Clinical Validity of the Lung Cancer Biomarkers Identified by Bioinformatics Analysis of Public Expression Data

Bumjin Kim,[1] Hyun Joo Lee,[2] Hye Young Choi,[3] Youngah Shin,[1] Seungyoon Nam,[1,4] Gilju Seo,[3] Dae-Soon Son,[3] Jisuk Jo,[3] Jaesang Kim,[1] Jinseon Lee,[3] Jhingook Kim,[2] Kwhanmien Kim,[2] and Sanghyuk Lee[1]

[1]Division of Life and Pharmaceutical Sciences, Ewha Womans University; [2]Department of Thoracic Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine; [3]Cancer Research Division, Center for Clinical Research, Samsung Biomedical Research Institute; and [4]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea

Abstract

Identification of molecular markers often leads to important clinical applications such as early diagnosis, prognosis, and drug targeting. Lung cancer, the leading cause of cancer-related deaths, still lacks reliable molecular markers. We have combined the bioinformatics analysis of the public gene expression data and clinical validation to identify biomarker genes for non–small-cell lung cancer. The serial analysis of gene expression and the expressed sequence tag data were meta-analyzed to produce a list of the differentially expressed genes in lung cancer. Through careful inspection of the predicted genes, we selected 20 genes for experimental validation using semiquantitative reverse transcriptase-PCR. The microdissected clinical specimens used in the study consisted of three groups: lung tissues from benign diseases and the paired (cancer and pathologic normal) tissues from non–small-cell lung cancer patients. After extensive statistical analyses, seven genes (*CBLC, CYP24A1, ALDH3A1, AKR1B10, S100P, PLUNC*, and *LOC147166*) were identified as potential diagnostic markers. Quantitative real-time PCR was carried out to additionally assess the value of the seven identified genes leading to the confirmation of at least two genes (*CBLC* and *CYP24A1*) as highly probable novel biomarkers. The gene properties of the identified markers, especially their relationship to lung cancer and cell signaling pathway regulation, further suggest their potential value as drug targets as well. [Cancer Res 2007;67(15):7431–8]

Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide. Compared with other major types of cancer such as colon, prostate, and breast cancers, the clinical outcome of conventional therapies such as surgery, radiotherapy, and chemotherapy still remains poor despite the major efforts to improve treatment methods during past decades. The most important reasons for the lack of improvement in prognosis are the difficulties in making the early-stage diagnosis of lung cancer and the high recurrence rate after curative treatments (1). Thus, identification and validation of diagnostic and prognostic biomarkers is tremendously important to improve the clinical outcome of lung cancer treatments (2).

The emergence of high-throughput molecular tools such as microarrays and large-scale databases of serial analysis of gene expression (SAGE) and expressed sequence tags (EST) brought a new paradigm for biomarker discovery. Numerous candidate biomarkers have been reported for risk assessment, screening, diagnosis, prognosis, and for selection and monitoring of therapies since the application of such new technologies and techniques had been initiated (3). For example, microarrays have been successfully applied to find new classes of diseases, to predict prognosis, and to identify diagnostic markers for early detection (4). Typically, these studies have used classifiers that consist of tens or hundreds of genes. However, it is still not clear whether or not such molecular signatures would be more effective than a few biomarkers of high sensitivity and specificity.

SAGE is a sequencing-based technique for quantitatively profiling the gene expression. Tag counts provide an estimation of the expression levels in SAGE. The merit of the technique is that it does not require any prior knowledge of the gene sequence, thus generating unbiased profiling even for unknown genes (5). Metadata analysis is relatively simple due to the simple nature of the data. Two research groups have used the SAGE technique to find the differentially expressed genes as biomarkers for lung cancer (6, 7).

EST is a relatively low-throughput technique as compared with the microarray or SAGE techniques. However, the vast amount of public data in dbEST makes it a valuable information source for studying gene expression pattern as can be seen in the case of the Cancer Genome Anatomy Project (8). In addition, several studies have reported the successful application of EST data to find tissue-specific and/or cancer-specific genes (9).

SAGE and EST data are often complementary in their library characteristics. Gene expression from EST is not quantitative because preparation of many libraries has included normalization or subtraction steps. SAGE provides a quantitative profile, but the number of public libraries is rather small compared with the EST cDNA libraries (~300 versus ~8,600 for human). Furthermore, the tag-to-gene assignment is not a trivial task in SAGE analysis. Bioinformatics methods that integrate these public data, taking merits and shortcomings of each into consideration, would significantly facilitate the discovery of biomarkers.

A major obstacle in finding biomarkers with using public expression data is the problem of sample heterogeneity. Many cDNA libraries from the Cancer Genome Anatomy Project are from

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

B. Kim and H.J. Lee contributed equally to this work.

Requests for reprints: Sanghyuk Lee, Division of Life and Pharmaceutical Sciences, Ewha Womans University, 11-1 Daehyun-dong, Seodaemun-gu, Seoul 120-750, Korea. Phone: 82-2-3277-2888; Fax: 82-2-3277-3760; E-mail: sanghyuk@ewha.ac.kr or Kwhanmien Kim, Department of Thoracic Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, 50 Ilwon-dong, Gangnam-gu, Seoul 135-710, Korea. Phone: 82-2-3410-3485; Fax: 82-2-3410-0089; E-mail: kwhanmien.kim@samsung.com.

doi:10.1158/0008-5472.CAN-07-0003

bulk samples and cell lines. The coverage of SAGE libraries is quite limited and, consequently, lacks statistical significance for valid biomarkers. For example, only six SAGE libraries are publicly available for lung tissues, and their preparation protocols significantly vary from one another. Thus, it is essential to critically validate the predicted candidate markers using a number of well-defined clinical samples. Investigating the molecular properties and pathways they are involved in may also be helpful for deciding the value of the genes as drug targets in addition to biomarkers.

