A Gene Expression Signature Predicts Survival of Patients with Stage I Non-Small Cell Lung Cancer

Yan Lu[1,2], William Lemon[1,2], Peng-Yuan Liu[1,2], Yijun Yi[1,2], Carl Morrison[3], Ping Yang[4], Zhifu Sun[4], Janos Szoke[5], William L. Gerald[5], Mark Watson[2,6], Ramaswamy Govindan[2,7], Ming You[1,2*]

1 Department of Surgery, Washington University School of Medicine, St. Louis, Missouri, United States of America, **2** The Alvin J. Siteman Cancer Center, Washington University School of Medicine, St. Louis, Missouri, United States of America, **3** Department of Pathology, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, United States of America, **4** Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America, **5** Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America, **6** Department of Pathology and Immunology, Washington University in St. Louis, St. Louis, Missouri, United States of America, **7** Department of Internal Medicine, Washington University in St. Louis, St. Louis, Missouri, United States of America

ABSTRACT

Background

Lung cancer is the leading cause of cancer-related death in the United States. Nearly 50% of patients with stages I and II non-small cell lung cancer (NSCLC) will die from recurrent disease despite surgical resection. No reliable clinical or molecular predictors are currently available for identifying those at high risk for developing recurrent disease. As a consequence, it is not possible to select those high-risk patients for more aggressive therapies and assign less aggressive treatments to patients at low risk for recurrence.

Methods and Findings

In this study, we applied a meta-analysis of datasets from seven different microarray studies on NSCLC for differentially expressed genes related to survival time (under 2 y and over 5 y). A consensus set of 4,905 genes from these studies was selected, and systematic bias adjustment in the datasets was performed by distance-weighted discrimination (DWD). We identified a gene expression signature consisting of 64 genes that is highly predictive of which stage I lung cancer patients may benefit from more aggressive therapy. Kaplan-Meier analysis of the overall survival of stage I NSCLC patients with the 64-gene expression signature demonstrated that the high- and low-risk groups are significantly different in their overall survival. Of the 64 genes, 11 are related to cancer metastasis *(APC, CDH8, IL8RB, LY6D, PCDHGA12, DSP, NID, ENPP2, CCR2, CASP8,* and *CASP10)* and eight are involved in apoptosis *(CASP8, CASP10, PIK3R1, BCL2, SON, INHA, PSEN1,* and *BIK)*.

Conclusions

Our results indicate that gene expression signatures from several datasets can be reconciled. The resulting signature is useful in predicting survival of stage I NSCLC and might be useful in informing treatment decisions.

The Editors' Summary of this article follows the references.

Funding: This work was supported by grants from the US National Institutes of Health (CA058554, CA093643, CA099187, ES012063, ES013340, and CA084354). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Academic Editor: Naftali Kaminski, University of Pittsburgh School of Medicine, United States of America

Citation: Lu Y, Lemon W, Liu PY, Yi Y, Morrison C, et al. (2006) A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PLoS Med 3(12): e467. doi:10.1371/journal.pmed.0030467

Received: May 16, 2006
Accepted: September 20, 2006
Published: December 26, 2006

Abbreviations: ADC, adenocarcinoma; DWD, distance-weighted discrimination; NSCLC, non-small cell lung cancer; QRT-PCR, quantitative real-time polymerase chain reaction; SCC, squamous cell carcinoma

* To whom correspondence should be addressed. E-mail: youm@msnotes.wustl.edu

Introduction

Lung cancer is the leading cause of cancer death for both men and women in the US [1]. The high mortality among patients with lung cancer is mainly due to the absence of an effective screening strategy to identify lung cancer at an early stage [2]. Thus, only ~25% of patients presenting with lung cancer are in a sufficiently early stage to be amenable to effective surgical treatment. Patients with stage I or II non-small cell lung cancer (NSCLC) have ~70% five-year survival after surgery alone compared to less than a 5% five-year survival for advanced lung cancer (stages IIIB and IV) [3]. Even with surgical resection, almost half of those with stage I or II disease eventually die from recurrences.

Treatment choices for patients with NSCLC depend on the stage at which the cancer is diagnosed. Patients diagnosed with stage I NSCLC usually receive surgical resection only [4]. Patients with stage IA (T1N0M0) undergo resection and are rarely treated with adjuvant chemotherapy. Patients with resected stage IB–III (any T any N M0 except T1N0M0) NSCLC show improved survival when given adjuvant chemotherapy [4].

No reliable clinical or molecular predictors of recurrent disease are currently available. Because of heterogeneity in recurrence rates among patients with the same stage of cancer, it is critical to isolate a reliable molecular signature in tumors that could be used to identify those who are likely to develop recurrent disease and would thus benefit from adjuvant therapy. Moreover, identification of genes and molecular pathways critical for development of metastasis could lead to advances in therapeutics.

Several studies based on microarray technology have been performed to determine genetic profiles predictive of survival in NSCLC and to develop genomic approaches for stratifying risk [5–8]. However, the identified survival-related genes lacked consistency among these studies, likely due to limited patient samples, disease heterogeneity, and/or technical factors such as differences in microarray platforms and specimen processing. In this study, we conducted a meta-analysis of seven datasets to search for differentially expressed genes related to survival time (under 2 years, i.e., short-term survival and over 5 years, i.e., long-term survival). The data analyzed include our own previously unpublished dataset.

Methods

Data Collection

Samples from Washington University. Thirty-six patients who underwent resection of stage IB NSCLC at Washington University School of Medicine (WUSM; St. Louis, Missouri, United States) were recruited for this study. These samples are referred to as dataset 1. Informed consent was obtained from the patients for tissue procurement prior to surgery and their medical records were maintained according to institutional guidelines and in conformance with HIPPA regulations. The overall survival data on all patients were censored on the date of the last follow-up visit or death from causes other than lung cancer. Tumor tissues were processed by the Human Tissue Bank and the Gene Chip Facility at WUSM according to standard operating procedures and protocols.

Briefly, frozen tissue samples at –80°C were pulverized and total cellular RNA was collected from each flash-frozen sample using TRIzol RNA isolation reagent (Invitrogen [http://www.invitrogen.com]). Total RNA was processed with a Qiagen (http://www.qiagen.com) RNeasy Mini kit. In vitro transcription-based RNA amplification was then performed on at least 8 μg of total RNA from each sample. Complementary DNA was synthesized using the T7-(dT)24 primer: 5′-GGCCAGTGAATTGTAATACGACT-CACTA-TAGGGAGGCGG-(dT)24–3′. The cDNA was processed using phase-lock gel (Fisher [http://www.fishersci.com; #E0032005101]) phenol/chloroform extraction. Next, in vitro transcriptional labeling with biotin was performed using the Enzo Bioarray Kit (Affymetrix [http://www.affymetrix.com; #900182]). The resulting cRNA was processed again using the Qiagen RNeasy Mini kit. Labeled cRNA was hybridized to HG__U95Av2 (Affymetrix) arrays according to manufacturer's instructions.

The raw fluorescence intensity data within CEL files were preprocessed with Robust Multichip Average (RMA) algorithm [9], as implemented with R packages from Bioconductor (http://www.bioconductor.org). This algorithm analyzes the microarray data in three steps: a background adjustment, quantile normalization, and finally summation of the probe intensities for each probe set using a log scale linear additive model for the log transform of (background corrected, normalized) PM intensities.

Samples from Mayo Clinic. Eighteen patients with stage I squamous cell carcinoma (SCC) were selected from the patients diagnosed with lung cancer from 1997 to 2001 who underwent resection at Mayo Clinic, Rochester, Minnesota, United States. These samples are referred to as dataset 2. All enrolled patients and use of their tissues in the study were approved by the institutional review board of Mayo Clinic. The resected tumors were flash-frozen to –80 °C within 30 min after the tissues were surgically removed. The RNA isolation, cRNA synthesis, and microarray hybridization were performed as described by Sun et al. [6]. The raw fluorescence intensity data within CEL files were also preprocessed with the RMA algorithm.

Samples from other groups. Dataset 3 was from Beer et al. [5] and includes 67 stage I primary lung adenocarcinomas (ADCs) (http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/index.html). Dataset 4 was from Bhattacharjee et al. [10] and includes 72 stage I lung ADCs (http://www.broad.mit.edu/mpr/lung/). Dataset 5 was from Borczuk et al. [8] and includes one squamous and three ADCs (http://hora.cpmc.columbia.edu/dept/pulmonary/5ResearchPages/Laboratories/powell%20supp1.htm). Dataset 6 was from Gerald et al. (unpublished data) and includes 63 stage I lung ADCs. Dataset 7 was from Bild et al. [11] and includes 33 squamous cell carcinomas and 31 ADCs (GEO accession number GSE3141). The raw data within the CEL files of these datasets were also preprocessed with the RMA algorithm.

All the samples used in our data analyses are listed in Table S1. Details of the clinical information for the subjects in each dataset are described in Table 1.

Data Processing

Gene matching. Because several different microarray platforms were used in these datasets, the probe sets should be matched to identical genes. The batch query tool provided by

Table 1. Clinical Summary of Patients in the Analyzed Datasets

Characteristics	Measurements	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7
Total samples	*n*	36	18	67	4	72	63	64
Mean age (range)	Years	66 (48–81)	70 (59–80)	64 (41–85)	76 (61–88)	64 (33–88)	65 (40–82)	ND
Sex	Male	20	10	25	2	29	27	ND
	Female	16	8	42	2	43	36	ND
Mean follow-up (days)	Total overall survival	1,369	1,301	1,310	303	1,403	1,387	1,139
	Total alive	1,570	1,813	1,430	303	1,805	1,441	1,414
	Total dead	665	660	924	ND	901	1,064	785
Stage	IA	0	7	44	1	33	25	30
	IB	36	11	23	3	39	38	27
Histological type	ADC	14	0	67	3	72	63	31
	SCC	18	18	0	1	0	0	33
	Other	4	0	0	0	0	0	0

ND, no data
doi:10.1371/journal.pmed.0030467.t001

Affymetrix (https://www.affymetrix.com/analysis/netaffx/batch_query.affx) was used for matching probe sets among datasets 1 to 7 [12]. Based on the latest UniGene clusters annotation provided by the manufacturer (NCBI Build 35.1), there were a total of 4,905 genes on all the five Affymetrix microarray systems HG_U95Av2, Hu6800, Hu133A, HG_U133AB, and Hu133plus2.

