

Supplementary Materials

Table S1. A 15-gene lung cancer prognostic signature..	2
Table S2. A 16-gene signature sharing common biological functions between 12- and 15-gene signatures (Table S5)..	3
Table S3. Multivariate Cox proportional analysis of 15- and 16-gene risk score with major clinical covariates in lung cancer survival on testing cohorts (DFCI and MSK)..	4
Table S4. Multivariate Cox proportional analysis of 15- and 16-gene risk score with all clinical covariates in lung cancer survival on testing cohorts (DFCI and MSK)..	5
Table S5. Comparison of biological functions between 12-gene signature and 15-gene signature with curated database.....	7
Table S6. 14 published lung cancer gene signatures evaluated in GSEA.	9
Table S7. Summary of gene selection and classification methods of molecular classifiers compared in Fig. 5.....	10
Table S8. Machine learning algorithm and genes used in chemoresponse prediction using 12-gene signature.....	11
Table S9. Sensitivity and specificity of the 12-, 15- and 16-gene prognostic models.	13
Figure S1. Gene set enrichment analysis of the 12-gene signature along with 14 published gene signatures for NSCLC.....	14
Figure S2. Evaluation of the 15-gene, 12-gene, and 16-gene prognostic models with molecular prognostic models presented by Shedden et al (2008)..	15
Figure S3. Comparison of gene expression patterns of the 15-gene signature measured with DNA microarray and RT-PCR microfluidic low density arrays (LDA).....	16

Table S1. A 15-gene lung cancer prognostic signature. This gene signature was identified using pooled-variance *t*-tests and RELIEF algorithm. The expression of the 15 genes were used as covariates in Cox model and median risk score from training set was used as the cutoff point.

Probe Set ID	Gene	Functions	Classification
204854_at	GPR162 /// LEPREL2	Collagen biosynthesis, folding, and assembly	Metabolism
206150_at	CD27	B-cell activation and immunoglobulin synthesis; signaling transduction	Oncogene
205171_at	PTPN4	Cell growth, differentiation, mitotic cycle, and oncogenic transformation	Oncogene
201107_s_at	THBS1	Cell-to-cell and cell-to-matrix interactions.	Oncogene
210762_s_at	DLC1	A candidate tumor suppressor gene	Oncogene
218340_s_at	UBA6	Ubiquitin-activating protein	Protein Degradation
211327_x_at	HFE	Iron absorption	Signaling Transduction
208772_at	ANKHD1	Unknown	Structure
211603_s_at	ETV4	Cellular movement	Transcription
207296_at	ZNF343	Unknown	Transcription
214717_at	DKFZp434H1419	Unknown	N/A
213779_at	EMID1	Unknown	N/A
215598_at	TTC12	Binding	N/A
201581_at	TXNDC13	Cell redox homeostasis, electron transport chain	N/A
205308_at	FAM164A	Unknown	N/A

Table S2. A 16-gene signature sharing common biological functions between 12- and 15-gene signatures (Supplementary Table 5). Cox model was fitted with these 16 gene expression levels and 75th percentile of the risk scores from training set was used as the cutoff.

Probe Set ID	Gene	Functions	Classification
206150_at	CD27	B-cell activation and immunoglobulin synthesis; signaling transduction	Oncogene
205171_at	PTPN4	Cell growth, differentiation, mitotic cycle, and oncogenic transformation	Oncogene
201107_s_at	THBS1	Cell-to-cell and cell-to-matrix interactions.	Oncogene
211327_x_at	HFE	Iron absorption	Signaling Transduction
211603_s_at	ETV4	Cellular movement	Transcription
201581_at	TXNDC13	Cell redox homeostasis, electron transport chain	N/A
212041_at	ATP6V0D1	Atpase	Metabolism
222078_at	PKLR	Pyruvate kinase	Metabolism
219808_at	SCLY	Catalyzes the decomposition of L-selenocysteine to L-alanine and elemental selenium	Metabolism
209420_s_at	SMPD1	Converts sphingomyelin to ceramide	Metabolism
210762_s_at	DLC1	A candidate tumor suppressor gene	Oncogene
204524_at	PDPK1	Cell signal protein	Oncogene
218833_at	ZAK	Cell signal protein	Oncogene
208855_s_at	STK24	Protein kinase	Signaling Transduction
208775_at	XPO1	Nuclear protein transport	Signaling Transduction
46142_at	LMF1	Maturation of specific proteins in the endoplasmic reticulum	Structure

Table S3. Multivariate Cox proportional analysis of 15- and 16-gene risk score with major clinical covariates in lung cancer survival on testing cohorts (DFCI and MSK).

