THE LANCET
Oncology

## Supplementary appendix

Supplement to: Johnson EC, Demontis D, Thorgeirsson TE, et al. A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Oncol* 2020; published online Oct 20. http://dx.doi.org/10.1016/S2215-0366(20)30339-4.

# Appendix

**Contents:**

# 1 Cohort descriptions

## 1.1 Case-control cohorts

### *iPSYCH*

*Sample description:* The goal of the iPSYCH data collection was to study the genetic underpinnings of 6 psychiatric disorders[1] (schizophrenia, bipolar disorder, major depressive disorder, ADHD, anorexia and autism spectrum disorder) and individuals were selected from a birth cohort comprising individuals born in Denmark between May 1, 1981, and December 31, 2005, who were residents in Denmark on their first birthday and who have a known mother (N = 1,472,762). Psychiatric cases were identified based on information in the Danish Psychiatric Central Research Register[2], and 30,000 randomly selected controls were identified from the same nationwide birth cohort in the Danish Civil Registration System. A biological sample from the study individuals were identified in the Danish New Born Screening Biobank[3], that contains blood spot samples from nearly all babies born in Denmark since 1981. DNA collection and genotyping was done as described previously[1,4,5].
The study was approved by the Danish Research Ethics Committee and the Danish Data Protection Agency.

*Cannabis abuse/dependence measure:* CUD cases were defined using ICD10 codes (F12.1-12.2) through information in the Danish Psychiatric Central Research Register as well as in the Danish National Patient Register using information up to 2016. Controls were individuals who did not have ICD10 codes related to CUD.

### *deCODE*

*Sample description:* Cases were drawn from the largest addiction treatment center in Iceland, the SAA-National Center of Addiction Medicine[6]. Controls were recruited as part of various genetic research programs at deCODE Genetics. Individuals diagnosed with substance use/abuse were excluded from controls. The deCODE Genetics study was approved by the Data Protection Commission of Iceland and the National Bioethics Committee of Iceland.

*Cannabis abuse/dependence measure:* Cannabis use disorder diagnoses in this treatment cohort were made, in the years 1977 – 2014, using DSM-IIIR, DSM-IV and DSM-5 criteria, by clinicians using the Diagnostic and Statistical Manual of Mental Disorders (DSM) system.

### *Comorbidity and Trauma Study (CATS)*

*Sample description:* This study consisted of opioid dependent individuals aged 18 and older recruited from opioid substitution therapy clinics in the greater Sydney area and genetically unrelated individuals with little or no lifetime opioid misuse from neighborhoods in geographic proximity to these clinics. All subjects were of European-Australian descent. Additional details are available in [7].

*Cannabis abuse/dependence measure:* All participants were assessed using a version of the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA). Cannabis abuse and dependence were defined using DSM-IV criteria. For the purposes of these analyses, controls were defined as anyone who did not meet criteria for cannabis abuse or dependence. No other comorbid diagnoses were excluded.

### *Christchurch Health and Development study (CHDS)*

*Sample description:* The Christchurch Health and Development study (CHDS)[8,9] is a longitudinal study of a birth cohort of 1,265 children collected in mid-1977 from urban Christchurch, New Zealand. Data on social circumstances, health, development and wellbeing of the participants was obtained from the cohort at birth, four months, one year, annually to age 16 years, and at 18, 21, 25, 30, and 35 years. All study information was collected on the basis of signed consent from study participants and all information is fully confidential. All aspects of the study have been approved by the Canterbury (NZ) Ethics Committee.

*Cannabis abuse/dependence measure*: At ages 18, 21, 25, 30 and 35 years cohort members were questioned about their substance use behaviours and problems associated with substance use since the previous assessment (alcohol, tobacco, cannabis, other illicit drugs), using the relevant sections of the Composite International Diagnostic Interview (CIDI) to assess DSM-IV symptom criteria for substance use disorders. Using this information, lifetime cannabis abuse or dependence was classified on the basis of whether the participant met DSM criteria for cannabis abuse or dependence at any assessment up to age 35.

### *Study of Addiction: Genetics and Environment (SAGE), Collaborative Genetic Study of Nicotine Dependence (COGEND) & Family Study of Cocaine Dependence (FSCD)*

*Sample description:* Subjects for the Study of Addiction: Genetics and Environment (SAGE) were selected from three large, complementary studies: COGA[10], Family Study of Cocaine Dependence (FSCD)[11], and the Collaborative Genetic Study of Nicotine Dependence (COGEND)[12]. We analyze these subsets separately and remove overlap between cohorts (Supplementary Methods). COGA participants were assessed using the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA). FSCD and COGEND participants were assessed using polydiagnostic instruments closely based on the SSAGA. Genotyping was conducted using the Illumina Human1Mv1_C BeadChips. Further details of the SAGE samples are available in [13].

*Cannabis abuse/dependence measure:* Cases reported a lifetime history of DSM-IV cannabis abuse or dependence. Genetically unrelated control subjects did not meet criteria for a diagnosis of cannabis abuse or dependence.

### *Gene-Environment-Development Initiative (GEDI) – Duke University (GSMS)*

*Sample description:* The Duke arm of the NIDA-funded Gene-Environment-Development Initiative (GEDI) combined existing phenotypic and environmental data from two large prospective studies, the Great Smoky Mountains Study (GSMS) and the Caring for Children in the Community (CCC) study. For each of the two population-based contributing studies, genome-wide genotyping was conducted using a common platform (Illumina Human660W-Quad v1), generating a total genotyped sample of ~1300 subjects. Further details of the GEDI-Duke sample are available in [14,15].

*Cannabis abuse/dependence measure:* Participants of both studies were assessed via structured interviewing using the Young Adult Psychiatric Assessment and its early life extension (i.e., YAPA and CAPA), yielding diagnoses and symptom scales for a wide range of substance use disorders (SUDs). Cannabis abuse or dependence was defined using DSM-IV criteria. For the purposes of these analyses, controls were defined as anyone who did not meet criteria for cannabis abuse or dependence. No other comorbid diagnoses were excluded.

### *Center on Antisocial Drug Dependence (CADD)*

*Sample description:* The sample of 1,901 unrelated adolescents was aggregated from several studies described elsewhere[16–19]. This cohort was over-selected for adolescent behavioral disinhibition, with half of the participants ascertained specifically from high-risk populations (i.e. recruited through substance abuse treatment, special schools, or involvement with the criminal justice system; see supplement of [20] for additional criteria for clinical probands). CADD GWAS participants were an average age of 16.5 (SD = 1.4, range = 13.0–19.9), 28.9% were female, and 37.3% of participants reported non-Caucasian ancestry.

*Cannabis abuse/dependence measure:* Lifetime cannabis abuse or dependence was assessed with the CIDI-SAM and defined as meeting cannabis abuse or dependence at any wave for this longitudinal study.

### *Alcohol Dependence in African Americans (ADAA)*

*Sample description:* Data from "Alcohol Dependence in African Americans: A Case-Control Genetic Study" (ADAA) was funded by NIH grant R01 AA017444. The data were collected between 2009 and 2013 and consisted of cases recruited from treatment centers in St. Louis, Missouri and controls screened for the absence of alcohol use disorder recruited from households selected from neighborhoods in proximity to neighborhoods of residence of case participants.

*Cannabis abuse/dependence measure:* Cases met criteria for DSM-IV cannabis abuse or dependence. Controls were individuals who did not meet criteria for cannabis abuse or dependence (DSM-IV).

## 1.2 Family-based cohorts

### *Brisbane Longitudinal Twin Study (BLTS)*

*Sample description:* Beginning in 1992, the Brisbane Longitudinal Twin Study (BLTS) consists of 3,561 individuals: 1,422 twin pairs and 717 additional siblings first enrolled at age 12 years and now aged 30 years and older[21] (see also [22]). The sample is: genetically informative (MZ and DZ twins, and often parents and siblings; genotyped for 610,000 common single nucleotide polymorphisms - SNPs); (b) large; (c) longitudinal with many participants having been assessed at 12, 14, 16 and 21 years of age; (d) well characterized for behavioral and brain-related outcomes; (e) rich in biological samples; and includes (f) a subgroup [n=969] who have undergone MRI scanning. As part of an ongoing US NIH/NIDA funded project beginning 2009, measures of lifetime cannabis use, abuse and dependence data are collected, along with diagnostic data for nicotine, alcohol, and other illicit substances, as well as pilot epidemiological data for ecstasy and methamphetamine use. The average age at interview is 25.65 years (SD=3.65, range=18-38yrs). The entire BLTS sample and 1,549 of their parents have GWAS data (Illumina 610k chip)[23] imputed on the GRCh37 assembly.

*Cannabis abuse/dependence measure:* DSM-IV cannabis abuse and/or dependence was coded as either the endorsement of one or more abuse criteria and/or the endorsement of three or more dependence criteria. Controls were any individuals who did not meet these criteria.

### *Gene-Environment-Development Initiative (GEDI) – Virginia Commonwealth University (VTSABD)*

*Sample description:* The VCU arm of the NIDA-funded Gene-Environment-Development Initiative (GEDI) combined existing phenotypic and environmental data from the Virginia Twin Study of Adolescent Behavioral Development (VTSABD) study, a population-based multi-wave, cohort-sequential twin study of adolescent psychopathology and its risk factors, and two follow-up studies, the Young Adult Follow Up (YAFU) and the Transitions to Substance Abuse (TSA) study. For each of the contributing studies, genome-wide genotyping was conducted using a common platform (Illumina Human660W-Quad v1), generating a total genotyped sample of ~900 subjects. Further details of the GEDI-VCU sample are available in [14,24].

*Cannabis abuse/dependence measure:* Participants were assessed via structured interviewing using the Child Adult Psychiatric Assessment (CAPA), a Structured Clinical Interview for DSM-IV (SCID)-based assessment of psycho-pathology in young adult twins for YAFU and the Life Experiences Interview (LEI) for TSA, yielding diagnoses and symptom scales for a wide range of substance use disorders (SUDs). Cannabis abuse/dependence was defined using DSM-IV criteria. No comorbid diagnoses were excluded.

### *Collaborative Study on the Genetics of Alcoholism*

*Sample description:* COGA is a multi-site study of alcohol dependent probands and their family members. Alcohol dependent probands were recruited from inpatient and outpatient facilities. Community probands and their family members were also recruited from a variety of sources. The full sample of 12,145 individuals were genotyped on four different genome-wide genotyping arrays[25]. Among these arrays, two to 127 samples were genotyped on at least two different arrays with pairwise concordance rates all > 99.18%. Due to the complex family structure and the reconstruction of pedigrees that occurred after Mendelian checks were performed, COGA data were not subjected to re-imputation using the Picopili pipeline[26].

*Cannabis abuse/dependence measure:* All participants were assessed using the Semi-Structured Assessment for the Genetics of Alcoholism[27,28]. Cases met criteria for a lifetime history of DSM-IV cannabis abuse or dependence. Individuals, including related and unrelated subjects, not meeting criteria for cannabis abuse/dependence were included as controls.

***Minnesota Center for Twin and Family Research (MCTFR)***

*Sample description:* The MCTFR is a community-based longitudinal sample including pedigrees designed to include two rearing parents and two offspring[29]. Assessments across subsets of the study varied but were readily harmonized to DSM-IIIR and DSM-IV diagnoses. As part of the Genes Environment Development Initiative (GEDI), genotyping was carried out using the Illumina Human660W-Quad array. The final GWAS sample included 1,631 genotyped spouse pairs and 1,404 families with genotyped parents and offspring (at least one).[30]

*Cannabis abuse/dependence measure:* Cases met criteria for DSM-IIIR cannabis abuse or dependence.

***Center for Education and Drug Abuse Research (CEDAR) – Substance Abuse and the Dopamine System Study (SADS)***

*Sample description:* Participants were recruited from the Pittsburgh, Pennsylvania, metropolitan area through newspaper advertisements, social service agencies, substance abuse treatment programs and various other media. For this project, the sample is drawn from two combined studies with distinct but related ascertainment schemes, from the same Greater Pittsburgh population, joined in the Substance Use Disorder Liability: Candidate System Genes study (R01 DA019157)[31]. CEDAR (P50 DA005605) is a longitudinal family/high-risk study of substance use disorder (SUD)[32]. Parents from a sample of nuclear families, ascertained in CEDAR through the father who did or did not have a DSM-III-R SUD (DSM-IV was introduced after this study started) related to illicit drugs (an illegal substance or nonmedical use of a prescribed psychoactive drug), provided a source for male and female cases and controls. All diagnoses have been revised using DSM-IV criteria; the SADS participants were also diagnosed accordingly. Control subjects had no substance (including alcohol) use disorder, or Axis I or II psychiatric disorder. Participants from the SADS study (R01 DA011922) were males 14-18 years of age having a DSM-IV diagnosis of substance dependence related to use of illicit drugs. In both CEDAR and SADS subsamples, probands having a psychiatric disorder other than SUD qualified for the study unless they had a lifetime history of psychosis or any other condition where valid reporting was uncertain. The vocabulary subscale of WISC-III (subjects below age 16) or WAIS-III (age 16 and older) was administered prior to implementation of the protocol and was required to be in the normal range (>70). Since psychiatric comorbidity is common among substance abusers, cases were not excluded for any Axis I or Axis II disorders. The CEDAR and SADS subjects were self-identified European-Americans from the same Greater Pittsburgh geographic area, and the genomic inflation factor based on all genotyped SNPs, evaluating the excess false-positive rate, was satisfactory at .9812.

