Published in final edited form as:
Hum Genet. 2011 February ; 129(2): 177–188. doi:10.1007/s00439-010-0911-7.

Uncovering hidden variance: pair-wise SNP analysis accounts for additional variance in nicotine dependence

Robert C. Culverhouse,
Division of General Medical Sciences, Department of Medicine, Washington University, Saint Louis, MO 63110, USA rculverh@wustl.edu

Division of Biostatistics, Washington University, Saint Louis, MO 63110, USA

Nancy L. Saccone,
Department of Genetics, Washington University, Saint Louis, MO 63110, USA

Jerry A. Stitzel,
Department of Integrative Physiology, University of Colorado, Boulder, CO 80309, USA

Jen C. Wang,
Department of Psychiatry, Washington University, Saint Louis, MO 63110, USA

Joseph H. Steinbach,
Department of Anesthesiology Basic Science Research, Washington University, Saint Louis, MO 63110, USA

Alison M. Goate,
Department of Genetics, Washington University, Saint Louis, MO 63110, USA

Department of Psychiatry, Washington University, Saint Louis, MO 63110, USA

Tae-Hwi Schwantes-An,
Department of Genetics, Washington University, Saint Louis, MO 63110, USA

Richard A. Grucza,
Department of Psychiatry, Washington University, Saint Louis, MO 63110, USA

Victoria L. Stevens, and
Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, GA 30329, USA

Laura J. Bierut
Department of Psychiatry, Washington University, Saint Louis, MO 63110, USA

Abstract

Results from genome-wide association studies of complex traits account for only a modest proportion of the trait variance predicted to be due to genetics. We hypothesize that joint analysis of polymorphisms may account for more variance. We evaluated this hypothesis on a case–control

Electronic supplementary material The online version of this article (doi:10.1007/s00439-010-0911-7) contains supplementary material, which is available to authorized users.

Conflict of interest Drs. LJ Bierut, AM Goate, and JC Wang are listed as inventors on a patent "Markers for Addiction" (US 20070258898) covering the use of certain SNPs, including rs16969968, in determining the diagnosis, prognosis, and treatment of addiction. Dr. N Saccone is the spouse of S Saccone, who is also listed on the above patent. Dr. Bierut acted as a consultant for Pfizer, Inc. in 2008.

smoking phenotype by examining pairs of nicotinic receptor single-nucleotide polymorphisms (SNPs) using the Restricted Partition Method (RPM) on data from the Collaborative Genetic Study of Nicotine Dependence (COGEND). We found evidence of joint effects that increase explained variance. Four signals identified in COGEND were testable in independent American Cancer Society (ACS) data, and three of the four signals replicated. Our results highlight two important lessons: joint effects that increase the explained variance are not limited to loci displaying substantial main effects, and joint effects need not display a significant interaction term in a logistic regression model. These results suggest that the joint analyses of variants may indeed account for part of the genetic variance left unexplained by single SNP analyses. Methodologies that limit analyses of joint effects to variants that demonstrate association in single SNP analyses, or require a significant interaction term, will likely miss important joint effects.

Introduction

A key challenge for genetic analysis today is to account for the bulk of the phenotypic variance in complex traits attributable to genetic factors. Traditional genetic analysis methods exploit the possibility that variation in a single genetic locus may result in detectable effects on a phenotype. These methods typically examine genotypes one at a time, and build additive (or log-additive) effects models from them. The approach has been successful, with genome-wide association studies (GWAS) alone identifying over 500 genetic variants contributing to disease (Hindorff et al. 2009). However, for many complex traits (e.g. obesity, smoking, diabetes), the variants identified by studies with large samples and dense genome-wide geno-typing account for only a modest fraction of the phenotypic variance estimated to be attributable to genetic contributions (Goldstein 2009; Hirschhorn 2009; Kraft and Hunter 2009). A variety of factors have been suggested to account for the "missing variance", including rare variants, variants not surveyed by current GWAS chips, structural variants (e.g. copy number variants such as insertion/deletions or copy neutral variation such as inversions and translocations), population heterogeneity, gene–gene interactions, and gene-environment interactions (Galvan et al. 2010; Manolio et al. 2009).

While all of these likely have some role in the "missing" genetic contribution to complex diseases, examining datasets for joint effects (gene–gene or gene–environment) may prove essential in order to understand genetic associations to complex phenotypes, in spite of the computational and analytical challenges. Biology provides well-documented examples of epistasis playing an important role in phenotypes ranging from gross morphology to longevity to efficiency of reproduction (Anholt et al. 2003; Gerke et al. 2009; Mackay 2010; Vieira et al. 2000; Wolf et al. 2005). Further, many important traits of medical interest (such as heart disease, hypertension, diabetes, cancer, and infection) arise from biological systems controlled by interacting genetic factors (Churchill et al. 2004; Lander and Schork 1994; Phillips 2008; Routman and Cheverud 1995; Schork 1997; Szathmary et al. 2001).

There are a variety of definitions for epistasis, but a common hallmark is that the effect of one genetic locus differs depending on the genetic background. In such cases, modeling genetic loci jointly may explain more of the trait variance than a model that includes only main effects. In the most extreme case, it may turn out that the importance of a genetic variant may be revealed as significant only in a joint analysis (Culverhouse et al. 2002).

In this study, we have taken a systematic approach to examine joint effects of variants marking the cholinergic nicotinic receptor subunit genes on smoking behavior. Nicotine dependence is a heritable behavior, with estimates of approximately 50% for the genetic contribution to phenotypic variance (Li et al. 2003; Sullivan and Kendler 1999). The $\alpha_5\alpha_3\beta_4$ cholinergic nicotinic receptor cluster on chromosome 15 is now unequivocally associated with nicotine dependence and the number of cigarettes smoked per day (Amos et al. 2008; Berrettini et al. 2008; Bierut et al. 2007; Bierut et al. 2008; Liu et al. 2010; Saccone et al. 2009b; Saccone et al. 2007; Stevens et al. 2008; Thorgeirsson et al. 2010; Tobacco and Genetics Consortium 2010; Weiss et al. 2008). For example, a recent report of a meta-analysis with 76,972 subjects (Thorgeirsson et al. 2010) achieves a p value 2.4×10^{-69} for a SNP in this cluster. Other cholinergic nicotinic receptors play a more modest role (Saccone et al. 2010b; Schlaepfer et al. 2008; Thorgeirsson et al. 2010).

Our goals in this study are to investigate the following questions:

1. Can examining variants in the cholinergic nicotinic receptor subunit genes jointly using the restricted partition method (RPM) account for more of the variance in nicotine dependence than univariate analyses?
2. Are joint effects detected by the RPM also detected by logistic regression, a traditional statistical approach?
3. Do the joint effects replicate in an independent sample?
4. How much does the explained variance increase when the joint effects are included?

