HHS Public Access

Author manuscript

Addict Biol. Author manuscript; available in PMC 2019 January 01.

Published in final edited form as:

Addict Biol. 2018 January ; 23(1): 461–473. doi:10.1111/adb.12489.

Whole Genome Sequence Study of Cannabis Dependence in Two Independent Cohorts

Ian R. Gizer, PhD[*], **Chris Bizon, PhD**[*], **David A. Gilder, MD**, **Cindy L. Ehlers, PhD**[+], and **Kirk C. Wilhelmsen, MD, PhD**[+]

Department of Psychological Sciences, University of Missouri, Columbia, MO 65211 (IRG), the Department of Molecular and Cellular Neurosciences, The Scripps Research Institute, La Jolla, CA 92037(CLE, DAG), and the Departments of Neurology and Genetics, University of North Carolina, Chapel Hill, NC 27599 (KCW, CB)

Abstract

Recent advances in genome wide sequencing techniques and analytical methods allow for more comprehensive examinations of the genome than microarray-based genome-wide association studies. The present report provides the first application of whole genome sequencing to identify low frequency variants involved in cannabis dependence across two independent cohorts. The present study used low-coverage whole genome sequence data to conduct set-based association and enrichment analyses of low frequency variation in protein-coding regions as well as regulatory regions in relation to cannabis dependence. Two cohorts were studied: a population-based Native American tribal community consisting of 697 participants nested within large multi-generational pedigrees and a family-based sample of 1832 predominantly European ancestry participants largely nested within nuclear families. Participants in both samples were assessed for DSM-IV lifetime cannabis dependence, with 168 and 241 participants receiving a positive diagnosis in each sample, respectively. Sequence kernel association tests identified one protein-coding region, *C1orf110* and one regulatory region in the *MEF2B* gene that achieved significance in a meta-analysis of both samples. A regulatory region within the *PCCB* gene, a gene previously associated with schizophrenia, exhibited a suggestive association. Finally, a significant enrichment of regions within or near genes with multiple splice variants or involved in cell adhesion or potassium channel activity were associated with cannabis dependence. This initial study demonstrates the potential utility of low pass whole genome sequencing for identifying genetic variants involved in the etiology of cannabis use disorders.

Corresponding Author: Cindy L. Ehlers PhD, 10550 N. Torrey Pines Road, SP30-1501, La Jolla, CA 92037. Telephone (858) 784-7058 cindye@scripps.edu.

* and + indicate that these sets of authors contributed equally to the present study.

Authors' contributions

IRG, CLE, and KCW were responsible for the study concept and design. CLE, DAG, and KCW were responsible for overseeing data collection. KCW and CB performed the genomic sequencing and variant calling. IRG, CB, CLE, and KCW assisted with data analysis and interpretation of findings. IRG drafted the manuscript. CB, CLE, and KCW provided critical revision of the manuscript for important intellectual content. All authors critically reviewed content and approved final version for publication.

The authors report no potential conflicts of interest.

Keywords

cannabis; genetic association; genetics; marijuana dependence; whole genome sequencing

Introduction

Cannabis is the most widely used illicit drug in the United States, with lifetime rates of cannabis use disorders ranging from 1.5%–2.9 % (Hasin et al., 2015). Cannabis use is associated with substantial physical and psychiatric sequelae in some users (Volkow et al., 2014 & 2016). Given that cannabis use is expected to increase in the United States with increased decriminalization and legalization, it is of great interest to identify etiological factors that may lead to dependence. Twin studies have shown a substantial genetic component to the etiology of cannabis use disorders, with a recent meta-analysis suggesting heritability estimates of 0.51 for males and 0.59 for females (Verweij et al., 2010). Nonetheless, molecular genetic studies attempting to identify the specific variants involved in this genetic risk have yielded primarily mixed results.

Candidate gene studies of the endocannabinoid receptor type 1 gene (*CNR1*) and/or the fatty acid amide hydrolase gene (*FAAH*), which is involved in endocannabinoid metabolism, have reported associations with cannabis use and dependence (e.g., Tyndale et al., 2007), marijuana withdrawal and craving (Haughey et al., 2008), and related phenotypes such as trait impulsivity (Ehlers et al., 2007). These results, however, have failed to achieve the current standards for genome-wide significance. Genome-wide association studies (GWAS) have identified novel loci related to cannabis use disorders, including ankyrin-repeat and fibronectin type III domain containing 1 gene, *ANKFN1* (Agrawal et al., 2011), and a gene cluster located on chromosome 17q24 (*c17orf58, BPTF, PPM1D* (Agrawal et al., 2014)). Nonetheless, these results also failed to achieve genome-wide significance. More recently, the first genome-wide significant associations were reported with loci located in or near three genes, *RP11-206M11.7*, *SLC35G1*, *CSMD1* (Sherva et al., 2016).

The application of next generation sequencing methods to the study of complex traits presents several potential advantages over GWAS microarrays that may allow for further progress to be made. First, GWAS microarrays have been primarily designed to measure common genetic variants (i.e., minor allele frequencies [MAF] > 0.05), and thus, are not well-positioned to capture genetic variants with lower allele frequencies (Nelson et al., 2013). As a result, rare variation has been cited by some as a potential source of the 'missing' heritability, which refers to the gap between heritability estimates of complex traits derived from twin studies and the proportion of variation in a trait explained by measured genetic variants in a GWAS, as well as the 'still missing' heritability, which refers to the gap between heritability estimates of complex traits derived from twin studies and those derived from GWAS using genomic similarity approaches (Wray and Maier, 2014), including traits such as cannabis initiation (Stringer et al., 2016). In contrast, sequencing technologies, which directly interrogate each variant, do not have this limitation. Second, common genetic variation captured by GWAS microarrays varies as a function of the ancestral group under study thus complicating the study of diverse ancestral groups (Cantor et al., 2010). More

specifically, these microarrays have typically been designed to capture common variation in individuals of specific ancestral groups (e.g., European ancestry). Thus, variants specific to populations outside of these groups (e.g., American Indians) may not be captured.

Given the described advantages, the present report utilized low-coverage whole genome sequencing (WGS) to identify low-frequency risk variants for cannabis use disorders that would typically not be captured by a GWAS microarray in two independent cohorts, a population-based sample of Native American Indians and a family-based study of individuals of predominantly European descent initially selected for alcohol dependence. The inclusion of a Native American cohort should also be emphasized given that this cohort represents an understudied population with some of the highest rates of substance use problems in the United States (Ehlers et al., 2004). The present study used WGS to study the relations of low frequency variants to Diagnostic and Statistical Manual of Mental Disorders - IV (DSM-IV) cannabis dependence using three primary analytic approaches: (1) set-based tests of association of low frequency variants (MAF < 0.02) located in protein-coding regions of the genome, (2) set-based tests of association of low frequency variants located in regulatory regions of the genome, and (3) enrichment analysis to evaluate whether genes related to associated sets shared similarities in structure or function.

Materials and Methods

Data were collected at The Scripps Research Institute and the Gallo Institute at the University of California at San Francisco (UCSF). Assessment procedures were approved by Institutional Review Boards at each institution. Data collection procedures by The Scripps Research Institute were also approved by a tribal group overseeing health issues for the communities where recruitment took place. Notably, human subject permissions and the wishes of the participating tribes do not allow study data to be entered into public databases. Ongoing management and analysis of data collected at the UCSF site was approved by the Institutional Review Board at the University of North Carolina at Chapel Hill. Participants at both sites were fully briefed on the nature of the study, provided written informed consent prior to enrollment, and were compensated for time spent in the study.

