Gene-expression profiles predict survival of patients with lung adenocarcinoma

David G. Beer1, Sharon L.R. Kardia2, Chiang-Ching Huang3, Thomas J. Giordano4, Albert M. Levin5, David E. Misek2, Lin Lin1, Guoan Chen1, Tarek G. Gharib1, Dafydd G. Thomas4, Michelle L. Lizyness4, Rork Kuick5, Satoru Hayasaka1, Jeremy M.G. Taylor3, Mark D. Iannettoni1, Mark B. Orringer1 & Samir Hanash5

Departments of 1Surgery, 2Epidemiology, 3Biostatistics, 4Pathology and 5Pediatrics, University of Michigan, Ann Arbor, Michigan, USA

Correspondence should be addressed to D.G.B.; email: dgbeer@umich.edu.

Published online: 15 July 2002, doi:10.1038/nm733

Lung cancer remains the leading cause of cancer death in industrialized countries. Most patients with non-small cell lung cancer (NSCLC) present with advanced disease, and despite recent advances in multi-modality therapy, the overall 10-year survival rate remains dismal 8–10%. However, a significant minority of patients (∼25–30%) with NSCLC have stage I disease and receive surgical intervention alone. Although 35–50% of patients with stage I disease will relapse within 5 years2–4, it is not currently possible to identify specific high-risk patients.

Adenocarcinoma is currently the predominant histological subtype of NSCLC (refs. 1,5,6). Although morphological assessment of lung carcinomas can roughly stratify patients, there is a need to identify patients at high risk for recurrent or metastatic disease. Preoperative variables that affect survival of patients with NSCLC have been identified7–10. Tumor size, vascular invasion, poor differentiation, high tumor-proliferative index and several genetic alterations, including K-ras (refs. 11,12) and p53 (refs. 10,13) mutations, have prognostic significance. Multiple independently assessed genes or gene products have also been identified tightly together within Cluster 1 (data not shown). We excluded if the measure of their 75th percentile value was less than 100. Although potentially resulting in the loss of some information, trimming in this manner decreased the possibility that the clustering algorithm would be strongly influenced by genes with little or no expression in these samples. Hierarchical clustering with the resulting 4,966 genes yielded 3 clusters of tumors (Fig. 1). All 10 non-neoplastic samples clustered tightly together within Cluster 1 (data not shown). We examined the relationships between cluster and patient and tumor characteristics (Fig. 1 and Supplementary Figure A online). There were associations between cluster and stage (P = 0.030) and between cluster and differentiation (P = 0.01). Cluster 1 contained the greatest percentage (42.8%) of well differentiated tumors, followed by Cluster 2 (27%) and Cluster 3 (4.7%). Cluster 3 contained the highest percentage of both poorly differentiated (47.6%) and stage III tumors (42.8%), yet contained 3 (14.3%) moderately differentiated and 1 (5%) well differentiated stage I tumors. Notably, 11 stage I tumors were present in Cluster 3, sug-
suggesting a common gene-expression profile for this subset of stage I and stage III tumors.

For patients with stage I and stage III tumors, the average ages were 68.1 and 64.5 years and the percentage of smokers was 88.9% and 89.5%, respectively. Marginally significant associations between cluster and smoking history were observed ($P = 0.06$). A significant relationship between histopathological classification and cluster was only discernable for bronchioalveolar adenocarcinomas (BAs), which were only present in Clusters 1 and 2 ($P = 0.0055$) and comprised 35.7% and 12.3% of tumors for Clusters 1 and 2, respectively.

We examined the heterogeneity in gene-expression profiles based on the trimmed data set among normal lung samples and stage I and stage III adenocarcinomas by calculating correlation coefficients between all pairs of samples. In contrast to normal lung samples that displayed highly similar gene-expression profiles (median correlation, 0.9), both stage I and III lung tumors demonstrated much greater heterogeneity in their expression profiles with lower correlation coefficients (median values, 0.82 and 0.79, respectively).

