Genome-Wide Association Study of Copy Number Variations (CNVs) with Opioid Dependence

Dawei Li[*,1,2,3,4], **Hongyu Zhao**[5,6], **Henry R Kranzler**[7], **Ming D Li**[8], **Kevin P Jensen**[1], **Tetyana Zayats**[1], **Lindsay A Farrer**[9] **and Joel Gelernter**[1,6,10]

[1]Department of Psychiatry, School of Medicine, Yale University, New Haven, CT, USA; [2]Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT, USA; [3]Department of Computer Science, University of Vermont, Burlington, VT, USA; [4]Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, VT, USA; [5]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA; [6]Department of Genetics, School of Medicine, Yale University, New Haven, CT, USA; [7]Department of Psychiatry, University of Pennsylvania School of Medicine and VISN 4 MIRECC, Philadelphia VAMC, Philadelphia, PA, USA; [8]Department of Psychiatry and Neurobehavioral Sciences, University of Virginia, Charlottesville, VA, USA; [9]Departments of Medicine (Biomedical Genetics), Neurology, Ophthalmology, Genetics and Genomics, Biostatistics, and Epidemiology, Boston University Schools of Medicine and Public Health, Boston, MA, USA; [10]VA Connecticut Healthcare Center, Department of Neurobiology, Yale University School of Medicine, New Haven, CT, USA

Single-nucleotide polymorphisms that have been associated with opioid dependence (OD) altogether account for only a small proportion of the known heritability. Most of the genetic risk factors are unknown. Some of the 'missing heritability' might be explained by copy number variations (CNVs) in the human genome. We used Illumina HumanOmni1 arrays to genotype 5152 African-American and European-American OD cases and screened controls and implemented combined CNV calling methods. After quality control measures were applied, a genome-wide association study (GWAS) of CNVs with OD was performed. For common CNVs, two deletions and one duplication were significantly associated with OD genome-wide (eg, $P = 2 \times 10^{-8}$ and OR (95% CI) = 0.64 (0.54–0.74) for a chromosome 18q12.3 deletion). Several rare or unique CNVs showed suggestive or marginal significance with large effect sizes. This study is the first GWAS of OD using CNVs. Some identified CNVs harbor genes newly identified here to be of biological importance in addiction, whereas others affect genes previously known to contribute to substance dependence risk. Our findings augment our specific knowledge of the importance of genomic variation in addictive disorders, and provide an addiction CNV pool for further research. These findings require replication.

Neuropsychopharmacology (2015) **40,** 1016–1026; doi:10.1038/npp.2014.290; published online 19 November 2014

INTRODUCTION

Substance dependence (SD) is a set of common, often chronic, psychiatric disorders characterized by physical and psychological addiction to alcohol or other drugs. In the United States, in 2001, the 1-year point prevalence of substance use disorders (excluding nicotine) was 9.35% (Grant *et al*, 2004). The consequences of substance use disorders are extremely costly to individuals and society (Li and Burmeister, 2009). Genetic factors are important in SD etiology; the major SD traits have moderate to high heritability (Goldman *et al*, 2005) based on convergent findings obtained through methodologically distinct approaches (Goldman *et al*, 2005; Li and Burmeister, 2009). Recent genome-wide association studies (GWASs) and

linkage studies have identified several regions harboring genes associated with addiction to various substances, including alcohol (Gelernter *et al*, 2014a), nicotine (Li and Burmeister, 2009), cocaine (Gelernter *et al*, 2014c), and opioids (Gelernter *et al*, 2014b). However, all of the single-nucleotide polymorphisms (SNPs) that have been identified as associated with SD account for a small proportion (2–15%) of known heritable risks for developing the disorders (Manolio *et al*, 2009). We hypothesized that some of the 'missing heritability' in SD might be explained by copy number variations (CNVs). In the present article, we focus this approach on the etiology of opioid dependence (OD). Opioids are among the most addictive substances known and OD is moderately heritable ($h^2 \sim 65\%$) (Goldman *et al*, 2005).

CNV, one type of structural variation, is the gain or loss of a relatively lengthy segment of DNA sequence. CNVs occur in the healthy human genome (Iafrate *et al*, 2004; Sebat *et al*, 2004) and 8% of individuals have a CNV of > 500 kilobase pairs (kb) (Itsara *et al*, 2009). The nucleotides encompassed by the CNVs annotated in the Database of Genomic Variants (DGV) (Iafrate *et al*, 2004)

*Correspondence: Dr D Li, Department of Microbiology and Molecular Genetics, University of Vermont, 95 Carrigan Drive, Stafford Hall, Burlington, VT 05405, USA, Tel: +1 802 656 9838, Fax: +1 802 656 8749, E-mail: dawei.li@uvm.edu

cover 35% of the total nucleotides of the human genome (Zhang et al, 2009); in comparison, SNPs account for <1% (although many CNVs have overestimated boundaries, and a greater proportion of smaller CNVs (<30 kb) are predicted to remain unidentified). The regions residing in CNVs include functional genes involved in the regulation of cell growth and metabolism, implicating vital roles for CNVs in the variability in human traits, disease risk, and evolution (Iafrate et al, 2004). CNVs can be familial (heritable) or de novo, contributing to the development of Mendelian, sporadic, and complex diseases (Zhang et al, 2009). For example, CNVs are responsible in part for the emergence of advantageous human traits such as cognitive capacity and endurance running (Dumas et al, 2007; Lupski, 2007), explained by evolutionary selection or genetic drift (Nguyen et al, 2008). CNVs also significantly influence human diversity and the predisposition to disease, modifying the penetrance of inherited diseases (Mendelian and polygenic) and phenotypic expression of sporadic traits (Lupski, 2006). Specific CNVs may affect inflammatory response, immunity, olfactory function, cell proliferation (Schaschl et al, 2009; Young et al, 2008), and consequently clinically important phenotypic variation. CNVs have been associated with a wide variety of health problems or traits, such as autoimmune diseases (Fanciulli et al, 2007), autism (Sebat et al, 2007), schizophrenia (Stefansson et al, 2008; The International Schizophrenia Consortium, 2008), lean body mass (Hai et al, 2011), obesity (Bochukova et al, 2010; Walters et al, 2010), and HIV/AIDS susceptibility (Gonzalez et al, 2005). The same, or similar, CNVs have been observed in more than one study for more than one trait, such as the 16p11.2 and 22q deletions with autism (Kumar et al, 2008; Marshall et al, 2008; Mefford et al, 2009; Sebat et al, 2007; Weiss et al, 2008) and the 15q13.3 and 1q21.1 deletions with schizophrenia (International Schizophrenia Consortium, 2008; Stefansson et al, 2008; Vrijenhoek et al, 2008; Wilson et al, 2006; Xu et al, 2008).

There is no published GWAS that has systematically evaluated CNVs in SD, although such variation may be important in regulating the phenotype. CNV research is still in its infancy, with several technical limitations, for example, CNV prediction by any particular method may yield false positives. To address this limitation, we implemented a combined CNV calling method based on two calling algorithms (Colella et al, 2007; Sanders et al, 2011; Wang et al, 2007) that have been evaluated previously using quantitative PCR (qPCR; Sanders et al, 2011). We found that the combined method yielded a significantly greater positive predictive rate (which, in this study, was 100% for a selected homozygous deletion) than single algorithm results. Because GWAS requires a large number of subjects, the sample size in this study is considerably larger than those in previous genetic studies of OD. We collected a total of 6950 subjects (5152 after quality control), including African-American (AA) and European-American (EA) drug-dependent cases and screened controls. Of this number, 2227 were diagnosed with OD, representing one of the largest known OD genetics cohorts. We assayed the samples using the Illumina HumanOmniQuad high-density SNP array platform. Our results revealed OD-associated CNVs encompassing (or close to) biologically important genes in addictions.

