nature medicine

Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer

Avrum Spira[1], Jennifer E Beane[2,8], Vishal Shah[2,8], Katrina Steiling[1], Gang Liu[1], Frank Schembri[1], Sean Gilman[3], Yves-Martine Dumas[1], Paul Calner[4], Paola Sebastiani[5], Sriram Sridhar[1], John Beamis[3], Carla Lamb[3], Timothy Anderson[6], Norman Gerry[7], Joseph Keane[4], Marc E Lenburg[7] & Jerome S Brody[1]

Lung cancer is the leading cause of death from cancer in the US and the world[1]. The high mortality rate (80–85% within 5 years) results, in part, from a lack of effective tools to diagnose the disease at an early stage[2–4]. Given that cigarette smoke creates a field of injury throughout the airway[5–11], we sought to determine if gene expression in histologically normal large-airway epithelial cells obtained at bronchoscopy from smokers with suspicion of lung cancer could be used as a lung cancer biomarker. Using a training set ($n = 77$) and gene-expression profiles from Affymetrix HG-U133A microarrays, we identified an 80-gene biomarker that distinguishes smokers with and without lung cancer. We tested the biomarker on an independent test set ($n = 52$), with an accuracy of 83% (80% sensitive, 84% specific), and on an additional validation set independently obtained from five medical centers ($n = 35$). Our biomarker had ~90% sensitivity for stage 1 cancer across all subjects. Combining cytopathology of lower airway cells obtained at bronchoscopy with the biomarker yielded 95% sensitivity and a 95% negative predictive value. These findings indicate that gene expression in cytologically normal large-airway epithelial cells can serve as a lung cancer biomarker, potentially owing to a cancer-specific airway-wide response to cigarette smoke.

Physicians increasingly encounter current and former smokers with clinical suspicion for lung cancer, on the basis of abnormal radiographic imaging and/or symptoms. Flexible bronchoscopy is a relatively noninvasive initial diagnostic test to use in this setting, involving cytologic examination of materials obtained from endobronchial brushings, bronchoalveolar lavage and endo- and transbronchial biopsies of the suspect area[12,13]. The sensitivity of bronchoscopy for lung cancer ranges from 30% for small peripheral lesions to 80% for centrally located endobronchial disease[14]. As a result, most patients require further invasive diagnostic tests, which delay treatment (median delay 3–7 months from first symptoms to diagnosis[15]) and generate additional costs and risks for complications.

Cigarette smoke creates a field of injury in all airway epithelial cells exposed to it. Previous studies have shown that noncancerous large-airway epithelial cells of current and former smokers with and without lung cancer exhibit allelic loss[6,7], p53 mutations[8], changes in promoter methylation[9] and increased telomerase activity[10]. Using DNA microarrays, we recently described smoking-induced changes in the gene expression of large-airway epithelial cells obtained during bronchoscopy from nonsmokers and from current and former smokers without lung cancer[11]. These studies led us to question whether profiles of gene expression in large-airway epithelial cells could provide insights into how individual smokers differ in their responses to cigarette smoke and whether such profiling might detect smokers in whom the mutagenic effects of cigarette smoke have resulted in lung cancer (given that only 10–15% of smokers develop lung cancer). A lung cancer diagnostic using this approach might eliminate the need for additional diagnostic tests that are costly, incur risk and prolong the diagnostic evaluation of suspect lung cancer patients.

Using Affymetrix HG-U133A microarrays, we performed gene-expression profiling of large-airway epithelial cell brushings obtained from current and former smokers who underwent flexible bronchoscopy, as a diagnostic study for clinical suspicion of lung cancer, between January 2003 and April 2005. Each individual was followed after bronchoscopy until a final diagnosis of lung cancer or not lung cancer was made. In our primary analysis, we included 129 subjects (60 smokers with lung cancer and 69 smokers without lung cancer) who had achieved final diagnoses as of May 2005 and had high-quality microarray data (**Supplementary Tables 1** and **2** online). Bronchial brushings yielded 90% epithelial cells, with the majority being ciliated or basal cells; no dysplastic or cancer cells were seen in representative brushings and there was no difference in the proportion of inflammatory cells between smokers with and without lung cancer (data not shown).

[1]The Pulmonary Center, Boston University Medical Center, 715 Albany Street, Boston, Massachusetts 02118, USA. [2]Bioinformatics Program, Boston University, 44 Cummington Street, Boston, Massachusetts 02215, USA. [3]Pulmonary Division, Department of Medicine, Lahey Clinic, 41 Mall Road, Burlington, Massachusetts 01805, USA. [4]St. James's Hospital and Trinity College, James's Street, Dublin, Ireland. [5]School of Public Health, Boston University, 715 Albany Street, Boston, Massachusetts 02118, USA. [6]Caritas St. Elizabeth's Medical Center and Tufts University School of Medicine, 736 Cambridge Street, Boston, Massachusetts 02135, USA. [7]Department of Genetics and Genomics, Boston University, 715 Albany Street, Boston, Massachusetts 02118, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to A.S. (aspira@bu.edu).

