Genome-wide association study of behavioral disinhibition in a selected adolescent sample

Jaime Derringer[1]*, Robin P. Corley[2], Brett C. Haberstick[2], Susan E. Young[2], Brittany Demmitt[2], Daniel P. Howrigan[3], Robert M. Kirkpatrick[4], William G. Iacono[5], Matt McGue[5], Matthew Keller[2], Sandra Brown[6], Susan Tapert[6], Christian J. Hopfer[7], Michael C. Stallings[2], Thomas J. Crowley[7], Soo Hyun Rhee[2], Ken Krauter[2], John K. Hewitt[2], Matthew B. McQueen[2]

[1] University of Illinois Urbana-Champaign, [2] University of Colorado Boulder, [3] Massachusetts General Hospital/Harvard Medical School, [4] Virginia Commonwealth University, [5] University of Minnesota, [6] University of California San Diego, [7] University of Colorado Denver

* Corresponding Author: Jaime Derringer, jderr@illinois.edu, Department of Psychology, University of Illinois Urbana-Champaign, Champaign IL USA 61820

Abstract

Behavioral disinhibition (BD) is a quantitative measure designed to capture the heritable variation encompassing risky and impulsive behaviors. As a result, BD represents an ideal target for discovering genetic loci that predispose individuals to a wide range of antisocial behaviors and substance misuse that together represent a large cost to society as a whole. Published genome-wide association studies (GWAS) have examined specific phenotypes that fall under the umbrella of BD (e.g. alcohol dependence, conduct disorder); however no GWAS has specifically examined the overall BD construct. We conducted a GWAS of BD using a sample of 1,901 adolescents over-selected for characteristics that define high BD, such as substance and antisocial behavior problems, finding no individual locus that surpassed genome-wide significance. Although no single SNP was significantly associated with BD, restricted maximum likelihood analysis estimated that 49.3% of the variance in BD within the Caucasian sub-sample was accounted for by the genotyped SNPs (p=0.06). Gene-based tests identified seven genes associated with BD ($p \leq 2.0 \times 10^{-6}$). Although the current study was unable to identify specific SNPs or pathways with replicable effects on BD, the substantial sample variance that could be explained by all genotyped SNPs suggests that larger studies could successfully identify common variants associated with BD.

Behavioral disinhibition (BD) is a latent quantitative measure designed to capture common variation shared across many harmful or dangerous behaviors including substance problems, antisocial or criminal behavior, and novelty seeking (Young et al. 2000). In addition, 60%-80% of variation in BD is attributed to additive genetic effects, making BD more heritable than many of the individual component behaviors used to formulate the latent BD construct (Young et al. 2000, Krueger et al. 2002, Hicks et al. 2013). To date, genome-wide association studies (GWAS) have been restricted to such individual component behaviors of BD (McGue et al. 2013), including use or abuse of various substances (e.g. alcohol (Bierut et al. 2010, Edenberg et al. 2010, Frank et al. 2012, Gelernter et al. 2014a, Kapoor et al. 2013, Schumann et al. 2011, Treutlein et al. 2009, Wang et al. 2012), tobacco (Bierut et al. 2007, Furberg et al. 2010, Liu et al. 2010, Thorgeirsson et al. 2010), cannabis (Agrawal et al. 2011, Verweij et al. 2013), methamphetamine (Uhl et al. 2008), opioids (Gelernter et al. 2014b, Nielsen et al. 2010), and cocaine (Gelernter et al. 2014c)), conduct disorder (Dick et al. 2011), adult antisocial behavior (Tielbeek et al. 2012), and related personality constructs such as excitement seeking (Terracciano et al. 2011). Although certain behaviors, most notably tobacco use (Bierut et al. 2007, Furberg et al. 2010, Liu et al. 2010, Thorgeirsson et al. 2010), have identified robust associations with specific variants, many GWAS fail to identify individual loci with genome-wide significant effects. This suggests that much of the heritability underlying each trait is unlikely the result of a small number of variants with large effects, and will require larger sample sizes in order to identify variants with small effects (Lee et al. 2011). GWAS of other phenotypes have identified significant replicated effects when large enough samples sizes have been amassed to provide adequate statistical power to identify variants despite very small effect sizes (e.g., accounting for 0.1% of the total variance, or less; Sullivan et al. 2011; Rietveld et al. 2013).

Increasing sample sizes is only one of a number of ways to increase statistical power. Improved phenotypic assessment and modeling could also provide increased statistical power for studies conducted in more moderately sized samples, and particularly for phenotypes that are presumed to be continuously distributed in the population (van der Sluis et al. 2012). BD is a prime example in this context, as relevant quantitative differences in phenotypic severity are maintained between individuals, whereas a case-control approach is fairly insensitive to these differences. However, one issue with searching for specific genetic influences on many continuous phenotypes, such as BD, is that the most severe, clinically significant levels are relatively rare in the general population, as they are located on the extreme ends of the distribution. Ascertaining samples specifically for individuals with extreme phenotypes may improve our ability to detect small genetic effects by increasing the sample variance. Therefore, an ideal sample might be considered one that is enriched (and well-measured) for extreme BD characteristics.

We report here results and characterization of the initial GWAS from the Center on Antisocial Drug Dependence (CADD), an adolescent sample over-selected for severe BD characteristics. Any genetic effects on BD are potentially attributable to many (i.e., thousands of) variants, each with a very, very small effect. Incorporating methods for aggregating effects across multiple variants, such as gene- and pathway-based analyses, can identify promising causal biological systems beyond the significance of any single variant. In addition to SNP level association, the current study applied gene-based, pathway-based, and genome-wide approaches to characterize genetic influences on BD in a diverse, clinically-oversampled, thoroughly phenotyped sample. By supplementing a GWAS with several methods of aggregating genetic evidence across many potentially associated variants, we sought to generate novel insights into

the potential genetic etiology of BD and identify promising candidates, either old or new, for future study.

Methods

Participants

Participants with genetic and relevant phenotypic data were ascertained from the CADD projects; full details of participant selection for inclusion in the GWAS sample are provided in the Supplemental materials. GWAS participants were drawn from several primary studies described elsewhere (Hartman et al. 2008, Petrill et al. 2003, Rhea et al. 2006, Stallings et al. 2005). The current sample of 1,901 unrelated adolescents was over-selected for adolescent BD, with half of the participants ascertained specifically from high-risk populations (i.e. recruited through substance abuse treatment, special schools, or involvement with the criminal justice system; see Supplement for additional criteria for clinical probands). CADD GWAS participants had an average age of 16.5 (SD=1.4, range=13.0-19.9), 28.9% were female, and 37.3% of participants reported non-Caucasian ancestry (primarily Hispanic or African; see Supplemental Table S1 for complete demographic statistics).

Phenotype

BD was defined as a composite measure of substance dependence vulnerability (assessed across 10 substances), novelty seeking, and conduct disorder symptoms. The BD phenotype has been previously examined within the CADD samples, including Young et al. (2000) demonstrating that the component measures have loadings ≥0.4 on a common, highly heritable BD latent factor, and linkage analyses by Stallings et al. (2003, 2005). A full description of construction of the BD phenotype is provided in the Supplement; Supplemental Figure S1 shows the distribution of BD in the CADD GWAS sample. Briefly, principal component scores were

normed to community-representative samples in CADD and applied to all CADD GWAS participants from both the community-representative (48.2%) and high-risk samples (51.8%). Average scores on the BD composite measure were 0.19 (SD=1.2, range=-1.9–5.0) for the community-representative participants and 2.76 (SD=1.2, range=-0.3–6.7) for the high-risk participants.

Genotyping

All participants were genotyped on the Affymetrix 6.0 platform (Affymetrix, Inc., Santa Clara CA), with a total of 696,388 autosomal SNPs available for analysis after quality control. Full details on processing and cleaning genotypes for the CADD GWAS sample is provided in the Supplement. Population stratification was examined by performing multidimensional scaling in PLINK (Purcell et al. 2007), in which ten ancestry dimensions were estimated. The first three dimensions notably captured genetic variation among individuals of self-reported African, Hispanic, and Asian ancestry, compared to a central (majority) node of individuals of self-reported European ancestry. Supplemental Figure S2 illustrates the first three ancestry dimensions within the CADD GWAS sample (along with individuals' self-reported ancestry).

Analyses

Genome-wide analysis was conducted as a linear regression of the additive effect of each SNP on BD in PLINK (Purcell et al. 2007). All autosomal SNPs that passed basic quality controls were tested for association with BD, and 10 ancestry dimensions were included as covariates. Age and sex were accounted for in the estimation of the BD phenotype. The criterion for individual SNP significance was set at $p<5\times10^{-8}$.

Genome-wide data from the CADD GWAS sample were further characterized using Genome-wide Complex Trait Analysis (GCTA; Yang et al. 2011). GCTA allows us to estimate

the proportion of variance in the phenotype that may be explained using all of the genotyped SNPs using restricted maximum likelihood (REML) analysis. While this method does not specifically identify any causal variants, it does estimate the total proportion of sample variance that may be explained by all of the genotyped SNPs.

Gene-based tests were conducted using VEGAS (Liu et al. 2010), which aggregates evidence of association across all SNPs within a gene. A total of 16,094 autosomal genes were tested for association with BD in CADD, based on the primary GWAS results, with a multiple-testing-corrected significance threshold set at $p<3.1\times10^{-6}$.

INRICH (Lee et al. 2012) was selected to conduct our pathway analyses as it is well-suited for testing both large (i.e., exploratory) and small (i.e., candidate) pathway sets. We took two, complementary approaches to pathway analysis: first, we sought to confirm previously proposed candidate gene pathways (Hodgkinson et al., 2008); second, we conducted an exploratory analysis aimed at identifying novel pathways involved in BD (The Gene Ontology Consortium, 2000). Additional details of the pathway analysis methods are discussed in the Supplemental Materials.

