8-26-2021

Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction

Richard Karlsson Linnér

Travis T. Mallard

Peter B. Barr

Sandra Sanchez-Roige

James W. Madole

See next page for additional authors

Recommended Citation

Karlsson Linnér, R., Mallard, T.T., Barr, P.B. et al. Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction. Nat Neurosci (2021). https://doi.org/10.1038/s41593-021-00908-3

Title: Multivariate genomic analysis of 1.5 million people identifies genes related to addiction, antisocial behavior, and health

Authors: Richard Karlsson Linnér [1^], Travis T. Mallard [2^], Peter B. Barr [3#], Sandra Sanchez-Roige [4,5#], James W. Madole [2], Morgan N. Driver [6], Holly E. Poore [7], Andrew D. Grotzinger [2], Jorim J. Tielbeek [8], Emma C. Johnson [9], Mengzhen Liu [10], Hang Zhou [11,12], Rachel L. Kember [13,14], Joëlle A. Pasman [15], Karin J.H. Verweij [16], Dajiang J. Liu [17,18], Scott Vrieze [10], COGA Collaborators, Henry R. Kranzler [13,14], Joel Gelernter [11,12,19,20], Kathleen Mullan Harris [21,22], Elliot M. Tucker-Drob [2,23], Irwin Waldman [7,24], Abraham A. Palmer [4,25,†], K. Paige Harden [2,23,†], Philipp D. Koellinger [1,26,†*], and Danielle M. Dick [3,6†*]

Affiliations:

[1]Department of Economics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands.

[2]Department of Psychology, University of Texas at Austin, Austin, TX, USA.

[3]Department of Psychology, Virginia Commonwealth University, Richmond, VA, USA.

[4]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA.

[5]Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA.

[6]Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, USA.

[7]Department of Psychology, Emory University, Atlanta, GA, USA.

[8]Department of Complex Trait Genetics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands.

[9]Department of Psychiatry, Washington University School of Medicine, Saint Louis, MO, USA.

[10]Department of Psychology, University of Minnesota, Minneapolis, MN, USA.

[11]Department of Psychiatry, Yale University School of Medicine, West Haven, CT, USA.

[12]Department of Psychiatry, VA CT Healthcare System, West Haven, CT, USA

[13]Center for Studies of Addiction, University of Pennsylvania School of Medicine, Philadelphia, PA, USA.

[14]Mental Illness Research Education and Clinical Center, Crescenz VA Medical Center, Philadelphia, PA, USA.

[15]Behavioural Science Institute, Radboud University Nijmegen, Nijmegen, Netherlands.

[16]Department of Psychiatry, University of Amsterdam, Amsterdam, Netherlands.

[17]Department of Public Health Sciences, Penn State University, Hershey, PA, USA.

[18]Institute of Personalized Medicine, Penn State University, Hershey, PA, USA.

[19]Department of Genetics, Yale University School of Medicine, West Haven, CT, USA.

[20]Department of Neuroscience, Yale University School of Medicine, West Haven, CT, USA.

[21]Department of Sociology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[22]Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[23]Population Research Center, University of Texas at Austin, Austin, TX, USA.

[24]Center for Computational and Quantitative Genetics, Emory University, Atlanta, GA, USA.

[25]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA.

[26]La Follette School of Public Affairs, University of Wisconsin-Madison, WI, USA.

*Correspondence to: ddick@vcu.edu, koellinger@wisc.edu

^Joint first authors, #Joint second authors, †Joint senior authors

Abstract

Behaviors and disorders related to self-regulation, such as substance use, antisocial conduct, and ADHD, are collectively referred to as *externalizing* and have a shared genetic liability. We applied a multivariate approach that leverages genetic correlations among externalizing traits for genome-wide association analyses. By pooling data from ~1.5 million people, our approach is statistically more powerful than single-trait analyses and identifies more than 500 genetic loci. The identified loci were enriched for genes expressed in the brain and related to nervous system development. A polygenic score constructed from our results captures variation in a broad range of behavioral and medical outcomes that were not part of our genome-wide analyses, including traits that until now lacked well-performing polygenic scores, such as opioid use disorder, suicide, HIV infections, criminal convictions, and unemployment. Our findings are consistent with the idea that persistent difficulties in self-regulation can be conceptualized as a neurodevelopmental condition.

Main

Behaviors and disorders related to self-regulation, such as substance use disorders or antisocial behaviors, have far-reaching consequences for affected individuals, their families, communities, and society at large[1,2]. Collectively, this group of correlated traits are classified as *externalizing*[3]. Twin-family studies have demonstrated that externalizing liability is highly heritable (~80%)[4,5], suggesting it will be as tractable to gene discovery as other complex traits or medical conditions[6]. To date, however, there have been no large-scale molecular genetic studies that utilize the extensive degree of genetic overlap among externalizing traits to aid gene discovery, as most studies have focused on individual disorders or diseases[7]. But for many high-cost, high-risk externalizing behaviors – opioid use disorder and suicide attempts being salient examples – there are too few cases available with genome-wide data to yield sufficient power for gene discovery[8,9].

A complementary strategy to the single-disease approach is to study the shared genetic architecture across traits in multivariate analyses, which boosts statistical power by pooling data across genetically correlated traits[10]. Multivariate approaches can utilize summary statistics from genome-wide association studies (GWAS), which are now widely available, to allow for the discovery of connections between phenotypes not naturally studied together because they span different domains, fields of study, or life stages. Conveniently, by adjusting for sample overlap, novel statistical methods can attain an even greater effective sample size by efficiently utilizing observations from overlapping studies. Elucidating the shared genetic basis of externalizing liability has the potential to advance our understanding of the biological processes related to behavioral undercontrol, and enables mapping the pathways by which genetic risk and socio-environmental factors interact to contribute to the development of different externalizing outcomes.

Here, we applied genomic structural equation modeling (Genomic SEM) to summary statistics from GWAS on multiple forms of externalizing behavior for which large samples were available[10]. This approach was grounded in the existing literature showing shared genetic liability across numerous externalizing disorders and with non-psychiatric variation in externalizing behavior[5,11]. We posited that applying this multivariate approach would lead to the identification of genetic variants associated with a broad array of externalizing phenotypes, as well as related behavioral, social, and medical outcomes that were not directly included in our genome-wide association analysis.

Results

Multivariate analysis of seven externalizing phenotypes identifies numerous genetic associations with a general liability to externalizing

Following our preregistered analysis plan (https://doi.org/10.17605/OSF.IO/XKV36, Supplementary Information section 1), we collated GWAS summary statistics from externalizing-related disorders and behaviors, with our final analysis using data from seven externalizing phenotypes with sample sizes >50,000 (**Table 1**): (1) attention-deficit/hyperactivity disorder (ADHD), (2) problematic alcohol use (ALCP), (3) lifetime cannabis use (CANN), (4) age at first sexual intercourse (FSEX), (5) number of sexual partners (NSEX), (6) general risk tolerance (RISK), and (7) lifetime smoking initiation (SMOK). All samples were of European ancestry. The GWAS protocol is described in Supplementary Information section 2 (**Supplementary Tables 1–4**).

Table 1. Summary of seven externalizing-related disorders and behaviors with GWAS summary statistics (N > 50,000)

Phenotype (abbreviation)	N	h^2 (*SE*)	λ_{GC}	Mean χ^2	Intercept	Ratio	Reference
Attention-deficit/hyperactivity disorder (ADHD)	53,293	.235 (*.015*)	1.253	1.297	1.034	.113	[12]
Problematic alcohol use (ALCP)	164,121	.055 (*.004*)	1.149	1.174	1.013	.073	[13,14]
Lifetime cannabis use (CANN)	186,875	.066 (*.004*)	1.230	1.267	1.026	.098	[15]
Age at first sexual intercourse (FSEX)	357,187	.115 (*.004*)	1.623	1.869	1.036	.041	[16]
Number of sexual partners (NSEX)	336,121	.097 (*.004*)	1.492	1.682	1.027	.041	[16]
General risk tolerance (RISK)	426,379	.053 (*.002*)	1.372	1.461	1.019	.041	[16]
Lifetime smoking initiation (SMOK)	1,251,809	.078 (*.002*)	2.328	3.152	1.126	.058	[17]

Notes: The statistics reported in this table were all estimated with LD Score regression[18]. Heritability (h^2) is on the observed scale[18]. λ_{GC} is the median χ^2 statistic divided by the expected median of the χ^2 distribution with 1 degree of freedom[19]. Mean χ^2 is the average χ^2 statistic. Intercept is the estimated LD Score regression intercept. Ratio measures stratification bias, defined as (Intercept – 1) / (Mean χ^2 – 1)[18].

