Shared Gene Expression Alterations in Nasal and Bronchial Epithelium for Lung Cancer Detection

Joseph F. Perez-Rogers, Joseph Gerrein, Christina Anderlind, Gang Liu, Sherry Zhang, Yuriy Alekseyev, Kate Porta Smith, Duncan Whitney, W. Evan Johnson, David A. Elashoff, Steven M. Dubinett, Jerome Brody, Avrum Spira*, Marc E. Lenburg*; for the AEGIS Study Team

Abstract

Background: We previously derived and validated a bronchial epithelial gene expression biomarker to detect lung cancer in current and former smokers. Given that bronchial and nasal epithelial gene expression are similarly altered by cigarette smoke exposure, we sought to determine if cancer-associated gene expression might also be detectable in the more readily accessible nasal epithelium.

Methods: Nasal epithelial brushings were prospectively collected from current and former smokers undergoing diagnostic evaluation for pulmonary lesions suspicious for lung cancer in the AEGIS-1 (n = 375) and AEGIS-2 (n = 130) clinical trials and gene expression profiled using microarrays. All statistical tests were two-sided.

Results: We identified 535 genes that were differentially expressed in the nasal epithelium of AEGIS-1 patients diagnosed with lung cancer vs those with benign disease after one year of follow-up (P < .001). Using bronchial gene expression data from the AEGIS-1 patients, we found statistically significant concordant cancer-associated gene expression alterations between the two airway sites (P < .001). Differentially expressed genes in the nose were enriched for genes associated with the regulation of apoptosis and immune system signaling. A nasal lung cancer classifier derived in the AEGIS-1 cohort that combined clinical factors (age, smoking status, time since quit, mass size) and nasal gene expression (30 genes) had statistically significantly higher area under the curve (0.81; 95% confidence interval [CI] = 0.74 to 0.89, P = .01) and sensitivity (0.91; 95% CI = 0.81 to 0.97, P = .03) than a clinical-factor only model in independent samples from the AEGIS-2 cohort.

Conclusions: These results support that the airway epithelial field of lung cancer–associated injury in ever smokers extends to the nose and demonstrates the potential of using nasal gene expression as a noninvasive biomarker for lung cancer detection.
over 95% were ultimately diagnosed as benign (2). While there are
guidelines for the management of pulmonary nodules (3),
unnecessary invasive procedures (including surgical lung biopsy) are
frequently performed on patients who are ultimately diag-
nosed with benign disease (4,5). There is a clear and growing need
to develop additional diagnostic approaches for evaluating pul-
monary lesions to determine which patients should undergo CT surveil-
ance or invasive biopsy.

Our group and others have established that airway gene
expression signatures can serve as biomarkers for the presence
of smoking-related lung diseases including lung cancer and
COPD (6–8). Previous work has shown that chronic exposure to
tobacco smoke results in an airway-wide field of injury with both reversible and irreversible alterations in bronchial airway
epithelial cell gene expression upon smoking cessation (9,10).
Importantly, gene expression profiles from cytotologically normal
bronchial epithelial cells obtained via endobronchial brushings can serve as a biomarker that distinguishes ever smokers with lung cancer from those with benign lung disease (6) indepen-
dently of clinical risk factors (11). More recently, Whitney et al.
reported a 23-gene bronchial genomic lung cancer classifier (12),
which was validated in two prospective clinical trials (13). While
the high sensitivity and negative predictive value of this classi-
fier suggest that it can be used to assign patients to CT surveil-
ance, it is dependent upon cells obtained during bronchoscopy.
Although bronchoscopy-related complications are uncommon
(14,15), bronchoscopy is not always chosen as a diagnostic
modality based on the size/location of the lung lesion, pretest risk of disease, patient preference, and degree of underlying lung disease.

