### Citation

### Published Version
doi://10.1371/journal.pone.0004653

### Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:5978725

### Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

---

**Genome-Wide and Candidate Gene Association Study of Cigarette Smoking Behaviors**

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.
Genome-Wide and Candidate Gene Association Study of Cigarette Smoking Behaviors

Neil Caporaso1,2*, Fangyi Gu2*, Nilanjan Chatterjee1, Jin Sheng-Chih3, Kai Yu1, Meredith Yeager1,5, Constance Chen2, Kevin Jacobs5,6, William Wheeler7, Maria Teresa Landi1, Regina G. Ziegler1, David J. Hunter1,2,5, Stephen Chanock4, Susan Hankinson2, Peter Kraft2,9, Andrew W. Bergen9, and John P. Neale1,2

1 Division of Cancer Epidemiology and Genetics, NCI, Bethesda, Maryland, United States of America, 2 Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America, 3 Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, Baltimore, Maryland, United States of America, 4 Channing Laboratory, Brigham and Women’s Hospital, Boston, Massachusetts, United States of America, 5 Core Genotyping Facility, Advanced Technology Center, National Cancer Institute, Gaithersburg, Maryland, United States of America, 6 BioInformed LLC, Gaithersburg, Maryland, United States of America, 7 Information Management Services, Inc., Rockville, Maryland, United States of America, 8 Molecular Genetics Program, Center for Health Sciences, SRI International, Menlo Park, California, United States of America

Abstract

The contribution of common genetic variation to one or more established smoking behaviors was investigated in a joint analysis of two genome-wide association studies (GWAS) performed as part of the Cancer Genetic Markers of Susceptibility (CGEMS) project in 2,329 men from the Prostate, Lung, Colon and Ovarian (PLCO) Trial, and 2,282 women from the Nurses’ Health Study (NHS). We analyzed seven measures of smoking behavior, four continuous (cigarettes per day [CPD], age at initiation of smoking, duration of smoking, and pack years), and three binary (ever versus never smoking, ≤10 versus >10 cigarettes per day [CPDBI], and current versus former smoking). Association testing for each single nucleotide polymorphism (SNP) was conducted by study and adjusted for age, cohabitation/marital status, education, site, and principal components of population substructure. None of the SNPs achieved genome-wide significance (p<10−5) in any combined analysis pooling evidence for association across the two studies; we observed between two and seven SNPs with p<10−3 for each of the seven measures. In the chr15q25.1 region spanning the nicotinic receptors CHRNA5 and CHRNA3, we identified multiple SNPs associated with CPD (p<10−3), including rs1051730, which has been associated with nicotine dependence, smoking intensity and lung cancer risk. In parallel, we selected 11,199 SNPs drawn from 359 a priori candidate genes and performed individual-gene and gene-group analyses. After adjusting for multiple tests conducted within each gene, we identified between two and five genes associated with each measure of smoking behavior. Besides CHRNA3 and CHRNA5, MAOA was associated with CPDBI (gene-level p<5.4×10−3), our analysis provides independent replication of the association between the chr15q25.1 region and smoking intensity and data for multiple other loci associated with smoking behavior that merit further follow-up.

Introduction

Cigarette smoking is a risk factor for more than two dozen diseases and the single biggest cause of preventable mortality worldwide [1]. Although public awareness of the dangers of smoking is widespread and public health measures such as public building smoking restrictions and increased cigarette taxes have had salutary effects on smoking rates, dependence on nicotine, the major psychoactive component in tobacco, induces most people who start smoking to continue to smoke in spite of their wish to quit [2]. Environmental influences on tobacco dependence including cultural perceptions and economics, low socioeconomic status, peer smoking and maternal smoking during pregnancy are well documented. Nevertheless, twin studies provide strong evidence that a range of diverse smoking phenotypes including age at initiation, intensity, and cessation have a substantial hereditary component [1,3,4,5,6]. Identifying the specific loci that influence smoking behaviors (including initiation, intensity and cessation) could lead to important etiological insights and facilitate the development of treatments to further reduce smoking related mortality.

Genome-wide linkage studies have identified chromosomal regions that may harbor loci contributing to one of many smoking behavior phenotypes: age at initiation [7,8]; some variant of CPD [9,10,11,12,13,14,15,16,17] DSM-IV Nicotine Dependence [11,17]; some variant of Ever-Never [9,17,18,19,20]; Fagerstrom test for nicotine dependence (FTQ, FTND) or Heaviness Smoking Index (HIS) [14,21,22]; Pack-years [9,23]; Current versus Former...
and withdrawal severity [11]. While some regions have shown suggestive linkage to smoking behaviors in multiple studies [24], linkage results have generally been heterogeneous and short on conclusive findings; to date no risk loci have been discovered that definitively account for linkage signals.