Promising results from the pathway analysis of the CADD sample were followed up in two additional samples: the Minnesota Center for Twin and Family Research (MCTFR; *N*=3,378), a community-based adolescent sample (McGue et al. 2013, Miller et al. 2012), and the Study of Addiction: Genes and Environment (SAGE; *N*=3,988), a clinically over-selected study of addiction (Bierut et al. 2010; dbGaP study accession: phs000092.v1.p1). A phenotype similar to BD as defined in the CADD sample was available in the MCTFR sample (Hicks et al. 2010; McGue et al. 2013). The phenotype analysed in the SAGE sample was the average number of dependence symptoms for substances that each participant used. Full description of the MCTFR

and SAGE samples is provided in the Supplement.

Results

Figure 1 summarizes the GWAS results for BD in the over-selected CADD sample. No individual SNP reached genome-wide significance ($p<5\times10^{-8}$), nor did any SNP reach genome-wide significance in the MCFTR or SAGE samples (see Supplemental Figure S3 for QQ plots of the GWAS results from each study). Results from loci reaching $p<5.0\times10^{-5}$ in CADD are summarized in Table 1 (full GWAS results are available from the first author on request). The most significant SNP in the CADD GWAS was rs7104461 ($p=5.8\times10^{-6}$), an intergenic SNP on chromosome 11 for which there are no previously reported associated phenotypes. While this SNP was not genotyped in either the MCTFR or SAGE samples, it is in linkage disequilibium with rs341058 ($r^2=1.0$ in 1000 Genomes Pilot 1 CEU sample, distance = 8721bp; Johnson et al. 2008), which was genotyped on both MCTFR and SAGE platforms and may serve as a proxy to compare results across samples. This proxy SNP was not associated with either adolescent BD in MCTFR ($p=0.30$) or adult substance dependence symptoms in SAGE ($p=0.87$).

Whole-genome SNP-heritability was estimated with GCTA in the CADD sample. SNPs genotyped in the current study explained 27.8% of the CADD sample variance in BD ($SE=0.15$, $p=0.03$). The point estimate of heritability remained fairly stable when the sample was restricted to individuals estimated to be <2.5% identical-by-state ($N=1148$, $V(G)/Vp=30.9\%$, $SE=0.28$, $p=0.10$) or those individuals with only Caucasian ancestry (as determined by an examination of ancestry component plots, $N=1031$, $V(G)/Vp=49.3\%$, $SE=0.31$, $p=0.06$).

Gene-based association tests identified seven genes as significant after Bonferroni correction for testing >16,000 genes: *MAGI2* ($p<1.0\times10^{-6}$), *NAV2* ($p<1.0\times10^{-6}$), *CACNA1C* ($p=1.0\times10^{-6}$), *PCDH9* ($p=1.0\times10^{-6}$), *MYO16* ($p=1.0\times10^{-6}$), *IQCH* ($p=2.0\times10^{-6}$), *DLGAP1*

($p<1.0\times10^{-6}$). We examined overlap of these novel "candidate" genes derived from the CADD GWAS with results from MCTFR and SAGE as a single "pathway" (i.e., gene set) in INRICH (Lee 2012). This allowed us to estimate whether specific genes identified in the CADD results overlapped with the low *p*-value genomic regions (i.e., loci tagged at $r^2>0.5$ by a SNP reaching GWAS $p<5\times10^{-3}$) in the MCTFR and SAGE results more than expected by chance. The CADD-identified gene set was not significant in analysis of either the MCTFR (0 regions overlapped genes identified in CADD, $p=1.0$) or SAGE samples (6 regions overlapped genes identified in CADD, $p=0.14$).

Supplemental Table S2 presents gene-based association test results for previously identified addiction candidate genes (Hodgkinson et al. 2008), none of which were significant after adjustment for multiple testing (minimum $p=1.4\times10^{-3}$). Supplemental Table S3 gives results for each of the addiction candidate gene sets tested in CADD. None of the addiction candidate gene sets showed evidence of greater-than-chance overlap with low *p*-value genomic regions in the CADD GWAS (minimum $p=5.0\times10^{-1}$).

Promising pathways emerging from our exploratory pathway analysis were defined as those meeting nominal significance before correcting for multiple testing in CADD and either MCTFR or SAGE samples (Empirical $p<0.05$). Two pathways met these criteria: visual perception (Empirical $p_{CADD}=0.038$, $p_{MCTFR}=0.012$, $p_{SAGE}=0.22$) and phosphatidylcholine biosynthetic process (Empirical $p_{CADD}=0.039$, $p_{MCTFR}=1.0$, $p_{SAGE}=0.026$). Neither pathway achieved marginal significance in any sample after correction for multiple testing (i.e., Corrected $p<0.10$). Supplemental Table S4 provides results from all 72 pathways meeting Empirical $p<0.05$ in CADD (from a total of 3440 pathways tested) that were subsequently tested in the MCTFR and SAGE samples.

Discussion

No SNP was significantly associated with BD in the CADD GWAS. This is not surprising, given the relatively small sample. GWAS of psychiatric and behavioral phenotypes that have successfully identified and replicated individual effects of common SNPs have relied on very large samples (Rietveld et al. 2013; Ripke et al. 2013). Despite the lack of significance of any individual SNP, GCTA REML analysis estimated that 49.3% (*SE*=0.31, *p*=0.06) of the Caucasian ancestry sub-sample variation in BD could be accounted for by all of the genotyped SNPs. Conversely, a similar study found no evidence of variance in early adolescent (12-year-old) non-substance behavioral problems being attributable to common variants (Trzaskowski et al. 2013). This may suggest qualitative differences between genetic effects on BD at different ages, an effect that has been reported from twin models of comorbidity between dependence on different substances (Vrieze et al. 2012), which is a marker of BD.

Gene-based tests identified seven genes associated with BD in the CADD sample. However, neither the genes nor pathways identified as marginally overrepresented in the CADD GWAS results showed evidence of replicable low-*p*-values in either the MCTFR or SAGE samples. Taken together, these findings suggest that discoverable effects of common SNPs underlie the genetic architecture of BD, although better-powered studies are required to identify the associated loci.

The comparisons made between datasets must be considered in light of several limitations of the current study. There are substantial differences among the examined samples in terms of age (CADD and MCTFR represent adolescent data, while SAGE was comprised of adults), sex composition (MCTFR and SAGE are split evenly by sex, while CADD has an overrepresentation of males due to the sampling scheme), and diversity of ancestry (MCTFR is

less diverse than either CADD or SAGE, which each have different representations of non-Caucasian ancestry groups). The sampling schemes of CADD and SAGE aimed to increase power to detect effects by oversampling extreme phenotype individuals, whereas the MCTFR study is closer to community-representative.

We sought to identify genetic influences on adolescent BD through a multifaceted approach. We initially characterized results from a standard GWAS by estimating the variance explained by common SNPs, and used gene- and pathway-based tests to identify potential novel candidate genes and pathways. Results from the estimation of sample variance explained by all genotyped SNPs and significant gene-based tests suggest there is a real genetic signal to be detected within the noise. However, the current sample is likely underpowered to detect realistic effect sizes of individual SNPs. Further, the lack of correspondence between pathway analyses in the CADD and replication samples may be due to limited power, or qualitative differences in the genetic effects on BD across different ages (adolescent versus adult) or sampling distributions (over-sampled for BD versus community-representative). Key to the search for causal genetic pathways underlying BD will be the availability of increasingly large, thoroughly phenotyped samples. Although the current analyses did not identify specific loci associated with BD, we demonstrate substantial heritability due to effects of common SNPs. Larger studies with appropriate phenotypes could well allow successful identification of common variants associated with BD.

Acknowledgments

The Center on Antisocial Drug Dependence (CADD) data reported here were funded by grants from the National Institute on Drug Abuse (P60 DA011015, R01 DA012845, R01 DA021913, R01 DA021905).

The Minnesota Center for Twin and Family Research (MCTFR) was supported in part by USPHS Grants from the National Institute on Alcohol Abuse and Alcoholism (AA09367 and AA11886), the National Institute on Drug Abuse (DA05147, DA13240, and DA024417), and the National Institute of Mental Health (MH066140).

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 datasets 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). 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" (HHSN268200782096C). The datasets used for the analyses described in this manuscript were

obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1 through dbGaP accession number phs000092.v1.p.

Jaime Derringer was supported by the National Institute of Mental Health (T32 MH016880).

References

Agrawal, A., Lynskey, M. T., Hinrichs, A., et al. (2011). A genome-wide association study of DSM-IV cannabis dependence. *Addiction biology*, *16*(3), 514-518.

Bierut, L. J., Agrawal, A., Bucholz, K. K., et al. (2010). A genome-wide association study of alcohol dependence. *Proceedings of the National Academy of Sciences*, *107*(11), 5082-5087.

Bierut, L. J., Madden, P. A., Breslau, N., et al. (2007). Novel genes identified in a high-density genome wide association study for nicotine dependence. *Human molecular genetics*, *16*(1), 24-35.

Bierut, L. J., Strickland, J. R., Thompson, J. R., et al. (2008). Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings. *Drug and alcohol dependence*, *95*(1), 14-22.

Collins, A. L., Kim, Y., Sklar, P., et al. (2012). Hypothesis-driven candidate genes for schizophrenia compared to genome-wide association results. *Psychological medicine*, *42*(3), 607.

Dick, D. M., Aliev, F., Krueger, R. F., et al. (2011). Genome-wide association study of conduct disorder symptomatology. *Molecular psychiatry*, *16*(8), 800-808.

Edenberg, H. J. (2002). The collaborative study on the genetics of alcoholism: an update. *Alcohol Research and Health*, *26*(3), 214-218.

Edenberg, H. J., Koller, D. L., Xuei, X., et al. (2010). Genome-Wide Association Study of Alcohol Dependence Implicates a Region on Chromosome 11. *Alcoholism: Clinical and Experimental Research, 34*(5), 840-852.

Frank, J., Cichon, S., Treutlein, J., et al. (2012). Genome-wide significant association between

alcohol dependence and a variant in the ADH gene cluster. *Addiction Biology, 17*(1), 171-180.

