Signatures of Environmental Exposures Using Peripheral Leukocyte Gene Expression: Tobacco Smoke

Johanna W. Lampe,[1] Sergey B. Stepaniants,[2] Mao Mao,[2] Jerald P. Radich,[1] Hongyue Dai,[2] Peter S. Linsley,[2] Stephen H. Friend,[2] and John D. Potter[1]

[1]Fred Hutchinson Cancer Research Center, Seattle, WA and [2]Rosetta Inpharmatics, LLC, Merck Research Laboratories, Kirkland, WA

Abstract

Functional biological markers of environmental exposures are important in epidemiological studies of disease risk. Such markers not only provide a measure of the exposure, they also reflect the degree of physiological and biochemical response to the exposure. In an observational study, using DNA microarrays, we show that it is possible to distinguish between 85 individuals exposed and unexposed to tobacco smoke on the basis of mRNA expression in peripheral leukocytes. Furthermore, we show that active exposure to tobacco smoke is associated with a biologically relevant mRNA expression signature. These findings suggest that expression patterns can be used to identify a complex environmental exposure in humans. (Cancer Epidemiol Biomarkers Prev 2004;13(3):445–453)

Introduction

One of the major preoccupations in epidemiology is the measurement of exposures such as infectious agents, foods and dietary constituents, recreational drugs, and environmental and workplace exposures and their associations with disease outcomes. Although some exposures, such as infectious exposures, have been well measured objectively, both directly and indirectly (via serological testing of antibodies), measurement of other classes of exposures has relied largely on self-report or complex environmental monitoring. Some behaviors and exposures are well measured using self-report (*e.g.*, smoking), but many others are much more difficult (*e.g.*, diet and workplace exposures).

Environmental exposures influence a variety of biological processes in quite distinct ways (*e.g.*, enzyme induction, oxidation, signal transduction, etc.). Many of these responses also influence gene expression. Therefore, given a sufficiently large set of biological data to interpret, signatures or patterns of gene expression might emerge that would allow identification and even quantitation of specific exposures. These complex patterns of gene expression can be measured by DNA microarrays (1). One of the most readily obtainable biological materials from the general population is blood. Although peripheral leukocytes see only some environmental exposures and have a limited repertoire of responses, they are exposed, nonetheless, to many of the same environmental agents to which target tissues are exposed. Thus, blood is a useful and convenient biological material in which to seek exposure signatures.

We hypothesized that peripheral leukocyte mRNA expression could be used as a sensor to detect environmental exposures in observational studies. To test this hypothesis, we measured the mRNA signatures of individuals exposed and unexposed to tobacco smoke on the basis that: (*a*) we could verify smoking exposure by both self-report and the measurement of plasma cotinine concentrations; and (*b*) if tobacco smoking did not yield a detectable and repeatable signature, other more subtle environmental exposures were even less likely to do so. We also reasoned that if we could detect a signature for tobacco, other exposures, behaviors, and characteristics could also be explored for characteristic signatures.

Methods

Study Population. The study population was drawn from individuals who responded to advertisements and fliers placed in various locations across the local community. In addition, 3000 letters were mailed to individuals selected randomly from the Department of Motor Vehicles Department of Licensure list and 1000 letters were sent to individuals over age 60. We selected participants to obtain an equal distribution of smokers and nonsmokers within age ranges and aimed to recruit 10 each of male/smokers, female/smokers, male/nonsmokers, and female/nonsmokers in the age categories

Received 7/2/03; revised 10/16/03; accepted 11/12/03.

Grant support: Fred Hutchinson Cancer Research Center and Rosetta Inpharmatics, LLC.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: J.W. Lampe, S.B. Stepaniants, and M. Mao contributed equally to this work.

Requests for reprints: John D. Potter, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, MP-900, PO Box 19024, Seattle, WA 98109. E-mail: jpotter@fhcrc.org

21–40 and 41–60 and 5 of each of those >60 years. Eligibility criteria included: at least 21 years of age; smoker (more than one-half pack or 10+ cigarettes/day) or nonsmoker; if female, not pregnant or lactating; no contraindication to blood draws; not on medications (including oral, intramuscular, and subdermal contraceptive methods, and over-the-counter medications taken on a regular basis). Individuals were eligible if using herbal or nutritional supplements (*e.g.*, vitamins), and those taking specific medications for a finite period became eligible for the study, 1 week after they stopped the medication. The study activities were approved by the Fred Hutchinson Cancer Research Center Institutional Review Board and informed, written consent was obtained from all study participants.

Sample and Data Collection. Each participant was scheduled for two blood draws, 1 week apart. Blood samples were collected in the morning, between 7 and 10 a.m., after a 10-h fast. Two 10-ml blood samples were collected into heparin-containing tubes for mRNA and additional samples were collected into an EDTA-containing tubes for plasma cotinine and WBC differential count. Data on demographics, diet and exercise, and smoking history were also obtained. Blood samples collected at visit 1 were analyzed for plasma cotinine and nicotine concentrations by gas chromatography/nitrogen phosphate detection (National Medical Services, Willow Grove, PA). Cotinine, the primary metabolite of nicotine, has a longer circulating half-life than the parent compound and is used widely as a marker of tobacco use (2). WBC differential counts were determined on a Sysmex NE-8000 hematology analyzer.