Participants

Native American Sample—Participants were recruited from 8 geographically contiguous Indian communities with a total population of ~3,000 individuals. Individuals of Native American heritage that were between the ages of 18 and 82 years were recruited to participate using a combination of a venue-based methods for sampling hard-to-reach populations (Muhib et al., 2001) and a respondent-driven procedure (Heckathorn, 1997), as reported previously (Ehlers et al., 2004). All subjects were assessed using the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA) to collect demographic information and make DSM-IV cannabis dependence diagnoses. The SSAGA is a reliable and valid polydiagnostic psychiatric interview that has been successfully used in Native American populations (Bucholz et al., 1994).

Of the 775 participants in the Native American sample, 697 were successfully sequenced and included in the association analyses. Sixty-seven samples could not be sequenced

because of insufficient or low quality deoxyribonucleic acid (DNA), and 11, none of whom met criteria for cannabis dependence, were excluded because of sample misidentification as assessed by comparisons of self-reported familial relations to kinship coefficients derived from genotypes using PREST (Sun et al., 2002). Among sequenced participants, 168 were diagnosed with cannabis dependence. The average age of the sample was 31.2±0.5 years and 57% were female, though cannabis dependent participants were more likely to be older and male (Table 1). Forty-two percent of participants reported at least 50% Native American heritage based on their federal Indian blood quantum (Bizon et al., 2014).

University of California at San Francisco (UCSF) Family Study Sample—The UCSF Family Study sample was recruited nationwide for inclusion in a study of the genetics of alcoholism and other substance dependence. Responding individuals were invited to participate if they met screening criteria for a lifetime alcohol dependence diagnosis and had at least one sibling or both parents available to participate. Permission was then obtained from the proband to invite relatives to participate by mail. Probands with serious drug addictions other than cannabis (e.g., stimulants, cocaine, or opiates) or a history of intravenous substance use were excluded. Also excluded were subjects reporting a current or past diagnosis involving psychotic symptoms, a life-threatening illness, or an inability to speak and read English.

Participants represent a subset of the UCSF Family Alcoholism Study. From the full sample, 1886 were successfully sequenced and provided complete phenotype data. From this subset, a further 54 were excluded because of sample misidentification, 10 of whom met criteria for cannabis dependence, resulting in a sample size of 1832. Within the final sample, 241 participants met criteria for cannabis dependence. The average age of the study sample was 49.0±13.1 years, 62% (n=1134) were female, and 95% reported Caucasian ethnicity (n=1737) with the remaining individuals reporting African-American (n=52), Native American (n=25), 'Other' (n=17), and Asian (n=1) ancestry. Cannabis dependent participants were more likely to be male, younger, unmarried, and have an annual income <$20,000 (see Table 1).

Sequencing

Blood derived DNA was sequenced using Illumina low-coverage WGS, and genotyped using an Affymetrix Exome1A chip to assess accuracy of genotype calls from sequence data. Pair-end sequencing was performed on HiSeq2000 sequencers (Illumina, San Diego, CA). For the Native American sample, approximately 80% of samples were sequenced at a coverage depth between 3X and 12X (range: 1X – 31X) with depth of coverage evenly distributed across the genome. For the UCSF Family Study sample, approximately 86% of the samples were sequenced at a coverage depth between 2X and 6X (range: 1X – 18X) with depth of coverage evenly distributed across the genome. Sequence reads were aligned using blocked multiple-sequence alignment (BMA), and realigned near indels with the Genome Analysis Toolkit (GATK). Because low-pass sequencing was used and to capitalize on the fact that participants were nested within families, variants were called using the LD-aware variant caller Thunder (Li, 2011) in order to increase the accuracy of the variant calls. Variant call quality was assessed through comparison of the sequencing results to exome array genotypes

for all subjects resulting in a 98% concordance rate in each sample. Because the number of rare variants identified within a single study increases with sample size due to their low frequency, this low-coverage approach allowed for the sequencing of a greater number of participants and inclusion of a larger number of rare variants. This led to a greater number of identified rare variants in the UCSF Family study relative to the Native American sample, given the former's larger size. The general methodology for sequencing and variant calling in this sample has been previously published (Bizon et al., 2014).

Data Analysis

To assess ancestry and admixture proportions in the Native American sample, we used a supervised clustering approach that used the algorithm implemented in ADMIXTURE (Alexander et al., 2009) in conjunction with a reference panel containing genotype information at about 300k strand-unambiguous SNPs. The ancestry estimates were then further refined through a noise reduction approach via bootstrapping (Libiger and Schork, 2013) and identified as corresponding to the four major continental populations: African, East Asian, European, and Native American. For the UCSF Family Study, ancestry proportions were obtained using principal components analysis (Price et al., 2006). Regression analyses examining the relations between these ancestry proportions and the cannabis diagnosis were nonsignificant (p-values: 0.366–0.890).

Participants in both samples were nested in families. Thus, the linear mixed model approach implemented in EMMAX (Kang et al., 2010) was used to control for population substructure and genetic relatedness in all analyses. This approach calculates a genetic similarity matrix for all pairwise combinations of study participants using measured genotype data and includes this matrix as a random effect in the mixed model to partition variance in the phenotype that can be attributed to familial relatedness and population substructure. Ancestry estimates obtained as described above were included as covariates to fully account for potential inflation in the test statistics along with gender, age, and age-squared.

For the set-based analyses, low frequency variants (MAF < 0.02) were analyzed using the optimized sequence kernel association test (SKAT-O; Lee et al., 2012a) implemented within the EMMAX framework. Only sets for which at least 1% of participants in each sample carried a low frequency variant were included in the analysis. Specifically, if only a small proportion of the study participants carry a rare variant within a given set (i.e., < 1% of the study sample), the resulting analysis would be similar to analyzing a low frequency variant in isolation, and thus, will be particularly susceptible to chance fluctuations in the data, given that the result is based on only a handful of observations. For sets based on the coding regions of genes, all variants within the protein coding region were retained, including nonsynonymous, synonymous, start codon loss or gain, and stop codon loss or gain variants. For sets based on regulatory elements, the decision was made to focus on regulatory elements that showed evidence of persisting across tissue types, given that little is known regarding the etiology of cannabis dependence and alterations in gene expression both inside and outside of the central nervous system are of potential relevance. We considered restricting this analysis to only regulatory elements influencing genes expressed in the central nervous system; however, many regions were located in close proximity to more than

one gene or in intergenic regions, making it difficult to assign many of these regions to a specific gene. Because the primary effect of including non-CNS expressed genes in the analysis was the adoption of a more conservative significance threshold, the decision was made to include all regulatory regions in the analysis. Thus, data from the NIH Roadmap Epigenomics project focused on identifying epigenetic marks across tissues were accessed and used to define the boundaries of the regulatory elements (Roadmap Epigenomics Consortium et al., 2015); HoneyBadger2 dataset accessed on 12/15/2015 from http://www.broadinstitute.org/~meuleman/reg2map/). Because some of these elements are quite small (150 bp) and frequently lie close together, elements within 2 kb of each other were combined into a single element.

All analyses were performed separately for each cohort within the EPACTS software pipeline (Kang, 2014). A weighted Z-score approach (Stouffer et al., 1949) was used to combine results across samples. Enrichment analyses were conducted using the Database for Annotation, Visualization and Integrated Discovery (DAVID; Huang et al., 2009) in which genes with p-values < 0.005 were evaluated to determine whether they were over-represented in any gene categories. These analyses were conducted using a custom background that included only those genes tested in the set based tests of coding regions and regulatory elements, respectively.

