2012

Using Genetic Information in Risk Prediction for Alcohol Dependence

Jia Yan
Virginia Commonwealth University

USING GENETIC INFORMATION IN RISK PREDICTION FOR ALCOHOL DEPENDENCE

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

Jia Yan
Bachelor of Arts, Rutgers University, 2006

Director: Danielle M. Dick, Ph.D.
Associate Professor, Departments of Psychology, Psychiatry, and Human and Molecular Genetics

Virginia Commonwealth University
Richmond, VA
September 2012

Table of Contents

Acknowledgments

I am fortunate to have had an incredibly nurturing atmosphere in which to grow as a scientist and a genetic counselor here at Virginia Commonwealth University, thanks to the support of many individuals.

First and foremost, I would like to thank my advisor, Dr. Danielle Dick, whose zeal for research and gift for teaching has provided me with a strong framework of support and guidance throughout my graduate training. From sitting down and looking through scripts and results to talking about the conclusions, she has taught me how to think critically about multiple sources of information and to see the bigger pictures. I can't keep count of the number of times I left a discussion with her completely pumped and excited about research. I am especially thankful for the way in which she embraced dual degree training with her support of the program and the integration of clinical genetics and gene-finding research on alcohol dependence in this dissertation project.

I owe a great deal of gratitude to my committee members, Drs. Kenneth Kendler, Brion Maher, John Quillin, Todd Webb, and Timothy York. I have learned so much from each one of them from meetings, classes, and during our individual discussions and work on analyses. I am thankful for their time and thoughtful reflection on our projects and devotion to being wonderful

teachers. They have taught me so much about how to think and how to ask questions, and subsequently have helped me to grow as a researcher.

I want to thank members of my lab: Drs. Fazil Aliev, Shawn Latendresse, Alexis Edwards, Jessica Salvatore, Seung Bin Cho, Jackie Meyers (fellow graduate student and one of my best friends from the beginning through it all), and new graduate students, Jim Clifford, Megan Cooke, and Neeru Goyal. Particularly, Shawn first helped me become immersed in our research when I joined the lab and Fazil has devoted countless hours to explaining the mathematical concepts behind our work, helping me learn about our datasets and how to perform numerous statistical analyses.

I also want to thank members of the Department of Human and Molecular Genetics, the Clinical Genetics Group and my fellow genetic counseling students, and the Virginia Institute for Psychiatric and Behavioral Genetics for all of their teaching, feedback, supervision, ideas, questions, and suggestions. I especially want to thank the director of the graduate program, Dr. Rita Shiang, and the director of the genetic counseling program, Ms. Rachel Gannaway, who were instrumental in guiding and supporting me throughout the dual degree program. I also want to thank my fellow students, Jackie, Rosie, Tim, Laura, Amy, Rachel, Lori, Vern, Belal, Sami, and Ankita for all of their help during graduate school, and for their friendship.

Finally, I would like to thank my family. My parents, Wen-Bin Yan and Xiao-Juan Guan, and brother, James, have all inspired me with their pursuit of science, and have helped instill in me a love of learning early on. My fiancé, Brian, has been my steadfast source of strength over these years. I don't know where I would be without their love and support.

List of Tables

List of Figures

Abstract

USING GENETIC INFORMATION IN RISK PREDICTION FOR ALCOHOL DEPENDENCE

By Jia Yan, B.A.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Human and Molecular Genetics and Master of Science in Genetic Counseling at Virginia Commonwealth University.

Virginia Commonwealth University, 2012.

Major Director: Danielle M. Dick, PhD
Associate Professor of Psychiatry, Psychology, and Human and Molecular Genetics

Family-based and genome-wide association studies (GWAS) of alcohol dependence (AD) have reported numerous associated variants. The clinical validity of these variants for predicting AD compared to family history has not yet been reported. These studies aim to explore the aggregate impact of multiple genetic variants with small effect sizes on risk prediction in order to provide a clinical interpretation of genetic contributions to AD. Data simulations showed that given AD's prevalence and heritability, a risk prediction model incorporating all genetic contributions would have an area under the receiver operating characteristic curve (AUC) approaching 0.80, which is often a target AUC for screening. Adding additional environmental factors could increase the AUC to 0.95. Using the Collaborative Study on the Genetics of Alcoholism (COGA) and the Study of Addiction: Genes and Environment (SAGE) GWAS samples, we used several different sources to capture genetic information associated with AD in discovery samples, and then tested genetic sum scores created based on this information for predictive accuracy in validation

samples. Scores were assessed separately for single nucleotide polymorphisms (SNPs) associated in candidate gene studies and in GWAS analyses. Candidate gene sum scores did not exhibit significant predictive accuracy, but SNPs meeting less stringent *p*-value thresholds in GWAS analyses did, ranging from mean estimates of 0.549 for SNPs meeting $p<0.01$ to 0.565 for SNPs meeting $p<0.50$. Variants associated with subtypes of AD showed that there is similarly modest and significant predictive ability for an externalizing subtype. Scores created based on all individual SNP effects in aggregate across the entire genome accounted for 0.46%-0.57% of the variance in AD symptom count, and have AUCs of 0.527 to 0.559. Additional covariates and environmental factors that are correlated with AD increased the AUC to 0.865. Family history was a better classifier of case-control status than genetic sum scores, with an AUC of 0.686 in COGA and 0.614 in SAGE. This project suggests that SNPs from candidate gene studies and genome-wide association studies currently have limited clinical validity, but there is potential for enhanced predictive ability with better detection of genetic factors contributing to AD.

Chapter 1: Introduction

Background and significance

Alcohol dependence (AD) is a complex psychiatric condition that is influenced by both genetic and environmental factors (Stacey et al., 2009). It affects 4-5% of individuals at any given time in the United States and accounts for 10% of disability-adjusted life years lost (Hasin et al., 2007; Rehm et al., 2009). It results in numerous unintentional and intentional injuries and impacts other diseases such as maternal and perinatal disorders, liver cirrhosis, cancer, diabetes mellitus, cognitive impairments, and cardiovascular diseases (Rehm et al., 2009; Stavro et al., 2012). The World Health Organization estimated that harmful alcohol use results in 20-30% of liver cirrhosis, liver and esophageal cancer, epilepsy, homicide, and motor vehicle accidents worldwide (World Health Organization, 2004). The substantial contribution of AD to the global burden of disease makes efforts to identify differential susceptibility to AD an important public health need (Rehm et al., 2009). Based on twin studies, AD has an estimated heritability of around 50-60% for both men and women (Heath et al., 1997; Kendler et al., 1992; Prescott and Kendler, 1999). Gene-finding studies have reported numerous genetic loci associated with alcohol dependence. The existence of public interest in genetic counseling and genetic testing for alcohol dependence stresses the importance of coupling gene-finding studies with the evaluation of predictive accuracy and clinical

utility of genetic information for alcohol dependence (Gamm et al., 2004b; Khoury et al., 2009). This study explores the clinical validity of using information about specific genetic variants, family history, and additional factors such as marital status, religious attendance, educational attainment, and income, in risk prediction for alcohol dependence.

Alcohol dependence is defined by the DSM-IV-TR as three or more of the following symptoms over a twelve-month period: tolerance, withdrawal, excessive consumption, inability to reduce alcohol use, spending a great deal of time on obtainment of alcohol, giving up or reducing important social, occupational, or recreational activities due to alcohol use, and continued use despite adverse physical or psychological consequences (4th ed.; *DSM–IV*; American Psychiatric Association, 1994). Development of AD involves initiation of use and a process by which compulsive behavior arises following controlled drinking initiation. Addiction has been described as encompassing three stages characterized by aspects of impulsivity and compulsion: "binge/intoxication", "withdrawal/negative affect", and "preoccupation/anticipation", or craving (Koob and Volkow, 2010). AD is a prevalent disorder in the United States. According to the National Institute on Alcohol Abuse and Alcoholism (NIAAA)'s latest National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) for Alcohol Use and Alcohol Use Disorders in 2001-2002, the 12-month prevalence of alcohol dependence in the United States was 3.81%, with a prevalence rate of 5.42% in males and 2.32% in females (Grant et al., 2004). The lifetime prevalence of alcohol dependence has been reported to range from 5.4% in the National Comorbidity Survey (NCS) (Kessler et al., 2005) to 12.5% in NESARC, with alcohol use disorders twice as common in men as

they are in women (Hasin et al., 2007). A study from the National Longitudinal Alcohol Epidemiological survey showed that of the general population, approximately 40% reported having some history of alcoholism in their family and approximately 7-9% of the population reported having both first and second-degree relatives with AD (Gamm et al., 2004a; Grant, 2000) . The World Health Organization estimates that more than 200 million people in the world are affected with AD (Ginter and Simko, 2009).

Genetics of alcohol dependence

Human linkage and association studies and animal studies have shown numerous genetic variants and key pathways associated with AD and other substance use disorders, with multiple genetic contributions of small effect contributing to risk (Gelernter and Kranzler, 2009; Kalsi et al., 2009). Many studies of the genetics of alcohol dependence have revealed phenotypic and etiological complexities, finding genetic influences across a variety of alcohol phenotypes, including alcohol dependence that is co-morbid with other drug dependence and externalizing and internalizing disorders such as conduct disorder, adult antisocial personality disorder, and major depressive disorder, and intermediate phenotypes and alcohol-related traits including impulsivity, sensation-seeking, and behavioral disinhibition. Studies show unique and shared genetic etiology for these comorbid phenotypes, as well as genetic variants that contribute to intermediate phenotypes and alcohol-related traits, which could in turn influence risk for dependence (Kendler et al., 2012; Dick et al., 2008a; Kertes et al., 2011; Knopik et al., 2004; Buscemi and Turchi, 2011).

Among the first studies investigating the etiology of alcohol dependence were

family studies that reported evidence for familial transmission of AD (Merikangas, 1990; Merikangas et al., 1985; Radouco-Thomas et al., 1979). Because family members can share both genetic and environmental factors, intergenerational family studies of AD cannot distinguish clearly between the two. Twin studies, however, are genetically informative in delineating the etiological contributions of genetic and environmental factors. These studies compare the phenotypic similarities and differences between monozygotic twins, who share 100% of their genetic variation, and dizygotic twins, who share on average 50% of their genetic variation. Both types of twin pairs also have a common shared environment as well as unique environmental factors specific to each twin. These studies have found that about 50-60% of the variability in alcohol dependence is attributable to additive genetic factors, and that the rest of the variability in AD is due primarily to unique unshared environmental factors (Kendler et al., 1992; Heath et al., 1997; Prescott and Kendler, 1999).

The search for specific genetic loci contributing to alcohol dependence began with linkage studies, which investigate the co-segregation of genetic markers with a disease or trait within a family. Linkage studies assess the occurrence of co-segregation more than expected by Mendel's Law of Independent Assortment, with the assumption that segregation between a genetic marker and disease status occurs when the marker has a close physical distance to the disease locus and therefore lower likelihood of separation from the disease locus during meiotic recombination. Because the chromosomal location of genetic markers used in linkage studies are known, linkage study results help localize disease regions (White et al., 1989; Botstein et al., 1980). Studies of large, densely affected families and sibling pairs with alcohol dependence have uncovered regions

across multiple chromosomes that showed evidence for linkage. Genome-wide linkage studies in the Collaborative Study on the Genetics of Alcoholism (COGA) sample have found linkage of alcohol dependence diagnoses to chromosome 4q near the *ADH* gene cluster, and also on chromosomes, 1, 2, 3, 7, and 8 (Reich et al., 1998; Nurnberger et al., 2001; Schuckit et al., 2001; Foroud et al., 2000). A study by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) identified a region on chromosome 11 and also a region on chromosome 4q in a sample from a Southwest American Indian tribe, near the *GABRB1* gene (Long et al., 1998). The 4q region was further supported in the Irish Affected Sib Pair Study of Alcohol Dependence (IASPSAD) sample (Prescott et al., 2006).

Linkage studies uncover chromosomal regions at low resolutions, on the order of 1 centimorgan (cM) or roughly 1 megabase (Mb), and are more suited for diseases of Mendelian etiology, which have single-gene contributions of larger effect sizes. Follow-up studies fine-mapping chromosomal regions using association analyses are necessary to discover specific genes (Boehnke, 1994; Ciaranello and Ciaranello, 1991). Association analyses, on the other hand, are designed to discover specific alleles that are correlated with disease status. Association studies are based on the observation that patterns of linkage disequilibrium (LD), or correlations among alleles more than expected by chance, are maintained across the genome. These correlations exist often because of close physical distance, which reduces the likelihood that recombination events between the two loci occur across generations. A map of the LD structure of the human genome has allowed for LD-based tagging of the entire genome using common single nucleotide polymorphisms (SNPs). Haplotype blocks consisting of groups of alleles that are

correlated and inherited together can be captured by a reduced number of informative SNPs. These SNPs have been used to test for associations with disease status, with the idea that SNPs that are associated with disease status are either in LD with causal loci (indirect association), or are themselves causative (direct association) (International HapMap Consortium, 2003) .

Candidate gene studies of alcohol dependence have investigated the correlation between specific genetic variants and alcohol dependence. Candidate genes are typically selected as positional candidates, or genes that are located in or near linkage regions for alcohol-related phenotypes, and/or functional candidates, which are genes involved in specific biological pathways that are hypothesized to influence risk for addiction (Gelernter and Kranzler, 2009; Zhu and Zhao, 2007). Association methods include family-based association, which assesses for increased frequency of transmission of specific alleles from parents to affected offspring using the transmission disequilibrium test and population-based association, which tests for allele frequency differences between cases and matched-controls for a disorder using regression methods (Buscemi and Turchi, 2011). Many candidate genes have been found to be associated with AD. A number of promising candidates that have been replicated in independent samples have emerged. Some of the candidate genes and pathways that are currently thought to be involved in AD include the following:

Alcohol Metabolism

The most robustly replicated candidate genes with the largest estimated effect sizes encode enzymes that play primary roles in alcohol metabolism: alcohol and aldehyde

dehydrogenases. The breakdown of ethanol occurs primarily in the liver, the first step of which involves oxidation of ethanol to acetaldehyde, a reaction that is catalyzed by alcohol dehydrogenases (ADH). The second step is catalyzed by aldehyde dehydrogenases (ALDH), which results in oxidation of acetaldehyde to acetate. The accumulation of acetaldehyde leads to adverse physiological reactions to alcohol, such as nausea, flushing, and tachycardia (Edenberg, 2007a). Variants of the *ADH* and *ALDH* genes that confer differences in the elimination of alcohol and the accumulation of acetaldehyde – and subsequently symptoms following alcohol consumption – have been found to influence risk for alcohol dependence (Strat et al., 2008). Both coding and noncoding variations have been associated with AD, with allele frequencies varying across populations of different ancestry.

The *ALDH2*2* Glu504Lys allele results in nearly inactive ALDH2, and therefore lack of conversion of acetaldehyde to acetate, resulting in nausea, tachycardia, and in particular, a severe flushing response following alcohol intake (Edenberg, 2007a). Carriers of this allele have been found to have a significantly decreased risk for AD (heterozygotes in East Asian populations have been found to have a fivefold reduction in risk for AD) (Mathews et al., 2012; Chen et al., 1999). The frequency of *ALDH2*2* is common in East Asian populations, but rare in European and African populations (Oota et al., 2004). Variants in the *ALDH1A1* gene have also been associated with AD and drinking behavior in different populations, including American Indian, Finnish, and East Indian samples (Ehlers et al., 2004; Lind et al., 2008; Moore et al., 2007). *ADH1B*2* and *ADH1C*1,* which encode enzymes that have greater activity in the conversion of alcohol into acetaldehyde, have been shown to have protective effects against AD in East Asian

populations, perhaps due to the increased feelings of nausea resulting from toxicity of a larger quantity of acetaldehyde (Choi et al., 2005; Crabb et al., 1989; Crabb et al., 1993; Edenberg, 2007b; Li et al., 2011). *ADH1B*2* has also been found to have a protective effect among individuals of Jewish descent (Hasin et al., 2002a; Hasin et al., 2002b) and in a group of Mexican American men in the United States (Konishi et al., 2004). The variant has been associated in African American individuals, though it has a smaller frequency in the population (Whitfield, 2002). Studies of the seven *ADH* genes in European Americans have found association of *AHD1B* (Whitfield, 2002) and of *ADH4* variants with AD (Edenberg et al., 2006; Luo et al., 2005b; Edenberg, 2007a; Guindalini et al., 2005). More recently, despite the substantially lower frequency of the *ADH1B*2* allele in European American populations, *ADH1B*2* has been found also to have a protective effect on alcohol dependence in European Americans and African Americans (Bierut et al., 2012). In a German sample, a genome-wide significant finding was found for a SNP between the *ADH1B* and *ADH1C* genes that is in LD with the functional *ADH1C* Arg272Gln variant, which has been previously associated with alcohol consumption (Frank et al., 2012; Macgregor et al., 2009). Convergent evidence from linkage, candidate gene, and genome-wide association studies, coupled with knowledge of the biological function of ADHs and animal and expression studies, support the importance of *ADH* genes for AD (Ehlers et al., 2010; Chen et al., 2005).

Reward Pathways

Reward pathways have been found to be involved in alcohol initiation, tolerance, preference, consumption, abuse, and dependence (Strat et al., 2008). Genes that play a

role in neurotransmitter systems, including ones involving dopamine, gamma-aminobutyric acid (GABA), opioids, glutamate, and serotonin, have been key candidate genes for AD (Palmer et al., 2012).

GABA, the major inhibitory neurotransmitter in the central nervous system, has been implicated in risk for AD. Variants in the *GABA* inhibitory pathway have been associated with AD across multiple samples, particularly for the $GABA_A$ receptor, *GABRA2*. In the COGA high-density family sample with multiple first-degree relatives diagnosed with AD, Edenberg et al. found evidence for association of multiple SNPs in the *GABRA2* gene with alcohol dependence and increased power in the beta frequency band measured by electroencephalography, which is an endophenotype for AD (Edenberg et al., 2004). This association has been replicated in several additional studies (Covault et al., 2004; Lappalainen et al., 2005; Drgon et al., 2006; Fehr et al., 2006; Soyka et al., 2008; Enoch et al., 2006; Bierut et al., 2010). The lack of association between *GABRA2* and AD has also been reported (Matthews et al., 2007). This non-replication has been attributed to a difference in phenotype; the sample in which the negative finding was seen had minimal comorbidity with other drug dependence and psychiatric phenotypes (Matthews et al., 2007). In fact, further study showed that *GABRA2* was associated with AD that is comorbid with other drug dependence (Agrawal et al., 2006). *GABRG3, GABRA1, GABRA6,* and *GABRB1* are several additional GABA receptor genes reported to be associated with AD (Dick et al., 2004; Dick et al., 2006b; Noble et al., 1998; Song et al., 2003).

Genes encoding dopamine receptors have been associated with AD, including the dopamine D2 receptor, *DRD2* and the dopamine D4 receptor gene, *DRD4*. An increased

frequency of the *DRD2* A1 allele of the Taq1A restriction fragment length polymorphism has been associated with AD. There have been a number of studies investigating this association, with mixed results. The first reports of association showed that the Taq1A1 restriction fragment length polymorphism was associated with AD in a sample of postmortem brain tissue in severe alcoholics and controls (Blum et al., 1990; Blum et al., 1991). Since then, there have been a number of studies that replicated this finding in independent associations with AD (Comings et al., 1994; Blum et al., 1990; Blum et al., 1991; Noble et al., 1991; Amadeo et al., 1993; Amadeo et al., 2000; Foley et al., 2004; Hietala et al., 1997; Ishiguro et al., 1998; Konishi et al., 2004; Kono et al., 1997). However, there have also been many failures to replicate (Arinami et al., 1993; Bolos et al., 1990; Cook et al., 1992; Chen et al., 2001; Cruz et al., 1995; Edenberg et al., 1998; Gelernter and Kranzler, 1999; Gelernter et al., 1991; Goldman et al., 1992; Lee et al., 1999; Lobos and Todd, 1998; Lu et al., 1996; Sander et al., 1995; Sander et al., 1999; Suarez et al., 1994; Turner et al., 1992). The discordance across studies has been attributed to differences in phenotypic severity, co-occurrence of other phenotypes such as polysubstance abuse and impulsive and compulsive behaviors - coined "reward deficiency syndrome" by Blum et al. (Blum et al., 1996), and population stratification (Dick et al., 2007d). A later study found that the Taq1A polymorphism that had been thought to be located in the *DRD2* gene was actually located 10 kb downstream in the *ANKK1* gene (Neville et al., 2004). A comprehensive study of SNPs across both *DRD2* and *ANKK1* found associations for SNPs in both genes, with stronger evidence for SNPs in the 5' region of *ANKK1,* particularly for AD with medical complications (Dick et al., 2007d).

Several genes related to serotonin (5-hydroxytryptamine, 5-HT) have been suggested to play a role in AD. Specifically, a functional insertion-deletion variant in the serotonin transporter protein (5-HTT)-linked promoter region (5-HTTLPR) affects regulation of 5-HT levels and has been associated with AD; however, results have been controversial, with a large number of both positive and negative findings (reviewed in Dick and Foroud, 2003). A meta-analysis of 17 studies, comprised of 3,489 alcoholics and 2,325 controls, showed that the short (S) allele was associated with AD, with an odds ratio of 1.18 (Feinn et al., 2005). A gain-of-function 5-HT3 receptor gene (*HTR3B*) variant has also been associated with alcohol dependence in a treatment-seeking sample of individuals of African descent (Enoch et al., 2011).

The muscarinic acetylcholine receptor M2 (*CHRM2*) has been associated with AD and the endophenotype event-related oscillations (EROs) (Wang et al., 2004). Evidence for this association has been seen in other independent samples (Luo et al., 2005a; Kendler et al., 2011). Further study has suggested that *CHRM2* is particularly important in AD that is comorbid with other drug dependence (Dick et al., 2007a) and associated with the severity of alcohol dependence (Jung et al., 2011). It has also been implicated in risk during adolescence, showing an association with adolescent substance use and behavioral disinhibition (Hendershot et al., 2011) and interaction with parental monitoring in risk for externalizing (Dick et al., 2011). Variants in the gene have also been associated with nicotine dependence (Mobascher et al., 2010), major depressive disorder (Wang et al., 2004) and IQ (Dick et al., 2007c).

Genome-wide association studies of alcohol dependence

More recently, a number of genome-wide association studies (GWAS) have been performed for alcohol dependence and alcohol-related phenotypes. GWA-studies, which assess common markers across the genome for association with a common disorder, are an a priori approach to gene finding (Visscher et al., 2012). Compared with a linkage study, which is more suited to detecting loci with larger effects sizes, an association study is a more powerful method that requires fewer markers and a smaller sample size to detect common variants of small effect for diseases. For example, in a nonparametric linkage analysis of affected sib pairs, an allele with moderate frequency (0.10-0.50) and modest genotypic relative risk (GRR) of 1.5 would have a probability of allele sharing between siblings of only 50.5% - 51%, which is close to the null hypothesis of 0.50. In an association study, however, the degree of overtransmission from heterozygous parents to affected offspring for an allele of this effect would be around 60%. The number of families required to detect an allele with a GRR of 1.5 effect using linkage analysis would be on the order of 17,000-67,000, compared with only 949-2218 for an association study to detect the same effect (Risch and Merikangas, 1996).

Genome-wide association studies for AD have supported previous candidate gene studies, as well as reported many new genes and pathways in risk for alcohol-related phenotypes (Treutlein and Rietschel, 2011b; Treutlein et al., 2009; Frank et al., 2012; Bierut et al., 2010; Edenberg et al., 2010; Agrawal et al., 2011; Heath et al., 2011; Lind et al., 2010; Schumann et al., 2011; Wang et al., 2012; Wang et al., 2011; Zuo et al., 2012; Kendler et al., 2011; Zuo et al., 2011). Although many of the reported genome-wide association studies to date have reported variants that did not meet the genome-wide significance threshold of $p < 5x10^{-8}$, many have reported variants that were associated

with low *p*-values ($p < 1x10^{-5}$). Additional details about specific AD GWAS are summarized in Chapter 4. Briefly, several results from AD GWAS reports include the following: two correlated SNPs in the 3' flanking region of the peroxisomal trans-2-enoyl-CoA reductase gene (*PECR)* (Treutlein et al., 2009); a group of chromosome 11 genes (*SLC22A18, PHLDA2, NAP1L4, SNORA54, CARS*, and *OSBPL5*) (Edenberg et al., 2010); the semaphorin 3E gene (*SEMA3E)* (Lind et al., 2010)*; MARK1,* which is involved in phosphorylation of microtubule-associated proteins (Lind et al., 2010)*; DDX6,* which encodes a putative RNA helicase, and *KIAA1409*, which is thought to be part of a sodium channel complex (Lind et al., 2010); The *KIAA0040* gene was associated with AD in both Zuo et al.'s study (2012) and Wang et al.'s meta-analysis (2011); *THSD7B*, *NRD1,* and *PKNOX2* in Wang et al. (2011). Studies of quantitative traits such as alcohol consumption have identified a genome-wide significant association with the *AUTS2* gene (Schumann et al., 2011) and evidence of association for the *TMEM108* and *ANKS1A* genes (Heath et al., 2011). In a study of an alcohol factor score, Kendler et al. (2011) found the most significant SNP to be *KCNMA1, AKAP9,* and *PIGG* in the EA sample and CEACAM6, KCNQ5, SLC35B4, and MGLL in the AA sample, and found support for previously associated candidate genes for *ADH1C, NFKB1,* and *ANKK1* in the EA sample and *ADH5, POMC*, and *CHRM2* in the AA sample (Kendler et al., 2011).

Environmental factors influencing risk for alcohol dependence

Additional factors that contribute to AD have been implicated in numerous studies. A study investigating risk factors predicting problem drinking in a sample of 30-year old Danish men identified low birth weight, number of life crises in childhood, ratings of

childhood unhappiness and antisocial personality disorder as powerful independent predictors of problem drinking, accounting for 46% of the variance in problem drinking (Knop et al., 2003). Childhood maltreatment has also been implicated as an environmental risk for substance use disorders such as AD implicated in numerous studies, including physical abuse, neglect, and a particularly specific risk for substance use disorders in cases of child sexual abuse (Clark and Winters, 2002; Dinwiddie et al., 2000; Kendler et al., 2000; McLaughlin et al., 2010; Nelson et al., 2002; Sartor et al., 2007). Religiosity has also been shown to have a protective main effect on risk for substance use disorders (Kendler et al., 2003a; Koopmans et al., 1999) . In independent samples, educational attainment has been found to be associated with AD (Grant et al., 2012). Socioregional residence has been shown to influence both religiosity and alcohol use (Dick et al., 2001). Marital status has also been shown to be associated with AD (Dick et al., 2006a). In data from the National Longitudinal Alcohol Epidemiology Study and the National Epidemiologic Study on Alcohol and Related Conditions, marital status and educational attainment were associated with alcohol dependence and income was associated with alcohol abuse (Caetano et al., 2011). Hicks et. al. assessed six environmental risk factors – academic achievement and engagement, antisocial and pro-social peer affiliations, mother-child and father-child relationship problems, and stressful life events – and found that each risk factor had a significant correlation with externalizing disorders in adolescence such as substance use disorders and antisocial behavior (Hicks et al., 2009).

These studies stress the important role that specific environmental variables play in the risk for alcohol use behaviors and AD. In addition to having a main effect on AD

risk, many of these "environmental" factors, have a moderating effect on, and a correlation with, genetic risk factors for AD (Dick et al., 2001; Hicks et al., 2009; Caetano et al., 2011; Dick et al., 2006a; Koopmans et al., 1999). Ultimately, a combination of multiple genetic and environmental factors should be used to predict and treat disease.

Psychiatric genetic counseling and testing

The field of genetic counseling for single gene, Mendelian disorders of high penetrance utilizes a large range of testing options that often have high clinical validity. In contrast, genetic counseling for phenotypes of complex etiology, including alcohol dependence, schizophrenia, and the majority of cancer and autism cases, is limited to a general discussion of genetic and environmental contributions to pathogenesis and the use of empiric risk estimates rather than direct genetic testing (Harper, 2004). Psychiatric genetic counseling is different from genetic counseling for single gene conditions of high penetrance in regard to both the degree of uncertainty and the availability of testing. Despite these differences, the principles of genetic counseling for each are the same. As defined by the National Society of Genetic Counselors (National Society of Genetic Counselors' Definition Task Force et al., 2006):

> Genetic counseling is the process of helping people understand and adapt to the medical, psychological and familial implications of genetic contributions to disease. This process integrates:
> - Interpretation of family and medical histories to assess the chance of disease occurrence or recurrence.
> - Education about inheritance, testing, management, prevention, resources and research.

- Counseling to promote informed choices and adaptation to the risk or condition.

Individuals seek genetic counseling for psychiatric disorders due to a variety of reasons, including finding out a cause for the disorder and obtaining risk assessments for the recurrence of a disorder. Affected individuals may be concerned with the risk of passing the disorder to their children. Individuals and their families could be struggling to understand the etiology of the condition. Family members of an individual with a psychiatric condition may be concerned with their own risks for developing the condition, as well as the risks for their children. They may face psychosocial issues specific to having a family member with a disorder, such as the "survivor guilt" sometimes experienced by siblings and other family members without symptoms. Parents of affected individuals may also harbor guilt related to the belief that they played a part in causing the illness in their children. Families may face stigma in society. Having a better understanding of the psychiatric condition affecting their family could help them better develop coping strategies and form behaviorally adaptive practices (Austin and Honer, 2007).

Genetic counseling seeks to address these issues. Risk assessment during a genetic counseling session involves gathering a targeted family history to trace psychiatric features, along with distinct physical and cognitive features associated with some psychiatric conditions, through a three-generation pedigree (Peay et al., 2008). The nature of uncertainty in psychiatric phenotypes is addressed through education about the environmental and genetic contributions to psychiatric conditions. Currently, population-based empiric risk estimates based on family history and degree of relatedness to an

affected individual, are quoted during counseling about recurrence risks (Harper, 1998). Table 1.1 and 1.2 below illustrate the empiric risk to relatives of persons with specific psychiatric disorders.

Table 1.1 Empiric risk for common psychiatric disorders in first-degree relatives

Psychiatric Disorder	General Population	First-degree relative
Schizophrenia	1%	5-16%
Bipolar Disorder	1-5%	4-18% (BPD) 9-25% (UPD)
Major Depression	5-35% (females) 5-15% (males)	10-25%
Obsessive Compulsive Disorder	1-3%	10%
Panic Disorder	2-6%	8-31%

Adapted from Hill and Sahhar, 2006
BPD = bipolar depression; UPD = unipolar depression

Table 1.2 Empiric risk for schizophrenia in relatives of a person with schizophrenia

Relationship to person with schizophrenia	Lifetime risk
General population	1%
First-degree relative	
Identical twin	40-48%
Fraternal twin	10-17%
Sibling	9%
Parent	6-13%
Offspring	13%
Second-degree relative	
Aunt/uncle	2%
Niece/nephew	4%
Granchild	5%
Third-degree relative	
First cousin	2%

Adapted from (Finn and Smoller, 2006)

Traditional empiric risk for AD for family members of an individual affected with AD

from several studies are reviewed in Table 1.3 (from (Merikangas, 1990).

Table 1.3 Familial empiric risk for AD

Relationship to person with AD	Recurrence risk
Sibling	11%
Sister	1-8%
Brother	11.8-12.4%
Parent	29.8%
Mother	1.6-6%
Father	16.1-22%
Parents and siblings as a group	35.6%
Grandfathers	11%

There exist limitations to risk assessment that focuses primarily on empiric risk estimates derived from family history information obtained in family-based population studies. As is the case with any risk estimate derived from a population sample, empiric risk may not be applicable for a specific individual due to differences in both genetic and environmental background, particularly since empiric risk can vary widely across multiple studies. Furthermore, empiric risk may not be available for families with multiple psychiatric phenotypes or across all family relationships (Austin and Peay, 2006). Genetic information that has better-characterized risk estimates and is more specific to the individual may provide more accurate recurrence risk assessments than family history alone.

Attitudes toward genetic counseling and testing for psychiatric disorders in general

A number of studies investigating attitudes towards genetic counseling in individuals

who are affected and their families have shown both a desire for genetic counseling and an interest in genetic testing (Peay and Sheidley, 2008; Austin and Honer, 2007; Hill and Sahhar, 2006). A survey of 31 individuals with bipolar disorder showed that more than 75% of them wanted genetic counseling. More than 70% of 48 family members of individuals with schizophrenia wanted genetic counseling (Austin and Honer, 2007). Several studies assessing interest in genetic testing in independent samples of individuals affected with a range of psychiatric disorders and their family members show upwards of more than 80% of individuals possessing a desire to test for genes implicated in psychiatric disorders. In a survey of 48 members of families with multiple affected members, 83% wanted genetic testing for genes of small effect. In regard to prenatal testing for psychiatric conditions, of 65 members of the Alliance for the Mentally Ill, 77% believed that it should be available for bipolar disorder, 85% for schizophrenia and autism, 70% for attention deficit disorder, and 55% for panic disorder. Even if there is an absence of childhood preventative treatment for a disorder, 68% of the 48 members of the bipolar support group sample, which included families and friends of affected individuals, endorsed testing for children (reviewed in Smoller et al, pp 30-31, 2008). Further investigation into whether or not individuals were interested in genetic testing for major depressive disorder revealed that individuals were more likely to be interested in testing if they had a personal history of mental illness, a greater than average self-estimated risk for depression, had perceived benefits for genetic testing, and – unexpectedly, believed that evidence for a genetic component for mental illness would increase social stigma (Wilde et al., 2011). Potential for discrimination and interference with privacy decreased interest in a genetic test for major depressive disorder (Wilde et al., 2010).

Attitudes towards genetic testing for alcohol dependence

Survey studies specific to interest in risk assessment for alcohol dependence suggest that there may be considerable interest in genetic counseling and potential genetic testing to determine personal risk for alcohol dependence (Gamm et al., 2004b). Of the general population, 60-77% has been reported to share a belief that AD is "a lot" or to "some" extent due to genetic effects (Gamm et al., 2004b). In one study of 27 individuals with at least one first degree relative with alcohol dependence and an average of three additional second and third degree relatives with alcohol dependence, 63% said that they would choose to undergo a genetic test to determine their own risk for alcohol dependence if a genetic test for alcohol dependence were available. Of interested individuals, 59% believed that testing would lead to better prevention or treatment and 48% believed that it would help address their concerns about their own children's risk (Gamm et al., 2004b). This research on testing attitudes for AD and other psychiatric disorders reveal a substantial population that wants to know genetic information. This research reveals a need for the careful evaluation of the clinical utility of genetic information and subsequent education about genetic testing.

Research on genetic testing acquisition has shown one caveat about predicting testing uptake before a test becomes available: although there appears to be an interest in testing for psychiatric conditions, actual decisions to have testing may not be as high as predicted once testing does become available. In the case of Huntington disease, a smaller proportion of at-risk individuals pursued testing after the gene was found and testing became available than the proportion that had been estimated to be interested in

testing from survey studies performed before the availability of testing (Evers-Kiebooms and Decruyenaere, 1998). Evers-Kiebooms and Decruyenaere showed in 1989 in a large survey in Belgium that 66% of individuals at risk for Huntington disease and 74% of their partners had indicated that they wanted to make use of testing (Evers-Kiebooms et al., 1989). An assessment of testing published in 1998 about a decade after genetic testing for Huntington disease became available showed that testing uptake was actually 6-20% in at-risk individuals across different populations (Evers-Kiebooms and Decruyenaere, 1998).

How accurate risk prediction could influence management and prevention

If clinically valid variants were established for AD, the next step would be to assess the clinical utility of genetic testing for AD. Genetic heterogeneity and genetic testing based on information from genetic association studies can create increased uncertainty and confusion if the usefulness of testing an individual for susceptibility genes of small effect is not addressed. When determining whether individuals would benefit from genetic testing for a psychiatric disorder, a primary question to ask is whether knowledge of particular genetic risk factors changes management in a meaningful way. The hope is that risk prediction will help tailor individual treatment for disorders in a number of ways.

Currently, a proportion of pharmacological treatment for psychiatric disorders consists of trial and error. Pharmacogenetics utilizing panels of variants that predict treatment metabolism and response may reduce delay in treatment and remove additional toxicity from medication due to inappropriate dosage and drug type (Smoller et al., 2008). Heterogeneity in causes within different individuals with alcohol dependence may

prompt differences in treatment. Several genetic variants have been shown to have potential pharmacogenetic utility for individualized therapy for AD. For example, naltrexone, an opioid receptor antagonist that is used as pharmacological treatment for alcohol dependence targeted towards mitigating the rewarding response to alcohol, has been shown to be differentially effective based on *OPRM1* genotype. *OPRM1 77G* carriers have been suggested to have increased mesolimbic dopamine activity in response to alcohol and subsequently a greater treatment response to naltrexone (Heilig et al., 2011).

Additionally, Wray et al. asserted that environmental interventions might have the greatest impact on risk reduction in those who have the greatest risk to begin with (Wray et al., 2008). Therefore, identification of individuals at the highest level of genetic risk for targeted intervention may be an effective risk-reducing strategy for a disorder for which interventions are available. Knowing risk may help categorize individuals into groups of clinical significance for targeted treatment. In the case of AD, previous research has suggested that risk variants for AD may confer additional risks for trajectories of externalizing behavior across development (Dick et al., 2009). *GABRA2* has been associated with externalizing trajectories and was shown to interact with parental monitoring (Dick et al., 2009). Early prediction of AD may therefore also lead to the prediction and interventions for additional categories of risk and phenotypes across the lifespan.

Knowledge about genetic information for a disorder has been suggested to have a potentially unique impact on an individual's physical and emotional response to behavioral recommendations, known as the adherence response (McBride et al., 2012).

Testing that provides individualized knowledge of genetic susceptibility may increase motivation to make behavioral changes compared with not having information on genetic testing. While individuals who seek out genetic tests have been shown to be knowledgeable about and motivated to improve health-promoting behaviors, less is known about precisely how best to customize interventions based on genetic information for individuals across a spectrum of interest in genetic testing (McBride et al., 2012). Studies on smoking cessation rates show limited change in cessation after individuals receive knowledge about genetic testing information, with either no change in smoking rates, or decreases in smoking only during a finite period directly following education about genetic testing information. These earlier studies focused on single-gene variants that may confer increased susceptibility for lung cancer. More recent preliminary studies have shown that individuals who received information on more risk variants resulted in a greater likelihood of quitting smoking than individuals who received feedback about fewer risk variants (McBride et al., 2010). For alcohol dependence, a clinical scenario involving testing for multiple variants would be more likely than testing only for a few single gene variants, as AD is a complex trait with multiple genetic influences. The question of whether or not testing for multiple genetic variants that increase susceptibility to AD would affect behavioral outcomes would need further evaluation.

Ethical, legal, and social implications

Genetic information is unique in that it may have implications for the health of not only the individual being tested, but also that of the individual's family and future progeny. Issues of autonomy may come into play when one person's decision to test or not to test

affects a group of individuals. For the proband, the desire to maintain personal privacy in regard to medical information may conflict with a duty to warn family members about information that may affect their health. The right *not* to know genetic information may be violated in relatives of an individual who decides to have testing.

Genetic information specific to multi-allelic disorders of complex etiology may have less predictive implications for family members than genetic information for higher penetrance Mendelian disorders with single causative alleles. However, intuitive knowledge of the familial nature of a complex disorder may still exert a blanket of influence on perceptions and decision-making in relatives based on genetic testing in one proband. Ultimately, decisions made using genetic information by individuals may influence the greater community and population (Smoller, 2008). Individuals who elect genetic testing could face the possibility of being labeled and treated with stigma. Unaffected individuals with an increased risk resulting from a genetic test may experience psychological distress as well as experience discrimination from society, insurance companies, and employers, including "anticipatory stigma", which means discrimination to the same extent as affected individuals of unaffected individuals who are perceived to be at risk for a disorder (Austin and Honer, 2007). Thus, the ethical issues that accompany genetic testing for psychiatric and other complex disorders must be addressed alongside research advances.

There has been a recent emergence of direct to consumer (DTC) personal genomics testing for many multifactorial disorders, including addiction, despite limited information about the clinical validity and utility of genetic variants associated with these disorders (Mathews et al., 2012). General public perceptions of the clinical utility of direct-to-

consumer genomic profiling have been shown to be more optimistic than those of genetics professionals (Leighton et al., 2012; McGuire et al., 2009; Wilde et al., 2011). Public interest in genetic testing may be due in part to a misunderstanding of how predictive genetics can be for complex disorders (Lawrence and Appelbaum, 2011). The potential harm of inaccurate information for consumers emphasizes the need to couple gene-finding efforts with rigorous evaluation of predictive accuracy, and subsequent education for the general public about genomic testing (Khoury et al., 2009).

Genetic risk prediction studies

There exists debate about whether aggregate profiles of associated markers could be used to predict risk for complex diseases (Janssens et al., 2006; Lee et al., 2008; Evans et al., 2009). Previous efforts to study risk prediction for complex disorders have assessed the predictive ability of genetic sum scores based on number of risk alleles that have been associated with a particular disorder.

Previous studies have investigated the potential for risk prediction for a number of common complex disorders. A clinical test is evaluated within the A.C.C.E framework, based on the test's *a*nalytical validity, *c*linical validity, *c*linical utility, and *e*thical, legal, and social implications surrounding testing (Khoury et al., 2009). Analytical validity is the reliability and accuracy of the test measure itself, such as the accuracy of genotyping calls on a SNP array. Clinical validity is the degree to which the test can explain and predict risk for a disorder. Measures such as sensitivity, or the probability of a positive test among individuals with a disorder, and specificity, or the probability of a negative test among individuals without the disorder, are indicators of a test's clinical validity.

Clinical utility is the benefits and limitations of the test in changing management of a disorder for an individual. Clinical utility can encompasses changes in screening procedures, treatment, and preventative behavioral or pharmaceutical measures. It can also represent personal utility, in which knowledge about the test results alone can make a difference in an individual's perspective in a beneficial way (Foster et al., 2009). For example, an individual who may be suffering from self-blame for having a psychiatric disorder may benefit from knowledge that genetics played a role in the disorder's etiology. This individual may be better able to cope with the disorder, even if there are no direct changes in treatment or prevention based on the genetic test results alone (Khoury et al., 2009; Foster et al., 2009).

