Prognostic Gene Expression Signature for Squamous Cell Carcinoma of Lung

Chang-Qi Zhu, Dan Strumpf, Chun-Yan Li, Qing Li, Ni Liu, Sandy Der, Frances A. Shepherd, Ming-Sound Tsao, and Igor Jurisica

Abstract

Purpose: This study aimed to identify and validate a gene expression signature for squamous cell carcinoma of the lung (SQCC).

Experimental Design: A published microarray dataset from 129 SQCC patients was used as a training set to identify the minimal gene set prognostic signature. This was selected using the MAximizing R Square Algorithm (MARS), a novel heuristic signature optimization procedure based on goodness-of-fit (R square). The signature was tested internally by leave-one-out-cross-validation (LOOCV), and then externally in three independent public lung cancer microarray datasets: two datasets of non–small cell lung cancer (NSCLC) and one of adenocarcinoma (ADC) only. Quantitative-PCR (qPCR) was used to validate the signature in a fourth independent SQCC cohort.

Results: A 12-gene signature that passed the internal LOOCV validation was identified. The signature was independently prognostic for SQCC in two NSCLC datasets (total n = 223) but not in ADC. The lack of prognostic significance in ADC was confirmed in the Director’s Challenge ADC dataset (n = 442). The prognostic significance of the signature was validated further by qPCR in another independent cohort containing 62 SQCC samples (hazard ratio, 3.76; 95% confidence interval, 1.10-12.87; P = 0.035).

Conclusions: We identified a novel 12-gene prognostic signature specific for SQCC and showed the effectiveness of MARS to identify prognostic gene expression signatures.

Identifying gene expression signatures that capture altered key pathways/regulators in carcinogenesis may lead to the discovery of molecular subclasses and predict patient outcomes (1). Several prognostic gene expression signatures have been published for non–small cell lung cancer (NSCLC; refs. 2–8) and its adenocarcinoma (ADC) subtype (9–12). Fewer studies have been done to identify prognostic signatures specific for lung squamous cell carcinoma (SQCC; refs. 13, 14), but their validation in independent cohorts or datasets has been limited.

Factors such as patient/sample heterogeneity, small sample size, variation in microarray platforms, and RNA preparation and hybridization protocols could all contribute to difficulties in the validation of gene expression signatures. In addition, the loss of information through arbitrary exclusion of patients or genes prior to analysis may play an important role. Supervised data mining methodology assigns cases into good and poor prognosis subgroups at specified time points (13, 15). This arbitrary assignment of a cutoff to split good and poor prognosis cases could be problematic due to the nonlinear relationships between gene expression and patient survival. Other investigators have compared two extremes in outcome (very early death versus long survival; refs. 3, 12); however, this approach may result in significant information loss, for almost half of the cases with intermediate survival are excluded from analysis, thereby leading to high finite sample variation (16) and making the cohort under study less representative. This variability explains, therefore, why the validation of the identified signature published to date has been challenging.

It is estimated that most tissues express only 30% to 40% of genes (17) or 10,000 to 15,000 genes (18). Furthermore, among the expressed genes from similar tissue types, only a small fraction is differentially expressed. Only...
Translational Relevance

Lung squamous cell cancer (SQCC) is the second most frequent lung cancer. Treatment for this malignancy is determined by stage, although this system lacks precision at an individual patient level with respect to predicting likelihood of recurrence and death from cancer. Tumor phenotype has been shown to be driven by the underlying gene expression. The present study aimed to identify a gene expression signature capturing the altered key pathways in cancer biology for SQCC patients. We identified a 12-gene signature prognostic for SQCC that is independent of stage. Importantly, its prognostic independence has been consistently validated either in silico or by quantitative PCR in multiple independent SQCC cohorts. This signature enables us to identify low- and high-risk patients with distinct survival outcomes that potentially may direct postoperative therapy. The signature has the potential to play an important role in refining lung cancer management.

these differentially expressed genes distinguish one phenotype from another. In an attempt to compensate for this in genome-wide microarray studies, some investigators have excluded genes with low expression or low variation prior to signature selection (3, 8–10). This approach may result in the exclusion of potentially important low-expression but key regulatory genes, leading to another potential source of information loss. In addition, signatures are generated using a forced forward inclusion procedure predetermined by the rank of significance of the gene (8, 9) or the bootstrap score (13), regardless of whether the included gene contributes to the classification ability of the signature. The lack of heuristic measures in these methods potentially reduces the robustness of these signatures.

