Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome

Emily S. Wan1,∗, Weiliang Qiu1, Andrea Baccarelli2, Vincent J. Carey1, Helene Bacherman1, Stephen I. Rennard3, Alvar Agusti4, Wayne Anderson5, David A. Lomas6 and Dawn L. DeMeo1

1Channing Laboratory and the Division of Pulmonary and Critical Care Medicine, Brigham and Women’s Hospital, Boston, MA, USA, 2Exposure Epidemiology and Risk Program, Harvard School of Public Health, Boston, MA, USA, 3Section of Pulmonary and Critical Care, University of Nebraska Medical Center, Omaha, NE, USA, 4Thorax Institute, Hospital Clinic, IDIBAPS, Barcelona and CIBERES, Fundacio Caubet-Cimera, Palma de Mallorca, Spain, 5GlaxoSmithKline Research and Development, Research Triangle Park, NC, USA and 6Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK

Received November 16, 2011; Revised March 22, 2012; Accepted April 2, 2012

The impact of cigarette smoking can persist for extended periods following smoking cessation and may involve epigenetic reprogramming. Changes in DNA methylation associated with smoking may help to identify molecular pathways that contribute to the latency between exposure and disease onset. Cross-sectional cohort data from subjects in the International COPD Genetics Network (n = 1085) and the Boston Early-Onset COPD study (n = 369) were analyzed as the discovery and replication cohorts, respectively. Genome-wide methylation data on 27 578 CpG sites in 14 475 genes were obtained on DNA from peripheral blood leukocytes using the Illumina HumanMethylation27K Beadchip in both cohorts. We identified 15 sites significantly associated with current smoking, 2 sites associated with cumulative smoke exposure, and, within the subset of former smokers, 3 sites associated with time since quitting cigarettes. Two loci, factor II receptor-like 3 (F2RL3) and G-protein-coupled receptor 15 (GPR15), were significantly associated in all three analyses and were validated by pyrosequencing. These findings (i) identify a novel locus (GPR15) associated with cigarette smoking and (ii) suggest the existence of dynamic, site-specific methylation changes in response to smoking which may contribute to the extended risks associated with cigarette smoking that persist after cessation.

INTRODUCTION

Tobacco smoking is a leading cause of death and disability worldwide. Smoking attributable deaths are projected to increase over the next two decades and will account for 10% of all deaths globally by 2030 (1). The impact of smoking on cardiovascular (2–4), respiratory (5–9) and malignant (10–12) diseases can persist for extended periods after smoking cessation and may involve epigenetic reprogramming. Identifying molecular pathways that contribute to the delay between exposure and disease may offer opportunities for targeted diagnostic and therapeutic interventions.

In mammals, DNA methylation occurs at the 5′ carbon of cytosine residues in the context of CpG dinucleotides (13) (e.g. a ‘CpG site’) and is a highly plastic, tissue-specific phenomenon which can reflect exposure to exogenous stimuli such as smoking (14). Surprisingly, the number of studies examining the effects of smoking on DNA methylation in blood has been limited. Although relative global hypomethylation has been associated with prenatal exposure to tobacco smoke (15–17), the majority of studies have not demonstrated a significant impact of adult smoking on global DNA methylation content in blood (18–22) (whether assessed directly or through surrogates such as LINE-1 or Alu measurements).

∗To whom correspondence should be addressed at: 181 Longwood Avenue, Boston, MA 02115, USA. Tel: +1 6175252112; Fax: +1 6175250958; Email: reesw@channing.harvard.edu

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
Limited numbers of studies have investigated differential methylation in response to cigarette smoking in candidate gene promoters in peripheral blood leukocytes; these studies have been largely restricted to malignancy-related genes and have enrolled small numbers of subjects (23–26). In contrast, Breitling et al. (14) recently reported the results of a genome-wide analysis in peripheral blood leukocytes using a discovery cohort of 177 subjects followed by replication in a second cohort (n = 316); they identified a previously unsuspected CpG site located within the coagulation factor II receptor-like 3 (F2RL3) locus as being strongly associated with current smoking.

