Nicotine & Tobacco Research, 2017, 426–434
doi:10.1093/ntr/ntw260
Original investigation
Advance Access publication October 17, 2016

Original investigation

A Novel Tobacco Use Phenotype Suggests the 15q25 and 19q13 Loci May be Differentially Associated With Cigarettes per Day and Tobacco-Related Problems

Leah S. Richmond-Rakerd MA[1,2], Jacqueline M. Otto MA[1,2], Wendy S. Slutske PhD[1,2], Cindy L. Ehlers PhD[3], Kirk C. Wilhelmsen PhD[4], Ian R. Gizer PhD[1,2]

[1]Department of Psychological Sciences, University of Missouri, Columbia, MO; [2]Alcoholism Research Center at Washington University School of Medicine, St. Louis, MO; [3]Department of Molecular and Cellular Neurosciences (CLE), The Scripps Research Institute, La Jolla, CA; [4]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC

Corresponding Author: Ian R. Gizer, PhD, Department of Psychological Sciences, University of Missouri, 210 McAlester Hall, Columbia, MO 65211, USA. Telephone: 573-882-5427; Fax: 573-882-7710; E-mail: gizeri@missouri.edu

Abstract

Introduction: Tobacco use is associated with variation at the 15q25 gene cluster and the cytochrome P450 (CYP) genes CYP2A6 and CYP2B6. Despite the variety of outcomes associated with these genes, few studies have adopted a data-driven approach to defining tobacco use phenotypes for genetic association analyses. We used factor analysis to generate a tobacco use measure, explored its incremental validity over a simple indicator of tobacco involvement: cigarettes per day (CPD), and tested both phenotypes in a genetic association study.

Methods: Data were from the University of California, San Francisco Family Alcoholism Study (n = 1942) and a Native American sample (n = 255). Factor analyses employed a broad array of tobacco use variables to establish the candidate phenotype. Subsequently, we conducted tests for association with variants in the nicotinic acetylcholine receptor and CYP genes. We explored associations with CPD and our measure. We then examined whether the variants most strongly associated with our measure remained associated after controlling for CPD.

Results: Analyses identified one factor that captured tobacco-related problems. Variants at 15q25 were significantly associated with CPD after multiple testing correction (rs938682: p = .00002, rs1051730: p = .0003, rs16969968: p = .0003). No significant associations were obtained with the tobacco use phenotype; however, suggestive associations were observed for variants in CYP2B6 near CYP2A6 (rs45482602: ps = .0082, .0075) and CYP4Z2P (rs10749865: ps = .0098, .0079).

Conclusions: CPD captures variation at 15q25. Although strong conclusions cannot be drawn, these finding suggest measuring additional dimensions of problems may detect genetic variation not accounted for by smoking quantity. Replication in independent samples will help further refine phenotype definition efforts.

Implications: Different facets of tobacco-related problems may index unique genetic risk. CPD, a simple measure of tobacco consumption, is associated with variants at the 15q25 gene cluster.

Additional dimensions of tobacco problems may help to capture variation at 19q13. Results demonstrate the utility of adopting a data-driven approach to defining phenotypes for genetic association studies of tobacco involvement and provide results that can inform replication efforts.

Introduction

In the United States, tobacco consumption results in more than 480 000 premature deaths and productivity losses of $289 billion annually.[1] Despite its health and financial costs, many individuals continue to use tobacco: in 2014, estimated smoking prevalence in the United States was 16.8%.[2] Genetic factors influence tobacco involvement, and heritability estimates for smoking-related outcomes are in excess of 50%.[3–6] Identifying genes that contribute to tobacco use can help clarify its biological underpinnings and identify individuals at risk for problems.

Variants in genes that encode the nicotinic acetylcholine receptor (CHRN) subunits are robustly associated with smoking and have been implicated in a number of outcomes, including age of initiation,[7,8] subjective response,[9,10] nicotine dependence (ND),[11,12] lung cancer,[13,14] cotinine levels,[15,16] and exhaled carbon monoxide.[17]

Cytochrome P450 (CYP) genes CYP2A6 and CYP2B6 also demonstrate replicable associations. CYP2A6 and CYP2B6 are associated with smoking cessation success,[18–21] and CYP2A6 is linked with cigarette consumption and dependence.[22–24] Many tobacco-related measures are related to the CHRN and CYP genes; however, cigarettes per day (CPD) is commonly used and is a strong candidate for genetic association studies. Genome-wide meta-analyses have found CPD to be associated with the CHRNA5/A3/B4 gene cluster[25–28] and variants at CHRNB3, CHRNA6, CYP2A6, and CYP2B6.[25]

Despite the many tobacco-related outcomes associated with the CHRN and CYP genes, few studies have adopted a data-driven approach to defining phenotypes for association analyses. This may increase power to detect variants for tobacco involvement. It may also change the associations obtained. For instance, CPD is often used as a proxy for ND; however, Rice et al.[29] found that in a multiethnic sample, dependence as assessed by the Fagerström Test for Nicotine Dependence (FTND)—but not CPD—was associated with a genetic locus in the region of CHRNB3. Hancock et al.,[30] in the largest GWAS meta-analysis of ND to date, observed associations between FTND scores and variants in CHRNA4 that have not been associated with CPD. Thus, different phenotypes may show stronger associations with different variants. Encompassing more dimensions of use should help identify variants that overlap with those obtained in analyses of CPD and variants specific to other facets of use. This can improve our understanding of the mechanisms underlying genotype-phenotype associations.

We aimed to characterize a broad spectrum of tobacco involvement. We therefore surveyed a number of behaviors, including age of onset, duration of use, and disorder. Factor analysis is ideal for defining the latent structure of interrelated items. Behavior genetic studies have demonstrated its utility for characterizing substance use phenotypes for genetic analyses. For instance, Lessov et al.[31] used phenotypic and genetic factor analyses of the *DSM-IV* ND criteria and the Heaviness of Smoking Index to identify a highly heritable dependence phenotype. To our knowledge, however, no such approach has been used to define tobacco use phenotypes for molecular genetic studies.

This study adopted a data-driven approach to define a candidate tobacco use phenotype for genetic association studies. We conducted phenotypic analyses in two independent samples to establish the measure. Subsequently, we evaluated the utility of this measure by conducting single variant tests for genetic association, focusing on a set of variants within the CHRN and CYP genes. We first conducted tests with CPD. Next, we examined whether the genetic variants most strongly associated with our phenotype remained associated after controlling for CPD. This was done to determine the incremental validity of our measure over a simple and commonly used measure of tobacco involvement.

Methods

Participants

Data were collected at the University of California, San Francisco (UCSF) and The Scripps Research Institute (La Jolla, CA). Assessment procedures were approved by each organization's institutional review boards. Participants were fully briefed on the study and provided informed consent. Management and analysis of data collected at UCSF were approved by the institutional review board at the University of North Carolina at Chapel Hill. Data collection at The Scripps Research Institute was also approved by the Indian Health Council.

