JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Report: Regular Manuscript

Genetic Architecture and Molecular Neuropathology of Human Cocaine Addiction

https://doi.org/10.1523/JNEUROSCI.2879-19.2020

Cite as: J. Neurosci 2020; 10.1523/JNEUROSCI.2879-19.2020

Received: 3 December 2019
Revised: 4 May 2020
Accepted: 10 May 2020

Genetic Architecture and Molecular Neuropathology of Human Cocaine Addiction

Spencer B. Huggett[1,2] and Michael C. Stallings[1,2]

1. *Department of Psychology and Neuroscience, University of Colorado, Boulder, CO, USA*
2. *Institute for Behavioral Genetics, University of Colorado Boulder, CO, USA*

Corresponding Author: Spencer B. Huggett

Emails: sphu3837@colorado.edu & Michael.Stallings@colorado.edu

Abstract: # of words - 250

Introduction: # of words - 650

Discussion: # of words – 1,446

# of pages (text) - 24

# of tables: 2; # of figures: 7

Acknowledgements
We would like to acknowledge the funding that supported this work from the National Institute of Health (T32 DA 17637-14) and National Institute on Drug Abuse (DA035804). We appreciate Dr. Joel Gelernter and Dr. Henry Kranzler for giving us access to their genome-wide association summary statistics on cocaine dependence and the Gene Expresion Omnibus (GEO) for providing open access to the single-cell RNA-sequencing data used in this study.

Conflict of Interest[1]

Keywords: Cocaine dependence, Cocaine use disorder, Genome-wide association study (GWAS), RNA sequencing, Multi-Ancestry, GWAS Follow-up

[1]The authors declare no competing financial interests[1]

Genetic Architecture and Molecular Neuropathology of Human Cocaine Addiction

Spencer B. Huggett[1,2] and Michael C. Stallings[1,2]

[1]Department of Psychology and Neuroscience, University of Colorado
[2]Institute for Behavioral Genetics, University of Colorado
Supported in part by: NIDA P60 DA11015

Abstract

We integrated genomic and bioinformatic analyses, utilizing data from the largest genome-wide association study (GWAS) of cocaine dependence (CD; n=6,546; 82.37% with CD; 57.39% male) and the largest post-mortem gene-expression sample of individuals with cocaine use disorder (CUD; n=36; 51.35% with CUD; 100% male). Our genome-wide analyses identified one novel gene (*NDUFB9*) associated with the genetic predisposition to CD in African-Americans. The genetic architecture of CD was similar across ancestries. Individual genes associated with CD demonstrated modest overlap across European and African-Americans, but the genetic liability to CD converged on many similar tissue types (brain, heart, blood, liver) across ancestries. In a separate sample, we investigated the neuronal gene expression associated with CUD by using RNA sequencing of dorsal-lateral pre-frontal cortex neurons. We identified 133 genes differentially expressed between CUD cases and cocaine-free controls, including previously implicated candidates for cocaine use/addiction (*FOSB, ARC, KCNJ9/GIRK3, NR4A2, JUNB* and *MECP2*). Differential expression analyses significantly correlated across European and African-Americans. While genes significantly associated with CD via genome-wide methods were not differentially expressed, two of these genes (*NDUFB9* and *C1qL2*) were part of a robust gene co-expression network associated with CUD involved in neurotransmission (GABA, acetylcholine, serotonin and dopamine) and drug addiction. We then used a "guilt-by-association" approach to unravel the biological relevance of *NDUFB9* and *C1qL2* in the context of CD. In sum, our study furthers the understanding of the genetic architecture and molecular neuropathology of human cocaine addiction and provides a framework for translating biological meaning to otherwise obscure genome-wide associations.

Significance Statement:

Our study: **1)** further clarifies the genetic and neurobiological contributions to cocaine addiction, **2)** provides a rapid approach for generating testable hypotheses for specific candidates identified by genome-wide research and **3)** investigates the cross-ancestral biological contributions to cocaine use disorder/dependence for individuals of European- and African-American ancestries.

Introduction

Neuroscience research has facilitated the identification of genes studied in hypothesis-driven human genetic research, often called "candidate gene studies." The candidate gene literature proposes numerous genetic associations with cocaine use/addiction from genomic variants within genes from neurotransmitter systems. However, some experts question the validity of candidate gene research due to a lack of reproducibility (Colhoun et al. 2003; Munafo, 2009) and encourage the use of hypothesis-free genome-wide methods.

Genome-wide association studies (GWASs) have identified thousands of genetic variants associated with human traits. However, linking molecular mechanisms to GWAS findings is challenging. Significant GWAS results do not generally conform to *a priori* candidate genes and often tag non-protein coding genomic regions (Maurano et al. 2012). Therefore individual gene variants from GWASs are rarely interpreted with concrete mechanisms. Experimental laboratory studies have unraveled mechanisms for a few GWAS findings (Claussnitzer et al. 2015; Sekar et al. 2016), but these studies are expensive and time intensive, so it is not feasible to apply this line of research for all GWAS findings. Systematic approaches are needed to prioritize individual genes from GWASs for follow-up investigation in specific tissues or cell types. Another important caveat of GWAS research is that most findings are based on individuals of European ancestry/ethnicity (Martin et al. 2019), highlighting a priority to investigate the genetic basis of traits among non-Europeans.

GWASs have discovered four significant genes contributing to predisposition to cocaine dependence (CD; DSM-IV): *FAM53B, KCTD20*, *STK38* and *C1qL2* (Gelernter et al. 2014; Huggett & Stallings, 2019). The relevance of these genes with CD is not fully understood. Follow-up investigation in mice revealed that *Fam53b* might influence cocaine self-administration via midbrain co-expression with *Cyfip2* (Dickson et al. 2016), a gene, which influences cocaine-induced sensitization (Kumar et al. 2013). Similarly, our previous work found that *KCTD20* was associated with human cocaine abuse/dependence through a hippocampal gene co-expression network implicated in synaptic plasticity (Huggett & Stallings, 2019). This work provides a "guilt-by-association" approach to infer the role of newly associated disease genes and helps contextualize and interpret otherwise ambiguous genetic associations. Given the surplus of publically available bioinformatic data, systems-based computational follow-up may be a fruitful line of inquiry that could help translate biological meaning to obscure genetic associations.

Despite the rising rates of cocaine and drug-related overdoses in the United States (US; NIDA, 2018) post-mortem brain data on substance use disorders remain limited. The largest cocaine-related human brain sample used RNA-sequencing (RNA-seq) on dorsolateral PFC (dlPFC) neurons from individuals of mixed ancestries (Ribeiro et al. 2017). The PFC is a critical region for the neuropathology of cocaine addiction and plays a role in decision-making, salience attribution and promotes inhibitory control over drug addiction (Goldstein et al. 2011). Rodent models suggest that PFC glutamate neurons provide "top-down" control of reward circuitry and increase motivation to seek/use cocaine (Kalivas et al. 2005), but little

is known regarding the neuro-adaptations underlying PFC dysfunction in human cocaine addicts. Ribeiro *et al*. identified associations of various immediate early genes (*FOS, JUN* and *JUNB*) with dlPFC neuro-adaptations of cocaine use disorder (CUD; DSM-V) and found one gene co-expression network associated with CUD that was enriched for neuroplasticity processes and GWAS associations for Body-Mass index and obesity. Notably, while, genome-wide research has begun to disentangle the genetic architecture of human traits across ancestries (Peterson et al. 2019), we are aware of no transcriptome-wide studies characterizing potential similarities/differences across ancestries/ethnicities. Future research is warranted to clarify the links between the genetic risk for substance abuse and the neurobiological characteristics of the addicted brain, while also investigating how gene expression generalizes across ethnicities.

This study aimed to unravel the genetic architecture and molecular neuropathology of human cocaine addiction. Integrating genomic and bioinformatic methods, we identified specific genes and tissues associated with the predisposition to CD and characterized PFC neuro-adaptations associated with CUD. We translated findings across ancestries and methods and sought to make human genetic findings more relevant for neuroscientists.

Materials and Methods

Genome-wide Analyses

Sample

We used case-control GWAS summary statistics from Gelernter et al. 2014, which was based on data from 3,370 African-Americans (44.18% female; M_{age} =

41.71) and 3,176 European-Americans (40.96% female; M_{age} = 37.35). Participants were a part of the Study of Addiction: Genetics and Environment (SAGE) or recruited via clinical settings in the northeastern US. Genome-wide analyses were performed separately by ancestry to account for population stratification. GWAS summary statistics corrected for relatedness via generalized estimating equations and adjusted for 3 ancestral principal components, age and sex, but not co-occurring substance abuse nor other psychiatric co-morbidities. All participants reported trying cocaine and 90.39% of African-Americans and 73.96% of European-Americans had a lifetime diagnosis of CD (3+ Symptoms of DSM-IV criteria). In a portion of this sample, measurements of CD yielded high internal reliability k > 0.80 (Pierucci-Lagha et al. 2005), indicating reliable trait measurement. Stringent quality control was applied to the genotypic data of all subjects and imputation was performed using the 1000 Genomes reference panel.

Experimental Design and Analysis:

Gene-Based Associations

To detect specific protein-coding genes underlying the predisposition of CD, we conducted Multi-marker Analysis of GenoMic Annotation (de Leeuw et al. 2015; MAGMA, *v*1.06) gene-based association tests by submitting summary statistics to the Functional Mapping and Annotation (FUMA, *v*1.1.2) GWAS pipeline (Watanabe et al. 2017). Contrary to GWAS, which performs millions of regressions for all common gene variants across the genome, gene-based associations perform one regression per protein-coding gene and therefore reduce the multiple testing burdens of GWAS and offer more interpretable results. Most protein-coding genes

have a multitude of gene variants. MAGMA gene-based tests use a principal components analysis to reduce the numerous variants for a certain gene into a single signal, which is then associated with the trait (de Leeuw et al. 2015). Our gene-based analyses included all single nucleotide polymorphisms within protein-coding regions of the genome (Ensembl v85) that had a minor allele frequency > 1%. In total, our gene-based tests included 18,122 genes for the African-American sample and 18,220 genes in European-American sample (18,903 shared genes). We compared the results of our previously published gene-based test of CD in European-Americans (Huggett & Stallings, 2019) to the African-American sample and used a Bonferonni correction for multiple testing to determine genome-wide significance ($p < 2.7e-6$). Note that this standard Bonferronni *p*-value correction (FUMA default) demarks a less significant threshold than the original GWAS (Gelernter et al. 2014; $p < 5.0e-8$) due to the reduction of tests performed (~18,000 versus ~9 million).

To interrogate specific alleles underlying the genetic predisposition to CD, we investigated specific single nucleotide polymorphisms (SNPs) driving genome-wide significant associations. First, we reported the lead SNP from each genomic region, the total SNPs within each gene as well as the number of parameters for each gene, which reflects independent linkage disequilibrium blocks within genes. 'Causal' SNPs are more likely to confer a biological consequence in protein or transcript function. Leveraging DNA sequencing data from 71,702 individuals, we queried the Genome Aggregation Database (v3; Karczewski et al. 2019; https://gnomad.broadinstitute.org/) for missense mutations, or SNPs that code for

an amino acid substitution, among genome-wide significant gene-based test results. To determine whether a missense mutation was significantly associated with CD, we used a Bonferonni correction for all missense variants within each gene. We also estimated the relationship between particular missense mutations and a gene's lead SNP using LDlink (Machiela &Chanock, 2015; https://ldlink.nci.nih.gov/), which computes linkage disequilibrium between loci by ancestry.

Our study then refined the focus of gene-based associations with CD from a genome-wide perspective to a candidate systems approach, selecting genes from typically studied neurotransmitter systems. In total, these analyses included 130 genes from GABA, glutamate, acetylcholine, endocannabanoid, dopamine, epinephrine/norepinephrine and serotonin systems encompassing synthesis, vesicular transport, receptors, degradation and reuptake genes. Since these classical genes rarely surpass conservative genome-wide significance thresholds, we assessed whether these hypothesis-driven neurotransmitter genes surpassed a nominally significant *p*-value threshold ($p < 0.05$) as typically employed in the candidate gene literature. Collapsing across ancestries, we tested whether these candidate neurotransmitter genes were enriched for being nominally associated with CD using a Fisher's exact test.