RNA Isolation. Blood samples for mRNA were processed within 3 h of being drawn. Following lysis of erythrocytes and removal of cell debris, leukocytes were isolated from whole blood after dextran separation. Total RNA was isolated using TRIzol reagent (Invitrogen, Carlsbad, CA) and extracted according to manufacturer's protocol with one adjustment: following addition of isopropanol, RNA was precipitated on ice (4°C) for at least 10 min instead of 15–30°C for 10 min. The pellet was resuspended in 100 μl Total RNase/DNase-free water.

RNA was purified with RNase-free DNase Sets and RNeasy Kits (Qiagen, Valencia, CA). All spins were performed at 8000 rpm for 1 min (Eppendorf centrifuge, model 5415C, Brinkmann Instruments, Westbury, NY) unless noted otherwise. Total RNA was mixed with 350 μl of RLT buffer with β-mercaptoethanol (Sigma, St. Louis, MD) and 250 μl 100% ethanol (Aaper Alcohol and Chemical Co., Shelbyville, KY). This mixture was transferred to an RNeasy Mini column and centrifuged. Flowthrough was reloaded back into the column and centrifuged once again. The column was washed with 350 μl RW1 buffer. DNase I mixture (10 μl DNase I and 70 μl RDD buffer per sample) was added directly onto the membrane, and the column was left at room temperature for 15 min. Following DNase I treatment, the column was treated with additional RW1 buffer (350 μl) and centrifuged and then washed with 500 μl 80% ethanol. Following the wash, the column with bound RNA was centrifuged for 2 min at 14,000 rpm to remove any traces of ethanol. Total RNA was eluted into a collection tube with two subsequent 50-μl aliquots DEPC-treated water (centrifugation at 14,000 rpm for 1 min). Total RNA concentration was determined by A_{260nm} reading on an Eppendorf BioPhotometer (Brinkmann Instruments).

cRNA Labeling and Expression Profiling. cDNA was produced from 5 μg total RNA by reverse transcription (RT) using Moloney murine leukemia virus (MMLV) RTase and then transcribed into cRNA by *in vitro* transcription (IVT) using T7 RNA polymerase. 5-(3-Aminoallyl)uridine 5′-triphosphate (Sigma) was incorporated into cRNA in the IVT reaction. For cRNA labeling, the allylamine-derivatized cRNA products were reacted with *N*-hydroxysuccinimide esters of Cy3 or Cy5 dyes (Amersham Pharmacia Biotech, Piscataway, NJ) as described (3). Five micrograms of Cy-labeled cRNA from one leukocyte sample were mixed with the same amount of reverse color Cy-labeled product from a pool, which consisted of an equal amount of cRNA from each of seven individuals (men and women; smoking status unknown) unrelated to the study participants. Arrays were run on samples collected at visit 1 and on a subset of samples collected from 17 individuals at visit 2. The resulting labeled probes were hybridized to hu25k oligonucleotide microarrays. All hybridizations were done in duplicate with fluor reversal on two microarrays to compensate for potential biases due to the different chemical properties of Cy3 and Cy5 dyes. The arrays were scanned to detect the level of gene expression for 21,000 genes as described previously (4).

Data Analysis. For the analyses, we used gene expression data from 65 individuals (32 smokers and 33 nonsmokers) as a training set to select and optimize a set of reporter genes. We used the array data for 20 other participants (10 smokers and 10 nonsmokers, equally distributed by sex) as a test set (TEST1) and the follow-up visit 2 samples for 17 of these 20 individuals as a separate test set (TEST2). Thus, TEST1 and TEST2 were derived from the same individuals at two time points, chosen as such to determine whether the signature was stable over the time.

This study relied on a cRNA pool of seven unrelated individuals rather than a pool of all of the study participants. The latter approach is often used in array work and affords greater power to discriminate between groups; however, it also requires that a sufficient amount of RNA be available from all participants to contribute to the pool. In our study, the smoker/nonsmoker discrimination power has the potential to be limited due to the gene signature resulting from the unrelated pool. To reduce this undesirable effect, we preprocessed the three sets (training set, TEST1, and TEST2) by subtracting the average gene expression value from each gene across individuals in the training set ($n = 65$). This preprocessing is similar to centering or re-ratioing of the samples to the newly formed mathematical pool formed as a geometric mean of all individual samples within the set. This established the same baseline across the three sets.

Using data from the training set, first, we selected a set of signature genes that satisfied the following criteria: $|\text{xdev}| > 2.5$ (P value ~ 0.01) and abs($\log_{10}$ ratio) > 0.3 in

more than three individuals. The resulting 861 genes were pared down further to 857 by requiring them to have more than 95% of valid entries. These genes were ordered in descending fashion based on their absolute correlation coefficient to the plasma cotinine level. We began optimization of the reporter set by selecting the top 5 reporters and incrementing from the top of the list one reporter at a time. At each step, one profile was left out of our training set. The remaining 64 profiles were used to compute smoker and nonsmoker expression templates by averaging gene expression values across smokers and nonsmokers in the remaining set. The profile was classified as smoker or nonsmoker if it correlated more strongly to smokers or nonsmokers, respectively. This procedure was carried out for all 65 profiles in the training set and the total misclassification was computed as the sum of Type 1 and Type 2 errors. The optimization was designed to pick the set of reporters for which the total misclassification was minimal.

