Genetic risk factors for alcohol and nicotine behaviors

Whitney Elizabeth Melroy-Greif

University of Colorado at Boulder, wmelroy827@gmail.com

Follow this and additional works at: https://scholar.colorado.edu/iphy_gradetds

Part of the Genetics Commons

Recommended Citation
https://scholar.colorado.edu/iphy_gradetds/40
Genetic risk factors for alcohol and nicotine behaviors

By

Whitney Elizabeth Melroy-Greif

B.S. California Polytechnic State University San Luis Obispo, 2009

A dissertation submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
Department of Integrative Physiology
2015
This thesis entitled: “Genetic risk factors for alcohol and nicotine behaviors”
Written by Whitney Elizabeth Melroy-Greif
has been approved for the Department of Integrative Physiology

________________________________________
(Dr. Marissa A. Ehringer, Ph.D.)

________________________________________
(Dr. Jerry A. Stitzel, Ph.D.)

________________________________________
(Dr. Matthew B. McQueen, Sc.D.)

________________________________________
(Dr. Michael C. Stallings, Ph.D.)

________________________________________
(Dr. John E. Hokanson, M.P.H., Ph.D.)

Date:_______

The final copy of this thesis has been examined by the signatories, and we find that both the content and form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB protocol # 1008.34 and 14-0420
Alcohol and tobacco use are highly heritable and widespread problems. There is a substantial genetic correlation between alcohol dependence (AD) and nicotine dependence (ND), suggesting the same genetic factors explain much of the variation in these disorders. The purpose of these studies was to investigate aspects of genetic influence on alcohol and nicotine behaviors in human populations.

Using single nucleotide polymorphisms (SNPs), two candidate gene studies were performed to assess associations with alcohol abuse and dependence (AAD), alcohol consumption (AC), and conduct disorder (CD). The first study involved a replication of SNPs in GABRA2; although one SNP was associated with AAD and CD, previous findings with another well-studied SNP were not replicated. In the second study, GRM7 was investigated. Both the single SNP analyses as well as the gene-based test employed failed to find an association between common variants in GRM7 and AC.

In a separate series of analyses, gene-based tests were utilized to test for association with drug behaviors in a manually-curated set of genes. The first series of analyses focused on ND and cigarettes per day (CPD) in European Americans (EAs) and African Americans (AAs). Although different genes were identified for CPD and ND for each ethnic group, genes associated with CPD in EAs tended to be associated with ND in AAs, and vice versa. The second series of analyses focused on AD in EAs. Although many of the genes associated with AD in EAs were different from those associated with nicotine behaviors in EAs, one gene, DNAJA3, was associated with both drug behaviors in EAs.

Targeted sequencing of genes showing preliminary association with AD or ND was performed in a small sample. Rare variants identified in each gene were collapsed into sets to
test for association. One set in FGFR2 was associated with AD after correction for multiple testing.

In conclusion, SNP associations, GWAS, and sequencing studies can be utilized to investigate genetic risk factors for AD and ND and suggest future avenues for investigation. These studies demonstrate the importance of large samples and using multiple analytical approaches in order to weigh the reproducibility of each study.
ACKNOWLEDGEMENTS

I would like to thank my graduate advisor, Dr. Marissa Ehringer, for her exceptional guidance and support. I am extremely fortunate to have been able to work with such a successful woman in so many ways. I would also like to thank the rest of my committee for their assistance and help in various aspects of my training.

I would like to thank the past and present members of the Genetics of Substance Abuse Laboratory: Drs. Holly Stephens, Helen Kamens, Xavi Gallego, Matthew Powers, and Todd Darlington for being understanding friends and mentors that I could ask for help from; Kristen Burrows for always listening; and Jill Miyamato for taking care of day-to-day activities. Finally, thanks to the other graduate students I have worked with and everyone at the Institute for Behavioral Genetics; research is a collaborative effort and many people at the Institute are indispensible for the work they do.

Finally, I am utterly indebted to my friends and family for their constant support and encouragement: my parents and in-laws for their relentless belief in me; my sister for pushing me to be the best role model I can be; Meghan, Lindsay, and Lauren for always being there to talk and laugh at my jokes; and Eric, for being an amazing and understanding partner who can always see the best in a situation and make me laugh no matter the circumstances.
CONTENTS

CHAPTER I

I. INTRODUCTION.................................................................1

Overview .................................................................................1
Specific aims............................................................................1
Background and significance..................................................2

Alcohol and tobacco use.........................................................2

Evidence for shared and unique genetic influences between alcohol and nicotine use ..................................................3

Genetic variation in the human genome.................................4

Methodology for studying human genetic variation.................6

Association studies.................................................................6
Candidate gene studies............................................................7
Genome-wide association studies...........................................7
Sequencing..............................................................................8
Problems with human genetic association studies ..................10

Physiology of nicotinic acetylcholine receptors .....................10

Nicotinic acetylcholine receptors...........................................10

Muscle nicotinic receptors......................................................11

Neuronal nicotinic receptors..................................................11

Upregulation of nAChRs........................................................13

nAChR subunit mRNA is unchanged in response to nicotine.....16

Proposed mechanisms of nAChR upregulation.......................17

Nicotinic acetylcholine receptors and human drug use............23

CHAPTER II

II. EXAMINATION OF GENETIC VARIATION IN GABRA2 WITH CONDUCT DISORDER AND ALCOHOL ABUSE AND DEPENDENCE IN A LONGITUDINAL STUDY........28
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>28</td>
</tr>
<tr>
<td>Materials and methods</td>
<td>30</td>
</tr>
<tr>
<td>Samples</td>
<td>30</td>
</tr>
<tr>
<td>Colorado Center on Antisocial Drug Dependence</td>
<td>31</td>
</tr>
<tr>
<td>Genetics of Antisocial Drug Dependence</td>
<td>31</td>
</tr>
<tr>
<td>SNP selection</td>
<td>32</td>
</tr>
<tr>
<td>Genotyping</td>
<td>33</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>33</td>
</tr>
<tr>
<td>Data descriptive</td>
<td>33</td>
</tr>
<tr>
<td>Analysis of the CADD sample</td>
<td>34</td>
</tr>
<tr>
<td>Analysis of the GADD sample</td>
<td>35</td>
</tr>
<tr>
<td>Combined analysis of the CADD and GADD samples</td>
<td>37</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>37</td>
</tr>
<tr>
<td>Correction for multiple testing</td>
<td>37</td>
</tr>
<tr>
<td>Results</td>
<td>37</td>
</tr>
<tr>
<td>Discussion</td>
<td>43</td>
</tr>
</tbody>
</table>

**CHAPTER III**

**III. NO EVIDENCE FOR ASSOCIATION OF COMMON VARIANTS IN GRM7 WITH ALCOHOL CONSUMPTION**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>49</td>
</tr>
<tr>
<td>Materials and methods</td>
<td>51</td>
</tr>
<tr>
<td>Samples</td>
<td>51</td>
</tr>
<tr>
<td>SNP selection</td>
<td>53</td>
</tr>
<tr>
<td>Genotyping</td>
<td>54</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>55</td>
</tr>
<tr>
<td>Phenotypic descriptive</td>
<td>55</td>
</tr>
</tbody>
</table>
Statistical genetics analysis.................................................................56
Results.................................................................................................57
SNP analysis.........................................................................................57
Gene-based test.....................................................................................60
Discussion............................................................................................62

CHAPTER IV

IV. EXAMINATION OF THE INVOLVEMENT OF CHOLINERGIC-ASSOCIATED GENES IN NICOTINE BEHAVIORS IN EUROPEAN AND AFRICAN AMERICANS..........67

Introduction..........................................................................................67
Materials and methods........................................................................70
Samples.................................................................................................70
Analyses.................................................................................................70
Meta-analysis.........................................................................................71
Comparison with previous meta-analyses.............................................72
Results.....................................................................................................72
Cigarettes per day and nicotine dependence in EAs.............................73
Cigarettes per day and nicotine dependence in AAs.............................76
Discussion............................................................................................78

CHAPTER V

V. EXAMINATION OF THE INVOLVEMENT OF CHOLINERGIC-ASSOCIATED GENES IN ALCOHOL DEPENDENCE IN EUROPEAN AMERICANS...........................86

Introduction..........................................................................................86
Materials and methods........................................................................89
Samples.................................................................................................89
SNP mapping.........................................................................................90
Statistical analysis................................................................................90
Phenotypic descriptive.........................................................................90
Statistical genetics analysis ................................................................. 90
Post hoc SNP comparisons ................................................................... 91
Results ................................................................................................... 92
Discussion ............................................................................................ 95

CHAPTER VI

VI. TARGETED GENE SEQUENCING ................................................... 99

Introduction ......................................................................................... 99
Materials and methods ........................................................................ 100
Samples ............................................................................................... 100
Sequencing ......................................................................................... 101
Mapping, variant detection, and annotation ....................................... 102
Statistical analysis ............................................................................. 102
  Phenotypes ....................................................................................... 102
    GRM7 ............................................................................................. 103
    FGFR2 ........................................................................................... 103
    CAMK2A and FYN .......................................................................... 103
  Association analysis ......................................................................... 104
Results .................................................................................................. 105
Discussion ........................................................................................... 114

CHAPTER VII

VII. CONCLUSION .................................................................................. 118

REFERENCES ..................................................................................... 121
LIST OF TABLES

CHAPTER I
Table 1. Association of nAChR markers with phenotypes for human drug use..................................23

CHAPTER II
Table 1. Genotypic characteristics of the CADD and GADD samples for each ethnic group......39
Table 2. Results from the FBAT analysis in the CADD.................................................................42
Table 3. Results from the FBAT analysis using subjects in the GADD.........................................42
Table 4. Results from the FBAT analysis using subjects in the CADD and GADD combined sample.................................................................................................................................43

CHAPTER III
Table 1. Demographic characteristics of the CADD and GADD samples.................................58
Table 2. Genotypic characteristics of EA and Hispanic subjects in the CADD and GADD samples................................................................................................................................................58
Table 3. Results from the FBAT and QTDT analyses in the CADD..............................................59
Table 4. Results from the FBAT and QTDT analyses in the GADD .............................................60
Table 5. Phenotypic characteristics of the samples utilized in the gene-based test.................60
Table 6. Gene-based test results from the JAG analysis..............................................................61
Table 7. Gene-based test results from the PLINK analysis.........................................................61

CHAPTER IV
Table 1. Study and sample characteristics of the GWAS utilized.................................................73
Table 2. Results from the JAG and PLINK tests with CPD in EAs.............................................74
Table 3. Results from the JAG and PLINK tests with FTND in EAs............................................76
Table 4. Results from the JAG and PLINK tests with CPD in AAs.............................................77
Table 5. Results from the JAG and PLINK tests with FTND in AAS............................................78
CHAPTER V

Table 1. Polymorphisms in the CHRN genes associated with alcohol behaviors..........................87
Table 2. Phenotypic characteristics for each sample......................................................................92
Table 3. Results from the JAG and PLINK analysis in each study..................................................92

CHAPTER VI

Table 1. Preliminary results showing association with FGFR2 and AD using the PLINK set test in CADD-GADD...........................................................................................................100
Table 2. Target regions for each gene.............................................................................................102
Table 3. Phenotypic descriptive for the subjects in each sequencing analysis.................................105
Table 4. Results for each set from the SKAT-O test in GRM7.........................................................107
Table 5. Results for each set from the SKAT-O test in FGFR2.......................................................111
Table 6. Results for the CAMK2A and FYN sets from the SKAT-O analysis.................................114
LIST OF FIGURES

CHAPTER II

Figure 1. Study sample characteristics of the CADD and GADD.................................................. 38

Figure 2. Linkage Disequilibrium (LD) plot generated by Haploview (Barrett et al., 2005) for the SNPs chosen in the CADD sample.................................................................................................................. 40

Figure 3. Linkage Disequilibrium (LD) plot generated by Haploview (Barrett et al., 2005) for the SNPs chosen in the GADD sample.................................................................................................................. 41

CHAPTER IV

Figure 1. Cellular processes proposed to play a role in nAChR upregulation.............................. 69

CHAPTER V

Figure 1. Schematic of the SNPs associated with AD and their genomic location on the Human Genome Build hg19........................................................................................................................................... 94

CHAPTER VI

Figure 1. Map of variants identified in the sequenced region around GRM7................................. 106

Figure 2. Map of variants identified in the sequenced region around FGFR2................................. 110

Figure 3. Base pair position on the Human Genome Build hg19 of the variants in set 9 by –log_{10} of the p-value after association with AD................................................................................................. 112

Figure 4. Map of variants identified in the sequenced region around CAMK2A............................ 113

Figure 5. Map of variants identified in the sequenced region around FYN................................. 113
CHAPTER 1
INTRODUCTION

Overview

Tobacco and alcohol use are serious and worldwide problems. These disorders are heritable in human populations and evidence suggests they may share some common genetic liabilities. This doctoral thesis investigates the role of single nucleotide polymorphisms with alcohol and tobacco phenotypes through a variety of genetic methodologies: candidate gene studies, gene-based approaches using Genome-Wide Association Studies, and sequencing studies. Although a candidate gene in both the gamma-aminobutyric acid and glutamatergic systems will be examined, a large portion of this thesis will focus on genes implicated in the cholinergic system.

Specific aims

Aim One: Previous studies have shown associations between single nucleotide polymorphisms (SNPs) in gamma-aminobutyric acid (GABA) A receptor, alpha 2 (GABRA2) and adolescent conduct disorder and alcohol dependence in adulthood, but not adolescent alcohol dependence. Aim one was intended as a replication and extension of this work, focusing on adolescent conduct disorder, adolescent alcohol abuse and dependence, and adult alcohol abuse and dependence.

Aim Two: Animal studies have provided evidence for the role of glutamate receptor, metabotropic 7 (GRM7) in alcohol behaviors; aim two served to investigate the role of common SNPs in GRM7 with human alcohol behaviors. First, we genotyped 2 SNPs in GRM7 and tested for association with alcohol consumption in our two in-house samples. Second, using four Genome-Wide Association Studies (GWAS) we performed a gene-based test to assess the hypothesis that variation throughout the gene is involved in alcohol behaviors.
Aim Three: Examination of a manually curated set of genes believed to play a role in the development nicotine dependence. Two gene-based approaches were utilized to test these genes for association with cigarettes per day and nicotine dependence in European and African American populations using various GWAS acquired through dbGaP and collaborations.

Aim Four: Based on high comorbidity rates between alcohol and nicotine use, the gene set established to test for nicotine behaviors was tested for association with alcohol dependence in European Americans.

Aim Five: Several genes showing preliminary evidence for association with alcohol or nicotine behaviors were sequenced as a part of a larger study. In the last part of this dissertation the sequencing and analysis of these genes will be discussed.

Background and significance

Alcohol and tobacco use

Alcohol use is a severe and widespread problem. Worldwide, 3.3 million people die each year due to the harmful use of alcohol, representing 5.9% of worldwide deaths. Furthermore, 5.1% of the global burden of disease and injury is attributed to alcohol and recent causal relationships have been established between harmful drinking and occurrence of infectious diseases such as tuberculosis and HIV/AIDS (WHO).

Tobacco is similarly debilitating; as of May 2012, tobacco was estimated to kill up to half of its users. Of the nearly 6 million people that die per year due to tobacco smoke, 5 million are either users or ex-users, and 600,000 are non-smokers exposed to second hand smoke (WHO). In the U.S. alone, 1 in 5 deaths are attributable to smoking (CDC), and an additional 6.8 million people suffer from a serious illness caused by smoking (CDC). Unless the rate of smoking declines, the annual death toll as a consequence of smoking will exceed roughly 8 million by 2030 (WHO).
Evidence for shared and unique genetic influences between alcohol and nicotine use

Many researchers in the field of behavior genetics assess the heritability of complex, or non-mendelian, traits. The focus of these studies is monozygotic and dizygotic (MZ and DZ, respectively) twins. MZ, or identical, twins are from the same zygote and are genetically identical. DZ, or fraternal, twins are from 2 separate zygotes that have been separately fertilized and, on average, share 50% of their genetic variation. Although studies utilizing twins have several limitations, including the equal environments assumption, the possible lack of generalizability to the population, and inability to account for epistasis, they are very useful in determining if a trait is genetically influenced, as well as how much genetic influence a trait has.

Behavior geneticists use genetic model fitting to compare correlations between MZ and DZ twins. Essentially, a model that fits the raw data is established that determines the genetic and environmental components of a trait. In particular, these models can address the question of whether genetic effects are acting on a trait and, if so, the magnitude of these effects (Eaves et al., 1997). Although model fitting can reveal much about the nature of the data, the essential point to remember in the context of this dissertation is if a trait has genetic influences, trait correlations between MZs should be more similar than those between DZs. In this manner we can estimate the heritability for a trait of interest.

The proportion of phenotypic variance that can be accounted for by genetic differences between individuals is referred to as heritability and can be divided into two categories: broad sense heritability, and narrow sense heritability. Broad sense heritability refers to the proportion of phenotypic variance that can be accounted for by all the genetic influences between individuals. Narrow sense heritability refers to the proportion of phenotypic variance that can be accounted for by additive genetic influences between individuals, and is what is measured in behavior genetics. Although we will be focusing specifically on genetic influences on a trait, it is important to remember that finding genetic influence on a complex trait does not mean the
environment is not important; the field of behavior genetics can also estimate the magnitude of environmental influences on traits.

Behavior genetics has clearly demonstrated that alcohol and nicotine behaviors are both highly heritable and genetically correlated. Twin studies have estimated that nicotine dependence in men and women is 56-72% heritable (Haberstick et al., 2007; Kendler et al., 1999; Lessov et al., 2004; Maes et al., 2004; True et al., 1999). Similarly, alcohol dependence is estimated to be 49-65% heritable (Heath et al., 1997; Prescott et al., 1999; True et al., 1996; True et al., 1999). Furthermore, there is a substantial genetic correlation between alcohol and tobacco problem use in adolescence (Young et al., 2006) and regular use in young adults (Koopmans et al., 1997), as well as lifetime nicotine and alcohol dependence in adulthood (True et al., 1999). A recent study found that 56% and 51% of the variation in alcohol and nicotine dependence, respectively, was driven by additive genetic factors and that overlapping genetic factors between the two behaviors account for 41% of the variation (Vrieze et al., 2013). In other words, the same genetic factors can explain much of the variation in alcohol and nicotine dependence.

Although nicotine and alcohol dependence share underlying genetic factors, there is evidence that each behavior is also influenced to some extent by unique genetic influences (Hicks et al., 2011; Palmer et al., 2012). This indicates that while there may be some genetic variation that puts individuals at risk for both alcohol and nicotine dependence, there may exist some genetic variation that is specific to each disorder.

**Genetic variation in the human genome**

In 2001, the human genome was fully sequenced by two groups with different methods, both of which came to the same primary conclusion. Although the human genome is comprised of 3.2 billion base pairs, or nucleotide pairs, only about 1.5% of the genome codes for genes, corresponding to roughly 30,000 protein-coding genes. Within each gene are introns, non-
coding regions, as well as coding regions, or exons. There are two primary types of additive variation in the human genome: structural variants (SVs), such as insertions, deletions, and copy number variants; and single nucleotide polymorphisms (SNPs), or a mutation at a single base pair. At the time the human genome was initially sequenced, 1.4 million SNPs were identified (Lander et al., 2001; Venter et al., 2001). Although only complex disorders will be discussed in this dissertation, additive genetic variation contributes to both complex and Mendelian disorders. There are also non-additive sources of genetic variation in the genome. These include epistasis, or the interaction between two or more genes, as well as epigenetics, or heritable changes not due to change in the DNA sequence but by altering gene expression.

In 2012, 1,092 subjects from 14 populations were sequenced as part of the 1,000 Genomes Project. With the addition of subjects from different ethnic backgrounds and a large sample size, 38 million SNPs were identified, including rare variants with a frequency of 1%, along with 1.4 million short insertions and deletions, and more than 14,000 large deletions (Abecasis et al., 2012).

The body of this dissertation will focus on SNPs. Although there is more genetic variation in the genome itself, including SVs, SNPs are easily genotyped and much of the data on SNP genotyping has been made available to the public. SNPs in coding regions can either modify the amino acid sequence of the protein or not, and are labeled non-synonymous or synonymous, respectively. Although SNPs in coding regions are likely targets for association studies, only an estimated 1.5% of the genome is protein coding (Lander et al., 2001). SNPs are commonly found in introns as well as between genes (referred to as intergenic regions). Although often overlooked in studies, these regions can be important in the regulation of gene expression and bear consideration for risk in complex disorders. Throughout this dissertation, the major allele of the SNP will represent the allele that is present in over 50% of the sample, and the minor allele of the SNP will represent the allele that is less common in the sample.
**Methodology for studying human genetic variation**

**Association studies**

This dissertation will discuss various methods for performing an association test between a variant and phenotype of interest. Association studies can be performed using case/control data (in which subjects are unrelated), or family-based data. With case/control data, the allele frequencies between cases and controls are compared in order to assess if the frequency of an allele is more common in cases compared to controls; if so, this indicates association. Association studies can also be performed using quantitative phenotypes, most commonly using a regression-based approach to test whether or not the SNP is associated with the phenotype of interest. In family-based studies, the observed number of alleles transmitted to the offspring is compared with those expected in Mendelian transmissions; association is assumed if the number of alleles transmitted to affected offspring is greater than would be expected by chance (Laird *et al.*, 2006).

Although both case/control and family-based approaches have similar power to detect association, there are several strengths and weaknesses associated with each approach. First, family-based studies are robust to population stratification (the difference in allele frequencies in distinct populations as a result of different ancestry) but because they examine allele transmission within, and not between, a family, it is often difficult to ascertain family members to participate and involves extra costs and genotyping. Case/control studies are not robust to population stratification, but often require less resources than family-based studies (Laird *et al.*, 2006).

The approaches discussed above, case/control and family-based studies, are both utilized in this dissertation. The methods used to test genetic variation in the human genome for association with complex traits are discussed below: candidate gene studies, genome-wide association studies, and sequencing.
Candidate gene studies

Candidate gene studies seek to study the genetic influences on a complex trait based on a priori knowledge. Briefly, the methodology for these studies is: generate a hypothesis and identify genes that might have a role in the aetiology of the disease, identify the hypothesized causal variant (e.g. one that causes a change in the protein or expression of the gene) or a variant in linkage disequilibrium (LD, the non-random association of alleles at a particular locus) with the causal variant, genotype the targeted variant, and finally use statistical methods to test whether there is an association between the variant and the disease. Two main criticisms of these studies are that significant findings may not be replicated (possibly due to small sample sizes and low power), and that because they are based on the ability to predict functional genes and variants, current knowledge may be insufficient to pick a candidate gene for any particular disease. Many critics argue that agnostic approaches (like the one described in the next section) are more likely to yield results than hypothesis-driven approaches like candidate gene studies (Tabor et al., 2002).

Genome-wide association studies

While single candidate gene studies certainly have their merit, the notion that multiple gene effects lead to quantitative traits was developed in the early 20th century (Fisher, 1918; Wright, 1921). Accordingly, in the early 21st century Genome-Wide Association Studies (GWAS) were developed. These studies assay anywhere from several hundred to a million SNPs in any number of individuals and perform an association test between each SNP and the phenotype of interest. Although these tests are referred to as “Genome-Wide,” they rely on the concept of LD. There are an estimated 10 million common SNPs, or those with minor allele frequency greater than 5%, but due to LD in the genome we can capture much of the genetic variation of these common SNPs by genotyping roughly 1 million SNPs (Gabriel et al., 2002).
Although early GWAS were generally considered a “failure” in that few SNPs reached significance after correction for multiple testing, these studies were greatly underpowered and subsequent GWAS with ample sample size have been able to produce results. A large problem with the early GWAS was lack of power; in order to identify the effect size expected for a given SNP, a much larger sample size was needed (Wang et al., 2005). One possible solution to this problem was to use multiple GWAS, or do a meta-analysis of several studies, in order to achieve a large sample size. This was made possible by the National Center for Biotechnology Information’s database of Genotypes and Phenotypes (dbGaP), developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype (Mailman et al., 2007).

Although GWAS studies have implicated many SNPs of small effect in multiple genes that confer risk for complex traits (Manolio, 2010), there is still an issue of what is termed the “missing heritability” (Manolio et al., 2009). In short, although twin studies have estimated that nicotine and alcohol dependence are approximately 56-72% (Haberstick et al., 2007; Kendler et al., 1999; Lessov et al., 2004; Maes et al., 2004; True et al., 1999) and 49-65% heritable (Heath et al., 1997; Prescott et al., 1999; True et al., 1996; True et al., 1999), respectively, all variants tagged on GWAS chips only account for 36% and 21% of the variation in nicotine and alcohol behaviors, respectively (Vrieze et al., 2013). There are many explanations for the missing heritability: SVs not captured on current arrays, low power to detect gene-gene interactions, and rare variants that are poorly detected by GWAS (Manolio et al., 2009). It is also possible that heritability is over-estimated by twin studies.

**Sequencing**

As mentioned above, one possibility to account for the “missing heritability” in complex traits is the presence of rare variants that are not tagged in GWAS. Thus, researchers began sequencing regions of the genome in order to identify and test rare variants for association with
complex phenotypes. Although there are many types of sequencing in the field of genetics, including sequencing for DNA methylation and RNA sequencing, the focus on this dissertation is DNA. The following paragraphs will briefly cover methods for DNA sequencing and the selection of regions for DNA sequencing.

Since the development of Sanger sequencing in 1974, there have been rapid increases in DNA sequencing technology, leading us to what we refer to as “next-generation” sequencing. In 1974 Frederick Sanger developed Sanger sequencing, a chain-termination method in which there are four reactions into which only one of the four chain-terminating nucleotides is added. When the chain-terminating nucleotide binds to the elongating DNA strand, the strand is terminated and can later be run on a gel to detect bands of DNA and infer which nucleotide resides in each position (Sanger et al., 1977). Although very accurate, this method is time consuming and would limit how much DNA can be sequenced. Recently, high throughput next-generation technologies have been developed. While there are many high throughput sequencing platforms, one of the most frequently used is illumina. This platform uses reversible terminator chemistry in which all four fluorescently labeled nucleotides are included in the reaction. These labeled nucleotides will be incorporated into a growing DNA strand and terminate replication; the fluorescent label is then detected and a cleavage site cuts the fluorescent label from the nucleotide, unblocking replication of the DNA strand so the next fluorescently labeled nucleotide can bind (Bentley et al., 2008).

With the advent of high throughput sequencing technologies, researchers have been able to sequence large regions of the genome relatively quickly. However, many scientists choose to target a specific region of the genome to sequence. There are pros and cons to each approach. When considering whole genome sequencing, the goal is often to find genes associated with a phenotype of interest; this can be costly when it comes to how many reads you get at each location (how many times each smaller DNA sequence aligns at a particular locus), how you prioritize variants to test, and multiple testing correction. Whole exome
sequencing is a popular alternative, in which just the exons of the genome are sequenced. This approach lends itself to easy prioritization of the variants to test, and is less costly than whole genome sequencing. However, variants not in exons are excluded. It is highly likely that variants outside of exons, such as those in introns or those either up- or downstream of genes contribute to complex diseases (Schork et al., 2013). Candidate gene sequencing allows the sequencing of candidate genes and can incorporate regions up- and downstream of the gene. These studies are less costly and often allow a larger number of reads to be generated at each locus and are particularly useful when the hypothesis involves a particular gene.

Problems with human genetic association studies

There are many limitations of human genetic association studies. A primary criticism is that many studies lack the power to detect the small effect we know SNPs have. One major limitation is lack of reliability; namely, that a candidate gene study or GWAS may find one SNP associated with a phenotype, but this result is not replicated with a different sample. This can be due to low power, variations in study design (e.g. differences in sample ascertainment or definition of the phenotype), or population stratification. Finally, selection of polymorphisms can play a role in that polymorphisms are assumed to be in LD with causal variants. However, LD is often different between populations; thus, chosen polymorphisms may or may not be in LD with the causal variant in different populations (Tabor et al., 2002).

Physiology of nicotinic acetylcholine receptors

Nicotinic acetylcholine receptors

Although there are many compounds in tobacco smoke, nicotine is considered to be the major addictive component of tobacco smoke (Gunby, 1988; Rose, 2006; Stolerman et al., 1995). Nicotine binds muscle and neuronal nicotinic acetylcholine receptors (nAChRs) in the peripheral and central nervous systems (Gotti et al., 2009). Specifically, the behavioral changes
associated with nicotine exposure are thought to be due to the ability of nicotine to cross the blood-brain barrier and bind neuronal nAChRs in the brain (Clarke, 1987). Nicotinic acetylcholine receptors are members of a family of ligand-gated ion channels (for a review see [Gotti et al., 2009]). They are pentameric receptors composed of various subunits ranging from alpha1-alpha10 (α1-α10) and beta1-beta4 (β1-β4) (discussed in more detail below) clustered around a central ion pore. Nicotine produces rewarding effects by interacting with nAChRs in the brain’s reward system, comprised of dopaminergic neurons that originate in the ventral tegmental area (VTA) which project to regions such as the nucleus accumbens (NAc), hippocampus, amygdala, and prefrontal cortex (PFC) (Changeux, 2010). Upon stimulation by nicotine, the five subunits undergo a conformational change causing the central pore to open, allowing extracellular ions to enter the cell. A large focus of the Ehringer lab is to evaluate how variations in the nAChR genes confer risk to alcohol and nicotine dependence; although the first two aims presented in this dissertation focus on genes outside the nAChRs, the last two aims presented pertain to these receptors and their characteristics and thus an introduction is merited.

Muscle nicotinic receptors

Muscle nAChRs all have the same pentameric stoichiometry of (α1)2β1δγ (fetal-type) or (α1)2 β1δε (adult-type) (Le Novere et al., 2002). Although primarily known to be expressed in muscle tissue, previous work has suggested that subunits of these receptors are also present in mammalian ciliary ganglia (Pugh et al., 1995), vestibular and cochlear hair cells (Scheffer et al., 2007), as well as brain regions such as the cortex, hippocampus, and cerebellum (Ghedini et al., 2010). In addition, a splice variant of the α1 subunit was shown to expressed in numerous locations including brain, kidney, heart, liver, lung, and thymus in humans (Talib et al., 1993).
Neuronal nAChRs are comprised of combinations of the α2-α10 and β2-β4 subunits with variable stoichiometry that are expressed in the central and peripheral nervous systems, located both presynaptically and postsynaptically (Millar et al., 2009). All of the subunits are expressed in the mammalian central nervous system with the exception of the α8 subunit, which is only expressed in chicks (Britto et al., 1992; Schoepfer et al., 1990). Neuronal nAChRs can assemble in two ways; they can form homomeric receptors in which all five subunits are the same, or they can form heteromeric receptors in which the five subunits that make up the functional receptor are different.

In the mammalian system, α7 and α9 subunits form homomeric receptors when expressed as isolated subunits in cells (Couturier et al., 1990; Elgoyhen et al., 1994; Schoepfer et al., 1990). The α10 subunit, although it cannot form a functional homomeric receptor, is most closely related to the α9 subunit and forms a functional receptor when coexpressed with the α9 subunit (Elgoyhen et al., 1994; Lustig et al., 2001; Sgard et al., 2002). In the central nervous system, the α9 and α10 subunits have been primarily identified in the inner ear cells, particularly the cochlear and vestibular hair cells (Elgoyhen et al., 1994; Elgoyhen et al., 2001; Luo et al., 1998; Lustig et al., 1999; Lustig et al., 2001; Vetter et al., 1999), as well as the pituitary (Sgard et al., 2002). The α7 subunit, as opposed to the α9 and α10 subunits, is widely expressed throughout the mammalian central nervous system (reviewed in [Millar et al., 2009]).

