A 12-Gene Set Predicts Survival Benefits from Adjuvant Chemotherapy in Non–Small Cell Lung Cancer Patients

Hao Tang, Guanghua Xiao, Carmen Behrens, et al.


<table>
<thead>
<tr>
<th>Updated version</th>
<th>Access the most recent version of this article at: doi:10.1158/1078-0432.CCR-12-2321</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary Material</td>
<td>Access the most recent supplemental material at: <a href="http://clincancerres.aacrjournals.org/content/suppl/2013/01/28/1078-0432.CCR-12-2321.DC1.html">http://clincancerres.aacrjournals.org/content/suppl/2013/01/28/1078-0432.CCR-12-2321.DC1.html</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cited Articles</th>
<th>This article cites by 41 articles, 14 of which you can access for free at: <a href="http://clincancerres.aacrjournals.org/content/19/6/1577.full.html#ref-list-1">http://clincancerres.aacrjournals.org/content/19/6/1577.full.html#ref-list-1</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citing articles</td>
<td>This article has been cited by 4 HighWire-hosted articles. Access the articles at: <a href="http://clincancerres.aacrjournals.org/content/19/6/1577.full.html#related-urls">http://clincancerres.aacrjournals.org/content/19/6/1577.full.html#related-urls</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E-mail alerts</th>
<th>Sign up to receive free email-alerts related to this article or journal.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reprints and Subscriptions</td>
<td>To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at <a href="mailto:pubs@aacr.org">pubs@aacr.org</a>.</td>
</tr>
<tr>
<td>Permissions</td>
<td>To request permission to re-use all or part of this article, contact the AACR Publications Department at <a href="mailto:permissions@aacr.org">permissions@aacr.org</a>.</td>
</tr>
</tbody>
</table>
A 12-Gene Set Predicts Survival Benefits from Adjuvant Chemotherapy in Non–Small Cell Lung Cancer Patients

Hao Tang1,2, Guanghua Xiao1,2, Carmen Behrens8, Joan Schiller3,4, Jeffrey Allen1,2, Chi-Wan Chow8, Milind Suraokar8, Alejandro Corvalan8, Jianhua Mao10, Michael A. White3,5, Ignacio I. Wistuba8,9, John D. Minna4,6,7, and Yang Xie1,2,3

Center, Dallas; Departments of Thoracic/Head and Neck Medical Oncology and Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390. Phone: 214-648-5178; Fax: 214-648-1663; E-mail: Yang.Xie@UTSouthwestern.edu

Published OnlineFirst January 28, 2013; DOI: 10.1158/1078-0432.CCR-12-2321

Abstract

Purpose: Prospectively identifying who will benefit from adjuvant chemotherapy (ACT) would improve clinical decisions for non–small cell lung cancer (NSCLC) patients. In this study, we aim to develop and validate a functional gene set that predicts the clinical benefits of ACT in NSCLC.

Experimental Design: An 18-hub-gene prognosis signature was developed through a systems biology approach, and its prognostic value was evaluated in six independent cohorts. The 18-hub-gene set was then integrated with genome-wide functional (RNAi) data and genetic aberration data to derive a 12-gene predictive signature for ACT benefits in NSCLC.

Results: Using a cohort of 442 stage I to III NSCLC patients who underwent surgical resection, we identified an 18-hub-gene set that robustly predicted the prognosis of patients with adenocarcinoma in all validation datasets across four microarray platforms. The hub genes, identified through a purely data-driven approach, have significant biological implications in tumor pathogenesis, including including NFK2-1, Aurora Kinase A, PRCI, CDKN3, MBIP, and RRM2. The 12-gene predictive signature was successfully validated in two independent datasets (n = 90 and 176). The predicted benefit group showed significant improvement in survival after ACT (UT Lung SPORE data: HR = 0.34, P = 0.017; JBR.10 clinical trial data: HR = 0.36, P = 0.038), whereas the predicted nonbenefit group showed no survival benefit for 2 datasets (HR = 0.80, P = 0.70; HR = 0.91, P = 0.82).

Conclusions: This is the first study to integrate genetic aberration, genome-wide RNAi data, and mRNA expression data to identify a functional gene set that predicts which resectable patients with non–small cell lung cancer will have a survival benefit with ACT. Clin Cancer Res; 19(6); 1577–86. © 2013 AACR.

Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide (1). Even after an apparent complete resection of NSCLC, 33% of patients with pathologic stage IA and 77% with stage IIIA disease die within 5 years of diagnosis. Several randomized trials have shown that there is a survival benefit with ACT in resected NSCLC (2–6). However, the effect is modest—only 4% to 15% improvement in 5-year survival, while at the same time, such treatment may cause serious adverse effects (5, 7). Because the response to standard chemotherapy in lung cancer varies, it would be very helpful to prospectively identify the subgroup(s) of patients who are unlikely to benefit from ACT, and therefore, can be spared the side effects of unnecessary treatment.

Recently, several groups have developed gene expression signatures aiming to classify lung cancer patients into groups with distinct clinical outcomes (8–21). However, most current molecular signatures for lung cancer are prognostic only, and do not provide any estimation as to whether a patient would benefit from ACT. In addition, the signatures often contain large numbers of genes, with limited information about the functional importance of the genes. All of these problems limit the clinical application of those signatures. In this study, we used a systems biology approach to construct a survival-related gene network in NSCLC and identified 18 “hub” genes, which consistently coexpressed with many survival-related genes and hence play important roles in multiple biological processes. Here we show that the 18-hub-gene set is functionally important and predicts the overall prognosis of NSCLC patients with stage I to III disease. Our previous RNAi screening study
Translational Relevance

Randomized clinical trials have showed the survival benefit of adjuvant chemotherapy (ACT) in resected non–small cell lung cancer (NSCLC). Because the response to standard chemotherapy in lung cancer varies, it would be very helpful to prospectively identify patients who will benefit from ACT to guide the treatment plan. In this study, we used a systems biology approach to identify an 18-hub-gene signature that can robustly predict the prognosis of patients with early-stage adenocarcinoma of the lung. Furthermore, we integrated these hub genes with genetic aberration and genome-wide RNAi functional data to derive a 12-gene set that is predictive for survival benefit with ACT. This 12-gene predictive signature has been validated in 2 independent NSCLC cohorts. As this predictive signature contains a small set of genes and has shown robust predictive power across platforms and studies, it may have therapeutic utility in determining which early-stage NSCLC patients would benefit from ACT.

Materials and Methods

Patients and samples

**UT Lung SPORE cohort.** Patients were eligible to enter the study if they underwent curative resection for NSCLC at MD Anderson Cancer Center between December 1996 and June 2007. Those with radiation therapy were excluded from the study. All tissue samples were obtained by surgical resection from patients who had provided written informed consent. Tissues were stored at −140°C after being snap frozen in liquid nitrogen. Serial sectioning of each sample was used to histologically evaluate tumor and malignant cells content before RNA extraction (25). The primary tumor tissues from 176 patients were randomly selected from the UT Lung SPORE tumor collection based on stringent, predefined quality control procedures, including the presence of ≥70% tumor tissue and ≥50% malignant cells in the frozen tissue used for RNA extraction. In this cohort, 133 patients are adenocarcinomas (ADCs) and 43 patients are squamous cell carcinomas (SCCs); 49 patients received ACT (mainly carboplatin plus taxanes) and 127 patients did not receive ACT. The clinical information and gene expression data for the UT Lung SPORE cohort were deposited in GEO database (GSE42127).

**Samples from other groups.** In addition to the UT Lung SPORE data, 7 public NSCLC microarray datasets (10, 13, 17, 26–29) were used in this study. The National Cancer Institute Director’s Challenge Consortium study (Consortium dataset; ref. 13), which is the largest independent public available lung cancer microarray dataset and involves 442 resected ADCs, was used as the training set. Six datasets were used to validate the prognosis signature: UT Lung SPORE data, GSE3141 (ADC, n = 58; SCC, n = 53), GSE8894 (ADC, n = 62; SCC, n = 76), GSE11969 (ADC, n = 90; SCC, n = 35), GSE13213 (ADC, n = 117), GSE4573 (SCC, n = 129). Among these 6 datasets, 3 (GSE13213, GSE8894, and GSE11969) are Asian cohorts. Two datasets were used to validate the predictive signature: UT Lung SPORE data and GSE14814 that includes 90 samples (49 patients with vinorelbine plus cisplatin ACT and 41 patients without ACT) collected from the JBR.10 trial. Table 1 provides detailed information on these datasets. Because 43 of 133 samples in the original JBR.10 dataset (GSE14814) were also included in the Consortium data (training set), these 43 samples were excluded from the JBR.10 dataset to ensure the independence between the training and validation sets.