**Northern-blot and immunohistochemistry analyses**

Of the 4,966 genes examined, 967 differed significantly between Northern-blot and immunohistochemistry analyses. The mRNA from 20 of the normal lung and tumor samples were arbitrarily selected to verify the microarray expression data. The mRNA from 20 of the normal lung and tumor samples was examined by northern-blot hybridization with probes for insulin-like growth factor–binding protein 3 (IGFBP3), cystatin C and HSP-70 to determine whether mRNA overexpression was reflected by an increase of their corresponding proteins in tumors.

**Immunohistochemistry** was performed for IGFBP3, cystatin C and HSP-70 to determine whether mRNA overexpression was reflected by an increase of their corresponding proteins in tumors.

![Image](https://example.com/image1.png)

**Fig. 2** Validation analyses of gene-expression profiling. **a**, Northern-blot analysis of selected candidate genes for verification of data obtained from oligonucleotide arrays. The same sample RNA for the 4 uninvolved lung, 8 stage I and 8 stage III tumors was used for the northern-blot and oligonucleotide array analyses. **b**, Correlation analysis of quantitative data obtained from oligonucleotide arrays and northern blots measured by integrated phosphorimager-based signals for the IGFBP3 and LDH-A genes. The ratio of IGFBP3, cystatin C and LDH-A mRNA to 28S rRNA was determined. The relative values for each gene from each sample are shown. n, non-neoplastic normal lung; 1, stage I tumors; 3, stage III tumors. **c**, Immunohistochemical analysis of IGFBP-3, HSP-70 and cystatin C in lung and lung adenocarcinomas. Cytoplasmic IGFBP-3 immunoreactivity in a neoplastic gland (tumor L22) with prominent apical staining (blue reactant staining, arrow, upper left). Diffuse cytoplasmic HSP-70 immunoreactivity (tumor L27), yet stromal elements show no reactivity (upper right). Normal lung parenchyma (lower left) shows cytoplasmic cystatin C immunoreactivity in alveolar pneumocytes (arrow) and intra-alveolar macrophages but tumor (L90) shows diffuse cytoplasmic cystatin C immunoreactivity with prominent apical staining (lower right). Magnification, ×200.
Immunoreactivity for both IGFBP-3 and HSP-70 (Fig. 2c) was detected in the cytoplasm of the adenocarcinomas, with little detectable reactivity in the stromal or inflammatory cells. Cystatin C was detected in alveolar pneumocytes and intra-alveolar macrophages in non-neoplastic lung parenchyma and also consistently in the cytoplasm of neoplastic cells.

Gene-expression profiles predict survival
As expected, Kaplan–Meier survival curves (Fig. 3a) and log-rank tests indicated poorer survival among stage III compared with stage I adenocarcinomas (P = <0.0001). Two statistical approaches were used to determine whether gene-expression profiles could predict survival using the data set of 4,966 genes. In one approach, equal numbers of randomly assigned stage I and stage III tumors constituted training (n = 43) and testing (n = 43) sets. In the training set, the top 10, 20, 50 or 75 genes were used to create risk indices that were evaluated for their association with survival using the 50th, 60th or 70th percentile cutoff points to categorize patients into high or low groups. The results were similar across cutoff points but the 50-gene risk index had the best overall association with survival in the training set.

After conservatively choosing the 60th percentile cutoff point from the training set, we then applied this risk index and cutoff point to the testing set. The risk index of the top 50 genes correctly identified low- and high-risk individuals within the independent testing set (P = 0.024) (Fig. 3b and Supplementary Methods online). Notably, 11 stage I tumors were included in the high-risk subgroup. When this risk assignment was then conditionally examined for stage progression (Fig. 3c), low- and high-risk groups among stage I tumors were found to differ (P = 0.028) in their survival.

