Gene Expression Signature Predicts Recurrence in Lung Adenocarcinoma

Jill E. Larsen,[1,2] Sandra J. Pavey,[2,3] Linda H. Passmore,[1] Rayleen V. Bowman,[2] Nicholas K. Hayward,[2,3] and Kwun M. Fong[1,2]

Abstract

Purpose: Improving outcomes for early-stage lung cancer is a major research focus at present because a significant proportion of stage I patients develop recurrent disease within 5 years of curative-intent lung resection. Within tumor stage groups, conventional prognostic indicators currently fail to predict relapse accurately.

Experimental Design: To identify a gene signature predictive of recurrence in primary lung adenocarcinoma, we analyzed gene expression profiles in a training set of 48 node-negative tumors (stage I-II), comparing tumors from cases who remained disease-free for a minimum of 36 months with those from cases whose disease recurred within 18 months of complete resection.

Results: Cox proportional hazards modeling with leave-one-out cross-validation identified a 54-gene signature capable of predicting risk of recurrence in two independent validation cohorts of 55 adenocarcinomas [log-rank $P = 0.039$; hazard ratio (HR), 2.2; 95% confidence interval (95% CI), 1.1-4.7] and 40 adenocarcinomas (log-rank $P = 0.044$; HR, 3.3; 95% CI, 1.4-7.9). Kaplan-Meier log-rank analysis found that predicted poor-outcome groups had significantly shorter survival, and furthermore, the signature predicted outcome independently of conventional indicators of tumor stage and node stage. In a subset of earliest stage adenocarcinomas, generally expected to have good outcome, the signature predicted samples with significantly poorer survival.

Conclusions: We describe a 54-gene signature that predicts the risk of recurrent disease independently of tumor stage and which therefore has potential to refine clinical prognosis for patients undergoing resection for primary adenocarcinoma of the lung.

Lung cancer remains the leading cause of cancer death in Western countries, with an overall 5-year survival of 10% to 15% (1). Complete surgical resection remains the most effective treatment, but despite clinical and technical advancements, the outcome of lung cancer has not improved significantly during the last 20 years. Treatment failure is frequently attributable to the presence of undetectable and unpredictable micrometastases (2). Generally, early-stage tumors have better clinical outcome with stage I non–small cell lung carcinomas (NSCLC) having a 5-year survival of ~65% (3–5). However, disease recurrence has been reported to occur in up to 30% to 40% of stage I NSCLCs (6, 7), indicating that despite aiding treatment planning, there are limitations in current clinical staging techniques.

NSCLCs constitute ~80% of all lung cancers, with small cell carcinomas making up the remaining 20%. The NSCLC group can be further divided into histologic subtypes with adenocarcinoma, squamous cell carcinoma (SCC), and large cell carcinoma being the most common, accounting for 40%, 27%, and 8% of all lung cancers, respectively (8). We previously found a gene expression signature to predict tumor recurrence in lung SCCs (9) and, consequently, sought to determine if a recurrence signature also existed in lung adenocarcinomas.

Gene expression profiling has been used to characterize prognosis in lung cancer (10–22), mostly with survival rather than tumor recurrence as an end point. During the conduct of our studies, three other research groups have reported expression profiles associated with recurrence, in a small SCC cohort (18), and two cohorts of mixed adenocarcinoma and SCC (15, 21). We studied tumor gene expression profiles in a large, well-defined cohort of primary node-negative lung adenocarcinomas. Despite all of these patients being pathologically classed

Authors' Affiliations: [1]Department of Thoracic Medicine, The Prince Charles Hospital, Brisbane, Australia, [2]School of Medicine, University of Queensland, and [3]Human Genetics Laboratory, Queensland Institute of Medical Research, Herston, Australia

Received 10/19/06; revised 2/11/07; accepted 2/28/07.

Grant support: National Health and Medical Research Council (338200) and The Prince Charles Hospital Research Foundation. J.E. Larsen is the recipient of an Edwin Tooth PhD Scholarship of the University of Queensland. K.M. Fong is the recipient of a Queensland Smart State Fellowship. N.K. Hayward is a recipient of a National Health and Medical Research Council Fellowship.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Requests for reprints: Jill E. Larsen, Department of Thoracic Medicine, The Prince Charles Hospital, Brisbane 4032, Australia. Phone: 61-7-3139-4110; Fax: 61-7-3139-4957; E-mail: Jill_E_Larsen@health.qld.gov.au.

doi:10.1158/1078-0432.CCR-06-2525

as stage I/II by the tumor-node-metastasis (TNM) staging system of Mountain (4), they displayed considerable prognostic heterogeneity. We aimed to identify a clinically useful classification signature that could predict the likelihood of tumor recurrence. Such a signature could ultimately aid clinicians in making treatment decisions, for example, in the selection of patients most likely to benefit from adjuvant chemotherapy.

Materials and Methods

Most of the methodology used in this study has been described previously (9). Brief outlines are given below, but detailed information can be obtained from cited article (9) or accompanying supporting material.

Sample collection and selection. Fresh-frozen primary lung tumor tissue specimens were collected from consecutive patients undergoing curative surgical resection at The Prince Charles Hospital between 1990 and 2004. Study inclusion criteria were as follows: primary NSCLC of the adenocarcinoma histologic subtype (mixed histology excluded); no nodal metastases at surgery (pathologically N0 stage); tumor H&E examination showed at least 50% tumor cells; surgical bronchial margins were free of disease, and resection was considered complete; no neo- and/or adjuvant radiation or chemotherapy; and fitted to one of our two disease recurrence outcome criteria: nonrecurrence, clinically disease-free for at least 36 months following surgery; or recurrent disease, unambiguous clinical, imaging, or histopathologic evidence of recurrence of the original primary lung cancer in a local or distant metastatic site occurring between 3 and 18 months postresection. The threshold of 36 months for nonrecurrence cases was selected because the majority of patients develop disease recurrence within this period of time (23) and to allow for comparison with other similarly designed studies (22).

KRAS *mutation assay.* Genomic DNA was isolated from normal lung tissue as described previously (24), and the mutation status of codon 12 and codon 13 of *KRAS* was determined using RFLP analysis and sequencing as previously described (25, 26).

Microarray analysis. Total RNA was extracted from each tumor sample and compared with Universal Human Reference RNA (Stratagene) on a commercially available 22K Human Oligo Microarray printed by the British Columbia Gene Array Facility[4] using the Operon Human Genome Oligo Set v2.0[5] containing 21,329 70-mer probes representing ~14,200 named transcripts. Microarray experiments conformed to the Minimum Information about a Microarray Experiment guidelines.[6] Arrays were scanned and quantified, and data were normalized and filtered on probe signal and quality. Of the 21,329 probes present on the array, 18,250 passed filtering criteria. The data discussed in this publication have been made available in the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) public repository[7] and are accessible through GEO Series accession number GSE5843.