**RNA extraction and microarray profiling.** The frozen tissues specimens were processed on the cryostat to generate multiple 5-μm-thick sections for subsequent homogenization using an electric homogenizer. Before RNA extraction, histology sections were stained and reviewed to assess the percentage of tumor. Total RNA was extracted using TRIReAGENT (Life Technologies) according to manufacturer’s protocol. The nanodrop spectrophotometer (Thermo Fisher) was used to estimate the concentration of RNA, whereas the quality of the RNA was assessed on Nano Series II RNA LAB-chips using Agilent Bioanalyzer 2100 (Agilent Technologies, Inc.). All samples selected for RNA profiling have an RNA integrity number (RIN) ≥5. Total RNA was processed for analysis on the Illumina Human-6 V3 arrays according to Illumina protocols for first- and second-strand synthesis, biotin labeling, and fragmentation.

**Microarray data preprocessing.** The UT Lung SPORE Illumina bead array data were processed using Model-Based Background Correction (MBCB) method (30). For the Consortium and GSE14814 datasets, the raw Affymetrix cel data were downloaded from caArray database and gene expression omnibus (GEO), respectively. Both datasets were then preprocessed by the robust multiarray average (RMA) algorithm and quantile–quantile normalization (31). For datasets that did not provide raw data file (GSE3141, GSE4573, and GSE8894)
Table 1. Clinical characteristics of patients in the validation datasets

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td>Female</td>
<td>83 (47.2%)</td>
<td>57 (48.7%)</td>
<td>48 (32.2%)</td>
<td>34 (24.6%)</td>
<td>–</td>
<td>47 (36.4%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td>93 (52.8%)</td>
<td>60 (51.3%)</td>
<td>101 (67.8%)</td>
<td>104 (75.4%)</td>
<td>–</td>
<td>82 (63.6%)</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td>I</td>
<td>112 (63.6%)</td>
<td>79 (67.5%)</td>
<td>78 (52.3%)</td>
<td>–</td>
<td>62 (55.9%)</td>
<td>73 (56.6%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II</td>
<td>32 (18.2%)</td>
<td>13 (11.1%)</td>
<td>26 (17.4%)</td>
<td>–</td>
<td>33 (25.6%)</td>
<td>45 (50.0%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>III</td>
<td>30 (17.0%)</td>
<td>25 (21.4%)</td>
<td>45 (30.2%)</td>
<td>–</td>
<td>23 (17.8%)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IV</td>
<td>1 (0.6%)</td>
<td>–</td>
<td>–</td>
<td>138 (100%)</td>
<td>49 (44.1%)</td>
<td>–</td>
</tr>
<tr>
<td>Histology</td>
<td></td>
<td>ADCs</td>
<td>133 (75.6%)</td>
<td>117 (100%)</td>
<td>90 (60.4%)</td>
<td>62 (44.9%)</td>
<td>58 (52.3%)</td>
<td>28 (31.1%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCCs</td>
<td>43 (24.4%)</td>
<td>–</td>
<td>35 (23.5%)</td>
<td>76 (55.1%)</td>
<td>53 (47.7%)</td>
<td>129 (100%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Others</td>
<td>–</td>
<td>–</td>
<td>24 (16.1%)</td>
<td>–</td>
<td>–</td>
<td>10 (11.1%)</td>
</tr>
<tr>
<td>Median follow-up</td>
<td></td>
<td></td>
<td>47.4</td>
<td>68</td>
<td>78</td>
<td>41.8</td>
<td>31.1</td>
<td>34.2</td>
</tr>
<tr>
<td>Platform</td>
<td></td>
<td></td>
<td>Illumina</td>
<td>Agilent 44K</td>
<td>Agilent 21.6K custom array</td>
<td>Affy U133</td>
<td>Affy. U133</td>
<td>Affy. U133A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Human-WG6 V3</td>
<td></td>
<td></td>
<td>Plus_2</td>
<td>Plus_2</td>
<td></td>
</tr>
</tbody>
</table>

Survival analysis

Overall survival time was calculated from the date of surgery until death or last follow-up contact. Survival curves were estimated using the Kaplan–Meier method (32) and were compared using log-rank test. Univariate and multivariate survival analyses were done using Cox proportional–hazards model (33). Meta-analysis was used to combine the results across different test sets. It was done using the R package meta (34). The overall combined estimate of the HR was estimated based on their values and standard errors in individual validation set.

