Robust Gene Expression Signature from Formalin-Fixed Paraffin-Embedded Samples Predicts Prognosis of Non-Small-Cell Lung Cancer Patients

Yang Xie1,2, Guanghua Xiao1, Kevin R. Coombes6, Carmen Behrens7, Luisa M. Solis8, Gabriela Raso8, Luc Girard4,5, Heidi S. Erickson7, Jack Roth9, John V. Heymach7, Cesar Moran8, Kathy Danenberg10, John D. Minna2,3,4,5, Ignacio I. Wistuba7,8.

Authors’ Affiliations: Departments of 1Clinical Sciences, 2Simmons Cancer Center, 3Internal Medicine, 4Pharmacology, and 5Hamon Center for Therapeutic Oncology Research The University of Texas Southwestern Medical Center, Dallas, Texas; Departments of 6Bioinformatics and Computational Biology, 7Thoracic/Head and Neck Medical Oncology, 8Pathology and 9Thoracic and Cardiovascular Surgery, The University of Texas M. D. Anderson Cancer Center, Houston, Texas; and 10Response Genetics, Inc, Los Angeles, California.

Grant support: This study was supported in part by grants from the Department of Defense (W81XWH-07-1-0306 03 to J.D.M., and I.I.W.), the Specialized Program of Research Excellence in Lung Cancer Grant (P50CA70907 to J.D.M., J. R., and I.I.W.), the National Cancer Institute (1R01CA152301-01 to Y.X. and I.I.W., Cancer Center Support Grant CA-16672), National Institute of Health (NIH UL1 RR024982 to G.X).

Running title: Robust prognosis signature from fixed tissue

Abbreviations: Non-small-cell-lung-cancer: NSCLC; Formalin-fixed paraffin-embedded: FFPE; Robust gene set: RGS

Key words: Lung Cancer Prognosis, Gene Expression Signature, Formalin Fixed Paraffin Embedded Samples

Requests for reprints: Ignacio I. Wistuba, Departments of Pathology and Thoracic/Head and Neck Medical Oncology, Unit 85, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030. Phone: 713-563-9184; Fax: 713-792-0309; E-mail: iiwistuba@mdanderson.org.
Statement of Translational Relevance.

This paper is the first study to develop a robust prognosis signature for non-small cell lung cancer (NSCLC) based on genome-wide expression profiling of clinically available formalin-fixed and paraffin-embedded (FFPE) samples. Although clinical FFPE tumor samples are widely available, the genome-wide expression profiling of FFPE samples has been hampered due to the degradation of RNAs extracted from them. In this paper, we show that NSCLC FFPE-derived signature is strongly associated with patients’ clinical outcome, is independent of clinical prognostic variables, and can be validated in several independent studies. We showed that, after strict quality control and analysis procedures, genome-wide profiling of FFPE samples can actually provide an unique opportunity to identify a set of genes whose expression level is less sensitive to the environmental changes. This gene signature is more robust across different platforms and studies, which is critical for the successful application of gene signatures in real clinical settings.
ABSTRACT

Purpose. The requirement of frozen tissues for microarray experiments limits the clinical usage of genome-wide expression profiling using microarray technology. The goal of this study is to test the feasibility of developing lung cancer prognosis gene signatures using genome-wide expression profiling of formalin-fixed paraffin-embedded (FFPE) samples, which are widely available and provide a valuable rich source for studying the association of molecular changes in cancer and associated clinical outcomes.

Experimental Design. We randomly selected 100 Non-Small-Cell lung cancer (NSCLC) FFPE samples with annotated clinical information from the UT-Lung SPORE Tissue Bank. We micro dissected tumor area from FFPE specimens, and used Affymetrix U133 plus 2.0 arrays to attain gene expression data. After strict quality control and analysis procedures, a supervised principal component analysis was used to develop a robust prognosis signature for NSCLC. Three independent published microarray data sets were used to validate the prognosis model.