We describe here a bioinformatics method and clinical validation to identify diagnostic marker genes for lung cancer. The SAGE and EST data were meta-analyzed to produce a list of differentially expressed genes in lung cancers. A systematic examination of the annotated gene properties led to 20 genes which were subjected to experimental validation using clinical specimens from lung cancer patients. Semiquantitative reverse transcriptase-PCR (RT-PCR) followed by extensive statistical analyses established seven genes (*CBLC, CYP24A1, ALDH3A1, AKR1B10, S100P, PLUNC*, and *LOC147166*) as the potential diagnostic markers for lung cancer. Quantitative real-time PCR experiments were carried out with additional samples for the seven identified markers as the final validation step. We further describe the molecular properties of these genes, especially their relationship to lung cancer and regulatory signaling pathways, to examine their value beyond diagnostic or prognostic biomarkers.

Materials and Methods

We used the ECgene system throughout the bioinformatics analyses. ECgene is a novel algorithm that combines genome-based EST clustering and graph-based transcript assembly procedures (10). Its EST clustering is conceptually similar to the genome-based version of UniGene clustering. The main advantage of ECgene is that it produces transcript models and subclustering according to alternative splicing. Such mRNA models are useful for assigning genes to SAGE tags.

The clustering algorithm is more conservative in terms of selecting valid alignments and splice site sharing. Approximately 6% to 7% of ESTs in the UniGene do not satisfy our conservative criteria of alignment. Additional 5% or so of ESTs belong to different EST clusters, most of which turned out to be neighboring clusters. We found that such ESTs aligned with the opposite strand or in the intronic regions without sharing any splice sites. Approximately 11% to 12% difference in EST members leads to a substantial difference in the final result. The schematic overview of our approach is given in Fig. 1.

Analysis of SAGE libraries. The SAGE Genie at the National Cancer Institute Cancer Genome Anatomy Project contains six SAGE libraries from lung tissues—three normal and three cancer-related libraries (11). Two of the normal libraries were from cell lines and two of the cancer libraries were from microdissected samples of lung adenocarcinoma. That some of the samples were not from primary tissues whereas others were prepared in bulk tissue form implies a lack of sample homogeneity. For the initial screening of the differentially expressed genes, we nevertheless pooled all six libraries into two groups: normal and cancer. The total number of tag counts in each library was normalized to 1 million and the average value in each group was used for a statistical test. Fisher's exact test was used to identify the differentially expressed tags. Tag-to-gene assignment is not trivial when interpreting the SAGE result. EST-based method used in the SAGEmap suffers from sequencing errors, alternative polyA site usage, and lack of proper annotations (12). We used the ECgene mRNA models for reliable tag-to-gene assignment. ECgene assembly procedure solves most of the problems associated with the EST-based methods described above (10).

EST analysis. We used the EST clusters from ECgene version 1.2 based on National Center for Biotechnology Information (NCBI) human genome 35. Approximately 7 million human EST sequences from more than 8,600

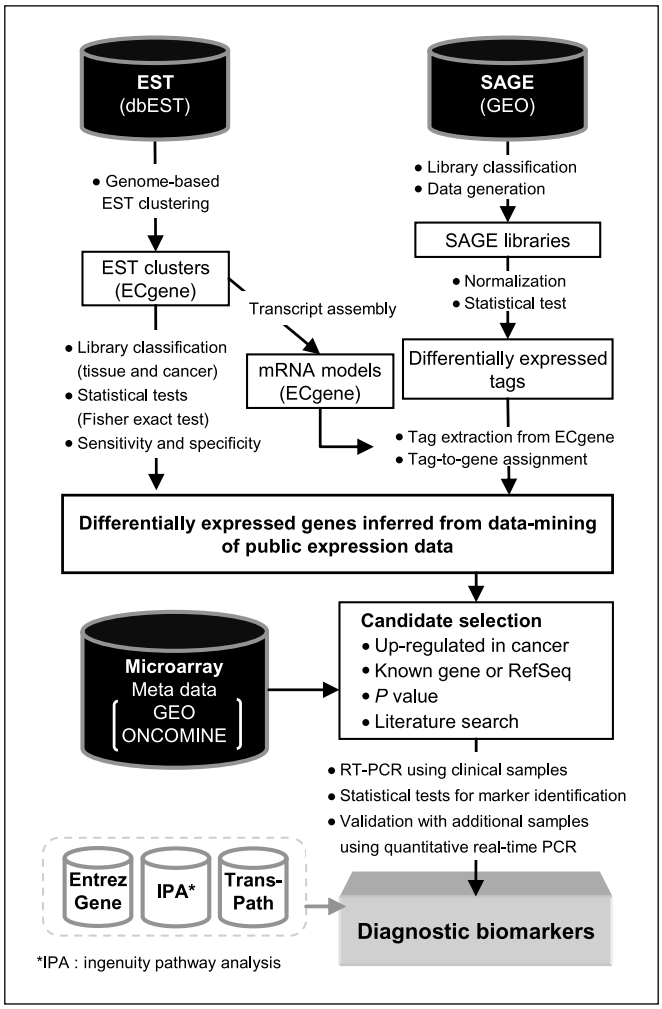

Figure 1. Schematic diagram of the overall procedure.

cDNA libraries are currently available in the GenBank. The total number of lung ESTs from 304 cDNA libraries is 401,672 as of August 2006. The diverse source of samples is one of the main advantages for meta-analysis of EST data.

Each EST cluster was tested for its differential expression in the lung tissue and in the cancerous tissue. Tissue specificity was tested separately from cancer specificity using Fisher's exact test. This approach has the advantage of retaining clusters with limited numbers of lung cancer ESTs that actually has a number of ESTs from normal lung tissue and other types of cancer. Gene expression is often deregulated with tumor progression. It would be reasonable to decide the cancer specificity regardless of their tissue origin.

Sample description. The primary lung tumor tissues were obtained from non–small-cell lung cancer patients who had undergone curative surgery. Our study was approved by the institutional review board, and the written informed consents were obtained from all patients (Institutional Review Board no. 2004-10-18). Lung cancer tissues were obtained from nonnecrotic tumor area. In each case, the normal parenchyma of the same lobe not continuous with the tumor was selected as "pathologically normal" tissue. All the specimens were soaked in liquid nitrogen immediately after resection and stored at −70°C.

