Full length article

Post-GWAS analysis of six substance use traits improves the identification and functional interpretation of genetic risk loci

Andries T. Mareesa,b,c,*, Eric R. Gamazond,e, Zachary Gerringb, Florence Vorspanf,g, Josh Fingala, Wim van den Brinka, Dirk J.A. Smia, Karin J.H. Verweijh,a, Henry R. Kranzleri, Richard Shervaj, Lindsay Farrerj, International Cannabis Consortium1, Joel Gelernterk, Eske M. Derksa,b

a Department of Psychiatry, Amsterdam UMC, Amsterdam Neuroscience, University of Amsterdam, Meibergdreef 9, Amsterdam, the Netherlands
b QIMR Berghofer, Translational Neurogenomics group, Brisbane, Australia
c Department of Economics, School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
d Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, United States
e Clare Hall, University of Cambridge, Cambridge, CB3 9AL, United Kingdom
f Assistance Publique – Hôpitaux de Paris, Hôpital Fernand Widal, Département de Psychiatrie et de Médecine Addictologique, 200 rue du Faubourg Saint Denis, 75010 Paris, France
g Inserm umr-s 1144, Université Paris Descartes, Université Paris Diderot, 4 avenue de l’Observatoire, 75006 Paris, France
h Behavioural Science Institute, Radboud University, Montessorilaan 3, 6525 HR Nijmegen, the Netherlands
i Center for Studies of Addiction, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine and Crescenz VAMC, Philadelphia, PA 19104, United States
j Section of Biomedical Genetics, Department of Medicine, Boston University School of Medicine, Boston, MA, United States
k Department of Psychiatry, Genetics, and Neuroscience, Yale University School of Medicine, New Haven, CT, United States

ARTICLE INFO

Keywords:
Addiction
eQTLs
Functional annotation
GTEx
Substance use
S-PrediXcan

ABSTRACT

Background: Little is known about the functional mechanisms through which genetic loci associated with substance use traits ascertain their effect. This study aims to identify and functionally annotate loci associated with substance use traits based on their role in genetic regulation of gene expression.

Methods: We evaluated expression Quantitative Trait Loci (eQTLs) from 13 brain regions and whole blood of the Genotype-Tissue Expression (GTEx) database, and from whole blood of the Depression Genes and Networks (DGN) database. The role of single eQTLs was examined for six substance use traits: alcohol consumption (N=537,349), cigarettes per day (CPD; N=263,954), former vs. current smoker (N=312,821), age of smoking initiation (N=262,990), ever smoker (N=632,802), and cocaine dependence (N=4,769). Subsequently, we conducted a gene level analysis of gene expression on these substance use traits using S-PrediXcan.

Results: Using an FDR-adjusted p-value<0.05 we found 2,976 novel candidate genetic loci for substance use traits, and identified genes and tissues through which these loci potentially exert their effects. Using S-PrediXcan, we identified significantly associated genes for all substance use traits.

Discussion: Annotating genes based on transcriptomic regulation improves the identification and functional characterization of candidate loci and genes for substance use traits.

1. Introduction

In recent years, large-scale genome-wide association studies (GWAS) of substance use traits (i.e., substance use disorders and quantitative measures of substance use) have been conducted (Gelernter et al., 2014; Liu et al., 2019). These GWAS revealed multiple genome-wide significant loci (p < 5.0 -10^-8). However, the GWAS approach faces two major challenges. First, the effects of individual single nucleotide polymorphisms (SNPs) are generally small, and a typical GWAS in which single-variant tests of association are performed is generally underpowered to detect trait-associated SNPs with small effect sizes. One possible solution to this problem is to increase sample
sizes by combining samples in GWAS meta-analyses (Evangelou and Ioannidis, 2013), but even in these studies, the identified loci explain a modest proportion of the trait variance in substance use traits (Liu et al., 2019; Pasman et al., 2018). A second limitation of the GWAS approach is that the functional relevance of identified SNPs remains unclear. The majority (~93%) of trait-associated SNPs are located in non-coding regions of the genome (Maurano et al., 2012) suggesting that these SNPs act through the regulation of gene expression rather than by altering the protein product. Furthermore, due to extensive linkage disequilibrium (LD) in the genome, GWAS alone are unable to distinguish causal variants from correlated non-functional variants within an LD block. Constrained by these challenges, GWAS alone have been largely unsuccessful in elucidating the biological mechanisms involved in most substance use traits.

To facilitate the identification of biological mechanisms underlying substance use traits, it is essential to study the genetic regulation of gene expression in relevant tissues, i.e., human brain tissue from specific brain regions. The relevance of the brain for understanding the etiology of substance use (and disorders) is supported by numerous gene expression studies, which aimed to explore the role of epigenetic regulation and the transcriptional machinery in the addicted brain. In their review of these studies, Zhou et al. (Zhou et al., 2014) summarized the findings of post-mortem human studies, including those that report differences in gene expression in various brain regions between cases with substance use disorder vs. non-addicted controls. Overall, the reviewed studies suggest a prominent role for mechanisms involved in transcriptomic regulation (Albertson et al., 2004, 2006; Celentano et al., 2009; Zhou et al., 2011).