Results. This study demonstrated that the robust gene signature derived from genome-wide expression profiling of FFPE samples is strongly associated with lung cancer clinical outcomes, can be used to refine the prognosis for stage I lung cancer patients and the prognostic signature is independent of clinical variables. This signature was validated in several independent studies and was refined to a 59-gene lung cancer prognosis signature.

Conclusions. We conclude that genome-wide profiling of FFPE lung cancer samples can identify a set of genes whose expression level provides prognostic information across different platforms and studies, which will allow its application in clinical settings.
INTRODUCTION

Lung Cancer is the leading cause of death from cancer for both men and women in the United States and in most parts of the world, with a 5-year survival rate of 15%\(^1\). Non-small-cell lung cancer (NSCLC) is the most common cause of lung cancer death, accounting for up to 85% of such deaths\(^2\). Clinical-pathologic staging is the standard prognosis factor for lung cancer used in clinical practice, but does not capture the complexity of the disease so that heterogeneous clinical outcomes within the same stage are commonly seen. Several randomized clinical trials showed that adjuvant chemotherapy improves survival in resected NSCLC\(^3\)\(^-\)\(^7\). The effect of adjuvant chemotherapy on prolonging survival is modest - only 4-15% improvement in 5-year survival, while such treatment is associated with serious adverse effects\(^6\),\(^8\). Therefore, it is of considerable clinical importance to have a robust and accurate prognostic signature for lung cancer, especially in early stage lung cancer to improve the current clinical decisions on whether an individual lung cancer patient should receive adjuvant chemotherapy or not.

Genome-wide expression profiles have been used to identify gene signatures to classify lung cancer patients with different survival outcomes\(^9\)-\(^16\). However, the requirement of frozen tissues for microarray experiments limits the clinical usage of these gene signatures. Furthermore, prognostic gene signatures for NSCLC developed by different groups show minimal overlap, and are often difficult to reproduce by independent groups\(^17\)-\(^18\). To address the problem of requirement for frozen issues, we designed this study to test the feasibility of developing lung cancer prognosis gene signatures using genome-wide expression profiling of formalin-fixed paraffin-embedded (FFPE) samples, which are widely available and provide a valuable rich
source for studying the association of molecular changes in cancer and associated clinical outcomes. We derived a prognosis signature for NSCLC from FFPE samples and validated it in several independent studies. To facilitate other researchers to reproduce all results in this study, we have provided a literate programming R package.

MATERIALS AND METHODS

Tissue specimens. The overall study design and the flow chart of the derivation and validation of the robust gene signature are described in Figure 1. We randomly selected 100 NSCLC FFPE samples with annotated clinical information from the UT-Lung SPORE Tissue Bank from 2001-2005. From these samples, 75 samples passed the mRNA quality control criteria (Supplementary Methods). Among these 75 samples, 48 samples are adenocarcinomas and 27 are squamous cell carcinomas. The median follow-up time is 2.8 years and the maximum follow-up time is 6.9 years; the characteristics of these patients are summarized in Supplementary Table 1. The samples were obtained under approval of the institutional review boards at M.D. Anderson Cancer center.

Sample microdissection and RNA extraction. FFPE tumor specimens were cut into serial sections with a thickness of 10 μm. For the pathological diagnosis, one slide was stained with H&E and evaluated by a pathologist. Other sections were stained with nuclear fast red (NFR, American MasterTech Scientific Inc., Lodi, CA) to enable visualization of histology. Tumor tissue was isolated using manual macro-dissection when the tumor area was > 0.5 x 0.5 mm or laser capture microdissection (P.A.L.M.
Microlaser Technologies AG, Munich, Germany) in cases of smaller tumor areas. At least 50 mm\(^2\) of tumor tissue was collected from each FFPE block. The extraction of RNA from tissue samples was done by a proprietary procedure of Response Genetics, Inc. (United States Patent Application 20090092979) designed to optimize the yield of higher molecular weight RNA fragments from FFPE specimens.