We hypothesized that a number of smoking metrics will influence DNA methylation patterns. To test this hypothesis, we sought to identify CpG sites that demonstrate differential methylation with respect to the current smoking status, smoking exposure (pack-years) and time since quitting in a large, well-characterized cohort of former and current smoking subjects from the International COPD Genetics Network (ICGN) (27,28) (n = 1085) followed by replication in the Boston Severe Early-Onset COPD study (EOCOPD) (n = 369) (29). In addition to independently validating the previously reported association of the F2RL3 locus with current smoking, we identify a novel association at the G-protein-coupled receptor 15 (GPR15) locus. We extend our analyses to investigate associations with cumulative smoking exposure (pack-years) and present evidence that suggests the existence of dynamic DNA methylation changes in response to smoking.

RESULTS

The characteristics of the subjects analyzed in the discovery (ICGN) and replication (EOCOPD) cohorts are summarized in Table 1. By design, COPD subjects in the EOCOPD cohort had a lower mean age and FEV1% predicted than COPD subjects in the ICGN cohort. The proportion of male subjects was greater in the ICGN cohort.

Fifteen CpG sites were significantly associated with current smoking in our cohorts (Table 2). The difference in the mean methylation between current smokers and former and never smokers is illustrated for each cohort. The median and interquartile ranges for each of the 15 sites are listed in Supplementary Material, Table S1. Because our discovery cohort compared current versus former smokers, we repeated our analysis after excluding 68 never smokers from our replication cohort (EOCOPD). Nine sites were significantly associated in this subgroup analysis, eight of which overlap with the primary analysis (Supplementary Material, Table S2). For the two most highly associated sites in both analyses, F2RL3 and GPR15, relative hypomethylation was observed with current smoking. Notably, these two CpG sites were also significantly associated with cumulative exposure to cigarette smoke (Table 3). Increasing cumulative exposure to cigarette smoke is associated with relative hypomethylation at both significant loci (Fig. 1).

Although data in our cohort are cross-sectional, we analyzed the time since quitting cigarettes in former smokers as a surrogate to evaluate the longitudinal effects of smoking on DNA methylation. The significant differentially methylated sites are listed in Table 4 along with the difference in the mean methylation between the highest and lowest quartiles of time since quitting. The three significant loci in the time since quitting analysis after adjustment for obstructive lung disease (EOCOPD, Early-Onset COPD; FEV1% predicted, forced expiratory volume in the first second expressed as % predicted of normal based on population specific prediction equations. Data are expressed as mean (SD) or n (%).

| Variable in current or former smokers only (In ICGN, n = 685, four subjects missing data). |
| Variable defined as FEV1/FVC <0.7. |

on DNA methylation. The significant differentially methylated sites are listed in Table 4 along with the difference in the mean methylation between the highest and lowest quartiles of time since quitting. The three significant loci in the time since quitting analysis after adjustment for obstructive lung disease; notably, for each site, current smoking was associated with decreased methylation, whereas time since smoking cessation was associated with increased methylation (Fig. 2). Among former smokers in the ICGN cohort, exploratory analyses revealed a significant interaction between cumulative smoke exposure and the time since quitting under a multiplicative model for F2RL3 and GPR15 (Table 5). This is illustrated qualitatively for the F2RL3 locus in Figure 3; subjects with the highest cumulative smoke exposure and shortest duration of smoking cessation had the lowest mean methylation, whereas subjects with the lowest cumulative smoke exposure and longest duration of cessation had the highest mean methylation.

Because our cohorts were ascertained on the presence or absence of COPD, we repeated our analyses with the addition of a covariate to adjust for obstructive lung disease. Using the same criteria for significance employed in our original analysis, 13 of the 15 sites remained significantly associated with current smoking. Both F2RL3 and GPR15 remained significantly associated with cumulative smoke exposure as measured in pack-years. F2RL3 remained significant in the time since quitting analysis after adjustment for obstructive lung disease.