Variable*	P value	Hazard Ratio (95% CI)^ψ	
<i>Analysis without risk score</i>			
Gender (Male)	0.22	1.34	(0.84-2.16)
Age at diagnosis (>60)	0.08	1.61	(0.95-2.74)
Tumor Stage			
Stage II	6.25E-05	2.91	(1.72-4.91)
Stage III	1.09E-05	4.16	(2.20-7.85)
<i>Analysis with 15-gene risk score</i>			
Gender (Male)	0.20	1.36	(0.85-2.18)
Age at diagnosis (>60)	0.08	1.60	(0.94-2.74)
Tumor Stage			
Stage II	1.32E-04	2.80	(1.65-4.74)
Stage III	4.82E-05	3.73	(1.98-7.05)
15-gene risk score	2.84E-04	1.99	(1.37-2.89)
<i>Analysis with 16-gene risk score</i>			
Gender (Male)	0.11	1.49	(0.92-2.41)
Age at diagnosis (>60)	0.18	1.44	(0.84-2.48)
Tumor Stage			
Stage II	5.36E-05	2.97	(1.75-5.03)
Stage III	7.52E-07	5.19	(2.70-9.96)
16-gene risk score	6.24E-07	2.50	(1.33-3.59)

* Gender was binary variable (0 for female and 1 for male); age at diagnosis was a binary variable (0 for < 60 years old and 1 otherwise); tumor stage was categorical variable of 3 categories (Stage I [as the reference group], Stage II, and Stage III). Risk score was continuous variable, and the hazard ratio represents the relative risk between the mean risk scores of high- and low-risk groups.

^ψ denotes confidence interval.

Table S4. Multivariate Cox proportional analysis of 15- and 16-gene risk score with all clinical covariates in lung cancer survival on testing cohorts (DFCI and MSK).

Variable*	P value	Hazard Ratio (95% CI)^ψ	
<i>Analysis without risk score</i>			
Gender (Male)	0.43	1.22	(0.74-1.99)
Age at diagnosis (>60)	0.05	1.70	(0.99-2.92)
Race			
Others/Unknown	0.28	0.43	(0.09-1.97)
White	0.10	0.28	(0.06-1.28)
Tumor Grade			
Moderately differentiated	0.14	0.53	(0.23-1.24)
Poorly differentiated	0.70	1.17	(0.53-2.61)
Smoking History			
Smokers	0.62	0.84	(0.43-1.66)
Unknown	0.91	0.89	(0.11-7.10)
Tumor Stage			
Stage II	3.31E-04	2.72	(1.57-4.69)
Stage III	2.38E-05	4.93	(2.35-10.33)
<i>Analysis with 15-gene risk scores</i>			
Gender (Male)	0.36	1.26	(0.77-2.06)
Age at diagnosis (>60)	0.04	1.75	(1.02-3.01)
Race			
Others/ Unknown	0.38	0.50	(0.11-2.31)
White	0.14	0.32	(0.07-1.45)
Tumor differentiation			
Moderately differentiated	0.16	0.55	(0.24-1.27)
Poorly differentiated	0.99	0.99	(0.44-2.23)
Smoking History			
Smokers	0.93	0.97	(0.49-1.91)
Unknown	0.85	1.22	(0.15-9.89)
Tumor Stage			
Stage II	2.61E-04	2.76	(1.60-4.77)
Stage III	5.19E-05	4.66	(2.21-9.82)
15-gene risk score	2.47E-03	1.81	(1.23-2.65)
<i>Analysis with 16-gene risk score</i>			
Gender (Male)	0.17	1.42	(0.86-2.35)
Age at diagnosis (>60)	0.09	1.63	(0.93-2.85)
Race			
Others/ Unknown	0.22	0.38	(0.08-1.77)
White	0.05	0.22	(0.05-1.00)
Tumor differentiation			
Moderately differentiated	0.16	0.55	(0.23-1.28)
Poorly differentiated	0.96	1.02	(0.45-2.30)
Smoking History			