Distance-weighted discrimination. Systematic differences from different datasets were remarkable, which would compromise the integrity of the data from different laboratories. To integrate the gene expression data from datasets 1 to 5, the distance-weighted discrimination (DWD) method (https://genome.unc.edu/pubsup/dwd/index.html) [13] was used to identify and adjust systematic differences that were present within these microarray datasets. The DWD method corrects for systematic biases across microarray batches by finding a separating hyperplane between the two batches and adjusting the data by projecting the different batches on the DWD plane, finding the batch mean, and then subtracting out the DWD plane multiplied by this mean [13]. All of the 197 samples from the five datasets were broken into two sub-branches, each of which was composed of samples from all of the five datasets (Figure S1). Poolability tests were performed to examine if these DWD-transformed gene expression data from different resources were poolable [14]. We randomly reshuffled data resources and generated 100 replicates of simulated data. We then compared the number of p-values below certain thresholds with the expected counts obtained by simulations that take into account the distributions of the DWD-transformed gene expression data and sample size in our study.

Data Analysis

To preselect survival-related genes, ANOVA analysis was applied to 88 patients in datasets 1 to 5 who died within 2 years or survived beyond 5 years after surgery. Empirical p-values for each gene were obtained through 10,000 permutation tests. Genes with significant survival effects ($p < 0.01$) were selected for Cox proportional hazards regression analyses. Multivariate Cox proportional hazards regression analyses (adjusted for age, gender, cancer subtype, and cancer stage) with 10,000 bootstrap resampling were performed for each survival-related gene using all of 197 samples in datasets 1 to 5. The proportional hazards assumption for variables such as age, sex, cancer subtype, and cancer stage was investigated by examining the scaled Schoenfeld residuals. Sex and cancer stage generally displayed a significant deviation from this assumption. Therefore, these two variables were taken as strata and others as covariates in our Cox proportional hazards model. The plot of global p-values obtained by testing the proportional hazards assumption for all survival-related genes showed that the model used in our survival analysis was statistically warranted (Figure S2). The genes were ranked according to the bootstrap frequencies of $p < 0.01$ for their expression in regression models.

To identify a gene signature predictive of survival outcome, survival analyses were performed on all 197 samples in datasets 1 to 5. Partial Cox regression was performed to construct predictive components, and time-dependent ROC curve analysis was applied to evaluate the results [15]. The risk scores were calculated by a linear combination of the gene expression values for the selected genes, weighted by their estimated regression coefficients. All the samples were classified into high or low risk groups according to the risk scores. To choose an appropriate subset of genes for a common signature, we performed a forward selection procedure: (1) increase one gene each time based on the rank of genes that were identified in the above bootstrap analyses; (2) perform the partial Cox regression analysis and obtain the prediction accuracy using the chosen subset of genes; and (3) repeat steps 1 and 2 until the prediction accuracy is maximized. Kaplan-Meier survival plots, Mantel-Haenszel log rank tests, and time-dependent ROC analysis were implemented to assess the classification models according to the risk scores.

Hierarchical clustering based on a centered Pearson correlation coefficient algorithm and an average linkage method were used to show the expression patterns of survival-related genes in datasets 1 to 5.

All of the data analyses were implemented using the R statistical package [16]. A more detailed description of the data analyses is provided (Protocol S1).

Table 2. Genes Related to Survival

Expression	Gene	Function	*p*-Value
Overexpressed in low-risk patients	*ABI2*	Cell migration	0.0023
	APC[a]	Cell adhesion	0.0037
	ARHGEF1[a]	Cell proliferation	0.0031
	BCL2[a]	Antiapoptosis	0.0002
	BNIP1	Antiapoptosis	0.0004
	C21orf33	Cell wall	0.0080
	CASP10[a]	Induction of apoptosis	0.0082
	CASP8[a]	Regulation of apoptosis	0.0070
	CDH8[a]	Cell adhesion	0.0007
	CHRNA2	Signal transduction	0.0100
	DPH2L1	Cell proliferation	0.0049
	DTNA[a]	Signal transduction	0.0003
	DYRK1A	Transferase activity	0.0019
	ENPP2[a]	Cell motility, chemotaxis	0.0036
	GOLGA1[a]	Golgi autoantigen	0.0079
	GPLD1	Cell matrix adhesion	0.0037
	IL8RB[a]	Cell motility, chemotaxis	0.0090
	ITGB3	Cell matrix adhesion	0.0052
	ITSN1[a]	Calcium ion binding	0.0010
	LST1	Immune response	0.0090
	MAPK14	MAP kinase activity	0.0081
	NIFUN	Metal ion binding	0.0066
	NNT	Electron transport	0.0070
	NTRK3[a]	Cell differentiation	0.0048
	PPOX[a]	Electron transport	0.0030
	PTGER3	Cell death	0.0080
	RAB28[a]	GTPase activity	0.0044
	RAD9A	Regulation of cell cycle	0.0075
	RAE1[a]	Cytoskeleton	0.0021
	RBPMS	RNA processing	0.0070
	SELL	Cell motility	0.0060
	SLC15A1	Oligopeptide transport	0.0080
	SLC17A4	Sodium ion transport	0.0100
	SON[a]	Antiapoptosis	0.0027
	SUOX	Metal ion binding	0.0070
	TIAL1	Induction of apoptosis	0.0060
	UBE2I	Ubiquitin cycle	0.0040
	CCR2[a]	C-C chemokine receptor activity	0.0001
	CDC2L5	Positive regulation of cell proliferation	0.0044
	CUGBP2	Nuclear mRNA splicing, via spliceosome	0.0045
	DBH	Catecholamine biosynthesis	0.0031
	FLT3	Positive regulation of cell proliferation	0.0067
	FUCA1[a]	Carbohydrate metabolism	0.0026
	GALGT	Carbohydrate metabolism	0.0014
	GCS1	Carbohydrate metabolism	0.0044
	GGCX	γ-glutamyl carboxylase activity	0.0034
	GLG1	Fibroblast growth factor binding	0.0100
	GM2A	Sphingolipid catabolism	0.0100
	GNAT2[a]	G protein-coupled receptor protein signaling pathway	0.0027
	GNRH1	Negative regulation of cell proliferation	0.0083
	GTF2I	Transcription factor activity	0.0042
	INSR	Epidermal growth factor receptor activity	0.0070
	MAP4K1[a]	MAP kinase kinase kinase kinase activity	0.0060
	MAPK10	MAP kinase kinase activity	0.0090
	MEF2C[a]	Transcription factor activity	0.0032
	MLLT10[a]	Transcription factor activity	0.0016
	NFATC3	Transcription factor activity	0.0090
	NR1H4[a]	Transcription factor activity	0.0000
	PBP[a]	Serine-type endopeptidase inhibitor activity	0.0059
	PIGC[a]	Transferase activity, transfers glycosyl groups	0.0100
	PIK3R1[a]	Phosphatidylinositol 3-kinase activity	0.0050
	PKNOX1[a]	Transcription factor activity	0.0031
	PRKACA[a]	cAMP-dependent protein kinase activity	0.0001
	RPS14	Structural constituent of ribosome	0.0073
	SAA4	Lipid transporter activity	0.0044
	SLC35B1	UDP-galactose transporter activity	0.0080
	SNX1[a]	Intracellular protein transport	0.0055
	SSR2	Cotranslational protein-membrane targeting	0.0046
	SUPT4H1	Positive regulation of transcription	0.0003

Table 2. Continued.

Expression	Gene	Function	*p*-Value
	TMED9	Intracellular protein transport	0.0059
	TMSB4X[a]	Regulation of actin cytoskeleton	0.0027
	TRA2A[a]	Nuclear mRNA splicing, via spliceosome	0.0000
	ZNFN1A1[a]	Regulation of transcription, DNA-dependent	0.0002
Overexpressed in high-risk patients	*ABCA2*	ATPase activity	0.0024
	ABCC1[a]	Response to drug	0.0001
	ADAM17[a]	Cell-cell signaling	0.0020
	BLM[a]	DNA repair	0.0100
	C4orf10		0.0071
	CHERP[a]	Neurogenesis	0.0013
	CRABP1[a]	Signal transduction	0.0100
	HMGB2	DNA repair	0.0090
	INHA[a]	induction of apoptosis	0.0083
	LY6D[a]	Cell adhesion	0.0100
	NID[a]	Cell matrix adhesion	0.0009
	NOTCH3[a]	Cell differentiation	0.0011
	PCDHGA12[a]	Cell adhesion	0.0023
	PFN2[a]	Actin cytoskeleton	0.0074
	PSEN1[a]	Antiapoptosis	0.0013
	RIF1	ATPase activity	0.0085
	SEMA3F	Extracellular space	0.0100
	SH3GL2	Transferase activity	0.0100
	SMC1L1[a]	Chromatin binding	0.0036
	STC1[a]	Cell division	0.0090
	SUMO1	Ubiquitin cycle	0.0048
	TLK1	Cell cycle	0.0068
	TOP3B	DNA modification	0.0033
	UBE3A	Ubiquitin cycle	0.0039
	UGP2	Kinase activity	0.0059
	ZWINTAS[a]	Cell cycle	0.0019
	ARL4A[a]	Small GTPase-mediated signal transduction	0.0012
	BIK[a]	Induction of apoptosis	0.0040
	CBX3	Regulation of transcription, DNA-dependent	0.0090
	DSP[a]	Cell-cell adherens junction	0.0020
	ETV6	Transcription factor activity	0.0088
	FBN2[a]	Extracellular matrix structural constituent	0.0083
	GABRA3	γ-aminobutyric acid signaling pathway	0.0045
	GLI2[a]	Transcription factor activity	0.0085
	HNRPD[a]	Regulation of transcription, DNA-dependent	0.0031
	IRS1[a]	Insulin receptor binding	0.0072
	LARS2[a]	Leucine-tRNA ligase activity	0.0020
	LHCGR	Lutropin-choriogonadotropic hormone receptor activity	0.0039
	OLFM1	Latrotoxin receptor activity	0.0100
	PLEC1[a]	Structural constituent of cytoskeleton	0.0030
	PPP2R4	Phosphatase activator activity	0.0026
	PYGL[a]	Glycogen metabolism	0.0025
	SLC2A1[a]	Carbohydrate transport	0.0024
	SLC7A1[a]	Basic amino acid transporter activity	0.0016
	SMARCA3[a]	Chromatin modification	0.0066
	TFAM	DNA-dependent DNA replication	0.0014
	UPK2[a]	Membrane organization and biogenesis	0.0055
	VARS2	Translational elongation	0.0072
	VGLL1[a]	Transcription regulator activity	0.0032
	ZNF154[a]	Regulation of transcription, DNA-dependent	0.0044
	ZNF410[a]	Regulation of transcription, DNA-dependent	0.0004