Identification of a robust set of survival genes
Although predictive of patient survival, a single training-testing set may not provide the most robust set of genes due to random sampling issues. Therefore, a 'leave-one-out' cross-validation approach was used to identify genes associated with survival from all 86-tumor samples. We first developed a 50-gene risk index in an independent cohort of 84 Massachusetts-based lung adenocarcinomas that are significantly different (P = 0.003). Among the 62 stage I lung adenocarcinomas in the Massachusetts sample, the high- and low-risk groups differed significantly (P = 0.006).

After conservatively choosing the 60th percentile cutoff point from the training set, we then applied this risk index and cutoff point to the testing set. The risk index of the top 50 genes correctly identified low- and high-risk individuals within the independent testing set (P = 0.024) (Fig. 3b and Supplementary Methods online). Notably, 11 stage I tumors were included in the high-risk subgroup. When this risk assignment was then conditionally examined for stage progression (Fig. 3c), low- and high-risk groups among stage I tumors were found to differ (P = 0.028) in their survival.

Identification of a robust set of survival genes
Although predictive of patient survival, a single training-testing set may not provide the most robust set of genes due to random sampling issues. Therefore, a 'leave-one-out' cross-validation approach was used to identify genes associated with survival from all 86-tumor samples. We first developed a 50-gene risk index in an independent cohort of 84 Massachusetts-based lung adenocarcinomas that are significantly different (P = 0.003). Among the 62 stage I lung adenocarcinomas in the Massachusetts sample, the high- and low-risk groups differed significantly (P = 0.006).

After conservatively choosing the 60th percentile cutoff point from the training set, we then applied this risk index and cutoff point to the testing set. The risk index of the top 50 genes correctly identified low- and high-risk individuals within the independent testing set (P = 0.024) (Fig. 3b and Supplementary Methods online). Notably, 11 stage I tumors were included in the high-risk subgroup. When this risk assignment was then conditionally examined for stage progression (Fig. 3c), low- and high-risk groups among stage I tumors were found to differ (P = 0.028) in their survival.

Identification of a robust set of survival genes
Although predictive of patient survival, a single training-testing set may not provide the most robust set of genes due to random sampling issues. Therefore, a 'leave-one-out' cross-validation approach was used to identify genes associated with survival from all 86-tumor samples. We first developed a 50-gene risk index in an independent cohort of 84 Massachusetts-based lung adenocarcinomas that are significantly different (P = 0.003). Among the 62 stage I lung adenocarcinomas in the Massachusetts sample, the high- and low-risk groups differed significantly (P = 0.006).

After conservatively choosing the 60th percentile cutoff point from the training set, we then applied this risk index and cutoff point to the testing set. The risk index of the top 50 genes correctly identified low- and high-risk individuals within the independent testing set (P = 0.024) (Fig. 3b and Supplementary Methods online). Notably, 11 stage I tumors were included in the high-risk subgroup. When this risk assignment was then conditionally examined for stage progression (Fig. 3c), low- and high-risk groups among stage I tumors were found to differ (P = 0.028) in their survival.
Table 1  Selected examples of the top 100 genes from cross-validation