MATERIALS AND METHODS

Subjects

The 6950 subjects were recruited at the Yale University School of Medicine (APT Foundation, New Haven), the University of Connecticut Health Center (Farmington), the Medical University of South Carolina (Charleston), the University of Pennsylvania School of Medicine (Philadelphia), McLean Hospital (Harvard Medical School, Belmont), and the University of Virginia (UVA) School of Medicine (Charlottesville). Subjects, except those recruited at UVA, were ascertained using Diagnostic and Statistical Manual of Mental Disorders-fourth edition (DSM-IV) criteria (American Psychiatric Association, 1994) for all major psychiatric traits, including opioid, cocaine, or alcohol dependence. Subjects were interviewed using the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSAD-DA) (Gelernter et al, 2005; Pierucci-Lagha et al, 2007). Control subjects had no diagnosed substance use or major psychiatric disorders. Subjects from the UVA site were from the Mid-South Case Control (MSCC) study on smoking dependence, where each subject was screened for multiple addictions and other psychiatric disorders. The only control subjects from the MSCC study who were used for this study were those who were screened to exclude those with substance use or psychiatric disorders (Cui et al, 2013). Details including sample size for each recruiting site (excluding UVA subjects) are provided elsewhere (Gelernter et al, 2014c). After a complete description of the study, written informed consent was obtained from each subject, as approved by the institutional review board at each site. Certificates of confidentiality for the work were obtained from both the National Institute on Drug Abuse (NIDA) and the National Institute on Alcohol Abuse and Alcoholism (NIAAA).

Genotyping

DNA was extracted from immortalized cell lines, blood, or saliva. The subjects were genotyped using the Illumina HumanOmni1-Quad platform with 1 140 419 predesigned probes (Illumina, San Diego, California) (Hodgkinson et al, 2008). Genotyping was conducted at the Yale Center for Genome Analysis and the Center for Inherited Disease Research (CIDR). For quality control, 141 HapMap samples were genotyped simultaneously with our samples (Supplementary Table 1), and two samples (NA10851 and NA11995) were included on every plate. Genotype data will be available through the database of Genotypes and Phenotypes (dbGaP).

CNV Calling

Raw intensity at each probe locus was first analyzed using the algorithms implemented in the Illumina GenomeStudio genotyping module, including intensity normalization, clustering, genotype calling, and internal quality control. The Hidden Markov Models implemented in PennCNV (Wang et al, 2007) and QuantiSNP (Colella et al, 2007) were adopted to infer CNVs by integrating multiple sources of information, for example, SNP allelic ratio distribution and signal intensity. GNOSIS (Sanders et al, 2011) was applied

to replicate the calls from PennCNV and QuantiSNP (but not used for association analyses). For homozygous deletions (0-copy), an independent calling algorithm implemented in CNVision (Sanders *et al*, 2011) was also adopted. This method looks for a probe with LRR < -3 and continues until it encounters a probe with LRR > -1.

Quantitative PCR Validation of CNV

TaqMan real-time qPCR (Alkan *et al*, 2011) was used to validate the samples with a CNV called by the Illumina genotyping platform. In this study, the TaqMan qPCR (Sanders *et al*, 2011) validation experiments were conducted using CNV assays from Applied Biosystems (ABI, Foster City, CA) for an arbitrarily picked CNV (detected by the combined methods) that occurred in 23 subjects. The comparative C_T method ($\Delta\Delta C_T$) of relative quantification (Livak and Schmittgen, 2001) was applied. Genomic DNA of individuals with and without predicted homozygous deletions was amplified in quadruplicate (Supplementary Materials).

Sample-Based Quality Controls

A total of 6950 samples were successfully genotyped. Blind duplicate reproducibility rate was 99.99% based on the genotypes of 70 duplicate sample pairs. The genotype concordance of 141 HapMap samples was 99.7%. The genotype missing rate for the raw data was 0.23% (chromosome Y excluded). We removed 364 samples with low-intensity quality, discrepant sex information, unusual X- and Y-chromosome patterns, or unexpected duplicated DNA based on the quality control functions of the genotyping array or suggested by the array provider (Supplementary Table 2). Samples were also excluded if they had low quality inferred by either PennCNV or QuantiSNP or were duplicate samples. We only analyzed the unrelated AA and EA samples. Other quality control procedures are described in the quality control section of the Supplementary Materials. A total of 5389 samples remained after the sample-based quality control analyses (Supplementary Table 3).

The quality control procedure was effective in excluding poor- or low-quality samples. For example, before the quality controls were applied, histograms showed that the CNV count per sample differed substantially from a normal distribution with an extremely long tail. Specifically, the CNV count (per sample) at the richest observation (minimum CNV count–maximum CNV count, abbreviated as modal number or mode (min–max)) was 1002 (31–16 336) and the arithmetic mean ± SD was 1220 ± 1286. However, after our sample-based quality controls were applied, the CNV counts followed a normal distribution with a mode (min–max) of 940 (322–2345) and mean of 1044 ± 303. Figure 1a and b shows the distributions of the CNV counts before and after our sample-based quality controls, respectively (after merging the CNVs from three methods). Supplementary Figures 1 and 2, the corresponding plots before merging the CNVs, provide stronger evidence that our quality controls improved the data quality. Similar improvement patterns were observed in the 0-copy deletions predicted by the homozygous deletion algorithm, implying that most of the outliers were removed

(Supplementary Figures 3 and 4). Thus, it appears that the quality control procedures removed the majority of samples with poor-quality CNV data.

In addition, the CNV counts (mean) per sample were 961 ± 312, 399 ± 86, and 88 ± 38 based on QuantiSNP, PennCNV, and GNOSIS (Supplementary Figures 5–7, respectively). In DGV, 22 ± 3% of the CNVs were reported as common CNVs (Supplementary Figure 8). The ethnic distributions of the samples have been described in our previous study (Li *et al*, 2012). An average of 50 ± 5 ancestry informative markers (AIMs) (Sanders *et al*, 2011) were used to infer sample ancestry, and the samples with potential population stratification issues (non-European or non-African ancestry) were removed from the analyses (Supplementary Figure 9).

CNV-Based Quality Controls

The following criteria were applied to filter possibly unreliable CNV calls further. Only CNVs that (1) overlapped two or more probes and (2) were commonly identified by PennCNV and QuantiSNP were included. CNVs with an overlap of >50% were considered to be the same CNV (Sanders *et al*, 2011). CNVs that were called as deletions by one method but inferred as duplications by another, or vice-versa, were excluded. For the homozygous deletion method, only CNVs that overlapped two or more probes and had LogR < -5 were included. Supplementary Table 4 shows the criteria used for CNV-based quality controls. Overall, 162 871 CNVs were identified with 95% of detected CNVs < 60 kb ranging from 17 to 9 937 527 bp in length (mean = 18 442 ± 129 188 bp) in AAs and 83 669 CNVs with 95% of detected CNVs < 60 kb ranging from 17 to 25 678 802 bp (mean = 16 591 ± 206 680 bp) in EAs. Each CNV spanned 20 probes, on average. The CNV counts per sample were 46 ± 14 in AAs (Figure 1c) and 49 ± 16 in EAs (Figure 1d) after both sample- and CNV-based quality controls. The frequencies (mean) of the filtered CNVs were 0.61 ± 2.72% in AAs (Supplementary Figure 10) and 0.86 ± 3.84% in EAs (Supplementary Figure 11). For the filtered homozygous deletions, the frequencies were 0.42 ± 1.48% in AAs (Supplementary Figure 12) and 0.81 ± 2.49% in EAs (supplementary Figure 13). The total sample size was 5152 after both sample-based and CNV-based quality controls.

