Gene expression predictors of breast cancer outcomes

*Erich Huang, Skye H Cheng, Holly Dressman, Jennifer Pittman, Mei Hua Tsou, Cheng Fang Horng, Andrea Bild, Edwin S Iversen, Ming Liao, Chii Ming Chen, Mike West, Joseph R Nevins, Andrew T Huang

Summary

Background Correlation of risk factors with genomic data promises to provide specific treatment for individual patients, and needs interpretation of complex, multivariate patterns in gene expression data, as well as assessment of their ability to improve clinical predictions. We aimed to predict nodal metastatic states and relapse for breast cancer patients.

Methods We analysed DNA microarray data from samples of primary breast tumours, using non-linear statistical analyses to assess multiple patterns of interactions of groups of genes that have predictive value for the individual patient, with respect to lymph node metastasis and cancer recurrence.

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

Interpretation Multiple aggregate measures of profiles of gene expression define valuable predictive associations with lymph node metastasis and disease recurrence for individual patients. Gene expression data have the potential to aid accurate, individualised, prognosis. Importantly, these data are assessed in terms of precise numerical predictions, with ranges of probabilities of outcome. Precise and statistically valid assessments of risks specific for patients, will ultimately be of most value to clinicians faced with treatment decisions.

Lancet 2003; **361:** 1590–96
See Commentary page 1576

Koo Foundation Sun Yat-Sen Cancer Centre, Taipei, Taiwan (S H Cheng MD, M-H Tsou MD, C-F Horng MD, C-M Chen MS, A T Huang MD)**; Departments of Molecular Genetics and Microbiology** (E Huang MD PhD, H Dressman PhD, A Bild PhD, J R Nevins PhD)**, Medicine** (A T Huang MD)**, and Biostatistics and Bioinformatics** (E S Iversen PhD)**, and Institute of Statistics and Decision Sciences, Duke University** (J Pittman PhD, E S Iversen PhD, M Liao BS, M West PhD), **Durham, NC, USA; and Howard Hughes Medical Institute** (J R Nevins PhD)**, Durham, NC, USA**

Correspondence to: Dr Joseph R Nevins, Department of Molecular Genetics and Microbiology, Duke University Medical Centre, Durham, NC 27708, USA
(e-mail: j.nevins@duke.edu)

Introduction

Calibration of therapeutic intervention for an individual's outlook is central to effective oncological treatment. In breast cancer, invasion into axillary lymph nodes is the most important prognostic factor.[1,2] Dissection of axillary nodes is therefore crucial in therapeutic decision making. New, less invasive, methods for assessment of lymph node status—such as sentinel node biopsy—are gaining acceptance,[1] but clinico-pathological indices, such as the presence or absence of positive axillary nodes, remain the best way to classify patients into broad subgroups by recurrence and survival.[3–5] In patients with no detectable lymph-node involvement, a population thought to be at low-risk, between 22% and 33% develop recurrent disease after a 10-year follow-up.[6] Identification of individuals in this group who are at risk for recurrence cannot be done at present.

Diagnosis of lymph-node status is important in accurate prediction of disease course and recurrence of breast cancer. Although clinical predictors are useful, they are not accurate enough for prediction in the individual patient. Genomic measures of gene expression provide new information to identify patterns of gene activity that subclassify tumours.[7–10] Such patterns might correlate with biological and clinical properties of the tumours, so we could usefully investigate whether, and how, such data might add predictive value to clinical predictors. Credible assessment of predictors is critical to establish reproducible results, and a key step towards integration of complex genomic data into outlook for individual patients.[11-14]

Here, we move towards this goal by looking at gene expression patterns that predict involvement of the lymph node and recurrence of breast cancer in defined patient subgroups. We focus on predictions for the individual patient and aim to provide quantitative measures—in respect of probabilities of clinical phenotype and disease outcome—that summarise the genomic information relevant to such prediction.

Methods

Procedures

The analyses detailed here comply with MIAME (minimal information about a microarray experiment)-guidelines established by the Microarray gene expression data society (www.mged.org). The analysis used 89 tumour samples for comparative measurements of gene expression. Our goal was to identify gene expression patterns that are characteristic of particular sets of tumour samples within the group. These samples represent a heterogeneous population, and were selected on the basis of clinical parameters and outcomes, to generate cases suitable for two focused studies. Table 1 shows details of clinical characteristics of the 89 patients. For the lymph node study, external validation was done with prediction of outcomes for a subset of tumours from our previous breast

GLOSSARY

K-MEANS CLUSTERING
Standard clustering of genes into a number, (k) of separate groups. Correlation-based k-means defines groups to include genes most highly related in terms of correlation calculated across samples.

SINGULAR VALUE DECOMPOSITION
A mathematical procedure by which trends in large datasets can be noted.

GENE-SPECIFIC NOISE
Variation in measurements of gene expression that is not due to underlying causes across a set of genes, and mainly relate to experimental and processing errors.

RECURSIVE PARTITIONING
The reduction of the sample to a set of subsamples by successive binary divisions.

PARSIMONIOUS
A parsimonious statistical model is the least complex of candidate models that fit and predict a data set.