Materials and methods

Collaborative Genetic Study of Nicotine Dependence data

Individuals from the Collaborative Genetic Study of Nicotine Dependence (COGEND) used in these analyses were of European descent (N = 2,062). Recruitment of subjects started with community-based telephone-screening of approximately 50,000 people aged 25–44. The screening identified individuals who had smoked at least 100 cigarettes in their lifetime (smokers). The smokers then answered further questions about their smoking behavior, including the Fagerström Test for Nicotine Dependence (FTND) (Heatherton et al. 1991). This assessment identified nicotine dependent subjects who were current smokers with a score of 4 or higher on the FTND (cases). Non-dependent control subjects smoked at least 100 cigarettes, but were required to have a lifetime FTND score of 0 or 1 even during the heaviest period of smoking. Nicotine dependent cases and non-dependent controls who smoked were invited to participate in this genetic study of smoking. COGEND recruited subjects from three urban areas in the USA: St. Louis (N = 1,434), Detroit (N = 581), and Minneapolis (N = 47). The COGEND study was carried out with approval from the appropriate institutional review boards and with informed consent from all participants.

The SNPs examined in this study represent variants in 16 cholinergic nicotinic receptor genes, with comprehensive coverage of the $\alpha_5\alpha_3\beta_4$ cluster on chromosome 15. The COGEND data contain genotypes for 210 SNPs in cholinergic nicotinic receptor genes. Before analysis, genotypic data were thinned using a pair-wise threshold of $r^2 < 0.95$ based on linkage disequilibrium in COGEND to reduce the number of redundant pair-wise tests. This resulted in an analysis dataset of 127 SNPs. The SNPs were not thinned further because a goal of these analyses was to be sensitive to epistasis or other multi-locus effects that might display little to no marginal effects, and whose effects might not be adequately captured using tags with lower correlations.

All genotype data have undergone extensive quality control measures (Bierut et al. 2007; Saccone et al. 2009a). Self-reported race was verified with an EIGENSTRAT (Price et al. 2006) analysis including HapMap data. EIGENSTRAT analysis also found no significant population differences between sites. Map positions and annotations were obtained from the Genome Reference Consortium Homo sapiens assembly GRCh37.

American Cancer Society data

A subset of individuals from the American Cancer Society (ACS) Cancer Prevention Study-II Nutrition cohort (*N* ~ 184,000) was used as a replication dataset. DNA was collected from 109,380 individuals from this cohort. Smoking behavior information was obtained from questionnaires administered in 1982, 1992, and 1997. Based on the responses to these questionnaires, a case/control phenotype (heavy smokers/light smokers) was developed as a proxy for nicotine dependence and non-dependence (Stevens et al. 2008). Inclusion in the ACS dataset used for our analysis required subjects to report smoking at least 100 cigarettes lifetime. Cases (heavy smokers) also needed to report in at least one interview smoking at least 30 cigarettes per day for at least 5 consecutive years. Controls (light smokers), in addition to smoking more than 100 cigarettes lifetime, smoked for at least 1 year, but endorsed the lightest smoking rate available on each of the three questionnaires (<5 cigarettes/day in 1982 and 1992, <10 cigarettes/day in 1997). After ascertainment for heavy and light smoking, 1,500 heavy smokers and 1,500 light smokers were selected for genotyping. Of these, 2,847 smokers (1,452 heavy smokers and 1,395 light smokers) passed genotyping quality control and were included in this study. Both current and former smokers were included in the analysis.

SNPs genotyped in the ACS subjects included those in the neuronal nicotinic acetylcholine receptor genes with $p < 0.05$ in the initial univariate analysis of the COGEND study of nicotine dependence (Bierut et al. 2007) and additional haplotype-tagging SNPs chosen using the method of Gabriel et al. (2002). After filtering for genotype quality control, the ACS data used in our analyses contain a total of 61 cholinergic nicotinic receptor subunit SNPs (a subset of the 210 COGEND cholinergic nicotinic receptor subunit SNPs). Detailed descriptions of the ACS sample ascertainment and genotyping quality control are described in Stevens et al. (2008).

Summary information about age and sex in the two datasets is given in Table 1.

Joint analysis

A variety of methods can be used to investigate how multiple genetic variants contribute jointly to a dichotomous trait. We chose to use the restricted partition method (RPM), a statistical approach designed to identify combinations of qualitative genetic and environmental factors (e.g. genotypes, categorical or dichotomized environmental exposures) contributing to a quantitative or binary trait (Culverhouse et al. 2007; Culverhouse et al. 2004) and logistic regression modeling, a more traditional statistical approach. The RPM is agnostic regarding a specific genetic model (e.g. additive, dominant, or recessive in the case of a single SNP) and was specifically designed to be sensitive even if the contribution from a combination of factors is predominantly presented as an interaction displaying little or no marginal effects. It is an exploratory method that uses the data to determine the number of distinct risk classes, with the aim of determining if modeling predictors jointly accounts for more of the variation in phenotype than summing their individual effects. It does not specifically identify interactions or test them for significance.

Intuitively, the RPM is a multi-locus ‘measured genotype’ (Boerwinkle et al. 1986) approach, assessing the mean trait values or prevalence information for different multi-locus genotypes. Genotypes (groups) are iteratively merged if a multiple comparisons test indicates that they are not significantly different. The algorithm halts when either all genotypes are merged (indicating no significant association between these genotypes and the phenotype) or when the remaining groups are significantly different from each other. In this case, the resulting model is evaluated for the proportion of trait variance explained by the groups. Significance of the full model (not just the interaction) is evaluated using permutation tests because the goal is to identify contributors to phenotypic variation, not to identify or test interactions. The benefits of focusing on the full model, rather than on specific interactions, have been demonstrated by extensive evaluation in a variety of simulated models (Marchini et al. 2005). We note that if the RPM model explains none of the variance (i.e. all the genotypes are merged into a single group), then the *p* value is necessarily = 1, as no permuted dataset could result in a lower explained variance. A detailed description of the RPM can be found in Culverhouse et al. (2004).

Application of the RPM to the COGEND sample (Hypothesis generation)

Applying the RPM to the dichotomous nicotine dependence phenotype among smokers in COGEND allowed the development of hypotheses that joint effects of specific genetic loci (tagged by SNPs genotyped in COGEND) increased the variance explained. To minimize the chance of reporting the same signal multiple times, we report two pair-wise results as independent only if at least one SNP from the first pair is not in linkage disequilibrium ($r^2 <$ 0.4) with either of the SNPs in the second pair.

