Supporting Information

Hsu et al. 10.1073/pnas.0900827106

SI Methods

Gene Expression Analysis. Statistical modeling representing activation of the individual transcription factors (*NKX2–8*, *PAX9*, and *TTF-1*) was performed using metagene construction and binary prediction analysis, as described previously (1–3). Specifically, the gene expression data of the individual transfectants were used to serve as a training set to create a signature that distinguished the transfectant group from the control/normal group in a supervised analysis using Bayesian regression methodologies to develop a probit model predictive of transcription factor expression. This method generates a model in which a restricted set of differentially expressed genes could distinguish the comparison groups and a signature is obtained that summarizes the constituent of genes as a single expression profile and is described as the top principal components of that set of genes.

Gene expression signatures that reflect the activity of a given transcription factor (*TTF-1*, *NKX2–8*, or *PAX9*) were then applied to a clinically annotated data set of 91 tumor samples (GSE3141) to predict patterns of transcription factor activation. When predicting the activation of transcription factors across the NSCLC samples, gene selection and identification based on the training data set were used to compute metagene values using the principal components of the training data and tumor expression data. The metagene is defined as the dominant singular factor (principal component) that represents the dominant average pattern of expression for a cluster of genes. Using the training set of expression vectors (of values across metagenes) described above to represent 2 biological states, a binary probit regression model of predictive probabilities for each of the 2 states (transcription factor activation or not) for each case was estimated using Bayesian methods (4). To prevent overfitting of the model, a leave-one-out cross-validation analysis was performed to test the stability and predictive capability of the model. Predictions of transcription factor pathway activation within the tumor samples were evaluated using methods previously described (5) to produce estimated relative probabilities and associated measures of uncertainty of transcription factor activation across the validation set. A priori to survival analyses, an estimated probability of 0.5 was classified as high probability of transcription factor activation and a probability of

Gene Expression Levels of *NKX2–8, TTF-1,* and *PAX9*. Gene expression levels of *TTF-1* (Probe IDs 207771_at, 207772_s_at), *NKX2–8* (Probe ID 207451_at), and *PAX9* (207059_at) were determined using the annotated Probe ID as specified (Affymetrix) and determined on a clinically annotated data set of 91 tumor samples (GSE3141). The mean values of the expression levels for *NKX2–8*, *TTF-1*, and *PAX9* were determined and samples with a gene expression level above the mean were identified as having high expression of a transcription factor and samples with a gene expression level below the mean were identified as having low expression of a transcription factor.

Cross-Platform Analysis. To map the probe sets across different generations of Affymetrix gene chip array, we used an in-house program, Chip Comparer (http://tenero.duhs.duke.edu/genearray/perl/chip/chipcomparer.pl). First, each probe set ID in the specified Affymetrix gene chips was mapped to the corresponding LocusLink (EntrezGene) ID by parsing local copies of the LocusLink and UniGene databases to identify the GenBank accession numbers and LocusLink IDs associated with that probe set. Second, probe sets from different gene chips sharing the same LocusLink ID were matched, as described previously (1).

Univariate and Multivariate Analyses. In an effort to fully understand the prognostic significance of clusters representing coactivation of the transcription factors on survival in patients with early stage NSCLC, univariate and multivariate analyses were performed with the use of the Cox proportional hazards model. Multivariate models include continuous covariates for age, gender, and tumor size and dichotomous covariates for lymph node status and histologic subtype. Hazard ratios and 95% confidence intervals are reported with respect to the hierarchical cluster with lowest level of survival. *P*-values are based on likelihood-ratio tests, and analyses were performed using the statistical package R (6).

