nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01851-6

Multi-ancestry meta-analysis of tobacco use disorder identifies 461 potential risk genes and reveals associations with multiple health outcomes

Received: 24 March 2023

Accepted: 21 February 2024

Published online: 17 April 2024

Check for updates

Sylvanus Toikumo[1,2,23], Mariela V. Jennings[3,23], Benjamin K. Pham[3], Hyunjoon Lee[4], Travis T. Mallard[5,6,7,8], Sevim B. Bianchi[3], John J. Meredith[3], Laura Vilar-Ribó[9], Heng Xu[3], Alexander S. Hatoum[10], Emma C. Johnson[10], Vanessa K. Pazdernik[11], Zeal Jinwala[2], Shreya R. Pakala[3], Brittany S. Leger[3,12], Maria Niarchou[13], Michael Ehinmowo[14], Penn Medicine BioBank*, Greg D. Jenkins[11], Anthony Batzler[11], Richard Pendegraft[11], Abraham A. Palmer[3,15], Hang Zhou[16,17], Joanna M. Biernacka[11,18], Brandon J. Coombes[11], Joel Gelernter[16,17], Ke Xu[16,17], Dana B. Hancock[19], Nancy J. Cox[20], Jordan W. Smoller[5,6,7,8], Lea K. Davis[4,13,20], Amy C. Justice[17,21,22], Henry R. Kranzler[1,2], Rachel L. Kember[1,2] & Sandra Sanchez-Roige[3,13,15] ✉

Tobacco use disorder (TUD) is the most prevalent substance use disorder in the world. Genetic factors influence smoking behaviours and although strides have been made using genome-wide association studies to identify risk variants, most variants identified have been for nicotine consumption, rather than TUD. Here we leveraged four US biobanks to perform a multi-ancestral meta-analysis of TUD (derived via electronic health records) in 653,790 individuals (495,005 European, 114,420 African American and 44,365 Latin American) and data from UK Biobank ($n_{combined}$ = 898,680). We identified 88 independent risk loci; integration with functional genomic tools uncovered 461 potential risk genes, primarily expressed in the brain. TUD was genetically correlated with smoking and psychiatric traits from traditionally ascertained cohorts, externalizing behaviours in children and hundreds of medical outcomes, including HIV infection, heart disease and pain. This work furthers our biological understanding of TUD and establishes electronic health records as a source of phenotypic information for studying the genetics of TUD.

Tobacco use disorder (TUD) is the most prevalent substance use disorder (SUD) in the world, with 85% of smokers meeting criteria for TUD (also known as nicotine dependence)[1,2]. TUD is a problematic pattern of tobacco use that leads to clinically significant impairment or distress[2]. Individuals with nicotine dependence often experience withdrawal symptoms when they stop smoking. As a result, they often have substantial difficulty quitting and continue to smoke despite negative mental, social and medical consequences. Tobacco smoking

A full list of affiliations appears at the end of the paper. ✉e-mail: sanchezroige@ucsd.edu

is the leading cause of preventable death worldwide, causing 6 million annual premature deaths[3], and is also highly associated with other worldwide leading contributors of morbidity and mortality, including lung cancer, chronic obstructive pulmonary disease, cardiovascular disease, mood disorders and other SUDs[4–6]. Unfortunately, available preventative and treatment options for TUD have low success rates[7].

Genetic factors influence smoking behaviours, with twin-heritability estimates ranging from ~30% to 70% (refs. 8–12). Recently, genome-wide association studies (GWAS) have expanded in size (~2.5 million) and yielded hundreds of new loci for smoking-related behaviours (summarized in Supplementary Table 1), primarily for nicotine consumption[13]. These GWAS have revealed pervasive pleiotropy, with Mendelian randomization (MR) analyses highlighting potential causal effects of regular tobacco smoking on health outcomes (for example, cardiovascular health[14], cancer risk[14] and bone mineral density[15]), numerous other SUDs (for example, alcohol[14], cannabis[16] and opioid use disorder (OUD)[17]) and psychiatric and related conditions (for example, major depressive disorder[18], suicide-related behaviours[19] and loneliness[20]).

While these studies have been immensely successful, they have not focused on TUD itself, which consists of several components that begin with smoking initiation and regular use and develop into problematic use, dependence, cessation and relapse. As a result, relatively little is known about the specific genes that confer risk for the development of TUD and associated conditions. One of the main roadblocks to progress in identifying risk-conferring genes has been the lack of sufficiently large samples with misuse phenotypes. This is an important limitation because previous studies have shown that the genetic architecture of substance use is different from that of misuse[21–26]. The largest GWAS of nicotine dependence, comprising 58,000 European- and African-ancestry smokers, using the self-reported Fagerström test for nicotine dependence (FTND), identified only five loci[27]. In addition, while there have been nicotine-dependence GWAS in individuals of ancestries other than European[28] (Supplementary Table 1 for full list), sample sizes for diverse populations have been limited ($n < 12{,}000$).

The use of electronic health records (EHR) is a relatively untapped, cost-effective strategy for characterizing smoking-related phenotypes, including TUD. EHR-defined TUD generally relies on International Classification of Disease (ICD) diagnostic codes, which can be aggregated into 'phecodes' that require the presence of an ICD code on two or more separate visits. TUD diagnostic codes are effective identifiers of smoking status[29]. A key consideration, and the one we examine in this study, is the use of TUD phecodes in large-scale GWAS to boost power and improve our ability to identify new loci for TUD[29–31]. To address this question, we performed a multi-ancestral meta-analysis of TUD comprising 653,790 individuals of European (EUR), African American (AA) and Latin American (LA) ancestry recruited from several biobanks within the PsycheMERGE network[32] (Vanderbilt University Medical Center's biobank (BioVU) $n_{EUR} = 46{,}905$; Mass General Brigham Biobank (MGBB) $n_{EUR} = 22{,}268$; Penn Medicine BioBank (PMBB)[33] $n_{EUR} = 28{,}999$, $n_{AA} = 10{,}088$; Million Veteran Program (MVP) $n_{EUR} = 396{,}833$, $n_{AA} = 104{,}332$, $n_{LA} = 44{,}365$) and combined with existing data from the UK Biobank (UKBB; $n_{EUR} = 244{,}890$), which used a less stringent definition. In secondary analyses, we further characterized the genetic architecture of TUD, examined pleiotropy with other psychiatric and medical outcomes and harnessed the data to reveal potential medications for treating this serious psychiatric condition.

Results

Cohort and phenotype descriptions

We included individuals from eight cohorts across five different sites (Fig. 1a for an overview of the cohorts; Supplementary Table 2 for sample sizes). The methods to ascertain cases were identical for seven of these cohorts. Individuals were identified as cases if they met criteria for a TUD phecode (a TUD ICD9 or ICD10 code on two or more separate visits, described in Supplementary Table 3); controls were screened for the absence of a TUD diagnostic code. We benchmarked the TUD-EHR definition against self-reported smoking questionnaire data and other comorbid ICD codes (Supplementary Table 4). Across contributing biobanks, cases were enriched for ever-smokers (92–99%), with only a few (<2%) cases self-identifying as never-smokers (Supplementary Table 5). In contrast, a smaller proportion of controls were ever-smokers (17–56%), with a larger proportion self-identifying as never-smokers (39–73%). Attempts at smoking cessation were reported by 15–25% of controls and 65–95% of cases. Controls were comparable to cases for age and sex but reported much lower prevalences of other substance and psychiatric disorders than did cases. Thus, almost all TUD cases have evidence of being either former or current smokers on the basis of available self-report data.

SNP heritability and genetic correlations across sites

After applying similar data quality controls, we conducted within-cohort association analyses using logistic regression and relevant covariates (Methods). We estimated the proportion of variance attributable to the measured common variants (single nucleotide polymorphism (SNP) heritability, h^2_{SNP}) to be ~5–15% (based on liability scale, assuming a lifetime risk of 12.5%; Fig. 1b and Supplementary Table 6), which is consistent with previous nicotine-related GWAS[13,27]. Genetic correlations, r_g, across sites and ancestries were mostly high and positive ($r_g > 0.51$, $P < 1.56 \times 10^{-2}$, EUR sites; $r_g = 0.93$, $P = 0.45$, AA sites; cross-ancestry $r_g = 0.74–0.84$, $P < 3.90 \times 10^{-4}$; Fig. 1b and Supplementary Table 6), serving as the basis for ancestry-specific and multi-ancestry meta-analyses and suggesting that the genetic architecture of TUD is similar across ancestries.

Multi-ancestry meta-analyses of TUD

The primary multi-ancestry meta-analysis of 20,801,211 imputed SNPs ($\lambda_{GC} = 1.141$; Fig. 2) was performed on seven cohorts from four US biobanks, comprising 653,790 individuals with TUD phecode data available, with 75.71% EUR, 17.50% AA and 6.79% LA.

We identified 120 genome-wide significant (GWS) ($P < 5.00 \times 10^{-8}$) lead SNPs ($r^2 < 0.1$) located in 88 independent loci (Supplementary Table 7). All GWS loci had been reported by previous smoking GWAS (Supplementary Table 7), including aspects of smoking initiation (88 of 88), age of initiation (14 of 88), consumption (38 of 88), cessation (48 of 88) and nicotine dependence (1 of 88; Supplementary Figs. 2 and 3). While all these loci were recently discovered in a GWAS of 3.4 million individuals in the GSCAN study[13], here we reproduce some of the GSCAN findings with a considerably smaller sample size (Supplementary Fig. 3).

Our analyses provide corroborative support for nicotinic acetylcholine receptor genes as risk genes for smoking-related traits: *CHRNA5* (rs576982, $P = 3.4 \times 10^{-19}$, chr. 15; this region includes rs16969968, a well-established functional missense polymorphism (D398N) in *CHRNA5*, $P = 2.47 \times 10^{-12}$), *CHRNB2* (rs45490696, $P = 1.45 \times 10^{-9}$, chr. 1), *CHRNA2* (rs2741339, $P = 5.21 \times 10^{-17}$, chr. 8) and *CHRNA4* (rs2273500, $P = 2.84 \times 10^{-22}$, chr. 20). Second, we identified associations with variants in several genes that modulate dopaminergic transmission, such as the dopamine receptor D2 (*DRD2*: rs34632468, $P = 1.04 \times 10^{-11}$ and rs4936277, $P = 1.81 \times 10^{-9}$, chr. 11), known for its relationship with dopamine and reward[34], previously associated with nicotine dependence[35] and implicated in a recent large-scale GWAS of addiction[36]; dopamine beta-hydroxylase (*DBH*: rs2007153, $P = 9.35 \times 10^{-21}$ and rs2519155, $P = 7.25 \times 10^{-13}$, chr. 9), which encodes an enzyme necessary to convert dopamine to norepinephrine and has been consistently implicated in smoking behaviours[13,37]; lysine demethylase 4A (*KDM4A:* rs489319, $P = 1.61 \times 10^{-11}$, chr. 1), previously found to interact with dopaminergic agents and implicated in problematic opioid use[38]; phosphodiesterase 4B (*PDE4B*: rs7528604, $P = 5.68 \times 10^{-10}$, chr. 1), which has regulatory effects on dopaminergic pathways and has been implicated in GWAS

https://doi.org/10.1038/s41562-024-01851-6

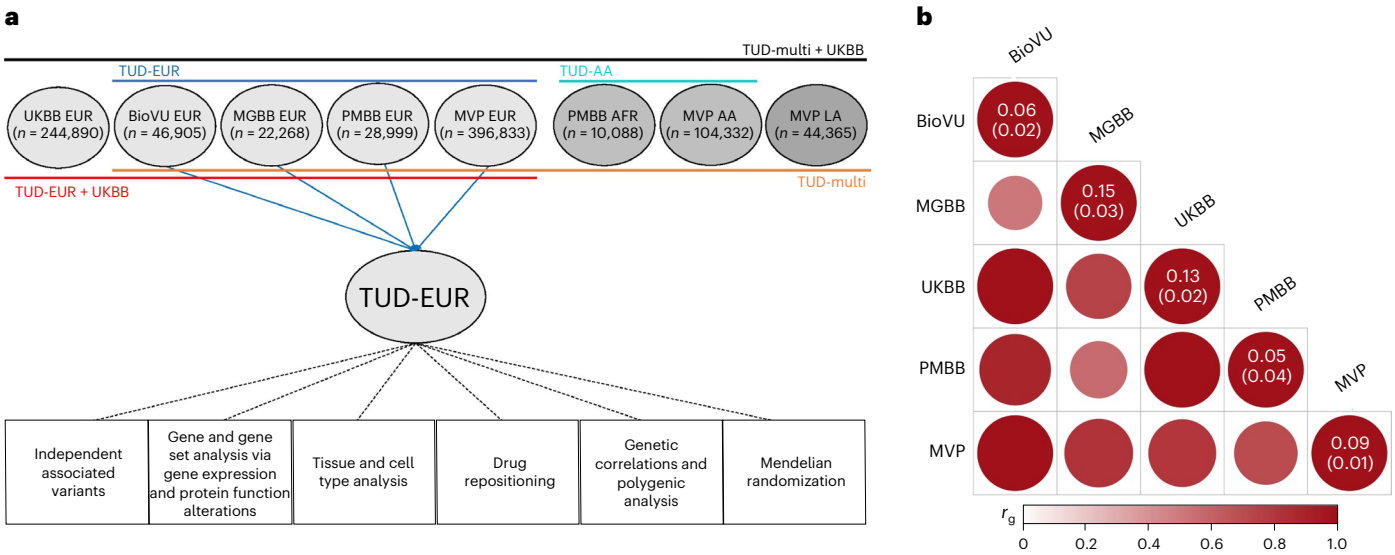

Fig. 1 | Overview of the cohorts, analysis pipeline and genetic correlations among the sites. a, We conducted independent GWAS of TUD cases and controls in individuals of European (EUR) ancestry across four PsycheMERGE sites (BioVU, Vanderbilt University Medical Center's biobank; MGBB, Mass General Brigham Biobank; PMBB, Penn Medicine BioBank; and MVP, Million Veteran Program) and performed a GWAS meta-analysis (TUD-EUR); these summary results were used for all secondary analyses. For African American (AA), we conducted GWAS meta-analysis of TUD cases and controls from the PMBB and MVP cohorts (TUD-AA). For Latin American (LA), we conducted GWAS of TUD cases and controls from the MVP cohort. Next, we performed a multi-ancestral GWAS meta-analysis (TUD-multi), which combined the results from all seven cohorts. We also obtained summary statistics from UK Biobank (UKBB), which used a less stringent case definition in individuals of EUR ancestry, and performed a GWAS meta-analysis within EUR individuals (TUD-EUR + UKBB) and across ancestries (TUD-multi + UKBB). Supplementary Table 2 summarizes the datasets used for the analyses. We subjected the TUD-EUR summary statistics to several secondary analyses to characterize the genetic architecture of TUD. **b**, LDSC genetic correlations (r_g) for TUD between EUR sites were positive and high, ranging from 0.51 to unity (two-sided *P* values are provided in Supplementary Table 6), with most CI overlapping (Supplementary Fig. 1). LDSC genetic correlation for TUD between the two AA samples was strongly positive ($r_g = 0.93$) but not significant ($P = 0.45$). LDSC SNP heritability estimates (h^2_{SNP} 5–15%) are shown in the diagonal.

of externalizing behaviours[39], smoking initiation[37,39] and general liability for addiction[36]; and neural cell adhesion molecule 1, *NCAM1* (rs9919558, $P = 4.44 \times 10^{-12}$, chr. 11), which modulates dopamine signalling[40,41] and has been associated with several smoking-related traits[35,37]. We also identified an association with a deleterious (Combined Annotation Dependent Deletion = 18.9)[42] SNP (rs986391, $P = 3.08 \times 10^{-14}$, chr. 5) in the *TENM2* gene, recently implicated in smoking initiation (SmkInit), cigarettes smoked per day (CPD) and smoking cessation (SmkCess)[13].