However, in general differential expression analysis between cases and controls does not reveal the direction of causality, i.e., it does not answer the question whether altered gene expression in a brain region represents susceptibility to substance dependence or whether excessive substance use is responsible for the altered gene expression. Because a large proportion of disease-associated variants exert their effects by regulating gene expression (Maurano et al., 2012), we applied an integrative approach that combines transcriptome data with summary statistics from recent GWAS to explore the role of regulatory genetic variants in substance use traits. Genetic variants that are associated with messenger RNA (mRNA) expression levels of one or more genes, in one or more tissues, are known as “expression quantitative trait loci” (eQTLs). The genes that are under the influence of at least one eQTL, in one or more tissues are called eGenes. Two types of eQTLs have been identified: cis-eQTLs, influencing expression levels of genes on the same locus (located at ±1 MB from the gene), and trans-eQTLs that have their effect on genes at a different locus (e.g., those on a different chromosome). Because the statistical power to detect trans-eQTLs is low at current sample sizes, we focused only on cis-eQTLs. We determined the eQTL status of 13 human brain tissues and whole blood using the Genotype-Tissue Expression (GTEx) project, the most comprehensive eQTL database available to date, in terms of the diversity of tissues included (Consortium, 2018). GTEx (V7) provides samples from 53 different tissues, including 13 from the brain, obtained from post-mortem adult subjects. In addition, we performed the same procedure for whole blood eQTLs from the Depression Genes and Networks (DGN) database (Battle et al., 2014).

In the current study, we investigated associations between cis-eQTLs and substance use traits, by filtering GWAS variants on eQTL status for specific brain regions and whole blood. This approach has several benefits. First, it reduces the multiple-testing burden by focusing specifically on variants that are involved in genetic regulation, thereby facilitating the identification of novel candidate variants which might be especially beneficial for relatively small GWAS studies. Second, functional interpretation of eQTLs is relatively straightforward because an eQTL variant tags a causal regulatory variant of an eGene. Therefore, the eGene targeted by the eQTL is a sensible target for future follow-up studies. Third, by investigating whether eQTLs in particular tissues (e.g., the brain or specific regions within the brain) are significant, we obtain information on which (brain) tissues are involved in substance use traits at a genetic level. Fourth, this method provides causal associations between gene-expression and substance use, because eQTL-gene expression associations have been evaluated in healthy individuals, and are thus not influenced by substance abuse. In addition to this single variant approach, we conducted an integrative analysis to investigate expression on a gene level using S-Predixcan (Barbeira et al., 2018), which can provide further mechanistic insights into the substance use traits under investigation.

We analyzed eQTLs in 13 brain regions (i.e., amygdala, anterior cingulate cortex, caudate, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, putamen, nucleus accumbens, substantia nigra, and cervical spine) and in whole blood as a comparison tissue. The role of eQTLs was examined for six substance use traits. We used summary statistics of the discovery samples obtained from major GWAS for these traits (Gelernter et al., 2014; Liu et al., 2019). The substance use traits examined in the current study were included if sufficiently large GWAS (i.e., >4,000 subjects) are available. The aims of this study were: i) to identify novel genetic loci associated with these substance use traits; ii) to improve the functional characterization of novel and known genetic loci; and iii) to obtain information about the mediating effects of gene expression levels on substance use traits by using a gene-level association approach.

2. Materials and methods

2.1. GWAS summary statistics

We obtained summary statistics from previously conducted GWAS meta-analyses, all of which were based on large samples; alcohol consumption (N = 537,349), CPD (N = 263,954), former vs. current smoker (N = 312,821), age of smoking initiation (N = 262,990), ever smoker (N = 632,802), and cocaine dependence (N = 4,769). Information on subjects, sample preparation, and analytic methods can be found in the original articles of the corresponding GWAS (Gelernter et al., 2014; Liu et al., 2019). For some samples, genome-wide data was only provided for the discovery samples, explaining the difference in the number of subjects between the current study and the samples described in the original papers. Detailed information about the samples is provided in Supplementary Table 1.

2.2. eQTL data

Details concerning the GTEx data used for this study are described elsewhere (Aguet et al., 2016; Ardlie et al., 2015). To filter and annotate SNPs of the GWAS summary statistics with eQTL information, we downloaded significant SNP-gene associations (FDR-adjusted p-value < 0.05; i.e., eQTL-eGene associations) for 13 brain tissues: amygdala, anterior cingulate cortex, caudate, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, nucleus accumbens, putamen, spine (cervical), substantia nigra, and whole blood from the GTEx Portal V7 (http://www.gtexportal.org/home/GTEXAnalysis_v7_eQTL.tar). The sample sizes varied across the 13 brain tissues (N ranges from 80 to 154); for whole blood the SNP-gene associations were obtained from 369 individuals (Aguet et al., 2016). Detailed information about the GTEx sample size per tissue and the number of significant eQTLs and eGenes per tissue can be found in Supplementary Table 2. DGN (N = 922) (Battle et al., 2014) data consists of a single tissue: whole blood. This data was used as an independent sample to conduct additional eQTL informed analyses. For the DGN database we used the same method as for GTEx database described above.
2.3. Statistical analyses