Pyrosequencing was successfully completed in the ICGN cohort for the F2RL3 and GPR15 loci; completion rates were 99.6 and 96.9%, respectively. The distributions of the pyrosequencing values with respect to the current smoking status are illustrated in Supplementary Material, Figure S1. Student’s t-test comparing DNA methylation by the current smoking status remained highly statistically significant for

Table 1. Cohort characteristics

<table>
<thead>
<tr>
<th></th>
<th>ICGN</th>
<th>EOCOPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>1085</td>
<td>369</td>
</tr>
<tr>
<td>Age (years)</td>
<td>57.3 (8.1)</td>
<td>47.5 (7.1)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>590 (54.4%)</td>
<td>132 (35.8%)</td>
</tr>
<tr>
<td>Female</td>
<td>509 (45.6%)</td>
<td>237 (64.2%)</td>
</tr>
<tr>
<td>Pack-years</td>
<td>41.7 (26.2)</td>
<td>28.7 (23.6)</td>
</tr>
<tr>
<td>Age started smoking (years)</td>
<td>16.3 (4.1)</td>
<td>16.3 (4.1)</td>
</tr>
<tr>
<td>Never smokers</td>
<td>0</td>
<td>68 (18.4%)</td>
</tr>
<tr>
<td>Current smokers</td>
<td>396 (36.5%)</td>
<td>103 (27.9%)</td>
</tr>
<tr>
<td>Former smokers</td>
<td>689 (63.5%)</td>
<td>198 (53.7%)</td>
</tr>
<tr>
<td>Time since quitting (years)</td>
<td>12.3 (10.4)</td>
<td>7.7 (7.5)</td>
</tr>
<tr>
<td>Obstructive lung diseasea</td>
<td>Yes = 657</td>
<td>No = 428</td>
</tr>
<tr>
<td>FEV1% predicted</td>
<td>45.4 (17.7)</td>
<td>104.6 (15.8)</td>
</tr>
<tr>
<td></td>
<td>38.0 (27.6)</td>
<td>89.7 (16.6)</td>
</tr>
</tbody>
</table>

ICGN, International COPD Genetics Network; EOCOPD, Early-Onset COPD; FEV1% predicted, forced expiratory volume in the first second as % predicted of normal based on population specific prediction equations.
DNA methylation were also confirmed.

Complex chronic exposures such as cigarette smoking can pose challenges in the analysis and interpretation of DNA methylation data. However, the careful analysis of large data sets with detailed phenotype information, along with rigorous statistical thresholds for significance, can reveal novel loci and pathways which may be involved in mediating the effects of smoking in the development of disease. Significantly associated sites from our analyses have been annotated to genes involved in a wide range of functions including cell signaling/apoptosis (FASLG), neurodevelopment (LRRN3, CNTNAP2) and alcohol metabolism (ADHFE1). Myosin light chain kinase (MYLK) has been ascribed putative roles in endothelial (30) and epithelial (31–34) barrier functions; variants in this gene have been associated with acute lung injury (35,36), asthma (37–39) and familial aortic dissections (40). Interestingly, cg09837977 (LRRN3), cg12044210 and cg21917349 (APBA2), and cg10161121 (FASLG) map to loci that have previously been included in a smoking ‘quit-success’ genotype score (41).