Furthermore, we identified variants in *GRM8* (glutamate metabotropic receptor 8; rs2157752, $P = 5.32 \times 10^{-9}$, chr. 7), important for mediating reward-related learning and memory, and in *BDNF* (rs6265, $P = 7.98 \times 10^{-10}$, chr. 11), a candidate gene in genetic studies of SUDs given its role in synaptogenesis and memory. None of the lead SNPs showed evidence of heterogeneity across cohorts, on the basis of the I^2 index (Supplementary Fig. 4). Combining these data with UKBB (which uses a less stringent TUD definition, TUD-multi + UKBB) yielded fewer lead SNPs (Supplementary Table 8).

Ancestry-specific meta-analyses of TUD

We conducted within-ancestry meta-analyses of EUR (TUD-EUR) and AA (TUD-AA) using a sample-size weighted fixed effects model and a GWAS of LA (TUD-LA).

TUD-EUR included 11,422,241 imputed SNPs in a cohort of 163,734 TUD cases and 331,271 controls, which is 8.5 times larger than the total sample size of previous nicotine-dependence GWAS[27]. Observable inflation is attributable to polygenic signal rather than population stratification or other confounding (linkage disequilibrium score regression (LDSC) intercept 1.049, s.e. = 0.012) and we did not identify evidence of heterogeneity (I^2) across the cohorts (Supplementary Fig. 6). The TUD-EUR meta-analysis yielded a significant h^2_{SNP} estimate of 11.70% (s.e. = 0.005; Supplementary Table 9) and identified 74 GWS significant lead SNPs located in 63 independent loci (Fig. 2b and Supplementary Table 10). Fourteen of these loci were ancestry specific in EUR and not GWS in the multi-ancestry GWAS. Among the 63 independent loci, 13 were fine-mapped to a credible set (posterior inclusion probability >0.50), of which 6 harboured known protein-coding genes (*CHRNB2*, *GALNT10*, *FAM168A*, *SPATS2*, *SYT17* and *ASIC2*; Supplementary Table 11).

Combining these data with those of UKBB in a secondary GWAS (TUD-EUR + UKBB) yielded very similar results (for example, similar h^2_{SNP} estimate of 9.30% and r_g estimate of 0.99, s.e. = 0.001; lead SNPs and independent loci presented in Supplementary Table 12). Considering the similarity between the primary and secondary GWAS, all downstream analyses used the EUR GWAS for the most stringent TUD definition (TUD-EUR), which excluded the UKBB sample.

The TUD-AA meta-analysis yielded a significant h^2_{SNP} estimate of 11.09% (s.e. = 0.014; Supplementary Table 9) and two independent loci (Supplementary Table 13), one on chr. 9 (rs2007153, $P = 1.17 \times 10^{-8}$) in *DBH*, which is new for the AA population, and another on chr. 20 (rs6011779, $P = 9.27 \times 10^{-9}$) in the *CHRNA4* gene, replicating a finding from a previous multi-ancestral (EUR + AA) GWAS of smoking[27]. Multi-ancestry fine-mapping analyses using PAINTOR corroborated the region in chr. 9, identifying two potential causal variants in this locus (Supplementary Table 14). The TUD-LA GWAS yielded a significant h^2_{SNP} estimate of 8.14% (s.e. = 0.02; Supplementary Table 9) but did not identify any GWS loci (Fig. 2), presumably due to the smaller sample size.

Integration of GWAS results with functional genomic data

To further our biological interpretation of the TUD-EUR GWAS results and prioritize potential candidate genes and proteins, we performed several in silico downstream analyses using MAGMA[41,42], H-MAGMA[43], S-MultiXcan/S-PrediXcan[44], transcriptome-wide association studies (TWAS)[45] and proteome-wide association studies (PWAS)[45].

First, we conducted gene-based analyses via MAGMA[41,42], which mapped SNP-level associations to 91 significant genes ($P < 2.63 \times 10^{-6}$), 20 (21.62%) of which replicated genes near or in GWS loci (for example, *CHRNA3*, *CHRNA4*, *KDM4A* and *DBH*; Supplementary Table 15).

https://doi.org/10.1038/s41562-024-01851-6

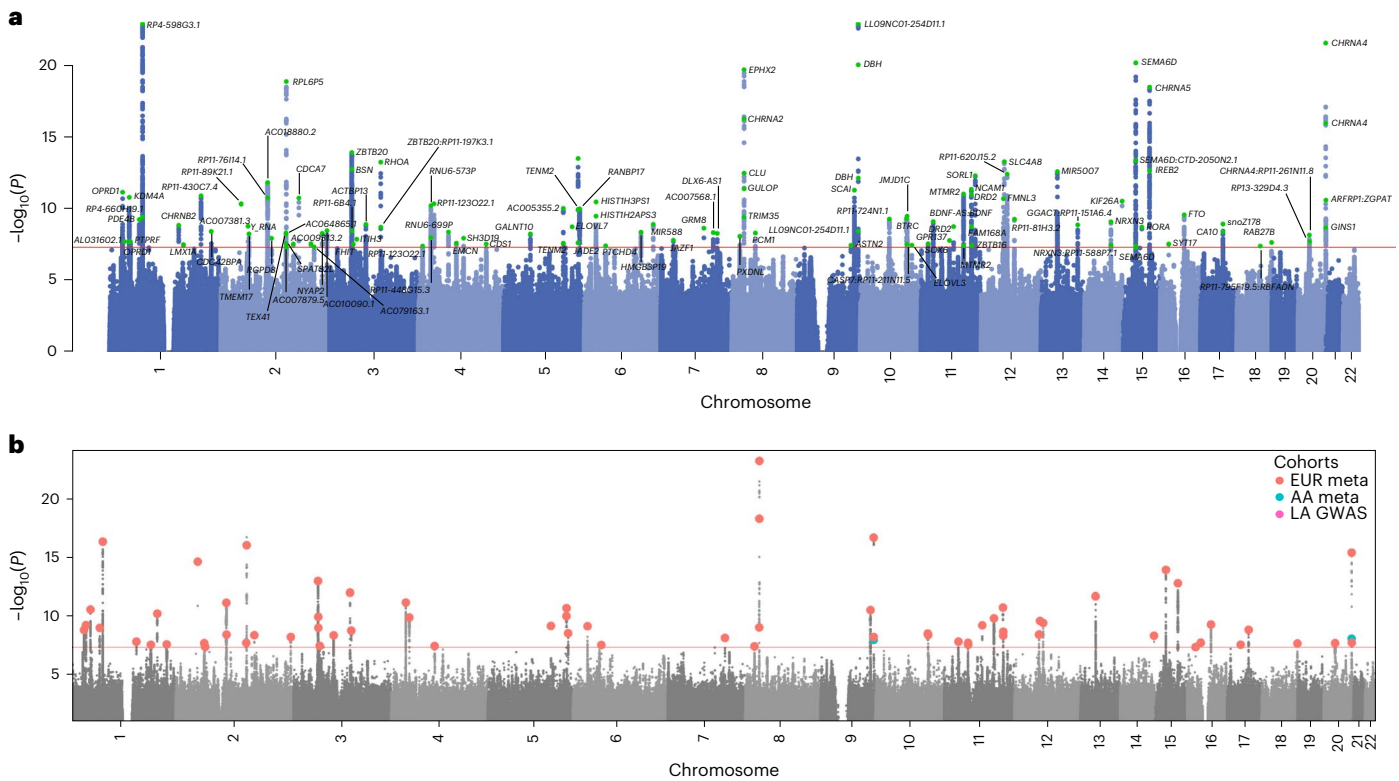

Fig. 2 | Manhattan and porcupine plots for the TUD-multi meta-analysis and ancestry-specific GWAS. a, TUD-multi identified 88 independent risk loci, all of which were recently identified by the GSCAN study. **b**, Porcupine plot of ancestry-specific meta-analyses identified 63 loci in the EUR cohort (red) and 2 loci in the AA cohort (blue). No significant associations were detected in the LA cohort. We used a sign test to examine the 74 EUR lead SNPs in the AA and LA cohorts, of which 57 and 53, respectively, were directly analysed or had proxy SNPs in these populations (Supplementary Table 10). Most SNPs had the same direction of effect in both populations (AA = 45 out of 57, LA = 41 out of 53; sign test AA $P = 1.31 \times 10^{-5}$, LA $P = 8.17 \times 10^{-5}$; Supplementary Fig. 5). All statistical tests used were two-sided.

To identify neurobiologically relevant target genes, we incorporated TUD GWAS data with chromatin interaction profiles from human brain tissue using Hi-C coupled MAGMA (H-MAGMA)[43]. These analyses identified 1,017 unique gene–tissue pairs associated with TUD ($P < 9.44 \times 10^{-7}$), a significant proportion of which showed cell-type (15.63% cortical neurons, 16.42% iPSC-derived neurons, 20.75% midbrain dopaminergic neurons and 14.25% iPSC-derived astrocytes) or developmental stage-specific (15.73% fetal and 17.21% adult) expression (Supplementary Table 16).

Using S-MultiXcan to predict the effect of common SNP variation on gene expression in several brain tissues, we detected significant associations for 46 genes (Supplementary Table 17), with effects dispersed across 13 brain regions (amygdala, anterior cingulate cortex, basal ganglia (caudate, nucleus accumbens and putamen), cortex and frontal cortex, cerebellar hemisphere, cerebellum, hippocampus, hypothalamus, spinal cord, substantia nigra). Inspection of region-specific results via S-PrediXcan identified 25 genes which were consistently upregulated (*GPX1*, *PPP6C*, *GMPPB*, *WDR6*, *QRICH1*, *NICN1*, *ARFRP1*, *METTL21B*, *RNF123*, *CCDC88B*, *HIST1H2BD*, *CCDC71* and *PSMA4*) or downregulated (*CHRNA2*, *AMT*, *P4HTM*, *NCKIPSD*, *ATP23*, *DALRD3*, *MST1*, *RHCE*, *TSFM*, *RBM6*, *TRIM35* and *PHACTR4*) in more than one brain region (Supplementary Table 18).

Next, we assessed differential transcriptomic and proteomic regulation of TUD risk loci in the dorsolateral prefrontal cortex by performing TWAS (mRNA and splicing) and PWAS, respectively. Associations across these three regulatory models identified 59 TUD risk genes and proteins (32, mRNA expression; 13, splicing expression; 14, proteome expression; Supplementary Tables 19 and 20), 51 of which were unique. Colocalization analysis identified four genes and proteins (NT5C2, GPX1, ABHD12 and RHCE) associated with TUD via their regulation of brain expression levels and protein abundance (PP4 > 0.80, Supplementary Table 21 and Supplementary Fig. 7).

Overall, after controlling for several comparisons, these analyses identified 461 unique genes with statistical evidence of association with TUD (Fig. 3a and Supplementary Table 22). Of these, 159 genes converged across at least two methods and 2 genes (*GPX1* and *GMPPB*) converged across all six methods and replicated previous GSCAN findings. A total 110 (23.86%) of the 461 genes identified via these analyses were identified by the GWS loci and 2 were new TUD genes not identified in previous FTND or GSCAN analyses (*PTCHD4* and *THUMPD3*), which prompt new hypotheses to be tested experimentally.

Tissue and cell-type analyses

To identify relevant tissues implicated in TUD, we performed various SNP (LDSC partitioned heritability) and gene-wide (MAGMA) analyses. We performed partitioned heritability in LDSC to evaluate the enrichment of the genome-wide findings in over 50 functional genomic annotations (and across tissues, as described below). In the baseline LDSC model, conserved and regulatory functional annotations were significantly enriched (Supplementary Fig. 8 and Supplementary Table 23 for full list).

Tissue enrichment analyses in MAGMA use gene expression data from GTEx (v.8). In addition to non-brain tissues (that is, cardiovascular, haematopoietic, adrenal pancreas and other, $P < 3.37 \times 10^{-5}$; Supplementary Table 24), we detected significant enrichment mostly in the brain ($P = 1.53 \times 10^{-15}$), spanning several brain regions, including the hippocampus, the limbic system and frontal cortex (Supplementary Tables 25 and 26 and Fig. 3b,c), most of which were also implicated