Consistent with twin studies[4,5], the genetic correlations among the seven discovery phenotypes were moderate to high (**Figure 1A** and **Supplementary Table 5**). Using Genomic SEM[10] (Supplementary Information section 3), which is unbiased by sample overlap and differences in sample sizes in the discovery phenotypes, we formally modeled the genetic covariances among the seven phenotypes and found that a common factor model fits the data best. This common factor, which we refer to as *EXT*, captures a shared genetic liability to the seven externalizing traits that we included in our analyses (**Figure 1B** and **Supplementary Table 7**).

We then extended Genomic SEM to estimate genetic correlations between *EXT* and 92 preregistered phenotypes with GWAS summary statistics that were not included among the seven discovery phenotypes (**Extended Data Fig. 1** and **Supplementary Table 8**). The genetic correlations indicate convergent and discriminant validity of the common *EXT* factor (**Figure 1C**): As anticipated, *EXT* showed strong genetic correlations with drug exposure (r_g = .91), antisocial behavior (r_g = .65), motor impulsivity (r_g = .70), failures to plan (r_g = .70),

and (lack of) agreeableness (r_g = –.79), a personality trait characterized by kindness and cooperativeness that has been found to be low in individuals displaying antisocial behavior. *EXT* was also strongly correlated with suicide attempts (r_g = .68). *EXT* showed more modest inverse correlations with educational attainment (r_g = –.32) and intelligence (r_g = –.23), indicating that the latent factor is not simply reflecting genetic influences on cognitive ability. Finally, there was a strong genetic correlation with the Townsend index (r_g = .71), a measure of neighborhood deprivation that reflects high concentrations of unemployment, household overcrowding, and low concentrations of home- and car-ownership[20].

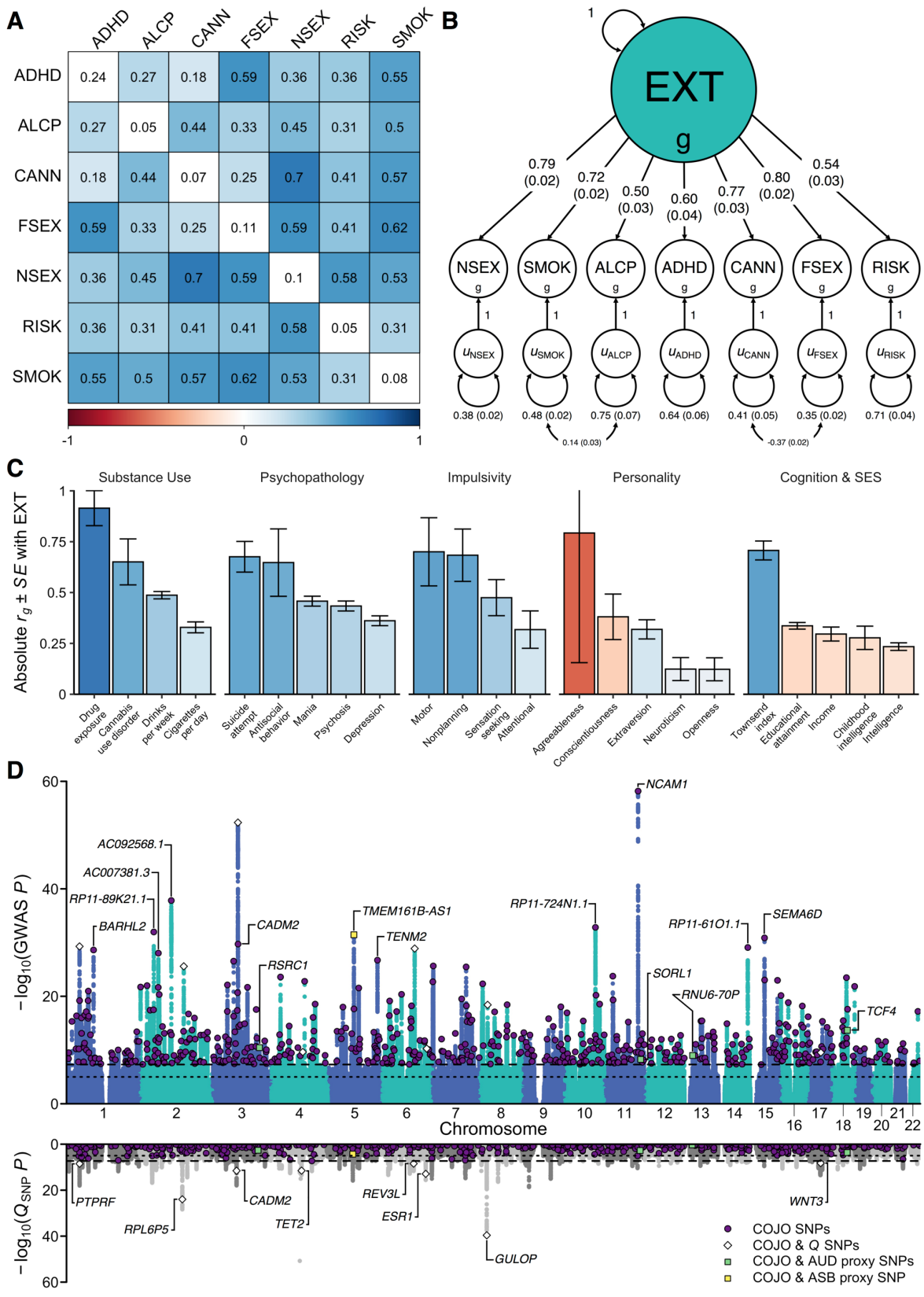

Figure 1 | Multivariate genome-wide analyses with Genomic SEM. (**A**) Pair-wise genetic correlations (r_g) among seven discovery phenotypes, with observed-scale SNP heritabilities (h^2) on the diagonal. (**B**) Path diagram of a confirmatory factor model estimated with Genomic SEM. The parameter estimates were

standardized, and standard errors are presented in parentheses. (**C**) Absolute value genetic correlations, $|r_g|$, between the genetic externalizing factor (*EXT*) and phenotypes selected to establish convergent and discriminant validity, where blue and red bars represent positive and negative genetic correlations, respectively. Standard errors are presented as error bars. (**D**) GWAS associations (top panel) and Q_{SNP} tests of heterogeneity (bottom panel) for *EXT*. Purple dots represent 579 *EXT* lead SNPs that are conditionally and jointly associated (COJO) at genome-wide significance (two-sided test $P < 5\times10^{-8}$). White diamonds represent eight of the 579 SNPs that also show significant Q_{SNP} heterogeneity. Four green and one yellow squares represent five out of the 579 SNPs that also were Bonferroni-significant proxy-phenotype associations with alcohol use disorder (AUD) and antisocial behavior (ASB), respectively. ADHD is attention deficit hyperactivity disorder, ALCP is problematic alcohol use, CANN is lifetime cannabis use, EXT is externalizing, FSEX is age at first sex, NSEX is number of sexual partners, RISK is general risk tolerance, SMOK is lifetime smoking initiation.

We next used Genomic SEM[10] to perform a GWAS on the shared genetic liability *EXT* (**Figure 1D** and **Extended Data Fig. 2**) (Supplementary Information section 3.4). This analysis estimated single-nucleotide polymorphism (SNP) associations directly with the *EXT* factor, with an effective sample size of N = 1,492,085 individuals. These analyses are different in their approach and substantially increase sample size, statistical power, and the range of findings compared to previous work[21] (Supplementary Information section 2.2.1). After applying conditional and joint multiple-SNP analysis (COJO) on a set of near-independent, genome-wide significant (two-sided test $P < 5\times10^{-8}$) lead SNPs[22], we identified 579 conditionally and jointly associated SNPs (**Supplementary Table 9**), meaning they were significantly associated with *EXT* even after statistically adjusting for each other and other lead SNPs. Of the 579 *EXT* SNPs and their correlates within linkage disequilibrium (LD) regions ($r^2 > 0.1$), 121 (21%) were new loci, not previously associated with any of the seven externalizing behaviors/disorders that went into the Genomic SEM model, and 41 (7%) can be classified as entirely novel, as they have not been reported previously for any trait in the GWAS literature.