Smokers	0.53	0.81	(0.41-1.59)
Unknown	0.96	0.94	(0.12-7.54)
Tumor Stage			
Stage II	2.37E-04	2.79	(1.62-4.83)
Stage III	2.09E-06	6.34	(2.96-13.58)
16-gene risk score	7.49E-07	2.45	(1.72-3.50)

* Gender was binary variable (0 for female and 1 for male); age ge at diagnosis was a binary variable (0 for < 60 years old and 1 otherwise); race was a categorical variable of 3 categories (African American [as the reference group], White, and Others [composed of Asian (5) , Hawaiian or Pacific Islander (1), and unknown]); tumor grade was categorical variable of 3 categories (Well [as the reference group], Moderately, and Poorly differentiate); Smoking history was a categorical variable of 3 categories (Non-smokers, Smokers, and Unknown); tumor stage was categorical variable of 3 categories (Stage I [as the reference group], Stage II, and Stage III). Risk score was continuous variable, and the hazard ratio represents the relative risk between the mean risk scores of high- and low-risk groups.

^ψ denotes confidence interval.

Table S5. Comparison of biological functions between 12-gene signature and 15-gene signature with curated database. The biological functions were obtained using Ingenuity Pathway Analysis (IPA).

Category	Category	12-gene	15-gene	Common
Diseases and Disorders	Cancer			✓
	Cardiovascular Disease		✓	
	Connective Tissue Disorders		✓	
	Dermatological Diseases and Conditions		✓	
	Genetic Disorder			✓
	Hematological Disease			✓
	Hepatic System Disease			✓
	Immunological Disease			✓
	Infection Mechanism	✓		
	Inflammatory Disease		✓	
	Inflammatory Response		✓	
	Metabolic Disease			✓
	Neurological Disease			✓
	Reproductive System Disease			✓
	Respiratory Disease			✓
Skeletal and Muscular Disorders		✓		
Molecular and Cellular Functions	Amino Acid Metabolism			✓
	Antigen Presentation		✓	
	Carbohydrate Metabolism		✓	
	Cell Cycle		✓	
	Cell Death			✓
	Cell Morphology		✓	
	Cell Signaling			✓
	Cell-To-Cell Signaling and Interaction		✓	
	Cellular Assembly and Organization			✓
	Cellular Compromise		✓	
	Cellular Development			✓
	Cellular Function and Maintenance			✓
	Cellular Growth and Proliferation			✓
	Cellular Movement			✓
	DNA Replication, Recombination, and Repair	✓		
	Drug Metabolism		✓	
	Gene Expression	✓		
	Lipid Metabolism			✓
	Molecular Transport			✓
	Nucleic Acid Metabolism		✓	
	Post-Translational Modification			✓
	Protein Synthesis		✓	
	Protein Trafficking		✓	
RNA Trafficking	✓			

	Small Molecule Biochemistry			✓
Physiological System Development and Function	Cardiovascular System Development and Function		✓	
	Cell-mediated Immune Response		✓	
	Hematological System Development and Function		✓	
	Immune Cell Trafficking		✓	
	Nervous System Development and Function		✓	
	Organ Development		✓	
	Skeletal and Muscular System Development and Function			✓
	Tissue Development		✓	
	Tumor Morphology		✓	
	Visual System Development and Function		✓	

Table S6. 14 published lung cancer gene signatures evaluated in GSEA.

Signature Name (GSEA)	First Author	Publication PubMed ID	No. of Signature Genes/Probes	No. of Genes matched in GSEA (By gene symbol)
Beer_50g	Beer, DG	PMID:12118244	50	45
Bhattacharjee_150g	Bhattacharjee, A	PMID:11707567	150	130
Boutros_6g	Boutros, PC	PMID:19196983	6	6
Chen_5g	Chen, HY	PMID:17202451	5	5
Guo_35g	Guo, L	PMID:16740756	35	34
Lau_3g	Lau, SK	PMID:18065728	3	3
Lu_64g	Lu, Y	PMID:17194181	64	62
Potti_133g	Potti, A	PMID:16899777	133	129
Raponi_50g	Raponi, M	PMID:16885343	50	44
Shedden_MA	Shedden, K	PMID:18641660	13830	8319
Shedden_MB	Shedden, K	PMID:18641660	52	50
Shedden_MC	Shedden, K	PMID:18641660	26	23
Shedden_MD	Shedden, K	PMID:18641660	42	34
Shedden_MH	Shedden, K	PMID:18641660	313	244

Table S7. Summary of gene selection and classification methods of molecular classifiers compared in Fig. 5. Gene signatures A-N were reported in (Shedden et al, 2008).