Given concordant response of nasal and bronchial epithelium to tobacco smoke (16) and the validated performance of the bron-
chial genomic classifier for lung cancer, we sought to test the hypoth-
thesis that cancer-associated expression differences might also be detectable in nasal epithelium and related to those found in bronchial epithelium. Detecting cancer-associated gene
expression in nasal epithelium would suggest its potential to serve as a less invasive biopsy specimen for lung cancer detection and potentially expand the clinical settings in which airway gene
expression could be used for this purpose.

Methods

For a complete description of the methods, please see the
Supplementary Materials and Methods (available online).

Experimental Design

Patients were enrolled at 28 medical centers in the United
States, Canada, and Europe as part of two prospective studies
within the Airway Epithelial Gene Expression in the Diagnosis of Lung Cancer (AEGIS) clinical trials (registered as NCT01309087 and
NCT00746759). All study protocols were approved by the institu-
tional review board at each medical center, and written
informed consent was obtained from all patients prior to enroll-
ment. Inclusion and exclusion criteria have been previously
described (13). All patients were current or former cigarette
smokers (>100 cigarettes in their lifetime) undergoing broncho-
scopy as part of their diagnostic workup for clinical suspicion
of lung cancer, and all samples were collected prospectively
prior to diagnosis. Patients were followed for up to one year
postbronchoscopy until a final diagnosis of lung cancer or
benign disease was made. Lung tumor stage was assessed using
the TNM staging system (17). In this study, 554 nasal epithelium
samples were selected and profiled using microarrays (see
Supplementary Materials and Methods, available online). We
were limited by patients with a benign diagnosis and matched
them approximately 2:3 with patients diagnosed with lung can-
cer. Microarray data generated from bronchial epithelium sam-
plies from 299 patients together with their clinical annotations
were obtained from Whitney et al. (12).

Microarray Processing

All procedures were performed as previously described (7) using
Affymetrix Gene 1.0 ST microarrays. CEL files from nasal and
bronchial samples passing quality control (Supplementary
Materials and Methods, available online) were processed sepa-
rately using the Robust Multichip Average (RMA) algorithm (18)
and the standard Affymetrix Chip Definition File to estimate
gene expression signal. ComBat (19) was used to correct for
batch effects.

Characterization of Cancer Associated Genes in Nasal
Epithelium

Cancer-associated gene expression profiles in nasal epithelium
were identified using linear models (20) that corrected for smoking
status, pack-years, sex, age, and RNA quality (RIN). Functional
enrichment of the most differentially expressed genes (P < .001) was
assessed using the Reactome and Gene Ontology (GO) databases and
Enrichr (21). Similarities between cancer-associated gene
expression profiles in nasal and bronchial epithelium were assessed
using a preranked gene set enrichment analysis (GSEA) (22).

The expression of 11 gene clusters previously identified as
being associated with cancer in the bronchial epithelium (12)
was summarized in AEGIS-1 nasal samples by taking the mean
of the cluster genes per sample. The association of each cluster
mean with the presence or absence of cancer was computed
using a Welch t test. Finally, the correlation of scores from a
previously reported bronchial genomic lung cancer classifier (12)
was evaluated in matched bronchial and nasal samples from
AEGIS-1 (n = 157 patients). These analyses are described in detail
in the Supplementary Materials and Methods (available online).

Derivation of Nasal Clinico-Genomic Lung Cancer
Classifier

We derived a clinical factor lung cancer classifier using a
training set of AEGIS-1 patients (n = 517) using logistic regres-
sion to combine patient age, smoking status (current, former),
time since quit (<15 years, >15 years, unknown), and mass size
(<3 cm, ≥3 cm, infiltrates). Similarly, we derived a clinico-
genomic classifier using penalized logistic regression to combine
all of the variables in the clinical factor model plus the score
from a nasal lung cancer gene expression classifier. This gene
expression classifier was derived in the nasal training set
(n = 375) using a weighted voting algorithm. The most differen-
tially expressed genes by moderated t test were included in
this model. The total number of genes included in this classifier
was optimized in cross-validation. Specifically, we selected the
smallest number of genes that maximized cross-validation
performance. A more detailed description is provided in the
Supplementary Materials and Methods (available online).
Statistical Analysis