<table>
<thead>
<tr>
<th>Gene name</th>
<th>% Change in tumor</th>
<th>% Change in stage I versus stage III t-test</th>
<th>Coefficient $\beta$</th>
<th>Unigene comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASP4</td>
<td>-6%</td>
<td>57%</td>
<td>0.0022</td>
<td>Apoptosis-related</td>
</tr>
<tr>
<td>P63</td>
<td>37%</td>
<td>43%</td>
<td>0.0010</td>
<td></td>
</tr>
<tr>
<td>KRT7</td>
<td>126%</td>
<td>55%</td>
<td>0.0003</td>
<td></td>
</tr>
<tr>
<td>LAMB1</td>
<td>-20%</td>
<td>60%</td>
<td>0.0027</td>
<td></td>
</tr>
<tr>
<td>BMP2</td>
<td>-21%</td>
<td>47%</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>CDC6</td>
<td>1070%</td>
<td>148%</td>
<td>0.0124</td>
<td></td>
</tr>
<tr>
<td>S100P</td>
<td>1572%</td>
<td>77%</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>STX1A</td>
<td>54%</td>
<td>26%</td>
<td>0.0031</td>
<td></td>
</tr>
<tr>
<td>ADM</td>
<td>39%</td>
<td>117%</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>AKAP 12</td>
<td>-47%</td>
<td>214%</td>
<td>0.0010</td>
<td></td>
</tr>
<tr>
<td>ARHE</td>
<td>87%</td>
<td>87%</td>
<td>0.0092</td>
<td></td>
</tr>
<tr>
<td>GRB7</td>
<td>15%</td>
<td>15%</td>
<td>0.0030</td>
<td></td>
</tr>
<tr>
<td>VEGF</td>
<td>85%</td>
<td>85%</td>
<td>0.0013</td>
<td></td>
</tr>
<tr>
<td>WNT10B</td>
<td>31%</td>
<td>20%</td>
<td>0.0022</td>
<td></td>
</tr>
<tr>
<td>ADM</td>
<td>8%</td>
<td>51%</td>
<td>0.0008</td>
<td></td>
</tr>
<tr>
<td>ERBB2</td>
<td>92%</td>
<td>120%</td>
<td>0.0013</td>
<td></td>
</tr>
<tr>
<td>FXYD3</td>
<td>111%</td>
<td>73%</td>
<td>0.0046</td>
<td></td>
</tr>
<tr>
<td>SLC20A1</td>
<td>58%</td>
<td>66%</td>
<td>0.0021</td>
<td></td>
</tr>
<tr>
<td>CSTB</td>
<td>50%</td>
<td>34%</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>CTLA1</td>
<td>-10%</td>
<td>67%</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>CYP24</td>
<td>N/A</td>
<td>2%</td>
<td>0.0008</td>
<td></td>
</tr>
<tr>
<td>FUT3</td>
<td>114%</td>
<td>-1%</td>
<td>0.0043</td>
<td></td>
</tr>
<tr>
<td>MLN64</td>
<td>32%</td>
<td>80%</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>PDE7A</td>
<td>33%</td>
<td>-35%</td>
<td>-0.0187</td>
<td></td>
</tr>
<tr>
<td>PGGL</td>
<td>-64%</td>
<td>-170%</td>
<td>-0.0011</td>
<td></td>
</tr>
<tr>
<td>SLC1A6</td>
<td>-32%</td>
<td>86%</td>
<td>0.0069</td>
<td></td>
</tr>
<tr>
<td>COPEB</td>
<td>-33%</td>
<td>25%</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>CRK</td>
<td>32%</td>
<td>48%</td>
<td>0.0098</td>
<td></td>
</tr>
<tr>
<td>RELA</td>
<td>-7%</td>
<td>20%</td>
<td>0.0034</td>
<td></td>
</tr>
<tr>
<td>KIAA0005</td>
<td>40%</td>
<td>45%</td>
<td>0.0010</td>
<td></td>
</tr>
<tr>
<td>MGB1</td>
<td>125%</td>
<td>459%</td>
<td>0.0018</td>
<td></td>
</tr>
</tbody>
</table>

Bolded genes were also significant for survival in 43 tumor training set (Fig. 3b).
Fig. 4  Gene expression patterns of top survival genes.  

(a) Gene-expression patterns determined using agglomerative hierarchical clustering of the 86 lung adenocarcinomas against the 100 survival-related genes (Table 1) identified by the training-testing, cross-validation analysis. Substantially elevated (red) or decreased (green) expression of the genes is observed in individual tumors. Some tumors (black arrow and expanded area) show extremely elevated expression of specific genes.  

