Determination of Genetic Mechanisms Underlying Smoking Addiction

Jiekun (Jackie) Yang
Taigu, Shanxi, China

B.S., Renmin University of China, 2009
M.A., University of Wisconsin-Madison, 2011

A Dissertation presented to the Graduate Faculty of the
University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Biochemistry and Molecular Genetics

University of Virginia
May, 2016

To my parents.

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

– Marie Curie

Acknowledgements

My deepest thanks go to my mentor, Dr. Ming Li. He brought me in the human genetics field, and guided my research presented in this dissertation. His enthusiasm about science and sense of duty as a mentor inspire me to work hard and grow as a scientist. I am deeply grateful for all the help and support I have received from Dr. Li.

I also owe much to Drs. Tanseli Nesil, Junran Cao, and Shaolin Wang from the Li Lab for making great contributions to my research, and at the same time being great friends. I have learned a lot about the field from different angles because of their distinct expertise.

Thanks also to Drs. Stefan Bekiranov, Charles Farber, Steve Rich, and Michael Timko for the opportunity to have valuable scientific discussions at the committee meetings. Their insights made Chapters 4 and 5 further enhanced.

My sincere thanks also go to our collaborators, Drs. Jennie Ma, Thomas Payne from the University of Mississippi Medical Center, and David Goldman from NIAAA of NIH. Without their efforts, we could not obtain the precious samples and data for our studies.

I would also thank Drs. Joel Hockensmith, Wendy Lynch, Ms. Debbie Sites, and Ms. Sandra Foster, who took care of all the administrative problems, so I could focus on my research.

Finally, I would like to thank my family and friends for being extremely supportive. The tons of love I received from them give me courage and confidence to pursue my dream of becoming a scientist.

Abstract

Smoking poses significant threats to public health. Despite 50 years of prevention efforts, smoking remains the greatest cause of preventable diseases and deaths. Even though today's users smoke fewer cigarettes than those 50 years ago, they are at higher risk of developing lung cancer because of changes in cigarettes. Our group and others have shown strong evidence for the involvement of genetics in nicotine dependence (ND), with an average heritability of 0.56. This dissertation contributes to our understanding of the genetic structure of smoking from four perspectives. In the first study, leveraging computational efficiency of the GPU-based Generalized Multifactor Dimensionality Reduction (GMDR-GPU) program, we detected variants in genes encoding the 5-$HT3_{AB}$ receptors (*HTR3A* and *HTR3B*) and the serotonin transporter (*SLC6A4*) interactively affecting etiology of alcohol, cocaine, and nicotine dependence, although their individual effect was weak. In the second study, targeted next-generation sequencing was used to discover rare variants from ND candidate genes. Although none of the genotyped common variants showed significant association with different smoking measures, the weighted sum statistic (WSS) and combined sum test results indicated that rare variants alone or combined with common variants in a subset of candidate genes contribute significantly to the risk of ND. In the third study, we developed an ND genetic susceptibility map based on the results obtained by the approaches commonly used in recent years, which include genome-wide linkage, candidate gene association, GWAS, and targeted sequencing studies. Converging and diverging results from these empirical approaches have elucidated a preliminary genetic

architecture of this intractable psychiatric disorder and yielded new hypotheses on ND etiology. In the final study, *cis*-expression and methylation quantitative trait loci (eQTL and mQTL) were mapped for the candidate genes ascertained in the third study using human brain tissues. Among the one eQTL and two mQTLs determined, for the first time we showed that the minor allele of one variation significantly decreased methylation levels at one gene, reduced expression levels of another, and lowered percentage of smokers all in a dominant way for the same cohort. The studies presented in this dissertation provide a variety of novel insights into the genetic mechanisms of smoking addiction; and perhaps more importantly, new ideas and methods to study other complex traits/diseases are generated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of genetic studies on smoking

Smoking kills more than 6 million people annually worldwide.[1] It causes more than 480,000 deaths each year, about one of every five deaths in the United States.[2] Although it has been shown that smoking can damage every part of the body and leads to cancers and chronic diseases, in 2014, an estimated 40 million adults in the United States still smoke cigarettes.[3] Nicotine dependence (ND), with an average heritability of 0.56,[4] is the primary factor maintaining smoking behavior.[5]

Systematic genetic investigation on smoking started from genome-wide linkage studies. After reporting the first genome-wide linkage scan in the African American (AA) family[6] and then in the European American (EA) family samples,[7] our group analyzed all reported linkage peaks for various ND assessments, such as smoking quantity, Fagerström Test for Nicotine Dependence (FTND), or habitual smoking from more than 20 studies in 13 samples. Following

the rigorous criteria of Lander and Kruglyak,[8] we nominated 13 regions, located on chromosomes 3–7, 9–11, 17, 20, and 22, as having "significant" or "suggestive" linkage for ND in at least two independent samples.[9] Although these studies showed the locations of risk loci for ND, no relevant risk variants/genes were identified from most of these linkage peaks at that time.

Along with technological advances, by following up on the identified linkage peaks and biological functions underlying ND, variants within candidate genes were examined for association with smoking. For example, our group performed linkage-based candidate gene association studies within the detected linkage peaks on chromosomes 9, 11, and 17 and found that *GABBR2*[10] and *NTRK2*[11] on chromosome 9, *ARRB1* on chromosome 11, and *ARRB2* on chromosome 17[12] are significantly associated with ND in at least one ethnic sample (AA or EA). Concurrently, we performed biology-based candidate gene association studies and demonstrated that *CHAT*,[13] *CHRNB2*,[14] *COMT*,[15] *DRD1*,[16] *DRD3*,[17] and *TAS2R38*[18] are significantly associated with ND in our AA or EA samples or both.

Additionally, genome-wide association studies (GWASs) were employed to identify risk variants for ND. The most replicable and significant GWAS finding on ND is the association of variants in the *CHRNA5*/*A3*/*B4* cluster on chromosome 15 with ND[19-30] and lung cancer.[31, 32] Significant association of variants in *CHRNB3*/*A6* and *CYP2A6*/*B6* with ND were also reported in GWASs.[24, 33, 34] However, although many genes were reported to be associated with ND through linkage- and biology-based association studies, only a few reached genome-wide significance or were detected in GWAS on ND. And the few GWAS triumphs stand in contrast to the limited heritability they explain; e.g., the most significant single nucleotide polymorphism (SNP) in

CHRNA3 accounted for only 0.5% of the variance in cigarettes smoked per day (CPD) in a meta-analysis of 73,853 subjects.[23] The "missing heritability" issue emerged, and researchers suggested that many factors may result in it.[35] This was the basic situation of the ND genetics field, when I joined the research group.

1.2 Dissertation rationale

This dissertation attempts to probe answers for three important questions of the field. First, what factors may contribute to "missing heritability" of ND? As Zuk *et al*.[36] specified, the proportion of heritability explained by a set of variants is the ratio of (*i*) the heritability due to these variants (numerator), estimated directly from their observed effects, to (*ii*) the total heritability (denominator), inferred directly from population data. And they showed that a substantial portion of missing heritability could arise from overestimation of the denominator, specifically, estimates of total heritability implicitly assume the trait involves no genetic interactions (epistasis).

However, epistasis does exist for ND, because previous studies from our group implicated genes encoding the 5-$HT3_{AB}$ receptors (*HTR3A* and *HTR3B*) and the serotonin transporter (*SLC6A4*) interactively in alcohol, cocaine, and nicotine dependence. Further, we were wondering if epistasis among the three genes remains in subjects with multiple addictions (comorbidity), which is a common phenomenon in the addiction field, and what the interaction models look like if epistasis is detected. **Chapter 2** approaches this problem by using the Study of Addiction: Genetics and Environment (SAGE) data, and discovers significant interaction models, where most of the SNPs included for each addictive phenotype are either overlapped

or in high linkage disequilibrium for both AA and EA samples.[37] While **Chapter 2** explores the "missing heritability" issue from the denominator side by extending one example of epistasis from subjects with single addiction to multiple addictions, **Chapter 3** focuses on the numerator to discover new susceptibility variants.

By that time, because efforts have largely focused on common genetic variants, one hypothesis is that much of the "missing heritability" is due to rare genetic variants.[38] Although it is also recognized that rare alleles are more likely to be deleterious and are represented disproportionately among disease alleles,[39] studies on rare variants in ND have been limited mostly to nAChR subunit genes.[40-43] Thus, in **Chapter 3**, we concentrate on a group of 30 genes implicated in ND and other addictions.[44] After a targeted sequencing of the genes followed by association analysis of common and rare variants in our family and case-control samples, we find that rare variants alone or combined with common variants in a subset of biological candidate genes contribute substantially to the risk of ND. Thus effects of rare variants are missing from the numerator of heritability explained. **Chapters 2** and **3** collectively illustrate that epistasis and rare variants contribute to the "missing heritability" issue of ND.

Second, what have we known about ND genetics and what will be our next step? For the past decade, experimental approaches for ND genetics, along with technological advancements and studies on other complex diseases/traits, have evolved, from genome-wide linkage study to candidate gene association study, and from GWAS to targeted sequencing. However, we do not know whether the discoveries from all of these approaches consistent with one another or not, and if we should focus on results obtained from "newer" approaches; e.g., GWAS, and abandon findings from "older" ones, such as genome-wide linkage study, in the literature sea. In **Chapter**

4, we address these issues by developing a ND genetic susceptibility map based on results obtained by the approaches commonly used in recent years as mentioned above. Converging and diverging results from these empirical approaches have elucidated a preliminary genetic architecture of this intractable psychiatric disorder and yielded new hypotheses on ND etiology. More importantly, the insights we obtained by putting together results from diverse approaches can be applied to other complex diseases/traits.[45]

Third, what are the mechanistic steps between genetic variation and ND? Although quite a few susceptibility genes have been identified to influence smoking, as shown in **Chapter 4**, the mechanistic steps between genetic variation and smoking-related traits are generally not understood. With the rise of massively parallel sequencing technologies, multiple layers of gene regulation, including chromatin states and transcription factor (TF) binding footprints, profiles or different epigenetic marks, and posttranscriptional modifications, have been characterized, which enable us to probe regulatory variants' control of transcriptional processes through multiple aspects of gene regulation.[46, 47] Additionally, most identified variants within the ND genetic loci are located in noncoding regions based on association study results. Thereby in **Chapter 5**, for the first time, we link genetic variation, DNA methylation, mRNA expression, and smoking status together using the same participants, which depicts a regulatory mechanism from *cis*-methylation quantitative trait loci (mQTL) to phenotypic manifestation of smoking. Moreover, different regulatory effect patterns of low-frequency and common variants on mRNA expression and DNA methylation, respectively, are observed. Experiments to test and verify causal effects among these layers of regulation are warranted.

The unifying thread of this dissertation is answering the three key questions to gain new insights into genetic mechanisms underlying smoking addiction. Indeed, the four chapters (**Chapters 2** to **5**) contribute to our understanding of the complex genetic structure for ND.

Chapter 2

Interaction among serotonin receptor and transporter genes

2.1 Abstract

Previous studies have implicated genes encoding the 5-$HT3_{AB}$ receptors (*HTR3A* and *HTR3B*) and the serotonin transporter (*SLC6A4*), both independently and interactively, in alcohol (AD), cocaine (CD), and nicotine dependence (ND). However, whether these genetic effects also exist in subjects with comorbidities remains largely unknown. We used 1,136 African-American (AA) and 2,428 European-American (EA) subjects from the Study of Addiction: Genetics and Environment (SAGE) to determine associations between 88 genotyped or imputed variants within *HTR3A*, *HTR3B,* and *SLC6A4* and three types of addictions, which were measured by DSM-IV diagnoses of AD, CD, and ND and the Fagerström Test for Nicotine Dependence (FTND), an independent measure of ND commonly used in tobacco research. Individual SNP-based association analysis revealed a significant association of rs2066713 in *SLC6A4* with FTND in AA

(β = -1.39; *P* = 1.6E-04). Haplotype-based association analysis found one major haplotype formed by SNPs rs3891484 and rs3758987 in *HTR3B* that was significantly associated with AD in the AA sample, and another major haplotype T-T-G, formed by SNPs rs7118530, rs12221649, and rs2085421 in *HTR3A*, which showed significant association with FTND in the EA sample. Considering the biologic roles of the three genes and their functional relations, we used the GPU-based Generalized Multifactor Dimensionality Reduction (GMDR-GPU) program to test SNP-by-SNP interactions within the three genes and discovered two- to five-variant models that have significant impacts on AD, CD, ND, or FTND. Interestingly, most of the SNPs included in the genetic interaction model(s) for each addictive phenotype are either overlapped or in high linkage disequilibrium for both AA and EA samples, suggesting these detected variants in *HTR3A*, *HTR3B,* and *SLC6A4* are interactively contributing to etiology of the three addictive phenotypes examined in this study.

2.2 Introduction

Serotonin (5-hydroxytryptamine; 5-HT) is a neurotransmitter that mediates rapid excitatory responses through ligand-gated channels (5-HT3 receptors). The 5-HT3 receptors, unlike other serotonergic receptor classes, which are G protein-coupled,[48-50] belong to the superfamily of nicotinic acetylcholine (nACh), subtype A of the γ-aminobutyric acid ($GABA_A$) and glycine receptors.[51] The serotonin-gated ion channel conducts primarily Na^+ and K^+, resulting in rapid neuronal depolarization followed by a rapid desensitization and the release of stored neurotransmitter, which suggests a potentially important role for this receptor system in neuronal circuitry involved in drug abuse.[52] Further, 5-HT3 receptors are co-localized with nACh

receptors on nerve terminals in several brain pathways of reward processing, including dopaminergic terminals in the striatum.[53] Although there is no evidence that they interact physically, cross-regulation may take place at a downstream molecular level.[53-55] Besides the potentially important role of 5-HT3 receptors in the development of nicotine dependence (ND), they can be potentiated through acute exposure to alcohol at concentrations that produce intoxication.[56, 57]

Whereas 5-HT3 receptors assembled by 5-$HT3_A$ subunits are uniformly located in various parts of the central and peripheral nervous systems, transcripts of the 5-$HT3_A$ and 5-$HT3_B$ subunits are coexpressed in the amygdala, caudate, and hippocampus, areas implicated in alcohol, nicotine, and other drug addictions, and form pharmacologically more potent heteropentameric receptors compared with the 5-$HT3_A$ homomeric structures.[58-60] The genes encoding the 5-$HT3_A$ and 5-$HT3_B$ receptor subunits (namely, *HTR3A* and *HTR3B*) lie in a 90-kb region on chromosome 11q23.1.[61]

Serotonin transporters (SERTs), one major class of monoamine transporters, which regulate the availability of 5-HT in the synaptic cleft through re-uptake, is encoded by the *SLC6A4* gene on chromosome 17q11.2.[62] *SLC6A4* spans 37.8 Kb and is composed of fourteen exons encoding a protein of 630 amino acids.[63] Alternate promoters in combination with differential splicing involving exon 1A, B, and C in specific tissues, and alternate polyadenylation site usage resulting in multiple mRNA species are likely participants in the regulation of SERT expression in humans.[64, 65] SERTs mediate antidepressant action and behavioral effects of cocaine and amphetamines.[62] Sequence variations in *SLC6A4* have been associated with several

neuropsychiatric conditions, including major depressive disorders, anxiety-related personality traits, and antidepressant response.[66-68]

In addition, previous association studies have posited a significant role for *HTR3A*, *HTR3B,* and *SLC6A4* in AD,[60, 69] cocaine dependence (CD),[60] and ND,[70] both independently and through gene-by-gene interactions. Importantly, both studies reported by Seneviratne *et al*.[69] and Yang *et al*.[70] indicated significant interactive effects of genetic variations in *HTR3A*, *HTR3B*, and *SLC6A4* in influencing the etiology of AD and ND, even though both individual SNP- and haplotype-based association analyses revealed only weak association of variants in the three genes with AD and ND. Our group has also reported that a combined five-marker genotype panel in *HTR3A*, *HTR3B* and *SLC6A4* can be used to predict the outcome of treatment of alcohol dependence with the 5-HT3 antagonist ondansetron.[71, 72] Thus, the objective of this study was to determine whether there exist significant interactive effects between the three genes in subjects with multiple addictions of both African- and European-American origin.

2.3 Subjects and methods

2.3.1 Subjects

SAGE is a population-based study with 4,032 subjects of either European (EA) or African American (AA) descent. Participants were selected from three large complementary datasets: the Collaborative Study on the Genetics of Alcoholism (COGA),[73] the Collaborative Genetic Study of Nicotine Dependence (COGEND),[34] and the Family Study of Cocaine Dependence (FSCD).[74] All subjects included in these studies include comprehensive demographic information such as age, sex, and ethnicity. Genotyping was performed on the Illumina Human 1M platform

with 1,040,107 SNPs available for each DNA sample. For a detailed description of this GWAS dataset, please see the paper by Bierut *et al*.[75]

According to the quality control (QC) report of the GENEVA alcohol-dependence project accompanying the dataset, stringent QC criteria were applied to all the samples. After removal of subjects with abnormal chromosomes 11 or 17 (such as aneuploidy and mosaic cell populations), related individuals, Hispanics, 3,564 (54.8% females) samples were retained for all analyses in this study. Among these samples, 2,428 (56.1% females) were EA and 1,136 (52.1% females) were AA. According to the principal component (PC) analysis results from the original study, PC1 separates the self-identified black and white subjects very well, while PC2 separates the Asian HapMap samples and the self-identified Hispanic subjects from the others; meanwhile, similar results were seen with analyses using two principal components indexing continuous variation and self-reported race as categorical variables.[75] Since Hispanic subjects were removed, self-identified racial groups were used to distinguish AA from EA in all analyses.

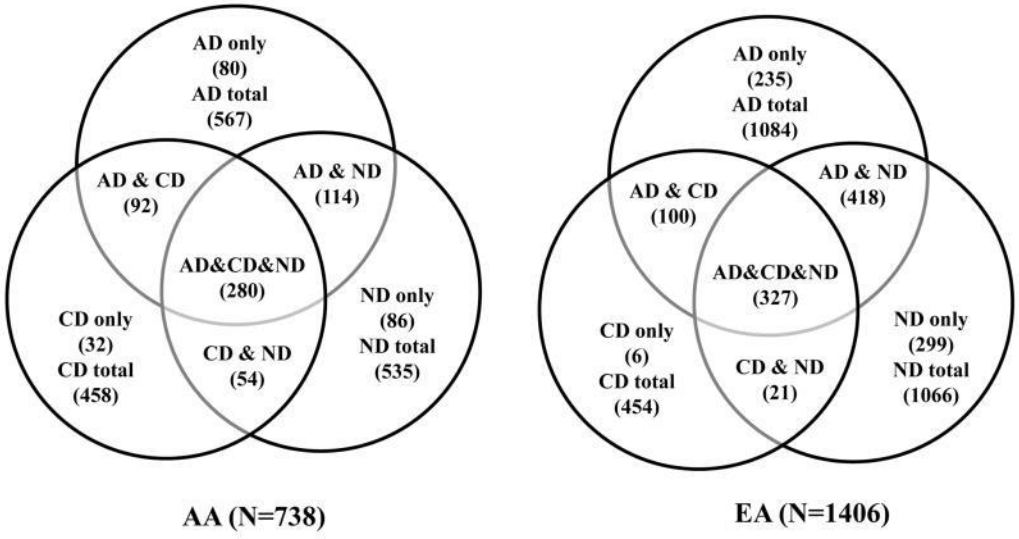
The dependence status of each subject for nicotine, alcohol, and cocaine were assessed by the DSM-IV criteria, which were obtained from the original dataset. In addition, the Fagerström Test for Nicotine Dependence (FTND) score of each subject was chosen as an independent measure of ND, because it is one of the commonly used measures in ND research, thus providing a means of comparing results from different studies.[70] The detailed characteristics of the AA and EA samples are summarized in **Table 2.1** and **Figure 2.1**.

Table 2.1: Characteristics of the SAGE AA and EA samples used in the study.

Characteristic	African-American	European-American
Sample size	1,136	2,428
Age, years (SD)	40.2 (7.4)	38.4 (9.7)
Female (%)	592 (52.1)	1,362 (56.1)
DSM-IV alcohol dependence (%)	567 (49.9)	1,084 (44.6)
DSM-IV cocaine dependence (%)	458 (40.3)	454 (18.7)
DSM-IV nicotine dependence (%)	535 (47.1)	1,066 (43.9)
FTND score (SD)	4.93 (2.32)	5.06 (2.76)
No addiction (%)	397 (34.9)	1,018 (41.9)
One type of addiction (%)	198 (17.4)	540 (22.2)
Two types of addiction (%)	260 (22.9)	539 (22.2)
Three types of addiction (%)	280 (24.6)	327 (13.5)

SD standard deviation.

Figure 2.1: *Venn diagrams* showing numbers of subjects with either sole or multiple addictions in the SAGE AA and EA samples.



Numbers in parentheses stand for sample sizes of either sole or multiple addictions. Numbers at the bottom of the figure are the total sample size for AAs and EAs, respectively. *AD* Alcohol Dependence; *CD* Cocaine Dependence; *ND* Nicotine Dependence. There are one AA and four EAs with CD missing.

2.3.2 Imputation and SNP selection

In the SAGE data, there were 27 genotyped SNPs across the *HTR3B* gene region, which included the functional SNP rs1176744 (Tyr129Ser) and the missense variant rs17116138 (Val183Ile). Of the 37 SNPs within the *HTR3A* gene region, there was a coding synonymous variant, rs1176713 (Leu465Leu). For the *SLC6A4* gene, 17 SNPs were genotyped, including rs6352 Lys605Asn, which changes an amino acid. All these SNPs follow the Hardy-Weinberg Equilibrium.

Although the 81 genotyped SNPs in SAGE well covered the three genes, in order to include as many SNPs as possible from related research papers,[60, 69, 70] we performed imputation for four SNPs in *HTR3B* and three SNPs in *HTR3A* using the 1000 Genomes AFR and EUR data as references for the AA and EA samples, respectively, with the MaCH program.[76, 77] Both reference panels were accessed through the 1000 Genomes Browser (http://browser.1000genomes.org/index.html). The r^2 values, which measure the imputation quality, for six out of the seven imputed SNPs (rs33940208 was excluded from further analysis because of low imputation quality) are > 0.8 for the EA sample. There are two SNPs (rs3758987 and rs4938056) with r^2 values between 0.7 and 0.8 for the AA sample; however, their minor allele frequencies are more than 35%, which guarantees their imputation qualities with comparatively low r^2 values.[77] A detailed list of genotyped and imputed SNPs is provided in **2.8 Supplementary Data**.

2.3.3 Statistical analysis

Individual SNP- and haplotype-based association analysis

Individual SNP-based association analyses with AD, CD, and ND were performed using logistic regression models, while FTND was analyzed using linear regression models implemented in PLINK.[78] Additive, dominant, and recessive models were all tested for each SNP, adjusted for sex, age, study (whether the subject was from COGEND, COGA, or FSCD), and two other dependence statuses that are not used as the response variable in the AA and EA samples. For example, if ND/FTND was used as the dependent variable, sex, age, study, AD, and CD were included as covariates in the logistic/linear regression model. Pair-wise linkage disequilibrium (LD) and haplotype blocks were assessed by Haploview (v. 4.2),[79, 80] and their associations with the four phenotypic measures were analyzed using Haplo Stats (v.1.6.3) through computing score statistics with the same covariates and genetic models used as the individual SNP-based association analysis.[81]

Statistically significant results for individual SNPs and major haplotypes (frequency $\geqslant$ 5%) were selected after controlling for Family Wise Error Rate (FWER) using Bonferroni correction. The three genetic models and the four phenotypic measures are highly related, with the correlation coefficient between AD and CD being 0.487, AD and ND 0.453, and CD and ND 0.352. To reduce the probability of producing false-negative results and at the same time to increase statistical power, less stringent Bonferroni-corrected *P* values were used to select significant associations, which were corrected for the number of SNPs or haplotypes, but not phenotypes or genetic models (as they are highly correlated to each other). Uncorrected *P* values are presented throughout the manuscript.

SNP-by-SNP interaction analysis of *HTR3A*, *HTR3B*, and *SLC6A4* variants

For the SNP-by-SNP interaction analysis of *HTR3A*, *HTR3B*, and *SLC6A4*, we performed exhaustive searches for two- to five-way interactions using the GMDR-GPU program,[82] which not only scales genetic and/or environmental factor numbers up to the GWAS level, but also runs much faster than the earlier version of the GMDR program[83] by employing more efficient computational implementation.[82] Similar to the association analysis described above, by taking sex, age, and two-dependence status as covariates, and one other dependence status as phenotype for the AA and EA samples, GMDR-GPU calculates a "score" statistic for each subject based on a generalized linear model under different distributions.[83] Specifically, we assumed that binary traits (AD, CD, and ND) follow a Bernoulli distribution and FTND follows a normal distribution in our gene-by-gene interaction analysis using GMDR-GPU.[82]

The best statistical SNP-by-SNP interaction model for a given order of interaction was determined by three factors: (1) the cross-validation consistency (CVC) statistics for the selected SNP combinations; (2) the prediction accuracies and the significance level or *P* value, which is determined by 10^7 permutation tests based on the observed testing accuracies; and (3) interaction analysis results of the SNP combination with all four phenotypes examined.[82] Please see the supplementary note for a detailed description of the GMDR-GPU program.

2.4 Results

2.4.1 Individual SNP-based association analysis

One SNP among the 88 variants tested for the three genes remained significant after Bonferroni correction ($P < 5.68E-04$), which is rs2066713 in *SLC6A4* with a *P* value of 1.6E-04 and beta value of -1.39 for FTND under the recessive model in the AA sample. The other seven SNPs presented

in **Table 2.2** showed marginal associations (*P* < 0.01) with AD, CD, ND, or FTND in either the AA or the EA sample. Within *HTR3B*, three SNPs were marginally associated with AD or FTND under the recessive model: rs12276717 showed marginal association (OR = 0.2; *P* = 0.005) with AD in the AA sample; and both rs1672717 and rs720396 were associated (β = -0.58; *P* = 0.004 and β = -0.53; *P* = 0.005, respectively) with FTND in the EA sample. Of the *HTR3A* SNPs, rs11214796 was marginally associated (β = 0.34; *P* = 0.01) with FTND in the AA sample under the additive model; rs1563533 showed marginal association (OR = 1.7; *P* = 0.004) with AD in AAs under the dominant model. For EAs, rs1020715 and rs2364857 were associated with CD and ND, respectively, with *P* values of 0.004 (OR = 16) under the additive model and 0.005 (OR = 0.7) under the dominant model. Among these seven marginal associations, rs1020715 is questionable given its low minor allele frequency (0.002).

Table 2.2: SNPs with *P* values < 0.01 in individual SNP association analyses with AD, CD, ND and FTND in AA and EA samples.

Sample	Gene	dbSNP ID	Minor allele (MAF)	DSM-IV						FTND	
				AD		CD		ND			
				OR	*P*	OR	*P*	OR	*P*	BETA	*P*
AA	***HTR3B***	rs12276717	C (0.148)	0.2	0.005[r]	0.8	0.273[d]	0.8	0.195[d]	0.49	0.419[r]
	HTR3A	rs11214796	T (0.403)	0.8	0.132[d]	0.8	0.304[r]	0.9	0.648[r]	0.34	0.010[a]
		rs1563533	A (0.157)	1.7	0.004[d]	0.8	0.274[d]	0.7	0.065[d]	1.13	0.041[r]
	SLC6A4	rs2066713	T (0.262)	1.2	0.128[a]	1.3	0.480[r]	0.8	0.353[r]	**-1.39**	**1.6E-04[r]**
EA	***HTR3B***	rs1672717	C (0.389)	0.9	0.447[d]	0.9	0.455[r]	1.0	0.659[r]	-0.58	0.004[r]
		rs720396	C (0.431)	1.2	0.261[r]	0.7	0.109[r]	1.0	0.906[d]	-0.53	0.005[r]
	HTR3A	rs1020715	T (0.002)	0.5	0.490[a]	16	0.004[a]	0.6	0.583[a]	-0.40	0.768[a]
		rs2364857	C (0.112)	1.2	0.184[a]	0.8	0.144[d]	0.7	0.005[d]	-0.21	0.181[a]

MAF minor allele frequency; *AD* DSM-IV alcohol dependence; *CD* DSM-IV cocaine dependence; *ND* DSM-IV nicotine dependence; *OR* odds ratio; *BETA* beta coefficient. Significant associations are given in bold. Superscripts following *P* values indicate genetic models used for analysis: *a* additive; *d* dominant; and *r* recessive. For AD, CD and ND, logistic regression models were implemented; for FTND, linear regression models were used. Sex, age, study and two out of three addiction status were used as covariates while the other one was used as dependent variable in all statistical models. See "Subjects and methods section" for details.

2.4.2 Haplotype-based association analysis

According to the haplotype block definition of Gabriel *et al*.,[80] there are 15 and 13 LD blocks in the AA and EA samples, respectively, within *HTR3A* and *HTR3B*, whereas 2 blocks were found in the *SLC6A4* region for both AA and EA samples. We used Haplo Stats to perform haplotype-based association analyses for all major haplotypes (frequency $\geqslant$ 5%) in each above-mentioned LD block with the four phenotypic measures in AA and EA samples.

In AAs, there was one major haplotype C-C, formed by SNPs rs3891484 and rs3758987, located in the 5′ region of *HTR3B* (LD block 2 in **Figure 2.2**) that was significantly associated with AD (frequency = 11.9%; $P = 0.002$) under the dominant model. This association remained significant after Bonferroni correction among the 17 major haplotypes in AAs ($P < 0.003$). Besides this haplotype, there were two haplotypes with P values < 0.01: (1) G-G-G-T-G-T-C-G-C, formed by SNPs rs17116138, rs2276307, rs11214775, rs3782025, rs1176735, rs1672717, rs17614942, rs7943062, and rs7945926 (LD block 5 within *HTR3B* in **Figure 2.2**), with a frequency of 12.6%, that was marginally associated with AD under the dominant model ($P = 0.004$) and ND under the additive model ($P = 0.009$); and (2) C-C-C-T-A, formed by SNPs rs6354, rs25528, rs2066713, rs8071667, and rs16965623 (LD block 2 of *SLC6A4* in **Figure 2.3**), with a frequency of 23.7%, that showed a marginal association with AD under the additive model ($P = 0.005$).

Figure 2.2: LD structures for *HTR3B* and *HTR3A* SNPs in the SAGE AA sample.

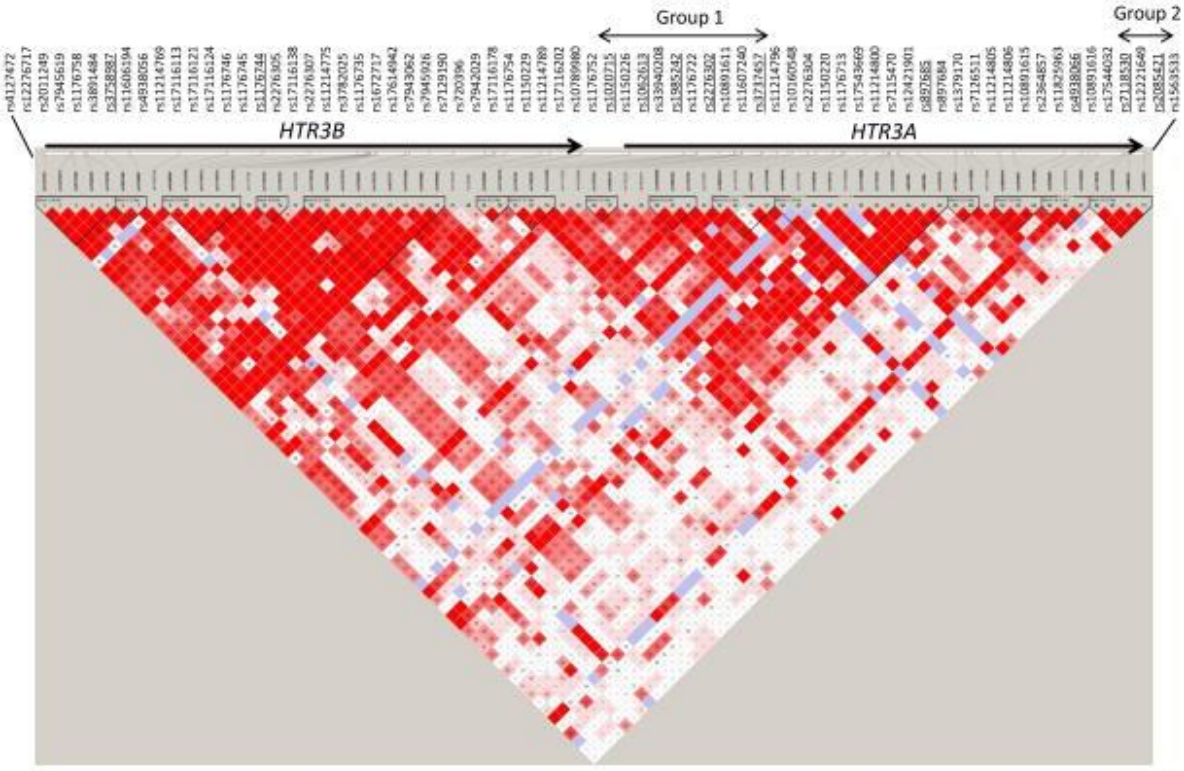
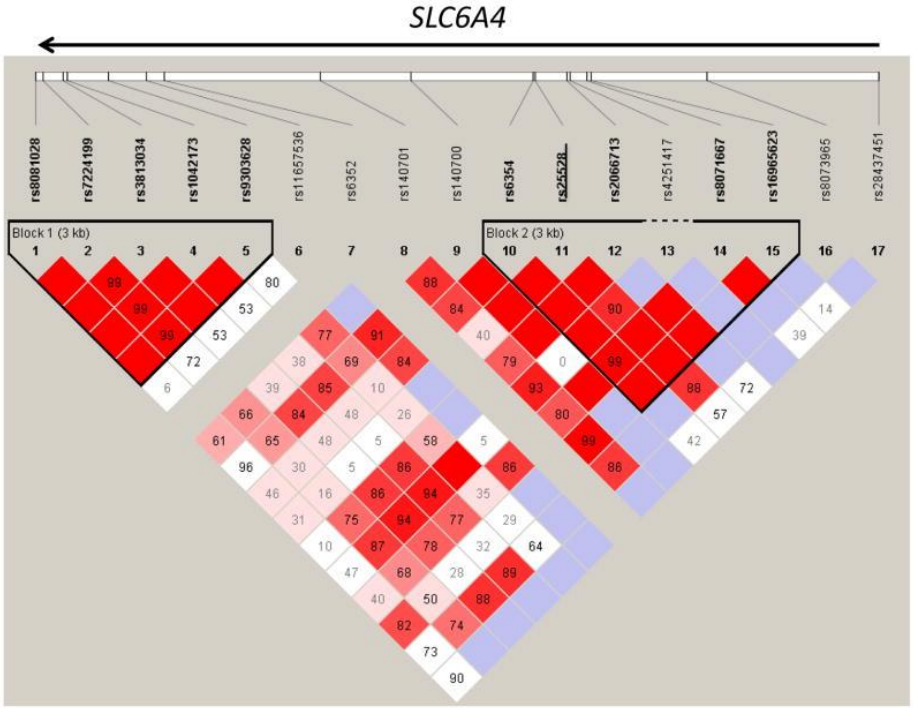


Figure 2.3: LD structure for *SLC6A4* SNPs in the SAGE AA sample.



Haploview (v. 4.2)[79] was used to calculate all *D*′ values, and haplotype blocks were defined according to Gabriel *et al*.[80] The number in *each box* represents the *D*′ value for each SNP pair surrounding that box. The *arrow* on top of the figure represents the gene transcription direction from 5'- to 3'-end. SNPs involved in later interactive models are *underlined* and grouped according to *D*′ values >0.7. Please refer to **Figure 2.6** for more information about SNP groups.

For the EA sample, we found one haplotype, T-T-G, formed by SNPs rs7118530, rs12221649, and rs2085421 (LD block 13 within *HTR3A* in **Figure 2.4**) significantly associated with FTND under the additive model (frequency = 60.5%; $P = 0.002$), which remained significant after Bonferroni correction for 15 major haplotypes ($P < 0.003$). The global P value of this haplotype was 0.008 under the recessive model, suggesting marginal association with FTND. There are three other haplotypes showing marginal significance in EAs: (1) rs11214769 and rs1176744 (LD block 2 within *HTR3B* in **Figure 2.4**) with ND (P global = 0.007) under the additive model; (2) A-G, formed by SNPs rs7942029 and rs17116178 (LD block 6 within *HTR3B* in **Figure 2.4**), with CD under the dominant model (frequency = 73%; $P = 0.007$); and (3) A-A-T-G-C, formed by SNPs rs6354, rs25528, rs2066713, rs425147, and rs8071667 (LD block 2 of *SLC6A4* in **Figure 2.5**), with FTND under the recessive model (frequency = 39.5%; $P = 0.007$). The detailed results of the haplotype-based association analyses in AAs and EAs are presented in **Tables 2.3** and **2.4**, respectively.

Figure 2.4: LD structures for *HTR3B* and *HTR3A* SNPs in the SAGE EA sample.

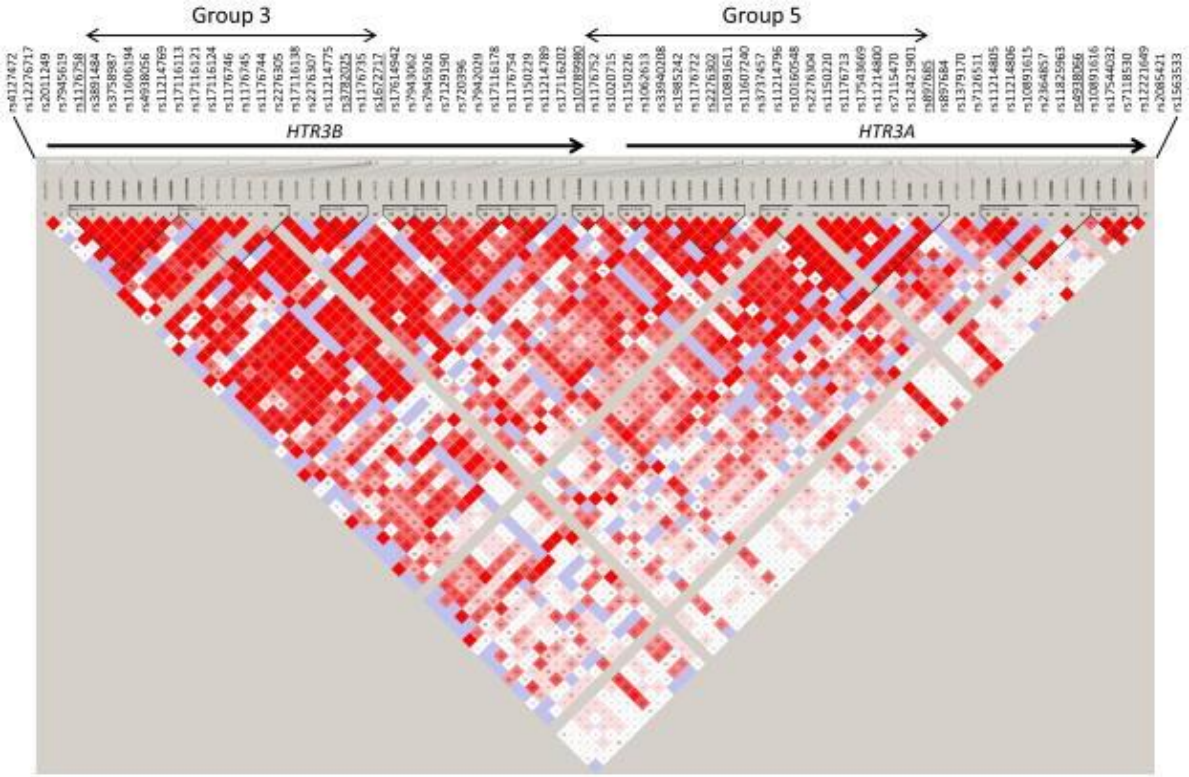
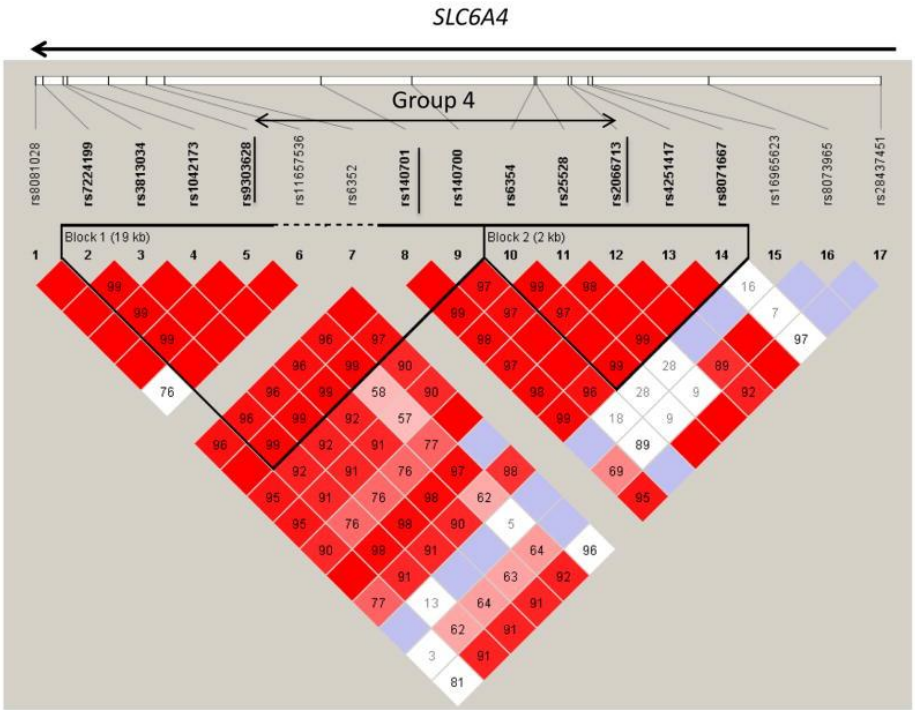


Figure 2.5: LD structure for *SLC6A4* SNPs in the SAGE EA sample.



Haploview (v. 4.2)[79] was used to calculate all *D*′ values, and haplotype blocks were defined according to Gabriel *et al*.[80] The number in *each box* represents the *D*′ value for each SNP pair surrounding that box. The *arrow* on top of the figure represents the gene transcription direction from 5'- to 3'-end. SNPs involved in later interactive models are *underlined* and grouped according to *D*′ values > 0.7. Please refer to **Figure 2.7** for more information about SNP groups.

Table 2.3: Major haplotypes (frequency ≥ 5%) associated with AD, CD, ND or FTND at $P < 0.01$ level in AA sample.

rs3891484-rs3758987 (*HTR3B*)			DSM-IV									FTND		
			AD			CD			ND					
1	2	Freq	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global
T	T	0.633	-1.028	0.304^{r}	0.012^{d}	1.795	0.073^{a}	0.083^{r}	-0.668	0.504^{d}	0.130^{d}	-1.310	0.190^{a}	0.064^{r}
T	C	0.245	-1.501	0.133^{r}		-2.360	0.018^{r}		2.111	0.035^{r}		1.885	0.059^{r}	
C	C	0.119	3.025	**0.002^{d}**		-0.341	0.733^{r}		-1.978	0.048^{a}		-1.817	0.069^{r}	

rs17116138-rs2276307-rs11214775-rs3782025-rs1176735-rs1672717-rs17614942-rs7943062-rs7945926 (*HTR3B*)										DSM-IV									FTND		
										AD			CD			ND					
1	2	3	4	5	6	7	8	9	Freq	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global
G	A	A	T	G	T	C	A	T	0.141	-1.460	0.144^{a}	0.067^{d}	1.832	0.067^{a}	0.346^{r}	-1.309	0.191^{d}	0.162^{a}	-0.697	0.486^{r}	0.512^{d}
G	A	G	T	G	T	C	G	C	0.129	0.663	0.507^{r}		-0.622	0.534^{d}		0.659	0.510^{d}		0.457	0.647^{d}	
G	A	A	T	G	T	C	G	C	0.128	-1.118	0.264^{d}		-1.346	0.178^{d}		1.429	0.153^{d}		-0.642	0.521^{a}	
G	G	G	T	G	T	C	G	C	0.126	2.917	0.004^{d}		1.131	0.258^{r}		-2.618	0.009^{a}		1.131	0.258^{d}	
G	A	G	C	A	T	C	G	C	0.114	1.583	0.113^{a}		-0.587	0.557^{d}		0.309	0.757^{a}		-1.477	0.140^{r}	
G	A	G	C	G	T	C	G	C	0.091	0.775	0.439^{r}		0.343	0.731^{a}		-0.965	0.335^{a}		-0.734	0.463^{r}	
G	A	G	C	G	T	A	G	C	0.090	-0.576	0.564^{r}		-1.213	0.225^{r}		0.474	0.635^{d}		1.784	0.074^{d}	
A	A	G	C	A	T	C	G	C	0.083	-1.717	0.086^{a}		1.143	0.253^{r}		1.737	0.082^{d}		-0.769	0.442^{r}	
G	A	G	C	G	C	C	G	C	0.080	-0.785	0.433^{d}		1.195	0.232^{r}		0.784	0.433^{r}		1.253	0.210^{r}	

rs6354-rs25528-rs2066713, rs8071667-rs16965623 (*SLC6A4*)						DSM-IV									FTND		
						AD			CD			ND					
1	2	3	4	5	Freq	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global
A	A	T	C	A	0.262	-1.631	0.103^{r}	0.019^{d}	-0.556	0.578^{r}	0.897^{d}	-0.812	0.417^{d}	0.474^{r}	-0.827	0.408^{d}	0.215^{d}
C	C	C	T	A	0.237	-2.795	0.005^{a}		0.910	0.363^{d}		1.442	0.149^{a}		0.608	0.543^{d}	
A	C	C	C	A	0.212	-1.134	0.257^{r}		0.362	0.717^{r}		0.307	0.759^{r}		-0.992	0.321^{r}	
A	A	C	C	A	0.118	1.637	0.102^{r}		0.437	0.662^{r}		-1.783	0.075^{r}		-1.704	0.088^{r}	
C	C	C	C	A	0.089	1.802	0.072^{d}		-0.644	0.520^{d}		-0.747	0.455^{a}		-1.301	0.193^{r}	
A	A	C	C	G	0.081	-0.313	0.754^{d}		-1.029	0.304^{d}		1.095	0.273^{d}		2.119	0.034^{d}	

Significant associations are given in bold. Superscripts following *P* values indicate genetic models used for analysis: *a* additive; *d* dominant; and *r* recessive. Sex, age, study and two out of three addiction status were adjusted while the other one was used as dependent variable in all statistical models.

Table 2.4: Major haplotypes (frequency ≥ 5%) associated with AD, CD, ND or FTND at $P < 0.01$ level in EA sample.

rs11214769-rs1176744 (*HTR3B*)			DSM-IV									FTND		
			AD			CD			ND					
1	2	Freq	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global
A	T	0.674	2.365	0.018^{r}	0.058^{r}	-1.159	0.246^{r}	0.492^{r}	-1.698	0.090^{r}	0.007^{a}	0.880	0.379^{d}	0.625^{d}
G	G	0.288	-1.608	0.108^{d}		1.150	0.250^{a}		-0.696	0.486^{r}		-0.312	0.755^{r}	

rs7942029-rs17116178 (*HTR3B*)			DSM-IV									FTND		
			AD			CD			ND					
1	2	Freq	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global
A	G	0.730	-1.582	0.114^{r}	0.190^{r}	2.717	0.007^{d}	0.055^{d}	-1.317	0.188^{d}	0.512^{r}	0.343	0.731^{d}	0.389^{a}
G	G	0.169	1.036	0.300^{d}		-1.356	0.175^{a}		-0.665	0.506^{r}		1.000	0.317^{a}	
A	T	0.100	1.670	0.095^{r}		-2.211	0.027^{r}		1.353	0.176^{r}		-1.132	0.258^{r}	

rs7118530-rs12221649-rs2085421 (*HTR3A*)				DSM-IV									FTND		
				AD			CD			ND					
1	2	3	Freq	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global
T	T	G	0.605	1.252	0.211^{r}	0.361^{d}	1.728	0.084^{d}	0.339^{d}	-2.254	0.024^{r}	0.101^{r}	-3.119	**0.002^{a}**	0.008^{r}
C	T	A	0.228	-1.733	0.083^{d}		-1.434	0.152^{r}		0.635	0.526^{d}		2.530	0.011^{a}	
C	C	A	0.138	0.768	0.442^{a}		-0.780	0.435^{r}		1.016	0.310^{r}		2.052	0.040^{r}	

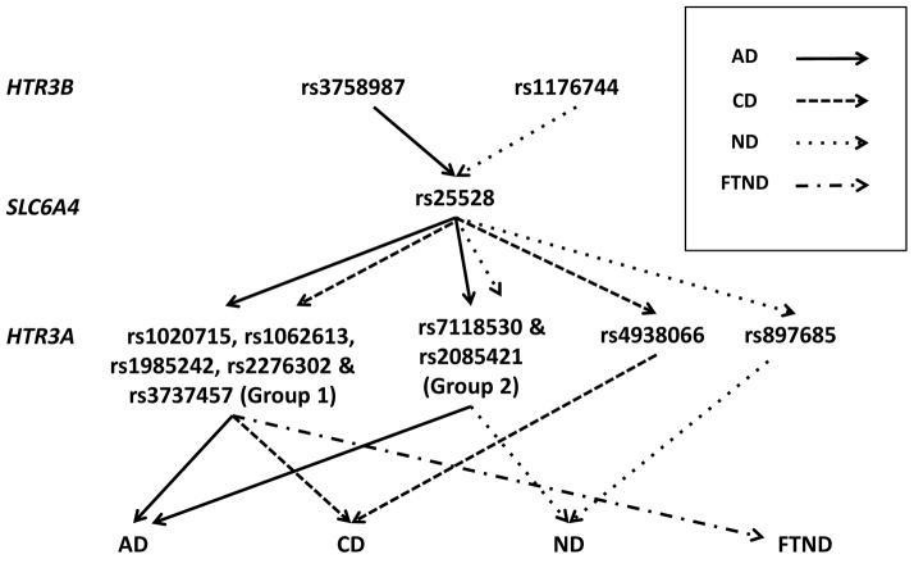
rs6354-rs25528-rs2066713-rs425147-rs8071667 (*SLC6A4*)						DSM-IV									FTND		
						AD			CD			ND					
1	2	3	4	5	Freq	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global	Hap score	*P* hap	*P* global
A	A	T	G	C	0.395	-1.241	0.215^{d}	0.735^{r}	-1.626	0.104^{a}	0.082^{a}	-1.296	0.195^{r}	0.720^{r}	-2.719	0.007^{r}	0.015^{r}
A	A	C	G	C	0.324	0.848	0.396^{a}		0.266	0.790^{a}		0.429	0.668^{a}		0.807	0.420^{d}	
C	C	C	G	T	0.179	0.736	0.462^{d}		0.231	0.817^{a}		0.388	0.698^{r}		2.230	0.026^{r}	
A	A	C	A	C	0.093	1.273	0.203^{r}		1.418	0.156^{a}		-1.177	0.239^{a}		-1.180	0.238^{a}	

Significant associations are given in bold. Superscripts following *P* values indicate genetic models used for analysis: *a* additive; *d* dominant; and *r* recessive. Sex, age, study and two out of three addiction status were adjusted while the other one was used as dependent variable in all statistical models.

2.4.3 SNP-by-SNP interaction analysis of *HTR3A*, *HTR3B*, and *SLC6A4*

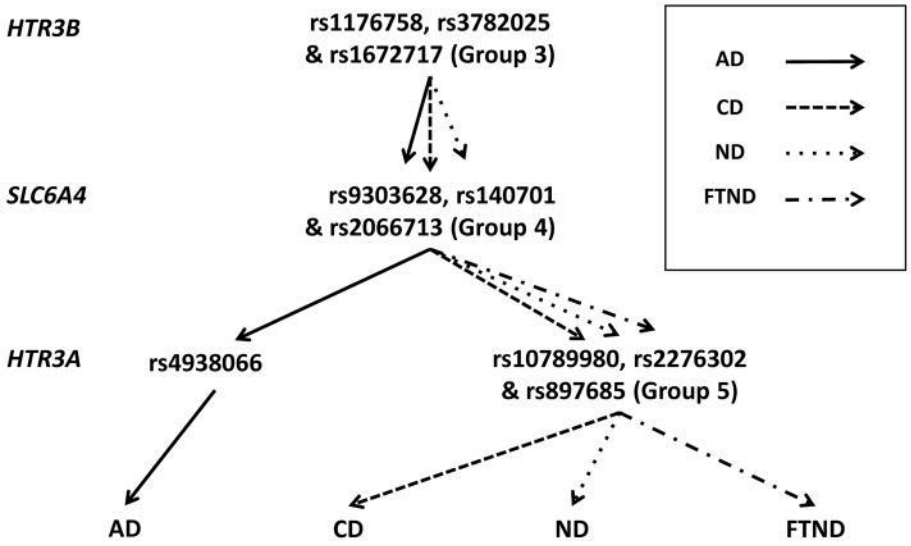
Two previous studies reported by our group indicated that there exist significant epistatic effects among *HTR3A*, *HTR3B*, and *SLC6A4* in both AAs and EAs in either AD or ND.[69, 70] As shown in **Table 2.5**, we determined the best interaction models for AD, CD, ND, and FTND based on CVC > 7 of 10, prediction accuracy > 55% and empirical *P* value < 0.005 for each model based on 10^7 permutation tests. Although the SNPs involved in different interaction models were not exactly the same, **Figures 2.6** and **2.7** show great overlaps and correlations among SNPs based on the LD structure of those SNPs included in each model.

Figure 2.6: Summary of detected interaction models in the SAGE AA sample.



GMDR-GPU[82] was used to perform exhaustive searches for two- to five-way interaction models. The best interaction models for AD, CD, ND, and FTND shown in the figure were determined based on CVC > 7 of 10 and prediction accuracy > 55%. The *P* value associated for each model shown here was < 0.005 based on 10^7 permutation tests. Interaction models with different phenotypes involved overlapped and highly correlated SNPs, which were grouped together. Group 1 includes rs1020715, rs1062613, rs1985242, rs2276302 and rs3737457; Group 2 includes rs7118530 and rs2085421. Pair-wise *D*′ values of adjacent SNPs within each group are > 0.7. SNP combinations for different phenotypes are presented by different types of *arrows*.

Figure 2.7: Summary of detected interaction models in the SAGE EA sample.



GMDR-GPU[82] was used to perform exhaustive searches for two- to five-way interaction models. The best interaction models for AD, CD, ND, and FTND shown in the figure were determined based on CVC > 7 of 10 and prediction accuracy > 55%. The *P* value associated for each model shown here was < 0.005 based on 10^7 permutation tests. Interaction models with different phenotypes involved overlapped and highly correlated SNPs, which were grouped together. Group 3 includes rs1176758, rs3782025 and rs1672717; Group 4 includes rs9303628, rs140701 and rs2066713; Group 5 includes rs10789980, rs2276302 and rs897685. Pair-wise *D*' values of adjacent SNPs within each group are > 0.7. SNP combinations for different phenotypes are presented by different types of *arrows*.

Table 2.5: Detected best SNP combinations in *HTR3A, HTR3B* and *SLC6A4* associated with AD, CD, ND or FTND based on cross validation consistency (CVC), prediction accuracy and empirical *P* value from 10^7 permutations in AA and EA samples.

Sample	SNP combination	Phenotype	CVC	Prediction accuracy	*P* value based on 10^7 permutations
AA	*HTR3A*: rs1020715, rs2276302, rs2085421 *HTR3B*: rs3758987 *SLC6A4*: rs25528	DSM-IV AD	8/10	57.63%	**0.00014**
	HTR3A: rs1985242, rs4938066 *SLC6A4*: rs25528	DSM-IV CD	7/10	58.74%	**0.0017**
	HTR3A: rs897685, rs7118530 *HTR3B*: rs1176744 *SLC6A4*: rs25528	DSM-IV ND	7/10	59.66%	**0.000057**
	HTR3A: rs1062613, rs3737457	FTND	10/10	60.34%	**0.000017**
EA	*HTR3A*: rs4938066 *HTR3B*: rs3782025 *SLC6A4*: rs2066713	DSM-IV AD	9/10	56.05%	**0.00074**
	HTR3A: rs2276302 *HTR3B*: rs1672717 *SLC6A4*: rs140701	DSM-IV CD	7/10	58.27%	**0.00029**
	HTR3A: rs897685 *HTR3B*: rs1176758 *SLC6A4*: rs9303628	DSM-IV ND	8/10	55.18%	**0.00075**
	HTR3A: rs10789980 *SLC6A4*: rs9303628	FTND	7/10	55.25%	**0.0039**

2.5 Discussion

Through individual SNP- and haplotype-based association analysis and SNP-by-SNP interaction analysis on AD, CD, ND, and FTND, we identified significant independent and interactive effects among 88 genotyped and imputed variants within *HTR3A*, *HTR3B*, and *SLC6A4* in the AA and EA samples. These findings confirm our hypothesis that interactive effects exist between 5-HT3 receptors and transporters in governing trans-synaptic serotonergic signaling underlying the pathophysiology of multiple addictions in two ethnic groups.

On the individual polymorphic level, rs2066713 was significantly associated with FTND in the AA sample. As a tag SNP located in the alternative splicing region of *SLC6A4* involving noncoding exons 1A and 1B, it is likely to regulate expression of the gene in humans, because exon 1B is surrounded by several consensus sites for transcription factors AP-1, AP-2, CREB/ATF, and NF-κB.[64] Rs2066713 was reported to be associated with schizophrenia in a South Indian population[84] and with autism in Caucasian samples.[85] On the haplotypic level, two SNPs located in the 5′-region of *HTR3B* (rs3891484 and rs3758987) and three SNPs located in the 3′-region of *HTR3A* (rs7118530, rs12221649, and rs2085421) are associated with AD in the AA sample and FTND in the EA sample, respectively. However, these association signals are not as strong as the SNP-by-SNP interaction results we obtained.

In the AA sample, there are 12 SNPs included in the four interaction models of AD, CD, ND, and FTND, which can be treated as seven groups based on *D*′ values. Rs25528 is the only SNP located in the 5′-region of *SLC6A4* in the AD, CD, and ND interaction models, which is in strong LD with rs2066713 and also locates in the alternative splicing region of *SLC6A4*. It has been reported to be significantly associated with the Beck Depression Inventory.[86] The two

distinct SNPs within *HTR3B* are rs3758987 and rs1176744. Rs3758987 locates in the 5′-UTR region of *HTR3B*, whereas the non-synonymous SNP rs1176744 results in a tyrosine-to-serine change at the 129th amino acid residue of 5-HT3B. This amino acid substitution significantly increases the maximum response of 5-HT3$_{AB}$ to serotonin, slows its deactivation and desensitization kinetics 20- and 10-fold, respectively, and confers a seven-fold increase in the receptors' mean open time.[87] There are two SNP groups and two individual SNPs in *HTR3A* involved in the AA interaction models, as shown in **Figure 2.6**. Group 1 spans from the 5′-UTR to the intron region of *HTR3A*, which covers five SNPs (rs1020715, rs1062613, rs1985242, rs2276302, and rs3737457). Rs1020715 and rs1062613 are translation regulatory variants located in an open reading frame upstream of the translation initiation site of *HTR3A* mRNA.[88] Rs2276302, together with rs3737457, is part of a haplotype reported to be associated with heroin addiction in AAs.[89] Rs897685, rs4938066, and Group 2 are all located in the 3′-UTR region of *HTR3A*.

