Lung adenocarcinoma five-gene signature

A five-gene and corresponding-protein signature for stage-I lung adenocarcinoma prognosis

Humam Kadara1, Carmen Behrens1, Ping Yuan2, Luisa Solis2, Diane Liu3, Xuemin Gu3, John D. Minna4, J. Jack Lee3, Edward Kim1, Waun-Ki Hong1, Ignacio I. Wistuba1,2†* and Reuben Lotan1†*.

Departments of 1Thoracic/Head and Neck Medical Oncology, 2Pathology and 3Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston TX and 4Hamon Center for Therapeutic Oncology the University of Texas Southwestern, Dallas, TX.

† These authors have contributed equally to this work and both should be considered as corresponding authors.

*To whom correspondence should be addressed: Reuben Lotan, Ph.D., Department of Thoracic/Head and Neck Medical Oncology, The University of Texas M.D. Anderson Cancer Center, Tel: 713-792-8467, Fax: 713-745-5656, Email: rlotan@mdanderson.org; or Ignacio I. Wistuba, M.D., Departments of Pathology and Thoracic/Head and Neck Medical Oncology, the University of Texas M.D. Anderson Cancer Center, Tel: 713-563-9184, Fax: 713-563-1848, Email: iiwistuba@mdanderson.org.

Keywords: Lung adenocarcinoma, NSCLC, gene signature, prognosis

Running title: Lung adenocarcinoma five-gene signature

Grant support

This work was supported in part by the Jeffrey Lee Cousins Fellowship in Lung Cancer Research (to HK), Department of Defense (grant W81XWH-04-1-0142 to W.K.H., R.L. and I.I.W.), Jimmy Lane Hewlett Fund for Lung Cancer Research (to R.L. and I.I.W) and NCI lung cancer SPORE (P50 CA70907 to J.D.M. and I.I.W.).
Lung adenocarcinoma five-gene signature

**Statement of Translational Relevance**

Identification of molecular prognostic markers will improve the clinical management of NSCLC, the leading cause of cancer-related deaths in the United States and worldwide. We derived a five-gene and corresponding immunohistochemical protein signature and tested their prognostic capacities in various publicly available microarray datasets and in an independent set of NSCLC histological tissue specimens, respectively. Both the five-gene transcript and corresponding protein signatures effectively predicted poor survival of all-stages or stage-I lung adenocarcinoma but not of squamous cell carcinoma patients. Moreover, the FILM protein signature specifically identified a subgroup of non-treated stage-IB lung adenocarcinoma patients with poor prognosis. These findings suggest that the derived five-gene signature may be assessed for expression by different methods (transcript versus protein) and for identifying early stage lung adenocarcinoma patients with poor prognosis that will benefit from adjuvant therapy following resective surgery.
Abstract

PURPOSE: Identification of effective markers for outcome is expected to improve the clinical management of non-small cell lung cancer (NSCLC). Here, we assessed in NSCLC the prognostic efficacy of genes, which we had previously found to be differentially expressed in an in vitro model of human lung carcinogenesis. EXPERIMENTAL DESIGN: Prediction algorithms and risk-score models were applied to the expression of the genes in publicly available NSCLC expression datasets. The prognostic capacity of the immunohistochemical expression of proteins encoded by these genes was also tested using formalin-fixed paraffin-embedded (FFPE) tissue specimens from 156 lung adenocarcinomas and 79 squamous cell carcinomas (SCCs). RESULTS: The survival of all-stages (p<0.001, HR=2.0) or stage-I (p<0.001, HR=2.84) adenocarcinoma patients that expressed the five-gene in vitro lung carcinogenesis model (FILM) signature was significantly poorer than that of patients who did not. No survival differences were observed between SCCs predicted to express or lack FILM signature. Moreover, all stages (p<0.001, HR=1.95) or stage-I (p=0.001, HR=2.6) adenocarcinoma patients predicted to be at high risk by FILM transcript exhibited significantly worse survival than patients at low risk. Furthermore, the corresponding protein signature was associated with poor survival (all stages, p<0.001, HR=3.6; stage-I, p<0.001, HR=3.5; stage-IB, p<0.001, HR=4.6) and mortality risk (all stages, p=0.001, HR=4.0; stage-I, p=0.01, HR=3.4; stage-IB, p<0.001, HR=7.2) in lung adenocarcinoma patients. CONCLUSIONS: Our findings highlight a gene and corresponding protein signature with effective capacity for identification of stage-I lung adenocarcinoma patients with poor prognosis that are likely to benefit from adjuvant therapy.
Introduction

Lung cancer is the leading cause of cancer-related deaths in the United States and worldwide (1). NSCLC accounts for the majority of the lung cancer cases and is comprised of two major subtypes, lung adenocarcinomas and SCCs (2). The average 5-year relative survival rate among NSCLC patients is only 15% (2-4). Mortality due to NSCLC is high because most cancers are diagnosed after regional or distant spread of the disease (3, 4). However, even the 5-year survival rate of stage I NSCLC patients (30-50%) is among the worst for early-stage disease of all other malignancies (4, 5). Therefore, identification of markers for early prediction of outcome is warranted for better clinical management of NSCLC patients, and in particular for those with early stage disease.

Although surgical resection is the first treatment, adjuvant therapy has been shown to improve the survival of lung cancer patients (6, 7). Several factors have been proposed to elevate risk and justify the use of adjuvant therapy for lung cancer patients of which the TNM staging is the most effective standard (5). However the potential benefits of adjuvant therapy are contentious, in particular in stage-I lung cancer patients (6, 7) as there are few if not any established clinical criteria to separate stage-I NSCLCs (8). Therefore, it is possible that additional molecular factors might help identify early or stage-I lung cancer patients with poor prognosis that may need to receive therapy versus patients with good prognosis that could be spared adjuvant therapy.

