

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

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

13 <sup>1</sup>Mental Illness Research, Education and Clinical Center, Crescenz VAMC, Philadelphia, PA,  
14 USA; <sup>2</sup>Department of Psychiatry, University of Pennsylvania Perelman School of Medicine,  
15 Philadelphia, PA, USA; <sup>3</sup>Department of Psychiatry, University of California San Diego, San  
16 Diego, CA, USA; <sup>4</sup>Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic  
17 Medicine, Massachusetts General Hospital, Boston, MA, USA; <sup>5</sup>Department of Psychiatry,  
18 Harvard Medical School, Boston, MA, USA; <sup>6</sup>Psychiatric Genetics Unit, Group of Psychiatry,  
19 Mental Health and Addiction, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de  
20 Barcelona, Barcelona, Spain; <sup>7</sup>Department of Psychiatry, Washington University School of  
21 Medicine, Saint Louis, Missouri, USA; <sup>8</sup>Department of Quantitative Health Sciences, Mayo  
22 Clinic, Rochester, MN, USA; <sup>9</sup>Program in Biomedical Sciences, University of California San  
23 Diego, La Jolla, CA, USA; <sup>10</sup>Vanderbilt Genetics Institute, Vanderbilt University Medical Center,  
24 Nashville, TN, USA; <sup>11</sup>Department of Psychology, University of Ibadan, Nigeria; <sup>12</sup>Department of  
25 Psychiatry, Yale University School of Medicine, New Haven, CT, USA; <sup>13</sup>Veterans Affairs  
26 Connecticut Healthcare System, West Haven, CT, USA; <sup>14</sup>Department of Psychiatry &  
27 Psychology, Mayo Clinic, Rochester, MN, USA; <sup>15</sup>Behavioral and Urban Health Program,  
28 Behavioral Health and Criminal Justice Division, RTI International, Research Triangle Park, NC,  
29 USA; <sup>16</sup>Department of Medicine, Division of Genetic Medicine, Vanderbilt University, Nashville,  
30 TN, USA; <sup>17</sup>Yale University School of Public Health, New Haven, CT, USA; <sup>18</sup>Veterans Affairs  
31 Connecticut Healthcare System, West Haven, CT, USA; <sup>19</sup>Yale University School of Medicine,  
32 New Haven, CT, USA.

33 \* These authors contributed equally

34 **Penn Medicine BioBank** is provided in the Supplementary Material

35 **Million Veteran Program** is provided in the Supplementary Material

36 **PsycheMERGE Substance Use Disorders Group (alphabetically):** Sevim B Bianchi,

37 Brandon J Coombes, Richard C Crist, Renata Cupertino, Lea K Davis, Mariela V Jennings,

38 Rachel L Kember, Hyunjoon Lee, Travis T Mallard, Maria Niarchou, Melissa N Poulsen, Sandra

39 Sanchez-Roige, Jordan W Smoller, Vanessa Troiani, Colin G Walsh

40 **Corresponding author:** Sandra Sanchez-Roige, PhD, Department of Psychiatry, University of

41 California San Diego, San Diego, CA, USA. Email: [sanchezroige@ucsd.edu](mailto:sanchezroige@ucsd.edu)

42 **ABSTRACT**

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

55

56 Tobacco use disorder (**TUD**) is the most prevalent substance use disorder in the world,  
57 with a high proportion of smokers meeting criteria for nicotine dependence.<sup>1,2</sup> Nicotine  
58 dependent individuals often experience withdrawal symptoms when they stop smoking. As a  
59 result, they often have substantial difficulty quitting and continue to smoke despite negative  
60 mental, social, and medical consequences. Tobacco smoking is the leading cause of  
61 preventable death worldwide, causing 6 million annual premature deaths,<sup>3</sup> and is also highly  
62 associated with other worldwide leading contributors of morbidity and mortality, including lung  
63 cancer, chronic obstructive pulmonary disease, cardiovascular disease, mood disorders, and  
64 other substance use disorders.<sup>4-6</sup> Unfortunately, available preventative and treatment options  
65 for TUD have low success rates.<sup>7</sup>

66 Genetic factors influence smoking behaviors, with twin-heritability estimates ranging  
67 from ~30-70%.<sup>8-12</sup> Recently, genome-wide association studies (**GWAS**) have expanded in size  
68 (N~2.5M) and yielded hundreds of novel loci for smoking-related behaviors (summarized in  
69 **Supplementary Table 1**), primarily for nicotine *consumption*.<sup>13</sup> These GWAS have revealed  
70 pervasive pleiotropy, with Mendelian randomization (**MR**) analyses highlighting potential causal  
71 effects of regular tobacco smoking on health outcomes (e.g., cardiovascular health,<sup>14</sup> cancer  
72 risk,<sup>14</sup> bone mineral density<sup>15</sup>), numerous other substance use disorders (e.g., alcohol,<sup>14</sup>  
73 cannabis<sup>16</sup> and opioid use disorders<sup>17</sup>), and psychiatric and related conditions (e.g., major  
74 depressive disorder,<sup>18</sup> suicide-related behaviors,<sup>19</sup> loneliness<sup>20</sup>).

75 While these studies have been immensely successful, they have not focused on TUD  
76 itself. As a result, relatively little is known about the specific genes that confer risk for the  
77 development of TUD and associated conditions. One of the major roadblocks to progress in  
78 identifying risk-conferring genes has been the lack of sufficiently large samples with *misuse*  
79 phenotypes. This is an important limitation because prior studies have shown that the genetic  
80 architecture of substance use is largely different from that of misuse.<sup>21-26</sup> The largest GWAS of

81 nicotine dependence, comprising 58,000 European- and African-ancestry smokers, using the  
82 self-reported Fagerström Test for Nicotine Dependence (**FTND**), identified only five loci.<sup>27</sup> In  
83 addition, while there have been nicotine dependence GWAS in individuals of ancestries other  
84 than European<sup>28</sup> (**Supplementary Table 1** for full list), sample sizes for diverse populations  
85 have been limited (N<12K).

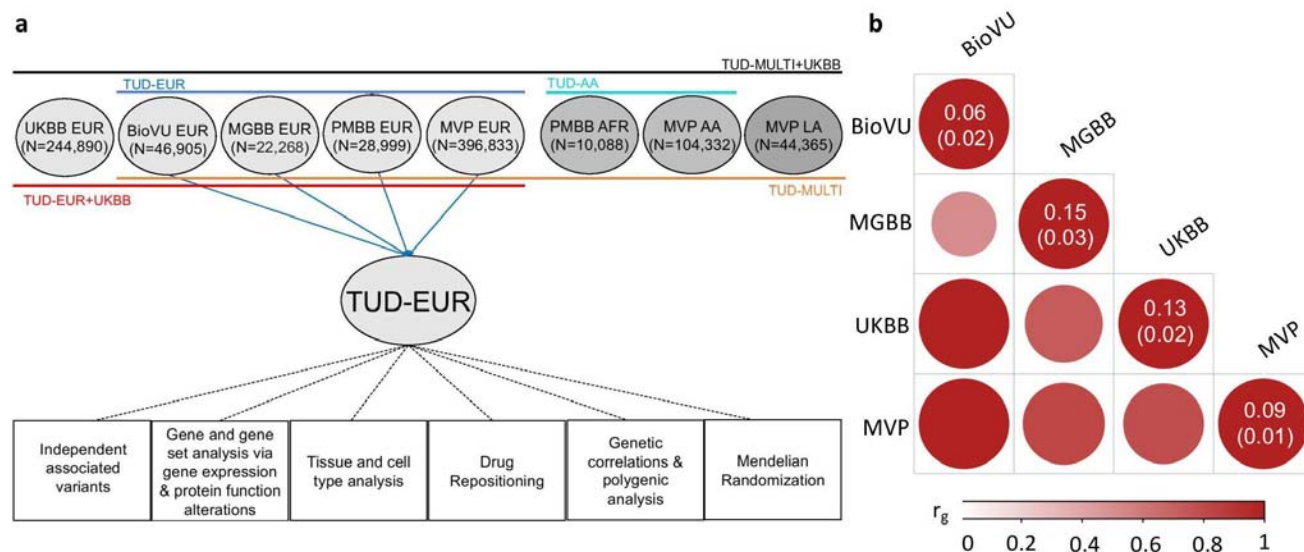
86 The use of electronic health records (**EHR**) is a relatively untapped, cost-effective  
87 strategy for characterizing smoking-related phenotypes, including TUD. EHR-defined TUD  
88 generally relies on International Classification of Disease (**ICD**) diagnostic codes, which can be  
89 aggregated into “phecodes” that require the presence of an ICD code on two or more separate  
90 visits. TUD diagnostic codes are effective identifiers of smoking status.<sup>29</sup> A key consideration,  
91 and the one we examine in this study, is the utility of TUD phecodes for use in large-scale  
92 GWAS to boost power and improve our ability to identify novel loci for TUD.<sup>29–31</sup> To address this  
93 question, we performed a multi-ancestral meta-analysis of TUD comprising 898,680 individuals  
94 of European (**EUR**), African American (**AA**) and Latin American (**LA**) ancestry recruited from  
95 multiple biobanks within the PsycheMERGE network<sup>32</sup> (Vanderbilt University Medical Center’s  
96 biobank, **BioVU**, N<sub>EUR</sub>=46,905; Mass General Brigham Biobank, **MGBB**, N<sub>EUR</sub>=22,268; Penn  
97 Medicine BioBank, **PMBB**,<sup>33</sup> N<sub>EUR</sub>=28,999, N<sub>AA</sub>=10,088; Million Veteran Program, **MVP**,  
98 N<sub>EUR</sub>=396,833, N<sub>AA</sub>=104,332, N<sub>LA</sub>=44,365), and combined with existing data from the UK  
99 Biobank (**UKBB**, N<sub>EUR</sub>=244,890), which used a less stringent definition. In secondary analyses,  
100 we further characterized the genetic architecture of TUD, examined pleiotropy with other  
101 psychiatric and medical outcomes, and harnessed the data to reveal new potential medications  
102 for treating this serious psychiatric condition.

103

104

## 105 Results

106 **Cohort Descriptions and Phenotype Validation.** We included individuals from eight cohorts  
 107 across five different sites (**Figure 1a** for an overview of the cohorts; **Supplementary Table 2** for  
 108 sample sizes). The methods to ascertain cases were identical for seven of these cohorts.  
 109 Individuals were identified as cases if they met criteria for a TUD phecode (a TUD ICD9 or  
 110 ICD10 code on two or more separate visits, described in **Supplementary Table 3**); controls  
 111 were screened for the absence of a TUD diagnosis. We benchmarked the TUD-EHR definition  
 112 against self-reported smoking questionnaire data and other comorbid ICD codes  
 113 (**Supplementary Table 4**). Across contributing biobanks, cases were enriched for ever smokers  
 114 (92-99%), with only a minor proportion (<2%) of cases self-identifying as never-smokers  
 115 (**Supplementary Table 5**). In contrast, a smaller proportion of controls were ever smokers (17-  
 116 56%), with a larger proportion self-identifying as never-smokers (39-73%). Attempts at smoking  
 117 cessation were reported by 15-25% of controls and 65-95% of cases. Controls were comparable  
 118 to cases on age and sex but reported much lower prevalences of other substance and  
 119 psychiatric disorders than cases. Thus, almost all TUD cases have evidence of being either  
 120 former or current smokers based on available self-report data.



121

122 **Figure 1. Overview of the cohorts and analysis pipeline (a) and genetic correlations**  
123 **among the sites (b).** (a) We conducted independent GWAS of TUD cases and controls in  
124 individuals of European (EUR) ancestry across four PsycheMERGE sites (BioVU, MGBB,  
125 PMBB, and MVP) and performed a GWAS meta-analysis (“TUD-EUR”); these summary results  
126 were used for all secondary analyses. For African American (AA), we conducted GWAS meta-  
127 analysis of TUD cases and controls from the PMBB and MVP cohorts (“TUD-AA”). For Latin  
128 American (LA), we conducted GWAS of TUD cases and controls from the MVP cohort. Next, we  
129 performed a multi-ancestral GWAS meta-analysis (“TUD-multi”), which combined the results  
130 from all seven cohorts. We also obtained summary statistics from UKBB, which used a less  
131 stringent case definition in individuals of EUR ancestry and performed a GWAS meta-analysis  
132 within EUR individuals (“TUD-EUR+UKBB”) and across ancestries (“TUD-multi+UKBB”).  
133 **Supplementary Table 2** summarizes the datasets used for the analyses. We subjected the  
134 TUD-EUR summary statistics to several secondary analyses to characterize the genetic  
135 architecture of TUD. (b) LDSC genetic correlations for TUD between all different EUR sites  
136 were positive and high, ranging from 0.51 to unity. LDSC genetic correlation for TUD across the  
137 two AA sites was strongly positive (0.86) but not significant ( $p=0.38$ ). We do not report  $r_g$   
138 between PMBB and other sites, because the  $h^2_{\text{SNP}}$  of TUD in PMBB was not significant ( $h^2_{\text{SNP}}$   
139  $=0.90$ ,  $\text{SE}=1.30$ ). LDSC SNP-heritability estimates ( $h^2_{\text{SNP}}$  6-15%) are shown in the diagonal.  
140 UKBB=UK Biobank, BioVU=Vanderbilt University Medical Center’s biobank, MGBB=Mass  
141 General Brigham Biobank, PMBB=Penn Medicine Biobank, MVP=Million Veteran Program.