Differences in clinical covariates between patients with and without lung cancer were assessed using Fisher’s exact test (categorical variables) or Welch t test (continuous variables). Differential expression analyses were performed using linear modeling (limma R package [20]) or Welch t tests unless otherwise specified. For the differential expression analysis, a two-sided P value of less than .001 was considered evidence of statistically significant differential expression. Correlation coefficients were calculated using Pearson’s product-moment coefficient. Classification performance was assessed using standard measures including receiver operating characteristic (ROC) curve, the area under the curve (AUC), sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV). Differences between ROC curve AUC values were assessed using DeLong’s test [23] for correlated ROC curves. Operating points for binary classification were chosen as the threshold that maximized sensitivity while maintaining 50% specificity in the training set. Differences in sensitivity and specificity between models were assessed using McNemar’s chi-square test for count data [24]. Statistical differences in NPV between models were assessed using the generalized score statistic [25] for paired analyses or a proportions test for unpaired analyses. All confidence intervals (CIs) are reported as two-sided binomial 95% confidence intervals. All statistical tests were two-sided, and a P value of less than .05 was considered statistically significant.

Results

Study Population

Five hundred fifty-four nasal epithelial samples were selected for microarray profiling from a larger pool of RNA samples collected prospectively from patients with suspect lung cancer enrolled in the AEGIS clinical trials [13]. Four hundred twenty-four of these samples were from patients enrolled in the AEGIS-1 trial, and 130 were from patients in the AEGIS-2 trial (Supplementary Figure 1, available online). Thirty-one patients from the AEGIS-1 cohort had an indeterminate cancer diagnosis at one year post–sample collection or were lost to follow-up and were removed from the study. Additionally, 18 nasal microarray samples from the AEGIS-1 data set that did not meet minimum quality standards were also removed. No samples were removed from the AEGIS-2 data set. The remaining 375 samples from the AEGIS-1 cohort were used as a training set in which all data analyses and biomarker derivation steps were performed, while the 130 samples from the AEGIS-2 cohort were used solely to validate the predictive model (Table 1). The distribution of cancer stages was slightly skewed toward later-stage cancers in the validation set (Supplementary Table 1, available online). Lung cancer patients tended to have larger nodules than patients with benign diagnoses in both the training and validation sets (P < .001 for both comparisons) (Supplementary Table 2, available online) while patient age was statistically significantly higher among cancer patients in the training set (P < .001). The gene expression data from these samples has been deposited in the NCBI Gene Expression Omnibus under accession number GSE80796.

Lung Cancer–Associated Gene Expression in Nasal Epithelium

We identified 535 genes that were differentially expressed in the nasal epithelium of AEGIS-1 patients diagnosed with lung cancer vs those with benign disease after one year of follow-up. Additionally, 18 nasal microarray samples from the AEGIS-1 data set that did not meet minimum quality standards were also removed. No samples were removed from the AEGIS-2 data set. The remaining 375 samples from the AEGIS-1 cohort were used as a training set in which all data analyses and biomarker derivation steps were performed, while the 130 samples from the AEGIS-2 cohort were used solely to validate the predictive model (Table 1). The distribution of cancer stages was slightly skewed toward later-stage cancers in the validation set (Supplementary Table 1, available online). Lung cancer patients tended to have larger nodules than patients with benign diagnoses in both the training and validation sets (P < .001 for both comparisons) (Supplementary Table 2, available online) while patient age was statistically significantly higher among cancer patients in the training set (P < .001). The gene expression data from these samples has been deposited in the NCBI Gene Expression Omnibus under accession number GSE80796.