With the advent of microarray and high-throughput technology, several early studies have shown the significant association of gene expression profiles and signatures with the survival and outcome of NSCLC patients (9-12). Additionally, gene expression signatures have been derived for the prognosis of early (stages I and II) (13, 14) and all stages NSCLC patients (15-17), lung adenocarcinomas alone (18, 19) and to predict recurrence of NSCLC disease (20, 21). Moreover, a five-gene signature identified from a set of 672 genes differentially expressed among invasive lung adenocarcinoma cell lines identified NSCLC patients at high risk of poor survival (16). We have previously studied the gene expression profiles of normal, premalignant and tumorigenic lung epithelial cells constituting an in vitro
Lung adenocarcinoma five-gene signature model of lung carcinogenesis and highlighted prominent gene expression profiles and pathways relevant to the survival of lung adenocarcinoma patients (22). Specifically, we identified the progressive modulation of six key genes, ubiquitin conjugating enzyme 2C ($UBE2C$), minichromosome maintenance ($MCM$) 2 and 6, targeting protein for Xklp2 (TPX2), flap structure-specific endonuclease 1 ($FEN1$) and stratifin ($SFN$), among the cell lines and the significant association of their expression with the survival of lung adenocarcinoma patients (22).

In this study, we sought to assess the prognostic efficacy of the aforementioned six genes in NSCLC. We analyzed the association of the expression of the genes at the transcript level with patient survival in several publicly available microarray datasets of lung adenocarcinoma and SCC. We derived a five-gene signature that is predictive of survival of early stage lung adenocarcinoma but not SCC patients. Moreover, we analyzed the proteins encoded by the five-gene signature using immunohistochemistry in an independent FFPE tissue microarray (TMA) NSCLC set, and found that it was also effective in predicting the survival of lung adenocarcinoma patients including those with stage I disease. Moreover, risk of mortality assessed by both the transcript and protein versions of the derived five-gene expression signature was an independent predictor of poor survival in lung adenocarcinoma patients. Importantly, the FILM protein signature was capable of identifying stage-IB lung adenocarcinoma patients with dismal prognosis warranting the need of adjuvant therapy in these patients.

**Methods**

**Analysis of publicly available NSCLC microarray datasets**

To assess the expression and clinical relevance of the aforementioned six genes in human NSCLC we used publicly available lung adenocarcinoma microarray data sets from the studies by Shedden et al (National Cancer Institute (NCI) Director’s Challenge, n=442) (19), Bild et al (Duke cohort, n=58) (15) and Bhattacharjee et al (Harvard cohort, n=125) (10). We also used publicly
Lung adenocarcinoma five-gene signature available microarray datasets generated from lung SCC patient samples from the studies by Raponi et al (n=130) (17) and Bild et al (n=53) (15). The raw microarray data for the Director’s Challenge study were obtained from the National Cancer Institute Cancer Array database, experiment ID 1015945236141280:1 (https://caarraydb.nci.nih.gov/caarray). Raw data from all the published studies were obtained from the Gene Expression Omnibus (GEO). Raw microarray data from all data sets were analyzed using the BRB-ArrayTools v.3.8.0 developed by Dr. Richard Simon and BRB-ArrayTools Development Team (23). Robust multi-array analysis (RMA) was used for normalization of gene expression data in R language environment (24). Probe sets for UBE2C, MCMs 2 and 6, FEN1, TPX2 and SFN in the microarray platforms used in the different studies (Affymetrix HG-U95Av2, HG-U133A and HG-U133 plus 2.0) were identified using NetAffx™ from Affymetrix (http://www.affymetrix.com/analysis/index.affx).

**Prediction of class**

To predict the class of independent patient cohorts, we adopted a previously developed model using six algorithms, compound covariate predictor (CCP), linear discriminator analysis (LDA), nearest neighbors 1 and 3 (NN-1 and NN-3), nearest centroid (NC) and support vector machines (SVM) (23, 25). Lung adenocarcinomas from the datasets in the study by the Shedden et al (19) study were designated together as a training set (n=442). Class separation into high versus low expression (Figure 1A) was maintained following analysis of the six genes in the 442 lung adenocarcinomas by hierarchical cluster analysis with average linkage (data not shown) using Cluster version 2.11. Samples from the Duke (15) and Harvard (10) cohorts were used as a combined lung adenocarcinoma test set and median-centered independently (n=183), whereas SCC samples from the studies by Raponi et al (17) and Bild et al (15) were designated as an SCC validation set (n=183) (Figure 1A). Assessment of the classification efficacy of the six genes as well as their capacity in estimating the probability of the identity of a particular sample was performed using a leave-one-out-cross-validation (LOOCV) approach with random
Lung adenocarcinoma five-gene signature permutation for accuracy estimation as previously described (23, 25). Five (all but $SFN$) of the six genes were capable of significant class prediction in training statistically assessed by a univariate t-test with a statistical cut-off of $p<10^{-5}$ (Figure 1A). Furthermore and compared to the other five genes, $SFN$ was down-regulated in tumors relative to normal tissue, exhibited relatively lower mean expression in tumors and increased misclassification. The six classification algorithms were then applied to the indicated test sets and the different patient groups or arms, predicted by the derived FILM signature, were analyzed for statistically significant differences in survival by Kaplan-Meier method for estimation of survival probability and log-rank tests in R language environment.