UCSF Family Alcoholism Study Sample

Participants are a subset of the UCSF Family Alcoholism Study who reported European ancestry and exposure to tobacco as defined by smoking more than 100 cigarettes, 30 cigars or pipes, or 30 pouchfuls of snuff or chewing tobacco in their lifetime. This threshold was derived by the authors of the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA[32]) and is consistent with the criterion employed by the Centers for Disease Control to classify smokers.[2] The lifetime smoking threshold was considered the minimal level of exposure to cigarettes; thus, individuals who did not meet this threshold were not administered questions concerning tobacco-related problems.

The UCSF Family Alcoholism Study[33] consists of 2524 participants from 890 families (average size = 2.8 members). It was a nationwide study on the genetics of alcoholism and other substance dependence designed to recruit small family pedigrees enriched for alcohol dependence. Probands were invited to participate if they met criteria for alcohol dependence in their lifetime and had at least one sibling or both parents available to participate. Probands were excluded if they reported serious drug addictions; history of intravenous substance use; current or past diagnosis of schizophrenia, bipolar disorder, or other psychiatric illness involving psychotic symptoms (those with depressive and anxiety disorders were not excluded); life-threatening illness; or an inability to speak English. Permission was then obtained from the proband to invite relatives to participate. Of the 2524 participants, 1841 (39.6% male, mean age = 48.7 years [*SD* = 13.3]) reported tobacco exposure and were included in this report.

One hundred one individuals did not provide data concerning tobacco exposure. In most cases, these individuals did not exhibit a sufficient level of smoking to continue that interview section. For a small number of participants, however, their reason for failing to respond and/or their responses to prior questions were deemed clinically relevant, and they were administered subsequent items. Thus, they provided a modest amount of data. For instance, of these respondents, 3.0%–27.7% endorsed tobacco-related problems. They were included in analyses, for a total sample of 1942 individuals.

Native American Sample

Participants were recruited from eight contiguous reservations with a total population of approximately 3000. Participants were recruited using a venue-based method for sampling hard-to-reach populations[34,35] and a respondent-driven procedure.[36,37] To be included in the study, an individual had to be at least 1/16th Native American Heritage, 18–70 years old, and mobile enough to be transported to The Scripps Research Institute. Participants were included in this study if they reported prior exposure to tobacco as defined by having smoked at least 100 cigarettes in their lifetime. Of the 418 participants administered the SSAGA tobacco assessment, 253 (46.3% male, mean age = 34.7 years [*SD* = 14.6]) reported exposure. Two individuals provided no data regarding exposure. To remain consistent with analyses conducted in the UCSF sample, these participants were included, for a total sample of 255 individuals.

Differences exist between the UCSF and Native American samples (eg, cultural norms concerning substance use, sample recruitment/inclusion procedures). Thus, we anticipated variability across cohorts with regard to nature and severity of tobacco involvement. However, a primary goal of this analysis was to develop a factor solution that might replicate across diverse samples. Replication analyses within the Native American cohort were considered important in helping to produce a generalizable phenotype. However, only the UCSF sample was employed for the association analysis, as the small number of individuals (n = 191) with genotype and phenotype data in the Native American sample provided limited power for a replication analysis.

Measures

Semi-Structured Assessment for the Genetics of Alcoholism

Both samples were administered a modified version of the SSAGA. The present study used demographic variables and data pertaining to tobacco use. Criteria from the *DSM-IV*, ICD-10, and FTND definitions of ND were included. These included a detailed assessment of withdrawal. Additional indicators included the ages of onset and offset of milestones (eg, daily smoking and dependence) and duration of use (eg, length of time smoked daily, duration of abstinence). CPD was operationalized as a continuous measure.

To survey a broad scope of tobacco-related behaviors, we adopted an over-inclusive method of variable selection. As many items as possible were identified, and prior to analyses, redundant variables were consolidated or removed. 42 items were included (Supplementary Tables S1a–S2).

Genotyping: UCSF Family Alcoholism Study Sample

The Affymetrix Axiom Exome Genotyping Array (Affymetrix Inc.) was used for genotyping. We focused on a final set of 231 single nucleotide variants (SNVs) within the CHRN and CYP genes. We included variants with prior evidence of association with tobacco involvement (eg, CHRNA6, CYP2A6, CYP2B6, CHRNB1, CHRNA4, and the CHRNA5/A3/B4 gene cluster[12,20,22,25,26,30,38–42]) and SNPs not previously implicated (Supplementary Table S3). This was for several reasons. First, we predicted that a comprehensive approach would help identify unique genetic signals. For instance, Saccone et al.,[12] in an analysis of all nAChr subunits, identified an association between ND and the CHRND-CHRNG gene cluster (a finding that has been replicated with cotinine levels[43]). Second, we included a variety of smoking behaviors, which might help identify a range of genetic variation. Third, we employed an exome genotyping array that captures rare variants. Lastly, many CYP variants are clustered within regions, sometimes making it difficult to identify the causal SNP. Thus, we included variants that exhibit broader effects (eg, CYP1A1, which includes polymorphisms associated with caffeine consumption[44,45] and lung cancer[46]) and other novel SNVs. Since not all SNPs had evidence for prior association (and we could not conduct a replication), we adhered firmly to the multiple testing correction as the minimum p value necessary for statistical significance.

Genotyping quality control was conducted using PLINK[47] and degree of relatedness estimations were conducted using PREST.[48] 36 individuals were removed due to unresolved pedigree errors, six due to unresolved discrepant sex codes, and five due to low genotype call rates (<95%). Of the pool of originally selected 2085 SNVs, 1675 did not vary and were excluded. An additional 58 were excluded due to low genotype call rates (<95%), and three were excluded due to deviations from Hardy-Weinberg equilibrium (p < 1e-05). Cross-referencing allele frequencies with the European samples for the 1000 Genomes Project[49] resulted in the exclusion of 113 SNVs whose allele frequencies differed more than 0.20 from this reference panel, leaving a final set of 231 SNVs. Of the 1841 individuals who reported tobacco exposure, 1308 had valid genotype data.

Statistical Analysis

Phenotypic Analyses

Phenotypic analyses were conducted in Mplus version 7 using the method of maximum likelihood with robust standard errors.[50,51] A clustering variable (the family number for each participant) was included. To evaluate the phenotypic structure of the items, a series of factor analyses were conducted. First, the UCSF sample was randomly split into two halves; one dataset was used for the exploratory factor analysis (EFA) and the other was used for the confirmatory factor analysis (CFA). Subsequently, we cross-validated the factor solution obtained in the UCSF sample in the Native American sample. Employing the items from the initial factor analysis, we conducted an EFA and CFA in the Native American sample using the same split-half procedure. Two sets of analyses were run: one including CPD and one excluding CPD. A geomin rotation solution was employed. Model fit was evaluated using the Akaike Information Criterion (AIC[52]), the Bayesian Information Criterion (BIC[53]), and Log-Likelihood (LL) values.