Received 5 August 2006; accepted 30 January 2007; published online 4 March 2007; doi:10.1038/nm1556

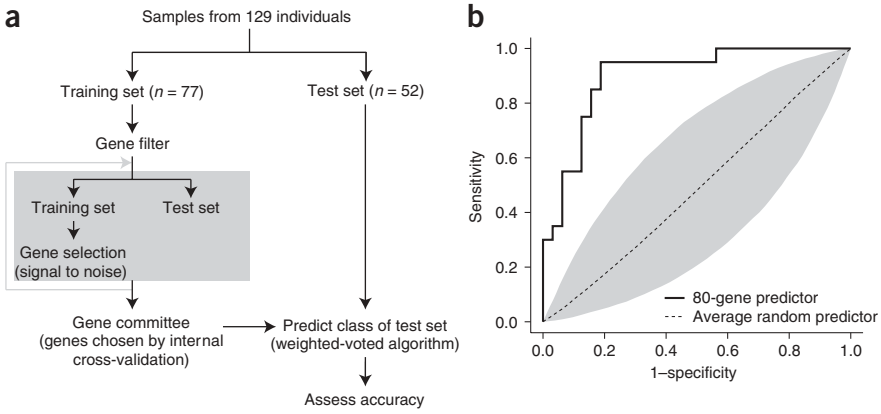

Figure 1 Development and performance of an airway biomarker for lung cancer. (**a**) Class prediction methodology. A total of 129 samples were separated into a training set and a test set. For the final gene committee, the most frequently chosen 40 upregulated and 40 downregulated probesets were selected by internal cross-validation using the training set samples. The weighted-voting algorithm using this committee was then used to predict the cancer status of independent test set samples that were not used for any part of the predictor discovery process. (**b**) Biomarker performance. The biomarker's sensitivity was determined as a function of specificity. For this analysis, noncancer predictions were multiplied by –1 to create a continuous scale. The solid black line represents the performance of the airway gene-expression biomarker on the test samples. The dotted black line represents the average performance of 1,000 biomarkers derived from training sets in which we randomized the cancer status of the samples. The upper and lower bounds of the shaded region represent the average performance for the top and bottom half of random biomarkers (based on area under the curve, AUC). There was a significant difference between the AUC of the actual biomarker and that of the random biomarkers ($P = 0.004$).

To develop a gene-expression biomarker of lung cancer, 60% of samples ($n = 77$), representing a spectrum of clinical risk for lung cancer, were randomly assigned to a training set (**Supplementary Methods** online); the remaining 52 samples were used as an independent test set. Using only the training set samples and a weighted-voting algorithm[16], we identified an 80-probeset biomarker that distinguished smokers with and without lung cancer in the training set (**Fig. 1** and **Supplementary Methods**). The accuracy, sensitivity and specificity of this biomarker on the independent test set samples were 83% (43 of 52), 80% (16 of 20) and 84% (27 of 32), respectively. The accuracy of the biomarker was independent of tumor location within the lung relative to the site of bronchial brushing (data not shown). Expression levels of the biomarker genes showed largely consistent differences between individuals with and without lung cancer (**Fig. 2a** and **Supplementary Table 3** online). Principal component analysis (PCA) of cancer samples according to the expression of these 80 probesets did not reveal cell type–specific or stage-specific gene-expression differences (data not shown). The accuracy of this biomarker was similar when used with expression levels derived from probe-level data using the MAS 5.0 algorithm (as opposed to Robust Multichip Average (RMA) algorithm) or when the Prediction Analysis for Microarrays (PAM) class prediction algorithm[17] was used (**Supplementary Methods**). The differential expression of seven genes in the biomarker was confirmed by RT-PCR (**Supplementary Fig. 1** online), and epithelial cell localization for two biomarker genes (*IL8* and *CD55*) was confirmed by immunohistochemistry (**Supplementary Fig. 2** online).

To evaluate the robustness of the biomarker, we compared the 80-probeset classifier to three different types of randomized classifiers (**Supplementary Methods**). The performance of the biomarker in classifying test set samples was significantly better than that of classifiers derived from training sets with randomized disease status labels ($P = 0.004$; **Fig. 1b**) or classifiers composed of randomly selected probesets ($P = 0.007$; **Supplementary Table 4** online). In addition, gene-expression differences related to differences in cumulative smoking history between smokers with and without lung cancer did not contribute to the biomarker's accuracy (**Supplementary Methods**). Finally, biomarker performance was insensitive to the particular composition of the training or test set, as 1,000 different training and test sets derived from the 129-sample set produced biomarkers with similar performance (**Supplementary Table 4** and **Supplementary Fig. 3** online).