The reporters were then used to predict smoking status in the entire training set and test sets (TEST1 and TEST2) using the smoker and nonsmoker expression templates computed from the training set as described above. We determined the sensitivity (*i.e.*, proportion of true positives correctly identified by the test) and specificity (*i.e.*, proportion of true negatives that are correctly identified by the test) of the gene profile using cotinine as the gold standard to define smoking status. Because smoking is often associated with other behaviors (*e.g.*, higher alcohol intake, lower exercise) which themselves have the capacity to influence gene expression, we attempted to determine whether these other exposures affected expression of the identified cotinine-associated genes. Given the small sample size, we used two approaches. First, we stratified individuals by sex, exercise, and aspirin use and determined whether we could discriminate between smokers and nonsmokers within the subgroups. Second, we determined the power of the reporter genes to predict differences in sex, exercise frequency, and aspirin use. (Additional data analysis details are provided in Appendix A.)

Results

We obtained blood samples and peripheral leukocyte RNA samples without evidence of degradation on 41 self-identified smokers (16 women and 25 men) and 44 self-identified nonsmokers (21 women and 23 men). We used plasma cotinine concentrations to define smoking status. There were 6 individuals whose self-identification of smoking status was discrepant with this classification. On the basis of plasma cotinine concentrations, three of the self-identified nonsmokers who had detectable cotinine (>30 ng/ml) were reclassified as smokers, and three of the smokers who had plasma cotinine concentrations of 0 ng/ml were reclassified as nonsmokers. Participant characteristics are presented in Table 1.

Categorization on the Basis of Differentially Expressed Genes. On the basis of the gene selection process, 861 signature genes were retained initially. Genes were rank ordered in descending order by the absolute value of correlation to plasma cotinine concentration. Optimization of the set of reporter genes was performed incrementally from the top of the gene list with the aim to minimize the total misclassification error in the training set. Thirty-six top candidate genes were selected on the basis of lowest error rate (Fig. 1); 27 genes were associated directly and 9 genes were associated inversely with plasma cotinine concentrations. Among the 36 genes, 26 were characterized genes (Table 2) and

Table 1. Characteristics of study participants

	Men		Women	
	Smoker	Nonsmoker	Smoker	Nonsmoker
	n = 26	n = 22	n = 16	n = 21
Age (years)	44.6 (13.1)[a]	42.8 (16.0)	44.0 (10.6)	39.5 (9.6)
Ethnicity (% Caucasian)	65	77	81	57
Weight (kg)	86.9 (17.1)	88.5 (14.4)	79.4 (18.1)	66.9 (12.4)
Height (cm)	178.1 (7.5)	178.9 (6.5)	162.9 (6.4)	164.7 (6.9)
BMI (kg/m^2)	27.4 (5.2)	27.7 (4.6)	30.1 (7.2)	24.7 (4.6)
Self-reported, current cigarette smoking, pack/day	1.1 (0.6)	0.1 (0.5)	1.0 (0.5)	0.1 (0.2)
Aspirin use (%)	30.8	22.7	18.7	9.5
Alcohol intake, g/day	7.5 (21.5)	16.5 (28.6)	7.8 (13.4)	2.9 (4.9)
Vegetable intake, servings/day	2.1 (1.5)	1.8 (1.3)	2.0 (2.2)	2.0 (0.9)
Exercise, regular moderate (%)	34.6	36.4	25.0	33.3
Plasma cotinine (ng/ml)	217 (106)	0	241 (85)	0
Plasma nicotine (ng/ml)	12 (10)	0	13 (13)	0
White blood cell differentials				
Polymorphonucleocytes (%)	56 (12)	55 (11)	58 (11)	59 (12)
Lymphocytes (%)	33 (11)	33 (10)	32 (11)	32 (10)
Monocytes (%)	7 (3)	6 (5)	7 (3)	5 (3)
Eosinophils (%)	2 (2)	5 (6)	2 (2)	2 (2)
Basophils (%)	1 (1)	1 (1)	1 (1)	1 (1)

Note: Plasma cotinine concentrations were used to classify individuals as smokers and nonsmokers.
[a]Mean (SD).

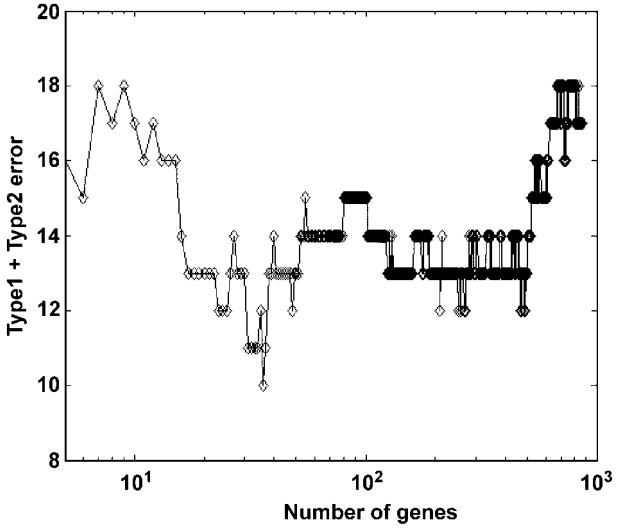


Fig. 1. Optimization of the number of reporter genes that minimized Type 1 and Type 2 error resulted in 36 reporter genes.

10 were ESTs, partial cDNAs, and hypothetical proteins. The absolute correlation of each of these individual genes with the plasma cotinine level was greater than 0.34 ($P < 0.003$; Table 2).