Problems can arise when conducting factor analyses with dichotomous items, including generation of spurious factors due to nonlinear associations between items and latent variables, distortion of the correlation matrix due to differing response proportions,[54] and reduced power of fit indices.[55] To address these issues we used Mplus, to allow for nonlinear relations between items and latent variables and use of estimators robust to deviations from normality.[56,57] Because statistical

power can also be reduced when analyzing categorical variables, we employed a large sample for the initial EFA and CFA.

Factor scores for association analyses were derived using multiple-group CFA, followed by multiple indicators multiple causes (MIMIC) models to test for differential item functioning (DIF) across samples.[58] A p value of .001 was adopted for tests of DIF to control for experiment-wise error given the large number of tests, and because chi-square tests evaluating model fit become more biased toward complex models as sample sizes increase. To control for sample-specific effects and increase the phenotype's generalizability, factor scores accounted for DIF. Scores were derived from a model in which the loadings and thresholds of items exhibiting DIF were freed across samples.

Single Variant Association Tests

Ancestry estimations were calculated from variants with a minor allele frequency ≥ 0.01 using principal components analysis[59] in the GCTA software.[60] The resulting estimates correlated highly with self-reported ancestry (first eigenvector and European ancestry in full sample: r = 0.718; second eigenvector and African ancestry (excluding European ancestry individuals): r = 0.792). Thus, these estimates were used as covariates to control for population substructure. Single variant association tests were conducted for CHRN and CYP variants with minor allele frequency ≥ 0.01 using EPACTS.[61] Models included sex, age, age-squared, and the first three eigenvectors generated from the PCA as covariates.

The first analysis examined the main effect of CPD. The second two sets of analyses examined the main effect of each factor score before and after controlling for CPD. EPACTS excludes individuals who are missing on covariates; therefore, the unadjusted models were run using only individuals with CPD data. Of the 1297 lifetime smokers with CPD data, 1253 had factor score data and were included in the unadjusted models.

Results

Phenotypic Analyses

Exploratory Factor Analysis

The first EFA–CFA included CPD. The EFA in the UCSF sample provided modest support for a two-factor solution. It yielded a better fit to the data (LL = −30623.93, AIC = 61517.85, BIC = 61746.69) than the single-factor solution (LL = −32041.47, AIC = 64270.94, BIC = 64728.83), and although the second factor contained only four items, three loaded highly and the factors were weakly correlated (r = 0.11). Further, the item content of each factor was distinct; the first consisted predominantly of items concerning tobacco-related problems, while the second consisted of items regarding length of tobacco use and ages of milestones. We therefore specified a two-factor solution for the CFA. Prior to the CFA, eight items with loadings below 0.30 and one item with a cross-loading were removed, and the two-factor structure was reconfirmed. Two items with low loadings were removed during the CFA. The final solution contained 31 items (27 items on the "tobacco use problems" factor and four items on the "age" factor; see Supplementary Tables S4a–S5b, S8a, and S8b). Of the "problems" items, 26 exhibited moderate to high loadings (range = 0.46–0.91). Of the "age" items, three exhibited high loadings (range = 0.82–0.98). CPD loaded onto the "problems" factor (loading = 0.37).

The EFA in the Native American sample provided limited support for a two-factor solution. It yielded a better fit to the data (LL = −3811.46, AIC = 7892.91, BIC = 8276.88) than the one-factor solution (LL = −3954.28, AIC = 8096.57, BIC = 8363.92); however, the second factor contained only three items. Our primary aim was to identify the "problems" factor obtained in the UCSF analysis, and the items exhibiting significant loadings in the single-factor solution largely replicated this factor. Thus, we specified a one-factor solution for the CFA. Seven and five items with loadings below 0.30 were removed during the EFA and CFA, respectively. Thirty items were retained in the final solution, of which 28 exhibited moderate to high loadings (range = 0.46–0.88). CPD was retained (loading = 0.34; see Supplementary Tables S6a–S7b, S9a, and S9b).

Multiple-Group Analysis

Items retained on the "problems" factors in the UCSF and Native American samples were compared. Five items retained in only one sample were excluded, resulting in a set of 26 items. When the EFA–CFA analyses were re-run excluding CPD, results were consistent, with the exception that the item "smoking caused nervousness, jitteriness, or emotional problems" was retained on the first factor in the UCSF sample (loading = 0.32). It was excluded from the multiple-group analysis as it was not retained in the Native American sample. Standardized loadings for the factor solutions in both samples are presented in Tables 1 and 2.

An initial one-factor model was fit to the data, allowing the item loadings and thresholds to be freely estimated for 25 of the 26 items (including CPD) and 24 of the 25 items (excluding CPD). The loading and threshold for the remaining item were constrained across groups for model identification. Constraining the loadings and thresholds resulted in a significant decrement in fit (including CPD: χ^2 = 274.92, df = 50, p < .0001; excluding CPD: χ^2 = 229.47, df = 48, p < .0001). Thus, we tested for DIF as a function of sample.

Differential Item Functioning

Seven and six items showed evidence of DIF when including and excluding CPD, respectively (Supplementary Tables S10 and S11). The items were consistent across analyses. Factor scores were derived from multiple-group models in which the loading and threshold for each item exhibiting DIF were freed across samples. The factor scores were almost perfectly correlated (r = .999).

Sensitivity Analyses

In the UCSF analysis, the variables that loaded onto the second factor were continuous. We explored whether operationalizing the variables categorically might change the pattern of loadings. We dichotomized the variables at their median value and re-ran the EFA. Very similar results were observed, with the exception that the variable pertaining to the age of ND onset no longer loaded significantly onto the second factor (loading = 0.18).

Relation With ND

We examined associations between the factor scores and individuals' scores on the FTND (UCSF sample: mean = 4.1 (SD = 2.6), range = 0–10; Native American sample: mean = 3.0 (SD = 2.7), range = 0–10). The correlations between FTND scores and the factor scores including and excluding CPD were 0.64 (p < .0001) and 0.62 (p < .0001), respectively. Thus, although there was a high degree of overlap, our scores included information not captured by diagnostic criteria.