[a] Included in group of 64 genes chosen for calculating risk scores of overall survival.

doi:10.1371/journal.pmed.0030467.t002

Quantitative RT-PCR Analysis

Using the samples from dataset 1, the relative expressions of nine randomly selected genes associated with survival were determined by quantitative RT-PCR (QRT-PCR) analysis as described in a previous report [17]. Primers for the QRT-PCR analysis (Table S2) were designed using Primer Express software version 2.0 (Applied Biosystems [http://www.appliedbiosystems.com]). Amplification of each target DNA was performed with SYBR Green master mix in Bio-Rad (http://www.bio-rad.com) Single Color Real-Time PCR Detection System according to the protocols provided. The control gene β-*actin* and the target genes amplified with

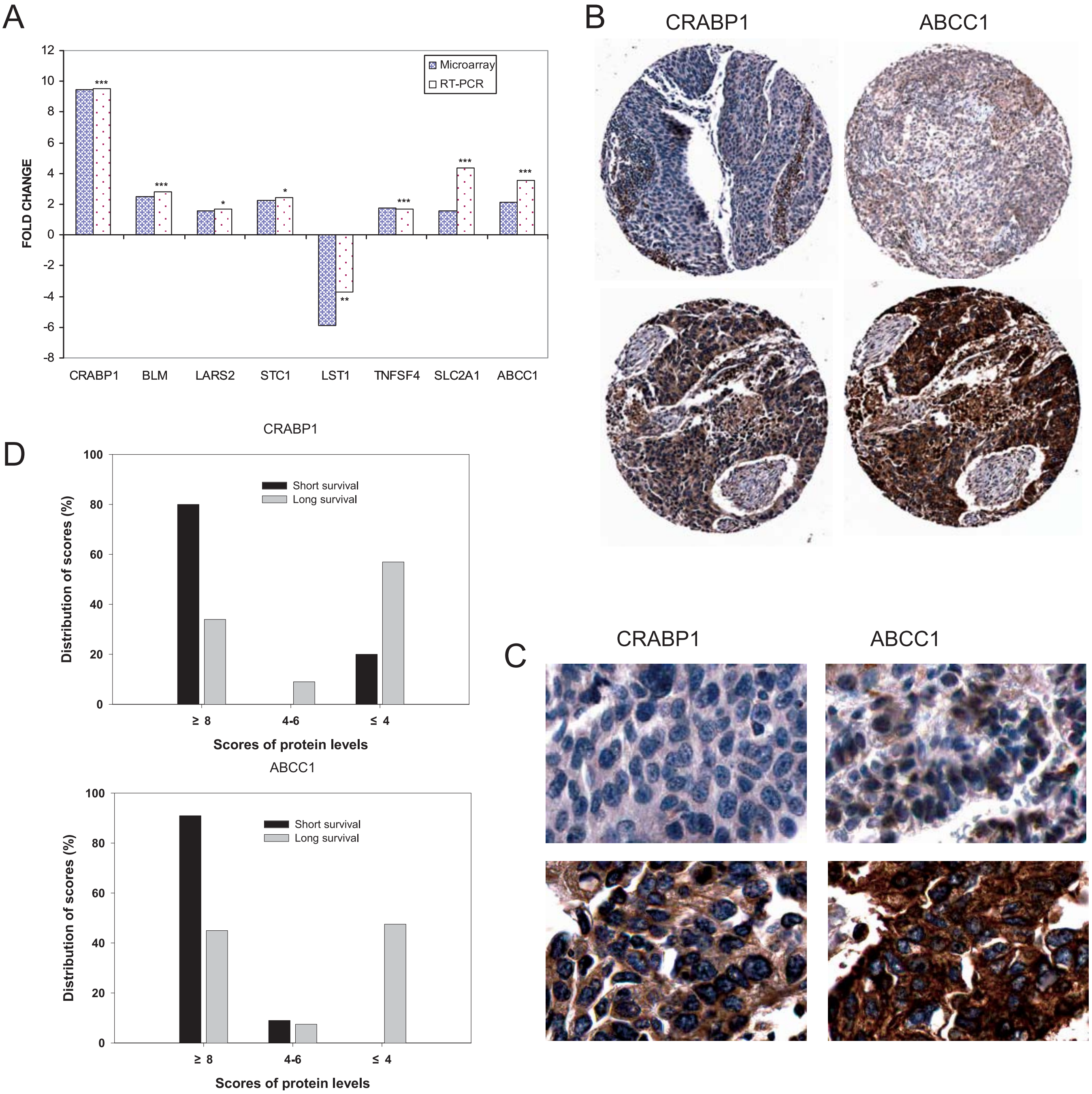

Figure 1. Validation Analyses of Gene Expression Profiling

(A) QRT-PCR validations of several candidate survival-related genes. Bars represent fold changes for the selected genes with differential expression between long- (>5 y) and short-term survival (<2 y) patients. Positive fold change represents up-regulated, and negative fold change represents down-regulated in short-term survival patients. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.005$.

(B and C) Immunostaining analysis of CRABP1 and ABCC1 expression in long- and short- term survival lung cancer patients. Low magnification (B) and 40× (C). Positive CRABP1 immunoreactivity was observed in cytoplasm of an acinar ADC (lower left photomicrographs of B and C) from short-term survival patients, and no CRABP1 reactivity was seen in a lung ADC from a long-term survival patient (upper left). Strong ABCC1 membranous staining (lower right) in tumor cells from short-term survival patients was observed, and weak ABCC1 reactivity was seen in a lung ADC from a long-term survival patient (upper right).

(D) Distribution of CRABP1 and ABCC1 protein levels in short- and long-term survival patients.

doi:10.1371/journal.pmed.0030467.g001

equal efficiencies. To assess whether two amplicons have the same efficiency, the variation of ΔC_T ($C_{T,\text{target}} - C_{T,\beta\text{-actin}}$, where C_T is cycle number at which the fluorescence signal exceeds background) with template dilution was evaluated [18]. The fold change of gene expression in long-term survival patients relative to short-term survival patients was calculated as $2^{-\Delta\Delta CT}$ ($\Delta\Delta C_T = \Delta C_{T\ \text{long}} - \Delta C_{T\ \text{short}}$). ANOVA was performed to determine differences among the groups.

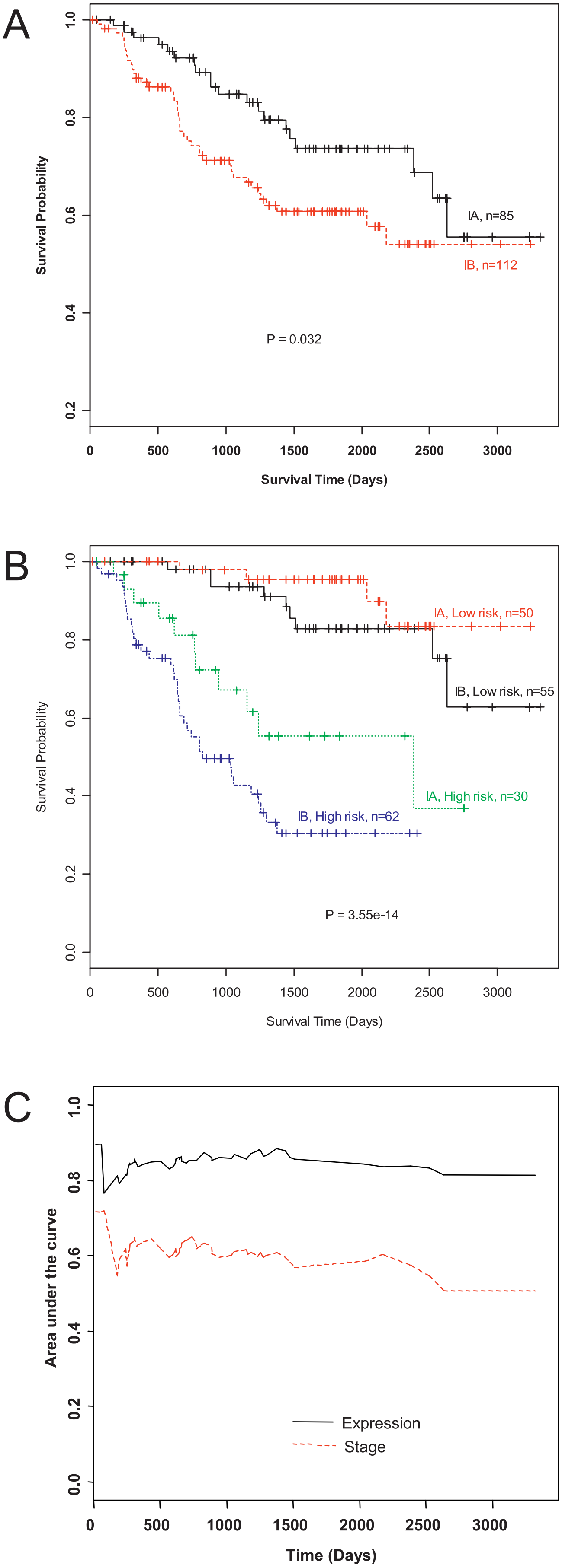

Figure 2. Survival Analyses of Stage I NSCLC

(A) Kaplan-Meier survival curves for patients with stage IA and with IB NSCLC.