Statistical Analyses

The filtered CNVs were projected to each probe and summarized by two-by-two tables (eg, CNVs overlapping each position in cases and controls). For each table, Fisher's exact test was applied to calculate the *P*-value and odds ratio (OR) with 95% confidence intervals (CIs) as the primary analysis. In this study, only the CNVs of >1000 bp were analyzed. For CNVs with both deletions (0 or 1 copy) and duplications (3 or 4 copies), association tests were also carried out for each category separately. Each race group (EA and AA) was confirmed using AIMs, and analyzed separately. The combined analyses of AAs and EAs were performed via meta-analysis together with heterogeneity analysis under a random effect model considering the direction of effects (Cao *et al*, 2014; Li and He, 2008). When
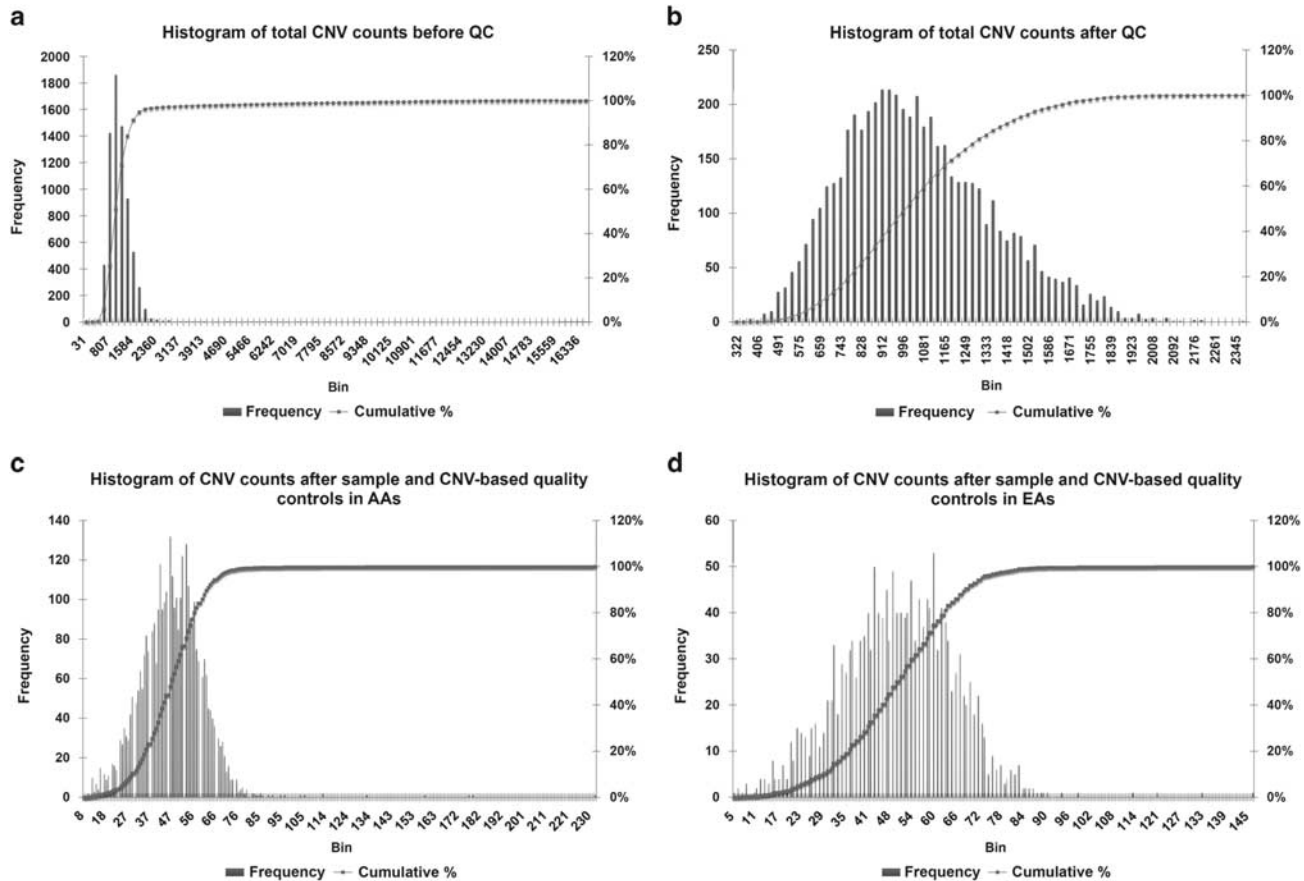
Figure 1 (a, b) Distributions of the CNV counts per sample *before* and *after* sample-based quality control, respectively. The merged CNV calls are used. (c, d) Distributions of the CNV counts per sample after both sample- and CNV-based quality controls in AAs and EAs, respectively. The modal numbers were 46 (8–231) with a mean of 46 ± 14 in AAs (c) and 59 (5–145) with a mean of 49 ± 16 in EAs (d).

a particular variant was only observed in either cases or controls alone, the Mantel–Haenszel exact analysis was adopted. After quality controls only a total of 321 unique CNV regions were recurrent (consistently called) with frequencies >1% in both AAs and EAs. Thus, the genome-wide Bonferroni significance threshold was set at $P < 0.05/321 = 0.00016$ for the association analyses of common CNVs. PLINK (Purcell *et al*, 2007) was used to map the significant CNVs on known genes, cytobands, CNV, and InDel regions, to measure the burden (rate) differences in cases *vs* controls, and to replicate the results from Fisher's exact tests. For example, we identified all the start and stop positions of the segments, calculated the CNVs that overlap each of the loci (and a 20-kb window around the locus), and then performed region-based analyses, that is, CNVs in cases that overlap known gene, cytological chromosome band, or CNV/Indel region *vs* those in controls (label-swapping max(T) permutation was used to empirically estimate significance, and correct for multiple testing). The tests were two sided. Statistical power analysis (Faul *et al*, 2007) showed that the filtered sample size had >99.9% power to detect an effect size of 0.1 (small) with significance level $\alpha = 0.05$, and 1 degree of freedom. The GeneMania (Warde-Farley *et al*, 2010) was used to map the genes related to identified CNVs to gene networks.

RESULTS

Burden Analyses in OD

Global burden analyses of the frequency differences between cases and controls could provide overall evidence of association. After both sample- and CNV-based quality controls, we analyzed a total of 5152 samples, including 547 AA and 1054 EA cases with OD and 2944 AA and 607 EA screened controls with no diagnosis of OD or other SD (Supplementary Table 5). The burden analyses (Supplementary Table 6) showed that the OD cases contained slightly fewer CNVs than controls (the average CNV counts per sample were 44.8 and 46.3 in the AA cases and controls ($P = 0.02$ based on *t*-test) as well as 48.3 and 50.5 in EAs ($P = 0.004$), respectively). The average length per CNV was 17.2 and 15.7 kb in the AA cases and controls and 16.4 and 14.9 kb in EAs, respectively. The same patterns were found when only CNVs intersecting with known genes or only homozygous deletions were analyzed. Furthermore, when only homozygous deletions were considered, the CNVs in the cases contained more genes (statistically insignificant) as those in the controls (the numbers of genes per total CNV kb were 4.1 and 1.8 in the AA cases and controls and 2.8 and 1.0 in EAs, respectively).

