Multi-ancestry meta-analysis of tobacco use disorders based on electronic health record data prioritizes novel candidate risk genes and reveals associations with numerous health outcomes

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

[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, San Diego, CA, USA; [4]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; [5]Department of Psychiatry, Harvard Medical School, Boston, MA, USA; [6]Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addiction, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain; [7]Department of Psychiatry, Washington University School of Medicine, Saint Louis, Missouri, USA; [8]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA; [9]Program in Biomedical Sciences, University of California San Diego, La Jolla, CA, USA; [10]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA; [11]Department of Psychology, University of Ibadan, Nigeria; [12]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA; [13]Veterans Affairs Connecticut Healthcare System, West Haven, CT, USA; [14]Department of Psychiatry & Psychology, Mayo Clinic, Rochester, MN, USA; [15]Behavioral and Urban Health Program, Behavioral Health and Criminal Justice Division, RTI International, Research Triangle Park, NC, USA; [16]Department of Medicine, Division of Genetic Medicine, Vanderbilt University, Nashville, TN, USA; [17]Yale University School of Public Health, New Haven, CT, USA; [18]Veterans Affairs Connecticut Healthcare System, West Haven, CT, USA; [19]Yale University School of Medicine, New Haven, CT, USA.

* These authors contributed equally

Penn Medicine BioBank is provided in the Supplementary Material

Million Veteran Program is provided in the Supplementary Material

PsycheMERGE Substance Use Disorders Group (alphabetically): Sevim B Bianchi, Brandon J Coombes, Richard C Crist, Renata Cupertino, Lea K Davis, Mariela V Jennings, Rachel L Kember, Hyunjoon Lee, Travis T Mallard, Maria Niarchou, Melissa N Poulsen, Sandra Sanchez-Roige, Jordan W Smoller, Vanessa Troiani, Colin G Walsh

Corresponding author: Sandra Sanchez-Roige, PhD, Department of Psychiatry, University of California San Diego, San Diego, CA, USA. Email: sanchezroige@ucsd.edu

ABSTRACT

Tobacco use disorder (**TUD**) is the most prevalent substance use disorder in the world. Genetic factors influence smoking behaviors, and although strides have been made using genome-wide association studies (**GWAS**) to identify risk variants, the majority of variants identified have been for nicotine consumption, rather than TUD. We leveraged five biobanks to perform a multi-ancestral meta-analysis of TUD (derived via electronic health records, **EHR**) in 898,680 individuals (739,895 European, 114,420 African American, 44,365 Latin American). We identified 72 independent risk loci; integration with functional genomic tools uncovered 330 potential risk genes, primarily expressed in the brain. TUD was genetically correlated with smoking and psychiatric traits from traditionally ascertained cohorts, externalizing behaviors in children, and hundreds of medical outcomes, including HIV infection, heart disease, and pain. This work furthers our biological understanding of TUD and establishes EHR as a source of phenotypic information for studying the genetics of TUD.

Tobacco use disorder (**TUD**) is the most prevalent substance use disorder in the world, with a high proportion of smokers meeting criteria for nicotine dependence.[1,2] Nicotine dependent individuals 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 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 substance use disorders.[4–6] Unfortunately, available preventative and treatment options for TUD have low success rates.[7]

Genetic factors influence smoking behaviors, with twin-heritability estimates ranging from ~30-70%.[8–12] Recently, genome-wide association studies (**GWAS**) have expanded in size (N~2.5M) and yielded hundreds of novel loci for smoking-related behaviors (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 (e.g., cardiovascular health,[14] cancer risk,[14] bone mineral density[15]), numerous other substance use disorders (e.g., alcohol,[14] cannabis[16] and opioid use disorders[17]), and psychiatric and related conditions (e.g., major depressive disorder,[18] suicide-related behaviors,[19] loneliness[20]).

While these studies have been immensely successful, they have not focused on TUD itself. 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 major 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 prior studies have shown that the genetic architecture of substance use is largely 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<12K).

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 utility of TUD phecodes for use in large-scale GWAS to boost power and improve our ability to identify novel loci for TUD.[29–31] To address this question, we performed a multi-ancestral meta-analysis of TUD comprising 898,680 individuals of European (**EUR**), African American (**AA**) and Latin American (**LA**) ancestry recruited from multiple 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 new potential medications for treating this serious psychiatric condition.

Results

Cohort Descriptions and Phenotype Validation. We included individuals from eight cohorts across five different sites (**Figure 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 diagnosis. 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 minor proportion (<2%) of 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 on age and sex but reported much lower prevalences of other substance and psychiatric disorders than cases. Thus, almost all TUD cases have evidence of being either former or current smokers based on available self-report data.

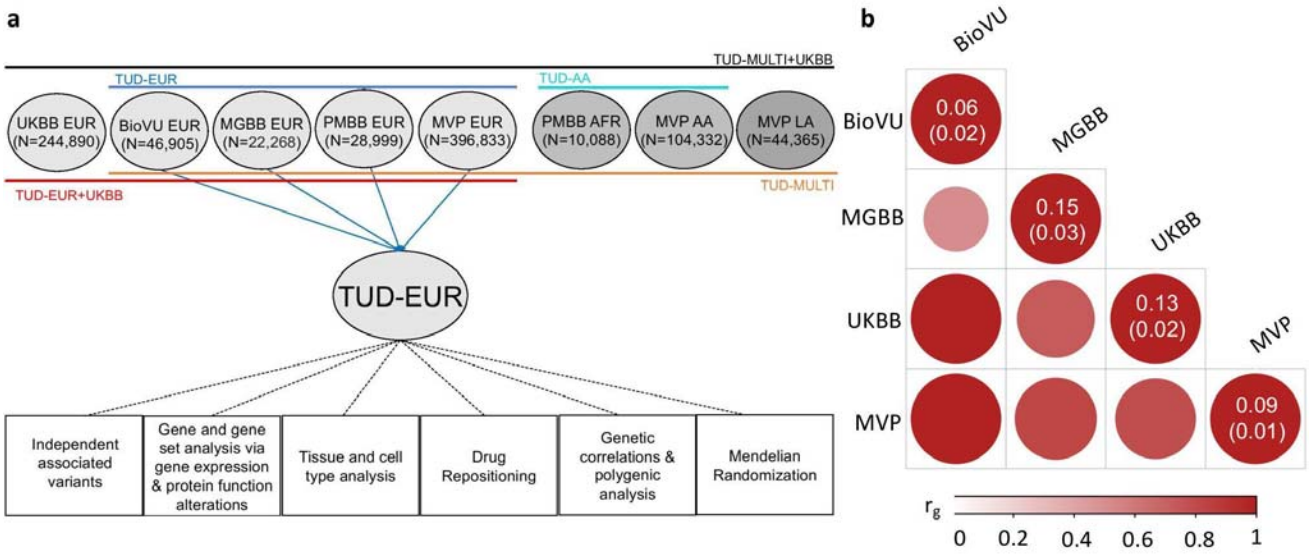

Figure 1. Overview of the cohorts and analysis pipeline (a) and genetic correlations among the sites (b). (**a**) We conducted independent GWAS of TUD cases and controls in individuals of European (EUR) ancestry across four PsycheMERGE sites (BioVU, MGBB, PMBB, and MVP) and performed a GWAS meta-analysis ("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 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 for TUD between all different EUR sites were positive and high, ranging from 0.51 to unity. LDSC genetic correlation for TUD across the two AA sites was strongly positive (0.86) but not significant (p=0.38). We do not report r_g between PMBB and other sites, because the h^2_{SNP} of TUD in PMBB was not significant (h^2_{SNP} =0.90, SE=1.30). LDSC SNP-heritability estimates (h^2_{SNP} 6-15%) are shown in the diagonal. UKBB=UK Biobank, BioVU=Vanderbilt University Medical Center's biobank, MGBB=Mass General Brigham Biobank, PMBB=Penn Medicine Biobank, MVP=Million Veteran Program.

Significant 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 (SNP-heritability, h^2_{SNP}) to be ~6-15% (based on liability scale, assuming a lifetime risk of 12.5%; **Figure 1b**, **Supplementary Table 6**), which is consistent with prior nicotine-related GWAS.[13,27] Genetic correlations across sites and ancestries were high and positive (r_g=0.51-1.24, p<1.80E-02, EUR sites; r_g=0.86, p=0.38, AA sites; cross-ancestry r_{gs}=0.72-0.84, p<7.80E-04; **Figure 1b**, **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 implicate biological underpinnings of TUD. The primary multi-ancestry meta-analysis of 29,448,768 imputed SNPs (lambda λ_{GC}=1.107, **Figure 2**) was performed on seven cohorts, comprising 653,790 individuals, with 75.71% EUR, 17.50% AA, and 6.79% LA.

We identified 97 GWS (*p*<5.00E-08) lead SNPs (r^2<0.1) located in 72 independent loci (**Supplementary Table 7**). All genome-wide significant loci had been previously reported by prior smoking GWAS (**Supplementary Table 7**), including aspects of smoking initiation (7/72), consumption (22/72), cessation (42/72) and nicotine dependence (1/72; **Supplementary Figure 1**). 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.