https://doi.org/10.1038/s41562-024-01851-6

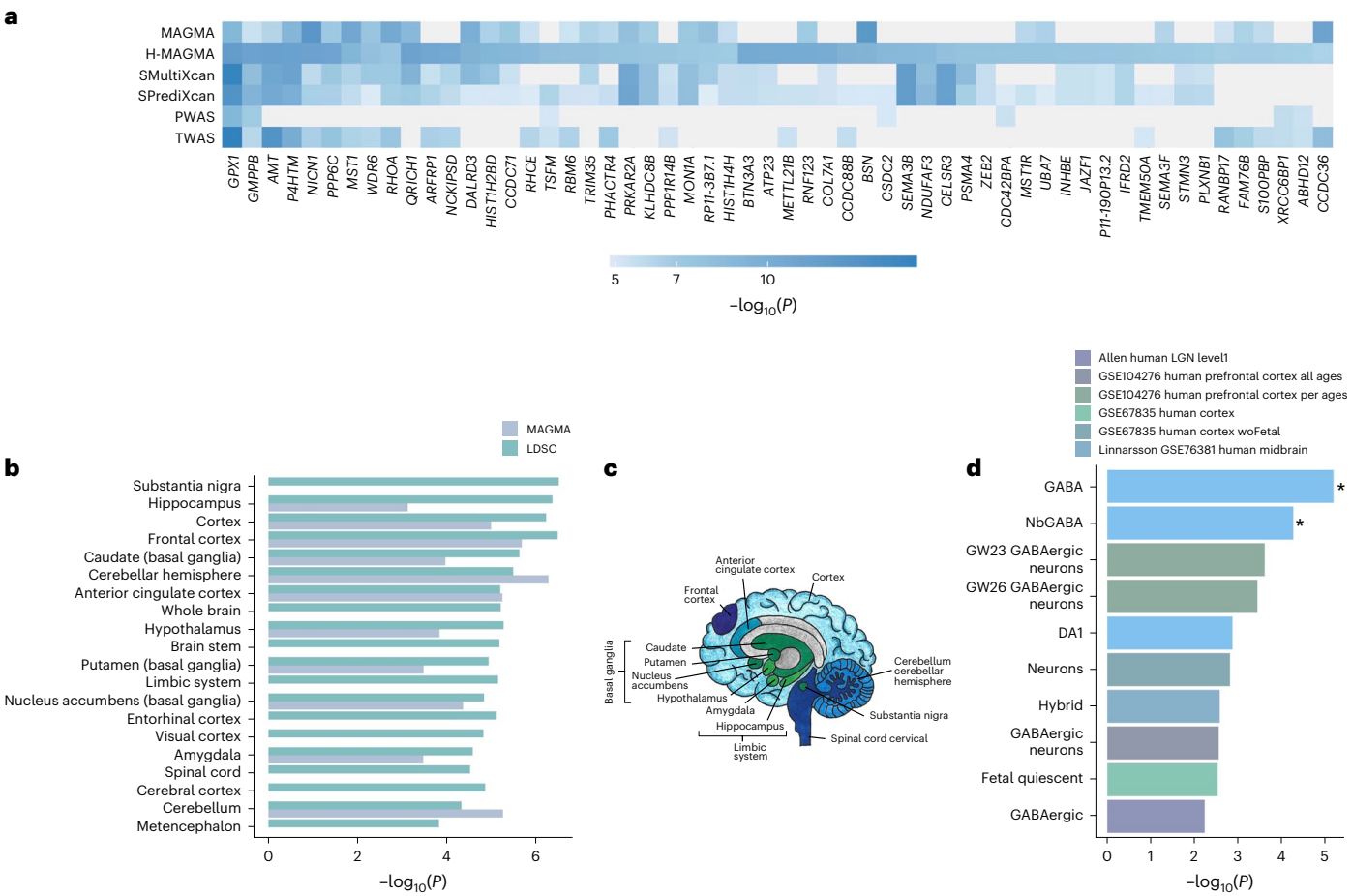

Fig. 3 | Integration with functional genomic data implicated 461 unique TUD candidate risk genes. a, Of 461 associated genes, 56 converged with at least three methods and were dispersed throughout the chromosomes. **b**, LDSC (SNP-based) and MAGMA tissue-specific gene expression of TUD risk genes reveals substantial brain enrichment (Supplementary Tables 25 and 26). Only tissues that survived multiple testing are plotted (MAGMA, two-sided $P < 9.26 \times 10^{-4}$; LDSC, $P < 2.44 \times 10^{-4}$). **c**, The genetic findings across several levels of analysis (LDSC, MAGMA, S-MultiXcan and BrainXcan) implicated brain regions exhibiting anatomical differences in cases. **d**, Cell-type-specific expression of TUD risk genes. Results from MAGMA property analyses and gene expression using human scRNA-seq datasets (Supplementary Table 28 for full list). After multiple testing corrections for all datasets, only genes expressed in GABAergic neurons were associated with TUD (Supplementary Table 28). The asterisks denote independent cell-type associations across datasets.

in S-MultiXcan (Supplementary Table 17). Correlating the effects of SNP variation with brain imaging traits via BrainXcan identified similar results, including significant ($P < 1.92 \times 10^{-4}$) associations with decreased grey matter volume in the right ventral striatum (Supplementary Table 27).

Next, we used FUMA to examine cell-type-specific gene expression associated with TUD, leveraging single-cell RNA sequencing (scRNA-seq) datasets. After multiple correction testing across datasets, we identified a significant association between TUD risk and cell-type-specific gene expression in GABAergic neurons for individual human scRNA-seq datasets (Linnarsson, midbrain, GABA: $P < 5.03 \times 10^{-3}$; nbGABA: $P < 4.29 \times 10^{-2}$; Fig. 3d and Supplementary Table 28). These results did not survive conditional analyses within and across datasets.

Gene set and pathway analyses

We used MAGMA[41,42] to conduct a gene-wise TUD analysis and to test for enrichment of pathways curated from several sources. After correcting for several comparisons, 13 related pathways and biological processes were significantly enriched for genes associated with TUD ($P < 2.65 \times 10^{-6}$; Supplementary Table 29). Associations implicated fundamental processes related to nicotine response (for example, high calcium and sodium permeable nicotinic acetylcholine receptors, $P = 6.03 \times 10^{-15}$; behavioural response to nicotine, $P = 5.81 \times 10^{-13}$), regulation of postsynaptic nicotinic acetylcholine receptors ($P = 1.32 \times 10^{-10}$) and nicotine effect on dopaminergic neurons ($P = 1.87 \times 10^{-6}$), among others.

Drug repurposing

Linking transcriptome-wide patterns to perturbagens that pass the blood–brain barrier from the Library of Integrated Network-Based Cellular Signatures (LINCS)[36] database identified 235 medications approved by the US Food and Drug Administration (Supplementary Table 30). Of the 235 identified medications, 20 targeted at least one mapped/independent gene from our GWAS (Fig. 4). The medications that significantly reversed (Bonferroni $P < 6.03 \times 10^{-5}$) the transcriptional profile associated with TUD included varenicline (a well-known therapeutic for SmkCess), sodium channel blockers (for example, amiloride) and compounds which are used to treat conditions that commonly co-occur with TUD, such as antipsychotics (for example, clozapine), dopaminergic agents (for example, ropinirole), opioids (for example, nalbuphine) and antidepressants (for example, amoxapine), among others (Supplementary Table 30).

An additional drug repositioning analysis using DRUGSETS identified three significant (Bonferroni $P < 6.80 \times 10^{-5}$) medications: varenicline, cytisine and galantamine (Supplementary Table 30).

https://doi.org/10.1038/s41562-024-01851-6

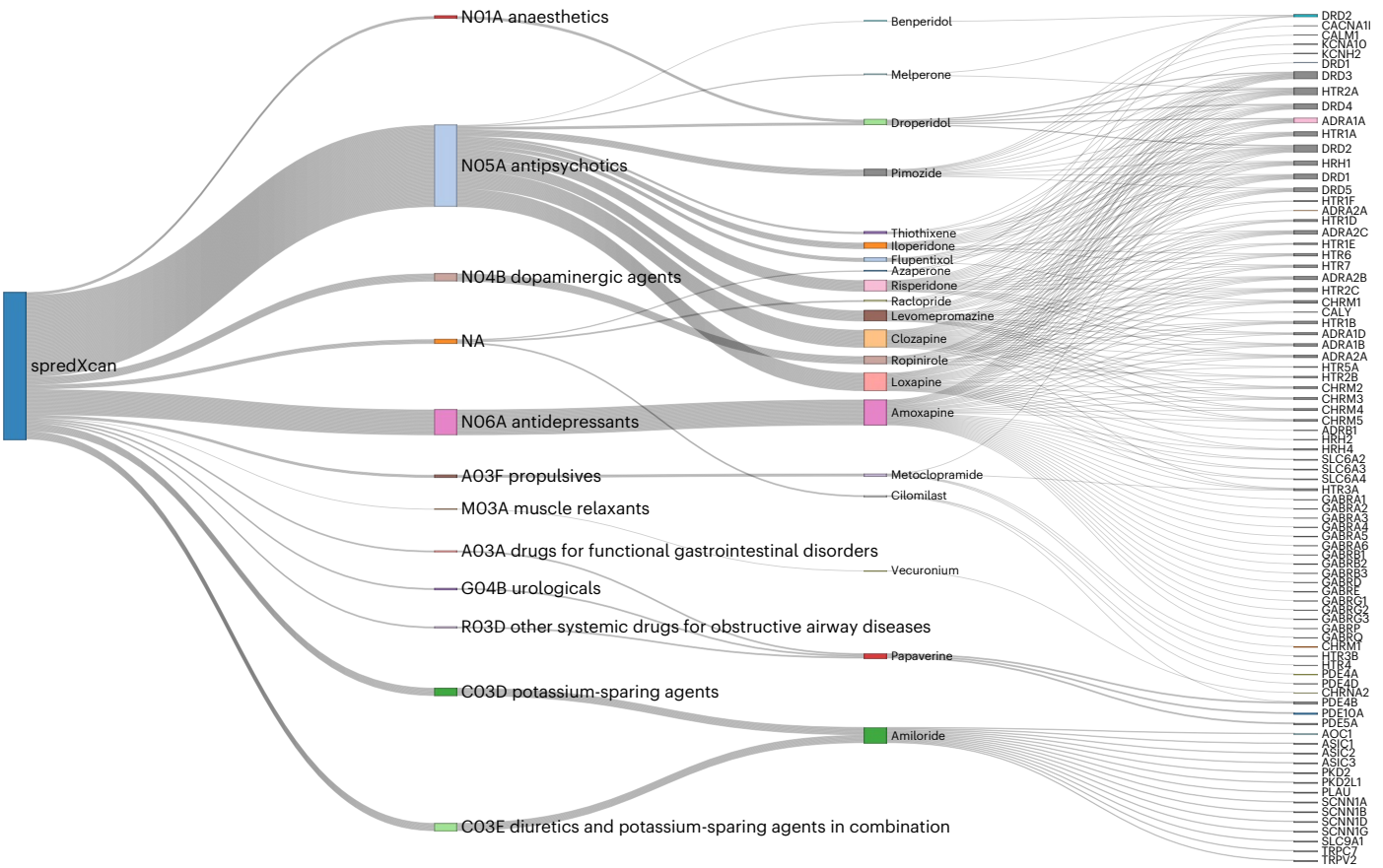

Fig. 4 | Sankey diagram showing drug-repurposing results from S-PrediXcan brain tissues. Twenty medications/perturbagens grouped by ATC category membership from the LINCS[36] database. ATC categories connected to perturbagen edges represent corresponding ATC category membership. Perturbagens connected to gene target edges are associated with the reversal of the TUD transcriptomic profile from S-PrediXcan brain tissue results. Only medications that targeted at least one mapped/independent gene from our GWAS are plotted.

Genetic correlation with other traits

We estimated pairwise r_g with TUD for 113 published phenotypes using LDSC[46]. TUD showed false discovery rate (FDR)-significant correlations r_g with 76 traits (Fig. 5b and Supplementary Table 31). As expected, the strongest positive correlations were with smoking-related traits (for example, age of smoking initiation (AgeSmkInit) $r_g = -0.59$, s.e. = 0.03; SmkInit $r_g = 0.81$, s.e. = 0.02; CPD $r_g = 0.44$, s.e. = 0.03; SmkCess $r_g = 0.66$, s.e. = 0.02; FTND $r_g = 0.63$, s.e. = 0.06; Fig. 5a) and other substance use traits (for example, cannabis use disorder (CUD) $r_g = 0.64$, s.e. = 0.04; drinks per week $r_g = 0.36$, s.e. = 0.02; OUD $r_g = 0.47$, s.e. = 0.04). TUD clustered with addiction traits rather than consumption phenotypes (Supplementary Fig. 9).

TUD was also genetically associated with 59 other psychiatric and medical conditions (Fig. 5b and Supplementary Table 31). There were significant positive r_g with psychiatric traits (for example, externalizing $r_g = 0.71$, s.e. = 0.02; ADHD $r_g = 0.50$, s.e. = 0.03; post-traumatic stress disorder $r_g = 0.31$, s.e. = 0.08) and risky behavioural traits, including younger age of first sex ($r_g = -0.50$, s.e. = 0.03). We also found positive r_g with health outcomes (for example, coronary artery disease $r_g = 0.26$, s.e. = 0.03; waist-to-hip ratio $r_g = 0.26$, s.e. = 0.02; multisite chronic pain $r_g = 0.36$, s.e. = 0.03) and several social determinants of health, such as the Townsend deprivation index ($r_g = 0.61$, s.e. = 0.07). There were negative r_g with socioeconomic variables, including educational attainment ($r_g = -0.53$, s.e. = 0.02) and household income ($r_g = -0.49$, s.e. = 0.03) and with intelligence ($r_g = -0.28$, s.e. = 0.02). Conditioning on alcohol, cannabis or OUD did not substantially modify the magnitude or direction of these associations (Supplementary Table 32). Virtually all r_g estimates for other phenotypes were greater with TUD than CPD (Fig. 5c) and FTND (Fig. 5d) but not SmkInit (Fig. 5e).

Among AA samples, there were significant r_g with smoking trajectories and other substance use traits (OUD $r_g = 0.44$, s.e. = 0.11; maximum habitual alcohol consumption $r_g = 0.77$, s.e. = 0.19).

Phenome-wide association analyses

To further explore pleiotropic effects, we performed a series of phenome-wide association studies (PheWAS) of TUD polygenic scores (PGS) in other EHR and clinical cohorts of adults and a young population-based cohort. We performed these analyses within ancestries.

