Variants in two adjacent genes, \textit{EGLN2} and \textit{CYP2A6}, influence smoking behavior related to disease risk via different mechanisms

A. Joseph Bloom\textsuperscript{1,*}, Timothy B. Baker\textsuperscript{2}, Li-Shiun Chen\textsuperscript{1}, Naomi Breslau\textsuperscript{3}, Dorothy Hatsukami\textsuperscript{4}, Laura J. Bierut\textsuperscript{1} and Alison Goate\textsuperscript{1}

\textsuperscript{1}Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA, \textsuperscript{2}University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA, \textsuperscript{3}Department of Epidemiology, Michigan State University, East Lansing, MI, USA and \textsuperscript{4}University of Minnesota, Minneapolis, MN, USA

Received April 11, 2013; Revised August 27, 2013; Accepted September 3, 2013

Genome-wide significant associations with cigarettes per day (CPD) and risk for lung cancer and chronic obstructive pulmonary disease (COPD) were previously reported in a region of 19q13, including \textit{CYP2A6} (nicotine metabolism enzyme) and \textit{EGLN2} (hypoxia response). The associated single nucleotide polymorphisms (SNPs) were assumed to be proxies for functional variation in \textit{CYP2A6}. Here, we demonstrate that when \textit{CYP2A6} and \textit{EGLN2} genotypes are analyzed together, the key \textit{EGLN2} variant, rs3733829, is not associated with nicotine metabolism independent of \textit{CYP2A6}, but is nevertheless independently associated with CPD, and with breath carbon monoxide (CO), a phenotype associated with cigarette consumption and relevant to hypoxia. SNPs in \textit{EGLN2} are also associated with nicotine dependence and with smoking efficiency (CO/CPD). These results indicate a previously unappreciated novel mechanism behind genome-wide significant associations with cigarette consumption and disease risk unrelated to nicotine metabolism.

INTRODUCTION

Despite reduced prevalence, tobacco use remains the largest cause of preventable mortality in the USA and worldwide (1). Smoking phenotypes are highly complex but nevertheless strongly influenced by heritable factors (2–4), and therefore genetic studies provide a tool to reveal their underlying biology and improve smoking cessation treatment. Although multiple unbiased genetic studies have been performed and many candidates otherwise investigated, thus far genes in only two pathways have consistently demonstrated associations with smoking behaviors (5–7). The loci most strongly and consistently associated with nicotine dependence, cigarettes per day (CPD), and smoking cessation, are the nicotinic receptor genes \textit{CHRNA5–CHRNA3–CHRNB4} (direct targets of nicotine in the central nervous system) at 15q25(5,7–12), and the primary nicotine metabolism gene \textit{CYP2A6} (13,14), at 19q13. \textit{CYP2A6} was identified in two large genome-wide association studies (GWAS) of CPD in European subjects. In one study, a single nucleotide polymorphism (SNP), rs4105144, located 5′ of \textit{CYP2A6}, showed the strongest association with CPD in that chromosomal region (7). The same study also reported a nominal association between rs4105144 and lung cancer (7). rs4105144 is in tight linkage disequilibrium ($D^′ = 1$) with most of the important functional polymorphisms in \textit{CYP2A6} (5,19). A large body of evidence has shown that nicotine metabolism predicts smoking behaviors including cessation.
(20–27), and variation in hepatic nicotine metabolism is strongly
determined by CYP2A6 genotype (14,28). However, as we previ-
ously reported, rs3733829 appears to be significantly associated
with cigarette consumption independent of CYP2A6 genotype
(15). Importantly, rs3733829 is not associated with metabolism
of nicotine to cotinine independent of CYP2A6 genotype (15)—
i.e. rs3733829 does not appear to account for additional variation
in CYP2A6 or otherwise influence nicotine metabolism.
rs3733829 is also associated with dichotomous nicotine depend-
ence (19), unlike SNP in CYP2A6 (15), further evidence that varia-
tion in CYP2A6 and EGLN2 affect smoking behaviors via
different mechanisms.

