Research Paper

Molecular gene signature and prognosis of non–small cell lung cancer

Poyin Huang[1,2,3,4], Chiou-Ling Cheng[5], Ya-Hsuan Chang[6], Chia-Hsin Liu[6], Yi-Chiung Hsu[6], Jin-Shing Chen[7], Gee-Chen Chang[8,9,10], Bing-Ching Ho[5], Kang-Yi Su[5], Hsuan-Yu Chen[6], Sung-Liang Yu[5,11,12,13,14]

[1]Department of Neurology, Kaohsiung Municipal Hsiao-Kang Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[2]Department of Neurology, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

[3]Ph.D. Program in Translational Medicine, Kaohsiung Medical University and Academia Sinica, Taiwan

[4]Department of Neurology, Faculty of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

[5]Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, National Taiwan University, Taipei, Taiwan

[6]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

[7]Department of Traumatology, National Taiwan University Hospital, Taipei, Taiwan

[8]Faculty of Medicine, School of Medicine, National Yang-Ming University, Taipei, Taiwan

[9]Division of Chest Medicine, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan

[10]Comprehensive Cancer Center, Taichung Veterans General Hospital, Taichung, Taiwan

[11]Center of Genomic Medicine, National Taiwan University, Taipei, Taiwan

[12]Center for Optoelectronic Biomedicine, College of Medicine, National Taiwan University, Taipei, Taiwan

[13]Graduate Institute of Pathology, College of Medicine, National Taiwan University, Taipei, Taiwan

[14]Department of Laboratory Medicine, National Taiwan University Hospital, Taipei, Taiwan

Correspondence to: Hsuan-Yu Chen, **email:** hychen@stat.sinica.edu.tw
Sung-Liang Yu, **email:** slyu@ntu.edu.tw

Keywords: *non–small cell lung cancer, TaqMan Low-Density Array, risk score, gene signature, prognosis*

Received: *March 28, 2016* ***Accepted:*** *June 30, 2016* ***Published:*** *July 16, 2016*

ABSTRACT

The current staging system for non–small cell lung cancer (NSCLC) is inadequate for predicting outcome. Risk score, a linear combination of the values for the expression of each gene multiplied by a weighting value which was estimated from univariate Cox proportional hazard regression, can be useful. The aim of this study is to analyze survival-related genes with TaqMan Low-Density Array (TLDA) and risk score to explore gene-signature in lung cancer. A total of 96 NSCLC specimens were collected and randomly assigned to a training (*n* = 48) or a testing cohort (*n* = 48). A panel of 219 survival-associated genes from published studies were used to develop a 6-gene risk score. The risk score was used to classify patients into high or low-risk signature and survival analysis was performed. Cox models were used to evaluate independent prognostic factors. A 6-gene signature including ABCC4, ADRBK2, KLHL23, PDS5A, UHRF1 and ZNF551 was identified. The risk score in both training (HR = 3.14, 95% CI: 1.14–8.67, *p* = 0.03) and testing cohorts (HR = 5.42, 95% CI: 1.56–18.84, *p* = 0.01) was the independent prognostic factor. In merged public datasets including GSE50081, GSE30219, GSE31210, GSE19188, GSE37745, GSE3141 and GSE31908, the risk score (HR = 1.50, 95% CI: 1.25–1.80, *p* < 0.0001) was also the independent prognostic factor. The risk score generated from expression of a small number of genes did perform well in predicting overall survival and may be useful in routine clinical practice.

INTRODUCTION

Lung cancer, predominantly non–small-cell lung cancer (NSCLC), is the most common cause of cancer deaths worldwide [1]. The relapse rate among patients with early-stage NSCLC is 40% within 5 years after potentially curative treatment [2]. The current TNM staging system provides guidance to the arrangement of initial treatments [3]. It is also a valuable indicator for predicting patient survival. However, for lung cancer, this system does not perform as well as in other cancers since 40% of patients with early stage in lung cancer relapsed within five years [4]. Thus, the current staging system for NSCLC is inadequate for predicting the outcome of treatment.

Results of molecular research may improve the management of patients. Advances in genomics and proteomics have generated many candidate markers with potential clinical value [5]. Gene expression profiling by microarray or real-time RT-PCR can be useful tools for identifying genes involved in the etiology or progression of cancer [5–17], as well as many other diseases [18–23]. The molecular signature can provide additional information for treatment decision. For example, the molecular status of the epidermal growth factor receptor (EGFR) in lung cancer [24–25] has been shown to have influence on the prognosis of patients and may indicate the introduction of alternative therapies. The combined use of TNM staging system and gene signature may enhance the prediction accuracy of survival and help avoid unnecessary treatment. Thus gene signature can provide extra information beyond the staging system. Furthermore, the gene signature also can be used to identify patients who are responsive to chemotherapy. After stratification by excision repair cross-complementation group 1 (ERCC1) expression, patients who received cisplatin-based chemotherapy had prolonged survival only for those with negative ERCC1 status but not for those with positive ERCC1 status [26–27]. A similar result was found in the study of gefinitib in lung cancer [28]. Lung adenocarcinoma with EGFR activating mutations had a higher response rate.

The genes that significantly correlated with clinical outcomes can be used to derive a predictive model for patients' survival. There were many algorithms available in the literature for predictive model developments [10, 29]. However, the complex structures of most algorithms or models have substantially reduced their potential in clinical application. But these limitations of microarray do not wipe out completely its benefits in exploratory studies wherein it is used as a screening tool. In contrast, RT-PCR is a faster and more stable assay, which is more suitable for clinical practice [30–33]. Although many studies reported that one single gene could predict clinical outcomes successfully, these findings need to be validated in more validation cohorts. A single gene may exhibit a strong association with clinical outcome before it was used as a classifier [34, 35]. In order for a single-gene-based classifier to reach a high accuracy level (i.e. sensitivity = 0.8 and specificity = 0.9), the odds ratio of the predictor gene needs to be as high as 228 [34, 35]. A combination of several potential genes may help to surpass this limitation. The small effect of each gene can be cumulated to improve the overall predictive power. For instance, the risk score, a linear combination of weighted gene expression, can be useful if properly constructed [4, 7, 31, 32, 36].