Until very recently, candidate gene association studies have focused on genes in a few candidate pathways. A 'reward deficiency syndrome' has been postulated as one unifying theme to account for the role of diverse neurotransmitters in nicotine dependence [25,26,27], and consequently many studies have evaluated genes in opioid [28,29], serotonergic [30,31,32], dopaminergic [26,33,34,35], drug metabolizing enzyme [36,37,38,39,40,41] and nicotinic and muscarinic cholinergic receptor pathways [42,43]. Results from these studies have been largely equivocal, due to small sample sizes in individual studies, incomplete and non-overlapping genetic coverage, differences in measures of smoking behavior, or differences in genetic and environmental backgrounds. It is also highly probable that many of the loci that influence smoking behavior lie outside of the previously-studied candidate regions.

A recent genome-wide association study of over 13,000 smokers identified a region on chromosome 15q25.1 associated with smoking intensity (number of cigarettes smoked per day) [44]. This region, spanning the nicotinic acetylcholine receptors, CHRNA5 and CHRNA3, and CHRNA4 and CHRNA7 and was also identified in a recent GWAS of dichotomized smoking intensity [45], and in two genome-wide association scans of lung cancer [46,47]. It was unclear whether the association between SNPs in this region and lung cancer was due to a genetic effect on smoking behavior, an independent effect on lung carcinogenesis, or both [48]. Two recent candidate gene studies together including almost 5000 smokers both found SNPs in nicotinic receptors including the chr15p25.1 nicotinic receptor loci to be associated with nicotine dependence [49,50].

To identify loci associated with smoking initiation, intensity and cessation we performed a genome-wide association study (GWAS) using data from subjects genotyped as part of the Cancer Genetic Markers of Susceptibility (CGEMS) project, including 2,617 ever-smokers [51,52]. In addition to single-marker tests of association in the GWAS, we also report results from gene- and gene-group-level tests of association of 359 candidate genes in 30 functional groups.

### Methods

#### Samples and genotypes

Subjects were drawn from two previous genome-wide association studies (GWAS), performed as part of the Cancer Genetic Markers of Susceptibility (CGEMS) project [51,52]. Data on smoking behaviors were available on 2,060 men from the Prostate, Lung, Colon and Ovarian Trial (PLCO) (1,172 prostate cancer cases and 1,157 controls) and on 2,282 postmenopausal women (1,145 with breast cancer and 1,142 controls) from the Nurses' Health Study (NHS). All subjects were of self-reported European ancestry, which was consistent with genetic analyses of population structure [33]. Samples from the PLCO were genotyped using the Illumina HumanHap 300 k and HumanHap 240 k platforms [52]; those from the NHS were genotyped using the Illumina HumanHap 550 k platform [51]. Genotyping was performed at the same laboratory and similar genotyping quality control (QC) procedures were used in each study. Individual samples were removed if more than 10% of SNPs failed genotyping, and individual SNPs were removed if more than 10% of samples failed. The average call rate for both PLCO and NHS samples was 99.8%. Combined genome-wide analyses were restricted to directly-genotyped SNPs with minor allele frequencies above 1% in each study (ca. 510,000 SNPs). Additional description of these studies is available in previous reports [51,52].

#### Adjustment for population stratification

For both PLCO and NHS, analyses of population stratification were conducted using approximately 10,000 unlinked SNP markers [51,52]. Individual European, Asian and African admixture proportions were estimated by STRUCTURE [53] applied to CGEMS data augmented by the HapMap CEU, CHB+JPT and YRI samples. Subjects with significant non-European admixture were excluded for PLCO and NHS. Residual within-Europe population stratification was estimated using the top three (PLCO) or four (NHS) principal components of genetic variation, as calculated using EIGENSTRAT [54].

### Smoking behaviors

We selected four continuous and three binary smoking behaviors for analysis (Table 1). The continuous measures were cigarettes smoked per day (CPD), age at smoking initiation (SMKAGE), duration of smoking (SMKDU) and pack-years (PKYRS), the binary measures were ever versus never smoking.