The rest of the neuronal nAChR subunits form heteromeric receptors consisting of α and β subunits (Anand et al., 1991; Boulter et al., 1987; Deneris et al., 1988; Wada et al., 1989). For heteromeric receptors, subunit composition varies, as does assembly; furthermore, the variable stoichiometry confers differences in calcium permeability, as well as agonist and antagonist sensitivity between different receptors (Luetje et al., 1991) (reviewed in [Millar et al., 2009]). Studies have shown that α2-α4 and β2 and β4 subunits can form functional receptors when expressed as a pair-wise combination of one α subunit and one β subunit; functional expression has been observed for α2β4, α3β2, α3β4, α4β2, and α4β4 combinations (Duvoisin et al., 1989;
Papke et al., 1989). The β3 and α5 subunits cannot form functional nAChRs unless expressed in combination with another α and β subunit pair. Electrophysiological approaches have been used to distinguish triple pair subunits from pair subunits and subsequently identified receptors composed of the α3β2α5 (Gerzanich et al., 1998; Wang et al., 1996), α3β4α5 (Fucile et al., 1997; Gerzanich et al., 1998; Wang et al., 1996), and α4β2α5 (Ramirez-Latorre et al., 1996) subunits. In addition, the β3 subunit has been shown to co-assemble into a functional receptor composed of α3β3β4 subunits in oocytes (Groot-Kormelink et al., 1998), and may promote expression and stability of human α6 subunits in transfected cell lines (Tumkosit et al., 2006). Expression studies have also demonstrated that the α6 subunit assembles into functional triplet receptors, including α6β3β4 (Kuryatov et al., 2000; Tumkosit et al., 2006), as well as α3α6β4 and α3α6β2 (Fucile et al., 1998; Kuryatov et al., 2000) subunit combinations. Lastly, α6 and β2 combinations have also been observed, particularly as α4α6β2β3 and α6β2β3 receptors (Salminen et al., 2004). While neuronal nAChRs are widespread throughout the central and peripheral nervous systems, α4β2 (Anand et al., 1991; Flores et al., 1992; Wada et al., 1989) receptors are the most abundant heteromeric receptors in the brain and have the highest affinity for nicotine (Wonnacott, 1990).

Upregulation of nAChRs

Nicotinic receptors have been referred to as a paradox (Wonnacott, 1990), as they do not behave as other receptors do under stimulation by an agonist or antagonist. Note that an agonist binds and activates a receptor whereas an antagonist binds and blocks receptor activation. Numerous studies have shown that chronic stimulation of a receptor by an antagonist causes upregulation, or an increased number of receptors, while stimulation of a receptor by an agonist causes downregulation, or a decreased number of receptors (Creese et al., 1981; Wonnacott, 1990). However, this phenomenon is not what you see with nAChRs in that stimulation of nAChRs by an agonist causes upregulation of receptor numbers. Of paramount
importance is that upregulation, for the purposes of this discussion, refers to an increase in receptor numbers, and the studies mentioned below have focused exclusively on measuring protein levels to assess the number of nAChRs.

Early studies on the brains from rats (Schwartz et al., 1983) and mice (Marks et al., 1983) repeatedly exposed to nicotine have shown that there is an increase in $[^3\text{H}]$acetylcholine and $(-)^[^3\text{H}]$Nicotine binding, two radiolabeled nAChR agonists. Schwartz and Kellar showed this in 1983 when they treated rats with 2 mg/kg of nicotine twice a day for 10 days. The result was an increase of $[^3\text{H}]$acetylcholine binding in cerebral cortex of rats (Schwartz et al., 1983).

Similarly, Marks et al. (1983) chronically infused mice with 5.0mg/kg/hr nicotine for 8-10 days (it is worth noting here that mice are less sensitive to the effects of nicotine than rats are (Matta et al., 2007), accounting for the difference in nicotine treatments in these two experiments) and reported an increase in $(-)^[^3\text{H}]$Nicotine binding in the cortex, midbrain, hindbrain, hippocampus, and hypothalamus, but not striatum (Marks et al., 1983). These results were confirmed in human post mortem brains in 1988; $(-)^[^3\text{H}]$Nicotine binding was measured in smokers and, compared to nonsmokers, smokers showed increased $(-)^[^3\text{H}]$Nicotine binding in most grey areas including the hippocampus, cortex, gyrus rectus, and median raphe nucleus, with the exception of the medulla oblongata which showed no changes in $(-)^[^3\text{H}]$Nicotine binding. Importantly, in the areas where smoking was associated with increased $(-)^[^3\text{H}]$Nicotine binding, radiolabeled ligand binding increased by 50-100%. The study examined 12 smokers who smoked 7-20 cigarettes per day in the two years prior to death and who had smoked within 48 hours of their death, and 18 nonsmokers who were either nonsmokers or smokers who quit smoking 5 years prior to death (Benwell et al., 1988). In all three of the above studies, increased radiolabeled ligand binding was confirmed to be due to an increase in receptor number, rather than an increase in binding affinity (Benwell et al., 1988; Marks et al., 1983; Schwartz et al., 1983). Thus, under chronic nicotine exposure, nAChRs undergo a nicotine-induced upregulation of receptor numbers.
Subsequent studies provided evidence that when nicotine exposure is ceased, decreased binding of (-)[³H]Nicotine occurs after 7-10 days in mice (Marks et al., 1985), 15-20 days in rats (Collins et al., 1990), and between 21 days and 2 months in humans (Breese et al., 1997; Mamede et al., 2007). In humans, nAChR upregulation was additionally shown to be dose-dependent (Breese et al., 1997; Perry et al., 1999). Gopalakrishnan et al. (1997) detected a significant increase in [³H]cytisine binding upon exposure to nicotine at concentrations as low as 100 nM, and a maximal 15-fold increase in binding at concentrations of 10 µM of nicotine the human embryonic kidney cell line HEK293 (Gopalakrishnan et al., 1997). This is particularly interesting because smokers have an average serum concentration of 100-200 nM nicotine after smoking (Benowitz, 1996; Henningfield et al., 1993), suggesting that nAChR upregulation could occur in smokers in between, as well as during, periods of smoking.

As expected, because at baseline human brains have different levels of nicotine binding in different regions of the brain (Court et al., 1995), the magnitude of increase in nicotine binding sites is not universal among all areas of the brain (Collins et al., 1989; Marks et al., 1992). Regions of the thalamus appear most resistant to nicotine-induced upregulation of nAChR numbers in rodent models (Kellar et al., 1989; Pauly et al., 1991). In humans, the medulla oblongata, as well as different layers of the cortex and hippocampus are resistant to, or undergo nAChR upregulation at variable rates (Benwell et al., 1988; Perry et al., 1999). As nAChR subunits have different expression levels in each region of the brain, this suggests that individual neuronal nicotinic receptor subtypes could undergo differential upregulation in different brain regions. In addition to upregulation of α4β2 subunit-containing receptors (Flores et al., 1992; Zhang et al., 1995), α3, albeit requiring a higher dose of nicotine, as well as α7, to a lesser extent, subunits also upregulate (Olale et al., 1997; Peng et al., 1997; Wang et al., 1996). Although the typical serum concentration of nicotine in smokers has little effect on muscle nAChRs (Lindstrom, 1996), upregulation of ganglionic α3β4 and α7 subunits, as well as the muscle type nAChR subunits α1, β1, γ, and δ, has been observed on the cell surface after a 48
hour exposure to nicotine at concentrations as low as 1µM (Ke et al., 1998). These findings confirm results from two previous studies on muscle nAChRs (Luther et al., 1989; Siegel et al., 1988), although earlier studies found that nAChRs in muscle cells are downregulated 40% by chronic exposure to nicotinic agonists (Appel et al., 1981; Gardner et al., 1979; Noble et al., 1978) (for a review see [Lindstrom, 1996]). Ke et al. (1998) also found that different concentrations of nicotine upregulate different receptor subunits to varying degrees; α3β4 subunits were more prone to upregulation than muscle nAChRs. This suggests that, while certain subunits are more susceptible than others to upregulation, predisposition to nicotine-induced upregulation is a characteristic of all nAChR subunits (Ke et al., 1998).

\textit{nAChR subunit mRNA is unchanged in response to nicotine}

Surprisingly, nicotinic receptor subunit mRNA is unchanged in response to nicotine. Early studies by Marks and colleagues showed that while there is an upregulation in nAChR protein levels, nAChR subunit mRNA remains steady and is likely not responsible for the upregulation of the nAChR subunits Chrna2, Chrna3, Chrna5, or Chrb4 in the mouse brain (Marks et al., 1992). These results were replicated in subsequent studies performed in mouse cell lines (Bencherif et al., 1995; Peng et al., 1994; Zhang et al., 1995), human cell lines (Peng et al., 1997) and rat brain (Bencherif et al., 1995). Ke et al. (1998) expanded the aforementioned work in cultured human cell lines and reported that changes in nAChR numbers are not attributable to changes in \textit{CHRNA1, CHRNA3, CHRNA5, CHRNA7, CHRNA2, CHRNA4, CHRN5, or CHRNA3} mRNA levels (Ke et al., 1998). Although there have been many studies on nAChR subunit mRNA in human brains, these studies have primarily focusing on nAChR expression changes in Parkinson’s and Alzheimer’s Diseases, as well as in aging brains, and to date only one study has been conducted to examine the correlation between nAChR subunit protein and mRNA in smokers. In 2003, Mousavi et al. found that protein levels of the α4 and α7 nAChR subunits were significantly increased in the temporal cortex of smokers compared to
those of nonsmokers; complementary to the results in rodents and cell lines, there were no differences in CHRNA4 and CHRNA7 mRNA in the temporal cortex of smokers as compared to those of nonsmokers (Mousavi et al., 2003). Thus, between studies in cell lines, as well as rodent and human brains, there are solid findings indicating that the nicotine-induced upregulation of nAChR numbers is independent of transcriptional events.

Proposed mechanisms of nAChR upregulation

Many studies have been conducted to explore the relationship between upregulation of nAChRs and nicotine exposure. Nicotine can pass through the plasma membrane of a cell and has been found in high concentration in cell organelles, as measured in the submaxillary glands of rats by Putney and Borzelleca in 1971 (Putney et al., 1971). Although nicotine can both bind to surface nAChRs and enter the cell to cause effects, studies by Peng et al. (1994) and Whiteaker et al. (1998) have indicated that nicotine acting on surface nAChRs, rather than intracellular nAChRs, leads to upregulation, although the increase in receptors is accredited to recruitment from a pre-existing pool of intracellular nAChRs. This conclusion was based on evidence that ligands unable to cross the cell membrane still produced an upregulation of nAChR numbers (Peng et al., 1994; Whiteaker et al., 1998). There are several theories as to how nAChR upregulation occurs discussed below, some of which are not necessarily mutually exclusive.

Early work by Bencherif et al. (1995) suggested that the increase in nAChR numbers comes from a pool of assembled receptors within the cell based on the observation that exposure to nicotine had no effect on the degradation of nAChRs (a decrease in degradation being expected should altered receptor turnover be the suspected mechanism for upregulation). This lead to the hypothesis that the increase in nicotine binding sites is due to a reservoir of nAChRs existing within the cell and upon exposure to nicotine, the intracellular nAChRs migrate to the surface of the cell. The observation that nAChR upregulation had an upward limit of 150%
suggests the intracellular nAChR reservoir is finite and that there are about 1.5 intracellular
nAChRs per surface nAChR. The authors hypothesized that the intracellular reservoir could
exist due to several different mechanisms and set the stage for subsequent research (Bencherif
et al., 1995).

One theory proposed by Gopalakrishnan et al. is that two second messenger pathways
influence nAChR upregulation, possibly through altering translation rates. The study used
HEK293 cells stably expressing human α4 and β2 subunits that lacked transcriptional regulatory
elements. Treatment with nicotine alone produced an upregulation of receptor number but
treatment with cyclohexamide, a protein synthesis inhibitor, failed to produce the expected
nicotine-induced upregulation of nAChR numbers. Treatment with the second messenger
analog dibutyl cyclic adenosine monophosphate (cAMP) enhanced the nicotine-induced
upregulation of nAChRs, while treatment with protein kinase C (PKC) inhibitors failed to produce
the nicotine-induced upregulation of nAChRs. This supports the theory that the nicotine-induced
nAChR upregulation, while independent of transcriptional events, may depend on translational
events, and suggests that the upregulation could be due to two second messenger pathways,
one involving PKC and one involving cAMP, either individually or in conjunction with one
another (Gopalakrishnan et al., 1997). A following study by Nashmi et al. (2003) confirmed this
work in HEK293 cells transfected with fluorescently labeled α4 and β2 subunits by showing
increased assembly of α4β2 nAChRs after the activation of PKC (Nashmi et al., 2003).

In 2005, Vallejo et al. proposed that nAChR upregulation is due to a change in
stoichiometry of receptors already present on the cell surface. Using a biotinylation assay they
determined that 20% of agonist binding sites are on the cell surface and 80% are intracellular
and furthermore that nicotine exposure did not alter the ratio of intracellular to surface receptors.
The cells were exposed to brefeldin A, a compound that interrupts metabolic processing in the
Golgi apparatus and thus trafficking to the cell surface, and showed that exposure to brefeldin A
and nicotine still resulted in an upregulation of nAChR numbers, suggesting that nAChR
upregulation is independent of intracellular receptor trafficking. The authors concluded that although the number of surface receptors is not changing, the stoichiometry of receptors changes from a resting state receptor that does not bind nicotinic agonists with high affinity to an active receptor that binds nicotinic agonists with high affinity under chronic nicotine exposure (Vallejo et al., 2005).

Several studies have provided evidence that upregulation of nAChRs is due to mechanisms that increase assembly of intracellular nAChRs and decrease the rate of turnover of surface nAChRs (Kuryatov et al., 2005; Peng et al., 1994; Wang et al., 1998). Work with a mouse cell line expressing α4 and β2 subunits lacking nAChR-specific promoters concluded that nicotine was still able to produce an upregulation of nAChR numbers. Furthermore, when the protein synthesis inhibitor cyclohexamide was expressed with nicotine, (-)-[^3H]Nicotine binding remained higher than control cells for a longer time, indicating that nAChR degradation was prolonged (Peng et al., 1994). This work was replicated with the α3 and β2 subunits, and additionally found a very small increase of β2 subunits within the cell despite the 5 fold increase in α3β2 receptors on the cell surface, suggesting that the rate of assembly of the complete nicotinic receptor increases with exposure to nicotine (Wang et al., 1998). Kuryatov et al. (2005) showed in a human cell line expressing α4 and β2 subunits that nicotine and nicotinic ligands (such as epibatidine and acetylcholine) could enter the lumen of the endoplasmic reticulum (ER), where a plethora of unassembled nAChR subunits reside, as shown by a Western blot. The authors also identified what was termed “assembly intermediates” that were likely either α4 or β2 subunits bound to chaperones and hypothesized that these nicotinic ligands act as chaperones, bind nAChR binding sites, and trigger maturation of subunits and receptor assembly (Kuryatov et al., 2005).

Another theory, proposed by Darsow et al. in 2005, is that nAChR upregulation is a consequence of increased nAChR exocytic trafficking to the cell surface. Although a stark contrast to the findings above (Kuryatov et al., 2005; Peng et al., 1994; Wang et al., 1998),
Darsow et al. hypothesized that transport through the secretory pathway from the ER is necessary for nAChR upregulation. After translation, membrane-bound proteins are transported through the secretory pathway to the plasma membrane; if trafficking through the secretory pathway is necessary for upregulation then disruption of this pathway should result in a lack of nicotine-induced nAChR upregulation. The authors disrupted the secretory pathway with brefeldin A, a compound that disrupts trafficking from the Golgi apparatus, and found that cells exposed to brefeldin A alone show a decrease in nAChR surface numbers and cells exposed to both brefeldin A and nicotine showed no upregulation of surface nAChRs. This result, however, was not seen when the authors measured intracellular nAChRs (Darsow et al., 2005), indicating that transport through the secretory pathway is necessary for upregulation of surface nAChRs but there is another mechanism responsible for increasing the intracellular concentration of nAChRs.

Similar to some of the work described above, several studies have proposed nAChR upregulation is a cause of increased subunit maturation and receptor assembly in the ER (Harkness et al., 2002; Nashmi et al., 2003; Sallette et al., 2005). Evidence for this theory has come from various studies using transfected α4 and β2 subunits; in one such study, the amount of total β2 subunit protein folded into correct conformation was influenced by both co-assembled partner subunits (the α4 subunit) and by chronic nicotine treatment. Because the amount of total α4 subunit protein folded into correct conformation was not affected by chronic nicotine treatment, this suggested that nAChR upregulation was not due to increase in total protein but how readily the subunits mature and fold in the ER (Harkness et al., 2002). Subsequent studies concluded that increased assembly of nAChRs occurs in the soma of neurons, rather than the dendritic processes, and nAChRs must be subsequently trafficked to the dendritic processes as dendritic processes have a larger overall concentration of nAChRs (Nashmi et al., 2003). Additional evidence supporting this theory showed that α4 and β2 subunits undergo complex oligosaccharide glycosylation, a modification known to only occur in the ER and the Golgi
apparatus, before they become heteromeric complexes and, furthermore, these glycosylations were found on all pentameric complexes, indicating that the receptor must be completely assembled into a 5-subunit pentameric complex in order to exit the ER and Golgi. Nicotine exposure was shown to increase the amount of complex oligosaccharides on nAChR subunits (Sallette et al., 2005). Taken together, these studies indicate that nAChR subunits undergo maturation and subsequently are assembled into pentamers in the ER of the cell soma and translocated to the cell surface, and this process could escalate upon exposure to nicotine (Harkness et al., 2002; Nashmi et al., 2003; Sallette et al., 2005).

Yet another theory poses that subunit degradation is blocked and subunits are consequently recycled to form new nicotinic receptors (Christianson et al., 2004; Ficklin et al., 2005; Rezvani et al., 2007). Treatment with proteasomal inhibitors has been shown to increase nAChR numbers and assembly in several cells lines (Christianson et al., 2004). Ubiquilin-1, a ubiquitin-like protein that can interact with both the proteasome and ubiquitin ligases in protein degradation, co-immunoprecipitated with the α3 subunit in human cell lines as well as neurons from the mouse superior cervical ganglion (SCG) and when ubiquilin-1 was transfected into human cell lines the number of surface nAChRs was reduced. These experiments were repeated in cultured SCG neurons endogenously expressing nAChRs; while neurons not exposed to nicotine and injected with a ubiquilin-1 lentivirus showed no changes in α3 subunit levels, neurons exposed to nicotine after injection with a ubiquilin-1 lentivirus showed a decrease in α3 subunit levels (Ficklin et al., 2005). Another study by Rezvani et al. in 2007 showed that nicotine itself can act as a partial proteasome inhibitor. In the PFC of C57BL/6J mice, nicotine causes an increase in ubiquitinated proteins, including the α7 subunit. The authors tested proteasomal activity of the proteasome and showed that treatment with nicotine lead to a decrease in proteasomal activity. Furthermore, this decrease in activity was not due to a decrease in expression of the proteasome, as nicotine was found to increase expression of proteasomal subunits. Thus, the decrease in activity of the proteasome must be due to an
interaction with nicotine (Rezvani et al., 2007). Subsequent work by the same group focused on UBXN2A, a protein containing an ubiquitin-like domain that binds to ubiquitin. This protein was also observed binding to α3 subunits using a yeast 2-hybrid screen and further studies showed that over expression of UBXN2A resulted in an increase in α3β2 nAChRs, as compared to normal expression of UBXN2A. In addition, expression of UBXN2A was found to decrease levels of α3 subunit ubiquitination (Rezvani et al., 2009).

While the above studies have supported various theories of nicotine-induced nAChR upregulation, many of them are not mutually exclusive. It is important to note that there are receptor-specific assembly folding factors that fold specific nAChR subunits present in some mammalian cells lines and not others (Sweileh et al., 2000), and different nAChR upregulation mechanisms may be impacted by cell line specific folding factors. Recently, work by Govind et al. (2012) has proposed that nAChR upregulation is not due to a single process as suggested by some of the above studies, but multiple processes occurring at different rates. By measuring the kinetics of upregulation, the authors surmised that upregulation occurred at two different rates, an initial fast component of upregulation that saturated after 4 hours, and a slower component that increased with continued nicotine exposure. Binding of conformation-dependent antibodies occurred at the rate of the fast component of upregulation, and immunoblot revealed that nAChR numbers were unchanged. This suggests that the fast component of upregulation results from nicotine-induced conformational changes (Govind et al., 2012), as seen by Vallejo et al. in 2005 (Vallejo et al., 2005). Higher numbers of nAChR subunits were observed during the slower component of upregulation, and proteasomal inhibition produced even higher levels of subunits in the cell. The authors concluded that the slower component of upregulation was due to decreased proteasomal subunit degradation, as previously seen (Rezvani et al., 2007), as well as increased subunit assembly, as earlier work had suggested (Harkness et al., 2002; Nashmi et al., 2003; Sallette et al., 2005). These results suggest that there are many
mechanisms, each acting at a different rate, that contribute to nAChR upregulation (Govind et al., 2012).

**Nicotinic acetylcholine receptors and human drug use**

There have been numerous associations between the CHRN genes and various drug phenotypes (Table 1).

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Gene(s)</th>
<th>Sample(s)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CHRNA4</td>
<td>Families from Anhui Province, China</td>
<td>(Feng et al., 2004)</td>
</tr>
<tr>
<td>AD&lt;sup&gt;b&lt;/sup&gt;</td>
<td>CHRNA4</td>
<td>Unrelated Korean males</td>
<td>(Kim et al., 2004)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA4</td>
<td>Subjects of European or African ancestry from southern US (TN, MS, or AK)</td>
<td>(Li et al., 2005)</td>
</tr>
<tr>
<td>Subjective responses to alcohol and tobacco, past 6 month use alcohol</td>
<td>CHRNA4, CHRN2&lt;sup&gt;2&lt;/sup&gt;</td>
<td>CADD&lt;sup&gt;c&lt;/sup&gt;</td>
<td>(Ehringer et al., 2007)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA3&lt;sup&gt;4&lt;/sup&gt;-B4</td>
<td>CADD, NYS&lt;sup&gt;e&lt;/sup&gt;</td>
<td>(Schlaepfer et al., 2008)</td>
</tr>
<tr>
<td>Age of initiation of alcohol and tobacco</td>
<td>CHRNA6-B3</td>
<td>CADD, Add Health&lt;sup&gt;f&lt;/sup&gt;</td>
<td>(Zeiger et al., 2008)</td>
</tr>
<tr>
<td>CPD&lt;sup&gt;g&lt;/sup&gt;</td>
<td>CHRNA5-A3</td>
<td>Three European populations</td>
<td>(Berrettini et al., 2008)</td>
</tr>
<tr>
<td>AD</td>
<td>CHRNA5-A3</td>
<td>COGA&lt;sup&gt;h&lt;/sup&gt;</td>
<td>(Wang et al., 2009)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA5-A3</td>
<td>&gt;11,000 Icelandic and European samples</td>
<td>(Thorgeirsson et al., 2008)</td>
</tr>
<tr>
<td>&quot;Pleasurable buzz&quot; during early experimentation with smoking</td>
<td>CHRNA5</td>
<td>American Caucasians and African Americans</td>
<td>(Sherva et al., 2008)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA5-A3-B4</td>
<td>COGA</td>
<td>(Bierut et al., 2008)</td>
</tr>
<tr>
<td>Age-dependent ND</td>
<td>CHRNA5-A3-B4</td>
<td>Three American cohorts</td>
<td>(Weiss et al., 2008)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA6-B3</td>
<td>NYS</td>
<td>(Hoft et al., 2009b)</td>
</tr>
<tr>
<td>Condition</td>
<td>CHRNA5-A3-B4, CHRNA6-B3, CHRND-G</td>
<td>COGEND</td>
<td>(Saccone et al., 2009a)</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------------</td>
<td>--------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Cocaine dependence</td>
<td>CHRNA5</td>
<td>FSCD, COGA</td>
<td>(Grucza et al., 2008)</td>
</tr>
<tr>
<td>Heavy smoking</td>
<td>CHRNA5-A3-B4</td>
<td>American Cancer Society CPS-II Cohort, and the CPS-II Nutrition Cohort</td>
<td>(Stevens et al., 2008)</td>
</tr>
<tr>
<td>Level of response to alcohol</td>
<td>CHRNA5-A3-B4</td>
<td>San Diego Sibling Pair</td>
<td>(Joslyn et al., 2008)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA5-A3-B4</td>
<td>Two Texas samples</td>
<td>(Spitz et al., 2008)</td>
</tr>
<tr>
<td>Smoking cessation</td>
<td>CHRNA5, CHRNA2</td>
<td>Pharmacogenetic trial of bupropion for smoking cessation</td>
<td>(Conti et al., 2008)</td>
</tr>
<tr>
<td>Smoking cessation</td>
<td>CHRNB2</td>
<td>European pregnant women</td>
<td>(Heitjan et al., 2008)</td>
</tr>
<tr>
<td>Smoking cessation during pregnancy</td>
<td>CHRNA5-A3-B4</td>
<td>Cancer Genetic Markers of Susceptibility</td>
<td>(Caporaso et al., 2009)</td>
</tr>
<tr>
<td>CPD</td>
<td>CHRNA5-A3-B4</td>
<td>NYS Family Study</td>
<td>(Hoft et al., 2009a)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA5-A3-B4</td>
<td>Utah and Wisconsin cohorts</td>
<td>(Baker et al., 2009)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA5-A3-B4</td>
<td>African Americans subjects from COGEND</td>
<td>(Saccone et al., 2009b)</td>
</tr>
<tr>
<td>ND, AD</td>
<td>CHRNA5-A3-B4</td>
<td>Virginia Adult Twin Study</td>
<td>(Chen et al., 2009b)</td>
</tr>
<tr>
<td>Heavy alcohol use</td>
<td>CHRNA6</td>
<td>Spanish population</td>
<td>(Landgren et al., 2009)</td>
</tr>
<tr>
<td>CPD</td>
<td>CHRNA3</td>
<td>Tobacco and Genetics Consortium, ENGAGE, Ox-GSK, GSK, ENGAGE</td>
<td>(Furberg et al., 2010)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNB3-A6, CHRNA4, CHRNB1, CHRNA10, CHRND-G</td>
<td>COGEND</td>
<td>(Saccone et al., 2010b)</td>
</tr>
<tr>
<td>CPD</td>
<td>CHRNA5-A3-B4, CHRNB3-A6</td>
<td>ENGAGE</td>
<td>(Thorgeirsson et al., 2010)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNB3-A6, CHRNA4, CHRNB1, CHRNA10, CHRND-G</td>
<td>COGEND</td>
<td>(Wessel et al., 2010)</td>
</tr>
<tr>
<td>ND</td>
<td>Neuronal CHRN genes</td>
<td>Group Health</td>
<td>(Li et al., 2010)</td>
</tr>
<tr>
<td>Smoking initiation, smoking cessation</td>
<td>CHRNA5-A3-B4</td>
<td>Korean subjects</td>
<td>(Ehringer et al., 2010)</td>
</tr>
<tr>
<td>Smoking quantity</td>
<td>CHRNA5-A3-B4</td>
<td>Os-GSK</td>
<td>(Liu et al., 2010)</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------</td>
<td>--------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Smoking quantity</td>
<td>CHRNA5-A3-B4</td>
<td>CGASP$^1$</td>
<td>(Saccone et al., 2010a)</td>
</tr>
<tr>
<td>ND, cocaine dependence, AD</td>
<td>CHRNA5-A3-B4</td>
<td>Five American cohorts (Uconn, Yale, MUSC, UPENN, McLean)</td>
<td>(Sherva et al., 2010)</td>
</tr>
<tr>
<td>Opioid dependence severity, ND severity</td>
<td>CHRNA5-A3-B4</td>
<td>Outpatients in treatment for opioid dependence</td>
<td>(Erlich et al., 2010)</td>
</tr>
<tr>
<td>Subjective response to nicotine</td>
<td>CHRNA2</td>
<td>Daily smokers from the Denver/Boulder area</td>
<td>(Hoft et al., 2011)</td>
</tr>
<tr>
<td>Smoking quantity, response to smoking cessation therapy</td>
<td>CHRNA5-A3-B4</td>
<td>Two studies, MT1 and MT2</td>
<td>(Sarginson et al., 2011)</td>
</tr>
<tr>
<td>CPD, ND</td>
<td>CHRNA4</td>
<td>Five American cohorts (Uconn, Yale, MUSC, UPENN, McLean)</td>
<td>(Han et al., 2011)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA5-A3</td>
<td>Virginia Twin Registry</td>
<td>(Maes et al., 2011)</td>
</tr>
<tr>
<td>Smoking quantity</td>
<td>CHRNA5-A3-B4</td>
<td>Three Italian populations</td>
<td>(Sorice et al., 2011)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA4</td>
<td>Five American cohorts (Uconn, Yale, MUSC, UPENN, McLean)</td>
<td>(Xie et al., 2011)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA5-A3-B4, CHRN5-A6</td>
<td>COGEND</td>
<td>(Haller et al., 2012)</td>
</tr>
<tr>
<td>ND, CPD</td>
<td>CHRNA5-B3</td>
<td>SAGE$^m$</td>
<td>(Rice et al., 2012)</td>
</tr>
<tr>
<td>ND, regular drinking</td>
<td>CHRNA5-A3-B4</td>
<td>Finnish Twin Cohort</td>
<td>(Broms et al., 2012)</td>
</tr>
<tr>
<td>Objective measures of tobacco exposure</td>
<td>CHRNA5-A3</td>
<td>Six European studies</td>
<td>(Munafo et al., 2012)</td>
</tr>
<tr>
<td>Onset of habitual smoking</td>
<td>CHRNA5-A3-B4</td>
<td>COGA</td>
<td>(Kapoor et al., 2012)</td>
</tr>
<tr>
<td>Genetic vulnerability to smoking in early-onset smokers</td>
<td>CHRNA5</td>
<td>Meta-Analysis</td>
<td>(Hartz et al., 2012)</td>
</tr>
<tr>
<td>Smoking quantity</td>
<td>CHRNA5-A3</td>
<td>STOMP$^g$ Genetics Consortium</td>
<td>(David et al., 2012)</td>
</tr>
<tr>
<td>General substance use initiation</td>
<td>CHRNA5-A3-B4</td>
<td>Add Health</td>
<td>(Lubke et al., 2012)</td>
</tr>
<tr>
<td>Pack years of smoking</td>
<td>CHRNA5-A3</td>
<td>Chinese Han population of chronic obstructive pulmonary disease patients</td>
<td>(Zhou et al., 2012)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA5-A3-B4, CHRN5-A6</td>
<td>COGEND</td>
<td>(Chen et al., 2012a)</td>
</tr>
<tr>
<td>CPD, age of smoking cessation, response to pharmacologic therapy</td>
<td>CHRNA5-A3-B4</td>
<td>Atherosclerosis Risk in Communities, University of Wisconsin Transdisciplinary Tobacco Use Research Center</td>
<td>(Chen et al., 2012b)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA4, CHRN B2</td>
<td>Japanese</td>
<td>(Chen et al., 2013)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA4, CHRN B2</td>
<td>CADD, GADD°</td>
<td>(Kamens et al., 2013)</td>
</tr>
<tr>
<td>Alcohol use</td>
<td>CHRNA5-A3-B4</td>
<td>National FINRISK Study, Health 2000 Survey</td>
<td>(Hallfors et al., 2013)</td>
</tr>
<tr>
<td>Response to smoking cessation therapies</td>
<td>CHRNA5-A3-B4</td>
<td>Eight clinical trials for smoking cessation</td>
<td>(Bergen et al., 2013)</td>
</tr>
<tr>
<td>Smoking quantity, snus consumption</td>
<td>CHRNA5-A3-B4</td>
<td>Nord-Trøndelag Health Study</td>
<td>(Gabrielsen et al., 2013)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRN B3</td>
<td>Subjects of European, African, and Asian ancestry from MSTCC°, SAGE, CGEMS°, and KARE° studies</td>
<td>(Cui et al., 2013)</td>
</tr>
<tr>
<td>Nicotine intake</td>
<td>CHRNA5-A3-B4</td>
<td>Alaskan natives</td>
<td>(Zhu et al., 2013)</td>
</tr>
<tr>
<td>Onset of regular smoking</td>
<td>CHRNA5-A3-B4</td>
<td>Meta-Analysis</td>
<td>(Stephens et al., 2013)</td>
</tr>
<tr>
<td>Cocaine use disorder</td>
<td>CHRN B3</td>
<td>SAGE</td>
<td>(Sadler et al., 2014)</td>
</tr>
<tr>
<td>CPD</td>
<td>CHRNA5-A3-B4, CHRN B3-A6, CHRNA2</td>
<td>SECASP°</td>
<td>(Cannon et al., 2014)</td>
</tr>
<tr>
<td>Frequency of binge drinking</td>
<td>CHRNA4</td>
<td>SECASP</td>
<td>(Coon et al., 2014)</td>
</tr>
<tr>
<td>ND</td>
<td>CHRNA2, CHRNA6</td>
<td>European and African Americans from TN, AR, MI, and MS; SAGE</td>
<td>(Wang et al., 2014)</td>
</tr>
<tr>
<td>CPD</td>
<td>CHRN B4</td>
<td>COGEND</td>
<td>(Haller et al., 2014)</td>
</tr>
<tr>
<td>Opioid dependence and withdrawal</td>
<td>CHRNA3</td>
<td>SAGE</td>
<td>(Muldoon et al., 2014)</td>
</tr>
<tr>
<td>Dizziness at smoking initiation</td>
<td>CHRNA3, CHRNA4, CHRNA6, CHRNA7</td>
<td>Nicotine Dependence in Teens</td>
<td>(Pedneault et al., 2014)</td>
</tr>
<tr>
<td>Smoking cessation</td>
<td>CHRNA5-A3-B4</td>
<td>Two smoking cessation trials in African Americans: the Nicotine Gum Study, and Buproprion Study</td>
<td>(Zhu et al., 2014)</td>
</tr>
<tr>
<td>Use of smokeless tobacco</td>
<td>CHRNA5-A3</td>
<td>Two studies conducted in India: the International multicenter oral cancer study, and the Mumbai study</td>
<td>(Anantharaman et al., 2014)</td>
</tr>
<tr>
<td>Severe withdrawal in a subgroup of smokers with higher lifetime prevalence of depression</td>
<td>CHRNA4</td>
<td>Treatment-seeking smokers from 5 Hungarian cessation centers</td>
<td>(Lazary et al., 2014)</td>
</tr>
<tr>
<td>COPD, lung cancer, heavy smoking (pack)</td>
<td>CHRNA7</td>
<td>Southern and eastern Han Chinese population</td>
<td>(Yang et al., 2014)</td>
</tr>
</tbody>
</table>
Twin studies have estimated that ND and smoking quantity are roughly 56-72% and 51-61% heritable in men and women, respectively (Broms et al., 2006; Kendler et al., 1999; Lessov et al., 2004; True et al., 1999; Vink et al., 2004); although there have been many associations between the CHRN genes and drug use phenotypes, the most well-replicated of these associations (between a SNP in CHRNA5 and smoking behaviors) account for a very small proportion of the variance in smoking (roughly 1 cigarette per day) (Berrettini et al., 2012; Saccone et al., 2010b). Based on previous work that provides evidence that common SNPs explain a large proportion of the heritability in height, Crohn’s disease, bipolar disorder, and type I diabetes (Lee et al., 2011; Yang et al., 2010), we hypothesize that common SNPs also explain a proportion of the variation in drug behaviors. In fact, common variation can account for up to 36% and 21% of the variation in nicotine and alcohol behaviors, respectively (Vrieze et al., 2013). The goal of aims three and four was to identify genes from a biological standpoint that may contribute risk to nicotine and alcohol behaviors and test these genes for association with these behaviors using publically available GWAS data.
CHAPTER II

EXAMINATION OF GENETIC VARIATION IN GABRA2 WITH CONDUCT DISORDER AND ALCOHOL ABUSE AND DEPENDENCE IN A LONGITUDINAL STUDY

This work was published in Behavior Genetics (Melroy et al., 2014)

Introduction

Alcohol use is a severe and worldwide problem; statistics from the World Health Organization showed that, as of February 2011, the harmful use of alcohol results in 2.5 million deaths each year (WHO). Alcohol is the world’s 3rd largest risk factor for disease, including cardiovascular disease, liver cirrhosis, and various cancers. Furthermore, 320,000 young people between the ages of 15 and 29 die from alcohol-related causes annually (Stevens et al., 2009). Many current behavior genetic approaches seek to identify some of the underlying genetic factors that confer higher risk to genetically influenced disorders; here, we examine genetic variation in gamma aminobutyric acid receptor alpha 2 (GABRA2) for association with phenotypes relating to alcohol use disorders using two independently ascertained samples.

