An Integrative Systems-based Analysis of Substance Use: eQTL-informed Gene-based Tests, Gene Networks, and Biological Mechanisms

Zachary F Gerring[1] Ph.D., Angela Mina Vargas[1] Ph.D., Eric R Gamazon[2-4] Ph.D., Eske M Derks[1] Ph.D.

[1] Translational Neurogenomics Laboratory; QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia
[2] Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[3] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA
[4] Clare Hall, University of Cambridge, Cambridge, United Kingdom

Corresponding author:
Zachary Gerring
Translational Neurogenomics Laboratory
300 Herston Road
email: zachary.gerring@qimrberghofer.edu.au

Formatted: Numbering: Continuous

Deleted: Gene-based

Formatted: Font: 11 pt, Font color: Auto

Abstract

Genome-wide association studies have identified multiple genetic risk factors underlying susceptibility to substance use, however the functional genes and biological mechanisms remain poorly understood. The discovery and characterisation of risk genes can be facilitated by the integration of genome-wide association data and gene expression data across biologically relevant tissues and/or cell types to identify genes whose expression is altered by DNA sequence variation (expression quantitative trait loci; eQTLs). The integration of gene expression data can be extended to the study of genetic co-expression, under the biologically valid assumption that genes form co-expression networks to influence the manifestation of a disease or trait. Here, we integrate genome-wide association data with gene expression data from 13 brain tissues to identify candidate risk genes for 8 substance use phenotypes. We then test for the enrichment of candidate risk genes within tissue-specific gene co-expression networks to identify modules (or groups) of functionally related genes whose dysregulation is associated with variation in substance use. We identified six gene modules in brain that were enriched with gene-based association signals for substance use phenotypes. For example, a single module of 29 co-expressed genes was enriched with gene-based associations for smoking cessation and biological pathways involved in the immune response. Our study demonstrates the utility of eQTL and gene co-expression analysis to uncover novel biological mechanisms for substance use traits.

Deleted: gene-based

Deleted: gene-based

Introduction

Substance use is linked to hundreds of diseases and adverse societal outcomes (1). A reduction in the prevalence of substance use will therefore not only reduce the global burden of disease but reduce costs to individual sufferers, their families, and society. Substance use encompasses a range of behaviours (e.g. alcohol consumption, tobacco smoking, and cannabis use), each of which is moderately heritable with a highly polygenic background, where hundreds to thousands of genetic variants contribute to disease risk. Genome-wide association studies (GWAS) have identified hundreds of genomic regions that contain genetic risk variants (or single nucleotide polymorphisms [SNPs]) robustly associated with substance use traits, including, for example, alcohol use (2) and dependence (3) (SNP heritability [h^2_{SNP}]: 9-12%), tobacco smoking (4) (h^2_{SNP}: 1-4%, and cannabis use (h^2_{SNP}: 11%) (5). However, the functional interpretation of these regions remains largely unknown due in part to the complex local correlation structure of the genome (linkage disequilibrium) and complex interaction patterns between genes, known as the "co-localisation problem" (6), making causal gene identification challenging. Single Nucleotide Polymorphisms, or genetic variants, may affect the expression of one or more genes or a broader network of genes within a disease relevant tissue or cell type (7). We and others have linked genetic variants to changes in gene expression, known as expression quantitative trait loci (eQTL), to identify individual risk genes as well as groups of correlated genes (risk modules) for mental health (8) and substance use disorders (9). The advantage of this approach is co-expressed genes can be causal for a trait without being influenced by the same genetic variant, thereby increasing the genomic search space for higher-order biological associations. In the present study, we will extend our earlier work (9) by generating gene co-expression modules characterized by correlated levels of gene expression. We will subsequently test for the enrichment of GWAS signals of 8 substance use traits within these co-expression modules.

Different methods exist to integrate genetic and transcriptomic information with a primary distinction between studies that use single-variant approaches (i.e., evaluating the impact of a single variant on gene expression) (10) versus gene-based approaches that combine information across multiple SNPs (i.e., imputation of gene expression at a gene-based level) (11,12). Irrespective of the applied methodology, eQTL analyses are usually based on reference datasets in which genetic and transcriptomic information has been collected in disease-relevant tissues.

Deleted: is

Deleted:

Formatted: Subscript

Deleted: gene-based

Deleted: gene-based

For example, Genotype-Tissue Expression (GTEx) project (version 7) contains genotype data linked to gene expression across 53 tissues from 714 donors, including 13 brain tissues from 216 donors. GTEx and other tissue-specific eQTL datasets represents a valuable resource with which to study gene expression and its relationship with genetic variation (13).

The integration of genetic variation and tissue-specific gene expression data has been used to prioritise functional gene candidates for substance use traits in disease-relevant tissues (i.e. brain tissue). For example, a secondary analysis of a nicotine dependence GWAS found an intronic SNP that regulates the expression of *DNMT3B* in brain (14), while a similar analysis of cannabis dependence found genetic variation associated with the expression of *CHRNA2* in brain (15). In many instances, the functional gene candidate was not the gene most proximate to the associated risk variant; a GWAS of alcohol consumption, for example, identified risk variants within the gene *KLB* that affected the expression of two distantly located genes *RCF1* and *RPL9* (16). Indeed, associations in which the nearest gene is not the functional candidate is widespread in substance use traits, where some 66% of trait associated eQTLs in GTEx targeted genes other than their most proximal gene (9).

