

Gene Expression Signatures for Predicting Prognosis of Squamous Cell and Adenocarcinomas of the Lung

Mitch Raponi,¹ Yi Zhang,¹ Jack Yu,¹ Guoan Chen,² Grace Lee,¹ Jeremy M.G. Taylor,^{3,5} James MacDonald,⁵ Dafydd Thomas,⁴ Christopher Moskaluk,⁶ Yixin Wang,¹ and David G. Beer^{2,4}

¹Veridex, LLC, a Johnson & Johnson company, San Diego, California; Departments of ²Surgery, ³Biostatistics, and ⁴Pathology, and ⁵Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan; and ⁶Department of Pathology, University of Virginia, Charlottesville, Virginia

Abstract

Non-small-cell lung cancers (NSCLC) compose 80% of all lung carcinomas with squamous cell carcinomas (SCC) and adenocarcinoma representing the majority of these tumors. Although patients with early-stage NSCLC typically have a better outcome, 35% to 50% will relapse within 5 years after surgical treatment. We have profiled primary squamous cell lung carcinomas from 129 patients using Affymetrix U133A gene chips. Unsupervised analysis revealed two clusters of SCC that had no correlation with tumor stage but had significantly different overall patient survival ($P = 0.036$). The high-risk cluster was most significantly associated with down-regulation of epidermal development genes. Cox proportional hazard models identified an optimal set of 50 prognostic mRNA transcripts using a 5-fold cross-validation procedure. Quantitative reverse transcription-PCR and immunohistochemistry using tissue microarrays were used to validate individual gene candidates. This signature was tested in an independent set of 36 SCC samples and achieved 84% specificity and 41% sensitivity with an overall predictive accuracy of 68%. Kaplan-Meier analysis showed clear stratification of high-risk and low-risk patients [log-rank $P = 0.04$; hazard ratio (HR), 2.66; 95% confidence interval (95% CI), 1.01-7.05]. Finally, we combined the SCC classifier with our previously identified adenocarcinoma prognostic signature and showed that the combined classifier had a predictive accuracy of 71% in 72 NSCLC samples also showing significant differences in overall survival (log-rank $P = 0.0002$; HR, 3.54; 95% CI, 1.74-7.19). This prognostic signature could be used to identify patients with early-stage high-risk NSCLC who might benefit from adjuvant therapy following surgery. (Cancer Res 2006; 66(15): 7466-72)

Introduction

Lung cancer is the leading cause of cancer deaths in developed countries and accounts for a million deaths each year worldwide. An estimated 171,900 new cases are detected in the United States,

accounting for ~13% of all cancer diagnoses. Non-small-cell lung cancers (NSCLC) compose the majority (~80%) of bronchogenic carcinoma with a lesser fraction being small-cell lung carcinomas (SCLC). The three main subtypes of NSCLC are adenocarcinoma (40%), squamous cell carcinoma (SCC; 40%), and large-cell cancer (20%). Adenocarcinoma has replaced SCC as the most frequent histologic subtype over the last 25 years. The treatment options for most patients with NSCLC are not different between either of these histologic subtypes. Currently, many early-stage lung cancer patients, such as those with stage IA disease, receive only surgery whereas other patients generally receive adjuvant chemotherapy and, possibly, radiation therapy following surgery (Lung Cancer Treatment Guidelines for Patients, National Comprehensive Cancer Network, May 2005 version 2).

The overall 10-year survival rate of patients with NSCLC is a dismal 8% to 10%. Approximately 25% to 30% of patients with NSCLC have stage I disease and, of these, 35% to 50% will relapse within 5 years after surgical treatment (1, 2). It is currently not possible to identify those patients that are of high risk of relapse. The ability to identify high-risk patients among the stage I disease group will allow for the consideration of additional therapeutic intervention. This potentially could lead to an improved survival in these patients. Indeed, recent clinical trials have shown that adjuvant therapy following resection of lung tumors can lead to improved survival in early-stage NSCLC (3, 4). Specifically, Kato et al. (3) showed that adjuvant chemotherapy with uracil-tegafur improved survival among patients with completely resected pathologic stage I lung adenocarcinoma, particularly with T₂ disease. It was also found that early-stage patients who received vinorelbine and cisplatin after surgery had an overall survival of 94 months compared with 73 months in those patients who did not receive the adjuvant therapy (4).

Microarray gene expression profiling has recently been used to define prognostic signatures in patients with lung adenocarcinomas (5-9); however, whereas several studies have investigated diagnostic profiles in lung SCC, there have been no similar large studies that have investigated gene expression profiles of prognosis in this subtype of NSCLC (10). Here, we report the profiling of a large set of 130 lung SCC samples using Affymetrix U133A GeneChips. Hierarchical clustering and Cox modeling identified genes that correlate with patient prognosis. Unsupervised clustering has also identified a SCC subtype showing an aggressive clinical profile. In addition, we have combined this SCC signature with our previous prognostic signature for lung adenocarcinoma to generate a classifier to capture the majority of NSCLC (5). The combined signature generated from both SCC and adenocarcinoma data sets was validated in an independent set of 72 NSCLC samples (11). These signatures and further understanding of their associated

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

The data discussed in this publication have been deposited in National Center for Biotechnology Information Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) and are accessible through GEO Series accession number GSE4573.