Our analyses provide corroborative support for nicotinic acetylcholine receptor genes as risk genes for smoking-related traits: *CHRNA5* (rs576982, *p*=1.60E-17, chr. 15; this region includes rs16969968, a well-established functional missense polymorphism [D398N] in *CHRNA5*, *p*=4.93E-11), *CHRNB2 (*rs45490696, *p*=7.61E-09, chr. 1), *CHRNA2* (rs2741339, *p*=2.86E-20, chr. 8), and *CHRNA4* (rs2273500, *p*=7.34E-22, chr. 20). Second, we identified associations with variants in several genes that modulate dopaminergic transmission, such as the dopamine receptor D2 (*DRD2*: genomic position 113334227, *p*=1.04E-11, and rs4936277, *p*=1.81E-09, chr.11), known for its relationship with dopamine and reward,[34] previously associated to nicotine dependence[35] and implicated in a recent large-scale GWAS of addiction;[36] dopamine beta-hydroxylase (*DBH*: rs2007153, 2.55E-16, and rs2519155, *p*=8.74E-12, chr.9), which encodes an enzyme necessary to convert dopamine to norepinephine and has been consistently implicated in smoking behaviors;[13,37] lysine demethylase 4A (*KDM4A:*

rs489319, *p*=1.47E-10, chr. 1), previously found to interact with dopaminergic agents and implicated in problematic opioid use;[38] phosphodiesterase 4B (*PDE4B*: rs7528604, *p*=5.68E-10, chr. 1), which has regulatory effects on dopaminergic pathways and has been implicated in GWAS of externalizing behaviors,[39] smoking initiation,[37,40] and general liability for addiction;[36] and neural cell adhesion molecule 1, *NCAM1* (rs4144892, *p*=5.44E-12, chr. 11), which modulates dopamine signaling[41] and has been associated with several smoking-related traits.[35,37] We also identified an association with a deleterious (CADD=23.1)[42] coding SNP (rs61738568, *p*=2.08E-08, chr. 16) in the *FBRS* gene, recently implicated in smoking initiation.[13]

Furthermore, we identified variants in *GRM8* (Glutamate Metabotropic Receptor 8; rs2157752, *p*=2.79E-08, chr.7), important for mediating reward-related learning and memory, and in *BDNF* (rs6265, *p*=7.18E-09, chr. 11), a candidate gene in genetic studies of substance use disorders given its role in synaptogenesis and memory. None of the lead SNPs showed evidence of heterogeneity across cohorts, based on the I^2 index (**Supplementary Figure 2**). Combining these data with UKBB (which uses a less stringent TUD definition, TUD-multi+UKBB) yielded very similar results (i.e., comparable number of lead SNPs, with an addition of three independent loci: *GALNT10**rs11952152, *PXDNL**rs4873592 and *snoU13**rs830432; **Supplementary Table 8**).

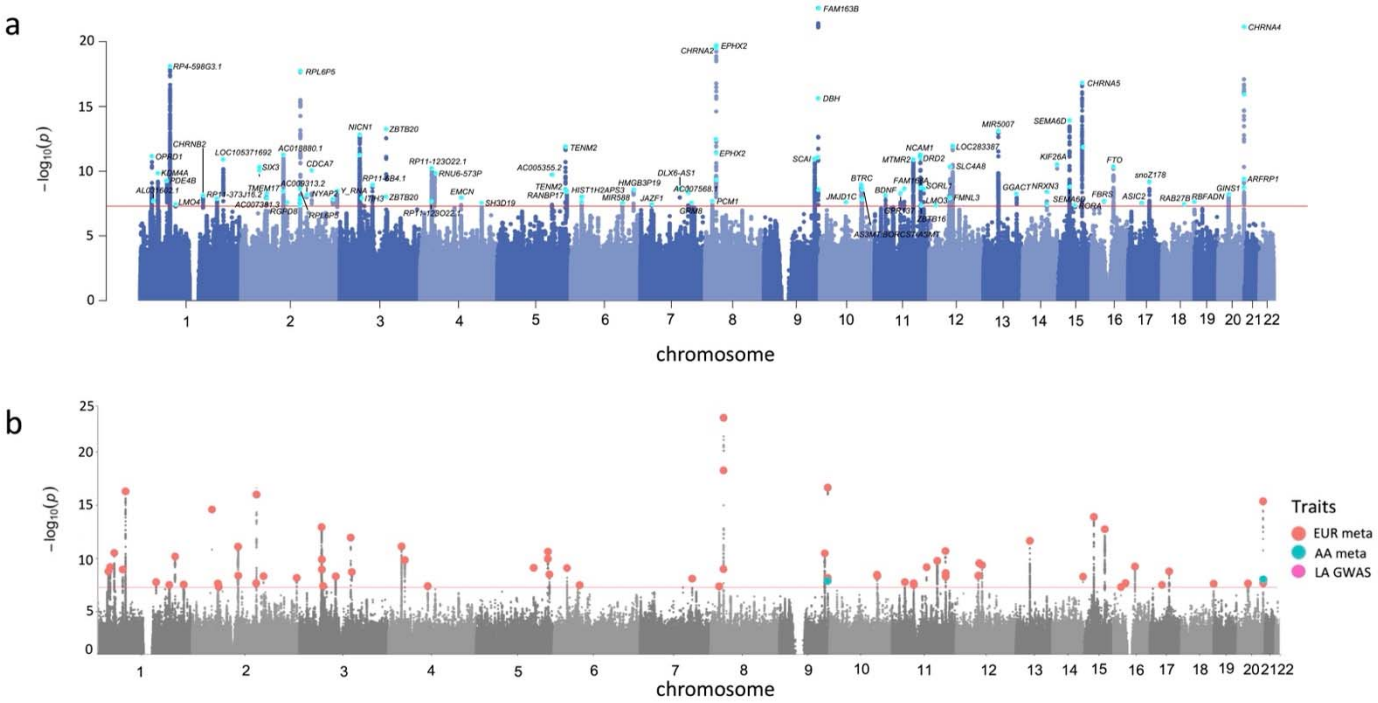

Figure 2. Manhattan and porcupine plots for the TUD-multi meta-analysis and ancestry-specific GWAS. (**a**) TUD-multi identified 72 independent risk loci, all of which were recently replicated by the GSCAN study. (**b**) Porcupine plot of ancestry-specific meta-analyses identified 55 loci in the European cohort (EUR, in red), and 2 loci in the African cohort (AA, in blue). No significant associations were detected in the Latin American (LA) cohort.

Within-ancestry meta-analyses identify ancestry-specific loci associated with TUD. We conducted within-ancestry meta-analyses of EUR (TUD-EUR) and AA (TUD-AA) using an sample-size weighted fixed effects model, and a GWAS of LA (TUD-LA).

TUD-EUR included 19,096,380 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 (LDSC intercept 1.052, SE=0.012) and we did not

identify evidence of heterogeneity (I^2) across the cohorts (**Supplementary Figure 3**). The TUD-EUR meta-analysis yielded a significant h^2_{SNP} estimate of 7.10% (SE=0.003, **Supplementary Table 9**), and identified 68 GWS significant lead SNPs located in 55 independent loci (**Figure 2B**; **Supplementary Table 10**). Ten of these loci were ancestry specific in EUR and not GWS in the multi-ancestry GWAS. Among the 55 independent loci, 8 were fine-mapped to a credible set (posterior inclusion probability > 0.50), of which 6 harbored known protein coding genes (*ZBTB20, HIST1H2BH, BDNF, SLC4A8, KIF26A, ASIC2*; **Supplementary Table 11**).

Again, combining these data with those of UKBB in a secondary GWAS (TUD-EUR+UKBB) yielded very similar results (e.g., similar h^2_{SNP} estimate of 7.00%; with the addition of three independent loci - *LOC105373664**rs6430094, *GALNT10**rs7737824, and *CHRNA4**rs6011779, **Supplementary Table 12**). Considering the similarity in the number of loci identified 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 11.30% (SE=0.015, **Supplementary Table 9**), and 2 independent loci (**Supplementary Table 13**), one on chr. 9 (rs2007153, p=1.17E-08) in *DBH*, which is novel for the AA population, and another on chr. 20 (rs6011779, p=9.27E-09) in the *CHRNA4* gene, replicating a finding from a prior multi-ancestral (EUR+AA) GWAS of smoking.[27] Multi-ancestry fine-mapping analyses using PAINTOR corroborated the region in chr. 9, identifying two putative causal variants in this locus (**Supplementary Table 14**). The TUD-LA GWAS yielded a significant h^2_{SNP} estimate of 8.10% (SE=0.017, **Supplementary Table 9**) but did not identify any GWS loci (**Figure 2**), presumably due to the smaller sample size.

Integration with functional genomic data implicates hundreds of novel TUD candidate risk genes. To further our biological interpretation of the TUD-EUR GWAS results and prioritize causal genes and proteins, we performed multiple *in silico* downstream analyses using MAGMA,[43,44] H-MAGMA,[45] S-MultiXcan/S-PrediXcan,[46] TWAS,[47] and PWAS.[47]

First, we conducted gene-based analyses via MAGMA,[43,44] which mapped SNP-level associations to 86 significant genes (p<2.63E-06), 83 (90.69%) of which replicated genes near or in GWS loci (e.g., *CHRNA3*, *CHRNA4*, *CHRNA5*, *BDNF*, *PTPRF, KDM4A*, *DBH*; **Supplementary Table 15**).

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).[45] These analyses identified 746 unique gene-tissue pairs associated with TUD (p<9.44E-07), a significant proportion of which showed cell-type (16.49% cortical neurons, 16.75% iPSC derived neurons, 20.78% midbrain dopaminergic neurons, 13.00% iPSC derived astrocytes) or developmental stage (15.55% fetal, 17.43% adult) specific expression (**Supplementary Table 16**).