(B) Kaplan-Meier survival curves for stage IA and IB patients defined by having positive (high-risk) or negative (low-risk) risk scores of overall survival. The risk scores were estimated with seven principle components based on the model built by 64 survival-related genes identified in five datasets.

(C) Area under the ROC curve for survival models based on stage information or expression data, respectively.

doi:10.1371/journal.pmed.0030467.g002

A *p*-value of less than 0.05 was considered to indicate statistical significance.

Tissue Microarray

Lung tissues of 60 stage I NSCLC patients (including 12 patients dead by 2 years after surgical resection and 48 alive for more than 5 years) were collected during surgery between 1985 and 1999 at the Arthur James Cancer Hospital of the Ohio State University Medical School (Columbus, Ohio, United States). All tissues were fixed in formalin and embedded in paraffin. A patient tissue microarray was constructed from these tissues for examination of CRABP1 and ABCC1 immunoreactivity in short- and long-term survival patients. All antibodies were antigen-retrieved in a vegetable steamer with TRS, pH 6.1 (Dako [http://www.dako.com]), staining was performed on a Dako autostainer, and all primary incubations were for 1 h at room temperature. For CRABP1 (Abcam [http://www.abcam.com, #Ab2816], dilution 1:1000), detection kit used was LSAB+ (Dako). For ABCC1 (AXXORA [http://www.axxora.com, ALX-801-007-C125], dilution 1:50), the detection kit used was Vectastain Elite (Vector Labs [http://www.vectorlabs.com]). The immunohistochemical staining images were scanned using an ImageScope (Aperio [http://www.aperio.com]). The percentage of positive cancer cells was scored on a semiquantitative scale as 0 (0%), 1 (1%–25%), 2 (25%–50%), 3 (50%–75%), and 4 (75%–100%). Intensity was scored as 1 (weak), 2 (intermediate), and 3 (strong). Results were calculated by multiplying the score of percentage of positive cells (P) by the intensity (I). The maximum score was 12. Two investigators did the evaluation of immunostaining results independently. Student's t-test was used in statistical analyses.

Results

Differentially Expressed Genes Associated with Survival

Tables 2 and S3 list the genes related to overall survival in the combined data ($p < 0.01$). As shown in Table 2, we observed relatively consistent changes for both genes whose expression in low-risk patients is higher than in high-risk patients and genes whose expression in high-risk patients is higher than in low-risk patients. Since we did not use data from normal paired lungs in these analyses, it is not clear whether these genes are all overexpressed in both low-risk and high-risk patients. Therefore, there are at least four possibilities of gene-expression patterns: (1) one group of genes overexpressed in low-risk patients and another group of genes overexpressed in high-risk patients; (2) one group of genes overexpressed in high-risk patients and another group of genes underexpressed in high-risk patients; (3) one group of genes overexpressed in low-risk patients and another

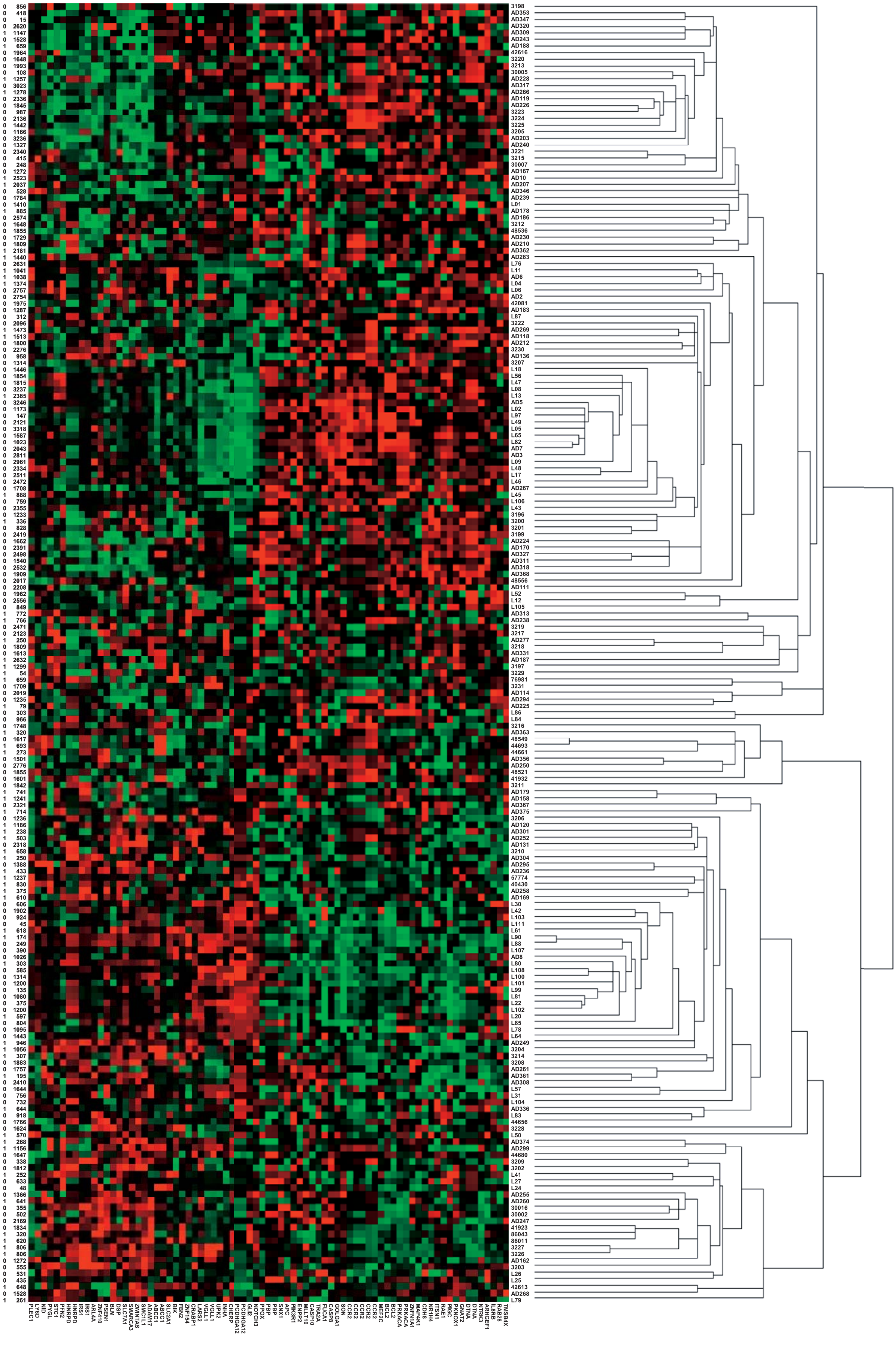
PLEC1
LY6D
NID
PYGL
STC1
PFN2
HNRPD
HNRPD
IRS1
IRS1
ARL4A
ZNF410
PSEN1
BLM
DSP
SLC7A1
SMARCA3
ZWINTAS
SMC1L1
ADAM17
ABCC1
ABCC1
SLC2A1
BIK
FBN2
ZNF154
CRABP1
LARS2
VGLL1
VGLL1
UPK2
INHA
CHERP
PCDHGA12
PCDHGA12
GLI2
NOTCH3
PPOX
PBP
PBP
SNX1
APC
PIK3R1
ENPP2
MLLT10
CASP10
TRA2A
FUCA1
CASP8
GOLGA1
SON
CCR2
CCR2
CCR2
CCR2
CCR2
MEF2C
BCL2
BCL2
PRKACA
PRKACA
ZNFN1A1
MAP4K1
CDH8
NR1H4
ITSN1
RAE1
PIGC
PKNOX1
GNAT2
DTNA
DTNA
NTRK3
ARHGEF1
IL8RB
RAB28
TMSB4X

Figure 3. Gene Expression Patterns of 64 Top Survival Genes for 197 NSCLC Patients from Datasets 1 to 5
Patients were generally classified into two groups (short-term versus long-term survival) with distinct expression patterns. The first column on the left represents patient status: 0, alive; 1, dead; the second column on the left represents follow-up time (days).
doi:10.1371/journal.pmed.0030467.g003

group of genes underexpressed in low-risk patients; or (4) a mixture of all three scenarios. In order to clarify this issue, we have begun to systematically acquire microarray information from all paired normal lungs in an attempt to determine possible expression patterns of these survival-related genes by comparing normal gene expression levels between low-risk and high-risk patients. The results from this ongoing study will be included a future publication.