**Microarray data preprocessing and quality control.** Total RNA was processed for analysis on the Affymetrix U133 plus 2.0 arrays according to Affymetrix protocols for first- and second-strand synthesis, biotin labeling and fragmentation. The quality control procedure for microarray data analysis was based on the percentage of present calls calculated by the MAS5 package. We selected arrays with at least 15\% of probe sets present; 55 out of 75 arrays passed this quality control criterion and will be used for the analysis. We selected probe sets that are present on all 55 arrays; 1400 genes past this criterion. These 1400 genes were referred as the robust gene set (RGS) since the mRNA expression of these genes are robust to FFPE processing. The 55 samples and the 1400 genes were used to develop gene signatures.

After microarray analysis QC, we used the RMA background correction algorithm \(^{19}\) to remove non-specific background noise. A robust regression model \(^{20}\) was fitted to the probe level data, and the fitted expression values for the probes at the 3’ end were used to summarize the probe set expression values. Quantile-quantile normalization was used to normalize all the arrays. Consortium microarray raw data \(^{13}\) was downloaded from the National Cancer Institute’s caArray database and preprocessed by RMA background correction and quantile-quantile normalization. All gene-expression values were log-transformed (on a base 2 scale).
**Supervised classification using supervised principal component analysis.**

Classification was performed using supervised principal component analysis \(^{21-22}\), a widely used classification method in biomedical research \(^{23-26}\). As a supervised classification method, each prediction model was trained in a training dataset and then the performance was tested in an independent test dataset. We used an R package (version 2.81), Superpc (version 1.05), to implement the prediction algorithm, and the default parameters were used. The implementation details can be found in the **Supplementary Sweave Report**. The training and testing sets for each prediction model are summarized in **Supplementary Table 2**.

**Survival analysis.** Overall survival time was calculated from the date of surgery until death or the last follow-up contact. Survival curves were estimated using the product-limit method of Kaplan-Meier \(^{27}\) and were compared using the log-rank test. The maximum follow-up time for the FFPE patient cohort is less than 7 years, while some patients in the consortium cohort have been followed for up to 17 years. To avoid the extrapolation of the prediction model, the comparison of survival time between predicted groups are truncated at 7 years. The analysis results without truncation can be seen in **Supplementary Sweave Report**. Univariate and multivariate Cox proportional-hazards analysis \(^{28}\) were also performed, with survival as the dependent variable.
RESULTS

The robust gene set defines two tumor groups. The expression of these 1400 genes divided the 55 patients into two groups based on unsupervised clustering analysis (with euclidean distance and complete linkage for the hierarchical clustering algorithm) (Figure 2). Interestingly, group 1 has significantly shorter survival time compared to group 2 (Figure 2b, HR=3.6, $P=0.017$) and multivariate Cox proportional-hazards analysis showed that the association between RGS groups and survival ($P=0.012$) is independent of stage. Notably, group 1 was dominated by squamous cell carcinoma (23/28), whereas group 2 was dominated by adenocarcinomas (25/27) ($P<0.0001$) (Supplementary Table 3). The other clinical characteristics including gender, age and smoking status were not significantly different between the two groups. To explore whether the association between RGS groups and survival is due to the histology difference between two groups, we drew Kaplan-Meier curves by both histology and RGS groups (Supplementary Figure 1) and it shows clearly that RGS can distinguish high and low risk groups within both adenocarcinoma and squamous groups indicating the association of RGS groups and survival is independent of histology groups.

We used gene set enrichment analysis (GSEA) to identify the enriched gene sets in both RGS groups. Interestingly, an estrogen receptor (ER) negative signature in breast cancer $^{29}$ is enriched in RGS group 1, meanwhile, an ER positive signature in breast cancer $^{29}$ is enriched in RGS group 2 (Figures 2c and 2d) indicating the relationship between the ER signatures and the RGS groups. The other enriched
gene sets are summarized in **Supplementary Table 4**; notably, genes enriched in group 1 are also enriched in mouse neural stem cells and embryonic stem cells.