Validation of our recurrence signature was done in two independent test sets using publicly available microarray data (refs. 10, 27; Table 1). Samples were censored if they had <3 years follow-up or had distant metastases at surgical resection (TNM1). The first validation test set[8] (10) comprised 55 adenocarcinomas (40 TNM stage I, 10 stage II, and 5 stage III), and the second validation test set (ref. 27; GEO Series accession number GSE31514) comprised 40 adenocarcinomas (30 TNM stage I, 7 stage II, and 3 stage III). ANOVA was done to normalize experimental differences between our data set and the test data sets because the latter were from external sources on different microarray platforms (Affymetrix U95A, ref. 10; and Affymetrix U133Plus 2.0, ref. 27, respectively).

Statistical analysis. Cox proportional hazards modeling (28) with leave-one-out cross-validation (LOOCV) was used to identify genes in which expression levels correlated with tumor recurrence (BRB ArrayTools Version 3.5; developed by Dr. Richard Simon and Amy Peng Lam[9]). In our training set of 48 adenocarcinoma samples, we used the end point of time to recurrence, defined as the time from surgery to tumor recurrence (local, regional, or distant) and applied LOOCV to ensure the selection of robust genes (and not biased by single samples). Briefly, 48 iterations of Cox modeling were done so that each sample was left out once with the significance of each gene in relation to time to recurrence calculated at each iteration. *P* values for each gene were then averaged and ranked to identify genes that consistently, and robustly, correlated with time to recurrence. Recurrence of adenocarcinomas following surgery was univariately associated with *KRAS* mutation status and tumor (TNM) stage. Therefore, *KRAS* mutation status and tumor stage were included as covariates in the development of the predictor to ensure selection of genes in which expression adds to the prediction of tumor recurrence by these two factors. We selected genes that met set criteria (average $P < 0.01$) to identify the 54 genes that comprise our recurrence signature.

The signature was validated in two independent test sets of 55 and 40 adenocarcinomas, respectively. Cox proportional hazards modeling was used to classify patients in each of the data sets as likely or not to develop recurrence. A percentile risk index for each patient was calculated based on the expression levels of the signature. High risk (of recurrence) was defined as above 50%. Kaplan-Meier survival plots and log-rank tests done in SPSS Version 11.5 (SPSS Inc.) were used to assess the differences in survival of the predicted good- and poor-outcome groups.

Hierarchical clustering was done using the average linkage method, centering on genes with bootstrapping of 1,000 iterations (BRB ArrayTools Version 3.5). Distributions of clinical and pathologic parameters were analyzed using χ^2, *t* test, or log-rank tests as appropriate.

Results

Tumor characteristics of the training set. To identify a gene expression signature of tumor recurrence, a training set of 48 pathologically staged N0 lung adenocarcinomas was analyzed: 16 with stage IA (T1N0M0) disease, 30 with stage IB (T2N0M0), and 2 with stage IIB (T3N0M0). The demographics of the patients and tumors in the training and test sets are outlined in Table 1 with detailed information for each patient in the training set given in Supplementary Table S1. All patients had a minimum follow-up of 60 months or until death. As expected, stratification by conventional clinical and pathologic factors identified patient groups with significant (or near-significant) differences in outcomes (Kaplan-Meier analyses, Supplementary Fig. S1). Earliest stage IA tumors had a significantly better outcome in terms of overall survival and time to recurrence when compared with stages IB and IIB tumors combined (log-rank $P = 0.0283$ and 0.0029, respectively; Supplementary Fig. S1C and D, respectively). Additionally, tumors without pleural invasion had a longer time to recurrence (log-rank $P = 0.0202$; Supplementary Fig. S1H). *KRAS* mutation (codon 12 or 13) was significantly associated with shorter time to recurrence (log-rank $P = 0.01$; Supplementary Fig. S1J), but not with overall survival (Supplementary Fig. S1I). Most survival

[4] http://prostatelab.org/arraycentre/index.html
[5] http://www.operon.com
[6] http://www.mged.org/Workgroups/MIAME/miame_checklist.html
[7] http://www.ncbi.nlm.nih.gov/geo/
[8] http://www.broad.mit.edu/mpr/lung/
[9] http://linus.nci.nih.gov/~brb/tool.htm

Table 1. Clinical, pathologic, and prognostic characteristics of adenocarcinoma cases in training and test sets

	Training set		Independent test sets	
	All tumors	**No recurrence/ recurrence**	**Bhattacharjee et al. (10)**	**Bild et al. (27)**
Patient demographics				
Total number of samples	48	25/23	55	40
Age, median	64	64/64	63	67
Sex, *n* (%)				
Male	34 (71)	18/16	23 (42)	23 (58)
Female	14 (29)	7/7	32 (58)	17 (43)
Smoking status, *n* (%)				
Never	1 (2)	1/0	NA	NA
Former	10 (21)	4/6	NA	NA
Current	37 (77)	20/17	NA	NA
Pack years, median	45	41/51	45	NA
KRAS mutation, *n* (%)*	20 (42)	6/14		
Codon 12 mutant	15 (31)	5/10	12 (22)	NA
Codon 13 mutant	5 (10)	1/4	NA	NA
Primary tumor				
TNM stage, *n* (%) †				
IA T1N0M0	16 (33)	13/3	18 (33)	20 (50)
IB T2N0M0	30 (63)	12/18	22 (41)	10 (25)
IIA T1N1M0	0 (0)	0/0	3 (5)	2 (5)
IIB				
T3N0M0	2 (4)	0/2	3 (5)	5 (13) ‡
T2N1M0	0 (0)	0/0	4 (7)	
IIIA				
T1N2M0	0 (0)	0/0	2 (4)	3 (7) ‡
T2N2M0	0 (0)	0/0	3 (5)	
Maximum tumor measurement (mm), median	33	30/35	36	
Differentiation, *n* (%)				
P	19 (40)	10/9	24 (44)	NA
M	15 (31)	6/9	27 (49)	NA
W	14 (29)	9/5	4 (7)	NA
Tumor invasion, *n* (%)				
Lymphatic	9 (19)	4/5	NA	NA
Vascular	19 (40)	9/10	NA	NA
Pleural	18 (38)	6/12	NA	NA
Clinical investigation, *n* (%)				
Upper Abdominal CT scan	48 (100)	25/23	NA	NA
Bone scan	31 (65)	16/15	NA	NA
Chest CT scan	48 (100)	25/23	NA	NA
Head CT scan	28 (58)	15/13	NA	NA
Positron emission tomography scan	7 (15)	4/3	NA	NA
Prognostic parameters				
Recurrence site, *n* (%)				
No evidence of disease (NED)	22 (46)	22/0	22 (31)	NA
Lung, primary site	2 (4)	1/1	0 (0)	NA
Lymph node (loco-regional)	4 (8)	0/4	8 (11)	NA
Lung, distant	6 (13)	1/5	16 (23)	NA
Bone	3 (6)	0/3	5 (7)	NA
Brain	4 (8)	1/5	10 (14)	NA
Liver	2 (4)	0/2	4 (6)	NA
Adrenal	1 (2)	0/1	1 (1)	NA
Chest wall	1 (2)	0/1	4 (6)	NA
Pancreas	1 (2)	0/1	1 (1)	NA
Survival Status, *n* (%) §				
Alive	26 (54)	3/23	25 (45)	17 (43)
Dead	22 (46)	20/2	30 (55)	23 (57)