EHR cohorts. We conducted PheWAS with EHR data to test the association between polygenic risk for TUD and liability for thousands of other medical conditions, including TUD, in another independent site—the Mayo Clinic Biobank. As expected, TUD PGS was strongly associated with TUD ($P = 1.90 \times 10^{-145}$; Supplementary Table 34 and Fig. 6a), explaining 7.3% of the (Nagelkerke's R^2) variance. Additional significant ($P < 3.24 \times 10^{-5}$) associations included four traits in the SUDs domain (for example, alcohol-related disorders, OR = 1.33, $P = 6.30 \times 10^{-26}$), ten psychiatric conditions (for example, depression, OR = 1.09, $P = 4.31 \times 10^{-11}$) and medical conditions strongly associated with TUD (for example, chronic airway obstruction, OR = 1.25, $P = 1.60 \times 10^{-32}$). Most of these associations remained significant after accounting for TUD diagnosis (Supplementary Table 34). We also noted associations across several other medical categories, including endocrine/metabolic (for example, morbid obesity, OR = 1.12, $P = 3.53 \times 10^{-13}$; type 2

https://doi.org/10.1038/s41562-024-01851-6

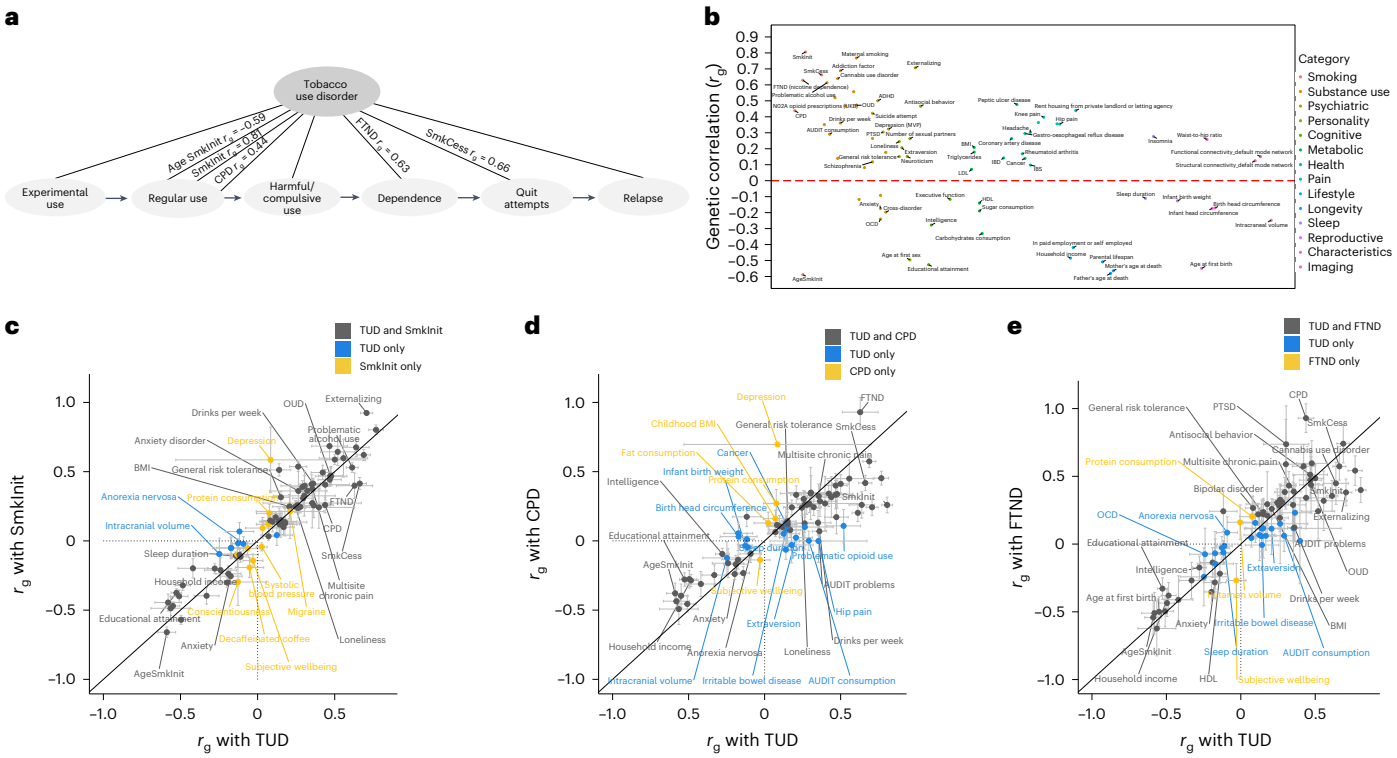

Fig. 5 | FDR-significant genetic correlations between TUD-EUR and 113 complex traits, including smoking and related phenotypes. a, TUD consists of several components, progressing from experimental use to regular use, compulsive use, cessation and relapse. Therefore, high genetic correlations (r_g) are to be expected between the age of smoking initiation (AgeSmkInit), smoking initiation (SmkInit), cigarettes per day (CPD), smoking cessation (SmkCess)[13], nicotine dependence measured using the Fagerström test for nicotine dependence (FTND)[27] and TUD (see Supplementary Table 31 for full results). **b**, Genetic correlations with an extended list of traits from publicly available GWAS. Traits with positive r_g values are plotted above the line; traits with negative r_g values below the line. All r_g are significant using a 5% FDR correction for multiple testing. **c**–**e**, Systematic comparison of significant genetic correlation estimates between TUD and SmkInit (**c**), CPD (**d**) and FTND (**e**) reveals overlapping (black dots) and trait-specific (blue and yellow dots) relations between TUD and these other smoking phenotypes. The r_g estimates were generally higher for TUD than CPD—even with a smaller sample size (TUD, n = 495,005; CPD, n = 784,353)—and FTND. On the contrary, r_g were generally smaller for TUD than SmkInit, possibly because of the larger sample for SmkInit (n = 3,383,199) than TUD. Overall, these results indicate that these smoking behaviours, including SmkInit, CPD, FTND and TUD, represent both unique and interrelated polygenic influences, which are complementary to those associated with other complex behaviours and disorders at the genetic level. ADHD, attention-deficit hyperactivity disorder; PTSD, posttraumatic stress disorder; BMI, body mass index; HDL, high density lipoprotein; LDL, low density lipoprotein; OCD, obsessive compulsive disorder.

diabetes, OR = 1.09, $P = 1.48 \times 10^{-9}$), digestive (for example, diseases of oesophagus, OR = 1.07, $P = 1.47 \times 10^{-10}$), circulatory (for example, ischaemic heart disease, OR = 1.09, $P = 1.56 \times 10^{-11}$) and neurologic (for example, pain, OR = 1.07, $P = 4.33 \times 10^{-8}$), among others (Supplementary Table 34). Compared to FTND PGS, TUD PGS were more strongly associated across virtually all domains, including TUD (Fig. 6a). After conditioning on PGS for other smoking variables (CPD, SmkInit and FTND), TUD PGS was still significantly associated with TUD and 14 other mental and medical traits (Supplementary Table 34). We repeated the TUD PGS analyses in a BioVU cohort of AA individuals using the TUD-AA meta-analysis results. As expected, TUD was the strongest (OR = 1.20, $P = 2.81 \times 10^{-6}$) association (Supplementary Table 35).

Yale-Penn sample. We next extended the analyses to a deeply characterized sample recruited for genetic studies of SUDs—the Yale-Penn sample[47]. We examined the association between PGS for TUD and hundreds of other traits derived from a comprehensive psychiatric interview, the semi-structured assessment for drug dependence and alcoholism (SSADDA). TUD-EUR and TUD-AA PGS were strongly associated with nicotine dependence as defined via a Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnosis in both the EUR (OR = 1.83, $P = 3.51 \times 10^{-49}$; Fig. 6b and Supplementary Table 36) and AA cohorts (OR = 1.13, $P = 7.13 \times 10^{-4}$), respectively, although the latter association did not survive multiple testing correction (Supplementary Table 37). In the EUR cohort, we also noted significant associations between TUD-EUR PGS and 224 other phenotypes, including 162 in the substance-related domain (44 opioid-related, 31 cocaine-related, 25 alcohol-related, 23 tobacco-related, 14 sedative-related, 13 cannabis-related, 10 other and 2 stimulant-related) and 62 in other domains (13 medical, 34 psychiatric (9 PTSD, 11 depression, 7 antisocial personality, 3 suicide, 2 ADHD and 2 conduct disorder), 9 environmental and 6 demographic phenotypes). Again, compared to FTND PGS, TUD-EUR PGS was more strongly associated across virtually all domains, including nicotine dependence (Nagelkerke's R^2 = 0.101 versus 0.062; Supplementary Table 36). After conditioning on PGS for other smoking variables (CPD, SmkInit and FTND), TUD PGS was still significantly associated with 11 smoking-related traits and 50 other mental and medical conditions (Supplementary Table 36), again emphasizing the value of collecting information on later stages of vulnerability or more severe phenotypes, such as TUD.

Adolescent brain cognitive development cohort. Last, we extended our polygenic analyses to a drug naive developmental sample (9–11 years of age at recruitment; analytic n = 62–5,556)—the adolescent brain cognitive development (ABCD) cohort. We concentrated on 12 traits that showed significant genetic correlations in the adult samples (Supplementary Table 38 and Fig. 6c). Although tobacco exposure was uncommon in this paediatric population (2.30% prevalence),

https://doi.org/10.1038/s41562-024-01851-6

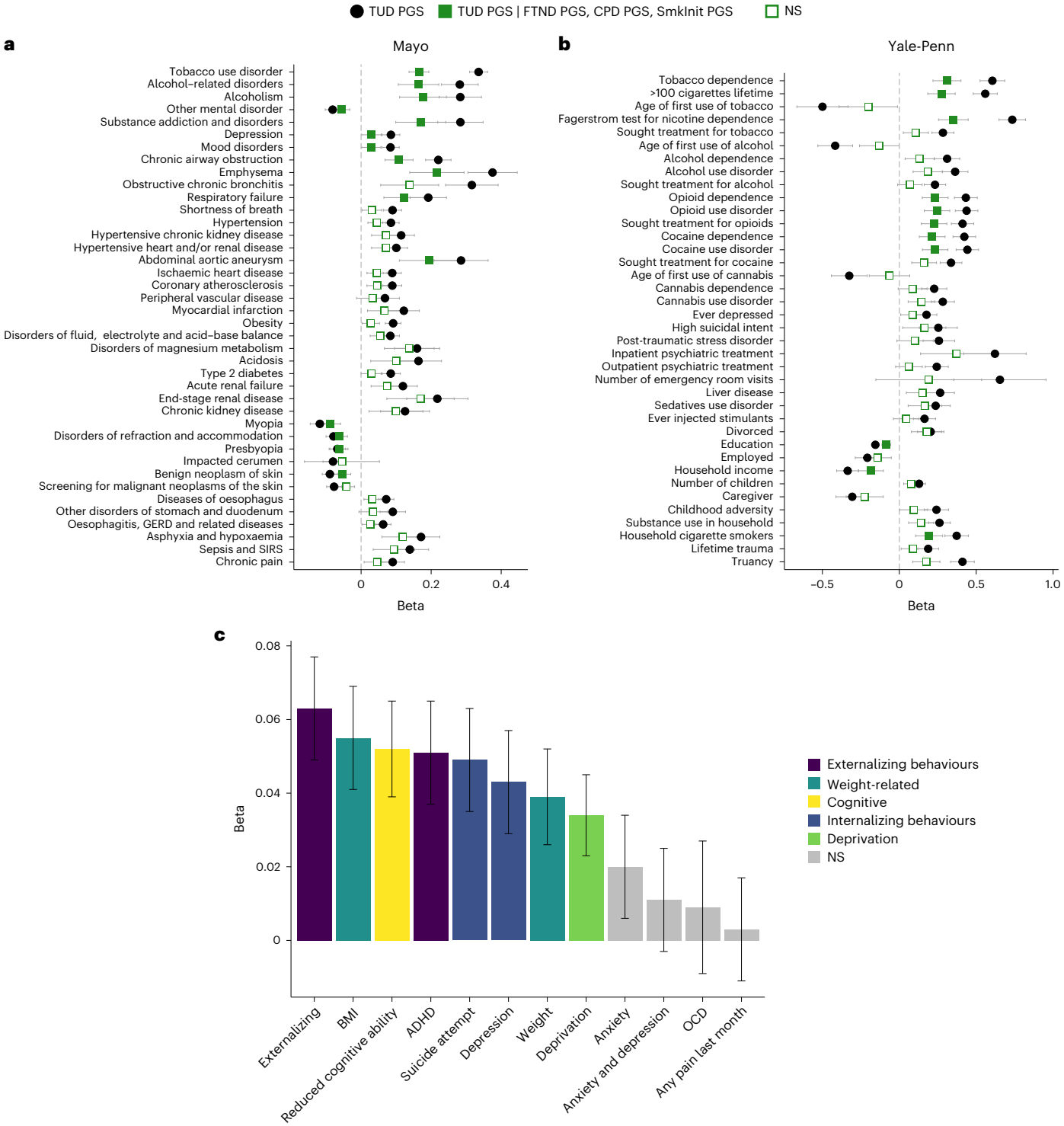

Fig. 6 | Phenome-wide association studies of TUD PGS. a–**c**, TUD PGS PheWAS in the Mayo Clinic (**a**), Yale-Penn (**b**) and ABCD European (**c**) cohorts. Only selected Bonferroni-significant traits are shown. In **a** and **b**, association of TUD PGS (black) is conditioned on PGS for FTND, CPD and SmkInit (green). Values represent β and s.e. The exact values for each association and extended lists of traits can be found in Supplementary Tables 34, 36 and 38. The number of observations used in **c** is shown in Supplementary Table 38. NS, not significant.

externalizing behaviours, which emerge in childhood and are strong correlates of substance use, were available. After multiple testing correction, TUD PGS was significantly ($P < 4.00 \times 10^{-3}$) associated with externalizing behaviours (Child Behaviour Check List (CBCL) externalizing scores, $\beta = 0.07$, $P = 1.21 \times 10^{-6}$; CBCL ADHD scores, $\beta = 0.06$, $P = 4.97 \times 10^{-5}$), as well as internalizing (suicide attempt $\beta = 0.05$, $P = 1.52 \times 10^{-3}$; CBCL depression scores $\beta = 0.05$, $P = 1.11 \times 10^{-3}$); cognitive ability ($\beta = 0.06$, $P = 8.35 \times 10^{-6}$); neighbourhood deprivation ($\beta = 0.04$, $P = 1.05 \times 10^{-3}$); and weight-related phenotypes (BMI $\beta = 0.06$, $P = 1.61 \times 10^{-5}$; weight $\beta = 0.04$, $P = 2.77 \times 10^{-3}$). Notably, these children were not chronically exposed to tobacco; therefore, we would speculate that these associations are not a consequence of smoking but rather may underlie overlapping genetic architectures among the traits studied that predate use of tobacco.

Bidirectional Mendelian randomization analyses

We used MR analyses to test directional causal relationships between significantly genetically correlated traits ($n = 31$) and TUD among EURs only resulting from the small sample size and limited statistical power in other populations (Supplementary Table 39). There was a positive causal effect of TUD on cross-disorder (inverse-variance weighted (IVW) $\beta = 0.93$, s.e. = 0.02, $P = 5.06 \times 10^{-10}$, 95% confidence interval (CI) = 0.64–1.22). Seven traits showed significant causal effects on TUD. Specifically, we observed a negative causal effect of education attainment (IVW $\beta = -0.25$, s.e. = 0.02, $P = 2.02 \times 10^{-39}$, 95% CI = −0.29–[−0.22]) and a positive causal effect of drinks per week (IVW $\beta = 0.22$, s.e. = 0.02, $P = 8.53 \times 10^{-28}$, 95% CI = 0.18–0.26), depression (IVW $\beta = 0.09$, s.e. = 0.01, $P = 1.35 \times 10^{-12}$, 95% CI = 0.06–0.11), BMI (IVW $\beta = 0.10$, s.e. = 0.01, $P = 1.85 \times 10^{-38}$, 95% CI = 0.08–0.11), externalizing (IVW $\beta = 0.48$, s.e. = 0.02, $P = 3.38 \times 10^{-131}$, 95% CI = 0.44–0.52), opioid prescriptions (IVW $\beta = 0.04$, s.e. = 0.01, $P = 2.33 \times 10^{-5}$, 95% CI = 0.02–0.06) and OUD (IVW $\beta = 0.06$, s.e. = 0.01, $P = 1.70 \times 10^{-7}$, 95% CI = 0.04–0.08) on TUD.

Discussion

Uncovering the genetic underpinnings of individual differences in TUD liability can advance diagnosis, prevention and treatment efforts for a disorder of enormous public health significance. GWAS have uncovered several associations with tobacco use but findings for tobacco dependence or disorder have been limited because of the difficulty of characterizing large numbers of individuals using a gold-standard research or clinical diagnosis. Here, we present a multi-ancestry GWAS of TUD using data from EHR, as a complementary strategy for ascertainment. EHR-biobanks are the result of years of work recruiting, consenting and genotyping individuals. As a result, researchers can now conduct studies such as the one reported here, gathering data for 898,680 individuals in <4 months, to identify biology for disorders. The number of GWAS signals, enrichment in relevant pathways and tissues and genetic overlap with nicotine-related traits provide proof of principle that EHR can serve as a complementary tool to study TUD genetics.