For the single variant analyses we extracted those SNPs that are eQTLs from the summary statistics of all GWAS. This extraction was performed for each tissue. We applied the Benjamini-Hochberg False Discovery Rate (FDR) approach (Benjamini and Hochberg, 1995), which controls the expected proportion of false positives among all signals with a FDR value below a fixed threshold, to determine significance. Using the GTEx database, we applied FDR on all 14 tissues combined for each of the six substance use traits, which provided us with six sets of eQTLs with an FDR value. The single tissue DGN data set was analysed separately, also using FDR. Subsequently, we determined significant trait associations within those eQTL-sets using a FDR threshold of $q=0.05$.

Significant eGenes were identified by applying the FDR 0.05 threshold on the eQTLs, and linking the eQTLs with significant trait associations to the targeted eGenes. Since a single genetic locus can include multiple eQTLs that are in high LD, we clumped eQTLs with significant trait associations per tissue using PLINK2 (Chang et al., 2015). As clumping cut-offs, we used an $R^2$ of $>0.1$ and a physical distance of 1000 kb, which generated a list of “index” eQTLs (i.e., independent eQTLs), which we considered as distinct loci. Due to the limited power all identified index eQTL, eGenes, and tissues should be considered as candidate findings. Statistical analyses for the single variants were performed using the open-source programming language R (https://www.r-project.org/). For visualization, we used the R-library “ggplot2”. To generate a heatmap, a matrix was generated based on the R squared of the number of index eQTLs that met the trait-association threshold of an FDR-adjusted $p$-value $<0.05$. To investigate gene expression levels we used S-PrediXcan (Barbeira et al., 2018), which integrated eQTL information from summary statistics of the substance use GWAS in an aggregated manner. S-PrediXcan estimates gene expression weights by training a linear prediction model in samples with both gene expression and SNP genotype data. Subsequently these weights are used to predict gene expression from GWAS summary statistics, while incorporating the variance and co-variance of SNPs from an LD reference panel. The current study used expression weights for the 14 tissues central in this study from the GTEx Project (V7) and whole blood from the DGN cohort (Battle et al., 2014; Gamazon et al., 2018), and LD information from the 1000 Genome Project Phase 3 (Delaneau et al., 2014). These data were processed with beta values and standard errors from substance use summary statistics to estimate the expression-GWAS association statistic. A Bonferroni correction was used to determine the transcriptome-wide significant threshold (adjusting for all tissues and genes per trait).

3. Results

The GWAS data sets of the six substance use traits used in this study showed different levels of genetic signal when plotting all SNPs (Supplementary Fig. 1). Alcohol consumption, CPD, ever smoker, age of smoking initiation, and former vs. current smoker showed significant trait associations (FDR-adjusted $p$-value $<0.05$). In contrast, cocaine dependence did not reveal any trait associated SNPs in the full GWAS. These results are largely in line with the original GWAS reports, and serve to compare the full (i.e., ‘uninformed’) GWAS against the eQTL informed GWAS analyses central in the current study.

After extracting GTEx eQTLs from the full GWAS summary statistics, which reduced the multiple testing burden and focused on SNPs with strong prior functional support, significant trait associations (FDR-adjusted $p$-value $<0.05$) were observed. Specifically, this was observed in all tissues, for the following substance use traits: alcohol consumption, CPD, ever smoker, age of smoking initiation, and former vs. current smoker (Fig. 1, Table 1, and Supplementary Table 3), but no significant trait associations were found for any tissue for cocaine dependence. Using this methodology, we identified 2,976 (GTEx) and 811 (DGN) novel loci (i.e., loci not identified in the uninformed GWAS of the six substance use traits; FDR-adjusted $p$-value $<0.05$; Table 2). Moreover, these eQTLs show evidence of exerting their effects in the various tissues through numerous eGenes (Table 1, Supplementary Tables 3 and 4). Consistent over all substance use traits, we observed that eGenes targeted by index eQTLs were generally different from the nearest gene (Table 1 and Supplementary Table 3). In fact, in 66.3 % of the cases the nearest gene was not the eGene in the GTEx informed analyses (i.e., analyses of GWAS summary statistics focusing only on GTEx eQTLs). This suggests that proximity is a limited measure in determining the functional relevance of a gene.

For alcohol consumption, we found eQTLs with significant trait associations in all 14 GTEx tissues (Fig. 1), and identified considerably more significant eGenes than for the other substance use traits (Supplementary Table 5). In total, we identified 949 novel index eQTLs for alcohol consumption, i.e. these SNPs were not significant in the uninformed GWAS.

For CPD, we found significant trait associations in all 14 GTEx tissues. We identified 509 index eQTLs which were not identified in the uninformed GWAS for CPD (Table 2). Many of the strongest associations were found in brain, and were located within the gene cluster: IREB2-CHRNA3-CHRNA5-CHRN4-HYKK-PSMA4 (Table 1, Supplementary Table 3). For the other smoking traits (i.e., former vs. current smoker, age of smoking initiation, ever vs. never smoker) we identified 117, 66, and 1,335 novel loci, respectively (Table 2). For these traits the IREB2-CHRNA3-CHRNA5-CHRN4-HYKK-PSMA4 gene cluster showed considerably weaker associations (Table 1, Supplementary Table 3). Furthermore, the GTEx eQTL informed GWAS for cocaine dependence did not reveal any significant trait associations, similar to the uninformed GWAS.