While single-eQTL approaches have improved the functional annotation of individual SNPs, more recent approaches combine eQTL information across multiple SNPs in close proximity to a gene. These methods either impute gene expression levels using a reference dataset (11,12) or incorporate eQTL information within a gene-based test (17). We recently developed a gene-based test called eMAGMA, which performs gene-level testing by combining GWAS summary statistics, tissue-specific cis-eQTL information, and reference linkage disequilibrium data (8). eMAGMA and other gene-based approaches, such as S-PrediXcan which imputes genetically-regulated gene expression levels using GWAS summary statistics, are more powerful than single-eQTL annotation (18) and may integrate tissue-specific gene expression information for the discovery of pathogenic and/or surrogate tissues. For example, a gene-based analysis of six substance use traits reported altered genetically regulated gene expression in case samples, with many candidate risk genes either unique to brain or whole blood (9). These results suggest many regulatory effects for substance use traits manifest in a subset of disease-relevant tissues such as brain, however some effects may be shared across tissues and detected in other tissues with larger eQTL reference set samples sizes, such as whole blood. By collapsing multiple SNPs to individual functionally relevant genes, these approaches also facilitate the identification of shared mechanisms underlying substance use traits; gene-based analyses of

Deleted: gene-based

Deleted: n

Deleted: for disease

Deleted: These

Deleted: gene-based

Deleted: gene-based

Deleted: gene-based

lifetime cannabis use (5) and alcohol consumption (16) both identified *CADM2* as a candidate risk gene, suggesting shared mechanisms underlying these traits.

Genetic studies suggest substance use is highly polygenic; many genes are likely to interact with one-another in complex tissue- or cell-type specific networks influence substance use risk. Gene co-expression network analysis describes the relationship between genes in terms of their pairwise correlation, where highly correlated genes may share a functional relationship (i.e. highly correlated genes are likely to be involved in the same biological process). A genetic perturbation that affects the expression of a single gene within co-expression network may therefore alter the activity of a wider set of genes. We recently applied this heuristic in a gene co-expression network analysis of major depression, and identified novel gene candidates and gene modules both associated with major depression and disease-relevant biological pathways (8).

In the present study, we aim to improve our understanding of the biological mechanisms underlying 8 substance use phenotypes by exploring associations with co-expression networks derived from human brain samples. First, to identify candidate causal genes, we will integrate GWAS summary statistics for each phenotype with eQTL information from brain tissues in GTEx using a novel gene-based method called eMAGMA (17). Second, we will explore the gene-based overlap of associations across substance use phenotypes. Finally, we will use a gene co-expression network analysis to identify modules of genes enriched with gene-based association signals, before using biological pathway analysis to characterise the substance use risk modules.

Deleted: gene-based

Deleted: gene-based

Deleted: gene-based

Methods

The Genotype-Tissue Expression (GTEx) project

Fully processed, filtered and normalised gene expression data for 13 brain tissues (Supplementary Table 1) were downloaded from the Genotype-Tissue Expression project portal (version 7) (http://www.gtexportal.org). Only genes with ten or more donors were included. Other inclusion criteria for expressed genes were expression estimates > 0.1 Reads Per Kilobase of transcript (RPKM) and an aligned read count of six or more within each tissue. Within each tissue, the distribution of RPKMs in each sample was quantile-transformed based

on the average empirical distribution observed across all samples. Expression levels for each gene in each tissue were subsequently transformed to the quantiles of the standard normal distribution.

Genome-wide association study of substance use traits

We downloaded GWAS summary statistics for 8 substance use traits (smoking age of initiation, cigarettes per day, drinks per week, smoking cessation, smoking initiation, alcohol use disorder, alcohol dependence, and lifetime cannabis use) listed in Table 1. Detailed methods, including a description of population cohorts, quality control of raw SNP genotype data, and association analyses for substance use GWAS are provided in their respective publications (2–5).

eQTL-informed gene-level analysis of substance use GWAS signals

We identified and prioritised risk genes for each substance use phenotype using eMAGMA (17) and S-PrediXcan, both of which integrate GWAS summary statistics with eQTL information from the GTEx project. eMAGMA assigns SNPs within or near target genes based on significant (FDR<0.05) SNP-gene associations in GTEx. Gene-based statistics were subsequently computed using the sum of the assigned SNP –log(10) *P* values while accounting for Linkage Disequilibrium. S-PrediXcan, on the other hand, imputes genetically-regulated gene expression from training models to estimate the phenotype-expression association, while also controlling for Linkage Disequilibrium. For both approaches, we used gene expression data for 13 brain tissues generated from GTEx (v7), and LD information from the 1000 Genomes Project Phase 3 (19). For each tissue, we corrected for multiple testing using Bonferroni correction based on the number of genes per tissue (Supplementary Table 1). Due to correlated expression across tissues, no correction for the number of phenotypes studied (N=8) was performed.

Fine-mapping of causal gene sets

S-PrediXcan and other transcriptomic approaches may yield false-positive gene-trait associations due to correlation (LD) among SNPs used to generate the eQTL weights in the predication models (20). We used fine-mapping of causal gene sets (FOCUS) to appropriately model the impact of gene-trait correlations on the S-PrediXcan expression weights and assign a causal probability to each gene within substance use risk loci (20). We built a multi-tissue

Deleted: association

Deleted: ,

Deleted: s

Deleted: to

Deleted: Gene-based

Deleted: We

Formatted: Font: Not Bold

Formatted: Font: Not Bold

Formatted: Font: Not Bold

Formatted: Font: Not Bold

Deleted:

Formatted: Font: Not Bold

Formatted: Font: Not Bold

Formatted: Font: Not Bold

eQTL database using GTEx v7 brain tissues (https://github.com/bogdanlab/focus/) to use as the eQTL weights database, and LD information from the 1000 Genomes Project Phase 3 (19) as reference genotypes.