In the EA sample, as shown in **Figure 2.7**, three SNP groups and one individual SNP are included in the four interaction models for AD, CD, ND, and FTND. Of the three SNPs included in Group 3, rs1176758 is located in the 5′-UTR region of *HTR3B*; Ducci *et al*.[90] reported that the intronic SNP rs3782025 was associated with alcohol use disorders + co-morbid antisocial personality disorder in Finns; rs1672717 was significantly associated with the intensity of nausea and vomiting among cancer patients treated with opioids.[91] Although the functionality of the intronic SNPs rs3782025 and rs1672717 is not clear, the strength of interactive effects between SNPs within the gene and addictions is likely to be similar, as *HTR3B* is covered by only one LD block in Caucasians according to the HapMap data. Group 4 includes three correlated

SNPs: rs9303628, rs140701, and rs2066713. The first two SNPs are located in the intron regions of *SLC6A4*, which may represent new regulatory variants or indicate that they reside in LD with such a variant. Rs2066713 has shown an independent effect on FTND, and rs25528, a SNP in strong LD with rs2066713, is the major interactive signal of *SLC6A4* in the AA sample. The variants of *HTR3A* involved in the interaction models are Group 5 (rs10789980, rs2276302, and rs3737457) and rs4938066. Three of the four SNPs are overlapped in the AA sample, whereas rs10789980 locates within the same open reading frame of *HTR3A* as rs1020715 and rs1062613.

By further examination of SNPs detected in these interactive models for the AA and EA samples, we found that there is one locus in HTR3B, one locus in SLC6A4 and two separate loci in HTR3A that collaboratively contribute to AD, CD, ND and FTND in the EA sample. One locus (rs4938066 in **Figure 2.7**) in HTR3A specifically influences AD, while the other locus (Group 5 in **Figure 2.7**) affects CD, ND and FTND, which may suggest different receptor variations in AD subjects comparing with CD and ND participants that couple with transporter changes in order to take effect. However, relationship among the four interactive models in the AA sample is not as obvious as it is in the EA sample. Also, **Figure 2.6** shows the trend that more loci in HTR3A are involved in the interactive models of the four phenotypes.

Previous studies by our group have shown interaction effects among *HTR3A*, *HTR3B*, and *SLC6A4* in AD and ND samples.[69, 70] However, most case subjects included in the studies reported by Yang *et al*.[70] and Seneviratne *et al*.[69] have primarily only one type of addiction. Thus, this study has extended such an interaction effect among the three genes to subjects with multiple addictive phenotypes (AD, CD, and ND). This strongly implies that variants in the three

genes have significant epistatic effects influencing, not only in one type of addiction, but also in multiple addictions, although limited SNPs with major effects on the three genes were revealed by our previous studies[69, 70] and this one. Result consistency among the three studies were even found at the SNP level. Seneviratne *et al*.[69] revealed two four-variant models carried a risk for AD, which include rs10160548 in *HTR3A*, rs1176744 and rs3782025 in *HTR3B*, and 5'-HTTLPR and rs1042173 in *SLC6A4*. Yang *et al*.[70] showed significant interactions among rs1062613 and rs10160548 in *HTR3A*, rs1176744 in *HTR3B*, and 5'-HTTLPR and rs1042173 in *SLC6A4* in affecting ND.Rs1176744 in *HTR3B* overlaps among the three studies, which makes a residue change from Tyrosine to Serine. Besides rs1176744, this study has rs3782025 of *HTR3B* in common with Seneviratne *et al*.[69] and rs1062613 of *HTR3A* with Yang *et al*.[70] These three SNPs may be important serotonin-receptor- and transporter-function-modifying gene variants or in strong linkage with such variants.

One possible explanation for all these findings is that increased synaptic 5-HT, caused by limited SERT re-uptake abilities, coupled with increased 5-$HT3_{AB}$ receptor responsiveness to 5-HT results in enhanced dopamine transmission in the reward pathway that is associated with a greater risk of multiple addictions. To take it further, cocaine inhibits SERT re-uptake;[92] alcohols increase the maximal efficacy of dopamine activation of 5-HT3 receptors;[93] both nicotine and cocaine compete with serotonin for the 5-HT3 receptor site that controls channel opening.[94]

This hypothesis is supported by the study results of SERT deficient mice. Researchers found that 5-HT3 receptors are upregulated in frontal cortex (+ 46%), parietal cortex (+ 42%), and in stratum oriens of the CA3 region of the hippocampus (+ 18%) of SERT knockout mice.[95] Mutations that result in reduced or absent SERT function in mice have led to increased anxiety

and stress-related behaviors. Although the effects are not as robust as those in the experimental mice, SERT-function-modifying gene variants in humans influence many of the same phenotypes.[96]

Considering other studies using the SAGE dataset, Bierut *et al*.[75] published the first and major genome-wide association study of alcohol dependence, within which they found fifteen SNPs yielded $P < 10^{-5}$ among 948,658 SNPs analyzed. Although the best *P* value of our single SNP- and haplotype-based association analyses is at 10^{-4} level, on the one hand, we only analyzed 88 SNPs applying candidate gene approach; on the other hand, the significant interactive effect among *HTR3A*, *HTR3B* and *SLC6A4* may represent a way to disentangle the influence of comorbid substance-use disorders.

In a study in AA males, Enoch *et al*.[60] showed that rs1176744 in *HTR3B* influenced alcohol dependence. In our analyses, however, we did not detect an independent effect of rs1176744; instead, we found that rs1176744 together with rs11214769 formed a major haplotype, which was significantly associated with ND in the EA sample, and together with rs897685 and rs7118530 in *HTR3A* and rs25528 in *SLC6A4* showed a significant interactive effect on ND in the AA sample. The explanation may lie in sex differences and multiple addictions.

The primary reason for us not pooling the AA and EA samples is that minor allele frequencies of most SNPs are very different for the two ethnic groups, and we wondered such pooling might yield false-positive results.[97] Considering the genetic heterogeneity of AA and EA populations, analyzing them separately may also reduce uncertainty and confidence interval

width. Another reason is that genetic association findings in two diverse samples are providing independent replication.

This study should be considered in the context of its limitations. A functional promoter polymorphism, 5-HTTLPR in *SLC6A4* has been reported to have mixed associations with alcohol, cocaine, heroin, or nicotine dependence.[69, 70, 98-102] However, limited by the original GWAS data of SAGE, we do not have this polymorphism available and are not able to test its associations with the four phenotypes and interactions with other variants. This was also one of the reasons that we chose to replicate our previous findings in alcohol and nicotine dependence at gene level instead of single SNP level, since the previous interactive signals were mainly driven by 5'-HTTLPR from analyzing fewer variants compared with this study.[69, 70] At the same time, by following this approach, we detected a new variant group (rs2066713 and rs25528) in *SLC6A4* that contributes both independently and interactively with variants in *HTR3A* and *HTR3B* to the four addictive phenotypes. These two variants are in strong LD with each other and reside in the alternative splicing region involving noncoding exons 1A, 1B and 1C, which may account for another major interactive signal in *SLC6A4*.

We also acknowledge that gene-by-gene interaction detected by this study through genetic epidemiological approach remains to be further tested experimentally in future.[103] Even though further improvement of the GMDR-GPU is still needed, the GMDR has been successful in identifying the significant interaction of *CHRNA4* with *CHRNB2*,[104] *NTRK2* with *BDNF*,[104] and *GABBR1* with *GABBR2*[105] in ND, of *LEPR* and *ADRB2* in obesity,[106] and of *HNF4A* and *KCNJ11* in type 2 diabetes (T2D),[107] to name a few.

In summary, we showed significant interactive effects among *HTR3A*, *HTR3B*, and *SLC6A4* in AA and EA subjects with multiple addictions. Such findings not only corroborate the findings from our previous studies on single-agent addictions, but also conform with the increasingly appreciated epistatic effects of variants in complex trait studies, which may account for the mysterious missing heritability.[36, 104]

2.6 Chapter acknowledgments

This chapter was adapted from Yang and Li.[37]

2.7 Supplementary note about GMDR method and program

Multifactor dimensionality reduction (MDR) uses a nonparametric and genetic model-free approach to address concerns about inaccurate parameter estimates and low power for identifying interactions in relatively small sample sizes (Hahn, Ritchie, and Moore 2003). Compared with other MDR algorithms, one of the major advantages of GMDR developed by our group is the allowance of covariate justification, which calculates a “score” statistic (residuals of logistic regression for binary traits and residuals of linear regression for quantitative traits) for each subject.[83] Advanced users of GMDR-GPU program also have the option of providing the scores directly to the program so they can use their own regression models to calculate the scores.[82] After appropriate justification of covariates, GMDR-GPU trains and ranks all SNP combinations for a given order following a cross-validation framework. Specifically, the data are randomly divided into *K* (default = 10) partitions of equal size for *K*-fold cross-validation.

Accordingly, *K* training sets are formed where each set consists of all but one of the *K* data partitions. Within each training set, the genotypes of all the SNP combinations are classified as high-risk or low-risk cells according to the genotype and score data; i.e., the justified phenotypic data; and all the SNP combinations are ranked by their training accuracies. Those combinations with the highest training accuracies are then selected based on the cross-validation consistency (CVC), which is defined by the number of times the particular combination is selected from all the training sets. The higher the CVC, the more robust the SNP combination as a predictive interaction model. After identifying the candidate interaction models, their prediction accuracies are calculated by averaging their corresponding testing accuracies among all the data partitions that are not contained in the training sets. The significance or *P* value is determined by a permutation test based on the prediction accuracy.

2.8 Supplementary data

Supplementary Table: Information on SAGE-Genotyped and Imputed Variants across *HTR3B*, *HTR3A*, and *SLC6A4*

Gene	dbSNP ID	Chromosomal location*	Physical location	Minor allele (MAF) in AA	Minor allele (MAF) in EA	Imputation quality
HTR3B	rs4127472	113765417	5'-UTR	T (0.086)	T (0.001)	
	rs12276717	113768414	5'-UTR	C (0.148)	C (0.009)	
	rs2011249	113768638	5'-UTR	T (0.211)	T (0.205)	
	rs7945619	113770022	5'-UTR	A (0.155)	A (0.131)	
	rs1176758	113770355	5'-UTR	C (0.092)	C (0.398)	
	rs3891484	113772589	5'-UTR	C (0.122)	C (0.135)	
	rs3758987	113775275	5'-UTR	C (0.364)	C (0.285)	r^2(AA) = 0.7032 r^2(EA) = 0.8811
	rs11606194	113780981	intron	C (0.02)	C (0.08)	
	rs4938056	113786539	intron	C (0.355)	T (0.425)	r^2(AA) = 0.7169 r^2(EA) = 0.8182
	rs11214769	113790668	intron	G (0.246)	G (0.289)	
	rs17116113	113796084	intron	C (0.083)	C (0.001)	
	rs17116121	113801668	intron	A (0.089)	A (0.028)	
	rs17116124	113802435	intron	C (0.089)	C (0.028)	
	rs1176746	113802601	intron	T (0.144)	T (0.393)	
	rs1176745	113802934	intron	T (0.117)	T (4.12E-4)	
	rs1176744	113803028	Y129S (NP_006019.1)	G (0.46)	G (0.325)	
	rs2276305	113803104	A154A (NP_006019.1)	A (0.117)	A (0.009)	r^2(AA) = 0.9183 r^2(EA) = 0.9221
	rs17116138	113803666	V182I (NP_006019.1)	A (0.089)	A (0.028)	
	rs2276307	113803887	intron	G (0.126)	G (0.229)	

Gene	dbSNP ID	Chromosomal location*	Physical location	Minor allele (MAF) in AA	Minor allele (MAF) in EA	Imputation quality
	rs11214775	113807181	intron	A (0.279)	A (0.3)	
	rs3782025	113807607	intron	C (0.467)	C (0.461)	
	rs1176735	113809982	intron	A (0.198)	A (0.001)	
	rs1672717	113812733	intron	C (0.084)	C (0.389)	
	rs17614942	113816377	intron	A (0.09)	A (0.067)	r^2(AA) = 0.8308 r^2(EA) = 0.96
	rs7943062	113817286	3'-UTR	A (0.144)	A (0.171)	
	rs7945926	113817468	3'-UTR	T (0.149)	T (0.149)	
	rs7129190	113818752	3'-UTR	C (0.444)	C (0.485)	
	rs720396	113822745	3'-UTR	C (0.233)	C (0.431)	
	rs7942029	113823865	3'-UTR	G (0.305)	G (0.17)	
	rs17116178	113827326	3'-UTR	T (0.152)	T (0.101)	
	rs1176754	113827650	3'-UTR	C (0.411)	C (0.11)	
HTR3A	rs1150229	113833152	5'-UTR	C (0.091)	T (0.498)	
	rs11214789	113839272	5'-UTR	C (0.104)	C (0.14)	
	rs17116202	113840761	5'-UTR	A (0.083)	A (0.001)	
	rs10789980	113841003	5'-UTR	A (0.413)	G (0.495)	
	rs1176752	113843477	5'-UTR	A (0.336)	A (0.074)	
	rs1020715	113845161	5'-UTR	T (0.318)	T (0.002)	
	rs1150226	113845541	5'-UTR	T (0.314)	T (0.088)	
	rs1062613	113846006	5'-UTR	T (0.458)	T (0.225)	
	rs33940208	113846077	L16L (NP_000860.2)	T (0.143)	T (0.012)	r^2(AA) = 0.7614 r^2(EA) = 0.628
	rs1985242	113848273	intron	T (0.344)	A (0.322)	r^2(AA) = 0.8957 r^2(EA) = 0.9583
	rs1176722	113848474	intron	A (0.093)	A (0.143)	

Gene	dbSNP ID	Chromosomal location*	Physical location	Minor allele (MAF) in AA	Minor allele (MAF) in EA	Imputation quality
	rs2276302	113850140	intron	A (0.493)	G (0.31)	
	rs10891611	113851413	intron	C (0.415)	C (0.145)	
	rs11607240	113851853	intron	T (0.018)	T (0.072)	
	rs3737457	113853699	intron	A (0.155)	A (0.012)	
	rs11214796	113854679	intron	T (0.403)	C (0.207)	
	rs10160548	113856681	intron	T (0.263)	G (0.313)	
	rs2276304	113857843	intron	T (0.002)	NA	
	rs1150220	113857886	intron	A (0.121)	A (0.194)	r^2(AA) = 0.9677 r^2(EA) = 0.9859
	rs1176713	113860425	L458L (NP_000860.1)	C (0.305)	C (0.207)	
	rs17543669	113862457	3'-UTR	C (0.013)	C (0.055)	
	rs11214800	113862930	3'-UTR	A (0.099)	C (0.488)	
	rs7115470	113863131	3'-UTR	A (0.107)	A (0.138)	
	rs12421901	113863209	3'-UTR	T (0.113)	T (0.012)	
	rs897685	113864109	3'-UTR	G (0.409)	G (0.102)	
	rs897684	113864370	3'-UTR	T (0.302)	T (0.001)	
	rs1379170	113866551	3'-UTR	G (0.285)	A (0.306)	
	rs7126511	113870883	3'-UTR	G (0.215)	G (0.339)	
	rs11214805	113872040	3'-UTR	G (0.2)	G (0.008)	
	rs11214806	113872627	3'-UTR	C (0.061)	C (0.1)	
	rs10891615	113875853	3'-UTR	T (0.5)	T (0.326)	
	rs2364857	113877557	3'-UTR	C (0.32)	C (0.112)	
	rs11825963	113883337	3'-UTR	A (0.226)	A (0.002)	
	rs4938066	113885412	3'-UTR	A (0.437)	G (0.322)	
	rs10891616	113886236	3'-UTR	C (0.033)	NA	
	rs17544032	113887138	3'-UTR	T (0.18)	T (0.215)	

Gene	dbSNP ID	Chromosomal location*	Physical location	Minor allele (MAF) in AA	Minor allele (MAF) in EA	Imputation quality
	rs7118530	113890125	3'-UTR	C (0.249)	C (0.366)	
	rs12221649	113890353	3'-UTR	C (0.05)	C (0.138)	
	rs2085421	113890708	3'-UTR	A (0.498)	A (0.394)	
	rs1563533	113892617	3'-UTR	A (0.157)	A (0.028)	
SLC6A4	rs8081028	28523314	3'-UTR	A (0.118)	A (0.031)	
	rs7224199	28523726	3'-UTR	G (0.493)	T (0.449)	
	rs3813034	28524804	3'-UTR	C (0.22)	C (0.448)	
	rs1042173	28525011	3'-UTR	G (0.222)	G (0.448)	
	rs9303628	28527228	intron	T (0.35)	C (0.481)	
	rs11657536	28529242	intron	A (0.006)	A (0.026)	
	rs6352	28530193	K604N (NP_001036.1)	NA	C (4.13E-4)	
	rs140701	28538532	intron	A (0.304)	A (0.415)	
	rs140700	28543389	intron	A (0.046)	A (0.091)	
	rs6354	28549898	5'-UTR	C (0.326)	C (0.187)	
	rs25528	28549978	5'-UTR	A (0.46)	C (0.187)	
	rs2066713	28551665	5'-UTR	T (0.262)	T (0.396)	
	rs4251417	28551858	5'-UTR	A (0.014)	A (0.094)	
	rs8071667	28552773	5'-UTR	T (0.236)	T (0.178)	
	rs16965623	28552986	5'-UTR	G (0.081)	G (0.001)	
	rs8073965	28559182	5'-UTR	T (0.014)	T (0.04)	
	rs28437451	28568301	5'-UTR	A (0.003)	A (0.022)	

* chromosomal locations are in reference to the GRCh37.p5 assembly; *MAF* minor allele frequency; r^2 values of the 7 imputed SNPs were obtained using MaCH with reference to the 1000 Genomes AFR and EUR panels, respectively, for the AA and EA samples.

Chapter 3

Rare variant effects for candidate genes

3.1 Abstract

Genetic and functional studies have revealed that both common and rare variants of several nicotinic acetylcholine receptor (nAChR) subunits are associated with nicotine dependence (ND). In this study, we identified variants in 30 candidate genes including nicotinic receptors in 200 sib pairs selected from the Mid-South Tobacco Family (MSTF) population with equal numbers of African Americans (AAs) and European Americans (EAs). We selected 135 of the rare and common variants and genotyped them in the Mid-South Tobacco Case-Control (MSTCC) population, which consists of 3088 AAs and 1430 EAs. None of the genotyped common variants showed significant association with smoking status (smokers vs. non-smokers), Fagerström Test for Nicotine Dependence (FTND) scores, or indexed cigarettes per day (CPD) after Bonferroni correction. Rare variants in *NRXN1*, *CHRNA9*, *CHRNA2*, *NTRK2*, *GABBR2*,

GRIN3A, *DNM1*, *NRXN2*, *NRXN3*, and *ARRB2* were significantly associated with smoking status in the MSTCC AA sample, with weighted sum statistic (WSS) *P* values ranging from 2.42×10^{-3} to 1.31×10^{-4} after 10^6 phenotype rearrangements. We also observed a significant excess of rare nonsynonymous variants exclusive to EA smokers in *NRXN1*, *CHRNA9*, *TAS2R38*, *GRIN3A*, *DBH*, *ANKK1*/*DRD2*, *NRXN3*, and *CDH13* with WSS *P* values between 3.5×10^{-5} and 1×10^{-6}. Variants rs142807401 (A432T) and rs139982841 (A452V) in *CHRNA9* and variants V132L, V389L, rs34755188 (R480H), and rs75981117 (N549S) in *GRIN3A* are of particular interest because they are found in both the AA and EA samples. A significant aggregate contribution of rare and common coding variants in *CHRNA9* to the risk for ND (SKAT-C *P* = 0.0012) was detected by applying the combined sum test in MSTCC EAs. Together, our results indicate that rare variants alone or combined with common variants in a subset of 30 biological candidate genes contribute substantially to the risk of ND.

3.2 Introduction

In recent years, candidate gene and genome-wide association studies (GWAS) have identified several common genetic variants associated with the risk of nicotine dependence (ND). These genes include the nicotinic acetylcholine receptor (nAChR) subunit genes *CHRNA5*, *CHRNA3*, and *CHRNB4* (clustered on human chromosome 15q) and the *CHRNA6* and *CHRNB3* genes (clustered on chromosome 8p).[23-25] Examples of findings involving genes other than nicotinic receptors are the nicotine metabolism gene *CYP2A6*,[24] the dopamine receptor gene *DRD2* and its closely linked gene *ANKK1*,[108, 109] the dopamine hydroxylase gene *DBH*,[110] the brain-derived neurotrophic factor gene *BDNF*,[110, 111] and the synaptic maintenance gene *NRXN1*.[34, 112]

However, the variants of these susceptibility genes can explain only a small to modest part of the estimated heritability for ND; e.g., alleles of the *CHRNA5-CHRNA3-CHRNB4* nAChR gene cluster explain < 1% of the variance in the amount smoked.[21] On the other hand, there is increasing evidence that both common and rare or low-frequency genetic variants are playing a significant role in the involvement of each susceptibility gene for ND and other complex human diseases.[113-115]

Several studies have revealed that rare variants of nAChR subunits are associated with ND both genetically and functionally. Wessel *et al*.[40] investigated the contribution of common and rare variants in 11 *nAChR* genes to Fagerström Test for Nicotine Dependence (FTND) scores in 448 European-American (EA) smokers who participated in a smoking cessation trial. Significant association was found for common and rare variants of *CHRNA5* and *CHRNB2*, as well as for rare variants of *CHRNA4*. Xie *et al*.[41] followed up on the *CHRNA4* finding by sequencing exon 5, where most of the rare nonsynonymous variants were detected, in 1000 ND cases and 1000 non-ND comparison subjects with equal numbers of EAs and African Americans (AAs), and reported that functional rare variants within *CHRNA4* might reduce ND risk. Recently, Haller *et al*.[42] detected protective effects of rare missense variants at conserved residues in *CHRNB4* and examined functional effects of the three major association signal contributors (T375I and T91I in *CHRNB4* and R37H in *CHRNA3*) *in vitro*, the minor alleles of which increased cellular response to nicotine. However, like the other two studies, Haller *et al*.[42] limited their sequencing targets to *nAChR* subunits.

To address whether genes other than *nAChR* subunit genes having common variants associated with ND also contain rare ND susceptibility variants, this study was conducted with

the goal of determining both the individual and the cumulative effects of rare and common variants in genes/regions implicated in ND candidate gene studies and/or GWAS through pooled sequencing of a subset of our Mid-South Tobacco Family (MSTF) samples followed by conducting validation in an independent case-control sample. Additionally, we implemented a three-step strategy to identify association signals of rare and common variants within the same genomic region. First, we evaluated each common variant individually with a univariate statistic; i.e., logistic and linear regression models. Second, rare variants were grouped by genomic regions and analyzed using burden tests, i.e., the weighted sum statistic (WSS);[116] third, we tested for combined effects of rare and common variants with a unified statistical test that allows both types of variants to contribute fully to the overall test statistic.[117]

3.3 Materials and methods

3.3.1 Subjects

Four hundred subjects (200 sib pairs) were selected for variant discovery from the MSTF population based on ethnic group (AAs or EAs), smoking status (smokers or non-smokers), and FTND scores (light smokers: FTND < 4 or heavy smokers: FTND ≥ 4). The reasons for us to choose participants from our family study as discovery samples for deep-sequencing analysis were based on the following two main factors. First, recent studies have shown that rare variants are enriched in family data. If one family member has a copy of a rare allele, half of the siblings are expected to carry it, and hence, variants that are rare in the general population could be very commonly presented in certain families.[118] Second, family-based designs are advantageous for their robustness to population stratification. Participants in this family-based

study were recruited between 1999 and 2004 primarily from the Mid-South states within the USA. More detailed descriptions of demographic and clinical data for these participants can be found in **Supplementary Table 1** and previous publications from our group.[14, 112, 119, 120]

Subjects used for variant validation and analysis were recruited from the same geographical area during 2005–2011 as part of the Mid-South Tobacco Case-Control (MSTCC) study under the same recruitment criteria used for the MSTF sample except the subjects were required to be biologically unrelated to each other. Written informed consent was obtained from all participants under the aegis of a human research protocol approved by the IRB of the University of Virginia and University of Mississippi Medical Center. Questionnaires assessing various smoking-related behaviors and other characteristics of interest were administered to participants. Individuals exhibiting substance dependence or abuse other than for alcohol were excluded. The MSTCC sample included 3088 unrelated AAs (1454 smokers and 1634 non-smokers) and 1430 unrelated EAs (758 smokers and 672 non-smokers). All smokers had smoked at least 100 cigarettes in their lifetimes, while non-smokers were required to have smoked 1-99 cigarettes in their lifetimes, but had no tobacco use in the past year. The ND of each smoker was assessed by the FTND, a commonly used measure, as well as indexed cigarettes per day (CPD) based on a 0 to 3 scale (0: 1–10 CPD, 1: 11–20 CPD, 2: 21–30 CPD and 3: > 30 CPD). Detailed characteristics of the MSTCC AA and EA samples are summarized in **Table 3.1**.

Table 3.1: Demographic and phenotypic characteristics of MSTCC AA and EA samples.

Characteristic	AA (*N* = 3088)		EA (*N* = 1430)	
	Smokers	Non-smokers	Smokers	Non-smokers
Sample size	1454	1634	758	672
Female (%)	681 (46.8)	962 (58.9)	380 (50.1)	451 (67.1)
Age, years (SD)	43.6 (12.5)	42.1 (14.2)	41.6 (12.2)	45.1 (14.9)
Indexed CPD (SD)	1.9 (0.4)	NA	1.9 (0.5)	NA
FTND Score (SD)	8.6 (1.2)	NA	8.0 (1.9)	NA

Indexed CPD and FTND scores are for smokers only. Indexed CPD: 0 (1–10 CPD), 1 (11–20 CPD), 2 (21–30 CPD), 3 (> 30 CPD). FTND score: possible range 0-10. *AA* African American; *CPD* cigarettes per day; *EA* European American; *FTND* Fagerström test for nicotine dependence ; *MSTCC* Mid-South Tobacco Case-Controls; *NA* not applicable; *SD* standard deviation.

3.3.2 Sequencing and genotyping

We used a customized capture panel of 30 targeted genes, which included *nAChR* subunit genes and several neurotransmitter receptor and metabolism genes. Almost all of these genes have been reported by our or other research groups to be associated with at least one ND measure in either AA or EA samples. Please refer to **Table 3.2** for the detailed gene list and related references. The coding regions, UTR regions, and flanking sequences of these genes were covered by the Agilent SureSelect Capture panel (250 kb). We divided the 400 samples from the MSTF study into eight pools based on ethnic group, smoking status, and FTND scores to conduct high-throughput sequencing (50 samples/pool).[121] The concentration of each DNA sample was first measured using the QuantiT™ dsDNA assay kit (Life Technologies, Carlsbad, CA) and then 50 DNA samples were pooled together in an equimolar amount as suggested by manufacturers. Each pooled DNA sample was subsequently subjected to library preparation, targeted capture,

and high-throughput sequencing analysis (72 bp paired-end) according to the protocols provided by manufacturers. Base quality recalibration and alignment were performed using Burrows-Wheeler Aligner (BWA)[122] referencing hg19. We used Syzygy[113] to call variants from the pooled targeted resequencing data.

Together, about 62 GB (868 million reads) of raw sequencing data was obtained from deep-sequencing analysis of the eight pooled DNA samples, with an average of 108 million reads per pooled DNA sample. After appropriate quality control and data filtering, more than 80% of the raw sequencing data was successfully mapped to hg19. A total of 147 million reads were mapped to the targeted regions, which were 100% covered with a median coverage of 106 × for each individual DNA sample. Minor allele frequencies (MAF) were calculated for 25 common variants within coding regions and compared with our previous genotyping results based on the *Taq*Man® assay for individual DNA samples, which revealed that the MAF correlations between results of the two methods are 0.97 for AA samples and 0.90 for EA samples.[121]

After removing intronic and synonymous variants, we identified a total of 430 putative functional variants with a minimum read of more than 500 and an MAF of more than 0.75% from our deep-sequencing analysis of pooled DNA samples. Next, based on their SIFT[123] and PolyPhen[124] scores and MAF rankings, we selected 130 variants, which included 118 rare and 12 common variants, for further validation using independent MSTCC samples. Additional 62 common variants were also chosen from literature on association studies of the 30 genes for validation, based on the fact that they had been reported to be nominally or significantly associated with different ND measures (for a detailed list of these reports, please see **Table**

3.2). Selection of the 130 rare and common variants was based on the SIFT[123] and PolyPhen[124] predictions with the following criteria: 1) all premature stop codons; 2) damaging variants presented in either smoker or non-smoker samples; and 3) damaging and benign variants with an MAF ratio > 1.5 between the smoker and non-smoker samples with the goal of increasing statistical power to detect significant single nucleotide polymorphisms (SNPs) between the two groups. These SNPs were genotyped on the *Taq*Man® OpenArray® genotyping system (Life Technologies, Carlsbad, CA) for the case-control samples. All experiments related to deep sequencing and genotyping validation were performed in the Laboratory of Neurogenetics at the NIAAA, NIH.

Table 3.2: Biological information on rare and common variants of 30 candidate genes.

Gene	(SNP type)/amino acid change	dbSNP ID	Chr.	Hg19 position	Allele 1/ Allele 2	Allele 1 freq. in AA (%)	Allele 1 freq. in EA (%)	PhyloP score	SIFT prediction	Polyphen prediction	Ref.
CHRNB2	E34G	rs200223952	1	154541974	G/A	0	0.07	4.33	TOLERATED	BENIGN	[14]
	Y178*	–		154543833	A/C	0.02	0	0.08	Premature stop codon		
NRXN1	T274P	rs77665267	2	50280522	G/T	0.03	0.07	5.01	DAMAGING	BENIGN	[112]
	R206L	–		50280725	A/C	0.11	0.04	4.46	DAMAGING	PROBABLY DAMAGING	
	(Intron)	rs10208208		50593914	T/G	14.82	2.32	0.05			
	(Intron)	rs10490227		50659515	T/C	23.63	13.73	-0.43			
	(Intron)	rs6721498		50713012	G/A	49.37	52.02	-1.17			
	Y367*	–		50765614	T/G	0.02	0	-0.12	Premature stop codon		
	S62*	–		50850606	T/G	0.02	0	6.04	Premature stop codon		
	(Intron)	rs2193225		51079482	C/T	21.54	50.36	-1.19			
CHRNA1†	E436D	rs61737716	2	175613317	A/C	0	0.04	0.65	TOLERATED	PROBABLY DAMAGING	
DRD3†	G9S	rs6280	3	113890815	T/C	26.30	63.00	0.11	TOLERATED	BENIGN	[17]
	(Intron)	rs7638876		113894300	T/C	19.20	62.20	-2.06			
	(Intergenic)	rs9825563		113900220	G/A	48.98	32.71	-0.30			
CHRNA9	A312T	rs56210055	4	40356031	A/G	7.19	0.85	6.20	DAMAGING	POSSIBLY DAMAGING	
	A315V	rs55633891		40356041	T/C	15.07	12.55	4.48	DAMAGING	BENIGN	
	A432T	rs142807401		40356391	A/G	0.06	0.07	4.45	TOLERATED	BENIGN	
	A452V	rs139982841		40356452	T/C	0.14	0.04	6.02	DAMAGING	PROBABLY DAMAGING	
DRD1	(Intergenic)	rs265975	5	174862195	C/T	35.36	60.71	-0.31			[16]
	(3′ UTR)	rs686		174868700	A/G	43.08	63.57	-0.26			
	R226W	–		174869427	A/G	0.02	0	2.92	DAMAGING	PROBABLY DAMAGING	
	(5′ UTR)	rs4532		174870150	C/T	11.44	33.39	-0.80			
DDC	(Intron)	rs1451371	7	50553051	C/T	30.62	47.20	0.72			[119]
	(Intron)	rs3735273		50596864	T/C	36.10	20.96	-0.41			
	E61D	rs11575292		50611601	A/C	1.38	0.18	0.39	TOLERATED	PROBABLY DAMAGING	
	(Intron)	rs921451		50623285	C/T	22.22	30.50	-0.14			
TAS2R38	R274C	rs114288846	7	141673087	A/G	1.93	0.11	0.48	DAMAGING	PROBABLY DAMAGING	[18]
	V262A	rs1726866		141672705	A/G	32.88	49.82	0.88	TOLERATED	BENIGN	

Gene	(SNP type)/amino acid change	dbSNP ID	Chr.	Hg19 position	Allele 1/ Allele 2	Allele 1 freq. in AA (%)	Allele 1 freq. in EA (%)	PhyloP score	SIFT prediction	Polyphen prediction	Ref.
	W135G	rs139843932		141672670	C/A	0.78	0.04	2.68	DAMAGING	PROBABLY DAMAGING	
CHRNA2	S488*	–	8	27320497	T/G	0.02	0.07	1.74	Premature stop codon		
	R121L	–		27324833	A/C	0.02	0	1.50	DAMAGING	POSSIBLY DAMAGING	
	T22I	rs2472553		27328511	A/G	16.62	13.32	-0.33	TOLERATED	BENIGN	
CHRNB3	(Intergenic)	rs10958725	8	42524584	G/T	30.63	74.85	0.35			[33]
	(Intergenic)	rs10958726		42535909	T/G	39.79	75.02	-0.13			
	(Intergenic)	rs4736835		42547033	C/T	34.85	74.75	-1.27			
	(Intergenic)	rs6474412		42550498	T/C	34.78	74.54	-1.64			
	(5′ UTR)	rs4950		42552633	A/G	27.16	73.95	-0.40			
	(Intron)	rs13280604		42559586	A/G	27.40	73.94	0.43			
	(Intron)	rs6474415		42562938	A/G	23.03	73.66	-0.69			
	H410Y	–		42587678	T/C	0.05	0	2.31	DAMAGING	POSSIBLY DAMAGING	
	K451E	rs35327613		42591735	G/A	4.91	0.25	1.60	TOLERATED	BENIGN	
NTRK2	L140F	rs150692457	9	87322819	C/G	0.35	0	0.64	DAMAGING	PROBABLY DAMAGING	[11]
	(Intron)	rs1187272		87404086	A/G	37.19	66.53	2.11			
	C623*	–		87563481	A/C	0.02	0	-0.14	Premature stop codon		
GABBR2	P742Q	–	9	101068407	T/G	0.03	0	5.15	DAMAGING	PROBABLY DAMAGING	[10]
	G671C	-		101068621	A/C	0.02	0	5.07	DAMAGING	PROBABLY DAMAGING	
	(Intron)	rs2491397		101205162	T/C	44.61	51.63	0.70			
	(Intron)	rs2184026		101304348	T/C	6.31	22.78	-0.78			
	A120A	rs3750344		101340316	C/T	26.07	18.20	0.33			
GRIN3A	(Intron)	rs11788456	9	104348150	G/A	45.09	44.93	0.20			[125]
	(Intron)	rs17189632		104368002	A/T	36.72	43.59	0.11			
	N549S	rs75981117		104433048	C/T	0.11	0.51	3.22	DAMAGING	POSSIBLY DAMAGING	
	R480H	rs34755188		104433255	T/C	0.33	1.88	4.16	DAMAGING	PROBABLY DAMAGING	
	V389L	–		104449017	A/C	0.02	0.04	4.27	DAMAGING	POSSIBLY DAMAGING	

Gene	(SNP type)/amino acid change	dbSNP ID	Chr.	Hg19 position	Allele 1/ Allele 2	Allele 1 freq. in AA (%)	Allele 1 freq. in EA (%)	PhyloP score	SIFT prediction	Polyphen prediction	Ref.
	V132L	–		104499868	A/C	0.02	0.04	3.84	DAMAGING	PROBABLY DAMAGING	
DNM1	L16M	rs61757224	9	130965795	A/C	0.05	0.19	1.07	DAMAGING	PROBABLY DAMAGING	126
	S126*	–		130981002	A/C	0.05	0	5.99	Premature stop codon		
	R228L	–		130982360	T/G	0.03	0	6.05	DAMAGING	PROBABLY DAMAGING	
	Y231*	–		130982464	A/C	0.02	0	2.36	Premature stop codon		
	F336F	rs3003609		130984755	T/C	11.29	54.62	-0.03			
DBH	(Intergenic)	rs3025343	9	136478355	A/G	2.03	10.37	0.63			24
	I340T	rs182974707		136509437	C/T	0.06	0.04	1.84	TOLERATED	POSSIBLY DAMAGING	
	A362V	rs75215331		136513028	T/C	0.06	0.07	5.39	DAMAGING	PROBABLY DAMAGING	
	Y389*	–		136513110	A/C	0.06	0	1.76	Premature stop codon		
	T395P	–		136513126	C/A	0.02	0	4.42	DAMAGING	PROBABLY DAMAGING	
	G482R	rs41316996		136521654	A/G	0.06	0.32	1.89	DAMAGING	PROBABLY DAMAGING	
	R549C	rs6271		136522274	T/C	1.58	6.39	1.68	DAMAGING	PROBABLY DAMAGING	
CHAT	(5′ UTR)	rs1880676	10	50824117	A/G	4.91	23.21	2.03			13
	A120T	rs3810950		50824619	A/G	4.89	23.19	0.88	TOLERATED	BENIGN	
	E188G	rs75011234		50827946	G/A	0.33	0.36	1.64	DAMAGING	PROBABLY DAMAGING	
	L243F	rs8178990		50830171	T/C	1.14	4.98	2.26	DAMAGING	PROBABLY DAMAGING	
	G284S	rs146236256		50833616	A/G	0	0.04	5.92	DAMAGING	PROBABLY DAMAGING	
	P299L	rs868749		50833662	T/C	0.02	0.04	6.01	DAMAGING	PROBABLY DAMAGING	
LOC100188947†	(Intron)	rs1329650	10	93348120	T/G	9.50	26.86	-2.00			23
	(Intron)	rs1028936		93349797	C/A	8.10	18.32	-0.33			
CHRNA10	R421C	rs2231548	11	3687429	A/G	1.22	0.07	2.36	DAMAGING	PROBABLY DAMAGING	

Gene	(SNP type)/amino acid change	dbSNP ID	Chr.	Hg19 position	Allele 1/ Allele 2	Allele 1 freq. in AA (%)	Allele 1 freq. in EA (%)	PhyloP score	SIFT prediction	Polyphen prediction	Ref.
	R351W	rs139793380		3687639	A/G	0.10	0.04	1.65	DAMAGING	PROBABLY DAMAGING	
	L348R	rs147150654		3687647	C/A	0.21	0	3.01	DAMAGING	PROBABLY DAMAGING	
	V248L	rs2231542		3688615	A/C	0.08	0	1.19	DAMAGING	POSSIBLY DAMAGING	
	W86G	–		3690532	C/A	0.02	0.04	4.55	DAMAGING	PROBABLY DAMAGING	
	E85G	rs77958837		3690534	C/T	0.02	0	2.80	TOLERATED	BENIGN	
	T77N	rs55719530		3690558	T/G	1.67	1.92	3.86	DAMAGING	PROBABLY DAMAGING	
BDNF	V74M	rs6265	11	27679916	T/C	3.13	14.36	3.63	DAMAGING	POSSIBLY DAMAGING	[111]
	(Intron)	rs6484320		27703188	T/A	7.48	18.11	1.01			
	E6K	rs66866077		27720937	T/C	1.27	5.84	0	DAMAGING	BENIGN	
	(Intron)	rs2030324		27726915	G/A	47.29	49.27	0.12			
	(Intron)	rs7934165		27731983	A/G	47.22	49.12	1.07			
NRXN2	T1371P	–	11	64390287	G/T	0.51	0.76	4.09	DAMAGING	PROBABLY DAMAGING	
	V53G	–		64410118	C/A	0.03	0	2.81	DAMAGING	POSSIBLY DAMAGING	
	E267G	–		64457927	C/T	0.02	0	2.37	DAMAGING	POSSIBLY DAMAGING	
ARRB1†	H198P	–	11	74989678	G/T	0.02	0.04	4.33	DAMAGING	POSSIBLY DAMAGING	[12]
ANKK1	C52W	rs111789052	11	113258762	G/C	6.09	0.40	2.25	DAMAGING	PROBABLY DAMAGING	[109]
	R122H	rs35877321		113264382	A/G	0.22	1.27	0.70	DAMAGING	POSSIBLY DAMAGING	
	R185Q	rs115800217		113265724	A/G	10.26	0.94	0.62	TOLERATED	PROBABLY DAMAGING	
	R237*	rs56047699		113266815	T/C	0.02	0.04	1.36	Premature stop codon		
	S313*	-		113268045	A/C	0.19	0.11	-0.43	Premature stop codon		
	G318R	rs11604671		113268059	A/G	10.92	42.70	-0.01	TOLERATED	BENIGN	
	P351S	rs186633697		113269742	T/C	0.21	0	0.03	TOLERATED	BENIGN	

Gene	(SNP type)/amino acid change	dbSNP ID	Chr.	Hg19 position	Allele 1/ Allele 2	Allele 1 freq. in AA (%)	Allele 1 freq. in EA (%)	PhyloP score	SIFT prediction	Polyphen prediction	Ref.
	E376K	rs56299709		113269817	A/G	1.24	0.04	3.29	TOLERATED	PROBABLY DAMAGING	
	R445C	rs78229381		113270024	T/C	6.16	0.58	0.66	DAMAGING	PROBABLY DAMAGING	
	E458G	rs184645039		113270064	G/A	0.54	0.11	4.23	DAMAGING	PROBABLY DAMAGING	
	H490R	rs2734849		113270160	G/A	16.71	43.24	-1.12	TOLERATED	BENIGN	
	E587*	rs113005509		113270450	T/G	2.28	0.14	0.84	Premature stop codon		
	Q657*	rs202222056		113270660	T/C	0.49	0	1.08	Premature stop codon		
	R734C	–		113270891	T/C	0.03	0.11	0.06	DAMAGING	PROBABLY DAMAGING	
DRD2	E181*	–	11	113286325	A/C	0.02	0.04	3.68	Premature stop codon		[109]
	(Intron)	rs2075654		113289066	T/C	4.25	19.73	0.49			
	(Intron)	rs2075652		113294898	A/G	4.87	1.12	-0.01			
	(Intron)	rs4586205		113307129	T/G	35.61	71.86	-0.88			
NRXN3	Y234*	rs199840331	14	79181259	A/C	0.02	0	-0.33	Premature stop codon		
	G696*	–		79433576	T/G	0.16	0.04	6.33	Premature stop codon		
	T99P	–		79933611	C/A	0.05	0.04	5.18	DAMAGING	POSSIBLY DAMAGING	
CHRNA5	(Intron)	rs588765	15	78865425	T/C	29.46	38.84	-0.27			[28]
	V134I	rs2229961		78880752	A/G	0.40	0.95	5.99	DAMAGING	PROBABLY DAMAGING	
	K167R	rs80087508		78882233	G/A	1.87	0.11	5.01	DAMAGING	PROBABLY DAMAGING	
	D398N	rs16969968		78882925	A/G	6.01	29.51	3.19	TOLERATED	BENIGN	
CHRNA3	(3′ UTR)	rs578776	15	78888400	G/A	46.33	65.19	0.09			[28]
	H217Y	rs72650603		78894335	A/G	0.05	0.22	6.42	DAMAGING	PROBABLY DAMAGING	
	Y215Y	rs1051730		78894339	A/G	12.81	30.20	2.54			
	(Intron)	rs6495308		78907656	C/T	29.74	29.26	-1.56			
	R37H	rs8192475		78911230	T/C	1.04	4.40	3.28	DAMAGING	POSSIBLY DAMAGING	
CHRNB4	R497C	–	15	78917483	A/G	0.05	0	-1.82	DAMAGING	PROBABLY DAMAGING	[28]
	F462V	–		78917588	C/A	0.02	0	4.75	DAMAGING	PROBABLY DAMAGING	

Gene	(SNP type)/amino acid change	dbSNP ID	Chr.	Hg19 position	Allele 1/ Allele 2	Allele 1 freq. in AA (%)	Allele 1 freq. in EA (%)	PhyloP score	SIFT prediction	Polyphen prediction	Ref.
	R349C	rs56235003		78921602	A/G	0.10	0.61	1.40	DAMAGING	PROBABLY DAMAGING	
	P145A	–		78922214	C/G	0.02	0	5.76	DAMAGING	PROBABLY DAMAGING	
	S140G	rs56218866		78922229	C/T	4.25	0.83	2.22	TOLERATED	POSSIBLY DAMAGING	
	T91I	rs12914008		78923505	A/G	0.73	3.58	1.72	TOLERATED	BENIGN	
	N41S	rs75495090		78927863	C/T	1.40	0.22	4.40	DAMAGING	PROBABLY DAMAGING	
CDH13	N39S	rs72807847	16	82892037	G/A	3.18	0.83	1.25	TOLERATED	BENIGN	
	V464I	rs200591230		83711918	A/G	0.02	0.07	3.39	DAMAGING	PROBABLY DAMAGING	
ARRB2	T84P	–	17	4619841	C/A	0.02	0	4.29	DAMAGING	PROBABLY DAMAGING	[12]
	H281Q	–		4622686	A/C	0.03	0	-0.47	DAMAGING	POSSIBLY DAMAGING	
CHRNA4†	(3′ UTR)	rs2236196	20	61977556	A/G	35.26	73.67	-0.24			[14]
	P457L	rs201739273		61981180	A/G	0.03	0	0.49	DAMAGING	POSSIBLY DAMAGING	
	(Intron)	rs2273504		61988061	A/G	15.84	17.85	-0.53			

AA African American; *Chr.* chromosome; *EA* European American; *Freq.* frequency; *Ref.* reference; *SNP* single nucleotide polymorphism; *UTR* untranslated region. † none or only one rare variant validated in this gene, so burden rare variant analysis was not applicable; – not reported in dbSNP database by 2/17/2014. SNP positions are based on human genome reference assembly build 37.1 (hg19). PhyloP score is basewise vertebrate conservation score.

3.3.3 Data analysis

We arbitrarily used a 5% MAF threshold to define rare and common variants for all samples. Conservation status was determined by the basewise vertebrate conservation PhyloP score.[127] A site was defined as conserved when its PhyloP score was ≥ 2, corresponding to a *P* value of 0.01. Both SIFT[123] and PolyPhen[124] were used to predict the effect of nonsynonymous variants on protein structure and function. SIFT yields two predictions: tolerated and damaging, and PolyPhen offers three: benign, possibly damaging, and probably damaging. Since all samples used in this study were recruited from the same geographical region of Mississippi following exactly same inclusion and exclusion criteria, significant population stratification was not detected between smokers and non-smokers in either AAs or EAs based on principle component analysis of 49 and 51 common variants included in this study, respectively, for each ethnic group (**Supplementary Figure 2**) and other genotyping results on the same samples (data not shown).

For common variants, we performed individual SNP-based association analysis with smoking status using logistic regression models and with FTND and indexed CPD using linear regression models as implemented in PLINK.[78] Additive, dominant, and recessive genetic models were tested for each SNP, adjusted for sex and age in the AA and EA samples separately. All common variants were in Hardy-Weinberg equilibrium within population.

As reported that grouping rare variants together would increase statistical power for association analysis, we used the WSS pooling method[116] to test for association of rare variants with smoking status. This method is applicable to genomic regions with at least two rare nonsynonymous variants. In most cases, one genomic region contained a single gene, the exceptions being the *ANKK1/DRD2* and *CHRNA5/A3/B4* gene clusters. The WSS method can only

accommodate binary response variables because of its intrinsic characteristics.[116] In WSS, rare variant counts within the same genomic region for each individual are accumulated rather than collapsed, as implemented in the Cohort Allelic Sums Test (CAST).[128] This method puts greater weight on alleles with lower frequencies in controls, which have a higher tendency to be functional both biologically and statistically. Scores for all subjects are then ordered, and the WSS is computed as the sum of ranks for all cases. Variants over-represented in cases will have larger WSS values. Then 10^6 permutations were performed to determine *P* values for each genomic region. Limited by computational burden, 10^8 permutations were implemented only when 10^6 phenotype rearrangements were insufficient to acquire an exact *P* value.

After obtaining association results for common and rare variants separately, we evaluated the cumulative effects of both rare and common variants on smoking status using the combined sum tests (i.e., SKAT-C and Burden-C) and adaptive sum tests (i.e., SKAT-A and Burden-A) with age and sex controlled.[117] Smoking status was used as the sole response variable for the following two reasons: 1) to keep analysis results consistent with rare variant analysis; 2) the other two phenotypes (FTND and indexed CPD) are available for smokers only, using of which means excluding around half of the samples and rare variants presented only in non-smoker samples. The combined sum tests choose the weight parameter in such a manner that rare and common variants contribute equally to the overall test statistic. In contrast, the adaptive sum tests are more powerful if the overall effect sizes of rare and common variants are very different, for example, when a trait is associated only with rare or common variants in the region. Because the relative contribution of rare and common variants to ND risk is unknown, we used both tests to estimate their combined effects. Burden and variance-component (e.g.,

SKAT) tests are two major types of group-wise association tests proposed for rare variant analysis, which in this case were extended to accommodate combined analysis of rare and common variants by adjusting the weighting scheme. Only genomic regions with at least one rare and one common variant can be analyzed by this approach.

To determine the effect directions of significant results obtained from the above group-wise tests, we performed case control-based association analysis for each rare variant using PLINK.[78] Then rare variants were separated into two groups based on their estimated odds ratios (OR): if OR > 1, the rare variant was predicted to increase smoking risk; if OR < 1, the rare variant was considered to be protective. However, limited by low frequencies of the rare variants and our moderate sample size used in this study, OR was not available for every rare variant, which happened mostly for rare variants with fewer copies of the minor allele. In this case, we roughly assigned the variant to the risk group if more minor alleles were in smokers; otherwise, to the protective group. For collapsing methods, such as the WSS test, the statistical power decreases dramatically as the proportion of functional variants excluded from the analysis increases.[129] Also, because most of the genes or genomic regions investigated in this study have only 2 to 4 rare variants, splitting them based on their effect directions would provide little information about association with the phenotype of interest given our sample sizes.[130]

As a result, we only performed effect direction specific combined and adaptive sum tests, not WSS, as described above to further characterize cumulative variant effect directions. Even though we put rare and common variants with the same effect direction together, some of the groups still had limited number of variants. For groups with one rare variant and one

common variant, SKAT-C and Burden-C tests are equivalent, so do SKAT-A and Burden-A tests; if only rare or common variants exist in a group, SKAT-C will provide the same results as SKAT-A, which also applies to Burden-C and Burden-A; in cases of only one rare or common variant, all four tests are equivalent to logistic regression analysis.

Bonferroni corrections were used to select significant association results for all analyses. Uncorrected *P* values are presented throughout the manuscript.

3.4 Results

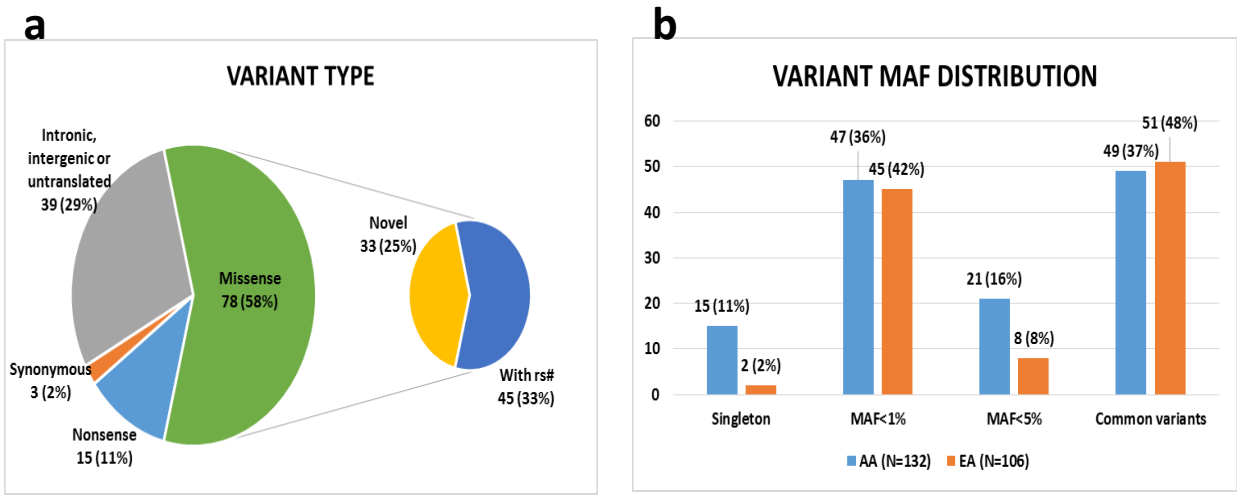
3.4.1 Description of variants and their functionality prediction

There existed 135 out of the 192 variants selected for validation in the MSTCC samples based on genotyping results, which include 33 novel variants (25%; without rs numbers in the dbSNP database as searched on 2/17/2014) in 30 candidate genes (**Table 3.2**). As shown in **Figure 3.1a**, 58% of these variants (*N* = 78) are missense; 11% (*N* = 15) are nonsense–premature stop codons; and 2% (*N* = 3) are synonymous; the remaining 29% (*N* = 39) are from intronic, intergenic, or untranslated regions. Of the 93 nonsynonymous variants, 79 (85%) were predicted to be damaging by PolyPhen, SIFT, or both. The prediction concordance rate between SIFT and PolyPhen programs was 51% (69/135); 14 of 69 were predicted as tolerated by SIFT and benign by PolyPhen; the remaining 55 were predicted to be damaging by SIFT and possibly or probably damaging by PolyPhen. All 33 novel variants were nonsynonymous; they will be mentioned as amino acid change throughout the manuscript. Additionally, 55% of the coding variants were located at conserved sites (53/96; PhyloP score ≥ 2)[127] compared with only 5% of non-coding

variants (2/39). The proportion of conserved sites is significantly different among the coding and non-coding variants (Fisher's Exact $P = 1.59 \times 10^{-8}$).

Of the validated variants, 67% are rare (91/135; MAF < 5%) in AAs, EAs, or both (**Table 3.2**), many appearing only once in all individuals (17/91 = 19% are singletons) and 7 appearing once only in both the AA and EA samples. Among the 44 common variants, 77% (34/44) belong to non-coding regions compared with 5% (5/91) of the 91 rare variants (Fisher's exact test $P = 8.82 \times 10^{-18}$), which is consistent with data from exome sequencing studies that non-synonymous coding variants are significantly skewed toward low frequencies.[131] **Figure 3.1b** compares the MAF distributions in the AA and EA samples for different MAF groups, revealing a higher percentage of singletons and rare variants with MAF between 1% and 5%, and a lower proportion of common variants in the AA sample relative to the EA sample.

Figure 3.1: Descriptive statistics of the 135 validated variants.



(a) Proportions of different variant types. Almost 70% of the validated variants lead to amino acid changes. All novel-identified variants (without rs# in dbSNP database by 2/17/2014) are missense. **(b) The MAF distribution of variants for the AA and EA samples.** The four categories are singleton-only one copy of a rare allele identified in the AA and EA samples, MAF < 1%, MAF < 5%, and common variants. The AA sample has more singletons and low-frequency variants (1% < MAF < 5%) and fewer common variants than the EA sample.

3.4.2 Association analysis results for common variants

There are 24 SNPs across 12 genes (*DRD3*, *CHRNA9*, *DRD1*, *DDC*, *CHRNB3*, *NTRK2*, *GABBR2*, *BDNF*, *ANKK1*, *DRD2*, *CHRNA3*, and *CHRNA4*) and one genomic region (LOC100188947) that show nominally significant association ($P < 0.05$) with smoking status, FTND, or indexed CPD in the AA sample (**Supplementary Table 2**). Of them, rs1051730 in *CHRNA3* has the lowest *P* value, 0.0016 (OR = 2.45; 95% confidence interval [CI] = 1.41, 4.26), which is nominally associated with smoking status under the recessive model. Twenty-one SNPs of 8 genes (*NRXN1*, *CHRNA9*, *TAS2R38*, *CHRNB3*, *NTRK2*, *DBH*, *CHAT*, *BDNF*, and *CHRNA3*) and one genomic region (LOC100188947) are nominally associated with the three phenotypes in the EA sample. Both rs1726866 of *TAS2R38* and rs2030324 of *BDNF* have the smallest *P* value, 0.0017, in the EA sample. The SNP rs1726866 shows nominal damaging effects toward FTND (beta = 0.30; 95% CI = 0.11, 0.49) under the additive model, while rs2030324 nominally protects against FTND (beta = -0.51; 95% CI = -0.83 ~ 0.19) under the recessive model.

The SNPs rs55633891 in *CHRNA9*, 5 SNPs (rs10958725, rs10958726, rs4736835, rs6474412, and rs13280604) in *CHRNB3*, rs1187272 in *NTRK2*, rs1329650 in LOC100188947, and rs6484320 in *BDNF* show nominally significant associations in both the AA and EA samples (**Supplementary Table 2**). However, none of these SNPs survives Bonferroni correction (threshold of significance for AAs = 1.13×10^{-4} for 49 variants, 3 genetic models, and 3 phenotypes; for EAs = 1.09×10^{-4} for 51 variants, 3 genetic models, and 3 phenotypes). Of note, some variants have MAF > 5% in only one sample, which were not called common variants based on our definition, but we performed individual variant analysis for these SNPs.

3.4.3 Association analysis results for rare variants

By using the WSS method, 10 genes (*NRXN1*, *CHRNA9*, *CHRNA2*, *NTRK2*, *GABBR2*, *GRIN3A*, *DNM1*, *NRXN2*, *NRXN3*, and *ARRB2*) are significantly associated with smoking status in the AA sample (**Table 3.3**), with *P* values ranging from 1.31×10^{-4} for *CHRNA2* to 2.42×10^{-3} for *GRIN3A* based on 10^6 permutations. The family-wise error rate (FWER) for 19 genomic regions or genes tested in AAs, which contain at least two nonsynonymous rare variants, is 2.63×10^{-3} (0.05/19). There are 7 genes (*NRXN1*, *CHRNA9*, *TAS2R38*, *GRIN3A*, *DBH*, *NRXN3*, and *CDH13*) and 1 gene cluster (*ANKK1/DRD2*) showing significant associations, at *P* values between 1×10^{-6} (*DBH* and *NRXN3*) and 3.5×10^{-5} (*CDH13*) in the EA sample based on 10^6 or 10^8 permutations (i.e., permuting subjects' smoker/non-smoker status for 10^6 or 10^8 times; see **Table 3.3**). With 11 genes tested for EAs, the FWER threshold is 4.55×10^{-3} (0.05/11). *TAS2R38* ($P = 2 \times 10^{-6}$), *NRXN3* ($P = 1 \times 10^{-6}$), and *CDH13* ($P = 3.5 \times 10^{-5}$) are the three genes that required 10^8 permutations in order to obtain a reliable *P* value.

The genes *NRXN1*, *CHRNA9*, *GRIN3A*, and *NRXN3* have significantly larger WSS values in both AAs and EAs. *NRXN1* has two nonsynonymous substitutions (R206L and rs77665267) and two premature stop codons (S62* and Y367*) in the AA sample ($P = 2.28 \times 10^{-4}$), while only R206L and rs77665267 were detected in the EA sample ($P = 2 \times 10^{-6}$). The two nonsynonymous variants (rs142807401 and rs139982841) of *CHRNA9* are found in both the AA ($P = 3.81 \times 10^{-4}$) and EA ($P = 8 \times 10^{-6}$) samples, as are the four SNPs (V132L, V389L, rs34755188, and rs75981117) of *GRIN3A* ($P = 2.42 \times 10^{-4}$ in AAs; $P = 8 \times 10^{-6}$ in EAs). For *NRXN3*, there are two premature stop codons (rs199840331 and G696*) and one nonsynonymous variant (T99P) included in the

analysis for AA subjects ($P = 2.17 \times 10^{-4}$) and one premature stop codon (G696*) and one nonsynonymous variant (T99P) included in the analysis for EA subjects ($P = 1 \times 10^{-6}$).

Table 3.3: Significant rare variant association results using weighted sum statistic (WSS) in AA and EA samples.