The single variant eQTL analyses informed by the DGN data, showed 224 significant index eQTLs for alcohol consumption, 136 for CPD, 401 for ever smoker, 25 for age of smoking initiation, and 35 for former vs. current smoker, which were not found by the uninformed GWAS (Table 2 and Supplementary Table 5). While the genes most proximal to significant index eQTL based on DGN were occasionally similar as found in the GTEx analyses, they often targeted different
comparing the S-PrediXcan results informed by GTEx and DGN, we were either unique to brain or to whole blood (Table 4). Moreover, analyses showed evidence of tissue specificity, as many of our findings under investigation (Table 3). The identified genes of the S-PrediXcan ever smoker, we found differentially expressed genes in all tissues (Supplementary Table 6). For alcohol consumption, CPD, and smoking initiation, ever smoker, and former vs. current smoker (Tables 3, 4 and Supplementary Table 7). The overarching aim of this study was to explore whether functionally annotating SNPs using information on their role in the regulation of gene expression facilitates the identification and functional interpretation of candidate loci involved in substance use traits. Furthermore, we aimed to identify genes with differentially genetically regulated levels of gene expression associated with substance use traits. Using the GTEx database we explored the role of regulatory genetic variants in 14 different tissues and tested association with six substance use traits. In addition, we conducted independent eQTL analyses using the DGN database. We identified index eQTLs with significant trait associations (FDR-adjusted p-value < 0.05) for five of the six traits examined: alcohol consumption, CPD ever smoker, age of onset of smoking, former vs. current smoker. For cocaine dependence, no GTEx eQTLs with significant trait associations were found. Compared to the full (i.e., eQTL-informed) GWAS results, functional annotation of eQTLs improved the power to detect significant trait associations. Overall, informed by GTEx, this method allowed us to identify 2,976 novel index eQTLs, we observed many findings to be unique to a specific reference panel (Supplementary Table 7).

4. Discussion

The overarching aim of this study was to explore whether functionally annotating SNPs using information on their role in the regulation of gene expression facilitates the identification and functional interpretation of candidate loci involved in substance use traits. Furthermore, we aimed to identify genes with differentially genetically regulated levels of gene expression associated with substance use traits. Using the GTEx database we explored the role of regulatory genetic variants in 14 different tissues and tested association with six substance use traits. In addition, we conducted independent eQTL analyses using the DGN database. We identified index eQTLs with significant trait associations (FDR-adjusted p-value < 0.05) for five of the six traits examined: alcohol consumption, CPD ever smoker, age of onset of smoking, former vs. current smoker. For cocaine dependence, no GTEx eQTLs with significant trait associations were found. Compared to the full (i.e., eQTL-informed) GWAS results, functional annotation of eQTLs improved the power to detect significant trait associations. Overall, informed by GTEx, this method allowed us to identify 2,976 novel index eQTLs, we interpret these index eQTLs as candidate loci for substance use traits. These candidate loci were not previously detected by the uninformed GWAS of the six substance use traits. EGenes. Comparing DGN to GTEx whole blood, different eGenes were targeted by the same eQTL in 65.6% of the cases.

The S-PrediXcan analyses, informed by the GTEx eQTLs, revealed large numbers of significantly associated genes (corrected for the number of genes and tissues) for alcohol consumption, CPD, age of smoking initiation, ever smoker, and former vs. current smoker (Tables 3, 4 and Supplementary Table 6). For alcohol consumption, CPD, and ever smoker, we found differentially expressed genes in all tissues under investigation (Table 3). The identified genes of the S-PrediXcan analyses showed evidence of tissue specificity, as many of our findings were either unique to brain or to whole blood (Table 4). Moreover, comparing the S-PrediXcan results informed by GTEx and DGN, we observed many findings to be unique to a specific reference panel (Supplementary Table 7).

### Table 1

**Overview of the five strongest eQTL-eGene relations per substance use trait, identified by the current study.**

<table>
<thead>
<tr>
<th>Trait</th>
<th>N unique novel loci</th>
<th>GTEX informed</th>
<th>DGN informed</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPD</td>
<td>509</td>
<td>214</td>
<td></td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>949</td>
<td>136</td>
<td></td>
</tr>
<tr>
<td>Ever smoker</td>
<td>1335</td>
<td>401</td>
<td></td>
</tr>
<tr>
<td>Age of initiation</td>
<td>66</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Former vs. current smoker</td>
<td>117</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Cocaine dependence</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2976</td>
<td>811</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2

**Overview of novel significant index eQTLs which were not identified by the uninformed GWAS (FDR adj. p-value < 0.05).**