142

143 **Significant SNP-heritability and genetic correlations across sites.** After applying similar  
144 data quality controls, we conducted within-cohort association analyses using logistic regression  
145 and relevant covariates (**Methods**). We estimated the proportion of variance attributable to the  
146 measured common variants (SNP-heritability,  $h^2_{\text{SNP}}$ ) to be ~6-15% (based on liability scale,  
147 assuming a lifetime risk of 12.5%; **Figure 1b, Supplementary Table 6**), which is consistent with  
148 prior nicotine-related GWAS.<sup>13,27</sup> Genetic correlations across sites and ancestries were high and  
149 positive ( $r_g=0.51-1.24$ ,  $p<1.80\text{E-}02$ , EUR sites;  $r_g=0.86$ ,  $p=0.38$ , AA sites; cross-ancestry  
150  $r_{gs}=0.72-0.84$ ,  $p<7.80\text{E-}04$ ; **Figure 1b, Supplementary Table 6**), serving as the basis for



151 ancestry-specific and multi-ancestry meta-analyses, and suggesting that the genetic  
152 architecture of TUD is similar across ancestries.

153 **Multi-ancestry meta-analyses implicate biological underpinnings of TUD.** The  
154 primary multi-ancestry meta-analysis of 29,448,768 imputed SNPs ( $\lambda_{GC}=1.107$ , **Figure**  
155 **2**) was performed on seven cohorts, comprising 653,790 individuals, with 75.71% EUR, 17.50%  
156 AA, and 6.79% LA.

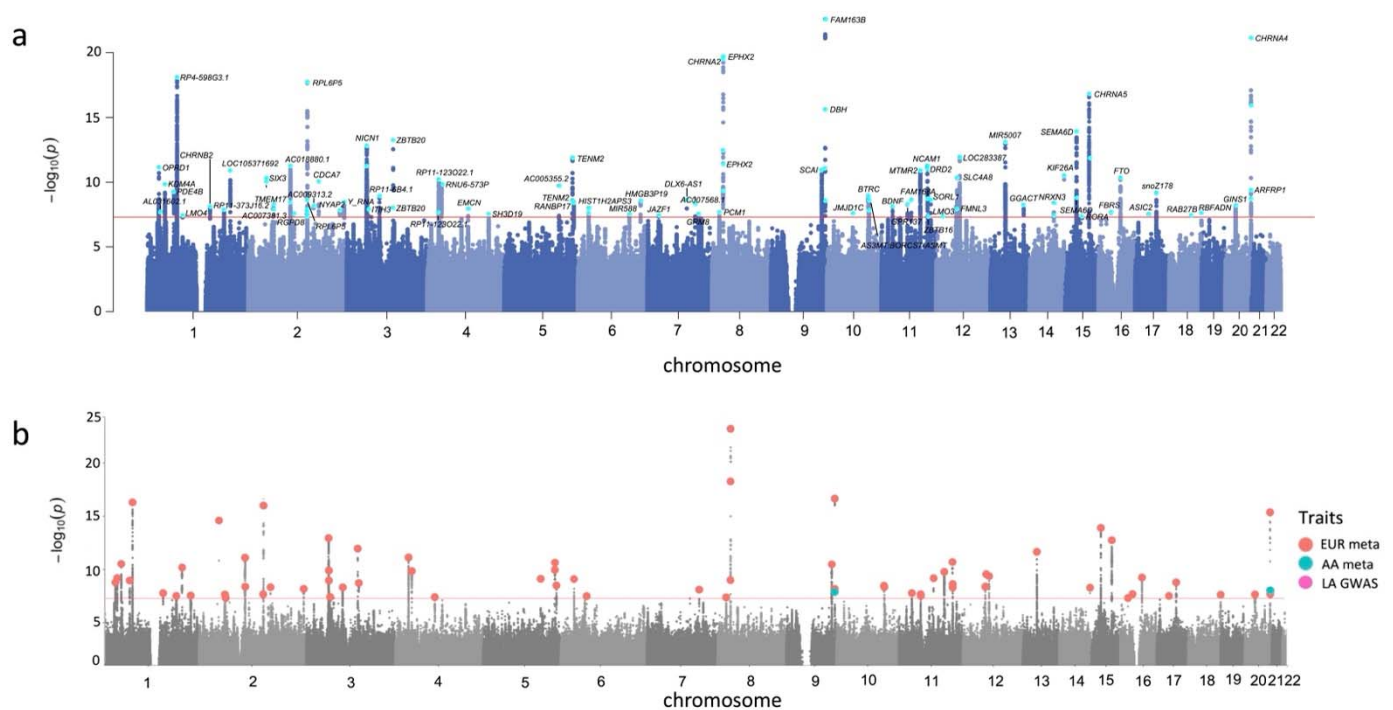
157 We identified 97 GWS ( $p<5.00E-08$ ) lead SNPs ( $r^2<0.1$ ) located in 72 independent loci  
158 (**Supplementary Table 7**). All genome-wide significant loci had been previously reported by  
159 prior smoking GWAS (**Supplementary Table 7**), including aspects of smoking initiation (7/72),  
160 consumption (22/72), cessation (42/72) and nicotine dependence (1/72; **Supplementary Figure**  
161 **1**). While all these loci were recently discovered in a GWAS of 3.4 million individuals in the  
162 GSCAN study,<sup>13</sup> here we reproduce some of the GSCAN findings with a considerably smaller  
163 sample size.

164 Our analyses provide corroborative support for nicotinic acetylcholine receptor genes as  
165 risk genes for smoking-related traits: *CHRNA5* (rs576982,  $p=1.60E-17$ , chr. 15; this region  
166 includes rs16969968, a well-established functional missense polymorphism [D398N] in  
167 *CHRNA5*,  $p=4.93E-11$ ), *CHRNA2* (rs2741339,  $p=2.86E-20$ , chr. 8), and *CHRNA4* (rs2273500,  $p=7.34E-22$ , chr. 20). Second, we identified  
168 associations with variants in several genes that modulate dopaminergic transmission, such as  
169 the dopamine receptor D2 (*DRD2*: genomic position 113334227,  $p=1.04E-11$ , and rs4936277,  
170  $p=1.81E-09$ , chr.11), known for its relationship with dopamine and reward,<sup>34</sup> previously  
171 associated to nicotine dependence<sup>35</sup> and implicated in a recent large-scale GWAS of  
172 addiction;<sup>36</sup> dopamine beta-hydroxylase (*DBH*: rs2007153,  $2.55E-16$ , and rs2519155,  $p=8.74E-$   
173  $12$ , chr.9), which encodes an enzyme necessary to convert dopamine to norepinephrine and has  
174 been consistently implicated in smoking behaviors;<sup>13,37</sup> lysine demethylase 4A (*KDM4A*:



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

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



193

194 **Figure 2. Manhattan and porcupine plots for the TUD-multi meta-analysis and ancestry-**  
195 **specific GWAS. (a)** TUD-multi identified 72 independent risk loci, all of which were recently  
196 replicated by the GSCAN study. **(b)** Porcupine plot of ancestry-specific meta-analyses identified  
197 55 loci in the European cohort (EUR, in red), and 2 loci in the African cohort (AA, in blue). No  
198 significant associations were detected in the Latin American (LA) cohort.

199

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

203 TUD-EUR included 19,096,380 imputed SNPs in a cohort of 163,734 TUD cases and  
204 331,271 controls, which is 8.5 times larger than the total sample size of previous nicotine  
205 dependence GWAS.<sup>27</sup> Observable inflation is attributable to polygenic signal rather than  
206 population stratification or other confounding (LDSC intercept 1.052, SE=0.012) and we did not

207 identify evidence of heterogeneity ( $I^2$ ) across the cohorts (**Supplementary Figure 3**). The TUD-  
208 EUR meta-analysis yielded a significant  $h^2_{SNP}$  estimate of 7.10% (SE=0.003, **Supplementary**  
209 **Table 9**), and identified 68 GWS significant lead SNPs located in 55 independent loci (**Figure**  
210 **2B; Supplementary Table 10**). Ten of these loci were ancestry specific in EUR and not GWS in  
211 the multi-ancestry GWAS. Among the 55 independent loci, 8 were fine-mapped to a credible set  
212 (posterior inclusion probability > 0.50), of which 6 harbored known protein coding genes  
213 (*ZBTB20*, *HIST1H2BH*, *BDNF*, *SLC4A8*, *KIF26A*, *ASIC2*; **Supplementary Table 11**).

214 Again, combining these data with those of UKBB in a secondary GWAS (TUD-  
215 EUR+UKBB) yielded very similar results (e.g., similar  $h^2_{SNP}$  estimate of 7.00%; with the addition  
216 of three independent loci - *LOC105373664*\*rs6430094, *GALNT10*\*rs7737824, and  
217 *CHRNA4*\*rs6011779, **Supplementary Table 12**). Considering the similarity in the number of  
218 loci identified between the primary and secondary GWAS, all downstream analyses used the  
219 EUR GWAS for the most stringent TUD definition (TUD-EUR), which excluded the UKBB  
220 sample.

221 The TUD-AA meta-analysis yielded a significant  $h^2_{SNP}$  estimate of 11.30% (SE=0.015,  
222 **Supplementary Table 9**), and 2 independent loci (**Supplementary Table 13**), one on chr. 9  
223 (rs2007153,  $p=1.17E-08$ ) in *DBH*, which is novel for the AA population, and another on chr. 20  
224 (rs6011779,  $p=9.27E-09$ ) in the *CHRNA4* gene, replicating a finding from a prior multi-ancestral  
225 (EUR+AA) GWAS of smoking.<sup>27</sup> Multi-ancestry fine-mapping analyses using PAINTOR  
226 corroborated the region in chr. 9, identifying two putative causal variants in this locus  
227 (**Supplementary Table 14**). The TUD-LA GWAS yielded a significant  $h^2_{SNP}$  estimate of 8.10%  
228 (SE=0.017, **Supplementary Table 9**) but did not identify any GWS loci (**Figure 2**), presumably  
229 due to the smaller sample size.

230

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

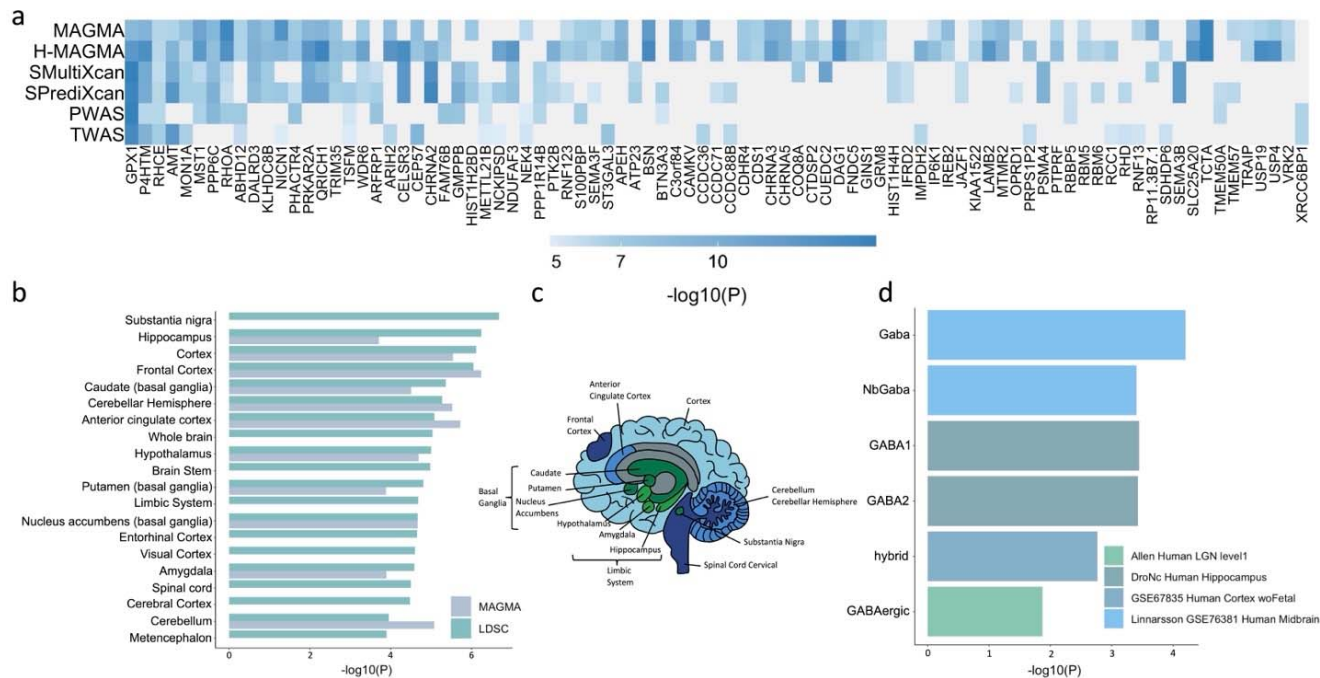
235 First, we conducted gene-based analyses via MAGMA,<sup>43,44</sup> which mapped SNP-level  
236 associations to 86 significant genes ( $p < 2.63E-06$ ), 83 (90.69%) of which replicated genes near  
237 or in GWS loci (e.g., *CHRNA3*, *CHRNA4*, *CHRNA5*, *BDNF*, *PTPRF*, *KDM4A*, *DBH*;  
238 **Supplementary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

239 To identify neurobiologically relevant target genes, we incorporated TUD GWAS data  
240 with chromatin interaction profiles from human brain tissue using Hi-C coupled MAGMA (H-  
241 MAGMA).<sup>45</sup> These analyses identified 746 unique gene-tissue pairs associated with TUD  
242 ( $p < 9.44E-07$ ), a significant proportion of which showed cell-type (16.49% cortical neurons,  
243 16.75% iPSC derived neurons, 20.78% midbrain dopaminergic neurons, 13.00% iPSC derived  
244 astrocytes) or developmental stage (15.55% fetal, 17.43% adult) specific expression  
245 (**Supplementary Table 16**).