The ability of a clinical test to distinguish between individuals with and without a disease is typically assessed based on the test's sensitivity and specificity. A frequent measure of clinical validity is the receiver operating characteristic (ROC) curve (Figure 1.1), which plots the sensitivity vs. 1-specificity for every cut-off of a continuous predictor to distinguish between presence and absence of a disease diagnosis (Spitalnic, 2004). The area under the ROC curve (AUC) for a continuous predictor corresponds to the probability that an individual with the disease would have a higher measured or predicted risk than an individual without the disease, and therefore reflects the proportion of individuals classified correctly as cases or controls based on the predictor. This measure of concordance is also known as the c statistic. An AUC of 0.50 means that the predictor can accurately classify 50% of individuals, or no greater than chance, whereas an AUC of 1.0 means that the predictor can correctly classify 100% of individuals,

corresponding to perfect discriminative ability. An AUC of 0.80 is typically used as a target cut-off for screening and 0.99 for diagnosis (Janssens et al., 2006).

The AUC is a measure of the ability to discriminate between a case and a control, as opposed to the positive predictive value (PPV), which is a direct measure of whether a person with a positive test result will develop a disease (Cook, 2007). A predictor that has perfect discriminatory accuracy will have a single cut-off on the upper leftmost point on the ROC curve (Figure 1.1) that corresponds to 100% sensitivity, or 100% true positive rate, and 100% specificity, or 100% true negative rate (Attia et al., 2009).

Risk Score	Sensitivity (True Positive Rate)	Specificity	1-Specificity (False Postive Rate)
0	100%	0%	100%
1	96%	56%	44%
2	88%	84%	16%
3	68%	94%	6%
4	40%	98%	2%
5	0%	100%	0%

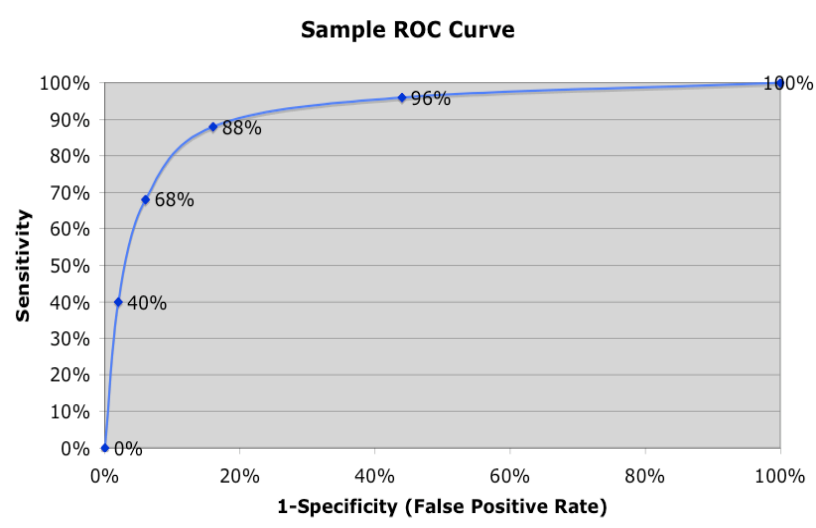


Figure 1.1 Sample ROC curve for a hypothetical continuous predictor
Percentages on the curve represent the sensitivity and 1-specificity corresponding to every risk score cut-off.

Using simulated data of 100,000 individuals with an incidence of 10% for coronary heart disease (CHD), van der Net et al. (2009) calculated the area under the receiver operating characteristic curve (AUC) to determine the ability of genetic risk profiles to discriminate between individuals who will and will not develop CHD. Using ten identified variants with odds ratios that varied from 1.13 to 1.42, the AUC was 0.59 (van der Net et al., 2009). A study by Jakobsdottir et al. showed that a model of 12 SNPs for type 2 diabetes had an AUC of 0.64, a model of 2 SNPs for prostate cancer had an AUC of 0.56, and a model of 5 SNPs for Crohn's disease had an AUC of 0.66 (Jakobsdottir et al., 2009).

In addition to assessing the ability of using only genetic markers to predict disease, the predictive value that a group of genetic markers adds to current clinical predictors of disease has been evaluated (Pencina et al., 2008; Greenland, 2008). Gail (2008) showed that adding seven SNPs identified from candidate gene studies and genome-wide association studies with per allele odds ratios (ORs) ranging from 1.07 to 1.20 to the clinically used National Cancer Institute's Breast Cancer Risk Assessment (BCRAT) tool, which takes into account family history, personal history of breast biopsies, and age at menarche and first live born for the prediction of breast cancer, only improved the AUC from 0.604 to 0.632. This increase was less than that of adding mammographic density to the BCRAT (Gail, 2008). In the case of Alzheimer disease, the addition of the *APOE* genotype to existing clinical criteria produced only a nominal increase that was not statistically significant (Attia et al., 2009). Talmud et al. (2010) showed that sum scores of risk alleles for 20 SNPs associated with type II diabetes did not appear to add to the

phenotype-based risk models, Cambridge risk score and Framingham offspring risk score, in discrimination for type II diabetes.

One reason that genetic variants may not add significantly to risk models that take into account other, more easily measured, risk factors is that they may also contribute to the very risk factors that are part of the original prediction model (Janssens and van Duijn, 2008). Therefore, adding the variants does not add additional information to the model. For example, in the Whitehall II prospective cohort study on type II diabetes (Talmud et al., 2010), a risk model containing both genetic and clinical predictors was assessed, but this risk score could be confounded by the inclusion of family history along with genetic sum scores in the models. In light of this, the study showed whether genetic variants could add additional information despite possible correlation with family history information (Talmud et al., 2010). Another reason is that the ROC may be different for populations of patients with different genetic and environmental backgrounds, in which case the sensitivity and specificity of a test may not be the same within each background profile. Finally, a prediction model that consists primarily of genetic variants has a maximum AUC constrained by the heritability of the trait as well as the disease prevalence in a population (Wray et al., 2009). This stresses the importance of taking into account other medical and environmental predictors when assessing the utility of adding genetic components of disease to risk prediction.

Prior analysis comparing the use of odds ratios to the AUC shows that an OR of 16 is needed for an AUC of 0.84 (Pepe et al., 2004). Many predictors of multifactorial common diseases do not have ORs of this magnitude. A combination of a large number of multiple predictors, however, may have greater discriminative power. For example, the

Framingham score, an established and validated discriminator of risk of cardiovascular disease that discriminates disease status with an AUC of 0.80, includes a number of risk factors that have ORs of less than 2.2, much lower than that of the overall score itself; none of the factors have enough discriminatory ability individually (Pepe et al., 2004). The ORs of associated SNPs for a complex disease such as alcohol dependence in no way come close to an OR of 16; SNPs associated with AD often have ORs less than 1.30. A panel of SNPs taken together, on the other hand, despite individually small ORs for each SNP, may have better discriminatory ability.

Project rationale and design

This project examined genetic and environmental variables to create risk profiles for alcohol dependence, in an effort to provide a clinical interpretation of current research on alcohol dependence in the context of risk prediction. The project first studied the potential predictive power for AD, and prediction of other complex multifactorial psychiatric disorders such as major depressive disorder and schizophrenia, by examining the effect of specific characteristics of disease and genetic predictors on clinical discrimination using simulated data. Existent data was then used to test how different methods of capturing genetic information could be used to predict clinical status for alcohol dependence. We assessed several different sources of predictive information encompassing genetic and other clinical predictors: data from previously associated candidate genes from the literature, family history information, data from genome-wide association studies (GWAS), and data from additional environmental and clinical variables in addition to genetic information.

Information about genetic variants contributing to alcohol dependence came from the Collaborative Study on the Genetics of Alcoholism (COGA), which is a National Institute on Alcohol Abuse and Alcoholism (NIAAA) sponsored project aimed at identifying genes involved in alcohol dependence, with 10 collaborative sites across the United States. Both a high-density family-based association sample and a case-control genome-wide association study sample have been ascertained. COGA has previously reported positive family-based association results for alcohol dependence with 114 SNPs in 21 genes using the high-density family sample. Many of these genes have also been associated with alcohol dependence in other studies. We created genetic sum scores based on risk alleles of associated SNPs in these genes. We then compared the sum score with family history in its ability to discriminate between cases and controls for alcohol dependence in a subset of the COGA GWAS sample that is independent of the gene-finding family sample and in a subset of independent individuals in the Study of Addiction: Genes and Environment (SAGE) GWAS sample, which is a separate case-control study for alcohol dependence that also contains individuals with cocaine and nicotine dependence. Next, the impact of GWAS results was assessed using both the COGA and SAGE GWAS samples. The effects of SNPs across the genome were explored by creating genetic sum scores based on subsets of SNPs meeting varying *p*-values and creating genetic sum scores consisting of the individual effects of all genotyped markers across the genome. Finally, we assessed risk prediction using different gene-finding designs based on phenotypes targeted towards reducing the heterogeneity of a binary alcohol dependence phenotype and increasing study power to detect small effects by studying alcohol dependence symptom count and alcohol

dependence subtypes. Together, these studies aimed to combine genetic and environmental variables associated with alcohol dependence in order to evaluate how both could lead to better risk prediction.

Chapter 2: Assessment of predictive ability of genetic information for common psychiatric disorders in simulated data

Abstract

Simulation studies were conducted to determine the maximum discriminatory accuracy of genetic information for models of three psychiatric disorders using receiver operating characteristic (ROC) curve analysis. The models broadly reflected the heritabilities and lifetime prevalences for major depressive disorder (MDD), alcohol dependence (AD), and schizophrenia (SCZ). We found that the highest areas under the ROC curve (AUCs) were obtained from polygenic scores created based on results from a gene-finding discovery sample of 10,000 cases and 10,000 controls, for all three disorders. For a model based on schizophrenia, with a heritability of 80% and prevalence of 1%, the AUC just passed 0.90. For major depressive disorder, with a heritability of 30% and prevalence of 13%, the AUC approached 0.80. If all genetic contributions are included in a prediction model for AD, given AD's heritability of around 50%, AUCs approached 80%. Adding environmental risk effects increased the maximum AUC to 0.95 (Maher *et al.*, in preparation).

Introduction and background

The question of whether or not genetic information can be used to predict risk for complex disorders has been addressed in multiple studies using both existent data and data simulations (Janssens et al., 2006; Purcell et al., 2009; Wray et al., 2007; Wray et al., 2010; Evans et al., 2009). For multifactorial disorders of complex etiology, a single genetic variant is likely to have a small effect on the phenotype, and therefore would likely not have significant predictive accuracy. We know that because the heritability of complex disorders such as alcohol dependence is not 100%, the predictive accuracy of genetic information alone for alcohol dependence would not be 100%. Many studies of risk prediction using genetic information have used a "genomic profiling" approach of combining multiple SNPs that have been associated with the disorder in question into polygenic scores based on the number of risk alleles carried by an individual (Manolio, 2010; Khoury et al., 2004; Janssens and van Duijn, 2009). The purpose of this study was to assess the potential maximum discriminatory accuracy, as measured by the area under the receiver operating characteristic curve (AUC), for alcohol dependence (AD), major depressive disorder (MDD), and schizophrenia (SCZ) using information from known true loci through a genomic profiling approach.

Janssens et al. (2006) used data simulations to determine the potential AUCs that could be reached for complex disorders based on multiple genetic variants. They assessed the impact of number of genes involved, risk allele frequency, disease prevalence, heritability, and odds ratios of risk genotypes. They found that high AUCs could be reached for several different models. For a group of variants that explained 30% of the phenotypic variance, the maximum AUC was 0.83 for a disease with 30% prevalence and

0.97 for a disease with 1% prevalence. Wray et al. (2007) commented that the study by Janssens et al. did not incorporate error into an individual's true genetic risk, but instead calculated the correlation between genetic risk and disease status to be equal to the square root of the broad-sense heritability on the observed scale. Wray et al. presented a different model for the predictive ability of genome-wide scores based on disease heritability of 0.10-0.20 and prevalence of 0.05 and 0.10 using simulated data. They created for these parameters models that differed by the mean and maximum relative risk (RR) and maximum proportion of genetic variance explained by one locus, and calculated the expected number of loci contributing to the complex phenotype using this information. The number of loci is proportional to RR, heritability, and prevalence of a disorder. They found that the predictive accuracy of genetic risk was highest when 10,000 cases and controls were used for a model with heritability of 10% and prevalence of 5% caused by 100 loci with RR of 1.15, the accuracy of prediction was 0.97 when calculating the correlation between logarithms of the true and predicted probability of the disease based on the following:

$$\frac{P(D \mid G)P(G)}{P(D)} = P(G \mid D)$$, where D = disease, G = genotype

→ Posterior probability of disease, given genotype:

$$P(D \mid G) = \frac{P(G \mid D)P(D)}{P(G)}$$

These simulation studies show that in aggregate, information from SNPs may account for more of the variability in disease, and therefore be more predictive of diseases in independent samples. We do not know whether a disease model specific to the epidemiological model of psychiatric disorders such as alcohol dependence, major

depressive disorder, and schizophrenia has the potential for predictive accuracy. Accordingly, we examined the effect of specific characteristics of disease on predictive power for alcohol dependence, schizophrenia, and major depressive disorder using simulated data. Accordingly, data simulations were implemented to mirror polygenic etiological models. Polygenic scores were simulated for each sample based on specific disease attributes, including heritability, prevalence, allele frequency, genotypic relative risk (GRR), *p*-value threshold used to select associated SNPs in discovery samples, genetic correlation between discovery and validation samples, and the sample sizes of the discovery and target samples. We addressed whether it is possible to achieve an AUC that is generally accepted as a screening threshold, 0.80 (Janssens et al., 2006), using genetic information, and then assessed the maximum AUC for a model for alcohol dependence with the addition of an environmental effect.

One environmental effect that we modeled that has been shown to increase risk for alcohol dependence was child sexual abuse. According to the Centers for Disease Control and Prevention (CDC) Adverse Childhood Experiences (ACE) Study of 9,367 females and 7,970 males, 20.7% of participants experienced child sexual abuse (http://www.cdc.gov/ace/index.htm). Child sexual abuse has been shown to lead to increased risk for alcohol and substance use disorders (Table 2.1). The environmental risk factor effect from child sexual abuse was used in the model to determine how much predictive accuracy could be obtained from combining genetic information with effects of environmental factors. Table 2.1 summarizes studies reporting the effect sizes that child sexual abuse has on alcohol and substance use disorders.

Table 2.1 Childhood sexual abuse and risk for alcohol dependence

Study	Trauma	Odds Ratio	Details about study
(Dinwiddie et al., 2000)	CSA	2.81 (CI = 1.89-4.17) in females	Prevalence of CSA was 5.9% in females and 2.5% in male
	CSA	1.91 (CI = 1.08-3.39) in males	
	CSA	3.28 in F, 3.79 M	ORs after controlling for parental alcohol problems and birth cohort
(Kendler et al., 2000)	Any CSA	~3 for AD	Prevalence of CSA in study subjects: 17-21%, females only; see table for CIs
	CSA with intercourse	4-6.5 for AD	
(McLaughlin et al., 2010)	CSA	2.02 (CI = 1.45–2.80) for AD	Adjusted for co-twin AD status, zygosity, and interaction between zygosity and co-twin AD status
(Nelson et al., 2002)	CSA, with intercourse	3.6 for AD	Prevalence of having at least 1 CSA in study subject = 16.7% in females, 5.4% in males
	CSA, no intercourse	1.81 for AD	
	Sexual abuse	1.6 for SUD	OR for SUD was highest for sexual abuse
(Sartor et al., 2007)	CSA	1.47 for alcohol consumption	

Brief summary from several studies on CSA (child sexual abuse): ORs roughly ranged from ~1.4-6.5 AD = alcohol dependence; SUD = substance use disorder

Methods

In collaboration with Dr. Brion Maher, we simulated discovery samples, calculated the number of true loci based on different genetic and epidemiological models of disease, created polygene scores based on associated loci at varying *p*-value thresholds, and then assessed for discriminatory accuracy for the disease of the scores using receiver operating characteristic (ROC) curve analyses. We investigated the following models for effects of

polygene scores on discriminatory ability, based on prior epidemiological studies: schizophrenia and bipolar disorder, with a prevalence of 1% (Perala et al., 2007; Weissman et al., 1996) and heritability of 80% (Sullivan et al., 2003); major depressive disorder, with a prevalence of 13% (Hasin et al., 2005) and a heritability of 30% (Sullivan et al., 2000), and alcohol dependence, with a heritability of 50% (Kendler et al., 1992; Heath et al., 1997) and a prevalence of 13% (Hasin et al., 2007; Kessler et al., 1994). The total number of independent SNPs assessed in the Stage I discovery sample was 100,000. The mean AUC in the Stage II validation sample was calculated over 100 iterations for each model.

We varied the minor allele frequency to be 0.05, 0.1, 0.2, 0.3, and 0.4. The GRR was ranged from 1.0 to 2.0. The *p*-value thresholds for significance in Stage I disease for selection of SNPs to predict risk in Stage II disease used were Pt < $1x10^{-4}$, Pt<0.001, Pt<0.01, Pt=0.05, Pt=0.1, Pt=0.2, Pt=0.3, Pt=0.4, and Pt=0.5. The discovery and target sample sizes included 1,000 cases and 1,000 controls, 2,000 cases and 2,000 controls, 5,000 cases and 5,000 controls, and 10,000 cases and 10,000 controls. The number of disease loci was calculated based on the equation derived by Wray et al. (2007):

$$N = \frac{(\log(h^2 + (1 - \mathrm{h}^2)K) - \log(K))}{\{2[\log(1 + \mathrm{MAF}(\mathrm{GRR} * 2 - 1)) - \log((1 + \mathrm{MAF}(\mathrm{GRR} - 1))2)]\}}$$

where N = number of disease loci, h^2 is the heritability of the disease, K is the prevalence of the disease, MAF is the minor allele frequency of the variant, and GRR is the genotype relative risk, or the ratio of disease risks between those with and those without the susceptibility genotype(s) for the disease loci.

The baseline genetic risk for a homozygous genotype for the normal allele, *b,* was calculated as:

$$b = \frac{K}{\{(1-\mathrm{MAF})^2 + 2(1-\mathrm{MAF})(\mathrm{MAF})(\mathrm{GRR}) + \mathrm{MAF}^2(1+2(\mathrm{GRR}-1))\}}$$

The probability of affection status, given genotype is:

$$P\,(D \mid G) = b\,(1+2(\mathrm{GRR}-1))$$

Using Bayes' Theorem, the probability of genotype, given affection status, is:

$$P(G|D) = \sqrt{\frac{P(D|G)MAF^2}{K}}$$

Mitra's non-centrality parameter was calculated based on sample size and case-control minor allele frequency for association tests (Mitra, 1958) . Varying critical values were set for each *p*-value cutoff. The mean polygenic sum scores and the variance of the polygenic sum scores were calculated from results for each significance threshold, based on the probability of genotypes given affection status, minor allele frequency, and proportion of null loci vs. disease loci at each threshold. The polygenic scores were then used to assess discriminatory accuracy in the Stage 2 target/validation case-controls samples. The AUCs for the polygenic scores were calculated based on the Wilcoxon rank-sum test. All data simulations and analyses were performed in SAS (SAS Institute Inc., Cary, NC, USA).

Results

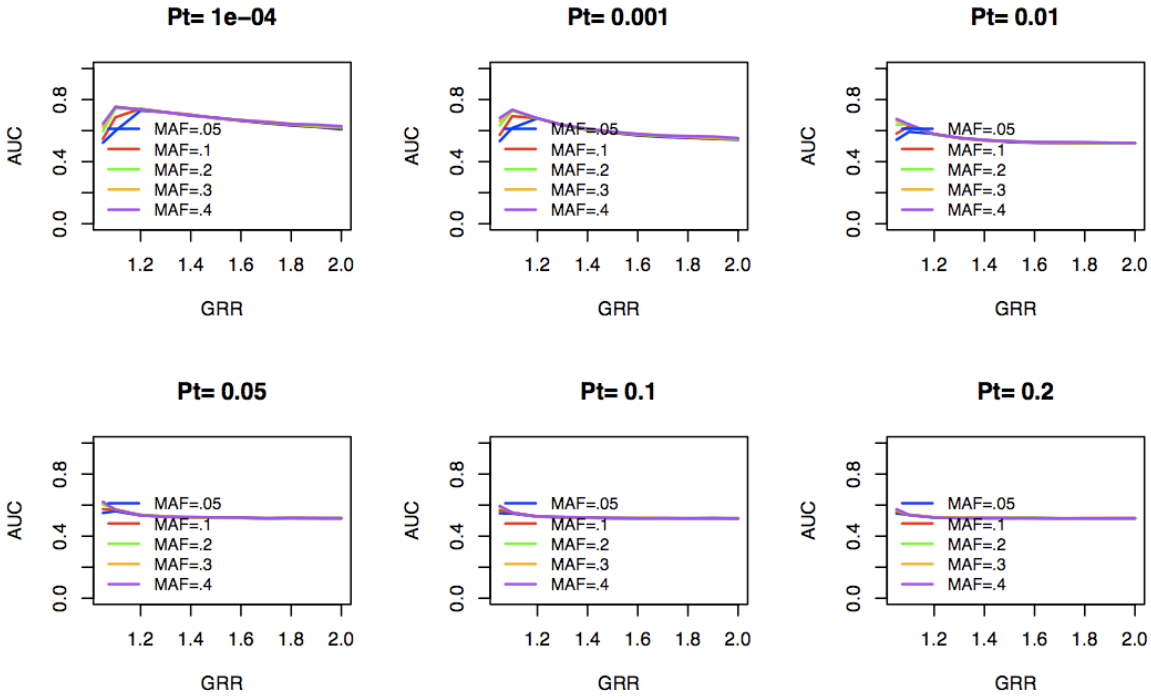
Results showed that as GRR increased, the number of susceptibility loci decreased and as heritability increased, so did the maximum AUC. At a Stage I discovery phase with 10,000 cases and 10,000 controls, AUC was highest and was just over 0.90 for a model with heritability of 0.80. Similar patterns were shown for the models for alcohol dependence, with a heritability of 0.50 (AUCs approaching 0.80), and major depressive disorder, with a heritability of 0.30, though the AUC is not as high (AUCs approaching 0.80, but not as high as those for AD). As *p*-value thresholds used to select SNPs to create polygenic scores become less stringent, at $p<1x10^{-4}$, the AUC was highest than for more liberal *p* – value thresholds. Adding the effect of environmental risk factors increased the AUC substantially.

Figure 2.1 illustrates the AUCs corresponding to each polygenic score for models for major depressive disorder, alcohol dependence, and schizophrenia. The genotype relative risk (GRR) is plotted on the x-axis. Plots are shown separately for scores composed of variants with different minor allele frequencies. Rho represents the genetic correlation between the discovery (Stage 1) and target (Stage 2) sets. In the scenarios shown below in which the same phenotype is assessed for Stage 1 and Stage 2 datasets, rho is 1 between the discovery and target samples. Shown below are the maximum AUC estimates obtained from the models. The results from the largest Stage 1 discovery case-control set of 10,000 cases and 10,000 controls had the highest AUCs. Changes in the number of Stage 2 validation cases-control set did not change the AUC estimates. Smaller discovery sample sizes reduced the maximum AUC.

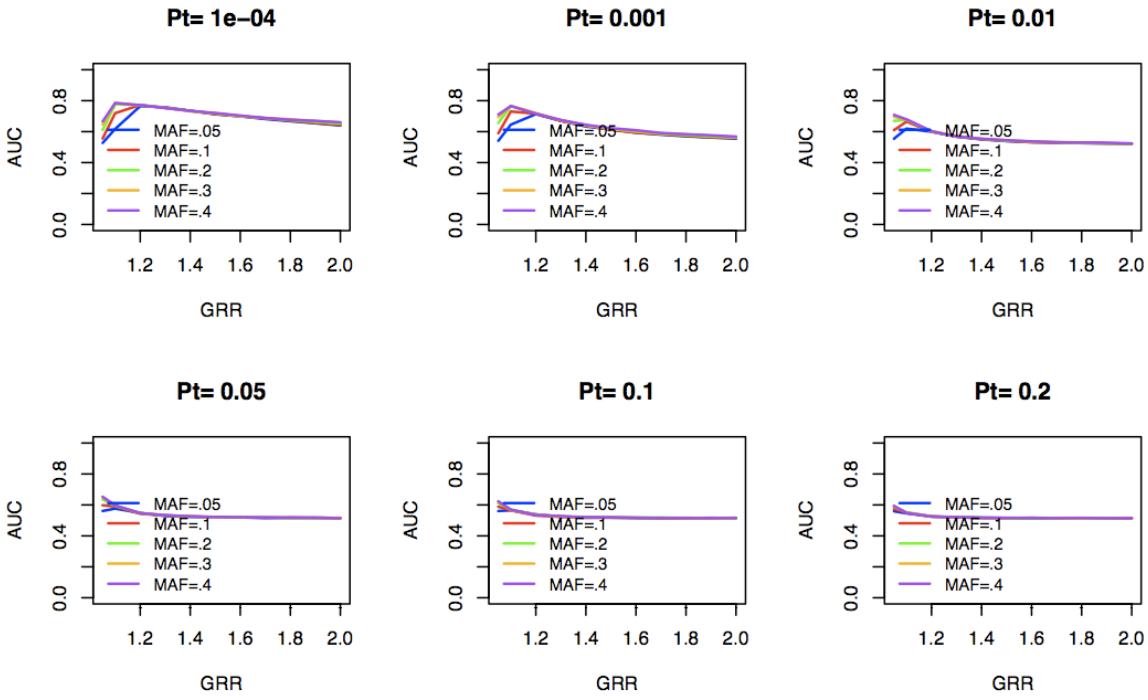
Figure 2.1 AUC estimates for polygenic scores for three different psychiatric disorders.

Details specific to each disorder are listed, including heritability and prevalence. Polygenic scores selected at different p-value significance thresholds are plotted in separate panels for each condition. The number of independent SNPs represents SNPs used in the association analyses in Stage 1 samples. The number of Stage 1 (discovery sample) and Stage 2 (target sample) cases and controls are specified. The y-axis plots the AUCs corresponding to each model. The x-axis lists genotypic relative risk (GRR). Results are plotted separately for varying minor allele frequencies (MAF).

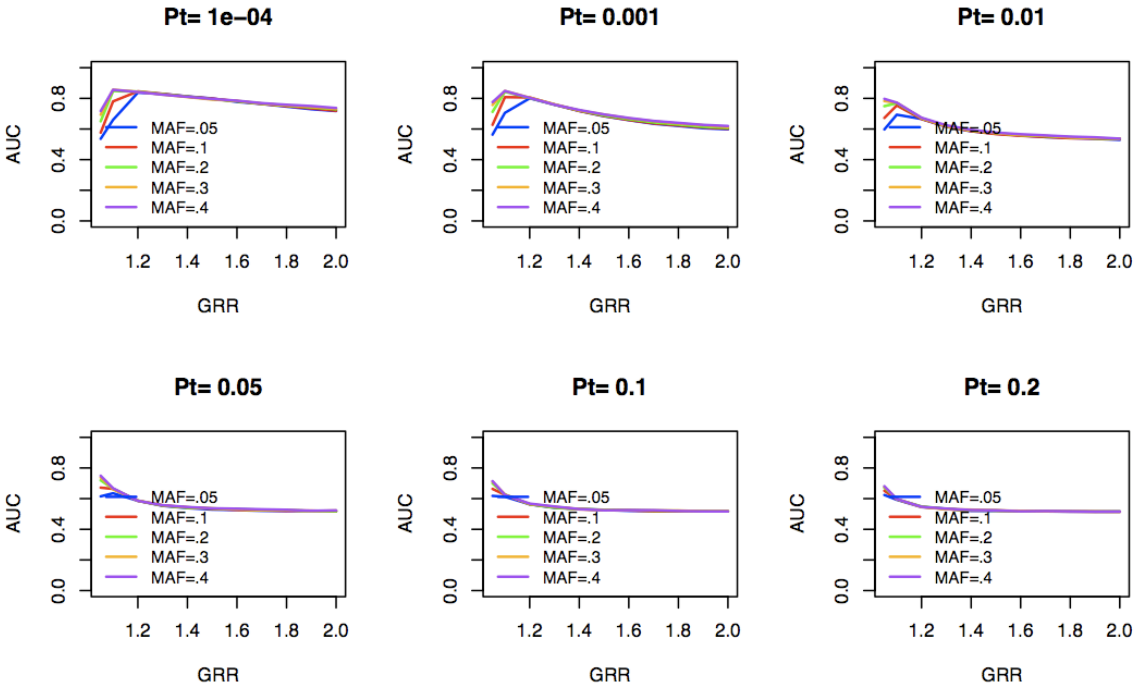
Major Depressive Disorder
10,000 Stage 1 Cases, 10,000 Stage 1 Controls
1,000 Stage 2 Cases, 1,000 Stage 2 Controls
100,000 Independent SNPs, rho = 1, h^2 = 30%, Prevalence = 13%



Alcohol Dependence
10,000 Stage 1 Cases, 10,000 Stage 1 Controls
1,000 Stage 2 Cases, 1,000 Stage 2 Controls
100,000 Independent SNPs, rho = 1, h^2 = 50%, Prevalence = 13%



Schizophrenia
10,000 Stage 1 Cases, 10,000 Stage 1 Controls
1,000 Stage 2 Cases, 1,000 Stage 2 Controls
100,000 Independent SNPs, rho = 1, h^2 = 80%, Prevalence = 1%



Based on the model for AD, Table 2.2 summarizes AUCs resulting from prediction using polygenic scores that include increasing proportions of true loci, and conversely, decreasing proportions of null loci for a hypothetical model containing 100 true loci, with minor allele frequency of 0.30 and genotypic relative risk of 1.20. When a polygenic score consisting of 100% true loci and 0% null loci is included as the sole predictor for AD, with a prevalence of 13% and heritability of 50%, then the score has an AUC of approximately 0.78, under the maximum AUC conditions described shown in Figure 2.1. Figure 2.2 illustrates the AUCs of scores that include both genetic and environmental effects for a model including all 100 true loci.

Table 2.2 AUCs for polygenic scores consisting of mixture of true and null loci.

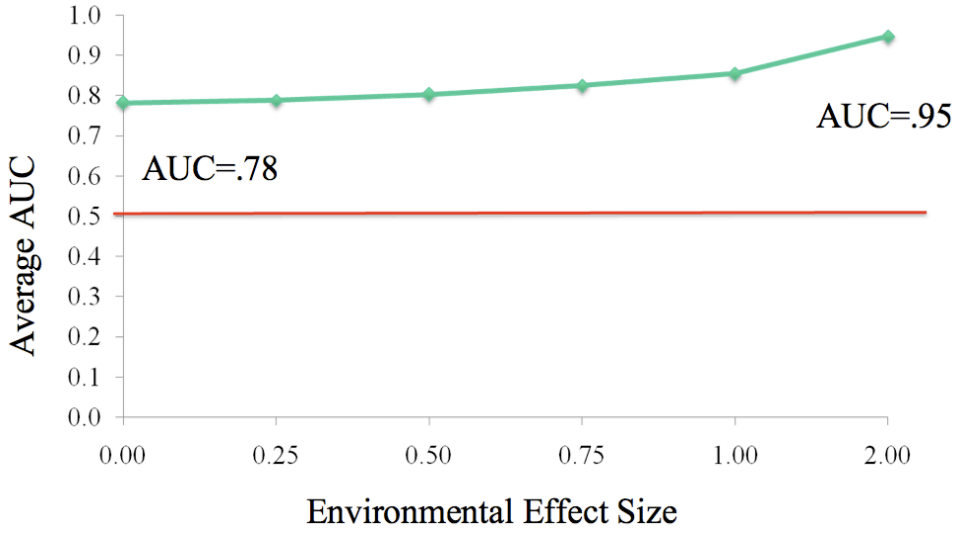
Proportion of true loci	N true loci	N null loci	AUC
1	100	0	0.77980444
0.9	90	10	0.75846888
0.8	80	20	0.7340018
0.7	70	30	0.71060652
0.6	60	40	0.68497332
0.5	50	50	0.65534604
0.4	40	60	0.63169552
0.3	30	70	0.59919648
0.2	20	80	0.56812712
0.1	10	90	0.53579312

This table displays the corresponding AUCs for increasing proportions of true loci (*N true loci)* compared with number of null loci *(N null loci)* contributing to AD for a model in which there exists 100 true loci.

Figure 2.2 Plot of AUC estimates for polygenic scores combined with environmental effects.

The y-axis shows the AUC estimates for the polygenic scores for alcohol dependence. The green line in this figure plots the AUCs corresponding to a genetic score composed of 100% of 100 true loci, with the addition of environmental factors that contribute to alcohol dependence. Plotted on the x-axis are increasing effect sizes for environmental factors. The red line plots the AUC = 0.50 point, which is equivalent to the AUC a score

that can classify accurately 50% of cases and controls, or predicting no better than chance.



Discussion

Data simulations presented here show that using a highly-powered sample with a large discovery sample size to detect a greater proportion of true loci contributing to three common psychiatric disorders produces higher AUCs for polygenic scores created based on SNPs meeting more stringent *p*-value thresholds of $p < 10^{-4}$. Discovery sample sizes of 10,000 cases and 10,000 controls resulted in polygenic scores that had the highest discriminatory accuracy, and showed that discovery samples of this magnitude are necessary for AUCs approaching 0.80. As expected, AUCs of polygenic scores were highest for the most heritable condition, schizophrenia, lower for alcohol dependence, and lowest for the condition with the lowest heritability of the three, major depressive

disorder. AUCs of greater than 80% were achieved for schizophrenia. For alcohol dependence and major depressive disorder, AUCs approached 0.80.

These results differ from results using existent data from the Welcome Trust Case Control Consortium (WTCCC) and the International Schizophrenia Consortium (ISC) in that SNPs in the latter two studies that were selected based on the most liberal thresholds actually resulted in the highest proportion of variance accounted for in the trait in question. The reason for this difference lies in failure to fully correct for population stratification in the real data. Population stratification can occur if cases and controls differ in a variable other than disease status, and concurrently differ in frequencies of alleles that are correlated with this additional variable. If this occurs, allele frequency differences between cases and controls that are attributable to differences in the additional variable could mistakenly be attributed to disease status if differences in the third variable are not corrected for. For example, if there exist differences in ancestry between cases and controls and these differences are not accounted for, alleles attributable simply to ancestry differences could be spuriously associated with the disease phenotype. In the case of WTCCC, because population stratification was not fully accounted for, only more liberal significance thresholds would be able to incorporate a greater proportion of true loci in the polygenic score. Scores composed of variants meeting more stringent *p*-value thresholds in this case would incorporate more spurious results due to population stratification that would therefore fail to replicate in independent samples (Evans et al., 2009).

As the size of the discovery sample increases, there is more power to detect individual SNPs of small effects. Similar to our results, Purcell et al. (2009) showed

through simulations that in cases of a larger discovery sample size with more power, a more stringent *p*-value threshold would be able to select SNPs that account for more of the variance in a trait in a target sample than SNPs selected at the same threshold in a smaller discovery sample with less power. In a large sample, a less stringent threshold would include more null loci that outweigh the true loci that have been detected in the sample at more stringent thresholds, and therefore SNPs selected at more liberal thresholds would actually account for less variance in a trait in a target sample (Purcell et al., 2009). This is consistent with our results in larger discovery sample sizes of 10,000 cases and 10,000 controls, which had the highest AUCs for the most stringent *p*-value thresholds. The fact that AUCs increased with increasing discovery sample sizes, but did not change with increasing validation sample sizes stresses the importance of developing sample sizes with high power to detect true associations before assessing for replication or clinical validation in independent samples. Smaller discovery sample sizes have insufficient power to detect true risk loci with small GRRs. Collaborative projects and consortia are underway for psychiatric conditions, making studies with sample sizes of 10,000 cases and 10,000 controls a reality, particularly in the Psychiatric GWAS consortium, which is a large-scale collaboration studying five major psychiatric diseases: ADHD, autism, bipolar disorder, major depressive disorder, and schizophrenia (Sullivan, 2010). The largest study to date for an alcohol-related trait consists of 12 European American samples, including a total of 21,607 individuals, with a replication sample of 21,185 individuals. This study reported an association between alcohol consumption and the *autism susceptibility candidate 2* gene (*AUTS2)* and was one of the first studies to report a genome-wide significant finding for an alcohol-related trait (Schumann et al.,

2011). However, the combined sample for alcohol consumption was population-based; none of the 12 samples was ascertained based on alcohol dependence diagnosis. Currently, many alcohol dependence samples have modest sizes; many include fewer than 1,000 alcohol-dependent cases. Alcohol dependence consortia combining multiple samples are currently underway. In Europe, the Alcohol-GWAS (AlcGen) Consortium was created as a part of the European Network on Genomic and Genetic Epidemiology (ENGAGE) Program, which studies a variety of common complex diseases (http://www.euengage.org/). In the United States, a meta-analysis of alcohol consumption is being undertaken using European American samples in the National Cancer Institute (NCI; N = 17,000) and the Gene Environment Association Studies Consortium (GENEVA; N = 17,000), which includes the collaborative alcohol dependence sample, the Study of Addiction: Genes and Environment (SAGE; N = 4121) (Agrawal et al., 2012) .

The results of these data simulations show that there is potential for discriminatory accuracy that reaches higher AUCs typical of screening tools for psychiatric conditions with similar genetic models to the one presented here for schizophrenia. For alcohol dependence, when all genetic contributions are known and used to predict risk in independent samples, then the AUC could potentially approach 0.80. Adding larger environmental effects such as the one modeled by the lower end estimate of the effect of child sexual abuse on AD (odds ratio ~2) increases the AUC for classifying AD even more to 0.95, which exceeds the 0.80 marker of a good screening tool. As we continue to pursue the identification of risk factors for alcohol dependence, we will learn more in depth about the genetic architecture of alcohol dependence. This

knowledge may in turn lead to better risk assessment using genetic and environmental factors.

Chapter 3: Genetic risk prediction using candidate gene variants and family history

Abstract

A number of studies investigating the clinical utility of genetic variants associated with complex disorders have illustrated the limitations and potential benefits of using genetic information in risk prediction for complex traits (Evans et al., 2009). The focus of this study was to assess the clinical validity of previously published genetic variants associated with alcohol dependence (AD) in predicting risk for AD in an independent sample. The predictive ability of these variants in aggregate was compared to family history. Using the Collaborative Study on the Genetics of Alcoholism (COGA) and the Study of Addiction: Genes and Environment (SAGE) genome-wide association study (GWAS) samples, we performed receiver operating characteristic (ROC) curve analysis to estimate the ability of a panel of SNPs to correctly classify cases and controls for DSM-IV AD. Specifically, sum scores of risk alleles were generated for a panel of 22 semi-independent SNPs correlated at r^2 <0.50, covering 15 genes, and a panel of 18 SNPs, correlated at r^2<0.25, covering 15 genes, that had reported associations with alcohol dependence in the COGA high-density family-based association sample. We identified a subset of individuals consisting of 627 cases and 454 controls from the COGA GWAS sample and 610 cases and 992 controls from the SAGE GWAS sample

that were not part of the original family-based association sample. We then performed ROC analysis for the sum scores in these subsets. These analyses did not result in significant discriminative ability for the sum scores; the area under the ROC curve (AUC) for the panel of SNPs correlated at $r^2<0.50$ was 0.498 (95% CI = 0.463, 0.533, p = 0.915) in COGA and 0.496 (95% CI = 0.466, 0.525, p = 0.782) in SAGE. For the SNPs panel correlated at $r^2<0.25$, the AUC was 0.491 (95% CI = 0.456, 0.525, p = 0.595) in COGA and 0.492 (95% CI = 0.462, 0.521, p = 0.583) in SAGE. These results suggest that the SNPs are not predicting better than chance. The presence or absence of family history for AD was a better classifier of case control status in the COGA sample, with an AUC of 0.686 (95% CI = 0.654, 0.718, $p < 0.001$) and 0.587 (95% CI = 0.558, 0.617, $p <$ 0.001) for a paternal history of AD-related traits in SAGE. This study shows that these SNPs currently have limited clinical validity and illustrates the need for further expansion of prediction panels for a complex disorder that encompasses both environmental and genetic risk factors of small effect such as AD.

Introduction

Alcohol dependence (AD) is a complex psychiatric disorder with approximately 50-60% heritability (Kendler et al., 1992; Heath et al., 1997; Gelernter and Kranzler, 2009) and 12.5% lifetime prevalence in the United States (Hasin et al., 2007). Numerous genetic variants have been reported to be associated with AD. Many of these gene-finding measures were carried out in the Collaborative Study on the Genetics of Alcoholism (COGA), which is a large-scale collaborative study consisting of families with individuals who meet both DSM-IIIR and Feighner criteria for AD recruited from alcohol treatment centers across the United States. Family-based association studies in a high-density subset of the COGA sample consisting of families with 3 or more first-degree relatives who meet lifetime criteria for AD yielded many associated genes, many of which have been replicated in other studies (Edenberg and Foroud, 2006).