Identification of gene expression modules

Gene co-expression modules were constructed for 13 individual brain tissues using the weighted gene co-expression network analysis (WGCNA) package in R (21). An unsigned pairwise correlation matrix – using Pearson's product moment correlation coefficient – was calculated. An appropriate "soft-thresholding" value, which emphasizes strong gene-gene correlations at the expense of weak correlations, was selected for each tissue by plotting the strength of correlation against a series (range 2 to 20) of soft threshold powers. The correlation matrix was subsequently transformed into an adjacency matrix. Matrices are characterised by nodes (corresponding to genes) and edges (corresponding to the connection strength between genes). Each adjacency matrix was normalised using a topological overlap function. Hierarchical clustering was performed using average linkage, with one minus the topological overlap matrix as the distance measure. The hierarchical cluster tree was cut into gene modules using the dynamic tree cut algorithm (22), with a minimum module size of 30 genes. We amalgamated modules if the correlation between their eigengenes – defined as the first principal component of their genes' expression values – was greater or equal to 0.8.

Gene-set analysis of gene co-expression modules

To identify gene co-expression modules enriched with substance risk genes, we performed gene-set analysis of eMAGMA gene-level results in the derived tissue-specific gene co-expression modules using the gene-set analysis function in MAGMA v1.06 (17,23). The competitive analysis tests whether the genes in a gene-set (i.e. gene co-expression module) are more highly associated with risk genes than other genes while accounting for gene size and gene density. We applied an adaptive permutation procedure (23) (N=10,000 permutations) to obtain P values corrected for multiple testing. The 1000 Genomes European reference panel (Phase 3) was used to account for Linkage Disequilibrium between SNPs. For each tissue and gene-based enrichment method, a quantile-quantile plot of observed versus expected *P* values was generated to assess inflation in the test statistic.

Characterisation of gene expression modules

Formatted: Font: Not Bold

Formatted: Font: Not Bold

Formatted: Font: Not Bold

Deleted: ¶

Deleted: gene-based

Gene expression modules enriched with substance use GWAS association signals were assessed for enrichment of biological pathways and processes using g:Profiler (https://biit.cs.ut.ee/gprofiler/) (24). Ensembl gene identifiers within substance use gene modules were used as input; we tested for the over-representation of module genes in Gene Ontology (GO) biological processes. We set the statistical domain scope (i.e. reference gene set) to "only annotated genes". The g:Profiler algorithm uses a Fisher's one-tailed test for gene pathway enrichment; the smaller the P value, the lower the probability a gene belongs to both a co-expression module and a biological term or pathway purely by chance. Multiple testing correction was done using g:SCS; this approach accounts for the correlated structure of GO terms and biological pathways, and corresponds for an experiment-wide threshold of α=0.05.

Preservation of gene co-expression networks across tissues

To examine the tissue-specificity of modular enrichments and biological pathways, we assessed the preservation (i.e. replication) of network modules across GTEx brain tissues using the "modulePreservation" R function implemented in WGCNA (25). Briefly, the module preservation approach takes as input "reference" and "test" network modules and calculates statistics for three preservation classes: i) density-based statistics, which assess the similarity of gene-gene connectivity patterns between a reference network module and a test network module; ii) separability-based statistics, which examine whether test network modules remain distinct in reference network modules; and iii) connectivity-based statistics, which are based on the similarity of connectivity patterns between genes in the reference and test networks. We report the "Zsummary" statistic as a measure of preservation. A Zsummary value greater than 10 suggests there is strong evidence a module is preserved between the reference and test network modules, while a value between 2 and 10 indicates weak to moderate preservation and a value less than 2 indicates no preservation.

Results

Study cohorts

The substance use phenotypes included in our study are presented in Table 1. The GSCAN (GWAS and Sequencing Consortium of Alcohol and Nicotine use) analysis of 5 substance use phenotypes in 1.2 million individuals contributed the largest number of significant loci (566 variants in 406 loci) for our study. All of the included studies, with the exception of alcohol

dependence from the Psychiatric Genomics Consortium, used samples derived from the UK Biobank and/or 23andMe. Over half of the significant loci across the 8 phenotypes were related to smoking initiation, which contained the largest number of samples (N=1,232,091).

Gene-based tests of association

To identify genes whose expression is influenced by genetic variation underlying disease risk, we performed eMAGMA using GWAS summary statistics and gene expression information from 13 brain tissues in GTEx v7 (Table S2). We identified 272 unique gene-based associations across all brain tissues (after Bonferroni correction for the number of genes in each tissue) (Supplementary Table 2). The number of significant genes for each phenotype was a function of GWAS sample size; 118 genes in 13 brain tissues associated with smoking initiation (GWAS N samples=1,232,091), while a single significant gene was associated with alcohol dependence (GWAS N samples = 46,568).