Until today, many studies demonstrate the gene signatures for survival in lung cancer by using public database [33, 37, 38]. However, there is no consistent gene-signature generated from the current studies. Hence, the aim of our study is to analyze the huge published survival-related genes measured by TaqMan Low-Density Array and risk score to explore the robust lung cancer gene-signature in lung cancer. The identification of these gene signatures will help scientists to develop not only robust gene signature for prognosis prediction but also the potential druggable targets for lung cancer.

RESULTS

Among the 96 NSCLC patients, 50 (52.08%) have adenocarcinoma and 46 (47.92%) have squamous cell carcinoma. There is one missing data in stage, thus 58 (61.05%) patients have stage I, 14 (14.74%) have stage II and 23 (24.21%) have stage III NSCLC. These patients had not received adjuvant chemotherapy with a median follow-up time of 32.13 months (range 3.83 to 109.33). Other basic characteristics of the patients were shown in Table 1. Among the 258 selected genes, 24 (LOC158381, ALPPL2, C3orf45, CALCA, CASR, CCKBR, CYP3A4, CYP3A43, DEFA6, FEV, FGF4, FLJ16124, FLJ21511, IGLL1, IL11, LCT, MEP1B, PDIA2, PTBP1, SLC1A7, SLC26A3, TBC1D29, MYCNOS, C5orf24) of them were excluded from the panel because more than half of the patients (> 48 patients) have undetermined gene expression level. Univariate Cox regression analysis was performed to find genes associated with survival and a total of 6 genes such as ABCC4, ADRBK2, KLHL23, PDS5A, UHRF1 and ZNF551 were found to have significant associations with overall survival (Table 2) Pathways analyses showed that there were no interactions between these genes and only ADRBK2 was shown to be significantly involved in Cardiac β-adrenergic Signaling pathway. These 6 genes were not involved in the same pathway thus collinearity should not be a concern. For each patient, a risk-score according to a linear combination of the expression level of these 6 genes was used to classify patients into high or low risk signature. In the training group (n = 48), 24 patients are low risk and 24 patients are high risk. In univariate Cox model, risk score classifying patients into high

Table 1: Basic clinical characteristics of the study population

	ALL (n = 96)		Training group (n = 48)		Testing group (n =48)		p–value
Gender							1.000
Female	22	22.92	11	22.92	11	22.92	
Male	74	77.08	37	77.08	37	77.08	
Histology							1.000
SCC	46	47.92	23	47.92	23	47.92	
Adenocarcinoma	50	52.08	25	52.08	25	52.08	
Stage							0.339
I/II	72	75.79	38	80.85	34	70.83	
III	23	24.21	9	19.15	14	29.17	
Stage							0.374
I	58	61.05	32	68.09	26	54.17	
II	14	14.74	6	12.77	8	16.67	
III	23	24.21	9	19.15	14	29.17	
Age							0.207
mean, sd	67.11	10.54	68.47	10.95	65.75	10.04	

Data are presented as mean ± standard deviation or n (%); p value by χ^2 or two-tailed t test. One patient has missing data in stage (n = 95).

or low risk signature (HR = 2.94, 95% CI: 1.26–6.85, p–0.01) was associated with patient survivals. In multivariate Cox model, risk score classifying patients into high or low risk signature (adjusted HR = 3.14, 95% CI: 1.14–8.67, p = 0.03), along with NSCLC stage (adjusted HR = 4.66, 95% CI: 1.51–14.39, p = 0.01), are both independent prognostic factors (Table 3 and Figure 1). In the testing group (n = 48), 31 patients are

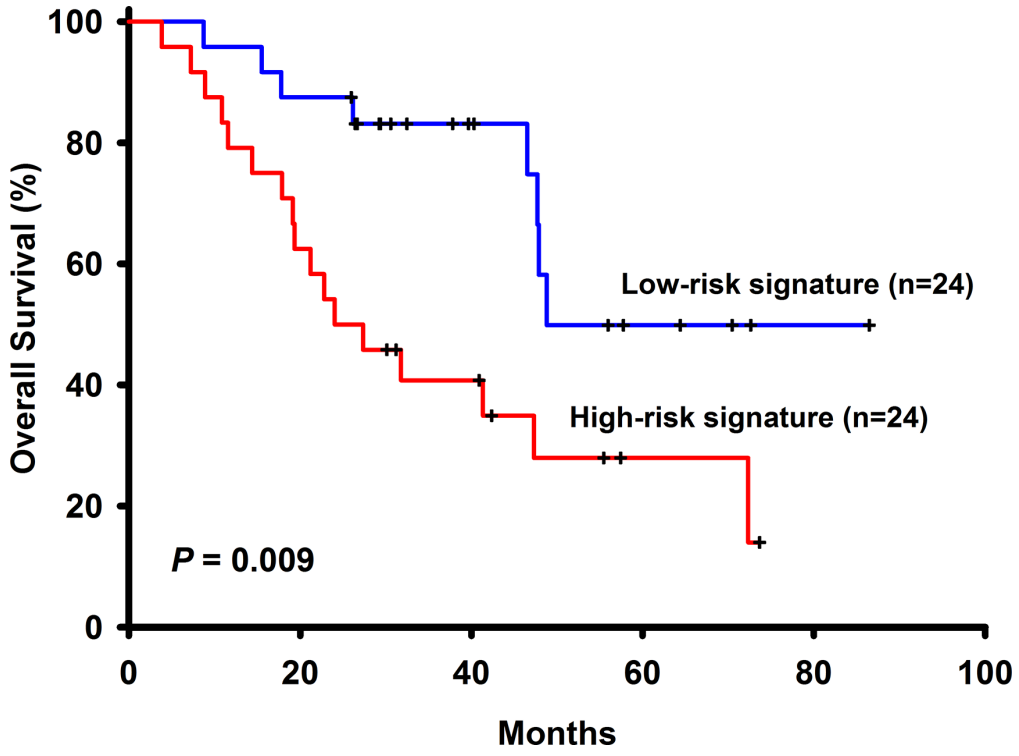

Figure 1: The survival analysis of the 6-gene signautre in the training cohort.

