A Training-Testing Approach to the Molecular Classification of Resected Non-Small Cell Lung Cancer


Vanderbilt-Ingram Cancer Center and Department of Medicine [N. Y., K. Y., T. F. D., S. N., D. P. C.], Department of Preventive Medicine [P. L., Y. S.], Department of Pathology [M. E., A. G., R. J.], Department of Cardiac and Thoracic Surgery [J. R. R.], and Department of Molecular Physiology and Biophysics [S. L., J. H. M.], Vanderbilt University School of Medicine, Nashville, Tennessee 37232-6838; Cardiovascular Surgical Associates, Saint Thomas Hospital, Nashville, Tennessee 37205 (J. C. N.); and Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center, Dallas, Texas 75235 [J. D. M.]

ABSTRACT

Purpose: RNA expression patterns associated with non-small cell lung cancer subclassification have been reported, but there are substantial differences in the key genes and clinical features of these subsets casting doubt on their biological significance.

Experimental Design: In this study, we used a training-testing approach to test the reliability of cDNA microarray-based classifications of resected human non-small cell lung cancers (NSCLCs) analyzed by cDNA microarray.

Results: Groups of genes were identified that were able to differentiate primary tumors from normal lung and lung metastases, as well as identify known histological subgroups of NSCLCs. Groups of genes were identified to discriminate sample clusters. A blinded confirmatory set of tumors was correctly classified by using these patterns. Some histologically diagnosed large cell tumors were clearly classified by using these patterns. These results demonstrate that gene expression profiling can identify molecular classes of resected NSCLCs that correctly classify a blinded test cohort, and correlates with and supplements standard histological evaluation.

INTRODUCTION

Lung cancer represents a challenging clinical problem in most of the developed countries. The number of deaths from lung cancer in the United States is more than the next four most common cancers combined. Despite the best current treatment, the overall 5-year survival after diagnosis is only 10–15%. Improvements in prevention, early detection, prognosis, and therapy have been difficult to achieve. Clinically, lung cancers display a broad range of clinical behaviors ranging from slowly progressing to rapidly fatal, they can be highly metastatic or only locally invasive, and they may display responsiveness or resistance to therapy (1); the molecular basis of these variations in behavior is completely unknown.

The classification of lung cancers has traditionally been based primarily on light microscopic morphological findings. According to the current histological lung cancer classification proposed by the WHO in 1981, lung cancers can be divided into two broad groups, small cell lung cancer, accounting for 20–25% of bronchogenic carcinomas, and NSCLC,2 accounting for almost all of the remaining cases. NSCLC has three major subgroups: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma (2). Even within the subgroup of NSCLC there is a great degree of heterogeneity in behavior, and the histological subclassifications for NSCLCs have no predictive use and all are treated identically despite decades of research.

It is clear that each tumor has unique genetic differences, and it is hypothesized that these differences determine its biological behavior. A large effort has been made by many laboratories to study many individual candidate genetic abnormalities in an attempt to develop molecular markers for lung cancer classification and prognosis, but after hundreds of such studies, none of these single markers are of any real clinical utility. Even today, all NSCLCs are usually treated identically, stage for stage, and no molecular marker is used for routine therapeutic decisions. Thus, it is becoming clear that complex biological behaviors of tumors will only be explainable by complex patterns of multiple markers.

Microarray technology has enabled expression analysis of thousands of genes at one time, allowing insight into complex gene expression patterns and perturbations (3). To date, microarray technology has been successfully applied to a wide

2 The abbreviations used are: NSCLC, non-small cell lung cancer; WFCCM, Weighted Flexible Compound Covariate Method; SAM, Significance Analysis of Microarrays; ACTN4, α-actinin-4.
variety of malignant diseases, such as leukemia, lymphoma, colon cancer, melanoma, ovarian cancer, breast cancer, hepatocellular carcinoma, and prostate cancer (4–10). These studies have succeeded in identifying dozens of crucial genes that are up- or down-regulated in certain types of malignant cells or tissues. In lung cancer, several groups have reported microarray-based subclassifications of lung adenocarcinomas, but these studies differ from each other in significant ways, and none of these studies have tested their patterns by using blinded sets of tumors (11–13).

To begin to explore lung cancer molecular profiles that predict biological behavior, e.g., histological subtype or survival of lung cancer patients, we applied cDNA microarray technology to the study of a set of freshly resected human lung cancers and used multiple statistical methods to correlate differentially expressed genes with histology and clinical outcome. We then successfully tested this classification pattern with an independent cohort of tumors.

