Effects of cigarette smoke on the human airway epithelial cell transcriptome

Avrum Spira**,†, Jennifer Beane†, Vishal Shah†, Gang Liu*, Frank Schembri*, Xueimei Yang*, John Palma§, and Jerome S. Brody*

*Pulmonary Center and Department of Medicine, Boston University School of Medicine, 715 Albany Street, Boston, MA 02118; †Bioinformatics Program, College of Engineering, Boston University, 44 Cummings Street, Boston, MA 02215; and ‡Affymetrix Inc., 3380 Central Expressway, Santa Clara, CA 95051

Cigarette smoke is the major cause of lung cancer, the leading cause of cancer death, and of chronic obstructive pulmonary disease, the fourth leading cause of death in the United States. Using high-density gene expression arrays, we describe genes that are normally expressed in a subset of human airway epithelial cells obtained at bronchoscopy (the airway transcriptome), define how cigarette smoking alters the transcriptome, and detail the effects of variables, such as cumulative exposure, age, sex, and race, on cigarette smoke-induced changes in gene expression. We also determine which changes in gene expression are and are not reversible when smoking is discontinued. The persistent altered expression of a subset of genes in former smokers may explain the risk these individuals have for developing lung cancer long after they have discontinued smoking. The use of gene expression profiling to explore the normal biology of a specific subset of cells within a complex organ across a broad spectrum of healthy individuals and to define the reversible and irreversible genetic effects of cigarette smoke on human airway epithelial cells has not been previously reported.

Materials and Methods

Study Population and Sample Collection. We recruited nonsmoking and smoking subjects (n = 93) to undergo fiberoptic bronchoscopy at Boston Medical Center between November 2001 and June 2003. Nonsmoking volunteers with significant environmental cigarette exposure and subjects with respiratory symptoms or regular use of inhaled medications were excluded. For each subject, a detailed smoking history was obtained including number of pack-years, number of packs per day, age started, age quit, and environmental tobacco exposure.

Bronchial airway epithelial cells were obtained from brushings of the right mainstem bronchus taken during fiberoptic bronchoscopy with an endoscopic cytobrush (Cellebrity Endoscopic Cytology Brush, Boston Scientific, Boston). The brushes were immediately placed in TRIZol reagent (Invitrogen) after removal from the bronchoscope and kept at ~80°C until RNA isolation was performed. RNA was extracted from the brushes by using TRIZol Reagent (Invitrogen) according to the manufacturer’s protocol, with a yield of 8–15 μg of RNA per patient. Integrity of the RNA was confirmed by running it on a RNA-denaturing gel. Epithelial cell content of representative bronchial brushing samples was quantified by cytocentrifugation (Cytospin, ThermoShandon, Pittsburgh) of the cell pellet and staining with a cytokeratin antibody (Signet Laboratories, Dedham, MA). The study was approved by the Institutional Review Board of Boston University Medical Center, and all participants provided written informed consent.

Microarray Data Acquisition and Preprocessing. We obtained a sufficient quantity of good-quality RNA for microarray studies from 85 of the 93 subjects recruited into our study. Six to eight micrograms of total RNA was processed, labeled, and hybridized to Affymetrix HG-U133A GeneChips containing ~22,500 human transcripts (for detailed protocol, see Supporting Text, which is published on the PNAS website, http://www.pnas.org/cgi/doi/10.1073/pnas.0401422101).

