Three-Gene Expression Signature Predicts Survival in Early-Stage Squamous Cell Carcinoma of the Lung

Marcin Skrzypski,1 Ewa Jassem,1 Miquel Taron,2 Jose Javier Sanchez,3 Pedro Mendez,2 Witold Rzyman,1 Grazyna Gulida,1 Dan Raz,5 David Jablons,5 Mariano Provencios4 Bartomeu Massuti,6 Imane Chaib,2 Laia Perez-Roca,2 Jacek Jassem,1 and Rafael Rosell2

Abstract Purpose: Adjuvant treatment may improve survival in early-stage squamous cell carcinoma (SCC) of the lung; however, the absolute gain is modest and mainly limited to stage II-IIIA. Current staging methods are imprecise indications of prognosis, but high-risk patients can be identified by gene expression profiling and considered for adjuvant therapy.

Experimental Design: The expression of 29 genes was assessed by reverse transcriptase quantitative PCR in frozen primary tumor specimens obtained from 66 SCC patients who had undergone surgical resection. Expression values were dichotomized using the median as a cutoff value. We used a risk score to develop a gene expression model for the prediction of survival.

Results: The univariate analysis of gene expression in the training cohort identified 10 genes with significant prognostic value: CSF1, EGFR, CA IX, PH4, KIAA0974, ANLN, VEGFC, NTRK1, FNI, and INR1. In the multivariate Cox model, CSF1 (hazard ratio, 3.5; P = 0.005), EGFR (hazard ratio, 2.7; P = 0.02), CA IX (hazard ratio, 0.2; P < 0.0001), and tumor size >4 cm (hazard ratio, 2.7; P = 0.02) emerged as significant markers for survival. The high prognostic value of a risk score based on the expression of the three genes (CSF1, EGFR, and CA IX) was positively validated in a separate cohort of 26 patients in an independent laboratory (P = 0.05).

Conclusions: The three-gene signature is strongly associated with prognosis in early-stage SCC. Positive independent validation suggests its suitability for selecting SCC patients with an increased risk of death who might benefit from adjuvant treatment.

Approximately 40% of stage I and 60% of stage II non–small-cell lung cancer (NSCLC) patients will die within 5 years after curative pulmonary resection, mainly due to the development of distant metastases (1). These patients may well derive a notable benefit from adjuvant chemotherapy. Whereas at present, there are no reliable clinical predictors of relapse after surgery in early-stage NSCLC, transcriptional analysis of primary tumors has identified gene expression profiles strongly related to disease recurrence in adenocarcinoma (2–8) and, to a lesser extent, in squamous cell carcinoma (SCC; refs. 5–8).

A meta-analysis (7) of data sets from seven microarray studies (2, 4, 9, 10) identified a 64-gene-expression signature that predicted survival with 85% accuracy; however, only 52 SCCs were included in these analyses (9, 10). A recent study including 51 stage I–II lung SCCs identified a 111-gene signature with a 72% predictive accuracy for disease recurrence (11). However, the lack of commonality of many genes identified between the published prognostic signatures may well reflect the fact that numerous gene expression signatures may be capable of predicting outcome in NSCLC (12). Reverse transcriptase real-time quantitative PCR (RT-QPCR) also allows for accurate and reproducible RNA quantification. The expression pattern of eight genes determined by RT-QPCR correlated with survival in lung adenocarcinoma (13). Furthermore, a five-gene signature, including genes engaged in tumor-macrophage interaction related to tumor progression, was predictive for survival in adenocarcinoma and SCC of the lung (14).

Increasing evidence suggests that SCC and adenocarcinoma of the lung are molecularly two separate entities (15). These two histologic NSCLC subtypes differ in expression levels of 33 receptor tyrosine kinases. For example, adenocarcinomas express higher levels of ERBB2 and ERBB3, whereas EPHB6 levels are higher in SCCs, and neurotrophic tyrosine receptor kinase 2 is absent in adenocarcinomas but present in SCCs (16). In addition, the expression of eight genes involved in...
DNA repair and metastasis was higher in SCCs than in adenocarcinomas (17).

To construct a reliable prognostic gene signature that could identify lung SCC patients with a high risk of death, we examined, by RT-QPCR, the expression of 29 genes related to disease progression and metastasis. We included genes with high discriminatory power (PH4, CD47 KIAA0794; ref. 18) and those concordantly up- or down-regulated (VEGF C, SERPIN1, Rho GDIβ, STX1A; ref. 12) in microarray studies. EGFR, NTRK1, and INR1 were also included based on the fact that the hazard risk for metastasis development was increased in early-stage NSCLC tumors with overexpression of these receptor tyrosine kinase genes by RT-PCR (16). The rationale for other genes included in the analysis is included in the Supplementary Text S1.