Table 2: 6-gene signature identified from Cox model of training group (n = 48)

Variable	coefficient	HR	HR, 95% CI		p-value
ABCC4-Hs00988734_m1	−0.22096	0.802	0.693	0.927	0.0029
ADRBK2-Hs01007260_m1	−0.52732	0.59	0.408	0.853	0.0051
KLHL23;PHOSPHO2-Hs00376354_m1	−0.64501	0.525	0.335	0.822	0.0049
PDS5A-Hs00374857_m1	−0.62813	0.534	0.361	0.79	0.0017
UHRF1-Hs01086727_m1	0.45151	1.571	1.129	2.185	0.0073
ZNF551-Hs00292939_m1	−0.28384	0.753	0.621	0.912	0.0037

low risk and 17 patients are high risk. Identical to the result in the training group, in univariate Cox model, risk score classifying patients into high or low risk signature (HR = 2.77, 95% CI:1.12–6.85, p = 0.03) was associated with patient survivals. In multivariate Cox model, risk score classifying patients into high or low risk signature (adjusted HR = 5.42, 95% CI: 1.56–18.84, p = 0.01), along with NSCLC stage (adjusted HR = 11.18, 95% CI: 3.43–36.40, p < 0.001), are both independent prognostic factors (Figure 2).

To further validate our findings, the risk score derived from 6 genes associated with overall survival was applied directly to merged public datasets including GSE50081, GSE30219, GSE31210, GSE19188, GSE37745, GSE3141 and GSE31908 GSE3141. The basic characteristics of the validation cohort were shown in Supplementary Table 1. In this dataset, result of the survival analysis showed that patients with high risk signature had shorter overall survival (p < 0.0001) (Figure 3). In univariate Cox model, the 6-gene risk signature was a risk factor of patients' survival (HR = 1.74, 95% CI: 1.47–2.05, p < 0.0001). In multivariate Cox model, the 6-gene risk signature (adjusted HR = 1.50, 95% CI: 1.25–1.80, p < 0.0001), histology (adjusted HR = 0.65, 95% CI: 0.54–0.78, p < 0.0001) and gender (adjusted HR = 1.43, 95% CI: 1.17–1.74, p = 0.0005) are independent prognostic factors. Overall, the risk score, based on a linear combination of the expression level of 6 genes, which classified patients into high or low risk signature, is consistently an independent prognostic factor associated with NSCLC patient survivals.

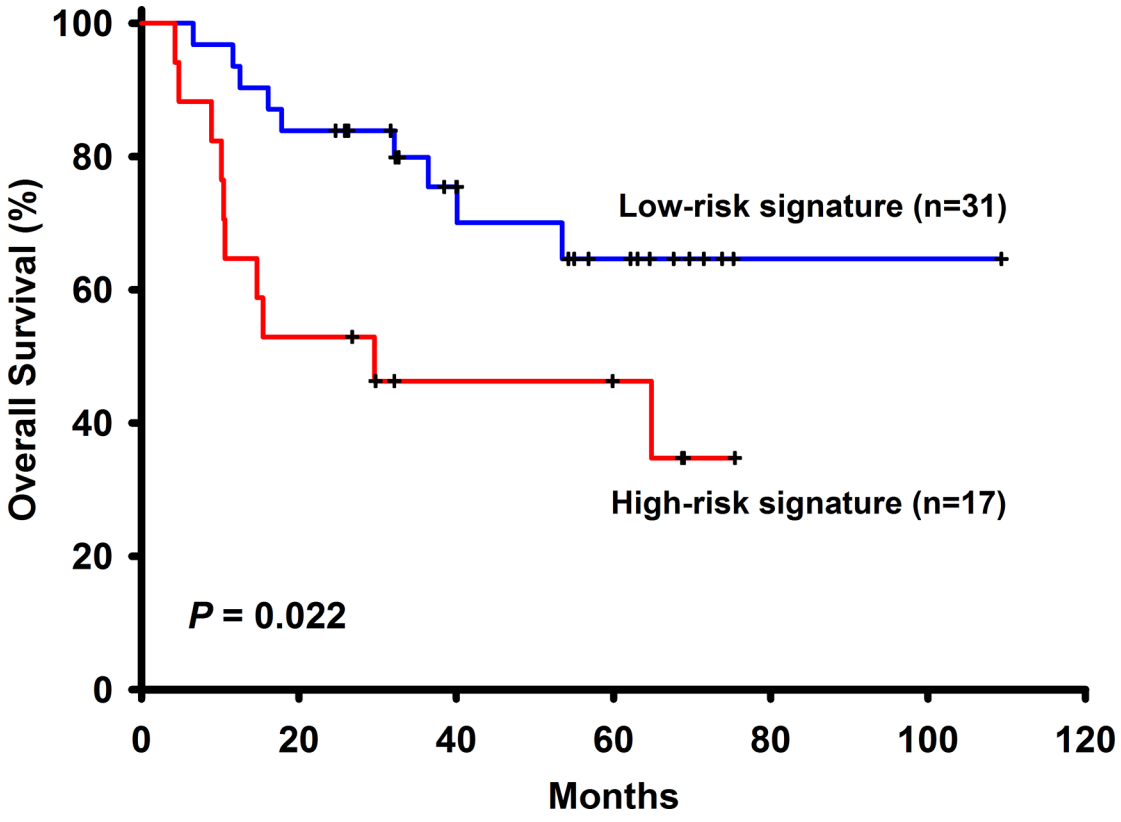

Figure 2: The survival analysis of the 6-gene signautre in the testing cohort.