Table 1. Standardized Loadings for the One-Factor Solution Including CPD

Item	UCSF	NA
Smoke in forbidden places	0.58 [0.53, 0.63]	0.67 [0.57, 0.78]
Smoke when ill	0.63 [0.59, 0.68]	0.61 [0.48, 0.74]
Chain smoke	0.46 [0.41, 0.52]	0.56 [0.43, 0.68]
Reduced activity engagement	0.62 [0.56, 0.67]	0.53 [0.33, 0.72]
Smoke in larger amounts or over longer periods than intended	0.59 [0.54, 0.64]	0.59 [0.47, 0.71]
Run out of cigarettes sooner than expected	0.54 [0.49, 0.59]	0.63 [0.50, 0.75]
Smoke in dangerous places	0.50 [0.45, 0.55]	0.52 [0.37, 0.67]
Desire to quit	0.55 [0.50, 0.60]	0.59 [0.45, 0.73]
Ability to quit	0.67 [0.63, 0.71]	0.58 [0.42, 0.74]
Inability to quit	0.67 [0.63, 0.71]	0.68 [0.52, 0.84]
Withdrawal: Irritability	0.88 [0.85, 0.91]	0.79 [0.70, 0.88]
Withdrawal: Nervousness or anxiety	0.88 [0.85, 0.90]	0.88 [0.81, 0.95]
Withdrawal: Restlessness	0.87 [0.84, 0.90]	0.87 [0.79, 0.94]
Withdrawal: Concentration problems	0.83 [0.80, 0.86]	0.85 [0.76, 0.94]
Withdrawal: Depression	0.75 [0.72, 0.79]	0.74 [0.60, 0.88]
Withdrawal: Appetite increase or weight gain	0.52 [0.48, 0.57]	0.55 [0.42, 0.68]
Withdrawal: Trouble sleeping	0.78 [0.74, 0.82]	0.85 [0.75, 0.95]
Withdrawal: Craving	0.76 [0.73, 0.80]	0.80 [0.71, 0.89]
Four withdrawal symptoms within 24 h of quitting/cutting down	0.91 [0.88, 0.93]	0.89 [0.78, 1.00]
Role interference from withdrawal symptoms	0.78 [0.74, 0.82]	0.57 [0.29, 0.85]
Continue use to avoid withdrawal symptoms	0.75 [0.71, 0.79]	0.72 [0.58, 0.87]
Continue despite health problems from smoking	0.47 [0.42, 0.52]	0.46 [0.29, 0.62]
Continue despite smoking-exacerbated illness	0.58 [0.53, 0.62]	0.58 [0.43, 0.73]
Tolerance	0.55 [0.50, 0.61]	0.67 [0.55, 0.79]
Time to first cigarette	0.53 [0.48, 0.58]	0.65 [0.53, 0.77]
CPD	0.37 [0.33, 0.42]	0.32 [0.21, 0.43]

CPD = cigarettes per day; NA = Native American sample; UCSF = University of California, San Francisco Family Alcoholism Study sample. Standardized loadings derived from the multiple group confirmatory factor analysis. The positive loadings observed for time to first cigarette and ability to quit are due to reverse-scoring the variables. 95% confidence limits presented in brackets.

Table 2. Standardized Loadings for the One-Factor Solution Excluding CPD

Item	UCSF	NA
Smoke in forbidden places	0.58 [0.53, 0.63]	0.59 [0.45, 0.72]
Smoke when ill	0.62 [0.58, 0.67]	0.60 [0.47, 0.73]
Chain smoke	0.45 [0.39, 0.50]	0.55 [0.42, 0.68]
Reduced activity engagement	0.61 [0.55, 0.67]	0.51 [0.32, 0.71]
Smoke in larger amounts or over longer periods than intended	0.59 [0.54, 0.64]	0.59 [0.47, 0.71]
Run out of cigarettes sooner than expected	0.54 [0.49, 0.59]	0.62 [0.49, 0.75]
Smoke in dangerous places	0.49 [0.44, 0.54]	0.51 [0.36, 0.66]
Desire to quit	0.55 [0.50, 0.60]	0.59 [0.45, 0.73]
Ability to quit	0.55 [0.50, 0.60]	0.58 [0.42, 0.74]
Inability to quit	0.67 [0.62, 0.71]	0.67 [0.51, 0.84]
Withdrawal: Irritability	0.89 [0.86, 0.91]	0.80 [0.71, 0.89]
Withdrawal: Nervousness or anxiety	0.88 [0.86, 0.91]	0.88 [0.81, 0.95]
Withdrawal: Restlessness	0.87 [0.85, 0.90]	0.87 [0.79, 0.94]
Withdrawal: Concentration problems	0.83 [0.80, 0.86]	0.85 [0.77, 0.94]
Withdrawal: Depression	0.76 [0.72, 0.80]	0.75 [0.61, 0.88]
Withdrawal: Appetite increase or weight gain	0.52 [0.48, 0.57]	0.55 [0.42, 0.68]
Withdrawal: Trouble sleeping	0.79 [0.75, 0.82]	0.85 [0.75, 0.94]
Withdrawal: Craving	0.76 [0.73, 0.80]	0.80 [0.71, 0.89]
Four withdrawal symptoms within 24 h of quitting/cutting down	0.91 [0.89, 0.94]	0.89 [0.78, 1.00]
Role interference from withdrawal symptoms	0.78 [0.74, 0.81]	0.57 [0.29, 0.84]
Continue use to avoid withdrawal symptoms	0.75 [0.71, 0.79]	0.72 [0.58, 0.87]
Continue despite health problems from smoking	0.46 [0.41, 0.51]	0.45 [0.28, 0.61]
Continue despite smoking-exacerbated illness	0.57 [0.52, 0.62]	0.57 [0.42, 0.72]
Tolerance	0.55 [0.50, 0.60]	0.66 [0.54, 0.78]
Time to first cigarette	0.52 [0.47, 0.57]	0.64 [0.51, 0.76]

CPD = cigarettes per day; NA = Native American sample; UCSF = University of California, San Francisco Family Alcoholism Study sample. Standardized loadings derived from the multiple group confirmatory factor analysis. The positive loadings observed for time to first cigarette and ability to quit are due to reverse-scoring the variables. 95% confidence limits presented in brackets.

Single Variant Association Tests

Using the Genetic type I error calculator,[62] we computed the significance threshold necessary to control the type I error rate at .05 across the 231 SNVs, while accounting for correlations among variants. The critical *p* value was .00034. Although this was the required threshold for statistical significance for all variants, we examined results for SNPs that did not reach this threshold, but were within genes for which there was prior evidence for association with smoking. For those variants, we adopted a liberal *p* value cutoff of .10. Table 3 displays the variants that met this threshold.

The strongest associations for CPD were obtained for an intronic variant (rs938682; $p = .00002$) and a synonymous variant (rs1051730; $p = .0003$) in CHRNA3 and a missense mutation in CHRNA5 (rs16969968; $p = .0003$). The latter two variants (rs1051730 and rs16969968) were in near perfect linkage disequilibrium. These variants were not associated with our factor scores. Thus, in this sample, SNPs within the 15q25 gene cluster related more strongly to smoking quantity than tobacco-related problems. No other variants were significantly associated with CPD; however, the significance levels of variants within several previously implicated genes fell below .10. These included the missense SNPs rs148166815 ($p = .0586$) and rs28399435 ($p = .0588$) in CYP2A6 and a synonymous variant in CHRNA4 (rs2273506; $p = .0578$).