(b) An outlier gene-expression pattern (≥5 times the interquartile range among all samples) is observed for the erbB2 and Reg1A genes (top left and right, respectively). The S100P and crk genes (bottom left and right, respectively) show a graded pattern of expression related to patient survival. 

(c) The number of outliers per person identified in the top 100 genes plotted by survival distribution.
this difference did not reach statistical significance among all patients ($P = 0.25$), between patients within tumor clusters ($P = 0.41$) or when analyzed separately among stage I ($P = 0.22$) and stage III ($P = 0.53$) patients. Nuclear accumulation of p53 was detected in 17.9% stage I and in 22.2% stage III tumors. No significant relationship was observed for p53 staining and patient survival, cluster or tumor stage.

Confirmation using an independent set of adenocarcinomas
The robustness of our 50-gene risk index in predicting survival in lung adenocarcinomas was tested using oligonucleotide gene-expression data obtained from a completely independent (Massachusetts-based) sample of 84 lung adenocarcinomas (62 stage I, 14 stage II and 8 stage III; ref. 21, and dataset A at www.genome.wi.mit.edu/MPR/lung). To ensure equivalent power for testing and comparability of samples, the criteria for including tumors in the analysis were 40% or greater tumor cellularity, no mixed histology (that is, adenosquamous) and patient survival information. To obtain comparative gene-expression measures between the two data sets, gene sequences present on the U95A and HuGeneFL array were examined, and expression data for our top 50 cross-validation genes for all 84 Massachusetts samples were obtained and processed$^{24}$ (see also Supplementary Methods online). When we examined the risk assignment of these 84 samples, employing the identical cutoff point used for the 86 Michigan-based lung samples, we observed low- and high-risk groups (Fig. 3g; $P = 0.003$). Notably, among the 62 stage I tumors, high- and low-risk groups were observed that differed significantly ($P = 0.006$) in their survival (Fig. 3h).

Survival genes had graded and outlier expression patterns
A statistical and graphical analysis of the 100 survival-related genes (Table 1) clustered against all 86 tumors revealed individual tumors with substantially elevated expression in both a limited and larger number of genes (Fig. 4a). Among these genes, we observed two distinct patterns of expression related to patient survival. One pattern, designated ‘outlier’, included genes showing substantially elevated expression (greater than five times the interquartile range among all samples), whereas the other pattern, designated ‘graded’, was characterized by continuously distributed expression with patient survival (Fig. 4b). The erbB2 and Reg1A genes are examples of outlier expression patterns and $S100P$ and crk genes of graded patterns. The number of outliers per person in the top 100 genes was identified and plotted according to survival times and events (Fig. 4c). Both stage I and stage III lung adenocarcinomas showed outlier gene patterns and 10 tumors contained 3 or more outlier genes.

Because gene amplification may result in increased gene expression, the nine genes with outlier expression patterns (erbB2, SLClA6, Wnt 1, MGB1, Reg1A, AKAP12, PACE, CYP24, KYNU) and one gene with a graded expression pattern (KRT18) were examined using quantitative genomic PCR to evaluate genomic copy number (Fig. 5a). Gene amplification of erbB2 (17q12) was detected in tumor L94, which had the highest erbB2 mRNA expression (Fig. 4a). Gene amplification was not detected for any of the other seven tested genes in tumor L94, as well as in other tumors. The two genes most frequently demonstrating the outlier pattern in these lung adenocarcinomas were KYNU and CYP24, and were present in 10 and 9 tumors, respectively. CYP24 has been described as a gene amplified and overexpressed in breast cancer$^{22}$, and these results indicate elevated expression in lung adenocarcinoma.

