NIH Public Access

Author Manuscript

J Thorac Cardiovasc Surg. Author manuscript; available in PMC 2008 January 13.

Published in final edited form as:

J Thorac Cardiovasc Surg. 2006 September ; 132(3): 621–627.

Differential diagnosis of solitary lung nodules with gene expression ratios

Gavin J. Gordon, PhD[a], **Levi A. Deters, BS**[a], **Matthew D. Nitz, BS**[a], **Barry C. Lieberman, BA**[a], **Beow Y. Yeap, ScD**[b], and **Raphael Bueno, MD**[a]

a *Thoracic Surgery Oncology Laboratory and Division of Thoracic Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Mass*

b *Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Mass*

Abstract

Objective—We have developed a new technique that uses the ratios of select gene expression levels to translate complex genomic data into simple clinically relevant tests for the diagnosis and prognosis of cancer. We determined whether select gene pair ratio combinations can be used to detect and diagnose lung cancer with high accuracy and sensitivity.

Methods—We used gene expression profiling data to train a ratio-based predictor model to discriminate among a set of samples (n = 145 total) composed of normal lung, small cell lung cancer, adenocarcinoma, squamous cell carcinoma, and pulmonary carcinoid (the training set). We then examined the optimal test in an independent set of samples (the test set, n = 122). Finally, we used one aspect of the test to determine whether the gene ratio technique was capable of detecting cancer in specimens from fine-needle aspirations performed ex vivo with normal lung (n = 14) and suspected tumor nodules (n = 15) acquired at our institution.

Results—We found that a ratio-based test with 23 genes could be used to classify training set samples with 90% accuracy. This same test was similarly accurate (88%) when applied to the test set of samples. We also found that this test was 87% and 100% accurate at detecting cancer in normal and tumorous fine-needle aspiration specimens, respectively.

Conclusion—The gene expression ratio diagnostic technique is likely to aid in the differential diagnosis of solitary lung nodules in patients with suspected cancer and may also prove useful bin developing lung cancer screening strategies that incorporate analysis of fine-needle aspiration specimens.

Lung cancer is one of the most common cancers in the Western world and the number one cause of cancer deaths for both men and women in the United States. As many as 80% of patients with lung cancer have non–small cell lung cancer, a histologic category of primary lung cancer that includes adenocarcinoma, squamous cell carcinoma (SCC) and large cell carcinoma. Most of the remainder have small cell lung cancer (SCLC). Other types of primary lung malignancies include pulmonary carcinoid (1% of all primary lung malignancies).[1,2]

Most patients with lung cancer are seen with advanced disease not amenable to surgical therapy. However, screening with spiral computed tomography (CT) for lung cancer is a technique rapidly gaining popularity in the United States, with the goal of identifying lung cancer at early stages, when it is far more likely to be curable with surgery.[3] Initial studies of this new screening technology have demonstrated a high incidence of nonmalignant nodules in the lungs

Address for reprints: Gavin J. Gordon, PhD, or Raphael Bueno, MD, Brigham and Women's Hospital, Division of Thoracic Surgery, 75 Francis St, Boston, MA 02115 (E-mail: ggordon@partners.org, rbueno@partners.org).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

of former smokers. The preliminary recommendations are to measure radiographic volume change of all subcentimeter nodules at 3-month intervals and to obtain biopsy samples of any growing nodule. Biopsy to obtain definitive diagnosis of any noncalcified nodule greater than 1 cm is also advised.[3–6] Biopsies can be accomplished surgically with video-assisted thoracoscopic surgery or by transthoracic fine-needle aspiration (FNA).

Percutaneous CT-guided transthoracic FNA of lung nodules is a safe and well-accepted cytopathologic diagnostic technique that has been applied to lesions as small as 5 mm. There are very few false-positive cytologic diagnoses, but the false-negative rate has been reported to approach 30%.[7] The ability of a cytologist to make a correct diagnosis depends on the quality of cells obtained and the preservation of tissue architecture. Cytologic diagnosis by FNA is also hindered by the frequent inability of the cytologist to determine the type of cancer found in the pulmonary nodule and to differentiate metastatic cancer to the lung from primary lung cancer.[8] As a consequence, the clinical diagnostic strategy in the management of many newly discovered pulmonary nodules is to surgically remove those nodules for which a definitive benign histologic typing has not been obtained or to monitor all subcentimeter nodules with interval CT scans and remove them surgically if they grow.[9]