Most of these genes are related to cell adhesion, cell motility, cell proliferation, and apoptosis. Notably, several genes have been reported to be associated with cancer survival *(APC, IRS1, SLC2A1, BCL2, ABCC1, FLT3, RAD9A, Inhibin A, NTRK3, CASP8,* and *CASP10).* The *APC* gene plays a role in NSCLC, and high *APC* promoter methylation is significantly associated with poor survival in NSCLC [19]. *IRS1* is a high-risk classifier gene associated with cancer death within 12 months [8]. In another study, *BCL2* was observed to be up-regulated in a group of long-term survival patients with NSCLC [20]. *ABCC1* expression levels have been shown to be an independent predictor for disease-free survival in adult acute myeloid leukemia [21]. Acute myeloid leukemia patients with *FLT3* mutations tend to have poor prognoses [22]. In addition, *RAD9A*, which is involved in DNA repair, was found to be increased in radioresistant cells over radiosensitive cells [23]. *Inhibin A* was found to be overexpressed in two cases of primary clear cell renal cell carcinoma (2/16 [13%]) and three cases of metastatic clear cell renal cell carcinoma (3/5 [60%]) [24]. The expression of CASP8 and CASP10 was frequently decreased at the mRNA and protein levels in lung cancer progression [25]. We found that the genes encoding these two caspases were up-regulated in long-term survival patients. High *NTRK3* mRNA expression generally presages longer survival [26]. Four survival genes *(TMSB4X, INHA, FUCA1,* and *STC1)* were previously identified by the cross-validation procedure in dataset 3 [5]. Not surprisingly, several genes were reported to be involved in cancer progression, survival, or cancer subtypes in the original reports. For example, *ATP2B1, AKAP12, TNFAIP6, RGS16, HSPA8, RPS3, ADM,* and *P2RX5* are survival-associated genes [8]. *MUC1* may play a role in progression and invasiveness of colorectal carcinomas [27]. Finally *AGT, XBP1,* and *PODXL* are overexpressed in ADC compared with SCC [28].

Validation of Selected Genes

To validate the microarray gene expression results from the meta-analysis, the relative expression of nine genes associated with overall survival *(CRABP1, BLM, ABCC1, SLC2A1, TNFSF4, BCL2, LST1, STC1,* and *LARS2)* was determined by QRT-PCR analysis on the samples from dataset 1. We confirmed the expression results for all these nine genes except *BCL2* ($p \leq 0.05$) (Figure 1A).

The patient tissue microarray of 60 completely independent patients was interrogated for *CRABP1* and *ABCC1* to determine if mRNA changes were correlated with increased protein expression in lung ADCs from patients with short-term survival. CRABP1 staining was observed in the cytoplasm of tumor cells in most lung tumor tissues (Figure 1B and 1C). CRABP1 exhibited stronger staining in tissues of short-term survival patients than in those of long-term survival patients (Figures 1B, 1C, and S3), the scores for short- and long-term survival were 8.8 ± 3.1(mean ± SD, same hereafter) and 4.9 ± 2.8 ($p < 0.0001$), respectively. In short-term survival patients, 80% and 20% of the samples had scores of 8 or higher and 4 or lower in CRABP1 immunostaining, respectively; in contrast, in long-term survival patients, 34% and 59% of the samples had scores of 8 or higher and 4 or lower, respectively (Figure 1D). Similar trends were also observed in mRNA levels in our samples (see Figure 1A). ABCC1 showed either membranous or cytoplasmic staining in tumor cells of tissues of both short- and long-term survival NSCLC patients (Figures 1B, 1C, and S3). A significant increase in scores of ABCC1-positive staining was also present in tissues of short-term survival patients; the scores in tissues of short-term and long-term survival patients were 10 ± 2.3 and 6.6 ± 3.1 ($p = 0.002$), respectively. In short-term survival patients, 91% and 9% of the samples had ABCC1 immunostaining scores of 8 or higher and 4 or lower, respectively; in long-term survival patients, 45% and 48% of the samples had scores 8 or higher and 4 or lower, respectively. The results indicate that expression of these two proteins is consistent with the results from both microarray and RT-PCR analyses. Higher protein levels of CRABP1 and ABCC1 tend to increase risk of short survival of stage I NSCLC patients.

Identification of a Gene Expression Signature for Survival

Next, we determined if a subset of the genes related to overall survival can be used to predict survival of patients with stage I NSCLC. Risk scores were derived from survival analyses of all 197 samples in datasets 1 to 5 with the partial Cox regression. Kaplan-Meier survival analyses were performed after the samples were classified into high- and low-risk groups according to the risk scores. As shown in Figure 2A, Kaplan-Meier survival curves indicated poorer survival in stage IB than in stage IA NSCLC ($p = 0.032$). To determine whether gene expression profiles could accurately predict overall survival, the risk scores calculated by the 64 genes (listed in Table 2) were used to classify all of the samples from datasets 1 to 5 into two groups as high and low risk groups. Kaplan-Meier analysis using expression profiles demonstrated that the high and low-risk groups were significantly different in their overall survival ($p < 0.001$) (Figure 2B). A comparison of Figure 2A and 2B clearly shows that the gene expression signature has higher classification power than the staging method. The former has a larger area between the two risk groups and a smaller p-value from the Mantel-Haenszel log rank test. Figure 2C shows the time-dependent area under the ROC curves based on the stage information or the estimated risk scores of the patients. We observed that the Cox model with gene expression data gave the better predictive performance with the areas under the ROC curve close to 80%. The

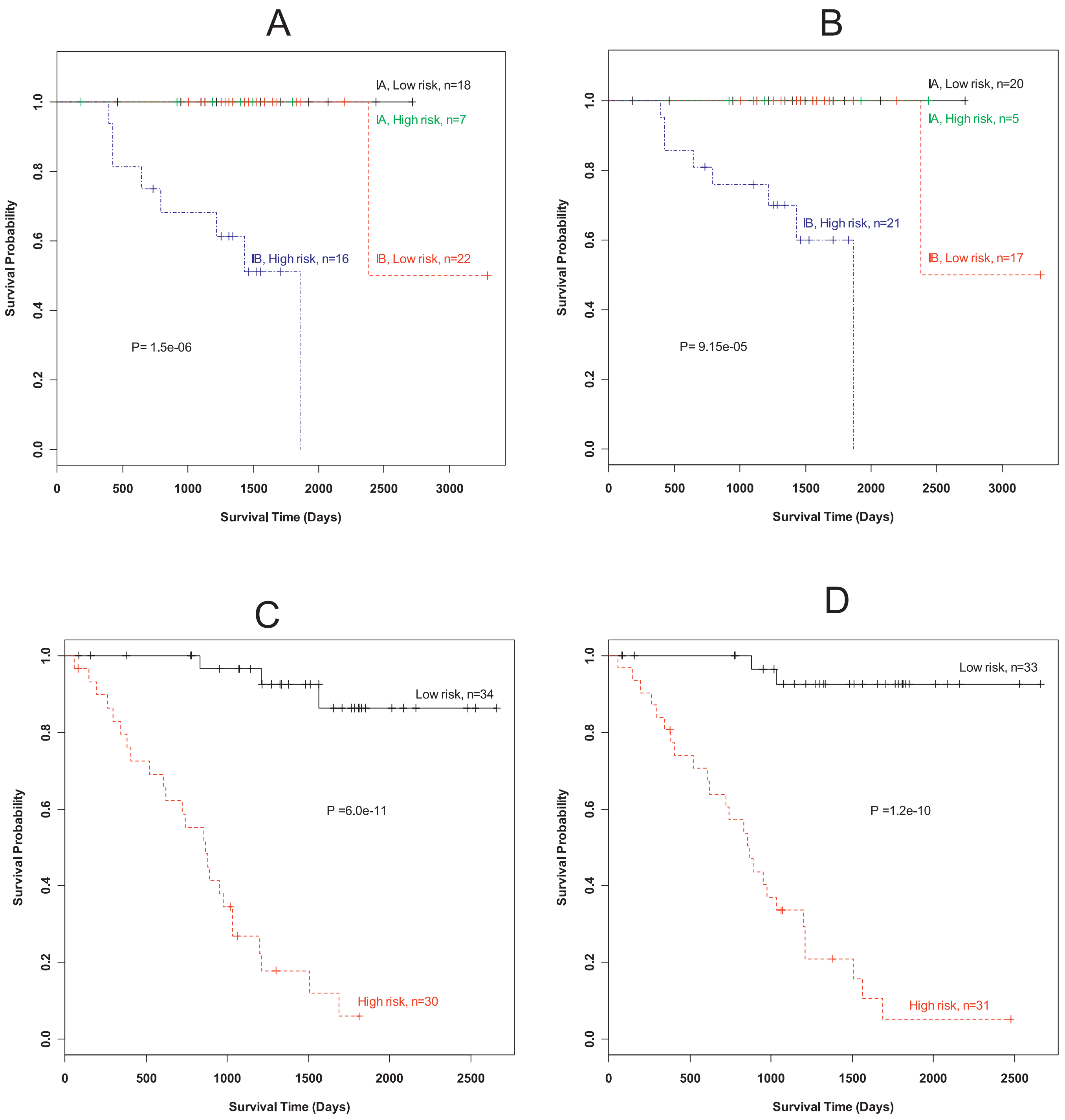

Figure 4. Comparison of the Prediction Accuracy of Lung Cancer Survival Using Our 64-Gene Signature and a Different 50-Gene Signature

(A and B) Kaplan-Meier survival curves for dataset 6 under our 64-gene signature (A) and the 50-gene signature from Beer et al. [5] (B). Scores were estimated using two principle components.

(C and D) Kaplan-Meier survival curves for dataset 7 using our 64-gene signature (C) and the 50-gene signature from Beer et al. [5] (D). Scores were estimated using eight principle components.

doi:10.1371/journal.pmed.0030467.g004

Cox model with stage information, in contrast, resulted in areas under the curve below 60%.

Patients with a postoperative survival of at least 5 years and those who died within 2 years after resection were selected for estimating the predictive power of Kaplan-Meier survival analysis using expression profiles. According to the risk scores by the partial Cox regression approach, 77 of the 88 patients were classified correctly (87% accuracy) (Table S4). Gene expression patterns were determined using hierarchical clustering of the 197 NSCLC samples against the 64 top survival-related genes (Figure 3). Short- and long-term survival NSCLC patients had distinct expression patterns among the 64 genes that were used for establishing

Table 3. The Signature Genes of Survival Identified in Our Meta-Analysis Overlap Those in Previous Studies of Lung Cancer Survival

Gene	Methods of Analysis	References
APC	Hypermethylation, quantitative fluorogenic RT-PCR	[19]
BCL2	Hemotoxylin-eosin staining	[39]
	Meta-analysis	[40]
FBN2	RT-PCR and methylation-specific PCR	[41]
TMSB4X	Microarray	[5]
FUCA1	Microarray	[5]
STC1	Microarray	[5]
SLC2A1	Immunohistochemical analysis	[33]
	Immunohistochemical analysis	[42]
	Microarray	[5]
INHA	Microarray	[5]
ABCC1	Review	[43]

doi:10.1371/journal.pmed.0030467.t003

a gene expression signature predictive of stage I NSCLC survival.