Molecular Classifier*	Number of signature genes	Gene selection method(s)	Classification method(s)
Shedden A	~ 9591 Genes	Clustering analysis	Ridged Cox proportional hazard model
Shedden C	23 Genes	SAM, Maximizing Chi-Square analysis (MCA, univariate Cox model and k-mean clustering)	Binary Tree-Structured Vector Quantization (BTSVQ)
Shedden D	37 Genes	SAM, Maximizing Chi-Square analysis (MCA, univariate Cox model and k-mean clustering)	Binary Tree-Structured Vector Quantization (BTSVQ)
Shedden E	1 Gene	Gene Expression Fold Change	Post-hoc split of expression of one gene
Shedden F	42 Genes	Univariate Cox Model	Principal Components and Cox Model
Shedden G	38 Genes	Univariate Cox Model	Principal Components and Cox Model
Shedden H	252 Genes	Scoring and filtering on set of mitosis genes	Majority vote
Shedden J	5 Genes	Univariate Cox model (Chen et al, NEJM 07)	Ridged Cox proportional hazard model
Shedden K	16 Genes	Univariate Cox model (Chen et al, NEJM 07)	Ridged Cox proportional hazard model
Shedden L	9 Genes (from 80 Genes)	Principal Components (Potti et al, NEJM 06)	Ridged Cox proportional hazard model
Shedden M	45 Genes (from 80 Genes)	Principal Components (Potti et al, NEJM 06)	Ridged Cox proportional hazard model
Shedden N	80 Genes	Principal Components (Potti et al, NEJM 06)	Ridged Cox proportional hazard model
12-gene	12 Genes	t-test, SAM, RELIEFF	Naïve Bayes

*Gene signatures A-H were identified in (Shedden et al, 2008). Gene signatures J and K were identified in (Chen et al, 2007). Gene signatures L, M, and N were identified in (Potti et al, 2006).

Table S8. Machine learning algorithm and genes used in chemoresponse prediction using 12-gene signature.

Anti-cancer Agent	Machine learning algorithm	Genes Selected	Resistant lung cancer cell lines	Sensitive lung cancer cell lines
Carboplatin	RBF Network (seed = 2)	ATP6V0D1 CCDC99 FAM164A LMF1 PDPK1 PKLR SCLY SMPD1 STK24 XPO1	LC:EKVX LC:NCI_H322M	LC:NCI_H460 LC:NCI_H522 (LC:NCI_H23 not included due to missing values)
Paclitaxel	IBK (k=3)	CCDC99 DLC1 LMF1 PKLR SMPD1 XPO1 ZAK	LC:HOP_92 LC_EKVX	LC:NCI_H460 LC:NCI_H522
Cisplatin	Decorate (PART as base learner)	ATP6V0D1 CCDC99 FAM164A LMF1	LC:NCI_H226 LC:EKVX LC:NCI_H322M	LC:HOP_62 LC:NCI_H460 (LC:NCI_H23 not included due to missing values)
Etoposide	AdaBoostM1 (seed = 2, Random Tree as base learner)	CCDC99 LMF1 SCLY STK24 XPO1	LC:EKVX LC:NCI_H322M	LC:HOP_62 LC:NIC_H460
Erlotinib	RBF Network	DLC1 LMF1 XPO1 SMPD1 STK24 PDPK1 ZAK PKLR CCDC99	LC:NCI_H226 (LC:NCI_H23 not included due to missing values)	LC:EKVX LC:NCI_H322M LC:NCI_H522

Gefitinib	Multilayer	ATP6V0D1	LC:A549	LC:EKVX
	Perceptron (seed=2, learning rate=0.4)	SMPD1	LC:HOP_62	LC:NCI_H322M
		XPO1	LC:HOP_92	
		PKLR	LC:NCI_H226	
		STK24	(LC:NCI_H23 not	
		SCLY	included due to	
		missing values)		

Table S9. Sensitivity and specificity of the 12-, 15- and 16-gene prognostic models.