MATERIALS AND METHODS

Tissues and Cell Lines. Lung cancer, normal lung, and nonlung tumor tissues in excess of what was necessary for diagnostic purposes were obtained <15 min after removal from the patient and placed in RNAlater (Ambion, Austin, TX) before being snap frozen in liquid nitrogen. Before and independent of the molecular analysis, the lung cancer patients were assigned a Tumor-Node-Metastasis postsurgical stage score according to the current international lung cancer staging system, and then classified histologically into adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and small cell carcinoma using standard WHO criteria (2). The cancers were additionally classified as well, moderately, or poorly differentiated. A pool of RNAs isolated from six different lung cancer cell lines was prepared as a common reference to represent the common histological subtypes of lung tumor. This RNA pool provided us with a large amount of reliable and consistent RNA likely to contain the vast majority of genes expressed in lung cancer and enabled us to compare gene expression patterns across tumor samples. Six lung cancer cell lines of different histological subtypes, H23, H69, H157, H441, H596, and H727, were selected and purchased from American Type Culture Collection. The purified cDNA was dried to the manufacturer’s instructions. The purified cDNA was used as a template with NaOH, the labeled single-stranded cDNA was hydrolyzed using RNaseA, and the labeled cDNA was applied to the slide, and sealed under a coverslip. The slide was placed in a humidified chamber at 65°C for 14–16 h. After hybridization, the slide was washed once with PBS and resuspended in hybridization solution [3× SSC/0.2% SDS/1 μg/μl yeast tRNA/1 μg/μl poly(dA)], heat denatured, and applied to the slide, and sealed under a coverslip. The slide was placed in a humidified chamber at 65°C for 14–16 h. After hybridization, the slide was washed in 2× SSC/0.1% at 55°C for 5 min, 1× SSC at room temperature for 5 min, and 0.1× SSC at room temperature for 5 min. The slide was dried and scanned by a GenePix 4000B scanner (Axon Instruments, Inc., Foster City, CA). The resulting image was analyzed using GenePix Pro software (Axon Instruments, Inc.). To assay for artifacts caused by specific dye combinations, we performed two reciprocal hybridizations on different arrays for each sample (switching the dyes between test sample and reference cDNAs). When extracting the data from the original pictures captured by the scanner, spots with obvious blemishes or spots of which the diameters were <60 μm were flagged out at the initial step and excluded for additional analysis. Also, 96 spots that had not been assigned to any unigene cluster were excluded from this analysis. Normalization was performed based on the premise that the arithmetic mean of the ratios from every spot that performed well should be equal to 1 (14). In brief, nonflagged spots that had a ratio between 0.1 and 10 were selected. The log value for the expression ratio of each selected spot was

\[ \log \frac{\text{test sample}}{\text{reference sample}} \]

The quantity and the quality of RNA were assessed by UV spectrophotometer and electrophoresis on a 1% formaldehyde agarose gel.

Microarray Preparation. 5184 cDNA inserts were PCR amplified from the sequence verified human clones purchased from Research Genetics (Rockville, MD) to represent 4827 unique genes. Aliquots of all of the PCR products were examined by agarose gel electrophoresis. The 681 products that did not amplify or contained multiple bands were not used for data analysis. For arraying, the PCR amplified cDNA inserts were resuspended in 3× SSC and arrayed onto the poly-L lysine coated glass slides by Stanford type microarrayer robot at the Vanderbilt Microarray Shared Resource. Printed slides, VMSR human 5k cDNA microarrays, were postprocessed by UV exposure. The complete gene list and chip production information is available on our web site.4 Just before hybridization, slides were prehybridized in 5× SSC, 0.1% SDS, and 1% BSA (Sigma, St. Louis, MO) for 45 min at 42°C. After washing five times by dipping in MilliQ purified water at room temperature, slides were dipped into isopropanol and air-dried.

Microarray Analysis. Fluorescently labeled cDNAs were made from 50 to 100 μg of sample and reference total RNA through an anchored oligodeoxynucleotidyl acid [5'-T(20)VN-3' (V = any nucleotide except T, n = any nucleotide)] primed reverse transcriptase reaction. The labeling reactions were done in the presence of 100 ng/μl anchored primer, 200 μM each of dATP, dGTP, and dTTP; 120 μM dCTP; 10 mM DTT; 200 units of SuperScript II (Life Technologies, Inc.); and 120 μM Cy3-dCTP or Cy5-dCTP (Amersham Pharmacia, Piscataway, NJ) in a 30-μl solution. After hydrolyzing the RNA template with NaOH, the labeled single-stranded cDNA was purified by QiAquick PCR purification kit (Qiagen) according to the manufacturer’s instructions. The purified cDNA was dried in a SpeedVac and resuspended in hybridization solution [3× SSC/0.2% SDS/1 μg/μl yeast tRNA/1 μg/μl poly(dA)], heat denatured, and applied to the slide, and sealed under a coverslip. The slide was placed in a humidified chamber at 65°C for 14–16 h. After hybridization, the slide was washed in 2× SSC/0.1% at 55°C for 5 min, 1× SSC at room temperature for 5 min, and 0.1× SSC at room temperature for 5 min. The slide was dried and scanned by a GenePix 4000B scanner (Axon Instruments, Inc., Foster City, CA). The resulting image was analyzed using GenePix Pro software (Axon Instruments, Inc.). To assay for artifacts caused by specific dye combinations, we performed two reciprocal hybridizations on different arrays for each sample (switching the dyes between test sample and reference cDNAs). When extracting the data from the original pictures captured by the scanner, spots with obvious blemishes or spots of which the diameters were <60 μm were flagged out at the initial step and excluded for additional analysis. Also, 96 spots that had not been assigned to any unigene cluster were excluded from this analysis. Normalization was performed based on the premise that the arithmetic mean of the ratios from every spot that performed well should be equal to 1 (14). In brief, nonflagged spots that had a ratio between 0.1 and 10 were selected. The log value for the expression ratio of each selected spot was