www.pnas.org/cgi/doi/10.1073/pnas.0401422101

PNAS | July 6, 2004 | vol. 101 | no. 27 | 10143–10148

© 2004 by The National Academy of Sciences of the USA
The normal large-airway transcriptome was defined by the genes whose median \( P_{\text{detection}} \) value was <0.05 across all 23 healthy never smokers (7,119 genes expressed across the majority of subjects) and a subset of these 7,119 genes whose \( P_{\text{detection}} \) value was <0.05 in all 23 subjects (2,382 genes expressed across all subjects). The coefficient of variation for each gene in the transcriptome was calculated as the standard deviation divided by the mean expression level multiplied by 100 for that gene across all nonsmoking individuals. To identify functional categories that were over- or underrepresented within the airway transcriptome, GOMINER software (16) was used to functionally classify the genes expressed across all nonsmokers (2,382 probe sets) by the molecular function categories within the ontology. Multiple linear regressions were performed on the top 10% most variable probe sets (712 probe sets, as measured by the coefficient of variation) in the normal airway transcriptome (7,119 probe sets) to study the effects of age, gender, and race on gene expression (see supporting information).

To examine the effect of smoking on the airway, a two-sample Student's \( t \) test was used to test for genes differentially expressed between current smokers (\( n = 34 \)) and never smokers (\( n = 23 \)). To quantify how well a given gene's expression level correlates with the number of pack-years of smoking among current smokers, Pearson correlation coefficients were calculated (see supporting information). For multiple comparison correction, a permutation test was used to assess the significance of our \( P \) value threshold for any given gene's comparison between two groups \([P_{\text{FDR}} \text{ value}] \) or between a clinical variable \([P_{\text{correlation}} \text{ value}] \) (see supporting information for details). To further characterize the behavior of current smokers, 2D hierarchical clustering of all never smokers and current smokers using the genes that were differentially expressed between current vs. never smokers was performed. Hierarchical clustering of the genes and samples was performed by using log-transformed z-score normalized data with a Pearson correlation (uncentered) similarity metric and average linkage clustering with CLUSTER and TREEVIEW software.

Multidimensional scaling and principal component analysis were used to characterize the behavior of current smokers (\( n = 34 \)) across all 23 subjects. In addition, we executed an unsupervised hierarchical clustering analysis of all 18 former smokers according to the expression of the genes differentially expressed between current and never smokers. To identify genes irreversibly altered by cigarette smoking, we performed Student's \( t \) test on the \( z \) scores of all probesets between former smokers (\( n = 18 \)) and never smokers (\( n = 23 \)) across the genes that were considered differentially expressed between current and never smokers (see supporting information for details).

Given the invasive nature of the bronchoscopy procedure, we were unable to recruit age-, race-, and gender-matched patients for the smoker vs. nonsmoker comparison. Because of baseline differences in age, gender, and race between never- and current-smoker groups (see Table 4, which is published as supporting information on the PNAS web site), we performed an analysis of covariance (ANCOVA) to test the effect of smoking status (never or current) on gene expression while controlling for the effects of age (the covariate). In addition, a two-way ANOVA was performed to test the effect of smoking status (never or current) on gene expression while controlling for the fixed effects of race (encoded as three racial groups: Caucasian, African American, and other) and gender and the interaction terms of status:race or status:gender. Both the ANCOVA and two-way ANOVA were performed with PARTek 5.0 software.

**Quantitative PCR Validation.** Quantitative real-time PCR was used to confirm the differential expression of a select number of
genes. Primer sequences were designed with PRIMER EXPRESS software (Applied Biosystems). Forty cycles of amplification, data acquisition, and data analysis were carried out in an ABI Prism 7700 Sequence Detector (Applied Biosystems). All real-time PCR experiments were carried out in triplicate on each sample (see supporting information for protocol details).

**Additional Information.** Additional information from this study, including the raw image data from all microarray samples (DAT files), expression levels for all genes in all samples (stored in a relational database), user-defined statistical and graphical analysis of data, and clinical data on all subjects are available at http://pulm.bumc.bu.edu/aged. Data from our microarray experiments have also been deposited in National Center for Biotechnology Information Gene Expression Omnibus (accession no. GSE994).