Table 3: Multivariate Cox model of training, testing and validation cohorts

Variable	Hazard Ratio (95% CI)	*P* Value
Training cohort		
High-risk six-gene signature	3.14 (1.14–8.67)	0.027
Male	1.74 (0.39–7.82)	0.47
Older age	1.02 (0.97–1.07)	0.43
Adenocarcinoma	2.65 (0.95–7.43)	0.06
Tumor stage III	4.66 (1.51–14.39)	0.01
Testing cohort		
High-risk six-gene signature	5.42 (1.56–18.84)	0.008
Male	7.23 (0.81–64.97)	0.08
Older age	1.01 (0.94–1.08)	0.83
Adenocarcinoma	0.97 (0.35–2.70)	0.95
Tumor stage III	11.18 (3.43–36.40)	< .0001
Validation cohort		
High-risk six-gene signature	1.50 (1.25–1.80)	< .0001
Adenocarcinoma	0.65 (0.54–0.78)	< .0001
Male	1.43 (1.17–1.74)	0.0005

DISCUSSION

NSCLC is a heterogeneous disease resulting from multiple somatic mutations. Due to the complexity, it is less likely that a single gene expression pattern could be effectively used to predict the clinical course and outcome of NSCLC for all patients [15]. Instead, multiple sets of gene expression patterns may exist in tumors. Thus, it is believed that multiple sets of gene expression signatures that can be used for outcome prediction exist in NSCLC [32–33]. Despite the breakthrough in next-generation sequencing technology, microarray technologies are still

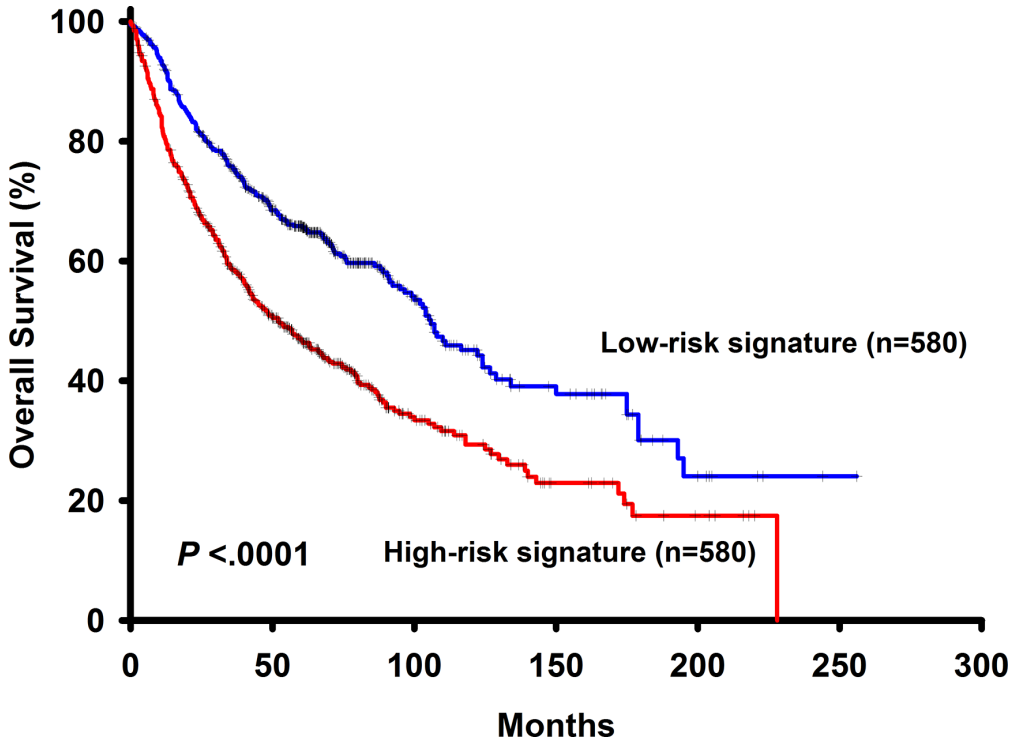

Figure 3: The survival analysis of the 6-gene signautre in the validation dataset.

useful platforms for biological exploration. Lung cancer has been among the earliest and most intensely studied diseases using microarray platforms [39]. Two very recent studies have used microarray technologies to derive a robust prognostic gene expression signature for early stage lung adenocarcinoma [40] and identify a 17 gene expression signature that distinguishes lymphangiogenic from non-lymphangiogenic NSCLC cell lines [41]. Molecular signatures help to reveal the biologic spectrum of lung cancers, throw light on the pathogenetic alterations in gene expressions and cellular pathways, identify prognostic and predictive gene signatures, customize therapies, identify new therapeutic targets and evaluate new drugs [39]. The small effect of each gene can be cumulated and a combination of several potential genes may help to improve the overall predictive power. In this study , we use the risk score algorithm to combine several potential genes to surpass the limitation of using a single gene expression pattern to predict NSCLC outcome.