There is strong evidence that adolescent conduct disorder (CD) is influenced substantially by genetic factors (Gelhorn et al., 2005; Rose et al., 2004; Slutske et al., 1997). CD has been shown to be a robust predictor of both concurrent and future alcohol dependence (Kuperman et al., 2001; Moss et al., 2001; Palmer et al., 2013; White et al., 2001) and there is evidence for shared genetic influence between these disorders (Button et al., 2007; Kendler et al., 2003; Slutske et al., 1998). Twin and family studies have also shown evidence for a general vulnerability to substance use disorders (Button et al., 2006), and, more recently, suggested that a general externalizing liability accounts for much of the genetic risk in substance use disorder and behavior disinhibition phenotypes (Hicks et al., 2011). Therefore, one hypothesis is that the development of traits that are influenced by genetic factors can vary over time, where CD may
be an adolescent manifestation of genetic factors that predispose to adult alcohol dependence (Dick et al., 2006).

The gamma aminobutyric acid, GABA, neurotransmitter is the predominant inhibitory neurotransmitter in the central nervous system and is a reasonable target for candidate gene studies on alcoholism. Not only are GABA receptors present in the mesolimbic dopamine pathway (Johnson et al., 1992; Steffensen et al., 1998), widely believed to play a role in the development of addiction, but studies using both rodent and human brain samples have shown that long-term ethanol exposure, as well as ethanol withdrawal, causes alterations in GABA(A) receptor subunit expression (Devaud et al., 1997; Devaud et al., 1996; Dodd et al., 1992; Grobin et al., 1998; Lewohl et al., 1997; Matthews et al., 1998; Mitsuyama et al., 1998). GABA can also modulate emotion and response to stress, further implicating this neurotransmitter system in drug behaviors (Herman et al., 2004; Martijena et al., 2002).

Results from the Collaborative Study on the Genetics of Alcoholism demonstrated highly significant associations between alcohol dependence and single nucleotide polymorphisms (SNPs) in GABRA2 (Edenberg et al., 2004). Numerous studies have replicated these associations with GABRA2 and alcohol dependence (Bierut et al., 2010a; Covault et al., 2004; Covault et al., 2008; Drgon et al., 2006; Enoch et al., 2009; Fehr et al., 2006; Ittiwut et al., 2012; Lappalainen et al., 2005; Li et al., 2014; Olfson et al., 2012; Philibert et al., 2009; Soyka et al., 2008; Villafuerte et al., 2012), although one study has found this effect in the opposite direction (Lind et al., 2008). While these studies have primarily focused on adult alcohol dependence, two studies have examined variation in GABRA2 for adolescent alcohol dependence and CD. These studies found no association with GABRA2 and adolescent alcohol dependence, but found evidence for an association with CD in adolescents (Dick et al., 2006; Sakai et al., 2010). These results are not altogether surprising since estimates of heritability of alcohol dependence are roughly 64% in adulthood (Heath et al., 1997), yet no genetic effects are present in adolescence, with variance in adolescent alcohol dependence being attributed to environmental
influences (Rose et al., 2004). It is worth noting however, that a recent longitudinal study showed genetic influences ranging from 35-40% on general substance use disorders between ages 14-17 that decreased with age (Vrieze et al., 2012).

The study presented here was designed as a replication and extension of the Sakai et al. (2010) study, which found some evidence for association with adolescent CD in a subset of the Colorado Center on Antisocial Drug Dependence (CADD) sample (although it did not survive statistical correction for multiple testing). In the current study, the sample size has been expanded and an independent replication sample is now available. We now have an additional wave of data collection; thus adolescent alcohol problems and CD could be evaluated at the first wave of data collection (adolescence and young adulthood) and alcohol abuse and dependence (AAD) subsequently examined in the second wave of data collection (adulthood).

Furthermore, previous studies have focused on the diagnosis of alcohol dependence as a categorical phenotype (Bierut et al., 2010a; Covault et al., 2004; Covault et al., 2008; Dick et al., 2006; Drgon et al., 2006; Edenberg et al., 2004; Enoch et al., 2009; Fehr et al., 2006; Ittiwut et al., 2012; Lappalainen et al., 2005; Li et al., 2014; Olfson et al., 2012; Sakai et al., 2010; Soyka et al., 2008). In this study we examined the sum of AAD symptoms as a quantitative trait. There are three reasons for this approach: 1) item-response theory work that has shown that abuse and dependence symptoms provide overlapping information on severity (Gelhorn et al., 2008; Langenbucher et al., 2004; Martin et al., 2006; Saha et al., 2006), 2) this analysis is clinically relevant given the field has moved toward new DSM V criteria that merged abuse and dependence symptoms, and 3) continuous variables can provide a more accurate estimate of the phenotype.

**Materials and methods**

**Samples**
The sample consisted of non-Hispanic European American (EA) subjects and Hispanic subjects drawn from the Colorado Center on Antisocial Drug Dependence (CADD). The CADD is a longitudinal study currently in its 3rd wave of data collection consisting of four separate samples. The Family Study (FS) is a sample of clinical adolescent probands, ascertained while in treatment for antisocial drug dependence, their family members, and a matched set of control families (Stallings et al., 2005; Stallings et al., 2003). The Colorado Adoption Project (Petrill SA, 2003), the Colorado Longitudinal Twin Study (Rhea et al., 2006), and the Colorado Community Twin Study (Rhea et al., 2006) represent community unselected samples that were used in this study for phenotypic standardization (see below). Only subjects from the FS were included in the molecular genetic study. Briefly, clinical probands were recruited from treatment facilities in the Denver area. Probands were selected from individuals who had consecutive admissions to the treatment facilities between February of 1993 and June of 2001. Controls were recruited from the community and matched to the clinical probands based on age, gender, ethnicity, and zip code. All individuals living in the same household as the proband were asked to participate in the study, which created family-based data. For this study data from both wave 1, data collection for which began in 1997 and ended in 2002, and wave 2, data collection for which began in 2002 and ended in 2008, were used. When assessed, buccal cell DNA was collected from subjects who gave voluntary consent. All recruitment, assessment, and DNA collection procedures were approved by the University of Colorado's IRB.

Genetics of Antisocial Drug Dependence

As an independent replication sample, EA subjects and Hispanic subjects from the Genetics of Antisocial Drug Dependence (GADD) were assessed, as previously described (Kamens et al., 2013). Probands in Denver, CO, and San Diego, CA, were identified from treatment programs, involvement with the criminal justice system, or special schools who met at
least one of the criteria for having a substance use disorder (other than nicotine dependence) and CD. Siblings of the proband, as well as one or both biological parents, were included in the sample as well. The GADD is also a longitudinal study in the 2nd wave of data collection; data from wave 1, collected between 2001 and 2006, and wave 2, collection of which began in 2009 and is ongoing, were utilized in the analyses. DNA was obtained with consent through either buccal cells or blood. The University of California and the University of Colorado IRBs approved all subject recruitment, assessment, and DNA collection procedures.

**SNP selection**

The candidate SNPs were identified through review of primary literature, focusing on SNPs that have been previously associated with alcohol dependence and CD in order to replicate previous findings (Bierut et al., 2010a; Covault et al., 2004; Covault et al., 2008; Dick et al., 2006; Edenberg et al., 2004; Enoch et al., 2009; Fehr et al., 2006; Ittiwut et al., 2012; Lappalainen et al., 2005; Li et al., 2014; Lind et al., 2008; Philibert et al., 2009; Sakai et al., 2010; Soyka et al., 2008). Two of the most well-replicated SNPs for alcohol dependence, rs279858 and rs279871, were chosen in addition to 3 other SNPs from GABRA2: rs567926, rs279845, and rs9291283. These three later SNPs have been associated also with alcohol dependence in the literature (Bierut et al., 2010a; Covault et al., 2004; Covault et al., 2008; Edenberg et al., 2004; Fehr et al., 2006; Li et al., 2014; Lind et al., 2008; Philibert et al., 2009; Soyka et al., 2008), but with less support than rs279858 and rs279871. Rs567926, rs279845, and rs9291283 were not in high linkage disequilibrium (LD) at $r^2 < 0.8$ with rs279858 during preliminary examination of the haplotype structure of GABRA2 in the Hapmap sample of Utah residents with ancestry from northern and western Europe using Haploview (Barrett et al., 2005). Genomic DNA extracted from buccal or blood cells was amplified with primer extension preamplification (Anchordoquy et al., 2003) or using the REPLI-g kit according to the manufacturers protocol (Qiagen, Valencia, California). Genotyping was performed in the CADD
initially and it was observed that two of the SNPs, rs279858 and rs279871, were in high LD ($r^2 > 0.8$ in both Hispanics and EAs) and thus rs279871 was not genotyped in the GADD.

**Genotyping**

SNP genotyping was performed with TaqMan®® assays for allelic discrimination according to manufacturer's instructions (Applied Biosystems, Foster City, California). Two thousand four hundred and ninety six subjects from CADD and 3,072 subjects from GADD were genotyped. Polymerase Chain Reaction (PCR) reactions were performed with the Biomek® 3000 Laboratory Automation Workstation (Beckman Coulter Inc, Brea, California) and the Dual 384-Well GeneAmp®® PCR system 9700 (Applied Biosystems, Foster City, California). To analyze the amplified plates a 7900 Real-Time PCR System (Applied Biosystems, Foster City, California) was used. Based on all the SNPs that have been previously genotyped in these samples from the CADD (33) and GADD (12), DNA samples for subjects with overall call rates <90% were excluded. Three hundred and eighty four samples were genotyped twice for each SNP to determine concordance between replicate reactions; the percent of discordant calls for SNPs in the CADD and GADD, respectively, are shown as follows: rs567926 (0.78%, 0.00%), rs279858 (0.78%, 0.26%), rs279871 (0.52%, N/A), rs279845 (1.30%, 0.00%), rs9291283 (0.26%, 0.78%). Genotype clusters from the amplified 384 well plates were auto-called by the Applied Biosystems TaqMan®® Genotyper software (Applied Biosystems, Foster City, California) and subsequently visually examined by two independent lab personnel to verify calls. Final calls were determined when both laboratory personnel agreed; if they did not agree genotypic data were excluded.

**Statistical analysis**

Data descriptive
Mendelian errors were identified using FBAT (Rabinowitz et al., 2000). If Mendelian errors were detected, the SNP genotype for that family was removed. Using Haploview (Barrett et al., 2005), pairwise LD ($r^2$) and Hardy-Weinberg equilibrium (HWE) were evaluated.

**Analysis of the CADD sample**

Three phenotypes were examined: lifetime sum CD symptoms in adolescence/young adulthood, sum AAD symptoms in adolescence/young adulthood, and sum AAD symptoms in adulthood. CD was assessed in adolescents using the Diagnostic Interview Schedule for Children (DISC) (Shaffer et al., 1993). Early study participants were evaluated with DISC 2.3, which assessed DSM IIIR diagnoses. Later participants were assessed using DISC IV, which assessed DSM IV diagnoses. CD was assessed for subjects over the age of 18 utilizing the Diagnostic Interview Schedule (DIS) (Robins et al., 1981), and, similar to the DISC, began with DSM IIIR diagnoses and finished with DSM IV diagnoses. DSM IV defined AAD were assessed with the Composite International Diagnostic Interview – Substance Abuse Module (CIDI-SAM) (Cottler et al., 1990). As comorbidity is high in this sample, subjects with comorbid drug use were not excluded (Stallings et al., 2003). Each variable was standardized and analyzed as described below.

**Conduct Disorder:** For the adolescent/young adult lifetime CD symptoms score, data were drawn from wave 1. The majority of subjects were assessed using DSM IV criteria but a small proportion were assessed using DSM IIIR criteria. Sum symptom counts for CD were standardized based on the distribution of symptoms in the CADD community sample. A linear regression was performed using Statistical Analysis System (SAS) 9.3 software (SAS Institute Inc., Cary, NC) to determine residuals from sex, age, and age squared. The coefficients from the CADD community sample were applied to the CADD clinical sample (i.e. Z scores of clinical subjects were expressed as deviations from the means of the community samples). In this case, clinical subjects assessed using DSM IV criteria were standardized to the community subjects.
assessed with DSM IV criteria, and the same procedure was used for subjects assessed with DSM IIIR criteria. After standardization, only phenotypic data from subjects between the ages of 10 and 25 were included in the analysis. The analysis was performed on 1,789 subjects and run as described in a later section.

Adolescent/young adult alcohol abuse and dependence: For the adolescent/young adult AAD symptoms sum score, data were drawn from wave 1 of data collection. All sum AAD counts were assessed using DSM IV criteria. Only phenotypic data for subjects who had previously used alcohol were included; in other words, if a subject had never had a drink of alcohol they were excluded from the analysis. Of the 492 adolescent subjects with no dependence symptoms, 25.6% of them endorsed an abuse criterion, strengthening the argument for inclusion of abuse symptoms. Sum symptom counts were standardized in the same manner described for CD above, using residuals and coefficients derived in the community sample and applying these to the clinical subjects. After standardization, only phenotypic data from subjects between the ages of 10 and 25 were included for analysis. The analysis was performed on 1,199 subjects.

Adult alcohol abuse and dependence: For adult AAD symptoms sum score, data were drawn from wave 2. All sum AAD counts were assessed using DSM IV criteria. Only phenotypic data for subjects who had previously used alcohol were included. Of the 246 adult subjects with no dependence symptoms, 25.6% of them endorsed an abuse criterion. Sum symptom counts were standardized using the CADD community sample as above for CD and adolescent AAD. For the analysis only subjects assessed at wave 2 that were also included in the adolescent/young adult analyses of CD and AAD were analyzed. The analysis was performed on 703 subjects.

Analysis of the GADD sample
As in the CADD, CD was assessed in adolescents using the DISC and in subjects over 18 using the DIS. All study participants were evaluated using DSM IV diagnoses. DSM IV defined AAD were assessed with the CIDI-SAM. Subjects with comorbid drug use were not excluded. Each variable was standardized and analyzed as described below.

**Conduct disorder:** For lifetime CD symptoms, data were drawn from wave 1. All of the subjects were assessed using DSM IV criteria. The GADD sample includes only clinical subjects, so the coefficients from the standardization of the DSM IV CD symptoms from the CADD community sample were applied to the GADD sample (i.e. Z scores of GADD subjects were expressed as deviations from the means of the CADD community samples). After standardization, only phenotypic data from subjects between the ages of 10 and 25 were analyzed. The analysis was performed on 1,540 subjects.

**Adolescent/young adult alcohol abuse and dependence:** Adolescent/young adult AAD symptoms were drawn from wave 1 of data collection. All items were assessed using DSM IV criteria. As in the CADD, only phenotypic data for subjects who had previously used alcohol were included. Of the 572 adolescent subjects with no dependence symptoms, 10.5% of them endorsed an abuse criterion. For standardization, the coefficients from the standardization of the adolescent AAD symptoms from the CADD community sample were applied to the GADD sample. After standardization, only phenotypic data from subjects between the ages of 10 and 25 were included. The analysis was performed on 1,186 subjects.

**Adult alcohol abuse and dependence:** Adult AAD symptoms were drawn from wave 2. All items were assessed using DSM IV criteria. Only phenotypic data for subjects who had previously used alcohol were included. Of the 150 adult subjects with no dependence symptoms, 30.7% of them endorsed an abuse criterion. For standardization, the coefficients from the standardization of the adult AAD symptoms from the CADD community sample were applied to the GADD sample. For the analysis only standardized phenotypic data from the
subjects assessed at wave 2 that were concurrently included in the GADD adolescent/young adult analyses of CD and AAD were analyzed. The analysis was performed on 873 subjects.

Combined analysis of the CADD and GADD samples

To increase statistical power, the GADD and CADD samples were combined after the data were cleaned in each sample (e.g. removal of Mendelian errors and phenotypic standardization). The analysis of the combined sample was run as described below.

Statistical analysis

The data were analyzed using an additive genetic model in a family-based association test performed in FBAT (Rabinowitz et al., 2000). FBAT builds on the original Transmission Disequilibrium Test (TDT) (Spielman et al., 1993), where alleles transmitted to extreme offspring are compared to the expected distribution of alleles among offspring under Mendel’s law of segregation and conditioning. Three phenotypes were analyzed using an additive test in FBAT: adolescent/young adult CD, adolescent/young adult AAD, and adult AAD.

Correction for multiple testing

We used the spectral decomposition (SNPSpD) method (Nyholt, 2004) to estimate the minimum p-value required to keep experimental type I error < 0.05. This method is used as a correction for multiple testing for SNPs in linkage disequilibrium with each other. All Hispanic and EA family members were included to define a new p-value used to correct for multiple testing. The spectral decomposition was run using each sample separately to provide a corrected p-value for each analysis of the CADD, GADD, and CADD/GADD combined samples.

Results
The phenotypic characteristics between the CADD clinical and GADD samples are comparable as shown in Figure 1 below.

![Figure 1](image)

**Figure 1.** Study sample characteristics of the CADD and GADD. Each phenotypic variable is shown with the respective wave the data were drawn from. As CADD is a sample comprised of clinical and control subjects, variable descriptives for each are shown separately. The mean ± the standard deviation is shown for age, lifetime CD symptoms, and lifetime adolescent/young adult and adult AAD symptoms for each variable used in the analysis, as well as the sample size for each variable.

Although the same subjects were used in the adolescent/young adult CD and AAD analyses, only phenotypic data from subjects that had used alcohol were included in the adolescent/young adult AAD analysis, accounting for the difference in sample size between the two adolescent/young adult analyses. This was done because most of the previous studies that
found associations between *GABRA2* and alcohol dependence used alcohol dependent cases and matched controls in their analyses (Bierut *et al.*, 2010a; Covault *et al.*, 2004; Covault *et al.*, 2008; Drgon *et al.*, 2006; Edenberg *et al.*, 2004; Enoch *et al.*, 2009; Fehr *et al.*, 2006; Ittiwut *et al.*, 2012; Lappalainen *et al.*, 2005; Olfson *et al.*, 2012; Soyka *et al.*, 2008), thus only subjects that had used alcohol were included in the present study in order to minimize sample differences between the current and past studies. Although our adolescent/young adult analyses include subjects of a substantially large age range, this is the approximate age range for probands in wave 1 of the CADD and GADD samples.

The minor alleles were identical in EAs and Hispanics and the minor allele frequencies were similar for rs567926, rs279858, rs279845, and rs9291283 (see Table 1 below).

**Table 1.** Genotypic characteristics of the CADD and GADD samples for each ethnic group. Location refers to base pair position on chromosome 4 from the UCSC genome browser. Alleles shown are major/minor allele. MAF: minor allele frequency. HWE refers to the Hardy-Weinberg equilibrium p-value. Due to high LD between rs279858 and rs279871, rs279871 was not genotyped in the GADD sample.

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>SNP</th>
<th>Location</th>
<th>Alleles</th>
<th>MAF</th>
<th>HWE</th>
<th>Alleles</th>
<th>MAF</th>
<th>HWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Hispanic European American</td>
<td>rs567926</td>
<td>46241769</td>
<td>A/G</td>
<td>0.48</td>
<td>0.87</td>
<td>A/G</td>
<td>0.42</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>rs279858</td>
<td>46314593</td>
<td>T/C</td>
<td>0.48</td>
<td>0.86</td>
<td>T/C</td>
<td>0.42</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>rs279871</td>
<td>46307533</td>
<td>T/C</td>
<td>0.48</td>
<td>0.82</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>rs279845</td>
<td>46329723</td>
<td>T/A</td>
<td>0.49</td>
<td>0.98</td>
<td>T/A</td>
<td>0.44</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>rs9291283</td>
<td>46371833</td>
<td>G/A</td>
<td>0.27</td>
<td>0.87</td>
<td>G/A</td>
<td>0.25</td>
<td>0.98</td>
</tr>
<tr>
<td>Hispanic</td>
<td>rs567926</td>
<td>46241769</td>
<td>A/G</td>
<td>0.50</td>
<td>0.05</td>
<td>A/G</td>
<td>0.44</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>rs279858</td>
<td>46314593</td>
<td>T/C</td>
<td>0.50</td>
<td>0.14</td>
<td>T/C</td>
<td>0.46</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>rs279871</td>
<td>46307533</td>
<td>T/C</td>
<td>0.50</td>
<td>0.07</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>rs279845</td>
<td>46329723</td>
<td>A/T</td>
<td>0.49</td>
<td>0.17</td>
<td>T/A</td>
<td>0.46</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>rs9291283</td>
<td>46371833</td>
<td>G/A</td>
<td>0.22</td>
<td>0.21</td>
<td>G/A</td>
<td>0.17</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Hispanic and EA subjects were combined for all analyses since the allele frequencies did not differ significantly. There were no significant deviations from the Hardy-Weinberg equilibrium at p < 0.05.

In EA and Hispanic subjects in the CADD, the SNPs were somewhat correlated with an $r^2 > 0.50$, with the exception of rs9291283 (Figure 2).
**Figure 2.** Linkage Disequilibrium (LD) plot generated by Haploview (Barrett et al., 2005) for the SNPs chosen in the CADD sample. Numbers in the boxes are $r^2 \times 100$. **(Left)** LD plot for EAs. Note that rs279858 and rs279871 are in high LD. **(Right)** LD plot for Hispanics.

As mentioned previously, rs279858 and rs279871 were highly correlated at $r^2 > 0.80$. Similar LD patterns were observed in EAs and Hispanics in the GADD sample as in the CADD sample (Figure 3), which is in agreement with the similarity in allele frequencies.
Figure 3. Linkage Disequilibrium (LD) plot generated by Haploview (Barrett et al., 2005) for the SNPs chosen in the GADD sample. Numbers in the boxes are $r^2 \times 100$. (Left) LD plot for EAs. (Right) LD plot for Hispanics.

In addition, the $r^2$ value between rs9291283 and all other SNPs examined was less than 0.10, indicating this SNP represents an independent signal in the gene.

As expected from previous analyses of these SNPs, no significant associations were detected with AAD in adolescence/young adulthood in the CADD, GADD, or combined sample (results not shown).

Rs9291283 was suggestively associated with AAD in adulthood in the CADD sample (Table 2).
Table 2. Results from the FBAT analysis in the CADD. Only the results for the risk allele are shown. The number of informative families used in the analysis is shown to the right of the p-value.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>CD at adolescence/young adulthood</th>
<th>AAD at adulthood</th>
<th>SNP</th>
<th>Risk Allele</th>
<th>Z score</th>
<th>P-value</th>
<th># Fam</th>
<th>Risk Allele</th>
<th>Z score</th>
<th>P-value</th>
<th># Fam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G</td>
<td>0.387</td>
<td>0.699</td>
<td>214</td>
<td>G</td>
<td>1.625</td>
<td>0.104</td>
<td>164</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.180</td>
<td>0.857</td>
<td>232</td>
<td>C</td>
<td>0.525</td>
<td>0.599</td>
<td>170</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>1.283</td>
<td>0.199</td>
<td>202</td>
<td>A</td>
<td>1.100</td>
<td>0.271</td>
<td>159</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>0.997</td>
<td>0.319</td>
<td>172</td>
<td>A</td>
<td>1.859</td>
<td>0.063</td>
<td>129</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The sign of the Z score indicates which allele is over transmitted; a positive Z score indicates an allele is over transmitted and is interpreted as the risk allele.

Results from the spectral decomposition to correct for multiple SNP testing indicated the significance threshold required to keep Type I error rate at 5% is 0.017. One should bear in mind an additional correction for the four multiple phenotypes, but since the phenotypes are correlated we present results as comparison to the 0.017 cut-off. Although rs279871 was included in the spectral decomposition and FBAT analysis it was not significantly associated with the phenotypes of interest (results not shown). In this analysis rs9291283 showed a trend toward association with AAD in adulthood at p = 0.063, with the A allele conferring risk to AAD.

In the GADD sample, rs9291283 was associated with CD (Table 3).

Table 3. Results from the FBAT analysis using subjects in the GADD. Only the results for the risk allele are shown. The number of informative families used in the analysis is shown to the right of the p-value.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>CD at adolescence/young adulthood</th>
<th>AAD at adulthood</th>
<th>SNP</th>
<th>Risk Allele</th>
<th>Z score</th>
<th>P-value</th>
<th># Fam</th>
<th>Risk Allele</th>
<th>Z score</th>
<th>P-value</th>
<th># Fam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G</td>
<td>1.503</td>
<td>0.133</td>
<td>242</td>
<td>G</td>
<td>0.925</td>
<td>0.355</td>
<td>132</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.885</td>
<td>0.376</td>
<td>258</td>
<td>C</td>
<td>0.863</td>
<td>0.388</td>
<td>146</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>1.050</td>
<td>0.294</td>
<td>258</td>
<td>A</td>
<td>0.700</td>
<td>0.484</td>
<td>142</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>2.809</td>
<td>0.005</td>
<td>184</td>
<td>A</td>
<td>0.734</td>
<td>0.463</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The sign of the Z score indicates which allele is over transmitted; a positive Z score indicates an allele is over transmitted and is interpreted as the risk allele.

Indicates significant p-values after correction for multiple SNP testing at p = 0.019.
Results from spectral decomposition indicated a multiple SNP significance threshold required to keep Type I error rate at 5% is 0.019. In this sample, rs9291283 was significantly associated with CD in adolescence/young adulthood at \( p = 0.005 \), a finding that was not seen in the CADD sample. Again, the A allele was shown to be over-transmitted to extreme offspring.

In the combined CADD and GADD analysis, rs9291283 was nominally associated with AAD in adulthood (\( p = 0.057 \); Table 4).

**Table 4.** Results from the FBAT analysis using subjects in the CADD and GADD combined sample. Only the results for the risk allele are shown. The number of informative families used in the analysis is shown to the right of the p-value.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Risk Allele</th>
<th>CD at adolescence/young adulthood</th>
<th>AAD at adulthood</th>
<th>Risk Allele</th>
<th>Z score</th>
<th>P-value</th>
<th># Fam</th>
<th>Z score</th>
<th>P-value</th>
<th># Fam</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs567926</td>
<td>G</td>
<td>0.934</td>
<td>0.350</td>
<td>456</td>
<td>G</td>
<td>1.809</td>
<td>0.070b</td>
<td>296</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs279858</td>
<td>C</td>
<td>0.787</td>
<td>0.431</td>
<td>490</td>
<td>C</td>
<td>0.973</td>
<td>0.331</td>
<td>316</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs279845</td>
<td>A</td>
<td>0.071</td>
<td>0.943</td>
<td>460</td>
<td>A</td>
<td>1.270</td>
<td>0.204</td>
<td>301</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs9291283</td>
<td>A</td>
<td>1.494</td>
<td>0.135</td>
<td>356</td>
<td>A</td>
<td>1.902</td>
<td>0.057b</td>
<td>224</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) The sign of the Z score indicates which allele is over transmitted; a positive Z score indicates an allele is over transmitted and is interpreted as the risk allele.

\(^b\) Indicates trending p-values at \( p < 0.1 \).

Results from the spectral decomposition set the multiple SNP significance threshold required to keep Type I error rate at 5% at 0.019. In this analysis rs9291283 showed a trend toward association with AAD in adulthood at \( p = 0.057 \), replicating the trending association between rs9291283 and adult AAD in the CADD sample. However, neither of these associations was statistically significant.

**Discussion**

The goals of this study were to expand on previous work demonstrating strong associations of *GABRA2* gene SNPs with alcohol dependence in adulthood (Bierut *et al.*, 2010a; Covault *et al.*, 2004; Covault *et al.*, 2008; Drgon *et al.*, 2006; Edenberg *et al.*, 2004;
Enoch et al., 2009; Fehr et al., 2006; Itiwut et al., 2012; Lappalainen et al., 2005; Li et al., 2014; Lind et al., 2008; Offson et al., 2012; Philibert et al., 2009; Soyka et al., 2008; Villafuerte et al., 2012), and replicate work showing a suggestive association with CD in younger samples (Dick et al., 2006; Sakai et al., 2010). This study was an important extension of the Sakai et al. (2010) study; not only was adolescent CD and AAD assessed in a larger sample, but AAD in adulthood as well. This study design enabled the examination of genetic influences on the development of three traits over time.

