Gene Expression Signature Predicts Recurrence in Lung Adenocarcinoma

Jill E. Larsen,1,2 Sandra J. Pavéy,2,3 Linda H. Passmore,3 Rayleen V. Bowman,2 Nicholas K. Hayward,2,3 and Kwun M. Fong1,2

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 micrometa-

---

Authors’ Affiliations: 1Department of Thoracic Medicine, The Prince Charles Hospital, Brisbane, Australia, 2School of Medicine, University of Queensland, and 3Human 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.

© 2007 American Association for Cancer Research. doi:10.1158/1078-0432.CCR-06-2525
Recurrence Prediction in Lung Adenocarcinoma

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 adjacent 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 using the Operon Human Genome Oligo Set v2.0 containing 21,329 70-mer probes representing 14,200 named transcripts. Microarray experiments conformed to the Minimum Information about a Microarray Experiment guidelines. 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 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 (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). 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 χ2, 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. S1I), but not with overall survival (Supplementary Fig. S1I). Most survival
Table 1. Clinical, pathologic, and prognostic characteristics of adenocarcinoma cases in training and test sets

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

NOTE: Univariate analyses were done to identify significant associations between parameters and recurrence phenotype in training set (χ², 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/II; $P = 0.041$, $\chi^2$) and KRAS mutation status ($P = 0.008$, $\chi^2$), with stage IB/IIIB 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.
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 \( [\text{log-rank} \ P = 0.039; \text{hazard ratio (HR)}, 2.20; 95\% \text{ confidence interval (CI)}, 1.02-4.73; \text{Fig. 4A}] \). 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; \text{HR}, 2.93; 95\% \text{CI}, 1.12-7.65\}), \text{N stage (Wald } P = 0.055; \text{HR}, 2.13; 95\% \text{CI}, 0.98-4.60), \text{and KRAS mutation (Wald } P = 0.016; \text{HR}, 2.96; 95\% \text{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 \( (\text{log-rank} \ P = 0.004; \text{HR}, 3.30; 95\% \text{CI}, 1.39-7.86; \text{Fig. 4C}) \) on log-rank analysis, and this was independent of overall TNM stage \( (\text{Wald } P = 0.010; \text{HR}, 3.50; 95\% \text{CI}, 1.35-9.05) \) 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

![Figure 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.](image-url)
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 Duoh 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


Imaging, Diagnosis, Prognosis

References


Acknowledgments

The authors thank Drs. Belinda Clarke and Edwina Duoh 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


Acknowledgments

The authors thank Drs. Belinda Clarke and Edwina Duoh 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


Acknowledgments

The authors thank Drs. Belinda Clarke and Edwina Duoh 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


Acknowledgments

The authors thank Drs. Belinda Clarke and Edwina Duoh 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


Acknowledgments

The authors thank Drs. Belinda Clarke and Edwina Duoh 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


Acknowledgments

The authors thank Drs. Belinda Clarke and Edwina Duoh 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

Gene Expression Signature Predicts Recurrence in Lung Adenocarcinoma


Updated version
Access the most recent version of this article at:
http://clincancerres.aacrjournals.org/content/13/10/2946

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2007/05/16/13.10.2946.DC1

Cited articles
This article cites 45 articles, 18 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/13/10/2946.full#ref-list-1

Citing articles
This article has been cited by 20 HighWire-hosted articles. Access the articles at:
http://clincancerres.aacrjournals.org/content/13/10/2946.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.