Published in final edited form as:

Lung Cancer. 2007 February ; 55(2): 157–164.

Distinguishing Lung Tumours From Normal Lung Based on a Small Set of Genes

Tatiana Dracheva^a, Reena Philip^b, Wenming Xiao^b, Alexander G. Gee^c, John McCarthy^c, Ping Yang^d, Yue Wang^b, Gang Dong^b, Hongjun Yang^b, and Jin Jen^a

^aLaboratory of Population Genetics, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892

^bMetriGenix US, Inc., Baltimore, MD 21227

^cAnVil Incorporated, Burlington, MA 01803

^dMayo Clinic College of Medicine, Department of Health Sciences Research, 200 First St. SW, Rochester, MN 55905

Abstract

Identifying specific molecular markers and developing sensitive detection methods are two of the fundamental requirements for detection and differential diagnosis of cancer. Toward this goal, we first performed cDNA array analysis using 65 non-small cell lung cancer and non-involved normal lung tissues. We then used several complementary statistical and analytical methods to examine gene expression profiles generated by us and others from four independent sets of normal and neoplastic lung tissues. We report here that several sets of roughly 20 genes were sufficient to provide a robust distinction between normal and neoplastic tissues of the lung. Next we assessed the predictive ability of these gene sets by using Flow-Thru Chips® (FTC) (MetriGenix, Baltimore, MD) containing 20 genes to screen 48 primary lung tumours and normal lung tissues. Gene expression changes detected by FTC distinguished lung cancers from the normal lung tissues by using an RNA amount equivalent to that present in as few as 300 cells. We also used an independent set of 24 genes and showed that their expression profile was equally effective when measured by quantitative polymerase chain reaction (Q-PCR). Our results demonstrate that lung cancers can be identified based on the expression patterns of just 20 genes and that this approach is applicable for cancer diagnosis, prognosis, and monitoring using small amount of tumor or biopsy samples.

Keywords

cancer detection; class prediction; expression profiling; gene chip; non–small cell lung cancer; Flow-Thru Chip

Corresponding author: Dr. Jin Jen, 41 Library Dr., Rm. D702, Bethesda, MD 20892, USA. Email: jenj@mail.nih.gov.
Present addresses: FDA/CDRH/OIVD, 2098 Gaither Rd., Rockville, MD 20850 (RP); BIMAS, Center for Information Technology, NIH, Bethesda, MD 20892 (WX); University of Massachusetts–Lowell, Department of Computer Science, 198 Riverside St., Lowell, MA 01854 (AG); Pharsight Corporation, Strategic Consulting Services, Mountain View, CA 94041 (JMC) Alexandria Research Institute, Alexandria, VA 22314 (YW) National Institute of Allergy and Infectious Diseases/NIH, 6610 Rockledge Dr., Bethesda, MD 20892-6601 (GD) Tianjin Biochip Corporation, Tianjin, China 300457 (HY)

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest statement: None declared.

Cancer is a genetic disease that involves the alteration of many genes [1]. Identifying specific genetic changes and using them as molecular markers have offered powerful new possibilities in which cancer diagnosis and prognosis can be improved [2]. In the past several years, studies have shown that molecular signatures can predict therapeutic response as well as clinical outcome in cancer patients [3-8].

In lung cancer, several gene expression profiling-based studies involving normal and neoplastic lung tissue have shown that 1) gene expression profiles are distinct for different types of lung cancer and normal tissue [9-11]; 2) the set of genes associated with a tumour type reflects the biological basis of the cancer and provide a molecular signature for the disease [10,12]; and 3) gene expression profile could provide prognostic as well as predictive value about the cancer subtype or metastatic potential [13,14]. However, because of the varied platforms and the different types of tissues used for each analysis, addressing the robustness of these predictive gene sets and estimating the minimum number of genes necessary in such sets to accomplish these goals has been difficult.

