

Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Chen H-Y, Yu S-L, Chen C-H, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007;356:11-20.

Supplementary Appendix

Contents

I. Supplementary Methods

II. Supplementary Tables

III. Supplementary Figure Legend

IV. References

I. Supplementary Methods

Real-Time Reverse Transcription Polymerase Chain Reaction (Real-Time RT-PCR)

To validate the altered expression genes found by microarray analyses, real-time RT-PCR was performed for 16 genes plus one control, using pre-designed gene-specific TaqMan® probes and primer sets purchased from Applied Biosystems (Applied Biosystems, Branchburg, NJ) (Hs00154054_m1 for *ANXA5*, Hs00265843_m1 for *DLG2*, Hs00185667_m1 for *ZNF264*, Hs00169257_m1 for *DUSP6*, Hs00286741_m1 for *CPEB4*, Hs00178427_m1 for *LCK*, Hs00234829_m1 for *STAT1*, Hs00231302_m1 for *RNF4*, Hs00180031_m1 for *IRF4*, Hs01013132_m1 for *STAT2*, Hs00300159_m1 for *HGF*, Hs00176538_m1 for *ERBB3*, Hs00169714_m1 for *NF1*, Hs00234508_m1 for *FRAP1*, Hs00202450_m1 for *MMD*, Hs00234864_m1 for *HMMR*, and Hs00427620_m1 for *TBP*). Real-time RT-PCR amplification was carried out using Taqman One-Step RT-PCR Master Mix Reagent (Applied Biosystems, Branchburg, NJ) on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Branchburg, NJ), according to the manufacturer's instructions. Gene expression was quantified relative to the expression of *TBP* using Sequence Detector Software (Applied Biosystems, Branchburg, NJ) and the relative quantification method.

Data Pretreatment

For the analysis of whether the genes expression signature could predict clinical outcomes of non-small cell lung cancer, the samples were randomly

separated into training and testing datasets. There were no significant differences in clinicopathologic features between the two sets (Supplementary Table 1).

The averaged intensity for each gene expression was used for the subsequent analyses. To reduce the variation arising from experimental results derived from different microarrays, the intensity values of the spots from each microarray were rescaled using a quantile normalization method.¹ To reduce background noise, any background intensity lower than 3,000 was assigned the value of 3000.² After background correction, each microarray expression value was transformed to the logarithm of base 2. We used the coefficient of variation (CV) of 3% as the threshold to select those genes with largest variability in expression levels; genes with $CV < 3\%$ were excluded from further analyses. Finally, there were 485 genes out of 672 genes with $CV > 3\%$.

After logarithm-transformation, the intensity of each gene expression was transformed to an ordinal coding level according to the ranking of the gene expression level among the total expressions of 485 genes in 125 patients ($485 \times 125 = 60,625$ observations). The gene expression was coded as 1, 2, 3, or 4, respectively if the gene expression was ranked $\leq 25^{\text{th}}$, $> 25^{\text{th}}$ and $\leq 50^{\text{th}}$, $> 50^{\text{th}}$ and $\leq 75^{\text{th}}$, or $> 75^{\text{th}}$ percentile of total gene expressions.

Risk Score Calculation

There were 16 genes significantly correlated with patient survival through univariate Cox regression analysis, in which the regression coefficients were as follows: *NF1*, 0.47; *HGF*, 0.51; *HMMR*, 0.52; *IRF4*, 0.52; *ZNF264*, 0.55; *ERBB3*, 0.55; *STAT2*, 0.59; *CPEB4*, 0.59; *RNF4*, 0.65; *DUSP6*, 0.75; *MMD*, 0.92; *DLG2*,

1.32; *ANXA5*, -1.09; *LCK*, -0.84; *FRAP1*, -0.77; and *STAT1*, -0.58. Then a patient's risk score was derived by a summation of each gene expression level times its corresponding coefficient as follows: Risk score= (0.47 × *NF1* value) + (0.51 × *HGF* value) + (0.52 × *HMMR* value) + (0.52 × *IRF4* value) + (0.55 × *ZNF264* value) + (0.55 × *ERBB3* value) + (0.59 × *STAT2* value) + (0.59 × *CPEB4* value) + (0.65 × *RNF4* value) + (0.75 × *DUSP6* value) + (0.92 × *MMD* value) + (1.32 × *DLG2* value) - (1.09 × *ANXA5* value) - (0.84 × *LCK* value) - (0.77 × *FRAP1* value) - (0.58 × *STAT1* value). The risk score was used to classify patients into high or low risk signature, in which a high risk score indicated a poorer survival for patients. To avoid the effect of extreme values and set the number of patients in the two groups (high vs. low risk signature) equal in the training dataset, the 50th percentile (median) was chosen as the cut-off value. In the testing dataset, both the regression coefficients of risk score and the cut-off value derived from the training dataset were applied directly.