Results

Analysis of Protein-Coding Regions

We first conducted set-based analyses of the genes in the RefSeq database. 12,662 genes were identified that met the described inclusion criteria, resulting in a critical p-value of 3.9e–6 to determine significance. A single gene that met this criterion was identified in meta-analysis, a coiled-coil domain containing protein on chromosome 1, *C1orf110* (p=3.20e–6) (see Table 2; also see Supplementary Table 1 for top results in each cohort and Supplementary Figures 1 and 2 for q-q plots of the distribution of *p*-values in each cohort). Notably, these analyses were repeated when restricting the UCSF Family Study sample to participants that were of predominantly European ancestry (>70%; n=1641) as indicated by the principal components analysis. The number of excluded participants was greater than the number of participants that self-reported non-European ancestry due to discordances between self-report and genetically-derived ancestry estimates. The meta-analytic result for *C1orf110* was similar (p=4.66e–6), but no longer met the significance threshold after dropping these participants. A similar, though slightly larger, drop in the p-value was observed when the same number of randomly selected European ancestry participants were dropped from the analysis (p=2.14e–5). Together, these results suggest the change in significance likely reflects a decrease in statistical power following the exclusion of almost 200 participants rather than unaccounted for population substructure. Figure 1 displays the layout of the gene and the location of the variants included in the test for each sample. To further explore the relations of these variants to cannabis dependence, the Polyphen, SIFT, and Provean databases were queried to estimate the likely impact of these variants on gene function. The resulting values are shown in Table 3 alongside the single variant EMMAX results, which provide an indication of the direction and magnitude of their relation with

cannabis dependence. An examination of this table suggests that, within the UCSF Family Study sample, variants predicted to negatively impact protein function show stronger relations to the cannabis dependence diagnosis. The results are less clear for the Native American sample; however, the one variant that does show a relation with cannabis dependence (rs187742957) is predicted to negatively impact the encoded protein.

We then conducted an enrichment analysis based on the SKAT-O results using DAVID. This analysis suggested an over-representation of genes in several pathways related to potassium channel activity, including potassium ion transport (GO:0006813: Benjamini-Hochberg corrected p=0.013), cation channel complex (GO:0034703: Benjamini-Hochberg corrected p=0.036), voltage-gated potassium channel complex (GO:0008076: Benjamini-Hochberg corrected p=0.040), potassium channel complex (GO:0034705: Benjamini-Hochberg corrected p=0.040), and potassium channel activity (GO:0005267: Benjamini-Hochberg corrected p=0.043). The genes contributing to these results included *KCNK17, RYR2, SLC24A3, KCNJ4,* and *SLC24A2.*

Analysis of Regulatory Elements

We then conducted set-based analyses of regulatory elements identified by the NIH Roadmap Epigenomics project (Roadmap Epigenomics Consortium et al., 2015) that were consistently observed across tissue types. 165,586 elements were identified that met the inclusion criteria for this analysis resulting in a critical p-value of 3.0e–7 that was used to determine significance. One regulatory element met this threshold (Table 4; also see Supplementary Table 2 for top results in each cohort and Supplementary Figures 3 and 4 for q-q plots of the distribution of *p*-values in each cohort). Variants contained within the 1st intron of the myocyte enhancer factor 2B gene (*MEF2B*) showed a significant relation with cannabis dependence across samples (combined p-value = 1.28E–08). As shown in Figure 2, the significant region is characterized by the presence of both H3K9me3 and H3K36me histone marks, a pattern that has been associated with the preservation of exons from recombination and regulation of alternative splicing. A query of the Ensemble database does include a *MEF2B* transcript (ENST00000409447) that is annotated as having an untranslated exon in this region (bottom of Figure 2). This analysis was repeated when restricting the UCSF Family Study sample to those participants of predominantly European ancestry and again dropping the same number of European ancestry participants at random. The p-value for the association with the *MEF2B* element fell by an order of magnitude and became nonsignificant when non-European ancestry participants were removed from the analysis (p=3.52e–7). The p-value dropped by half an order of magnitude when a random set of European ancestry participants were removed from the analysis (p=7.80e–8), but continued to meet the significance threshold. Given the relative similarity in the decline of the p-value, it is likely that these changes resulted from reduced statistical power rather than unaccounted for population stratification.

A suggestive association was also observed for a regulatory element within the 3rd intron of the propionyl-CoA carboxylase beta subunit gene (*PCCB*; combined p-value = 1.55E–06). As shown in Figure 3, this regulatory element is characterized by the presence of H3K4me1 histone marks as well as the absence of H3K27ac marks, which have been suggested to

reflect poised enhancers that become active during cell differentiation and reflect activity of the gene in the cell's adult state. Data from the Roadmap Epigenomics project suggest the presence of these marks in relation to this intronic regulatory region is most pronounced in brain tissues, suggesting a role for *PCCB* in neural development and function. This association became slightly stronger when the UCSF Family Study sample was restricted to participants of European ancestry (p=3.51e–7) and weaker when dropping a random set of European ancestry participants (p=3.04e–5).

Enrichment analyses were then conducted using DAVID by assigning individual regulatory elements to the nearest gene within 5000 kilobases. This analysis suggested an over-representation of genes that exhibit multiple splice variants (uniprot keywords: alternative splicing - Benjamini-Hochberg corrected p=4.2e–07; uniprot sequence annotation: splice variant - Benjamini-Hochberg corrected p=4.6e–06) as well as genes related to cell adhesion (Panther BP00124 - Benjamini-Hochberg corrected p=0.010) and neurogenesis (Panther BP00199 - Benjamini-Hochberg corrected p=0.042).

Discussion

The present report represents, to our knowledge, the first study to utilize next generation sequencing technology to identify low-frequency genetic variants related to cannabis dependence. Three approaches were taken to accomplish this: (1) set-based tests of association of low frequency (MAF<0.02) coding variants, (2) set-based tests of association of low frequency variants within regulatory regions identified by the NIH Roadmap Epigenomics project, and (3) enrichment analysis to evaluate whether specific gene- or regulatory-based sets were associated with cannabis dependence. Analyses were conducted in two distinct cohorts, a Native American community sample and participants from the UCSF Family Alcoholism study, and results were combined to identify variants and genes that confer risk for cannabis dependence across samples.

The set-based analyses of low frequency variation in protein-coding regions yielded a single genome-wide significant association between cannabis dependence and the coiled-coil domain-containing protein gene (*C1orf110*) on chromosome 1. Adding to the strength of this finding, the pattern of results for the individual variants within the coding regions of this gene was highly concordant with predictions regarding the impact of the individual variants on *C1orf110* function, at least in the UCSF study sample. Though the result was weaker in the Native American sample, the one variant that did show a relation with cannabis dependence was predicted to negatively impact the *C1orf110* protein. Interpretation of this result is somewhat complicated given that little is known about the *C1orf110* protein. Nonetheless, the gene does appear to be a downstream target of the NAD-dependent deacetylase sirtuin-2, which plays an important role in cellular responses to oxidative stress (Liu et al., 2013). Notably, SIRT-2 shows altered expression in the hippocampus following repeated exposure to Δ^9-THC (Quinn et al., 2008), and thus, may provide an avenue of research for future studies investigating how *C1orf110* may be related to cannabis dependence.