Gene	AA Sample		EA Sample	
	SNPs	Permuted *P* value	SNPs	Permuted *P* value
NRXN1	rs77665267 (p.T274P) - (p.R206L) - (p.Y367*) - (p.S62*)	**2.28×10^{-4}**	rs77665267 (p.T274P) - (p.R206L)	**2×10^{-6}**
CHRNA9	rs142807401 (p.A432T) rs139982841 (p.A452V)	**3.81×10^{-4}**	rs56210055 (p.A312T) rs142807401 (p.A432T) rs139982841 (p.A452V)	**8×10^{-6}**
TAS2R38	rs139843932 (p.W135G) rs114288846 (p.R274C)	0.5346	rs139843932 (p.W135G) rs114288846 (p.R274C)	**2×10^{-6}**†
CHRNA2	- (p.S488*) - (p.R121L)	**1.31×10^{-4}**	NA	NA
NTRK2	rs150692457 (p.L140F) - (p.C623*)	**4.25×10^{-4}**	NA	NA
GABBR2	- (p.P742Q) - (p.G671C)	**1.58×10^{-4}**	NA	NA
GRIN3A	rs75981117 (p.N549S) rs34755188 (p.R480H) - (p.V389L) - (p.V132L)	**2.42×10^{-3}**	rs75981117 (p.N549S) rs34755188 (p.R480H) - (p.V389L) - (p.V132L)	**8×10^{-6}**
DNM1	rs61757224 (p.L16M) - (p.S126*) - (p.R228L) - (p.Y231*)	**3.53×10^{-4}**	NA	NA
DBH	rs182974707 (p.I340T) rs75215331 (p.A362V) - (p.Y389*) - (p.T395P) rs41316996 (p.G482R) rs6271 (p.R549C)	0.2427	rs182974707 (p.I340T) rs75215331 (p.A362V) rs41316996 (p.G482R)	**1×10^{-6}**

Gene	AA Sample		EA Sample	
	SNPs	Permuted *P* value	SNPs	Permuted *P* value
NRXN2	- (p.T1371P) - (p.V53G) - (p.E267G)	**1.49×10^{-3}**	NA	NA
ANKK1/DRD2	rs35877321 (p.R122H) rs56047699 (p.R237*) - (p.S313*) rs186633697 (p.P351S) rs56299709 (p.E376K) rs184645039 (p.E458G) rs113005509 (p.E587*) rs202222056 (p.Q657*) - (p.R734C) - (p.E181*)	0.8114	rs111789052 (p.C52W) rs35877321 (p.R122H) rs115800217 (p.R185Q) rs56047699 (p.R237*) - (p.S313*) rs56299709 (p.E376K) rs78229381 (p.R445C) rs184645039 (p.E458G) rs113005509 (p.E587*) - (p.R734C) - (p.E181*)	**6×10^{-6}**
NRXN3	rs199840331 (p.Y234*) - (p.G696*) - (p.T99P)	**2.17×10^{-4}**	- (p.G696*) - (p.T99P)	**1×10^{-6}**†
CDH13	rs72807847 (p.N39S) rs200591230 (p.V464I)	0.5231	rs72807847 (p.N39S) rs200591230 (p.V464I)	**3.5×10^{-5}**†
ARRB2	- (p.T84P) - (p.H281Q)	**1.32×10^{-4}**	NA	NA

AA African American; *EA* European American; *NA* not applicable; that is, without two rare nonsynonymous variants in gene or region; *SNP* single nucleotide polymorphism; WSS, weighted sum statistic. Permuted *P* value = value based on 10^6 permutations; - = not reported in dbSNP database by 2/17/2014. SNPs included in both AA and EA rare variant analysis are underlined. Significant association *P* values after correction for multiple testing ($p < 2.63 \times 10^{-3}$ for AA sample and $p < 4.55 \times 10^{-3}$ for EA sample) are given in bold. See "Materials and methods section" for details. † *P* value based on 10^8 permutations.

3.4.4 Association analysis results for rare and common variants

CHRNA9, with two rare variants (rs142807401 and rs139982841) and two common variants (rs56210055 and rs55633891), and *DRD1*, with one rare variant (R226W) and three common variants (rs265975, rs686, and rs4532), are nominally associated with smoking status after correcting for sex and age in the AA sample (**Table 3.4**). The *P* values are 0.0495 for *CHRNA9* using Burden-A method and 0.0458 using Burden-C, and 0.0430 using Burden-A for *DRD1*. All four variants of *CHRNA9* result in amino acid changes, among which rs56210055 has an MAF of 7.19% in AAs, but only 0.85% in EAs. So in the EA sample, with three rare variants (rs56210055, rs142807401, and rs139982841) and one common variant (rs55633891), *CHRNA9* shows significant association, with *P* values of 0.0012, 0.0032, 0.0036, and 0.0080 using SKAT-C, Burden-C, SKAT-A, and Burden-A, respectively (**Table 3.4**). The first three *P* values survive multiple testing correction for 12 genes, which have at least one rare and one common variant and were eligible to be included in this analysis in the EA sample (0.05/12 = 0.0042). Both rare and common variants of *CHRNA9* contribute to the risk for ND in EAs and possibly in AAs.

Nominal significant associations were also detected in effect direction separated analysis for *NRXN1*, *CHRNA9*, *DRD1*, *ANKK1*/*DRD2*, and *CHRNA5*/*A3*/*B4* (**Table 3.4**). Two rare variants (rs77665267 and rs10208208) and one common variant (rs10490227) of *NRXN1* in EAs show a *P* value of 0.0362 using the Burden-A method, indicating a possible combined risk effect of the three variants. The common variant, rs10490227, did not show any significant association with smoking status in individual SNP-based analysis; however, it is nominally associated with FTND (**Supplementary Table 2**). For *CHRNA9*, its nominal association in AAs seems to be mainly driven by one rare variant (rs142807401) and two common variants

(rs56210055 and rs55633891) with decreased probability of smoking. SNPs rs142807401 and rs55633891 have opposite effects in the EA sample, which suggests population specific effects or is simply caused by the rough assignment of effect directions as described in the Materials and Methods section. Three out of the four variants in *DRD1*, which increase smoking risk, results in a nominal association in the AA sample (Burden-C $P = 0.0393$ and Burden-A $P = 0.0372$).

Burden-C and Burden-A methods worked as expected for the effect direction separated analysis according to their theoretical designs and assumptions. Besides *NRXN1*, *CHRNA9* and *DRD1*, these two methods discovered nominal associations between the two genomic regions (*ANKK1*/*DRD2* and *CHRNA5*/*A3*/*B4*) that contain the most variants in this study and smoking status in the AA samples as well. Eight rare variants and one common variant in *ANKK1*/*DRD2* together decrease smoking risk, while eight rare variants and two common variants in *CHRNA5*/*A3*/*B4* display an opposite effect (**Table 3.4**).

For groups with rare variants only, the combined and adaptive sum tests revealed nominal associations between *TAS2R38*, *GRIN3A*, *DNM1*, *DBH* and smoking status, respectively, in either AAs or EAs (**Supplementary Table 4**). This can be seen as a confirmation of the association signals detected by the WSS method. Non-significant association results for rare variant analysis and rare and common variant combined analysis are presented in **Supplementary Tables 3** and **4**.

Table 3.4: Significant combined and adaptive sum test results of cumulative rare- and common-variant effects on smoking status in AA and EA samples.

Gene	AA Sample					EA Sample				
	Rare variant(s)	Common variant(s)	Effect direction	*P* value†		Rare variant(s)	Common variant(s)	Effect direction	*P* value†	
				Separated	Pooled				Separated	Pooled
NRXN1	rs77665267 (p.T274P)	rs10208208	↑	0.5553	0.5537	**rs77665267 (p.T274P)**	**rs10490227**	↑	0.1016	0.1062
		rs6721498		0.3023	0.9368				0.0527	0.0772
	- (p.S62*)			0.5354	0.5398	**rs10208208**			0.0738	0.0683
				0.2996	0.9275				**0.0362**	0.0505
	- (p.R206L)	rs10490227	↓	0.4244		- (p.R206L)	rs6721498	↓	0.4648	
	- (p.Y367*)	rs2193225		0.1991			rs2193225		0.2849	
				0.5232					0.4506	
				0.3547					0.4596	
CHRNA9	**rs139982841 (p.A452V)**		↑	0.2381	0.0706	**rs142807401 (p.A432T)**	**rs55633891 (p.A315V)**	↑	**0.0143**	**0.0012**
				0.2381	0.0766				**0.0143**	**0.0032**
				0.2381	0.1448				**0.0246**	**0.0036**
				0.2381	**0.0495**				**0.0246**	**0.0080**
	rs142807401 (p.A432T)	**rs56210055 (p.A312T)**	↓	**0.0381**		**rs56210055 (p.A312T)**		↓	**0.0119**	
				0.0143					**0.0072**	
		rs55633891 (p.A315V)		0.0942		**rs139982841 (p.A452V)**			**0.0119**	
				0.0353					**0.0072**	
DRD1	**- (p.R226W)**	**rs265975**	↑	0.0572	0.1121			NA		
		rs4532		**0.0393**	**0.0458**					
				0.0549	0.1224					
				0.0372	**0.0430**					
		rs686	↓	0.8101						
				0.8101						
				0.8101						
				0.8101						
ANKK1/ DRD2	rs56299709 (p.E376K)	rs111789052 (p.C52W)	↑	0.4445	0.3915	rs111789052 (p.C52W)	rs2075654	↑	0.3828	0.5566
				0.1179	0.0950		rs4586205		0.1708	0.8627
	- (p.R734C)	rs115800217 (p.R185Q)		0.3915	0.4850	rs35877321 (p.R122H)			0.2649	0.4306
	- (p.E181*)			0.1370	0.1361				0.1303	0.8619
	rs2075654									

Gene	AA Sample					EA Sample				
	Rare variant(s)	Common variant(s)	Effect direction	*P* value†		Rare variant(s)	Common variant(s)	Effect direction	*P* value†	
				Separated	Pooled				Separated	Pooled
		rs11604671 (p.G328R) rs78229381 (p.R445C) rs2734849 (p.H490R)				rs56299709 (p.E376K) rs78229381 (p.R445C) rs184645039 (p.E458G) - (p.R734C)				
	rs35877321 (p.R122H) rs56047699 (p.R237*) - (p.S313*) rs186633697 (p.P351S) rs184645039 (p.E458G) rs113005509 (p.E587*) rs202222056 (p.Q657*) rs2075652	**rs4586205**	↓	0.4965 **0.0371** 0.5029 **0.0273**		rs115800217 (p.R185Q) rs56047699 (p.R237*) - (p.S313*) rs113005509 (p.E587*) - (p.E181*) rs2075652	rs11604671 (p.G328R) rs2734849 (p.H490R)	↓	0.7714 0.3605 0.8413 0.2840	
CHRNA5/ A3/B4	**rs72650603 (p.H217Y) rs8192475 (p.R37H) - (p.R497C) - (p.F462V) rs56235003 (p.R349C) - (p.P145A)**	**rs16969968 (p.D398N) rs1051730 (p.Y215Y)**	↑	0.0901 **0.0323** 0.0701 0.0596	0.2406 0.4546 0.1398 0.4909	rs2229961 (p.V134I) rs72650603 (p.H217Y) rs8192475 (p.R37H)	rs588765 rs16969968 (p.D398N) rs1051730 (p.Y215Y)	↑	0.7389 0.2600 0.6989 0.3322	0.7570 0.9566 0.7185 0.9078

Gene	AA Sample					EA Sample				
	Rare variant(s)	Common variant(s)	Effect direction	*P* value†		Rare variant(s)	Common variant(s)	Effect direction	*P* value†	
				Separated	Pooled				Separated	Pooled
	rs12914008 (p.T91I) rs75495090 (p.N41S)									
	rs2229961 (p.V134I) rs80087508 (p.K167R) rs56218866 (p.S140G)	rs588765 rs578776 rs6495308	↓	0.5737 0.4017 0.4497 0.4983		rs80087508 (p.K167R) rs56235003 (p.R349C) rs56218866 (p.S140G) rs12914008 (p.T91I) rs75495090 (p.N41S)	rs578776 rs6495308	↓	0.6270 0.3729 0.5879 0.5716	

AA African American; *EA* European American; *NA* not applicable; *OR* odds ratio; SNP single nucleotide polymorphism; *P* values for each gene or region were obtained from four statistical methods; i.e., SKAT-C, Burden-C, SKAT-A, and Burden-A; † = p values from top to bottom for each gene or region were obtained in the abovementioned order. Only genes or regions with at least one rare and one common variant were eligible for the pooled analysis. ↑ = variants increase smoking risk estimated from individual variant-based odds ratios (if available) or minor allele counts in Cases and Controls, ↓ = variants decrease smoking risk; effect direction specific tests were applied with p values listed under "Separated". SNP rs numbers are based on dbSNP database (accessed on 2/17/2014). SNPs included in both AA and EA samples for this analysis are underlined. Nominal significant associations ($P < 0.05$) for both "Pooled" and "Separated" analyses are given in bold, including *P* values, SNP, and gene names. See the section of "Materials and methods" for details.

3.5 Discussion

Although none of the 44 common variants showed significant association with any of the three nicotine phenotypes (smoking status, FTND, and indexed CPD) after Bonferroni correction in this study, rare variants in 10 genes (*NRXN1*, *CHRNA9*, *CHRNA2*, *NTRK2*, *GABBR2*, *GRIN3A*, *DNM1*, *NRXN2*, *NRXN3*, and *ARRB2*) in the AA sample and 7 genes (*NRXN1*, *CHRNA9*, *TAS2R38*, *GRIN3A*, *DBH*, *NRXN3*, and *CDH13*) plus 1 gene cluster (*ANKK1*/*DRD2*) in the EA sample are significantly associated with smoking status using the WSS method. Further, we also detected a significant cumulative effect of both rare and common variants in *CHRNA9* that contribute to smoking status with age and sex controlled in the EA sample when applying both the combined and the adaptive sum test.

Among the common variants that are nominally associated with any of the three ND measures, SNP rs1051730 is of great interest. This SNP has the smallest common variant association *P* value in the AA sample, which has been reported as the most significant genome-wide association in meta-analyses of subjects of European ancestry ($P = 2.75 \times 10^{-73}$).[24, 25, 110, 132] Another was rs16969968, the most robust genetic finding on chromosome 15q25 in subjects of European ancestry, with a *P* value of 5.57×10^{-72}.[24, 25, 110, 132] Although we did not find significant associations for these two SNPs in our EA sample, which is likely attributable to the small sample size (758 smokers vs. 672 non-smokers), the nominally significant association presented for the AA sample is of interest, providing an independent replication of the association of this SNP with smoking in our independent samples.

HapMap data show that rs1051730 and rs16969968 are in strong linkage disequilibrium in European and Asian populations but not in AAs ($r^2 = 0.40$).[30] In a meta-analysis of AA

samples, Chen *et al*.[30] found that rs16969968 is more strongly associated with heavy smoking (*P* = 0.0011) than is rs1051730 (*P* = 0.011). In our AA sample, however, only rs1051730 is nominally associated with smoking status (*P* = 0.0016; OR = 2.45; 95% CI = 1.41, 4.26) under the recessive model even though the correlation coefficient between rs1051730 and rs16969968 is 0.42; this is consistent with the HapMap data. As a coding synonymous variant, rs1051730 is expected to have less functional significance than rs16969968, a missense mutation (aspartate to asparagine). So while the functional significance of rs16969968 has been demonstrated *in vitro*[133] and to some extent via α5 knockout mouse models that show a role for the gene,[134] the functional relevance of rs1051730 is undetermined. Based on our study result, we suspect that rs1051730 is in linkage disequilibrium (LD) with another functional missense variant with a large effect but low MAF, other than rs16969968, in our AA sample; or it changes *CHRNA3* expression in a significant way.

For rare variants, although we have 10 and 8 genomic regions significantly associated with smoking status in the AA and EA samples, respectively, the two ethnic samples provide replication for each other only for four genes that overlapped across the samples: *NRXN1*, *CHRNA9*, *GRIN3A*, and *NRXN3*. Among the four genes, *CHRNA9* and *GRIN3A* have rare nonsynonymous variants that are seen in both populations, which could be of importance in an evolutionary functional context because of the implication that they are ancient. Because *CHRNA9* is also significantly and nominally associated with smoking status for rare and common variant combined analysis in both the EA and AA sample, it will be discussed first.

CHRNA9, which codes for nAChR α9, is located on chromosome 4p15.1-p14 and contains five exons and four introns.[135] The protein is composed of 479 amino acids

(UniProtKB/Swiss-Prot ID: Q9UGM1; RefSeq ID: NP_060051) and contains two highly conserved domains, which are the neurotransmitter-gated ion-channel ligand binding domain (aa 31–236) and the neurotransmitter-gated ion-channel transmembrane region (aa 244–457).[136] The nAChR α9 can form homo- or hetero-oligomeric cation-selective channels in conjunction with nAChR α10[137] and is usually expressed in the cochlea, keratinocytes, pituitary gland, B-cells, and T-cells.[137-139] Both α9 and α10 nAChR subunits also are coexpressed in dorsal root ganglion neurons.[140]

The four variants in *CHRNA9* that contribute to the association signals are rs56210055 (p.A312T), rs55633891 (p.A315V), rs142807401 (p.A432T), and rs139982841 (p.A452V). All have PhyloP Scores > 4 (**Table 3.2**). Both ala[312] and ala[315] lie within a transmembrane region composed of 22 amino acids (aa 302–323), whereas ala[432] and ala[452] are located within the cytoplasmic region (aa 324–457). The rs139982841 variant has also been identified in lung cancer tissues in the catalogue of somatic mutations in cancer (COSM587183).

Other researchers have reported nominally significant association of *CHRNA9* (rs4861065) with ND in a female Israeli sample[141] and of *CHRNA9* (rs766988 and rs4861065) with response inhibition, as well as of *CHRNA9* (rs4861065) with selective attention in a subset of the same sample, in which neurocognitive functions are putatively implicated in ND susceptibility.[142] Chikova *et al*.[143] revealed that rs56159866 and rs6819385 in *CHRNA9* are associated with an increased risk of lung cancer, while three SNPs, rs55998310, rs56291234, and rs182073550 (single nucleotide deletion) protect against lung cancer.

All these SNPs are either synonymous variations or within intronic or UTR regions, and therefore lack any obvious direct functional effect but may affect protein production at the

transcriptional and/or translational levels or simply manifest association through linkage disequilibrium with other functional variants. In contrast, the four variants we reported in this study all cause amino acid changes, among which rs56210055 (p.A312T) and rs55633891 (p.A315V) may affect nAChR stability or the permeability of the ion channel, while rs142807401 (p.A432T) and rs139982841 (p.A452V) may influence downstream signaling characteristics based on the amino acid locations they affect. Based on the effect direction specific analysis results shown in **Table 3.4**, these four variants may have a mixture of risk and protective effects in affecting smoking risk. Thus, future functional studies are warranted for these four SNPs in *CHRNA9*.

GRIN3A is localized on chromosome 9q34 and consists of nine exons,[144] which code for glutamate receptor ionotropic NMDA 3A (GluN3A). The deduced protein contains 1115 amino acids (UniProtKB/Swiss-Prot ID: Q8TCU5; RefSeq ID: NP_597702.2) and shows 92.7% identity to rat NMDA receptor 3A.[144] Functional NMDA receptors are heterotetramers composed of two ζ subunits (GluN1) and two ε subunits (GluN2A, GluN2B, GluN2C, or GluN2D) or third subunits (GluN3A or GluN3B), which serve critical functions in neuronal development, functioning, and degeneration of the mammalian central nervous system.[145] GluN3A suppresses NMDA receptor functions in a dominant-negative way.[146, 147] GluN3A-containing NMDA receptors display reduced Ca^{2+} permeability and low sensitivity to Mg^{2+} blockade.[148, 149] The transcript of *GRIN3A* was detected by *in situ* hybridization in human fetal spinal cord and forebrain.[150]

All four substituted amino acids, val^{132}, val^{389}, arg^{480}, and asn^{549}, are located in the extracellular region of GluN3A and are conserved, with PhyloP scores > 3 (**Table 3.2**). We have previously reported common variants of *GRIN3A* significantly associated with different ND

measures in the MSTF population.[125] Different variants within *GRIN3A* have also been associated with Alzheimer's disease[151] and schizophrenia.[152] The recent work by Takata *et al*.[152] identified disease association of a missense variant in *GRIN3A* (p.R480G, rs149729514; P = 0.00042; OR = 1.58) in a Japanese schizophrenia case-control cohort. This association was supported by their meta-analysis with independent Han-Chinese case-control and family samples (combined $P = 3.3 \times 10^{-5}$). However, as the authors suggested, the *GRIN3A* R480G variant was not detected in AA and EA populations, and thus it seems to be Asian specific.

In this study, instead of finding the glycine substitution at residue 480, we identified a histidine substitution at the same position of GluN3A in both AAs and EAs. The ingenious connection between the two studies confers great functional importance for this residue not only in ND, but also in other psychiatric disorders. Another variant, rs75981117 (p.N549S), is an N-linked glycosylation site on GluN3A, which could be important for both the structure and function of the protein. SNPs rs75981117 (p.N549S), rs34755188 (p.R480H), and V389L together show a nominal protective effect against smoking risk in AAs (**Supplementary Table 4**). The functional importance of the four variants may show in ND-related mouse models, as Marco *et al*.[153] recently discovered that overexpression of GluN3A in mouse striatum mimicked the synapse loss observed in Huntington's disease mouse models, whereas genetic deletion of GluN3A prevented synapse degeneration, ameliorated motor and cognitive decline, and reduced striatal atrophy and neuronal loss in the YAC128 Huntington's disease mouse model.

Because of space limitations, we cannot elaborate on the potential functional importance of the rare variants we identified in *NRXN1*, *CHRNA2*, *TAS2R38*, *NTRK2*, *GABBR2*, *DNM1*, *DBH*, *NRXN2*, *ANKK1*/*DRD2*, *NRXN3*, and *CDH13* here. To interpret the results of this

study more appropriately, four main limitations need to be considered. First, rare variants are usually population specific, or even sample specific, which, on one hand, makes replication very difficult and on the other hand, reveals that the rare variants identified in this study are just a starting point. Association studies of these biological candidate genes in other populations and samples are thus warranted. Second, we limited our search to biological candidate genes, which makes these findings not surprising at the gene level. If we are to uncover new genes, more comprehensive and hypothesis-free analyses, particularly genome-wide sequencing analyses of rare variants, are needed. Third, although none of the 44 common variants showed significant association with any of the three nicotine phenotypes after Bonferroni correction in this study, it does not mean common variants in general are not important in affecting smoking risk. The primary reason for our failure of identifying significant association of these common variants with ND measures is more likely related to the sample size used in our study, especially for EAs with a sample size of 1430. Another reason may be the selection of these common variants from our previous studies, 30 and 7 of which showed nominal or significant associations in preceding analysis of MSTF and MSTCC samples, respectively (**Supplementary Table 2**). Nineteen out of the 30 common variants chosen based on former MSTF study results were found nominally associated with at least one of the three ND measures in either AA or EA case-control samples; however, all 7 common variants selected from a meta-analysis including MSTCC samples showed nominal significance in this study composed solely of MSTCC subjects. This phenomenon suggests a possible effect of different study designs – family and case control, which is influenced by diverse inheritance patterns of multi-factorial quantitative genetic traits and environmental factors during development. Regression toward the mean

effect cannot be excluded completely for the common variants selected from the previous family studies, either. Fourth, it is hard to dissect the contribution of each rare variant and relative contribution between rare and common variants, hampered by our sample size and the statistical methods we currently apply.

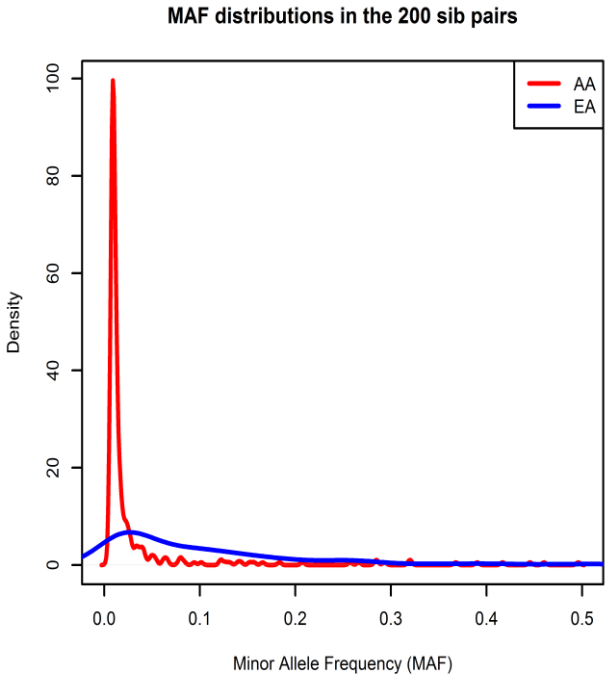
We used one type of burden test; that is, WSS,[116] to accumulate counts of rare variants in separate genomic regions and then examined their overrepresentation in cases vs. controls. The burden test is a compromise between extremely low allele frequency and limited statistical power, which enables detection of pooled rare variant effects but is incapable of disentangling individual effects of rare variants. For combined analysis of rare and common variants, we implemented the combined and adaptive sum tests;[117] the former assumes equal contribution of rare and common variants, and the latter presumes rare variants have different effects than common variants. Without knowing the relative contribution of rare and common variants to any trait of interest, we highly encourage applying both tests to analyze the same dataset as used in this study. We also performed effect direction specific analyses to examine combined effect directions of rare and common variants. Because of the limited number of rare variants available for each gene or genomic region and expected substantial power loss of burden tests when functional variants are excluded, this analytical strategy was only applied to the combined and adaptive sum tests. Nominal association results detected provide evidence for combined effect direction speculation of the variant groups; however, no significant association was discovered. This strategy will be more effective with increased number and more accurate classification of rare variants available.

This study demonstrates for the first time the contribution of common and particularly rare variants, within a subset of biological candidate genes besides *nAChR* subunit genes, to the risk for ND. Our findings about these variants, especially rs56210055 (p.A312T), rs55633891 (p.A315V), rs142807401 (p.A432T), and rs139982841 (p.A452V) in *CHRNA9* and V132L, V389L, rs34755188 (p.R480H), and rs75981117 (p.N549S) in *GRIN3A* are interesting and encouraging and deserve further study using both *in vitro* and *in vivo* approaches.

3.6 Chapter acknowledgments

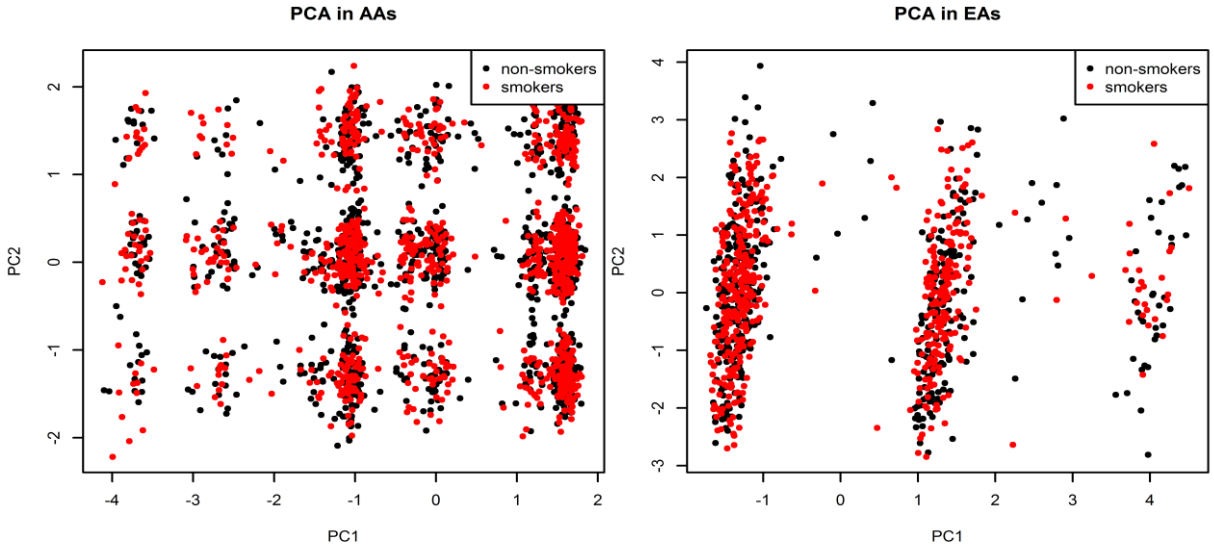
This chapter was adapted from Yang *et al*.[44]

3.7 Supplementary data



Supplementary Figure 1: Density plot of estimated minor allele frequency (MAF) distributions from pooled-sequencing analysis of the 200 sib pairs. Much more rare variants were discovered in AAs compared with EAs.

Supplementary Figure 2: Scatter plots of principal components 1 and 2 for smokers and non-smokers based on principal component analysis of 49 and 51 common variants available for AAs and EAs, respectively. Smokers and non-smokers are uniformly mixed within the two ethnic groups. Significant population stratification was not detected between smokers and non-smokers.



Supplementary Table 1: Demographic and phenotypic characteristics of 200 MSTF sib pairs.

Characteristic	AA (*N* = 200)		EA (*N* = 200)	
	Smokers	Non-smokers	Smokers	Non-smokers
Sample size	100	100	100	100
Female	59	80	62	83
Age, years (SD)	38.5 (13.4)	35.8 (13.0)	37.8 (10.1)	52.9 (14.6)
Indexed CPD (SD)	1.7 (0.8)	NA	1.9 (0.8)	NA
FTND score (SD)	7.3 (1.7)	NA	7.1 (1.7)	NA

Indexed CPD and FTND scores are for smokers only. Indexed CPD: 0 (1–10 CPD), 1 (11–20 CPD), 2 (21–30 CPD), 3 (> 30 CPD). FTND score: possible range 0-10. *AA* African American; *CPD* cigarettes per day; *EA* European American; *FTND* Fagerström test for nicotine dependence ; *MSTF* Mid-South Tobacco Family; *NA* not applicable; *SD* standard deviation.

Supplementary Table 2: Common variant analysis results with smoking status, FTND and indexed CPD in AA and EA samples.

Gene	dbSNP ID	A1	AA Sample							EA Sample						
			Freq (A1)	Smoking Status		FTND		Indexed CPD		Freq (A1)	Smoking Status		FTND		Indexed CPD	
				OR	*P*†	BETA	*P*	BETA	*P*		OR	*P*	BETA	*P*	BETA	*P*
NRXN1	rs10208208*	T	14.82%	1.0060	0.9358	-0.0669	0.2463	-0.0206	0.2614	2.32%			NA			
				1.0040	0.9637	-0.0811	0.2232	-0.0232	0.2729							
				1.0300	0.8959	-0.0599	0.7436	-0.0318	0.5843							
	rs10490227*	T	23.63%	0.9326	0.2438	0.0084	0.8657	0.0222	0.1565	13.73%	1.0320	0.7712	-0.3033	**0.0255**	0.0312	0.4299
				0.9215	0.2640	-0.0181	0.7661	0.0222	0.2427		1.0130	0.9175	-0.3297	**0.0431**	0.0308	0.5139
				0.9027	0.5116	0.1444	0.2737	0.0499	0.2284		1.2280	0.5404	-0.6222	0.1171	0.0813	0.4800
	rs6721498**	A	50.63%	1.0500	0.3416	0.0566	0.1739	0.0044	0.7389	47.98%	0.9211	0.2890	0.1310	0.1810	0.0095	0.7362
				1.0010	0.9929	0.0509	0.4544	-0.0065	0.7625		0.8125	0.0955	0.1444	0.3499	-0.0161	0.7178
				1.1400	0.1205	0.1010	0.1392	0.0183	0.3954		0.9959	0.9748	0.2143	0.2006	0.0470	0.3321
	rs2193225**	C	21.54%	0.9667	0.5788	-0.0434	0.3832	0.0071	0.6529	50.36%	0.9857	0.8511	-0.2689	**0.0061**	-0.0366	0.1960
				0.9548	0.5332	-0.0665	0.2722	0.0021	0.9140		0.9640	0.7704	-0.3377	**0.0337**	-0.0449	0.3299
				0.9816	0.9091	0.0104	0.9374	0.0406	0.3354		0.9981	0.9882	-0.3837	**0.0174**	-0.0538	0.2503
DRD3	**rs6280****	T	26.30%	0.9685	0.5799	0.0011	0.9824	0.0128	0.3942	63.00%	0.9549	0.5643	-0.0601	0.5528	0.0055	0.8543
				0.9550	0.5258	0.0693	0.2469	0.0110	0.5583		0.9181	0.4504	-0.0421	0.7686	0.0333	0.4255
				0.9835	0.9061	-0.2581	**0.0272**	0.0348	0.3446		0.9868	0.9329	-0.1538	0.4451	-0.0446	0.4486
	rs7638876*	T	19.20%	0.9809	0.7608	-0.0558	0.2849	0.0055	0.7387	62.20%	0.9821	0.8232	-0.0188	0.8552	0.0040	0.8933
				0.9518	0.5168	-0.0183	0.7725	0.0047	0.8145		0.9901	0.9308	-0.0122	0.9334	0.0258	0.5399
				1.1210	0.5148	-0.3250	**0.0228**	0.0172	0.7009		0.9513	0.7516	-0.0493	0.8092	-0.0352	0.5512
	rs9825563*	G	48.98%	0.9732	0.5882	0.0142	0.7282	0.0038	0.7672	32.71%	1.1140	0.2083	0.0551	0.6133	0.0089	0.7751
				0.8970	0.1806	0.0429	0.5172	0.0043	0.8371		1.1060	0.3653	0.1236	0.3847	0.0301	0.4643
				1.0400	0.6362	-0.0059	0.9314	0.0061	0.7759		1.2650	0.2173	-0.0783	0.7371	-0.0394	0.5590
CHRNA9	**rs56210055**	A	7.19%	0.9010	0.2868	0.1715	**0.0293**	0.0414	0.0984	0.85%			NA			
				0.8476	0.1190	0.1709	0.0524	0.0337	0.2300							
				2.0320	0.1095	0.4816	0.0994	0.1999	**0.0315**							
	rs55633891	T	15.07%	0.8676	**0.0492**	0.0648	0.2906	0.0011	0.9552	12.55%	1.3520	**0.0107**	-0.0566	0.6874	-0.0533	0.1832
				0.8742	0.0930	0.0579	0.3873	-0.0043	0.8391		1.3670	**0.0181**	-0.0006	0.9972	-0.0534	0.2523
				0.6469	0.1031	0.2506	0.2979	0.0723	0.3415		1.9800	0.1126	-0.5331	0.2154	-0.1344	0.2788
DRD1	**rs265975***	C	35.36%	1.1150	**0.0411**	-0.0236	0.5848	-0.0083	0.5434	60.71%	0.9144	0.2560	0.0466	0.6493	0.0245	0.4046
				1.1010	0.1868	-0.0392	0.5159	-0.0241	0.2053		0.8873	0.2956	-0.0038	0.9793	0.0342	0.4171
				1.2710	**0.0275**	-0.0140	0.8701	0.0160	0.5543		0.8884	0.4295	0.1781	0.3617	0.0294	0.6041
	rs686**	A	43.08%	0.9878	0.8101	-0.0029	0.9434	0.0340	**0.0095**	63.57%	1.0560	0.5047	0.0029	0.9690	-0.0283	0.3403

Gene	dbSNP ID	A1	AA Sample							EA Sample						
			Freq (A1)	Smoking Status		FTND		Indexed CPD		Freq (A1)	Smoking Status		FTND		Indexed CPD	
				OR	*P*†	BETA	*P*	BETA	*P*		OR	*P*	BETA	*P*	BETA	*P*
				0.9603	0.5971	-0.0200	0.7475	0.0352	0.0751		1.0310	0.7844	-0.0444	0.7584	-0.0318	0.4461
				1.0190	0.8411	0.0192	0.7965	0.0604	**0.0109**		1.1660	0.3476	0.1038	0.6098	-0.0481	0.4139
	rs4532**	C	11.44%	1.1640	0.0548	0.0395	0.5265	0.0070	0.7232	33.39%	1.1530	0.0853	0.0093	0.9202	-0.0163	0.5827
				1.1300	0.1640	0.0439	0.5378	0.0047	0.8344		1.1370	0.2517	0.0010	0.9943	-0.0148	0.7189
				2.1030	**0.0152**	0.0649	0.7553	0.0374	0.5685		1.3780	0.0669	0.0407	0.8459	-0.0360	0.5532
DDC	rs1451371*	C	30.62%	1.0170	0.7576	-0.0035	0.9357	-0.0113	0.4170	47.20%	1.0390	0.6268	-0.1401	0.1444	-0.0253	0.3694
				1.0350	0.6295	-0.0175	0.7668	-0.0091	0.6266		0.9279	0.5440	-0.0842	0.5875	-0.0537	0.2328
				0.9876	0.9141	0.0289	0.7616	-0.0294	0.3286		1.2190	0.1384	-0.3139	0.0569	-0.0123	0.7967
	rs3735273*	T	36.10%	0.9494	0.3199	0.0141	0.7421	0.0357	**0.0080**	20.96%	1.0140	0.8831	-0.0134	0.9182	-0.0008	0.9825
				0.9156	0.2263	0.0298	0.6189	0.0531	**0.0049**		1.0110	0.9207	-0.0167	0.9084	0.0013	0.9748
				0.9732	0.7954	-0.0046	0.9578	0.0350	0.1978		1.0490	0.8585	-0.0057	0.9867	-0.0135	0.8917
	rs921451*	C	22.22%	1.0440	0.4655	0.0077	0.8720	0.0229	0.1300	30.50%	1.0830	0.3510	-0.0405	0.7143	-0.0023	0.9425
				1.0530	0.4847	0.0565	0.3574	0.0384	**0.0468**		1.0760	0.5108	-0.0622	0.6596	-0.0053	0.8967
				1.0670	0.6568	-0.1621	0.1694	-0.0039	0.9157		1.2040	0.3386	-0.0155	0.9485	0.0042	0.9518
TAS2R38	**rs1726866***	A	32.88%	0.9801	0.7091	-0.0175	0.6947	0.0040	0.7735	49.82%	1.0710	0.3710	0.3014	**0.0017**	0.0422	0.1316
				0.9722	0.6964	-0.0590	0.3228	-0.0035	0.8507		1.0540	0.6756	0.3243	**0.0417**	0.0500	0.2810
				0.9801	0.8608	0.0703	0.4602	0.0275	0.3597		1.1430	0.2913	0.4929	**0.0018**	0.0646	0.1598
CHRNA2	rs2472553	A	16.62%	1.0200	0.7747	-0.0382	0.4994	0.0115	0.5191	13.32%	1.1720	0.1702	-0.0011	0.9994	0.0197	0.6297
				1.0140	0.8550	-0.0419	0.5146	0.0161	0.4281		1.1790	0.2003	-0.0044	0.9778	0.0284	0.5380
				1.0970	0.6803	-0.0614	0.7383	-0.0097	0.8670		1.4110	0.4250	0.0412	0.9316	-0.0296	0.8308
CHRNB3	**rs10958725‡‡**	G	30.63%	0.9755	0.6487	0.1025	**0.0222**	0.0213	0.1294	74.85%	0.8535	0.0775	-0.0527	0.6604	-0.0525	0.1256
				0.9404	0.3972	0.1410	**0.0191**	0.0322	0.0879		0.8789	0.2497	-0.0679	0.6342	-0.0526	0.2050
				1.0480	0.6927	0.1153	0.2382	0.0164	0.5931		0.6279	**0.0374**	-0.0384	0.9015	-0.1160	0.1997
	rs10958726‡	G	60.21%	1.0200	0.6973	0.0836	**0.0469**	0.0258	0.0519	24.98%	0.8525	0.0760	-0.0539	0.6533	-0.0553	0.1050
				0.9885	0.8772	0.1102	0.0731	0.0453	**0.0198**		0.8741	0.2275	-0.0619	0.6625	-0.0552	0.1792
				1.0960	0.3481	0.1130	0.1529	0.0164	0.5108		0.6350	**0.0454**	-0.0732	0.8163	-0.1236	0.1764
	rs4736835‡‡	C	34.85%	1.0310	0.5576	0.0995	**0.0191**	0.0216	0.1113	74.75%	0.8644	0.1041	-0.0497	0.6792	-0.0498	0.1436
				1.0290	0.6902	0.1452	**0.0141**	0.0346	0.0666		0.8921	0.3058	-0.0657	0.6437	-0.0488	0.2353
				1.0680	0.5422	0.1016	0.2367	0.0154	0.5745		0.6397	**0.0468**	-0.0253	0.9355	-0.1150	0.2036
	rs6474412‡	T	34.78%	1.0340	0.5228	0.0910	**0.0341**	0.0182	0.1790	74.54%	0.8562	0.0841	-0.0134	0.9177	-0.0445	0.1919
				1.0430	0.5623	0.1368	**0.0215**	0.0323	0.0850		0.8856	0.2788	-0.0247	0.8629	-0.0410	0.3205
				1.0500	0.6543	0.0816	0.3500	0.0057	0.8371		0.6228	**0.0344**	0.0327	0.9168	-0.1153	0.2015

Gene	dbSNP ID	A1	AA Sample							EA Sample						
			Freq (A1)	Smoking Status		FTND		Indexed CPD		Freq (A1)	Smoking Status		FTND		Indexed CPD	
				OR	*P*†	BETA	*P*	BETA	*P*		OR	*P*	BETA	*P*	BETA	*P*
	rs4950‡	A	27.16%	1.0990	0.0954	0.0379	0.4069	0.0016	0.9116	73.95%	0.8411	**0.0489**	-0.0557	0.6375	-0.0579	0.0836
				1.1010	0.1818	0.0669	0.2583	0.0105	0.5742		0.8679	0.2040	-0.0792	0.5769	-0.0552	0.1775
				1.2190	0.1418	-0.0109	0.9180	-0.0249	0.4550		0.6112	**0.0206**	-0.0119	0.9681	-0.1394	0.1058
	rs13280604‡	A	27.40%	1.1270	**0.0341**	0.0423	0.3545	0.0032	0.8227	73.94%	0.8445	0.0532	-0.0491	0.6774	-0.0571	0.0866
				1.1350	0.0796	0.0760	0.2007	0.0118	0.5296		0.8713	0.2164	-0.0748	0.5983	-0.0553	0.1769
				1.2640	0.0783	-0.0137	0.8953	-0.0195	0.5542		0.6176	**0.0220**	0.0096	0.9741	-0.1341	0.1161
	rs6474415‡	A	23.03%	1.0550	0.3682	0.0399	0.4029	-0.0052	0.7273	26.34%	0.8382	**0.0423**	-0.0356	0.7636	-0.0519	0.1189
				1.0320	0.6652	0.0638	0.2904	-0.0017	0.9269		0.8686	0.2064	-0.0616	0.6640	-0.0481	0.2406
				1.2340	0.1610	-0.0001	0.9993	-0.0253	0.4950		0.6005	**0.0139**	0.0382	0.8960	-0.1303	0.1230
NTRK2	**rs1187272****	A	37.19%	0.9853	0.7772	0.0169	0.6987	0.0051	0.7079	66.53%	0.9587	0.6204	-0.0030	0.9747	0.0314	0.3432
				1.0290	0.7029	0.0098	0.8722	-0.0204	0.2880		1.1170	0.3216	0.0270	0.8503	0.0650	0.1170
				0.8937	0.2764	0.0467	0.5896	0.0612	**0.0244**		0.6055	**0.0072**	-0.1087	0.6770	-0.0523	0.4900
GABBR2	rs2491397*	T	44.61%	1.0300	0.5699	0.0207	0.6275	-0.0083	0.5403	51.63%	1.0450	0.5653	-0.0242	0.8140	-0.0034	0.9029
				1.0760	0.3470	0.0416	0.5212	-0.0105	0.6076		0.9939	0.9605	-0.0014	0.9929	0.0360	0.4305
				0.9920	0.9290	0.0082	0.9125	-0.0114	0.6306		1.1390	0.3106	-0.0618	0.7016	-0.0470	0.3131
	rs2184026*	T	6.31%	1.1780	0.1137	-0.0942	0.2624	-0.0251	0.3425	22.78%	1.0730	0.4473	-0.0724	0.5388	0.0132	0.7036
				1.2420	**0.0498**	-0.1134	0.1940	-0.0286	0.2979		1.1290	0.2861	-0.0073	0.9595	0.0261	0.5304
				0.5037	0.2081	0.4015	0.4322	0.0514	0.7490		0.9277	0.7596	-0.4935	0.1244	-0.0352	0.7047
	rs3750344**	C	26.07%	1.0820	0.1916	0.0184	0.7070	-0.0011	0.9437	18.20%	0.9696	0.7719	0.1304	0.3359	0.0017	0.9657
				1.0940	0.2190	0.0093	0.8759	0.0018	0.9241		0.9633	0.7506	0.0758	0.6153	-0.0113	0.7953
				1.1260	0.4561	0.0802	0.5247	-0.0154	0.7003		0.9967	0.9931	0.9070	0.0662	0.1435	0.3144
GRIN3A	rs11788456**	G	45.09%	1.0250	0.6325	0.0342	0.4143	0.0107	0.4129	44.93%	0.9465	0.4919	-0.1141	0.2616	0.0062	0.8322
				1.0040	0.9634	0.0076	0.9067	0.0101	0.6189		0.9220	0.5034	-0.1382	0.3638	-0.0004	0.9922
				1.0720	0.4328	0.0944	0.1960	0.0199	0.3853		0.9415	0.6669	-0.1633	0.3562	0.0197	0.7005
	rs17189632**	A	36.72%	1.0230	0.6754	-0.0040	0.9271	0.0003	0.9834	43.59%	0.9545	0.5551	0.0171	0.8605	0.0078	0.7820
				1.0580	0.4468	0.0411	0.5000	-0.0001	0.9939		0.8205	0.0964	0.0320	0.8301	0.0129	0.7651
				0.9725	0.7936	-0.1028	0.2447	0.0015	0.9580		1.1410	0.3539	0.0112	0.9487	0.0074	0.8835
DNM1	rs3003609**	T	11.29%	0.9030	0.1943	-0.0137	0.8359	0.0072	0.7312	54.62%	0.9501	0.5035	0.1516	0.1313	-0.0079	0.7862
				0.8964	0.2182	-0.0431	0.5613	0.0098	0.6775		1.0740	0.5487	0.1465	0.3366	0.0169	0.7018
				0.8204	0.4679	0.2591	0.2697	-0.0071	0.9242		0.7818	0.0658	0.2735	0.1220	-0.0474	0.3553
DBH	**rs3025343**	A	2.03%	NA						10.37%	1.1490	0.2840	0.3652	**0.0231**	0.0223	0.6318
											1.1710	0.2558	0.3800	**0.0274**	0.0288	0.5648

Gene	dbSNP ID	A1	AA Sample							EA Sample						
			Freq (A1)	Smoking Status		FTND		Indexed CPD		Freq (A1)	Smoking Status		FTND		Indexed CPD	
				OR	*P*†	BETA	*P*	BETA	*P*		OR	*P*	BETA	*P*	BETA	*P*
											1.0380	0.9485	0.6212	0.3881	-0.0541	0.7953
	rs6271	T	1.58%			NA				6.39%	1.0600	0.7096	0.2012	0.6595	0.0301	0.6056
											1.0680	0.6980	0.0596	0.7814	0.0300	0.6299
											1.0400	0.9539	0.8229	0.3871	0.0845	0.7592
CHAT	**rs1880676***	A	4.91%			NA				23.21%	1.0830	0.3728	0.2065	0.0733	-0.0174	0.6023
											1.1620	0.1841	0.2805	**0.0496**	-0.0316	0.4452
											0.9158	0.6913	0.1466	0.6168	0.0205	0.8094
	rs3810950*	A	4.89%			NA				23.19%	1.0780	0.4032	0.2048	0.0751	-0.0147	0.6581
											1.1550	0.2037	0.2793	0.0502	-0.0279	0.4994
											0.9110	0.6738	0.1420	0.6276	0.0220	0.7947
LOC100188947	**rs1329650**	T	9.50%	1.1790	0.0542	-0.0783	0.2472	-0.0297	0.1626	26.86%	0.9893	0.9012	-0.1778	0.1154	-0.0535	0.0991
				1.1500	0.1370	-0.0625	0.4112	-0.0317	0.1849		1.0580	0.6096	-0.3308	**0.0194**	-0.0974	**0.0176**
				2.2250	**0.0263**	-0.3654	0.1291	-0.0586	0.4400		0.7806	0.2213	0.1719	0.5190	0.0415	0.5915
	rs1028936	C	8.10%	1.1810	0.0755	-0.0794	0.2837	-0.0263	0.2617	18.32%	1.0390	0.6931	-0.1461	0.2415	-0.0651	0.0683
				1.1690	0.1150	-0.0821	0.3025	-0.0295	0.2407		1.1100	0.3812	-0.2175	0.1464	-0.0956	**0.0270**
				1.9790	0.1554	-0.1667	0.6147	-0.0144	0.8907		0.8040	0.3967	0.0258	0.9393	0.0006	0.9950
BDNF	**rs6265***	T	3.13%			NA				14.36%	0.8439	0.1320	0.3437	**0.0231**	0.0801	0.0652
											0.8594	0.2405	0.4137	**0.0138**	0.1015	**0.0355**
											0.5601	0.1216	0.1311	0.8118	-0.0266	0.8666
	rs6484320*	T	7.48%	1.1010	0.3154	-0.0831	0.2876	0.0352	0.1541	18.11%	0.9429	0.5600	0.3663	**0.0061**	0.0837	**0.0306**
				1.1150	0.2894	-0.1223	0.1432	0.0263	0.3196		1.0050	0.9694	0.4171	**0.0052**	0.1118	**0.0101**
				1.0310	0.9445	0.5189	0.1542	0.2646	**0.0213**		0.5652	0.0674	0.3897	0.3863	-0.0557	0.6706
	rs66866077	T	1.27%			NA				5.84%	0.7953	0.1616	-0.0212	0.9234	-0.0700	0.2719
											0.7974	0.1964	-0.0398	0.8647	-0.0852	0.2077
											0.5011	0.3700	0.3783	0.7300	0.1384	0.6628
	rs2030324*	G	47.29%	1.0550	0.2923	-0.0003	0.9937	-0.0068	0.6026	49.27%	1.0900	0.2784	-0.2558	**0.0123**	-0.0499	0.0915
				1.0170	0.8372	-0.0020	0.9765	-0.0126	0.5483		1.2300	0.1036	-0.1508	0.3683	-0.0333	0.4918
				1.1440	0.1184	0.0012	0.9858	-0.0054	0.8082		1.0150	0.9076	-0.5121	**0.0017**	-0.0964	**0.0419**
	rs7934165*	A	47.22%	1.0610	0.2477	-0.0009	0.9836	-0.0066	0.6139	49.12%	1.0930	0.2602	-0.2639	**0.0099**	-0.0555	0.0604
				1.0120	0.8830	0.0040	0.9524	-0.0115	0.5828		1.2350	0.0958	-0.1835	0.2737	-0.0433	0.3699
				1.1690	0.0717	-0.0068	0.9224	-0.0060	0.7867		1.0190	0.8837	-0.5037	**0.0021**	-0.1016	**0.0325**
ANKK1	**rs111789052**	G	6.09%	1.0900	0.4104	0.0822	0.3210	0.0365	0.1630	0.40%				NA		

Gene	dbSNP ID	A1	AA Sample							EA Sample						
			Freq (A1)	Smoking Status		FTND		Indexed CPD		Freq (A1)	Smoking Status		FTND		Indexed CPD	
				OR	*P*†	BETA	*P*	BETA	*P*		OR	*P*	BETA	*P*	BETA	*P*
				1.0380	0.7420	0.0654	0.4773	0.0386	0.1850							
				3.6440	**0.0246**	0.4277	0.1770	0.0774	0.4394							
	rs115800217	A	10.26%	0.9977	0.9784	0.0805	0.2429	0.0209	0.3376	0.94%	NA					
				0.9953	0.9591	0.0908	0.2272	0.0240	0.3128							
				1.0270	0.9364	0.0688	0.8057	0.0120	0.8920							
	rs11604671**	A	10.92%	1.1900	**0.0288**	0.0379	0.5503	-0.0089	0.6567	42.70%	1.0020	0.9801	0.1154	0.2540	0.0117	0.6869
				1.2270	**0.0225**	0.0190	0.7903	-0.0113	0.6185		1.1460	0.2416	0.1257	0.4012	0.0282	0.5149
				1.1730	0.5666	0.2800	0.2067	-0.0005	0.9938		0.8294	0.1776	0.1925	0.2940	-0.0033	0.9506
	rs78229381	T	6.16%	1.0530	0.6209	0.0452	0.5879	0.0326	0.2187	0.58%	NA					
				0.9967	0.9763	0.0600	0.5141	0.0329	0.2614							
				4.6380	**0.0181**	-0.0647	0.8444	0.0877	0.4031							
	rs2734849**	G	16.71%	1.0810	0.2446	0.0760	0.1614	-0.0059	0.7307	43.24%	0.9899	0.8949	0.1162	0.2521	0.0099	0.7336
				1.0760	0.3481	0.0748	0.2438	-0.0087	0.6693		1.1130	0.3618	0.1285	0.3926	0.0269	0.5358
				1.2460	0.2766	0.1924	0.2257	0.0025	0.9598		0.8336	0.1891	0.1892	0.3014	-0.0072	0.8920
DRD2	rs2075654*	T	4.25%	NA						19.73%	1.0110	0.9106	-0.1032	0.4016	0.0262	0.4610
											0.1162	0.9998	-0.1942	0.1879	0.0095	0.8239
											1.0910	0.7519	0.2327	0.4923	0.1488	0.1282
	rs4586205*	T	35.61%	0.9875	0.8098	0.0062	0.8853	0.0260	0.0541	71.86%	1.0160	0.8537	-0.0336	0.7708	-0.0369	0.2507
				0.9325	0.3370	-0.0248	0.6797	0.0224	0.2371		1.1130	0.3325	-0.0314	0.8233	-0.0453	0.2668
				1.1010	0.3632	0.0761	0.3774	0.0592	**0.0293**		0.7802	0.2015	-0.0714	0.7852	-0.0490	0.5184
CHRNA5	rs588765	T	29.46%	1.0000	0.9931	0.0414	0.3713	0.0080	0.5843	38.84%	1.1000	0.2285	0.1293	0.2057	0.0211	0.4681
				0.9838	0.8209	0.0627	0.2907	0.0135	0.4702		1.1220	0.3123	0.2673	0.0674	0.0710	0.0930
				1.0580	0.6662	0.0171	0.8727	-0.0013	0.9698		1.1580	0.3368	0.0032	0.9864	-0.0444	0.4184
	rs16969968*	A	6.01%	1.1800	0.1180	-0.0513	0.5375	-0.0082	0.7536	29.51%	1.0360	0.6844	0.0560	0.6047	0.0075	0.8138
				1.1610	0.1837	-0.0613	0.4959	-0.0078	0.7830		1.0540	0.6330	0.0500	0.7229	0.0150	0.7127
				2.2940	0.1330	0.0209	0.9541	-0.0296	0.7958		1.0160	0.9378	0.1391	0.5803	-0.0086	0.9057
CHRNA3	rs578776*	A	53.67%	1.0150	0.7677	0.0363	0.3843	0.0062	0.6402	34.81%	0.9206	0.3072	-0.0887	0.4040	-0.0060	0.8416
				1.0060	0.9386	0.0417	0.5233	0.0117	0.5724		0.9058	0.3757	-0.0789	0.5803	-0.0022	0.9582
				1.0380	0.6724	0.0566	0.4279	0.0041	0.8541		0.8797	0.4388	-0.1892	0.3768	-0.0206	0.7398
	rs1051730*	A	12.81%	1.1920	**0.0197**	-0.0279	0.6364	-0.0085	0.6486	30.20%	1.0460	0.6006	0.0491	0.6492	0.0051	0.8728
				1.1430	0.1138	-0.0508	0.4580	-0.0137	0.5259		1.0650	0.5671	0.0479	0.7349	0.0099	0.8088
				2.4450	**0.0016**	0.0948	0.6060	0.0162	0.7796		1.0360	0.8585	0.1079	0.6638	-0.0045	0.9497

Gene	dbSNP ID	A1	AA Sample							EA Sample						
			Freq (A1)	Smoking Status		FTND		Indexed CPD		Freq (A1)	Smoking Status		FTND		Indexed CPD	
				OR	*P*†	BETA	*P*	BETA	*P*		OR	*P*	BETA	*P*	BETA	*P*
	rs6495308*	C	29.74%	0.9000	0.0571	-0.0277	0.5503	0.0016	0.9115	29.26%	0.9437	0.4918	-0.2183	**0.0452**	-0.0357	0.2525
				0.8875	0.0972	0.0097	0.8705	0.0017	0.9274		0.9130	0.4108	-0.1776	0.2129	-0.0335	0.4146
				0.8365	0.1558	-0.1829	0.0907	0.0032	0.9257		0.9745	0.8915	-0.5619	**0.0192**	-0.0806	0.2447
CHRNA4	rs2236196	A	35.26%	1.0240	0.6456	-0.0342	0.4264	-0.0027	0.8441	73.67%	1.0040	0.9668	0.1344	0.2460	-0.0115	0.7329
				0.9679	0.6554	-0.0025	0.9668	-0.0042	0.8258		1.0880	0.4486	0.1550	0.2752	-0.0151	0.7126
				1.1840	0.1134	-0.1327	0.1242	-0.0022	0.9348		0.7324	0.1615	0.2030	0.4998	-0.0085	0.9219
	rs2273504	A	15.84%	0.9493	0.4590	-0.0753	0.1884	-0.0256	0.1591	17.85%	1.0810	0.4485	0.0235	0.8496	0.0023	0.9526
				0.9380	0.4249	-0.0295	0.6523	-0.0150	0.4686		1.1390	0.2754	0.0267	0.8591	-0.0006	0.9886
				0.9694	0.8911	-0.5553	**0.0028**	-0.1480	**0.0120**		0.8394	0.5758	0.0449	0.9127	0.0269	0.8218

A1 allele 1; *Freq (A1)* frequency of allele 1; *OR* odds ratio; *BETA* beta coefficient; NA not applicable; † *P* values from top to bottom for each SNP were obtained under additive, dominant and recessive genetic model, respectively; */** SNPs showing nominal/significant associations in previous studies of MSTF samples; ‡/‡‡ SNPs showing nominal/significant associations in previous meta-analysis including MSTCC samples (please refer to **Table 3.2** for publication references). This table shows individual variant-based analysis results of SNPs with minor allele frequency (MAF) > 5%. Nominally significant associations ($P < 0.05$) are given in bold, including *P* values, SNP and gene names.

Supplementary Table 3: Non-significant rare variant association results using the weighted sum statistic (WSS) in AA and EA samples.

Gene	AA sample		EA sample	
	SNP	Permuted *P* value	SNP	Permuted *P* value
CHRNB3	- (p.H410Y) rs35327613 (p.K451E)	0.1009	NA	
CHAT	rs3810950 (p.A120T) rs75011234 (p.E188G) rs8178990 (p.L243F) rs868749 (p.P299L)	0.0611	rs868749 (p.P299L) rs75011234 (p.E188G) rs8178990 (p.L243F) rs146236256 (p.G284S)	0.0887
CHRNA10	rs2231548 (p.R421C) rs139793380 (p.R351W) rs55719530 (p.T77N) rs147150654 (p.L348R) rs2231542 (p.V248L) - (p.W86G) rs77958837 (p.E85G)	0.3839	rs139793380 (p.R351W) rs55719530 (p.T77N) rs2231548 (p.R421C) - (p.W86G)	0.6766
BDNF	rs6265 (p.V74M) rs66866077 (p.E6K)	0.4503	NA	
CHRNA5/CHRNA3/ CHRNB4	rs72650603 (p.H217Y) rs8192475 (p.R37H) - (p.R497C) - (p.F462V) rs56235003 (p.R349C) - (p.P145A) rs12914008 (p.T91I) rs75495090 (p.N41S) rs2229961 (p.V134I) rs80087508 (p.K167R) rs56218866 (p.S140G)	0.5565	rs2229961 (p.V134I) rs72650603 (p.H217Y) rs8192475 (p.R37H) rs80087508 (p.K167R) rs56235003 (p.R349C) rs56218866 (p.S140G) rs12914008 (p.T91I) rs75495090 (p.N41S)	0.2363

Permuted *P* value = *P* value based on 10^6 permutations; - = not reported in dbSNP database by 2/17/2014; NA = not applicable, i.e., without two rare nonsynonymous variants in the gene. SNPs included in both AA and EA rare variant analyses are underlined. See the section of 'Materials and methods' for details.