Genes that have been associated with AD in COGA include genes involved in alcohol metabolism, such as the alcohol dehydrogenase and aldehyde dehydrogenase genes (*ADH* and *ALDH)* (Edenberg et al., 2006; Luo et al., 2005b; Edenberg, 2007a; Guindalini et al., 2005) . Genes encoding subunits of receptors that respond to gamma-aminobutyric acid (GABA), the major inhibitory neurotransmitter in the central nervous system, including *GABRB3*, *GABRG3, and GABRA2,* as well as dopamine receptor gene *DRD2* and the neighboring gene *ANKK1*, have been associated with AD, providing evidence that variation affecting reward pathways could be involved in susceptibility to AD (Edenberg et al., 2004; Covault et al., 2004; Lappalainen et al., 2005; Drgon et al., 2006; Fehr et al., 2006; Soyka et al., 2008; Enoch, 2008; Enoch et al., 2006; Dick et al., 2007d; Dick et al., 2004; Noble et al., 1998; Song et al., 2003). Additional genes

encoding receptors such as the nicotinic receptor *CHRNA5* (Wang et al., 2009; Saccone et al., 2007), the opioid receptor genes *PDYN* and *OPRK1* (Xuei et al., 2006; Xuei et al., 2007; Gerra et al., 2007; Williams et al., 2007), the muscarinic receptor *CHRM2* (Wang et al., 2004; Luo et al., 2005a; Kendler et al., 2011), and the neurokinin receptor *TACR3* (Foroud et al., 2008), and *ACN9* (Dick et al., 2008b), which is involved in gluconeogenesis have all been associated with AD in COGA. Table 3.1 lists the genes that have been associated with AD in the COGA high-density subset, with its corresponding COGA family-based association study and replication studies.

Table 3.1 Genes Associated with Alcohol Dependence in COGA

Study	Gene	Replication
(Edenberg et al., 2004)	*GABRA2*	(Covault et al., 2004; Fehr et al., 2006; Lappalainen et al., 2005; Soyka et al., 2008; Enoch et al., 2006; Drgon et al., 2006)
(Dick et al., 2004)	*GABRB3* and *GABRG3*	(Noble et al., 1998; Song et al., 2003) (*GABRB3)*
(Wang et al., 2004)	*CHRM2*	(Luo et al., 2005a; Kendler et al., 2011)
(Hinrichs et al., 2006)	*TAS2R16*	
(Wang et al., 2009)	*CHRNA5*	(Saccone et al., 2007)

(Xuei et al., 2006)	*PDYN* and *OPRK1*	(Williams et al., 2007; Gerra et al., 2007)
(Edenberg et al., 2006)	*ADH* genes: *ADH4, ADH1A, ADH1B*	(Luo et al., 2005b; Guindalini et al., 2005)
(Edenberg et al., 2008)	*NFKB1*	(Kendler et al., 2011)
(Foroud et al., 2008)	*TACR3*	
(Dick et al., 2008b)	*ACN9*	
(Dick et al., 2007)	*ANKK1/DRD2*	(Comings et al., 1994; Blum et al., 1990; Blum et al., 1991; Noble et al., 1991; Amadeo et al., 1993; Amadeo et al., 2000; Foley et al., 2004; Hietala et al., 1997; Ishiguro et al., 1998; Konishi et al., 2004; Kono et al., 1997; Dick et al., 2007d)

The association of genetic variants with complex disease has spurred dialogue on and assessment of risk prediction using genetic information for common multifactorial disorders (Jostins and Barrett, 2011; Janssens et al., 2006). For some complex disorders, such as diabetes, risk algorithms based on clinical measures such as the Cambridge and Framingham risk score have a high degree of clinical validity for screening; the area under the receiver operating characteristic (ROC) curve (AUC) exceeds 0.80 for type II diabetes. In these cases, genetic information has not been shown to add predictive value (Talmud et al., 2010). A risk model for alcohol dependence based on clinical variables does not exist. The process of genetic counseling for a complex psychiatric disorder such

as alcohol dependence involves helping individuals understand, manage and cope with genetic risk so that they have less anxiety and more empowerment over what many consider to be a devastating disorder over which one has little control (Peay and Sheidley, 2008).

A discussion for an individual who is concerned about risk for alcohol dependence may focus on current knowledge about the etiology of alcohol dependence and a detailed history of clinical and sub-clinical features for alcohol dependence and possible co-occurring conditions in both sides of the family. Risk assessment combines family history, environmental risk factors, and empiric risk estimates for alcohol dependence across family studies (Peay and Sheidley, 2008).

Current risk assessment does not include genetic testing for common variants and the predictive value of genetic testing for alcohol dependence has yet to be determined. This study investigates whether a panel of candidate gene SNPs that have been associated with alcohol dependence can be used in risk prediction for alcohol dependence. Most of the genetic variants contributing to complex disorders such as type II diabetes and alcohol dependence have small effect sizes. This, along with the fact that alcohol dependence has both genetic and environmental risk factors, means that any one SNP alone is not expected to be a good predictor of alcohol dependence. This study aims to explore the aggregate impact of multiple genetic variants with small effect sizes on clinical risk prediction.

This study also assesses the validity of family history in predicting risk for alcohol dependence. Of the general population, approximately 40% report some family history of AD and approximately 7-9% of the population report having both first and

second-degree relatives with AD (Gamm et al., 2004). Family history can account for more of the latent genetic vulnerability and parental environment contributing to alcohol dependence that is not captured on panels of candidate gene SNPs alone. Research has shown that family history can be a powerful tool for prognosis and prediction and can have practical clinical utility for both complex diseases and Mendelian syndromes. It can be used to help predict illness severity and stratify individuals into specific prognosis groups with distinct treatment and prevention needs (Pyeritz, 2012; Odgers et al., 2007; Milne et al., 2009).

In this study, we created additive genetic sum scores based on genetic variants in candidate gene studies from the high-density family-based association analyses in COGA that are summarized in Table 3.1. We determined the allele conferring increased risk for AD in the high density family-based association sample and then created sum scores by adding the number of risk alleles carried by individuals in two independent samples: the portion of the COGA Genome-Wide Association Study (GWAS) sample that was not part of the COGA high-density family-based association sample and a portion of the Study of Addiction: Genes and Environment (SAGE) GWAS sample that is independent of the COGA sample. We also explored the effect on risk prediction when family history information and genetic information were combined. We tested whether individual variants may add more specific information for an individual's risk profile, beyond that of the latent genetic factors captured by family history alone, and therefore increase risk prediction. Furthermore, because the variants associated with AD from candidate gene studies were associated in more densely affected samples with multiple affected family members, we assessed whether the candidate gene sum scores would be more informative

in determining affection status in the context of a positive family history of AD compared with a negative family history of AD. In these analyses, we stratified the sample into positive and negative family history of AD before performing subsequent ROC curve analyses.

Materials and methods

Sample and measures

COGA family-based association analysis sample

COGA is a large-scale multi-center family study with 10 collaborative sites across the United States. The sample consists of families containing probands meeting both DSM-IIIR and Feighner criteria for AD ascertained since 1989 from outpatient and inpatient alcohol treatment centers at 7 sites across the United States: Indiana University, State University of New York Health Science Center, University of Connecticut, University of Iowa, University of California/San Diego and Washington University in St Louis, and Howard University (Begleiter et al., 1995) . Families were interviewed using a poly-diagnostic instrument, the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA), which assesses Feighner, DSM-IIIR, DSM-IV, and ICD-10 criteria for major psychiatric disorders (Feighner et al., 1972; American Psychiatric Association, 1987; American Psychiatric Association, 2000; World Health Organization, 2008) . More than 1300 probands with AD have been recruited. Unaffected subjects were defined as individuals who drank but did not meet criteria for AD or other substance abuse disorders (Wang et al., 2008). In order to obtain normative measures and provide a comparative

general population sample, unscreened control families were selected from the community through a variety of methods (Edenberg et al., 2005; Edenberg and Foroud, 2006). A subset of the COGA sample was identified as a group of high-density families with 3 or more first-degree relatives who met lifetime criteria for AD. This sample consists of more than 300 extended families for a total of more than 3000 individuals (Edenberg and Foroud, 2006).

SNPs included in this analysis were selected from 9 COGA papers reporting family-based association analyses for AD using individuals from the high-density subset (Table 3.1). The number of individuals included varied across studies: association analyses that encompassed all ancestries ranged from 2139 to 2310 individuals from 262 families; 35 of these families, comprising a total of 298 individuals, are of African American (AA) ancestry. Analyses conducted in the European American (EA) subset ranged from 1172 to 1923 individuals from 217-219 families. Genotyping for these individuals is described in detail in the original COGA papers. Briefly, SNPs within and flanking candidate genes were selected from public databases including dbSNP (http://www.ncbi.nlm.nih.gov/SNP), HapMap (http://www.hapmap.org), and LocusLink (http://www.ncbi.nim.nih.gov/LocusLink/refseq.html). Genotyping was done using a modified single nucleotide extension reaction, with allele detection by mass spectrometry (Sequenom MassArray system; Sequenom, San Diego, CA, USA). SNPs were in Hardy Weinberg Equilibrium. Genotypes were checked for Mendelian inheritance using programs including PEDCHECK. USERM13 was used to calculate marker allele frequencies and heterozygosities (Edenberg et al., 2008).

COGA GWAS sample

A case-control sample of 1945 phenotyped subjects was selected from the larger COGA sample for genome-wide association studies. Cases had a lifetime diagnosis of AD by DSM-IV criteria. Controls reported consuming alcohol but did not have a diagnosis of AD or alcohol abuse by any of the diagnostic criteria assessed by SSAGA and did not meet diagnostic criteria for dependence on cocaine, marijuana, opioids, sedatives, or stimulants. Controls could not share a known common ancestor with a case and were preferentially selected to be above the age of 25 years. 1081 individuals in the COGA GWAS EA sample were independent of the COGA family sample.

Genotyping was completed using the Illumina Human 1M DNA Analysis BeadChip at the Center for Inherited Disease Research (CIDR). Additional details on the COGA GWAS sample can be found in Edenberg et al. (2010).

SAGE GWAS sample

The Study of Addiction: Genes and Environment (SAGE) is part of the Gene Environment Association Studies initiative of the National Human Genome Research Institute to identify genetic contributions to addiction through large-scale genome-wide association studies. The entire SAGE sample consists of 4,121 cases and unrelated controls from subsets of three large studies on addiction: the Family Study of Cocaine Dependence (FSCD), the Collaborative Genetic Study of Nicotine Dependence (COGEND), and COGA. All cases in SAGE have DSM-IV lifetime diagnosis of AD. Controls were exposed to alcohol. Some controls met criteria of nicotine dependence based on the Fagerström Test for nicotine dependence, but none met criteria for a DSM-

IV lifetime dependence diagnosis for alcohol, marijuana, cocaine, opiates or other drug. The FSCD and COGEND portions of the SAGE GWAS EA sample were extracted for use as a second independent sample to assess for discriminative ability.

Genotyping for all three studies that are part of the SAGE GWAS sample was completed at CIDR using the Illumina Human 1M DNA Analysis BeadChip. Additional details on the SAGE GWAS sample can be found in Bierut et al. (2010).

Family history measures

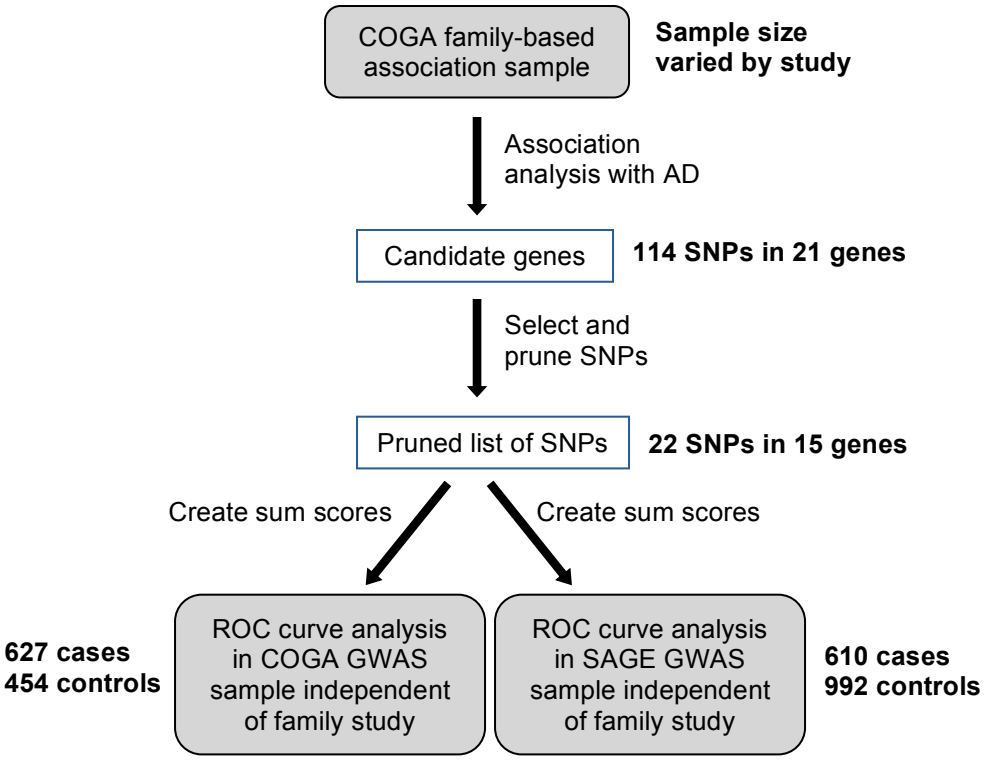
Family history information for the COGA GWAS sample was obtained for both cases and controls as a dichotomous "yes" / "no" variable for any existence of a family history of AD in relatives, as reported by the subject. The SAGE GWAS sample included a "yes" / "no" variable about history of AD in specifically the proband's mother and father. The presence or absence of family history was primarily used as a binary variable in order to reflect clinical scenarios in which an individual is asked whether or not she or he has a family history of alcohol dependence. Both COGA and SAGE had information about parental history of AD based on a more inclusive, or "relaxed" criterion and a more stringent "strict" criterion. Family history information was also expanded further into an ordinal variable based on parental history that addressed whether an individual had 0, 1, or 2 parents with a history of alcohol dependence.

Data analyses

Data analysis overview

SNPs that were previously associated with AD in candidate gene studies in the COGA high-density family-based association sample were used to create a candidate gene sum score to assess prediction of AD in independent individuals from the COGA and SAGE GWAS samples. Because the SNPs contributing to the candidate gene sum scores were previously reported to be associated in either all-EA or >85% EA samples, a genetic sum score based on results from these studies would possibly be applicable primarily to EA individuals. Therefore, we assessed predictive ability in the EA subsets of the COGA and SAGE samples. We first determined the alleles conferring risk for AD using the family-based association study, and then used the alleles to create additive genetic sum scores to assess for risk in the COGA and SAGE GWAS samples (Figure 3.1).

Figure 3.1 Study Overview



Family-based association analysis

In order to create genetic sum scores for prediction of alcohol dependence, risk alleles needed to be determined for each SNP that had been associated with alcohol dependence in prior reports. Because the exact allele conferring increased risk for alcohol dependence was not explicitly reported for each COGA paper, analyses were repeated for each study using the Pedigree Disequilibrium Test (PDT) in UNPHASED (PDTPHASE), as described in the original COGA papers (Table 3.1). The statistic used for association, specific measure for alcohol dependence, and the population selected for analyses varied across the published COGA studies. For example, the study reported by Edenberg et al. (2004) used the PDT average statistic, which averages all association statistics across families, to study the association of *GABRA2* SNPs with AD diagnosed using DSM-IV criteria, whereas the study by Xuei et al. (2006) also reported the PDT average statistic, but examined the association of *PDYN* with AD based on DSM-III and Feighner criteria. In contrast to the PDT statistic used by Xuei et al. and Edenberg et al., the study by Dick et al. (2007) used the PDT sum statistic, which places greater weight on families with more informative trios and discordant siblings. All results in our study were generated to match the statistic, diagnosis, and population used by each previously published COGA candidate gene study. The PDT average statistic was used in the majority of the COGA candidate gene studies. Wang et al. (2004) reported the PDT sum statistic in addition to the PDT average statistic. In order to match the majority of studies, statistics used in our association analysis consisted of the PDT average and/or sum statistics. In addition to PDT, Dick et al. (2004) also performed a classic TDT analysis in TRANSMIT using one trio selected from each COGA family. Wang et al. (2008) used the family-based

association test (FBAT); risk alleles for these analyses were obtained via correspondence with Dr. Wang. In the family-based association analyses, multiple outcomes were used across studies. Depending on the diagnosis used in the study, the DSM-IV, DSM-IIIR, or DSM-IIIR + Feighner Criteria for the COGA definition of alcohol dependence were used as outcomes.

SNP selection

Several criteria were used in the selection of the panel of SNPs used for classification of alcohol dependence status. An initial list of 114 SNPs across 21 genes was generated based on prior association with alcohol dependence according to either DSM-IV or COGA (DSM-IIIR + Feighner) criteria. A smaller proportion of the sample had early-onset (≤ 22 years of age) alcohol dependence (N = 454 in COGA and N=811 in SAGE). SNPs associated only with early onset alcohol dependence were removed from the list so that SNPs in the candidate gene panel would be applicable to the wider range of ages of individuals in the full COGA and SAGE validation samples. The age of onset for AD in the independent COGA GWAS sample ranged from 12 to 65 years of age. Age of onset in the FSCD and COGEND portion of the SAGE GWAS sample ranged from 13 to 55 years of age. Because assessment of discriminative accuracy was to be performed in European American individuals, SNPs that were associated only in the African American subset were removed from the list. Forty-two of the SNPs showing association in the original papers were present on the Illumina Human 1M DNA Analysis BeadChip. Proxy SNPs on the Illumina chip with an $r^2 > 0.70$ were found for 32 SNPs based on LD calculations in the HapMap CEU data using Haploview (Barrett et al., 2005) and Plink

version 1.07 (Purcell et al, 2007). Thirty-two of the SNPs did not have proxies. Seven of these SNPs had proxies in the list of COGA family sample SNPs for which proxy SNPs existed on the Illumina chip, based on LD calculations using Haploview. The final list contained 81 SNPs.

In order that genes with a large number of associated SNPs in high LD were not disproportionately represented in the risk panel, we generated a list of semi-independent SNPs for the panel. We explored the effect of pruning SNPs based on different r^2 thresholds; SNPs that were more correlated than these thresholds were removed from the list. We first assessed the use of a more inclusive panel with a pruning threshold of $r^2 < 0.50$, and then used a more stringent threshold of $r^2 < 0.25$. All SNPs in this panel were included in the HapMap list, but not all SNPs were part of the Illumina 1M SNP chip. In order to create the panel without using information from the independent validation samples, LD estimations used for pruning the SNPs were based on the HapMap Phase 3 CEU data rather than data from the COGA GWAS sample. Calculations for LD were performed using the Plink version 1.07 LD function (Purcell et al., 2007). Selection of which SNP of a pair of correlated SNPs to remove depended on a ranked list of SNPs based on the level of significance from the family-based association results and how closely the SNP on the Illumina chip matched the original family-based SNP. SNPs were rank-listed in the following order:

1. SNPs from the family studies with exact matches in the COGA GWAS list of SNPs, ranked in descending order for their corresponding p-values from the family studies.

2. Proxy SNPs in the COGA GWAS sample for the SNPs in the family sample, with $r^2 > 0.70$ based on HapMap data, listed in descending order first for r^2 and then for p-value of the original SNP in the family study.
3. Proxies in the COGA GWAS sample to proxy SNPs in the COGA family SNPs (in HapMap) to the remaining list of SNPs that are not in HapMap. Ranked in the same way as the above proxy SNPs, listed in descending order first for r^2 and then for p-value of the original SNP in the family study.

Both pruning thresholds of $r^2 < 0.50$ and $r^2 < 0.25$ created a set of 15 genes from the original 21 genes, primarily due to correlations among the *ADH* SNPs. Table 3.2 summarizes the list of SNPs after pruning for LD based on a threshold of $r^2 < 0.50$. Pruning resulted in a set of 22 SNPs from a threshold of $r^2<0.50$ and 18 SNPs from a threshold of $r^2<0.25$. SNPs selected at the threshold of $r^2 < 0.25$ were the same as SNPs selected based on $r^2 < 0.50$, with the exception of 4 SNPs that were pruned out based on this more stringent threshold: rs2235749 and rs6045819 in *PDYN,* rs7794886 in *ACN9,* and rs997917 in *OPRK1*.

Table 3.2 Pruned set of candidate gene SNPs at $r^2 < 0.50$.

SNP	Status	Gene	COGA family study p-value	MAF Fam	MAF COGA	MAF SAGE	Risk Allele
rs10499934	In_sample	*ACN9*	0.003	0.23	0.22	0.23	A
rs12671685	In_sample	*ACN9*	0.027	0.11	0.12	0.11	A
rs7794886	In_sample	*ACN9*	0.006	0.35	0.36	0.35	T
rs4147531	In_sample	*ADH1A*	0.007	0.43	0.46	0.47	C
rs1229982	In_sample	*ADH1B*	0.048	0.22	0.20	0.19	T
rs1126672	In_sample	*ADH4*	0.010	0.29	0.28	0.29	C

rs17115439	In_sample	*ANKK1*	0.096	0.33	0.32	0.32	C
rs680244	In_sample	*CHRNA5*	0.114	0.42	0.41	0.42	G
rs1799978	In_sample	*DRD2*	0.168	0.06	0.05	0.05	G
rs279858	In_sample	*GABRA2*	0.010	0.38	0.42	0.42	A
rs1897356	In_sample	*GABRB3*	0.020	0.17	0.15	0.15	C
rs16918941	In_sample	*OPRK1*	0.023	0.06	0.06	0.07	G
rs6985606	In_sample	*OPRK1*	0.004	0.48	0.50	0.48	T
rs997917	In_sample	*OPRK1*	0.011	0.27	0.29	0.27	C
rs1997794	In_sample	*PDYN*	0.011	0.37	0.36	0.35	C
rs2235749	In_sample	*PDYN*	0.010	0.27	0.27	0.26	A
rs6045819	In_sample	*PDYN*	0.038	0.10	0.12	0.12	G
rs11722288	In_sample	*TACR3*	0.022	0.29	0.29	0.29	G
rs3762894	In_sample	*ADH4*	0.050	0.16	0.15	0.16	C
rs1391175	Use_proxy (rs13120165)	*GABRG1*	0.036	0.06	0.03	0.03	A
rs3097490	Use_proxy (rs1571281)	*GABRG3*	0.137	0.44	0.44	0.46	G
rs324640	Use_proxy (rs324649)	*CHRM2*	0.038	0.43	0.42	0.42	T

"Status" indicates whether or not the SNP was directly genotyped on the Illumina 1M SNP chip or a proxy SNP was used. The SNP numbers are SNPs from candidate gene studies, with proxy SNPs indicated as such in the "status" column. The COGA family-based association *p*-values from the previously published studies are listed. MAF Fam shows the minor allele frequency of the SNP in the COGA family-based candidate gene association sample. MAF COGA and MAF SAGE correspond to the MAF in the COGA and SAGE GWAS samples, respectively. The risk allele corresponds to the GWAS alleles matched by allele frequency to the risk allele in the family-based candidate gene association sample.

Genetic sum scores

Additive genetic risk scores were created using the --score option in PLINK v1.07 (Purcell et al, 2007). This follows an additive model for risk variants, as described by Evans et al. (2009):

$$N(risk) = \frac{\sum x_i}{n}$$

x_i = # of risk alleles (0, 1, 2) at SNP_i
n = number of nonmissing genotypes

The number of risk alleles for each candidate gene SNP was added and then divided by the total number of non-missing genotypes to create a normalized allele count for each individual. Because odds ratios associated with the risk alleles varied across family-based analyses in COGA and replication studies, the additive score was created without weighting alleles by effect size. The risk allele in the SAGE and COGA samples was determined by matching by frequency with alleles that were associated with AD in the family sample. Minor allele frequencies across the family-based association sample, COGA GWAS sample, and SAGE GWAS sample were similar for each SNP (Table 3.2).

Association analysis of candidate gene sum score with AD:

The genetic sum scores were tested for association with DSM-IV AD in the case-control COGA and SAGE samples using logistic regression with sex as a covariate in COGA and sex, age quartiles, and study site as covariates in SAGE. The GWAS association models were selected to follow the methods used in the previously reported primary COGA and SAGE GWAS analyses (Bierut et al., 2010; Edenberg et al., 2010). In addition to testing the aggregate genetic sum scores for association with AD in the sample used for prediction, the individual SNPs contributing to the scores were each tested for association with AD. All candidate gene SNPs were tested for association before performing LD-based pruning in order to assess the overall replication of the candidate gene SNPs in the independent COGA and SAGE GWAS samples. All association analyses were completed in the case-control samples using logistic regression using an additive model in PLINK v1.07. In order to assess replication across ancestries, and to account for the ethnicity differences in samples used in the previously reported candidate gene studies, association

analyses were performed in both the EA subset of the sample and the entire sample, which includes individuals of non-EA ancestry. Association analyses in the entire GWAS samples that included individuals of non-EA ancestry included molecularly derived principal components factor covariates, PC1 and PC2, distinguishing primarily between European and African ancestry.

ROC curve analyses

Discriminatory accuracy of genetic sum scores and family history was measured using ROC curve analysis in SPSS/PASW v17.0 (SPSS Inc., Chicago IL) with alcohol dependence as the binary outcome.

The genetic sum scores and family history variables were used as the predictors. Additionally, in order to assess whether a panel of SNPs associated with AD in a high-density family sample would be more informative in predicting risk for individuals with a positive family history of AD compared with individuals without a known family history, ROC analysis for the genetic sum score was also split by presence of family history. For example, in COGA, AUC was calculated for risk panels separately for those with a positive family history of alcohol dependence and those with a negative family history of alcohol dependence. In SAGE, analysis was split based on a variable created for positive family history or negative family history in the proband's parents based on strict standard criteria.

We assessed the value of combining information from the candidate gene panel with family history, as family history and the candidate gene sum scores were not correlated (Pearson's r = 0.021, n = 1081 for the candidate gene sum scores pruned at

both $r^2 < 0.50$ [p=0.490] and $r^2 < 0.25$ [p=0.480] in COGA, Pearson's r = -0.038, n=1566 for the candidate gene sum scores pruned at $r^2 < 0.50$ [p=0.137] in SAGE, and Pearson's r = -0.017, n=1566 for the candidate gene sum scores pruned at $r^2 < 0.25$ [p=0.497] in SAGE). Predicted probabilities from logistic regression using genetic sum scores + family history information were calculated and then used as continuous predictors of alcohol dependence in order to determine whether or not genetic sum scores added to family history information in risk prediction.

Results

Family-based association analysis

The re-run family-based association analyses resulted in p-values that matched those of the previously published studies for the majority of SNPs (Table 3.2). Several SNPs had different p-values in our repeat analysis due to differences in the sample inclusion criteria used in the COGA papers compared with our analysis. Because the exact individuals included was not explicitly reported in the studies, we did not have exact p-value matches in our analysis for several of the SNPs from the COGA family studies; however, p-values for the SNPs remained significant, with the exception of one SNP in *ANKK1,* for which our p-value was considerably different and not significant. This SNP was therefore not included in these analyses. For the two studies that used TRANSMIT and FBAT for association, Dick et al., 2004 and Wang et al., 2008, we used information about the risk allele obtained directly from Dr. Wang and matched the association results using PDTPHASE for several of the SNPs for Dr. Dick's study.

ROC curve and logistic regression analysis

ROC curve analysis showed that neither of the genetic sum scores created based on the pruning thresholds of $r^2 < 0.50$ and $r^2 < 0.25$ had an AUC estimate that reached statistical significance at $p < 0.05$ for discrimination of alcohol dependence status in the COGA or SAGE GWAS samples. The family history variables, however, did produce statistically significant AUCs. ROC curve analysis results for discriminative ability of family history compared with the genetic sum scores are summarized in Table 3.3 for the COGA GWAS sample. Estimates in the SAGE GWAS sample are summarized in Table 3.4. The distribution of genetic sum scores was similar in cases and controls in both the COGA and SAGE GWAS samples. Figure 3.2 displays the distributions of the candidate gene sum scores separately for cases and controls in the COGA GWAS sample.

Table 3.3 AUC Estimates of Predictors in the COGA GWAS Sample

				Asymptotic 95% Confidence Interval	
Diagnostic Classifier	AUC	Std. Error[a]	Asymptotic Sig.[b]	Lower Bound	Upper Bound
Family history	0.686	0.016	< 0.001	0.654	0.718
SCORE25 [c]	0.491	0.018	0.595	0.456	0.525
SCORE50 [d]	0.498	0.018	0.915	0.463	0.533

a. Under the nonparametric assumption
b. Null hypothesis: true area = 0.5
c. Genetic sum score based on pruned list of COGA variants at an r^2 of 0.25
d. Genetic sum score based on pruned list of COGA variants at an r^2 of 0.50

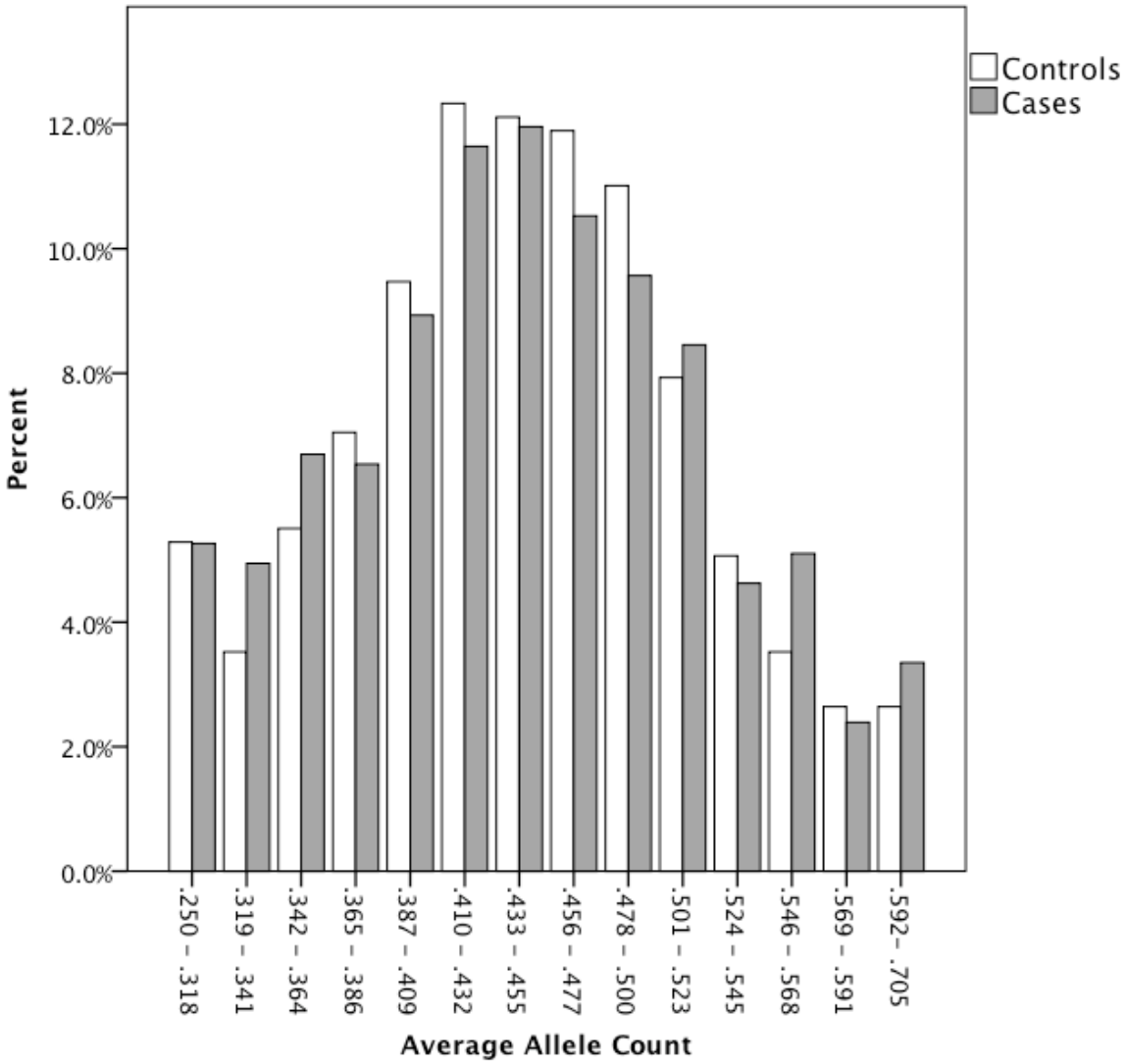
Table 3.4 AUC Estimates of Predictors in the SAGE GWAS Sample

				Asymptotic 95% Confidence Interval	
Diagnostic Classifier	Area	Std. Error[a]	Asymptotic Sig.[b]	Lower Bound	Upper Bound
SCORE25 [c]	0.492	0.015	0.583	0.462	0.521
SCORE50 [d]	0.496	0.015	0.782	0.466	0.525
History of alcoholism in mother-relaxed [e]	0.556	0.015	< 0.001	0.527	0.586
History of alcoholism in mother-strict [f]	0.547	0.015	0.002	0.518	0.577

History of alcoholism in father-relaxed [g]	0.587	0.015	< 0.001	0.558	0.617
History of alcoholism in father-strict [h]	0.582	0.015	< 0.001	0.552	0.612
History of AD in either mother or father (relaxed) [i]	0.614	0.015	< 0.001	0.584	0.643

a. Under the nonparametric assumption
b. Null hypothesis: true area = 0.5
c. Genetic sum score based on pruned list of COGA variants at an r^2 of 0.25
d. Genetic sum score based on pruned list of COGA variants at an r^2 of 0.50
e.- i. Family history predictors based on a binary absence or presence of parental family history of alcohol dependence.

Figure 3.2 Distribution of genetic sum scores based on candidate gene SNPs pruned at r^2<0.50 in cases and controls for AD in the COGA GWAS sample



Summary of distributions:

	N	Min	Max	Mean	Std. Deviation
Controls	454	0.25	0.705	0.45552	0.075985
Cases	627	0.25	0.659	0.45517	0.078184

COGA sample is independent of the COGA high-density family-based association sample. The figure shows the frequency of normalized allele counts in bins after sum scores were created by adding the number of risk alleles of SNPs associated with AD in candidate gene studies, and then dividing by the number of non-missing genotypes for each individual. The table summarizes the mean sum score, and range for the sum score separately for cases and controls.

Logistic regression showed that none of the candidate gene sum scores created in COGA and SAGE was significantly associated with alcohol dependence ($p = 0.940$ for the candidate gene sum score pruned at $r^2 < 0.50$, $p = 0.753$ for the score pruned at $r^2 < 0.25$ in the COGA GWAS sample, $p = 0.627$ for the candidate gene sum score pruned at $r^2 < 0.50$, and $p = 0.501$ for the score pruned at $r^2 < 0.25$ in the SAGE GWAS sample). Many of the individual SNPs contributing to the genetic sum scores were not significantly associated with alcohol dependence. Of the 22-SNP panel pruned at $r^2 < 0.50$, two of the SNPs met a nominal threshold of $p < 0.10$ in the independent European American subset of the COGA sample, one in the *ADH4* gene and one in the *ANKK1* gene. Two SNPs, one in the *TACR3* gene and one in the *GABRB3* gene, met this threshold in the SAGE GWAS European American sample, and one SNP in the *GABRG1* gene had a *p*-value < 0.05 in the sample. In the entire COGA GWAS sample including individuals of African American and other ancestries, one SNP in the *DRD2* gene had a *p*-value < 0.05 and the same *ADH4* SNP nominally associated with AD in the EA sample had a *p*-value < 0.10 in the COGA sample. In the entire SAGE GWAS sample, the same *GABRB3* and *GABRG1* SNPs met a threshold of $p < 0.05$. The *p*-values resulting from logistic regression analyses for individual SNPs contributing to the candidate gene sum

score panels post-LD-based pruning are shown in Table 3.5. The expanded list of all SNPs prior to pruning based on LD included a greater number of SNPs that met the nominal replication *p*-values of 0.05 and 0.10. Table 3.6 shows SNP association results of all SNPs before performing LD-based pruning.

Table 3.5 The association of individual SNPs contributing to candidate gene sum scores in COGA and in SAGE GWAS samples.

CHR	SNP	Gene	P-val COGA EA	P-val COGA All	P-val SAGE EA	P-val SAGE All
4	rs13120165	*GABRG1*	0.104	0.842	0.388	0.569
4	rs279858	*GABRG1*	0.681	0.429	**0.024	**0.017
4	rs1126672	*ADH4*	*0.073	*0.069	0.922	0.449
4	rs3762894	*ADH4*	0.510	0.128	0.592	0.501
4	rs4147531	*ADH1A*	0.571	0.940	0.806	0.508
4	rs1229982	*ADH1B*	0.104	0.337	0.604	0.594
4	rs11722288	*TACR3*	0.121	0.148	*0.061	0.266
7	rs10499934	*ACN9*	0.859	0.385	0.224	0.412
7	rs7794886	*ACN9*	0.941	0.476	0.590	0.512
7	rs12671685	*ACN9*	0.746	0.452	0.252	0.307
7	rs324649	*CHRM2*	0.868	0.610	0.429	0.121
8	rs997917	*OPRK1*	0.937	0.989	0.956	0.954
8	rs16918941	*OPRK1*	0.516	0.712	0.773	0.499
8	rs6985606	*OPRK1*	0.495	0.439	0.851	0.522
11	rs17115439	*ANKK1*	*0.077	0.238	0.964	0.825
11	rs1799978	*DRD2*	0.133	**0.040	0.239	0.480
15	rs1897356	*GABRB3*	0.570	0.847	*0.064	**0.048
15	rs1571281	*GABRG3*	0.296	0.749	0.926	0.905
15	rs680244	*CHRNA5*	0.779	0.923	0.909	0.239
20	rs2235749	*PDYN*	0.696	0.680	0.513	0.381
20	rs6045819	*PDYN*	0.840	0.687	0.652	0.535
20	rs1997794	*PDYN*	0.255	0.655	0.470	0.833

P-values are shown for logistic regression results of each individual SNP for association with AD. P-val COGA_EA indicates results of association analyses in the European American subset of the COGA GWAS sample that is independent of the COGA high-density family-based association sample. P-val SAGE_EA reflects association results in the FSCD and COGEND portion of the SAGE European American sample. COGA_All and SAGE_All show results in samples that are included in the EA portion of the COGA high-density family-based association sample, as well as independent individuals of other ancestries. ** SNPs with $p < 0.05$ for association with AD; * SNPs with $p < 0.10$

Table 3.6 Results of logistic regression for AD for all SNPs associated with AD in candidate gene family-based association studies

Original.SNP	Panel.SNP	Status	Chr	BP	Minor Allele	P_COGA_ea	P_COGA_all	P_SAGE_ea	P_SAGE_all	Gene
rs1391175	rs13120165	Use_proxy	4	45796306	G	0.1037	0.842	0.3882	0.5686	GABRG1
rs572227	rs572227	In_sample	4	45946150	A	0.7217	0.4717	0.0362*	0.0257*	GABRA2
rs548583	rs548583	In_sample	4	45958101	T	0.5835	0.3469	0.038*	0.0352*	GABRA2
rs540363	rs502038	Use_proxy	4	45975075	G	0.5948	0.3551	0.0456*	0.0517	GABRA2
rs279858	rs279858	In_sample	4	46009350	G	0.6811	0.4289	0.0235*	0.0172*	GABRA2
rs279843	rs279843	In_sample	4	46019961	T	0.7184	0.5615	0.0261*	0.0958	GABRA2
rs279841	rs279841	In_sample	4	46035520	A	0.6439	0.4576	0.0491*	0.0367*	GABRA2
rs189957	rs189957	In_sample	4	46041436	G	0.4301	0.7495	0.037*	0.0632	GABRA2
rs530329	rs10805145	Use_proxy	4	46053088	C	0.4286	0.6303	0.0466*	0.0622	GABRA2
rs1042365	rs2602891	Use_proxy	4	100262307	C	0.0774	0.0747	0.9414	0.4309	ADH4
rs1042364	rs1042364	In_sample	4	100264597	A	0.0745	0.0723	0.9414	0.4335	ADH4
rs2602866	rs1126673	Not_inHapmap	4	100264639	G	0.1373	0.3802	0.6655	0.7133	ADH intergenic region
rs1126672	rs1126672	In_sample	4	100266835	T	0.0727	0.0693	0.9219	0.4489	ADH4
rs4699714	rs4699714	In_sample	4	100279561	G	0.0745	0.0744	0.9809	0.5225	ADH4
rs3762894	rs3762894	In_sample	4	100285107	C	0.5103	0.1281	0.592	0.5013	ADH4
rs1984362	rs4699716	Use_proxy	4	100285148	A	0.0863	0.1423	0.9277	0.7161	ADH intergenic region
rs4147531	rs4147531	In_sample	4	100431220	T	0.5705	0.94	0.8057	0.5075	ADH1A
rs1826909	rs1826909	In_sample	4	100436766	T	0.2009	0.6125	0.7163	0.6092	ADH intergenic region
rs1353621	rs1353621	In_sample	4	100460598	G	0.168	0.5802	0.6477	0.7146	ADH1B
rs1159918	rs1159918	In_sample	4	100462032	T	0.0664	0.347	0.9637	0.3006	ADH1B
rs1229982	rs1229982	In_sample	4	100462955	T	0.1043	0.3366	0.6042	0.5942	ADH1B
rs980455	rs980455	In_sample	4	103637989	G	0.3076	0.2597	0.6982	0.1601	NFKB1
rs3774932	rs1598856	Use_proxy	4	103665145	T	0.6549	0.9748	0.6178	0.8151	NFKB1
rs230530	rs230530	In_sample	4	103673010	C	0.456	0.618	0.7002	0.7004	NFKB1
rs230529	rs230528	Use_proxy	4	103676615	C	0.1332	0.204	0.6217	0.5606	NFKB1
rs1801	rs1598859	Use_proxy	4	103725482	C	0.0249*	0.0372*	NA	NA	NFKB1
rs2765	rs2765	In_sample	4	104730215	G	0.1863	0.4915	0.1639	0.1436	TACR3
rs11722288	rs11722288	In_sample	4	104752399	A	0.1209	0.1481	0.0615	0.2657	TACR3
rs1917939	rs1917939	In_sample	7	96554800	A	0.5073	0.8577	0.3394	0.5938	ACN9
rs1343646	rs1343646	In_sample	7	96562388	A	0.9556	0.5776	0.2418	0.504	ACN9
rs10246622	rs4729330	Use_proxy	7	96574166	C	0.8368	0.5738	0.8085	0.6379	ACN9
rs10499934	rs10499934	In_sample	7	96575512	G	0.8593	0.3854	0.2239	0.4119	ACN9
rs7794886	rs7794886	In_sample	7	96586948	C	0.9407	0.4764	0.5895	0.5122	ACN9
rs12056091	rs12056091	In_sample	7	96596607	A	0.9702	0.7922	0.5824	0.7877	ACN9
rs12670377	rs12670377	In_sample	7	96624561	A	0.9721	0.9737	0.4452	0.3343	ACN9
rs12671685	rs12671685	In_sample	7	96644065	G	0.7456	0.4523	0.2523	0.3065	ACN9
rs1204014	rs1204014	In_sample	7	122422079	A	0.8391	0.328	0.1785	0.833	TAS2R16
rs846664	rs846664	In_sample	7	122422409	G	0.8966	0.4627	0.1133	0.2031	TAS2R16
rs324640	rs324649	Use_proxy	7	136344049	T	0.868	0.61	0.4286	0.1207	CHRM2
rs12548098	rs6651353	Use_proxy	8	54314821	G	0.3401	0.5794	0.8225	0.5328	OPRK1
rs997917	rs997917	In_sample	8	54314931	C	0.9367	0.9885	0.9563	0.9544	OPRK1
rs6473797	rs6473798	Use_proxy	8	54315658	T	0.815	0.729	0.9625	0.7883	OPRK1
rs16918941	rs16918941	In_sample	8	54323255	G	0.5159	0.7118	0.7727	0.4989	OPRK1
rs6985606	rs6985606	In_sample	8	54323669	C	0.4945	0.4389	0.8509	0.5222	OPRK1
rs17115439	rs17115439	In_sample	11	112769482	T	0.0765	0.238	0.9637	0.8252	ANKK1/DRD2
rs4938012	rs4938013	Use_proxy	11	112769680	A	0.0663	0.0661	0.9109	0.2684	ANKK1/DRD2
rs4938016	rs4938016	In_sample	11	112775225	G	0.1387	0.0852	0.8175	0.3065	ANKK1/DRD2
rs1799978	rs1799978	In_sample	11	112851561	G	0.1331	0.0396*	0.2393	0.4799	DRD2
rs1897356	rs1897356	In_sample	15	24416628	C	0.5703	0.8469	0.0645	0.0484*	GABRB3
rs3101636	rs3101636	In_sample	15	25444308	G	0.2567	0.7063	0.8957	0.9355	GABRG3
rs3097490	rs1571281	Use_proxy	15	25445018	A	0.2956	0.7494	0.9261	0.9045	GABRG3
rs2303879	rs1571280	Use_proxy	15	25445120	C	0.2714	0.7127	0.9261	0.9045	GABRG3
rs140679	rs140679	In_sample	15	25446271	C	0.2999	0.1681	0.8263	0.44	GABRG3
rs1979906	rs8053	Use_proxy	15	76628275	T	0.8052	0.8565	0.9814	0.7392	CHRNA5
rs680244	rs680244	In_sample	15	76658343	A	0.7785	0.9227	0.9086	0.2394	CHRNA5
rs621849	rs621849	In_sample	15	76659916	G	0.7514	0.8954	0.982	0.2406	CHRNA5
rs1051730	rs1051730	In_sample	15	76681394	T	0.2268	0.4858	0.7837	0.5422	CHRNA5
rs6495307	rs3743077	Use_proxy	15	76681951	A	0.6386	0.7045	0.9063	0.5149	CHRNA5
rs2235749	rs2235749	In_sample	20	1907939	A	0.6961	0.6801	0.5127	0.3808	PDYN
rs910080	rs910080	In_sample	20	1908226	G	0.637	0.9049	0.6398	0.4707	PDYN
rs10485703	rs10485703	In_sample	20	1908313	G	0.6939	0.3576	0.5716	0.3264	PDYN
rs6045819	rs6045819	In_sample	20	1909134	G	0.84	0.6871	0.6519	0.5345	PDYN
rs6045868	rs6045868	In_sample	20	1915278	A	0.6654	0.8556	0.8009	0.827	PDYN
rs10854244	rs6045912	Use_proxy	20	1922008	A	0.1127	0.466	0.5405	0.7815	PDYN
rs1997794	rs1997794	In_sample	20	1922858	C	0.2548	0.6547	0.4695	0.8328	PDYN

KEY:

p-value < 0.05

p-value < 0.10

P_COGA_ea COCA GWAS p-value from association in European American individuals

P_COGA_all COCA GWAS p-value from association in entire sample

P_SAGE_ea SAGE GWAS p-value from association in European American individuals

P_SAGE_all SAGE GWAS p-value from association in entire sample

Family history expanded results

ROC curve analysis based on presence or absence of family history

ROC curve analyses stratified by presence of family history of AD did not result in significant AUCs or differences in discriminatory accuracy of genetic sum scores in the positive or negative family history groups. Tables 3.7a and 3.7b show the AUC results and level of significance for the genetic sum scores in discriminating between cases and controls by presence or absence of family history in the COGA and SAGE GWAS datasets.