Lung Cancer Cell Lines and Drug Sensitivity Assays. The NSCLC cell lines (H522, H1703, H23, H1568, H661, H2073, H2085, H838, H520, H1650, H2030, H226, H1573, H1437, A549, H1563, H1395, H1651, H1944, H460, H1666, H1838, H2170, H2126, H2405, H1793, H358, H1373, H2291, H322M, HCC2935, H596, H441, H2122, H1975, H647, H2228, and HCC4006) were grown as recommended by the supplier (ATCC). These cell lines were used in drug sensitivity assays to examine the association between patterns of coactivation of transcription factors and sensitivity to cisplatin, using cell proliferation experiments as described previously (7). Briefly, optimal cell number and linear range of drug concentration were first determined for each cell line and drug. Cells were plated in drug-free media at a concentration of 3,000–7,000 cells/well in tissue-culture-treated 96-well plates. Five replicate wells were plated for each planned drug concentration. Control wells were additionally plated containing cells in growth media without drug and wells with growth media without cells. Plates were incubated for 24 h at 37 °C. After 24 h, each cell line was exposed to a series of increasing cisplatin (Duke University pharmacy storeroom) concentrations and cell cytotoxicity was assessed with propidium iodide (Sigma-Aldrich) staining at days 0 and 5 (FLUOstar Optima, BMG Labtech). EC_{50} (GraphPad Prism, GraphPad Software) for cisplatin was defined for each cell line in 2–5 independent replicate experiments.

1. Bild AH, et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439:353–357.
2. Potti A, et al. (2006) A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N Engl J Med* 355:570–580.
3. Hsu DS, et al. (2007) Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer. *J Clin Oncol* 25:4350–4357.
4. Team RDC (2006) R: A language and environment for statistical computing. Available at R Foundation for Statistical Computing, http://www.R-project.org.
5. Riedel RF, et al. (2008) A genomic guided approach to identify molecular pathways associated with chemotherapy resistance. *Mol Cancer Ther* 7:3141–3149.
6. Team RDC (2006) R: A language and environment for statistical computing. Available at R Foundation for Statistical Computing, http://www.R-project.org.
7. Riedel RF, et al. (2008) A genomic guided approach to identify molecular pathways associated with chemotherapy resistance. *Mol Cancer Ther* 7:3141–3149.

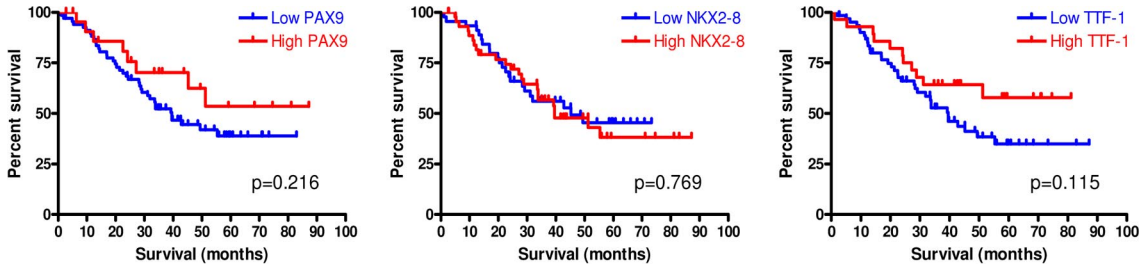

Fig. S1. Kaplan–Meier survival analysis of patients with lung cancer differentiated by gene expression level of individual transcription factors (*TTF-1*, *NKX2–8*, or *PAX9*) shows no significant difference in survival.

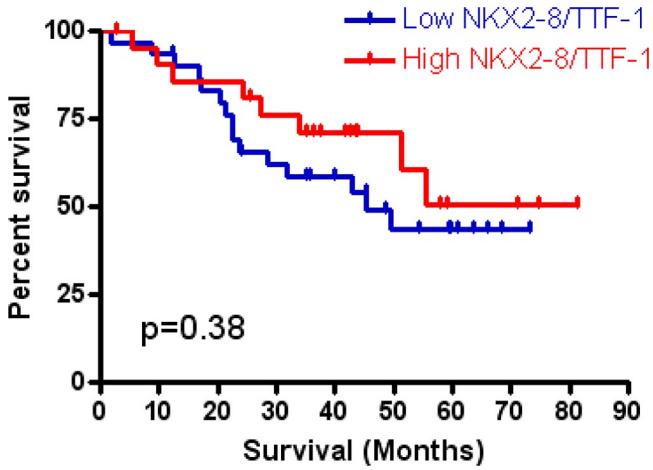

Fig. S2. Kaplan–Meier survival analysis of patients with lung cancer differentiated by combinatory gene expression level of transcription factors (high or low *NKX2–8/TTF-1*) shows no significant difference in survival.