There was no overlap in significant eMAGMA associations across all phenotypes, and only modest overlap between phenotype pairs. For example, 27 genes were significantly associated with both alcohol use disorder and the number of drinks per week (Table 3). There was a high correlation between the number of samples for each tissue and significant gene-based associations (Pearson's r = 0.87). Cerebellum accounted for the largest number of significant associations (N associations=183) and also contained the largest number of post-mortem brain samples (N samples=154). We compared the number of significant associations from the eMAGMA analysis with previous findings from conventional MAGMA and S-PrediXcan (Supplementary Table 3). The total number of eMAGMA associations is smaller than the number of significant conventional MAGMA associations, but larger than the number of S-PrediXcan associations. Genes found by eMAGMA but not conventional MAGMA or S-PrediXcan by phenotype are shown in Supplementary Table 4.

The gene CADM2, which has been linked to behavioural undercontrol, was associated with 4 substance use phenotypes (drinks per week, alcohol use, smoking initiation, and cannabis use). Furthermore, the effect direction of CADM2 was consistent across phenotypes (Supplementary Table 5). Another four genes (*AMT*, *CHRNA2*, *GPX1*, *KANSL1*) were significant across three phenotypes (cigarettes per day, age of smoking initiation, and smoking cessation (Table 4). Overall, we found moderate correlation of eMAGMA Z-scores between phenotype pairs (Supplementary Table 6 and Supplementary Figure 1), with the strongest correlations between

Deleted: studies

Deleted: eQTL-informed gene-based

Deleted: The number of significant genes

Deleted: unique

Deleted: There was some gene-based overlap across phenotypes; f

Deleted: The brain tissue with the largest number of associations was c

Deleted: which

Deleted: There was a high correlation between the number of samples for each tissue and significant gene-based associations (Pearson's r = 0.87).

Deleted: We

Deleted: previous reports on

Deleted: gene-based test results

Deleted: findings

Deleted: Table S3

Deleted:

Deleted: However, it should be noted as a proportion of unique genes tested (eMAGMA only tests eGenes while S-PrediXcan only tests genes with reliably imputed genetically-regulated gene expression), S-PrediXcan and eMAGMA yield relatively more gene-based associations than conventional MAGMA.¶
¶

Deleted: risk taking behaviour and a care-free personality

Formatted: Not Highlight

Deleted: and therefore represents a promising candidate for future functional studies on across substance use traits

Deleted:

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Deleted:

Formatted: Not Highlight

(AUDIT) and drinks per week (Pearson's r = 0.263, P < 2.22 × 10^{-16}) and smoking initiation and drinks per week (r = 0.220, P = 3.95 × 10^{-14}).

Fine-mapping further prioritises genes within GWAS risk loci

We applied the fine-mapping of causal gene sets (FOCUS) algorithm to prioritise genes within GWAS risk loci. All of the phenotypes, with the exception of alcohol dependence, contained "credible" genes (that is, genes most likely to be causal for a given phenotype). We identified a total of 269 unique credible genes across 77 distinct loci for 7 substance use phenotypes. Smoking initiation had the largest number of loci with credible genes (N=42 loci containing 117 credible genes), followed by cigarettes per day (N=19 loci containing 46 credible genes). Candidate casual genes with the highest posterior inclusion probability (PIP) included *FPGT* (S-PrediXcan Z score -6.33; PIP: 1) for smoking initiation; *ZNF780B* (S-PrediXcan Z score 5.37; PIP 1) for smoking cessation; *RFC1* (S-PrediXcan Z score –9.41; PIP: 1) for drinks per week; *SNRPA* (S-PrediXcan Z score -9.44; PIP: 1) for cigarettes per day; *CADM2* (S-PrediXcan Z score 4.38; PIP: 0.624) for lifetime cannabis use; *GRK4* (S-PrediXcan Z score -4.7; PIP: 0.542) for age of smoking initiation; and *FAM180B* for alcohol use disorder (S-PrediXcan Z score -5.74; PIP: 0.749) . A full list of credible genes for each phenotype is provided in Supplementary Table 7. We assessed the overlap in credible genes across phenotypes. A total of 43 credible genes were prioritised in more than one phenotype (Supplementary Table 8). Interestingly, the genes *SNRPA* and *ZNF780B* had posterior probabilities close to or equal to 1 for both smoking cessation and cigarettes per day, while the S-PrediXcan Z scores for these genes had opposite effect directions. This is consistent with the inverse relationship between the phenotypes, and provides strong evidence of their involvement in substance use risk.

Network-based enrichment of substance use risk genes

We tested for the enrichment of gene-based association signals in brain tissue-dependent gene co-expression networks. Age of initiation of smoking (AOI), drinks per week (DPW), and smoking cessation (SMC) each showed enrichment of gene-based association signals within two modules. The module DPW-1 had the largest number of gene-based associations with a nominal P value < 0.05 (N=9 genes; 22.5%), followed by the module DPW-2 (N=27; 16.5%) (Supplementary Table 9). The module DPW-2 also harboured two genes—*TUFM* (nucleus