246 Using S-MultiXcan to predict the effect of common SNP variation on gene expression in  
247 multiple brain tissues, we detected significant associations for 34 genes (**Supplementary Table**  
248 **17**), with effects dispersed across 12 brain regions (cerebellum, anterior cingulate cortex, basal  
249 ganglia [nucleus accumbens and putamen], cortex and frontal cortex, amygdala, hypothalamus,  
250 substantia nigra, spinal cord, cerebellar hemisphere, spinal cord). Inspection of region-specific  
251 results via S-PrediXcan identified five genes that were consistently upregulated (*GPX1*, *PPP6C*,  
252 *GMPPB*, *WDR6*) or downregulated (*CHRNA2*) in more than one brain region (**Supplementary**  
253 **Table 18**).

254           Next, we assessed differential transcriptomic and proteomic regulation of TUD risk loci in  
255 the dorsolateral prefrontal cortex (DLPFC) by performing TWAS (mRNA and splicing) and  
256 PWAS, respectively. Associations across these three regulatory models identified 43 TUD  
257 unique risk genes (34, mRNA expression; 15, splicing expression; 14, proteome expression;  
258 **Supplementary Tables 19 and 20**). Colocalization analysis identified five genes and proteins  
259 (NT5C2, GPX1, NEK4, ABHD12, RHCE) associated with TUD via their regulation of brain  
260 expression levels and protein abundance (PP4 >0.80, **Supplementary Table 21**,  
261 **Supplementary Figure 4**).

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



269

270 **Figure 3. Integration with functional genomic data implicated 330 unique TUD candidate**

271 **risk genes.** (a) Of 330 associated genes, 87 converged with at least 2 methods, and were

272 dispersed throughout the chromosomes. (b) LDSC (SNP-based) and MAGMA tissue-specific

273 gene expression of TUD risk genes reveals substantial brain enrichment. (c) The genetic

274 findings across multiple levels of analysis (LDSC, MAGMA, MultiXcan, BrainXcan) implicated

275 brain regions exhibiting anatomical differences in cases. (d) Cell type-specific expression of

276 TUD risk genes. Results from MAGMA property analyses and gene expression using human

277 single-cell RNA-sequencing datasets (**Supplementary Table 28** for full list). After multiple

278 testing correction for all datasets, only genes expressed in GABAergic neurons were associated

279 with TUD (**Supplementary Table 28**).

280

281 **Tissue and cell-type analyses of TUD identify enrichment in brain tissue and**

282 **GABAergic neurons.** To identify relevant tissues implicated in TUD, we performed various

283 SNP (LDSC partitioned heritability) and gene-wide (MAGMA) analyses. We performed

284 partitioned heritability in LDSC to evaluate the enrichment of the genome-wide findings in over

285 50 functional genomic annotations (and across tissues, as described below). In the baseline

286 LDSC model, conserved and regulatory functional annotations were significantly enriched  
287 (**Supplementary Figure 5** and **Supplementary Table 23** for full list).

288 Tissue enrichment analyses in MAGMA use gene expression data from GTEx (v8). In  
289 addition to non-brain tissues (i.e., cardiovascular, hematopoietic, adrenal pancreas, and other,  
290  $p < 3.37E-05$ , **Supplementary Table 24**), we detected significant enrichment mostly in the brain  
291 ( $p = 1.53E-15$ ), spanning multiple brain regions, including the hippocampus, the limbic system,  
292 frontal cortex (**Supplementary Tables 25-26, Figure 3b-c**), most of which were also implicated  
293 in S-MultiXcan (**Supplementary Table 17**). Correlating the effects of SNP variation with brain  
294 imaging traits via BrainXcan identified similar results, including significant ( $p < 1.92E-04$ )  
295 associations with decreased gray matter volume in the right ventral striatum (**Supplementary**  
296 **Table 27**).

297 Next, we used FUMA to examine cell-type specific gene expression associated with  
298 TUD, leveraging single-cell RNA-sequencing (sc-RNA seq) datasets. We identified a significant  
299 association ( $p < 0.05$ ) between TUD risk and cell-type specific gene expression in GABAergic  
300 neurons for individual human sc-RNA seq datasets (Linnarsson, midbrain,  $p < 3.94E-04$ ; Allen  
301 Brain Atlas, dorsal lateral geniculate nucleus,  $p = 1.34E-02$ ; DroNc-seq, hippocampus,  $p < 3.74E-$   
302  $04$ ; **Figure 3d; Supplementary Table 28**). These results did not survive conditional analyses  
303 within and across datasets.

304

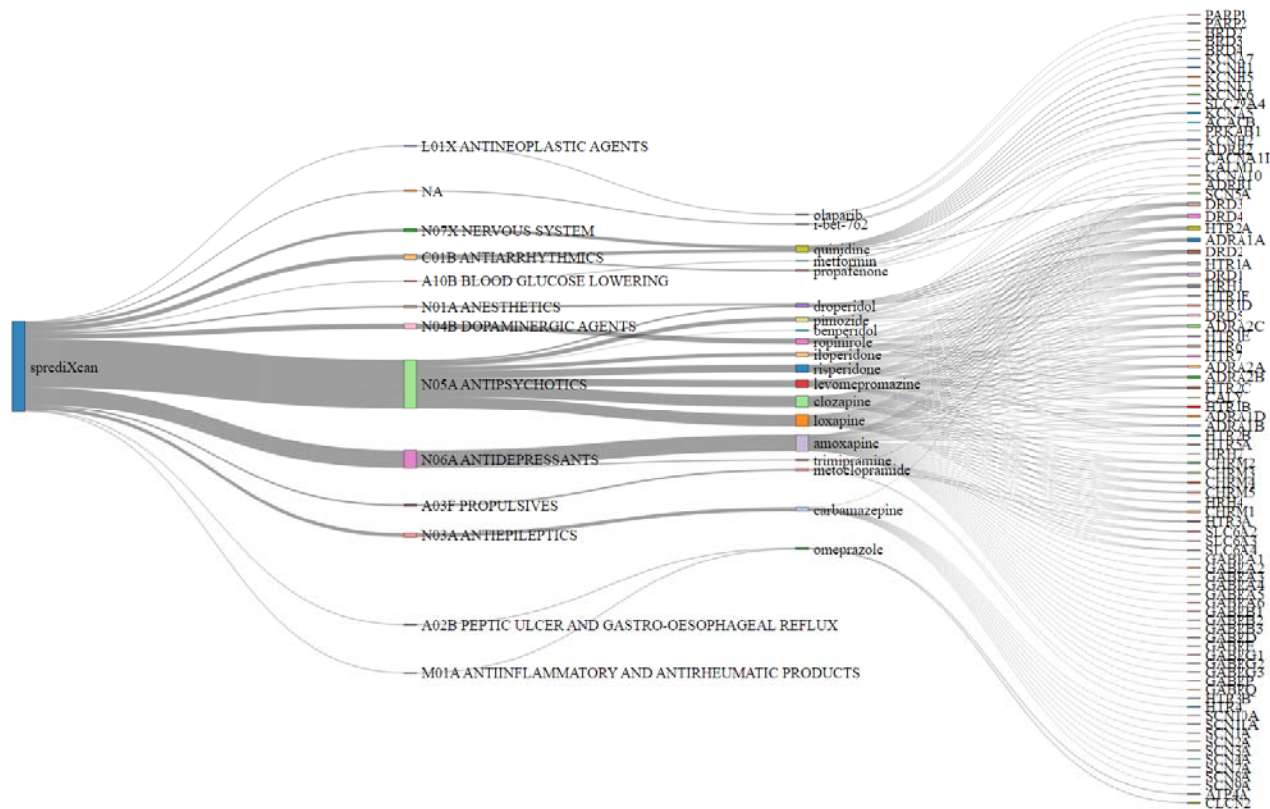
305 **Implications for TUD biology based on gene-set and pathway analyses.** We used  
306 MAGMA<sup>43,44</sup> to conduct a gene-wise TUD analysis and to test for enrichment of pathways  
307 curated from multiple sources. After correcting for multiple comparisons, 25 related pathways  
308 and biological processes were significantly enriched for genes associated with TUD ( $p < 2.76E-$   
309  $06$ ; **Supplementary Table 29**). Associations implicated fundamental processes related to



310 nicotine response (e.g., high calcium and sodium permeable nicotinic acetylcholine receptors,  
311  $p=4.66E-16$ ; behavioral response to nicotine,  $p=5.97E-16$ ), regulation of postsynaptic signaling  
312 ( $p=2.61E-08$ ), and maintenance of synapse structure ( $p=9.26E-07$ ), among others.

313

314 **Drug Repurposing.** Linking transcriptome-wide patterns to perturbagens that pass the blood-  
315 brain barrier from the Library of Integrated Network-Based Cellular Signatures (LINCS)<sup>36</sup>  
316 database identified 293 FDA approved medications approved by the U. S. Food and Drug  
317 Administration (**Supplementary Table 30**). 31 of the 293 identified medications targeted at least  
318 one mapped/independent gene from our GWAS. The medications that significantly reversed  
319 (Bonferroni  $p<6.03E-05$ ) the transcriptional profile associated with TUD included varenicline (a  
320 well-known therapeutic for smoking cessation), sodium channel blockers (e.g., amiloride), and  
321 compounds that are used to treat conditions that commonly co-occur with TUD, such as  
322 antipsychotics (e.g., clozapine), dopaminergic agents (e.g., ropinirole), opioids (e.g.,  
323 nalbuphine), and antidepressants (e.g., amoxapine), among others (**Figure 4**).