Univariate and pair-wise RPM analyses were performed for the 127 nicotinic receptor SNPs that passed quality control for the combined COGEND data. Because we were analyzing a binary trait, the explained variance, V_E, was defined by $V_E = 1 - \frac{\sum N_i p_i q_i}{N_{pq}}$ where *N* is the total sample size, *p* is the proportion of cases, and *q* is the proportion of controls, with the corresponding subscripted variables defined similarly for the resulting RPM model groups. We note that this is always less than or equal to the variance accounted for by the full model (i.e. all cells kept separate). Thus, the merging of cells by the RPM method provides some protection from over-fitting and the corresponding inflation of estimated explained variance. Sampling variability and sparseness can be expected to result in “noise” in the estimates of explained phenotypic variance. A test of the RPM using real genotypes with simulated phenotypes (Culverhouse et al. 2009) suggests that pair-wise results accounting for less than 0.5% of the trait variance may have an increased risk of being false positives. Because of this, we restricted our 2-SNP signals of interest to those explaining more than 1.0% of the trait variance.

Hypothesis testing

We took two approaches to further examine the joint effects identified by the RPM in the COGEND data—testing using another analytic method, logistic regression, and replication in an independent dataset.

Logistic regression testing—All the top SNP-pairs identified by the RPM were evaluated using logistic regression to determine if the signals could be identified by this alternative analytic method. The SNPs were treated as class variables. Results from univariate models and joint models with and without an interaction term were compared. Logistic regression analyses were performed using SAS® software version 9.1 (SAS Institute Inc., Cary, NC).

Replication testing—Replication using independent data is the gold standard. For this, we used the ACS data that had the heavy-smoking/light-smoking phenotype as proxy for the FTND derived nicotine dependence/non-dependence phenotype from COGEND. Our criteria for deciding that an RPM result is recapitulated in the independent dataset involve two factors. First, the joint RPM model should account for substantially more of the trait variance than the sum of the univariate RPM models in the independent dataset (we used an arbitrary threshold of 15% increase or greater). The second condition we require for replication is that the RPM model produced in the replication data should be consistent with the RPM model in the initial dataset. By this we mean that, in general, genotypes that are of high risk in the initial model should be high risk in the replication, and the same for low risk genotypes. We note that although this second replication condition is heuristic, (1) it is a higher dimensional analog for the routine procedure of checking univariate results to ensure that allelic effect is in the same direction in the initial and replication datasets, and (2) it provides an additional way to reject replication and cannot increase the number of false positives. Because only a limited number of the COGEND SNPs were genotyped in the ACS data, this testing for replication was only available for a few signals identified in the COGEND data. For one of the COGEND SNPs not genotyped in the ACS data, rs3743075, linkage disequilibrium information from both COGEND and HapMap suggests that rs514743 is an appropriate surrogate (r^2 = 0.97 in COGEND data, r^2 = 0.92 in the HapMap release 22 CEU data). The use of rs514743 as a proxy for rs3743075 allowed us to evaluate one additional signal from the COGEND data in the ACS data.

Results

Univariate results

A complete list of the 127 SNPs used in this study, along with the results of the univariate RPM analyses on the hypothesis generating data (COGEND) can be found in Table S1 (supplementary information). The phenotype was case–control status for nicotine dependence. The table lists the nicotine receptor gene cluster associated with each SNP, the chromosome, the base-pair location, the proportion of variance attributable to the SNP, and an empirical *p* value for association. The permutation-based empirical *p* values resulted from a maximum of 100,000 permutations of the phenotype for each SNP. The predictors were the SNPs treated as 3-level categorical variables. The results are consistent with the primary analyses (univariate) from the first-stage analysis of the COGEND data using logistic regression methods (Bierut et al. 2007;Saccone et al. 2009a).

Pair-wise results

There were a total of 8,001 SNP-pairs analyzed in the hypothesis generating dataset (COGEND). We present our top results for each of two rankings: the first based on total explained variance, the second based on increase in explained variance by the joint analysis of the pair compared to the sum of the univariate effects. Table 2 reports the top six results in terms of total variance explained. For each result, we list the chromosomes, gene clusters, and rs numbers for the two SNPs representing the signal. This information is followed by the percentage of the phenotypic variance accounted for by each of the SNPs under univariate RPM analyses and the variance explained under RPM analyses combining both SNPs in the model. All of the pair-wise RPM models in this table had empirical *p* values of $\leq 10^{-6}$ (based on results from 1,000,000 permutations each; 10 times as many as were used for the univariate estimates).

Each of the joint effect signals in Table 2 involves the SNP rs16969968. This SNP provides the strongest signal in the univariate RPM analysis of the COGEND data, accounting for 1.22% of the trait variance. Because selection focused on maximizing explained variance, it is not surprising to find rs16969968 in each of the signals. We note that RPM models for two of the secondary SNPs accounted for none of the trait variance in the single SNP analyses; however, when jointly analyzed with rs16969968, additional trait variance is explained.

The RPM results are followed in Table 2 by the results from a logistic regression analysis. The first column under “Logistic Regression results” contains the *p* values associated with the 2 degree of freedom (df) univariate model for each of the two SNPs. The second column contains the *p* value for the 4 df interaction term in the model containing both SNPs and the interaction variable. Univariate RPM and logistic regression results are consistent (cf supplementary Table S1): the variant rs16969968, which explains the most variance in the univariate RPM analysis, is the most statistically significant SNP in the logistic regression analysis; and the SNPs that explain none of the variance in the RPM analysis all have *p* values >0.2 in the univariate logistic regression analysis. The first reported signal, involving rs16969968 and rs2133965 (which account for nearly 90% more of the variance when analyzed jointly), shows a significant interaction *p* value ($p = 0.002$) using logistic regression. None of the other logistic regression interaction terms in this table were significant.

Four of the signals listed in Table 2 were testable in the ACS data, and details of the replication analysis are summarized in Table 3. The format is similar to that of Table 2. Our criteria for replication are that the RPM model accounts for at least a 15% increase in explained variance and that the RPM models produced in the two datasets should be consistent. Three of the four signals replicate. All of these signals displayed >20% increase in explained variance, well above our threshold for replication. For three of the four, the RPM models in the ACS data were similar to those found for the COGEND data, meeting our criteria for replication. The fourth RPM model was not sufficiently similar to the one proposed in the COGEND data for us to consider it to be a replication.

The variant rs1696998, though still the most significant SNP in the ACS data, only accounts for 0.87% of the trait variance under a univariate RPM analysis. All of the secondary SNPs account for none of the variance under univariate RPM analysis of these data. As in the COGEND data, the univariate *p* values are consistent between the RPM and logistic regression analyses, and the interaction terms were not significant under the logistic regression analysis (although the final pair, rs16969968 and rs514743, displayed a nominal *p* value of 0.05 for the interaction term).

Our second ranking of signals, based on increased explained variance under an RPM analysis, was aimed at identifying pairs with joint effects much larger than the sum of the individual effects. We note that this is a filtered ranking because of our requirement that any reported pair-wise signal explain at least 1% of the phenotypic variance. The RPM and logistic regression results for this set of pairs are summarized in Table 4. Once again, all of the pair-wise RPM models had empirical *p* values of no more than 10^{-6} (based on results from 1,000,000 permutations each). Because this ranking was designed to find pairs with large synergistic effects, it is perhaps not surprising that the logistic regression analysis found the interaction terms for each of these signals to be nominally significant, with uncorrected *p* values ranging from 0.04 to 2.2×10^{-5}. None of these pairs were testable in the independent ACS data because in each case, at least one of the SNPs was not genotyped in the ACS data and no good surrogate could be identified.