Supplementary Table 4: Non-significant combined and adaptive sum test results of cumulative rare- and common-variant effects on smoking status in AA and EA samples.

Gene	AA Sample					EA Sample				
	Rare variant(s)	Common variant(s)	Effect direction	*P* value†		Rare variant(s)	Common variant(s)	Effect direction	*P* value†	
				Separated	Pooled				Separated	Pooled
DDC		rs1451371	↑	0.7211	0.2125		rs1451371	↑	0.5797	0.3459
		rs921451		0.3975	0.2399		rs921451		0.1814	0.2291
				0.7211	0.1795				0.5797	0.2840
				0.3975	0.1790				0.1814	0.2724
	rs11575292 (p.E61D)	rs3735273	↓	0.1467		rs11575292 (p.E61D)	rs3735273	↓	0.3594	
				0.1467					0.3594	
				0.1975					0.2764	
				0.1975					0.2764	
TAS2R38	rs139843932 (p.W135G)		↑	0.1421	0.3060		rs1726866 (p.V262A)	↑	0.3712	0.0761
				0.0514	0.1397				0.3712	0.1083
	rs114288846 (p.R274C)			0.1421	0.2603				0.3712	0.0938
				0.0514	0.1015				0.3712	0.1341
		rs1726866 (p.V262A)	↓	0.7090		rs139843932 (p.W135G)		↓	**0.0123**	
				0.7090					**0.0110**	
				0.7090		rs114288846 (p.R274C)			**0.0123**	
				0.7090					**0.0110**	
CHRNA2		rs2472553 (p.T22I)	↑	0.7734	0.7077		rs2472553 (p.T22I)	↑	0.1469	0.1792
				0.7734	0.4655				0.1469	0.1792
				0.7734	0.7344				0.1469	0.2721
				0.7734	0.4096				0.1469	0.2721
	- (p.S488*)		↓	0.4845		- (p.S488*)		↓	0.3695	
	- (p.R121L)			0.2291					0.3695	
				0.4845					0.3695	
				0.2291					0.3695	
CHRNB3	- (p.H410Y)	rs10958726	↑	0.3497	0.4704	rs35327613 (p.K451E)		↑	0.8075	0.1680
		rs4736835		0.4017	0.6462				0.8075	0.1678
		rs6474412		0.3765	0.4282				0.8075	0.1126
		rs4950		0.4447	0.5934				0.8075	0.1126
		rs13280604								

Gene	AA Sample					EA Sample				
	Rare variant(s)	Common variant(s)	Effect direction	*P* value†		Rare variant(s)	Common variant(s)	Effect direction	*P* value†	
				Separated	Pooled				Separated	Pooled
		rs6474415								
	rs35327613 (p.K451E)	rs10958725	↓	0.8252 0.8252 0.8788 0.8788			rs10958725 rs10958726 rs4736835 rs6474412 rs4950 rs13280604 rs6474415	↓	0.0565 0.0565 0.0565 0.0565	
NTRK2	rs150692457 (p.L140F) - (p.C623*)	rs1187272	↓	NA	0.2961 0.2337 0.2408 0.1821			NA		
GABBR2	- (p.P742Q) - (p.G671C)	rs2491397 rs2184026 rs3750344 (p.A120A)	↑	NA	0.3785 0.1201 0.3757 0.1220			NA		
GRIN3A	- (p.V132L)	rs11788456 rs17189632	↑	0.5704 0.4206 0.4422 0.4419	0.3566 0.1328 0.2359 0.1094	rs75981117 (p.N549S)		↑	0.7562 0.7562 0.7562 0.7562	0.6119 0.3223 0.6852 0.4677
	rs75981117 (p.N549S) rs34755188 (p.R480H) - (p.V389L)		↓	0.1323 **0.0303** 0.1323 **0.0303**		rs34755188 (p.R480H) - (p.V389L) - (p.V132L)	rs11788456 rs17189632	↓	0.5010 0.2041 0.4942 0.3291	
DNM1	- (p.S126*) - (p.R228L)		↑	**0.0480** **0.0154** **0.0480** **0.0154**	0.1079 0.2385 0.1824 0.3505	rs61757224 (p.L16M)		↑	0.7267 0.7267 0.7267 0.7267	0.7977 0.7977 0.7994 0.7994

Gene	AA Sample					EA Sample				
	Rare variant(s)	Common variant(s)	Effect direction	*P* value[†]		Rare variant(s)	Common variant(s)	Effect direction	*P* value[†]	
				Separated	Pooled				Separated	Pooled
	rs61757224 (p.L16M) - (p.Y231*)	rs3003609 (p.F336F)	↓	0.3497 0.2414 0.3504 0.3502			rs3003609 (p.F336F)	↓	0.5611 0.5611 0.5611 0.5611	
DBH			NA			rs41316996 (p.G482R)	rs3025343 rs6271 (p.R549C)	↑	0.5497 0.4514 0.6736 0.5449	0.4944 0.6153 0.5923 0.5454
						rs182974707 (p.I340T) rs75215331 (p.A362V)		↓	0.1593 **0.0451** 0.1593 **0.0451**	
CHAT			NA			rs868749 (p.P299L)	rs1880676 rs3810950 (p.A120T)	↑	0.4681 0.4675 0.6125 0.6127	0.6154 0.4470 0.6476 0.4995
						rs75011234 (p.E188G) rs8178990 (p.L243F) rs146236256 (p.G284S)		↓	0.5901 0.2291 0.5901 0.2291	
BDNF	rs6265 (p.V74M)	rs6484320 rs2030324 rs7934165	↑	0.3884 0.1876 0.5011 0.1973	0.4071 0.1192 0.5012 0.1993			NA		
	rs66866077 (p.E6K)		↓	0.3421 0.3421 0.3421 0.3421						

Gene	AA Sample					EA Sample				
	Rare variant(s)	Common variant(s)	Effect direction	*P* value†		Rare variant(s)	Common variant(s)	Effect direction	*P* value†	
				Separated	Pooled				Separated	Pooled
CHRNA4	rs201739273 (p.P457L)	rs2236196	↑	0.8791	0.8558			NA		
				0.8791	0.9640					
				0.8762	0.8623					
				0.8762	1.0000					
		rs2273504	↓	0.4605						
				0.4605						
				0.4605						
				0.4605						

NA Not Applicable. *P* values for each gene or region were obtained from four different statistical methods, i.e., SKAT-C, Burden-C, SKAT-A and Burden-A; † = P values from top to bottom for each gene or region were obtained in the abovementioned order. Only genes or regions with at least one rare variant and one common variant were eligible to be included in the pooled analysis. ↑ = Variants increased estimated smoking risk according to individual variant-based odds ratios (if available) or minor allele counts in Cases and Controls, ↓ = Variants decreased smoking risk; effect direction specific tests were applied with *P* values listed under "Separated". SNP rs numbers are based on dbSNP database (accessed on 2/17/2014). SNPs included in both AA and EA samples for this analysis are underlined. Nominal significant associations ($P < 0.05$) for "Separated" analysis with only rare variants are given in bold. See the section of 'Materials and methods' for details.

Chapter 4

Nicotine dependence susceptibility map

4.1 Abstract

Experimental approaches to genetic studies of complex traits evolve with technological advancements. How do discoveries using different approaches advance our knowledge of the genetic architecture underlying complex diseases/traits of interest? Do most of the findings of "newer" techniques; e.g., genome-wide association study (GWAS), provide more information than "older" ones, such as genome-wide linkage study? In this review, we address these issues by developing a nicotine dependence (ND) genetic susceptibility map based on the results obtained by the approaches commonly used in recent years, which include genome-wide linkage, candidate gene association, GWAS, and targeted sequencing studies. Converging and diverging results from these empirical approaches have elucidated a preliminary genetic architecture of this intractable psychiatric disorder and yielded new hypotheses on ND etiology.

The insights we obtained by putting together results from diverse approaches can be applied to other complex diseases/traits. In sum, developing a genetic susceptibility map and keeping it updated is an effective way to keep track of what we know about a disease/trait and what the next steps might be with new approaches.

4.2 Introduction

Along with technological advancements, experimental approaches for the genetic study of complex diseases/traits have evolved, from genome-wide linkage study to candidate gene association study, and from genome-wide association study (GWAS) to targeted sequencing. With improvements in accuracy, coverage, and cost, whole-exome and whole-genome sequencing studies seem to be the next mainstream approaches. Are the discoveries from all of these approaches consistent with one another? Should we focus on results obtained from "newer" approaches; e.g., GWAS, and abandon findings from "older" ones, such as genome-wide linkage study, in the literature sea? How can we make the findings guide our understanding of the genetic architecture of the disease/trait in question? In this review, we use nicotine dependence (ND) as an example to investigate these issues.

Tobacco smoking poses significant threats to public health and kills more than 6 million people annually worldwide, making it one of the three leading components of the global disease burden in 2010.[1] Despite 50 years of prevention efforts, smoking remains the greatest cause of preventable diseases and deaths; each year, nearly 500,000 Americans die prematurely from smoking, and more than 16 million Americans suffer from a disease caused by smoking. Even though today's users smoke fewer cigarettes than those 50 years ago, they

are at higher risk of developing adenocarcinoma, possibly because of ventilated filters and greater amounts of tobacco-specific nitrosamines in cigarettes.[2]

Since the 1980s, a broad scientific consensus has been established that nicotine dependence (ND) is the primary factor maintaining smoking behavior.[5] We and others have shown strong evidence for the involvement of genetics in ND, with an average heritability of 0.56.[4, 154] In the past dozen years, considerable efforts have been undertaken to identify the genetic factors underlying ND. However, only three widely accepted "successes;" i.e., the neuronal nicotinic acetylcholine receptor gene clusters on chromosomes 15 (*CHRNA5*/*A3*/*B4*)[19, 21, 23-25, 27-30, 132, 133, 155-157] and 8 (*CHRNB3*/*A6*)[19, 24, 33, 132, 158-161] and the genes encoding nicotine-metabolizing enzymes on chromosome 19 (*CYP2A6*/*A7*),[23, 24, 162-164] meet community standards for significance and replication.[165] These few triumphs stand in contrast to the limited heritability they explain; e.g., the most significant synonymous single-nucleotide polymorphism (SNP) rs1051730 ($P = 2.75 \times 10^{-73}$) in *CHRNA3* accounted for only 0.5% of the variance in cigarettes smoked per day (CPD) in a meta-analysis of 73,853 subjects.[23] Researchers have suggested that "missing heritability" is merely hidden and additional loci can be discovered in GWAS with larger samples,[166, 167] not to mention that the largest ND GWAS to date included 143,023 subjects,[23] and lots of relevant genetic loci have been revealed with other experimental approaches, such as genome-wide linkage, hypothesis-driven candidate gene association, and targeted sequencing studies. Despite the fact that many non-GWAS findings have an uncertain yield or failed to be replicated, sorting out genetic loci with evidence from multiple approaches is not only essential but also more cost effective than pursuing a formidable sample size for GWAS.

In this communication, we first review the literature on ND genetics (all smoking-related phenotypes were included) using different approaches by highlighting the converging results from different approaches and then offer new hypotheses that have emerged across the allelic spectrum, including common and rare variants. These findings provide insights into the preliminary genetic architecture of ND: data that are essential for guiding future research. Crucially, we show that developing a genetic susceptibility map with data from various approaches is an effective way for knowledge integration, research progress evaluation, and research direction forecast.

4.3 Genome-wide linkage studies

For many years, linkage analysis was the primary approach for the genetic mapping of both Mendelian and complex traits with familial aggregation.[168, 169] However, it was largely supplanted by the wide adoption of GWAS since the middle 2000s. In 2008, we published a comprehensive review on more than 20 published genome-wide linkage studies for smoking behavior and identified 13 regions, located on chromosomes 3–7, 9–11, 17, 20, and 22, suggestively or significantly linked with various ND measurements in at least two independent samples.[9] Since then, only one genome-wide linkage study has been reported, by Hardin *et al*.,[170] and the same linkage region (6q26) was detected as in their previous analysis using the same sample, but with a different phenotype.[171] In addition, Han *et al*.[172] conducted a meta-analysis of 15 genome-wide linkage scans of smoking behavior and identified two suggestive (5q33.1–5q35.2 and 17q24.3–q25.3) and one significant (20q13.12–q13.32) linkage region. In fact, the regions on chromosomes 5 and 20 expand two of the regions reported in our 2008

review; the region on chromosome 17 reported by Han *et al*.[172] verified one of the regions detected only in one sample before 2008, which makes a new nominated linkage peak (**Table 4.1**).[9] **Figure 4.1** shows updated linkage results after incorporating the findings reported after 2008 by Han *et al*.[172]

Table 4.1: Information on the nominated linkage regions updated based on Li.[9]

Chr.	Marker or marker region	Position	Chr. bands	Phenotype
3	D3S1763-D3S1262	167,239,681-186,223,727	3q26-q27	DSM-IV ND, SQ
4	D4S403-D4S2632, D4S244	13,750,828-65,491,728	4p15-q13.1	FTND, CPD
5 (Region 1)	D5S1969, D5S647, D5S428	53,242,832-85,410,963	5q11.2-q14	SQ, smoking status, FTND
5 (Region 2)*	D5S400, D5S1354	149,800,001-179,631,902	5q33.1-q35*	FTND, CPD
6	D6S1009, D6S1581-D6S281, D6S446	137,302,085-170,552,657	6q23.3-q27	smoking status, FTND, withdrawal severity
7	D7S486, D7S636	115,894,675-150,699,599	7q31.2-q36.1	FTND, DSM-IV
9 (Region 1)	D9S2169-D9S925, D9S925-D9S319	5,200,390-29,560,115	9p21-p24.1	FTND, HSI, SQ
9 (Region 2)	D9S257-D9S910, D9S283, D9S64, D9S1825	90,290,735-127,888,281	9q21.33-q33	SQ, FTND, smoking status
10	D10S1432, D10S2469/CYP17, D10S597, D10S1652-D10S1693, D10S129-D10S217	64,407,495-129,540,525	10q21.2-q26.2	SQ, FTND, smoking status
11	D11S4046, D11S4181, D11S2362-D11S1981, D11S1999-D11S1981, D11S2368-D11S2371, D11S1392-D11S1344, D11S1985-D11S2371	1,963,635-73,505,374	11p15-q13.4	FTND, SQ
17 (Region 1)	GATA193, D17S974-D17S2196, D17S799-D17S2196, D17S799-D17S1290	10,518,666-56,331,730	17p13.1-q22	CPD, SQ, HSI
17 (Region 2)*	D17S968	67,100,001-81,195,210	17q24.3-q25.3*	smoking status
20*	D20S119-D20S178, D20S481-D20S480	43,648,850-58,400,000	20q13.12-q13.32*	CPD, SQ
22	D22S345-D22S315, D22S315-D22S1144	24,488,587-27,683,302	22q11.23-12.1	CPD, age at first cigarette

Chr. chromosome; *SQ* smoking quantity; *HSI* heaviness of smoking index. This table was modified based on Table 3 of Li.[9] * denotes linkage regions expanded or newly ascertained after evaluating results published after our 2008 review. Genomic positions for microsatellite markers and corresponding chromosome bands were obtained through the UCSC Genome Browser (http://genome.ucsc.edu/), which are in the GRCh37/hg19 assembly.

4.4 Hypothesis-driven candidate gene association studies

Candidate gene association studies usually have moderate sample size and are much cheaper than GWAS, where the genes examined are selected according to the linkage/GWAS study results or biological hypotheses. However, because of population heterogeneity issues and liberal statistical thresholds (compared with GWAS) that often are applied, hypothesis-driven candidate gene association studies generally are considered to have an uncertain yield.[173] On the other hand, the abundant results obtained from this approach provide more in-depth exploration of potential targets and offer valuable replication for the other unbiased approaches; e.g., genome-wide linkage study and GWAS.

To eliminate concerns about potential false-positive results, especially for studies reported in earlier years, we focused primarily on the genes showing significance in at least two independent studies with a sample size of ≥ 1,000 or within (or close to) nominated linkage regions or overlapping with GWAS results but with a sample size of ≥ 500 based on the statistical thresholds set by each study. Because the reported sex-average recombination rate is 1.30 ± 0.80 cM/Mbp,[174] here we defined candidate genes within 2 megabases (Mbp) of any linkage region as "within" and 2-5 Mbp as "close to". The sample size requirement was determined with the following parameters: two-tailed $\alpha = 0.05$, population risk of 0.30, minor allele frequencies of 0.20, genotypic relative risk of 1.3 with an approximate odds ratio (OR) of 1.5 or 0.7, which is similar to the statistics usually found in candidate gene association studies. For a statistical power of 0.80 ($\beta = 0.20$) using the allelic test, the minimum sample size for a case-control study is 1,062, with equal numbers of cases and controls. Of the reported 201 candidate gene association studies, only 88 have had a sample size of 1,000 or more.

Considering the detected power of 0.54 for a sample size of 500 under the dominant genetic model, we also included genes implicated in studies with 500-1,000 subjects, given that the genes were located in a nominated linkage peak[9] or overlapped with GWAS signals. In total, 34 genetic loci with 43 genes met the criteria (**Table 4.2** and **Figure 4.1**), which were assigned to the following four gene groups. For details on those studies that failed to pass the thresholds but show positive associations, please see **Supplementary Table 1**.

4.4.1 Neurotransmitter system genes

Dopaminergic system: The dopaminergic system has long been acknowledged to play a critical role in nicotine addiction.[175] The most studied gene in this system is *DRD2*, located on chromosome 11q23.2 within a modest linkage peak.[176] The intriguing polymorphism *Taq*1A is located in *ANKK1* near *DRD2*, leading to an amino acid change in *ANKK1*.[177] Several other variants and haplotypes in regions adjacent to *DRD2*, within *TTC12* and *ANKK1*, or downstream of *DRD2* have been associated with smoking-related phenotypes.[19, 108, 109, 178, 179] Besides *DRD2*, a modest number of studies have shown significant associations between ND traits and other dopamine receptor genes, such as *DRD1*[16] and *DRD4*,[180-182] and genes involved in dopamine metabolism, including dopamine β-hydroxylase (*DBH*),[19, 183, 184] DOPA decarboxylase (*DDC*),[119, 185] and catechol-O-methyl transferase (*COMT*).[15, 186-190] All of these genes are within or close to the nominated linkage peaks[9] except for *DBH* and *DDC*, which have received support from GWAS results[23] and as an ND-associated gene from two independent studies with sample sizes ≥ 1,000,[119, 185] respectively.

Huang *et al*.[17] implicated *DRD3* as a susceptibility gene for ND, but this result has not yet been replicated. Meanwhile, Stapleton *et al*.[191] showed a significant association of a dopamine transporter gene (*SLC6A3*) with smoking cessation in a meta-analysis of 2,155 subjects (80% of European ancestry), although this finding received only weak support from another study on age at smoking initiation in 668 Asians.[192] This gene group includes two others, protein phosphatase 1 regulatory subunit 1B (*PPP1R1B*) and μ-opioid receptor (*OPRM1*), on the basis of their functional connections with dopamine in studies of other addictive substances. *PPP1R1B*, also known as dopamine- and cAMP-regulated neuronal phosphatase (*DARPP-32*), encodes a key phosphoprotein involved in the regulation of several signaling cascades for dopaminoceptive neurons in several areas of the brain, which also is required for the biochemical effects of cocaine.[193] Activation of OPRM1 in the ventral tegmental area suppresses the activity of inhibitory GABAergic interneurons, resulting in disinhibition of dopamine neurons and dopamine release from terminals in the ventral striatum.[194] *OPRM1* A118G variation is a genetic determinant of the striatal dopamine response to alcohol in men,[194] with a preliminary study of tobacco smoking confirming this result.[195] Although we believe in the importance of the above-mentioned genes in ND based on rigorous scientific evidence, the inconsistent results are worth further examination.[196-200]

GABAergic and serotonergic systems: For the GABAergic system, variants in the $GABA_B$ receptor subunit 2 (*GABBR2*),[10] $GABA_A$ receptor-associated protein (*GABARAP*),[201] and $GABA_A$ receptor subunits alpha-2 (*GABRA2*) and -4 (*GABRA4*)[19, 202, 203] were significantly associated with different ND phenotypes. Cui *et al*.[204] reviewed the significance of the GABAergic system in ND and alcohol dependence. In addition, the serotonergic system is implicated in susceptibility to

ND because nicotine increases serotonin release in the brain, and symptoms of nicotine withdrawal are associated with diminished serotonergic neurotransmission.[205] Genes encoding serotonin receptor 3A, ionotropic (*HTR3A*),[70] 5A, G protein-coupled (*HTR5A*),[19] and serotonin transporter (*SLC6A4*)[206-208] showed a significant association with smoking-related behaviors. All of these seven genes of the GABAergic and serotonergic systems are within or close to the nominated linkage peaks,[9] which strengthens the validity of the identified associations, although two studies reported negative results for association between serotonin transporter gene (*SLC6A4*) and smoking behavior.[209, 210] Another gene worth mentioning for this group is serotonin receptor 2A, G protein-coupled (*HTR2A*), which is within a modest linkage peak (13q14) suggested by Li *et al*.[6] and was significantly associated with smoking status in a Brazilian sample of 625 subjects.[211] Replication in larger samples is needed to confirm association of this gene with ND.

Glutamatergic system and related genes: Two glutamate receptors, ionotropic, NMDA 3A (*GRIN3A*), within the nominated linkage peak on 9q21.33-q33,[9] and NMDA 2B (*GRIN2B*), suggested by one GWAS[212] and close to a modest linkage peak on 12p13.31-13.32,[7] were significantly associated with scores on the Fagerström Test for Nicotine Dependence (FTND).[213, 214] More genes in the glutamatergic system, such as *GRIN2A*, *GRIK2*, *GRM8*, and *SLC1A2*, showed suggestive association with smoking behavior in the GWAS reported by Vink *et al*.[212] but without significant replication in candidate gene association studies. Accumulating evidence suggests that blockade of glutamatergic transmission attenuates the positive reinforcing and incentive motivational aspects of nicotine, inhibits the reward-enhancing and conditioned

rewarding effects of nicotine, and blocks nicotine-seeking behavior.[215] More attention may be paid to this neurotransmitter system in the future.

In the catch-all part, after showing suggestive association in the first ND GWAS,[34] neurexin 1 (*NRXN1*) association has been replicated in two independent studies with more than 2,000 subjects of three ancestries: African, Asian, and European.[112, 216] Although neurexin 3 (*NRXN3*) also showed a significant association with the risk of being a smoker,[217] this finding has not been verified in any other ND samples, and *NRXN3* is not within any detected linkage peak.[9] Neurexins are cell-adhesion molecules that play a key role in synapse formation and maintenance and have been implicated in polysubstance addiction.[218]

4.4.2 Nicotinic receptor (nAChR) subunit and other cholinergic system genes

As nAChR subunit gene clusters on chromosomes 15 (*CHRNA5*/*A3*/*B4*) and 8 (*CHRNB3*/*A6*) are major discoveries from ND GWAS, their candidate association results will be discussed together with the GWAS results. Significant association of variants in two other subunit genes (*CHRNA4* and *CHRNB1*) did not approach genome-wide significance ($p < 5\times10^{-8}$), but they are both close to the nominated linkage peaks.[9] Association of *CHRNA4* with ND, close to the nominated linkage peak on 20q13.12–13.32,[9] has been demonstrated in five independent studies (**Table 4.2**).[14, 214, 219-221] Variants within *CHRNB1*, located close to the nominated linkage peak on 17p13.1-q22,[9] also are significantly associated with FTND and CPD scores.[214, 222] Two other genes encoding nAChR subunits, *CHRNB2* and *CHRNA2*, although associated with ND-related phenotypes in two studies,[223, 224] are not within any detected linkage peaks and have no replication studies reported with the required sample size. Thus, these two genes are

considered to have only weak evidence of involvement and therefore are not included in **Figure 4.1** and **Table 4.2**. Besides nAChR subunit genes, cholinergic receptors, muscarinic 1 (*CHRM1*) and 2 (*CHRM2*), were found to be significantly associated with CPD and FTND, respectively.[214, 222] They are within nominated linkage peaks as well.[9] However, because of the inadequacy of knowledge of their biological functions compared with other neuronal nAChR subunit genes, they have been less investigated.

4.4.3 Nicotine metabolism genes

Of the nicotine metabolism genes, those encoding nicotine-metabolizing enzymes (*CYP2A6* and *CYP2B6*) have been extensively investigated.[225] Six studies have provided consistent evidence that variants leading to reduced or absent CYP2A6 activity are associated with various smoking-related phenotypes, including the nicotine metabolite ratio,[226] time to relapse,[163] exhaled carbon monoxide (CO),[164] initial subjective response to nicotine,[208] FTND,[19] and CPD.[227] All six samples consisted of subjects of European descent (**Table 4.1**). The negative result of *CYP2A6* in the 2004 meta-analytic review contrasts with the findings from more recent studies, which we believe offer stronger statistical evidence.[228] Such significant association of variants in the *EGLN2-CYP2A6-CYP2B6* region with ND is corroborated by GWAS results, as discussed in the next section.[24, 162]

4.4.4 MAPK signaling pathway and other genes

Although space limitations do not permit an exhaustive review, we want to acknowledge studies implicating involvement of other genes in ND, these genes including brain-derived

neurotrophic factor (*BDNF*),[111, 229] neurotrophic tyrosine kinase, receptor type 2 (*NTRK2*),[11] arrestin, beta 1 (*ARRB1*),[12] *MAP3K4*,[214] *SHC3*,[230] dynamin 1 (*DNM1*),[126] taste receptor type 2, member 38 (*TAS2R38*),[18] amyloid beta precursor protein-binding, family B, member 1 (*APBB1*),[231] *PTEN*,[232] and neuregulin 3 (*NRG3*).[233] It is worth noting that the first five genes listed all belong to the MAPK signaling pathway, which was identified as significantly enriched in involvement with four drugs subject to abuse, namely, cocaine, alcohol, opioids, and nicotine.[234]

Table 4.2: Significant candidate gene association study results for ND-related phenotypes.

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
Neurotransmitter system genes										
Dopaminergic system										
TTC12-ANKK1-DRD2	11q23.2	Within the modest linkage peak on 11q23 (0 bp)[176]	638 (270 AAs+368 EAs)	rs2303380-rs4938015-rs11604671 (*TTC12*)-(*ANKK1*)-(*ANKK1*)			0.01	OR = 1.6	Regular smoking initiation	[178]
			752 (European)	rs1800497 (*Taq*1A)	113270828	Missense (*ANKK1*)	4.0×10^{-3} (interaction with sex and treatment)		Smoking cessation	[235]
			755 (European)	rs1800497 (*Taq*1A)	113270828	Missense (*ANKK1*)	0.04 (interaction with treatment)		Smoking cessation	[236]
			782 (European)	rs1799732 (-141C Ins/Del)	113346251:113346252	Near Gene-5 (*DRD2*)	0.01 (interaction with treatment)		Smoking cessation	[237]
			1,026 (European)	rs1800497 (*Taq*1A)	113270828	Missense (*ANKK1*)	$<1.0 \times 10^{-8}$		Smoking status	[238]
			1,615 (854 AAs+761 EAs)	rs4938012	113259654	5′-UTR (*ANKK1*)	8.0×10^{-6} (pooled)		DSM-IV ND	[108]
			1,900 (European and other)	rs1800497 (*Taq*1A)	113270828	Missense (*ANKK1*)	0.01 (interaction with ADHD symptoms)		Initial subjective response to nicotine	[208]
			1,929 (European)	rs4245150	113364647	Intergenic	0.01		FTND≥4 vs. FTND=0 in smokers	[19]
				rs17602038	113364691	Intergenic	0.01			
			2,037 (1,366 AAs+671 EAs)	rs2734849	113270160	Missense (*ANKK1*)	5.3×10^{-4} (AA)		HSI	[109]
			4,762 (European)	rs10502172	113199146	Intronic (*TTC12*)	9.1×10^{-6}	OR = 1.3	Smoking status	[179]
DRD1	5q35.2	Within the nominated linkage peak on 5q34-q35 (0 bp)	2,037 (1,366 AAs+671 EAs)	rs686	174868700	3′-UTR	4.8×10^{-3} (AA)		FTND	[16]

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
DRD4	11p15.5	Within the nominated linkage peak on 11p15-q13.4 (1.3 Mbp)	720 (European)	VNTR		Exon 3	0.03		Smoking cessation	180
			839 (59% European)	VNTR		Exon 3	2.0×10^{-3} (interaction with neuroticism)	OR = 3.5	Progression to ND	181
			2,274 (European)	VNTR		Exon 3	6.0×10^{-3}	$\beta = 0.1$	CPD	182
DBH	9q34.2	GWAS[23]	793 (European)	rs1541333	136511385	Intronic	4.0×10^{-4} (interaction with ND)		Smoking cessation	183
			1,608 (European)	rs3025382	136502321	Intronic	3.3×10^{-4}		FTND ≥4 vs.0 in smokers	214
			1,929 (European)	rs4531	136509370	Missense	5.1×10^{-3}		FTND ≥4 vs.0 in smokers	19
			2,521 [239]	rs5320	136507473	Missense	7.0×10^{-3} (male)		CPD	184
DDC	7p12.1		1,590 (854 AAs+736 EAs)	rs12718541	50550144	Intronic	2.0×10^{-4} (pooled)		FTND	185
			2,037 (1,366 AAs+671 EAs)	rs921451	50623285	Intronic	0.01 (EA)		CPD	119
COMT	22q11.21	Close to the nominated linkage peak on 22q11.23-q12.1 (4.5 Mbp)	511 (81 AAs+430 EAs)		rs737865-rs165599		4.3×10^{-3} (EA)		Smoking cessation	186
			614 (91% European)	rs4680	19951271	Missense	<0.05	OR = 2.1	Increased smoking	187
			657 (European)	rs4680	19951271	Missense	0.02 (male)		Smoking status	188
			2,037 (1,366 AAs+671 EAs)	rs4680	19951271	Missense	9.0×10^{-3} (EA)		CPD	15
			6,310 (European)	rs4680	19951271	Missense	3.0×10^{-3}	OR = 0.7	Smoking cessation	189
			13,312 (European)	rs4680	19951271	Missense	7.0×10^{-3} (meta)	OR = 1.1	Smoking status before pregnancy	190
PPP1R1B	17q12	Within the nominated	2,037 (1,366 AAs+671 EAs)	rs2271309-rs907094-rs3764352-rs3817160			0.01 (EA)		CPD	240

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
		linkage peak on 17p13.1-q22 (0 bp)								
OPRM1	6q25.2	Within the nominated linkage peak on 6q23.3-q27 (0 bp)	710 (European)	rs1799971	154360797	Missense	5.0×10^{-3} (interaction with sex)		Smoking cessation	241
			1,929 (European)	rs510769	154362019	Intronic	9.8×10^{-3}		FTND ≥4 vs. 0 in smokers	19
GABAergic system										
GABBR2	9q22.33	Within the nominated linkage peak on 9q21.33-q33 (0 bp)	1,276 (793 AAs+483 EAs)	rs1435252	101103591	Intronic	3.0×10^{-3} (EA)		CPD	10
				rs3750344	101340316	Synonymous	3.0×10^{-3} (EA)			
DLG4-GABARAP	17p13.1	Within the nominated linkage peak on 17p13.1-q22 (0 bp)	2,037 (1,366 AAs+671 EAs)	rs222843	7145981	nearGene-5 (*GABARAP*)	9.0×10^{-3} (EA)		FTND	201
GABRA2-GABRA4	4p12	Within the nominated linkage peak on 4p15-q13.1 (0 bp)	1,929 (European)	rs3762611	46997288	nearGene-5 (*GABRA4*)	9.0×10^{-4}	OR = 0.5	FTND ≥4 vs. 0 in smokers	19, 202, 203
Serotonergic system										
HTR3A	11q23.2	Within the modest linkage peak on 11q23 (0 bp)[176]	2,037 (1,366 AAs+671 EAs)	rs1150226-rs1062613-rs33940208-rs1985242-rs2276302-rs10160548			2.0×10^{-3} (AA)		HSI	70
HTR5A	7q36.2	Close to the nominated linkage peak on 7q31.2-q36.1 (4.2 Mbp)	1,929 (European)	rs6320	154862621	Synonymous	6.5×10^{-3}		FTND ≥4 vs. 0 in smokers	19
SLC6A4	17q11.2	Within the nominated linkage peak on	782 (European)	5-HTTLPR+intronic VNTR			1.0×10^{-4}	OR = 1.4	Smoking status	206

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
		17p13.1-q22 (0 bp)	1,098 (41% European)	5-HTTLPR			$<1.0 \times 10^{-3}$ (interaction with peer smoking)	HR = 5.7	Regular smoking initiation	207
			1,900 (European and other)	5-HTTLPR			0.02 (interaction with ADHD symptoms)		Initial subjective response to nicotine	208
Glutamatergic system & other										
GRIN3A	9q31.1	Within the nominated linkage peak on 9q21.33-q33 (0 bp)	2,037 (1,366 AAs+671 EAs)	rs17189632	104368002	Intronic	2.0×10^{-4} (pooled)		FTND	213
GRIN2B	12p13.1	GWAS,[212] close to the modest linkage peak on 12p13.31-13.32 (3.6 Mbp)[7]	1,608 (European)	rs17760877	13819473	Intronic	1.5×10^{-5} (interaction with age of onset)		FTND ≥4 vs. 0 in smokers	214
NRXN1	2p16.3	GWAS[34]	2,037 (1,366 AAs+671 EAs)	rs6721498	50713012	Intronic	8.6×10^{-6} (AA)		FTND	112
			2,516 [239]	rs2193225	51079482	Intronic	6.0×10^{-3}		Smoking status	216
Nicotinic receptor (nAChR) subunit & other cholinergic system genes										
CHRNA5-CHRNA3-CHRNB4	15q25.1	GWAS[21, 23-25, 27]	516 (European)	rs1051730	78894339	Synonymous (*CHRNA3*)	3.0×10^{-6}	$\beta = 0.3$	Cotinine level	156
			965 (European)	rs578776	78888400	3′-UTR (*CHRNA3*)	8.0×10^{-3}		Neural response	242
			1,073 (European)	rs16969968-rs680244 (*CHRNA5*)-(*CHRNA5*)			2.7×10^{-3} (interaction with treatment)	OR = 3.1	Smoking cessation	243
			1,030 (European)	rs1051730	78894339	Synonymous (*CHRNA3*)	4.0×10^{-3}	$\beta = 0.1$	CPD	244
			1,075 (775 EAs+169 Hispanics+131 others)	rs1948	78917399	3′-UTR (*CHRNB4*)	$<1.0 \times 10^{-3}$	HR = 1.3	Age of initiation	245

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
			1,118 (European)	rs3743078	78894759	Intronic (*CHRNA3*)	1.0×10^{-4} (ADHD patients)	OR = 1.8	Smoking status	[246]
			1,450 (European)	rs16969968	78882925	Missense (*CHRNA5*)	2.0×10^{-3} (adolescents who tried smoking before 18)	OR = 2.4	Smoking status	[247]
			1,608 (European)	rs578776	78888400	3′-UTR (*CHRNA3*)	3.8×10^{-4}		FTND ≥4 vs. 0 in smokers	[214]
			1,689 (European)	rs1051730	78894339	Synonymous (*CHRNA3*)	0.01 (meta)	OR = 0.8	Smoking cessation	[248]
			1,929 (European)	rs578776	78888400	3′-UTR (*CHRNA3*)	1.1×10^{-4}	OR = 1.3	FTND ≥4 vs. 0 in smokers	[19, 249]
			1,936 (815 discovery+1,121 replication)	rs16969968	78882925	Missense (*CHRNA5*)	2.8×10^{-3} (replication)		FTND	[250]
			2,038 (European)	rs16969968	78882925	Missense (*CHRNA5*)	7.7×10^{-3} (interaction with peer smoking)		FTND ≥4 vs. 0 in smokers	[251]
			2,047 (European)	rs16969968	78882925	Missense (*CHRNA5*)	5.2×10^{-8}	$\beta = 0.2$	FTND	[227]
			2,206 (European)	rs16969968	78882925	Missense (*CHRNA5*)	4.4×10^{-3} (interaction with childhood adversity in male)	OR = 1.8	DSM-IV ND	[252]
			2,284 (European)	rs17487223	78923987	Intronic (*CHRNB4*)	1.0×10^{-3}		Habitual smoking	[133]
			2,474 (European)	rs1051730	78894339	Synonymous (*CHRNA3*)	3×10^{-4}	OR = 1.3	CPD during pregnancy	[253]
			2,633 (European)	rs1051730	78894339	Synonymous (*CHRNA3*)	<0.01 (NRT at 6 months)	OR = 2.5	Smoking cessation	[254]

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
			2,772 (710 AAs+2,602 EAs)	rs16969968	78882925	Missense (*CHRNA5*)	4.5×10^{-8}	OR = 1.4	FTND ≥4 vs. 0 in smokers	26
			2,827 (European)	rs680244-rs569207-rs16969968-rs578776-rs1051730 (*CHRNA5*)-(*CHRNA5*)-(*CHRNA5*)-(*CHRNA3*)-(*CHRNA3*)			2.0×10^{-5} (age of daily smoking ≤16)	OR = 1.8	FTND≤4 vs. FTND≥6	157
			2,847 (European)	rs3743078	78894759	Intronic (*CHRNA3*)	5.0×10^{-9}	OR = 0.7	Heavy vs. light smokers	255
				rs11637630	78899719	Intronic (*CHRNA3*)	5.0×10^{-9}	OR = 0.7		
			4,150 (European)	rs1051730	78894339	Synonymous (*CHRNA3*)	5.7×10^{-3}	OR = 1.3	CPD≤10 vs. CPD>10	256
			4,153 (European)	rs16969968	78882925	Missense (*CHRNA5*)	5.0×10^{-3}		FTND	257
			4,762 (European)	rs1051730	78894339	Synonymous (*CHRNA3*)	1.1×10^{-5}	OR = 1.3	Smoking status	179
			8,842 [239]	rs951266	78878541	Intronic (*CHRNA5*)	1.0×10^{-3} (male)	OR = 1.7	Indexed CPD	29
				rs11072768	78929478	Intronic (*CHRNB4*)	1.0×10^{-3} (male)	OR = 1.2	Smoking status	
			32,587 (10,912 AAs+6,889 Asians+14,786 European)	rs16969968	78882925	Missense (*CHRNA5*)	1.1×10^{-17} (meta)	OR = 1.3	CPD≤10 vs. CPD≥20	30
			32,823 (European)	rs1051730	78894339	Synonymous (*CHRNA3*)	$<1.0 \times 10^{-3}$		Pack-years	258
			33,348 (European)	rs16969968	78882925	Missense (*CHRNA5*)	0.01	OR = 1.5 (early-onset) OR = 1.3 (late-onset)	CPD ≤10 vs. >20	259
			38,617 (European)	rs16969968	78882925	Missense (*CHRNA5*)	6.0×10^{-31}	OR = 1.3	CPD ≤10 vs. >20	260
CHRNB3-CHRNA6	8p11.21	GWAS[24, 34, 161]	965 (European)	rs4950	42552633	5′-UTR (*CHRNB3*)	$<1.0 \times 10^{-4}$ (patients with	OR = 1.5	Smoking status	261

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
							Parkinson's disease)			
			1,051 (132 AAs+860 EAs+28 Hispanics+31 others)	rs7004381	42551161	nearGene-5 (*CHRNB3*)	2.4×10^{-3} (pooled)		Quit attempt	159
			1,076 (189 AAs+631 EAs+154 Hispanics+102 others)	rs892413	42614378	Intronic (*CHRNA6*)	$<1.0 \times 10^{-3}$ (interaction with ADHD symptoms)	β = -0.3	CPD	262
			1,929 (European)	rs13277254	42549982	nearGene-5 (*CHRNB3*)	4.0×10^{-5}	OR = 1.4	FTND ≥4 vs. 0 in smokers	249
			2,047 (European)	rs6474412	42550498	nearGene-5 (*CHRNB3*)	1.3×10^{-4}	β = -0.2	WISDM tolerance	227
			2,580 (74% European)	rs4950	42552633	5′-UTR (*CHRNB3*)	$<1.0 \times 10^{-3}$		Initial subjective response to nicotine	160
				rs13280604	42559586	Intronic (*CHRNB3*)	$<1.0 \times 10^{-3}$			
			5,092 (1,661 AAs+3,431 EAs)	rs13273442	42544017	nearGene-5 (*CHRNB3*)	8.6×10^{-5} (meta)	OR = 0.8	FTND ≥4 vs. 0 or 1 in smokers	158
			22,654 (4,297 AAs+9,515 EAs+8,842 Asians)	rs4736835	42547033	nearGene-5 (*CHRNB3*)	5.1×10^{-8} (meta)	β = 0.16	FTND, indexed CPD	33
CHRNA4	20q13.33	Close to the nominated linkage peak on 20q13.12-13.32 (3.6 Mbp)	621 (Asian male)	rs1044397	61981104	Synonymous	$<1.0 \times 10^{-3}$		FTND	219
			1,608 (European)	rs2236196	61977556	3′-UTR	9.3×10^{-4}		FTND ≥4 vs. 0 in smokers	214
			2,037 (1,366 AAs+671 EAs)	rs2236196	61977556	3′-UTR	9.0×10^{-4} (AA female)		FTND	14

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
			3,695 (2,394 EAs+1,301 Hispanics)	rs1044396	61981134	Missense	0.02 (pooled)		DSM-IV ND symptom count	[221]
			5,561 (European)	rs2236196	61977556	3′-UTR	2.3×10^{-3}	$\beta = 0.1$	FTND	[220]
CHRNB1	17p13.1	Close to the nominated linkage peak on 17p13.1-q22 (3.2 Mbp)	1,608 (European)	rs17732878	7362359	nearGene-3	1.7×10^{-3}		FTND ≥4 vs. 0 in smokers	[214]
			2,037 (1,366 AAs+671 EAs)	rs2302763	7359277	Intronic	0.01 (EA)		CPD	[222]
CHRM1	11q12.3	Within the nominated linkage peak on 11p15-q13.4 (0 bp)	2,037 (1,366 AAs+671 EAs)	rs2507821-rs4963323-rs544978-rs542269-rs2075748-rs1938677			8.0×10^{-3} (AA)		CPD	[222]
CHRM2	7q33	Within the nominated linkage peak on 7q31.2-q36.1 (0 bp)	1,608 (European)	rs1378650	136705151	nearGene-3	2.1×10^{-3}		FTND ≥4 vs. 0 in smokers	[214]
Nicotine metabolism genes										
EGLN2-CYP2A6-CYP2B6	19q13.2	GWAS[23, 24, 162, 263]	545 (European)	rs1801272	41354533	Missense (*CYP2A6*)	$<1.0 \times 10^{-4}$		Nicotine metabolite ratio	[226]
				rs28399433	41356379	nearGene-5 (*CYP2A6*)	$<1.0 \times 10^{-4}$			
				*CYP2A6*12*		crossover with *CYP2A7*	$<1.0 \times 10^{-4}$			
				*CYP2A6*1B*		conversion	$<1.0 \times 10^{-4}$			
			709 (European)	genotype-based metabolism (*CYP2A6*)			2.0×10^{-8} (interaction with treatment)	HR = 0.4	Time to relapse	[163]
			1,355 (European)	rs3733829	41310571	Intronic (*EGLN2*)	3.8×10^{-5}	$\beta = 2.0$	Carbon monoxide (CO)	[164]
			1,900 (European and other)	rs1801272	41354533	Missense (*CYP2A6*)	0.02 (interaction		Initial subjective	[208]

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
							with ADHD symptoms)		response to nicotine	
			1,929 (European)	rs4802100	41496025	nearGene-5 (*CYP2B6*)	6.8×10^{-3}		FTND ≥4 vs. 0 in smokers	19
			2,047 (European)	rs3733829	41310571	Intronic (*EGLN2*)	1.5×10^{-3}	$\beta = 0.1$	CPD	227
MAPK signaling pathway & other genes										
BDNF	11p14.1	GWAS,[23] within the nominated linkage peak on 11p15-q13.4 (0 bp)	628 [239]	rs6265	27679916	Missense	<0.05 (male)		Age of initiation	229
			2,037 (1,366 AAs+671 EAs)	rs6484320-rs988748-rs2030324-rs7934165			9.0×10^{-4} (EA)		CPD	111
NTRK2	9q21.33	GWAS,[212] close to the nominated linkage peak on 9q21.33-33 (2.7 Mbp)	2,037 (1,366 AAs+671 EAs)	rs1187272	87404086	Intronic	1.0×10^{-3} (EA)		HSI	11
ARRB1	11q13.4	Within the nominated linkage peak on 11p15-q13.4 (1.5 Mbp)	2,037 (1,366 AAs+671 EAs)	rs528833-rs1320709-rs480174-rs5786130-rs611908-rs472112			8.0×10^{-4} (EA)		FTND	12
MAP3K4	6q26	Within the nominated linkage peak on 6q23.3-q27 (0 bp)	1,608 (European)	rs1488	161538250	3′-UTR	2.7×10^{-4}	OR = 1.4	FTND ≥4 vs. 0 in smokers	214
SHC3	9q22.1	Within the nominated linkage peak on 9q21.33-33 (0 bp)	2,037 (1,366 AAs+671 EAs)	rs1547696	91694120	Intronic	9.0×10^{-3} (pooled)		CPD	230
DNM1	9q34.11	Close to the nominated linkage peak on 9q21.33-33 (3.1 Mbp)	2,037 (1,366 AAs+671 EAs)	rs3003609	130984755	Synonymous	3.1×10^{-3} (EA)		CPD	126

Gene	Chr.	Linkage (distance) /GWAS	Sample size	Variant	Position	Variant Type	*P* value	Effect size	Phenotype	Refs
TAS2R38	7q34	Within the nominated linkage peak on 7q31.2-q36.1 (0 bp)	567 (European)	Haplotype conferring intermediate taste sensitivity (AAV)			1.0×10^{-3}		Smoking status	[264]
			2,037 (1,366 AAs+671 EAs)	Taster (PAV) and non-taster (AVI) haplotypes			3.0×10^{-3} (AA female)		CPD	[18]
APBB1	11p15.4	Within the nominated linkage peak on 11p15-q13.4 (0 bp)	2,037 (1,366 AAs+671 EAs)	rs4758416	6434149	Intronic	3.0×10^{-3} (pooled)		CPD	[231]
PTEN	10q23.1	Within the nominated linkage peak on 10q21.2-q26.2 (0 bp)	688 (European)	rs1234213	89689321	Intronic	2.0×10^{-4}		Smoking status	[232]
NRG3	10q23.1	Within the nominated linkage peak on 10q21.2-q26.2 (0 bp)	614 (European)	rs1896506	83874383	Intronic	4.0×10^{-4}		Smoking cessation	[233]

Smoking status smokers vs. non-smokers; *HSI* heaviness of smoking index (0-6 scale); *FTND* Fagerström Test for Nicotine Dependence (0-10 scale); *CPD* cigarettes smoked per day; *indexed CPD*[29, 33] CPD categorized as non-smoking, <10, 11-20, 21-30, and >31 CPD; *habitual smoking*[133] ever smoking 20 CPD for 6 months or more; heavy vs. light smokers:[255] heavy smokers defined as smoking at least 30 CPD for at least 5 years, and light smokers defined as smoking <5 CPD for at least 1 year; *WISDM* Wisconsin inventory of smoking dependence motives; *NRT* nicotine replacement therapy; *AA* African American; *EA* European American; *VNTR* variable number tandem repeat; *5-HTTLPR* serotonin-transporter-linked polymorphic region; *OR* odds ratio; *HR* hazard ratio; *ADHD* attention-deficit/hyperactivity disorder; *bp* base pair; *Mbp* megabase pair. Genes in this table were significantly associated with ND-related phenotypes in at least two hypothesis-driven candidate gene association studies with a sample size of more than 1,000, or in studies with a sample size of 500 or more but overlapped with linkage or GWAS findings. The "Linkage (distance)/GWAS" column indicates whether a gene is within (< 2 Mbp) or close to (2 - 5 Mbp) any reported linkage region (**Table 4.1**) or found significant in GWASs. All the linkage peaks are based on the review by Li in 2008[9] unless otherwise noted. Distances between candidate genes and closeby linkage regions are in parentheses. For genes with more than one significant variant in a particular study, only the variant(s) with the smallest *P* value(s) is presented, and only the most significant *P* value is shown for each variant if multiple phenotypes were tested in different ethnic/gender groups. Corresponding ethnic/gender group or special analysis methods, such as meta-analysis and interaction, for each *P* value are noted in parentheses right after. Variants composing the most significant haplotype are given if none of the single variants tested was statistically significant. Corresponding effect sizes are provided whenever available. The general term "smoking cessation" was used in the "Phenotype" column for ease of summarization, which represents abstinence at different time points for different studies. Variant positions are based on NCBI Build 37/hg19. For loci with multiple genes, symbols of the gene variants are indicated in parentheses following the variant type.

4.5 Genome-wide association studies

Since the first GWAS published in 2005,[265] this technique using millions of SNPs became the preferred mapping tool for complex diseases/traits.[168] As of October 2015, nine published GWASs and meta-GWASs have yielded 11 genetic loci carrying genome-wide significant variants (GWS; $P < 5 \times 10^{-8}$) associated with relevant ND phenotypes in subjects of European, African, and East Asian ancestries (**Table 4.3** and **Figure 4.1**). However, only three loci were replicated in more than two independent GWASs or meta-GWASs, among which the *CHRNA5*/*A3*/*B4* gene cluster has the most evidence of significance.

Before the GWAS reports, Saccone *et al*.[19] reported significant association of a 3'-UTR variant (rs578776) in *CHRNA3* with dichotomized FTND in smokers in a candidate gene association study with 348 genes. Then, in the GWAS era, five variants in this region reached genome-wide significance in five GWAS and meta-GWAS,[21, 23-25, 27] among which four (rs1051730, rs16969968, rs64952308, and rs55853698) were found to be significant in Europeans, and one (rs2036527) was significantly associated with CPD in AAs. The SNPs rs1051730, rs16969968, and rs55853698 are close-tagging proxies of each other (all pairwise $r^2 > 0.96$), [25] and rs2036527 also is correlated with rs1051730.[27] All the r^2s reported in the main text were extracted from the original studies. Thus, these variants were predicted to either tag or potentially cause the principal risk for high smoking quantity attributable to the 15q25 locus, with approximately one CPD step increase for each risk allele.[23, 25, 27] Although the synonymous SNP rs1051730 (Y188Y) in *CHRNA3* showed the strongest association, the nonsynonymous SNP rs16969968 (D398N) in *CHRNA5* and rs55853698 in the 5'-UTR of *CHRNA5* hold more promise for functional importance. In the European samples, conditional on rs16969968 or rs55853698,

residual association was detected at rs588765, tagging high expression of *CHRNA5* and rs6495308 within *CHRNA3* as showing significant association with CPD unconditionally. Liu *et al*.[25] discovered better model fitting when conditioning on rs55853698 and rs6495308 compared with rs16969968 and rs588765 using the Bayesian information criteria (BIC). Both rs588765 and rs6495308 were reported to be in low LD with each other (r^2 = 0.21) and both to be in only modest LD with the principal SNPs (maximum r^2 = 0.47) in subjects of European ancestry.[25] However, in the AA samples, no second association signal was detected in this region after conditioning on rs2036527, suggesting that rs20356527 and correlated SNPs in populations with African ancestry define a single common haplotype.[27] At the same time, the finding of importance of this gene cluster has been replicated by candidate gene association studies in persons of Asian ancestry[29, 30] and different ND phenotype-cotinine concentration,[156] neural response,[242] smoking cessation,[243, 248, 254] age of initiation,[245] and CPD during pregnancy.[253] The two most replicated variants in candidate gene association studies, rs16969968 and rs1051730, are consistent with the GWAS results. Please refer to **Table 4.2** for details.

The three GWS SNPs on chromosome 8p11 in samples with African and European ancestries—rs13280604, rs6474412, and rs1451240—are in perfect LD with each other[24, 161] and also with a variant (rs13277254) suggestively associated with ND status of smokers in the first ND GWAS.[34] As noted by Rice *et al*.,[161] although the dichotomized FTND appeared to have an equivalent relation with rs1451240 across ethnicities, the relation between this SNP and CPD was much weaker in AAs than in EAs. The other two SNPs were both significantly associated with CPD in Europeans.[24] These associated SNPs are either intergenic or intronic, which may tag

causal variation(s) within the LD block that contains *CHRNB3* and *CHRNA6*, or regulate the expression of the two genes directly. Significant association of variants in *CHRNB3* and *CHRNA6* with ND was also confirmed in eight candidate gene association studies with diverse population ancestries and smoking traits (**Table 4.2**).[33, 158-160, 227, 249, 261, 262] Cui *et al*.[33] obtained a close to GWS *meta-p* value for an upstream variant of *CHRNB3* (rs4736835) in a candidate gene association study of 22,654 subjects with African, European, and East Asian ancestries.

The last region detected by more than one GWAS or meta-GWAS is on chromosome 19q13.2 and includes genes such as *CYP2A6*/*A7*/*B6*, *EGLN2*, *RAB4B*, and *NUMBL*. Thorgeirsson *et al*.[24] identified rs4105144 and rs7937 as significantly associated with CPD in European samples. These two SNPs were reported to be in LD with each other (r^2 = 0.32 and D' = 0.82 in the HapMap CEU samples). Rs4105144 was also in LD with *CYP2A6*2* (rs1801272; r^2 = 0.13 and D' = 1.0 in the HapMap CEU samples), which reduces CYP2A6's enzymatic activity.[24] The SNP identified by the Tobacco and Genetics Consortium[23] (rs3733829) lies between these sites and was reported to show moderate LD with rs4105144 and rs7937. Besides association signals in samples with European ancestry, Kumasaka *et al*.[162] found a copy-number variant (CNV; rs8102683) with a strong effect on CPD (β = -4.00) in a Japanese population, and another significantly associated SNP (rs11878604; β = -2.69) located 30 kb downstream of the *CYP2A6* gene after adjustment of the CNV. Rs8102683 shared a common deletion region with other CNVs ranging from the 3' end of the *CYP2A6* gene to the 3' end of the *CYP2A7* gene; however, this common deletion was not significant in the European population.[162] Very recently, Loukola *et al*.[263] conducted the first GWAS on nicotine metabolite ratio (NMR) and detected 719 GWS

SNPs within this region. Strikingly, the significant *CYP2A6* variants explain a large fraction of variance (up to 31%) in NMR in their study samples.

All the other signals reported by only one GWAS or meta-GWAS can be found in **Table 4.3** and **Figure 4.1**, among which a missense variant rs6265 in *BDNF* was significantly associated with smoking initiation and an intergenic variant rs3025343 close to *DBH* was implicated in smoking cessation.[23] It is worth noting that GWASs without GWS variant identification still render valuable information in determining susceptibility loci for ND. The first ND GWAS, performed by Bierut *et al*.,[34] nominated *NRXN1* in the development of ND, which was validated by a subsequent candidate gene association study.[112] By using a network-based genome-wide association approach, Vink *et al*.[212] discovered susceptibility genes encoding groups of proteins, such as glutamate receptors, proteins involved in tyrosine kinase receptor signaling, transporters, and cell-adhesion molecules, many of which were confirmed in later candidate gene association studies.[11, 213] Please refer to **Supplementary Table 1** for a list of GWASs without GWS results.

Table 4.3: Significant genome-wide association study (GWAS) findings for ND-related phenotypes.

Population	Phenotype	Nearest gene	Chr.	SNP [Effect Allele]	Physical Position	Variant Type	Sample size	*P* value	Effect size	Refs
European	CPD	*CHRNA5/A3/B4*	15q25.1	rs1051730[A]	78894339	Synonymous	73,853	2.8×10^{-73}	$\beta = 1.02$	21, 23-25
				rs16969968[G]	78882925	Missense	73,853	5.6×10^{-72}	$\beta = 1.00$	23, 25
				rs6495308[T]	78907656	Intronic	136,090	5.8×10^{-44}	$\beta = 0.73$	25
				rs55853698	78857939	5′-UTR	136,090	1.3×10^{-16}		25
		CYP2A6, *EGLN2*, *RAB4B*	19q13.2	rs4105144[C]	41358624	Intergenic	83,317	2.2×10^{-12}	$\beta = 0.39$	24
				rs7937[T]	41302706	3′-UTR	86,319	2.4×10^{-9}	$\beta = 0.24$	24
				rs3733829[G]	41310571	Intronic	73,853	1.0×10^{-8}	$\beta = 0.33$	23
		LOC100188947	10q23.32	rs1329650[G]	93348120	Intronic	73,853	5.7×10^{-10}	$\beta = 0.37$	23
				rs1028936[A]	93349797	Intronic	73,853	1.3×10^{-9}	$\beta = 0.45$	23
		PDE1C	7p14.3	rs215605[G]	32336965	Intronic	77,012	5.4×10^{-9}	$\beta = 0.26$	24
		CHRNB3/A6	8p11.21	rs13280604[A]	42559586	Intronic	76,670	1.3×10^{-8}	$\beta = 0.31$	24
				rs6474412[T]	42550498	Intergenic	84,956	1.4×10^{-8}	$\beta = 0.29$	24
	FTND	*CACNA2D1*	7q21.11	rs13225753	82158523	Intergenic	4,117	3.5×10^{-8}	NA	266
	Smoking initiation	*BDNF*	11p14.1	rs6265[C]	27679916	Missense	143,023	1.8×10^{-8}	OR = 1.06	23
	Smoking cessation	*DBH*	9q34.2	rs3025343[G]	136478355	Intergenic	64,924	3.6×10^{-8}	OR = 1.12	23
	NMR	*CYP2A6*, *CYP2B6*, *CYP2A7*, *EGLN2*, *NUMBL*	19q13.2	rs56113850[C]	41353107	Intronic	1,518	5.8×10^{-86}	$\beta = -0.65$	263
African American	CPD	*CHRNA5/A3/B4*	15q25.1	rs2036527[A]	78851615	Intergenic	15,554	1.8×10^{-8}	$\beta < 1.00$	27
	FTND	*C14orf28*	14q21.2	rs117018253	45337321	Intergenic	3,529	4.7×10^{-10}	NA	266
		CSGALNACT1, *INTS10*	8p21.3	rs6996964	19623911	Intergenic	3,529	1.1×10^{-9}	NA	266
		DLC1	8p22	rs289519	13237048	Intronic	3,529	4.5×10^{-8}	NA	266
European & African American	Dichotomized FTND	*CHRNB3*	8p11.21	rs1451240[A]	42546711	Intergenic	4,200	6.7×10^{-16}	OR = 0.65	161
Japanese	CPD	*CYP2A6*, *CYP2A7*	19q13.2	rs8102683[0 copy]	41363765	CNV	17,158	3.8×10^{-42}	$\beta = -4.00$	162
				rs11878604[C]	41333284	Intergenic	17,158	9.7×10^{-30}	$\beta = -2.69$	162

CNV copy number variation; *CPD* cigarettes smoked per day; *dichotomized Fagerström Test for Nicotine Dependence (FTND)* scores ≥ 4 vs. < 4; *NA* not available; *NMR* nicotine metabolite ratio; *OR* odds ratio; s*moking cessation* whether regular smokers had quit at the time of interview; *smoking initiation* ever versus never began smoking. This table focuses on results achieving genome-wide significance (GWS). We used the significance threshold of 5×10^{-8}. The most significant GWAS finding from different studies for any specific variant is given. If numerous tightly mapped markers showed GWS in one study, only the most significant one is provided. Variant positions are based on NCBI Build 37/hg19. For many studies, it was not possible to extract the exact sample size used for each locus, so the sample sizes above are approximate. Effect sizes refer to beta coefficients for CPD and NMR, and odds ratios for smoking initiation and cessation.