Requests for reprints: Mitch Raponi, Veridex LLC, a Johnson & Johnson company, Molecular Diagnostics, 3210 Merryfriend Row, San Diego, CA 92121. Phone: 858-320-3319; E-mail: mraponi1@vruxus.jnj.com.

©2006 American Association for Cancer Research.
doi:10.1158/0008-5472.CAN-06-1191

biological information could be used to better manage patient treatment following initial surgery.

Materials and Methods

Detailed procedures, of what is briefly described here, can be found as Supporting Information.

Patient population. One hundred-thirty fresh frozen, surgically resected SCC lung tissue from 129 individual patients (LS-71 and LS-136 were duplicate samples from different areas of the same tumor) from all stages of lung SCC were evaluated in this study. Supplementary Table S1 lists the clinical data associated with all lung SCC samples used in this study. The 86 lung adenocarcinoma samples have previously been described (5). Table 1 shows the clinical information associated with both adenocarcinoma and SCC samples used for identifying and testing the prognostic signatures. Patients were censored from statistical analysis if they were alive but had <3 years of clinical follow-up. Censored patients were excluded from the receiver operating characteristic (ROC) analysis and the calculation of sensitivity and specificity but they were included in the Cox regression for marker discovery. The independent NSCLC data set was composed of 36 lung adenocarcinoma (27 stage I) and 36 lung SCC samples (25 stage I) with at least 3 years of follow-up (Gene Omnibus data set: GSE3141; ref. 11).

Microarray analysis. Total RNA was hybridized to the Affymetrix U133A GeneChip as previously described (12). Microarray data were extracted using the Affymetrix MAS 5 software. The CEL files of the external validation data sets were downloaded from <http://data.cgt.duke.edu/oncogene.php> (11). Because the validation data sets were from external sources on different platforms (U133Plus 2.0), ANOVA was used to

normalize the batch effect such as different sample preparation methods, different RNA extraction methods, different hybridization protocols, and different scanners. The SCC data discussed in this publication have been deposited in National Center for Biotechnology Information Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) and are accessible through GEO Series accession number GSE4573.

Real-time quantitative reverse transcription-PCR. Total RNA samples were first quality tested by capillary electrophoresis using a Bioanalyzer (Agilent, Palo Alto, CA). For aRNA, the Ribobest 1-Round Aminoallyl-aRNA Amplification Kit (Epicentre, Madison, WI) was used. Real-time quantitative reverse transcription-PCR (RT-PCR) analyses were done on the ABI Prism 7900HT sequence detection system (Applied Biosystems, Foster City, CA).

Immunohistochemistry. Immunohistochemistry was done on tissue microarrays containing 60 lung SCCs. These tumors were from a subset of the ones analyzed by microarray. The 60 samples chosen for the tissue microarray were randomly selected and included 31 stage I, 9 stage II, and 19 stage III SCC samples. Areas of the tumor that best represented the overall morphology were selected for generating a tissue microarray block as previously described (13). All controls stained negative for background.

Statistical analysis. Three statistical methods were done to identify the optimal prognostic gene signature: Cox proportional-hazard regression modeling, bootstrapping, and a 5-fold cross-validation. Hierarchical clustering was done to identify major clusters of patients and investigate their association with patient co-variables. The method for developing a risk index is similar to that previously described (14). Pathway analysis was done by first mapping genes to the Biological Process categories of Gene Ontology and then calculating the significance of overrepresented categories in the selected gene list.

Table 1. Profile of adenocarcinoma and SCC lung patients used to identify and test prognostic signatures