NOTE: Univariate analyses were done to identify significant associations between parameters and recurrence phenotype in training set (χ^2, *t* test, or log-rank).
Abbreviation: NA, not available.
* *P* value = 0.010.
† *P* value = 0.039.
‡ TNM staging unknown, tumor stage data available only.
§ *P* value < 0.001.

curves converge after 5 years, likely reflecting the competing risk of comorbid diseases in an older patient population (mean age, 63 ± 10 years). Age, sex, smoking history, tumor size, differentiation, and tumor invasion (lymphatic, vascular, or perineural) were not associated with time to recurrence or overall survival (Table 1).

Identification of adenocarcinoma subsets. To determine if any clinical or biological subsets existed in our training set of 48 adenocarcinoma samples due to gene expression profiles, we firstly did unsupervised hierarchical clustering. A final filtered gene list of 7,321 probes was used after excluding probes with low log-ratio variation ($P > 0.01$) from the initial filtered list of 18,250 probes (filtered on microarray quality). Adenocarcinomas clustered into two distinct groups of 22 and 26 samples (Fig. 1), with a robustness index of 0.80 over 100 permutations, indicating high reproducibility. The clusters differed significantly in TNM stage (stage IA versus IB/IIB; $P = 0.041$, χ^2) and *KRAS* mutation status ($P = 0.008$, χ^2), with stage IB/IIB tumors and mutant *KRAS* tumors being more prevalent in cluster 1. The overrepresentation of these two prognostic indicators in cluster 1 may contribute to the nonsignificant trend toward shorter overall survival and time to recurrence observed for cluster 1 in Kaplan-Meier analysis ($P > 0.05$; Fig. 2A and B, respectively). No significant association was identified between the two clusters and recurrence phenotype, recurrence site (none, local, distant), tumor size, differentiation, smoking status and history, sex, or age.

A supervised analysis of the two adenocarcinoma subsets identified by unsupervised clustering was done to identify those genes significantly differentially expressed between the two clusters and, therefore, most likely contributing to subset formation. This identified 1,455 probes differentially expressed between clusters 1 and 2 ($P < 0.05$; *t* test). Gene ontology analysis was done to determine if the ontologies represented by these 1,455 genes were significantly different from random selections of 1,455 genes on the array that passed the filters described above. Supplementary Table S2 lists the significantly overrepresented ontologies, with a ratio of observed genes/expected genes greater than 2-fold. In addition to many cellular metabolism and energy ontologies, the Ras protein signal transduction ontology was overrepresented, with six members, *RAB6A*, *RASSF2*, *G3BP*, *RASGRP1*, and *RASGRP4*, significantly differentially expressed.