To determine whether the graded or outlier gene-expression patterns also occur at the protein-expression level, 10 of the 100

---

![Fig. 5](image.png)

**Fig. 5** Gene amplification and protein expression of survival-related genes. **a**, Analysis of potential gene amplification for 9 genes showing outlier expression patterns in the lung tumors (erbB2, SLClA6, Wnt 1, MGB1, Reg1A, AKAP12, PACE, CYP24 and KYNU) and examined using quantitative genomic PCR. A gene showing graded expression pattern (KRT18), and one gene with a similar chromosome location as PACE, were used as controls. Only erbB2 and Reg1A are shown. An esophageal adenocarcinoma with high-level genomic amplification of erbB2 was used as a positive control and normal esophagus DNA was used as a negative control (C0). PCR fragments sizes were 343 bp for GAPDH, 166 bp for erbB2 and 126 bp for Reg1A. DNA is from normal lung (N) and tumor(T) from each patient (for example L37). **b**, Immunohistochemical analysis of survival related genes with lung adenocarcinoma microarrays using the tumors from this study. The transmembrane erbB2 protein (top left) expression is substantially increased in tumor L94 containing the amplified erbB2 gene (Fig. 4a and b). Expression of VEGF (top right) and $S100P$ (bottom left) was located within the neoplastic cells and the pattern of immunoreactivity was consistent with the graded expression pattern demonstrated by their mRNA profiles. Expression of the oncogene crk (bottom right) was abundantly expressed in neoplastic lung cells. Magnification, $\times 400$ (erbB2); $\times 200$ (VEGF, $S100P$ and crk).
top survival genes (Table 1) for which specific antibodies were available were chosen for immunohistochemical analysis using lung-tumor arrays from this study (Fig. 5b). Expression of membrane erbB2 protein was substantially increased in the erbB2-amplified tumor L94 and very low levels of expression were present in other tumors, consistent with mRNA-expression measurements (Fig. 4a and b). CDC6 protein expression was also substantially higher in tumor L94, consistent with mRNA levels (data not shown). Expression of vascular endothelial growth factor (VEGF) and S100P (Fig. 5b), as well as cytokeratin 18 (KRT18), cytokeratin 7 (KRT7) and fas-associated death domain (FADD) protein (data not shown), was located within the lung tumor cells and consistent with the graded expression pattern of the mRNA profiles. The oncogene crk showed both graded mRNA as well as a graded protein-expression pattern with survival, and was abundantly expressed in the tumor cells (Fig. 5b). These results indicate that many survival-associated genes are expressed at the protein level and demonstrate similar mRNA and protein-expression patterns.

Discussion

We used several approaches for the analysis of gene-expression data related to clinicopathological variables and patient survival. One approach, hierarchical clustering, was used to examine similarities among lung adenocarcinomas in their patterns of gene expression. Previous studies of lung tumors21,22 have also used this method to describe subclasses of lung tumors. Here, we found three clusters that showed significant differences with respect to tumor stage and tumor differentiation. This suggests, as expected, that tumors with similar histological features of differentiation demonstrate similarities in gene expression. This feature also partly underlies the observed statistical association of tumor stage and cluster, as many of the higher-stage tumors, often poorly differentiated and previously associated with a reduced survival23,24, were located in Cluster 3. Although this cluster contained the highest percentage of stage III tumors, it also contained a nearly equal mixture of stage I and stage III tumors and not all tumors were poorly differentiated. This indicates that a subset of stage I lung adenocarcinomas share gene-expression profiles with higher-stage tumors. Notably, 10 of the 11 stage I tumors found in Cluster 3 were the high-risk stage I tumors identified using the risk index in the ‘leave-one-out’ cross-validation.