We applied a new heuristic MAximizing R Square Analysis (MARS) algorithm to identify a novel 12-gene prognostic signature from a microarray dataset of 129 SQCC patient samples (13). To show the robustness of this signature, we tested it in multiple independent cohorts of histology-defined patients with microarray or quantitative reverse transcriptase-PCR (qPCR).

Materials and Methods

Datasets

Four large, publicly available NSCLC microarray datasets were used: 129 SQCC samples collected between October 1991 and July 2002 from Molecular Diagnostics, Veridex LLC (UM; ref. 13), 85 NSCLC samples (44 SQCC and 41 ADC) samples from Duke Lung Cancer Prognostic Laboratory at Duke University (Duke; ref. 3), 138 NSCLC samples (76 SQCC and 62 ADC) from Sungkyunkwan University (SKKU) without preoperative therapy (7), and 327 ADC samples from the National Cancer Institute Director's Challenge Consortium for the Molecular Classification of ADC (DCC; ref. 11). UM was used as the training set, whereas the remaining three datasets served as independent test sets. In addition, qPCR validation of the signature was carried out in 62 SQCC samples (tumor cellularity $\geq 60\%$) from the University Health Network (UHN) without preoperative or postoperative adjuvant therapy. Patient demographics of the five independent datasets are shown in Table 1. The primary survival end point was 5-year survival (in UM, Duke, DCC) and overall survival for UHN or disease-free survival (SKKU). Additional details are provided online as “Supplementary Materials.”

Data preprocessing

Microarray data were obtained directly from the authors (13) or downloaded from the web (3, 7). Raw .cel files from UM and Duke were preprocessed by the Robust Multichip Average (RMA) algorithm using RMAexpress v0.5 (19), and log2 transformed. Because there were no raw data under the accession number GSE8894 (7), the GCRMA preprocessed data from SKKU were downloaded instead. Probe-sets (ps) were annotated using NetAffx v4.2 annotation tool (20). Only probe-sets with “grade A” annotation were used for signature optimization. Preprocessed data were standardized by Z-score transformation, which centered the expression level to mean zero and SD of 1 (21); risk score was the Z-score weighted by the coefficient of the Cox proportional model (4, 9).

Univariate analysis and signature selection

The association of the expression of individual probe-sets with 5-year survival was evaluated by Cox proportional hazards regression. A stringent threshold of $P < 0.005$ was set for preselecting the candidate probe-sets chosen for signature optimization (22).

Signature optimization was conducted by an exclusion followed by an inclusion selection procedure (Fig. 1A). The exclusion procedure took all probe-sets (96 ps) that met preselection criteria. Each probe-set was excluded one at a time, and this procedure was repeated until there was only one probe-set left. The inclusion procedure started with the probe-set left by the exclusion procedure. Each probe-set was added one at a time, and this procedure was repeated until all probe-sets were included. Then the probe set against $R^2$ was plotted and a set of minimum number of probe-sets (12 ps) having the largest $R^2$ was identified (Fig. 1B). Principal component analysis (PCA) and a stepwise selection in the Cox model were then used to synthesize gene expression information of the selected signature for dimensionality reduction, removing possible collinear expression of genes, and constructing a parsimonious multivariate model. The risk score for validation was the product of Z-score of gene expression and the coefficient of selected principal components (PC) in the multivariate model (Supplementary Table S1). The signature was then verified by an internal validation by the leave-one-out-cross-validation (LOOCV).
Validation of the expression signature

The signature generated from the training set and verified by LOOCV was then validated in silico in three independent datasets from Duke (3), SKKU (7), and DCC (11), and then by qPCR in our own UHN cohort. The independent prognostic value of the signature was tested in the multivariate Cox proportional hazard model, and was adjusted for stage, age, and sex. Statistical analyses were done using SAS v9.1 (SAS Institute).