In this study, we first identified the minimum number of genes sufficient to distinguish lung tumours from adjacent normal lung tissue by using independent training and testing sample sets. We then confirmed the robust nature of such distinction in another set of training and testing samples. We show that one can distinguish lung tumour from normal tissues by using several different sets of approximately 20 genes. We confirmed the robustness of this method by using both Flow-Thru Chips (FTCs) (MetriGenix, Baltimore, MD) containing 20 genes and Q-PCR ABI Micro Fluidic Cards (MFC) (Applied Biosystems, Foster City, CA) containing 24 genes. Our results indicate that as few as 300 cells are sufficient to identify the neoplastic state of lung.

Materials and Methods

Samples and Datasets for Array Analyses

A total of 65 tumour and normal total RNA samples were included in Set 1 (Table 1) and used for the GeneChip analysis. These samples were either purchased from BioChain (Hayward, CA) or obtained from John Hopkins Hospital (Baltimore, MD) as previously described [10]. GeneChip probe arrays U95A and U95Av2 were obtained from Affymetrix (Santa Clara, CA). RNA samples were prepared and hybridized by following standard Affymetrix protocol. Two internal GeneChip standards, GAPDH and β -actin, were used to ensure sample and procedure quality control.

Gene expression data for Set 2 were obtained from the Lung Cancer Genomics Web site (<http://research.dfci.harvard.edu/meyersonlab/lungca/data.html>)[11]. Sets 3 and 4 contained GeneChip probe array U95A-E data from 78 and 37 samples, respectively, and were provided by Gene Logic (Gaithersburg, MD). In the FTC versus Taqman comparative analyses, four normal lung, three squamous-cell carcinoma, and two adenocarcinoma samples for were purchased from Clontech (Mountain View, CA) and the American Type Culture Collection (Manassas, VA). Affymetrix data used for comparisons with the FTC results were provided by Gene Logic and included array data for 90 normal lung samples, 39 adenocarcinoma samples, and 34 squamous-cell carcinoma samples. For MFC validation, we used an independent set of 17 pairs of fresh-frozen primary tumour and nontumour lung tissues from the Mayo Clinic. All samples were obtained following protocols approved by the respected Institutional Review Boards.

Normalization and Filtering

For the analysis purposes each probe on the Affymetrix chip was treated as a separate gene and referred to as such throughout the text. The mean intensity for each gene in Set 1 was normalized using the Affymetrix MicroArray Suite software. Specifically, we set all intensities less than 45 to 45 (45 is the median of positive intensity values less than 100) and then median-centered the intensity of each array at 200. Genes were excluded if intensity values were equal to or less than 100 for that gene or had an absolute call of “absent” in all samples. Intensity scores were then log transformed. For the eight samples that had repeats, we averaged each twin sample by taking the mean of the two log-transformed intensities. This processing method resulted in 9918 analyzable genes for 65 arrays used in Set 1. Since many of the 9918 genes had small variation and contributed mostly to background noise, we filtered out genes with small variances using the following rule. A gene is filtered out if $(n - 1) \times s_i < \mu \times \chi^2(.99, n - 1)$, where n is the number of samples (65); s_i is the sample variance for gene i , where i ranges from 1 to 9918; μ is the median of the sample variances $s_1, \dots, s_{9918}$; and $\chi^2(.99, n - 1)$ is the 99th percentile of the chi-square distribution with $n - 1$ degrees of freedom. A total of 3652 genes remained after such filtering. Set 2 was normalized exactly the same as for Set 1. For Set 3 and Set 4, we were able to arrive at approximately 2000 genes that were most differentially expressed by filtering based on a standard t statistic with probabilities less than 0.05 after Bonferroni correction.

Clustering, Multidimensional Scaling and Class Prediction

We performed unsupervised cluster analysis by using a hierarchical agglomerative clustering algorithm with average linkage and Pearson correlation metric to study the relationship between samples [15,16]. We also used principal-component analysis (PCA) to demonstrate the similarity in expression profiles among the various samples [10,16].