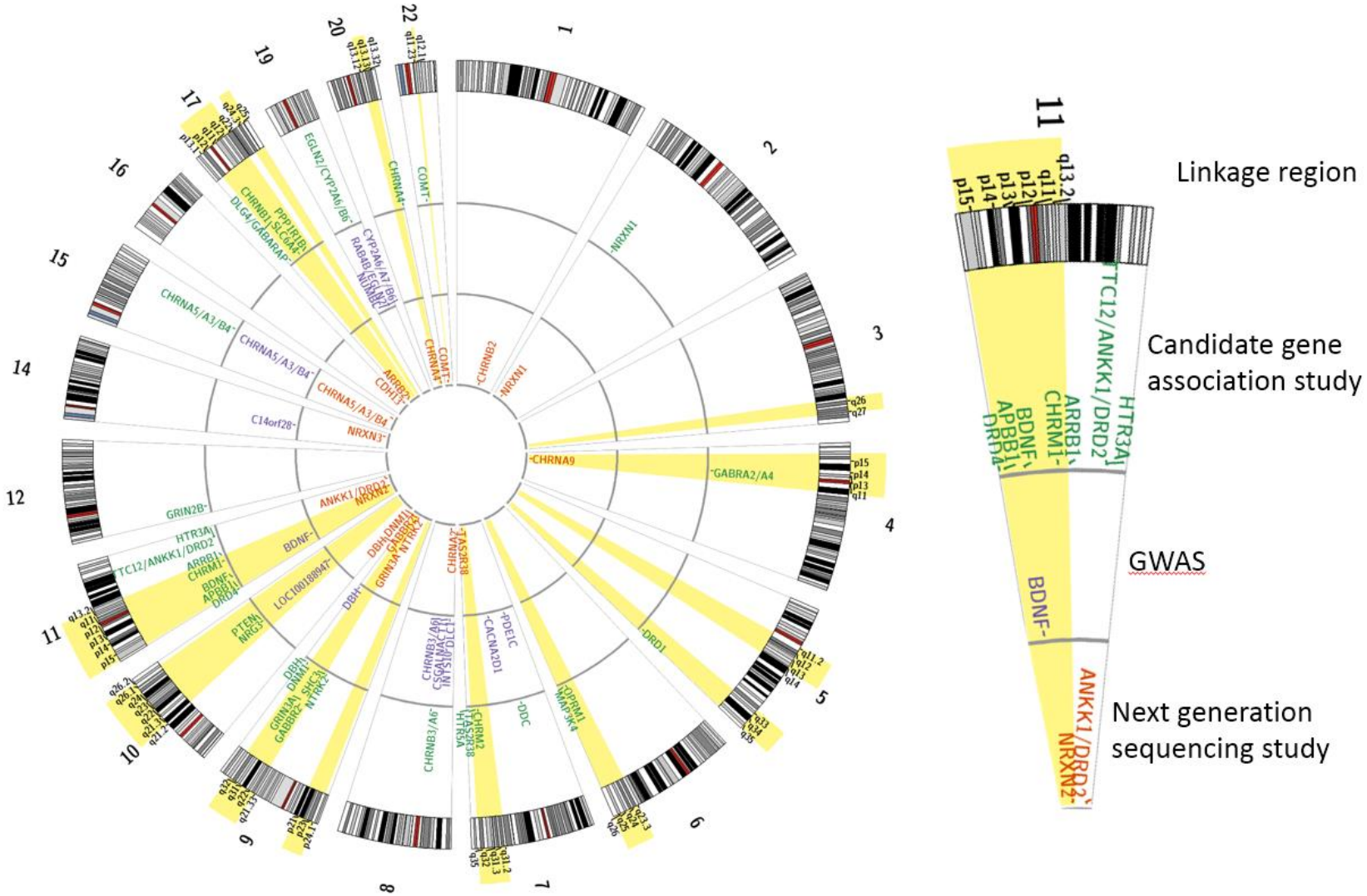
4.6 Targeted sequencing studies

As the "missing heritability" issue emerged in each field, researchers suspected that much of the missing heritability is attributable to genetic variants that are too rare to be detected by GWAS but may have relatively large effects on risk and thus are important to study using next-generation sequencing technologies.[267] Both population genetic theories and empirical studies of several complex traits suggest that rare alleles are enriched for functional and deleterious effects and thus are disproportionately represented among disease alleles.[39]

For the field of ND genetics, rare variant investigation started with the nAChR subunit genes, which not only are biologically important but also have yielded the most replicable results in both GWASs and candidate gene association studies, as presented above. Wessel *et al*.[40] first examined the contribution of common and rare variants in 11 nAChR genes to FTND in 448 EA smokers, which revealed significant effects of common and rare variants combined in *CHRNA5* and *CHRNB2*, as well as of rare variants only in *CHRNA4*. Xie *et al*.[41] followed up on the *CHRNA4* finding by sequencing exon 5, where most of the nonsynonymous rare variants were detected, in 1,000 ND cases and 1,000 non-ND controls with equal numbers of EAs and AAs. They discovered that functional rare variants within *CHRNA4* may reduce ND risk. Also, Haller *et al*.[42] detected protective effects of missense rare variants at conserved residues in *CHRNB4*. They examined *in vitro* the functional effects of the three major association signal contributors (i.e., T375I and T91I in *CHRNB4* and R37H in *CHRNA3*), finding that the minor alleles of the studied SNPs increased cellular response to nicotine. The two rare variants in *CHRNB4* were confirmed to augment nicotine-mediated α3β4 nAChR currents in hippocampal neurons, as did a third variant, D447X, in the report of Slimak *et al*.[268] The fourth SNP they analyzed, R348C,

reduced nicotine currents. They also observed that habenular expression of the β4 gain-of-function allele T374I resulted in strong aversion to nicotine in mice, whereas transduction of the β4 loss-of-function allele R348C failed to induce nicotine aversion. Later, Doyle *et al*.[43] reported an interesting rare variant in *CHRNA5* that could result in nonsense-mediated decay of aberrant transcripts in 250 AA heavy smokers. And recently, Yang *et al*.[44] performed a targeted sequencing study with the goal of determining both the individual and the cumulative effects of rare and common variants in 30 candidate genes implicated in ND. Rare variants in *NRXN1*, *CHRNA9*, *CHRNA2*, *NTRK2*, *GABBR2*, *GRIN3A*, *DNM1*, *NRXN2*, *NRXN3*, and *ARRB2* were found to be significantly associated with smoking status in 3,088 AA samples, and a significant excess of rare variants exclusive to EA smokers was observed in *NRXN1*, *CHRNA9*, *TAS2R38*, *GRIN3A*, *DBH*, *ANKK1*/*DRD2*, *NRXN3*, and *CDH13*. All the 18 genetic loci implicated in targeted sequencing studies are marked in **Figure 4.1**.

Figure 4.1: The ND genetic susceptibility map with nominated linkage peaks and candidate genes, as suggested by genome-wide linkage, hypothesis-driven candidate gene association (CAS), genome-wide association (GWAS), and targeted sequencing (next-generation sequencing; NGS) studies. Linkage peaks are highlighted in light gray; CAS, GWAS, and NGS results are presented as gene names at the outer, middle, and inner rings, respectively.



4.7 Implications

According to our list, 242 candidate gene association, 22 genome-wide linkages, 18 GWAS, and 5 targeted sequencing, making a total of 287 studies, have been conducted in the ND genetics field so far. The numbers for genome-wide linkage and candidate gene association studies before 2004 are based on Li[9] and Munafò *et al*.,[269] respectively. As a summary and refining of the 287 ND genetic studies, we developed an ND genetic susceptibility map with 14 linkage regions and 47 unique loci of 60 susceptibility genes altogether (**Figure 4.1**).

Genome-wide linkage and GWAS are unbiased exploratory approaches. By comparing their results, we found that only two GWS signals are within the nominated linkage peaks, which are *LOC100188947*, and *BDNF*.[9, 270] The other nine loci, including the three most replicable ones, are outside of the linkage peaks, and the rest of the twelve linkage regions do not contain GWS signals (**Tables 4.1** and **4.2**). This discrepancy reflects the different natures of the two genome-wide study approaches. Genome-wide linkage studies usually investigate sparse microsatellites (between 200-500) segregated with the trait of interest in different families, while GWAS takes advantage of dense common variants (10^4-10^6) and thousands of unrelated individuals. Because distinct characteristics of family and case control samples and known locus heterogeneity for ND, we may not expect same sets of susceptibility alleles prominent in both samples. The large nominated linkage regions tagged by microsatellites may implicate common variants, which may be detected in GWAS, such as the two overlapping loci, aggregate of rare variants, or any structural variants within the region. However, even if a linkage region is driven by common variants, we may not locate them in GWAS due to the stringent *p* values applied. And the latter two cases are clearly beyond the capabilities of GWAS.

The presence of GWS signals outside of linkage peaks might also result from the lack of power for linkage studies to detect weak genetic effects exhibited by the loci involved in complex diseases compared with association studies.[271] As we can see, unbiased approaches are powerful by marking areas for us in the genome; nevertheless, the areas they indicate are often large and may not be complete. In this case, hypothesis-driven studies are useful and necessary tools to not only scrutinize marked areas, but also explore promising false negative results and biologically plausible targets.

Both candidate gene association and targeted sequencing studies serve this purpose. Candidate gene association studies replicated and extended 5 of the 11 GWAS results, i.e., *CHRNB3*/*A6*, *DBH*, *BDNF*, *CHRNA5*/*A3*/*B4*, and *EGLN2*/*CYP2A6*/*B6*. For the other 29 non-GWS candidate genetic loci, 20 and 7 were selected from within and close to linkage peaks, respectively, the exceptions being *NRXN1* and *DDC* (**Table 4.2**), which reminds us of the importance of examining suggestive results in GWAS,[34] the other two examples being *GRIN2B* and *NTRK2*,[212] and biologically plausible genes, separately. Although we have localized candidate genes within most of the nominated linkage regions, three linkage peaks, on chromosomes 3q26-q27, 5q11.2-q14, 9p21-p24.1, and 17q24.3-q25.3, are still empty, suggesting there are novel susceptibility genes to be discovered in the future. Overlaps and distinctions from the two unbiased approaches and the significant number of loci reproduced or proposed in candidate gene studies suggested that, we have many more study targets with good statistical evidence besides the three most replicable GWAS loci. The fourth "immature" approach is also hypotheses-driven and has verified the importance of rare variants in ND genetics.[40-42, 44] Besides the demonstrated aggregate effects of rare variants in 12 genetic loci

implicated in previous studies, biological candidates showing equivocal or no association beforehand were found to be significantly associated with ND-related phenotypes, such as *CHRNB2*, *CHRNA9*, *CHRNA2*, *NRXN2*, *NRXN3*, and *CDH13*, among which *CHRNA9* and *NRXN2* are within linkage regions.[9, 44] Thus, we believe whole-exome and whole-genome sequencing studies focusing on rare variants, as the third unbiased experimental approach, will reveal new susceptibility genes/variants and further dissect the existing targets.

It is worth noting that to establish a positive replication of a genotype-phenotype association, in replication studies, every effort should be made to analyze phenotypes comparable to those reported in the initial study.[165] However, the ND genetics studies mentioned above have utilized a plethora of smoking related phenotypes. In general, they can be classified into four groups: 1) categorical variables along smoking trajectories, e.g., smoking initiation, status, and cessation; 2) ND assessed using DSM-IV or FTND; 3) smoking quantity like CPD; and 4) endophenotypes such as NMR, cotinine and CO levels, or functional imaging results. At least two out of the four groups of phenotypes have been used in genome-wide linkage studies (**Table 4.1**),[9] candidate gene association studies (**Table 4.2**), and GWASs (**Table 4.3**). Due to primarily sample source and size requirement differences, DSM or FTND ascertained ND definitions were commonly used in linkage studies, while CPD is more often applied in GWASs. For candidate gene association studies, more comprehensive smoking profiles were usually available and thus tested for association with positive results from unbiased studies as replication, or more importantly, extension by using different phenotypes (**Table 4.2**), because there is considerable evidence that the various smoking measures are not highly related to one another.[272] Even for measures with relatively high correlation, such as

FTND and CPD, the slight change of phenotype from FTND-based ND to CPD changes study results.[161] Therefore, although several loci showed associations with different phenotypes, such as *TTC12-ANKK1-DRD2*, *CHRNA5/A3/B4*, and *CYP2A6/B6* (**Tables 4.2** and **4.3**), we should not expect positive associations with one phenotype to be reproduced in samples with other phenotypes, and it is important to keep in mind that a small change in phenotype may expose previously undiscovered variants, which underlie different biological processes and may have specific roles in distinguishing phenotypes.[161]

Additionally, gene–gene and gene–environment interactions are two pieces of information missing from the current map because of the small number of reported studies. We expect more results in these two areas will come out with the development of efficient algorithms and become important parts of the susceptibility map. It is also worth noting that half of the 48 ND loci were significantly associated with alcohol-related phenotypes, and ~30% were involved in illicit drug dependence (**Supplementary Table 2**), suggesting that the 60 genes on the ND map are good candidates for addiction studies of other drugs as well.

4.8 Future directions

Technological advances enable the development of different experimental approaches. A genetic susceptibility map, as put together in this review, contains scientific evidence from diverse approaches and can serve as a draft of the "parts list" to be updated periodically until complete.[173] We hope such an enumeration will catalyze an array of specific, targeted, and nuanced scientific studies, as suggested by Sullivan *et al*.;[173] e.g., calculating the heritability explained by the 47 genetic loci, replicating association signals currently lacking in evidence,

identifying causal variant(s) within each locus through expression data integration and functional characterization, elucidating biological mechanisms between the genotype and ND, exploring gene–gene and gene–environment interactions, understanding the part played by epigenetic modifications, developing and evaluating treatment prediction models, and so forth.

Although the sample size of candidate gene association studies has increased over the years (**Supplementary Figure 1a**), genetic power calculation and corresponding sample size ascertainment should always be a top priority before conducting genetic studies. Additionally, only 18% and 10% of the 287 studies investigated subjects with African and Asian ancestries, respectively, compared with 69% for European ancestry (**Supplementary Figure 1b**). Studying different populations is necessary to understand the genetic causes of ND in various ethnic groups. Concurrently, given the importance of rare variants suggested by targeted sequencing study results, thorough and well-powered genomic evaluations at the lower end of the allelic spectrum are needed. Whole-exome and whole-genome sequencing studies with enough statistical rigor would enable a substantial update of the ND genetic susceptibility map in the near future.

However, it is important to acknowledge that the genetic liability accounted for by each of the 47 loci is low, considering their respective effect sizes, which may also explain why they can be identified through one type of unbiased study, but not the other. Anticipating future studies on predictive power of these loci cumulatively, we are inclined to project that the heritability explained will still be limited, which renders the susceptibility map only as a beginning. Functional studies have been carried out for a few variations with certain or uncertain smoking associations as well (**Table 4.4**). Nevertheless the *TTC12-ANKK1-DRD2*

cluster shows consistent association with smoking-related behaviors (**Table 4.2**), function of the most prominent variation in this region-*Taq1A* is still largely unknown; on the other hand, we have known the molecular and neurobehavioral functional consequences of *BDNF* met66val polymorphism (rs6265) for more than a decade,[273] but its association with ND phenotypes is not very strong (**Table 4.2**). Combining the susceptibility map results with relevant functional annotations will certainly facilitate our determination of variations bearing higher translational values.[274] All in all, this map empowers us to sift through existing accomplishments and ponder future research strategies, an approach that may serve as a useful tool for other complex diseases/traits.

Table 4.4: Functional studies of variations associated with smoking in the 47 ND susceptibility loci.

Chr.	Gene	Experiment	Variation [Effect Allele]	Effect	Ref.
1	*CHRNB2*	*in vitro* gene expression assay	rs2072658 [A]	reduced expression	275
6	*OPRM1*	PET brain imaging	rs1799971 [G]	binding potential & receptor availability change	195, 276, 277
8	*CHRNA2*	electrophysiology assay	rs141072985 rs56344740 rs2472553	nAChR function change	278, 279
	CHRNB3	*in vitro* gene expression assay	rs6474413 [C]	reduced expression	280
		ChIP & *in vitro* gene expression assay	rs4950 [G]	eliminated TF binding & reduced promoter activity	261
9	*DNM1*	*in vitro* gene expression assay	rs3003609 [T]	reduced expression	126
11	*BDNF*	fMRI, ^{1}H-MRSI & immuno-enzyme assays	rs6265	different brain activation, BDNF secretion & subcellular distribution	273
	DRD4	fMRI	exon 3 VNTR	different brain activation	281
15	*CHRNA5/A3/B4*	imaging series of *in vitro* assays	rs16969968 [A]	brain circuit strength prediction altered response to nicotine agonist	133, 282, 283

Chr.	Gene	Experiment	Variation [Effect Allele]	Effect	Ref.
		electrophysiology & FLEXstation		lower Ca permeability & increased short-term desensitization	
17	*SLC6A4*	*in vitro* gene expression assay in situ hybridization SPECT imaging	5-HTTLPR	transcriptional efficiency & expression change	284-286
19	*CYP2A6/B6*	Please refer to Tricker[287] for a comprehensive summary.			
20	*CHRNA4*	electrophysiology assay	exon 5 haplotype	different receptor sensitivity	288
22	*COMT*	enzyme activity assay	rs4680 [A]	less enzyme activity	289

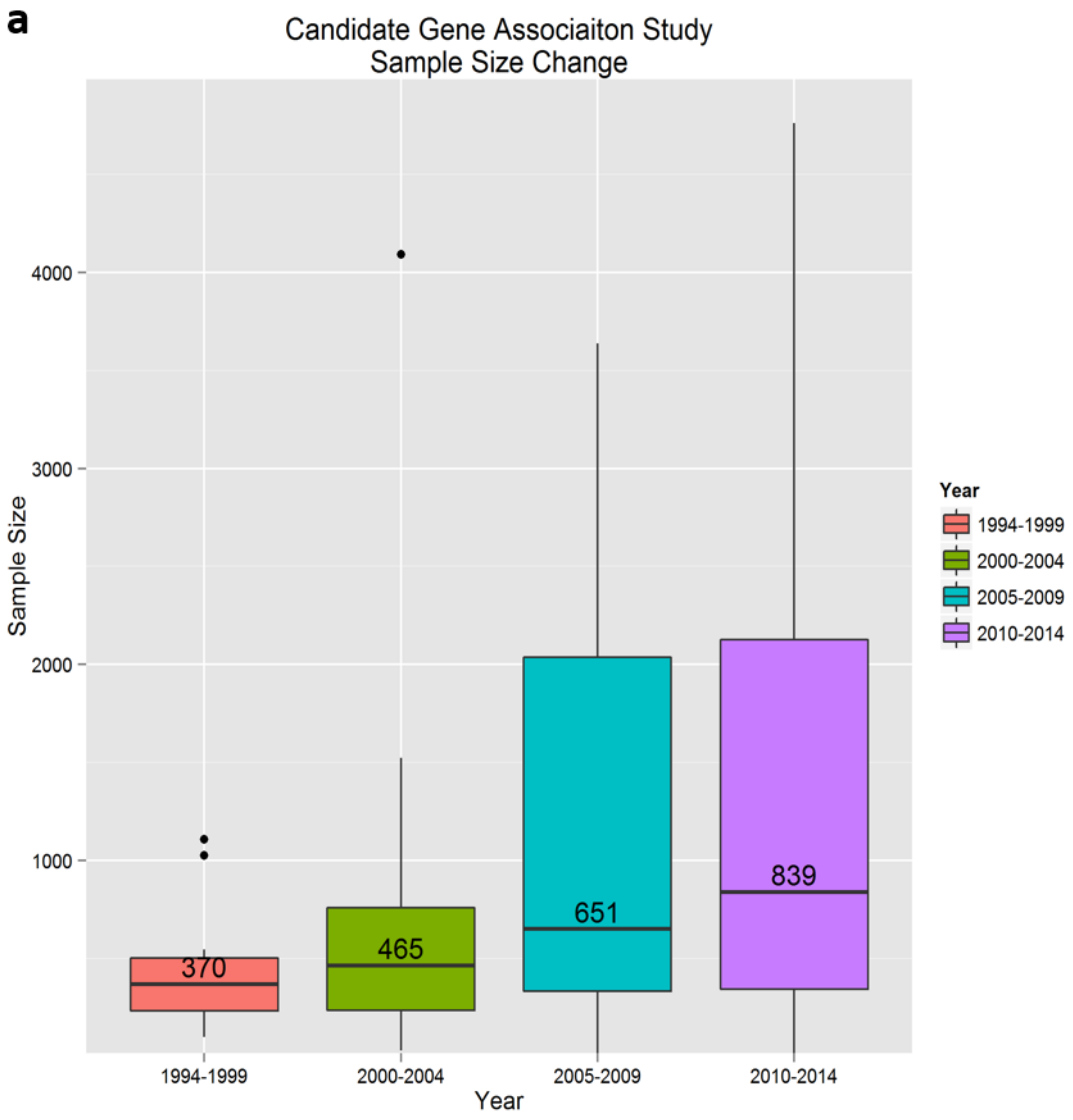
PET positron emission tomography; *ChIP* chromatin immunoprecipitation; *fMRI* functional magnetic resonance imaging; *^{1}H-MRSI* ^{1}H magnetic resonance spectroscopic imaging; *SPECT* single-photon emission computed tomography; *nAChR* nicotinic acetylcholine receptor.

4.9 Chapter acknowledgments

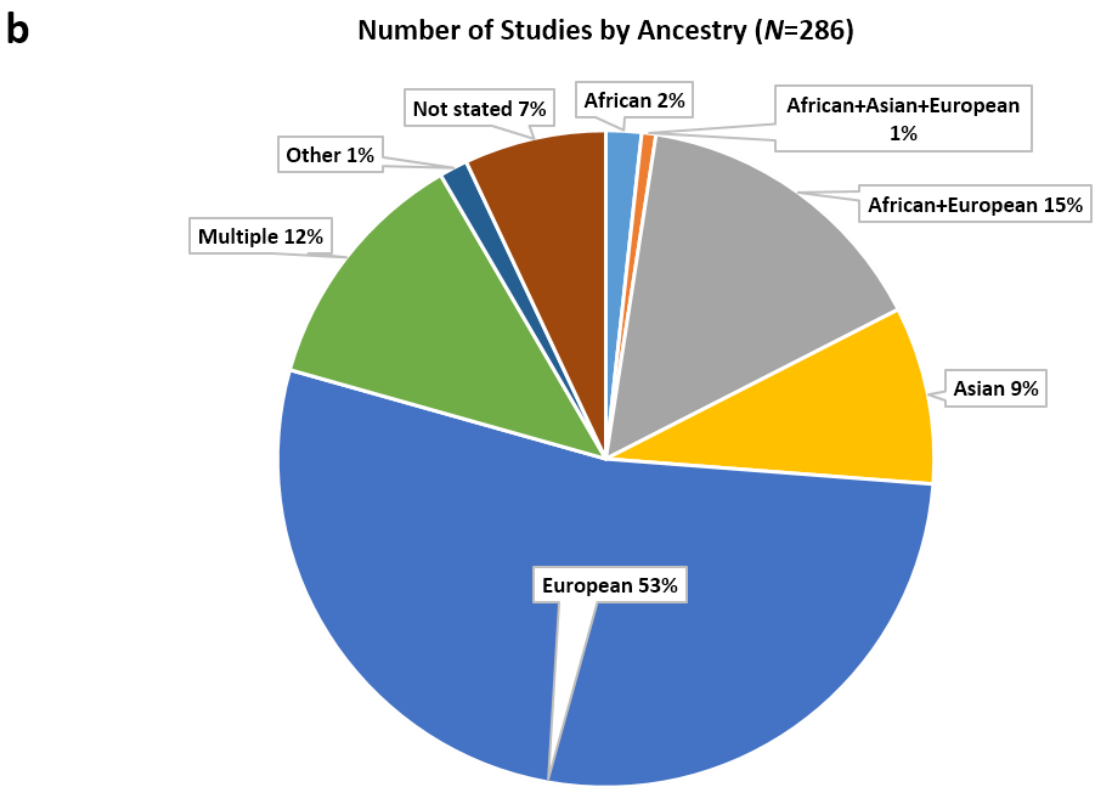
This chapter was adapted from Yang and Li.[45]

4.10 Supplementary data

Supplementary Figure 1: Extent of previous studies.



(a) **Sample size change for the 242 candidate gene association studies from 1994 to 2014.** Boxplot and sample size medians are given for every five years. The Y axis is scaled for the best representation. A portion of the outliers are hidden.



(**b**) **Percentages of the 286 ND genetic studies investigating subjects of particular ancestries.** The category of "multiple" represents samples involving more than one ethnic group but without separate analysis, the majority of which usually had a single ancestry. There are 1% (African + Asian + European) and 15% (African + European) of the studies focusing on subjects from three and two ancestries, respectively.

Supplementary Table 1. A list of: **1)** ND GWASs without GWS results; **2)** candidate gene association studies with sample sizes ≥ 1000, but without replication studies of comparable sizes; **3)** candidate gene association studies with sample sizes between 500 and 1000, but the susceptibility genes implicated not within the nominated linkage peaks or GWAS regions; and **4)** candidate gene association studies with sample sizes ≥ 500, but showed negative results for all or part of the genes tested.

Category	Reference	Population	Sample Size	Phenotype	Gene(s) Implicated/Studied
1) ND GWASs without GWS results	Bierut et al.[34]	European	1,929	FTND≥4 vs. FTND=0 in smokers	*NRXN1*, *CHRNB3*
	Berrettini et al.[20]	European	15,000	CPD	*CHRNA5/A3*
	Uhl et al.[290]	European	550	Smoking cessation	
	Vink et al.[212]	European	11,360	Smoking status	*GRIN2B*, *GRIN2A*, *GRIK2*, *GRM8*, *NTRK2*, *GRB14*, *SLC1A2*, *SLC9A9*, *CDH23*
	Liu et al.[239]	European+African	9,714	Smoking status, CPD, HSI, FTND	*IL15*
	Drgon et al.[291]	European	480	Smoking status	
	Drgon et al.[292]	European	262	Smoking status	
	Uhl et al.[293]	European	369	Smoking status	
	Hamidovic et al.[294]	European+African	680	Serum cotinine level	*IDE*, *LOC101928077*
2) CAS ($N \geq$ 1000), but without good replication	Stapleton *et al*.[191]	80% European	2,155	Smoking cessation	*SLC6A3*
	Ehringer *et al*.[223]	European+African+ Hispanic	1,068	Subjective response to nicotine	*CHRNB2*
	Huang *et al*.[17]	European+African	2,037	CPD, HSI, FTND	*DRD3*
	Sun *et al*.[12]	European+African	2,037	CPD, HSI, FTND	*ARRB2*
	Wei *et al*.[13]	European+African	2,037	CPD, HSI, FTND	*CHAT*
	Jackson *et al*.[295]	European	3,969	FTND	*CAMK4*
	Docampo *et al*.[217]	European	1,596	FTND, smoking status	*NRXN3*
	Mutschler *et al*.[296]	European	1,094	Smoking status	*NPY*
	Wang *et al*.[224]	European+African	7,186	CPD, FTND	*CHRNA2*
3) CAS (500 ≤ $N \leq$ 1000), but genes not overlapped with linkage or GWAS results	Chen *et al*.[297]	European	688	Progression to ND	*EPAC*
	Ling *et al*.[192]	Asian	668	Age of initiation	*SLC6A3*
	do Prado-Lima *et al*.[211]	Brazilian (multi-ethnic)	625	Smoking status	*HTR2A*
	Chen et al.[298]	European	688	Smoking status	*RHOA*, *YWHAG*
	Ton *et al*.[299]	93% European	593	Smoking cessation	*CCK*
	O'Gara *et al*.[300]	European+African	583	Smoking cessation	*SLC6A3*
	Rovaris *et al*.[301]	European	627	Smoking status, CPD, FTND	*NR3C2*, *NR3C1*
	Chen *et al*.[302]	Asian	558	FTND	*CHRNB2*

Category	Reference	Population	Sample Size	Phenotype	Gene(s) Implicated/Studied
4) CAS ($N \geq 500$), but with negative results	Smits *et al*.[303]	European+African+ Asian	20,938	Smoking status, CPD	*CYP1A1*, *GSTM1*, *GSTT1*, *NAT2*, *GSTP1*
	Carter *et al*.[228]	not stated	4,091	Smoking status, CPD	*CYP2A6*
	Saadat *et al*.[304]	European	683	Smoking status	*GSTM1*, *GSTT1*
	Huang *et al*.[196]	European	1,518	Smoking status, cotinine level	*MAOA*, *DBH*, *DRD4*, *HTR2A*
	Rodriguez *et al*.[305]	European	3,637	Age of initiation	*TPH*
	Trummer *et al*.[209]	European	2,844	Smoking status, FTND, CPD etc.	*SLC6A4*
	Ton *et al*.[197]	93% European	593	Smoking cessation	*TH*, *DRD2*, *DRD3*, *DRD4*, *SLC6A3*, *COMT*
	David *et al*.[210]	European	1,398	Smoking relapse	*TPH1*, *SLC6A4*, *HTR1A*
	Siiskonen *et al*.[306]	European	6,358	Smoking status and cessation	*NR3C1*
	Munafo *et al*.[307]	European	2,437	Smoking status	*DRD2*
	Breitling *et al*.[198]	European	1,443	Smoking cessation	*COMT*
	Munafo *et al*.[308]	European	887	Smoking cessation	*DRD2*
	Breitling *et al*.[309]	European	1,446	Smoking status	*DDC*, *DRD2*, *SLC6A3*
	Hubacek *et al*.[310]	European	2,559	Smoking status and cessation, CPD	*FTO*
	Spruell *et al*.[311]	European	925	Smoking cessation, cotinine level	*CHRNA4*, *CHRNB2*
	Marteau *et al*.[199]	European	633	NRT consumption	*OPRM1*
	Munafo *et al*.[200]	European	598	Smoking cessation	*OPRM1*
	Chenoweth *et al*.[312]	African	667	Nicotine metabolic ratio, CPD, nicotine equivalents in urine	*POR*, *FMO3*

Supplementary Table 2. A list of the genes on the ND genetic susceptibility map, which are also implicated in other drug addictions.

Category	Gene	Phenotype*	Reference(s)
Alcohol	*BDNF*	AD	313
	CDH13	Comorbid AD/ND	314
	CHRM2	AD	315
	CHRNA4	Subjective response to alcohol	275
		Binge drinking	316
	CHRNA5/A3/B4	AD	317-319
	CHRNB3/CHRNA6	Alcohol consumption	320, 321
		AD	319
	CHRNB2	Subjective response to alcohol	223, 275
	COMT	Alcohol consumption	322
	DBH	AD	323
	DDC	Maximum number of drinks	324
	DRD1	AD	325
	DRD4	AD	325
	GABRA2	Alcoholism	326
		AD	313, 327-330
	GRIN2B	Alcoholism	331
	HTR3A	AD	37
	NRXN1	AD	332
	NRXN3	AD	333
	NTRK2	AD	313
	OPRM1	Alcohol sensitivity	334
	PPP1R1B	AD	313
	SHC3	Alcohol sensitivity	335
	SLC6A4	AD	37, 98, 336
	TAS2R38	Alcohol consumption	337
	TTC12/ANKK1/DRD2	Alcoholism	338
		AD	339, 340
		Comorbid alcohol/drug dependence	341
Illicit drugs	*BDNF*	Subjective and physical response to amphetamine	342
	CDH13	Methamphetamine dependence	343
	CHRNA2	Antisocial drug dependence	344
	CHRNA5/A3/B4	CD	318, 319, 345
	CHRNB3/A6	CD	319
	COMT	Cocaine-induced paranoia	346
	DBH	Cocaine-induced paranoia	347
	DRD2	Polysubstance abuse	348
		Heroin abuse (nasal inhalation)	349
		Amphetamine dependence	350
	DRD4	HD	351, 352
	GABRA2	Marijuana and illicit drug dependence	353

Category	Gene	Phenotype*	Reference(s)
		Cocaine cue-reactivity	[354]
	GRIN2B	CD	[331]
	HTR3A	CD	[37]
	OPRM1	HD	[355]
		Substance dependence (heroin/cocaine)	[356]
		Antisocial drug dependence	[344]
		Cocaine cue-reactivity	[354]
	SLC6A4	HD	[357]
		CD	[37]

* *AD* alcohol dependence; *CD* cocaine dependence; *HD* heroin dependence; *ND* nicotine dependence.

Chapter 5

Expression and methylation quantitative trait loci

5.1 Abstract

Although quite a few susceptibility genes have been identified to influence smoking, the mechanistic steps between genetic variation and smoking-related traits are generally not understood. In this study, we mapped *cis*-expression and methylation quantitative trait loci (eQTL and mQTL) for 57 smoking candidate genes using the BrainCloud cohort (*N* = 94 African Americans [AAs] and 84 European Americans [EAs] for expression; *N* = 31 AAs and 29 EAs for methylation). A eQTL, with a range of 75 Kb between the two eQTL variants furthest away, was identified to significantly affect *EGLN2* expression in the EA sample, where multiple associated low-frequency variants were previously found affecting smoking quantity in aggregate. Two mQTLs with ranges of 121 and 35 Kb were detected for CpG sites in *NRXN1* and *CYP2A7*, respectively. Particularly we showed for the first time that the minor allele of one variant

(rs3745277), which is located in *CYP2A7P1* (downstream of *CYP2B6*) and receives strong biological evidence from the Roadmap Epigenomics and ENCODE projects, significantly decreased methylation levels at the CpG site for *CYP2A7* (cg25427638; $P = 5.31\times10^{-7}$), reduced RNA expression levels of *CYP2B6* ($P = 0.03$), and lowered percentage of smokers (8.8% vs. 42.3%, *OR* [95% *CI*] = 0.14 [0.02 – 0.62], $P = 4.47 \times 10^{-3}$) in a dominant way for the same cohort. Conditional analysis indicated negligible contribution of smoking on either cg25427638 methylation or *CYP2B6* expression beyond the genetic variation effect discovered. Our findings link genetic variation, DNA methylation, mRNA expression, and smoking status together using the same participants, which depicts a regulatory mechanism from mQTL to phenotypic manifestation for the first time. Additionally, this study demonstrates different regulatory effects of low-frequency and common variants on mRNA expression and DNA methylation, respectively. Experiments to test and verify causal effects among these layers of regulation are thus warranted.

5.2 Introduction

Tobacco smoking is the leading preventable cause of death in the US,[2] and one of the three leading components of the global disease burden.[1] It causes more than 480,000 deaths each year, about one of every five deaths in the US,[2] and kills more than 6 million people annually worldwide.[1] Even though we have known that smoking can cause cancer almost anywhere in the body, harms every organ, and affects a person's overall health,[2] in 2014, an estimated 40 million adults in the US still currently smoke cigarettes.[3] Nicotine dependence (ND), with an

average heritability of 0.56,[4] is the primary factor maintaining smoking behavior[5] and predicting failures of smoking cessation.[358]

With the efforts to identify genetic factors underlying smoking for decades, we have obtained 14 linkage regions and 47 unique loci of 60 susceptibility genes using stringent sample size, replication, and *P* value criteria.[45] Most identified variants within these genetic loci are located in noncoding regions based on association study results,[45] which is in line with observations from other fields: variation at noncoding regulatory sequences contributes to the genetics of complex traits.[359] However, until recently, we have known little about the mechanisms by which most regulatory variants act.[46] With the rise of massively parallel sequencing technologies, recent studies have characterized multiple layers of gene regulation, including chromatin states, transcription factor (TF) binding footprints, profiles or different epigenetic marks, and posttranscriptional modifications, which enable us to probe regulatory variants' control of transcriptional processes through multiple aspects of gene regulation.[46, 47] *Cis*-expression (eQTL) and methylation quantitative trait loci (mQTL) mapping with these molecular phenotypes have been successfully used to nominate SNPs, which would then be tested for their associations with different types of drug addictions including alcohol,[360] heroin,[361] and smoking.[362]

Hancock *et al*.[362] successfully found several SNPs associated with both *CHRNA5* methylation and expression in addiction-relevant brain regions and with risk of ND across diverse ancestry groups. However, unlike the comparatively well-characterized chromosome 15q25.1 region, eQTL and mQTL mapping for the other ND candidate genes in human brain tissues have not been done yet. Furthermore because of the uniqueness of human brain

specimen collection, datasets with regulatory phenotypes in this particular tissue type are limited. And for the few datasets publicly available, as Hancock *et al*.[362] pointed out, linking DNA methylation, mRNA expression and ND in the same participants is usually not feasible. Here for the first time we connected genetic variation, DNA methylation, mRNA expression, and smoking status together using one cohort. Although smoking status is not as widely used in the smoking genetics field as ND-related phenotypes, they have been proven to efficiently tag some of the same linkage regions and susceptibility genes as ND traits.[45]

Specifically for this study, we had three objectives: 1) to pinpoint *cis*-regulatory loci (eQTL and mQTL) for a collection of ND susceptibility genes; 2) to annotate significant *cis*-eQTL and mQTL variants by integrating regulatory features from the Roadmap Epigenomics,[363] ENCODE,[364] and Genotype-Tissue Expression (GTEx)[365] Projects; and 3) to link *cis*-regulatory variants, DNA methylation, mRNA expression and smoking status together for exploration of possible biological mechanisms involved. We focus solely on mechanisms by which variants affect regulation of nearby genes (i.e., putatively in *cis*) in this study, as these are better understood and, moreover, likely represent the first step in most *trans*-acting QTLs as well.[46] Another reason is that, with the current sample size, this study does not have sufficient power to reliably identify *trans*-regulatory variants.

5.3 Materials and methods

5.3.1 The BrainCloud cohort and study samples

We used the BrainCloud cohort (http://braincloud.jhmi.edu/) to find variations associated with expression and methylation levels in 57 ND susceptibility genes.[45] The variation genotypes

(Illumina Human1M-Duo and HumanHap650Y arrays) were imputed with the 1000 Genomes Project phase 3 reference panel; and mRNA (Illumina Human 49 K Oligo array) and DNA methylation levels (Illumina HumanMethylation27 array) were available from post-mortem prefrontal cortex of human participants who had no neuropathological or neuropsychiatric diagnoses, no reported alcohol or drug abuse and no positive toxicology result.[366, 367] Please refer to **Chapter 4** for detailed information on the 57 genes. The gene cluster *CHRNA5/A3/B4* has been studied elsewhere and thus excluded from this study.[362] We obtained these data via the database of Genotypes and Phenotypes (dbGaP; accession number phs000417.v2.p1) and the BrainCloud project website (http://braincloud.jhmi.edu/). Gene expression data are also available through the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO series number GSE30272).

In the original dataset, SNP genotypes are available on 270 participants, while mRNA expression and DNA methylation data are available on subsets of 269 and 108 subjects, respectively. However, as reported by Colantuoni *et al*.[366] and Numata *et al*.,[367] fundamentally distinct expression and methylation profiles within fetal, infant and childhood development were observed, followed by continuous progressions of change throughout the rest of the lifetime. To eliminate significant effects of developmental life stages, and be consistent with common age onset for smoking, only postchildhood subjects (ages older than 10 years) were selected for analysis in this study, which resulted in 178 (94 African Americans [AAs] and 84 European Americans [EAs]) and 60 (31 AAs and 29 EAs) samples with expression and methylation data, respectively. Sample characteristics are presented in **Table 5.1**. All the 60 subjects with methylation levels also have transcriptional data available.

Table 5.1: Characteristics of study participants.

Sample	Expression		Methylation	
	AA	EA	AA	EA
Sample Size: *N*	94	84	31	29
Age (years): mean (SD)	39.8 (15.9)	36.8 (18.3)	44.2 (18.9)	43.3 (19.5)
Female: *N* (%)	34 (36.2)	25 (29.8)	15 (48.4)	13 (44.8)
Smoker: *N* (%)	30 (31.9)	16 (19.0)	8 (25.8)	6 (20.7)
PMI (hours): mean (SD)	33.3 (15.6)	27.4 (15.1)	33.8 (15.4)	29.0 (15.3)
PH: mean (SD)	6.6 (0.3)	6.5 (0.3)	6.6 (0.3)	6.5 (0.3)
RIN: mean (SD)	8.1 (0.7)	8.2 (0.8)	8.1 (0.6)	8.0 (1.0)

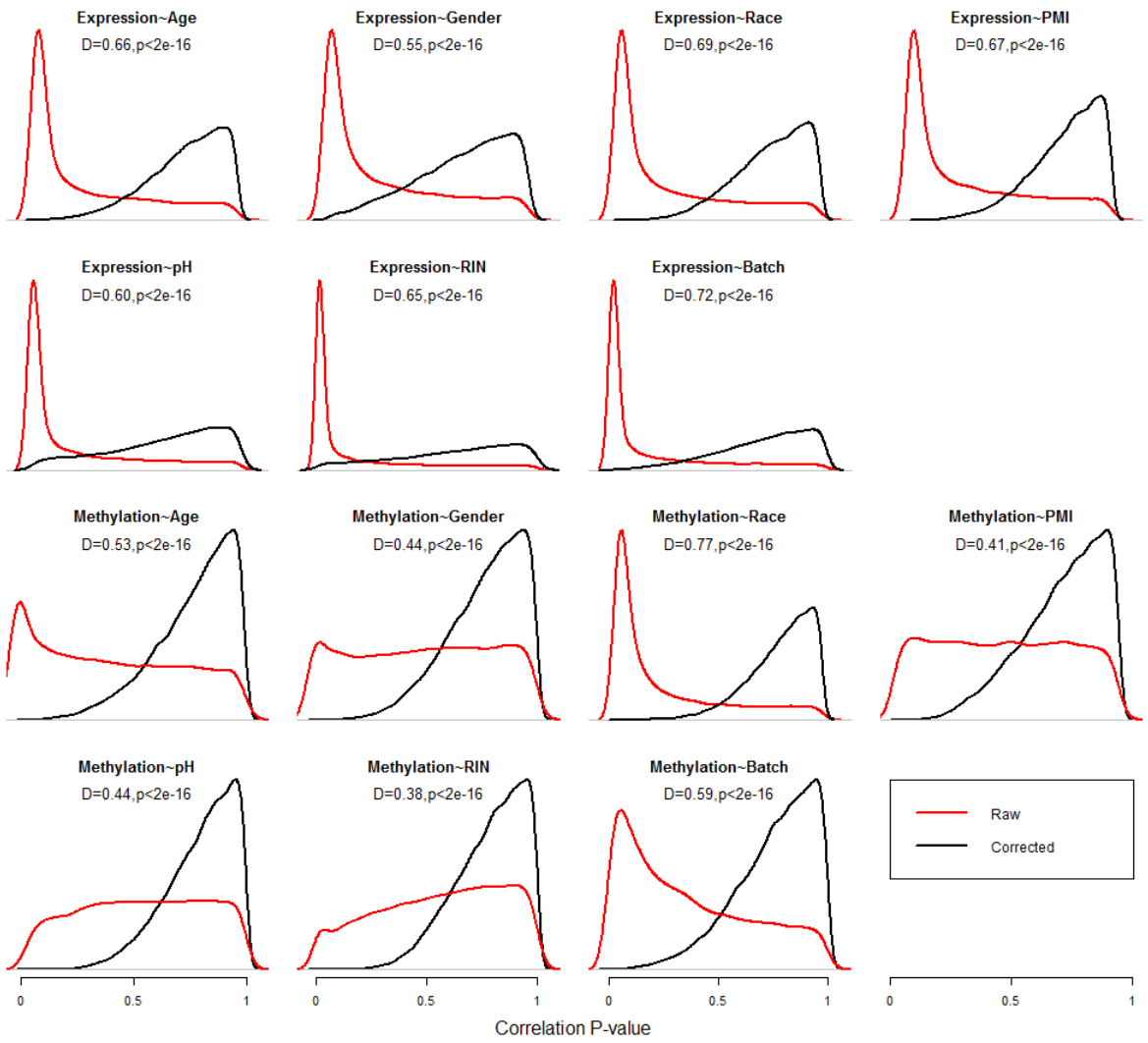
AA African American; *EA* European American; *SD* standard deviation; *PMI* post-mortem interval; *RIN* RNA integrity number.

5.3.2 Genome-wide covariate and surrogate variable analysis

We analyzed the contribution of each of the demographic (age, sex and race) and technical (PMI [post-mortem interval], pH, RIN [RNA integrity number] and batch) covariates to the genome-wide expression and methylation levels. The Kolmogorov-Smirnov (KS) test was used to capture the difference of the correlation p value distributions for each covariate. **Figure 5.1** shows the correlation p values between a specific covariate and all the probes based on the expression or methylation data before (red curve, raw) and after (black curve, corrected) adjustment of all the covariates listed above. Except for confounding of demographic variables, particular attention was paid to post-mortem quantitative factors including PMI, tissue pH status, RIN, and batch effects, the importance of which have been demonstrated in brain tissues.[368] The distinct distributions of p values before and after covariates correction, and the significant KS test results strongly suggest that adjusting for the covariates is necessary for the downstream analysis steps. Because of the significant allele frequency differences observed between the AA and EA samples (**Table 5.2**), a robust linear regression model for covariates

correction was implemented in each ethnic group separately, using R command like the following: rlm(expression/methylation~age+sex+PMI+pH+RIN+batch). We performed surrogate variable analysis on the obtained residuals across the genome, using the R package "sva".[369, 370] No surrogate variable was found in either AAs or EAs, which verified the elimination of known and unknown factors after the above covariates adjustment. These residuals, instead of the raw levels, were used for further analysis.

Figure 5.1: Correcting for covariate effects on the expression and methylation data.



Each plot shows the distribution of the correlation *P* values (X-axis) between a specific covariate and all the probes based on the expression or methylation data before (red curve, raw) and after (black curve, corrected) the covariates adjustment. The Kolmogorov-Smirnov (KS) test was used to capture the difference of the correlation *P* value distributions. The D statistics and *P* values of the KS test are shown here. Y-axis denotes the fraction of probes showing correlation with a given covariate.

5.3.3 Selection of probes and genotype imputation

One or more expression probes existed for 51 out of the 57 genes. For the 17 genes with more than one probes, the one covering all transcripts of the target gene and, at the same time, showing the highest intensity, was selected for all but two genes (*COMT* and *DLC1*), for which we made our decisions solely based on probe position and intensity, respectively. All the 51 selected expression probes contained no SNP in their respective sequences (**Supplementary Table 1**). For the methylation data, 107 CpG sites were identified and measured for 50 of the 57 genes (**Supplementary Table 1**). Among these sites, 74 (69%) are within CpG islands. The genotype imputation interval for each gene was determined based on the widest genomic range set by the gene start and end positions, and the CpG site position(s), plus 1Mb region on both sides.[367, 371-373] If the intervals for several adjacent genes were overlapped with each other, an all-inclusive interval was chosen on the corresponding chromosome for imputation (**Supplementary Table 2**).

Genotype imputation in each interval was conducted with reference to the 1000 Genomes Phase 3 integrated variant set release haplotype panel (October, 2014 release available at https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html), using IMPUTE2.[374] As Howie *et al*.[375] suggested, an ancestrally inclusive reference panel was used to improve imputation accuracy, especially for low-frequency variants (minor allele frequency [MAF] < 5%). The default settings in IMPUTE2 were used. We used a cutoff of 0.3 for the 'info' metric, comparable to the r-squared metrics reported by other programs like MaCH and Beagle,[376] to remove poorly imputed variations.[372, 377] Specific numbers of variations before and after imputation for each interval are shown in **Supplementary Table 2**. To have the same set of

variations for both ethnic groups, any variant loci with: (a) call rate of less than 95%, and (b) Hardy-Weinberg equilibrium (HWE) p value of less than 0.001 was excluded in the AA and EA combined sample with expression or methylation data.

5.3.4 Association QTL analysis and multiple testing correction

Genotype dosage data, including the imputed variations after quality control, were analyzed for association with expression and methylation phenotypes (the residuals described in the preceding section) using PLINK.[78] Linear regression analysis was performed to test for correlation between the residuals and the number of minor alleles for each variation via an additive genetic model. From this analysis, an asymptotic p value from the Wald statistic was obtained as a measure of association for each variation with any expression or methylation level of a given gene.

To correct for the large number of variations tested per phenotype, a region-wide empirical p value was computed for the asymptotic p value for each variation by using 1,000 max(T) permutations of label swapping provided within PLINK.[78] Swapping labels allows the linkage disequilibrium (LD) of the genomic regions being tested to be maintained across the observed and permuted samples. To correct for the number of phenotypes being investigated for expression or methylation, a false discovery rate (FDR) threshold was calculated based on the region-wide empirical p values, using the *fwer2fdr* function of the "multtest" package in R, where augmentation multiple testing procedure (AMTP) adjusted *P* values were obtained by adding suitably chosen null hypotheses to the set of null hypotheses already rejected by the initial permutation procedure. It is worth noting that variations within sequences of probes may

cause differential hybridization and inaccurate expression and possible methylation measurements. To exclude this confounding factor, because there is no variation within the expression probes in our case, only significant mQTL variants were removed from the results if they are at the CpG sites under investigation.

5.3.5 Variant annotation and *post hoc* analysis

For loci with aggregated QTL variants ($N \geq 10$), LD blocks were defined following Gabriel *et al*.[80] in Haploview (https://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview; **Figure 5.2** and **Supplementary Figures 1-3**). Tracks from the 1000 Genomes Browser (http://browser.1000genomes.org/index.html) were added to illustrate relative positions of the QTL variants with respect to their corresponding expression or methylation probes and other regulatory features found in multiple cell lines, such as promoter regions and transcription factor binding sites. Within each LD block identified, if genotyped variants were available, they were presented in **Tables 5.2-5.4**; otherwise, imputed variants with multiple pieces of evidence from HaploReg v4.1[378] (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) were displayed. HaploReg annotations for all significant QTL variants were included in **Supplementary Table 3**.

Since the *cis*-regulatory variants in *NRXN1*, *CYP2A7* and *EGLN2* account for the majority of the total significant QTL variants identified, which are more likely to be true signals,[379] we conducted post hoc analysis for these three genes. Pearson product-moment correlations between corresponding methylation and expression levels were examined. Although smoking

related phenotypes are limited for this cohort with only “smoke at death” and “smoking history” variables available, we assigned smoking status for each subject based on these two phenotypes: one person was classified as a smoker if he/she smoked at death with a positive or missing smoking history, or he/she did not smoke at death but with a positive smoking history; one person was only assigned to the non-smoker group if both his/her smoking traits were negative. Thirty-six subjects with missing smoking history plus missing or negative smoking at death were assorted as having missing smoking status. Please find specific numbers (percentages) of smokers in **Table 5.1**. No difference of expression or methylation profiles on the genome-wide level was detected between smokers and non-smokers in this cohort.

Student’s t-test was used to compare methylation or expression level difference between subjects with distinct race, smoking status, or genotypes. Fisher’s exact test was implemented in the analysis of contingency tables formed by any combination of subjects’ ethnic identity, smoking status, and variant minor allele copies. Because smoking can induce certain gene expression[380] and modify DNA methylation at particular genes,[381] we compared nested linear regression models using ANOVA to exclude influence of smoking on expression and methylation measurements for certain gene(s)/ CpG site(s), where significant association between smoking status and expression or methylation levels was found. Since outliers for expression and methylation data existed (**Figure 5.4-5.5**, **Supplementary Figures 4-6**), we confirmed the student’s t-test results using robust statistical methods based on Wilcox’ WRS functions as implemented in the R package “WRS2”.[382] All the above mentioned statistical tests were performed in R.

5.4 Results

5.4.1 *cis*-mQTL mapping

There are 138 *cis*-mQTL variants found for ten genes with region-wide significance, all of which are ethnic-specific except for *CYP2A7* (**Table 5.2**). Phenotype-wide significant variants account for 102 of them. Among the 138 region-wide significant variants, 42 and 68 are for two CpG sites in *NRXN1* (cg10917619) and *CYP2A7* (cg25427638), respectively. The *cis*-mQTL for *NRXN1*, with all the associated variants in complete LD, is EA-specific, while the one for *CYP2A7* is observed in both the AA and EA samples. Although the *cis*-mQTL variants for *CYP2A7* formed more than one haplotype block, the D' values among them are very high. Corresponding LD plots are in **Supplementary Figures 1** and **3**. Furthermore, 57 of the 68 region-wide significant variants in the *CYP2A7* mQTL showed phenotype-wide significance in both ethnic groups.

Annotations from HaploReg v4.1[378] corroborated the regulatory potential of the significant *cis*-mQTL variants for *NRXN1* and *CYP2A7* (**Table 5.2** and **Supplementary Table 3**). Thirteen of the 42 variants for *NRXN1* showed significant association with either enhancer or both enhancer and promoter histone marks in human brain tissues, based on core 15-state model predictions from the Roadmap Epigenomics Project.[363] A core set of 5 chromatin marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) assayed in 127 epigenomes was concatenated to train a ChromHMM model and compute posterior probabilities of 15 chromatin states for each variant in the Project.[363] Additionally almost all of the 42 variants except 4 (rs13031157, rs17573587, rs17514717, and rs13023341) were predicted to alter one or more of the regulatory motifs indicative of TF binding sites,[383] among which rs17514766 was detected to be bound with SP1 in H1-hESC cells.[364] The SP1 protein is a TF that binds to GC-rich

motifs of many promoters. Moreover, 16 of the 42 variants showed 1 to 3 independent QTL hits in the genome-wide repository of associations between SNPs and phenotypes (GRASP; http://grasp.nhlbi.nih.gov/Overview.aspx),[384] either as mQTL variants for the same CpG site (cg10917619) in temporal cortex, frontal cortex, or caudal pons,[372] or as loci significantly associated with blood metabolite concentrations or ratios.[385] Four variants, rs6545187, rs7574611, rs13031157, and rs7594170, significantly affected *NRXN1* expression in nerve tibial tissues based on the GTEx Project results.[365] Among them, rs7574611 was discovered as both a *cis*-mQTL and eQTL variant, and rs7594170 was significantly associated with both serum concentration of 17-dimethylurate and *NRXN1* expression levels in independent studies.[365, 372, 385] According to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) functional annotation, all of the 42 variants within a range of 121 Kb (distance between the two mQTL variants furthest away), are intronic except for rs67661616, which is within the 5'-UTR region of *NRXN1* (**Supplementary Figure 1**).

Unlike the *cis*-mQTL variants for *NRXN1*, a considerable proportion (18%) of the 68 variants for *CYP2A7* had significant associations with promoter or enhancer histone marks or DNase I hypersensitive sites in human induced pluripotent stem cells (iPSCs), embryonic stem cells (ESCs), or H1 derived neuronal progenitor cultured cells (ESDRs) instead of relevant brain tissues (**Table 5.2** and **Supplementary Table 3**).[363] However, similar as *NRXN1*, 61 of the 68 variants were predicted to change one or more of the regulatory motifs.[383] Eight variants affected TF binding in ChIP-Seq experiments of the ENCODE Project.[364] The involved TFs include glucocorticoid receptor (GR), transcriptional repressor CTCF, double-strand-break repair protein RAD21, transcriptional co-activating protein P300, signal transducer and activator of

transcription 3 (STAT3), estrogen receptor alpha (ERalpha_a), forkhead box protein A1 (FOXA1), and trans-acting T-cell-specific TF (GATA3). Regarding the GRASP results,[384] associations between 15 variants and methylation levels of cg25427638 were replicated in cerebellum, caudal pons, and frontal and temporal cortex tissues of an independent study with 150 neurologically normal Caucasian subjects.[372] The 68 variants overlap with *CYP2B6* and *CYP2A7P1* genes in a 35 Kb region (**Supplementary Figure 3**). If we look at the associated variants' effect sizes as measured by R^2, the square of the correlation coefficient, the effects of these *cis*-mQTL variants are larger in AAs than in EAs (**Table 5.2**).

Table 5.2: Significant CpG-variant pairs with region- or pheno-wide *cis* associations in AA or EA samples.

Chr	Gene	CpG	Variant ID	Variant Position	Distance to CpG	Imputed?	A1/A2	AA (N = 31)				EA (N = 29)				HaploReg Annotation
								A1 Freq	R^2	Region-wide *P*	Pheno-wide *P*	A1 Freq	R^2	Region-wide *P*	Pheno-wide *P*	
2	*NRXN1*	cg10917619*	rs7567632	51198384	-57243	N	C/T	0.10	0.13	1	1	0.19	0.51	**3.30E-02**	5.39E-02	mQTL in frontal & temporal cortex[372]
			rs6545187	51200546	-55081	N	C/T	0.23	0.23	9.98E-01	1	0.40	0.55	**1.50E-02**	**3.00E-02**	eQTL in nerve tibial[365]
			rs7574611	51214993	-40634	N	A/C	0.16	0.27	9.85E-01	1	0.34	0.70	**9.99E-04**	**2.00E-03**	DNase in 7 brain regions & fetal brain female;[363] mQTL in frontal & temporal cortex;[372] eQTL in nerve tibial[365]
			rs7594170	51237767	-17860	N	G/A	0.35	0.20	1	1	0.40	0.61	**4.00E-03**	**7.99E-03**	Enhancer in brain germinal matrix;[363] eQTL in nerve tibial[365]
			rs2163018	51304238	48611	N	T/C	0.11	0.14	1	1	0.19	0.51	**3.30E-02**	6.59E-02	mQTL in frontal & temporal cortex[372]
7	*PDE1C*	cg22131691	rs73303752	31495561	-615427	Y	G/A	0.11	0.55	**3.80E-02**	6.67E-02	0	NA	1	1	Enhancer in fetal brain[363]
	CHRM2	cg04748704	rs35259000	137153061	599818	Y	C/CTA	0.13	0.54	**1.50E-02**	**2.35E-02**	0.07	0	1	1	DNase in fetal brain female[363]
			rs1528099	137167984	614741	N	A/G	0.26	0.50	**3.80E-02**	6.99E-02	0.07	0	1	1	
	TAS2R38	cg25481253	rs6949261	141980921	307537	Y	G/C	0.42	0.46	**5.00E-02**	6.99E-02	0.09	0.03	1	1	
		cg03017475	rs1110074	141210054	-464287	N	T/C	0.24	0	1	1	0.47	0.46	**1.90E-02**	**3.13E-02**	
			rs4726505	141949743	275402	Y	A/C	0.41	0.50	**1.70E-02**	**3.20E-02**	0.09	0	1	1	
8	*CHRNA2*	cg02953306	rs78908411	27331607	-5019	Y	T/A	0.07	0.55	**8.99E-03**	**1.80E-02**	0	NA	1	1	Enhancer in neuron cells[363]
9	*DBH*	cg25020204	rs117956167	136466632	-33602	Y	G/C	0.02	0	1	1	0.04	0.52	**4.30E-02**	6.59E-02	Enhancer in brain germinal matrix[363]
10	*PTEN*	cg01228636	rs140133143	89751172	129399	Y	AT/A	0	NA	1	1	0.04	0.53	**2.50E-02**	**4.40E-02**	
			rs118039301	89759243	137470	Y	A/C	0	NA	1	1	0.04	0.53	**2.50E-02**	**5.00E-02**	
		cg16687447**	rs185979581	88920089	-703247	Y	T/C	0.04	0.83	**6.99E-03**	**1.40E-02**	0	NA	1	1	Regulatory motifs altered
11	*CHRM1*	cg13530039	rs2736595	61720538	-969019	Y	G/A	0.23	0.53	**3.50E-02**	**3.40E-02**	0.79	0.05	1	1	Enhancer in brain substantia nigra[363]
			rs2524294	61721234	-968323	Y	G/A	0.23	0.53	**3.50E-02**	**4.49E-02**	0.79	0.05	1	1	Active enhancer flank in 6 brain regions[363]
14	*NRXN3*	cg15572745	rs144623790	78026973	-843259	Y	AAT/A	0.25	0.01	1	1	0.05	0.46	**4.40E-02**	6.59E-02	
19	*CYP2A7*	cg25427638*	rs10500282	41508442	119085	N	C/T	0.39	0.56	**2.00E-03**	**4.00E-03**	0.29	0.27	6.45E-01	7.87E-01	
			rs11673270	41520844	131487	N	C/A	0.48	0.49	**8.99E-03**	**1.80E-02**	0.26	0.32	3.86E-01	5.05E-01	
			rs3745275	41531705	142348	N	A/G	0.42	0.59	**9.99E-04**	**2.00E-03**	0.31	0.48	**1.40E-02**	**2.80E-02**	mQTL in caudal pons, cerebellum, frontal & temporal cortex[372]
			rs3745277	41531915	142558	N	A/G	0.48	0.54	**3.00E-03**	**5.99E-03**	0.22	0.43	5.79E-02	8.79E-02	Proteins bound (H1-hESC): CTCF & RAD21;[364]

Chr	Gene	CpG	Variant ID	Variant Position	Distance to CpG	Imputed?	A1/A2	AA (N = 31)				EA (N = 29)				HaploReg Annotation
								A1 Freq	R^2	Region-wide *P*	Pheno-wide *P*	A1 Freq	R^2	Region-wide *P*	Pheno-wide *P*	
			rs10409701	41537868	148511	N	A/G	0.40	0.73	**9.99E-04**	**2.00E-03**	0.21	0.52	**8.99E-03**	**1.80E-02**	mQTL in caudal pons, cerebellum, frontal & temporal cortex[372] mQTL in caudal pons, cerebellum, frontal & temporal cortex[372]

AA African American; *EA* European American; *Imputed* whether the variant is imputed or not, "Y" means yes, "N" means no; *A1 Freq* allele frequency of A1; R^2 regression r-squared; *Region-wide P* corrected empirical *P* value based on 10^3 max(T) permutations with correction for the number of variants tested for *cis* associations at this CpG site; *Pheno-wide P* region-wide *P* value after correcting for the 107 CpG sites tested using augmentation multiple testing procedure. Genomic positions are based on the NCBI Build 37/hg19 assembly. Significant region- and pheno-wide *P* values are given in bold. Variant annotations were obtained from HaploReg v4.1 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php). The * symbol means only genotyped variants are given, which tag other imputed ones in the same LD block (**Supplementary Figures 1 and 3**). The ** symbol indicates only imputed variant(s) with the strongest support from HaploReg are given, while other imputed ones tagging the same mQTL can be found in **Supplementary Figure 2**.