Table 1. Clinical and demographic characteristics of patients who contributed nasal epithelial samples

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>AEGIS-1 training set (n = 375)</th>
<th>AEGIS-2 validation set (n = 130)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer Status, No.*</td>
<td>.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>243</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>Benign Disease</td>
<td>132</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>Smoking Status, No.*</td>
<td>.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>140</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Former</td>
<td>235</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>Sex, No.*</td>
<td>.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>237</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>138</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Cumulative smoke exposure (SD, No.), pack-yy†</td>
<td>.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39.0 (26.9, 371)</td>
<td>34.8 (30.7, 130)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time since quit (SD, No.), y†</td>
<td>.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.6 (12.9, 309)</td>
<td>9.4 (13.4, 120)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (SD), y†</td>
<td>.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59.5 (10.4)</td>
<td>61.7 (11.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesion size, No.*,†</td>
<td>.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;3 cm</td>
<td>171</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>≤3 cm</td>
<td>142</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>Infiltrate</td>
<td>44</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>18</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Lesion location, No.*,§</td>
<td>.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central</td>
<td>134</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Peripheral</td>
<td>114</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Central and peripheral</td>
<td>100</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>27</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Lung cancer histological type, No.*,‖</td>
<td>.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small cell</td>
<td>40</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Non–small cell</td>
<td>200</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>90</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>Squamous</td>
<td>72</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Large cell</td>
<td>9</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Not specified</td>
<td>29</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Diagnosis of benign condition, No.*</td>
<td>.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infection</td>
<td>105</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Sarcoidosis</td>
<td>36</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Not specified</td>
<td>21</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>48</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

*P value calculated using two-sided Fisher’s Exact test.
†P value calculated using two-sided Student’s t test.
‡P value calculated comparing >3 cm vs ≤3 cm vs infiltrates.
§P value calculated comparing central vs peripheral vs central and peripheral.
‖P value calculated comparing non–small cell vs small cell.

(P < .001) (Figure 1; Supplementary File 1, available online). Genes downregulated in patients with lung cancer were enriched for genes associated with DNA damage, regulation of apoptosis, and processes involved in immune system activation including the interferon-gamma signaling pathway and antigen presentation (Table 2). Among genes that were upregulated in lung cancer patients, we found enrichment for genes involved in endocytosis and ion transport (Table 2). A complete list of statistically significantly enriched pathways and GO categories (FDR < 0.05) is shown in Supplementary Files 2–4 (available online).

Similarities in Nasal and Bronchial Cancer–Associated Gene Expression

To determine if a shared pattern of cancer-related gene expression might exist between the nose and bronchus, we leveraged...
microarray data from 299 bronchial epithelium samples obtained from AEGIS-1 patients (12). One hundred fifty-seven of the 299 bronchial samples came from the same patients as those in our nasal training set (Supplementary Table 3 and Supplementary Figure 2, available online). Using bronchial gene expression data from the AEGIS-1 patients, we found statistically significant concordant cancer-associated gene expression alterations between the two airway sites ($P < .001$). GSEA (22) revealed (Figure 2A) that the genes upregulated in nasal epithelium of patients with lung cancer were among the genes most upregulated in bronchial epithelium of patients with cancer ($P < .001$) and that a similar relationship exists for genes downregulated in patients with cancer between the nose and bronchus ($P < .001$). The expression of the most concordantly differentially expressed genes is shown in Figure 2B and highlighted in Supplementary File 1 (available online).

To further explore the hypothesis of a shared field of nasal and bronchial lung cancer–associated injury, we examined the nasal expression of 232 genes with lung cancer–associated expression in bronchial epithelium (12). Whitney et al. grouped these genes into 11 clusters, and we found that the mean expression of eight of the 11 clusters was statistically significantly associated with lung cancer ($P < .05$) in our training set (Table 3), including gene clusters enriched for genes involved in cell cycle, response to retinoic acid, and the innate immune response. Based on the concordant expression of cancer-associated genes
in bronchial and nasal epithelium, we computed a “bronchial” lung cancer classifier score (12) for the nasal training set samples and found that they were highly correlated with the scores computed from matched bronchial samples (n = 157, R = 0.70, P < .001) (Supplementary Figure 3, available online). Taken together, these results suggest that lung cancer–associated gene expression differences are similar in nasal and bronchial epithelium.