Adenocarcinomas and squamous carcinomas are distinct disease entities with different gene expression patterns, thus using independent prognostic signatures for squamous carcinomas and adenocarcinomas should be more biologically significant and less affected by genetic heterogeneities [32, 37]. Generally, the gene signature selected from one cell type is predictive for that specific cell type. Whether gene signatures from different cell types can be predictive for each other in NSCLC is still an unidentified issue, with implications as to how gene expression signatures could be translated into clinical practice. One previous study has shown that a prognostic signature may not be cell type specific and that a universal signature reflecting tumor aggressiveness and subsequent clinical outcome may exist across histologic cell types [37]. This would be of clinical importance because unified gene signatures would dramatically simplify the prognosis evaluation process for different types of carcinoma. Since a gene signature selected from adenocarcinoma or squamous cell carcinoma may be predictive for both histologic subtypes, specific prognostic signatures for NSCLC will be more attractive due to their broader applicability [37–38]. Thus, patients of squamous carcinoma and adenocarcinoma are equally represented in this study so that cancer specific, rather than histology specific signatures could be generated by our experimental method.

There are some potential clinical implications of our findings. First of all, they provide a prognostic tool. Furthermore, they might identify targets for molecular therapies and aid in directing the options of therapeutic regime. Their potential exploitation for the identification of novel therapeutic targets is directly linked to whether the gene panel is associated with lung carcinogenesis. However, expression profile analysis generally can not lead to immediate biological implications [24–25]. As might be expected, these genes, if without an established role in the pathogenesis of NSCLC, may be proved to be of no prognostic value when used individually. However, although the mechanisms by which some of these genes affect patient survival are not clarified, at least it seemed that the expression pattern of these genes might have some crucial information for NSCLC prognosis. In order to understand how these 6-signature identified in this study may influence survival in patients with NSCLC, information on the functions of the encoded proteins was obtained from the GeneCards database (http://www.genecards.org/) and described as follows: PDS5A is a cell cycle related gene. PDS5A has been shown to improve cell proliferation in G2/M phase. It may contribute to tumorigenesis by upregulating p63 and promoting cell cycle progression. PDS5A is a nuclear protein and involves in the establishment, maintenance and dissolution of sister chromatid cohesion. Altered expression levels of PDS5A have been observed in tumors of the breast, kidney, esophagus, stomach, liver and colon and in malignant gliomas [42]. ABCC4 is a protein coding gene and the protein encoded is a member of the superfamily of ATP-binding cassette transporters. These proteins transport various molecules across extra and intracellular membranes. This protein also involved in multidrug resistance. Diseases associated with ABCC4 include lung cancer and hemostasis is involved in its associated pathways. ABCC4 was highly expressed in lung cancer cell lines. ABCC4 expression was markedly downregulated in A549 and 801D cells using the RNA interference technique. Suppression of ABCC4 expression inhibited cell growth. The percentage of cells in G1 phase was increased when ABCC4 expression was suppressed. Phosphorylation of retinoblastoma protein was weakened, originating in the downregulation of ABCC4. ABCC4 mRNA was highly expressed in lung cancer tissue and lung cancer cell lines. ABCC4 is a potential target for lung cancer therapy [43]. ADRBK2 encodes the beta adrenergic receptor kinase which specifically phosphorylates the agonist occupied form of the beta adrenergic and related G protein coupled receptors. The existence of this receptor kinase may serve to desensitize synaptic receptors. Endocytosis and chemokine signaling are involved in its associated pathways. The β2-adrenergic receptor is most frequently involved in carcinogenic processes. Earlier studies have established a relation between the β2-adrenergic receptor and various characteristics of cancer including cell proliferation, apoptosis, chemotaxis, metastasis, tumor growth and angiogenesis [44]. KLHL23 has strong similarities with gene KLHL7 and not much information could be obtained regarding KLHL23. KLHL7 antibodies are associated with various types of cancer, such as ovary, rectum, colon, lung and prostate cancer. However, the function of KLHL7 is unknown [45]. UHRF1 can influence the cell cycle control mainly through the epigenetic silencing of relevant tumor suppressors. UHRF1 mRNA was overexpressed

in NSCLC tumor tissues in comparison to their normal adjacent tissue. UHRF1 is a key epigenetic switch, which controls cell cycle in NSCLC through its ability to sustain the transcriptional silencing of tumor suppressor genes by maintaining their promoters in a hypermethylated status [46]. ZNF551 is a protein coding gene and may be involved in transcriptional regulation, no other information could be obtained from the literature. Further research on their potential functional role in modulating the survival of NSCLC patients is warranted [47].

In conclusion, we selected 6 genes that were associated with NSCLC patients' survival from expression data obtained by TLDA. The risk score algorithm was then successfully used to categorize patients into better and poorer prognosis since that the risk score algorithm did perform well in predicting overall survival for NSCLC in validation dataset. The result is of clinical interest that the risk score generated from expression of a relatively small number of genes [48] may be useful in routine clinical practice. Additionally, pair-samples were not available for analysis. However, since the aim of this study is to develop gene signature for prognosis prediction, the influence of not using normal tissues for analyses may be attenuated.

MATERIALS AND METHODS

Patients and tissue specimens

The frozen specimens of tumor tissue from 96 patients who underwent surgical resection of NSCLC at the Taichung Veterans General Hospital and the National Taiwan University Hospital were enrolled in this study Squamous carcinoma ($n = 46$) and adenocarcinoma ($n = 50$) histology are represented equally so that cancer-specific, rather than histology-specific markers may be elicited by the experimental method [46–47]. None of the patients underwent radiotherapy or chemotherapy prior to surgery. These patients had not received adjuvant chemotherapy with a median follow-up time of 32.13 months (range 3.83 to 109.33). Complete clinical information such as age, gender, stage, histological cell type, follow-up time, and survival status were collected. The present study was approved by the Ethics Committee and written informed consent was acquired from each patient.