As the test set samples used to determine biomarker accuracy were collected at the same institutions and during the same time period as the training samples used to derive the biomarker, we tested the biomarker on an independently collected prospective validation set ($n = 40$). Five of these samples did not pass our array quality filter and were excluded from further analysis. In this validation set, there were no significant differences in age or cumulative tobacco exposure between smokers with and without cancer ($P > 0.05$; **Supplementary Table 1**), and the set also included samples from an additional medical center. The biomarker accurately classified 28 of 35 (80%) samples from the validation set (83% sensitive and 76% specific), and the expression of biomarker probesets in these samples was similar to that in the original test set (**Fig. 2b**). We also determined the biomarker's diagnostic yield on all individuals recruited into the prospective series regardless of the quality of samples obtained from them (**Supplementary Methods**).

To investigate whether the biomarker genes identified in cytologically normal large-airway epithelial cells are differentially expressed in actual lung cancer tissue, we evaluated the biomarker for its ability to distinguish between normal and cancerous lung tissue in two previously published microarray datasets[18,19]. In one[18], the airway biomarker classified normal lung tissue from smokers without cancer and lung tumor tissue from smokers with 90% accuracy (**Supplementary Methods**). PCA also revealed differences in gene expression across the biomarker probesets between normal and tumor tissue in this dataset ($P = 0.026$; **Fig. 3**). In the other dataset, which contained samples of lung tumors from smokers with squamous cell carcinoma and histologically normal samples from lung tissue adjacent to these tumors[19], all samples were classified as being from smokers with cancer; moreover, the expression of biomarker probesets was similar between tumor and adjacent normal tissue samples (**Supplementary Methods**). In addition, we tested our biomarker on two other large U133A microarray datasets of lung cancer samples and correctly classified 99% (129 of 130) of samples from one dataset[20] and 90% (178 of 198) from the other[21].

To determine whether the large-airway gene-expression biomarker offers a diagnostic advantage over traditional cytopathology of cells obtained at bronchoscopy, we compared the accuracy of these tests and also investigated the accuracy of a diagnostic that combines the results from both tests. Bronchoscopy, in which the cytopathology of cells obtained from endoscopic brushing, washings or biopsy of the

