GENE EXPRESSION PREDICTORS OF BREAST CANCER OUTCOMES

(revision of:

PREDICTION OF BREAST CANCER STATES AND OUTCOMES BY
INCORPORATING GENE EXPRESSION PROFILES)

Erich Huang\textsuperscript{2}, Skye H Cheng\textsuperscript{1}, Holly Dressman\textsuperscript{2}, Jennifer Pittman\textsuperscript{5}, Mei-Hua Tsou\textsuperscript{1}, Cheng-Fang Horng\textsuperscript{1}, Andrea Bild\textsuperscript{2}, Edwin S Iversen\textsuperscript{4,5}, Ming Liao\textsuperscript{5}, Chii-Ming Chen\textsuperscript{1}, Mike West\textsuperscript{5}, Joseph R Nevins\textsuperscript{2,6} and Andrew T Huang\textsuperscript{1,3}

\textsuperscript{1}Koo Foundation Sun Yat-Sen Cancer Center, Taipei, Taiwan
\textsuperscript{2}Department of Molecular Genetics and Microbiology, Duke University Medical Center
\textsuperscript{3}Department of Medicine, Duke University Medical Center
\textsuperscript{4}Department of Biostatistics and Bioinformatics, Duke University Medical Center
\textsuperscript{5}Institute of Statistics and Decision Sciences, Duke University
\textsuperscript{6}Howard Hughes Medical Institute
SUMMARY

Background The integration of currently accepted risk factors with genomic data carries the promise of focusing the practice of medicine on the individual patient. Such integration requires interpreting the complex, multivariate patterns in gene expression data, and evaluating their capacity to improve clinical predictions. We do this here, in a study of predicting nodal metastatic states and relapse for breast cancer patients.

Methods DNA microarray data from samples of primary breast tumors were analyzed using non-linear statistical analyses to evaluate multiple patterns of interactions of groups of genes that have predictive value, at the individual patient level, with respect to lymph node metastasis and cancer recurrence.

Findings We identify aggregate patterns of gene expression (metagenes) that associate with lymph node status and recurrence, and that are capable of honestly predicting outcomes in individual patients with about 90% accuracy. The identified metagenes define distinct groups of genes, suggesting different biological processes underlying these two characteristics of breast cancer. Initial external validation comes from similarly accurate predictions of nodal status of a small sample in a quite distinct population group.

Interpretation Multiple aggregate measures of gene expression profiles define valuable predictive associations with lymph node metastasis and disease recurrence for the individual patient. These results indicate the potential for gene expression data to aid in achieving more accurate individualized prognosis. Importantly, this is evaluated in terms of precise numerical predictions, via ranges of probabilities of outcome, for the individual patient. Such precise and statistically valid assessments of patient-specific risk will ultimately be of most value to clinical practitioners faced with treatment decisions.
INTRODUCTION

Calibrating therapeutic intervention to an individual’s prognosis is central to effective oncologic treatment. In breast cancer, invasion into axillary lymph nodes is the most significant prognostic factor in breast cancer (1;2). Dissection of axillary nodes is consequently a crucial component of the therapeutic decision-making process. Newer, less invasive modalities for assessing lymph node status, such as sentinel node biopsy, are gaining acceptance (1), but it remains clear that clinico-pathologic parameters such as the presence or absence of positive axillary nodes represent the best means available to classify patients into broad subgroups by recurrence and survival (3-5). Even so it remains an imperfect tool. Among patients with no detectable lymph node involvement, a population thought to be in a low-risk category, between 22 and 33% develop recurrent disease after a 10-year follow-up (6). Properly identifying individuals out of this group who are at risk for recurrence is beyond current capabilities.

The question of lymph node diagnosis is part of the broader issue of more accurately predicting breast cancer disease course and recurrence. Though current clinical predictors are useful, they are just not accurate enough for prediction at the individual patient level. Genomic measures of gene expression, using microarrays and other technologies, provide new information that is now understood to identify patterns of gene activity that sub-classify tumors (7-10). Such patterns may correlate with the biological and clinical properties of the tumors, so it is of interest to investigate whether, and how, such data might add predictive value to current clinical predictors. Credible predictive evaluation is critical in establishing reproducible results and a key step towards integrating complex genomic data into prognoses for the individual patient (11-14).
The studies we report here move towards this goal in studying gene expression patterns predictively related to lymph node involvement and breast cancer recurrence in defined patient subgroups. We focus on predictions for the individual patient and aim to provide quantitative measures – in terms of probabilities of clinical phenotype and disease outcome -- that summarize the genomic information relevant to predicting at the individual level.