Quantitative reverse transcriptase-PCR

qPCR was done on the ABI 7900HT real-time PCR system (AppliedBioSystems). Primers were designed primarily within the target sequences of the selected probe-sets; however, if no primer could be found in this area, primers were designed in the CDS of the target gene (Supplementary Table S2). PCR reaction optimization has been described previously (21). Five nanograms of cDNA were used for each reaction, and expression level was quantified using 2ΔΔCt method after normalization to three housekeeping genes (TBP, BAT1, and ACTB; Supplementary Table S3).

Protein-protein interaction network construction and analysis

Protein-protein interaction (PPI) data were obtained by querying the Interologous Interaction Database (http://ophid.utoronto.ca/i2d; I2D v1.71; ref. 23). The interacting proteins were then used to query the same database to determine whether any interactions were present among them. The PPI network was visualized and annotated using NAviGaTOR v2.1.13 (http://ophid.utoronto.ca/navigator; ref. 24).

Gene Ontology term and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis

GoStat (25) was used to evaluate Gene Ontology (GO) term representation enrichment. For KEGG pathways (http://www.genome.jp/kegg/; ref. 26) representation enrichment analysis, Fisher’s exact test was used and the significance was corrected by the Bonferroni method. Representation enrichment in the PPI network of the three signature probe-sets against 1,000 randomly generated PPI networks was tested by Student’s t-test.

Results

New prognostic gene expression signature for lung SQCC

The steps leading to signature identification and subsequent validation are represented schematically in Fig. 1A. In total there were 22,215 ps on the U133A chip, 19,619 with grade A annotation. Univariate analysis identified 96 ps that were significantly associated with overall survival at

Table 1. Demographic data for patients in the five datasets

<table>
<thead>
<tr>
<th></th>
<th>UM</th>
<th>Duke</th>
<th>SKKU</th>
<th>DCC*</th>
<th>UHN</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>129</td>
<td>89</td>
<td>138</td>
<td>327</td>
<td>62</td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;65</td>
<td>52</td>
<td>33</td>
<td>79</td>
<td>152</td>
<td>20</td>
</tr>
<tr>
<td>≥65</td>
<td>77</td>
<td>56</td>
<td>59</td>
<td>175</td>
<td>42</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>82</td>
<td>54</td>
<td>104</td>
<td>172</td>
<td>41</td>
</tr>
<tr>
<td>Female</td>
<td>47</td>
<td>35</td>
<td>34</td>
<td>155</td>
<td>21</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IA</td>
<td>27</td>
<td>37</td>
<td>16</td>
<td>108</td>
<td>12</td>
</tr>
<tr>
<td>IB</td>
<td>46</td>
<td>30</td>
<td>72</td>
<td>120</td>
<td>25</td>
</tr>
<tr>
<td>IIA</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>IIB</td>
<td>27</td>
<td>13</td>
<td>18</td>
<td>42</td>
<td>16</td>
</tr>
<tr>
<td>IIIA</td>
<td>17</td>
<td>3†</td>
<td>16</td>
<td>31</td>
<td>5†</td>
</tr>
<tr>
<td>IIIB</td>
<td>6†</td>
<td>1†</td>
<td>10†</td>
<td>8†</td>
<td>0†</td>
</tr>
<tr>
<td>IV</td>
<td>0</td>
<td>1†</td>
<td>0</td>
<td>0</td>
<td>0‡</td>
</tr>
<tr>
<td>Histology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>0</td>
<td>43</td>
<td>62</td>
<td>327</td>
<td>0</td>
</tr>
<tr>
<td>SQ</td>
<td>129</td>
<td>46</td>
<td>76</td>
<td>0</td>
<td>62</td>
</tr>
<tr>
<td>Platform</td>
<td>U133A</td>
<td>U133 +2</td>
<td>U133 +2</td>
<td>U133A</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The values represent number of patients and comparative percentage in parentheses. Abbreviations: AD, adenocarcinoma of lung; SQ, squamous cell carcinoma of lung.