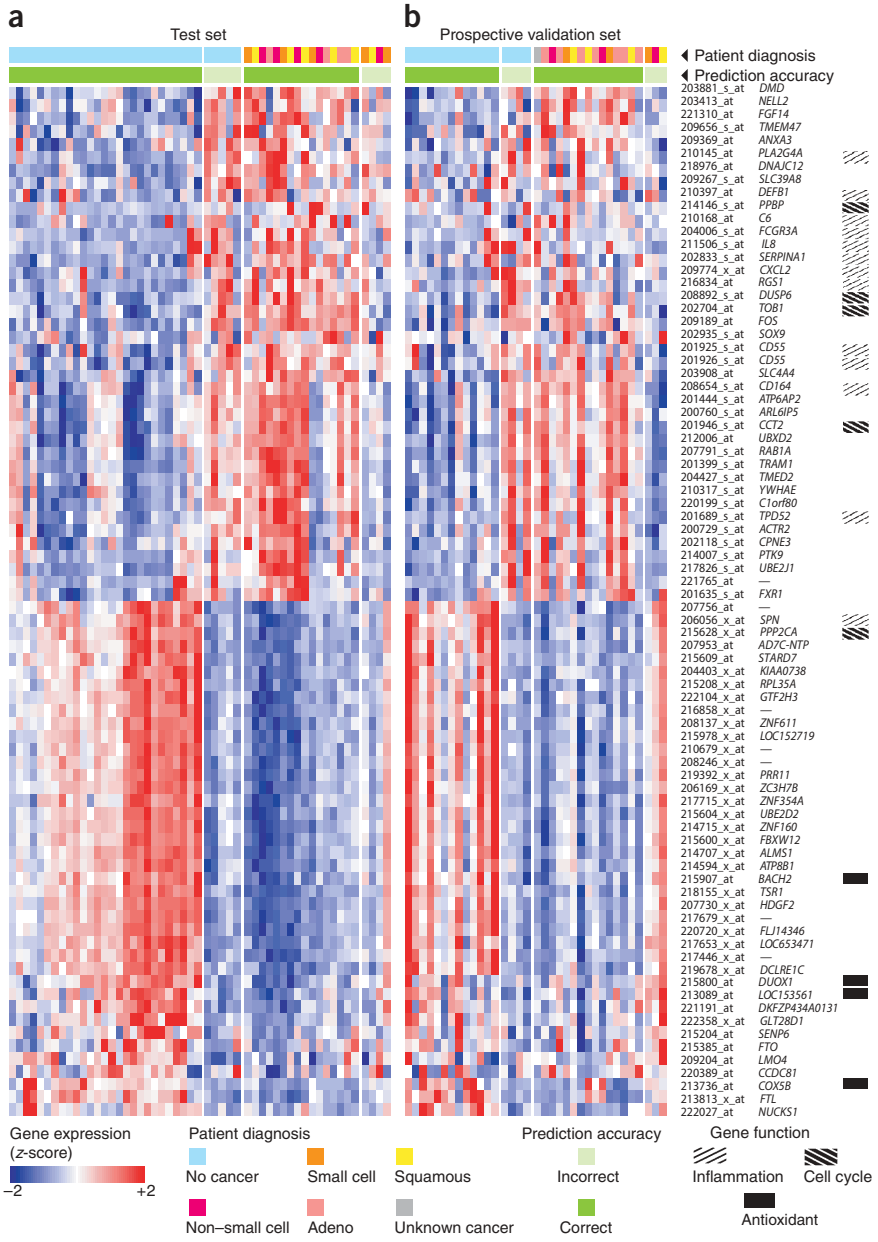

Figure 2 Hierarchical clustering of biomarker probeset expression in two independent test sets. (**a**,**b**) Expression levels of the biomarker probesets in the 52 test set samples and the 35 prospective validation set samples were normalized by *z*-score and are organized from top to bottom by hierarchical clustering. The Affymetrix HG-U133A probeset ID and HUGO symbol are given to the right of each gene along with functional annotation of select genes (cross-hatched boxes). The samples are organized from left to right by diagnosis (that is, whether the patient had a clinical diagnosis of cancer). Within these two groups, the samples are organized by the accuracy of the class prediction (samples classified incorrectly are on the right for each group of patients, shown in light green). Classification was correct for 43 of 52 (83%) test samples and 27 of 35 (80%) prospective validation set samples.

affected region was assessed, diagnosed cancer in 32 of 60 (53%) subjects with lung cancer in the primary dataset and yielded a definitive diagnosis of a noncancer pathology (for example, tuberculosis) in 5 of 69 smokers without lung cancer—meaning that the results of bronchoscopy were not diagnostic of either cancer or a noncancer pathology in 92 samples. Among nondiagnostic bronchoscopies ($n = 92$), the gene-expression biomarker's accuracy was 85% (89% sensitive, 83% specific). By combining the tests such that a diagnosis of lung cancer from either cytopathology or the biomarker indicated lung cancer, we achieved 95% diagnostic sensitivity (57 of 60) across all cancer subjects in the training and test sets. With a disease prevalence of ~50% in this cohort, negative cytopathology and a negative biomarker prediction resulted in a 95% negative predictive value for disease (**Fig. 4a**,**b**). In the prospective validation samples ($n = 35$), bronchoscopy was 44% sensitive. Combining bronchoscopy with the gene-expression biomarker in the prospective validation set samples resulted in 94% (17 of 18) sensitivity, a 93% negative predictive value and an 81% positive predictive value (**Fig. 4c**,**d**).

As the high mortality rate for lung cancer stems at least in part from the failure to achieve early diagnosis, we examined the performance of the biomarker in early-stage disease. In our primary dataset, we found that the biomarker was ~90% sensitive for stage 1 lung cancer whereas the sensitivity of routine bronchoscopic studies was ~35% (**Supplementary Fig. 4** online). In the prospective validation set, the biomarker correctly classified each of seven samples from smokers with stage 1 or stage 2 lung cancer.

In addition to serving as a diagnostic biomarker, airway gene expression in smokers with and without lung cancer can provide insight into the nature of the airway pathophysiology in smokers with lung cancer. The airway biomarker contains genes that are also differentially expressed in lung cancer tissue (**Fig. 3**), even though the predominant epithelial cells of upper airways being sampled in this study (ciliated cells) differ from those in which most lung cancers occur (glandular, squamous and neuroendocrine cells). This suggests that the biomarker measures a common cancer-specific gene-expression pattern that occurs throughout the respiratory tract epithelium. Our finding that tumor location relative to the site of bronchial brushing had no effect on classifier accuracy suggests that the changes in airway gene expression between smokers with and without lung cancer are unlikely to be caused directly by the tumor itself. It is therefore possible that airway cancer-specific gene-expression changes may occur prior to the appearance of frank malignancy.