METHODS

MIAME (minimum information about a microarray experiment)-compliant information regarding the analyses performed here, as defined in the guidelines established by MGED (www.mged.org), is detailed in the following sections.

**Experimental design.** The analysis involved the use of a total of 89 tumor samples (details described below) for comparative gene expression measurements. The goal of the analysis was to identify those gene expression patterns characteristic of particular sets of tumor samples within the group based on the statistical analysis methods described below. These samples represent a heterogeneous population, and were selected based on clinical parameters and outcomes with the view to generating cases suitable for two focused studies, as reported here. Details of clinical characteristics of the 89 patients are provided in Table 1. For the lymph node study, external validation involved predicting outcomes on a subset of tumors from our previous Duke breast cancer study; full clinical and protocol details of this study are as previously reported (11). Each sample was hybridized once.
Samples used, extract preparation, and labeling. The 89 samples are from primary tumor biopsies at the Koo Foundation Sun Yat-Sen Cancer Center (KF-SYSCC) in Taipei, collected and banked between 1991-2001. Samples were collected under Duke (IRB# 3157-01) and KF-SYSCC (9/21/01, see Supplementary Material) Institutional Review Board guidelines. Total RNA was extracted from tumor tissue with Qiagen RNEasy kits, and assessed for quality with an Agilent Lab-on-a-Chip 2100 Bioanalyzer. Hybridization targets (probes for hybridization) were prepared from total RNA according to standard Affymetrix protocols.

Hybridization procedures and parameters. The amount of starting total RNA for each reaction was 20 µmcg. Briefly, first strand cDNA synthesis was generated using a T7-linked oligo-dT primer, followed by second strand synthesis. An in vitro transcription reaction was performed to generate the cRNA containing biotinylated UTP and CTP, which was subsequently chemically fragmented at 95°C for 35 min. The fragmented, biotinylated cRNA was hybridized in MES buffer (2-[N-morpholino]ethansulfonic acid) containing 0.5 mg/ml acetylated bovine serum albumin to Affymetrix GeneChip Human U95Av2 arrays at 45°C for 16hr, according to the Affymetrix protocol (www.affymetrix.com and www.affymetrix.com/products/arrays/specific/hgu95.affx). The arrays contain over 12,000 genes and ESTs. Arrays were washed and stained with streptavidin-phycoerythrin (SAPE, Molecular Probes). Signal amplification was performed using a biotinylated anti-streptavidin antibody (Vector Laboratories, Burlingame, CA) at 3 µmcg/ml. This was followed by a second staining with SAPE. Normal goat IgG (2 mg/ml) was used as a blocking agent.
**Measurement data and specifications.** Scans were performed with an Affymetrix GeneChip scanner and the expression value for each gene was calculated using the Affymetrix Microarray Analysis Suite (v5.0), computing the expression intensities in ‘signal’ units defined by software. Scaling factors were determined for each hybridization based on an arbitrary target intensity of 500. Scans were rejected if the scaling factor exceeded a factor of 25, resulting in only one reject. Files containing the computed single intensity value for each probe cell on the arrays (CEL files), files containing experimental and sample information (control info files), and files providing the signal intensity values for each probe set, as derived from the Affymetrix Microarray Analysis Suite (v5.0) software (pivot files), can be found in the Supplementary Material on the project web site.

**Array design.** All assays employed the Affymetrix Human U95Av2 GeneChip. The characteristics of the array are detailed on the Affymetrix web site [www.affymetrix.com/products/arrays/specific/hgu95.affx](http://www.affymetrix.com/products/arrays/specific/hgu95.affx).

**Statistical analysis.** Analysis uses predictive statistical tree models (15). This begins by applying k-means correlation-based clustering following an initial screen to remove genes varying at low levels, targeting a large number of clusters that are then used to generate a corresponding number of *metagene* patterns. Each metagene is the dominant singular factor (principal component) within a cluster, evaluated using the singular value decomposition (SVD). We identify 496 such factors this way, each representing the key common pattern of expression of the genes in the corresponding cluster. This strategy extracts multiple such patterns while reducing dimension and smoothing out gene-specific noise through the aggregation within clusters. Formal predictive analysis then
uses these metagenes in a Bayesian classification tree analysis. This generates multiple recursive partitions of the sample into subgroups (the “leaves” of the classification tree), and associates Bayesian predictive probabilities of outcomes with each subgroup. The analysis is applicable to even very small samples, and is developed to generate parsimonious models that are automatically resistant to over-fitting, as detailed previously (15). Overall predictions for an individual sample are then generated by averaging predictions, with appropriate weights, across many such tree models. We perform iterative out-of-sample, cross-validation predictions: leaving each tumor out of the data set one at a time, refitting the model (both the metagene factors and the partitions used) from the remaining tumors, and then predicting the hold-out case. This rigorously tests the predictive value of a model and mirrors the real-world prognostic context where prediction of new cases as they arise is the major goal.