Deleted: alcohol use
Formatted: Superscript
Deleted:
Formatted: Font: 12 pt, Bold
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Justified, Line spacing: 1.5 lines
Deleted: a
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt, Italic
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted ...
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted ...
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted ...
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted ...
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted ...
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted ...
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Not Highlight
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted ...
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted ...
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: 12 pt
Deleted: gene-based
Deleted: each
Deleted: gene-based
Deleted: (Table 5)
Formatted: Font: Italic
Formatted: Font: 12 pt, Font color: Auto

accumbens basal ganglia; $P=1.07 \times 10^{-10}$) and *RPL9* (nucleus accumbens basal ganglia: $P=2.08 \times 10^{-7}$)—with significant (Bonferroni-corrected) eMAGMA associations, highlighting their potentially coordinated association with drinks per week. Furthermore, the genes *RPS26* and *SNF8* had nominally significant eMAGMA P values in the modules *DPW-2* (*RPS26*, $P=0.0316$; *SNF8*, $P=0.0006$) and AOI-2 (*RPS26*, $P=5.55 \times 10^{-5}$; *SNF8*, $P=0.0007$), suggesting some shared modular activity across substance use phenotypes (Supplementary Table 9). A biological category association analysis of the enriched modules identified processes related to RNA processing (module AOI-2; $P=5.12 \times 10^{-8}$); GABA synthesis, release, reuptake and degradation (module DPW-1; $P=1.39 \times 10^{-6}$) and the immune response (module SMC-1; $P=1.64 \times 10^{-67}$) (Table 5 and Supplementary Table 10). We extracted eMAGMA associations for genes that intersect both the enriched module and significant biological pathways (Supplementary Table 11). Several biological pathways had a relatively large proportion of nominally significant eMAGMA associations. For example, 4 out of 8 overlapping module genes for the AOI-2 pathway "metabolism of RNA" contained eMAGMA P values < 0.05 (Supplementary Table 12). These data support the involvement of the gene co-expression modules in substance use, although the overlap between eMAGMA associations and biological pathways is modest for several phenotype modules (e.g. DPW-1 "neurotransmitter transport" contains 2 genes with eMAGMA associations, one of which has a nominal P value < 0.05).

There was strong preservation (Z score > 10) of gene connectivity structure within significant modules across brain tissues (Figure 1), however DPW-2 (anterior cingulate cortex enriched with developmental and neurotransmitter pathways) had slightly lower preservation compared to other tissue modules. These data suggest modules and pathway enrichments may be generalised across tissue types for substance use traits and provide further support to maximise tissue sample size for a single brain tissue/region rather than maximising brain region coverage.

Discussion

Genetic risk factors for substance use alter the expression of target genes, which may in turn influence the activity of highly co-expressed (but not necessarily co-regulated) genes in a tissue-specific manner. We used expression quantitative trait loci from 13 brain tissues in a novel gene-based test (eMAGMA) to identify candidate risk genes for 8 substance abuse traits. The risk genes were subsequently tested for enrichment in tissue-specific gene co-expression networks to identify groups of highly correlated genes associated with substance abuse and improve the biological interpretation of gene-based associations. We identified 272 gene-based

Formatted: Font: Not Italic
Formatted: Font: Italic
Formatted: Superscript
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Deleted: enrichment
Deleted: gene expression
Deleted: age of initiation of smoking;
Deleted: 1
Formatted: Font color: Black
Formatted: Font: Italic, Font color: Black
Formatted: Font color: Black
Formatted: Font: 12 pt, Font color: Auto
Formatted: Font: Italic
Deleted: Gene expression (transcription); adjusted $P=0.0355$), cellular metabolism (e.g. smoking cessation; module SMC-2; Cellular metabolic process, $P=0.0443$) and protein transport (e.g. drinks per week DPW-1; neurotransmitter transport, $P=0.0053$).

Deleted: Section Break (Next Page)
Formatted: Space Before: Auto, Line spacing: 1.5

Deleted: gene-based

Deleted: gene-based
Deleted: 324
Deleted: gene-based

associations across 8 substance use traits, many of which were associated with multiple traits. Candidate risk genes for 3 substance use traits (age of initiation, drinks per week, and smoking cessation) were enriched in at least one co-expression module, which contained genes involved in gene expression and cellular metabolism. These results demonstrate the utility of integrating genetic, gene expression, and gene co-expression data for the biological interpretation of complex traits such as substance use.

Our gene-level (eMAGMA) approach annotates target genes by assigning genetic variants to genes based on tissue-specific eQTL information before testing for the enrichment of GWAS signals in target genes. The number of significant gene-level associations across the 8 substance use traits ranged from 1 (alcohol dependence) to 118 (smoking initiation). The number of associations was a function of GWAS sample size, highlighting the importance of sample size in genetic studies of complex traits. In a comparison of eMAGMA and other gene-based methods, eMAGMA performed similarly to S-PrediXcan in terms of number of significant associations, while it shows a 1.2 to 7-fold reduction compared to MAGMA gene-based test results (17) (Table S6). The latter finding is not unexpected since the total number of tested genes in eMAGMA (i.e., genes of which gene expression is controlled by at least one eQTL) is substantially lower than the total number of protein-coding genes (e.g. the number of tested genes in amygdala using eMAGMA is 1301 versus 18,128 tested genes using conventional MAGMA). However, while eMAGMA identifies fewer genes than its conventional MAGMA counterpart, the gene candidates are directly linked to the regulation of gene expression in a particular tissue and thereby offer a biologically meaningful substrate for follow-up analyses.