Our clinical samples initially consisted of 11 adenocarcinoma, 11 squamous cell carcinoma, and 10 lung tissues from benign lung diseases. The detailed characteristics of patients and the pathology are provided in Table 1. Paired samples were obtained from primary lung cancer and the

adjacent normal tissues for validation tests. Semiquantitative RT-PCR experiments were carried out for initial screening of biomarkers from 20 candidate genes using these 22 paired samples. The 10 lung tissues from benign lung diseases were used as the control for noncancerous disease states. In the final validation step, the real-time quantitative RT-PCR was done for 7 biomarkers on 36 paired samples, which included 14 additional paired samples (7 adenocarcinoma and 7 squamous cell carcinoma) and the 22 original paired samples.

RNA extraction. The frozen tissues were microdissected and lightly stained with hematoxylin to identify the portion consisting of >90% cancer cells. The total RNA was extracted with an RNeasy kit (Qiagen). The RNA quality was assessed with the use of a Nanodrop spectrometer (Nanodrop Technologies) and by electrophoresis on 1% agarose gel that contained 0.6 mol/L formaldehyde and ethidium bromide.

RT-PCR experiment. The total RNA was isolated from 20-μm-thick cryostat sections. Reverse transcription was carried out with 1 μg of RNA,

Table 1. Clinical characteristics of patients with lung disease

No.	Sex	Age (y)	Diagnosis	Operations	pStage	Smoking
Benign lung disease						
1	F	53	Bullous disease	Wedge resection		−
2	M	18	Pneumothorax	Wedge resection		−
3	M	44	Behçet's disease	Lobectomy		+
4	F	50	NTM	Lobectomy		−
5	F	18	Endotracheal tuberculosis	Lobectomy		−
6	M	61	NTM	Lobectomy		+
7	M	37	Aspergilloma	Lobectomy		+
8	M	41	Bronchiectasia	Lobectomy		−
9	F	50	Tuberculosis	Pneumonectomy		−
10	M	25	CCAM	Lobectomy		+
Non–small-cell lung cancer						
1	M	59	Adenocarcinoma	Lobectomy	1B	−
2	M	67	Adenocarcinoma	Lobectomy	1B	+
3	M	51	Adenocarcinoma	Lobectomy	1B	+
4	M	73	Adenocarcinoma	Lobectomy	1A	−
5	M	55	Adenocarcinoma	Lobectomy	1A	+
6	M	63	Adenocarcinoma	Bilobectomy	1B	+
7	F	63	Adenocarcinoma	Lobectomy	1B	−
8	F	61	Adenocarcinoma	Lobectomy	1B	−
9	F	64	Adenocarcinoma	Lobectomy	1A	−
10	F	59	Adenocarcinoma	Lobectomy	1A	−
11	M	52	Adenocarcinoma	Lobectomy	1B	+
12*	M	72	Adenocarcinoma	Lobectomy	1B	+
13*	F	43	Adenocarcinoma	Lobectomy	1B	−
14*	F	51	Adenocarcinoma	Lobectomy	1A	−
15*	M	74	Adenocarcinoma	Lobectomy	1B	+
16*	M	58	Adenocarcinoma	Lobectomy	2A	+
17*	M	71	Adenocarcinoma	Lobectomy	1B	+
18*	F	76	Adenocarcinoma	Lobectomy	1B	−
19	M	71	Squamous cell carcinoma	Lobectomy	1B	+
20	M	52	Squamous cell carcinoma	Bilobectomy	3B	+
21	M	71	Squamous cell carcinoma	Lobectomy	1B	+
22	M	74	Squamous cell carcinoma	Lobectomy	3B	+
23	M	68	Squamous cell carcinoma	Lobectomy	1B	+
24	M	51	Squamous cell carcinoma	Sleeve lobectomy	2A	+
25	M	62	Squamous cell carcinoma	Pneumonectomy	3B	+
26	M	55	Squamous cell carcinoma	Lobectomy	1B	−
27	M	66	Squamous cell carcinoma	Bilobectomy	2B	+
28	M	80	Squamous cell carcinoma	Lobectomy	1B	+
29	M	63	Squamous cell carcinoma	Pneumonectomy	3B	+
30*	M	70	Squamous cell carcinoma	Lobectomy	1B	+
31*	M	58	Squamous cell carcinoma	Lobectomy	1B	+
32*	M	60	Squamous cell carcinoma	Lobectomy	1B	+
33*	M	56	Squamous cell carcinoma	Lobectomy	1B	+
34*	F	38	Squamous cell carcinoma	Lobectomy	1B	−
35*	M	65	Squamous cell carcinoma	Pneumonectomy	2B	+
36*	M	69	Squamous cell carcinoma	Lobectomy	3A	+

Abbreviations: NTM, nontuberculous mycobacteria; CCAM, congenital cystic adenomatoid malformation; No., numbers correspond to the patient ID.
*Samples which were used only in quantitative real-time RT-PCR.

50 μmol/L oligodT (20), 250 μmol/L deoxynucleotide triphosphate (dNTP), 10 units of Moloney murine leukemia virus reverse transcriptase III (Invitrogen), and 5× reverse transcriptase buffer in a total volume of 50 μL. The PCR reactions were done using 1 μL (20 ng) of each cDNA, 250 nmol/L dNTP, 10 pmol/L of each primer, and 2 units of i-StarTag polymerase (Intron Biotechnology, Inc.). The sequences of oligonucleotide primers used in the experiment are listed in Supplementary Table S1. The results were expressed as the ratio of the relative levels.