Approximately 1.5% of the human genome is transcriptional-
ly responsive to hypoxic conditions (29,30), including carbon
monoxide (CO) (31) and cigarette smoke exposure (32).
EGLN2 a.k.a. PHD1 a.k.a. hypoxia-inducible factor prolyl
hydroxylase (HIF-PHI) is a key component of the cellular
oxygen-sensing pathway that regulates the expression of many
downstream genes (33–36). EGLN2 is one of three similar
genes that act in the hypoxia–response pathway and is widely
expressed in tissues including brain, lung and muscle. EGLN
gene products respond to hypoxia by acting upon multiple
targets (37–40), but primarily they modify the key transcription
regulator hypoxia-inducible factor (HIF). Under normal condi-
tions the EGLNs hydroxylate prolyl residues on HIFα subunits
(HIF1α, 2α or 3α) targeting them for ubiquitin-mediated prote-
olysis. EGLN2 is not required for survival (38,39), but, uniquely,
EGLN2 deficiency causes acute hypoxia tolerance in skeletal muscle and reduced exercise performance due to a
shift from oxidative to anaerobic metabolism (42).

Here, we use exhaled CO, a sensitive measure of cigarette con-
sumption highly relevant to hypoxia and disease risk, to evaluate
the independent effects of genetic variation in CYP2A6 and
EGLN2. Our results indicate hitherto unappreciated GWAS sig-
nificant associations (5,18) and open up a novel biological
pathway to investigation regarding variation in smoking behav-
ior and related disease risk.

RESULTS

The CYP2A6 locus is highly heterogeneous and no single SNP
can act as a proxy for CYP2A6 activity. Therefore, in initial ana-
lyses we used a predictive model of CYP2A6 activity based on
CYP2A6 genotype to capture the genetic diversity of this locus
in a single variable, as previously described (15,28). In a single
SNP analysis, the EGLN2 SNP rs3733829 is significantly associ-
ated with categorical CPD in the Transdisciplinary Tobacco Use
Research Center (UW-TTURC) subjects (P = 0.026, β = 0.06 ± 0.03, n = 1395, Table 1).
However, it does not remain significantly associated (P =

![Figure 1. Relative positions on chromosome 19 of CYP2A6, EGLN2 and RAB4B
genes and variants associated with cigarette consumption and related disease in GWAS (5,18).](image-url)
0.11) in a multivariate model that also includes the CYP2A6 variable, while the CYP2A6 variable itself is strongly associated with CPD in the multivariate model \( (P = 7.1 \times 10^{-4}, \beta = 0.11 \pm 0.03) \). By comparison, both rs3733829 and the CYP2A6 variable are strongly associated with exhaled CO in single variable analyses \( (P = 3.8 \times 10^{-5}, \beta = 1.98 \pm 0.48 \text{ ppm}, \ P = 2.3 \times 10^{-5}, \beta = 2.07 \pm 0.49 \text{ ppm}, \) respectively, \( n = 1355 \), Table 1) and are independently associated with CO in a multivariate analysis \( (P = 1.2 \times 10^{-3}, \beta = 1.59 \pm 0.49 \text{ ppm} \) and \( P = 3.6 \times 10^{-4}, \beta = 1.77 \pm 0.49 \text{ ppm} \) respectively, Table 1). rs3733829 also remains significantly associated with CO \( (P = 1.0 \times 10^{-3}, \beta = 1.69 \pm 0.50 \text{ ppm}) \) in a multivariate model that includes all of the CYP2A6 reduced function alleles common in Europeans \( (CYP2A6*1A, *2, *4, *9 \text{ and } *12) \) as separate variables (Table 2). This is in contrast to the lack of a significant association between rs3733829 and nicotine metabolism in a parallel multivariate analysis of the previously described nicotine metabolism data (28) in which the CYP2A6 reduced function alleles show large and highly significant associations (Table 3). rs7937 (minor allele frequency = 49%), a variant in the 3’ UTR of RAB4B (Fig. 1) in linkage disequilibrium with rs3733829 \( (R^2 = 0.34, D’ = 0.86 \text{ in TTURC European Americans, Fig. 2}) \), was also previously identified as GWAS significantly associated with both CPD and COPD (7,18). rs7937 is nominally associated with CO \( (P = 0.011) \) in TTURC, but not independent of rs3733829 \( (P = 0.8, \) Table 1). rs7937 was not significantly associated with CPD.