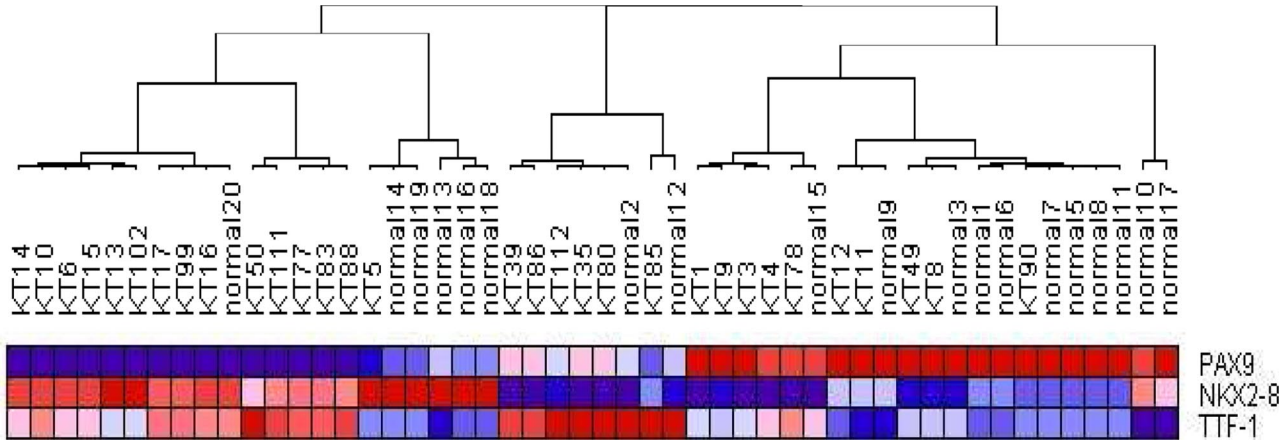

Fig. S3. Prediction of *TTF-1*, *NKX2–8*, and *PAX9* pathway status in a mouse lung cancer model. A set of previously published mouse Affymetrix expression data from normal and lung tumor tissue with spontaneous activating KRAS mutations was used to validate the oncogenic relevance of *TTF-1*/*NKX2–8* coactivation. The predicted probabilities of pathway activity in the normal tissue and tumors are shown (red, high level of activation; blue, low level of activation).

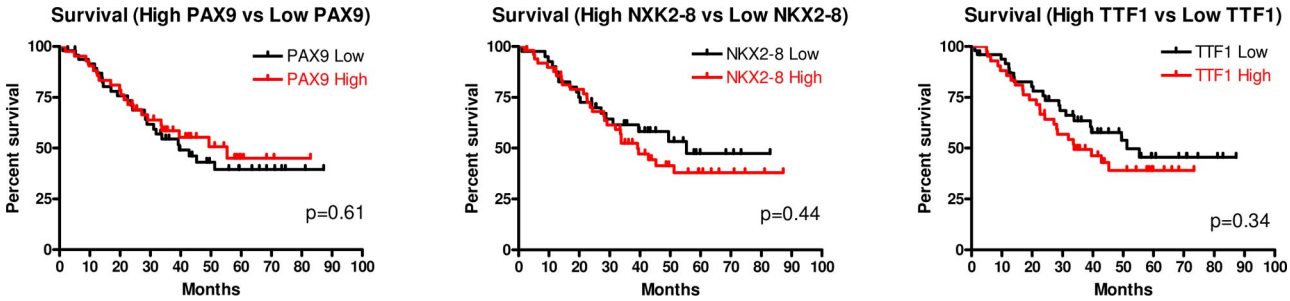

Fig. S4. Kaplan–Meier survival analysis of patients with lung cancer differentiated by activation of individual transcription factors (*TTF-1*, *NKX2–8*, or *PAX9*) shows no significant difference in survival.

