Title

Multivariate genetic analyses of 2.2 million individuals reveal broad and substance-specific pathways of addiction risk

Journal

Nature Mental Health, 4(4)

ISSN

2731-6076

Authors

Poore, Holly E

Chatzinakos, Chris

Leger, Brittany

et al.

Publication Date

2026

DOI

10.1038/s44220-026-00608-6

Peer reviewed

nature mental health

Article

https://doi.org/10.1038/s44220-026-00608-6

Multivariate genetic analyses of 2.2 million individuals reveal broad and substance-specific pathways of addiction risk

Received: 7 January 2025

Accepted: 13 February 2026

Published online: 20 March 2026

Check for updates

Holly E. Poore[1] ✉, Chris Chatzinakos[2], Brittany Leger[3], Jean Gonzalez[3,4], Travis T. Mallard[5,6], Fazil Aliev[1], Alexander Hatoum[7], Irwin D. Waldman[8], Sandra Sanchez-Roige[3,4,9], Abraham A. Palmer[3,4], K. Paige Harden[10,11], Danielle M. Dick[1,12] ✉ & Peter B. Barr[2,12]

Ongoing efforts to identify genes involved in substance use disorders (SUDs) often focus on individual disorders despite high rates of co-occurrence with each other and other externalizing traits. Here we investigate whether incorporating data on other externalizing traits can boost power to detect without sacrificing specificity of SUD genetic signal. We used multivariate genomic analyses and downstream biological annotation and genetic association analyses to explore this question. We found that joint analysis of SUDs and other externalizing traits resulted in increased insights into the neurobiology of broad and substance-specific SUD risk. We found no evidence of loss of specificity for SUD genetic signal but note improvements in our ability to characterize the neurobiology of broad and substance-specific SUD genetic effects. Our findings suggest that genetic risk for SUDs operates largely via pathways shared with other behaviors characterized by behavioral disinhibition, with additional substance-specific risk, and that modeling this shared disposition improves gene discovery.

Substance use disorders (SUDs) are common[1] and associated with substantial cost to the individuals, their families and society[2]. They are also moderately heritable (~40–60%)[3–5], leading to several ongoing efforts to identify genes that confer risk for SUDs. Most of these studies focus on specific drugs (for example, smoking[6] or alcohol separately). However, SUDs frequently co-occur, and twin studies indicate that this overlap is largely due to shared genetic influences. This shared genetic influence, which broadly impacts SUD risk, accounts for up to 74–80% of the genetic influences on alcohol use disorders (AUDs) and 62–74% of genetic influences on other SUDs[5,7], with the remaining genetic risk being substance specific. More recently, multivariate genome-wide association studies (GWAS) have been applied to SUDs, providing further evidence that most genomic risk is shared[8–11].

We also know from epidemiological and twin studies that SUDs share phenotypic and genetic influences with other outcomes characterized by behavioral disinhibition; SUDs load together with childhood conduct disorders, adult antisocial behavior and personality traits on a common underlying genetic factor[12,13]. In twin studies, this

[1]Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ, USA. [2]Department of Psychiatry and Behavioral Science, SUNY Downstate Health Sciences University, Brooklyn, NY, USA. [3]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA. [4]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA. [5]Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [6]Department of Psychiatry, Harvard Medical School, Cambridge, MA, USA. [7]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA. [8]Department of Psychology, Emory University, Atlanta, GA, USA. [9]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [10]Department of Psychology, University of Texas at Austin, Austin, TX, USA. [11]Population Research Center, University of Texas at Austin, Austin, TX, USA. [12]These authors jointly supervised this work: Danielle M. Dick, Peter B. Barr. ✉e-mail: holly.poore@rutgers.edu; danielle.m.dick@rutgers.edu

underlying latent factor is highly heritable (~80%), more so than any of the disorders or traits studied individually. In the psychological literature, this spectrum of behaviors and disorders, including SUDs, is typically referred to as the externalizing spectrum[14,15].

Despite this strong evidence that SUDs and other traits related to behavioral disinhibition can be modeled with a common dimension of phenotypic and genotypic variation, most gene identification efforts for SUDs study each disorder in isolation or focus on a single dimension of SUD risk[8,10]. The question continues to arise: is the genetic variation captured by externalizing distinct from genetic risk for SUDs or are the genetic influences on SUDs largely those that impact multiple externalizing conditions? The answer to this question can be used to inform the best way to study the etiology of SUDs. If genetic variance between SUDs and externalizing traits is mostly shared, we can use well-powered externalizing GWAS results to boost signal for SUDs. However, if genetic variation is sufficiently distinct, we risk masking SUD-specific genetic signal in a joint analysis.

As a first step in addressing this question, our group recently published a study using Genomic SEM[16] to test alternative models of the relationship between four SUDs, for which large GWAS are now available, and six other externalizing traits used in a previous GWAS of externalizing[17]. The two best fitting models were (1) a common factor in which all SUDs and other externalizing traits loaded onto a single factor, indicating that genetic risk for SUDs is largely indistinguishable from broad externalizing risk, and (2) a two-factor model in which the factor representing shared SUD risk was correlated with a factor representing behavioral disinhibition, suggesting that there was substantial overlap between SUDs and other externalizing traits, but some specificity for SUDs remained. Either way, results from this study suggested that the genetic influences on these two sets of phenotypes were highly correlated. Although both models fit well, the correlation between the two dimensions in the second model was ~0.90, suggesting that they share the majority of their genetic influences.

In the current study, we carry forward both models to conduct multivariate genomic analyses from data on >2.2 million individuals on these three factors. We use Genomic SEM and a series of downstream biological annotation and genetic association analyses to answer three main questions. (1) Does joint analysis of behavioral disinhibition traits and SUDs result in increased power and/or decreased specificity to detect genetic effects for SUDs? (2) What can we learn about the neurobiology specific to each SUD (that is, after accounting for what SUDs share with other externalizing indicators)? (3) What do we learn about risk for addiction broadly when we analyze these traits together?

Results

Multivariate GWAS

We used Genomic SEM[16] to perform multivariate GWAS on latent factors that posit different ways that genetic influences impact SUDs and correspond to two models previously estimated by this group[18] (Supplementary Table 1a,b). The first model hypothesizes that a shared set of genes influences SUDs along with other externalizing phenotypes, such as attention deficit hyperactivity disorder (ADHD) and substance initiation (see below for further discussion of each phenotype). This shared genetic dimension is shown in box A of Fig. 1. This model also facilitates examination of the genetic influences on each SUD after accounting for what it shares with other SUDs and externalizing traits (boxes B–E in Fig. 1). These influences reflect residual, substance-specific genetic effects after accounting for what is shared with each of the other externalizing indicators, hereafter referred to as 'residual SUD'). The second model hypothesizes that separate, but correlated, dimensions of genetic risk influence SUDs and other externalizing traits, here referred to as behavioral disinhibition (boxes F and G in Fig. 1). In our previous work[18], both models emerged as viable candidates; thus, in this study, we carry both forward into the multivariate GWAS analyses to compare their ability to facilitate gene identification for SUDs.

As this study is a direct extension of and requires comparison to results from our previous work[17,18], we retain the GWAS used in those studies as indicators in the current models. These indicators include ADHD[19], cannabis initiation (CANN)[20], smoking initiation (SMOK)[21], age at first sexual intercourse (FSEX)[22], number of sexual partners (NSEX)[22], risk taking propensity (RISK)[22], problematic alcohol use (PAU)[23], problematic tobacco use (PTU)[9], opioid use disorder (OUD)[24] and cannabis use disorder (CUD)[25]. Owing to concerns about the use of FSEX as an externalizing indicator[26,27], we ran sensitivity analyses leaving out this indicator, which indicated that it did not exert undue influence on the composition of the factor (see Supplementary Section 1 and Supplementary Table 2 for more details). All GWAS included only individuals whose genomes were most similar to those from reference panels sampled from Europe (hereafter referred to as 'EUR') with a total sample size of 2,219,357 unique individuals (Supplementary Table 3). After merging across the 10 sets of summary statistics, 5,963,905 single nucleotide polymorphisms (SNPs) were available for analysis.

We identified 708 loci in the single-factor externalizing model, in which SUDs were modeled as part of the externalizing spectrum (Supplementary Table 1a). In the two-factor model, we identified 631 and 48 genomic risk loci (679 total) for the behavioral disinhibition and SUD factors, respectively (Fig. 2 and Supplementary Table 4; see Supplementary Section 2 for additional details). Of the 708 externalizing loci and their correlates within linkage disequilibrium (LD) regions ($r^2 > 0.1$), 187 (26%) were not identified in the previous externalizing[17] or addiction risk[8] GWAS and 403 (57%) were not previously associated with a substance use trait in the GWAS literature (as reported in the NHGRI-EBI GWAS Catalog[28] version e114_r2025-06-27). Novel variants include rs11692435 mapped to *ACTR1B*, which is involved in a variety of cellular functions and has been previously implicated in alcoholic gastritis, alcohol consumption and smoking initiation, and rs10145520 mapped to *RAB2B*, which is involved in GTP binding and has been previously implicated in smoking initiation and educational attainment.