Quantitative real-time RT-PCR experiment. The cDNA used for quantitative real-time RT-PCR was prepared in the same manner as in the semiquantitative RT-PCR experiment. We used the FAM dye–labeled TaqMan MGB probes (Applied Biosystems). Each probe was designed to be specific to the seven final candidate genes (see above). In addition, TATA box binding protein–specific probe was used as internal control. The PCR reaction mixture consists of the reverse transcription product, TaqMan 2× Universal PCR Master Mix, and the appropriate 20× TaqMan Gene expression assay mix containing primers and probe for the gene of interest. Cycle variables for the PCR reaction were 50°C for 2 min (UNG activation) and then 95°C for 10 min, followed by 40 cycles of a denaturing step at 95°C for 15 s and an annealing/extension step at 60°C for 60 s. All reactions were run in triplicates. The relative expression values of each gene to internal control gene were analyzed using the equation 2^{-dC_T}, where $dC_T = (C_{T\text{target gene}} - C_{T\text{internal control gene}})$ (ref. 13).

Statistical analyses for biomarker evaluation. We measured the DNA band intensity using the BioRAD densitometer (Bio-Rad). The ANOVA statistical tests with the Tukey-Kramer multiple comparison method were done to find the differentially expressed genes between the benign disease tissues and cancer tissues. The paired *t* test was done to identify the differentially expressed genes between the pathologic normal and cancer tissues from each of the patients. Differences were considered significant at $P < 0.05$.

We also carried out the feature selection procedures that are frequently used to identify important features in classification problems. All samples were classified into the normal and cancer classes. The gene expression values of 20 genes were used as an input data. We applied three feature selection methods to identify the most important genes for classifying normal and tumor samples. The support vector machine classifier, χ^2 test statistics, and gain ratio methods were used as implemented in the WEKA package (14). Parameters were set as the default values in the WEKA and the 5-fold cross-validation was done with the constraint seed of 1. The search method was set as the ranker.

Results and Discussion

Identification of Candidate Genes from Public Gene Expression Data

The public SAGE and EST data were analyzed according to the bioinformatics pipeline in Fig. 1 to find the candidate genes for subsequent statistical analyses. The details were given in Materials and Methods. Statistical tests produced several hundreds of candidates that were differentially expressed in lung cancers at $P = 0.01$. Because the sample amount of the clinical specimen was limited, we selected 20 genes for experimental validation, taking the characteristics of the data and gene properties into consideration.

The SAGE data yielded the following 10 genes: *MUC5AC, TFF3, PLUNC, CYBA, CGI-38, GBA, S100P, s-TIM, s-C20orf85*, and *CYP24A1* (details given in Supplementary Table S2). We used the following criteria for gene selection: (*a*) the availability of full-length clones; (*b*) no ambiguity in the tag-to-gene assignment; (*c*) the *P* values and the real tag counts in six libraries; (*d*) the gene properties; and (*e*) the literature survey.

Similarly, another set of 10 genes resulted from the EST analyses: *ALDH3A1, TRIM16, AKR1B10, T, LOC147166, FOXA2, SCTR, DRD2, CBLC*, and *GLP2R* (Supplementary Table S3). We used the criteria of (*a*) the number of ESTs and cDNA libraries, (*b*) multiplicity of exons, (*c*) the percentage of lung and cancer ESTs and libraries, and (*d*) the gene properties. In contrast to the SAGE data, the vast number of cDNA libraries makes it possible to use the tissue and cancer specificities as filtering criteria.

We specifically looked for genes that were up-regulated in the cancer samples because most of the known biomarkers for cancer diagnostics are the overexpressed ones (15). The reliability of gene annotation was also taken into account. To identify markers with testable biological functions, the unknown genes and immune-related genes were excluded in spite of their potential as good biomarkers. Full lists of the differentially expressed genes from SAGE and EST data are available in Supplementary Tables S7 and S8.

Comparing the candidate gene lists from SAGE and EST reveals that the overlap is not significant. This is frequently seen when the number of available libraries for one of the two data set is small. Only six libraries are in public for lung SAGE. A close examination of SAGE candidate genes shows that four genes (*PLUNC, CGI-38, s-C20orf85*, and *GBA*) would have been in the EST candidates without application of the specificity criteria.

Comparison with microarray data. Results from microarray experiments are the most abundant form of gene expression data. The expression level of the 20 candidate genes was compared with the public microarray data. We simply used the Gene Expression Omnibus (GEO) database from the NCBI (16) and the ONCOMINE database developed by Chinnaiyan's group (17).

The GEO serves as a public repository for a wide variety of high-throughput gene expression data from microarray, SAGE, and proteomic methods. We looked for the data sets whose characteristics are similar to our study design, comparing the normal and tumor tissues from lung cancer patients. The GEO search resulted in a data set (GDS1312) revealing differential expression between paired samples from 10 squamous cell carcinoma patients (18). The data set showed that 3 of the 10 SAGE candidates (*CYP24A1, S100P*, and *PLUNC*) and 4 of the 10 EST candidates (*CBLC, ALDH3A1, AKR1B10*, and *TRIM16*) were overexpressed in the tumor samples. However, several genes (*CGI-38, CYBA, TFF3, GBA*, and *FOXA2*) were down-regulated in the tumor samples contrary to our prediction.

The ONCOMINE provides a comprehensive interpretation of published microarray experiments including three pioneering works on lung cancer (19–21). Beer et al. compared nonneoplastic lung tissues with lung adenocarcinoma (10 nonneoplastic versus 86 adenocarcinoma tissues). Bhattacharjee et al. compared 17 normal lung tissues with 139 lung adenocarcinoma samples. Analysis of the data showed that three genes (*S100P, s-TIM*, and *TFF3*) were overexpressed in lung tumor samples compared with normal samples. *CYBA* was underexpressed in lung cancer samples in ONCOMINE database as well.

Comparison with the microarray data indicates that our prediction based on SAGE and EST data agrees to a significant extent with the microarray data. However, there exists substantial difference between three types of high-throughput expression data. This implies that experimental validation using well-defined clinical samples is an essential step for definitively identifying biomarkers.