We tested a compound covariate predictor to determine whether we could use patterns of gene expression to sort samples into two classes on the basis of normal or tumour sample and then squamous or adenocarcinoma tumour subtype [17]. We then estimated the misclassification rate by using leave-one-out cross-validation and used random permutations of the class membership indicators to determine the statistical significance of the results.

Performance Evaluation

We determined gene set sizes via a repeatable two-step process, which involved selecting and then evaluating predictor genes. Because the combinatorial scale of trying all possible gene sets requires significant time and computational power, we used sample gene sets defined by three different gene selection methods, as described before [18]. Gene sets were systematically drawn from the reduced set of 162 most significant differentially expressed gene probes representing 159 unique genes (Supplemental Table I). First, we defined and analyzed the results from 10 independent random gene sets. This step provided a lower predictive bound for each gene set size. Second, using a standard F test for statistical significance, we chose only genes that demonstrated high statistical significance. Finally, we included the RadViz™ technique developed at AnVil (Burlington, MA), which is based on a pairwise t statistic [18]. Applying these three approaches to the available expression data, we were able to generate gene sets that ranged from 1 to 100 genes.

To evaluate the resulting sets of genes, we applied a collection of predictive algorithms to each set by using a 10-fold cross-validation testing approach. Initial comparison of both 10-fold and hold-one-out cross-validation showed that both methods produce essentially the same predictive accuracy. The predictive algorithms used in this analysis included but were not limited to variations of neural networks, support vector machines, Naïve Bayes, and k th-nearest

neighbors, as implemented in the Weka software application (<http://www.cs.waikato.ac.nz/~ml>) [19].

FTC Design and Analysis

FTCs were produced at Metrigenix according to standard protocols [20]. The probes consisted of 60-mer oligonucleotides were designed for each gene with a specified ΔG , GC-content, and melting temperature. Each probe sequence was 58–62 nucleotides long and was located within the first 500 bases of the 3' end of the gene. Also, each probe sequence was BLAST searched against sequence databases to ensure minimal similarity to any gene in the genomic database. The dynamic range of the detection was 2.5 logs with cRNA generated from 1 μ g of total RNA (data not shown).

Total RNA samples were used to isolate cRNA as per published protocols [21]. The cRNA was isolated with biotinylated CTP and UTP nucleotides for use in subsequent streptavidin–horseradish peroxidase staining to indirectly detect hybridization by chemiluminescence. In a standard experiment, 10 μ g of cRNA was used in hybridization buffer containing 1 $\times$ MES (morpholine ethanesulfonic acid), 0.88 M NaCl, 0.02 M EDTA, 0.5% sarcosine, 33% formamide, and 50 μ g of herring sperm DNA. Samples were first denatured at 95°C for 5 minutes and then vortexed and spun down for 2 minutes at maximum speed before being added to the FTC. Hybridization was performed at 37°C for 16 hours with continuous rotation of the cartridge in an incubator. After hybridization, samples were removed and the FTC was washed with 1 $\times$ SSPE (0.18 M NaCl, 10 mM NaH₂PO₄, and 1 mM EDTA pH 7.7) for 5 minutes, followed by blocking for 5 minutes with 1% goat serum. Staining was performed using 0.75 ng of streptavidin–horseradish peroxidase (Pierce) in 1 $\times$ SSPE for 5 minutes. The FTC was flushed again with 1 $\times$ SSPE before the detection step. Detection and imaging were performed as previously described [20]. All the steps except the hybridization step were performed at room temperature. The PCI imaging software (C-Imaging, Cranberry Township, PA) was used for image capture and GenePix Pro 4 (Axon Instruments, Union City, CA) was used for image analysis. The variation of the signal intensities in repeated experiments was less than 3%.