**Supplementary information**

Additional information, including full details of all metegenes and complete details of the statistical tree methodology, is available at the project web site: [http://cgt.duke.edu/](http://cgt.duke.edu/)

**Role of funding source**

Partial support came from Koo Foundation Sun Yat-Sen Cancer Center (KFSYSCC) Research Fund. Several coauthors are personnel at KFSYSCC, and were involved in clinical data collection and design, and the writing and submission of this report.
RESULTS

*Gene expression patterns in primary breast tumors that predict lymph node metastasis*

The first study compares traditional “low-risk” versus “high-risk” patients, primarily based on lymph node status in order to evaluate the predictive associations of gene expression patterns with aggressive versus more benign tumors. Among ER positive individuals, the “high-risk” clinical profile is represented by advanced lymph node metastases (10 or more positive nodes); the “low-risk” profile identifies node-negative women of age greater than 40 years with tumor size below 2cm, precisely as currently used in clinical prognostic practice (15). Our data provides expression profiles on 18 high-risk and 19 low-risk cases (37 of the 89 total in Table 1) to which we applied the Bayesian statistical tree analysis. Figure 1 displays summary predictions from the resulting total of 37 cross-validation analyses. For each individual tumor, this graph illustrates the predicted probability for “high-risk” versus “low-risk” (red versus blue) together with an approximate 90% confidence interval, based on analysis of the 36 remaining tumors performed successively 37 times as each tumor prediction is made. It is important to recognize that each sample in the data set, when assayed in this manner, constitutes a validation set that accurately assesses the robustness of the predictive model. The metagene model accurately predicts nodal metastatic potential; about 90% (with 95% CI 79-99%) of cases are accurately predicted based on a simple threshold at 0.5 on the estimated probability in each case. Case number 7 is in the intermediate zone, exhibiting patterns of expression of the selected metagenes that relate equally well to those of “high” and “low-risk” cases, while case 22 is a clinical “high-risk” case with genomic expression patterns that relate more closely to “low-risk” cases. In contrast, node negative patients 5 and 11 have gene expression patterns more strongly indicative of “high-risk”,

8
and are key cases for follow-up investigations. The details of clinical information in these apparently discordant cases are shown in Table 2.

Clinical features of these few cases are illuminating, and suggestive of how a broader investigation of clinical data combined with molecular model-based predictions may aid in the eventual decision-making process. Case 22 did in fact recur, 6 years post-surgery; this patient’s classification as high-risk for recurrence based on purely clinical parameters was moderated by a lower risk based on metagenes, as demonstrated by this patient having survived recurrence-free for a longer time. Thus the lower probability prediction assigned to patient 22 based on the gene expression profiles is reflected in the clinical behavior of her disease. The clinically “low-risk” patient 7 recurred at 31 months, and patient 11 at 38 months, whereas case 5 is currently disease-free after only 12 months of follow-up. Cases 7 and 11 thus partly corroborate the predictions based on genomic criteria. With such predictions as part of a prognostic model, more intensive or innovative post-surgical therapy would have been indicated for these two cases.

A critical aspect of the analyses described here is allowing the complexity of distinct gene expression patterns to enter the predictive model. Tumors are graphed against metagene levels for three of the highest scoring metagene factors (Figure 2). This analysis highlights the need to analyze multiple aspects of gene expression patterns. For example, if the low-risk cases 1, 3 and 11 are assessed against metagene 146 alone, their levels are more consistent with high-risk cases. However, when additional dimensions are considered, the picture changes. The second frame (upper right) shows that low-risk is consistent with low levels of metagene 130 or high levels of metagene 146; hence, cases 1 and 3 are not inconsistent in the overall pattern, though case 11 is consistent. An analysis that selects one set of genes, summarized here as one metagene, as a “predictor”
would be potentially misleading, as it ignores the broader picture of multiple interlocked genomic patterns that together characterize a state. In the predictions, these two metagenes play key roles: low levels of metagene 146 coupled with higher levels of metagene 130 are strongly predictive of high-risk cases. Metagene 330 also plays a role and it is the combined use of multiple metagenes, in the context of the tree selection model building process, that ultimately yields a pattern that has the capacity to accurately predict the clinical outcome.