Gene expression profiling with microarrays and complex bioinformatics tools has been used successfully to diagnose cancer and predict disease-related outcome for multiple neoplasms, including lung cancer.[10–12] Unfortunately, these models are difficult to assess clinically because they rely on the measurement of expression levels of relatively large numbers of genes with costly data-acquisition platforms and sophisticated algorithms and software. We recently described a method for translating gene expression profiling data into clinically relevant tests with ratios of gene expression for multiple cancers.[13–17] Here we report the discovery of differentially expressed genes among normal lung and multiple types of lung cancer. We then used these genes in the development of a gene ratio method for the differential diagnosis of lung cancer or pulmonary nodule. Finally, we provide evidence suggesting that this technique may complement ongoing lung cancer screening strategies through the analysis of FNA samples.

Methods

Solid Tumor and FNA Samples

Solid samples and ex vivo FNA specimens were collected from 15 consecutive patients undergoing lung resection for cancer at Brigham and Women's Hospital. As soon as the specimen was surgically removed from the patient, it was taken to the frozen-section room, where, under the supervision of the pathologist, FNA was obtained from the palpable nodule through the surrounding pulmonary parenchyma by means of equipment and protocols identical to those currently used by cytopathologists at our institution (3-mL syringe and 22-gauge needle 1.5 inches long). Aspirated FNA material was immediately placed in RNA extraction buffer (Trizol reagent; Invitrogen Corporation, Carlsbad, Calif) for processing. The diagnostic criterion standard was obtained after pathologic analysis of the solid tumor sample. Studies that used human tissues were approved by and conducted in accordance with the policies of the Institutional Review Board at Brigham and Women's Hospital.

Gene Expression Profiling Data

Microarray data for normal and tumorous tissues were obtained from two sources. Gene expression data for the training set of samples (n = 145 total) were obtained with Affymetrix high-density oligonucleotide microarrays (U95A chip; Affymetrix, Santa Clara, Calif) with probe sets representing approximately 12,000 genes and consisting of normal lung (n = 13) and the following primary tumors: SCLC (n = 7), lung adenocarcinoma (n = 89), lung SCC (n

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

= 24), and pulmonary carcinoid (n = 12).[18] Gene expression data for all additional primary and metastatic tumor samples (the test set) were acquired from a single source with the same Affymetrix U95A microarray.[19] Primary tumors of the test set consisted of lung SCC (n = 13) and the following adenocarcinomas: prostate (n = 24), colon (n = 20), breast (n = 25), gastroesophageal (n = 12), pancreatic (n = 6), and lung (n = 13). Metastatic tumors in the test set (n = 9) included those arising from breast, colon, prostate, lung, and kidney tumors.

Data and Statistical Analysis

To train an expression ratio-based predictor model, we used an approach similar to that in previous published studies.[13–17] We performed five separate analyses to determine differences in gene expression patterns between two groups composed of multiple combinations of tissues chosen from the 145 training set samples. In each of the five training subsets, one group was composed of all available samples of a single tissue type, whereas the other group consisted of a random sampling of all remaining tissue types, with equal representation according to the remaining tissue type with the smallest number of samples. For example, the lung adenocarcinoma training subset (n = 117 total) examined differences in gene expression between two groups, lung adenocarcinoma (n = 89) and not lung adenocarcinoma (n = 28, consisting of 7 samples each of the other four tissue types according to the total number of SCLC tissues). This process was repeated sequentially for the remaining training subsets: SCLC (n = 55 total), normal lung (n = 41 total), lung SCC (n = 72 total), and pulmonary carcinoid (n = 64 total). This experimental design resulted in five training sets with unique sample numbers (and membership) and was used to discover optimal discriminating genes in an unbiased fashion while ensuring equal representation among multiple tissue types.