We identified 14, 23 and 1 genomic risk loci for the residual PAU, PTU and CUD phenotypes, respectively, whereas the residual OUD GWAS did not identify any genome-wide significant hits (Supplementary Table 5). Many of these loci were mapped to genes involved in pharmacokinetic (for example, metabolism) or pharmacodynamic (for example, drug targets) processes for specific substances. For example, residual PAU loci were mapped to genes in the alcohol dehydrogenase family (*ADH1B*, *ADH4*, *ADH5* and *ADH6*), which is involved in metabolism of alcohol, and loci for residual PTU were mapped to genes in the nicotinic acetylcholine receptor family of proteins (*CHRNA3*, *CHRNA5*, *CHRNA6*, *CHRNB3* and *CHRNB4*).

Biological annotation

We used four methods to identify genes associated with the three latent genomic factors: multi-marker analysis of genomic annotation (MAGMA, version 1.08)[29] (Supplementary Tables 6 and 7), MetaXcan[30] (Supplementary Table 8), summary-based Mendelian randomization (SMR[31]; Supplementary Table 9) and Hi-C-coupled MAGMA (H-MAGMA[32]; Supplementary Tables 10 and 11). MAGMA tissue expression analyses are shown in Supplementary Section 3, Supplementary Figs. 1–3 and Supplementary Tables 12 and 13.

Identifying genes for latent genomic factors. To account for differences in gene-based mapping methods, we further evaluated the intersection of genes that were identified in all four gene-based analyses. The intersection of genes identified by these four methods resulted in 113 genes for externalizing, and 97 and 4 (101 total) genes for behavioral disinhibition and SUDs, respectively (Fig. 3a and Supplementary Table 14). We next identified genes that were unique to each factor (that is, associated with 1 and not the other 2 factors) and found 37 unique genes for externalizing and 21 for behavioral disinhibition and 3 for SUDs in the 2-factor model (Fig. 3b). Of the 37 unique externalizing genes, 30 (81%)

https://doi.org/10.1038/s44220-026-00608-6

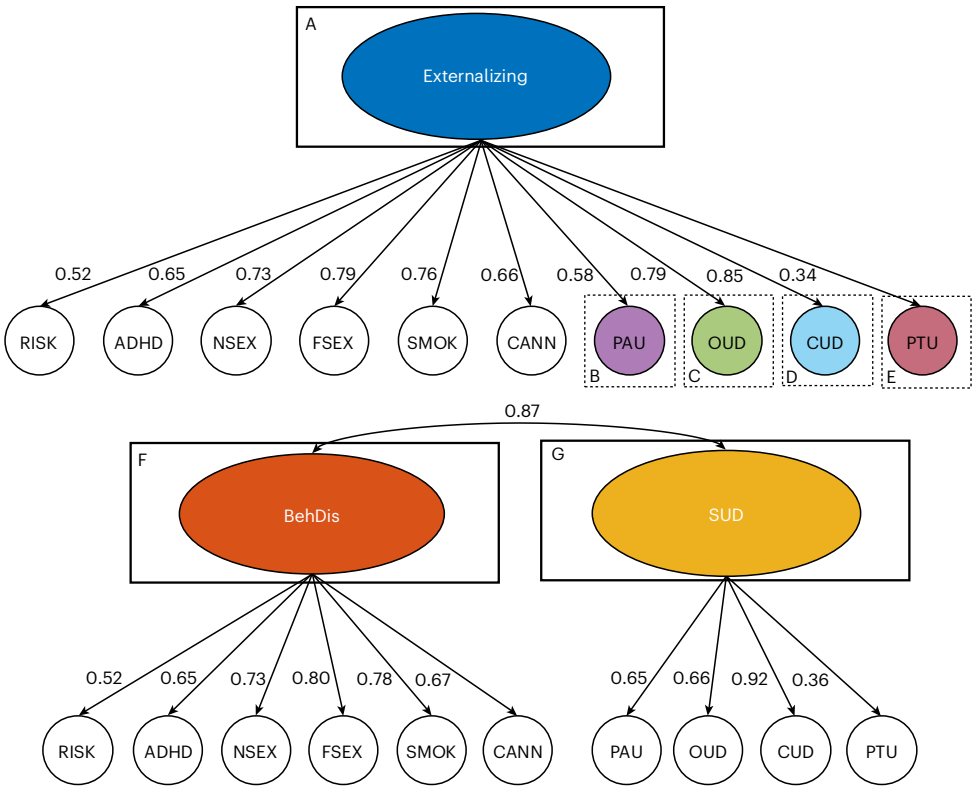

Fig. 1 | Path diagrams of models used in the current analyses. Box A represents the broad externalizing factor onto which all behavioral disinhibition and SUD phenotypes load. Boxes B–E represent residual SUD phenotypes. Boxes F and G represent narrower factors reflecting behavioral disinhibition (BehDis) and SUD phenotypes, respectively. Single-headed arrows indicate factor loadings, whereas the double-headed arrow indicates a correlation between the two factors.

have been previously associated with a substance use phenotype (for example, initiation, consumption or use disorder)[28], including *SMIM19*, which has been associated with nicotine dependence, and *KLHL29*, which has been associated with cannabis, nicotine and alcohol use. Of the three unique SUD genes, *PPP6C* has been associated with CUD, AUD and OUD.

Identifying residual SUD genetic influences. Only genes for residual PAU and PTU were identified in multiple methods (and only using MAGMA and H-MAGMA methods), resulting in 21 and 25 genes, respectively (Supplementary Table 15). We saw some substance specificity in the genes associated with the residual SUD phenotypes. Most genes identified for the residual SUDs were previously associated with a substance use phenotype, and many were associated only with a phenotype for that specific substance. For example, of the 21 genes associated with residual PAU, the 13 genes that were previously associated with a substance use trait were only associated with an alcohol and not with other substances (excluding multivariate analyses of multiple substances).

Network analysis. We used gene-network analysis[33,34] for each factor to identify genes that are part of the implicated biological networks, even if they are not directly identified by the GWAS (Fig. 4a; see Supplementary Section 4 and Supplementary Fig. 4 for more details). The externalizing and behavioral disinhibition networks were similar to one another, whereas the SUD network shared few genes, likely owing to the limited power of the SUD gene set. We then used hierarchical clustering[35] and gene ontology[36,37] enrichment to identify the gene communities within these networks (Supplementary Table 16). We found that the externalizing network contains gene communities that function in postsynaptic protein localization, organization and signaling, as well as vesicle-mediated transport and cytoskeleton-dependent transport (Fig. 4b). We further validated our findings using annotations in the GWAS Catalog. We found that the externalizing network was significantly enriched for genes identified in psychiatric disorders ($P = 2.2 \times 10^{-7}$), substance-related disorders ($P = 2.5 \times 10^{-10}$) and substance abuse ($P = 4.5 \times 10^{-14}$; Supplementary Table 17). The behavioral disinhibition network showed a similar pattern of enrichment, whereas the SUD network was not significantly enriched for genes identified in any traits in the GWAS Catalog.

Drug repurposing. We next mapped genes associated with factors and residual SUDs to the Drug–Gene Interaction Database to identify druggable targets, with a particular emphasis on drugs with FDA approval for use in medication-assisted treatment (MAT) of SUDs (Supplementary Table 18). For the factors, we used genes identified in at least three of the four gene-based methods. Given their lower power and the exploratory nature of these analyses, we used genes identified in any gene-based method for the residual SUD effects. We identified 118 druggable targets resulting in 1,024 US Food and Drug Administration (FDA)-approved drug–gene interactions for externalizing. Identified drugs included those used in MAT, including naltrexone, methadone hydrochloride, disulfiram, varenicline, baclofen and acamprosate calcium. Genes identified for behavioral disinhibition and SUDs resulted in 96 and 3 druggable targets and 704 and 11 drug–gene interactions, respectively. Drugs used in MAT were also identified for each factor, but fewer types of drug and at a lower rate compared with externalizing. We found 16 and 28 druggable targets and 181 and 231 drug–gene interactions for residual PAU and PTU, respectively (Supplementary Table 19). We found that the associations for residual PAU were specific; the only MAT drugs identified, disulfiram and naltrexone, are used to treat PAU. By contrast, drugs used in MAT identified for residual PTU spanned a wider range of SUDs (for example, varenicline, methadone hydrochloride and disulfiram).