<table>
<thead>
<tr>
<th>Tissue</th>
<th>PVALUE</th>
<th>FDR</th>
<th>GWAS effect</th>
<th>GTEx slope</th>
<th>Proximal gene</th>
<th>eGene</th>
<th>Distance to TSS (kb)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Blood</td>
<td>1.10E-46</td>
<td>3.59E-40</td>
<td>-2.84E-02</td>
<td>-3.58E-01</td>
<td>ENSG00000134962</td>
<td>ENSG0000035928</td>
<td>46</td>
</tr>
<tr>
<td>Cerebellar Hemisphere</td>
<td>3.33E-33</td>
<td>4.18E-28</td>
<td>-2.33E-02</td>
<td>5.00E-01</td>
<td>ENSG0000084734</td>
<td>ENSG0000234072</td>
<td>152</td>
</tr>
<tr>
<td>Whole Blood</td>
<td>3.33E-33</td>
<td>4.18E-28</td>
<td>-2.33E-02</td>
<td>-1.39E-01</td>
<td>ENSG0000084734</td>
<td>ENSG0000115216</td>
<td>80</td>
</tr>
<tr>
<td>Cerebellar Hemisphere</td>
<td>3.33E-33</td>
<td>4.18E-28</td>
<td>-2.33E-02</td>
<td>5.00E-01</td>
<td>ENSG0000084734</td>
<td>ENSG0000234072</td>
<td>152</td>
</tr>
<tr>
<td>Whole Blood</td>
<td>3.33E-33</td>
<td>4.18E-28</td>
<td>-2.33E-02</td>
<td>-1.39E-01</td>
<td>ENSG0000084734</td>
<td>ENSG0000115216</td>
<td>80</td>
</tr>
</tbody>
</table>

* Distance of index eQTL to transcription start site (TSS) of target gene (eGene).
disorders (Acheson et al., 2009; Cheetham et al., 2014; Hanson et al., 2010; Herting et al., 2011, 2010; Sjoerds et al., 2013). Interestingly, these brain regions have – among other traits – been implicated in individual differences in alcohol consumption. It should be noted, however, that previous research indicates that these regulation hotspots should be interpreted with caution due to the complex correlation structure of gene expression, which could lead to false positive associations (de Koning and Haley, 2005; Peng et al., 2007).

The associations found for CPD on chromosome 15q25.1 were magnitudes stronger than the ones found for the other substance use traits examined in the current study (i.e., 10^{-40} vs 10^{-211}). The index eQTLs with the strongest associations, all located within the gene cluster, predominantly targeted CHRNA5. This gene cluster has been associated many times in the literature with smoking severity (Barrie et al., 2016; Furberg et al., 2016; Herting et al., 2011, 2010; Sjoerds et al., 2013). Interesting findings were index eQTLs in the putamen, caudate and cervical spine targeting alcohol dehydrogenase 1C (ADH1C) to be significantly associated with alcohol consumption. ADH1C was previously reported to be associated with alcohol dependence and consumption (Clarke et al., 2017; Frank et al., 2012; Park et al., 2013; Treutlein et al., 2009). Moreover, we identified multiple index eQTLs targeting many different genes in various tissues on chromosome 17q21.31 for alcohol consumption. Chromosome 17q21.31 has been described as one of the genome’s most structurally complex and evolutionary dynamic regions, and genes in this region have – among other traits – been implicated in alcohol use (Liu et al., 2019; Nelson et al., 2016; Pennisi, 2008). This underscores the complexity of this region and of the potential importance of gene regulatory mechanisms (Louro et al., 2009). Our results, in combination with the emerging literature, suggest that this region includes multiple functional genetic variants that contribute to individual differences in alcohol consumption. The Bonferroni corrected threshold for transcriptome-wide significance (adjusted for all tissues and genes per trait) is 9.7e-7 for the GTEx analyses. Genquant gene

## Table 3
Overview of the five strongest GTEx informed S-PrediXcan gene associations per substance use trait, genes printed in bold indicate significance.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene name</th>
<th>Zscore</th>
<th>Effect size</th>
<th>pvalue</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000035928.10</td>
<td>RFC1</td>
<td>-8.96</td>
<td>-3.37E-02</td>
<td>3.21E-19</td>
<td>Cerebellar Hemisphere</td>
</tr>
<tr>
<td>ENSG00000238088.3</td>
<td>LRR2C7A2</td>
<td>-8.61</td>
<td>2.79E-02</td>
<td>7.47E-18</td>
<td>Cerebellum</td>
</tr>
<tr>
<td>ENSG00000262539.1</td>
<td>RP11-259G18.3</td>
<td>-8.56</td>
<td>-1.87E-02</td>
<td>1.08E-17</td>
<td>Cerebellar Hemisphere</td>
</tr>
<tr>
<td>ENSG00000214425.2</td>
<td>LRR2C7A4P</td>
<td>8.54</td>
<td>1.89E-02</td>
<td>1.36E-17</td>
<td>Cerebellar Hemisphere</td>
</tr>
<tr>
<td>ENSG00000263503.1</td>
<td>RP11-707023.5</td>
<td>-8.51</td>
<td>-2.35E-02</td>
<td>1.75E-17</td>
<td>Cerebellar Hemisphere</td>
</tr>
</tbody>
</table>