The selection of predictor genes for use in expression ratio-based diagnosis was performed essentially as described,[15,16] with minor modifications. With a 2-sided Student (parametric) *t* test, we identified statistically significant (see Table 1 for exact *P* values) genes with inversely correlated average expression levels between both groups in each of the five training subsets. We then filtered the resulting gene lists to find those genes with at least a 2-fold difference in average expression levels between groups. To minimize the effects of background noise, the list of distinguishing genes was additionally refined by requiring that the mean expression level (Affymetrix average difference) be greater than 500 in at least one of the two groups, similar to previous studies.[15,16] A large number of genes were found to fit the filtering criteria in each of the training subsets. To further reduce the number of genes, we randomly chose for additional study a total of 8 genes from the among the most statistically significant differentially expressed genes in each training subset. Four of these genes were expressed at relatively higher levels in a single tissue type, and 4 were expressed at relatively higher levels in the remaining tissue types combined. There was a single exception: in the normal lung training subset, only 3 genes were expressed at relatively higher levels in all abnormal tissues. In one training subset (lung SCC), there was a single case of duplication among the genes chosen for further analysis, considering that (1) we randomly chose additional genes for study, (2) we initially identified genes strictly on the basis of their unique Affymetrix probe set identifiers (and not gene name), and (3) the same gene can be represented by multiple Affymetrix probe sets.

All possible nonredundant gene pair expression ratios were separately calculated for each sample in all five training subsets by placing single genes overexpressed in each tissue type in the numerator and single genes overexpressed in the combination of all other tissues in the denominator. All negative Affymetrix average difference values (undetectable) were arbitrarily assigned an expression level of 1 to facilitate meaningful comparisons. The identity of samples used to generate the gene lists was then predicted in a binary manner with ratio values and a threshold equal to 1. For example, in the lung adenocarcinoma analysis, individual samples with ratio values greater than 1 and less than 1 were predicted to be lung adenocarcinoma and

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

not lung adenocarcinoma, respectively. A final diagnostic call was made from the value of the geometric mean of the three most accurately predictive individual ratios with the same criteria. Use of the geometric mean value for multiple ratios has the effect of giving equal weights to ratio changes of identical magnitude but opposite direction on the log scale. Finally, all five tissue-specific 3-ratio tests were used to comprehensively analyze the 145 training set samples. We hypothesized that the identity of any given sample would be the tissue with the greatest geometric mean. No-calls were made conservatively when no tissue-specific geometric mean was greater than 1. In these cases, a diagnosis was attempted with a m a jority rules voting approach[15] by considering only the direction (and not the magnitude) of all 15 individual tissue ratios relative to the threshold value. The test set of samples was analyzed in exactly the same manner with identical Affymetrix probe set identifiers. The classification accuracy of the model in a subset of the test set and in FNA samples was assessed with an exact 1-sample binomial test. The *P* values are reported under the null hypothesis of differential diagnosis randomly assigned with equal probability of 0.5 based on 1-sided tests to reject lower levels of accuracy. The 95% confidence interval (CI) for proportions is based on the exact binomial distribution. All calculations and statistical comparisons were generated with S-PLUS20, except the exact binomial procedures, which were computed with Stata 7 (Stata Corporation, College Station, Tex).

Real-time Quantitative Reverse Transcriptase–Polymerase Chain Reaction

Real-time quantitative reverse transcriptase–polymerase chain reaction (RT-PCR) was performed as described with 2 μg of total RNA.[16] Primer sequences (synthesized by Invitrogen) used for RT-PCR were as follows (forward and reverse, respectively): *MFAP4* (5′-ACTTCTCCATCTCCCCGAAC-3′ and 5′-TGGTAGGACAGGGAGTCACC-3′), *PRDX2* (5′-AGACAATGGAATGGCAGCTT-3′ and 5′-TGCCCAGAAGTGGCATTAGT-3′), *AGER* (5′-TCCACTGGATGAAGGATGGT-3′ and 5′-CAGCTGTAGGTTCCCTGGTC-3′) and *SSR4* (5′-GGAGCAGGATGCGTATAGGA-3′ and 5′-TCTGACTGCACAGATTCTTGG-3′).

Results

Identification of Predictor Genes and Generation of a Model for the Differential Diagnosis of Lung Nodules With Gene Expression Ratios

We discovered a total of 39 predictor genes that fit the filtering criteria and were chosen for further analysis (Table 1). We calculated and assessed for classification accuracy a total of 16 possible individual gene pair ratios for all training subsets except normal lung. We calculated 12 possible gene pair ratios for this subset, since only three genes (not four) were expressed at relatively higher levels in all abnormal tissues. We have previously shown that optimal classification accuracy with expression ratio-based methods can be achieved in most circumstances by combining the expression data from the three most accurate individual gene pair ratios.[15,16] We therefore obtained a combined score (geometric mean, see Methods section) for each of the training subset samples with the three most accurate ratios from each training subset. We found that we could identify these samples with very high accuracy (Table 2).