Nasal Gene Expression as an Independent Predictor of Lung Cancer Status

To determine if nasal gene expression could serve as a predictor of lung cancer status, we selected the 30 most statistically significantly differentially expressed genes (P < .001) from among the 535 genes with cancer-associated nasal gene expression for use in a weighted-voting biomarker (Supplementary File 5, available online). The biomarker panel size of 30 genes was chosen as the smallest number of genes that achieved maximal performance in cross-validation. This biomarker had an AUC of 0.69 (n = 375, 95% CI = 0.63 to 0.75, P < .001) in cross-validation in the training set. Twenty-two of the 30 genes were also statistically significantly correlated between matched bronchial and nasal samples (mean R = 0.29, range = 0.16–0.49, P < .05).

In order to evaluate the potential for the nasal gene expression biomarker to add to clinical risk factors for lung cancer detection, we developed a clinical risk factor model and tested whether incorporating the gene-expression biomarker enhances its performance. In choosing which clinical risk factors to include, we relied on a study in which Gould et al. identified smoking status, time since quit, age, and mass size as important clinical risk factors of lung cancer for patients with solitary pulmonary nodules (26). Patient age, smoking status (current, former), time since quit (≤ 15 years, > 15 years, unknown), and categorized mass size (< 3 cm, ≥ 3 cm, infiltrates) were used to create a clinical risk factor model for lung cancer using logistic regression. The training set for this model consisted of the nasal training set used to derive the gene expression classifier as well as clinical data from an additional 142 patients from the AEGIS-1 cohort for a total training set of 517 patients for the clinical model (see Supplementary Figure 2, available online). A

Table 2. Functional characterization of genes with cancer-associated expression in nasal epithelium

<table>
<thead>
<tr>
<th>Genes</th>
<th>False discovery rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Downregulated genes (n = 492)</strong></td>
<td></td>
</tr>
<tr>
<td>DNA damage</td>
<td></td>
</tr>
<tr>
<td>Signal transduction involved in mitotic DNA integrity checkpoint (GO:1902400)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ubiquitin-dependent degradation of Cyclin D1 (reactome)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Regulation of apoptosis (reactome)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GI/S DNA damage checkpoints (reactome)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Immune system activation</td>
<td></td>
</tr>
<tr>
<td>Antigen presentation and processing of exogenous antigen (GO:0019884)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Interferon-gamma signaling (reactome)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Upregulated genes (n = 43)</strong></td>
<td></td>
</tr>
<tr>
<td>Ion transport</td>
<td></td>
</tr>
<tr>
<td>Response to magnesium ion (GO:0032026)</td>
<td>0.01</td>
</tr>
<tr>
<td>Regulation of endocytosis (GO:0030100)</td>
<td>0.01</td>
</tr>
<tr>
<td>Positive regulation of release of calcium ion into cytosol (GO:0010524)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Figure 2. Concordance between cancer-associated gene expression in bronchial and nasal epithelium. A) The 535 genes with cancer-associated expression in nasal epithelium were split into up- and downregulated gene sets, and we examined their distribution within all genes ranked from most downregulated (left) to most upregulated (right) in the bronchial epithelium of patients with cancer using gene set enrichment analysis. We found that the genes with increased expression in nasal epithelium are enriched among the genes that are most induced in the bronchial epithelium of patients with cancer (top; P < .001 by a two-sided permutation-based Kolmogorov-Smirnov-like test [22]) while the reverse was true for genes with decreased expression in nasal epithelium (bottom; P < .001 by a two-sided permutation-based Kolmogorov-Smirnov-like test [22]). Genes included in the core enrichment are shown in the green box. B) Heatmaps and hierarchical clustering of the core enrichment genes in nasal (left) and bronchial (right) samples. All statistical tests were two-sided.