Materials and Methods

Patients. The study comprised two patient cohorts. The training cohort included 66 patients with lung SCC who underwent curative pulmonary resection between 2000 and 2004 at the hospital of the University of Gdansk, Poland. Stage I-IIIA patients fulfilling the following criteria were enrolled: complete pulmonary resection, availability of snap-frozen tumor sample (tumor specimen containing at least 60% of tumor tissue in the section), and at least a 3-year follow-up for nonrelapsed patients. In all instances, complete mediastinal lymphadenectomy was done and the surgical margins were free of tumor (R0). None of the patients received chemotherapy. The median follow-up was 37 months. The validation cohort comprised 26 stage I-II SCC patients who underwent curative pulmonary resection between 1997 and 2004 at the hospital of the University of California San Francisco. None of the patients received chemotherapy. The median follow-up was 56 months. Patient characteristics for both cohorts are summarized in Table 1, and complete details are shown in Supplementary Table S1.

Gene expression analysis. Tumor samples were obtained during surgery as blocks of 1 cm³ and snap frozen in liquid nitrogen. Tissues were stored in -80°C until total RNA was extracted with AllPrep kits (Qiagen). Only tumor samples containing >60% of tumor cells on a microscopic section were eligible for further processing. The concentration of RNA was assessed in Nano-drop and the quality of obtained RNA was confirmed on agarose gel. cDNA was synthesized from 1 μg of total RNA using the High-Capacity cDNA Archive Kit (Applied Biosystems). Quantitative RT-PCR reactions of 29 genes (Supplementary Table S2) were run using Applied Biosystems TaqMan Low Density Arrays (microfluidic cards) in an ABI PRISM 7900 HT Sequence Detection System (Applied Biosystems). One channel of a microfluidic card was loaded with a mix of 55 μL TaqMan Universal PCR Master Mix (Applied Biosystems) and 55 μL of a cDNA template corresponding to 100 ng of total RNA.

Relative gene expression values were calculated by the ∆ΔCt method (19) using the Sequence Detection System 2.1 software (Applied Biosystems). The ∆ΔCt method gives the amount of target gene normalized to an endogenous reference gene and relative to a calibrator sample (reference for all samples). The raw gene expression values were normalized according to the expression of ribosomal 18S RNA. The normalized expression of each gene was calibrated by its expression in an RNA pool (made by pooling equal amounts of total RNA from each sample). molecular analyses were carried out at the Medical Oncology Service, Catalan Institute of Oncology (Badalona, Spain), for the training cohort and at the University of California (San Francisco, CA) for the validation cohort.

Statistical analysis. We categorized SCC tumors as having good or poor prognosis based on the gene expression profile. The method consisted of five steps: (a) the gene expression was coded as 0 (low) or 1 (high), using the median expression value as a cutoff point; (b) a univariate Cox model with overall survival as the dependent variable was constructed and categorized with gene expression levels as independent variables; (c) the genes that were significant in the univariate analysis were included in a multivariate Cox proportional hazards model for survival; (d) the risk score for each patient was calculated; (e) the curves for overall survival were obtained with the Kaplan-Meier product limit method according to the median value of the risk score for the entire population. Data on patients were analyzed from the date of surgery to the time of relapse or death, or the date on which data were censored. Comparisons were made with the two-sided log-rank test. Patient characteristics were compared between the two risk groups using the Mann-Whitney test. The χ² and Fisher’s exact tests were used to compare categorical variables. The Spearman rho method was used to correlate expression levels of the genes examined. The Bonferroni method was used to adjust for multiple comparisons. All statistical analyses were carried out at a 5% level of significance and with a power of 80%, using the Statistical Package for the Social Sciences, version 13 (SPSS, Inc.).

Results

The overall median survival for the 66 patients in the training cohort was 38.2 months [95% confidence interval (95% CI), 26.6-47 months]. When patients were divided by disease stage, median survival was 41.3 months for stage I patients and 33.6 months (95% CI, 24.9-42.2 months) for stage II patients. Primary tumor size was ≤4 cm in 27 patients and >4 cm in the remaining 39. Thirty-three patients developed distant metastases after surgery, and the remaining 33 patients were free of distant metastasis after a median follow-up of 37 months (range, 24-64 months). Seven of the 33 patients who developed metastasis also had a local relapse.