*One case in DCC has no stage.
†Not included in analysis.
The exclusion selection procedure started with these 96 ps and by stepwise exclusion, probe set 211514_at was identified as its last one. This was followed by the inclusion procedure using 211514_at as its starting probe-set. The procedure included one probe-set at a time until all 96 ps were included. The exclusion procedure identified the largest $R^2$ of 0.77 with a combination of 12 ps (12-gene; Fig. 1B). PCA analysis and the multivariate Cox proportional hazard model with stepwise selection revealed that four PCs were significantly associated with survival at $P < 0.05$ (Supplementary Table S1). Subsequent LOOCV identified a predicting error of the signature of only 4.7% (6 cases). Thus, the 12-gene combination was established as the prognostic gene signature (Table 3).

When the risk score was dichotomized at the optimal cutoff (-0.056, Supplementary Table S1), the 12-gene signature classified 63 and 66 SQCC patients into low- and high-risk groups, respectively. High-risk SQCC had significantly poorer survival than the low-risk group (HR, 2.91; 95% CI, 1.17-7.24; $P = 0.022$; Fig. 2A), whereas the survival difference between the different risk groups for the ADC patients was not significant (HR, 1.87; 95% CI, 0.92-3.82; $P = 0.54$; Supplementary Fig. S1A). Stratified analysis by stage showed that the high-risk group classified by the signature had poorer survival in both stage I (HR, 1.87; 95% CI, 0.65-5.43; $P = 0.247$; Fig. 2B) and II SQCC (HR, 7.69; 95% CI, 0.87-67.67; $P = 0.066$; Fig. 2C). The high-risk group was associated with older patients (age $\geq 65$ years, $P = 0.02$) but not with stage or sex (data not shown). Furthermore, multivariate analysis showed that the signature was an independent prognostic factor in SQCC (HR, 15.18; 95% CI, 6.04-38.11; $P < 0.0001$; Supplementary Table S4).

In silico validation of the new 12-gene signature
We first tested the 12-gene signature in the Duke 89 NSCLC dataset (46 SQCC and 43 ADC). Four patients with stage III-IV (2 ADC and 1 SQCC in stage III and 1 SQCC in stage IV) were excluded from further analysis (Table 1). When the risk score was dichotomized at -0.056, the signature classified 25 and 19 of 44 SQCC and 13 and 28 of 41 ADC into low- and high-risk groups, respectively. High-risk SQCC had significantly poorer survival than the low-risk group (HR, 2.91; 95% CI, 1.17-7.24; $P = 0.022$; Fig. 2A), whereas the survival difference between the different risk groups for the ADC patients was not significant (HR, 1.87; 95% CI, 0.92-3.82; $P = 0.54$; Supplementary Fig. S1A). Stratified analysis by stage showed that the high-risk group classified by the signature had poorer survival in both stage I (HR, 1.87; 95% CI, 0.65-5.43; $P = 0.247$; Fig. 2B) and II SQCC (HR, 7.69; 95% CI, 0.87-67.67; $P = 0.066$; Fig. 2C). The high-risk group was associated with older patients (age $\geq 65$ years, $P = 0.02$) but not with stage or sex (data not shown). Furthermore, multivariate analysis showed that the signature was an independent prognostic factor in SQCC (HR, 3.05; 95% CI, 1.14-8.21; $P = 0.027$) but not in ADC (HR, 1.73; 95% CI, 0.59-5.12; $P = 0.322$; Table 2 and Supplementary Table S5) after adjustment for stage, age, and sex.