Tissue Enrichment

To identify tissues underlying the genetic pathophysiology of CD, we performed tissue enrichment analyses. These analyses assess *where* genes underlying the predisposition of a trait might be exerting a functional role. Tissue enrichment analyses identify which tissues a list of input genes demonstrate

differential expression (up or down-regulated in a tissue compared to all other tissues). We assessed tissue enrichment in 53 tissues from hundreds of healthy human samples (GTEx Consortium, 2013) and performed analyses separately by ethnicity - including genes nominally associated with CD (unadjusted $p < 0.05$; 901 genes in African-Americans and 1008 genes in European-Americans). Tissue enrichment analyses used competitive hyper-geometric tests to compare a specific tissue type versus all other tissues and incorporated a Bonferoni multiple testing correction to ascertain significantly enriched tissues ($p < 0.05/53$).

Neuron Specific RNA-seq Analyses

Sample

Next, we performed neuron-specific RNA-sequencing (RNA-seq) analyses using publically available data from the largest post-mortem human brain study on cocaine use (GEO: GSE99349). Cocaine users (n = 19; 100% male; 6 African-Americans, 6 European-Americans and 7 Hispanics or Latinos) died from the toxic effects of chronic cocaine abuse and met criteria for cocaine use disorder (CUD; DSM-V criteria). Age and race matched cocaine-free controls (n = 17; 7 African-Americans, 4 European-Americans and 6 Hispanics or Latinos) were selected from homicides, accidental or cardiac-related deaths and had negative urine screens for common drugs before death. Cases and controls did not significantly differ on post-mortem index (PMI), RNA integrity (RIN), age, or brain pH level, all $|t| > 1.69$, all $p > 0.100$.

Data Preparation

For more details on the sample, tissue preparation, RNA extraction, library construction and RNA-seq protocol see Ribiero et al. (2017). Briefly, dlPFC tissue was extracted from the middle frontal gyrus at the lateral portion of Broadmann's area 46. Fluorescent activated cell sorting dissociated dlPFC cell types and neuronal nuclei were isolated/extracted via the mouse anti-NeuN antibody. RNA isolation was conducted via the Zymo Directzol RNA miniprep kit (Zymo Research; R2050). Indexed libraries were constructed using 10 ng of nuclear RNA from each sample with the Contech SMARTer Stranded Total RNA-seq library preparation kit (Clontech/Takara; 634839). Paired end (2x125) RNA-seq was performed using the Illumina Hiseq-2000 (Liu et al. 2011) and resulted in an average of 50,925,315 read pairs per sample.

We pre-processed the RNA-seq data from Ribiero et al. via Trimmomatic v 0.36 to eliminate short and low quality reads (Phred score < 20 or < 100 bases) as well as Illumina adapters, which resulted in an average of 30,486,006 read pairs per sample. We then aligned the RNA-seq data to the hg19 reference genome via the Spliced Transcripts Alignment to a Reference (STAR; Dobin et al. 2013). On average we had 26,476,583 (*SD* = 6,173,119) uniquely mapped read pairs per sample, with a mean alignment rate of 86.84% (*SD* = 5.86%) and observed no significant differences in read alignment between cases and controls, $t = 0.668$, $p = 0.509$. Our study used HTseq (Anders et al. 2015) to transfer mapped reads into discrete genes/transcripts.

Our re-analysis of the Ribiero et al. (2017) data differed in two ways. First, we defined differentially expressed genes with an adjusted *p*-value threshold of

Benjamini-Hochberg False Discovery Rate (BH-FDR) < 0.05. Second, we normalized RNA-seq data with SCnorm – a method that utilizes quantile regression and seems to properly handle data derived from single cell types (Bacher et al. 2017). RNA-seq approaches from a single cell (type) differ from regular RNA-seq due to the presence of technical noise (i.e., zero-inflated read counts of genes not expressed in sequenced cells) and may require sensitive statistical care. To test whether SCnorm increased power, we assessed the number of differentially expressed genes identified from this technique compared to a standard normalization method (DESeq2 scale factors). Without covariates, we found just 6 differentially expressed genes/transcripts ($p_{adj} < 0.05$) using the standard scale factor approach, but identified 250 differentially expressed genes/transcripts ($p_{adj} < 0.05$) via the SCnorm technique. Additionally, we found appreciable evidence for zero-inflated read counts and discovered that SCnorm successfully accommodated for this noise (data available upon request), perhaps stemming from non-neuronal genes/transcripts. Accordingly, the lowest decile of normalized read counts were enriched for cortical astrocytes ($p_{adj} = 6.92e\text{-}4$) and oligodendrocytes ($p_{adj} = 0.002$), but not cortical neuronal cell types (all $p_{adj} > 0.999$) as observed from a cell specific expression analysis (Doughtrey et al. 2010). Thus, we normalized the RNA-seq data with SCnorm (for our differential expression analyses) as it appeared to properly account for technical artifacts and afforded increased statistical power.

Experimental Design and Analysis:

Differential Expression

We utilized DESeq2 (Love et al. 2014) to assess differentially expressed genes/transcripts and to investigate the association of differential expression analyses across ancestries. We used the full sample to identify differentially expressed genes/transcripts (49,496 total genes/transcripts), which controlled for RIN, PMI, age, race (European-American = -1; Hispanic = 0; African-American = 1), blood alcohol content, smoking status (smokers = 1; non-smokers = 0) and hidden batch effects (2 surrogate variables via the svaseq package; Leek, 2014).

To complement our genome-wide analyses, we investigated the association between ancestry specific differential expression results from African-American (n = 13) and European-American (n = 10) subsamples. Due to low sample size for ancestry specific differential expression analyses, we did not control for all possible confounds, but adjusted for two common and salient covariates (PMI and age). Log fold change estimates from differential expression analyses are estimated with noise – especially among lowly expressed genes/transcripts. To accommodate for this error/noise and enable transcriptome-wide investigation (e.g., low and high expressed genes/transcripts), our cross-ancestry RNA-seq analysis focused on test statistics from differential expression analyses (DESeq2 Wald statistics), which account for log fold change effect size and standard error for individual genes/transcripts. Additionally, we selected the genes/transcripts with a differential expression Wald-statistic > |2| in either European- or African-American specific analyses (705 genes/transcripts) and investigated the cross-ancestry correlation of cocaine-related gene/transcripts.

Gene Co-expression Networks

Next, our study utilized a systems-genetics approach to model clusters of genes derived from correlated RNA expression (gene co-expression networks/gene networks). The reader should note that these analyses do not determine gene co-expression networks *a priori*, but rather create gene networks from the observed RNA-seq data. Specifically, we conducted a signed weighted gene co-expression network analysis (WGCNA; Langfelder & Hovarth, 2008), using the same input parameters as our previous work (Huggett & Stallings, 2019). Briefly, we filtered genes/transcripts based on expression level, such that we only included genes/transcripts with an average baseline expression > 1 read count per sample, which resulted in a total of 15,178 genes/transcripts for WGCNA modeling. Our WGCNA approach computed Pearson Product-Moment correlations of normalized RNA expression (log_2-counts per million) of all WGCNA genes/transcripts with themselves and weighted these correlations by raising them to the (default) power of 12, which satisfied WGCNA distribution assumptions (scale free topology = 0.84). Then, using a dynamic tree-cutting algorithm we split clusters of correlated/co-expressed genes into defined WGCNA gene co-expression networks (minimum module size = 50).

To validate our WGCNA gene networks, we utilized a Z-summary module preservation statistic (Langfelder et al. 2011). Z-summary statistics above 10 indicate gene networks are highly robust and reproducible and Z-summary statistics greater than 2 suggest WGCNA gene networks are weak to moderately reproducible. Our validation approach was based on previous work (Vanderlinden et al. 2013) that incorporates within sample and out of sample gene network

validation technique. Our within sample gene network validation analysis is indicative of WGCNA network *stability* and compared the WGCNA networks from the current study to 100 bootstrapped samples from the same dataset (human dlPFC neurons; n = 37). Then to assess whether gene networks were *robust* in a separate sample, we tested if our constructed WGCNA gene networks were reproducible in an independent sample using RNA-seq data of hippocampal tissue from human cocaine users/addicts and controls (Zhou et al. 2011).

Similar to previous research (Ponomarev et al. 2012), we used an effect-size based approach leveraging test statistics from our full sample differential expression analysis (DESeq2 Wald statistics) to associate gene co-expression networks with CUD. That is, we calculated the average absolute value of Wald statistics for all genes/transcripts within each defined gene co-expression network. The directions of associations were determined by assessing whether mean effect sizes for gene networks were positive or negative. We ascertained significant gene networks via 100,000 permutations. That is, our permutations re-sampled the absolute values of Wald statistics from all WGCNA genes to approximate a null distribution. We then derived *p*-values by determining the probability that a gene co-expression network had an average absolute Wald statistic in relation to what is expected under the null. We defined a significant association of a gene network with CUD, if it survived a Bonferonni correction ($p < 0.05/\#$ of WGCNA gene networks) *and* demonstrated enrichment for differentially expressed genes (FDR < 0.05).

Functional Annotation

We functionally annotated our RNA-seq results via the Database for Annotation, Visualization and Integrated Discovery (DAVID v6.8; Huang et al. 2009) and queried for enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, biological processes (BPs), and/or molecular functions (MFs). To control for false positives, we required significant enrichment to survive correction for multiple testing (FDR < 0.05) and adjusted for the 'background distribution' by incorporating a list of genes that were included for each analysis. We uploaded our results to GeneWeaver (https://www.geneweaver.org/; Baker et al. 2012), which can be found by searching the reported id numbers (GS#).

Results

Genome-wide Analyses

Gene-Based Associations

To identify specific genes underlying the predisposition to CD, we conducted gene-based association tests. **Figure 1** shows the Miami plot visualizing the results of our gene-based associations with cocaine dependence (CD) for African- and European-Americans. Extending our previous gene-based associations with CD among European-Americans (Huggett & Stallings, 2019), we identified one novel genome-wide significant association with CD in African-Americans (p = 8.27e-07), the NADH: ubiquinone oxioreductase subunit B9 gene (*NDUFB9*), but not in European-Americans (p = 0.910). The *NDUFB9* gene is part of the inner mitochondrial membrane and plays a role in oxidative phosphorylation, but the relevance of this gene in the context of cocaine addiction is not known, warranting further investigation. To investigate specific loci underlying our genome-wide

significant associations, we reported each region's most significant SNP (lead SNP) and examined missense mutations for each gene associated with CD. After correction for multiple testing, we found significant associations between a missense mutation in the *NDUFB9* gene (rs34095749) with CD in African-Americans and a missense mutation in the *KCTD20* gene (rs2239808) with CD in European-Americans (see **Table 1**).

Utilizing a nominally significant threshold ($p < 0.05$), our gene-based test found 901 and 1,008 genes associated with CD in African-Americans (see GS357670) and European-Americans (see GS357669), respectively. We found a small, but significant, association between gene-based associations (*Z*-statistics) across European and African Americans ($B = 0.017$, s.e. = 0.008, p = 0.024; $R^2 = 0.0002$) and observed 59 genes ($p < 0.05$) that were nominally associated with CD in both ancestries.

Next, we investigated gene-based associations with CD for the 130 candidate neurotransmitter system genes commonly studied with cocaine use/addiction. Of these genes, we found ten nominally significant associations with CD from GABA, glutamate, endocannabanoid, serotonin, norepinephrine and acetylcholine genes (see **Figure 2**). The most significant candidate genetic association with CD (in African-Americans) came from the *CHRNB4* gene ($\#_{SNPs} = 422$, $Z = 4.00$, $p = 3.199e-05$), which resides in a validated gene cluster for CD (Gruzca et al. 2008) as well as nicotine dependence (Saccone et al. 2009). Despite dopamine's prominence in the candidate gene literature, we found no dopamine genes to be associated with CD (all $p > 0.108$). Candidate neurotransmitter genes were not enriched to be (nominally)

associated with CD, OR = 0.73, 95% CI [0.34, 1.40], p = 0.465. In other words, candidate neurotransmitter genes were no more likely to be (nominally) associated with CD than we would expect by chance.