TREE MODELS
A statistical tree model is a recursive partition of a sample data set into a set of subsamples defined as the terminal nodes of a series of successive binary partitions of the data. The "tree" begins with all data in a root node, and the node is split to separate the samples into two subsamples. Each of these is then candidate for a further binary split. The method by which nodes are split (or not), and by which statistical inference is made, is part of the model specification.

HOLD-OUT CASE
In a cross-validation analysis, a sample observation taken out of the data set and predicted on the basis of the model fitted to the remaining data. In one-at-a-time cross-validation, each observation is treated as a single hold-out case in a succession of independent re-analyses of the remaining data.

cancer study, full clinical and protocol details of which are as previously reported.[11] Every sample was hybridised once.

The 89 samples were obtained at biopsy of primary tumour at the Koo Foundation Sun Yat-Sen Cancer Centre (KF-SYSCC), Taipei, collected and banked between 1991 and 2001. Samples were taken (http://image.thelancet.com/extras/02art11142webappendix.pdf) according to institutional review board guidelines. Total RNA was extracted from tumour tissue with Qiagen RNEasy kits (Qiagen, Valencia, CA, USA), and assessed for quality with an Agilent Lab-on-a-Chip 2100 Bioanalyzer (Agilent, Palo Alto, CA, USA).

The amount of starting total RNA for each reaction was 20 μg. Synthesis of the first strand of cDNA was done by a T7-linked oligo-dT primer, followed by second strand synthesis. An in-vitro transcription reaction was used to generate the cRNA containing biotinylated uracil triphosphate and cytosine triphosphate, which was then chemically fragmented at 95°C for 35 min. The fragmented, biotinylated cRNA was hybridised, in MES buffer (2-[N-morpholino]ethansulfonic acid) containing 0·5 mg/mL acetylated bovine serum albumin, to human U95Av2 gene chip arrays (Affymetrix, Santa Clara, CA, USA), containing over 12 000 genes and expressed sequence tags (EST), at 45°C for 16 h, according to the manufacturer's instructions.[15] Arrays were washed and stained with streptavidin-phycoerythrin (SAPE, Molecular Probes Eugene, OR, USA). Signal amplification was done by a biotinylated anti-streptavidin antibody (Vector Laboratories, Burlingame, CA, USA) at a concentration of 3 μg/mL. This step was followed by a second staining with SAPE. Goat IgG (2 mg/mL) was used as a blocking agent.

Scans were done with an Affymetrix GeneArray scanner, and the expression value for every gene was calculated using the Affymetrix microarray suite (version 5.0), computing the expression intensities in signal units defined by the software. Scaling factors were determined for each hybridisation on the basis of 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 every probe cell on the arrays (CEL files), those with experimental and sample information (control information files), and files providing the signal intensity values for every probe set, as derived from the Affymetrix software (pivot files), can be found at http://www.cagp.duke.edu.

	Number (%)
Age (years)	
<40	27 (30%)
41–50	26 (29%)
51–60	19 (21%)
>60	17 (19%)
Histology type	
Infiltrating ductal carcinoma	78 (88%)
Infiltrating lobular carcinoma	2 (2%)
Papillary carcinoma	2 (2%)
Tubular carcinoma	1 (1%)
Cribriform carcinoma	1 (1%)
Apocrine carcinoma	1 (1%)
Others (mixed histological findings)	4 (4%)
Pathological tumour size (cm)	
<1 cm	6 (7%)
1–2 cm	31 (35%)
2–5 cm	47 (53%)
>5 cm	5 (6%)
Lymph node positive (number)	
0	19 (21%)
1–3	52 (58%)
4–9	0 (0%)
>10	18 (20%)
Nuclear grade	
Grade I	15 (17%)
Grade II	24 (27%)
Grade III	50 (56%)
LVI (peritumoral and intratumoral)	
Absent	35 (39%)
Focal	16 (18%)
Prominent	38 (43%)
ER status	
Positive	74 (83%)
Negative	15 (17%)

ER=oestrogen receptor; LVI=lymphatic vessel invasion.

Table 1: **Clinical characteristics of 89 patients**

To extend this analysis to an independent data set, we used a small but relevant subset of patient samples from our previous (US) study.[11] Compared with the Asian cohort reported here, the patients in the US study were generally much older with larger tumours at surgery. Very few women have a large number (>9) of lymph nodes. To generate meaningful numbers of cases, we reassigned the risk criteria by ignoring age, reducing the number of positive nodes for the high-risk group and substantially increasing the maximum tumour size for the low-risk group. On this basis 13 patients met these criteria.