Table 3.7a COGA GWAS Sample

Family History	Classifier	AUC	p-value *	95% CI
No	24 SNPs (r^2 threshold = 0.50)	0.521	0.4	0.472, 0.571
	18 SNPs (r^2 threshold = 0.25)	0.517	0.493	0.468, 0.567
Yes	24 SNPs (r^2 threshold = 0.50)	0.454	0.109	0.4, 0.508
	18 SNPs (r^2 threshold = 0.25)	0.439	**0.036**	0.385, 0.494

*null hypothesis: AUC=0.50

Table 3.7b SAGE GWAS Sample

Family History	Classifier	AUC	p-value *	95% CI
No	24 SNPs (r^2 threshold = 0.50)	0.497	0.864	0.462, 0.535
	18 SNPs (r^2 threshold = 0.25)	0.492	0.648	0.457, 0.527
Yes	24 SNPs (r^2 threshold = 0.50)	0.520	0.553	0.453, 0.586
	18 SNPs (r^2 threshold = 0.25)	0.506	0.857	0.439, 0.573

*null hypothesis: AUC=0.50

A "yes" for family history in the SAGE GWAS sample means that either the participant's mother or father has a personal history of AD, by strict standard criteria and a "no" means that neither the participant's mother nor father has a personal history of AD by strict standard criteria.

Analyses exploring the combination of family history and genetic sum scores showed nominal, but not significant, improvements in discriminatory accuracy. For example, we found that the AUC for family history increased nominally from 0.686 to 0.690 in COGA after adding the candidate gene sum score pruned at $r^2<0.50$. Table 3.7 summarizes the AUC estimates for the original family history binary variable, the family history ordinal variables, and the family history variables plus the genetic sums scores. A test of the statistical difference for the family history alone vs. family history plus candidate gene sum score pruned at $r^2<0.50$ performed using DeLong's method for comparing ROC curves using the pROC package in R 2.10.2 showed that the difference between the two ROC curves was not significant ($z = 0.4508$, $p = 0.6521$).

Table 3.7 Summary of expanded family history analyses

Predictor	AUC	AUC FH+SCORE50	AUC FH+SCORE25
Results in the COGA GWAS sample independent of the family sample:			
SCORE50 = SNPs with r2 < 0.50	0.498	-	-
SCORE25 = SNPs with r2 < 0.25	0.491	-	-
Original FH binary variable "No": N = 536 "Yes": N = 545 (any FH) Total N = 1081	0.686	0.690	0.693
Ordinal FH var based on relaxed criteria "No": N = 850 (maternal and paternal Hx only) "1 parent": N = 147 "Both parents": N = 24 Total N = 1021	0.621	0.620	0.624
Ordinal FH var based on strict criteria "No": N = 850 "1 parent": N = 141 "Both parents": N = 22 Total N = 1013	0.618	0.617	0.621

Results in the SAGE sample independent of the COGA sample:			
SCORE50 = SNPs with r2 < 0.50	0.496	-	-
SCORE25 = SNPs with r2 < 0.25	0.492	-	-
Original FH binary variable "No": N = 1246 "Yes": N = 356 (based on relaxed criteria) Total N = 1602	0.614	0.614	0.618
Ordinal FH var based on relaxed criteria "No": N = 1246 "1 parent": N = 309 "Both parents": N = 47 Total N = 1602	0.617	0.616	0.620
Ordinal FH var based on strict criteria "No": N = 1246 "1 parent": N = 303 "Both parents": N = 30 Total N = 1579	0.614	0.615	0.617

The AUC estimates of the family history variables alone and family history (FH) with the addition of genetic sum scores are shown. SCORE50 refers to the genetic sum scores created based on variants pruned at $r^2 < 0.50$ and SCORE25 indicates a score based on variants pruned at $r^2 < 0.25$. The original AUC measures for the genetic sum scores are included in the table for comparison purposes.

Discussion

This study aimed to evaluate the clinical validity of genetic variants that have been associated with AD by exploring the aggregate effect of associated SNPs on risk prediction for AD. Prior studies on the clinical use of genetic information in predicting risk for other complex disorders have investigated the effect of genetic sum scores in risk assessment and shown significant, but small, AUCs. In our study, genetic sum scores were created based on results from SNPs that were associated with AD in family-based candidate gene studies. ROC curve analysis was used to assess the ability of the sum scores to classify cases and controls for AD. Results did not show significant AUCs for

the candidate gene sum scores, suggesting that these sum scores are not predicting better than chance. The individual variants contributing to the sum scores did not yield significant results in the independent samples in which discriminative ability was assessed. Because of the lack of replication for individual SNPs and sum score associations with AD, AUC estimates were not significant. Family history, on the other hand, did have significant discriminative ability for AD.

This assessment of discriminatory accuracy shows that these panels of SNPs currently have limited clinical validity. One reason that many of the candidate gene SNPs did not replicate in the independent samples used to assess for clinical validity could be due to heterogeneity across samples; different genetic variants may contribute to risk in different populations containing varying subsets of alcohol-dependent individuals. Therefore, genetic risk could be unique to the samples used in these association analyses. For example, several variants have been found to have stronger association with AD in individuals with co-occurring drug dependence. Dick et al. showed that *CHRM2* is associated with a form of AD that is comorbid with drug dependence, but not with AD alone (2007a). Individuals with this comorbidity were also found to have more severe alcohol problems. In another case, Foroud et al. found that SNPs in *TACR3* that were associated with AD in EA COGA families had the strongest association in individuals with more severe AD and comorbid cocaine dependence (2008). Furthermore, Agrawal et al. showed that *GABRA2* is associated with AD only in individuals with comorbid drug dependence. When these individuals were removed from the analysis, no association remained (2006). A next step in developing genetic risk models for AD would be to assess for prediction for different subtypes of AD. SNPs from primary analyses in the

family-based portion of the study may not have replicated in independent COGA and SAGE GWAS individuals due to sampling differences between the GWAS sample and the family-based association sample. One possibility is that the high-density family-based sample may be more severely affected than a case-control sample and therefore show differences in underlying genetic etiology. There was not a significant difference in mean DSM-IV symptom count for AD between the COGA high-density family-based sample, (mean=5.33, SD=1.82), and the SAGE (mean=4.87, SD=1.51) or COGA GWAS (mean=5.45, SD=1.42) samples. Although the difference in symptom count between the two samples was not significant, there is a nominal difference in symptom count; larger sample sizes may have more power to detect a difference between the two samples. Furthermore, severity of alcohol dependence may differ in ways beyond criterion count. For example, the severity of the individual symptoms themselves may differ between individuals with the same symptom count. This difference may manifest in ways beyond individual symptom count, such as the extent of tolerance and withdrawal, duration of symptoms, and number of episodes.

Many of the candidate gene SNPs used to compute the genetic sum score in the GWAS sample displayed allelic effects that were in the opposite direction. Prior literature has reported significant association in both directions for the same genetic variant in different samples. For some variants, the direction of effect for some loci could be different in different samples due to heterogeneity across samples. Differences in population structure may correspond to allele frequency differences across samples so that different variants are in LD with the causal variant in distinct samples (Zuo et al., 2012). Differences in phenotype between the samples may mean that alleles could

increase risk specifically for one phenotype in one sample, and not increase risk for a different phenotype in another sample. For example, the *GABRA2* gene has been associated with AD in different samples, but the allele conferring risk is different in different samples – in some, the major allele was associated with increased AD and in others, the minor homologous allele was associated. Further investigation suggests that the risk allele for *GABRA2* may vary across the studies due to differences in a co-occurring phenotype with AD. Trait anxiety, or harm avoidance based on the Tridimensional Personality Questionnaire (TPQ), has been suggested to have an influence on whether a *GABRA2* allele would increase risk for AD, with the major haplotype associated with AD in individuals with alcohol dependence who have high trait anxiety, and the minor haplotype associated with AD in individuals with low trait anxiety, and intermediate frequency haplotypes to be associated with unaffected status (Enoch, 2008; Enoch et al., 2006). In the COGA high-density family-based association sample, individuals with alcohol dependence have been shown to have higher trait anxiety than individuals without alcohol dependence (Ducci et al., 2007; Enoch, 2008).

These results show that family history is a better classifier than current conceptualizations of SNP panels, based on candidate genes for AD. Family history is currently likely a better predictor than this panel of SNPs because it accounts for more of the latent genetic factors contributing to AD, whereas the contribution to risk of the panel of SNPs is less clear. Family history also contains non-genetic predictors, which could account for a significant proportion of the risk as well, as family history could influence to some extent the environment that an individual is exposed to during development. Family studies show some evidence for influence of parental alcohol dependence on risk

for substance use disorders in children, or cultural transmission (Koopmans and Boomsma, 1996; Newlin et al., 2000). Furthermore, the etiology of AD may be different for one family versus another. Therefore, risk prediction based on one individual's family history may encompass genetic factors that are more specific to that individual than a general panel of SNPs, which may not explain risk for the particular subgroup to which that individual belongs. The nominal increase of the AUC after adding candidate gene SNP scores that are not correlated with family history to the family history variables suggests that variants associated with AD may provide additional risk information to family history alone.

Importantly, before assessment on clinical validity is made, the contribution of genetic sum scores, rather than individual associated SNPs, must be determined. Because variants contributing to AD have small effect sizes, and the outcome used in the association studies is a dichotomous diagnosis rather than a continuous outcome, larger sample sizes are needed for increased power to detect causal variants that replicate across studies (Bierut et al., 2010). GWA studies have shown replication of many of the SNPs associated with AD in the COGA candidate gene studies (Edenberg et al., 2010; Bierut et al., 2010), the results of which are shown in Table 3.6; however, in an effort to create SNPs that capture unique information by pruning them based on LD, some of the replicated SNPs were not included in the model. SNPs that were included represented ones with the lowest *p*-values from the family-based candidate gene association studies, and ones that are captured on the current GWAS arrays. To explore the effect that having more replicated variants, despite correlation between the variants, on a genetic sum score's predictive accuracy, we created an expanded candidate gene sum score without

pruning. This expanded score did not have a significant AUC (AUC = 0.493, p = 0.642); the AUC was not significantly different from the two candidate gene sum scores pruned at $r^2 < 0.50$ (AUC = 0.498 in COGA and 0.496 in SAGE) and $r^2 < 0.25$ (AUC = 0.491 in COGA and 0.492 in SAGE). Continued investigation of the genetics of AD will further refine the prediction model to include SNPs that have replicated and that capture unique associations.

In the COGA and SAGE GWAS samples, previous analyses have demonstrated that the missense SNP rs1229984 in the *ADH1B* gene encoding alcohol dehydrogenase was associated with AD at $p < 5x10^{-8}$ (Bierut et al., 2012). This variant, previously well-recognized for its protective influence on alcoholism in Asian populations, has also been found to exert an influence on alcoholism risk in European Americans and African Americans. However, it is fairly uncommon in non-Asian samples (<5%) and is poorly captured by content on commercially available GWAS platforms such as the Illumina platform used in the COGA and SAGE GWAS samples, due to lack of LD with neighboring SNPs. We assessed the discriminatory accuracy of this *ADH1B* SNP for AD and found that it alone has an AUC of 0.538 (p = 7.58 x 10^{-4}) in COGA. The inclusion of this SNP in the candidate gene sum score increased the AUC from 0.498 to 0.503, but this AUC was not significant, presumably partly due to the very low allele frequency in this population. This suggests that including known variants that replicate in the validation sample used for prediction could have a greater AUC, but a panel of several dozen SNPs may still include false positives in addition to true findings of small effect. Noise from null loci could outweigh effects from true loci in a small panel of SNPs,

which would decrease the predictive accuracy of the aggregate SNP panel. Expanding the panel to include additional replicated true variants could increase the AUC further.

The maximum AUC for a risk model containing only genetic variants is constrained by the heritability of the trait, as well as the disease prevalence in a population (Wray et al., 2010). As heritability of a disease goes down and as prevalence goes up, the maximum AUC goes down (Wray et al., 2010). This stresses the importance of taking into account other factors contributing to the variability in AD for risk prediction, particularly since AD is a fairly prevalent disorder. Additional measures to increase power may include reducing heterogeneity by refining the phenotype used as the outcome in the association study (Bierut et al., 2010). Large-scale meta-analyses, along with expanded individual association studies for AD, may improve the detection of disease variants.

We do not yet have enough information about the specific variants contributing to AD to use genetic data for clinical risk prediction. Family history is currently a better predictor of alcohol dependence, though a variant that was associated in the GWAS sample used for prediction was shown to have a significant AUC. This study suggests that expanding the number of replicated variants associated with AD would account for a greater portion of the genetic variance for AD and therefore improve risk prediction. Because AD also has a substantial unique environmental etiology in addition to genetic, a prediction tool based on genetic information alone would not have the highest AUC; the addition of environmental factors would account for more of the variability in AD and therefore a model that takes into consideration both could have better predictive ability. Data simulations that we have conducted show that adding environmental effects could

potentially raise the predictive accuracy to an AUC of 0.95 (Maher et al., in preparation). While genetic information may be of limited clinical validity at the moment, as we continue to identify genes successfully, and incorporate information from both genetic and environmental risk factors, there is potential for future clinical validity.

Chapter 4: Risk prediction using information from genome-wide association studies for AD

Abstract

Genome-wide association studies (GWAS) of alcohol dependence (AD) have reported numerous variants. The clinical validity of these genetic variants to discriminate between cases and controls for DSM-IV AD has not been reported. The Collaborative Study on the Genetics of Alcoholism (COGA) and the Study of Addiction: Genes and Environment (SAGE) GWAS samples were used to examine the aggregate impact of multiple genetic variants with small effect sizes on clinical risk prediction for AD using receiver operating characteristic (ROC) curve analysis. In these analyses, subsets of the COGA and SAGE samples were used as gene discovery and validation samples in two sets of analyses, in which genetic sum scores were created by adding risk alleles of associated SNPs in discovery samples and then assessed for their ability to discriminate between cases and controls in independent validation samples. SNPs from GWAS analysis that met nominal association levels in two discovery subsets and SNPs from GWAS analysis that met varying "significance" criteria based on *p*-value thresholds from 0.0001 to 0.5 were assessed separately for predictive accuracy. ROC curve analysis using scores created from semi-replicated SNPs did not result in significant discriminatory ability for the genetic sum scores, suggesting that the SNPs are not predicting better than chance. SNPs

that met less stringent *p*-value thresholds of 0.01 to 0.50 in GWAS analyses did yield significant area under the ROC curve (AUC) estimates, ranging from mean AUC estimates of 0.549 for SNPs with $p < 0.01$ to 0.565 for SNPs with $p < 0.50$. This study shows that these SNPs from GWAS analyses account for some of the risk in AD, but have limited clinical validity. This illustrates the need for further development of prediction panels that incorporate replicated variants contributing to risk for AD.

Introduction

A number of genome-wide association studies have been performed for alcohol dependence (AD) and alcohol-related phenotypes. (Treutlein and Rietschel, 2011b; Treutlein et al., 2009; Frank et al., 2012; Bierut et al., 2010; Edenberg et al., 2010; Agrawal et al., 2011; Heath et al., 2011; Lind et al., 2010; Schumann et al., 2011; Wang et al., 2012; Wang et al., 2011; Zuo et al., 2012; Kendler et al., 2011; Zuo et al., 2011). Genome-wide associations studies (GWAS) have the benefit of a hypothesis-free approach to finding variants associated with common diseases without prior information on putative chromosomal regions or genes. The studies could provide coverage of common markers across the genome based on correlation due to linkage disequilibrium (LD) between loci (Visscher et al., 2012). Prior to the technical feasibility of the GWAS era, Risch and Merikangas projected that using association studies to study common variants that contribute to common diseases would be more powerful and require fewer markers and sample sizes to detect small effects than using linkage studies, which are more suited to detecting loci with larger effects sizes (Risch and Merikangas, 1996).

For alcohol dependence, many of the reported genome-wide association studies to date have reported variants that did not meet the genome-wide significance threshold of $p < 5x10^{8}$ based on Bonferroni correction for one million tests for a GWAS using one million markers, though many have shown variants that were associated with low *p*-values ($p < 1x10^{-5}$). Two studies have reported genome-wide significant findings for variants that have been replicated in other samples (Zuo et al., 2012; Lind et al., 2010). The results of GWA studies for alcohol dependence have supported previous candidate

gene associations and implicated many new genes and pathways in risk for alcohol-related phenotypes.

Among the few studies that have reported genome-wide significant findings, two studies found significant associations of the alcohol dehydrogenase (*ADH)* genes at this threshold. Frank et al. (2011) found rs1789891, located between the *ADH1B* and *ADH1C* genes, to be associated with AD in a treatment-based sample of 1333 male individuals with severe AD and 2168 controls, all of German descent. Bierut et al. (2012) reported genome-wide support for the low-frequency rs1229984 SNP in the *ADH1B* gene in the SAGE GWAS sample, totaling 2298 individuals with AD and 3334 controls without dependence, and including individuals of both European and African ancestry. This study provided new support for association of the *ADH1B* variant that was previously limited to individuals of East Asian descent.

In the first reported GWAS of AD, Treutlein et al. (Treutlein et al., 2009) performed a case-control study in which cases were recruited from treatment centers in Germany. They further assessed their top results in a follow-up sample and found evidence for two correlated SNPs in the 3' flanking region of the peroxisomal trans-2-enoyl-CoA reductase gene (*PECR)* located in the 2q35 region. Regions on chromosome 2q have previously been implicated in linkage studies of alcohol-related phenotypes (Schuckit et al., 2001; Nurnberger et al., 2001; Dick et al., 2010). Thus, this finding provided additional support for possible involvement of genes in this region for alcohol dependence. Primary analyses in the COGA GWAS sample did not result in SNPs that met genome-wide significance, but convergent evidence from the case-control GWAS sample, COGA family-based association sample, and gene expression analyses supported

association of a group of chromosome 11 genes (*SLC22A18, PHLDA2, NAP1L4, SNORA54, CARS*, and *OSBPL5*), particularly for early-onset AD (Edenberg et al., 2010). The primary SAGE GWAS did not result in genome-wide significant association with AD, but showed modest replication for the previously implicated *GABRA2* gene; all *GABRA2* SNPs were nominally associated at $p < 0.05$ in the SAGE sample (Bierut et al., 2010). In meta-analysis of an Australian and Dutch sample, the top hit was the semaphorin 3E gene (*SEMA3E),* which is involved in synaptic specificity of motor circuits in mice. Gene network analyses revealed evidence for ion channel and cell adhesion molecule genes in this study (Lind et al., 2010). Lind et al. also found SNPs that met genome-wide significance for association with comorbid alcohol/nicotine dependence in *MARK1,* which is involved in phosphorylation of microtubule-associated proteins, *DDX6,* which encodes a putative RNA helicase, and *KIAA1409*, which is thought to be part of a sodium channel complex. The *KIAA0040* gene was associated with AD in both Zuo et al.'s study (2012) and Wang et al.'s meta-analysis (2011). Wang et al. also found an association between AD and *THSD7B* and *NRD1,* and found replication of *PKNOX2*. Studies of quantitative traits such as alcohol consumption have identified a genome-wide significant association with the *AUTS2* gene (Schumann et al., 2011) and evidence of association for the *TMEM108* and *ANKS1A* genes (Heath et al., 2011). In a study of an alcohol factor score based on DSM-like symptoms, Kendler et al. (2011) found the most significant SNP to be *KCNMA1, AKAP9,* and *PIGG* in the EA sample and CEACAM6, KCNQ5, SLC35B4, and MGLL in the AA sample. They also found support for previously associated candidate genes for *ADH1C, NFKB1,* and *ANKK1* in the EA sample and *ADH5, POMC*, and *CHRM2* in the AA sample (Kendler et al., 2011).

ROC curve analyses of prior complex diseases have shown modest predictive ability of genetic sum scores (Jostins and Barrett, 2011). Several studies have shown that when a greater number of genetic variants meeting more liberal *p*-value thresholds were included in an aggregate genetic sum score, then the score accounted for more of the variability in phenotype compared with a score consisting of fewer variants meeting more stringent *p*-value thresholds. Purcell et al. showed in the International Schizophrenia Consortium sample of 3,322 cases with schizophrenia and 3,587 controls that thousands of variants associated at less stringent *p*-value thresholds of $p < 0.10$, $p < 0.20$, $p < 0.30$, $p < 0.40$, and $p < 0.50$, accounted for more of the variance in schizophrenia. Of the different *p*-value threshold scores, the scores created from SNPs meeting increasingly large *p*-value thresholds accounted for more of the variance in schizophrenia. A threshold of $p < 0.50$ explained the most phenotypic variance – about 3% – in schizophrenia in an independent target sample (Purcell et al., 2009). This corresponds to an AUC of 0.65 in discriminating case-control status for schizophrenia (Jostins and Barrett, 2011). The score was also found to explain 1-2% of the variance in bipolar disorder, but did not account for a significant proportion of phenotypic variance in non-psychiatric disorders, supporting a shared polygenic component between schizophrenia and bipolar disorder (Purcell et al., 2009). Genetic risk profile studies for depression and anxiety showed similar polygenic models. Demirkan et al. (2011) used results from the Genetic Association Information Network (GAIN) major depressive disorder (MDD) GWA study to select SNPs meeting varying *p*-value thresholds ranging from $p < 0.00001$ with incremental threshold changes including $p < 0.0001$, 0.001, 0.01, 0.1, and 0.2, up until $p < 1.0$. They created genetic sum scores using these SNPs and tested them for association

with MDD in independent samples. Risk scores associated with MDD at $p < 0.1$ to $p < 0.4$ in the GAIN-MDD discovery sample explained significantly 0.65% of the variance in MDD in an independent sample. They found that risk scores created based on SNPs meeting $p<0.1$ to $p<1.0$ in the GAIN-MDD discovery sample explained 1-2.1% of the variance in anxiety in an independent sample, with increasing variance explained with each incremental increase in the *p*-value threshold used to select the discovery SNPs (Demirkan et al., 2011).

The method of selecting variants based on less stringent thresholds has also been used previously to assess risk prediction for the disease cohorts in the Wellcome Trust Case Control Consortium (WTCCC), many of which had previously associated variants (Evans et al., 2009). Evans et al. examined the predictive ability of genome-wide information for the 7 common diseases in the WTCCC: bipolar disorder, coronary heart disease, hypertension, Crohn's disease, rheumatoid arthritis, type I diabetes, and type II diabetes. Because the effect sizes of alleles contributing to these complex diseases are often in the range of 1.1-2, and are therefore likely to have individually small effects on prediction, they created genetic sum scores composed of many SNPs at liberal significance thresholds. They found that many of the genome-wide scores produced significant AUCs, with an AUC of 0.549 for bipolar disorder to an AUC of 0.784 for type I diabetes. AUCs were highest for disorders with known genetic regions of larger effect, such as the involvement of variants in the MHC region in type I diabetes and Crohn's disease. Genome-wide scores added to the discriminatory accuracy of known variants, particularly for diseases in which the effects of known variants were smaller. Scores based on SNPs meeting more liberal thresholds had the best discriminatory accuracy for

disease that did not have known large-effect loci, such as bipolar disorder, coronary heart disease, hypertension, and type II diabetes. For some conditions, the AUC peaked at SNPs selected at $p < 0.50$, and for others, the AUC was higher at $p < 0.80$. This suggests that more liberal thresholds capture more of the polygenic effects, but as *p*-value thresholds continue to increase, the polygenic scores may be too diluted by null effects to have increasing discriminatory accuracy (Evans et al., 2009).

These studies show that although effect sizes of common SNPs are individually too small to meet genome-wide significance thresholds, selecting SNPs at more liberal *p*-value thresholds would include a greater proportion of true loci that could in aggregate account for variance in a complex polygenic trait, despite noise from null loci. Most of the genetic variants contributing to AD have small effect sizes. A genetic sum score composed of many genetic variants at liberal thresholds would provide an aggregate predictor without necessitating knowledge of specific true loci. This study explored the cumulative impact of multiple genetic variants with small effect sizes from genome-wide association studies, with a focus on risk prediction, in order to provide a clinical assessment of genetic contributions from GWAS data to AD.

We used results from the Collaborative Study on the Genetics of Alcoholism (COGA) and the Study of Addiction: Genes and Addiction (SAGE) GWAS samples to capture genetic effects on alcohol dependence in order to predict risk in independent sample subsets. We first created genetic sum scores created based on semi-replicated variants that met nominal *p*-value thresholds in two separate halves of the SAGE sample, with the idea that SNPs that had replicated may be more likely to represent "true positives" and enhance predictive ability of genetic sum scores composed of these replicated SNPs. We

then assessed genetic sum scores based on variants that met varying *p*-value thresholds. We combined the COGA and SAGE GWAS samples and then split the combined sample randomly in half. One half of the combined sample was used as a discovery sample and the other half a validation sample. We created genetic sum scores based on SNPs that met a range of *p*-value thresholds from $p < 0.0001$ to $p < 0.50$ in the discovery sample and then assessed for prediction in the validation sample, testing that using SNPs meeting less stringent *p*-value thresholds may better detect variants of small effect.

Materials and methods

Sample and measures

COGA GWAS sample

The Collaborative Study on the Genetics of Alcoholism (COGA) is a large-scale multi-center family study developed to identify genes that contribute to alcohol-related outcomes. The sample consists of families containing probands meeting both DSM-IIIR and Feighner criteria for alcohol dependence who were ascertained from outpatient and inpatient alcohol treatment centers at six sites across the United States. Families reported information about family history and were interviewed using a poly-diagnostic instrument, the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA), which assesses Feighner, DSM-IIIR, DSM-IV, and ICD-10 criteria. A case-control sample of 1945 phenotyped subjects was formed by COGA for a genome-wide association study (GWAS). Cases had a lifetime diagnosis of alcohol dependence by DSM-IV criteria. Controls reported consuming alcohol but did not have a diagnosis of alcohol dependence or abuse by any of the diagnostic criteria assessed by SSAGA

(Feighner, DSM-IIIR, DSM-IV, and ICD-10) or DSM-IIIR or DSM-IV criteria for cocaine, marijuana, opioids, sedatives, or stimulants. They could not share a known common ancestor with a case and were preferentially selected to be above the age of 25 years.

Genotyping was completed using the Illumina Human 1M DNA Analysis BeadChip at the Center for Inherited Disease Research (CIDR). DNA was extracted from blood and lympoblastoid cell lines. The Ilumina Infinium II assay protocol was followed with hybridization to Illumina Human 1M BeadChips (Illumina, San Diego, CA). A total of 1,069,796 SNPs were used, with a mean spacing of 2.4 kb (Edenberg et al., 2010). The dataset had a total of 1,041,465 SNPs with genotypes that had Gencall quality scores of 0.15 or higher. Samples with genotypes for less than 98% of SNPs were removed. All samples were screened for cryptic relatedness and population stratification. Principal components analysis clustered samples along HapMap reference panels. Principal components-derived covariates were created to separate the sample into individuals of European American and African American descent. Additional details about the quality control steps taken to process the genotypic information in the COGA dataset can be found in (Edenberg et al., 2010).

SAGE GWAS sample

The Study of Addiction: Genes and Environment (SAGE) is part of the Gene Environment Association Studies (GENEVA) initiative of the NHGRI to identify genetic contributions to addiction through large-scale genome-wide association studies of cases and controls. The SAGE sample consists of 4,121 cases and unrelated controls from

subsets of three large studies on addiction: the Family Study of Cocaine Dependence (FSCD), the Collaborative Genetic Study of Nicotine Dependence (COGEND), and COGA. Individuals from FSCD with cocaine dependence were recruited from treatment units for chemical dependency in the St. Louis metropolitan area. Age, race, sex, and residency-matched controls were recruited through the community-based Missouri Family Registry. The COGEND study is a community-based study of participants recruited in St. Louis, MO and Detroit, MI. Although the SAGE sample consists of three samples that were ascertained differently, cases in SAGE are defined as having DSM-IV lifetime diagnosis of alcohol dependence, and all participants in SAGE were assessed using the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA). Controls were exposed to alcohol. Some controls met criteria for nicotine dependence based on the Fagerström Test for nicotine dependence, but none met criteria for a DSM-IV lifetime dependence diagnosis for alcohol, marijuana, cocaine, opiates or other drug.

Parallel to the COGA GWAS, genotyping was completed using the Illumina Human 1Mv1_C DNA Analysis BeadChip and the Ilumina Infinium II assay protocol (Illumina, San Diego, USA) at the Center for Inherited Disease Research (CIDR). The quality control process involved checking for Mendelian errors, batch effects, cryptic relatedness, potential chromosomal anomalies, and deviation from Hardy Weinberg equilibrium. Additional details about the sample can be found in the primary SAGE GWAS manuscript by Bierut et al. (Bierut et al., 2010).

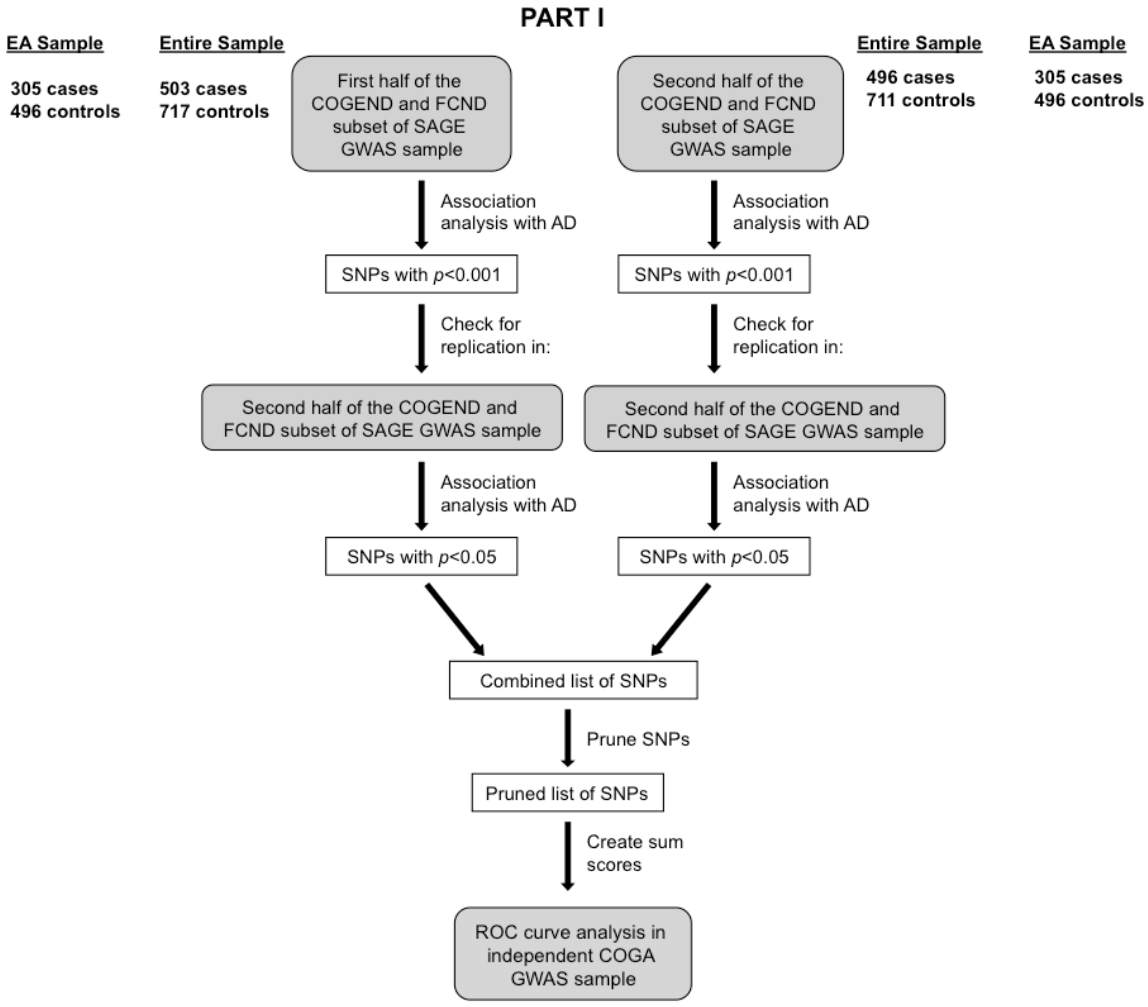
Data analysis

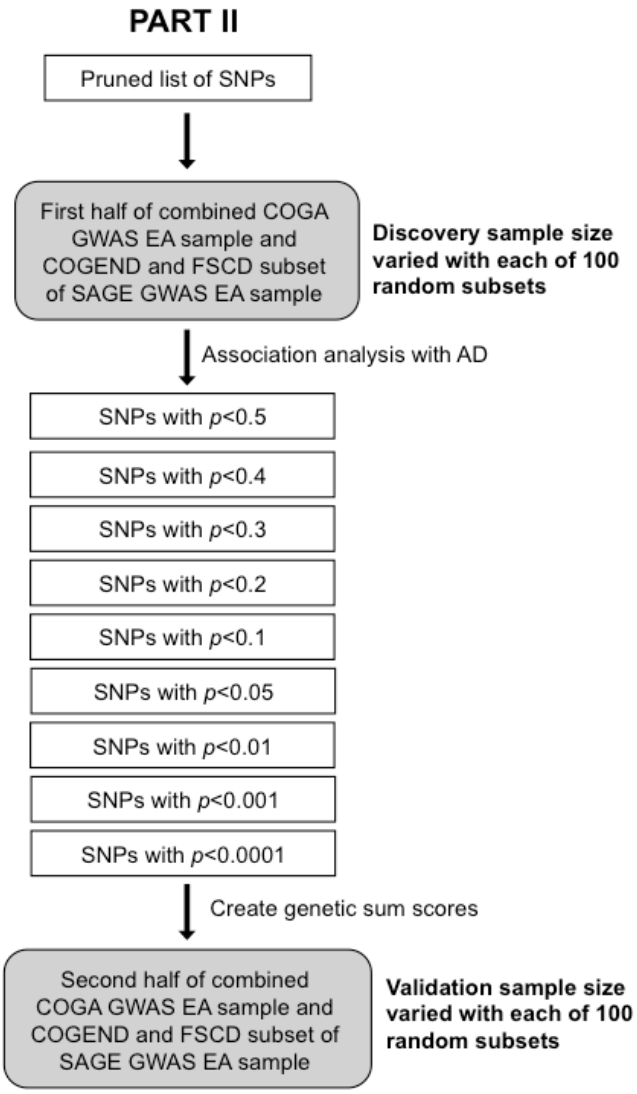
Analyses overview

Figure 4.1 illustrates the overview of the analyses, which were broken down into two parts:

Part I, which evaluated panels of SNPs that have been replicated in two samples, and Part II, which assessed the discriminatory accuracy of SNPs meeting varying significance thresholds in genome-wide association analyses.

Figure 4.1 Study overview





Part I. SAGE GWAS SNP panel with replication

Sample selection and association analyses

In order to select an independent gene-finding sample, the FSCD and COGEND subset of the SAGE GWAS sample without COGA individuals was selected for this part of the study. This selected subset was split randomly in half. Association analysis was performed in each half of the sample subset using logistic regression with covariates for sex, age, and the FSCD and COGEND study site variables.

Because removing COGA individuals and then splitting the sample in half substantially reduced the SAGE sample size, all analyses were completed in the entire sample including individuals of different ancestries, as well as in the subset of European American individuals. For association analysis in the entire sample, molecularly derived principal components factors for ethnicity, PC1 and PC2, were added as covariates. Figure 4.1 summarizes the number of individuals included in the analyses separately for the entire sample and for the EA sample. The entire SAGE sample consisted of 2484 individuals (1450 controls and 1034 cases). Of the entire SAGE GWAS sample, 1425 individuals who were part of COGA were removed from the entire SAGE sample. The remaining FSCD and COGEND subset of the entire SAGE sample consisted of 1220 individuals for one half of the subset and 1207 individuals for the other half of the subset. The SAGE GWAS European American sample contained 801 individuals for each half of the subset after removing COGA individuals and then performing a 50% split.

SNP selection and pruning

SNPs that met a *p*-value threshold of $p < 0.001$ in the first half of the SAGE sample and a threshold of $p < 0.05$ in the second half of the sample, and vice versa, were selected for further analysis (Figure 4.1). The direction of effect of the minor allele on alcohol dependence risk was matched for SNPs meeting both *p*-value thresholds. Because a nominal *p*-value threshold of $p < 0.001$ was used in the discovery subsample, many of the SNPs meeting this threshold may be false positive associations. It can therefore be expected that the direction of an associated null allele may flip in the second SAGE subsample. SNPs that share the direction of effect across both SAGE subsamples are

more likely to be true findings. SNPs that did not have the same direction of effect in the two halves of the sample were eliminated from the combined list of semi-replicated SNPs. The combined list of SNPs that showed association and had the same direction of effect across both halves of the discovery sample was then pruned based on an r^2 threshold of 0.50 using an LD-based pruning function in PLINK version 1.07 (Purcell et al., 2007). This method calculated pairwise genotypic correlations for the list of SNPs. One of each pair of SNPs with correlations greater than an r^2 of 0.50 was removed. Because LD estimates are more accurate in larger samples, LD calculations for SNP pruning were performed in the complete SAGE GWAS sample, including individuals from COGA.