To evaluate how well we succeeded in increasing the explained variance in phenotype, we evaluated two logistic regression models for nicotine dependence in the COGEND dataset: one with interactions and one without. Predictors included the 11 SNPs from the 3 two-locus models that were validated in the ACS data and the 4 two-locus models that had logistic regression interaction term *p* values <0.01: rs16969968, rs2133965, rs3787138, rs13277524, rs3743075 from Table 2 and rs2292977, rs17483548, rs667282, rs680244, rs1500948, rs2611603 from Table 4. The logistic regression model without interactions accounted for 5.15% of the trait variance and resulted in a *c* statistic of 0.622. In contrast, the model including the same predictors, but adding the 7 pair-wise interaction terms listed in the tables, accounted for 9.49% of the trait variance and resulted in a c statistic of 0.662.

Discussion

Our goal was to determine if examining genetic variants jointly could account for substantially more of the variance in phenotype than would be estimated by summing the results of univariate analyses. We identified several variants that jointly explain more phenotypic variance than they do individually, and three of the four tested signals replicate in independent data. Our results highlight two important lessons: (1) joint effects need to be investigated across all SNPs, not just those with main effects, and (2) it is not sufficient to look at the significance of an interaction term in a logistic regression model to determine if there is a joint effect.

The prime illustration for both of these points is the joint effect of the loci tagged by rs16969968 and rs3743075: as illustrated in Table 2, univariate analysis of the COGEND data indicates that the genetic locus tagged by rs16969968 displays a highly significant association to nicotine dependence, and the second locus, tagged by rs3743075, does not display evidence of association to nicotine dependence under univariate analyses. However, when the two genetic loci are analyzed jointly using the RPM, this pair accounts for much more of the variance in the dichotomous nicotine dependence/non-dependence pheno-type than the sum of the univariate effects. This same pattern of results is seen in the ACS data: univariate analysis finds that the locus tagged by rs16969968 displays a substantial univariate effect on the dichotomous heavy versus light smoking phenotype, and the second locus is not associated with the phenotype in a univariate analysis. Again, a joint RPM analysis of the two SNPs accounts for considerably more of the phenotype variance than the sum of the univariate results. Although analysis by logistic regression also displayed the increase in explained variance (results not shown), the interaction term was not significant. As a result, we would have missed this joint effect had we required a significant interaction coefficient.

The case for these two loci having a synergistic joint effect has been greatly strengthened by results very recently reported by the Consortium for the Genetic Analysis of Smoking Phenotypes (CGASP) (Saccone et al. 2010a). The CGASP subjects consist of current and former smokers of European ancestry ($N = 38{,}617$), including COGEND and ACS subjects. A case/control phenotype of heavy versus light smoking was analyzed. In univariate analysis, the locus tagged by rs16969968 was highly associated with the heavy smoking phenotype ($p = 5.96 \times 10^{-31}$), while the second locus, tagged by rs588765, was considerably less significant ($p = 4.54 \times 10^{-4}$). When the two loci were analyzed jointly, each became more significant: $p = 3.52 \times 10^{-36}$ for the locus tagged by rs16969968, and $p = 6.03 \times 10^{-9}$ (passing the threshold for genome-wide significance) for the locus tagged by rs588765. The logistic regression interaction term in this analysis was not significant. The explanation for these results is that the risk alleles for the two loci are negatively correlated. As a result, in a univariate analysis, the association of rs16969968 to nicotine dependence is present, but dampened, while the effect of the second locus is almost completely masked until jointly analyzed.

Although we cannot know how often this kind of joint effect might occur, we find this result, displayed in COGEND data, ACS data, and the larger CGASP data, a strong argument that tests for interactions not be limited to loci with substantial main effects, and that the consideration of joint effects should not be limited to ones giving rise to a statistically significant interaction term. It is an advantage of the RPM, focused on identifying sources of trait variance and not specifically interactions, that it identifies such pairs in an automated fashion, without the need to examine each model individually (an impractical approach when evaluating many thousands or millions of multi-locus models).

Although a replicable joint effect does not require a significant logistic regression interaction term, it appears that if the proportional increase in explained variance is great enough (e.g. over 80%), logistic regression can detect a significant interaction term. The first SNP pair in Table 2, rs16969968 and rs2133965, with nearly 90% increase in explained variance, was the only pair from this table for which the logistic regression interaction term was associated with a p value <0.05. In contrast, all five pairs from Table 4, which focused on large proportional increases in explained variance, had interaction p values <0.05.

The variants rs16969968, a non-synonymous coding SNP in *CHRNA5*, and rs1051730, a synonymous SNP, are of particular interest because these SNPs are strongly associated with nicotine dependence and smoking behaviors (Amos et al. 2008; Berrettini et al. 2008; Liu et al. 2010; Thorgeirsson et al. 2008; Thorgeirsson et al. 2010; Tobacco and Genetics Consortium 2010). These two SNPs are highly correlated ($r^2 = 0.991$) in our data, and we included rs16969968 in our list of 127 to be tested. We selected this SNP because it results in an amino acid change, and in vitro studies demonstrate that it alters receptor function (Bierut et al. 2008). Because SNPs were thinned to include only those with $r^2 < 0.95$, rs1051730 was not included in our primary analyses. We confirmed, with secondary univariate and joint analyses using rs1051730, that the results were very similar to those for rs16969968. (An analog of Table 2, replacing rs16969968 throughout with rs1061730, can be found in Supplemental Table S2.) We cannot say which, if either, of these two SNPs is biologically linked to smoking behavior. We can only say that they are representatives of a cluster of variants in tight linkage disequilibrium that display these effects. Further research in the lab will be required to identify the causative variants definitively.

We focused on cholinergic nicotinic receptor subunit genes because the protein products physically combine to form biologically active receptors that bind nicotine. For example, the $\alpha4$ and $\alpha5$ nicotinic acetylcholine receptor (nAChR) subunits combine with the nAChR $\beta2$ subunit to form an $\alpha4_2\beta2_2\alpha5_1$ receptor that is expressed in various brain regions including the mesolimbic reward pathway (McClure-Begley et al. 2009; Salminen et al. 2004; Zoli et al. 2002). Thus, it is plausible that variants in and around the genes that form these subunits may alter the subunit makeup of nicotinic receptors or the relative expression of the receptors. Of course, joint effects for these genes need not be limited to changes that alter an amino acid, but can include variants that alter splicing, mRNA expression, stability, or other regulatory factors. Our analytic evidence of joint effects can suggest models that can be tested biologically in the laboratory.