**ARTICLE**


5 of 9

Downloaded from https://academic.oup.com/jnci/article-abstract/109/7/djw327/3053477
by Boston University user
on 07 November 2017
The performance of the clinical and clinicogenomic models were chosen to maximize training set sensitivity with specificity of 50% or greater for both models. The addition of cancer-associated gene expression to the clinical risk factor model increased sensitivity from 0.79 (95% CI = 0.67 to 0.88) to 0.91 (95% CI = 0.81 to 0.97, \( P = .03 \)) and negative predictive value from 0.73 (95% CI = 0.58 to 0.84) to 0.85 (95% CI = 0.69 to 0.94, \( P = .03 \)) (Table 4). The negative likelihood ratio of the clinicogenomic classifier was consistent between training (0.18; 95% CI = 0.12 to 0.28) and validation (0.18; 95% CI = 0.08 to 0.39) sets. Additionally, in subjects with either lesion size less than 3 cm or peripheral lesions, the clinicogenomic model had a negative predictive value of 0.85 (95% CI = 0.65 to 0.96) or 0.93 (95% CI = 0.66 to 1.00), respectively (Supplementary Table 4, available online).

Table 3. Aggregate expression of lung cancer gene clusters from bronchial epithelium in nasal epithelial samples

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Function</th>
<th>No. of probesets</th>
<th>Direction in cancer</th>
<th>( P^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1†</td>
<td>Innate immune</td>
<td>25</td>
<td>Down</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>2†</td>
<td>Mitotic cell cycle</td>
<td>47</td>
<td>Down</td>
<td>.05</td>
</tr>
<tr>
<td>3</td>
<td>Inflammation</td>
<td>45</td>
<td>Down</td>
<td>.09</td>
</tr>
<tr>
<td>4†</td>
<td>Resp. retinoic acid/cell cycle</td>
<td>34</td>
<td>Up</td>
<td>.004</td>
</tr>
<tr>
<td>5</td>
<td>NA</td>
<td>10</td>
<td>Up</td>
<td>.36</td>
</tr>
<tr>
<td>6</td>
<td>NA</td>
<td>21</td>
<td>Down</td>
<td>.02</td>
</tr>
<tr>
<td>7†</td>
<td>Submucosal gland markers</td>
<td>20</td>
<td>Up</td>
<td>.01</td>
</tr>
<tr>
<td>8</td>
<td>n/a</td>
<td>15</td>
<td>Up</td>
<td>.003</td>
</tr>
<tr>
<td>9†</td>
<td>Xenobiotic detoxification</td>
<td>7</td>
<td>Down</td>
<td>.15</td>
</tr>
<tr>
<td>10†</td>
<td>Cartilaginous markers</td>
<td>4</td>
<td>Down</td>
<td>.05</td>
</tr>
<tr>
<td>11</td>
<td>NA</td>
<td>1</td>
<td>Down</td>
<td>.03</td>
</tr>
</tbody>
</table>

\( P^* \) value of two-sided \( t \) test measuring the difference in mean average expression of all genes in a cluster between cancer and benign nasal sample in the AEGIS-1 cohort.

†In bronchial genomic classifier described by Whitney et al. 2015 (12).

**Discussion**

In this study, we have established that the lung cancer-associated airway “field of injury” detectable in bronchial airway epithelium \((6,11,12)\) extends to the nasal epithelium. These findings strengthen the “field of injury” hypothesis in which there are gene expression alterations in normal-appearing epithelial cells throughout the entire airway of smokers with lung cancer and, intriguingly, suggest the potential for lung cancer biomarkers based on nasal epithelial gene expression.
While previous studies have validated the existence of bronchial airway gene expression alterations in patients with lung cancer and demonstrated their clinical utility in lung cancer detection (13), little is known about the physiological processes responsible for this “field of injury.” One hypothesis for the presence of lung cancer–associated alterations in nasal and bronchial gene expression is that the subset of smokers who develop lung cancer exhibit a distinct genomic response to tobacco smoke exposure throughout all airway epithelial cells, consistent with the “etiologic field effect” described by Lochhead et al. for colon and other cancer types (27). This paradigm suggests that the airway gene-expression signature is a risk marker for lung cancer as opposed to a direct consequence of the presence of lung cancer based on local or systemic factors produced by the tumor or its microenvironment (ie, the “conventional field effect” defined by Lochhead et al. [27]). Consistent with the etiologic field effect hypothesis, we observed a concordant downregulation of genes associated with immune system activation in patients with lung cancer in both bronchial and nasal epithelium, which might suggest that an impaired immune response sets the stage for tumorogenesis in the lung microenvironment. Alternatively, despite the distance to the tumor, these cancer-associated gene expression differences may be a direct result of factors secreted by the tumor or its microenvironment, or some other consequence of the presence of the tumor consistent with the “conventional field effect” described above.

