Copy Number Variations in 6q14.1 and 5q13.2 are Associated with Alcohol Dependence


Background: Excessive alcohol use is the third leading cause of preventable death and is highly correlated with alcohol dependence, a heritable phenotype. Many genetic factors for alcohol dependence have been found, but many remain unknown. In search of additional genetic factors, we examined the association between Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) alcohol dependence and all common copy number variations (CNVs) with good reliability in the Study of Addiction: Genetics and Environment (SAGE).

Methods: All participants in SAGE were interviewed using the Semi-Structured Assessment for the Genetics of Alcoholism, as a part of 3 contributing studies. A total of 2,610 non-Hispanic European American samples were genotyped on the Illumina Human 1M array. We performed CNV calling by CNVPartition, PennCNV, and QuantiSNP, and only CNVs identified by all 3 software programs were examined. Association was conducted with the CNV (as a deletion/duplication) as well as with probes in the CNV region. Quantitative polymerase chain reaction (qPCR) was used to validate the CNVs in the laboratory.

Results: CNVs in 6q14.1 ($p = 1.04 \times 10^{-5}$) and 5q13.2 ($p = 3.37 \times 10^{-5}$) were significantly associated with alcohol dependence after adjusting multiple tests. On chromosome 5q13.2, there were multiple candidate genes previously associated with various neurological disorders. The region on chromosome 6q14.1 is a gene desert that has been associated with mental retardation and language delay. The CNV in 5q13.2 was validated, whereas only a component of the CNV on 6q14.1 was validated by qPCR. Thus, the CNV on 6q14.1 should be viewed with caution.

Conclusions: This is the first study to show an association between DSM-IV alcohol dependence and CNVs. CNVs in regions previously associated with neurological disorders may be associated with alcohol dependence.

Key Words: Copy Number Variations, Alcohol Dependence, CNV Accuracy.

During 2001 to 2005, excessive alcohol use contributed to about 79,000 deaths and 2.3 million years of potential life lost in the United States (Kanny et al., 2011). Excessive alcohol consumption, the third leading cause of preventable death in the United States, can cause damage to the central and peripheral nervous system, and to nearly every organ system in the body (Cuan and De Belleroche, 2002; Testino, 2008). It is also strongly correlated with alcohol dependence, a serious psychiatric disorder that affects about 12% of American adults across their lifetime (Hasin et al., 2007). Alcohol dependence is a common, complex disease characterized by compulsive and uncontrolled alcohol consumption despite its negative effects on the drinker’s health, relationships, productivity, and social standing.

There is robust evidence for heritable influences on the liability to alcohol dependence (Bierut et al., 1998). Siblings of alcohol-dependent individuals have a 3- to 8-fold increased risk of developing alcoholism (Reich et al., 1998) with twin studies revealing the heritability of alcohol dependence to be $\approx 50\%$ (Heath et al., 1997; Kendler et al., 1994; Knopik et al., 2004). Given its serious public health impact (Braillon and Dubois, 2005) and the strong evidence for its biological underpinnings, numerous linkage and association studies have been targeted at gene identification for alcohol dependence (Bierut et al., 2010; Edenberg and Foroud, 2006; Hill et al., 2004; Long et al., 1998; Reich et al., 1998). Recently, several genome-wide association studies (GWAS) queried the genome for association (Bierut et al., 2010; Edenberg et al., 2010; Heath et al., 2011; Treutlein et al., 2007).
COPY NUMBER VARIATIONS WITH ALCOHOL DEPENDENCE

2009). Results surpassed genome-wide significance in 1 study of early-onset male alcoholics (Treutlein et al., 2009), but across the multiple efforts, effect sizes were small and did not replicate. This has generated considerable interest in the examination of other possible contributors to the “missing heritability” for alcohol dependence. One such contributor is copy number variations (CNVs).