The set-based analyses of regulatory regions also yielded a single genome-wide significant association. Variation within a regulatory region located in the first intron of *MEF2B* was associated with risk for cannabis dependence. This element was characterized by H3K9me3 and H3K36me histone marks, suggesting the region may be important for protecting exons from recombination and regulation of alternative splicing (Schor et al., 2009). In contrast to variants in protein-coding regions where *in silico* tools (e.g., Polyphen, SIFT) can be used to make strong predictions regarding their impact on function, similar predictions regarding the impact of variants in regulatory regions on expression are not currently available, limiting the ability to determine which variants in the region might be most strongly related to cannabis dependence risk. Nonetheless, the *MEF2B* protein, and the MEF protein family in general, are of direct relevance to substance use phenotypes. *In vitro* studies have demonstrated that these proteins play important roles in synapse formation and plasticity (Flavell et al., 2006). In the anterior cingulate and hippocampus, where *MEF2B* is prominently expressed, increased MEF expression leads to decreases in dendritic spine density and has been shown to result in disrupted memory formation (Rashid et al., 2014). In the nucleus accumbens, suppression of MEF2 proteins is necessary for the observed increase in dendritic spine density that follows cocaine administration (Pulipparacharuvil et al., 2008). Thus, it may be that variation in *MEF2B* influences memory formation and associative learning with respect to substance use.

Though not significant, a second regulatory element yielded suggestive evidence for association with cannabis dependence risk. This element was located in an intron of the *PCCB* gene, which encodes for the propionyl-CoA carboxylase (PCC) enzyme beta subunit, an enzyme involved in the metabolism of several amino acids, lipids and cholesterol. Loss of function mutations in this gene lead to propionic acidemia, a condition that leads to the toxic buildup of propionyl-CoA in the nervous system and can lead to developmental delays, intellectual disability, in some cases, psychosis (Dejean de la Bâtie et al., 2014). Of note, *PCCB* is also located in a region that was associated with increased risk for schizophrenia in the largest genome-wide association study of that disorder conducted to date (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), which is of interest given the putative link between adolescent cannabis use and increased risk for psychosis (Volkow et al., 2016).

In comparing the set-based tests of protein-coding and regulatory regions, a higher degree of concordance was observed across study samples for the latter tests. It is likely that risk variants unique to each population exist, which could explain the discrepant results for the analysis of protein-coding regions. Nonetheless, previous reviews suggest that such differences are unlikely to account for a substantial proportion of the heritability of substance use phenotypes (Ehlers and Gizer, 2013), and would not explain the higher concordance rates for the analysis of regulatory elements. An alternative explanation, though speculative, is that the difference in consistency may be a reflection of the contributions of variation in protein-coding relative to regulatory regions to the etiology of complex phenotypes such as marijuana dependence. Previous studies have indicated that variants associated with complex traits contained in the GWAS catalog are more likely to be located within intronic and other regulatory regions of a gene rather than the protein-coding region (Nicolae et al., 2010). As a result, it is possible that the greater consistency observed across

samples for the set-based tests of regulatory elements may reflect a greater proportion of 'true' signals, relative to those in the protein-coding regions.

With respect to the significant results reported in the present study, it should be noted that the SKAT-O test analyzes the correlations among individual variant score tests within a set, and it cannot provide an estimate of explained variance attributable to the combined set of variants. Nonetheless, the effects sizes for the individual variants within sets ranged from R^2 = 0.0 – 0.015, suggesting effect sizes comparable to those found in GWAS. This has important implications for future studies seeking to relate low frequency variation to complex psychiatric traits such as substance use.

As noted, rare variation has been cited by some as a potential source of the 'missing' heritability of complex traits (Wray and Maier, 2014). The present report suggests that low frequency variants of modest-to-large effect are unlikely to play an outsized role in the etiology of cannabis dependence at the population level, though this does not rule out the possibility that such variants may have larger effects within families. Studies of *de novo* variation in autism and schizophrenia suggest that moderately penetrant rare variants related to these disorders may be distributed across a broad set of genes (Neale et al., 2012). If a similarly polygenic architecture can be assumed for substance use disorders, this suggests that large samples will be required to detect associations with low frequency variants even when set-based aggregation approaches, such as the SKAT-O test, are implemented.

This may also explain why there was little overlap observed with previous GWAS of cannabis dependence (Agrawal et al., 2011, 2014; Sherva et al., 2016). Although some discrepancies should be anticipated given the focus of GWAS on common variation and the focus of the present report on low-frequency variants, the genes involved should show some degree of replication if they are associated with risk for the disorder (Visscher et al., 2012). As sample sizes examining each type of variation increase allowing for more powerful tests of association, it would be expected that results from these studies would converge on an overlapping set of genes involved in risk for cannabis dependence.

Pathway analyses allow for the aggregation of even larger variant sets relative to set-based analyses of specific genomic elements, and thus, represent another approach for studying low frequency variation in relation to complex traits. Analyses conducted in the present study revealed the strongest relations between cannabis dependence and two gene-sets characterized by genes with multiple splice variants. This is of particular interest given that variants influencing the alternative splicing of a gene are overrepresented within the NHGRI GWAS catalog (Lee et al., 2012b). Additionally, alternative splicing processes are critically involved in human brain development (Johnson et al., 2009), and there is emerging evidence that these processes may be disrupted in psychiatric disorders, including schizophrenia (Barry et al., 2014) and substance use disorders (Moyer et al., 2011). Finally, previous GWAS have shown relations between genes included in potassium channel cocaine dependence (Gelernter et al., 2014b) and opioid dependence (Gelernter et al., 2014a) and between substance use disorders and cell adhesion genes (e.g., Edwards et al., 2015), providing further evidence supporting the role of these gene pathways in the etiology of cannabis use disorders.

Given the promising nature of the described results, it is important to note that the modest sample size, though large for a whole genome sequencing study, represents a limitation of the present study. As stated, the significant associations of the protein-coding regions of *C1orf110* and the regulatory element within *MEF2B* became nonsignificant when the UCSF Family Study sample was restricted to European ancestry participants, but very similar results were observed when a random set of European ancestry individuals were dropped from the analysis. This suggests the changes in p-values resulted from the ~10% reduction in sample size, and that the overall results were not inflated by the inclusion of ancestrally diverse individuals. Further, a post-hoc power analysis was conducted, indicating that the set-based tests of protein-coding regions had power=0.56 and the tests of regulatory regions had power=0.47 to detect a significant effect[1], which were reduced when restricting the UCSF Family Study sample to European ancestry individuals. Together, this highlights an important difficulty in conducting whole genome sequencing studies, given the associated costs, and emphasizes the need for large-scale collaborative efforts such as those that have been developed for GWAS.

In conclusion, it is important to emphasize the unique nature of the present study. To our knowledge, this is the first study to use whole genome sequence data to conduct genome-wide analyses of low frequency variation in relation to cannabis dependence. Thus, the comprehensive nature of the data analyzed presents an important advance over previous studies conducted using genotyping microarrays. This is particularly relevant to the study of Native Americans given that they represent a historically under-studied population. By using a whole genome sequencing approach, the present report provides initial data suggesting that functional variation in *C1orf110* and variation potentially involved in the regulation of *MEF2B* and *PCCB* expression play an important role in the etiology of cannabis dependence. Nonetheless, replication will be an important to step in evaluating the robustness of the reported results.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grants from the National Institute of Drug Abuse (NIDA; DA030976) to IRG, CLE, and KCW and from the National Institute of Alcohol Abuse and Alcoholism (NIAAA) and the National Center on Minority Health and Health Disparities (NCMHD) (5R37 AA010201) to CLE. Additional funding provided by the State of California and the Ernest Gallo Clinic and Research Center for medical research on alcohol and substance abuse through the University of California at San Francisco (UCSF) to KCW. The authors wish to thank Piotr A Mieczkowski, Ewa P Malc, Phil Owen, and Scott A Chasse for their assistance in sample preparation and performing the sequencing, the technical support of Corinne Kim, Philip Lau, Evelyn Phillips, Gina Stouffer and Derek Wills for assistance in data collection and analysis, and Shirley Sanchez for assistance in manuscript editing.