https://doi.org/10.1038/s44220-026-00608-6

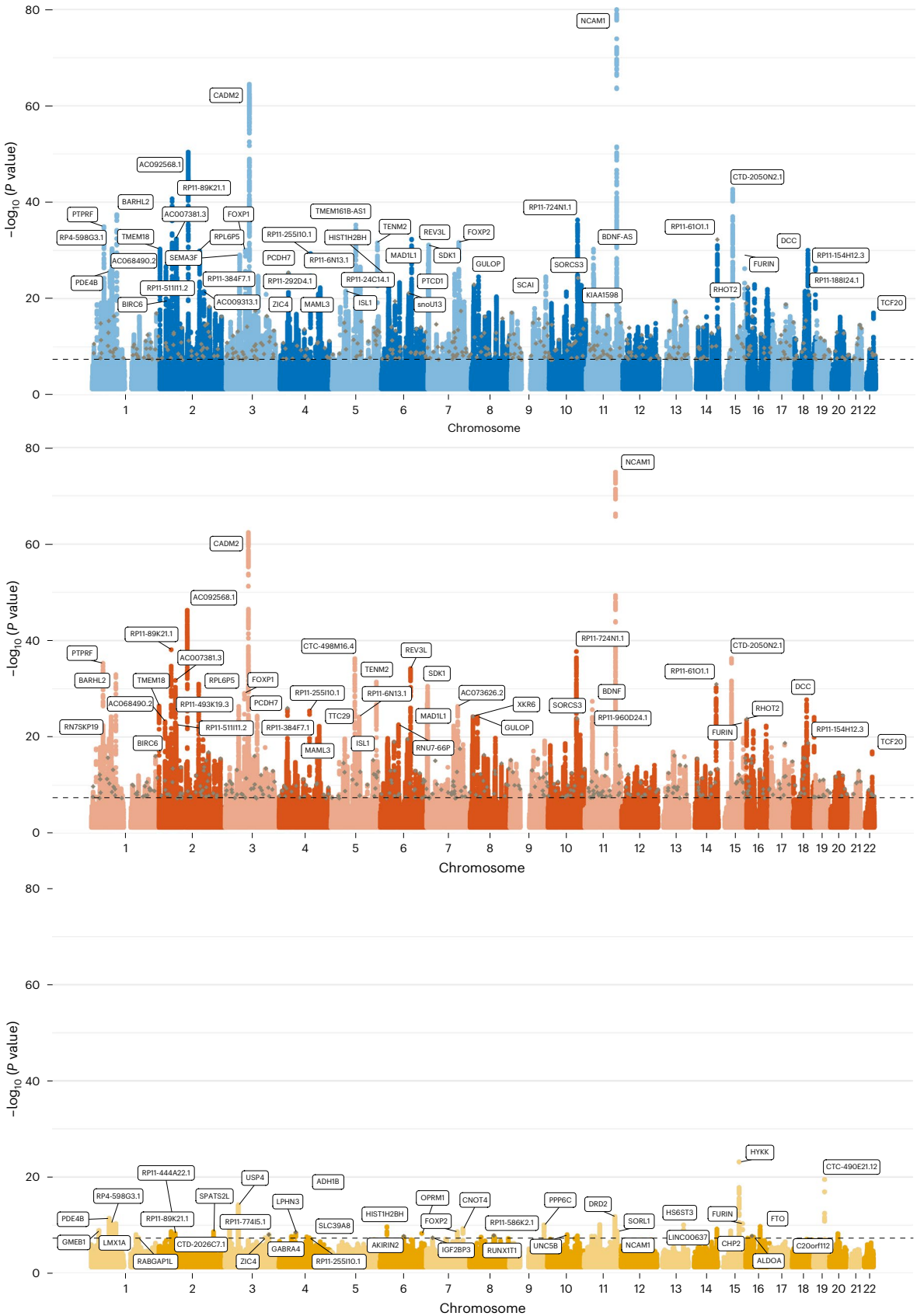

Fig. 2 | Manhattan plots of (top to bottom) externalizing, behavioral disinhibition and SUD. Brown points represent novel SUD loci (loci not previously associated with a substance use trait). Top loci are mapped to the nearest gene using ANNOVAR[63] annotation. The dotted line represents genome-wide significance at $P < 5 \times 10^{8}$ (Bonferroni correction in a two-sided test).

https://doi.org/10.1038/s44220-026-00608-6

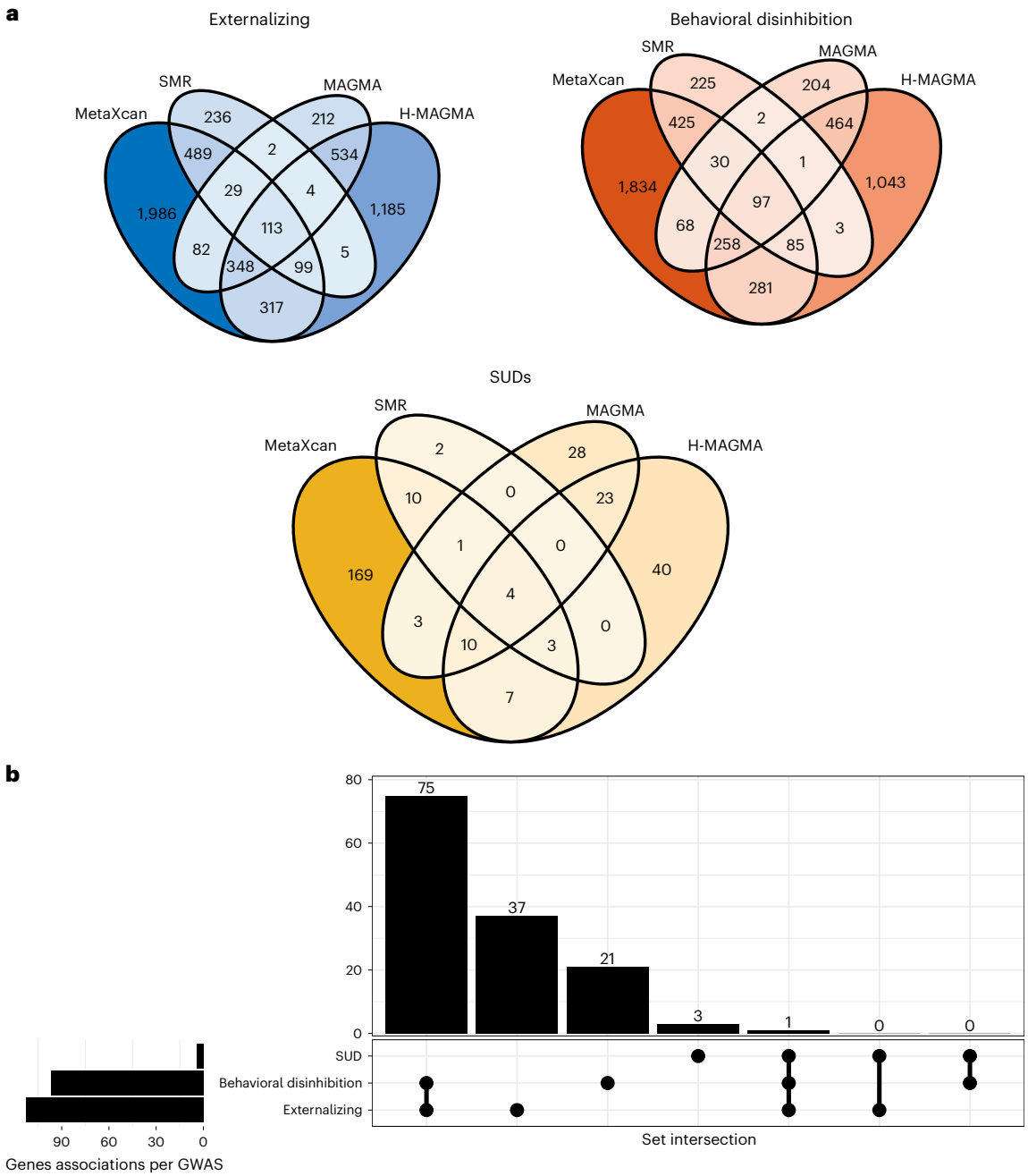

Fig. 3 | Gene based results for externalizing, behavioral disinhibition, and SUD. a, Venn diagrams showing the overlap of genes identified by the MetaXcan, SMR, MAGMA and H-MAGMA analyses for externalizing (blue), behavioral disinhibition (orange) and SUD (yellow). **b**, UpSet plot showing intersecting sets of high-confidence genes (those identified by all three gene-based methods) across externalizing, behavioral disinhibition and SUD.

Genetic correlations

Next, we further characterized the residual SUDs by estimating the genetic correlations of each of the four univariate (original) SUD GWAS, and the GWAS of the residual SUDs (after removing genetic influences shared with other externalizing conditions in the externalizing model) with external correlates (Fig. 5a and Supplementary Tables 20–23). Two notable patterns emerged. First, residual SUD effects appeared to capture substance-specific effects in an expected manner. For example, all four univariate SUDs were significantly genetically correlated with maximum alcoholic beverage (r_G values ranged from 0.30 to 0.76); however, after accounting for variance shared with other externalizing traits, only residual PAU remained significantly correlated ($r_G = 0.41$, $P < 0.001$). Second, many residual SUD genetic effects retained significant associations with other forms of psychopathology, highlighting the complex associations between SUDs and both internalizing and thought disorder psychopathology. In other words, the residual genetic effects reflect not only substance-specific effects such as drug metabolism but also genetic influences that impact that SUD through other mechanisms not shared with externalizing.