CNVs are duplications or deletions of a particular segment of an individual’s genome and reflect inherent structural instability in the architecture of the genome. They are prevalent forms of common genetic variation and can have a substantial influence on gene expression levels (Perry et al., 2007). For instance, Mendelian disorders such as Williams–Beuren syndrome (caused by a deletion at chromosome region 7q11.23) and Charcot–Marie–Tooth neuropathy type 1A (caused by duplication at chromosome region 17p11.2) (Krajewski et al., 2000; Martens et al., 2008) are attributable to CNVs. Despite the deleterious effects of CNVs and their links to disease, few studies have examined CNVs in the context of psychiatric illness, particularly alcohol dependence. This is primarily because of the inherent challenges involved in the identification of what constitutes a CNV. While traditional methods of CNV identification involve laboratory-based experiments, they can also be identified (or “called”) using GWAS data where a series of single-nucleotide polymorphisms (SNPs) or “intensity” probes are interrogated for their occurrence in a state other than the expected disomic (i.e., 2 copy) state. Typically, the intensity of the probe signal that is expected when 2 copies of the probe are present is compared with the observed intensity that is expected to be enhanced for duplications, or suppressed for deletions. These probes are routinely included in GWAS chips, and thus, as GWAS technology became more accessible, there was an upswing in CNV identification efforts. However, this method of CNV calling from GWAS microarrays can be associated with relatively high error rates. For instance, in a previous study, we demonstrated the relatively modest concordance in CNV detection using 3 widely utilized software packages with varying algorithms. In that study, we implemented statistical measures that enhance the reliability of the detected CNVs using multiple algorithms and further validated the CNVs identified using statistical programs by quantitative polymerase chain reaction (qPCR) in the laboratory (Lin et al., 2011).

Other challenges of CNV detection include (i) size of the CNV, with smaller CNVs (<10 kb) being harder to detect, (ii) number of CNV probes in the region of the CNV, with fewer probes resulting in greater noise, (iii) the general quality of the data (including artifacts in the SNP data) and genomic waves (intensity variations in normalized GWAS data), (iv) ethnic variations, and (v) source of the sample that was genotyped—for instance, it is now well known that deletions and duplications can arise in DNA drawn from cell lines (i.e., extracting cells from a DNA source and maintaining them in laboratory cultures to enhance longevity), and thus, CNV detection using cell cultures requires caution. Yet, if attention is paid to these challenges, CNVs represent a unique route for inquiry into the genetic architecture of alcohol dependence.

There continues to be a great deal of progress in statistical methods for CNV detection. In tandem, there is growing excitement about the association between these CNVs and human behavior, and the extent to which these intriguing variations in the human genome may contribute to that elusive “missing heritability” in complex behavioral phenotypes and psychiatric illness. While there has been some promise in studies of autism, and intellectual disabilities (Glessner et al., 2009; Pinto et al., 2010), as well as schizophrenia and bipolar disorder (Lachman et al., 2007; Stefansson et al., 2008), research on CNVs in studies of addiction, particularly alcohol dependence, is lacking. In this study, we examine the CNVs for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) alcohol dependence in a large sample of European American subjects.

MATERIALS AND METHODS

Samples

Data were drawn from the Study of Addiction: Genetics and Environment (SAGE) (Bierut et al., 2010). SAGE is 1 study of the Gene Environment Association Studies (GENEVA) project (Cornelis et al., 2010). Unrelated cases and controls for the SAGE sample were drawn from 3 contributing projects: the Collaborative Study on the Genetics of Alcoholism (COGA), the Collaborative Study on the Genetics of Nicotine Dependence (COGEND), and the Family Study of Cocaine Dependence (FSCD). While the contributing studies originally ascertained subjects for alcohol dependence (COGA), nicotine dependence (COGEND: based on a Fagerstrom Test for Nicotine Dependence score of 4 or greater in current smokers, controls being smokers), and cocaine dependence (FSCD), the subset of cases selected for genotyping in SAGE were uniformly defined as those meeting criteria for DSM-IV alcohol dependence (N = 1,899), whereas controls (N = 1,946) were individuals who reported drinking alcohol but did not meet criteria, during their lifetime, for alcohol dependence. Note that even though the parent projects were family based, the series of cases and controls for SAGE, and consequently for these analyses, were selected to be unrelated individuals. Of these, 1,186 cases and 1,397 controls are of self-reported non-Hispanic European American descent. All participants agreed to share their DNA and phenotypic information for research purposes and provided written informed consent following instructions from institutional review boards at all data collection sites.

Measures. A lifetime diagnosis of DSM-IV alcohol dependence was determined via self-reported interview information collected using the Semi-Structured Assessment for the Genetics of Alcoholism. Controls were individuals who had drunk alcohol at least once in their lifetime but did not meet criteria for alcohol dependence.