For FTC experiments, we used an array representing 20 lung cancer genes, housekeeping genes, and other control genes. The signal-to-noise ratio of each gene was normalized to that of GAPDH. Fold change was calculated between tumour and normal samples for each gene.

Taqman Analysis

The primers and probes for Taqman (Q-RT-PCR) analysis were designed using the Primer Express software from Applied Biosystems (Foster City, CA). The samples were run on ABI 7700 Taqman under the following standard protocol: step 1- 2 minutes at 50°C; step 2- 30 minutes at 60°C and 5 minutes at 95°C; step 3- 20 seconds at 94°C and 1 minute at 60°C (repeat step 3 for 40 cycles); reaction volume was 25 μ l. For Taqman analysis, tumour versus normal fold changes were calculated using comparative the C_T method by normalizing to GAPDH.

Quantitative Polymerase Chain Reaction Analysis on MFC

Probes and primers for the 24 of the 162 most differentially expressed genes were randomly selected and applied onto an ABI custom MFC with S18 primers and probe (ABI probe 4342379-18S) included as a control. Total RNA from 17 tumour and normal pairs was extracted with the standard Trizol procedure, and then an ABI High-Capacity cDNA Archive kit was used to synthesize cDNA for quantitative polymerase chain reaction (Q-PCR) analysis on the MFC by using standard ABI protocol.

Results and Discussion

Identifying the Minimum Number of Genes Required for Lung Cancer Classification

Our previous studies and those of others have shown that molecular profiles differ by the cellular subtype of the tumour. Whether this distinction could be accomplished using a small number of genes that are readily testable in a clinical setting is unclear. To address this issue, we selected gene expression profiles that were generated using oligonucleotide-based GeneChip arrays and divided them as either training or testing sets to identify the most informative genes. The first test set (Set 1) contained gene expression data on 65 different samples of either lung tumour or normal lung tissue that we generated for this study (Table I). After filtering out those genes with similar expression patterns among all tissues, 3652 genes were analyzed and a set of 162 of the most differentially expressed genes was identified (Supplemental Table I). To equally represent adenocarcinoma and squamous cell carcinoma in the “tumour” class prediction, we used t-test to select 100 gene probes each most differently expressed between adenocarcinoma and normal samples and between squamous cell carcinoma and normal samples. These two sets of genes shared 38 probes, giving the combined final set of 162 probes, representing 159 unique genes. We treated each of those probes as a separate gene. PCA using the 162 genes showed distinctive separation among the samples of different tissues: normal, adenocarcinoma, and squamous-cell carcinoma (Fig. 1). The covariate predictor based on these genes correctly classified 62 (95%) of 65 samples. All squamous-cell carcinoma samples were classified correctly. Overall accuracy of this predictor was approximately 95% (Table II). Although a high level of accuracy was achieved, this result was expected since the leave-one-out method used for validation assumes that the training set of samples is different but biologically similar to the test set. To further evaluate the accuracy of the predictor, we applied this predictor to a second set (Set 2) of expression data from an independent study (Table I). This data set included 183 normal and lung cancer samples; the predictor had only six misclassifications with a correct classification rate of 96% (Table II). Interestingly, our predictor also correctly classified all six small-cell lung cancer samples as cancer, even though the training set did not include small-cell lung cancer samples.