Experimental Validation of the Candidate Biomarker Genes

The validation procedure consists of two steps. Initially, semiquantitative RT-PCR experiments were done for the 20 candidate

genes using 22 paired samples and 10 inflammatory tissues. Seven genes were selected from extensive statistical analyses of RT-PCR results. Subsequently, quantitative real-time RT-PCR experiments were carried out for the seven genes using 36 paired samples, which included the original 22 paired samples and additional 14 paired samples.

RT-PCR results. We carried out semiquantitative RT-PCR experiments for the 20 candidate genes selected from the SAGE and EST analyses. Among the 10 genes from EST analysis, 6 genes (*T, FOXA2, SCTR, GLP2R, TRIM16*, and *DRD2*) were immediately excluded from further validation efforts because these genes were not detected in any of the lung tissue. This may reflect the fact that many EST libraries were normalized or subtracted to detect even the genes with extremely low level expression. Most ESTs for two of the genes (*SCTR* and *GLP2R*) in fact are from normalized libraries. The reason for the discrepancy for other genes is not clear. We finally evaluated 14 genes in 10 inflammatory tissues and 22 paired cancer tissues by RT-PCR (Fig. 2).

The results of the RT-PCR experiment were quantified and appropriate statistical tests were done to identify the differentially expressed genes as described in Materials and Methods. Table 2 shows the number of samples in which each gene showed a positive expression (for details, see Supplementary Tables S4 and S5). From the three types of samples used in this study—benign disease tissues, pathologic normal tissues, and lung cancer tissues—we made two separate statistical comparisons. The first was to compare the lung cancer tissues with the benign lung tissues. Using benign lung tissues from other patients as a control ensures that the samples were genuinely normal notwithstanding that the differential expression might be due to the genetic differences among individuals. The ANOVA statistical comparison between the tumor tissues and the benign disease tissues identified four genes as overexpressed in tumor tissues: *CBLC* ($P = 0.007$), *S100P* ($P = 0.012$), *CYP24A1* ($P = 0.022$), and *ALDH3A1* ($P = 0.024$). Detailed histologic examinations indicated that the first three genes (*CBLC, S100P*, and *CYP24A1*) were overexpressed in adenocarcinoma tissue and *ALDH3A1* was overexpressed in squamous cell carcinoma tissue. Second, we compared lung cancer tissues with the pathologic normal tissues from the same patient. The genetic background was identical although the pathologic normal tissues might be different from the genuine normal condition. The paired *t* test was done to identify seven genes that were overexpressed in tumor tissues compared with the paired pathologic normal tissues: *CBLC* ($P < 0.001$), *S100P* ($P = 0.005$), *CYP24A1* ($P = 0.002$), *ALDH3A1* ($P = 0.001$), *PLUNC* ($P = 0.003$), *AKR1B10* ($P = 0.012$), and *LOC147166* ($P = 0.012$). Importantly, four genes obtained from the first comparison showed differential expression in this paired comparison as well, and three additional genes were identified.

We also applied the feature selection methods that are frequently used to select discriminatory features for classification problems in the field of the machine learning. Samples were divided into two classes, and three methods were applied to select genes that effectively discriminates the normal and tumor samples. Details are given in Materials and Methods. Table 3 shows the six highest ranking genes from each method of 14 genes from the semi-quantitative RT-PCR results. Top three genes were invariant to the choice of algorithms, which indicated that *CBLC, ALDH3A1*, and *CYP24A1* are potentially excellent biomarkers. *PLUNC* and *LOC147166* ranked between the fourth and the sixth candidates in all three methods, suggestive of their potential.

Quantitative real-time RT-PCR results. Quantitative real-time RT-PCR experiments were done for seven genes (*CBLC, S100P, CYP24A1, ALDH3A1, PLUNC, AKR1B10*, and *LOC147166*) that showed significant differences in RT-PCR experiments. Patient samples include the initial 22 and 14 additional paired sets. Details of clinical characteristics are provided in Table 1, and the experimental result is given in Supplementary Table S6.

The results of the experiment according to the paired *t* test indicate that five genes—*CBLC* ($P = 0.002$), *S100P* ($P = 0.031$), *CYP24A1* ($P = 0.027$), *AKR1B10* ($P = 0.035$), and *LOC147166* ($P = 0.007$)—were significantly overexpressed in tumor samples. In terms of the disease type, *CBLC* was significant in both tumor tissues. *S100P* was overexpressed in adenocarcinoma tissues, whereas *ALDH3A1, AKR1B10*, and *LOC147166* were overexpressed in squamous cell carcinoma tissues.

Figure 3 shows the box plot analysis of the real-time RT-PCR results. For each gene, median fold change and distribution were analyzed for all samples together and separately for the two different cancer types. Consistent with the *t* test, *CBLC* stands out as the most probable biomarker for both types of lung cancers. *CYP24A1* also seems to be a viable general biomarker for lung cancers. For adenocarcinoma, *PLUNC* and *S100P* also seem to be potential biomarkers, whereas *AKR1B10* and *ALDH3A1* seem to be likely candidates for squamous cell carcinoma. Importantly, when quantitated separately, the 22 original samples from which the

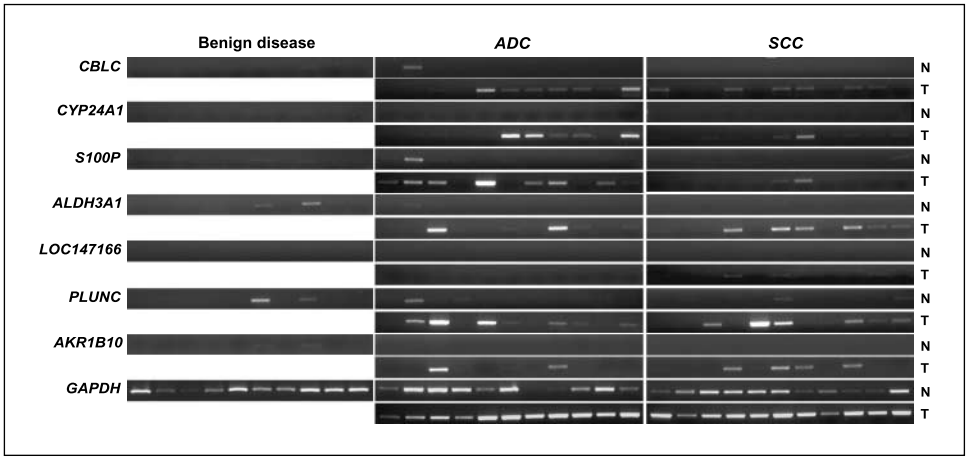

Figure 2. The semiquantitative RT-PCR results of the overexpressed genes. Seven genes are significantly overexpressed in tumor tissues as compared with the paired pathologic normal tissues in non–small-cell lung cancer patients. Four genes among these are also significantly overexpressed in tumor tissues compared with the benign disease tissues. *T,* tumor; *N,* pathologic normal; *ADC,* adenocarcinoma; *SCC,* squamous cell carcinoma.