Using S-MultiXcan to predict the effect of common SNP variation on gene expression in multiple brain tissues, we detected significant associations for 34 genes (**Supplementary Table 17**), with effects dispersed across 12 brain regions (cerebellum, anterior cingulate cortex, basal ganglia [nucleus accumbens and putamen], cortex and frontal cortex, amygdala, hypothalamus, substantia nigra, spinal cord, cerebellar hemisphere, spinal cord). Inspection of region-specific results via S-PrediXcan identified five genes that were consistently upregulated (*GPX1*, *PPP6C*, *GMPPB*, *WDR6*) or downregulated (*CHRNA2*) 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 (DLPFC) by performing TWAS (mRNA and splicing) and PWAS, respectively. Associations across these three regulatory models identified 43 TUD unique risk genes (34, mRNA expression; 15, splicing expression; 14, proteome expression; **Supplementary Tables 19** and **20**). Colocalization analysis identified five genes and proteins (NT5C2, GPX1, NEK4, ABHD12, RHCE) associated with TUD via their regulation of brain expression levels and protein abundance (PP4 >0.80, **Supplementary Table 21**, **Supplementary Figure 4**).

Overall, after controlling for multiple comparisons, these analyses identified 330 unique genes with statistical evidence of association with TUD (**Figure 3a**, **Supplementary Table 22**). Of these, 87 converged across at least 2 methods, and 3 (*GPX1*, *P4HTM* and *RHCE*) converged across all six methods. 304 (92.12%) of the 330 genes identified via these analyses were not identified by the GWS loci; 75 (22.72%) were novel TUD genes not identified in prior FTND or GSCAN analyses (e.g., other genes from the KDMA family [*KDM4D*, *KDM4F*, *KDM4E*], *SLC9A2, NFKB2*), which prompt novel hypotheses to be tested experimentally.

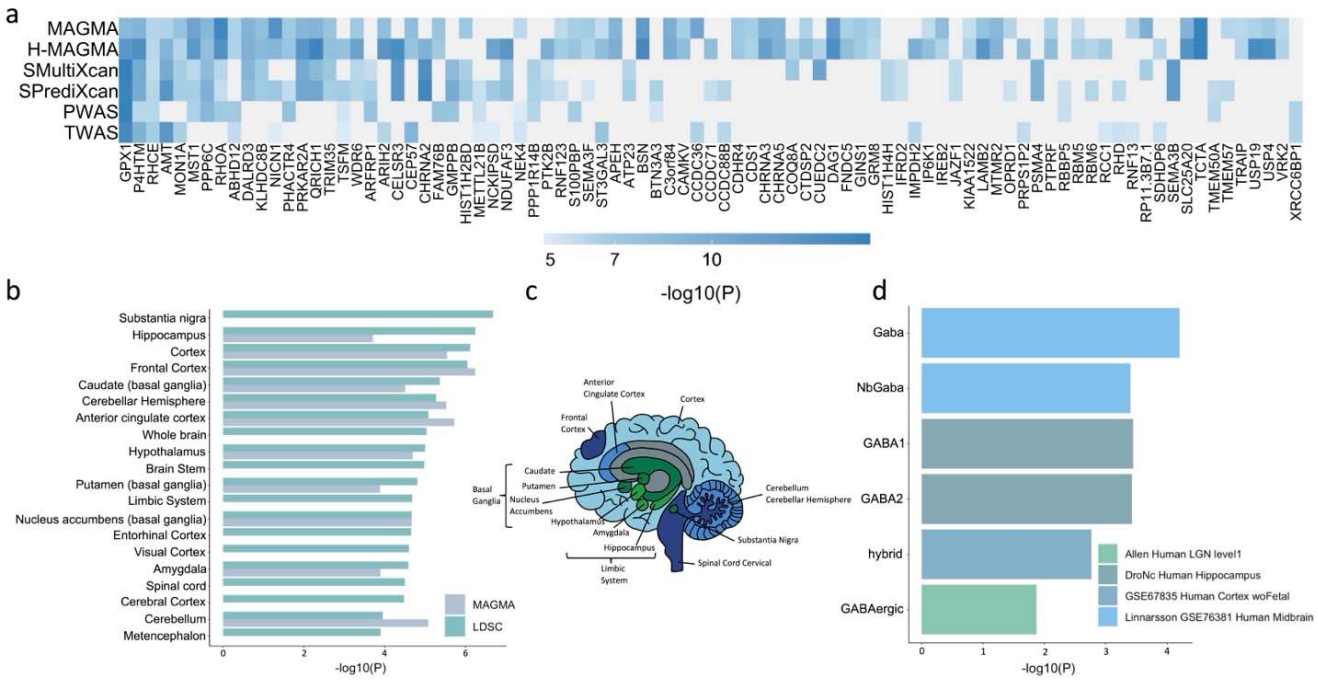

Figure 3. Integration with functional genomic data implicated 330 unique TUD candidate risk genes. (**a**) Of 330 associated genes, 87 converged with at least 2 methods, and were dispersed throughout the chromosomes. (**b**) LDSC (SNP-based) and MAGMA tissue-specific gene expression of TUD risk genes reveals substantial brain enrichment. (**c**) The genetic findings across multiple levels of analysis (LDSC, MAGMA, MultiXcan, 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 single-cell RNA-sequencing datasets (**Supplementary Table 28** for full list). After multiple testing correction for all datasets, only genes expressed in GABAergic neurons were associated with TUD (**Supplementary Table 28**).

Tissue and cell-type analyses of TUD identify enrichment in brain tissue and GABAergic neurons. 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 Figure 5** and **Supplementary Table 23** for full list).

Tissue enrichment analyses in MAGMA use gene expression data from GTEx (v8). In addition to non-brain tissues (i.e., cardiovascular, hematopoietic, adrenal pancreas, and other, p<3.37E-05, **Supplementary Table 24**), we detected significant enrichment mostly in the brain (p=1.53E-15), spanning multiple brain regions, including the hippocampus, the limbic system, frontal cortex (**Supplementary Tables 25-26**, **Figure 3b-c**), most of which were also implicated in S-MultiXcan (**Supplementary Table 17**). Correlating the effects of SNP variation with brain imaging traits via BrainXcan identified similar results, including significant (p<1.92E-04) associations with decreased gray 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 (sc-RNA seq) datasets. We identified a significant association (p<0.05) between TUD risk and cell-type specific gene expression in GABAergic neurons for individual human sc-RNA seq datasets (Linnarsson, midbrain, p<3.94E-04; Allen Brain Atlas, dorsal lateral geniculate nucleus, p=1.34E-02; DroNc-seq, hippocampus, p<3.74E-04; **Figure 3d**; **Supplementary Table 28**). These results did not survive conditional analyses within and across datasets.

Implications for TUD biology based on gene-set and pathway analyses. We used MAGMA[43,44] to conduct a gene-wise TUD analysis and to test for enrichment of pathways curated from multiple sources. After correcting for multiple comparisons, 25 related pathways and biological processes were significantly enriched for genes associated with TUD (p<2.76E-06; **Supplementary Table 29**). Associations implicated fundamental processes related to

nicotine response (e.g., high calcium and sodium permeable nicotinic acetylcholine receptors, *p*=4.66E-16; behavioral response to nicotine, *p*=5.97E-16), regulation of postsynaptic signaling (*p*=2.61E-08), and maintenance of synapse structure (*p*=9.26E-07), 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 293 FDA approved medications approved by the U. S. Food and Drug Administration (**Supplementary Table 30**). 31 of the 293 identified medications targeted at least one mapped/independent gene from our GWAS. The medications that significantly reversed (Bonferroni *p*<6.03E-05) the transcriptional profile associated with TUD included varenicline (a well-known therapeutic for smoking cessation), sodium channel blockers (e.g., amiloride), and compounds that are used to treat conditions that commonly co-occur with TUD, such as antipsychotics (e.g., clozapine), dopaminergic agents (e.g., ropinirole), opioids (e.g., nalbuphine), and antidepressants (e.g., amoxapine), among others (**Figure 4**).