The contrast in the associations of rs3733829 with CO and CPD, respectively, led us to investigate a previously used measure of smoking efficiency, CO/CPD (43,44). rs3733829 is significantly associated with CO/CPD \( (P = 5.8 \times 10^{-3}, \beta = 0.58 \pm 0.21 \text{ ppm/CPD}) \), as is the key functional SNP in the nicotinic subunit gene CHRNA5, rs16969968 (9), \( (P = 6.6 \times 10^{-4}, \beta = 0.78 \pm 0.21 \text{ ppm/CPD}) \). The genotype-predicted CYP2A6 metric is not significantly associated with CO/CPD in these subjects \( (P = 0.07, \) Table 1) nor were individual CYP2A6 alleles in a multivariate regression analysis (Table 4); however, the effect sizes for the complete loss-of-function alleles \( (CYP2A6*2, *4, *12) \) were consistent with each other (Table 4), and failure to detect a significant association with CO/CPD may be due to lack of power. Power to detect an association with CO/CPD (assuming a minor allele frequency of 3%) was >97% for an effect size of 1.0 ppm/CPD (see Table 4), but only 59% for an effect size of 0.5 ppm/CPD.

**DISCUSSION**

Several candidate pathways have been investigated regarding their influence upon smoking behaviors, but only the nicotinic receptor subunit and nicotine metabolism genes have shown consistent evidence of association (5–7). Unaccounted for

---

**Table 2.** Multivariate analysis of associations between CO and EGLN2/CYP2A6 genotype in TTURC

<table>
<thead>
<tr>
<th>Variable</th>
<th>( n^a )</th>
<th>MAF(^b)</th>
<th>( \beta \pm \text{SE}^c )</th>
<th>( P)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP2A6*1A</td>
<td>523</td>
<td>0.19</td>
<td>-0.48 ± 0.63</td>
<td>0.4</td>
</tr>
<tr>
<td>CYP2A6*2</td>
<td>78</td>
<td>0.03</td>
<td>-2.62 ± 1.43</td>
<td>0.1</td>
</tr>
<tr>
<td>CYP2A6*4</td>
<td>52</td>
<td>0.02</td>
<td>-2.18 ± 1.47</td>
<td>0.1</td>
</tr>
<tr>
<td>CYP2A6*9</td>
<td>204</td>
<td>0.08</td>
<td>-1.15 ± 0.85</td>
<td>0.2</td>
</tr>
<tr>
<td>CYP2A6*12</td>
<td>69</td>
<td>0.03</td>
<td>-3.92 ± 1.49</td>
<td>0.008</td>
</tr>
<tr>
<td>EGLN2 rs3733829</td>
<td>957</td>
<td>0.35</td>
<td>1.69 ± 0.50</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Associations between exhaled CO and genetic variables modeled as components of a multivariate model including age, sex and study as covariates in 1355 subjects (2710 chromosomes).

\(^a\) Number of minor allele haplotypes.

\(^b\) Minor allele frequency.

\(^c\) Effect size ± standard error in ppm.