Polygenic scores

We calculated polygenic scores (PGS) from GWAS of each latent genomic factor (externalizing, behavioral disinhibition and SUD) among EUR individuals from the Collaborative Study on the Genetics of Alcoholism (COGA; $N = 7{,}530$), a multi-site family-based study[38], and All of Us ($N_{max} = 77{,}442$), a national cohort study[39,40]. We also calculated PGS

https://doi.org/10.1038/s44220-026-00608-6

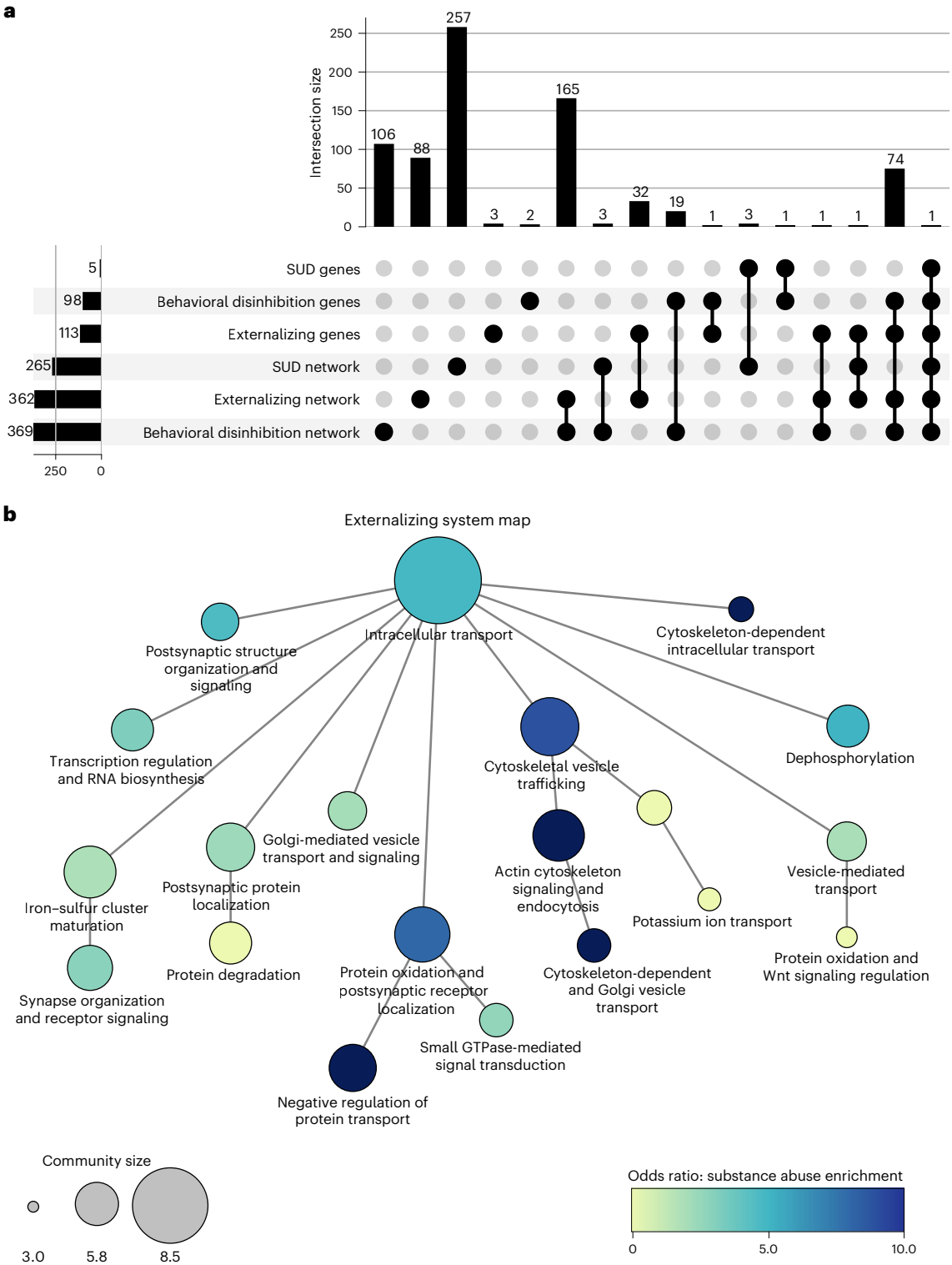

Fig. 4 | Network annotation. a, UpSet plot showing the overlap in genes identified from gene mapping algorithms (Fig. 3a) and network analysis. **b**, System map for the externalizing network. Community labels represent aggregated significant gene ontology enrichment labels, and node size indicates log-transformed community size, measured by the number of genes. Color indicates the odds ratio of the enrichment of that community for genes annotated in the GWAS Catalog for substance abuse. Only labeled communities and those connecting labeled communities are shown.

for the residual SUD effects (for alcohol, opioid, tobacco and cannabis use disorders) in COGA. These analyses allowed us to compare the total variance explained by each of the PGS and identify the specificity, or loss thereof, in polygenic prediction of SUDs. In COGA, we report associations with AUD, CUD, OUD and symptoms from the Fagerstrom Test for Nicotine Dependence (FTND)[41]. In All of Us, we report associations with AUD, tobacco use disorder (TUD) and drug use disorder, which includes all other illicit substances other than alcohol and tobacco.

Factor PGS. Externalizing PGS (EXT_{PGS}) explained the most variance in all SUDs across both samples except AUD diagnosis in All of Us, which was slightly better predicted by SUD_{PGS}. In COGA, EXT_{PGS} explained

https://doi.org/10.1038/s44220-026-00608-6

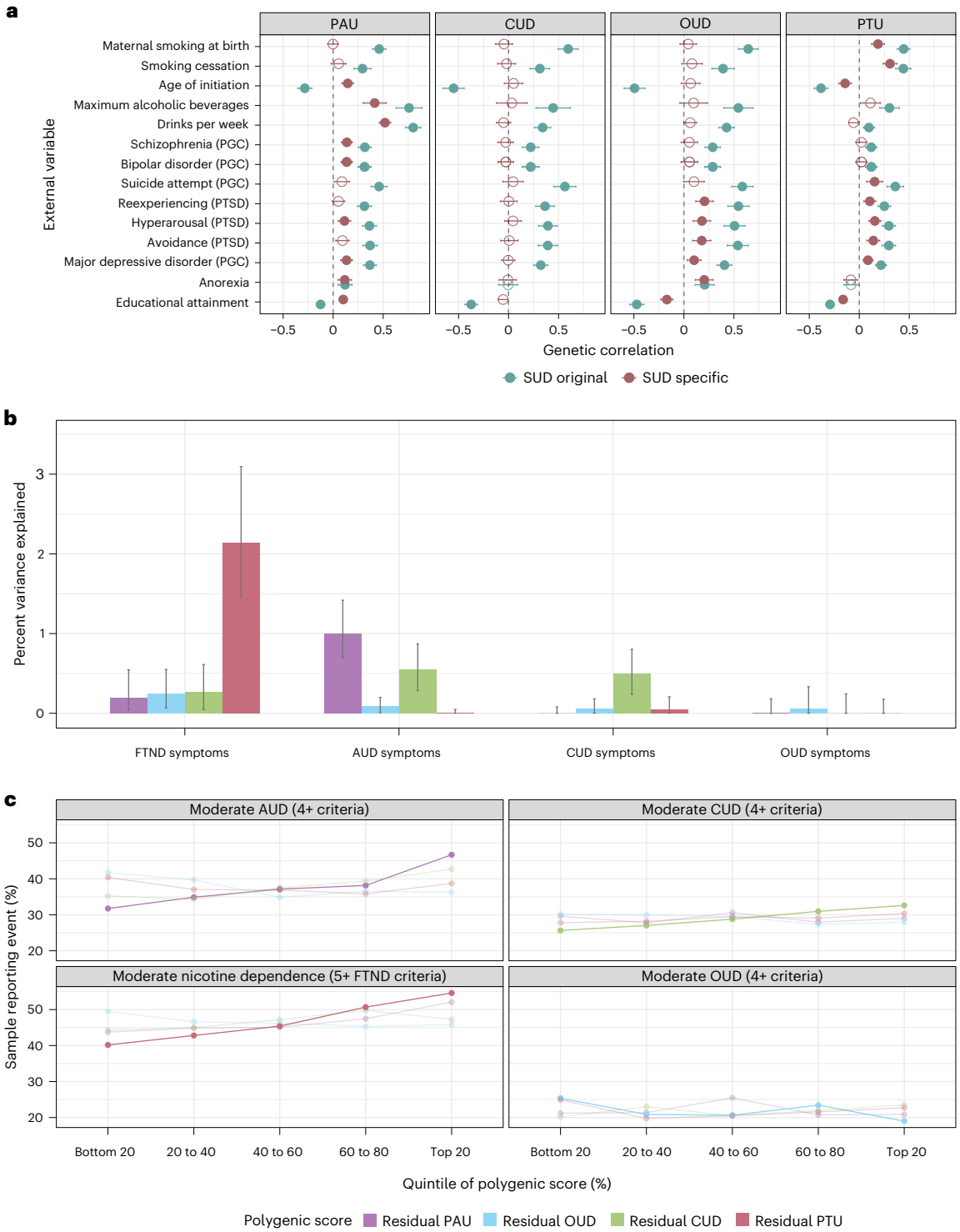

Fig. 5 | Genetic association analyses with residual SUDs. a, Genetic correlations and their 95% confidence intervals of the original SUD indicator GWAS (blue) and the residual SUD genetic effects (red) with relevant outcomes. Filled-in dots indicate correlations that were statistically significant after Bonferroni correction, whereas open dots indicate correlations that were not statistically significant. Sample sizes underlying the genome-wide association summary statistics used for these analyses are shown in Supplementary Tables 20–23. **b**, R^2 estimates and their 95% confidence intervals of variance explained by residual SUD PGS in SUDs in COGA (N = 3,671 for FTND symptoms, N = 7,252 for AUD symptoms, N = 5,014 for CUD symptoms and N = 1,663 for OUD symptoms). **c**, Corresponding relative risk of moderate SUD across quintiles of PGS. Matched SUD phenotype and residual SUD PGS are shown in a darker color relative to unmatched PGS.