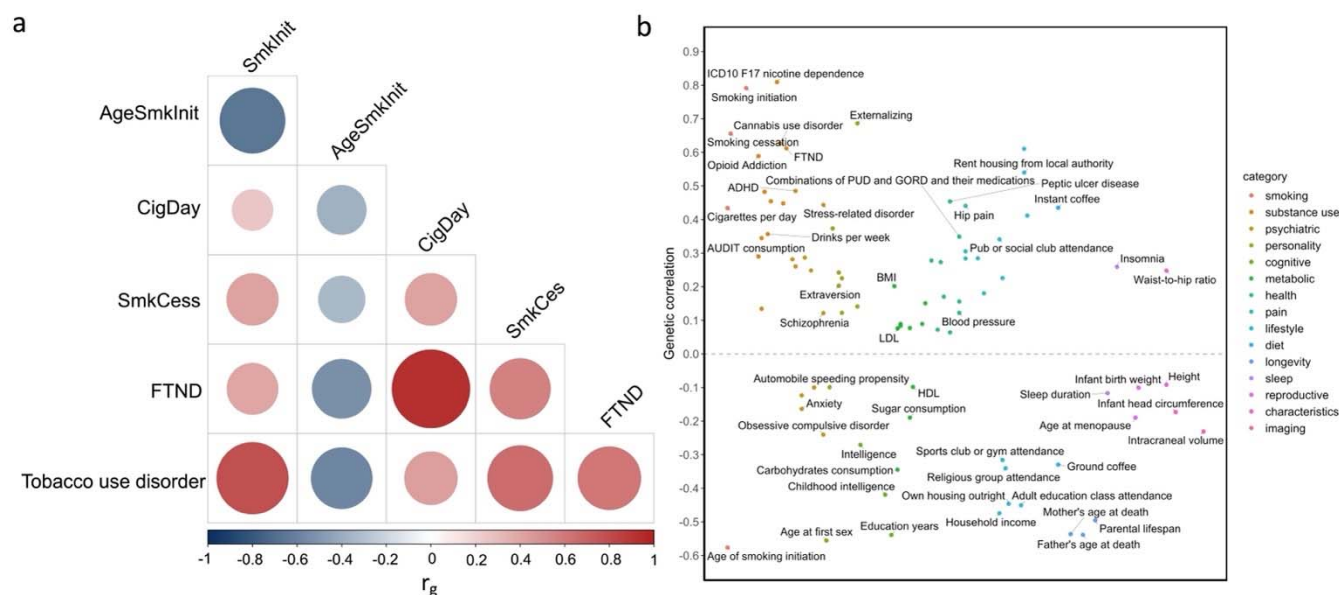
324

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

331

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

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



340  
 341 **Figure 5. FDR-significant genetic correlations between TUD-EUR and 115 complex traits,**  
 342 **including smoking and related phenotypes (b).** (a) Genetic correlations ( $r_g$ ) between age of  
 343 smoking initiation (AgeSmkInit), cigarettes per day (CigDay), smoking cessation (SmkCess),  
 344 nicotine dependence via the Fagerström Test for Nicotine Dependence (FTND), and tobacco  
 345 use disorder (see **Supplementary Table 31** for full results). (b) Genetic correlations with an  
 346 extended list of traits from publicly available GWAS. Traits with positive  $r_g$  values are plotted  
 347 above the line; traits with negative  $r_g$  values below the line. All  $r_g$ s are significant using a 5%  
 348 FDR correction for multiple testing. AgeSmkInit, age of smoking initiation smoking; CigDay,  
 349 cigarettes smoked per day; SmkCess, smoking cessation;<sup>13</sup> FTND, Fagerstrom Test for Nicotine  
 350 Dependence.<sup>27</sup>

351  
 352  
 353 TUD was also genetically associated with psychiatric and medical conditions (**Figure 5b,**  
 354 **Supplementary Table 31**). There were significant positive  $r_g$  with psychiatric traits (e.g.,  
 355 externalizing  $r_g=0.69$ ,  $SE=0.02$ ; ADHD  $r_g=0.49$ ,  $SE=0.04$ ; stress-related disorder  $r_g=0.44$ ,  
 356  $SE=0.04$ ) and risky behavioral traits, including lower age of first sex ( $r_g=-0.56$ ,  $SE=0.02$ ). We  
 357 also found positive  $r_g$  with health outcomes (e.g., coronary artery disease  $r_g=0.27$ ,  $SE=0.03$ ;

358 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  
359 several social determinants of health, such as the Townsend deprivation index ( $r_g=0.61$ ,  
360  $SE=0.07$ ). There were negative  $r_g$  with socioeconomic variables, including years of education  
361 ( $r_g=-0.54$ ,  $SE=0.02$ ) and household income ( $r_g=-0.47$ ,  $SE=0.04$ ) and with childhood intelligence  
362 ( $r_g=-0.42$ ,  $SE=0.07$ ). Conditioning on alcohol, cannabis, or opioid use disorders did not  
363 substantially modify the magnitude or direction of these associations (**Supplementary Table**  
364 **32**). Virtually all  $r_g$  estimates for other phenotypes were greater with TUD than cigarettes per day  
365 (**Supplementary Figure 7**) and FTND (**Supplementary Figure 8**).

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

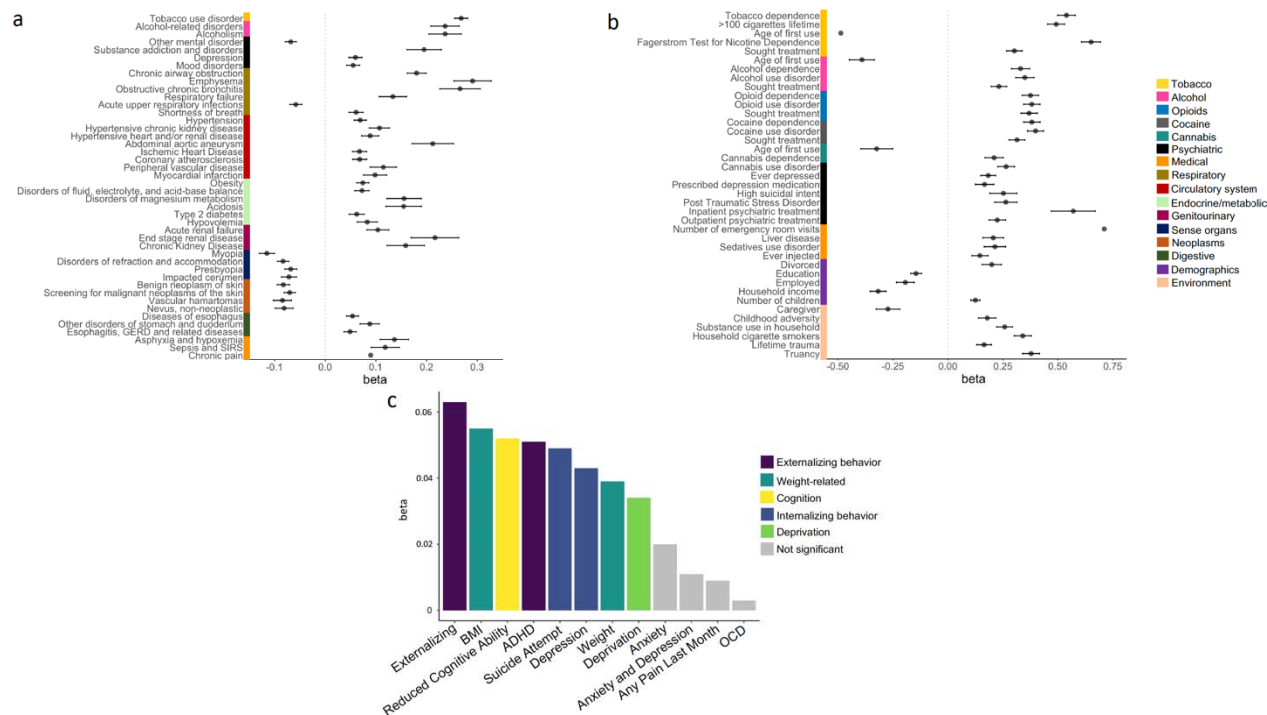
371 **Phenome-wide association analyses.** To further explore pleiotropic effects, we performed a  
372 series of phenome-wide association studies (**PheWAS**) of TUD polygenic scores (**PGS**) in other  
373 EHR and clinical cohorts of adults, and a young population-based cohort. We performed these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

375 *EHR cohorts.* We conducted PheWAS with EHR data to test the association between polygenic  
376 risk for TUD and liability for thousands of other medical conditions, including TUD, in another  
377 independent site, Mayo Clinic. As expected, TUD PGS was strongly associated with TUD-EHR  
378 ( $p=1.60E-88$ , **Supplementary Table 34, Figure 6a**), explaining 6.3% of the (Nagelkerke's  $R^2$ )  
379 variance. Additional significant ( $p<7.25E-05$ ) associations included other substance use  
380 disorders (e.g., alcohol-related disorders,  $OR=1.27$ ,  $p=1.53E-17$ ), medical conditions strongly  
381 associated with TUD (e.g., chronic airway obstruction,  $OR=1.19$ ,  $p=5.73E-21$ ) and other  
382 psychiatric conditions (e.g., depression,  $OR=1.06$ ,  $p=8.60E-06$ ). These remained significant

383 after accounting for TUD diagnosis (**Supplementary Table 34**). We also noted associations  
384 across multiple other medical categories, including endocrine/metabolic (e.g., morbid obesity,  
385 OR=1.09,  $p=1.06E-07$ ; type 2 diabetes, OR=1.06,  $p=1.78E-05$ ), digestive (e.g., diseases of  
386 esophagus, OR=1.05,  $p=3.45E-06$ ), circulatory (e.g., ischemic heart disease, OR=1.07,  
387  $p=7.43E-07$ ) and neurologic (e.g., pain, OR=1.07,  $p=9.95E-07$ ), among others (**Supplementary**  
388 **Table 34**). Compared to FTND PGS, TUD PGS were more strongly associated across virtually  
389 all domains, including TUD (**Supplementary Figure 9**). We repeated the TUD PGS analyses in  
390 a BioVU cohort of AA individuals using the TUD-AA meta-analysis results. As expected, TUD  
391 was the strongest and most significant (OR=1.19,  $p=4.03E-06$ ) association (**Supplementary**  
392 **Table 35**).

393 *Yale-Penn sample.* We next extended the analyses to a deeply characterized sample recruited  
394 for genetic studies of substance use disorders: the Yale-Penn sample.<sup>49</sup> We examined the  
395 association between PGS for TUD and hundreds of other traits derived from a comprehensive  
396 psychiatric interview, the Semi-Structured Assessment for Drug Dependence and Alcoholism  
397 (**SSADDA**). TUD-EUR and TUD-AA PGS were strongly associated with many substance use  
398 traits, including nicotine dependence as defined via a Diagnostic and Statistical Manual of  
399 Mental Disorders (**DSM**) diagnosis in both the EUR (OR=1.71,  $p=2.51E-41$ ; **Figure 6b**;  
400 **Supplementary Table 36**) and AA cohorts (OR=1.12,  $p=8.10E-04$ ), respectively, although the  
401 latter association did not survive multiple testing correction (**Supplementary Table 37**). Again,  
402 compared to FTND PGS, TUD-EUR PGS was more strongly associated across virtually all  
403 domains, including nicotine dependence (Nagelkerke's  $R^2=0.101$  vs 0.062; **Supplementary**  
404 **Table 38, Supplementary Figure 10**), again emphasizing the value of collecting information on  
405 later stages of vulnerability or more severe phenotypes, such as TUD.

406



407  
 408 **Figure 6. TUD PGS PheWAS in the (a) Mayo Clinic, (b) Yale-Penn, and (c) ABCD**  
 409 **European cohorts.** Only selected FDR-significant traits are shown. The exact values for each  
 410 association and extended lists of traits can be found in **Supplementary Tables 34, 36 and 39.**  
 411  
 412 *Adolescent Brain Cognitive Development (ABCD) cohort.* Lastly, we extended our polygenic  
 413 analyses to a drug naïve developmental sample (9-11 years of age at recruitment; analytic  
 414 N=52 to 5,556). We concentrated on 12 traits that showed significant genetic correlations in the  
 415 adult samples (**Supplementary Table 39, Figure 6c**). Although tobacco exposure was  
 416 uncommon in this pediatric population (2.30% prevalence), externalizing behaviors, which  
 417 emerge in childhood and are strong correlates of substance use, were available. After correcting  
 418 for multiple testing, TUD PGS was significantly ( $p < 4.00E-03$ ) associated with externalizing  
 419 behaviors (i.e., Child Behavior Check List [CBCL] externalizing scores,  $\beta = 0.06$ ,  $p = 3.15E-06$ ;  
 420 CBCL ADHD scores,  $\beta = 0.05$ ,  $p = 1.58E-04$ ), as well as internalizing (i.e., suicide attempt,  $\beta = 0.05$ ,  
 421  $p = 7.01E-04$ , CBCL depression scores,  $\beta = 0.04$ ,  $p = 1.73E-03$ ), cognitive ability ( $\beta = 0.003$ ,  
 422  $p = 6.35E-05$ ), neighborhood deprivation ( $\beta = 0.03$ ,  $p = 2.53E-03$ ), and weight-related phenotypes



423 (i.e., BMI,  $\beta=0.06$ ,  $p=4.44E-05$ ; weight,  $\beta=0.04$ ,  $p=2.98E-03$ ). Notably, these children were not  
424 chronically exposed to tobacco; therefore, we would speculate that these associations are not a  
425 consequence of smoking but rather may underlie overlapping genetic architectures among the  
426 traits studied that predate use of tobacco.

427

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

434

## 435 **Discussion**

436 Uncovering the genetic underpinnings of individual differences in TUD liability can  
437 advance diagnosis, prevention, and treatment efforts for a disorder of enormous public health  
438 significance. GWAS have uncovered multiple associations with tobacco use, but findings for  
439 tobacco dependence or disorder have been limited due to the difficulty of characterizing large  
440 numbers of individuals using a gold-standard research or clinical diagnosis. Here we present the  
441 first multi-ancestry GWAS of TUD using data from EHR, as a complementary strategy for  
442 ascertainment. In less than four months, and leveraging data from the PsycheMERGE  
443 consortium, we gathered TUD-EHR data for 898,680 individuals. The number of GWAS signals,  
444 enrichment in relevant pathways and tissues, and genetic overlap with nicotine-related traits  
445 provide proof of principle that EHR can serve as a complementary tool to study TUD genetics.