Our findings demonstrate that TUD, as defined via EHR, was genetically correlated with traits derived from traditionally ascertained cohorts, including nicotine dependence via FTND and SmkCess, providing clear evidence that the signal captured by TUD phecodes is valid. Of note, the genetic correlation between TUD and CPD was relatively modest ($r_g = 0.44$), suggesting that the genetic architectures of consumption and misuse are only partially overlapping, consistent with previous GWAS of alcohol and cannabis use and misuse (for example, refs. 23,26,48). This contrasts with earlier observations for FTND and CPD, for which the genetic correlation was almost at unity ($r_g = 0.95$)[27]. This shows that TUD captures features beyond the frequency of smoking or severity of nicotine dependence. Although FTND and TUD were more strongly genetically correlated ($r_g = 0.63$), in general, we observed that TUD PGS was more predictive of DSM-defined tobacco dependence and a plethora of comorbid traits in the Yale-Penn sample, than FTND PGS. The only exception was for smoke after waking, which was more strongly associated with FTND PGS, probably because time to first cigarette is one of the FTND items. TUD was highly correlated ($r_g = 0.81$) with regular cigarette use (that is, smoking at least 100 cigarettes in a lifetime, previously referred to as 'smoking initiation')[13], which is expected as nicotine is a highly addictive substance, with 85% of smokers meeting criteria for TUD[1,2]. However, our polygenic findings demonstrate that TUD explains additional variance above and beyond that accounted for by other smoking traits (SmkInit, CPD and FTND). This emphasizes the need to measure the full spectrum of addiction liability[49], from regular use to more severe phenotypes, such as TUD, to account for the distinct biological factors relevant at each stage.

Common SNPs were able to account for a fraction (12%) of the overall heritability of TUD (40–60%) as determined by previous family and twin studies[9,11]. The multi-ancestral meta-analysis identified 88 independent loci, 18 times the number previously reported for nicotine dependence[27]. These include corroborative support for the involvement of nicotinic acetylcholine receptor genes (*CHRNA5-A3-B4*, *CHRNB2*, *CHRNA2* and *CHRNA4*), which have been consistently associated with smoking behaviours[20], particularly in studies of self-reported CPD[13]. Other variants identified were in genes that modulate dopaminergic and glutamatergic neurotransmission, compromising reward-based learning and facilitating drug-seeking behaviour, and in *BDNF*, which is involved in memory consolidation processes[50] and a well-studied candidate gene in addiction[51]. These and other candidates supported by TUD (for example, *PDE4B*) were genetically correlated with other addiction phenotypes[36], emphasizing the shared neurobiological mechanisms of addiction.

Downstream analyses prioritized genes and drug candidates that could be used for follow-up mechanistic studies in model organisms. Specifically, we identified 'core' genes that could be 'pleiotropic hotspots' associated with several traits. One was glutathione peroxidase-1 (*GPX1*), which is involved in oxidative stress. Intriguingly, it has been reported that glutathione peroxidase-1 protects against lung inflammation induced by smoking in mice and agents that mimic this action (for example, ebselen), which restore *GPX1* activity in situations of extreme oxidative stress, can protect from lung inflammation induced by smoking[52]. Another was *GMPPB*, which has been associated with accelerated lung aging and e-cigarette smoking[53]. *NT5C2* is involved in maintaining cellular nucleotide balance and was associated with schizophrenia[54] and smoking behaviours in an exome-wide association study[55]. These genes showed a consistent association based on colocalization analyses (here and previously[56]), suggesting that they could confer TUD risk by modulating regulated gene expression and protein abundance in the brain.

The enrichment of TUD in brain tissues further supports TUD as a brain disorder, long supported by neuroscience and more recently by genetics[57]. We provide suggestive evidence for the involvement of the cerebellum in TUD, along with other regions that have long been studied in relation to addiction such as the fronto-striatal loop, hippocampus and amygdala[58].

Genetic correlations revealed substantial levels of pleiotropy with traits that often co-occur with TUD, including other substance use and psychiatric disorders. These associations were particularly evident in the Yale-Penn sample[47], which has comprehensive phenotypic data for SUDs. In adult patients from the Mayo Clinic, we replicated the associations with substance and other psychiatric disorders, extending them to medical disorders, such as HIV, heart disease and pain, some of which (for example, respiratory conditions) probably reflect chronic smoking. The positive associations between genetic liability for TUD and other outcomes, such as BMI and other internalizing/externalizing problems in tobacco-naive children (ABCD), may also reflect true biological relationships. Although we are far from untangling this complex web of genetic and non-genetic correlations, the extensive phenotypic spectrum associated with TUD is undeniable.

Currently, developing new therapeutics for TUD is viewed as risky because of a lack of high-quality targets, historically low success rates and unintended side effects. Although genes identified in our GWAS, including *CHRNA5*, *CHRNA4* and *CHRNB2*, might moderate the effect of varenicline, a SmkCess treatment that operates as a partial agonist at the nicotine acetylcholine a2b4 receptor[59], varenicline (along with other medications such as nicotine replacement therapies) has limited efficacy or adverse effects[60,61]. In a proof-of-principle study, ref. 62 identified several repurposing candidates for treating psychiatric disorders by connecting imputed transcriptomic profiles from GWAS data to drug-induced gene expression profiles. Using this approach, we identified hundreds of potential drug candidates predicted to significantly reverse the TUD transcriptomic profile. These included norepinephrine re-uptake inhibitors (for example, amoxapine)

and antipsychotics (for example, clozapine), pointing to convergent molecular mechanisms between TUD and other psychiatric disorders which are the usual target of these agents, replicating previous observations[63,64]. The potential therapeutic utility of anti-inflammatory and blood glucose-lowering medications was also suggested by our analyses, in addition to an anti-Parkinson medication known to interact with dopaminergic activity (biperiden) and one which acts both as an antagonist of acetylcholinesterase and an agonist of nicotinic receptors (galantamine), as shown in recent independent studies[64,65]. Although, so far, no repurposed drugs have been developed for treating SUDs on the basis of GWAS data, this is an important potential path forward, particularly for SUDs, where few effective pharmacotherapies are available.

Future research may address some of the limitations of our study. Previous work has demonstrated that ICD codes have a low sensitivity for current tobacco use but may have a reasonable specificity for this common behaviour[66]. Our results appeared to be robust to moderate levels of misclassification, particularly in controls, as detected by the pairing with self-reported questionnaire data. Our results also appeared to be robust to moderate levels of cross-cohort heterogeneity, including potential differences in diagnostic practices and different levels of misdiagnosis of control populations across sites. Although studies that systematically evaluate the effect of removing potentially misclassified individuals are needed, we chose not to remove them in this study because not all individuals had concomitant survey data available. This questionnaire data, along with other forms of EHR data (for example, clinical notes), may help capture additional phenotypes, including the response to treatment or the ability to successfully quit smoking without formal treatment. We have highlighted potential differences of traits ascertained by ICD codes as a limitation of our study. The r_g results revealed high levels of association between TUD and hundreds of other traits. However, the extent to which TUD shares biological underpinnings with other traits and diseases may also be influenced by potential misdiagnosis, ascertainment and cross-trait assortative mating, among many other factors[67]. Longitudinal data from EHR, with data collection spanning the period previous to and following the onset of substance use and SUD, are particularly valuable for studying the timing of onset, within-person change and application of time-varying effects, which will help to differentiate causation from correlational findings. The advent of single-cell transcriptomics, larger quantitative trait loci (QTL) databases in more specific cell types and the inclusion of more ancestrally diverse samples, as well as samples with varying sociocultural context from different geographic regions beyond the United States and United Kingdom, will improve the interpretability of associated loci. Although we have included diverse cohorts, our study lacked many major ancestral groups such as East Asians and South Asians. Furthermore, other forms of genetic variation, such as rare single variants[68,69] or structural polymorphisms[70], are likely to account for much of the 'missing heritability' in genetic risk for TUD. Last, tobacco use can be greatly affected by environmental factors[12], such as cultural context, public health policies and characteristics related to socioeconomic status[71]. Together with the existing body of literature[72–75], the strong genetic correlations between TUD and environmental influences, such as Townsend deprivation index, educational attainment and prenatal smoking, underscore the importance of considering environmental moderators in understanding the complex aetiology of TUD. There is a great need in the field, therefore, to systematically assess sociocultural factors in healthcare settings[76].

In sum, this work demonstrates that EHR is a viable and cost-efficient complementary approach to rigorous clinical ascertainment for genetic studies of TUD, similar to other SUD traits. At various levels of analysis, this study identifies and prioritizes previously unidentified genes of potential interest. TUD shares biological processes common to many SUDs and is highly correlated with many psychiatric and medical disorders. We anticipate that these results can be combined with previous smoking GWAS in larger multivariate analyses to elucidate the full spectrum of smoking behaviours and accelerate gene discovery for TUD.

Methods

Ethics

This study complies with all relevant ethical regulations. The project was approved by the Institutional Review Board (IRB) from Vanderbilt University Medical Center (VUMC) (nos. 160302, 172020 and 190418), MGBB (no. 2018P002642), PMBB (no. 813913), the Central VA and site-specific IRBs (MVP) and the Mayo Clinic.

Smoking phenotypes and cohorts

We defined cases as patients who received at least two TUD ICD9 or ICD10 codes (corresponding to the phecode definition) in their medical records and controls as patients who had no TUD diagnosis codes (Supplementary Table 2). In UKBB only, cases were defined as having one ICD10 code for TUD and controls had none[41]. Additionally, we required controls to be 18 years of age or older at time of analysis (April 2022). Patients younger than 18 years were excluded because they may not yet have reached the age of TUD diagnosis. We examined the sensitivity of our TUD phenotyping using the patients' self-reported tobacco use via survey data when available (Supplementary Table 3, list of smoking traits).

Our data sources included registries from five health systems linked to biobanks: BioVU, MGBB, PMBB, MVP and UKBB. There were 46,905 (EUR) patients from VUMC, 22,268 (EUR) patients from MGBB, 39,087 patients from PMBB (28,999 EUR and 10,088 AA), 545,530 patients from MVP (396,833 EUR, 104,332 AA and 44,365 LA) and 244,890 participants from UKBB. Details of each registry, including demographics and data sources, are listed in Supplementary Table 2.

Genotyping, imputation and GWAS

For all cohorts, the initial GWAS analyses were conducted within genetic ancestral groups. Genetic ancestral groups were determined for BioVU[77], MGBB and PMBB on the basis of principal component analysis[78] and comparison to known ancestries in the 1000 Genomes Project Phase 3 (ref. 79) reference panel. In MVP, genetic ancestral groups were determined by harmonizing genetic ancestry and self-identified ancestry (HARE)[80], which also defines genetic ancestry on the basis of the 1000 Genomes reference panel. Further details on genotyping, phasing and imputation[81] for each site are included in the Supplementary Information. GWAS analyses were performed within each ancestral group using SAIGE v.0.44.6.5 (ref. 82) or PLINK v.2.0 (refs. 83,84) and a logistic regression. For the BioVU (7,167 cases and 39,738 controls), MGBB (6,708 cases and 15,560 controls) and UKBB (10,287 cases and 234,603 controls) cohorts, there were GWAS for only the EUR ancestral group. In PMBB, in addition to the EUR sample (3,088 cases and 25,911 controls) we conducted an additional GWAS of the African ancestral group sample (1,722 cases and 8,366 controls). In MVP, in addition to the EUR sample (146,771 cases and 250,062 controls), we performed additional GWAS of the AA (43,743 cases and 60,589 controls) and the LA (12,277 cases and 32,088 controls) ancestral groups. Each of the univariate GWAS covaried for ten genetic ancestry principal components (PC), age, sex, number of ICD codes and length of record. The summary statistics for TUD in UKBB were downloaded from the GWAS atlas (https://atlas.ctglab.nl/traitDB/3439).

SNP heritability

We estimated h^2_{SNP} based on the liability scale (population prevalence estimates of 0.125) for common SNPs mapped to HapMap3 (ref. 85) using LDSC[46]. For AA and LA, we created in-sample linkage disequilibrium (LD) scores derived from the MVP genotype data using cov-LDSC[86].

Meta-analyses and independent variants

Meta-analyses were conducted using a sample-size weighted method in METAL[87], assuming shared risk effects across ancestries. Effective sample sizes (n_{eff}), calculated using the formula: 4/(1/n_case + 1/n_control), were used to compensate for the imbalance in the ratio of cases to controls. n_{eff} were used in all meta-analyses and all downstream analyses.

We conducted five meta-analyses of TUD GWAS summary statistics across the following datasets: (1) within-ancestry meta-analysis for EUR samples in BioVU, MGBB, PMBB, MVP and an additional meta-analysis including UKBB, (2) within-ancestry meta-analysis for AA in MVP and Penn and (3) multi-ancestry meta-analysis across EUR (BioVU, MGBB, PMBB and MVP), AA (PMBB and MVP) and LA (MVP) datasets and an additional meta-analysis including UKBB. Inflation of test statistics due to polygenicity or cryptic relatedness was assessed using the LDSC attenuation ratio ((LDSC intercept − 1)/(mean of association chi-square statistics − 1)). Resulting GWS loci were defined as those with $P < 5.00 \times 10^{-8}$ with LD $r^2 > 0.1$, within a 1 Mb window, based on the structure of the Haplotype Reference Consortium (HRC) multi-ancestry reference panel for the multi-ancestry meta-analysis or the HRC ancestry-appropriate reference panel otherwise. GWS loci were examined for heterogeneity across cohorts via the I^2 inconsistency metric.

To identify TUD risk loci and lead SNPs, we performed LD clumping in FUMA[41] using a range of 3 Mb, $r^2 > 0.1$ and the respective ancestry 1000 Genome phase 3 reference panel[79]. Genomic risk loci that were located <1 Mb apart were incorporated into a single locus. For loci that harboured several variants, we used COJO in GCTA[88] to define independent variants by conditioning them on the most significant variant within each locus. Following conditioning, significant variants ($P < 5.00 \times 10^{-8}$) were considered independent.

We determined credible variants among the independent variants by merging risk variants within 1 Mb of the lead variant and fine-mapped the resulting region with 95% credible sets using FINEMAP[89]. Posterior inclusion probability ranges from 0 to 1 with values closer to 1 indicating greater causal probability. We implicated a putative causal variant if it accounted for >50% of the posterior probability in the 95% credible set.

Multi-ancestry fine-mapping analyses

We used PAINTOR v.3.1 (ref. 90) to perform multi-ancestry fine mapping for the two risk loci identified in both the TUD-EUR and TUD-AA metaGWAS. For each locus, we extracted SNPs with an absolute value of Z-score >3.9 within a 1 Mb region of the lead SNP. As suggested by PAINTOR, we created the AA and EUR LD matrices using the 1000 Genome phase 3 reference panel[79]. We calculated the probability of each SNP being the causal variant, assuming that each locus has two causal variants.