Furberg, H., Kim, Y., Dackor, J., et al. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*, *42*(5), 441-U134.

Gelernter, J., Kranzler, H. R., Sherva, R., et al. (2014a). Genome-wide association study of alcohol dependence: significant findings in African-and European-Americans including novel risk loci. *Molecular Psychiatry, 19*, 41-49.

Gelernter, J., Kranzler, H. R., Sherva, R., et al. (2014b). Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. *Biological Psychiatry, 76*(1), 66-74.

Gelernter, J., Sherva, R., Koesterer, R., et al. (2014c). Genome-wide association study of cocaine dependence and related traits: FAM53B identified as a risk gene. *Molecular Psychiatry, 19*(6), 717-723.

Hartman, C. A., Gelhorn, H., Crowley, T. J., et al. (2008). Item Response Theory Analysis of DSM-IV Cannabis Abuse and Dependence Criteria in Adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(2), 165-173.

Hicks, B. M., Foster, K. T., Iacono, W. G., et al. (2013). Genetic and Environmental Influences on the Familial Transmission of Externalizing Disorders in Adoptive and Twin Offspring. *JAMA psychiatry*, *70*(10), 1076-1083.

Hicks, B. M., Schalet, B. D., Malone, S. M., et al. (2011). Psychometric and genetic architecture of substance use disorder and behavioral disinhibition measures for gene association studies. *Behav Genet*, *41*(4), 459-475.

Hodgkinson, C. A., Yuan, Q., Xu, K., et al. (2008). Addictions biology: haplotype-based analysis

for 130 candidate genes on a single array. *Alcohol and Alcoholism*, *43*(5), 505-515.

Johnson, A. D., Handsaker, R. E., Pulit, S. L., et al. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, *24*(24), 2938-2939.

Kapoor, M., Wang, J. C., Wetherill, L., et al. (2013). A meta-analysis of two genome-wide association studies to identify novel loci for maximum number of alcoholic drinks. *Human Genetics, 132*(10), 1141-1151.

Krueger, R. F., Hicks, B. M., Patrick, C. J., et al. (2002). Etiologic connections among substance dependence, antisocial behavior and personality: Modeling the externalizing spectrum. *J Abnorm Psychol*, *111*(3), 411.

Lee, P. H., O'Dushlaine, C., Thomas, B., et al. (2012). INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics*, *28*(13), 1797-1799.

Lee, S. H., Wray, N. R., Goddard, M. E., et al. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet*, *88*(3), 294-305.

Li, X., Basu, S., Miller, M. B., et al. (2011). A rapid generalized least squares model for a genome-wide quantitative trait association analysis in families. *Human heredity*, *71*(1), 67-82.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., et al. (2010). A versatile gene-based test for genome-wide association studies. *Am J Hum Genet*, *87*(1), 139-145.

Liu, J. Z., Tozzi, F., Waterworth, D. M., et al. (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nature Genetics, 42*(5), 436-440.

McGue, M., Zhang, Y., Miller, M. B., et al. (2013). A genome-wide association study of behavioral disinhibition. *Behav Genet*, *43*(5), 363-373.

Miller, M. B., Basu, S., Cunningham, et al. (2012). The Minnesota Center for Twin and Family Research genome-wide association study. *Twin research and human genetics*, *15*(6), 767-774.

Nielsen, D. A., Ji, F., Yuferov, V., et al. (2010). Genome-wide association study identifies genes that may contribute to risk for developing heroin addiction. *Psychiatric genetics*, *20*(5), 207-214.

Petrill, S., Plomin, R., DeFries, J. C., Hewitt, J. K., editors (2003) Nature, Nurture, and the Transition to Adolescence. New York: Oxford University Press.

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

Rhea, S. A., Gross, A. A., Haberstick, B. C., et al. (2006). Colorado twin registry. *Twin Research and Human Genetics*, *9*(06), 941-949.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., ... & McMahon, G. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *science*, *340*(6139), 1467-1471.

Ripke, S., O'Dushlaine, C., Chambert, K., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet*, *45*(10), 1150-1159.

Ruano, D., Abecasis, G. R., Glaser, B., et al. (2010). Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. *Am J Hum Genet*, *86*(2), 113-125.

Schumann, G., Coin, L. J., Lourdusamy, A., et al. (2011). Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. *Proceedings of the National Academy of Sciences,*

108(17), 7119-7124.

Stallings, M. C., Corley, R. P., Dennehey, B., et al. (2005). A genome-wide search for quantitative trait loci that influence antisocial drug dependence in adolescence. *Archives of General Psychiatry*, *62*(9), 1042-1051.

Sullivan, P. (2011). Don't give up on GWAS. *Molecular psychiatry*, 17(1), 2-3.

Terracciano, A., Esko, T., Sutin, A. R., et al. (2011). Meta-analysis of genome-wide association studies identifies common variants in CTNNA2 associated with excitement-seeking. *Translational psychiatry*, *1*(10), e49.

The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. *Nat Genet*, *25*(1), 25-29.

Thorgeirsson, T. E., Gudbjartsson, D. F., Surakka, I., et al. (2010). Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet*, *42*(5), 448-453.

Tielbeek, J. J., Medland, S. E., Benyamin, B., et al. (2012). Unraveling the genetic etiology of adult antisocial behavior: A genome-wide association study. *PloS one*, *7*(10), e45086.

Treutlein, J., Cichon, S., Ridinger, M., et al. (2009). Genome-wide association study of alcohol dependence. *Archives of general psychiatry*, *66*(7), 773-784.

Trzaskowski, M., Dale, P. S., & Plomin, R. (2013). No genetic influence for childhood behavior problems from DNA analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(10), 1048-1056.

Uhl, G. R., Drgon, T., Liu, Q. R., et al. (2008). Genome-wide association for methamphetamine dependence: convergent results from 2 samples. *Archives of general psychiatry*, *65*(3), 345-355.

Van der Sluis, S., Posthuma, D., Nivard, M. G., et al. (2013). Power in GWAS: lifting the curse

of the clinical cut-off. *Molecular psychiatry*, *18*, 2-3.

Verweij, K. J., Vinkhuyzen, A. A., Benyamin, B., et al. (2013). The genetic aetiology of cannabis use initiation: a meta□analysis of genome□wide association studies and a SNP□based heritability estimation. *Addiction biology*, *18*(5), 846-850.

Vrieze, S. I., Hicks, B. M., Iacono, W. G., et al. (2012). Decline in genetic influence on the co-occurrence of alcohol, marijuana, and nicotine dependence symptoms from age 14 to 29. *American Journal of Psychiatry*,*169*(10), 1073-1081.

Wang, J. C., Foroud, T., Hinrichs, A. L., et al. (2012). A genome-wide association study of alcohol-dependence symptom counts in extended pedigrees identifies C15orf53. *Molecular Psychiatry, 18*(11), 1218-1224.

Yang, J., Lee, S. H., Goddard, M. E., et al. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*, *88*(1), 76-82.

Young, S. E., Stallings, M. C., Corley, R. P., et al. (2000). Genetic and environmental influences on behavioral disinhibition. *American journal of medical genetics*, *96*(5), 684-695.

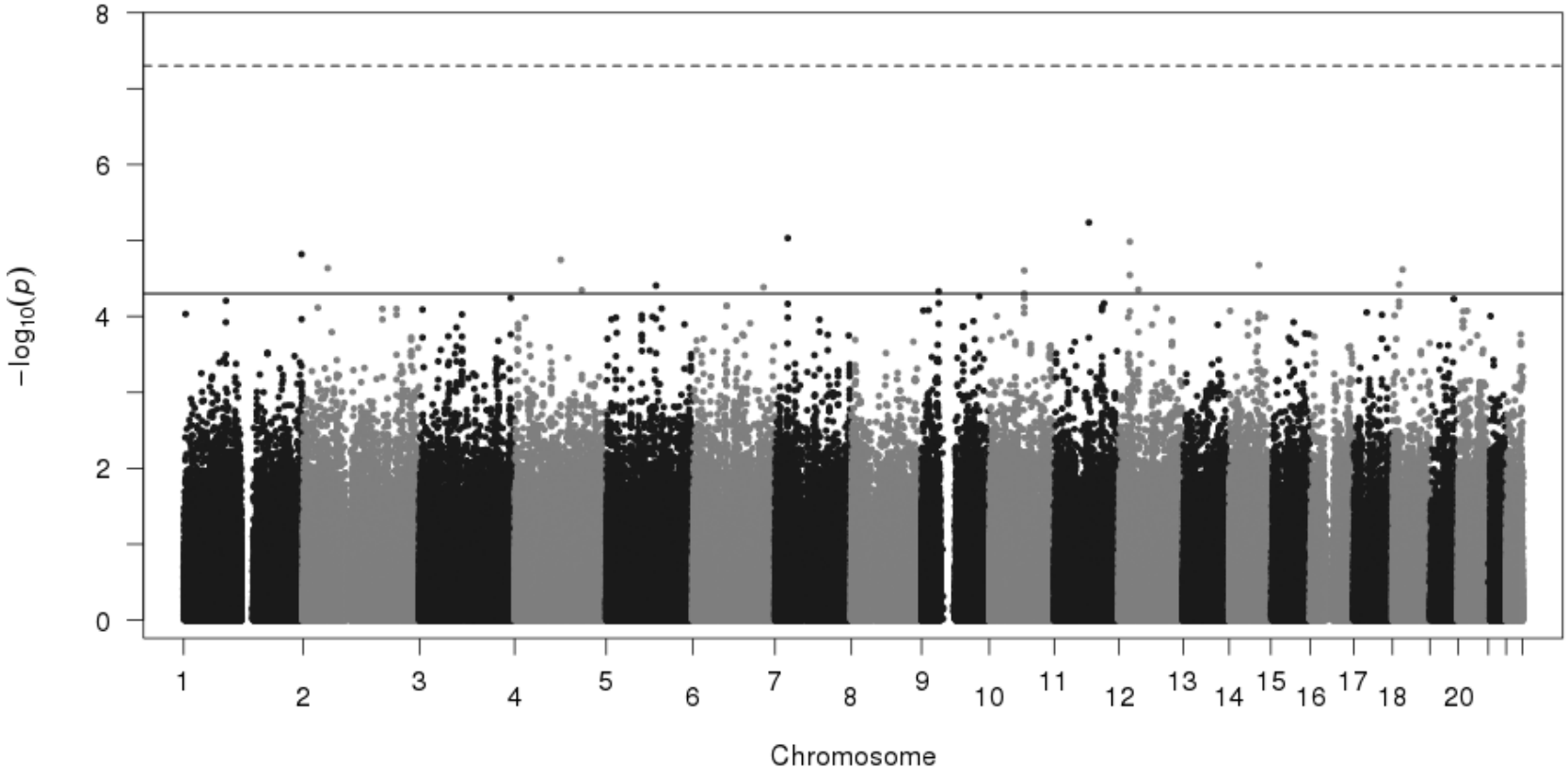
Yu, K., Li, Q., Bergen, A. W., et al. (2009). Pathway analysis by adaptive combination of P-values. *Genetic epidemiology*, *33*(8), 700-709.