#### Table 1. The distribution of smoking behaviors and covariates in NHS and PLCO CGEMS samples.

<table>
<thead>
<tr>
<th>Smoking Behaviors or Covariates</th>
<th>NHS (n = 2282)</th>
<th>PLCO (n = 2060)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPD (Cigarettes per day, mean±SD)</td>
<td>18.5±10.5</td>
<td>22.0±13.6</td>
</tr>
<tr>
<td>SMKAGE (Age of initiation, years, mean±SD)</td>
<td>19.6±3.7</td>
<td>18.1±5.0</td>
</tr>
<tr>
<td>SMKDU (Duration, years, mean±SD)</td>
<td>25.3±15.5</td>
<td>27.7±13.8</td>
</tr>
<tr>
<td>PKYRS (Pack-years, mean±SD)</td>
<td>24.9±22.2</td>
<td>31.8±27.3</td>
</tr>
<tr>
<td>EVNV (N, %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smokers</td>
<td>1038 (45)</td>
<td>687 (33)</td>
</tr>
<tr>
<td>Ever smokers</td>
<td>1244 (55)</td>
<td>1373 (67)</td>
</tr>
<tr>
<td>CIGSTAT (N, %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current smokers</td>
<td>137 (11)</td>
<td>212 (15)</td>
</tr>
<tr>
<td>Former smokers</td>
<td>1107 (89)</td>
<td>1161 (85)</td>
</tr>
<tr>
<td>CPDBI (N, %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤10 CPD</td>
<td>906 (75)</td>
<td>1167 (85)</td>
</tr>
<tr>
<td>&gt;10 CPD</td>
<td>305 (25)</td>
<td>202 (15)</td>
</tr>
<tr>
<td>Living Status (N, %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alone/not married</td>
<td>457 (20)</td>
<td>271 (13)</td>
</tr>
<tr>
<td>Not alone/married</td>
<td>1825 (80)</td>
<td>1789 (87)</td>
</tr>
<tr>
<td>Education (N, %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No college</td>
<td>0</td>
<td>838 (41)</td>
</tr>
<tr>
<td>Some college</td>
<td>1591 (70)</td>
<td>409 (20)</td>
</tr>
<tr>
<td>Bachelors</td>
<td>463 (20)</td>
<td>392 (19)</td>
</tr>
<tr>
<td>Masters or above</td>
<td>228 (10)</td>
<td>421 (20)</td>
</tr>
<tr>
<td>Age (N, %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55–59 years</td>
<td>111 (5)</td>
<td>400 (19)</td>
</tr>
<tr>
<td>60–64</td>
<td>437 (19)</td>
<td>648 (31)</td>
</tr>
<tr>
<td>65–69</td>
<td>411 (18)</td>
<td>651 (32)</td>
</tr>
<tr>
<td>70–74</td>
<td>581 (26)</td>
<td>361 (18)</td>
</tr>
<tr>
<td>&gt;74</td>
<td>742 (32)</td>
<td>0</td>
</tr>
</tbody>
</table>

* descriptive statistics for smoking behaviors included ever smokers only.
doi:10.1371/journal.pone.0004653.t001
status (EVNV), smokers who quit versus those who did not (CIGSTAT), and a binary measure of smoking intensity (CPDBI, defined as ten or more cigarettes per day versus fewer than ten). Only current or former smokers were included in the analyses involving the smoking phenotypes with the exception of EVNV which included never smokers.

All of these behaviors were measured by baseline questionnaire (BQ) in the PLCO (administered from 1994–2001) [53]. Age at initiation was defined as the age when a subject started smoking “regularly for six months or longer.” Former smokers were defined as ever-smokers who did not smoke regularly at BQ and were asked to report the age at which they stopped smoking regularly. Ever smokers were asked to provide information on the number of cigarettes they smoked per day, in categories (1–10, 11–20, 21–30, 31–40, 41–60, 61–80, over 80). For continuous analyses we assigned subjects to the midpoint of their category (or 90 cigarettes per day for over 80). Duration was derived from data on age at smoking initiation and age at cessation. Pack years was derived by converting cigarettes per day to packs per day (CPD/20) and multiplying this figure by duration.