**Construct and validate RGS prognosis signatures.**

*FFPE samples training to testing.* The strong associations between RGS groups and survival outcomes motivated us to explore whether RGS expression profile can be used to construct prognosis signature. We randomly divided 55 patients into training (25 samples) and testing (30 samples) sets, and constructed a prediction model using 1400 robust gene expression values in the training set through a supervised principal component approach \(^\text{21}\). **Figure 3a** shows that the predicted low risk group has significant longer survival time than the predicted high risk group (*P*=0.013) in the testing set. To test if this association was not random, we randomly split the data into training and testing sets 200 times, repeated the same prediction and testing procedures for each set, and found that the prognosis performance of RGS signature is significantly better than random (*P*=0.02).

*Frozen samples training to testing.* We then tested whether this robust gene set can be used to construct prognosis signature in frozen samples. The largest independent public available lung cancer microarray data set is the recently published NCI Director’s Consortium for study of lung cancer involving 442 resected adenocarcinomas \(^\text{13}\). From that study, Affymetrix U133A microarray data for the 1012 robust genes were excerpted with 388 less genes than our FFPE data due to the microarray platform difference. We used the same training and testing strategy as in the original analyses of these data \(^\text{13}\) for constructing and validating prognosis signature through supervised principal component approach. The training set included...
samples from University of Michigan Cancer Center (UM) and Moffitt Cancer Center (HLM), and the testing set included the Memorial Sloan-Kettering Cancer Center (MSK) and Dana-Farber Cancer Institute (CAN/DF) samples. This analysis revealed that the predicted low risk group has significant longer survival time than the predicted high risk group (HR=2.44, \( P=0.00014 \)) in the testing dataset (Figure 3b).

**FFPE to frozen samples and vice versa.** Next, we used our FFPE and the consortium datasets as frozen samples to investigate whether the prediction model built from one type of sample can be validated in another type of sample. Again, the same supervised principal component method was used to construct the prediction model. The prediction model built from FFPE samples can significantly distinguish the high and low risk groups in frozen samples (Figure 3c, HR= 1.95, \( P=5.4\times10^{-7} \)), and the prediction model built from frozen samples can also distinguish the high and low risk groups in FFPE samples but with marginal significance (Figure 3d, HR= 3.59, \( P=0.068 \)). We also tested the performance of FFPE prediction model on four individual datasets in consortium study and found that the predicted low risk groups have longer survival time compared to the predicted high risk groups for all sets: MSKCC dataset (median survival time 6.5 vs. 3.3 years; HR= 2.31, \( P=0.0093 \)), DFCI dataset (median survival time 5.9 vs. 0.9 years; HR= 2.62 \( P=0.0076 \)), HLM dataset (median survival time 3.4 vs. 2.2 years; HR= 1.25, \( P=0.4 \)) and MI data set (median survival time 5.4 vs. 2.2 years; HR= 1.98, \( P=0.0011 \)) (Supplementary Figure 2). Next, we compared the performance of RGS signature with previous published lung cancer prognosis signatures using the same consortium dataset as testing set. Shedden et al showed that the hazard ratios for Method A signature (the best signature in Shedden et al) and Chen et al {Chen, 2007 #13} signatures range from 1.10-1.83 for MSK test set. While
the hazard ratio for our RGS signature is 2.89 on the same MSK test set. For the CAN/DF test set, the hazard ratios range from 1.76-2.30 using the published signatures, while the hazard ratio for our RGS signature on the same CAN/DF test set is 2.39. Therefore, the prognosis performance of RGS prognosis is at least as good as other published signatures in the microarray dataset.

**The RGS prognosis signature is independent of clinical variables.** To test whether RGS is an independent prognosis signature, we fitted a multivariate Cox regression model including RGS risk scores, age, gender, stage, smoking status, adjuvant chemotherapy usage and clinical sites as co-variables for the consortium data set. The RGS risk scores were calculated from the prediction model built from the FFPE samples set. Table 1 shows that the RGS signature is significantly associated with the survival time after adjusting for other clinical variables (HR=1.3, \( P=0.007 \)). Pathological stages based on international staging system is the most widely used and important prognosis variable for lung cancer patients, here we tested whether RGS signature can further refine the prognosis within each stage. The RGS prognosis signature from FFPE samples was tested within each stage of the consortium dataset. The results show clearly that the RGS signature is significantly associated with survival outcome within each stage (Figure 3e to g; HR= 1.54, \( P=0.036 \) for stage I, HR= 1.81, \( P=0.022 \) for stage II and HR= 1.90, \( P=0.021 \) for stage III), indicating that the RGS signature can refine the prognosis for lung cancer patients. The RGS prognosis signature from FFPE samples was further tested for patients with or without adjuvant chemotherapy separately, and the results show that the RGS signature is significantly associated with survival for both groups (Supplementary Figure 3 a,b; HR=1.95,
P=0.015 for patients with chemotherapy, HR=1.99, P=0.00062 for patients without chemotherapy).