In considering the SNPs in high LD and comparing results to previous reports, our findings are in agreement with Dick et al. (2006) and Sakai et al. (2010) in not finding an association with adolescent AAD and rs279871 or rs279858 in GABRA2, yet our results contrast with theirs in findings with CD. In the GADD sample, rs9291283 was associated with CD in adolescence/young adulthood (p = 0.005). This result was not seen in the CADD sample, or in the combined CADD and GADD sample. This may be due in part to the finding that, for CD, the G allele was over-transmitted to affected offspring in the CADD sample. Although this finding was not significant, it could explain why, when the two samples were combined, rs9291283 was not associated with CD despite the fact it was the strongest association in the study. In the Dick et al. (2006) and Sakai et al. (2010) studies, rs279871 was found to be associated with adolescent CD, yet here neither rs279871 nor the SNP we examined in high LD (\( r^2 > 0.8 \)) with this SNP, rs279858, was associated with CD. Interestingly, the Sakai et al. (2010) study used two association approaches: a case control and a family based approach. While the case control approach indicated evidence of an association between rs279871 and adolescent CD, the family based approach did not. This is particularly note-worthy in that there was some overlap between our wave 1 sample and the sample used by Sakai; eight hundred and thirty four Hispanic and EA subjects were used in both the present study and the study by Sakai et al. (2010). In addition to using different statistical approaches, non-replication may be related to the inclusion of both adolescents and young adults since both previous studies used only
adolescents. Furthermore, both Dick et al. (2006) and Sakai et al. (2010) used primarily DSM IIIR criteria for establishing CD symptoms while our study used primarily DSM IV criteria. Although the DSM IV and DSM IIIR classifications contain many of the same criteria for CD, the DSM IV, compared to DSM IIIR, introduced two new criteria, staying out late without permission (prior to age 13) and bullying, and also required that truancy began before age 13 to be counted. Recently however, a study by Dick and colleagues found evidence for association between rs279871 and subclinical self-reports of externalizing behavior, yet no evidence for association between rs279871 and DSM IV CD symptoms in adults or adolescents (Dick et al., 2013), lending support to our findings.

Rs9291283 was nominally associated with adult AAD in the CADD and combined CADD and GADD samples. Although rs9291283 has been previously shown to be nominally associated with alcohol dependence (Bierut et al., 2010a; Covault et al., 2004; Soyka et al., 2008), rs279858 has been the most well replicated SNP in GABRA2 associated with adult alcohol dependence (Bierut et al., 2010a; Covault et al., 2004; Edenberg et al., 2004; Fehr et al., 2006; Lappalainen et al., 2005; Li et al., 2014; Villafuerte et al., 2012), a finding which failed to replicate in our study. This result may be related to slight differences in the alcohol phenotypes examined. Most of the previous studies (Bierut et al., 2010a; Covault et al., 2004; Covault et al., 2008; Drgon et al., 2006; Edenberg et al., 2004; Enoch et al., 2009; Fehr et al., 2006; Ittiwut et al., 2012; Lappalainen et al., 2005; Li et al., 2014; Offson et al., 2012; Philibert et al., 2009; Soyka et al., 2008) have focused specifically on alcohol dependence. However, recent item response theory work has shown that AAD symptoms provide overlapping information on severity (Gelhorn et al., 2008; Langenbucher et al., 2004; Martin et al., 2006; Saha et al., 2006), so we included both alcohol abuse and dependence scores in our analysis. Although many studies have replicated GABRA2 SNP associations with adult alcohol dependence, there have been a few studies whose findings parallel our own (Lydall et al., 2011; Matthews et al., 2007; Onori et al., 2010) or contrast in directionality from the majority (Lind et
Matthews et al. (2007) found no association between rs279871 and rs279858 with alcohol dependence in a linkage study. Similarly, Lydall et al. (2011) found no association between rs279871, rs279858, or rs9291283 and alcohol dependence using a case control design. Of particular interest is the study by Onori et al. (2010) examining alcohol use disorders; no significant associations were found with rs567926, rs279871, rs279858, or rs279845. Finally, Lind et al. (2008) found evidence for association with rs279858 with quantitative sum alcohol dependence symptoms, as well as a single principal component factor score from the DSMIV alcohol dependence symptoms. However, the direction of effect was opposite from that previously observed and replicated (Bierut et al., 2010a; Covault et al., 2004; Edenberg et al., 2004; Fehr et al., 2006; Lappalainen et al., 2005; Li et al., 2014; Villafuerte et al., 2012) and furthermore did not survive correction for multiple testing. Thus the overall results for association of this region with alcohol disorders remain mixed.

There are some important strengths, as well as limitations, in our study. In particular, the use of a family-based design protects against population stratification. With the exception of the Sakai et al. (2010) study, prior analyses used primarily subjects of one ethnicity. While many ethnicities have been represented, including European American, Japanese, and African American populations, our study represents one of the first to include subjects of two ethnicities. Furthermore, the inclusion of the Hispanics provided a larger sample size. While it is possible that distributional differences in CD and AAD may be culturally influenced and affect our results, we do not believe this is the case because the distributions for CD and adult AAD between EAs and Hispanics were comparable in our samples. In addition, we used a continuous variable to examine CD and AAD, rather than dichotomous variable. Continuous variables provide a more accurate estimate of the phenotype, as dichotomizing variables can cause a loss of information. Moreover, our incorporation of both AAD symptoms is unique. This study, however, is not without limitations. Specifically, there is low power to detect and replicate genetic associations when the CADD and GADD samples are analyzed separately, a fact that could explain false
positives. In addition, there might exist unidentified ascertainment or sample characteristics differences between the CADD and GADD samples, again leading to inconsistent results. High comorbid drug use is also present in our samples, which could confound results. Although we attempted to capture all the genetic signals in the region, some may have been missed. Furthermore, it is becoming clear that the genetic risk for alcohol disorders is likely due to common variants in many genes, each of small effect, as well as rare variants with potentially large effects (Enoch, 2013). The advent of next generation sequencing methods may allow identification of novel rare variants in future studies.

The analyses presented here were run in several different ways, all of which showed consistent results. Although we limited our sample size in the adult AAD analyses by only including those subjects who where included in the adolescent/young adult CD and AAD, when we did not impose this restriction and included all subjects from waves 1 and 2 the results yielded the same findings. The goal of limiting the samples was to ensure that each subject in the analysis was represented at each time point. Furthermore, when subjects who had never had alcohol were included in the analysis, the results were comparable.

The ideal end goal of many of these human molecular genetic studies is to inform future pharmacogenetic studies of the role of certain polymorphisms in genetically influenced disorders. While it is likely to take time for these approaches to be implemented in clinical practice, molecular genetic studies serve to identify possible targets for novel treatments. Although our findings provide limited evidence for a role of GABRA2 in CD and AAD, GABRA2 is also a likely candidate for disorders such as anxiety, general substance abuse, depression, and schizophrenia (Engin et al., 2012). For example, a recent haplotype analysis of rs9291283, the SNP with which we found the strongest evidence of association with CD in the GADD sample, revealed a significant association with cocaine addiction, and it has been postulated that rs894269, a SNP in high LD with rs9291283, lies in a cis-enhancer region of the gene. Future studies examining the possible function of rs894269 should be pursued in order to
examine whether it may represent the underlying causal variant contributing the association seen with its correlates, including rs9291283 (Dixon et al., 2010).

In conclusion, our results provide limited support for an association between rs9291283 and CD in adolescence/young adulthood and AAD in adulthood. Unfortunately, we did not replicate previous findings showing associations between rs279871 and rs279858 with adolescent CD and adult AAD, respectively. However, perhaps the more important implication here is the importance of replication and combining studies. There is an increasing literature on negative and non-replicated findings and the field of behavior genetics has yet to identify the reasons for this (Hewitt, 2012). There are many possibilities, including differences in sample ascertainment, assessment, as discussed above. It is equally likely that many of the initial positive associations are false positives, and this study demonstrates the utility of including an independent replication sample in the study design, so that the results from an individual sample can be weighted in the context of their reproducibility.
CHAPTER III

NO EVIDENCE FOR ASSOCIATION OF COMMON VARIANTS IN \textit{GRM7} WITH ALCOHOL CONSUMPTION

This work has been resubmitted to \textit{Alcohol}

\textbf{Introduction}

There is strong evidence that alcohol consumption (AC) phenotypes are more heritable in adults than in adolescents, with estimates of heritability ranging from 31-51\% (Hansell \textit{et al.}, 2008). Furthermore, studies have indicated that the same genetic influences are acting on both alcohol dependence (AD) and AC (Kendler \textit{et al.}, 2010), suggesting that heavy AC in adolescence can lend itself to the development of alcohol abuse and potentially AD in adulthood. While Genome-Wide Associations studies (GWAS) have identified several genes associated with AC (Baik \textit{et al.}, 2011; Chen \textit{et al.}, 2012\textit{c}; Heath \textit{et al.}, 2011; Kapoor \textit{et al.}, 2013; Pan \textit{et al.}, 2013; Pei \textit{et al.}, 2012; Quillen \textit{et al.}, 2014; Schumann \textit{et al.}, 2011), these associations often have small effect, indicating that AC is a complex problem influenced by numerous genes, many of which are likely still unidentified.

Recent work has identified the mouse glutamate metabotropic receptor 7 (\textit{Grm7}) as a candidate gene for AC (Vadasz \textit{et al.}, 2007). This receptor affects alcohol drinking patterns and neurological processes. \textit{Grm7} knockout mice show increased AC, while mice overexpressing this gene drink less alcohol. Furthermore, there are significant differences in \textit{Grm7} mRNA abundance between congenic and background strains in addiction related brain regions. In this experiment, the congenic strain created contained a small segment of the inbred BALB/cJ mouse strain chromosome 6 (on which \textit{Grm7} resides) on a C57BL/6By background strain. Comparison between the congenic and background strains revealed that congenic strains consumed alcohol less than the background strains, providing further evidence that \textit{Grm7}
affects alcohol consumption (Gyetvai et al., 2011). Moreover, viral-mediated knockdown of Grm7 in the Nucleus Accumbens (NAc), a region strongly implicated in drug reward, was also shown to increase ethanol consumption and preference in rats (Bahi, 2013).

Pharmacological interventions in rodent models have also supported a role for Grm7 in the rewarding effects of drugs. Administration of the Grm7 agonist AMN082 (Mitsukawa et al., 2005) was shown to decrease gamma aminobutyric acid (GABA) and increase glutamate levels in the NAc (Li et al., 2008b). The same agonist was later shown to decrease cocaine self-administration and cocaine-induced enrichment of electrical brain-stimulation reward in rats, suggesting that Grm7 is involved in cocaine reinforcement (Li et al., 2009b). In addition, AMN082 was shown to decrease ethanol consumption and preference in rats (Bahi et al., 2012), and subsequent work has shown that although AMN082 had no effect on extinction of ethanol-induced conditioned place preference, it reduced ethanol-induced conditioned place preference reinstatement. This suggests that Grm7 activation can lessen the reinstatement of the rewarding effects of alcohol (Bahi, 2012). These studies provide strong evidence for Grm7 as a potential target for alcohol cessation therapies.

Although there has been work examining the effect of human glutamate metabotropic receptor 7 (GRM7) polymorphisms on human substance use disorders, the majority of the work has focused on psychiatric disorders (Saus et al., 2010), including bipolar disorder (Alliey-Rodriguez et al., 2011), schizophrenia (Ganda et al., 2009; Ohtsuki et al., 2008; Shibata et al., 2009), major depressive disorder (Breen et al., 2011; Hamilton, 2011; Muglia et al., 2010; Pergadia et al., 2011; Shyn et al., 2011), panic disorder (Otowa et al., 2009), autism spectrum disorders (Yang et al., 2013), and ADHD (Elia et al., 2011; Park et al., 2013). Although not genome wide significant, studies have found evidence for association between GRM7 and nicotine dependence and cessation success (Uhl et al., 2010) and alcoholism (Li et al., 2011). Furthermore, variation in GRM7 was identified as a leading edge gene in a Gene Set Enrichment analysis analyzing subjects’ level of response to alcohol (Joslyn et al., 2010). Two
SNPs in GRM7 were chosen in the present study to test for association with AC: rs3749380, based on previous associations with schizophrenia (Ohtsuki et al., 2008); and rs1485175 as a negative control.

The goal of the current study was two-fold: 1) to follow up on previous studies of GRM7 by examining two common single nucleotide polymorphisms (SNPs) in GRM7 for association with AC in two independent samples, and 2) to use a gene-based approach to test all available common SNPs in GRM7 for association with AC. GRM7 is roughly 833kb in length (Kent et al., 2002). The expanse of the gene makes a comprehensive examination of which genetic signals may be associated with alcohol behaviors difficult. For that reason the initial SNP analysis was followed up using a more extensive approach by investigating all common variants available in existing dbGAP (Mailman et al., 2007) datasets.

Materials and methods

Samples

For the single SNP analyses two independent samples were used. The discovery samples consisted of 1,803 non-Hispanic European American (EA) subjects and 1,159 Hispanic subjects drawn from the Colorado Center on Antisocial Drug Dependence (CADD). Currently in it's 3rd wave of data collection, the CADD is a longitudinal study. It consists of 4 separate samples. The Family Study (FS) is a sample of clinical adolescent probands, ascertained while in treatment for antisocial drug dependence, their family members, and a matched set of control families (Stallings et al., 2005; Stallings et al., 2003). The Colorado Adoption Project (CAP) (Petrill SA, 2003), the Colorado Longitudinal Twin Study (CLTS) (Rhea et al., 2006), and the Colorado Community Twin Study (CCTS) (Rhea et al., 2006) are community unselected samples that were used in this study for phenotypic standardization (see below). Only subjects from the FS were included in this molecular genetic study. Ascertainment of this sample is described as follows; clinical probands were recruited and selected from individuals who had
consecutive admissions to treatment facilities in the Denver area between February of 1993 and June of 2001. Control subjects were recruited from the community and matched to the clinical probands based on age, gender, ethnicity, and zip code. All family members living in the same household as the proband were asked to participate in the study, which created family-based data. To increase similarity between this study and findings by Vadasz et al. (Vadasz et al., 2007), the most recent wave of data collection, representing probands past the age of initiation, was used; if the subject had not been assessed at wave 3, data from wave 2 were used, and if there were no data from wave 2, wave 1 data were used. Specifically, the work by Vadasz and colleagues identifying Grm7 as a candidate gene for AC utilized adult mice (Vadasz et al., 2007). The first wave of data collection from the CADD recruited adolescents aged 13-19; each subsequent wave of data collection began roughly 5-7 years after the previous one. Thus, using the most recent wave of data collection ensured that most subjects included in this analysis were adults. When assessed, buccal cell DNA was collected from subjects who provided informed voluntary consent. All recruitment, assessment, and DNA collection procedures were approved by the University of Colorado’s IRB and are in accordance with the Helsinki Declaration of 1975, as revised in 1983.

Families of EA and Hispanic subjects from the Genetics of Antisocial Drug Dependence (GADD) were assessed, as previously described (Kamens et al., 2013) to provide an independent replication sample. One thousand forty nine EA and 516 Hispanic subjects were used in the analysis. The GADD sample is a high-risk sample; probands, aged 14-19 years old, from Denver, CO, and San Diego, CA, were identified from treatment programs, involvement with the criminal justice system, or special schools. Probands were required to meet at least one of the criteria for both a substance use disorder (other than nicotine dependence) and conduct disorder. Siblings of the proband, as well as one or both biological parents, were recruited. The GADD is a longitudinal study in its 2nd wave of data collection; as in the CADD, data from wave 2 were preferentially used, and data from wave 1 were included if the subject
did not have wave 2 data. DNA was acquired with consent through either buccal cells or blood. All procedures are in accordance with the Helsinki Declaration of 1975, as revised in 1983. The University of California and the University of Colorado IRBs approved all subject recruitment, assessment, and DNA collection procedures.

For the gene-based test, data for EAs from four Genome-Wide Association Studies (GWAS) were collected: The Study of Addition: Genetics and Environment (SAGE) (Bierut et al., 2010a), Genome-wide association study of Alcohol use and alcohol use disorder in Australian Twin-families (OZ-ALC) (Grant et al., 2007; Knopik et al., 2004; Saccone et al., 2007b), the Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al., 2002), and a GWAS comprised of unrelated EA subjects from the CADD and GADD samples (CADD-GADD) (Derringer et al., 2015). As mentioned above, the CADD and GADD samples include probands targeted for drug use, their families and matched control families, as well as community twin samples. MESA is a community-based sample, whereas SAGE and OZ-ALC were ascertained based on substance dependence. All GWAS datasets were imputed to Phase 1 of the 1,000 Genomes data (version 3) (Abecasis et al., 2012) and were cleaned using standard quality control procedures. These included removing variants with low imputation accuracy (< 0.9), low minor allele frequency (MAF < 0.01), and variants out of Hardy-Weinberg equilibrium (HWE, p < 0.001).

**SNP selection**

For the single SNP analysis, two SNPs were selected based on MAF and location; rs3749380 and rs1485175 are common SNPs (MAF > 5%), have been previously examined in GWAS, and code for synonymous changes in exons 1 and 8, respectively. Additional SNPs could not be examined due to budgetary constraints. Preliminary examination of the haplotype structure of GRM7 in Hapmap (http://hapmap.ncbi.nlm.nih.gov/) using Haploview (Barrett et al., 2005) revealed that these two SNPs were not in linkage disequilibrium (LD) at $r^2 = 0.0$. Although
rs3749380 did not tag any other SNPs in the region, rs1485175 was indicated to tag eight other SNPs in the gene.

For the gene-based test, SNPs were annotated to GRM7 using the hg19 build. Variants 20kb up- and downstream of the gene were included, as evidence has suggested most genetic variation lies in this region (Veyrieras et al., 2008). Only variants present in all four GWAS were included in the final analysis.

**Genotyping**

SNP genotyping for rs3749380 and rs1485175 was performed using the TaqMan® assays for allelic discrimination according to manufacturer’s instructions (Applied Biosystems, Foster City, CA). Two thousand four hundred and ninety six subjects from CADD and 3,072 subjects from GADD were genotyped. Polymerase Chain Reaction (PCR) reactions were performed with the Biomek® 3000 Laboratory Automation Workstation (Beckman Coulter Inc, Brea, CA) and the Dual 384-Well GeneAmp® PCR system 9700 (Applied Biosystems, Foster City, CA). A 7900 Real-Time PCR System (Applied Biosystems, Foster City, CA) was used to analyze the amplified plates. Taking into account all the SNPs that have been previously genotyped in the CADD (33) and GADD (12), SNP genotypes for subjects with overall call rates <90% were excluded. Three hundred and eighty four samples were genotyped twice for each SNP to determine concordance between replicate reactions; the percent of discordant calls for SNPs in the CADD and GADD, respectively, were as follows: rs3749380 (0.78%, 0.00%), rs1485175 (1.04%, 0.52%). Genotype clusters from the amplified plates were auto-called by the Applied Biosystems TaqMan® Genotyper software (Applied Biosystems, Foster City, CA). Genotype calls were also visually examined by two independent lab personnel to confirm calls. Final calls were determined when both laboratory personnel agreed; if they did not agree genotypic data were excluded. Mendelian errors were identified using FBAT (Rabinowitz et al., 2000). If Mendelian errors were detected, the SNP genotype for that family was removed.
Using previous SNPs genotyped in the CADD sample (in the α4 (N = 10) and β2 (N = 10) subunits of the nicotinic acetylcholine receptor genes) 3 families in the CADD sample were identified as having Mendelian errors in 3 or more SNPs in both genes; the SNP genotypes for these families were also removed. Using Haploview (Barrett et al., 2005), pairwise linkage disequilibrium (LD) ($r^2$), MAF, and HWE were evaluated.

**Statistical analysis**

*Phenotypic descriptive*

For the single SNP analysis the phenotype examined was AC, defined here as the maximum number of drinks consumed in 24 hours over one’s lifetime, as assessed with the Composite International Diagnostic Interview – Substance Abuse Module (CIDI-SAM) (Cottler et al., 1990). Only phenotypic data for subjects who had previously used alcohol were included; in other words, if a subject had never had a drink of alcohol they were excluded from the analysis. AC was standardized based on the distribution of AC in the CADD community sample (CAP, CLTS, and CCTS samples, described above). A linear regression was performed using Statistical Analysis System (SAS) 9.3 software (SAS Institute Inc., Cary, NC) to determine residuals from sex, age, and age squared. The coefficients from the CADD community sample were applied to the CADD clinical sample (i.e. Z scores of clinical subjects were expressed as deviations from the means of the community samples), as well as the GADD sample.

Although AC was defined differently in each GWAS obtained from dbGAP, all attempts were made to harmonize the phenotype between samples. As the CADD-GADD GWAS is longitudinal, phenotypes were collected the same as for the SNP study; data were preferentially drawn from wave 3, from wave 2 if the subject was not assessed at wave 3, and from wave 1 if the subject did not have data at either later wave. In the CADD-GADD, as well as SAGE and OZ-ALC samples, AC was defined as the maximum number of drinks consumed in 24 hours over one’s lifetime and assessed with the Composite International Diagnostic Interview –
Substance Abuse Module (CIDI-SAM) (Cottler et al., 1990) and versions of the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA) (Bucholz et al., 1994), respectively. All subjects from these samples were included since information about whether each subject had tried alcohol was not always available. AC was defined in MESA as the largest number of drinks consumed in 24 hours over the past month in current drinkers (non-current drinkers were not asked the question). AC was log transformed in order to normalize the phenotype. Age and sex, as well as the first 5 principal components to control for ancestry, were used as covariates in the analysis. As the SAGE sample is comprised of three sub studies with slightly different ascertainment, study was also included as a covariate.

**Statistical genetics analysis**

Rs3749380 and rs1485175 were tested for association using an additive genetic model in a family-based association test performed in FBAT (Rabinowitz et al., 2000). FBAT builds on the original Transmission Disequilibrium Test (TDT) (Spielman et al., 1993), where alleles transmitted to affected offspring are compared to the expected distribution of alleles among offspring under Mendel’s law of segregation and conditioning. For results with $p < 0.05$, correction for multiple testing was implemented using the Min-p test in FBAT, which accounts for correlation between SNPs. This test provides a $p$-value for the most significant result based on the number of SNPs tested; if this $p$-value is less than 0.05, the result is considered significant after correction for multiple SNP testing at $p < 0.05$. This correction does not account for testing of multiple ethnic groups. In order to contrast findings using a different family based test, analyses were repeated using the within-family association and linkage test implemented in Quantitative Transmission Disequilibrium Test (QTDT) (Abecasis et al., 2000). Correction for multiple testing was implemented in QTDT, where the output shows the risk allele and Bonferroni significance level for the most significant association based on the number of SNPs tested. As with the FBAT correction, this correction does not adjust for multiple ethnic groups.
tested. While both tests examine the transmission of alleles within a family, thus controlling for population stratification, QTDT uses regression-based framework to test for association where the phenotype is the outcome. Conversely, FBAT uses the genotype as the outcome; using the additive model, the outcome is allele count for each offspring.

The gene-based test was carried out using two analytical approaches. Joint Association of Genetic Variants (JAG; available at http://ctglab.nl/software/jag/) (Lips) uses raw data and invokes PLINK (Purcell et al., 2007) to run a genome-wide association on each SNP in the gene. By summing the \(-\log_{10}\) of each SNP p-value a multivariate test statistic is created. Permutations of the phenotype are run and summed as described above. The empirical p-value for each gene \(P_{\text{emp}}\) is the number of times the sum of the \(-\log_{10}\) of the permuted p-value exceeds or equals the sum of the \(-\log_{10}(p\text{-value})\) from the genome-wide association. The second analytical approach used to confirm findings from the JAG analyses was the set-test implemented in PLINK (Purcell et al., 2007). Briefly, for each gene PLINK determines which SNPs are in LD above a certain threshold \(r^2 = 0.5\), performs a standard SNP regression, and selects up to \(N\) “independent” SNPs with \(p < 0.05\). From this subset, the statistic for each gene is calculated as the mean of these SNP statistics. Keeping LD between SNPs constant, the phenotype is permuted. The empirical p-value \(P_{\text{emp}}\) is the number of times the permuted gene statistic exceeds the original statistic for that gene. The \(P_{\text{emp}}\) for each sample from each JAG and PLINK analysis were combined using the weighted Z-score method (Whitlock, 2005).

**Results**

**SNP analysis**

The phenotypic characteristics between the CADD and GADD samples were comparable (Table 1).
Although the GADD sample is somewhat younger than the CADD sample, AC and AD means and distributions were comparable between the clinical samples. Although the results shown pertain to subjects who have used alcohol, the analyses were repeated using all available subjects and the results were comparable.

The minor alleles and MAF for rs3749380 and rs1485175 were similar between EA and Hispanic subjects in each sample (Table 2).

### Table 1. Demographic characteristics of the CADD and GADD samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>EAs</th>
<th>Hispanics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Age (^a)</td>
</tr>
<tr>
<td>CADD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community Adolescents</td>
<td>349</td>
<td>24.4±5.3</td>
</tr>
<tr>
<td>Adults</td>
<td>399</td>
<td>48.3±6.3</td>
</tr>
<tr>
<td>Clinical Adolescents</td>
<td>596</td>
<td>24.4±6.2</td>
</tr>
<tr>
<td>Adults</td>
<td>458</td>
<td>46.4±9.0</td>
</tr>
<tr>
<td>GADD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical Adolescents</td>
<td>843</td>
<td>21.4±4.1</td>
</tr>
<tr>
<td>Adults</td>
<td>206</td>
<td>45.4±6.0</td>
</tr>
</tbody>
</table>

\(^a\) Mean ± standard deviation  
\(^b\) Alcohol consumption; mean ± standard deviation

### Table 2. Genotypic characteristics of EA and Hispanic subjects in the CADD and GADD samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>SNP</th>
<th>Location (^a)</th>
<th>Alleles (^b)</th>
<th>MAF (^c)</th>
<th>HWE (^d)</th>
<th>Alleles</th>
<th>MAF</th>
<th>HWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CADD</td>
<td>rs3749380</td>
<td>6903047</td>
<td>C/T</td>
<td>0.40</td>
<td>0.36</td>
<td>C/T</td>
<td>0.40</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>rs1485175</td>
<td>7620539</td>
<td>C/T</td>
<td>0.45</td>
<td>0.88</td>
<td>C/T</td>
<td>0.41</td>
<td>0.32</td>
</tr>
<tr>
<td>GADD</td>
<td>rs3749380</td>
<td>6903047</td>
<td>C/T</td>
<td>0.38</td>
<td>0.74</td>
<td>C/T</td>
<td>0.42</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>rs1485175</td>
<td>7620539</td>
<td>C/T</td>
<td>0.45</td>
<td>0.19</td>
<td>C/T</td>
<td>0.46</td>
<td>0.09</td>
</tr>
</tbody>
</table>

\(^a\) On chromosome 3, from the UCSC genome browser  
\(^b\) Major/minor allele  
\(^c\) Minor allele frequency  
\(^d\) Hardy-Weinberg equilibrium p-value
The discordant genotype call rate was lower for both SNPs in the GADD than in the CADD sample because many of the DNA samples for the GADD were drawn from blood, as opposed to saliva as in the CADD, and DNA samples from blood generally have higher genotyping success (Philibert et al., 2008). The genotype for rs3749380 was removed for 22 and 8 families in the CADD and GADD, respectively, and in 24 and 8 families for rs1485175 in the CADD and GADD, respectively, after Mendelian errors were detected in these families. There were no significant deviations from the HWE at p < 0.05. Furthermore, both SNPs represent independent signals in GRM7 and were not correlated ($r^2 = 0$).

In the EA group from the CADD sample, rs3749380 was significantly associated with AC at p = 0.010 and p = 0.002 using FBAT and QTDT, respectively (Table 3).

Table 3. Results from the FBAT and QTDT analyses in the CADD.

<table>
<thead>
<tr>
<th>Group</th>
<th>SNP</th>
<th>Risk allele</th>
<th>Z-score</th>
<th>P-value</th>
<th># Fam</th>
<th>SNP</th>
<th>P-value</th>
<th># Probands</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAs</td>
<td>rs3749380</td>
<td>T</td>
<td>2.574</td>
<td>0.010$^d$</td>
<td>115</td>
<td>rs3749380</td>
<td>0.002$^a$</td>
<td>247</td>
</tr>
<tr>
<td></td>
<td>rs1485175</td>
<td>C</td>
<td>0.667</td>
<td>0.505</td>
<td>123</td>
<td>rs1485175</td>
<td>0.722</td>
<td>261</td>
</tr>
<tr>
<td>Hispanics</td>
<td>rs3749380</td>
<td>T</td>
<td>1.806</td>
<td>0.071</td>
<td>78</td>
<td>rs3749380</td>
<td>0.060</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>rs1485175</td>
<td>T</td>
<td>0.720</td>
<td>0.472</td>
<td>75</td>
<td>rs1485175</td>
<td>0.304</td>
<td>143</td>
</tr>
</tbody>
</table>

$^a$ A positive z-score indicates the allele is over-transmitted and is interpreted as the risk allele.

$^b$ The number of informative families included in the analysis.

$^c$ The number of informative probands included in the QTDT test.

$^d$ Significant after correction for multiple testing at p = 0.020.

$^e$ Overall Bonferroni significance level for this association is p = 0.031.

These associations survived correction for multiple testing with the Min-p test p-value of 0.020 for FBAT, and Bonferroni significance level of 0.031 in QTDT. The T allele was the risk allele. In Hispanics, rs3749380 was nominally associated with AD at p = 0.071 and 0.060 using FBAT and QTDT, respectively (Table 3). Again the T allele was the risk allele.

Contrary to the results in the CADD sample, neither rs3749380 nor rs1485175 was associated with AC in EAs from the GADD. However, the T allele of rs1485175 was nominally associated with increased risk of AC at p = 0.035 in Hispanics (Table 4).
Table 4. Results from the FBAT and QTDT analyses in the GADD.

<table>
<thead>
<tr>
<th>Group</th>
<th>SNP</th>
<th>Risk allele</th>
<th>Z-score(^a)</th>
<th>P-value</th>
<th># Fam(^b)</th>
<th>SNP</th>
<th>P-value</th>
<th># Probands(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAs</td>
<td>rs3749380</td>
<td>C</td>
<td>1.499</td>
<td>0.134</td>
<td>133</td>
<td>rs3749380</td>
<td>0.705</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>rs1485175</td>
<td>C</td>
<td>1.311</td>
<td>0.190</td>
<td>133</td>
<td>rs1485175</td>
<td>0.804</td>
<td>59</td>
</tr>
<tr>
<td>Hispanics</td>
<td>rs3749380</td>
<td>T</td>
<td>0.558</td>
<td>0.577</td>
<td>77</td>
<td>rs3749380</td>
<td>NT(^d)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>rs1485175</td>
<td>T</td>
<td>2.111</td>
<td>0.035(^e)</td>
<td>73</td>
<td>rs1485175</td>
<td>NT(^d)</td>
<td>-</td>
</tr>
</tbody>
</table>
\(^a\) A positive z-score indicates the allele is over-transmitted and is interpreted as the risk allele.
\(^b\) The number of informative families included in the analysis.
\(^c\) The number of informative probands included in the QTDT test.
\(^d\) Not Tested due to limited number of probands.
\(^e\) Not significant after correction for multiple testing at p = 0.070.

This association did not survive correction for multiple testing after the Min-p test p-value of 0.070. Results using QTDT confirmed the lack of findings with EAs in the GADD sample; due to the limited number of probands, QTDT was unable to perform an association using Hispanic subjects in the GADD (Table 4).

**Gene-based test**

Phenotypic characteristics for the samples utilized in the gene-based test are given in Table 5.

Table 5. Phenotypic characteristics of the samples utilized in the gene-based test.