In contrast to previous analyses of lung adenocarcinomas21,22, we validated the expression data from the arrays. The strong correlation of northern-blot analysis and oligonucleotide-array data for gene expression in the same samples (Fig. 2b) indicates that these studies provide robust gene-expression estimates. Immunohistochemistry using the same tumor samples in tissue arrays demonstrates protein expression within the lung tumor cells. Together, these studies indicate that many of the genes identified using gene-expression profiles are likely relevant to lung adenocarcinoma. For example, IGFBP3 gene expression is increased in lung adenocarcinomas (Fig. 2c). IGFBP3 protein modulates the autocrine or paracrine effects of insulin-like growth factors, elevated IGFBP3 expression is observed in colon cancer25, and increased serum IGFBP3 is associated with progression in breast cancer26. Heat-shock protein 70 (HSP-70) is increased in lung adenocarcinomas of smokers27 and is associated with increased metastatic potential in breast cancer26. Increased serum lactate dehydrogenase is correlated with tumor stage and tumor burden28, and cystatin C, a cysteine protease inhibitor expressed in human lung cancers29, is prognostic in some cancers30. The decreased expression of this protease inhibitor may affect the invasive properties of the tumor cell.

The cross-validation analytical strategy we used is particularly informative for these types of gene-expression analyses for disease outcome31,32, and identification of cross-validated genes with a larger tumor cohort may help refine this risk index for use in a clinical setting. The gene-expression data also provide opportunities to observe overarching patterns that advance our understanding of associations between genes and disease. For example, the top 100 survival genes include those involved in signaling, cell cycle and growth, transcription, and translation and metabolism. Expression of many of these genes is likely a function of increased proliferation and metabolism in the more aggressive tumors. Some genes, such as erbB2 and Reg1A (Fig. 4a and b), were highly overexpressed in a few patients having poor survival. In one tumor, the erbB2 gene was amplified (Fig. 5a), demonstrating that genomic changes may underlie the overexpression of a subset of these outlier genes. Immunohistochemistry confirmed protein overexpression in this patient’s tumor (Fig. 5b). Notably, seven of the eight outlier genes were not amplified, indicating that other mechanisms underlie the increased mRNA expression of these survival-related genes.

Most genes showed a graded relationship between expression and patient survival. Genes such as that encoding VEGF, known to be strongly associated with survival in lung cancer35,36, were identified as related to patient survival in our study. VEGF demonstrated a graded expression pattern, as did the S100P and crk oncogene (Fig. 5b). S100P is a calcium-regulated protein not previously reported in lung cancer. The crk gene, the cellular homolog of the v-crk oncogene, is a member of a family of adaptor proteins involved in signal transduction and interacts directly with c-Jun N-terminal kinase 1 (JNK1)38. Although crk has not been shown to have a role lung cancer, its role in the MAP-kinase pathway, which leads to activation of matrix metalloproteinase secretion and cell invasion39, indicates potential involvement in the the tumor cell invasion or metastasis of some lung adenocarcinomas. Among the many genes identified in this study, like crk, that may be causally involved in lung cancer progression (Table 1), some were related to survival in many patients, and others in only smaller subsets of patients. This result is consistent with the complex molecular architecture of tumors in general, the heterogeneity of lung adenocarcinomas in particular and the multiple mechanisms underlying tumor-cell survival, invasion and metastasis39.

Our results demonstrate that a gene-expression risk profile—based on the genes most associated with patient survival—can distinguish stage I lung adenocarcinomas and differentiate prognoses. The particular genes that define the clusters, or are associated with survival, likely reflect the characteristics of the particular tumors included in the analysis. Current therapy for patients with stage I disease usually consists of surgical resection without adjuvant treatment32,3. Clearly, the identification of a high-risk group among patients with stage I disease would lead to consideration of additional therapeutic intervention for this group, possibly leading to improved survival of these patients.

Methods

Patient population. Sequential patients seen at the University of Michigan Hospital between May 1994 and July 2000 for stage I or stage III lung adenocarcinoma were evaluated for this study. Consent was received and the project was approved by the local Institutional Review Board. Primary tumors and adjacent non-neoplastic lung tissue were obtained at the time of
surgery. Peripheral portions of resected lung carcinomas were sectioned, evaluated by a study pathologist and compared with routine H&E sections of the same tumors, and utilized for mRNA isolation. Regions chosen for analysis contained a tumor cellularity greater than 70%, no mixed histology, potential metastatic origin, extensive lymphatic infiltration or fibrosis. Tumors were histopathologically divided into two categories based on their growth pattern: bronchial-derived, if they exhibited invasive features with architectural destruction, and bronchioloalveolar, if they exhibited preservation of the lung architecture. All stage I patients received only surgical resection with intra-thoracic nodal sampling and no other treatments. Stage III patients received surgical resection plus chemotherapy and radiotherapy.