In the NHS, SMKAGE was measured as at BQ; all other behaviors were measured using cumulative information from the BQ (administered in 1976) and subsequent follow-up questionnaires (one every two years). The majority of NHS subjects (2149) had smoking data available up through the 2002 questionnaire. For those few women (133) with no smoking data available from the 2002 follow-up cycle, we used data from the latest available follow-up. Age at initiation was defined as the age when a subject started smoking cigarettes “regularly.” Former smokers were defined as ever-smokers who identified themselves as non-smokers on any questionnaire (and did not identify as a smoker on any subsequent questionnaire). Age at cessation was explicitly asked in NHS BQ. For women who quit smoking after the BQ, age at cessation was inferred as the median age between the questionnaire that identified her as a smoker. Prior to 1982, current or former smokers were asked to write in the average number of cigarettes they smoked per day; subsequent questionnaires captured information about smoking intensity in categories (1–4, 5–14, 15–24, 25–34, 35–44, and over 44 cigarettes per day). Pack years was estimated as the sum of the products of smoking intensity (categories were assigned midpoint values, e.g. 5–14 was coded as 10 cigarettes per day, or 0.5 packs per day) at questionnaire k times the interval between questionnaire k and questionnaire k+1 (or half that interval for women who were smokers at questionnaire k and non-smokers at questionnaire k+1). Smoking duration was calculated as the sum of all intervals where a woman was a smoker. Average cigarettes per day (the CPD variable used in this GWAS) was calculated as pack-years divided by smoking duration.

Association analyses

Continuous phenotypes were log transformed to achieve approximate normality and SNP genotypes were coded as counts of minor alleles. For each study, we defined any phenotype that was more than three standard deviations from the mean to be an outlier. Outliers that were above (below) the mean were then truncated to the 99th (1st) percentile of the raw distribution. We tested for association between each SNP marker and each continuous phenotype using linear regression, adjusted for study center (PLCO) or geographic region (NHS); age at smoking assessment in five-year bins (baseline for PLCO or last available follow-up for NHS); marital status (married versus not; PLCO) or living arrangement (living alone or with others, NHS); education (4 categories PLCO, 5 categories NHS); prostate (PLCO) or breast (NHS) cancer case-control status; and selected principal components of genetic variation. For binary traits, we used unconditional logistic regression, adjusted for the same covariates. These tests were conducted separately for PLCO and NHS. For SNPs that passed QC filters and had minor allele frequency above 1% in both studies, we combined evidence for association across PLCO and NHS using weighted Z-scores [56]. Heterogeneity in SNP-smoking behavior associations across study was assessed using Q and F statistics [57]. Power calculations were performed using Quanto [http://hydra.usc.edu/GxE]/[58].

Analyses of candidate genes and candidate gene groups

We selected 359 genes for additional analyses, based on their hypothesized relationship to smoking behavior. 349 of these genes were previously selected by the NICSNP Candidate Gene Committee [50]. We added ten candidate genes identified from two recent GWAS of dichotomized measures of nicotine dependence (CTNNB1, FBLX17, FTO, NRX1, PBX2, TRPC7, FTS13A) and CPD (DGE1, RORB, SLC03A4) [45,59]. For each candidate gene we tested the null hypothesis that no SNP within 20 kb upstream of the start of transcription and 10 kb downstream of the stop of transcription (based on NCBI build 35.1) was associated with smoking behaviors using a parametric permutation procedure that allows for covariate adjustment. We compared the smallest observed p-value for any SNP in the candidate gene region to the empirical null distribution of the smallest p-value based on 20,000 random permutations. This approach provides a gene-level p-value that is adjusted for both the number of SNPs in the gene region and their linkage disequilibrium structure.

Candidate genes (and the SNPs in the corresponding gene regions) were grouped based on known functional similarity. We used a slightly modified version of the groups developed by the NICSNP Candidate Gene Committee, (Table S1). To test for association between SNPs in a group and smoking behaviors, we used a modified rank truncated product method [60] which compares the product of the ten smallest gene-level p-values over all the genes in the group to its simulated null distribution. Such a group or pathway level analysis potentially has more power to detect associations when a group containing multiple susceptibility genes each has modest evidence for association [61].

Chr15q25.1 SNP imputation

Multiple SNPs in the 15q25.1 region have been shown to be associated with CPD, nicotine dependence, or risk of lung cancer [44,47,49,50,59]. We directly genotyped some of these SNPs and could impute others using the observed genotypes in PLCO and NHS samples and the phased HapMap CEU samples (Release 21). Imputation, restricted to the region of chromosome 15q23, was performed for each study separately using the Hidden Markov Model implemented in Mach 1.0 [62]. All of the imputed genotypes had high quality scores (R-squares>0.8 for 95% of SNPs in the region).