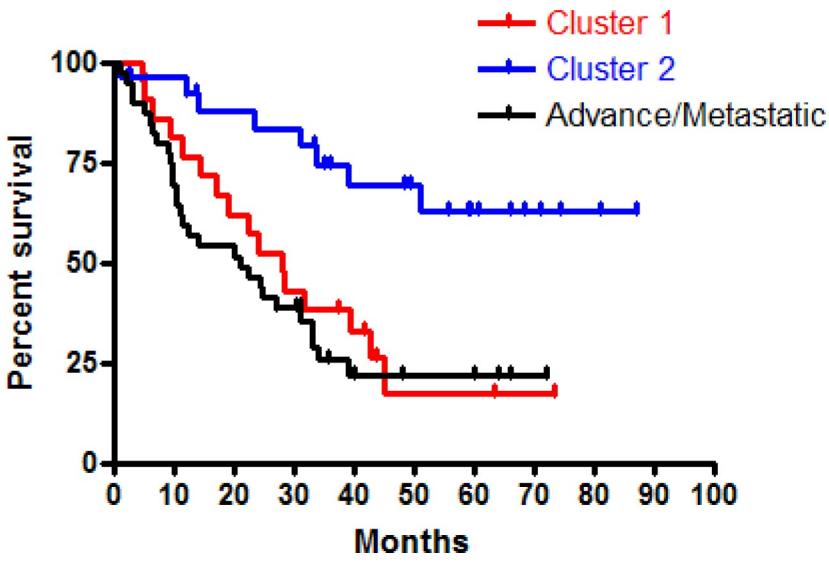

Fig. S5. Kaplan–Meier analysis demonstrating that the survival pattern of the poor prognosis cluster (*NKX2–8* and *TTF-1* coactivated) of patients with early NSCLC mimics the survival pattern of patients with stage IV/advanced disease.