\[ \log \frac{\text{test sample}}{\text{reference sample}} \]

The quantity and the quality of RNA were assessed by UV spectrophotometer and electrophoresis on a 1% formaldehyde agarose gel.

Microarray Preparation. 5184 cDNA inserts were PCR amplified from the sequence verified human clones purchased from Research Genetics (Rockville, MD) to represent 4827 unique genes. Aliquots of all of the PCR products were examined by agarose gel electrophoresis. The 681 products that did not amplify or contained multiple bands were not used for data analysis. For arraying, the PCR amplified cDNA inserts were resuspended in 3× SSC and arrayed onto the poly-L lysine coated glass slides by Stanford type microarrayer robot at the Vanderbilt Microarray Shared Resource. Printed slides, VMSR human 5k cDNA microarrays, were postprocessed by UV exposure. The complete gene list and chip production information is available on our web site.4 Just before hybridization, slides were prehybridized in 5× SSC, 0.1% SDS, and 1% BSA (Sigma, St. Louis, MO) for 45 min at 42°C. After washing five times by dipping in MilliQ purified water at room temperature, slides were dipped into isopropanol and air-dried.

Microarray Analysis. Fluorescently labeled cDNAs were made from 50 to 100 μg of sample and reference total RNA through an anchored oligodeoxynucleotidyl acid [5'-T(20)VN-3' (V = any nucleotide except T, n = any nucleotide)] primed reverse transcriptase reaction. The labeling reactions were done in the presence of 100 ng/μl anchored primer, 200 μM each of dATP, dGTP, and dTTP; 120 μM dCTP; 10 mM DTT; 200 units of SuperScript II (Life Technologies, Inc.); and 120 μM Cy3-dCTP or Cy5-dCTP (Amersham Pharmacia, Piscataway, NJ) in a 30-μl solution. After hydrolyzing the RNA template with NaOH, the labeled single-stranded cDNA was purified by QiAquick PCR purification kit (Qiagen) according to the manufacturer’s instructions. The purified cDNA was dried in a SpeedVac and resuspended in hybridization solution [3× SSC/0.2% SDS/1 μg/μl yeast tRNA/1 μg/μl poly(dA)], heat denatured, and applied to the slide, and sealed under a coverslip. The slide was placed in a humidified chamber at 65°C for 14–16 h. After hybridization, the slide was washed in 2× SSC/0.1% at 55°C for 5 min, 1× SSC at room temperature for 5 min, and 0.1× SSC at room temperature for 5 min. The slide was dried and scanned by a GenePix 4000B scanner (Axon Instruments, Inc., Foster City, CA). The resulting image was analyzed using GenePix Pro software (Axon Instruments, Inc.). To assay for artifacts caused by specific dye combinations, we performed two reciprocal hybridizations on different arrays for each sample (switching the dyes between test sample and reference cDNAs). When extracting the data from the original pictures captured by the scanner, spots with obvious blemishes or spots of which the diameters were <60 μm were flagged out at the initial step and excluded for additional analysis. Also, 96 spots that had not been assigned to any unigene cluster were excluded from this analysis. Normalization was performed based on the premise that the arithmetic mean of the ratios from every spot that performed well should be equal to 1 (14). In brief, nonflagged spots that had a ratio between 0.1 and 10 were selected. The log value for the expression ratio of each selected spot was

\[ \log \frac{\text{test sample}}{\text{reference sample}} \]
determined. The average of all log values (AvgLog) was calculated. The normalized ratio for each spot was calculated as the normalized ratio = original ratio: $10^{\text{AvgLog}}$. The data from spots of which the fluorescence intensity in each channel was $<1.4$ times the local background were deleted for subsequent analysis. We also excluded data that was only available from one of the reciprocal experiments or where data from reciprocal experiments differed by $>2$-fold. After this data filtering, we took the average of the data from the reciprocal experiments in the logarithmic field. Finally, data from spots that had $>70\%$ of data present across the samples in the training cohort were used for additional statistical analysis. Thus, in the final data table, we had data from 3811 spots that represented 3647 genes. We used every gene on this table for the search for differentially expressed genes among the training sample subtypes. The original data tables and more information regarding with our analysis are available as the supplementary information on our website.3