We determined the sensitivity (*i.e.*, proportion of true positives correctly identified by the test) and specificity (*i.e.*, proportion of true negatives that are correctly identified by the test) of the gene profile using cotinine as the gold standard to define smoking status (Table 3 and Figs. 2 and 3). Of the six misclassified smokers in the training set, two had presented themselves at the clinic as nonsmokers; these individuals had cotinines in the range of 30–50 ng/ml. In TEST1 and TEST2, one smoker (the same individual) was classified as a nonsmoker. In a cross-validation analysis using the training set, the overall (Type 1 + Type 2) error rate was 26%.

Is There Evidence of Confounding by Other Exposures? We tested for confounding (5), using two approaches. First, we stratified individuals by sex, exercise, and aspirin use and determined whether we could discriminate between smokers and nonsmokers within the subgroups. In sets TEST1 and TEST2, smokers and nonsmokers were still suitably classified (Type 1 + Type 2 error ≤20%) within strata based on sex, exercise, aspirin use, and alcohol intake. Second, we

Table 2. List of genes identified by hu25k that were significantly correlated with plasma cotinine concentrations ($P < 0.003$)

Correlation coefficient	Accession no.	Gene name	Gene description
0.699	Contig57903_RC		*Homo sapiens* cDNA FLJ13545 fis, clone PLACE1006867
0.601	NM_017933	*FLJ20701*	hypothetical protein FLJ20701
0.593	Contig38824_RC		ESTs
0.488	Contig40530		ESTs
0.483	NM_000104	*CYP1B1*	cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile)
0.453	NM_006344	*HML2*	macrophage lectin 2 (calcium dependent)
0.450	NM_000647	*CCR2*	chemokine (C-C motif) receptor 2
0.440	NM_003283	*TNNT1*	troponin T1, skeletal, slow
0.430	NM_018041	*FLJ10254*	*Homo sapiens* cDNA FLJ10254 fis, clone HEMBB1000848
0.423	NM_012307	*EPB41L3*	erythrocyte membrane protein band 4.1-like 3
0.413	NM_013956	*NRG1*	neuregulin 1
0.411	AI205537_RC		TBP-interacting protein
0.411	NM_000576	*IL1B*	interleukin 1, β
0.406	NM_003315	*DNAJC7*	DnaJ (Hsp40) homologue, subfamily C, member 7
0.402	NM_013962	*NRG1*	neuregulin 1
0.381	Contig38043_RC		ESTs, Weakly similar to hypothetical protein FLJ12547 [*Homo sapiens*]
0.374	NM_004385	*CSPG2*	chondroitin sulfate proteoglycan 2 (versican)
0.372	NM_004566	*PFKFB3*	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3
0.371	NM_013958	*NRG1*	neuregulin 1
0.361	NM_014020	*LR8*	LR8 protein
0.359	Contig35145_RC		cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)
0.358	NM_004505	*USP6*	ubiquitin specific protease 6 (Tre-2 oncogene)
0.355	D26362	*BRD3*	bromodomain containing 3
0.353	NM_018487	*HCA112*	hepatocellular carcinoma-associated antigen 112
0.349	X03084	*C1QB*	complement component 1, q subcomponent, β polypeptide
0.346	Contig65401_RC		hypothetical protein MGC26963
0.344	AB028947	*KIAA1024*	KIAA1024 protein
−0.354	NM_009590	*AOC2*	amine oxidase, copper containing 2 (retina-specific)
−0.355	NM_020366	*RPGRIP1*	retinitis pigmentosa GTPase regulator interacting protein 1
−0.358	NM_017709	*FLJ20202*	hypothetical protein FLJ20202
−0.370	NM_003596	*TPST1*	tyrosylprotein sulfotransferase 1
−0.375	Contig23408_RC		ESTs
−0.392	NM_012337	*NESG1*	nasopharyngeal epithelium specific protein 1
−0.417	AF050145	*IDS*	iduronate 2-sulfatase (Hunter syndrome)
−0.440	NM_016022	*LOC51107*	CGI-78 protein
−0.444	NM_002586	*PBX2*	pre-B-cell leukemia transcription factor 2

Table 3. Sensitivity and specificity of the 36 marker genes in distinguishing smokers and nonsmokers

Set	Nonsmokers	Smokers	Type I error	Type II error	Sensitivity (95% CI)	Specificity (95% CI)
Training	33	32	4	6	81% (66–93)	88% (72–97)
TEST1	10	10	0	1	90% (56–100)	100% (69–100)
TEST2	7	10	0	1	90% (56–100)	100% (59–100)

tested the power of the 36 reporter genes to predict exposures or behaviors plausibly associated (either directly or inversely) with smoking, for example, exercise, aspirin use, vitamin-supplement use, alcohol intake, vegetable intake. Whereas the 36 reporter genes differentiated smokers and nonsmokers with overall (Type 1 + Type 2) error rates of 5% and 6% in TEST1 and TEST2, respectively, using the data to predict each of the other exposures resulted in error rates of >30%; an exception was exercise for which overall error rates of 30% and 11% were observed for TEST1 and TEST2, respectively. Although these attempts to understand the relationships between smoking-associated exposures and identified smoking-related genes are rudimentary, these results show that plasma cotinine is a better predictor of the 36 reporter genes than is any other plausible variable.

Discussion

With this study, we have demonstrated that we can distinguish between individuals exposed and unexposed to tobacco smoke on the basis of mRNA expression in peripheral leukocytes. Furthermore, we have shown that active exposure to tobacco smoke is associated with a biologically relevant mRNA expression signature. Previously, several studies have shown that, experimentally, *in vitro* and in animals, single exposures [*e.g.*, heat shock (6) and 17β-estradiol (7)] alter gene expression in peripheral leukocytes. Our study shows that the use of expression patterns can be expanded to identify a complex exposure in human observational studies.