Tissue Enrichment

To find tissues implicated in the genetic etiology of CD, we performed tissue specificity/enrichment analyses of the genes nominally associated with CD. Genes nominally associated with CD were enriched among numerous tissue types (see **Figure 3**). Despite minimal overlap of individual genes associated with CD across ancestries, 70.37% of significantly enriched tissues in African-Americans were also significantly enriched in European-Americans ($p_{adj} < 0.05$). Tissue overlap across ancestry exceeded what we would expect by chance alone, OR = 3.65, 95% CI [1.05, 13.70], p = 0.029. The replicated tissue types across ancestries tag plausibly implicated tissues in the genetic etiology of CD, including: heart, liver, blood and most brain regions and highlight various tissues for follow-up investigation.

Neuron Specific RNA-seq Analyses

Differential Expression

To follow-up genome-wide associations with CD, we used publically available data from dlPFC neurons from individuals with cocaine use disorder (CUD; n = 19) and cocaine-free controls (n=17; Ribeiro et al. 2017). After successful data normalization and adjustment for covariates, we found 133 differentially expressed genes/transcripts (all $p_{adj} < 0.05$; see **Figure 4** and GS357661). Similar to Ribeiro et al. 2017, 42.86% of differentially expressed genes/transcripts were non-coding, and of these, pseudogenes were the most abundant - including 15 pseudogenes of

mitochondrial inner membrane genes. Given that most non-coding transcripts lack detailed functional characterization, perhaps it is not surprising that differentially expressed genes/transcripts were not enriched for any KEGG pathways, BPs or MFs, all $p_{adj} > 0.089$, although we did identify some typical candidates for cocaine use/dependence. That is, consistent with previous research, we found increased expression of *FOSB* (Larson et al. 2010), *JUNB* (Guez-Barber et al. 2011), *ARC* (Zavala et al. 2008; Salery et al. 2017), *MECP2 (*Im et al. 2010; Deng et al. 2014*), NR4A2* (Lopez et al. 2019), *KCNJ9/GIRK3* (Rifkin et al. 2018; McCAll et al. 2019), *MAPK1* (Cahill et al. 2016) and *CAMK2N1* (Ribeiro et al. 2018). These genes represent various "immediate early genes" whose expression is induced by cocaine, intracellular signaling cascades that modulate neural responsiveness, and nuclear epigenetic transcripts that perturb the expression of numerous genes. No genome-wide significant association with CD (*FAM53B, C1qL2, KCTD20, STK38* or *NDUFB9)* was significantly differentially expressed in dlPFC neurons, all | $\log_2$ fold change | < 0.411, all $p > 0.026$, all $p_{adj} > 0.341$.

To complement our genome-wide analyses, we explored whether neuro-transcriptomic associations with CUD generalized across European- and African-Americans. After covariate adjustment, we found one gene that was significantly differentially expressed in European-Americans (*PAX8-AS1*, $\log_2$ fold change = -7.90, p_{adj} = 4.36e-5). In African-Americans, we found 37 significant differentially expressed genes – the most significant was the *CHRNG* gene ($\log_2$ fold change = 30.00, p_{adj} = 4.29e-30). While the top associations were different across ancestry, we found that the transcriptome-wide differential expression results significantly

correlated across ancestries, r = 0.174, p < 2.e-16 (see **Figure 5**). This association persisted after selecting cocaine-related genes/transcripts (r = 0.332, p < 2.e-16; 705 genes/transcripts).

Gene Co-expression Networks

Next, we modeled systems of co-expressed genes (gene networks) using weighted gene co-expression network analysis (WGCNA). Similar to previous WGCNA results with these data (Ribeiro et al. 2017), we constructed 12 gene co-expression networks – each of which is arbitrarily assigned to a color. To evaluate the stability and validity of our gene co-expression networks, we used a standard network preservation technique (Langfelder et al. 2011) to assess within sample and out of sample gene network reproducibility. Our analyses suggest that our WGCNA networks were highly reproducible/valid within sample (e.g., stable; all Z-summary > 16.19) and, except for the tan (Z-summary = 9.75) and yellow gene networks (Z-summary = 0.47), were valid in an independent RNA-seq sample (e.g., robust; all Z-summary > 12.67) of hippocampal tissue from human cocaine users and controls (Huggett & Stallings, 2019).

After validating our WGCNA gene co-expression networks, we associated these networks with CUD. Using an effect-size based approach and permuting p-values, we found 6 gene networks associated with CUD, all p_{adj} < 0.049 (see **Figure 6A**). We subsequently tested significantly associated WGCNA gene networks for enrichment of differentially expressed genes/transcripts (133 genes). We found significant enrichment of differentially expressed genes among one WGCNA network, the *blue gene network* (2,735 genes; see **Figure 6B** and GS357662). Thus,

412 the blue gene network was robustly associated with CUD and selected for follow-up
413 investigation.

414 The blue gene network recapitulated many molecular processes and was
415 significantly enriched for 13 KEGG pathways (p_{adj} < 0.05; see **Figure 6C**). Similar to
416 Ribeiro et al. (2017), our blue gene network was enriched for neuroplasticity
417 processes and also over-represented for various *neurotransmitt*er *signaling*
418 pathways, morphine addiction, intracellular signaling and circadian entrainment.
419 Note that other KEGG drug addiction pathways (nicotine addiction, alcoholism and
420 cocaine addiction) approached significant enrichment (all p: 0.007-0.054; all p_{adj}:
421 0.051-0.201). Of note, the blue gene network was enriched for the 130 candidate
422 neurotransmitter system genes, OR = 2.51, 95% CI [1.61, 3.83], p = 3.175e-05.

423 We then assessed the overlap between genetic predispositions to CD and the
424 blue gene network robustly associated with CUD. Of the five genome-wide
425 significant associations with CD, our analyses identified the *NDUFB9* and *C1qL2*
426 genes to be central entities (> 50%tile of module membership (kME), kME > 0.58) of
427 the blue gene network. To better understand the role of *NDUFB9* and *C1qL2* in the
428 context of cocaine addiction, we explored/visualized their co-expression patterns
429 with the blue network genes annotated for *neurotransmitt*er *signaling* and *drug*
430 *addictions*, see **Figure 7**. Of particular note, we found that our data-derived blue
431 gene network recapitulated previously established connections between *FOSB* and
432 *JUN* genes, which are thought to perpetuate chronic cocaine/drug seeking behavior
433 (Nestler et al. 2002) and further highlights the validity of the co-expression patterns
434 from this gene network.

We used co-expression patterns in the blue gene network to better understand biological functions of *NDUFB9* and *C1qL2* with cocaine use via a "guilt-by-association" approach (Oliver, 2000). Guilt-by-association analyses are commonly used to unravel the biological role of new disease genes and are based on the principle that if genes are highly associated with each other (e.g., co-expressed) they are more likely to share a function (van Dam et al. 2018). In our guilt-by-association technique, we selected the most highly co-expressed genes (weighted $r >$ 0.05 or raw $r > 0.78$) separately for *NDUFB9* and *C1qL2* within the blue gene network, and then investigated enrichment for biological processes, molecular functions and KEGG pathways. Our guilt-by-association analyses indicated that the *NDUFB9* gene might play a role in cell death, synaptic plasticity and cell adhesion (see **Table 2**, all $p_{adj} < 0.037$). Highly co-expressed genes with *C1qL2* were significantly enriched for: neurotransmitter signaling, drug response, synaptic plasticity, cell proliferation and neurodevelopment (see **Table 2**, all $p_{adj} < 0.045$).

Discussion

We extend previous genome-wide research (Gelernter et al. 2014; Huggett & Stallings, 2019) that identified four genes significantly associated with cocaine dependence (CD; *C1qL2*, *FAM53B, KCTD20, STK38*) by discovering one novel gene (*NDUFB9*) implicated in the genetic liability to CD for African-Americans. Our study highlights associations between two missense mutations and CD that may interfere with the product of the *NDUFB9* and *KCTD20* genes. Similar to other psychiatric genetic research (Johnson et al. 2017; Border et al. 2019), we found minimal evidence indicating that genes from candidate neurotransmitter systems contribute

458 to the genetic predisposition of CD. Genome-wide significant genes associated with
459 CD were *not* differentially expressed in dlPFC neurons between individuals with
460 CUD and cocaine-free controls. However, *NDUFB9* and *C1qL2* were central parts of a
461 gene co-expression network associated with CUD and exhibited co-expression with
462 relevant drug addiction genes. So while most GWAS findings tend not correspond
463 with pre-hypothesized targets, they may still play a broader role in biologically
464 relevant systems. Similarly, genome-wide associations with psychiatric traits
465 (including alcohol dependence) demonstrated appreciable overlap with PFC gene
466 co-expression networks associated with these traits and corresponded to neuronal,
467 synaptic and mitochondrial functions (Gandal et al. 2018; Kapoor et al. 2019).

468 Our study suggests common biological contributions to cocaine addiction
469 across ancestries/ethnicities. Similar to other substance dependence research
470 (Brick et al. 2019), we found that the individual genetic predispositions to CD
471 demonstrated (modest) genetic overlap across African-Americans and European-
472 Americans. Robust across ancestries, we discovered that the genetic liability of CD
473 manifested as a multi-organ phenomenon involving the heart, liver, blood and brain.
474 Using RNA-seq from PFC neurons, we identified convergence of cocaine-related
475 gene expression across African-Americans and European-Americans, albeit with
476 small to moderate effect sizes. One potential reason for the modest magnitudes of
477 these associations is that rates of psychiatric co-morbidities between cocaine
478 abusing European-American and African-American seem to differ (Petry, 2003), but
479 many other factors could be at play. To our knowledge, this is the first study to

assess the cross-ancestry transcriptome-wide neurodiversity/similarity for a psychiatric trait - making interpretations difficult.

We found evidence of disrupted GABA, but not glutamate, neurotransmitter signaling in dlPFC neurons of human cocaine addicts (blue gene network; see **Fig 6**). PFC GABAergic signaling is sparsely studied in rodent models of cocaine use, but some evidence suggests that GABA regulates prefrontal disinhibition (Cass et al. 2013). We discovered that various GABA genes (*GABBR2, GABRA1, GABRA4, GABRB2, GABARAPL1, GABARAPL2*) were core elements ("hub genes"; top 10% of gene network connectivity) of PFC network function for individuals with CUD. That is, GABAergic genes demonstrated very high co-expression/connectivity patterns with other genes in the blue gene network; suggesting GABAergic transmission plays a critical, yet unappreciated, modulatory role of PFC neurons in disordered cocaine use.

The blue gene network associated with CUD also suggests that catecholamine; acetylcoholine and endocannabanoid signaling play an important role in the PFC for the neuropathology of CUD. Specifically, PFC *DRD5* activity may mediate executive functioning (Carr et al. 2017) and impulsive decision-making (Loos et al. 2010) as well as *HTR1A* and *ADRA1A* could regulate PFC glutamate and/or GABA transmission and various cocaine-related behaviors (Mitrano et al. 2012; Howell et al. 2014). Particular nicotinic (*CHRNA6, CHRNB2*) and muscarinic acetylcholine subunit genes (*CHRM1, CHRM4*) we found to be associated with CUD have previously been implicated in rodent cocaine research (Carrigan et al. 2007; Dencker et al. 2012; Sanjakdar et al. 2015) and might govern selective attention and

promote incentive salience to drugs/drug-related cues (Williams et al. 2008). Lastly, cocaine has found to alter expression of endocannabanoid genes/receptors in the mouse PFC (Bystrowska et al. 2019), which could facilitate the strength of connections between PFC neurons (Kasanetz et al. 2013). Our results provide an in human corroboration of specific genes and pathways commonly studied in animal models of cocaine use and/or drug-related behaviors.