between 1.6% (CUD) and 3.3% (AUD) of the variance and SUD_{PGS} explained 0.39% (OUD) and 2.9% (FTND symptoms; Supplementary Table 24). In All of Us, EXT_{PGS} explained between 3.1% (AUD) and 7.1% (TUD) of the variance in SUD diagnosis and SUD_{PGS} accounted for 3.1% (AUD) and 4.4% (drug use disorder; Supplementary Table 25).

Residual SUD PGS. We next compared the variance explained by the residual SUD PGS in SUDs in COGA (Fig. 5a). Here, too, we saw evidence of substance-specific associations such that FTND, AUD and CUD symptoms were most strongly associated with the PGS that corresponds to their substance (for example, $resPTU_{PGS}$ was most strongly associated

with FTND symptoms). In Fig. 5b, we show the rates of moderate SUD across quintiles of each residual PGS, with the matched SUD phenotype and residual SUD PGS shown in a darker color. Here we see that, especially for residual PAU and PTU PGS, the prevalence of the disorder is lowest in the lowest quintile and increases in a more linear fashion across quintiles of risk in matched relative to unmatched PGS.

Discussion

Despite evidence that SUDs share phenotypic[42] and genetic[8,12,18] variance with each other and with other externalizing traits, gene identification efforts continue to study individual SUDs in isolation or treat them as distinct from externalizing. This leaves potentially important genetic variance untagged and may, in part, explain the slow progress in gene discovery for certain SUDs relative to other complex behavioral phenotypes[43]. The role of externalizing in SUDs is also important for prevention, intervention and treatment. If SUD genetic risk largely reflects broad externalizing risk, it suggests that individuals at risk for or currently experiencing an SUD are also at risk for other externalizing problems, which should be taken into account in prevention and intervention efforts. Here we drew from the phenotypic classification and twin literature on the nature of genetic influences on SUDs[5,7,12] and capitalized on recent advances in multivariate statistical genetic methods[16], to model the shared genetic architecture of SUDs and other externalizing phenotypes with the goal of improving gene discovery for SUDs.

We performed multivariate GWAS, biological annotation and genetic association analyses using the factors and residual SUD genetic effects identified in our study to address our three main questions related to improving insights into the neurobiology of broad and specific genetic effects on SUD. Our results demonstrate that modeling genetic covariance of SUDs alongside related externalizing traits improves gene discovery for SUDs and did not result in loss of specificity to detect SUD-specific effects. We base this argument on the observation that, across analyses, the broad externalizing factor was better powered to detect genetic variants for and account for variance in SUDs relative to the more specific factors and that analyzing SUDs separately did not result in novel findings related to SUDs.

Analysis of the residual SUD genetic effects revealed several interesting insights into their neurobiology. First, many genes identified for the residuals were involved in substance-specific pharmacokinetic and pharmacodynamic processes (for example, *ADH1B* and *CHRNA5*), potentially indicating that genetic liability to addiction is made up of risk general to externalizing and risk related to an individuals' biological sensitivity to each specific substance. Second, genetic correlation analyses suggested that most cross-substance genetic associations (for example, correlations between CUD and alcohol phenotypes) are explained by shared externalizing variance. In addition, most SUDs retained significant associations with other forms of psychopathology (for example, internalizing and thought disorders), highlighting the complex interface between SUDs and a broad range of psychopathology. Finally, the residual PGS showed substance-specific prediction such that the residual PGS for a specific SUD best predicted that SUD in an independent sample. This highlights the potential translational utility of broad and specific PGS whereby a broader metric of risk captures an individual's general liability to addiction, whereas specific metrics can provide insight into risk for problems with specific substances. Relative to the previous externalizing and addiction risk GWAS[8,17], we identified 187 novel genetic variants using our expanded externalizing model, including several previously implicated in substance use traits. Our expanded externalizing factor also picked up genes not identified by the other factors and which were enriched for substance use outcomes, highlighting the benefit of joint analysis for these phenotypes.

These findings should be interpreted in light of a few limitations. First and foremost, our analyses include only data from participants of European ancestry, thereby limiting the generalizability of our results. This was a practical choice as we relied on previously published GWAS and sufficiently powered GWAS of non-European ancestry samples are not yet available for these outcomes. Second, although our sensitivity analyses indicated that this phenotype did not exert undue influence on the composition of externalizing, we note that sources of individual differences in age at first sex might differ across populations[44,45]. Third, results for individual residual analyses are dependent on both the power of the original GWAS indicator and the loadings of that indicator on its factor, which determines how much variance is leftover to analyze. Finally, although SUDs manifest the strongest relationships with other externalizing phenotypes, they also have a complex interface with many other forms of psychopathology, including internalizing[46,47], which are not modeled in the current study.

Conclusion

Our analyses demonstrate that genetic influences on SUDs operate through broad externalizing and substance-specific pathways. Our findings suggest that specific genetic influences are largely related to pharmacokinetic and pharmacodynamic processes involved in the body's response to the substance. Genetic propensity for the development of an SUD, therefore, can be conceptualized as a combination of risk for behavioral disinhibition and physiological response to the individual substance, which may make one more like to initiate and maintain its use. We further demonstrate the value of studying SUDs in the context of other externalizing traits and future studies aimed at gene discovery for SUDs should continue to leverage their genetic covariance to improve power for gene identification and downstream biological annotation and prediction analyses.

Methods

This research was reviewed and approved by the Institutional Review Board at Rutgers University (Pro2022000138).

Multivariate GWAS

This study was approved by the Rutgers University Institutional Review Board. We used Genomic SEM[16] to estimate SNP effects in two models (Fig. 1). In our two-factor model, the SUD factor included PAU[48], PTU[9], OUD[24] and CUD[25]. The remaining non-SUD indicators from the original externalizing model (ADHD[19], RISK[22], NSEX[22], FSEX[22], SMOK[21] and CANN[20]) loaded onto a behavioral disinhibition factor (boxes B and C in Fig. 1). In our common factor model, all behavioral disinhibition phenotypes and SUDs loaded onto a common externalizing factor (box A in Fig. 1). We retained the same GWAS as were used in the respective original multivariate GWAS except in the case of OUD, for which a better powered GWAS was available.

Clumping was performed in Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) version 1.6.1 (ref. 49) using an r^2 threshold of ≥0.6 to define independent significant SNPs, a second threshold of $r^2 \geq 0.1$ to define lead SNPs, and a maximum distance between LD blocks of 250 kb to merge into a locus.

To characterize differences in statistical power among the multivariate GWAS, we examine the mean χ^2, λ_{GC} and number of genome-wide significant risk loci for each factor and residual SUD (Supplementary Table 1c). We evaluated the novelty of our findings by comparing the genomic risk loci identified in our analyses with (1) loci identified in the original externalizing and addiction risk GWAS and (2) loci previously identified for substance use phenotypes in the GWAS literature. This latter test was performed by comparing the genomic risk loci for our factors and correlated SNPs ($r^2 > 0.1$) with those in the NHGRI-EBI GWAS Catalog[28] (version e114_r2025-06-27). Finally, we assessed the relative performance of our two models by comparing the degree of heterogeneity of SNP effects. We did so by calculating Q_{SNP} heterogeneity statistics, which can be used to identify SNPs that have an effect on one or more indicator phenotypes that is better explained by pathways independent of the factor. In other words, the Q_{SNP} test is

designed to ensure that SNPs are not being detected with the shared factor primarily owing to their association with one of the constituent indicators. If a factor is truly capturing the majority of genetic variance shared among its indicator phenotypes, there will be few Q_{SNP} loci relative to the number of factor loci.