No associations with our factor scores reached significance after multiple testing correction. However, suggestive associations were observed for a missense mutation in CYP2B6 (rs45482602; *p*s = .0075, .0082), a gene previously implicated in smoking-related outcomes,[19,21,25] and these persisted after controlling for CPD (*p*s = .0098, .0100). Suggestive associations were also obtained with a novel intronic variant at CYP4Z2P (rs10749865; *p*s = .0079, .0098). This signal appeared specific to tobacco-related problems, as the SNP was not related to CPD and effects were largely unchanged after adjusting for CPD (*p*s = .0019).

Discussion

We employed a data-driven approach to define a novel candidate phenotype for genetic association studies of tobacco involvement. Using data from two samples, phenotypic and genetic analyses were conducted to establish the generalizability of the phenotype and explore its incremental validity over a commonly used and simple measure of tobacco involvement: CPD. Analyses identified one factor that captured tobacco-related problems. Replicating prior research, CPD was associated with variants in CHRNA3 and CHRNA5. No significant associations were obtained for our factor scores; however, suggestive association was observed with variants in two CYP genes that were unrelated to CPD.

Variants within the 15q25 gene cluster were most strongly related to CPD and displayed specificity to this phenotype. These included SNPs in CHRNA3 (rs938682 and rs1051730) and CHRNA5 (rs16969968). rs938682 is in strong linkage disequilibrium with rs1051730, which has been previously associated with CPD.[25,26,28] Further, rs1051730 is in near perfect linkage disequilibrium with rs16969968, which is a top hit in analyses of CPD.[28,63–65] rs16969968 has also been associated with tobacco-related biomarkers such as cotinine levels[16] and exhaled carbon monoxide,[17] as well as with nicotine-related functional effects (see Wen et al.[66] for a review). Present findings support the utility of CPD in studies of genetic variation at 15q25, and suggest that additional domains of problems provide no incremental information concerning variation at these regions.

No variants reached significance when associations were tested with the factor scores. However, suggestive association was observed for a missense variant in CYP2B6 (rs45482602; minor allele frequency = 0.01 in the present sample). CYP2B6 metabolizes bupropion, which is used as a smoking cessation aid.[20] It has been implicated in a GWAS of nicotine metabolite ratio.[67] Certain types of problems may be more strongly related to nicotine metabolism than others. For instance, some studies have found relations between nicotine metabolite ratio and physiological symptoms such as withdrawal[68,69] and craving.[70] Therefore, current findings suggest that measures of tobacco-related problems rather than smoking quantity may display specificity to variants in CYP2B6. Given the lack of significant association, however, this remains unclear. Further, it should be noted that the suggestive association with rs45482602 may have resulted from linkage disequilibrium with other variants in the 19q13 locus (particularly those within CYP2A6, which is located near CYP2B6 and expression of which has been shown to be influenced by variants near rs45482602[71]).

An intronic variant at CYP4Z2P (rs10749865) showed suggestive evidence for association with our phenotypes. No prior studies of tobacco use have found relations with this gene. Nonetheless, two lines of evidence suggest how rs10749865 might influence risk for tobacco involvement. First, rs10749865 has been identified as an eQTL for CYP4B1,[71] which is downstream of CYP4Z2P. CYP4B1 is expressed in the surface epithelium[72] and submucosal gland ducts[73] of the lungs and has been related to COPD.[74,75] Second, variants in CYP4Z2P represent QTLs for fatty acid metabolites (tetradecanedioate and hexadecanedioate) measured in blood.[76] These accumulate in the lung tissue of individuals suffering from pulmonary arterial hypertension,[77] which can result from cigarette smoke exposure.[78,79] These lines of research suggest a possible relation between rs10749865 and tobacco use; however, given the lack of significant association with this variant and limited prior findings, strong conclusions cannot be drawn. Replication will be necessary.

Limitations

Several limitations should be considered. The first concerns generalizability, as most UCSF sample participants were Caucasian. Application of present findings to other racial/ethnic groups may thus be limited. However, the generalizability of our measure was increased by replicating phenotypic analyses in a Native American cohort. Second, the UCSF sample was recruited based on alcohol dependence status, and high levels of alcohol involvement exist within the Native American cohort. Therefore, the factor structure obtained may partly reflect risk for comorbid alcohol and tobacco use. This may also limit power to detect associations with tobacco involvement within the UCSF sample. Relatedly, although suggestive associations were observed with our factor scores, no effects were statistically significant, which may be due to insufficient statistical power. Third, we were unable to replicate the association analyses due to the limited number of individuals with genotype data in the Native American sample. Replication is therefore warranted. It should be noted, however, that the effect sizes obtained for our factor scores were consistent with those observed in association studies of complex traits. Lastly, the genotyping array used was designed to capture rare exonic variation and may not capture all non-exonic regulatory variants.