The combination of genomic and bioinformatics techniques may help contextualize and interpret nebulous genetic associations with human traits. *NDUFB9* is a subunit of the inner mitochondrial complex I. Evidence indicates that cocaine inhibits complex I of the inner mitochondrial membrane (Cunha-Oliveira et al. 2013), which is similar to other genetic associations with substance use/dependence that implicate binding targets of specific drugs. Mitochondrial complex I is thought to mediate altered energy metabolism and cocaine-induced neurotoxicity (Dey et al. 2007; Periera & Cunha-Oliveira et al. 2017) and is consistent with our "guilt-by-association" results suggesting *NDUFB9* may be involved in neuro-degeneration and ATP production (oxidative phosphorylation). Additionally, analogous to research highlighting the role of mitochondria in drug addiction (Sadakierska-Chudy et al. 2014), our guilt-by-association analyses suggest that *NDUFB9* could be involved in cell death, synaptic plasticity and calcium signaling. *NDUFB9* is not the only mitochondrial gene implicated in cocaine addiction. We found 26 different mitochondrial inner membrane genes within the blue gene network associated with CUD, including 12 *NDUF* subunits – suggesting links between cocaine use and broad mitochondrial functioning. Accordingly,

various mitochondrial genes have demonstrated associations with human cocaine abuse/dependence in the dlPFC (Lehrmann et al. 2013), hippocampus (Zhou et al. 2011) and midbrain (Bannon et al. 2014). Despite the mounting evidence, very little is known regarding the relation between mitochondrial genes and cocaine or drug use behavior. One study indicates that mitochondrial genes may contribute to cocaine withdrawal, as they observed differential expression of 40 mitochondrial genes in the PFC of mice experiencing protracted abstinence after chronic high doses of cocaine use (Li et al. 2017).

The *C1qL2* gene is secreted from the innate immune system and is thought to modulate trans-synaptic glutamatergic connections (Evans et al. 2019). Similar to previous work (Matsuda et al. 2017), we identified *C1qL2* to be co-expressed with *C1qL3* and found that *C1qL2* may regulate glutamate receptor signaling (see **Table** 2). Extending this research, we hypothesize and provide novel evidence that *C1qL2* may be involved in broader neurotransmitter signaling (GABA, acetylcholine and serotonin), ion transport (K^+/Na^+), neuro-development and various drug addiction pathways. *C1qL2* may be a particularly tantalizing candidate for follow-up, as it is implicated in typical biological processes underlying cocaine use, is highly conserved across species and is differentially expressed in the hippocampus in mouse models of cocaine use (Walker et al. 2018). Overall, we prioritize a specific cell-type for follow-up investigation (neurons) and propose specific biological roles/hypotheses for otherwise obscure genomic associations with cocaine addiction.

548 This study should be interpreted with the following limitations. While, we
549 used the largest GWAS of cocaine addiction to date, our (highly) selected sample had
550 uneven case/control ratios and was not large by contemporary standards and thus
551 the estimates from this study were approximate. The gene-based associations we
552 observed with CD barely surpassed genome-wide significance, warranting larger
553 studies to replicate these findings. Although, including only individuals who were
554 exposed to cocaine may have enhanced the power to identify genes associated with
555 CD (Cabana-Dominguez et al. 2019; Polimanti et al. 2020). Our tissue enrichment
556 findings indicated plausible tissue types for cocaine addiction, suggesting the
557 importance of follow-up among multiple tissue types, however, not all tissues
558 seemed directly relevant for CD (e.g., muscle/skeletal) and certain genes may exert
559 tissue specific functions. Tissue-enrichment analyses utilized GTEx samples, which
560 included mostly Caucasian individuals and may complicate our cross-ancestry
561 comparisons. Our RNA-seq design cannot disentangle whether findings are
562 attributed to chronic cocaine use, acute cocaine toxicity, or psychiatric co-
563 morbidities; but it is reassuring to detect some usual suspects in the realm of
564 cocaine addiction. While our RNA-seq results are theoretically specific to neurons,
565 they do not distinguish between types of neurons and also included various
566 genes/transcripts that are non-neuronal (e.g., glial genes).
567 In conclusion, our study translates genetic findings across methods and
568 ancestries using independent samples. We identified significant overlap across
569 ancestries for trancriptomic, but not genomic, associations with cocaine addiction.
570 Neurotransmitter genes generally demonstrated little contribution to the genetic

architecture of CD, *but* were prominent features underlying the neuropathology of CUD. That is, DNA variation within classical neurotransmitter genes was not typically associated with the genetic predisposition to cocaine addiction, but disrupted systems of neurotransmitter genes were associated with cocaine addiction via neuronal RNA expression/co-expression. Significant genome-wide associations with CD were linked to broad systems of genes/transcripts in PFC neurons from individuals with CUD. Ultimately, our study represents a proof-of-principle that utilizes hypothesis-free methods for generating testable hypotheses regarding the role of genes detected by GWASs and shows the promise of multi-omic analyses. We believe that this line of research provides an important alternative approach for validating genetic associations especially when no genomic replication data exists. Our study may also serve a supplemental purpose for neuroscientists and experimental researchers to help refine particular genes in specific tissues/cell-types for follow-up investigation, while also providing tangible molecular interpretations for otherwise obscure genes identified by genome-wide association analyses.

Figures/Table Legends

Figure 1- Legend: Miami plot visualizing results from gene-based association analyses. Each dot represents an individual protein-coding gene, x-axis denotes chromosome number and the y-axis shows the $-\log_{10}$ *p*-value. African-American results are displayed on top and European-American results (from Huggett & Stallings, 2019) are shown on bottom. Dashed-red line represents genome-wide significance and dashed brown line represents the unadjusted/nominal *p*-value threshold < 0.05. Red dots are genes nominally significant in both African and European-Americans.

Figure 2- Legend: Miami plot showing the associations of 130 genes from candidate neurotransmitter systems. Each gene is color coded by neurotransmitter type and the different shapes represent the different parts of the system. The x-axis denotes chromosome number and the y-axis shows the $-\log_{10}$ *p*-value with African-Americans displayed on top and European-Americans shown on bottom. The dashed-red line represents the Bonferonni correction for multiple testing ($p < 0.05 / 130$) and the dashed-brown line represents the unadjusted/nominal *p*-value threshold < 0.05.

Figure 3- Legend: Shows the implicated tissue types based on genes nominally associated with cocaine dependence (CD) separately by ancestry. The x-axis shows all tissue types (GTEx) sorted alphabetically and the y-axis represents the $-\log_{10}$ *p*-value. Solid boxes denote results from the African-American analysis and dashed boxes show European-American results from (Huggett & Stallings, 2019). Red bars show replicated tissue types that were significantly enriched ($p_{adj} < 0.05$) across both ancestries. The labels of replicated tissues are emphasized in bold text on the x-axis.

Figure 4- Legend: Volcano plot showing genes/transcripts that are expressed differently in human PFC neurons between controls (n = 17) and individuals with cocaine use disorder (CUD; n = 19). Each dot represents a gene/transcript. The x-axis denotes the $\log_2$ fold change with positive values corresponding to increased expression in those with CUD. The y-axis shows the $-\log_{10}$ FDR adjusted *p*-value and all genes above the dashed-red line survive correction for multiple testing (133 gene/transcripts; $p_{adj} < 0.05$). We labeled all genes significantly associated with the genetic predisposition to CD and highlighted significantly differentially expressed genes/transcripts previously implicated in cocaine use and the most abundant non-coding transcripts (pseudogenes).

Figure 5– Legend: Heat scatter plot depicting the correlation of neuronal dlPFC gene expression associated with cocaine use disorder (CUD) from African-Americans (n = 13) and European-Americans (n = 10). The x-axis shows the *Wald* statistics from the European-American differential expression analysis and the y-axis represents the Wald statistics from the African-American differential expression analysis. Each dot represents a specific gene/transcript and the bright red color shows the highest

frequency, whereas the light purple/pink indicates the lowest frequency of genes/transcripts. The dashed black line highlights the Pearson Product correlation of gene expression across ethnicities (r = 0.174, p < 2e-16).

Figure 6– Legend: **A)** The x-axis shows the twelve WGCNA gene co-expression networks. The y-axis shows the absolute value of Wald statistics (from whole sample differential expression analysis) of all genes within a defined/discrete WGCNA network. The directions of associations were determined by assessing whether mean effect sizes for gene networks were positive or negative. All WGCNA gene networks to the right of the dashed red line are significantly associated with cocaine use disorder (p_{adj} < 0.05). **B)** The six associated WGCNA gene networks were subsequently tested for enrichment of the 133 differentially expressed genes in dlPFC neurons. The y-axis represents the odd's ratio calculated by a two-sided Fisher's exact test. Only the blue gene network demonstrated significant enrichment and was selected for follow-up investigation. **C)** Potential functions of blue gene network via functional annotation analysis of pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG; all p_{adj} < 0.05). We picked 30 representative functions/pathways and grouped them into five domains, which are labeled by colors.

Figure 7– Legend: Shows the genes from the blue gene network significantly enriched for *drug addiction* and *neurotransmission* from Kyoto Encyclopedia of Genes and Genomes (KEGG, 2019) and their relation to the genes associated with the predisposition to CD (in triangles). Co-expression patterns with *NDUFB9* and *C1qL2* are highlighted in red. Only co-expression patterns above a weighted r > 0.05 are shown. Genes in cyan show increased expression in dlPFC neurons for those with CUD and magenta represents decreased expression.

Table 1- Legend: We collapsed *KCTD20* & *STK38* into a single category because they stem from the same genomic signal. AA stands for African-American and EA stands for European-American ancestry. The # of parameters represents the amount of independent signals tested within a protein coding gene and differ across ancestries due to disparate LD patterns. We estimated the linkage disequilibrium patterns of missense variants with lead SNPs using LDlink, and selecting the African-American and CEU reference panels. Note that *C1qL2* only had one missense mutation, but was not tested included in the genome-wide association study on cocaine dependence due to low minor allele frequency across ancestries (< 1%).

Table 2- Legend: Our "guilt-by-association" approach assesses the function of genes/transcripts that are highly co-expressed with *NDUFB9* and *C1qL2* in the blue gene network associated with CUD and assesses their enrichment for biological processes, molecular functions and KEGG pathways using DAVID (Huang et al. 2009). We selected the most highly co-expressed genes with *NDUFB9* (300 genes/transcripts) and *C1qL2* (694 genes/transcripts) in the blue gene network by using an arbitrary co-expression threshold of: weighted r > 0.05 (raw r > 0.78).

References

587 Anders, S., Pyl, P. T., & Huber, W. (2015). Genome analysis HTSeq — a Python framework to work with
588 high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169.
589 https://doi.org/10.1093/bioinformatics/btu638

590 Bacher, R., Chu, L. F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., … Kendziorski, C. (2017).
591 SCnorm: Robust normalization of single-cell RNA-seq data. Nature Methods, 14(6), 584–586.

592 Baker, E. J., Jay, J. J., Bubier, J. A., Langston, M. A., & Chesler, E. J. (2012). GeneWeaver : a web-based
593 system for integrative functional genomics. *Nucleic Acids Research*, *40*, 1067–1076.
594 https://doi.org/10.1093/nar/gkr968

595 Bannon, M. J., Johnson, M. M., Michelhaugh, S. K., Hartley, Z. J., Halter, S. D., David, J. A., … Schmidt,
596 C. J. (2014). A molecular profile of cocaine abuse includes the differential expression of genes that
597 regulate transcription, chromatin, and dopamine cell phenotype. Neuropsychopharmacology, 39(9),
598 2191–2199.

599 Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., & Keller, M. C. (2018).
600 No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major
601 Depression Across Multiple Large Samples. The American Journal of Psychiatry, (July), 1–12.

602 Brick, L. A., Keller, M. C., Knopik V. S., & Palmer. R. H. C. (2017). Shared additive genetic variation for
603 alcohol dependence among subjects of African and European ancestry. *Addiction Biology*, 24(1),
604 132-144.

605 Bystrowska, B., Frankowska, M., Smaga, I., Niedzielska-Andres, E., Pomierny-Chamioło, L., & Filip, M.
606 (2019). Cocaine-Induced Reinstatement of Cocaine Seeking Provokes Changes in the
607 Endocannabinoid and N-Acylethanolamine Levels in Rat Brain Structures. Molecules, 24(6), 1125.