**Table 2.** Significant differently methylated CpG sites by current smoking

<table>
<thead>
<tr>
<th>CpG</th>
<th>Gene</th>
<th>ICGN Change in mean methylation</th>
<th>FDR, P-value</th>
<th>EOCOPD Change in mean methylation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg03636183</td>
<td>F2RL3</td>
<td>−0.08</td>
<td>2.36 × 10⁻⁶</td>
<td>−0.06</td>
<td>6.87 × 10⁻¹¹</td>
</tr>
<tr>
<td>cg19859270</td>
<td>GPR15</td>
<td>−0.02</td>
<td>1.92 × 10⁻²</td>
<td>−0.02</td>
<td>3.18 × 10⁻⁷</td>
</tr>
<tr>
<td>cg099837977</td>
<td>LRRN3</td>
<td>−0.02</td>
<td>3.88 × 10⁻¹</td>
<td>−0.02</td>
<td>5.45 × 10⁻⁵</td>
</tr>
<tr>
<td>cg01500140</td>
<td>LIM2</td>
<td>0.01</td>
<td>1.88 × 10⁻¹⁰</td>
<td>0.01</td>
<td>3.83 × 10⁻⁵</td>
</tr>
<tr>
<td>cg13247990</td>
<td>MYLK</td>
<td>0.01</td>
<td>3.24 × 10⁻⁷</td>
<td>0.01</td>
<td>2.42 × 10⁻⁴</td>
</tr>
<tr>
<td>cg01988129</td>
<td>ADHFE1</td>
<td>0.01</td>
<td>4.16 × 10⁻⁶</td>
<td>0.01</td>
<td>4.12 × 10⁻⁵</td>
</tr>
<tr>
<td>cg16254309</td>
<td>CNTNAP2</td>
<td>−3.36 × 10⁻³</td>
<td>1.74 × 10⁻⁵</td>
<td>0.01</td>
<td>3.48 × 10⁻⁴</td>
</tr>
<tr>
<td>cg18881723</td>
<td>SLAMF1</td>
<td>−0.01</td>
<td>1.69 × 10⁻⁴</td>
<td>−0.01</td>
<td>3.56 × 10⁻⁴</td>
</tr>
<tr>
<td>cg12044210</td>
<td>APBA2</td>
<td>1.74 × 10⁻³</td>
<td>3.47 × 10⁻⁴</td>
<td>0.01</td>
<td>5.51 × 10⁻⁴</td>
</tr>
<tr>
<td>cg21917349</td>
<td>APBA2</td>
<td>1.66 × 10⁻³</td>
<td>5.49 × 10⁻⁴</td>
<td>0.01</td>
<td>5.73 × 10⁻⁵</td>
</tr>
<tr>
<td>cg15691199</td>
<td>CEBPE</td>
<td>−3.87 × 10⁻³</td>
<td>2.89 × 10⁻³</td>
<td>−3.22 × 10⁻³</td>
<td>2.66 × 10⁻⁴</td>
</tr>
<tr>
<td>cg24262469</td>
<td>TIPARP</td>
<td>−1.96 × 10⁻³</td>
<td>0.01</td>
<td>0.01</td>
<td>7.27 × 10⁻⁵</td>
</tr>
<tr>
<td>cg05445326</td>
<td>TM4SF19</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>7.00 × 10⁻⁴</td>
</tr>
<tr>
<td>cg15258980</td>
<td>ARHGAP25</td>
<td>−3.11 × 10⁻³</td>
<td>0.02</td>
<td>−4.49 × 10⁻³</td>
<td>4.14 × 10⁻⁴</td>
</tr>
<tr>
<td>cg10161121</td>
<td>FASLG</td>
<td>−0.01</td>
<td>−2.81 × 10⁻³</td>
<td>0.01</td>
<td>7.93 × 10⁻⁴</td>
</tr>
</tbody>
</table>

ICGN, International COPD Genetics Network; EOCOPD, Early-Onset COPD; FDR, false-discovery rate.

*Defined as sites with an FDR adjusted P-value < 0.05 in the discovery cohort (ICGN) and a one-sided P-value < 10⁻³ in the replication cohort (EOCOPD).

*Change in mean methylation = (mean methylation in current smokers) – (mean methylation in former and never smokers). Methylation values may range from 0 (no methylation) to 1 (fully methylated).

*One-sided P-value.

*In model adjusted for covariates direction of effect is positive.

**Table 3.** Significant differently methylated CpG sites associated with cumulative smoking exposure (pack-years)

<table>
<thead>
<tr>
<th>CpG</th>
<th>Gene</th>
<th>ICGN Mean methylation in highest pack-years quartile</th>
<th>ICGN Mean methylation in lowest pack-years quartile</th>
<th>Difference FDR, P-value</th>
<th>EOCOPD Mean methylation in highest pack-years quartile</th>
<th>EOCOPD Mean methylation in lowest pack-years quartile</th>
<th>Difference P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg03636183</td>
<td>F2RL3</td>
<td>0.76</td>
<td>0.84</td>
<td>−0.07</td>
<td>1.91 × 10⁻⁸</td>
<td>0.79</td>
<td>0.05</td>
</tr>
<tr>
<td>cg19859270</td>
<td>GPR15</td>
<td>0.89</td>
<td>0.90</td>
<td>−0.01</td>
<td>6.35 × 10⁻⁶</td>
<td>0.85</td>
<td>0.04</td>
</tr>
</tbody>
</table>

ICGN, International COPD Genetics Network; EOCOPD, Early-Onset COPD; FDR, false-discovery rate.

Mean methylation values may range from 0 (no methylation) to 1 (fully methylated).

Mean value of pack-years by quartile in ICGN: Quartile 1 = 14.4, Quartile 2 = 30.3, Quartile 3 = 44.4, Quartile 4 = 77.9.