**Computation of risk scores**

We generated a model for estimation of mortality risk similar to what was described earlier (26). Using gene expression data from training cohorts during LOOCV, the Cox regression coefficients were computed for $UBE2C$, $MC2$ and 6, $FEN1$ and $TPX2$. A risk score was then derived for each patient in the training set by calculating the summation of the products of the Cox coefficients and normalized centered expression of each gene. Patients were dichotomized into high-risk and low-risk groups using the 50th percentile (median) cut-off of the risk score as the threshold value or were divided into tertiles or three groups (low, intermediate and high risk). In all cases, the coefficient and the threshold value derived from the training cohort were directly applied to the gene expression data from validation cohorts. Kaplan-Meier and log-rank test were then applied to assess the significance of prognostic difference between the risk groups without or with the inclusion of age and gender clinical covariates as previously described (27).

**Tissue microarray**

For this study, we obtained archived FFPE samples from surgically resected lung cancer specimens from the lung cancer tissue bank at The University of Texas M. D. Anderson Cancer Center.
Lung adenocarcinoma five-gene signature (Houston, TX). The tissue specimens originally had been collected between 2003 and 2005 and had been classified using the 2004 World Health Organization (WHO) classification system as described before (28). The tissue microarray analyzed in this study is comprised of 235 NSCLC tumor specimens (156 lung adenocarcinomas and 79 SCCs) obtained from patients, who underwent surgery at the same institution from 2003-2005 and who did not receive any adjuvant or neo adjuvant therapy, under a protocol that was approved by the MD Anderson Cancer Center institutional review board. Detailed clinical and pathologic information was available for most of these cases and included patients’ demographic data, smoking history (never smokers or ever smokers, patients who had smoked at least 100 cigarettes in their lifetime) and pathologic tumor-node-metastasis (TNM) staging (29). After histological examination of NSCLC specimens, the NSCLC TMAs were constructed by obtaining three 1-mm in diameter cores from each tumor at three different sites (periphery, intermediate and central tumor sites). The TMAs were prepared with a manual tissue arrayer (Advanced Tissue Arrayer ATA100, Chemicon International, Temecula, CA).

Immunohistochemical analysis of the protein signature

Immunohistochemistry was done on histology sections of FFPE tissue samples, using the purified primary anti-human antibodies against UBE2C (1:500 dilution) (Boston Biochem, Cambridge, MA), FEN1 (1:50 dilution) (BD Biosciences, San Jose, CA), MCM2, MCM6, SFN (all 1:100 dilution) (Novus Biologicals, Littleton, CO) and TPX2 (1:400 dilution) (Novus Biologicals). The sections were deparaffinized, hydrated, subjected to antigen retrieval by heating in a steamer for 20 min with 10 mmol/L sodium citrate (pH 6.0), and then incubated in peroxidase blocking reagent (DAKO). Sections were then washed with Tris-containing buffer and incubated overnight at 4°C with the primary antibodies. Subsequently, the sections were washed and incubated with secondary antibodies using the Evision plus labeled polymer kit (DAKO) for 30 min followed by incubation with avidin-biotin-peroxidase complex (DAKO) and development with diaminobenzidine chromogen for 5 min. Finally,
Lung adenocarcinoma five-gene signature
the sections were rinsed in distilled water, counterstained with hematoxylin (DAKO), and mounted on
glass slides before evaluation under the microscope. FFPE samples processed similarly, except for the
omission of the primary antibody, were used as negative controls. Experienced lung cancer pathologists
blinded to the clinical data examined the immunostainings jointly at the same time using light
microscopy to generate one set of readings (P.Y. and I.I.W.). The antigens studied exhibited mainly
nuclear immunoreactivity. The immunostainings were quantified using a four-value intensity score (0,
1+, 2+, and 3+) and the percentage (0-100%) of the extent of reactivity in each core. The final score was
then obtained by multiplying the intensity and reactivity extension values (range, 0-300) as previously
reported (22, 28, 30).

To understand the association of the expression of the studied proteins with the survival of
NSCLC patients, a combined immunoreactivity score for each patient was computed by simple addition
of the individual final immunoreactivity scores for each of the analyzed antigens. Lung adenocarcinoma
and SCC patients were then dichotomized into a high and low expression group using the 50th percentile
(median) cut-off of the combined final immunoreactivity score value. Alternatively for confirmation,
patient samples were clustered by average linkage using Cluster v 2.11 program following median-
centering of the antigens’ immunoreactivity scores. Clusters were then identified and visualized using
TreeView programs (Michael Eisen Laboratory, Lawrence Berkeley National Laboratory and University
of California, Berkeley; http://rana.lbl.gov/EisenSoftware.htm). To further validate the prognostic
relevance of the protein signature, lung adenocarcinomas were randomly divided into a training set
(n=78) and a complete (n=78) or stage I only test set (n=62) and a mortality risk was estimated similar to
what was described above using Cox coefficients in the training set and protein expression within each
set. Statistically significant differences in the survival of the clusters, expression or risk groups were
analyzed by the Kaplan-Meier and log-rank tests in R language environment without and with the
inclusion of and adjustment for age and gender as previously described (27).
Lung adenocarcinoma five-gene signature

Results

Derivation of a five-gene signature predictive of survival in lung adenocarcinoma

We have previously identified genes that are expressed differentially among cells constituting an in vitro model of lung carcinogenesis (22). Functional pathways analysis of genes differentially expressed between the previously studied lung tumorigenic (1170-I) and normal lung epithelial cells aided us to identify a significantly modulated gene-interaction network comprised of six key genes based on level of modulation and number of interactions with neighboring molecules (22). Here, we sought to test the relevance of these six genes to NSCLC prognosis. The lung adenocarcinoma microarray datasets from the NCI Director’s challenge study (Shedden et al) (19) were used as a training set (n=442). Lung adenocarcinomas from the Duke (15) and Harvard (10) cohorts served as a combined adenocarcinoma validation set, whereas SCCs from the studies by Raponi et al (17) and Bild et al (15) served as an SCC test set (Figure 1A). An LOOCV approach, described in more detail in the Methods section of this manuscript, was used to train the six genes based on class separation of high versus low expression identified by cluster analysis (not shown). The number of genes was reduced to five following application of a t-test p<10^-5 statistical cut-off to minimize misclassification during LOOCV (Figure 1A) giving rise to a five-gene signature which we have designated as FILM. The FILM signature was efficacious as indicated by the sensitivity and specificity of the six prediction algorithms (Supplementary Table 1), which demonstrated that overall survival of all stages (p<0.001) and stage-I only (p<0.001) lung adenocarcinoma patients predicted to express higher levels of the FILM signature was significantly poorer compared to patients with lower expression of FILM using the linear-discriminator analysis method (LDA) for cross-validation (Figure 1B). Similar findings were obtained using the other methods outlined in Supplementary table 1 (data not shown).