Biological annotation

Gene-based methods. We used four methods to identify genes associated with the three latent genomic factors. First, we used multi-marker analysis of genomic annotation (MAGMA, version 1.08)[29], in which genome-wide SNPs were mapped to 18,235 protein-coding genes from Ensembl v102, and SNPs within each gene were jointly tested for association with each factor. We evaluated Bonferroni-corrected significance adjusted for the number of genes (one-sided $P < 2.74 \times 10^{-6}$). Second, we used Hi-C-coupled MAGMA (H-MAGMA)[32], which builds on MAGMA by leveraging chromatin interaction profiles from human brain tissue. We evaluated Bonferroni-corrected two-sided *P*-value thresholds, adjusted for multiple testing within each analysis (two-sided $P < 1.72 \times 10^{-7}$ for residual PAU and $P < 1.62 \times 10^{-7}$ for all other phenotypes). Next, we used MetaXcan[30] to conduct a transcriptome-wide association study (TWAS) using genetically regulated expression models from GTEx v8 (ref. 50). This analysis leveraged GWAS summary statistics to estimate gene–trait associations. Within-tissue Bonferroni correction was applied to identify statistically significant TWAS genes. Finally, we used summary-data-based Mendelian randomization (SMR)[31] to (1) test the extent to which gene expression mediated the relationship between SNPs and the phenotype and (2) identify genes that are more likely to be functionally relevant to the phenotype using the heterogeneity in dependent instruments (HEIDI) test. The HEIDI test distinguishes causality and pleiotropy models, in which the effect of a genetic variant on a trait is mediated by gene expression or where the genetic variant has direct effects on both the trait and expression, from the linkage model, in which associations between gene expression and trait are due to LD between two distinct causal variants. Evidence of causal or pleiotropic effects suggests that the gene should be prioritized for follow-up analyses. We identified genes of interest as those that met SMR test Bonferroni correction significance threshold and had a HEIDI test two-sided $P > 0.05$. We also used MAGMA to identify genes associated with each residual SUD phenotype.

Network analysis. We used gene-network analysis to compare the biological processes underlying externalizing, behavioral disinhibition and SUDs. Using a random-walk[33] algorithm within the PCNet2.0 (ref. 34) interactome, we generated gene networks for each latent genomic factor. As described previously[17,51], network propagation was computed using NetColoc[33], using an alpha = 0.5 and a *z*-score cut-off of 3. We generated a multiscale systems map for each network using the Hierarchical community Decoding Framework (HiDeF v1.1 Beta) algorithm as implemented in Cytoscape's Python package CDAPS[52], using a maximum resolution of 2, and enrichment for Gene Ontology (GO) terms limited to terms with 50–1,000 annotations, a minimum of 3 genes overlapping and enrichment two-sided $P < 0.05$. All significant annotations for each community were manually aggregated into a single label for each community. GWAS Catalog enrichment was calculated using Experimental Factor Ontology-based phenotype grouping for the GWAS Catalog, downloaded from https://www.ebi.ac.uk/gwas/docs/file-downloads on 7 August 2024. Enrichment was calculated via hypergeometric test against all PCNet2.0 nodes, and Bonferroni corrected for all traits in the Experimental Factor Ontology.

Drug repurposing analyses. Factor genes identified by at least three methods and residual genes identified in at least one method were mapped to ref. 53. We queried broadly for all FDA-approved drugs as well as looked specifically for following FDA-approved drugs used in the MAT of any SUD: buprenorphine, naltrexone, naloxone hydrochloride, methadone hydrochloride, acamprosate calcium, disulfiram, varenicline, bupropion hydrochloride and baclofen.

Genetic correlations

Genetic correlations of the univariate (that is, original) SUDs and residual SUDs with 39 external correlates were estimated in Genomic SEM. External correlates were chosen based on their putative relevance to SUDs and included variables from the domains of cognitive traits, other forms of psychopathology, personality and substance use. Residual correlations were estimated from the externalizing model, in which loadings of SUDs on the common externalizing factor and their genetic correlation with an external correlate were simultaneously estimated. Statistical significance was determined using a Bonferroni correction.

Polygenic scores

Target samples. We calculated polygenic scores among EUR individuals from the COGA ($N = 7{,}530$) and All of Us ($N_{max} = 77{,}442$).

COGA. The COGA is a large, multi-site, family-based study incorporating multiple generations with the goal of studying etiological influences on AUDs[38]. Individuals in treatment for AUD and their families were recruited into the study along with a smaller number of community-based comparison families beginning in 1991. Data were collected over four waves, culminating in data available for those included in the initial assessment, follow-up of individuals in the first wave, a prospective wave, which focused on longitudinal assessments of youth in case and comparison families, and an ongoing wave of participants now in midlife and later life stages. The present study includes data from individuals in waves 1–3 who were of EUR ancestry and had GWAS data available.

DNA samples were genotyped using the Illumina Human1 M array (Illumina), the Illumina Human OmniExpress 12V1 array (Illumina), the Illumina 2.5 M array (Illumina) or the Smokescreen genotyping array (Biorealm LLC)[54]. Data were imputed to 1000 Genomes (phase 3) and SNPs with a minor allele frequency <0.01, were genotyped at a rate <0.95, or that violated Hardy–Weinberg equilibrium were excluded. Principal components (PCs) were calculated using Eigenstrat[55] and 1000 Genomes (phase 3, version 5). The first ten PCs, age and sex were used as covariates in the polygenic score analyses (PGS). Additional details about genotyping and quality control procedures have been published elsewhere[56,57].

We report associations with problematic substance use (symptoms of AUD, CUD, OUD and other SUD as well as the FTND[41]).

All of Us. All of Us is an observational cohort study of diverse, adult (age ≥18) participants, intended to be representative of the United States with the goal of studying the effects of lifestyle, environment and genomics to improve health outcomes[39,40,58]. Data collection began in May 2018 and is ongoing. Participants are recruited through healthcare organizations, other community enrollment sites or by enrolling directly online. Participants give informed consent and complete a basic health survey online and are then invited to undergo a physical examination and to provide a biospecimen at an affiliated healthcare site. If consent is provided, additional information is collected by linking participants' electronic health records (EHR) and by completing periodic surveys assessing lifestyle, well-being, physical health and behavior. Participants in release 7 (6 May 2018 to 23 February 2023) with linked EHR and whole-genome sequencing data were included in the current study.

Participants were genotyped using the Illumina Global Diversity Array. The All of Us Research Program completed genetic ancestry classification analyses using a random forest classifier to predict ancestry from PCs. PCs, age and sex were used as covariates in the PGS analyses. Outcome SUDs include AUD, TUD and other drug disorder (that is, use disorder for all other drugs) present in the EHR. Diagnoses for all

disorders were based on phecodes, which are clusters of related billing codes from the International Statistical Classification of Diseases 9th and 10th revisions. Individuals with two or more phecodes related to single disorders were categorized as cases.

Calculation of PGS. We used PRS-CS[59] to adjust original GWAS beta weights for linkage disequilibrium and Plink2[60] to construct each PGS from these weights. We evaluated the incremental R^2/pseudo-R^2 (ΔR^2) attained by adding the polygenic score to a regression with baseline covariates (for example, age, sex and ancestry PCs). We used least squares regression for continuous outcomes and logistic regression for categorical ones and adjusted the standard errors in COGA to account for the family structure. Regression models were estimated using the estimatr package in R[61]. We estimated 95% confidence intervals for ΔR^2 using bootstrapping (1,000 iterations) in the boot R package[62].

Statistics and reproducibility

This study used secondary data analysis, and no new data were generated. We primarily used genome-wide association summary statistics drawn from previously published studies. We included GWAS that are well powered ($N > 50{,}000$) and, in cases where more than one GWAS of the same phenotype is available, chose the most recent and largest available. Target samples for PGS analyses were chosen owing to the availability of relevant outcome phenotypes and sufficient power (≥90% to detect an R^2 of 1%). No data from these target samples were excluded from the analysis.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The summary statistics used for the current study are from previously published studies and no new data were collected. These studies are described in the text. Some of these studies include data that have restricted access to protect the privacy of the study participants. The full sets of externalizing and residual SUD summary statistics generated by this study can be made available to qualified investigators who enter into an agreement with 23andMe that protects participant confidentiality, and who have access to MVP data. Once these requests have been satisfied, investigators may request to use our summary statistics via the procedures detailed at https://externalizing.rutgers.edu/request-data/.

Code availability

We used the following programs to complete our analyses: Genomic SEM (https://github.com/GenomicSEM/GenomicSEM), FUMA (https://fuma.ctglab.nl/home), MAGMA (https://cncr.nl/research/magma/), H-MAGMA (https://github.com/thewonlab/H-MAGMA), MetaXcan (https://github.com/hakyimlab/MetaXcan), SMR (https://yanglab.westlake.edu.cn/software/smr), Cytoscape (https://cytoscape.org/), PRS-CS (https://github.com/getian107/PRScs), Plink (https://www.cog-genomics.org/plink/), R (https://cran.r-project.org/) and R packages estimatr (https://cran.r-project.org/web/packages/estimatr/index.html) and boot (https://cran.r-project.org/web/packages/boot/index.html). Scripts used to run these analyses are available by request from the first author.