Table 1. Top associated loci from the CADD GWAS.

	Index SNP description				*Association*		*Locus description*					
Index SNP	*A1*	*A2*	*MAF*	*HWE P*	*Beta*	*P*	*Chr*	*Start*	*Stop*	*N*	*Kb*	*Genes*
rs4654186	C	T	0.467	5.7E-3	-0.248	1.5E-5	1	246168017	246230729	6	62.7	*SMYD3*
rs11562945	A	G	0.029	2.1E-1	0.720	2.3E-5	2	51671180	51671180	1	0.0	
rs2114532	A	C	0.025	2.8E-2	0.780	1.8E-5	4	96129959	96130145	2	0.2	*UNC5C*
rs17050678	A	G	0.202	6.2E-1	-0.289	4.5E-5	4	139945108	140143272	9	198.2	*PPP1R14BP3,ELF2,CCRN4L*
rs10464002	T	C	0.164	1.8E-1	0.317	3.9E-5	5	103549477	103657217	4	107.7	
rs10485364	T	G	0.011	2.0E-1	1.183	4.1E-5	6	147106285	147106285	1	0.0	*LOC729176,ADGB*
rs12666574	C	T	0.326	7.9E-1	-0.272	9.3E-6	7	26491618	26526960	2	35.3	*LOC441204*
rs10972581	A	G	0.407	2.7E-1	-0.238	4.7E-5	9	35779571	35870001	6	90.4	*TMEM8B,SPAG8,OR13J1, OR13E1P,NPR2,LOC100128136, HINT2,FP588,FAM221B*
rs10509330	T	C	0.289	6.2E-1	-0.265	2.5E-5	10	72898795	72944032	12	45.2	
rs7104461	A	C	0.132	1.6E-1	-0.392	5.8E-6	11	72394446	72404958	2	10.5	*RPS12P20,PDE2A,ARAP1*
rs901625	G	T	0.419	5.7E-1	0.251	1.0E-5	12	22787796	22857819	10	70.0	*RPS27P22,ETNK1*
rs11175260	C	T	0.070	2.9E-1	-0.459	4.4E-5	12	40387225	40601708	7	214.5	*SLC2A13,RPL30P13,LRRK2*
rs1652591	T	C	0.417	4.8E-1	-0.250	2.1E-5	14	82542483	82559572	3	17.1	
rs7233911	A	G	0.293	8.2E-1	0.261	3.8E-5	18	13927552	13931827	3	4.3	*RPL36AP49,MC2R*
rs16940157	A	C	0.004	1.0E+0	1.849	2.4E-5	18	21149803	21149803	1	0.0	*NPC1*

Note. Index SNP = most significant SNP tagging the LD block; A1 = tested minor allele; A2 = alternate major allele; MAF = minor allele frequency; HWE P = Hardy-Weinberg Equilibrium p-value; Beta = linear regression coefficient; P = association p-value; Chr, Start, Stop = location of the LD block tagged by the Index SNP; N = number of tested SNPs in the LD block; Kb = size of the LD block; Genes = genes overlapping the LD block.

Figure 1. Plot of -$\log_{10}(p)$ from the CADD GWAS, arranged by chromosomal location. The top (dashed) horizontal line indicates genome-wide significance at $p=5\times10^{-8}$; the lower (solid) line marks $p=5\times10^{-5}$ (loci described in Table 1).



Supplementary Methods and Results

Supplementary Tables and Figures

Supplemental Methods and Results

Center on Antisocial Drug Dependence (CADD)

Phenotype definition. We created a behavioral disinhibition (BD) composite score for every subject aged 13-19 from the studies that are part of the Center on Antisocial Drug Dependence (Hartman 2008 et al., Petrill et al. 2003, Rhea 2006 et al., Stallings et al. 2005). BD was defined by three measures: 1) dependence vulnerability (Stallings et al. 2003, 2005); 2) Lifetime conduct disorder symptom counts from either the Diagnostic Interview Schedule (DIS; Robins et al. 1981) or the Diagnostic Interview Schedule for Children (DISC; Shaffer et al. 2000); and 3) novelty seeking from the Tridimensional Personality Questionnaire (TPQ; Cloninger et al. 1991) or the Temperament and Character Inventory (TCI; Cloninger et al.1994). We did not include self-reported ADHD because it was unavailable for a sizable portion of the sample. Each measure was adjusted for type of test (DSM-IIIR or DSM-IV for symptom measures, TPQ or TCI for novelty seeking), type of administration (paper-and-pencil or computer-administered), and age and age-squared within sex within only the community-selected samples. Regression weights were then applied to all subjects, including the clinical samples and their siblings. We then standardized the scores again, to create equivalent scores for each sex with a mean of 0 and standard deviation of 1 within the community subjects. Principal component scores were used to create a composite BD score from the three normed measures (dependence vulnerability, conduct disorder, and novelty seeking). Finally, the saved first component was standardized within sex from the community norms. The principal component score coefficients were 0.499 for conduct disorder, 0.490 for dependence vulnerability, and 0.290 for novelty seeking.

Criteria for clinical probands. “Clinical” probands (i.e., participants selected as part of a

high-risk BD sample) were defined as those who met the following criteria at study intake: (1) age 14-19 years, (2) full-scale IQ >80, (3) had one or more lifetime non-tobacco substance dependence symptoms, (4) had one or more lifetime Conduct Disorder symptoms, (5) were more than one standard deviation above the community mean on a composite measure of combined conduct problems and substance dependence symptoms (Stallings et al., 2005), and (6) were ascertained as a high risk sample (i.e. recruited through substance abuse treatment, special schools, or involvement with the criminal justice system). In addition, for participants 17 years of age or younger, (7) valid written consent from parent, guardian, or custodial agency, together with assent from the subject, were required. At the time of first contact, exclusion criteria for probands were: (1) psychosis; (2) current serious risk of suicide, violence, or fire setting (though many probands do have these problems in their past histories); (3) insufficient English skills for assenting/consenting or completing interviews.

Selection of participants for genotyping. The majority of participants identified for genotyping were recruited at project centers in Boulder, Colorado, USA and Denver, Colorado, USA. A smaller sample of individuals with extremely high BD phenotypes was recruited from a project site in San Diego, California, USA. Within the families of clinical probands, we selected subjects with a BD score equal to or greater than 1. We then ranked within family for highest BD score and selected the individual with the highest score for genotyping. In the community-based families, we first identified families where at least one family member had a BD score greater than the clinical proband family average (2.634) and selected the most extreme individual for genotyping. From the remaining non-clinical families, we prioritized families from the Longitudinal Twin Study (Rhea et al., 2006;2013), Colorado Twin Study with executive function data (Rhea et al., 2006;2013), and Family Study (Stalling et al., 2003, 2005) community samples

(to maximize availability of phenotypic measures for future study), as well as families that reported non-Caucasian or Hispanic ancestry and gave greater priority to males (to roughly sex-match the heavily male clinical sample). We then included a random number within the priority score and selected the individual with the highest priority from within each family, resulting in essentially a random draw within families. Additional family members were genotyped as time and finances allowed. If genotyping of a target proband failed but genotyping for a sibling succeeded, the sibling replaced the target proband in the final CADD GWAS sample.

Genotyping and quality control. In total, we attempted genotyping of 2776 samples from 1985 families (where "family" includes biologically related individuals and/or sample duplicates). 43% of samples came from genotyped families of two or more samples (average family size = 3.0). Of these genotyped individuals, 68% provided saliva while the remaining participants provided whole-blood samples. Individual samples were randomized into batches of 48 subjects each. All participants were genotyped on the Affymetrix 6.0 array (Affymetrix, Inc., Santa Clara CA). Intensity data were normed separately by sample type (blood versus saliva) and any samples with a quality control score less than 0.8 in Birdseed was removed. Genotype calls were made using BEAGLECALL (Browning & Yu 2009) with 6 iterations, each with increasingly stringent calling parameters. Potential cryptic relatedness and sample contamination were further investigated among the Caucasian samples using pi-hat estimates from PLINK (Purcell et al. 2007). Genotypes from 83 chips were dropped due to identity or quality problems (including excess heterozygosity, mismatch between genotyped and self-reported sex, and apparent unrelatedness with duplicate or family samples). This represents an overall sample failure rate around 3%. An additional 24 chips whose identities could not be positively confirmed were dropped from a single batch with a high rate of misidentified chips. After full

quality control for sample failures, contamination, low call rates, and potential sample mix-ups, and dropping related individuals and sample duplicates, our final sample included 1,901 unrelated individuals with estimated behavioral disinhibition (BD) scores. Of the final BD sample, 34.7% had available duplicate, family-member, or previously completed genotypes available, allowing positive confirmation of identities of these samples.