Validation of Minimum Gene Prediction Using Independent Sample Sets

We attempted to verify the number of genes required for the predictor by using a variety of different analytical methods. Our results demonstrate that a predictor composed of several different sets of 20 most differentially expressed genes could essentially provide the same predictive accuracy (Table II). This result suggests that there are multiple sets of genes that differentiate between normal and tumour samples and that several sets of about 20 genes each is sufficient to predict this difference. Even for the sets randomly selected from the list of 162 genes, almost perfect classification was achieved with accuracy greater than 95% for the gene sets of approximately 20 or more genes (Fig. 2). In addition, an in-depth analysis of the two-class normal-versus-cancer problem demonstrated that as few as two genes could distinguish between tissue types (Fig. 2), and as few as 15 genes could distinguish among tumour subtypes [18]. Using linear discriminate analysis tools available directly from GeneLogic, we also observed the robust nature of such gene expression-based predictors in an independent set of 78 (Set 3) normal lung and tumour tissue samples. Again, a set of 20 genes, independent from 162, could also distinguish between normal and tumour lung tissues with high accuracy in both the training set (Fig. 1B, Table II, Set 3) and testing samples (Table II, Set 4). More significantly, the separation between different types of tissues was better when using 20 genes comparing to 162 genes (Fig. 1A and B), because larger sets of genes usually incorporate more variance (noise) in the analysis, making the clusters less distinct than those obtained using smaller gene sets.

Distinguishing Lung Cancer from Nondiseased Lung Using FTC and Q-PCR

Automated and objective diagnoses of cancer will provide a major advantage in distinguishing tumour from the normal and may help improve cancer prognosis in clinical settings. Therefore, we constructed a lung cancer array FTC containing sense-strand oligonucleotide probes designed to hybridize with the antisense sequence of the targets. Unlike conventional two-dimensional planar surface microarray technologies, FTC technology uses a three-dimensional porous substrate with well-defined microchannels connecting its upper and lower faces such that fluid can flow through the chip [20]. This technique improves sensitivity and specificity of the hybridization signal, which is directly proportional to the relative concentration of the target in the testing sample.

Fig. 3 compares gene expression data from FTC with GeneChip and Q-RT-PCR (Taqman). Comparing the three methods indicated that FTC results were reproducible and highly sensitive for most genes. When the predictive value of FTC for lung cancer diagnosis was tested using a panel of 48 primary tumour and normal lung samples (a subset of Set 1 including five normal lung, 19 adenocarcinoma, and 24 squamous-cell carcinoma samples), clear distinctions were observed among different groups of samples (Fig. 4) when the signal-to-noise values of the 20 genes were normalized to that of GAPDH. In our hands, the signal distribution on the FTC usually allows immediate classification of the sample (Examples shown in Fig. 5).

Furthermore, we also determined whether we could obtain this classification by using an alternative method such as Q-PCR and whether the use of different samples could affect this distinction. We generated a lung cancer-specific MFC by using a subset of 24 genes from the set of 162 and tested its ability to distinguish lung tumours from normal lung tissue in the same patient. A total of 34 paired tumour/normal samples were tested, and a paired *t*-test showed that 21 of 24 genes demonstrated statistically significant difference in expression between normal and tumour samples (Table III).

Biology of Genes that Distinguish Lung Cancer from non-diseased Lung

To better understand the biological meaning of the genes from the most significant 162-gene list, we used Ingenuity Pathways Analysis software to analyze the pathways and functions in which they were involved. As expected, many of the genes that provided the most significant distinction between normal lung and tumour tissue were those that were involved in cellular growth and proliferation (45 genes), cell death (41 genes), and cellular movement (34 genes) (Supplemental Table II). Although some gene expression differences may simply reflect the difference between an epithelially-derived tumour and normal lung tissue, which may contain cells derived from multiple lineages, many genes in this list are known to be involved in different types of cancer, including that of the lung. Among them are *DLC1*, *FGFR2*, *NME1*, *TFAP2A*, and *TGFB2*, known to be involved in breast cancer [22-25]; *COL1A2*, *SLC16A4*, and *TTK*, associated with gastric carcinoma [26,27]; and *CACNA2D2*, *SEMA3B*, *TP73L* (*p63*), and *DLC1*, involved in cell death, apoptosis, and colony formation of lung cancer cell lines [28-31]. However, it is worth noting that although having expression differences in many genes, as few as 20 genes is sufficient to reliably predict the neoplastic state of the lung samples. We believe this reflects the differences in cellular origin and metabolism between the tumour and non-tumour lung as well as the extensive genetic changes that are commonly associated with human cancers. Recently, several studies have identified relatively small gene sets that can predict the survival of patients with NSCLC [32,33]. The FTC and MFC methods described here demonstrate the potential to develop similar simple tests that can be readily used for lung cancer prognosis based on a minimal set of predictive genes.