Genomic SEM was used to perform SNP-level tests of heterogeneity (Q_{SNP}; Supplementary Information section 3.5.1) that investigate whether each SNP had consistent, pleiotropic effects on the seven input phenotypes that effectively operate via the shared genetic liability *EXT* (**Extended Data Fig. 2**). Only 1% (8/579) of the 579 *EXT* SNPs were significant (one-sided Q_{SNP} $P < 5\times10^{-8}$) in Q_{SNP} tests (**Figure 1D**; **Supplementary Table 9**), providing further evidence that the genetic variants we identified primarily index a unitary dimension of genetic externalizing liability rather than representing an amalgamation of variants with divergent associations across the discovery phenotypes. The genome-wide Q_{SNP} analysis was adequately powered (mean $\chi^2 = 1.864$; **Extended Data Fig. 2**), and as expected, it identified heterogeneity in regions of the genome not associated with *EXT*. The strongest Q_{SNP} and most salient example of a trait-specific association is SNP rs1229984 (one-sided Q_{SNP} $P = 1.67\times10^{-51}$). This particular SNP, located in the gene *ADH1B*, is a known missense variant with a well-established role in alcohol metabolism[23], and it was not associated with *EXT* (two-sided $P = 0.022$) but only with problematic alcohol use (two-sided $P = 6.43\times10^{-57}$).

Because the discovery stage effectively exhausted large study cohorts available for strict replication, we instead performed a series of preregistered quasi-replication analyses, which have previously been applied successfully in the GWAS setting[24,25]. Further below, we additionally perform holistic quasi-replication of the 579 *EXT* SNPs in polygenic score analyses (also in within-family models). For SNP-level quasi-replication analyses of the 579 SNPs (Supplementary Information section 4), a three-step holistic method tested their association with two independent, GWAS meta-analyses on externalizing phenotypes: (1) alcohol use disorder (r_g with *EXT* = 0.52; N = 202,004), and (2) antisocial behavior (r_g with *EXT* = 0.69; N = 32,574). First, we tested whether the 579 SNPs (or an LD proxy for missing SNPs, $r^2 > 0.8$) showed sign concordance, *i.e.*, the same direction of effect between *EXT* and alcohol use disorder or antisocial behavior: 75.4% of SNPs showed sign concordance with

alcohol use disorder (two-sided test $P = 6.84\times10^{-36}$) and 66.9% with antisocial behavior (two-sided test $P = 1.39\times10^{-15}$) (**Extended Data Fig. 3**). For the second and third tests, we generated empirical null distributions for the two phenotypes by randomly selecting 250 near-independent ($r^2 < 0.1$) SNPs per each of the 579 SNPs, matched on allele frequency. In the second test, a greater proportion of the 579 SNPs were nominally associated ($P < 0.05$) with the two phenotypes compared to their empirical null distributions: 124 (21.4% vs. 6.6%) with alcohol use disorder (two-sided $P = 1.87\times10^{-31}$) and 58 (10.5% vs. 4.7%) with antisocial behavior ($P = 1.64\times10^{-8}$). In the third test, the 579 SNPs were jointly more strongly enriched for association with alcohol use disorder (one-sided Mann-Whitney test $P = 5.89\times10^{-26}$) and antisocial behavior ($P = 1.10\times10^{-5}$) compared to their empirical null distributions. Overall, the quasi-replications consistently suggested that the GWAS of *EXT* is not spurious overall, and that it is enriched for genetic signal with phenotypes of central importance to the literature on externalizing.

Bioinformatic analyses highlight relevant neurodevelopmental and biological processes

We performed a series of bioinformatic analyses to explore the biological processes underlying externalizing liability (Supplementary Information section 6, **Supplementary Tables 9–10**, and **21–29**; **Extended Data Figs. 5–8**). Consistent with the idea that persistent difficulties in self-regulation can be conceptualized as a neurodevelopmental condition[26,27], MAGMA gene-property analyses suggested an abundance of enrichment in genes expressed in brain tissues, particularly during prenatal developmental stages (**Extended Data Fig. 7**), with the strongest enrichment seen in the cerebellum, followed by frontal cortex, limbic system tissues, and pituitary gland tissues (**Extended Data Fig. 6**). Furthermore, MAGMA gene-set analysis identified gene sets related to neurogenesis, nervous system development, and synaptic plasticity, among other gene-sets related to neuronal function and structure.

Because of the strong polygenic signal identified in the GWAS of *EXT*, four different gene-based analyses identified an abundance of implicated genes (>3,000): (1) functional annotation of the 579 SNPs to their nearest gene with FUMA[28], which suggested 587 genes; (2) MAGMA gene-based association analysis[29], which identified 928 Bonferroni-significant genes (one-sided test $P < 2.74\times10^{-6}$); (3) H-MAGMA[30], a method that assigns non-coding SNPs to cognate genes based on chromatin interactions in adult brain tissue and which identified 2,033 Bonferroni-significant genes (one-sided test $P < 9.84\times10^{-7}$); and (4) S-PrediXcan[31], which uses transcriptome-based analyses of predicted gene expression in 13 brain tissues and which identified 348 Bonferroni-significant gene-tissue pairs (two-sided test $P < 2.73\times10^{-7}$).

We found 34 genes that were consistently identified in all four methods, while 741 overlapped across two or more methods (**Supplementary Table 29**; **Extended Data Fig. 8**). Several of the 34 implicated genes are novel discoveries for the psychiatric/behavioral literature and have previously been identified only in relation to biomedical disease. Such

discoveries include *ALMS1* (previously associated with kidney function and urinary metabolites[32]), and *ERAP2* (blood protein levels and autoimmune disease [33,34]). Other genes among the 34 have previously been identified in GWAS of behavioral or psychiatric traits: Cell Adhesion Molecule 2 (*CADM2*, previously identified in GWAS related to self-regulation, including drug use and risk tolerance[16,35]), Zic Family Member 4 (*ZIC4*, associated with brain volume[36]), Gamma-Aminobutyric Acid Type A Receptor Subunit Alpha 2 (*GABRA2*; the site of action for alcohol and benzodiazepines, extensively studied in relation to alcohol dependence[37,38], and proposed candidate gene for many psychiatric disorders[39,40]), *NEGR1* (neuronal growth regulator, associated with intelligence and educational attainment[25,41]), and Paired Basic Amino Acid Cleaving Enzyme (*FURIN*, associated with schizophrenia, risk tolerance, and trans-diagnostic vulnerability to psychiatric disorders[42,43]).

Genetic risk scores explain substantial variation in behavioral, psychiatric, and social outcomes

We created a genome-wide polygenic score for *EXT,* adjusted for LD[44,45], among subjects from two European-ancestry datasets selected for their detailed phenotypes related to externalizing outcomes (Supplementary Information section 5): (1) the National Longitudinal Study of Adolescent to Adult Health (Add Health; N = 5,107), a U.S.-based study of adolescents who were recruited from secondary schools in the mid-1990s; (2) the Collaborative Study on the Genetics of Alcoholism (COGA; N = 7,594), a U.S.-based study focused on understanding genetic contributions to alcohol use disorders.

To investigate the validity of *EXT*, in each sample, we fit a latent factor model to phenotypic data corresponding to the seven Genomic SEM phenotypes (**Extended Data Fig. 4** and **Supplementary Table 13)**. Controlling for age, sex, and ten principal components of genetic ancestry, the *EXT* polygenic score was strongly associated with the latent phenotypic

factor in both data sets ($\beta_{\text{Add Health}}$ = 0.33, 95% CI, 0.30 to 0.36, ΔR^2 = 10.5%; β_{COGA} = 0.30, 95% CI, 0.27 to 0.34, ΔR^2 = 8.9%; **Figure 2A** and **Supplementary Table 14**). The variance explained by the *EXT* polygenic score (ΔR^2 ~ 8.9–10.5%) is commensurate with many conventional variables used in social science research, including parental socioeconomic status, family income or structure, and neighborhood disadvantage/disorder[46–48]. Next, as further quasi-replication, in each sample we created a polygenic score using only the 579 *EXT* SNPs. This polygenic score was associated with the latent phenotypic externalizing factor in both samples, explaining ~3–4% of the variance ($\beta_{\text{Add Health}}$ = 0.20, 95% CI, 0.17 to 0.23; β_{COGA} = 0.17, 95% CI, 0.13 to 0.20; **Supplementary Table 14**).