We have presented evidence that examining factors jointly, including SNPs displaying little to no univariate association to the phenotype, may uncover interactions and other synergistic effects that account for a sizable portion of the “missing” genetic variance remaining after univariate analyses of well-powered GWAS (Goldstein 2009; Hirschhorn 2009; Kraft and Hunter 2009). Although this study on the genetics of nicotine dependence focused on only a small portion of the genome, we found that examining nicotinic receptor variants in a pair-wise manner increased the proportion of variation explained. The challenge with such an approach, particularly if applied to data with 1,000,000 genotyped polymorphisms, is the number of tests. Even if limited to all pair-wise models, such analyses would be computationally expensive and statistically intractable. The problem becomes more extreme if analyses involve more than two factors at a time.

Two approaches for addressing this problem have been applied in this study. One approach to address both the computational and statistical problem of the large number of tests is to filter the polymorphisms to be analyzed based on biological plausibility. In this case, we limited our analyses to polymorphisms located in or in linkage dis-equilibrium with nicotinic receptor subunit genes and filtered so that no two SNPs in our analysis had r^2 >

0.95. Other filtering approaches could make use of additional biological information such as evolutionary conservation, biological pathways, and results from previous studies.

An approach to address the statistical problem of the large number of tests is to use a second, independent dataset. Although filtering may provide considerable help with the computation challenges, the number of tests will likely remain very large and the effects to be detected relatively modest. As a result, the issue of statistical significance will likely remain even after judicious a priori filtering of predictors. One way to address this is to use the initial data for hypothesis generation and an independent dataset to test a small number of hypotheses. If the key biological contributors to the phenotype are identified in the initial data, we would expect that the signal would replicate in an ethnically matched, independent replication sample.

The focus of this study on pair-wise analyses is one of its limitations. The reasons for this choice include limiting the large number of statistical tests performed, the theoretical results suggesting that factors not detectible with pair-wise analyses are unlikely to display large effects (Culverhouse et al. 2002), and the fact that our hypothesis-generating sample was of modest size for examining joint effects. We recognize that higher-order joint effects with three or more variants may play important roles in the etiology of complex disease, but larger datasets are needed to evaluate such complex multi-SNP effects. A second limitation of the study is that we were not able to test all of the identified signals in independent data.

This study was also limited to two analytic methods for examining joint genetic effects. A variety of methods are available to investigate how multiple genetic variants contribute jointly to a dichotomous trait. In addition to the RPM and the traditional logistic regression, there are other partition-based methods [e.g. the Combinatorial Partition Method (CPM) (Nelson et al. 2001), multifactor dimensionality reduction (MDR) (Hahn et al. 2003), the Generalized MDR (GMDR) (Lou et al. 2007)], as well as information theory based methods [e.g. k-way interaction information (KWII) and total correlation information (TCI) (Chanda et al. 2007)], and a variety of other approaches. Because these approaches each have their advantages for detecting joint effects, an analysis using multiple methods can identify additional signals that might have been missed by a single method, as well as provide increased confidence in signals that are identified by multiple methods.

The results of this study support three ideas which we believe are important for future studies. First, this approach identified several variants that have minimal evidence of association when tested individually, but which have larger effects when examined in combination with other genetic predictors. Second, requiring a significant interaction term in logistic regression testing may not detect variants that jointly explain more of the variance. Finally, increases in explained variance can be obtained by including only a small number of interactions.

We note that as more researchers begin specifically looking for joint effects, the reports of interactions have increased. This is particularly true for genetic studies of smoking. In addition to the CGASP result mentioned above, several other studies have reported interesting results from the joint analysis of multiple loci chr15q25 in samples of European descent (Li et al. 2010a; Liu et al. 2010; Thorgeirsson et al. 2010; Tobacco and Genetics Consortium 2010), African Americans (Li et al. 2010a), and Koreans (Li et al. 2010b). Additional joint genetic effects related to smoking have been reported in other candidate regions, including interactions among variants in GABBR1 and GABBR2 related to nicotine dependence in African and European Americans (Li et al. 2009) and an interaction between CYP2A6 and MAOA affecting smoking in Chinese (Tang et al. 2009).

As researchers struggle to make the best use of the wealth of genetic data now available, a key goal is to account for more of the phenotypic variance that is expected to be attributable to genetics. The results of this study suggest that pair-wise examination of genetic data can be a useful tool for achieving this goal, uncovering substantial additional genetic contribution to phenotypic variance through joint effects of multiple loci.

Web resources

EIGENSTRAT software: http://genepath.med/harvard.edu/~reich/EIGENSTRAT.htm. R software: http://www.r-project.org/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

In memory of Theodore Reich, founding Principal Investigator of COGEND, we are indebted to his leadership in the establishment and nurturing of COGEND and acknowledge with great admiration his seminal scientific contributions to the field. Lead investigators directing data collection are Laura Bierut, Naomi Breslau, Dorothy Hatsukami, and Eric Johnson. The authors thank Heidi Kromrei and Tracey Richmond for their assistance in data collection and Brian Suarez for his thoughtful comments on the manuscript. This work was supported by National Institutes of Health grants K25 GM69590 from the National Institute of General Medical Sciences, R03 DA023166 from the National Institute on Drug Abuse (NIDA), and P01 CA89392 from the National Cancer Institute, as well as IRG-58-010-50 from the American Cancer Society. Geno-typing services for this project were provided by Perlegen Sciences under NIDA Contract HHSN271200477471C. Additional genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096. Phenotypic and genotypic data are stored in the NIDA Center for Genetic Studies (NCGS) at http://zork.wustl.edu/ under NIDA Contract HHSN271200477451C (PIs J Tischfield and J Rice).

References

Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 2008;40:616–622. [PubMed: 18385676]

Anholt RR, Dilda CL, Chang S, Fanara JJ, Kulkarni NH, Ganguly I, Rollmann SM, Kamdar KP, Mackay TF. The genetic architecture of odor-guided behavior in Drosophila: epistasis and the transcriptome. Nat Genet 2003;35:180–184. [PubMed: 12958599]

Berrettini W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H, Waterworth D, Muglia P, Mooser V. Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. Mol Psychiatry 2008;13:368–373. [PubMed: 18227835]

Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, Swan GE, Rutter J, Bertelsen S, Fox L, Fugman D, Goate AM, Hinrichs AL, Konvicka K, Martin NG, Montgomery GW, Saccone NL, Saccone SF, Wang JC, Chase GA, Rice JP, Ballinger DG. Novel genes identified in a high-density genome wide association study for nicotine dependence. Hum Mol Genet 2007;16:24–35. [PubMed: 17158188]