The SKKU dataset (7) included 138 stage I-III NSCLC (76 SQCC and 62 ADC) patients profiled using the U133
plus 2 chip. This is the only NSCLC microarray dataset from Asia. Validation of our signature used recurrence-free survival as this is the only end point reported for this study. Because the GEO database has no raw data, we downloaded the expression data which was already GCRMA-preprocessed and log2 transformed. Gene expression level was Z-score transformed and risk score was derived using the formula listed in Supplementary Table S1. The 12-gene signature classified 41 and 35 of 76 SQCC and 27 and 35 of 62 ADC into low- and high-risk groups, respectively. Significantly shortened recurrence-free survival was observed in the high-risk group in the SQCC (HR, 2.46; 95% CI, 1.26-4.79; \( P = 0.008 \); Fig. 2D) but not in the ADC (HR, 1.43; 95% CI, 0.70-2.90; \( P = 0.323 \); Supplementary Fig. S1B). Stratified analysis by stage showed that the signature worked in stage I (HR, 2.52; 95% CI, 0.93-6.78; \( P = 0.068 \); Fig. 2E) and stage II and III (HR, 6.20; 95% CI, 1.84-20.86; \( P = 0.003 \); Fig. 2F). The risk groups were not significantly associated with stage, sex, and age (data not shown). Multivariate analysis showed that the signature was independently prognostic in SQCC (HR, 2.77; 95% CI, 1.34-5.73; \( P = 0.006 \)) but not in ADC (HR, 1.92; 95% CI, 0.91-4.05; \( P = 0.086 \); Table 2 and Supplementary Table S5) after adjustment for stage, age, and sex.

To determine further whether the signature was prognostic in ADC, the 12-gene signature was tested in the largest available ADC microarray dataset from the NIH Director's

### Table 2. Validation of the 12-gene signature

<table>
<thead>
<tr>
<th></th>
<th>Squamous cell carcinoma</th>
<th>Adenocarcinoma</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n )</td>
<td>HR (95% CI)</td>
</tr>
<tr>
<td>Duke</td>
<td>44</td>
<td>3.05 (1.14-8.21)</td>
</tr>
<tr>
<td>SKKU</td>
<td>76</td>
<td>2.77 (1.34-5.73)</td>
</tr>
<tr>
<td>DCC</td>
<td>327</td>
<td>1.23 (0.85-1.78)</td>
</tr>
<tr>
<td>Quantitative reverse transcriptase PCR validation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UHN</td>
<td>62</td>
<td>3.76 (1.10-12.87)</td>
</tr>
</tbody>
</table>

NOTE: The prognostic effect of the MARSA 12-gene signature was adjusted for stage, patients' age, and sex. Abbreviation: \( n \), number of patients.