	Training set			Testing set		
	SCC (129)	Adenocarcinoma (86)	Total (215)	SCC (36)	Adenocarcinoma (36)	Total (72)
Age \pm SD (y)	67 \pm 10	64 \pm 10	66 \pm 10	66 \pm 11	64 \pm 9	65 \pm 10
Sex						
Male	82 (64%)	35 (41%)	117 (54%)	24 (67%)	20 (56%)	44 (61%)
Female	47 (36%)	51 (59%)	98 (46%)	12 (33%)	16 (44%)	28 (39%)
Stage						
I	73 (56%)	67 (78%)	140 (65%)	25 (69%)	27 (75%)	52 (72%)
II	33 (26%)	0 (0%)	33 (15%)	9 (25%)	7 (19%)	16 (23%)
III	23 (18%)	19 (22%)	42 (20%)	1 (3%)	2 (6%)	3 (4%)
IV	0 (0%)	0 (0%)	0 (0%)	1 (3%)	0 (0%)	1 (1%)
T stage						
T ₁	33 (26%)	48 (56%)	81 (38%)	NA	NA	NA
T ₂	75 (58%)	30 (35%)	105 (49%)	NA	NA	NA
T _{3/4}	21 (16%)	8 (9%)	29 (13%)	NA	NA	NA
Lymph node						
Positive, 46 (36%)	17 (20%)	63 (29%)	NA	NA	NA	NA
Negative, 83 (64%)	69 (80%)	152 (71%)	NA	NA	NA	NA
Grade						
Well	15 (12%)	24 (28%)	39 (18%)	NA	NA	NA
Moderate	58 (45%)	41 (48%)	99 (46%)	NA	NA	NA
Poor	56 (43%)	21 (24%)	77 (36%)	NA	NA	NA
Died <3 y						
Yes	17 (47%)	17 (47%)	34 (47%)	17 (47%)	17 (47%)	34 (47%)
No	19 (53%)	19 (53%)	38 (53%)	19 (53%)	19 (53%)	38 (53%)
Censored	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

NOTE: NA, not available. Censored, patients were censored from statistical analysis if they were alive but had <3 years of clinical follow-up. Censored patients were excluded from the ROC analysis and the calculation of sensitivity and specificity but they were included in the Cox regression for marker discovery.

Results

Unsupervised hierarchical clustering identifies an aggressive subtype of SCC. One hundred-thirty total RNA samples from 129 lung SCC patients were amplified to aRNA and hybridized to Affymetrix U133A GeneChips. These samples represented all stages of SCC with 56% of tumors coming from stage I patients. We profiled all stages of SCC lung tumors because our previous data in lung adenocarcinoma identified prognostic signatures irrespective of stage and a comprehensive analysis of lung SCC tumors has not previously been done. Probe sets were filtered from the data set if they were not called present in at least 10% of all samples (including normal). The data set was further filtered by removing probe sets (coefficient of variance, <30%) that had low variation of expression across all samples.

As expected, Kaplan-Meier analysis indicated that stage III patients had significantly worse overall survival compared with stages I and II (Supplementary Fig. S1A). Stage II patients could be seen to have a worse overall survival compared with those patients with stage I only within the first year of follow-up. It was also noted that the overall survival rate in this population of older patients (mean age, 68 ± 10 years) converged beyond 5 years for all stages (data not shown). This is likely due to confounding diseases in this population. Because the majority of patients relapse within the first 3 years, we used the 3-year time point as a cutoff in our survival analyses (15).

We asked the question whether there are gene expression profiles that correlate with outcome irrespective of stage of disease. To this end, we did both unsupervised and supervised statistical analyses. The 130 lung SCC samples were clustered based on unsupervised hierarchical clustering of the remaining 11,101 genes. Samples that originated from different areas of the same tumor (LS71 and LS136) clustered next to each other and their correlation was 0.92, indicating that the gene expression profiling measures were reproducible (data not shown). The mean value of these duplicate samples was used in further analyses. Interestingly, the resulting 129 lung SCC samples clustered into two main groups: one containing 55 patients and the other with 74 patients (Fig. 1A). Similar clusters were observed when the analysis was done with 14K, 11K, 9K, or 7K genes (data not shown). No significant association between the two clusters and tumor stage, differentiation, or patient gender was identified. There were approximately equal proportions of each stage present in both clusters (cluster 1 consisted of 31 stage I, 15 stage II, and 9 stage III patients; cluster 2 consisted of 42 stage I, 18 stage II, and 14 stage III patients). However, the patients in clusters 1 and 2 showed significantly different overall survival (Fig. 1B; $P = 0.036$), indicating that gene expression profiles, irrespective of stage, existed that were associated with overall survival (Fig. 1B).

The unsupervised hierarchical clustering described above identified two groups of patients that differed significantly in their overall survival. We identified 121 genes (non-unique) of which expression levels were significantly different between the high-risk and low-risk groups ($P < 0.001$; mean difference >3-fold; Supplementary Table S2). Interestingly, the majority of these genes (118) were down-regulated in the high-risk group. Analysis of Gene Ontology functional categories revealed several groups of overrepresented genes (Supplementary Table S3). The top category included genes associated with epidermal development function, including keratins and small proline-rich proteins (Table 2). When the 20 probe sets specific for the genes involved in epidermal differentiation (based on Gene Ontology functions) were used to