**Results and Discussion**

**Study Population and Replicate Samples.** Microarrays from 75 subjects passed the quality-control filters described above and are included in this study. Demographic data on these subjects, including 23 never smokers, 34 current smokers, and 18 former smokers, are presented in Table 4. Bronchial brushings yielded 90% epithelial cells, as determined by cytokeratin staining, with the majority being ciliated cells. Samples taken from the right and left main bronchi in the same individual were highly reproducible with an $R^2$ value of 0.92, as were samples from the same individual taken 3 months apart with an $R^2$ value of 0.85 (see Fig. 5 and Table 5, which are published as supporting information on the PN web site).

**The Normal Airway Transcriptome.** A total of 7,119 genes were expressed at measurable levels in the majority of never smokers, and 2,382 genes were expressed in all 23 healthy never smokers. Expression levels of the 7,119 genes varied relatively little; 90% had a coefficient of variation (standard deviation from the mean) of $<50\%$ (see Fig. 6, which is published as supporting information on the PN web site). Only a small part of the variation between subjects could be explained by age, gender, or race on multiple linear regression analysis (see Table 6, which is published as supporting information on the PN web site).

Table 1 depicts the GOMINER molecular functions (16) of the 2,382 genes expressed in large-airway epithelial cells of all healthy never smokers. Genes associated with oxidant stress, ion and electron transport, chaperone activity, vesicular transport, ribosomal structure, and binding functions are overrepresented. Genes associated with transcriptional regulation, signal transduction, pores and channels, and immune, cytokine, and chemokine genes are underrepresented. Upper airway epithelial cells, at least in healthy subjects, appear to serve as an oxidant detoxifying defense system for the lung but serve few other complex functions in the basal state.

**Effects of Cigarette Smoking on the Airway Transcriptome.** Smoking altered the airway epithelial cell expression of numerous genes. Ninety-seven genes were found to be differentially expressed by Student’s $t$ test between current and never smokers at $P < 1.06 \times 10^{-5}$. This $P_{(\text{test})}$ value threshold was selected based on a permutation analysis performed to address the multiple comparison problem inherent in any microarray analysis (see supporting information for further details). We chose a very stringent multiple-comparison correction and $P_{(\text{test})}$ value threshold to identify a subset of genes altered by cigarette smoking with only a small probability of having a false positive. Of the 97 genes that passed the permutation analysis, 68 (73%) represented increased gene expression among current smokers. The greatest increases were in genes that coded for xenobiotic functions such as CYP1B1 (30-fold) and DBDD (5-fold), antioxidants such as GPX2 (3-fold) and ALDH1A1 (6-fold), and genes involved in electron transport such as NADPH (4-fold). In addition, several cell adhesion molecules, including CEACAM6 (2-fold) and claudin 10 (3-fold), were increased in smokers, perhaps in response to the increased permeability that has been found upon exposure to cigarette smoke (17). Genes that decreased included TU3A (4-fold), MMP10 (2-fold), HLF (2-fold), and CX3CL1 (2-fold). In general, genes that were increased in smokers tended to be involved in regulation of oxidant stress and glutathione metabolism, xenobiotic metabolism, and secretion. Expression of several putative onco genes (pirin, CA12, and CEACAM6) were also increased. Genes that decreased in smokers tended to be involved in regulation of inflammation, although expression of several putative tumor suppressor genes (TU3A, SLIT1 and -2, and GAS6) were decreased. Changes in the expression of select genes were confirmed by real-time RT-PCR (see Fig. 7, which is published as supporting information on the PN web site).