In this study, we examined gene expression in total leukocytes, rather than specific cellular subsets. Total numbers of peripheral leukocytes characteristically have been shown to differ by smoking status (8, 9), and differential counts have been reported to differ by number of cigarettes smoked (8) or to remain unchanged (10). Recently, interindividual variation in expression patterns also has been shown to be influenced by peripheral leukocyte composition (11). Consequently, in theory, observed differences in gene expression between smokers and nonsmokers may reflect differences in the percentage of each cell type. In our sample, there were no significant differences in the major cell type distributions by smoking status (Table 1); thus, this factor is unlikely to be a major explanation for the differences we observed in gene expression.

Many of the reporter genes that were associated with plasma cotinine concentrations in our study have not been described in relation to cigarette smoking. However, several reporter genes, such as *IL-1β* and *CYP1B1*, are associated with the pathophysiology of smoking-induced injury. Exposure to cigarette smoke is an oxidant burden, not only at the initial site of contact (*i.e.*, respiratory epithelium), but also in peripheral leukocytes, where oxidative damage and polycyclic aromatic hydrocarbon-adducts are readily detectable in smokers (12, 13). Few studies have examined the effects of cigarette smoking on leukocyte gene expression in circulation; however, smoking has well-established immune and inflammatory effects in bronchial tissue and bronchoalveolar macrophages. Differences in cytokine profiles in bronchoalveolar lavage fluid, including IL-1β, have been reported for smokers and nonsmokers (14, 15). Cigarette smoke upregulates expression of proinflammatory cytokines such as IL-1β, which in turn induce expression of a wide variety of genes. Similarly, expression of the biotransformation enzyme CYP1B1 is increased with active cigarette smoking and polycyclic aromatic hydrocarbon exposure in bronchoalveolar macrophages (16). Although one study, using quantitative reverse transcription-PCR, was unable to detect an effect of smoking on CYP1B1 mRNA levels in blood mononuclear cells (17), our results suggest that, with the higher number of cigarettes smoked by participants in our study, CYP1B1 in total leukocytes is responsive to cigarette smoke exposure.

The physiological and biochemical effects associated with cigarette smoking and the constituents of tobacco smoke support the relevance of several other reporter genes. For example, 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase-3 (PFKFB3), positively associated with plasma cotinine concentrations, is a key enzyme that regulates glycolysis in mammalian cells. It is induced by hypoxia and IL-1β through stabilization of hypoxia-inducible factor-1α (18). Iduronate 2-sulfatase, an enzyme responsible for degradation of dermatan sulfate and heparan sulfate, constituents of mucopolysaccharides, was inversely associated with plasma cotinine. Reduced expression of iduronate 2-sulfatase could extend the half-life of mucins in smokers.

From a biological standpoint, many of the identified genes associated with smoking in this study, especially those related to immune function and inflammation, can be attributed readily to cigarette smoke exposure. The absence of data on the association between expression of particular genes and cigarette smoking may reflect a lack of information on some of the pathways in relation to smoke exposure; it also may reflect the complexity of studying the effect of tobacco smoke on gene expression in intact humans. First, the pathways affected by tobacco smoke are numerous and interconnected, and some genes may be influenced by the expression of other genes. Second, tobacco smoke is a complex mixture of compounds, each of which is likely to have multiple targets (2). Third, tobacco use clusters with other human behaviors, which, if not acknowledged and accounted for, leads to confounding and biased estimates of association. Thus, the expression profile associated with smoking may also be associated with a pattern of other behaviors (*e.g.*, higher alcohol intake, lower exercise) that by themselves may influence gene expression. Nonetheless, these findings show that it is reasonable to explore whether these other exposures and behaviors