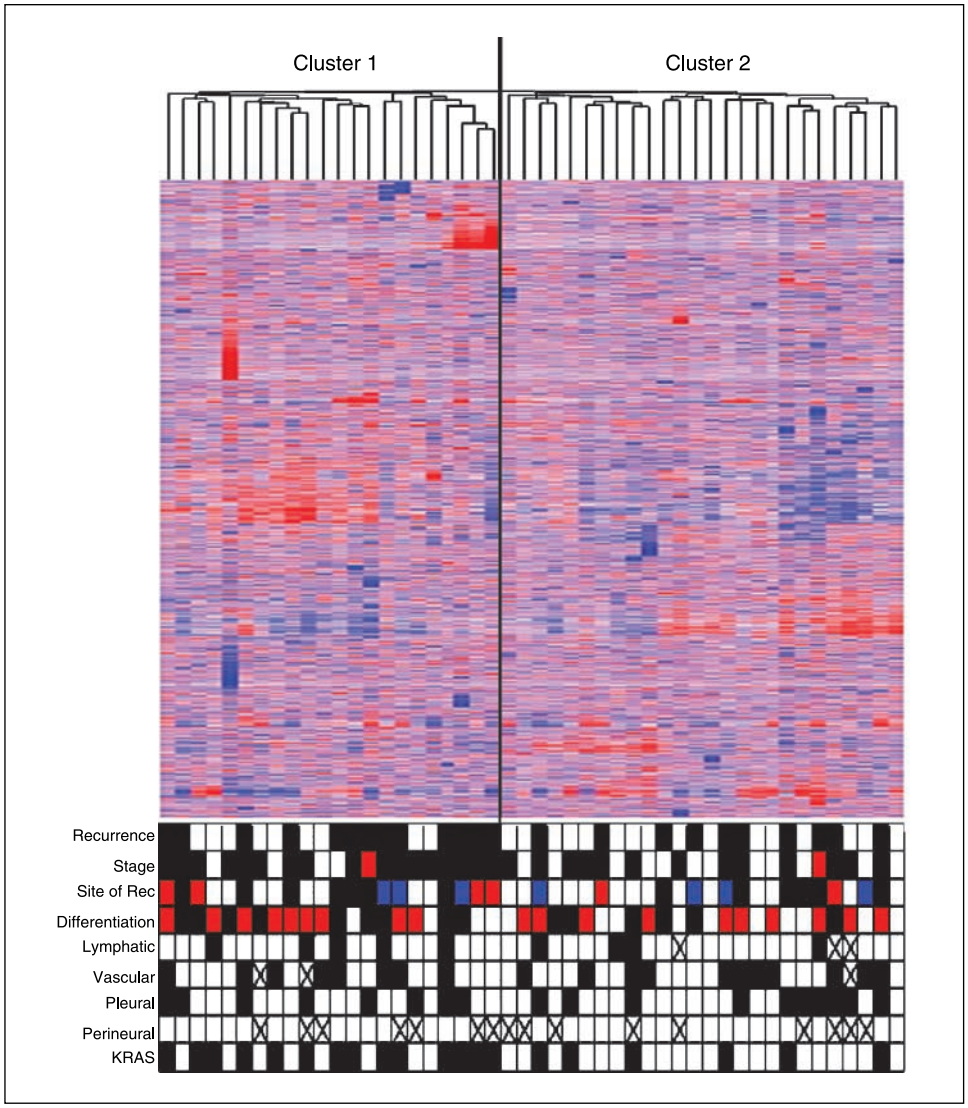

Fig. 1. Unsupervised analysis of the training set of 48 adenocarcinoma samples identifies two clinically relevant subsets of adenocarcinoma. Unsupervised hierarchical two-dimensional clustering of 48 adenocarcinomas (using Pearson correlation with 1,000 bootstrap iterations using a filtered list of 7,321 probes) identified two distinct clusters of 22 and 26 samples. Each column is a sample, and each row is a gene. Heat map indicates level of gene expression; red, high expression; blue, low expression. Prognostic parameters color-coded beneath heat map; recurrence (*white*, no recurrence; *black*, recurrence); TNM stage (*white*, stage I; *black*, stage II; *red*, stage III); site of recurrence (*white*, none; *black*, distant; *red*, local); Differentiation (*white*, well; *black*, moderate; *red*, poor); tumor invasion (lymphatic, vascular, pleural, perineural; *white*, no; *black*, yes; *X*, not determined); *KRAS* mutation status (*white*, wild type; *black*, mutant).

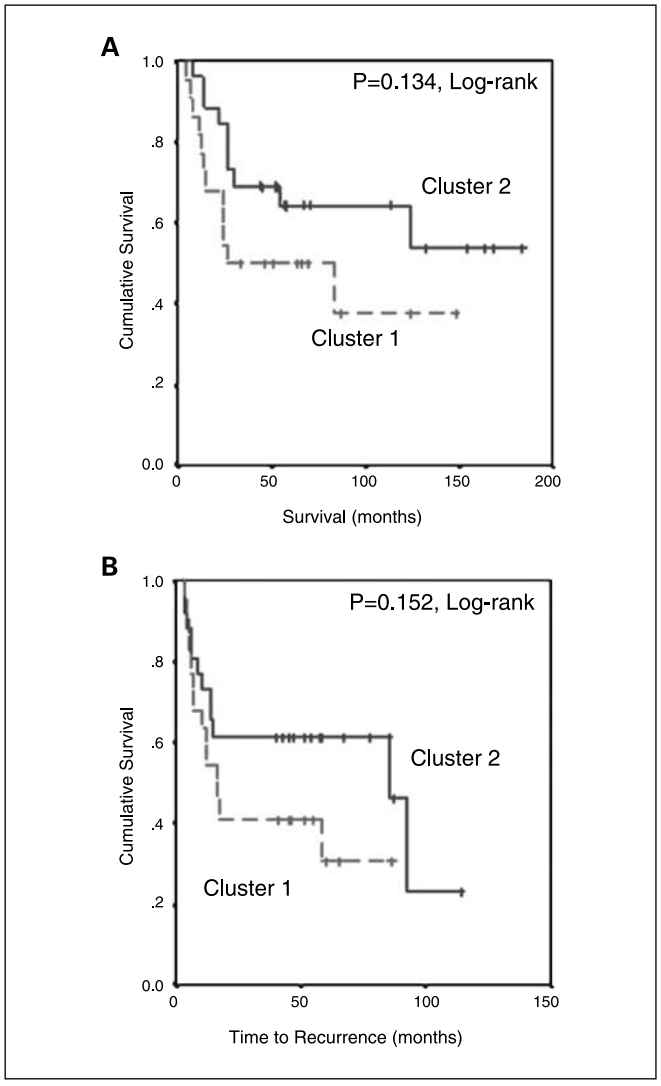

Fig. 2. *A,* Kaplan-Meier analysis of the training set of 48 adenocarcinoma samples. *B,* Kaplan-Meier analysis (log-rank) of unsupervised clusters in relation to time to recurrence. Tick marks, patients whose data were censored at last follow-up.

Identification of adenocarcinoma recurrence signature in training set. Next, to identify a manageable, robust set of genes in which expression could predict primary tumors likely to develop recurrent disease, we employed Cox proportional hazard modeling (28) with LOOCV. Using the filtered set of 7,321 probes in our training set of 48 adenocarcinoma samples, we selected the top 54 genes (genes with an average significance level <0.01) in which expression was significantly associated with the time to recurrence (Supplementary Table S3). Two factors significantly associated with disease recurrence in this cohort (*KRAS* mutation status and tumor stage) were included as covariates in the development of the predictor to ensure the selection of genes in which expression added further to the prediction of recurrent disease by these two variables alone. Gene ontology analysis of the 54 genes found overrepresentation of several recurrence-associated biological processes, including angiogenesis, apoptosis, cell cycle, cell motility and adhesion, and signal transduction and cell communication.

Hierarchical two-dimensional clustering analysis was done on the 54 genes to examine intrinsic gene expression patterns in the training set of 48 adenocarcinomas (Fig. 3A). The clustering identified two distinct categories of genes in the 54-gene signature: those in which expression correlated with disease recurrence and those in which expression correlated with disease-free survival (no recurrence). The 54 genes separated adenocarcinomas into two clusters of tumors with significantly different times to recurrence by Kaplan-Meier (log-rank $P = 0.0001$) analysis (Fig. 3B).

Validation of adenocarcinoma recurrence signature in independent test sets. To determine if the 54-gene signature could predict patients likely to develop tumor recurrence in independent samples, we applied it to two publicly available adenocarcinoma data sets, comprising 55 stages I to III adenocarcinomas (10) and 40 stages I to III adenocarcinomas (27) after exclusion of cases with <3 years follow-up. The Bhattacharjee data set included recurrence annotation where patients were classified as with or without recurrence or, in some cases, undetermined. In the absence of recurrence annotation in the Bild data set, overall survival was used as the clinical end point based on the strong correlation between time to recurrence and survival in lung cancer patients. A percentile risk index for each patient was calculated based on the expression levels of the 54-gene signature: poor outcome was defined as risk >50%, and good outcome was defined as risk ≤50%. Cox proportional hazards modeling was used to classify patients in each test data set. The predictive accuracy of the recurrence signature was determined against patient status 3 years after surgery: recurrence/no recurrence (Bhattacharjee data set) or dead/alive (Bild data set).