**Table 3.** Multivariate analysis of associations between nicotine metabolism and EGLN2/CYP2A6 genotype in COGEND

<table>
<thead>
<tr>
<th>Variable</th>
<th>( n^a )</th>
<th>MAF(^b)</th>
<th>( \beta \pm \text{SE}^c )</th>
<th>( P)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP2A6*1A</td>
<td>57</td>
<td>0.15</td>
<td>-0.04 ± 0.01</td>
<td>4.6 \times 10^{-6}</td>
</tr>
<tr>
<td>CYP2A6*2</td>
<td>12</td>
<td>0.03</td>
<td>-0.18 ± 0.02</td>
<td>4.5 \times 10^{-13}</td>
</tr>
<tr>
<td>CYP2A6*4</td>
<td>6</td>
<td>0.02</td>
<td>-0.21 ± 0.03</td>
<td>5.6 \times 10^{-13}</td>
</tr>
<tr>
<td>CYP2A6*9</td>
<td>24</td>
<td>0.06</td>
<td>-0.07 ± 0.01</td>
<td>2.5 \times 10^{-6}</td>
</tr>
<tr>
<td>CYP2A6*12</td>
<td>9</td>
<td>0.02</td>
<td>-0.18 ± 0.02</td>
<td>3.2 \times 10^{-14}</td>
</tr>
<tr>
<td>EGLN2 rs3733829</td>
<td>137</td>
<td>0.36</td>
<td>0.01 ± 0.01</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Associations between the percent of deuterated nicotine metabolized to cotinine in plasma 30 min after oral nicotine administration and genetic variables modeled as components of a multivariate model including age and sex as covariates in 189 subjects (378 chromosomes).

\(^a\) Number of minor allele haplotypes.

\(^b\) Minor allele frequency.

\(^c\) Effect size ± standard error in the phenotype (cotinine/(cotinine + nicotine)).

---

Figure 2. Linkage disequilibrium \( (R^2 \text{ and } D’) \) plot of rs3733829 in EGLN2, rs7937 in RAB4B and functional variants in CYP2A6 among TTURC European American Subjects. Numbers represent the \( R^2 \) value expressed as a percentile. Red squares represent pairs with logarithm of odds (LOD) scores for linkage disequilibrium \( ≥2 \text{ and } D’ = 1.0 \), pink squares represent LOD \( ≥2 \text{ and } D’ < 1.0 \), blue squares represent \( D’ = 1 \) but LOD \(< 2 \), white squares represent LOD \(< 2 \) and \( D’ < 1.0 \). Plot generated using HaploView v4.2.
propose that biological differences associated with rs3733829 genotype likely influence smoking behaviors via a novel mechanism unrelated to nicotine metabolism (15), and paved the way to confirm this independent influence using a more suitable phenotype. Thus, our results arise from the convergence of two different approaches to pursuing the biological mechanisms underlying genetic correlates of smoking behavior and associated disease risk: a thorough dissection of the CYP2A6 locus and its influence on nicotine metabolism, and the use of a sensitive and appropriate measure of cigarette consumption and smoke exposure—exhaled CO.

CPD is the most commonly used and easily collected measure of cigarette consumption, but suffers from errors in self-report (47,48) and cannot capture differences in smoking topography, i.e. depth of inhalation, number of puffs per cigarette, etc. As such, correlations between CPD and biomarkers of smoke exposure are modest; CPD explains less than half the variance in saliva levels of the primary metabolite of nicotine, cotinine (49–51), which can vary by greater than an order of magnitude among smokers who report smoking the same number of CPD (50). Breath CO, a byproduct of combustion, has a relatively short half-life (1–4 h (52)) making it sensitive to time of last cigarette (43), but this problem can be reduced by using a check-in process before CO collection to avoid immediate prior smoking, as in this study. CO is significantly correlated with plasma cotinine (53,54) and smoking rate (53,55), and it is more highly correlated with a key biomarker of carcinogen exposure, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (56), than CPD, indicating that CO may capture aspects of smoking behavior relevant to cancer risk that are incompletely reflected through CPD. In this study, CO may also have the advantage of being particularly sensitive to differences in factors associated with hypoxia response, which might be more likely to influence smoking topography than smoking frequency (CPD).

CO is a measure of total smoke exposure; it essentially sums all factors contributing to exposure including number of cigarettes consumed (CPD) and the efficiency with which they are consumed (reflecting differences in smoking topography: number of puffs per cigarettes, depth of inhalation, etc.). By using CO/CPD, we have attempted to examine the efficiency of smoking as it is defined on the basis of yield of CO/cigarette smoked. This permits a rough distinction between factors that drive the frequency of smoking versus those that determine the smoking intensity, which has been related to tobacco dependence and withdrawal magnitude (57–59). It has been argued that variation in the nicotinic receptor gene CHRNA5 influences cancer risk via pathways other than those that mediate changes in smoking behavior (60–62), because the association between variation in CHRNA5 and lung cancer is largely independent of CPD. Our results are not conclusive as to whether greater CYP2A6 activity and the shorter nicotine half-life experienced among smokers who report smoking the same number of CPD, indicate that CO may capture aspects of smoking behavior relevant to cancer risk that are incompletely reflected through CPD. In this study, CO may also have the advantage of being particularly sensitive to differences in factors associated with hypoxia response, which might be more likely to influence smoking topography than smoking frequency (CPD).