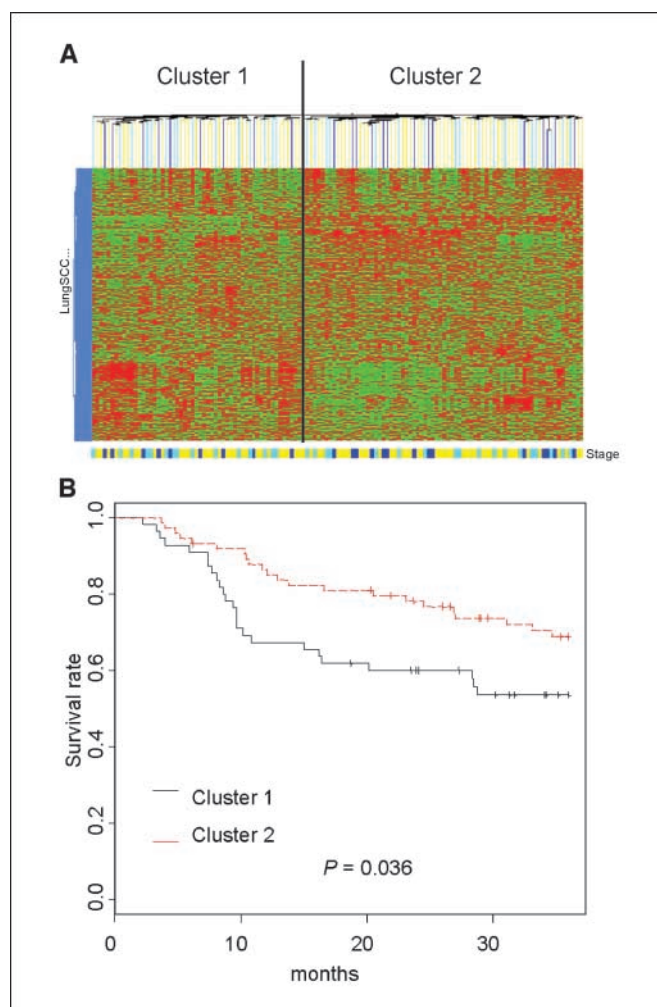


Figure 1. Unsupervised hierarchical clustering identifies two clinically relevant subsets of SCC. A, hierarchical clustering of 129 SCC samples. Each column is a sample and each row a gene transcript. Red and green, high and low gene expressions, respectively, as compared across all samples. Vertical lines, color-coded by stage: yellow, stage 1; light blue, stage 2; dark blue, stage 3. B, overall survival of patients based on cluster.

cluster the patient samples, the two prognostically differentiated groups were maintained (data not shown). These data indicate that there were two major subtypes of SCC, one of which had a profile lacking expression of epidermal differentiation genes, possibly indicating a less differentiated and more aggressive subgroup of tumors.

Interplatform validation of gene expression data. To confirm the results obtained using microarrays, we did TaqMan quantitative RT-PCR on four randomly selected genes (*FGFR2*, *KRT13*, *NTRK2*, and *VEGF*) using RNA from the same 129 tumor samples. The correlation between the platforms ranged from 0.71 to 0.96, further indicating that the Affymetrix generated expression data were reproducible (Supplementary Fig. S2). Immunohistochemistry was then done on tissue microarrays from a subset of these tumors to confirm that the expression of several of these proteins originated within tumor cells. Supplementary Fig. S3 showed expression of several keratins and the tyrosine kinase protein *FGFR2* within SCC cells rather than primarily from stromal cells. We also investigated the association of *FGFR2* protein expression

Table 2. Probe sets associated with epidermal differentiation

Probe set ID	Gene title	Gene symbol	Mean (A)	Mean (B)	Mean ratio (A/B)
206581_at	basonuclin 1	BNC1	6.51	8.12	0.33
204136_at	collagen, type VII, α 1	COL7A1	7.75	9.66	0.27
204636_at	collagen, type XVII, α 1	COL17A1	6.43	8.97	0.17
207935_s_at	keratin 13	KRT13	6.75	10.55	0.07
209351_at	keratin 14	KRT14	6.56	11.04	0.04
204734_at	keratin 15	KRT15	7.96	11.90	0.07
209800_at	keratin 16	KRT16	5.87	10.51	0.04
205157_s_at	keratin 17	KRT17	8.36	12.75	0.05
212236_x_at	keratin 17	KRT17	10.21	13.56	0.10
201820_at	keratin 5	KRT5	9.87	14.04	0.06
210355_at	parathyroid hormone-like hormone	PTH1H	6.94	9.19	0.21
206300_s_at	parathyroid hormone-like hormone	PTH1H	7.33	9.51	0.22
211756_at	parathyroid hormone-like hormone	PTH1H	7.51	9.94	0.19
213796_at	small proline-rich protein 1A	SPRR1A	5.50	11.46	0.02
214549_x_at	small proline-rich protein 1A	SPRR1A	8.44	11.14	0.15
205064_at	small proline-rich protein 1B (cornifin)	SPRR1B	8.38	12.06	0.08
208539_x_at	small proline-rich protein 2A	SPRR2A	6.25	10.69	0.05

NOTE: Mean values (cluster A and B) are log values.

and overall survival, and this supported a trend towards low FGFR2 expression and poor survival ($P = 0.02$) in the SCC tumors and was consistent with data obtained using mRNA expression.