## Alcohol consumption

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene name</th>
<th>Zscore</th>
<th>Effect size</th>
<th>pvalue</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000041357.11</td>
<td>PSMA4</td>
<td>30.86</td>
<td>3.39E-01</td>
<td>4.04E-209</td>
<td>Whole Blood</td>
</tr>
<tr>
<td>ENSG00000166275.11</td>
<td>C1orf2</td>
<td>-7.99</td>
<td>3.15E-02</td>
<td>1.35E-15</td>
<td>Cerebellar Hemisphere</td>
</tr>
<tr>
<td>ENSG00000166275.11</td>
<td>C1orf3</td>
<td>10.28</td>
<td>2.91E-02</td>
<td>2.90E-15</td>
<td>Cerebellum</td>
</tr>
<tr>
<td>ENSG00000235266.1</td>
<td>RP11-753C18.8</td>
<td>-7.73</td>
<td>-3.0E-02</td>
<td>9.42E-15</td>
<td>Whole Blood</td>
</tr>
<tr>
<td>ENSG0000017385.11</td>
<td>LEPRB</td>
<td>7.68</td>
<td>4.13E-02</td>
<td>1.63E-14</td>
<td>Frontal Cortex</td>
</tr>
</tbody>
</table>

## Age of initiation

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene name</th>
<th>Zscore</th>
<th>Effect size</th>
<th>pvalue</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000120903.6</td>
<td>CHRNA2</td>
<td>5.71</td>
<td>1.61E-02</td>
<td>1.10E-08</td>
<td>Cerebellum</td>
</tr>
<tr>
<td>ENSG00000120903.6</td>
<td>CHRNA2</td>
<td>5.44</td>
<td>5.12E-02</td>
<td>5.42E-08</td>
<td>Caudate</td>
</tr>
<tr>
<td>ENSG00000271643.1</td>
<td>RP11-10C24.3</td>
<td>-4.88</td>
<td>4.75E-02</td>
<td>1.01E-17</td>
<td>Cerebellum</td>
</tr>
<tr>
<td>ENSG00000197386.6</td>
<td>HTT</td>
<td>-4.88</td>
<td>-8.73E-02</td>
<td>1.08E-06</td>
<td>Whole Blood</td>
</tr>
<tr>
<td>ENSG00000120903.6</td>
<td>CHRNA2</td>
<td>4.84</td>
<td>1.57E-02</td>
<td>1.28E-06</td>
<td>Cerebellum</td>
</tr>
</tbody>
</table>

## Ever smoker

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene name</th>
<th>Zscore</th>
<th>Effect size</th>
<th>pvalue</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000041357.11</td>
<td>PSMA4</td>
<td>6.16</td>
<td>5.10E-02</td>
<td>7.29E-10</td>
<td>Whole Blood</td>
</tr>
<tr>
<td>ENSG00000233622.1</td>
<td>CYP27BP</td>
<td>5.35</td>
<td>3.97E-02</td>
<td>8.98E-08</td>
<td>Whole Blood</td>
</tr>
<tr>
<td>ENSG00000232630.1</td>
<td>PRO51P2</td>
<td>-5.10</td>
<td>-3.44E-02</td>
<td>3.46E-07</td>
<td>Whole Blood</td>
</tr>
<tr>
<td>ENSG00000269585.1</td>
<td>EGLN2</td>
<td>-5.02</td>
<td>-1.51E-02</td>
<td>4.96E-07</td>
<td>Whole Blood</td>
</tr>
<tr>
<td>ENSG00000126215.9</td>
<td>XRCC3</td>
<td>5.03</td>
<td>2.31E-02</td>
<td>5.10E-07</td>
<td>Whole Blood</td>
</tr>
</tbody>
</table>

## Former vs. Current smoker

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene name</th>
<th>Zscore</th>
<th>Effect size</th>
<th>pvalue</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000257941.1</td>
<td>RP11-290L1.4</td>
<td>3.77</td>
<td>1.73</td>
<td>1.65E-04</td>
<td>Frontal Cortex</td>
</tr>
<tr>
<td>ENSG00000152056.12</td>
<td>AP153</td>
<td>3.63</td>
<td>2.44</td>
<td>2.81E-04</td>
<td>Whole Blood</td>
</tr>
<tr>
<td>ENSG00000271179.1</td>
<td>RP11-629P16.1</td>
<td>-3.62</td>
<td>-1.29</td>
<td>2.95E-04</td>
<td>Nucleus accumbens</td>
</tr>
<tr>
<td>ENSG00000164344.11</td>
<td>KLKB1</td>
<td>-3.60</td>
<td>-4.00E1</td>
<td>3.14E-04</td>
<td>Whole Blood</td>
</tr>
<tr>
<td>ENSG00000169291.5</td>
<td>SHER</td>
<td>-3.57</td>
<td>-1.46</td>
<td>3.62E-04</td>
<td>Whole Blood</td>
</tr>
</tbody>
</table>

consumption, significant index eQTLs were observed in all brain regions and in whole blood. Most of these brain regions have previously been demonstrated to play a role in the susceptibility to alcohol use disorders (Acheson et al., 2009; Cheetham et al., 2014; Hanson et al., 2010; Herting et al., 2011, 2010; Sjoerds et al., 2013). The eQTL results, in combination with the emerging literature, suggest that this region includes multiple functional genetic variants that contribute to individual differences in alcohol consumption. It should be noted, however, that previous research indicates that these regulation hotspots should be interpreted with caution due to the complex correlation structure of gene expression, which could lead to false positive associations (de Koning and Haley, 2005; Peng et al., 2007).