Table 3. Results of Single Variant Association Tests

SNP	Alleles		Chr	Position	Gene	CPD			Factor score with CPD						Factor score without CPD					
									Unadjusted			Adjusted			Unadjusted			Adjusted		
	Effect	Other				β (*SE*)	R^2	*p*	β (*SE*)	R^2	*p*	β (*SE*)	R^2	*p*	β (*SE*)	R^2	*p*	β (*SE*)	R^2	*p*
rs938682	A	G	15	78896547	CHRNA3	2.86 (0.67)	0.014	.00002	0.10 (0.04)	0.004	.0259	0.02 (0.04)	0.000	.6539	0.09 (0.04)	0.003	.0382	0.02 (0.04)	0.000	.6481
rs16969968	G	A	15	78882925	CHRNA5	−2.27 (0.62)	0.010	.0003	−0.03 (0.04)	0.001	.4144	0.03 (0.04)	0.000	.4614	−0.03 (0.04)	0.000	.5222	0.03 (0.04)	0.001	.4430
rs1051730	G	A	15	78894339	CHRNA3	−2.25 (0.62)	0.010	.0003	−0.03 (0.04)	0.001	.4433	0.03 (0.04)	0.001	.4306	−0.02 (0.04)	0.000	.5544	0.03 (0.04)	0.001	.4136
rs2472553	G	A	8	27328511	CHRNA2	−1.95 (0.83)	0.004	.0194	−0.05 (0.05)	0.001	.3665	−0.003 (0.05)	0.000	.9452	−0.04 (0.05)	0.001	.4155	−0.003 (0.05)	0.000	.9533
rs1825089	C	A	15	78997562	CHRNB4	−3.41 (1.78)	0.003	.0550	−0.13 (0.11)	0.001	.2696	−0.02 (0.10)	0.000	.8720	−0.11 (0.11)	0.001	.3240	−0.01 (0.11)	0.000	.9020
rs2273506	G	A	20	61990939	CHRNA4	−2.00 (1.06)	0.003	.0578	−0.12 (0.07)	0.003	.0662	−0.06 (0.06)	0.001	.2873	−0.12 (0.07)	0.002	.0830	−0.06 (0.06)	0.001	.3095
rs148166815	A	G	19	41351309	CYP2A6	−11.89 (6.28)	0.003	.0586	−0.02 (0.40)	0.000	.9594	0.30 (0.37)	0.001	.4096	0.01 (0.40)	0.000	.9781	0.30 (0.37)	0.001	.4151
rs28399435	C	T	19	41356246	CYP2A6	−2.95 (1.56)	0.003	.0588	−0.18 (0.10)	0.003	.0710	−0.09 (0.09)	0.001	.3154	−0.18 (0.10)	0.003	.0789	−0.09 (0.09)	0.001	.3089
rs45482602	A	C	19	41515255	CYP2B6	1.33 (2.45)	0.000	.5860	0.42 (0.16)	0.006	.0082	0.37 (0.14)	0.005	.0100	0.42 (0.16)	0.006	.0075	0.38 (0.15)	0.005	.0098
rs10749865	A	G	1	47309265	CYP4Z2P	0.16 (0.66)	0.000	.8031	−0.11 (0.04)	0.005	.0098	−0.12 (0.04)	0.008	.0019	−0.11 (0.04)	0.006	.0079	−0.12 (0.04)	0.008	.0019
rs12908877	G	A	15	32323454	CHRNA7	−0.84 (0.64)	0.001	.1890	−0.08 (0.04)	0.003	.0534	−0.06 (0.04)	0.002	.1336	−0.08 (0.04)	0.003	.0567	−0.06 (0.04)	0.002	.1352
rs79220301	A	G	17	7352012	CHRNB1	0.03 (2.43)	0.000	.9899	0.28 (0.16)	0.003	.0769	0.27 (0.14)	0.003	.0525	0.28 (0.16)	0.003	.0730	0.28 (0.14)	0.003	.0537

Adjusted = adjusted for CPD, Chr = chromosome; CPD = cigarettes per day; *SE* = standard error; Unadjusted = factor score alone. The critical *p* value for statistical significance was .00034; however, also included are results for variants within genes with prior evidence of association with tobacco use outcomes. The *p* value cutoff for inclusion of these variants was .10. rs1051730 is in near perfect linkage disequilibrium with rs16969968.

Conclusions

Notwithstanding limitations, the present study provides important results that can inform phenotype selection in association analyses of tobacco use. Findings replicate research suggesting that CPD detects variation at 15q25. Although strong conclusions cannot be drawn regarding the incremental validity of our phenotype, findings suggest that measuring additional dimensions of problems may capture variation in CYP genes not accounted for by CPD. Future studies employing larger samples should aim to replicate the present findings. Continued research exploring the relative utility of CPD and additional dimensions of tobacco use will help to further refine phenotype definition efforts.

Supplementary Material

Supplementary Tables S1–S11 can be found online at http://www.ntr.oxfordjournals.org

Funding

This work was supported by grants from the National Institutes of Health from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) to LSR-R (AA023419), the National Institute on Drug Abuse to IRG, CLE, and KCW (DA030976), and the NIAAA and the National Center on Minority Health and Health Disparities to CLE (AA010201). Additional funding was provided by the State of California and the Ernest Gallo Clinic and Research Center for Medical Research on Alcohol and Substance Abuse through the University of California at San Francisco to KCW.

Declaration of Interests

None declared.

References

1. U.S. Department of Health and Human Services. *The Health Consequences of Smoking–50 Years of Progress: A Report of the Surgeon General*. Atlanta, GA: U.S. Department of Health and Human Services, CDC; 2014.
2. Jamal A, Homa DM, O'Connor E, et al. Current cigarette smoking among adults–United States, 2005–2014. *Morb Mortal Wkly Rep*. 2015;64:1233–1240.
3. Koopmans, JR, Slutske WS, Heath AC, et al. The genetics of smoking initiation and quantity smoked in Dutch adolescent and young adult twins. *Behav Genet*. 1999;29(6):383–393. doi:10.1023/A:1021618719735
4. Li MD, Cheng R, Ma JZ, et al. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction*. 2003;98(1):23–31. doi:10.1046/j.1360-0443.2003.00295.x
5. Maes HH, Sullivan PF, Bulik et al. A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence. *Psychol Med*. 2004;34(7):1–11. doi:10.1017/S0033291704002405
6. Vink JM, Willemsen G, Boomsma. DI Heritability of smoking initiation and nicotine dependence. *Behav Genet*. 2005;35(4):397–406. doi:10.1007/s10519-004-1327-8
7. Lubke GH, Stephens SH, Lessem JM, et al. The CHRNA5/A3/B4 gene cluster and tobacco, alcohol, cannabis, inhalants and other substance use initiation: replication and new findings using mixture analyses. *Behav Genet*. 2012;42(4):636–646. doi:10.1007/s10519-012-9529-y
8. Schlaepfer IR, Hoft NR, Collins AC, et al. The CHRNA5/A3/B4 gene cluster variability as an important determinant of early alcohol and tobacco initiation in young adults. *Biol Psychiat*. 2008;63(11):1039–1046. doi:10.1016/j.biopsych.2007.10.024
9. Ehringer MA, Clegg HV, Collins AC, et al. Association of the neuronal nicotinic receptor β2 subunit gene (CHRNB2) with subjective responses