**Refine to 59-gene prognosis signature.** Among all the RGS genes, 131 genes are associated with survival (P<0.05) in the FFPE dataset, and 365 genes are associated with overall survival (P<0.05) in the consortium dataset by univariate Cox regression analysis. There is significant overlap between these two gene lists (Figure 4a; 59 common genes; P=0.0008, hyper-geometric test). More significant genes were found in the consortium data compared to the FFPE data, which is likely due to the larger sample size (n=442) of the consortium dataset compared to the FFPE dataset sample size (n=55). Surprisingly, hazard ratios from the two datasets are very consistent with each other. All 59 genes have the same direction of effects (positive or negative) on the survival between the two data sets and the hazard ratios from two datasets are highly correlated (Pearson’s correlation = 0.86) (Figure 4b), indicating the high consistency of expressions of these genes across datasets. These results motivated us to hypothesize that these 59 genes (Supplementary Table 5) alone can be used for lung cancer prognosis. To test this hypothesis, we applied supervised principal component analysis to these 59 genes using the FFPE dataset to construct a 59-gene prognosis signature. Because the selection of these 59 genes used information from both FFPE and consortium datasets, we used another two independent lung cancer datasets including the Bild et al. (n=111) dataset and the Bhattacharjee et al. dataset (n=117) downloaded from the literature to validate our 59-gene signature. The 59-gene prediction model built from FFPE samples can significantly distinguish the high and low risk groups for both the Bhattacharjee et al and Bild et al. data sets (Figure 5a,
HR=1.81, $P=0.016$ and Figure 5c, HR=2.10, $P=0.02$, respectively). Furthermore this signature can also significantly distinguish the high and low risk groups within stage I patients for both datasets (Figure 5b 5d) indicating this 59-gene signature can refine the prognosis for lung cancer patients within stage I patients. Due to the small sample size for stage II and stage III patients in Bild et al. and Bhattacharjee et al. studies, the 59-gene prognosis signature was not tested for stage II and stage III patients. We also found that 59-gene prediction model built from the consortium dataset can also distinguish the high and low risk groups for the Bild et al. and Bhattacharjee et al datasets (Supplementary Figure 4 a-d).

To understand the potential biological relevance of these 59 genes significantly associated with survival in the FFPE and consortium data sets, we used Ingenuity Pathway Analysis (IPA) to explore which known regulatory networks are enriched in this 59-gene set. IPA analysis revealed the most significant molecular networks to be cancer, tumor morphology, and respiratory disease. This network (Figure 4c) includes 14 genes of the 59-gene set and is centered on transcription factors $HNF4A$, $HNF1A$, and $ONECUT1$ ($HNF6A$). This hepatocellular network has been implicated in hepatocellular carcinoma as determined by in vitro study and molecular interactions in this network are putatively involved in lung cancer survival.