Association Analyses of OD

Individually significant common CNVs. Genome-wide association analyses were carried out to compare the CNV counts between OD cases and controls individually for each CNV. A P-value below the genome-wide significance level is evidence supporting an association of a CNV with OD. Overall, three CNVs, a chromosome 18q12.3 deletion ($P(Z) = 2 \times 10^{-8}$), a chromosome Xq28 deletion ($P(Z) = 3 \times 10^{-6}$), and a chromosome 1q21.3 duplication ($P(Z) = 9 \times 10^{-7}$), were genome-wide significantly associated with OD in both the AA and EA populations (Table 1). The genome-wide threshold was 0.00016 (based on the number of unique genome regions of the common CNVs that were used for the association analyses). Evidence for significant association was found for the 18q12.3 deletion with a protective effect on OD (OR = 0.59 (0.47–0.75) and $P = 3 \times 10^{-6}$ in AAs; OR = 0.68 (0.54–0.86) and P = 0.0008 in EAs; and OR = 0.64 (0.54–0.74) and $P(Z) = 2 \times 10^{-8}$ in the meta-analysis of combined samples). Interestingly, the reciprocal CNV (duplication) of the exact same region showed an opposite (risk) effect (OR = 5.40 (0.72–40.45)), and the P-value was 2×10^{-6} when both deletions and duplications were analyzed for trend in AAs. This deletion is located between the *LOC647946* and *KC6* genes. (*LOC647946* is an uncharacterized noncoding RNA and is a predicted top target of the motif CDC5L.p2 of the cell division cycle 5-like gene (Suzuki *et al*, 2009), and *KC6* was found to be associated with childhood obesity (Bradfield *et al*, 2012) and multiple blood and metabolism-related traits.) On the same cytoband (~3970 kb distance), 18q12.3, another deletion, showed suggestive evidence of a protective effect (OR = 0.49 (0.33–0.73) and $P(Z) = 0.0005$ in the combined samples; Table 2). This intergenic deletion maps between the *SETBP1* (where *microRNA 4319* is located) and *SLC14A2* genes. The latter gene was reported to be associated with metabolic syndrome and related traits (Tsai *et al*, 2010).

We also found evidence of genome-wide association in an Xq28 deletion with a risk effect (OR = 4.19 (2.03–8.49) and $P = 5 \times 10^{-5}$ in AAs; the deletion was only observed in cases (P = 0.03) in EAs; OR = 4.68 (2.38–9.2) and $P(Z) = 3 \times 10^{-6}$ in the combined samples. All of the Xq28 deletions (46 samples: 16/547 cases and 21/2944 controls in AAs; 9/1054 cases and 0/607 controls in EAs) had the same number of probes (15), and were called by all three methods. This deletion is located between the *HMGB3* and *GPR50* genes, and the latter gene was reported to be associated with bipolar affective disorder in multiple populations (Macintyre *et al*, 2010; Thomson *et al*, 2005), autism spectrum disorders (Chaste *et al*, 2010), and circulating triglyceride and HDL levels (Bhattacharyya *et al*, 2006).

We also found genome-wide significant association for a 1q21.3 duplication with a risk effect (OR = 1.58 (1.27–1.96) and $P = 4 \times 10^{-5}$ in AAs; OR = 1.7 (1.11–2.66) and P = 0.01 in EAs; and OR = 1.6 (1.33–1.94) and $P(Z) = 9 \times 10^{-7}$ in the combined samples). Its reciprocal CNV (deletion) at the same position consistently showed an opposite (protective) effect (OR = 0.55 (0.34–0.91) and $P(Z) = 0.02$ in the combined samples. This CNV, inferred based on 49 probes (10 or more probes being sufficient for confidence in the CNV inference), intersects with the exons of two cornified envelope genes (*LCE3B* and *LCE3C*). Deletion of *LCE3B* and

Table 1 CNVs Significantly Associated with Opioid Dependence

Location (chromosome: start–end)	AA				EA				Meta-analysis		Genes or adjacent genes[a]	Region	Cytoband	Note
	Case (547)	Control (2944)	OR (95% CI)	P	Case (1054)	Control (607)	OR (95% CI)	P	OR (95% CI)	P				
Chr18: 37118901–37122002	109 (0.20)	869 (0.30)	0.59 (0.47–0.75)	**2.9 × 10⁻⁶** [b]	239 (0.23)	183 (0.30)	0.68 (0.54–0.86)	0.0008	0.64 (0.54–0.74)	**2.2 × 10⁻⁸**	LOC647946, KC6	Intergenic	18q12.3	Deletion
Chr1: 150826015–150853563	144 (0.26)	543 (0.18)	1.58 (1.27–1.96)	**3.9 × 10⁻⁵** [c]	91 (0.09)	32 (0.05)	1.7 (1.11–2.66)	0.01 [d]	1.6 (1.33–1.94)	**9.4 × 10⁻⁷** [e]	LCE3B, LCE3C	Exon	1q21.3	Duplication
ChrX: 150019751–15002135I	16 (0.03)	21 (0.01)	4.19 (2.03–8.49)	**5.5 × 10⁻⁵**	9 (0.01)	0 (0)	Inf (1.14–Inf)	0.03	4.68 (2.38–9.2)	**2.7 × 10⁻⁶**	HMGB3, GPR50	Intergenic	Xq28	Deletion (males)

The numbers in the brackets after sample sizes are corresponding frequencies. The P-values of <0.00016 (genome-wide significance) are in bold.
[a] If a CNV is in intergenic region, the two adjacent genes are shown. The hg18 assembly (NCBI 36 assembly) was used to define the genome coordinates throughout the study.
[b] P = 0.05 and OR = 5.4 (0.72–40.45) for the reciprocal CNV (duplication) and $P = 2.4 \times 10^{-6}$ for the trend analysis of deletion and duplication.
[c] P = 0.5 and OR = 0.68 (0.26–1.51) for the reciprocal CNV (deletion) and $P = 9 \times 10^{-5}$ for the trend analysis of deletion and duplication.
[d] P = 0.03 and OR = 0.49 (0.24–0.96) for the reciprocal CNV (deletion) and P = 0.004 for the trend analysis of deletion and duplication.
[e] P = 0.02 and OR = 0.55 (0.34–0.91).