Gene-expression profiling and K-ras mutation analysis. RNA isolation, cRNA synthesis and gene-expression profiling were performed as described40. Details of gene annotation and K-ras mutation analysis are provided in supplementary information.

Northern-blot analysis. Total cellular RNA (10 μg) was separated in 1.2% agarose-formaldehyde gels and vacuum-transferred to Gene Screen Plus (NEL Life Science Products, Boston, Massachusetts). Hybridization conditions and probe labeling were as described40. Individual sequence-validated (Huntsville, Alabama). The human histone H4 cDNA and the 28S ribosomal RNA 26-mer oligonucleotide probe were prepared and labeled as described40.

Gene-amplification analysis. 11 genes were selected for the analysis of genetic alterations. Primers were designed using PrimerSelect 4.05 Windows software (DNASTAR, Madison, Wisconsin), avoiding pseudogenes or potential homologous regions. Forward and reverse primers for the genes are provided (Supplementary Methods online). Quantitative genomic-PCR was then applied and analyzed as described40.

Immunohistochemical staining. The H&E-stained slides of all primary lung tumors were used to identify the most representative regions of each tumor and a tissue microarray (TMA) block was constructed as described40. Immunohistochemistry (IHC) was performed using both routine and sections from the TMA block as described40. Detailed methods and the concentrations used for all antibodies are provided in the Supplementary Methods.

Statistical methods. t-tests were used to identify differences in mean gene-expression levels between comparison groups. Agglomerative hierarchical clustering46 was applied using the average linkage method to investigate whether there was evidence for natural groupings of tumor samples based on correlations between gene-expression profiles. To investigate the robustness of the clustering inference, gene-expression values were perturbed by adding random Gaussian error of magnitude obtained from a duplicate sample to each data point and then reclustered to determine concordance in the tumor’s class membership. Pearson, $\chi^2$ and Fisher’s exact tests were used to assess whether cluster membership was associated with physical and genetic characteristics of the tumors.

To determine whether gene-expression profiles were associated with variability in survival times, 2 separate but complementary approaches were used. In the first approach, the 86 tumors were randomly assigned to equivalent training and testing sets consisting of equal numbers of stage I and III tumors in order to validate a novel risk-index function that captured the effect of many genes at once. In the second approach, cross-validation was used to more robustly identify the genes associated with survival. Briefly, a ‘leave-one-out’ cross-validation procedure in which 85 of the 86 tumors (the training set) was used to identify genes that were univariately proportional-hazard regression modeling46, was weighted by their estimated regression coefficients. Kaplan–Meier survival plots and log-rank tests were then used to assess whether the risk-index assignment to high/low categories was validated in the test set. A more detailed description is provided (Supplementary Methods online).

Note: Supplementary information is available on the Nature Medicine website.

Acknowledgments
We thank D. Sanders for technical assistance; D. Sing for assistance with the figures; and G. Omenn for critical reading of this manuscript. This work was supported by National Cancer Institute grant: U19 CA-8593 and the Tissue Core of the University of Michigan Comprehensive Cancer Center (NIH CA-46952).

Competing interests statement
The authors declare that they have no competing financial interests.

RECEIVED 5 APRIL; ACCEPTED 14 JUNE 2002

27. Vadgama J.V. et al. Plasma insulin-like growth factor-I and serum IGF-binding protein 3 can be associated with the progression of breast cancer, and predict the risk of recurrence and the probability of survival in African-American and Hispanic