Mechanistically, it is intriguing that a number of genes with important roles in cancer-related processes are among the differentially expressed genes. Of the genes that were downregulated in patients with lung cancer, CASP10 and CD177 were among the most correlated genes between bronchial and nasal epithelium and are associated with the induction of apoptosis and activation of the immune response, respectively. We also identified a number of genes involved in the p53 pathway that were downregulated in patients with lung cancer, including BAK1, ST14, CD82, and MUC4. BAK1 is associated with the induction of apoptosis (28,29) and has been previously shown to be a number of genes involved in the p53 pathway that were downregulated in patients with non–small cell lung cancer (NSCLC) (30). ST14 has been described as a tumor suppressor in breast cancer and its overexpression associated with the inhibition of tumor cell migration and cell invasion (31). The downregulation of CD82, which is a metastasis suppressor in prostate cancer (32), has been shown to be correlated with poor survival in patients with lung adenocarcinoma (33). MUC4, whose downregulation has been associated with increased tumor stage and poorer overall survival, has also been shown to play an oncogenic role in multiple cancers and is a tumor suppressor in NSCLC, acting as a modifier of p53 expression (34).

From a clinical perspective, we found that the addition of lung cancer–associated gene expression to established clinical risk factors improved the sensitivity and negative predictive value for detecting lung cancer; these are the key performance metrics for driving potential clinical utility in this setting (eg, allowing physicians to avoid unnecessary invasive procedures in those with benign disease). This provides the first proof of concept for the use of nasal gene expression for lung cancer detection. We elected to establish the presence of a nasal field of lung cancer–associated injury using samples from the AEGIS trial given the unique availability of matched bronchial samples, despite the fact that these patients were undergoing bronchoscopy for suspect lung cancer. The demonstration of the added value of nasal gene expression for lung cancer detection in this setting sets the stage for the development of nasal gene expression biomarkers for lung cancer in other clinical settings where bronchoscopy is not frequently used because of lesion size/location, risk of complications, or cost. In particular, it will now be of interest to develop nasal biomarkers for patients with small peripheral nodules found incidentally or via screening as our current bronchoscopy-based cohort is enriched for patients with centrally located lesions. In the clinical setting of patients with small peripheral nodules, we envision that a nasal biomarker for lung cancer with a low negative likelihood ratio (on par with the NLR we observed for the nasal biomarker in the AEGIS samples) could be used to identify nodule patients who are at low risk of malignancy and can be managed by CT surveillance.

The importance and potential impact of this study derive from several key strengths. First, the patients came from a large number of academic and community hospitals and reflect a variety of practice settings and different geographical locales; thus the diversity of alternative benign diagnoses is represented. Second, the training and validation sets came from two separate clinical trials, which minimizes the potential for the model to depend on locally confounding variables. Third, the samples were prospectively collected and cancer status was unknown at the time of collection. Fourth, we have shown that nasal gene expression identifies a source of lung cancer risk that is independent of major clinical risk factors. There were also a number of important limitations to this study. Nasal samples used in this study were collected from patients undergoing bronchoscopy for clinical suspicion of lung cancer. As a result, our cohort was enriched for patients with larger nodules and an elevated pretest risk of lung cancer. Further, the size of our independent validation set limited our ability to assess the subgroup performance of our biomarker in patients with small and/or peripheral nodules, a clinical setting where the test may have greater clinical utility. Lastly due to