Gene-based and pathway analyses

We conducted bioannotation and bioinformatic analyses to further characterize the loci identified by the TUD GWAS (Supplementary Methods). We used the default version (v.1.3.6a) of the FUMA web-based platform[41] to identify independent SNPs ($r^2 < 0.10$) and to study their functional consequences. We also used MAGMA v.1.08 (refs. 41,42) to perform competitive gene set and pathway analyses. SNPs were mapped to 19,532 protein-coding genes from Ensembl (build 85). We applied a Bonferroni correction based on the total number of genes tested ($P < 2.63 \times 10^{-6}$). Gene sets were obtained from Msigdb v.7.0 ('Curated gene sets' and 'GO terms'). We also used Hi-C coupled MAGMA (H-MAGMA[43]) to assign non-coding (intergenic and intronic) SNPs to genes on the basis of their chromatin interactions. Exonic and promoter SNPs were assigned to genes on the basis of physical position. H-MAGMA uses four Hi-C datasets, which were derived from fetal brain, adult brain, iPSC-derived neurons and iPSC-derived astrocytes (https://github.com/thewonlab/H-MAGMA). We applied a Bonferroni correction based on the total number of gene–tissue pairs tested ($P < 9.44 \times 10^{-7}$).

S-MultiXcan/S-PrediXcan

We used S-MultiXcan v.0.7.0 (an extension of S-PrediXcan v.0.6.2; ref. 44) to identify specific expression QTL-linked genes associated with TUD. This approach uses genetic information to predict transcript abundance in 13 brain tissues and tests whether the predicted transcripts correlate with TUD. S-PrediXcan uses precomputed tissue weights from the genotype–tissue expression (GTEx) v.8 project database (https://www.gtexportal.org/) as the reference transcriptome dataset. For S-PrediXcan and S-MultiXcan analyses, we chose to use sparse (elastic net) prediction models, which are available at http://predictdb.hakyimlab.org/. We applied a conservative Bonferroni correction based on the total number of gene–tissue pairs tested (14,198 gene–tissue pairs tested; $P < 3.52 \times 10^{-6}$).

PWAS/TWAS

To identify proteins whose genetically regulated expression is associated with TUD, we performed PWAS analyses by integrating TUD GWAS summary statistics and precomputed protein QTLs from discovery (Banner)[91,92] and validation (ROSMAP)[93,94] datasets using the FUSION pipeline (http://gusevlab.org/projects/fusion/)[45]. Next, TWAS was performed using gene and splicing expression profiles measured in the adult dorsolateral prefrontal cortex and gene expression profiles from the frontal cortex. Human brain transcriptome data, used as expression reference panels, were obtained from the CMC[93] and GTEx frontal cortex v.7 (refs. 45,95). All tests were Bonferroni corrected for multiple testing ($\alpha = 0.05/n$ genes tested).

Of the overlapping findings across independent TWAS or PWAS datasets, colocalization analysis (in FUSION[45,96]) was used to determine whether SNPs mediate the association with TUD via effects on gene and protein expression. A posterior colocalization probability of 80% was used to indicate a shared causal signal.

Partitioning heritability enrichment

We used LDSC to partition TUD-EUR h^2_{SNP} and examined the enrichment on the basis of several functional genomic annotation models[97,98]. In the baseline model, we examined 75 overlapping functional annotations comprising genomic, epigenomic and regulatory features. We also analysed ten overlapping cell-type groups derived from 220 cell-type-specific annotations in four histone marks: methylated histone H3 Lys4 (H3K4me1), trimethylated histone H3 Lys4 (H3K4me3), acetylated histone H3 Lys4 (H3K4ac) and H3K27ac. Enriched cell-type categories were analysed on the basis of annotations obtained from H3K4me1-imputed, gapped peak data generated by the Roadmap Epigenomics Mapping Consortium[99]. We removed multi-allelic and major histocompatibility complex region variants and only report categories enriched after Bonferroni correction.

Tissue enrichment analysis

We used the LDSC package to conduct cell-type-specific heritability analysis[98]. In this analysis, we applied stratified LD score regression on the TUD-EUR meta-analysis summary statistics with sets of specifically expressed genes in various tissues from GTEx[95,100,101] to identify TUD-relevant tissues. We applied a conservative Bonferroni correction based on the number of tissues simultaneously tested (205 tissues tested, $P < 2.44 \times 10^{-4}$). We also used MAGMA v.1.08 gene–property analysis of expression data from GTEx (54 tissue types) and BrainSpan (29 brain samples at different age) in FUMA v.1.3.6a (ref. 102) to test the relationships between tissue-specific gene expression profiles and TUD-gene associations.

Cell-type-specific expression of TUD risk genes

We performed cell-type-specific analyses implemented in FUMA, using data from nine scRNA-seq datasets from human brain (datasets listed in the Supplementary Information). The method uses MAGMA gene–property analysis to test for association between cell-specific gene

expression and TUD-gene association[41]. Conditional analyses for multiple testing are applied to correct for all tested cell types across datasets.

BrainXcan

We used the BrainXcan package (https://github.com/hakyimlab/brainxcan)[103] to predict the association between the TUD phenotype and brain features. This approach uses genetically determined brain image-derived phenotypes (IDPs) to test brain region association with the TUD phenotype via linear regression. IDPs were constructed by training genetic predictors on original IDPs from magnetic resonance imaging via ridge regression[103]. IDPs were retrieved from the BrainXcan database (https://zenodo.org/record/4895174). Only significant IDP associations with TUD that survived a Bonferroni correction are reported (93 IDPs tested; $P < 1.92 \times 10^{-4}$).

Drug repurposing

Our signature matching technique used data from the LINCs L1000 database. The LINCs L1000 database catalogues in vitro gene expression profiles (signatures) from thousands of compounds in >80 human cell lines (level 5 data from phase I GSE92742 and phase II GSE70138). We selected compounds that were currently FDA approved or in clinical trials (via https://clue.io/repurposing#download-data; updated 24 March 2020). Our analyses included signatures of 829 chemical compounds (590 FDA approved, 239 in clinical trials) in five neuronal cell lines (NEU, NPC, MNEU.E, NPC.CAS9 and NPC.TAK), a total of 3,897 signatures.

We matched in vitro medication signatures with TUD signatures from brain tissue transcriptome-wide association analyses (conducted using S-PrediXcan). This consisted of amygdala, anterior cingulate cortex BA24, caudate basal ganglia, cerebellar hemisphere, cerebellum, cortex, frontal cortex BA9, hippocampus, hypothalamus, nucleus accumbens basal ganglia, putamen basal ganglia, substantia nigra and pituitary brain regions. We computed weighted Pearson correlations between transcriptome-wide brain associations and in vitro L1000 compound signatures[36], weighting each gene by its proportion of heritability explained, using the metafor package (v.3.8-1) in R. We treated each L1000 compound as a fixed effect incorporating the effect size (rweighted) and sampling variability (se2r_weighted) from all signatures of a compound (for example, across all time points, cell lines and doses). Brain region was included as a random effect to account for any tissue-specific heterogeneity. Both the genes for the transcriptome-wide association analysis input and the medications from our drug repurposing analyses were required to survive a Bonferroni correction for multiple testing (transcriptome-wide correction = 0.05/14,199 = 3.52×10^{-6}; Perturbagen correction = 0.05/3,897 = 1.28×10^{-5}).

We applied an additional drug repositioning method, DRUGSETS[104]. Data were drawn from the Clue Repurposing Hub and the Drug Gene Interaction Database. Drug–gene sets were created for 1,201 drugs with genes whose protein products are targeted by or interact with that specific drug. Competitive gene set analysis was performed using MAGMA v.1.08 (refs. 41,42) while conditioning on a gene set of all drug target genes in the data ($n = 2,281$) to test for significant associations between drug–gene sets and TUD. We applied a Bonferroni correction for the number of drug–gene sets tested ($P < 0.05/735 = 6.80 \times 10^{-5}$).

Genetic correlation analyses

We estimated the within-ancestry r_g for TUD using LDSC[46] and the cross-ancestry r_g for TUD across population groups using POPCORN[46]. We used the ancestry-specific 1000 Genomes Project phase 3 (ref. 80) data as the LD references.

We used local LDSC[46] to calculate genetic correlations (r_g) between TUD and 113 other traits or diseases[46]. Local traits were selected on the basis of previously known phenotypic associations between TUD and other SUD phenotypes and related traits (for example, CUD and various measures of impulsivity). We used the standard Benjamini–Hochberg FDR correction (FDR 5%) to correct for multiple testing. We also calculated a Bonferroni correction for 113 comparisons ($P < 4.42 \times 10^{-4}$); however, this correction is overly conservative because many of the traits tested are highly correlated with one another. For AA individuals, we calculated r_g between TUD and 11 published traits using in-sample LD scores derived from the MVP genotype data using cov-LDSC[86].

mtCOJO

We used mtCOJO[105] to individually condition the TUD-EUR summary statistics on loci associated with other comorbid traits, including alcohol dependence, CUD and OUD. This analysis allowed us to examine whether the genetic associations with TUD would be preserved when controlling for those covariate phenotypes. To test as many SNPs while preserving computational efficiency, we used P value thresholds of 5.00×10^{-6}, 5.00×10^{-8}, 5.00×10^{-6}, respectively, for alcohol dependence, CUD and OUD. We then computed genetic correlations using the TUD summary statistics adjusted for the covariates of interest.

Unsupervised learning to determine TUD clustering

Previous studies have shown that consumption and misuse/dependence phenotypes have a distinct genetic architecture. To explore whether the TUD meta-analysis clustered more with consumption or misuse/dependence phenotypes, we used a data-driven unsupervised machine learning method known as agglomerative hierarchical clustering analysis (HCA)[106]. HCA forms clusters iteratively by creating groups and successively joining or splitting those groups on the basis of a prespecified algorithm[106]. Agglomerative nesting (AGNES) is a bottom-up process focused on individual traits to structure. Agglomerative clustering was chosen as this allowed us to compare different algorithms to maximize for the dissimilarity on each branch, with Ward's minimum variance method performing best. All models were fit in R using the cluster package (v.2.1.4)[106].

The product of HCA is a dendrogram, formed with several brackets called 'branches'. Phenotypes on the same branch are more similar to each other on the basis of their pairwise genetic associations with each other and with all other phenotypes on that branch. Branches can form sub-branches of more specific clustering. The genetic correlations of former smoker and smoking initiation were reversed to show the intuitive effects against the other traits in the dendrogram.

Phenome-wide association studies

Mayo Clinic Biobank. We performed a PheWAS in the Mayo Clinic Biobank[107]. Phecodes were ascertained using EHR data from 57,001 patients from the Mayo Clinic Biobank. Genotyping details are included in the Supplementary Information. PGS were calculated using LDpred2 (ref. 108) using the auto feature in the bigsnpr (v.1.10.4) R package. To evaluate the unique contribution of polygenic scores for TUD in relation to other smoking behaviours, we calculated PGS for SmkInit, CPD[13] and FTND[27] and ran additional PheWAS of TUD covarying for SmkInit, CPD and FTND PGS.

Yale-Penn. We performed PheWAS in the Yale-Penn sample[47]; which is a genotyped[109] and deeply phenotyped cohort using the SSADDA, a detailed psychiatric instrument used to assess physical, psychosocial and psychiatric manifestations of SUDs and comorbid psychiatric traits[110,111]. This comprehensive interview includes more than 3,500 items representing lifetime diagnostic criteria for the DSM-IV (ref. 112), DSM-5 (ref. 113) SUDs and DSM-IV (ref. 112) psychiatric disorder history.

PGSs were calculated using PRS-Continuous shrinkage software (PRS-CS)[114]. We used the default setting in PRS-CS to estimate the shrinkage parameters and fixed the random seed to 1 for reproducibility. To identify associations between the PGS for TUD and clinical phenotypes, we performed a PheWAS by fitting logistic regression models for binary phenotypes and linear regression models for continuous phenotypes. Analyses were conducted using the PheWAS v.0.12 R

package[115] adjusting for sex, median age and the first ten PCs within each genetic ancestry. We performed sensitivity analyses by covarying for SmkInit, CPD[13] and FTND[27] PGS. Bonferroni correction was applied for each ancestral-specific analysis to account for multiple testing ($P < 7.25 \times 10^{-5}$).

Adolescent brain cognitive development. We performed polygenic analyses in the ABCD sample[116]. Again using PRS-CS[117], we fitted a fixed effects model in the ABCD European subsample (wave 3 for phenotypes and wave 3 for genotypes), controlling for first ten PCs, age, sex, site, as fixed effect covariates and family ID as random effects covariates. We included 12 measures which showed significant r_g in the adults datasets and were available in this cohort; these included two binary phenotypes (pain, 'any pain last month'; and suicide attempt, 'description') and ten continuous measures (from CBCL[118]—'CBCL Externalizing', 'CBCL ADHD', 'CBCL Depression', 'CBCL AnxDep'; 'CBCL AnxDis'; 'CBCL OCD'; cognitive ability via the National Institutes of Health (NIH) cognitive toolbox total score[119]; BMI; weight; deprivation). Results were corrected for multiple testing ($P < 4.0 \times 10^{-3}$). Additional genotyping, quality control and statistical details are described in the Supplementary Information.

Mendelian randomization

Two-sample MR[120,121] was used to evaluate the potential causal association between TUD and genetically correlated traits using samples of EUR ancestry only (without UKBB). Of the 76 traits that showed significant genetic correlations (Supplementary Table 31), we removed 45 that were phenotypically similar (for example, BMI and obesity). From each category, we selected those traits with higher r_g. Therefore, we tested 31 traits for a causal relationship with TUD. We inferred causality bidirectionally using three methods: weighted median, IVW and MR-Egger, followed by a pleiotropy test using the MR-Egger intercept[122,123]. Instrumental variants were those associated with the exposure after clumping ($r^2 = 0.01$) and at $P < 1.0 \times 10^{-5}$. We considered causal effects as those for which at least two MR tests were significant after Bonferroni correction ($P = 0.05/31 = 1.61 \times 10^{-3}$) and that showed no evidence of violation of the horizontal pleiotropy test (MR-Egger intercept $P > 0.05$).

Statistics and reproducibility

All statistical analyses performed as part of this study have been described in the Methods. No statistical method was used to predetermine sample size. Randomization and blinding did not apply.