*Cannabis abuse/dependence measure:* Lifetime cannabis abuse and dependence were diagnosed using an expanded version of the Structured Clinical Interview for DSM-III-R-outpatient version (SCID-OP). Controls did not meet criteria for cannabis abuse or dependence.

### *Yale-Penn 1*

*Sample description:* Yale-Penn subjects were recruited in the eastern US, predominantly in Connecticut and Pennsylvania. They were administered the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA)[33] to derive DSM-IV diagnoses of lifetime alcohol and drug dependence (and other major psychiatric traits). The study received IRB approval from all participating institutions and written informed consent was obtained from all study participants. Additional information is available in the relevant GWAS publications (e.g. [34–37]).

*Cannabis abuse/dependence measure:* DSM-IV diagnoses from the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA).

### *Yale-Penn 2*

*Sample description:* Participants of Yale-Penn 2 were recruited and ascertained following the same protocol as Yale-Penn 1, described above, with a larger proportion of samples coming from unrelated individuals rather than families. Genotyping was performed using Illumina HumanCoreExome. Participants were grouped separately from Yale-Penn 1 based on the epoch of recruitment and the platform used for genotyping. Written informed consent was obtained from subjects as approved at each site by the respective institutional review boards, and certificates of confidentiality were obtained from NIDA and NIAAA.

*Cannabis abuse/dependence measure:* DSM-IV diagnoses of lifetime cannabis dependence were derived from the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA)[34].

### *Australian Alcohol and Nicotine Studies (OZ-ALC-NAG)*

*Sample description:* Participants were recruited from twins and their relatives who had participated in questionnaire- and interview-based studies on alcohol and nicotine use and alcohol-related events or symptoms (as described in [38]). They were living in Australia and of predominantly European ancestry.

*Cannabis abuse/dependence measure:* Assessed using DSM-IV criteria.

### *Irish Affected Sib Pair Study of Alcohol Dependence (IASPSAD)*

*Sample description:* Participants in the Irish Affected Sib Pair Study of Alcohol Dependence (IASPSAD)[39] were recruited in Ireland and Northern Ireland between 1998 and 2002. Briefly, probands were ascertained in community alcoholism treatment facilities and public and private hospitals. Probands were eligible for inclusion if they met DSM-IV criteria for lifetime alcohol dependence and if all four grandparents had been born in Ireland, Northern Ireland, Scotland, Wales, or England. Probands, siblings, and parents were interviewed by clinically trained research interviewers, most of whom had extensive clinical experience with alcoholism. We assessed lifetime history of alcohol and drug dependence using a modified version of the Semi-Structured Assessment of the Genetics of Alcoholism (SSAGA) interview, version II[27], demographic characteristics, other comorbid conditions, alcohol-related traits, personality features, and clinical records. All participants provided informed consent. We included 815 probands and siblings in genotyping. Controls were genotyped from 2,048 DNA samples from healthy, unpaid volunteers donating blood at the Irish Blood Transfusion Service and obtained from the Trinity College Biobank https://www.tcd.ie/ttmi/facilities/trinity-biobank/ at Trinity College Dublin. Biobank controls were eligible if they denied any problems with alcohol or history of mental illness and if all four grandparents had been born in Ireland, Northern Ireland, Scotland, Wales, or England. Information about age and sex was available for these subjects.

*Cannabis abuse/dependence measure:* DSM-IV criteria for lifetime cannabis abuse and dependence. Because of the sample source, controls were not formally screened for cannabis use.

### 1.3 Summary statistics cohorts

***National Longitudinal Study of Adolescent to Adult Health (Add Health)***

*Sample description:* The National Longitudinal Study of Adolescent to Adult Health (Add Health) is an ongoing, nationally-representative longitudinal cohort study of 20,000+ adolescents followed into adulthood for 20+ years across five interview waves from 1994-2018. Extensive longitudinal social, behavioral, environmental, and biological data are available, and the design included an embedded genetic subsample of MZ and DZ twins, full sibs, half sibs, and unrelated adolescents in the same household. Genome-wide data are available on 9,975 individuals using two Illumina platforms (Human Omni1-Quad BeadChip, Human Omni-2.5 Quad BeadChip) consisting of 631,990 SNPs. Add Health is a multiracial and multiethnic sample with substantial numbers of individuals with Hispanic and Asian ancestry. For more information about the design of Add Health see [40,41].

*Cannabis abuse/dependence measure:* Lifetime DSM-IV cannabis abuse and dependence was assessed using a questionnaire modeled on the Composite-International Diagnostic Interview, Substance Abuse Module (CIDI-SAM).

## 2 Quality control

### 2.1 Case-control cohorts

Quality control (QC) was performed separately for each case/control cohort using ricopili[42] (https://github.com/Nealelab/ricopili).

Following the standardized ricopili pipeline, variants in each cohort were first filtered for call rate (<5% missingness), followed by individual-level filters for call rate (<2% missingness) and heterozygosity ($|F_{het}| > .20$). If chromosome X variants were available for the cohort, sex checks were also performed to ensure concordance with reported sex. Variants were then filtered for call rate (<2% missingness), differential missingness between cases and controls (absolute difference < 2%), invariant markers, and departure from Hardy-Weinberg equilibrium in cases ($P > 1e-10$) or controls ($P > 1e-6$). In cohorts involving multiple genotyping batches, variants were also filtered for association with batch controlling for the phenotype.

Quality Control (QC) was performed prior to estimation of relatedness and principal components (described below). In cases of cryptic relatedness or ancestry outliers, QC was repeated after outlier removal to ensure no additional variants or individuals failed QC after removal of the affected individuals.

Stringent quality control was applied to the iPSYCH cohort and only samples with an individual call rate (>0.95) and genotypes with high call rate (>0.98), no strong deviation from Hardy–Weinberg equilibrium ($p > 1e{-}6$ in controls or $p > 1e{-}10$ in cases) and low heterozygosity rates ($F_{het} < 0.2$) were included.

Quality control for the deCODE Genetics sample was performed as described previously[43].

## 2.2 Family-based cohorts

QC for family-based cohorts was performed using picopili[26] (https://github.com/Nealelab/picopili). This QC, imputation, and analysis pipeline was developed for the current analysis with the aim of paralleling the functionality of ricopili (https://github.com/Nealelab/ricopili) with appropriate modifications for the analysis of family-based GWAS cohorts.

QC of the family-based cohorts applied the same basic filters as the case/control QC pipeline (i.e. call rates, heterozygosity, discordant sex checks, differential missingness, and departure from Hardy-Weinberg equilibrium). Where applicable, tests were based on allele frequencies computed from founders in the family-based cohort using PLINK 1.9[44]. In addition, family-based cohorts were QCed to remove individuals or variants with excessive Mendelian error rates. After QC, remaining Mendelian errors were set to missing.

As in the case/control cohorts, QC was repeated after stratification by ancestry and removal of ancestry outliers and instances of cryptic relatedness.

For COGA, with its complex family structure and genotyping on different arrays, detailed QC are described in Lai et al. 2019[25]. Initial QC used a set of 47,000 high quality variants genotyped on all arrays to assess duplicate samples and confirm the pedigree structure. Family structures were altered as needed, and genotypes were checked for Mendelian inconsistencies using Pedcheck[45]; inconsistencies were set to missing. Based on the first two PCs, each individual was then assigned a race classification (AFR, EUR, and Other). Families were assigned a family-based race, according to the majority of individual-based race in that family. Following this QC, all samples were imputed to 1000 Genomes using the cosmopolitan reference panel (Phase 3, version 5, NCBI GRCh37) using SHAPEIT2[46] then Minimac3[47] within each array. Imputed variants with $R^2 < 0.30$ were excluded, and genotype probabilities were converted to genotypes if probabilities $\geq 0.90$. Pedcheck[45] was used again to detect and clean Mendelian inconsistences for imputed variants. All genotyped and imputed variants with missing rates <25%, MAF $\geq$ 1% and HWE p values > 1e-6 were included in family-based analyses.

## 2.3 Summary statistics cohorts

For cohorts contributing summary statistics, pre-imputation QC was performed by the respective studies according to their chosen analysis protocols. For Add Health, mismatches on heterozygosity and sex were removed but no additional sample filtering was conducted prior to imputation.

# 3 Principal components analysis and relatedness estimation

## 3.1 Case-control and family-based cohorts

Principal components analysis (PCA) and relatedness estimation were performed within each cohort using a more stringently QCed set of variants, as detailed in [26] (for PGC) and [48] (for iPSYCH). PCA and relatedness estimation for the family-based cohorts was performed using picopili (https://github.com/Nealelab/picopili).

After stratification by ancestry, the full ricopili pipeline of QC, relatedness estimation, and PCA was repeated within each ancestry stratum of each cohort. Remaining PCA outliers within each ancestry group were removed as necessary.

In iPSYCH, relatedness and population stratification were evaluated using a set of high-quality markers (genotyped autosomal markers with a $MAF > 0.05$, Hardy–Weinberg equilibrium $P > 1e{-4}$ and SNP call rate $> 0.98$), which were pruned for linkage disequilibrium ($r^2 < 0.075$) resulting in a set of 37,425 pruned markers. (Markers located in the long-range linkage disequilibrium regions defined by Price et al.[49] were excluded.) Genetic relatedness was estimated using PLINK v.1.9[44,50] to identify first- and second-degree relatives ($\pi > 0.2$); one individual was excluded from each related pair (cases preferred over controls). Genetic outliers were identified for exclusion based on principal component analysis (PCA) using EIGENSOFT 7.2.1.[51,52] A genetically homogenous sample was defined based on a subsample of individuals being Danes for three generations (identified based on register information about the birth country of the individuals, their parents and grandparents). The subsample of Danes was used to define the center based on the mean values of principal components 1 and 2. Subsequently, principal components 1 and 2 were used to define a genetically homogenous population by excluding individuals outside an ellipsoid with the axes greater than 6 SD from the mean. After outlier exclusion, PCA was redone and the principal components from this analysis were included in the association analysis.

Details regarding the deCODE sample are in 6.4.

## 3.2 Summary statistics cohorts

In Add Health, a genetic relationship matrix (GRM) was computed in GCTA[53] to account for admixture within specified ancestral groups.

# 4 Imputation

## 4.1 Case-control cohorts

Imputation of case/control cohorts was performed using ricopili[42] (https://github.com/Nealelab/ricopili) for case-control data.

Prior to imputation, each cohort was aligned to 1000 Genomes Project Phase 3 reference data[54,55]. LiftOver[56] to human genome build hg19 was performed if needed, and matching of chromosome, position, and alleles to the reference data was verified. To assist with strand flips and strand ambiguous SNPs, allele frequencies were also checked against 1000 Genomes reference data. For European ancestry cohorts, SNPs were excluded if their allele frequency difference by more than 0.15 from 1000 Genomes European ancestry individuals; for African ancestry individuals, SNPs were filtered for allele frequency differences greater than 0.25 compared to 1000 Genomes African ancestry individuals. The looser threshold was specified in African ancestry cohorts to account for varying degrees of admixture, and generally yielded higher quality imputation results (data not shown).

After alignment to the 1000 Genomes Project Phase 3 reference[55], each cohort was phased using SHAPEIT[57] and imputed using IMPUTE2[58,59]. Imputation dosages and best-guess genotypes were saved for analysis. PCA was performed within each cohort using best-guess genotypes to compute principal components (PCs) for use as covariates in GWAS following the same procedure described above. For this post-imputation PCA, best-guess genotypes were strictly filtered for quality (call rate > 99% for genotype calls with posterior probabilities > 0.8, MAF > 5%) and more stringently pruned for LD (pairwise $r^2 < 0.1$, and removal of additional previously-identified regions of high LD[49]).

In the iPSYCH cohort, genotypes were phased and imputed using phase 3 of the 1000 Genomes Project[54] imputation reference panel and SHAPEIT[46] and IMPUTE2[59].