### Table 4. Classifier performance in the validation set (n = 130)

<table>
<thead>
<tr>
<th>Biomarker performance metric</th>
<th>Clinical model</th>
<th>Clinicogenomic model</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area under the curve (95% CI)*</td>
<td>0.74 (0.66 to 0.83)</td>
<td>0.81 (0.74 to 0.89)</td>
<td>.01</td>
</tr>
<tr>
<td>Sensitivity (95% CI)†</td>
<td>0.79 (0.67 to 0.88)</td>
<td>0.91 (0.81 to 0.97)</td>
<td>.03</td>
</tr>
<tr>
<td>Specificity (95% CI)¶</td>
<td>0.58 (0.45 to 0.70)</td>
<td>0.52 (0.39 to 0.64)</td>
<td>.42</td>
</tr>
<tr>
<td>Negative predictive value (95% CI)‡</td>
<td>0.73 (0.58 to 0.84)</td>
<td>0.85 (0.69 to 0.94)</td>
<td>.03</td>
</tr>
<tr>
<td>Positive predictive value (95% CI)‡</td>
<td>0.66 (0.54 to 0.76)</td>
<td>0.66 (0.55 to 0.76)</td>
<td>.97</td>
</tr>
<tr>
<td>Accuracy (95% CI)§</td>
<td>0.68 (0.60 to 0.76)</td>
<td>0.72 (0.63 to 0.79)</td>
<td>.68</td>
</tr>
</tbody>
</table>

*P value comparing models calculated using Delong’s two-sided test. CI = confidence interval.
†P value comparing models calculated using McNemar’s two-sided chi-square test.
‡P value comparing models calculated using two-sided generalized score statistic.
§P value comparing models calculated using two-sided Fisher Exact test.
our nested case/control study design, the disease prevalence in our cohort is not representative of the intended use clinical setting.

Together, the findings demonstrate the existence of a cancer-associated airway field of injury that can be measured in nasal epithelium, a biosample that can be collected noninvasively with little instrumentation or advanced training. Moreover, we find that nasal gene expression contains information about the presence of cancer that is independent of standard clinical risk factors, suggesting that nasal epithelial gene expression might aid in lung cancer detection. These findings, in particular the high NPV of the nasal clinicogenic biomarker, suggest the potential to rule out lung cancer and set the stage for efforts to develop nasal gene expression biomarkers that might have clinical utility in avoiding unnecessary invasive procedures in settings where bronchoscopy is not used as a diagnostic procedure, including small peripheral nodules.

**Funding**

This research was supported by grants from the National Institutes of Health (NIH) Early Detection Research Network (U01CA152751 and U01CA214182 to AS, MEL, SD, and DE), the Department of Defense (DOD W81XWH-11-2-0161 to AS and ML), and the Boston University Coulter Award (0-057-281-A594-5 to AS and MEL).

This study was sponsored by Veracyte, Inc.

**Notes**

The study sponsor was not involved in data generation, analysis, or interpretation. The study sponsor did review a final draft of the manuscript.

The authors would like to thank Adam Gower for his implementation of the weighted voting algorithm and Yaron Gesthalter for his input and revisions to early drafts of the manuscript.

JF, AS, and MEL conceived the study. JPR performed the primary analyses. JG and CA performed preliminary analyses. GL, SZ, and YA performed the experiments. WEJ, DE, and MEL advised statistical analyses. KP and DW provided the samples. JPR, JG, CA, GL, SZ, YA, WEJ, JB, KP, DW, SD, AS, and MEL wrote the manuscript. All authors provided critical input on the manuscript and approved it for publication in its final form.

KP and DW are employees of Veracyte, Inc. AS and MEL are consultants to Veracyte, Inc. Boston University owns patents related to the subject matter of this manuscript.

Microarray CEL files and gene expression data used in this manuscript have been deposited in the NCBI Gene Expression Omnibus under accession number GSE80796.

References