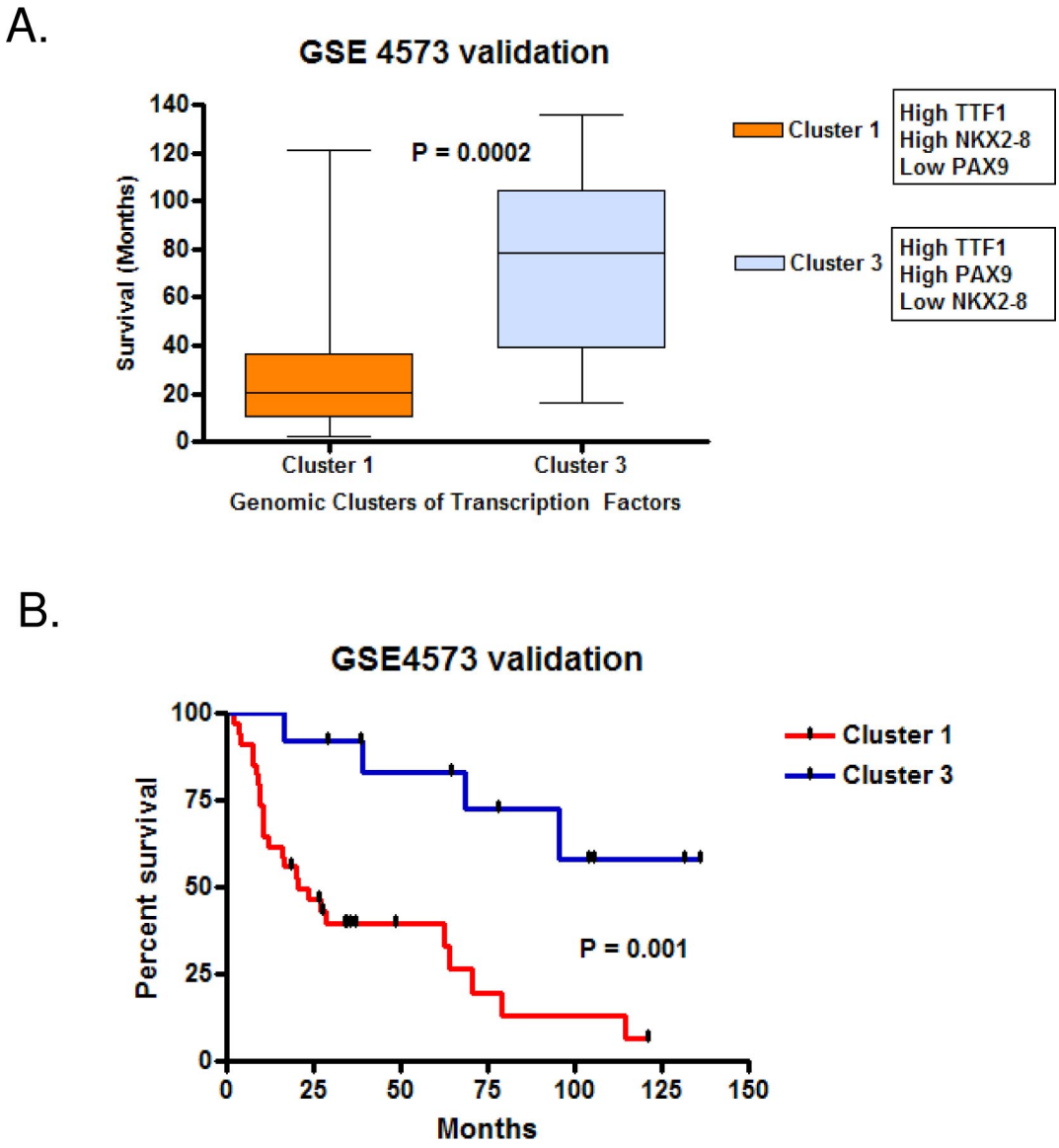

Fig. S6. (*A*) Box and whisker plots of the probability of survival plotted against the probabilities of samples having high activation of *TTF-1* and *NKX2–8* (cluster 1) and samples having high activation of *TTF-1* and *PAX9* (cluster 3) on the GSE4573 data set. (*B*) Kaplan–Meier survival analysis on the GSE4573 data set shows that cluster 1 has a poorer prognosis than cluster 3