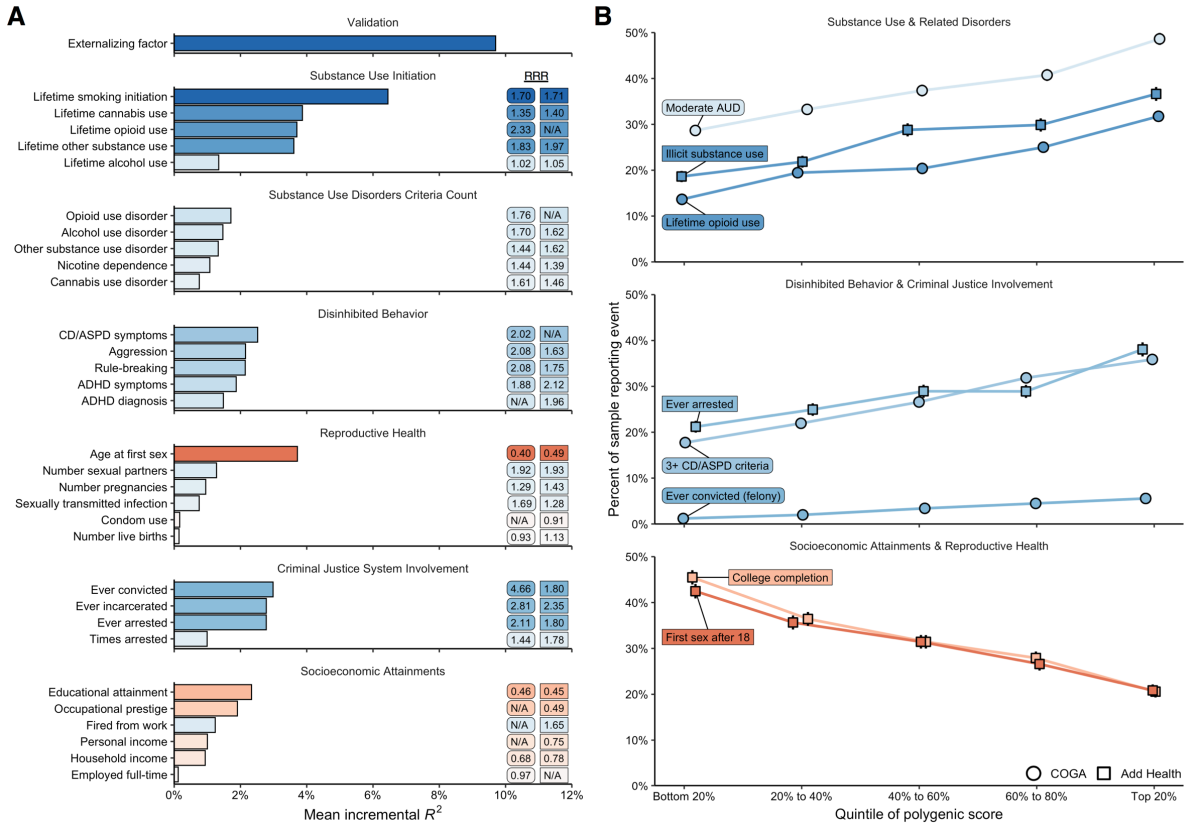

Figure 2 | Polygenic score associations with behavioral, psychiatric, and social outcomes in the independent Add Health (*N* = 5,107) and COGA (*N* = 7,594) datasets. (**A**) Bar charts illustrating the mean proportion of variance (incremental R^2, or ΔR^2) explained by the polygenic score. Blue and red bars indicate positive and negative associations, respectively. Relative risk ratios (RRRs), comparing individuals in the lowest

20% to those in the highest 20% of the polygenic score distribution, are reported for Add Health and COGA in square and round boxes, respectively. (**B**) Line charts illustrating the relative risks across quintiles of the polygenic score for eight illustrative outcomes: (1) meeting 4 or more criteria for alcohol use disorder (AUD), (2) lifetime use of an illicit substance other than cannabis, (3) lifetime opioid use, (4) ever being arrested, (5) meeting 3 or more criteria for conduct disorder (CD) or antisocial personality disorder (ASPD), (6) ever being convicted of a felony, (7) completing college, and (8) first sexual intercourse at the age of 18 or older. 95% confidence intervals are presented with error bars for each quintile.

We next explored to what extent polygenic scores for *EXT* were associated with childhood externalizing disorders and a variety of specific phenotypes that reflect difficulty with self-regulation or its social consequences (**Figure 2B** and **Supplementary Tables 16–19**). Polygenic scores for *EXT* explained significant variance (ΔR^2) in criteria counts of ADHD (mean ΔR^2 = 1.65%), conduct disorder (CD; mean ΔR^2 = 3.1%), and oppositional defiant disorder (ODD; ΔR^2 = 1.96%), as well as in phenotypes categorized as substance use initiation (mean ΔR^2 = 1.3–6.5%), substance use disorders (mean ΔR^2 = 0.8–1.7%), disinhibited behaviors (mean ΔR^2 = 1.5–2.5%), criminal justice system involvement (mean ΔR^2 = 1.0–3.0%), reproductive health (mean ΔR^2 = 0.3–3.7%), and socioeconomic attainment (mean ΔR^2 = 0.1–2.3%). Many of the phenotypes – such as opioid use disorder criteria count, conduct disorder and antisocial personality disorder criteria count, lifetime history of arrest or incarceration, and lifetime history of being fired from work, were not included in our Genomic SEM analyses; however, our *EXT* polygenic score is notable in capturing appreciable variance in phenotypes that are still lacking large GWAS samples (a striking example being opioid use disorder[8]). The associations between the *EXT* polygenic score and this broad range of phenotypes represents an affirmative test of the hypothesis that genetic variants associated with externalizing liability generalize to a wide variety of behavioral and social outcomes related to behavioral undercontrol.