Genetic sum scores and ROC curve analyses

Additive genetic sum scores of risk alleles were created in the COGA GWAS sample based on pruned SNPs from SAGE GWAS results. Because the odds ratios varied across the SAGE discovery samples, risk alleles were not weighted. The genetic sum score was then used to classify case-control status in individuals from the COGA GWAS sample using ROC curve analyses. Association between alcohol dependence and the genetic sum scores and individual pruned SNPs was performed in the COGA sample.

Part II. GWAS results from varying *p*-value thresholds

Sample selection

In Part II of the analyses, a combined GWAS sample was created by merging the COGA and SAGE GWAS samples. This combined COGA-SAGE GWAS sample was then split

into discovery and validation subsamples. The discovery and validation subsets of the sample were created based on an initial combined sample in an attempt to reduce heterogeneity across the discovery and validation samples compared with using discovery and validation samples gathered using different ascertainment procedures. In order to control for differences in association between African American and European American subjects, analysis for this part of the study was performed only in the European American subset. The FSCD and COGEND subset of the SAGE EA sample was combined with the COGA GWAS EA sample. Because the COGA GWAS sample contains individuals who are not part of the COGA subset of the entire SAGE sample, combining the COGA GWAS sample with the nicotine and cocaine studies created a larger GWAS sample than the SAGE GWAS sample alone. This combined sample allowed for more power when it was split in half into discovery and validation samples. Controls who endorsed 3 or more symptoms for DSM-IV AD, but did not cluster within a 12-month period, were removed from the combined sample, as these individuals could still represent increased genetic risk for alcohol dependence (N = 49, all from the SAGE GWAS sample). The combined sample included 2951 individuals, comprised of 1495 cases and 1456 controls (Table 4.1).

Table 4.1 Summary of cases and controls by study for combined COGA and SAGE GWAS sample

	DSM-IV AD		
Study	Controls	Cases	Total
COGA	552	846	1398
COGEND	702	335	1037
FSCD	241	275	516
Total	1495	1456	2951

The FSCD and COGEND subset of the SAGE EA sample and COGA GWAS EA combined sample was split randomly in half so that each half contained 50% of cases and 50% of controls. In order to account for chance effects, repeated random sub-sampling cross-validation was implemented by performing this subsetting procedure100 times to obtain 100 subsamples in which analyses were completed.

SNP pruning

Before logistic regression analyses were performed, the LD-based pruning function in PLINK version 1.07 was used to prune the 1,041,983 SNPs genotyped in the combined sample. The SNPs were pruned at $r^2 < 0.50$ using a sliding window of 50 base pairs shifted by 5 base pairs following each pruning step. Pruning resulted in 385,060 of the original SNPs pruned in, which represented 36.95% of all SNPs in the combined sample.

Association analyses

Association was performed in PLINK version 1.07 using logistic regression under an additive model with sex and a dummy-coded site covariate distinguishing between the COGA, FSCD, and COGEND study sites. Figure 4.1 lists the *p*-value thresholds used to select SNPs from association results in the first half of the sample.

Genetic sum scores and ROC curve analyses

Because both the COGA and SAGE GWAS samples had the same SNPs genotyped, and were confirmed to share the direction of the genotyped strand, GWAS results were

matched directly by allele. Genetic sum scores were created for autosomal SNPs. Scores of total allele count were weighted by the natural log of the odds ratio for each reference minor allele, and then divided by the number of non-missing genotypes for each individual using PLINK version 1.07:

$$\text{Genetic sum score} = \frac{\sum[x_i \times \ln(OR_i)]}{N} \frac{\sum[x_i \times \ln(OR_i)]}{N}$$

where $x_i x_i$ is the number of reference alleles at the *i*th SNP, *OR* is the corresponding odds ratio, and *N* is the number of non-missing genotypes for each individual.

Discriminatory accuracy of genetic sum scores was measured using ROC curve analysis in the caTools package (Tuszynski, 2011) in R version 2.12.2 (R Development Core Team, 2011). The *p*-values associated with the AUCs for these sum scores were calculated based on the Wilcoxon rank-sum test using R version 2.12.2. Following completion of the 100 iterations of the subsampling procedure, the mean of the AUC estimates and confidence intervals of mean estimates were calculated using SPSS/PASW v17.0 (SPSS Inc., Chicago IL).

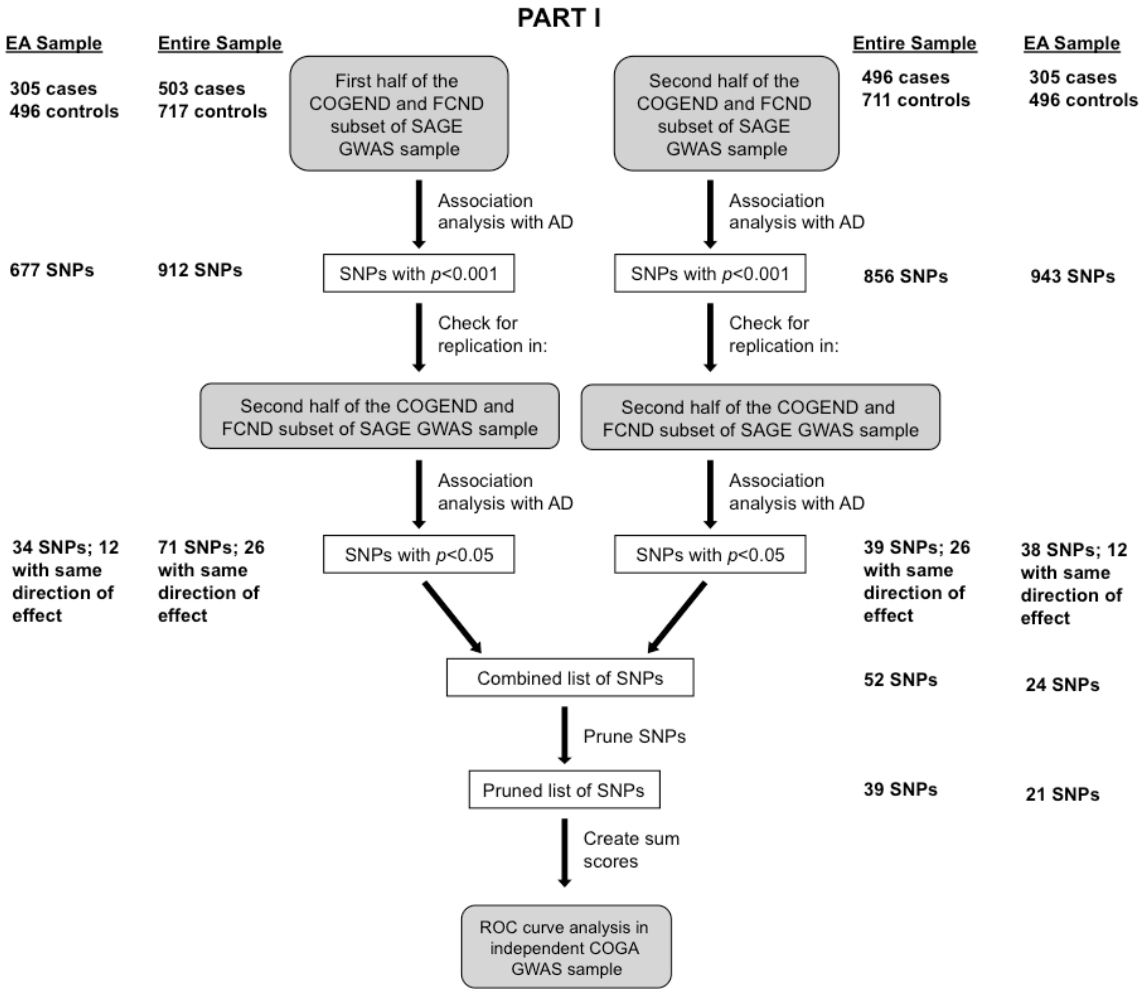
Results

Figure 4.2 summarizes the resultant number of SNPs at each step of analysis for both the EA sample and the entire sample. GWAS analyses of the entire SAGE sample including all individuals across ethnicities in Part I of the study resulted in 52 SNPs that met significance criteria at $p < 0.001$ in one discovery sample and $p < 0.05$ in the second discovery sample and had the same direction of effect in both discovery samples. After

LD-based pruning at $r^2 < 0.50$, 39 SNPs remained. ROC curve analysis showed that the genetic sum score created based on these 39 SNPs did not have significant discriminatory accuracy in the COGA GWAS sample (Table 4.2). Association analysis of this genetic sum score with alcohol dependence in the entire COGA sample using logistic regression did not result in a significant association ($p = 0.206$). One SNP out of the 39 used for prediction had a p-value < 0.05 for association with alcohol dependence: rs1950231 on chromosome 14 ($p = 0.0496$).

Analyses in the EA subset of the SAGE GWAS sample resulted in 24 SNPs that met significance criteria and shared direction of effect in both of the SAGE discovery sample subsets. After pruning, 21 SNPs remained. The genetic sum score created using these 21 SNPs was not a significant classifier for case-control status in the COGA GWAS sample. The AUC estimate for the sum score is shown in Table 4.2. The test of association of the genetic sum score with alcohol dependence in European-American subset of the COGA sample using logistic regression was not significant ($p = 0.176$). None of the 21 SNPs in the panel used for prediction in the European-American subset of the sample was associated with alcohol dependence at $p < 0.05$.

Figure 4.2 Number of SNPs resulting from GWAS analyses with semi-replicated SNPs



Gray boxes show samples used for each step of analyses. White boxes display the selection criteria for SNPs at each step. The number of SNPs resulting from each step of analysis is shown in separate columns for the EA sample and for the entire SAGE GWAS sample.

Table 4.2 ROC curve analysis results of semi-replicated SNPs from GWAS analyses

Diagnostic Classifier	AUC	Std. Error [a]	Asymptotic Sig. [b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
39 SNPs in all individuals [c]	0.521	0.014	0.126	0.494	0.548
21 SNPs in EA individuals [d]	0.520	0.016	0.203	0.489	0.551

a. Under the nonparametric assumption
b. Null hypothesis: true area = 0.5

c. Genetic sum score created using GWAS results in the entire sample including European American, African American, and other ancestries
d. Genetic sum score created using GWAS results in the European American sample

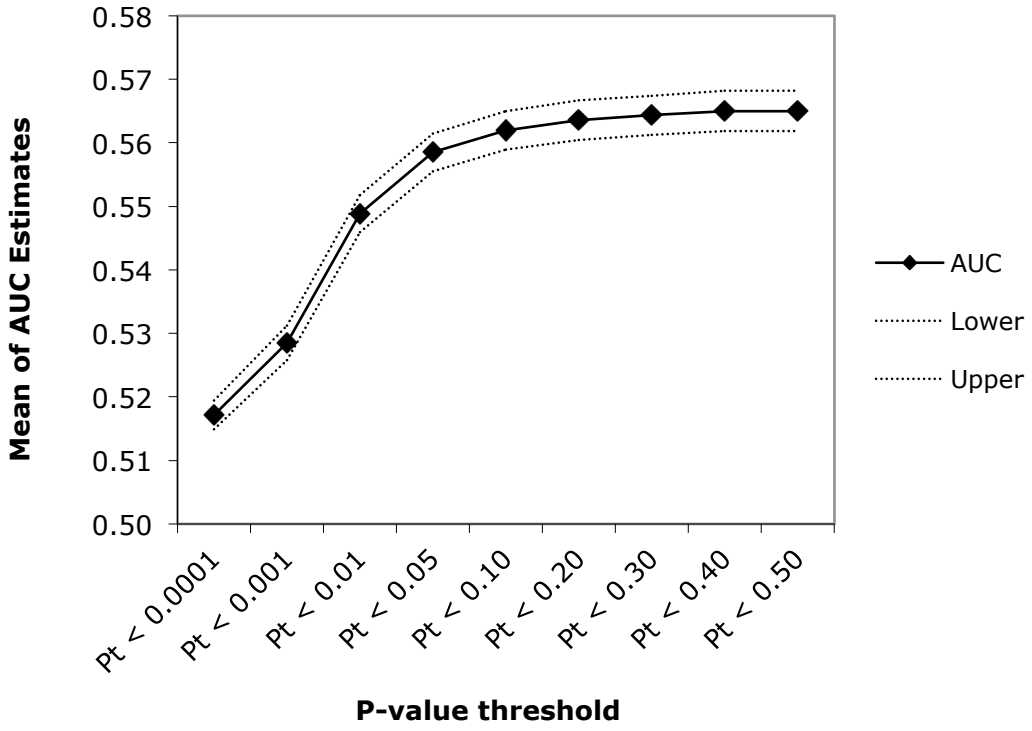
In Part II of the study assessing SNPs that met varying *p*-value thresholds, SNPs meeting increasingly stringent significance thresholds of $p < 0.001$ and lower did not have significant AUCs in ROC curve analyses. Subsets of SNPs meeting more liberal *p*-value thresholds of $p < 0.01$ and greater had AUC estimates that were significant ($p < 0.05$ for AUC). Table 4.5 summarizes mean AUC estimates for each set of SNPs meeting *p*-value thresholds across the 100 random divisions of the SAGE-COGA combined sample. The median of the *p*-values associated with each AUC estimate was determined (Table 4.5) because the distribution of these *p*-values was significantly skewed across the subsets of SNPs. Figure 4.3 illustrates the AUC estimates of genetic sum scores created based on varying *p*-value thresholds. Although the significance threshold at which the AUC value peaked varied across each random sample subset, AUC point estimates showed an increasing trend as the *p*-value threshold used for SNP selection became less stringent.

Table 4.5 Results of SNP subsets from varying *P*-value thresholds

P-value threshold score	N subsets	Mean AUC	95% Confidence Interval		Median *p*-value for AUC
			Lower	Upper	
P-value < 0.50	100	0.565	0.562	0.568	1.37E-05
P-value < 0.40	100	0.565	0.562	0.568	1.42E-05
P-value < 0.30	100	0.564	0.561	0.567	1.82E-05
P-value < 0.20	100	0.564	0.561	0.567	2.62E-05
P-value < 0.10	100	0.562	0.559	0.565	4.81E-05
P-value < 0.05	100	0.559	0.556	0.562	1.04E-04
P-value < 0.01	100	0.549	0.546	0.552	0.00167
P-value < 0.001	100	0.528	0.526	0.531	0.0632
P-value < 0.0001	100	0.517	0.515	0.519	0.291

Summary statistics for 100 random 50% splits of the combined COGA-SAGE sample into discovery samples and validation samples. Sum scores were created based on SNPs meeting each *p*-value threshold, by adding minor alleles weighted by the log of the odds ratio for AD. Confidence intervals are based on 100 AUC estimates from 100 separate sum score calculations at each *p*-value threshold. Median *p*-value threshold was calculated because distributions of *p*-values were skewed.

Figure 4.3 Mean AUC estimates for varying P-value thresholds



The mean of all 100 AUC estimates for sum scores created using SNPs that meet different *p*-value thresholds in discovery samples is plotted here in the solid line. Dashed lines represent the upper and lower bounds of the 95% confidence intervals of the mean of the AUC estimates.

Discussion

This study aimed to evaluate the clinical validity of genetic variants that have been associated with alcohol dependence by exploring the aggregate effect of associated SNPs on risk prediction for alcohol dependence. Prior studies on the clinical use of genetic information in predicting risk for other complex disorders have investigated the effect of genetic sum scores in risk assessment and shown significant, but small, AUCs. In our study, genetic sum scores were created based on results from two sources of genome-wide association results: SNPs from a semi-replicated list of variants that were associated with alcohol dependence in two "separate" GWAS samples and SNPs that met varying *p*-value significance thresholds in GWAS analyses. ROC curve analysis was used to assess the ability of the sum scores to classify cases and controls for alcohol dependence. The scores created based on semi-replicated SNPs at nominal *p*-value thresholds of $p < 0.001$ and $p < 0.05$ in the two separate discovery and replication samples in Part I of the study did not show significant AUCs or significant association with alcohol dependence in the independent clinical validation sample. Results from Part II of the study showed significant, albeit small, AUC estimates for sum scores based on SNPs that met *p*-value thresholds ranging from $p < 0.10$ to $p < 0.50$. Significant AUC estimates were under 0.60.

These results support a polygenic model involving hundreds of variants of small effect contributing to risk for AD that are consistent with other findings on alcohol phenotypes and other complex traits (Heath et al., 2011; Purcell et al., 2009; Frank et al., 2012). Less stringent thresholds allowed for the selection of more true findings with

effect sizes that would not otherwise have reached genome-wide significance. Combining nominally associated SNPs in aggregate improved clinical validity because these true loci could outweigh noise from null loci.

In Part I of the study, we created discovery and replication samples by splitting just the FSCD and COGEND portion of the SAGE GWAS sample in half, and then assessed for clinical validity in the COGA GWAS sample. This selection method studied the COGA and SAGE GWAS samples as distinct populations in order to find more variants that are associated in samples that are ascertained differently. Variants that replicated across the discovery sample, FSCD/COGEND, and the validation sample, COGA, would possibly contribute to AD risk in more general populations than variants that replicated in samples with similar population structures, such as those that replicated in both halves of SAGE. In Part II of the study, we combined the COGA and SAGE samples before performing subsampling to create samples with similar population structure across discovery and validation sets in order to address heterogeneity across samples. Of the list of SNPs in Part I of the study that met nominal significance criteria in both halves of the SAGE sample, the majority of SNPs did not share the same direction of effect suggesting that many of these results could be false positives. It is expected that many SNPs meeting the nominal *p*-values would represent type I error, particularly given the high number of tests performed in GWAS analyses. The replication step was an attempt to filter out SNPs that had opposite directions of effect in order to retain a greater proportion of SNPs that could be true positives.

The finding that genetic sum scores created from SNPs meeting less stringent *p*-value thresholds were significantly associated with AD and had significant discriminative

ability suggests that varying p-value thresholds could better detect variants of small effect. The samples used in this study did not have enough power to detect the entire range of small effect sizes for individual variants assessed in these analyses at a genome-wide significance level. Splitting the COGA-SAGE combined sample further reduced power. The numbers of loci meeting each *p*-value threshold were close to what would be expected by chance when selecting lists of SNPs at each threshold. Therefore, a genetic sum score created based on these thresholds would encompass SNPs that may not contribute to risk for AD. For these false positives, weighting by the log of the OR obtained from logistic regression in the discovery samples for these SNPs could in fact be weighting by the opposite direction of effect that some of the SNPs have in the validation sample. This in turn would decrease the association between the genetic sum score and alcohol dependence in the validation sample, and therefore the AUC. As sample sizes increase for studies of alcohol dependence, and as meta-analyses combine results across all genome-wide association studies of AD, a more precise odds ratio could be obtained and more true loci may be found.

The polygenic nature of the AD indicates a spectrum of allele frequencies contributing to AD. Larger sample sizes are necessary for detecting smaller effects without including null markers at the same significance thresholds (Park et al., 2010) . This would allow the creation of genetic sum scores diluted by fewer null effects. The markers used in current GWAS platforms are common variants with minor allele frequencies greater than 1% that also capture multiple variants in LD with the variants directly genotyped on the SNP chip. There is evidence for alleles associated with AD with low frequency not captured on the GWAS platforms that still have a significant

effect on AD in the population. A previous report in the COGA and SAGE GWAS samples demonstrated that the Arg48His variant, rs1229984 in the *ADH1B* gene encoding alcohol dehydrogenase was associated with AD at $p < 5x10^{-8}$ with a relatively large effect size (Bierut et al., 2012). In this study, a meta-analysis across COGA, FSCD, and COGEND showed that the allele encoding His48 had a significantly protective effect on alcohol dependence (OR = 0.34, $P = 6.6x10^{-9}$). This variant was previously well-recognized for its protective influence on alcohol dependence in Asian populations, but had low frequency in European Americans (MAF = 3-4% in the COGA and SAGE GWAS EA samples) and African Americans (MAF = 1-2% in COGA and SAGE GWAS AA samples). It is poorly captured by commercially available GWAS platforms such as the Illumina platform used in the COGA and SAGE GWAS samples, due to lack of LD with neighboring SNPs. Using the targeted genotyping data for the *ADH1B* SNP available in the COGA GWAS sample, we assessed the discriminatory accuracy for AD and found that rs1229984 alone has an AUC of 0.538 ($p = 7.58 \times 10^{-4}$) in COGA. Additional investigation of variants with lower frequency and expanded genetic association studies to include more variants not captured on GWAS arrays would allow for the inclusion of additional associated SNPs into a predictive score that may have better clinical validity.

The results of this particular study, along with prior genome-wide association studies of alcohol dependence, reveals that the genetic architecture of alcohol dependence includes many common alleles of small effect that may in aggregate account for variability in AD. These results provide additional support for the theory of polygenic inheritance for a disease model for alcohol dependence. This information, coupled with

further studies on the nature of variants associated with AD, may help increase understanding of the biology of AD and how to utilize associated variant effects in risk prediction and treatment for AD.

Chapter 5: Estimating the genome-wide effect of common polygenic variation and environmental factors on risk prediction for alcohol dependence symptom count

Abstract

This study assessed the extent to which common genetic variation contributes to variability in alcohol dependence (AD) symptom count, and how well aggregated effects of single nucleotide polymorphisms (SNPs) on AD symptom count predict risk for AD in independent samples. We used the *g*enome-*w*ide *c*omplex *t*rait *a*nalysis (GCTA) tool developed by Yang et al. (2011) to estimate the proportion of variance in AD symptom count accounted for by genotyped SNPs in the Collaborative Study on the Genetics of Alcoholism (COGA) genome-wide association study (GWAS) sample and the Study of Addiction: Genes and Environment (SAGE) GWAS sample. We used the COGA and SAGE samples reciprocally as discovery and validation samples. We first estimated SNP effects using the discovery sample and then created additive genetic sum scores in the validation sample, weighted by the discovery SNP effects. The genetic sum scores were then assessed for their contributions to the variance in AD symptom count and the accuracy with which they predicted AD in the validation sample. The proportion of variance accounted for by SNPs across the genome was 53.19% in COGA and not

significant in SAGE. The predictive accuracy for AD, measured by the area under the receiver operating characteristic curve (AUC) was 0.549 in COGA and 0.527 in SAGE. Both GCTA sum scores were significantly associated with AD symptom count in the replication samples, accounting for 0.46% of the variance in SAGE and 0.57% of the variance in COGA. Including additional covariates associated with AD was able to account for an additional 18.80% of the variance in symptom count.

Introduction

Alcohol dependence is a complex disorder that encompasses numerous medical, social, and psychiatric problems and has an estimated lifetime risk of 12.5-14% (Hasin et al., 2007; Kessler et al., 1994). About 50-60% of the variability in alcohol dependence is attributed to genetic factors (Kendler et al., 1992; Heath et al., 1997). Alcohol dependence, along with a vast number of other common, complex traits, has been investigated using several genome-wide association studies (GWAS) over the past several years. Numerous associated variants have been reported from GWASs of alcohol dependence, though few have reached genome-wide significance levels (Treutlein and Rietschel, 2011a; Frank et al., 2012; Wang et al., 2011; Zuo et al., 2012; Zuo et al., 2011; Schumann et al., 2011). To date, more than 2,000 novel variants have been identified as associated with complex disease (NHGRI catalogue www.genome.gov/gwastudies) (Hindorff LA, MacArthur J (European Bioinformatics Institute), Wise A, Junkins HA, Hall PN, Klemm AK, and Manolio TA) . For the majority of these traits, the amount of phenotypic variance accounted for by discovered loci is substantially lower than the estimated heritabilities for the traits based on twin and family studies (Visscher et al., 2012). A number of explanations have been attributed to the problem of "missing heritability" for common traits, including the contribution of epistasis, gene-environment interactions, epigenetics, and rare variants not captured on current GWAS arrays. Additionally, part of the missing heritability has been described as hidden heritability attributed to effects that are in fact captured on current GWAS platforms – common alleles with effects too small to be detected by genome-wide significance thresholds used in GWA studies (Gibson, 2010; Manolio et al., 2009).

Evidence for a polygenic etiology exists for many complex traits, leading to the implication that a proportion of the missing heritability could be accounted for by the aggregate effect of common SNPs already genotyped on current GWAS arrays (Gibson, 2010; International Schizophrenia Consortium et al., 2009). The Genome-Wide Complex Trait Analysis (GCTA) method developed by Yang et al. (2011) uses a mixed linear model to estimate the proportion of phenotypic variance accounted for by SNPs in total. The method models all common SNPs genotyped in GWAS together by using restricted maximum likelihood (REML) to provide an unbiased estimate of the variance explained by all SNPs. In the mixed linear model, SNP effects are treated as random variables, with additional covariates treated as fixed effects (Yang et al., 2010; Visscher et al., 2010). Prior GWAS evidence for height in 183,727 individuals showed 180 associated variants that explained 10% of the phenotypic variation in height, which is substantially lower than the estimated 80% heritability for height based on twin and family studies (Lango Allen et al., 2010). By modeling all 294,831 SNPs genotyped in the GWAS sample together, Yang et al. showed that 45% of the variance in height could be attributed to the aggregated effect of the genotyped SNPs. They found that when they accounted for incomplete LD between causal SNPs and genotyped SNPs, they were able to explain the remaining genetic variance in height (Yang et al., 2010; Visscher et al., 2010).

In our analyses, the European American (EA) portions of the COGA sample and the SAGE GWAS sample without COGA GWAS individuals were used as discovery and replication samples. We used the GCTA software tool developed by Yang et al. (http://gump.qimr.edu.au/gcta/) to estimate the proportion of variance in DSM-IV alcohol dependence (AD) symptom count explained by common SNPs separately for the COGA

GWAS sample and the SAGE GWAS sample (Yang et al., 2011). We then used the best linear unbiased prediction (BLUP) solutions for individual SNP effects to create genetic sum scores weighted by the SNP effects in a second sample, either COGA or SAGE, that was independent of the sample used to estimate the random SNP effects.

We also performed linear regression for AD symptom count in COGA and then generated genetic sum scores based on SNPs meeting varying *p*-value thresholds. The GCTA genetic sum scores from the mixed linear model and the genetic sum scores from association using linear regression were assessed for association with alcohol dependence symptom count and for predictive accuracy for alcohol dependence in an independent sample.

Finally, we incorporated several environmental risk factors that have been shown to influence risk for alcohol-related phenotypes into the prediction models. Specifically, religiosity has been associated with decreased risk for substance use disorders (Kendler et al., 2003; Koopmans et al., 1999). In independent samples, educational attainment has been found to be associated with AD (Grant et al., 2012). Marital status has also been shown to be associated with risk for AD (Dick et al., 2006). In data from the National Longitudinal Alcohol Epidemiology Study and the National Epidemiologic Study on Alcohol and Related Conditions, marital status and educational attainment were associated with alcohol dependence and income was associated with alcohol abuse (Caetano et al., 2011). We added these additional variables to baseline models including the genetic sum scores into risk models for AD symptom count and AD diagnosis in order to assess the contributions of different predictors for AD.

Materials and methods

Sample Selection

The European American (EA) subsets of the Collaborative Study on the Genetics of Alcoholism and the Study of Addiction: Genes and Environment GWAS samples were used reciprocally as independent discovery and validation samples, after removing overlap between the two samples. Both were described in detail previously in Chapter 4 and in the original COGA and SAGE GWAS reports (Edenberg et al., 2010; Bierut et al., 2010). The entire SAGE EA sample consists of 1165 cases and 1376 unrelated controls (N = 2541). For this study, the COGA portion of the SAGE EA GWAS dataset was removed (N = 939, with 555 cases and 384 controls) in order to use the FSCD (N = 519, with 275 cases and 244 controls) and COGEND (N = 1083, with 335 cases and 748 controls) portions of the SAGE dataset and the COGA GWAS dataset as independent subsets. The total number of individuals from COGA was 1398 individuals (847 cases and 552 controls) and the total number from SAGE was 1602 individuals (610 cases and 992 controls).

Data Analysis

Discovery sample analysis using the GCTA method

Figure 5.1 summarizes the study flow. The GCTA tool was used to estimate the phenotypic variance explained by autosomal SNPs using the REML method for AD

symptom count in the COGA and SAGE GWAS EA samples as described in Yang et al. (2011). The tool uses a mixed linear model, with covariates treated as fixed effects and genetic factors estimated as random effects. The phenotypic variance in AD symptom count attributed to common SNPs was calculated based on a genetic relationship matrix (GRM) across all individuals in the GWAS sample; the quantitative phenotype is regressed on genetic similarity in the mixed linear model. Prior to estimating the variance accounted for by SNPs using the genetic relationship matrix, we followed Yang et al's procedure to implement a genetic relationship cut-off of 0.025 – corresponding to cousins 2-3 times removed – in order to remove potential shared environmental and latent genetic factors in more closely related individuals, which may account for additional proportions of the phenotypic variance beyond that of the genotyped SNPs. Following pruning of the genetic relationship matrix at a relationship of 0.025, 1261 individuals remained of the 1398 individuals in the COGA EA sample and 1524 of the 1602 individuals remained in the SAGE EA sample. Two additional ancestry outliers from SAGE and two from COGA were removed from the samples prior to estimation of variance components.

The model used in the GCTA tool is represented by the following equation:

$$y = X\beta + g + \varepsilon \ \ with \ V = A\sigma_g^2 + I\sigma_\varepsilon^2$$

where y is the phenotypic value, β is the fixed effects (i.e. covariates), g represents the total genetic effects of individuals with $g \sim N(0, A\sigma_g^2).\sigma_g^2$). V represents the variance of $y.y$. A represents the genetic relationship matrix calculated from genotype data in the sample. $\sigma_g^2\sigma_g^2$ corresponds to the variance explained by all SNPs, estimated using restricted maximum likelihood. I is an

NxN identity matrix and $\sigma_\varepsilon^2 \sigma_\varepsilon^2$ represents variance explained by residual effects (Yang et al., 2011).

In our analyses, the following covariates were included as fixed effects in the mixed linear model: age, sex, 20 eigenvectors, and year of birth. GCTA estimation of the variance accounted for by SNPs in the SAGE GWAS sample included study site as an additional covariate in order to account for site differences between the FSCD and COGEND samples. The 20 eigenvectors were estimated using the GCTA tool and were included in the model because the GCTA method could be particularly sensitive to population stratification (Browning and Browning, 2011). Population stratification can occur if cases and controls differ in frequencies of alleles due to variables other than disease status that happen also to differ between cases and controls. If population structure differences between cases and controls are not accounted for, alleles attributable simply to ancestry differences could be spuriously associated with the disease phenotype. For example, in the case of height, spurious associations may occur if there are sub-populations with different ancestries in the study sample and individuals from one ancestry group happen to differ in the height phenotype compared with individuals from another group, and there are allele frequency differences between these subgroups. These alleles could very well have a contribution to height in the sample population; however, they could also have frequency differences across the sub-populations simply because individuals in different populations with distinct ancestral backgrounds often have different allele frequencies and LD structure. A plot of the first two eigenvectors indicated that the European American sample in COGA is not entirely homogeneous (Figure 1). Age, sex and year of birth were included as covariates to account for

differences in drinking patterns between cohorts and sex, and to control for possible changes in lifetime AD symptom count endorsement with increasing age.

Figure 5.1 Overview of Study Design

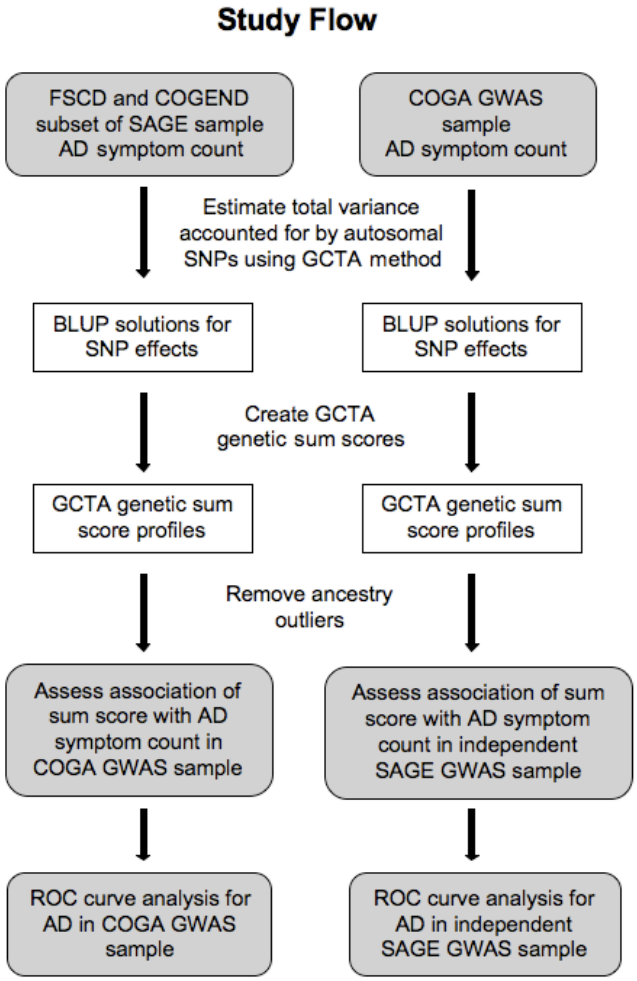
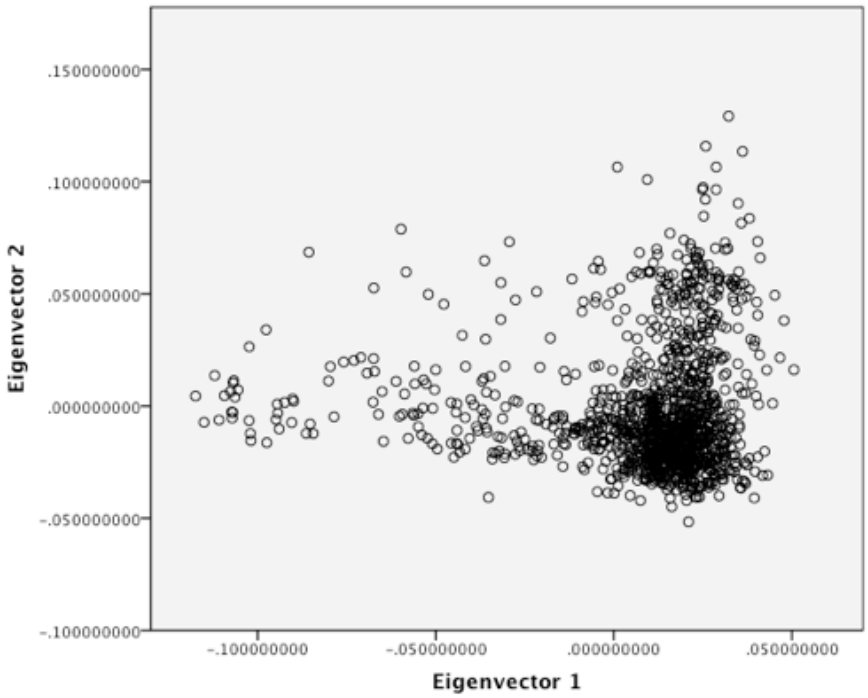


Figure 5.2 Principal components analysis plot of 1st eigenvector and 2nd eigenvector in the EA subset of the COGA GWAS sample



The number of alcohol dependence symptoms coded for each individual represented the maximum number of alcohol dependence symptoms that the individual ever endorsed across interview waves. The age variable corresponded to the age at the interview during which the maximum symptom count was endorsed and was included as a covariate because maximum lifetime symptom count could increase as an individual ages and has had more time to experience symptoms. Furthermore, an individual who endorses 7 symptoms at a young age may represent different etiology for AD compared with an individual who endorses the same number of maximum symptoms at an age that is several decades older. Year of birth, although correlated with age at interview, was included as a continuous variable in the model in order to control for cohort effects, as

patterns of AD symptom counts have differed across cohort years, particularly for women.

Validation sample replication of GCTA genetic sum scores

The random effects of the SNPs were predicted in COGA and SAGE using the best linear unbiased prediction (BLUP) method. The BLUP solutions for the individual SNP effects were calculated based on these random effects. In the set of analyses using COGA as the discovery sample, GCTA genetic sum scores of minor alleles of each genotyped SNP, weighted by the BLUP solution of each SNP effect from COGA, were created in the FSCD and COGEND portion of the SAGE EA GWAS sample using the --profile function in PLINK version 1.07 (Purcell et al., 2007). The GCTA genetic sum scores were then assessed for association with alcohol dependence symptom count using linear models in the independent SAGE sample. These data analysis steps were then repeated using SAGE as the discovery sample and COGA as the replication sample. Linear models in the replication sample included sex, age at interview, and year of birth as covariates in both COGA and SAGE, with the addition of study site as a covariate for SAGE analyses.

In the analyses using the SAGE sample as a discovery sample and the COGA sample as the replication sample, additional covariates associated with alcohol dependence including religious attendance, marital status, educational attainment, and income were available in the COGA sample, which allowed for the inclusion of these additional variables in the linear model. All linear models were performed using R version 2.12.2 (R Core Development Team, 2011).

Analysis of SNPs meeting varying *p*-value thresholds in GWAS of AD symptoms

Analyses on alcohol dependence symptom count were also performed using a linear regression approach in the COGA EA sample in order to compare the results to the GCTA genetic sum scores. Linear regression was performed using Plink version 1.07 (Plink et al., 2007) in the COGA EA sample with alcohol dependence symptoms as the outcome and age, age at interview, and year of birth as covariates. SNPs were pre-pruned at an $r^2 < 0.50$ before they were included in the analyses, resulting in 386,545 SNPs. Autosomal SNPs were selected from the results. Varying *p*-value thresholds were used to select SNPs at $p < 0.0001$, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, .0.4, and 0.5. Genetic sum scores were created in the independent FSCD and COGEND portions of the SAGE GWAS sample with each subset of SNPs by adding minor alleles weighted by standardized Betas obtained from the linear regression results in COGA. The proportion of variance accounted for in AD symptom count by the genetic sum scores was assessed in the FSCD/COGEND validation sample. Linear models in this assessment included the same covariates as the ones used to assess the GCTA genetic sum scores.

Clinical validity assessment

The GCTA genetic sum scores created based on the BLUP solutions for SNP effects and the genetic sum scores created based on linear regression results in COGA were assessed for predictive ability for alcohol dependence in the independent SAGE sample. Discriminatory accuracy was measured using the area under the receiver operating characteristic curve (AUC) in SPSS/PASW version 17.0 (SPSS Inc., Chicago IL).

Results

GCTA variance components estimation

Figure 5.3a and 5.3b displays the distribution of alcohol dependence symptom counts across the COGA and SAGE samples, respectively, separately for cases and controls. The proportion of phenotypic variance in alcohol symptom count that was accounted for by common SNPs in COGA was 53.19% (SE = 25.7%). The proportion of variance accounted for by common SNPs in SAGE was not significant at 0.0001% (SE = 24.60%).

Figure 5.3a. Alcohol dependence symptom count in the SAGE GWAS sample. Distribution of alcohol dependence symptom count is shown, separated by case and control status. Blue bars represent percentage of the controls endorsing the symptom count labeled on the x-axis. Green bars represent percentage of the cases. Several individuals who endorsed 3 or 4 symptoms were classified as controls because the symptoms did not cluster in a 12-month period.

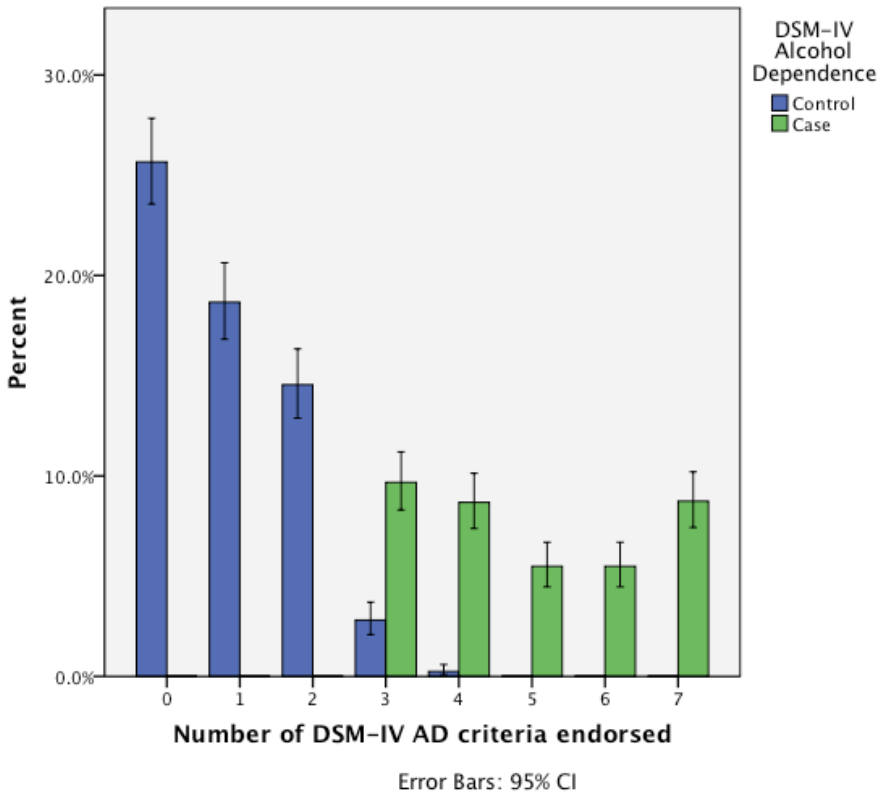
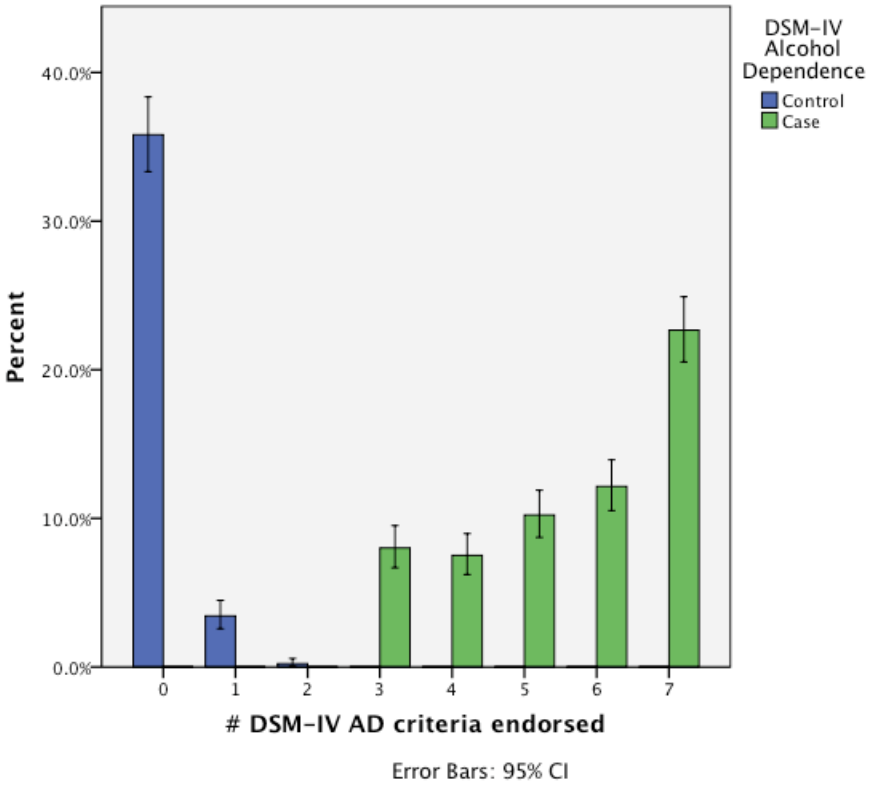


Figure 5.3b Alcohol dependence symptom count in the COGA GWAS sample. Distribution of alcohol dependence symptom count is show below, separated by case (green) and control (blue) status. The COGA GWAS sample was preferentially selected in order to maximize difference in phenotype between the cases and controls, as is shown by the larger difference in symptom count frequencies between cases and controls. Only three individuals in COGA endorsed 2 symptoms.