Identification of a SCC prognostic gene signature. To identify a robust set of mRNA transcripts that could stratify patients into good and poor prognostic groups, we employed Cox proportional hazard modeling. The 129 SCC samples were analyzed using a 5-fold cross-validation to determine the optimal number of gene transcripts to be used in the prognostic signature (Fig. 2A). When increasing numbers of gene transcripts were plotted versus the mean performance from 100 five-fold cross-validations, as measured by the average area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curves and using the death within 3 years as the defining point, it could be seen that the signature performance began to plateau at around 50 probe sets (Fig. 2B). Therefore, the top ranked 50 probe sets were used in the classifier (Supplementary Table S4).

An independent set of 72 NSCLC samples comprising 36 adenocarcinoma and 36 SCC samples, with at least 3 years of follow-up, was available to test our classifiers (Gene Omnibus data set: GSE3141; ref. 11). When our SCC classifier was applied to the 36 independent SCC samples, an ROC analysis showed an AUC of 0.68 (Fig. 3A). To define a cutoff to apply to the test set, we did another 100 five-fold cross-validations in the training set. When plotted against the percentile of the risk index, an optimal cutoff of 68% was identified to reach the minimum mean error rate (Supplementary Fig. S4A). The performance of the signature with this cutoff showed a specificity of 84% [95% confidence interval (95% CI), 0.62-0.94] and a sensitivity of 41% (95% CI, 0.22-0.64) using death within 3 years as the defining point. The Kaplan-Meier analysis showed a significant difference in overall survival of the stratified patients [log-rank $P = 0.04$; hazard ratio (HR), 2.66; 95% CI, 1.01-7.05; Fig. 3B]. In the undiagnosed population, the death rate at 3 years was 47.2% compared with 38.5% and 70.0% in the good and poor prognostic patient groups, respectively. Further, when only the relevant stage I patient population was analyzed, a

similar level of stratification was found with a predictive value of 64% (Fig. 3C) and similar stratification of good and poor outcome groups with a HR of 2.78 (95% CI, 0.81-9.58; log-rank $P = 0.09$; Fig. 3D).

Performance of combined lung SCC and adenocarcinoma classifiers. We previously identified a 50-gene signature that could predict outcome in lung adenocarcinoma patients (Supplementary Table S5; ref. 5). Forty-seven of these gene transcripts were mapped to the U133Plus GeneChip used in the validation data set from Duke University. When tested in the 36 independent adenocarcinoma samples, this classifier showed an AUC 0.83 in the ROC analysis (Supplementary Fig. S5A). Again, using a cutoff defined in the training set (70th percentile for adenocarcinoma; Supplementary Fig. S4B), analysis of overall survival showed a significant

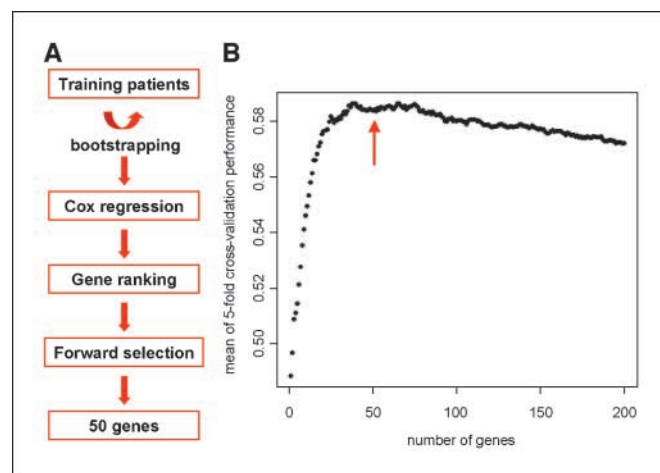


Figure 2. Selection of the classifiers for lung SCC and adenocarcinoma prognosis. A, the strategy for selecting a minimal number of prognostic gene transcripts in both the adenocarcinoma and SCC data sets. B, results of a five-fold cross-validation when an increasing number of gene transcripts are used in the classifier for SCC.