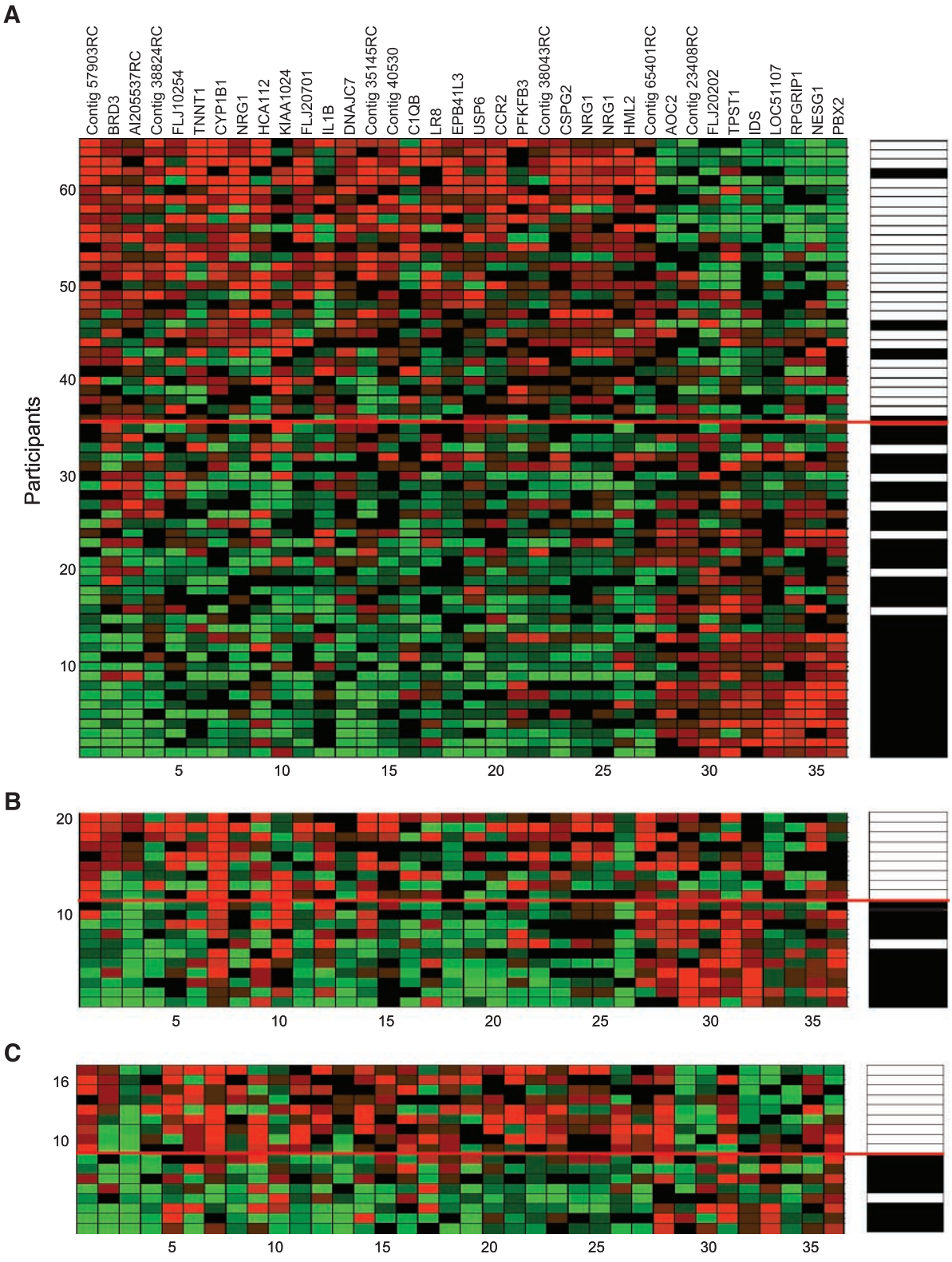


Fig. 2. Expression data matrices across 36 optimal reporter genes for cotinine exposure and categorization of individuals in the training set (**A**; $n = 65$) and TEST1 (**B**; $n = 20$) and TEST2 (**C**; $n = 17$) test sets using the 36 reporter genes. Reporter genes are ordered in the matrix based on the expression difference between nonsmoker and smoker templates in the training set. The smoking status for each participant is shown in the *right panel*: *white* indicates participants who were classified as smokers based on plasma cotinine concentrations >30 ng/ml; *black* indicates participants classified as nonsmokers with plasma cotinine of 0 ng/ml. The *red line* indicates equal correlation of profiles to both the smoker and nonsmoker templates. The color range of log ratios is from −0.3 to 0.3 for *bright green* and *bright red*, respectively.

also have readable signatures. Proper study of the relationships between exposures and gene expression will include both observational and experimental studies in humans. Such studies, optimally, will have sufficient power to detect the whole spectrum of differences and sufficient data on other exposures to allow proper control of confounding (5). Finally, our findings raise interesting, but as yet unresolved, implications for the use of readily obtained surrogate tissues in studying the relationship between exposures and biology relevant to the progression of human cancer.