The associations found for CPD on chromosome 15q25.1 were magnitudes stronger than the ones found for the other substance use traits examined in the current study (i.e., 10^{-40} vs 10^{-211}). The index eQTLs with the strongest associations, all located within the gene cluster, predominantly targeted CHRNA5. This gene cluster has been associated many times in the literature with smoking severity (Barrie et al., 2016; Furberg et al., 2016). Our findings extend the previously observed genetic association by showing that the CPD-associated eQTLs regulate gene expression of CHRNA3 and CHRNA5 in striatal brain areas. A study by Barrie et al., in which GTEx tissue-specific eQTLs from the gene cluster CHRNA5/
**Table 4**

Number of gene discoveries from S-PrediXcan analyses and an overview of the number of significant associations in brain tissues, and whole blood.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Alcohol consumption</th>
<th>CPD</th>
<th>Ever vs. never smoker</th>
<th>Age of smoking onset</th>
<th>Former vs. current smoker</th>
<th>Cocaine dependence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Amygdala</td>
<td>18</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Anterior cingulate cortex</td>
<td>21</td>
<td>6</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caudate</td>
<td>25</td>
<td>8</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cerebellar hemisphere</td>
<td>29</td>
<td>7</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>34</td>
<td>8</td>
<td>27</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Cortex</td>
<td>39</td>
<td>6</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Frontal cortex</td>
<td>18</td>
<td>3</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hippocampus</td>
<td>15</td>
<td>7</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hypothalamus</td>
<td>17</td>
<td>4</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nucleus accumbens</td>
<td>21</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Putamen</td>
<td>17</td>
<td>6</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Spine</td>
<td>12</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Substantia nigra</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Whole blood</td>
<td>27</td>
<td>14</td>
<td>21</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>n unique genes ‘total’</td>
<td>71</td>
<td>35</td>
<td>78</td>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>n unique genes ‘brain’ ‘total’</td>
<td>61</td>
<td>27</td>
<td>67</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>n unique genes ‘brain’ ‘only’</td>
<td>10</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>n unique genes ‘whole blood’ ‘total’</td>
<td>44</td>
<td>21</td>
<td>57</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

n unique genes ‘total’ = the number of unique genes that are significant in one or more tissues; n unique genes; n genes ‘brain’ = the number of genes that are significant in one or more brain tissues; n genes ‘brain’ = the number of genes that are significant in one or more brain tissues and not significant in whole blood; n unique genes ‘whole blood’ only, the number of genes that are significant in whole blood, but not significant in any of the brain tissues.

**CHRNA3/CHRN4** were explored to investigate their role in nicotine dependence, found significant stratial eQTLs targeting the same eGenes (CHRNA3, CHRNA5, RP11-650L12.2) as we found here for CPD (Barrie et al., 2016). The stratial eQTLs identified in the current study and the findings of Barrie et al. highlight the importance of genetic regulation in the striatum for smoking behaviors. For the other smoking related traits findingsof Barrie et al. highlight the importance of genetic regulation in
eQTLs. In fact, in 65.6 percent another eGene was targeted. This is probably due to the fact that both GTEx and DGN are underpowered, and thus providing incomplete, not overlapping, eQTL information.

The findings and conclusions of this study should be interpreted in view of some key limitations. Despite the comparatively large sample sizes of the GWAS, our study may still be underpowered to detect small genetic effects, which is especially true for the cocaine dependence sample. In addition, although GTEx and DGN belong to the most comprehensive genetic expression databases, the statistical power for eQTL discovery is still modest for some tissues (Ardlie et al., 2015). The GTEx brain sample sizes are smaller than those for whole blood, which is reflected in fewer identified index eQTLs in the separate brain tissues. It is therefore likely that we tested only a subset of the total number of true eQTLs for the various tissues. Our analyses focus on the role of eQTLs in a wide range of tissues, however, recently it has been shown that eQTL effects may differ between cell types within a specific tissue (van der Wijst et al., 2018). Therefore, to better understand the role of gene expression in substance use traits future studies should focus on cell type specific analyses (van der Wijst et al., 2018). Furthermore, the genes identified with the single variant and S-PrediXcan analyses should be seen as ‘candidates’ as correlated levels of gene expression may be observed in high LD genomic regions which makes it challenging to identify the true causal genes (Wainberg et al., 2019). Moreover, the GTEx data was composed of subjects of European ancestry, while the GWAS results used in the current study for and cocaine dependence are based on European and African American subjects. Since eQTLs may not completely overlap across ethnic populations, this may have reduced our ability to detect novel loci. However, with sample size always being a limiting factor in complex trait GWAS, we included the
combined sample of European and African American subjects for cocaine dependence to improve statistical power. By no means we claim to present the full set of eQTLs, eGenes, and tissues involved in the substance use traits under investigation. Therefore, the index eQTLs, eGenes, and tissues identified by this study to be involved in substance use traits, should be seen as ‘candidates’. Moreover, independent replication of these novel candidate loci is necessary before strong conclusions can be drawn regarding the role of these loci in substance use traits.