Reporting summary

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

Data availability

Summary statistics can be accessed at the PsycheMERGE website (https://psychemerge.com) or by emailing the corresponding author (sanchezroige@ucsd.edu). The following datasets were retrieved for secondary analyses: Ensembl build 85 (https://www.ebi.ac.uk/about/news/updates-from-data-resources/ensembl-version-85/), Msigdb v.7.0 (https://data.broadinstitute.org/gsea-msigdb/msigdb/release/7.0/), Genotype–Tissue Expression (GTEx) v.8 project database (https://www.gtexportal.org/), PredictDB Data Repository (http://predictdb.hakyimlab.org/), BrainQTL (http://predictdb.hakyimlab.org/), BrainXcan database (https://zenodo.org/record/4895174), LINCs L1000 database (https://commonfund.nih.gov/LINCS), Drug Gene Interaction Database (https://repo-hub.broadinstitute.org/repurposing#download-data), 1000 Genomes Project phase 3 (https://internationalgenome.org/data-portal/sample), BrainSpan (http://www.brainspan.org/), H-MAGMA four Hi-C datasets provided with the software (https://github.com/thewonlab/H-MAGMA/tree/master/Input_Files) and PredictDB Data Repository (http://predictdb.org/).

Code availability

All software used to generate results has been previously published and corresponding citations are provided in Methods; that is, SAIGE v.0.44.6.5, PLINK v.1.9/v.2.080, LDSC v.1.0.1, cov-LDSC (https://github.com/yang-luo-lab/cov-ldsc), METAL 2020-05-05 (https://github.com/statgen/METAL), FUMA v.1.3.6a (https://fuma.ctglab.nl/), COJO in GCTA v.1.94.1, FINEMAP v.1.4.2, PAINTOR v.3.1, MAGMA v.1.08, H-MAGMA v.1.08, S-MultiXcan v.0.7.0, S-PrediXcan v.0.6.2, BrainXcan (https://github.com/hakyimlab/brainxcan), metafor package (v.3.8-1), DRUGSETS (https://github.com/nybell/drugsets), POPCORN (https://github.com/brielin/Popcorn), mtCOJO in GCTA v.1.94.1, cluster package v.2.1.4, PheWAS package v.0.12, LDpred2 from the bigsnpr package v.1.10.4, PRS-CS/PRS-CSx v.1.0.0 (https://github.com/getian107) and MendelianRandomization package v.0.9.0.

References

1. *Health Effects of Cigarette Smoking* (CDC, 2021); www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm
2. Oliver, J. A. & Foulds, J. Association between cigarette smoking frequency and tobacco use disorder in U.S. adults. *Am. J. Prev. Med.* **60**, 726–728 (2021).
3. *The Top 10 Causes of Death* (WHO, 2020); www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
4. Benowitz, N. L. & Liakoni, E. Tobacco use disorder and cardiovascular health. *Addiction* **117**, 1128–1138 (2022).
5. Kalman, D., Morissette, S. B. & George, T. P. Co-morbidity of smoking in patients with psychiatric and substance use disorders. *Am. J. Addict.* **14**, 106–123 (2005).
6. McRobbie, H. & Kwan, B. Tobacco use disorder and the lungs. *Addiction* **116**, 2559–2571 (2021).
7. Ziedonis, D., Das, S. & Larkin, C. Tobacco use disorder and treatment: new challenges and opportunities. *Dialogues Clin. Neurosci.* **19**, 271–280 (2017).
8. Kendler, K. S., Schmitt, E., Aggen, S. H. & Prescott, C. A. Genetic and environmental influences on alcohol, caffeine, cannabis and nicotine use from early adolescence to middle adulthood. *Arch. Gen. Psychiatry* **65**, 674–682 (2008).
9. Do, E. K. et al. Genetic and environmental influences on smoking behavior across adolescence and young adulthood in the Virginia twin study of adolescent behavioral development and the transitions to substance abuse follow-up. *Twin Res. Hum. Genet.* **18**, 43–51 (2015).
10. Agrawal, A., Budney, A. J. & Lynskey, M. T. The co-occurring use and misuse of cannabis and tobacco: a review. *Addiction* **107**, 1221–1233 (2012).
11. Agrawal, A. et al. The genetics of addiction—a translational perspective. *Transl. Psychiatry* **2**, e140–e140 (2012).
12. Sullivan, P. F. & Kendler, K. S. The genetic epidemiology of smoking. *Nicotine Tob. Res.* **1**, S51–S57 (1999).
13. Saunders, G. R. B. et al. Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* **612**, 720–724 (2022).
14. Larsson, S. C. & Burgess, S. Appraising the causal role of smoking in several diseases: a systematic review and meta-analysis of Mendelian randomization studies. *eBioMedicine* **82**, 104154 (2022).
15. Yuan, S., Michaëlsson, K., Wan, Z. & Larsson, S. C. Associations of smoking and alcohol and coffee intake with fracture and bone mineral density: a Mendelian randomization study. *Calcif. Tissue Int.* **105**, 582–588 (2019).
16. Mahedy, L. et al. Testing the association between tobacco and cannabis use and cognitive functioning: findings from an observational and Mendelian randomization study. *Drug Alcohol Depend.* **221**, 108591 (2021).
17. Zhou, H. et al. Association of *OPRM1* functional coding variant with opioid use disorder: a genome-wide association study. *JAMA Psychiatry* **77**, 1072 (2020).

18. Wootton, R. E. et al. Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. *Psychol. Med.* **50**, 2435–2443 (2020).
19. Harrison, R., Munafò, M. R., Davey Smith, G. & Wootton, R. E. Examining the effect of smoking on suicidal ideation and attempts: triangulation of epidemiological approaches. *Br. J. Psychiatry* **217**, 701–707 (2020).
20. Xu, K. et al. Genome-wide association study of smoking trajectory and meta-analysis of smoking status in 842,000 individuals. *Nat. Commun.* **11**, 5302 (2020).
21. Sanchez-Roige, S. et al. Genome-wide association study of alcohol use disorder identification test (AUDIT) scores in 20 328 research participants of European ancestry: GWAS of AUDIT. *Addict. Biol.* **24**, 121–131 (2019).
22. 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).
23. Mallard, T. T. & Sanchez-Roige, S. Dimensional phenotypes in psychiatric genetics: lessons from genome-wide association studies of alcohol use phenotypes. *Complex Psychiatry* **7**, 45–48 (2021).
24. Mallard, T. T. et al. Item-level genome-wide association study of the alcohol use disorders identification test in three population-based cohorts. *Am. J. Psychiatry* https://doi.org/10.1176/appi.ajp.2020.20091390 (2021).
25. Sanchez-Roige, S. & Palmer, A. A. Emerging phenotyping strategies will advance our understanding of psychiatric genetics. *Nat. Neurosci.* **23**, 475–480 (2020).
26. Johnson, E. C. et al. A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Psychiatry* **7**, 1032–1045 (2020).
27. Quach, B. C. et al. Expanding the genetic architecture of nicotine dependence and its shared genetics with multiple traits. *Nat. Commun.* **11**, 5562 (2020).
28. Hancock, D. B., Markunas, C. A., Bierut, L. J. & Johnson, E. O. Human genetics of addiction: new insights and future directions. *Curr. Psychiatry Rep.* **20**, 8 (2018).
29. Sanchez-Roige, S., Cox, N. J., Johnson, E. O., Hancock, D. B. & Davis, L. K. Alcohol and cigarette smoking consumption as genetic proxies for alcohol misuse and nicotine dependence. *Drug Alcohol Depend.* **221**, 108612 (2021).
30. DeBoever, C. et al. Assessing digital phenotyping to enhance genetic studies of human diseases. *Am. J. Hum. Genet.* **106**, 611–622 (2020).
31. Sanchez-Roige, S. & Palmer, A. A. Electronic health records are the next frontier for the genetics of substance use disorders. *Trends Genet.* **35**, 317–318 (2019).
32. Zheutlin, A. B. et al. Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 106,160 patients across four health care systems. *Am. J. Psychiatry* **176**, 846–855 (2019).
33. Verma, A. et al. The Penn medicine bioBank: towards a genomics-enabled learning healthcare system to accelerate precision medicine in a diverse population. *J. Pers. Med.* **12**, 1974 (2022).
34. Roughley, S., Marcus, A. & Killcross, S. Dopamine D1 and D2 receptors are important for learning about neutral-valence relationships in sensory preconditioning. *Front. Behav. Neurosci.* **15**, 740992 (2021).
35. Gelernter, J. et al. Haplotype spanning TTC12 and ANKK1, flanked by the DRD2 and NCAM1 loci, is strongly associated to nicotine dependence in two distinct American populations. *Hum. Mol. Genet.* **15**, 3498–3507 (2006).
36. Hatoum, A. S. et al. Multivariate genome-wide association meta-analysis of over 1 million subjects identifies loci underlying multiple substance use disorders. *Nat. Mental Health.* **1**, 210–223 (2023).
37. 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).
38. Sanchez-Roige, S. et al. Genome-wide association study of problematic opioid prescription use in 132,113 23andMe research participants of European ancestry. *Mol. Psychiatry* **26**, 6209–6217 (2021).
39. 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).
40. Xiao, M.-F. et al. Neural cell adhesion molecule modulates dopaminergic signaling and behavior by regulating dopamine D2 receptor internalization. *J. Neurosci.* **29**, 14752–14763 (2009).
41. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. *Nat. Commun.* **10**, 3222 (2019).
42. Leeuw, C. A., de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
43. 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).
44. 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).
45. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
46. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
47. Kember, R. L. et al. Phenome-wide association analysis of substance use disorders in a deeply phenotyped sample. *Biol. Psychiatry* **93**, 536–545 (2023).
48. Sanchez-Roige, S., Palmer, A. A. & Clarke, T.-K. Recent efforts to dissect the genetic basis of alcohol use and abuse. *Biol. Psychiatry* **87**, 609–618 (2020).
49. McLellan, A. T., Koob, G. F. & Volkow, N. D. Preaddiction—a missing concept for treating substance use disorders. *JAMA Psychiatry* **79**, 749–751 (2022).
50. Miranda, M., Morici, J. F., Zanoni, M. B. & Bekinschtein, P. Brain-derived neurotrophic factor: a key molecule for memory in the healthy and the pathological brain. *Front. Cell. Neurosci.* **13**, 363 (2019).
51. Barker, J. M., Taylor, J. R., De Vries, T. J. & Peters, J. Brain-derived neurotrophic factor and addiction: pathological versus therapeutic effects on drug seeking. *Brain Res.* **1628**, 68–81 (2015).
52. Duong, C. et al. Glutathione peroxidase-1 protects against cigarette smoke-induced lung inflammation in mice. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **299**, L425–L433 (2010).
53. Scieszka, D. et al. Subchronic electronic cigarette exposures have overlapping protein biomarkers with chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. *Am. J. Respir. Cell Mol. Biol.* **67**, 503–506 (2022).
54. Aberg, K. A. et al. A comprehensive family-based replication study of schizophrenia genes. *JAMA Psychiatry* **70**, 573 (2013).
55. Erzurumluoglu, A. M. et al. Meta-analysis of up to 622,409 individuals identifies 40 new smoking behaviour associated genetic loci. *Mol. Psychiatry* **25**, 2392–2409 (2020).
56. Toikumo, S., Xu, H., Gelernter, J., Kember, R. L. & Kranzler, H. R. Integrating human brain proteomic data with genome-wide association study findings identifies new brain proteins in substance use traits. *Neuropsychopharmacology* **47**, 2292–2299 (2022).

57. Kember, R. L. et al. Cross-ancestry meta-analysis of opioid use disorder uncovers new loci with predominant effects in brain regions associated with addiction. *Nat. Neurosci.* **25**, 1279–1287 (2022).
58. Koob, G. F. & Volkow, N. D. Neurobiology of addiction: a neurocircuitry analysis. *Lancet Psychiatry* **3**, 760–773 (2016).
59. King, D. P. et al. Smoking cessation pharmacogenetics: analysis of varenicline and bupropion in placebo-controlled clinical trials. *Neuropsychopharmacology* **37**, 641–650 (2012).
60. King, A. C. et al. Effects of naltrexone on smoking cessation outcomes and weight gain in nicotine-dependent men and women. *J. Clin. Psychopharmacol.* **32**, 630–636 (2012).
61. Carpenter, M. J. et al. Clinical strategies to enhance the efficacy of nicotine replacement therapy for smoking cessation: a review of the literature. *Drugs* **73**, 407–426 (2013).
62. So, H.-C. et al. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. *Nat. Neurosci.* **20**, 1342–1349 (2017).
63. Sey, N. Y. A. et al. Chromatin architecture in addiction circuitry identifies risk genes and potential biological mechanisms underlying cigarette smoking and alcohol use traits. *Mol. Psychiatry* **27**, 3085–3094 (2022).
64. Chen, F. et al. Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. *Nat. Genet.* **55**, 291–300 (2023).
65. Jamali, Q. Galantamine as a treatment option for nicotine addiction. *J. Smok. Cessat.* **2021**, 9975811 (2021).
66. McGinnis, K. A. et al. Using the biomarker cotinine and survey self-report to validate smoking data from United States Veterans Health Administration electronic health records. *JAMIA Open* **5**, ooac040 (2022).
67. Border, R. et al. Cross-trait assortative mating is widespread and inflates genetic correlation estimates. *Science* **378**, 754–761 (2022).
68. Brazel, D. M. et al. Exome ChIP meta-analysis fine maps causal variants and elucidates the genetic architecture of rare coding variants in smoking and alcohol use. *Biol. Psychiatry* **85**, 946–955 (2019).
69. Jang, S.-K. et al. Rare genetic variants explain missing heritability in smoking. *Nat. Hum. Behav.* **6**, 1577–1586 (2022).
70. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).
71. Hiscock, R., Bauld, L., Amos, A., Fidler, J. A. & Munafò, M. Socioeconomic status and smoking: a review. *Ann. NY Acad. Sci.* **1248**, 107–123 (2012).
72. Pasman, J. A. et al. Genetic risk for smoking: disentangling interplay between genes and socioeconomic status. *Behav. Genet.* **52**, 92–107 (2022).
73. Treur, J. L. et al. Testing familial transmission of smoking with two different research designs. *Nicotine Tob. Res.* **20**, 836–842 (2018).
74. Meyers, J. L. et al. Interaction between polygenic risk for cigarette use and environmental exposures in the Detroit Neighborhood Health Study. *Transl. Psychiatry* **3**, e290 (2013).
75. Pasman, J. A., Verweij, K. J. H. & Vink, J. M. Systematic review of polygenic gene-environment interaction in tobacco, alcohol and cannabis use. *Behav. Genet.* **49**, 349–365 (2019).
76. Sanchez-Roige, S., Kember, R. L. & Agrawal, A. Substance use and common contributors to morbidity: a genetics perspective. *EBioMedicine* **83**, 104212 (2022).
77. Dennis, J. K. et al. Clinical laboratory test-wide association scan of polygenic scores identifies biomarkers of complex disease. *Genome Med.* **13**, 6 (2021).
78. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
79. The 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
80. Fang, H. et al. Harmonizing genetic ancestry and self-identified race/ethnicity in genome-wide association studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).
81. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
82. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
83. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
84. 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).
85. Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E. T. & Schaffner, S. F. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
86. Luo, Y. et al. Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Hum. Mol. Genet.* **30**, 1521–1534 (2021).
87. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
88. 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).
89. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
90. Kichaev, G. & Pasaniuc, B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).
91. Beach, T. G. et al. Arizona study of aging and neurodegenerative disorders and brain and body donation program. *Neuropathology* **35**, 354–389 (2015).
92. Wingo, T. S. et al. Brain proteome-wide association study implicates new proteins in depression pathogenesis. *Nat. Neurosci.* **24**, 810–817 (2021).
93. Wingo, A. P. et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nat. Genet.* **53**, 143–146 (2021).
94. Bennett, D. A. et al. Religious orders study and rush memory and aging project. *J. Alzheimers Dis.* **64**, S161–S189 (2018).
95. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
96. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
97. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
98. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
99. Bernstein, B. E. et al. The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
100. Fehrmann, R. S. N. et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47**, 115–125 (2015).
101. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
102. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).