RNA extraction and mRNA isolation

Total RNA isolation was performed by using RNazol B reagent. Cells were lysed directly in a culture T-flask (Falcon) by adding 1.5 ml of RNazol B reagent (1 ml/100 mm^2) to a 150 mm^2 tissue flask, and total RNAs from the cells were obtained by using RNAzol™ B to extract total RNAs and 100% isopropanol to precipitate RNAs. OligotexTM mRNA Midi Kit (QIAGEN) was used to gain mRNAs from the total RNAs mentioned above.

Gene expression profiling

TaqMan low-density RT PCR arrays (Applied Biosystems) were designed to determine expression of the selected genes. Total RNA (0.5 µg) extracted was reverse transcribed with 200 U Superscript II RT (Invitrogen) and 250 ng random hexamers. A reaction mix containing 75 ng of cDNA and 50 µl of 2× PCR Master Mix (Euregentec) was added to a TaqMan microfluidic card. Reverse transcription and rt PCR was performed in a 7900HT RT PCR System (Applied Biosystems). The cycling program used was 50°C for 30 min, 94.5°C for 15 min, then 40 cycles of 96°C for 30 s and 59.7°C for 1 min. The expression level of each gene was measured in triplicate, and a panel of reference genes (ACTB, B2M, GAPDH, GUSB, HMBS, HPRT1, IPO8, PGK1, POLR2A, PPIA, TBP, TFRC, UBC, YWHAZ, 18S) was used. The average Ct value of each target gene was normalized against the geometric mean of the Ct values of the 15 reference genes. Relative gene expression was expressed as $2^{-\Delta Ct}$.

Candidate gene selection

A panel of genes correlated with survival for lung cancer were collected from published studies. The including criteria of studies are journal quality, microarray study, and study of meta-analysis. A total of 258 genes (including 15 endogenous control genes, Supplementary Table 2) correlated with patients' survival published on Journal of Clinical Oncology and Journal of Clinical Investigation were selected [32–33, 37–38]. Among these 258 genes, 24 undetermined (> 48 patients) genes and 15 reference genes were excluded from further analyses. Thus, a total of 219 genes remained.

Training and testing group

The flow chart of this study was shown in Supplementary Figure 1. The 96 specimens were randomly assigned to a training cohort ($n = 48$) or a testing cohort ($n = 48$). Using the training cohort, based on the expression level of each gene, univariate Cox model was performed. Thus, a total of 219 models were developed. In each attempt to acquire ideal gene signature, 6 genes with p value < 0.01 in the univariate Cox model were selected to form the risk score.

Risk score algorithm

To investigate the effectiveness of candidate genes as a gene signature for clinical outcome prediction, a mathematical formula for survival prediction was constructed, taking into account both the strength and the positive or negative association of each gene with survival. More specifically, we assigned each patient, a risk-score

according to a linear combination of the expression level of the genes, weighted by the regression coefficients derived from aforementioned univariate Cox regression analyses [7, 31]. There were 6 genes significantly had high or low risk for patient survival through univariate Cox regression analysis and the regression coefficients were as follows: ABCC4, −0.22096; ADRBK2, −0.52732; KLHL23, −0.64501; PDS5A, −0.62813; UHRF1, 0.45151; ZNF551, –0.28384 (Table 2). Then a patient's risk score was derived by a summation of each gene expression level times its corresponding coefficient. Risk score can be expressed as: −(0.22096 × ABCC4 value) − (0.52732 × ADRBK2 value) − (0.64501 × KLHL23 value) − (0.62813 × PDS5A value) + (0.45151 × UHRF1 value) − (0.28384 × ZNF551 value). The risk score was used to classify patients into high or low risk signature based on median as cut-off point. In the testing dataset, both the regression coefficients of risk score and the cut-off value derived from the training dataset were applied directly.

Ingenuity pathway analysis

The 6 survival-associated genes were further analysed for biological function or involvement in different pathways using Ingenuity Pathway Analysis software (IPA, Igenuity Systems, USA).

Survival analysis

The Kaplan-Meier method was used to estimate overall survival. Differences in survival between the two groups were analyzed using the log-rank test. Both univariate and multivariate Cox proportional hazard regression analyses were used to evaluate independent prognostic factors associated with patient survivals. Age, gender, histology and stage were used as covariates.

Validation cohort

Public datasets including GSE50081, GSE30219, GSE31210, GSE19188, GSE37745, GSE3141 and GSE31908 were merged for validation analysis. These datasets used Affymetrix U133 Plus 2.0 as platform and there is a total of 1160 NSCLC patients available for analysis. The compatible probesets of acquired 6-gene signature were identified and the gene expression levels were obtained (Supplementary Table 3). Variables with more than 10% missing data were not added in the multivariate analysis.

ACKNOWLEDGMENTS

We thank the NCFPB Integrated Core Facility for Functional Genomics, NRPB Pharmacogenomics Lab, MOST and the Microarray Core Facility of the National Taiwan University Center of Genomic Medicine for technical support and Mathematics in Biology Group of Institute of Statistical Science AS supported genomic, biological, and bioinformatics work.

CONFLICTS OF INTEREST

None.

FUNDING

Supported by grants from Academia Sinica, Institute of Statistical Science AS, AS-100-TP-AB2, NSC 98- 3112-B-001-034, NSC 99-2314-B-001-003-MY3, NSC 100-2325-B-001-027, NSC 101-2325-B-002-071, NSC 102-2325-B-002-078, NSC 101-2319-B-002-002, NSC 102-2319-B-002-002, NSC 102-2911-I-002-303, NSC 101-2911-I-002-303, NSC 102-2911-I-002-303, DOH101- TD-B-111-001, 102R7557, NSC 102-2923-B-002-004, NSC 103-2923-B-002-003, MOST 103-2325-B-002-026, MOST 104-2923-B-002-003, MOST 104-2314-B-075A-012, MOST104-2911-I-002-302 and Taiwan Biosignature Project of Lung Cancer.