	Sensitivity (% of correctly predicted high-risk patients)				Specificity (% of correctly predicted low-risk patients)			
	<i>n</i>	12-gene	15-gene	16-gene	<i>n</i>	12-gene	15-gene	16-gene
<i>3-year survival as the cutoff (high-risk: death within 3-y; low-risk: alive after 3-y)</i>								
UM & HLM	95	73.65	76.84	47.37	152	59.21	64.47	87.50
MSK	23	86.96	82.61	60.87	71	57.75	50.70	70.42
DFCI	22	68.18	86.36	54.55	55	76.36	47.27	81.82
<i>5-year survival as the cutoff (high-risk: death within 5-y; low-risk: alive after 5-y)</i>								
UM & HLM	125	72.80	72.80	44.80	104	66.35	69.23	93.27
MSK	34	70.59	67.65	50.00	31	48.39	41.94	67.74
DFCI	28	64.29	78.57	50.00	36	77.78	47.22	86.11
<i>2.5-year and 5-year survival as the high- and low-risk cutoffs (high-risk: death within 2.5-y; low-risk: alive after 5-y)</i>								
UM & HLM	84	75.00	77.38	48.81	104	66.35	69.23	93.27
MSK	21	95.24	85.71	66.67	31	48.39	41.94	67.74
DFCI	20	70.00	85.00	55.00	36	77.78	47.22	86.11

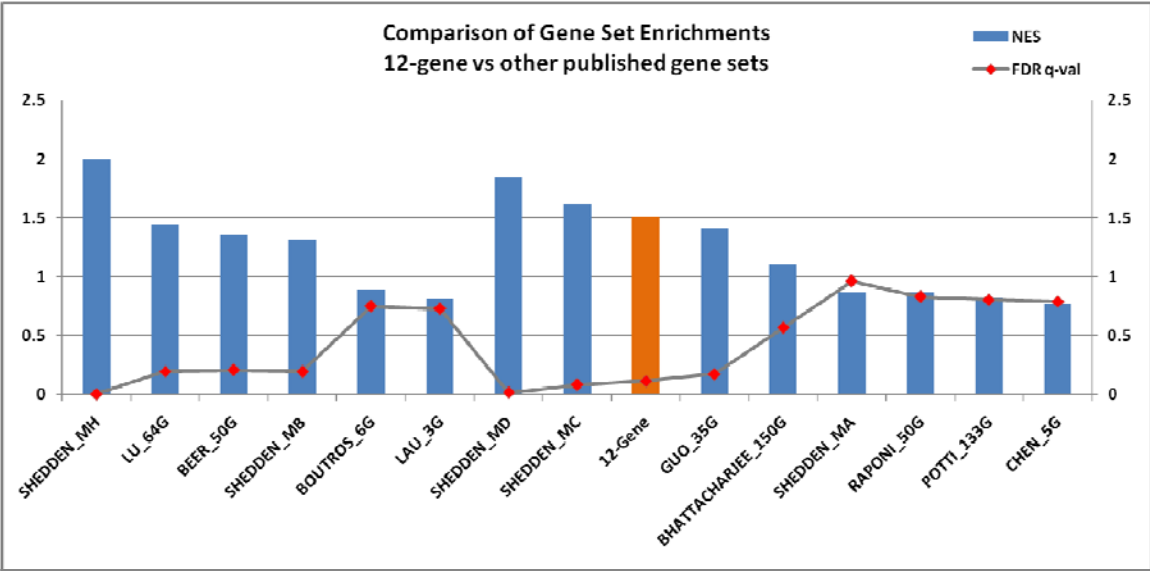


Figure S1. Gene set enrichment analysis of the 12-gene signature along with 14 published gene signatures for NSCLC. A summary of the 14 gene signatures analyzed is listed in Supplementary Table 6.