**DISCUSSION**

In this study, we tested the feasibility of deriving a lung cancer prognosis gene signature from formalin-fixed paraffin-embedded tumor samples based on genome-wide mRNA expression profiling. Although RT-PCR methods have been used to
measure gene expression level from FFPE samples \(^{33-35}\), the selection of genes for testing are limited to the current knowledge base which is incomplete and inconsistent \(^{36}\). Due to degradation and chemical alteration of RNA extracted from FFPE samples, the use of microarray analysis of gene expression from FFPE samples has been hampered \(^{36}\). New technology and methodologies developed to extract RNA from FFPE samples coupled with new array platforms have made it possible to measure gene expression from FFPE samples \(^{33}, 37-40\). A recent study demonstrated the feasibility of using DNA-mediated annealing, selection, extension and ligation (DASL) arrays with 6100 preselected genes to profile mRNA expression from hepatocellular carcinoma tissue \(^{41}\). No prognosis signature for other types of cancer has been developed using microarray analysis of gene expression from FFPE extracted RNA. In this study, we built a robust gene signature for NSCLC based on microarray analysis of FFPE samples. We claim this is a robust gene signature because it has been validated in 6 independent published datasets including 4 sets from the consortium study and 2 additional studies from DFCI and Duke. We also built a prediction model using the same set of robust genes from frozen samples and validated the model in both frozen and FFPE samples.

Most published gene signatures identified from different studies are usually very different and with little overlap. However, we found that there is significant overlap among the robust genes associated with survival outcomes between the FFPE dataset and the consortium dataset \((P=0.008)\). More impressively, the hazard ratios, indicating the strength of the association of genes expression and survival time, are highly consistent between two independent datasets. Our interpretation for this consistency across studies is that the gene expression variation across studies is a major
contribution to signature differences across studies. In this study, we used strict quality steps to exclude genes that were not expressed in our FFPE samples. This allowed for analysis of the remaining genes which had more stable expression patterns and were more robust to environment changes. Validation of our novel 59-gene signature prognostic for NSLC survival in two additional independent datasets further confirmed the robustness of these genes.

By grouping our RGS of 1400 genes by gene expression, we found that the group expression levels correlated with survival. Interestingly, group 1 had a shorter survival and contained an ER negative breast cancer signature. Group 2 had a longer survival and contained an ER positive breast cancer signature. This correlation with ER status and survival has been demonstrated previously in breast cancer and shown to have predictive power for prognosis. In addition to ER status, the RGS groups were separated by the presence of stem cell signatures (embryonic stem cell signature and neural stem cell signature), with group 1 (shorter survival) having two stem cell signatures whereas group 2 (longer survival) did not. The embryonic stem cell signature has previously been shown to be associated with poor prognosis of NSCLC. In addition, in mouse models, a hematopoietic and neural stem cell–like signature in primary tumors has been shown to be a predictor of poor prognosis in 11 types of cancer, including lung. These ER status and stem cell signature data support our RGS expression groupings and their correlation with survival prognosis.

Besides the prognostic signature, the predictive signatures to determine the optimal chemotherapy regimen for individual patients also have tremendous clinical benefit. Tumor samples from clinical trials data are important to develop predictive signatures to reduce the selection bias for evaluating treatment efficacy within...
signature groups. However, very limited frozen tumor samples are available from completed clinical trials. Our study demonstrated the feasibility of using FFPE samples for genome-wide mRNA profiling. Therefore, this study provides an important step to construct and validate predictive signatures for chemotherapy response using the available FFPE samples from clinical trials in the future.
FIGURE LEGENDS

Figure 1. (a) Flow chart of the derivation and validation of the robust gene signature from formalin-fixed and paraffin-embedded samples collected from M.D. Anderson UT-Lung Cancer SPORE tissue bank. (b) Flow chart of the derivation and validation of 59-gene prognosis signature.