Minnesota Center for Twin and Family Research (MCTFR)

Sample. MCTFR is a family-based study, including parents and their biological or adoptive offspring. While the MCTFR is broadly community-representative in terms of being unselected for specific behavioral outcomes, a subset termed the "Enrichment Study" (comprising 17% of the GWAS sample analyzed here) was over-selected for twin pairs where at least one of the pair demonstrated childhood behavioral problems (such as symptoms of ADHD, conduct disorder, or academic disengagement). These participants were recruited to increase sampling of individuals at risk for later disinhibition-related outcomes (Keyes et al. 2009). For the current project, we utilized data from the MCTFR offspring generation only, at their assessment around age 17, to improve the potential similarity in genetic etiology between the MCTFR and CADD samples. That is, while the MCTFR participants represent a less severe sample in which to examine BD, the availability of nearly identical phenotypes (e.g., inclusion of a non-substance measure of disinhibition) at similar developmental stages (i.e., adolescence) in both the CADD and MCTFR samples may improve our ability to make interpretable comparisons between samples. MCTFR participants had an average age of 17.9 (SD=0.78, range=16-21), 53.3% were female, and individuals included in the GWAS were restricted to those of non-Hispanic Caucasian ancestry.

Phenotype. The target phenotype was a higher-order factor score of BD, defined using

both substance and non-substance antisocial behaviors, as described in (Hicks et al. 2011; McGue et al. 2013).

Genotyping. Participants were genotyped on the Illumina Human660W-Quad array (Illumina, Inc., San Diego, CA), with a total of 515383 autosomal SNPs available after quality control. Ten ancestry principle components were estimated to account for population stratification. Details of the quality control procedures and ancestry principal components estimation in the MCTFR are described in Miller et al. (2012).

Study of Addiction: Genes and Environment (SAGE)

Sample. SAGE was designed as a study of adult alcohol dependence, including 3988 unrelated genotyped individuals (50% of whom were selected as alcohol dependent cases) drawn from three primary studies of alcohol (COGA, Edenberg et al. 2002), nicotine (COGEND, Bierut et al. 2007), and cocaine dependence (FSCD, Bierut et al. 2008). SAGE participants had an average age of 39.0 (SD=9.1, range=18-77), 54.3% were female, and 35.6% of participants reported Hispanic and/or African ancestry, with the remainder of the sample being non-Hispanic Caucasian. Although the available phenotype and age of assessment differ substantially between the CADD and SAGE studies, the similar sampling methods aimed at over-including individuals in the high extreme, clinically significant range may provide better power to detect genetic effects, to the extent that these effects would be more difficult to detect in a community sample where there is less variance in BD or substance dependence.

Phenotype. Based on available data, the target phenotype was the average number of substance dependence symptoms endorsed by a participant for any substance they reported having ever used, including alcohol, nicotine, cannabis, cocaine, opiates, and other drugs.

Genotyping. All participants in SAGE were genotyped on the Illumina Human1Mv1_C

BeadChip array (Illumina, Inc., San Diego, CA), with a total of 917694 autosomal SNPs available after quality control. Details of the quality control procedures in SAGE are described in Bierut et al. (2010). SAGE genotypes were accessed via the National Center for Biotechnology Information's database of Genotypes and Phenotypes (dbGaP; study accession: phs000092.v1.p1). Genomic ancestry components were estimated using the same method applied in the CADD sample, extracting 10 genomic ancestry dimensions by performing multidimensional scaling in PLINK.

Pathway analysis

Pathway analyses provide a promising avenue for identifying and validating candidate biological systems involved in the etiology of psychiatric and behavioral phenotypes. Pathway approaches seek to demonstrate whether regions of significance in GWAS results tend occur in genes clustered into pre-defined "pathways". For some phenotypes, we may have strong hypotheses about which biological pathways or sets of candidate genes are likely to be involved in the etiology of a disease; for others, we may be searching for new candidates. Even in cases where prior evidence suggests strong candidate genes or pathways, these are not always borne out in thorough large-scale analyses.

Pathway analysis of genome-wide data is a rapidly developing area, in terms of both theory and application. We took two, complementary approaches to pathway analysis: first, we sought to confirm previously proposed candidate gene pathways; second, we conducted an exploratory analysis aimed at identifying novel pathways involved in BD. In the confirmatory analysis, we tested a predefined set of pathways composed of genes widely theorized to influence addiction (a component of BD strongly associated with the overarching phenotype). In the exploratory analysis, we sought to identify novel pathways for BD by comparing the CADD

GWAS results to pathways defined in the Gene Ontology database (GO; The Gene Ontology Consortium 2000). For any pathways identified as marginally significant in the CADD GWAS results, we then sought replication of pathway association in both the MCTFR and SAGE samples.

Pathway analysis generally proceeds along two steps: first we identify genomic regions showing association with the phenotype; and secondly we test whether these regions overlap with genes clustered within pathways more than would be expected by chance. We defined low *p*-value genomic regions as those that included one or more SNPs with association $p<5\times10^{-3}$, and extending to surrounding SNPs meeting $r^2>0.5$ with the index SNP (the lowest-*p*-value SNP in the region) and association $p<5\times10^{-2}$. These low *p*-value genomic regions were estimated from each study's GWAS results in PLINK.

INRICH (Lee et al. 2012) was selected to conduct our pathway analyses as it is well-suited for testing both large (i.e., exploratory) and small (i.e., candidate) pathway sets and assesses overlap between pathways and associated genomic regions, rather than specific SNPs. This makes it easier to compare results from samples genotyped on different platforms, without either removing non-overlapping SNPs or requiring an intermediate step of imputation. INRICH's two-step permutation procedure produces two *p*-values: an *empirical P* that takes into account genomic coverage of both the pathway and the tested SNPs, and a *corrected P*, adjusted for testing multiple pathways. INRICH was run with standard settings (with 10^6 permutations in the first step, and 10^4 in the second), and gene locations were defined by NCBI build 37.2.

Candidate genes. We sought to test whether candidate gene sets including commonly studied genes for addiction and related behaviors were over-represented among the low *p*-value genomic regions in the CADD GWAS. These candidate gene sets were defined by autosomal

genes identified by Hodgkinson et al. (2008), sorted into 13 subgroups on the basis of the biological system with which the gene is primarily associated (for example, serotonergic, dopaminergic, opioid, etc. or "other"). In addition, we tested an omnibus candidate gene set comprised of all 127 autosomal genes (regardless of biological system assignment). While the candidate gene sets highlighted here do not represent a comprehensive list of all candidate genes for addiction, or even all genes potentially interacting within the identified candidate pathways, they are intended as a clearly defined and widely used list of candidate genes for addiction and related phenotypes, such as BD, that may be illustrative for purposes of comparison to the results of exploratory gene- and pathway-based tests. All 13 candidate gene sets, plus the omnibus set of candidate genes, were tested for over-representation of low *p*-value regions in the CADD GWAS results. None of the candidate gene sets showed evidence of greater-than-chance overlap with low *p*-value genomic regions in the CADD GWAS (Supplemental Table S3).

Exploratory pathway analysis. We next tested all pathways in the Gene Ontology database (GO; The Gene Ontology Consortium 2000) that included between 5 and 200 genes. GO is a freely available database of pathways defined by known biological function. It is not limited to specific phenotypes, and therefore provides a range of pathways for identification of novel pathways related to BD. A total of 3440 pathways tested were tested for overrepresentation of low *p*-value regions in the CADD GWAS results. 72 pathways meeting empirical $p<0.05$ (before multiple testing correction) were identified in CADD and subsequently tested for replication in MCTFR and SAGE.

Replication in MCTFR and SAGE. Pathways that showed nominal significance in CADD (defined by Empirical $p<0.05$) were tested for replication using the same testing parameters in INRICH, to estimate whether low *p*-value genomic regions from the MCTFR or SAGE GWAS

results overlapped with pathways identified in CADD more than expected by chance. Promising pathways emerging from our exploratory analysis were defined as those meeting nominal significance before correcting for multiple testing in CADD and either MCTFR or SAGE samples (Empirical $p<0.05$). Two pathways met these criteria: visual perception (Empirical $p_{CADD}=0.038$, $p_{MCTFR}=0.012$, $p_{SAGE}=0.22$) and phosphatidylcholine biosynthetic process (Empirical $p_{CADD}=0.039$, $p_{MCTFR}=1.0$, $p_{SAGE}=0.026$). INRICH pathway analysis results for all 72 pathways tests in CADD, MCTFR, and SAGE are presented in Supplemental Table S4).

Method comparison within CADD. Because pathway analysis of SNP data is still rapidly developing, no single method has yet been established as a best practice approach. As such, we sought to compare the primary results from INRICH to an alternate approach. The candidate-gene-based pathways were tested via adaptive rank truncated product (ARTP, Yu et al. 2009), which utilizes a permutation approach modeling the raw genotypic and phenotypic data (compared to INRICH, which utilizes GWAS p-values). ARTP provides two approaches to estimating pathway p-values: a gene-based approach and a SNP-based approach (i.e., irrespective of gene membership). We ran ARTP with 1000 permutations in the CADD sample for each pathway identified as promising in the exploratory pathway analysis conducted in INRICH (that is, those meeting Empirical $p<0.05$ in CADD and either MCTFR or SAGE samples). *P*-values were comparable between pathway analysis methods for both the visual perception ($p_{INRICH}=0.038$, $p_{ARTP\text{-}gene}=0.071$, $p_{ARTP\text{-}SNP}=0.067$) and phosphatidylcholine biosynthetic process ($p_{INRICH}=0.039$, $p_{ARTP\text{-}gene}=0.084$, $p_{ARTP\text{-}SNP}=0.040$) pathways.

References

Bierut, L. J., Agrawal, A., Bucholz, K. K., et al. (2010). A genome-wide association study of alcohol dependence. *Proceedings of the National Academy of Sciences*, *107*(11), 5082-5087.

Bierut, L. J., Madden, P. A., Breslau, N., et al. (2007). Novel genes identified in a high-density genome wide association study for nicotine dependence. *Human molecular genetics*, *16*(1), 24-35.