608 Cabana-domínguez, J., Shivalikanjli, A., & Fernàndez-castillo, N. (2019). Progress in
609 Neuropsychopharmacology & Biological Psychiatry Genome-wide association meta-analysis of
610 cocaine dependence : Shared genetics with comorbid conditions. *Progress in*
611 *Neuropsychopharmacology & Biological Psychiatry*, *94*, 109667.
612 https://doi.org/10.1016/j.pnpbp.2019.109667

613 Cahill, M. E., Bagot, R. C., Gancarz, A. M., Walker, D. M., Sun, H. S., Wang, Z. J., … Nestler, E. J.
614 (2016). Bidirectional Synaptic Structural Plasticity after Chronic Cocaine Administration Occurs
615 through Rap1 Small GTPase Signaling. Neuron, 89(3), 566–582.

616 Carr, G. V., Maltese, F., Sibley, D. R., Weinberger, D. R., & Papaleo, F. (2017). The dopamine D5 receptor
617 is involved in working memory. Frontiers in Pharmacology, 8(OCT), 1–9.

618 Carrigan, K. A., & Dykstra, L. A. (2007). Behavioral effects of morphine and cocaine in M1 muscarinic
619 acetylcholine receptor-deficient mice. Psychopharmacology, 191(4), 985–993.

620 Cass, D. K., Thomases, D. R., Caballero, A., & Tseng, K. Y. (2013). Developmental Disruption of Gamma-
621 Aminobutyric Acid Function in the Medial Prefrontal Cortex by Noncontingent Cocaine Exposure
622 during Early Adolescence. Biological Psychiatry, 74(7), 490–501.

623 Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., … Kellis, M. (2015).
624 FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. The New England Journal of
625 Medicine, 373(10), 895–907.

626 Colhoun, H. M., McKeigue, P. M., Smith, G. D. (2003). Problems of reporting genetic associations with
627 complex outcomes. The Lancet, 361:865-872.

628 Consortium, T. Gte. (2013). The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project.
629 Nature Genetics, 45(6), 580–585.

630 Crunelle, C. L., Kaag, A. M., van Wingen, G., van den Munkhof, H. E., Homberg, J. R., Reneman, L., &
631 van den Brink, W. (2014). Reduced Frontal Brain Volume in Non-Treatment-Seeking Cocaine-
632 Dependent Individuals: Exploring the Role of Impulsivity, Depression, and Smoking. Frontiers in
633 Human Neuroscience, 8(January), 1–7.

634 Cunha-Oliveira, T., Rego, A. C., Cardoso, S. M., Borges, F., Swerdlow, R. H., Macedo, T., & de Oliveira,
635 C. R. (2006). Mitochondrial dysfunction and caspase activation in rat cortical neurons treated with
636 cocaine or amphetamine. Brain Research, 1089(1), 44–54.

637 Cunha-Oliveira, T., Silva, L., Silva, A. M., Moreno, A. J., Oliveira, C. R., & Santos, M. S. (2013).
638 Mitochondrial complex I dysfunction induced by cocaine and cocaine plus morphine in brain and
639 liver mitochondria. Toxicology Letters, 43, 2013.

640 de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized Gene-Set
641 Analysis of GWAS Data. PLoS Computational Biology, 11(4), 1–19.

642 Dencker, D., Weikop, P., Sørensen, G., Woldbye, D. P. D., Wörtwein, G., Wess, J., & Fink-Jensen, A.
643 (2012). An allosteric enhancer of M4 muscarinic acetylcholine receptor function inhibits behavioral
644 and neurochemical effects of cocaine. Psychopharmacology, 224(2), 277–287.

645 Deng, J. V., Wan, Y., Wang, X., Cohen, S., Wetsel, W. C., Greenberg, M. E., … West, A. E. (2014).
646 MeCP2 Phosphorylation Limits Psychostimulant-Induced Behavioral and Neuronal Plasticity.
647 Journal of Neuroscience, 34(13), 4519–4527.

648 Dey, S., & Snow, D. M. (2007). Cocaine exposure in vitro induces apoptosis in fetal locus coeruleus
649 neurons through TNF-α-mediated induction of Bax and phosphorylated c-Jun NH2-terminal kinase.
650 Journal of Neurochemistry, 103(2), 542–556.

651 Dickson, P. E., Miller, M. M., Calton, M. A., Bubier, J. A., Cook, M. N., Goldowitz, D., … Mittleman, G.
652 (2016). Systems genetics of intravenous cocaine self-administration in the BXD recombinant inbred
653 mouse panel. Psychopharmacology, 5(6), 1–8.

654 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013).
655 STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21.

656 Doughetry, J. D., Schmidt, E. F., Nakajima, M., & Heintz, N. (2010). Analytical approaches to RNA
657 profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Research,*
658 38(13), 4218-4230.

659 Evans, A. J., Gurung, S., Henley, J. M., Nakamura, Y., & Wilkinson, K. A. (2019). Exciting Times: New
660 Advances Towards Understanding the Regulation and Roles of Kainate Receptors. Neurochemical
661 Research, 44(3), 572–584.

662 Gelernter, J., Sherva, R., Koesterer, R., Almasy, L., Zhao, H., Kranzler, H. R., & Farrer, L. (2014).
663 Genome-wide association study of cocaine dependence and related traits: FAM53B identified as a
664 risk gene. Molecular Psychiatry, 19(6), 717–723.

665 Goldstein, R. Z., & Volkow, N. D. (2011). Dysfunction of the prefrontal cortex in addiction: neuroimaging
666 findings and clinical implications. Nat Rev Neurosci, 12(11), 652–669.

667 Grucza, R. A., Wang, J. C., Stitzel, J. A., Hinrichs, A. L., Scott, F., Saccone, N. L., … Bierut, L. J. (2008).
668 A Risk Allele for Nicotine Dependene in CHRNA5 Is a Protective Allele for Cocaine Dependence.
669 Biological Psychiatry, 64(11), 922–929.

670 Guez-Barber, D., Fanous, S., Golden, S. A., Schrama, R., Koya, E., Stern, A. L., … Hope, B. T. (2011).
671 FACS Identifies Unique Cocaine-Induced Gene Regulation in Selectively Activated Adult Striatal
672 Neurons. Journal of Neuroscience, 31(11), 4251–4259.

673 Howell, L. L., & Cunningham, K. A. (2014). Serotonin 5-HT2 Receptor Interactions with Dopamine
674 Function: Implications for Therapeutics in Cocaine Use Disorder. Pharmacological Reviews, 67(1),
675 176–197.

676 Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene
677 lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(2), 44–57.
678 https://doi.org/10.1038/nprot.2008.211

679 Huggett, S. B., & Stallings, M. C. (2019). Cocaine'omics: Genome-wide and transcriptome-wide analyses
680 provide biological insight into cocaine use and dependence. Addiction Biology, (January), 1–10.

681 Im, H.-I., Hollander, J. A., Bali, P., & Kenny, P. J. (2010). MeCP2 controls BDNF expression and cocaine
682 intake through homeostatic interactions with microRNA-212. Nat Neuroscience, 5(6), 1–8.

683 Johnson, E. C., Border, R., Melroy-greif, W. E., Leeuw, C. De, Ehringer, M. A., & Keller, M. C. (2018).
684 associated with schizophrenia than non-candidate genes. Biological Psychiatry, 82(10), 702–708.

685 Kalivas, P. W., Volkow, N., & Seamans, J. (2005). Unmanageable motivation in addiction: A pathology in
686 prefrontal-accumbens glutamate transmission. Neuron, 45(5), 647–650.

687 Kapoor, M., Wang, J. C., Farris, S. P., Liu, Y., McClintick, J., Gupta, I., … Goate, A. (2019). Analysis of
688 whole genome-transcriptomic organization in brain to identify genes associated with alcoholism.
689 Translational Psychiatry, 9(1).

690 Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Wang, Q., Collins, R. L., … Banks, E.
691 (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-
692 function intolerance across human protein-coding genes. *bioRxiv*.

693 Kasanetz, F., Lafourcade, M., Deroche-Gamonet, V., Revest, J., Berson, N., Balado, E., … Manzoni, O.
694 (2013). Prefrontal synaptic markers of cocaine addiction-like behavior in rats. Molecular Psychiatry,
695 18, 729–737.

696 Kuleshov, M. V, Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., … Ma'ayan, A.
697 (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic
698 Acids Research, 44(W1), W90-7.

699 Kumar, V., Kim, K., Joseph, C., Kourrich, S., Yoo, S.-H., Huang, H. C., … Takahashi, J. S. (2013).
700 C57BL/6N Mutation in Cytoplasmic FMRP interacting protein 2 Regulates Cocaine Response.
701 Science, 342, 1508–1512.

702 Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis.
703 BMC Bioinformatics, 9, 559.

704 Langfelder, P., Luo, R., Oldham, M. C., & Horvath, S. (2011). Is my network module preserved and
705 reproducible? PLoS Computational Biology, 7(1).

706 Larson, E. B., Akkentli, F., Edwards, S., Graham, D. L., Simmons, D. L., Alibhai, I. N., … Self, D. W.
707 (2010). Striatal regulation of ΔfosB, FosB, and cFos during cocaine self-administration and
708 withdrawal. Journal of Neurochemistry, 115(1), 112–122.

709 Leek, J. T. (2014). Svaseq: Removing batch effects and other unwanted noise from sequencing data.
710 Nucleic Acids Research, 42(21), e161.

711 Lehrmann, E., Oyler, J., Vawter, M. P., Hyde, T. M., Kolachana, B., Kleinman, J. E., … Freed, W. J.
712 (2003). Trancriptional profiling in the human prefrontal cortex: Evidence for two activational states
713 associated with cocaine abuse. Pharmacogenetics, 3(1), 27–40.

714 Li, M., Xu, P., Xu, Y., Teng, H., Tian, W., Du, Q., & Zhao, M. (2017). Dynamic expression changes in the
715 transcriptome of the prefrontal cortex after repeated exposure to cocaine in mice. Frontiers in
716 Pharmacology, 8(MAR), 1–10.

717 Loos, M., Pattij, T., Janssen, M. C. W., Counotte, D. S., Schoffelmeer, A. N. M., Smit, A. B., … Van
718 Gaalen, M. M. (2010). Dopamine receptor D1/D5gene expression in the medial prefrontal cortex
719 predicts impulsive choice in rats. Cerebral Cortex, 20(5), 1064–1070.

720 Lopez, A. J., Hemstedt, T. J., Jia, Y., Hwang, P. H., Campbell, R. R., Kwapis, J. L., … Wood, M. A.
721 (2019). Epigenetic regulation of immediate-early gene Nr4a2/Nurr1 in the medial habenula during
722 reinstatement of cocaine-associated behavior. Neuropharmacology, 153(March), 13–19.

723 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for
724 RNA-seq data with DESeq2. Genome Biology, 15(12), 550.

725 Machiela, M. J., & Chanock, S. J. (2015). LDlink : a web-based application for exploring population-
726 specific haplotype structure and linking correlated alleles of possible functional variants.
727 *Bioinformatics*, *31*(21), 3555–3557. https://doi.org/10.1093/bioinformatics/btv402

728 Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Current clinical
729 use of polygenic scores will risk exacerbating health disparities. Nature Genetics, 51(4), 584–591.

730 Matsuda, K. (2017). Synapse organization and modulation via C1q family proteins and their receptors in
731 the central nervous system. Neuroscience Research, 116, 46–53.

732 Maurano, M., Maurano, M., Humbert, R., Rynes, E., Thurman, R., Haugen, E., … Stamatoyannopoulos, J.
733 (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.
734 Science, 337(6099), 1190–1195.

735 McCall, N. M., Fernandez de Velasco, E. M., & Wickman, K. (2019). GIRK channel activity in dopamine
736 neurons of the ventral tegmental area bi-directionally regulates behavioral sensitivity to cocaine. The
737 Journal of Neuroscience, 39(19), 3101–3118.

738 Michael J. Gandal, Jillian R. Haney, Neelroop N. Parikshak, Virpi Leppa, Steve Horvath, D. H. G. (2018).
739 Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap.
740 Science, 697(February), 693–697.

741 Mitrano, D. A., Schroeder, J. P., Smith, Y., Cortright, J. J., Bubula, N., Vezina, P., & Weinshenker, D.
742 (2012). Alpha-1 adrenergic receptors are localized on presynaptic elements in the nucleus accumbens
743 and regulate mesolimbic dopamine transmission. Neuropsychopharmacology, 37(9), 2161–2172.

744 Munafo, M. R., & Flint, J. (2009). Replication and heterogeneity in gene x environment interaction studies.
745 International Journal of Neuropsychopharmacology, 12, 727–729.