Finally, to systematically and comprehensively analyze the entire cohort of training set samples in parallel, we calculated all 15 ratios (from Table 2) for every training set sample (n = 145 total) and predicted a tissue type according to the criteria stated in the Methods section. Not surprisingly, we discovered that expression ratio diagnosis with these 23 genes was highly accurate: 90% (130/145, 95% CI 84%–94%, $P < 10^{-6}$) of the samples were correctly predicted with 5 errors and 10 no-calls. The no-calls included adenocarcinoma (n = 6), pulmonary carcinoid (n = 1), and SCC (n = 3). Importantly, no tumors were called normal. Of the 13 initial

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

no-calls, 3 were resolved on further analysis (as described in the Methods section) and 10 samples remained without definitive diagnosis. The 10 final no-call samples had very low, or more frequently undetectable, expression levels for multiple genes used in the analysis, which likely reflects microarray defects or artifacts related to sample preparation. Interestingly, we found that12% of samples (17/145) had multiple (n = 2 in all cases) combined scores greater than 1. Of these, 88% (15/17) were called correctly, with an average 7-fold difference between both combined scores (range 1.3–46.8), with the larger of the two combined scores used to assign a diagnosis.

Verification of Expression Level Ratios As a Diagnostic Tool

Next we tested the ability of these five highly accurate expression ratio combinations to diagnose cancer in a separate cohort of 113 primary tumors and 9 metastatic tumors (the test set) for which expression profiling data were available.[19] A total of 26 samples (n = 13 each of primary lung adenocarcinoma and lung SCC) were directly relevant to the validation of the model developed here because they were obtained from primary lung lesions. The remaining tumors were adenocarcinomas originating from tissues other than lung or represented metastatic disease, and we used these samples to test multiple hypotheses. We first hypothesized that adenocarcinomas of diverse origin are more similar to one another than to any of the other four tissue types examined in this study with respect to global gene expression patterns, specifically the 23 genes used in the expression ratio diagnostic model. We also hypothesized that the diagnostic model developed here would be equally applicable in analyzing metastatic tumors. To perform this analysis, we used the expression values for all 23 diagnostic genes to calculate the five most accurate 3-ratio combinations and predicted the identity of all 122 samples with exactly the same criteria as before. In this analysis, the classification accuracy for all adenocarcinomas was evaluated without respect to tissue type of origin. The results for the classification of primary tumors (n = 113) are presented in Table 3. Overall, our model was 88% (107/122, 95% CI 81%–93%, $P < 10^{-6}$) accurate in identifying the tumor type of test set samples and was 88% (100/113, 95% CI 81%–94%, $P < 10^{-6}$) and 78% (7/9, 95% CI 40%–97%, $P = .090$) accurate within the subsets of primary and metastatic tumors, respectively. Specifically, we found that we could accurately (26/28 or 93%, 95% CI 76%–99%) and significantly ($P = 2 \times 10^{-6}$) predict the identity of primary lung tumors and successfully diagnose both metastatic lung tumors.

Analysis of Ex Vivo FNA Samples

There is ample evidence that material obtained from FNA is sufficient for both microarray and RT-PCR analysis.[12,21–24] We therefore performed a study to test directly the potential of gene ratios to accurately detect cancer in FNA-derived samples. We used simulated ex vivo FNAs in these initial studies to determine whether sample-acquisition procedures introduced variability and possibly affected the outcome of the test. Pathologic analysis of a portion of the solid specimen classified these 15 nodules as non–small cell lung cancer (n = 6), SCC (n = 3), SCLC (n = 1), pulmonary carcinoid (n = 2), bronchoalveolar carcinoma (n = 1), adenocarcinoma (n = 1), and benign fibrosis (n = 1). With quantitative RT-PCR, we obtained gene expression data for the four genes comprising the normal lung versus lung tumor test (from Table 2) from each FNA specimen. Then we calculated all three ratios in this test and determined whether each individual sample was normal lung or lung tumor, according to the criteria detailed here. We found that ratio-based testing could accurately identify FNA samples ($P = 10^{-6}$) from nonmalignant specimens (87% specificity, 13/15, 95% CI 60%–98%) and tumor specimens (100% sensitivity, 14/14, (95% CI 81%–100%). Importantly, neither of the two misclassified samples were tumor tissues (false-negative results). Interestingly, one preoperatively suspected tumor FNA sample was predicted to be nonmalignant (normal lung) by the ratio-based cancer detection test. On final pathologic review, it was determined that this nodule was in fact a fibrotic mass and not a tumor.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Discussion