For the deCODE Genetics cohort, variant imputation was performed based on the IMPUTE HMM model and long-range phasing, as described previously[43]

### 4.2 Family-based cohorts

Family-based cohorts were imputed using picopili[26] (https://github.com/Nealelab/picopili) paralleling the same procedure described above for case/control cohorts. (COGA was treated separately, as described above and in Lai et al. (2019)[25].) Each cohort was matched to the 1000 Genomes Project Phase 3[54] imputation reference data following the same set of heuristics as are implemented in ricopili. Pre-phasing and imputation were then performed with SHAPEIT[57] and IMPUTE2[58,59] with two primary changes to accommodate the family data. First, phasing was performed for each chromosome rather than in 3 MB genomic chunks in order to assist in identifying any long regions of haplotype sharing between family members. Second, the duoHMM algorithm in SHAPEIT[60] was enabled to allow use of pedigree information in refining haplotype calls.

After imputation, best-guess genotypes were called (minimum posterior probability > 0.8) and QCed for call rate (missingness < 2%), INFO score > 0.6, and allele frequency > 0.005. (Additional filtering was applied prior to meta-analysis, see below.) Any apparent Mendelian errors in the imputed pedigrees were set as missing. After QC, post-imputation PCA was then performed to compute PCs for use as covariates in the GWAS using the same protocol as the PCA performed in the family-based cohorts prior to imputation (see above).

### 4.3 Summary statistics cohorts

Add Health data were imputed using the Haplotype Reference Consortium[61] on the Michigan Imputation Server[47].

# 5 Cross-cohort relatedness and ancestry confirmation

After imputation, QCed best-guess genotypes from each cohort were merged to allow filtering for cryptic relatedness between cohorts. Imputed genotypes were filtered for allele frequency and imputation quality (i.e. INFO score, call rate at posterior probability > 0.80) within each cohort, and then merged and filtered to variants passing QC across cohorts. As in the within-cohort relatedness checks, the passing variants were then pruned for LD and used to estimate genetic relatedness between all pairs of individuals. Relatedness among EUR cohorts was estimated using PLINK[44], while relatedness with AFR cohorts was estimated using REAP[62] to account for varying admixture.

In cases of observed cross-cohort cryptic relatedness ($\pi > 0.1$), individuals were removed from each related pair as in the within-cohort relatedness filtering. In order to maximize effective sample size, priority was given to keeping individuals with a CUD diagnosis, individuals in cohorts with small sample sizes, and individuals who were part of a pedigree in a family-based study. Individuals with cryptic relatedness to a large number of other samples were prioritized for removal. Instances of known overlap between the cohorts (e.g. among the cohorts in SAGE) were also verified and filtered accordingly.

Unrelated individuals were also used to verify ancestry assignment of the EUR and AFR cohorts, respectively, by merging the cohorts of each ancestry with 1000 Genomes Project reference samples and performing PCA.

**Table 1** reports final sample sizes for analysis after filtering for cross-cohort relatedness. GWAS were performed separately in each cohort (and for EUR and AFR within a cohort) using the set of individuals who passed this relatedness check.

# 6 Genome-wide association

## 6.1 Case-control cohorts

Genome-wide association studies (GWAS) were performed in each case/control cohort using PLINK[44]. Logistic regression was performed to test association between CUD and the imputed additive dosage of each variant, controlling for sex and principal components (PCs).

The number of PCs included as covariates to control for confounding from population structure varied by ancestry and sample size. In EUR cohorts, the number of PC covariates was determined by cohort sample size in order to reflect differential power of PCA to detect true population structure[63]. Specifically, in EUR cohorts with fewer than 2000 samples or fewer than 500 cases, the first 5 PCs were included as covariates; larger cohorts included the first 10 PCs. The number of cases was included as a criterion to prevent over-fitting to PCs in large cohorts with strongly skewed case/control ratios.

In AFR cohorts, we included as covariates the top PCs associated with genome-wide population structure, as opposed to local ancestry tracts[26], up to a maximum of 5 or 10 PCs based on the same sample size thresholds as in EUR cohorts. In practice, this resulted in the use of between 1 and 5 PCs in each cohort.

The association analysis in the iPSYCH cohort was done using logistic regression and imputed marker dosages. The following covariates were used: principal components 1–4, and additionally one principal component from the PCA associated with case-control status; the 19 data processing waves; and diagnosis of major psychiatric disorders studied by iPSYCH. Results for 9,729,295 markers were generated; subsequently, markers with an imputation INFO score < 0.7 (n = 608,367), markers with a MAF < 0.01 (n = 10,220) and multiallelic markers (n = 143,083) were removed. In total, after filtering, 8,969,939 markers remained for further analysis. All analyses of the iPSYCH sample were performed at the secured national GenomeDK high-performance computing cluster in Denmark.

### 6.2 Family-based cohorts

GWAS was performed in each family-based cohort using best-guess imputed genotypes for each variant. The association model used to test association for each variant was selected based on the complexity of the pedigree structure in each cohort's family-based design. Cohorts with a simple pedigree structure were tested using generalized estimating equations (GEE). Cohorts with more complex pedigrees that performed poorly in the GEE model (e.g., COGA) were tested using generalized linear mixed models (GLMM). Both models are described in detail in [26]. Sex and PC covariates were included following the same protocol as described above for case/control cohorts.

*GWAS of Case-control Individuals*

In addition to the primary family-based analyses, a subset of unrelated individuals was selected from each family-based cohort to perform a conventional case/control GWAS. Unrelated individuals were chosen to maximize the effective sample size for case/control analysis for CUD within each cohort. GWAS was then performed using logistic regression with the best-guess imputed genotypes in PLINK[44]. Sex and PC covariates were included following the same protocol as the case/control GWAS, as described above.

### 6.3 Summary statistics cohorts

GWAS was performed within the Add Health cohort following standard protocols (analyzed using a mixed linear model association framework within GCTA[53,64] with sex as a covariate.)

### 6.4 Population stratification in deCODE

In the deCODE sample, population stratification was accounted for by dividing by an inflation factor estimated from LD score regression (LDSR)[65]. Price et al.[66] found that "the divergence time of Icelandic regions has been too short for differential selective forces to have had a significant impact on allele frequencies", and "A consequence of these findings is that whenever λ is close to 1 in a disease association study involving the Icelandic population, false positive associations due to population stratification can be conclusively ruled out. If λ is greater than 1, then dividing association statistics by λ will still prevent false positive associations." Because county of origin predicts more variance due to genetic drift than genetic PCs, this covariate was included in the logistic regression models in lieu of PCs.

## 7 Genome-wide meta-analysis

We performed two batches of primary meta-analyses. First, we perform meta-analysis of all samples (including related individuals and summary statistic cohorts). Second, we perform meta-analysis of case-control genotyped samples only (i.e. excluding family-based samples and summary statistic cohorts) within the PGC, meta-analyzing with iPSYCH and deCODE. The full set of meta-analysis designs is described in **Table S1**.

### 7.1 Meta-analysis with related samples

The primary discovery meta-analysis was performed using all available EUR-only samples, including related individuals and summary statistic cohorts (17,068, cases, 357,219 controls). These meta-analyses were performed using p-values with weights defined by the effective sample size of each cohort. These weights were defined to account for the differences in case/control balance and degree of relatedness within each cohort, while allowing meta-analysis without comparable effect size estimates from the GLMM or summary statistic cohorts.

For meta-analysis, results from each cohort were filtered for imputation INFO score (> 0.8), minor allele frequency (> 1%), and expected minor allele count (MAC) in cases and controls (> 5). In the PGC, the summary statistics from each sample were also filtered to only report results for variants present in an effective sample size > 1000 and > 15% of the maximum effective sample size for the meta-analysis. GWAS results from the AddHealth summary statistics cohort were filtered according to the same criteria after being aligned to match the same genomic reference as the genotyped cohorts (e.g. matching rsids, positions, and alleles), except for INFO score > 0.8, as this information was not available.

The final meta-analysis summary statistics were further filtered to only report results for variants present in an effective sample size > 1000 and > 15% of the maximum effective sample size for the meta-analysis. The summary statistics were also filtered such that a SNP had to be present in at least two of the three contributing GWAS (deCODE, iPSYCH, and PGC) for further annotation.

In addition to this primary meta-analysis within ancestries, meta-analysis was also performed across ancestries (20,916 cases, 363,116 controls). In the trans-ancestral meta-analysis, long, highly significant, likely false-positive indels in the deCODE summary statistics were excluded, but SNPs were not required to be present in at least two of the three samples, as only the PGC contained AFR-ancestry individuals.

### 7.2 Meta-analysis with case-control genotyped samples

Meta-analysis of case-control genotyped samples was performed using conventional inverse-variance weighted fixed effects meta-analysis in METAL[67]. This analysis excluded the summary statistic cohort (AddHealth) and restricted the family-based cohorts from the PGC to unrelated individuals only. Meta-analysis was performed for only the European (EUR) cohorts. Total sample sizes for this meta-analysis were 14,080 cases, 343,726 controls in EUR cohorts.

This analysis was primarily intended to provide estimates of variant effect sizes, as well as the computation of polygenic risk scores. This restricted set of samples is necessary for estimation of effect sizes because many of the summary statistic cohorts relied on GWAS with a linear rather than logistic link function and thus do not have comparable effect sizes to the genotyped cohorts, and because effects sizes are unavailable for the family-based cohorts with complex pedigrees analyzed using the GLMM score test.

## 8 FUMA annotation

In FUMA[68] v1.3.5e, we specified that the annotation should only include SNPs independent at $R^2 < 0.1$ as "independent significant SNPs", rather than the default parameter of $R^2 < 0.6$. We used the 1000 Genomes Phase 3 reference panel, and used data from BrainSpan[69], for gene mapping using MAGMA[70], data from PsychENCODE[71], Common Mind Consortium[72], BRAINEAC[73], and GTEx[74] for eQTL mapping, and PsychENCODE and Hi-C datasets from Giusti-Rodriguez et al.[75] for chromatin interaction mapping.

## 9 H-MAGMA analyses

Besides the classic MAGMA analyses conducted via FUMA, we also used an alternative approach, Hi-C coupled MAGMA[76] (H-MAGMA). H-MAGMA takes into account long-range regulatory interaction effects to assign non-coding SNPs (intergenic and intronic) to genes based on their chromatin interactions (exonic and promoter SNPs are still assigned to genes based on genomic location). Four Hi-C datasets were used: one for fetal brain tissue[77], one for adult brain tissue[78], one for iPSC-derived astrocytes[79], and one for iPSC-derived neurons[79] (all available for download: https://github.com/thewonlab/H-MAGMA).

## 10 Genetic correlation analyses

**Table S2** outlines the measures that were selected for genetic correlation analyses using LD Score Regression[65,80] (LDSR). LDSR analyses were performed using the subset of SNPs available in HapMap3 and LD scores that were computed using the 1000 Genomes Project Phase 3 reference panel for European populations (pre-computed scores available for download were taken from https://data.broadinstitute.org/alkesgroup/LDSCORE/). To minimize the multiple testing burden, we only examined 22 robustly heritable measures that have been previously associated with CUD from a phenotypic perspective. Within each domain, we prioritized for GWAS with higher heritability reflecting adequate power. For psychiatric disorders, schizophrenia, bipolar disorder, major depression, posttraumatic stress disorder and ADHD have been previously associated with cannabis use and CUD[81]. There has also been some evidence of an association between cannabis *use* and anorexia nervosa[82]; thus we included this disorder as well. Despite associations between CUD and anxiety disorders, summary statistics were not available for download from the current largest GWAS of anxiety, and previous studies were not well-powered enough for LDSR. For substance use and disorder, no other illicit drugs were included as those GWAS do not demonstrate adequate power[83]. We included alcohol use disorder[84], cigarettes per day[85], drinks per week[85], the Fagerström Test for Nicotine Dependence[83], and smoking initiation[85] (as well as a comparison with cannabis use[81]). The personality measure of risk tolerance is amongst the strongest phenotypic correlates of substance use[86]. We also included age at first birth, which is an index of reproductive tempo and has been associated with precocious engagement in sexual activity as well as experimentation with substances[87]. Several psychosocial indices were studied. We included correlations with body mass index, which has been repeatedly linked to cannabis use in epidemiological analyses with associations attributed to peripheral effects of THC on satiety and energy regulation[88]. Likewise, several epidemiological studies indicate sleep disruption in heavy cannabis users, hence we selected chronotype as the most well-powered index of sleep, and also because it was previously found to be associated with substance use, including cannabis use[89]. In addition, measures of socio-economic status, such as educational attainment and neighborhood deprivation, were also included as they have been repeatedly linked to addictions[26]. The summary statistics for Townsend Deprivation Index and age at first birth were taken from a phenome-wide analysis conducted on 361,194 genotyped individuals in the UK Biobank by the Neale lab (https://github.com/Nealelab/UK_Biobank_GWAS). Finally, as the present study examines the association between CUD and brain volume, we assessed the genetic correlation with intracranial volume as well as volume of the nucleus accumbens, caudate and putamen, all striatal regions that have been robustly implicated in the neural circuitry underpinning addictions, as well as the hippocampal region.