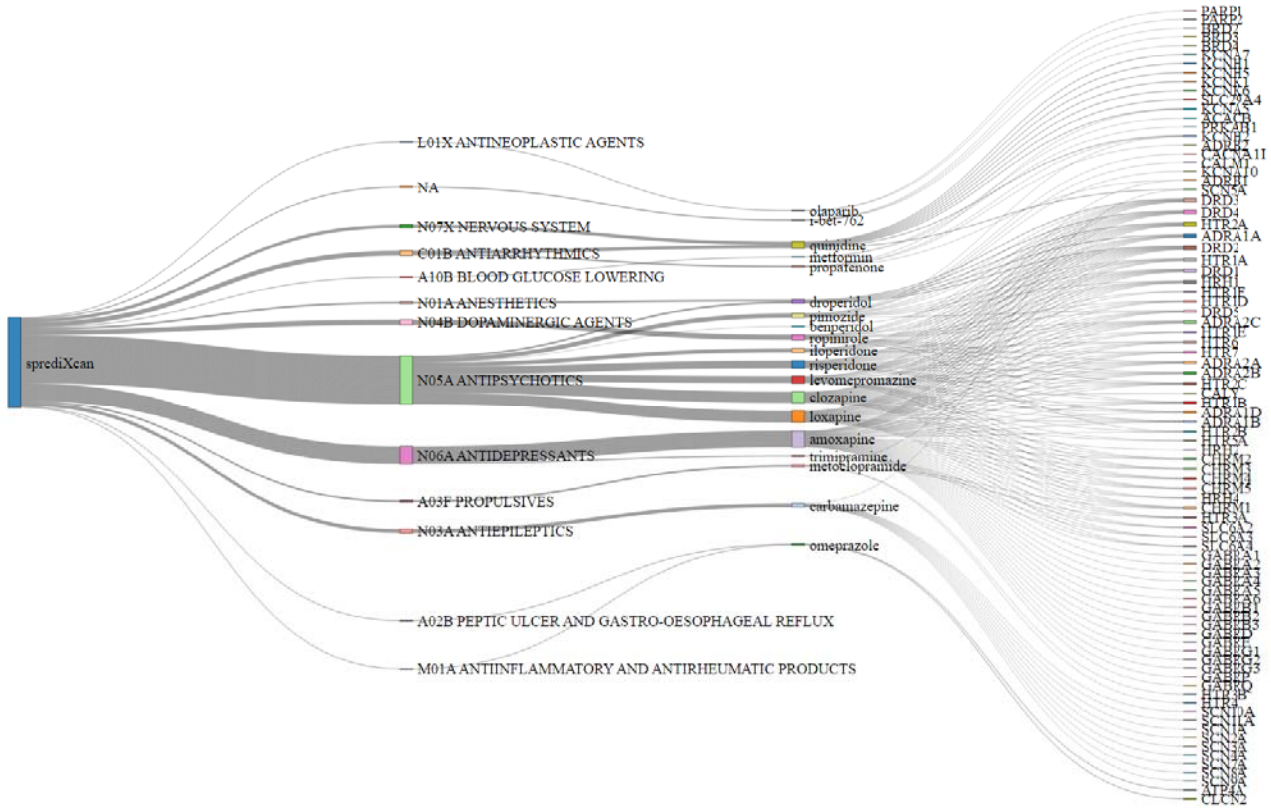

Figure 4. Sankey Diagram showing Drug Repurposing results from S-PrediXcan Brain Tissues. 293 medications/perturbagens grouped by ATC category membership from the Library of Integrated Network-Based Cellular Signatures (LINCS)[36] database. ATC category connected to perturbagen edges represent corresponding ATC category membership. Perturbagen connected to gene target edges are associated with the reversal of the TUD transcriptomic profile from S-PrediXcan Brain Tissue results.

Genetic correlation with other traits. We estimated pairwise r_g with TUD for 115 published phenotypes using LDSC.[48] TUD showed FDR-significant correlations r_g with 85 traits (**Figure 5**; **Supplementary Table 31**). As expected, the strongest positive correlations were with smoking-related traits (e.g., smoking initiation r_g=0.79, SE=0.02; smoking cessation r_g=0.66, SE=0.03; cigarettes per day r_g=0.43, SE=0.03; FTND r_g=0.61, SE=0.06; **Figure 5a**) and other substance use traits (e.g., cannabis use disorder r_g=0.63, SE=0.04; drinks per week r_g=0.36, SE=0.03;

opioid use disorder (**OUD**) r_g=0.44, SE=0.07). TUD clustered with addiction traits rather than consumption phenotypes (**Supplementary Figure 6**).

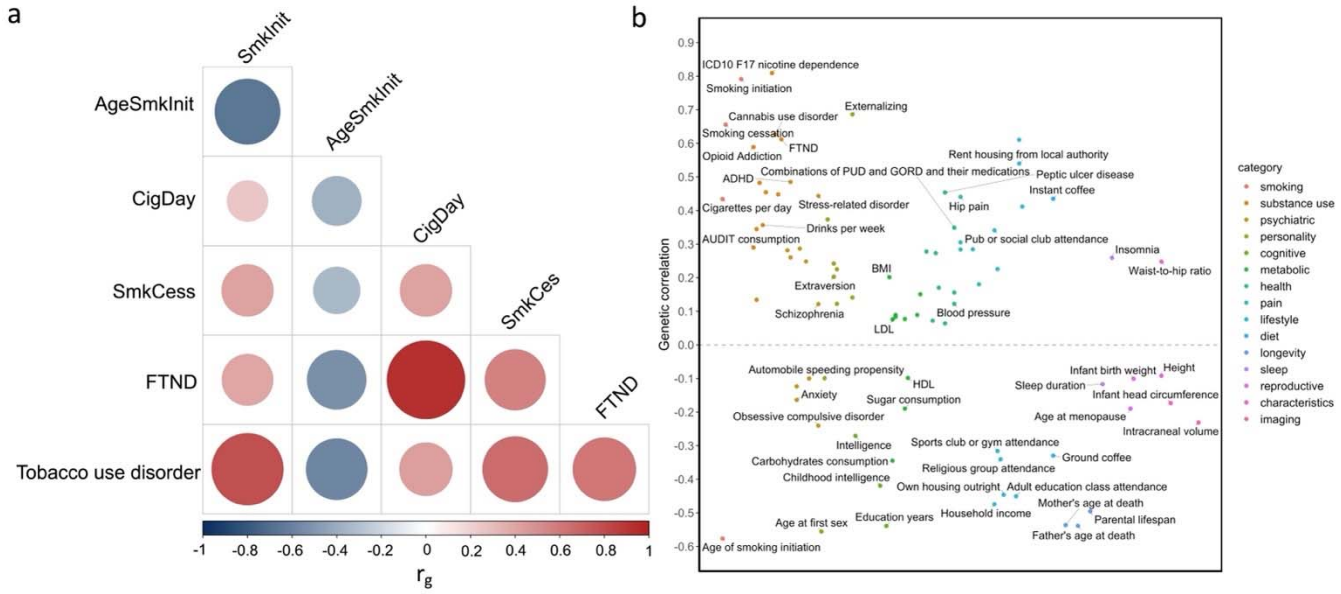

Figure 5. FDR-significant genetic correlations between TUD-EUR and 115 complex traits, including smoking and related phenotypes (b). (**a**) Genetic correlations (r_g) between age of smoking initiation (AgeSmkInit), cigarettes per day (CigDay), smoking cessation (SmkCess), nicotine dependence via the Fagerström Test for Nicotine Dependence (FTND), and tobacco use disorder (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 d below the line. All r_gs are significant using a 5% FDR correction for multiple testing. *AgeSmkInit*, age of smoking initiation smoking; *CigDay*, cigarettes smoked per day; *SmkCess*, smoking cessation;[13] *FTND*, Fagerstrom Test for Nicotine Dependence.[27]

Tud was also genetically associated with psychiatric and medical conditions (**Figure 5b**, **Supplementary Table 31**). There were significant positive r_g with psychiatric traits (e.g., externalizing r_g=0.69, SE=0.02; ADHD r_g=0.49, SE=0.04; stress-related disorder r_g=0.44, SE=0.04) and risky behavioral traits, including lower age of first sex (r_g=-0.56, SE=0.02). We also found positive r_g with health outcomes (e.g., coronary artery disease r_g=0.27, SE=0.03;

waist-to-hip ratio r_g=0.25, SE=0.02; hip pain r_g=0.44, SE=0.04; knee pain r_g=0.31, SE=0.04) and several social determinants of health, such as the Townsend deprivation index (r_g=0.61, SE=0.07). There were negative r_g with socioeconomic variables, including years of education (r_g=-0.54, SE=0.02) and household income (r_g=-0.47, SE=0.04) and with childhood intelligence (r_g=-0.42, SE=0.07). Conditioning on alcohol, cannabis, or opioid use disorders 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 cigarettes per day (**Supplementary Figure 7**) and FTND (**Supplementary Figure 8**).

Among AA datasets, there were significant r_g with smoking trajectories and other substance use traits (OUD r_g=0.42, SE=0.06; maximum habitual alcohol consumption r_g=0.78, SE=0.2). Nominal associations (p<0.05) were observed for smoking initiation (r_g=0.35, SE=0.13), depression (r_g=0.42, SE=0.2) and type 2 diabetes (r_g=-0.23, SE=0.1; **Supplementary Table 33**).

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, Mayo Clinic. As expected, TUD PGS was strongly associated with TUD-EHR (p=1.60E-88, **Supplementary Table 34**, **Figure 6a**), explaining 6.3% of the (Nagelkerke's R^2) variance. Additional significant (p<7.25E-05) associations included other substance use disorders (e.g., alcohol-related disorders, OR=1.27, p=1.53E-17), medical conditions strongly associated with TUD (e.g., chronic airway obstruction, OR=1.19, p=5.73E-21) and other psychiatric conditions (e.g., depression, OR=1.06, p=8.60E-06). These remained significant

after accounting for TUD diagnosis (**Supplementary Table 34**). We also noted associations across multiple other medical categories, including endocrine/metabolic (e.g., morbid obesity, OR=1.09, *p*=1.06E-07; type 2 diabetes, OR=1.06, *p*=1.78E-05), digestive (e.g., diseases of esophagus, OR=1.05, *p*=3.45E-06), circulatory (e.g., ischemic heart disease, OR=1.07, *p*=7.43E-07) and neurologic (e.g., pain, OR=1.07, *p*=9.95E-07), among others (**Supplementary Table 34**). Compared to FTND PGS, TUD PGS were more strongly associated across virtually all domains, including TUD (**Supplementary Figure 9**). 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 and most significant (OR=1.19, *p*=4.03E-06) association (**Supplementary Table 35**).

Yale-Penn sample. We next extended the analyses to a deeply characterized sample recruited for genetic studies of substance use disorders: the Yale-Penn sample.[49] 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 many substance use traits, including nicotine dependence as defined via a Diagnostic and Statistical Manual of Mental Disorders (**DSM**) diagnosis in both the EUR (OR=1.71, *p*=2.51E-41; **Figure 6b**; **Supplementary Table 36**) and AA cohorts (OR=1.12, *p*=8.10E-04), respectively, although the latter association did not survive multiple testing correction (**Supplementary Table 37**). 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 vs 0.062; **Supplementary Table 38**, **Supplementary Figure 10**), again emphasizing the value of collecting information on later stages of vulnerability or more severe phenotypes, such as TUD.