Table 2 CNVs Suggestively Associated with Opioid Dependence

Location (Chromosome: start–end)	AA				EA				Meta-analysis		Genes or adjacent genes[a]	Region	Cytoband	Note
	Case (547)	Control (2944)	OR (95% CI)	P	Case (1054)	Control (607)	OR (95% CI)	P	OR (95% CI)	P				
ChrX: 154135529–154135411	11	15	4.01 (1.65–9.39)	0.001	8	1	4.63 (0.62–205.82)	0.2	4.08 (1.96–8.49)	0.00017	VBP1, RAB39B	Intergenic	Xq28	Deletion (males)
Chr19: 32468601–32680227	0	41	0 (0–0.5)	0.002	0	2	0 (0–3.06)	0.1	0 (0–0.36)	0.0002	LOC100101266, LOC148189	Intergenic	19q12	Duplication
Chr20:29267954–29297768	0	20	0 (0–1.09)	0.06	5	13	0.22 (0.06–0.66)	0.002	0.18 (0.05–0.5)	0.0002	FRG1B, DEFB115	Intergenic	20q11.21	Both
Chr10: 122216951–122218524	16	144	0.59 (0.32–0.99)	0.04[b]	95	82	0.63 (0.46–0.88)	0.005	0.62 (0.47–0.81)	0.0005	PPAPDC1A	Intron	10q26.12	Deletion
Chr18: 41092137–41092686	19	186	0.53 (0.31–0.87)	0.007	12	17	0.4 (0.17–0.9)	0.02	0.49 (0.33–0.73)	0.0005	SETBP1, SLC14A2	Intergenic	18q12.3	Deletion
Chr6: 161190060–161198759	2	46	0.23 (0.03–0.89)	0.03	16	22	0.41 (0.2–0.82)	0.01	0.35 (0.19–0.63)	0.0007	PLG, MAP3K4	Intergenic	6q26	Both
Chr6: 74649290–74650963	7	12	3.17 (1.05–8.77)	0.02[c]	24	4	3.51 (1.2–13.98)	0.02	3.31 (1.64–6.69)	0.0008	CD109, COL12A1	Intergenic	6q13	Duplication

The suggestive threshold was $P < 0.001$.
$P(Q) > 0.1$ for all the meta-analyses. Both indicates combined duplication and deletion. The full versions of this table are shown in Supplementary Tables 7, 8, and 16.
[a]If a CNV is in intergenic region, the two adjacent genes are shown.
[b]$P > 0.05$ and OR = 3.6 (0.3–31.47) for the reciprocal CNV (duplication) and $P = 0.04$ for the trend analysis of deletion and duplication.
[c]$P > 0.05$ and OR = 0 (0–8.16) for the reciprocal CNV (deletion) and $P = 0.03$ for the trend analysis of deletion and duplication.

LCE3C was associated with chronic hand eczema (Molin et al, 2011), psoriasis(de Cid et al, 2009; Riveira-Munoz et al, 2011), rheumatoid arthritis (Docampo et al, 2010), and systemic lupus erythematosus (Lu et al, 2011).

Some CNVs showed *P*-values above the genome-wide significance level but still at a stringent ('suggestive') level, making them potentially interesting for further investigation. In this study, two deletions and two duplications showed suggestive or marginal associations with OD ($0.00017 < P(Z) < 0.001$; Table 2). For example, a 10q26.12 deletion showed association with a protective effect (OR = 0.62 (0.47–0.81) and $P(Z) = 0.0005$ in the combined samples. Duplication of the exact same region showed an opposite effect (OR = 3.6 (0.3–31.47); and $P = 0.04$ when both deletions and duplications were analyzed for trend in AAs. This CNV intersects with the intronic region of the *PPAPDC1A* gene that encodes phosphatidate phosphatase and is conserved in many species from chimpanzee to rice. Evidence of association was also found for a 6q13 duplication with a risk effect (OR = 3.31 (1.64–6.69) and $P(Z) = 0.0008$ in the combined samples. Its reciprocal CNV (deletion) of the same region was only identified in AA controls ($P > 0.05$), again consistently showing an opposite effect. This CNV is located between the *CD109* and *COL12A1* genes that were found to be associated with oral cancers (Hagiwara et al, 2008) and fibroma (Yasuda et al, 2009), respectively. We also observed a duplication on 6q26 with a suggestive protective effect (OR = 0.35 (0.19–0.63) and $P(Z) = 0.0007$ in the combined samples. The 6q26 duplication is between the *PLG* and *MAP3K4* genes. The *MAP3K4* gene was reported to play an important role in nicotine dependence (Grucza et al, 2010).

Rare and unique CNVs with large effects. Some CNVs revealed no *P*-values below the genome-wide or suggestive significance level but showed large effect sizes (ie, ORs), often because of low CNV frequencies (ie, few total observations of variant alleles). These CNVs might be of clinical interest, pending confirmation. Overall, we observed dozens of rare and unique CNVs with potentially large effect size (Supplementary Table 7). Among them, three CNVs, an Xq28 deletion ($P(Z) = 0.0002$), a 19q12 duplication ($P(Z) = 0.0002$), and a 20q11.21 CNV ($P(Z) = 0.0002$), showed suggestive ($P < 0.001$) associations (Table 2). In addition, four deletions (2q32.1, 4q34.1, 9p21.3, and 10q21.3) and three duplications (1p11.1–1p11.2, 12p11.21, and 12p13.31)) showed large risk effects (OR > 3) that were statistically replicated in both AAs and EAs. For example, the deletion on 2q32.1 showed ORs = 8.11 and nominally 'infinity' (the upper limit could not be estimated because the CNVs were observed only in 3 of 1054 cases but not in any of the 607 controls) in AAs and EAs, respectively; and OR = 9.39 (1.39–105.31) and $P(Z) = 0.009$ in the combined samples. On the other hand, we found 10 deletions and 4 duplications with large protective effects (OR < 0.3) that were replicated in both AAs and EAs. All of the 14 CNVs were identified only in the controls in AAs, EAs, and both populations together. For example, for the 3p26.2 deletion with 34 probes, 11 and 5 CNVs were found in the AA and EA controls, respectively, but none in the AA or EA cases ($P = 0.001$). For the 19q12 duplication, 41 and 2 CNVs were

identified in the AA and EA controls, respectively, but not in the cases in either population ($P = 0.0002$). The Supplementary Table 7 also shows additional CNVs (including low-quality CNVs) that were uniquely observed in either cases or controls in both populations, including 24 deletions and 3 duplications unique to the cases and 3 duplications unique to the controls. Both AAs and EAs showed the same patterns. For example, a 3q12.2 duplication, which intersects with the exons of the *TFG* gene, was identified only in the cases and not in the controls in both populations.

For the homozygous deletions alone (inferred by the 0-copy algorithm; Supplementary Table 8), we identified 23 deletions with medium-to-large risk effects (OR >1.5) and 10 unique (exclusive to) in the cases and 26 with medium-to-large protective effects (OR <0.6) and 27 unique in the controls that were replicated in both populations. For example, a homozygous deletion on 19p13.2, which intersects with the intronic region of the *KANK2* gene, was unique in the cases. PLINK generated consistent association results (Supplementary Tables 9–13).

Burden and Association Analyses of Alcohol, Cocaine, Opioids, Cannabis, and Nicotine Dependence

Because some of our OD patients were diagnosed with dependence on multiple substances, we carried out similar analyses by selecting from among the OD patients a subset of subjects with more severe addictive disorders, that is, we identified 118 AA and 214 EA cases with comorbid alcohol, cocaine, opioid, marijuana, and nicotine dependence and 1372 AA and 56 EA screened controls with no diagnosed dependence on any of the five substances (Supplementary Table 14). The CNV burden analyses (Supplementary Table 15) showed consistent patterns: in AAs, the average length of the homozygous deletions in the cases was 1.64 times longer than those in the controls (1.0 and 0.6 kb, respectively, $P > 0.05$); in EAs, the case group had, on average, 5 more CNVs (50 and 45, respectively, $P = 0.04$) or 2 more homozygous deletions (5 and 3, respectively, $P = 0.003$) than the control group. The total length per sample of the homozygous deletions in the cases was twice that in the controls (3.7 and 1.8 kb, respectively, $P = 0.02$). More interestingly, the homozygous deletions in the cases contained more genes than those in the controls in EAs (1.7 genes per sample in cases *vs* 1.1 genes per sample in controls, $P = 0.03$; or 7.48 genes per total CNV kb in cases *vs* 1.15 in controls, $P > 0.05$). In addition, we observed a similar trend in data from the Study of Addiction: Genetics and Environment (dbGaP Study Accession: phs000092.v1.p1).