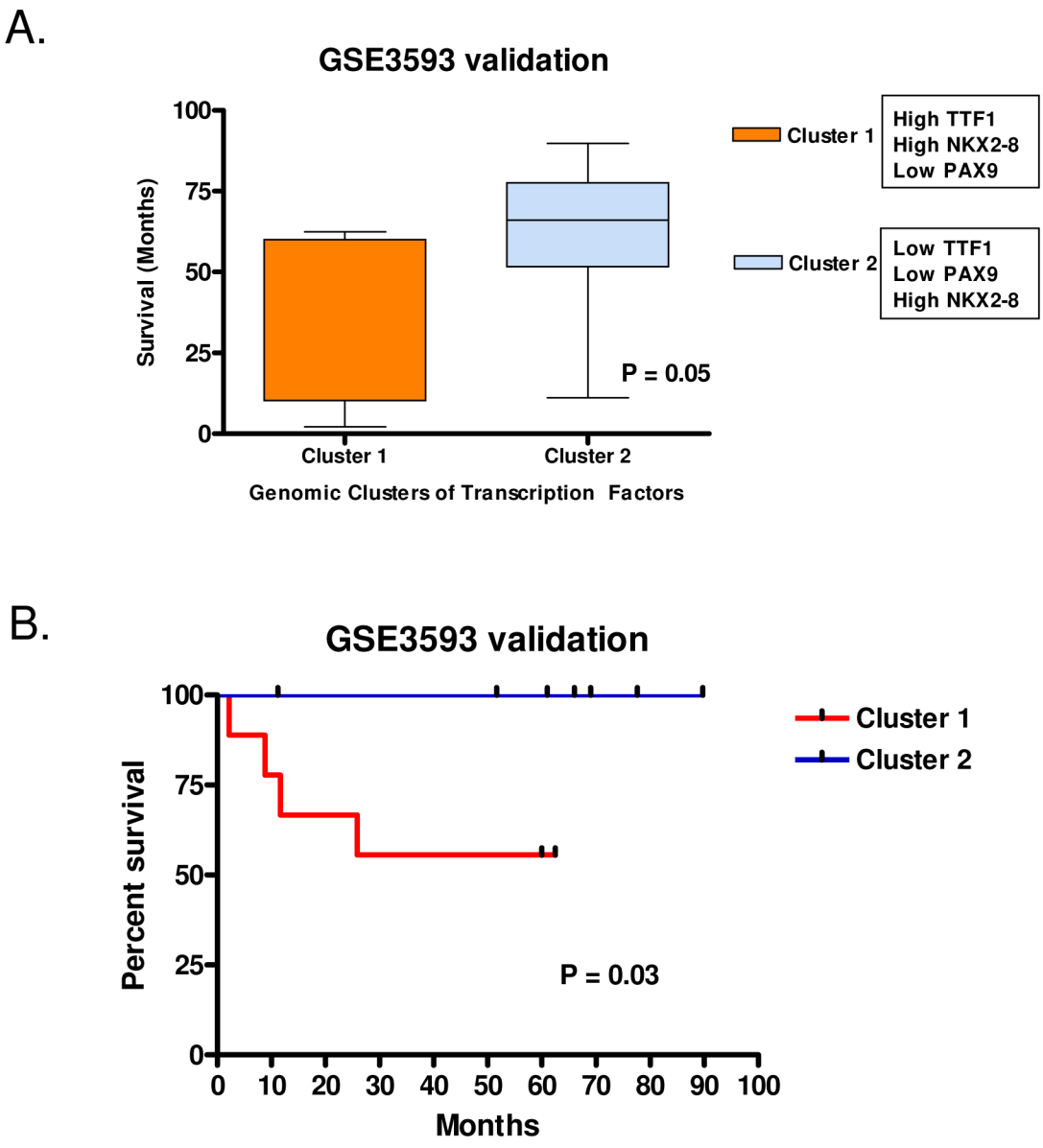

Fig. S7. (*A*) Box and whisker plots of the probability of survival plotted against the probabilities of samples having high activation of *TTF-1* and *NKX2–8* (cluster 1) and samples having high activation of *NKX2–8* alone (cluster 2) on the GSE3593 data set. (*B*) Kaplan–Meier survival analysis on the GSE3593 data set shows that cluster 1 has a poorer prognosis than cluster 2.

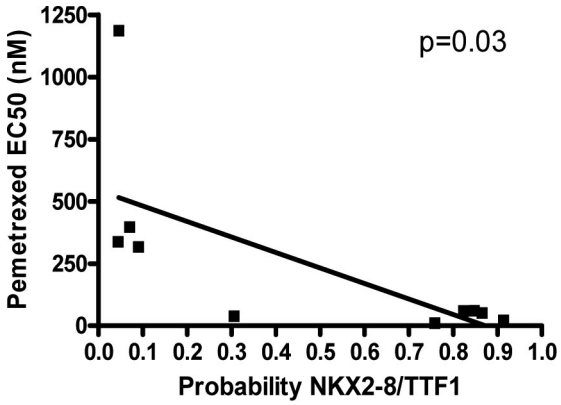

Fig. S8. Linear regression analysis demonstrates a significant relationship between the EC_{50} of pemetrexed and the predicted probability of transcription factor coactivation of *NKX2–8* and *TTF-1*.