Bierut, L. J., Strickland, J. R., Thompson, J. R., et al. (2008). Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings. *Drug and alcohol dependence*, *95*(1), 14-22.

Browning, B. L., & Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet*, *85*(6), 847-861.

Cloninger, C. R., Przybeck, T. R., Svrakic, D. M., & Wetzel, R. D. (1994). The Temperament and Character Inventory (TCI): a guide to its development and use. St. Louis, MO: Center for Psychobiology of Personality, Washington University.

Cloninger, C. R., Przybeck, T. R., & Svrakic, D. M. (1991). The tridimensional personality questionnaire: US normative data. *Psychological reports*, *69*(3), 1047-1057.

Edenberg, H. J. (2002). The collaborative study on the genetics of alcoholism: an update. *Alcohol Research and Health*, *26*(3), 214-218.

Hartman, C. A., Gelhorn, H., Crowley, T. J., et al. (2008). Item Response Theory Analysis of DSM-IV Cannabis Abuse and Dependence Criteria in Adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(2), 165-173.

Keyes, M. A., Malone, S. M., Elkins, I. J., et al. (2009). The enrichment study of the Minnesota twin family study: increasing the yield of twin families at high risk for externalizing psychopathology. *Twin Research and Human Genetics*, *12*(05), 489-501.

Lee, P. H., O'Dushlaine, C., Thomas, B., & Purcell, S. M. (2012). INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics*, *28*(13), 1797-1799.

McGue, M., Zhang, Y., Miller, M. B., et al. (2013). A genome-wide association study of behavioral disinhibition. *Behav Genet*, *43*(5), 363-373.

Petrill S., Plomin R., DeFries J. C., Hewitt J. K., editors (2003) Nature, Nurture, and the Transition to Adolescence. New York: Oxford University Press.

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

Rhea, S. A., Gross, A. A., Haberstick, B. C., & Corley, R. P. (2006). Colorado twin registry. *Twin Research and Human Genetics*, *9*(06), 941-949.

Rhea, S. A., Gross, A. A., Haberstick, B. C., & Corley, R. P. (2013). Colorado Twin Registry—An Update. *Twin research and human genetics*, *16*(1), 351-7.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health diagnostic interview schedule: its history, characteristics, and validity. *Archives of general psychiatry*, *38*(4), 381-389.

Shaffer, D., Fisher, P., Lucas, et al. (2000). NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(1), 28-38.

Stallings, M. C., Corley, R. P., Dennehey, B., et al. (2005). A genome-wide search for quantitative trait loci that influence antisocial drug dependence in adolescence. *Archives of General Psychiatry*, *62*(9), 1042-1051.

Stallings, M. C., Corley, R. P., Hewitt, J. K., et al. (2003). A genome-wide search for quantitative trait loci influencing substance dependence vulnerability in adolescence. *Drug and alcohol dependence*, *70*(3), 295-307.

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1), 25-29.

Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., et al. (2009). Pathway analysis by adaptive combination of P-values. *Genetic epidemiology*, *33*(8), 700-709.

Supplemental Table S1. Sample demographics

	CADD	MCTFR	SAGE
N	1901	3378	3988
Phenotype [M(SD)]	1.52 (1.77)	-0.11 (0.34)	2.09 (1.72)
Age [years M(SD)]	16.5 (1.44)	17.9 (0.78)	39.1 (9.12)
Sex			
Male	71.1%	46.7%	45.9%
Female	28.9%	53.3%	54.1%
Self-reported ancestry			
White	59.4%	100.0%	65.1%
Hispanic	25.1%	0.0%	3.5%
Black	6.0%	0.0%	31.3%
Native American	2.2%	0.0%	0.0%
Asian / Pacific Islander	2.1%	0.0%	0.0%
Other / mixed / none reported	5.2%	0.0%	0.1%
Sample representation			
Community	48.2%	83.0%	48.8%
Clinical (see Notes)	51.8%	17.0%	51.2%

Notes.

Phenotype:

CADD = Behavioral disinhibition factor score;
MCTFR = Behavioral disinhibition factor score;
SAGE = average substance dependence symptoms.

Self-reported ancestry: For all studies, individuals were categorized as Hispanic if they reported any Hispanic ancestry; individuals were categorized by their non-Hispanic racial category if they did not report Hispanic ancestry.

Sample representation:

Community in CADD and MCTFR indicates community-representative sample;
Community in SAGE are individuals who do not qualify for DSM-IV substance dependence;
Clinical in CADD are individuals ascertained from substance abuse treatment, special schools, or involvement with the criminal justice system;
Clinical in MCTFR are individuals over-selected for increased rates of ADHD, conduct disorder, and academic disengagement;
Clinical in SAGE are individuals who meet DSM-IV substance dependence.

Supplemental Table S2. Gene-based *p*-values for candidate genes (Hodgkinson et al. 2008).

Gene	*P*	*Gene*	*P*	*Gene*	*P*
GRIK1	1.4E-3	*OPRM1*	3.3E-1	*AVPR1A*	5.4E-1
CHRM3	2.7E-3	*PENK*	3.4E-1	*FOSL1*	5.5E-1
CNR1	6.5E-3	*CHRNA4*	3.5E-1	*GABRA4*	5.5E-1
TH	8.5E-3	*CRH*	3.5E-1	*CARTPT*	5.5E-1
CRHBP	1.7E-2	*TAC1*	3.6E-1	*CHRM2*	5.5E-1
MAPK14	4.9E-2	*GABRG3*	3.6E-1	*OPRL1*	5.6E-1
DBI	5.4E-2	*DRD5*	3.6E-1	*SLC32A1*	5.6E-1
ADRA1A	5.6E-2	*ADH4*	3.6E-1	*MPDZ*	5.8E-1
ALDH1A1	6.2E-2	*NGF*	3.7E-1	*NPY5R*	5.8E-1
GRM1	1.0E-1	*ADRA2A*	3.7E-1	*CCKAR*	5.9E-1
PRKCE	1.1E-1	*GAD2*	3.7E-1	*GLRA1*	6.2E-1
ADH1C	1.1E-1	*PNOC*	3.8E-1	*CHRM1*	6.3E-1
BDNF	1.2E-1	*OPRD1*	3.8E-1	*DDC*	6.4E-1
DRD2	1.3E-1	*SLC6A4*	3.8E-1	*NPY*	6.5E-1
CAT	1.4E-1	*GABRA6*	3.9E-1	*POMC*	6.6E-1
ADH7	1.5E-1	*CYP2E1*	4.0E-1	*ALDH2*	6.6E-1
ARRB2	1.7E-1	*HTR2B*	4.1E-1	*JUN*	6.8E-1
PPP1R1B	1.7E-1	*CCKBR*	4.1E-1	*ADH1A*	7.0E-1
ADH1B	1.8E-1	*SLC6A7*	4.2E-1	*GABRB1*	7.2E-1
FOSL2	1.9E-1	*GRIN2A*	4.4E-1	*HCRT*	7.6E-1
GABRG2	2.0E-1	*CLOCK*	4.4E-1	*MAPK1*	7.8E-1
GRIN2C	2.0E-1	*COMT*	4.4E-1	*DRD3*	7.9E-1
PDYN	2.0E-1	*GAL*	4.5E-1	*CCK*	8.0E-1
GPHN	2.0E-1	*GABRB2*	4.5E-1	*ADRA2B*	8.1E-1
SLC6A3	2.1E-1	*GABRD*	4.5E-1	*FEV*	8.2E-1
NTSR2	2.2E-1	*GABRA2*	4.5E-1	*NR3C1*	8.3E-1
HTR2A	2.2E-1	*HTR3A*	4.5E-1	*FAAH*	8.4E-1
NPY2R	2.3E-1	*ADH6*	4.5E-1	*HTR3B*	8.4E-1
GRIN2B	2.4E-1	*SLC18A2*	4.7E-1	*HTR1A*	8.5E-1
SLC6A13	2.4E-1	*DBH*	4.7E-1	*NTRK2*	8.7E-1
AVPR1B	2.6E-1	*MAPK3*	4.7E-1	*CDK5R1*	8.7E-1
ADCY7	2.6E-1	*CRHR2*	4.9E-1	*OPRK1*	8.8E-1
CHRM5	2.7E-1	*CHRNB2*	5.0E-1	*DRD4*	9.0E-1
ADRA2C	2.8E-1	*HTR1B*	5.0E-1	*ADRB2*	9.2E-1
GAD1	2.8E-1	*GRIN1*	5.0E-1	*TPH1*	9.3E-1
NTSR1	3.0E-1	*CREB1*	5.0E-1	*LEP*	9.3E-1
CSNK1E	3.0E-1	*SLC6A2*	5.1E-1	*ADH5*	9.5E-1
DRD1	3.1E-1	*SLC6A11*	5.2E-1	*GSK3B*	9.6E-1
GLRB	3.3E-1	*TPH2*	5.3E-1	*SLC29A1*	9.9E-1
OXT	3.3E-1	*GABRB3*	5.3E-1	*FOS*	1.0E+0
CRHR1	3.3E-1	*NPY1R*	5.4E-1	*CHRM4*	1.0E+0

Supplemental Table S3. Candidate gene pathways tested in CADD. Candidate genes and pathway assignments are based on Hodgkinson et al. (2008). *Intervals* indicates the total number of genomic regions included in the pathway definition, and *Overlap* indicates the number of those pathway intervals that overlap low *p*-value genomic regions in the CADD GWAS. Empirical *p* takes into account genomic coverage of the pathway and tested SNPs, and the Corrected *p*-value has been adjusted for multiple testing.