Table 4. Multivariate analysis of associations between (CO/CPD) and EGLN2/ CYP2A6 genotype in TTURC

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>MAFb</th>
<th>β ± SEc</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP2A6*1A</td>
<td>521</td>
<td>0.19</td>
<td>0.14 ± 0.27</td>
<td>0.6</td>
</tr>
<tr>
<td>CYP2A6*2</td>
<td>78</td>
<td>0.03</td>
<td>−0.69 ± 0.62</td>
<td>0.3</td>
</tr>
<tr>
<td>CYP2A6*4</td>
<td>52</td>
<td>0.02</td>
<td>−0.82 ± 0.64</td>
<td>0.2</td>
</tr>
<tr>
<td>CYP2A6*9</td>
<td>204</td>
<td>0.08</td>
<td>0.03 ± 0.37</td>
<td>0.9</td>
</tr>
<tr>
<td>CYP2A6*12</td>
<td>69</td>
<td>0.03</td>
<td>−0.72 ± 0.65</td>
<td>0.3</td>
</tr>
<tr>
<td>EGLN2 rs3733829</td>
<td>955</td>
<td>0.35</td>
<td>0.55 ± 0.22</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Associations between exhaled CO/CPD and genetic variables modeled as components of a multivariate model including age, sex and study as covariates in 1353 subjects (2706 chromosomes).

In a previously described experiment, CYP2A6 haplotypes explained >70% of the variance in the metabolism of nicotine to cotinine (15,28). Thus, we possessed substantial power to detect genetic effects on nicotine metabolism. Using CYP2A6 haplotypes as covariates, there was >85% power to detect significant (P < 0.05, n = 189 subjects) effects as small as 2% (0.02 on a scale of 0.0–1.0) and >99% power to detect effects of 3% for hypothetical genetic variants with the same minor allele frequency as rs3733829 (36% in TTURC European American subjects). For comparison, the variable with the smallest effect size in the published genetic model of nicotine metabolism, rs1137115, a synonymous variant that alters CYP2A6 mRNA splicing efficiency (46), has an effect size >0.04 (28). Our inability to detect an independent effect of rs3733829 on nicotine metabolism via this highly sensitive analysis led us to
The clear demonstration of an association between smoking intensity and EGLN2, a hypoxia sensor upstream of a large portion of the human transcriptome (29,30), invites further investigation of this pathway and the mechanism underlying its genetic association. rs3733829 has been reported to be associated with EGLN2 mRNA levels in monocytes and lymphocytes (16,17). Bioinformatic tools do not predict a clear functional consequence of rs3733829 or other tightly linked SNPs; however, multiple SNPs in high LD with rs3733829 occur in the 5′UTR or promoter region and may impact EGLN2 transcription. Further studies will be necessary to determine the specific mechanisms by which polymorphisms in EGLN2 alter gene function. That the effect of this genetic variation is strong enough to be detected through a complex behavioral phenotype also indicates that this variation may be associated with robust differences in general cellular hypoxia response. Such differences may also affect other important processes such as wound healing, tumor growth or ischemia recovery.