Results

The distributions of the smoking behaviors and demographic covariates included in the analysis of the NHS and PLCO datasets are presented in Table 1. The men in the PLCO sample have smoking behaviors that are more prevalent and more severe (greater frequency of ever, current and heavy smokers, earlier age of onset, longer duration, and greater pack-years) than do the women in the NHS sample. The direction and significance of
correlations among the smoking phenotypes within the dataset are similar, with all correlations highly significant (P<0.0001), except for the correlation between age at initiation and cigarettes per day in the NHS sample (Table S2).

Quantile-quantile plots of the −log_{10} p-values for SNP association with smoking behaviors (Figure S1) showed no evidence for systematic bias (each genomic inflation factor λ<1.02). None of the SNPs achieved genome-wide significance (p<10^{-5}) in any combined analysis pooling evidence for association across the two studies (Figures 1 and 2). Table 2 lists detailed results for SNPs with a combined-analysis p-value<10^{-5} for each smoking behavior. For the combined GWAS analysis of the seven smoking behaviors, the most significant SNP smoking behavior association result is rs6437740 with CPD (P=2.4×10^{-7}). Including this result, there are 8 gene regions and 3 genomic regions with predicted but not verified coding regions associated with SNPs in the group of SNP smoking behavior results with P<10^{-5} (Table 2). We observed no evidence for systematic heterogeneity in results between studies, and no single SNP showed evidence for heterogeneity by sex at the genome-wide significance level (see summary of Q statistics in Figure S2).

We analyzed 359 candidate genes previously nominated as candidate genes for nicotine dependence [50], or identified in GWAS studies of dichotomized nicotine dependence and CPD [45,59] and gene-level results are summarized in Table 3. Of note, rs3027409 in the MAOA candidate gene region had a p-value of 6.7×10^{-6} for association with CPDBI, which led to a gene-level p-value smaller than 5.4×10^{-5}, the smallest a priori candidate gene association we observed. Nine candidate gene groups were associated with at least one smoking behavior at the 0.10 level (Table 4) with two (Nicotinic Receptors and Voltage-Dependent Calcium-Activated Potassium Channels) associated with a smoking behavior (CPD) at the 0.01 level.

Figure 3 and Table S3 summarize the associations between genotyped and imputed SNPs and CPD in PLCO and NHS smokers for the chr15q25.1 region spanning CHRNA3 and CHRNA5. The strongest association signal we observe in this region is at rs2036527 (combined P=8×10^{-5}), located 10,051 base pairs 3′ of PSMA4 and 6,290 base pairs 5′ of CHRNA5 in a region of strong linkage disequilibrium spanning both genes.

**Discussion**

We performed genome wide association analyses for seven related smoking behaviors in two datasets totaling 4,611 individuals and 2617 ever smokers. We selected smoking behaviors with established hereditary components [4,5,21,63] and public health relevance[64,65]. To the best of our knowledge this study represents the first genome-wide association study of duration of smoking, pack years, and age at initiation of smoking. The sample size is also larger than most published candidate gene association studies of smoking behavior [66] and two previous genome-wide association studies of smoking behaviors [45,59].

Although we did not discover novel genome-wide significant (p<10^{-7}) associations, we did find additional evidence for an association between genetic variants in the chr15q25.1 region and number of cigarettes smoked per day. Candidate gene analyses also provided suggestive evidence for association between variants in the MAOA gene region and the smoking behavior cigarettes per day.

The lack of genome-wide significant results suggests that common variants have at most a modest influence on smoking behavior. We had adequate power to detect a variant that

---

**Figure 1.** −log_{10} p-values for association with seven continuous (Figure 1) and categorical (Figure 2) smoking behaviors. P-values are based on the combined evidence for association from both PLCO and NHS.

doi:10.1371/journal.pone.0004653.g001
explained even 2.5% of the variation in cigarettes per day. We had 61% power in the NHS sample and 71% power in the PLCO sample to detect such a variant at the $10^{-7}$ level; the power of the combined analysis was greater than 99%. Conversely, the lack of genome-wide significant findings does not rule out the existence of (many) common variants with small individual effects on smoking behavior, since our power to detect any one is small. Even with our relatively large sample size, our power to detect a variant similar to the 15q25.1 SNP rs1051730 (which was estimated to explain about 0.7% of the trait variance [44] at the genome-wide significance level) was only 8.5% for the combined analysis (and less than 1% for either study alone).