<table>
<thead>
<tr>
<th>Study</th>
<th>N</th>
<th>Age(^a)</th>
<th>% Male</th>
<th>AC (raw)(^b)</th>
<th>AC (log)(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CADD-GADD</td>
<td>1031</td>
<td>23.81±4.25</td>
<td>70.71%</td>
<td>5.17±6.91</td>
<td>0.72±0.32</td>
</tr>
<tr>
<td>MESA</td>
<td>897</td>
<td>62.51±10.16</td>
<td>54.18%</td>
<td>2.63±2.33</td>
<td>0.62±0.19</td>
</tr>
<tr>
<td>OZ-ALC</td>
<td>2000</td>
<td>44.03±9.53</td>
<td>50.25%</td>
<td>18.79±14.41</td>
<td>1.21±0.32</td>
</tr>
<tr>
<td>SAGE</td>
<td>2247</td>
<td>38.19±9.60</td>
<td>44.68%</td>
<td>19.47±18.87</td>
<td>1.21±0.32</td>
</tr>
</tbody>
</table>
\(^a\) Mean ± standard deviation.
\(^b\) Raw values for AC; mean ± standard deviation.
\(^c\) Log\(_{10}\) transformed AC used in the analysis; mean ± standard deviation.

For the analysis, a random subset of unrelated subjects was selected from the OZ-ALC study. The OZ-ALC and SAGE samples had expectedly high scores for AC since these samples were ascertained based on alcohol and other drug use. The community based MESA sample had low
AC reporting. Finally, the CADD-GADD GWAS sample showed slightly higher AC values than MESA but markedly lower values than those samples ascertained on alcohol and drug use only.

Nine hundred and two SNPs in GRM7 were tested for association with AC using a gene-based test. Results indicated that variation in GRM7 is not associated with AC with combined \( p = 0.743 \) using the JAG test (Table 6).

### Table 6. Gene-based test results from the JAG analysis.

<table>
<thead>
<tr>
<th>Study</th>
<th>sumlogReal(^a)</th>
<th>( P_{\text{emp}} )</th>
<th>var(Perms)(^b)</th>
<th>mean(Perms)(^c)</th>
<th>nEff(^d)</th>
<th>Combined p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CADD-GADD</td>
<td>321.51</td>
<td>0.750</td>
<td>9921.65</td>
<td>391.92</td>
<td>15</td>
<td>0.743</td>
</tr>
<tr>
<td>MESA</td>
<td>275.26</td>
<td>0.916</td>
<td>9441.08</td>
<td>392.15</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>OZ-ALC</td>
<td>333.77</td>
<td>0.692</td>
<td>9567.68</td>
<td>390.33</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>SAGE</td>
<td>410.65</td>
<td>0.365</td>
<td>9405.54</td>
<td>391.37</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) The test statistic (\( \Sigma \log_{10}(P) \)) of the original data

\(^b\) The variance of the distribution of the test statistic from the permutations

\(^c\) The mean of the test statistics from the permutations

\(^d\) The effective number of SNPs

This result was replicated with combined \( p = 0.812 \) using the PLINK test (Table 7).

### Table 7. Gene-based test results from the PLINK analysis.

<table>
<thead>
<tr>
<th>Study</th>
<th>nSIG(^a)</th>
<th>iSIG(^b)</th>
<th>( P_{\text{emp}} )</th>
<th>SNPs(^c)</th>
<th>Minor allele</th>
<th>( \beta )</th>
<th>P-value</th>
<th>Combined p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CADD</td>
<td>16</td>
<td>5</td>
<td>0.601</td>
<td>rs3804849(^d)</td>
<td>G</td>
<td>0.040</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs4054135(^d)</td>
<td>C</td>
<td>0.035</td>
<td>0.014</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs1356268</td>
<td>G</td>
<td>-0.075</td>
<td>0.026</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs3804826(^a)</td>
<td>A</td>
<td>0.028</td>
<td>0.044</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs9844393</td>
<td>G</td>
<td>0.036</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>MESA</td>
<td>2</td>
<td>2</td>
<td>0.809</td>
<td>rs9828691</td>
<td>G</td>
<td>0.029</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs7611549</td>
<td>A</td>
<td>0.018</td>
<td>0.036</td>
<td></td>
</tr>
<tr>
<td>OZ-ALC</td>
<td>8</td>
<td>2</td>
<td>0.852</td>
<td>rs162773(^a)</td>
<td>C</td>
<td>-0.019</td>
<td>0.020</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs3804861</td>
<td>T</td>
<td>0.022</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs9872244(^f)</td>
<td>G</td>
<td>0.029</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs28437102(^f)</td>
<td>G</td>
<td>0.036</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs9813021</td>
<td>A</td>
<td>0.041</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs28378912</td>
<td>A</td>
<td>0.019</td>
<td>0.021</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs17693618</td>
<td>G</td>
<td>0.030</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs7634808</td>
<td>G</td>
<td>0.019</td>
<td>0.032</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs7634694</td>
<td>G</td>
<td>0.017</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs1356268</td>
<td>G</td>
<td>-0.043</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>SAGE</td>
<td>71</td>
<td>8</td>
<td>0.462</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Total number of SNPs below p-value threshold (\( p < 0.05 \))

\(^b\) Number of significant SNPs also passing LD criteria (\( r^2 > 0.5 \))
The correlation between the JAG and PLINK p-values for each study was high ($r = 0.76$).

**Discussion**

The goals of this study were to evaluate the evidence for association of *GRM7* with AC based on findings that *Grm7* is a candidate gene for AC in a mouse model (Vadasz *et al.*, 2007). To our knowledge, this is one of the first studies to both examine individual polymorphisms in *GRM7* as well as perform a gene-based test for association with AC. Although one prior study has identified a differentially inherited SNP module, a collaboration of SNPs contributing to the disease synergistically that is differentially inherited in affected and unaffected sibpairs, in *GRM7* in alcoholics, the two SNPs in the module were in introns and not in LD with the SNPs examined in the present study (Li *et al.*, 2011). Our gene-based test is an important extension of this work and represents an independent study because none of the four previously studied SNPs were included in our gene-based test as they were not present in all four GWAS utilized.

Consistent with our hypothesis based on these two SNPs selected from the schizophrenia study (Ohtsuki *et al.*, 2008), the T allele of rs3749380 was associated in EAs with AC in the CADD sample using FBAT and QTDT while the negative control SNP rs1485175 showed no evidence for association. Neither SNP was associated with the phenotype in the EA GADD sample. This discrepancy could be due to lack of power to detect true associations, or undetected differences between the CADD and GADD samples. The GADD sample did not include any control families, so this difference in study design may be reflected in differential transmission effects. Since this SNP has not been examined in any previous drug behavior studies, its role in alcohol behaviors cannot be compared to others; however, the T allele of
rs3749380 was nominally associated with schizophrenia in two independent Japanese populations, and was shown to have lower promoter activity than the C allele using a luciferase assay (Ohtsuki et al., 2008). In addition, this SNP has been associated with panic disorder in a Japanese population (Otowa et al., 2009).

Rs1485175 was nominally associated with AC in Hispanics from the GADD sample and rs3749380 was trending toward association with AC in Hispanics from the CADD sample. These associations do not survive correction for multiple testing and the lack of congruity between the samples suggests lower power to detect true associations and heterogeneity in the Hispanic sample; however, these results do not rule out a role for GRM7 in risk for alcohol behaviors in Hispanics. In all analyses, the T allele was over-transmitted to affected offspring. While rs1485175 has been included in several GWAS of schizophrenia (Ganda et al., 2009; Ohtsuki et al., 2008; Shibata et al., 2009), it was not significantly associated with schizophrenia, nor has it previously been associated with alcohol behaviors.

A gene-based test was run to assess whether collapsing variants of potentially small effect and testing them together may provide evidence for association between GRM7 and AC. Two gene-based tests were run on each of four GWAS datasets and the p-values from each analysis were combined; no association was found between GRM7 and AC using either analytical method. Using JAG, the effective number of SNPs tested was calculated, a measure that takes LD into account to estimate how many overall signals were tested. In the MESA, OZ-ALC, and SAGE samples 16 signals were included; in the CADD-GADD sample, 15 signals were included. This difference is likely caused by the samples having different underlying LD structure. Notably, the two SNPs previously examined were not included in the gene-based test. Although the gene-based test encompasses more genetic variation than single SNP studies, the exclusion of rs3749380 and rs1485175 highlight the limitation of focusing only on current GWAS to perform human genetic studies and boosts the importance of single SNP studies.
While two analytical methods were used for both the SNP analysis and gene-based test and results were consistent, there are some differences between the two methods worth noting. In the single-SNP analyses, both FBAT and the within-family QTDT association test control for population stratification since they use a family-based approach. However, the QTDT test employs a regression-based approach where the phenotype is the outcome, which may explain the slightly different results obtained using the two methods. Under a normally distributed phenotype with no stratification, QTDT can be more powerful. However, our Hispanic sample is highly admixed, which could lead to erroneous results. Nevertheless, for testing a continuous trait in the entire sample with a phenotype of mean 0, heritability of 0.1 and frequency of 0.3, both tests can have roughly 0.75 power (Lange et al., 2002). For the gene-based tests, JAG and PLINK consolidate SNP associations throughout a gene, but JAG takes into account every SNP and performs best when there are many signals in the gene contributing to risk for the phenotype. Conversely, PLINK selects only the most significant signals to contribute to the test statistic and therefore performs better with a gene with a few very strong signals.

There are several limitations that must be considered while interpreting our findings. Specifically, there may be low power to detect and replicate genetic SNP associations of very small effect in the CADD and GADD samples, which could lead to false negatives. Also, the multiple testing correction only adjusted for the number of SNPs tested, not the number of phenotypes or samples tested. There might exist unidentified ascertainment or sample characteristic differences between the CADD and GADD samples, or the GWAS samples utilized. In the CADD and GADD samples, selective attrition could result in bias, particularly if subjects who did not return at wave 2 or 3 failed to do so because of incarceration or confinement in a treatment facility for drug use. The phenotypic measure of AC has poor distributional properties, and duration was not taken into account; future studies may benefit from imposing a duration limit or averaging typical and peak drinking levels to get a smoother distribution. This study also did not consider gene by SNP or SNP by SNP interactions.
There are several strengths of our study that help mitigate some of these limitations. In particular, the use of a family-based design protects against population stratification so should minimize false positive associations. Furthermore, our SNP analysis included an important replication sample that was recruited and assessed in a manner similar to the first sample. Although it is likely that neither of the two SNPs examined is a causal variant, a gene-based test was used to try to cover most of the variation in the gene. However, we only analyzed SNPs that were common between all four GWAS (in order to combine p-values across datasets), potentially ignoring important signals. Despite this limitation, the gene-based test represents an effort to characterize all common variation in GRM7 in relation to alcohol behaviors using currently available datasets. The use of several samples resulted in a large overall sample size. SNP-set based approaches can increase power to detect genetic loci with individually small effects by consolidating SNP associations (Purcell et al., 2009; Wang et al., 2011).

Although this study focused on GRM7, there are several other glutamate metabotropic receptors implicated in alcohol behaviors that warrant discussion. Activation of glutamate metabotropic receptor 5 (Grm5) has direct excitatory effects and potentiates N-methyl-D-aspartate receptor currents (Awad et al., 2000). In addition, Grm5 antagonists have been shown to decrease ethanol seeking and relapse (Backstrom et al., 2004) and self-administration (Cowen et al., 2005; Schroeder et al., 2005) in rats, as well as in mice (Hodge et al., 2006). This is the opposite effect seen with Grm7, where overexpression leads to decreased ethanol consumption and knocking out the gene results in increased alcohol consumption (Gyetvai et al., 2011). Although the association was not replicated, GRM5 was also associated with AD in a study of adolescents assessed for risky drinking (Schumann et al., 2008). In addition, several SNPs in GRM8 have been associated with alcohol dependence (Chen et al., 2009a). Recently, a pre-mature stop codon in the glutamate metabotropic receptor 2 (Grm2) gene was linked to increased alcohol consumption and preference in rats. Pharmacological blockade of the receptor further increased alcohol consumption in rats, and Grm2 knockout mice displayed
elevated alcohol consumption (Zhou et al., 2013), paralleling the effect seen with Grm7 knockout mice (Gyetvai et al., 2011). Although this provides evidence that Grm2 mediates alcohol behaviors, the total loss of function of Grm2 seen in rodent models is most likely rare in human populations. Regardless, GRM5, GRM8, and GRM2 are interesting candidates for alcohol treatment (Holmes et al., 2013; Joslyn et al., 2010).

In conclusion, this work provides preliminary evidence for an association between rs3749380 and AC in EAs. However, this association did not replicate in an independent sample and a gene-based test did not support an association between GRM7 and AC. Importantly, the gene-based test did not include rs3749380 or rs1485175. This emphasizes the importance of single SNP analyses as current GWAS datasets often lack potentially important signals. The availability of an independent replication sample and publically available GWAS on dbGaP, as well as the use of two different analytical tools, allowed us to weigh findings in the context of their reproducibility across samples and between analyses. Although our results do not provide support for further examination of common variation in GRM7 in human alcohol behaviors, it remains possible that variation in regulatory regions far upstream from the gene or rare variants may play a role in risk for AC. As mentioned above, Grm5, GRM8, and Grm2 have also been implicated in alcohol consumption (Awad et al., 2000; Backstrom et al., 2004; Chen et al., 2009a; Cowen et al., 2005; Hodge et al., 2006; Schroeder et al., 2005; Zhou et al., 2013), suggesting that the glutamate system is a likely target for alcohol cessation therapies and emphasizes the importance of continued investigation of these genes in alcohol behaviors.
CHAPTER IV

EXAMINATION OF THE INVOLVEMENT OF CHOLINERGIC-ASSOCIATED GENES IN NICOTINE BEHAVIORS IN EUROPEAN AND AFRICAN AMERICANS

This work is in preparation

Introduction

Cigarette smoking is a harmful and detrimental habit. Between the years of 2000 and 2004, cigarette smoking and exposure to tobacco resulted in at least 443,000 premature deaths and $96.8 billion in productivity losses annually in the United States (Prevention, 2008). As of May 2012, the World Health Organization estimated that 6 million people die per year due to tobacco smoke, and, unless the rate of smoking declines, the annual death toll as a consequence of smoking will exceed 8 million by 2030.

Nicotine has long since been considered the major addictive component in tobacco smoke (Gunby, 1988; Stolerman et al., 1995) and exerts its effect by binding nicotinic acetylcholine receptors (nAChRs, encoded by the CHRN genes) in the peripheral and central nervous systems (Gotti et al., 2009). Accordingly, numerous studies have identified polymorphisms in several CHRN genes associated with nicotine dependence (ND) and cigarettes per day (CPD). The most well-replicated of these associations lies within a cluster of 3 CHRN genes on chromosome 15q25 (CHRNA3-A5-B4) (Berrettini et al., 2012), although several other receptors have been associated with these phenotypes. Most notably, CHRNB3-A6 on chromosome 8p11 has also been consistently associated with CPD (Thorgeirsson et al., 2010) and ND (Bierut et al., 2007; Haller et al., 2012; Hoft et al., 2009b; Rice et al., 2012; Saccone et al., 2009a; Saccone et al., 2010b; Saccone et al., 2007a). Other CHRN genes associated with CPD and ND include CHRND-G, CHRN1B, CHRNA10, CHRNA4, and CHRNB2.
Twin studies have estimated that ND and smoking quantity are roughly 56-72% and 51-61% heritable in men and women, respectively (Broms et al., 2006; Kendler et al., 1999; Lessov et al., 2004; True et al., 1999; Vink et al., 2004), yet associations between the CHRN genes and ND and CPD account for a very small proportion of the variance in smoking (roughly 1 cigarette per day) (Berrettini et al., 2012; Saccone et al., 2010b). Based on work that provides evidence that common single nucleotide polymorphisms (SNPs) explain a large proportion of the heritability in height, Crohn’s disease, bipolar disorder, and type I diabetes (Lee et al., 2011; Yang et al., 2010), we hypothesize that common SNPs also explain a proportion of the variation in smoking quantity and ND. The goal of this study was to utilize set-based approaches to test for association between a targeted group of SNPs and nicotine behaviors. SNP-set based approaches can increase power to detect genetic loci with individually small effects by consolidating SNP associations, and help to prioritize associations based on biological relevance (Purcell et al., 2009; Wang et al., 2011).

Upon chronic exposure to nicotine, nAChRs undergo an upregulation of receptor number (Benwell et al., 1988; Marks et al., 1983; Schwartz et al., 1983) that is independent of upregulation of CHRN mRNA (Marks et al., 1992). Although there are many theories as to how this upregulation occurs, including increased receptor trafficking (Darsow et al., 2005), decreased subunit degradation (Rezvani et al., 2009; Rezvani et al., 2007), increased nAChR subunit maturation and folding (Harkness et al., 2002; Nashmi et al., 2003; Sallette et al., 2005), and increased translation and 2nd messenger signaling (Gopalakrishnan et al., 1997), researchers have yet to test several of the proteins suggested to impact upregulation in concert. We posit that gene products that interact with nAChRs are involved in nAChR upregulation and thus play a role in the development of ND.
Through manual curation of the literature, and in conversation with experts in the field, we have assembled a list of genes that encode proteins known to interact with nAChRs or play a role in their downstream signaling (Figure 1).

![Figure 1. Cellular processes proposed to play a role in nAChR upregulation. The number of genes identified that play a role in each process is shown in parentheses (several genes were proposed to be involved in more than one cellular process). ER: endoplasmic reticulum.](image)

The goal of this study was to assess the evidence for association with each gene and CPD and ND in African and European Americans (AA and EA, respectively). Although many studies use smoking quantity as a proxy for ND, research has demonstrated unique genetic effects on CPD, suggesting that a smoking quantity measure may not serve as a simple proxy for the genetic influences on ND (Haberstick et al., 2007). These analyses were carried out in several Genome-Wide Association Studies (GWAS) and results were later combined as a meta-analysis.
Materials and methods

Samples

Subjects with European (EA) and African ancestry (AA) were used from the following GWAS: the Study of Addiction: Genetics and Environment (SAGE) (Bierut et al., 2010b), the Genetic Epidemiology of COPD (COPDGene) (Regan et al., 2010), and an in-house GWAS comprised of unrelated Caucasian subjects from the Colorado Center on Antisocial Drug Dependence and the Genetics of Antisocial Drug Dependence (CADD-GADD) (Derringer et al., 2015). The SAGE study is comprised of 3 sub-studies ascertained on cocaine, alcohol, or nicotine dependence. The CADD-GADD sample is a mixture of various samples including probands targeted for drug use, their families and matched control families, as well as community twin samples. Finally, the COPDGene study consists of smokers who have smoked at least 10 pack-years. All GWAS were imputed to Phase 1 of the 1,000 Genomes dataset (version 3) (Abecasis et al., 2012) and were cleaned using standard quality control procedures: SNPs with low imputation accuracy ( < 0.9), low minor allele frequency (MAF) ( < 0.01), and out of Hardy-Weinburg equilibrium (HWE, p < 0.001) were excluded. We analyzed two phenotypes when available from each GWAS: quantitative ND symptoms as assessed by the Fagerström Test for Nicotine Dependence (FTND) (Heatherton et al., 1991); and CPD, binned as follows: 0-10 cigarettes per day, 11-20, 21-30, 31-40, and 41 or more. Only subjects who had smoked 100 cigarettes or more in their lifetime were included.

Analyses

SNPs were annotated to our genes of interest using the hg19 build. SNPs 20kb up- and downstream of each gene were included in the annotation, as evidence has suggested the majority of genetic variants that influence expression are located within 20kb of a gene (Veyrieras et al., 2008). For the analyses in EAs, only SNPs present in the SAGE, COPDGene, and CADD samples were included. Similarly, for the AA analyses, only SNPs present in the AA
SAGE and COPDGene samples were included. Age, sex, and the first 5 principal components from each GWAS were included as covariates in each analysis. Additionally, study was used as a covariate in the SAGE analysis. Ten thousand permutations were performed for each test.

**Joint Association of Genetic Variants (JAG; available at [http://ctglab.nl/software/jag/](http://ctglab.nl/software/jag/))** (Lips): JAG uses raw data and invokes PLINK (Purcell et al., 2007) to run a genome-wide association on each SNP included in the gene. A multivariate SNP test statistic is calculated by summing the $-\log_{10}$ of each SNP p-value and the empirical p-value for each gene is calculated by summing the $-\log_{10}(p$-value) for each permutation of the phenotype. The empirical p-value ($P_{\text{emp}}$) represents the number of times the $\Sigma -\log_{10}(p$-value) exceeds or equals the $-\log_{10}(p$-value) from the genome-wide association. $P_{\text{emp}}$ tests the hypothesis that the multivariate pattern of p-values of all SNPs in a gene is significantly different than what is expected under the null hypothesis of no association. This analysis was applied to our data in order to test the association between individual genes and CPD and FTND.

**PLINK (Purcell et al., 2007):** A secondary analytical approach was used to confirm findings from the analyses in JAG using the set-test implemented in PLINK, a whole genome association analyses toolkit. Briefly, for each gene, PLINK determines which SNPs are in linkage disequilibrium (LD) above a certain threshold $R$ (we used $r^2 = 0.5$) and perform a standard single SNP regression. Up to $N$ "independent" SNPs with p-values below 0.05 (user-defined) are selected for each gene. From this subset of SNPs, the statistic for each set is calculated as the mean of these single SNP statistics. The phenotype is permuted, keeping LD between SNPs constant. $P_{\text{emp}}$ is the number of times the permuted set-statistic exceeds the original one for that gene.

**Meta-analysis**

The $P_{\text{emp}}$ from each analysis using JAG and the $P_{\text{emp}}$ from each analysis using PLINK were combined using the weighted Z-score method (Whitlock, 2005). As the PLINK test only
includes SNPs with p < 0.05 in the creation of the test statistic, an automatic p-value of 1 is generated if a gene does not have any SNPs with p < 0.05. These p-values were changed to NA and excluded from the meta-analyses. Correction for multiple testing was implemented using the Bonferroni correction, adjusted solely for how many genes were tested, and not for the number of phenotypes or ethnic groups tested. Regardless, this number is likely to be conservative as several of the genes tested may be correlated.

Comparison with previous meta-analyses

Output from the PLINK set test also includes the SNPs that passed the p-value and LD thresholds in each gene (if any). The p-values of the significant SNPs in genes associated with CPD in EAs from the present study were compared to the p-values for those SNPs (if present) in the TAG consortium (2010) for CPD. Using SNP Annotation and Proxy Search (SNAP) (Johnson et al., 2008), pairwise r² between SNPs was determined using a sample of Utah residents with ancestry from northern and western Europe (CEU) from the 1000 Genomes Pilot 1 (1000G) data (Abecasis et al., 2012).

Results

Study and sample characteristics for the COPDGene, SAGE, and CADD-GADD samples are given in Table 1.
The COPDGene sample was the oldest, and the CADD-GADD the youngest. EAs in the COPDGene and SAGE had similar smoking patterns, as did AAs in the COPDGene and SAGE samples. The AAs had somewhat lower CPD and higher FTND than their EA counterparts.

**Cigarettes per day and nicotine dependence in EAs**

Fifteen thousand eight hundred and ninety three SNPs from our gene set were tested for association with CPD in EAs from the CADD-GADD, COPDGene, and SAGE samples. The p-values were combined using the weighted Z-score method. Eight genes had $p < 0.05$ using both JAG and PLINK: *CHRNA3, CHRNA5, CHRNB4, NCAM1, DAG1, DNAJA3, CHRNE,* and *MAPRE1* (Table 2).

### Table 1. Study and sample characteristics of the GWAS utilized.

<table>
<thead>
<tr>
<th>Study</th>
<th>Age* (N)</th>
<th>Sex (% male)</th>
<th>CPD*</th>
<th>Age (N)</th>
<th>Sex (% male)</th>
<th>FTND*</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA</td>
<td>COPDGene</td>
<td>62.09±8.84 (6670)</td>
<td>52.37%</td>
<td>25.84±11.44</td>
<td>57.50±7.86 (2568)</td>
<td>53.35%</td>
</tr>
<tr>
<td></td>
<td>SAGE</td>
<td>35.74±6.92 (1255)</td>
<td>37.83%</td>
<td>25.90±19.94</td>
<td>37.53±8.65 (1673)</td>
<td>43.93%</td>
</tr>
<tr>
<td></td>
<td>CADD</td>
<td>23.44±4.40 (588)</td>
<td>73.98%</td>
<td>17.96±11.54</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>AA</td>
<td>COPDGene</td>
<td>54.68±7.21 (3300)</td>
<td>55.94%</td>
<td>21.30±10.40</td>
<td>53.34±6.04 (2567)</td>
<td>57.78%</td>
</tr>
<tr>
<td></td>
<td>SAGE</td>
<td>39.30±6.81 (594)</td>
<td>47.47%</td>
<td>23.64±17.39</td>
<td>40.03±7.13 (779)</td>
<td>49.81%</td>
</tr>
</tbody>
</table>

*Mean ± standard deviation
Table 2. Results from the JAG and PLINK tests with CPD in EAs. The p-value for each test is shown for each study as well as the combined p-value for genes with p < 0.05 in either analysis.

<table>
<thead>
<tr>
<th>Gene</th>
<th>COPDGene</th>
<th>SAGE</th>
<th>CADD-GADD</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_j^a$</td>
<td>$P_p^b$</td>
<td>$P_j$</td>
<td>$P_p$</td>
</tr>
<tr>
<td>CHRNA3</td>
<td>&lt;1.00E-4</td>
<td>1.00E-4</td>
<td>0.085</td>
<td>NA</td>
</tr>
<tr>
<td>NCAM1</td>
<td>0.003</td>
<td>0.054</td>
<td>0.164</td>
<td>0.155</td>
</tr>
<tr>
<td>CHRNA5</td>
<td>0.001</td>
<td>5.00E-4</td>
<td>0.875</td>
<td>NA</td>
</tr>
<tr>
<td>DAG1</td>
<td>0.005</td>
<td>0.019</td>
<td>0.593</td>
<td>NA</td>
</tr>
<tr>
<td>DNAJA3</td>
<td>0.005</td>
<td>0.12</td>
<td>0.554</td>
<td>NA</td>
</tr>
<tr>
<td>CHRN4B</td>
<td>0.002</td>
<td>1.00E-4</td>
<td>0.948</td>
<td>NA</td>
</tr>
<tr>
<td>CHRNE</td>
<td>0.010</td>
<td>5.00E-4</td>
<td>0.180</td>
<td>0.152</td>
</tr>
<tr>
<td>MAPRE1</td>
<td>0.015</td>
<td>0.022</td>
<td>0.749</td>
<td>NA</td>
</tr>
<tr>
<td>EPHA4</td>
<td>0.056</td>
<td>0.275</td>
<td>0.039</td>
<td>0.413</td>
</tr>
<tr>
<td>UNC50</td>
<td>0.040</td>
<td>0.079</td>
<td>0.196</td>
<td>NA</td>
</tr>
<tr>
<td>LAMA5</td>
<td>0.032</td>
<td>0.167</td>
<td>0.405</td>
<td>NA</td>
</tr>
<tr>
<td>APC</td>
<td>0.046</td>
<td>0.082</td>
<td>0.190</td>
<td>0.208</td>
</tr>
<tr>
<td>PRNP</td>
<td>0.897</td>
<td>NA</td>
<td>0.008</td>
<td>0.015</td>
</tr>
<tr>
<td>LRP4</td>
<td>0.191</td>
<td>0.027</td>
<td>0.146</td>
<td>0.124</td>
</tr>
<tr>
<td>PPP3CA</td>
<td>0.480</td>
<td>NA</td>
<td>0.156</td>
<td>0.048</td>
</tr>
<tr>
<td>CHRN4B3</td>
<td>0.807</td>
<td>NA</td>
<td>0.017</td>
<td>0.018</td>
</tr>
<tr>
<td>CACNA2D1</td>
<td>0.105</td>
<td>0.015</td>
<td>0.960</td>
<td>0.740</td>
</tr>
<tr>
<td>SRC</td>
<td>0.787</td>
<td>NA</td>
<td>0.176</td>
<td>0.084</td>
</tr>
<tr>
<td>IQGAP1</td>
<td>0.035</td>
<td>0.040</td>
<td>0.775</td>
<td>NA</td>
</tr>
<tr>
<td>DNM1L</td>
<td>0.205</td>
<td>NA</td>
<td>0.024</td>
<td>0.041</td>
</tr>
<tr>
<td>SYT1</td>
<td>0.572</td>
<td>0.047</td>
<td>0.979</td>
<td>NA</td>
</tr>
</tbody>
</table>

$^a$ $P_{emp}$ for the JAG test
$^b$ $P_{emp}$ for the PLINK test
$^c$ Significant after multiple testing correction at p < 5.00E-4

Only the association with CHRNA3 survived correction for multiple testing at p < 5.00E-4 in both JAG and PLINK, although the associations with CHRNA5, CHRN4B4, and CHRNE survived correction for multiple testing using the PLINK test. Three genes were associated with CPD at p < 0.05 in one method and trending toward association at p < 0.1 in the other method: IQGAP1, APC, and UNC50. The correlations between the p-values in JAG and PLINK for COPDGene, SAGE, and CADD were moderate: 0.66, 0.76, and 0.45, respectively.

The p-values for the SNPs that passed the p-value and LD thresholds in PLINK for the genes associated with CPD using JAG and PLINK at p < 0.05 from each study were compared to the p-values for those SNPs in the TAG consortium. Only one SNP in NCAM1, rs17115184, was associated with CPD at p = 0.016 in the TAG study. Although this SNP was not associated
with CPD in CADD-GADD or SAGE, it was associated with CPD in COPDGene at p = 0.032. As expected, several signals in *CHRNA3-A5-B4* were associated with CPD in both TAG and the COPDGene and CADD samples. Two SNPs in *DAG1* were associated with CPD at p < 0.05 in the COPDGene sample; while one of these SNPs was in the TAG data yet not associated with CPD, the second SNP was not in TAG. Two SNPs in *CHRNE* were associated with CPD: one from COPDGene and one from SAGE. The SNP associated with CPD in the SAGE study, rs74550041, was in high LD with a SNP in the TAG study but the association was not replicated. The SNP associated with CPD from COPDGene was not in the TAG study, nor was it in the 1000G data so no LD information was available. One SNP in *DNAJA3* was associated with CPD in COPDGene; this SNP was in moderate LD (r² = 0.52-0.84) with four SNPs in TAG, although none were associated with CPD at p < 0.05. Finally, three SNPs in COPDGene and one in CADD-GADD were associated with CPD from *MAPRE1*. Two of the SNPs associated with CPD in the COPDGene study were also associated with CPD from TAG at p < 0.05: rs4911262 and rs612005. The SNP associated in the CADD-GADD study, rs6058892, was in high LD (r² from 0.84 – 1) with seven SNPs in the TAG, all trending toward association at p < 0.1. This SNP was also nominally associated with CPD in the COPDGene study (p = 0.043).

Fifteen thousand eight hundred and ninety three SNPs set were tested for association with FTND in EAs from the COPDGene and SAGE samples. The p-values were combined using the weighted Z-score method. Four genes were associated at p < 0.05 using both analytical methods: *CHRNA3, CHRNA9, DNAJA3*, and *RPSN* (Table 3).
Table 3. Results from the JAG and PLINK tests with FTND in EAs. The p-value for each test is shown for each study as well as the combined p-value for genes with p < 0.05 in either analysis.