Our approach enables the study of tissue-specific gene expression changes underlying substance abuse traits. The majority of the significant associations were detected in cerebellum, a region that has been implicated in addiction (26). While a robust functional mechanism specific to cerebellum has not been established, a recent study in mice showed that the cerebellum controls the reward circuitry and social behaviour through direct projections from the deep cerebellar nuclei to the brain's reward center (i.e., the ventral tegmental area) (27). This suggests changes in gene expression in cerebellum precipitate behavioural changes related to substance use. It should be noted, however, that cerebellar gene-based associations may be proxy associations for a causal tissue or cell type, given cerebellum has the largest number of brain tissue samples in GTEx thereby increasing statistical power to identify gene associations.

Deleted: ¶

Formatted: Indent: First line: 0", Space Before: Auto, Line spacing: 1.5 lines

Deleted: gene-based

Deleted: gene-based

Deleted: 3

Deleted: Of note

Deleted: even though the total number of significant associations may be reduced compared to MAGMA,

Deleted: the value of this approach can be found in the improved functional interpretation of significant eMAGMA associations. ¶
¶

Deleted: gene-based

Deleted: ¶

Previous studies showed moderate to large correlations of additive genetic effects across substance use traits (28,29). We aimed to investigate whether the genetic correlations would be recapitulated in terms of gene-level associations. Indeed, we observed substantial overlap for some trait combinations with high genetic correlations. For example, 82% of the genes that were significantly associated with alcohol use disorder were also linked to the number of drinks per week. This is higher than the genetic correlation (r_g) between the two phenotypes (r_g=?), which may be the result of eMAGMA assigning different genetic variants underlying each phenotype to the same gene, increasing the overlap between phenotypes. However, it is difficult to compare the level of overlap in gene-level associations, which relate to specific loci, and genetic correlations, which measures genome-wide significant correlations. Interestingly, gene-level associations for lifetime cannabis use showed substantial overlap with drinks per week (32% overlap) and smoking initiation (27% overlap). One of the genes contributing to the genetic overlap is *CADM2*, which was found to be associated with 4 out of 8 traits (i.e., alcohol consumption, alcohol use disorder, smoking initiation, and cannabis use). *CADM2* was previously found to be associated with a broad profile of risk-taking behaviour and behavioural under-control (30). Furthermore, CADM2-knockout mice have increased locomotor activity and reduced body weight, suggesting an important role in behavioural regulation and energy homeostasis (31). The robust association between *CADM2* expression and multiple substance use traits highlights the need for future functional studies to further explore the functional gene mechanisms.

We also detected the susceptibility locus at a chromosome 3p21.31 gene cluster for smoking-related phenotypes: smoking initiation, cigarettes per day, and smoking cessation. The cluster covers 7 genes with eMAGMA associations (*AMT, GPX1, NCKIPSD, P4HTM, WDR6, DALRD3*, and *CCDC71*), several of which have been related to intelligence and cognitive functional measurement (32). None of the predicted expression models in our fine-mapping (FOCUS) analysis explained the observed S-PrediXcan associations for these genes, meaning a putative causal gene could not be prioritised in the locus. This is most likely due to high linkage disequilibrium at the locus. Nonetheless, these associations are consistent with the highly negative genetic correlation of smoking-related phenotypes with years of education (4). Other overlapping gene-based associations included *MAPT* and *CRHR1* for AUDIT (alcohol use disorder) and drinks per week. These genes are located within a common inversion polymorphism at chromosome 17q21.31, which is related to alterations in tissue-specific gene expression (33) and neurodegenerative disorders such as Parkinson's disease and Alzheimer's

Formatted: Space Before: Auto, Line spacing: 1.5

Deleted: AUD

Formatted: Highlight

Formatted: Not Superscript/ Subscript, Highlight

Deleted: This is in line with the high genetic correlation (r_g) between total AUDIT scores and alcohol consumption (r_g=0.96) (2).

Deleted: a risk-taking, optimistic, and care-free personality

Deleted: Our finding of

Deleted: consistent

Deleted: provides additional evidence that this gene plays a causal role and

Deleted: mechanisms

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

disease (34,35). However, a causative role of individual genes within this locus in substance use has not been established and cannot be inferred from the present data.

Our network-based approach identified gene co-expression networks enriched with GWAS signals of age of smoking initiation, alcohol consumption, and smoking cessation. The implicated modules were enriched in biological pathways related to cellular metabolism ("cellular metabolic process", nucleus accumbens basal ganglia, P=0.0443) and gene expression ("RNA processing", spinal cord cervical C-1, $P=6 \times 10^{-4}$), among others. The terms "gene expression" and "RNA processing" are difficult to interpret because they involve every processes in which a stretch of DNA is converted into a mature gene product. "Cellular metabolism", while similarly broad in biological pathways, encompasses all chemical reactions involving the breakdown of drug compounds and alcohols and would therefore be expected to be associated with substance use. Interestingly, a module enriched with risk genes associated with drinks per week (DPW-1) was associated with the biological process "GABA synthesis, release, reuptake and degradation". Alcohol directly binds to gamma-aminobutyric acid (GABA) receptors, causing the release of the inhibitory neurotransmitter GABA and inducing the sedative effects associated with alcohol use (36). Our findings represent some of the first evidence that alternations in genetically regulated expression in anterior cingulate cortex may influence alcohol consumption behaviour through changes in the brain's reward circuitry and warrant follow-up validation studies.