The notion of a cancer-specific airway-wide response to tobacco smoke is strengthened by the types of genes that comprise the biomarker (**Supplementary Table 3**). Genes functioning in inflammation, cell cycle progression and signaling predominated among those that were upregulated in smokers with cancer, whereas genes involved in antioxidant defense, ubiquitination and DNA repair predominated among those that were downregulated. A number of genes associated with the RAS oncogene pathway, including *RAB1A*

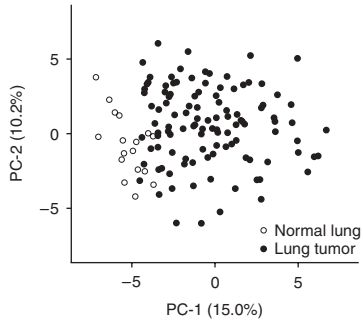

Figure 3 Principal component analysis (PCA) of airway biomarker gene expression in lung tissue samples. The 80 biomarker probesets were mapped to 64 probesets in a HG-U95Av2 microarray dataset of lung cancer and normal lung tissue[18]. The normal lung samples separate from lung cancer samples along the first principal component (*t*-test, *P*-value = 0.026), indicating that cancer status is a major source of variation in the expression of biomarker probesets.

and *FOS* (both implicated in tumorgenesis[22,23]), were upregulated. We found that a number of key antioxidant defense genes were decreased in airway epithelium of smokers with lung cancer, including *BACH2*, which encodes a transcription factor that promotes cell apoptosis in response to oxidative stress[24], and the genes encoding dual oxidase-1 and a DNA repair enzyme, DNA repair protein 1C. The classifier also contained several proinflammatory genes, including those encoding interleukin-8 and β-defensin 1 (both implicated in lung cancer[25,26]), which were upregulated in smokers with lung cancer. Higher levels of these chronic inflammatory mediators may result in increased oxidative stress and contribute to lung tumor promotion and progression[27].

Whereas the biomarker contains a number of inflammatory genes, several lines of evidence suggest that that the biomarker detects cancer-specific gene-expression differences in epithelial cells. Cytologic review of select airway brushings revealed that greater than 90% of cells were epithelial in origin and that there was no difference in the proportion of inflammatory cells between smokers with and without lung cancer. Further, we found that expression levels of inflammatory cell–specific genes did not make a distinction between samples from smokers with and without cancer (**Supplementary Fig. 2**). Finally, previous studies[28] have shown that many of the inflammatory genes in our biomarker are expressed in airway epithelium, and the Human Gene Atlas study[29] has shown that several have relatively high expression levels in bronchial epithelial cells (**Supplementary Methods**). Immunohistochemistry for two of the inflammatory genes in our biomarker demonstrated that they are expressed in airway epithelial cells (**Supplementary Fig. 2**).

In summary, our study has identified an airway gene-expression biomarker that has the potential to have an impact on the diagnostic evaluation of smokers with suspect lung cancer. These individuals often undergo flexible bronchoscopy as an initial diagnostic test, in which the cytopathology of cells obtained from the lower airway is examined. Gene-expression profiling can be performed on histologically normal upper airway epithelial cells obtained at the time of the bronchoscopy in a simple and noninvasive fashion, prolonging the procedure by only 3–5 min. Our data suggest that combining cytopathology with the gene-expression biomarker improves the diagnostic sensitivity of the overall bronchoscopy procedure (from 53% to 95%). In the setting of our study, where disease prevalence was 50%, a negative bronchoscopy and negative biomarker for lung cancer resulted in a 95% negative predictive value, potentially allowing these individuals to be followed nonaggressively with serial imaging studies. For individuals with a negative bronchoscopy and positive gene-expression signature, the positive predictive value was ~70%; these individuals would probably require further invasive testing to confirm the presumptive lung cancer diagnosis. However, compared to bronchoscopy alone, the strong negative predictive value of the combined cytopathology and gene-expression biomarker test should substantially reduce the number of individuals requiring further invasive diagnostic testing.

The notion of a cancer-specific airway-wide injury suggests that cancer-specific alterations in gene expression that occur as a result of smoking might precede the development of lung cancer. If this is true, the lag between alterations in gene expression and the appearance of