We integrated genes (n=584) we had previously found to be differentially expressed between normal and lung tumorigenic cells (22) with the Shedden et al dataset which in turn subdivided patients into two clusters with significant differences in survival (data not shown). To compare to the
Lung adenocarcinoma five-gene signature performance of the FILM signature, we also analyzed the 584 genes using similar approaches and derived a five-gene signature by recursive feature selection and a fold difference in expression of at least two between two classes identified by cluster analysis (data not shown). The recursive feature-generated five-gene signature exhibited reduced sensitivity and specificity compared to the FILM signature (data not shown). These findings demonstrate the effectiveness of the FILM signature, despite being selected a priori, in predicting the survival of lung adenocarcinoma patients.

The FILM signature does not predict survival of lung SCC patients

We then examined the capacity of the FILM signature to predict survival in lung SCC. A similar strategy to that depicted in Figure 1A was employed except that a pooled SCC test set (n=183) was used. All six prediction algorithms depicted the inability of the FILM signature to predict survival in lung SCC (Figure 1C and data not shown). Similar results were obtained when only lung SCC patients were used in the training and test sets (data not shown). These findings exemplify the prognostic specificity of the FILM signature for lung adenocarcinoma.

Mortality risk assessed by the FILM expression signature predicts survival in all stages or stage-I lung adenocarcinoma

We then sought to further validate the robustness of the FILM expression signature in predicting survival in lung adenocarcinoma. We used a strategy similar to what was described before (26) to estimate mortality risk based on computation of risk scores. Lung adenocarcinomas from the Shedden et al (19) study were used as a training set whereas adenocarcinomas from the Harvard (10) and Duke (15) cohorts were pooled as a test set (Figure 2A). We developed risk scores for patients using the Cox regression coefficients of the genes in the FILM signature and their normalized expression data in the training cohort and patients were dichotomized by the median (50%) risk score (Figure 2B) or were divided into tertiles and three risk groups (low, intermediate and high; Supplementary Figure 1) as
Lung adenocarcinoma five-gene signature described before (19). Lung adenocarcinoma patients identified to be at high risk based on FILM signature exhibited significantly worse survival than patients at low risk ($p=5.9 \times 10^{-5}$) (Figure 2C). For validation of the risk score model, Cox regression coefficients and dichotomization cut-off threshold generated from the training cohort were directly applied to the validation set ($n=183$). All stages ($p<0.001$) or stage-I ($p=0.001$) lung adenocarcinoma patients predicted to be at high risk displayed significantly worse survival (Figure 2D). Similar results were obtained when patients in the training sets were divided into tertile risk groups (Supplementary Figure 1) along with improved capacity of the FILM transcript signature to separate stage-I but not all stages lung adenocarcinoma patients with poor survival from those with excellent survival (Supplementary Figures 1C and 1D). Multivariate Cox proportional hazard regression analyses of patients divided into three risk groups revealed that mortality risk assessed by the FILM transcript signature was, along with stage, an independent predictor of survival ($p=0.03$, HR=1.8, 95% CI=1.06-3.1) in all stages lung adenocarcinoma patients (Supplementary Table 2). Importantly, in stage-I only patients, high risk estimated by the FILM signature was an independent predictor of survival ($p<0.001$ HR=4.5, 95% CI 1.98-10.16) similar in significance to stage IA versus IB disease ($p=0.001$, HR=0.135, 95% CI 0.04-0.45) (Supplementary Table 2). Survival probability analysis was then adjusted by inclusion of clinical covariates, age and gender, in all stages and stage-I test set patients similar to what was described earlier by Shedden et al (19). Inclusion of age and gender as covariates, enhanced the capacity of the FILM transcript signature to separate patients in the test set with good survival from those with poorer survival more notably when analyzing events after seventy months and in particular in stage-I patients (Supplementary Figures 2B and 2C). In contrast, the FILM signature performed similarly with or without clinical covariates when analyzed in the Director’s challenge study utilized as a training set (Supplementary Figure 2A).

We then asked whether there is an association between mortality risk assessed by the FILM signature and survival in stage IA and stage IB patients separately. Stage IB but not IA lung adenocarcinoma patients, from the datasets of the NCI Director’s challenge, Duke and Harvard cohorts...
Lung adenocarcinoma five-gene signature combined, at high risk of mortality as predicted by the FILM signature exhibited significantly poorer overall survival than patients at low risk (p<0.001, HR=2.1, 95% CI=1.4-3.2) (Supplementary Figure 3A). All stage IA and IB patients were included in the analysis as the therapy status of patients in the Duke and Harvard cohorts is not available. We found similar results, albeit less effective, when we analyzed stage IA and IB lung adenocarcinoma patients from the NCI Director’s challenge datasets dataset that are known to have not been treated with any form of therapy (Supplementary Figure 3B). These findings further validate the robustness of the FILM expression signature in predicting survival in lung adenocarcinoma and demonstrate the signature’s capacity in identifying a subpopulation of stage-I patients with poor prognosis.