[1]Power analysis was based on the specified p-value thresholds of 3.9e–6 and 3.0e–7 for the protein-coding and regulatory element set-based tests, respectively, a region of 4000 bps, the assumption that 50% of variants were causal with 10% of causal variants exhibiting a protective effect, and a maximum Odds Ratio for a single variant of 5.0.

References

Agrawal A, Lynskey MT, Bucholz KK, Kapoor M, Almasy L, Dick DM, Edenberg HJ, Foroud T, Goate A, Hancock DB, Hartz S, Johnson EO, Hesselbrock V, Kramer JR, Kuperman S, Nurnberger JI Jr, Schuckit M, Bierut LJ. DSM-5 cannabis use disorder: A phenotypic and genomic perspective. Drug Alcohol Depend. 2014; 134:362–369. [PubMed: 24315570]

Agrawal A, Lynskey MT, Hinrichs A, Grucza R, Saccone SF, Krueger R, Neuman R, Howells W, Fisher S, Fox L, Cloninger R, Dick DM, Doheny KF, Edenberg HJ, Goate AM, Hesselbrock V, Johnson E, Kramer J, Kuperman S, Nurnberger JI Jr, Pugh E, Schuckit M, Tischfield J, Rice JP, Bucholz KK, Bierut LJ. Consortium G. A genome-wide association study of DSM-IV cannabis dependence. Addict Biol. 2011; 16:514–518. [PubMed: 21668797]

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19:1655–1664. [PubMed: 19648217]

Barry G, Briggs J, Vanichkina D, Poth E, Beveridge N, Ratnu V, Nayler S, Nones K, Hu J, Bredy T, Nakagawa S, Rigo F, Taft RJ, Cairns MJ, Blackshaw S, Wolvetang EJ, Mattick JS. The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. Mol Psychiatry. 2014; 19:486–494. [PubMed: 23628989]

Bizon C, Spiegel M, Chasse SA, Gizer IR, Li Y, Malc EP, Mieczkowski PA, Sailsbery JK, Wang X, Ehlers CL, Wilhelmsen KC. Variant calling in low-coverage whole genome sequencing of a Native American population sample. BMC Genomics. 2014; 15:85. [PubMed: 24479562]

Bucholz KK, Cadoret R, Cloninger CR, Dinwiddie SH, Hesselbrock VM, Nurnberger JI Jr, Reich T, Schmidt I, Schuckit MA. A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. J Stud Alcohol Drugs. 1994; 55:149–158.

Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am J Hum Genet. 2010; 86:6–22. [PubMed: 20074509]

Dejean de la Bâtie C, Barbier V, Valayannopoulos V, Touati G, Maltret A, Brassier A, Arnoux J, Grevent D, Chadefaux B, Ottolenghi C, Canouï P, de Lonlay P. Acute Psychosis in Propionic Acidemia 2 Case Reports. J Child Neurol. 2014; 29:274–279. [PubMed: 24334345]

Edwards AC, Aliev F, Wolen AR, Salvatore JE, Gardner CO, McMahon G, Evans DM, Macleod J, Hickman M, Dick DM, Kendler KS. Genomic influences on alcohol problems in a population-based sample of young adults. Addiction. 2015; 110:461–470. [PubMed: 25439982]

Ehlers CL, Gizer IR. Evidence for a genetic component for substance dependence in Native Americans. Am J Psychiatry. 2013; 170:154–164. [PubMed: 23377636]

Ehlers CL, Slutske WS, Lind PA, Wilhelmsen KC. Association between single nucleotide polymorphisms in the cannabinoid receptor gene (CNR1) and impulsivity in southwest California Indians. Twin Res Hum Genet. 2007; 10:805–811. [PubMed: 18179391]

Ehlers CL, Wall TL, Betancort M, Gilder DA. Clinical course of alcoholism in 243 Mission Indians. Am J Psychiatry. 2004; 7(1):1204–1210.

Flavell SW, Cowan CW, Kim T-K, Greer PL, Lin Y, Paradis S, Griffith EC, Hu LS, Chen C, Greenberg ME. Activity-dependent regulation of MEF2 transcription factors suppresses excitatory synapse number. Science. 2006; 311:1008–1012. [PubMed: 16484497]

Gelernter J, Kranzler HR, Sherva R, Koesterer R, Almasy L, Zhao H, Farrer LA. Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. Biol Psychiatry. 2014a; 76:66–74. [PubMed: 24143882]

Gelernter J, Sherva R, Koesterer R, Almasy L, Zhao H, Kranzler H, Farrer L. Genome-wide association study of cocaine dependence and related traits: FAM53B identified as a risk gene. Mol Psychiatry. 2014b; 19:717–723. [PubMed: 23958962]

Haughey HM, Marshall E, Schacht JP, Louis A, Hutchison KE. Marijuana withdrawal and craving: influence of the cannabinoid receptor 1 (CNR1) and fatty acid amide hydrolase (FAAH) genes. Addiction. 2008; 103:1678–1686. [PubMed: 18705688]

Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. Soc Probl. 1997; 44:174–199.

Hasin DS, Saha TD, Kerridge BT, Goldstein RB, Chou SP, Zhang H, Jung J, Pickering RP, Ruan WJ, Smith SM, Huang B, Grant BF. Prevalence of marijuana use disorders in the United States between 2001–2002 and 2012–2013. JAMA Psychiatry. 2015; 72:1235–1242. [PubMed: 26502112]

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4:44–57. [PubMed: 19131956]

Johnson MB, Kawasawa YI, Mason CE, Krsnik Ž, Coppola G, Bogdanović D, Geschwind DH, Mane SM, State MW, Šestan N. Functional and evolutionary insights into human brain development through global transcriptome analysis. Neuron. 2009; 62:494–509. [PubMed: 19477152]

Kang, HM. Efficient and Parallelizable Association Container Toolbox (EPACT). University of Michigan Center for Statistical Genetics; 2014. Available at: http://genome.sph.umich.edu/wiki/EPACTS [Accessed 06.13.16]

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–354. [PubMed: 20208533]

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Team ELP, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012a; 91:224–237. [PubMed: 22863193]

Lee Y, Gamazon ER, Rebman E, Lee Y, Lee S, Dolan ME, Cox NJ, Lussier YA. Variants affecting exon skipping contribute to complex traits. PLoS Genet. 2012b; 8:e1002998. [PubMed: 23133393]

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27:2987–2993. [PubMed: 21903627]

Libiger O, Schork NJ. A Method for Inferring an Individual's Genetic Ancestry and Degree of Admixture Associated with Six Major Continental Populations. Front Genet. 2013; 3:322. [PubMed: 23335941]

Liu J, Wu X, Wang X, Zhang Y, Bu P, Zhang Q, Jiang F. Global gene expression profiling reveals functional importance of SIRT2 in endothelial cells under oxidative stress. Int J Mol Sci. 2013; 14:5633–5649. [PubMed: 23478437]

Moyer RA, Wang D, Papp AC, Smith RM, Duque L, Mash DC, Sadee W. Intronic polymorphisms affecting alternative splicing of human dopamine D2 receptor are associated with cocaine abuse. Neuropsychopharmacology. 2011; 36:753–762. [PubMed: 21150907]

Muhib FB, Lin LS, Stueve A, Miller RL, Ford WL, Johnson WD, Smith PJ. A venue-based method for sampling hard-to-reach populations. Public Health Rep. 2001; 116(Suppl 1):216–222. [PubMed: 11889287]