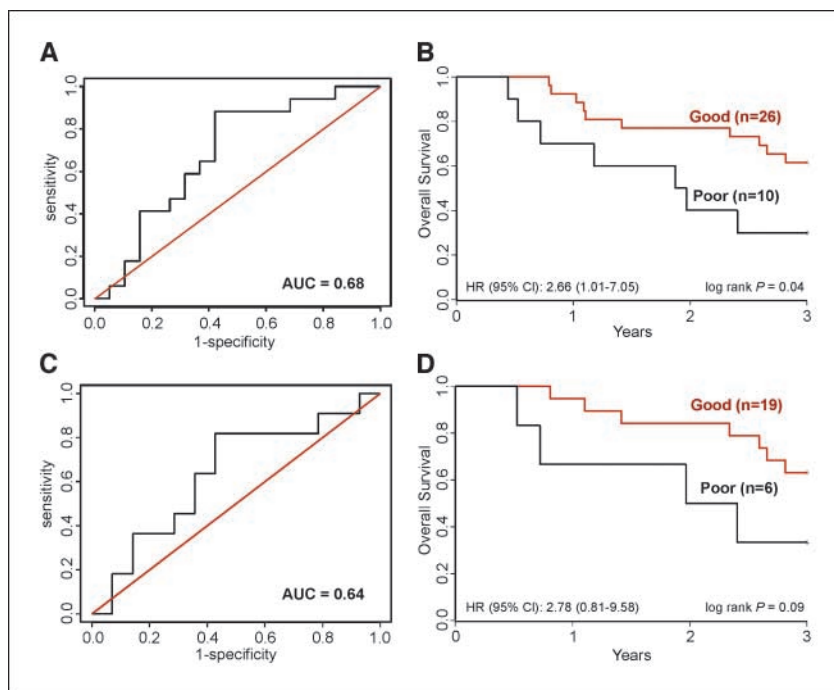


Figure 3. Validation of a 50-gene transcript SCC classifier in an independent testing set. *A*, ROC analysis of the 50-gene transcript signature in 36 independent lung SCC samples. *B*, associated Kaplan-Meier analysis of 36 patients. *C*, ROC analysis for 25 stage I patients; *D*, associated Kaplan-Meier analysis among the stage I patients only.

difference between the two groups (log-rank $P = 0.0008$; HR, 8.33; 95% CI, 1.89-36.6) with a specificity of 68% (95% CI, 0.46-0.85) and a sensitivity of 88% (95% CI, 0.66-0.98; Supplementary Fig. S5B). Similar performance was also seen in the 27 patients with stage I lung adenocarcinoma (Supplementary Fig. S5C and D).

To summarize the prognostic value of these two classifiers in NSCLC, we analyzed the independent SCC and adenocarcinoma samples as a single data set. Thus, taken together, the SCC and adenocarcinoma classifiers were tested in the complete set of 72 independent NSCLC samples. The resulting ROC curve showed an AUC of 0.71 (Fig. 4A). Kaplan-Meier analysis showed a significant stratification in overall survival (log-rank $P = 0.0002$; HR, 3.54; 95% CI, 1.74-7.19; Fig. 4B). The overall specificity and sensitivity of the two classifiers were 77% (95% CI, 0.59-0.88) and 64% (95% CI, 0.43-0.80), respectively. We then tested the performance of the two classifiers in the subset of 52 independent NSCLC stage I samples and the ROC curve showed an AUC of 0.70 (Fig. 4C). Similarly, Kaplan-Meier analysis showed a significant stratification in overall survival in this clinically relevant group of patients (log-rank $P = 0.0012$; HR, 3.86; 95% CI, 1.61-9.24; Fig. 4D). These data indicate that the combination of the two classifiers can identify early-stage NSCLC patients with either a good or poor prognosis.

Discussion

The identification of prognostic markers for NSCLC could assist in the clinical management of patients suffering from this aggressive disease (10). SCC represents one of the major histologic subtypes of NSCLC and, like adenocarcinoma, has a poor clinical outcome. Current clinical practice involves treatment of patients with stages IB and above with surgery and adjuvant chemotherapy; however, most patients that are stage IA patients receive only surgical resection. As a significant proportion of stage I patients will relapse within 3 years, identification of early-stage patients with a poor prognosis could delineate the appropriate candidates

for adjuvant therapy. Several genomic and proteomic approaches have been done to identify signatures that can more accurately stratify NSCLC patients (reviewed in ref. 10). However, the majority of these studies have focused on adenocarcinoma whereas those that have addressed SCC have examined relatively small numbers of SCC samples (6, 7, 16-18). The current study aimed to identify a novel SCC classifier and to test whether a more general NSCLC prognostic signature can be used for such purpose.

We first analyzed a large set of 129 lung SCC tumors and found evidence of a novel subset of tumors associated with a more aggressive clinical behavior. Hierarchical clustering showed two major subtypes of SCC, one of which was at a significantly higher risk of death within 3 years. There was no association with the SCC subtypes and tumor grade (differentiation). However, the expression profile associated with the poor prognostic subgroup was indicative of a dedifferentiated phenotype. Specifically, genes related to epidermal differentiation were down-regulated in the poor prognostic group, including many cytokeratins and small proline-rich proteins. Immunohistochemistry showed that several of the cytokeratins and tyrosine kinase receptor FGFR2 were associated with tumor cells and not the stroma. We also found that for FGFR2, protein expression correlated with clinical outcome in our SCC samples. Our finding that some tyrosine kinase receptors are down-regulated in aggressive SCC is consistent with the finding of Muller-Tidow et al. (19) who mapped expression of 56 known tyrosine kinase receptors in 70 early-stage NSCLC samples, including 31 SCCs. In that study, reduced expression of FGFR2 and NTRK2 showed a trend towards decreased risk of distant metastasis.