In this report, we applied the gene ratio technique[13–17] to the detection and diagnosis of lung cancer with a combined approach consisting of multiple data-acquisition platforms and sample sources. The primary strength of this technique is that quantitative RT-PCR can be used to validate the model in additional samples and without further reference to training samples with data acquired on the same platform. In this way, it becomes possible for other clinical investigators without access to complex bioinformatics tools to reproduce initial findings in large numbers of samples worldwide. Here for the first time we extend this technique to the classification of more than two tissue types and provide evidence strongly suggesting that this technique is equally useful in the analysis of FNA-derived material.

Strategies to reduce mortality from lung cancer include the development and implementation of an effective screening system such as spiral CT for at-risk populations. Even as this technique is being studied, it is also being rapidly implemented by physicians, in many cases on demand by patients willing to bear the cost. Spiral CT of the chest can be excessively sensitive, and it is generally estimated that only 10% of nodules detected in the lungs of smokers are actually cancerous.[3,4,25] FNA of newly discovered pulmonary nodules is an attractive technique, but unfortunately is currently limited by the size of the nodule and the accuracy of cytopathologic examination, specifically in distinguishing between false- and true-negative results, which may account for up to a third of all biopsy specimens.[7] The major problem is that a negative cytologic result is simply a negative result in the majority of cases. This is often due to inadequate sampling or lack of sufficient cytologic features to call the sample a tumor.[8] The gene ratio method can potentially address several of these clinical insufficiencies. For example, it could add a genomic component to the diagnosis that requires only the extraction of very small quantities of tumor RNA (tissue), thereby facilitating the acquisition of samples that would otherwise not demonstrate cytologically diagnostic tumor cells. Also, in the concept of gene ratio-based analysis, a diagnosis of nonmalignant is actually a positive diagnosis of benign tissue and not necessarily just a negative result. In our studies, both of the misclassified FNA samples proved to be normal lung, a disappointing finding considering that the virtue of FNA cytopathologic examination is its low false-positive rate. Since samples of normal lung tissue were harvested from the same patient in an area proximal to the suspected tumor, the margin could have been contaminated with histologically undetectable tumor cells, or transformed epithelial cells that have not yet formed a tumor. Alternatively, the misclassification could have resulted from inherent biologic variability reflected in gene expression. Unfortunately, as in similar pilot studies,[12] sufficient material was not available to conduct cytologic analyses, which might have addressed some of these possibilities. To examine all these hypotheses systematically, we are conducting additional studies to refine the list of discriminating genes and prospectively obtaining consent and collecting FNA material linked to cytologic findings for use in follow-up studies.

Our experiments used an ideal scenario (an ex vivo FNA) to test the ability of multiple distinguishing genes to classify normal and malignant tissues accurately in the context of a gene expression ratio-based model. Even though the syringe, needle gauge, and biopsy technique were all similar to those typically used by cytopathologists at our institution, before implementation this technique will require rigorous testing to take into account additional clinical parameters, such as patient movement. Considering that the ex vivo FNA specimen was acquired through the surrounding pulmonary parenchyma and was still accurate at detecting tumor, we believe that the genes as reported will be suitable for use in actual FNA specimens. Recent work by other investigators has demonstrated the general feasibility of using transthoracic CT–guided FNA biopsy to obtain material with RNA suitable for even stringent applications, such as gene expression profiling with microarrays.[12]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

The genes used in this study have also been partially validated by another group of investigators who used a single gene pair ratio (*RAGE*/cyclin-B2) to detect lung cancer.[26] We independently found that *RAGE* (also known as *AGER*) is overexpressed in normal lung relative to tumor, and thus it is part of our normal lung test (Table 2). We also found that cyclin-B2 was statistically significantly ($P = .013$) downregulated in normal lung relative to tumors. However, cyclin-B2 was not among our final list of discriminating genes, probably because of fundamental differences in experimental designs, because we examined a broader number of tumor types and used multiple genes and ratios.