to alcohol and nicotine. *Am J Med Genet*. 2007;144B(5):596–604. doi:10.1002/ajmg.b.30464
10. Ehringer MA, McQueen MB, Hoft NR, et al. Association of CHRN genes with "dizziness" to tobacco. *Am J Med Genet B*. 2010;153B(2):600–609. doi:10.1002/ajmg.b.31027
11. Bierut LJ, Madden PAF, Breslau N, et al. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet*. 2007;16(1):24–35. doi:10.1093/hmg/ddl441
12. Saccone NL, Saccone SF, Hinrichs AL, et al. Multiple distinct risk loci for nicotine dependence identified by dense coverage of the complete family of nicotinic receptor subunit (CHRN) genes. *Am J Med Genet B*. 2009;150B(4):453–466. doi:10.1002/ajmg.b.30828
13. Qu X, Wang K, Dong W, et al. Associations between two CHRNA3 variants and susceptibility of lung cancer: a meta-analysis. *Sci Rep*. 2016;6:20149. doi:10.1038/srep2014
14. Timofeeva MN, Hung RJ, Rafnar T, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14900 cases and 29485 controls. *Hum Molec Genet*. 2012;21(22):4980–4995. doi:10.1093/hmg/dds334
15. Munafò MR, Timofeeva MN, Morris RW, et al. Association between genetic variants on chromosome 15q25 locus and objective measures of tobacco exposure. *J Natl Cancer Inst*. 2012;104(10):740–748. doi:10.1093/jnci/djs191
16. Ware JJ, Chen X, Vink J, et al. Genome-wide meta-analysis of cotinine levels in cigarette smokers identifies locus at 4q13.2. *Sci Rep*. 2016;6:20092. doi:10.1038/srep20092
17. Bloom AJ, Hartz SM, Baker TB, et al. Beyond cigarettes per day: a genome-wide association study of the biomarker carbon monoxide. *Ann Am Thorac Soc*. 2014;11(7):1003–1010. doi:10.1513/AnnalsATS.201401-010OC
18. Chen L-S, Bloom AJ, Baker TB, et al. Pharmacotherapy effects on smoking cessation vary with nicotine metabolism gene (CYP2A6). *Addiction*. 2014;109(1):128–137. doi:10.1111/add.12353.
19. Tomaz PRX, Santos JR, Issa JS, et al. CYP2B6 rs2279343 polymorphism is associated with smoking cessation success in bupropion therapy. *Eur J Clin Pharmacol*. 2015;71(9):1067–1073. doi:10.1007/s00228-015-1896
20. Ray R, Tyndale RF, Lerman C. Nicotine dependence pharmacogenetics: role of genetic variation in nicotine-metabolizing enzymes. *J Neurogenet*. 2009;23(3):252–261. doi:10.1080/01677060802572887
21. Lee AM, Jepson C, Hoffman E, et al. CYP2B6 genotype alters abstinence rates in a bupropion smoking cessation trial. *Biol Psychiatry*. 2007;62(6):635–641. doi:10.1016/j.biopsych.2006.10.005
22. Bloom AJ, Harari O, Martinez M, et al. Use of a predictive model derived from in vivo endophenotype measurements to demonstrate associations with a complex locus, CYP2A6. *Hum Mol Genet*. 2012;21(13):3050–3062. doi:10.1093/hmg/dds114
23. Malaiyandi V, Sellers EM, Tyndale RF. Implications of CYP2A6 genetic variation for smoking behaviors and nicotine dependence. *Persp Clin Pharmacol*. 2005;77(3):145–158. doi:10.1016/j.clpt.2004.10.011
24. Tyndale RF, Sellers EM. Variable CYP2A6-mediated nicotine metabolism alters smoking behavior and risk. *Drug Metab Dispos*. 2001;29(4):548–552.
25. Thorgeirsson TE, Gudbjartsson DF, Surakka I, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet*. 2010;42(5):448–454. doi:10.1038/ng.573
26. Liu JZ, Tozzi F, Waterworth DM, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet*. 2010;42(5):436–440. doi:10.1038/ng.572
27. David SP, Hamidovic GK, Chen AK, et al. Genome-wide meta-analyses of smoking behaviors in African Americans. *Transl Psychiat*. 2012;2:e119. doi:10.1038/tp.2012.41
28. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*. 2010;42(5):441–447. doi:10.1038/ng.571
29. Rice JP, Hartz S, Agrawal A, et al. CHRNB3 is more strongly associated with FTCD-based nicotine dependence than cigarettes per day: phenotype definition changes GWAS results. *Addiction*. 2012;107(11):2019–2028. doi:10.1111/j.1360-0443.2012.03922.x
30. Hancock DB, Reginsson GW, Gaddis NC, et al. Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence. *Transl Psychiatry*. 2015;5:e651. doi:10.1038/tp.2015.149
31. Lessov CN, Martin NG, Statham DJ, et al. Defining nicotine dependence for genetic research: evidence from Australian twins. *Psychol Med*. 2004;34(5):865–879. doi:10.1017/S0033291703001582
32. Bucholz KK, Cadoret R, Cloninger CR, et al. A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. *J Stud Alcohol*. 1994;55(2):149–158. doi:10.15288/jsa.1994.55.149
33. Vieten C, Seaton KL, Feiler HS, et al. The University of California, San Francisco Family Alcoholism Study. I. Design, Methods, and Demographics. *Alcoholism Clin Exp Res*. 2004;28(10):1509–1516. doi:10.1097/01.ALC.0000142261.32980.64
34. Kalton G, Anderson DW. Sampling rare populations. *J R Stat Soc*. 1986;149(1):65–82. doi:10.2307/2981886
35. Muhib FB, Lin LS, Stueve A, et al. A venue-based method for sampling hard-to-reach populations. *Public Health Rep*. 2001;116(suppl 1):216–222. doi:10.1093/phr/116.S1.216
36. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Prob*. 1997;44(2):174–199. doi:10.2307/3096941
37. Ehlers CL, Wall TL, Betancourt M, et al. The clinical course of alcoholism in 223 Mission Indians. *Am J Psychiat*. 2004;161(7):1204–1210.
38. Han S, Gelernter J, Luo X, et al. Meta-analysis of 15 genome-wide linkage scans of smoking behavior. *Biol Psychiatry*. 2010;67(1):12–19. doi:10.1016/j.biopsych.2009.08.028
39. Lou X-Y, Ma JZ, Payne TJ, et al. Gene-based analysis suggests association of the nicotinic acetylcholine receptor β1 subunit (CHRNB1) and M1 muscarinic acetylcholine receptor (CHRM1) with vulnerability for nicotine dependence. *Hum Mol Genet*. 2006;120(3):381–389. doi:10.1007/s00439-006-0229-7
40. Philibert RA, Todorov A, Andersen A, et al. Examination of the Nicotine Dependence (NICSNP) Consortium findings in the Iowa adoption studies population. *Nicotine Tob Res*. 2009;11(3):286–292. doi:10.1093/ntr/ntn034
41. Minicã CC, Mbarek H, Pool R, et al. Pathways to smoking behaviours: biological insights from the Tobacco and Genetics Consortium meta-analysis [published online ahead of print March 29, 2016]. *Mol Psychiatry*. doi:10.1038/mp.2016.20
42. Thorgeirsson TE, Steinberg S, Reginsson GW. A rare missense mutation in CHRNA4 associates with smoking behavior and its consequences [published online ahead of print March 8, 2016]. *Mol Psychiatry*. doi:10.1038/mp.2016.13
43. Keskitalo-Vuokko K, Pitkäniemi J, Broms U, et al. Associations of nicotine intake measures with CHRN genes in Finnish smokers. *Nicotine Tob Res*. 2011;13(8):686–690. doi:10.1093/ntr/ntr059
44. Amin N, Byrne E, Johnson J, Chenevix-Trench G, et al. Genome-side association analysis of coffee drinking suggests association with CYP1A1/CYP1A2 and NRCAM. *Mol Psychiatr*. 2012;17:1116–1129. doi:10.1038/mp.2011.101
45. McMahon G, Taylor AE, Smith GD, et al. Phenotype refinement strengthens the association of AHR and CYP1A1 genotype with caffeine consumption. *PLOS One*. 2014;9(7):e103448. doi:10.1371/journal.pone.0103448
46. Chen Z, Li Z, Niu X, et al. The effect of CYP1A1 polymorphisms on the risk of lung cancer: a global meta-analysis based on 71 case-control studies. *Mutagenesis*. 2011;26(3):437–446. doi:10.1093/mutage/ger002
47. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–575. doi:10.1086/519795
48. Sun L, Wilder K, McPeek MS. Enhanced pedigree error detection. *Hum Hered*. 2002;54(2):99–110. doi:10.1159/000067666
49. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;490:56–65. doi:10.1038/nature11632