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin C-F, Stevens C, Wang L-S, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH Jr, Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012; 485:242–245. [PubMed: 22495311]

Nelson SC, Doheny KF, Pugh EW, Romm JM, Ling H, Laurie CA, Browning SR, Weir BS, Laurie CC. Imputation-Based Genomic Coverage Assessments of Current Human Genotyping Arrays. G3 (Bethesda). 2013; 3:1795–1807. [PubMed: 23979933]

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010; 6:e1000888. [PubMed: 20369019]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]

Pulipparacharuvil S, Renthal W, Hale CF, Taniguchi M, Xiao G, Kumar A, Russo SJ, Sikder D, Dewey CM, Davis MM, Greengard P, Naim AC, Nestler EJ, Cowan CW. Cocaine regulates MEF2 to control synaptic and behavioral plasticity. Neuron. 2008; 59:621–633. [PubMed: 18760698]

Quinn HR, Matsumoto I, Callaghan PD, Long LE, Arnold JC, Gunasekaran N, Thompson MR, Dawson B, Mallet PE, Kashem MA, Matsuda-Matsumoto H, Iwazaki T, McGregor IS. Adolescent rats find repeated Δ9-THC less aversive than adult rats but display greater residual cognitive deficits and changes in hippocampal protein expression following exposure. Neuropsychopharmacology. 2008; 33:1113–1126. [PubMed: 17581536]

Rashid A, Cole C, Josselyn S. Emerging roles for MEF2 transcription factors in memory. Genes Brain Behav. 2014; 13:118–125. [PubMed: 23790063]

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, Jager PLD, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]

Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–427. [PubMed: 25056061]

Schor IE, Rascovan N, Pelisch F, Alló M, Kornblihtt AR. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. Proc Natl Acad Sci, USA. 2009; 106:4325–4330. [PubMed: 19251664]

Sherva R, Wang Q, Kranzler H, Zhao H, Koesterer R, Herman A, Farrer LA, Gelernter J. Genome-wide Association Study of Cannabis Dependence Severity, Novel Risk Variants, and Shared Genetic Risks. JAMA Psychiatry. 2016; 73:472–480. [PubMed: 27028160]

Stouffer, SA., Suchman, EA., DeVinney, LC., Star, SA., Williams, RM, Jr. The American Soldier, Studies in Social Psychology in World War II Vol.1: Adjustment during Army Life. Princeton University Press; Princeton NJ: 1949.

Stringer S, Minică CC, Verweij KJH, Mbarek H, Bernard M, Derringer J, van Eijk KR, Isen JD, Loukola A, Maciejewski DF, Mihailov E, van der Most PJ, Sánchez-Mora C, Roos L, Sherva R, Walters R, Ware JJ, Abdellaoui A, Bigdeli TB, Branje SJT, Brown SA, Bruinenberg M, Casas M, Esko T, Garcia-Martinez I, Gordon SD, Harris JM, Hartman CA, Henders AK, Heath AC, Hickie IB, Hickman M, Hopfer CJ, Hottenga JJ, Huizink AC, Irons DE, Kahn RS, Korhonen T, Kranzler HR, Krauter K, van Lier PAC, Lubke GH, Madden PAF, Mägi R, McGue MK, Medland SE, Meeus WHJ, Miller MB, Montgomery GW, Nivard MG, Nolte IM, Oldehinkel AJ, Pausova Z, Qaiser B, Quaye L, Ramos-Quiroga JA, Richarte V, Rose RJ, Shin J, Stallings MC, Stiby AI, Wall TL, Wright MJ, Koot HM, Paus T, Hewitt JK, Ribasés M, Kaprio J, Boks MP, Snieder H, Spector T, Munafò MR, Metspalu A, Gelernter J, Boomsma DI, Iacono WG, Martin NG, Gillespie NA, Derks EM, Vink JM. Genome-wide association study of lifetime cannabis use based on a large meta-analytic sample of 32 330 subjects from the International Cannabis Consortium. Transl Psychiatry. 2016; 6:e769. [PubMed: 27023175]

Sun L, Wilder K, McPeek MS. Enhanced pedigree error detection. Hum Heredity. 2002; 54:99–110. [PubMed: 12566741]

Tyndale RF, Payne JI, Gerber AL, Sipe JC. The fatty acid amide hydrolase C385A (P129T) missense variant in cannabis users: Studies of drug use and dependence in caucasians. Am J Med Genet B Neuropsychiatr Genet. 2007; 144B:660–666. [PubMed: 17290447]

Verweij KJ, Zietsch BP, Lynskey MT, Medland SE, Neale MC, Martin NG, Boomsma DI, Vink JM. Genetic and environmental influences on cannabis use initiation and problematic use: a meta-analysis of twin studies. Addiction. 2010; 105:417–430. [PubMed: 20402985]

Visscher PM, Goddard ME, Derks EM, Wray NR. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. Mol Psychiatry. 2012; 17:474–485. [PubMed: 21670730]

Volkow ND, Baler RD, Compton WM, Weiss SR. Adverse health effects of marijuana use. New Eng J Med. 2014; 370:2219–2227. [PubMed: 24897085]

Volkow ND, Swanson JM, Evins AE, DeLisi LE, Meier MH, Gonzalez R, Bloomfield MAP, Curran HV, Baler R. Effects of cannabis use on human behavior, including cognition, motivation, and psychosis: a review. JAMA Psychiatry. 2016; 73:292–297. [PubMed: 26842658]

Wray NR, Maier R. Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability. Curr Epidemiol Rep. 2014; 1(4):220–227.

Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, Koebbe BC, Nielsen C, Hirst M, Farnham P, Kuhn RM, Zhu J, Smirnov I, Kent WJ, Haussler D, Madden PAF, Costello JF, Wang T. The human epigenome browser at Washington University. Nat Methods. 2011; 8:989–990. [PubMed: 22127213]

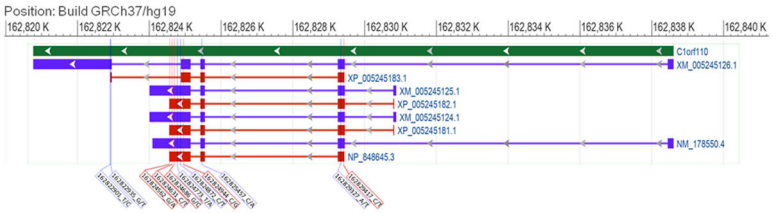

Figure 1.

Diagram of *C1orf110*, in green, and it's sequenced transcripts in alternating red and blue and labeled by their NCBI Reference Sequence IDs. The location of the analyzed coding variants are depicted below in red if they were predicted to impact protein function and blue if they were predicted to have no impact by the Provean, SIFT, and Polyphen tools.