In conclusion, we have produced evidence strongly suggesting that FNA specimens are suitable for gene ratio-based detection and diagnosis of lung cancer, and we are now conducting prospective studies to validate these initial proof-of-principle experiments. We ultimately view this technique as an adjunct and extension to current cytopathologic techniques in the evaluation of suspect lung nodules. Whereas cytopathologists require the preservation of tissue architecture and intact cells for definitive diagnosis, our proposed analysis only requires intact tumor RNA. Furthermore, other gene ratio-based tests, such as for the prognosis of lung cancer,[14] may also be applicable to the analysis of FNA specimens to aid in tailoring the best therapy for the patient in whom cancer is detected and diagnosed. This approach may ultimately allow clinicians to tailor the therapy of individual cancer patients.[20]

Acknowledgements

Supported in part by grants to G.J.G. from the Cancer Research and Prevention Foundation and the National Cancer Institute (CA105249) and to R.B. from the Brigham Surgical Group Foundation, The Milton Fund of Harvard Medical School, CIMIT, and the National Cancer Institute (CA098501).

References

1. Bueno R, Wain JC, Wright CD, Moncure AC, Grillo HC, Mathisen DJ. Bronchoplasty in the management of low-grade airway neoplasms and benign bronchial stenoses. Ann Thorac Surg 1996;62:824–8. [PubMed: 8784013]
2. Solaini L, Bagioni P, Prusciano F, Di Francesco F, Poddie DB. Video-assisted thoracic surgery (VATS) lobectomy for typical bronchopulmonary carcinoid tumors. Surg Endosc 2000;14:1142–5. [PubMed: 11148784]
3. Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, et al. Early lung cancer action project: a summary of the findings on baseline screening. Oncologist 2001;6:147–52. [PubMed: 11306726]
4. Henschke CI, Naidich DP, Yankelevitz DF, McGuinness G, McCauley DI, Smith JP, et al. Early lung cancer action project: initial findings on repeat screenings. Cancer 2001;92:153–9. [PubMed: 11443621]
5. Gorlova O, Kimmel M, Henschke C. Modeling of long-term screening for lung carcinoma. Cancer 2001;92:1531–40. [PubMed: 11745232]
6. Miettinen O, Henschke C. CT screening for lung cancer: coping with nihilistic recommendations. Radiology 2001;221:592–6. [PubMed: 11719649]
7. Calhoun P, Feldman P, Armstrong P, Black W, Pope T, Minor G, et al. The clinical outcome of needle aspirations of the lung when cancer is not diagnosed. Ann Thorac Surg 1986;41:592–6. [PubMed: 3718033]
8. Silverman J. Inflammatory and neoplastic processes of the lung: differential diagnosis and pitfalls in FNA. Diagn Cytopathol 1995;13:448–62. [PubMed: 8834320]
9. Aberle D, Gamsu G, Henschke C, Naidich D, Swensen S. A consensus statement of the Society of Thoracic Radiology: screening for lung cancer with helical computed tomography. J Thorac Imaging 2001;16:65–8. [PubMed: 11149694]
10. Beer DG, Kardia SL, Huang CC, Giordana TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24. [PubMed: 12118244]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

11. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 2002;99:13784–9.
12. Borczuk AC, Shah L, Pearson GD, Walter KL, Wang L, Austin JH, et al. Molecular signatures in biopsy specimens of lung cancer. Am J Respir Crit Care Med 2004;170:167–74. [PubMed: 15087295]
13. Gordon GJ, Rockwell GN, Godfrey PA, Jensen RV, Glickman JN, Yeap BW, et al. Validation of genomics-based prognostic tests in malignant pleural mesothelioma. Clin Cancer Res 2005;11:4406–14. [PubMed: 15958625]
14. Gordon GJ, Richards WG, Sugarbaker DJ, Jaklitsch MT, Bueno R. A prognostic test for adenocarcinoma of the lung from gene expression profiling data. Cancer Epidemiol Biomarkers Prev 2003;12:905–10. [PubMed: 14504202]
15. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswami S, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 2002;62:4963–7. [PubMed: 12208747]
16. Gordon GJ, Hsiao LL, Jensen RV, Gullans SR, Blumenstock JE, Richards WG, et al. Using gene expression ratios to predict outcome among patients with mesothelioma. J Natl Cancer Inst 2003;95:598–605. [PubMed: 12697852]
17. Bueno R, Loughlin KR, Powell MH, Gordon GJ. A diagnostic test for prostate cancer from gene expression profiling data. J Urol 2004;171:903–6. [PubMed: 14713850]
18. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma sub-classes. Proc Natl Acad Sci U S A 2001;98:13790–5. [PubMed: 11707567]
19. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, et al. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 2001;61:7388–93. [PubMed: 11606367]
20. Venables, WN.; Riley, BD. Modern applied statistics with S-Plus. New York: Springer; 1997.
21. Crnogorac-Jurcevic T, Efthimiou E, Capelli P, Blaveri E, Baron A, Terris B, et al. Gene expression profiles of pancreatic cancer and stromal desmoplasia. Oncogene 2001;20:7437–46. [PubMed: 11704875]
22. Assersohn L, Gangi L, Zhao Y, Dowsett M, Simon R, Powles TJ, et al. The feasibility of using fine needle aspiration from primary breast cancers for cDNA microarray analyses. Clin Cancer Res 2002;8:794–801. [PubMed: 11895911]
23. Dey P, Luthra UK, George SS, Prasad A. C-erbB-2 expression and DNA ploidy in breast cancer on fine-needle aspiration cytology material. Diagn Cytopathol 2000;22:261–2. [PubMed: 10787152]
24. Sotiriou C, Powles TJ, Dowsett M, Jazaeri AA, Feldman AL, Assersohn L, et al. Gene expression profiles derived from fine needle aspiration correlate with response to systemic chemotherapy in breast cancer. Breast Cancer Res 2002;4:R3. [PubMed: 12052255]
25. Sagawa M, Saito Y, Sato M, Sugita M, Endo C, Takahashi S, et al. Lung cancer screening—its efficacy and limitations. Gan To Kagaku Ryoho 2002;29:1894–9. [PubMed: 12465387]
26. Hofmann HS, Hansen G, Burdach S, Bartling B, Silber RE, Simm A. Discrimination of human lung neoplasm from normal lung by two target genes. Am J Respir Crit Care Med 2004;170:516–9. [PubMed: 15331390]

Abbreviations and Acronyms

CI
: confidence interval

CT
: computed tomography

FNA
: fine-needle aspiration

RT-PCR

reverse transcriptase–polymerase chain reaction

SCC

squamous cell carcinoma

SCLC

small cell lung cancer

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 1

Lung nodule diagnostic genes

GenBank accession no.	*P* value	Ratio[*]	Description (LocusLink ID)
			Normal lung[†]
L38486	6.7×10^{-16}	179	Microfibrillar-associated protein 4 *(MFAP4)*
AI312905	1.6×10^{-14}	439	*EST*
M91211	7.8×10^{-14}	15	Advanced glycosylation end product-specific receptor *(AGER)*
D88587	2.5×10^{-11}	14	Ficolin *(FCN3)*
J04111	4.4×10^{-7}	0.35	v-*jun* Avian sarcoma virus 17 oncogene homolog *(JUN)*
L19185	2.7×10^{-5}	0.34	Peroxiredoxin 2 *(PRDX2)*
Z69043	9.5×10^{-5}	0.46	Signal sequence receptor, delta *(SSR4)*
			Lung adenocarcinoma
AB000712	2.4×10^{-9}	94	Claudin 4 *(CLDN4)*
AF001294	4.0×10^{-9}	3.6	Tumor suppressing subtransferable candidate 3 *(TSSC3)*
J05581	1.7×10^{-7}	4.1	Mucin 1, transmembrane *(MUC1)*
M18728	3.0×10^{-7}	5.4	Carcinoembryonic antigen-related cell adhesion molecule 6 *(CEACAM6)*
Z78388	7.3×10^{-10}	0.31	Neuronal protein *(NP25)*
M25756	1.2×10^{-7}	0.03	Secretogranin II *(SCG2)*
U15979	1.4×10^{-7}	0.02	Delta-like homolog, *Drosophila (DLK1)*
L07335	2.1×10^{-7}	0.22	Sex-determining region Y-box 2 *(SOX2)*
			Lung squamous cell carcinoma
M21389	8.2×10^{-11}	30	Keratin 5 *(KRT5)*
Y16961	1.8×10^{-8}	23	Tumor protein 63 kd with strong homology to p53 *(TP63)*
L42611[‡]	3.9×10^{-8}	18	Keratin 6A *(KRT6A)*
U97105	8.7×10^{-7}	0.37	Dihydropyrimidinase-like 2 *(DPYSL2)*
AF004563	1.7×10^{-5}	0.48	Syntaxin binding protein 1 *(STXBP1)*
AF042792	8.0×10^{-5}	0.32	Calcium channel, voltage-dependent *(CACNA2D2)*
U43203	1.8×10^{-4}	0.26	Thyroid transcription factor 1 *(TITF1)*
			Small cell lung cancer
D82345	1.6×10^{-8}	16	Thymosin, beta *(TMSNB)*
AA203476	5.6×10^{-8}	9	Pituitary tumor transforming 1 *(PTTG1)*
U73379	1.1×10^{-7}	17	Ubiquitin-conjugating enzyme E2C *(UBE2C)*
D00762	6.6×10^{-7}	11	Proteasome subunit, alpha type, 3 *(PSMA3)*
L25080	8.4×10^{-5}	0.28	*ras* Homolog gene family, member A *(ARHA)*
X05409	8.0×10^{-4}	0.2	Aldehyde dehydrogenase 2 *(ALDH2)*
X59798	.002	0.24	Cyclin D1 *(CCND1)*
X04366	.008	0.25	Calpain 1 *(CAPN1)*
			Pulmonary carcinoid
L18983	$1.6\ 3\ 10^{-10}$	11	Protein tyrosine phosphatase, receptor type, N *(PTPRN)*
AB014558	$7.8\ 3\ 10^{-10}$	102	Cryptochrome 2, photolyase-like *(CRY2)*
U96750	$4.6\ 3\ 10^{-7}$	350	*ras* Homolog gene family, member I *(ARHI)*
Y00064	$5.8\ 3\ 10^{-7}$	118	Chromogranin B *(CHGB)*
AA203487	$5.4\ 3\ 10^{-8}$	0.03	CD68 antigen *(CD68)*
AA631972	$1.1\ 3\ 10^{-7}$	0.01	Natural killer cell transcript 4 *(NK4)*
X67325	$1.2\ 3\ 10^{-7}$	0.15	Interferon, alpha-inducible protein 27 *(IFI27)*
X62744	$3.5\ 3\ 10^{-7}$	0.01	Major histocompatibility complex, class II, DM alpha *(HLA-DMA)*