Linear models

GCTA genetic sum scores in the SAGE sample

The base linear model including sex, year of birth, age, and the site covariate distinguishing between COGEND and FSCD in SAGE showed that all covariates were significantly associated with alcohol dependence symptom count (Table 5.1). The adjusted r^2 of the base model showed that the model accounted for 12.46% of the variance in alcohol dependence symptom count. The GCTA genetic sum scores followed a normal distribution (Figure 5.4). After the GCTA genetic sum score was added to the base model, the score was significantly associated with alcohol dependence symptom count ($F_{1,1594} = 5.609$, $p = 0.00376$). The proportion of variance in alcohol dependence

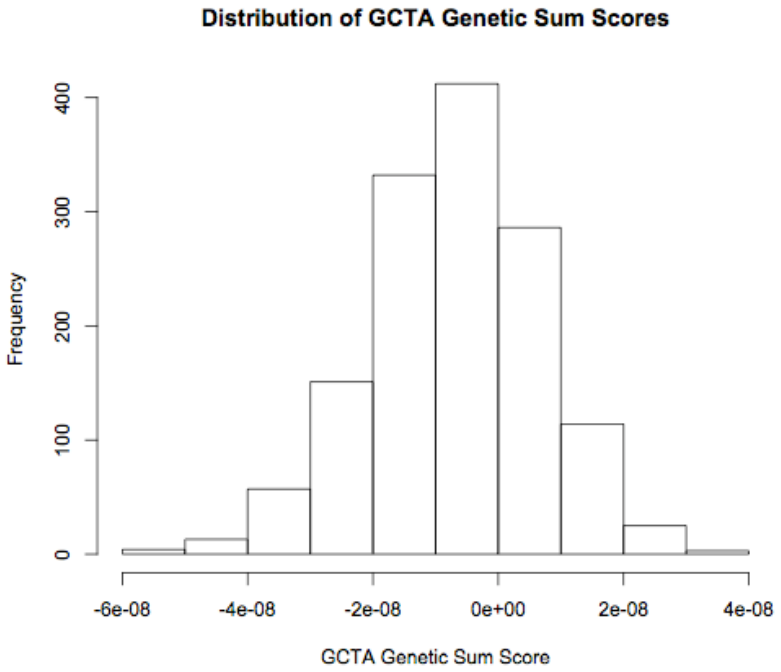
symptoms accounted for by the GCTA genetic sum score was 0.46%. The new model including the GCTA genetic sum score had an adjusted r^2 of 12.87%. A comparison of Beta estimates showed that the GCTA genetic sum score had a lower Beta compared with the other predictors.

P-value threshold analyses in the SAGE sample

In the assessment of SNPs that met varying *p*-value thresholds in association with AD symptom count in COGA, SNPs meeting increasingly stringent significance thresholds of $p < 0.0001$ and lower did not account for a significant proportion of the variance in SAGE. Genetic sum scores created based SNPs that met a p-value cutoff of 0.001 to 0.5 in COGA did account for a significant proportion of the AD symptom count variance in SAGE (Table 5.2).

Each of the *p*-value threshold genetic sum scores was modestly correlated with the GCTA genetic sum score ($r = \sim 0.3$, $p < 2.2 \times 10^{-16}$). When both the genetic sum score selected based on $p < 0.5$ and the GCTA sum score were added to the linear model in SAGE, in combination the two genetic sum scores were able to account for a total of 1.52% of the variance in AD symptom count. However, the GCTA genetic sum score was no longer significant when the additional score was added ($p = 0.153$), although the *p*-value threshold score remained significant ($p = 7.29 \times 10^{-6}$).

Figure 5.4 Distribution of GCTA genetic sum scores in SAGE.
Sum scores plotted here are created based on the sum of the minor allele for each genotyped SNP weighted by the COGA-derived BLUP solutions.



GCTA genetic sum score analyses in the COGA sample

The base linear model including sex, year of birth, and age showed that all covariates were significantly associated with alcohol dependence symptom count in the COGA GWAS sample (Table 5.1). The base model accounted for 21.07% of the variance in alcohol dependence symptom count. The GCTA genetic sum score was added to the base model, and was shown to be associated with alcohol dependence symptom count ($F_{1,1391}$ = 10.282, p = 0.001374). The proportion of variance in alcohol dependence symptoms accounted for by the GCTA genetic sum score was 0.52%. The new model including the GCTA genetic sum score had an adjusted r^2 of 21.59%.

After incorporating into the model additional covariates that have previously been found to account for some of the variance in alcohol dependence, including religious

service attendance, income, educational attainment, and marital status, these covariates together accounted for an additional 18.80% of the variance in alcohol dependence symptom count. The final model accounted for 39.38% of the phenotypic variance in alcohol symptom count. The GCTA sum score was still significant after adding the additional covariates (p = 0.00360). A comparison of Beta estimates showed that the GCTA genetic sum score had a higher Beta compared with several other covariates, including age, year of birth, religious attendance, income, and educational attainment.

Table 5.1 Summary of linear models in SAGE and COGA including GCTA sum score. All variables are centered in order to compare Beta estimates. The GCTA genetic sum score was z-transformed.

Summary of linear models in SAGE based on COGA-derived SNP effects

	Estimate	S.E.	t-value	*p*-value
Base model with covariates				
Intercept	2.424	0.053	46.043	< 2e-16
sex	-0.874	0.110	-7.968	3.05E-15
study site	0.761	0.120	6.336	3.06E-10
year of birth	-0.431	0.055	-7.826	9.08E-15
age	-0.438	0.056	-7.832	8.71E-15
Model summary: $F_{4,1595}$=57.9, p-value<2.2e-16, Multiple r2=0.1268, Adj. r2=0.1246				
Model with covariates + GCTA genetic sum scores				
Intercept	2.420	0.053	46.15	< 2e-16
sex	-0.877	0.109	-8.018	2.05E-15
study site	0.757	0.120	6.315	3.49E-10
year of birth	-0.432	0.055	-7.865	6.73E-15
age	-0.439	0.056	-7.877	6.17E-15
GCTA genetic score (z)	0.152	0.053	2.902	0.00376
Model summary: $F_{5,1594}$=48.22, p-value<2.2e-16, Multiple r2=0.1314, Adj. r2=0.1287				
Proportion of variance in AD symptom count explained by GCTA genetic sum score= 0.46%				

Summary of linear models in COGA based on SAGE-derived SNP effects

	Estimate	S.E.	t-value	*p*-value
Base model with covariates				
Intercept	3.4026	0.0689	49.377	< 2e-16

sex	-2.4155	0.1382	-17.476	< 2e-16
year of birth	0.0775	0.0242	3.209	1.36E-03
age	0.0300	0.0252	1.191	2.34E-01
Model summary: $F_{3,1392}$=125.1, p-value<2.2e-16, Multiple r^2=0.2124, Adj. r^2=0.2107				
Model with covariates + GCTA genetic sum scores				
Intercept	3.4026	0.0687	49.5410	< 2e-16
sex	-2.4076	0.1378	-17.4740	< 2e-16
year of birth	0.0832	0.0241	3.4490	5.79E-04
age	0.0357	0.0251	1.4200	1.56E-01
GCTA genetic score (z)	0.2210	0.0689	3.2070	1.37E-03
Model summary: $F_{4,1391}$=97.02, p-value<2.2e-16, Multiple r^2=0.2181, Adj. r^2=0.2159				
Proportion of variance in AD symptom count explained by GCTA genetic sum score= 0.57%				
Model with covariates + GCTA genetic sum scores + additional variables				
Intercept	2.4413	0.1046	23.3430	< 2e-16
sex	-1.9245	0.1448	-13.2860	< 2e-16
year of birth	0.1509	0.0347	4.3500	1.50E-05
age	0.1296	0.0358	3.6220	3.07E-04
GCTA genetic score	0.2102	0.0720	2.9180	3.60E-03
Religious attendance	-0.0095	0.0019	-4.8730	1.27E-06
Current income	-0.1235	0.0389	-3.1710	0.001565
Highest school grade	-0.1393	0.0346	-4.0240	6.16E-05
Marital status2	0.9239	0.6533	1.4140	0.157616
Marital status3	2.0339	0.3404	5.9750	3.20E-09
Marital status4	1.8790	0.1974	9.5170	< 2e-16
Marital status5	1.3563	0.2390	5.6760	1.80E-08
Model summary: $F_{11,1005}$=61.01, p-value<2.2e-16, Multiple r^2=0.4004, Adj. r^2=0.3938				
Proportion of variance in AD symptom count explained by additional variables = 18.8%				

Marital status is dummy-coded with “married” as reference; marital status2 = widowed, marital status3 = divorced, marital status4 = separated, marital status5 = never married

Table 5.2 Summary of linear models in SAGE and COGA using genetic sum scores created based on SNPs meeting varying *p*-value thresholds.

Validation sample proportion of variance accounted for by sum score:
Summary of linear models in SAGE based on COGA-derived results

P-value threshold score	Variance accounted for in AD sx count	p-value from linear model	AUC estimate for AD	p-value for AUC
P-value < 0.50	0.0142	3.12E-07	0.570	2.88E-06
P-value < 0.40	0.0140	4.03E-07	0.569	3.97E-06
P-value < 0.30	0.0136	5.79E-07	0.567	6.41E-06

P-value < 0.20	0.0131	9.60E-07	0.561	4.48E-05
P-value < 0.10	0.0133	7.76E-07	0.557	0.0001123
P-value < 0.05	0.0144	2.67E-07	0.559	7.25E-05
P-value < 0.01	0.0105	1.13E-05	0.546	0.001993
P-value < 0.001	0.0054	1.67E-03	0.536	0.01578
P-value < 0.0001	0.0005	3.55E-01	0.509	0.5343

Baseline models were the same as the models shown in Table 5.1. Reported are the change in r^2 attributed to the genetic sum score, and the associated p-value in the linear model. AUCs for discriminative accuracy for alcohol dependence are reported with associated p-values.

Risk prediction assessment

Clinical validity determined for the GCTA genetic sum score using the receiver operating characteristic curve showed an AUC of 0.527 (p = 0.070) for the GCTA genetic sum score in discrimination for case-control status of alcohol dependence in SAGE. An assessment of the predicted probabilities of the logistic regression model for alcohol dependence including age, sex, study site, and year of birth covariates showed that the covariates have greater discriminative accuracy than the GCTA genetic sum scores (AUC = 0.690, p <0.001). The model with covariates and the GCTA genetic sum score had a nominal increase in AUC compared with the covariates only model (AUC = 0.692, p <0.001). (Figure 5.5a).

The p-value threshold scores had significant AUCs in ROC curve analyses for SNPs that had met more liberal p-value thresholds of $p < 0.001$ and greater in the COGA discovery sample (Table 5.2). AUC point estimates showed an increasing trend as the p-value threshold used for SNP selection became more liberal.

Clinical validity analyses in COGA for the GCTA genetic sum score using the ROC curve showed a significant AUC of 0.559 (p = 0.00194) for the GCTA genetic sum score in discrimination for case-control status of alcohol dependence. An assessment of the predicted probabilities of the logistic regression model for alcohol dependence

including age, sex, and year of birth covariates showed that the covariates have greater discriminative accuracy than the GCTA genetic sum scores (AUC = 0.765). The model with covariates and the GCTA genetic sum score had a nominal increase in AUC (AUC = 0.771). The incorporation of additional variables increased the AUC further to 0.865. All AUC estimates were significant (Figure 5.5b).

Figure 5.5a Discriminatory accuracy in SAGE

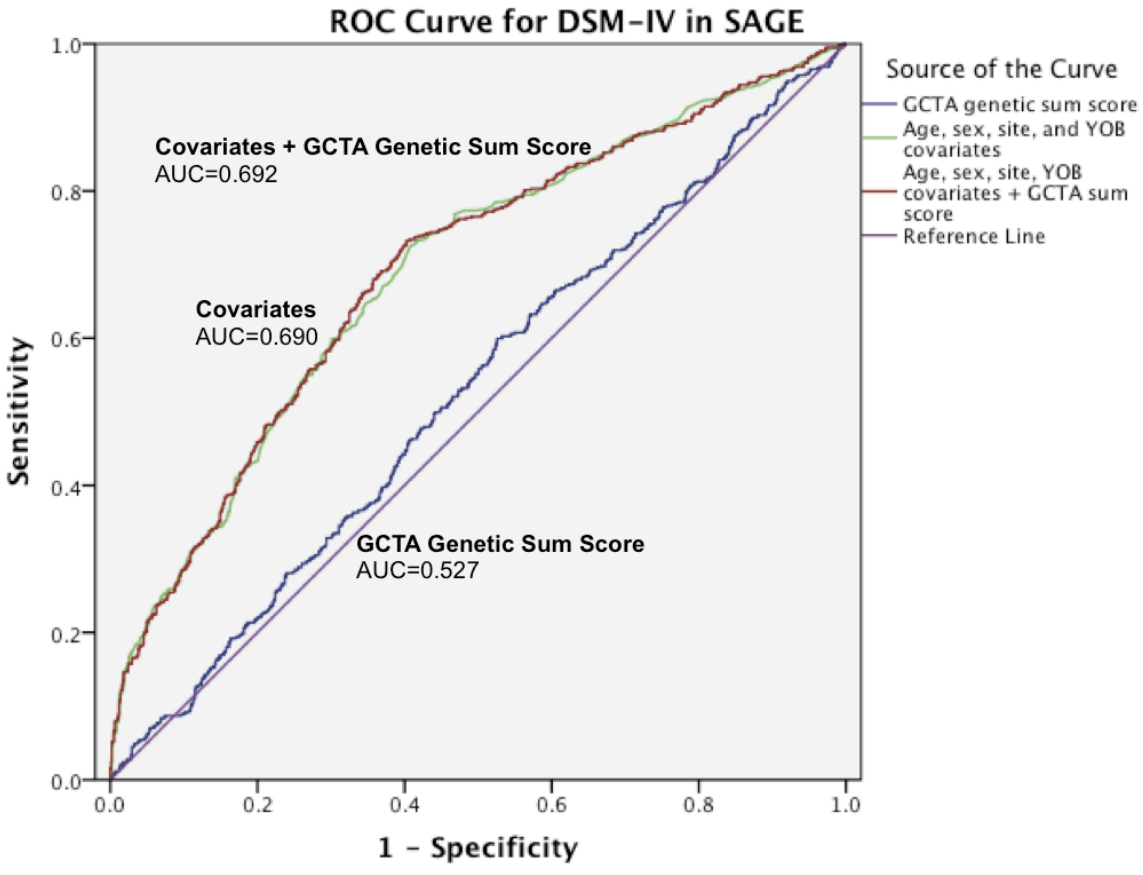
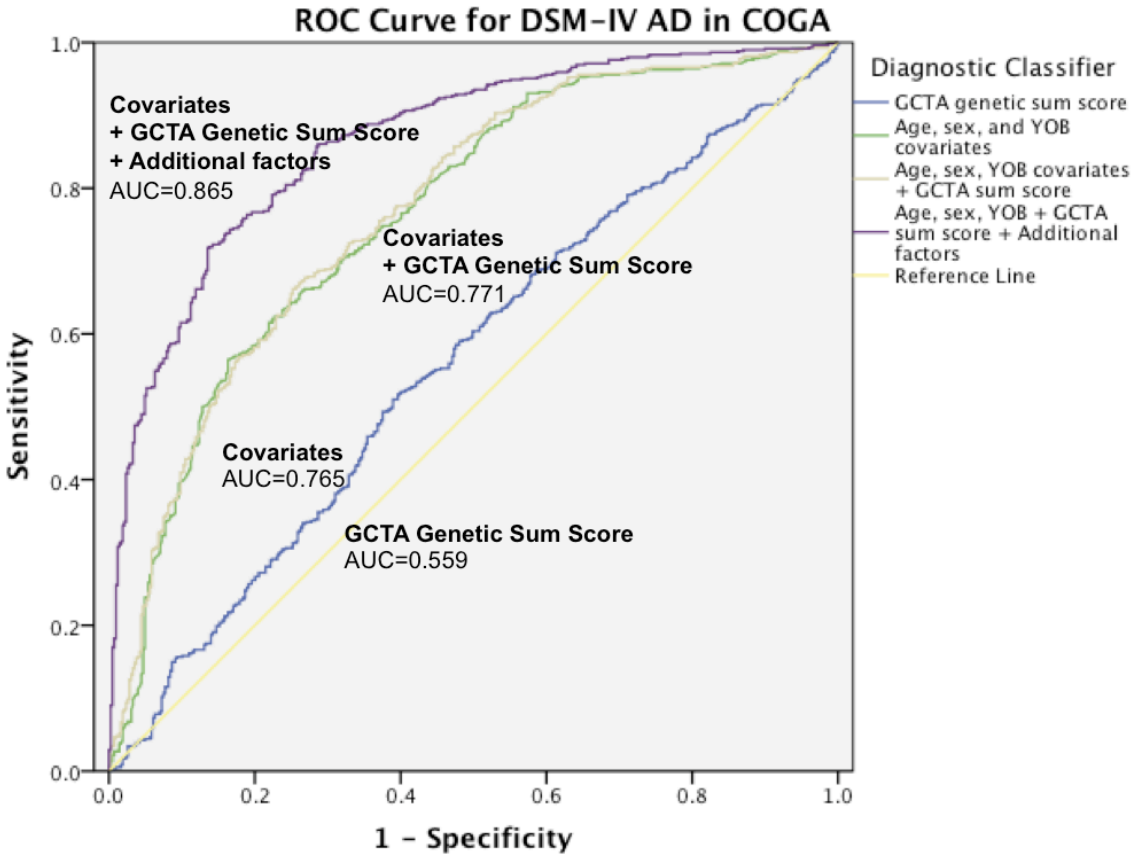


Figure 5.5b Discriminatory accuracy in COGA



Discussion

Previous candidate gene and genome-wide association studies of alcohol dependence and other traits have reported a number of associated variants. Many of these variants only account for a small fraction of the phenotypic variability in the complex trait. This study aimed to assess the proportion of phenotypic variance explained by common SNPs for DSM-IV alcohol dependence symptom count and to assess for replication and clinical validity for alcohol dependence symptom count and alcohol dependence in independent validation samples.

The results of the GCTA mixed linear model showed a significant proportion of the variance in alcohol dependence symptom count accounted for by all GWAS SNPs in COGA of about 53%, which is within the range of the estimated heritability of 50-60% for alcohol dependence based on twin studies (Kendler et al., 1992; Heath et al., 1997; Prescott and Kendler, 1999). The results from replication analyses showed significant replication of the resulting GCTA genetic sum scores created using the BLUP solutions for individual SNP effects in both the COGA and SAGE replication samples. Clinical validity for models including the GCTA genetic sum score showed small AUCs for the GCTA genetic sum score in discrimination for case-control status of alcohol dependence in both COGA and SAGE. An assessment of the predicted probabilities of the logistic regression model for alcohol dependence including age, sex, site, and year of birth covariates showed that the covariates have greater discriminative accuracy than the GCTA genetic sum scores.

Prior results in these analyses without year of birth as a covariate resulted in a substantially higher estimate of heritability attributed to the common SNPs on the array.

In COGA, the proportion of phenotypic variance in alcohol dependence explained by common SNPs was 75.64% (SE = 25.51%) in the discovery sample and 0.75% in the SAGE validation sample, in contrast to the 53.19% (SE = 25.8%) and 0.46% in the analyses that included year of birth as a covariate. This contrast in results emphasized the importance of accounting for a cohort effect in the model for alcohol dependence symptom count, particularly since drinking patterns in the United States have changed over time.

Genetic sum scores created based on linear regression results also accounted for a significant proportion of the variance in alcohol dependence symptom count. Genetic sum scores created based on a *p*-value cutoff of 0.001 to 0.5 in COGA accounted for a significant proportion of the AD symptom count variance in SAGE, with scores based on $p < 0.001$ accounting for about 0.54% of the variance and scores based on $p < 0.1$-0.5 accounting for 1-1.4% of the variance in alcohol dependence symptom count. In combination, the genetic sum score selected based on $p < 0.5$ and the GCTA genetic sum score based on a discovery COGA sample were able to account for 1.52% of the variance in AD symptom count in SAGE. The observation that the GCTA genetic sum score was no longer significant after adding the genetic sum score created based on SNPs with $p < 0.50$ in linear regression analyses and the modest correlation between the two scores suggests that there are some shared polygenic effects captured by both genetic sum scores.

Results in COGA showed that the variance accounted for by genome-wide SNPs in aggregate is substantial at about 53%. The GCTA genetic sum score, however, accounted for a much smaller proportion of variance (0.46%) in the independent SAGE

sample. This shows that there could be a great degree of prediction error of individual SNPs effects estimated by the GCTA tool. The sample sizes used in this study were limited compared with the previous study on height in which the method was first demonstrated, which had close to 4,000 individuals. In order to assess the impact of sample size on variance estimate, the original study sampled 4 replicates of 1,000, 2,000, and 3,000 individuals. They found that the average estimates were the same across samples, but the standard error increased with decreasing sample size (Yang et al., 2010).

Although common SNPs as estimated by the GCTA tool did not in aggregate account for a significant proportion of the variance in AD symptom count in SAGE, the GCTA genetic sum scores created based on the BLUP solution of these individual SNP effects in SAGE did significantly account for 0.57% of the variance in COGA. In order to assess whether these could be spurious results, further follow-up analyses were performed in which the GCTA tool was applied to a non-heritable quantitative phenotype in SAGE. A random continuous phenotype was simulated in the SAGE sample to determine whether a negative finding in the discovery sample using the GCTA tool could still create SNP effects that accounted for a significant proportion of the variance in AD symptom count in an independent sample. Similar to results in real SAGE data, aggregate SNPs did not account for a significant proportion of variance in the continuous phenotype in SAGE. Unlike the results of real SAGE data, the GCTA genetic sum score created based on the BLUP solutions of SNP effects was not significantly associated with AD symptom count in COGA ($p = 0.554$). This suggests that although the aggregate SNPs did not account for a significant proportion of the phenotypic variance in AD symptom

count in SAGE, individual SNP effects were still able to predict some degree of risk in the independent COGA sample.

Results in SAGE were substantially different from COGA. One reason for this discrepancy may be that individuals in COGA and SAGE differ in phenotypic severity and therefore underlying etiology for AD. COGA was clinically ascertained from treatment centers for alcohol dependence throughout the United States. The cocaine-dependent individuals from the FSCD sample in SAGE were also clinically ascertained from chemical dependency centers; however, the nicotine dependence COGEND study in the SAGE sample was designed as a community-based sample. The COGEND study makes up the majority (about 2/3) of the SAGE sample. Because most of the SAGE sample is community-based, the sample may represent a different range of phenotypes with distinct risk for AD compared with the COGA sample. In fact, the COGA sample was ascertained specifically to maximize difference in symptom count between cases and controls, and therefore have fewer individuals who are controls endorsing the middle-range 2 symptoms compared with SAGE, which includes individuals across the range of symptoms, including several controls who endorsed 3 AD symptoms, but did not cluster in a 12-month period, which is necessary for an AD diagnosis. Furthermore, as shown by Figure 5.4a and 5.4b, COGA has a greater number of individuals endorsing higher symptom counts than SAGE. COGA may therefore have been more enriched to estimate genetic effects for AD with higher symptom count.

These results support a polygenic model of risk for alcohol dependence symptoms that is consistent with prior studies on psychiatric and other common complex disorders (International Schizophrenia Consortium et al., 2009; Heath et al., 2011; Gibson, 2010).

Although the GCTA genetic sum scores have replicated in the independent samples, the proportion of variance accounted for the sum scores were less than that accounted for by other variables such as religious attendance, income, educational attainment, and marital status. The reason that the additional factors accounted for more of the variance in symptom count could be due partly to the clinical nature of the ascertainment of the COGA sample from treatment centers for alcohol dependence. These variables may therefore have different contributions to alcohol-related outcomes in the COGA sample than to alcohol-related traits in a general population.

This study shows that common variants, in aggregate, account for a significant, but small, proportion of the variance in alcohol dependence symptom count. Genome-wide association studies for alcohol-related phenotypes have provided more information about the genetic architecture of AD. That there has not been an emergence of single large-effect alleles accounting for a large proportion of variance in alcohol use phenotypes, but rather multiple loci accounting for a small proportion of the phenotypic variability suggests that a polygenic model could potentially improve risk prediction. Prior studies by Aulchenko et al. for height found through simulation that when predicting phenotypic extremes such as 1% of the highest and lowest values with an AUC of 0.80, a genetic score needs to explain 17% of the variance in height (Aulchenko et al., 2009). When predicting the phenotype with an AUC of 0.95, then genetic scores needed to explain 53% of the variance in height. Our simulation studies conducted specific to alcohol dependence suggest that when we have more exhaustively identified genes contributing to the genetic susceptibility toward alcohol dependence, there is the potential for AUCs approaching 80% to be reached with genetic information (Maher et al., in

preparation; Chapter 2). Although genetic information has limited clinical validity at the moment, we may have the potential for future clinical validity if we assess many genetic variants and environmental factors together.

Chapter 6: Genetic risk prediction for alcohol dependence subtypes

Abstract

Alcohol dependence (AD) is a complex psychiatric condition with a great deal of phenotypic and etiologic heterogeneity. Multiple subtypes of AD have been described, including an internalizing subtype that is often comorbid with major depressive disorder and anxiety and an externalizing subtype that is often comorbid with other drug dependence, conduct disorder, and adult antisocial personality disorder. Twin studies have suggested that part of the co-occurrence of these phenotypes is due to shared genetic factors. In this study, the Collaborative Study on the Genetics of Alcoholism (COGA) and Study of Addiction Genes and Environment (SAGE) genome-wide association study (GWAS) samples were used to investigate the etiology of phenotypes correlated with AD and risk prediction for an internalizing subtype of AD, AD with major depressive disorder (MDD), and an externalizing subtype of AD, AD with conduct disorder or with illicit drug dependence. Results showed that sum scores of individual SNP effects derived for AD symptom count also accounted for significant proportions of variance in correlated phenotypes that did not appear to be driven solely by phenotypic correlation with AD symptoms. Assessment of risk prediction for AD subtypes showed increasing, but modest, areas under the receiver operating characteristic curve (AUCs) of 0.547 to 0.5610 for SNPs meeting $p < 0.05$ to $p < 0.50$ respectively, and non-significant results for

MDD, which may be due to low power in this sample. This study suggests that the shared genetic variance between AD-related phenotypes could be due in part to aggregated genome-wide common polygenic variance of small effect, but that prediction of subtypes is modest.

Introduction

Alcohol dependence (AD) is a complex psychiatric condition with a large degree of phenotypic and etiologic heterogeneity. A number of other psychiatric phenotypes often co-occur with alcohol dependence, prompting the notion of subtypes of AD. There exists extensive history for exploration of alcohol-related subtypes. In particular, Cloninger et al. has described Type I alcoholism, which has later age of onset after 25 years, lower novelty-seeking and antisocial behavior, anxious personality traits, and higher harm avoidance than Type II alcoholism, which is characterized by earlier onset, higher novelty-seeking and antisocial behavior, and lower harm avoidance (Cloninger et al., 1988; Cloninger et al., 1981) . Another subtype that has been widely described is Babor's Type A and Type B typology (1992), in which Type A is characterized by later onset, less severe dependence, with fewer alcohol-related problems and childhood risk factors, and less comorbidity with other psychiatric disorders (Babor et al., 1992b; Babor et al., 1992a). Studies using latent class analysis of AD phenotypes have shown the following classes: a mild class with low likelihood of comorbid psychopathology, a severe class characterized by high probability of comorbidity with psychopathology, and a class with high probabilities of major depressive disorder (Sintov et al., 2010). A study by Del Boca and Hesselbrock showed a mild class, a severe class, an internalizing class with high probabilities of depression and anxiety, and an externalizing class with high levels of antisocial personality disorder (Del Boca and Hesslebrock, 1996).

Researchers have debated whether the comorbidity between AD and major depressive disorder (MDD) is attributed to a causation model in which major depressive disorder increases risk for AD, and/or vice versa, or whether an additional factor

influences risk for both (Nurnberger et al., 2002; Lyons et al., 2006). Studies have suggested that MDD and AD are genetically related. Family studies show that the co-occurrence of AD and depression occurs across relatives. First-degree relatives of alcoholic probands in the Collaborative Study on the Genetics of Alcoholism (COGA) have been described to have an increased occurrence of depressive syndrome, or depression that may or may not occur with increased alcohol consumption (Nurnberger et al., 2002). Twin studies have found a genetic correlation of approximately 0.4-0.6 between major depressive disorder and alcohol dependence (Kendler et al., 1993) . Linkage studies have identified that the same chromosomal region was linked to both AD and MDD, suggesting that a common locus may increase risk for either AD or MDD (Nurnberger et al., 2002). Candidate gene association studies have discovered associations with AD that were particularly strong for AD that is comorbid with major depressive disorder, compared with AD alone (Wang et al., 2004; Dick et al., 2007d). Recently, a genome-wide association of comorbid AD and MDD in COGA reported top results in several genes that had not been previously implicated, as well as multiple pathways, including glutaminergic genes. The majority of results were shown to be different between the comorbid phenotype and AD without MDD (Edwards et al., 2012).

Twin research has shown evidence for shared genetic contributions across alcohol dependence and externalizing psychopathologies. Kendler et al., (2003) studied the contributions of genetic and environmental factors to common psychiatric disorders and found that 69% of the heritability of alcohol dependence was accounted for by a common genetic factor contributing to a group of externalizing phenotypes, which included other drug dependence, antisocial personality disorder, and conduct disorder (Kendler et al.,

2003b) . Candidate gene studies have reported alcohol dependence associations that are stronger for, or unique to, AD with comorbid drug dependence, conduct disorder, and antisocial personality disorder, illustrating genetic contributions that may be specific to these externalizing phenotypes. Foroud et al. found that SNPs in *TACR3* that were associated with AD in EA COGA families had the strongest association in individuals with more severe AD and comorbid cocaine dependence (2008). Dick et al. showed that variants in *CHRM2* is associated with a form of AD that is comorbid with drug dependence, but not with AD alone, and that the risk allele for *CHRM2* based on association of adult AD conferred increased risk for adolescent externalizing under conditions of low parental monitoring (Dick et al., 2007b; Dick et al., 2011). Agrawal et al. found *GABRA2* to be associated with AD only in individuals with comorbid drug dependence; when these individuals were removed from the analysis, no association remained (2006). Variants associated in *GABRA2* with AD have been further characterized in a developmental sample and were found to be associated with trajectories of externalizing (Dick et al., 2009).

The convergence of phenotypic, family, twin, and molecular genetics studies suggests that distinct etiological contributions may underlie risk for internalizing and externalizing subtypes of alcohol dependence. The studies described previously here in Chapters 3 and 4 investigated the prediction of a binary diagnosis of AD. Using genetic information from specific candidate genes did not result in significant predictive accuracy for AD and GWAS results showed significant, but modest discriminatory accuracy for AD. One reason for the low predictive accuracy could be that AD is a heterogeneous phenotype, and an AD diagnosis that is comorbid with another condition may have

different underlying risk than AD in general. Accordingly, genetic variants associated with an AD subtype may be more predictive for that subtype.

In order to further characterize the comorbidity of alcohol-related phenotypes, and to assess risk prediction for variants contributing to comorbid phenotypes, we performed two sets of analyses. In the first part of the study, we assessed the extent to which genome-wide SNP effects overlapped between phenotypes that are correlated with DSM-IV AD symptom count in order to assess the genetic overlap due to common variants between correlated phenotypes. We determined the proportion of variance accounted for in traits that are correlated with AD symptom count by a genome-wide genetic sum score estimated in the Collaborative Study on the Genetics of Alcoholism (COGA) GWAS sample using the genome-wide complex trait analysis (GCTA) tool (Yang et al., 2011).

In the second part of this study, we used the COGA and the Study of Addiction: Genes and Addiction (SAGE) GWAS samples to capture genetic effects on subtypes of alcohol dependence in order to predict risk in AD subtypes in independent sample subsets. We combined the COGA and SAGE GWAS samples and then split the combined sample randomly in half. One half of the combined sample was used as a discovery sample and the other half a validation sample. We then performed GWAS analyses in the discovery sample separately for an externalizing and an internalizing subtype of alcohol dependence. The prior studies described in Chapters 4 and 5 showed that effect sizes of common SNPs for AD are individually small and that current studies have been underpowered to detect these small effect sizes at genome-wide significance thresholds. Selecting SNPs in aggregate across the genome and at more liberal *p*-value thresholds would include a greater proportion of true loci that could in aggregate account

for a significant, albeit small, proportion of the variance in AD, despite noise from null loci (Evans et al., 2009; Purcell et al., 2009). Therefore, in this study, subtypes of AD were assessed using varying significance levels in GWAS analyses, particularly as the sample size was reduced with the selection of subtypes, though power to detect loci could potentially increase with more homogeneous phenotypes. Variants meeting these thresholds were then assessed for discriminatory accuracy for AD subtypes.

Materials and methods

Sample selection

In the first part of the study, we performed all analyses in the COGA GWAS EA sample. In the second part of the study, we combined the COGA and SAGE GWAS samples, after removing overlap between the two samples, and further categorized the sample into case-control status for an internalizing subtype and an externalizing subtype of alcohol dependence. We subsequently split the sample in half so that each half contained 50% of cases and 50% of controls for the externalizing and internalizing phenotypes. In order to reduce heterogeneity, the European American portion of the combined sample was used in our analyses.

The internalizing subtype was defined as meeting DSM-IV criteria for alcohol dependence and DSM-IV criteria for a major depressive episode. We included both the "dirty" (depressive episode experienced with drugs and/or alcohol) and "clean" diagnosis (not under the influence of drugs or alcohol) for a major depressive episode, but removed individuals who met criteria for a major depressive episode due to bereavement. Table

6.1 summarizes the number of individuals who met criteria for the comorbid internalizing phenotype in each study.

The externalizing phenotype was defined by AD with illicit drug dependence and/or with conduct disorder. Controls were selected to have no illicit drug dependence, conduct disorder, or alcohol dependence. Individuals who had nicotine dependence were also removed from the control group, as nicotine dependence may still encompass some degree of shared genetic risk with an externalizing phenotype. Table 6.2 summarizes the number of individuals who met criteria for the comorbid externalizing phenotype in each study.

Table 6.1 Internalizing subtype sample size by study

	COGA	COGEND	FSCD	Total
Control (no AD or MDD)	461	671	202	1334
Case (AD + MDD)	379	97	139	615
Total	840	768	341	1949

AD = alcohol dependence; MDD = major depressive disorder

Table 6.2 Externalizing subtype sample size by study

	COGA	COGEND	FSCD	Total
Control (no ND, CD, or DD)	518	750	234	1502
Case (AD + CD or DD)	545	136	251	932
Total	1063	886	485	2434

AD = alcohol dependence; ND = nicotine dependence; CD = conduct disorder; DD = illicit drug dependence

Data analysis

GCTA correlated phenotypes analyses

The GCTA tool developed by Yang et al. (2011) was used to estimate the phenotypic variance explained by autosomal SNPs using the restricted maximum likelihood (REML) method for alcohol dependence symptom count in the COGA GWAS EA sample. The GCTA tool, described in detail in Chapter 5, uses a mixed linear model, with covariates treated as fixed effects and common SNPs genotyped on the GWAS array estimated as random effects. The phenotypic variance in alcohol dependence symptom count attributed to common SNPs was calculated based on a genetic relationship matrix across all individuals in the GWAS sample after pruning the genetic relationship matrix at a cut-off of 0.025 in order to remove potential shared environmental and latent genetic factors in more closely related individuals that could account for additional proportions of the variance beyond the genotyped SNPs. In prior analyses, we used the GCTA tool to estimate the proportion of variance in DSM-IV alcohol dependence symptom count explained by common SNPs in European American subset of the COGA GWAS sample. The following covariates were included as fixed effects in the mixed linear model: age, sex, 20 eigenvectors, and year of birth. Age, sex and year of birth were included as covariates to account for differences in drinking patterns between cohorts and sex, and to control for possible increasing lifetime alcohol dependence symptom count endorsement with increasing age.

Here, we used the best linear unbiased prediction (BLUP) solutions for individual SNP effects estimated in the COGA GWAS sample to create genetic sum scores weighted by the SNP effects within the COGA GWAS sample. The GCTA genetic sum scores were then assessed for association with AD symptom count and the following correlated phenotypes within the COGA GWAS sample: the maximum number of drinks in 24 hours that the study participant reported consuming, antisocial personality disorder symptom count, marijuana dependence symptom count, conduct disorder symptom count, cocaine dependence symptom count, opioid dependence symptom count, other drug dependence symptom count, and number of depressive symptoms, all measured by the DSM IV (American Psychiatric Association, 2000), and the Fagerstrom Test for Nicotine Dependence (FTND) score.

We also assessed two additional phenotypes to use as controls that are hypothesized not to have shared genetic variance with alcohol dependence – one that was correlated with alcohol dependence symptom count and one that was not. For the trait that had a correlation with AD symptom count, height was assessed. For the uncorrelated trait, a random, normally distributed quantitative phenotype was simulated in the COGA dataset. We compared the results of these two phenotypes with those of the psychiatric and substance use phenotypes.

A linear model was used to assess association of GCTA genetic sum scores in correlated phenotypes using R version 2.12.2 (R Core Development Team, 2011). Linear models included sex, age at interview, study site, and year of birth as covariates.

AD subtypes risk prediction analyses

We created genetic sum scores based on SNPs that met *p*-value thresholds from $p < 0.0001$ to $p < 0.50$ in the discovery sample and then assessed for prediction in the validation sample, separately for the internalizing and externalizing phenotypes. Association analysis was performed in the discovery sample using logistic regression with covariates for sex and the COGA, FSCD and COGEND study site variables in PLINK v1.07 (Purcell et al., 2007).

Prior to association, SNPs were pre-pruned based on an r^2 threshold of 0.50 using an LD-based pruning function in PLINK version 1.07. This method calculated pairwise genotypic correlations for the list of SNPs. One of each pair of SNPs with correlations greater than an r^2 of 0.50 was removed. LD calculations for SNP pruning were performed in the combined COGA and SAGE GWAS sample, sans overlap. Pruning resulted in 385,060 SNPs kept for analyses.

Genetic sum scores and ROC curve analyses

Because both the COGA and SAGE GWAS samples had the same SNPs genotyped, and were confirmed to share the direction of the genotyped strand, GWAS results were matched directly by allele. Genetic sum scores were created for autosomal SNPs. Scores of total allele count were weighted by the natural log of the odds ratio for each reference minor allele, and then divided by the number of non-missing genotypes for each individual using PLINK version 1.07:

$$\text{Genetic sum score } = \frac{\sum[x_i \times \ln(OR_i)]}{N} \frac{\sum[x_i \times \ln(OR_i)]}{N}$$

where $x_i x_i$ is the number of reference alleles at the *i*th SNP, *OR* is the corresponding odds ratio, and *N* is the number of non-missing genotypes for each individual.

Discriminatory accuracy of genetic sum scores was measured using ROC curve analysis in the caTools package in R version 2.12.2 (Tuszynski, 2011; R Core Development Team, 2011). The *p*-values associated with the AUCs for these sum scores were calculated based on the Wilcoxon rank-sum test using R version 2.12.2.

Results

Assessment of correlated phenotypes in the COGA GWAS sample

GCTA genetics sum scores created within COGA showed significant variance accounted for in correlated phenotypes of GCTA genetic sum scores created based on AD symptom count (Table 6.3). The amount of variance accounted for by the GCTA genetic sum score was not directly proportional to the correlation between alcohol dependence symptom count and the second phenotype. For example, number of depressive symptoms had a lower correlation with AD symptom count than conduct disorder symptom count, but the GCTA genetic sum score derived from AD symptom count accounted for more of the variance in number of depressive symptoms than in conduct disorder symptom count. Furthermore, opioid dependence symptom count was less correlated with AD symptom count than height, but the GCTA genetic sum score accounted for a significant proportion

of the variance in opioid dependence symptom count, compared with no significant proportion of the variance in height. The GCTA genetic sum score did not account for a significant proportion of the variance in the random continuous phenotype.