Figure 2. Microarray analysis of the gene-expression profiles from formalin-fixed and paraffin-embedded (FFPE) lung tumor samples. (a) Unsupervised cluster analysis of the 55 FFPE lung cancer patient cohort using the expression profile of 1400 robust genes that pass the microarray quality control criterion. Vertical and horizontal axes represent robust genes and lung cancer patient clusters, respectively. (b) Kaplan-Meier plot showing the association of the expression of robust genes with patient survival P-values were obtained using the log-rank test. Red color represents sample Cluster I and black color represents sample Cluster II defined by unsupervised clustering algorithm using robust gene profiling data. • indicates censored samples. Gene set enrichment analysis found that the ER negative signature derived from breast cancer patients is enriched in group 1 defined by RGS expression (c), and the ER positive signature derived from breast cancer patients is enriched in group 2 defined by RGS expression (d). The y axis shows running enrichment scores for the specific gene set on the 1400 pre-ranked genes. The x axis shows the rank in the ordered dataset. The vertical lines represent the locations of the genes that are in the specific gene set.
**Figure 3.** Kaplan-Meier plots showing the predictive power of the robust gene signatures. 55 FFPE tumor samples from M. D. Anderson Cancer Center were randomly divided into training (25 samples) and testing (30 samples) sets (a). Independent validation of the robust gene signature in the 442-frozen-sample cohort from multi-institute consortium. The microarray data sets were divided into two groups, one for the training and the other for the testing cohort according to the original paper (b). The training data is 55 FFPE tumor samples and the testing data set is 442-frozen-sample cohort from multi-institute consortium. The testing was done for all patients (c), stage I patients (e), stage II patients (f) and stage III patients (g) separately. The training data is the consortium dataset with 442 frozen samples and the testing data is 55 FFPE samples from M.D. Anderson Cancer Center (d). $P$ values were obtained by the log-rank test. Red and black lines represent predicted high- and low-risk groups, respectively. • indicates censored samples.

**Figure 4.** Comparison of individual gene effect across FFPE samples from M, D. Anderson Cancer Center and 442 frozen samples from consortium. (a) Venn-diagram of genes associated with overall survival ($P<0.05$ in univariate Cox regression models). It shows 59 genes are significantly associated with survival in both FFPE data and consortium data. (b) The hazard ratios from univariate Cox regression models for the 59 genes common in both sets are consistent between FFPE set and consortium set. (c) Regulatory gene and protein interaction networks defined by the 59 predictors. Computational molecular interaction network prediction based on genes and proteins associated with the significant pathways in the Ingenuity Pathways Knowledge Base (IPKB) by Ingenuity Pathways Analysis (IPA). Interactions between the different nodes
are given as solid (direct interaction) and dashed (indirect interaction) lines (edges) with various colors for the different interaction types. This network received the highest score by IPA and is mostly centered on the transcription factors \textit{HNF4A} and \textit{HNF1A}, and \textit{ONECUT1}. The shaded genes are the genes belonging to 59-gene signature.

**Figure 5.** Kaplan-Meier plots showing the predictive power of the 59-gene signature for two independent validation sets. The training data is 55 FFPE tumor samples from M.D. Anderson Cancer Center and the testing data set is frozen samples from lung cancer patients from Bhattacharjee et al \cite{31} dataset (a), the stage I patients from Bhattacharjee et al dataset (b), frozen samples from lung cancer patients from Bild et al \cite{9} dataset (c), and the stage I patients from Bild et al dataset (d). \textit{P} values were obtained by the log-rank test. Red and black lines represent predicted high- and low-risk groups, respectively. • indicates censored samples.
SUPPLEMENTARY FIGURES

Supplementary Figure 1. Kaplan-Meier survival plots for different histology (a) and histology combined with RGS groups defined by expression of 1400 robust genes in M.D. Anderson Cancer Center patients (b). P values were obtained by log-rank test. Red and black lines represent predicted high- and the low-risk groups, respectively. Solid and dash lines represent adenocarcinoma and squamous groups, respectively. • indicates censored samples.

Supplementary Figure 2. Kaplan-Meier plots showing the predictive power of the robust gene signature from different sets. The training data is 55 FFPE tumor samples and the testing data sets are frozen samples from 4 institutions in the consortium data: MSKCC (a), DFCI (b), MI (c) and HLM (d). P-values were obtained by the log-rank test. Red and black lines represent predicted high- and low-risk groups, respectively. • indicates censored samples.

Supplementary Figure 3. Kaplan-Meier plots showing the predictive power of the robust gene signature for patients with (a) and without (b) chemotherapy. The training data is 55 FFPE tumor samples and the testing data sets are frozen samples the consortium data. P-values were obtained by the log-rank test. Red and black lines represent predicted high- and low-risk groups, respectively. • indicates censored samples.