Eight diagnostic genes for each tissue type were chosen for further study. Four genes were expressed at relatively higher levels in a single tissue type, and four genes were expressed at relatively higher levels in an equal number of all other tissue types combined.

[*] Average expression level in a single tissue type/average expression level in all other tissue types combined.

[†] Only three genes (not four) were overexpressed in all abnormal tissues and fit the filtering criteria.

[‡] This gene was represented by multiple probe sets on the expression profiling platform of the training set (see Methods section).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

TABLE 2

Diagnostic accuracy of 3-ratio combinations in training subsets

3-Ratio test*	Accuracy
Normal lung: *MFAP4\PRDX2, AGER\PRDX2, AGER\SSR4*	100% (41/41)
Lung adenocarcinoma: *CLDN4\SOX2, CLDN4\NP25, TSSC3\SOX2*	89% (104/117)
Lung SCC: *KRT5\CACNA2D2, KRT6A\TITF1, KRT6A\STXBP1*	90% (65/72)
SCLC: *PTTG1\CCND1, PSMA3\ALDH2, TMSNB\ALDH2*	100% (55/55)
Pulmonary carcinoid: *PTPRN\NK4, CRY2\CD68, CRY2\HLA-DMA*	98% (63/64)

The three most accurate individual ratios identified in each training subset were combined and used to obtain a diagnostic call for samples contained within each subset.

*Genes are represented by LocusLink symbols found in Table 1.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 3

Distribution of class predictions for test set samples

Actual identity of unknown sample	Predicted class*					
	LADCA	LSCC	SCLC	PC	NL	No call
LADCA	12	1				
LSCC	1	12				
PR	24					
CO	19	1				
BR	20	1	1	3		
GA	9	1	1	1		
PA	4	1		1		

The value in each box is the number of primary tumors from the test set (n = 113) predicted with a given identity from the expression ratio-based model developed in the training set of samples. *LADCA*, Lung adenocarcinoma; *LSCC*, lung SCC; *PC*, pulmonary carcinoid; *NL*, normal lung; *PR*, prostate adenocarcinomas; *CO*, colon adenocarcinomas; *BR*, breast adenocarcinomas; *GA*, gastroesophageal adenocarcinomas; *PA*, pancreatic adenocarcinomas.