Decision Tree Model

To classify patients using 5-gene based expression profile by real-time RT-PCR, the decision tree model with univariate-splits was used in this study (Supplementary Fig. 1). One training set of samples (patients) with known class labels (high or low-risk) was used to train the tree model such that a decision tree with one decision rule accompanying each intermediate node is built for future

samples prediction. At each intermediate node, only one of the 5 genes was used to split that node into two daughter nodes. One of the 5 genes is selected with a cut-off value for partitioning all the patients in the current node into two as pure as possible daughter nodes among all potential gene/cut-off combinations. This splitting process is repeated for each intermediate node until no further split is necessary (either class labels are pure or expression profiles are homogeneous enough). To avoid over-fitting (since one can keep on splitting till there is only one patient per terminal node), a pruning method called minimum error was used to prune off unstable leaves of the tree. A goodness function was set as information gain of splitting and a leaf impurity percentage was set as global with 1% of misclassification. More detailed descriptions on the above processes can be found in the instruction manual of software (Avadis™) (Strand Genomic, Redwood, CA).³

4

Our rationale for using decision tree method here rather than the risk score method was that the former has been shown to be relatively easy to use when the number of genes was small as well as able to capture much of the relevant covariate structure, especially the complex interaction and non-linearity involved in the RT-PCR expression levels.

II. Supplementary Tables

Supplementary Table 1. Clinicopathologic features of 125 NSCLC patients

Characteristic	Training set	Testing set	P value
	No. of patients (%)	No. of patients (%)	
	n=63	n=62	
Age (mean±SD)	65.9±9.6	65.7±9.6	0.56 [†]
Gender			
Male	48 (76.2)	53 (85)	0.26 [‡]
Female	15 (23.8)	9 (15)	
Stage			
I	25 (39.7)	23 (37.1)	0.08 [§]
II	10 (15.9)	20 (32.3)	
III	28 (44.4)	19 (30.6)	
Primary tumor			
T1 and T2	44 (69.8)	46 (74.2)	0.69 [‡]
T3 and T4	19 (30.2)	16 (25.8)	
Regional lymph nodes			
N0	27 (42.9)	33 (53.2)	0.28 [‡]
N1, N2, and N3	36 (57.1)	29 (46.8)	
Cell type			
Adenocarcinoma	34 (54)	26 (41.9)	0.36 [§]
Squamous cell carcinoma	24 (38)	28 (45.1)	
Others	5 (8)	8 (13)	

[†]t test.

[‡]Fisher's exact test.

[§]Chi-squared test.

Supplementary Table 2. Clinical characteristics of patients in the independent set of published microarray data

Characteristic	Patients with high risk gene signature patients (%)	Patients with low risk gene signature (%)	P value
	n=62	n=24	
Age (mean±SD)	63.7±9.9	63.7±8.5	0.998 [†]
Gender			
Male	25 (40)	10 (42)	1 [‡]
Female	37 (60)	14 (58)	
Stage			
I	49 (79)	18 (75)	0.774 [‡]
III	13 (21)	6 (25)	

[†]t test

[‡]Fisher's exact test

III. Supplementary Figure Legend

Supplementary Figure 1. 5-gene signature and decision tree to classify NSCLC

patients into high and low risk gene signature groups.

Fifty seven patients with high risk and 44 patients with low risk gene signatures predicted by risk score model based on microarray gene

expressions were re-classified by 5-gene signature and decision tree

analysis. The gene expression of each patient was measured by

real-time RT-PCR.. Circles represent internal nodes. Beneath each node

is the gene whose expression level is used to split the node and the

cut-off value is displayed on the arrow to the right. Inside each node is

the node number (top), the number of original high risk gene signature

patients (middle), and the number of original low risk gene signature

patients (bottom). Boxes are terminal nodes. The patients were

re-assigned to high risk gene signature (total number of patients in pink

boxes, n=59) or low risk gene signature (total number of patients in blue

boxes, n=42) by decision tree analysis There are totally four patients

that are misclassified by the decision tree analysis as compared to the risk

score model.

IV. References

1. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;19:185-93.
2. Chen JJ, Lin YC, Yao PL, et al. Tumor-associated macrophages: the double-edged sword in cancer progression. *J Clin Oncol* 2005;23:953-64.
3. Shafer JC, Agrawal R, Mehta: M. SPRINT: A Scalable Parallel Classifier for Data Mining. In: *Proc of the 22th Int'l Conference on Very Large Databases*; 1996; India; 1996.
4. Avadis™. Avadis™ user manual. USA: Strand Genomic Pvt Ltd; 2004.