## 11 Latent causal variable analyses

We conducted latent causal variable (LCV) analyses on CUD and the top genetically correlated traits: educational attainment, age at first birth, Townsend Deprivation Index (TDI), smoking initiation, and ADHD. LCV is an approach related to Mendelian Randomization but potentially more robust in the presence of genetic correlation[102]. Additionally, LCV permits us to remove the bias of sample overlap among the GWAS datasets investigated.

## 12 Polygenic risk score analyses

### 12.1 UK Biobank

The UK Biobank is a national volunteer health resource which gathers genotypic and phenotypic data on a representative population of the United Kingdom[90,91]. Cannabis use was ascertained in the UK Biobank using an online mental health questionnaire[92] which was completed by 157,366 individuals over a 1-year period in 2017. Using a white British unrelated subset of the UK Biobank (removing third degree relatives or closer (using a kinship coefficient > 0.044)) Field 20453 "Have you taken CANNABIS (marijuana, grass, hash, ganja, blow, draw, skunk, weed, spliff, dope), even if it was a long time ago?" and field 20454 "Considering when you were taking cannabis most regularly, how often did you take it?" were used to create a numerical variable corresponding to 0 – Never used cannabis (N=85214), 1 – Used cannabis 1-2 times but not daily (N=9410), 2 – Used cannabis 3-10 times but not daily (N=5566), 3 – Used cannabis 11-100 times but less than daily (N=4316), 4 – Used cannabis over 100 times or daily (N=2692). Polygenic risk scores were created in the imputed UK Biobank data using PRSice-2[93], based on effect sizes derived from the EUR meta-analysis of unrelated, genotyped individuals ($N_{CUD} =$ 14,080, $N_{control} =$ 343,726). Nine PRS were generated using different p-value thresholds ($p_T$) in the discovery GWAS ($p \leq 0.0001$, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 1.0 (all LD-independent SNPs)).

Using linear regression, cannabis use frequency was tested for association with PRS using age, sex, and 20 genetic principal components as covariates. Variance explained was calculated by subtracting the adjusted $R^2$ from a null model (linear model excluding PRS) from the adjusted $R^2$ from the full model (including PRS).

This research was conducted using the UK Biobank Resource, application numbers 4844 and 58146 (each site conducted analyses independently).

### 12.2 BioVU biobank

Polygenic scores for CUD were computed using the PRS-CS[94] “auto” version (i.e., the global shrinkage parameter phi was learnt from the data in a Bayesian approach) for each of the 66,915 genotyped individuals of European descent in BioVU. Genotyping and QC of this sample have been described elsewhere[95,96]. In the genotyped BioVU sample, a logistic regression model was fitted to each of 1,335 case/control phenotypes to estimate the odds of each diagnosis given the CUD polygenic score, after adjustment for sex, median age of the longitudinal EHR measurements, top 10 principal components of ancestry. The disease phenotypes included 171 circulatory system, 170 genitourinary, 169 endocrine/metabolic, 162 digestive, 140 neoplasms, 132 musculoskeletal, 127 sense organs, 126 injuries & poisonings, 90 dermatologic, 85 respiratory, 84 neurological, 76 mental disorders, 69 infectious diseases, 62 hematopoietic, 56 congenital anomalies, 49 symptoms, 46 pregnancy complications. We required the presence of at least two International Classification of Disease (ICD) codes that mapped to a PheWAS disease category (Phecode Map 1.2 (https://phewascatalog.org/phecodes) to assign "case" status[97]. We analyzed 1,335 phecodes (a "phecode" is a phenotype code, created by aggregating one or more related ICD codes into distinct diseases) with at least 100 cases, and used a Bonferroni-corrected phenome-wide significance threshold of $\alpha < .05 / 1335 = 3.74e-5$. PheWAS analyses were run using the PheWAS R package v0.12.[98]

### 12.3. Adolescent Brain and Cognitive Development (ABCD)

Data from the ongoing Adolescent Brain Cognitive Development (ABCD) study[99] (data release 2.0.1; https://abcdstudy.org/) were used to test whether CUD PRS are associated with brain structure among 4,539 cannabis-naïve (via self-report or toxicology) children of European ancestry (mean age = 9.93±0.63 years; 46.82% girls). All parents provided written informed consent, and all children provided verbal assent to a research protocol approved by the institutional review board at each data collection site (N = 22) throughout the United States (https://abcdstudy.org/sites/abcd-sites.html). Genetic QC followed the Ricopili pipeline[42].Total bilateral white matter volume, gray matter volume, and intracranial volume were estimated using FreeSurfer[100] 5.3. PRS from the CUD GWAS were generated at nine p-value thresholds (i.e., PT = 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, and 1). Linear mixed-effects models were used to include scanner (for imaging analyses) and family as nested random effects, conducted using the lme4 package in R[101], version 3.6.0. All analyses also included the following fixed effect covariates: first 20 ancestry principal components, age, sex, age by sex, parents combined income, caregiver education, genotyping batch, caregiver's marital status, prenatal cannabis exposure before and after knowledge of pregnancy, and twin status. Multiple testing was accounted for by applying random field theory correction across p-value thresholds.

## 13 Supplemental tables

Link to downloadable online Excel spreadsheet of supplemental tables: Online Supplemental Tables

**CONTENTS:**

**Table S5.** All SNPs within the two genomic risk loci with $p < 0.05$ for the European-ancestry meta-analysis.
**Table S6.** All genes mapped by MAGMA for the EUR-ancestry summary statistics.
**Table S7.** Results of pathway analysis through PASCAL (alpha level after multiple testing correction = 4.6e-5).
**Table S8.** Significant findings from S-PrediXcan Analyses.
**Table S9.** All S-PrediXcan results.
**Table S10.** Hi-C coupled MAGMA (H-MAGMA) results for adult brain tissue. Genes above dashed line pass multiple testing corrections.
**Table S11.** Hi-C coupled MAGMA (H-MAGMA) results for fetal brain tissue. Genes above dashed line pass multiple testing corrections.
**Table S12.** Hi-C coupled MAGMA (H-MAGMA) results for iPSC-derived astrocytes. Genes above dashed line pass multiple testing corrections.
**Table S13.** Hi-C coupled MAGMA (H-MAGMA) results for iPSC-derived neurons. Genes above dashed line pass multiple testing corrections.
**Table S14.** Genetic correlations with traits of interest from LD Score Regression.
**Table S15.** Comparison of genome-wide SNPs in Pasman et al. cannabis use study with the current CUD GWAS meta-analysis.
**Table S16.** Results for top two SNPs after conditioning CUD on cannabis use using mtCOJO.
**Table S17.** Genetic correlations between cannabis use and traits of interest from LD Score Regression.
**Table S18.** Comparison of genetic correlations between relevant traits and CUD vs. CUD after covarying on cannabis use.
**Table S19.** Results of latent causal variable analyses with CUD and top correlated traits.
**Table S20.** Association between CUD PRS and cannabis use in the UK Biobank (N=107,198) at 9 different p-value threshold cut-offs. Most significant association is indicated in bold text.
**Table S21.** Results of PRSet gene-set enrichment test.
**Table S22.** Association between CUD PRS and phecodes in the BioVU biobank.
**Table S23.** Association between CUD PRS and phecodes in the BioVU biobank, after conditioning the CUD summary statistics for smoking intitiation loci using mtCOJO.
**Table S24.** Association between CUD PRS and phecodes in the BioVU biobank, covarying for tobacco use disorder (TUD) phecode.
**Table S25.** Association between PRS (for both CUD and cannabis use) and white matter volume in the ABCD sample at 9 different p-value threshold cut-offs. Most significant association is indicated in bold text.
**Table S26.** Association between PRS (for both CUD and cannabis use) and gray matter volume in the ABCD sample at 9 different p-value threshold cut-offs. Most significant association is indicated in bold text.
**Table S27.** Results for top two SNPs after conditioning CUD on schizophrenia using mtCOJO.

## 14 Supplemental figures

**CONTENTS:**

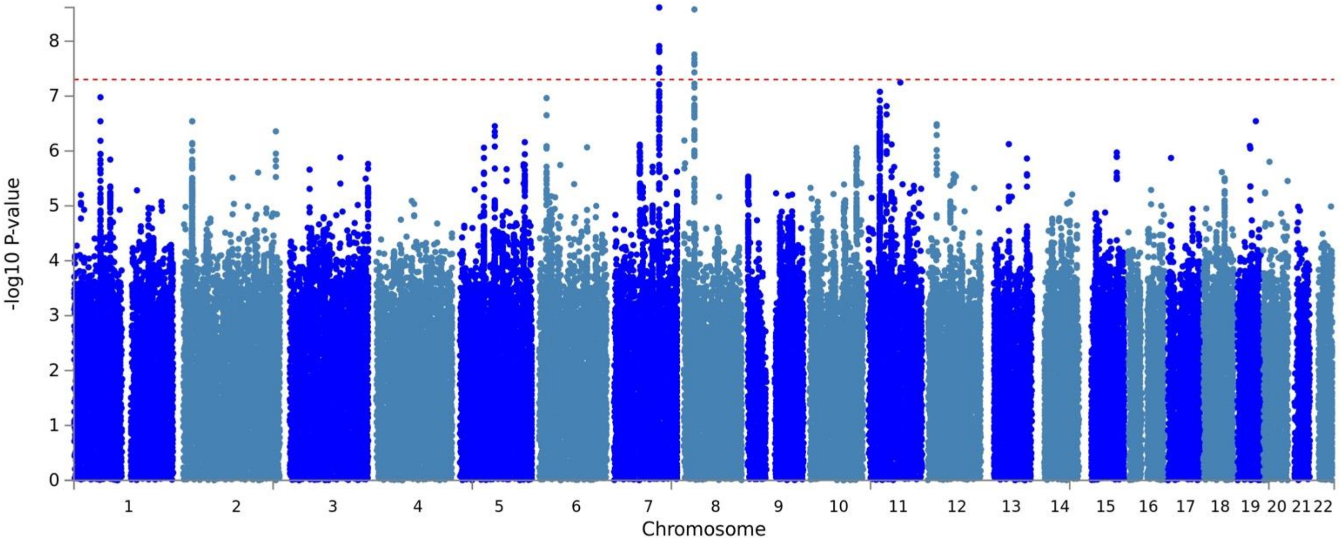
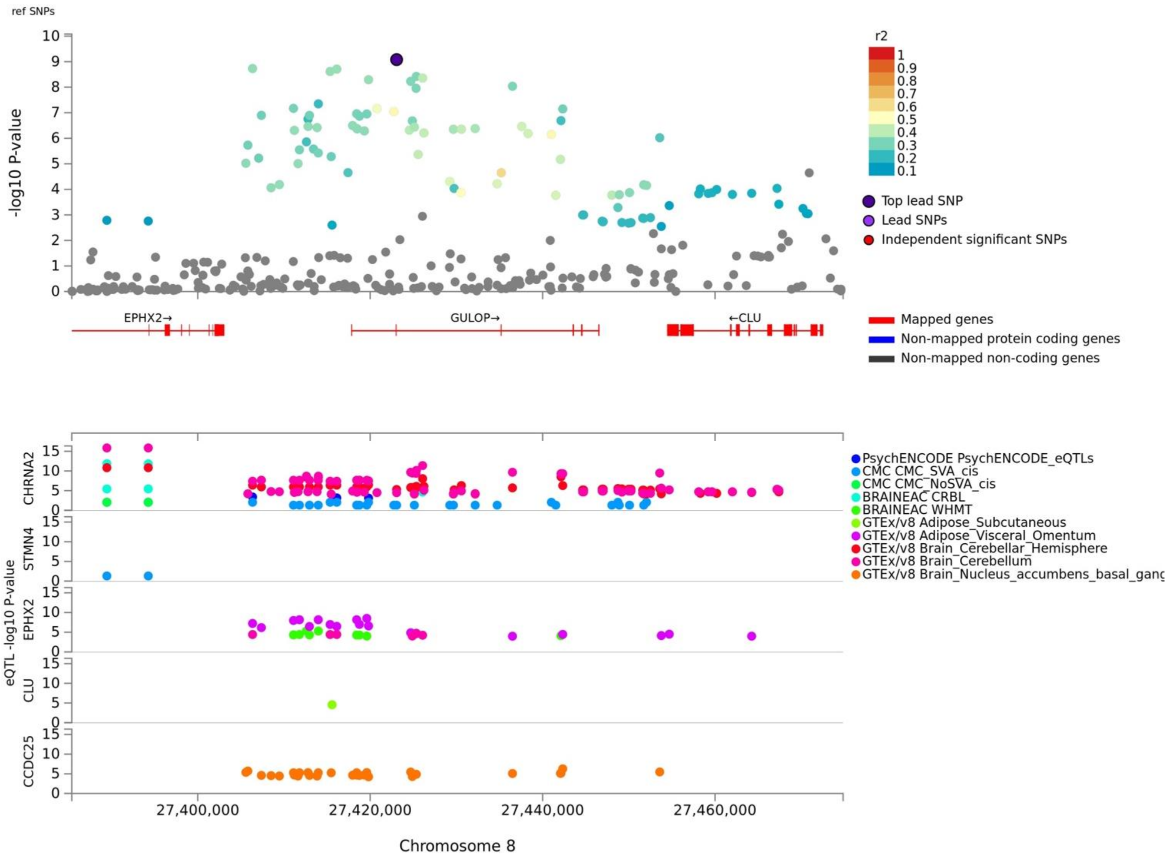
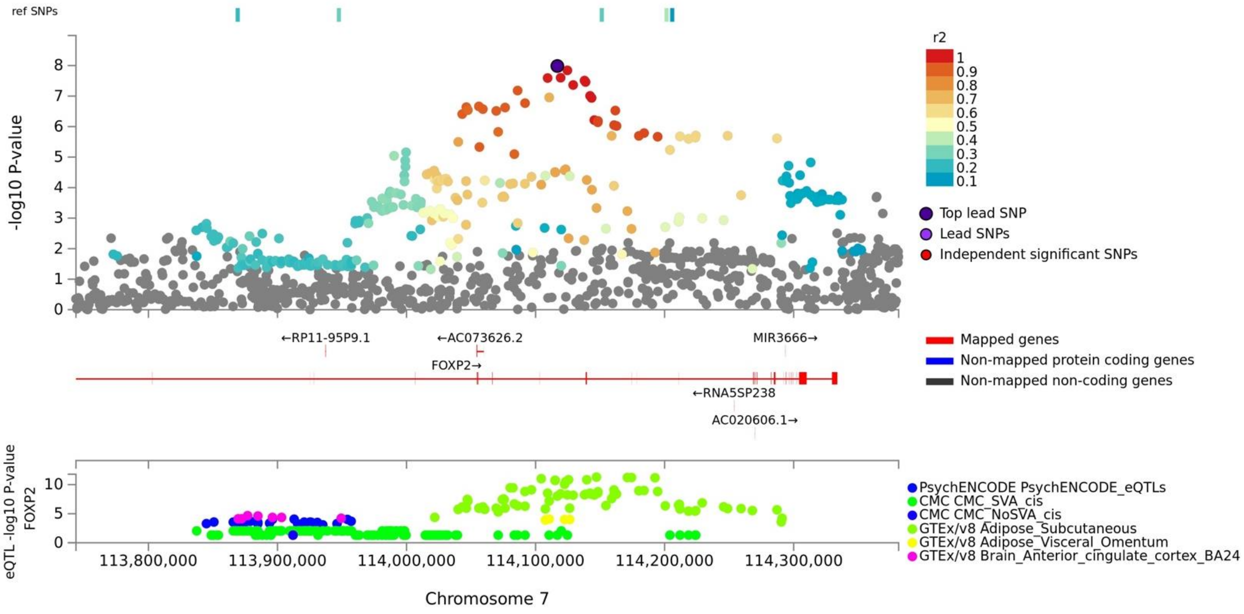