Gene network analysis

The lung cancer survival-related gene network was constructed using the Consortium dataset. The association between the expression level of each probeset and survival time was evaluated using multivariate Cox model adjusted for age, cancer stage, and sample processing sites. The false discovery rate (FDR) was calculated from a β-uniform mixture model (35). All probesets that passed the FDR criteria (FDR < 10%) were included in gene network analysis. When there are multiple probesets corresponding to a single gene, the expression levels from the probesets were averaged to derive the gene-level expression. The Sparse Partial Correlation Estimation (SPACE) algorithm (36) was used to construct the network of survival-associated genes using their expression values in the Consortium dataset. From the constructed gene network, genes with at least 7 connections to other genes were identified as “hub” genes.

Prediction methods

Supervised principal component analysis (37, 38) was applied to construct the prediction model, which is based on the linear combinations of gene expressions of the provided gene set in the training dataset. Then we apply the risk prediction model to the test set, and derive a risk score for each samples based on their gene expressions. The test set samples are divided into 2 equal-sized risk groups based on the median of the predicted risk scores. For the prediction model we used the first 3 principal components, which the default parameter of the program with prediction (superPC R package). The training and validation strategy is illustrated in Fig. 1. Please see the Supplementary SWEAVE report for all analysis details including the models, parameters, and procedures for this study.

Results

Identification of an 18-hub-gene set

From the Consortium dataset, we identified 797 genes (Fig. 1) whose expression levels were associated with patients' overall survival time (FDR < 10%). Next, we constructed a lung cancer survival-related gene network (see Materials and Methods) based on expression changes of these 797 genes across 442 lung cancer samples in the Consortium dataset (Fig. 2A). We identified 18 hub genes that are connected with at least 7 other genes in the constructed network. Among these 18 genes (summarized in Fig. 2B), RRM2, AURKA, PRC1, and CDKN3 are associated with poor prognosis, whereas the remaining 14 genes are associated with good prognosis.
Prognosis performance of the 18-hub-gene set

Robustness of the prognostic signature. A prognostic signature was developed using the expression of the 18-hub-gene set and patients' survival outcomes from the Consortium dataset (training set) based on the superPC method. The prognostic signature was evaluated in ADC patients from 5 independent validation sets across 4 different microarray platforms, including: UT Lung SPORE (Illumina-6 V3), GSE13141 and GSE8894 (Affymetrix U133Plus2), GSE11969 (Agilent 21.6K custom array), and GSE13213 (Agilent 44K). Patients receiving ACT were excluded from the validation sets. Remarkably, the prognostic signature consistently predicted overall survival in all 5 validation sets. The predicted high-risk group has significantly worse survival outcomes than the predicted low-risk group: GSE3141 \( n = 58, HR = 2.06; 95\% \text{ CI}, 1.01–4.2; P = 0.042 \), UT Lung SPORE \( n = 94, HR = 2.85; 95\% \text{ CI}, 1.36–5.97; P = 0.0038 \), GSE8894 \( n = 62, HR = 3.73; 95\% \text{ CI}, 1.45–9.59; P = 0.0034 \), GSE11969 \( n = 90, HR = 1.87; 95\% \text{ CI}, 0.99–3.53; P = 0.049 \), GSE13213 \( n = 117, HR = 2.74; 95\% \text{ CI}, 1.51–4.98; P = 0.0058 \). Because most of the public datasets did not provide complete demographic information, we conducted multivariate survival analysis in UT Lung SPORE data. The predicted high-risk group has significantly worse survival outcomes than the predicted low-risk group (HR = 2.93; 95\% CI, 1.25–6.88; P = 0.0137) after adjusting for stage, age, and gender (Supplementary Table S2). Furthermore, the 18-hub-gene signature consistently predicted the prognosis of patients with stage I disease: GSE13141 \( n = 30, HR = 3.88; 95\% \text{ CI}, 1.18–12.8; P = 0.016 \), UT Lung SPORE \( n = 67, HR = 3.18; 95\% \text{ CI}, 1.14–8.84; P = 0.019 \).
GSE11969 (n = 52, HR = 2.85; 95% CI, 0.99–8.23; P = 0.043), GSE13213 (n = 79, HR = 5.31; 95% CI, 1.99–14.2; P = 0.00020; Fig. 3B).