Confirmation of the Gene Expression Signature in Independent Datasets

The robustness of the 64-gene expression signature in predicting survival in lung cancer was further tested with oligonucleotide gene expression data obtained from two completely independent datasets—dataset 6 (63 stage I lung ADC including nine long-term survival patients and five short-term survivors) and dataset 7 (64 stage I lung ADC and SCC, including eight long-term survival patients and twelve short-term survivors). When we examined the risk assignment of the samples in these two datasets using the risk scores based on our 64-gene signature, high- and low-risk groups were observed that differed significantly in survival in datasets 6 and 7. One of 14 individuals was classified incorrectly (93% accuracy) using the 64-gene signature in dataset 6 ($p < 0.001$; Figure 4A). In dataset 7, we correctly classified all 20 patients who survived for at least 5 years or died within 2 years using the 64-gene signature ($p < 0.001$; Figure 4C). We also examined the risk assignment of the samples in these two datasets using the risk scores based on the 50-gene signature reported by Beer et al. [5]. Classification was less accurate with this gene signature: three patients who lived for more than 5 years in dataset 6 were classified into the low-risk group according the risk scores calculated by our gene signature, but all of these patients were classified into the high-risk group under the Beer et al. [5] gene signature (Figure 4A and 4B). Also, in dataset 7 one patient surviving for more than 2,476 days was classified into the high-risk group under the Beer et al. [5] gene signature (Figure 4D).

Discussion

In this study, we combined several lung cancer gene expression studies based on Affymetrix microarray platforms into a single, homogeneous dataset by using the DWD method. The increased sample size was intended to reduce false positives and increase statistical power in detecting survival-related genes. This improved dataset enabled us to identify a gene expression signature consisting of 64 genes that can accurately predict which stage I lung cancer patients may experience poor survival following resection.

In general, the pathologic diagnosis used to classify a lung tumor is combined with the stage of the cancer to predict patient survival and direct therapy [29]. Unfortunately, current methods of classification and staging are not completely reliable or sufficiently precise [2], and no reliable markers exist to predict the presence of micrometastasis or outcome in patients with resected NSCLC. It is not unusual for patients with lung cancers of identical histology, differentiation, location, and stage to have major differences in survival or response to therapy [29]. Some patients diagnosed with stage I NSCLC survive after surgery for some time, whereas others do not. This prognostic variability makes the results of this study important. Patients whose early-stage tumors contain signatures predicting short survival times would benefit from the aggressive therapies currently given only to those with later-stage cancers.

In this study, we included cancer subtypes as a factor in the ANOVA model to choose survival-related genes, and we adjusted cancer subtypes in the multivariate Cox proportional hazards regression analyses. Therefore, the gene expression signatures identified in the current study should be suitable for both lung ADC and SCC. This generalization of our gene signatures was also demonstrated in two independent large datasets—dataset 6 (63 stage I lung ADC including nine long-term and five short-term survival patients) and dataset 7 (64 stage I lung ADC and SCC, including eight long-term and 12 short-term survival patients). Our gene signatures can accurately predict patient survivals in these two datasets with mixed stage I lung cancer subtypes. To our knowledge, such signatures have not been convincingly reported previously, and we propose that they should be used to inform the clinical management of lung cancer patients.

Our survival gene signatures consist of genes that are involved in cancer metastasis such as cell adhesion *(APC, CDH8, DSP, LY6D, PCDHGA12,* and *NID),* cell motility *(IL8RB, ENPP2,* and *CCR2),* and inflammation and immune response *(CASP8* and *CASP10).* In addition, seven of the genes are related to apoptosis *(INHA, PSEN1, CASP8, CASP10, PIK3R1, BCL2,* and *BIK)* and another five are related to transport mechanisms *(ABCC1, ITSN1, CRABP1, SLC2A1,* and *ZWINTAS).* Nine of the signature genes were previously identified as lung cancer survival factors (Table 3), and 29 genes have been associated with survival in other cancer types including breast carcinoma, brain cancer, and gastric cancer (Table 4).

ABCC1 and *SLC2A1* are particularly attractive biomarkers for survival in NSCLC. The protein encoded by *ABCC1* is a member of the superfamily of ATP-binding cassette transporters. ATP-binding cassette proteins transport various molecules across extra- and intracellular membranes. This full transporter is a member of the multidrug resistance-associated protein subfamily, and it functions as a multispecific organic anion transporter, with oxidized glutathione, cysteinyl leukotrienes, and activated aflatoxin B1 as substrates. This protein also transports glucuronides and sulfate conjugates of steroid hormones and bile salts. ABCC1 overexpression is associated with DNA aneuploid carcinomatous cells in NSCLC [30]. SLC2A1 is a major glucose transporter, which is an integral membrane glycoprotein

Table 4. The Signature Genes of Survival Identified in Our Meta-Analysis Are Also Involved in Survival of Other Cancer Types

Gene	Cancer Type	Measure	Survival Time (y)	Class 1 (*n*)	Class 2 (*n*)	t-Test	*p*-Value[a]	Reference
CRABP1	Brain	Survival	3	Alive (13)	Dead (56)	−2.907	0.005	[44]
	Brain	Survival	3	Alive (9)	Dead (20)	−2.107	0.045	[45]
BLM	Breast carcinoma	Disease-free survival	5	No disease (196)	Relapse (79)	−4.363	<0.001	[46]
ABCC1	Gastric cancer	Survival	3	Alive (23)	Dead (40)	−3.033	0.005	[47]
	Brain	Survival	3	Alive (13)	Dead (56)	−2.806	0.009	[44]
	Breast carcinoma	Disease-free survival	5	No disease (180)	Relapse (93)	−2.857	0.005	[48]
SLC2A1	Breast carcinoma	Disease-free survival	5	No disease (196)	Relapse (79)	−3.409	<0.001	[46]
BCL2	Breast carcinoma	Disease-free survival	5	No disease (196)	Relapse (79)	4.243	<0.001	[46]
	Breast carcinoma	Relapse-free survival	5	No disease (11)	Relapse (35)	3.022	0.007	[49]
LST1	Breast carcinoma	Disease-free survival	5	No disease (180)	Relapse (93)	3.568	<0.001	[48]
	Breast carcinoma	Metastases	5	Negative (51)	positive (46)	2.806	0.006	[50]
STC1	Brain	Survival	3	Alive (9)	Dead (20)	−3.762	<0.001	[45]
	Brain	Survival	3	Alive (13)	Dead (56)	−3.634	<0.001	[44]
CDH8	Brain	Survival	3	Alive (13)	Dead (56)	4.139	<0.001	[44]
ENPP2	Breast carcinoma	Disease-free survival	5	No Disease (180)	Relapse (93)	2.742	0.007	[48]
ITSN1	Brain	Survival	3	Alive (13)	Dead (56)	2.351	0.033	[44]
	Breast carcinoma	Disease-free survival	5	No Disease (180)	Relapse (93)	2.753	0.007	[48]
PBP	Brain	Survival	3	Alive (13)	Dead (56)	3.328	0.003	[44]
	Medulloblastoma	Survival	5	Alive (21)	Dead (18)	3.505	0.001	[51]
	Breast carcinoma	Disease-free survival	5	No disease (180)	Relapse (93)	3.019	0.003	[48]
	Breast carcinoma	Metastases	5	Negative (51)	Positive (46)	2.273	0.025	[50]
	Breast Ductal Carcinoma	Survival	5	Alive (70)	Dead (28)	2.072	0.043	[52]
PIK3R1	Brain	Survival	3	Alive (13)	Dead (56)	3.42	0.002	[44]
	Brain	Survival	3	Alive (9)	Dead (20)	2.249	0.033	[45]
	Breast carcinoma	Disease-free survival	5	No disease (196)	Relapse (79)	2.871	0.005	[46]
	Breast carcinoma	Disease-free survival	5	No disease (180)	Relapse (93)	2.711	0.007	[48]
DSP	Breast carcinoma	Metastases	5	Negative (51)	Positive (46)	−3.338	0.001	[50]
	Breast ductal carcinoma	Survival	5	Alive (70)	Dead (28)	−2.038	0.045	[52]
CHERP	Brain	Survival	3	Alive (9)	Dead (20)	−3.075	0.007	[45]
	Brain	Survival	3	Alive (13)	Dead (56)	−2.429	0.029	[44]
SLC7A1	Breast carcinoma	Disease-free survival	5	No disease (196)	Relapse (79)	−3.476	<0.001	[46]
	Breast carcinoma	Disease-free survival	5	No disease (180)	Relapse (93)	−2.72	0.007	[48]
	Breast carcinoma	Metastases	5	Negative (51)	Positive (46)	−3.263	0.002	[50]
HNRPD	Brain	Survival	3	Alive (13)	Dead (56)	−2.158	0.038	[44]
IRS1	Clear cell renal cell carcinoma	Survival	3	Alive (7)	Dead (33)	−3.316	0.004	[53]
	Brain	Survival	3	Alive (13)	Dead (56)	−4.539	<0.001	[44]
CASP10	Brain	Survival	3	Alive (13)	Dead (56)	3.269	0.004	[44]
ARHGEF1	Breast carcinoma	Disease-free survival	5	No disease (180)	Relapse (93)	2.634	0.009	[48]
PRKACA	Diffuse large B cell lymphoma	Survival	5	Alive (86)	Dead (121)	3.387	<0.001	[54]
PYGL	Brain	Survival	3	Alive (13)	Dead (56)	−4.539	<0.001	[44]
GLI2	Gastric cancer	Survival	3	Alive (23)	Dead (40)	−2.731	0.009	[47]
ZWINTAS	Breast carcinoma	Disease-free survival	5	No disease (196)	Relapse (79)	−5.985	<0.001	[46]
	Breast carcinoma	Disease-free survival	5	No disease (180)	Relapse (93)	−2.825	0.005	[48]
	Breast carcinoma	Metastases	5	Negative (51)	positive (46)	−3.248	0.002	[50]
PLEC1	Clear cell renal cell carcinoma	Survival	3	Alive (7)	Dead (33)	−3.316	0.004	[53]
VGLL1	Breast carcinoma	Disease-free survival	5	No disease (196)	Relapse (79)	−3.202	0.002	[46]
	Breast carcinoma	Metastases	5	Negative (51)	positive (46)	−2.327	0.023	[50]
	Breast carcinoma	Relapse-free survival	5	No Disease (11)	Relapse (35)	−2.144	0.043	[49]
LY6D	Brain	Survival	3	Alive (13)	Dead (56)	−3.13	0.006	[44]
APC	Brain	Survival	3	Alive (13)	Dead (56)	3.402	0.002	[44]
	Brain	Survival	3	Alive (9)	Dead (20)	2.337	0.032	[45]
SMC1L1	Breast carcinoma	Disease-free survival	5	No disease (180)	Relapse (93)	−3.843	<0.001	[48]
	Breast carcinoma	Disease-free survival	5	No disease (196)	Relapse (79)	−3.374	<0.001	[46]
IL8RB	Breast carcinoma	Disease-free survival	5	No disease (180)	Relapse (93)	2.064	0.04	[48]