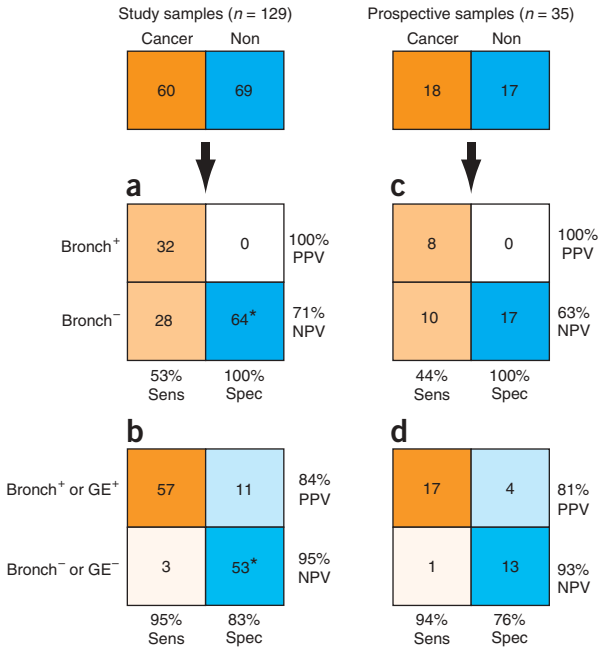

Figure 4 Diagnostic utility of bronchoscopy and the gene-expression biomarker. (**a**) Bronchoscopy results for the individuals in the primary dataset. Only 32 of 60 smokers with lung cancer had bronchoscopies that diagnosed lung cancer. Bronchoscopy resulted in the diagnosis of a noncancer pathology in five samples (excluded from the boxes labeled with an asterisk but included in the calculation of negative predictive value), and was nondiagnostic in the remaining 92 samples. This resulted in 92 smokers for whom further diagnostic tests were required in order to rule in or rule out the presence of lung cancer. (**b**) Combined test results for the primary dataset. A combined test where a cancer diagnosis from either bronchoscopy or gene expression is considered diagnostic of lung cancer achieved a sensitivity of 95% (57 of 60 cancer subjects) with only a 17% false-positive rate (11 of the 64 noncancer subjects). (**c**) Bronchoscopy results for the 35 individuals in the prospective study. Of 18 with lung cancer, 8 had bronchoscopies that were diagnostic of lung cancer. The remaining 27 samples had bronchoscopies that were negative for lung cancer and all other noncancer pathologies. (**d**) Combined test results for the 35 prospective study subjects. Combining bronchoscopy with gene expression resulted in a sensitivity of 94% (17 of 18 cancer subjects). The shading of the contingency table boxes reflects the fraction of each sample type in each quadrant. 'Cancer' and 'Non' headings indicate patients with and without cancer, respectively. Bronch$^+$ and Bronch$^-$, diagnosed as having or not having cancer, respectively, by cytopathology of bronchoscopic material; GE$^+$ and GE$^-$, diagnosed as having or not having cancer, respectively, by the gene-expression biomarker; NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity; Spec, specificity.

lung cancer could contribute to the biomarker's false-positive rate in our cross-sectional study. A longitudinal study will be needed to assess whether false-positive biomarker diagnoses represent smokers at higher risk for developing lung cancer. If this is the case, our biomarker might be useful as a screening tool for lung cancer among healthy smokers and may have the potential to identify high-risk smokers who would derive the most benefit from chemoprophylaxis.

METHODS

Patient population. We recruited current and former smokers ($n = 208$) undergoing flexible bronchoscopy (as a diagnostic study for clinical suspicion of lung cancer) between January 2003 and April 2005 at four institutions: Boston University Medical Center, Boston Veterans Administration, Lahey Clinic and St. James's Hospital (**Supplementary Methods**). We classified subjects as having lung cancer if their bronchoscopy or subsequent lung biopsy yielded lung tumor cells. Subjects were classified to the noncancer group if the bronchoscopy or subsequent lung biopsy yielded a non-lung-cancer pathology or if their radiographic abnormality resolved on follow-up chest imaging. Individuals without final diagnoses as of May 2005 were excluded from this primary dataset. The study was approved by the Institutional Review Boards of all medical centers, and all participants provided written informed consent.

Airway epithelial cell collection. Following routine diagnostic bronchoscopy studies, we obtained bronchial airway epithelial cells from the uninvolved right mainstem bronchus with an endoscopic cytobrush (Cellebrity Endoscopic Cytobrush, Boston Scientific). If a suspicious lesion (endobronchial or submucosal) was seen in the right mainstem bronchus, brushings were obtained from the uninvolved left mainstem bronchus. The brushes were immediately placed in TRIzol reagent (Invitrogen) and stored at –80 °C. RNA was extracted from the brushes using TRIzol per the manufacturer's protocol; RNA yields were 8–15 μg. We confirmed RNA integrity by performing a denaturing gel electrophoresis. We determined the epithelial cell content and morphology of representative samples by cytocentrifugation (ThermoShandon Cytospin) of the cell pellet and cytokeratin antibody staining (Signet); these results were reviewed by a pathologist.

Microarray data acquisition and preprocessing. We processed, labeled and hybridized 6–8 μg of total RNA to Affymetrix HG-U133A GeneChips containing 22,215 probesets as described previously[11]. We obtained sufficient quantities of high-quality RNA for microarray studies from 152 of 208 samples. The quantity of RNA obtained per sample increased during the course of the study (because of an increase in the number of airway brushings performed); 90% of samples from the latter half of the study were included in the microarray analysis. We obtained probe-level data using the Robust Multichip Average (RMA) algorithm[30] to maximize correlation between technical replicates (**Supplementary Methods**). We used a *z*-score metric, which correlates with the percent probesets present metric, to filter out arrays of poor quality (**Supplementary Methods**), leaving 129 samples available for analysis.