The corresponding FILM protein signature also predicts survival of non-treated all stages or stage-I lung adenocarcinoma patients

We then explored the prognostic capacity of the FILM signature in NSCLC at the protein level because the expression of transcripts and corresponding encoded proteins do not always correlate (31). We analyzed the protein expression of the FILM signature by immunohistochemistry analysis in FFPE histological tissue specimens obtained from 156 and 79 lung adenocarcinoma and SCC patients, respectively, and who did not receive any therapy before or after tumor resection (Figure 3A). A combined total immunoreactivity score for FILM protein expression was computed, and patients were then dichotomized based on the median FILM protein score. In accordance with the findings obtained with the transcript signature, all-stages (p<0.001, HR=3.6, 95% CI=1.9-6.8) or stage-I only (p<0.001, HR=3.5, 95% CI=1.7-7.5) lung adenocarcinoma patients with higher expression of FILM protein exhibited significantly poorer survival compared to patients with lower expression. In contrast, the protein signature was again not prognostic in lung SCC (Figure 3B). To confirm the prognostic capacity of the FILM protein signature, all stages (Figure 3C) or stage-I (Figure 3D) lung adenocarcinoma patients were analyzed by hierarchical cluster analysis by average linkage based on the centered
Lung adenocarcinoma five-gene signature expression of the immunoreactivity scores of FILM protein expression. Two clusters were identified with dissimilar expression of FILM protein. All stages or stage-I only lung adenocarcinoma patients in the cluster with higher FILM protein (high cluster) exhibited significantly poorer survival compared to patients in the low FILM protein cluster (both p<0.001) (Figures 3C and 3D).

We then determined to further confirm the robustness of the FILM protein signature in predicting the survival of lung adenocarcinoma patients. Lung adenocarcinoma patients were randomized into a training (n=78) set and an all-stages (n=78) or stage-I only (n=62) test set (Figure 4A). As performed with the gene expression signature, risk scores were computed based on the FILM proteins centered immunoreactivity scores and Cox coefficients in the training set (Figure 4A). Patients were then dichotomized based on the median (50%) risk score into high versus low risk groups (Figure 4B) or into tertiles (low, intermediate and high risk groups; data not shown) and subsequently analyzed for survival differences. Lung adenocarcinoma patients in the training set and predicted to be at high mortality risk based on FILM protein exhibited significantly poorer survival than patients at low risk (p=0.01) (Figure 4C). Cox regression coefficients and dichotomization cut-off threshold generated from the training set were then directly applied to the test set (n=78) or only stage-I patients within the set (n=62) and analyzed similarly. In accordance, all stages (p=0.001, HR=4.0, 95% CI=1.6-9.8) or stage-I (p=0.01, HR=3.4, 95% CI=1.4-9.1) lung adenocarcinoma patients predicted to be at high mortality risk based on FILM protein exhibited significantly poor survival (Figure 4D and Table 1). Separation of patients by the FILM protein risk score into tertiles did not increase its capacity for survival prediction (data not shown). Moreover, multivariate Cox hazard regression analyses demonstrated that risk assessed by FILM protein was an independent predictor of survival in all stages (p=0.005, HR=5.78, 95% CI=1.7-20) almost similar in connotation to the most significant variable, age (p=0.002, HR=1.1, 95% CI=1.03-1.15) (Table 1). In addition, FILM protein risk was also an independent, and better than IB stage, predictor of survival in stage I patients (p=0.01, HR=6, 95% CI=1.4-26.2) (Table 1). Inclusion of clinical covariates (age and gender) enhanced the capacity of the FILM protein signature to separate
Lung adenocarcinoma five-gene signature patients with poor survival from those with excellent survival (Supplementary Figure 4). When comparing at 50 months follow up and onwards and following adjustment of survival probability plots for age and gender, no events or deaths were noted in low risk patients in contrast to survival analysis without the clinical covariates.

Since the FILM protein signature was effective in stage-I lung adenocarcinoma prognosis, we tested its prognostic capacity in stage IA and stage IB patients separately. When patients were dichotomized based on mortality risk computed by FILM FFPE protein signature, stage IB but not stage IA patients with higher risk exhibited significantly poorer survival than patients at low risk (p<0.001, HR=4.2, 95% CI=1.8-9.9) (Figure 5A). Similarly, when patients were divided into tertile groups of low, intermediate and high risk, FILM protein signature was able to further separate stage-IB patients with poor survival from those with excellent survival (p<0.001, HR=7.26, 95% CI=2.40-21.98) (Figure 5B). These results demonstrate that the FILM protein signature, like its transcript version, is valuable for predicting the survival of lung adenocarcinoma patients and that the protein signature may be valuable for identifying stage-I or -IB patients who may benefit from adjuvant therapy.

Discussion

We previously identified genes differentially expressed among cells constituting an *in vitro* model of lung carcinogenesis (22). Functional pathways analysis of the differentially expressed genes highlighted a significantly modulated gene-interaction network comprised of six key genes, *UBE2C*, *MCM2*, *MCM6*, *FEN1*, *TPX2* and *SFN*, and that were associated with poor survival in lung adenocarcinoma. In this study, we sought to validate and further examine the prognostic capacity of these genes using multiple methods, various publicly available expression datasets of lung adenocarcinoma and SCCs as well as an additional set of FFPE NSCLCs. We derived a five-gene signature, FILM (*UBE2C*, *MCM2*, *MCM6*, *FEN1*, *TPX2*), that was predictive of poor survival in lung adenocarcinoma including those with stage-I disease. In contrast, the FILM signature was not prognostic
Lung adenocarcinoma five-gene signature in lung SCCs. In addition, we developed a risk model based on the FILM signature with and without inclusion of clinical covariates (age and gender) and demonstrated that FILM transcript signature mortality risk score was predictive of poor survival in lung adenocarcinoma. Moreover, we analyzed and validated the expression of the protein version of the FILM classifier by immunohistochemistry in a series of FFPE NSCLCs and found that FILM protein signature, like the transcript signature, was associated with poor survival in lung adenocarcinomas and not SCCs. Furthermore, risk assessed by the FILM immunohistochemical protein signature was a significant predictor of poor survival in all stages or stage-I lung adenocarcinomas. Lastly, we demonstrated that the FILM protein signature effectively identified a subset of non-treated stage IB lung adenocarcinomas with poor prognosis that may benefit from adjuvant therapy.