<table>
<thead>
<tr>
<th>Gene</th>
<th>COPDGene</th>
<th>SAGE</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P_j^a</td>
<td>P_p^b</td>
<td>P_j</td>
</tr>
<tr>
<td>CHRNA3</td>
<td>0.001</td>
<td>0.003</td>
<td>0.031</td>
</tr>
<tr>
<td>CHRNA9</td>
<td>0.032</td>
<td>0.021</td>
<td>0.196</td>
</tr>
<tr>
<td>DNAJA3</td>
<td>0.003</td>
<td>0.004</td>
<td>0.737</td>
</tr>
<tr>
<td>RAPSN</td>
<td>0.167</td>
<td>0.167</td>
<td>0.028</td>
</tr>
<tr>
<td>FYN</td>
<td>0.329</td>
<td>0.010</td>
<td>0.634</td>
</tr>
<tr>
<td>CHRNA5</td>
<td>0.238</td>
<td>0.003</td>
<td>0.435</td>
</tr>
<tr>
<td>DMN1L</td>
<td>0.715</td>
<td>NA</td>
<td>0.002</td>
</tr>
<tr>
<td>CHRNB4</td>
<td>0.325</td>
<td>0.002</td>
<td>0.477</td>
</tr>
<tr>
<td>CHRNE</td>
<td>0.506</td>
<td>NA</td>
<td>0.013</td>
</tr>
<tr>
<td>CAMK2A</td>
<td>0.191</td>
<td>0.035</td>
<td>0.134</td>
</tr>
<tr>
<td>CAMK2G</td>
<td>0.004</td>
<td>0.007</td>
<td>0.863</td>
</tr>
<tr>
<td>SEC24D</td>
<td>0.429</td>
<td>0.011</td>
<td>0.844</td>
</tr>
<tr>
<td>LRP4</td>
<td>0.142</td>
<td>0.034</td>
<td>0.272</td>
</tr>
<tr>
<td>CRELD2</td>
<td>0.372</td>
<td>NA</td>
<td>0.123</td>
</tr>
<tr>
<td>CHRNB3</td>
<td>0.460</td>
<td>NA</td>
<td>0.024</td>
</tr>
<tr>
<td>STAT3</td>
<td>0.623</td>
<td>NA</td>
<td>0.253</td>
</tr>
<tr>
<td>SRC</td>
<td>0.333</td>
<td>NA</td>
<td>0.128</td>
</tr>
<tr>
<td>CAMK2B</td>
<td>0.046</td>
<td>0.045</td>
<td>0.550</td>
</tr>
<tr>
<td>NRXN1</td>
<td>0.461</td>
<td>0.111</td>
<td>0.826</td>
</tr>
</tbody>
</table>

^a P_emp for the JAG test
^b P_emp for the PLINK test
^c Significant after multiple testing correction at p < 5.00E-4

Only the association with CHRNA3 using JAG survived correction for multiple testing at p < 5.00E-4. Although not significant using JAG, FYN was associated with FTND using PLINK at p = <1.00E-4. This association also survived correction for multiple testing. Two additional genes were nominally associated with FTND: CAMK2G at p = 0.053 and p = 0.007 in JAG and PLINK, respectively; and CAMK2A at p = 0.091 and p = 0.006 in JAG and PLINK, respectively. The correlations between the p-values for JAG and PLINK in the COPDGene and SAGE samples were 0.68 and 0.40, respectively.

Cigarettes per day and nicotine dependence in AAs

Fourteen thousand six hundred and two SNPs from our set of genes were tested for association with CPD in AAs from the SAGE and COPDGene samples and the p-values
combined using the weighted Z-score method. PICK1, CHRNA9, and CRELD2 were associated with CPD using both JAG and PLINK at p < 0.05 (Table 4).

**Table 4.** Results from the JAG and PLINK tests with CPD in AAs. The p-value for each test is shown for each study as well as the combined p-value for genes with p < 0.05 in either analysis.

<table>
<thead>
<tr>
<th>Gene</th>
<th>COPDGene</th>
<th>SAGE</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P_j a</td>
<td>P_p b</td>
<td>P_j</td>
</tr>
<tr>
<td>PICK1</td>
<td>0.020</td>
<td>0.066</td>
<td>0.044</td>
</tr>
<tr>
<td>CHRNA9</td>
<td>0.012</td>
<td>0.048</td>
<td>0.448</td>
</tr>
<tr>
<td>MAPK1</td>
<td>0.019</td>
<td>0.121</td>
<td>0.190</td>
</tr>
<tr>
<td>NTRK2</td>
<td>0.014</td>
<td>0.274</td>
<td>0.666</td>
</tr>
<tr>
<td>DAG1</td>
<td>0.055</td>
<td>0.132</td>
<td>0.006</td>
</tr>
<tr>
<td>UBQLN1</td>
<td>0.023</td>
<td>0.067</td>
<td>0.499</td>
</tr>
<tr>
<td>CRELD2</td>
<td>0.053</td>
<td>0.074</td>
<td>0.021</td>
</tr>
<tr>
<td>CAV3</td>
<td>0.688</td>
<td>NA</td>
<td>0.013</td>
</tr>
</tbody>
</table>

a P_emp for the JAG test  
b P_emp for the PLINK test

MAPK1, DAG1, and UBQLN1 were associated with CPD using JAG at p = 0.014, p = 0.022, and p = 0.025, respectively, and trending toward association using PLINK at p < 0.1. None of the associations remained significant after the Bonferroni correction for multiple testing at p < 5.00E-4. The correlations between the p-values in JAG and PLINK were 0.58 for both the COPDGene and SAGE studies.

The same set of SNPs was tested for association with FTND and the p-values combined in AAs from the SAGE and COPDGene samples. MAPRE1, IQGAP1, UNC50, PDIA3, DNAJA3, and RER1 were associated with FTND at p < 0.05 using JAG and PLINK (Table 5).
Table 5. Results from the JAG and PLINK tests with FTND in AAs. The p-value for each test is shown for each study as well as the combined p-value for genes with p < 0.05 in either analysis.

<table>
<thead>
<tr>
<th>Gene</th>
<th>COPDGene</th>
<th>SAGE</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(P_p)</td>
<td>(P_p)</td>
<td>(P_p)</td>
</tr>
<tr>
<td>MAPRE1</td>
<td>0.008</td>
<td>0.023</td>
<td>0.500</td>
</tr>
<tr>
<td>IQGAP1</td>
<td>0.023</td>
<td>0.041</td>
<td>0.130</td>
</tr>
<tr>
<td>UNC50</td>
<td>0.009</td>
<td>0.011</td>
<td>0.878</td>
</tr>
<tr>
<td>PDIA3</td>
<td>0.027</td>
<td>0.028</td>
<td>0.374</td>
</tr>
<tr>
<td>DNAJA3</td>
<td>0.022</td>
<td>0.011</td>
<td>0.818</td>
</tr>
<tr>
<td>RER1</td>
<td>0.012</td>
<td>0.018</td>
<td>0.961</td>
</tr>
<tr>
<td>CHRN4B1</td>
<td>0.462</td>
<td>NA</td>
<td>0.067</td>
</tr>
<tr>
<td>TMX3-54495</td>
<td>0.077</td>
<td>0.015</td>
<td>0.669</td>
</tr>
<tr>
<td>CACNA2D1</td>
<td>0.213</td>
<td>0.007</td>
<td>0.705</td>
</tr>
<tr>
<td>GAB2</td>
<td>0.780</td>
<td>NA</td>
<td>0.624</td>
</tr>
</tbody>
</table>

\(^a\) \(P_{emp}\) for the JAG test
\(^b\) \(P_{emp}\) for the PLINK test

Although these results were consistent, none survived correction for multiple testing at p < 5.00E-4. The correlations between the p-values in JAG and PLINK were 0.47 and 0.62 for the COPDGene and SAGE samples, respectively.

Discussion

While the CHRN gene cluster on chromosome 15q25 comprised of CHRNA3, CHRNA4, and CHRNA5 showed the strongest associations with nicotine behaviors in EAs in this study, replicating previous associations with these genes and CPD (2010; Berrettini et al., 2008; Cannon et al., 2014; Caporaso et al., 2009; Frederiksen et al., 2015; Gabrielsen et al., 2013; Liu et al., 2010; Saccone et al., 2010a; Sarginson et al., 2011; Sorice et al., 2011; Stevens et al., 2008; Thorgeirsson et al., 2010) and ND (Baker et al., 2009; Bierut et al., 2007; Bierut et al., 2008; Broms et al., 2012; Chen et al., 2009b; Erlich et al., 2010; Haller et al., 2012; Maes et al., 2011; Saccone et al., 2009a; Saccone et al., 2009b; Sherva et al., 2010; Spitz et al., 2008; Thorgeirsson et al., 2008; Weiss et al., 2008; Wessel et al., 2010), several genes not previously implicated in human populations were associated with CPD and FTND. NCAM1, DAG1, DNAJA3 (discussed in a later section), and MAPRE1 were associated with CPD and RAPSN
with FTND. Neural cell adhesion molecule 1, NCAM1 or NCAM, had been shown to cluster with nAChRs and known nAChR-associated proteins and may influence the stability of nAChR clusters (Conroy et al., 2000; Moscoso et al., 1995b). In rodents, nicotine exposure decreased expression of NCAM (Abrous et al., 2002; Shingo et al., 2005). Finally, human genetic studies have identified a linkage peak for the region containing NCAM1 for ND in EAs (Gelernter et al., 2007) and subsequent studies have supported evidence for association between NCAM1 and ND (Gelernter et al., 2006), alcohol dependence (Yang et al., 2007), heroin dependence (Nelson et al., 2013), and general drug dependence (Yang et al., 2008). Dystroglycan 1, or DAG1, forms a complex with several other proteins to provide a link between the neuromuscular junction and extracellular matrix of the cell. It has been shown to co-localize with nAChRs throughout maturation of the rat neuromuscular junction (Bewick et al., 1996), as well as enhance synaptic differentiation and stability (Cote et al., 1999; Grady et al., 2000), particularly though association with known proteins that cluster nAChRs at the synapse (Apel et al., 1995; Cartaud et al., 1998; Fuhrer et al., 1999; Jacobson et al., 2001; Tremblay et al., 2006; Yamada et al., 1996). Additionally, dystroglycans are transported together with nAChRs and rapsyn in to the postsynaptic membrane in post-golgi vesicles (Marchand et al., 2001). MAPRE1, or EB1, is a microtubule-associated protein whose interaction with known nAChR-associated proteins is important for clustering and surface insertion and stabilization of α3-containing nAChRs (Rosenberg et al., 2008; Temburni et al., 2004). RAPSN encodes the protein rapsyn, a protein consistently implicated in clustering (Apel et al., 1995; Gautam et al., 1999; Gillespie et al., 1996; Qu et al., 1996), phosphorylation (Apel et al., 1997; Mohamed et al., 1999), localization (Kassner et al., 1998), turnover (Wang et al., 1999), trafficking (Han et al., 2000; Marchand et al., 2000), and degradation (Phillips et al., 1997) of nAChRs at the neuromuscular junction. Furthermore, rapsyn interacts with dystroglycan at the postsynaptic membrane (Apel et al., 1995; Cartaud et al., 1998; Fuhrer et al., 1999).
For two of the genes discussed above, *NCAM1* and *MAPRE1*, single SNP associations observed in the present study with CPD were also observed in the TAG consortium. While these signals did not survive correction for multiple testing in either analysis, these signals warrant further investigation.

Unexpectedly, *CHRNE*, which codes for the ε subunit in the adult muscle nAChR, was associated with CPD in EAs using both analytical methods and implicated in FTND in the SAGE study. The nAChR subunits in muscle cells were included in the present study for several reasons: first, Ghedini et al. (2010) detected positive polymerase chain reaction (PCR) bands for the α1, β1, δ, and ε subunits in the cortex, hippocampus, and cerebellum of adult mice (Ghedini et al., 2010), suggesting these muscle nAChR subunits may be expressed in various brain regions of mammals; second, several proteins that interact with nAChRs were identified in *Caenorhabditis elegans*, in which there is little distinction between muscle and neuronal nAChRs; and third, recent research suggests that nicotine has a rapid, transient action on the afferents of sensory nerves that mediate the acute neural effects of nicotine and, based on these results, hypothesized that nicotine in the periphery produces a rapid sensory signal to the central nervous system that is followed by a slower, more prolonged direct drug action in the brain and that these two actions produce a distinct conditioning in the brain as a result of nicotine exposure (Kiyatkin, 2014). Although the typical serum concentration of nicotine in smokers has little effect on muscle nAChRs (Lindstrom, 1996), upregulation of the muscle type nAChR subunits α1, β1, γ, and δ, has been observed in vitro after a 48 hour exposure to nicotine at concentrations as low as 1µM (Ke et al., 1998). However, evidence for muscle type nAChR upregulation remains mixed (Appel et al., 1981; Gardner et al., 1979; Luther et al., 1989; Noble et al., 1978; Siegel et al., 1988). This is the first evidence for association of *CHRNE* with nicotine behaviors in human populations to our knowledge.

Unlike in EAs, the *CHRN* genes, particularly *CHRNA3-A5-B4*, were not the strongest association with CPD and FTND in AAs. None of the cluster genes were associated with CPD in
either the SAGE or COPDGene samples. This finding is in agreement with findings on CPD in AAs in the COPDGene sample (Frederiksen et al., 2015). However, one previous study by David et al. (2012) with a large sample size of over thirty-two thousand found an association between CHRNA3-A5-B4 with CPD. However, only one SNP, residing in the 5′-distal enhancer region of CHRNA5, exceeded genome-wide significance (David et al., 2012). It is possible that since our analyses sum individual SNP statistics to get an overall gene p-value, this SNP was overshadowed by SNPs of lower significance. In our study, CHRNA3, CHRNB4, and CHRNA5 were nominally associated with FTND in the SAGE sample using PLINK, although there was no evidence for association between the three genes and FTND in the COPDGene sample. While this suggests heterogeneity between the two samples, it aligns with what has been previously seen in the literature. Saccone et al. (2009) reported an association between SNPs in CHRNA5 and ND in the Collaborative Genetic Study of Nicotine Dependence (COGEND), a part of the SAGE study (Saccone et al., 2009b), yet no association between CHRNA3, CHRNB4, or CHRNA5 with ND was seen in the COPDGene sample (Frederiksen et al., 2015).

CHRNA9 was associated with FTND in EAs and CPD in AAs, possibly connecting vertigo and smoking behaviors. Vertigo has long sense been known to be a disorder of the inner ear (Ciuman, 2013). In the central nervous system, the α9 and α10 subunits, encoded by the CHRNA9 and CHRNA10 genes, respectively, have been primarily identified in the inner ear cells, particularly the cochlear and vestibular hair cells (Elgoyhen et al., 1994; Elgoyhen et al., 2001; Luo et al., 1998; Lustig et al., 1999; Lustig et al., 2001; Vetter et al., 1999). The α10 subunit, although it cannot form a functional homomeric receptor, is most closely related to the α9 subunit and forms a functional receptor when coexpressed with the α9 subunit (Elgoyhen et al., 1994; Lustig et al., 2001; Sgard et al., 2002). Molecules known to contribute to nAChR formation at the neuromuscular junction have been identified in inner ear cells, and have also been shown to interact with α9 subunits; one such molecule is rapsyn, which was associated with FTND in EAs in the present study (Osman et al., 2008). Although the primary role identified
for nAChR function within the inner ear is related to efferent auditory neuronal activity (Lustig, 2006), the evidence for protein-protein interactions with the α9α10 receptor suggests this receptor may play a role in other physiological processes, such as vertigo, outside of the auditory system. Furthermore, CHRNA10 has been associated with dizziness in a human sample (Ehringer et al., 2010). Although CHRNA9 has been previously associated with ND in female Israeli students (Greenbaum et al., 2006), as well as with lung cancer in Caucasians (Chikova et al., 2012), this is the first association to our knowledge with CHRNA9 in AAs.

Two additional genes were nominally associated with CPD in AAs using JAG and PLINK: PICK1, and CRELD2. Protein interacting with protein kinase C α, PICK1, has been shown to bind α7 nAChRs and play a role in clustering (Baer et al., 2007), as well as in anchoring proteins to the neuromuscular junction through its interaction with erb-b3 receptor tyrosine kinase 2 (ERBB2) (Huang et al., 2002; Jaulin-Bastard et al., 2001). ERBB2, a receptor for neuregulin, is enriched at synaptic sites in myotubues (Jo et al., 1995; Trinidad et al., 2000; Zhu et al., 1995) and increasing evidence has suggested a role for ERBB2 in the synaptic expression of nAChRs (Jo et al., 1995; Leu et al., 2003; Meier et al., 1998; Moscoso et al., 1995a). CRELD2, cysteine-rich with EGF-like domains 2, is an endoplasmic reticulum (ER) protein whose expression is induced by nicotine (Hosur et al., 2009). CRELD2 has been shown to co-localize with α4β2 nAChRs in the ER and evidence suggests it can regulate the membrane transport (Ortiz et al., 2005) and expression of α4β2 nAChRs (Hosur et al., 2009; Ortiz et al., 2005).

Several different genes than those associated with CPD were associated with FTND in AAs: MAPRE1, IQGAP1, UNC50, PDIA3, DNAJA3, and RER1. As mentioned above, MAPRE1 influences nAChR clustering, surface expression, and stabilization via interactions with known nAChR-associated proteins (Rosenberg et al., 2008; Temburni et al., 2004). IQGAP1, a scaffolding protein widely expressed in different cell types, is thought to play a role in neuronal development, synaptic plasticity, and nervous system disorders (Jausoro et al., 2012).
Additionally, IQGAP1 was found to be present in a multi-molecular complex that stabilized and linked the necessary postsynaptic components to the cytoskeleton for expression of α3-containing nAChRs (Rosenberg et al., 2008). UNC50, or UNCL, is an inner nuclear membrane protein involved in the cell-surface expression of nAChRs; it has been shown to increase the expression of surface nAChRs in mammalian cells up to 1.6 fold (Fitzgerald et al., 2000). PDIA3, or ERp57, is an ER chaperone that associates with newly synthesized nAChR subunits. These associations are long-lived and prolong subunit lifetime by approximately 10 fold (Wanamaker et al., 2007). RER1, a putative gogli-ER retrieval receptor, controls surface expression of nAChRs at the neuromuscular junction by localizing unassembled α-containing receptors to the early secretory pathway (Valkova et al., 2011).

DNAJA3 was associated with both nicotine behaviors in EAs and with FTND in AAs. These associations were driven by the COPDGene study in all cases. Rs2270365 was associated with CPD in EAs (although SNPs in LD with this SNP were not associated with CPD in TAG), rs11076831 with FTND in EAs, and rs7201012 with FTND in AAs. In the CEU population of 1000G these SNPs were moderately correlated, suggesting one signal in the gene is driving the association with CPD and FTND; rs2270365 and rs11076831 were correlated at $r^2 = 0.52$, rs2270365 and rs7201012 at $r^2 = 0.61$, and rs11076831 and rs7201012 at $r^2 = 0.78$. DNAJA3 (or TID1), is a protein that stimulates heat shock protein chaperones, has been shown to bind a major component of the agrin receptor and colocalize with nAChRs at the neuromuscular junction. Knockdown of DNAJA3 impaired neuromuscular transmission and caused dispersal of nAChR clusters. Overexpression of DNAJA3 induced phosphorylation and clustering of nAChRs (Linnoila et al., 2008).

It is interesting that several genes associated with CPD in EAs were implicated in FTND in AAs (e.g. MAPRE1, IQGAP1, UNC50, and DNAJA3) and vice versa. These findings might indicate underlying genetic differences in the development of smoking behaviors in EAs and AAs. AAs tend to start smoking later in life (Chatila et al., 2004; Duncan et al., 2012; Finkenauer
et al., 2009; White et al., 2004) and have a lower lifetime prevalence of ND compared to EAs (Edens et al., 2010). Additionally, AAs report higher cravings and more pleasurable sensations after smoking (Carter et al., 2010; Finkenauer et al., 2009), lower rates of regular smoking (Chatila et al., 2004; Edens et al., 2010; Ellickson et al., 2004; Finkenauer et al., 2009; Hahn et al., 1990; Kabat et al., 1991; Moolchan et al., 2006; Signorello et al., 2009; White et al., 2004), higher nicotine intake per cigarette (Perez-Stable et al., 1998), and slower metabolism rates of nicotine (Benowitz et al., 2011; Perez-Stable et al., 1998; Rubinstein et al., 2013; Signorello et al., 2009; St Helen et al., 2013) compared to EAs. Finally, EAs have higher smoking cessation rates than AAs (Hahn et al., 1990). Together with our data, this suggests differences in smoking patterns between EAs and AAs may be partially genetically driven.

Other than the CHRNA3-A5-B4 findings in EAs, our results show little overlap in the genes associated with CPD and FTND within each ethnic group. As research has demonstrated both unique and common genetic influences on CPD and ND (Haberstick et al., 2007), we expected to see some overlap between the two phenotypes. This discrepancy may be due to the fact that FTND was only measured in current smokers in the COPDGene sample and our results could reflect differences in smoking cessation as well.

There are some differences in the analyses done in JAG verses the set-based test in PLINK. Primarily, JAG performs better when there are many signals in the gene that contribute risk to the phenotype because the p-values for all the SNPs are simply summed to create the gene p-value. Conversely, the set-based test in PLINK selects only the most significant signals to make the set statistic. Thus, a gene with a few very strong signals would appear significant using the PLINK set-test but not the JAG gene test, yet a gene with many weak signals would appear significant using the JAG gene test but not the PLINK set-test. Nevertheless, the utilization of two different tests serves as confirmation that a particular gene is worth future investigation and enables us to refine the list of future targets for nicotine studies.

While not without limitations, several strengths of the present study help mitigate these
limitations. Although JAG was not included, one study found the PLINK set test to be the most powerful of 7 studied algorithms for pathway analyses (Gui et al., 2011). SNP-set based approaches can increase power to detect genetic loci with individually small effects by consolidating SNP associations and, in addition to the use of two analytical methods, can help to prioritize associations based on biological relevance (Purcell et al., 2009; Wang et al., 2011). A meta-analysis approach was used as opposed to a mega-analysis so associations with genes across studies and analytical methodologies could be compared. By curating a unique list of genes to test, we avoided many potential limitations when examining already curated gene sets like KEGG and GO; for example, when examining gene sets from KEGG or GO, genes that have been well studied and more thoroughly annotated. As of 2011, only about 5,000 human genes had been annotated to KEGG pathways (Wang et al., 2011). However, despite being able to curate a unique set of genes, this study was limited to testing SNPs annotated on the hg19 build and genes annotated on all GWAS utilized. In addition, these studies would benefit from a priori knowledge of each gene, in that we could target certain SNPs from each gene to test and inform us to which analysis would be most appropriate to use for a particular gene.

In conclusion, we have identified in human samples several genes previously thought to be involved in nAChR upregulation that were associated with FTND and CPD. The list of genes tested is by no means an exhaustive list, and simply served to test the hypothesis that SNPs in the identified genes are associated with nicotine behaviors, suggesting these SNPs may alter nAChR function and/or expression through protein-protein interactions. It is likely that genes on our list not associated with nicotine phenotypes in the current study may nevertheless be affecting nAChR upregulation in a manner not directly driven by SNPs; for example, increased transcription factor binding, or enhancer effects. Further addition of independent samples would aid in strengthening these associations, or replication in a group of unrelated GWAS. This work highlights the value of collaboration between neuroscience experts and statistical geneticists to identify and characterize novel genetic associations.
CHAPTER V
EXAMINATION OF THE INVOLVEMENT OF CHOLINERGIC-ASSOCIATED GENES IN ALCOHOL DEPENDENCE IN EUROPEAN AMERICANS

This work is in preparation

Introduction

Alcoholism is a damaging and detrimental disease. Worldwide, 3.3 million people die each year due to the harmful use of alcohol. This number represents 5.9% of deaths worldwide. Furthermore, approximately 25% of total deaths in people aged 20-30 are alcohol-attributable. Alcohol causes 5.1% of the global burden of disease and injury, as measured in disability-adjusted life years. Finally, recent causal relationships have been established between harmful drinking and the incidence of infectious diseases such as tuberculosis and HIV/AIDs (WHO).

There is strong evidence that alcohol dependence (AD) is influenced by genetic factors. Estimates from twin studies estimate that AD is 49-64% heritable in adult men and women (Heath et al., 1997; Prescott et al., 1999; True et al., 1999; Vrieze et al., 2013). Furthermore, 21% of the heritability in AD can be explained by common single nucleotide polymorphisms (SNPs) (Vrieze et al., 2013).

There is heavy overlap between alcohol and tobacco behaviors (DiFranza et al., 1990). Additionally, twin studies have shown a substantial genetic correlation between lifetime alcohol and nicotine dependence (True et al., 1999). Evidence from electrophysiological, pharmacological, and neurochemical studies has suggested some of this common genetic liability may lie in the nicotinic acetylcholine receptors (nAChRs, encoded by the CHRN genes) (Larsson et al., 2004). In line with these findings, numerous human genetic association studies have found evidence for involvement of nAChRs with alcohol behaviors (Table 1).
Table 1. Polymorphisms in the *CHRN* genes associated with alcohol behaviors.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Gene(s)</th>
<th>Sample(s)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD&lt;sup&gt;a&lt;/sup&gt;</td>
<td><em>CHRNA4</em></td>
<td>Unrelated Korean males</td>
<td>(Kim <em>et al</em>., 2004)</td>
</tr>
<tr>
<td>Subjective responses to alcohol and tobacco, past 6 month use alcohol</td>
<td><em>CHRNA4</em>, <em>CHRNB2</em></td>
<td>CADD&lt;sup&gt;b&lt;/sup&gt;</td>
<td>(Ehringer <em>et al</em>., 2007)</td>
</tr>
<tr>
<td>Age of initiation of alcohol and tobacco</td>
<td><em>CHRNA3</em>-B4</td>
<td>CADD, NYS&lt;sup&gt;c&lt;/sup&gt;</td>
<td>(Schlaepfer <em>et al</em>., 2008)</td>
</tr>
<tr>
<td>AD</td>
<td><em>CHRNA5</em>-A3</td>
<td>COGA&lt;sup&gt;d&lt;/sup&gt;</td>
<td>(Wang <em>et al</em>., 2009)</td>
</tr>
<tr>
<td>Level of response to alcohol</td>
<td><em>CHRNA5</em>-A3-B4</td>
<td>San Diego Sibling Pair</td>
<td>(Joslyn <em>et al</em>., 2008)</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td><em>CHRN3</em>-A6</td>
<td>NYS Family Study</td>
<td>(Hoft <em>et al</em>., 2009a)</td>
</tr>
<tr>
<td>ND&lt;sup&gt;e&lt;/sup&gt;, AD</td>
<td><em>CHRNA5</em>-A3</td>
<td>Virginia Adult Twin Study</td>
<td>(Chen <em>et al</em>., 2009b)</td>
</tr>
<tr>
<td>Heavy alcohol use</td>
<td><em>CHRNA6</em></td>
<td>Spanish population</td>
<td>(Landgren <em>et al</em>., 2009)</td>
</tr>
<tr>
<td>ND, cocaine dependence, AD</td>
<td><em>CHRNA5</em>-A3-B4</td>
<td>Five American cohorts</td>
<td>(Sherva <em>et al</em>., 2010)</td>
</tr>
<tr>
<td>ND, regular drinking</td>
<td><em>CHRNA5</em>-A3-B4</td>
<td>Finnish Twin Cohort</td>
<td>(Broms <em>et al</em>., 2012)</td>
</tr>
<tr>
<td>General substance use initiation</td>
<td><em>CHRNA5</em>-A3-B4</td>
<td>Add Health</td>
<td>(Lubke <em>et al</em>., 2012)</td>
</tr>
<tr>
<td>Alcohol use</td>
<td><em>CHRNA5</em>-A3-B4</td>
<td>National FINRISK Study, Health 2000 Survey</td>
<td>(Hoflfors <em>et al</em>., 2013)</td>
</tr>
<tr>
<td>Frequency of binge drinking</td>
<td><em>CHRNA4</em></td>
<td>SECASP</td>
<td>(Coon <em>et al</em>., 2014)</td>
</tr>
</tbody>
</table>

<sup>a</sup> AD: Alcohol Dependence

<sup>b</sup> CADD: Center on Antisocial Drug Dependence

<sup>c</sup> NYS: National Youth Survey

<sup>d</sup> COGA: Collaborative Study on the Genetics of Alcoholism

<sup>e</sup> ND: Nicotine Dependence

An interesting characteristic of nAChRs is their upregulation in response to chronic nicotine exposure. The major addictive component in tobacco smoke is nicotine (Gunby, 1988; Stolerman *et al*., 1995). Nicotine exerts its effect by binding nAChRs located in the peripheral and central nervous systems (Gotti *et al*., 2009). Upregulation of nAChRs after exposure to tobacco (Benwell *et al*., 1988; Marks *et al*., 1983; Schwartz *et al*., 1983) has been consistently shown in several models. Furthermore, this phenomenon is independent of *CHRN* mRNA (Marks *et al*., 1992). There are many theories as to how this upregulation occurs: increased receptor trafficking (Darsow *et al*., 2005), decreased subunit degradation (Rezvani *et al*., 2009; Rezvani *et al*., 2007), increased nAChR subunit maturation and folding (Harkness *et al*., 2002;
Nashmi et al., 2003; Sallette et al., 2005), and increased translation and 2nd messenger signaling (Gopalakrishnan et al., 1997). However, researchers have yet to test several of the proteins suggested to impact upregulation in conjunction. Through manual curation of the literature and in conversation with experts a list of genes has been developed whose encoded proteins interact with nAChRs (see Figure 1 in Chapter IV on page 69 for a schematic of how these proteins cooperate with nAChRs).

Despite the number of associations between the CHRN genes and alcohol behaviors, the evidence that alcohol induces upregulation of nAChRs remains mixed. Early studies in rats showed that ethanol treatment decreased nAChR binding sites in the hippocampus but increased binding sites in the hypothalamus and thalamus (Yoshida et al., 1982). This aligned with findings that mice treated with ethanol showed changes in nicotine binding dependent on strain and region (Booker et al., 1997). Unlike nicotine, chronic ethanol exposure was shown to alter mRNA levels in different cell lines, as well as hinder nicotine-induced upregulation of nAChRs (Gorbounova et al., 1998). However, a subsequent study found an initial decrease in nAChR expression after exposure to ethanol followed by an increase in nAChR expression in cell lines. This pattern was replicated when cells were exposed to both nicotine and ethanol and resulted in higher and more long-lasting nAChR upregulation than cells being exposed to nicotine alone (Dohrman et al., 2003). Contrary to this finding, another group found that concurrent nicotine and ethanol exposure elicited nAChR upregulation, yet no effects of ethanol alone on nAChR binding or upregulation were seen in mice, suggesting upregulation of nAChRs after nicotine and ethanol exposure could have been due to nicotine alone (Ribeiro-Carvalho et al., 2008; Ribeiro-Carvalho et al., 2009). Finally, the number of nAChR binding sites did not differ in the frontal cortex or thalamus of alcoholics compared to age-matched controls (Hellstrom-Lindahl et al., 1993). Although the studies presented above do not have consistent findings, the associations between nAChRs and alcohol behaviors suggest a role for the nAChR neurotransmitter system in alcoholism.
Based on work presented above suggesting that a large proportion of the heritability of AD can be explained by common SNPs (Vrieze et al., 2013), we hypothesize that common SNPs in cholinergic-associated genes explain a large fraction of the variation in AD. The goal of the present study was to use the list of cholinergic-associated genes developed through manual curation of the literature and assess the evidence for association with each gene and AD in European Americans (EAs) using set-based approaches. SNP-set based approaches are a significant technique that can increase the power to detect genetic loci with small effects by combining individual SNP associations (Purcell et al., 2009; Wang et al., 2011).

Materials and methods

Samples

Data for EAs from four Genome-Wide Association Studies (GWAS) were obtained: The Study of Addition: Genetics and Environment (SAGE) (Bierut et al., 2010a), Genome-wide association study of Alcohol use and alcohol use disorder in Australian Twin-families (OZ-ALC) (Grant et al., 2007; Knopik et al., 2004; Saccone et al., 2007b), Alcohol dependence GWAS in European- and African-Americans (AD-GWAS) (Gelernter et al., 2005), and a GWAS comprised of unrelated EA subjects from the CADD and GADD samples (CADD-GADD) (Derringer et al., 2015). The CADD and GADD samples include probands targeted for drug use, their families and matched control families, as well as community twin samples. SAGE and AD-GWAS were ascertained based on substance dependence. The OZ-ALC study sampled target families with a family history of AD and matched control families without a family history of AD. One subject was randomly selected from each family from the OZ-ALC study to include in the analysis. All GWAS datasets were imputed to Phase 1 of the 1,000 Genomes data (version 3) (Abecasis et al., 2012) and were cleaned using standard quality control procedures. These included
removing variants with low imputation accuracy (< 0.9), low minor allele frequency (MAF < 0.01), and variants out of Hardy-Weinberg equilibrium (HWE, p < 0.001)

**SNP mapping**

SNPs were annotated to genes in the set using the hg19 build. Variants 20kb up- and downstream of the gene were included, as previous evidence has suggested most genetic variation lies in this region (Veyrieras et al., 2008). Only variants present in all four GWAS were included in the final analysis.