Table 6.3 Summary of variance accounted for by GCTA genetic sum score in COGA

Phenotype	Correlation with AD Sx Count *	Proportion of variance accounted for by COGA aggregate genetic sum score	*p-value*
AD Symptom Count	1*	67.72%	*p < 2e-16*
Maximum number of drinks in 24 hours	0.694*	24.07%	*p < 2e-16*
Antisocial Personality Disorder Symptom Count	0.668*	19.11%	*p < 2e-16*
Marijuana dependence symptom count	0.485*	10.79%	*p < 2e-16*
Conduct Disorder Symptom Count	0.476*	7.50%	*p < 2e-16*
Cocaine dependence symptom count	0.466*	10.88%	*p < 2e-16*
Other drug dependence symptom count	0.437*	12.01%	*p < 2e-16*
FTND Score	0.387*	11.21%	*p < 2e-16*
Number of depressive symptoms	0.377*	13.62%	*p < 2e-16*
Height	0.294*	0.10%	*p = 0.10*
Opioid dependence symptom count	0.291*	6.30%	*p < 2e-16*
Random, normally-distributed quantitative pheno	0.006 (*p* = 0.823)	3.88×10^{-5}	*p = 0.83*

Assessment of comorbid phenotypes in the COGA and SAGE GWAS sample

Table 6.4 summarizes the resulting AUCs for genetic sum scores created based on SNPs at each *p*-value for the comorbid subtypes. AUC estimates showed increasing AUCs for genetic sum scores created based on SNPs meeting increasingly liberal *p*-value thresholds for the externalizing phenotype. AUC estimates for the internalizing phenotype were not significant and did not show a consistent pattern across *p*-value threshold cut-offs.

Table 6.4 Summary of AUC estimates for AD subtypes

P-value threshold	Binary AD	Externalizing AD Subtype	P-value for AUC	Internalizing AD Subtype	P-value for AUC
P-value < 0.50	0.565	0.5610	5.86E-04	0.5192	0.3393
P-value < 0.40	0.565	0.5599	7.28E-04	0.5221	0.2709
P-value < 0.30	0.564	0.5629	3.89E-04	0.5232	0.2468
P-value < 0.20	0.564	0.5628	3.97E-04	0.5229	0.2540
P-value < 0.10	0.562	0.5598	7.39E-04	0.5260	0.1950
P-value < 0.05	0.559	0.5465	0.0088	0.5337	0.0929
P-value < 0.01	0.549	0.5238	0.1793	0.5274	0.1716
P-value < 0.001	0.528	0.5063	0.7223	0.5340	0.0901
P-value < 0.0001	0.517	0.5060	0.7331	0.5066	0.7432

Externalizing AD Subtype = alcohol dependence that is comorbid with drug dependence or conduct disorder; Internalizing AD Subtype = alcohol dependence that is comorbid with major depressive disorder.

Discussion

This study aimed to evaluate the clinical validity of genetic variants that have been associated with alcohol dependence subtypes by exploring the aggregate effect of associated SNPs on risk prediction for alcohol dependence subtypes. Assessing risk prediction for a diagnosis of alcohol dependence has shown limited predictive ability using candidate gene information and GWAS results for AD. In this study, we first assessed the underlying genetic overlap between correlated phenotypes with DSM-IV alcohol dependence symptom count. We then assessed risk for AD subtypes, with the idea that risk prediction for a disorder may improve if specific predictors are identified to be more informative for a subset of individuals.

In order to further examine the genetic overlap between traits that have been suggested to have shared genetic variance with AD, we assessed whether we could quantify this genetic correlation using aggregated effects of common SNPs across the genome. In this part of the study, the GCTA genetic sum score created based on AD symptom count accounted for a significant proportion of the phenotypic variance in multiple correlated phenotypes thought to be etiologically related to AD. As expected, the GCTA genetic sum score accounted for a significant and substantial proportion of the variance in AD symptom count (67.72%), though not all of the variance. One important limitation to assessing genetic risk within the same sample in which the genetic sum score weights were estimated is that results would be largely inflated; independent samples are necessary for replication. A randomly simulated uncorrelated phenotype was created and found to be associated with the GCTA genetic sum score, which suggests that the sum score is not explaining risk indiscriminately for phenotypes within the discovery sample. In order to control for inflation, model was also run on height so that we could determine whether the GCTA genetic sum score accounted for variance in a correlated phenotype simply because its correlation with AD symptom count rendered it a proxy for AD symptom count rather than because there exists shared etiology between the two phenotypes. Height is not thought to have shared etiology with AD, but is correlated with AD symptom count. Its correlation with AD symptoms is driven by sex, for which there is a higher prevalence of AD among males compared with females (Hasin et al., 2007). There is no reported evidence of common genetic factors contributing to both sex determination and alcohol dependence and genetic contributions to AD have been estimated to be the same in males and females (Young-Wolff et al., 2012). Results

showed that the AD symptom count-derived GCTA genetic sum score did not account for a significant proportion of variance in height ($r^2 = 0.10\%$, $p = 0.10$). This supports the notion that significant association of the GCTA genetic sum score with correlated phenotypes may mean that common polygenic variation contributes to shared etiology between the two phenotypes.

Results showed that phenotypic correlation did not directly correspond with polygenic sharing. This provides possible insight into the extent to which polygenic variation is shared between AD symptoms and correlated disorders. Although several phenotypes had similar correlations, the proportion of variance accounted for by the GCTA genetic sum score was not the same. For example, conduct disorder symptom count, had a higher phenotypic correlation with AD symptom count compared with number of depressive symptoms, but the GCTA genetic sum score derived from AD symptom count accounted for more of the variance in number of depressive symptoms compared with conduct disorder symptom count. This suggests that a greater proportion of polygenic risk may be shared between major depressive disorder symptom count and AD symptom count than between conduct disorder symptom count and AD.

Risk prediction assessment of an externalizing AD subtype showed significant AUCs that increased with genetic sum scores created based on SNPs meeting less stringent *p*-value thresholds. Results for the internalizing phenotype, however, did not show a consistent significant pattern of AUCs across *p*-value threshold cut-offs. The reason could be that the COGA-SAGE combined GWAS sample size was greatly reduced in the analyses. Selecting only individuals meeting the subtype-defined phenotypes reduced the sample, and splitting the samples into discovery and validation

sets further reduced the sample sizes. Although a more narrowly defined phenotype could confer increased power for a genetic association study, a reduction in sample size could result in a corresponding reduction in power.

The results of this study support a shared genetic component to alcohol-related phenotypic correlations that are consistent with twin and family studies. The substantial proportion of variance accounted for by AD symptom count-derived scores in the correlated phenotypes suggests that a proportion of the shared genetic variance is due to aggregated genome-wide common polygenic variance of small individual effect. Risk prediction for AD subtypes resulted in similar results to an AD binary diagnosis, with non-significant results for MDD, likely due to limited power. Assessment in larger samples may help uncover variants that are informative for predicting and possibly diagnosing subtypes of AD.

Chapter 7: Conclusions

A greater understanding of the biology of alcohol dependence and its interplay with the environment is key to the advent of better treatment, prediction, and prevention of alcohol dependence. Advances in gene identification are now yielding replicable susceptibility loci for many complex traits, including alcohol dependence. However, each accounts for only a small proportion of the phenotypic variance of a trait. These studies assessed the predictive ability of currently known associated genetic variants, new aggregate measures to capture genetic information across the genome, and environmental factors associated with alcohol dependence phenotypes, with the aim of providing a clinical interpretation of the current body of knowledge of factors contributing to risk for alcohol dependence.

Based on data simulation results, there is potential for alcohol dependence risk prediction; using only information on genetic contributions to alcohol dependence, it is possible to have discriminatory accuracy close to 80%. A risk prediction algorithm based on genetic information alone is not expected to reach an AUC of 100% because environment also plays a substantial role in risk for alcohol dependence. Adding the effects of environmental risk factors show that there is the potential for this AUC to increase to 0.95. Results demonstrate that as more causative loci are uncovered, and as environmental factors for alcohol dependence are found, then genetic susceptibility

testing and environmental risk prediction for alcohol dependence may have high clinical validity.

The series of studies reported here showed that risk prediction for alcohol dependence is currently limited. Genetic sum scores based on candidate gene SNPs did not have significant predictive value as assessed by receiver operating characteristic curve analyses. Aggregate genetic sum scores including information on more variants across the genome did, however, have significant discriminatory accuracy for predicting alcohol dependence. This was particularly true for variants that met varying *p*-value thresholds that were less stringent – a subsequent genetic sum score based on these nominally associated SNPs had a higher AUC than SNPs meeting more stringent thresholds. We found that a similar pattern was observed for repeated subsampling cross-validation procedures. We also found that results were similar for alcohol dependence internalizing and externalizing subtypes, as well as alcohol dependence symptom count, though our sample size for the internalizing subtype may have been too small to detect a significant effect. We also assessed the predictive ability of individual effects estimated based on the best linear unbiased prediction (BLUP) solutions of SNP effects using restricted maximum likelihood to estimate the phenotypic variance accounted for AD symptom count using the GCTA method developed by Yang et al. (2011). Scores based on markers across the genome accounted for a significant, but small proportion of the variance in alcohol dependence symptom count. A polygenic theory of etiology exists for many disorders of complex, multifactorial etiology (Fisher 1918). These results provide evidence for a polygenic contribution to alcohol dependence.

One reason that candidate gene sum scores associated in different populations did not in aggregate replicate in the COGA and SAGE GWAS samples may be due to phenotypic and genetic heterogeneity across the samples. In these analyses, several of the candidate gene variants had an opposite direction of effect across the candidate gene family-based association sample and the GWAS samples. Population differences between the high-density COGA sample and the GWAS samples could account for differences in direction of effect. According to Zuo et al., two different populations with distinct structures could share causal variants, but the variants in LD with these causal variants, which are the ones detected in genetic association studies, could be different across distinct populations (Zuo et al., 2012). To address this issue, they suggested that focus be placed on identifying risk regions, rather than individual variants, as these individual risk markers could be different in distinct populations, but be tagging the same causal variants. A next step would be to investigate prediction using regions with subsets of variants rather than replicating all candidate gene variants, as different variants from different regions may have different directions of effect and varying discriminatory accuracy for specific populations.

In comparison with other risk prediction models, the AUCs of genetic sum scores are similar to AUCs of some clinical prediction models, showing that despite low clinical validity, genetic information for AD is in some circumstances comparative to existing clinical risk models for complex diseases. For example, the Gail model, or the Breast Cancer Risk Assessment Tool, is a commonly used tool to estimate 5-year and lifetime risk for invasive breast cancer and to help determine whether or not chemopreventative therapy is warranted. It assesses risk factors such as age, race, family history, personal

reproductive history including age at menarche and age at first live-born, and medical history such as atypical hyperplasia on biopsy and number of biopsies performed. A Gail model five-year absolute risk estimate of breast cancer of 1.67% is used as the cut-off in FDA guidelines for use of selective estrogen receptor modulators such as tamoxifen or raloxifen in risk reduction for breast cancer. The model has been validated at the population level, by comparing the expected (E) number of women who develop breast cancer based on the model with the observed number of women in the population who develop breast cancer (O). The calibration performance of the risk tool, or E/O ratio, has been found to be >93% in U.S. women (Spiegelman et al., 1994; Rockhill et al., 2001; Bondy et al., 1994; Costantino et al., 1999). Evaluation of clinical validity at the individual level using ROC curve analyses, however, showed that this model has poor classifier ability, with an AUC of 0.557 - 0.60 (Gail, 2008; Mealiffe et al., 2010; Wacholder et al., 2010; Rockhill et al., 2001; Anothaisintawee et al., 2012). Despite the low AUC, the Gail model is used as a screening tool in conjunction with personalized risk-benefits analysis of the clinical utility of chemopreventative therapy, and has been found to be particularly useful for younger women with increased risk based on the Gail model (Gail et al., 1999). Large prospective studies, such as the Framingham Heart Study, have developed extensive clinical risk algorithms, such as the Framingham Risk Score, which uses clinical risk factors including age, sex, LDL cholesterol, HDL cholesterol, total cholesterol, dyslipidemia, blood pressure, treatment for hypertension, and smoking to predict risk for cardiovascular disease. The Framingham Risk Score has been found to have AUCs of 0.72 – 0.86 for predicting cardiovascular disease mortality (Siontis et al., 2012). The Cambridge risk score for predicting risk for type II diabetes

includes family history, age, sex, drug treatment, smoking status, and body mass index and has an AUC of 0.78 for classifying cases and controls for type II diabetes (Talmud et al., 2010). Once more risk factors are found for AD, similar large, prospective population-based studies will be necessary for validation of risk prediction algorithms for alcohol-related phenotypes.

Our findings stress that despite interest in genetic testing, and the availability of testing through direct to consumer (DTC) avenues, genetic testing for AD is not yet ready to be applied in a clinical setting. Without additional studies of clinical utility and validation of risk models for AD, the low clinical validity of genetic variants for AD connotes limited predictive value for risk assessment and decision-making based on information about specific genetic variants alone. Family history is shown here to be a better predictor of alcohol dependence than information from SNPs in aggregate. Despite its higher AUC, family history still does not have a discriminative accuracy that is close to the typical target AUC of 0.80 for screening tools. Therefore, the prediction of alcohol dependence has room to improve if it is to be used as a screening tool with high clinical validity. Expanded family history information across multiple relationships may have better predictive ability. In our analyses, an ordinal family history variable had a nominally higher AUC compared with a binary family history variable based on the same criteria for parental history of AD. Family history containing information about relatives other than mother and father of a proband had a higher AUC than information about parental AD history only. The cost of taking a sufficient family history must also be weighed against the cost of having a genetic test, which should in turn be weighed against the use of environmental factors, as well as a combination of all three predictors –

specific genetic variants, environmental factors, and family history information. As sequencing costs decrease, future cost-benefit analyses may show that for some conditions with well-validated genetic information, the price of analyzing a panel of SNPs from sequencing data may be lower than taking a detailed family and/or clinical history to obtain the same risk estimates. Perhaps an even more effective way to predict risk would be to combine family history information with genetic information in order to uncover the degree of penetrance for a specific set of alleles, and to find out which genetic markers are most informative for a particular family based on its history (Ruderfer et al., 2010).

Continued studies on the clinical validity of genetic risk factors in cross-sectional and retrospective studies may inform prospective clinical studies. Developmental studies assessing the discriminatory accuracy of phenotypes across the lifespan can be assessed. Since the maximum AUC of a disorder is dependent on its prevalence, and prevalence of AD is age-dependent, the ability of genetic marker to discriminate affection status would vary with age. Further research could show that environmental prediction and intervention may be most helpful in particular subgroups. Genetic information may allow for identification of subgroups that will respond better to prevention and treatment measures, such as targeting alcohol metabolism with withdrawal phenotypes. Finally, with knowledge of genetic contributions to AD, studies should help determine the best way in which to communicate risk for a complex disorder such as AD in a clinically meaningful way that minimizes fatalism and maximizes healthy behavior and psychosocial adaptation.

Future extensions of this work would include the incorporation of gene-gene and gene-environment interactions to create an additively coded composite risk score weighted by the effect size associated with the interactions in both simulated data and real data. Many of the "environmental" variables included in these analyses have been shown to moderate genetic contributions to AD, and therefore augment genetic influences under some conditions and mask them under other conditions. Consequently, these interactions could affect how informative genetic variants may be for a given individual, and may depend on existing background environmental risk for that individual. Data mining techniques such as tree-based regression and classification are approaches designed for incorporating all of the complex predictor interactions and main effects to predict risk for a disorder. Several data mining techniques have been applied to GWAS data. Pirooznia et al. (2012) used Bayesian networks, support vector machine, random forest, radial basis function network, and logistic regression methods to investigate predictors of bipolar disorder compared with the polygenic scores created based on varying *p*-value thresholds. They found that the AUCs of data mining approaches did not reach the AUC of the polygenic scoring approach, with the exception of results from Bayesian networks, which produced an AUC of 0.550 compared with the AUC of 0.549 for a polygenic score (Pirooznia et al., 2012). Another recent study by Wong et al. (2012) used the advance recursive partition approach (ARPA), a classificatory multidimensional tree technique, to assess autosomal non-synonymous SNPs in 188 genes, comorbid conditions, and covariates to predict major depressive disorder (MDD). They found a tree structure that replicated in an independent sample and had a predictive accuracy of AUC = 0.63 for MDD (Wong et al., 2012). Unlike traditional gene-finding approaches using

association analyses, data mining approaches focus specifically on optimizing risk prediction by modeling multiple correlations and interactions between variables that predict risk significantly for an outcome of interest. Using an a priori set of expanded, less correlated genetic markers to assess for a set of predictors in a data mining framework that incorporates interactions between markers may improve risk prediction for AD. Furthermore, including higher order interactions between genetic and environmental variables could improve prediction using data mining, particularly in light of the suggested importance of gene-environment interactions in risk for AD.

In summary, these studies aimed to find ways of utilizing the growing body of knowledge in the development of alcohol dependence to find clinically meaningful profiles of risk, and to place risk factors for AD in a clinical context. To date, the current status of psychiatric disorders and other complex diseases have limited genetic prediction options. Although variants from association studies are currently not sufficient to act as predictors of disease status, data simulations show that by using larger studies, markers that serve as good classifiers may be found. Growing interest in genetic counseling and testing to determine and manage risk for complex phenotypes calls for continuous evaluation of clinical validity and utility that should be concomitant with focused investigation of the genetic influences on these traits. The hope is that by combining a number of risk variants, more of the variance explained by genetic factors can be accounted for even though the variance explained independently by each variant is small (Wray et al., 2008). The increasing number of gene-identification studies, coupled with decreasing costs of high-throughput genotyping and whole genome sequencing, may allow for better detection of markers covering a range of frequencies and effects on

alcohol dependence. By continuing to find biomarkers for disease risk and progression, we may eventually be able to develop models for better risk prediction, diagnosis, prognosis, treatment, and prevention.

Literature Cited

Literature Cited

Agrawal A, Edenberg HJ, Foroud T, Bierut LJ, Dunne G, Hinrichs AL, Nurnberger JI, Crowe R, Kuperman S, Schuckit MA, Begleiter H, Porjesz B, Dick DM. 2006. Association of GABRA2 with drug dependence in the collaborative study of the genetics of alcoholism sample. Behav Genet 36:640-650.

Agrawal A, Freedman ND, Bierut LJ. 2011. Genome-wide association studies of alcohol intake--a promising cocktail? Am J Clin Nutr 93:681-683.

Agrawal A, Freedman ND, Cheng YC, Lin P, Shaffer JR, Sun Q, Taylor K, Yaspan B, Cole JW, Cornelis MC, DeSensi RS, Fitzpatrick A, Heiss G, Kang JH, O'Connell J, Bennett S, Bookman E, Bucholz KK, Caporaso N, Crout R, Dick DM, Edenberg HJ, Goate A, Hesselbrock V, Kittner S, Kramer J, Nurnberger JI,Jr, Qi L, Rice JP, Schuckit M, van Dam RM, Boerwinkle E, Hu F, Levy S, Marazita M, Mitchell BD, Pasquale LR, Bierut LJ, GENEVA Consortium. 2012. Measuring alcohol consumption for genomic meta-analyses of alcohol intake: Opportunities and challenges. Am J Clin Nutr 95:539-547.

Amadeo S, Abbar M, Fourcade ML, Waksman G, Leroux MG, Madec A, Selin M, Champiat JC, Brethome A, Leclaire Y. 1993. D2 dopamine receptor gene and alcoholism. J Psychiatr Res 27:173-179.

Amadeo S, Noble EP, Fourcade-Amadeo ML, Tetaria C, Brugiroux MF, Nicolas L, Deparis X, Elbaz A, Zhang X, Ritchie T, Martin PV, Mallet J. 2000. Association of D2 dopamine receptor and alcohol dehydrogenase 2 genes with polynesian alcoholics. Eur Psychiatry 15:97-102.

American Psychiatric Association. 1987. Diagnostic and statistical manual of mental disorders, 3rd edition. (revised). American Psychiatric Press: Washington, DC .

American Psychiatric Association. 2000. Diagnostic and statistical manual of mental disorders, 4th edition, text revision. Washington, DC: American Psychiatric Association.

Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A. 2012. Risk prediction models of breast cancer: A systematic review of model performances. Breast Cancer Res Treat 133:1-10.

Arinami T, Itokawa M, Komiyama T, Mitsushio H, Mori H, Mifune H, Hamaguchi H, Toru M. 1993. Association between severity of alcoholism and the A1 allele of the dopamine D2 receptor gene TaqI A RFLP in japanese. Biol Psychiatry 33:108-114.

Attia J, Ioannidis JP, Thakkinstian A, McEvoy M, Scott RJ, Minelli C, Thompson J, Infante-Rivard C, Guyatt G. 2009. How to use an article about genetic association: C: What are the results and will they help me in caring for my patients? JAMA 301:304-308.

Aulchenko YS, Struchalin MV, Belonogova NM, Axenovich TI, Weedon MN, Hofman A, Uitterlinden AG, Kayser M, Oostra BA, van Duijn CM, Janssens AC, Borodin PM. 2009. Predicting human height by victorian and genomic methods. Eur J Hum Genet 17:1070-1075.

Austin JC, Honer WG. 2007. The genomic era and serious mental illness: A potential application for psychiatric genetic counseling. Psychiatr Serv 58:254-261.

Austin JC, Peay HL. 2006. Applications and limitations of empiric data in provision of recurrence risks for schizophrenia: A practical review for healthcare professionals providing clinical psychiatric genetics consultations. Clin Genet 70:177-187.

Babor TF, Dolinsky ZS, Meyer RE, Hesselbrock M, Hofmann M, Tennen H. 1992a. Types of alcoholics: Concurrent and predictive validity of some common classification schemes. Br J Addict 87:1415-1431.

Babor TF, Hofmann M, DelBoca FK, Hesselbrock V, Meyer RE, Dolinsky ZS, Rounsaville B. 1992b. Types of alcoholics, I. evidence for an empirically derived typology based on indicators of vulnerability and severity. Arch Gen Psychiatry 49:599-608.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265.

Begleiter H, Reich T, Hesselbrock V, Porjesz B, Li TK, Schuckit M, Edenberg HJ, Rice J. 1995. The collaborative study on the genetics of alcoholism. Alcohol Health Res World 19:228-236.

Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S, Hinrichs AL, Almasy L, Breslau N, Culverhouse RC, Dick DM, Edenberg HJ, Foroud T, Grucza RA, Hatsukami D, Hesselbrock V, Johnson EO, Kramer J, Krueger RF, Kuperman S, Lynskey M, Mann K, Neuman RJ, Nothen MM, Nurnberger JI,Jr, Porjesz B, Ridinger M, Saccone NL, Saccone SF, Schuckit MA, Tischfield JA, Wang JC, Rietschel M, Goate AM, Rice JP, Gene, Environment Association Studies Consortium. 2010. A genome-wide association study of alcohol dependence. Proc Natl Acad Sci U S A 107:5082-5087.

Bierut LJ, Goate AM, Breslau N, Johnson EO, Bertelsen S, Fox L, Agrawal A, Bucholz KK, Grucza R, Hesselbrock V, Kramer J, Kuperman S, Nurnberger J, Porjesz B, Saccone NL, Schuckit M, Tischfield J, Wang JC, Foroud T, Rice JP, Edenberg HJ. 2012. ADH1B is associated with alcohol dependence and alcohol consumption in populations of european and african ancestry. Mol Psychiatry 17:445-450.

Blum K, Noble EP, Sheridan PJ, Finley O, Montgomery A, Ritchie T, Ozkaragoz T, Fitch RJ, Sadlack F, Sheffield D. 1991. Association of the A1 allele of the D2 dopamine receptor gene with severe alcoholism. Alcohol 8:409-416.

Blum K, Noble EP, Sheridan PJ, Montgomery A, Ritchie T, Jagadeeswaran P, Nogami H, Briggs AH, Cohn JB. 1990. Allelic association of human dopamine D2 receptor gene in alcoholism. JAMA 263:2055-2060.

Blum K, Sheridan PJ, Wood RC, Braverman ER, Chen TJ, Cull JG, Comings DE. 1996. The D2 dopamine receptor gene as a determinant of reward deficiency syndrome. J R Soc Med 89:396-400.

Boehnke M. 1994. Limits of resolution of genetic linkage studies: Implications for the positional cloning of human disease genes. Am J Hum Genet 55:379-390.

Bolos AM, Dean M, Lucas-Derse S, Ramsburg M, Brown GL, Goldman D. 1990. Population and pedigree studies reveal a lack of association between the dopamine D2 receptor gene and alcoholism. JAMA 264:3156-3160.

Bondy ML, Lustbader ED, Halabi S, Ross E, Vogel VG. 1994. Validation of a breast cancer risk assessment model in women with a positive family history. J Natl Cancer Inst 86:620-625.

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331.

Browning SR, Browning BL. 2011. Population structure can inflate SNP-based heritability estimates. Am J Hum Genet 89:191-3; author reply 193-5.

Buscemi L, Turchi C. 2011. An overview of the genetic susceptibility to alcoholism. Med Sci Law 51 Suppl 1:S2-6.

Caetano R, Baruah J, Chartier KG. 2011. Ten-year trends (1992 to 2002) in sociodemographic predictors and indicators of alcohol abuse and dependence among whites, blacks, and hispanics in the united states. Alcohol Clin Exp Res 35:1458-1466.

Chen CC, Lu RB, Chen YC, Wang MF, Chang YC, Li TK, Yin SJ. 1999. Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism. Am J Hum Genet 65:795-807.

Chen HJ, Tian H, Edenberg HJ. 2005. Natural haplotypes in the regulatory sequences affect human alcohol dehydrogenase 1C (ADH1C) gene expression. Hum Mutat 25:150-155.

Chen WJ, Chen CH, Huang J, Hsu YP, Seow SV, Chen CC, Cheng AT. 2001. Genetic polymorphisms of the promoter region of dopamine D2 receptor and dopamine transporter genes and alcoholism among four aboriginal groups and han chinese in taiwan. Psychiatr Genet 11:187-195.

Choi IG, Son HG, Yang BH, Kim SH, Lee JS, Chai YG, Son BK, Kee BS, Park BL, Kim LH, Choi YH, Shin HD. 2005. Scanning of genetic effects of alcohol metabolism gene (ADH1B and ADH1C) polymorphisms on the risk of alcoholism. Hum Mutat 26:224-234.

Ciaranello RD, Ciaranello AL. 1991. Genetics of major psychiatric disorders. Annu Rev Med 42:151-158.

Clark DB, Winters KC. 2002. Measuring risks and outcomes in substance use disorders prevention research. J Consult Clin Psychol 70:1207-1223.

Cloninger CR, Bohman M, Sigvardsson S. 1981. Inheritance of alcohol abuse. cross-fostering analysis of adopted men. Arch Gen Psychiatry 38:861-868.

Cloninger CR, Sigvardsson S, Gilligan SB, von Knorring AL, Reich T, Bohman M. 1988. Genetic heterogeneity and the classification of alcoholism. Adv Alcohol Subst Abuse 7:3-16.

Comings DE, Muhleman D, Ahn C, Gysin R, Flanagan SD. 1994. The dopamine D2 receptor gene: A genetic risk factor in substance abuse. Drug Alcohol Depend 34:175-180.

Cook BL, Wang ZW, Crowe RR, Hauser R, Freimer M. 1992. Alcoholism and the D2 receptor gene. Alcohol Clin Exp Res 16:806-809.

Cook NR. 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115:928-935.

Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS. 1999. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 91:1541-1548.

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

Crabb DW, Dipple KM, Thomasson HR. 1993. Alcohol sensitivity, alcohol metabolism, risk of alcoholism, and the role of alcohol and aldehyde dehydrogenase genotypes. J Lab Clin Med 122:234-240.

Crabb DW, Edenberg HJ, Bosron WF, Li TK. 1989. Genotypes for aldehyde dehydrogenase deficiency and alcohol sensitivity. the inactive ALDH2(2) allele is dominant. J Clin Invest 83:314-316.

Cruz C, Camarena B, Mejia JM, Paez F, Eroza V, Ramon De La Fuente J, Kershenobich D, Nicolini H. 1995. The dopamine D2 receptor gene TaqI A1 polymorphism and alcoholism in a mexican population. Arch Med Res 26:421-426.

Del Boca FK, Hesslebrock MN. 1996. Gender and alcoholic subtypes. Alcohol Health Res. World 20:56-62.

Demirkan A, Penninx BW, Hek K, Wray NR, Amin N, Aulchenko YS, van Dyck R, de Geus EJ, Hofman A, Uitterlinden AG, Hottenga JJ, Nolen WA, Oostra BA, Sullivan PF, Willemsen G, Zitman FG, Tiemeier H, Janssens AC, Boomsma DI, van Duijn CM, Middeldorp CM. 2011. Genetic risk profiles for depression and anxiety in adult and elderly cohorts. Mol Psychiatry 16:773-783.

Dick DM, Agrawal A, Schuckit MA, Bierut L, Hinrichs A, Fox L, Mullaney J, Cloninger CR, Hesselbrock V, Nurnberger JI,Jr, Almasy L, Foroud T, Porjesz B, Edenberg H,

Begleiter H. 2006a. Marital status, alcohol dependence, and GABRA2: Evidence for gene-environment correlation and interaction. J Stud Alcohol 67:185-194.

Dick DM, Agrawal A, Wang JC, Hinrichs A, Bertelsen S, Bucholz KK, Schuckit M, Kramer J, Nurnberger J,Jr, Tischfield J, Edenberg HJ, Goate A, Bierut LJ. 2007a. Alcohol dependence with comorbid drug dependence: Genetic and phenotypic associations suggest a more severe form of the disorder with stronger genetic contribution to risk. Addiction 102:1131-1139.

Dick DM, Aliev F, Kramer J, Wang JC, Hinrichs A, Bertelsen S, Kuperman S, Schuckit M, Nurnberger J,Jr, Edenberg HJ, Porjesz B, Begleiter H, Hesselbrock V, Goate A, Bierut L. 2007c. Association of CHRM2 with IQ: Converging evidence for a gene influencing intelligence. Behav Genet 37:265-272.

Dick DM, Aliev F, Wang JC, Grucza RA, Schuckit M, Kuperman S, Kramer J, Hinrichs A, Bertelsen S, Budde JP, Hesselbrock V, Porjesz B, Edenberg HJ, Bierut LJ, Goate A. 2008a. Using dimensional models of externalizing psychopathology to aid in gene identification. Arch Gen Psychiatry 65:310-318.

Dick DM, Aliev F, Wang JC, Saccone S, Hinrichs A, Bertelsen S, Budde J, Saccone N, Foroud T, Nurnberger J,Jr, Xuei X, Conneally PM, Schuckit M, Almasy L, Crowe R, Kuperman S, Kramer J, Tischfield JA, Hesselbrock V, Edenberg HJ, Porjesz B, Rice JP, Bierut L, Goate A. 2008b. A systematic single nucleotide polymorphism screen to fine-map alcohol dependence genes on chromosome 7 identifies association with a novel susceptibility gene ACN9. Biol Psychiatry 63:1047-1053.

Dick DM, Edenberg HJ, Xuei X, Goate A, Kuperman S, Schuckit M, Crowe R, Smith TL, Porjesz B, Begleiter H, Foroud T. 2004. Association of GABRG3 with alcohol dependence. Alcohol Clin Exp Res 28:4-9.

Dick DM, Latendresse SJ, Lansford JE, Budde JP, Goate A, Dodge KA, Pettit GS, Bates JE. 2009. Role of GABRA2 in trajectories of externalizing behavior across development and evidence of moderation by parental monitoring. Arch Gen Psychiatry 66:649-657.

Dick DM, Meyers J, Aliev F, Nurnberger J,Jr, Kramer J, Kuperman S, Porjesz B, Tischfield J, Edenberg HJ, Foroud T, Schuckit M, Goate A, Hesselbrock V, Bierut L. 2010. Evidence for genes on chromosome 2 contributing to alcohol dependence with conduct disorder and suicide attempts. Am J Med Genet B Neuropsychiatr Genet 153B:1179-1188.

Dick DM, Meyers JL, Latendresse SJ, Creemers HE, Lansford JE, Pettit GS, Bates JE, Dodge KA, Budde J, Goate A, Buitelaar JK, Ormel J, Verhulst FC, Huizink AC. 2011. CHRM2, parental monitoring, and adolescent externalizing behavior: Evidence for gene-environment interaction. Psychol Sci 22:481-489.

Dick DM, Plunkett J, Wetherill LF, Xuei X, Goate A, Hesselbrock V, Schuckit M, Crowe R, Edenberg HJ, Foroud T. 2006b. Association between GABRA1 and drinking behaviors in the collaborative study on the genetics of alcoholism sample. Alcohol Clin Exp Res 30:1101-1110.

Dick DM, Rose RJ, Viken RJ, Kaprio J, Koskenvuo M. 2001. Exploring gene-environment interactions: Socioregional moderation of alcohol use. J Abnorm Psychol 110:625-632.

Dick DM, Wang JC, Plunkett J, Aliev F, Hinrichs A, Bertelsen S, Budde JP, Goldstein EL, Kaplan D, Edenberg HJ, Nurnberger J,Jr, Hesselbrock V, Schuckit M, Kuperman S, Tischfield J, Porjesz B, Begleiter H, Bierut LJ, Goate A. 2007d. Family-based association analyses of alcohol dependence phenotypes across DRD2 and neighboring gene ANKK1. Alcohol Clin Exp Res 31:1645-1653.

Dinwiddie S, Heath AC, Dunne MP, Bucholz KK, Madden PA, Slutske WS, Bierut LJ, Statham DB, Martin NG. 2000. Early sexual abuse and lifetime psychopathology: A co-twin-control study. Psychol Med 30:41-52.

Drgon T, D'Addario C, Uhl GR. 2006. Linkage disequilibrium, haplotype and association studies of a chromosome 4 GABA receptor gene cluster: Candidate gene variants for addictions. Am J Med Genet B Neuropsychiatr Genet 141B:854-860.

Ducci F, Enoch MA, Funt S, Virkkunen M, Albaugh B, Goldman D. 2007. Increased anxiety and other similarities in temperament of alcoholics with and without antisocial personality disorder across three diverse populations. Alcohol 41:3-12.

Edenberg HJ. 2007a. The genetics of alcohol metabolism: Role of alcohol dehydrogenase and aldehyde dehydrogenase variants. Alcohol Res Health 30:5-13.

Edenberg HJ. 2007b. The genetics of alcohol metabolism: Role of alcohol dehydrogenase and aldehyde dehydrogenase variants. Alcohol Res Health 30:5-13.

Edenberg HJ, Bierut LJ, Boyce P, Cao M, Cawley S, Chiles R, Doheny KF, Hansen M, Hinrichs T, Jones K, Kelleher M, Kennedy GC, Liu G, Marcus G, McBride C, Murray SS, Oliphant A, Pettengill J, Porjesz B, Pugh EW, Rice JP, Rubano T,

Shannon S, Steeke R, Tischfield JA, Tsai YY, Zhang C, Begleiter H. 2005. Description of the data from the collaborative study on the genetics of alcoholism (COGA) and single-nucleotide polymorphism genotyping for genetic analysis workshop 14. BMC Genet 6 Suppl 1:S2.

Edenberg HJ, Dick DM, Xuei X, Tian H, Almasy L, Bauer LO, Crowe RR, Goate A, Hesselbrock V, Jones K, Kwon J, Li TK, Nurnberger JI,Jr, O'Connor SJ, Reich T, Rice J, Schuckit MA, Porjesz B, Foroud T, Begleiter H. 2004. Variations in GABRA2, encoding the alpha 2 subunit of the GABA(A) receptor, are associated with alcohol dependence and with brain oscillations. Am J Hum Genet 74:705-714.

Edenberg HJ, Foroud T. 2006. The genetics of alcoholism: Identifying specific genes through family studies. Addict Biol 11:386-396.

Edenberg HJ, Foroud T, Koller DL, Goate A, Rice J, Van Eerdewegh P, Reich T, Cloninger CR, Nurnberger JI,Jr, Kowalczuk M, Wu B, Li TK, Conneally PM, Tischfield JA, Wu W, Shears S, Crowe R, Hesselbrock V, Schuckit M, Porjesz B, Begleiter H. 1998. A family-based analysis of the association of the dopamine D2 receptor (DRD2) with alcoholism. Alcohol Clin Exp Res 22:505-512.

Edenberg HJ, Koller DL, Xuei X, Wetherill L, McClintick JN, Almasy L, Bierut LJ, Bucholz KK, Goate A, Aliev F, Dick D, Hesselbrock V, Hinrichs A, Kramer J, Kuperman S, Nurnberger JI,Jr, Rice JP, Schuckit MA, Taylor R, Todd Webb B, Tischfield JA, Porjesz B, Foroud T. 2010. Genome-wide association study of alcohol dependence implicates a region on chromosome 11. Alcohol Clin Exp Res 34:840-852.

Edenberg HJ, Xuei X, Chen HJ, Tian H, Wetherill LF, Dick DM, Almasy L, Bierut L, Bucholz KK, Goate A, Hesselbrock V, Kuperman S, Nurnberger J, Porjesz B, Rice J, Schuckit M, Tischfield J, Begleiter H, Foroud T. 2006. Association of alcohol dehydrogenase genes with alcohol dependence: A comprehensive analysis. Hum Mol Genet 15:1539-1549.

Edenberg HJ, Xuei X, Wetherill LF, Bierut L, Bucholz K, Dick DM, Hesselbrock V, Kuperman S, Porjesz B, Schuckit MA, Tischfield JA, Almasy LA, Nurnberger JI,Jr, Foroud T. 2008. Association of NFKB1, which encodes a subunit of the transcription factor NF-kappaB, with alcohol dependence. Hum Mol Genet 17:963-970.

Edwards AC, Aliev F, Bierut LJ, Bucholz KK, Edenberg H, Hesselbrock V, Kramer J, Kuperman S, Nurnberger JI,Jr, Schuckit MA, Porjesz B, Dick DM. 2012. Genome-

wide association study of comorbid depressive syndrome and alcohol dependence. Psychiatr Genet 22:31-41.

Ehlers CL, Spence JP, Wall TL, Gilder DA, Carr LG. 2004. Association of ALDH1 promoter polymorphisms with alcohol-related phenotypes in southwest california indians. Alcohol Clin Exp Res 28:1481-1486.

Ehlers CL, Walter NA, Dick DM, Buck KJ, Crabbe JC. 2010. A comparison of selected quantitative trait loci associated with alcohol use phenotypes in humans and mouse models. Addict Biol 15:185-199.

Enoch MA. 2008. The role of GABA(A) receptors in the development of alcoholism. Pharmacol Biochem Behav 90:95-104.

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

Enoch MA, Schwartz L, Albaugh B, Virkkunen M, Goldman D. 2006. Dimensional anxiety mediates linkage of GABRA2 haplotypes with alcoholism. Am J Med Genet B Neuropsychiatr Genet 141B:599-607.

Evans DM, Visscher PM, Wray NR. 2009. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet 18:3525-3531.

Evers-Kiebooms G, Decruyenaere M. 1998. Predictive testing for huntington's disease: A challenge for persons at risk and for professionals. Patient Educ Couns 35:15-26.

Evers-Kiebooms G, Swerts A, Cassiman JJ, Van den Berghe H. 1989. The motivation of at-risk individuals and their partners in deciding for or against predictive testing for huntington's disease. Clin Genet 35:29-40.

Fehr C, Sander T, Tadic A, Lenzen KP, Anghelescu I, Klawe C, Dahmen N, Schmidt LG, Szegedi A. 2006. Confirmation of association of the GABRA2 gene with alcohol dependence by subtype-specific analysis. Psychiatr Genet 16:9-17.

Feighner JP, Robins E, Guze SB, Woodruff RA,Jr, Winokur G, Munoz R. 1972. Diagnostic criteria for use in psychiatric research. Arch Gen Psychiatry 26:57-63.

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

Finn CT, Smoller JW. 2006. Genetic counseling in psychiatry. Harv Rev Psychiatry 14:109-121.

Foley PF, Loh EW, Innes DJ, Williams SM, Tannenberg AE, Harper CG, Dodd PR. 2004. Association studies of neurotransmitter gene polymorphisms in alcoholic caucasians. Ann N Y Acad Sci 1025:39-46.

Foroud T, Edenberg HJ, Goate A, Rice J, Flury L, Koller DL, Bierut LJ, Conneally PM, Nurnberger JI, Bucholz KK, Li TK, Hesselbrock V, Crowe R, Schuckit M, Porjesz B, Begleiter H, Reich T. 2000. Alcoholism susceptibility loci: Confirmation studies in a replicate sample and further mapping. Alcohol Clin Exp Res 24:933-945.

Foroud T, Wetherill LF, Kramer J, Tischfield JA, Nurnberger JI,Jr, Schuckit MA, Xuei X, Edenberg HJ. 2008. The tachykinin receptor 3 is associated with alcohol and cocaine dependence. Alcohol Clin Exp Res 32:1023-1030.

Foster MW, Mulvihill JJ, Sharp RR. 2009. Evaluating the utility of personal genomic information. Genet Med 11:570-574.

Frank J, Cichon S, Treutlein J, Ridinger M, Mattheisen M, Hoffmann P, Herms S, Wodarz N, Soyka M, Zill P, Maier W, Mossner R, Gaebel W, Dahmen N, Scherbaum N, Schmal C, Steffens M, Lucae S, Ising M, Muller-Myhsok B, Nothen MM, Mann K, Kiefer F, Rietschel M. 2012. Genome-wide significant association between alcohol dependence and a variant in the ADH gene cluster. Addict Biol 17:171-180.

Gail MH. 2008. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst 100:1037-1041.

Gail MH, Costantino JP, Bryant J, Croyle R, Freedman L, Helzlsouer K, Vogel V. 1999. Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. J Natl Cancer Inst 91:1829-1846.