Bierut LJ, Stitzel JA, Wang JC, Hinrichs AL, Grucza RA, Xuei X, Saccone NL, Saccone SF, Bertelsen S, Fox L, Horton WJ, Breslau N, Budde J, Cloninger CR, Dick DM, Foroud T, Hatsukami D, Hesselbrock V, Johnson EO, Kramer J, Kuperman S, Madden PA, Mayo K, Nurnberger J Jr, Pomerleau O, Porjesz B, Reyes O, Schuckit M, Swan G, Tischfield JA, Edenberg HJ, Rice JP, Goate AM. Variants in nicotinic receptors and risk for nicotine dependence. Am J Psychiatry 2008;165:1163–1171. [PubMed: 18519524]

Boerwinkle E, Chakraborty R, Sing CF. The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann Hum Genet 1986;50:181–194. [PubMed: 3435047]

Chanda P, Zhang A, Brazeau D, Sucheston L, Freudenheim JL, Ambrosone C, Ramanathan M. Information-theoretic metrics for visualizing gene-environment interactions. Am J Hum Genet 2007;81:939–963. [PubMed: 17924337]

Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W, Bleich A, Bogue M, Broman KW, Buck KJ, Buckler E, Burmeister M, Chesler EJ, Cheverud JM, Clapcote S, Cook MN, Cox RD, Crabbe JC, Crusio WE, Darvasi A, Deschepper CF, Doerge RW, Farber CR, Forejt J, Gaile D, Garlow SJ, Geiger H, Gershenfeld H, Gordon T, Gu J, Gu W, de Haan G, Hayes NL, Heller C, Himmelbauer H, Hitzemann R, Hunter K, Hsu HC, Iraqi FA, Ivandic B, Jacob HJ, Jansen RC, Jepsen KJ, Johnson DK, Johnson TE, Kempermann G, Kendziorski C, Kotb M, Kooy RF, Llamas B, Lammert F, Lassalle JM, Lowenstein PR, Lu L, Lusis A, Manly KF, Marcucio R, Matthews D, Medrano JF, Miller DR, Mittleman G, Mock BA, Mogil JS, Montagutelli X, Morahan G, Morris DG, Mott R, Nadeau JH, Nagase H, Nowakowski RS, O'Hara BF, Osadchuk AV, Page GP, Paigen B, Paigen K, Palmer AA, Pan HJ, Peltonen-Palotie L, Peirce J, Pomp D, Pravenec M, Prows DR, Qi Z, Reeves RH, Roder J, Rosen GD, Schadt EE, Schalkwyk LC, Seltzer Z, Shimomura K, Shou S, Sillanpaa MJ, Siracusa LD, Snoeck HW, Spearow JL, Svenson K, et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 2004;36:1133–1137. [PubMed: 15514660]

Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 2002;70:461–471. [PubMed: 11791213]

Culverhouse R, Klein T, Shannon W. Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 2004;27:141–152. [PubMed: 15305330]

Culverhouse R, Hinrichs AL, Jin CH, Suarez BK. Gene x gene and gene x environment interactions for complex disorders. BMC Proc 2007;1(Suppl 1):S72. [PubMed: 18466574]

Culverhouse R, Jin W, Jin CH, Hinrichs A, Suarez BK. Power and false positive rates for the Restricted Partition Method (RPM) in a large candidate gene dataset. BMC Proc 2009;3(Suppl 7):S74. [PubMed: 20018069]

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. Science 2002;296:2225–2229. [PubMed: 12029063]

Galvan A, Ioannidis JP, Dragani TA. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. Trends Genet 2010;26:132–141. [PubMed: 20106545]

Gerke J, Lorenz K, Cohen B. Genetic interactions between transcription factors cause natural variation in yeast. Science 2009;323:498–501. [PubMed: 19164747]

Goldstein DB. Common genetic variation and human traits. N Engl J Med 2009;360:1696–1698. [PubMed: 19369660]

Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 2003;19:376–382. [PubMed: 12584123]

Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom KO. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. Br J Addict 1991;86:1119–1127. [PubMed: 1932883]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009;106:9362–9367. [PubMed: 19474294]

Hirschhorn JN. Genomewide association studies—illuminating biologic pathways. N Engl J Med 2009;360:1699–1701. [PubMed: 19369661]

Kraft P, Hunter DJ. Genetic risk prediction—are we there yet? N Engl J Med 2009;360:1701–1703. [PubMed: 19369656]

Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994;265:2037–2048. [PubMed: 8091226]

Li MD, Cheng R, Ma JZ, Swan GE. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. Addiction 2003;98:23–31. [PubMed: 12492752]

Li MD, Mangold JE, Seneviratne C, Chen GB, Ma JZ, Lou XY, Payne TJ. Association and interaction analyses of GABBR1 and GABBR2 with nicotine dependence in European- and African-American populations. PLoS One 2009;4:e7055. [PubMed: 19763258]

Li MD, Xu Q, Lou XY, Payne TJ, Niu T, Ma JZ. Association and interaction analysis of variants in CHRNA5/CHRNA3/CHRNB4 gene cluster with nicotine dependence in African and European Americans. Am J Med Genet B Neuropsychiatr Genet 2010a;153B:745–756. [PubMed: 19859904]

Li MD, Yoon D, Lee JY, Han BG, Niu T, Payne TJ, Ma JZ, Park T. Associations of variants in CHRNA5/A3/B4 gene cluster with smoking behaviors in a Korean population. PLoS One 2010b; 5(8):e12183. [PubMed: 20808433]

Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, Berrettini W, Knouff CW, Yuan X, Waeber G, Vollenweider P, Preisig M, Wareham NJ, Zhao JH, Loos RJ, Barroso I, Khaw KT, Grundy S, Barter P, Mahley R, Kesaniemi A, McPherson R, Vincent JB, Strauss J, Kennedy JL, Farmer A, McGuffin P, Day R, Matthews K, Bakke P, Gulsvik A, Lucae S, Ising M, Brueckl T, Horstmann S, Wichmann HE, Rawal R, Dahmen N, Lamina C, Polasek O, Zgaga L, Huffman J, Campbell S, Kooner J, Chambers JC, Burnett MS, Devaney JM, Pichard AD, Kent KM, Satler L, Lindsay JM, Waksman R, Epstein S, Wilson JF, Wild SH, Campbell H, Vitart V, Reilly MP, Li M, Qu L, Wilensky R, Matthai W, Hakonarson HH, Rader DJ, Franke A, Wittig M, Schafer A, Uda M, Terracciano A, Xiao X, Busonero F, Scheet P, Schlessinger D, St Clair D, Rujescu D, Abecasis GR, Grabe HJ, Teumer A, Volzke H, Petersmann A, John U, Rudan I, Hayward C, Wright AF, Kolcic I, Wright BJ, Thompson JR, Balmforth AJ, Hall AS, Samani NJ, Anderson CA, Ahmad T, Mathew CG, Parkes M, Satsangi J, Caulfield M, Munroe PB, Farrall M, Dominiczak A, Worthington J, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 2010;42:436–440. [PubMed: 20418889]

Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet 2007;80:1125–1137. [PubMed: 17503330]

Mackay TF. Mutations and quantitative genetic variation: lessons from Drosophila. Philos Trans R Soc Lond B Biol Sci 2010;365:1229–1239. [PubMed: 20308098]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature 2009;461:747–753. [PubMed: 19812666]

Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005;37:413–417. [PubMed: 15793588]

McClure-Begley TD, King NM, Collins AC, Stitzel JA, Wehner JM, Butt CM. Acetylcholine-stimulated [3H]GABA release from mouse brain synaptosomes is modulated by alpha4beta2 and alpha4alpha5beta2 nicotinic receptor subtypes. Mol Pharmacol 2009;75:918–926. [PubMed: 19139153]

Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 2001;11:458–470. [PubMed: 11230170]

Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 2008;9:855–867. [PubMed: 18852697]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909. [PubMed: 16862161]

Routman EJ, Cheverud JM. Gene effects on a quantitative trait: two-locus epistatic effects measured at microsatellite markers and at estimated QTL. Evolution 1995;51:1654–1662.

Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau O, Swan GE, Goate AM, Rutter J, Bertelsen S, Fox L, Fugman D, Martin NG, Montgomery GW, Wang JC, Ballinger DG, Rice JP, Bierut LJ. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum Mol Genet 2007;16:36–49. [PubMed: 17135278]

Saccone NL, Saccone SF, Hinrichs AL, Stitzel JA, Duan W, Pergadia ML, Agrawal A, Breslau N, Grucza RA, Hatsukami D, Johnson EO, Madden PA, Swan GE, Wang JC, Goate AM, Rice JP, Bierut LJ. 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 Neuropsychiatr Genet 2009a;150B:453–466. [PubMed: 19259974]

Saccone NL, Wang JC, Breslau N, Johnson EO, Hatsukami D, Saccone SF, Grucza RA, Sun L, Duan W, Budde J, Culverhouse RC, Fox L, Hinrichs AL, Steinbach JH, Wu M, Rice JP, Goate AM, Bierut LJ. The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. Cancer Res 2009b; 69:6848–6856. [PubMed: 19706762]

Saccone NL, Culverhouse RC, Schwantes-An TH, Cannon DS, Chen X, Cichon S, Giegling I, Han S, Han Y, Keskitalo-Vuokko K, Kong X, Landi MT, Ma JZ, Short SE, Stephens SH, Stevens VL, Sun L, Wang Y, Wenzlaff AS, Aggen SH, Breslau N, Broderick P, Chatterjee N, Chen J, Heath AC, Heliovaara M, Hoft NR, Hunter DJ, Jensen MK, Martin NG, Montgomery GW, Niu T, Payne TJ, Peltonen L, Pergadia ML, Rice JP, Sherva R, Spitz MR, Sun J, Wang JC, Weiss RB, Wheeler W, Witt SH, Yang BZ, Caporaso NE, Ehringer MA, Eisen T, Gapstur SM, Gelernter J, Houlston R, Kaprio J, Kendler KS, Kraft P, Leppert MF, Li MD, Madden PA, Nothen MM, Pillai S, Rietschel M, Rujescu D, Schwartz A, Amos CI, Bierut LJ. Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. PLoS Genet 2010a;6(8):e1001053. [PubMed: 20700436]

Saccone NL, Schwantes-An TH, Wang JC, Grucza RA, Breslau N, Hatsukami D, Johnson EO, Rice JP, Goate AM, Bierut LJ. Multiple cholinergic nicotinic receptor genes affect nicotine dependence risk in African and European Americans. Genes Brain Behav 2010b;6:741–750.

Salminen O, Murphy KL, McIntosh JM, Drago J, Marks MJ, Collins AC, Grady SR. Subunit composition and pharmacology of two classes of striatal presynaptic nicotinic acetylcholine receptors mediating dopamine release in mice. Mol Pharmacol 2004;65:1526–1535. [PubMed: 15155845]

Schlaepfer IR, Hoft NR, Ehringer MA. The genetic components of alcohol and nicotine co-addiction: from genes to behavior. Curr Drug Abuse Rev 2008;1:124–134. [PubMed: 19492010]

Schork NJ. Genetics of complex disease: approaches, problems, and solutions. Am J Respir Crit Care Med 1997;156:S103–S109. [PubMed: 9351588]

Stevens VL, Bierut LJ, Talbot JT, Wang JC, Sun J, Hinrichs AL, Thun MJ, Goate A, Calle EE. Nicotinic receptor gene variants influence susceptibility to heavy smoking. Cancer Epidemiol Biomarkers Prev 2008;17:3517–3525. [PubMed: 19029397]

Sullivan PF, Kendler KS. The genetic epidemiology of smoking. Nicotine Tob Res 1999;1(Suppl 2):S51–S57. discussion S69–70. [PubMed: 11768187]

Szathmary E, Jordan F, Pal C. Molecular biology and evolution. Can genes explain biological complexity? Science 2001;292:1315–1316. [PubMed: 11360989]

Tang X, Guo S, Sun H, Song X, Jiang Z, Sheng L, Zhou D, Hu Y, Chen D. Gene-gene interactions of CYP2A6 and MAOA polymorphisms on smoking behavior in Chinese male population. Pharmacogenet Genomics 2009;19:345–352. [PubMed: 19415821]

Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, Stacey SN, Bergthorsson JT, Thorlacius S, Gudmundsson J, Jonsson T, Jakobsdottir M, Saemundsdottir J, Olafsdottir O, Gudmundsson LJ, Bjornsdottir G, Kristjansson K, Skuladottir H, Isaksson HJ, Gudbjartsson T, Jones GT, Mueller T, Gottsater A, Flex A, Aben KK, de Vegt F, Mulders PF, Isla D, Vidal MJ, Asin L, Saez B, Murillo L, Blondal T, Kolbeinsson H, Stefansson JG, Hansdottir I, Runarsdottir V, Pola R, Lindblad B, van Rij AM, Dieplinger B, Haltmayer M, Mayordomo JI, Kiemeney LA, Matthiasson SE, Oskarsson H, Tyrfingsson T, Gudbjartsson DF, Gulcher JR, Jonsson S, Thorsteinsdottir U, Kong A, Stefansson K. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452:638–642. [PubMed: 18385739]

Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, Sulem P, Rafnar T, Esko T, Walter S, Gieger C, Rawal R, Mangino M, Prokopenko I, Magi R, Keskitalo K, Gudjonsdottir IH, Gretarsdottir S, Stefansson H, Thompson JR, Aulchenko YS, Nelis M, Aben KK, den Heijer M, Dirksen A, Ashraf H, Soranzo N, Valdes AM, Steves C, Uitterlinden AG, Hofman A, Tonjes