The 18-hub-gene prognostic signature is ADC specific. ADC and SCC are 2 major NSCLC histology subtypes with fundamentally different molecular makeup (39). Because the 18-hub-gene prognostic signature was derived from a cohort of ADC patients only, we wanted to determine whether it was specific for ADC or could also predict prognosis for SCC patients. We tested the 18-hub-gene prognostic signature in SCC patients from GSE3141 (n = 53), UT Lung SPORE (n = 33), GSE8894 (n = 76), GSE11969 (n = 35), GSE4573 (n = 129). The results (Fig. 3C) show that the signature does not predict survival in any of the 5 datasets. Note that 4 datasets (GSE3141, SPORE, GSE8894, and GSE13213) have both ADC and SCC patients, and the 18-hub-gene signature has significant prognostic values in all ADC subcohorts, but not in any SCC subcohorts. These results show that the 18-hub-gene prognostic signature is ADC specific (P = 0.00047 for interaction between histology and signature).

GSE11969 (n = 52, HR = 2.85; 95% CI, 0.99–8.23; P = 0.043), GSE13213 (n = 79, HR = 5.31; 95% CI, 1.99–14.2; P = 0.00020; Fig. 3B).

The 18-hub-gene prognostic signature is ADC specific. ADC and SCC are 2 major NSCLC histology subtypes with fundamentally different molecular makeup (39). Because the 18-hub-gene prognostic signature was derived from a cohort of ADC patients only, we wanted to determine whether it was specific for ADC or could also predict prognosis for SCC patients. We tested the 18-hub-gene prognostic signature in SCC patients from GSE3141 (n = 53), UT Lung SPORE (n = 33), GSE8894 (n = 76), GSE11969 (n = 35), GSE4573 (n = 129). The results (Fig. 3C) show that the signature does not predict survival in any of the 5 datasets. Note that 4 datasets (GSE3141, SPORE, GSE8894, and GSE13213) have both ADC and SCC patients, and the 18-hub-gene signature has significant prognostic values in all ADC subcohorts, but not in any SCC subcohorts. These results show that the 18-hub-gene prognostic signature is ADC specific (P = 0.00047 for interaction between histology and signature).

The 18-hub-gene set has better performance than top-ranked genes.

Selecting an optimal small set of genes from a large candidate gene list is a critical step for developing clinically practical molecular assays. The most widely used ranking based approaches (10) select genes with the most prominent P values obtained from individual gene-based testing. We derived an 18-top-ranked-gene set containing 18 genes with the most significant association with the survival outcome based on the expression of the 18 hub genes divided the patients into an ADC dominated group and a SCC-dominated group (Supplementary Fig. S1).

15 of the 18 hub genes express differently between ADC and SCC patients, and unsupervised clustering analysis based on the expression of the 18 hub genes divided the patients into an ADC dominated group and a SCC-dominated group (Supplementary Fig. S1).

The 18-hub-gene set has better performance than top-ranked genes.

Selecting an optimal small set of genes from a large candidate gene list is a critical step for developing clinically practical molecular assays. The most widely used ranking based approaches (10) select genes with the most prominent P values obtained from individual gene-based testing. We derived an 18-top-ranked-gene set containing 18 genes with the most significant association with the survival outcome based on the expression of the 18 hub genes divided the patients into an ADC dominated group and a SCC-dominated group (Supplementary Fig. S1).

15 of the 18 hub genes express differently between ADC and SCC patients, and unsupervised clustering analysis based on the expression of the 18 hub genes divided the patients into an ADC dominated group and a SCC-dominated group (Supplementary Fig. S1).

The 18-hub-gene set has better performance than top-ranked genes.

Selecting an optimal small set of genes from a large candidate gene list is a critical step for developing clinically practical molecular assays. The most widely used ranking based approaches (10) select genes with the most prominent P values obtained from individual gene-based testing. We derived an 18-top-ranked-gene set containing 18 genes with the most significant association with the survival outcome based on the expression of the 18 hub genes divided the patients into an ADC dominated group and a SCC-dominated group (Supplementary Fig. S1).