Gamm JL, Nussbaum RL, Biesecker BB. 2004a. Genetics and alcoholism among at-risk relatives I: Perceptions of cause, risk, and control. Am J Med Genet A 128A:144-150.

Gamm JL, Nussbaum RL, Bowles Biesecker B. 2004b. Genetics and alcoholism among at-risk relatives II: Interest and concerns about hypothetical genetic testing for alcoholism risk. Am J Med Genet A 128A:151-155.

Gelernter J, Kranzler H. 1999. D2 dopamine receptor gene (DRD2) allele and haplotype frequencies in alcohol dependent and control subjects: No association with phenotype or severity of phenotype. Neuropsychopharmacology 20:640-649.

Gelernter J, Kranzler HR. 2009. Genetics of alcohol dependence. Hum Genet 126:91-99.

Gelernter J, O'Malley S, Risch N, Kranzler HR, Krystal J, Merikangas K, Kennedy JL, Kidd KK. 1991. No association between an allele at the D2 dopamine receptor gene (DRD2) and alcoholism. JAMA 266:1801-1807.

Gerra G, Leonardi C, Cortese E, D'Amore A, Lucchini A, Strepparola G, Serio G, Farina G, Magnelli F, Zaimovic A, Mancini A, Turci M, Manfredini M, Donnini C. 2007. Human kappa opioid receptor gene (OPRK1) polymorphism is associated with opiate addiction. Am J Med Genet B Neuropsychiatr Genet 144B:771-775.

Gibson G. 2010. Hints of hidden heritability in GWAS. Nat Genet 42:558-560.

Ginter E, Simko V. 2009. Alcoholism: Recent advances in epidemiology, biochemistry and genetics. Bratisl Lek Listy 110:307-311.

Goldman D, Dean M, Brown GL, Bolos AM, Tokola R, Virkkunen M, Linnoila M. 1992. D2 dopamine receptor genotype and cerebrospinal fluid homovanillic acid, 5-hydroxyindoleacetic acid and 3-methoxy-4-hydroxyphenylglycol in alcoholics in finland and the united states. Acta Psychiatr Scand 86:351-357.

Grant BF. 2000. Estimates of US children exposed to alcohol abuse and dependence in the family. Am J Public Health 90:112-115.

Grant JD, Scherrer JF, Lynskey MT, Agrawal A, Duncan AE, Haber JR, Heath AC, Bucholz KK. 2012. Associations of alcohol, nicotine, cannabis, and drug use/dependence with educational attainment: Evidence from cotwin-control analyses. Alcohol Clin Exp Res 36:1412-1420.

Greenland P. 2008. Comments on 'evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. pencina, R. B. D'agostino sr, R. B. D'agostino jr, R. S. vasan, statistics in medicine (DOI: 10.1002/sim.2929). Stat Med 27:188-190.

Guindalini C, Scivoletto S, Ferreira RG, Breen G, Zilberman M, Peluso MA, Zatz M. 2005. Association of genetic variants in alcohol dehydrogenase 4 with alcohol dependence in brazilian patients. Am J Psychiatry 162:1005-1007.

Harper P. 2004. Practical genetic counseling 6thEdition. Woburn, MA: Hodder Arnold.

Hasin D, Aharonovich E, Liu X, Mamman Z, Matseoane K, Carr And LG, Li TK. 2002a. Alcohol dependence symptoms and alcohol dehydrogenase 2 polymorphism: Israeli ashkenazis, sephardics, and recent russian immigrants. Alcohol Clin Exp Res 26:1315-1321.

Hasin D, Aharonovich E, Liu X, Mamman Z, Matseoane K, Carr L, Li TK. 2002b. Alcohol and ADH2 in israel: Ashkenazis, sephardics, and recent russian immigrants. Am J Psychiatry 159:1432-1434.

Hasin DS, Goodwin RD, Stinson FS, Grant BF. 2005. Epidemiology of major depressive disorder: Results from the national epidemiologic survey on alcoholism and related conditions. Arch Gen Psychiatry 62:1097-1106.

Hasin DS, Stinson FS, Ogburn E, Grant BF. 2007. Prevalence, correlates, disability, and comorbidity of DSM-IV alcohol abuse and dependence in the united states: Results from the national epidemiologic survey on alcohol and related conditions. Arch Gen Psychiatry 64:830-842.

Heath AC, Bucholz KK, Madden PA, Dinwiddie SH, Slutske WS, Bierut LJ, Statham DJ, Dunne MP, Whitfield JB, Martin NG. 1997. Genetic and environmental contributions to alcohol dependence risk in a national twin sample: Consistency of findings in women and men. Psychol Med 27:1381-1396.

Heath AC, Whitfield JB, Martin NG, Pergadia ML, Goate AM, Lind PA, McEvoy BP, Schrage AJ, Grant JD, Chou YL, Zhu R, Henders AK, Medland SE, Gordon SD, Nelson EC, Agrawal A, Nyholt DR, Bucholz KK, Madden PA, Montgomery GW. 2011. A quantitative-trait genome-wide association study of alcoholism risk in the community: Findings and implications. Biol Psychiatry .

Heilig M, Goldman D, Berrettini W, O'Brien CP. 2011. Pharmacogenetic approaches to the treatment of alcohol addiction. Nat Rev Neurosci 12:670-684.

Hendershot CS, Bryan AD, Ewing SW, Claus ED, Hutchison KE. 2011. Preliminary evidence for associations of CHRM2 with substance use and disinhibition in adolescence. J Abnorm Child Psychol 39:671-681.

Hicks BM, South SC, Dirago AC, Iacono WG, McGue M. 2009. Environmental adversity and increasing genetic risk for externalizing disorders. Arch Gen Psychiatry 66:640-648.

Hietala J, Pohjalainen T, Heikkila-Kallio U, West C, Salaspuro M, Syvalahti E. 1997. Allelic association between D2 but not D1 dopamine receptor gene and alcoholism in finland. Psychiatr Genet 7:19-25.

Hill MK, Sahhar M. 2006. Genetic counselling for psychiatric disorders. Med J Aust 185:507-510.

Hindorff LA, MacArthur J (European Bioinformatics Institute), Wise A, Junkins HA, Hall PN, Klemm AK, and Manolio TA. A catalog of published genome-wide association studies. available at: Www.genome.gov/gwastudies. accessed [february 2012].

Hinrichs AL, Wang JC, Bufe B, Kwon JM, Budde J, Allen R, Bertelsen S, Evans W, Dick D, Rice J, Foroud T, Nurnberger J, Tischfield JA, Kuperman S, Crowe R, Hesselbrock V, Schuckit M, Almasy L, Porjesz B, Edenberg HJ, Begleiter H, Meyerhof W, Bierut LJ, Goate AM. 2006. Functional variant in a bitter-taste receptor (hTAS2R16) influences risk of alcohol dependence. Am J Hum Genet 78:103-111.

International HapMap Consortium. 2003. The international HapMap project. Nature 426:789-796.

International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748-752.

Ishiguro H, Arinami T, Saito T, Akazawa S, Enomoto M, Mitushio H, Fujishiro H, Tada K, Akimoto Y, Mifune H, Shioduka S, Hamaguchi H, Toru M, Shibuya H. 1998. Association study between the -141C ins/del and TaqI A polymorphisms of the dopamine D2 receptor gene and alcoholism. Alcohol Clin Exp Res 22:845-848.

Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE. 2009. Interpretation of genetic association studies: Markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 5:e1000337.

Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM. 2006. Predictive testing for complex diseases using multiple genes: Fact or fiction? Genet Med 8:395-400.

Janssens AC, van Duijn CM. 2008. Genome-based prediction of common diseases: Advances and prospects. Hum Mol Genet 17:R166-73.

Janssens AC, van Duijn CM. 2009. Genome-based prediction of common diseases: Methodological considerations for future research. Genome Med 1:20.

Jostins L, Barrett JC. 2011. Genetic risk prediction in complex disease. Hum Mol Genet 20:R182-8.

Jung MH, Park BL, Lee BC, Ro Y, Park R, Shin HD, Bae JS, Kang TC, Choi IG. 2011. Association of CHRM2 polymorphisms with severity of alcohol dependence. Genes Brain Behav 10:253-256.

Kalsi G, Prescott CA, Kendler KS, Riley BP. 2009. Unraveling the molecular mechanisms of alcohol dependence. Trends Genet 25:49-55.

Kendler KS, Bulik CM, Silberg J, Hettema JM, Myers J, Prescott CA. 2000. Childhood sexual abuse and adult psychiatric and substance use disorders in women: An epidemiological and cotwin control analysis. Arch Gen Psychiatry 57:953-959.

Kendler KS, Chen X, Dick D, Maes H, Gillespie N, Neale MC, Riley B. 2012. Recent advances in the genetic epidemiology and molecular genetics of substance use disorders. Nat Neurosci 15:181-189.

Kendler KS, Heath AC, Neale MC, Kessler RC, Eaves LJ. 1992. A population-based twin study of alcoholism in women. JAMA 268:1877-1882.

Kendler KS, Heath AC, Neale MC, Kessler RC, Eaves LJ. 1993. Alcoholism and major depression in women. A twin study of the causes of comorbidity. Arch Gen Psychiatry 50:690-698.

Kendler KS, Kalsi G, Holmans PA, Sanders AR, Aggen SH, Dick DM, Aliev F, Shi J, Levinson DF, Gejman PV. 2011. Genomewide association analysis of symptoms of

alcohol dependence in the molecular genetics of schizophrenia (MGS2) control sample. Alcohol Clin Exp Res 35:963-975.

Kendler KS, Liu XQ, Gardner CO, McCullough ME, Larson D, Prescott CA. 2003a. Dimensions of religiosity and their relationship to lifetime psychiatric and substance use disorders. Am J Psychiatry 160:496-503.

Kendler KS, Prescott CA, Myers J, Neale MC. 2003b. The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. Arch Gen Psychiatry 60:929-937.

Kertes DA, Kalsi G, Prescott CA, Kuo PH, Patterson DG, Walsh D, Kendler KS, Riley BP. 2011. Neurotransmitter and neuromodulator genes associated with a history of depressive symptoms in individuals with alcohol dependence. Alcohol Clin Exp Res 35:496-505.

Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Arch Gen Psychiatry 62:593-602.

Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen HU, Kendler KS. 1994. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the united states. results from the national comorbidity survey. Arch Gen Psychiatry 51:8-19.

Khoury MJ, McBride C, Schully SD, Ioannidis JP, Feero WG, Janssens AC, Gwinn M, Simons-Morton DG, Bernhardt JM, Cargill M, Chanock SJ, Church GM, Coates RJ, Collins FS, Croyle RT, Davis BR, Downing GJ, Duross A, Friedman S, Gail MH, Ginsburg GS, Green RC, Greene MH, Greenland P, Gulcher JR, Hsu A, Hudson KL, Kardia SL, Kimmel PL, Lauer MS, Miller AM, Offit K, Ransohoff DF, Roberts HS, Rasooly RS, Stefansson K, Terry SF, Teutsch SM, Trepanier A, Wanke KL, Witte JS, Xu J. 2009. The scientific foundation for personal genomics: Recommendations from a national institutes of health-centers for disease control and prevention multidisciplinary workshop. Genet Med .

Khoury MJ, Yang Q, Gwinn M, Little J, Dana Flanders W. 2004. An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. Genet Med 6:38-47.

Knop J, Penick EC, Jensen P, Nickel EJ, Gabrielli WF, Mednick SA, Schulsinger F. 2003. Risk factors that predicted problem drinking in danish men at age thirty. J Stud Alcohol 64:745-755.

Knopik VS, Heath AC, Madden PA, Bucholz KK, Slutske WS, Nelson EC, Statham D, Whitfield JB, Martin NG. 2004. Genetic effects on alcohol dependence risk: Re-evaluating the importance of psychiatric and other heritable risk factors. Psychol Med 34:1519-1530.

Konishi T, Luo HR, Calvillo M, Mayo MS, Lin KM, Wan YJ. 2004. ADH1B*1, ADH1C*2, DRD2 (-141C ins), and 5-HTTLPR are associated with alcoholism in mexican american men living in los angeles. Alcohol Clin Exp Res 28:1145-1152.

Kono Y, Yoneda H, Sakai T, Nonomura Y, Inayama Y, Koh J, Sakai J, Inada Y, Imamichi H, Asaba H. 1997. Association between early-onset alcoholism and the dopamine D2 receptor gene. Am J Med Genet 74:179-182.

Koob GF, Volkow ND. 2010. Neurocircuitry of addiction. Neuropsychopharmacology 35:217-238.

Koopmans JR, Boomsma DI. 1996. Familial resemblances in alcohol use: Genetic or cultural transmission? J Stud Alcohol 57:19-28.

Koopmans JR, Slutske WS, van Baal GC, Boomsma DI. 1999. The influence of religion on alcohol use initiation: Evidence for genotype X environment interaction. Behav Genet 29:445-453.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, Konig IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M,

Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpelainen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Pare G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietilainen KH, Pouta A, Ridderstrale M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kahonen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimaki T, Melander O, Mosley TH,Jr, Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tonjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Gronberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Volzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832-838.

Lappalainen J, Krupitsky E, Remizov M, Pchelina S, Taraskina A, Zvartau E, Somberg LK, Covault J, Kranzler HR, Krystal JH, Gelernter J. 2005. Association between alcoholism and gamma-amino butyric acid alpha2 receptor subtype in a russian population. Alcohol Clin Exp Res 29:493-498.

Lawrence RE, Appelbaum PS. 2011. Genetic testing in psychiatry: A review of attitudes and beliefs. Psychiatry 74:315-331.

Lee JF, Lu RB, Ko HC, Chang FM, Yin SJ, Pakstis AJ, Kidd KK. 1999. No association between DRD2 locus and alcoholism after controlling the ADH and ALDH genotypes in chinese han population. Alcohol Clin Exp Res 23:592-599.

Lee SH, van der Werf JH, Hayes BJ, Goddard ME, Visscher PM. 2008. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet 4:e1000231.

Leighton JW, Valverde K, Bernhardt BA. 2012. The general public's understanding and perception of direct-to-consumer genetic test results. Public Health Genomics 15:11-21.

Li D, Zhao H, Gelernter J. 2011. Strong association of the alcohol dehydrogenase 1B gene (ADH1B) with alcohol dependence and alcohol-induced medical diseases. Biol Psychiatry 70:504-512.

Lind PA, Eriksson CJ, Wilhelmsen KC. 2008. The role of aldehyde dehydrogenase-1 (ALDH1A1) polymorphisms in harmful alcohol consumption in a finnish population. Hum Genomics 3:24-35.

Lind PA, Macgregor S, Vink JM, Pergadia ML, Hansell NK, de Moor MH, Smit AB, Hottenga JJ, Richter MM, Heath AC, Martin NG, Willemsen G, de Geus EJ, Vogelzangs N, Penninx BW, Whitfield JB, Montgomery GW, Boomsma DI, Madden PA. 2010. A genomewide association study of nicotine and alcohol dependence in australian and dutch populations. Twin Res Hum Genet 13:10-29.

Lobos EA, Todd RD. 1998. Association analysis in an evolutionary context: Cladistic analysis of the DRD2 locus to test for association with alcoholism. Am J Med Genet 81:411-419.

Long JC, Knowler WC, Hanson RL, Robin RW, Urbanek M, Moore E, Bennett PH, Goldman D. 1998. Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an american indian population. Am J Med Genet 81:216-221.

Lu RB, Ko HC, Chang FM, Castiglione CM, Schoolfield G, Pakstis AJ, Kidd JR, Kidd KK. 1996. No association between alcoholism and multiple polymorphisms at the dopamine D2 receptor gene (DRD2) in three distinct taiwanese populations. Biol Psychiatry 39:419-429.

Luo X, Kranzler HR, Zuo L, Wang S, Blumberg HP, Gelernter J. 2005a. CHRM2 gene predisposes to alcohol dependence, drug dependence and affective disorders: Results from an extended case-control structured association study. Hum Mol Genet 14:2421-2434.

Luo X, Kranzler HR, Zuo L, Yang BZ, Lappalainen J, Gelernter J. 2005b. ADH4 gene variation is associated with alcohol and drug dependence: Results from family controlled and population-structured association studies. Pharmacogenet Genomics 15:755-768.

Lyons MJ, Schultz M, Neale M, Brady K, Eisen S, Toomey R, Rhein A, Faraone S, Tsuang M. 2006. Specificity of familial vulnerability for alcoholism versus major depression in men. J Nerv Ment Dis 194:809-817.

Macgregor S, Lind PA, Bucholz KK, Hansell NK, Madden PA, Richter MM, Montgomery GW, Martin NG, Heath AC, Whitfield JB. 2009. Associations of ADH and ALDH2 gene variation with self report alcohol reactions, consumption and dependence: An integrated analysis. Hum Mol Genet 18:580-593.

Manolio TA. 2010. Genomewide association studies and assessment of the risk of disease. N Engl J Med 363:166-176.

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

Mathews R, Hall W, Carter A. 2012. Direct-to-consumer genetic testing for addiction susceptibility: A premature commercialisation of doubtful validity and value. Addiction .

Matthews AG, Hoffman EK, Zezza N, Stiffler S, Hill SY. 2007. The role of the GABRA2 polymorphism in multiplex alcohol dependence families with minimal comorbidity: Within-family association and linkage analyses. J Stud Alcohol Drugs 68:625-633.

McBride CM, Bryan AD, Bray MS, Swan GE, Green ED. 2012. Health behavior change: Can genomics improve behavioral adherence? Am J Public Health 102:401-405.

McBride CM, Koehly LM, Sanderson SC, Kaphingst KA. 2010. The behavioral response to personalized genetic information: Will genetic risk profiles motivate individuals and families to choose more healthful behaviors? Annu Rev Public Health 31:89-103.

McGuire AL, Diaz CM, Wang T, Hilsenbeck SG. 2009. Social networkers' attitudes toward direct-to-consumer personal genome testing. Am J Bioeth 9:3-10.

McLaughlin KA, Green JG, Gruber MJ, Sampson NA, Zaslavsky AM, Kessler RC. 2010. Childhood adversities and adult psychiatric disorders in the national comorbidity survey replication II: Associations with persistence of DSM-IV disorders. Arch Gen Psychiatry 67:124-132.

Mealiffe ME, Stokowski RP, Rhees BK, Prentice RL, Pettinger M, Hinds DA. 2010. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. J Natl Cancer Inst 102:1618-1627.

Merikangas KR. 1990. The genetic epidemiology of alcoholism. Psychol Med 20:11-22.

Merikangas KR, Leckman JF, Prusoff BA, Pauls DL, Weissman MM. 1985. Familial transmission of depression and alcoholism. Arch Gen Psychiatry 42:367-372.

Milne BJ, Caspi A, Harrington H, Poulton R, Rutter M, Moffitt TE. 2009. Predictive value of family history on severity of illness: The case for depression, anxiety, alcohol dependence, and drug dependence. Arch Gen Psychiatry 66:738-747.

Mitra SK. 1958. On the limiting power function of the frequency chi-square test. Annals of Mathematical Statistics 29:1221-1233.

Mobascher A, Rujescu D, Mittelstrass K, Giegling I, Lamina C, Nitz B, Brenner H, Fehr C, Breitling LP, Gallinat J, Rothenbacher D, Raum E, Muller H, Ruppert A, Hartmann AM, Moller HJ, Gal A, Gieger C, Wichmann HE, Illig T, Dahmen N, Winterer G. 2010. Association of a variant in the muscarinic acetylcholine receptor 2 gene (CHRM2) with nicotine addiction. Am J Med Genet B Neuropsychiatr Genet 153B:684-690.

Moore S, Montane-Jaime K, Shafe S, Joseph R, Crooks H, Carr LG, Ehlers CL. 2007. Association of ALDH1 promoter polymorphisms with alcohol-related phenotypes in trinidad and tobago. J Stud Alcohol Drugs 68:192-196.

National Society of Genetic Counselors' Definition Task Force, Resta R, Biesecker BB, Bennett RL, Blum S, Hahn SE, Strecker MN, Williams JL. 2006. A new definition of genetic counseling: National society of genetic counselors' task force report. J Genet Couns 15:77-83.

Nelson EC, Heath AC, Madden PA, Cooper ML, Dinwiddie SH, Bucholz KK, Glowinski A, McLaughlin T, Dunne MP, Statham DJ, Martin NG. 2002. Association between self-reported childhood sexual abuse and adverse psychosocial outcomes: Results from a twin study. Arch Gen Psychiatry 59:139-145.

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

Newlin DB, Miles DR, van den Bree MB, Gupman AE, Pickens RW. 2000. Environmental transmission of DSM-IV substance use disdorders in adoptive and step families. Alcohol Clin Exp Res 24:1785-1794.

Noble EP, Blum K, Ritchie T, Montgomery A, Sheridan PJ. 1991. Allelic association of the D2 dopamine receptor gene with receptor-binding characteristics in alcoholism. Arch Gen Psychiatry 48:648-654.

Noble EP, Zhang X, Ritchie T, Lawford BR, Grosser SC, Young RM, Sparkes RS. 1998. D2 dopamine receptor and GABA(A) receptor beta3 subunit genes and alcoholism. Psychiatry Res 81:133-147.

Nurnberger JI,Jr, Foroud T, Flury L, Meyer ET, Wiegand R. 2002. Is there a genetic relationship between alcoholism and depression? Alcohol Res Health 26:233-240.

Nurnberger JI,Jr, Foroud T, Flury L, Su J, Meyer ET, Hu K, Crowe R, Edenberg H, Goate A, Bierut L, Reich T, Schuckit M, Reich W. 2001. Evidence for a locus on chromosome 1 that influences vulnerability to alcoholism and affective disorder. Am J Psychiatry 158:718-724.

Odgers CL, Milne BJ, Caspi A, Crump R, Poulton R, Moffitt TE. 2007. Predicting prognosis for the conduct-problem boy: Can family history help? J Am Acad Child Adolesc Psychiatry 46:1240-1249.

Oota H, Pakstis AJ, Bonne-Tamir B, Goldman D, Grigorenko E, Kajuna SL, Karoma NJ, Kungulilo S, Lu RB, Odunsi K, Okonofua F, Zhukova OV, Kidd JR, Kidd KK.

2004. The evolution and population genetics of the ALDH2 locus: Random genetic drift, selection, and low levels of recombination. Ann Hum Genet 68:93-109.

Palmer RH, McGeary JE, Francazio S, Raphael BJ, Lander AD, Heath AC, Knopik VS. 2012. The genetics of alcohol dependence: Advancing towards systems-based approaches. Drug Alcohol Depend .

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. 2010. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42:570-575.

Peay HL, Sheidley BR. 2008. Principles of genetic counseling. In: Smoller JSW, Sheidley BR, Tsuang MT, editors. Psychiatric genetics: Applications in clinical practice. Washington, D.C., London: American Psychiatric Publishing. p 27-46.

Peay HL, Veach PM, Palmer CG, Rosen-Sheidley B, Gettig E, Austin JC. 2008. Psychiatric disorders in clinical genetics I: Addressing family histories of psychiatric illness. J Genet Couns 17:6-17.

Pencina MJ, D'Agostino RB S, D'Agostino RB,Jr, Vasan RS. 2008. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Stat Med 27:157-72; discussion 207-12.

Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. 2004. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 159:882-890.

Perala J, Suvisaari J, Saarni SI, Kuoppasalmi K, Isometsa E, Pirkola S, Partonen T, Tuulio-Henriksson A, Hintikka J, Kieseppa T, Harkanen T, Koskinen S, Lonnqvist J. 2007. Lifetime prevalence of psychotic and bipolar I disorders in a general population. Arch Gen Psychiatry 64:19-28.

Pirooznia M, Seifuddin F, Judy J, Mahon PB, Bipolar Genome Study (BiGS) Consortium, Potash JB, Zandi PP. 2012. Data mining approaches for genome-wide association of mood disorders. Psychiatr Genet 22:55-61.

Prescott CA, Kendler KS. 1999. Genetic and environmental contributions to alcohol abuse and dependence in a population-based sample of male twins. Am J Psychiatry 156:34-40.

Prescott CA, Sullivan PF, Kuo PH, Webb BT, Vittum J, Patterson DG, Thiselton DL, Myers JM, Devitt M, Halberstadt LJ, Robinson VP, Neale MC, van den Oord EJ, Walsh D, Riley BP, Kendler KS. 2006. Genomewide linkage study in the irish affected sib pair study of alcohol dependence: Evidence for a susceptibility region for symptoms of alcohol dependence on chromosome 4. Mol Psychiatry 11:603-611.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575.

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P, International Schizophrenia Consortium. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748-752.

Pyeritz RE. 2012. The family history: The first genetic test, and still useful after all those years? Genet Med 14:3-9.

Radouco-Thomas S, Garcin F, Laperriere A, Marquis PA, Lambert J, Denver J, Lacerte M, Lacroix D, Radouco-Thomas C. 1979. Genetic epidemiology and the prevention of functional mental disorders and alcoholism: Family study and biological predictors. Prog Neuropsychopharmacol 3:165-189.

Rehm J, Mathers C, Popova S, Thavorncharoensap M, Teerawattananon Y, Patra J. 2009. Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders. Lancet 373:2223-2233.

Reich T, Edenberg HJ, Goate A, Williams JT, Rice JP, Van Eerdewegh P, Foroud T, Hesselbrock V, Schuckit MA, Bucholz K, Porjesz B, Li TK, Conneally PM, Nurnberger JI,Jr, Tischfield JA, Crowe RR, Cloninger CR, Wu W, Shears S, Carr K, Crose C, Willig C, Begleiter H. 1998. Genome-wide search for genes affecting the risk for alcohol dependence. Am J Med Genet 81:207-215.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516-1517.

Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. 2001. Validation of the gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 93:358-366.

Ruderfer DM, Korn J, Purcell SM. 2010. Family-based genetic risk prediction of multifactorial disease. Genome Med 2:2.

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

Sander T, Harms H, Podschus J, Finckh U, Nickel B, Rolfs A, Rommelspacher H, Schmidt LG. 1995. Dopamine D1, D2 and D3 receptor genes in alcohol dependence. Psychiatr Genet 5:171-176.

Sander T, Ladehoff M, Samochowiec J, Finckh U, Rommelspacher H, Schmidt LG. 1999. Lack of an allelic association between polymorphisms of the dopamine D2 receptor gene and alcohol dependence in the german population. Alcohol Clin Exp Res 23:578-581.

Sartor CE, Lynskey MT, Bucholz KK, McCutcheon VV, Nelson EC, Waldron M, Heath AC. 2007. Childhood sexual abuse and the course of alcohol dependence development: Findings from a female twin sample. Drug Alcohol Depend 89:139-144.

Schuckit MA, Edenberg HJ, Kalmijn J, Flury L, Smith TL, Reich T, Bierut L, Goate A, Foroud T. 2001. A genome-wide search for genes that relate to a low level of response to alcohol. Alcohol Clin Exp Res 25:323-329.

Schumann G, Coin LJ, Lourdusamy A, Charoen P, Berger KH, Stacey D, Desrivieres S, Aliev FA, Khan AA, Amin N, Aulchenko YS, Bakalkin G, Bakker SJ, Balkau B, Beulens JW, Bilbao A, de Boer RA, Beury D, Bots ML, Breetvelt EJ, Cauchi S, Cavalcanti-Proenca C, Chambers JC, Clarke TK, Dahmen N, de Geus EJ, Dick D, Ducci F, Easton A, Edenberg HJ, Esko T, Fernandez-Medarde A, Foroud T, Freimer NB, Girault JA, Grobbee DE, Guarrera S, Gudbjartsson DF, Hartikainen AL, Heath AC, Hesselbrock V, Hofman A, Hottenga JJ, Isohanni MK, Kaprio J, Khaw KT, Kuehnel B, Laitinen J, Lobbens S, Luan J, Mangino M, Maroteaux M, Matullo G, McCarthy MI, Mueller C, Navis G, Numans ME, Nunez A, Nyholt DR, Onland-Moret CN, Oostra BA, O'Reilly PF, Palkovits M, Penninx BW, Polidoro S, Pouta A, Prokopenko I, Ricceri F, Santos E, Smit JH, Soranzo N, Song K, Sovio U, Stumvoll M, Surakk I, Thorgeirsson TE, Thorsteinsdottir U, Troakes C, Tyrfingsson T, Tonjes A, Uiterwaal CS, Uitterlinden AG, van der Harst P, van der

Schouw YT, Staehlin O, Vogelzangs N, Vollenweider P, Waeber G, Wareham NJ, Waterworth DM, Whitfield JB, Wichmann EH, Willemsen G, Witteman JC, Yuan X, Zhai G, Zhao JH, Zhang W, Martin NG, Metspalu A, Doering A, Scott J, Spector TD, Loos RJ, Boomsma DI, Mooser V, Peltonen L, Stefansson K, van Duijn CM, Vineis P, Sommer WH, Kooner JS, Spanagel R, Heberlein UA, Jarvelin MR, Elliott P. 2011. Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. Proc Natl Acad Sci U S A 108:7119-7124.

Sintov ND, Kendler KS, Young-Wolff KC, Walsh D, Patterson DG, Prescott CA. 2010. Empirically defined subtypes of alcohol dependence in an irish family sample. Drug Alcohol Depend 107:230-236.

Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. 2012. Comparisons of established risk prediction models for cardiovascular disease: Systematic review. BMJ 344:e3318.

Song J, Koller DL, Foroud T, Carr K, Zhao J, Rice J, Nurnberger JI,Jr, Begleiter H, Porjesz B, Smith TL, Schuckit MA, Edenberg HJ. 2003. Association of GABA(A) receptors and alcohol dependence and the effects of genetic imprinting. Am J Med Genet B Neuropsychiatr Genet 117B:39-45.

Soyka M, Preuss UW, Hesselbrock V, Zill P, Koller G, Bondy B. 2008. GABA-A2 receptor subunit gene (GABRA2) polymorphisms and risk for alcohol dependence. J Psychiatr Res 42:184-191.

Spiegelman D, Colditz GA, Hunter D, Hertzmark E. 1994. Validation of the gail et al. model for predicting individual breast cancer risk. J Natl Cancer Inst 86:600-607.

Spitalnic S. 2004. Test properties 2: Likelihood ratios, bayes' formula, and receiver operating characteristic curves. Hospital Physician 40:53-58.

Stacey D, Clarke TK, Schumann G. 2009. The genetics of alcoholism. Curr Psychiatry Rep 11:364-369.

Stavro K, Pelletier J, Potvin S. 2012. Widespread and sustained cognitive deficits in alcoholism: A meta-analysis. Addict Biol .

Strat YL, Ramoz N, Schumann G, Gorwood P. 2008. Molecular genetics of alcohol dependence and related endophenotypes. Curr Genomics 9:444-451.

Suarez BK, Parsian A, Hampe CL, Todd RD, Reich T, Cloninger CR. 1994. Linkage disequilibria at the D2 dopamine receptor locus (DRD2) in alcoholics and controls. Genomics 19:12-20.

Sullivan PF. 2010. The psychiatric GWAS consortium: Big science comes to psychiatry. Neuron 68:182-186.

Sullivan PF, Kendler KS, Neale MC. 2003. Schizophrenia as a complex trait: Evidence from a meta-analysis of twin studies. Arch Gen Psychiatry 60:1187-1192.

Sullivan PF, Neale MC, Kendler KS. 2000. Genetic epidemiology of major depression: Review and meta-analysis. Am J Psychiatry 157:1552-1562.

Talmud PJ, Hingorani AD, Cooper JA, Marmot MG, Brunner EJ, Kumari M, Kivimaki M, Humphries SE. 2010. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. BMJ 340:b4838.

Treutlein J, Cichon S, Ridinger M, Wodarz N, Soyka M, Zill P, Maier W, Moessner R, Gaebel W, Dahmen N, Fehr C, Scherbaum N, Steffens M, Ludwig KU, Frank J, Wichmann HE, Schreiber S, Dragano N, Sommer WH, Leonardi-Essmann F, Lourdusamy A, Gebicke-Haerter P, Wienker TF, Sullivan PF, Nothen MM, Kiefer F, Spanagel R, Mann K, Rietschel M. 2009. Genome-wide association study of alcohol dependence. Arch Gen Psychiatry 66:773-784.

Treutlein J, Rietschel M. 2011a. Genome-wide association studies of alcohol dependence and substance use disorders. Curr Psychiatry Rep 13:147-155.

Treutlein J, Rietschel M. 2011b. Genome-wide association studies of alcohol dependence and substance use disorders. Curr Psychiatry Rep 13:147-155.

Turner E, Ewing J, Shilling P, Smith TL, Irwin M, Schuckit M, Kelsoe JR. 1992. Lack of association between an RFLP near the D2 dopamine receptor gene and severe alcoholism. Biol Psychiatry 31:285-290.

Tuszynski J. 2011. caTools: Tools: Moving window statistics, GIF, Base64, ROC AUC, etc.. R package version 1.12. http://CRAN.R-project.org/package=caTools.

van der Net JB, Janssens AC, Sijbrands EJ, Steyerberg EW. 2009. Value of genetic profiling for the prediction of coronary heart disease. Am Heart J 158:105-110.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. Am J Hum Genet 90:7-24.

Visscher PM, Yang J, Goddard ME. 2010. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by yang et al. (2010). Twin Res Hum Genet 13:517-524.

Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, Thun MJ, Cox DG, Hankinson SE, Kraft P, Rosner B, Berg CD, Brinton LA, Lissowska J, Sherman ME, Chlebowski R, Kooperberg C, Jackson RD, Buckman DW, Hui P, Pfeiffer R, Jacobs KB, Thomas GD, Hoover RN, Gail MH, Chanock SJ, Hunter DJ. 2010. Performance of common genetic variants in breast-cancer risk models. N Engl J Med 362:986-993.

Wang JC, Grucza R, Cruchaga C, Hinrichs AL, Bertelsen S, Budde JP, Fox L, Goldstein E, Reyes O, Saccone N, Saccone S, Xuei X, Bucholz K, Kuperman S, Nurnberger J,Jr, Rice JP, Schuckit M, Tischfield J, Hesselbrock V, Porjesz B, Edenberg HJ, Bierut LJ, Goate AM. 2009. Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence. Mol Psychiatry 14:501-510.

Wang JC, Hinrichs AL, Stock H, Budde J, Allen R, Bertelsen S, Kwon JM, Wu W, Dick DM, Rice J, Jones K, Nurnberger JI,Jr, Tischfield J, Porjesz B, Edenberg HJ, Hesselbrock V, Crowe R, Schuckit M, Begleiter H, Reich T, Goate AM, Bierut LJ. 2004. Evidence of common and specific genetic effects: Association of the muscarinic acetylcholine receptor M2 (CHRM2) gene with alcohol dependence and major depressive syndrome. Hum Mol Genet 13:1903-1911.

Wang KS, Liu X, Zhang Q, Pan Y, Aragam N, Zeng M. 2011. A meta-analysis of two genome-wide association studies identifies 3 new loci for alcohol dependence. J Psychiatr Res 45:1419-1425.

Wang KS, Liu X, Zhang Q, Wu LY, Zeng M. 2012. Genome-wide association study identifies 5q21 and 9p24.1 (KDM4C) loci associated with alcohol withdrawal symptoms. J Neural Transm 119:425-433.

Weissman MM, Bland RC, Canino GJ, Faravelli C, Greenwald S, Hwu HG, Joyce PR, Karam EG, Lee CK, Lellouch J, Lepine JP, Newman SC, Rubio-Stipec M, Wells JE, Wickramaratne PJ, Wittchen H, Yeh EK. 1996. Cross-national epidemiology of major depression and bipolar disorder. JAMA 276:293-299.

White R, Lalouel JM, Leppert M, Lathrop M, Nakamura Y, O'Connell P. 1989. Linkage maps of human chromosomes. Genome 31:1066-1072.

Whitfield JB. 2002. Alcohol dehydrogenase and alcohol dependence: Variation in genotype-associated risk between populations. Am J Hum Genet 71:1247-50; author reply 1250-1.

Wilde A, Meiser B, Mitchell PB, Hadzi-Pavlovic D, Schofield PR. 2011. Community interest in predictive genetic testing for susceptibility to major depressive disorder in a large national sample. Psychol Med 41:1605-1613.

Wilde A, Meiser B, Mitchell PB, Schofield PR. 2010. Public interest in predictive genetic testing, including direct-to-consumer testing, for susceptibility to major depression: Preliminary findings. Eur J Hum Genet 18:47-51.

Williams TJ, LaForge KS, Gordon D, Bart G, Kellogg S, Ott J, Kreek MJ. 2007. Prodynorphin gene promoter repeat associated with cocaine/alcohol codependence. Addict Biol 12:496-502.

Wong ML, Dong C, Andreev V, Arcos-Burgos M, Licinio J. 2012. Prediction of susceptibility to major depression by a model of interactions of multiple functional genetic variants and environmental factors. Mol Psychiatry 17:624-633.

World Health Organization. 2004. Global status report on alcohol 2004. Rep. World Health Organization, Geneva.

World Health Organization. 2008. ICD-10: International statistical classification of diseases and related health problems (10th rev. ed.). New York, NY.

Wray NR, Goddard ME, Visscher PM. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 17:1520-1528.

Wray NR, Goddard ME, Visscher PM. 2008. Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev 18:257-263.

Wray NR, Yang J, Goddard ME, Visscher PM. 2010. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 6:e1000864.

Xuei X, Dick D, Flury-Wetherill L, Tian HJ, Agrawal A, Bierut L, Goate A, Bucholz K, Schuckit M, Nurnberger J,Jr, Tischfield J, Kuperman S, Porjesz B, Begleiter H,

Foroud T, Edenberg HJ. 2006. Association of the kappa-opioid system with alcohol dependence. Mol Psychiatry 11:1016-1024.

Xuei X, Flury-Wetherill L, Bierut L, Dick D, Nurnberger J,Jr, Foroud T, Edenberg HJ. 2007. The opioid system in alcohol and drug dependence: Family-based association study. Am J Med Genet B Neuropsychiatr Genet 144B:877-884.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565-569.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: A tool for genome-wide complex trait analysis. Am J Hum Genet 88:76-82.

Young-Wolff KC, Chereji E, Prescott CA. 2012. Heritability of alcohol dependence is similar in women and men. Evid Based Ment Health 15:57.

Zhu M, Zhao S. 2007. Candidate gene identification approach: Progress and challenges. Int J Biol Sci 3:420-427.

Zuo L, Gelernter J, Zhang CK, Zhao H, Lu L, Kranzler HR, Malison RT, Li CS, Wang F, Zhang XY, Deng HW, Krystal JH, Zhang F, Luo X. 2012. Genome-wide association study of alcohol dependence implicates KIAA0040 on chromosome 1q. Neuropsychopharmacology 37:557-566.

Zuo L, Zhang CK, Wang F, Li CS, Zhao H, Lu L, Zhang XY, Lu L, Zhang H, Zhang F, Krystal JH, Luo X. 2011. A novel, functional and replicable risk gene region for alcohol dependence identified by genome-wide association study. PLoS One 6:e26726.

Appendix

COGA AND SAGE ACKNOWLEDGMENTS

The Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, H. Edenberg, L. Bierut, includes ten different centers: University of Connecticut (V. Hesselbrock); Indiana University (H.J. Edenberg, J. Nurnberger Jr., T. Foroud); University of Iowa (S. Kuperman, J. Kramer); SUNY Downstate (B. Porjesz); Washington University in St. Louis (L. Bierut, A. Goate, J. Rice, K. Bucholz); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield); Southwest Foundation (L. Almasy), Howard University (R. Taylor) and Virginia Commonwealth University (D. Dick). Other COGA collaborators contributing to the GWAS data set include: L. Bauer (University of Connecticut); D. Koller, S. O'Connor, L. Wetherill, X. Xuei (Indiana University); Grace Chan (University of Iowa); N. Manz, M. Rangaswamy (SUNY Downstate); A. Hinrichs, J. Rohrbaugh, J-C Wang (Washington University in St. Louis); A. Brooks (Rutgers University); and F. Aliev (Virginia Commonwealth University). A. Parsian and M. Reilly are the NIAAA Staff Collaborators. We continue to be inspired by our memories of Henri Begleiter and Theodore Reich, founding PI and Co-PI of COGA, and also owe a debt of gratitude to other past organizers of COGA, including Ting-Kai Li, currently a consultant with

COGA, P. Michael Conneally, Raymond Crowe, and Wendy Reich, for their critical contributions. This national collaborative study is supported by NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA). (2/2010)

The COGA GWAS sample genotyping was funded through U10 AA008401 from the NIAAA, National Institutes of Health, Bethesda, MD, USA, U01HG004438, and NIH GEI, and HHSN268200782096C, NIH contract "High throughput genotyping for studying the genetic contributions to human disease". The genotyping center was the Johns Hopkins University Center for Inherited Disease Research (CIDR), Baltimore, MD, USA

SAGE is part of the Gene Environment Association Studies initiative (GENEVA) funded by the National Human Genome Research Institute. Funding sources are: U01 HG004422. National Human Genome Research Institute, Bethesda, MD, USA; U10 AA008401. National Institute on Alcohol Abuse and Alcoholism and National Institute on Drug Abuse, Bethesda, MD, USA; P01 CA089392. National Cancer Institute, Bethesda, MD, USA; R01 DA013423. National Institute on Drug Abuse, Bethesda, MD, USA; and R01 DA019963. National Institute on Drug Abuse, Bethesda, MD, USA

Funding sources did not have a role in the design, conduct, or reporting of these studies.

Vita

Jia Yan was born on January 18, 1983 in Dong Sheng, Inner Mongolia, China and is an American citizen. She graduated from West Windsor-Plainsboro High School South in 2001. She received a B.A. from Rutgers University in New Brunswick, NJ with majors in Genetics and Psychology, and worked for SERV Behavioral Health System, Inc. in Cranbury, NJ as a counselor in psychiatric group homes prior to starting graduate school. In 2007, she joined the Department of Human and Molecular Genetics at Virginia Commonwealth University in Richmond, VA to work towards an M.S. in Genetic Counseling and Ph.D. in Human and Molecular Genetics.