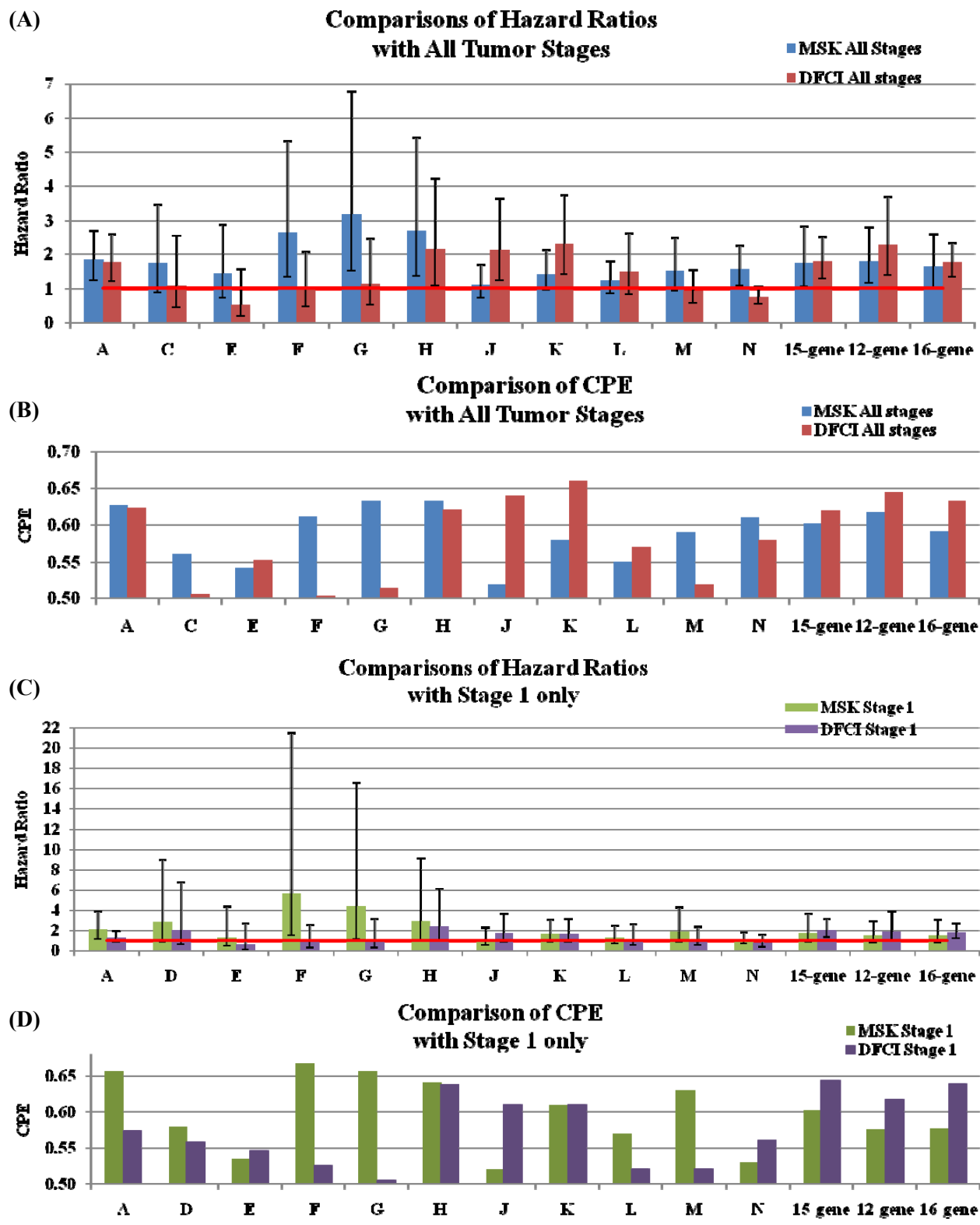


Figure S2. Evaluation of the 15-gene, 12-gene, and 16-gene prognostic models with molecular prognostic models presented by Shedden et al (2008). Hazard ratio (A, C) and concordance probability estimate (CPE) (B, D) were compared on patients in all stages (A, B) and stage I (C, D) of lung cancer. Error bars in (A) and (C) represent 95% confidence interval of hazard ratio.