In conclusion, our study demonstrated several important findings: 1) roughly 20 genes can distinguish the neoplastic phenotype of non-small-cell lung cancer from normal lung tissue;

2) several sets of these genes can provide the same discriminatory power; 3) several methods used for measuring gene expression such as Q-PCR and FTC technology can rapidly detect and analyze gene expression levels by using a critical set of genes; and 4) a custom-based array chip or MFC is potentially useful in monitoring gene expression profiles of different tissue types for cancer diagnostics and monitoring disease progression in clinical settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

We thank Dr. Arindam Bhattacharjee and Dr. Matthew Meyerson for sharing with us their gene expression data, Dr. Joanna Shih for statistical guidance, and Mr. Gabe Waggoner for editing.

References

1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484–7. [PubMed: 7570003]
2. Sidransky D. Emerging molecular markers of cancer. *Nat Rev Cancer* 2002;2:210–9. [PubMed: 11990857]
3. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 2002;346:1937–47. [PubMed: 12075054]
4. Holtkamp N, Reuss DE, Atallah I, et al. Subclassification of nerve sheath tumors by gene expression profiling. *Brain Pathol* 2004;14:258–64. [PubMed: 15446580]
5. Murakawa K, Tada M, Takada M, et al. Prediction of lymph node metastasis and perineural invasion of biliary tract cancer by selected features from cDNA array data. *J Surg Res* 2004;122:184–94. [PubMed: 15555617]
6. Iizuka N, Oka M, Yamada-Okabe H, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet* 2003;361:923–9. [PubMed: 12648972]
7. Ye QH, Qin LX, Forgues M, et al. Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nat Med* 2003;9:416–23. [PubMed: 12640447]
8. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–6. [PubMed: 11823860]
9. 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. [PubMed: 9865724]
10. Nacht M, Dracheva T, Gao Y, et al. Molecular characteristics of non-small cell lung cancer. *PNAS* 2001;98:15203–8. [PubMed: 11752463]
11. 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* 2001;98:13790–5. [PubMed: 11707567]
12. Meyerson M, Franklin WA, Kelley MJ. Molecular classification and molecular genetics of human lung cancers. *Semin Oncol* 2004;31:4–19. [PubMed: 14981576]
13. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8:816–24. [PubMed: 12118244]
14. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. *Nat Genet* 2003;33:49–54. [PubMed: 12469122]
15. Johnson, RA.; Wichern, DW. *Applied Multivariate Statistical Analysis*. Prentice-Hall; Englewood Cliffs, New Jersey: 1982. Englewood Cliffs, New Jersey: Prentice-Hall
16. Gordon, AD. *Classification*. Chapman and Hall/CRC Press; London: 1999.