The robustness of the FILM transcript signature was validated by deriving a risk score model using the genes’ Cox coefficients and expression in the training set similar to what was described before for predicting recurrence in tamoxifen-treated node-negative breast cancer patients (26). Following direct application of the same Cox coefficients and dichotomization threshold from the training set onto an independent test set, mortality risk assessed by the FILM expression signature was an independent predictor of poor survival in all stages or stage-I lung adenocarcinomas. It is worthwhile to mention that mortality risk assessed by FILM signature was significantly predictive of poor survival regardless of the identity of the training and validation cohorts (data not shown). Moreover, these findings were not replicated in lung SCCs. It is unclear why the FILM signature is only prognostic in lung adenocarcinoma but not SCCs. One possible explanation is that the molecular constituents of the FILM signature are significantly up-regulated in 1170-I lung adenocarcinoma forming cells (32) compared to normal bronchial epithelial cells (22). However, one cannot neglect the significant expression of the genes in lung SCCs. For example, TPX2 protein was shown to be up-regulated in preneoplastic lesions representing lung SCC pathogenesis as well as in tumors (33). In addition, we have previously reported the significant up-regulation of UBE2C immunohistochemical protein in lung SCCs compared to normal.
Lung adenocarcinoma five-gene signature bronchial epithelia (22). Interestingly, UBE2C protein immunohistochemical expression was significantly higher in lung SCCs compared to adenocarcinomas (22). It is plausible to suggest that increased expression of the FILM signature renders lung adenocarcinomas but not SCCs more clinically aggressive.

The prognostic effectiveness of the FILM expression signature was validated at an additional level by assessing the immunohistochemical expression of proteins encoding the FILM classifier in an independent set of FFPE NSCLCs. We deemed this approach to be important as the levels of proteins need not to match with the expression level of transcripts (31, 34). In clear accordance to the transcript signature, the FILM protein signature was prognostic in lung adenocarcinoma (all stages or stage-I) but not in SCC patients and an independent predictor of poor survival in stage-I lung adenocarcinomas. However, it is noteworthy that the semi-quantitative and subjective nature of immunohistochemistry analyses, unlike automated quantitative determination of histological protein expression (35), poses a limitation on using an immunohistochemical protein signature for prognostic purposes such as independent validation by other groups or studies. Moreover, a shortcoming of a protein prognostic classifier from paraffin embedded tissues is the possible variability in readings among different pathologists using the same or different tissue sets. Nevertheless, we attempted to overcome this potential limitation by averaging the score of three tissue cores representing three different sites per tumor specimen to generate final immunohistochemical scores. Moreover, we demonstrated the prognostic capacity of the FILM protein signature using various analytical methods; patient dichotomization based on median immunohistochemical score, hierarchical clustering of patients and derivation of a risk score model. In addition, the prognostic capacities of each protein within the FILM signature were analyzed independently and deemed in all cases weaker in comparison to the combined FILM protein signature. Furthermore and importantly, the results obtained using the transcript and protein versions of the FILM signature were nearly identical demonstrating the reliability and biological significance of this signature in lung adenocarcinoma prognosis.
Lung adenocarcinoma five-gene signature

We had previously derived global differential expression profiles between normal and tumorigenic lung epithelial cells that are also widely expressed in lung adenocarcinomas from the Shedden et al dataset (22). It is plausible that different five-gene signatures developed from all genes differentially expressed among the in vitro model of lung carcinogenesis may as well be equally or more efficacious in prognosis compared to the FILM signature. To test this hypothesis, we derived a five-gene signature by recursive feature selection and a fold difference in expression of at least two between the classes identified by cluster analysis of the Shedden et al dataset following integration of the genes (data not shown). The FILM signature, despite being originally based on genes selected a priori, was superior to the recursive feature-generated five-gene signature in sensitivity and specificity as well as in predicting poor survival in lung adenocarcinoma. The effectiveness of the FILM signature in lung adenocarcinoma prognosis may be due to the nature of its genes that are proliferation-related despite acting at different points and through distinct mechanisms in cell cycle control (33, 36-42). It is noteworthy that in a large meta analysis using breast tumors from several publicly available breast cancer expression datasets, Wirapati et al demonstrated that all nine prognostic signatures compared exhibited similar prognostic capacities which were mostly driven by proliferation-related genes (43, 44). Moreover, in the NCI Director’s Challenge study by Shedden et al, the most effective gene classifiers in survival prediction (methods A and H as designated by the authors) were mainly comprised of proliferation related genes and performed well with or without clinical covariates. Interestingly, all genes within the FILM signature were found in methods A and H developed by Shedden et al which may explain why the FILM transcript signature performed similarly with or without clinical covariates when analyzed in the Director’s Challenge study.