50. Muthén LK, Muthén BO. *Mplus User's Guide*. 7th ed. Los Angeles, CA: Muthén & Muthén; 1998–2012.
51. Muthén B, Muthén L, Asparouhov T. Estimator choice with categorical outcomes. 2015. www.statmodel.com/download/EstimatorChoices.pdf. Accessed May 24, 2016.
52. Akaike H. A new look at statistical model identification. *IEEE T Automat Contr*. 1974;AU-19:719–722. doi:10.1109/TAC.1974.1100705
53. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461–464. doi:10.1214/aos/1176344136
54. Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. *Psychol Assessment*. 2000;12(3):287–297. doi:10.1037//1040-3590.12.3.287
55. Yu C-Y. *Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models With Binary and Continuous Outcomes [Unpublished doctoral dissertation]*. University of California, Los Angeles; 2002.
56. Kline RB. *Principles and Practice of Structural Equation Modeling*. 2nd ed. New York, NY: The Guilford Press; 2011.
57. Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*. 1984;49(1):115–132. doi:10.1007/BF02294210
58. Jöreskog KG, Goldberger AS. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *J Am Stat Assoc*. 1975;70(351a):631–639. doi:10.2307/2285946
59. Price AL, Patterson NL, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904–909. doi:10.1038/ng1847
60. Yang J, Lee SH, Goddard ME, et al. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76–82. doi:10.1016/j.ajhg.2010.11.011
61. Kang HM. *Epacts: Efficient and Parallelizable Association Container Toolbox*. University of Michigan: Department of Biostatistics and Center for Statistical Genetics; 2012. www.sph.umich.edu/csg/kang/epacts/. Accessed February 1, 2016.
62. Li M-X, Yeung JMY, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet*. 2012;131(5):747–756. doi:10.1007/s00439-011-1118-2
63. Chen L-S, Saccone NL, Culverhouse RC, et al. Smoking and genetic risk variation across populations of European, Asian, and African-American ancestry – a meta-analysis of chromosome 15q25. *Genet Epidemiol*. 2012;36(4):340–351. doi:10.1002/gepi.21654
64. Saccone NL, Culverhouse RC, Schwantes-An T-H, et al. Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet*. 2010;6(8):e1001053. doi:10.1371/journal.pgen.1001053
65. Gabrielsen ME, Romundstad P, Langhammer A, et al. Association between a 15q25 gene variant, nicotine-related habits, lung cancer and COPD among 56307 individuals from the HUNT study in Norway. *Eur J Hum Genet*. 2013;21:1293–1299. doi:10.1038/ejhg.2013.26
66. Wen L, Jiang K, Yuan W, et al. Contribution of variants in CHRNA5/A3/B4 Gene Cluster on Chromosome 15 to tobacco smoking: from genetic association to mechanism. *Mol Neurobiol*. 2016;53(1):472–484. doi:10.1007/s12035-014-8997-x
67. Loukola A, Buchwald J, Gupta R, et al. A genome-wide association study of a biomarker of nicotine metabolism. *PLoS Genet*. 2015;11(9):e1005498. doi:10.1371/journal.pgen.1005498
68. Hendricks PS, Delucchi KL, Benowitz NL, et al. Clinical significance of early smoking withdrawal effects and their relationships with nicotine metabolism: preliminary results from a pilot study. *Nicotine Tob Res*. 2014;16(5):615–620. doi:10.1093/ntr/ntt204
69. Rubinstein ML, Benowitz NL, Auerback GM, et al. Rate of nicotine metabolism and withdrawal symptoms in adolescent light smokers. *Pediatrics*. 2008;122(3):e643–e647. doi:10.1542/peds.2007-3679
70. Lerman C, Tyndale R, Patterson F, et al. Nicotine metabolite ratio predicts efficacy of transdermal nicotine for smoking cessation. *Clin Pharmacol Ther*. 2006;79(6):600–608. doi:10.1016/j.clpt.2006.02.006
71. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2012;45(6):580–585. doi:10.1038/ng.2653
72. Hackett NR, Butler MW, Shaykhiev R, et al. RNA-Seq quantification of the human small airway epithelium transcriptome. *BMC Genomics*. 2012;13:82. doi:10.1186/1471-2164-13-82
73. Fischer AJ, Goss KL, Scheetz TE, et al. Differential gene expression in human conducting airway surface epithelia and submucosal glands. *Am J Resp Cell Mol*. 2009;40(2):189–199. doi:10.1165/rcmb.2008-0240OC
74. Choi JS, Lee WJ, Baik SH, et al. Array CGH reveals genomic aberrations in human emphysema. *Lung*. 2009;187(3):165–172. doi:10.1165/rcmb.2008-0240OC
75. Wu X, Sun X, Chen C, et al. Dynamic gene expressions of peripheral blood mononuclear cells in patients with acute exacerbation of chronic obstructive pulmonary disease: a preliminary study. *Crit Care*. 2014;18:508. doi:10.1186/s13054-014-0508-y
76. Shin SY, Fauman EB, Petersen AK, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet*. 2014;46(6):543–550. doi:10.1038/ng.2982
77. Zhao Y, Peng J, Lu C, et al. Metabolomic heterogeneity of pulmonary arterial hypertension. *PloS One*. 2014;9(2):e88727. doi:10.1371/journal.pone.0088727
78. Keusch S, Hildenbrand FF, Bollmann T, et al. Tobacco smoke exposure in pulmonary arterial and thromboembolic pulmonary hypertension. *Respiration*. 2014;88(1):38–45. doi:10.1159/000359972
79. Schiess R, Senn O, Fischler M, et al. Tobacco smoke: a risk factor for pulmonary arterial hypertension?: a case-control study. *CHEST*. 2010;138(5):1086–1092. doi:10.1378/chest.09-2962