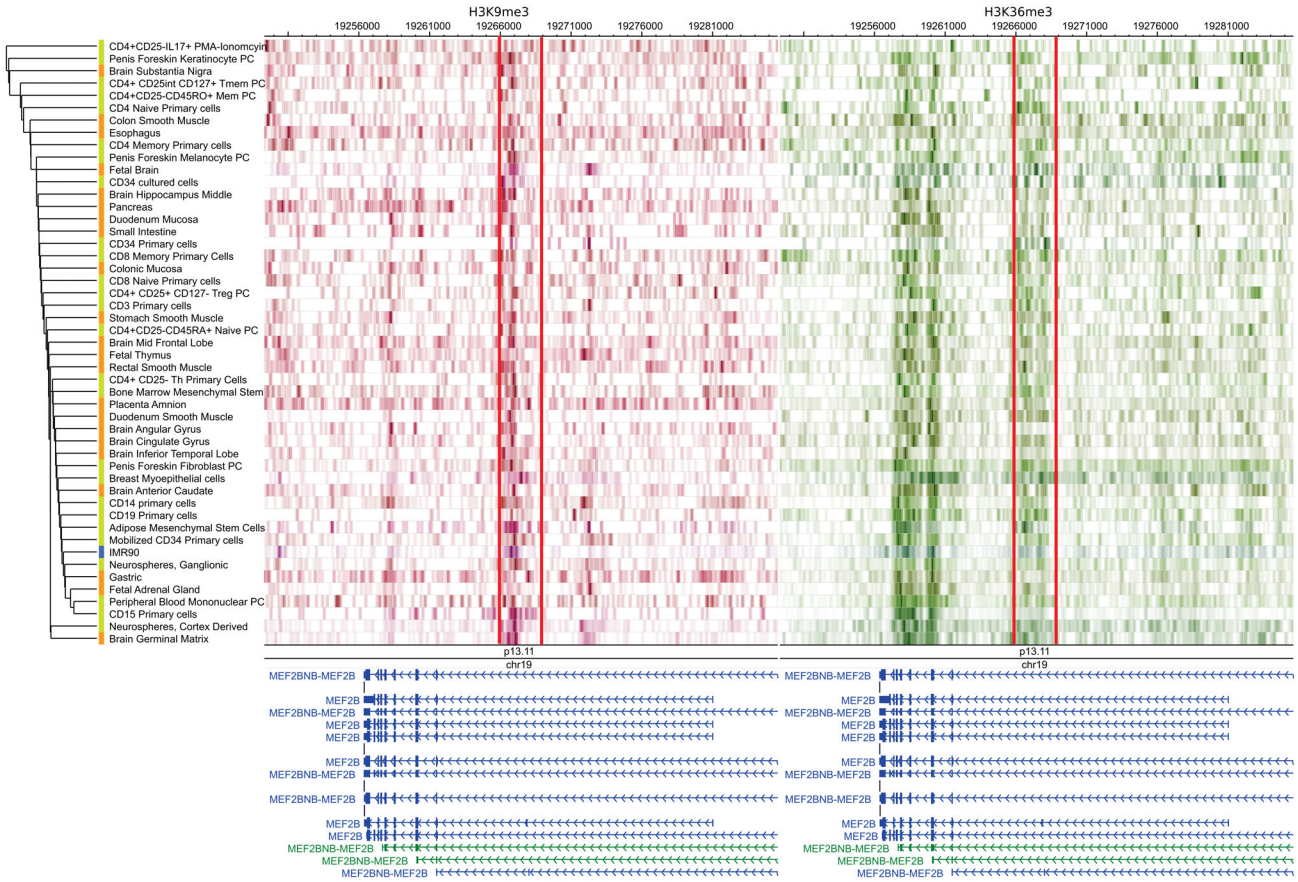

Figure 2.

Depiction of H3K9me3 (left) and HK36me3 (right) histone marks in the *MEF2B* region across tissue types included in the NIH Epigenomics Roadmap project. The y-axis indicates the tissue studied. The x-axis indicates physical position, and the illustrations below are gene diagrams based on observed transcripts of the respective genes. Shaded areas indicate increased presence of the respective histone mark. Vertical red lines indicate the boundaries of the associated regulatory element. PC = primary cells. Image was generated using the WashU Epigenome Browser (Zhou et al., 2011).

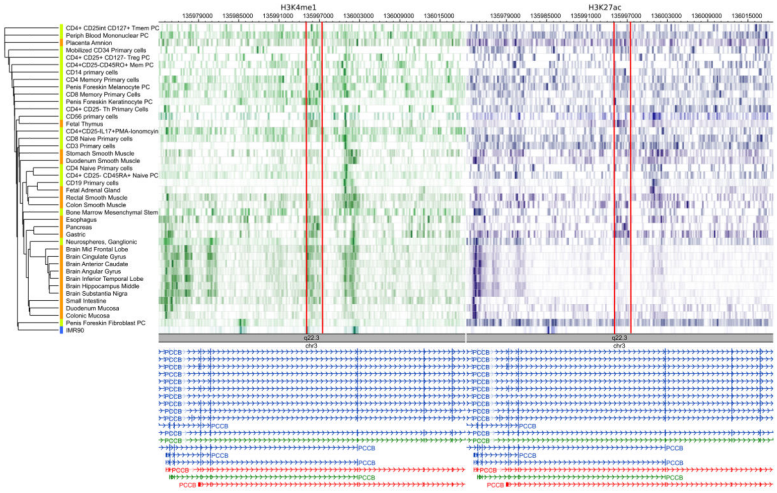

Figure 3.

Depiction of H3K4me1 (left) and HK27ac (right) histone marks in the *PCCB* region across tissue types included in the NIH Epigenomics Roadmap project. The y-axis indicates the tissue studied. The x-axis indicates physical position, and the illustrations below are gene diagrams based on observed transcripts of the respective genes. Shaded areas indicate increased presence of the respective histone mark. Vertical red lines indicate the boundaries of the associated regulatory element. PC = primary cells. Image was generated using the WashU Epigenome Browser (Zhou et al., 2011).

Table 1

Demographic and phenotypic characteristics of study participants with available sequence data[1].

	Native American Sample			UCSF Family Study Sample		
	Cannabis Dependence Diagnosis[1]			Cannabis Dependence Diagnosis[1]		
	No (n=529)	Yes (n=168)	Total (n=697)	No (n=1591)	Yes (n=241)	Total (n=1832)
Age (yrs ±SE)	32.4 ±0.61	28.1 ±0.78	31.2 ±0.5*	50.3 ±13.0	40.5±10.2	49.0±13.1*
Income >$20,000	254 (52%)	88 (57%)	342 (53%)	1257 (79%)	177 (73%)	1434 (78%)*
Female Gender	321 (61%)	78 (46%)	399 (57%)*	1012 (64%)	122 (51%)	1134 (62%)*
Married	91 (17%)	23 (14%)	114 (16%)	932 (59%)	122 (51%)	1054 (58%)*

Notes:

* - indicates significant difference between diagnostic groups ($p<0.05$);

1 - cells indicate number of participants and percent of sample in parantheses unless otherwise indicated.

Table 2

Association Results (p<0.001) for Set-Based Analysis of Coding Regions and Cannabis Dependence.