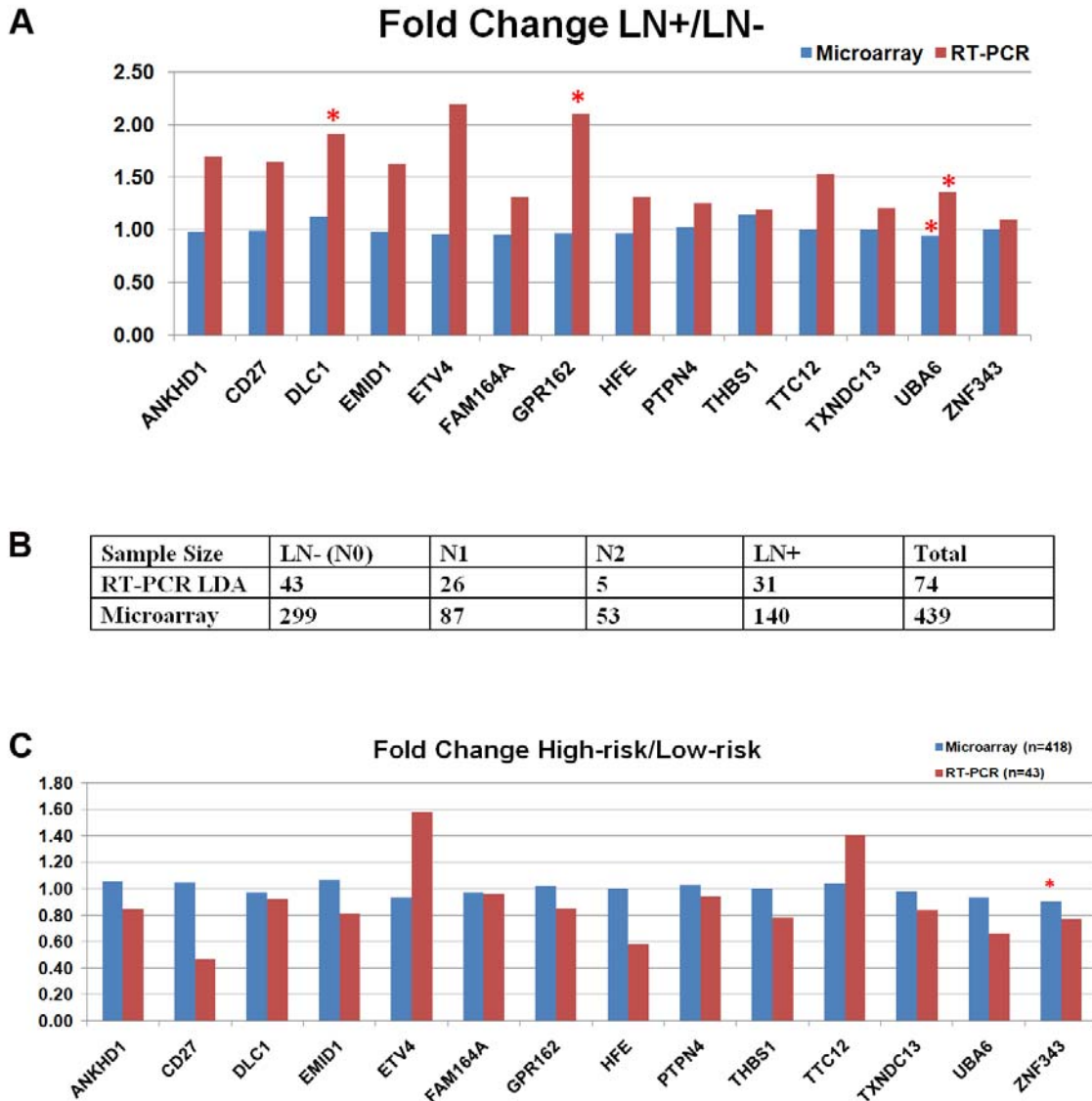


Figure S3. Comparison of gene expression patterns of the 15-gene signature measured with DNA microarray and RT-PCR microfluidic low density arrays (LDA). Gene expression fold change in lymph node positive (LN+) patients vs. lymph node negative (LN-) patients was compared (A). Samples included in the fold change comparison are summarized in (B). On patient with follow-up information, gene expression fold change in high-risk patients vs. low-risk patients at 3-year period after surgery was also compared (C). The RT-PCR data were normalized with POLR2A in a sample-wise manner. DNA microarray data were obtained from Shedden et al (2008). Red asterisk (*) above the bar indicates the gene was differentially expressed *t*-test ($P < 0.05$).