**Figure S1. SNP-level Manhattan plot for CUD (Trans-ancestral).**


ref SNPs
-log10 P-value
10
9
8
7
6
5
4
3
2
1
0
r2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Top lead SNP
Lead SNPs
Independent significant SNPs
EPHX2→
GULOP→
←CLU
Mapped genes
Non-mapped protein coding genes
Non-mapped non-coding genes
CHRNA2
STMN4
EPHX2
CLU
CCDC25
eQTL -log10 P-value
15
10
5
0
PsychENCODE PsychENCODE_eQTLs
CMC CMC_SVA_cis
CMC CMC_NoSVA_cis
BRAINEAC CRBL
BRAINEAC WHMT
GTEx/v8 Adipose_Subcutaneous
GTEx/v8 Adipose_Visceral_Omentum
GTEx/v8 Brain_Cerebellar_Hemisphere
GTEx/v8 Brain_Cerebellum
GTEx/v8 Brain_Nucleus_accumbens_basal_ganc
27,400,000
27,420,000
27,440,000
27,460,000
Chromosome 8

**A.**


ref SNPs
-log10 P-value
8
7
6
5
4
3
2
1
0
r2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Top lead SNP
Lead SNPs
Independent significant SNPs
←RP11-95P9.1
←AC073626.2
FOXP2→
MIR3666→
←RNA5SP238
AC020606.1→
Mapped genes
Non-mapped protein coding genes
Non-mapped non-coding genes
eQTL -log10 P-value
FOXP2
10
5
0
PsychENCODE PsychENCODE_eQTLs
CMC CMC_SVA_cis
CMC CMC_NoSVA_cis
GTEx/v8 Adipose_Subcutaneous
GTEx/v8 Adipose_Visceral_Omentum
GTEx/v8 Brain_Anterior_cingulate_cortex_BA24
113,800,000
113,900,000
114,000,000
114,100,000
114,200,000
114,300,000
Chromosome 7

**B.**

**Figure S2. Regional association plots for genome-wide significant loci in the EUR-only GWAS meta-analysis. A.** Regional association plot of chromosome 8 risk locus, with eQTLs displayed in bottom panel **B.** Regional association plot of chromosome 7 risk locus, showing eQTLs in bottom panel.

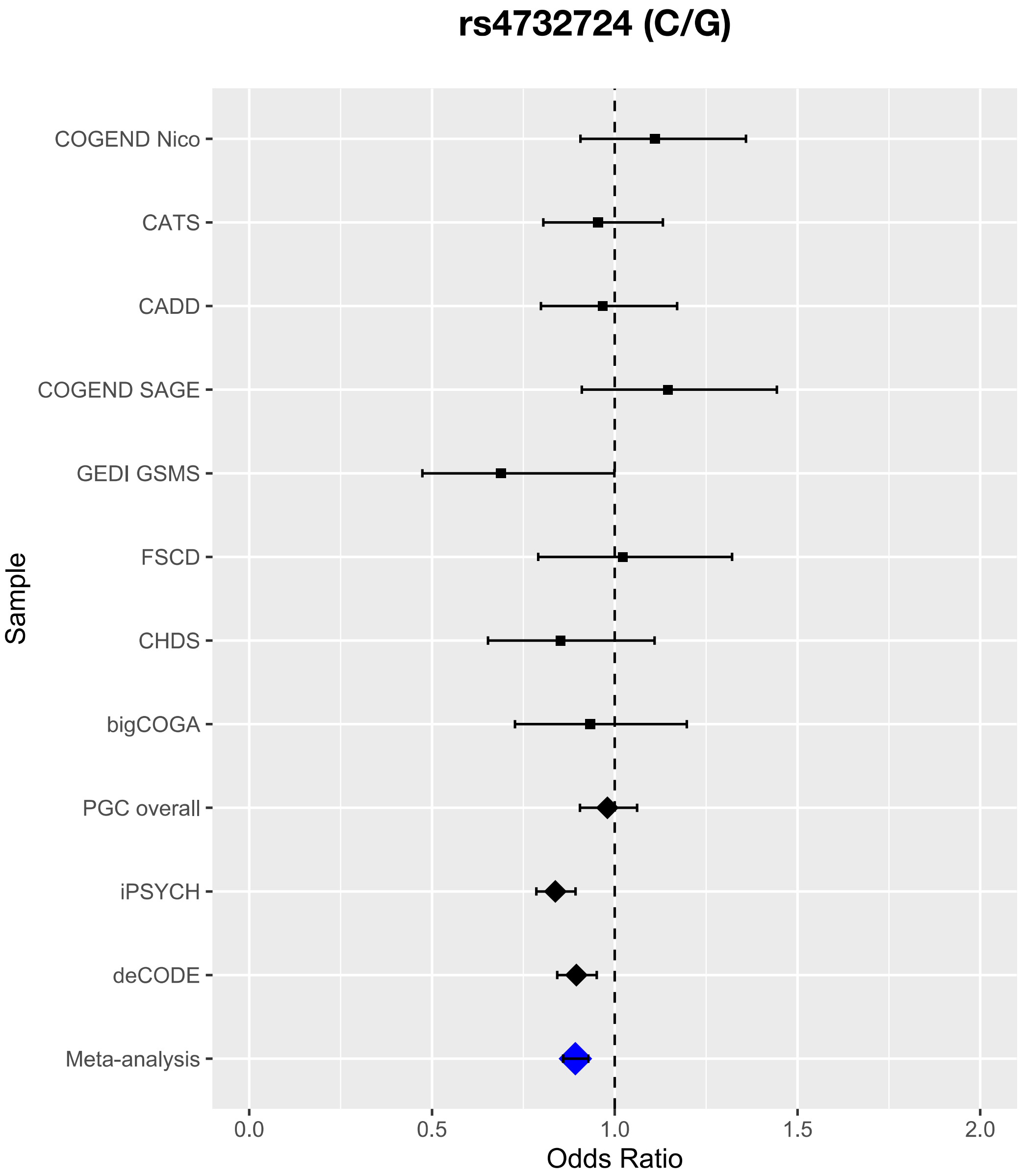


**Figure S3. Forest plot of study-specific and meta-analytic association statistics** (odds ratios with 95% confidence intervals) in the European ancestry genotyped case-control individuals at the lead SNP, rs4732724, at the genomic risk locus on chromosome 8.

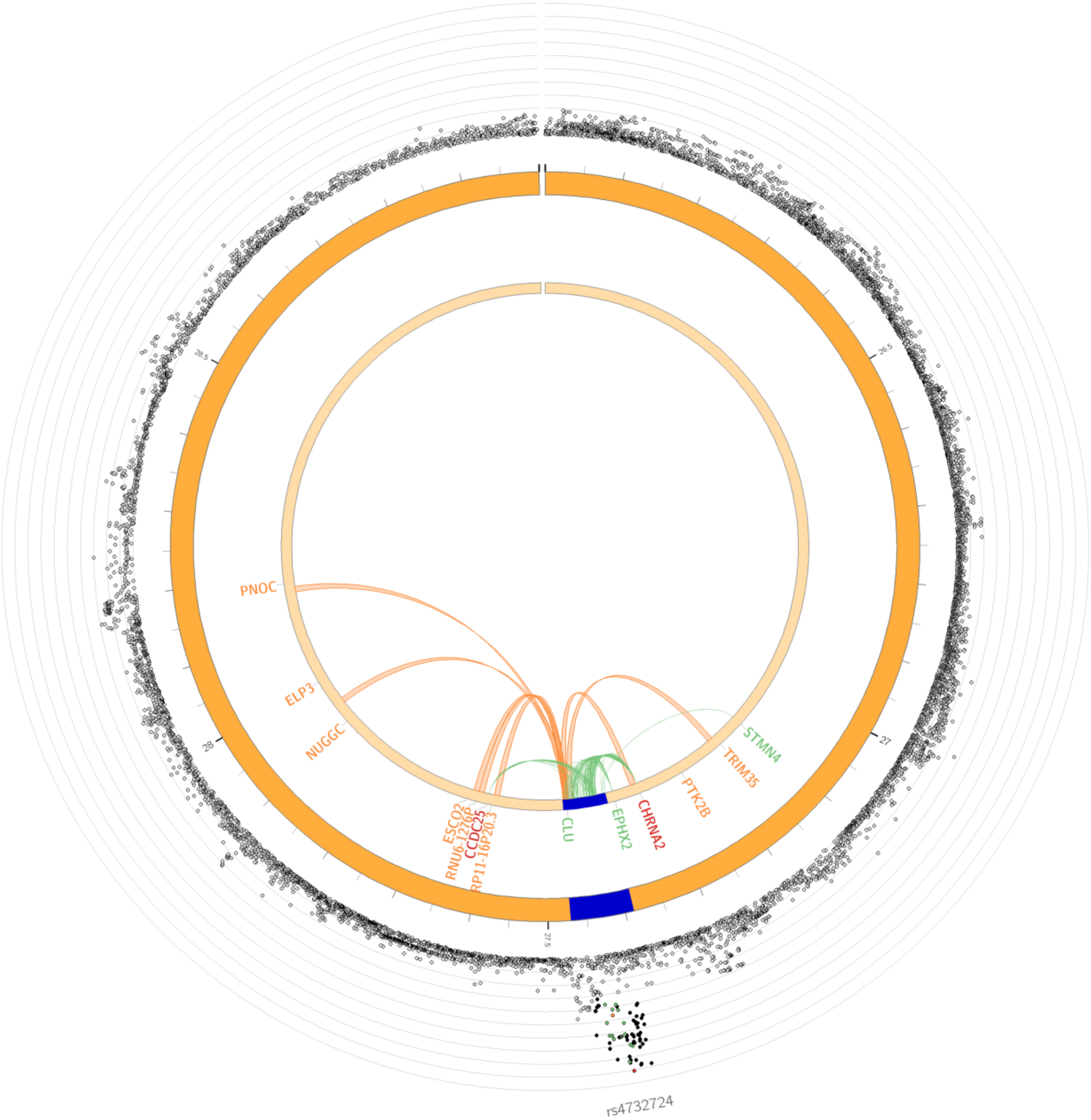


**Figure S4. Circos plot showing eQTL and chromatin interactions at the genomic risk locus on chromosome 8.** The most outer layer of the plot shows the Manhattan plot of SNP associations (-$\log_{10}$(p-value)), but only SNPs with $p < 0.05$ are displayed. The rsID of the top SNPs in each risk locus are displayed in the most outer layer. The second layer shows the genomic risk locus highlighted in blue. Only mapped genes by either chromatin interaction and/or eQTLs are displayed. If the gene is mapped only by chromatin interactions or only by eQTLs, it is colored orange or green, respectively. When the gene is mapped by both, it is colored red. Chromatin interaction links are colored orange, while eQTL links are colored green.