446 Our findings demonstrate that TUD-EHR was genetically correlated with traits derived  
447 from traditionally ascertained cohorts, including nicotine dependence via FTND and smoking  
448 cessation, providing clear evidence that the signal captured by TUD phecodes is valid. Of note,  
449 the genetic correlation between TUD-EHR and other smoking behaviors, such as number of  
450 cigarettes smoked per day (**CPD**), although significant and positive, was moderate in magnitude  
451 ( $r_g=0.43$ ), suggesting that the genetic architectures of consumption and misuse may be distinct.  
452 This is in contrast to earlier observations for FTND and CPD, where the genetic correlation was  
453 almost at unity ( $r_g=0.95$ ).<sup>27</sup> This shows that TUD captures features beyond the frequency of  
454 smoking or severity of nicotine dependence. Although FTND and TUD were more strongly  
455 correlated ( $r_g=0.61$ ), in general, we observed that TUD PGS was more predictive of DSM-  
456 defined tobacco dependence and a plethora of comorbid traits in the Yale-Penn sample, than  
457 FTND PGS. The only exception was for time-to-first cigarette in the morning, which was more  
458 strongly associated with FTND PGS, likely because time-to-first cigarette is one of the FTND  
459 items. Overall, this emphasizes the need to continue measuring the full spectrum of addiction  
460 liability,<sup>50</sup> such as CPD, and more severe phenotypes, such as TUD, to account for the distinct  
461 biological factors relevant at each stage.

462 Common SNPs were able to account for a fraction (7%) of the overall heritability of TUD  
463 (40-60%) as determined by prior family and twin studies.<sup>9,11</sup> The multi-ancestral meta-analysis  
464 identified 72 independent loci, 13 times the number previously reported for nicotine  
465 dependence.<sup>27</sup> These include corroborative support for the involvement of nicotinic acetylcholine  
466 receptor genes (*CHRNA5-A3-B4*, *CHRNA2*, *CHRNA4*), which have been consistently  
467 associated with smoking behaviors,<sup>20</sup> particularly in studies of self-reported CPD.<sup>13</sup> We also  
468 identified polymorphisms in genes implicated in nicotine clearance, like *CYP2A6*, previously  
469 linked to heavy smoking.<sup>51</sup> Other variants identified are in genes that modulate dopaminergic  
470 and glutamatergic neurotransmission, compromising reward-based learning and facilitating

471 drug-seeking behavior, and in *BDNF*, which is involved in memory consolidation processes,<sup>52</sup>  
472 and a well-studied candidate gene in addiction.<sup>53</sup> These and other candidates supported by  
473 TUD (e.g., *PDE4B*) were genetically correlated with other addiction phenotypes,<sup>36</sup> emphasizing  
474 the shared neurobiological mechanisms of addiction.

475 Downstream analyses prioritized genes and drug candidates that could be used for  
476 follow-up mechanistic studies in model organisms. Specifically, we identified “core” genes that  
477 could be “pleiotropic hotspots” associated with multiple traits. One was glutathione peroxidase-1  
478 (*GPX1*), which is involved in oxidative stress. Intriguingly, it has been reported that glutathione  
479 peroxidase-1 protects against lung inflammation induced by smoking in mice, and agents that  
480 mimic this action (e.g., ebselen), which restore GPX1 activity in situations of extreme oxidative  
481 stress, can protect from lung inflammation induced by smoking.<sup>54</sup> Another was *GMPPB*, which  
482 has been associated with accelerated lung aging and e-cigarette smoking.<sup>55</sup> *NT5C2* is involved  
483 in maintaining cellular nucleotide balance, and was associated with schizophrenia<sup>55</sup> and  
484 smoking behaviors in an exome-wide association study.<sup>56</sup> These genes showed a consistent  
485 causal effect based on colocalization analyses (here and previously<sup>57</sup>), suggesting that they  
486 could confer TUD risk by modulating regulated gene expression and protein abundance in the  
487 brain.

488 The enrichment of TUD in brain tissues further supports TUD as a brain disorder, long  
489 supported by neuroscience and more recently by genetics.<sup>58</sup> We provide suggestive evidence  
490 for the involvement of the cerebellum in TUD, along with other regions that have long been  
491 studied in relation to addiction such as the fronto-striatal loop, hippocampus, and amygdala.<sup>59</sup>

492 Genetic correlations revealed substantial levels of pleiotropy with traits that often co-  
493 occur with TUD, including other substance use and psychiatric disorders. These associations  
494 were particularly evident in the Yale-Penn sample,<sup>49</sup> which has comprehensive phenotypic data  
495 for substance use disorders. In adult patients from the Mayo Clinic, we replicated the

496 associations with substance and other psychiatric disorders, extending them to medical  
497 disorders, such as HIV, heart disease, and pain, some of which, like respiratory conditions,  
498 likely reflect chronic smoking. The positive associations between genetic liability for TUD and  
499 other outcomes, such as BMI or other internalizing/externalizing problems in tobacco-naive  
500 children (ABCD), may also reflect true biological relationships. Although we are far from  
501 untangling this complex web of genetic and non-genetic correlations, the extensive phenotypic  
502 spectrum associated with TUD is undeniable.

503         Currently, developing new therapeutics for TUD is viewed as risky because of a lack of  
504 high-quality targets, historically low success rates, and unintended side effects. Although genes  
505 identified in our GWAS, including *CHRNA7*, *CHRNA5*, *CHRNA4*, and *CHRNA2*, might moderate  
506 the effect of varenicline, a smoking cessation treatment that operates as a partial agonist at the  
507 nicotine acetylcholine  $\alpha 2\beta 4$  receptor,<sup>60</sup> varenicline (along with other medications such as  
508 nicotine replacement therapies) has limited efficacy or adverse effects.<sup>61,62</sup> In a proof-of-principle  
509 study, So et al.<sup>63</sup> identified several repurposing candidates for treating psychiatric disorders by  
510 connecting imputed transcriptomic profiles from GWAS data to drug-induced gene expression  
511 profiles. Using this approach, we identified hundreds of potential drug candidates predicted to  
512 significantly reverse the TUD transcriptomic profile. These included norepinephrine reuptake  
513 inhibitors (e.g., amoxapine) and antipsychotics (e.g., clozapine), pointing to convergent  
514 molecular mechanisms between TUD and other psychiatric disorders that are the usual target of  
515 these agents, replicating prior observations.<sup>64</sup> The potential therapeutic utility of anti-  
516 inflammatory or blood glucose lowering medications were also suggested by our analyses.  
517 Although, to date, no repurposed drugs have been developed for treating SUDs based on  
518 GWAS data, this is an important potential path forward, particularly for SUDs, where few  
519 effective pharmacotherapies are available.

520 Future research may address some of the limitations of our study. Prior work has  
521 demonstrated that ICD codes have a low sensitivity for current tobacco use, but may have a  
522 reasonable specificity for this common behavior.<sup>65</sup> Our results appeared to be robust to  
523 moderate levels of misclassification, particularly in controls, as detected by the pairing with self-  
524 reported questionnaire data. Although studies that systematically evaluate the effect of  
525 removing potentially misclassified individuals are needed, we chose not to remove them in this  
526 study because not all individuals had concomitant survey data available. This questionnaire  
527 data, along with other forms of EHR data (e.g., clinical notes), may help capture additional  
528 phenotypes, including the response to treatment or the ability to successfully quit smoking  
529 without formal treatment. Longitudinal data from EHR, with data collection spanning the period  
530 prior to and following the onset of substance use and SUD, are particularly valuable for studying  
531 the timing of onset, within-person change, and application of time-varying effects, which will help  
532 to differentiate causation from correlational findings. The advent of single-cell transcriptomics,  
533 larger QTL databases in more specific cell types, and the inclusion of more ancestrally diverse  
534 samples will improve the interpretability of associated loci. Although we have included diverse  
535 cohorts, our study lacked many major ancestral groups such as East Asians and South Asians.  
536 Lastly, other forms of genetic variation, such as rare single variants<sup>66</sup> or structural  
537 polymorphisms<sup>67</sup> are likely to account for much of the “missing heritability” in genetic risk for  
538 TUD.

539 In sum, this work demonstrates that EHR is a viable and cost efficient complementary  
540 alternative to rigorous clinical ascertainment for genetic studies of TUD, similar to other SUD  
541 traits. At various levels of analysis, this study identifies and prioritizes previously unidentified  
542 genes of potential interest. TUD shares biological processes common to many SUDs and is one  
543 among a number of highly correlated psychiatric and medical disorders. We anticipate that

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

546

## 547 **Methods**

548 **Smoking phenotypes and cohorts.** We defined cases as patients who received at least two  
549 TUD ICD-9 or -10 codes (corresponding to the phecode definition) in their medical records, and  
550 controls as patients who had no TUD diagnosis code (**Supplementary Table 2**). In UKBB only,  
551 cases were defined as having 1 ICD-10 code for TUD, and controls had none.<sup>43</sup> Additionally, we  
552 required controls to be 18 years of age or older at time of analysis (04/2022). Patients younger  
553 than 18 years were excluded because they may not yet have reached the age of TUD  
554 diagnosis. We examined the sensitivity of our TUD phenotyping using the patients' self-reported  
555 tobacco use survey when available (**Supplementary Table 3**, list of smoking traits).

556 Our data sources included registries from five health systems linked to biobanks:  
557 Vanderbilt University Medical Center's (VUMC) biobank (BioVU), Mass General Brigham  
558 Biobank (MGBB), Penn Medicine BioBank (PMBB), Million Veteran Program (MVP), and UK  
559 Biobank (UKBB). There were 46,905 (EUR) patients from VUMC, 22,268 (EUR) patients from  
560 MGBB, 39,087 patients from PMBB (28,999 EUR and 10,088 AA), 545,530 patients from MVP  
561 (396,833 EUR, 104,332 AA, 44,365 LA), and 244,890 participants from UKBB. Details of each  
562 registry, including demographics and data sources, are listed in the **Supplementary Table 2**.

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

568 group sample, and in MVP we performed additional GWAS of the African American ancestral  
569 sample and the Latin American ancestral group sample. Each of the univariate GWAS covaried  
570 for 10 genetic ancestry principal components, age, sex, number of ICD codes and length of  
571 record. The summary statistics for TUD in UKBB were downloaded from the GWAS atlas  
572 (<https://atlas.ctglab.nl/traitDB/3439>).

573 *BioVU*. We used de-identified clinical data from individuals in BioVU. Genotype data  
574 were generated using the Illumina Multi-Ethnic Genotype Array (MEGAEX) for 72,824  
575 individuals. Details on the quality control process have been described elsewhere.<sup>70</sup> Genotypes  
576 were filtered for SNP (<0.95) and individual (<0.98) call rates, sex discrepancies, and excessive  
577 heterozygosity ( $|F_{het}| > 0.2$ ).<sup>71</sup> The sample was then filtered for cryptic relatedness by removing  
578 one individual of each pair for which  $\text{pihat} > 0.2$ . PCA using FlashPCA2 combined with CEU, YRI  
579 and CHB reference sets from 1000 Genomes Project Phase 3<sup>72</sup> was conducted to determine  
580 European Ancestry. We confirmed the absence of genotyping batch effects using 'batch' as the  
581 phenotype. We imputed genotypes using the Michigan Imputation Server with the reference  
582 panel from the Haplotype Reference Consortium. SNPs were filtered for imputation quality ( $R^2$   
583  $> 0.3$  or INFO  $> 0.95$ ) and converted to hard calls. We restricted the analyses to autosomal SNPs  
584 with minor allele frequency  $> 0.01$ . We removed SNPs that differed by  $> 10\%$  from the 1000  
585 Genomes Project phase 3 CEU set<sup>72</sup> and those with a Hardy Weinberg Equilibrium  $p < 1.00E-10$ .  
586 The resulting data set contained hard-called SNP information for 9,386,383 SNPs in 72,824  
587 individuals of European Ancestry. Controls were also required to have 3 or more years of  
588 medical history with VUMC. These procedures resulted in a total sample of 7,167 cases and  
589 39,738 controls in BioVU. The project was approved by the VUMC Institutional Review Board  
590 (IRB #160302, #172020, #190418).

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

593 rate <0.98, and those that were duplicated across batches, monomorphic, not confidently  
594 mapped to a genomic location, or associated with genotyping batch. Sample-level quality  
595 control filters were applied to remove individuals with a call rate <0.98, excessive autosomal  
596 heterozygosity ( $\pm 3$  standard deviations from the mean), or discrepant self-reported and  
597 genetically inferred sex. PCs of ancestry were calculated using the 1000 Genomes Phase 3  
598 dataset as a reference panel. The Michigan Imputation Server was then used to impute missing  
599 genotypes with the Haplotype Reference Consortium dataset serving as the reference panel.  
600 Imputed genotype dosages were converted to hard-call format and subjected to further quality  
601 control, where SNPs were removed if INFO score <0.8, MAF <0.01, HWE  $p < 1.00E-10$ , or  
602 missingness (variant call rate <0.98). Only unrelated individuals ( $\pi\text{-hat} < 0.2$ ) of European  
603 ancestry were included in the present study. These procedures yielded a final analytic sample of  
604 6,708 cases and 15,560 in the MGBB. The project was approved by the MGBB Institutional  
605 Review Board (IRB #2018P002642).