Statistical Data Analysis. The statistical analyses for the microarray data were focused on the following steps: (a) selecting the important genes that were differentially expressed among the histological groups; (b) using the class prediction model based on the WFCCM (15, 16) to verify if the genes selected in step one have the statistical significant prediction power on the training samples; (c) applying the prediction model generated from step two to a set of blinded samples for examining the prediction power on the blinded samples; and (d) using the agglomerative hierarchical clustering algorithm (17) to investigate the pattern among the statistically significant discriminator genes as well as the biological status.

The selection of important genes was based on SAM (18), Weighted Gene Analysis (10), and the $t$ test, and the cutoff points for each method were 3.7, 3.0, and $P < 0.0001$, respectively. The cutoff points were determined based on the significance as well as the prediction power of each method. The gene was on the final list if it met at least one of these three selection criteria.

The WFCCM (15, 16) was used in the class-prediction model based on the selected genes. This method was designed to combine the most significant genes associated with the biological status from each analysis method, e.g., SAM, Weighted Gene Analysis, Info Score (20), and $t$ test. In other words, the WFCCM is an extension of the compound covariate method, which allows for the consideration of more than one statistical analysis method into the compound covariate, and it reduces the dimensionality of the problem using a new covariate obtained as a weighted sum of the most important predictors. The WFCCM for tumor sample $i$ is defined as $WFCCM(i) = \sum_j [x_j(S_j)^{1/2}] [W_i][x_j]$, where $x_j$ is the log-ratio measured in tissue sample $i$ for gene $j$, $S_j$ is the standardized statistic, e.g., $t$-statistic, for statistical analysis method $j$. $W_i$ is the weight of gene $j$, which is defined as $W_j = [(\sum_k I_{jk}^2/K) (1 - \text{Info Score}_j)]$, where $I_{jk} = 1$, if the gene $j$ was statistically significant in method $k$; and $I_{jk} = 0$, if the gene $j$ is not statistically significant in method $k$.

The class-prediction model was applied to determine whether the patterns of gene expression could be used to classify tissue samples into two classes according to the chosen parameter, e.g., normal tissue versus tumor tissue. We estimated the misclassification rate using leave-one-out cross-validated class prediction method based on the WFCCM. This leave-one-out cross-validated method was processed in four steps. First, WFCCM was applied to calculate the single compound covariate for each tissue sample based on the significant genes. Second, one tissue sample was selected and removed from the data set, and the distance between two tissue classes for the remaining tissue samples was calculated. Third, the removed tissue sample was classified based on the closeness of the distance of two tissue classes. Fourth, steps 2 and 3 were repeated for each tissue sample. To determine whether the accuracy for predicting membership of tissue samples into the given classes (as measured by the number of correct classifications) was better than the accuracy that could be attained for predicting membership into random grouping of the tissue samples, we created 5000 random data sets by permuting class labels among the tissue samples. The cross-validated class prediction was performed on the resulting data sets, and the percentage of permutations that resulted in as few or fewer misclassifications as for the original labeling of samples was reported. If $<0.05$ of the permutations resulted in as few or fewer misclassifications, the accuracy of prediction into the given classes was considered significant.

The prediction of the blinded samples was completed using the method described above. The blinded sample was classified based on the closeness of the distance of two tissue classes, which was determined using the WFCCM.

The agglomerative hierarchical clustering algorithm (17) was applied to investigate the pattern among the statistically significant discriminator genes as well as the biological status using the software of Eisen et al. (21).

Survival was estimated with the Kaplan-Meier method, and differences between groups were compared with the log-rank test.

Sequence Verification of cDNAs. For the differentially expressed genes, the cDNA inserts were verified by DNA sequencing using vector-specific modified M13 primers, forward primer (5′-GTTCCTCCACGTCCACGTTG-3′) or reverse primer (5′-TGACGCGATTAAACCATTCACACAG-3′). Cycle sequencing reactions were performed with fluorescent-labeled nucleotides at the DNA sequencing shared resource in Vanderbilt Ingram Cancer Center. Sequence database searches were performed with Basic Local Alignment Search Tool (BLAST) sequence comparison programs at National Center for Biotechnology Information.5