SNPs at the nicotinic receptor candidate genes CHRNA3 (chr15q25.1) and CHRNA1 (chr2q31.1) are associated in the CGEMS sample with three smoking behaviors: CPD, PKYRS and SMKAGE (Table 3). Another candidate gene association study investigating 348 of 359 candidate genes included in this study [50] evaluated association with a dichotomized nicotine dependence phenotype, and identified nicotinic receptor SNPs associated with FTND, including rs578776 and rs1051730 within CHRNA3, and rs16969968 within CHRNA5. Nicotinic receptors are also associated with CPD in the candidate gene group analysis as the most significantly associated gene group, and also with the phenotype SMKDU (Table 4).

Finally, we combined our chr15q25.1 results with data from three other published reports (Table S3) [44,46,47]. The SNP rs1051730, found within CHRNA5 (Ex5+268), was highly statistically significantly associated with CPD ($p=5 \times 10^{-32}$); the SNP rs8034191 (LOC123688 IVS2+256) was also highly statistically significantly associated with CPD ($p=2 \times 10^{-28}$). These SNPs were evaluated using a total of 26,789 (rs1051730) or 24,891 (rs8034191) smokers from this study and two other reports. The CHRNA5 SNP rs16969968 (Ex5-54, D398N) was significantly associated ($p<.01$) with CPD in this study but not an earlier, smaller study; combined evidence for association in 3,464 smokers remained significant ($p=2 \times 10^{-3}$). Comparative judgments of the relative importance of the individual SNPs are not possible due to the different sample sizes, the strong LD among the SNPs and the inability to adjust for the effects of the other SNPs in this meta-analysis.

Our candidate gene analyses identified an association (rs3027409, $p=5.4 \times 10^{-5}$) between genetic variation in MAOA and a dichotomized measure of smoking intensity (10 or less cigarettes smoked per day versus more than 10). This was the only gene-level result that remained significant after Bonferroni correction for the number of genes tested, which we regard as a conservative multiple-testing correction. This association is notable because of the role of the monoamine oxidases in the regulation of catecholamines and the inhibition of monoamine oxidases A and B by tobacco smoke [67]. There is substantial evidence that smoking results in reduced levels of the monoamine oxidase enzymes [67,68] and subsequent reduced catabolism of dopamine likely contributes to the reinforcing and motivating effects of smoking. Investigation of MAO-related polymorphisms in relation to alcoholism [69,70], Parkinson disease [71,72,73] and smoking [34,67,70,74,75,76,77] have yielded mixed results; our results suggest further investigation of this X-chromosome locus is warranted.

The gene group analysis that we performed provides one way to summarize the statistical evidence for association between a trait and multiple genetic variants across groups of genes that share sequence similarity and function. Nicotinic cholinergic receptors
and voltage-dependent calcium-activated potassium channels were significantly associated with CPD (gene group P<0.01). We have previously discussed nicotinic receptor findings above. The association of rs7050529 (IVS3+286 of TRPC5) with CPD (Table 2) is notable as a closely related family member, TRPC7, was previously significantly associated with nicotine dependence [59]. The transient receptor potential cation family is a superfamily of 28 genes coding for cationic ion channels responding to temperature, endogenous and exogenous organic compounds, Ca2+ flux, and mechanical stimuli, and are expressed in nearly every tissue [78]. This study, the NICSNP study and Feng et al., 2006 [79] have identified significant associations between five Transient Receptor Family Potential (TRP) subfamilies and nicotine related behaviors in the canonical (this study Table 2, [59] Table 1, and [79]) and vanilloid subfamilies (this study Table 3, and Saccone et al., 2007, Supplementary Material [50]). Recently, Gu et al., 2005 [80] have shown that vanilloid subfamily members are expressed in the lung and are responsible for the pulmonary chemoreflex response, suggesting further study of these TRP subfamilies and their potential role in smoking behavior and downstream consequences may be fruitful. The cytochrome P450, cell cycle control, and alcohol dehydrogenase candidate genes groups also exhibited nominally significant (0.01 < Ppermuted < 0.05) associations with smoking behaviors (Table 4). The cytochrome P450 results may have been driven by association between SNPs at CYP2B6 with EVNV, CYP2A6 and SMKAGE (Table 3). These results are consistent with evidence for the relationship between CYP2A6 genetic variation and both nicotine metabolism [81,82,83] and smoking behavior [41,84].