Supplementary Figure 4. Kaplan-Meier plots showing the predictive power of the 59-gene signature for two independent validation sets. The training data is the 442-frozen-sample cohort from multi-institute consortium dataset, and the testing data set is frozen
samples from Bhattacharjee et al dataset 31 (a), the stage I patients from Bhattacharjee et al dataset (b), frozen samples from lung cancer patients from Bild et al dataset 9 (c), and the stage I patients from Bild et al dataset (d). P values were obtained by the log-rank test. Red and black lines represent predicted high- and low-risk groups, respectively. • indicates censored samples.
REFERENCES

Table 1. The association between patients’ characteristics and RGS risk scores and survival time for consortium patients based on multivariate Cox regression model. (RGS scores were calculated from the prediction model built from MDACC FFPE samples)

<table>
<thead>
<tr>
<th>Variables</th>
<th>HR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGS risk scores</td>
<td>1.300 (1.074, 1.574)</td>
<td>0.0070</td>
</tr>
<tr>
<td>Gender (Female vs. Male)</td>
<td>0.803 (0.576, 1.119)</td>
<td>0.19</td>
</tr>
<tr>
<td>Age (continuous in unit of 10 years)</td>
<td>1.571 (1.321, 1.868)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Smoking (Current/Former vs. Never)</td>
<td>1.356 (0.791, 2.322)</td>
<td>0.27</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage II vs. stage I</td>
<td>2.116 (1.433, 3.126)</td>
<td>0.0002</td>
</tr>
<tr>
<td>Stage III vs. stage I</td>
<td>4.855 (3.164, 7.449)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Adjuvant Chemotherapy (Yes vs. No)</td>
<td>1.688 (1.172, 2.431)</td>
<td>0.0049</td>
</tr>
<tr>
<td>Study sites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFCI vs. MI</td>
<td>1.295 (0.741, 2.264)</td>
<td>0.36</td>
</tr>
<tr>
<td>HLM vs. MI</td>
<td>1.632 (1.094, 2.434)</td>
<td>0.016</td>
</tr>
<tr>
<td>MSKCC vs. MI</td>
<td>0.657 (0.419, 1.031)</td>
<td>0.068</td>
</tr>
</tbody>
</table>
**Figure 1**

A

- MDACC samples (N = 100)
- RNA QC
- FFPE Microarray experiment (N = 75)
- Analysis QC
- FFPE Data (N = 55)
- Prediction analysis
  - RGS signature
  - Predict
  - Consortium overall (Figure 3C)
  - Consortium by stage (Figure 3 E-G)
  - Consortium by site (Figure S2 A-D)

B

- All probe sets on Affymetrix U133 plus 2.0 arrays
- Microarray QC
  - 1400 robust genes
  - Associate with survival in FFPE data
  - Associate with survival in Consortium data
  - 131 genes
  - 365 genes
  - Overlap
  - 59-gene signature
    - Validation of 59-gene signature
    - Predict
      - Bhattacharjee et al (Figure 5A)
      - Bhattacharjee et al Stage I (Figure 5B)
      - Bild et al (Figure 5C)
      - Bild et al Stage I (Figure 5D)
Figure 2

A

Group 1

Group 2

B

Survival

Year

\( p = 0.017 \)

Group 1

Group 2

C

Enrichment plot: BRCA_ER_NEG

D

Enrichment plot: BRCA_ER_POS
Figure 5

A 59 genes, FFPE to Bhattacharjee et al dataset

B 59 genes, FFPE to Bhattacharjee et al dataset, stage I

C 59 genes, FFPE to Bild et al dataset

D 59 genes, FFPE to Bild et al dataset, stage I

Survival vs. Year

p = 0.016

p = 0.039

p = 0.02

p = 0.028
Clinical Cancer Research

Robust Gene Expression Signature from Formalin-Fixed Paraffin-Embedded Samples Predicts Prognosis of Non-Small-Cell Lung Cancer Patients

Yang Xie, Guanghua Xiao, Kevin Coombes, et al.

Clin Cancer Res Published OnlineFirst July 8, 2011.