Overall, the association analyses (Supplementary Table 16) showed that 10 duplications and 15 deletions were observed only in the cases; and 9 duplications and 7 deletions were observed only in the controls. These unique CNVs were rare and replicated in both AAs and EAs. The smallest $P(T)$ was 0.002 (an Xq21.1 deletion) among the risk CNVs and the smallest $P(T)$ was 0.008 (a 18p11.32 deletion) among the protective CNVs. Compared with the results for OD, the results for dependence on all five substances appeared to show larger effect sizes and gene enrichment scores; however, the sample size of severe cases limited the statistical power, resulting in fewer signals (genome-wide *P*-values) observed.

Summary of Genes or Regions Involved

To summarize, when all of these CNVs (Supplementary Tables 7, 8, and 16) were combined, 110 regions (including low-quality CNVs) were identified. A total of 194 genes were involved (ie, the two adjacent genes were used when a CNV was in an intergenic region), and 17 genes were observed multiple times (Supplementary Table 17); for example, CNVs in the intergenic region *LOC100101266-LOC148189* were observed five times, and CNVs in each of the two regions of *DDX12-KLRB1* and *RIOK2-RGMB* were observed three times. Gene network analyses showed that the majority of these genes were strongly connected based on known protein–protein physical interaction, colocalization, shared protein domain, coexpression, and genetic interaction information (Supplementary Figure 14). Some of the genes have been reported to be associated with alcohol dependence (*MMADHC-TRNAE38P* (Heath *et al*, 2011)) or alcohol and nicotine codependence (*KCND2* (Zuo *et al*, 2012)) in the SNP-based GWAS literature (Supplementary Table 18). We compared the genes that were affected by (intersected or were close to) CNVs identified in this study and those that were affected by SNPs and pathways identified in our published OD GWAS study (Gelernter *et al*, 2014b). We found that three genes, *CTNNA3*, *PTPRC*, and *PTPRD*, were implicated at least modestly in both studies, with the first gene encoding a cadherin-associated catenin protein and the latter two encoding protein tyrosine phosphatases, with all three proteins being related to the plasma membrane. These results are shown in Supplementary Table 19.

DISCUSSION

We carried out a genome-wide CNV study of OD in a sample of 5152 EAs and AAs. In the course of the study, we implemented combined CNV calling methods. Our selected CNV calling algorithms have previously been validated by a large number of TaqMan qPCR experiments (Sanders *et al*, 2011) and, in this study, we successfully replicated the experiments in our own data (the results showed reaction efficiency of $>95\%$ and an R^2 value of 0.98 for both amplification targets. Consistent with our prediction, there was no amplification observed for the tested CNV in any of the 23 subjects). We identified dozens of CNVs, with three of them being genome-wide significant. These CNVs (common, rare, and unique) showed strong associations (eg, $P(Z) = 2 \times 10^{-8}$), large risk or protective effects, or both. Both duplications and deletions were observed in four common CNVs; consistently, the duplications showed risk effects whereas the deletions of the same regions showed protective effects, suggesting that more copies in these regions result in higher risk, a hypothesis that should be investigated. We observed a few CNVs only in OD cases (or cases addicted to all five substances) or only in controls. The majority of the observations were replicated in two independent populations, AAs and EAs. Some of our identified CNVs contain genes that were previously reported to be associated with SD (eg, the *MAP3K4* gene was previously reported to be associated with nicotine dependence (Grucza *et al*, 2010); CNV in the *MAP3K3* gene was recently reported as a mutational mechanism in

schizophrenia (Rippey *et al*, 2013)), whereas some others harbor new genes of potential biological importance in addiction.

Regarding the Xq28 deletion (within the intergenic region between *HMGB3* and *GPR50*), all 46 of the subjects whom we predicted to carry this deletion were males (deletions on X chromosome are generally be more consequential in males than females). The following may partially explain this observation. (1) We hypothesize that the primary, observable form of this Xq28 deletion is '1-copy' loss, resulting in females primarily who have a single copy (and 1-copy lost (or deleted)) and 2 copies (no copy is lost) whereas the males can only have 0-copy (1-copy is lost) and 1-copy (no copy is lost). Because the CNV calling algorithms were much more sensitive at distinguishing 0-copy and 1-copy from 2-copies, male hemizygotes were to be more easily detected than female heterozygotes (ie, no copies of the variable segment distinguished from 1-copy *vs* a difference in intensity between 1-copy and 2-copy). Furthermore, after we removed the low-confidence CNVs, which were more likely to be 1-copy compared with 0-copy deletions, in quality control, we only observed 0-copy genotypes for this deletion site. This hypothesis needs experimental validation. (2) This deletion might be associated with an X chromosome-linked disease. For example, the Fragile X mental retardation protein (FMRP) was found to influence the development of addiction-related behaviors (Smith *et al*, 2014); the Fragile X syndrome gene, *FMR1*, is 3 million base pairs from this deletion. (3) This observation might also be because of chance or an unknown biological mechanism. For instance, a GWAS (Kennedy *et al*, 2012) showed that *MAMLD1*, 586 kb upstream from the deletion, was associated with immune response to smallpox vaccine; the gamma-aminobutyric acid receptor subunit gene, *GABRE*, is 851 kb downstream from the deletion.

The microarray-based CNV calling methods assume a diploid genome; however, CNVs tend to reside in repetitive sequences and have a positive correlation with segmental duplications. With an uncertain signal-to-noise ratio (McCarroll *et al*, 2008), CNV (particularly duplication) detection becomes difficult and can be unreliable (Alkan *et al*, 2011) when the breakpoints lie in duplicated regions (Alkan *et al*, 2011). Consequently, identifying accurate boundaries and copy numbers require careful calling strategies. Our major CNV findings were outside of repetitive regions (those in segmental duplications are marked in Supplementary Tables 7, 8, and 16). Furthermore, we identified consensus CNV calls from multiple independent algorithms designed specifically for Illumina platforms and optimized parameters in conjunction with manual curation and experimental validation. As shown in the quality control section, our combined methods and stringent quality controls significantly improved the calling accuracy. However, as a tradeoff, many low-confidence samples and CNVs were excluded from our analyses, resulting in the observation that the CNV frequencies in our samples (mean = 0.7%) were lower than those reported in the DGV or other studies. Although this might produce overall genome-wide bias in ways that could not be directly characterized, it also resulted in a set of retained CNV calls in which we could be highly confident.

CNVs encompass more total nucleotides and arise *de novo* more frequently (ie, higher locus-specific mutation rate (Zhang *et al*, 2009)) than SNPs. CNVs play a major role in human evolution, genetic diversity, and susceptibility to diseases (Stankiewicz and Lupski, 2010). CNVs caused by genomic rearrangements can have direct effects on phenotypes through mechanisms such as (1) gene dosage, (2) gene interruption, (3) gene fusion (hybridization of multiple separate genes—fusion genes are often oncogenes), (4) position effects (effects on expression or regulation of a nearby gene outside of the CNV region that may account for some of our identified intergenic CNVs), (5) unmasking of recessive alleles or functional polymorphisms, and (6) transvection effects (Lupski and Stankiewicz, 2005; Zhang *et al*, 2009).