Recently, Inamura et al. (18) also found two subclasses present in 48 lung SCC samples using a cDNA array platform. In their analyses, several biological process categories were identified as being enriched in SCC compared with normal lung tissue, including epidermal development. They also showed that the two subgroups showed different clinical outcomes; however, their

analysis did not associate the epidermal development category with these clusters. This may be due to the different platforms used (different genes analyzed), their small data set, and the different algorithms for selecting functional categories. Further, the authors did not identify a prognostic classifier and, unfortunately, the raw data were not available from their study for independent analysis. Nevertheless, their report does confirm our finding of two expression-based SCC subtypes with distinct clinical outcomes.

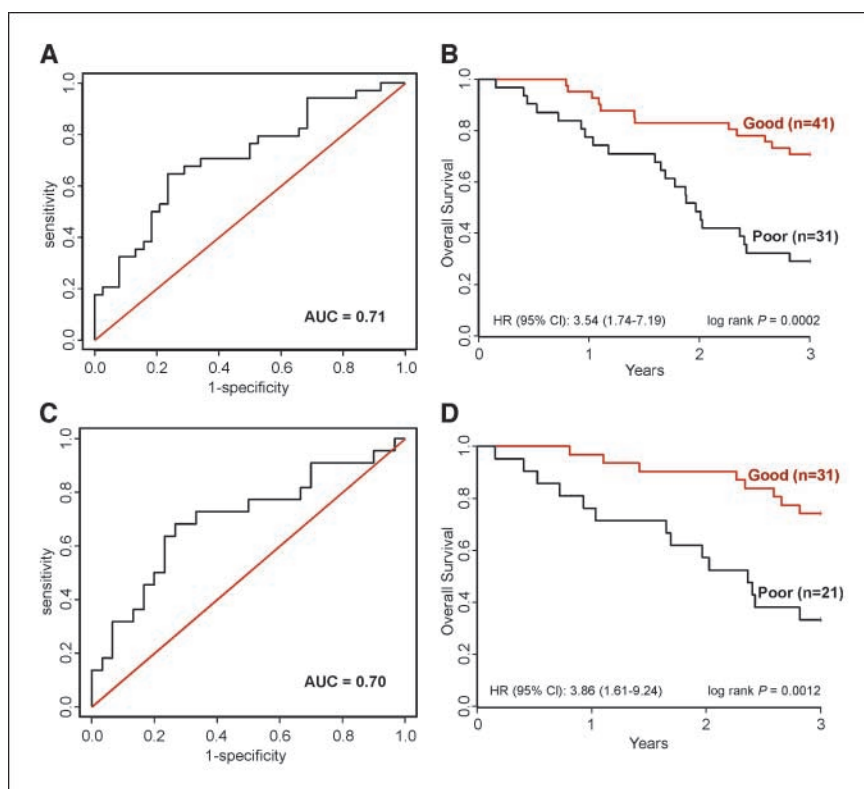
To develop the most robust prognostic signature for SCC, we did Cox proportional hazard modeling on all 129 SCC samples. An optimal set of 50 gene transcripts was identified and then tested in a completely independent set of 36 SCC samples from an independent institution (11). The performance of this classifier provided an overall predictive accuracy of 68%. Using a cutoff defined in the training set, a specificity of 84% and a sensitivity of 41% were achieved and Kaplan-Meier analysis showed significant stratification of high-risk and low-risk patients (log-rank $P = 0.04$; HR, 2.66; 95% CI, 1.01-7.05). Whereas only 25 patients samples were available that had stage I SCC and 3-year follow-up, a similar trend was also seen when analyzed with this classifier ($P = 0.09$). We believe that further analysis of larger lung SCC test sets will allow a refinement of this classifier and provide better understanding of the biological basis for associations with clinical outcome.

To determine whether an expression-based classifier can capture the majority of all NSCLCs, we combined the classifier obtained from the 129 SCC samples with our previously identified 50-gene classifier for 86 lung adenocarcinomas (5). The two classifiers were tested in an independent set of 52 stage I NSCLC samples (36 adenocarcinoma and 36 SCC; ref. 11). This dual signature clearly stratified good and poor prognostic patients with a HR of 3.86 (95% CI, 1.61-9.24; log-rank $P = 0.0012$).