Table 5.3: Significant expression probe-variant pairs with region- or pheno-wide *cis* associations in the AA sample.

Chr	Gene	Expression Probe	Variant ID	Variant Position	Distance to TSS	Imputed?	A1/A2	AA (N = 94)				HaploReg Annotation
								A1 Freq	R^2	Region-wide *P*	Pheno-wide *P*	
5	*DRD1*	HEEBO-029-HCC29B22 (hHC010798)	rs147731662	175600872	729709	Y	T/C	0.01	0.21	**4.60E-02**	9.19E-02	
7	*CACNA2D1*	HEEBO-026-HCC26O7 (hHC009943)	rs73386029	82036439	-36592	Y	C/A	0.02	0.25	**1.70E-02**	**3.40E-02**	
	CHRM2	HEEBO-014-HCC14F4 (hHC005116)	rs324650	136693661	140262	Y	A/T	0.67	0.17	**4.20E-02**	8.39E-02	
	HTR5A	HEEBO-023-HCC23I11 (hHC008651)	rs139998364	155628744	766134	Y	A/G	0.01	0.46	**1.80E-02**	**3.60E-02**	Enhancer in neuronal progenitor cells[363]
9	*DNM1*	HEEBO-058-HCC58E9 (hHC021993)	rs3824415	130145624	-820039	Y	A/G	0.22	0.19	**3.40E-02**	6.79E-02	eQTL for *SLC2A8* in cerebellum & temporal cortex[386] and nerve tibial[365]
14	*C14orf28*	HEEBO-013-HCC13H6 (hHR004782)	rs78410784	45639985	273478	Y	G/C	0.08	0.18	**3.60E-02**	7.19E-02	

AA African American; *TSS* transcriptional start site; *Imputed* whether the variant is imputed or not, "Y" means yes, "N" means no; *A1 Freq* allele frequency of A1; R^2 regression r-squared; *Region-wide P* corrected empirical *P* value based on 10^3 max(T) permutations with correction for the number of variants tested for *cis* associations with this expression probe; *Pheno-wide P* region-wide *P* value after correcting for the 51 expression probes tested using augmentation multiple testing procedure. Genomic positions are based on the NCBI Build 37/hg19 assembly. Significant region- and pheno-wide p values are in bold. Variant annotations were obtained from HaploReg v4.1 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php).

Table 5.4: Significant expression probe-variant pairs with region- or pheno-wide *cis* associations in the EA sample.

Chr	Gene	Expression Probe	Variant ID	Variant Position	Distance to TSS	Imputed?	A1/A2	EA (N = 84)				HaploReg Annotation
								A1 Freq	R^2	Region-wide *P*	Pheno-wide *P*	
11	*BDNF*	HEEBO-047-HCC47K19 (hHC017923)	rs72887755	27801789	58493	Y	A/G	0.02	0.22	**3.50E-02**	6.19E-02	
			rs116860953	27930226	186930	Y	G/A	0.02	0.21	**3.10E-02**	**3.70E-02**	
19	*EGLN2*	HEEBO-060-HCC60O16 (hHC023008) **	rs4802088	41255768	-49377	Y	T/C	0.03	0.30	**2.00E-03**	**4.00E-03**	Flanking active TSS in neuronal progenitor & neuron cells, 8 brain regions and fetal brain;[363] Enhancer in astrocytes primary cells;[364] DNase in neuronal progenitor cells[363] dbSNP annotation: 5′-UTR of *C19orf54*
			rs34406232	41305530	385	Y	A/C	0.02	0.29	**3.00E-03**	**5.99E-03**	Active TSS in neuronal progenitor & neuron cells, 7 brain regions and fetal brain female,[363] and also astrocytes primary cells;[364] Bivalent/poised TSS in brain germinal matrix;[363] DNase in neuronal progenitor[363] & astrocytes primary cells;[364] Protein bound (H1-hESC): POL2[364] Motif changed: Evi-1[383] dbSNP annotation: 5′-UTR of *EGLN2*

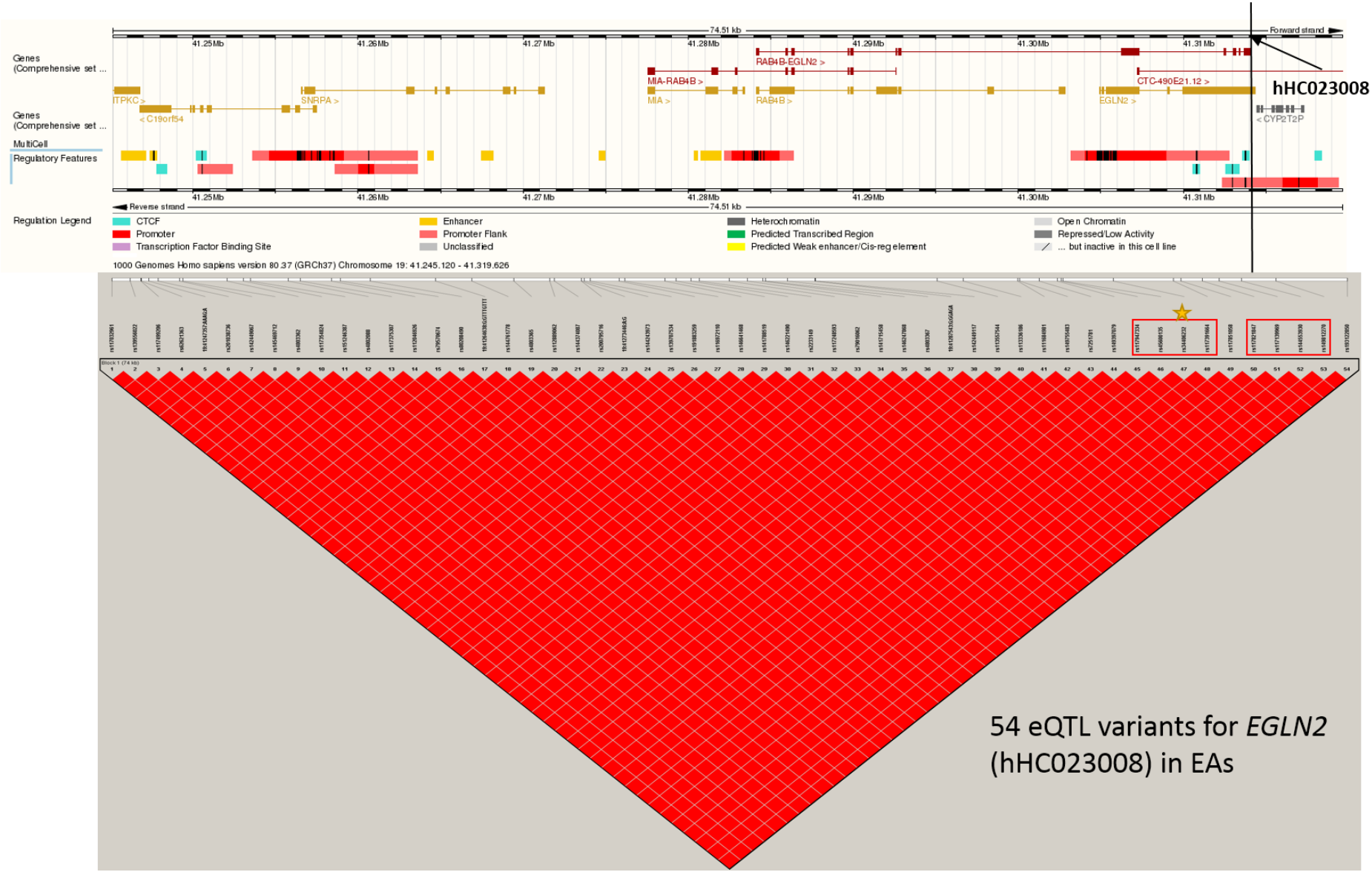
EA African American; *TSS* transcriptional start site; *Imputed* whether the variant is imputed or not, "Y" means yes, "N" means no; *A1 Freq* allele frequency of A1; R^2 regression r-squared; *Region-wide P* corrected empirical *P* value based on 10^3 max(T) permutations with correction for the number of variants tested for *cis* associations with this expression probe; *Pheno-wide P* region-wide *P* value after correcting for the 51 expression probes tested using augmentation multiple testing procedure. Genomic positions are based on the NCBI Build 37/hg19 assembly. Significant region- and pheno-wide p values are in bold. Variant annotations were obtained from HaploReg v4.1 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php). The ** symbol indicates only imputed variants with the strongest support from HaploReg are given, while other imputed ones tagging the same eQTL can be found in **Figure 5.2**.

5.4.2 *cis*-eQTL mapping

Six *cis*-eQTL variants were detected in the AA sample for six different genes based on the region-wide significance level (**Table 5.3**). Two of the 6 variants passed the phenotype-wide significance threshold: rs73386029 for *CACNA2D1* and rs139998364 for *HTR5A*. Fifty-six region-wide significant *cis*-eQTL variants were identified in the EA sample for two genes, *BDNF* and *EGLN2*, among which 54 variants were significantly associated with the expression level of *EGLN2*, measured by probe hHC023008 (**Table 5.4**). Fifty-three of the 54 variants for *EGLN2*, forming one LD block, showed pheno-wide significance as well (**Figure 5.2**).

Among the 54 *cis*-eQTL variants for *EGLN2*, 23 indicated chromatin states of promoter, enhancer, or DNase I activity cluster in human brain tissues, as annotated by HaploReg v4.1 (**Table 5.4** and **Supplementary Table 3**).[378] One or more regulatory motifs were changed for 48 of the 54 variants.[383] Bindings of proteins, such as DNA polymerase II (POL2) and nuclear phosphoprotein c-Myc, were influenced by 15 variants in different cell lines from the ENCODE Project.[364] However, none of the variants has been replicated as QTL variants for methylation, metabolic trait, or expression levels in any independent studies yet to our knowledge. The *EGLN2 cis*-eQTL has a span of 75 Kb, ranging from *ITPKC* to downstream of *EGLN2* based on both RefSeq (http://www.ncbi.nlm.nih.gov/refseq/) and GENCODE (http://www.sanger.ac.uk/science/tools) gene annotations (**Figure 5.2**). It is worth noting that all the associated variants within this *cis*-eQTL have low minor allele frequencies, between 1% and 5%, and their effects on mRNA expression in terms of R^2 are smaller than *cis*-mQTL variants' effects on DNA methylation (**Table 5.4**), both of which may lead to the lack of replication.

Figure 5.2: Linkage disequilibrium (LD) among the 54 eQTL variants for *EGLN2* in the BrainCloud EA sample.



The LD plot was drawn using Haploview (https://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview). Gene annotation and regulatory features tracks were obtained from the 1000 Genomes Browser (http://browser.1000genomes.org/index.html). The black vertical bar indicates position of the expression probe hHC023008. The red boxes highlight the eight rare variants, which were found to be collectively affecting smoking quantity by Clark *et al*.[387] The star marks the variant with the strongest biological evidence based on HaploReg v4.1 results (http://www.broadinstitute.org/mammals/haploreg/haploreg.php).

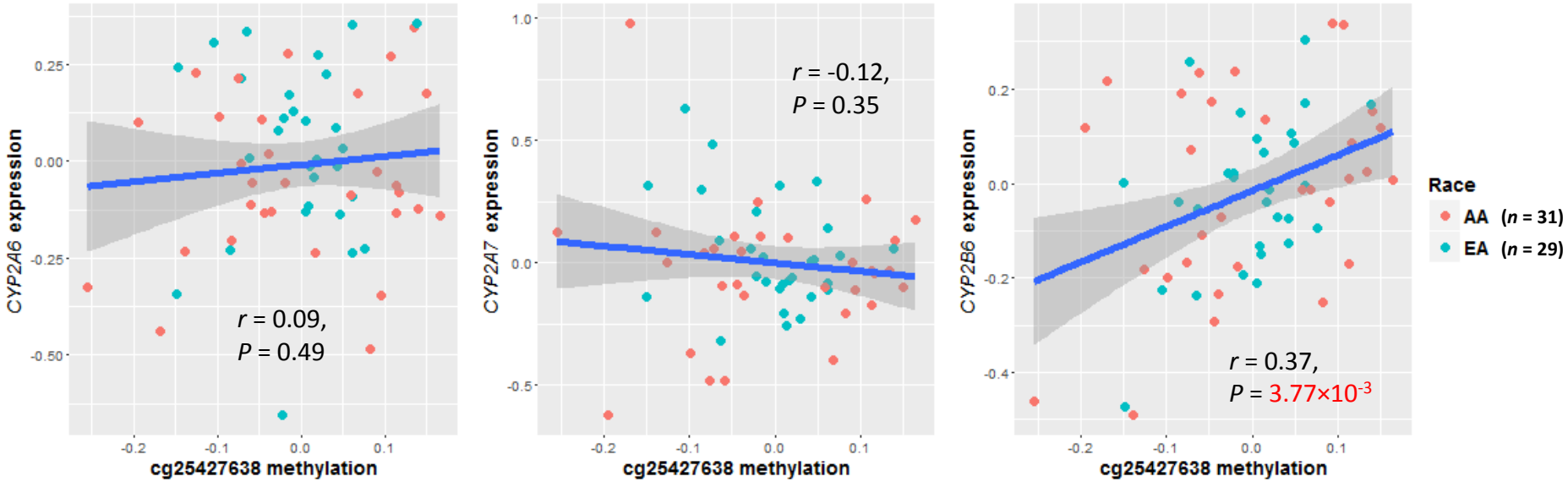
5.4.3 *Post hoc* analysis for *NRXN1*, *CYP2A7*, and *EGLN2*

Correlation between methylation and expression

Although previous studies have shown it is unlikely that there is a direct causal link between changes in the regulatory mechanism and differences in gene expression levels,[46, 371] the intuitive interpretation of a variant that belongs to an mQTL is that genetic variation results in a change in methylation, which in turn results in a change of the expression of a nearby gene. Because both methylation and expression levels were ethnicity-corrected, and the effect of race is not significant for the two measurements in the AA and EA pooled sample confirmed by t test results (n = 60, P > 0.05), to enlarge sample size, the pooled sample was used to test correlations between methylation levels of the two CpG sites for *NRXN1* and *CYP2A7*, and corresponding expression of their nearby gene(s).

For *NRXN1*, we did not observe a significant correlation between methylation levels at cg10917619 and *NRXN1* transcription (r = -0.26, P = 0.21; **Supplementary Figure 4a**). For the CpG site of *CYP2A7* (cg25427638), because three members of the cytochrome P450 gene family (*CYP2A6*, *CYP2B6*, and *CYP2A7*) are located next to each other with high sequence homology in this region, correlations between cg25427638 and expression levels of the three genes were tested. Methylation levels at cg25427638 were not significantly associated with *CYP2A6* (r = 0.09, P = 0.49) or *CYP2A7* expression (r = -0.12, P = 0.35), but significantly associated with expression of *CYP2B6* (r = 0.37, P = 3.77×10^{-3}; **Figure 5.3**). The significant positive correlation observed between cg25427638 and *CYP2B6* may indicate additional regulatory steps involved, which has been reported in other studies.[362, 371, 372]

Figure 5.3: Correlations between cg25427638 methylation and expression levels of *CYP2A6*, *CYP2A7*, and *CYP2B6* in the pooled sample, respectively.



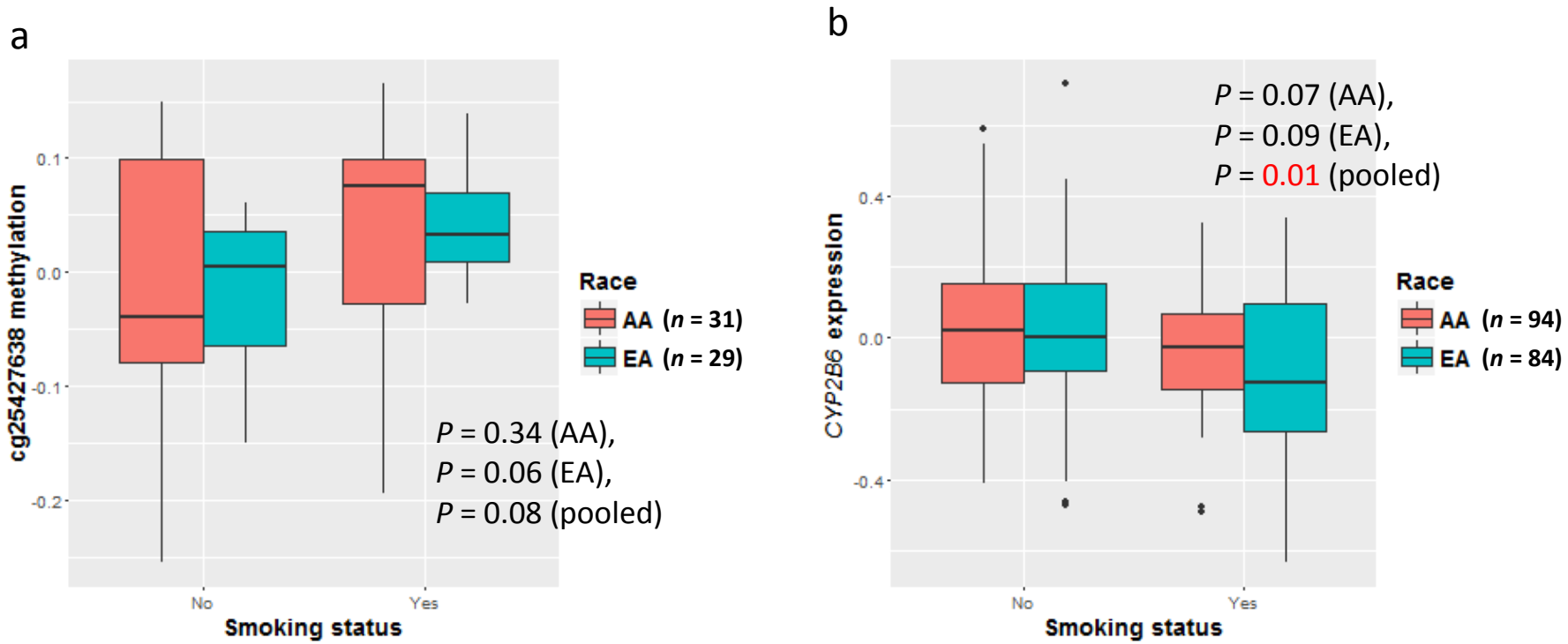
Each panel shows the scatterplot of cg25427638 methylation (X-axis) and expression level of one gene (Y-axis), superimposed with a linear regression line and its standard error region in the pooled sample. Data points are color-coded for AAs and EAs, with sample size included in the legend. Corresponding Pearson product-moment correlation and associated *P* value for the pooled sample are shown in each panel. Significant *P* value ($P < 0.05$) is marked in red.

Methylation and expression difference between smokers and non-smokers

For the next step, by utilizing the smoking status classification, we tried to detect whether there were any differences for cg10917619 (*NRXN1*) and cg25427638 (*CYP2A7*) methylation or *CYP2B6* and *EGLN2* expression between smokers and non-smokers. *CYP2B6* was examined due to its significant association with cg25427638 methylation. Ethnicity had nominally significant effect on smoking status in the pooled sample with only expression data (n = 178, P = 0.06), but not in the smaller sample with both expression and methylation levels (n = 60, P = 0.76). Results for the AA, EA, and pooled samples were all reported for the following analysis.

Methylation levels for cg10917619 were not significantly different between smokers and non-smokers in AA (P = 0.27), EA (P = 0.61), or the pooled sample (P = 0.78; **Supplementary Figure 4b**). However, methylation levels at cg25427638 showed nominally significant difference between smokers and non-smokers in the EA (P = 0.06) and pooled samples (P = 0.08), not in AAs (P = 0.34; **Figure 5.4a**). Because elimination of ethnicity effect on methylation measurements was confirmed previously, this sample difference is more likely due to sampling errors. For *CYP2B6* expression, significant difference between the two smoking status was found in the pooled sample (P = 0.01), but only nominal significance was detected for the AA (P = 0.07) and EA (P = 0.09) samples, respectively, due to the sample size issue (**Figure 5.4b**). Insignificant results were observed for expression levels of *EGLN2* (P = 0.96, 0.21, and 0.51 for the AA, EA, and pooled samples, respectively; **Supplementary Figure 5**).

Figure 5.4: Comparison of (a) cg25427638 methylation, and (b) *CYP2B6* expression between smokers and non-smokers.



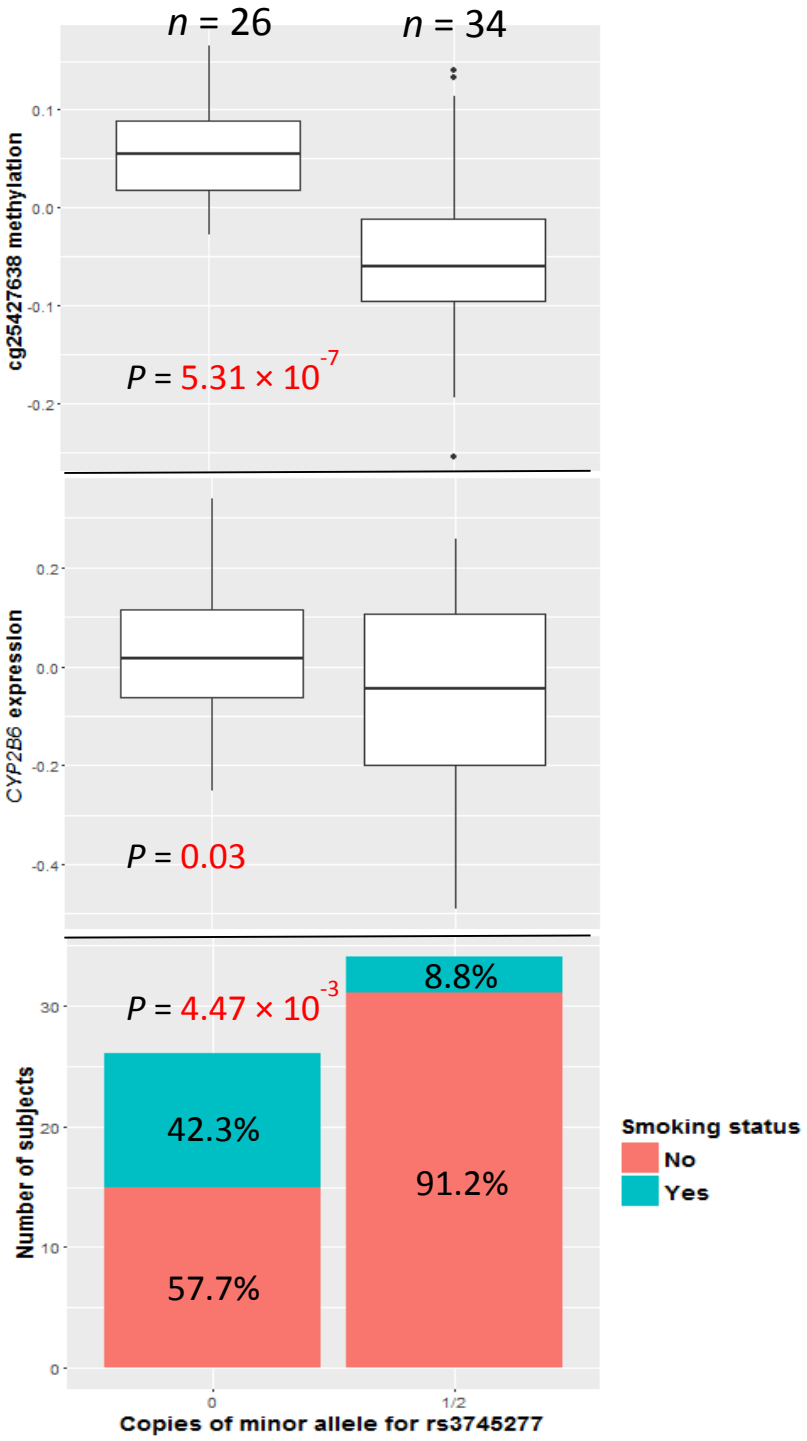
The two boxplots show methylation difference at cg25427638 (panel a) and expression difference for *CYP2B6* (panel b) for the two smoking status in each ethnic group, AA or EA. *P* values from Student's t test in the AA, EA, and pooled samples are also shown in the figure. Due to the existence of outliers, robust statistical methods were implemented to confirm the t test results. Corresponding sample size is included in the legend. Significant *P* value ($P < 0.05$) is marked in red.

Connection of genetic variation, methylation, expression, and smoking status

With clustering of *cis*-mQTL variants in *CYP2B6* and *CYP2A7P1* genes, significant correlation between the mQTL associated methylation site at *CYP2A7* (cg25427638) and *CYP2B6* expression, significantly different cg25427638 methylation and *CYP2B6* expression between smokers and non-smokers, we picked one genotyped variant (rs3745277), which had strong biological evidence based on HaploReg v4.1[378] results (**Table 5.2**), to check the effect of its minor allele on cg25427638 methylation, *CYP2B6* expression, and smoking status simultaneously. As shown in **Figure 5.5**, subjects with one or two copies of the minor allele (A) of rs3745277 (dominant effect) had a significant decrease in their methylation levels at cg25427638 ($P = 5.31 \times 10^{-7}$), a weak decrease in their expression levels of *CYP2B6* ($P = 0.03$), and a significantly lower percentage of smokers (8.8% vs. 42.3%, *OR* [95% *CI*] = 0.14 [0.02–0.62], $P = 4.47 \times 10^{-3}$). We also performed same sets of analysis for AAs and EAs separately, where similar patterns of results were observed (**Supplementary Figure 6**).

Linear regression models including smoking status as a covariate did not significantly improve model fits between copy numbers of rs3745277 minor allele and *CYP2B6* expression (ANOVA $P = 0.99$) or cg25427638 methylation (ANOVA $P = 0.87$), which means the influence of smoking on expression of *CYP2B6* and methylation at the cg25427638 site in the pooled sample is likely to be minimal. Same insignificant results ($P > 0.05$) were found when we performed separate analysis in the AA and EA samples.

Figure 5.5: Comparisons of cg25427638 methylation, *CYP2B6* expression, and smoker percentage between subjects with zero copy of rs3745277 minor allele verse one and two copies combined.



The two boxplots indicate methylation levels at cg25427638 (top) and expression amounts for *CYP2B6* (middle), respectively, for subjects with 0 or 1/2 copies of rs3745277 minor allele. Student's t test results are included in these two panels. Due to the existence of outliers as shown in the top panel, robust statistical methods were implemented to confirm the t test results. The barplot at the bottom illustrates smoker and non-smoker percentages for each genotype group. Fisher's exact test was performed to obtain the *P* value for this panel. Corresponding sample size for each minor allele copy group is included on the top of the figure. All the *P* values are significant ($P < 0.05$) and marked in red.

5.5 Discussion

In this work, we mapped *cis*-eQTL and mQTL for 57 smoking susceptibility genes in human brain. Six (for *DRD1*, *CACNA2D1*, *CHRM2*, *HTR5A*, *DNM1*, and *C14orf28*) and two (for *BDNF* and *EGLN2*) *cis*-eQTLs were detected in the AA and EA samples, respectively (**Tables 5.3** and **5.4**). Eight *cis*-mQTLs were found for seven genes (*PDE1C*, *CHRM2*, *TAS2R38*, *CHRNA2*, *PTEN*, *CHRM1*, and *CYP2A7*) in the AA sample, with two for two different CpG sites of *TAS2R38*. And six *cis*-mQTLs were located for six different genes (*NRXN1*, *TAS2R38*, *DBH*, *PTEN*, *NRXN3*, and *CYP2A7*) in the EA sample (**Table 5.2**). Only the *cis*-mQTL for *CYP2A7* is the same across the two ethnic groups. As reported by others that same or overlapped QTLs associated with both DNA methylation and mRNA expression are uncommon,[372, 377] we did not observe overlap for the *cis*-eQTLs and mQTLs identified for the 57 candidate genes. Among the QTLs determined, the *cis*-eQTL for *EGLN2* accounts for 54 (87%) out of the 62 total significant variants for all the *cis*-eQTLs; the *cis*-mQTLs for *NRXN1* and *CYP2A7*, respectively, make up 42 (30%) and 68 (49%) out of the 138 significant *cis*-mQTL variants. Based on the most recent 1000 Genomes Project results, where we obtained the imputation reference panel, overall, a typical eQTL signal comprised 67 associated variants.[379] Same aggregation phenomenon was observed in the original BrainCloud studies.[366, 367] Thus we conducted post hoc analysis for *EGLN2*, *NRXN1*, and *CYP2A7*, which are more likely to be true signals. The following discussion will also focus on the three genes.

All the 54 significantly associated variants within the *cis*-eQTL of *EGLN2* had strong biological evidence to affect gene expression according to HaploReg v4.1[378] results (**Supplementary Table 3**). Rs34406232, for example, is at active transcriptional start site (TSS)

not only in neuronal progenitor, neuron, and astrocytes primary cells,[363, 364] but also in eight different human brain regions including dorsolateral prefrontal cortex.[363] It is also within a DNase I hypersensitive site, which is characterized by open and accessible chromatin for active genes. More importantly, in H1-hESC cells, this variant interacted with POL2,[364] and changes a regulatory motif for Evi-1,[383] which positively regulates transcription from RNA POL2 promoter. Additionally, rs34406232 resides in the 5'-UTR region of *EGLN2* based on dbSNP annotation, while recent studies have suggested that transcriptional regulation near the 5' ends of genes (rather than RNA decay) might be exerting the strongest amount of control on gene expression levels.[46] All of these evidence makes rs34406232 highly likely to be a functional regulatory variant for *EGLN2* expression (**Table 5.4**). However, because all the 54 *cis*-eQTL variants are in complete LD with each other (**Figure 5.2**), and they may cause changes in *EGLN2* mRNA expression individually or collectively, we cannot pin down rs34406232 as the causal variant.

When we compared *EGLN2* expression levels between smokers and non-smokers, no significant difference was found (**Supplementary Figure 5**), which is not surprising, because variants within *EGLN2* were reported to be associated with cigarettes smoked per day (CPD), breath carbon monoxide (CO), ND, smoking efficiency (CO/CPD), and nicotine metabolite ratio (NMR),[23, 164, 263] but not directly with smoking status. Bloom *et al*.[164] indicated that multiple SNPs in high LD with rs3733829, a GWAS hit but without clear functional consequence prediction,[23] reside in the 5'-UTR or promoter region and may impact *EGLN2* transcription in subjects of European descent. It is even more interesting that Clark *et al*.[387] targeted captured the *EGLN2* region followed by next-generation deep sequencing (mean coverage 78 ×) in 363 individuals of European ancestry, and identified variant sets with regulatory annotations like

gene promoters, and chromatin states involving or flanking active TSSs significantly associated with CPD. Eight of the 54 *cis*-eQTL variants for *EGLN2* in this study were detected by Clark *et al*. with similar minor allele frequencies,[387] and were included in both individual and sets of variants analysis. However, significant association with CPD was only found for different variant sets. These results gave us great confidence in the imputation quality of IMPUTE2[375] for low-frequency variants, and make us prone to believe that the *cis*-eQTL variants determined here affect *EGLN2* transcription in aggregate, which then influences CPD or other smoking traits mentioned above.

Although the 42 *cis*-mQTL variants for *NRXN1* were overrepresented with enhancer and promoter histone marks in human brain tissues,[363] 36% of the variants were replicated as mQTL variants with the same CpG site in three human brain regions by an independent study,[372] and four of them affected *NRXN1* expression in nerve tibial tissues,[365] none of these variants were identified as *cis*-eQTL variants in this study (**Supplementary Table 3**). We did not detect correlation between methylation levels at cg10917619 and *NRXN1* expression, or a difference of methylation at this site between smokers and non-smokers, either (**Supplementary Figure 4**). The two nonsynonymous rare variants (p.R206L and rs77665267[p.T274P]), reported by our group having an aggregate effect on smoking status in 1430 unrelated EA subjects, are 903 Kb away from the *cis*-mQTL, which are more likely to affect protein coding directly.[44] Rs2193225 and one major haplotype including this SNP, 32 Kb downstream from the *cis*-mQTL, were significantly associated with different measures of ND in an EA family sample with 671 subjects.[112] Bierut *et al*.[34] nominated three SNPs (rs12623467, rs12467557, and rs10490162) within the mQTL to be associated with ND status in smokers. However, none of these variants

were in LD with the *cis*-mQTL either in the BrainCloud EA or the 1000 Genomes EUR sample. Thus the result for *NRXN1* in this study is no more than a *cis*-mQTL for cg10917619, whose involvement in mRNA expression and ultimately smoking is obscure. Its connection with previous association results and the underlying biological mechanism for those association signals need further investigation.

The *cis*-mQTL for *CYP2A7* was identified in both the AA and EA samples. Fifteen of the variants were replicated with the same CpG site in four human brain regions by an independent study.[372] But unlike the QTLs for *EGLN2* and *NRXN1*, this QTL is 116 Kb away from the CpG site (cg25427638) for *CYP2A7* (**Supplementary Figure 3**). Yet architectural proteins including cohesin (subunit RAD21) and CTCF were found to be bound at several variant loci of the *cis*-mQTL, which can bring together enhancers and gene promoters that may be located far apart in linear sequence.[388] When we tried to correlate methylation at cg25427638 with gene expression, *CYP2B6* instead of *CYP2A7* and *CYP2A6* showed significance (r = 0.37, $P = 3.77 \times 10^{-3}$; **Figure 5.3**). The specific mechanisms among the *cis*-mQTL, cg25427638 methylation in *CYP2A7*, and *CYP2B6* expression are not clear. Further, *CYP2B6* expression was significantly different between smokers and non-smokers, while methylation levels of cg25427638 indicated nominally significant difference (**Figure 5.4**). Observation of these differences may benefit from larger sample size, since no difference was found in frontal cortex of a previous study with 26 subjects.[389] By using one genotyped variant with strong biological evidence (rs3745277; **Table 5.2**), we demonstrated its effect on cg25427638 methylation, *CYP2B6* expression, and smoking status simultaneously (**Figure 5.5**). Other variants within the *cis*-mQTL are expected to have the same regulatory effects on methylation, expression, and phenotypic layers, due to high LD

among the 68 variants. Same as the *cis*-eQTL for *EGLN2*, we cannot determine causal variant(s) within the region.

Several variants in this region were reported to be associated with either CPD or nicotine metabolism, which includes rs4105144, a CPD GWAS hit,[24] rs7260329, an NMR GWAS hit,[263] and rs8109525, associated with nicotine metabolism independent of the well-studied nonsynonymous variants rs3211371, rs3745274, and rs2279343 (*CYP2B6*5* and **6*).[390] Additionally, 16 CpG sites within 19q13 were affected by the NMR GWAS hit (rs7260329) using whole blood DNA,[263] and rs8109525 was associated with differences in *CYP2B6* mRNA expression in liver biopsy samples.[390] However, neither rs7260329 nor rs8109525 was in LD with the *cis*-mQTL in the BrainCloud cohort while rs4105144 was not available. All the three variants were not in LD with the mQTL in either the 1000 Genomes AFR or EUR samples. Regulatory differences among tissues (blood, brain, and liver in this case) exist.[46] For the three functional variants (rs3211371, rs3745274, and rs2279343), unfortunately, they were not available in this study either. Even though rs2279343 and rs3211371 were not in LD with the mQTL in AFR and EUR samples of the 1000 Genomes Project, rs3745274 was found to be in high LD ($r^2 \geq 0.8$) with 4 and 8 mQTL variants in the 1000 Genomes AFR and EUR samples, respectively. Because rs3745274 and rs2279343 were reported to be linked, and these functional variants were detected to affect expression and activity of CYP2B6 in liver,[391] LD between rs3745274 and the mQTL is at least dubious, and should not influence our interpretation of the *cis*-mQTL in human brain.

Although the role of CYP2B6 in hepatic nicotine metabolism to cotinine is minor (~10%) relative to CYP2A6, it is expressed in the brains of both nonhuman primates and humans, thus

potentially modulating central nervous system nicotine metabolism and the duration of action of nicotine in the brain.[392] Recently Garcia *et al*.[393] showed that a mechanism-based inhibitor selective for CYP2B, C8-xanthate, increased the percentage of rats that acquired self-administration from 40% after vehicle pretreatment to 100%, with no difference in peripheral nicotine levels measured at the end of behavior (rats were given intracerebroventricular pretreatment with C8-xanthate/ASCF and underwent intravenous nicotine self-administration). This strongly corroborates our observation of decreased *CYP2B6* expression in smokers compared with non-smokers. As chronic nicotine treatment could induce *CYP2B6* expression in African Green monkey brain, especially CYP2B6 protein levels were induced 1.5-fold in the frontal cortex,[380] basal *CYP2B6* expression levels for subjects carrying 1 or 2 copies of rs3745277 or other *cis*-mQTL variants were anticipated to be even lower. Lee *et al*.[394] discovered that phenobarbital treatment significantly induced CYP2B6 protein levels in all African Green monkey brain regions including frontal cortex. Combining this observation with another finding from Garcia *et al*.[393] that C8-xanthate increased the number of sessions required to meet extinction criteria, phenobarbital used as precision medicine for subjects carrying minor alleles of the *cis*-mQTL variants is likely to better treat smoking.

There are five limitations for this study. First is sample size. Although even using modest sample sizes (60-100 individuals), early studies found a large number of genetic associations with differences in gene regulation.[46] We acknowledge that more *cis*-eQTLs and mQTLs are anticipated if larger sample sizes are available. Second is its inability to map *trans*-regulatory variants, because heritability studies suggest that more than half of the genetically explained variance in gene expression is due to *trans*-acting variants.[395] But reliable detection of *trans*-

QTLs has been challenging in humans due to the smaller effect sizes of *trans*-acting variants compared with *cis*-QTLs and a higher statistical penalty for multiple testing.[46] Besides enlarging sample sizes, focusing on QTLs affecting the expression levels of putative *trans*-regulatory elements (thereby minimizing the number of tests performed) might be another promising approach.[46] Third, other mechanisms exist by which variants can affect gene expression beside DNA methylation, which we were not able to investigate here, such as transcriptional elongation (by POL2 travelling rates), mRNA processing, modification, and degradation, defects in polyadenylation, and targeting by miRNAs. Fourth, although post-mortem prefrontal cortex grey matter tissue homogenates were used to measure mRNA expression and DNA methylation levels in the BrainCloud cohort, a highly relevant brain region for studying smoking,[396-398] projects have found that active regulatory regions and non-coding transcripts are often cell- and tissue-specific, and regulatory differences between tissues or different cell types are of a much larger magnitude.[46] Considering physiological response to nicotine is a complex process involving multiple brain regions,[399] and the brain is built from a large number of cell types,[400] in-depth investigation of gene regulation in different brain regions and cell types is anticipated, for example by using the PsychENCODE data.[401] Last the smoking status phenotype derived from two variables in the original study is primitive. Limited sample size of this study may magnify the sampling error for this phenotype, and statistical power may be reduced to connect molecular measurements with smoking status, compared with other more refined and commonly used ND traits, e.g., CPD, FTND and ND classification based on Diagnostic and Statistical Manual of Mental Disorders (DSM).

Despite limitations, this is the first study integrating data for genetic variation, DNA methylation, mRNA expression, and smoking phenotype together in the same cohort. The connection we found among *cis*-mQTL, CpG site methylation in *CYP2A7*, *CYP2B6* expression, and smoking status received vigorous support from projects enabling parallel sequencing of regulatory features, i.e., the Roadmap Epigenomics,[363] ENCODE,[364] and GTEx projects.[365] Further studies are warranted to test and verify the order of occurrence for these regulatory layers, or weave other factors into the picture such as chromatin accessibility, histone modifications, and TF binding using one cohort if possible. Through this way, we are not only able to learn about mechanistic steps between genetic variation and smoking effectively, but also narrow down functional, causal variant list efficiently.

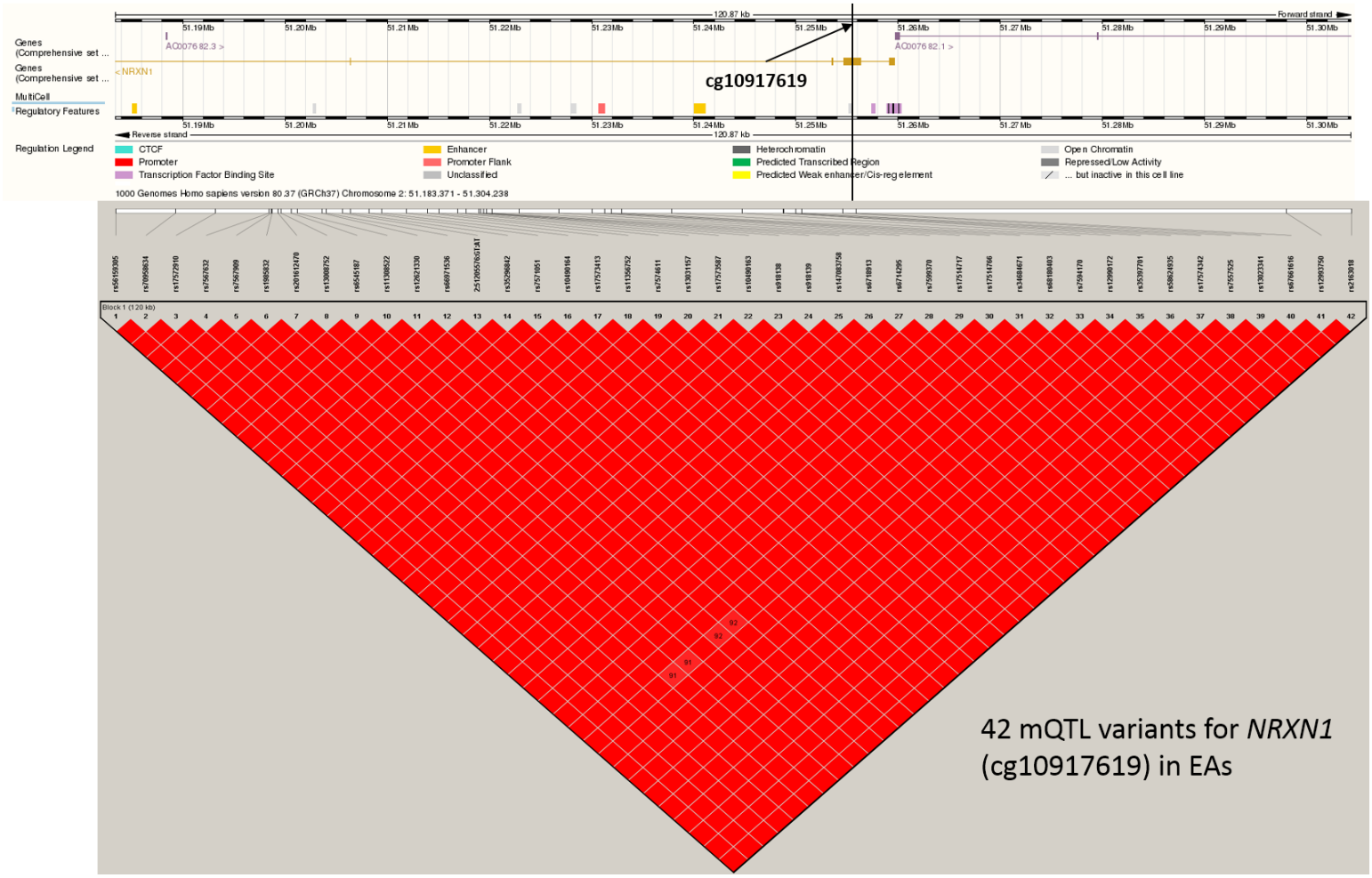
5.6 Chapter acknowledgements

This chapter was adapted from a manuscript in preparation, authored by Jiekun Yang and Ming Li. Stefan Bekiranov, Charles Farber, Stephen Rich, and Michael Timko provided constructive conversations that aided in the completion of this work.

5.7 Supplementary data

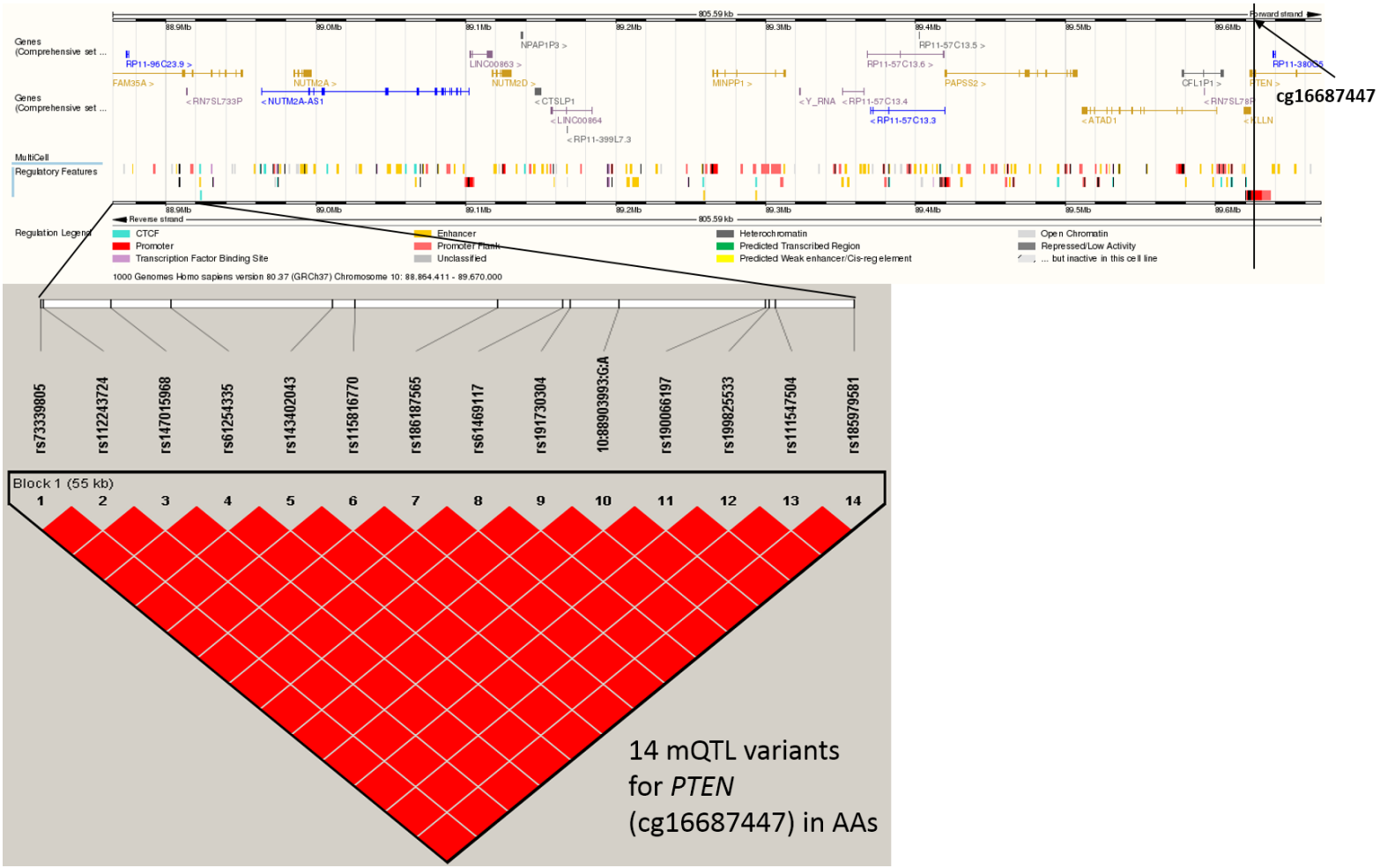
Supplementary Table 3 will be in a separate file named Yang_Jiekun_2016_Sup1.xlsx.

Supplementary Figure 1: Linkage disequilibrium (LD) among the 42 mQTL variants for *NRXN1* in the BrainCloud EA sample.



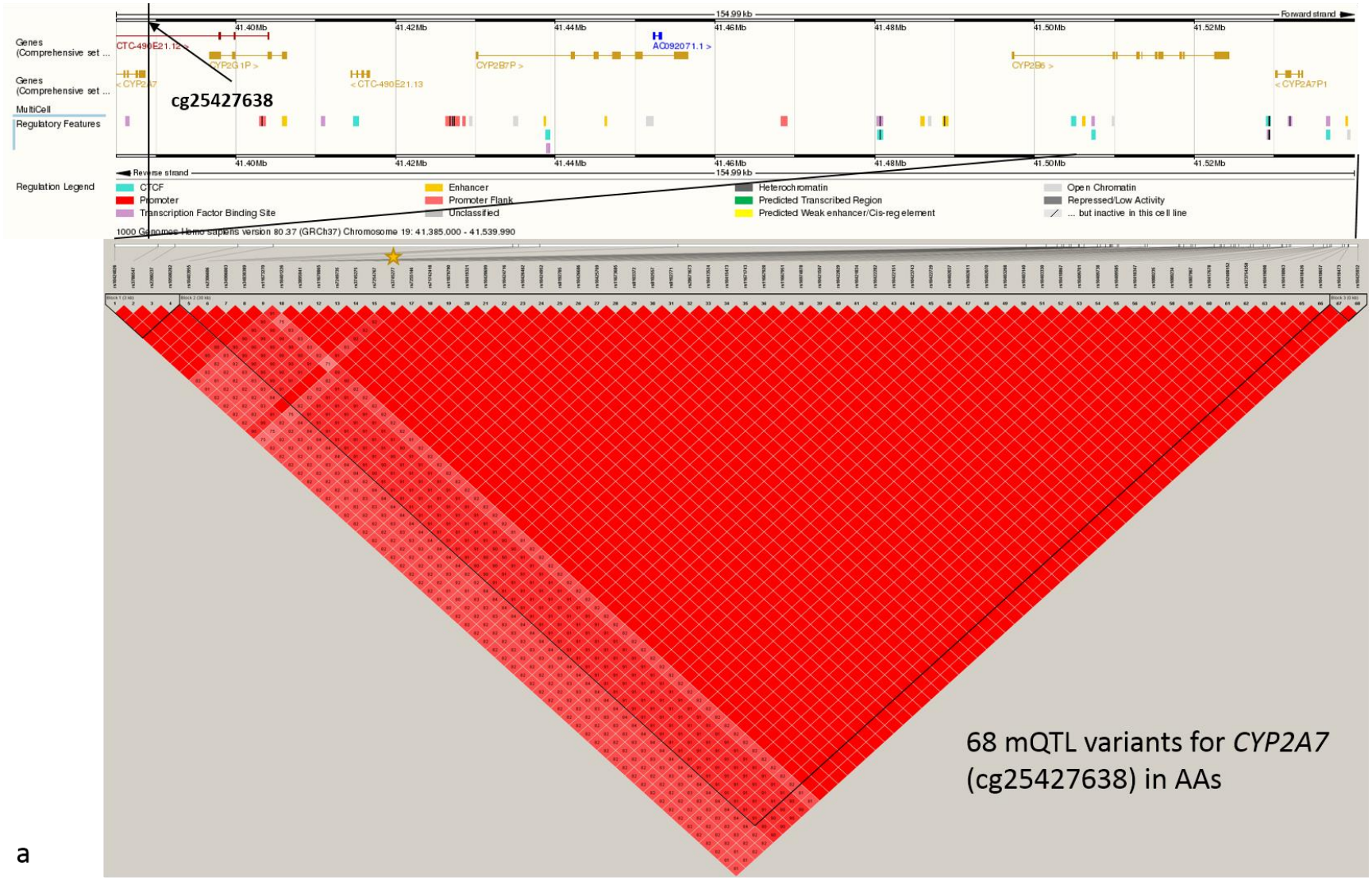
The LD plot was drawn using Haploview (https://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview). Gene annotation and regulatory features tracks were obtained from the 1000 Genomes Browser (http://browser.1000genomes.org/index.html). The black vertical bar indicates position of the methylation probe cg10917619.

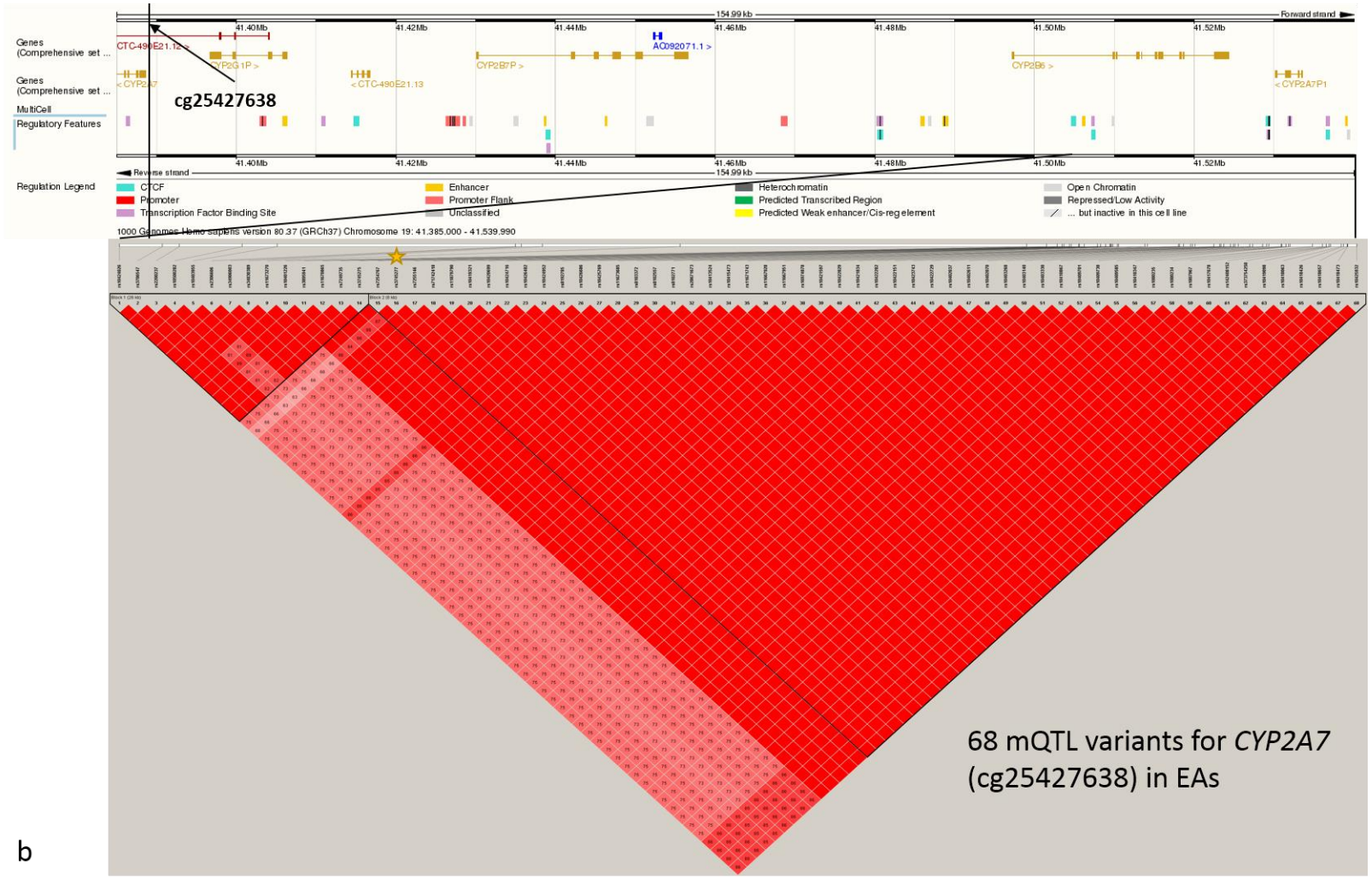
Supplementary Figure 2: Linkage disequilibrium (LD) among the 14 mQTL variants for *PTEN* in the BrainCloud AA sample.



The LD plot was drawn using Haploview (https://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview). Gene annotation and regulatory features tracks were obtained from the 1000 Genomes Browser (http://browser.1000genomes.org/index.html). The black vertical bar indicates position of the methylation probe cg16687447.

Supplementary Figure 3: Linkage disequilibrium (LD) among the 68 mQTL variants for *CYP2A7* in the BrainCloud AA (panel a) and EA (panel b) sample.