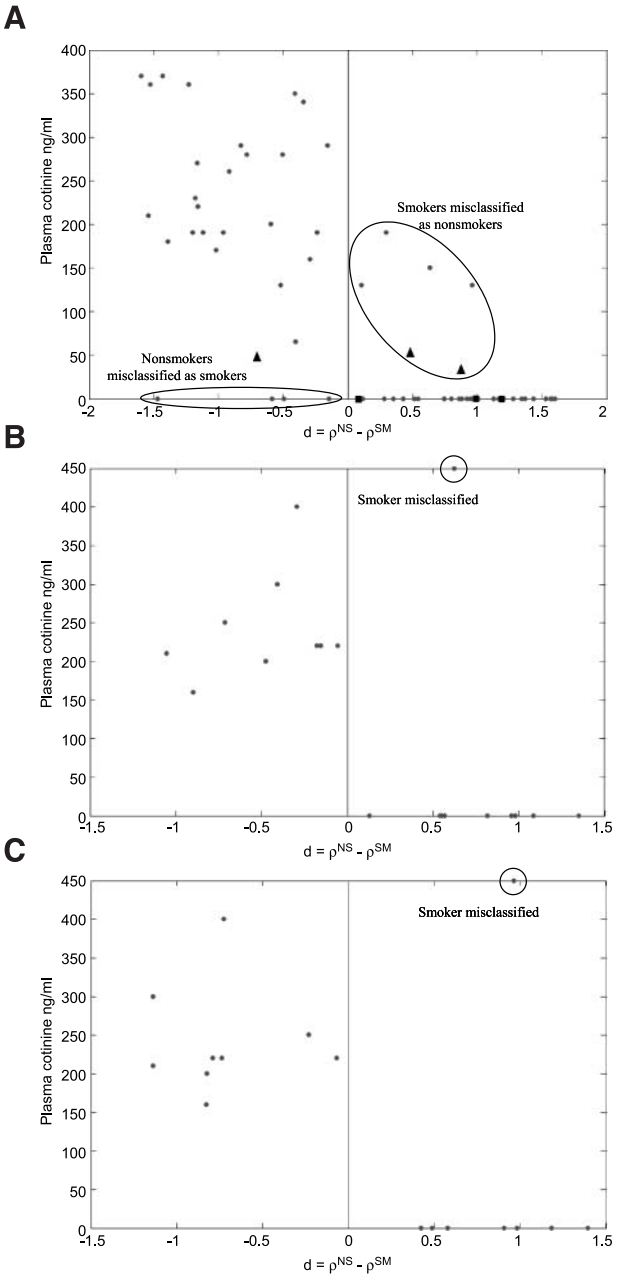


Fig. 3. For each profile P in the training and test sets, a cosine correlation to the T^{SM} and T^{NSM}, weighted by the spread of the template, is computed (see Appendix A). This generates two correlation values for each profile ρ^{SM} and ρ^{NSM} to smoker and nonsmoker templates, respectively. The classification metric is defined as $d = \rho^{NSM} - \rho^{SM}$. The scatter plots of this metric against plasma cotinine concentration in the three data sets, **(A)** training set, **(B)** TEST1, and **(C)** TEST2, are displayed. In theory, the smokers should be in the part of the plot where $d < 0$, while the nonsmokers where $d > 0$. The *points* that fail this classification are the false positive and false negatives. The six participants in the training set for whom self-report and plasma cotinine concentrations were discrepant are represented as *triangles* (self-reported nonsmokers with detectable cotinine) and *squares* (self-reported smokers with 0 ng/ml cotinine).

Appendix A

Supplementary Material

1. Re-ratioing to the mean template from 65 individuals

For a given gene, let X_I and σ_I denote its expression log-ratio for person I = 1,...,N, N = 65 and its error, respectively. For the expression of each gene, one could compute the mean and its error across 65 individuals as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_I$$

$$\sigma_\mu = std(X_I)/\sqrt{N-1}$$

Equation 1

We preprocess each gene's expression by subtracting the mean μ and propagating error as follows

$$Y_I = X_I - \mu$$

$$\epsilon_I = \sqrt{\sigma_I^2 + \sigma_\mu^2}$$

Equation 2

where Y_I and ε_I are the expression and the error of the gene with respect to its average expression μ. The P value for the significance of expression is computed as

$$xdev_I = \frac{Y_I}{\varepsilon_I}$$

$$P_I = Erfc\left(\frac{1}{\sqrt{2}}|xdev_I|\right)$$

Equation 3

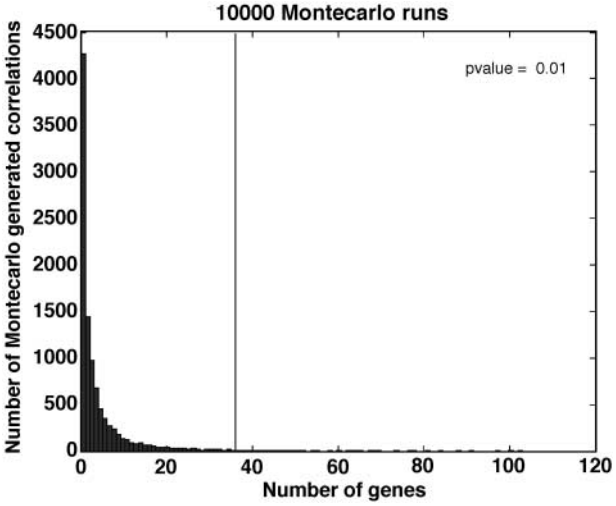
2. Correlation to the cotinine profile

For each gene, correlation to the cotinine profile was computed using Pearson correlation coefficient.

3. Significance of the selected predictor genes

Thirty-six reporter genes were checked for significance by performing Monte Carlo simulations. The correlations of the 36 reporter genes to the cotinine profile were equal or above ~0.34. The originally selected 861 signature genes were correlated to the randomly permuted cotinine profile using 10,000 Monte Carlo runs. For each such run, we counted how many genes out of 861 had correlation coefficients greater or equal to 0.34. We then counted the number of runs that resulted in the number of genes greater or equal to 36—the number of genes that

had positive predictive value. One hundred ten of the 10,000 Monte Carlo runs resulted in ≥36 genes with a correlation coefficient ≥0.34, yielding a P value of 0.01 (figure).



The originally selected 861 signature genes were correlated to the randomly permuted cotinine profile using 10,000 Monte Carlo runs. For each run, we determined how many genes out of 861 had a correlation coefficient ≥ 0.34 and counted the number of runs that resulted in the number of genes ≥ to 36—the number of genes that had positive predictive value.

4. Similarity metric for profile classification

For the classification, we have used correlation-based similarity metric. Smoker and nonsmoker expression templates and their spreads have been computed as averages over smokers and nonsmokers in the training set as (example only for smokers)

$$T_i^{SM} = \frac{1}{N_{Smokers}} \sum_{j \in Smokers} LR_{ij}$$

$$stdT_i^{SM} = \sqrt{\frac{1}{N_{Smokers} - 1} \sum_{j \in Smokers} (LR_{ij} - T_i^{SM})^2}$$

$$i = 1, 2, \ldots, 36$$

Equation 4

where LR_{ij} are log ratios of gene i in the experiment j and we have selected 36 genes based on the optimization process.