Genotyping and Quality Control

The Center for Inherited Disease Research at Johns Hopkins University genotyped all samples on the Illumina Human 1M array (Illumina, Inc., San Diego, CA). An extensive data-cleaning effort had been made to ensure data quality. These procedures included, but not limited to, using HapMap controls, detection of gender mis-annotation and chromosomal anomalies, cryptic relatedness, population structure, batch effects, Mendelian error detection, and duplication error detection. A detailed description of data-cleaning effort is described elsewhere (Bierut et al., 2010, unpublished data).
CNV Calling

The Illumina 1M array has a total of 1,072,820 probes, predominantly indexed by polymorphic SNPs. Of these probes, 23,812 are non-SNP “intensity-only” markers for CNV detection. All of the 1,072,820 probes were used for the CNV analyses. Three common programs were used to call CNVs: CNVPartition, PennCNV (Wang et al., 2007), and QuantiSNP (Colella et al., 2007). Genomic waves were also adjusted when we called CNVs by PennCNV and QuantiSNP (Diskin et al., 2008). Both PennCNV and QuantiSNP report a metric score for quality control purposes. As recommended by QuantiSNP documentation, we removed all CNV calls that had log Bayes factor <10, as well as poor-quality based on quality control measures for CNV analysis, following the approaches described in our previous work (Lin et al., 2011). In total, we genotyped 2,583 non-Hispanic European American samples in SAGE, and among them, 95 samples failed to pass quality controls for CNV analysis.

Comparative Statistics

The CNV calls from different programs were compared against each other. In our previous work, we have demonstrated that a CNV that is confirmed by all 3 CNV calling programs has a higher reproducibility rate, and thus a higher reliability (Lin et al., 2011). Therefore, we required that only CNVs detected by all 3 programs would be studied.

Association Analysis

Logistic regressions were performed on all CNV regions. After identifying potential regions, individual dummy variables for duplications and deletions were created to dissect the association signal with DSM-IV alcohol dependence. Several covariates were included in the model based on the previous GWAS of these data (Bierut et al., 2010), including sex, age, and 2 principal components indexing continuous ancestral genetic variation. We also included a dummy variable to indicate the source of DNA (cell line vs. whole blood). In addition, we ensured that these potentially confounding variables were not directly associated with the identified CNVs.

CNVs with Different Starting and Ending Points

Even when all 3 programs detect a CNV, they often report different starting and ending points for the same CNV segment, which leads to computational challenges in combining CNV reports. There is a lack of consensus in the research community regarding this issue, and therefore, in addition to studying the CNV as a deletion or duplication, we adopted an additional straightforward approach for association. First, SNP probes and intensity-only probes were used to detect CNVs by multiple programs. Second, a change of copy number at a particular probe was considered detected when all CNV programs reported CNV segments that cover the probe. Third, association between alcohol dependence and each probe (assigned the same copy number as the CNV) was examined. For instance, if a CNV (duplication or deletion) was detected in region X, using probes (SNP or intensity probes) A, B, C, D, E, F, and G by 3 different programs, then the results from the 3 programs for each probe were compared against each other (see Fig. S1). If agreement was reached among the 3 programs, then the CNV for these probes (probes D and E) were confirmed and would be used in the following analysis. If there was disagreement among the 3 programs, then a missing value was assigned to these probes (probes B, C, and F).

Validation

CNVs identified by 3 independent programs were validated in subjects carrying the variant with qPCR. We selected a TaqMan CNV probe in the target region. The probe was presigned by Applied Biosystems (Applied Biosystems, Foster City, CA). Genomic DNA was analyzed with real-time PCR using an ABI-7900 Fast PCR system. Each real-time PCR run included within-plate duplicates. Correction for sample-to-sample variation was done by simultaneously amplifying a standard CNV reference assay, RNAse P. Real-time data were analyzed using the comparative C method (Muller et al., 2002). The C values of each sample were normalized with the C value for the RNAse P assay. Only the samples with a standard error <0.15 were analyzed. Copy numbers were calculated using ABI CopyCallerTM Software v1.0.

RESULTS

Sociodemographic Characteristics

Of the 2,583 non-Hispanic European American samples from SAGE, 95 failed to pass quality control for CNV analysis, leaving 1,140 cases and 1,348 controls. The mean age among subjects with alcohol dependence was 38.2 (SD = 10.0), and for controls, it was 39.0 (SD = 9.5). Sixty percent of the cases and 29.2% of the controls were men. As shown in Table 1, cases were more likely than controls to be dependent on nicotine and illicit drugs, including nicotine, cocaine, and marijuana. They were also more likely to meet criteria for a lifetime history of conduct disorder and major depression.

Alcohol History

Cases also reported an earlier age of heavy and regular alcohol use and, by definition, reported more alcohol symptoms (Table 2).

CNV Detection

Of the samples that passed quality control, we identified 1,139 CNV regions with length >50 kb and number of probes not <10 (Lin et al., 2011). Among them, only 141 CNV regions have frequency higher than 1%. All of these CNV regions had previously been documented in the database of genetic variants (Zhang et al., 2006). Level of agreement across the various algorithms was moderate (see Table S1). Thus, after adjusting for multiple tests, our significance threshold for association analyses is 0.05/141 = 3.54 × 10⁻⁴.