The findings of this study should be interpreted in view of the following limitations. First, although GTEx is one of the most comprehensive genetic expression databases available to date, the statistical power for eQTL discovery is still modest (37). We observed a strong correlation (Pearson's r = 0.87) between the post-mortem sample size and the number of gene discoveries suggesting that molecular studies of substance use phenotypes should maximise brain tissue sample. It should be noted, however, as the sample size of GTEx continues to increase the number of genes with significant eQTLs (eGenes) will plateau and further increases in sample size will have little impact on biological conclusions. Second, our analyses focus on the role of eQTLs in brain tissues while recent studies have shown that eQTL effects may differ between cell types within a specific tissue (38). Cerebellum, for example, contains the largest number of neurons in the human brain (39), potentially increasing the likelihood of identifying neuronal-specific pathways compared to other brain regions. Third, the identified genes should be seen as 'candidates' as correlated levels of gene expression in high LD

Deleted: ¶

Deleted: ¶

genomic regions makes it challenging to identify the true causal genes (40). Finally, our gene co-expression analyses rely on the stability (i.e. robustness) of gene networks both within and between tissues (8).

In summary, we assessed gene targets and biological pathways underlying 8 substance use traits. Our gene-based approach, eMAGMA, identified 272 candidate risk genes for substance use whose expression is altered in at least one of 13 brain tissues. We confirmed substantial gene-based overlap between substance use traits, with the highest overlap between drinks per well and alcohol use. The gene *CADM2*, recently associated with lifetime cannabis use, risk-taking behaviour, and a behavioural undercontrol, was associated with half of the substance based traits. We used gene co-expression networks in brain to identify broader, functionally related modules (groups) of genes potentially implicated in substance use. Six gene modules across 3 traits were enriched with gene-based associations. One of the associated co-expression modules, in anterior cingulate cortex, was enriched with biologically meaningful pathways related to GABA release and degradation, highlighting the utility of our approach in describing the molecular characteristics of substance use traits. The integration of summary statistics from larger GWAS of substance use traits with gene expression data from brain tissues, provided by GTEx (41) and other consortia (42), will facilitate the translation of statistical associations to the discovery of causal genes and molecular mechanisms.

Deleted: ¶

Deleted: gene-based

Deleted: 324

Deleted: gene-based

Deleted: carefree attitude

Deleted: gene-based

Funding:

E.R.G. is supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R35HG010718.

Disclosure:

E.R.G. receives an honorarium from Circulation Research of the AHA as member of the Editorial Board. He performed consulting for the City of Hope / Beckman Research Institute. All other authors report no conflicts of interest.

Figure legends:

Figure 1: Preservation of gene connectivity across co-expression modules enriched with gene-based association signals for substance use traits. Notes: A Z-summary value greater than 10 suggests there is strong evidence a module is preserved between the reference and test network modules, while a value between 2 and 10 indicates weak to moderate preservation and a value less than 2 indicates no preservation. Grey boxes indicate the tissue in which the significant association was found. AOI, age of smoking initiation; DPW, drinks per week; SMC, smoking cessation.

Formatted: Justified

Deleted: gene-based

Formatted: Font: (Default) Times New Roman, 12 pt

Deleted:

Formatted: Font: Not Bold

Formatted: Font: (Default) Times New Roman, 12 pt

References

1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet [Internet]. 2018 Nov 10;392(10159):1789–858. Available from: https://doi.org/10.1016/S0140-6736(18)32279-7

2. Sanchez-Roige S, Palmer AA, Fontanillas P, Elson SL, Adams MJ, Howard DM, et al. Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. Am J Psychiatry. 2018 Oct 19;176(2):107–18.

3. Walters RK, Polimanti R, Johnson EC, McClintick JN, Adams MJ, Adkins AE, et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. Nat Neurosci. 2018;21(12):1656–69.

4. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet. 2019;51(2):237–44.

5. Pasman JA, Verweij KJH, Gerring Z, Stringer S, Sanchez-Roige S, Treur JL, et al. GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia. Nat Neurosci. 2018;21(9):1161–70.

6. Gamazon ER, Zwinderman AH, Cox NJ, Denys D, Derks EM. Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. Nat Genet. 2019;51(6):933–40.

7. Gamazon ER, Segrè A V, van de Bunt M, Wen X, Xi HS, Hormozdiari F, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet. 2018;50(7):956–67.

8. Gerring ZF, Gamazon ER, Derks EM. A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. PLOS Genet. 2019 Jul 15;15(7):e1008245.

9. Marees AT, Gamazon ER, Gerring Z, Vorspan F, Fingal J, den Brink W van, et al. Post-GWAS analysis of six substance use traits improves the identification and functional interpretation of genetic risk loci: Full list of International Cannabis Consortium members. Drug Alcohol Depend. 2019;107703.

10. Hormozdiari F, van de Bunt M, Segrè A V., Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;99(6):1245–60.

11. Gamazon ER, Wheeler HE, Shah KP, Mozaffari S V, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–8.

12. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245–52.

13. eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. 2017 Oct 11;

14. Hancock DB, Guo Y, Reginsson GW, Gaddis NC, Lutz SM, Sherva R, et al. Genome-wide association study across European and African American ancestries identifies a SNP in DNMT3B contributing to nicotine dependence. Mol Psychiatry [Internet]. 2017/10/03. 2018 Sep;23(9):1911–9. Available from: https://pubmed.ncbi.nlm.nih.gov/28972577

15. Demontis D, Rajagopal VM, Thorgeirsson TE, Als TD, Grove J, Leppälä K, et al. Genome-wide association study implicates CHRNA2 in cannabis use disorder. Nat Neurosci [Internet]. 2019;22(7):1066–74. Available from: https://doi.org/10.1038/s41593-019-0416-1

16. Clarke T-K, Adams MJ, Davies G, Howard DM, Hall LS, Padmanabhan S, et al. Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). Mol Psychiatry [Internet]. 2017;22(10):1376–84. Available from: https://doi.org/10.1038/mp.2017.153

17. Gerring ZF, Mina-Vargas A, Derks EM. eMAGMA: An eQTL-informed method to identify risk genes using genome-wide association study summary statistics. bioRxiv. 2019 Jan 1;854315.