Mean value of pack-years by quartile in EOCOPD: Quartile 1 = 1.3, Quartile 2 = 20.0, Quartile 3 = 34.4, Quartile 4 = 60.2.

*Defined as sites with an FDR adjusted P-value < 0.05 in the discovery cohort (ICGN) and a one-sided P-value < 10⁻³ in the replication cohort (EOCOPD).

*One-sided P-value.

*Both F2RL3 (P-value 2.2 × 10⁻¹⁶) and GPR15 (P-value 1.5 × 10⁻¹²) and again demonstrated relative hypomethylation associated with current cigarette smoking for each locus. Consistent with the array-based data, a significant inverse correlation between DNA methylation and pack-years of smoking (P F2RL3 = 8.54 × 10⁻⁹ and P GPR15 = 7.68 × 10⁻⁸) and a significant positive correlation between time since quitting (P F2RL3 = 1.35 × 10⁻¹¹ and P GPR15 = 7.58 × 10⁻⁸) and DNA methylation were also confirmed.

**DISCUSSION**

Complex chronic exposures such as cigarette smoking can pose challenges in the analysis and interpretation of DNA methylation data. However, the careful analysis of large data sets with detailed phenotype information, along with rigorous statistical thresholds for significance, can reveal novel loci and pathways which may be involved in mediating the effects of smoking in the development of disease. Significantly associated sites from our analyses have been annotated to genes involved in a wide range of functions including cell signaling/apoptosis (FASLG), neurodevelopment (LRRN3, CNTNAP2) and alcohol metabolism (ADHFE1). Myosin light chain kinase (MYLK) has been ascribed putative roles in endothelial (30) and epithelial (31–34) barrier functions; variants in this gene have been associated with acute lung injury (35,36), asthma (37–39) and familial aortic dissections (40). Interestingly, cg09837977 (LRRN3), cg12044210 and cg21917349 (APBA2), and cg10161121 (FASLG) map to loci that have previously been included in a smoking ‘quit-success’ genotype score (41).

F2RL3 and GPR15, the two loci with the strongest associations with current smoking, have also been annotated...
to seemingly disparate functions. F2RL3, also known as protease-activated receptor 4, has established roles in platelet activation (42,43) and cell signaling (44); differential methylation at this site with respect to current smoking was first reported by Breitling et al. (14). GPR15 is a protein which shares sequence homology with the angiotensin II AT1 and AT2 receptors (45) and has been established as a co-receptor for the human immunodeficiency virus (46–48). Although there have been no previous statistically significant associations of this locus with smoking, cg19859270 was notably ranked second in the analysis conducted by Breitling et al. (14) and demonstrated a similar magnitude and direction of the effect.

We acknowledge that, despite robust statistical significance, the effect estimates of many of the loci reported in this manuscript are modest. Even the effect estimates for F2RL3 and GPR15 do not meet the 15% change often cited as a criterion for significance. However, considerable evidence supports the accuracy of our estimates. The differences in the mean methylation between current smokers and non-smokers for F2RL3 and GPR15 do not appear to be artifacts of the array-based platform as pyrosequencing produced similar differences in the mean methylation for each site. In addition, the differences in the mean methylation are comparable with those reported in the independent populations studied by Breitling et al. (14). We contend, therefore, that large-scale epigenetic studies are a nascent field and that the threshold of change in DNA methylation which correlates with functional significance on the population level has not been firmly established.

To examine the possible functional impact of differential methylation at loci reported in this manuscript, we conducted exploratory analyses using publicly available data sets to assess for changes in gene expression by the current smoking status in blood (49–53), histologically normal lung tissue (54) and bronchial epithelium (55) (Supplementary Material, Table S3). Using Student’s t-test to compare the processed, log-transformed expression values in current smokers versus former/never smokers, several loci demonstrated evidence of nominal association (unadjusted P-value < 0.05). Increased gene expression in current smokers was noted for GPR15 in one of two peripheral blood data sets with transcripts available, whereas FASLG, the sole overlapping locus where differential gene expression with respect to the smoking status has previously been reported in blood (50), was found to have decreased expression in smokers in three data sets. Interestingly, increased gene expression in current smokers at LRRN3, a locus which has not been previously reported in either methylation or expression studies, was noted in four of five peripheral blood data sets (with the fifth data set potentially limited by small sample size). FASLG and LRRN3 were not significantly associated in either lung or bronchial epithelial tissue while gene expression at MYLK, which was reported to demonstrate rapidly reversible changes in bronchial epithelium (55), did not vary by the current smoking status in blood. These differences may be attributable to the tissue specificity of gene expression and DNA methylation patterns.