MATERIALS AND METHODS

This study complies with the Code of Ethics of the World Medical Association. The University of Wisconsin-Madison and Washington University Human Studies Independent Review Boards approved these studies (approval number for Collaborative Genetic Study of Nicotine Dependence (COGEND) is 00-0203), and all subjects provided written informed consent. Subjects were recruited from three UW-TTURC randomized, placebo-controlled smoking cessation trials, ‘Dependence’ (65), ‘Ed.Sr’ (66) and ‘TTURC 2’ (67). All subjects analyzed were of European descent, at least 18 years of age, and smoked 10 or more CPD. TTURC participants completed baseline assessments of demographics, smoking history (including CPD) and tobacco dependence measured by the Fagerström Test of Nicotine Dependence (68), and provided breath samples for alveolar CO analysis (67). Baseline CO was collected prior to the use of any smoking cessation pharmacotherapy and prior to the quit attempt, and participants had received no instruction to cut down or modify their smoking prior to sample collection. One CO sample was analyzed from TTURC 2 study subjects and the mean of two samples collected on the same day was analyzed for the Dependence and Ed.Sr study subjects (overall mean CO for all subjects = 26.6 ± 12.3, range = 1110 ppm). CO samples were collected using standard assessment methods for smoking cessation clinical trials (69–72) and occurred after a check-in process that entailed a delay between smoking and CO collection to reduce distortion by immediate smoking. CPD was collected and analyzed as a four-level categorical variable (CPD ≤ 10, 10 < CPD ≤ 20, 20 < CPD ≤ 30 and CPD > 30) as in previous studies (69). To calculate CO/CPD, a measure of smoking efficiency (43,44), the four CPD categories were coded as 1, 2, 3 and 4, respectively.

The COGEND is a multisite project in the United States (12). The 189 subjects analyzed here were self-identified as being of European American ancestry, and race was previously verified using EIGENSTRAT (70). Sample demographics were previously described (8,12).

Genome-wide genotyping in TTURC was performed by the Center for Inherited Disease Research at Johns Hopkins University using the Illumina Omni2.5 microarray (www.illumina.com). Data cleaning was led by the GENEVA Coordinating Center at the University of Washington. Additional genotyping and gene copy number determination in the CYP2A6 locus was performed using a Taqman copy number assay (Hs00010002_cn, Applied Biosystems, Foster City, CA, USA) and custom designed Sequenom MassARRAY (Sequenom, San Diego CA, USA) and KASPar (KBioscience, Hoddesdon, Herts, UK) assays, as previously described (28). The predictive model of CYP2A6 activity based on CYP2A6 genotype was utilized to generate a continuous variable metric of nicotine metabolism for each subject, as previously described (15,28).

Association analyses were performed in R (R Foundation for Statistical Computing, Vienna, Austria), with sex, age and study as covariates, in genotyped subjects with available measures of CO or CPD. Power calculations were performed using Quanto 1.1 (71) LD was determined and LD plot generated using Haplo-View v4.2 (72).

ACKNOWLEDGEMENTS

The authors thank and mention the following: The Wisconsin State Laboratory of Hygiene provided considerable technical assistance in the form of DNA extraction. Research was also aided by the Wisconsin Partnership Program. Medication was provided to patients at no cost under a research agreement with GlaxoSmithKline. Investigators directing data collection for COGEND are Laura Bierut, Naomi Breslau, Dorothy Hatsukami and Eric Johnson; data management are organized by Nancy Saccone and John Rice; laboratory analyses are led by Alison Goate; data collection were supervised by Tracey Richmond.

Conflict of Interest statement. Drs. Goate and Bierut are listed as inventors on a patent (US 20070258898) covering the use of certain SNPs in determining the diagnosis, prognosis and treatment of addiction.

FUNDING

The National Institute of Mental Health (ST32MH1014677-33 to A.J.B.); the National Cancer Institute (P01 CA-089392 to L.J.B., and P50 CA-84724 and K05 CA-139871 to T.B.B.); the National Institute on Drug Abuse (K02 DA-021237 to L.J.B.) and the National Human Genome Research Institute (U01 HG-004422 to L.J.B.). Funding support for genotyping, which was performed at the Center for Inherited Disease Research (CIDR), was provided by X01 HG005274-01. CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C. Assistance with genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies (GENEVA) Coordinating Center (U01 HG004446). Funding support for collection of datasets and samples was provided by the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392) and the University of Wisconsin Transdisciplinary Tobacco Use Research Center (P50 DA019706, P50 CA084724). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.
REFERENCES