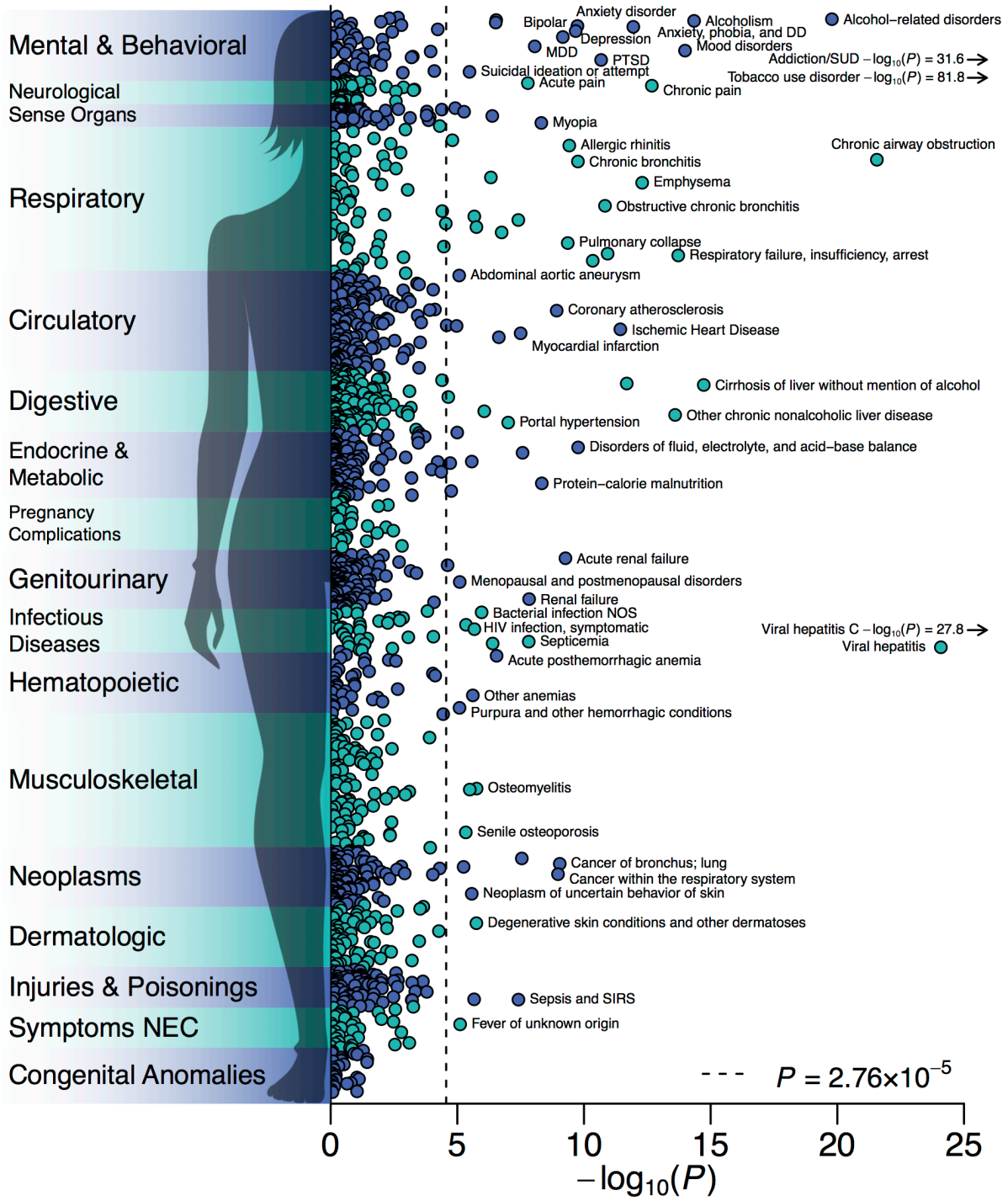

Figure 3 | Phenome-wide association study in the BioVU biorepository. $-\log_{10} P$ values of two-sided test for association of polygenic score for *EXT* with 1,335 medical outcomes were derived with logistic regression in up to 66,915 patients, adjusted for sex, median age in the EHR data, and the first 10 genetic PCs. The dashed line is the Bonferroni-corrected significance threshold; adjusted for the number of tested medical conditions. 84 medical conditions were Bonferroni-significant, while 255 conditions were significant at a false discovery rate less than 0.05. The labels for some conditions were omitted. The full results, including case-control counts, effect sizes, and standard errors, are reported in **Supplementary Table 20**.

To evaluate medical outcomes associated with genetic liability to externalizing, we conducted a phenome-wide association study (PheWAS) in 66,915 genotyped individuals of European-ancestry in the BioVU biorepository, a U.S.-based biobank of electronic health records from the Vanderbilt University Medical Center, spanning 1990 to 2017[49,50]. A logistic regression was fit to 1,335 case/control disease phenotypes. Of these, 255 disease phenotypes were associated with the *EXT* polygenic score at a false discovery rate less than 0.05 (**Figure 3** and **Supplementary Table 20**). The most abundant associations were with mental and behavioral disorders, such as substance use, mood disorders, suicidal ideation, and attempted suicide. Individuals with higher *EXT* polygenic scores also showed worse health in nearly every bodily system. They were more likely to suffer, for example, from ischemic heart disease, viral hepatitis C and HIV infection, type 2 diabetes and obesity, cirrhosis of liver, sepsis, and lung cancer. Notably, many of these medical outcomes are mediated by behaviors related to self-regulation, e.g., smoking, drinking, drug use, condomless sex, and overeating.

Within-family analyses demonstrate that polygenic associations are robust to confounding

Genetic associations detected in GWAS can be due to direct genetic effects, but can also be confounded by uncontrolled population stratification, indirect genetic effects mediated through the parental environment, and assortative mating[51,52]. While reducing statistical power, sibling comparisons overcome these methodological challenges, because meiosis randomizes genotypes to siblings[51,53]. We therefore conducted within-family analyses of polygenic score associations in the sibling sub-samples of Add Health (N = 994 siblings from 492 families) and COGA (N = 1,353 siblings from 621 families), as well as a sample of sibling pairs from the UK Biobank (N = 39,640), which were held-out from the discovery stage (Supplementary Information section 2.3.2).

In Add Health and COGA, the phenotypic factor derived from observations corresponding to the seven discovery phenotypes (see above) was regressed on the *EXT* polygenic scores in a within-family model (**Supplementary Table 15**). Parameter estimates from the within-family models ($\beta_{\text{Add Health}}$ = 0.12, 95% CI, 0.04 to 0.20; β_{COGA} = 0.14, 95% CI, 0.08 to 0.20) were slightly attenuated compared to OLS models without family-specific intercepts ($\beta_{\text{Add Health}}$ = 0.20, 95% CI, 0.16 to 0.24; β_{COGA} = 0.16, 95% CI, 0.12 to 0.20), but remained strong (Add Health β / β_{WF} = 1.667; COGA β / β_{WF} = 1.142) and statistically significant (two-sided test $P = 4.89\times10^{-3}$ and 1.87×10^{-6}, respectively). Additionally, the association of the quasi-replication polygenic score constructed with the 579 *EXT* SNPs did not attenuate in within-family models and remained significant (**Supplementary Table 15**).

In the UK Biobank sibling hold-out sample, we conducted polygenic score analyses of 33 phenotypes from the domains of risky behavior, reproductive health, cognitive ability, personality, and socioeconomic status (**Supplementary Table 19**). Similar to Add Health and COGA, within-family estimates were only modestly attenuated for risky behavior and reproductive health outcomes (mean β / β_{WF} = 1.079); however, effect-sizes in within-family models were substantially attenuated for cognitive ability and socioeconomic status outcomes (β / β_{WF} was 3.3 for educational attainment, 4.9 for household income, 2.1 for neighborhood deprivation). Overall, the *EXT* polygenic score remained significantly associated (two-sided test $P < 0.05$) with 21 outcomes, showing that our GWAS of externalizing captures direct genetic effects on behavioral health and is not solely a consequence of uncontrolled population stratification, indirect genetic effects, or other forms of environmental confounding.

Discussion

Externalizing disorders and behaviors are a widely prevalent cause of human suffering, but understanding of the molecular genetic underpinnings of externalizing has

lagged considerably behind progress made in other areas of medical and psychiatric genetics. For example, dozens of associated genetic loci have been discovered for schizophrenia (>100 loci)[54], bipolar disorder (30 loci)[55], and major depressive disorders (44 loci)[56], whereas recent GWASs of antisocial behavior[57], alcohol use disorders[58], and opioid use disorders[8] have identified only a very small number of significantly associated loci, if any at all. Here, we used multivariate genomic analyses to accelerate genetic discovery, identifying 579 genome-wide significant loci associated with a predisposition toward externalizing disorders and behaviors, 121 of which are entirely novel discoveries for any of the seven phenotypes analyzed. Our results demonstrate that moving beyond traditional disease classification categories can enhance gene discovery, improve polygenic scores, and provide information about the underlying pathways by which genetic variants impact clinical outcomes. GWAS efforts find almost ubiquitous genetic correlations across psychiatric disorders and diagnoses[59,60]; new analytic methods now allow us to capitalize on these genetic correlations. Pragmatically, non-disease phenotypes such as the ones we use here (*e.g.*, self-reported age at first sex) are often easier to measure in the general population than diagnostic status, making it easier to achieve large sample sizes. Expanding beyond individual diagnoses increases our ability to detect genes underlying human behavioral and medical outcomes of consequence.

Our results highlight again that there is no distinct line between the genetic study of biomedical conditions and the genetic study of social and behavioral traits[61]. Linking biology with socially-valued behavioral outcomes can be politically sensitive (**Box 1**)[62]. Polygenic scores created using our GWAS results were associated not just with psychiatric and substance use disorders, but also with correlated social outcomes, such as lower employment and greater criminal justice system involvement, as well as with biomedical conditions affecting nearly every system in the body. Considered together, our analyses demonstrate the

far-reaching toll of human suffering borne by people with high genetic liabilities to externalizing.

Box 1. Grappling with the Legacy of Eugenics

In 1912, Henry Goddard published what is now considered an infamous work of pseudoscience: *The Kallakak Family* traced several generations of a "feeble-minded" family to argue that not just intellectual ability, but also drunkenness, criminality, sexual promiscuity, and morality were hereditary [63]. On the basis of these pedigrees, Goddard recommended that the "feeble-minded" should be institutionalized and prohibited from reproducing. Horrifically, these recommendations were put into practice: Involuntary sterilization programs and other forms of state-sponsored violence targeting the poor and ethnic/racial minorities persisted for decades[64,65]. Even now, the danger of eugenics is not safely in the past. Modern genetics research is routinely appropriated by white supremacist movements to argue that racialized disparities in health, employment, and criminal justice system involvement are due to the genetic inferiority of people of color rather than environmental and historical disadvantages[66–68]. At the same time, failing to understand how genetic differences contribute to vulnerability to externalizing can increase stigma and blame for these behaviors[69,70]. Given the horrific legacy of eugenics, the ongoing reality of racism in the medical and criminal justice systems, and the importance of combatting stigma in psychiatric disorders, the scientific results we report here, which are, for technical reasons, limited to European individuals, must be interpreted with the utmost care. Please see our supporting materials at www.externalizing.org for more information.