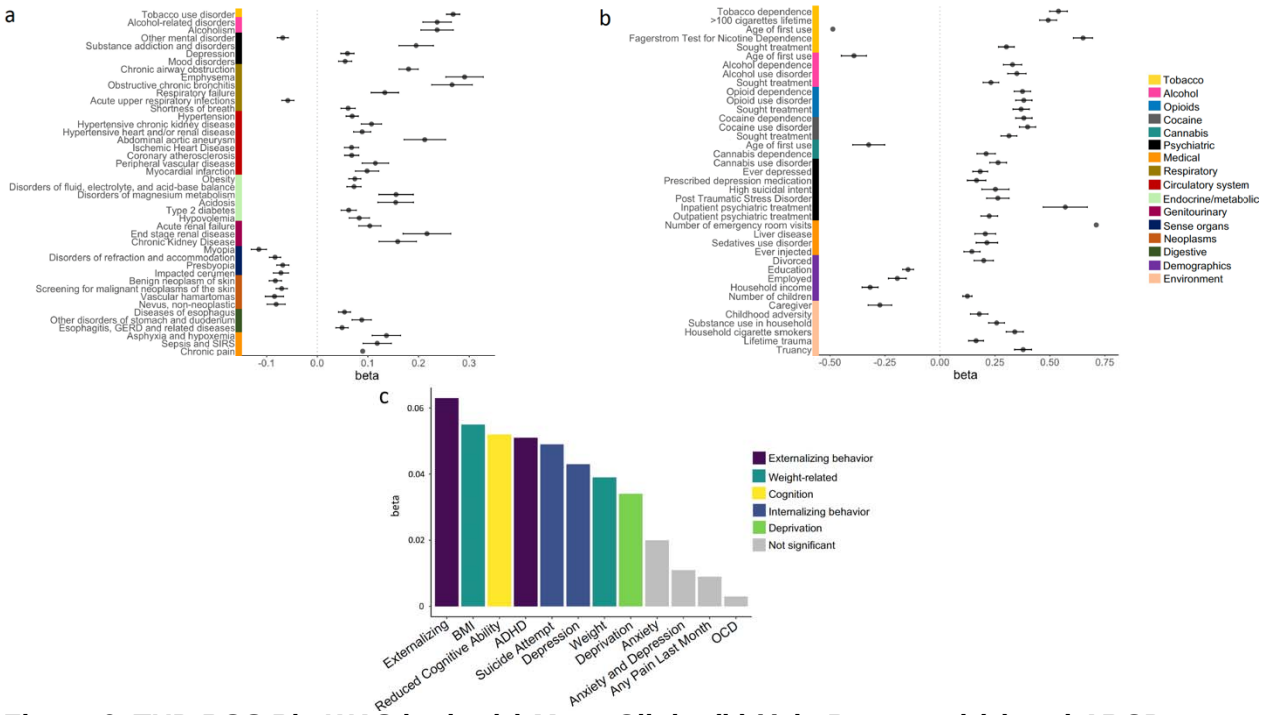

Figure 6. TUD PGS PheWAS in the (a) Mayo Clinic, (b) Yale-Penn, and (c) and ABCD European cohorts. Only selected FDR-significant traits are shown. The exact values for each association and extended lists of traits can be found in **Supplementary Tables 34, 36 and 39**.

Adolescent Brain Cognitive Development (ABCD) cohort. Lastly, we extended our polygenic analyses to a drug naïve developmental sample (9-11 years of age at recruitment; analytic N=52 to 5,556). We concentrated on 12 traits that showed significant genetic correlations in the adult samples (**Supplementary Table 39**, **Figure 6c**). Although tobacco exposure was uncommon in this pediatric population (2.30% prevalence), externalizing behaviors, which emerge in childhood and are strong correlates of substance use, were available. After correcting for multiple testing, TUD PGS was significantly (p<4.00E-03) associated with externalizing behaviors (i.e., Child Behavior Check List [**CBCL**] externalizing scores, β=0.06, p=3.15E-06; CBCL ADHD scores, β=0.05, p=1.58E-04), as well as internalizing (i.e., suicide attempt, β=0.05, p=7.01E-04, CBCL depression scores, β=0.04, p=1.73E-03), cognitive ability (β=0.003, p=6.35E-05), neighborhood deprivation (β=0.03, p=2.53E-03), and weight-related phenotypes

(i.e., BMI, β=0.06, *p*=4.44E-05; weight, β=0.04, *p*=2.98E-03). 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.

Causal relationships with TUD and bi-directional effects of TUD with other traits. We used MR analyses to test directional causal relationships between significantly genetically correlated traits (N=6) and TUD among EURs only due to the small sample size and limited statistical power in other populations (**Supplementary Table 40**). We observed a significant positive bidirectional causal effect between TUD and depression. TUD had a significant negative causal effect on educational attainment, drinks per week and ADHD.

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 multiple associations with tobacco use, but findings for tobacco dependence or disorder have been limited due to the difficulty of characterizing large numbers of individuals using a gold-standard research or clinical diagnosis. Here we present the first multi-ancestry GWAS of TUD using data from EHR, as a complementary strategy for ascertainment. In less than four months, and leveraging data from the PsycheMERGE consortium, we gathered TUD-EHR data for 898,680 individuals. 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-EHR was genetically correlated with traits derived from traditionally ascertained cohorts, including nicotine dependence via FTND and smoking cessation, providing clear evidence that the signal captured by TUD phecodes is valid. Of note, the genetic correlation between TUD-EHR and other smoking behaviors, such as number of cigarettes smoked per day (**CPD**), although significant and positive, was moderate in magnitude (r_g=0.43), suggesting that the genetic architectures of consumption and misuse may be distinct. This is in contrast to earlier observations for FTND and CPD, where 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 correlated (r_g=0.61), 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 time-to-first cigarette in the morning, which was more strongly associated with FTND PGS, likely because time-to-first cigarette is one of the FTND items. Overall, this emphasizes the need to continue measuring the full spectrum of addiction liability,[50] such as CPD, and 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 (7%) of the overall heritability of TUD (40-60%) as determined by prior family and twin studies.[9,11] The multi-ancestral meta-analysis identified 72 independent loci, 13 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*, *CHRNA4*), which have been consistently associated with smoking behaviors,[20] particularly in studies of self-reported CPD.[13] We also identified polymorphisms in genes implicated in nicotine clearance, like *CYP2A6*, previously linked to heavy smoking.[51] Other variants identified are in genes that modulate dopaminergic and glutamatergic neurotransmission, compromising reward-based learning and facilitating

drug-seeking behavior, and in *BDNF*, which is involved in memory consolidation processes,[52] and a well-studied candidate gene in addiction.[53] These and other candidates supported by TUD (e.g., *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 multiple 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 (e.g., ebselen), which restore GPX1 activity in situations of extreme oxidative stress, can protect from lung inflammation induced by smoking.[54] Another was *GMPPB*, which has been associated with accelerated lung aging and e-cigarette smoking.[55] *NT5C2* is involved in maintaining cellular nucleotide balance, and was associated with schizophrenia[55] and smoking behaviors in an exome-wide association study.[56] These genes showed a consistent causal effect based on colocalization analyses (here and previously[57]), 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.[58] 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.[59]

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,[49] which has comprehensive phenotypic data for substance use disorders. 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, like respiratory conditions, likely reflect chronic smoking. The positive associations between genetic liability for TUD and other outcomes, such as BMI or 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 *CHRNA7*, *CHRNA5*, *CHRNA4*, and *CHRNB2*, might moderate the effect of varenicline, a smoking cessation treatment that operates as a partial agonist at the nicotine acetylcholine a2b4 receptor,[60] varenicline (along with other medications such as nicotine replacement therapies) has limited efficacy or adverse effects.[61,62] In a proof-of-principle study, So et al.[63] 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 reuptake inhibitors (e.g., amoxapine) and antipsychotics (e.g., clozapine), pointing to convergent molecular mechanisms between TUD and other psychiatric disorders that are the usual target of these agents, replicating prior observations.[64] The potential therapeutic utility of anti-inflammatory or blood glucose lowering medications were also suggested by our analyses. Although, to date, no repurposed drugs have been developed for treating SUDs based on 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 limiations of our study. Prior work has demonstrated that ICD codes have a low sensitivity for current tobacco use, but may have a reasonable specificity for this common behavior.[65] Our results appeared to be robust to moderate levels of misclassification, particularly in controls, as detected by the pairing with self-reported questionnaire data. Although studies that systematically evaluate the effect of removing potentially missclassified 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 (e.g., clinical notes), may help capture additional phenotypes, including the response to treatment or the ability to successfully quit smoking without formal treatment. Longitudinal data from EHR, with data collection spanning the period prior 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 QTL databases in more specific cell types, and the inclusion of more ancestrally diverse samples 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. Lastly, other forms of genetic variation, such as rare single variants[66] or structural polymorphisms[67] are likely to account for much of the “missing heritability” in genetic risk for TUD.

In sum, this work demonstrates that EHR is a viable and cost efficient complementary alternative 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 one among a number of highly correlated psychiatric and medical disorders. We anticipate that

these results can be combined with prior smoking GWAS in larger multivariate analyses to elucidate the full spectrum of smoking behaviors and accelerate gene discovery for TUD.

Methods

Smoking phenotypes and cohorts. We defined cases as patients who received at least two TUD ICD-9 or -10 codes (corresponding to the phecode definition) in their medical records, and controls as patients who had no TUD diagnosis code (**Supplementary Table 2**). In UKBB only, cases were defined as having 1 ICD-10 code for TUD, and controls had none.[43] Additionally, we required controls to be 18 years of age or older at time of analysis (04/2022). 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 survey when available (**Supplementary Table 3,** list of smoking traits).