Association Between CNV's and Alcohol Dependence

Two CNV regions were significantly associated with alcohol dependence (Table 3): chromosome 6q14.1 (OR = 2.86, p = 1.04 × 10⁻⁶, n = 121 subjects with the duplication) and chromosome 5q13.2 (OR = 1.99, p = 3.37 × 10⁻⁴, n = 59 subjects with the duplication, and n = 58 subjects with deletions). The p-values for each probe in these 2 regions are listed in Tables S2 and S3. There were significant differences in the distribution of men and women across cases and
controls. While all analyses modeled sex as a covariate, we also conducted secondary analyses separately in men and women. Significant results for SNPs in the region of the CNV on 5q13.2 and 6q14.1 were replicated with the same direction of effect (albeit at less significant p-values; see Tables S4A, B and S5A, B).

Validation Using qPCR

For the CNV at 5q13.2, over 97% of these CNVs (called by all 3 algorithms) were confirmed as true CNVs using qPCR. Only 85% were validated when the CNV was called by 2 of 3 algorithms, and 38% when only 1 of 3 algorithms called it. Redoing analyses with successfully validated CNVs yielded identical results. However, for 6q14.1, while all deletions were confirmed, none of the duplications were reproduced via qPCR. This suggests that the result for 6q14.1 should be viewed with caution.

DISCUSSION

These analyses evaluated the association between CNVs and alcohol dependence among a relatively large sample of alcohol-dependent cases and nondependent alcohol-exposed controls. We found 2 regions significantly associated with alcohol dependence: Chr5: 68,921,426-70,412,247 and Chr6: 79,034,386-79,090,197. To our knowledge, this is the first study to connect CNVs and alcohol dependence.

The identified chromosomal regions have been previously associated with several neurological and other disorders.

<table>
<thead>
<tr>
<th>Table 1. Sociodemographic Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total no. (male)</td>
</tr>
<tr>
<td>Alcohol dependence</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Disorder and other psychiatric diagnoses were categorized by Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.

<table>
<thead>
<tr>
<th>Table 2. Alcohol History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of onset of regular drinking</td>
</tr>
<tr>
<td>Alcohol dependence</td>
</tr>
<tr>
<td>No dependence</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Regular drinking is defined as drinking once a month for 6 months or more.

<table>
<thead>
<tr>
<th>Table 3. Associations of CNVs and Alcohol Dependence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Chr5: 68,921,426-70,412,247</td>
</tr>
<tr>
<td>Chr6: 79,034,386-79,090,197</td>
</tr>
<tr>
<td>Chr6: 79,034,386-79,090,197</td>
</tr>
<tr>
<td>Chr6: 79,034,386-79,090,197</td>
</tr>
<tr>
<td>Chr6: 79,034,386-79,090,197</td>
</tr>
</tbody>
</table>

The starting and ending points are defined by the probes whose p-values have a clear deviation from the rest.

The p-value of this region is annotated by the most significant probe in this CNV.

This CNV was analyzed as ≤ 1 = deletion, 2 = normal diploid, and ≥ 3 = duplication. For cases: deletions: 4/22 had a dosage score of 0.667 (18/22 had a score of 1, or heterozygous for deletion); duplications: 4/22 had a dosage score of 3 (heterozygous), and the remainder had a mean score of 3.3. For controls: deletions: 11/36 had a dosage score of 0.667, remainder scoring 1; duplications: 13/22 had a dosage score of 3, 2/22 had a score of 3.3, and 1/22 had a score of 4 (homozygous duplication). CNVs, copy number variations.

OCLN is located <30 kb downstream of this region.
Chr5: 68,921,426-70,412,247 covers several genes, including *SMA4*, *SFR1*, *SFR1B*, *SMN2*, *SMA3*, *NAIP*, *GTF2H2*, *GTF2H2D*, and the downstream *OCLN*. Among them, *SMA3*, *SMA4*, and *SMN2* are known to be associated with spinal muscular atrophy (Also-Rallo et al., 2011; Wong et al., 2007). Recent research shows that the genes in this region have a function in the nervous system (Dachs et al., 2011), including *OCLN*, another candidate in this region, which is an integral membrane protein that is required for cytokine-induced regulation of the tight junction paracellular permeability barrier. Mutations in this gene are thought to be a cause of pseudo-*TORCH* syndrome, an autosomal recessive neurologic disorder that mimics the clinical characteristics (e.g., microcephaly, seizures, spasticity) attributable to congenital infections caused by *rubella*, *cytomegalovirus*, or *herpes simplex* (O’Driscoll et al., 2010). While the CNV in Chr6: 79,034,386-79,090,197 is located in a gene desert, there is evidence that suggests a link between chromosome region 6q14.1 and mild mental retardation, language delay, and minor dysmorphisms (Lespinasse et al., 2009; Puppala et al., 2011). Also, it is hypothesized that noncoding intergenic regions, such as this, may contain regulatory elements, such as enhancers and chromosome scaffold components that are capable of changing gene expression.