In recent reviews of studies geared towards the development of prognostic expression signatures in NSCLC, Subramanian and Simon have suggested guidelines to follow in the design and testing of an expression signature for lung cancer prognosis including the stratification of stage IA and stage IB NSCLC patients independently in an effort to identify patients who may benefit from adjuvant therapy.
Lung adenocarcinoma five-gene signature (6, 43, 44). Towards this, we had analyzed stage IA and stage IB patients alone using both the transcript and protein FILM signatures in publicly available microarray datasets and a tissue microarray of FFPE NSCLC specimens, respectively. We found that the FILM signature identified a subgroup of stage IB but not stage IA lung adenocarcinoma patients with poor prognosis, which was consistent as it was found by analyzing the transcript signature in publicly available microarray datasets and the parallel protein signature in a tissue microarray of FFPE specimens. However, it is noteworthy that we analyzed, as depicted in Supplementary Figure 3, the entire set of stage IA and IB lung adenocarcinomas from the Director’s challenge (19), Duke (15) and Harvard (10) datasets as the therapy status for all the patients from the two latter cohorts and some from the former is not known. Nevertheless, when analyzed only in patients from the Director’s challenge dataset that were not treated with any form of therapy, the FILM transcript signature was also prognostic, albeit less, in stage IB lung adenocarcinoma. Furthermore, we analyzed the FILM protein signature in a FFPE tissue microarray with more complete and annotated clinicopathological information and found that the FILM protein signature effectively identifies non-treated stage IB lung adenocarcinoma patients with poor prognosis pinpointing to a potential benefit of adjuvant treatment in this patient population. It is worthwhile to mention that the FILM protein signature, in comparison to its transcript counterpart, was more effective in identifying stage-I and in particular -IB lung adenocarcinoma patients with very poor survival and separating them from patients with excellent survival. Therefore, it is plausible to assume that the FILM protein signature may be clinically more useful for stage-I or specifically –IB lung adenocarcinoma prognosis. However and as mentioned before, it cannot be neglected that the relatively more qualitative nature of immunohistochemistry analysis of FFPE tissues and more difficult external validation compared to real-time PCR assessment of RNA poses a limitation on the translation of FFPE protein signatures to the clinic. Unless the capacity of the FILM protein signature is cross-validated in independent tissue microarrays and by different investigators, a combination of both mRNA and protein signatures may be powerful for stage-I lung adenocarcinoma prognosis.
Lung adenocarcinoma five-gene signature

In conclusion, our study describes the development and testing of an effective five-gene signature in lung adenocarcinoma prognosis. This expression signature has several unique and valuable attributes; 1) it identifies stage-I lung adenocarcinomas with poor prognosis; 2) exhibits prognostic specificity towards lung adenocarcinomas only; 3) the protein variant of the signature can effectively predict survival in all stages or stage I lung adenocarcinoma following analysis of FFPE tissue specimens by immunohistochemistry; and 4) the protein FFPE variant is effective in identifying non-treated stage-IB lung adenocarcinoma patients with dismal prognosis that are most likely to benefit from adjuvant therapy. Therefore, further independent studies are warranted to externally validate the potential clinical use of this five-gene signature, and in particular, the prognostic capacity of the FILM protein FFPE signature towards translation to the clinic for identifying non-treated stage-I and specifically IB lung adenocarcinoma patients with poor survival that will benefit from adjuvant therapy.
References

Lung adenocarcinoma five-gene signature


Lung adenocarcinoma five-gene signature


Lung adenocarcinoma five-gene signature


Lung adenocarcinoma five-gene signature

Figure legends

Figure 1. Development of a five-gene signature, FILM, effective in predicting the survival of all stages or stage-I lung adenocarcinoma patients. A. Derivation and validation of the FILM expression signature. Lung adenocarcinomas from the NCI Director’s challenge study were used as a training set (n=442) and adenocarcinoma samples from the Duke and Harvard cohorts were pooled as an adenocarcinoma test set (n=183). Lung SCCs from the studies by Bild et al and Raponi et al were pooled as an SCC test set (n=183). Gene selection by the prediction algorithms is detailed in the Materials and Methods section. B. All stages (n=183, left panel) or stage-I lung adenocarcinomas (n=110, right panel) and lung SCCs (n=183) (C) were dichotomized into groups of high (red) versus low (blue) expression of FILM signature expression as predicted by the linear discriminator analysis (LDA) method. Statistically significant differences in survival between the groups were assessed by the Kaplan-Meier method of estimation of survival probability and the long-rank test.

Figure 2. Risk assessed by FILM signature predicts poor survival in all stages or stage-I lung adenocarcinoma. A. Development of FILM risk score model. Lung adenocarcinomas from the NCI Director’s challenge study were used as training set (n=442) whereas samples from the Duke and Harvard cohorts were pooled as a test set. Patient risk scores were computed as detailed in the Materials and Methods section. B. Patients in the training set were dichotomized into high versus low risk groups based on the median (50%) value (dashed line). C. Kaplan-Meier plots depicting survival differences between high (black) and low (grey) risk groups in the training set. D. Gene Cox coefficients and dichotomization threshold from the training set were directly applied onto all stages and stage-I lung adenocarcinomas in the test set to separate patients into high (black) and low (grey) risk groups.

Figure 3. FILM immunohistochemical protein signature is significantly associated with poor survival. A. Representative photomicrographs depicting immunohistochemical expression of the
Lung adenocarcinoma five-gene signature indicated antigens in FFPE histological tissue specimens of lung adenocarcinoma. **B.** Following analysis of FILM protein expression by immunohistochemistry, lung adenocarcinomas ($n=156$) and SCCs ($n=79$) were dichotomized into groups with high (red) versus low (blue) expression of FILM protein based on the combined immunoreactivity score of the antigens. For further validation of the FILM protein prognostic signature, all stages ($n=156$) \((C)\) or stage I ($n=123$) \((D)\) lung adenocarcinomas were clustered following median-centering of the antigens’ immunoreactivity scores, after which survival differences between the high (red) and low (blue) FILM protein clusters were analyzed.