Our data sources included registries from five health systems linked to biobanks: Vanderbilt University Medical Center's (VUMC) biobank (BioVU), Mass General Brigham Biobank (MGBB), Penn Medicine BioBank (PMBB), Million Veteran Program (MVP), and UK Biobank (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, 44,365 LA), and 244,890 participants from UKBB. Details of each registry, including demographics and data sources, are listed in the **Supplementary Table 2**.

Genotyping, imputation, and GWAS. For all cohorts, the initial GWAS analyses were conducted within genetic ancestral groups. GWAS analyses were performed within each ancestral group using SAIGE version 0.44.6.5[68] or PLINK 2.0[69] and a logistic regression. For the BioVU, MGBB, and UKBB cohorts, there were GWAS for only the European ancestral group (**Supplementary Material**). In PMBB, we conducted additional GWAS of the African ancestral

group sample, and in MVP we performed additional GWAS of the African American ancestral sample and the Latin American ancestral group sample. Each of the univariate GWAS covaried for 10 genetic ancestry principal components, 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).

BioVU. We used de-identified clinical data from individuals in BioVU. Genotype data were generated using the Illumina Multi-Ethnic Genotype Array (MEGAEX) for 72,824 individuals. Details on the quality control process have been described elsewhere.[70] Genotypes were filtered for SNP (<0.95) and individual (<0.98) call rates, sex discrepancies, and excessive heterozygosity (|Fhet|>0.2).[71] The sample was then filtered for cryptic relatedness by removing one individual of each pair for which pihat>0.2. PCA using FlashPCA2 combined with CEU, YRI and CHB reference sets from 1000 Genomes Project Phase 3[72] was conducted to determine European Ancestry. We confirmed the absence of genotyping batch effects using 'batch' as the phenotype. We imputed genotypes using the Michigan Imputation Server with the reference panel from the Haplotype Reference Consortium. SNPs were filtered for imputation quality (R^2 >0.3 or INFO >0.95) and converted to hard calls. We restricted the analyses to autosomal SNPs with minor allele frequency >0.01. We removed SNPs that differed by >10% from the 1000 Genomes Project phase 3 CEU set[72] and those with a Hardy Weinberg Equilibrium p<1.00E-10. The resulting data set contained hard-called SNP information for 9,386,383 SNPs in 72,824 individuals of European Ancestry. Controls were also required to have 3 or more years of medical history with VUMC. These procedures resulted in a total sample of 7,167 cases and 39,738 controls in BioVU. The project was approved by the VUMC Institutional Review Board (IRB #160302, #172020, #190418).

MGBB. MGBB samples were genotyped using the Illumina Multi-Ethnic Global array with hg19 coordinates. Variant-level quality control filters were applied to remove variants with a call

rate <0.98, and those that were duplicated across batches, monomorphic, not confidently mapped to a genomic location, or associated with genotyping batch. Sample-level quality control filters were applied to remove individuals with a call rate <0.98, excessive autosomal heterozygosity (±3 standard deviations from the mean), or discrepant self-reported and genetically inferred sex. PCs of ancestry were calculated using the 1000 Genomes Phase 3 dataset as a reference panel. The Michigan Imputation Server was then used to impute missing genotypes with the Haplotype Reference Consortium dataset serving as the reference panel. Imputed genotype dosages were converted to hard-call format and subjected to further quality control, where SNPs were removed if INFO score <0.8, MAF <0.01, HWE p<1.00E-10, or missingness (variant call rate <0.98). Only unrelated individuals (pi-hat <0.2) of European ancestry were included in the present study. These procedures yielded a final analytic sample of 6,708 cases and 15,560 in the MGBB. The project was approved by the MGBB Institutional Review Board (IRB #2018P002642).

PMBB. PMBB samples were genotyped by the GSA genotyping array. Quality control removed SNPs with marker call rate <95% and sample call rate <90%, and individuals with sex discrepancies. Genotype phasing and imputation was performed on the TOPMed Imputation server.[73] The phasing was done using EAGLE (v2.4.1)[30] and imputation was performed using MINIMAC software.[73] IBD analysis was used to check for relatedness among imputed samples using PLINK 1.9. We randomnly removed one individual from each pair of related individuals (pi-hat <0.25). SNPs with an INFO score <0.3, MAF <0.01, a genotype call rate <0.95 or an HWE p<1.00E-6 were removed. To estimate genetic ancestry, PCs were calculated based on common SNPs between PMBB and the 1000 Genomes Project phase3[72] using the smartpca module of Eigensoft package.[74] Participants were assigned to an ancestry based on the distance of 10 PCs from the 1000 Genomes reference populations. The resulting dataset

included 10,088 AAs (cases=1,722) and 28,999 EURs (cases=3,088). The PMBB is approved under IRB protocol #813913.

MVP. MVP samples were genotyped using the Affymetrix Axiom Biobank Array. Samples were removed if they had extreme heterozygosity, call rate <98.5%, sex mismatch, or >7 relatives. SNPs were removed if they had call rate <0.98 or a Hardy–Weinberg equilibrium (HWE) threshold of p<1.00E-06. Genotype phasing and imputation was performed using SHAPEIT4 (v.4.1.3)[75] and Minimac4 software[73], respectively. Biallelic and non-biallelic SNPs were imputed using the African Genome Resources and 1000 Genomes reference panels.[72] Ancestry was defined for three mutually exclusive ancestral groups (European, African American, and Hispanic American) utilizing a previously defined approach harmonizing genetic ancestry and self-identified ancestry (HARE).[76] SNPs with imputation quality (INFO) score□<0.7, minor allele frequency (MAF, AA <□0.005; EUR□<□0.001; HIS□<□0.01), genotype call rate <0.95, and HWE p<1.00E-06 were removed. We also excluded one individual from each pair of related individuals (kinship >0.08, N=31,010). The final sample comprised 104,332 AAs (cases=43,743), 396,833 EURs (cases=146,771) and 44,365 LAs (cases=12,277). The Central VA Institutional Review Board (IRB) and site-specific IRBs approved the MVP study.

SNP-heritability (h^2_{SNP}). We estimated h^2_{SNP} based on the liability-scale (population prevalence estimates of 0.125) for common SNPs mapped to HapMap3[77] using LDSC.[48] For AA and LA, we created in-sample LD scores derived from the MVP genotype data using cov-LDSC.[78]

Meta-analyses and independent variants. Meta-analyses were conducted using a sample-size-weighted method in METAL,[79] 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 four 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 all datasets (AA [PMBB, MVP]; EUR [BioVU, MGBB, PMBB, MVP, UKBB]; HA [MVP]). 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 genome-wide significant (GWS) loci were defined as those with p<5.00E-08 with LD r^2>0.1, within a 1MB 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[43] using a range of 3 Mb, r^2 >0.1, and the respective ancestry 1000 Genome reference panel.[72] Genomic risk loci that were located <1Mb apart were incorporated into a single locus. For loci that harbored multiple variants, we used COJO in GCTA[80] to define independent variants by conditioning them on the most significant variant within each locus. Following conditioning, significant variants (p<5.00E-08) were considered independent.

We determined credible variants among the independent variants by merging risk variants within 1Mb of the lead variant and fine-mapped the resulting region with 95% credible sets using FINEMAP.[81] A posterior inclusion probability (PIP>0.5) was used to denoted causal signals.

Multi-ancestry fine-mapping analyses. We used PAINTOR v3.1[82] 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 larger than 3.9 within a 1Mb 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[72]. 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-EHR GWAS (**Supplementary Methods**). We used the default version (v1.3.6a) of the FUMA web-based platform[43] to identify independent SNPs (r^2<0.10) and to study their functional consequences. We also used MAGMA v1.08[43,44] 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<2.56E−06). Gene sets were obtained from Msigdb v7.0 ("Curated gene sets", "GO terms"). We also used Hi-C coupled MAGMA (H-MAGMA[45]) to assign non-coding (intergenic and intronic) SNPs to genes based on their chromatin interactions. Exonic and promoter SNPs were assigned to genes based on 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<9.55E−07).

S-MultiXcan/S-PrediXcan. We used S-MultiXcan v0.7.0 (an extension of S-PrediXcan v0.6.2[46]) to identify specific eQTL-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 pre-computed tissue weights from the Genotype-Tissue Expression (GTEx) v8 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<3.52E−06).

Partitioning Heritability Enrichment. We used LDSC to partition TUD-EUR h^2_{SNP} and examined the enrichment based on several functional genomic annotation models.[83,84] In the baseline model, we examined 75 overlapping functional annotations comprising genomic, epigenomic and regulatory features. We also analyzed 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 analyzed based on annotations obtained from H3K4me1-imputed, gapped peak data generated by the Roadmap Epigenomics Mapping Consortium.[85] 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 (https://www.nature.com/articles/s41588-018-0081-4). 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[86–88] to identify TUD-relevant tissues. We applied a conservative Bonferroni correction based on the number of tissues simultaneously tested (205 tissues tested, p<2.44E-04). We also used MAGMA v1.08 gene-property analysis of expression data from GTEx (54 tissue types) and BrainSpan (29 brain samples at different age) in FUMA v1.3.6a[75] 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 single-cell RNA sequencing data sets from human brain (data sets listed in the **Supplementary Material**). The method is described in detail in

Watanabe et al.,[43] and uses MAGMA gene-property analysis to test for association between cell specific gene expression and TUD-gene association. Conditional analyses for multiple testing are applied to correct for all tested cell types across datasets.

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 pQTLs from discovery (Banner)[89,90] and validation (ROSMAP)[91,92] datasets using the FUSION pipeline (http://gusevlab.org/projects/fusion/).[47] Next, TWAS was performed using gene and splicing expression profiles measured in the adult DLPFC and gene expression profiles from the frontal cortex. Human brain transcriptome data, used as expression reference panels, were obtained from the CMC[91] and GTEx frontal cortex v7.[47,86] 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[47,93]) was used to determine whether SNPs mediate the association with TUD via effects on gene and protein expression. A posterior colocalization probability (PP) of 80% was used to indicate a shared causal signal.