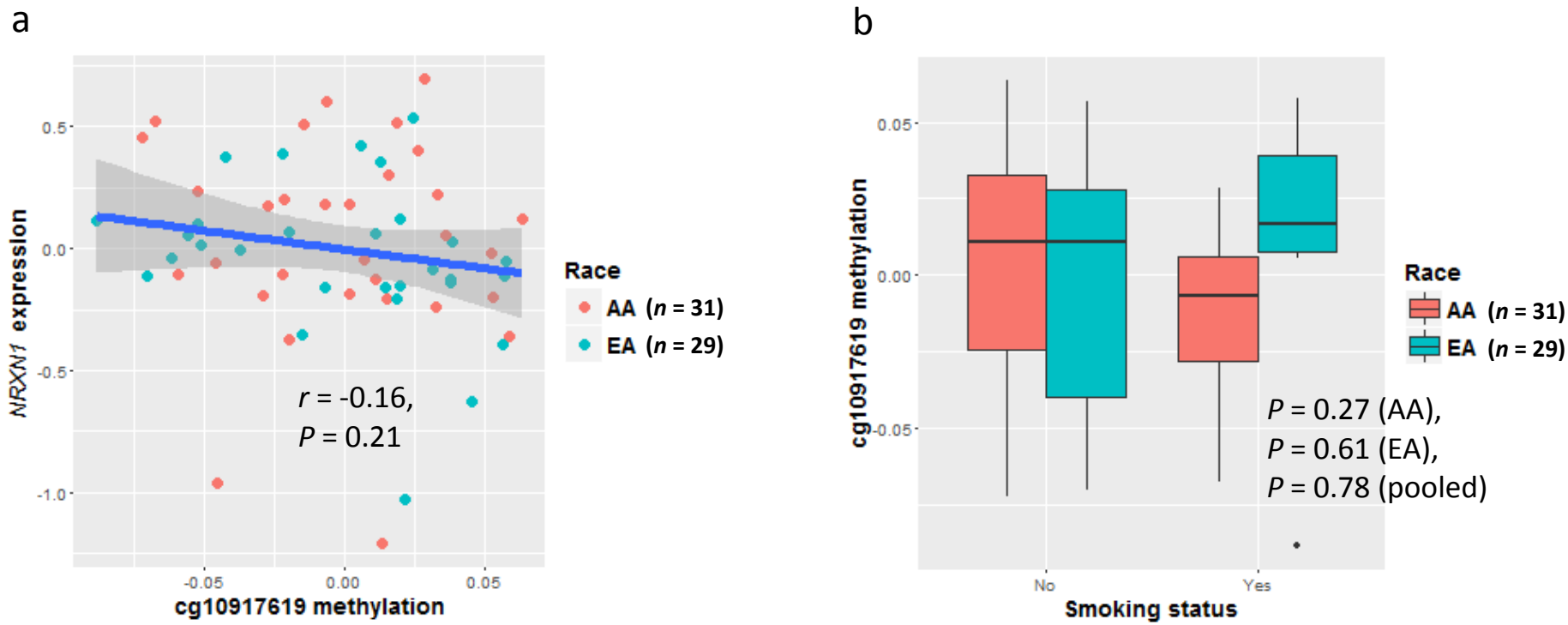




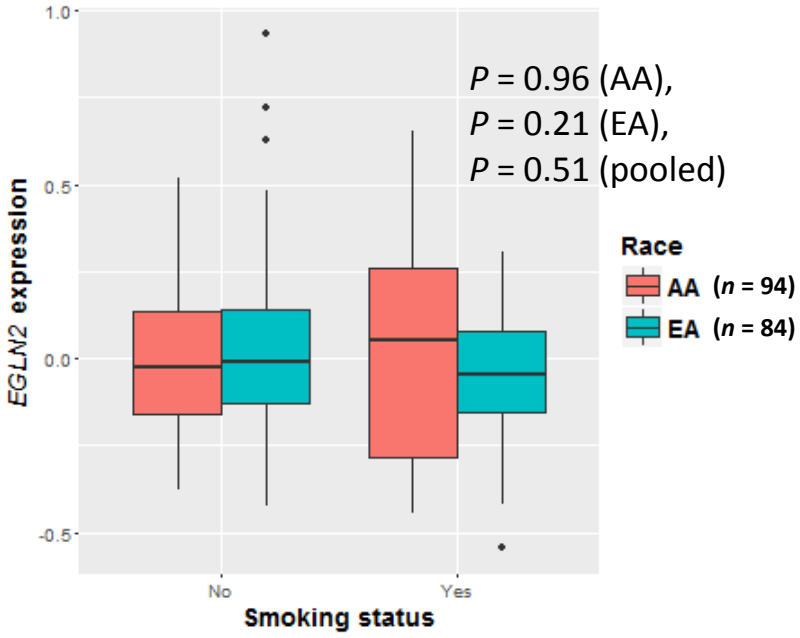
b

The LD plots were drawn using Haploview (https://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview). Gene annotation and regulatory features tracks were obtained from the 1000 Genomes Browser (http://browser.1000genomes.org/index.html). The black vertical bars indicate position of the methylation probe cg25427638. The star marks the variant with the strongest biological evidence based on HaploReg v4.1 results (http://www.broadinstitute.org/mammals/haploreg/haploreg.php).

Supplementary Figure 4: (a) Correlation between cg10917619 methylation and expression levels of *NRXN1* in the pooled sample. The scatterplot shows correlation between cg25427638 methylation (X-axis) and expression level of *NRXN1* (Y-axis), superimposed with a linear regression line and its standard error region in the pooled sample. Data points are color-coded for AAs and EAs. Corresponding sample size is included in the legend. Pearson product-moment correlation and its associated *P* value for the pooled sample are shown in the figure as well. **(b) Comparison of cg10917619 methylation levels between smokers and non-smokers.** The boxplot shows methylation difference at cg10917619 for the two smoking status in each ethnic group, AA or EA. *P* values from Student's t test in the AA, EA, and pooled samples are also shown in the figure. Due to the existence of outliers, robust statistical methods were implemented to confirm the t test results. Corresponding sample size is included in the legend.

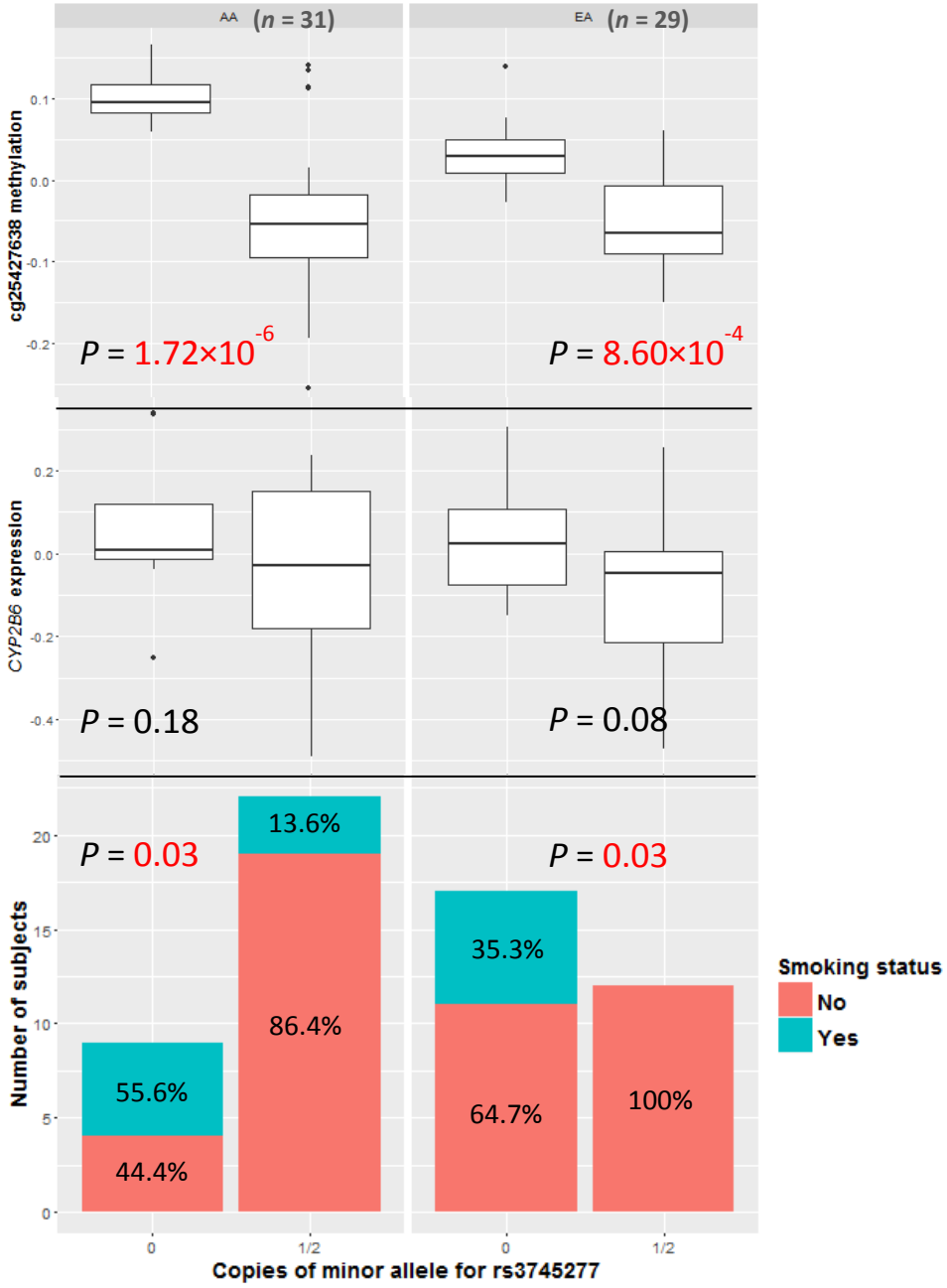


Supplementary Figure 5: Comparison of *EGLN2* expression levels between smokers and non-smokers.



The boxplot shows expression difference of *EGLN2* for the two smoking status in each ethnic group, AA or EA. *P* values from Student's t test in the AA, EA, and pooled samples are also shown in the figure. Due to the existence of outliers, robust statistical methods were implemented to confirm the t test results. Corresponding sample size is included in the legend.

Supplementary Figure 6: Comparisons of cg25427638 methylation, *CYP2B6* expression, and smoker percentage between subjects with zero copy of rs3745277 minor allele verse one and two copies combined in AAs or EAs separately.



The four boxplots indicate methylation levels at cg25427638 (top) and expression amounts for *CYP2B6* (middle), respectively, for subjects with 0 or 1/2 copies of rs3745277 minor allele in the AA (left) or EA (right) ethnic group. Student's t test results are included in these two panels. Due to the existence of outliers, robust statistical methods were implemented to confirm the t test results. The barplots at the bottom illustrate smoker and non-smoker percentages for each genotype group in AAs or EAs. Fisher's exact test was performed to obtain the *P* values for this panel. Corresponding sample size for each ethnic group is included on the top of the figure. Significant *P* values ($P < 0.05$) are marked in red.

Supplementary Table 1: Information on the 51 expression and 107 methylation probes for the 57 susceptibility genes.

Chr	Gene	Gene strand	Gene position	TSS coordinate	Expression			Methylation			
					Illumina ID	Oligo ID	Probe position	Probe ID	CpG position	SNP in CpG?	CpG island location
1	*CHRNB2*	+	154540256-154552353	154540257	HEEBO-065-HCC65I17	hHC024785	154548409-154548478	cg21052164	154539977	N	154539927-154540715
								cg00818872	154540271	N	
2	*NRXN1*	-	50145643-51259674	51255603	HEEBO-029-HCC29N11	hHC011075	50147539-50147608	cg16279786	51255306	N	51254462-51255631
								cg10917619	51255627	N	
4	*CHRNA9*	+	40337468-40356973	40337469	HEEBO-021-HCC21M22	hHC007990	40356890-40356959	cg10375110	40337443	N	
								cg23621817	40337853	N	
	GABRA2	-	46251581-46391396	46391945	HEEBO-029-HCC29D9	hHC010833	46252316-46252385	cg21820677	46392528	N	46391171-46392738
								cg15918284	46392657	N	
	GABRA4	-	46920916-46995580	46995580	HEEBO-008-HCC8H13	hHC002869	46921585-46921654	cg03593419	46995474	N	46994956-46995967
								cg16358826	46996264	N	46996078-46996300
5	*DRD1*	-	174867674-174871163	174871163	HEEBO-029-HCC29B22	hHC010798	174868285-174868354	cg16112129	174870968	N	174870500-174872470
								cg17307280	174871636	N	
6	*OPRM1*	+	154360442-154440594	154360443	HEEBO-045-HCC45I7	hHC017095	154412509-154412578	cg14262937	154360351	N	
								cg22719623	154360732	N	154360422-154361116
	MAP3K4	+	161412821-161538417	161412822	HEEBO-106-HCA106B14	hHA040358	161519390-161519459	cg22685251	161413652	Y	161412230-161413698
7	*PDE1C*	-	31792692-32338941	32110991	HEEBO-037-HCC37J17	hHC014057	31792718-31792787	cg22131691	32110988	N	32109753-32111164
								cg00546491	32111062	N	
	DDC	-	50526133-50633154	50628751	HEEBO-015-HCC15F9	hHC005505	50526277-50526346	cg04144768	50628154	N	
								cg03843951	50629634	N	
	CACNA2D1	-	81579417-82073031	82073031	HEEBO-026-HCC26O7	hHC009943	81579471-81579540	cg04008901	82072779	N	82071945-82073635
								cg06379754	82073573	N	
	CHRM2	+	136553415-136704999	136553399	HEEBO-014-HCC14F4	hHC005116	136704457-136704526	cg04748704	136553243	N	
								cg00973677	136553595	N	136553295-136556346
	TAS2R38	-	141672431-141673573	141673489		NA		cg25481253	141673384	Y	
								cg03017475	141674341	N	
	HTR5A	+	154862545-154877459	154862610	HEEBO-023-HCC23I11	hHC008651	154876848-154876917	cg15835825	154862030	N	
								cg25780543	154862770	N	
8	*DLC1*	-	13072081-13372395	13372395	HEEBO-092-HCA92G22	hHA035110	13072198-13072267	cg05226008	13372132	N	

Chr	Gene	Gene strand	Gene position	TSS coordinate	Expression			Methylation			
					Illumina ID	Oligo ID	Probe position	Probe ID	CpG position	SNP in CpG?	CpG island location
								cg00933411	13373090	N	
	CSGALNACT1	-	19261671-19460056	19460056	HEEBO-006-HCC6J20	hHC002156	19262195-19262264		NA		
	INTS10	+	19674917-19709586	19674918	HEEBO-025-HCC25M12	hHC009516	19709413-19709482		NA		
	CHRNA2	-	27317278-27336813	27336758		NA		cg02953306	27336626	Y	
								cg04953015	27338236	Y	
	CHRNB3	+	42552561-42592209	42552562	HEEBO-027-HCC27C4	hHC010036	42591751-42591820	cg00367281	42552470	Y	
								cg06840801	42552719	N	
	CHRNA6	-	42607779-42623619	42623619	HEEBO-047-HCC47C4	hHC017716	42608261-42608330	cg07906724	42623946	N	
9	*NTRK2*	+	87283465-87638505	87283466	HEEBO-013-HCC13A17	hHC004625	87636860-87636929	cg22402007	87282823	N	87282549-87285901
								cg09539438	87283789	N	
	SHC3	-	91620686-91793682	91793682		NA		cg13351583	91793648	N	91792490-91793763
								cg00420568	91794904	N	91794806-91795518
	GABBR2	-	101050365-101471175	101471479	HEEBO-023-HCC23B14	hHC008486	101050909-101050978	cg02058918	101470884	N	101470524-101472117
								cg07903918	101471986	N	
	GRIN3A	-	104331634-104500862	104500862	HEEBO-017-HCC17D10	hHC006226	104331813-104331882	cg08997253	104500729	N	104499137-104501229
								cg18794577	104501030	N	
	DNM1	+	130965662-131017527	130965663	HEEBO-058-HCC58E9	hHC021993	131008704-131008773	cg02494117	130965474	N	130965354-130966691
								cg13309018	130965761	N	
	DBH	+	136501484-136524466	136501485	HEEBO-064-HCC64K10	hHC024442	136523852-136523921	cg25020204	136500234	N	
								cg07824742	136501784	Y	136501723-136501928
10	*CHAT*	+	50822082-50873150	50817141	HEEBO-099-HCA99O4	hHA037972	50822104-50822173	cg12052765	50816963	N	50816963-50820806
								cg18592174	50817306	N	
	NRG3	+	83635069-84746935	83635070	HEEBO-025-HCC25D4	hHC009292	84745521-84745590		NA		
	PTEN	+	89623194-89728532	89623195	HEEBO-028-HCC28K6	hHR010614	89726288-89726357	cg21480743	89621419	N	89621218-89624183
								cg01228636	89621773	N	
								cg04738091	89622084	N	
								cg21573601	89622589	N	
								cg20849549	89623138	N	
								cg16687447	89623336	N	
								cg17489897	89623432	N	
								cg08859916	89624102	N	

Chr	Gene	Gene strand	Gene position	TSS coordinate	Expression			Methylation			
					Illumina ID	Oligo ID	Probe position	Probe ID	CpG position	SNP in CpG?	CpG island location
	LOC100188947	-	93066719-93371217	93371217		NA			NA		
11	*DRD4*	+	637304-640703	637305	HEEBO-064-HCC64C11	hHC024251	640629-640698	cg06825142	637170	N	636306-637963
	APBB1	-	6416354-6440300	6440644	HEEBO-060-HCC60F2	hHC022778	6416538-6416607	cg19327844	6440482	N	6439781-6440887
								cg05079045	6440803	N	
	BDNF	-	27676441-27681196	27743296	HEEBO-047-HCC47K19	hHC017923	27676838-27676907	cg27351358	27743258	N	27743206-27744913
								cg16257091	27743580	N	
	CHRM1	-	62676150-62689012	62689012	HEEBO-051-HCC51K5	hHC019445	62676441-62676510	cg00987015	62688751	N	
								cg13530039	62689557	N	
	NRXN2	-	64373645-64490660	64490660	HEEBO-042-HCC42K22	hHC016006	64373670-64373739	cg16718678	64490633	N	64490169-64491216
	ARRB1	-	74976481-75062873	75062875	HEEBO-053-HCC53J15	hHC020199	74979933-74980002		NA		
	TTC12	+	113185328-113244016	113185329	HEEBO-051-HCC51O24	hHC019560	113243931-113244000	cg12177743	113185079	N	113184925-113186066
								cg24264506	113185537	N	
	ANKK1	+	113258512-113271140	113258513	HEEBO-056-HCC56P9	hHC021489	113270929-113270998		NA		
	DRD2	-	113280316-113346001	113345881	HEEBO-051-HCC51C1	hHC019249	113280823-113280892	cg12758687	113346327	N	113344914-113346439
								cg21330703	113346388	N	
	HTR3A	+	113845909-113861034	113845910	HEEBO-106-HCA106I12	hHA040524	113857476-113857545	cg24134767	113845638	N	
12	*GRIN2B*	-	13714409-14133022	14133052	HEEBO-060-HCC60E20	hHC022772	13714915-13714984	cg04016326	14132940	N	
								cg13264741	14133426	N	14133083-14135400
14	*C14orf28*	+	45366506-45376460	45366507	HEEBO-013-HCC13H6	hHR004782	45376220-45376289		NA		
	NRXN3	+	78870092-80330760	78870093	HEEBO-027-HCC27P15	hHC010359	80328681-80328750	cg16372520	78869751	N	
								cg15572745	78870232	N	
16	*CDH13*	+	82660577-83830199	82660578	HEEBO-018-HCC18C20	hHC006596	83829828-83829897	cg01880569	82660328	N	82660309-82661822
								cg08977371	82660490	N	
								cg08747377	82660670	N	
								cg00806490	82660873	N	
								cg13759328	82661521	N	
								cg19369556	82661725	N	
								cg16777782	82671333	N	82670941-82671531
								cg19854301	82671450	N	
								cg02168291	82671520	N	
17	*ARRB2*	+	4613788-4624795	4613789	HEEBO-038-HCC38E24	hHC014328	4624637-4624706	cg03950654	4613328	N	

Chr	Gene	Gene strand	Gene position	TSS coordinate	Expression			Methylation			
					Illumina ID	Oligo ID	Probe position	Probe ID	CpG position	SNP in CpG?	CpG island location
								cg23779331	4614312	N	4612478-4614747
	DLG4	-	7093211-7123369	7123030	HEEBO-069-HCC69C6	hHC026166	7093896-7093965	cg12228229	7122261	N	
								cg02740128	7123860	N	7122940-7123898
	GABARAP	-	7143737-7145753	7145753	HEEBO-071-HCC71M23	hHR027191	7145576-7145645	cg25737491	7145532	N	7145374-7146069
								cg23983449	7146757	N	7146209-7146827
	CHRNB1	+	7348406-7360932	7348406		NA		cg04809787	7348339	N	7348187-7349319
								cg18884137	7348490	Y	
	SLC6A4	-	28521337-28562986	28562705		NA		cg22584138	28562220	N	28562022-28563220
								cg05016953	28562813	N	
	PPP1R1B	+	37783178-37792877	37783179	HEEBO-106-HCA106C7	hHA040375	37783656-37783725	cg00112517	37783011	N	
								cg08411435	37784024	N	37783122-37784064
19	*RAB4B*	+	41284170-41302847	41284177	HEEBO-064-HCC64A23	hHC024215	41292785-41292854	cg24958765	41283667	N	41283544-41284469
								cg13332130	41284234	N	
	EGLN2	+	41305047-41314337	41305145	HEEBO-060-HCC60O16	hHC023008	41314080-41314149	cg22499964	41304369	N	41304039-41304401
								cg22671726	41305423	N	41304543-41305934
	CYP2A6	-	41349442-41356352	41356340	HEEBO-071-HCC71I21	hHR027093	41349588-41349657	cg05910970	41355973	N	
								cg02043477	41357152	N	
	CYP2A7	-	41381343-41388657	41388657	HEEBO-066-HCC66P15	hHR025335	41381465-41381534	cg20075229	41388937	N	
								cg25427638	41389357	Y	
	CYP2B6	+	41497203-41524301	41497204	HEEBO-087-HCA87O20	hHA033380	41518580-41518649	cg10322876	41496749	N	
								cg19756068	41497222	N	
20	*CHRNA4*	-	61974664-61992695	61992748	HEEBO-048-HCC48I6	hHC018246	61976458-61976527	cg00318573	61993118	N	61992084-61993608
								cg08912400	61993427	N	
22	*COMT*	+	19929262-19957496	19929309	HEEBO-086-HCA86H14	hHA032822	19929315-19929384	cg15926585	19928445	Y	

Chr chromosome; *TSS* transcriptional start site; *NA* not available. Whenever available, information in this table are based on the BrainCloud and BrainCloudMethyl applications distributed through the BrainCloud project website (http://braincloud.jhmi.edu/), with genomic positions lifted from NCBI Build 36/hg18 to Build 37/hg19. Otherwise, gene positions and TSS coordinates were obtained from the UCSC Genome Browser (https://genome.ucsc.edu/) and the Database of Transcriptional Start Sites (http://dbtss.hgc.jp/), respectively. All the genomic positions in this table are based on the NCBI Build 37/hg19 assembly.

Supplementary Table 2: Information on the imputation intervals for each of the 57 genes.

Chr	Imputation interval	Gene(s) within the interval	Number of variations	
			Before imputation	After imputation (info≥0.3)
1	153539977-155552353	*CHRNB2*	358	15,235
2	49145643-52259674	*NRXN1*	1,162	36,333
4	39337443-41356973	*CHRNA9*	557	20,469
	45251581-47996264	*GABRA2*, *GABRA4*	518	26,244
5	173867674-175871636	*DRD1*	694	17,969
6	153360351-155440594	*OPRM1*	746	22,384
	160412821-162538417	*MAP3K4*	845	23,280
7	30792692-33338941	*PDE1C*	964	26,573
	49526133-51633154	*DDC*	607	21,068
	80579417-83073573	*CACNA2D1*	804	24,745
	135553243-137704999	*CHRM2*	660	19,723
	140672431-142674341	*TAS2R38*	491	17,695
	153862030-155877459	*HTR5A*	819	20,892
8	12072081-14373090	*DLC1*	930	32,877
	18261671-20709586	*CSGALNACT1*, *INTS10*	1,246	32,649
	26317278-28338236	*CHRNA2*	851	21,668
	41552470-43623946	*CHRNB3*, *CHRNA6*	301	19,679
9	86282823-88638505	*NTRK2*	683	21,899
	90620686-92794904	*SHC3*	575	21,137
	100050365-102471986	*GABBR2*	656	20,634
	103331634-105501030	*GRIN3A*	750	24,036
	129965474-132017527	*DNM1*	465	16,868
	135500234-137524466	*DBH*	769	20,960
10	49816963-51873150	*CHAT*	585	15,838
	82635069-85746935	*NRG3*	981	31,941
	88621419-90728532	*PTEN*	599	17,687
11	0-1640703	*DRD4*	317	16,202
	5416354-7440803	*APBB1*	1,023	22,709
	26676441-28743580	*BDNF*	509	16,158
	61676150-65490660	*CHRM1*, *NRXN2*	677	32,854
	73976481-76062873	*ARRB1*	524	19,121
	112258512-114861034	*TTC12*, *ANKK1*, *DRD2*, *HTR3A*	886	24,711
12	12714409-15133426	*GRIN2B*	889	23,136
14	44366506-46376460	*C14orf28*	441	19,237
	77869751-81330760	*NRXN3*	1,141	33,808
16	81660328-84830199	*CDH13*	2,033	49,608
17	3613328-5624795	*ARRB2*	660	20,915
	6093211-8360932	*DLG4*, *GABARAP*, *CHRNB1*	685	21,292
	27521337-29562986	*SLC6A4*	239	16,612
	36783011-38792877	*PPP1R1B*	369	16,949
19	40283667-42524301	*RAB4B*, *EGLN2*, *CYP2A6*, *CYP2A7*, *CYP2B6*	429	21,151
20	60974664-62993427	*CHRNA4*	444	20,255
22	18928445-20957496	*COMT*	469	16,689

Chr chromosome; *info* the metric used by IMPUTE2 to measure imputation quality. All the imputation intervals in this table are based on the NCBI Build 37/hg19 assembly.

Chapter 6

Future directions

Chapters 2 to **5** showed that: (*i*) for candidate genes, genes encoding the 5-$HT3_{AB}$ receptors (*HTR3A* and *HTR3B*) and the serotonin transporter (*SLC6A4*) specifically, gene-by-gene interactions (epistasis) contribute significantly to multiple addictions including smoking; (*ii*) in a subset of ND candidate genes, nonsynonymous rare variants independently or combined with common variants significantly affect smoking; (*iii*) genetic results from different experimental approaches are not always consistent, especially for the two unbiased genome-wide tests (genome-wide linkage studies and GWASs), and developing a genetic susceptibility map and keeping it updated is an effective way to keep track of what we know about the genetic architecture of a disease/trait and what the next steps might be with new experimental approaches available; and (*iv*) emerging molecular phenotypes, mRNA expression and DNA methylation levels in this case, enable us to bridge missing mechanistic steps between genetic variation and smoking traits, and we for the first time detect variants within one *cis*-mQTL that affect methylation, expression, and phenotypic manifestation simultaneously using the same participants.

These results not only corroborate complexity of the genetic structure underlying smoking, more importantly they advance our understanding about ND genetics by distilling comprehensive information (epistasis, rare variants, eQTLs, and mQTLs) from existing candidate genes innovatively, and making up a thorough genetic susceptibility map, which summarizes all the current findings and implicates paths to further discover new targets.

To extend these findings, I would like to continue my research in the following three directions (**Figure 6.1**): first, the epistasis and rare variant studies in this dissertation are limited to candidate genes, due to technological and cost reasons, respectively (**Chapters 2** and **3**).[37, 44] However, with the ever-growing computational power and plummeting sequencing cost, it is essential to quantify contribution of epistasis and rare variants to smoking heritability on the genome-wide level. To what extent epistasis inflates the denominator of heritability explained, how much rare variant effects add to the numerator, and whether these two factors can solve the "missing heritability" issue for ND remain to be determined.

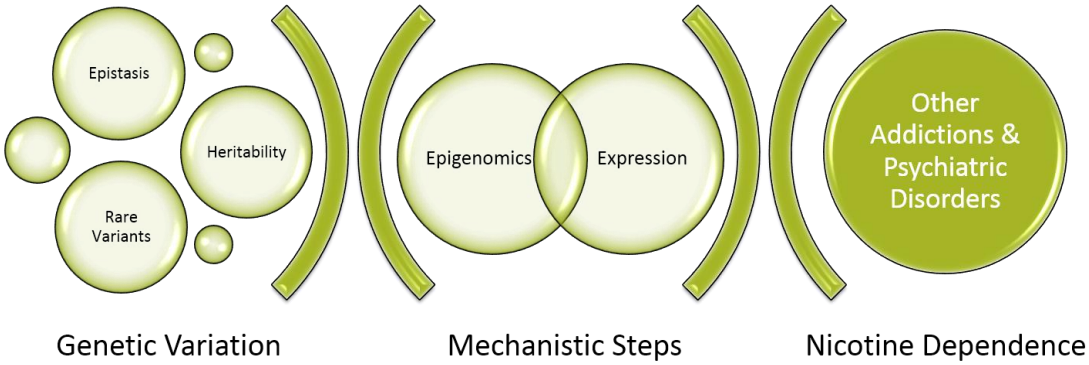
Second, we have only used mRNA expression and DNA methylation data for ND candidate genes in the prefrontal cortex region to explore mechanistic steps between genetic variation and smoking status, as a result of limited data and sample size issues (**Chapter 5**). In the future, with a large quantity of molecular phenotypes becoming available, such as histone modifications, DNase I hypersensitivity sites, and chromatin conformation change, in multiple brain regions and for specific cell types, we would have the capability to scrutinize characteristics of each regulation layer, relationships among different layers, and mechanisms through which they contribute to ND cooperatively. Beyond the "what" question (what are the genetic factors underlying smoking?), this will help us solve the "how" question (how do they

contribute to the risk of smoking?). Only with this level of understanding may we more confidently translate knowledge (what we know about ND genetics) to practice (efficient therapeutic treatments).

Third, smoking not only shares a significant portion of its genetic underpinning with other addictions, as seen in the epistasis and susceptibility map studies,[37, 45] but also has intricate connections with other psychiatric disorders, e.g., schizophrenia and autism spectrum disorder. An analysis of data from the National Comorbidity Study (NCS), a nationally representative survey of psychiatric disorders in the United States, found that 41% of people with a psychiatric disorder smoke, about twice the rate (22.5%) seen in those without psychiatric diagnoses. People with psychiatric disorders consume 44.3% of all cigarettes smoked in the United States.[402] To extend our research scope to this special population, and identify genetic factors leading to the comorbidity, are urgently needed to improve treatment for people with psychiatric disorders, because the high rate of smoking is an important factor in increased rates of physical illness and mortality in this group. And research indicates that treating multiple illnesses simultaneously in an integrated fashion is generally the best treatment approach for these patients.

All the above efforts are guiding our way to new therapeutic targets for smoking treatment and prevention that takes into account individual variability in genes, which echoes the Precision Medicine Initiative® (PMI) announced by President Obama on January 20, 2015. This new era of individualized care presents an opportunity for us to control smoking damage more effectively.

Figure 6.1: Three future research directions.



References

1. Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H *et al.* A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012; **380**(9859)**:** 2224-2260.

2. USDHHS. *The Health Consequences of Smoking—50 Years of Progress. A Report of the Surgeon General.* US Department of Health & Human Services, Center for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promortion: Atlanta, Georgia, 2014.

3. Prevention CfDCa. Current Cigarette Smoking Among Adults-United States, 2005-2014. *Morbidity and Mortality Weekly Report* 2015; **64**(44)**:** 1233-1240.

4. 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**(1)**:** 23-31.

5. Gunby P. Surgeon General emphasizes nicotine addiction in annual report on tobacco use, consequences. *Jama* 1988; **259**(19)**:** 2811.

6. Li MD, Payne TJ, Ma JZ, Lou XY, Zhang D, Dupont RT *et al.* A genomewide search finds major susceptibility loci for nicotine dependence on chromosome 10 in African Americans. *Am J Hum Genet* 2006; **79**(4)**:** 745-751.

7. Li MD, Ma JZ, Payne TJ, Lou XY, Zhang D, Dupont RT *et al.* Genome-wide linkage scan for nicotine dependence in European Americans and its converging results with African Americans in the Mid-South Tobacco Family sample. *Mol Psychiatry* 2008; **13**(4)**:** 407-416.

8. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11**(3)**:** 241-247.

9. Li MD. Identifying susceptibility loci for nicotine dependence: 2008 update based on recent genome-wide linkage analyses. *Hum Genet* 2008; **123**(2)**:** 119-131.

10. Beuten J, Ma JZ, Payne TJ, Dupont RT, Crews KM, Somes G *et al.* Single- and multilocus allelic variants within the GABA(B) receptor subunit 2 (GABAB2) gene are significantly associated with nicotine dependence. *Am J Hum Genet* 2005; **76**(5)**:** 859-864.

11. Beuten J, Ma JZ, Payne TJ, Dupont RT, Lou XY, Crews KM *et al.* Association of specific haplotypes of neurotrophic tyrosine kinase receptor 2 gene (NTRK2) with vulnerability to nicotine dependence in African-Americans and European-Americans. *Biol Psychiatry* 2007; **61**(1)**:** 48-55.

12. Sun D, Ma JZ, Payne TJ, Li MD. Beta-arrestins 1 and 2 are associated with nicotine dependence in European American smokers. *Mol Psychiatry* 2008; **13**(4)**:** 398-406.

13. Wei J, Ma JZ, Payne TJ, Cui W, Ray R, Mitra N *et al.* Replication and extension of association of choline acetyltransferase with nicotine dependence in European and African American smokers. *Hum Genet* 2010; **127**(6)**:** 691-698.

14. Li MD, Beuten J, Ma JZ, Payne TJ, Lou XY, Garcia V *et al.* Ethnic- and gender-specific association of the nicotinic acetylcholine receptor alpha4 subunit gene (CHRNA4) with nicotine dependence. *Hum Mol Genet* 2005; **14**(9)**:** 1211-1219.

15. Beuten J, Payne TJ, Ma JZ, Li MD. Significant association of catechol-O-methyltransferase (COMT) haplotypes with nicotine dependence in male and female smokers of two ethnic populations. *Neuropsychopharmacology* 2006; **31**(3)**:** 675-684.

16. Huang W, Ma JZ, Payne TJ, Beuten J, Dupont RT, Li MD. Significant association of DRD1 with nicotine dependence. *Hum Genet* 2008; **123**(2)**:** 133-140.

17. Huang W, Payne TJ, Ma JZ, Li MD. A functional polymorphism, rs6280, in DRD3 is significantly associated with nicotine dependence in European-American smokers. *Am J Med Genet B Neuropsychiatr Genet* 2008; **147B**(7)**:** 1109-1115.

18. Mangold JE, Payne TJ, Ma JZ, Chen G, Li MD. Bitter taste receptor gene polymorphisms are an important factor in the development of nicotine dependence in African Americans. *Journal of medical genetics* 2008; **45**(9)**:** 578-582.

19. Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA *et al.* Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet* 2007; **16**(1)**:** 36-49.

20. Berrettini W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H *et al.* Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol Psychiatry* 2008; **13**(4)**:** 368-373.

21. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008; **452**(7187)**:** 638-642.

22. Weiss RB, Baker TB, Cannon DS, von Niederhausern A, Dunn DM, Matsunami N *et al.* A candidate gene approach identifies the CHRNA5-A3-B4 region as a risk factor for age-dependent nicotine addiction. *PLoS Genet* 2008; **4**(7)**:** e1000125.

23. TAG. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; **42**(5)**:** 441-447.

24. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F *et al.* Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* 2010; **42**(5)**:** 448-453.

25. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010; **42**(5)**:** 436-440.

26. Saccone NL, Wang JC, Breslau N, Johnson EO, Hatsukami D, Saccone SF *et al.* The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. *Cancer research* 2009; **69**(17)**:** 6848-6856.

27. David SP, Hamidovic A, Chen GK, Bergen AW, Wessel J, Kasberger JL *et al.* Genome-wide meta-analyses of smoking behaviors in African Americans. *Translational psychiatry* 2012; **2:** e119.

28. 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* 2010; **153B**(3)**:** 745-756.

29. Li MD, Yoon D, Lee JY, Han BG, Niu T, Payne TJ *et al.* Associations of variants in CHRNA5/A3/B4 gene cluster with smoking behaviors in a Korean population. *PLoS One* 2010; **5**(8)**:** e12183.

30. Chen LS, Saccone NL, Culverhouse RC, Bracci PM, Chen CH, Dueker N *et al.* Smoking and genetic risk variation across populations of European, Asian, and African American ancestry--a meta-analysis of chromosome 15q25. *Genet Epidemiol* 2012; **36**(4)**:** 340-351.

31. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008; **40**(5)**:** 616-622.

32. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008; **452**(7187)**:** 633-637.

33. Cui WY, Wang S, Yang J, Yi SG, Yoon D, Kim YJ *et al.* Significant association of CHRNB3 variants with nicotine dependence in multiple ethnic populations. *Mol Psychiatry* 2013; **18**(11)**:** 1149-1151.

34. Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF *et al.* Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet* 2007; **16**(1)**:** 24-35.

35. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010; **11**(6)**:** 446-450.

36. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* 2012; **109**(4)**:** 1193-1198.

37. Yang J, Li MD. Association and interaction analyses of 5-HT3 receptor and serotonin transporter genes with alcohol, cocaine, and nicotine dependence using the SAGE data. *Hum Genet* 2014; **133**(7)**:** 905-918.

38. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* 2014; **111**(4)**:** E455-464.

39. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014; **15**(5)**:** 335-346.

40. Wessel J, McDonald SM, Hinds DA, Stokowski RP, Javitz HS, Kennemer M *et al.* Resequencing of nicotinic acetylcholine receptor genes and association of common and rare variants with the Fagerstrom test for nicotine dependence. *Neuropsychopharmacology* 2010; **35**(12)**:** 2392-2402.

41. Xie P, Kranzler HR, Krauthammer M, Cosgrove KP, Oslin D, Anton RF *et al.* Rare nonsynonymous variants in alpha-4 nicotinic acetylcholine receptor gene protect against nicotine dependence. *Biol Psychiatry* 2011; **70**(6)**:** 528-536.

42. Haller G, Druley T, Vallania FL, Mitra RD, Li P, Akk G *et al.* Rare missense variants in CHRNB4 are associated with reduced risk of nicotine dependence. *Hum Mol Genet* 2012; **21**(3)**:** 647-655.

43. Doyle GA, Chou AD, Saung WT, Lai AT, Lohoff FW, Berrettini WH. Identification of CHRNA5 rare variants in African-American heavy smokers. *Psychiatr Genet* 2014; **24**(3)**:** 102-109.

44. Yang J, Wang S, Yang Z, Hodgkinson CA, Iarikova P, Ma JZ *et al.* The contribution of rare and common variants in 30 genes to risk nicotine dependence. *Mol Psychiatry* 2015; **20**(11)**:** 1467-1478.

45. Yang J, Li MD. Converging findings from linkage and association analyses on susceptibility genes for smoking and other addictions. *Mol Psychiatr* 2016.

46. Pai AA, Pritchard JK, Gilad Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet* 2015; **11**(1)**:** e1004857.

47. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 2013; **342**(6159)**:** 744-747.

48. Cravchik A, Goldman D. Neurochemical individuality: genetic diversity among human dopamine and serotonin receptors and transporters. *Arch Gen Psychiatry* 2000; **57**(12)**:** 1105-1114.

49. Barnes NM, Hales TG, Lummis SC, Peters JA. The 5-HT3 receptor--the relationship between structure and function. *Neuropharmacology* 2009; **56**(1)**:** 273-284.

50. Boess FG, Martin IL. Molecular biology of 5-HT receptors. *Neuropharmacology* 1994; **33**(3-4)**:** 275-317.

51. Maricq AV, Peterson AS, Brake AJ, Myers RM, Julius D. Primary structure and functional expression of the 5HT3 receptor, a serotonin-gated ion channel. *Science* 1991; **254**(5030)**:** 432-437.

52. Grant KA. The role of 5-HT3 receptors in drug dependence. *Drug Alcohol Depend* 1995; **38**(2)**:** 155-171.

53. Nayak SV, Ronde P, Spier AD, Lummis SC, Nichols RA. Nicotinic receptors co-localize with 5-HT(3) serotonin receptors on striatal nerve terminals. *Neuropharmacology* 2000; **39**(13)**:** 2681-2690.

54. Dougherty JJ, Nichols RA. Cross-regulation between colocalized nicotinic acetylcholine and 5-HT3 serotonin receptors on presynaptic nerve terminals. *Acta Pharmacol Sin* 2009; **30**(6)**:** 788-794.

55. Yamauchi JG, Nemecz A, Nguyen QT, Muller A, Schroeder LF, Talley TT *et al.* Characterizing ligand-gated ion channel receptors with genetically encoded Ca2++ sensors. *PLoS One* 2011; **6**(1)**:** e16519.

56. Narahashi T, Kuriyama K, Illes P, Wirkner K, Fischer W, Muhlberg K *et al.* Neuroreceptors and ion channels as targets of alcohol. *Alcohol Clin Exp Res* 2001; **25**(5 Suppl ISBRA)**:** 182S-188S.

57. Sung KW, Engel SR, Allan AM, Lovinger DM. 5-HT(3) receptor function and potentiation by alcohols in frontal cortex neurons from transgenic mice overexpressing the receptor. *Neuropharmacology* 2000; **39**(12)**:** 2346-2351.

58. Davies PA, Pistis M, Hanna MC, Peters JA, Lambert JJ, Hales TG *et al.* The 5-HT3B subunit is a major determinant of serotonin-receptor function. *Nature* 1999; **397**(6717)**:** 359-363.

59. Dubin AE, Huvar R, D'Andrea MR, Pyati J, Zhu JY, Joy KC *et al.* The pharmacological and functional characteristics of the serotonin 5-HT(3A) receptor are specifically modified by a 5-HT(3B) receptor subunit. *J Biol Chem* 1999; **274**(43)**:** 30799-30810.

60. Enoch MA, Gorodetsky E, Hodgkinson C, Roy A, Goldman D. Functional genetic variants that increase synaptic serotonin and 5-HT3 receptor sensitivity predict alcohol and drug dependence. *Mol Psychiatry* 2011; **16**(11)**:** 1139-1146.

61. Miyake A, Mochizuki S, Takemoto Y, Akuzawa S. Molecular cloning of human 5-hydroxytryptamine3 receptor: heterogeneity in distribution and function among species. *Molecular pharmacology* 1995; **48**(3)**:** 407-416.

62. Ramamoorthy S, Bauman AL, Moore KR, Han H, Yang-Feng T, Chang AS *et al.* Antidepressant- and cocaine-sensitive human serotonin transporter: molecular cloning, expression, and chromosomal localization. *Proc Natl Acad Sci U S A* 1993; **90**(6)**:** 2542-2546.

63. Lesch KP, Balling U, Gross J, Strauss K, Wolozin BL, Murphy DL *et al.* Organization of the human serotonin transporter gene. *Journal of neural transmission General section* 1994; **95**(2)**:** 157-162.

64. Bradley CC, Blakely RD. Alternative splicing of the human serotonin transporter gene. *J Neurochem* 1997; **69**(4)**:** 1356-1367.

65. Ozsarac N, Santha E, Hoffman BJ. Alternative non-coding exons support serotonin transporter mRNA expression in the brain and gut. *J Neurochem* 2002; **82**(2)**:** 336-344.

66. Dong C, Wong ML, Licinio J. Sequence variations of ABCB1, SLC6A2, SLC6A3, SLC6A4, CREB1, CRHR1 and NTRK2: association with major depression and antidepressant response in Mexican-Americans. *Mol Psychiatry* 2009; **14**(12)**:** 1105-1118.

67. Lopez-Leon S, Janssens AC, Gonzalez-Zuloeta Ladd AM, Del-Favero J, Claes SJ, Oostra BA *et al.* Meta-analyses of genetic studies on major depressive disorder. *Mol Psychiatry* 2008; **13**(8)**:** 772-785.

68. McCauley JL, Olson LM, Dowd M, Amin T, Steele A, Blakely RD *et al.* Linkage and association analysis at the serotonin transporter (SLC6A4) locus in a rigid-compulsive subset of autism. *Am J Med Genet B Neuropsychiatr Genet* 2004; **127B**(1)**:** 104-112.

69. Seneviratne C, Franklin J, Beckett K, Ma JZ, Ait-Daoud N, Payne TJ *et al.* Association, Interaction, and Replication Analysis of Serotonin Receptor Subtypes A and B and Transporter Genes in Alcohol Dependence. *Submitted* 2013.

70. Yang Z, Seneviratne C, Wang S, Ma JZ, Payne TJ, Wang J *et al.* Serotonin transporter and receptor genes significantly impact nicotine dependence through genetic interactions in both European American and African American smokers. *Drug Alcohol Depend* 2013; **129**(3)**:** 217-225.

71. Johnson BA, Ait-Daoud N, Seneviratne C, Roache JD, Javors MA, Wang XQ *et al.* Pharmacogenetic approach at the serotonin transporter gene as a method of reducing the severity of alcohol drinking. *The American journal of psychiatry* 2011; **168**(3)**:** 265-275.

72. Johnson BA, Seneviratne C, Wang XQ, Ait-Daoud N, Li MD. Determination of genotype combinations that can predict the outcome of the treatment of alcohol dependence using the 5-HT(3) antagonist ondansetron. *The American journal of psychiatry* 2013; **170**(9)**:** 1020-1031.

73. Edenberg HJ, Bierut LJ, Boyce P, Cao M, Cawley S, Chiles R *et al.* Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. *BMC genetics* 2005; **6 Suppl 1:** S2.

74. Grucza RA, Wang JC, Stitzel JA, Hinrichs AL, Saccone SF, Saccone NL *et al.* A risk allele for nicotine dependence in CHRNA5 is a protective allele for cocaine dependence. *Biol Psychiatry* 2008; **64**(11)**:** 922-929.

75. Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E *et al.* A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci U S A* 2010; **107**(11)**:** 5082-5087.

76. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**(8)**:** 816-834.

77. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet* 2009; **10:** 387-406.

78. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**(3)**:** 559-575.

79. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**(2)**:** 263-265.

80. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B *et al.* The structure of haplotype blocks in the human genome. *Science* 2002; **296**(5576)**:** 2225-2229.

81. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002; **70**(2)**:** 425-434.

82. Zhu Z, Tong X, Liang M, Cui W, Su K, Li MD *et al.* Development of GMDR-GPU for gene-gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes. *PLoS One* 2013; **8**(4)**:** e61943.

83. Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC *et al.* 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**(6)**:** 1125-1137.

84. Vijayan NN, Iwayama Y, Koshy LV, Natarajan C, Nair C, Allencherry PM *et al.* Evidence of association of serotonin transporter gene polymorphisms with schizophrenia in a South Indian population. *J Hum Genet* 2009; **54**(9)**:** 538-542.

85. Ma DQ, Rabionet R, Konidari I, Jaworski J, Cukier HN, Wright HH *et al.* Association and gene-gene interaction of SLC6A4 and ITGB3 in autism. *Am J Med Genet B Neuropsychiatr Genet* 2010; **153B**(2)**:** 477-483.

86. Su S, Zhao J, Bremner JD, Miller AH, Tang W, Bouzyk M *et al.* Serotonin transporter gene, depressive symptoms, and interleukin-6. *Circulation Cardiovascular genetics* 2009; **2**(6)**:** 614-620.

87. Krzywkowski K, Davies PA, Feinberg-Zadek PL, Brauner-Osborne H, Jensen AA. High-frequency HTR3B variant associated with major depression dramatically augments the signaling of the human 5-HT3AB receptor. *Proc Natl Acad Sci U S A* 2008; **105**(2)**:** 722-727.

88. Niesler B, Flohr T, Nothen MM, Fischer C, Rietschel M, Franzek E *et al.* Association between the 5' UTR variant C178T of the serotonin receptor gene HTR3A and bipolar affective disorder. *Pharmacogenetics* 2001; **11**(6)**:** 471-475.

89. Levran O, Londono D, O'Hara K, Randesi M, Rotrosen J, Casadonte P *et al.* Heroin addiction in African Americans: a hypothesis-driven association study. *Genes Brain Behav* 2009; **8**(5)**:** 531-540.

90. Ducci F, Enoch MA, Yuan Q, Shen PH, White KV, Hodgkinson C *et al.* HTR3B is associated with alcoholism with antisocial behavior and alpha EEG power--an intermediate phenotype for alcoholism and co-morbid behaviors. *Alcohol* 2009; **43**(1)**:** 73-84.

91. Laugsand EA, Fladvad T, Skorpen F, Maltoni M, Kaasa S, Fayers P *et al.* Clinical and genetic factors associated with nausea and vomiting in cancer patients receiving opioids. *Eur J Cancer* 2011; **47**(11)**:** 1682-1691.

92. Torres GE, Gainetdinov RR, Caron MG. Plasma membrane monoamine transporters: structure, regulation and function. *Nature reviews Neuroscience* 2003; **4**(1)**:** 13-25.

93. Lovinger DM, Sung KW, Zhou Q. Ethanol and trichloroethanol alter gating of 5-HT3 receptor-channels in NCB-20 neuroblastoma cells. *Neuropharmacology* 2000; **39**(4)**:** 561-570.

94. Breitinger HG, Geetha N, Hess GP. Inhibition of the serotonin 5-HT3 receptor by nicotine, cocaine, and fluoxetine investigated by rapid chemical kinetic techniques. *Biochemistry* 2001; **40**(28)**:** 8419-8429.

95. Mossner R, Schmitt A, Hennig T, Benninghoff J, Gerlach M, Riederer P *et al.* Quantitation of 5HT3 receptors in forebrain of serotonin transporter deficient mice. *J Neural Transm (Vienna)* 2004; **111**(1)**:** 27-35.

96. Murphy DL, Lesch KP. Targeting the murine serotonin transporter: insights into human neurobiology. *Nature reviews Neuroscience* 2008; **9**(2)**:** 85-96.

97. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003; **361**(9357)**:** 598-604.

98. Feinn R, Nellissery M, Kranzler HR. Meta-analysis of the association of a functional serotonin transporter promoter polymorphism with alcohol dependence. *Am J Med Genet B Neuropsychiatr Genet* 2005; **133B**(1)**:** 79-84.

99. Gerra G, Garofano L, Santoro G, Bosari S, Pellegrini C, Zaimovic A *et al.* Association between low-activity serotonin transporter genotype and heroin dependence: behavioral and personality correlates. *Am J Med Genet B Neuropsychiatr Genet* 2004; **126B**(1)**:** 37-42.

100. Mannelli P, Patkar AA, Murray HW, Certa K, Peindl K, Mattila-Evenden M *et al.* Polymorphism in the serotonin transporter gene and response to treatment in African American cocaine and alcohol-abusing individuals. *Addict Biol* 2005; **10**(3)**:** 261-268.

101. Patkar AA, Berrettini WH, Hoehe M, Hill KP, Sterling RC, Gottheil E *et al.* Serotonin transporter (5-HTT) gene polymorphisms and susceptibility to cocaine dependence among African-American individuals. *Addict Biol* 2001; **6**(4)**:** 337-345.

102. Saiz PA, Garcia-Portilla MP, Florez G, Arango C, Corcoran P, Morales B *et al.* Differential role of serotonergic polymorphisms in alcohol and heroin dependence. *Prog Neuropsychopharmacol Biol Psychiatry* 2009; **33**(4)**:** 695-700.

103. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009; **10**(6)**:** 392-404.

104. Li MD, Lou XY, Chen GB, Ma JZ, Elston RC. Gene-Gene Interactions Among CHRNA4, CHRNB2, BDNF, and NTRK2 in Nicotine Dependence. *Biol Psychiat* 2008; **64**(11)**:** 951-957.

105. Li MD, Mangold JE, Seneviratne C, Chen GB, Ma JZ, Lou XY *et al.* Association and interaction analyses of GABBR1 and GABBR2 with nicotine dependence in European- and African-American populations. *PLoS One* 2009; **4**(9)**:** e7055.

106. Angeli CB, Kimura L, Auricchio MT, Vicente JP, Mattevi VS, Zembrzuski VM *et al.* Multilocus analyses of seven candidate genes suggest interacting pathways for obesity-related traits in Brazilian populations. *Obesity (Silver Spring)* 2011; **19**(6)**:** 1244-1251.

107. Neuman RJ, Wasson J, Atzmon G, Wainstein J, Yerushalmi Y, Cohen J *et al.* Gene-gene interactions lead to higher risk for development of type 2 diabetes in an Ashkenazi Jewish population. *PLoS One* 2010; **5**(3)**:** e9903.

108. Gelernter J, Yu Y, Weiss R, Brady K, Panhuysen C, Yang BZ *et al.* Haplotype spanning TTC12 and ANKK1, flanked by the DRD2 and NCAM1 loci, is strongly associated to nicotine dependence in two distinct American populations. *Hum Mol Genet* 2006; **15**(24)**:** 3498-3507.

109. Huang W, Payne TJ, Ma JZ, Beuten J, Dupont RT, Inohara N *et al.* Significant association of ANKK1 and detection of a functional polymorphism with nicotine dependence in an African-American sample. *Neuropsychopharmacology* 2009; **34**(2)**:** 319-330.

110. Tobacco-and-Genetics-Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; **42**(5)**:** 441-447.

111. Beuten J, Ma JZ, Payne TJ, Dupont RT, Quezada P, Huang W *et al.* Significant association of BDNF haplotypes in European-American male smokers but not in European-American female or African-American smokers. *Am J Med Genet B Neuropsychiatr Genet* 2005; **139B**(1)**:** 73-80.

112. Nussbaum J, Xu Q, Payne TJ, Ma JZ, Huang W, Gelernter J *et al.* Significant association of the neurexin-1 gene (NRXN1) with nicotine dependence in European- and African-American smokers. *Hum Mol Genet* 2008; **17**(11)**:** 1569-1577.

113. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 2011; **43**(11)**:** 1066-1073.

114. Asselbergs FW, Guo Y, van Iperen EP, Sivapalaratnam S, Tragante V, Lanktree MB *et al.* Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am J Hum Genet* 2012; **91**(5)**:** 823-838.

115. Diogo D, Kurreeman F, Stahl EA, Liao KP, Gupta N, Greenberg JD *et al.* Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. *Am J Hum Genet* 2013; **92**(1)**:** 15-27.

116. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; **5**(2)**:** e1000384.

117. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* 2013; **92**(6)**:** 841-853.

118. Shi G, Rao DC. Optimum designs for next-generation sequencing to discover rare variants for common complex disease. *Genet Epidemiol* 2011; **35**(6)**:** 572-579.

119. Ma JZ, Beuten J, Payne TJ, Dupont RT, Elston RC, Li MD. Haplotype analysis indicates an association between the DOPA decarboxylase (DDC) gene and nicotine dependence. *Hum Mol Genet* 2005; **14**(12)**:** 1691-1698.

120. Beuten J, Ma JZ, Payne TJ, Dupont RT, Crews KM, Somes G *et al.* Single- and Multilocus Allelic Variants within the GABAB Receptor Subunit 2 (GABAB2) Gene Are Significantly Associated with Nicotine Dependence. *Am J Hum Genet* 2005; **76**(5)**:** 859-864.

121. Wang S, Yang Z, Ma JZ, Payne TJ, Li MD. Introduction to deep sequencing and its application to drug addiction research with a focus on rare variants. *Molecular neurobiology* 2014; **49**(1)**:** 601-614.

122. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**(14)**:** 1754-1760.

123. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* 2009; **4**(7)**:** 1073-1081.

124. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P *et al.* A method and server for predicting damaging missense mutations. *Nature methods* 2010; **7**(4)**:** 248-249.

125. Ma JZ, Payne TJ, Nussbaum J, Li MD. Significant association of glutamate receptor, ionotropic N-methyl-D-aspartate 3A (GRIN3A), with nicotine dependence in European- and African-American smokers. *Hum Genet* 2010; **127**(5)**:** 503-512.

126. Xu Q, Huang W, Payne TJ, Ma JZ, Li MD. Detection of genetic association and a functional polymorphism of dynamin 1 gene with nicotine dependence in European and African Americans. *Neuropsychopharmacology* 2009; **34**(5)**:** 1351-1359.

127. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010; **20**(1)**:** 110-121.

128. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation research* 2007; **615**(1-2)**:** 28-56.

129. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008; **83**(3)**:** 311-321.

130. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012; **91**(2)**:** 224-237.

131. Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 2010; **42**(11)**:** 969-972.

132. Saccone NL, Schwantes-An TH, Wang JC, Grucza RA, Breslau N, Hatsukami D *et al.* Multiple cholinergic nicotinic receptor genes affect nicotine dependence risk in African and European Americans. *Genes Brain Behav* 2010; **9**(7)**:** 741-750.

133. Bierut LJ, Stitzel JA, Wang JC, Hinrichs AL, Grucza RA, Xuei X *et al.* Variants in nicotinic receptors and risk for nicotine dependence. *The American journal of psychiatry* 2008; **165**(9)**:** 1163-1171.

134. Fowler CD, Lu Q, Johnson PM, Marks MJ, Kenny PJ. Habenular alpha5 nicotinic receptor subunit signalling controls nicotine intake. *Nature* 2011; **471**(7340)**:** 597-601.

135. Lustig LR, Peng H. Chromosome location and characterization of the human nicotinic acetylcholine receptor subunit alpha (alpha) 9 (CHRNA9) gene. *Cytogenetic and genome research* 2002; **98**(2-3)**:** 154-159.

136. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic acids research* 2011; **39**(Database issue)**:** D225-229.

137. Sgard F, Charpantier E, Bertrand S, Walker N, Caput D, Graham D *et al.* A novel human nicotinic receptor subunit, alpha10, that confers functionality to the alpha9-subunit. *Molecular pharmacology* 2002; **61**(1)**:** 150-159.

138. Nguyen VT, Ndoye A, Grando SA. Novel human alpha9 acetylcholine receptor regulating keratinocyte adhesion is targeted by Pemphigus vulgaris autoimmunity. *The American journal of pathology* 2000; **157**(4)**:** 1377-1391.

139. Peng H, Ferris RL, Matthews T, Hiel H, Lopez-Albaitero A, Lustig LR. Characterization of the human nicotinic acetylcholine receptor subunit alpha (alpha) 9 (CHRNA9) and alpha (alpha) 10 (CHRNA10) in lymphocytes. *Life sciences* 2004; **76**(3)**:** 263-280.

140. Lips KS, Pfeil U, Kummer W. Coexpression of alpha 9 and alpha 10 nicotinic acetylcholine receptors in rat dorsal root ganglion neurons. *Neuroscience* 2002; **115**(1)**:** 1-5.

141. Greenbaum L, Kanyas K, Karni O, Merbl Y, Olender T, Horowitz A *et al.* Why do young women smoke? I. Direct and interactive effects of environment, psychological characteristics and nicotinic cholinergic receptor genes. *Mol Psychiatry* 2006; **11**(3)**:** 312-322, 223.

142. Rigbi A, Kanyas K, Yakir A, Greenbaum L, Pollak Y, Ben-Asher E *et al.* Why do young women smoke? V. Role of direct and interactive effects of nicotinic cholinergic receptor gene variation on neurocognitive function. *Genes Brain Behav* 2008; **7**(2)**:** 164-172.

143. Chikova A, Bernard HU, Shchepotin IB, Grando SA. New associations of the genetic polymorphisms in nicotinic receptor genes with the risk of lung cancer. *Life sciences* 2012; **91**(21-22)**:** 1103-1108.

144. Andersson O, Stenqvist A, Attersand A, von Euler G. Nucleotide sequence, genomic organization, and chromosomal localization of genes encoding the human NMDA receptor subunits NR3A and NR3B. *Genomics* 2001; **78**(3)**:** 178-184.

145. Lipton SA, Choi YB, Takahashi H, Zhang D, Li W, Godzik A *et al.* Cysteine regulation of protein function--as exemplified by NMDA-receptor modulation. *Trends in neurosciences* 2002; **25**(9)**:** 474-480.

146. Ciabarra AM, Sullivan JM, Gahn LG, Pecht G, Heinemann S, Sevarino KA. Cloning and characterization of chi-1: a developmentally regulated member of a novel class of the ionotropic glutamate receptor family. *J Neurosci* 1995; **15**(10)**:** 6498-6508.

147. Sucher NJ, Akbarian S, Chi CL, Leclerc CL, Awobuluyi M, Deitcher DL *et al.* Developmental and regional expression pattern of a novel NMDA receptor-like subunit (NMDAR-L) in the rodent brain. *J Neurosci* 1995; **15**(10)**:** 6509-6520.

148. Perez-Otano I, Schulteis CT, Contractor A, Lipton SA, Trimmer JS, Sucher NJ *et al.* Assembly with the NR1 subunit is required for surface expression of NR3A-containing NMDA receptors. *J Neurosci* 2001; **21**(4)**:** 1228-1237.

149. Sasaki YF, Rothe T, Premkumar LS, Das S, Cui J, Talantova MV *et al.* Characterization and comparison of the NR3A subunit of the NMDA receptor in recombinant systems and primary cortical neurons. *Journal of neurophysiology* 2002; **87**(4)**:** 2052-2063.

150. Eriksson M, Nilsson A, Froelich-Fabre S, Akesson E, Dunker J, Seiger A *et al.* Cloning and expression of the human N-methyl-D-aspartate receptor subunit NR3A. *Neuroscience letters* 2002; **321**(3)**:** 177-181.