**Statistical analysis**

**Phenotypic descriptive**

Quantitative AD symptoms, defined as a maladaptive pattern of drinking that leads to significant clinical impairment or distress, was assessed using either the Composite International Diagnostic Interview – Substance Abuse Module (CIDI-SAM) (Cottler et al., 1990) or versions of the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA) (Bucholz et al., 1994). As the CADD-GADD GWAS is longitudinal, phenotypes were preferentially drawn from wave 3, from wave 2 if the subject was not assessed at wave 3, and from wave 1 if the subject did not have data at either later wave. All subjects from each sample were included since information about whether each subject had tried alcohol was not always available. Age and sex, as well as the first 5 principal components to control for ancestry, were used as covariates in the analysis. Although the SAGE and AD-GWAS samples were part of 3 sub-studies ascertained on various substances, study was only available in the SAGE sample and thus was additionally used as a covariate.

**Statistical genetics analysis**
The gene-based test was carried out using two analytical approaches: Joint Association of Genetic Variants (JAG; available at http://ctglab.nl/software/jag/) (Lips) and the PLINK set-test (Purcell et al., 2007). JAG uses raw data and invokes PLINK to run a genome-wide association on each SNP in the gene. The \(-\log_{10}\) of each SNP p-value is summed to create a multivariate test statistic. Permutations of the phenotype are run and summed as described above. The empirical p-value for each gene (\(P_{\text{emp}}\)) is the number of times the sum of the \(-\log_{10}\) of the permuted p-value exceeds or equals the sum of the \(-\log_{10}(p\text{-value})\) from the SNP associations. The set-test implemented in PLINK was also used in order to compare results between the two analytical approaches. Briefly, for each gene PLINK determines which SNPs are in LD above a certain threshold (\(r^2 = 0.5\)), performs a standard SNP regression, and selects up to \(N\) “independent” significantly associated SNPs at \(p < 0.05\). The statistic for each gene is calculated as the mean of these independent SNP statistics. Keeping LD between SNPs constant, the phenotype is permuted. \(P_{\text{emp}}\) is the number of times the permuted statistic exceeds the original statistic for that gene. The \(P_{\text{emp}}\) for each sample from the JAG and PLINK analyses were combined separately using the weighted Z-score method to obtain an overall p-value derived from each analysis (Whitlock, 2005). As the PLINK analyses only perform permutation for SNPs with \(p < 0.05\), if a gene does not have any SNPs with \(p < 0.05\) an automatic p-value of 1 is returned. These p-values were changed to NA and excluded from the meta-analyses as no permutation testing was performed.

**Post hoc SNP comparisons**

The PLINK set-test output also supplies the SNPs that pass the p-value and LD thresholds (if any) for each gene. Using SNP Annotation and Proxy Search (SNAP) (Johnson et al., 2008), pairwise \(r^2\) was evaluated between SNPs implicated with AD in each dataset. Pairwise \(r^2\) was determined using a sample of Utah residents with ancestry from northern and western Europe from the 1000 Genomes Pilot 1 (1000G) data (Abecasis et al., 2012).
Results

Phenotypic characteristics for the each GWAS are given in Table 2.

Table 2. Phenotypic characteristics for each sample.

<table>
<thead>
<tr>
<th>Study</th>
<th>N</th>
<th>Age±</th>
<th>% Male</th>
<th>AD±</th>
</tr>
</thead>
<tbody>
<tr>
<td>CADD-GADD</td>
<td>1031</td>
<td>23.59±4.26</td>
<td>70.71%</td>
<td>2.03±2.12</td>
</tr>
<tr>
<td>AD-GWAS</td>
<td>516</td>
<td>38.85±10.83</td>
<td>62.40%</td>
<td>5.36±1.39</td>
</tr>
<tr>
<td>OZ-ALC</td>
<td>2016</td>
<td>44.20±9.63</td>
<td>50.10%</td>
<td>2.07±1.80</td>
</tr>
<tr>
<td>SAGE</td>
<td>2258</td>
<td>38.18±9.58</td>
<td>44.69%</td>
<td>2.79±2.51</td>
</tr>
</tbody>
</table>

*Mean ± standard deviation

Although the CADD-GADD sample was slightly younger than the 3 other GWAS utilized, mean AD scores were comparable to the SAGE and OZ-ALC samples. The AD-GWAS sample had the highest AD scores.

Twenty three thousand five hundred and thirty seven SNPs in our gene set were tested for association with AD using two gene-based tests. There were 4 genes associated with AD at p < 0.05 using both analytical methods: LAMA5, CHRNA1, RAC1, and DNAJA3 (Table 3).

Table 3. Results from the JAG and PLINK analyses in each study. The combined p-values from the meta-analyses for genes with p < 0.05 in either analytical test are shown.

<table>
<thead>
<tr>
<th>Gene</th>
<th>SAGE</th>
<th>CADD-GADD</th>
<th>AD-GWAS</th>
<th>OZ-ALC</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pj,a</td>
<td>Pp,a</td>
<td>Pj,b</td>
<td>Pp,b</td>
<td>Pj,c</td>
</tr>
<tr>
<td>LAMA5</td>
<td>0.107</td>
<td>0.196</td>
<td>0.119</td>
<td>0.154</td>
<td>0.543</td>
</tr>
<tr>
<td>CHRNA1</td>
<td>0.025</td>
<td>0.030</td>
<td>0.632</td>
<td>NA</td>
<td>0.565</td>
</tr>
<tr>
<td>RAC1</td>
<td>0.065</td>
<td>0.175</td>
<td>0.027</td>
<td>0.153</td>
<td>0.914</td>
</tr>
<tr>
<td>NRG1</td>
<td>0.014</td>
<td>0.338</td>
<td>0.299</td>
<td>0.264</td>
<td>0.270</td>
</tr>
<tr>
<td>UTRN</td>
<td>0.593</td>
<td>0.633</td>
<td>0.192</td>
<td>0.668</td>
<td>0.243</td>
</tr>
<tr>
<td>NCAM1</td>
<td>0.114</td>
<td>0.297</td>
<td>0.103</td>
<td>0.449</td>
<td>0.500</td>
</tr>
<tr>
<td>LAMA1</td>
<td>0.221</td>
<td>0.454</td>
<td>0.189</td>
<td>0.204</td>
<td>0.071</td>
</tr>
<tr>
<td>DNAJA3</td>
<td>0.402</td>
<td>NA</td>
<td>0.044</td>
<td>0.040</td>
<td>0.015</td>
</tr>
<tr>
<td>CACNA2D1</td>
<td>0.122</td>
<td>0.320</td>
<td>0.062</td>
<td>0.434</td>
<td>0.777</td>
</tr>
<tr>
<td>PPP1R1B</td>
<td>0.804</td>
<td>NA</td>
<td>0.206</td>
<td>0.122</td>
<td>0.724</td>
</tr>
<tr>
<td>NSF</td>
<td>0.165</td>
<td>NA</td>
<td>0.276</td>
<td>NA</td>
<td>0.008</td>
</tr>
<tr>
<td>CAV3</td>
<td>0.020</td>
<td>0.007</td>
<td>0.777</td>
<td>NA</td>
<td>0.688</td>
</tr>
<tr>
<td>CHRNBP2</td>
<td>0.353</td>
<td>NA</td>
<td>0.951</td>
<td>NA</td>
<td>0.093</td>
</tr>
<tr>
<td>CAMK2A</td>
<td>0.086</td>
<td>0.173</td>
<td>0.813</td>
<td>NA</td>
<td>0.201</td>
</tr>
<tr>
<td>CAMK2D</td>
<td>0.581</td>
<td>0.068</td>
<td>0.830</td>
<td>NA</td>
<td>0.082</td>
</tr>
<tr>
<td>DLG2</td>
<td>0.533</td>
<td>0.106</td>
<td>0.496</td>
<td>0.245</td>
<td>0.21</td>
</tr>
<tr>
<td>DAG1</td>
<td>0.362</td>
<td>NA</td>
<td>0.531</td>
<td>NA</td>
<td>0.044</td>
</tr>
<tr>
<td>DOK7</td>
<td>0.788</td>
<td>NA</td>
<td>0.506</td>
<td>NA</td>
<td>0.957</td>
</tr>
<tr>
<td>UBXN2A</td>
<td>0.711</td>
<td>NA</td>
<td>0.319</td>
<td>0.290</td>
<td>0.430</td>
</tr>
<tr>
<td>Gene</td>
<td>YWHAQ</td>
<td>MAPRE1</td>
<td>SEC24D</td>
<td>TRPC7</td>
<td>UBQLN1</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>0.920</td>
<td>0.355</td>
<td>0.848</td>
<td>0.559</td>
<td>0.480</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>0.217</td>
<td>NA</td>
<td>0.167</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>0.367</td>
<td>0.872</td>
<td>0.921</td>
<td>0.530</td>
<td>0.491</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>0.035</td>
<td>NA</td>
<td>0.145</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>0.818</td>
<td>0.780</td>
<td>0.724</td>
<td>0.030</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.098</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td>0.025</td>
<td>NA</td>
<td>0.116</td>
<td>0.235</td>
<td>0.138</td>
</tr>
<tr>
<td></td>
<td>0.035</td>
<td>0.131</td>
<td>0.038</td>
<td>0.219</td>
<td>0.172</td>
</tr>
<tr>
<td></td>
<td>0.419</td>
<td>0.727</td>
<td>0.697</td>
<td>0.267</td>
<td>0.172</td>
</tr>
<tr>
<td></td>
<td>0.035</td>
<td>0.727</td>
<td>0.038</td>
<td>0.045</td>
<td>0.045</td>
</tr>
</tbody>
</table>

\(^a\) \(p_{\text{emp}}\) for the JAG test
\(^b\) \(p_{\text{emp}}\) for the PLINK test

Although these associations did not survive Bonferroni correction for multiple testing at \(p < 5.00 \times 10^{-4}\), all of these associations, with the exception of \(\text{CHRNA1}\), were driven by more than 1 study. One additional gene, \(\text{NSF}\), was associated with AD at \(p = 0.007\) in PLINK and trending towards association in JAG at \(p = 0.083\) (Table 3). The correlations between the empirical p-values from JAG and PLINK for SAGE, CADD-GADD, AD-GWAS, and OZ-ALC studies were 0.44, 0.39, 0.55, and 0.38, respectively.

In \(\text{LAMA5}\), two of the four SNPs implicated in any dataset were not in the 1000G so we cannot be sure if they are in LD with the other SNPs. However, the two remaining SNPs, although not in LD at \(r^2 > 0.5\), were associated with AD at \(p < 0.05\) in two different studies (Figure 1A).
Figure 1. Schematic of the SNPs associated with AD and their genomic location on the Human Genome Build hg19. The solid gray bars represent each gene. The black arrow indicates the direction of transcription. SNPs associated with AD that passed the p-value and LD threshold in each gene are shown above the gene, color-coded by study: green for CADD-GADD, red for SAGE, blue for OZ-ALC, and gray for AD-GWAS. The p-values for the SNPs in the other three datasets are shown below the gene, again color-coded by study. R^2 values are from the 1000 Genomes Pilot 1 data. A: LAMA5; B: RAC1; C: DNAJA3.

Of the six SNPs associated across the four studies in RAC1, two were in LD at r^2 = 0.696.

These SNPs were associated with AD in the SAGE and OZ-ALC studies, and the SNP associated with AD in the OZ-ALC study was nominally associated with AD in the SAGE and CADD-GADD studies at p = 0.063 and p = 0.048, respectively (Figure 1B). DNAJA3 was only associated with AD in the CADD-GADD and AD-GWAS studies using PLINK; however, the SNPs associated with AD in CADD-GADD and AD-GWAS were in LD at r^2 = 0.676, suggesting

<table>
<thead>
<tr>
<th></th>
<th>rs1760073</th>
<th>rs4925380^a</th>
<th>rs2427301</th>
<th>rs13036495^a</th>
</tr>
</thead>
<tbody>
<tr>
<td>CADD-GADD:</td>
<td>0.136</td>
<td>0.018</td>
<td>0.048</td>
<td>0.230</td>
</tr>
<tr>
<td>SAGE:</td>
<td>0.170</td>
<td>0.149</td>
<td>0.063</td>
<td>0.097</td>
</tr>
<tr>
<td>OZ-ALC:</td>
<td>0.003</td>
<td>0.019</td>
<td>0.044</td>
<td>0.157</td>
</tr>
<tr>
<td>AD-GWAS:</td>
<td>0.453</td>
<td>0.673</td>
<td>0.784</td>
<td>0.815</td>
</tr>
</tbody>
</table>

^a Not in the 1000 Genomes Pilot 1 data
the same signal emerged in each study. However, these SNPs were not associated with AD in either the SAGE or OZ-ALC studies (Figure 1C).

Discussion

The present study sought to investigate associations in a manually curated set of genes whose encoded proteins have been shown to interact with nAChRs in various ways with AD. Our analyses revealed three genes associated with AD at $p < 0.05$ in both analytical methods used. Although these associations did not survive correction for multiple testing they are driven by more than one study, strengthening the evidence that these genes are important for alcohol behaviors in human populations. This is the first work, to our knowledge, investigating this particular set of genes for association with AD.

$LAMA5$, $RAC1$, and $DNAJA3$ were nominally associated with AD in more than one study and have been implicated in nAChR function. $LAMA5$ codes for one of the vertebrate laminin alpha chains. Laminins are important for the maturation, stabilization and clustering of nAChRs at the neuromuscular junction, both in conjunction with and independently of known nAChR-associated proteins (Burkin et al., 1998; Burkin et al., 2000; Denzer et al., 1997; Godfrey et al., 1988; Lee et al., 2002; Montanaro et al., 1998; Nishimune et al., 2008; Sugiyama et al., 1997; Vogel et al., 1983). Laminin receptors are also differentially expressed in multiple brain regions in rats exposed to nicotine (Hwang et al., 2006). $RAC1$, a GTP binding protein, is involved in a unique nAChR internalization pathway that acts independently of typical processes triggering internalization (Borroni et al., 2011; Kumari et al., 2008). Finally, $DNAJA3$ (or $TID1$), a protein that stimulates heat shock protein chaperones, has been shown to bind a major component of the agrin receptor and colocalize with nAChRs at the neuromuscular junction. Knockdown of $DNAJA3$ impaired neuromuscular transmission and caused dispersal of nAChR clusters. Overexpression of $DNAJA3$ induced phosphorylation and clustering of nAChRs (Linnoila et al., 2008).
From our post hoc analyses of the SNPs implicated in *LAMA5*, *RAC1*, and *DNAJA3*, several signals for future investigation emerged. Although the SNPs in *DNAJA3* associated with AD in the CADD-GADD and AD-GWAS were in high LD ($r^2 = 0.676$), this signal was not replicated in either SAGE or OZ-ALC. However, two SNPs in *LAMA5* were associated at $p < 0.05$ in two different datasets; these SNPs are not in high LD at $r^2 > 0.5$ but merit future investigation. Finally, two SNPs in high LD ($r^2 = 0.696$) were associated with AD in *RAC1*. One of these SNPs, rs836559, was associated in the CADD-GADD and OZ-ALC samples and trending in the SAGE sample. Although this SNP, nor the SNP in LD with it, was associated in the AD-GWAS sample, this sample had the smallest sample size and likely did not have enough power to detect the small effect size we expect from these variants. Future work should evaluate the functionality of this downstream region of *RAC1*.

There are some methodological differences in the JAG and PLINK analyses. As the p-values for all the SNPs are simply summed to create the gene p-value, JAG performs better when there are many signals in the gene that contribute risk to the phenotype. Conversely, the set-based test in PLINK selects only the most significant signals ($p < 0.05$) to calculate the set statistic. Thus, a gene with a few very strong signals would appear significant using the PLINK test but not the JAG test. A gene with many weak signals would appear significant using the JAG test but not the PLINK test. Regardless, the utilization of two methodologically different tests serves as confirmation that a particular gene is worth future investigation and enables us to refine the list of future targets for alcohol studies.

There are several strengths of the analyses presented here. First, of seven studied pathway analysis algorithms (none of which were JAG as it is a newer tool), the PLINK set test was found to be the most powerful (Gui et al., 2011). Second, SNP-set based approaches have several advantages to GWAS: by consolidating SNP associations they can increase power to detect genetic loci with individually small effects, and can help to prioritize associations based on biological relevance (Purcell et al., 2009; Wang et al., 2011). Third, two analytical
approaches were used to compare the reproducibility of each analysis. Fourth, using a meta-
analysis over a mega-analysis allowed the comparison of p-values across studies for individual
genes. Finally, by curating a unique list of genes to test we avoided potential limitations when
examining already curated gene sets like KEGG and GO who have an inherent bias for genes
that have been well studied and more thoroughly annotated. As of 2011, only about 5,000
human genes had been annotated to KEGG pathways (Wang et al., 2011).

The present study is not without limitations. Despite being able to curate a unique set of
genes, this study was limited to testing SNPs annotated on the hg19 build and genes annotated
on all GWAS utilized. The correlations between the empirical p-values from the JAG and PLINK
analyses for each study were moderate. This is likely because PLINK does not calculate a test
statistic or empirical p-value if a gene has no SNPs with p < 0.05. If these criteria were less
stringent the correlations between the two analyses would likely be more comparable.

While several genes were associated with AD in the present study, these results do not
rule out a role for the other genes tested in AD and future studies are warranted. Common
variations (like those in GWAS) account for 21% (slightly less than half) of the variation in AD
(Vrieze et al., 2013), suggesting that rare variants in these genes could play a role in AD.
Furthermore, our analyses only included variation 20kb up- and downstream of each gene and it
is possible that farther upstream and trans regulatory elements play a role in AD.

In conclusion, we have identified in human samples several genes previously thought to
be involved in nAChR upregulation that were associated with AD: LAMA5, RAC1, and DNAJA3.
In particular, a signal downstream of RAC1 was implicated in AD in three of the four datasets
examined and merits future investigation. The list of genes examined is by no means an
exhaustive list, and simply served to test the hypothesis that SNPs in the identified genes are
associated with AD. Our results suggest these SNPs in the three aforementioned genes may
alter nAChR function and/or expression through protein-protein interactions. As mentioned
above however, it is also likely that genes on our list not associated with AD in the current study
may nevertheless be affecting nAChR upregulation in a manner not directly driven by SNPs. As these associations did not survive correction for multiple testing, the inclusion of independent samples or replication in a group of unrelated GWAS would strengthen these associations. This work highlights the importance of neuroscience experts and statistical geneticists collaborating to identify and characterize novel genetic associations.
CHAPTER VI
TARGETED GENE SEQUENCING

Introduction

As an extension of previous work presented in this dissertation, several genes of interest were appended to an existing sequencing project and sequenced. The regions were sequenced using targeted gene sequencing, as opposed to Whole Genome Sequencing (WGS) or Whole Exome Sequencing (WES). There were several reasons for this: first, targeted gene sequencing is preferable when there is a specific hypothesis concerning a gene and a particular phenotype; second, the purpose of WGS and WES is gene-discovery and by definition all genes sequenced are tested for association, requiring a strict correction for multiple testing; third, targeted gene sequencing is cheaper than WGS and WES; and lastly, although WES is less expensive than WGS, functional regions (such as untranslated regions, or UTRs) are often not covered and there is increasing evidence that functional regions are important in polygenic traits (Schork et al., 2013).

Four genes were sequenced and analyzed: GRM7; fibroblast growth factor receptor 2 (FGFR2); Ca²⁺/calmodulin-dependent protein kinase II alpha (CAMK2A); and FYN, an Src family tyrosine kinase. Although the evidence suggesting a role for GRM7 in alcohol consumption was primarily negative, it is important to reiterate that GRM7 is a large gene and many signals were not covered by the GWAS examined in Chapter III. Furthermore, it is likely that both common and rare variation contributes to complex disorders and rare variation in GRM7 was not examined in Chapter III. FGFR2 was sequenced based on preliminary evidence that it was associated with alcohol dependence (AD) in the CADD-GADD GWAS using the set test in PLINK (for a description of the PLINK set test see Chapters III, IV, and V) (Table 1).
Table 1. Preliminary results showing association with FGFR2 and AD using the PLINK set test in CADD-GADD. The test was run using all available SNPs (including 20 kb up- and downstream of the gene).

<table>
<thead>
<tr>
<th>Set</th>
<th>nSNP</th>
<th>nSIG</th>
<th>iSIG</th>
<th>P_emp</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGFR2</td>
<td>128</td>
<td>16</td>
<td>10</td>
<td>0.055</td>
</tr>
</tbody>
</table>

SNP Statistics\(^d\)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Minor Allele</th>
<th>(\beta)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3135802</td>
<td>T</td>
<td>0.226</td>
<td>0.035</td>
</tr>
<tr>
<td>rs4647919</td>
<td>T</td>
<td>1.040</td>
<td>0.001</td>
</tr>
<tr>
<td>rs3135807</td>
<td>T</td>
<td>1.046</td>
<td>0.001</td>
</tr>
<tr>
<td>rs3135808</td>
<td>A</td>
<td>1.012</td>
<td>0.002</td>
</tr>
<tr>
<td>rs6585740</td>
<td>G</td>
<td>-0.259</td>
<td>0.021</td>
</tr>
<tr>
<td>rs2114685</td>
<td>G</td>
<td>-0.251</td>
<td>0.027</td>
</tr>
<tr>
<td>rs2556537</td>
<td>T</td>
<td>0.232</td>
<td>0.031</td>
</tr>
<tr>
<td>rs1613776</td>
<td>T</td>
<td>0.483</td>
<td>0.031</td>
</tr>
<tr>
<td>rs3135817</td>
<td>C</td>
<td>0.197</td>
<td>0.032</td>
</tr>
<tr>
<td>rs1649166</td>
<td>T</td>
<td>0.226</td>
<td>0.035</td>
</tr>
</tbody>
</table>

\(^a\) Total number of SNPs tested in set
\(^b\) Total number of SNPs below p-value threshold (p < 0.05)
\(^c\) Number of significant SNPs also passing LD criteria \((r^2 = 1)\)
\(^d\) List of significant SNPs passing LD criteria and SNP statistics from the PLINK test

FGFR2 was no longer significant after using SNPs common to the four GWAS used in Chapter V. This suggested the analysis missed important signals and sequencing would allow a more comprehensive examination of the gene. Lastly, FYN and CAMK2A were sequenced based on preliminary evidence for association with FTND using JAG and PLINK. These associations remained in the analysis presented in Chapter IV and can be referenced in Table 3 on page 76.

The purpose of this study was to evaluate the evidence for association with rare variants (those variants not tagged in Genome-Wide Association Studies with low minor allele frequency, MAF < 0.01 or 0.05) in our genes of interest and alcohol and nicotine behaviors. Below the methods and results for this study will be discussed.

Materials and methods

Samples

Subjects were selected from the CADD (Stallings et al., 2005; Stallings et al., 2003) and GADD (Kamens et al., 2013), discussed in previous chapters. Phenotypic data for all subjects was collected with the Composite International Diagnostic Interview-Substance Abuse Module.
(CIDI-SAM) (Cottler et al., 1990). From this questionnaire a measure of Dependence
Vulnerability (DV) was created. Mathematically, this variable is the number of DSM-IV
dependence criteria endorsed across ten classes of drugs, divided by the number of drugs used
(Stallings et al., 2003). Conduct Disorder (CD) was also assessed in these samples with the
Diagnostic Interview Schedule for DSM-IV (DIS) (Robins et al., 1981) for subjects over 18 years
of age, or the Diagnostic Interview Schedule for Children (DISC) (Shaffer et al., 1993) for
subjects under 18 years of age. Subjects were selected for sequencing based on a composite
score including both DV and CD as the purpose of the sequencing study was to assess the
genetic risk for different classes of drugs and the comorbidity between them. Using the residuals
from a linear regression in SAS (SAS Institute Inc., Cary, NC, USA), the score was further
standardized on sex, age, and age$^2$ (identical to the standardization procedures outlined in
Chapters II and III). In order to enrich the sample for ethnic diversity, ideally leading to the
identification of multiple novel genetic variants, individuals of non-European descent were up
weighted on the composite DV/CD score. For individuals with positive scores, 0.5 extra point
were given to those subjects who self-identified as Hispanic; an additional one point was added
for non-European subjects. The reverse procedure was performed for those subjects with
negative composite scores. The subjects to be sequenced were CADD or GADD subjects who
had been previously genotyped in a GWAS study (previously mentioned in Chapters IV and V)
(Derringer et al., 2015) and represented the extreme ends of the composite DV/CD score
(controls had z scores ≤ -0.66, cases had z scores ≥ 3.88). The final sample size was 100 cases
and 100 controls.

**Sequencing**

The Repli-G kit was used to amplify genomic DNA, according to manufacturer’s
instructions (Qiagen, Hilden, Germany). DNA samples were send to Centrillion Bioscience
(https://www.centrillionbio.com/) where a proprietary custom enrichment strategy was performed
for the regions of interest. Each gene, with the exception of *GRM7*, was targeted to include
5,000 base pairs up- and downstream of the coding region (Table 2).

**Table 2.** Target regions for each gene. Chromosome, start and end base pair (bp) were
taken using the Human Genome Build hg19 from the UCSC Genome Browser (Kent et al.,
2002).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>Start bp</th>
<th>End bp</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>GRM7</em></td>
<td>3</td>
<td>6852802</td>
<td>7788218</td>
<td>-935416</td>
</tr>
<tr>
<td><em>FGFR2</em></td>
<td>10</td>
<td>123238212</td>
<td>123362972</td>
<td>-124760</td>
</tr>
<tr>
<td>CAMK2A</td>
<td>5</td>
<td>149625251</td>
<td>149641208</td>
<td>-15957</td>
</tr>
<tr>
<td>FYN</td>
<td>6</td>
<td>112010578</td>
<td>112029237</td>
<td>-18659</td>
</tr>
</tbody>
</table>

A larger region was targeted upstream of *GRM7* as research has suggested a role for *cis-*
regulatory elements in alcohol behaviors (Vadasz et al., 2007). The targeted regions were
sequenced on an Illumina HiSeq 2000 (Illumina, San Diego, CA, USA).

**Mapping, variant detection, and annotation**

The Burrows-Wheeler Aligner (BWA, v0.7.3a-r367) (Li et al., 2009a) was used to map
raw sequencing reads to the Human Genome Build hg19/dbSNP Build 137. Duplicate reads
were identified with Picard (v1.85, [http://picard.sourceforge.net](http://picard.sourceforge.net)). Following the Broad Institute’s
nucleotide variants were classified using the Genome Analysis Toolkit (GATK, v2,3) (DePristo et
al., 2011; McKenna et al., 2010). In short, base quality scores were recalibrated based on
known variant locations after reads were realigned locally to recognized indels. The
UnifiedGenotyper tool in GATK was used to call variants and ANNOVAR
([www.openbioinformatics.org/annovar](http://www.openbioinformatics.org/annovar)) was used to annotate the final set of variants using the
Human Genome Build hg19/dbSNP Build 137.

**Statistical analysis**

*Phenotypes*
Although the subjects were chosen based on their DV/CD composite scores at wave 1, phenotypic data was preferably taken from the latest wave of data collection, wave 3, to ensure that subjects were past the age of initiation for most substances. If no data were available at wave 3, data from wave 2 were utilized, and if no data were available from wave 2, data from wave 1 were used. The phenotype for each analysis was chosen based on preliminary results.

**GRM7**

In Chapter III, results suggesting an association between rs3749380 and maximum number of drinks consumed in 24 hours (AC) for subjects who had drank alcohol in the CADD were presented; this result remained when all subjects were included in the analysis (e.g. subjects who had not consumed alcohol were included). The phenotype utilized in the present study was AC for all subjects, standardized in the whole sample and adjusted for age, sex, and age\(^2\) (as described in Chapter III). Subjects who had never consumed alcohol were included in order to maximize the sample size. The first two principal components were associated with AC using a regression and thus included in the analysis to adjust for population stratification.

**FGFR2**

Quantitative number of AD symptoms in all subjects, regardless of whether or not the subject had previously drank alcohol, was assessed for association with variants in FGFR2. As described in Chapter V, DSM IV AD was assessed using the Composite International Diagnostic Interview-Substance Abuse Module (CIDI-SAM) (Cottler et al., 1990). Although initial regression analysis of the principal components on the phenotype showed no association (over what is expected by chance) with AD, the analysis was run using age and sex as covariates, and repeated using age, sex, and the first 5 principal components as covariates.

**CAMK2A and FYN**
Although data presented in Chapter IV suggested evidence for association with FTND and these genes, FTND was not available in this dataset. A comparable phenotype, DSM IV nicotine dependence (ND), was assessed using the Composite International Diagnostic Interview-Substance Abuse Module (CIDI-SAM) (Cottler et al., 1990) and used in the present analysis. All subjects were included (e.g. subjects who had not smoked before were included to maximize the sample size). Although initial regression analysis of the principal components on the phenotype showed no association (over what is expected by chance) with ND, the analysis was run using age and sex as covariates, and repeated using age, sex, and the first 5 principal components as covariates.

**Association analysis**

The analysis was performed using SKAT-O (Lee et al., 2012), implemented in an R (Team, 2012) package. Two commonly used methods to analyze sequencing data are burden and non-burden tests. Burden tests assume all rare variants have the same direction of effect and magnitude on the phenotype; in other words, all rare variants in a region would be either deleterious or protective for the phenotype. The Sequence Kernel Association Test (SKAT) (Wu et al., 2011) was recently proposed to account for unusually distributed rare variants in sequencing analyses. Although less powerful than the burden test when variants are in the same direction this test is robust in the presence of both protective and deleterious variants. SKAT-O is an extension of SKAT in which an optimal test is derived from a burden and typical SKAT test (Lee et al., 2012).

As the sequenced regions were large, it was necessary to break each region into groups of variants to test in SKAT-O. Two grouping approaches were used: a sliding window approach and a functional approach. For the sliding window approach, variants were grouped in consecutive sets. Each set consisted of approximately 50 variants (with the exception of the last set) and overlapped by 10 variants. Only rare variants were included in the analysis because
the specific goal of this project was to characterize rare variation in the region. Only variants with MAF ≤ 1% in two outside sequencing projects, Phase 1 of the 1,000 Genomes project (1000G) (Abecasis et al., 2012), and the National Heart, Lung, and Blood Institute’s Grand Opportunity Exome Sequencing Project (ESP) (Tennessen et al., 2012), were included (where a variant not in either outside project was considered to have MAF = 0% and was thus included in the analysis). Variants with MAF > 5% in both cases and controls from the 200 sequenced individuals were excluded. For the functional approach, variants were selected in a similar but more lenient manner: variants with MAF ≤ 5% in the 1000G and ESP were included, and variants with MAF > 5% in both cases and controls from the 200 sequenced individuals were excluded. Functional groups were established as follows: variants in UTRs and up- and downstream regions; variants in exons; and variants in UTRs, up- and downstream regions, and exons. A non-coding RNA (ncRNA) sequenced in GRM7 was also included as its own functional group.

Results

As mentioned previously the sample was enriched for ethnic diversity, with 45-68% of cases and controls of European American descent (Table 3).