**External validation of lymph node metastasis predictors**

To extend this analysis to an independent data set, we used a small but relevant subset of the patient samples studied in a previous Duke breast cancer analysis (11). This is a limited initial study, but most supportive of the basic conclusion of predictive value of multiple metagene patterns. Relative to the Asian cohort, the Duke study patients had rather different characteristics: the racial difference, and the facts that the US women were generally much older and had much larger tumors at surgery. Further, the numbers of extreme (>9) lymph nodes are very small, so we relaxed the criteria for the two risk groups (ignoring age, reducing the number of positive nodes for the high-risk group, and substantially increasing the maximum tumor size for the low-risk group) in order to generate meaningful numbers of cases for study. This led to 6 low-risk cases (lymph node negative, ER+, tumor sizes less than 3.5cm which is the median size of the whole group) and 7 high-risk cases (at least 4 positive nodes, rather than 10). Additional complications are due to the fact that the expression data for this older study were obtained on an earlier Affymetrix microarray, so represent different though overlapping genes; full details of the process of mapping to the metagenes defined by the current study are provided as
Supplementary Material. In spite of these complications, and the resulting expectation that predictive accuracy would be reduced, the predictions based on precisely the model fitted to the Asian data are very accurate: one of the low-risks cases appears more consistent, in terms of metagene expression, with the high-risk cases, whereas the remaining 12 cases are very accurately predicted to lie within their defined risk groups. Interestingly, the apparently discrepant low-risk case (#42) has the largest tumor (3.5cm) of the group. Figure 3 exhibits the three key metagenes, in a format similar to Figure 2 but now including also these external validation cases, where concordance with the Asian samples is clear.

Gene expression patterns that predict recurrence of disease in breast cancer

The second analysis concerns 3 year recurrence following primary surgery among the challenging and varied subset of patients with 1-3 positive lymph nodes. Such patients typically receive adjuvant chemotherapy alone, and uniformly across this risk group, so that it is of interest to explain variations in outcome within this subgroup based on predictors other than treatment regimen. This is a critical subgroup as more than 20% suffer relapse within five years (5). Hence, improved prognosis for this heterogeneous group is of critical importance; patients identified with a high probability of relapse could be targeted for more intensive treatment. Our dataset provides expression profiles on 52 cases in this lymph node category (34 non-recurrent, 18 recurrent). The aggregate predictions from the sets of generated statistical tree models defines a rather accurate picture; once again, there is an approximate 90% (with 95% CI 82-99%) overall predictive accuracy in the 52 separate one-at-a-time, cross-validation prediction assessments (Figure 4).
Based on the gene expression analysis, the 3 year non-recurrent cases 6 and 23, having profiles more akin to recurrent cases, would be candidates for intensive treatment. These patients did receive adjuvant chemotherapy based on additional clinical risk factors (especially tumor size). Thus traditional clinical risk factors other than lymph node status also indicate higher risk of recurrence for these two cases, consistent with the molecular predictions. Each actually survived recurrence-free for over three years; case 6 recurred at 42 months and case 23 remains disease-free after over 6 years. Cases with low genomic criteria for recurrence would be 36, 38 and 42. They, however, each recurred within three years. These are cases that, under prognosis informed by only the genomic model, would have been indicated as more benign and not candidates for intensive treatment, whereas such a treatment might have proven to be more beneficial. Evidently, there is much yet to learn about the combinations of integrated genomic and clinical characteristics that will improve our capacity to identify such critical cases.

**Genes implicated in lymph node and recurrence studies**

Subsets of genes related to the metagene predictors of lymph node involvement are replete with those involved in cellular immunity including a high proportion of genes that function in the interferon pathway. They include genes that are induced by interferon such as various chemokines and chemokine receptors (Rantes, CXCL10, CCR2), other interferon-induced genes (IFI30, IFI35, IFI27, IFI44, IFIT1, IFIT4, IFITM3), as well as interferon effectors (2’-5’ oligoA synthetase), and genes encoding proteins mediating the induction of these genes in response to interferon (STAT1 and IRF1). This connection is intriguing given the role of interferon as a mediator of the anti-tumor response and, together with the fact that many genes involved in T cell function (TCRA, CD3D, IL2R,
MHC) are also included within the group that predict lymph node metastasis. Possibly, this may reflect the distinct nature of these tumors that have acquired a metastatic potential that elicits an anti-tumor response that is ultimately unsuccessful or an aberration of the normal anti-tumor response. Both of the key metagenes, 146 and 330, contain a number of these interferon related genes.