Table 2. Number of samples positive for the gene expression in the semiquantitative RT-PCR experiment

Genes	Gene-expressed sample number				
	Benign disease (total 10)	Adenocarcinoma (total 11)		Squamous cell carcinoma (total 11)	
		Normal	Tumor	Normal	Tumor
CBLC	1	1	7	0	8
CYP24A1	2	0	9	0	8
S100P	3	5	10	5	7
ALDH3A1	4	2	8	1	10
LOC147166	0	0	3	0	4
PLUNC	2	2	7	2	8
AKR1B10	2	2	4	0	4
CGI-38	9	0	0	0	0
MUC5AC	2	1	3	0	1
s-C20orf85	4	6	9	4	5
TFF3	6	3	3	5	3
CYBA	10	11	11	11	11
GBA	10	11	11	11	11
s-TIM	10	11	11	11	11

NOTE: The numbers in the table indicate the number of total samples where the candidate gene is expressed in the semiquantitative RT-PCR experiment.

7 candidate genes were derived and the new 14 samples which could be considered as the validation set showed little difference in terms of fold change.

The result of analyzing the real-time PCR data using the feature selection methods is also included in Table 3. Again, *CBLC* is the most significant marker, followed by *CYP24A1* and *AKR1B10*. *ALDH3A1* scored poorly because its expression was suppressed in several newly tested samples.

Critical evaluation of biomarker genes. Inspecting the details of gene expression shown in Table 2 and Fig. 3 provides deeper insights on the biomarker evaluation. The *CBLC* gene, the most significant marker according to the statistical tests, was expressed in just two normal samples (1 in 10 benign lung disease tissues, 1 in 11 pathologic normal tissues from the adenocarcinoma patients, and none from the squamous cell carcinoma patient tissues). In contrast, most cancer tissues showed a positive expression for this gene (7 of 11 in the adenocarcinoma tissues and 8 of 11 in the squamous cell carcinoma tissues). In 35 of 36 cases, real-time RT-PCR analysis showed elevated expression in tumor samples, which strongly indicates that *CBLC* is a highly specific and sensitive cancer biomarker.

CYP24A1 also showed similar degree of specificity and sensitivity. Interestingly, two inflammatory tissues from benign lung disease cases showed positive expression although *CYP24A1* was not expressed in any pathologically normal tissues from 22 cancer patients. This indicates that *CYP24A1* could be an excellent marker for distinguishing normal and tumor tissues in paired samples but its expression can be induced in other types of lung diseases. Although the result from the real-time RT-PCR experiment, with 4 of 33 cases showing suppressed expression in tumor

Table 3. Important genes obtained from various feature selection methods

Rank	Semiquantitative RT-PCR			Quantitative real-time RT-PCR		
	SVM	χ^2	Gain ratio	SVM	χ^2	Gain ratio
1	*CBLC*	ALDH3A1	ALDH3A1	CBLC	CBLC	CBLC
2	*ALDH3A1*	CYP24A1	CYP24A1	LOC147166	CYP24A1	CYP24A1
3	*CYP24A1*	CBLC	CBLC	S100P	AKR1B10	AKR1B10
4	*PLUNC*	S100P	LOC147166	CYP24A1	ALDH3A1	ALDH3A1
5	*CGI-38*	PLUNC	S100P	AKR1B10	PLUNC	PLUNC
6	*LOC147166*	LOC147166	PLUNC	ALDH3A1	LOC147166	LOC147166

NOTE: Three machine-learning approaches were applied with 5-fold cross-validation. The six highest ranking genes were listed. Gene expression values from the semiquantitative RT-PCR were obtained for 14 genes, whereas the real-time RT-PCR experiment data are available for only the 7 validated genes.

Abbreviation: SVM, support vector machine.