We restricted our association tests to nonrare (>1%) CNVs for 2 reasons. First, the traditional GWAS design has little power to detect rare genetic variants (<1%), and the case-control study design of this project cannot provide enough power to detect rare CNVs. Second, accuracy of CNV detection diminishes with decreasing frequency.

In addition, we required that all CNVs be reproduced by 3 independent programs, a step that increases confidence in the results but that raises the potential problem of the same CNV region being detected with different starting and ending points, which results in uncertainty on how to combine these different CNV calls. To avoid this controversy, we adopted an intuitive method where we tested each genetic marker instead of a particular CNV segment. A CNV status is assigned to a particular genetic marker when all programs report a CNV that covers this probe (Tables S1 and S2). We validated our findings by qPCR—while the region on chromosome 5q13.2 was successfully validated, the duplication on 6q14.1 did not. Thus, the finding on 6q14.1 should be viewed with caution and as such, it cannot be claimed as a true CNV until laboratory validation is successful. However, we cannot be certain that the absence of experimental validation necessarily negates a CNV called using statistical algorithms—in some instances, lack of suitable laboratory probes in the region of the CNV can result in lack of experimental validation. Nonetheless, this underscores the considerable importance of experimental validation of CNVs identified using software algorithms.

The frequency of the CNV deletion on 6q14.1 is high (48.5% in cases and 49.8% in controls). Figure S2 provides the distribution of deletions and duplications from the Database of Genomic Variants (DGV). CNVs with frequencies such as those noted for 5q13.2 are not uncommon. Additionally, based on other reports in the DGV, for the 5q13.2 region, both deletions and duplications have been identified but frequencies vary dramatically by sample size and ethnicity of studied individuals. For 6q14.1, the DGV does not report on additional studies with reasonably large or diverse samples for us to validate the frequency of the observed CNV—nonetheless, the DGV notes 228 other instances where CNVs of such high frequency have been previously identified (restricting sample size in studies to be >30).

Our finding of the association between these CNVs and alcohol dependence is encouraging because it identifies regions previously associated with neurologic disorders; however, these findings will require replication. Furthermore, whether these CNVs co-segregate with alcoholism in families remains to be explored—unfortunately, despite the cases and controls being drawn from family-based data, GWAS data and biological samples on parents are currently not available. Nonetheless, our study is among the first to examine the role of CNVs in the etiology of alcohol dependence. This reflects the exciting phase of the post-GWAS genomics era where the quest to articulate the genetic architecture of serious psychiatric problems like alcohol dependence moves beyond single SNP association to new frontiers, such as CNVs and rare variants.

**ACKNOWLEDGMENTS**

Funding support for the Study of Addiction: Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01 HG004422). SAGE is one of the genome-wide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of data sets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392), and the Family Study of Cocaine Dependence (FSCD; R01 DA013423, R01 DA019963). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, and the NIH contract “High throughput genotyping for studying the genetic contributions to human disease” (HHSS268200782096C). AA also acknowledges support from ABMRF/The Foundation for Alcohol Research.
REFERENCES


SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Comparing CNV calls from different programs.

Fig. S2. Distribution of CNV frequencies from the database of genomic variants.

Table S1. Level of agreement across 3 algorithms used to call CNVs.

Table S2. Associations of each probe and alcohol dependence in Chr5: 68,921,426-70,412,247.

Table S3. Associations of each probe and alcohol dependence in Chr6: 79,034,386-79,090,197.

Table S4. (A) Associations of each probe and alcohol dependence in Chr5: 68,921,426-70,412,247 in males. (B) Associations of each probe and alcohol dependence in Chr5: 68,921,426-70,412,247 in females.

Table S5. (A) Associations of each probe and alcohol dependence in Chr6: 79,034,386-79,090,197 in males. (B) Associations of each probe and alcohol dependence in Chr6: 79,034,386-79,090,197 in females.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for content or functionality of any Supporting Information.