### Table 3. Composition of the 12-gene signature

<table>
<thead>
<tr>
<th>Probe set</th>
<th>Gene symbol</th>
<th>Gene title</th>
<th>Rank of exp. (%)</th>
<th>Rank of SD (%)</th>
<th>Rank of sig. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>221775_x_at</td>
<td>RPL22</td>
<td>Ribosomal protein L22</td>
<td>117 (0.6)</td>
<td>12,095 (61.7)</td>
<td>79 (82.3)</td>
</tr>
<tr>
<td>211527_x_at</td>
<td>VEGFA</td>
<td>Vascular endothelial growth factor A</td>
<td>3,660 (18.7)</td>
<td>910 (4.6)</td>
<td>48 (50.0)</td>
</tr>
<tr>
<td>213524_s_at</td>
<td>GOS2</td>
<td>G2-G2,switch 2</td>
<td>4,403 (22.4)</td>
<td>365 (1.9)</td>
<td>69 (71.9)</td>
</tr>
<tr>
<td>218678_at</td>
<td>NES</td>
<td>Nestin</td>
<td>4,504 (23.0)</td>
<td>4,749 (24.2)</td>
<td>64 (66.7)</td>
</tr>
<tr>
<td>211282_x_at</td>
<td>TNFRSF25</td>
<td>Tumor necrosis factor receptor superfamily, member 25</td>
<td>7,582 (38.7)</td>
<td>6,614 (33.7)</td>
<td>59 (61.5)</td>
</tr>
<tr>
<td>36552_at</td>
<td>DKFZP586P0123</td>
<td>Hypothetical protein</td>
<td>9,094 (46.4)</td>
<td>11,934 (60.8)</td>
<td>31 (32.3)</td>
</tr>
<tr>
<td>221900_at</td>
<td>COL8A2</td>
<td>Collagen, type VIII, a 2</td>
<td>10,236 (52.2)</td>
<td>1,574 (8.0)</td>
<td>66 (68.8)</td>
</tr>
<tr>
<td>219604_s_at</td>
<td>ZNF3</td>
<td>Zinc finger protein 3</td>
<td>15,673 (79.9)</td>
<td>18,300 (93.3)</td>
<td>71 (74.0)</td>
</tr>
<tr>
<td>211514_at</td>
<td>RIPK5</td>
<td>Receptor interacting protein kinase 5</td>
<td>15,976 (81.4)</td>
<td>19,129 (97.5)</td>
<td>2 (2.1)</td>
</tr>
<tr>
<td>221909_at</td>
<td>RNFT2</td>
<td>Ring finger protein, transmembrane 2</td>
<td>16,306 (83.1)</td>
<td>2,740 (14.0)</td>
<td>3 (3.1)</td>
</tr>
<tr>
<td>201335_s_at</td>
<td>ARHGEF12</td>
<td>Rho guanine nucleotide exchange factor (GEF) 12</td>
<td>17,123 (87.3)</td>
<td>18,491 (94.3)</td>
<td>21 (21.9)</td>
</tr>
<tr>
<td>215172_at</td>
<td>PTPN20A/B</td>
<td>Protein tyrosine phosphatase, non-receptor type 20A/B</td>
<td>19,558 (99.7)</td>
<td>17,956 (91.5)</td>
<td>65 (67.7)</td>
</tr>
</tbody>
</table>

NOTE: Rank of expression level (rank of exp.) is from high to low, rank of SD is from large to small, and rank of signature level (rank of sig.) is from high to low.
Challenge Consortium study (11), which included 442 samples. Among them, 327 patients did not receive any adjuvant chemotherapy or radiotherapy and had follow-up longer than 1 month. The 12-gene signature was not prognostic (HR, 1.26; 95% CI, 0.87-1.81; \( P = 0.221 \); Supplementary Fig. S1C). Multivariate analysis showed that it was not an independent prognostic factor in ADC (HR, 1.23; 95% CI, 0.85-1.78; \( P = 0.267 \); Table 2). These data confirm that the signature was not prognostic in ADC.

qPCR validation in UHN SQCC cohort

qPCR validation of the 12-gene signature was done in an independent set of 62 snap-frozen SQCC samples from UHN without preoperative or postoperative therapy. Fold change was calculated using \( 2^{-\Delta\Delta Ct} \) method and then Z-score transformed. Risk score was generated using parameters listed in Supplementary Table S1. When risk score was dichotomized at -0.056, the 12-gene signature was able to separate 41 and 21 SQCC into low- and high-risk groups with significant difference in 5-year overall survival (HR, 4.00; 95% CI, 1.20-13.31; \( P = 0.024 \); Fig. 2G). Stratified analysis by stage revealed that the signature was able to separate low- and high-risk groups with different survival outcomes; however, the significance was marginal due to the small sample size (stage I: HR, 3.39; 95% CI, 0.66-17.47; \( P = 0.145 \); Fig. 2H; and stage II and III: HR, 5.33; 95% CI, 0.88-32.19; \( P = 0.069 \); Fig. 2I). The risk groups were not associated with stage, age, and sex (data not shown). Nevertheless, multivariate analysis again showed that the signature was an independent prognostic factor (HR, 3.76; 95% CI, 1.10-12.87; \( P = 0.035 \); Table 2).

Because smoking influences airway gene expression, we assessed the association of the signature with smoking status. There were 46 past and 15 current smokers, and

![Fig. 2.](image-url)
only 1 lifetime nonsmoker in the UHN cohort. The $\chi^2$ test indicated that the association between the signature and patients’ past/current smoking status was not significant ($P = 0.81$). A technical factor that may have an impact on the gene expression level is the tumor cellularity. We did a Pearson correlation analysis of the risk score (as a continuous variable) and tumor cellularity in the UHN cohort. It showed that the risk score was not significantly correlated with tumor percentage (Pearson $r = -0.08, P = 0.71$).