A, Kovacs P, Hottenga JJ, Willemsen G, Vogelzangs N, Doring A, Dahmen N, Nitz B, Pergadia ML, Saez B, De Diego V, Lezcano V, Garcia-Prats MD, Ripatti S, Perola M, Kettunen J, Hartikainen AL, Pouta A, Laitinen J, Isohanni M, Huei-Yi S, Allen M, Krestyaninova M, Hall AS, Jones GT, van Rij AM, Mueller T, Dieplinger B, Haltmayer M, Jonsson S, Matthiasson SE, Oskarsson H, Tyrfingsson T, Kiemeney LA, Mayordomo JI, Lindholt JS, Pedersen JH, Franklin WA, Wolf H, Montgomery GW, Heath AC, Martin NG, Madden PA, Giegling I, Rujescu D, Jarvelin MR, Salomaa V, Stumvoll M, Spector TD, Wichmann HE, Metspalu A, Samani NJ, Penninx BW, Oostra BA, Boomsma DI, Tiemeier H, van Duijn CM, Kaprio J, Gulcher JR, McCarthy MI, Peltonen L, Thorsteinsdottir U, Stefansson K. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet 2010;42:448–453. [PubMed: 20418888]

Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet 2010;42:441–447. [PubMed: 20418890]

Vieira C, Pasyukova EG, Zeng ZB, Hackett JB, Lyman RF, Mackay TF. Genotype-environment interaction for quantitative trait loci affecting life span in Drosophila melanogaster. Genetics 2000;154:213–227. [PubMed: 10628982]

Weiss RB, Baker TB, Cannon DS, von Niederhausern A, Dunn DM, Matsunami N, Singh NA, Baird L, Coon H, McMahon WM, Piper ME, Fiore MC, Scholand MB, Connett JE, Kanner RE, Gahring LC, Rogers SW, Hoidal JR, Leppert MF. A candidate gene approach identifies the CHRNA5-A3-B4 region as a risk factor for age-dependent nicotine addiction. PLoS Genet 2008;4:e1000125. [PubMed: 18618000]

Wolf JB, Leamy LJ, Routman EJ, Cheverud JM. Epistatic pleiotropy and the genetic architecture of covariation within early and late-developing skull trait complexes in mice. Genetics 2005;171:683–694. [PubMed: 16020793]

Zoli M, Moretti M, Zanardi A, McIntosh JM, Clementi F, Gotti C. Identification of the nicotinic receptor subtypes expressed on dopaminergic terminals in the rat striatum. J Neurosci 2002;22:8785–8789. [PubMed: 12388584]

Table 1

Age and sex information for the COGEND and ACS cases and controls

	COGEND				ACS			
	N	**% Female**	**Mean age**	**Age range**	***N***	**% Female**	**Mean age**	**Age range**
Cases	1,063	45.3	36.9	23–15	1,452	43.3	69.6	49–88
Controls	999	61.5	35.9	25–15	1,395	62.7	70.2	53–90

Table 2

Independent SNP pairs jointly accounting for the most phenotypic variance in RPM analyses of nicotine dependence case–control status in the hypothesis-generating COGEND data (N = 2,062)

CHR	Gene	SNP	COGEND RPM results		COGEND logistic regression results	
			% variance explained	% increase in explained variance[a]	Univariate *p* value for SNP	*p* value of interaction
15	CHRNA5	rs16969968	1.22		<0.0001	
15	CHRNA7	rs2133965	0.0		0.25	
		Combined	2.31	89		0.0020
15	CHRNA5	rs16969968	1.22		<0.0001	
20	CHRNA4	rs3787138	0.55		0.0037	
		Combined	2.06	16		0.40
15	CHRNA5	rs16969968	1.22		<0.0001	
8	CHRNB3	rs13277524	0.46		0.0009	
		Combined	1.95	16		0.55
15	CHRNA5	rs16969968	1.22		<0.0001	
2	CHRND	rs2767	0.24		0.08	
		Combined	1.95	34		0.17
15	CHRNA5	rs16969968	1.22		<0.0001	
11	CHRNA10	rs2231532	0.24		0.06	
		Combined	1.88	29		0.13
15	CHRNA5	rs16969968	1.22		<0.0001	
15	CHRNA3	rs3743075	0.00		0.36	
		Combined	1.83	50		0.27

[a] Percentage increase in the model containing both SNPs compared to the sum of the univariate effects

Table 3

Results in the replication data (ACS) for four signals identified in the initial data (COGEND) (N = 2,844)

CHR	Gene	SNP	ACS RPM results		ACS logistic regression results		RPM result replicates in ACS?
			% variance explained	% increase in explained variance[a]	Univariate *p* value for SNP	*p* value of interaction	
15	CHRNA5	rs16969968	0.87		<0.0001		
20	CHRNA4	rs3787138	0.00		0.12		
		Combined	1.14	31		0.61	Yes
15	CHRNA5	rs16969968	0.87		<0.0001		
8	CHRNB3	rs13277254	0.00		0.15		
		Combined	1.15	32		0.15	Yes
15	CHRNA5	rs16969968	0.87		<0.0001		
2	CHRND	rs2767	0.00		0.16		
		Combined	1.06	22		0.90	No
15	CHRNA5	rs16969968	0.87		<0.0001		
15	CHRNA5	rs514743[b]	0.00		0.89		
		Combined	1.22	40		0.05	Yes

[a] Percentage increase in the model containing both SNPs compared to the sum of the univariate effects

[b] rs3743075 was not genotyped in the ACS data. rs514743 is highly correlated with rs3743075 ($r^2 > 0.9$) and was used as a surrogate for the validation

Table 4

SNP pairs jointly accounting for more than 1% of the total variance and increasing the explained variance by at least 100% over the sum of the individual effects in RPM analyses of nicotine dependence case–control status in the hypothesis-generating COGEND data

CHR	Gene	SNP	COGEND RPM results		COGEND logistic regression results	
			% variance explained	% increase in explained variance[a]	Univariate *p* value for SNP	*p* value of interaction
8	CHRNA2	rs2292977	0.00		0.72	
15	IREB2	rs17483548	0.78		<0.0001	
		Combined	1.57	101		0.0015
15	CHRNA5	rs667282	0.68		0.0005	
15	CHRNA5	rs680244	0.00		0.18	
		Combined	1.52	124		0.0011
15	CHRNA7	rs1500948	0.00		0.80	
15	CHRNA7	rs2611603	0.00		0.29	
		Combined	1.38	Inf		<0.0001
2	CHRNG	rs1881492	0.53		0.0044	
8	CHRNA2	rs2472553	0.00		0.17	
		Combined	1.17	121		0.014
15	CHRNA7	rs2337980	0.0		0.38	
20	CHRNA4	rs2236196	0.52		0.0019	
		Combined	1.17	126		0.044

These SNP-pairs were not testable in ACS

[a] Percentage increase in the model containing both SNPs compared to the sum of the univariate effects