Microarray data class prediction analysis. We separated 129 samples (69 from smokers without cancer, 60 from smokers with lung cancer) into a training ($n = 77$) and a test set ($n = 52$) (**Supplementary Methods**). Using only the training set, we identified genes that were differentially expressed in smokers with cancer, using pack-years of cigarette smoke exposure as a covariate to control for differences in cumulative tobacco exposure in smokers with cancer. Genes that were differentially expressed in cancer samples ($P < 0.05$) were selected by internal cross-validation within the training set using the signal-to-noise metric[16]. Internal cross-validation was repeated 50 times, and the most frequently chosen 40 upregulated and 40 downregulated probesets were selected as the final gene committee (**Supplementary Methods**). This committee of 80 probesets was then used to predict the cancer status of independent samples using the weighted-voting algorithm.

Quantitative PCR validation. We used real-time PCR to confirm the differential expression of select biomarker genes (**Supplementary Methods** and **Supplementary Table 5** online).

Immunohistochemistry. Using immunofluorescence, we investigated the cell of origin for two of the biomarker genes, *CD55* and *IL8*. Bronchial brushings from smokers with and without lung cancer were fixed and assayed by indirect immunofluorescence with polyclonal antibodies to human CD55 and IL8 (**Supplementary Methods**).

Prospective validation set. We assembled a second set of samples ($n = 40$) from additional individuals recruited between May 2005 and December 2005 and from individuals whose diagnoses were pending as of May 2005 but became final after that date.

Predictor gene expression in lung tumor tissue. We evaluated the ability of the 80-gene lung-cancer biomarker to distinguish between normal and cancerous lung tissue from smokers, using an Affymetrix HG-U95Av2 dataset[18]. We identified 64 HG-U95Av2 probesets that corresponded to the 80 probesets in our biomarker and used the expression of these probesets to classify the tumor and normal lung samples (**Supplementary Methods**). We also used the 80-gene biomarker to classify samples in a HG-U133A microarray dataset of squamous cell carcinoma lung tumor samples and matching adjacent histologically normal lung from smokers[19]. Finally, we evaluated the ability of the 80-gene biomarker to predict lung cancer in two large HG-U133A microarray datasets containing only lung cancer tissue samples[20,21].

Statistical analysis. All data preprocessing, class prediction and statistical analyses were accomplished using R and BioConductor packages.

Additional information. Probe-level expression data (CEL files) from all microarray samples, probeset expression levels, additional analyses and anonymized patient clinical data (including results from other diagnostic tests) are available at http://pulm.bumc.bu.edu/CancerDx with tools for further analysis.

Accession codes. All microarray data have been submitted to the Gene Expression Omnibus (GEO) under accession number GSE4115.

Note: Supplementary information is available on the Nature Medicine website.

ACKNOWLEDGMENTS

We thank C. O'Hara for histologic review of our airway epithelial cell samples; J. Warrington, J. Palma and R. Lipshutz for their support in designing and implementing this study; M. Klempner and D. Center for their review of the manuscript; F. O'Connell, J. Lundebye and the Lung Cancer Multi-Disciplinary Team at St. James's Hospital; and the doctors and nurses of the bronchoscopy service at Boston Medical Center, St. James's Hospital and Lahey Clinic. Affymetrix Inc. provided the HG-U133A arrays for these studies. This work was supported by the Doris Duke Charitable Foundation (A.S.), US National Institutes of Health/National Institute of Environmental Health Sciences (ES10377 to J.S.B.) and National Institutes of Health/ National Cancer Institute (R21CA10650 to A.S.).

AUTHOR CONTRIBUTIONS

A.S. was responsible for the conception and design of this study and oversaw all aspects of the study including patient recruitment, experimental protocols and data analysis. J.E.B. contributed to the design of the analytic strategy and was responsible for the computational analysis of gene-expression data including preprocessing, class prediction and the connection to tumor tissue. V.S. contributed to the analysis of gene-expression and clinical data and optimization of the class prediction algorithm. K.S. was responsible for patient recruitment and for collection and analysis of clinical data on all subjects in this study. G.L. performed the microarray experiments and real-time PCR studies and was responsible for QRTPCR data analysis. F.S. performed the histologic studies of airway samples and the immunofluorescence studies. S.G. recruited subjects, collected samples and contributed to the analysis of clinical data on all subjects. Y.-M.D. was responsible for coordinating all patient recruitment and sample collection. P.C., J.B., C.L. and T.A. recruited subjects and collected samples at their respective institutions. P.S. contributed to the statistical analysis of the data. S.S. contributed to the development of the relational database. N.G. performed all microarray hybridizations. J.K. recruited subjects, collected samples and provided support in the design of the study. M.E.L. was responsible for conceptualizing many aspects of the analytic strategy and directed the computational analysis. J.S.B. was responsible for the conception and design of the study and oversaw the experimental studies and biological interpretation of the data. A.S., J.E.B., V.S.,

M.E.L. and J.S.B. were responsible for the writing of the manuscript and for the supplementary information.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Medicine* website for details).