Pathway	*Intervals*	*Overlap*	*Empirical p*	*Corrected p*	*Genes*
Omnibus	123	5	9.7E-1	9.8E-1	see below
Drug metabolism	11	0	1.0E+0	1.0E+0	
Dopamine	10	0	1.0E+0	1.0E+0	
Serotonin	9	0	1.0E+0	1.0E+0	
GABA	16	1	1.0E+0	1.0E+0	*GABRG3*
Opioid	8	0	1.0E+0	1.0E+0	
Glycine	3	0	1.0E+0	1.0E+0	
NMDA	6	0	1.0E+0	1.0E+0	
Cannabinoid	2	1	1.0E+0	1.0E+0	*CNR1*
Signal transduction	22	2	5.0E-1	8.4E-1	*NGF, PRKCE*
Cholinergic	7	0	1.0E+0	1.0E+0	
Stress	9	0	1.0E+0	1.0E+0	
Adrenergic	8	1	1.0E+0	1.0E+0	*ADRA1A*
Other	12	0	1.0E+0	1.0E+0	

Supplemental Table S4. Tests of all pathways meeting Empirical $p<0.05$ in CADD that were subsequently tested in the MCTFR and SAGE samples. Pathways are identified by their unique Gene Ontology (GO) path ID number. Intervals indicates the total number of genomic regions included in the pathway definition, and Overlap indicates the number of those pathway intervals that overlap low *p*-value genomic regions in each sample's GWAS. Empirical *p* takes into account genomic coverage of the pathway and tested SNPs, and the Corrected *p*-value has been adjusted for multiple testing. Pathways are sorted on increasing Empirical *p*-value in CADD.

			CADD			*MCTFR*			*SAGE*		
GO path ID	*Pathway description*	*Intervals*	*Overlap*	*Emp. P*	*Corr. P*	*Overlap*	*Emp. P*	*Corr. P*	*Overlap*	*Emp. P*	*Corr. P*
GO:0031398	positive regulation of protein ubiquitination	30	7	1.5E-4	1.4E-1	0	1.0E+0	1.0E+0	4	2.6E-1	1.0E+0
GO:0034080	CenH3-containing nucleosome assembly at centromere	22	5	6.9E-4	5.4E-1	0	1.0E+0	1.0E+0	2	5.2E-1	1.0E+0
GO:0019843	rRNA binding	21	4	2.0E-3	8.8E-1	0	1.0E+0	1.0E+0	2	2.2E-1	1.0E+0
GO:0004614	phosphoglucomutase activity	5	3	2.2E-3	9.1E-1	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0006334	nucleosome assembly	91	7	2.8E-3	9.5E-1	0	1.0E+0	1.0E+0	6	1.6E-1	1.0E+0
GO:0004445	inositol-polyphosphate 5-phosphatase activity	7	3	3.4E-3	9.7E-1	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0002039	p53 binding	26	5	4.3E-3	9.9E-1	0	1.0E+0	1.0E+0	3	4.8E-1	1.0E+0
GO:0050852	T cell receptor signaling pathway	74	10	5.2E-3	9.9E-1	2	2.8E-1	9.4E-1	7	6.3E-1	1.0E+0
GO:0033344	cholesterol efflux	21	4	5.6E-3	1.0E+0	0	1.0E+0	1.0E+0	2	4.9E-1	1.0E+0
GO:0006646	phosphatidylethanolamine biosynthetic process	5	2	6.0E-3	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0010811	positive regulation of cell-substrate adhesion	13	4	6.1E-3	1.0E+0	0	1.0E+0	1.0E+0	2	4.8E-1	1.0E+0
GO:0042632	cholesterol homeostasis	45	6	7.7E-3	1.0E+0	0	1.0E+0	1.0E+0	6	1.7E-1	1.0E+0
GO:0050900	leukocyte migration	104	11	8.7E-3	1.0E+0	2	4.4E-1	9.8E-1	9	7.0E-1	1.0E+0
GO:0051001	negative regulation of nitric-oxide synthase	7	2	8.9E-3	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0

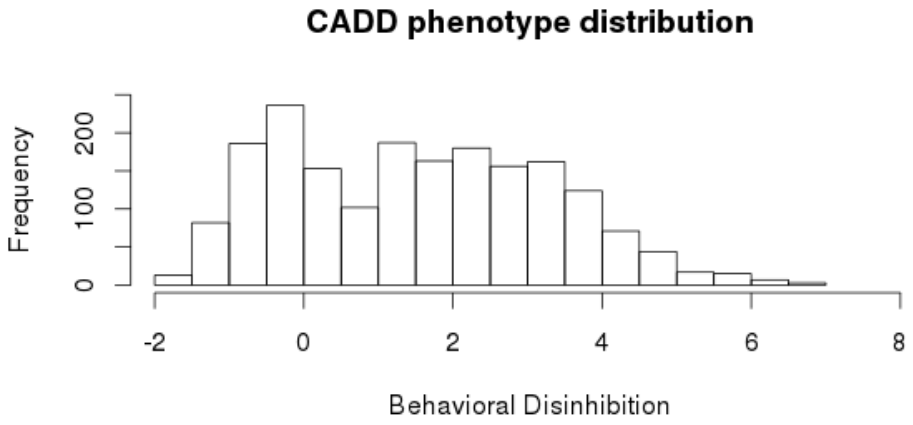
	activity										
GO:0016328	lateral plasma membrane	17	4	1.1E-2	1.0E+0	1	1.0E+0	1.0E+0	2	6.2E-1	1.0E+0
GO:0030897	HOPS complex	11	3	1.1E-2	1.0E+0	1	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0005978	glycogen biosynthetic process	12	3	1.1E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0017127	cholesterol transporter activity	12	3	1.2E-2	1.0E+0	0	1.0E+0	1.0E+0	3	9.4E-2	9.7E-1
GO:0000775	chromosome, centromeric region	54	6	1.2E-2	1.0E+0	0	1.0E+0	1.0E+0	4	6.4E-1	1.0E+0
GO:0032400	melanosome localization	5	2	1.2E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0043088	regulation of Cdc42 GTPase activity	7	3	1.2E-2	1.0E+0	0	1.0E+0	1.0E+0	2	2.3E-1	1.0E+0
GO:0019003	GDP binding	25	4	1.3E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0042393	histone binding	46	6	1.3E-2	1.0E+0	1	1.0E+0	1.0E+0	4	5.1E-1	1.0E+0
GO:0045120	pronucleus	7	2	1.4E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0042605	peptide antigen binding	12	3	1.7E-2	1.0E+0	1	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0004866	endopeptidase inhibitor activity	30	4	1.7E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0031267	small GTPase binding	8	3	1.8E-2	1.0E+0	0	1.0E+0	1.0E+0	3	7.9E-2	9.4E-1
GO:0000245	spliceosome assembly	19	3	1.8E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0000796	condensin complex	5	2	1.8E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0006301	postreplication repair	7	2	1.9E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0001931	uropod	6	2	1.9E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0006359	regulation of transcription from RNA polymerase III promoter	8	2	2.1E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0009311	oligosaccharide metabolic process	8	2	2.4E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0050999	regulation of nitric-oxide synthase activity	12	3	2.5E-2	1.0E+0	0	1.0E+0	1.0E+0	2	3.7E-1	1.0E+0
GO:0031013	troponin I binding	5	2	2.5E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0003711	transcription elongation regulator activity	9	2	2.5E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0008526	phosphatidylinositol transporter activity	5	2	2.6E-2	1.0E+0	1	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0007130	synaptonemal complex	8	2	2.8E-2	1.0E+0	1	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0

	assembly										
GO:0032436	positive regulation of proteasomal ubiquitin-dependent protein catabolic process	23	3	2.8E-2	1.0E+0	0	1.0E+0	1.0E+0	2	5.0E-1	1.0E+0
GO:0043691	reverse cholesterol transport	17	3	3.0E-2	1.0E+0	0	1.0E+0	1.0E+0	3	2.0E-1	1.0E+0
GO:0004835	tubulin-tyrosine ligase activity	14	3	3.0E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0005200	structural constituent of cytoskeleton	72	9	3.1E-2	1.0E+0	1	1.0E+0	1.0E+0	8	4.4E-1	1.0E+0
GO:0008360	regulation of cell shape	58	9	3.1E-2	1.0E+0	3	1.3E-1	8.7E-1	7	7.0E-1	1.0E+0
GO:0035024	negative regulation of Rho protein signal transduction	6	3	3.1E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0051403	stress-activated MAPK cascade	46	6	3.1E-2	1.0E+0	2	1.6E-1	9.0E-1	5	3.8E-1	1.0E+0
GO:0015804	neutral amino acid transport	7	2	3.1E-2	1.0E+0	1	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0042987	amyloid precursor protein catabolic process	8	2	3.2E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0018149	peptide cross-linking	20	4	3.2E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0000795	synaptonemal complex	11	2	3.4E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0008016	regulation of heart contraction	32	5	3.5E-2	1.0E+0	1	1.0E+0	1.0E+0	3	6.5E-1	1.0E+0
GO:0000405	bubble DNA binding	6	2	3.6E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0007520	myoblast fusion	11	4	3.8E-2	1.0E+0	0	1.0E+0	1.0E+0	3	4.3E-1	1.0E+0
GO:0007601	visual perception	194	18	0.0383	1	8	0.0119	0.2289	22	0.224	0.9998
GO:0006805	xenobiotic metabolic process	126	9	0.0383	1	3	0.2015	0.9302	15	0.2344	0.9998
GO:0019534	toxin transporter activity	5	2	0.0384	1	1	1	1	0	1	1
GO:0016045	detection of bacterium	9	2	0.0389	1	0	1	1	0	1	1
GO:0033700	phospholipid efflux	10	2	0.0389	1	0	1	1	2	0.1344	0.9918
GO:0006656	phosphatidylcholine biosynthetic process	12	2	0.0391	1	0	1	1	3	0.0257	0.5709
GO:0055091	phospholipid homeostasis	6	2	3.9E-2	1.0E+0	0	1.0E+0	1.0E+0	2	1.8E-1	1.0E+0
GO:0015631	tubulin binding	21	5	4.0E-2	1.0E+0	0	1.0E+0	1.0E+0	2	9.2E-1	1.0E+0
GO:0051098	regulation of binding	5	2	4.1E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0