Statistical analysis

Our analysis used predictive statistical tree models (unpublished). This model begins by applying k-means correlation-based clustering, after an initial screen to remove genes that show little variation, targeting numerous clusters that generate a corresponding number of metagene patterns. Every metagene is the dominant single factor (principal component) within a cluster, as assessed by the singular value decomposition. We identified 496 such factors this way, each representing the

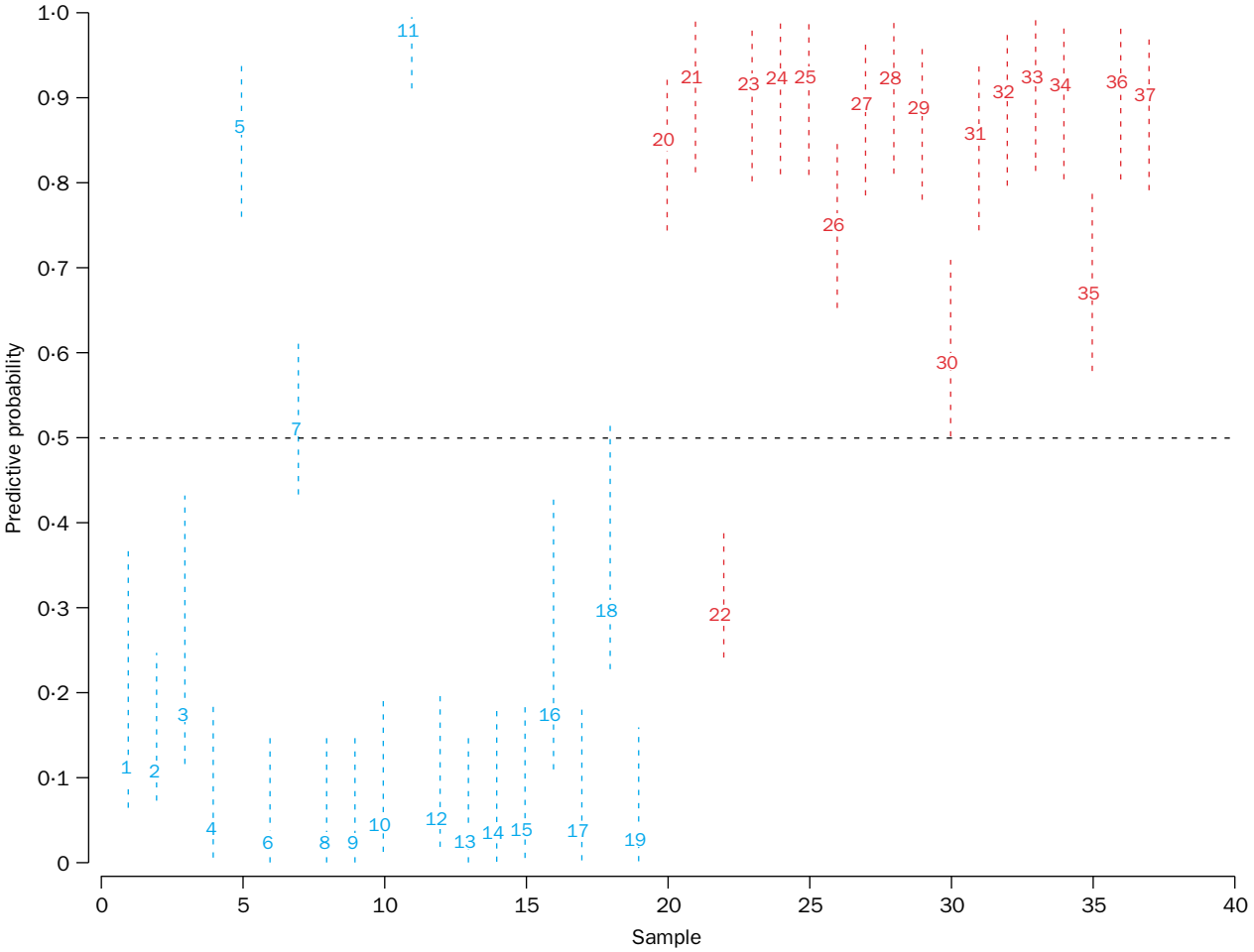

Figure 1: **Cross-validation probability predictions of lymph-node status in 37 tumours**
Tumour samples are plotted by index number, and the plotted numbers are marked on the vertical scale at the estimated predictive probabilities. Red=high-risk; blue=low-risk. Vertical lines=approximate 90% uncertainty intervals about these estimated probabilities.

key common pattern of expression of the genes in the corresponding cluster. This strategy helps to obtain many such patterns while reducing dimension and smoothing out gene-specific noise through the aggregation within clusters. For predictive analysis, we then used these metagenes in a Bayesian classification tree analysis. This method generates multiple recursive partitions of the sample into subgroups (the leaves of the classification tree) and associate Bayesian predictive probabilities of outcomes with each subgroup. The analysis is also applicable to very small samples, and was developed to generate parsimonious models that are automatically resistant to over-fitting (unpublished). Overall predictions for an individual sample are then produced by taking an average of predictions, with appropriate weights, across many such tree models. We did iterative out-of-sample, cross-validation predictions by leaving every tumour out of the data set one at a time, refitting the model (both the metagene factors and the partitions used) from the remaining tumours, and then predicting the hold-out case. This method rigorously tests the predictive value of a model and is much like the practical prognostic context in clinical practice, in which the primary goal is to predict new cases as they arise.