RESULTS

Initial Data Analysis and Sample Re-Evaluation. We first sought to establish whether molecular profiling of our tumor set could identify genes of which the expressions correlated with known light microscopic histological subgroups of lung cancer. Differentiation of NSCLC from small cell lung cancer or lung primary from metastasis to the lung from other organs is of significant clinical interest and sometimes problematic in practice. The ability of this technique to identify patterns of genes associated with these known histological subgroups

would also serve as a useful proof of principle. For this purpose, we analyzed 26 resected primary lung cancers, 3 normal lung tissues, and 2 metastatic lung tumors as our training cohort (Table 1, Supplementary Table 1). We compared the gene expression profiles of the following sample group pairs: normal lung tissues to tumor tissues, normal lung to primary lung tumors, and normal lung to metastatic lung tumors and NSCLCs. Furthermore, within the NSCLC group, we compared adenocarcinomas to nonadenocarcinomas, squamous cell carcinomas to nonsquamous cell carcinomas, and large cell carcinomas to non-large cell carcinomas. According to our statistical criteria (\( P \leq 0.0001 \) or absolute value of SAM \( \geq 3.75 \)), we were able to identify groups of genes of which the expression level best segregated samples into these groups except for the large cell carcinoma category (Supplementary Fig. 1, A–D, right gene lists).3

To examine how these gene expression patterns differentiated the sample groups, we used hierarchical clustering analysis and class-prediction models. Hierarchical clustering analysis classified the samples into subgroups, almost exactly as we expected (Fig. 1, A–D; Supplementary Fig. 1, A–D).3 However, there were several misclustered samples. For example, sample 32, thought to be a lung adenocarcinoma, clustered with normal lung as is shown in Fig. 1, A and B. Sample 5, of which the pathology report stated that it was a squamous cell carcinoma, clustered with adenocarcinoma as is shown in Fig. 1, C and D. Sample 22, stated to be a squamous cell carcinoma, clustered with adenocarcinoma as is shown in Fig. 1D. Next, we built a class prediction model system according to gene expression profile. Using our class prediction model system, most of the samples in our training cohort could be classified appropriately according to their gene expression pattern (Table 1). However, again, several samples (samples 32, 4, 5, 19, and 23) could not be classified into their stated histological groups.

Slides of the samples (samples 32, 5, 22, 4, 19, and 23) that had discrepancies between histological classification and gene expression classification were re-examined by a pathologist. Several light microscopic misclassifications were identified. According to histological re-evaluation, in sample 32, there was only \( \sim 15\% \) of tumor tissue, and the other \( 85\% \) consisted of fibrous tissue, explaining its clustering with the normal samples. Sample 5 showed nests with central necrosis suggesting cells with squamous differentiation. However, most of the tumor in this sample was very poorly differentiated lung cancer without intracellular bridges and without cytoplasmic keratinization, such that this sample might be better classified as large cell lung cancer. We also found that sample 22 was better classified as large cell carcinoma on review.

Through this initial analysis, we recognized the need for careful reverification of standard clinical histology and information on our samples, and all of the samples were rereviewed without finding any more major histological changes. In our review of clinical information, we found that sample 34 and sample 43 had received chemotherapy before surgery. For additional analysis, these two squamous cell carcinoma samples were analyzed in test cohort separately from other nonpretreated samples (Supplementary Table 1).3

### Identification of Genes Expressed in Sample Groups.

We reanalyzed the entire data set using the revised sample information (Supplementary Table 1).3 Again, we sought genes of which the expression levels best segregated samples into histological groups using the same statistical methods. Through this analysis, in most of the histological group comparisons, we identified a larger number of significant genes (Fig. 2, A–E; Supplementary Table 6).3 When we compared large cell lung carcinoma to non-large cell carcinoma in NSCLCs, we identified 2 statistically significant genes. In addition to the comparisons used in the initial analysis, we compared 3 more sample combinations within NSCLCs: adenocarcinomas to squamous cell carcinomas, adenocarcinomas to large cell carcinomas, and squamous cell carcinomas to large cell carcinomas. Although we could identify 27 genes of which expression levels differentiate adenocarcinoma from squamous cell carcinoma (Fig. 2F), we could not identify any gene of which the expression level differentiated large cell carcinomas from adenocarcinoma or large cell carcinomas from squamous cell carcinoma. The expression differences of several of these genes were also confirmed by Northern blotting (Supplementary Fig. 2, A and B).3

The proteins encoded by these differentially expressed genes in various lung cancer groups included transcription factors, proteins related to organic transport, and various metabolic cycles including the activated methyl cycle-related enzymes or structural components. Among the genes within these groupings were some of which the expression was already known to correlate with certain lung cancer subtypes. For example, the folate receptor is known to be expressed predominantly in adenocarcinomas of lung, whereas Mucin 1 is also known to be highly expressed in some lung adenocarcinomas and Keratins are known squamous differentiation markers (22–24).