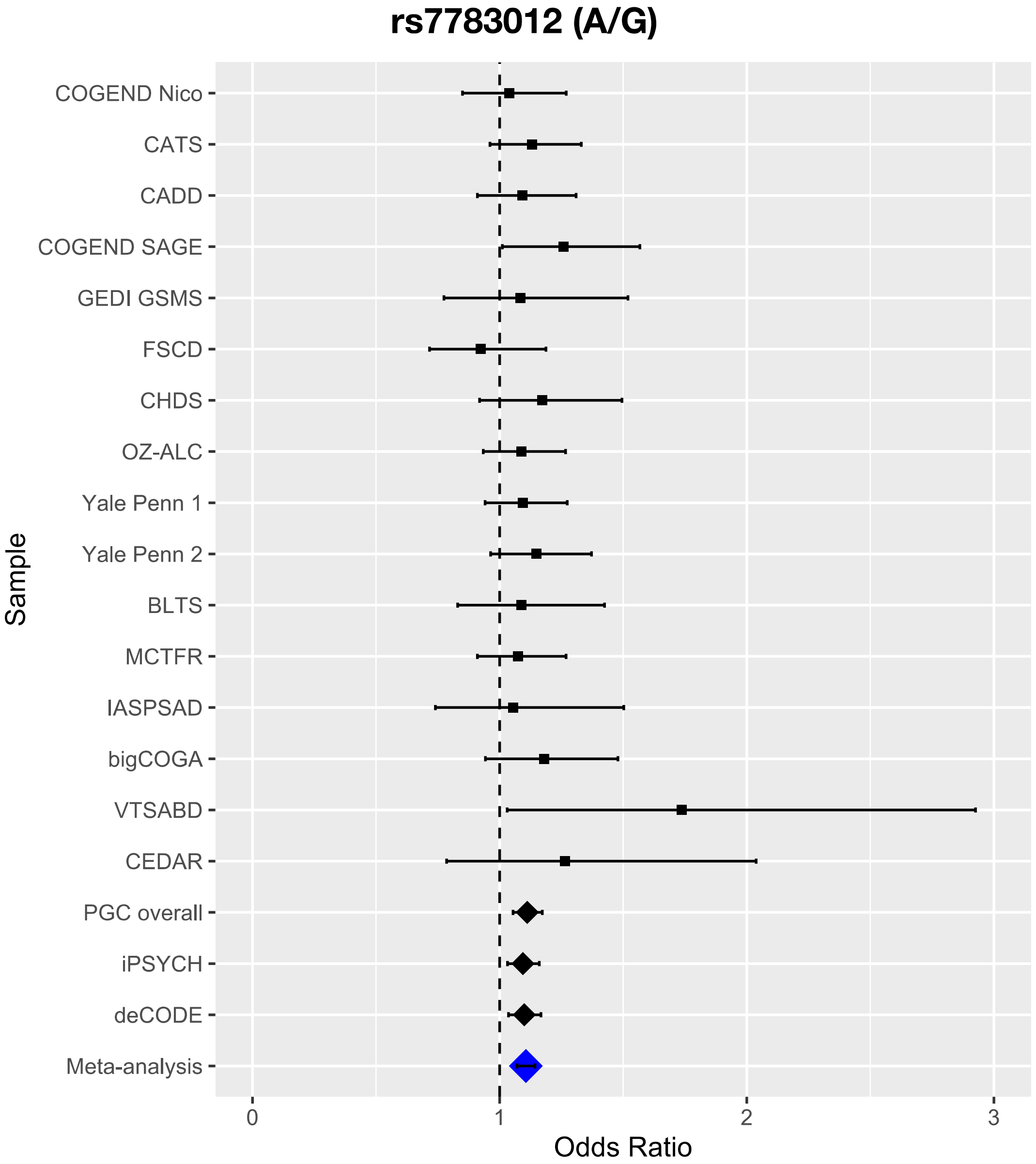


**Figure S5. Forest plot of study-specific and meta-analytic association statistics** (odds ratios with 95% confidence intervals) in the European ancestry genotyped case-control individuals at the lead SNP, rs7783012, at the genomic risk locus on chromosome 7.

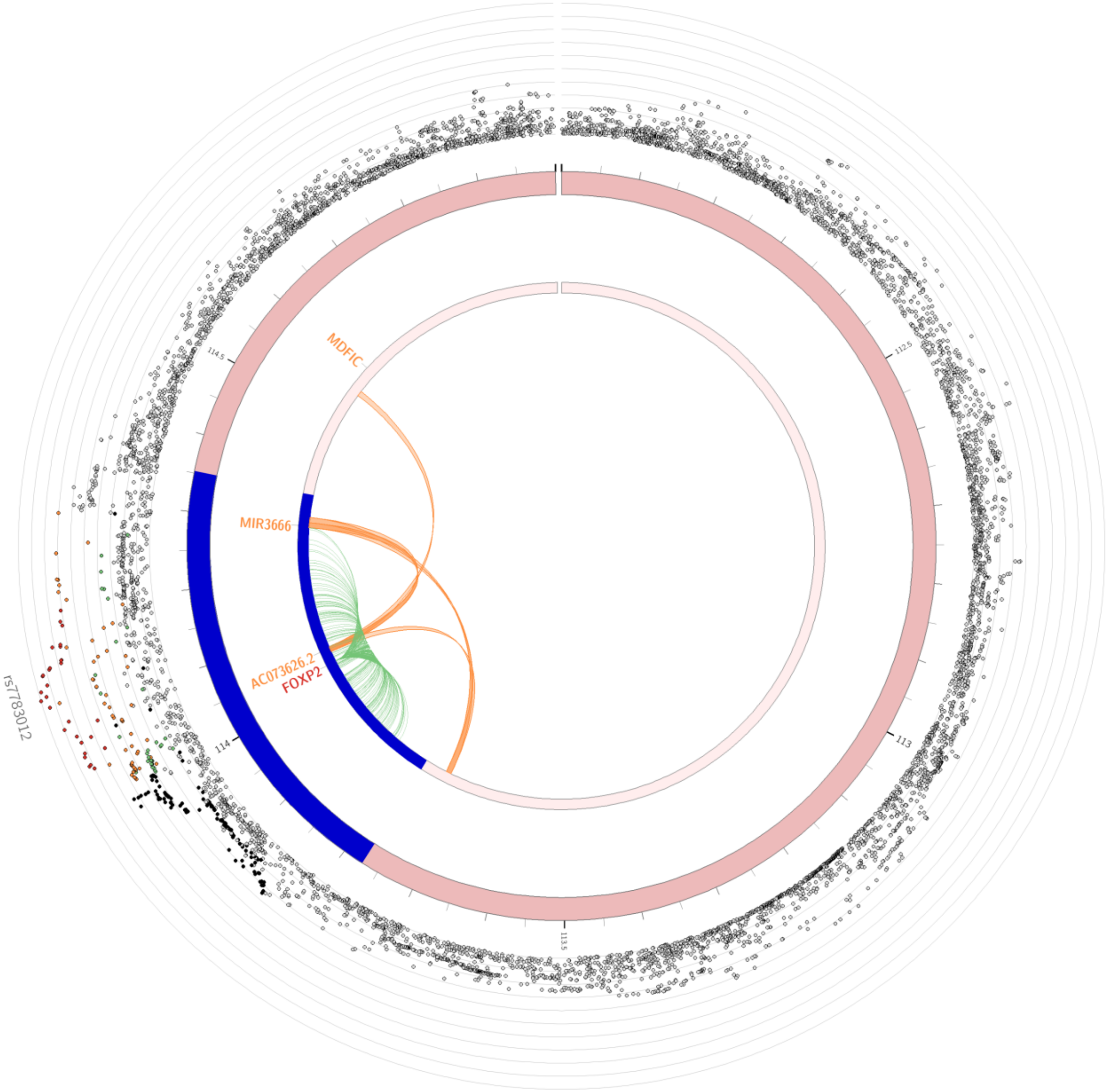


**Figure S6. Circos plot showing eQTL and chromatin interactions at the genomic risk locus on chromosome 7.** The most outer layer of the plot shows the Manhattan plot of SNP associations (-$\log_{10}$(p-value)), but only SNPs with $p < 0.05$ are displayed. The rsID of the top SNPs in each risk locus are displayed in the most outer layer. The second layer shows the genomic risk locus highlighted in blue. Only mapped genes by either chromatin interaction and/or eQTLs are displayed. If the gene is mapped only by chromatin interactions or only by eQTLs, it is colored orange or green, respectively. When the gene is mapped by both, it is colored red. Chromatin interaction links are colored orange, while eQTL links are colored green.