Table 1. GOMINER molecular functions of genes in airway epithelial cells

<table>
<thead>
<tr>
<th>Molecular functions</th>
<th>Overrepresented (cells/array)</th>
<th>Underrepresented (cells/array)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binding activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA binding</td>
<td>0.76 (273/366)</td>
<td></td>
</tr>
<tr>
<td>Translation</td>
<td>0.72 (72/101)</td>
<td></td>
</tr>
<tr>
<td>Transcription</td>
<td></td>
<td>0.30 (214/704)</td>
</tr>
<tr>
<td>GTP binding</td>
<td>0.55 (106/194)</td>
<td></td>
</tr>
<tr>
<td>GTPase</td>
<td>0.55 (83/152)</td>
<td></td>
</tr>
<tr>
<td>G nucleotide</td>
<td>0.52 (128/246)</td>
<td></td>
</tr>
<tr>
<td>Receptor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chaperone</td>
<td>0.62 (80/119)</td>
<td></td>
</tr>
<tr>
<td>Chemokine</td>
<td>0.24 (10/42)</td>
<td></td>
</tr>
<tr>
<td>Cytokine</td>
<td>0.20 (39/194)</td>
<td></td>
</tr>
<tr>
<td>Enzyme activity</td>
<td>0.46 (1346/2925)</td>
<td></td>
</tr>
<tr>
<td>Oxidoreductase</td>
<td>0.54 (225/417)</td>
<td></td>
</tr>
<tr>
<td>Isomerase</td>
<td>0.56 (48/82)</td>
<td></td>
</tr>
<tr>
<td>Signal transduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structural</td>
<td>0.46 (253/548)</td>
<td>0.29 (490/1716)</td>
</tr>
<tr>
<td>Transcription regulator</td>
<td></td>
<td>0.35 (321/917)</td>
</tr>
<tr>
<td>Transporter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carrier</td>
<td>0.48 (175/363)</td>
<td></td>
</tr>
<tr>
<td>Ion</td>
<td>0.56 (130/231)</td>
<td></td>
</tr>
<tr>
<td>Anion</td>
<td></td>
<td>0.26 (15/61)</td>
</tr>
<tr>
<td>Cation</td>
<td>0.64 (116/180)</td>
<td></td>
</tr>
<tr>
<td>Metal</td>
<td>0.68 (42/62)</td>
<td></td>
</tr>
<tr>
<td>Electron</td>
<td>0.58 (131/226)</td>
<td></td>
</tr>
<tr>
<td>Channel/pore</td>
<td>0.16 (43/269)</td>
<td></td>
</tr>
</tbody>
</table>

Major molecular functional categories and subcategories of 2,382 genes expressed in all never smoker subjects. Over- or underrepresentation of categories are determined by using Fisher’s exact test. The null hypothesis is that the number of genes in our flagged set belonging to a category divided by the total number of genes in the category is equal to the number of flagged genes not in the category divided by the total number of genes in the category. Equivalency in these two proportions is consistent with a random distribution of genes into functional categories and indicates no enrichment or depletion of genes in the category being tested. Categories considered to be statistically $P_{(\text{GOS})} < 0.05$ over- or underrepresented by GOMINER are shown. Cells/arrays refers to the ratio of the number of genes expressed in epithelial cells divided by the number of genes on the U133A array in each functional category. Actual numbers are in parentheses.
These three smokers, who were similar clinically to other smokers, also segregated in the same fashion when clusters were based on the top 361 genes differentially expressed between never and current smokers ($P < 0.001$). Expression of a number of redox-related and xenobiotic genes was not increased in these three smokers (subjects 147C, 164C, and 56C), and, therefore, their profile resembled that of never smokers despite their substantial and continuing exposure to cigarette smoke. Thus, these individuals failed to increase expression of a number of genes that serve as protective detoxification and antioxidant genes, potentially putting them at risk of more severe smoking-related damage. Whether these differences represent genetic polymorphisms, and whether these individuals represent the 10–15% of smokers who ultimately develop lung cancer is uncertain. However, one of these subjects (subject 147C) subsequently developed lung cancer during the 1-year follow-up, suggesting some link between the divergent patterns of gene expression and presence of or risk for developing lung cancer. Also, a subset of four additional current smokers clustered with current smokers but did not up-regulate expression of a cluster of predominantly redox/xenobiotic genes to the same degree as other smokers, although none of these smokers had developed lung cancer in 6 months of follow-up. In addition, a never smoker, 167N (blue box), is an outlier among never smokers and expresses a subset of genes at the level of current smokers. HUGO gene ID listed for all 97 genes. The functional classification of select genes is shown. Red, high level of expression; green, low level of expression; black, mean level of expression.