A strong correlation was observed between expression levels of 10 genes (CSF1, FN1, CA IX, PH4, KIAA0974, ANLN, VEGFC, ISNR, NTRK1, and EGFR); for example, high levels of PH4 were related to low or no expression of CA IX (p = -0.33; P = 0.007) and high levels of VEGF-C were related to high levels

<table>
<thead>
<tr>
<th>Table 1. Patient characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics</td>
</tr>
<tr>
<td>Age, y</td>
</tr>
<tr>
<td>Median (range)</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Sex, n (%)</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Race, n (%)</td>
</tr>
<tr>
<td>Caucasian</td>
</tr>
<tr>
<td>African-American</td>
</tr>
<tr>
<td>Other</td>
</tr>
<tr>
<td>Stage, n (%)</td>
</tr>
<tr>
<td>IA</td>
</tr>
<tr>
<td>IB</td>
</tr>
<tr>
<td>IIA</td>
</tr>
<tr>
<td>IIB</td>
</tr>
<tr>
<td>IIIA</td>
</tr>
<tr>
<td>Tumor diameter (cm)</td>
</tr>
<tr>
<td>Predictive accuracy of the three-gene signature (%)</td>
</tr>
</tbody>
</table>
of CSF-1 ($\rho = 0.64; P = 0.0001$; Supplementary Table S3). The expression levels of these 10 genes significantly correlated with overall survival in the univariate Cox regression analysis (Table 2; Supplementary Figs. S1-S10). The 10 genes were included in a multivariate Cox regression analysis together with stage, differentiation, primary tumor size ($\leq 4 \text{ cm}$ versus $>4 \text{ cm}$), sex, and age. The expression levels of these 10 genes significantly correlated with overall survival in the univariate Cox regression analysis (Table 2; Supplementary Table S3).

A risk score was generated by adding the z scores of the expression levels of each of the three genes multiplied by its corresponding coefficient as follows: risk score $= (0.93 \times \text{CSF}) + (1.1 \times \text{EGFR}) + (1.4 \times \text{CA IX})$. The risk score was used to classify patients into high ($>0$) or low ($\leq 0$) risk, where high risk meant poor survival. Overall median survival for high-risk patients was 24 months (95% CI, 17.1-30.9 months), whereas it was not reached in low-risk patients ($P = 0.0001$; Fig. 1). The risk score was able to predict survival with 70% accuracy. The 3-year survival was 62% for low-risk patients and 20% for high-risk patients. When patients were divided into two subgroups according to primary tumor size ($\leq 4 \text{ cm}$ versus $>4 \text{ cm}$), median survival was not reached for low-risk patients in either subgroup, whereas for high-risk patients, it was 27.5 months for those with tumors $\leq 4 \text{ cm}$ ($P = 0.005$) and 20 months (95% CI, 10-32.1 months) for those with tumors $>4 \text{ cm}$ ($P = 0.002$; Fig. 2). Ninety-three percent of the patients in the low-risk group and 100% of those in the high-risk group were classified correctly.

The prognostic value of the risk score was positively validated in an independent cohort of 26 lung SCC patients at the University of California at San Francisco. Median survival in the low-risk group was not reached, whereas it was 60 months (95% CI, 12.8-108.2 months) for the high-risk group ($P = 0.05$; Fig. 3). The accuracy of the risk score in the validation cohort was 70%.

As an ancillary analysis, we tested the sensitivity and specificity of the risk score for prediction of metastasis...
development. Sixty percent of the patients in the training cohort were selected at random for setting the threshold values, and the remaining 40% of the patients formed the validation group. The risk score showed 85% sensitivity and 60% specificity for prediction of metastasis development.

**Discussion**

NSCLC is a heterogeneous disease with varying outcome in patients within the same stage. Gene expression profiling by microarrays or RT-QPCR has been used to predict prognosis in NSCLC patients (2–8). In this first RT-QPCR–based study, including exclusively lung SCC patients, we have evaluated 29 genes on the basis of their potential clinical importance and constructed a prognostic three-gene expression signature (CSF-1, EGFR, and CA IX) for lung SCC. The prognostic value of the gene expression signature was validated in an independent cohort of patients with long follow-up.