Hierarchical cluster analysis showed clearer clustering for each histological group compared with the initial analysis (Fig. 2, A–D). However, large cell carcinomas were again difficult to
cluster into one group. When we clustered NSCLC samples according to the expression levels of genes of which the expression patterns correlated with adenocarcinoma, large cell carcinoma samples 4 and 5 clustered closely with adenocarcinomas (Fig. 2C). When we clustered NSCLC samples according to the expression levels of genes of which the expression patterns correlate with squamous cell carcinoma, large cell carcinoma samples 6, 14, and 19 clustered with squamous cell carcinomas (Fig. 2D). Even when we used the expression patterns of 2 genes identified to correlate significantly with large cell carcinoma clustering, the 2 large cell carcinoma samples 5 and 19 clustered with non-large cell carcinomas (Fig. 2E). Overall, our class prediction model system could classify almost all of the samples in the training cohort correctly except sample 5 (a large cell carcinoma), sample 23 (a squamous cell carcinoma), and sample 19 (a large cell carcinoma; Table 2).

**Class Prediction of Blinded Samples in Test Cohort.**

To confirm the relationship between the expression patterns of identified groups of genes and histological groups, we attempted to predict the histology of unknown samples in a test cohort using our class prediction model. For this purpose, we analyzed the additional data from 15 new samples in a blinded test cohort.

Through this analysis, we treated the 2 samples from patients who had received chemotherapy before surgery separately from the other samples (Supplementary Table 1).3 Class prediction of these blinded samples was performed based on the expression patterns of genes identified as correlated with known histological groups through the training cohort analysis (Table 3). We could successfully distinguish all of the tumor samples from normal lung correctly. We could also predict all of the primary lung tumors and metastatic lung tumors except sample 48. Sample 48, a metastatic lung tumor from a colon adenocarcinoma, was predicted to be a primary lung tumor and additionally predicted as a lung adenocarcinoma. Interestingly, this patient had multiple primary tumors, and this tumor was initially thought to be a primary lung tumor clinically, but ultimately judged to be a colon metastasis on clinical grounds. In the NSCLC group, most of the samples could be predicted correctly. However, it was again difficult to classify large cell carcinomas.

We also applied our classification system to the 2 pretreated squamous cell carcinoma samples, 34 and 43. Our prediction model could not identify these samples as squamous cell lung carcinoma correctly. Sample 34 was predicted as a primary NSCLC but was not assigned into any of the subtypes. Sample
43 was predicted not to be a primary lung cancer but more like normal lung or metastatic lung tumor. Interestingly, tumor 34 showed little response to therapy, whereas sample 43 had a good partial response clinically.

**Identification of Genes That Relate to Clinical Behavior.** Finally, we identified genes that correlated with the biological behavior of lung cancers. For this analysis, we used two kinds of clinical information, postoperative nodal metastatic status and overall survival (Supplementary Table 7). In the first analysis, we looked for genes that were differentially expressed between the nodal metastasis negative (N0, n = 17) and positive groups (N1 or N2, n = 14) in NSCLC samples (n = 31, data from sample 47 was removed through this analysis because of the duplication with sample 6; Supplementary Table 8). No gene was identified as differentially expressed between these groups in this dataset. Next, we tested for genes of which the expression level correlated with overall survival. When we compared the group with overall survival ≥1 year (n = 19) to the group who died within 1 year (n = 4; Supplementary Table 8), we identified one gene, ACTN4 gene (H50993), as highly expressed in the poor survival group. Then we compared the survival of ACTN4 low-expression group to that of ACTN4 high-expression group (Supplementary Table 9). As is shown in Fig. 3, the expression level of ACTN4 gene was a significant prognostic predictor of overall survival.

**Sequence Verification of cDNAs.** We reverified the sequences of cDNAs on our microarrays for the genes we identified as differentially expressed between sample groups through this study. Of 172 clones sequenced, 164 clones (95.3%) were verified to have the correct insert. Two of 8 clones that had wrong sequence, R18039 and N63941, had genomic sequences, different from accession IDs. Another 3 of those 8 clones, AA456400, AA495846, and H63865, had unknown sequences, different from accession IDs. The last 3 clones, H54093, AA422058, and AA211448, had cDNA sequences different from accession IDs (Fig. 2; Supplementary Fig. 1 and Supplementary Table 6).
DISCUSSION

We applied gene expression profiling approach to lung cancer classification. In this study, we used supervised methods to identify genes that correlated with certain biological features of a training cohort of tumors. Then, we attempted to examine the relationship between identified gene expression patterns and sample histologies by two statistical methods, hierarchical clustering analysis and a class prediction model system. A blinded confirmatory set of samples in test cohort was used to confirm our findings.