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



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

619 *MVP*. MVP samples were genotyped using the Affymetrix Axiom Biobank Array.  
620 Samples were removed if they had extreme heterozygosity, call rate <98.5%, sex mismatch, or  
621 >7 relatives. SNPs were removed if they had call rate <0.98 or a Hardy–Weinberg equilibrium  
622 (HWE) threshold of  $p < 1.00E-06$ . Genotype phasing and imputation was performed using  
623 SHAPEIT4 (v.4.1.3)<sup>75</sup> and Minimac4 software<sup>73</sup>, respectively. Biallelic and non-biallelic SNPs  
624 were imputed using the African Genome Resources and 1000 Genomes reference panels.<sup>72</sup>  
625 Ancestry was defined for three mutually exclusive ancestral groups (European, African  
626 American, and Hispanic American) utilizing a previously defined approach harmonizing genetic  
627 ancestry and self-identified ancestry (HARE).<sup>76</sup> SNPs with imputation quality (INFO)  
628 score  $\leq 0.7$ , minor allele frequency (MAF, AA  $\leq 0.005$ ; EUR  $\leq 0.001$ ; HIS  $\leq 0.01$ ),  
629 genotype call rate <0.95, and HWE  $p < 1.00E-06$  were removed. We also excluded one  
630 individual from each pair of related individuals (kinship >0.08, N=31,010). The final sample  
631 comprised 104,332 AAs (cases=43,743), 396,833 EURs (cases=146,771) and 44,365 LAs  
632 (cases=12,277). The Central VA Institutional Review Board (IRB) and site-specific IRBs  
633 approved the MVP study.

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

637 **Meta-analyses and independent variants**. Meta-analyses were conducted using a sample-  
638 size-weighted method in METAL,<sup>79</sup> assuming shared risk effects across ancestries. Effective  
639 sample sizes (N\_Eff), calculated using the formula:  $4/[1/n_{case} + 1/n_{control}]$ , were used to

640 compensate for the imbalance in the ratio of cases to controls.  $N_{Eff}$  were used in all meta-  
641 analyses and all downstream analyses.

642 We conducted four meta-analyses of TUD GWAS summary statistics across the  
643 following datasets: 1) within-ancestry meta-analysis for EUR samples in BioVU, MGBB, PMBB,  
644 MVP, and an additional meta-analysis including UKBB, 2) within-ancestry meta-analysis for AA  
645 in MVP and Penn, and 3) multi-ancestry meta-analysis across all datasets (AA [PMBB, MVP];  
646 EUR [BioVU, MGBB, PMBB, MVP, UKBB]; HA [MVP]). Inflation of test statistics due to  
647 polygenicity or cryptic relatedness was assessed using the LDSC attenuation ratio ((LDSC  
648 intercept - 1)/(mean of association chi-square statistics - 1)). Resulting genome-wide significant  
649 (GWS) loci were defined as those with  $p < 5.00E-08$  with LD  $r^2 > 0.1$ , within a 1MB window, based  
650 on the structure of the Haplotype Reference Consortium (HRC) multi-ancestry reference panel  
651 for the multi-ancestry meta-analysis, or the HRC ancestry-appropriate reference panel  
652 otherwise. GWS loci were examined for heterogeneity across cohorts via the  $I^2$  inconsistency  
653 metric.

654 To identify TUD risk loci and lead SNPs, we performed LD clumping in FUMA<sup>43</sup> using a  
655 range of 3 Mb,  $r^2 > 0.1$ , and the respective ancestry 1000 Genome reference panel.<sup>72</sup> Genomic  
656 risk loci that were located <1Mb apart were incorporated into a single locus. For loci that  
657 harbored multiple variants, we used COJO in GCTA<sup>80</sup> to define independent variants by  
658 conditioning them on the most significant variant within each locus. Following conditioning,  
659 significant variants ( $p < 5.00E-08$ ) were considered independent.

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

664 **Multi-ancestry fine-mapping analyses.** We used PAINTOR v3.1<sup>82</sup> to perform multi-ancestry  
665 fine mapping for the two risk loci identified in both the TUD-EUR and TUD-AA metaGWAS. For  
666 each locus, we extracted SNPs with an absolute value of Z-score larger than 3.9 within a 1Mb  
667 region of the lead SNP. As suggested by PAINTOR, we created the AA and EUR LD matrices  
668 using the 1000 Genome phase 3 reference panel<sup>72</sup>. We calculated the probability of each SNP  
669 being the causal variant, assuming that each locus has two causal variants.

670 **Gene-based and pathway analyses.** We conducted bioannotation and bioinformatic analyses  
671 to further characterize the loci identified by the TUD-EHR GWAS (**Supplementary Methods**).  
672 We used the default version (v1.3.6a) of the FUMA web-based platform<sup>43</sup> to identify  
673 independent SNPs ( $r^2 < 0.10$ ) and to study their functional consequences. We also used MAGMA  
674 v1.08<sup>43,44</sup> to perform competitive gene-set and pathway analyses. SNPs were mapped to 19,532  
675 protein-coding genes from Ensembl (build 85). We applied a Bonferroni correction based on the  
676 total number of genes tested ( $p < 2.56E-06$ ). Gene sets were obtained from Msigdb v7.0  
677 (“Curated gene sets”, “GO terms”). We also used Hi-C coupled MAGMA (H-MAGMA<sup>45</sup>) to  
678 assign non-coding (intergenic and intronic) SNPs to genes based on their chromatin  
679 interactions. Exonic and promoter SNPs were assigned to genes based on physical position. H-  
680 MAGMA uses four Hi-C datasets, which were derived from fetal brain, adult brain, iPSC-derived  
681 neurons, and iPSC-derived astrocytes (<https://github.com/thewonlab/H-MAGMA>). We applied a  
682 Bonferroni correction based on the total number of gene-tissue pairs tested ( $p < 9.55E-07$ ).

683 **S-MultiXcan/S-PrediXcan.** We used S-MultiXcan v0.7.0 (an extension of S-PrediXcan v0.6.2<sup>46</sup>)  
684 to identify specific eQTL-linked genes associated with TUD. This approach uses genetic  
685 information to predict transcript abundance in 13 brain tissues, and tests whether the predicted  
686 transcripts correlate with TUD. S-PrediXcan uses pre-computed tissue weights from the  
687 Genotype-Tissue Expression (GTEx) v8 project database (<https://www.gtexportal.org/>) as the  
688 reference transcriptome dataset. For S-PrediXcan and S-MultiXcan analyses, we chose to use

689 sparse (elastic net) prediction models, which are available at <http://predictdb.hakyimlab.org/>. We  
690 applied a conservative Bonferroni correction based on the total number of gene-tissue pairs  
691 tested (14,198 gene-tissue pairs tested;  $p < 3.52E-06$ ).

692 **Partitioning Heritability Enrichment.** We used LDSC to partition TUD-EUR  $h^2_{SNP}$  and  
693 examined the enrichment based on several functional genomic annotation models.<sup>83,84</sup> In the  
694 baseline model, we examined 75 overlapping functional annotations comprising genomic,  
695 epigenomic and regulatory features. We also analyzed ten overlapping cell-type groups derived  
696 from 220 cell-type-specific annotations in four histone marks: methylated histone H3 Lys4  
697 (H3K4me1), trimethylated histone H3 Lys4 (H3K4me3), acetylated histone H3 Lys4 (H3K4ac)  
698 and H3K27ac. Enriched cell-type categories were analyzed based on annotations obtained from  
699 H3K4me1-imputed, gapped peak data generated by the Roadmap Epigenomics Mapping  
700 Consortium.<sup>85</sup> We removed multi-allelic and major histocompatibility complex region variants,  
701 and only report categories enriched after Bonferroni correction.

702 **Tissue Enrichment Analysis.** We used the LDSC package to conduct cell type specific  
703 heritability analysis (<https://www.nature.com/articles/s41588-018-0081-4>). In this analysis, we  
704 applied stratified LD score regression on the TUD-EUR meta-analysis summary statistics with  
705 sets of specifically expressed genes in various tissues from GTEx<sup>86-88</sup> to identify TUD-relevant  
706 tissues. We applied a conservative Bonferroni correction based on the number of tissues  
707 simultaneously tested (205 tissues tested,  $p < 2.44E-04$ ). We also used MAGMA v1.08 gene-  
708 property analysis of expression data from GTEx (54 tissue types) and BrainSpan (29 brain  
709 samples at different age) in FUMA v1.3.6a<sup>75</sup> to test the relationships between tissue specific  
710 gene expression profiles and TUD-gene associations.

711 **Cell type-specific expression of TUD risk genes.** We performed cell-type specific analyses  
712 implemented in FUMA, using data from nine single-cell RNA sequencing data sets from human  
713 brain (data sets listed in the **Supplementary Material**). The method is described in detail in

714 Watanabe et al.,<sup>43</sup> and uses MAGMA gene-property analysis to test for association between cell  
715 specific gene expression and TUD-gene association. Conditional analyses for multiple testing  
716 are applied to correct for all tested cell types across datasets.

717 **PWAS/TWAS.** To identify proteins whose genetically regulated expression is associated with  
718 TUD, we performed PWAS analyses by integrating TUD GWAS summary statistics and  
719 precomputed pQTLs from discovery (Banner)<sup>89,90</sup> and validation (ROSMAP)<sup>91,92</sup> datasets using  
720 the FUSION pipeline (<http://gusevlab.org/projects/fusion/>).<sup>47</sup> Next, TWAS was performed using  
721 gene and splicing expression profiles measured in the adult DLPFC and gene expression  
722 profiles from the frontal cortex. Human brain transcriptome data, used as expression reference  
723 panels, were obtained from the CMC<sup>91</sup> and GTEx frontal cortex v7.<sup>47,86</sup> All tests were Bonferroni  
724 corrected for multiple testing ( $\alpha = 0.05/N$  genes tested).

725 Of the overlapping findings across independent TWAS or PWAS datasets, colocalization  
726 analysis (in FUSION<sup>47,93</sup>) was used to determine whether SNPs mediate the association with  
727 TUD via effects on gene and protein expression. A posterior colocalization probability (PP) of  
728 80% was used to indicate a shared causal signal.

729 **BrainXcan.** We used the BrainXcan package (<https://github.com/hakyimlab/brainxcan>)<sup>94</sup> to  
730 predict the association between the TUD phenotype and brain features. This approach uses  
731 genetically determined brain image-derived phenotypes (IDPs) to test brain region association  
732 with the TUD phenotype via linear regression. IDPs were constructed by training genetic  
733 predictors on original IDPs from MRI images via ridge regression.<sup>94</sup> IDPs were retrieved from  
734 the BrainXcan database (<https://zenodo.org/record/4895174>). Only significant IDP associations  
735 with TUD that survived a Bonferroni correction are reported (93 IDPs tested;  $p < 1.92E-04$ ).

736 **Drug repurposing.** Our signature matching technique used data from the Library of Integrated  
737 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,<sup>36</sup> 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 *metaphor* 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<sup>48</sup> and  
760 the cross-ancestry  $r_g$ s for TUD across population groups using POPCORN.<sup>48</sup> We used the  
761 ancestry-specific 1000 Genomes Project phase 3<sup>76</sup> data as the LD references.

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

771 **mtCOJO.** We used mtCOJO<sup>95</sup> to individually condition the TUD-EUR summary statistics on loci  
772 associated with other comorbid traits, including alcohol dependence, cannabis use disorder and  
773 opioid use disorder. This analysis allowed us to examine whether the genetic associations with  
774 TUD would be preserved when controlling for those covariate phenotypes. To test as many  
775 SNPs while preserving computational efficiency, we used a  $p$  value threshold of 5.00E-07,  
776 5.00E-07, 1.00E-07, respectively, for alcohol dependence, cannabis use disorder, and opioid  
777 use disorder. We then computed genetic correlations using the TUD summary statistics  
778 adjusted for the covariates of interest.

