Supplementary Materials and Methods

Independent validation of published signatures of Chen et al. and Lau et al.

The datasets used for the evaluation were stage I samples in the combined UM, HLM, MSK, and CAN/DF datasets reported by Shedden et al. (1). The signatures of Chen et al. and Lau et al.were originally developed using real time quantitative polymerase chain reaction data. We used the same pre-processing steps the authors reported for validating their signature on microarray data for this independent evaluation.

The five-gene signature of Chen et al.

All probe sets on the Affymetrix U133A array that detected the same gene names reported in Chen et al. (2) were identified from the Affymetrix annotation files (http://www.affymetrix.com), resulting in nine probe sets for the five genes and one probe set corresponding to the control gene, *TBP* . The probe sets identified were:

Gene symbol	**Gene name**	**Probe set**
DUSP6	Dual specificity phosphatase 6	208891_at, 208892_s_at, 208893_s_at
ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	202454_s_at
LCK	Lymphocyte-specific protein tyrosine kinase	204890_s_at, 204891_s_at
MMD	Monocyte to macrophage differentiation-associated	203414_at
STAT1	Signal transducer and activator of transcription 1	200887_s_at, 209969_s_at
TBP	TATA box binding protein	203135_at

Probe sets for the same gene were collapsed by averaging. The data was then log-transformed to base-2 scale after assigning a value of 1.1 to intensity values less than 1.1. The levels of expression of the five genes were then divided by the expression level of the control *TBP* gene to calculate relative levels of expression. The decision tree model reported by Chen et al. (2) was applied to these relative expression levels. Kaplan–Meier survival curves were constructed separately for stage IA and IB samples.

The three-gene signature of Lau et al.

Four housekeeping genes—*TBP*, *ACTB*, *B2M*, and *BAT1*—were used for the normalization. The probe sets corresponding to the three genes in the signature by Lau et al. (3) and the four housekeeping genes for normalization were identified from the Affymetrix annotation files (http://www.affymetrix.com) as:

Gene symbol	**Gene name**	**Probe set**
STX1A	Syntaxin 1A (brain)	204729_s_at
CCR7	Chemokine (C-C motif) receptor 7	206337_at
HIF1A	Hypoxia inducible factor 1, alpha subunit	200989_at
TBP	TATA box binding protein	203135_at
ACTB	Actin, beta	200801_x_at
B2M	Beta-2-microglobulin	201891_s_at
BAT1	HLA-B associated transcript 1	200041_s_at

The expression values were log-transformed to base-2 scale after adding a pseudo-count of 10. The normalization factor was calculated as the mean of the four housekeeping genes. Normalized expression values were obtained by subtracting this normalization factor from the initial expression levels. The normalized expression values were further median centered and then standardized to mean 0 and variance 1. The risk score and the risk groups were then calculated

on these standardized expression levels as outlined in Lau et al. (3). Kaplan–Meier survival curves were constructed separately for stage IA and IB samples. The survival curves were compared using the log-rank test.

Prognostic Models That Use Clinical Covariates

The training and the test datasets were the same as those reported by Shedden et al. (1), ie, University of Michigan Cancer Center (UM) and Mofitt Cancer Center (HLM) data for training and Memorial Sloan-Kettering Cancer Center (MSK) and Dana-Farber Cancer Institute (CAN/DF) data for validation. Only stage I data were considered. The training data were used to build a Cox proportional hazards regression model with age, stage (IA or IB), and adjuvant chemotherapy (yes or no) as covariates. The assumption of proportionality for all covariates was verified by testing for the statistical significance of correlations between the scaled Schoenfeld residuals and time and by graphically examining the scaled Schoenfeld residuals. The model was then evaluated on the test datasets. The median risk score was calculated for the training set and was used as the cutoff to stratify the patients in the test set into low- and high-risk groups. Kaplan–Meier survival curves were constructed for the high- and low-risk groups. The survival curves were compared using the log-rank test. All reported *P* values are two-sided. All analyses were performed by using R software version 2.8.0 (4).

Simulation Study

Lung cancer survival times for 129 patients were obtained from Bild et al. (5). Random gene expression profiles for these patients were generated from a standard normal distribution. A total of 5000 gene expression values were generated for each sample using R version 2.8.0 software

(4). Sixty percent of the patients were randomly assigned to the training set and the rest became part of the validation set. The entire simulation was repeated 10 times using different divisions of the data into training and validation sets.

No gene filtering or normalization was applied; all 5000 genes were used to build the model. BRB ArrayTools version 3.7.1 [http://linus.nci.nih.gov/BRB-ArrayTools.html; developed by Dr. Richard Simon and the BRB-Array Tools Development Team (6)] and R version 2.8.0 (4) was used for model development and analysis.

The algorithm used was the Survival Analysis Prediction Tool from BRB ArrayTools. Genes whose expression was statistically significantly associated with survival at $P \leq .001$ were selected by fitting Cox proportional hazards models to each gene in the training data. A Cox proportional hazards model was built using the principal components calculated from the statistically significant gene list using BRB ArrayTools. The principal components are combinations of the individual genes with coefficients determined from the training set. The smallest number of principal components that explained 75% of the variation in the statistically significant genes was used for the model building. The median risk score from the training data was used to stratify patients into high- or low-risk groups. Kaplan–Meier survival curves for the high- and low-risk groups were then computed and compared using the log-rank test. All reported *P* values are two-sided. R version 2.8.0 was used to construct and compare the Kaplan–Meier curves.

References

1. Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden K, Taylor JM, Enkemann SA, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med. 2008; 14(8):822-7.

2. Chen HY, Yu SL, Chen CH, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med. 2007; 356(1):11-20.

3. Lau SK, Boutros PC, Pintilie M, et al. Three-gene prognostic classifier for early-stage non small-cell lung cancer. J Clin Oncol. 2007; 25(35):5562-9.

4. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

5. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA Jr, Marks JR, Dressman HK, West M, Nevins JR. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006; 439(7074):353-7.

6. Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-ArrayTools. Cancer Inform. 2007; 2:11-17.