15 of the 18 hub genes express differently between ADC and SCC patients, and unsupervised clustering analysis based on the expression of the 18 hub genes divided the patients into an ADC dominated group and a SCC-dominated group (Supplementary Fig. S1).
Comparing the prognostic performances

Using the Consortium dataset as the training set, the prognosis performances of the 18-hub-gene set, 18-top-ranked-gene set, and 797-SR-gene set were compared in 5 independent validation sets for ADC patients. Figure 4A shows that the 18-hub-gene signature consistently predicted prognosis in all 5 validation sets (HR = 2.46; P = 1.74E-08 from meta-analysis), and outperformed the 18-top-ranked-gene signature (HR = 1.88, P = 4.45E-05 from meta-analysis), which predicted prognosis (with P-value < 0.05) in only 2 of 5 datasets. Furthermore, the 18-hub-gene signature has similar or even better prognostic performance than the 797-SR-gene signature (HR = 2.24; P = 2.72E-07; Fig. 4A). It suggests that the hub-gene approach can effectively reduce the number of genes in the signature without sacrificing the prediction performance.

Comparing the information content

We used information theory approach (see Supplementary Methods) to study the reason why the hub-gene approach works well. The 18-hub-gene set has significantly higher pair-wise mutual information distance (a measure for independency) than the 18-top-ranked-gene set (P = 1E-9; Fig. 4C), indicating that the hub gene set has lower information redundancy than the top-ranked-gene set. As a result, the 18-hub-gene set has much higher entropy (a measure for information content; Fig. 4D) and captures more variation across patient population (Fig. 4B) than the 18-top-ranked-gene set. In summary, the hub-gene approach can effectively retain information while largely reducing the number of genes in the signature, which is important for developing clinically practical assays.

Derivation of a 12-gene set

Figure 1B illustrates the procedures for deriving and validating the 12-gene signature. First, we found that 7 of 18 hub genes have significant genetic aberration in lung cancer using the Tumorscape program (http://www.broad-institute.org/tumorscape; Fig. 2B), including a key lung cancer driver gene (NKX2-1; ref. 23). Furthermore, 9 of 18 hub genes were “synthetic lethal” with paclitaxel for NSCLC (i.e., siRNA gene-specific knockdowns which killed NSCLC cells only in the presence of paclitaxel) based on our previous study (ref. 22; Fig. 2B). In total, 12 of 18 hub genes...
either have genetic aberration or are "synthetic lethal" for paclitaxel in lung cancer. These genes are DOCK9, RRM2, AURKA, HOPX, NKX2-1, TTC37, COL4A3, IFT57, C1orf116, HSD17B6, MBIP, and ATP8A1. We developed a prediction model (12-gene signature) using the expression of these 12 genes and patients' survival outcomes in the Consortium dataset (training set) based on the superPC model and tested its prognostic effects on 5 independent ADC cohorts. The predicted high-risk group has significantly worse survival outcomes than the predicted low-risk group in the testing cohorts (Supplementary Fig. S2), so this 12-gene signature can predict prognosis in early-stage NSCLC.

**The 12-gene signature predicts survival benefits from ACT in NSCLC.**

Because these 12 genes are "hubs" of the survival related genes, and play roles in cell response to chemotherapy drugs or have genetic aberrations in lung cancer, we hypothesize that this 12-gene set can predict survival benefits from ACT in NSCLC. To test this hypothesis, we tested whether the 12-gene signature can predict which patients would benefit from ACT using 2 independent validation sets: (1) 90 NSCLC samples from JBR.10 clinical trial (17) in which 49 patients received vinorelbine plus cisplatin ACT treatment and 41 patients did not receive ACT; (2) 176 NSCLC samples from UT Lung SPORE in which 49 patients received ACT (mainly Carboplatin plus Taxanes) and 127 patients did not receive ACT. Each patient in the validation sets was classified into a high- or low-risk group based on the 12-gene signature. Different from the prognosis biomarkers, no study has shown that the predictive biomarkers for chemotherapy are ADC or SCC specific. Therefore, we tested the 12-gene signature in all NSCLC patients as other predictive biomarker studies (8, 17, 40). For the JBR.10 dataset, the ACT-treated patients showed longer survival than those without ACT (HR = 0.36; 95% CI, 0.13–0.97; P = 0.038; Fig. 5A) in the high-risk group, whereas patients with ACT treatment had no significant survival.