In our study, the observed association between cell cycle control genes and quit status may be driven by association of SNPs at FBXL17 (gene-level, p = 0.021, rs1433050) and NFKB1 (rs10489113, gene-level P = 0.022). FBXL17 is one of 68 members of the human F-box protein superfamily, a large group of ubiquitin ligases [85]. Ubiquitin ligases function in the ubiquitin-proteasome complex, which regulates protein assembly, trafficking and degradation, a cellular activity itself regulated by nicotine [86].
FBXL17 was also identified in the NICSNP GWAS [59] as significantly associated with FNTD, via another SNP (rs10793832). None of the SNPs in the same high linkage-disequilibrium bin as rs10793832 (according to the Pelegen genome browser) were in high linkage disequilibrium with rs1433050, the FBXL17 SNP identified in this study. One SNP genotyped in this study (rs885624) was in the same LD block as rs10793832 but was not significantly associated with quit status in either this study alone or in the combined analysis (p = 0.39).

The finding that the alcohol dehydrogenases genes were significantly associated with the smoking behavior EVNV in this analysis (e.g., ADH4 gene-level P = 0.048 (rs3828541), and ADH6, gene-level P = 0.034 (rs3857224) suggests that genetic variation at these ADH loci may influence the establishment of smoking behavior. However this analysis did not control for alcohol consumption and so this finding should be considered preliminary.

Because of the large number of male and female smokers, we were able to conduct genome-wide association scans stratified by gender (study), and conduct a genome-wide association scan for differences in genetic effect between men and women. Such analyses are important, because the effect for some loci may differ between men and women or be restricted to one gender, e.g., due to differences in the environment. However, no SNPs achieved genome-wide significance for association with any smoking behavior in either study, and no SNP achieved genome-wide significance for heterogeneity in effect between men and women (between studies).

This study has several strengths. We performed a GWAS and candidate gene study investigating a variety of smoking behaviors.

<table>
<thead>
<tr>
<th>Smoking Behavior</th>
<th>Candidate Gene Region</th>
<th>N SNPs</th>
<th>Significant SNP</th>
<th>GWAS P weighted</th>
<th>Permutated P</th>
<th>beta PLCO</th>
<th>beta NHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPD</td>
<td>MAOA</td>
<td>8</td>
<td>rs2235186</td>
<td>3.8E-04</td>
<td>1.90E-03</td>
<td>0.03</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>TRPV1</td>
<td>28</td>
<td>rs4790520</td>
<td>9.52E-05</td>
<td>2.43E-03</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>CHRNA3</td>
<td>15</td>
<td>rs12914385</td>
<td>3.20E-04</td>
<td>3.05E-03</td>
<td>0.05</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>CHRNA5</td>
<td>8</td>
<td>rs1051730</td>
<td>6.13E-04</td>
<td>3.10E-03</td>
<td>0.04</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>FOSB</td>
<td>5</td>
<td>rs2238686</td>
<td>1.63E-03</td>
<td>6.86E-03</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>SMKAGE</td>
<td>SLC1A2</td>
<td>55</td>
<td>rs16927393</td>
<td>2.89E-05</td>
<td>1.10E-03</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>CYP2A6</td>
<td>2</td>
<td>rs8102683</td>
<td>1.57E-03</td>
<td>2.76E-03</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>RYR1</td>
<td>33</td>
<td>rs8107027</td>
<td>2.32E-04</td>
<td>7.14E-03</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>CHRNA1</td>
<td>8</td>
<td>rs2646165</td>
<td>1.45E-03</td>
<td>9.29E-03</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>SMKDU</td>
<td>GRM1</td>
<td>49</td>
<td>rs12197749</td>
<td>9.81E-05</td>
<td>3.24E-03</td>
<td>0.07</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>NCOA1</td>
<td>11</td>
<td>rs9309308</td>
<td>3.74E-04</td>
<td>3.48E-03</td>
<td>0.06</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>ADCY3</td>
<td>27</td>
<td>rs2278483</td>
<td>4.46E-04</td>
<td>7.29E-03</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td>PKYRS</td>
<td>MAOA</td>
<td>8</td>
<td>rs2235186</td>
<td>6.09E-05</td>
<td>1.90E-03</td>
<td>0.06</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>GRM6</td>
<td>12</td>
<td>rs1845940</td>
<td>3.68E-05</td>
<td>3.33E-04</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>PIK3C2G</td>
<td>84</td>
<td>rs2305220</td>
<td>8.04E-05</td>
<td>5.14E-03</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>GPR51</td>
<td>158</td>
<td>rs1000440</td>
<td>4.67E-05</td>
<td>5.33E-03</td>
<td>0.07</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>CHRNA3</td>
<td>15</td>
<td>rs12914385</td>
<td>1.04E-03</td>
<td>9.29E-03</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>EVNV</td>
<td>CYP2B6</td>
<td>13</td>
<td>rs2014141</td>
<td>7.31E-05</td>
<td>7.36E-04</td>
<td>0.11</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>SLC6A3</td>
<td>23</td>
<td>rs466049</td>
<td>2.61E-04</td>
<td>5.79E-03</td>
<td>0.13</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>CIGSTAT</td>
<td>2</td>
<td>rs12845178</td>
<td>1.20E-03</td>
<td>1.24E-03</td>
<td>0.12</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>NR3C2</td>
<td>82</td>
<td>rs10050229</td>
<td>1.47E-04</td>
<td>8.69E-03</td>
<td>0.32</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>CPDBI</td>
<td>MAOA</td>
<td>8</td>
<td>rs30072409</td>
<td>6.73E-06</td>
<td>5.36E-05*</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>ATM</td>
<td>9</td>
<td>rs469261</td>
<td>7.16E-04</td>
<td>4.81E-03</td>
<td>0.15</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>FOSB</td>
<td>5</td>
<td>rs2238686</td>
<td>2.22E-03</td>
<td>8.29E-03</td>
<td>0.42</td>
<td>0.20</td>
</tr>
</tbody>
</table>