References

1. Mojtabai, R. Estimating the prevalence of substance use disorders in the US using the benchmark multiplier method. *JAMA Psychiatry* **79**, 1074–1080 (2022).
2. Sacks, J. J., Gonzales, K. R., Bouchery, E. E., Tomedi, L. E. & Brewer, R. D. 2010 national and state costs of excessive alcohol consumption. *Am. J. Prev. Med.* **49**, e73–e79 (2015).
3. Verhulst, B., Neale, M. C. & Kendler, K. S. The heritability of alcohol use disorders: a meta-analysis of twin and adoption studies. *Psychol. Med.* **45**, 1061–1072 (2015).
4. Verweij, K. J. et al. Genetic and environmental influences on cannabis use initiation and problematic use: a meta-analysis of twin studies. *Addiction* **105**, 417–430 (2010).
5. Kendler, K. S., Jacobson, K. C., Prescott, C. A. & Neale, M. C. Specificity of genetic and environmental risk factors for use and abuse/dependence of cannabis, cocaine, hallucinogens, sedatives, stimulants, and opiates in male twins. *Am. J. Psychiatry* **160**, 687–695 (2003).
6. Saunders, G. R. B. et al. Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* **612**, 720–724 (2022).
7. Kendler, K. S., Myers, J. & Prescott, C. A. Specificity of genetic and environmental risk factors for symptoms of cannabis, cocaine, alcohol, caffeine, and nicotine dependence. *Arch. Gen. Psychiatry* **64**, 1313–1320 (2007).
8. Hatoum, A. S. et al. Multivariate genome-wide association meta-analysis of over 1 million subjects identifies loci underlying multiple substance use disorders. *Nat. Ment. Health* **1**, 210–223 (2023).
9. Hatoum, A. S. et al. The addiction risk factor: a unitary genetic vulnerability characterizes substance use disorders and their associations with common correlates. *Neuropsychopharmacology* https://doi.org/10.1038/s41386-021-01209-w (2021).
10. Khan, Y. et al. Transdiagnostic and disorder-level genome-wide association studies enhance precision of substance use and psychiatric genetic risk profiles in African and European ancestries. *Biol. Psychiatry* https://doi.org/10.1016/j.biopsych.2025.04.021 (2025).
11. Miller, A. P., Bogdan, R., Agrawal, A. & Hatoum, A. S. Generalized genetic liability to substance use disorders. *J. Clin. Investig.* https://doi.org/10.1172/JCI172881 (2024).
12. Krueger, R. F. et al. Etiological connections among substance dependence, antisocial behavior, and personality: modeling the externalizing spectrum. *J. Abnorm. Psychol.* **111**, 411–424 (2002).
13. Kendler, K. S. et al. A Swedish population-based multivariate twin study of externalizing disorders. *Behav. Genet.* **46**, 183–192 (2016).
14. Kotov, R. et al. The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *J. Abnorm. Psychol.* **126**, 454–477 (2017).
15. Krueger, R. F. et al. Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): II. Externalizing superspectrum. *World Psychiatry* **20**, 171–193 (2021).
16. Grotzinger, A. D. et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).
17. Karlsson Linnér, R. et al. Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction. *Nat. Neurosci.* **24**, 1367–1376 (2021).
18. Poore, H. E. et al. A multivariate approach to understanding the genetic overlap between externalizing phenotypes and substance use disorders. *Addict. Biol.* **28**, e13319 (2023).
19. Demontis, D. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
20. 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).
21. 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).
22. Karlsson Linnér, R. et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* **51**, 245–257 (2019).

23. Zhou, H. et al. Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nat. Neurosci.* **23**, 809–818 (2020).
24. Deak, J. D. et al. Genome-wide association study in individuals of European and African ancestry and multi-trait analysis of opioid use disorder identifies 19 independent genome-wide significant risk loci. *Mol. Psychiatry* **27**, 3970–3979 (2022).
25. Johnson, E. C. et al. A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Psychiatry* **7**, 1032–1045 (2020).
26. Harden, K. P. A sex-positive framework for research on adolescent sexuality. *Perspect. Psychol. Sci.* **9**, 455–469 (2014).
27. Lindberg, L. D., Maddow-Zimet, I. & Marcell, A. V. Prevalence of sexual initiation before age 13 years among male adolescents and young adults in the United States. *JAMA Pediatr.* **173**, 553–560 (2019).
28. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
29. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
30. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
31. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
32. Sey, N. Y. A. et al. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* **23**, 583–593 (2020).
33. Rosenthal, S. B. et al. Mapping the common gene networks that underlie related diseases. *Nat. Protoc.* **18**, 1745–1759 (2023).
34. Wright, S. N. et al. State of the interactomes: an evaluation of molecular networks for generating biological insights. *Mol. Syst. Biol.* **21**, 1–29 (2025).
35. Zheng, F. et al. HiDeF: identifying persistent structures in multiscale 'omics data. *Genome Biol.* **22**, 21 (2021).
36. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
37. The Gene Ontology Consortium et al. The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
38. Dick, D. M. et al. The collaborative study on the genetics of alcoholism: sample and clinical data. *Genes Brain Behav.* **22**, e12860 (2023).
39. The All of Us Research Program Genomics Investigators Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
40. The All of Us Research Program Investigators The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
41. Fagerstrom, K. Determinants of tobacco use and renaming the FTND to the Fagerstrom Test for Cigarette Dependence. *Nicotine Tob. Res.* **14**, 75–78 (2012).
42. Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D. & Kramer, M. D. Linking antisocial behavior, substance use, and personality: an integrative quantitative model of the adult externalizing spectrum. *J. Abnorm. Psychol.* **116**, 645–666 (2007).
43. Deak, J. D. & Johnson, E. C. Genetics of substance use disorders: a review. *Psychol. Med.* **51**, 2189–2200 (2021).
44. Carlson, M. D., Mendle, J. & Harden, K. P. Early adverse environments and genetic influences on age at first sex: evidence for gene × environment interaction. *Dev. Psychol.* **50**, 1532–1542 (2014).
45. Geronimus, A. T. Teenage childbearing as cultural prism. *Br. Med. Bull.* **69**, 155–166 (2004).
46. King, S. M., Iacono, W. G. & McGue, M. Childhood externalizing and internalizing psychopathology in the prediction of early substance use. *Addiction* **99**, 1548–1559 (2004).
47. Hussong, A. M., Jones, D. J., Stein, G. L., Baucom, D. H. & Boeding, S. An internalizing pathway to alcohol use and disorder. *Psychol. Addict. Behav.* **25**, 390–404 (2011).
48. Zhou, H. et al. Multi-ancestry study of the genetics of problematic alcohol use in over 1 million individuals. *Nat. Med.* **29**, 3184–3192 (2023).
49. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
50. GTEx Consortium The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
51. Wright, S. N. et al. Genome-wide association studies of human and rat BMI converge on synapse, epigenome, and hormone signaling networks. *Cell Rep.* **42**, 112873 (2023).
52. Singhal, A. et al. Multiscale community detection in Cytoscape. *PLoS Comput. Biol.* **16**, e1008239 (2020).
53. Cannon, M. et al. DGIdb 5.0: rebuilding the drug–gene interaction database for precision medicine and drug discovery platforms. *Nucleic Acids Res.* **52**, D1227–D1235 (2024).
54. Baurley, J. W., Edlund, C. K., Pardamean, C. I., Conti, D. V. & Bergen, A. W. Smokescreen: a targeted genotyping array for addiction research. *BMC Genom.* **17**, 145 (2016).
55. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
56. Lai, D. et al. Genome-wide association studies of alcohol dependence, DSM-IV criterion count and individual criteria. *Genes Brain Behav.* **18**, e12579 (2019).
57. Johnson, E. C. et al. The collaborative study on the genetics of alcoholism: genetics. *Genes Brain Behav.* **22**, e12856 (2023).
58. Ramirez, A. H. et al. The All of Us Research Program: data quality, utility, and diversity. *Patterns* **3**, 100570 (2022).
59. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
60. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
61. Blair, G., Coopper, J., Coppock, A., Humphreys, M. & Sonnet, L. estimatr: fast estimators for design-based inference. https://cran.r-project.org/web/packages/estimatr/index.html (2025).
62. Canty, A. & Ripley, B. boot: Bootstrap R (S-Plus) functions. https://cran.r-project.org/web/packages/boot/index.html (2024).
63. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

Acknowledgements

This work was supported by the National Institute on Alcohol Abuse and Alcoholism (R01AA015416 to D.M.D., K02AA018755 to D.M.D., U10AA008401 to D.M.D., P50AA022537 to D.M.D. and T32AA028254 to H.E.P.), the National Institute on Drug Abuse (K01DA059657 to H.E.P., R01DA050721 to D.M.D. and DP1DA054394 to S.S.-R.), the Tobacco-Related Disease Research Program (T32IR5226 to S.S.-R. and 28IR-0070C to A.A.P.) and the National Institute of General Medical Sciences (T32 GM008666 to B.L.).