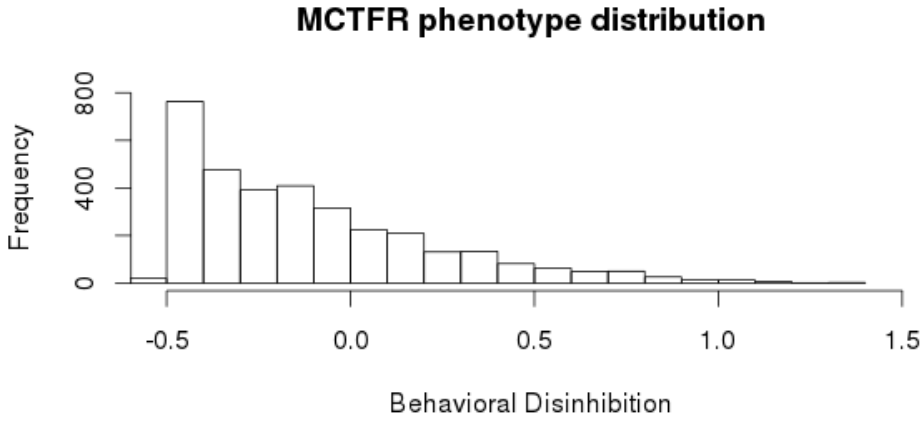
GO:0019885	antigen processing and presentation of endogenous peptide antigen via MHC class I	6	2	4.1E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0050700	CARD domain binding	7	2	4.3E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0042834	peptidoglycan binding	8	2	4.3E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0015175	neutral amino acid transmembrane transporter activity	9	2	4.5E-2	1.0E+0	1	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0030968	endoplasmic reticulum unfolded protein response	20	3	4.6E-2	1.0E+0	0	1.0E+0	1.0E+0	0	1.0E+0	1.0E+0
GO:0000793	condensed chromosome	20	3	4.6E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0006333	chromatin assembly or disassembly	37	4	4.6E-2	1.0E+0	1	1.0E+0	1.0E+0	3	6.5E-1	1.0E+0
GO:0007586	digestion	51	5	4.8E-2	1.0E+0	1	1.0E+0	1.0E+0	2	9.1E-1	1.0E+0
GO:0016565	general transcriptional repressor activity	9	2	4.8E-2	1.0E+0	0	1.0E+0	1.0E+0	2	1.1E-1	9.8E-1
GO:0008536	Ran GTPase binding	10	2	4.9E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0
GO:0008009	chemokine activity	46	3	4.9E-2	1.0E+0	0	1.0E+0	1.0E+0	1	1.0E+0	1.0E+0

Supplemental Figure S1. Phenotypic distributions in each study.

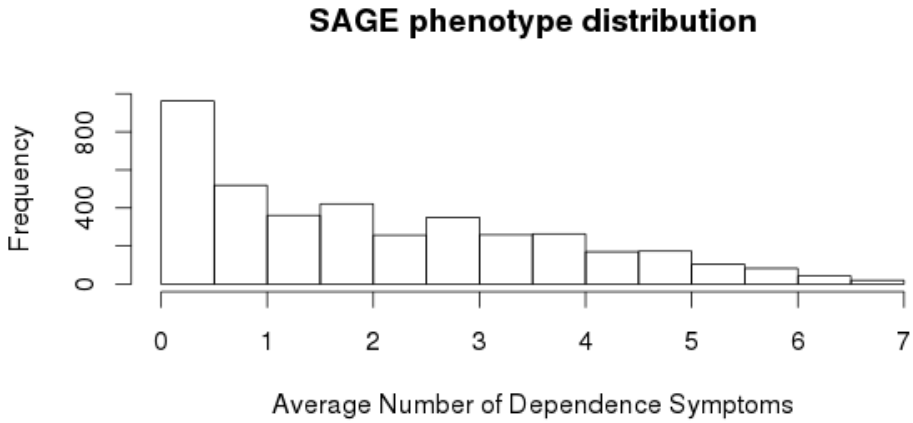
a.



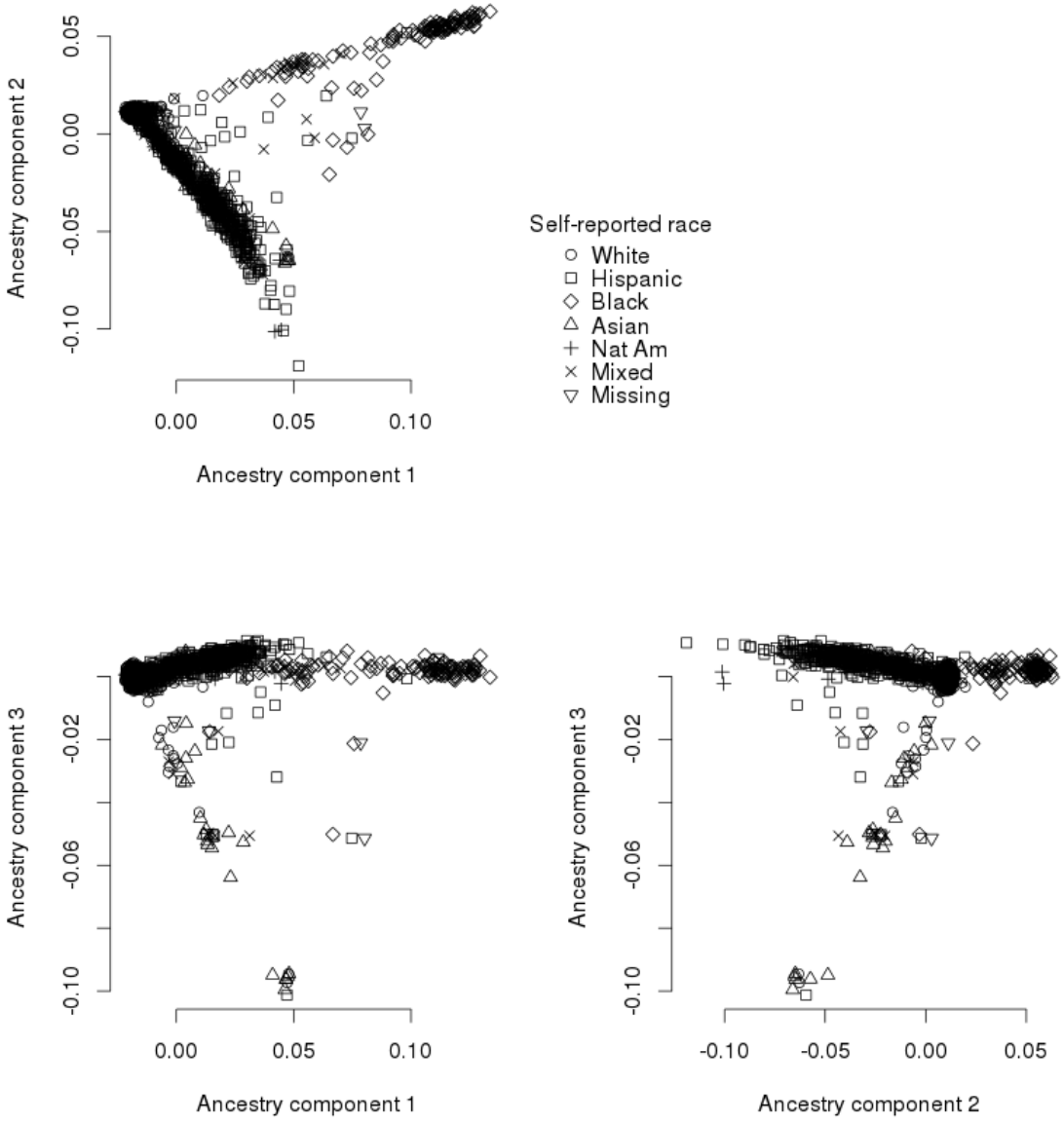
b.



c.

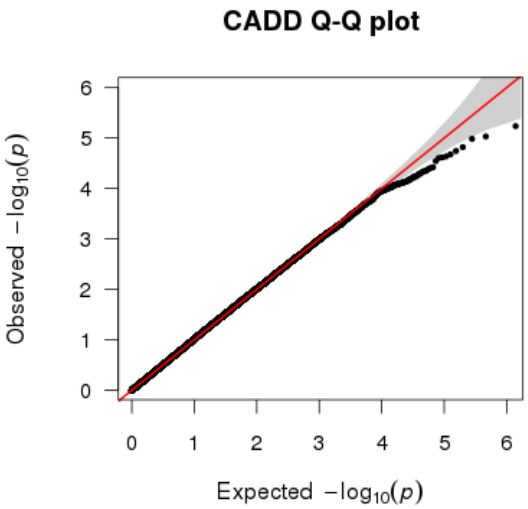


Supplemental Figure S2. Illustrations of the first three ancestry dimensions in CADD, with individuals labeled based on self-reported ancestry.

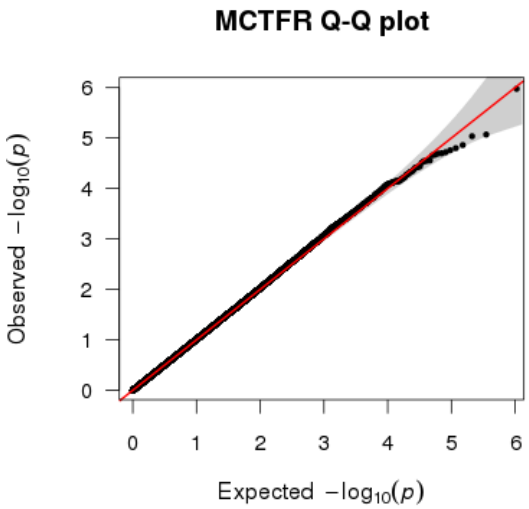


Supplemental Figure S3. Q-Q plots from each study's GWAS. The x-axis is the expected distribution of -$\log_{10}(p)$ while the y-axis is the observed distribution of -$\log_{10}(p)$ from the GWAS. The red horizontal line indicates p-values occurring exactly as expected by chance, and the grey shaded area indicates the 95% confidence interval for deviation from the expected distribution.

a.



b.



c.