Additional information, including full details of all metagenes (webtable 1: http://image.thelancet.com/extras/

Case number	Surgery	RT	CT	Histology	Tumour size (cm)	Nodes	ER	PR	Relapse
Asian study									
5	MRM	No	CMF	IDC	2	0	+++	++	NED, 12 months
7	MRM	No	No	IDC	1·7	0	+++	+++	Yes, 32 months
11	BCS	Yes	No	IDC	0·5	0	+	+++	Yes, 38 months
22	MRM	Yes	CEF	IDC	3	10	+	+	Yes, 75 months
US study									
38	MRM	No	No	TC	1·8	2	+	++	Yes, 11 months
23	MRM	No	CAF	IDC	3	1	–	–	NED, 74 months
6	MRM	No	CMF	ILC	3·1	2	+	+	Yes, 44 months
36	MRM	No	No	IDC	3·5	1	+	–	Yes, 6 months
42	MRM	No	CEF	IDC	3	2	+	+	Yes, 16 months

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; ER=oestrogen receptor; PR=progesterone receptor; CMF=cyclophosphamide, methotrexate, fluorouracil; CAF=cyclophosphamide, adriamycin, fluorouracil; CEF=cyclophosphamide, epirubicin, fluorouracil.

Table 2: **Clinical information for discordant cases**

02art11142webtable1.pdf) and complete details of the statistical tree methodology, are available at http://cagp.duke.edu.

Role of the funding source

Part funding came from KF-SYSCC Research Fund. Several investigators are personnel at KF-SYSCC, and took part in the study design, clinical data collection, and the writing and submission of this report. The sponsor had no role in study design, data collection, data analysis, data interpretation, or writing of this report.

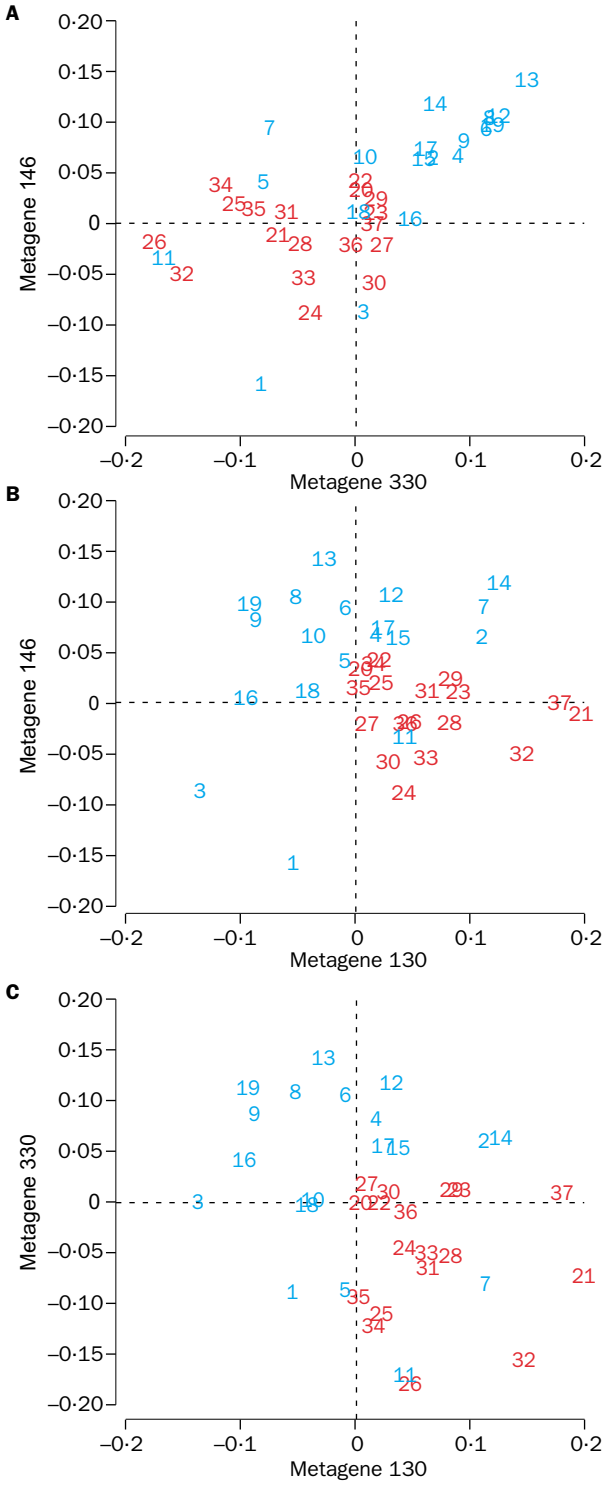

Figure 2: **Gene expression patterns from the major metagenes that predict lymph node status**

A=metagenes 146 and 330; B=metagenes 146 and 130; C=metagenes 330 and 130. Red=high-risk; blue=low-risk.