[a] *p*-Value indicates significance for t-test of difference in gene expression between two classes of patients.

doi:10.1371/journal.pmed.0030467.t004

involved in transporting glucose into most cells. Increased glucose transport in malignant cells has been associated with increased, deregulated expression of glucose transporter proteins that is characterized by the overexpression of SLC2A1 [31]. Differential expression levels of SLC2A1 have been observed between ADC and SCC [32], and overexpression of SLC2A1 in stage I NSCLC resulted in poor survival in another experiment [33]. Thus, these two genes could be targets of cancer therapy and prevention.

The survival-related genes identified in previous microarray studies of lung cancer patients failed to show consistency between studies, likely due to small patient sample numbers [34], and their predictive power was limited when tested in independent datasets. One solution to this problem is to integrate datasets from multiple studies to increase the sample size. Another problem is systematic biases due to different handling procedures in clinical studies, especially when samples/tumors are collected and processed at different institutions, using different microarray print batches, platforms, or array hybridization protocols. To integrate microarray datasets with different origins, distribution transformation methods, such as DWD, can be helpful. This method has been used previously to combine datasets from different batches into a single homogeneous dataset in head and neck SCC and breast carcinoma studies [35–38]. In our data analyses, we chose WUSM dataset as the reference batch, used the same mean and variance as reference, and then combined other datasets one by one. The hierarchical cluster analysis using the original nontransformed data classified the samples into five distinct groups according to data source rather than disease status (Figure S1A), demonstrating large systematic biases among the five studies. After DWD adjustment, however, all 197 samples from the five datasets were clustered into two sub-branches according to disease status rather than data source, each of which was composed of samples from all five datasets (Figure S1B). The batch differences disappeared; in this sense the samples from different datasets mixed well. Figure S4 also shows the effect of DWD adjustment using datasets 1 and 3. Principal component directions of adjusted data are markedly different from those of the raw data. We also performed poolability tests to examine if these DWD-transformed gene expression data from different resources are poolable. The results showed that the number of p-values falling in the tail in our data is similar to those in simulated data (Table S5). Therefore, the gene expression data from different resources are poolable after the DWD transformation, and thus can be combined for survival analysis. These results imply that the systematic biases were largely removed after DWD adjustment and thus the results from the integrated data should be robust.

Time to death due to cancer varies substantially among lung cancer patients. Studying censored survival time may be more informative than treating it as a binary or categorical variable. We applied multivariate Cox proportional hazards models with bootstrap resampling technology to the analysis of these censored survival data from different resource. Kaplan-Meier analysis using gene expression profiles demonstrated a significantly worse overall survival for high-risk patients compared to low-risk patients (Figure 2B), and using the 64-gene signature, we predicted the actual overall survival with greater than 85% accuracy. This new tool will help clinicians assess a patient's risk profile and to prescribe a course of treatment tailored to that profile. A patient whose cancer signature indicates that it is unlikely to metastasize would be spared the debilitating side effects of aggressive anticancer therapies, whereas a patient with an early but particularly aggressive tumor would be a candidate for aggressive treatment not usually given to early-stage patients, and thus experience improved survival.

Supporting Information

Figure S1. Hierarchical Clustering Analysis of Five Datasets

Analyses are shown for the raw data (A) and the DWD source and batch-adjusted data (B). Green, dataset 1 (WUSM); purple, dataset 2 (Mayo Clinic); blue, dataset 3 (Beer et al. [5]); red, dataset 4 (Bhattacharjee et al. [10]); yellow, dataset 5 (Borczuk et al. [8]).

Found at doi:10.1371/journal.pmed.0030467.sg001 (65 KB PDF).

Figure S2. Global p-Values from Tests of the Proportional Hazards Assumption for All Survival-Related Genes

Only two of 165 tests obtained global $p < 0.05$, indicating that proportional hazards models are statistically warranted for the survival analyses.

Found at doi:10.1371/journal.pmed.0030467.sg002 (72 KB PDF).

Figure S3. The Immunostaining Images of Lung Cancer Patient Tissue Microarray

The sections from short-term survival lung cancer patients are shown in box.

Found at doi:10.1371/journal.pmed.0030467.sg003 (1.3 MB PDF).

Figure S4. Principle Component Directions

Directions are given for the raw data (A) and the DWD source- and batch-adjusted data (B). Red, dataset 1 (WUSM); blue, dataset 3 (Beer et al. [5]).

Found at doi:10.1371/journal.pmed.0030467.sg004 (513 KB PDF).

Protocol S1. Detailed Description of the Data Analyses

Found at doi:10.1371/journal.pmed.0030467.sd001 (84 KB DOC).

Table S1. Sample Information on Datasets Used in the Meta-Analysis

Found at doi:10.1371/journal.pmed.0030467.st001 (73 KB XLS).

Table S2. Oligonucleotide Primers and Probes Used for RT-PCR Analysis

Found at doi:10.1371/journal.pmed.0030467.st002 (23 KB XLS).

Table S3. Detailed Information on Genes Related to Cancer Survival

Found at doi:10.1371/journal.pmed.0030467.st003 (32 KB XLS).

Table S4. Partial Cox Regression Classification of 197 stage I NSCLC patient using 64 Survival-Related Genes

Found at doi:10.1371/journal.pmed.0030467.st004 (32 KB XLS).

Table S5. Comparison of the Distribution of p-Values from Poolability Tests in the Real and Simulated Data

Found at doi:10.1371/journal.pmed.0030467.st005 (15 KB XLS).

Accession Numbers

Accession numbers for genes related to cancer survival in Table S3 can be found in Nucleotide (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide). The Gene Expression Omnibus (GEO [http://www.ncbi.nlm.nih.gov/geo]) accession number for microarray data from this study is GSE6253.

Acknowledgments

We thank Dr. David Beer at University of Michigan, Dr. Arindam Bhattacharjee at Harvard University, and Dr. Charles Powell at Columbia University for releasing microarray data. Some of the lung cancer tissues were procured from the tumor tissue bank at Washington University School of Medicine.

Author contributions. YL and MY designed the study. YL, WL, YY, PYL, and MY analyzed the data. ZS and RG enrolled patients. YL, PYL, YY, PY, MW, RG, and MY contributed to writing the paper. YL conducted statistical analyses and collected data from different institutes. YY and CM performed immunohistochemistry experiments and data analysis. ZS collected patient clinical information and did microarray experiments for the data from Mayo Clinic. MW designed, generated, and coordinated analysis of one of the datasets used in this study.

References

1. Spiro SG, Silvestri GA (2005) One hundred years of lung cancer. Am J Respir Crit Care Med 172: 523–529.
2. Bach PB, Kelley MJ, Tate RC, McCrory DC (2003) Screening for lung cancer: A review of the current literature. Chest 123: 72S–82S.
3. Mountain CF, Dresler CM (1997) Regional lymph node classification for lung cancer staging. Chest 111: 1718–1723.
4. El-Sherif A, Luketich JD, Landreneau RJ, Fernando HC (2005) New therapeutic approaches for early stage non-small cell lung cancer. Surg Oncol 14: 27–32.
5. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, et al. (2002) Gene-

expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 8: 816–824.