				p-value		
Gene	**Chrom.**	**Gene Boundaries**	**Variants Tested UCSF/NA**	**UCSF Family Study**	**Native American Sample**	**Meta-Analysis**
C1orf110	1	162822935-162829417	8/5	6.11E-06	0.066122	3.07E-06
MFAP3	5	153429301-153529504	12/4	0.001862	0.000445	1.33E-05
CRLF1	19	18709272-18710603	8/4	0.000403	0.045692	9.17E-05
ASB16	17	42248169-42255679	20/4	0.000824	0.022457	9.54E-05
RBM15B	3	51428887-51431437	5/8	0.000468	0.050031	0.000115
ITGA3	17	48141548-48166489	27/11	1.79E-06	0.7115	0.000117
DHDH	19	49436979-49447753	12/9	0.000473	0.05186	0.000121
HSPB7	1	16342103-16345866	8/5	5.13E-05	0.26399	0.000131
TCERG1L	10	132891438-133109602	15/9	0.001679	0.021181	0.000185
TAS2R46	12	11214145-11214654	5/2	0.001074	0.038906	0.000201
FAM117A	17	47788710-47799922	12/4	0.001986	0.021115	0.000219
CD1D	1	158151505-158153771	4/1	0.001917	0.022631	0.000224
NFYA	6	41046870-41060704	6/3	3.33E-06	0.75652	0.000237
ABCD3	1	94924174-94980750	9/3	0.0002	0.195	0.000256
GSX2	4	54966667-54968026	2/1	3.85E-05	0.45369	0.000286
NR1D1	17	38250184-38253469	16/6	0.001991	0.032335	0.000313
CSN3	4	71114793-71115159	3/1	1.32E-05	0.65071	0.00034
PRPF38A	1	52876825-52879611	3/1	0.001529	0.051599	0.000365
PM20D1	1	205799428-205819104	12/8	0.000827	0.096075	0.000379
IL29	19	39787153-39789115	4/2	0.001188	0.09609	0.000528
OR14A16	1	247978135-247978996	12/4	0.014232	0.00362	0.000549
CAPN3	15	42652065-42703537	33/11	0.03713	0.000408	0.000559
LCOR	10	98708995-98715545	7/2	0.000331	0.25824	0.000578
KCNK17	6	39267315-39282083	11/5	0.002893	0.043493	0.000583
U2AF1L4	19	36234762-36236120	9/3	0.000417	0.27462	0.000764
KCTD15	19	34290622-34303802	2/8	0.000362	0.31795	0.000851
ZNF615	19	52496149-52510548	18/12	0.009874	0.014692	0.000899

Gene	Chrom.	Gene Boundaries	Variants Tested UCSF/NA	*p*-value			
				UCSF Family Study	Native American Sample	Meta-Analysis
PCDHB3	5	140480275-140482620	14/11	0.009926	0.014658	0.000903
ZNF610	19	52848804-52870011	17/16	0.000718	0.22676	0.000918

Note. Chrom. = Chromosome.

Table 3

Associations of coding variants in *C1orf110* with Cannabis Dependence.

		UCSF Family Sample		Native American Sample		Functional Prediction Tools		
Variant	**rsNumber**	**Regression coefficient**	***p*-value**	**Regression coefficient**	***p*-value**	**PROVEAN***	**SIFT***	**Polyphen***
1:162822901_T/C	-			0.7241	0.08671	Neutral	-	Benign
1:162822935_G/T	rs547175693	0.08335	0.6747			Neutral	-	Benign
1:162824562_G/A	**rs187742957**	**0.162**	**0.01368**	**0.4956**	**0.01974**	**Deleterious**	**Deleterious**	**Probably damaging**
1:162824631_C/T	-			−0.2464	0.2522	Deleterious	Deleterious	Probably damaging
1:162824686_G/C	**-**	**0.5298**	**0.008009**			**Deleterious**	**Deleterious**	**Possibly damaging**
1:162824773_T/A	rs200349261	−0.137	0.5671			Neutral	Tolerated	Benign
1:162824872_C/T	rs142193912	−0.1838	0.5645			Neutral	Tolerated	Benign
1:162824944_C/G	rs200534408			−0.3181	0.4522	Deleterious	Deleterious	Possibly damaging
1:162825457_C/A	rs77227460	−0.01785	0.1913	0.05	0.23	Neutral	Tolerated	Benign
1:162829327_A/T	**rs78032080**	**0.7357**	**0.02298**			**Neutral**	**Tolerated**	**Benign**
1:162829417_C/T	**-**	**0.3475**	**0.014**			**Neutral**	**Tolerated**	**Possibly damaging**

Note: Bold text indicates variant that showed significant association with cannabis dependence using the single variant EMMAX test.

*- Terms in the cells below are the categorical predictions for each variant as defined by the authors of the corresponding algorithms.

Table 4

Association Results (p<0.001) for Set-Based Analysis of Regulatory Elements and Cannabis Dependence.

					p-value		
Chrom.	**Element Boundaries**	**Element Type**	**Nearest Gene**	**Variants Tested UCSF/NA**	**UCSF Family Study**	**Native American Sample**	**Meta-Analysis**
19	19266067-19269105	E/D	*MEF2B*	24/12	0.000026	0.0002	5.7e-08
3	135995054-135997390	E	*PCCB*	25/11	0.00000042	0.18494	1.6e-06
16	1908322-1911141	E	*MEIOB*	27/10	0.000004	0.05427	1.7e-06
3	90144843-90145482	E	Intergenic	7/4	0.000017	0.05421	6.1e-06
12	105745801-105747050	E	*C12orf75*	10/8	0.00016	0.00957	8.6e-06
12	21566456-21567243	E	Intergenic	5/3	0.0000615	0.050015	1.8e-05
16	34794333-34797676	E	Intergenic	27/17	0.001908	0.000839	1.96e-05
15	50435768-50441669	E/D	Intergenic	51/31	0.000103	0.036489	2.06e-05
3	69026981-69028664	E	*EOGT*	19/9	0.000112	0.035438	2.17e-05
7	37623970-37625039	E/D	Intergenic	8/7	0.0000428	0.10857	3.29e-05
12	5187603-5187997	E	Intergenic	10/5	0.0000125	0.22099	3.3e-05
14	52744303-52746459	E	*PTGDR*	28/14	0.00034	0.018931	3.42e-05
21	16893242-16894247	E	Intergenic	14/6	0.00011	0.06582	4.13e-05
13	20287789-20288090	E	*PSPC1*	6/3	0.001255	0.004876	4.26e-05
16	69121167-69127058	E/D	*TANGO6*	44/26	0.00391	0.000898	4.75e-05
5	143203591-143208616	E	*HMHB1*	39/25	0.000118	0.070201	4.76e-05
17	11085628-11086208	E	Intergenic	4/4	0.000607	0.014725	4.87e-05
18	12237289-12239721	E	Intergenic	24/12	0.000147	0.061474	4.97e-05
5	143191835-143192358	E	*HMHB1*	8/3	0.000991	0.008582	5.15e-05
5	153681692-153682998	E	*GALNT10*	9/4	0.000315	0.040161	6.48e-05
12	104212852-104215823	E	*NT5DC3*	19/10	0.000221	0.055937	6.53e-05
12	57664823-57667010	E	*R3HDM2*	16/6	0.00000166	0.61467	7.11e-05
7	92855686-92857842	P/E	*HEPACAM2*	16/9	0.0000259	0.25638	7.57e-05
2	61220032-61220394	D	*PUS10*	3/2	0.000856	0.017115	7.82e-05
5	73742592-73743242	E	Intergenic	5/2	0.0000747	0.15132	8.21e-05
1	204146269-204149731	E	Intergenic	38/24	0.0000512	0.19512	8.58e-05
4	20831675-20834764	E	*KCNIP4*	28/16	0.0000754	0.15723	8.71e-05

					p-value		
Chrom.	**Element Boundaries**	**Element Type**	**Nearest Gene**	**Variants Tested UCSF/NA**	**UCSF Family Study**	**Native American Sample**	**Meta-Analysis**
12	130158486-130161515	E	*TMEM132D*	30/13	0.000112	0.12168	8.73e-05
5	4700365-4704264	E/D	Intergenic	22/21	0.0000294	0.263	8.76e-05
5	135739166-135743009	E	Intergenic	38/8	0.002669	0.004284	8.8e-05
1	162824155-162827958	E	*C1orf110*	39/16	0.001827	0.008626	9.81e-05
9	124576924-124578213	E	Intergenic	14/9	0.000152	0.10766	9.81e-05

Note: Chrom. = Chromosome, D = Dyadic, E = Enhancer, P = Promoter.