. & Wahli,W. Determinants of vitellogenin B1 promoter architecture. HNF3 and estrogen responsive transcription within chromatin. *J. Biol. Chem.* **275**, 28291-28300 (2000).

40. Panet-Raymond, V., Gottlieb, B., Beitel, L.K., Pinsky, L. & Trifiro, M.A. Interactions between androgen and estrogen receptors and the effects on their transactivational properties. *Mol. Cell Endocrinol.* **167**, 139–150 (2000).
41. Gillman, T.A. & Pennefather, J.N. Evidence for the presence of both P1 and P2 purinoceptors in the rat myometrium. *Clin. Exp. Pharmacol. Physiol.* **25**, 592–599 (1998).
42. Schmidt, M. & Loffler, G. Induction of aromatase activity in human adipose tissue stromal cells by extracellular nucleotides – evidence for P2 purinoceptors in adipose tissue. *Eur. J. Immunol.* **252**, 147–154 (1998).
43. el-Tanani, M.K. & Green, C.D. Insulin/IGF-1 modulation of the expression of two estrogen-induced genes in MCF-7 cells. *Mol. Cell Endocrinol.* **121**, 29–35 (1996).
44. Gilbert, L. *et al.* A pilot study of pi-class glutathione S-transferase expression in breast cancer: correlation with estrogen receptor expression and prognosis in node-negative breast cancer. *J. Clin. Oncol.* **11**, 49–58 (1993).
45. Fanelli, M.A., Cuello Carrion, F.D., Dekker, J., Schoemaker, J. & Ciocca, D.R. Serological detection of heat shock protein hsp27 in normal and breast cancer patients. *Cancer Epidemiol. Biomarkers Prev.* **7**, 791–795 (1998).
46. Rochefort, H. Oestrogen- and anti-oestrogen-regulated genes in human breast cancer. *Ciba Found. Symp.* **191**, 254–265 (1995).
47. Prevot, D. *et al.* Relationships of the antiproliferative proteins BTG1 and BTG2 with CAF1, the human homolog of a component of the yeast CCR4 transcriptional complex:

involvement in estrogen receptor (alpha) signaling pathway. *J. Biol. Chem.* **in press**, (2001).

48. Carmeci, C. *et al.* Moesin expression is associated with the estrogen receptor negative breast cancer phenotype. *Surgery* **124**, 211–217 (1998).