A total of 32 (59%) of the 54 genes in the recurrence signature were represented on the early-generation U95A microarray (Affymetrix) used in the Bhattacharjee test set. Despite incomplete representation, the signature had an overall accuracy of 69% (79% sensitivity, 59% specificity) in predicting recurrence. Kaplan-Meier log-rank analysis confirmed that the predicted poor-outcome group had a significantly shorter time to recurrence [log-rank $P = 0.039$; hazard ratio (HR), 2.20; 95% confidence interval (95% CI), 1.02-4.73; Fig. 4A]. To determine independence of the prediction from clinical or biological factors, we adjusted for TNM stage, N stage, and *KRAS* mutation status in a Cox regression model. The signature remained significant for recurrence prediction after adjustment for all factors [TNM stage (Wald $P = 0.028$; HR, 2.93; 95% CI, 1.12-7.65), N stage (Wald $P = 0.055$; HR, 2.13; 95% CI, 0.98-4.60), and *KRAS* mutation (Wald $P = 0.016$; HR, 2.96; 95% CI, 1.22-6.99)].

In the second validation data set (Bild), 47 (87%) of the 54 genes in the signature were represented in the U133Plus 2.0 microarray (Affymetrix). The signature had an overall accuracy of 71% (65% sensitivity, 77% specificity). Improved accuracy in this data set may be due to the increased gene representation of the 54-gene signature. Predicted poor-outcome patients had a significantly worse overall survival (log-rank $P = 0.004$; HR, 3.30; 95% CI, 1.39-7.86; Fig. 4C) on log-rank analysis, and this was independent of overall TNM stage (Wald $P = 0.010$; HR, 3.50; 95% CI, 1.35-9.05) in Cox regression modeling. N stage and *KRAS* data were not available for adjustment.

Prediction of recurrence in earliest stage adenocarcinoma. A current limitation of clinical prognostic indicators is their

inability to predict those patients with early-stage disease who will unexpectedly develop disease recurrence. Therefore, we applied our recurrence signature to stage I adenocarcinomas in each of the two validation data sets to ascertain whether it has the potential to improve upon current prediction methods. In the Bhattacharjee cohort subset of stage I tumors (n = 40), the signature had an overall accuracy of 72% (73% sensitivity, 72% specificity): predicted high recurrence risk was associated with shorter observed time to recurrence (log-rank P = 0.021; HR, 3.47; 95% CI, 1.13-10.66; Fig. 4B), and the signature's performance

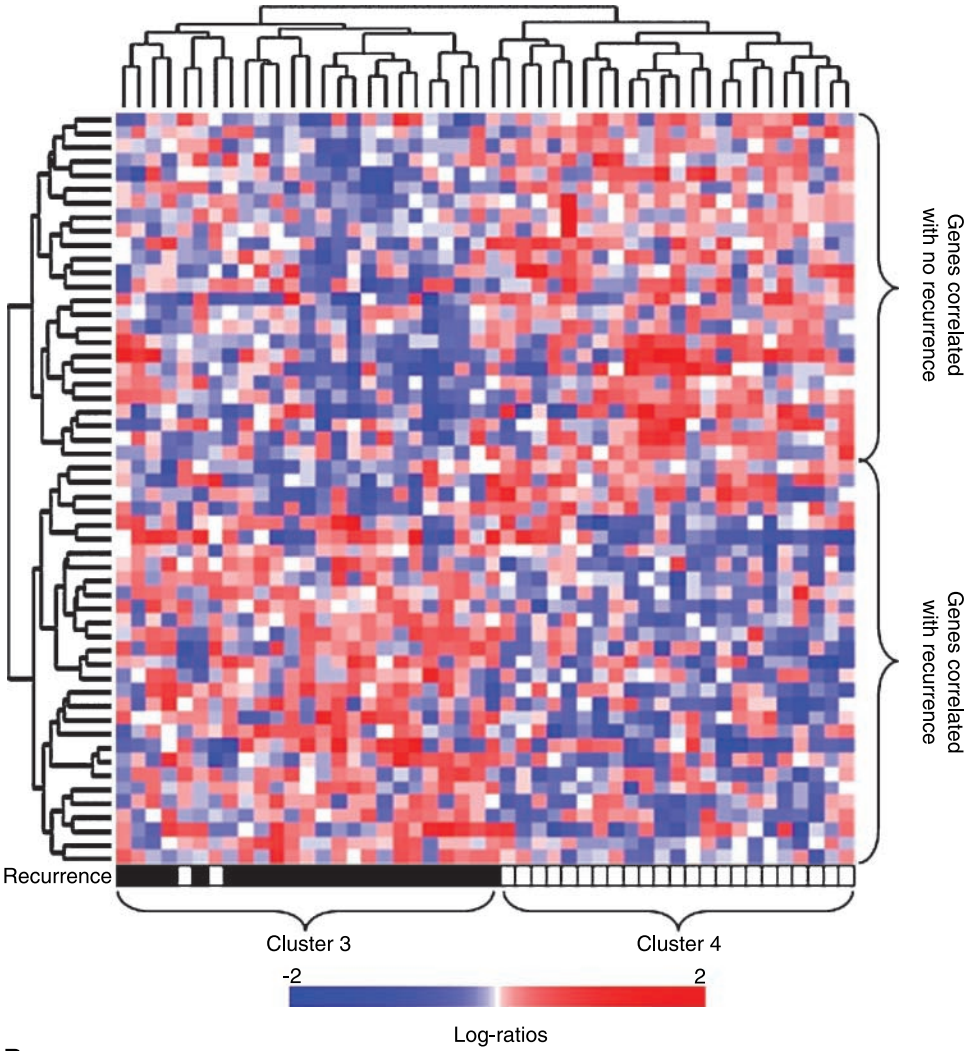

Fig. 3. Kaplan-Meier analysis of the training set of 48 adenocarcinoma samples. *A,* unsupervised hierarchical two-dimensional clustering of 48 adenocarcinoma samples in training set based on gene expression of 54 genes in recurrence signature. Each column is a sample, and each row is a gene. Heat map indicates level of gene expression; red, high expression; blue, low expression. Recurrence phenotype of samples is indicated underneath the heat map; white, no recurrence, black, recurrence. Hierarchical clustering was used to identify genes and samples with similar expression patterns. *B,* Kaplan-Meier (log-rank) analysis of two clusters of tumors identified by hierarchical clustering in terms of time to recurrence. Time is indicated in months on the x-axis, cumulative survival is indicated on the y-axis. Tick marks, patients whose data were censored at last follow-up.