746 Nestler, E. J., Barrot, M., & Self, D. W. (2002). FosB: A sustained molecular switch for addiction.
747 Proceedings of the National Academy of Sciences, 98(20), 11042–11046.

748 Oliver, S. (2000). Guilt-by-association goes global. Nature, 403(6770), 601–603.

749 Pereira, S. P., & Cunha-Oliveira, T. (2017). Role of Mitochondria on the Neurological Effects of Cocaine.
750 The Neuroscience of Cocaine: Mechanisms and Treatment. Elsevier Inc.

751 Peterson, R. E., et al. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations:
752 Opportunities, Methods, Pitfalls and Recommendations. Cell, 179:589-603.

753 Petry, N. M. (2003). A comparison of African American and non-Hispanic Caucasian cocaine-abusing
754 outpatients. *Drug and Alcohol Dependence*, *69*, 43–49.

755 Polimanti et al. (2020). Leveraging genome-wide data to investigate differences between opioid use vs.
756 opioid dependence in 41,176 individuals from the Psychiatric Genomics Consortium. *Molecular*
757 *Psychiatry.* https://doi.org/10.1038/s41380-020-0677-9.

758 Ponomarev, I., Wang, S., Zhang, L., Harris, R. A., & Mayfield, R. D. (2012). Gene Coexpression Networks
759 in Human Brain Identify Epigenetic Modifications in Alcohol Dependence. Journal of Neuroscience,
760 32(5), 1884–1897.

761 Ribeiro, E. A., Salery, M., Scarpa, J. R., Calipari, E. S., Hamilton, P. J., Ku, S. M., … Nestler, E. J. (2018).
762 Transcriptional and physiological adaptations in nucleus accumbens somatostatin interneurons that
763 regulate behavioral responses to cocaine. Nature Communications, 9(1).

764 Ribeiro, E. A., Scarpa, J. R., Garamszegi, S. P., Kasarskis, A., Mash, D. C., & Nestler, E. J. (2017). Gene
765 Network Dysregulation in Dorsolateral Prefrontal Cortex Neurons of Humans with Cocaine Use
766 Disorder. Scientific Reports, 7(1), 1–10.

767 Rifkin, R. A., Huyghe, D., Li, X., Parakala, M., Aisenberg, E., Moss, S. J., & Slesinger, P. A. (2018).
768 GIRK currents in VTA dopamine neurons control the sensitivity of mice to cocaine-induced
769 locomotor sensitization. Proceedings of the National Academy of Sciences, 115(40), E9479–E9488.

770 Saccone, N. L., Wang, J. C., Breslau, N., Johnson, E. O., Hatsukami, D., Saccone, S. F., … Bierut, L. J.
771 (2009). The CHRNA5-CHRNA3-CHRNB4 Nicotinic Receptor Subunit Gene Cluster Affects Risk
772 for Nicotine Dependence in African-Americans and in European-Americans. Cancer Research, (17),
773 6848–6857.

774 Sadakierska-Chudy, A., Frankowska, M., & Filip, M. (2014). Mitoepigenetics and drug addiction.
775 Pharmacology and Therapeutics, 144(2), 226–233.

776 Salery, M., Dos Santos, M., Saint-Jour, E., Moumné, L., Pagès, C., Kappès, V., … Vanhoutte, P. (2017).
777 Activity-Regulated Cytoskeleton-Associated Protein Accumulates in the Nucleus in Response to
778 Cocaine and Acts as a Brake on Chromatin Remodeling and Long-Term Behavioral Alterations.
779 Biological Psychiatry, 81(7), 573–584.

780 Sanjakdar, S. S., Maldoon, P. P., Marks, M. J., Brunzell, D. H., Maskos, U., McIntosh, J. M., … Damaj, M.
781 I. (2015). Differential roles of α6β2∗ and α4β2∗ neuronal nicotinic receptors in nicotine- and
782 cocaine-conditioned reward in mice. Neuropsychopharmacology, 40(2), 350–360.

783 Sekar, A., Bialas, A. R., de Rivera H., Hammond, T., Kamitaki, N., Tooley, K., Presumey, J., Baum M.,
784 Doren, V. V., Genovese, G., Rose, S. A., Handsaker, R. E., Schizophrenia Working Group of the
785 Psychiatric Genomics Consortium, Daly, M. J., Carroll, M. C., Stevens, B., McCarroll, S. A. (2016).
786 Schizophrenia risk from complex variation of complement component 4. Nature, 530:177-183.

787 van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018). Gene co-expression
788 analysis for functional classification and gene-disease predictions. Briefings in Bioinformatics, 19(4),
789 575–592.

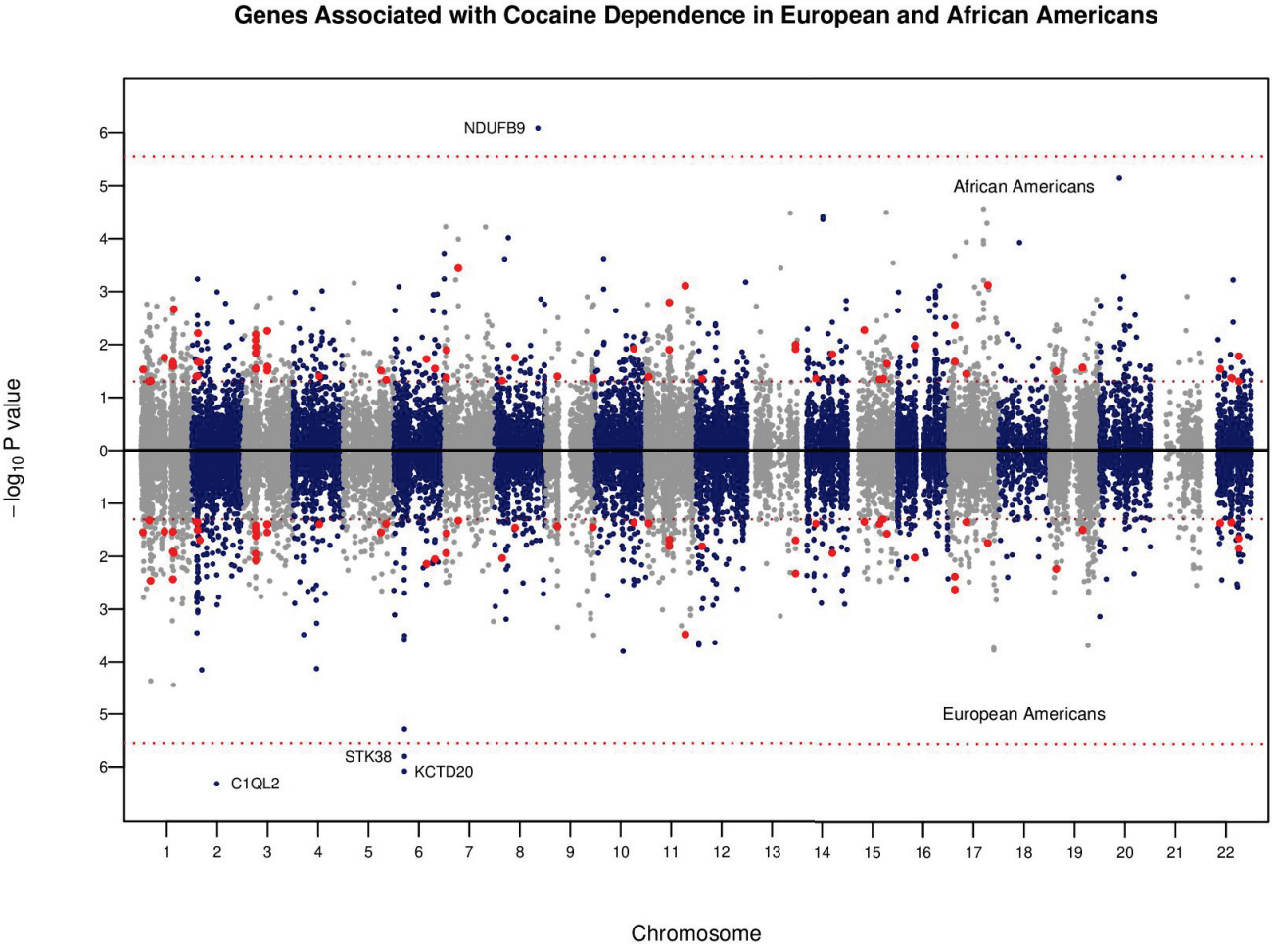
790 Walker, D. M., Cates, H. M., Loh, Y. H. E., Purushothaman, I., Ramakrishnan, A., Cahill, K. M., …
791 Nestler, E. J. (2018). Cocaine Self-administration Alters Transcriptome-wide Responses in the
792 Brain's Reward Circuitry. Biological Psychiatry, 84(12), 867–880.

793 Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). Functional mapping and
794 annotation of genetic associations with FUMA. Nature Communications, 8(1), 1826.

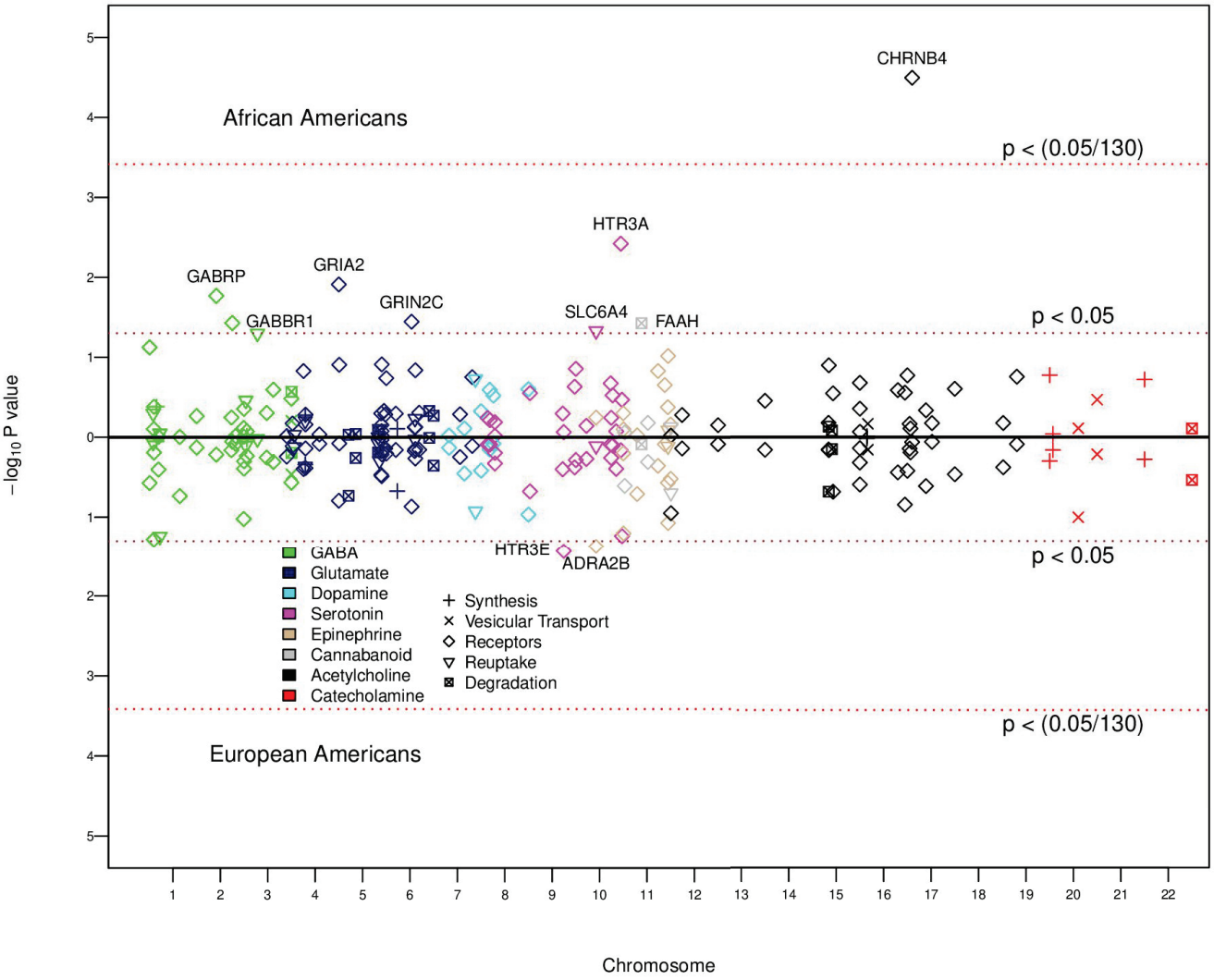
795 Williams, M., & Adinoff, B. (2008). The role of acetylcholine in cocaine addiction.
796 Neuropsychopharmacology, 33(8), 1779–1797.