**Figure 4.** Comparison of the 18-hub-gene set, 18-top-ranked-gene set, and 797-SR-gene signature. A, summary of the prognostic performance for 18 hub-gene set, 18 top-ranked-gene set, and 797-SR-gene set. The training data are the Consortium data; the validation sets include 5 different datasets. Overall HR and P-values were calculated from meta-analysis. B, expression variation across the population in the Consortium dataset based on principal component analysis. C, pair-wise mutual information distance based on expression values in the Consortium dataset. D, entropy of expression values in the Consortium dataset.
benefits \( [HR = 0.91; 95\% CI, 0.391–2.11; P = 0.823; \text{Fig. 5A}] \) in the low-risk group. Furthermore, the patients with ACT treatment even have worse survival outcomes in the first 21 months for the low-risk group. The signature has a similar predictive effect in the UT Lung SPORE data: the patients who received ACT had better overall survival in the high-risk group \( [HR = 0.34; 95\% CI, 0.13–0.86; P = 0.017; \text{Fig. 5B}] \), but not in the low-risk group \( [HR = 0.80; 95\% CI, 0.266–2.42; P = 0.70; \text{Fig. 5B}] \).

Discussion

This is the first study to use systems biology approaches to identify hub genes for prognostic and predictive signatures in lung cancer. Feature selection, which selects the most predictive genes while excluding the redundant genes to reduce the cost, is a critical step in developing a clinically practical molecular assay. A commonly used selection approach is based on ranking the performance of individual features (genes), and selecting the top-ranked features. However, the combination of top-ranked individual genes may not be optimal, because it does not consider relationship and potential information redundancy among genes.

In this study, we applied a systems biology approach to identify hub genes, which have 7 to 30 connections with other genes in the constructed survival-related network (Fig. 2), so the expression changes of these hub genes will affect many other genes and lead to substantial changes at the system level. This 18-hub-gene set has higher information content (Fig. 4B–D) than the 18-top-ranked-gene set and has remarkably robust prognosis performances across different datasets and microarray platforms. From the Molecular Signatures Database (MsigDB), we identified 4 lung cancer prognosis signatures derived from the same training dataset (the Consortium dataset). In addition, we identified another 4 NSCLC prognosis signatures with similar number of genes from the literatures \( (9, 17, 41, 42) \). We compared the prediction performances of the 18-hub-gene signature and the 8 prognosis signatures in GSE13213 \( (n = 117 \text{ for ADC}) \), which has the most ADC patients in our testing datasets, and the 18-hub-gene signature clearly outperforms all other 8 signatures (Supplementary Table S3). These results indicate that the hub genes capture the key mRNA expression information related to NSCLC patients’ survival.

The 18 hub genes, identified through a purely data-driven approach, have important biological implications in tumor development, including 7 cancer metastasis genes and 1 key lung cancer driver gene \( (\text{NKX2-1}) \), showing the biological relevance of this approach. To understand the potential biological and therapeutic relevance of the identified hub gene signature, we downloaded all the gene lists from the MSigDB C2 gene sets database, and evaluated the overlap between our signatures and the gene lists (Supplementary Table S4). Most notably, all of the hub genes have been identified in at least 1 gene list concerning cancer or carcinoma, whereas 7 genes are associated with cancer metastasis gene lists, and 6 genes are related to proliferation. The large overlap with cancer-associated gene lists implies that our prognostic gene signature is biologically relevant, and it is likely that the prognostic power is originated from their association with cancer metastasis or tumor cell proliferation. In particular, \( \text{NKX2-1} \) and \( \text{HOPX} \) are important for the activation of p53 pathways and potentially helpful in repressing lung ADC development \( (43, 44) \), and could be promising candidates for lung cancer therapy.

This is also the first study to integrate RNAi functional screening data \( (22) \) with mRNA expression and genetic aberration data \( (23, 24) \) to identify a gene signature that predicting the benefits of ACT in lung cancer. This 12-gene signature is predictive for ACT benefits in NSCLC for both paclitaxel or vinorelbine plus cisplatin (JBR.10 clinical trial cohort), and commonly used combinations such as carboplatin plus taxanes (UT Lung SPORE cohort). The 12-gene signature is both prognostic for ADC patients and predictive for ACT, so this signature has the potential to facilitate clinical decisions on using ACT for early-stage NSCLC patients. In addition, the 18-gene set is a stronger prognostic

![Figure 5. Validation of the 12-gene predictive signature in 2 independent data sets. A, JBR.10 clinical trial dataset. The high- and low-risk groups were defined by the 12-gene signature. The patients were divided into 2 equal-sized risk groups based on their estimated risk scores. In the high-risk group, patients with ACT (pink line) have significantly longer survival time than patients without ACT (observation group, blue line). In the low-risk group, patients with ACT (pink line) do not have significantly longer survival time than patients without ACT (observation group, blue line). B, UT Lung SPORE dataset. The risk groups were defined by the same 12-gene signature. In the high-risk group, patients with ACT (pink line) have significantly longer survival time than patients without ACT (observation group, blue line). In the low-risk group, patients with ACT (pink line) do not have significantly longer survival time than patients without ACT.]