There is little intersection between the lists of genes defined by key metagenes here and those from the Duke lymph node study (11), which is perhaps not surprising given the relative heterogeneity of the patients in the Duke study. However, when the method of analysis used previously (11) is reapplied to the restricted subset of 6 low versus 7 high risk cases identified in the external validation study reported above, the 100 genes that most strongly relate to the categorization of lymph node status do indeed overlap with the top few metagenes of the current study. In particular, these include several genes already noted that are involved in an interferon response (STAT1, MX1, IFIT1, ISG115, IFI27, and IFI44).

Genes implicated in recurrence prediction do not exhibit such a striking functional clustering but do include many examples previously associated with breast cancer. Moreover, this group of genes is clearly distinct set from those that predict lymph node involvement. They include genes associated with cell proliferation control, both cell cycle specific activities (CDKN2D, Cyclin F, E2F4, DNA primase, DNA ligase), more general cell growth and signaling activities (MK2, JAK3, MAPK8IP, and EF1α), and a number of growth factor receptors and G-protein coupled receptors, some of which have been shown to facilitate breast tumor growth (EpoR). Possibly, the poor prognosis with respect to survival reflects a more vigorous proliferative capacity of the tumor.
We conclude that genes implicated in the prediction of lymph node metastasis and overall recurrence of disease, although clearly representing interrelated phenomena, nevertheless reflect the participation of distinct biological processes. The modeling approach we take here is flexible in this regard. The tree models select only those metagenes that are most relevant to the prediction in hand.

**DISCUSSION**

Personalized medicine aims to characterize those variables unique to the individual that determine disease susceptibility, response to therapy, and eventual disease outcome. We address this in assessing complex, multivariate patterns in gene expression data from primary tumor biopsies, and in exploring the value of such patterns in predicting lymph node metastasis and relapse. The resulting predictive accuracy of about 90%, and additional understanding of individual outcomes generated by the analysis, confirm the utility of gene expression patterns as prognostic factors in breast cancer. We stress the focus on predictions made in terms of numerical probabilities of outcomes for individual patients, with associated measures of uncertainties.

The lymph node risk group analysis defines metagene patterns capable of predicting high versus low risk cases with good accuracy, in both internal and external validation studies. In reanalysis of the small subset of samples from our early study (11) that relate most closely to the risk categories defined in this current study, we find improved predictions relative to our earlier methods and also a number of genes, including interferon-induced genes and others, in common. This provides additional support for the biological relevance of the metagene predictors identified, and suggests
areas for further pathway studies. The concordance between genomic predictors found between the Asian and US samples, though preliminary, is also a positive finding.

A related recurrence study (13) defines a single summary of gene expression related to breast cancer recurrence (though not nodal metastasis), generating a 70 gene predictor. We have been unable to identify more than 17 of these 70 genes on the Affymetrix array used here, and none of these appears in the key metagenes in our recurrence study. It will be of some interest to develop serious comparative studies that deal with cross-technology issues, and to develop future studies that combine and compare alternative summary predictors of outcome. The analysis approach used in (13) follows our own earlier work (11) in developing a single predictor based on an initial screen for genes most correlated with outcome. One distinction of our current work relative to these prior studies is the view that multiple measures of gene expression – multiple metagenes – may be involved in explaining differences and defining predictions. Investigation several metagenes, defining distinct patterns in the data relevant to the outcome, show how the combined effect of several views of clinico-biological data can highlight the similarities between patients while also identifying their differences. The non-linear statistical analysis aids in the elucidation of such patterns as they shed light on individual cases, as well as providing for informed predictions based on multiple patterns.

This latter point relates to the broader question of utilizing gene expression profiles into prognostic settings. We believe that it is the integration of genomic data with clinical risk factors that will determine the strategy for treating patients as individuals with distinct genomic disease features. Genomic data will not replace traditional clinical risk factors but will add significant detail to this clinical information, especially in a context such as breast cancer where multiple, interacting biological and environmental processes
define physiological states, and individual dimensions provide only partial information. As one initial example, our recurrence study here focuses on the 1-3 positive lymph node group where the analysis defines metagenes optimized for prediction within that group; predicting other subgroups, such as higher-risk cases in terms of lymph node count or subgroups stratified by additional clinical factors, will involve exploration of metagenes that optimally relate to outcomes within those subgroups.