Our polygenic score for externalizing has one of the largest effect sizes of any polygenic score in psychiatric and behavioral genetics, accounting for 10% of the variance in

externalizing factor scores, and meaningful variance in outcomes as varied as opioid use, age at first sex, being fired from work, and being convicted of a crime. These effect sizes rival the associations observed with "traditional" covariates used in social science research. But, these effect sizes remain far below twin estimates of heritability for externalizing[5] and far below what is necessary to predict these outcomes for any individual[71,72]. Furthermore, while effect sizes were only modestly attenuated in within-family models of risky behavior and reproductive behavior, they were substantially attenuated in analyses of socioeconomic outcomes, indicating that substantial work remains to be done to clarify the association between externalizing genetics and socioeconomic inequality[51]. Additionally, application of these genetic discoveries to improve research and intervention will be limited as long as the samples available for genomics research fail to reflect the world's genetic diversity[73].

Finally, these results are *not* evidence that some people are genetically determined to experience certain life outcomes or are "innately" antisocial. Genetic differences are probabilistically associated with psychiatric, medical, and social outcomes, in part via environmental mechanisms that might differ across historical, political, and economic contexts[74]. For example, a policy change like decriminalization of cannabis use might mitigate associations between genetic vulnerabilities and criminal justice system involvement, because the state ceases to criminalize a behavior to which some individuals have a greater genetic susceptibility. At the same time, increased availability and decreased stigma may create environments more conducive to the development of substance problems among individuals who are genetically at risk[75]. The impact of genetic factors might also depend on other forms of social capital and privilege. For instance, childhood externalizing is associated with greater adult earnings, but only for children not raised in poverty[76,77]. The genetic differences identified here can thus be used in future research as a tool to trace how lifespan development is shaped via complex interactions between genetic predispositions,

environmental influences (*e.g.*, parenting, peer, and romantic relationships) and social institutions (*e.g.*, schools, jails, hospitals, creditors, and employers).

Online methods

The article is accompanied by Supplementary Information with further details. The study was performed according to a preregistered analysis plan (https://doi.org/10.17605/OSF.IO/XKV36), which specified that we would either generate new or collect existing single-phenotype genome-wide association study (GWAS) summary statistics on phenotypes related to the externalizing spectrum (Supplementary Information section 1). In the discovery stage, the summary statistics were to be analyzed with Genomic SEM with the aims of (a) estimating a genetic factor structure underlying externalizing liability, (b) identifying single-nucleotide polymorphisms (SNPs) and genes primarily involved in a shared genetic liability to externalizing, and (c) increasing the accuracy of genetic risk scores for specific externalizing phenotypes that are currently intractable to study in large samples. To ensure satisfying statistical power, we preregistered a minimum sample-size threshold of $N > 15{,}000$, and that additional exclusions would be based on displaying negligible or inaccurate SNP-based heritability or genetic covariance. The study did not manipulate an experimental condition, and thus, was neither randomized nor blinded.

Collecting existing single-phenotype GWAS on externalizing phenotypes

A detailed definition of "externalizing phenotypes" was preregistered to delimit the data collection of single-phenotype GWAS summary statistics (Supplementary Information section 2.1). Summary statistics from existing studies were either provided by or downloaded from the public repositories of 23andMe, the Psychiatric Genomics Consortium (PGC), the Million Veterans Program (MVP), the International Cannabis Consortium (ICC), the GWAS & Sequencing Consortium of Alcohol and Nicotine Use (GSCAN), the Social Science Genetics Association Consortium (SSGAC), the Genetics of Personality Consortium (GPC), and the Broad Antisocial Behavior Consortium (Broad ABC), see Supplementary Information section 2.2 for more details. All GWAS that were considered for inclusion are listed in **Supplementary Table 1**, and **Supplementary Table 2** reports the underlying studies that had contributed to the seven GWAS (or GWAS meta-analysis) that were included the final multivariate model specification (see below).

GWAS in UK Biobank (UKB)

New GWAS were estimated in UKB (Supplementary Information section 2.3), of which summary statistics for "age at first sexual intercourse" and "Alcohol Use Disorder Identification Test problem items" (AUDIT-P) were later included in the final multivariate model (see below). The GWAS were performed with linear mixed models (BOLT-LMM[78]) and were statistically adjusted for sex, birth year, sex-specific birth-year interaction dummies, genotyping array and batch, and 40 genetic principal components (PCs). Two partly overlapping hold-out subsamples of UKB participants were excluded from all single-phenotype GWAS summary statistics that included UKB data, and the participants were instead retained as an independent sample for polygenic score analyses (Supplementary Information section 2.3.2). Genetic relatives (pairwise KING coefficient ≥ 0.0442) of the held-out individuals were excluded from the study altogether to ensure independence between the discovery and follow-up analyses. Whenever an existing GWAS (or meta-analysis) was based on UKB data, we re-estimated the UKB component using the same phenotype definition as in the existing study, while excluding the held-out participants and their genetic relatives. See Supplementary Information section 2.3.2 for further details.

GWAS inclusion criteria, quality control, and meta-analysis

All GWAS were performed among individuals that (a) were of European ancestry, (b) were observed for all relevant covariates, (c) were successfully genotyped and passed standardized sample-level quality control (according to study-specific protocols[12–15,21,79]), and (d) were unrelated (unless a particular GWAS was estimated with linear mixed models). Genotypes were imputed with reference data from either the 1000 Genomes Consortium[80], the Haplotype Reference Consortium[81], the UK10K Consortium[82], or a combination thereof. We performed quality control of GWAS summary statistics with a whole-genome sequenced reference panel, assembled from 1000 Genomes Consortium[80] and UK10K Consortium[82] data (Supplementary Information section 2.4.1). Our quality-control procedure applied recommended[83] SNP-filtering to remove rare SNPs (minor allele frequency < 0.005), SNPs with an IMPUTE imputation quality (INFO) score less than 0.9, and otherwise low-quality variants (**Supplementary Table 3**). For a complete description of the quality-control procedure, see Supplementary Information section 2.4.

We performed sample-size weighted meta-analysis with METAL[84] (Supplementary Information section 2.5). Thereafter, we excluded any summary statistics that displayed

insufficient SNP-based heritability ($h^2 < 0.05$) or GWAS association signal ($\bar{\chi}^2 < 1.05$), estimated with LD Score regression[18,59]. At this stage, we had collected or generated well-powered summary statistics for eleven phenotype-specific GWAS (or meta-analysis) that satisfied our inclusion criteria and that were kept for exploratory factor analysis (**Supplementary Table 4**): (1) ADHD (N = 53,293), (2) age at first sexual intercourse (N = 357,187), (3) problematic alcohol use (N = 164,684), (4) automobile speeding propensity (N = 367,151), (5) alcoholic consumption (drinks per week; N = 375,768), (6) educational attainment (N = 725,186), (7) lifetime cannabis use (N = 186,875), (9) lifetime smoking initiation (N = 1,251,809), (9) general risk tolerance (N = 426,379), (10) irritability (N = 388,248), and (11) number of sexual partners (N = 336,121).

Exploratory factor analysis of genetic correlations

As an initial analysis to inform and guide the multivariate modeling process, we performed hierarchical clustering of a matrix with pair-wise LD Score genetic correlations (r_g) (Supplementary Information section 3). The GWAS effect-sizes of age at first sexual intercourse and educational attainment were reversed to anticipate positive correlations with externalizing liability. The 11 phenotypes displayed moderate-to-substantial genetic overlap with at least one other phenotype (max $|r_g|$ = 0.245–0.773), and the average $|r_g|$ across all pairwise correlations was 0.323 (**Supplementary Table 5**). Three clusters were identified: (1) attention deficit/hyperactivity disorder (ADHD), educational attainment (EDUC), age at first sexual intercourse (FSEX), irritability (IRRT), and smoking initiation (SMOK); (2) problematic alcohol use (ALCP), drinks per week (DRIN); and (3) lifetime cannabis use (CANN), automobile speeding propensity (DRIV), number of sexual partners (NSEX), general risk tolerance (RISK).