![Image of gene expression heatmap](image-url)

**Fig. 1.** Clustering of current- and never-smoker samples. Hierarchical clustering of current ($n = 34$) and never ($n = 23$) smokers according to the expression of the 97 genes differentially expressed between current and never smokers. Although current and never smokers separate into two groups, three current smokers appear to cluster with never smokers (yellow rectangle). Expression of several redox-related and xenobiotic genes in these subjects was not increased (brackets) and therefore resembled that of never smokers despite substantial smoke exposure. Also, a subset of current smokers (yellow circle) did not up-regulate expression of a number of predominantly redox/xenobiotic genes (white circle) to the same degree as other smokers. In addition, a never smoker, 167N (blue box), is an outlier among never smokers and expresses a subset of genes at the level of current smokers. HUGO gene ID listed for all 97 genes. The functional classification of select genes is shown. Red, high level of expression; green, low level of expression; black, mean level of expression.
7, which is published as supporting information on the PNAS web site), only five genes correlated with pack-years at the \( P < 3.1 \times 10^{-6} \) threshold (based on permutation analysis; see supporting information for details). These genes include cystatin, which has been shown to correlate with tumor growth and inflammation (18); HBP17, which has been shown to enhance fibroblast growth factor activity (19); and BRD2, which is a transcription factor that acts with E2F proteins to induce a number of cell cycle-related genes (20). Among the genes that were correlated at the \( P < 0.0001 \) level, a number of genes decreased with increasing cumulative smoking history, including genes that are involved in DNA repair (RPA1).

Because of baseline differences in age, sex, and race between never- and current-smoker groups, analysis of covariance and two-way ANOVA were performed to test the effect of smoking status on gene expression while controlling for the effects of age, gender, race, and two-way interactions. Many of the genes found to be modulated by smoking in this analysis were also found by using the simpler Student \( t \) test. Age and gender had little effect on gene expression changes induced by smoking, whereas race appeared to influence the effect of smoking on the expression of a number of genes. The ANOVA controlling for race yielded 16 genes not included in the set of 97 genes differentially expressed between current and never smokers (see Table 8, which is published as supporting information on the PNAS web site). Given the relatively small sample size for this subgroup analysis, these observations must be confirmed in a larger study but may account in part for the reported increased incidence of lung cancer in African-American cigarette smokers (21).

Thus, the general effect of smoking on large-airway epithelial cells was to induce expression of xenobiotic metabolism and redox stress-related genes and to decrease expression of some genes associated with regulation of inflammation. Several putative oncogenes were up-regulated and tumor suppressor genes were down-regulated, although their roles in smoking-induced lung cancer remain to be determined. Risk for developing lung...
cancer in smokers has been shown to increase with cumulative pack-years of exposure (22), and a number of putative oncogenes correlate positively with pack-years, whereas putative tumor suppressor genes correlate negatively.

It is unlikely that the alterations we observed in smokers were caused by a change in cell types obtained at bronchoscopy. Several dyein genes were expressed at high levels in never smokers in our study, consistent with the predominance of ciliated cells in our samples. The level of expression of various dyein genes, and therefore the balance of cell types being sampled, did not change in smokers. This finding is consistent with a previous study of antioxidant gene expression in airway epithelial cells from never and current smokers that showed no change in histologic types of cells obtained from smokers (8).

Our findings that drug metabolism and antioxidant genes are induced by smoking in airway epithelial cells is consistent with in vitro and in vivo animal studies (summarized in ref. 9). The high-density arrays used in our studies allowed us to define the effect of cigarette smoking on a large number of genes not previously described as being affected by smoking.