For each profile P in the training and test sets, a cosine correlation to the T^{SM} and T^{NSM}, weighted by the spread of the template, is computed as

$$\rho_k^{SM} = \frac{\sum_{i=1}^{36} P_{ik} \frac{T_i^{SM}}{stdT_i^{SM}}}{\sqrt{\sum_{i=1}^{36} P_{ik}^2 \sum_{i=1}^{36} \left(\frac{T_i^{SM}}{stdT_i^{SM}}\right)^2}}$$

Equation 5

Thus, we have two correlation values for each profile ρ^{SM} and ρ^{NSM} to smoker and nonsmoker templates, respectively. We can define the classification metric as

$$d = \rho^{NSM} - \rho^{SM}$$

Equation 6

The scatter plot of this metric against cotinine level in all three data sets is displayed in the figures. It is clear that most of the smokers should be in the part of the plot where $d < 0$, while most nonsmokers where $d > 0$. The points that fail this classification were our false positive and false negatives (see Fig. 3). Note, that 36 genes as well as smoker and nonsmoker templates are chosen from the training set and are used in the test sets.

5. Cross-validation

Using the originally defined 857 signature genes, cross-validation was performed based on the ''leave out one profile'' approach. At each step of cross-validation, a single expression profile was left out. The remaining 64 profiles were used to re-select reporters that had correlation coefficients to the cotinine level of greater or equal to 0.34. Smoker and nonsmoker templates were computed based on the remaining 64 profiles using this new reporter set. The profile left out was classified using the similarity metrics described above. The result of such cross-validation is a more conservative estimate of classification performance in the training set because it avoids the overtraining problem, that is, classifying based on the reporters and templates chosen from the same set.

Acknowledgments

We thank Yvonne Schwarz for her assistance with participant recruitment and sample collection, Jakub Stefka and Era Pogosova in the Radich laboratory for RNA isolation, and Chris Roberts, Michele Thorton, George Schreiber, and Rosetta's High Throughput Gene Expression Facility for the microarray runs.

References

1. Hughes TR, Marton MJ, Jones AR, et al. Functional discovery via a compendium of expression profiles. Cell, 2000;109–26.
2. IARC. IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans. Volume 38. Tobacco smoking. Switzerland: IARC; 1986.
3. Hughes TR, Mao M, Jones AR, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol, 2001;19:342–7.
4. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002;415:530–6.
5. Potter JD. At the interfaces of epidemiology, genetics and genomics. Nat Rev Genet, 2001;2:142–7.
6. Sonna LA, Gaffin SL, Pratt RE, Cullivan ML, Angel KC, Lilly CM. Effect of acute heat shock on gene expression by human peripheral blood mononuclear cells. J Appl Physiol, 2002;92:2208–20.
7. Rockett JC, Kavlock RJ, Lambright CR, et al. DNA arrays to monitor gene expression in rat blood and uterus following 17β-estradiol exposure: biomonitoring environmental effects using surrogate tissues. Toxicol Sci, 2002;69:49–59.
8. Schwartz J, Weiss ST. Host and environmental factors influencing the peripheral blood leukocyte count. Am J Epidemiol, 1991;134:1402–9.
9. Jensen EJ, Pedersen B, Frederiksen R, Dahl R. Prospective study on the effect of smoking and nicotine substitution on leucocyte blood counts and relation between blood leucocytes and lung function. Thorax, 1998;53:784–9.

10. Malech HL, Gallin JI. Neutrophils in human diseases. N Engl J Med, 1987;317:687–702.
11. Whitney AR, Diehn M, Popper SJ, et al. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci USA, 2003;100:1896–901.
12. Tang D, Santella RM, Blackwood AM, et al. A molecular epidemiological case-control study of lung cancer. Cancer Epidemiol Biomarkers & Prev, 1995;4:341–6.
13. Kiyosawa H, Suko M, Okudaira H, et al. Cigarette smoking induces formation of 8-hydroxydeoxyguanosine, one of the oxidative DNA damages in human peripheral leukocytes. Free Radic Res Commun, 1990;11:23–7.
14. Kuschner WG, D'Alessandro A, Wong H, Blanc PD. Dose-dependent cigarette smoking-related inflammatory responses in healthy adults. Eur Respir J, 1996;9:1989–94.
15. Mikuniya T, Nagai S, Tsutsumi T, et al. Proinflammatory or regulatory cytokines released from BALF macrophages of healthy smokers. Respiration, 1999;66:419–26.
16. Piipari R, Savela K, Nurminen T, et al. Expression of CYP1A1, CYP1B1 and CYP3A, and polycyclic aromatic hydrocarbon-DNA adduct formation in bronchoalveolar macrophages of smokers and non-smokers. Int J Cancer, 2000;86:610–6.
17. Dassi C, Signorini S, Gerthoux P, Cazzaniga M, Brambilla P. Cytochrome P450 1B1 mRNA measured in blood mononuclear cells by quantitative reverse transcription-PCR. Clin Chem, 1998;44: 2416–21.
18. Minchenko A, Leshchinsky I, Opentanova I, et al. Hypoxia-inducible factor-1-mediated expression of the 6-phoshofructo-2-kinase/fructose-2,6-bisphosphatase-3 (*PFKFB3*) gene: its possible effect in the Warburg effect. J Biol Chem, 2002;277:6183–7.