Reliably improved predictions of disease course, including lymph node metastasis or recurrence, will profoundly affect the clinical decision process. Several studies indicate that 22-33% of node negative tumors behave in a manner similar to node positive tumors (6). Whether an issue of timing or of the inability to recognize histopathologic involvement of tumor material in the lymph nodes, a capacity to identify these cases as requiring more intensive clinical intervention could lead to an improvement in cancer survival. Previous attempts to correlate characteristics of primary tumors such as S-phase fraction, tumor grade, ploidy, c-erbB-2 overexpression, and hormone receptor status with lymph node metastasis have proven unsuccessful (16-18). The ability to appropriately utilize gene expression profiles provides opportunity to add enormous additional detail to the few, currently used biological attributes in tumor characterization. Finally, genes implicated in these analyses generate information of value for future pathway studies, with the potential to identify new targets that may feed into improved therapeutic strategies as well as improved understanding of genes related to the biology of metastasis and tumor evolution.
Acknowledgements

Research was supported by Synpac (North Carolina) and the Koo Foundation Sun Yat-Sen Cancer Center Research Fund, and by NSF under grants NSF DMS-0102227 and NSF DMS-0112340. We appreciate the constructive comments of three anonymous reviewers of the original submission.

Conflict of interest statements

There are no conflicts of interest.

Reference List


FIGURE LEGENDS

Figure 1. Cross-validation probability predictions of lymph node status. Samples (tumors) are plotted by index number, and the plotted numbers are marked on the vertical scale at the estimated predictive probabilities of high-risk (red) versus low-risk (blue). Approximate 90% uncertainty intervals about these estimated probabilities are indicated by vertical dashed lines.

Figure 2. Gene expression patterns from the major metagenes that predict lymph node status. Levels of metagenes for samples are plotted by sample index number and by color (color coding as in Figure 1).

Figure 3. Gene expression patterns from the major metagenes that predict lymph node status from current and earlier Duke breast cancer study. Levels of metagenes as in Figure 2, with current study samples now colored cyan (low-risk) and magenta (high-risk). External validation samples from the 2001 Duke breast cancer study appear as red (high-risk) and blue (low-risk).

Figure 4. Cross-validation probability predictions of 3-year recurrence. Samples (tumors) are plotted by index number, and the plotted numbers are marked on the vertical scale at the estimated predictive probabilities of 3 year recurrence (red) versus 3 year recurrence free survival (blue). Approximate 90% uncertainty intervals about these estimated probabilities are indicated by vertical dashed lines.
Figure 1
Figure 3
Figure 4
Table 1. Clinical characteristics of patients in the study

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 40</td>
<td>27</td>
<td>30.3</td>
</tr>
<tr>
<td>41-50</td>
<td>26</td>
<td>29.2</td>
</tr>
<tr>
<td>51-60</td>
<td>19</td>
<td>21.4</td>
</tr>
<tr>
<td>&gt; 60</td>
<td>17</td>
<td>19.1</td>
</tr>
<tr>
<td><strong>Histology type</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infiltrating Ductal Carcinoma</td>
<td>78</td>
<td>87.6</td>
</tr>
<tr>
<td>Infiltrating Lobular Carcinoma</td>
<td>2</td>
<td>2.3</td>
</tr>
<tr>
<td>Papillary Carcinoma</td>
<td>2</td>
<td>2.3</td>
</tr>
<tr>
<td>Tubular Carcinoma</td>
<td>1</td>
<td>1.1</td>
</tr>
<tr>
<td>Cribriform Carcinoma</td>
<td>1</td>
<td>1.1</td>
</tr>
<tr>
<td>Apocrine Carcinoma</td>
<td>1</td>
<td>1.1</td>
</tr>
<tr>
<td>Others ( mixed of histologies)</td>
<td>4</td>
<td>4.5</td>
</tr>
<tr>
<td><strong>Pathological tumor size</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1 cm</td>
<td>6</td>
<td>6.8</td>
</tr>
<tr>
<td>1 – 2 cm</td>
<td>31</td>
<td>34.8</td>
</tr>
<tr>
<td>2 – 5 cm</td>
<td>47</td>
<td>52.8</td>
</tr>
<tr>
<td>&gt; 5 cm</td>
<td>5</td>
<td>5.6</td>
</tr>
<tr>
<td><strong>Lymph node positive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>19</td>
<td>21.4</td>
</tr>
<tr>
<td>1 – 3</td>
<td>52</td>
<td>58.4</td>
</tr>
<tr>
<td>4 – 9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>18</td>
<td>20.2</td>
</tr>
<tr>
<td><strong>Nuclear grade</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade I</td>
<td>15</td>
<td>16.8</td>
</tr>
<tr>
<td>Grade II</td>
<td>24</td>
<td>27.0</td>
</tr>
<tr>
<td>Grade III</td>
<td>50</td>
<td>56.2</td>
</tr>
<tr>
<td><strong>LVI (peritumoral and intratumoral)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>35</td>
<td>39.3</td>
</tr>
<tr>
<td>Focal</td>
<td>16</td>
<td>18.0</td>
</tr>
<tr>
<td>Prominent</td>
<td>38</td>
<td>42.7</td>
</tr>
<tr>
<td><strong>ER status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>74</td>
<td>83.1</td>
</tr>
<tr>
<td>Negative</td>
<td>15</td>
<td>16.9</td>
</tr>
</tbody>
</table>
Table 2. Clinical information on discordant cases