REFERENCES

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ. Cancer statistics, 2008. CA Cancer J Clin. 2008; 58:71–96.
2. Miller YE. Pathogenesis of lung cancer: 100 year report. Am J Respir Cell Mol Biol. 2005; 33:216–23.
3. Ludwig JA, Weinstein JN. Biomarkers in Cancer Staging, Prognosis and Treatment Selection. Nat Rev Cancer. 2005; 5:845–856.
4. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med. 2002; 8:816–24.
5. Mairinger FD, Ting S, Werner R, Walter RF, Hager T, Vollbrecht C, Christoph D, Worm K, Mairinger T, Sheu-Grabellus SY, Theegarten D, Wohlschlaeger J. Different micro-RNA expression profiles distinguish subtypes of neuroendocrine tumors of the lung: results of a profiling study. Mod Pathol. 2014; 27:1632–40.
6. Bullinger L, Döhner K, Bair E, Fröhling S, Schlenk RF, Tibshirani R, Döhner H, Pollack JR. Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. N Engl J Med. 2004; 350:1605–1616.
7. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, et al. A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer. N Engl J Med. 2007; 356:11–20.

8. Dave SS, Fu K, Wright GW, Lam LT, Kluin P, Boerma EJ, Greiner TC, Weisenburger DD, Rosenwald A, Ott G, Müller-Hermelink HK, Gascoyne RD, Delabie J, et al. Molecular Diagnosis of Burkitt's Lymphoma. N Engl J Med. 2006; 354:2431–2442.
9. Kim HS, Mitsudomi T, Soo RA, Cho BC. Personalized therapy on the horizon for squamous cell carcinoma of the lung. Lung Cancer. 2013; 80:249–55.
10. Zhou JX, Yang H, Deng Q, Gu X, He P, Lin Y, Zhao M, Jiang J, Chen H, Lin Y, Yin W, Mo L, He J. Oncogenic driver mutations in patients with non-small-cell lung cancer at various clinical stages. Ann Oncol. 2013; 24:1319–25.
11. Li T, Kung HJ, Mack PC, Gandara DR. Genotyping and genomic profiling of non-small-cell lung cancer: implications for current and future therapies. J Clin Oncol. 2013; 31:1039–49.
12. Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet. 2006; 7:200–210.
13. O'Shaughnessy JA. Molecular Signatures Predict Outcomes of Breast Cancer. N Engl J Med. 2006; 355:615–617.
14. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, et al. Genomic signatures to guide the use of chemotherapeutics. Nat Med. 2006; 12: 1294–1300.
15. Guo X, Li H, Fei F, Liu B, Li X, Yang H, Chen Z, Xing J. Genetic variations in SLC3A2/CD98 gene as prognosis predictors in non-small cell lung cancer. Mol Carcinog. 2015; 54:E52–60.
16. Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, Lu C, Woodgett J, Seiden I, Johnston M, Keshavjee S, Darling G, Winton T, et al. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. Cancer Res. 2002; 62:3005–8.
17. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med. 2008; 14:822–827.
18. Hakak Y, Walker JR, Li C, Wong WH, Davis KL, Buxbaum JD, Haroutunian V, Fienberg AA. Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. PNAS. 2001; 98:4746–4751.
19. Mehra MR, Uber PA, Walther D, Vesely M, Wohlgemuth JG, Prentice J, Tayama D, Billingham M. Gene Expression Profiles and B-Type Natriuretic Peptide Elevation in Heart Transplantation: More Than a Hemodynamic Marker. Circulation. 2006; 114:I-21–26.
20. Bull TM, Coldren CD, Geraci MW, Voelkel NF. Gene Expression Profiling in Pulmonary Hypertension. Proc Am Thorac Soc. 2007; 4:117–120.
21. Izuhara K, Saito H. Microarray-based identification of novel biomarkers in asthma. Allergol Int. 2006; 55:361–7.
22. Molina-Pinelo S, Gutiérrez G, Pastor MD, Hergueta M, Moreno-Bueno G, García-Carbonero R, Nogal A, Suárez R, Salinas A, Pozo-Rodríguez F, Lopez-Rios F, Agulló-Ortuño MT, Ferrer I, et al. MicroRNA-dependent regulation of transcription in non-small cell lung cancer. PLoS One. 2014; 9:e90524.
23. Hoffman PC, Mauer AM, Vokes EE. Lung cancer. Lancet. 2000; 355:479–85.
24. Yuan S, Yu SL, Chen HY, Hsu YC, Su KY, Chen HW, Chen CY, Yu CJ, Shih JY, Chang YL, Cheng CL, Hsu CP, Hsia JY, et al. Clustered genomic alterations in chromosome 7p dictate outcomes and targeted treatment responses of lung adenocarcinoma with EGFR-activating mutations. J Clin Oncol. 2011; 29:3435–42.
25. Chen HY, Liu CH, Chang YH, Yu SL, Ho BC, Hsu CP, Yang TY, Chen KC, Hsu KH, Tseng JS, Hsia JY, Chuang CY, Chang CS, et al. EGFR-activating mutations, DNA copy number abundance of ErbB family, and prognosis in lung adenocarcinoma. Oncotarget. 2016; 7:9017–25. doi: 10.18632/oncotarget.7029.
26. The International Adjuvant Lung Cancer Trial Collaborative G. Cisplatin-Based Adjuvant Chemotherapy in Patients with Completely Resected Non-Small-Cell Lung Cancer. N Engl J Med. 2004; 350:351–360.
27. Olaussen KA, Dunant A, Fouret P, Brambilla E, André F, Haddad V, Taranchon E, Filipits M, Pirker R, Popper HH, Stahel R, Sabatier L, Pignon JP, et al. DNA Repair by ERCC1 in Non-Small-Cell Lung Cancer and Cisplatin-Based Adjuvant Chemotherapy. N Engl J Med. 2006; 355:983–991.
28. Chou TY, Chiu CH, Li LH, Hsiao CY, Tzen CY, Chang KT, Chen YM, Perng RP, Tsai SF, Tsai CM. Mutation in the tyrosine kinase domain of epidermal growth factor receptor is a predictive and prognostic factor for gefitinib treatment in patients with non-small cell lung cancer. Clin Cancer Res. 2005; 11:3750–7.
29. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001; 7:673–9.
30. Ramaswamy S. Translating Cancer Genomics into Clinical Oncology. N Engl J Med. 2004; 350:1814–1816.
31. Lossos IS, Czerwinski DK, Alizadeh AA, Wechser MA, Tibshirani R, Botstein D, Levy R. Prediction of Survival in Diffuse Large-B-Cell Lymphoma Based on the Expression of Six Genes. N Engl J Med. 2004; 350:1828–1837.
32. Endoh H, Tomida S, Yatabe Y, Konishi H, Osada H, Tajima K, Kuwano H, Takahashi T, Mitsudomi T. Prognostic model of pulmonary adenocarcinoma by expression profiling of eight genes as determined by quantitative real-time reverse transcriptase polymerase chain reaction. J Clin Oncol. 2004; 22:811–9.