The prognostic power of our three-gene signature is comparable with that observed with recent RT-QPCR–based three-gene (20) and five-gene (14) signatures and with a five-microRNA signature (21). The 70% accuracy of our model is similar to the 68% accuracy of a 50-gene expression signature (6) and the 72% accuracy of a 111-gene expression signature (11), both derived from microarrays. However, the 50-gene signature was characterized by relatively low sensitivity (41%; ref. 6), which may limit its potential clinical application. In another microarray-based study (8), the accuracy of a “lung metagene” validated in two independent cohorts of NSCLC patients showed an overall accuracy of 72% and 79%. However, only 14 SCC patients were included in the first validation cohort, and the second validation cohort included only adenocarcinoma patients. A meta-analysis of lung cancer microarray studies, including both adenocarcinomas and SCCs, used extensive statistical methods for conversion of the gene expression values reported in the original studies and found a 64-gene expression signature with 87% accuracy in a subgroup of stage I patients (7).

Numerous gene expression signatures may be capable of predicting outcome in NSCLC, as has been shown in breast cancer.

**Table 3. Multivariate Cox model for overall survival in the training cohort**

<table>
<thead>
<tr>
<th>Hazard ratio (95% CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor size (cm)</td>
<td></td>
</tr>
<tr>
<td>≤4</td>
<td>1 (Reference)</td>
</tr>
<tr>
<td>&gt;4</td>
<td>2.7 (1.1-6.6)</td>
</tr>
<tr>
<td>CSF1</td>
<td></td>
</tr>
<tr>
<td>≤0.90</td>
<td>1 (Reference)</td>
</tr>
<tr>
<td>&gt;0.90</td>
<td>3.5 (1.5-8.5)</td>
</tr>
<tr>
<td>EGFR</td>
<td></td>
</tr>
<tr>
<td>≤0.76</td>
<td>1 (Reference)</td>
</tr>
<tr>
<td>&gt;0.76</td>
<td>2.7 (1.2-6.4)</td>
</tr>
<tr>
<td>CA IX</td>
<td></td>
</tr>
<tr>
<td>≤0.56</td>
<td>1 (Reference)</td>
</tr>
<tr>
<td>&gt;0.56</td>
<td>0.2 (0.07-0.43)</td>
</tr>
</tbody>
</table>

Fig. 1. Overall survival of patients in the training cohort according to their risk score.

Fig. 2. Overall survival of patients with tumors ≤4 cm (A) and >4 cm (B) according to their risk score.
Previous research by Bhattacharjee et al. (30) has identified a 11-cytokine gene signature, including CSF-1, with high prognostic value in lung adenocarcinoma. The expression of CSF-1, which is also known as macrophage colony-stimulating factor, is associated with dense infiltration by macrophages and is linked to tumor invasion and metastasis. This inflammatory reaction, which is mediated by macrophages, promotes tumor invasion and metastasis by means of a self-propagating chemotaxis loop involving reciprocal signaling between carcinoma cells expressing EGFR and macrophages expressing CSF-1. A recently reported 11-cytokine gene signature, including CSF-1, predicted outcome in 75% of lung adenocarcinoma patients (31). Patients whose early-stage tumors contain signatures predicting short survival may benefit from adjuvant treatment, including drugs that specifically target the CSF-1/EGFR chemotaxis loop (32).

High levels of CA IX expression formed part of a favorable prognostic three-gene signature in a microarray analysis of 116 acute myeloid leukemia patients (33). It seems that the overall expression of CA IX decreases with progression and development of metastases, raising the hypothesis that in the later stages of tumor growth, continued CA IX expression is no longer a requirement (34). (The other seven genes identified as prognostic markers in the univariate analysis are discussed in the Supplementary Text S1.)

New gene expression signatures based on functional criteria may provide complementary information that will help to refine a patient’s prognosis and inform therapeutic choices. The RT-QPCR assay is convenient in terms of laboratory work load and applicable for large-scale routine use, making it a viable alternative to more complex microarrays. The CSF-1/EGFR/CA IX gene expression signature can identify lung SCC patients with an increased risk of death and may be used to guide treatment strategies. A prospective randomized trial is warranted to validate this gene expression signature as a basis for selecting high-risk patients for adjuvant chemotherapy.

Disclosure of Potential Conflicts of Interest

Conflict of interest statement: None declared.

References

Three-Gene Expression Signature Predicts Survival in Early-Stage Squamous Cell Carcinoma of the Lung

Marcin Skrzypski, Ewa Jassem, Miquel Taron, et al.