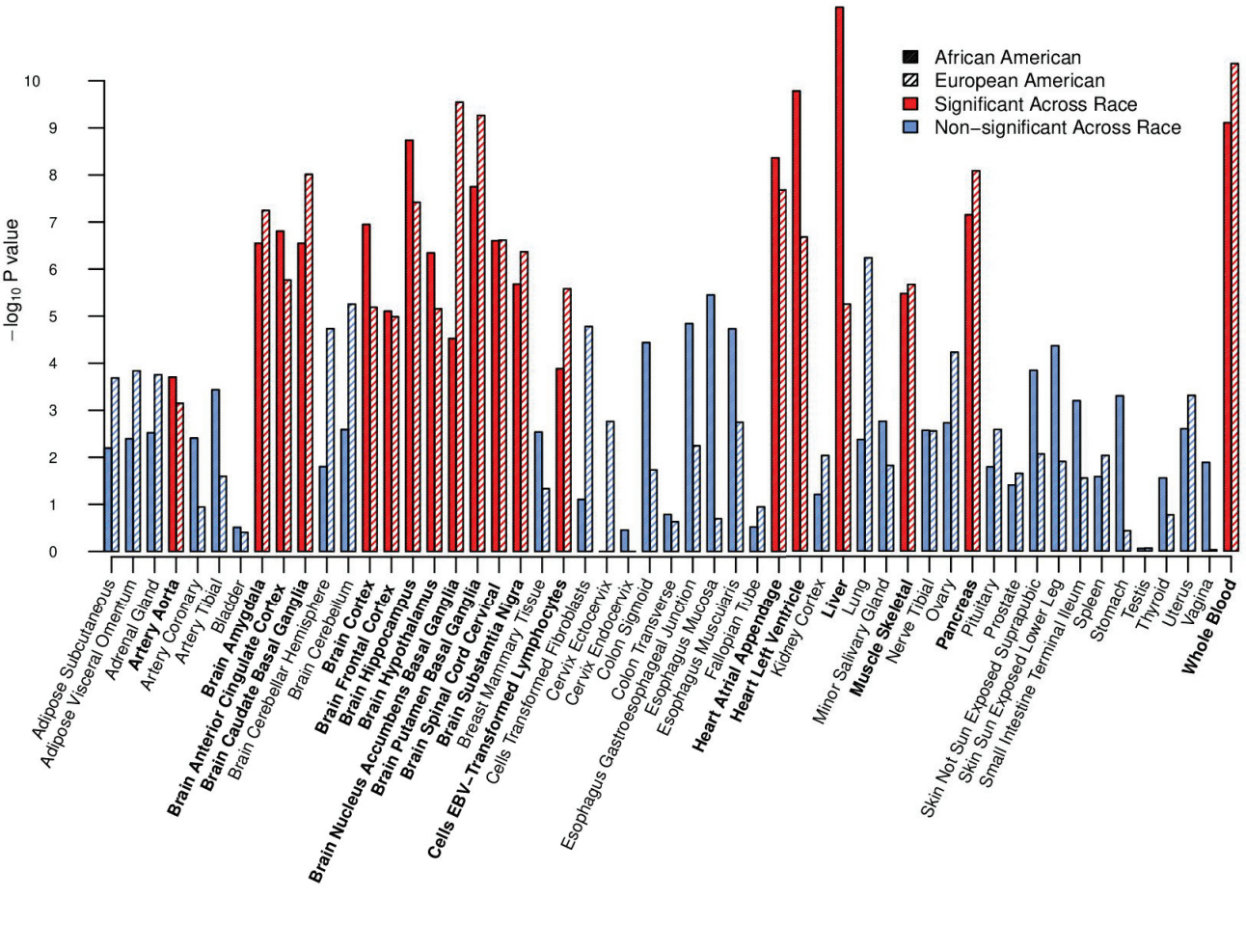
797 Zavala, A. R., Osredkar, T., Joyce, J. N., & Neisewander, J. L. (2008). Upregulation of Arc mRNA
798 Expression in the Prefrontal Cortex Following Cue0Induced Reinstatement of Extinguished Cocaine-
799 Seeking Behavior. Synapse, 62(6), 421–431.

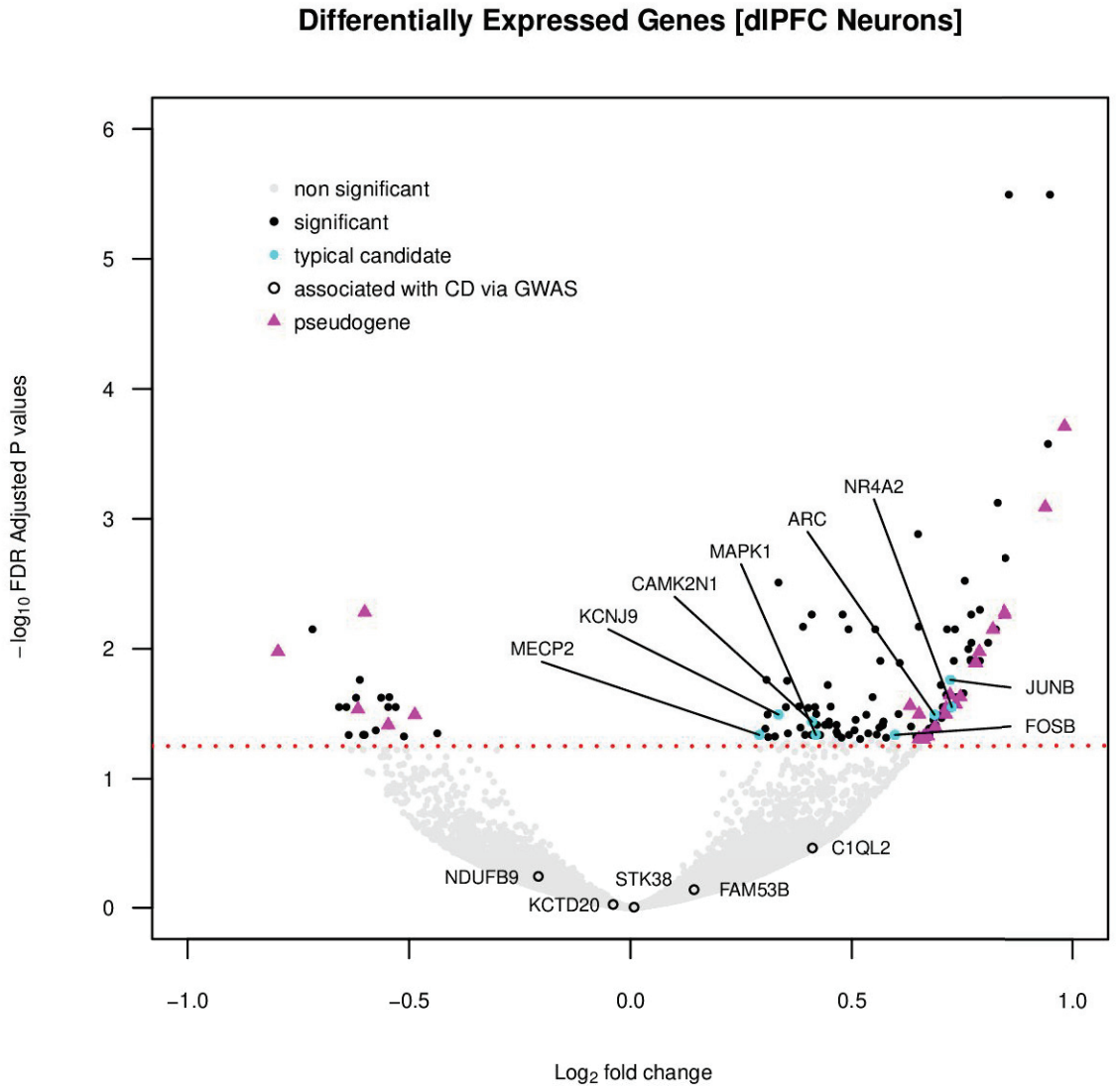
800 Zhou, Z., Yuan, Q., Mash, D. C., & Goldman, D. (2011). Substance-specific and shared transcription and
801 epigenetic changes in the human hippocampus chronically exposed to cocaine and alcohol.
802 Proceedings of the National Academy of Sciences, 108(16), 6626–6631.