Although our data are cross-sectional in nature, our analyses suggest that dynamic changes in DNA methylation in response to smoking may exist for F2RL3 and GPR15. The inverse correlation between increasing pack-years and methylation at F2RL3 and GPR15 is consistent with progressive demethylation in the setting of higher cumulative smoke exposure. Likewise, the positive correlation between time since quitting and methylation at these loci may be due to remethylation following smoking cessation.

The time course over which DNA methylation changes occur is not known and may vary by CpG site. The concept of differential responses to and recovery from exposure to cigarette smoke was well articulated in a gene expression study conducted in primary bronchial epithelial cells (55); while the majority of gene expression changes due to smoking were found to be rapidly reversible, a subset of loci demonstrated slow or irreversible changes. Studies performed in respiratory epithelial cell cultures exposed to cigarette smoke extract support the concept that distinct loci may be involved in short-term versus chronic exposure to cigarette smoke. In
this study, changes in DNA methylation were observable after only 9 months of cigarette smoke exposure (56); none of the differentially methylated sites overlapped with our loci.

To explore whether differential methylation at the loci reported in our study might be involved in rapidly reversible versus slowly reversible-type changes, we conducted additional subgroup analyses in a subset of the public gene expression data sets used in Supplementary Material, Table S3 which had data on former smokers. Although data on time since quitting was not available for former smokers in these cohorts, we hypothesized that genes which were differentially expressed between current and former smokers might represent relatively ‘rapidly reversible’ loci (Supplementary Material, Table S4). Conversely, we speculated that differential expression between former and never smokers might represent ‘slowly reversible’ or irreversible changes (Supplementary Material, Table S5). MYLK, a locus reported to have ‘rapidly reversible’ changes in gene expression in bronchial epithelial cells (55), conforms to our hypotheses well; differential expression is seen between current and former smokers, but not between former and never smokers (in bronchial epithelial cells). LRRN3, relatively hypomethylated in current smokers (Table 2) and relatively hypermethylated following smoking cessation (Table 4), appears to fit this speculative model of a ‘rapidly reversible’ locus with differential (increased) expression evident between current and former smokers in peripheral blood, but no differences between former and never smokers. In contrast, differential expression in peripheral blood at GPR15 and FASLG was observed in both the current/former and former/never smoker analyses and suggests that these loci may demonstrate ‘slowly reversible’ changes. Although intriguing, we acknowledge that only very limited conclusions can be drawn from these unadjusted exploratory analyses conducted in non-contemporaneous data sets and that additional studies are warranted.

We recognize the limitation of association studies as providing only correlative data; we cannot establish whether the changes in methylation reported herein are primary (i.e. established directly by smoking exposure) or secondary to more generalized processes (i.e. established as a consequence of perturbations in other cellular processes). Gene expression (55) and DNA methylation (56) changes may both persist in non-malignant tissues for extended periods following smoking cessation; the mechanism by which these changes are established and maintained is a topic of active research (57). Changes in methylation marks at specific loci may arise from a change in DNA methyltransferase expression, abundance or activity. Indeed, changes in DNA methyltransferase expression and levels have been reported in neurons (58) and human bronchial epithelial cell cultures (59) exposed to nicotine or cigarette smoke extract. However, we found no compelling evidence for differential expression in the DNA methyltransferases by current smoking status in blood (49–53,55), histologically normal lung tissue (54) or bronchial epithelium (55) based on our exploration of public data sets (Supplementary Material, Table S6). Alternative mechanisms, beyond changes in methyltransferase gene expression, may contribute to the establishment of locus-specific differential methylation observed in our study; additional in vitro and in vivo studies are needed to explore this question.