We thank The Externalizing Consortium. Principal investigators: D. M. Dick, P. Koellinger, K. P. Harden and A. A. Palmer. Lead analysts: R. K. Linnér, T. T. Mallard, P. B. Barr and S. Sanchez-Roige. Significant contributors: I. D. Waldman. The Externalizing Consortium has been supported by the National Institute on Alcohol Abuse and Alcoholism

(R01AA015416, administrative supplement to D.M.D.) and the National Institute on Drug Abuse (R01DA050721 to D.M.D.). Additional funding for investigator effort has been provided by K02AA018755, U10AA008401 and P50AA022537 to D.M.D., as well as a European Research Council Consolidator Grant (647648 EdGe to P. Koellinger). The content is solely the responsibility of the authors and does not necessarily represent the official views of the above funding bodies.
The Externalizing Consortium would like to thank the following groups for making the research possible: 23andMe Inc., Add Health, Vanderbilt University Medical Center's BioVU, Collaborative Study on the Genetics of Alcoholism (COGA), the Psychiatric Genomics Consortium's (PGC) Substance Use Disorders working group, UK10K Consortium, UK Biobank and Philadelphia Neurodevelopmental Cohort. All code necessary to replicate this study is available upon request.
We thank Million Veteran Program (MVP) staff, researchers and volunteers, who have contributed to MVP, and especially participants who previously served their country in the military and now generously agreed to enroll in the study.
We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study.
Finally, we thank the Collaborative Study on the Genetics of Alcoholism (COGA). Principal investigators B. Porjesz, V. Hesselbrock, and A. Agrawal; scientific director A. Agrawal; translational director D. Dick, including ten different centers: University of Connecticut (V. Hesselbrock); Indiana University (H. J. Edenberg, T. Foroud, Y. Liu and M. H. Plawecki); University of Iowa Carver College of Medicine (S. Kuperman and J. Kramer); SUNY Downstate Health Sciences University (B. Porjesz, J. Meyers, C. Kamarajan and A. Pandey); Washington University in St. Louis (L. Bierut, A. Agrawal and S. Hartz); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield, D. Dick, R. Hart and J. Salvatore); The Children's Hospital of Philadelphia, University of Pennsylvania (L. Almasy); Icahn School of Medicine at Mount Sinai (A. Goate and P. Slesinger); and Howard University (D. Scott). Other COGA collaborators include C. Holzhauer and M. Hesselbrock (University of Connecticut); J. Nurnberger Jr., L. Wetherill, X. Xuei, D. Lai and S. O'Connor (Indiana University); G. Chan (University of Iowa; University of Connecticut); D. B. Chorlian, J. Zhang, P. Barr, S. Kinreich, G. Pandey and Z. Neale (SUNY Downstate); N. Mullins (Icahn School of Medicine at Mount Sinai); A. Anokhin, K. Bucholz, F. Dong, A. Hatoum, E. Johnson, V. McCutcheon, J. Rice and S. Saccone (Washington University); F. Aliev, Z. Pang, S. Kuo and S. Brislin (Rutgers University); A. Merikangas (The Children's Hospital of Philadelphia and University of Pennsylvania); H. Chin and A. Parsian are the NIAAA staff collaborators. We continue to be inspired by our memories of H. Begleiter and T. Reich, founding principal investigator and co-principal investigator of COGA, and also owe a debt of gratitude to other past organizers of COGA, including T.-K. Li, P. M. Conneally, R. Crowe and W. Reich, for their critical contributions. This national collaborative study is supported by NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA).

Author contributions

H.E.P., D.M.D. and P.B.B. contributed to the conception and design. H.E.P., C.C., B.L., J.G. and P.B.B. were responsible for analysis of the data. H.E.P., C.C., B.L., J.G., T.T.M., FA, A.H., I.D.W., S.S.-R., A.A.P., K.P.H., D.M.D. and P.B.B. contributed to the interpretation of the data. H.E.P., B.L. and D.M.D. were responsible for drafting the original paper and H.E.P., C.C., B.L., J.G., T.T.M., F.A., A.H., I.D.W., S.S.-R., A.A.P., K.P.H., D.M.D. and P.B.B. contributed to revising and editing the paper. P.B.B. and D.M.D. were responsible for supervision of this work.

Competing interests

D.M.D. is a co-founder and Chief Scientific Officer for Thrive Genetics, Inc. She is on the Advisory Board for the Seek Women's Health Company. She has received royalties from Penguin Random House for her book, *The Child Code: Understanding Your Child's Unique Nature for Happier, More Effective Parenting*. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44220-026-00608-6.

Correspondence and requests for materials should be addressed to Holly E. Poore or Danielle M. Dick.

Peer review information *Nature Mental Health* thanks Romain Icick and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Holly Poore and Danielle Dick

Last updated by author(s): 01/13/2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No new data were collected and therefore no software was used for this purpose
Data analysis	No custom algorithms or software was developed in this study. Software and code for the genetic data analysis we report can be found at: Genomic SEM (https://github.com/GenomicSEM/GenomicSEM) v0.0.5 FUMA (https://fuma.ctglab.nl/home) v1.8.0 MAGMA (https://cncr.nl/research/magma/) v1.08 H-MAGMA (https://github.com/thewonlab/H-MAGMA) MetaXcan (https://github.com/hakyimlab/MetaXcan) v0.8.0 SMR (https://yanglab.westlake.edu.cn/software/smr) v1.4.0 Cytoscape (https://cytoscape.org/) v3.10.4 PRS-CS (https://github.com/getian107/PRScs) Plink (https://www.cog-genomics.org/plink/) v2.0 R v4.5.0 (https://cran.r-project.org/) and R packages estimatr v1.0.6 (https://cran.r-project.org/web/packages/estimatr/index.html) and boot v1.3.31 (https://cran.r-project.org/web/packages/boot/index.html)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The summary statistics used for the current study are from previously published studies and no new data were collected. These studies are described in the text. Some of these studies include data which have restricted access to protect the privacy of the study participants. The full sets of externalizing and residual SUD summary statistics generated by this study can be made available to qualified investigators who enter into an agreement with 23andMe that protects participant confidentiality and who have access to MVP data. Once these requests have been satisfied, investigators may request to use our summary statistics via the procedures detailed at https://externalizing.rutgers.edu/request-data/.

Restricted access individual-level phenotype and genetic data:
Collaborative Study on the Genetics of Alcoholism (COGA), dbGaP Study Accession: phs000763.v1.p1
All of Us (https://www.researchallofus.org/data-tools/data-access/)

Publicly available GWAS summary statistics:
GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN): https://conservancy.umn.edu/handle/11299/201564
International Cannabis Consortium (ICC): https://www.ru.nl/bsi/research/group-pages/substance-use-addiction-food-saf/vm-saf/genetics/international-cannabis-consortium-icc/
Psychiatric Genomics Consortium: https://www.med.unc.edu/pgc/download-results/
Social Science Genetic Association Consortium: https://www.thessgac.com

Restricted access GWAS summary statistics:
23andMe, Inc.: https://research.23andme.com/dataset-access/
Million Veterans Program: https://www.research.va.gov/mvp/

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We included data on all participants regardless of sex and gender. Sex and gender data were not available for published GWAS summary statistics. Biological sex was included as a covariate in all polygenic score analyses. Sex and gender were not considered in the study design and findings do not apply differently based on sex and gender. Disaggregated data for sex and gender were not collected.
Reporting on race, ethnicity, or other socially relevant groupings	All individuals were of European ancestry to account for differences in LD structure across ancestry and a lack of well powered GWAS available in other ancestry groups. Our results, therefore, are not representative beyond individuals of European ancestry.
Population characteristics	See behavioral and social sciences section
Recruitment	This study did not involve recruitment of new participants
Ethics oversight	All participants provided informed consent. This research was reviewed and approved the Institutional Review Board at Rutgers University (Pro2022000138).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Multivariate genome-wide association study in which genetic variants associated with broad factors representing externalizing,

Study description	behavioral disinhibition, SUDs, and residual SUDs were estimated. Subsequent bioannotation of these phenotypes was then performed using the software described above. All data are quantitative.
Research sample	Our total sample included 2,219,357 individuals of European ancestry drawn from the sources described above
Sampling strategy	Samples were drawn from GWAS summary statistics, information about which can be found in their respective publications. GWAS summary statistics were chosen from previously published studies (sources described above) that met certain inclusion criteria described elsewhere (https://www.nature.com/articles/s41593-021-00908-3). Phenotypic data for polygenic score analyses were drawn from the Collaborative Studies on the Genetics of Alcohol, a clinically ascertained sample (Max N = 7,252, mean age = 24.34, 52% female) and All of Us, a representative sample from the United States (Max N = 77,442, mean age = 50.90, 60% female).
Data collection	All analyses relied on previously collected data and no new data were collected for this study.
Timing	Not relevant as all data were previously published
Data exclusions	No data were excluded
Non-participation	No new data were collected for this study
Randomization	Not relevant as this study is not experimental

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*