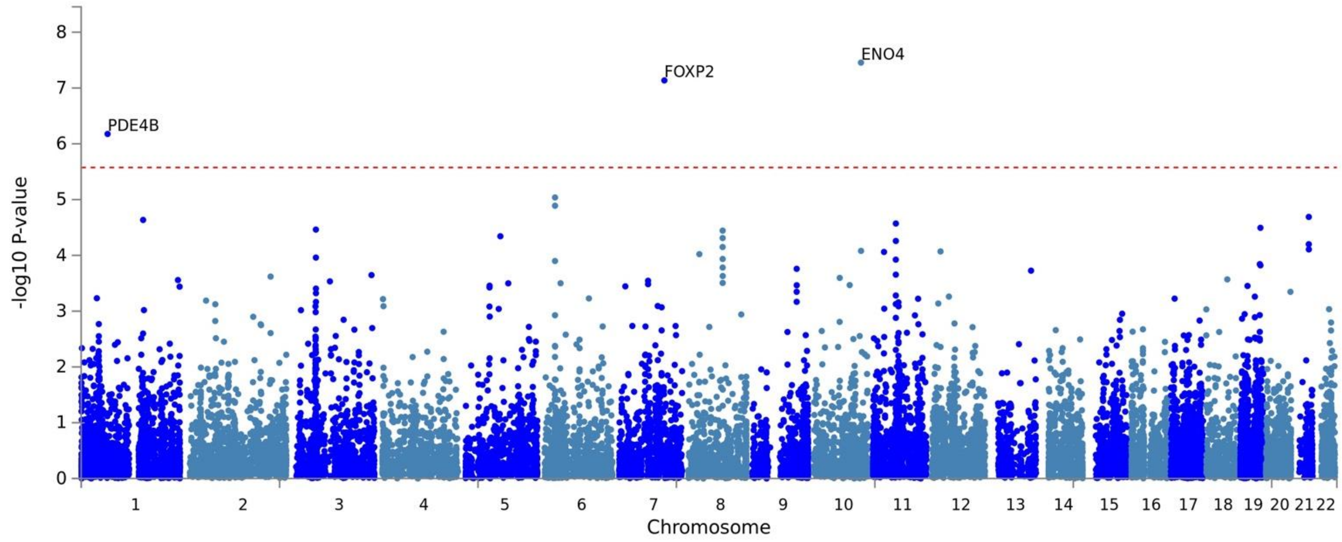


**Figure S7. Gene-level Manhattan plot (EUR-only).**

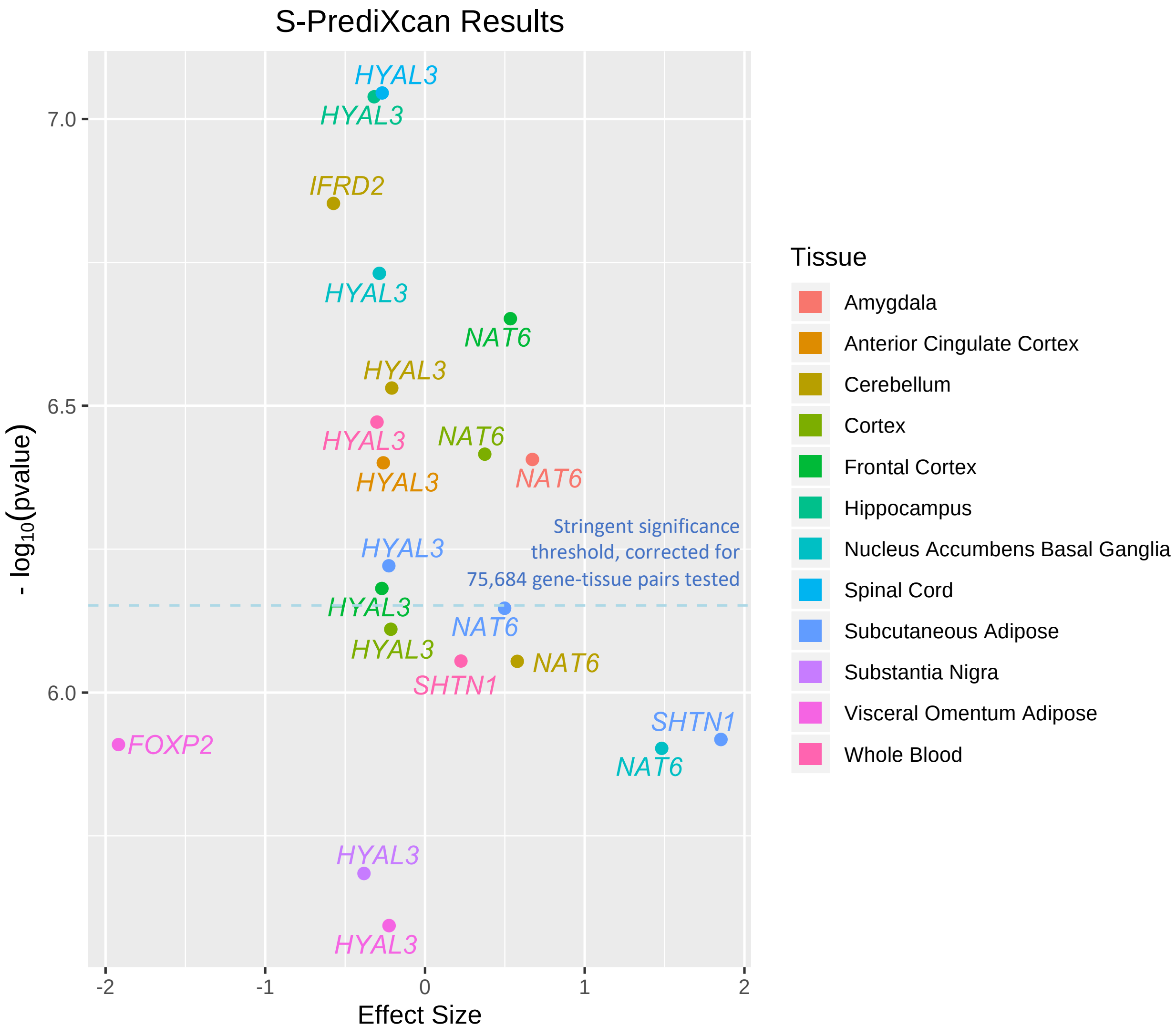
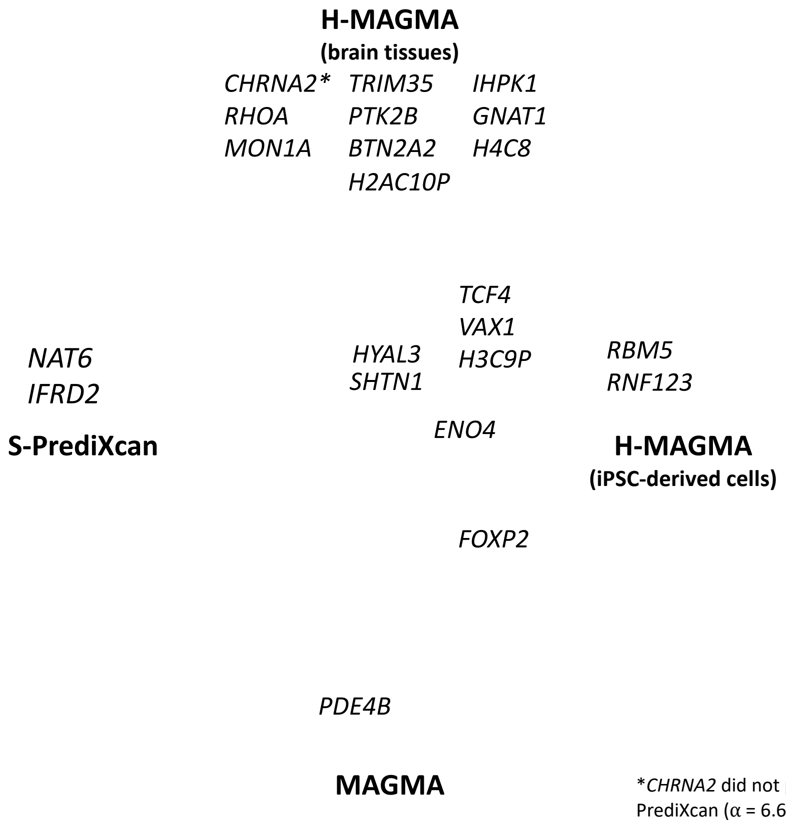


**Figure S8**. **S-PrediXcan results.** All genes shown are significant after correcting for the total number of genes tested (16,903), while genes above the dashed line are significant after correcting for the number of unique gene-tissue pairwise comparisons (75,684).



*CHRNA2 did not pass stringent multiple testing corrections in S-PrediXcan (α = 6.69e-7), but the p-value in the cerebellar hemisphere model was 1.29e-5, suggestive of multiple lines of evidence for this gene as well

**Figure S9. Overview of overlapping findings from different gene-based approaches.**

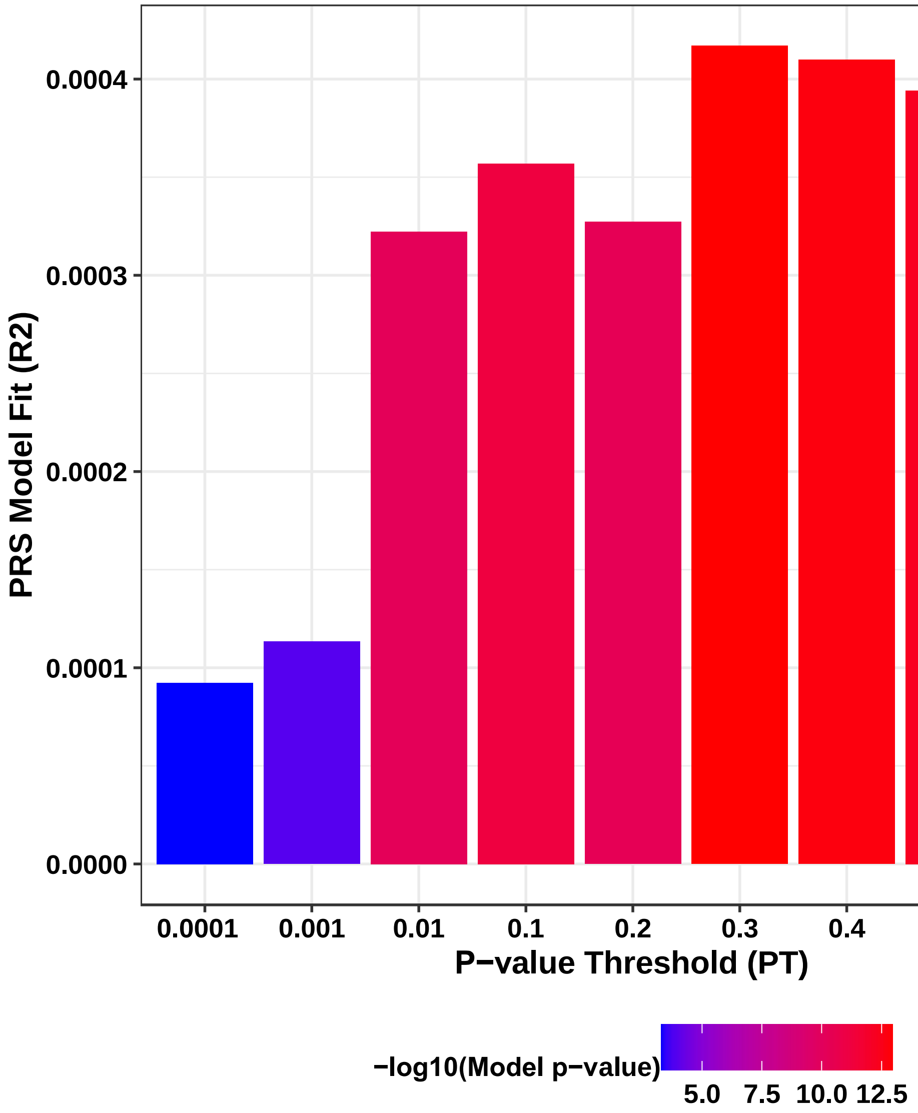


**Figure S10. Associations between CUD PRS and cannabis use frequency in the UK Biobank.** PRSice-2 was used to perform gene-set enrichment using gene sets and pathways from the Molecular Signatures Database (MSigDB[103]). (H) hallmark biological processes or states, (C1) positional sets from cytogenetic maps, (C2) chemical or genetic perturbations and canonical pathways, (C3) regulatory processes, (C4) computationally derived gene sets of cancer gene neighborhoods and modules, (C5) biological process, cellular component, and molecular function gene ontologies, (C6) oncogenic signatures, and (C7) immunologic signatures.