18. Fryett JJ, Inshaw J, Morris AP, Cordell HJ. Comparison of methods for transcriptome imputation through application to two common complex diseases. Eur J Hum Genet. 2018;26(11):1658–67.

19. Delaneau O, Marchini J, Consortium T 1000 GP. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun. 2014 Jun 13;5:3934.

20. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, et al. Probabilistic fine-mapping of transcriptome-wide association studies. Nat Genet. 2019;51(4):675–82.

21. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):1–13.

22. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. Bioinformatics. 2008;24(5):719–20.

23. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Comput Biol. 2015 Apr 17;11(4):e1004219.

24. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;44(Web Server issue):W83–9.

25. Langfelder P, Luo R, Oldham MC, Horvath S. Is My Network Module Preserved and Reproducible? PLoS Comput Biol. 2011;7(1):e1001057.

26. Moulton EA, Elman I, Becerra LR, Goldstein RZ, Borsook D. The cerebellum and addiction: insights gained from neuroimaging research. Addict Biol [Internet]. 2014 May;19(3):317–31. Available from: https://pubmed.ncbi.nlm.nih.gov/24851284

27. Carta I, Chen CH, Schott AL, Dorizan S, Khodakhah K. Cerebellar modulation of the reward circuitry and social behavior. Science (80-) [Internet]. 2019 Jan 18;363(6424):eaav0581. Available from: http://science.sciencemag.org/content/363/6424/eaav0581.abstract

28. Vink JM, Schellekens A. Relating addiction and psychiatric disorders. Sills J, editor. Science (80-) [Internet]. 2018 Sep 28;361(6409):1323 LP – 1324. Available from: http://science.sciencemag.org/content/361/6409/1323.abstract

29. Nivard MG, Verweij KJH, Minică CC, Treur JL, Derks EM, Stringer S, et al. Connecting the dots, genome-wide association studies in substance use. Mol Psychiatry [Internet]. 2016;21(6):733–5. Available from: https://doi.org/10.1038/mp.2016.14

30. Boutwell B, Hinds D, Team 23andMe Research, Tielbeek J, Ong KK, Day FR, et al. Replication and characterization of CADM2 and MSRA genes on human behavior. Heliyon [Internet]. 2017 Jul 26;3(7):e00349–e00349. Available from: https://pubmed.ncbi.nlm.nih.gov/28795158

31. Yan X, Wang Z, Schmidt V, Gauert A, Willnow TE, Heinig M, et al. Cadm2 regulates body weight and energy homeostasis in mice. Mol Metab [Internet]. 2018;8:180–8. Available from: http://www.sciencedirect.com/science/article/pii/S221287781730772X

32. Hill WD, Marioni RE, Maghzian O, Ritchie SJ, Hagenaars SP, McIntosh AM, et al. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. Mol Psychiatry [Internet]. 2019;24(2):169–81. Available from: https://doi.org/10.1038/s41380-017-0001-5

33. de Jong S, Chepelev I, Janson E, Strengman E, van den Berg LH, Veldink JH, et al. Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. BMC Genomics [Internet]. 2012;13(1):458. Available from: https://doi.org/10.1186/1471-2164-13-458

34. Skipper L, Wilkes K, Toft M, Baker M, Lincoln S, Hulihan M, et al. Linkage disequilibrium and association of MAPT H1 in Parkinson disease. Am J Hum Genet [Internet]. 2004/08/03. 2004 Oct;75(4):669–77. Available from: https://pubmed.ncbi.nlm.nih.gov/15297935

35. Myers AJ, Kaleem M, Marlowe L, Pittman AM, Lees AJ, Fung HC, et al. The H1c haplotype at the MAPT locus is associated with Alzheimer's disease. Hum Mol Genet. 2005 Aug;14(16):2399–404.

36. Banerjee N. Neurotransmitters in alcoholism: A review of neurobiological and genetic studies. Indian J Hum Genet [Internet]. 2014 Jan;20(1):20–31. Available from: https://pubmed.ncbi.nlm.nih.gov/24959010

37. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-) [Internet]. 2015 May 8;348(6235):648 LP – 660. Available from: http://science.sciencemag.org/content/348/6235/648.abstract

38. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L, et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet [Internet]. 2018;50(4):493–7. Available from: https://doi.org/10.1038/s41588-018-0089-9

39. Van Essen DC, Donahue CJ, Glasser MF. Development and Evolution of Cerebral and Cerebellar Cortex. Brain Behav Evol [Internet]. 2018;158–69. Available from: https://www.karger.com/DOI/10.1159/000489943

40. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51(4):592–9.

41. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv [Internet]. 2019 Jan 1;787903. Available from: http://biorxiv.org/content/early/2019/10/03/787903.abstract

42. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science (80-). 2018 Dec 14;362(6420):eaat8464.