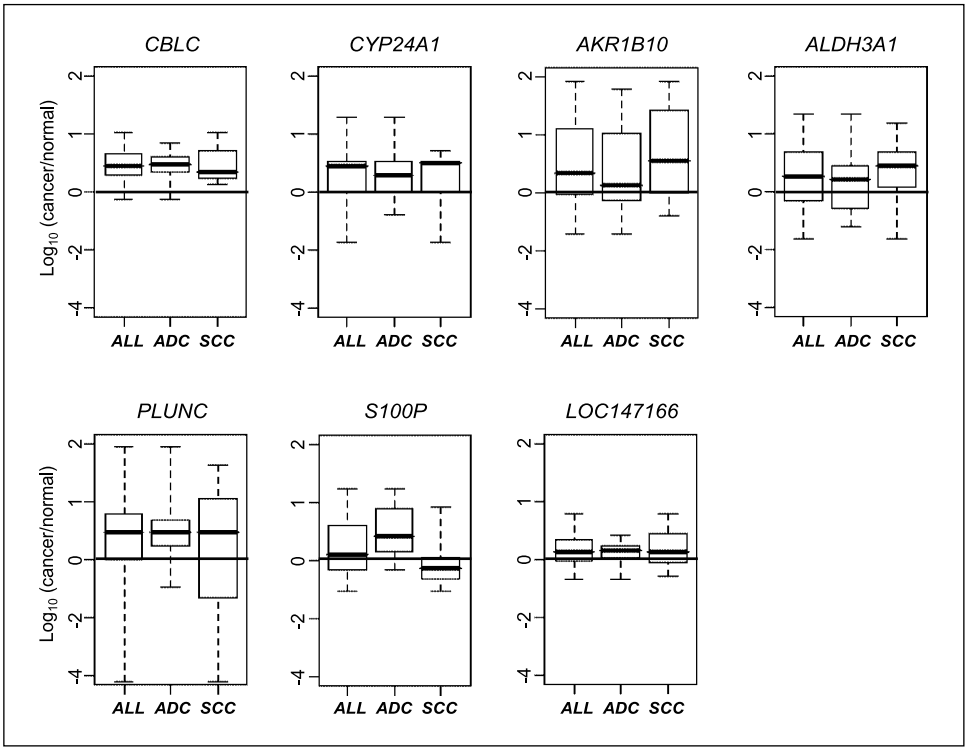

Figure 3. Box plot of quantitative real-time RT-PCR results. Fold changes among 36 paired samples (*ALL*), 18 adenocarcinoma samples (*ADC*), and 18 squamous cell carcinoma samples (*SCC*) are plotted in logarithmic scale. Score of zero in this scale implies no change, whereas score of 1 indicates a 10-fold change. The box area contains 50% of the data samples, and 99.3% of the samples are within the upper and lower boundary markers. The best biomarkers would show the highest score and smallest-sized box (e.g., *CBLC*).

samples, is not as impressive as the semiquantitative RT-PCR data, *CYP24A1* still seems to be a promising biomarker for lung cancers.

Although *S100P* showed a good statistical correlation, it does not seem to be a good marker because about half of the normal samples showed its expression. Furthermore, *S100P* scored poorly in feature selection methods. The box plot in Fig. 3 nevertheless indicates that *S100P* might be a good biomarker for adenocarcinoma subtype of lung cancer. *ALDH3A1* seemed to perform much better for its overall ability to distinguish cancer samples. The numbers were even more impressive for the squamous cell carcinoma patients (1 versus 10). Real-time RT-PCR data are consistent with these observations.

Among three additional genes obtained from the paired sample test, *AKR1B10* and *PLUNC* had a small tendency to be preferentially expressed in cancer samples than over normal samples, but their merit as biomarkers was not obvious in the RT-PCR result. However, real-time data strongly support *AKR1B10* as a promising candidate. In fact, all 14 additional samples showed increased expression in cancer tissues, whereas the original 22 samples had a mixed tendency. Similarly, *PLUNC* seems to have a fair potential as a biomarker for adenocarcinoma, with 16 of 18 samples showing elevated expression in cancer tissues.

In summary, we propose that the four genes (*CBLC, CYP24A, AKR1B10*, and *ALDH3A1*) that showed significant differences in both statistical tests and the RT-PCR validations are potential biomarkers for non–small-cell lung cancer patients. Two genes (*CBLC* and *CYP24A1*) are particularly promising. With respect to the histopathologic aspects, these genes were expressed in both adenocarcinoma and squamous cell carcinoma, indicating that they are not cancer type–specific markers.

Biological properties of the candidate genes. Biomarker discovery does not necessarily require understanding the biological function and regulatory mechanism of the candidate genes. However, molecular understanding of the biological function could still be worthwhile in that overexpression of these genes may be mechanistically linked to carcinogenesis. We therefore surveyed the literature and the knowledge databases such as Entrez Gene (22), Ingenuity Pathway Analysis (23), and TransPath Professional 7.3 (24) on the two most promising genes.

CBLC is a member of the Cbl family of multidomain signaling proteins with a tyrosine kinase binding domain and a RING finger domain, the latter of which interacts with the E2 ubiquitin conjugating enzymes of the ubiquitin pathway (25). Thus, the Cbl family gene products function as ubiquitin ligases toward activated protein tyrosine kinases such as *Src* (26) and *Lck* (27). *CBLC* is also known to bind to proteins with the Src homology-3 domain as well. It is recruited to the epidermal growth factor (EGF) receptor (EGFR) on EGF stimulation and increases ubiquitination of EGFR, thereby down-regulating EGFR signaling (28, 29). Mutations in the EGFR gene have been reported in non–small-cell lung cancer patients, especially in patients with adenocarcinoma, women, nonsmokers, and East Asians (30).

CYP24A1 is a member of the cytochrome *P*450 superfamily of enzymes involved in drug metabolism and synthesis of cholesterol, steroids, and other lipids. This mitochondrial protein initiates the degradation of 1,25-dihydroxyvitamin D3, the physiologically active form of vitamin D3. Albertson et al. (31) reported that gene copy number and expression are increased in breast cancer, and Mimori et al. (32) showed that its overexpression is linked to a poor prognosis for esophageal cancer. At the time of writing, Parise et al. (33) reported up-regulation of *CYP24A1* in non–small-cell lung cancer. The promoter region of *CYP24A1* contains two vitamin D response elements and an Ets-1 binding site (34). Its gene regulation is a complicated process involving vitamin D response, Ets-1, retinoid X receptor α, and various

mitogen-activated protein kinases such as extracellular signal–regulated kinase (ERK)-1 and ERK5. A number of studies have reported that Ets-1 is a proto-oncogene in various types of cancer.

Conclusion

In this study, we identified candidate biomarkers for lung cancers through bioinformatics analysis of the public SAGE and EST data and validated their potential using clinical specimens. *CBLC* and *CYP24A1* are two particularly promising biomarkers for non–small-cell lung cancer. Other genes (*ALDH3A1, AKR1B10*, and *LOC147166*) seem to have fair potential as well.

It is interesting to note the origin of biomarker genes. Two genes (*CYP24A1* and *S100P*) were derived from SAGE data and others (*CBLC, ALDH3A1, AKR1B10*, and *LOC147166*) were from the EST data. This implies different ranges of coverage for the two data sets and the benefits of using both types of data. Our study also shows that candidates from meta-analysis of the public expression data should be carefully tested through validation using clinical samples.