779 **Unsupervised learning to determine TUD clustering.** Previous studies have shown that  
780 consumption and misuse/dependence phenotypes have a distinct genetic architecture. To  
781 explore whether the TUD meta-analysis clustered more with consumption or  
782 misuse/dependence phenotypes, we used a data-driven unsupervised machine learning method  
783 known as agglomerative hierarchical clustering analysis (**HCA**).<sup>96</sup> HCA forms clusters iteratively  
784 by creating groups and successively joining or splitting those groups based on a prespecified  
785 algorithm.<sup>96</sup> Agglomerative nesting (AGNES) is a bottom-up process focused on individual traits  
786 to structure. Agglomerative clustering was chosen as this allowed us to compare different



787 algorithms to maximize for the dissimilarity on each branch, with Ward's minimum variance  
788 method performing best. All models were fit in R using the *cluster* package (version 2.1.4).<sup>96</sup>

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

### 795 **Phenome-wide association studies (PheWAS)**

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

812 (PI\_HAT>0.1875) or >25000 related samples (PI\_HAT>0.08); the relatedness analysis was  
813 performed iteratively until no such samples remained. For each pair with an estimated 2<sup>nd</sup>  
814 degree or higher relatedness, we removed the individual with shorter length of EHR. Finally,  
815 PRSs were calculated using LDpred2<sup>98</sup> I using the auto feature in the bigsnpr (v1.10.4) R  
816 package.

817 **Yale-Penn.** We performed PheWAS in the Yale-Penn sample,<sup>49</sup> which is a deeply phenotyped  
818 cohort using the Semi-Structured Assessment for Drug Dependence and Alcoholism, a detailed  
819 psychiatric instrument used to assess physical, psychosocial, and psychiatric manifestations of  
820 SUDs and comorbid psychiatric traits.<sup>99,100</sup> This comprehensive interview includes more than  
821 3,500 items representing lifetime diagnostic criteria for the DSM-IV,<sup>101</sup> DSM-5<sup>102</sup> SUDs and  
822 DSM-IV<sup>101</sup> psychiatric disorder history. Genotyping and quality control for this cohort have been  
823 extensively described.<sup>49,103</sup>

824 Using PRS-Continuous shrinkage software (PRS-CS),<sup>104</sup> PRSs were calculated for TUD.  
825 We used the default setting in PRS-CS to estimate the shrinkage parameters and fixed the  
826 random seed to 1 for reproducibility. To identify associations between the PRS for TUD and  
827 clinical phenotypes, we performed a PheWAS by fitting logistic regression models for binary  
828 phenotypes and linear regression models for continuous phenotypes. Analyses were conducted  
829 using the PheWAS v0.12 R package<sup>105</sup> adjusting for sex, median age and the first ten PCs  
830 within each genetic ancestry. Bonferroni correction was applied for each ancestral-specific  
831 analysis to account for multiple testing ( $p < 7.25E-05$ ).

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

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

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

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

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

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

861 supported by R01 MH113362. HRK was supported by the Veterans Integrated Service Network  
862 4 Mental Illness Research, Education and Clinical Center. RLK was supported by NIAAA K01  
863 AA028292. The content is solely the responsibility of the authors and does not necessarily  
864 represent the official views of the National Institutes of Health.

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

868 BioVU The dataset(s) used for the analyses described were obtained from Vanderbilt  
869 University Medical Center's BioVU which is supported by numerous sources: institutional  
870 funding, private agencies, and federal grants. These include the NIH funded Shared  
871 Instrumentation Grant S10RR025141; and CTSA grants UL1TR002243, UL1TR000445, and  
872 UL1RR024975. Genomic data are also supported by investigator-led projects that include  
873 U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378,  
874 U19HL065962, R01HD074711; and additional funding sources listed at  
875 <https://victr.vumc.org/biovu-funding/>.

876 This research is based on data from the Million Veteran Program, Office of Research  
877 and Development, Veterans Health Administration, and was supported by funding from the  
878 Department of Veterans Affairs Office of Research and Development, Million Veteran Program  
879 Grant #I01 BX004820. This publication does not represent the views of the Department of  
880 Veterans Affairs or the United States Government.

881 We acknowledge the Penn Medicine BioBank (PMBB) and the Mayo Clinic Biobank for  
882 providing data and thank the patient-participants of Penn Medicine and Mayo Clinic who  
883 consented to participate in this research program. We would also like to thank the Penn  
884 Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for

885 analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman  
886 School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National  
887 Center for Advancing Translational Sciences of the National Institutes of Health under CTSA  
888 award number UL1TR001878.

889 Data used in the preparation of this article were obtained from the Adolescent Brain  
890 Cognitive Development (ABCD) Study (<https://abcdstudy.org>), held in the NIMH Data Archive  
891 (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age  
892 9-10 and follow them over 10 years into early adulthood. The ABCD Study is supported by the  
893 National Institutes of Health and additional federal partners under award numbers  
894 U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117,  
895 U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123,  
896 U24DA041147, U01DA041093, and U01DA041025. A full list of supporters is available at  
897 <https://abcdstudy.org/federal-partners.html>. A listing of participating sites and a complete listing  
898 of the study investigators can be found at [https://abcdstudy.org/Consortium\\_Members.pdf](https://abcdstudy.org/Consortium_Members.pdf).  
899 ABCD consortium investigators designed and implemented the study and/or provided data but  
900 did not necessarily participate in analysis or writing of this report. This manuscript reflects the  
901 views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium  
902 investigators.

903 We would also like to thank The Externalizing Consortium for sharing the GWAS  
904 summary statistics of externalizing. The Externalizing Consortium: Principal Investigators:  
905 Danielle M. Dick, Philipp Koellinger, K. Paige Harden, Abraham A. Palmer. Lead Analysts:  
906 Richard Karlsson Linnér, Travis T. Mallard, Peter B. Barr, Sandra Sanchez-Roige. Significant  
907 Contributors: Irwin D. Waldman. The Externalizing Consortium has been supported by the  
908 National Institute on Alcohol Abuse and Alcoholism (R01AA015416 -administrative supplement),  
909 and the National Institute on Drug Abuse (R01DA050721). Additional funding for investigator

910 effort has been provided by K02AA018755, U10AA008401, P50AA022537, as well as a  
911 European Research Council Consolidator Grant (647648 EdGe to Koellinger). The content is  
912 solely the responsibility of the authors and does not necessarily represent the official views of  
913 the above funding bodies. The Externalizing Consortium would like to thank the following groups  
914 for making the research possible: 23andMe, Add Health, Vanderbilt University Medical Center's  
915 BioVU, Collaborative Study on the Genetics of Alcoholism (COGA), the Psychiatric Genomics  
916 Consortium's Substance Use Disorders working group, UK10K Consortium, UK Biobank, and  
917 Philadelphia Neurodevelopmental Cohort.

### 918 **Contributions**

919 The corresponding author (S.S-R) conceived the idea for the paper and wrote and edited the  
920 manuscript. Other contributing authors contributed analyses (S.T., M.V.J., B.P., H.L., T.T.L.,  
921 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.,  
922 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  
923 edited the paper. We thank Bryan Quach and Jesse Marks for their help in supplying portions of  
924 the data needed to create Supplementary Figure 1.

### 925 **Ethics declarations**

926 Dr. Palmer is on the scientific advisory board of Vivid Genomics for which he receives stock  
927 options. Dr. Smoller is a member of the Scientific Advisory Board of Sensorium Therapeutics  
928 (with equity) and has received grant support from Biogen, Inc. He is PI of a collaborative study  
929 of the genetics of depression and bipolar disorder sponsored by 23andMe for which 23andMe  
930 provides analysis time as in-kind support but no payments. Dr. Kranzler is a member of advisory  
931 boards for Dicerna Pharmaceuticals, Sophrosyne Pharmaceuticals, and Enthion  
932 Pharmaceuticals; a consultant to Sobrera Pharmaceuticals; the recipient of research funding  
933 and medication supplies for an investigator-initiated study from Alkermes; a member of the  
934 American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative, which was  
935 supported in the last three years by Alkermes, Dicerna, Ethypharm, Lundbeck, Mitsubishi,  
936 Otsuka, and Pear Therapeutics; and with Dr Gelernter, a holder of U.S. patent 10,900,082 titled:  
937 "Genotype-guided dosing of opioid agonists," issued 26 January 2021. The other authors  
938 declare no competing interests.





940 **References**

- 941 1. Centers for Disease Control and Prevention (CDC). *Health Effects of Cigarette Smoking*.  
942 [https://www.cdc.gov/tobacco/data\\_statistics/fact\\_sheets/health\\_effects/effects\\_cig\\_smokin](https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm)  
943 [g/index.htm](https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm) (2021).
- 944 2. Oliver, J. A. & Foulds, J. Association Between Cigarette Smoking Frequency and Tobacco  
945 Use Disorder in U.S. Adults. *Am. J. Prev. Med.* **60**, 726–728 (2021).
- 946 3. WHO. The top 10 causes of death. *World Health Organization* [https://www.who.int/news-](https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death)  
947 [room/fact-sheets/detail/the-top-10-causes-of-death](https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death).
- 948 4. Benowitz, N. L. & Liakoni, E. Tobacco use disorder and cardiovascular health. *Addiction*  
949 **117**, 1128–1138 (2022).
- 950 5. Kalman, D., Morissette, S. B. & George, T. P. Co-Morbidity of Smoking in Patients with  
951 Psychiatric and Substance Use Disorders. *Am. J. Addict. Am. Acad. Psychiatr. Alcohol.*  
952 *Addict.* **14**, 106–123 (2005).
- 953 6. Tobacco use disorder and the lungs - McRobbie - 2021 - *Addiction* - Wiley Online Library.  
954 <https://onlinelibrary.wiley.com/doi/10.1111/add.15309>.
- 955 7. Ziedonis, D., Das, S. & Larkin, C. Tobacco use disorder and treatment: new challenges  
956 and opportunities. *Dialogues Clin. Neurosci.* **19**, 271–280 (2017).
- 957 8. Kendler, K. S., Schmitt, E., Aggen, S. H. & Prescott, C. A. Genetic and Environmental  
958 Influences on Alcohol, Caffeine, Cannabis, and Nicotine Use From Early Adolescence to  
959 Middle Adulthood. *Arch. Gen. Psychiatry* **65**, 674–682 (2008).
- 960 9. Do, E. K. *et al.* Genetic and Environmental Influences on Smoking Behavior across  
961 Adolescence and Young Adulthood in the Virginia Twin Study of Adolescent Behavioral  
962 Development and the Transitions to Substance Abuse Follow-Up. *Twin Res. Hum. Genet.*  
963 *Off. J. Int. Soc. Twin Stud.* **18**, 43–51 (2015).

- 964 10. Agrawal, A., Budney, A. J. & Lynskey, M. T. The Co-occurring Use and Misuse of  
965 Cannabis and Tobacco: A Review. *Addict. Abingdon Engl.* **107**, 1221–1233 (2012).
- 966 11. Agrawal, A. *et al.* The genetics of addiction—a translational perspective. *Transl. Psychiatry*  
967 **2**, e140–e140 (2012).
- 968 12. Sullivan, P. F. & Kendler, K. S. The genetic epidemiology of smoking. *Nicotine Tob. Res.* **1**,  
969 S51–S57 (1999).
- 970 13. Saunders, G. R. B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol  
971 use. *Nature* **612**, 720–724 (2022).
- 972 14. Larsson, S. C. & Burgess, S. Appraising the causal role of smoking in multiple diseases: A  
973 systematic review and meta-analysis of Mendelian randomization studies. *eBioMedicine*  
974 **82**, (2022).
- 975 15. Yuan, S., Michaëlsson, K., Wan, Z. & Larsson, S. C. Associations of Smoking and Alcohol  
976 and Coffee Intake with Fracture and Bone Mineral Density: A Mendelian Randomization  
977 Study. *Calcif. Tissue Int.* **105**, 582–588 (2019).
- 978 16. Mahedy, L. *et al.* Testing the association between tobacco and cannabis use and cognitive  
979 functioning: Findings from an observational and Mendelian randomization study. *Drug*  
980 *Alcohol Depend.* **221**, 108591 (2021).
- 981 17. Zhou, H. *et al.* Association of *OPRM1* Functional Coding Variant With Opioid Use Disorder:  
982 A Genome-Wide Association Study. *JAMA Psychiatry* **77**, 1072 (2020).
- 983 18. Wootton, R. E. *et al.* Evidence for causal effects of lifetime smoking on risk for depression  
984 and schizophrenia: a Mendelian randomisation study. *Psychol. Med.* **50**, 2435–2443  
985 (2020).
- 986 19. Harrison, R., Munafò, M. R., Davey Smith, G. & Wootton, R. E. Examining the effect of  
987 smoking on suicidal ideation and attempts: triangulation of epidemiological approaches. *Br.*  
988 *J. Psychiatry* **217**, 701–707.