Table S1. Gene annotation (GATHER) of the genes constituting the *NKX2–8*, *PAX9*, and *TTF-1* signatures

No.	KEGG pathway annotation	*P*-value	FDR
	NKX2–8 signature		
1	path:hsa04010: MAPK signaling pathway	0.04	0.04
2	path:hsa04810: regulation of actin cytoskeleton	0.06	0.05
3	path:hsa04020: calcium signaling pathway	0.08	0.05
4	path:mmu04510: focal adhesion	0.1	0.06
5	path:hsa04070: phosphatidylinositol signaling system	0.29	0.21
6	path:hsa04080: neuroactive ligand-receptor interaction	0.3	0.21
7	path:hsa04620: Toll-like receptor signaling pathway	0.33	0.21
8	path:hsa00380: tryptophan metabolism	0.34	0.21
9	path:hsa04210: apoptosis	0.34	0.21
10	path:mmu00230: purine metabolism	0.38	0.22
11	path:hsa04350: TGF-β signaling pathway	0.4	0.22
12	path:hsa00562: inositol phosphate metabolism	0.46	0.25
13	path:hsa04630: Jak-STAT signaling pathway	0.49	0.25
14	path:rno04080: neuroactive ligand–receptor interaction	0.49	0.25
	PAX9 signature		
1	path:hsa04020: calcium signaling pathway	0.04	0.04
2	path:hsa04630: Jak-STAT signaling pathway	0.08	0.07
3	path:hsa04310: Wnt signaling pathway	0.1	0.08
4	path:hsa00190: oxidative phosphorylation	0.14	0.09
5	path:hsa04080: neuroactive ligand–receptor interaction	0.17	0.11
6	path:mmu00190: oxidative phosphorylation	0.23	0.15
7	path:mmu00230: purine metabolism	0.25	0.15
8	path:hsa04110: cell cycle	0.24	0.15
9	path:hsa00380: tryptophan metabolism	0.26	0.15
10	path:hsa04540: gap junction	0.27	0.15
11	path:mmu04080: neuroactive ligand–receptor interaction	0.3	0.17
12	path:hsa04350: TGF-β signaling pathway	0.31	0.17
13	path:mmu04060: cytokine–cytokine receptor interaction	0.44	0.25
14	path:mmu04070: phosphatidylinositol signaling system	0.43	0.25
15	path:hsa04910: insulin signaling pathway	0.42	0.25
16	path:hsa04610: complement and coagulation cascades	0.42	0.25
	TTF-1 signature		
1	path:rno04510: focal adhesion	0.01	0.01
2	path:hsa00190: oxidative phosphorylation	0.03	0.02
3	path:rno04010: MAPK signaling pathway	0.08	0.07
4	path:mmu04620: Toll-like receptor signaling pathway	0.08	0.07
5	path:mmu04060: cytokine–cytokine receptor interaction	0.12	0.09
6	path:rno04310: Wnt signaling pathway	0.11	0.09
7	path:hsa04810: regulation of actin cytoskeleton	0.19	0.14
8	path:rno04210: apoptosis	0.18	0.14
9	path:rno04512: ECM–receptor interaction	0.19	0.14
10	path:hsa00071: fatty acid metabolism	0.19	0.14
11	path:hsa00010: glycolysis/gluconeogenesis	0.21	0.14
12	path:hsa00350: tyrosine metabolism	0.22	0.15
13	path:hsa00500: starch and sucrose metabolism	0.22	0.15
14	path:rno04910: insulin signaling pathway	0.26	0.18
15	path:hsa04110: cell cycle	0.3	0.18
16	path:mmu04340: Hedgehog signaling pathway	0.29	0.18
17	path:mmu00010: glycolysis/gluconeogenesis	0.3	0.18
18	path:hsa04340: Hedgehog signaling pathway	0.26	0.18
19	path:hsa00330: arginine and proline metabolism	0.29	0.18
20	path:mmu04020: calcium signaling pathway	0.34	0.19
21	path:hsa00280: valine, leucine, and isoleucine degradation	0.35	0.19
22	path:mmu04530: tight junction	0.45	0.22
23	path:rno04330: Notch signaling pathway	0.43	0.22
24	path:hsa04630: Jak-STAT signaling pathway	0.48	0.22
25	path:hsa04350: TGF-β signaling pathway	0.41	0.22
26	path:mmu00650: butanoate metabolism	0.46	0.22
27	path:hsa00620: pyruvate metabolism	0.41	0.22
28	path:rno00564: glycerophospholipid metabolism	0.5	0.23

FDR, false discovery rate.