Potential limitations of this study include the inability to quantify the degree to which cell type heterogeneity in blood contributes to our results and the reliance upon self-reported smoking metrics. Epidemiological studies have found self-reported smoking to be reliable except in cases where non-smoking is perceived as socially desirable (59–61). Because misclassification due to under-reporting of current smoking would bias our results towards the null, we believe our results are robust even in the absence of biochemical confirmation of current smoking status (i.e. serum cotinine measurements).

Lastly, the frequency and impact of genetic and structural variants on the array-based platforms used to assess DNA methylation in our study is unknown and is an area of active research (62). We performed sequence alignment of the probes of significantly associated sites against a consensus human genome (NCBI37/hg19) and noted a number of single nucleotide polymorphisms, copy number variants and repetitive elements overlapping with our probes (Supplementary Material, Table S7). To assess the impact of these variants on our methylation data, we plotted the distribution of the raw methylation values of current and non-smokers (Supplementary Material, Fig. S2) and the distribution of the raw methylation values against the continuous variable of pack-years of smoking (Supplementary Material, Fig. S3). We saw no evidence of distinct subpopulations indicative of genetic or structural variants among significantly associated sites.

<table>
<thead>
<tr>
<th>CpG</th>
<th>Gene</th>
<th>ICGN</th>
<th>EOCOPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg03636183</td>
<td>F2RL3</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>cg19859270</td>
<td>GPR15</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td>cg09837977</td>
<td>LRRN3</td>
<td>0.84</td>
<td>0.82</td>
</tr>
</tbody>
</table>

ICGN, International COPD Genetics Network; EOCOPD, Early-Onset COPD; FDR, false-discovery rate

Mean value of time since quit (in years) by quartiles in ICGN: Quartile 1 = 1.5, Quartile 2 = 6.3, Quartile 3 = 14.2, Quartile 4 = 27.4.
Mean value of time since quit (in years) by quartiles in EOCOPD: Quartile 1 = 1.3, Quartile 2 = 3.5, Quartile 3 = 7.4, Quartile 4 = 18.7.

1Defined as sites with an FDR adjusted-P-value <0.05 in the discovery cohort (ICGN) and a one-sided P-value <10⁻² in the replication cohort (EOCOPD).

1One-sided P-value.
Additionally, despite the report of structural variants involving the GPR15 locus, the association of this locus remained robust when assessed through a second technology (pyrosequencing).

Given the profound physiological impact of smoking as well as considerable epidemiological evidence linking smoking to a number of systemic diseases beyond directly exposed organs, the effects of smoking on peripheral blood DNA methylation appear to impact a limited number of loci in our cohorts. In addition, since gene expression changes were not noted in our exploratory analyses at the majority of loci reported in our study, the impact of differential methylation at these loci remains to be established. Thus, the integration of epigenetic data with genetic and genomic lines of investigation will likely yield additional biological insights into clinical challenges such as smoking cessation and smoking-related complex diseases.

**MATERIALS AND METHODS**

**Subjects/cohorts**

The discovery cohort consisted of subjects participating in the ICGN (27,28), a family-based study which recruited probands in the age range of 45–65, with ≥5 pack-years of cigarette smoking and airflow limitation (defined as an FEV$_1$ < 60% predicted and an FEV$_1$/VC < 90% predicted). Subjects with alpha-1 antitrypsin genotypes PiZZ, Z-Null, Null–Null or SZ were excluded from the study. In addition, all probands were required to have at least one eligible sibling with ≥5 pack-years of cigarette smoking. All subjects completed spirometry, questionnaire data and had whole blood drawn for DNA analysis. Subjects included in the current analysis were of self-reported Caucasian ancestry. The protocol was approved by the local Internal Review Board at each of the recruitment sites and all subjects provided written informed consent. Details regarding the number of samples excluded at each step during data cleaning have been previously described (63) and are summarized in Supplementary Material, Methods; the number of subjects analyzed from this cohort was 1085.

The replication cohort consisted of subjects recruited from the Boston Severe EO COPD (29), a family-based study which recruited probands who had an FEV$_1$ < 40% predicted, by age < 53 in the absence of severe alpha-1 antitrypsin deficiency. Although extended pedigrees were recruited, DNA methylation data were obtained only on probands and their siblings. Subjects completed spirometry, questionnaire data and provided blood for DNA analysis. The protocol was
approved by the Brigham and Women’s Hospital Human Research Committee; all subjects provided written informed consent. Sample processing and quality control measures were identical to those employed for the ICGN cohort as outlined above; the number of subjects from the EOCOPD cohort included in our analysis was 369.