Three recent studies have analyzed lung cancer gene expression profiles by microarray technology (11–13). In two of these studies (11, 12), the investigators selected genes for analysis of which the expression was most similar within duplicate experiments yet varied widely among the other tumor samples. Then, using gene expression profiles of the selected genes across their tumor set, they clustered the samples by an unsupervised method to identify potentially novel classes in lung cancers. Each group discovered different candidate lung cancer subtypes within adenocarcinoma. Although the statistical approaches we used were different, several genes we identified as discriminators of histological groups were contained with the set of discriminant genes reported in those two papers. For example, four and a half LIM domain 1 was highly expressed in normal lung, and keratin 5 and bullous pemphigoid antigen 1 were highly expressed in squamous cell carcinoma in our study as was reported by Bhattacharjee et al. (12) Folate receptor, KIAA 1319 protein, and mucin 1 were highly expressed in adenocarcinoma compared with squamous cell carcinoma in our study as was also reported by Garber et al. (11). Besides these previously identified genes, we identified many genes that had not been reported previously to be differentially expressed in lung cancers. The expression level of these genes or the proteins encoded by these genes may be useful as novel biomarkers.

In our analyses, the large cell carcinoma group was extremely difficult to cluster into one group by its gene expression profile. When we attempted to identify the genes that correlated with large cell carcinomas, no or only a few genes were iden-
Lung Cancer Molecular Classification

Interestingly, this tumor ever, our prediction model predicted the colon adenocarcinoma hepatocellular carcinomas, as nonprimary lung tumors. How-

model accurately predicted 2 of these, the adrenal tumor and 75), and a hepatocellular carcinoma (sample 76). Our prediction confirm our prediction model system. In this test cohort, we had 3 kinds of metastatic tumors that were not in our training cohort, complete clinical information need to be analyzed to answer this question.

suggest not. Greater numbers of large cell carcinomas with small to make clear conclusions as to whether large cell carci-

noma clustered with adenocarcinomas in one of the adenocar-

large cell carcinoma cluster, and the remaining large cell carci-

expression profiles were correlated with survival of stage I recently, Beer (13) identified sets of genes of which the expres-

We also had 2 tumors from patients who had received preoperative therapy in our test cohort. Both of them had radiation and chemotherapy before surgery, and sample 34 did not respond significantly to this therapy, whereas tumor 43 responded well to treatment. Our prediction model classified sample 34 as primary lung carcinoma, but sample 43 was classified as a non-lung primary metastasis. This might reflect the expression changes associated with response to treatment and deserves additional investigation. Much more refinement of the model with a variety of lung and non-lung tumors, and with larger arrays will be required to obtain a more accurate histo-

ologic prediction model.

Whereas an interesting proof of principle, the potential clinical benefit of this technology will lie in its ability to predict the biological and clinical behavior of tumors. To identify such patterns, a supervised statistical method is required. In our study, we found the expression level of a novel marker ACTN4 was a significant prognostic predictor in these lung cancer patients. The ACTN4 gene product was originally identified through immunoscreening of monoclonal antibodies reactive with proteins up-regulated upon enhanced cell movement (25). Recently, Beer et al. (13) identified sets of genes of which the expression profiles were correlated with survival of stage I patients with lung adenocarcinomas. They chose these genes by leave-one-out and training-testing cross-validation methods, and

tified as candidates. Also, when we attempted to examine the relationships between gene expression patterns and sample histologies by using hierarchical clustering analysis and a class prediction model, large cell carcinomas were outliers. For example, large cell carcinoma samples 5 and 19 had gene expression profiles quite similar to adenocarcinoma and squamous cell carcinoma, respectively. These data suggest that poorly differ-

tiated tumors by light microscopic evaluation may be more related to tumors from either of these two groups than to each other. In the article from Garber et al. (11), they had four pure large cell carcinomas in their clustering analysis across 73 lung samples. Of those four large cell carcinomas, three large cell lung carcinomas clustered with adenocarcinomas to make a large cell carcinoma cluster, and the remaining large cell carcino-

were not significant. Greater numbers of large cell carcinomas with complete clinical information need to be analyzed to answer this question.

Our study is unique in that we used a blinded test cohort to confirm our prediction model system. In this test cohort, we had 3 kinds of metastatic tumors that were not in our training cohort, a colon adenocarcinoma (sample 48), an adrenal tumor (sample 75), and a hepatocellular carcinoma (sample 76). Our prediction model accurately predicted 2 of these, the adrenal tumor and hepatocellular carcinomas, as nonprimary lung tumors. How-

noma subgroups, adeno group 3 cluster. The numbers of large

large cell carcinomas in their clustering analysis across 73 lung

No. of differentially expressed genes

No. of misclassified samples (sample ID)

Probability of random permutations with misclassifications

No. of misclassified samples (sample ID)

Exact $P$ for testing $P = 0.5$

Table 2 Classification of 29 samples in revised training cohort according to gene expression pattern