Published online at http://www.nature.com/naturemedicine
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Parkin, D.M., Bray, F., Ferlay, J. & Pisani, P. Global cancer statistics, 2002. *CA Cancer J. Clin.* **55**, 74–108 (2005).
2. Hirsch, F.R., Merrick, D.T. & Franklin, W.A. Role of biomarkers for early detection of lung cancer and chemoprevention. *Eur. Respir. J.* **19**, 1151–1158 (2002).
3. Jett, J.R. Limitations of screening for lung cancer with low-dose spiral computed tomography. *Clin. Cancer Res.* **11**, 4988s–4992s (2005).
4. Macredmond, R. *et al.* Screening for lung cancer using low dose CT scanning: results of 2 year follow up. *Thorax* **61**, 54–56 (2006).
5. Auerbach, O., Hammond, E.C., Kirman, D. & Garfinkel, L. Effects of cigarette smoking on dogs. II. Pulmonary neoplasms. *Arch. Environ. Health* **21**, 754–768 (1970).
6. Powell, C.A., Klares, S., O'Connor, G. & Brody, J.S. Loss of heterozygosity in epithelial cells obtained by bronchial brushing: clinical utility in lung cancer. *Clin. Cancer Res.* **5**, 2025–2034 (1999).
7. Wistuba, I.I. *et al.* Molecular damage in the bronchial epithelium of current and former smokers. *J. Natl. Cancer Inst.* **89**, 1366–1373 (1997).
8. Franklin, W.A. *et al.* Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. *J. Clin. Invest.* **100**, 2133–2137 (1997).
9. Guo, M. *et al.* Promoter hypermethylation of resected bronchial margins: a field defect of changes? *Clin. Cancer Res.* **10**, 5131–5136 (2004).
10. Miyazu, Y.M. *et al.* Telomerase expression in noncancerous bronchial epithelia is a possible marker of early development of lung cancer. *Cancer Res.* **65**, 9623–9627 (2005).
11. Spira, A. *et al.* Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. USA* **101**, 10143–10148 (2004).
12. Postmus, P.E. Bronchoscopy for lung cancer. *Chest* **128**, 16–18 (2005).
13. Mazzone, P., Jain, P., Arroliga, A.C. & Matthay, R.A. Bronchoscopy and needle biopsy techniques for diagnosis and staging of lung cancer. *Clin. Chest Med.* **23**, 137–158 (2002).
14. Schreiber, G. & McCrory, D.C. Performance characteristics of different modalities for diagnosis of suspected lung cancer: summary of published evidence. *Chest* **123**, 115S–128S (2003).
15. Salomaa, E.R., Sallinen, S., Hiekkanen, H. & Liippo, K. Delays in the diagnosis and treatment of lung cancer. *Chest* **128**, 2282–2288 (2005).
16. Golub, T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
17. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567–6572 (2002).
18. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98**, 13790–13795 (2001).
19. Wachi, S., Yoneda, K. & Wu, R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**, 4205–4208 (2005).
20. Raponi, M. *et al.* Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res.* **66**, 7466–7472 (2006).
21. Potti, A. *et al.* A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N. Engl. J. Med.* **355**, 570–580 (2006).
22. Cheng, K.W., Lahad, J.P., Gray, J.W. & Mills, G.B. Emerging role of RAB GTPases in cancer and human disease. *Cancer Res.* **65**, 2516–2519 (2005).
23. Shimada, K. *et al.* Aberrant expression of RAB1A in human tongue cancer. *Br. J. Cancer* **92**, 1915–1921 (2005).
24. Kamio, T. *et al.* B-cell-specific transcription factor BACH2 modifies the cytotoxic effects of anticancer drugs. *Blood* **102**, 3317–3322 (2003).
25. Xie, K. Interleukin-8 and human cancer biology. *Cytokine Growth Factor Rev.* **12**, 375–391 (2001).
26. Arimura, Y. *et al.* Elevated serum beta-defensins concentrations in patients with lung cancer. *Anticancer Res.* **24**, 4051–4057 (2004).
27. Coussens, L.M. & Werb, Z. Inflammation and cancer. *Nature* **420**, 860–867 (2002).
28. Gudmundsson, G. & Hunninghake, G.W. Respiratory epithelial cells release interleukin-8 in response to a thermophilic bacteria that causes hypersensitivity pneumonitis. *Exp. Lung Res.* **25**, 217–228 (1999).
29. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
30. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).