The composition of the 12-gene signature

Table 3 shows the members of the 12-gene signature and their ranks of expression level, variance, and significance in the Veridex dataset (in decreasing order of importance). Notably, the expression level of individual genes varies greatly, from very high levels as for RPL22 (ranked in the top 0.6%) to extremely low levels for PTPN20A/B (ranked at 99.7%). The SD value also varies greatly, from very large as for G0S2 (ranked at 1.9% of the total) to very small for RIPK5 (ranked at 97.5% of the total). These data showed that the low-expression and low-variability genes were as important as those with higher expression and higher variability.

The GO (27) and KEGG pathway (26, 28) annotations revealed the involvement of several of the prognostic genes in signal transduction (e.g., VEGFA, TNFRSF25), cell cycle (e.g., VEGFA, G0S2), apoptosis (e.g., TNFRSF25), adhesion (e.g., COL8A2), and transcription and translation (ZNF3 and RPL22, respectively; Supplementary Table S6).

PPI network analysis

To assess the potential SQCC-specific biological relevance of the 12-gene signature genes further, we evaluated the functional relationship between our 12-gene signature and the reported Raponi (13) and Sun (8) 50-gene signatures (mapped to 12, 48, and 42 genes, respectively) through their corresponding PPI networks. We mapped 8 of 12 genes of the 12-gene signature, 35 of 48 and 31 of 42 for the Raponi and Sun signatures, respectively, to PPIs in the I2D 1.7 (23). Whereas the Raponi and Sun signatures had 10 overlapping probe-sets (9 genes), the 12-gene signature had no probe-sets genes overlapping with either of the 50-gene signatures. However, direct interactions between the signature genes/proteins or via shared interacting proteins were seen among these signatures, implying a rich shared functional milieu (Fig. 3). Annotation of the resulting PPI network with KEGG pathways indicated significant enrichment for proteins from the mitogen-activated protein kinase (MAPK) signaling pathway ($P = 0.019$; 80 of 1,075 proteins), which form direct interactions with 3, 14, and 9 genes/proteins of ours and the Raponi and Sun signatures, respectively (Supplementary Tables S6-S8).

Discussion

We describe here the MARSA algorithm, a heuristic signature selection method that includes only genes contributing to the separation ability of the signature. By applying the algorithm to the UM dataset, we identified a 12-gene prognostic signature. The prognostic value of the 12-gene signature was validated in silico in two independent SQCC microarray datasets (Duke HR, 3.05, 95% CI, 1.14-8.21; $P = 0.027$; SKKU HR, 2.73; 95% CI, 1.32-5.64; $P = 0.007$; Table 2) but not in the corresponding ADC datasets (Table 2). Further, we confirmed the absence of the prognostic value of the 12-gene signature in the largest available ADC dataset from DCC containing 442 ADC samples (Table 2). Importantly, qPCR validation in another independent cohort confirmed that the signature was an independent prognostic factor in SQCC (Table 2). Combined, our data strongly suggested that the 12-gene signature is a valuable prognostic factor for SQCC.

A prognostic factor provides information about a patient’s overall outcome that is independent of therapy, whereas a predictive biomarker gives information about the effect of a therapeutic intervention. When establishing a prognostic factor, it is best to evaluate outcome in patients who have not received systemic adjuvant therapy after surgery because chemotherapy has been shown to change the course of the disease and to improve overall survival significantly. In the present study, samples in the UM dataset were collected from patients between October 1991 and July 2002 who were not likely to have received adjuvant chemotherapy. Importantly, the signature was computationally confirmed in the two validation datasets (Duke and SKKU), which were used to explore prognostic signatures, and validated by qPCR in 62 patients in the UHN cohort of patients who had not received adjuvant therapy. Therefore, in these untreated patients our 12-gene signature is prognostic, and not predictive.