151. Liu HP, Lin WY, Liu SH, Wang WF, Tsai CH, Wu BT *et al.* Genetic variation in N-methyl-D-aspartate receptor subunit NR3A but not NR3B influences susceptibility to Alzheimer's disease. *Dementia and geriatric cognitive disorders* 2009; **28**(6)**:** 521-527.

152. Takata A, Iwayama Y, Fukuo Y, Ikeda M, Okochi T, Maekawa M *et al.* A population-specific uncommon variant in GRIN3A associated with schizophrenia. *Biol Psychiatry* 2013; **73**(6)**:** 532-539.

153. Marco S, Giralt A, Petrovic MM, Pouladi MA, Martinez-Turrillas R, Martinez-Hernandez J *et al.* Suppressing aberrant GluN3A expression rescues synaptic and behavioral impairments in Huntington's disease models. *Nature medicine* 2013; **19**(8)**:** 1030-1038.

154. Carmelli D, Swan GE, Robinette D, Fabsitz R. Genetic influence on smoking--a study of male twins. *The New England journal of medicine* 1992; **327**(12)**:** 829-833.

155. Berrettine W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H *et al.* alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol Psychiatr* 2008; **13**(4)**:** 368-373.

156. Keskitalo K, Broms U, Heliovaara M, Ripatti S, Surakka I, Perola M *et al.* Association of serum cotinine level with a cluster of three nicotinic acetylcholine receptor genes (CHRNA3/CHRNA5/CHRNB4) on chromosome 15. *Hum Mol Genet* 2009; **18**(20)**:** 4007-4012.

157. Weiss RB, Baker TB, Cannon DS, von Niederhausern A, Dunn DM, Matsunami N *et al.* A Candidate Gene Approach Identifies the CHRNA5-A3-B4 Region as a Risk Factor for Age-Dependent Nicotine Addiction. *Plos Genetics* 2008; **4**(7).

158. Culverhouse RC, Johnson EO, Breslau N, Hatsukami DK, Sadler B, Brooks AI *et al.* Multiple distinct CHRNB3-CHRNA6 variants are genetic risk factors for nicotine dependence in African Americans and European Americans. *Addiction* 2014; **109**(5)**:** 814-822.

159. Hoft NR, Corley RP, McQueen MB, Schlaepfer IR, Huizinga D, Ehringer MA. Genetic association of the CHRNA6 and CHRNB3 genes with tobacco dependence in a nationally representative sample. *Neuropsychopharmacology* 2009; **34**(3)**:** 698-706.

160. Zeiger JS, Haberstick BC, Schlaepfer I, Collins AC, Corley RP, Crowley TJ *et al.* The neuronal nicotinic receptor subunit genes (CHRNA6 and CHRNB3) are associated with subjective responses to tobacco. *Hum Mol Genet* 2008; **17**(5)**:** 724-734.

161. Rice JP, Hartz SM, Agrawal A, Almasy L, Bennett S, Breslau N *et al.* CHRNB3 is more strongly associated with Fagerstrom test for cigarette dependence-based nicotine dependence than cigarettes per day: phenotype definition changes genome-wide association studies results. *Addiction* 2012; **107**(11)**:** 2019-2028.

162. Kumasaka N, Aoki M, Okada Y, Takahashi A, Ozaki K, Mushiroda T *et al.* Haplotypes with copy number and single nucleotide polymorphisms in CYP2A6 locus are associated with smoking quantity in a Japanese population. *PLoS One* 2012; **7**(9)**:** e44507.

163. Chen LS, Bloom AJ, Baker TB, Smith SS, Piper ME, Martinez M *et al.* Pharmacotherapy effects on smoking cessation vary with nicotine metabolism gene (CYP2A6). *Addiction* 2014; **109**(1)**:** 128-137.

164. Bloom AJ, Baker TB, Chen LS, Breslau N, Hatsukami D, Bierut LJ *et al.* Variants in two adjacent genes, EGLN2 and CYP2A6, influence smoking behavior related to disease risk via different mechanisms. *Hum Mol Genet* 2014; **23**(2)**:** 555-561.

165. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G *et al.* Replicating genotype-phenotype associations. *Nature* 2007; **447**(7145)**:** 655-660.

166. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42**(7)**:** 565-569.

167. Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, Goddard ME *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* 2012; **44**(3)**:** 247-250.

168. Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* 2015; **16**(5)**:** 275-284.

169. Gelernter J. Genetics of complex traits in psychiatry. *Biol Psychiatry* 2015; **77**(1)**:** 36-42.

170. Hardin J, He Y, Javitz HS, Wessel J, Krasnow RE, Tildesley E *et al.* Nicotine withdrawal sensitivity, linkage to chr6q26, and association of OPRM1 SNPs in the SMOking in FAMilies (SMOFAM) sample. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2009; **18**(12)**:** 3399-3406.

171. Swan GE, Hops H, Wilhelmsen KC, Lessov-Schlaggar CN, Cheng LS, Hudmon KS *et al.* A genome-wide screen for nicotine dependence susceptibility loci. *Am J Med Genet B Neuropsychiatr Genet* 2006; **141B**(4)**:** 354-360.

172. Han S, Gelernter J, Luo X, Yang BZ. Meta-analysis of 15 genome-wide linkage scans of smoking behavior. *Biol Psychiatry* 2010; **67**(1)**:** 12-19.

173. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* 2012; **13**(8)**:** 537-551.

174. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* 2001; **409**(6822)**:** 951-953.

175. Dani JA. Roles of dopamine signaling in nicotine addiction. *Mol Psychiatry* 2003; **8**(3)**:** 255-256.

176. Gelernter J, Panhuysen C, Weiss R, Brady K, Poling J, Krauthammer M *et al.* Genomewide linkage scan for nicotine dependence: identification of a chromosome 5 risk locus. *Biol Psychiatry* 2007; **61**(1)**:** 119-126.

177. Neville MJ, Johnstone EC, Walton RT. Identification and characterization of ANKK1: a novel kinase gene closely linked to DRD2 on chromosome band 11q23.1. *Hum Mutat* 2004; **23**(6)**:** 540-545.

178. David SP, Mezuk B, Zandi PP, Strong D, Anthony JC, Niaura R *et al.* Sex differences in TTC12/ANKK1 haplotype associations with daily tobacco smoking in Black and White Americans. *Nicotine Tob Res* 2010; **12**(3)**:** 251-262.

179. Ducci F, Kaakinen M, Pouta A, Hartikainen AL, Veijola J, Isohanni M *et al.* TTC12-ANKK1-DRD2 and CHRNA5-CHRNA3-CHRNB4 influence different pathways leading to smoking behavior from adolescence to mid-adulthood. *Biol Psychiatry* 2011; **69**(7)**:** 650-660.

180. David SP, Munafo MR, Murphy MF, Proctor M, Walton RT, Johnstone EC. Genetic variation in the dopamine D4 receptor (DRD4) gene and smoking cessation: follow-up of a randomised clinical trial of transdermal nicotine patch. *Pharmacogenomics J* 2008; **8**(2)**:** 122-128.

181. Ellis JA, Olsson CA, Moore E, Greenwood P, Van De Ven MO, Patton GC. A role for the DRD4 exon III VNTR in modifying the association between nicotine dependence and neuroticism. *Nicotine Tob Res* 2011; **13**(2)**:** 64-69.

182. Das D, Tan X, Easteal S. Effect of model choice in genetic association studies: DRD4 exon III VNTR and cigarette use in young adults. *Am J Med Genet B Neuropsychiatr Genet* 2011; **156B**(3)**:** 346-351.

183. Leventhal AM, Lee W, Bergen AW, Swan GE, Tyndale RF, Lerman C *et al.* Nicotine dependence as a moderator of genetic influences on smoking cessation treatment outcome. *Drug Alcohol Depend* 2014; **138:** 109-117.

184. Ella E, Sato N, Nishizawa D, Kageyama S, Yamada H, Kurabe N *et al.* Association between dopamine beta hydroxylase rs5320 polymorphism and smoking behaviour in elderly Japanese. *J Hum Genet* 2012; **57**(6)**:** 385-390.

185. Yu Y, Panhuysen C, Kranzler HR, Hesselbrock V, Rounsaville B, Weiss R *et al.* Intronic variants in the dopa decarboxylase (DDC) gene are associated with smoking behavior in European-Americans and African-Americans. *Hum Mol Genet* 2006; **15**(14)**:** 2192-2199.

186. Berrettini WH, Wileyto EP, Epstein L, Restine S, Hawk L, Shields P *et al.* Catechol-O-methyltransferase (COMT) gene variants predict response to bupropion therapy for tobacco dependence. *Biol Psychiatry* 2007; **61**(1)**:** 111-118.

187. Amstadter AB, Nugent NR, Koenen KC, Ruggiero KJ, Acierno R, Galea S *et al.* Association between COMT, PTSD, and increased smoking following hurricane exposure in an epidemiologic sample. *Psychiatry* 2009; **72**(4)**:** 360-369.

188. Nedic G, Nikolac M, Borovecki F, Hajnsek S, Muck-Seler D, Pivac N. Association study of a functional catechol-O-methyltransferase polymorphism and smoking in healthy Caucasian subjects. *Neuroscience letters* 2010; **473**(3)**:** 216-219.

189. Omidvar M, Stolk L, Uitterlinden AG, Hofman A, Van Duijn CM, Tiemeier H. The effect of catechol-O-methyltransferase Met/Val functional polymorphism on smoking cessation: retrospective and prospective analyses in a cohort study. *Pharmacogenet Genomics* 2009; **19**(1)**:** 45-51.

190. Munafo MR, Freathy RM, Ring SM, St Pourcain B, Smith GD. Association of COMT Val(108/158)Met genotype and cigarette smoking in pregnant women. *Nicotine Tob Res* 2011; **13**(2)**:** 55-63.

191. Stapleton JA, Sutherland G, O'Gara C. Association between dopamine transporter genotypes and smoking cessation: a meta-analysis. *Addict Biol* 2007; **12**(2)**:** 221-226.

192. Ling D, Niu T, Feng Y, Xing H, Xu X. Association between polymorphism of the dopamine transporter gene and early smoking onset: an interaction risk on nicotine dependence. *J Hum Genet* 2004; **49**(1)**:** 35-39.

193. Farris SP, Harris RA, Ponomarev I. Epigenetic modulation of brain gene networks for cocaine and alcohol abuse. *Frontiers in neuroscience* 2015; **9:** 176.

194. Ramchandani VA, Umhau J, Pavon FJ, Ruiz-Velasco V, Margas W, Sun H *et al.* A genetic determinant of the striatal dopamine response to alcohol in men. *Mol Psychiatry* 2011; **16**(8)**:** 809-817.

195. Domino EF, Evans CL, Ni L, Guthrie SK, Koeppe RA, Zubieta JK. Tobacco smoking produces greater striatal dopamine release in G-allele carriers with mu opioid receptor A118G polymorphism. *Prog Neuropsychopharmacol Biol Psychiatry* 2012; **38**(2)**:** 236-240.

196. Huang S, Cook DG, Hinks LJ, Chen XH, Ye S, Gilg JA *et al.* CYP2A6, MAOA, DBH, DRD4, and 5HT2A genotypes, smoking behaviour and cotinine levels in 1518 UK adolescents. *Pharmacogenet Genomics* 2005; **15**(12)**:** 839-850.

197. Ton TG, Rossing MA, Bowen DJ, Srinouanprachan S, Wicklund K, Farin FM. Genetic polymorphisms in dopamine-related genes and smoking cessation in women: a prospective cohort study. *Behav Brain Funct* 2007; **3:** 22.

198. Breitling LP, Dahmen N, Illig T, Rujescu D, Nitz B, Raum E *et al.* Variants in COMT and spontaneous smoking cessation: retrospective cohort analysis of 925 cessation events. *Pharmacogenet Genomics* 2009; **19**(8)**:** 657-659.

199. Marteau TM, Aveyard P, Munafo MR, Prevost AT, Hollands GJ, Armstrong D *et al.* Effect on adherence to nicotine replacement therapy of informing smokers their dose is determined by their genotype: a randomised controlled trial. *PLoS One* 2012; **7**(4)**:** e35249.

200. Munafo MR, Johnstone EC, Aveyard P, Marteau T. Lack of association of OPRM1 genotype and smoking cessation. *Nicotine Tob Res* 2013; **15**(3)**:** 739-744.

201. Lou XY, Ma JZ, Sun D, Payne TJ, Li MD. Fine mapping of a linkage region on chromosome 17p13 reveals that GABARAP and DLG4 are associated with vulnerability to nicotine dependence in European-Americans. *Hum Mol Genet* 2007; **16**(2)**:** 142-153.

202. Agrawal A, Pergadia ML, Saccone SF, Hinrichs AL, Lessov-Schlaggar CN, Saccone NL *et al.* Gamma-aminobutyric acid receptor genes and nicotine dependence: evidence for association from a case-control study. *Addiction* 2008; **103**(6)**:** 1027-1038.

203. Agrawal A, Pergadia ML, Balasubramanian S, Saccone SF, Hinrichs AL, Saccone NL *et al.* Further evidence for an association between the gamma-aminobutyric acid receptor A, subunit 4 genes on chromosome 4 and Fagerstrom Test for Nicotine Dependence. *Addiction* 2009; **104**(3)**:** 471-477.

204. Cui WY, Seneviratne C, Gu J, Li MD. Genetics of GABAergic signaling in nicotine and alcohol dependence. *Hum Genet* 2012; **131**(6)**:** 843-855.

205. Iordanidou M, Tavridou A, Petridis I, Kyroglou S, Kaklamanis L, Christakidis D *et al.* Association of polymorphisms of the serotonergic system with smoking initiation in Caucasians. *Drug Alcohol Depend* 2010; **108**(1-2)**:** 70-76.

206. Kremer I, Bachner-Melman R, Reshef A, Broude L, Nemanov L, Gritsenko I *et al.* Association of the serotonin transporter gene with smoking behavior. *The American journal of psychiatry* 2005; **162**(5)**:** 924-930.

207. Daw J, Boardman JD, Peterson R, Smolen A, Haberstick BC, Ehringer MA *et al.* The interactive effect of neighborhood peer cigarette use and 5HTTLPR genotype on individual cigarette use. *Addictive behaviors* 2014; **39**(12)**:** 1804-1810.

208. Bidwell LC, Garrett ME, McClernon FJ, Fuemmeler BF, Williams RB, Ashley-Koch AE *et al.* A preliminary analysis of interactions between genotype, retrospective ADHD

symptoms, and initial reactions to smoking in a sample of young adults. *Nicotine Tob Res* 2012; **14**(2)**:** 229-233.

209. Trummer O, Koppel H, Wascher TC, Grunbacher G, Gutjahr M, Stanger O *et al.* The serotonin transporter gene polymorphism is not associated with smoking behavior. *Pharmacogenomics J* 2006; **6**(6)**:** 397-400.

210. David SP, Johnstone EC, Murphy MF, Aveyard P, Guo B, Lerman C *et al.* Genetic variation in the serotonin pathway and smoking cessation with nicotine replacement therapy: new data from the Patch in Practice trial and pooled analyses. *Drug Alcohol Depend* 2008; **98**(1-2)**:** 77-85.

211. do Prado-Lima PA, Chatkin JM, Taufer M, Oliveira G, Silveira E, Neto CA *et al.* Polymorphism of 5HT2A serotonin receptor gene is implicated in smoking addiction. *Am J Med Genet B Neuropsychiatr Genet* 2004; **128B**(1)**:** 90-93.

212. Vink JM, Smit AB, de Geus EJ, Sullivan P, Willemsen G, Hottenga JJ *et al.* Genome-wide association study of smoking initiation and current smoking. *Am J Hum Genet* 2009; **84**(3)**:** 367-379.

213. Ma JZ, Payne TJ, Li MD. Significant association of glutamate receptor, ionotropic N-methyl-D-aspartate 3A (GRIN3A), with nicotine dependence in European- and African-American smokers. *Hum Genet* 2010; **127**(5)**:** 503-512.

214. Grucza RA, Johnson EO, Krueger RF, Breslau N, Saccone NL, Chen LS *et al.* Incorporating age at onset of smoking into genetic models for nicotine dependence: evidence for interaction with multiple genes. *Addict Biol* 2010; **15**(3)**:** 346-357.

215. Li X, Semenova S, D'Souza MS, Stoker AK, Markou A. Involvement of glutamatergic and GABAergic systems in nicotine dependence: Implications for novel pharmacotherapies for smoking cessation. *Neuropharmacology* 2014; **76 Pt B:** 554-565.

216. Sato N, Kageyama S, Chen R, Suzuki M, Tanioka F, Kamo T *et al.* Association between neurexin 1 (NRXN1) polymorphisms and the smoking behavior of elderly Japanese. *Psychiatr Genet* 2010; **20**(3)**:** 135-136.

217. Docampo E, Ribases M, Gratacos M, Bruguera E, Cabezas C, Sanchez-Mora C *et al.* Association of neurexin 3 polymorphisms with smoking behavior. *Genes Brain Behav* 2012; **11**(6)**:** 704-711.

218. Liu QR, Drgon T, Walther D, Johnson C, Poleskaya O, Hess J *et al.* Pooled association genome scanning: validation and use to identify addiction vulnerability loci in two samples. *Proc Natl Acad Sci U S A* 2005; **102**(33)**:** 11864-11869.

219. Feng Y, Niu TH, Xing HX, Xu X, Chen CZ, Peng SJ *et al.* A common haplotype of the nicotine acetylcholine receptor alpha 4 subunit gene is associated with vulnerability to nicotine addiction in men. *American Journal of Human Genetics* 2004; **75**(1)**:** 112-121.

220. Breitling LP, Dahmen N, Mittelstrass K, Rujescu D, Gallinat J, Fehr C *et al.* Association of nicotinic acetylcholine receptor subunit alpha 4 polymorphisms with nicotine dependence in 5500 Germans. *Pharmacogenomics J* 2009; **9**(4)**:** 219-224.

221. Kamens HM, Corley RP, McQueen MB, Stallings MC, Hopfer CJ, Crowley TJ *et al.* Nominal association with CHRNA4 variants and nicotine dependence. *Genes Brain Behav* 2013; **12**(3)**:** 297-304.

222. Lou XY, Ma JZ, Payne TJ, Beuten J, Crew KM, Li MD. 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* 2006; **120**(3)**:** 381-389.

223. Ehringer MA, Clegg HV, Collins AC, Corley RP, Crowley T, Hewitt JK *et al.* Association of the neuronal nicotinic receptor beta 2 subunit gene (CHRNB2) with subjective responses to alcohol and nicotine. *Am J Med Genet B* 2007; **144B**(5)**:** 596-604.

224. Wang S, A DvdV, Xu Q, Seneviratne C, Pomerleau OF, Pomerleau CS *et al.* Significant associations of CHRNA2 and CHRNA6 with nicotine dependence in European American and African American populations. *Hum Genet* 2014; **133**(5)**:** 575-586.

225. Ray R, Tyndale RF, Lerman C. Nicotine dependence pharmacogenetics: role of genetic variation in nicotine-metabolizing enzymes. *Journal of neurogenetics* 2009; **23**(3)**:** 252-261.

226. Johnstone E, Benowitz N, Cargill A, Jacob R, Hinks L, Day I *et al.* Determinants of the rate of nicotine metabolism and effects on smoking behavior. *Clin Pharmacol Ther* 2006; **80**(4)**:** 319-330.

227. Chen LS, Baker TB, Grucza R, Wang JC, Johnson EO, Breslau N *et al.* Dissection of the phenotypic and genotypic associations with nicotinic dependence. *Nicotine Tob Res* 2012; **14**(4)**:** 425-433.

228. Carter B, Long T, Cinciripini P. A meta-analytic review of the CYP2A6 genotype and smoking behavior. *Nicotine Tob Res* 2004; **6**(2)**:** 221-227.

229. Zhang XY, Chen da C, Xiu MH, Luo X, Zuo L, Haile CN *et al.* BDNF Val66Met variant and smoking in a Chinese population. *PLoS One* 2012; **7**(12)**:** e53295.

230. Li MD, Sun D, Lou XY, Beuten J, Payne TJ, Ma JZ. Linkage and association studies in African- and Caucasian-American populations demonstrate that SHC3 is a novel susceptibility locus for nicotine dependence. *Mol Psychiatry* 2007; **12**(5)**:** 462-473.

231. Chen GB, Payne TJ, Lou XY, Ma JZ, Zhu J, Li MD. Association of amyloid precursor protein-binding protein, family B, member 1 with nicotine dependence in African and European American smokers. *Hum Genet* 2008; **124**(4)**:** 393-398.

232. Zhang L, Kendler KS, Chen X. Association of the phosphatase and tensin homolog gene (PTEN) with smoking initiation and nicotine dependence. *Am J Med Genet B Neuropsychiatr Genet* 2006; **141B**(1)**:** 10-14.

233. Turner JR, Ray R, Lee B, Everett L, Xiang J, Jepson C *et al.* Evidence from mouse and man for a role of neuregulin 3 in nicotine dependence. *Mol Psychiatry* 2014; **19**(7)**:** 801-810.

234. Li CY, Mao X, Wei L. Genes and (common) pathways underlying drug addiction. *PLoS computational biology* 2008; **4**(1)**:** e2.

235. Yudkin P, Munafo M, Hey K, Roberts S, Welch S, Johnstone E *et al.* Effectiveness of nicotine patches in relation to genotype in women versus men: randomised controlled trial. *BMJ* 2004; **328**(7446)**:** 989-990.

236. Johnstone EC, Yudkin PL, Hey K, Roberts SJ, Welch SJ, Murphy MF *et al.* Genetic variation in dopaminergic pathways and short-term effectiveness of the nicotine patch. *Pharmacogenetics* 2004; **14**(2)**:** 83-90.

237. Lerman C, Jepson C, Wileyto EP, Epstein LH, Rukstalis M, Patterson F *et al.* Role of functional genetic variation in the dopamine D2 receptor (DRD2) in response to bupropion and nicotine replacement therapy for tobacco dependence: results of two randomized clinical trials. *Neuropsychopharmacology* 2006; **31**(1)**:** 231-242.

238. Comings DE, Ferry L, Bradshaw-Robinson S, Burchette R, Chiu C, Muhleman D. The dopamine D2 receptor (DRD2) gene: a genetic risk factor in smoking. *Pharmacogenetics* 1996; **6**(1)**:** 73-79.

239. Liu YZ, Pei YF, Guo YF, Wang L, Liu XG, Yan H *et al.* Genome-wide association analyses suggested a novel mechanism for smoking behavior regulated by IL15. *Mol Psychiatry* 2009; **14**(7)**:** 668-680.

240. Beuten J, Ma JZ, Lou XY, Payne TJ, Li MD. Association analysis of the protein phosphatase 1 regulatory subunit 1B (PPP1R1B) gene with nicotine dependence in European- and African-American smokers. *Am J Med Genet B Neuropsychiatr Genet* 2007; **144B**(3)**:** 285-290.

241. Munafo MR, Elliot KM, Murphy MF, Walton RT, Johnstone EC. Association of the mu-opioid receptor gene with smoking cessation. *Pharmacogenomics J* 2007; **7**(5)**:** 353-361.

242. Nees F, Witt SH, Lourdusamy A, Vollstadt-Klein S, Steiner S, Poustka L *et al.* Genetic risk for nicotine dependence in the cholinergic system and activation of the brain reward system in healthy adolescents. *Neuropsychopharmacology* 2013; **38**(11)**:** 2081-2089.

243. Chen LS, Baker TB, Piper ME, Breslau N, Cannon DS, Doheny KF *et al.* Interplay of genetic risk factors (CHRNA5-CHRNA3-CHRNB4) and cessation treatments in smoking cessation success. *The American journal of psychiatry* 2012; **169**(7)**:** 735-742.

244. Falcone M, Jepson C, Benowitz N, Bergen AW, Pinto A, Wileyto EP *et al.* Association of the nicotine metabolite ratio and CHRNA5/CHRNA3 polymorphisms with smoking rate among treatment-seeking smokers. *Nicotine Tob Res* 2011; **13**(6)**:** 498-503.

245. Schlaepfer IR, Hoft NR, Collins AC, Corley RP, Hewitt JK, Hopfer CJ *et al.* The CHRNA5/A3/B4 gene cluster variability as an important determinant of early alcohol and tobacco initiation in young adults. *Biol Psychiatry* 2008; **63**(11)**:** 1039-1046.

246. Polina ER, Rovaris DL, de Azeredo LA, Mota NR, Vitola ES, Silva KL *et al.* ADHD diagnosis may influence the association between polymorphisms in nicotinic acetylcholine receptor genes and tobacco smoking. *Neuromolecular medicine* 2014; **16**(2)**:** 389-397.

247. Rodriguez S, Cook DG, Gaunt TR, Nightingale CM, Whincup PH, Day IN. Combined analysis of CHRNA5, CHRNA3 and CYP2A6 in relation to adolescent smoking behaviour. *J Psychopharmacol* 2011; **25**(7)**:** 915-923.

248. Munafo MR, Johnstone EC, Walther D, Uhl GR, Murphy MF, Aveyard P. CHRNA3 rs1051730 genotype and short-term smoking cessation. *Nicotine Tob Res* 2011; **13**(10)**:** 982-988.

249. Saccone NL, Saccone SF, Hinrichs AL, Stitzel JA, Duan W, Pergadia ML *et al.* Multiple distinct risk loci for nicotine dependence identified by dense coverage of the complete family of nicotinic receptor subunit (CHRN) genes. *Am J Med Genet B Neuropsychiatr Genet* 2009; **150B**(4)**:** 453-466.

250. Chen X, Chen J, Williamson VS, An SS, Hettema JM, Aggen SH *et al.* Variants in nicotinic acetylcholine receptors alpha5 and alpha3 increase risks to nicotine dependence. *Am J Med Genet B Neuropsychiatr Genet* 2009; **150B**(7)**:** 926-933.

251. Johnson EO, Chen LS, Breslau N, Hatsukami D, Robbins T, Saccone NL *et al.* Peer smoking and the nicotinic receptor genes: an examination of genetic and environmental risks for nicotine dependence. *Addiction* 2010; **105**(11)**:** 2014-2022.

252. Xie P, Kranzler HR, Zhang H, Oslin D, Anton RF, Farrer LA *et al.* Childhood Adversity Increases Risk for Nicotine Dependence and Interacts with alpha5 Nicotinic Acetylcholine Receptor Genotype Specifically in Males. *Neuropsychopharmacology* 2011.

253. Freathy RM, Ring SM, Shields B, Galobardes B, Knight B, Weedon MN *et al.* A common genetic variant in the 15q24 nicotinic acetylcholine receptor gene cluster (CHRNA5-CHRNA3-CHRNB4) is associated with a reduced ability of women to quit smoking in pregnancy. *Hum Mol Genet* 2009; **18**(15)**:** 2922-2927.

254. Bergen AW, Javitz HS, Krasnow R, Nishita D, Michel M, Conti DV *et al.* Nicotinic acetylcholine receptor variation and response to smoking cessation therapies. *Pharmacogenet Genomics* 2013; **23**(2)**:** 94-103.

255. Stevens VL, Bierut LJ, Talbot JT, Wang JC, Sun J, Hinrichs AL *et al.* Nicotinic receptor gene variants influence susceptibility to heavy smoking. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2008; **17**(12)**:** 3517-3525.

256. Sorice R, Bione S, Sansanelli S, Ulivi S, Athanasakis E, Lanzara C *et al.* Association of a variant in the CHRNA5-A3-B4 gene cluster region to heavy smoking in the Italian population. *European journal of human genetics : EJHG* 2011; **19**(5)**:** 593-596.

257. Maes HH, Neale MC, Chen X, Chen J, Prescott CA, Kendler KS. A twin association study of nicotine dependence with markers in the CHRNA3 and CHRNA5 genes. *Behavior genetics* 2011; **41**(5)**:** 680-690.

258. Rode L, Bojesen SE, Weischer M, Nordestgaard BG. High tobacco consumption is causally associated with increased all-cause mortality in a general population sample of 55,568 individuals, but not with short telomeres: a Mendelian randomization study. *Int J Epidemiol* 2014; **43**(5)**:** 1473-1483.

259. Hartz SM, Short SE, Saccone NL, Culverhouse R, Chen L, Schwantes-An TH *et al.* Increased genetic vulnerability to smoking at CHRNA5 in early-onset smokers. *Arch Gen Psychiatry* 2012; **69**(8)**:** 854-860.

260. Saccone NL, Culverhouse RC, Schwantes-An TH, Cannon DS, Chen X, Cichon S *et al.* Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet* 2010; **6**(8).

261. Bar-Shira A, Gana-Weisz M, Gan-Or Z, Giladi E, Giladi N, Orr-Urtreger A. CHRNB3 c.-57A>G functional promoter change affects Parkinson's disease and smoking. *Neurobiology of aging* 2014; **35**(9)**:** 2179 e2171-2176.

262. Lee CT, Fuemmeler BF, McClernon FJ, Ashley-Koch A, Kollins SH. Nicotinic receptor gene variants interact with attention deficient hyperactive disorder symptoms to predict smoking trajectories from early adolescence to adulthood. *Addictive behaviors* 2013; **38**(11)**:** 2683-2689.

263. Loukola A, Buchwald J, Gupta R, Palviainen T, Hallfors J, Tikkanen E *et al.* A Genome-Wide Association Study of a Biomarker of Nicotine Metabolism. *PLoS Genet* 2015; **11**(9)**:** e1005498.

264. Cannon DS, Baker TB, Piper ME, Scholand MB, Lawrence DL, Drayna DT *et al.* Associations between phenylthiocarbamide gene polymorphisms and cigarette smoking. *Nicotine Tob Res* 2005; **7**(6)**:** 853-858.

265. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**(5720)**:** 385-389.

266. Gelernter J, Kranzler HR, Sherva R, Almasy L, Herman AI, Koesterer R *et al.* Genome-wide association study of nicotine dependence in American populations: identification of novel risk loci in both African-Americans and European-Americans. *Biol Psychiatry* 2015; **77**(5)**:** 493-503.

267. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; **11**(6)**:** 415-425.

268. Slimak MA, Ables JL, Frahm S, Antolin-Fontes B, Santos-Torres J, Moretti M *et al.* Habenular expression of rare missense variants of the beta4 nicotinic receptor subunit alters nicotine consumption. *Frontiers in human neuroscience* 2014; **8:** 12.

269. Munafo MR, Clark TG, Johnstone EC, Murphy FG, Walton RT. The genetics basis for smoking behavior: A systematic review and meta-analysis. *Nicotine Tob Res* 2004; **6**(4)**:** 583-597.

270. TGC. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; **42**(5)**:** 441-447.

271. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**(5281)**:** 1516-1517.

272. Piper ME, McCarthy DE, Baker TB. Assessing tobacco dependence: a guide to measure evaluation and selection. *Nicotine Tob Res* 2006; **8**(3)**:** 339-351.

273. Egan MF, Kojima M, Callicott JH, Goldberg TE, Kolachana BS, Bertolino A *et al.* The BDNF val66met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell* 2003; **112**(2)**:** 257-269.

274. Ducci F, Goldman D. The genetic basis of addictive disorders. *The Psychiatric clinics of North America* 2012; **35**(2)**:** 495-519.

275. Hoft NR, Stitzel JA, Hutchison KE, Ehringer MA. CHRNB2 promoter region: association with subjective effects to nicotine and gene expression differences. *Genes Brain Behav* 2011; **10**(2)**:** 176-185.

276. Domino EF, Hirasawa-Fujita M, Ni L, Guthrie SK, Zubieta JK. Regional brain [(11)C]carfentanil binding following tobacco smoking. *Prog Neuropsychopharmacol Biol Psychiatry* 2015; **59:** 100-104.

277. Ray R, Ruparel K, Newberg A, Wileyto EP, Loughead JW, Divgi C *et al.* Human Mu Opioid Receptor (OPRM1 A118G) polymorphism is associated with brain mu-opioid receptor binding potential in smokers. *Proc Natl Acad Sci U S A* 2011; **108**(22)**:** 9268-9273.

278. Dash B, Li MD. Two rare variations, D478N and D478E, that occur at the same amino acid residue in nicotinic acetylcholine receptor (nAChR) alpha2 subunit influence nAChR function. *Neuropharmacology* 2014; **85:** 471-481.

279. Dash B, Lukas RJ, Li MD. A signal peptide missense mutation associated with nicotine dependence alters alpha2*-nicotinic acetylcholine receptor function. *Neuropharmacology* 2014; **79:** 715-725.

280. Kamens HM, Miyamoto J, Powers MS, Ro K, Soto M, Cox R *et al.* The beta3 subunit of the nicotinic acetylcholine receptor: Modulation of gene expression and nicotine consumption. *Neuropharmacology* 2015; **99:** 639-649.

281. Xu X, Clark US, David SP, Mulligan RC, Knopik VS, McGeary J *et al.* Effects of nicotine deprivation and replacement on BOLD-fMRI response to smoking cues as a function of DRD4 VNTR genotype. *Nicotine Tob Res* 2014; **16**(7)**:** 939-947.

282. Hong LE, Hodgkinson CA, Yang Y, Sampath H, Ross TJ, Buchholz B *et al.* A genetically modulated, intrinsic cingulate circuit supports human nicotine addiction. *Proc Natl Acad Sci U S A* 2010; **107**(30)**:** 13509-13514.

283. Kuryatov A, Berrettini W, Lindstrom J. Acetylcholine receptor (AChR) alpha5 subunit variant associated with risk for nicotine dependence and lung cancer reduces (alpha4beta2)(2)alpha5 AChR function. *Molecular pharmacology* 2011; **79**(1)**:** 119-125.

284. Hu XZ, Lipsky RH, Zhu G, Akhtar LA, Taubman J, Greenberg BD *et al.* Serotonin transporter promoter gain-of-function genotypes are linked to obsessive-compulsive disorder. *Am J Hum Genet* 2006; **78**(5)**:** 815-826.

285. Little KY, McLauglin DP, Ranc J, Gilmore J, Lopez JF, Watson SJ *et al.* Serotonin transporter binding sites and mRNA levels in depressed persons committing suicide. *Biol Psychiatry* 1997; **41**(12)**:** 1156-1164.

286. Heinz A, Jones DW, Mazzanti C, Goldman D, Ragan P, Hommer D *et al.* A relationship between serotonin transporter genotype and in vivo protein expression and alcohol neurotoxicity. *Biol Psychiatry* 2000; **47**(7)**:** 643-649.

287. Tricker AR. Nicotine metabolism, human drug metabolism polymorphisms, and smoking behaviour. *Toxicology* 2003; **183**(1-3)**:** 151-173.

288. Eggert M, Winterer G, Wanischeck M, Hoda JC, Bertrand D, Steinlein O. The nicotinic acetylcholine receptor alpha 4 subunit contains a functionally relevant SNP Haplotype. *BMC genetics* 2015; **16:** 46.

289. Chen J, Lipska BK, Halim N, Ma QD, Matsumoto M, Melhem S *et al.* Functional analysis of genetic variation in catechol-O-methyltransferase (COMT): effects on mRNA, protein, and enzyme activity in postmortem human brain. *Am J Hum Genet* 2004; **75**(5)**:** 807-821.

290. Uhl GR, Liu QR, Drgon T, Johnson C, Walther D, Rose JE *et al.* Molecular genetics of successful smoking cessation: convergent genome-wide association study results. *Arch Gen Psychiatry* 2008; **65**(6)**:** 683-693.

291. Drgon T, Montoya I, Johnson C, Liu QR, Walther D, Hamer D *et al.* Genome-wide association for nicotine dependence and smoking cessation success in NIH research volunteers. *Mol Med* 2009; **15**(1-2)**:** 21-27.

292. Drgon T, Johnson C, Walther D, Albino AP, Rose JE, Uhl GR. Genome-wide association for smoking cessation success: participants in a trial with adjunctive denicotinized cigarettes. *Mol Med* 2009; **15**(7-8)**:** 268-274.

293. Uhl GR, Drgon T, Johnson C, Walther D, David SP, Aveyard P *et al.* Genome-wide association for smoking cessation success: participants in the Patch in Practice trial of nicotine replacement. *Pharmacogenomics* 2010; **11**(3)**:** 357-367.

294. Hamidovic A, Goodloe RJ, Bergen AW, Benowitz NL, Styn MA, Kasberger JL *et al.* Gene-centric analysis of serum cotinine levels in African and European American populations. *Neuropsychopharmacology* 2012; **37**(4)**:** 968-974.

295. Jackson KJ, Sanjakdar SS, Chen X, Damaj MI. Nicotine reward and affective nicotine withdrawal signs are attenuated in calcium/calmodulin-dependent protein kinase IV knockout mice. *PLoS One* 2012; **7**(11)**:** e51154.

296. Mutschler J, Abbruzzese E, von der Goltz C, Dinter C, Mobascher A, Thiele H *et al.* Genetic variation in the neuropeptide Y gene promoter is associated with increased risk of tobacco smoking. *European addiction research* 2012; **18**(5)**:** 246-252.

297. Chen X, Wu B, Kendler KS. Association study of the Epac gene and tobacco smoking and nicotine dependence. *Am J Med Genet B Neuropsychiatr Genet* 2004; **129B**(1)**:** 116-119.

298. Chen X, Che Y, Zhang L, Putman AH, Damaj I, Martin BR *et al.* RhoA, encoding a Rho GTPase, is associated with smoking initiation. *Genes Brain Behav* 2007; **6**(8)**:** 689-697.

299. Ton TG, Rossing MA, Bowen DJ, Wilkerson HW, Farin FM. Cholecystokinin C-45T polymorphism and smoking cessation in women. *Nicotine Tob Res* 2007; **9**(1)**:** 147-151.

300. O'Gara C, Stapleton J, Sutherland G, Guindalini C, Neale B, Breen G *et al.* Dopamine transporter polymorphisms are associated with short-term response to smoking cessation treatment. *Pharmacogenet Genomics* 2007; **17**(1)**:** 61-67.

301. Rovaris DL, Mota NR, de Azeredo LA, Cupertino RB, Bertuzzi GP, Polina ER *et al.* MR and GR functional SNPs may modulate tobacco smoking susceptibility. *J Neural Transm* 2013; **120**(10)**:** 1499-1505.

302. Chen HI, Shinkai T, Utsunomiya K, Yamada K, Sakata S, Fukunaka Y *et al.* Possible association of nicotinic acetylcholine receptor gene (CHRNA4 and CHRNB2) polymorphisms with nicotine dependence in Japanese males: an exploratory study. *Pharmacopsychiatry* 2013; **46**(2)**:** 77-82.

303. Smits KM, Benhamou S, Garte S, Weijenberg MP, Alamanos Y, Ambrosone C *et al.* Association of metabolic gene polymorphisms with tobacco consumption in healthy controls. *International journal of cancer Journal international du cancer* 2004; **110**(2)**:** 266-270.

304. Saadat M, Mohabatkar H. Polymorphisms of glutathione S-transferases M1 and T1 do not account for interindividual differences for smoking behavior. *Pharmacol Biochem Behav* 2004; **77**(4)**:** 793-795.

305. Rodriguez S, Huang S, Chen XH, Gaunt TR, Syddall HE, Gilg JA *et al.* A study of TH01 and IGF2-INS-TH haplotypes in relation to smoking initiation in three independent surveys. *Pharmacogenet Genomics* 2006; **16**(1)**:** 15-23.

306. Siiskonen SJ, Visser LE, Tiemeier H, Hofman A, Lamberts SW, Uitterlinden AG *et al.* BclI glucocorticoid receptor polymorphism and smoking in the general population. *Addict Biol* 2009; **14**(3)**:** 349-355.

307. Munafo MR, Timpson NJ, David SP, Ebrahim S, Lawlor DA. Association of the DRD2 gene Taq1A polymorphism and smoking behavior: a meta-analysis and new data. *Nicotine Tob Res* 2009; **11**(1)**:** 64-76.

308. Munafo MR, Johnstone EC, Murphy MF, Aveyard P. Lack of association of DRD2 rs1800497 (Taq1A) polymorphism with smoking cessation in a nicotine replacement therapy randomized trial. *Nicotine Tob Res* 2009; **11**(4)**:** 404-407.

309. Breitling LP, Muller H, Illig T, Rujescu D, Winterer G, Dahmen N *et al.* Dopamine-related genes and spontaneous smoking cessation in ever-heavy smokers. *Pharmacogenomics* 2011; **12**(8)**:** 1099-1106.

310. Hubacek JA, Dlouha D, Lanska V, Adamkova V. Lack of an association between three tagging SNPs within the FTO gene and smoking behavior. *Nicotine Tob Res* 2012; **14**(8)**:** 998-1002.

311. Spruell T, Colavita G, Donegan T, Egawhary M, Hurley M, Aveyard P *et al.* Association between nicotinic acetylcholine receptor single nucleotide polymorphisms and smoking cessation. *Nicotine Tob Res* 2012; **14**(8)**:** 993-997.

312. Chenoweth MJ, Zhu AZ, Sanderson Cox L, Ahluwalia JS, Benowitz NL, Tyndale RF. Variation in P450 oxidoreductase (POR) A503V and flavin-containing monooxygenase (FMO)-3 E158K is associated with minor alterations in nicotine metabolism, but does not alter cigarette consumption. *Pharmacogenet Genomics* 2014; **24**(3)**:** 172-176.

313. Zhao Z, Guo AY, van den Oord EJ, Aliev F, Jia P, Edenberg HJ *et al.* Multi-species data integration and gene ranking enrich significant results in an alcoholism genome-wide association study. *BMC genomics* 2012; **13 Suppl 8:** S16.

314. Lind PA, Macgregor S, Vink JM, Pergadia ML, Hansell NK, de Moor MH *et al.* A genomewide association study of nicotine and alcohol dependence in Australian and Dutch populations. *Twin research and human genetics : the official journal of the International Society for Twin Studies* 2010; **13**(1)**:** 10-29.

315. Wang JC, Hinrichs AL, Stock H, Budde J, Allen R, Bertelsen S *et al.* Evidence of common and specific genetic effects: association of the muscarinic acetylcholine receptor M2 (CHRM2) gene with alcohol dependence and major depressive syndrome. *Hum Mol Genet* 2004; **13**(17)**:** 1903-1911.

316. Coon H, Piasecki TM, Cook EH, Dunn D, Mermelstein RJ, Weiss RB *et al.* Association of the CHRNA4 neuronal nicotinic receptor subunit gene with frequency of binge drinking in young adults. *Alcohol Clin Exp Res* 2014; **38**(4)**:** 930-937.

317. Wang JC, Grucza R, Cruchaga C, Hinrichs AL, Bertelsen S, Budde JP *et al.* Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence. *Mol Psychiatr* 2009; **14**(5)**:** 501-510.

318. Sherva R, Kranzler HR, Yu Y, Logue MW, Poling J, Arias AJ *et al.* Variation in nicotinic acetylcholine receptor genes is associated with multiple substance dependence phenotypes. *Neuropsychopharmacology* 2010; **35**(9)**:** 1921-1931.

319. Haller G, Kapoor M, Budde J, Xuei X, Edenberg H, Nurnberger J *et al.* Rare missense variants in CHRNB3 and CHRNA3 are associated with risk of alcohol and cocaine dependence. *Hum Mol Genet* 2014; **23**(3)**:** 810-819.

320. Landgren S, Engel JA, Andersson ME, Gonzalez-Quintela A, Campos J, Nilsson S *et al.* Association of nAChR gene haplotypes with heavy alcohol use and body mass. *Brain research* 2009; **1305 Suppl:** S72-79.

321. Hoft NR, Corley RP, McQueen MB, Huizinga D, Menard S, Ehringer MA. SNPs in CHRNA6 and CHRNB3 are associated with alcohol consumption in a nationally representative sample. *Genes Brain Behav* 2009; **8**(6)**:** 631-637.

322. Guillot CR, Fanning JR, Liang T, Berman ME. COMT Associations with Disordered Gambling and Drinking Measures. *Journal of gambling studies / co-sponsored by the National Council on Problem Gambling and Institute for the Study of Gambling and Commercial Gaming* 2015; **31**(2)**:** 513-524.

323. Preuss UW, Wurst FM, Ridinger M, Rujescu D, Fehr C, Koller G *et al.* Association of functional DBH genetic variants with alcohol dependence risk and related depression and suicide attempt phenotypes: results from a large multicenter association study. *Drug Alcohol Depend* 2013; **133**(2)**:** 459-467.

324. Pan Y, Luo X, Liu X, Wu LY, Zhang Q, Wang L *et al.* Genome-wide association studies of maximum number of drinks. *Journal of psychiatric research* 2013; **47**(11)**:** 1717-1724.

325. Prasad P, Ambekar A, Vaswani M. Case-control association analysis of dopamine receptor polymorphisms in alcohol dependence: a pilot study in Indian males. *BMC research notes* 2013; **6:** 418.

326. Edenberg HJ, Dick DM, Xuei XL, Tian HJ, Almasy L, Bauer LO *et al.* Variations in GABRA2, encoding the alpha 2 subunit of the GABA(A) receptor, are associated with alcohol dependence and with brain oscillations. *American Journal of Human Genetics* 2004; **74**(4)**:** 705-714.

327. Covault J, Gelernter J, Hesselbrock V, Nellissery M, Kranzler HR. Allelic and haplotypic association of GABRA2 with alcohol dependence. *Am J Med Genet B* 2004; **129B**(1)**:** 104-109.

328. Lappalainen J, Krupitsky E, Remizov M, Pchelina S, Taraskina A, Zvartau E *et al.* Association between alcoholism and gamma-amino butyric acid alpha 2 receptor subtype in a Russian population. *Alcoholism-Clinical and Experimental Research* 2005; **29**(4)**:** 493-498.

329. Fehr C, Sander T, Tadic A, Lenzen KP, Anghelescu I, Klawe C *et al.* Confirmation of association of the GABRA2 gene with alcohol dependence by subtype-specific analysis. *Psychiat Genet* 2006; **16**(1)**:** 9-17.

330. Lind PA, Macgregor S, Agrawal A, Montgomery GW, Heath AC, Martin NG *et al.* The role of GABRA2 in alcohol dependence, smoking, and illicit drug use in an Australian population sample. *Alcohol Clin Exp Res* 2008; **32**(10)**:** 1721-1731.

331. Enoch MA, Rosser AA, Zhou Z, Mash DC, Yuan Q, Goldman D. Expression of glutamatergic genes in healthy humans across 16 brain regions; altered expression in the hippocampus after chronic exposure to alcohol or cocaine. *Genes Brain Behav* 2014; **13**(8)**:** 758-768.

332. Yang HC, Chang CC, Lin CY, Chen CL, Fann CS. A genome-wide scanning and fine mapping study of COGA data. *BMC genetics* 2005; **6 Suppl 1:** S30.

333. Hishimoto A, Liu QR, Drgon T, Pletnikova O, Walther D, Zhu XG *et al.* Neurexin 3 polymorphisms are associated with alcohol dependence and altered expression of specific isoforms. *Hum Mol Genet* 2007; **16**(23)**:** 2880-2891.

334. Ray LA, Hutchison KE. A polymorphism of the mu-opioid receptor gene (OPRM1) and sensitivity to the effects of alcohol in humans. *Alcohol Clin Exp Res* 2004; **28**(12)**:** 1789-1795.

335. Morozova TV, Mackay TF, Anholt RR. Genetics and genomics of alcohol sensitivity. *Molecular genetics and genomics : MGG* 2014; **289**(3)**:** 253-269.

336. McHugh RK, Hofmann SG, Asnaani A, Sawyer AT, Otto MW. The serotonin transporter gene and risk for alcohol dependence: a meta-analytic review. *Drug Alcohol Depend* 2010; **108**(1-2)**:** 1-6.

337. Wang JC, Hinrichs AL, Bertelsen S, Stock H, Budde JP, Dick DM *et al.* Functional variants in TAS2R38 and TAS2R16 influence alcohol consumption in high-risk families of African-American origin. *Alcohol Clin Exp Res* 2007; **31**(2)**:** 209-215.

338. Noble EP, Blum K. Alcoholism and the D2 dopamine receptor gene. *Jama* 1993; **270**(13)**:** 1547-1548.

339. Munafo MR, Matheson IJ, Flint J. Association of the DRD2 gene Taq1A polymorphism and alcoholism: a meta-analysis of case-control studies and evidence of publication bias. *Mol Psychiatry* 2007; **12**(5)**:** 454-461.

340. Smith L, Watson M, Gates S, Ball D, Foxcroft D. Meta-analysis of the association of the Taq1A polymorphism with the risk of alcohol dependency: a HuGE gene-disease association review. *American journal of epidemiology* 2008; **167**(2)**:** 125-138.

341. Yang BZ, Kranzler HR, Zhao HY, Gruen JR, Luo XG, Gelernter J. Haplotypic Variants in DRD2, ANKK1, TTC12, and NCAM1 are Associated With Comorbid Alcohol and Drug Dependence. *Alcoholism-Clinical and Experimental Research* 2008; **32**(12)**:** 2117-2127.

342. Flanagin BA, Cook EH, Jr., de Wit H. An association study of the brain-derived neurotrophic factor Val66Met polymorphism and amphetamine response. *Am J Med Genet B Neuropsychiatr Genet* 2006; **141B**(6)**:** 576-583.

343. Uhl GR, Drgon T, Liu QR, Johnson C, Walther D, Komiyama T *et al.* Genome-wide association for methamphetamine dependence: convergent results from 2 samples. *Arch Gen Psychiatry* 2008; **65**(3)**:** 345-355.

344. Corley RP, Zeiger JS, Crowley T, Ehringer MA, Hewitt JK, Hopfer CJ *et al.* Association of candidate genes with antisocial drug dependence in adolescents. *Drug Alcohol Depend* 2008; **96**(1-2)**:** 90-98.

345. Grucza RA, Wang JC, Stitzel JA, Hinrichs AL, Saccone SF, Saccone NL *et al.* A Risk Allele for Nicotine Dependence in CHRNA5 Is a Protective Allele for Cocaine Dependence. *Biol Psychiat* 2008; **64**(11)**:** 922-929.

346. Ittiwut R, Listman JB, Ittiwut C, Cubells JF, Weiss RD, Brady K *et al.* Association between polymorphisms in catechol-O-methyltransferase (COMT) and cocaine-induced paranoia in European-American and African-American populations. *Am J Med Genet B Neuropsychiatr Genet* 2011; **156B**(6)**:** 651-660.

347. Kalayasiri R, Sughondhabirom A, Gueorguieva R, Coric V, Lynch WJ, Lappalainen J *et al.* Dopamine beta-hydroxylase gene (DbetaH) -1021C-->T influences self-reported paranoia during cocaine self-administration. *Biol Psychiatry* 2007; **61**(11)**:** 1310-1313.

348. Smith SS, O'Hara BF, Persico AM, Gorelick DA, Newlin DB, Vlahov D *et al.* Genetic vulnerability to drug abuse. The D2 dopamine receptor Taq I B1 restriction fragment length polymorphism appears more frequently in polysubstance abusers. *Arch Gen Psychiatry* 1992; **49**(9)**:** 723-727.

349. Li T, Liu X, Zhao J, Hu X, Ball DM, Loh el W *et al.* Allelic association analysis of the dopamine D2, D3, 5-HT2A, and GABA(A)gamma2 receptors and serotonin transporter genes with heroin abuse in Chinese subjects. *Am J Med Genet* 2002; **114**(3)**:** 329-335.

350. Nelson EC, Heath AC, Lynskey MT, Agrawal A, Henders AK, Bowdler LM *et al.* PTSD risk associated with a functional DRD2 polymorphism in heroin-dependent cases and controls is limited to amphetamine-dependent individuals. *Addict Biol* 2014; **19**(4)**:** 700-707.

351. Kotler M, Cohen H, Segman R, Gritsenko I, Nemanov L, Lerer B *et al.* Excess dopamine D4 receptor (D4DR) exon III seven repeat allele in opioid-dependent subjects. *Mol Psychiatry* 1997; **2**(3)**:** 251-254.

352. Li T, Xu K, Deng H, Cai G, Liu J, Liu X *et al.* Association analysis of the dopamine D4 gene exon III VNTR and heroin abuse in Chinese subjects. *Mol Psychiatry* 1997; **2**(5)**:** 413-416.

353. Agrawal A, Edenberg HJ, Foroud T, Bierut LJ, Dunne G, Hinrichs AL *et al.* Association of GABRA2 with drug dependence in the collaborative study of the genetics of alcoholism sample. *Behavior genetics* 2006; **36**(5)**:** 640-650.

354. Smelson D, Yu L, Buyske S, Gonzalez G, Tischfield J, Deutsch CK *et al.* Genetic association of GABA-A receptor alpha-2 and mu opioid receptor with cocaine cue-reactivity: evidence for inhibitory synaptic neurotransmission involvement in cocaine dependence. *The American journal on addictions / American Academy of Psychiatrists in Alcoholism and Addictions* 2012; **21**(5)**:** 411-415.

355. Bond C, LaForge KS, Tian M, Melia D, Zhang S, Borg L *et al.* Single-nucleotide polymorphism in the human mu opioid receptor gene alters beta-endorphin binding and activity: possible implications for opiate addiction. *Proc Natl Acad Sci U S A* 1998; **95**(16)**:** 9608-9613.

356. Hoehe MR, Kopke K, Wendel B, Rohde K, Flachmeier C, Kidd KK *et al.* Sequence variability and candidate gene analysis in complex disease: association of mu opioid receptor gene variation with substance dependence. *Hum Mol Genet* 2000; **9**(19)**:** 2895-2908.

357. Tan EC, Yeo BK, Ho BK, Tay AH, Tan CH. Evidence for an association between heroin dependence and a VNTR polymorphism at the serotonin transporter locus. *Mol Psychiatry* 1999; **4**(3)**:** 215-217.

358. Kozlowski LT, Porter CQ, Orleans CT, Pope MA, Heatherton T. Predicting smoking cessation with self-reported measures of nicotine dependence: FTQ, FTND, and HSI. *Drug Alcohol Depend* 1994; **34**(3)**:** 211-216.

359. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* 2013; **342**(6159)**:** 747-749.

360. Zuo L, Gelernter J, Zhang CK, Zhao H, Lu L, Kranzler HR *et al.* Genome-wide association study of alcohol dependence implicates KIAA0040 on chromosome 1q. *Neuropsychopharmacology* 2012; **37**(2)**:** 557-566.

361. Hancock DB, Levy JL, Gaddis NC, Glasheen C, Saccone NL, Page GP *et al.* Cis-Expression Quantitative Trait Loci Mapping Reveals Replicable Associations with Heroin Addiction in OPRM1. *Biol Psychiatry* 2015; **78**(7)**:** 474-484.

362. Hancock DB, Wang JC, Gaddis NC, Levy JL, Saccone NL, Stitzel JA *et al.* A multiancestry study identifies novel genetic associations with CHRNA5 methylation in human brain and risk of nicotine dependence. *Hum Mol Genet* 2015; **24**(20)**:** 5940-5954.

363. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518**(7539)**:** 317-330.

364. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**(7414)**:** 57-74.

365. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**(6235)**:** 648-660.

366. Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT *et al.* Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* 2011; **478**(7370)**:** 519-523.

367. Numata S, Ye T, Hyde TM, Guitart-Navarro X, Tao R, Wininger M *et al.* DNA methylation signatures in development and aging of the human prefrontal cortex. *Am J Hum Genet* 2012; **90**(2)**:** 260-272.

368. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* 2013; **153**(3)**:** 707-720.

369. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 2007; **3**(9)**:** 1724-1735.

370. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012; **28**(6)**:** 882-883.

371. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am J Hum Genet* 2010; **86**(3)**:** 411-419.

372. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 2010; **6**(5)**:** e1000952.

373. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 2014; **24**(1)**:** 14-24.

374. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**(6)**:** e1000529.

375. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* 2011; **1**(6)**:** 457-470.

376. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**(7)**:** 499-511.

377. Gamazon ER, Badner JA, Cheng L, Zhang C, Zhang D, Cox NJ *et al.* Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry* 2013; **18**(3)**:** 340-346.

378. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research* 2012; **40**(Database issue)**:** D930-934.

379. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO *et al.* A global reference for human genetic variation. *Nature* 2015; **526**(7571)**:** 68-74.

380. Lee AM, Miksys S, Palmour R, Tyndale RF. CYP2B6 is expressed in African Green monkey brain and is induced by chronic nicotine treatment. *Neuropharmacology* 2006; **50**(4)**:** 441-450.

381. Lee KW, Pausova Z. Cigarette smoking and DNA methylation. *Frontiers in genetics* 2013; **4:** 132.

382. Mair P, Schoenbrodt F, Wilcox R. WRS2: Wilcox robust estimation and testing. 2015.

383. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research* 2014; **42**(5)**:** 2976-2987.

384. Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* 2014; **30**(12)**:** i185-194.

385. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wagele B *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 2011; **477**(7362)**:** 54-60.

386. Zou F, Chai HS, Younkin CS, Allen M, Crook J, Pankratz VS *et al.* Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS genetics* 2012; **8**(6)**:** e1002707.

387. Clark SL, McClay JL, Adkins DE, Aberg KA, Kumar G, Nerella S *et al.* Deep Sequencing of Three Loci Implicated in Large-Scale Genome-Wide Association Study Smoking Meta-Analyses. *Nicotine Tob Res* 2015.

388. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014; **159**(7)**:** 1665-1680.

389. Miksys S, Lerman C, Shields PG, Mash DC, Tyndale RF. Smoking, alcoholism and genetic polymorphisms alter CYP2B6 levels in human brain. *Neuropharmacology* 2003; **45**(1)**:** 122-132.

390. Bloom AJ, Martinez M, Chen LS, Bierut LJ, Murphy SE, Goate A. CYP2B6 non-coding variation associated with smoking cessation is also associated with differences in allelic expression, splicing, and nicotine metabolism independent of common amino-acid changes. *PLoS One* 2013; **8**(11)**:** e79700.

391. Hofmann MH, Blievernicht JK, Klein K, Saussele T, Schaeffeler E, Schwab M *et al.* Aberrant splicing caused by single nucleotide polymorphism c.516G>T [Q172H], a marker of CYP2B6*6, is responsible for decreased expression and activity of CYP2B6 in liver. *The Journal of pharmacology and experimental therapeutics* 2008; **325**(1)**:** 284-292.

392. Tanner JA, Chenoweth MJ, Tyndale RF. Pharmacogenetics of nicotine and associated smoking behaviors. *Current topics in behavioral neurosciences* 2015; **23:** 37-86.

393. Garcia KL, Coen K, Miksys S, Le AD, Tyndale RF. Effect of Brain CYP2B Inhibition on Brain Nicotine Levels and Nicotine Self-Administration. *Neuropsychopharmacology* 2015; **40**(8)**:** 1910-1918.

394. Lee AM, Miksys S, Tyndale RF. Phenobarbital increases monkey in vivo nicotine disposition and induces liver and brain CYP2B6 protein. *British journal of pharmacology* 2006; **148**(6)**:** 786-794.

395. Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, Stefansson K. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet* 2011; **7**(2)**:** e1001317.

396. Yalachkov Y, Kaiser J, Naumer MJ. Brain regions related to tool use and action knowledge reflect nicotine dependence. *J Neurosci* 2009; **29**(15)**:** 4922-4929.

397. Wilson SJ, Sayette MA, Fiez JA. Prefrontal responses to drug cues: a neurocognitive analysis. *Nature neuroscience* 2004; **7**(3)**:** 211-214.

398. Goldstein RZ, Volkow ND. Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nature reviews Neuroscience* 2011; **12**(11)**:** 652-669.

399. Wise RA. Neurobiology of addiction. *Current opinion in neurobiology* 1996; **6**(2)**:** 243-251.

400. Kozlenkov A, Wang M, Roussos P, Rudchenko S, Barbu M, Bibikova M *et al.* Substantial DNA methylation differences between two major neuronal subtypes in human brain. *Nucleic acids research* 2015.

401. Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE *et al.* The PsychENCODE project. *Nature neuroscience* 2015; **18**(12)**:** 1707-1712.

402. Ziedonis D, Hitsman B, Beckham JC, Zvolensky M, Adler LE, Audrain-McGovern J *et al.* Tobacco use and cessation in psychiatric disorders: National Institute of Mental Health report. *Nicotine Tob Res* 2008; **10**(12)**:** 1691-1715.