Table 3. Phenotypic descriptive for the subjects in each sequencing analysis.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Age a</th>
<th>% Male</th>
<th>% Hispanic</th>
<th>% White</th>
<th>AC b</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GRM7</strong> Cases (N = 100)</td>
<td>20.7±4.16</td>
<td>68%</td>
<td>36%</td>
<td>45%</td>
<td>7.11±7.26</td>
</tr>
<tr>
<td>Controls (N = 100)</td>
<td>25.4±3.87</td>
<td>74%</td>
<td>23%</td>
<td>68%</td>
<td>2.33±3.11</td>
</tr>
<tr>
<td><strong>FGFR2</strong> Cases (N = 100)</td>
<td>20.8±4.19</td>
<td>68%</td>
<td>36%</td>
<td>45%</td>
<td>3.53±2.22</td>
</tr>
<tr>
<td>Controls (N = 100)</td>
<td>25.7±3.68</td>
<td>74%</td>
<td>23%</td>
<td>68%</td>
<td>0.85±1.17</td>
</tr>
<tr>
<td><strong>CAMK2A and FYN</strong> Cases (N = 100)</td>
<td>20.8±4.20</td>
<td>68%</td>
<td>36%</td>
<td>45%</td>
<td>3.87±2.09</td>
</tr>
<tr>
<td>Controls (N = 100)</td>
<td>25.8±3.47</td>
<td>74%</td>
<td>23%</td>
<td>68%</td>
<td>0.94±1.70</td>
</tr>
</tbody>
</table>

a Mean ± standard deviation
b Mean ± standard deviation, not adjusted for age, sex, and age²
Although the phenotypes for were preferably taken from wave 3, wave 2 if the subject was not assessed at wave 3, and finally wave 1 if the subject was not assessed at either wave 3 or 2, there is some difference in ages between the cases and controls in each analyses. This is due to the fact the \textit{GRM7} work was completed substantially earlier than the \textit{FGFR2} work and the \textit{FGFR2} work completed slightly before the \textit{CAMK2A/FYN} analyses. As wave 3 is currently ongoing, subjects are continuously added to the data and as a result the earlier work includes slightly younger subjects than the later work.

Eight thousand three hundred and eighty nine variants were identified in \textit{GRM7} (Figure 1).

\textbf{Figure 1.} Map of variants identified in the sequenced region around \textit{GRM7}. Arrows indicate the direction of transcription. Novel indicates the variant was not present in dbSNP 137, 1000G, or ESP. S: synonymous, NS: nonsynonymous.
The average read depth for the region was 458.5 after adjusting for differences in the total number of reads. The transition to transversion (Ti/Tv) ratio among all 8,389 variants was 1.81.

Of the 115 groups from the sliding window approach and 4 functional regions tested for association with AC using SKAT-O, none survived correction for multiple testing at Family-Wise Error Rate (FWER) = 0.05 (Table 4).

**Table 4.** Results for each set from the SKAT-O test in *GRM7.*

<table>
<thead>
<tr>
<th>Set</th>
<th># Markers</th>
<th># Tested Markers</th>
<th>P-value</th>
<th>P-value (log transformed AC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>50</td>
<td>50</td>
<td>0.265</td>
<td>0.158</td>
</tr>
<tr>
<td>Set 2</td>
<td>50</td>
<td>50</td>
<td>0.827</td>
<td>0.712</td>
</tr>
<tr>
<td>Set 3</td>
<td>50</td>
<td>50</td>
<td>0.619</td>
<td>0.666</td>
</tr>
<tr>
<td>Set 4</td>
<td>50</td>
<td>50</td>
<td>0.915</td>
<td>1.000</td>
</tr>
<tr>
<td>Set 5</td>
<td>50</td>
<td>50</td>
<td>0.674</td>
<td>0.362</td>
</tr>
<tr>
<td>Set 6</td>
<td>50</td>
<td>50</td>
<td>0.497</td>
<td>0.534</td>
</tr>
<tr>
<td>Set 7</td>
<td>50</td>
<td>50</td>
<td>0.121</td>
<td>0.236</td>
</tr>
<tr>
<td>Set 8</td>
<td>50</td>
<td>50</td>
<td>0.445</td>
<td>0.362</td>
</tr>
<tr>
<td>Set 9</td>
<td>50</td>
<td>50</td>
<td>0.382</td>
<td>0.414</td>
</tr>
<tr>
<td>Set 10</td>
<td>50</td>
<td>50</td>
<td>0.254</td>
<td>0.239</td>
</tr>
<tr>
<td>Set 11</td>
<td>50</td>
<td>50</td>
<td>0.082</td>
<td>0.199</td>
</tr>
<tr>
<td>Set 12</td>
<td>50</td>
<td>50</td>
<td>0.136</td>
<td>0.213</td>
</tr>
<tr>
<td>Set 13</td>
<td>50</td>
<td>50</td>
<td>0.230</td>
<td>0.170</td>
</tr>
<tr>
<td>Set 14</td>
<td>50</td>
<td>50</td>
<td>0.283</td>
<td>0.222</td>
</tr>
<tr>
<td>Set 15</td>
<td>50</td>
<td>50</td>
<td>0.276</td>
<td>0.211</td>
</tr>
<tr>
<td>Set 16</td>
<td>50</td>
<td>50</td>
<td>0.475</td>
<td>0.580</td>
</tr>
<tr>
<td>Set 17</td>
<td>50</td>
<td>50</td>
<td>0.237</td>
<td>0.369</td>
</tr>
<tr>
<td>Set 18</td>
<td>50</td>
<td>50</td>
<td>0.051</td>
<td>0.086</td>
</tr>
<tr>
<td>Set 19</td>
<td>50</td>
<td>50</td>
<td>0.009</td>
<td>0.008</td>
</tr>
<tr>
<td>Set 20</td>
<td>50</td>
<td>50</td>
<td>0.258</td>
<td>0.241</td>
</tr>
<tr>
<td>Set 21</td>
<td>50</td>
<td>50</td>
<td>0.473</td>
<td>0.519</td>
</tr>
<tr>
<td>Set 22</td>
<td>50</td>
<td>50</td>
<td>0.494</td>
<td>0.513</td>
</tr>
<tr>
<td>Set 23</td>
<td>50</td>
<td>50</td>
<td>0.130</td>
<td>0.250</td>
</tr>
<tr>
<td>Set 24</td>
<td>50</td>
<td>50</td>
<td>0.508</td>
<td>0.526</td>
</tr>
<tr>
<td>Set 25</td>
<td>50</td>
<td>50</td>
<td>0.704</td>
<td>0.674</td>
</tr>
<tr>
<td>Set 26</td>
<td>50</td>
<td>50</td>
<td>0.467</td>
<td>0.377</td>
</tr>
<tr>
<td>Set 27</td>
<td>50</td>
<td>50</td>
<td>0.378</td>
<td>0.344</td>
</tr>
<tr>
<td>Set 28</td>
<td>50</td>
<td>49</td>
<td>0.668</td>
<td>0.655</td>
</tr>
<tr>
<td>Set 29</td>
<td>50</td>
<td>50</td>
<td>0.287</td>
<td>0.422</td>
</tr>
<tr>
<td>Set 30</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.796</td>
</tr>
<tr>
<td>Set 31</td>
<td>50</td>
<td>50</td>
<td>0.134</td>
<td>0.135</td>
</tr>
<tr>
<td>Set 32</td>
<td>50</td>
<td>50</td>
<td>0.186</td>
<td>0.134</td>
</tr>
<tr>
<td>Set</td>
<td>X</td>
<td>Y</td>
<td>Value 1</td>
<td>Value 2</td>
</tr>
<tr>
<td>------</td>
<td>----</td>
<td>----</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>33</td>
<td>50</td>
<td>50</td>
<td>0.696</td>
<td>0.825</td>
</tr>
<tr>
<td>34</td>
<td>50</td>
<td>50</td>
<td>0.564</td>
<td>0.641</td>
</tr>
<tr>
<td>35</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.828</td>
</tr>
<tr>
<td>36</td>
<td>50</td>
<td>50</td>
<td>0.600</td>
<td>0.758</td>
</tr>
<tr>
<td>37</td>
<td>50</td>
<td>50</td>
<td>0.487</td>
<td>0.605</td>
</tr>
<tr>
<td>38</td>
<td>50</td>
<td>50</td>
<td>0.363</td>
<td>0.626</td>
</tr>
<tr>
<td>39</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.828</td>
</tr>
<tr>
<td>40</td>
<td>50</td>
<td>50</td>
<td>0.600</td>
<td>0.758</td>
</tr>
<tr>
<td>41</td>
<td>50</td>
<td>50</td>
<td>0.487</td>
<td>0.537</td>
</tr>
<tr>
<td>42</td>
<td>50</td>
<td>50</td>
<td>0.807</td>
<td>0.772</td>
</tr>
<tr>
<td>43</td>
<td>50</td>
<td>50</td>
<td>0.250</td>
<td>0.187</td>
</tr>
<tr>
<td>44</td>
<td>50</td>
<td>50</td>
<td>0.850</td>
<td>0.832</td>
</tr>
<tr>
<td>45</td>
<td>50</td>
<td>50</td>
<td>0.795</td>
<td>0.733</td>
</tr>
<tr>
<td>46</td>
<td>50</td>
<td>50</td>
<td>0.773</td>
<td>0.876</td>
</tr>
<tr>
<td>47</td>
<td>50</td>
<td>50</td>
<td>0.819</td>
<td>0.876</td>
</tr>
<tr>
<td>48</td>
<td>50</td>
<td>50</td>
<td>0.744</td>
<td>0.885</td>
</tr>
<tr>
<td>49</td>
<td>50</td>
<td>50</td>
<td>0.811</td>
<td>0.911</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>0.707</td>
<td>0.661</td>
</tr>
<tr>
<td>51</td>
<td>50</td>
<td>50</td>
<td>0.239</td>
<td>0.228</td>
</tr>
<tr>
<td>52</td>
<td>49</td>
<td>49</td>
<td>0.306</td>
<td>0.224</td>
</tr>
<tr>
<td>53</td>
<td>50</td>
<td>50</td>
<td>0.024</td>
<td>0.030</td>
</tr>
<tr>
<td>54</td>
<td>50</td>
<td>50</td>
<td>0.458</td>
<td>0.349</td>
</tr>
<tr>
<td>55</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.896</td>
</tr>
<tr>
<td>56</td>
<td>50</td>
<td>50</td>
<td>0.786</td>
<td>0.745</td>
</tr>
<tr>
<td>57</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>58</td>
<td>50</td>
<td>50</td>
<td>0.929</td>
<td>0.916</td>
</tr>
<tr>
<td>59</td>
<td>50</td>
<td>50</td>
<td>0.931</td>
<td>0.908</td>
</tr>
<tr>
<td>60</td>
<td>50</td>
<td>50</td>
<td>0.596</td>
<td>0.557</td>
</tr>
<tr>
<td>61</td>
<td>50</td>
<td>50</td>
<td>0.525</td>
<td>0.505</td>
</tr>
<tr>
<td>62</td>
<td>50</td>
<td>50</td>
<td>0.456</td>
<td>0.428</td>
</tr>
<tr>
<td>63</td>
<td>50</td>
<td>50</td>
<td>0.714</td>
<td>0.725</td>
</tr>
<tr>
<td>64</td>
<td>50</td>
<td>50</td>
<td>0.498</td>
<td>0.351</td>
</tr>
<tr>
<td>65</td>
<td>50</td>
<td>50</td>
<td>0.833</td>
<td>0.664</td>
</tr>
<tr>
<td>66</td>
<td>50</td>
<td>50</td>
<td>0.405</td>
<td>0.230</td>
</tr>
<tr>
<td>67</td>
<td>50</td>
<td>50</td>
<td>0.248</td>
<td>0.271</td>
</tr>
<tr>
<td>68</td>
<td>50</td>
<td>50</td>
<td>0.410</td>
<td>0.347</td>
</tr>
<tr>
<td>69</td>
<td>50</td>
<td>50</td>
<td>0.462</td>
<td>0.459</td>
</tr>
<tr>
<td>70</td>
<td>50</td>
<td>50</td>
<td>0.622</td>
<td>0.342</td>
</tr>
<tr>
<td>71</td>
<td>50</td>
<td>50</td>
<td>0.454</td>
<td>0.280</td>
</tr>
<tr>
<td>72</td>
<td>50</td>
<td>50</td>
<td>0.181</td>
<td>0.152</td>
</tr>
<tr>
<td>73</td>
<td>50</td>
<td>50</td>
<td>0.861</td>
<td>0.796</td>
</tr>
<tr>
<td>74</td>
<td>50</td>
<td>50</td>
<td>0.530</td>
<td>0.554</td>
</tr>
<tr>
<td>75</td>
<td>50</td>
<td>49</td>
<td>0.188</td>
<td>0.201</td>
</tr>
<tr>
<td>76</td>
<td>50</td>
<td>49</td>
<td>0.887</td>
<td>1.000</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.754</td>
<td>0.699</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.198</td>
<td>0.356</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.057</td>
<td>0.151</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.622</td>
<td>0.526</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.659</td>
<td>0.565</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.198</td>
<td>0.771</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.169</td>
<td>0.212</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.113</td>
<td>0.087</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.043</td>
<td>0.040</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.191</td>
<td>0.244</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.851</td>
<td>0.843</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.868</td>
<td>0.774</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.528</td>
<td>0.597</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.748</td>
<td>0.775</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>49</td>
<td>0.882</td>
<td>0.758</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.931</td>
<td>0.690</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.921</td>
<td>0.879</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.835</td>
<td>0.879</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.705</td>
<td>0.511</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.021</td>
<td>0.024</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.417</td>
<td>0.376</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.094</td>
<td>0.083</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.315</td>
<td>0.195</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.751</td>
<td>0.821</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.848</td>
<td>1.000</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.453</td>
<td>0.350</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.118</td>
<td>0.202</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.841</td>
<td>1.000</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.764</td>
<td>0.739</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.723</td>
<td>0.783</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.863</td>
<td>0.873</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.805</td>
<td>0.865</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.784</td>
<td>0.805</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.300</td>
<td>0.300</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.650</td>
<td>0.523</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.430</td>
<td>0.434</td>
</tr>
<tr>
<td>Set</td>
<td>50</td>
<td>50</td>
<td>0.112</td>
<td>0.210</td>
</tr>
<tr>
<td>Set</td>
<td>21</td>
<td>21</td>
<td>0.082</td>
<td>0.105</td>
</tr>
<tr>
<td>Funct1</td>
<td>16</td>
<td>16</td>
<td>0.853</td>
<td>0.907</td>
</tr>
<tr>
<td>Funct2</td>
<td>12</td>
<td>12</td>
<td>0.921</td>
<td>0.694</td>
</tr>
<tr>
<td>Funct3</td>
<td>28</td>
<td>28</td>
<td>1.000</td>
<td>0.901</td>
</tr>
<tr>
<td>Funct4</td>
<td>69</td>
<td>69</td>
<td>0.884</td>
<td>0.788</td>
</tr>
</tbody>
</table>

*a Several markers were not tested due to missingness*
Adjusted for age, sex, age^2, and the first two principal components

Log_{10} transformed phenotype, adjusted for age, sex, age^2, and the first two principal components

Functional grouping: UTR and up- and downstream regions

Functional grouping: exonic variants only

Functional grouping: UTR, exons, and up- and downstream regions

Functional grouping: ncRNA variants

Only binary variants were included in the analysis, resulting in a total of 8,365 grouped variants.

The analysis was repeated using a Log_{10} transformation of the phenotype, with similar results.

Eight hundred and fifty nine variants were identified around \textit{FGFR2} (Figure 2).

\textbf{Figure 2.} Map of variants identified in the sequenced region around \textit{FGFR2}. Arrows indicate the direction of transcription. Novel indicates the variant was not present in dbSNP 137, 1000G, or ESP. S: synonymous, NS: nonsynonymous.

The average read depth for the region was 551.8 after adjusting for differences in the total number of reads. The Ti/Tv ratio was 2.67.

Of the 13 groups from the sliding window approach and 3 functional groups, 1 group (set 9) was associated with AD with FWER = 0.05 after adjusting for age and sex (Table 5).
Table 5. Results for each set from the SKAT-O test in FGFR2.

<table>
<thead>
<tr>
<th>Set</th>
<th># Markers</th>
<th># Tested Markers</th>
<th>P-value&lt;sup&gt;a&lt;/sup&gt;</th>
<th>P-value&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>50</td>
<td>50</td>
<td>0.882</td>
<td>1.000</td>
</tr>
<tr>
<td>Set 2</td>
<td>50</td>
<td>50</td>
<td>0.915</td>
<td>0.698</td>
</tr>
<tr>
<td>Set 3</td>
<td>50</td>
<td>50</td>
<td>0.610</td>
<td>0.177</td>
</tr>
<tr>
<td>Set 4</td>
<td>50</td>
<td>50</td>
<td>0.543</td>
<td>0.687</td>
</tr>
<tr>
<td>Set 5</td>
<td>50</td>
<td>50</td>
<td>0.096</td>
<td>0.203</td>
</tr>
<tr>
<td>Set 6</td>
<td>50</td>
<td>50</td>
<td>0.769</td>
<td>0.898</td>
</tr>
<tr>
<td>Set 7</td>
<td>50</td>
<td>50</td>
<td>1.000</td>
<td>0.934</td>
</tr>
<tr>
<td>Set 8</td>
<td>50</td>
<td>50</td>
<td>0.357</td>
<td>0.339</td>
</tr>
<tr>
<td>Set 9</td>
<td>50</td>
<td>49</td>
<td>0.007&lt;sup&gt;g&lt;/sup&gt;</td>
<td>0.017&lt;sup&gt;h&lt;/sup&gt;</td>
</tr>
<tr>
<td>Set 10</td>
<td>50</td>
<td>50</td>
<td>0.182</td>
<td>0.187</td>
</tr>
<tr>
<td>Set 11</td>
<td>50</td>
<td>50</td>
<td>0.597</td>
<td>0.592</td>
</tr>
<tr>
<td>Set 12</td>
<td>50</td>
<td>50</td>
<td>0.699</td>
<td>0.902</td>
</tr>
<tr>
<td>Set 13</td>
<td>44</td>
<td>43</td>
<td>0.888</td>
<td>0.745</td>
</tr>
<tr>
<td>Funct1</td>
<td>27</td>
<td>27</td>
<td>0.855</td>
<td>0.740</td>
</tr>
<tr>
<td>Funct2</td>
<td>9</td>
<td>9</td>
<td>0.127</td>
<td>0.318</td>
</tr>
<tr>
<td>Funct3</td>
<td>36</td>
<td>36</td>
<td>0.750</td>
<td>0.934</td>
</tr>
</tbody>
</table>

<sup>a</sup> Several markers were not tested due to missingness  
<sup>b</sup> Adjusted for age and sex  
<sup>c</sup> Adjusted for age, sex, the first 5 principal components  
<sup>d</sup> Functional grouping: UTR and up- and downstream regions  
<sup>e</sup> Functional grouping: exonic variants only  
<sup>f</sup> Functional grouping: UTR, exons, and up- and downstream regions  
<sup>g</sup> Significant controlling for FWER = 0.05  
<sup>h</sup> Significant controlling for FWER = 0.5

Although no longer significant with FWER = 0.05, the same group remained significant after the inclusion of the first 5 principal components as covariates at FWER = 0.5. Of the 859 markers, 858 were biallelic and grouped for analysis.

Set 9 covered the base pairs from 123,312,481 – 123,325,169 on chromosome 10, according to the Human Genome Build hg19 from the UCSC genome browser (Kent et al., 2002). Of the 50 variants in the region, 3 were exonic and the remaining intronic. A linear regression using these variants were run in PLINK (Purcell et al., 2007) to test for an association with AD, after adjusting for age and sex. The results indicated 7 variants were driving the association at p < 0.018 (Figure 3), although not after the Bonferroni correction for multiple testing at p < 0.001.
Figure 3. Base pair position on the Human Genome Build hg19 of the variants in set 9 by $-\log_{10}$ of the p-value after association with AD.

Two hundred and thirty variants were identified in CAMK2A and FYN. All of the identified variants were bi-allelic. One hundred and five total variants were identified in CAMK2A (Figure 4).
Figure 4. Map of variants identified in the sequenced region around CAMK2A. Arrows indicate the direction of transcription. Novel indicates the variant was not present in dbSNP 137, 1000G, or ESP. S: synonymous, NS: nonsynonymous.

One hundred and twenty five variants were identified in FYN (Figure 5).

Figure 5. Map of variants identified in the sequenced region around FYN. Arrows indicate the direction of transcription. The red exon indicates a splice variant. Novel indicates the variant was not present in dbSNP 137, 1000G, or ESP. NS: nonsynonymous.
Both CAMK2A and FYN had high read depth after adjusting for the total number of mapped reads (444.0 and 602.3, respectively). The Ti/Tv ratio for the region covered by both genes was 3.60.

Since both CAMK2A and FYN are small genes, rather than employing the sliding window and functional approaches utilized above, each gene was tested as an individual set. Rare variants were selected as above; variants with MAF > 1% in the 1000G and ESP, as well as variants with MAF > 5% in both cases and controls were excluded. While CAMK2A showed no association with ND, FYN showed nominal association with ND at p < 0.054 but this association did not remain after the first 5 PCs were included as covariates (Table 6).

<table>
<thead>
<tr>
<th>Set</th>
<th># Markers</th>
<th># Tested Markers</th>
<th>P-value (adjusted for age, sex)</th>
<th>P-value (adjusted for age, sex, and 5 PCs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAMK2A</td>
<td>57</td>
<td>57</td>
<td>0.291</td>
<td>0.506</td>
</tr>
<tr>
<td>FYN</td>
<td>66</td>
<td>66</td>
<td>0.054</td>
<td>0.102</td>
</tr>
</tbody>
</table>

Discussion

This study represents the first attempt, to our knowledge, to characterize rare variants in GRM7, FGFR2, CAMK2A, and FYN. Although only sequenced in 200 individuals, all sequenced regions had high coverage. Only FGFR2 had a Ti/Tv ratio in between what is expected; the Ti/Tv ratio for SNPs, in exons, is around 3.0 and 2.0 for SNPs elsewhere (Bainbridge et al., 2011). However, these estimates are genome-wide and Ti/Tv ratios were calculated here for targeted gene sequencing. Nevertheless, all genes sequenced had a higher number of transitions than transversions, which is expected. Analyses in SKAT-O revealed putatively important regions of rare variants.

Although one SNP set, set 19, in GRM7 was nominally associated with AC this association did not survive correction for multiple testing. The set immediately upstream of set
19, set 18, was nominally associated with AC at p < 0.1, suggesting the region covered by these two sets may be in LD with a putative functional region.

One SNP set in FGFR2, set 9, was associated with AD with FWER = 0.05. A linear regression indicated that 7 SNPs were driving this association; of these 7 SNPs, 6 were in perfect LD. These 6 have MAF = 0.005 and are only present in one subject. Each variant had good coverage and the reads for each variant correspond to what is expected for a heterozygote (that is to say roughly 50% of the reads mapped to each allele), suggesting these variants are not due to sequencing error. This subject, a 15-year-old African American female of non-Hispanic descent, had high scores on AD.

Although FGFR2 was only nominally associated with AD in our preliminary analyses, evidence suggests a role for this gene in alcohol and nicotine behaviors. In smooth muscle cells, ethanol was shown to inhibit cell proliferation normally driven by FGFR2 ligands (Ghiselli et al., 2003). In vitro work has suggested FGFR2 acts through a signaling cascade involving L-type voltage gated Ca$^{2+}$ channels (VGCC) (Rosenthal et al., 2001) and although there is as of yet no direct link between FGFR2 and nAChRs, L-type VGCCs are modulated by both nicotine (Chang et al., 2001; Katsura et al., 2002) and alcohol (Dolin et al., 1987; Earl et al., 2011; Grant et al., 1993; Katsura et al., 2006; Katsura et al., 2005; Knott et al., 2002; Mah et al., 2011; Pietrzykowski et al., 2013; Walter et al., 1999; Wang et al., 1994; Wang et al., 1991a; Wang et al., 1991b) and may act in nAChR signaling pathways.

Despite evidence that CAMK2A plays a crucial role in learning and memory (Easton et al., 2013; Giese et al., 2013) as well as the development and maintenance of drug behaviors (Li et al., 2008a; Pierce et al., 1998; Tang et al., 2006), we did not find evidence for association of rare variants in CAMK2A with ND. Based on this, it is possible that testing for association of rare variants in CAMK2A and the DV phenotype might yield results. However, the gene-based test included regions 20kb up- and downstream of the gene, whereas our sequencing only included
regions 5kb up- and downstream of CAMK2A, so it is also possible that our sequencing study missed potentially important regulatory regions.

FYN was nominally associated with ND, although this association did not remain after correction for population stratification. Again, while it is possible our sequencing study missed regulatory regions as only 5kb up- and downstream of FYN were covered, it is also likely that with a larger sample size we would see a larger effect. Fyn and other Src Family Kinases (SFKs) are critical players in neuromuscular junction stabilization; normal activity of SFKs is required to maintain proper structure of the adult neuromuscular junction, and SFKs are responsible for the stabilization of the rapsyn-induced (Sadasivam et al., 2005) and agrin-induced (Smith et al., 2001) nAChR clusters. FYN is also known to phosphorylate several nAChR subunits (Fuhrer et al., 1996; Mohamed et al., 1999; Swope et al., 1994). Finally, evidence has suggested that FYN induces anchoring of nAChRs to the cytoskeleton in a rapsyn-dependent manner (Mohamed et al., 1999).

There are several limitations in the present study, most notably small sample size. Replication for all the findings in the present study is warranted. Additionally, the multiple testing corrections employed may have been too stringent as the groups of SNPs overlapped. Finally, although more of a caveat than a limitation, different GATK parameters may alter what is called a variant and none of the variants discussed above were confirmed with Sanger sequencing (Sanger et al., 1977). It is possible that some true variants were not called variants and vice versa.

The present study has several strengths. Although the sample size was small the regions sequenced had high read depth, suggesting the variants identified were not sequencing errors. Although a burden test would perform better if all the variants had the same direction of effect, preliminary analyses in FGFR2 (refer to Table 1 of the present Chapter on page 100) showed this not to be the case for this gene. Thus SKAT-O was a more powerful test to use.
Several future directions could strengthen this study. Sanger sequencing (Sanger et al., 1977) could be used to confirm the variants identified in this study and additional subjects could be sequenced in order to provide a replication sample. This study was designed to examine rare variation only, but common variants were also identified in each gene; as mentioned above, GWAS does not cover all the common variation in the genome and sequencing studies can further characterize common variation not covered in GWAS. Lastly, a sibling of the subject containing the 6 rare variants driving the FGFR2 association with AD was identified and can be sequenced in order to determine if those variants are de novo mutations. Although a parent of this subject would be ideal to sequence, no parental DNA is available in this sample.

The purpose of this study was to follow up on preliminary associations in genes with alcohol and tobacco behaviors and see if rare variants in these genes were also associated with these disorders. We found limited evidence for association with rare variants in GRM7 with AC. However, while twin studies estimate that AC is 43% heritable, additive SNP effects from GWAS account for 38% of this variation (Vrieze et al., 2013). While this does not rule out a role for rare variants in AC, it suggests regulatory elements upstream of GRM7, rather than rare variation, merit investigation. Similarly, this study found some evidence for association between rare variants in FYN, a critical player in stabilization of the neuromuscular junction, with ND, but no evidence for association with CAMK2A and ND. Finally, this is the first study to our knowledge to find association with rare variants in FGFR2 and AD. These findings should be followed up in a larger sample. This study provides a method for grouping variants into testable regions by taking into account prior sequencing studies.
CHAPTER VII
CONCLUSION

The studies presented here were designed to evaluate genetic influences on alcohol and nicotine behaviors in human populations. A common theme was the use of single nucleotide polymorphism (SNP) data to compare associations in different study populations and the implementation of different analytical methods to evaluate these associations. In particular, gene-based tests were used to test a set of genes for association with alcohol or nicotine behaviors to propose future candidates for research.

In Chapter II, SNPs previously associated with alcohol dependence (AD) and conduct disorder (CD) in GABRA2 were tested for alcohol abuse and dependence (AAD) as well as CD in two Colorado-based samples. Family based association tests were run using Hispanics and non-Hispanic European American subjects from two independent longitudinal samples. Our analysis provided nominal support for an association with rs9291283 and AAD in adulthood and CD in adolescence. Although the lack of findings between a well-studied SNP in GABRA2 with adolescent AAD was confirmed, we failed to replicate previous associations between this signal and adolescent CD and AAD in adulthood. Nevertheless, rs9291283 is an interesting candidate for further study as it has been associated with cocaine dependence and is in high linkage disequilibrium (LD) with a SNP in a cis-enhancer region of GABRA2.

In Chapter III, I utilized both individual SNP data and Genome-Wide Association Studies (GWAS) to test for association between GRM7 and alcohol consumption (AC). This work was proposed by a collaborator, Dr. Csaba Vadasz, and began by evaluating the evidence for association between two SNPs in GRM7 and AC using the same methods and samples as Chapter II. Although our analyses found evidence for association between rs3749380 and EAs and Hispanics in the CADD, these associations were not replicated in the GADD sample.
Similarly, rs1485175 was associated with AC in Hispanics from the GADD sample, but not the CADD sample. As GRM7 is a large gene and the two SNPs analyzed cover little of the variation in the gene, a gene-based test was run to evaluate evidence for association among all available SNPs in the gene. The test was run using four GWAS and the p-values between each study combined. Although our results showed no association between variation in GRM7 and AC, only common variants were studied and a role for rare variants or upstream regulatory elements was not ruled out.

Chapter IV focused on a manually curated set of genes selected based on their role in nicotinic acetylcholine receptor (nAChR) assembly, stability, signaling, trafficking, or function. Using gene-based tests, these analyses focused on testing for association between these genes and nicotine dependence (ND) and cigarettes per day (CPD) in GWAS with European and African Americans (EAs and AAs, respectively). As expected, the genes encoding the nAChRs were the strongest signal with these behaviors in EAs, but not AAs. Furthermore, while different genes were associated with CPD or ND for each ethnic group, genes associated with CPD in EAs tended to be associated with ND in AAs, and vice versa. These results suggest that genetic factors may drive some of the differences in smoking patterns between EAs and AAs.

Using the same collection of genes and methods utilized in Chapter IV, Chapter V focused on testing these genes for association with alcohol dependence (AD) as evidence has suggested the same genetic factors explain much of the variation in AD and ND. Although many of the genes associated with AD in EAs were different from those associated with ND or CPD in EAs, one gene, DNAJA3, was associated with both drug behaviors in EAs. This suggests there is at least one common mechanism underlying the development of alcohol and tobacco use disorders.

In a final project, targeted sequencing of genes showing preliminary association with alcohol or nicotine behaviors was performed in a small sample. Rare variants were identified in each gene and collapsed into sets to test for association. No statistically significant associations
were seen between AC and *GRM7* or *FYN* and *CAMK2A* with ND. One set in *FGFR2* was associated with AD after correction for multiple testing. Future studies could assess the functional relevance of this region. However, this study was underpowered and a replication study is necessary.

The purpose of human molecular genetic studies is to inform future pharmacogenetic studies of the role of certain variants in heritable disorders. While it is likely the implementation of this information into clinical practice will take time, these studies are critical to identify potential targets for novel treatments of these disorders. This dissertation utilized individual SNP associations, GWAS, and sequencing studies to investigate genetic risk factors for alcohol and nicotine behaviors and suggest future avenues for investigation. These studies not only highlight the importance of human molecular genetics for the characterization of future drug targets, but also demonstrate the importance of large samples and using multiple analytical approaches in order to weigh the reproducibility of each study.
REFERENCES


WHO | World Health Organization. WHO.


CDC Centers for Disease Control and Prevention. National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Department of Health and


Cowen MS, Djouma E, Lawrence AJ (2005). The metabotropic glutamate 5 receptor antagonist 3-[(2-methyl-1,3-thiazol-4-yl)ethynyl]-pyridine reduces ethanol self-administration in multiple


Li X, Gardner EL, Xi ZX (2008b). The metabotropic glutamate receptor 7 (mGluR7) allosteric agonist AMN082 modulates nucleus accumbens GABA and glutamate, but not dopamine, in rats. Neuropsychopharmacology 54(3): 542-551.


Smith CL, Mittleud P, Prescott ED, Fuhrer C, Burden SJ (2001). Src, Fyn, and Yes are not required for neuromuscular synapse formation but are necessary for stabilization of agrin-