17. Radmacher MD, Mcshane LM, Simon R. A paradigm for class prediction using gene expression profiles. *Journal of computational biology* 2002;9:505–11. [PubMed: 12162889]
18. McCarthy JF, Marx KA, Hoffman PE, et al. Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. *Annals of the New York Academy of Sciences* 2004;1020:239–62. [PubMed: 15208196]
19. Witten, IH.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann; San Francisco, California: 2000.
20. Cheek BJ, Steel AB, Torres MP, Yu YY, Yang H. Chemiluminescence detection for hybridization assays on the flow-thru chip, a three-dimensional microchannel biochip. *Anal Chem* 2001;73:5777–83. [PubMed: 11791544]
21. Lockhart DJ, Dong H, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675–80. [PubMed: 9634850]
22. Lucke CD, Philpott A, Metcalfe JC, et al. Inhibiting mutations in the transforming growth factor beta type 2 receptor in recurrent human breast cancer. *Cancer Res* 2001;61:482–5. [PubMed: 11212236]
23. Yuan BZ, Zhou X, Durkin ME, et al. DLC-1 gene inhibits human breast cancer cell growth and in vivo tumorigenicity. *Oncogene* 2003;22:445–5. [PubMed: 12545165]
24. Blume-Jensen P, Hunter T. Oncogenic kinase signalling. *Nature* 2001;411:355–65. [PubMed: 11357143]
25. Douglas DB, Akiyama Y, Carraway H, et al. Hypermethylation of a small CpGuanine-rich region correlates with loss of activator protein-2alpha expression during progression of breast cancer. *Cancer Res* 2004;64:1611–20. [PubMed: 14996719]
26. Oue N, Hamai Y, Mitani Y, et al. Gene expression profile of gastric carcinoma: identification of genes and tags potentially involved in invasion, metastasis, and carcinogenesis by serial analysis of gene expression. *Cancer Res* 2004;64:2397–405. [PubMed: 15059891]
27. Mori Y, Sato F, Selaru FM, et al. Instability typing reveals unique mutational spectra in microsatellite-unstable gastric cancers. *Cancer Res* 2002;62:3641–5. [PubMed: 12097267]
28. Zeng SX, Dai MS, Keller DM, Lu H. SSRP1 functions as a co-activator of the transcriptional activator p63. *Embo J* 2002;21:5487–97. [PubMed: 12374749]
29. Tomizawa Y, Sekido Y, Kondo M, et al. Inhibition of lung cancer cell growth and induction of apoptosis after reexpression of 3p21.3 candidate tumor suppressor gene SEMA3B. *Proc Natl Acad Sci U S A* 2001;98:13954–9. [PubMed: 11717452]
30. Yuan BZ, Jefferson AM, Baldwin KT, Thorgeirsson SS, Popescu NC, Reynolds SH. DLC-1 operates as a tumor suppressor gene in human non-small cell lung carcinomas. *Oncogene* 2004;23:1405–11. [PubMed: 14661059]
31. Carboni GL, Gao B, Nishizaki M, et al. CACNA2D2-mediated apoptosis in NSCLC cells is associated with alterations of the intracellular calcium signaling and disruption of mitochondria membrane integrity. *Oncogene* 2003;22:615–26. [PubMed: 12555074]
32. Guo L, Ma Y, Ward R, Castranova V, Shi X, Qian Y. Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. *Clin Cancer Res* 2006;12:3344–54. [PubMed: 16740756]
33. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res* 2006;66:7466–72. [PubMed: 16885343]

Figure 1. Principal-component analysis (PCA) of samples from Set 1 based on 162 selected genes (**A**) and Set 3 based on 20 selected genes (**B**). Blue, red, and green spheres represent squamous-cell carcinoma, adenocarcinoma, and normal lung tissue, respectively.

Figure 2.

Classification analysis that distinguishes normal lung tissue from lung cancer based on the 162 genes using three different gene selection approaches (shown in different colors) with increasing number of genes. Gray, blue, and yellow circles represent analysis using randomly generated gene sets, genes most significant by *F* test, and top automated RadViz™-selected genes, respectively.

Figure 3.

Comparison of relative expression for 20 selected genes. Each bar represents the average fold difference between gene expression of the particular gene in lung cancer samples versus that in normal lung tissue as measured by Q-RT-PCR (red), FTC (green), and Affymetrix GeneChip arrays (blue).

Figure 4. Principal-component analysis (PCA) of 48 samples from Set 1 based on 20 genes using FTC. Blue, red, and green spheres represent squamous-cell carcinoma, adenocarcinoma, and normal lung tissue, respectively.

Figure 5.