**Figure 4. Mortality risk by FILM protein signature predicts poor survival in non-treated all stages or stage-I lung adenocarcinomas.** **A.** Tissue microarray lung adenocarcinomas were randomly divided into a training ($n=78$) and all stages ($n=78$) or stage I test set ($n=62$) and risk scores were computed as detailed in the Materials and Methods section. **B.** Training set lung adenocarcinomas were then dichotomized into high and low risk groups based on the median (50%) value and analyzed for survival differences \((C)\). **D.** The univariate Cox coefficients and dichotomization threshold from the training set were directly applied onto the all stage and stage-I lung adenocarcinoma test sets to separate patients similarly into high (black and low (grey) risk groups which were then analyzed for differences in survival.

**Figure 5. FILM protein signature predicts poor survival in non-treated stage-IB lung adenocarcinoma.** Non-treated stage IA or stage IB tissue microarray FFPE lung adenocarcinomas were dichotomized into groups of \((A)\) high (black) versus low (grey) risk or \((B)\) divided into three groups of high (black), intermediate (dotted), low (grey) risk as predicted by the FILM protein signature. Survival differences between the different groups were statistically assessed by the Kaplan-Meier method and log-rank test.
Fig. 4

A

Training set (N=78)

risk-score prediction model

Entire validation set (N=78)

Stage-I only validation set (N=62)

Survival analysis

B

High risk

Low risk

Risk Score

5 4 3 2 1 0 1 2 3 4

D

stage I only

all stages

N=31

N=39

N=31

N=39

OS (months)

p=0.01

p=0.001

0.0 0.2 0.4 0.6 0.8 1.0

0 1 2 3 4 5 6 7

C

Probability

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8

0 10 20 30 40 50 60 70

OS (months)

p=0.01

Copyright © 2010 American Association for Cancer Research.
Table 1

FILM protein signature risk is an independent predictor of poor survival in non-treated all stage or stage-I lung adenocarcinoma

<table>
<thead>
<tr>
<th>All stages</th>
<th>Univariate</th>
<th></th>
<th>Multivariate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H.R. (95% C.I.)</td>
<td>p-value</td>
<td>H.R. (95% C.I.)</td>
<td>p-value</td>
</tr>
<tr>
<td>High risk</td>
<td>3.998 (1.575-9.846)</td>
<td>0.001</td>
<td>5.776 (1.668-20.002)</td>
<td>0.005</td>
</tr>
<tr>
<td>Stage III vs I, II or IV</td>
<td>2.505 (0.731-8.583)</td>
<td>0.144</td>
<td>7.011 (1.467-33.506)</td>
<td>0.018</td>
</tr>
<tr>
<td>Poor differentiation</td>
<td>1.401 (0.580-3.383)</td>
<td>0.453</td>
<td>0.814 (0.176-3.762)</td>
<td>0.793</td>
</tr>
<tr>
<td>Solid subtype</td>
<td>1.007 (0.994-1.020)</td>
<td>0.275</td>
<td>1.003 (0.983-1.023)</td>
<td>0.777</td>
</tr>
<tr>
<td>BAC subtype</td>
<td>0.997 (0.983-1.012)</td>
<td>0.71</td>
<td>1.014 (0.995-1.034)</td>
<td>0.14</td>
</tr>
<tr>
<td>Never vs ever smoker</td>
<td>0.969 (0.386-2.432)</td>
<td>0.946</td>
<td>1.366 (0.419-4.453)</td>
<td>0.605</td>
</tr>
<tr>
<td>Male gender</td>
<td>1.816 (0.823-4.003)</td>
<td>0.139</td>
<td>1.226 (0.491-3.062)</td>
<td>0.662</td>
</tr>
<tr>
<td>Age</td>
<td>1.032 (0.987-1.079)</td>
<td>0.161</td>
<td>1.088 (1.031-1.149)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage I</th>
<th>Univariate</th>
<th></th>
<th>Multivariate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H.R. (95% C.I.)</td>
<td>p-value</td>
<td>H.R. (95% C.I.)</td>
<td>p-value</td>
</tr>
<tr>
<td>High risk</td>
<td>3.359 (1.392-9.110)</td>
<td>0.011</td>
<td>6.049 (1.399-26.154)</td>
<td>0.01</td>
</tr>
<tr>
<td>Stage IB vs IA</td>
<td>2.556 (0.917-7.124)</td>
<td>0.072</td>
<td>3.154 (0.977-10.191)</td>
<td>0.055</td>
</tr>
<tr>
<td>Poor differentiation</td>
<td>2.135 (0.773-5.900)</td>
<td>0.143</td>
<td>1.034 (0.187-5.722)</td>
<td>0.97</td>
</tr>
<tr>
<td>Solid subtype</td>
<td>1.010 (0.996-1.025)</td>
<td>0.161</td>
<td>0.997 (0.972-1.026)</td>
<td>0.921</td>
</tr>
<tr>
<td>BAC subtype</td>
<td>0.996 (0.998-1.012)</td>
<td>0.646</td>
<td>1.007 (0.988-1.026)</td>
<td>0.459</td>
</tr>
<tr>
<td>Never vs ever smoker</td>
<td>0.728 (0.210-2.517)</td>
<td>0.616</td>
<td>0.770 (0.148-4.024)</td>
<td>0.757</td>
</tr>
<tr>
<td>Male gender</td>
<td>2.210 (0.871-5.604)</td>
<td>0.095</td>
<td>1.940 (0.597-6.289)</td>
<td>0.269</td>
</tr>
<tr>
<td>Age</td>
<td>1.037 (0.984-1.094)</td>
<td>0.173</td>
<td>1.060 (0.990-1.135)</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Univariate and multivariate Cox proportional Hazard regression analysis
H.R., hazard risk; C.I., confidence interval
A five-gene and corresponding-protein signature for stage-I lung adenocarcinoma prognosis

Humam Kadara, Carmen Behrens, Ping Yuan, et al.

Clin Cancer Res  Published OnlineFirst December 16, 2010.