Table S2. Common pathways exist between 3 independent NSCLC prognostic models

	NKX2–8 signature: KEGG pathway annotation	Lu *et al*. (2006)	Beer *et al*. (2002)	Potti *et al*. (2006)
1	path:hsa04010: MAPK signaling pathway	Present	Present	Present
2	path:hsa04810: regulation of actin cytoskeleton	Present	Present	Present
3	path:hsa04020: calcium signaling pathway			
4	path:mmu04510: focal adhesion	Present	Present	Present
5	path:hsa04070: phosphatidylinositol signaling system	Present		Present
6	path:hsa04080: neuroactive ligand–receptor interaction			
7	path:hsa04620: Toll-like receptor signaling pathway			
8	path:hsa00380: tryptophan metabolism			
9	path:hsa04210: apoptosis	Present	Present	Present
10	path:mmu00230: purine metabolism		Present	
11	path:hsa04350: TGF-β signaling pathway		Present	
12	path:hsa00562: inositol phosphate metabolism	Present		Present
13	path:hsa04630: Jak-STAT signaling pathway	Present	Present	Present
	TTF-1 signature: KEGG pathway annotation			
1	path:rno04510: focal adhesion	Present	Present	Present
2	path:hsa00190: oxidative phosphorylation		Present	Present
3	path:rno04010: MAPK signaling pathway	Present	Present	Present
4	path:mmu04620: Toll-like receptor signaling pathway			
5	path:mmu04060: cytokine–cytokine interaction	Present	Present	Present
6	path:rno04310: Wnt signaling pathway	Present	Present	
7	path:hsa04810: regulation of actin cytoskeleton	Present	Present	Present
8	path:rno04210: apoptosis	Present	Present	Present
9	path:rno04512: ECM–receptor interaction		Present	
10	path:hsa00071: fatty acid metabolism			
11	path:hsa00010: glycolysis/gluconeogenesis		Present	Present
12	path:hsa00350: tyrosine metabolism			
13	path:hsa00500: starch and sucrose metabolism		Present	
14	path:rno04910: insulin signaling pathway			Present
15	path:hsa04110: cell cycle	Present	Present	Present
16	path:mmu04340: Hedgehog signaling pathway			
17	path:mmu00010: glycolysis/gluconeogenesis		Present	Present
18	path:hsa04340: Hedgehog signaling pathway			
19	path:hsa00330: arginine and proline metabolism		Present	
20	path:mmu04020: calcium signaling pathway			
21	path:hsa00280: valine, leucine, and isoleucine degradation			
22	path:mmu04530: tight junction	Present	Present	Present
23	path:rno04330: Notch signaling pathway	Present		
24	path:hsa04630: Jak-STAT signaling pathway	Present	Present	Present
25	path:hsa04350: TGF-β signaling pathway		Present	
26	path:mmu00650: butanoate metabolism			
27	path:hsa00620: pyruvate metabolism			
28	path:rno00564: glycerophospholipid metabolism		Present	Present

Table S3. Demographic and clinical characteristics by data sets used in the analyses

	Discovery data set:	Validation data sets	
Characteristics	GSE3141	GSE3593	GSE4573
Sample size	91	84	130
Age (years)			
Median	67	66	67
Range	32–83	33–82	42–91
Sex (%)			
Male	56 (62)	56 (67)	82 (63)
Female	35 (38)	28 (33)	48 (37)
Stage (%)			
I	67 (74)	52 (62)	73 (56)
II	18 (20)	15 (18)	34 (26)
III	6 (6)	17 (20)	23 (18)
Histology (%)			
Adenocarcinoma	45 (49)	84 (100)	0 (0)
Squamous	46 (51)	0 (0)	130 (100)
Survival (months)			
Mean	35.4	52.3	45.9