According to the 'common disease, rare variant' hypothesis, many rare (unique or private) variants underlie susceptibility to complex conditions, and such CNVs would be of recent origin and likely to be highly penetrant (Cook and Scherer, 2008). This might be the case particularly for psychiatric disorders; for example, the frequency of the well-known 16p11.2 deletion was 1% in autism cases but 1×10^{-4} in controls (Weiss *et al*, 2008). Some of the highly penetrant CNVs that were identified in this study may contribute to the risk or severity of addictive disorders, as a consequence of loss, gain, or disruption of dosage-sensitive genes (Cook and Scherer, 2008).

CNV studies are important because they can also affect the interpretation of SNP genotyping. A deletion may cause contiguous SNPs to show loss of heterozygosity because hemizygous genotypes are called as homozygous (Wain *et al*, 2009). For example, if the minor allele A is present on one chromosome and the homologous chromosomal location is deleted, then only one allele is detected and the genotype is called as AA. This misrepresentation can cause apparent deviation from Hardy–Weinberg equilibrium, and a Mendelian transmission error. This circumstance has caused many SNPs in CNV regions to be excluded from the earliest genome-wide genotyping arrays (Cooper *et al*, 2008; McCarroll *et al*, 2008), yielding a paucity of conventional SNP probes in CNV-rich regions. Moreover, the location, size, and boundaries of the CNVs documented in public databases may be imprecise. Since the first-generation CNV map of the human genome was constructed in 2006 (Redon *et al*, 2006), no single human genome has been published that includes the complete spectrum of structural variation (Alkan *et al*, 2011), reflecting difficulties in the creation of accurate and complete sets of CNV calls. The array used in this study was intentionally designed with a large number of special intensity-only probes in CNV-rich regions. The newer generation of arrays, including the one that we used, have greater coverage and resolution (Conrad *et al*, 2008; McCarroll *et al*, 2008; Wain *et al*, 2009). Our results may provide a CNV candidate pool, notable for its genome-wide significant and large effects (eg, only observed in cases or controls), for further validation and genetic investigation of addiction and psychiatric illnesses.

We have completed GWASs that incorporated the sample described here for OD (Gelernter *et al*, 2014b) and several other traits, including alcohol dependence (Gelernter *et al*, 2014a), cocaine dependence (Gelernter *et al*, 2014c), and posttraumatic stress disorder (Xie *et al*, 2013). All of these

GWASs identified genome-wide-significant risk loci. These data have also contributed to analyses of the genetic architecture of alcohol dependence in the AA part of the sample (Yang *et al*, 2013). The basic SNP calls have extensive utility for GWASs and common-variant genetic risk score studies, as well as for use of intensity measures to estimate CNVs. As discussed above, although there was some overlap between possible risk genes identified in the present study and genes highlighted by pathway analysis in our previous OD GWAS, most of the major signals were unique to one or the other analysis methods. Our data thus weakly support convergence of mechanisms (SNP *vs* structural variation) affecting the same risk genes, and more strongly support the possibility that these mechanisms can modulate risk independently.

In conclusion, this study in AAs and EAs is the first genome-wide CNV association study of OD. We analyzed a large number of OD cases and screened controls, and our results suggested that many CNVs were likely to contribute to susceptibility or resistance to OD. The identification of these OD-associated or large-effect CNVs may enhance our understanding of the impact of genetic variation on the risk of opioid addiction. However, efforts to replicate these findings in larger, independent samples are warranted (Barnes *et al*, 2008; Wellcome Trust Case Control Consortium *et al*, 2010; Zhou and Stephens, 2012). Further investigation of the CNVs identified here in parents of probands to determine whether they are *de novo* or inherited and their pathogenic significance is also a logical next step in this line of inquiry.

REFERENCES

Alkan C, Coe BP, Eichler EE (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet* 12: 363–376.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders (DSM)*. 4th edn, American Psychiatric Press: Washington, DC.

Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D et al (2008). A robust statistical method for case-control association testing with copy number variation. *Nat Genet* 40: 1245–1252.

Bhattacharyya S, Luan J, Challis B, Keogh J, Montague C, Brennand J et al (2006). Sequence variants in the melatonin-related receptor gene (GPR50) associate with circulating triglyceride and HDL levels. *J Lipid Res* 47: 761–766.

Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K et al (2010). Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 463: 666–670.

Bradfield JP, Taal HR, Timpson NJ, Scherag A, Lecoeur C, Warrington NM et al (2012). A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat Genet* 44: 526–531.

Cao J, Liu X, Han S, Zhang CK, Liu Z, Li D (2014). Association of the HTR2A gene with alcohol and heroin abuse. *Hum Genet* 133: 357–365.

Chaste P, Clement N, Mercati O, Guillaume JL, Delorme R, Botros HG et al (2010). Identification of pathway-biased and deleterious melatonin receptor mutants in autism spectrum disorders and in the general population. *PLoS One* 5: e11495.

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P et al (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 35: 2013–2025.

Conrad DF, Pinto D, Feuk L, Redon R, Carter N, Lee C et al (2008): A comprehensive map of common copy number variation at 50bp resolution, and resulting biological insights. In 58th Annual Meeting of The American Society of Human Genetics, Philadelphia, PA, USA.

Cook EH Jr, Scherer SW (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature* 455: 919–923.

Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* 40: 1199–1203.

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

de Cid R, Riveira-Munoz E, Zeeuwen PL, Robarge J, Liao W, Dannhauser EN et al (2009). Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet* 41: 211–215.

Docampo E, Rabionet R, Riveira-Munoz E, Escaramis G, Julia A, Marsal S et al (2010). Deletion of the late cornified envelope genes, LCE3C and LCE3B, is associated with rheumatoid arthritis. *Arthritis Rheum* 62: 1246–1251.

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR et al (2007). Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res* 17: 1266–1277.

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L et al (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39: 721–723.

Faul F, Erdfelder E, Lang AG, Buchner A (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39: 175–191.

Gelernter J, Kranzler HR, Sherva R, Almasy L, Koesterer R, Smith AH et al (2014a). Genome-wide association study of alcohol dependence: significant findings in African- and European-Americans including novel risk loci. Mol Psychiatry 19: 41–49.

Gelernter J, Kranzler HR, Sherva R, Koesterer R, Almasy L, Zhao H et al (2014b). Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. Biol Psychiatry 76: 66–74.

Gelernter J, Panhuysen C, Weiss R, Brady K, Hesselbrock V, Rounsaville B et al (2005). Genomewide linkage scan for cocaine dependence and related traits: significant linkages for a cocaine-related trait and cocaine-induced paranoia. Am J Med Genet B Neuropsychiatr Genet 136B: 45–52.

Gelernter J, Sherva R, Koesterer R, Almasy L, Zhao H, Kranzler HR et al (2014c). Genome-wide association study of cocaine dependence and related traits: FAM53B identified as a risk gene. Mol Psychiatry 19: 717–723.

Goldman D, Oroszi G, Ducci F (2005). The genetics of addictions: uncovering the genes. Nat Rev Genet 6: 521–532.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G et al (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307: 1434–1440.

Grant BF, Stinson FS, Dawson DA, Chou SP, Dufour MC, Compton W et al (2004). Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Arch Gen Psychiatry 61: 807–816.

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

Hagiwara S, Murakumo Y, Sato T, Shigetomi T, Mitsudo K, Tohnai I et al (2008). Up-regulation of CD109 expression is associated with carcinogenesis of the squamous epithelium of the oral cavity. Cancer Sci 99: 1916–1923.

Hai R, Pei YF, Shen H, Zhang L, Liu XG, Lin Y et al (2011). Genome-wide association study of copy number variation identified gremlin1 as a candidate gene for lean body mass. J Hum Genet 57: 33–37.