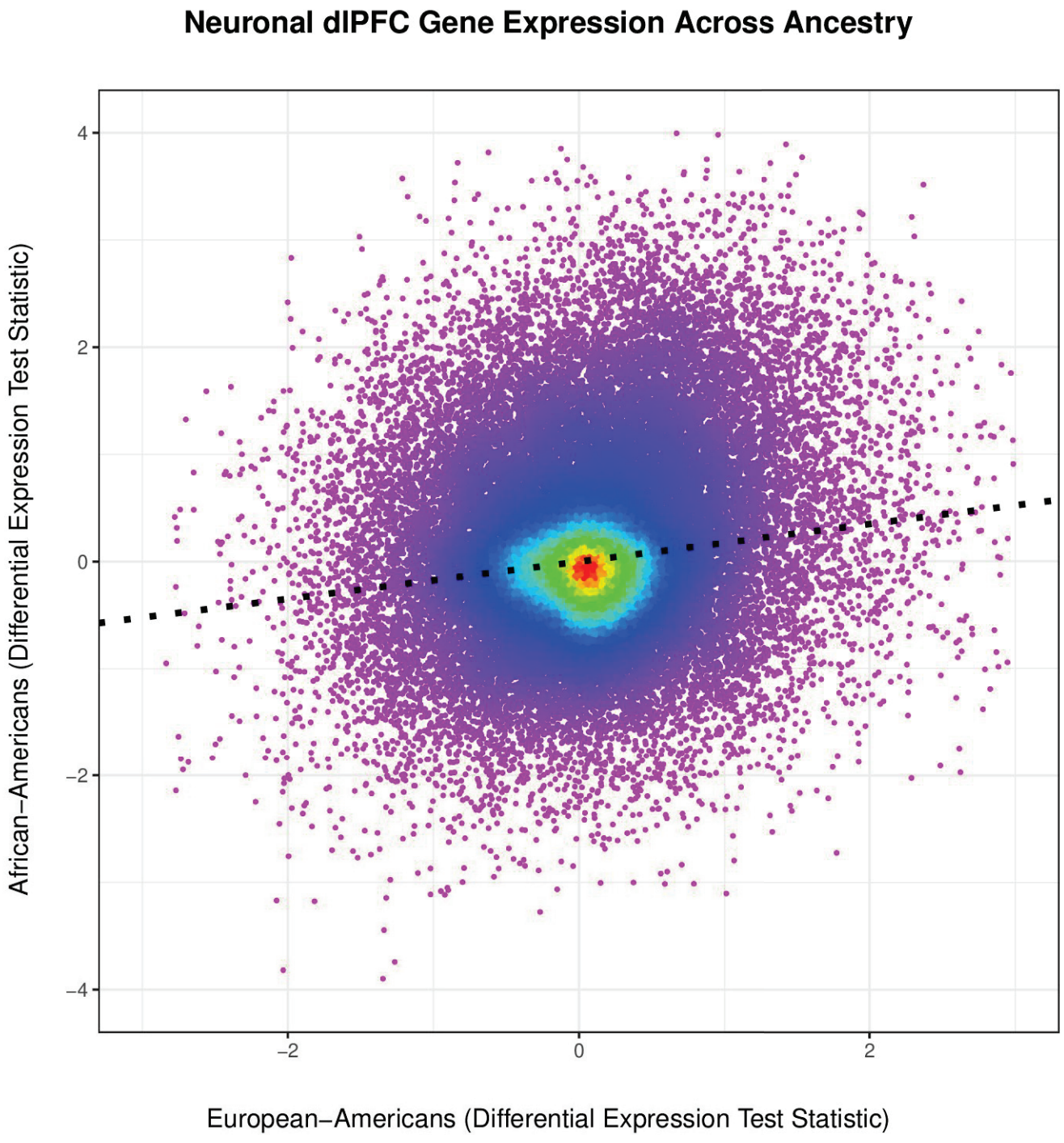


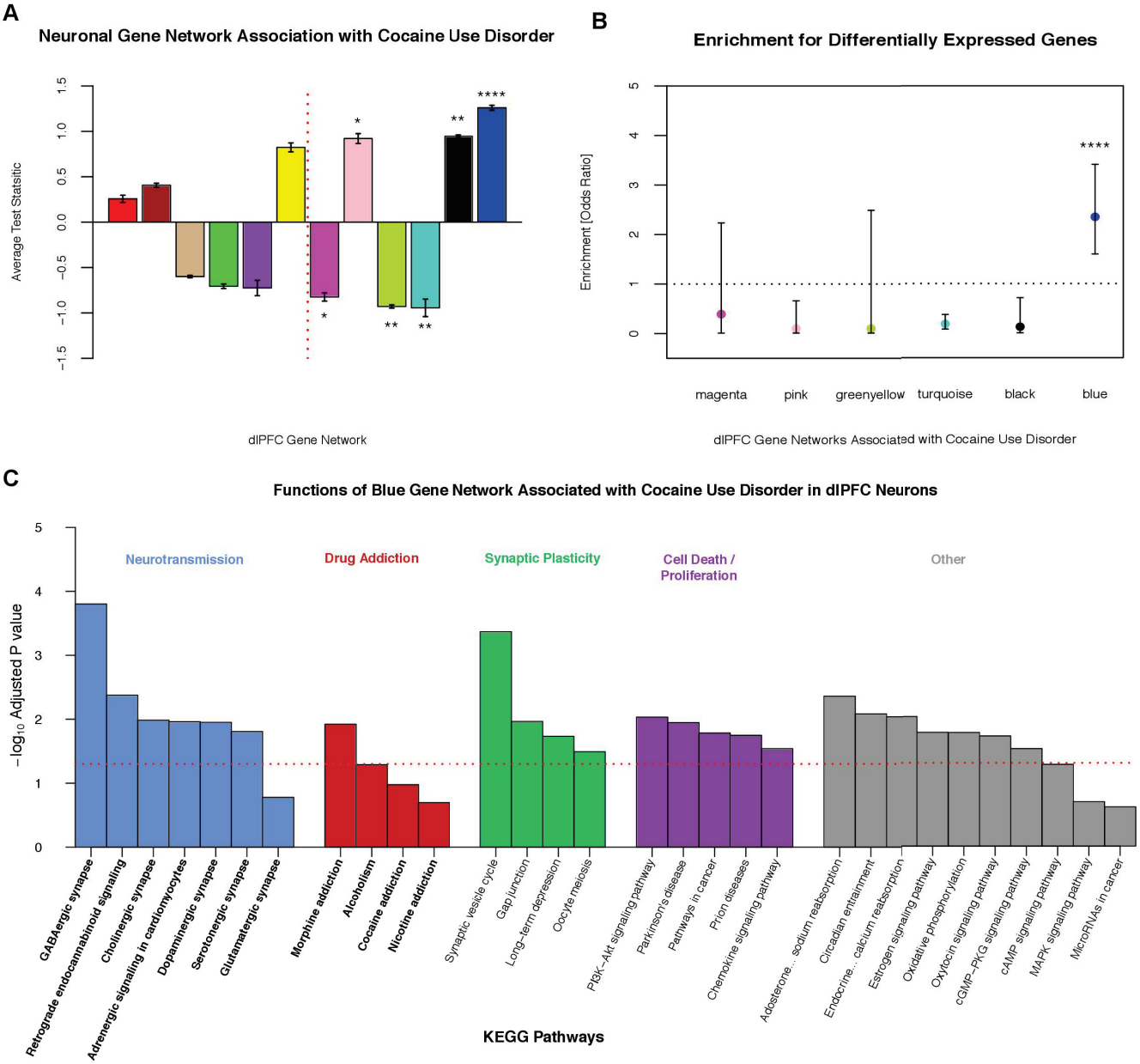
Gene–Based Tests of Cocaine Dependence: Candidate Neurotransmitter Systems











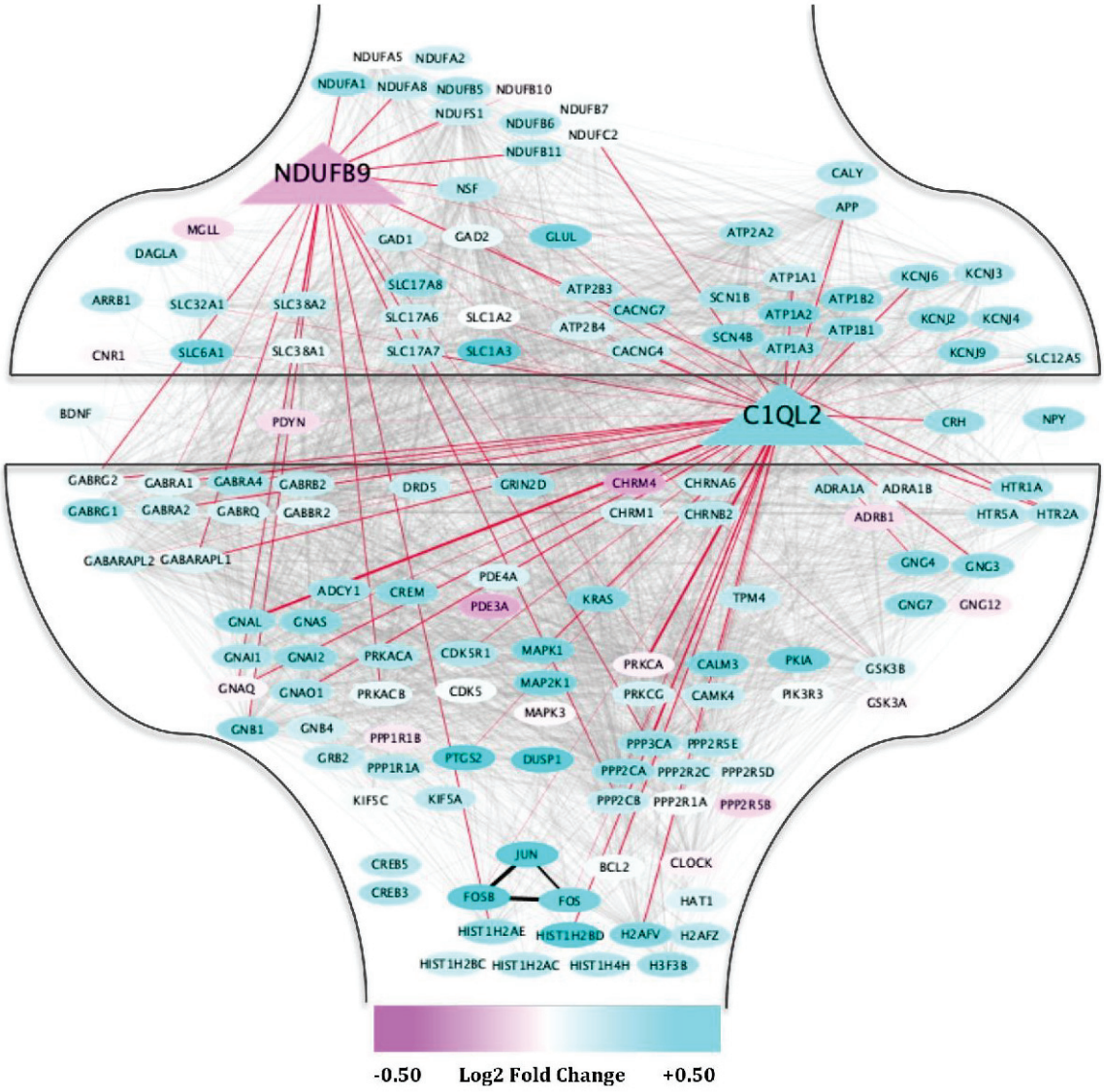


Table 1 Single Nucleotide Polymorphism (SNP) Associations with CD

SNPs Associated with the Genetic Predisposition to Cocaine Dependence by Ancestry						
Gene / Region:	*NDUFB9*		*C1qL2*		*KCTD20* & *STK38*	
Ancestry	AA	EA	AA	EA	AA	EA
#SNPs	174	205	10	74	215	174
# Parameters	51	29	4	14	38	17
Lead SNP	rs77422927		rs13020121		rs9470273	
Minor Allele	C		A		T	
Minor Allele Frequency	0.021	0.096	0.2247	0.30025	0.3192	0.21828
P_{SNP_Lead}	6.42E-06	0.963	0.902	2.22E-06	0.0253	1.42E-06
Direction of Effect	+	+	+	-	+	+
Missense SNP	rs34095749		NA		rs2239808 (KCTD20)	
Minor Allele	T		NA		C	
Minor Allele Frequency	0.013	0.050	NA	NA	0.4206	0.21439
Missense SNP	Proline_157_Serine		NA	NA	Serine_171_Threonine	
$P_{SNP\_Missense}$	0.00568	0.841	NA	NA	0.0501	1.28E-05
LD with Lead SNP (R^2)	0.4917	0.5154	NA	NA	0.609	0.9629
Direction of Effect	+	+	NA	NA	+	+

We collapsed *KCTD20* & *STK38* into a single category because they stem from the same genomic signal. AA stands for African-American and EA stands for European-American ancestry. The # of parameters represents the amount of independent signals tested within a protein coding gene and differ across ancestries due to disparate LD patterns. We estimated the linkage disequilibrium patterns of missense variants with lead SNPs using LDlink, and selecting the African-American and CEU reference panels. Note that *C1qL2* only had one missense mutation, but was not tested included in the genome-wide association study on cocaine dependence due to low minor allele frequency across ancestries (< 1%).

Table 2 "Guilt-By-Association" Analyses: Inferring Function of *NDUFB9* and *C1qL2* with CUD

Potential Functions of *NDUFB9* and *C1qL2*			
NDUFB9		*C1qL2*	
Biological Processes, Molecular Function or or KEGG Pathways	p_{adj}	Biological Processes, Molecular Function or or KEGG Pathways	p_{adj}
Neuro-degeneration / Cell Death		**Neurotransmitter Signaling**	
Phagosome acidification	0.0056	GABAergic synapse	2.4e-7
Parkinson's disease	0.0156	Serotonergic synapse	4.5e-4
Alzheimer's disease	0.0299	Cholinergic synapse	6.2e-4
Huntington's disease	0.0320	Glutamatergic synapse	0.0441
Synaptic Plasticity		**Ion Channels & Drug Addiction**	
Synaptic Vesicle Cycle	6.3e-4	Aldosterone-regulated sodium reabsorption	0.0013
Cadherin binding involved in cell-cell adhesion	0.0010	Alcoholism	0.0032
GTPase activity	0.0016	Response to drug	0.0041
GTP binding	0.0067	Nicotine addiction	0.0265
Cell to cell adhesion	0.0170	Potassium ion import	0.0428
Other Processes		**Neurodevelopment & Synaptic Plasticity**	
Oxidative Phosphorylation	2.6e-5	Small GTPase mediated signal transduction	2.2e-5
Protein Binding	7.9e-4	Positive regulation of cell proliferation	0.0177
Endocrine…regulated calcium reabsorption	0.0367	Nervous system development	0.0416

Our "guilt-by-association" approach assesses the function of genes/transcripts that are highly co-expressed with *NDUFB9* and *C1qL2* in the blue gene network associated with CUD and assesses their enrichment for biological processes, molecular functions and KEGG pathways using DAVID (Huang et al. 2009). We selected the most highly co-expressed genes with *NDUFB9* (300 genes/transcripts) and *C1qL2* (694 genes/transcripts) in the blue gene network by using an arbitrary co-expression threshold of: weighted $r > 0.05$ (raw $r > 0.78$).