<table>
<thead>
<tr>
<th>Classification (sample size)</th>
<th>No. of differentially expressed genes</th>
<th>No. of misclassified samples (sample ID)</th>
<th>Probability of random permutations with misclassifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>All samples</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal lung (3) vs. tumor (26)</td>
<td>54</td>
<td>0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Normal lung and metastatic tumor (5) vs. primary lung tumor (24)</td>
<td>62</td>
<td>0</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Normal lung (3) vs. NSCLC (23)</td>
<td>77</td>
<td>0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>NSCLC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adeno (8) vs. non-adeno (15)</td>
<td>6</td>
<td>1 (Non-adeno; sample 5)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Squamous (8) vs. non-squamous (15)</td>
<td>10</td>
<td>1 (Squamous; sample 23)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Large (7) vs. non-large (16)</td>
<td>2</td>
<td>1 (Large; sample 19)</td>
<td>=0.001</td>
</tr>
<tr>
<td>Adeno (8) vs. squamous (8)</td>
<td>27</td>
<td>0</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Table 3 The class prediction of 13 samples in test cohort according to gene expression pattern

<table>
<thead>
<tr>
<th>Classification (sample size)</th>
<th>No. of differentially expressed genes</th>
<th>No. of misclassified samples (sample ID)</th>
<th>Exact $P$ for testing $P = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All samples</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal lung (0) vs. tumor (13)</td>
<td>54</td>
<td>0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Normal lung and metastatic lung tumor (4) vs. primary lung tumor (9)</td>
<td>62</td>
<td>1 (Metastatic tumor; sample 48)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Normal lung (0) vs. NSCLC (9)</td>
<td>77</td>
<td>0</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>NSCLC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adeno (4) vs. non-adeno (5)</td>
<td>6</td>
<td>0</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Squamous (4) vs. non-squamous (5)</td>
<td>10</td>
<td>1 (Squamous; sample 42)</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Large (1) vs. non-large (8)</td>
<td>2</td>
<td>3 (Non-large; sample 40,41,49)</td>
<td>not significant</td>
</tr>
<tr>
<td>Adeno (4) vs. squamous (4)</td>
<td>27</td>
<td>0</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
used their model to predict survival. Unfortunately, the gene we identified as most correlated with survival, ACTN4, was not included in their list of 4966 analyzed genes. Therefore, we cannot determine the significance of ACTN4 in their data set. Although statistically significant, because our sample size was small and had only a short follow-up interval, the significance of ACTN4 expression in lung cancer survival needs to be reconfirmed by a larger cohort with a longer follow-up interval.

There has been a concern about potential errors in cDNA clones for microarray production (19). To increase our confidence in our findings, we sequenced every cDNA clone identified as important for each of the classes in our study. Eight clones of 172 (4.7%) showed incorrect sequences. This error rate is much lower than other studies reported previously and suggests that sequence verification is indispensable for verifying microarray data. Unfortunately, the data generated by the Beer et al. (13) and Bhattacharjee et al. (12) studies cannot be confirmed in this way because of limitations in proprietary oligonucleotide arrays.

Our results thus confirm that gene expression profiling may be an efficient tool for classifying lung tumors into biologically important and/or prognostic groups, and identifying genes associated with those distinctions. We demonstrate that these classifications can successfully classify blinded test cohorts of tumors. Using a small set of tumors and only limited clinical information, we successfully identified sets of genes distinguishing cancer from normal lung tissue, lung primary from metastasis, and with the known light microscopic histological groups. Our data also provide evidence that large cell carcinomas may not be a genetically distinct group, but may often represent tumors more closely related to adenocarcinomas or squamous cell carcinomas than to each other. In a few cases in our study, genetic predictions did not agree with the findings of traditional light microscopic evaluation and identified misclassifications in the original pathology reports. An unsuspected association of ACTN4 expression with survival was also identified and should be investigated additionally in larger numbers of patients. It is hoped that in the future, genetic classification will indicate other novel features of these tumors that were previously undetectable by standard procedures. The ability of our classification to correctly identify blinded tumor samples may suggest that the patterns we and others are observing may have real biological significance.

ACKNOWLEDGMENTS

We thank Mark McQuain, Melanie Robinson, Vicky Amann, and Dr. William Grady at the Vanderbilt University Medical Center, and Drs. Adi F. Gazdar, Shinichi Toyooka, and Kiyomi O. Toyooka at University of Texas Southwestern Medical Center for helpful support and thoughtful suggestions.

REFERENCES


A Training-Testing Approach to the Molecular Classification of Resected Non-Small Cell Lung Cancer

Noboru Yamagata, Yu Shyr, Kiyoshi Yanagisawa, et al.


Updated version
Access the most recent version of this article at:
http://clincancerres.aacrjournals.org/content/9/13/4695

Cited articles
This article cites 21 articles, 9 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/9/13/4695.full.html#ref-list-1

Citing articles
This article has been cited by 18 HighWire-hosted articles. Access the articles at:
http://clincancerres.aacrjournals.org/content/9/13/4695.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.