Results

In our previous study we compared low-risk versus high-risk patients, mainly based on lymph node status, for assessment of the predictive associations of gene expression patterns with aggressive versus benign tumours. In oestrogen receptor (ER) positive individuals the high-risk clinical profile is represented by advanced lymph node metastases (ten or more positive nodes); the low-risk profile identifies node negative women older than 40 years of age, with tumour size less

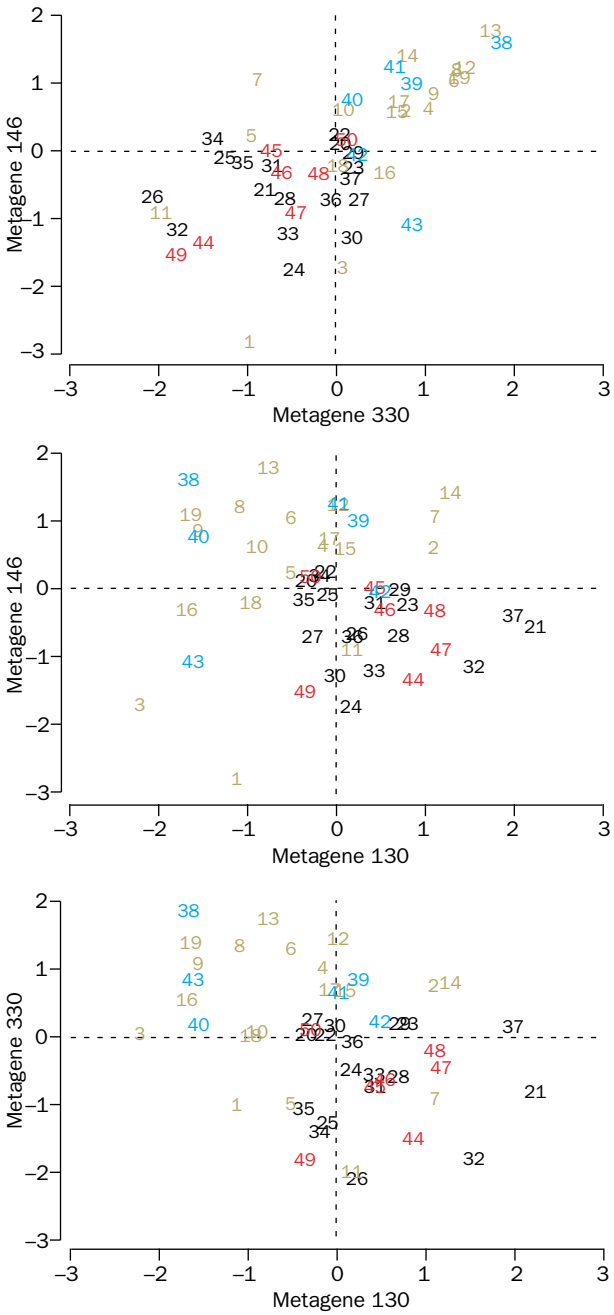

Figure 3: **Gene expression patterns from the major metagenes that predict lymph-node status from both present and previous studies**

Brown=current low-risk; black=current high-risk; red=previous high-risk; blue-previous low-risk.