<table>
<thead>
<tr>
<th>Case #</th>
<th>Surgery</th>
<th>RT</th>
<th>CT</th>
<th>Histology</th>
<th>Tumor size</th>
<th>Nodes</th>
<th>ER</th>
<th>PR</th>
<th>Relapse</th>
</tr>
</thead>
<tbody>
<tr>
<td>LN-5</td>
<td>MRM</td>
<td>N</td>
<td>CMF</td>
<td>IDC</td>
<td>2</td>
<td>0</td>
<td>+++</td>
<td>++</td>
<td>NED, 12 months</td>
</tr>
<tr>
<td>LN-7</td>
<td>MRM</td>
<td>N</td>
<td>No</td>
<td>IDC</td>
<td>1.7</td>
<td>0</td>
<td>+++</td>
<td>+++</td>
<td>Yes, 32 months</td>
</tr>
<tr>
<td>LN-11</td>
<td>BCS</td>
<td>Y</td>
<td>No</td>
<td>IDC</td>
<td>0.5</td>
<td>0</td>
<td>+</td>
<td>+++</td>
<td>Yes, 38 months</td>
</tr>
<tr>
<td>LN-22</td>
<td>MRM</td>
<td>Y</td>
<td>CEF</td>
<td>IDC</td>
<td>3</td>
<td>10</td>
<td>+</td>
<td>+</td>
<td>Yes, 75 months</td>
</tr>
<tr>
<td>Rec-38</td>
<td>MRM</td>
<td>N</td>
<td>No</td>
<td>TC</td>
<td>1.8</td>
<td>2</td>
<td>+</td>
<td>++</td>
<td>Yes, 11 months</td>
</tr>
<tr>
<td>Rec-23</td>
<td>MRM</td>
<td>N</td>
<td>CAF</td>
<td>IDC</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>NED, 74 months</td>
</tr>
<tr>
<td>Rec-6</td>
<td>MRM</td>
<td>N</td>
<td>CMF</td>
<td>ILC</td>
<td>3.1</td>
<td>2</td>
<td>+</td>
<td>+</td>
<td>Yes, 44 months</td>
</tr>
<tr>
<td>Rec-36</td>
<td>MRM</td>
<td>N</td>
<td>No</td>
<td>IDC</td>
<td>3.5</td>
<td>1</td>
<td>+</td>
<td>-</td>
<td>Yes, 6 months</td>
</tr>
<tr>
<td>Rec-42</td>
<td>MRM</td>
<td>N</td>
<td>CEF</td>
<td>IDC</td>
<td>3</td>
<td>2</td>
<td>+</td>
<td>+</td>
<td>Yes, 16 months</td>
</tr>
</tbody>
</table>

Abbreviations: MRM, modified radical mastectomy; RT, adjuvant Radiotherapy; CT, adjuvant chemotherapy; BCS, breast conserving surgery; NED, no evidence of disease; IDC, infiltrating ductal carcinoma; ILC, infiltrating lobular carcinoma; TC, tubular carcinoma
SUPPLEMENTARY INFORMATION

Supplementary material on genes and metagenes

Table 1. Genes associated with metagene predictors of lymph node metastasis.

Table 2. Genes associated with metagene predictors of breast cancer recurrence.

Table 3. Full list of genes defining all metagenes.

Supplementary material on statistical methods and data processing

Details on the specifics of data processing to evaluate metagene summaries for utilization in statistical analysis are provide here. Additional supplementary material includes the full technical report (ref. 15) that describes the statistical tree model methodology in complete detail.

Metagene summaries of gene expression profiles are obtained, for this breast cancer analysis, by combining standard clustering with also standard singular value decomposition (principal components) analysis. The precise steps taken in the study reported here are as follows:

- Raw data are the 12,625 signal intensity measures of expression of genes on the Affymetrix HU95aV2 DNA microarray, with signal intensities based on the Affymetrix V5 software then transformed to the log-base 2 scale. An initial screen reduces this to a total of 7,030 genes to remove sequences that vary at low levels or minimally. Specifically, this screens out genes whose expression levels
across all samples varies by less than two-fold, and whose maximum signal intensity value is lower than nine on a log-base 2 scale.