It is interesting to note that the performance of the adenocarcinoma classifier showed a better predictive accuracy than the SCC classifier (adenocarcinoma AUC = 0.83, SCC AUC = 0.68). This could be due to the heterogeneity of the SCC samples as indicated by the two distinct subgroups showing differing clinical outcomes in this tumor type. The independent SCC population may not have captured this heterogeneity because it was a relatively small sample size. Indeed, we did not find evidence of the two subgroups in this set of 36 samples (data not shown). Therefore, testing in a larger independent set of SCC samples will be required to further validate this signature.

This is the first large set of well pathologically and clinically annotated lung SCC to undergo gene expression profiling. Similarities and differences between several previous analyses of SCC are apparent; however, the value of large data sets of similarly analyzed samples will allow more complete analyses of groups and subgroups of these tumors and genes most associated with them. Ongoing collaborative studies on large numbers of lung adenocarcinomas by National Cancer Institute Director's Challenge investigators will provide similar data sets. The many specific individual genes and their potential involvement in lung tumor biology and clinical outcome will require more extensive analyses as we have recently shown for several such candidates (20-22). Prognostic genes acting in both SCC and adenocarcinoma is not surprising but defining the similarities and differences will allow greater understanding of this deadly disease and possibly better diagnostic or therapeutic approaches. In conclusion, our lung SCC and adenocarcinoma prognostic signatures, when used together in a combinatorial approach, identified early-stage NSCLC patients who have a poor prognosis. Subsequent studies will examine the predictive value of our classifier for possible recommendation of adjuvant therapy in those patients with early-stage lung cancer but seem to have biologically aggressive disease.

Figure 4. Validation of a combined 100-gene transcript NSCLC classifier in an independent testing set of NSCLC patients. *A*, ROC analysis of the 100-gene transcript signature in 72 NSCLC samples; *B*, associated Kaplan-Meier analysis. *C*, ROC analysis of the 100-gene transcript signature in 52 stage I NSCLC samples; *D*, associated Kaplan-Meier analysis.



Acknowledgments

Received 3/30/2006; revised 5/9/2006; accepted 5/18/2006.

Grant support: NIH through the University of Michigan Cancer Center grant CA46592 and the Virginia Tobacco Settlement Foundation (C. Moskaluk).

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.

We thank Drs. John F. Palma and Dimitri Talantov for critical comments on the manuscript.

References

- Naruke T, Goya T, Tsuchiya R, Suemasu K. Prognosis and survival in resected lung carcinoma based on the new international staging system. *J Thorac Cardiovasc Surg* 1988;96:440-7.
- Mountain CF, Lukeman JM, Hammar SP, et al. Lung cancer classification: the relationship of disease extent and cell type to survival in a clinical trials population. *J Surg Oncol* 1987;35:147-56.
- Kato H, Ichinose Y, Ohta M, et al. A randomized trial of adjuvant chemotherapy with uracil-tegafur for adenocarcinoma of the lung. *N Engl J Med* 2004;350:1713-21.
- Winton T, Livingston R, Johnson D, et al. Vinorelbine plus cisplatin vs. observation in resected non-small-cell lung cancer. *N Engl J Med* 2005;352:2589-97.
- Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8:816-24.
- 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.
- 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.
- 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.
- Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. *Nat Genet* 2003;33:49-54.
- Meyerson M, Carbone D. Genomic and proteomic profiling of lung cancers: lung cancer classification in the age of targeted therapy. *J Clin Oncol* 2005;23: 3219-26.
- 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.
- Wang Y, Jatkoe T, Zhang Y, et al. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. *J Clin Oncol* 2004;22:1564-71.
- Kononen J, Bubendorf L, Kallioniemi A, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998;4:844-7.
- Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. *J Comput Biol* 2002;9:505-11.
- Kiernan PD, Sheridan MJ, Byrne WD, et al. Stage I non-small cell cancer of the lung results of surgical resection at Fairfax Hospital. *Va Med Q* 1993;120:146-9.
- 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.
- Tomida S, Koshikawa K, Yatabe Y, et al. Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. *Oncogene* 2004;23:5360-70.
- 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.
- Muller-Tidow C, Diederichs S, Bulk E, et al. Identification of metastasis-associated receptor tyrosine kinases in non-small cell lung cancer. *Cancer Res* 2005; 65:1778-82.
- Gharib TG, Chen G, Wang H, et al. Proteomic analysis of cytokeratin isoforms uncovers association with survival in lung adenocarcinoma. *Neoplasia* 2002;4: 440-8.
- Miller CT, Chen G, Gharib TG, et al. Increased C-CRK proto-oncogene expression is associated with an aggressive phenotype in lung adenocarcinomas. *Oncogene* 2003;22:7950-7.
- Chen G, Bhojani MS, Heaford AC, et al. Phosphorylated FADD induces NF- κ B, perturbs cell cycle, and is associated with poor outcome in lung adenocarcinomas. *Proc Natl Acad Sci U S A* 2005;102:12507-12.