33. Bianchi F, Nuciforo P, Vecchi M, Bernard L, Tizzoni L, Marchetti A, Buttitta F, Felicioni L, Nicassio F, Di Fiore PP. Survival prediction of stage I lung adenocarcinomas by expression of 10 genes. J Clin Invest. 2007; 117:3436–3444.
34. Ware JH. The Limitations of Risk Factors as Prognostic Tools. N Engl J Med. 2006; 355:2615–2617.
35. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. Am J Epidemiol. 2004; 159:882–890.
36. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS. Multiple Biomarkers for the Prediction of First Major Cardiovascular Events and Death. N Engl J Med. 2006; 355:2631–2639.
37. Sun Z, Wigle DA, Yang P. Non-Overlapping and Non-Cell-Type-Specific Gene Expression Signatures Predict Lung Cancer Survival. J Clin Oncol. 2008; 26:877–883.
38. Lau SK, Boutros PC, Pintilie M, Blackhall FH, Zhu CQ, Strumpf D, Johnston MR, Darling G, Keshavjee S, Waddell TK, Liu N, Lau D, Penn LZ, et al. Three-Gene Prognostic Classifier for Early-Stage Non Small-Cell Lung Cancer. J Clin Oncol. 2007; 25:5562–5569.
39. Au JS, Cho WC, Chow SC, Yang M. DNA Expression Profi les of Non-Small Cell Lung Cancer. Chin J Lung Cancer. 2009; 12:8–15.
40. Krzystanek M, Moldvay J, Szüts D, Szallasi Z, Eklund AC. A robust prognostic gene expression signature for early stage lung adenocarcinoma. Biomark Res. 2016; 4:4.
41. Regan E, Sibley RC, Cenik BK, Silva A, Girard L, Minna JD, Dellinger MT. Identification of Gene Expression Differences between Lymphangiogenic and Non-Lymphangiogenic Non-Small Cell Lung Cancer Cell Lines. PLoS One. 2016; 11:e0150963.
42. Put N, Van Roosbroeck K, Vande Broek I, Michaux L, Vandenberghe P. PDS5A, a novel translocation partner of MLL in acute myeloid leukemia. Leuk Res. 2012; 36:e87–9.
43. Zhao X, Guo Y, Yue W, Zhang L, Gu M, Wang Y. ABCC4 is required for cell proliferation and tumorigenesis in non-small cell lung cancer. Onco Targets Ther. 2014; 21:343–51.
44. Pérez-Sayáns M, Somoza-Martín JM, Barros-Angueira F, Gayoso-Diz P, Otero-Rey EM, Gándra-Rey JM, García-García A. Activity of β2-adrenergic receptor in oral squamous cell carcinoma is mediated by overexpression of the ADRBK2 gene: a pilot study. Biotech Histochem. 2012; 87:179–86.
45. Bredholt G, Storstein A, Haugen M, Krossnes BK, Husebye E, Knappskog P, Vedeler CA. Detection of autoantibodies to the BTB-kelch protein KLHL7 in cancer sera. Scand J Immunol. 2006; 64:325–35.
46. Daskalos A, Oleksiewicz U, Filia A, Nikolaidis G, Xinarianos G, Gosney JR, Malliri A, Field JK, Liloglou T. UHRF1-mediated tumor suppressor gene inactivation in nonsmall cell lung cancer. Cancer. 2011; 117:1027–37.
47. Galvan A, Frullanti E, Anderlini M, Manenti G, Noci S, Dugo M, Ambrogi F, De Cecco L, Spinelli R, Piazza R, Pirola A, Gambacorti-Passerini C, Incarbone M, et al. Gene expression signature of non-involved lung tissue associated with survival in lungadenocarcinoma patients. Carcinogenesis. 2013; 34:2767–73.
48. Park YY, Park ES, Kim SB, Kim SC, Sohn BH, Chu IS, Jeong W, Mills GB, Byers LA, Lee JS. Development and validation of a prognostic gene-expression signature for lung adenocarcinoma. PLoS One. 2012; 7:e44225.
49. Villaruz LC, Burns TF, Ramfidis VS, Socinski MA. Personalizing therapy in advanced non-small cell lung cancer. Semin Respir Crit Care Med. 2013; 34:822–36.
50. Hensing T, Chawla A, Batra R, Salgia R. A personalized treatment for lung cancer: molecular pathways, targeted therapies, and genomic characterization. Adv Exp Med Biol. 2014; 799:85–117.