6. Sun Z, Yang P, Aubry MC, Kosari F, Endo C, et al. (2004) Can gene expression profiling predict survival for patients with squamous cell carcinoma of the lung? Mol Cancer 3: 35.
7. Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, et al. (2002) Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. Cancer Res 62: 3005–3008.
8. Borczuk AC, Shah L, Pearson GD, Walter KL, Wang L, et al. (2004) Molecular signatures in biopsy specimens of lung cancer. Am J Respir Crit Care Med 170: 167–174.
9. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.
10. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, et al. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 98: 13790–13795.
11. Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439: 353–357.
12. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, et al. (2003) NetAffx: Affymetrix probesets and annotations. Nucleic Acids Res 31: 82–86.
13. Benito M, Parker J, Du Q, Wu J, Xiang D, et al. (2004) Adjustment of systematic microarray data biases. Bioinformatics 20: 105–114.
14. Baltagi BH (2005) Econometric analysis of panel data. 3rd Ed. Hoboken, New Jersey: John Wiley and Sons. 314 pp.
15. Li H, Gui J (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics 20: I208–I215.
16. Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comput Graph Statist 5: 299–314.
17. Chaparro J, Reeds DN, Wen W, Xueping E, Klein S, et al. (2005) Alterations in thigh subcutaneous adipose tissue gene expression in protease inhibitor-based highly active antiretroviral therapy. Metabolism 54: 561–567.
18. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2(-\Delta\Delta C_T)$ method. Methods 25: 402–408.
19. Brabender J, Usadel H, Danenberg KD, Metzger R, Schneider PM, et al. (2001) Adenomatous polyposis coli gene promoter hypermethylation in non-small cell lung cancer is associated with survival. Oncogene 20: 3528–3532.
20. Mattern J, Koomagi R, Volm M (2002) Characteristics of long-term survivors of untreated lung cancer. Lung Cancer 36: 277–282.
21. Schaich M, Soucek S, Thiede C, Ehninger G, Illmer T (2005) MDR1 and MRP1 gene expression are independent predictors for treatment outcome in adult acute myeloid leukaemia. Br J Haematol 128: 324–332.
22. Drexler HG, Quentmeier H (2004) FLT3: Receptor and ligand. Growth Factors 22: 71–73.
23. Guo WF, Lin RX, Huang J, Zhou Z, Yang J, et al. (2005) Identification of differentially expressed genes contributing to radioresistance in lung cancer cells using microarray analysis. Radiat Res 164: 27–35.
24. Jung SM, Kuo TT (2005) Immunoreactivity of CD10 and inhibin alpha in differentiating hemangioblastoma of central nervous system from metastatic clear cell renal cell carcinoma. Mod Pathol 18: 788–794.
25. Shivapurkar N, Reddy J, Matta H, Sathyanarayana UG, Huang CX, et al. (2002) Loss of expression of death-inducing signaling complex (DISC) components in lung cancer cell lines and the influence of MYC amplification. Oncogene 21: 8510–8514.
26. Eberhart CG, Kratz J, Wang Y, Summers K, Stearns D, et al. (2004) Histopathological and molecular prognostic markers in medulloblastoma: c-Myc, N-myc, TrkC, and anaplasia. J Neuropathol Exp Neurol 63: 441–449.
27. Baldus SE, Monig SP, Huxel S, Landsberg S, Hanisch FG, et al. (2004) MUC1 and nuclear beta-catenin are coexpressed at the invasion front of colorectal carcinomas and are both correlated with tumor prognosis. Clin Cancer Res 10: 2790–2796.
28. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, et al. (2001) Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 98: 13784–13789.
29. Roukos DH, Kappas AM (2005) Perspectives in the treatment of gastric cancer. Nat Clin Pract Oncol 2: 98–107.
30. Doubre H, Cesari D, Mairovitz A, Benac C, Chantot-Bastaraud S, et al. (2005) Multidrug resistance-associated protein (MRP1) is overexpressed in DNA aneuploid carcinomatous cells in non-small cell lung cancer (NSCLC). Int J Cancer 113: 568–574.
31. Macheda ML, Rogers S, Best JD (2005) Molecular and cellular regulation of glucose transporter (GLUT) proteins in cancer. J Cell Physiol 202: 654–662.
32. Yamagata N, Shyr Y, Yanagisawa K, Edgerton M, Dang TP, et al. (2003) A training-testing approach to the molecular classification of resected non-small cell lung cancer. Clin Cancer Res 9: 4695–4704.
33. Younes M, Brown RW, Stephenson M, Gondo M, Cagle PT (1997) Overexpression of Glut1 and Glut3 in stage I nonsmall cell lung carcinoma is associated with poor survival. Cancer 80: 1046–1051.
34. Xu L, Tan AC, Naiman DQ, Geman D, Winslow RL (2005) Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics 21: 3905–3911.
35. Chung CH, Parker JS, Karaca G, Wu J, Funkhouser WK, et al. (2004) Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell 5: 489–500.
36. Hu Z, Fan C, Oh DS, Marron JS, He X, et al. (2006) The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 7: 96.
37. Oh DS, Troester MA, Usary J, Hu Z, He X, et al. (2006) Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. J Clin Oncol 24: 1656–1664.
38. Perreard L, Fan C, Quackenbush JF, Mullins M, Gauthier NP, et al. (2006) Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Res 8: R23.
39. Yilmaz A, Savas I, Dizbay Sak S, Gungor A, Kaya A, et al. (2005) Distribution of Bcl-2 gene expression and its prognostic value in non-small cell lung cancer. Tuberk Toraks 53: 323–329.
40. Martin B, Paesmans M, Berghmans T, Branle F, Ghisdal L, et al. (2003) Role of Bcl-2 as a prognostic factor for survival in lung cancer: A systematic review of the literature with meta-analysis. Br J Cancer 89: 55–64.
41. Chen H, Suzuki M, Nakamura Y, Ohira M, Ando S, et al. (2005) Aberrant methylation of FBN2 in human non-small cell lung cancer. Lung Cancer 50: 43–49.
42. Minami K, Saito Y, Imamura H, Okamura A (2002) Prognostic significance of p53, Ki-67, VEGF and Glut-1 in resected stage I adenocarcinoma of the lung. Lung Cancer 38: 51–57.
43. Yang P, Ebbert JO, Sun Z, Weinshilboum RM (2006) Role of the glutathione metabolic pathway in lung cancer treatment and prognosis: A review. J Clin Oncol 24: 1761–1769.
44. Freije WA, Castro-Vargas FE, Fang Z, Horvath S, Cloughesy T, et al. (2004) Gene expression profiling of gliomas strongly predicts survival. Cancer Res 64: 6503–6510.
45. Shai R, Shi T, Kremen TJ, Horvath S, Liau LM, et al. (2003) Gene expression profiling identifies molecular subtypes of gliomas. Oncogene 22: 4918–4923.
46. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999–2009.
47. Chen X, Leung SY, Yuen ST, Chu KM, Ji J, et al. (2003) Variation in gene expression patterns in human gastric cancers. Mol Biol Cell 14: 3208–3215.
48. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671–679.
49. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98: 10869–10874.
50. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415: 530–536.
51. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, et al. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415: 436–442.
52. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, et al. (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci U S A 100: 10393–10398.
53. Vasselli JR, Shih JH, Iyengar SR, Maranchie J, Riss J, et al. (2003) Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. Proc Natl Acad Sci U S A 100: 6958–6963.
54. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346: 1937–1947.

Editors' Summary

Background. Lung cancer is the commonest cause of cancer-related death worldwide. Most cases are of a type called non-small cell lung cancer (NSCLC) and are mainly caused by smoking. Like other cancers, how NSCLC is treated depends on the "stage" at which it is detected. Stage IA NSCLCs are small and confined to the lung and can be removed surgically; patients with slightly larger stage IB tumors often receive chemotherapy after surgery. In stage II NSCLC, cancer cells may be present in lymph nodes near the tumor. Surgery plus chemotherapy is the usual treatment for this stage and for some stage III NSCLCs. However, in this stage, the tumor can be present throughout the chest and surgery is not always possible. For such cases and in stage IV NSCLC, where the tumor has spread throughout the body, patients are treated with chemotherapy alone. The stage at which NSCLC is detected also determines how well patients respond to treatment. Those who can be treated surgically do much better than those who can't. So, whereas only 2% of patients with stage IV lung cancer survive for 5 years after diagnosis, about 70% of patients with stage I or II lung cancer live at least this long.

Why Was This Study Done? Even stage I and II lung cancers often recur and there is no accurate way to identify the patients in which this will happen. If there was, these patients could be given aggressive chemotherapy, so the search is on for a "molecular signature" to help identify which NSCLCs are likely to recur. Unlike normal cells, cancer cells divide uncontrollably and can move around the body. These behavioral differences are caused by changes in their genetic material that alter their patterns of RNA transcription and protein expression. In this study, the researchers have investigated whether data from several microarray studies (a technique used to catalog the genes expressed in cells) can be pooled to construct a gene expression signature that predicts the survival of patients with stage I NSCLC.

What Did the Researchers Do and Find? The researchers took the data from seven independent microarray studies (including a new study of their own) that recorded gene expression profiles related to survival time (less than 2 years and greater than 5 years) for stage I NSCLC. Because these studies had been done in different places with slightly different techniques, the researchers applied a statistical tool called distance-weighted discrimination to smooth out any systematic differences among the studies before identifying 64 genes whose expression was associated with survival. Most of these genes are involved in cell adhesion, cell motility, cell proliferation, and cell death, all processes that are altered in cancer cells. The researchers then developed a statistical model that allowed them to use the gene expression and survival data to calculate risk scores for nearly 200 patients in five of the datasets. When they separated the patients into high and low risk groups on the basis of these scores, the two groups were significantly different in terms of survival time. Indeed, the gene expression signature was better at predicting outcome than routine staging. Finally, the researchers validated the gene expression signature by showing that it predicted survival with more than 85% accuracy in two independent datasets.

What Do These Findings Mean? The 64 gene expression signature identified here could help clinicians prepare treatment plans for patients with stage I NSCLC. Because it accurately predicts survival in patients with adenocarcinoma or squamous cell cancer (the two major subtypes of NSCLC), it potentially indicates which of these patients should receive aggressive chemotherapy and which can be spared this unpleasant treatment. Previous attempts to establish gene expression signatures to predict outcome have used data from small groups of patients and have failed when tested in additional patients. In contrast, this new signature seems to be generalizable. Nevertheless, its ability to predict outcomes must be confirmed in further studies before it is routinely adopted by oncologists for treatment planning.

Additional Information. Please access these Web sites via the online version of this summary at http://dx.doi.org/10.1371/journal.pmed.0030467.

- US National Cancer Institute information on lung cancer for patients and health professionals.
- MedlinePlus encyclopedia entries on small-cell and non-small-cell lung cancer.
- Cancer Research UK, information on patients about all aspects of lung cancer.
- Wikipedia pages on DNA microarrays and expression profiling (note that Wikipedia is a free online encyclopedia that anyone can edit).