Effects of Smoking Cessation. Relatively little information is available about how smoking cessation alters the effects of smoking on airways. Cough and sputum production decreases rapidly in smokers with bronchitis who cease to smoke (23). The accelerated decline in forced expiratory volume, which characterizes smoking-induced obstructive pulmonary disease, reverses to an age-appropriate decline of forced expiratory volume when smoking is discontinued (24). However, the allelic loss in airway epithelial cells obtained at biopsy changes relatively little in former smokers, and the risk for developing lung cancer remains high for at least 20 years after smoking cessation (6).

Fig. 2A shows a multidimensional scaling plot of never and current smokers according to the expression of the 97 genes that distinguish current smokers from never smokers. Fig. 2B shows that former smokers who discontinued smoking <2 years before this study tend to cluster with current smokers, whereas former smokers who discontinued smoking for >2 years group more closely with never smokers. Hierarchical clustering of all 18 former smokers according to the expression of these same 97 genes also reveals two subgroups of former smokers, with the length of smoking cessation being the only clinical variable that was statistically different between the two subgroups (see Fig. 8, which is published as supporting information on the PNAS web site). Reversible genes were predominantly drug-metabolizing and antioxidant genes.

Thirteen genes did not return to normal levels in former smokers, even those who had discontinued smoking 20–30 years before testing (P < 9.8 × 10⁻14; threshold determined by permutation analysis). These genes include a number of poten-
tial tumor suppressor genes, e.g., TUS3A and CX3CL1, which are permanently decreased, and several putative oncogenes, e.g., CEACAM6 and HN1, which are permanently increased (see Fig. 3 and Table 9, which is published as supporting information on the PNAS web site). Three metallothionein genes remain decreased in former smokers. Metallothioneins have metal-binding, detoxification, and antioxidant properties and have been reported to affect cell proliferation and apoptosis (25). The metallothionein genes that remained abnormal in former smokers are located at 16q13, suggesting that this may represent a fragile site for DNA injury in smokers. The persistence of abnormal expression of select genes after smoking cessation may provide growth advantages to a subset of epithelial cells, allowing for clonal expansion and perpetuation of these cells years after smoking had been discontinued. These permanent changes might explain the persistent risk of lung cancer in former smokers.

Conclusions. We have, for the first time, characterized the genes expressed and, by extrapolation, defined the functions of a broad cross section of epithelial cells from a complex organ across a broad cross section of healthy individuals. Large-airway epithelial cells appear to serve antioxidant, metabolizing, and host-defense functions. Cigarette smoking, a major cause of lung disease, induces xenobiotic and redox-regulating genes and several oncogenes and decreases expression of several tumor suppressor genes and genes that regulate airway inflammation. We also identified a subset of smokers who respond differently to cigarette smoke and may be predisposed to its carcinogenic effects. Finally, we have explored the reversibility of altered gene expression when smoking was discontinued. The expression level of smoking-induced genes among former smokers began to resemble that of never smokers after 2 years of smoking cessation. Genes that reverted to normal within 2 years of cessation tended to serve metabolizing and antioxidant functions. Several genes, including potential oncogenes and tumor suppressor genes, failed to revert to never-smoker levels years after cessation of smoking. These later findings may explain the continued risk for developing lung cancer many years after individuals have ceased to smoke. In addition, results from this study raise the possibility that the airway gene expression profile in smokers may serve as a biomarker for lung cancer.

We thank Dr. David Center, Dr. Marc Lenburg, Dr. Mary Williams, and Garrett Frampton for their critical review of the manuscript. Affymetrix provided U133A arrays for these studies. This work was supported in part by a Doris Duke Charitable Foundation Clinical Scientist Development Award (to A.S.) and by National Institutes of Health Grants HL71771 and HL74079 (to J.S.B.) and E031077 (to J.S.B.).