Heath AC, Whitfield JB, Martin NG, Pergadia ML, Goate AM, Lind PA et al (2011). A quantitative-trait genome-wide association study of alcoholism risk in the community: findings and implications. Biol Psychiatry 70: 513–518.

Hodgkinson CA, Yuan Q, Xu K, Shen PH, Heinz E, Lobos EA et al (2008). Addictions biology: haplotype-based analysis for 130 candidate genes on a single array. Alcohol Alcohol 43: 505–515.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y et al (2004). Detection of large-scale variation in the human genome. Nat Genet 36: 949–951.

International Schizophrenia Consortium (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455: 237–241.

Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D et al (2009). Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84: 148–161.

Kennedy RB, Ovsyannikova IG, Pankratz VS, Haralambieva IH, Vierkant RA, Poland GA (2012). Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. Hum Genet 131: 1403–1421.

Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA et al (2008). Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet 17: 628–638.

Li D, He L (2008). Meta-study on association between the monoamine oxidase A gene (MAOA) and schizophrenia. Am J Med Genet B Neuropsychiatr Genet 147B: 174–178.

Li D, Zhao H, Kranzler HR, Oslin D, Anton RF, Farrer LA et al (2012). Association of COL25A1 with comorbid antisocial personality disorder and substance dependence. Biol Psychiatry 71: 733–740.

Li MD, Burmeister M (2009). New insights into the genetics of addiction. Nat Rev Genet 10: 225–231.

Livak KJ, Schmittgen TD (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 25: 402–408.

Lu X, Guo J, Zhou X, Li R, Liu X, Zhao Y et al (2011). Deletion of LCE3C_LCE3B is associated with rheumatoid arthritis and systemic lupus erythematosus in the Chinese Han population. Ann Rheum Dis 70: 1648–1651.

Lupski JR (2006). Genome structural variation and sporadic disease traits. Nat Genet 38: 974–976.

Lupski JR (2007). An evolution revolution provides further revelation. Bioessays 29: 1182–1184.

Lupski JR, Stankiewicz P (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet 1: e49.

Macintyre DJ, McGhee KA, Maclean AW, Afzal M, Briffa K, Henry B et al (2010). Association of GPR50, an X-linked orphan G protein-coupled receptor, and affective disorder in an independent sample of the Scottish population. Neurosci Lett 475: 169–173.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ et al (2009). Finding the missing heritability of complex diseases. Nature 461: 747–753.

Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J et al (2008). Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82: 477–488.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A et al (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40: 1166–1174.

Mefford HC, Cooper GM, Zerr T, Smith JD, Baker C, Shafer N et al (2009). A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. Genome Res 19: 1579–1585.

Molin S, Vollmer S, Weiss EH, Weisenseel P, Ruzicka T, Prinz JC (2011). Deletion of the late cornified envelope genes LCE3B and LCE3C may promote chronic hand eczema with allergic contact dermatitis. J Investig Allergol Clin Immunol 21: 472–479.

Nguyen DQ, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, Ponting CP (2008). Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome Res 18: 1711–1723.

Pierucci-Lagha A, Gelernter J, Chan G, Arias A, Cubells JF, Farrer L et al (2007). Reliability of DSM-IV diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (SSADDA). Drug Alcohol Depend 91: 85–90.

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

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD et al (2006). Global variation in copy number in the human genome. Nature 444: 444–454.

Rippey C, Walsh T, Gulsuner S, Brodsky M, Nord AS, Gasperini M et al (2013). Formation of chimeric genes by copy-number variation as a mutational mechanism in schizophrenia. Am J Hum Genet 93: 697–710.

Riveira-Munoz E, He SM, Escaramis G, Stuart PE, Huffmeier U, Lee C et al (2011). Meta-analysis confirms the LCE3C_LCE3B deletion as a risk factor for psoriasis in several ethnic groups and finds interaction with HLA-Cw6. *J Invest Dermatol* **131**: 1105–1109.

Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D et al (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**: 863–885.

Schaschl H, Aitman TJ, Vyse TJ (2009). Copy number variation in the human genome and its implication in autoimmunity. *Clin Exp Immunol* **156**: 12–16.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T et al (2007). Strong association of de novo copy number mutations with autism. *Science* **316**: 445–449.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P et al (2004). Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.

Smith LN, Jedynak JP, Fontenot MR, Hale CF, Dietz KC, Taniguchi M et al (2014). Fragile X mental retardation protein regulates synaptic and behavioral plasticity to repeated cocaine administration. *Neuron* **82**: 645–658.

Stankiewicz P, Lupski JR (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**: 437–455.

Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S et al (2008). Large recurrent microdeletions associated with schizophrenia. *Nature* **455**: 232–236.

Suzuki H, Forrest AR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM et al (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**: 553–562.

The International Schizophrenia Consortium (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**: 237–241.

Thomson PA, Wray NR, Thomson AM, Dunbar DR, Grassie MA, Condie A et al (2005). Sex-specific association between bipolar affective disorder in women and GPR50, an X-linked orphan G protein-coupled receptor. *Mol Psychiatry* **10**: 470–478.

Tsai HJ, Hsiao CF, Ho LT, Chuang LM, He CT, Curb JD et al (2010). Genetic variants of human urea transporter-2 are associated with metabolic syndrome in Asian population. *Clin Chim Acta* **411**: 2009–2013.

Vrijenhoek T, Buizer-Voskamp JE, van der Stelt I, Strengman E, Sabatti C, Geurts van Kessel A et al (2008). Recurrent CNVs disrupt three candidate genes in schizophrenia patients. *Am J Hum Genet* **83**: 504–510.

Wain LV, Armour JA, Tobin MD (2009). Genomic copy number variation, human health, and disease. *Lancet* **374**: 340–350.

Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J et al (2010). A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* **463**: 671–675.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF et al (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**: 1665–1674.

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P et al (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* **38**(Web Server issue): W214–W220.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R et al (2008). Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**: 667–675.

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V et al (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**: 713–720.

Wilson GM, Flibotte S, Chopra V, Melnyk BL, Honer WG, Holt RA (2006). DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. *Hum Mol Genet* **15**: 743–749.

Xie P, Kranzler HR, Yang C, Zhao H, Farrer LA, Gelernter J (2013). Genome-wide association study identifies new susceptibility loci for posttraumatic stress disorder. *Biol Psychiatry* **74**: 656–663.

Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* **40**: 880–885.

Yang C, Li C, Kranzler HR, Farrer LA, Zhao H, Gelernter J (2013). Exploring the genetic architecture of alcohol dependence in African-Americans via analysis of a genomewide set of common variants. *Hum Genet* **133**: 617–624.

Yasuda T, Nishio J, Sumegi J, Kapels KM, Althof PA, Sawyer JR et al (2009). Aberrations of 6q13 mapped to the COL12A1 locus in chondromyxoid fibroma. *Mod Pathol* **22**: 1499–1506.

Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ (2008). Extensive copy-number variation of the human olfactory receptor gene family. *Am J Hum Genet* **83**: 228–242.

Zhang F, Gu W, Hurles ME, Lupski JR (2009). Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**: 451–481.

Zhou X, Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**: 821–824.

Zuo L, Zhang F, Zhang H, Zhang XY, Wang F, Li CS et al (2012). Genome-wide search for replicable risk gene regions in alcohol and nicotine co-dependence. *Am J Med Genet B Neuropsychiatr Genet* **159B**: 437–444.

Supplementary Information accompanies the paper on the Neuropsychopharmacology website (http://www.nature.com/npp)