Variable definitions
Current smoking was defined as an affirmative response to the question: ‘Do you currently smoke cigarettes (as of 1 month ago)?’ The average number of cigarettes smoked per day was the self-reported number of cigarettes smoked per day during the period when the subject smoked. Pack-years was calculated according to the number of years smoked multiplied by the average number of cigarettes smoked per day divided by 20. In former smokers, the time since quitting was calculated as the subject’s age minus the age when they stopped smoking (in years). Obstructive lung disease was defined as an FEV₁/FVC ratio <0.7.

DNA extraction and bisulfite treatment
For both cohorts, DNA extraction from white blood cells was performed using the Gentra Puregene blood kit (Qiagen, Inc., Valencia, CA, USA). Bisulfite modification of 1 mg of DNA was performed using the 96-well EZ DNA Methylation Gold Deep-well kit (Zymo Research, Orange, CA, USA). The protocol is outlined in Supplementary Material, Methods. Less than 2% of DNA samples failed bisulfite conversion.

DNA methylation array
Both cohorts were assessed using the Illumina (San Diego, CA, USA) Infinium HumanMethylation27K BeadChip, which assays 27 578 CpG sites in 14 475 consensus coding sequences in the NCBI Database (Genome Build 36). Bisulfite converted DNA was amplified, fragmented and hybridized to locus-specific 50-mer probes on the arrays, followed by single base extension using biotin (C and G) or dinitrophenyl (A and T) labeled nucleotides, per routine. All arrays were imaged using an Illumina BeadArray Reader. Beadstudio software (methylation module version 3.2) was used to assemble the workspace. Each locus results in an intensity value for methylation (M) and unmethylated (U) alleles. The methylation status was expressed as a beta value (β), which represents a ratio of the M to U fluorescent signals such that $\beta = \text{Max}(M,0)/[\text{Max}(M,0)+\text{Max}(U,0)+100]$. Using this metric, DNA methylation is represented by a variable between 0 (no methylation) to 1 (complete methylation). Raw average beta values were exported and analyzed without normalization. CpG sites will be discussed using the industry assigned identifier (‘cg’) as outlined in the Illumina HumanMethylation27K
Quality control measures for array data are described in Supplementary Material, Methods.

**Pyrosequencing**

We performed pyrosequencing using the Pyromark96MD (Qiagen, Inc.) for technical validation of cg03636183 (F2RL5) and cg19859270 (GPR15) in the ICGN cohort. Details are provided in Supplementary Material. CpG methylation percentages were calculated using the formula C/(C + T) x 100 as implemented in the Pyromark CpG software (version 1.0.11).

**Statistical analysis**

Surrogate variable analysis (64,65), which identifies and estimates the effects of known and unknown sources of variation in high-throughput experiments, was performed to account for batch effects. To account for familial correlations in the data, generalized linear mixed models (66–69) as implemented in R (release 2.12) were used to analyze the methylation data. We used an identity link function in our models and the penalized quasi-likelihood method (66) to estimate the model parameters. The Illumina ‘beta’ value was modeled as the dependent variable, whereas the smoking metric (current smoking, pack-years and time since quitting cigarettes) was included as the independent variable. All models were adjusted for age, gender and significant surrogate variables. CpG sites with a false-discovery rate (FDR) adjusted P-value < 0.05 (70) in the discovery cohort (ICGN) were examined using the same models in the replication cohort (EOCOPD). CpG sites demonstrating a P-value < 10^{-3} in the replication cohort with the same direction of effect were considered significant. To assess the impact of COPD on our results, we repeated our analysis with the addition of a covariate to account for obstructive lung disease.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG online.

**FUNDING**

This work was supported by the National Institutes of Health (R01 HL089438, R01 HL075478, T32 HL007427, K12 HL089990, P01 HL105339); the Doris Duke Foundation Clinician Scientist Development Award (to D.L.D.); and the National Institute of Environmental Health Sciences-Harvard School of Public Health New Investigator Fund (ES00002 to A.B.). None of the funding sources played a role in the management, analysis or interpretation of the data, nor did they influence the preparation, review or decision to submit for publication.

**REFERENCES**