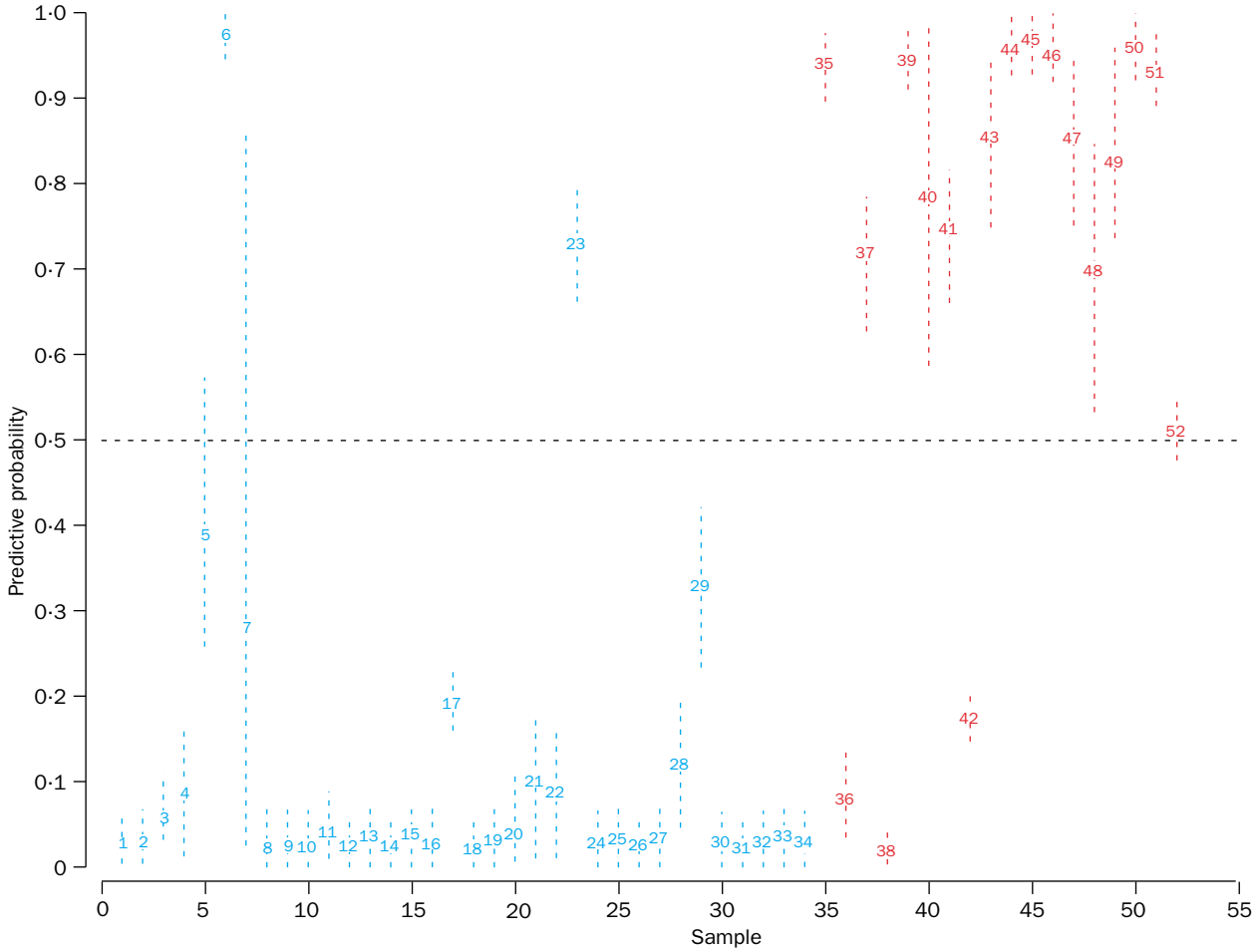

Figure 4: **Cross-validation probability predictions of 3-year recurrence**
Tumour samples are plotted by index number, and the plotted numbers are marked on the vertical scale at the estimated predictive probabilities of 3 year recurrence (red) versus 3 year recurrence free survival (blue). Vertical dashed lines=approximate 90% uncertainty intervals for these estimated probabilities.

than 2 cm. These definitions are those used for prognosis assessment in practice (unpublished). Our data provide 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 shows summary predictions from the resulting total of 37 cross-validation analyses. For individual tumours, this graph illustrates the predicted probability for high-risk versus low-risk together with a 90% confidence interval, based on analysis of the 36 remaining tumours, done successively for all 37 tumours. The samples, when assayed in this way, constitute a validation set that accurately assesses the robustness of the predictive model. The metagene model predicts nodal metastatic potential; about 90% (95% CI 79–99%) of cases are correctly predicted on the basis of a threshold of 0·5 of the estimated probability in all cases. Case 7 is in the intermediate zone, showing patterns of expression of the selected metagenes that relate equally well to those of high-risk and low-risk, whereas case 22 is a clinical high-risk case with genomic expression patterns that relate more closely to low-risk. By contrast, node negative (low-risk) patients 5 and 11 have gene expression patterns that are strongly indicative of high-risk, and are key cases for follow-up investigations. Table 2 shows the details of clinical information in these apparently discordant cases.

Our findings based on the clinical features of these few cases suggest how a broad investigation of both clinical data and predictions based on molecular models could aid clinical decisions. Case 22 did in fact recur, 6 years postsurgery; this patient's classification as high-risk for recurrence based on purely clinical indices was moderated by a lower risk based on metagenes, as shown by their long recurrence-free survival. Thus, the prediction of low probability of recurrence assigned to patient 22 on the basis of gene expression profiles, is corroborated by the clinical features of her disease. In the clinically low-risk patient 7, disease recurred at 32 months, and in patient 11 at 38 months, whereas case 5 was free of disease after 12 months of follow-up. Cases 7 and 11 thus partly corroborate the predictions based on genomic criteria. If such predictions were part of the prognostic model, more intensive postsurgical treatment than they received would have been indicated for these two patients.

A critical part of the analyses is that the complexity of distinct gene expression patterns is allowed to be entered into the predictive model. Tumour metagene values were plotted on a graph for three of the highest scoring metagene factors (figure 2). This analysis points to the need to analyse many features of gene expression patterns. For example, if the low-risk cases 1, 3, and 11 are assessed against metagene 146 alone (figure 2, A), their values are more consistent with high-risk rather than low-risk cases. However, when additional dimensions are considered, the picture changes. The figure also shows that low-risk is consistent with low levels of metagene 130 or high levels of metagene 146 (figure 2, B); thus, cases 1 and 3 are more-

or-less inconsistent in the overall pattern, although case 11 is wholly consistent. An analysis that selects one set of genes, summarised here as one metagene, as a predictor might be potentially misleading, because it ignores many interlinked genomic patterns that together characterise a state. These two metagenes (ie, 130 and 146) have key roles, and low values of metagene 146 coupled with high values of metagene 130 strongly predicts high-risk cases. Metagene 330 also plays a part (figure 2, C) and it is the combination of multiple metagenes, in the context of the tree selection model building process, that ultimately has the ability to accurately predict the clinical outcome.

The results from a subset of the patient samples investigated our previous study (in the US),[11] although a small initial study, lends support to the predictive value of multiple metagene patterns. The change of criteria led to six low-risk cases (lymph node negative, ER positive, tumour sizes less than 3·5 cm, which is the median size of the whole group) and seven high-risk cases (at least four positive nodes, rather than ten). A drawback to this comparison was that the expression data for the previous study were obtained with an earlier Affymetrix microarray, and so represent different though overlapping genes, but the predictions based on the model fitted to our new data are accurate. Full details of the process of mapping to the metagenes defined by the present study are provided in the appendix.