5. Conclusions

Our findings suggest that there is great value in utilizing brain and whole blood eQTL annotations for enhancing the discovery of novel genetic susceptibility loci for substance use traits. The tissue-focused GTEx eQTL analyses revealed 2,976 index eQTLs which were not identified in the discovery samples using the same threshold (FDR adjusted p < 0.05), implying that these candidate loci might be interesting for further research. In addition, the functional annotation of GWAS data revealed that most of the identified candidate eGenes targeted by the trait-associated index eQTLs are not the nearest genes, underscoring the importance of studying genetic regulation of gene expression for functional annotation of genetic loci. Finally, the S-PrediXcan results validated some of the candidate findings of the single variant analyses. Both the single variant and gene-level analyses confirm the conclusions of Gamazon et al. regarding the importance of multiple tissue eQTL investigation (Gamazon et al., 2018), as our analyses identified many interesting (novel) candidate eGenes and tissues for substance use traits. In conclusion, annotating genes based on transcriptomic regulation in brain and non-brain tissues improves both the identification of novel candidate genes and the functional characterization of genetic risk factors for substance use traits.

Author disclosures

ATM and EMD are supported by the Foundation Volksbond Rotterdam. ATM is supported by the Netherlands Organization of Scientific Research (NWO Vidi grant 016.Vidi.185.044, PI T.J. Galama). ERG is supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R35HG010718. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. FY is supported by the Investissements d’Avenir program managed by the ANR under reference ANR-11-IDEX-0004-02. KJHV is supported in part by a 2014 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation. ERG benefited from a Clare Hall Fellowship at the University of Cambridge. The funding sources had no involvement in study design; in the collection, analysis and interpretation of the data; in the writing of the report or the decision to submit for publication.

Author contributions

ATM performed the analyses, prepared the first draft of the manuscript, and completed the final version of this manuscript. ERG designed the methodology, provided scripts, and contributed to the writing process. ZG conducted the S-PrediXcan analyses, and contributed to the writing process. FY, DJAS, WvdB, and KJHV contributed in interpreting the results and to the writing process. JF performed a pilot study and performed part of the analyses. JG, RS, LF, and HRK participated in the data collection, and contributed to the writing process. EMD designed the methodology, contributed in the writing process, and supervised the project.

Dr. Kranzler has been a consultant, advisory board member or CME speaker for Indivior and Lundbeck. He is also a member of the American Society of Clinical Psychopharmacology’s Alcohol Clinical Trials Initiative (ACTIVE), which was supported in the last three years by AbbVie, Alkermes, Ethypharm, Indivior, Lilly, Lundbeck, Otsuka, Pfizer, and XenoPort.

Declaration of Competing Interest

Dr. Gamazon receives an honorarium from the journal Circulation Research of the American Heart Association, as a member of the Editorial Board. He also performs consulting on pharmacogenetic analysis with the City of Hope / Beckman Research Institute.

Dr. Kranzler has been a consultant, advisory board member or CME speaker for Indivior and Lundbeck. He is also a member of the American Society of Clinical Psychopharmacology’s Alcohol Clinical Trials Initiative (ACTIVE), which was supported in the last three years by AbbVie, Alkermes, Ethypharm, Indivior, Lilly, Lundbeck, Otsuka, Pfizer, and XenoPort.

Dr. van den Brink received speaker’s fees from Lundbeck, Indivior, Eli Lilly, and Pfizer and is a consultant to Indivior, Mundipharma, Novartis, Bioproject, D&A Pharmaceuticals, and Opiant Pharmaceuticals. No funding was received related to the present article.

The other authors declare no competing interests with regard to this study.

Appendix A


1Department of Complex Trait Genetics, VU Amsterdam, Center for Neurogenomics and Cognitive Research, Amsterdam, The Netherlands
2Department of Biological Psychology/Netherlands Twin Register, VU University, Amsterdam, The Netherlands
3The Hospital for Sick Children Research Institute, Toronto, Canada
4Department of Psychology, University of Illinois Urbana-Champaign, Champaign, Illinois, USA
5Department of Human Neurogenetics, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands
6Department of Psychology, University of Minnesota, Minneapolis, Minnesota, USA
7Department of Public Health, University of Helsinki, Hjelt Institute, Helsinki, Finland
8VU University, Department of Developmental Psychology and EMGO Institute for Health and Care Research, Amsterdam, The Netherlands
9Estonian Genome Center, University of Tartu, Tartu, Estonia
10Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

7
Appendix B. Supplementary data

Supplementary material related to this article can be found in the online version, at doi:10.1016/j.drugalcdep.2019.107703.

References