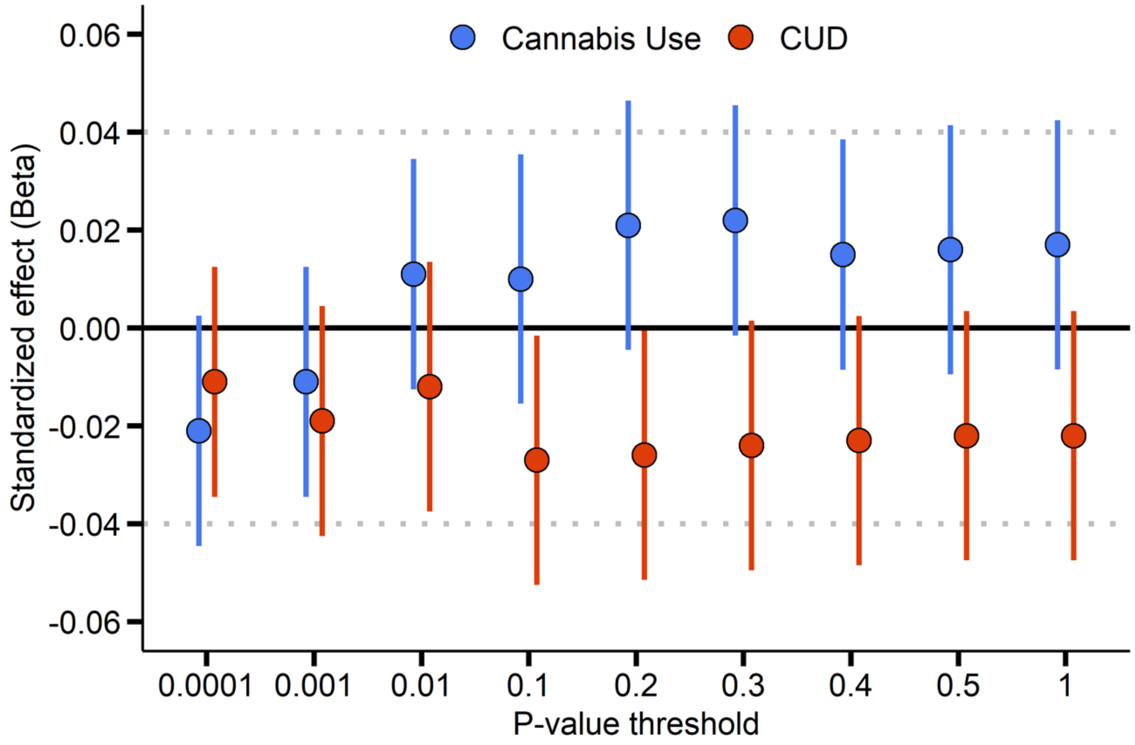


**Figure S11. Associations between total gray matter volume, CUD PRS, and cannabis use PRS.**

## 15 References

1. Pedersen, C. B. *et al.* The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6 (2018).
2. Mors, O., Perto, G. P. & Mortensen, P. B. The Danish psychiatric central research register. *Scand. J. Public Health* **39**, 54–57 (2011).
3. Nørgaard-Pedersen, B. & Hougaard, D. M. Storage policies and use of the Danish Newborn Screening Biobank. *J. Inherit. Metab. Dis. Off. J. Soc. Study Inborn Errors Metab.* **30**, 530–536 (2007).
4. Hollegaard, M. V *et al.* Robustness of genome-wide scanning using archived dried blood spot samples as a DNA source. *BMC Genet.* **12**, 58 (2011).
5. Børglum, A. D. *et al.* Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci. *Mol. Psychiatry* **19**, 325–333 (2014).
6. Tyrfingsson, T. *et al.* Addictions and their familiality in Iceland. *Ann. N. Y. Acad. Sci.* **1187**, 208–217 (2010).
7. Nelson, E. C. *et al.* Evidence of CNIH3 involvement in opioid dependence. *Mol. Psychiatry* **21**, 608 (2015).
8. Fergusson, D. M. & Horwood, L. J. The Christchurch Health and Development Study: review of findings on child and adolescent mental health. *Aust. N. Z. J. Psychiatry* **35**, 287–296 (2001).
9. Fergusson, D. M., Boden, J. M. & Horwood, L. J. Alcohol misuse and psychosocial outcomes in young adulthood: results from a longitudinal birth cohort studied to age 30. *Drug Alcohol Depend.* **133**, 513–519 (2013).
10. Begleiter, H. *et al.* The collaborative study on the genetics of alcoholism. *Alcohol Health Res. World* **19**, 228 (1995).
11. Bierut, L. J., Strickland, J. R., Thompson, J. R., Afful, S. E. & Cottler, L. B. Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings. *Drug Alcohol Depend.* **95**, 14–22 (2008).
12. Bierut, L. J. *et al.* Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum. Mol. Genet.* **16**, 24–35 (2006).
13. Bierut, L. J. *et al.* A genome-wide association study of alcohol dependence. *Proc. Natl. Acad. Sci.* **107**, 5082–5087 (2010).
14. Costello, E. J. *et al.* Genes, environments, and developmental research: methods for a multi-site study of early substance abuse. *Twin Res. Hum. Genet.* **16**, 505–515 (2013).
15. Adkins, D. E. *et al.* Genome-wide meta-analysis of longitudinal alcohol consumption across youth and early adulthood. *Twin Res. Hum. Genet.* **18**, 335–347 (2015).
16. Hartman, C. A. *et al.* Item response theory analysis of DSM-IV cannabis abuse and dependence criteria in adolescents. *J. Am. Acad. Child Adolesc. Psychiatry* **47**, 165–173 (2008).
17. Petrill, S. A., Plomin, R., DeFries, J. C. & Hewitt, J. K. *Nature, nurture, and the transition to early adolescence*. (Oxford University Press, 2003).
18. Rhea, S.-A., Gross, A. A., Haberstick, B. C. & Corley, R. P. Colorado twin registry. *Twin Res. Hum. Genet.* **9**, 941–949 (2006).
19. Stallings, M. C. *et al.* A genome-wide search for quantitative trait loci that influence

antisocial drug dependence in adolescence. *Arch. Gen. Psychiatry* **62**, 1042–1051 (2005).

20. Derringer, J. *et al.* Genome-wide association study of behavioral disinhibition in a selected adolescent sample. *Behav. Genet.* **45**, 375–381 (2015).
21. Gillespie, N. A. *et al.* The Brisbane Longitudinal Twin Study: pathways to Cannabis Use, Abuse, and Dependence project—current status, preliminary results, and future directions. *Twin Res. Hum. Genet.* **16**, 21–33 (2013).
22. Couvy-Duchesne, B., Davenport, T. A., Martin, N. G., Wright, M. J. & Hickie, I. B. Validation and psychometric properties of the Somatic and Psychological HEalth REport (SPHERE) in a young Australian-based population sample using non-parametric item response theory. *BMC Psychiatry* **17**, 279 (2017).
23. Medland, S. E. *et al.* Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am. J. Hum. Genet.* **85**, 750–755 (2009).
24. Maes, H. H., Silberg, J. L., Neale, M. C. & Eaves, L. J. Genetic and cultural transmission of antisocial behavior: an extended twin parent model. *Twin Res. Hum. Genet.* **10**, 136–150 (2007).
25. Lai, D. *et al.* Genome-wide association studies of alcohol dependence, DSM-IV criterion count and individual criteria. *Genes, Brain Behav.* **18**, e12579 (2019).
26. Walters, R. K. *et al.* Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656–1669 (2018).
27. Bucholz, K. K. *et al.* A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. *J. Stud. Alcohol* **55**, 149–158 (1994).
28. Hesselbrock, M., Easton, C., Bucholz, K. K., Schuckit, M. & Hesselbrock, V. A validity study of the SSAGA - A comparison with the SCAN. *Addiction* **94**, 1361–1370 (1999).
29. Vrieze, S. I. *et al.* Rare nonsynonymous exonic variants in addiction and behavioral disinhibition. *Biol. Psychiatry* **75**, 783–789 (2014).
30. Miller, M. B. *et al.* The Minnesota Center for Twin and Family Research Genome-Wide Association Study. *Twin Res. Hum. Genet.* **15**, 767–774 (2012).
31. Maher, B. S. *et al.* The AVPR1A gene and substance use disorders: association, replication, and functional evidence. *Biol. Psychiatry* **70**, 519–527 (2011).
32. Tarter, R. E. & Vanyukov, M. M. Introduction: Theoretical and operational framework for research into the etiology of substance use disorders. *J. Child Adolesc. Subst. Abuse* **10**, 1–12 (2001).
33. Pierucci-Lagha, A. *et al.* Reliability of DSM-IV diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (SSADDA). *Drug Alcohol Depend.* **91**, 85–90 (2007).
34. Gelernter, J. *et al.* Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. *Biol. Psychiatry* **76**, 66–74 (2014).
35. Gelernter, J. *et al.* Genome-wide association study of cocaine dependence and related traits: FAM53B identified as a risk gene. *Mol. Psychiatry* **19**, 717–723 (2014).
36. Gelernter, J. *et al.* Genome-wide association study of alcohol dependence: significant findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry* **19**, 41–49 (2014).
37. Sherva, R. *et al.* Genome-wide association study of cannabis dependence severity, novel risk variants, and shared genetic risks. *JAMA psychiatry* **73**, 472–480 (2016).

38. Heath, A. C. *et al.* A quantitative-trait genome-wide association study of alcoholism risk in the community: findings and implications. *Biol. Psychiatry* **70**, 513–518 (2011).
39. Prescott, C. A. *et al.* The Irish Affected Sib Pair Study of Alcohol Dependence: study methodology and validation of diagnosis by interview and family history. *Alcohol. Clin. Exp. Res.* **29**, 417–429 (2005).
40. Harris, K. M. *et al.* Cohort profile: The national longitudinal study of adolescent to adult health (add health). *Int. J. Epidemiol.* **48**, 1415-1415k (2019).
41. Harris, K. M., Halpern, C. T., Haberstick, B. C. & Smolen, A. The national longitudinal study of adolescent health (Add Health) sibling pairs data. *Twin Res. Hum. Genet.* **16**, 391–398 (2013).
42. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *bioRxiv* 587196 (2019).
43. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435 (2015).
44. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
45. O'Connell, J. R. & Weeks, D. E. PedCheck: A Program for Identification of Genotype Incompatibilities in Linkage Analysis. *Am. J. Hum. Genet.* **63**, 259–266 (1998).
46. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181 (2011).
47. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284 (2016).
48. Demontis, D. *et al.* Genome-wide association study implicates CHRNA2 in cannabis use disorder. *Nat. Neurosci.* 1 (2019).
49. Price, A. L. *et al.* Long-Range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 132–135 (2008).
50. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* (2007). doi:10.1086/519795
51. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904 (2006).
52. Galinsky, K. J. *et al.* Fast Principal-Component Analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).
53. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
54. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
55. Delaneau, O. *et al.* Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**, (2014).
56. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
57. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5 (2012).
58. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
59. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes.

*G3 Genes, Genomes, Genet.* **1**, 457–470 (2011).

60. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet* **10**, e1004234 (2014).
61. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279 (2016).
62. Thornton, T. *et al.* Estimating kinship in admixed populations. *Am. J. Hum. Genet.* **91**, 122–138 (2012).
63. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).
64. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* **46**, 100–106 (2014).
65. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
66. Price, A. L. *et al.* The Impact of Divergence Time on the Nature of Population Structure: An Example from Iceland. *PLOS Genet.* **5**, e1000505 (2009).
67. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
68. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
69. Ziats, M. N. & Rennert, O. M. Identification of differentially expressed microRNAs across the developing human brain. *Mol. Psychiatry* **19**, 848 (2014).
70. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* **11**, (2015).
71. Akbarian, S. *et al.* The psychencode project. *Nat. Neurosci.* **18**, 1707 (2015).
72. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience* **19**, 1442–1453 (2016).
73. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* **17**, 1418–1428 (2014).
74. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
75. Giusti-Rodriguez, P. *et al.* SU62 - A CHROMATIN CATALOG FOR THE INTERPRETATION OF GENETIC ASSOCIATIONS OF PSYCHIATRIC DISORDERS. *Eur. Neuropsychopharmacol.* **29**, S921–S922 (2019).
76. Nancy, Y., Fauni, H., Ma, W. & Won, H. Connecting gene regulatory relationships to neurobiological mechanisms of brain disorders. *bioRxiv* 681353 (2019).
77. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).
78. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science (80-. ).* **362**, eaat8464 (2018).
79. Rajarajan, P. *et al.* Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science (80-. ).* **362**, eaat4311 (2018).
80. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236 (2015).
81. Pasman, J. A. *et al.* GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia. *Nat. Neurosci.* **21**, 1161–

1170 (2018).

82. Munn-Chernoff, M. A. *et al.* Shared Genetic Risk between Eating Disorder- and Substance-Use-Related Phenotypes: Evidence from Genome-Wide Association Studies. *bioRxiv* 741512 (2019). doi:10.1101/741512
83. Hancock, D. B. *et al.* Genome-wide association study across European and African American ancestries identifies a SNP in DNMT3B contributing to nicotine dependence. *Mol. Psychiatry* **23**, 1911–1919 (2018).
84. Kranzler, H. R. *et al.* Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat. Commun.* **10**, 1499 (2019).
85. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
86. Volkow, N. D., Michaelides, M. & Baler, R. The neuroscience of drug reward and addiction. *Physiol. Rev.* **99**, 2115–2140 (2019).
87. Cederbaum, J. A., Jeong, C. H., Yuan, C. & Lee, J. O. Sex and substance use behaviors among children of teen mothers: A systematic review. *J. Adolesc.* **79**, 208–220 (2020).
88. Clark, T. M., Jones, J. M., Hall, A. G., Tabner, S. A. & Kmiec, R. L. Theoretical explanation for reduced body mass index and obesity rates in cannabis users. *Cannabis cannabinoid Res.* **3**, 259–271 (2018).
89. Ahrens, A. M. & Ahmed, O. J. Neural circuits linking sleep and addiction: animal models to understand why select individuals are more vulnerable to substance use disorders after sleep deprivation. *Neurosci. Biobehav. Rev.* (2019).
90. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
91. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* (2017).
92. Davis, K. A. S. *et al.* Mental health in UK Biobank: development, implementation and results from an online questionnaire completed by 157 366 participants. *BJPsych open* **4**, 83–90 (2018).
93. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).
94. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1–10 (2019).
95. Dennis, J. *et al.* Genetic risk for major depressive disorder and loneliness in sex-specific associations with coronary artery disease. *Mol. Psychiatry* 1–11 (2019).
96. Ruderfer, D. M. *et al.* Significant shared heritability underlies suicide attempt and clinically predicted probability of attempting suicide. *Mol. Psychiatry* 1–9 (2019).
97. Wei, W.-Q. *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Informatics Assoc.* **23**, e20–e27 (2016).
98. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).
99. Lisdahl, K. M. *et al.* Adolescent brain cognitive development (ABCD) study: overview of substance use assessment methods. *Dev. Cogn. Neurosci.* **32**, 80–96 (2018).
100. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).

101. Bates, D., Sarkar, D., Bates, M. D. & Matrix, L. The lme4 package. *R Packag. version* **2**, 74 (2007).

102. O'Connor, L. J. & Price, A. L. *Distinguishing genetic correlation from causation across 52 diseases and complex traits*. (Nature Publishing Group, 2018).

103. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545 LP – 15550 (2005).