One of the low-risk cases seemed consistent, in terms of metagene expression, with the high-risk cases, whereas the remaining 12 cases were accurately predicted to lie within their defined risk groups (webfigure: http://image.thelancet.com/extras/02art11142webfigure.pdf). The apparently discrepant low-risk case (case 42) had the largest tumour (3·5 cm) of the group. Figure 3 shows the three key metagenes, in a format similar to figure 2, but also includes these external validation cases, for whom concordance withthe 89 new samples is clear.

We did another analysis of recurrence 3 years after primary surgery, in the subset of patients with one to three positive lymph nodes. Our data set provided expression profiles for 52 cases in this lymph node category (34 non-recurrent, 18 recurrent). The aggregate predictions from the sets of generated statistical tree models defined an accurate picture; there was about 90% (with 95% CI 82–99%) overall accuracy in the 52 individual cross-validation prediction assessments (figure 4).

On the basis of gene-expression analysis, cases 6 and 23 (non-recurrent for at least 3 years), had profiles close to that of recurrent cases, and would be candidates for intensive treatment. These patients did receive adjuvant chemotherapy because of additional clinical risk factors (especially tumour size). Thus, clinical risk factors other than lymph-node status also indicate high risk of recurrence for these two cases, consistent with the molecular predictions. Each survived free of recurrence for more than 3 years; case 6 recurred at 44 months and case 23 remained disease-free for more than 6 years. Cases 36, 38, and 42 had low genomic criteria for recurrence, but they all recurred well within 3 years. In these three cases, if outlook had been determined only by the genomic model, disease would have been regarded as benign, and the patients would not have been thought candidates for intensive treatment that might have proven beneficial.

Subsets of genes related to the metagene predictors of lymph-node involvement contain many for cellular immunity (webtable 2: http://image.thelancet.com/extras/02art11142webtable2.pdf). They include genes that are induced by interferons, 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′ oligo A synthetase); and genes encoding proteins that mediate the induction of these genes in response to interferon (*STAT1* and *IRF1*).

Discussion

Our assessment of complex, multivariate patterns in gene-expression data from primary tumour biopsy specimens, and examination of the value of such patterns in prediction of lymph-node metastasis and relapse resulted in a predictive accuracy of about 90%. The analysis provided additional understanding of individual outcomes and confirmed the use of gene expression patterns as prognostic factors in breast cancer. The group analysis of lymph-node risk defines metagene patterns that can accurately predict high-risk versus low-risk cases, in both internal and external validation studies.

In reanalysis of the small subset of samples from our early study[11] that related most closely to the risk categories defined in this present study, we noted improved predictions relative to our earlier methods as well as several shared genes, including interferon-induced genes. Patients (with one to three positive lymph nodes) typically receive adjuvant chemotherapy alone so the explanation of variations in outcome within this subgroup (based on predictors other than treatment regimen), could prove useful. This is a critical subgroup because more than 20% have a relapse within 5 years.[5] Thus, improved outlook for this heterogeneous group is important because patients identified with a high probability of relapse can be targeted for intensive treatment. The concordance between genomic predictors found between the two sets of samples, though preliminary, is also a positive finding.

The connection between the metagene predictors and genes for interferons is intriguing in view of the role of interferons as mediators of the antitumour response and 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. This link might indicate the distinct nature of the tumours with metastatic potential that elicit an anti-tumour response that proves unsuccessful, or might indicate an aberration of the healthy anti-tumour response. Both the key metagenes 146 and 330 contain several of these interferon-related genes.

The key metagenes defined here, and those from the US study,[11] do not share many genes, which is perhaps not surprising in view of the relative heterogeneity of the patients in the previous study relative to that of the cohort reported here. However, when the method of analysis used previously[11] is reapplied to the restricted subset of seven low-risk versus seven high-risk cases identified, the 100 genes that most strongly relate to the categorisation of lymph-node status do indeed overlap with the top few metagenes of the present study. In particular, these 100 include several genes that are implicated in an interferon response (*STAT1*, *MX1*, *IFIT1*, *ISG115*, *IFI27*, and *IFI44*).

Genes implicated in recurrence prediction do not show such a striking functional clustering but do include many features previously associated with breast cancer (webtable 3: http://image.thelancet.com/extras/02art11142webtable3.pdf). Moreover, this group of genes is distinct from those that predict lymph node involvement. Its features include genes associated with cell proliferation control: (a) activities specific to cell-cyle (*CDKN2D*,

CCNF, *E2F4*, and the enzymes DNA primase, and DNA ligase), (b) general cell growth and signalling activities (*MAPKAPK2*, *JAK3*, *MAPK8IP*, and *EEF1A1*), and (c) several growth factor receptors and G-protein coupled receptors, some of which facilitate breast tumour growth (EpoR). Poor outlook with respect to survival is related to the vigorous proliferative ability of the tumour.