- 989 20. Xu, K. *et al.* Genome-wide association study of smoking trajectory and meta-analysis of  
990 smoking status in 842,000 individuals. *Nat. Commun.* **11**, 5302 (2020).
- 991 21. Sanchez-Roige, S. *et al.* Genome-wide association study of alcohol use disorder  
992 identification test (AUDIT) scores in 20 328 research participants of European ancestry:  
993 GWAS of AUDIT. *Addict. Biol.* **24**, 121–131 (2019).
- 994 22. Kranzler, H. R. *et al.* Genome-wide association study of alcohol consumption and use  
995 disorder in 274,424 individuals from multiple populations. *Nat Commun* **10**, 1499 (2019).
- 996 23. Mallard, T. T. & Sanchez-Roige, S. Dimensional Phenotypes in Psychiatric Genetics:  
997 Lessons from Genome-Wide Association Studies of Alcohol Use Phenotypes. *Complex*  
998 *Psychiatry* **7**, 45–48 (2021).
- 999 24. Mallard, T. T. *et al.* Item-Level Genome-Wide Association Study of the Alcohol Use  
1000 Disorders Identification Test in Three Population-Based Cohorts. *Am. J. Psychiatry*  
1001 *appi.ajp.2020.2* (2021) doi:10.1176/appi.ajp.2020.20091390.
- 1002 25. Sanchez-Roige, S. Emerging phenotyping strategies will advance our understanding of  
1003 psychiatric genetics. *Nat. Neurosci.* **23**, 6 (2020).
- 1004 26. Johnson, E. C. *et al.* A large-scale genome-wide association study meta-analysis of  
1005 cannabis use disorder. *Lancet Psychiatry* **7**, 1032–1045 (2020).
- 1006 27. Quach, B. C. *et al.* Expanding the genetic architecture of nicotine dependence and its  
1007 shared genetics with multiple traits. *Nat. Commun.* **11**, 5562 (2020).
- 1008 28. Hancock, D. B., Markunas, C. A., Bierut, L. J. & Johnson, E. O. Human Genetics of  
1009 Addiction: New Insights and Future Directions. *Curr. Psychiatry Rep.* **20**, 8 (2018).
- 1010 29. Sanchez-Roige, S., Cox, N. J., Johnson, E. O., Hancock, D. B. & Davis, L. K. Alcohol and  
1011 cigarette smoking consumption as genetic proxies for alcohol misuse and nicotine  
1012 dependence. *Drug Alcohol Depend.* **221**, 108612 (2021).

- 1013 30. DeBoever, C. *et al.* Assessing Digital Phenotyping to Enhance Genetic Studies of Human  
1014 Diseases. *Am. J. Hum. Genet.* **106**, 611–622 (2020).
- 1015 31. Sanchez-Roige, S. & Palmer, A. A. Electronic Health Records Are the Next Frontier for the  
1016 Genetics of Substance Use Disorders. *Trends Genet.* **35**, 317–318 (2019).
- 1017 32. Zheutlin, A. B. *et al.* Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia  
1018 in 106,160 Patients Across Four Health Care Systems. *Am. J. Psychiatry* **176**, 846–855  
1019 (2019).
- 1020 33. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning  
1021 Healthcare System to Accelerate Precision Medicine in a Diverse Population. *J. Pers. Med.*  
1022 **12**, 1974 (2022).
- 1023 34. Roughley, S., Marcus, A. & Killcross, S. Dopamine D1 and D2 Receptors Are Important for  
1024 Learning About Neutral-Valence Relationships in Sensory Preconditioning. *Front. Behav.*  
1025 *Neurosci.* **15**, (2021).
- 1026 35. Gelernter, J. *et al.* Haplotype spanning TTC12 and ANKK1, flanked by the DRD2 and  
1027 NCAM1 loci, is strongly associated to nicotine dependence in two distinct American  
1028 populations. *Hum. Mol. Genet.* **15**, 3498–3507 (2006).
- 1029 36. Hatoum, A. S. *et al.* Multivariate genome-wide association meta-analysis of over 1 million  
1030 subjects identifies loci underlying multiple substance use disorders. *Nat. Mental Health.* **1**,  
1031 210–223 (2023).
- 1032 37. 23andMe Research Team *et al.* Association studies of up to 1.2 million individuals yield  
1033 new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244  
1034 (2019).
- 1035 38. Sanchez-Roige, S. *et al.* Genome-wide association study of problematic opioid prescription  
1036 use in 132,113 23andMe research participants of European ancestry. *Mol. Psychiatry* **26**,  
1037 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., Umičević 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).

- 1063 49. Kember, R. L. *et al.* Phenome-wide Association Analysis of Substance Use Disorders in a  
1064 Deeply Phenotyped Sample. *Biol. Psychiatry* (2022) doi:10.1016/j.biopsych.2022.08.010.
- 1065 50. McLellan, A. T., Koob, G. F. & Volkow, N. D. Preadiction—A Missing Concept for Treating  
1066 Substance Use Disorders. *JAMA Psychiatry* **79**, 749–751 (2022).
- 1067 51. Brazel, D. M. *et al.* Exome Chip Meta-analysis Fine Maps Causal Variants and Elucidates  
1068 the Genetic Architecture of Rare Coding Variants in Smoking and Alcohol Use. *Biol.*  
1069 *Psychiatry* **85**, 946–955 (2019).
- 1070 52. Miranda, M., Morici, J. F., Zanoni, M. B. & Bekinschtein, P. Brain-Derived Neurotrophic  
1071 Factor: A Key Molecule for Memory in the Healthy and the Pathological Brain. *Front. Cell.*  
1072 *Neurosci.* **13**, (2019).
- 1073 53. Barker, J. M., Taylor, J. R., De Vries, T. J. & Peters, J. Brain-derived neurotrophic factor  
1074 and addiction: Pathological versus therapeutic effects on drug seeking. *Brain Res.* **1628**,  
1075 68–81 (2015).
- 1076 54. Duong, C. *et al.* Glutathione peroxidase-1 protects against cigarette smoke-induced lung  
1077 inflammation in mice. *Am. J. Physiol.-Lung Cell. Mol. Physiol.* **299**, L425–L433 (2010).
- 1078 55. Scieszka, D. *et al.* Subchronic Electronic Cigarette Exposures Have Overlapping Protein  
1079 Biomarkers with Chronic Obstructive Pulmonary Disease and Idiopathic Pulmonary  
1080 Fibrosis. *Am. J. Respir. Cell Mol. Biol.* **67**, 503–506 (2022).
- 1081 56. Erzurumluoglu, A. M. *et al.* Meta-analysis of up to 622,409 individuals identifies 40 novel  
1082 smoking behaviour associated genetic loci. *Mol. Psychiatry* **25**, 2392–2409 (2020).
- 1083 57. Toikumo, S., Xu, H., Gelernter, J., Kember, R. L. & Kranzler, H. R. Integrating human brain  
1084 proteomic data with genome-wide association study findings identifies novel brain proteins  
1085 in substance use traits. *Neuropsychopharmacology* **47**, 2292–2299 (2022).

- 1086 58. Kember, R. L. & et al. Cross-ancestry meta-analysis of opioid use disorder uncovers novel  
1087 loci with predominant effects on brain. *medRxiv* (2021)  
1088 doi:<https://doi.org/10.1101/2021.12.13.21267480>.
- 1089 59. Koob, G. F. & Volkow, N. D. Neurobiology of addiction: a neurocircuitry analysis. *Lancet*  
1090 *Psychiatry* **3**, 760–773 (2016).
- 1091 60. King, D. P. *et al.* Smoking Cessation Pharmacogenetics: Analysis of Varenicline and  
1092 Bupropion in Placebo-Controlled Clinical Trials. *Neuropsychopharmacology* **37**, 641–650  
1093 (2012).
- 1094 61. King, A. C. *et al.* Effects of Naltrexone on Smoking Cessation Outcomes and Weight Gain  
1095 in Nicotine-Dependent Men and Women. *J. Clin. Psychopharmacol.* **32**, 630–636 (2012).
- 1096 62. Carpenter, M. J. *et al.* Clinical Strategies to Enhance the Efficacy of Nicotine Replacement  
1097 Therapy for Smoking Cessation: A Review of the Literature. *Drugs* **73**, 407–426 (2013).
- 1098 63. So, H.-C. *et al.* Analysis of genome-wide association data highlights candidates for drug  
1099 repositioning in psychiatry. *Nat. Neurosci.* **20**, 1342–1349 (2017).
- 1100 64. Sey, N. Y. A. *et al.* Chromatin architecture in addiction circuitry identifies risk genes and  
1101 potential biological mechanisms underlying cigarette smoking and alcohol use traits. *Mol.*  
1102 *Psychiatry* **27**, 3085–3094 (2022).
- 1103 65. McGinnis, K. A. *et al.* Using the biomarker cotinine and survey self-report to validate  
1104 smoking data from United States Veterans Health Administration electronic health records.  
1105 *JAMIA Open* **5**, ooac040 (2022).
- 1106 66. Jang, S.-K. *et al.* Rare genetic variants explain missing heritability in smoking. *Nat. Hum.*  
1107 *Behav.* **6**, 1577–1586 (2022).
- 1108 67. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric  
1109 genetics. *Cell* **148**, 1223–1241 (2012).



- 1110 68. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in  
1111 large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
- 1112 69. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer  
1113 datasets. *GigaScience* **4**, 7 (2015).
- 1114 70. Dennis, J. K. *et al.* Clinical laboratory test-wide association scan of polygenic scores  
1115 identifies biomarkers of complex disease. *Genome Med.* **13**, 6 (2021).
- 1116 71. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based  
1117 Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 1118 72. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic  
1119 variation. *Nature* **526**, 68–74 (2015).
- 1120 73. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**,  
1121 1284–1287 (2016).
- 1122 74. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide  
1123 association studies. *Nat. Genet.* **38**, 904–909 (2006).
- 1124 75. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate,  
1125 scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
- 1126 76. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in  
1127 Genome-wide Association Studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).
- 1128 77. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in  
1129 diverse human populations. *Nature* **467**, 52–58 (2010).
- 1130 78. Luo, Y. *et al.* Estimating heritability and its enrichment in tissue-specific gene sets in  
1131 admixed populations. *Human Molecular Genetics* **30**, 1521–1534 (2021).
- 1132 79. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of  
1133 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

- 1134 80. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide  
1135 Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
- 1136 81. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-  
1137 wide association studies. *Bioinforma. Oxf. Engl.* **32**, 1493–1501 (2016).
- 1138 82. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-  
1139 Mapping Studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).
- 1140 83. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide  
1141 association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- 1142 84. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies  
1143 disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
- 1144 85. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat.*  
1145 *Biotechnol.* **28**, 1045–1048 (2010).
- 1146 86. The GTEx Consortium atlas of genetic regulatory effects across human tissues | Science.  
1147 [https://www.science.org/doi/10.1126/science.aaz1776?url\\_ver=Z39.88-](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%20pubmed)  
1148 [2003&rfr\\_id=ori:rid:crossref.org&rfr\\_dat=cr\\_pub%20%20pubmed.](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%20pubmed)
- 1149 87. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage  
1150 sensitivity in cancer. *Nat. Genet.* **47**, 115–125 (2015).
- 1151 88. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using  
1152 predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
- 1153 89. Beach, T. G. *et al.* Arizona Study of Aging and Neurodegenerative Disorders and Brain  
1154 and Body Donation Program. *Neuropathol. Off. J. Jpn. Soc. Neuropathol.* **35**, 354–389  
1155 (2015).
- 1156 90. Wingo, T. S. *et al.* Brain proteome-wide association study implicates novel proteins in  
1157 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 Extensions. (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).

- 1182 101. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*  
1183 *(DSM-IV)*. (American Psychiatric Association, 1994).
- 1184 102. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*  
1185 *(DSM-5)*. (American Psychiatric Association, 2013).
- 1186 103. Gelernter, J. *et al.* Genome-wide association study of alcohol dependence: significant  
1187 findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry* **19**,  
1188 41–49 (2014).
- 1189 104. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via  
1190 Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
- 1191 105. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-Wide Association Studies as a Tool  
1192 to Advance Precision Medicine. *Annu. Rev. Genomics Hum. Genet.* **17**, 353–373 (2016).
- 1193 106. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* **36**,  
1194 930–933 (2020).
- 1195 107. Ruan, Y. *et al.* Improving Polygenic Prediction in Ancestrally Diverse Populations. *medRxiv*  
1196 21 (2021) doi:<https://doi.org/10.1101/2020.12.27.20248738>.
- 1197 108. Rescorla, L. *et al.* Behavioral/Emotional Problems of Preschoolers Caregiver/Teacher  
1198 Reports From 15 Societies. *J. Emot. Behav. Disord.* **20**, 68–81 (2012).
- 1199 109. Akshoomoff, N. *et al.* NIH Toolbox Cognitive Function Battery (CFB): Composite Scores of  
1200 Crystallized, Fluid, and Overall Cognition. *Monogr. Soc. Res. Child Dev.* **78**, 119–132  
1201 (2013).
- 1202 110. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing  
1203 Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–  
1204 1739 (2017).

- 1205 111. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian  
1206 randomization: Using genes as instruments for making causal inferences in epidemiology.  
1207 *Stat. Med.* **27**, 1133–1163 (2008).
- 1208 112. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With  
1209 Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* **37**, 658–665 (2013).
- 1210 113. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in  
1211 Mendelian Randomization with Some Invalid Instruments Using a Weighted Median  
1212 Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
- 1213