103. Liang, Y. et al. BrainXcan identifies brain features associated with behavioral and psychiatric traits using large scale genetic and imaging data. Preprint at *medRxiv* https://doi.org/10.1101/2021.06.01.21258159 (2022).
104. Bell, N., Uffelmann, E., van Walree, E., de Leeuw, C. & Posthuma, D. Using genome-wide association results to identify drug repurposing candidates. Preprint at *medRxiv* https://doi.org/10.1101/2022.09.06.22279660 (2022).
105. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
106. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. cluster: Cluster analysis basics and extentions. R package version 2.1.4 (2013).
107. Bielinski, S. J. et al. Mayo Genome Consortia: a genotype–phenotype resource for genome-wide association studies with an application to the analysis of circulating bilirubin levels. *Mayo Clin. Proc.* **86**, 606–614 (2011).
108. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2020).
109. Gelernter, J. et al. Genome-wide association study of alcohol dependence: significant findings in African- and European-Americans including new risk loci. *Mol. Psychiatry* **19**, 41–49 (2014).
110. Pierucci-Lagha, A. et al. Diagnostic reliability of the Semi-structured Assessment for Drug Dependence and Alcoholism (SSADDA). *Drug Alcohol Depend.* **80**, 303–312 (2005).
111. 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).
112. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* (American Psychiatric Association, 1994).
113. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (American Psychiatric Association, 2013).
114. 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**, 1776 (2019).
115. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-wide association studies as a tool to advance precision medicine. *Annu. Rev. Genomics Hum. Genet.* **17**, 353–373 (2016).
116. Lam, M. et al. RICOPILI: Rapid imputation for COnsortias PIpeLIne. *Bioinformatics* **36**, 930–933 (2020).
117. Ruan, Y. et al. Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
118. Rescorla, L. et al. Behavioral/emotional problems of preschoolers caregiver/teacher reports from 15 societies. *J. Emot. Behav. Disord.* **20**, 68–81 (2012).
119. Akshoomoff, N. et al. NIH Toolbox Cognitive Function Battery (CFB): composite scores of crystallized, fluid and overall cognition. *Monogr. Soc. Res. Child Dev.* **78**, 119–132 (2013).
120. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).
121. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
122. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with several genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
123. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).

Acknowledgements

M.V.J., S.B.B., S.R.P. and S.S.-R. were supported by funds from the California Tobacco-Related Disease Research Program (grant nos T29KT0526 and T32IR5226). S.B.B. was also supported by P50DA037844. B.K.P., J.J.M. and S.S.-R. were supported by NIH/NIDA DP1DA054394. A.S.H. was supported by NIAAA AA030083. T.T.M. was supported by NHGRI T32HG010464. E.C.J. was supported by K01DA051759. J.G. was supported by VA Merit Award CX001849-01 and 5R01DA054869. D.B.H. was supported by R01 DA042090 and R01 DA051913. L.K.D. was supported by R01 MH113362. H.R.K. was supported by the Veterans Integrated Service Network 4 Mental Illness Research, Education and Clinical Center. R.L.K. was supported by NIAAA K01 AA028292. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. CTSA (SD, Vanderbilt Resources): The project described was supported by the National Center for Research Resources, grant UL1RR024975-01 and is now at the National Center for Advancing Translational Sciences, grant 2 UL1TR000445-06. BioVU: The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by numerous sources: institutional funding, private agencies and federal grants. These include the NIH funded Shared Instrumentation grant S10RR025141; and CTSA grants UL1TR002243, UL1TR000445 and UL1RR024975. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962 amd R01HD074711; and additional funding sources listed at https://victr.vumc.org/biovu-funding/. This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration and was supported by funding from the Department of Veterans Affairs Office of Research and Development, Million Veteran Program grant no. I01 BX004820. This publication does not represent the views of the Department of Veterans Affairs or the United States Government. We acknowledge the Penn Medicine BioBank and the Mayo Clinic Biobank for providing data and thank the patient-participants of Penn Medicine and Mayo Clinic who consented to participate in this research programme. We also thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol no. 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family and the National Center for Advancing Translational Sciences of the NIH under CTSA award no. UL1TR001878. Data used in the preparation of this article were obtained from the ABCD study (https://abcdstudy.org), held in the NIMH Data Archive. This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9–10 years and follow them over 10 years into early adulthood. The ABCD study is supported by the NIH and additional federal partners under award nos. U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, U24DA041147, U01DA041093 and U01DA041025. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/Consortium_Members. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. We also thank the Externalizing Consortium for sharing the GWAS summary statistics of externalizing. The Externalizing Consortium comprises: principal investigators D. M. Dick, P. Koellinger, K. P. Harden, A A.P.; lead analysts R. K. Linnér,

T.T.M., P. B. Barr and S.S.-R; contributor I. D. Waldman. The Externalizing Consortium has been supported by the National Institute on Alcohol Abuse and Alcoholism (R01AA015416—administrative supplement) and the National Institute on Drug Abuse (R01DA050721). Additional funding for investigator effort has been provided by K02AA018755, U10AA008401 and P50AA022537, 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, Add Health, Vanderbilt University Medical Center's BioVU, Collaborative Study on the Genetics of Alcoholism (COGA), the Psychiatric Genomics Consortium's SUDs working group, UK10K Consortium, UK Biobank and Philadelphia Neurodevelopmental Cohort. We thank B. Quach and J. Marks for their help in supplying portions of the data needed to create Supplementary Fig. 1.

Author contributions

S.S.-R. conceived the idea for the paper and wrote and edited the paper. S.T., M.V.J., B.K.P., H.L., T.T.M., S.B.B., L.V.-R., H.X., A.S.H., J.J.M., V.K.P., B.J.C. and R.LK. provided analyses. E.C.J., G.D.J., A.B., R.P., J.M.B., J.W.S., L.K.D., A.C.J. and R.L.K. contributed data. All contributing authors wrote and edited the paper.

Competing interests

J.W.S. is a member of the Scientific Advisory Board of Sensorium Therapeutics (with equity) and has received grant support from Biogen. He is Principal Investigator of a collaborative study of the genetics of depression and bipolar disorder sponsored by 23andMe for which 23andMe provides analysis time as in-kind support but no payments. H.R.K. is a member of advisory boards for Clearmind Medicine, Dicerna Pharmaceuticals, Sophrosyne Pharmaceuticals and Enthion Pharmaceuticals; a consultant to Sobrera Pharmaceuticals; the recipient of research funding and medication supplies for an investigator-initiated study from Alkermes; a member of the American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative, which was supported in the last 3 years by Alkermes, Dicerna, Ethypharm, Lundbeck, Mitsubishi, Otsuka and Pear Therapeutics; and, with J.G., is a holder of US patent 10,900,082 titled: 'Genotype-guided dosing of opioid agonists' issued 26 January 2021. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01851-6.

Correspondence and requests for materials should be addressed to Sandra Sanchez-Roige.

Peer review information *Nature Human Behaviour* thanks Robbee Wedow and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Mental Illness Research, Education and Clinical Center, Crescenz VAMC, Philadelphia, PA, USA. [2]Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [3]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA. [4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [5]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [6]Department of Psychiatry, Harvard Medical School, Boston, MA, USA. [7]Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [8]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Boston, MA, USA. [9]Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addiction, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain. [10]Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA. [11]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA. [12]Program in Biomedical Sciences, University of California San Diego, La Jolla, CA, USA. [13]Department of Medicine, Division of Genetic Medicine, Vanderbilt University, Nashville, TN, USA. [14]Department of Psychology, University of Ibadan, Ibadan, Nigeria. [15]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA. [16]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA. [17]Veterans Affairs Connecticut Healthcare System, West Haven, CT, USA. [18]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, USA. [19]RTI International, Research Triangle Park, NC, USA. [20]Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN, USA. [21]Yale University School of Public Health, New Haven, CT, USA. [22]Yale University School of Medicine, New Haven, CT, USA. [23]These authors contributed equally: Sylvanus Toikumo, Mariela V. Jennings. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: sanchezroige@ucsd.edu

Penn Medicine BioBank

Sylvanus Toikumo[1,2,23], Rachel L. Kember[1,2], Million Veteran Program* & PsycheMERGE Substance Use Disorder Workgroup*

Full lists of members and their affiliations appear in the Supplementary Information.

nature portfolio

Corresponding author(s): Sandra Sanchez-Roige

Last updated by author(s): Nov 25, 2023

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.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

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	Data collection was not performed as part of this study.
Data analysis	SAIGE v0.44.6.5, PLINK v1.9/v2.080, LDSC v1.0.1, cov-LDSC, METAL 2020-05-05, FUMA v1.3.6a, COJO/GCTA v1.94.1, FINEMAP v1.4.2, PAINTOR v3.1, MAGMA v1.08, H-MAGMA v1.08, S-MultiXcan v0.7.0, S-PrediXcan v0.6.2, BrainXcan, metafor v3.8-1, DRUGSETS, POPCORN, mtCOJO/GCTA v1.94.1, cluster v2.1.4, PheWAS v0.12, LDpred2/bigsnpr v1.10.4, PRS-CS/PRS-CSx v1.0.0, MendelianRandomization v 0.9.0.

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

GWAS summary statistics can be accessed at the PsycheMERGE website (https://psychemerge.com) or by emailing the corresponding author (sanchezroige@ucsd.edu). The following datasets were retrieved for secondary analyses: Ensembl build 85 (https://www.ebi.ac.uk/about/news/updates-from-data-resources/ensembl-version-85/), Msigdb v7.0 (https://data.broadinstitute.org/gsea-msigdb/msigdb/release/7.0/), Genotype-Tissue Expression (GTEx) v8 project database (https://www.gtexportal.org/), PredictDB Data Repository (http://predictdb.hakyimlab.org/), BrainQTL (http://predictdb.hakyimlab.org/), BrainXcan database (https://zenodo.org/record/4895174), Library of Integrated Network-based Cellular Signatures (LINCs) L1000 database (https://commonfund.nih.gov/LINCS), Drug Gene Interaction Database (https://repo-hub.broadinstitute.org/repurposing#download-data), 1000 Genomes Project phase 3 (https://internationalgenome.org/data-portal/sample), BrainSpan (http://www.brainspan.org/), H-MAGMA four Hi-C datasets provided with the software (https://github.com/thewonlab/H-MAGMA/tree/master/Input_Files), PredictDB Data Repository (http://predictdb.org/

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 have no information on gender; all analyses included sex as covariate.
Reporting on race, ethnicity, or other socially relevant groupings	Ancestry was determined though an analysis of genotype data, as detailed in the Methods. Sample size breakdown by ancestry is shown in Supplementary Table 5.
Population characteristics	Demographic information is detailed in Supplementary Tables 2 and 5.
Recruitment	Recruitment did not apply; we included data already available via electronic health records.
Ethics oversight	The project was approved by the VUMC, MGBB, PMBB and Mayo Clinic IRB. Central VA IRB and site-specific IRBs approved the MVP study.

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

Life sciences study design

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

Sample size	We used an opportunistic sample, reflecting our best efforts to obtain the highest numbers of subjects with phenotype/genotype data available for this study (N=898,680).
Data exclusions	All unrelated subjects that passed quality control and had available phenotype and genotype information were retained. All exclusion criteria are detailed in the Methods.
Replication	Genome-wide genetic correlations between all study cohorts and independent smoking cohorts were high, and our GWAS results replicated those from prior smoking GWAS
Randomization	Randomization does not apply to observational studies (i.e., GWAS) like ours.
Blinding	Blinding (i.e., masking the identify of the groups from researchers) does not apply to genetic studies like ours (i.e., GWAS).

Behavioural & social sciences study design

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

Study description	N/A
Research sample	N/A
Sampling strategy	N/A
Data collection	N/A
Timing	N/A
Data exclusions	N/A
Non-participation	N/A
Randomization	N/A

Ecological, evolutionary & environmental sciences study design

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

Study description	N/A
Research sample	N/A
Sampling strategy	N/A
Data collection	N/A
Timing and spatial scale	N/A
Data exclusions	N/A
Reproducibility	N/A
Randomization	N/A
Blinding	N/A

Did the study involve field work? ☐ Yes ☑ No

Field work, collection and transport

Field conditions	N/A
Location	N/A
Access & import/export	N/A
Disturbance	N/A

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

Antibodies

Antibodies used	N/A
Validation	N/A

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	N/A
Authentication	N/A
Mycoplasma contamination	N/A
Commonly misidentified lines (See ICLAC register)	N/A

Palaeontology and Archaeology

Specimen provenance	N/A
Specimen deposition	N/A
Dating methods	N/A

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	N/A

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

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	N/A
Wild animals	N/A
Reporting on sex	N/A
Field-collected samples	N/A
Ethics oversight	N/A

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	N/A
Study protocol	N/A
Data collection	N/A
Outcomes	N/A

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
X	☐	Public health
X	☐	National security
X	☐	Crops and/or livestock
X	☐	Ecosystems
X	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
X	☐	Demonstrate how to render a vaccine ineffective
X	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
X	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
X	☐	Increase transmissibility of a pathogen
X	☐	Alter the host range of a pathogen
X	☐	Enable evasion of diagnostic/detection modalities
X	☐	Enable the weaponization of a biological agent or toxin
X	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	N/A
Files in database submission	N/A
Genome browser session (e.g. UCSC)	N/A

Methodology

Replicates	N/A
Sequencing depth	N/A
Antibodies	N/A
Peak calling parameters	N/A
Data quality	N/A

Software

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	N/A
Instrument	N/A
Software	N/A
Cell population abundance	N/A
Gating strategy	N/A

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	N/A
Design specifications	N/A
Behavioral performance measures	N/A

Imaging type(s)	N/A
Field strength	N/A
Sequence & imaging parameters	N/A
Area of acquisition	N/A

Diffusion MRI ☐ Used ☑ Not used

Preprocessing

Preprocessing software	N/A
Normalization	N/A
Normalization template	N/A
Noise and artifact removal	N/A
Volume censoring	N/A

Statistical modeling & inference

Model type and settings	N/A
Effect(s) tested	N/A

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)

N/A

Correction

N/A

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

N/A

Graph analysis

N/A

Multivariate modeling and predictive analysis

N/A