We conclude that genes implicated in the prediction of lymph-node metastasis and overall recurrence of disease, although clearly representing interrelated events, nevertheless indicate the participation of distinct biological processes. The modelling approach we take here is flexible in this respect. The tree models select only those metagenes that are most relevant to the prediction.

Another recurrence study[13] defines one pattern of gene expression related to breast cancer recurrence (though not nodal metastasis) that generates a 70-gene predictor. We have been unable to identify any 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 might be useful to develop comparative studies that allow for cross-technology issues and that look at alternative summary predictors of outcome. van 't Veer and others'[13] analysis follows up our earlier work[11] in development of a single predictor through an initial screen for genes that have high correlation with outcome. One difference between this report and previous studies is our view that multiple measures of gene expression—multiple metagenes—might account for differences and define predictions.

Investigation of several metagenes and definition of distinct patterns in the data relevant to the outcome, show how the combination of various clinico-biological data can underscore both the similarities and the differences between patients. Non-linear statistical analysis helps us to understand such patterns and their relevance to individual cases. It also provides information that allows informed predictions based on multiple patterns that relate to the use of gene expression profiles in 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 substantial detail to this clinical information, especially in a disease such as breast cancer in which multiple, interacting biological and environmental processes define physiological states, and individual dimensions provide only some information. As an initial example, our recurrence study here focuses on the group with one to three positive lymph nodes, in which the analysis defines the most appropriate metagenes for prediction within that group. Prediction of other subgroups, such as high-risk cases in terms of lymph-node count, or subgroups stratified by additional clinical factors, will mean examination of metagenes that best relate to outcomes within those subgroups.

Improved predictions of disease course, including lymph-node metastasis or recurrence, will profoundly affect clinical decisions. Results of several studies show that 22–33% of node-negative tumours behave like node-positive tumours.[6] The ability to identify the cases that need intensive clinical intervention could lead to an improvement in cancer survival. Previous attempts to correlate characteristics of primary tumours such as S-phase fraction, tumour grade, ploidy, ERBB2 overexpression, and hormone receptor status with lymph-node metastasis have proven unsuccessful.[15-17] The ability to use appropriately profiles of gene expression could add enormous detail to the few known biological attributes in tumour characterisation. Finally, genes implicated in these analyses generate valuable information for future pathway studies, with the potential to identify new targets that might contribute to improved treatment as well as better understanding of genes that are related to metastasis and tumour growth.

Contributors

E Huang, A T Huang, J R Nevins, and M West designed the study. E Huang, H Dressman, S H Cheng, and M-H Tsou coordinated the study, data collection, and sample processing. H Dressman did the microarray analyses and, together with A Bild and J R Nevins, did the informatics analysis of the identified genes. J Pittman, E S Iversen, M Liao, and M West developed the statistical models and did the data analysis. S H Cheng, M-H Tsou, C-F Horng, and C-M Chen managed and analysed clinical information and, with J Pittman, developed the clinical data informatics. All investigators contributed to and reviewed the final report.

Conflict of interest statement

None declared.

Acknowledgments

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

References

1 Krag D, Weaver D, Ashikaga T, et al. The sentinel node in breast cancer - a multicenter validation study. *N Engl J Med* 1998; **339:** 941–46.

2 Singletary SE, Allred C, Ashley P, et al. Revision of the American Joint Committee on cancer staging system for breast cancer. *J Clin Oncol* 2002; **20:** 3628–36.

3 Overgaard M, Hansen PS, Overgaard J, et al. The Danish Breast Cancer Cooperative Group 82b Trial: Postoperative Radiotherapy in High-Rish Premenopausal Women with Breast Cancer Who Receive Adjuvant Chemotherapy. *N Engl J Med* 1997; **337:** 949–55.

4 Jatoi I, Hilsenbeck SG, Clark GM, Osborne CK. Significance of axillary lymph node metastasis in primary breast cancer. *J Clin Oncol* 1999; **17:** 2334–40.

5 Cheng SH, Tsou MH, Liu MC, et al. Unique features of breast cancer in Taiwan. *Breast Cancer Res Treat* 2000; **63:** 213–23.

6 Polychemotherapy for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. *Lancet* 2001; **352:** 930–42.

7 Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001; **98:** 13790–95.

8 Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; **403:** 503–11.

9 Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* 2000; **406:** 747–52.

10 Yeoh E-J, Ross ME, Shurtleff SA, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002; **1:** 133–43.

11 West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 2001; **98:** 11462–67.

12 Spang R, Zuzan H, West M, Nevins JR, Blanchette C, Marks J. Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biol* 2002; **2:** 369–81.

13 van'T Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415:** 530–36.

14 Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286:** 531–37.

15 Mittra I, MacRae KD. A Meta-analysis of reported correlations between prognostic factors in breast cancer: does axillary lymph node metastasis represent biology or chronology? *Eur J Cancer* 1991; **27:** 1574–83.

16 McGuire WL. Prognostic factors for recurrence and survival in human breast cancer. *Breast Cancer Res Treat* 1987; **10:** 5–9.

17 Tandon AK, Clark GM, Chamness GC, Ullrich A, McGuire WL. HER-2/neu oncogene protein and prognosis in breast cancer. *J Clin Oncol* 1989; **7:** 1120–28.