Downloaded from clinincancerres.aacrjournals.org on August 3, 2014. © 2013 American Association for Cancer Research.
signature for early-stage ADC patients, so the 18-gene signature could be helpful if the goal is to predict patients’ prognosis only. In addition, the EGFR mutation and ALK rearrangement could be important to patient response to chemotherapy, and further studies are needed to test how these mutations could affect the usage of the 12-gene signature.

Although this study shows the promising results and interesting functional relevance of the 12-gene signature, one limitation of this study is that the sample size is not big enough (45) to test the interaction between the signature and the treatment groups. Because the long-term survival outcome may be confounded by other nontreatment factors, we tested the interaction between signature groups and treatment using the survival in first 3 years after treatment. For JBR.10 data, the interaction between the 12-gene signature and the treatment groups is significant (P = 0.0005). The SPORE testing data is from a retrospective study. This dataset has limited sample size with treatments and the follow-up time is short, so the number of observed events is too small to reach the significant P-value for the interaction term. Therefore, a further prospective study with large sample size is needed to validate the 12-gene signature as a predictive signature.

In this study, the 18-hub-gene prognosis signature was validated in 6 independent datasets across 5 different microarray platforms (including Affymetrix U133Plus2, Affymetrix U133A, Illumina Human-6 V3, Agilent 21.6K custom arrays, and Agilent 44K), and the validation cohorts include 3 studies conducted in western countries and 3 studies conducted in Asia. The prognosis performances are consistent across these heterogeneous populations and experimental techniques. We tested the 12-gene predictive signature in 2 independent cohorts: the JBR.10 clinical trial and the UT Lung SPORE cohort. To our knowledge, this is the first study to include 2 validation datasets for predictive signatures in lung cancer. Zhu and colleagues [17] and Chen and colleagues [8] developed a predictive signature for ACT lung cancer, but it was only tested on the JBR.10 trial data. The UT Lung SPORE cohort used carboplatin plus taxanes-based ACT treatments and the microarray experiment platform is different. All these results show the robustness of the prognosis and predictive signatures developed from this study. To facilitate other researchers to reproduce the results in this study, we have provided a literate programming R package (SWEAVE report) in the Supplementary Material.

In summary, through systems biology approaches we have identified a robust 18-hub-gene signature for prognosis of resected NSCLC patients. Furthermore, we developed a 12-gene prognostic and predictive signature for ACT benefit in NSCLC patients using integrative analysis approaches. A prospective clinical study is needed to further validate the clinical value of the prognosis and predictive signatures in the decision-making process of ACT for resected NSCLC patients.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions

Conception and design: H. Tang, G. Xiao, C. Behrens, J. Schil, A. Corvalan, I. I. Wistuba, J. D. Minna, Y. Xie

Development of methodology: H. Tang, G. Xiao, A. Corvalan, J. D. Minna, Y. Xie

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): H. Tang, C. Behrens, C.-W. Chow, M. Suraoka, A. Corvalan, M. White, J. I. Wistuba, J. D. Minna, Y. Xie

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): H. Tang, G. Xiao, J. Schil, J. Allen, J. D. Minna, Y. Xie

Writing, review, and/or revision of the manuscript: H. Tang, G. Xiao, J. Schil, A. Corvalan, J.-H. Mao, M. White, J. I. Wistuba, J. D. Minna, Y. Xie

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): H. Tang, C. Behrens, J. Allen, C.-W. Chow, Y. Xie

Study supervision: H. Tang, G. Xiao, Y. Xie

Grant Support

This work was supported by NIH grants S10CA152301 (to H. Tang, Y. Xie, and I. I. Wistuba), University of Texas SPORE in Lung Cancer (P50CA70907 to J. D. Minna, Y. Xie, and I. I. Wistuba), P30CA142543 (to Y. Xie), 4R33GM072792 (to G. Xiao), NCI grant DMS-0907562 (to G. Xiao), NASA grant N00014-10-1-1036 (to I. I. Wistuba and J. D. Minna), Welch grant Welch Foundation I-1414 (to M. A. White), and CPRIT RP101251 (to Y. Xie).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received July 25, 2012; revised January 11, 2013; accepted January 21, 2013; published OnlineFirst January 28, 2013.

References