Examples of FTC analyses in primary tumour and normal samples. The FTC layout is as shown with each gene probe arrayed in quadruplets. Three representative chip images are shown after hybridization with normal lung tissue, adenocarcinoma, and squamous-cell carcinoma as indicated.

Table I

Number of samples used to select genes that distinguish between normal and tumor lung tissues.

Tissue and cancer type	Set1	Set2	Set3	Set4
Normal lung	17	17	33	20
Lung cancer	48	166	45	17
NSCLC	48	160	45	17
Adenocarcinoma	21	139	25	7
Squamous cell carcinoma	27	21	20	10
SCLC	0	6	0	0
Total samples	65	183	78	37

* NSCLC = non-small-cell lung cancer; SCLC = small-cell lung cancer.

Table II

Distinguishing lung tumors from normal samples using different sets of genes.

No. of genes in each set	Set No. and analysis type			
	Set 1 (n=65)	Set 2 (n=183)	Set 3 (n=78)	Set 4 (n=37)
162	Training 95% (62 [*])	Testing 96% (177)	Training	Testing
26	94% (61)	96% (174)		
20	92% (60)	96% (176)	100% (78)	97% (36)

*
Indicates the number of samples correctly classified.

Table III

Differential expression of 24 genes based on the ABI Micro Fluidic Card.

ABI probe	Gene name	C _T Tumor (Avg.)	C _T Normal (Avg.)	Paired t-test p
GDF10-Hs00192033_m1	growth differentiation factor 10	17.32	13.35	.0012
CDC2-Hs00364293_m1	cell division cycle 2, G1 to S and G2 to M	15.32	15.64	.0425
PAK1-Hs00176815_m1	p21/Cdc42/Rac1-activated kinase 1 (STE20 homolog, yeast)	14.24	14.77	.0298
VWF-Hs00169795_m1	von Willebrand factor	12.30	9.66	.0023
PTPRB-Hs00160781_m1	protein tyrosine phosphatase, receptor type, B	14.09	10.55	.0001
13CDNA73-Hs00224921_m1	hypothetical protein CG003	14.54	13.41	.2545
RASGRF1-Hs00182314_m1	Ras protein-specific guanine nucleotide-releasing factor 1	14.76	11.95	.0137
CDH5-Hs00174344_m1	cadherin 5, type 2, VE-cadherin (vascular epithelium)	13.20	9.71	.0001
RAMP2-Hs00237194_m1	receptor (calcitonin) activity modifying protein 2	13.89	9.74	.0002
NME1-Hs00264824_m1	non-metastatic cells 1, protein (NM23A) expressed in	11.23	12.05	.0001
SERPINB5-Hs00184728_m1	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 5	17.07	17.54	.1799
CCNE1-Hs00233356_m1	cyclin E1	15.99	17.22	.0010
HCK-Hs00176654_m1	hemopoietic cell kinase	12.66	11.29	.3611
NULL-Hs00405524_m1	KIAA1102 protein	10.78	8.49	.0082
LDB2-Hs00609711_m1	LIM domain binding 2	12.40	9.52	.0011
CCNB1-Hs00259126_m1	cyclin B1	14.87	16.00	.0024
TFAP2A-Hs00231461_m1	transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)	17.59	19.32	.0086
TCF21-Hs00162646_m1	transcription factor 21	14.34	10.81	.0014
TTK-Hs00177412_m1	TTK protein kinase	15.13	16.34	.0082
DLC1-Hs00183436_m1	deleted in liver cancer 1	12.21	9.34	.0021
TEK-Hs00176096_m1	TEK tyrosine kinase, endothelial	14.73	10.86	.0000
MAOB-Hs00168533_m1	monoamine oxidase B	14.59	12.39	.0459
FIGF-Hs00189521_m1	c-fos induced growth factor (vascular endothelial growth factor D)	15.73	11.65	.0025
KAL1-Hs00608006_m1	Kallmann syndrome 1 sequence	13.37	10.49	.0005