The cellular origin and pathogenesis of SQCC and ADC remain controversial. In contrast to ADC, SQCC tends to arise in the epithelium of large airways and its etiology is clearly linked to smoking, suggesting different pathogenic differences between the two lung cancer types (29). This is supported by differences in the occurrence of key genetic alterations in the two types of cancer (30). Although frequently mutated in ADC, KRAS (31, 32) and EGFR (33) mutations occur very infrequently in SQCC. In contrast, P53 mutation (32), TIMP3 (34) and HIF-1α (35) overexpressions occur more frequently in SQCC than in ADC of the lung. Moreover, gene expression profiling has shown distinctive patterns among the subtypes of NSCLC (36). Therefore, it may not be surprising that there could be gene signatures that are prognostic in SQCC but not in ADC patients.

Cancer phenotype is characterized by underlying gene expression. Thus, gene expression signatures may predict clinical outcome. The fact that our signature could be validated consistently in multiple independent SQCC cohorts suggests a notion that it might have captured a key gene expression program in squamous cancer biology. Indeed, many members of the 12-gene signature have been reported to be involved in processes underlying tumorigenesis, including tumor necrosis factor receptor superfamily
member 25 (TNFRSF25), triggering apoptosis and activating the transcription factor NF-κB in HEK293 or HeLa cells (37). RIPK5, a cell death inducer (38). Vascular endothelial growth factor (VEGF or VEGFA) has been extensively studied (39) and is a major regulator of tumor angiogenesis (40). ARHGEF4 (Rho guanine nucleotide exchange factor 4) is involved in G-protein mediated signaling, which has been implicated in regulating cell morphology and invasion (41). It has also been shown to interact directly with insulin-like growth factor receptor I (IGF-IR), providing a link between G protein-coupled and IGF-IR signaling pathways (ref. 42; Fig. 3). Inhibitors of IGF-IR are being studied in clinical trials in combination with chemotherapy and EGFR therapy, and preliminary results show high response rates in advanced NSCLC patients, especially of the SQCC subtype (43), although randomized trials to date have been negative.

Previous approaches to the identification of prognostic signatures filtered out low-expression or low-variance genes prior to signature selection. However, this might lead to the exclusion of low-expression but important genes in the signatures. In fact, one third of the genes (ARHGEF4, RIPK5, PTPN20A, and ZNF3) in the 12-gene signature had expression levels in the lowest 20% (from 79.9-99.7%), whereas their variation (SD) was in the lowest 10% (from 91.5-97.5%; Table 3) of all probe-sets. The consistent performance of the 12-gene signature in the training and test cohorts implied that these low-expressed and low-variable genes might have played important roles in tumor progression, and thus these genes must be included in signature selection.

In summary, MARSa is an effective approach to identify prognostic gene expression signatures, and this novel 12-gene prognostic signature seems specific for SQCC.

**Disclosure of Potential Conflicts of Interest**

No potential conflicts of interest were disclosed.
Acknowledgments

We thank Dr. Raponi and Veridex for providing the raw data of UM dataset, Dr. Keyue Ding of the National Cancer Institute of Canada Clinical Trials Group, and Dr. Geoffrey Liu and Ms. Melania Pintilie for helpful suggestions.

Grant Support

Canadian Cancer Society (NCIC 015184); Ontario Research Fund-Research Excellence (#03-020); Canada Foundation for Innovation (#12301, #203383); Canada Research Chair program; the Natural Science & Engineering Research Council of Canada (Grant #104105); Canadian Institutes for Health Research (CIHR 202370); and IBM. This research was funded in part by the Ontario Ministry of Health and Long Term Care (OMOMHTC). The views expressed do not necessarily reflect those of OMOMHTC.

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.

Received 03/10/2010; revised 07/30/2010; accepted 08/19/2010; published OnlineFirst 08/25/2010.

References


5046 Clin Cancer Res; 16(20) October 15, 2010 Clinical Cancer Research

Downloaded from clincancerres.aacrjournals.org on July 25, 2017. © 2010 American Association for Cancer Research.