BrainXcan. We used the BrainXcan package (https://github.com/hakyimlab/brainxcan)[94] 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 MRI images via ridge regression.[94] 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<1.92E-04).

Drug repurposing. Our signature matching technique used data from the Library of Integrated Network-based Cellular Signatures (LINCs) L1000 database. The LINCs L1000 database

738 catalogues in vitro gene expression profiles (signatures) from thousands of compounds in over
739 80 human cell lines (level 5 data from phase I: GSE92742 and phase II: GSE70138). We
740 selected compounds that were currently FDA approved or in clinical trials (via
741 https://clue.io/repurposing#download-data; updated 3/24/20). Our analyses included signatures
742 of 829 chemical compounds (590 FDA approved, 239 in clinical trials) in five neuronal cell-lines
743 (NEU, NPC, MNEU.E, NPC.CAS9 and NPC.TAK), a total of 3,897 signatures.

744 We matched in vitro medication signatures with TUD signatures from brain tissue
745 transcriptome-wide association analyses (conducted using S-PrediXcan). This consisted of
746 Amygdala, Anterior Cingulate Cortex BA24, Caudate Basal Ganglia, Cerebellar Hemisphere,
747 Cerebellum, Cortex, Frontal Cortex BA9, Hippocampus, Hypothalamus, Nucleus Accumbens
748 Basal Ganglia, Putamen Basal Ganglia, Substantia Nigra, and Pituitary brain regions. As
749 previously described,[36] we computed weighted Pearson correlations between transcriptome-
750 wide brain associations and in vitro L1000 compound signatures, weighting each gene by its
751 proportion of heritability explained, using the *metafor* package (version 3.8-1) in R. We treated
752 each L1000 compound as a fixed effect incorporating the effect size (rweighted) and sampling
753 variability (se2r_weighted) from all signatures of a compound (e.g., across all time points, cell
754 lines, doses). Brain region was included as a random effect to account for any tissue specific
755 heterogeneity. Both the genes for the transcriptome wide association analysis input and the
756 medications from our drug repurposing analyses were required to survive a Bonferroni
757 correction for multiple testing (transcriptome-wide correction=0.05/14,199=3.52E-06;
758 Perturbagen correction =0.05/3,897 =1.28E-05).

759 **Genetic correlation analyses.** We estimated the within-ancestry r_g*s* for TUD using LDSC[48] and
760 the cross-ancestry r_g*s* for TUD across population groups using POPCORN.[48] We used the
761 ancestry-specific 1000 Genomes Project phase 3[76] data as the LD references.

We used local LDSC[48] to calculate genetic correlations (r_g) between TUD and 115 other traits or diseases.[48] Local traits were selected based on previously known phenotypic associations between TUD and other substance use disorder phenotypes and related traits (e.g., cannabis use disorder, various measures of impulsivity). We used the standard Benjamini–Hochberg false discovery rate correction (FDR 5%) to correct for multiple testing. We also calculated a Bonferroni correction for 115 comparisons (p<4.35E−04); however, this correction is overly conservative because many of the traits tested are highly correlated with one another. For AAs, we calculated r_g between TUD and 11 published traits using in-sample LD scores derived from the MVP genotype data using cov-LDSC.[78]

mtCOJO. We used mtCOJO[95] to individually condition the TUD-EUR summary statistics on loci associated with other comorbid traits, including alcohol dependence, cannabis use disorder and opioid use disorder. 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 a p value threshold of 5.00E-07, 5.00E-07, 1.00E-07, respectively, for alcohol dependence, cannabis use disorder, and opioid use disorder. 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**).[96] HCA forms clusters iteratively by creating groups and successively joining or splitting those groups based on a prespecified algorithm.[96] 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 (version 2.1.4).[96]

The product of HCA is a dendrogram, formed with multiple brackets called "branches". Phenotypes on the same branch are more similar to each other based on their pairwise genetic associations with each other and with all other phenotypes on that branch. Branches can form subbranches of more specific clustering. The genetic correlations of CigarettesPerDay, FormerSmoker, and SmokingInitiation were reversed to show the intuitive effects against the other traits in the dendrogram.

Phenome-wide association studies (PheWAS)

Mayo Clinic Biobank. We performed a PheWAS in the Mayo Clinic Biobank (MCB).[97] Phecodes were ascertained using EHR data from 57,001 patients from the Mayo Clinic Biobank. EHR data for the participants was extracted on September 23, 2022 and included any diagnoses on or before April 6, 2020, the date patient consent was checked. The Institutional Review Board of Mayo Clinic approved this study. Samples were sequenced at the Regeneron Genetics Center (RGC) using a custom design that additionally augments the exome capture with "backbone" regions intended to measure common tagging variation for purposes of GWAS. The backbone regions are targeted at lower depth and undergo substantial post-processing using proprietary algorithms that can boost genotyping quality based on shared information via linkage disequilibrium and population allele frequencies. The resulting GxS data was run through the Mayo Clinic Genotype QC pipeline. In this QC pipeline, SNPs were excluded using filters for call rate (<95%), minor allele frequency (<0.5%), and Hardy-Weinberg Equilibrium (p<1.00E-06). Individuals were excluded for excessive missing genotypes (>5%), sex errors, or abnormal heterozygosity (<70% on multiple chromosomes). Cryptic relatedness analysis was performed in an interative process using PLINK and PRIMUS to estimate IBD sharing. Highly related samples were removed from the sample if they had >100 closely related samples

(PI_HAT>0.1875) or >25000 related samples (PI_HAT>0.08); the relatedness analysis was performed iteratively until no such samples remained. For each pair with an estimated 2nd degree or higher relatedness, we removed the individual with shorter length of EHR. Finally, PRSs were calculated using LDpred2[98] l using the auto feature in the bigsnpr (v1.10.4) R package.

Yale-Penn. We performed PheWAS in the Yale-Penn sample;[49] which is a deeply phenotyped cohort using the Semi-Structured Assessment for Drug Dependence and Alcoholism, a detailed psychiatric instrument used to assess physical, psychosocial, and psychiatric manifestations of SUDs and comorbid psychiatric traits.[99,100] This comprehensive interview includes more than 3,500 items representing lifetime diagnostic criteria for the DSM-IV,[101] DSM-5[102] SUDs and DSM-IV[101] psychiatric disorder history. Genotyping and quality control for this cohort have been extensively described.[49,103]

Using PRS-Continuous shrinkage software (PRS-CS),[104] PRSs were calculated for TUD. 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 PRS 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 v0.12 R package[105] adjusting for sex, median age and the first ten PCs within each genetic ancestry. Bonferroni correction was applied for each ancestral-specific analysis to account for multiple testing (p<7.25E-05).

Adolescent Brain Cognitive Development (ABCD). We performed polygenic analyses in the ABCD sample.[106] Again using PRS-CS,[107] we fitted a fixed effects model in the ABCD European subsample (wave 3 for phenotypes, wave 3 for genotypes), controlling for first 10 PCs, age, sex, site, as fixed effect covariates and family ID as random effects covariates. We included 12 measures that showed significant *rg* in the adults datasets and were available in this cohort;

these included 2 binary phenotypes (pain, "any pain last month"; and suicide attempt, "description"), and 10 continuous measures (from the CBCL child behavior checklist[108]- "CBCL Externalizing", "CBCL ADHD", "CBCL Depression", "CBCL ADHD", "CBCL AnxDep"; "CBCL AnxDis", "CBCL OCD"; cognitive ability via the NIH cognitive toolbox total score;[109] BMI; weight; deprivation). Results were corrected for multiple testing (p<4.0E-03). Additional genotyping, QC and statistical details are described in the **Supplementary Material**.

Mendelian Randomization. Two-sample Mendelian randomization[110,111] was used to evaluate the potential causal association between 6 genetically correlated traits and TUD using samples of European ancestry only (without UKBB). We inferred causality bidirectionally using three methods: weighted median, inverse-variance weighted (IVW) and MR-Egger, followed by a pleiotropy test using the MR Egger intercept.[112,113] Instrumental variants were those associated with the exposure after clumping (r^2 = 0.01) and at p<1.0E-05. We considered causal effects as those for which at least two MR tests were significant (p<0.05) and that showed no evidence of violation of the horizontal pleiotropy test (MR-Egger intercept p>0.05).

Data Availability. The full summary statistics from the meta-analyses will be available through dbGaP upon publication.

Code Availability. All software used to generate results has been previously published, and corresponding citations are provided in the Methods.

Acknowledgements. MVJ, SBB, and SSR were supported by funds from the California Tobacco-Related Disease Research Program (TRDRP; Grant Number T29KT0526 and T32IR5226). SBB and were also supported by P50DA037844. BP, JM and SSR were supported by NIH/NIDA DP1DA054394. ASH was supported by AA030083. TTM was supported by NHGRI T32HG010464. ECJ was supported by K01DA051759. JG was supported by VA Merit Award CX001849-01 and 5R01DA054869. DBH was supported by R01 DA042090. LKD was

supported by R01 MH113362. HRK was supported by the Veterans Integrated Service Network 4 Mental Illness Research, Education and Clinical Center. RLK 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 National Institutes of Health.

CTSA (SD, Vanderbilt Resources) The project described was supported by the National Center for Research Resources, Grant UL1 RR024975-01, and is now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06.

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, 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 #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 (PMBB) 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 program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for

analysis. The PMBB is approved under IRB protocol# 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 National Institutes of Health under CTSA award number UL1TR001878.