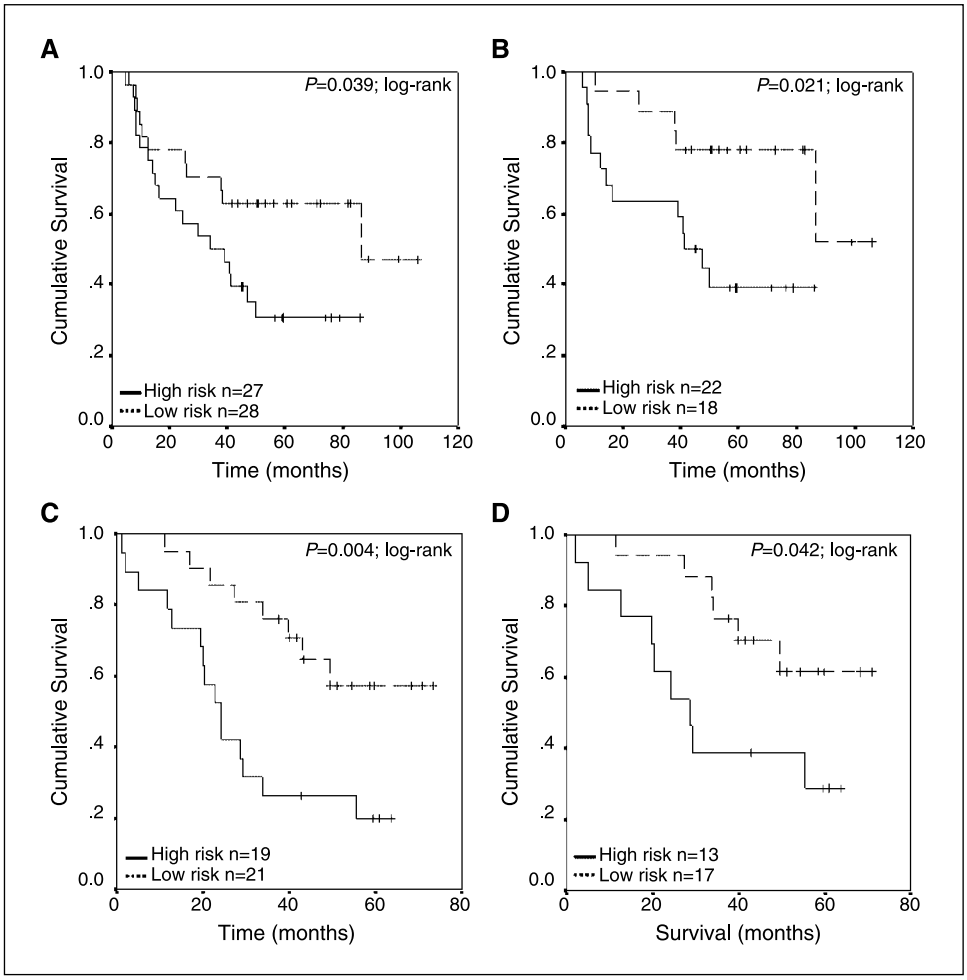

Fig. 4. Validation of the 54-gene classifier in two independent test sets of 55 adenocarcinomas and 40 adenocarcinomas. Comparison of Kaplan-Meier survival estimates (log-rank test) for predicted high (*full red line*) and low (*dashed blue line*) and high-risk patient groups using the 54-gene classifier in two validation data sets. Survival estimates in the Bhattacharjee test set samples in predicted high-and low-risk – outcome groups in terms of time to recurrence in all 55 samples (stages I-III; *A*) and 43 N0 stage samples (*B*). Survival estimates of the Bild test set samples comparing predicted high- and risk-outcome groups in terms of overall survival in all 40 (stages I-III; *C*) and 30 stage I samples (*D*). Tick marks, patients whose data were censored at last follow-up.

was independent of *KRAS* mutation status (Wald $P = 0.012$; HR, 5.88; 95% CI, 1.49-23.23) in the Cox regression analysis.

In the Bild cohort subset of stage I tumors ($n = 30$), the signature's predictive accuracy was 67% (60% sensitivity, 74% specificity), where predicted poor-outcome samples had significantly worse overall survival (log-rank $P = 0.042$; HR, 2.82; 95% CI, 1.01-8.03; Fig. 4D).

Discussion

We previously described a 111-gene signature of recurrence in lung SCCs capable of predicting outcome in independent samples (9). In this study, we sought to determine if a comparable signature existed in primary adenocarcinoma, the other major histologic subtype in lung cancer. The samples were pathologic stage N0, obtained from patients treated by curative intent surgical resection, and stratified for tumor recurrence.

To identify a gene expression signature in the primary tumor that could predict recurrence, we used Cox proportional modeling to identify a robust set of 54 genes in which expression correlated with disease recurrence. As expected, the gene ontology composition of the 54 genes has biological relevance to disease recurrence, such as angiogenesis, apoptosis, cell cycle and cell motility and adhesion. A potential issue in developing predictive signatures is overfitting to the training data set, resulting in a signature that reflects the characteristics of the training set samples and cannot accurately predict outcome in independent samples. Consequently, a critical test of prediction signatures is validation in independent data sets. We therefore used two independent data sets to validate our recurrence signature, and in both sets, the signature was an independent predictor of outcome, particularly in the earliest stage adenocarcinomas, which are expected to benefit most from curative resection.

Several studies have now used expression profiling to characterize prognosis in lung cancer (10–15, 17–22), where survival is the main outcome indicator rather than tumor recurrence. Although tumor recurrence and overall survival are closely related, it is not uncommon for some patients with localized recurrence to benefit from further surgery or radical treatment resulting in relatively prolonged survival. Furthermore, the overall survival (rather than disease-specific survival) end point is subject to influence by competing risks due to comorbidities. We considered prediction of tumor recurrence to be a superior end point to overall survival for our purpose of developing a gene expression signature to enable the identification of at-risk individuals who would perhaps benefit most from adjuvant treatment. In addition to our SCC study, three other studies have also used gene expression data to identify recurrence profiles (15, 18, 21). Two of these were unstratified for histopathologic subtype (15, 21), and the other used a small SCC cohort (18). As prognosis may differ

between the histologically and molecularly distinct SCC and adenocarcinoma subtypes, it is possible that generic mixed adenocarcinoma/SCC gene signatures may partly reflect gene expression differences between the subtypes of lung cancer (10, 11, 29). Consequently, we chose to exclusively analyze a cohort of histologically homogeneous adenocarcinomas to avoid potential confounding.