- The set of samples on these 7,030 genes are clustered using k-means correlated-based clustering. Any standard statistical package may be used for this; our analysis uses the xcluster software created by Gavin Sherlock at Stanford University (http://genome-www.stanford.edu/~sherlock/cluster.html). We defined a target of 500 clusters and the xcluster routine delivered 496 in this analysis.

- We extract the dominant singular factor (principal component) from each of the 496 clusters. Again, any standard statistical or numerical software package may be used for this; our analysis uses the reduced singular value decomposition function (svd) in Matlab (http://www.mathworks.com/products/matlab).

- These 496 metagene predictors are input to the tree model analysis as described in Pittman et al. 2002 (Ref 15) and available as a technical report in Supplementary Material. A key ingredient is the generalized likelihood ratio, or Bayes' factor, measure of association between metagenes and binary outcomes (Section 2.1 of the statistics paper). An initial ordering of metagenes is provided by the Bayes' factor values on all the data (at the root node of the tree). "Top" metagenes are those with highest Bayes' factor in this sense, and several "top" metagenes were selected to define the lists of genes (accompanying material) as described further below. Specifics parameters defined to create the precise tree models in the two breast examples are as follows (again with reference to Section 2 of the statistics paper). The tree model analysis as reported utilised a Bayes' factor threshold of 3 on the log scale, allowed up to 10 splits of the root node and then up to 4 at each
of nodes 1 and 2. Trees were allowed to grow to at most 2 levels consistent with the relatively small sample size of the data sets.

- Predictions for individual patients were performed as described in the paper: the analysis was repeated for each patient, holding out from the model fitting the expression and outcome data for that patient, and then developing the statistical tree model analysis based on only the remaining data. Then, the hold-out patient was predicted (using the statistical analysis as described in Section 2.4 and 2.5 of the statistics paper). We note that the model fitting, including the statistical evaluation of which metagenes are most predictive and the roles they play in the analysis (i.e., the “feature selection process”) is repeated anew for each of these analyses. Were this not done, and metagene selection based on all the data, then the predictions would appear much more accurate, but incorrectly and misleadingly so. This critical perspective, which we have terms “honest prediction” in the cross-validation context, is one we have taken pains to stress in our work (e.g., reference 11) and one that defines our approach to critical model evaluation when prediction is a primary focus.

- The lists of genes were generated precisely as follows, for each of the recurrence and metastasis analyses separately. From the statistical tree model fit to all the data, the "top" 4 metagenes were selected, based on the marginal Bayes' factor association measure as described. This defines 4 clusters of genes that are the initial basis of the list. The list was extended by adding in additional genes that are most highly correlated (standard linear correlation) with each of these 4 metagenes; the set of unique genes in the resulting lists are reported and form part
of this supplementary material, as are full details of all genes defining each of the 496 metagenes.

- In the lymph node metastasis external validation test, the predictions of the sample of cancers from the Duke 2001 PNAS study were performed directly using the tree model fitted only to the data from the current study (as described). That is, predictions were performed entirely out-of-sample with no modification at all to the definition of metagenes, the model or the details of analysis, so paralleling the "real life" circumstances of predicting new patients and providing a completely honest out-of-sample assessment of generalization and predictive validity.

- The metagene data for the Duke breast cancer samples used for external validation via out-of-sample prediction were evaluated as follows. The samples are from a 2000 study and gene expression profiles are on the early Affymetrix HU6800 array. The first step was then to identify all genes on that array (7,129 genes) that are also represented among the 12,625 genes on the U95av2 array. This was done using the chip-to-chip key available at the Affymetrix web site. This allows for the identification of genes on the HU6800 array that map to genes within each of the 496 metagene clusters from the current study. For example, the key metagenes 330, 146 and 130 have precisely 30, 37 and 8 genes, respectively; mapping these genes to the earlier HU6800 array identifies sets of 26, 42 and 4 genes, respectively (note that there are duplicates in some cases, as for metagene 146 here). These sets of genes on the HU6800 array define the metagene clusters and the corresponding value of the metagenes are evaluated precisely as described,
using the dominant singular factor (principal component) from each of the 496 clusters.

**Supplementary Figure.** Cross-validation and external validation probability predictions of lymph node status. Samples (tumors) are plotted by index number, and the plotted numbers are marked on the vertical scale at the estimated predictive probabilities of high-risk versus low risk. Color coding is as in Figure 3: predictions for the cases in the current study are the same in Figure 1, but now color coded as magenta (high-risk) and cyan (low risk), the cases from the Duke (PNAS 2001) study are correspondingly color coded red (high-risk) and blue (low-risk). Approximate 90% uncertainty intervals about these estimated probabilities are indicated by vertical dashed lines.
Supplementary Figure