Following the preregistration, exploratory factor analysis tested four different factor solutions, specifying 1...k + 1 factors (Supplementary Information section 3.2), where k corresponds to the number of clusters identified in the genetic correlation matrix, while retaining factors that explained at least 15% of the variance (a preregistered threshold). Exploratory factor analysis found that the fourth factor explained only 12.5% of the variance, and thus, the three-factor solution was considered the most appropriate exploratory model in terms of capturing variation (**Supplementary Table 6**). The pattern of factor loadings was consistent with the hierarchical clustering. However, as we detail in Supplementary Information section 3.2, the second and third factor mainly accounted for complex residual variation and divergent residual cross-trait correlations among the subset of phenotypes that

had the weakest loadings on the single common factor. Thus, we learned from the exploratory factor analysis that some of the 11 indicators may not be optimal for identifying a single common genetic liability to externalizing, and that a less complex model specification with fewer indicators would perhaps perform better than a three-factor model in the subsequent confirmatory factor analysis.

Confirmatory factor analyses with Genomic SEM

We formally modelled genetic covariances (rather than genetic correlations) and performed confirmatory factor analyses using the method genomic structural equation modeling (Genomic SEM)[10] (Supplementary Information section 3.3). Genomic SEM is unbiased by sample overlap and differences in sample size in the discovery phenotypes, and by applying to GWAS summary statistics it allows for genetic analyses of latent factors in larger samples than is typically possible with individual-level data[10]. We competed four models: (1) a common factor model with the aforementioned 11 phenotypes, (2) a correlated three-factors model with the 11 phenotypes (with and without cross-loadings), (3) a bifactor model with the 11 phenotypes, and (4) a revised common factor model that only included seven of the phenotypes that satisfied moderate-to-large (*i.e.*, $\geq .50$) loadings on the single latent factor in model (1) (**Supplementary Table 7**). We found that model (4) was the only model that closely approximated the observed genetic covariance matrix ($\chi^2(12) = 390.234$, AIC = 422.234, CFI = .957, SRMR = .079), fulfilled our preregistered model fit criteria, and coalesced with theoretical expectations of a general shared genetic liability to externalizing. This model was selected as our final factor specification, and we hereafter refer to it as "the latent genetic externalizing factor", or simply, "the externalizing factor" (*EXT*). To explore the convergent and discriminant validity of the externalizing factor, we estimated genetic correlations between the externalizing factor and 92 traits from various research domains (**Supplementary Table 8**).

Multivariate GWAS analyses with Genomic SEM

Using Genomic SEM, we performed multivariate GWAS analysis by estimating SNP associations with the externalizing factor (*EXT*), which is our main discovery analysis (Supplementary Information section 3.4). We estimated the effective sample size of the resulting "externalizing GWAS" to be N_{eff} = 1,492,085. The GWAS displayed strong association signal, with a mean χ^2 and genomic inflation factor (λ_{GC}) of 3.114 and 2.337, respectively. Analyses with LD Score regression suggest that the strong inflation observed in

the association test statistic is attributable to polygenicity rather than bias from population stratification[10,18], as the LD Score intercept and attenuation ratio were estimated to be 1.115 (SE = 0.019) and 0.054 (SE = 0.009), respectively.

A conventional "clumping" algorithm was applied to identify near-independent genome-wide significant lead SNPs (two-sided $P < 5\times10^{-8}$)[85], which were then subjected to "multi-SNP-based conditional & joint association analysis using GWAS summary data" (COJO) to estimate conditional SNP associations[22,86] (Supplementary Information section 3.4.2). We identified 579 lead SNPs that were conditionally and jointly associated with *EXT*. We performed lookups of these "579 *EXT* SNPs", as well as any correlated SNPs ($r^2 > 0.1$), in the NHGRI-EBI GWAS Catalog[7] (version e96 2019-05-03) to investigate whether the identified loci have previously been found associated with other traits at suggestive significance (two-sided $P < 1\times10^{-5}$). To evaluate whether each SNP acted through the externalizing factor, we estimated genome-wide Q_{SNP} heterogeneity statistics with Genomic SEM (Supplementary Information section 3.5.1). The null hypothesis of the Q_{SNP} test is that SNP effects on the constituent phenotypes operate (i.e., are statistically mediated) via the *EXT* factor, so a significant Q_{SNP} test indicates that SNP association is better explained by a trait-specific pathway independent of the *EXT* factor. The Q_{SNP} analysis was sufficiently powered to identify substantial heterogeneity in the genome (160 near-independent genome-wide significant Q_{SNP} loci), but reassuringly, did not identify heterogeneity among 99% (571/579) of the *EXT* SNPs. **Supplementary Table 9–10** reports the results of the externalizing GWAS and the heterogeneity analysis, together with bioannotation with "functional mapping and annotation of genetic associations" (FUMA)[28].

Proxy-phenotype and quasi-replication analysis

We performed a preregistered proxy-phenotype[87] and quasi-replication[24] analysis by investigating the 579 SNPs (k) for association in two independent, second-stage GWAS on (1) alcohol use disorder (N = 202,004, r_g = 0.52) and (2) antisocial behavior (N = 32,574, r_g = 0.69) (Supplementary Information section 4). For SNPs missing from the two second-stage GWAS, we analyzed highly correlated proxy SNPs ($r^2 > 0.8$). Significant proxy-phenotype associations were evaluated for Bonferroni-corrected significance (two-sided test $P < 0.05/k$). For the quasi-replication exercises, we generated empirical null distributions for the two second-stage GWAS by randomly selecting 250 near-independent ($r^2 < 0.1$) SNPs matched on MAF (± 1 percentage point) for each of the k SNPs. The quasi-replication approach was performed in three steps: (1) a binomial test of sign concordance, which tested whether the

direction of effect of the k SNPs were in greater concordance between the externalizing GWAS and each of the second-stage GWAS compared to what would be expected by chance ($H_0 = 0.5$); (2) a binomial test of whether a greater proportion of the k SNPs were nominally significant (two-sided $P < 0.05$) in the second-stage GWAS compared to the empirical null distribution; (3) a test of joint enrichment, performed as a non-parametric (one-sided) Mann-Whitney test of the null hypothesis that the P values of the k SNPs are derived from the empirical null distribution. We strongly rejected the null hypotheses of all quasi-replication tests, suggesting that the externalizing GWAS is not spurious overall and that it was more enriched for association with the second-stage phenotypes than their respective polygenic background GWAS signal (**Supplementary Table 11–12**).

Polygenic score analyses

We generated polygenic scores by summing genotypes weighed by the effect sizes estimated in the externalizing GWAS, among individuals of European ancestry in five independent study cohorts: (1) Add Health[88,89], (2) COGA[90–92], (3) PNC[93,94], (4) the UKB siblings hold-out cohort[95], and (5) the BioVU biorepository[96] (Supplementary Information section 5). In each dataset, we generated three scores, of which two were adjusted for linkage disequilibrium (LD): (1) PRS-CS[45], (2) LDpred (infinitesimal model)[44], and (3) unadjusted scores[97], while using SNPs that overlapped with the high-quality consensus set defined by the HapMap 3 Consortium[98]. Accuracy was evaluated as the incremental R^2/pseudo-R^2 (ΔR^2) attained by adding the polygenic score to a regression model with baseline covariates, in accordance with previous efforts[16,99]. The baseline model included covariates for sex, age, and genetic principal components (PCs), and genotyping array and batch. The choice of statistical model (e.g., OLS vs. logit) and adjustment of standard errors depended on (1) the distribution of the phenotype and (2) the structure of the data in the study cohort (independent vs. clustered observations), see Supplementary Information section 5.2.4 for further details. We estimated 95% confidence intervals for ΔR^2 using percentile method bootstrapping with 1000 iterations.

In Add Health and COGA, we performed out-of-sample validation of *EXT* by modeling a latent externalizing factor using phenotypic data corresponding to the seven Genomic SEM phenotypes (Supplementary Information section 5.2.3) (**Supplementary Table 13–14**). In Add Health, COGA, PNC, and the UKB siblings hold-out cohort, we performed exploratory polygenic score analyses with a wide range of preregistered phenotypes from the behavioral, psychiatric, and socioeconomic research domains

(**Supplementary Table 16–19**). We performed a phenome-wide association study (PheWAS) of medical outcomes in the BioVU biorepository by fitting a logistic regression to 1,335 case/control disease "phecodes"[100] (N = 66,915) (**Supplementary Table 20**).