Several factors may have limited the fidelity of the recurrence signature in the independent test samples. The first is the lack of perfect correlation of transcripts on differing platforms, which was more pronounced in the Bhattacharjee data set where 22 genes were absent, whereas only 7 were absent from the Bild data set. This difference in gene representation between the two test data sets may be a contributor to the increased accuracy of the 54-gene signature in the Bild data set. Our prediction method for validation used Cox proportional hazards modeling, where survival risk groups are constructed using the supervised principal component method (30). This reduces the signature to k "super-gene" expression levels (or principal components). Theoretically, despite absent genes in the independent data sets, the predictive ability of the gene signature can be maintained, provided that each principal component is sufficiently represented.

Another difficulty in our validation was the absence of recurrence annotation in the Bild test set, where we instead used survival data as a surrogate for recurrence. Although significantly different outcome groups were validated, we made an assumption that patients with poor survival developed disease recurrence, and those with good survival did not—which may not always be the case. Nevertheless, in independent samples, the signature predicted samples with significantly shorter outcome (either time to recurrence or overall survival), which was independent of and superior to (in stage I adenocarcinomas) conventional prognostic factors.

In comparing the 54-gene adenocarcinoma recurrence signature to our previously reported 111-gene SCC recurrence signature (9), no genes were common to both signatures, which was similarly reported in another study (22), in which a 50-gene SCC survival signature had no overlap with the previously reported 50-gene adenocarcinoma survival signature (12). This may be due, in part, to histologic differences; moreover, as a classifier or signature is a function that can transform the expression levels for a set of genes to a risk score or predicted class (31), lack of commonality in genes comprising published prognostic signatures for lung cancer to date may simply reflect the fact that several gene expression signatures are capable of predicting outcome in NSCLC, as recently shown in breast cancer (32–34). On the other hand, certain gene ontologies were shared between both adenocarcinoma and SCC recurrence signatures (Supplementary Fig. S2), including cell communication and signal transduction, as well as cell growth and movement, suggesting that gene expression signatures that predict the same outcome in the same disease need not comprise similar genes but rather similar biological processes or pathways.

In addition to comparing our 54-gene signature to our previously described 111-gene SCC signature, we also compared it to genes present in five NSCLC signatures that have been previously reported to predict either recurrence or overall survival (12, 18, 21, 22) with varying degrees of accuracy. Of the 260 genes that comprise these five published signatures, only three were common to more than one signature. Similarly, only 1 gene in our 54-gene signature was in common with these five published signatures. Lack of overlap in the composition of NSCLC prognostic signatures is not unexpected and agrees with a similar comparison of prognostic signatures in breast cancer (33) and likely reflects the fact that, like breast cancer, numerous gene expression signatures may be capable of predicting outcome in NSCLCs. This analysis confirms the promise of gene expression signatures to refine outcome prediction for lung cancer patients.

A recent approach taken by Raponi et al. (22) was to combine two individually developed survival signatures for adenocarcinoma and SCC to form a NSCLC signature, which had good predictive ability in mixed cohorts. When we combined our SCC recurrence signature (9) with our adenocarcinoma recurrence signature, however, we were not able to significantly predict recurrence or overall survival in a mixed cohort of 86 stages I to III adenocarcinomas and SCCs or a subset of 63 stage I tumors (ref. 27; Supplementary Fig. S3). This disparity may be due to chance, sample size, gene representation on the validation microarrays, or a reflection on the signatures themselves, indicating that further validation is required to determine if there truly exists a global NSCLC signature. Our recurrence signatures seem to be subtype specific, which supports substantial existing evidence of biological heterogeneity between subtypes, including gene expression profiles (10, 11, 29, 35), gene mutations (*KRAS* and *EGFR*; refs. 36, 37), and clinical outcome (23, 38).

Unsupervised hierarchical clustering of the 48 adenocarcinomas classified samples into subgroups that correlated with tumor stage and *KRAS* mutation status. It is possible that the higher frequency of higher stage tumors and mutant *KRAS* tumors in cluster 1 may have contributed to the observed trend of poorer outcome because both factors have been associated with poor survival in lung cancer (3–5, 36, 39–43). The unsupervised clustering indicates that the data set exhibited gene expression profiles for tumor stage and *KRAS* mutation status. Additionally, these factors were both significantly associated with the recurrence phenotype. Therefore, the potential confounding effects of these two factors were adjusted for in both the development of the recurrence signature and testing it in independent test sets of tumor samples. This ensured that the signature reflected gene expression changes in relation to the likelihood of tumor recurrence rather than other clinical and biological factors.

The Ras proteins are pivotal regulators of cellular proliferation, differentiation, motility, and apoptosis, with mutations in *KRAS* occurring in 20% to 30% of NSCLCs, more commonly in adenocarcinomas (44). *KRAS* mutations were a significant predictor of tumor recurrence in 244 NSCLCs (45), but not in two smaller studies (40, 46). A recent meta-analysis of 28 studies in NSCLC revealed a significant association between *KRAS* mutation and poorer survival in adenocarcinomas, but not SCCs (43). In this cohort, mutation status did not correlate with overall survival, but did correlate with time to recurrence.

In this study, we have shown the value of a genomic approach to identifying patients likely to develop tumor recurrence by characterizing an expression signature and validating it in two independent adenocarcinoma data sets derived from different microarray platforms. Together with other studies (12, 20–22),

these results strongly indicate the potential that gene expression signatures have to refine the prediction of recurrence and survival in early-stage lung cancers. If patients predicted to be at high risk of recurrent disease by genomic signatures are shown in clinical trials to be those who benefit most from adjuvant treatment, as is currently occurring in breast cancer (47), the clinical pay-off for genomic tumor analyses will have been realized.

Acknowledgments

The authors thank Drs. Belinda Clarke and Edwina Duhig and staff in Anatomical Pathology at The Prince Charles Hospital for aiding in collection and pathologic assessment of samples, staff in the Thoracic Research Laboratory at The Prince Charles Hospital for aid in specimen collection, Dr. Maree L. Colosimo for clinical data, and Dr. Ian A. Yang for general discussions on study design and analysis and comments to the manuscript.

References

1. Mathers C, Vos T, Stevenson C, Australian Institute of Health and Welfare, The burden of disease and injury in Australia: summary report. Canberra: Australian Institute of Health and Welfare; 1999. p 31.

2. Fidler IJ. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. Nat Rev Cancer 2003;3:453–8.

3. Martini N, Bains MS, Burt ME, et al. Incidence of local recurrence and second primary tumors in resected stage I lung cancer. J Thorac Cardiovasc Surg 1995; 109:120–9.