One of the major strengths of our study is the use of multiple clinical samples. Strong statistical support was thus possible although additional clinical samples should be used for further validation down the road. In addition, we tested only 20 genes in this study with several hundreds of candidates remaining to be examined. Biochemical studies for promising biomarkers are necessary as well to examine the potential of the candidate genes as drug targets. As additional expression data become available, it would be also be interesting to see if combinations of several differentially regulated genes could function with more sensitivity and specificity in the diagnosis and prognosis of lung cancers.

Acknowledgments

Received 1/3/2007; revised 5/6/2007; accepted 5/25/2007.

Grant support: Korea Science and Engineering Foundation through the bioinformatics research program (grant M10529000016-05N2900-01610), the NCRC program through the Center for Cell Signaling and Drug Discovery Research at Ewha Womans University (grant R15-2006-020), and Samsung Biomedical Research Institute through the SBRI program (grant C-A5-205-2).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

1. van Zandwijk N. New methods for early diagnosis of lung cancer. Lung Cancer 2002;38:S9–11.
2. Peto R, Lopez AD, Boreham J, Thun M, Heath C, Jr., Doll R. Mortality from smoking worldwide. Br Med Bull 1996;52:12–21.
3. Sienel W, Dango S, Ehrhardt P, Eggeling S, Kirschbaum A, Passlick B. The future in diagnosis and staging of lung cancer. Molecular techniques. Respiration 2006;73:575–80.
4. Petty RD, Nicolson MC, Kerr KM, Collie-Duguid E, Murray GI. Gene expression profiling in non-small cell lung cancer: from molecular mechanisms to clinical application. Clin Cancer Res 2004;10:3237–48.
5. Velculescu VE, Madden SL, Zhang L, et al. Analysis of human transcriptomes. Nat Genet 1999;23:387–8.
6. Hibi K, Liu Q, Beaudry GA, et al. Serial analysis of gene expression in non-small cell lung cancer. Cancer Res 1998;58:5690–4.
7. Nacht M, Dracheva T, Gao Y, et al. Molecular characteristics of non-small cell lung cancer. Proc Natl Acad Sci U S A 2001;98:15203–8.
8. Hess JL. The Cancer Genome Anatomy Project: power tools for cancer biologists. Cancer Invest 2003;21:325–6.
9. Brentani H, Caballero OL, Camargo AA, et al. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc Natl Acad Sci U S A 2003;100:13418–23.
10. Kim N, Shin S, Lee S. ECgene: genome-based EST clustering and gene modeling for alternative splicing. Genome Res 2005;15:566–76.
11. Boon K, Osorio EC, Greenhut SF, et al. An anatomy of normal and malignant gene expression. Proc Natl Acad Sci U S A 2002;99:11287–92.
12. Lash AE, Tolstoshev CM, Wagner L, et al. SAGEmap: a public gene expression resource. Genome Res 2000;10:1051–60.
13. Shih JY, Tsai MF, Chang TH, et al. Transcription repressor slug promotes carcinoma invasion and predicts outcome of patients with lung adenocarcinoma. Clin Cancer Res 2005;11:8070–8.
14. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
15. Dalton WS, Friend SH. Cancer biomarkers-an invitation to the table. Science 2006;312:1165–8.
16. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
17. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004;6:1–6.
18. Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics 2005;21:4205–8.
19. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24.
20. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 2001;98:13790–5.
21. Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 2001;98:13784–9.
22. Rioux PA, Gilbert WA, Littlejohn TG. A portable search engine and browser for the Entrez database. J Comput Biol 1994;1:293–5.
23. Kim Y, Ton TV, DeAngelo AB, et al. Major carcinogenic pathways identified by gene expression analysis of peritoneal mesotheliomas following chemical treatment in F344 rats. Toxicol Appl Pharmacol 2006;214:144–51.
24. Krull M, Pistor S, Voss N, et al. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. Nucleic Acids Res 2006;34:D546–51.
25. Thien CB, Langdon WY. Cbl: many adaptations to regulate protein tyrosine kinases. Nat Rev Mol Cell Biol 2001;2:294–307.
26. Kim M, Tezuka T, Tanaka K, Yamamoto T. Cbl-c suppresses v-Src-induced transformation through ubiquitin-dependent protein degradation. Oncogene 2004;23:1645–55.
27. Rao N, Miyake S, Reddi AL, et al. Negative regulation of Lck by Cbl ubiquitin ligase. Proc Natl Acad Sci U S A 2002;99:3794–9.
28. Keane MM, Ettenberg SA, Nau MM, et al. cbl-3: a new mammalian cbl family protein. Oncogene 1999;18:3365–75.
29. Courbard JR, Fiore F, Adelaide J, Borg JP, Birnbaum D, Ollendorff V. Interaction between two ubiquitin-protein isopeptide ligases of different classes, CBLC and AIP4/ITCH. J Biol Chem 2002;277:45267–75.
30. Toyooka S, Soh J, Shigematsu H, Aoe M, Date H. The impact and role of EGFR gene mutation on non-small cell lung cancer. Cancer Chemother Pharmacol 2006;58 Suppl 7:25–31.
31. Albertson DG, Ylstra B, Segraves R, et al. Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. Nat Genet 2000;25:144–6.
32. Mimori K, Tanaka Y, Yoshinaga K, et al. Clinical significance of the overexpression of the candidate oncogene CYP24 in esophageal cancer. Ann Oncol 2004;15:236–41.
33. Parise RA, Egorin MJ, Kanterewicz B, et al. CYP24, the enzyme that catabolizes the antiproliferative agent vitamin D, is increased in lung cancer. Int J Cancer 2006;119:1819–28.
34. Dwivedi PP, Hii CS, Ferrante A, et al. Role of MAP kinases in the 1,25-dihydroxyvitamin D3-induced transactivation of the rat cytochrome P450C24 (CYP24) promoter. Specific functions for ERK1/ERK2 and ERK5. J Biol Chem 2002;277:29643–53.