*Bonferroni P value, as 20,000 permutations were too few to accurately determine the p-value in that range.

1Gene regions can overlap, so some SNPs may be in multiple gene regions.

Table 4. Candidate Gene Groups with gene-level P < 0.10.
with public health importance for the first time in a sample unselected for smoking behaviors and/or smoking attributable disease. We confirm important findings from recent GWAS and candidate gene studies of nicotine dependence and CPD. Our sample size is relatively large, yet still not large enough to reliably detect variants with modest effects on smoking behaviors. The absence of selection bias in the cohort bases for the samples enhances generalizability to U.S. non-Hispanic whites although a modest limitation is that the education level in both cohorts is above average. By limiting analyses to subjects of European ancestry and adjusting for principal components of population structure, we minimized risk of false positives due to population stratification, but are not be able to detect SNP alleles associated with smoking behavior that are common in non-Europeans but rare among European-Americans. The smoking behavior characteristics for the two studies are quite similar after taking into account expected differences by gender (Table 1), and the correlation of smoking behaviors are similar within NHS and PLCO (see Table S1). The combined sample has the advantage of increased power and generalizability.

The diverse smoking behaviors we investigated represent the spectrum of key events in an individual’s smoking history from initiation (age at initiation, ever never smoking) thru establishment of dependency (smoking duration, smoking intensity, and pack

Figure 3. $-\log_{10}$ p-values for association with number of cigarettes smoked per day (CPD) for SNPs in the chr15q25.1 region. P-values are based on the combined evidence for association from both PLCO and NHS. Filled symbols denote genotyped SNPs; open symbols denote imputed SNPs. Black diamonds (squares) denote SNPs associated with continuous (binary) CPD in previous reports. Red circles denote SNPs associated with lung cancer in previous reports.

doi:10.1371/journal.pone.0004653.g003
years), to outcome (current versus former cigarette smoking status), with potential genetic influence at each stage. The finding that selected genes are associated with multiple phenotypes may represent both correlations among the phenotypes but also pleiotropic effects of the genes, and is a strength of the integrative approach [87]. Although we did not identify specific candidate regions that achieved the genomewide threshold of statistical significance, our study provides candidate genes for follow-up evaluation. Future GWAS studies with additional smoking behavioral measures, including nicotine dependence measures, the planned sharing of data across large consortia with increased sample size [88] and the functional analysis of individual SNPs [89], will be required to achieve the necessary power and specificity to understand SNP with low effects (OR <1.3), effects in subgroups, explore effect modification by demographic variables, and dissect pleiotropy.

Supporting Information

Table S1

| Found at: doi:10.1371/journal.pone.0004653.s001 (0.04 MB XLS) |

References

43. Lou XY, Ma JZ, Payne TJ, Beuten J, Crew KM, et al. (2006) Gene-based analysis suggests association of the nicotinic acetylcholine receptor beta1 subunit (CHRNB1) and M1 muscarinic acetylcholine receptor (CHRM1) with vulnerability for nicotine dependence. Hum Genet 120: 381–389.