4. Mountain CF. Revisions in the international system for staging lung cancer. Chest 1997;111:1710–7.

5. Hoffman PC, Mauer AM, Vokes EE. Lung cancer. Lancet 2000;355:479–85.

6. Flehinger BJ, Kimmel M, Melamed MR. The effect of surgical treatment on survival from early lung cancer. Implications for screening. Chest 1992;101:1013–8.

7. Strauss GM, Kwiatkowski DJ, Harpole DH, et al. Molecular and pathologic markers in stage I non – small-cell carcinoma of the lung. J Clin Oncol 1995;13:1265–79.

8. Feinstein MB, Bach PB. Epidemiology of lung cancer. Chest Surg Clin N Am 2000;10:653–61.

9. Larsen JE, Pavey SJ, Passmore LH, et al. Expression profiling defines a recurrence signature in lung squamous cell carcinoma. Carcinogenesis 2006; November 1. Epub ahead of print.

10. 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 U S A 2001;98: 13790–5.

11. Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 2001;98:13784–9.

12. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24.

13. Miura K, Bowman ED, Simon R, et al. Laser capture microdissection and microarray expression analysis of lung adenocarcinoma reveals tobacco smoking- and prognosis-related molecular profiles. Cancer Res 2002;62:3244–50.

14. Moran CJ, Arenberg DA, Huang CC, et al. RANTES expression is a predictor of survival in stage I lung adenocarcinoma. Clin Cancer Res 2002;8:3803–12.

15. Wigle DA, Jurisica I, Radulovich N, et al. Molecular profiling of non – small cell lung cancer and correlation with disease-free survival. Cancer Res 2002;62: 3005–8.

16. Chen G, Gharib TG, Wang H, et al. Protein profiles associated with survival in lung adenocarcinoma. Proc Natl Acad Sci U S A 2003;100:13537–42.

17. Gordon GJ, Richards WG, Sugarbaker DJ, Jaklitsch MT, Bueno R. A prognostic test for adenocarcinoma of the lung from gene expression profiling data. Cancer Epidemiol Biomarkers Prev 2003;12:905–10.

18. Sun Z, Yang P, Aubry MC, et al. Can gene expression profiling predict survival for patients with squamous cell carcinoma of the lung? Mol Cancer 2004;3:35–44.

19. Inamura K, Fujiwara T, Hoshida Y, et al. Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. Oncogene 2005;24:7105–13.

20. Guo L, Ma Y, Ward R, et al. Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. Clin Cancer Res 2006;12:3344–54.

21. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006;66: 7466–72.

22. Potti A, Mukherjee S, Petersen R, et al. A genomic strategy to refine prognosis in early-stage non – small-cell lung cancer. N Engl J Med 2006;355:570–80.

23. Hawson G, Zimmerman PV, Ford CA, Johnston NG, Firouz-Abadi A. Primary lung cancer: characterization and survival of 1024 patients treated in a single institution. Med J Aust 1990;152:230–4.

24. Larsen JE, Colosimo ML, Yang IA, et al. Risk of non – small cell lung cancer and the cytochrome P4501A1 $Ile^{462}Val$ polymorphism. Cancer Causes Control 2005;16:579–85.

25. Fong KM, Zimmerman PV, Smith PJ. KRAS codon 12 mutations in Australian non – small cell lung cancer. Aust N Z J Med 1998;28:184–9.

26. Hatzaki A, Razi E, Anagnostopoulou K, et al. A modified mutagenic PCR-RFLP method for K-ras codon 12 and 13 mutations detection in NSCLC patients. Mol Cell Probes 2001;15:243–7.

27. Bild AH, Yao G, Chang JT, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7.

28. Cox DR. Regression models and life-tables (with discussion). J R Statist Soc B 1972;34:187–220.

29. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001;98:15149–54.

30. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2004;2:511–22.

31. Simon R. Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. J Natl Cancer Inst 2006;98:1169–71.

32. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98:1183–92.

33. Fan C, Oh DS, Wessels L, et al. Concordance among gene-expression – based predictors for breast cancer. N Engl J Med 2006;355:560–9.

34. Naderi A, Teschendorff AE, Barbosa-Morais NL, et al. A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene 2007;26:1507–16.

35. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286:531–7.

36. Aviel-Ronen S, Blackhall FH, Shepherd FA, Tsao MS. K-ras mutations in non – small-cell lung carcinoma: a review. Clin Lung Cancer 2006;8:30–8.

37. Mitsudomi T, Kosaka T, Yatabe Y. Biological and clinical implications of EGFR mutations in lung cancer. Int J Clin Oncol 2006;11:190–8.

38. Beadsmoore CJ, Screaton NJ. Classification, staging and prognosis of lung cancer. Eur J Radiol 2003; 45:8–17.

39. Rosell R, Monzo M, Pifarre A, et al. Molecular staging of non – small cell lung cancer according to K-ras genotypes. Clin Cancer Res 1996;2:1083–6.

40. Moldvay J, Scheid P, Wild P, et al. Predictive survival markers in patients with surgically resected non – small cell lung carcinoma. Clin Cancer Res 2000;6:1125–34.

41. Fang D, Zhang D, Huang G, et al. Results of surgical resection of patients with primary lung cancer: a retrospective analysis of 1,905 cases. Ann Thorac Surg 2001;72:1155–9.

42. Shimizu K, Yoshida J, Nagai K, et al. Visceral pleural invasion classification in non – small cell lung cancer: a proposal on the basis of outcome assessment. J Thorac Cardiovasc Surg 2004;127:1574–8.

43. Mascaux C, Iannino N, Martin B, et al. The role of RAS oncogene in survival of patients with lung cancer: a systematic review of the literature with meta-analysis. Br J Cancer 2005;92:131–9.

44. Rodenhuis S, Slebos RJ, Boot AJ, et al. Incidence and possible clinical significance of K-ras oncogene activation in adenocarcinoma of the human lung. Cancer Res 1988;48:5738–41.

45. Kwiatkowski DJ, Harpole DH, Jr., Godleski J, et al. Molecular pathologic substaging in 244 stage I non – small-cell lung cancer patients: clinical implications. J Clin Oncol 1998;16:2468–77.

46. Baksh FK, Dacic S, Finkelstein SD, et al. Widespread molecular alterations present in stage I non – small cell lung carcinoma fail to predict tumor recurrence. Mod Pathol 2003;16:28–34.

47. Paik S. Molecular profiling of breast cancer. Curr Opin Obstet Gynecol 2006;18:59–63.