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers 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.pdf. 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 would also like to thank The Externalizing Consortium for sharing the GWAS summary statistics of externalizing. The Externalizing Consortium: Principal Investigators: Danielle M. Dick, Philipp Koellinger, K. Paige Harden, Abraham A. Palmer. Lead Analysts: Richard Karlsson Linnér, Travis T. Mallard, Peter B. Barr, Sandra Sanchez-Roige. Significant Contributors: Irwin 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, P50AA022537, as well as a European Research Council Consolidator Grant (647648 EdGe to 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 Substance Use Disorders working group, UK10K Consortium, UK Biobank, and Philadelphia Neurodevelopmental Cohort.

Contributions

The corresponding author (S.S-R) conceived the idea for the paper and wrote and edited the manuscript. Other contributing authors contributed analyses (S.T., M.V.J., B.P., H.L., T.T.L., S.B.B., L.V-R, H.X., A.H., J.J.M., V.P., G.Y., B.S.L., B.C., R.L.K), or data (E.C.J., MVP, G.D.J., A.B., R.P., R.B., A.A., J.B., J.W.S., L.K.D., A.C.J., R.L.K.). All contributing authors wrote and edited the paper. We thank Bryan Quach and Jesse Marks for their help in supplying portions of the data needed to create Supplementary Figure 1.

Ethics declarations

Dr. Palmer is on the scientific advisory board of Vivid Genomics for which he receives stock options. Dr. Smoller is a member of the Scientific Advisory Board of Sensorium Therapeutics (with equity) and has received grant support from Biogen, Inc. He is PI 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. Dr. Kranzler is a member of advisory boards for 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 three years by Alkermes, Dicerna, Ethypharm, Lundbeck, Mitsubishi, Otsuka, and Pear Therapeutics; and with Dr Gelernter, a holder of U.S. patent 10,900,082 titled: "Genotype-guided dosing of opioid agonists," issued 26 January 2021. The other authors declare no competing interests.

939

References

1. Centers for Disease Control and Prevention (CDC). *Health Effects of Cigarette Smoking*. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm (2021).
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. WHO. The top 10 causes of death. *World Health Organization* https://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. Am. Acad. Psychiatr. Alcohol. Addict.* **14**, 106–123 (2005).
6. Tobacco use disorder and the lungs - McRobbie - 2021 - Addiction - Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1111/add.15309.
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. Off. J. Int. Soc. Twin Stud.* **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. *Addict. Abingdon Engl.* **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 multiple diseases: A systematic review and meta-analysis of Mendelian randomization studies. *eBioMedicine* **82**, (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.

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* appi.ajp.2020.2 (2021) doi:10.1176/appi.ajp.2020.20091390.

25. Sanchez-Roige, S. Emerging phenotyping strategies will advance our understanding of psychiatric genetics. *Nat. Neurosci.* **23**, 6 (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**, (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. 23andMe Research Team *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).

1038 39. Linnér, R. K. Multivariate analysis of 1.5 million people identifies genetic associations with
1039 traits related to self-regulation and addiction. *Nat. Neurosci.* **24**, 27 (2021).
1040 40. Karlsson Linnér, R. *et al.* Multivariate analysis of 1.5 million people identifies genetic
1041 associations with traits related to self-regulation and addiction. *Nat. Neurosci.* (2021)
1042 doi:10.1038/s41593-021-00908-3.
1043 41. Xiao, M.-F. *et al.* Neural Cell Adhesion Molecule Modulates Dopaminergic Signaling and
1044 Behavior by Regulating Dopamine D2 Receptor Internalization. *J. Neurosci.* **29**, 14752–
1045 14763 (2009).
1046 42. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human
1047 genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
1048 43. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma,
1049 D. Genetic mapping of cell type specificity for complex traits. *Nat. Commun.* **10**, 3222
1050 (2019).
1051 44. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set
1052 Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
1053 45. Sey, N. Y. A. *et al.* A computational tool (H-MAGMA) for improved prediction of brain-
1054 disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* **23**,
1055 583–593 (2020).
1056 46. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene
1057 expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825
1058 (2018).
1059 47. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association
1060 studies. *Nat. Genet.* **48**, 245–252 (2016).
1061 48. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity
1062 in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

49. Kember, R. L. *et al.* Phenome-wide Association Analysis of Substance Use Disorders in a Deeply Phenotyped Sample. *Biol. Psychiatry* (2022) doi:10.1016/j.biopsych.2022.08.010.

50. McLellan, A. T., Koob, G. F. & Volkow, N. D. Preaddiction—A Missing Concept for Treating Substance Use Disorders. *JAMA Psychiatry* **79**, 749–751 (2022).

51. 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).

52. 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**, (2019).

53. 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).

54. 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).

55. 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).

56. Erzurumluoglu, A. M. *et al.* Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Mol. Psychiatry* **25**, 2392–2409 (2020).

57. Toikumo, S., Xu, H., Gelernter, J., Kember, R. L. & Kranzler, H. R. Integrating human brain proteomic data with genome-wide association study findings identifies novel brain proteins in substance use traits. *Neuropsychopharmacology* **47**, 2292–2299 (2022).

58. Kember, R. L. & et al. Cross-ancestry meta-analysis of opioid use disorder uncovers novel loci with predominant effects on brain. *medRxiv* (2021) doi:https://doi.org/10.1101/2021.12.13.21267480.

59. Koob, G. F. & Volkow, N. D. Neurobiology of addiction: a neurocircuitry analysis. *Lancet Psychiatry* **3**, 760–773 (2016).

60. King, D. P. *et al.* Smoking Cessation Pharmacogenetics: Analysis of Varenicline and Bupropion in Placebo-Controlled Clinical Trials. *Neuropsychopharmacology* **37**, 641–650 (2012).

61. 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).

62. 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).

63. So, H.-C. *et al.* Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. *Nat. Neurosci.* **20**, 1342–1349 (2017).

64. 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).

65. 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).

66. Jang, S.-K. *et al.* Rare genetic variants explain missing heritability in smoking. *Nat. Hum. Behav.* **6**, 1577–1586 (2022).

67. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).

68. 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).

69. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

70. Dennis, J. K. *et al.* Clinical laboratory test-wide association scan of polygenic scores identifies biomarkers of complex disease. *Genome Med.* **13**, 6 (2021).

71. 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).

72. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

73. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

74. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

75. 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).

76. 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).

77. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

78. Luo, Y. *et al.* Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Human Molecular Genetics* **30**, 1521–1534 (2021).

79. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

80. 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).

81. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinforma. Oxf. Engl.* **32**, 1493–1501 (2016).

82. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).

83. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

84. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

85. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).

86. The GTEx Consortium atlas of genetic regulatory effects across human tissues | Science. https://www.science.org/doi/10.1126/science.aaz1776?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed.

87. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47**, 115–125 (2015).

88. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).

89. Beach, T. G. *et al.* Arizona Study of Aging and Neurodegenerative Disorders and Brain and Body Donation Program. *Neuropathol. Off. J. Jpn. Soc. Neuropathol.* **35**, 354–389 (2015).

90. Wingo, T. S. *et al.* Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat. Neurosci.* **24**, 810–817 (2021).

1158 91. Wingo, A. P. *et al.* Integrating human brain proteomes with genome-wide association data
1159 implicates new proteins in Alzheimer's disease pathogenesis. *Nat. Genet.* **53**, 143–146
1160 (2021).
1161 92. Bennett, D. A. *et al.* Religious Orders Study and Rush Memory and Aging Project. *J.*
1162 *Alzheimers Dis. JAD* **64**, S161–S189 (2018).
1163 93. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic
1164 Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).
1165 94. Liang, Y. *et al.* BrainXcan identifies brain features associated with behavioral and
1166 psychiatric traits using large scale genetic and imaging data. 2021.06.01.21258159
1167 Preprint at https://doi.org/10.1101/2021.06.01.21258159 (2022).
1168 95. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances
1169 circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
1170 96. Maechler,M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. Cluster: Cluster Analysis
1171 Basics and Extentions. (2013).
1172 97. Bielinski, S. J. *et al.* Mayo Genome Consortia: A Genotype-Phenotype Resource for
1173 Genome-Wide Association Studies With an Application to the Analysis of Circulating
1174 Bilirubin Levels. *Mayo Clin. Proc.* **86**, 606–614 (2011).
1175 98. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics*
1176 **36**, 5424–5431 (2020).
1177 99. Pierucci-Lagha, A. *et al.* Diagnostic reliability of the Semi-structured Assessment for Drug
1178 Dependence and Alcoholism (SSADDA). *Drug Alcohol Depend.* **80**, 303–312 (2005).
1179 100. Pierucci-Lagha, A. *et al.* Reliability of DSM-IV Diagnostic Criteria Using the Semi-
1180 Structured Assessment for Drug Dependence and Alcoholism (SSADDA). *Drug Alcohol*
1181 *Depend.* **91**, 85–90 (2007).

101. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*. (American Psychiatric Association, 1994).

102. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. (American Psychiatric Association, 2013).

103. Gelernter, J. *et al.* Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry* **19**, 41–49 (2014).

104. 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).

105. 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).

106. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* **36**, 930–933 (2020).

107. Ruan, Y. *et al.* Improving Polygenic Prediction in Ancestrally Diverse Populations. *medRxiv* 21 (2021) doi:https://doi.org/10.1101/2020.12.27.20248738.

108. Rescorla, L. *et al.* Behavioral/Emotional Problems of Preschoolers Caregiver/Teacher Reports From 15 Societies. *J. Emot. Behav. Disord.* **20**, 68–81 (2012).

109. 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).

110. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).

111. 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).

112. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* **37**, 658–665 (2013).

113. 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).