We performed within-family analyses in data on full siblings in Add Health, COGA, and the UKB siblings hold-out cohort (Supplementary Information section 5.2.5). We analyzed 492 families in Add Health (N_{siblings} = 994), 621 families in COGA (N_{siblings} = 1,353), and 19,252 families in the UKB (N_{siblings} = 39,640). In Add Health and COGA, we applied OLS to test the externalizing polygenic score for association with a single outcome: the factor scores of the phenotypic externalizing factor (a continuous variable), while adjusting for family fixed-effects (i.e., family-specific dummy variables) (**Supplementary Table 15**). We then compared the magnitude of the within-family coefficient ($\hat{\beta}_{WF}$) to the coefficient of an OLS model without family-specific intercepts ($\hat{\beta}$). In the UKB siblings hold-out cohort, we performed an analogous within-family analysis of the exploratory phenotypes (**Supplementary Table 19**). We analyzed heteroskedasticity-consistent and cluster-robust standard errors, clustered at the family level.

Bioannotation

We performed a series of bioannotation and bioinformatic analyses to identify relevant biological pathways (Supplementary Information section 6). The method "functional mapping and annotation of genetic associations" (FUMA v1.3.5e)[28] was applied to explore the functional consequences of the 579 SNPs (**Supplementary Table 9**), which included ANNOVAR categories (*i.e.*, the functional consequence of SNPs on genes), Combined Annotation Dependent Depletion (CADD) scores (*i.e.*, a measure of how deleterious a SNP is; CADD > 12.37 is classified as deleterious), RegulomeDB scores (*i.e.*, a categorical score from 1a to 7 with 1a corresponding to the most biological evidence that the SNP is a regulatory element), mapping to expression quantitative trait loci (eQTLs), and chromatin states (values range from 1 to 15, with values 1 to 7 referring to an open chromatin state). The sources of the external reference data used by FUMA are described in ref.[28].

Gene-based association analyses was performed by applying the method "multi-marker analysis of genomic annotation" (MAGMA v1.07)[28,29] (Supplementary Information sections 6.1.2). The method accounts for LD, which was calculated using reference data from European-ancestry 1000 Genomes participants[80]. Genome-wide SNPs were first mapped to 18,093 protein-coding genes from Ensembl (build 85)[101], and the SNPs within each gene

were then jointly tested for association with *EXT*. We evaluated Bonferroni-corrected significance, adjusted for the number of tested genes (one-sided $P < 2.76\times10^{-6}$) (**Supplementary Table 21**). Next, MAGMA gene-set analysis was performed using 15,477 curated gene sets and Gene Ontology (GO)[102] terms obtained from the Molecular Signatures Database (MsigDB v7.0)[103]. We evaluated Bonferroni-corrected significance, adjusted for the number of tested gene sets (one-sided $P < 3.23\times10^{-6}$) (**Supplementary Table 22**). Lastly, a gene property analysis tested the relationships between 54 tissue-specific gene expression profiles and gene associations, while adjusting for the average expression of genes per tissue type as a covariate (**Supplementary Table 23**), and between brain gene expression profiles and gene associations across 11 brain tissues from BrainSpan[104] (**Supplementary Table 24**). Gene expression values were $\log_2$ transformed average Reads Per Kilobase Million (RPKM) per tissue type (after replacing RPKM > 50 with 50) based on GTEx RNA-seq data[105]. We evaluated Bonferroni-corrected significance, adjusted for the number of tested profiles (one-sided $P < 9.26\times10^{-4}$).

We used an extension of MAGMA: "Hi-C coupled MAGMA" or "H-MAGMA" [30], to assign non-coding (intergenic and intronic) SNPs to cognate genes based on their chromatin interactions. Exonic and promoter SNPs were assigned to genes based on physical position. We used four Hi-C datasets provided with the software, derived from adult brain[106], fetal brain[107], and iPSC derived neurons and astrocytes[108]. We evaluated Bonferroni-corrected significance, adjusted the number of tests within each of the four Hi-C datasets (one-sided $P < 9.83–9.86\times10^{-7}$) (**Supplementary Tables 25–28**).

The method S-PrediXcan v0.6.2[109] was used to analyze the association of *EXT* with gene expression levels in different brain tissues. We used pre-computed tissue weights from the Genotype-Tissue Expression (GTEx, v8) project database as the reference transcriptome dataset[105]. As input data, we used the *EXT* summary statistics, LD matrices of the SNPs (available at the PredictDB Data Repository, http://predictdb.org), and transcriptome tissue data related to 13 brain tissues: anterior cingulate cortex, amygdala, caudate basal ganglia, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, nucleus accumbens basal ganglia, putamen basal ganglia, spinal cord and substantia nigra. We evaluated transcriptome-wide significance at the two-sided test $P < 2.77\times10^{-7}$, which is the Bonferroni-corrected threshold adjusted for 13 tissues times 13,876 tested genes (180,388 gene-tissue pairs) (**Supplementary Table 29**). In **Supplementary Table 30** we summarize the genes findings across the bioannotation analyses.

Main References:

1. Richmond-Rakerd, L. S. *et al.* Clustering of health, crime and social-welfare inequality in 4 million citizens from two nations. *Nat. Hum. Behav.* **4**, 255–264 (2020).

2. Case, A. & Deaton, A. Mortality and Morbidity in the 21st Century. *Brookings Pap. Econ. Act.* **2017**, 397–476 (2017).

3. Achenbach, T. M. The classification of children's psychiatric symptoms: A factor-analytic study. *Psychol. Monogr. Gen. Appl.* **80**, 1–37 (1966).

4. Hicks, B. M., Krueger, R. F., Iacono, W. G., McGue, M. & Patrick, C. J. Family transmission and heritability of externalizing disorders: a twin-family study. *Arch. Gen. Psychiatry* **61**, 922–928 (2004).

5. Krueger, R. F. *et al.* Etiologic connections among substance dependence, antisocial behavior and personality: Modeling the externalizing spectrum. *J. Abnorm. Psychol.* **111**, 411–424 (2002).

6. Sullivan, P. F. & Geschwind, D. H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell* **177**, 162–183 (2019).

7. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2018).

8. Zhou, H. *et al.* Association of OPRM1 Functional Coding Variant With Opioid Use Disorder: A Genome-Wide Association Study. *JAMA Psychiatry* (2020). doi:10.1001/jamapsychiatry.2020.1206

9. Mullins, N. *et al.* GWAS of Suicide Attempt in Psychiatric Disorders and Association With Major Depression Polygenic Risk Scores. *Am. J. Psychiatry* **176**, 651–660 (2019).

10. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* (2019). doi:https://doi.org/10.1038/s41562-019-0566-x

11. Kendler, K. S. & Myers, J. The boundaries of the internalizing and externalizing genetic spectra in men and women. *Psychol. Med.* **44**, 647–655 (2013).

12. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).

13. Walters, R. K. *et al.* Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656–1669 (2018).

14. Sanchez-Roige, S. *et al.* Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *Am. J. Psychiatry* **176**, 107–118 (2018).

15. Pasman, J. A. *et al.* GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia. *Nat. Neurosci.* **21**, 1161–1170 (2018).

16. Karlsson Linnér, R. *et al.* Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* **51**, 245–257 (2019).

17. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).

18. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

19. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).

20. Townsend, P. *Health and deprivation: Inequality and the North*. (Croom Helm, 1988).

21. Karlsson Linnér, R. *et al.* Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* **51**, 245–257 (2019).

22. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1-3 (2012).

23. Hart, A. B. & Kranzler, H. R. Alcohol Dependence Genetics: Lessons Learned From Genomae-Wide Association Studies (GWAS) and Post-GWAS Analyses. *Alcohol. Clin. Exp. Res.* **39**, 1312–27 (2015).

24. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**, 624–633 (2016).

25. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).

26. Leshner, A. I. Addiction Is a Brain Disease, and It Matters. *Science* **278**, 45–47 (1997).

27. Moffitt, T. E. The Cambridge handbook of violent behavior and aggression. in *The Cambridge handbook of violent behavior and aggression* (eds. Flannery, D. J., Vazsonyi, A. T. & Waldman, I. D.) 49–74 (Cambridge University Press, 2007).

28. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

29. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, 1–19 (2015).

30. Sey, N. Y. A. *et al.* A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat.*

Neurosci. **23**, 583–593 (2020).

31. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

32. Jaykumar, A. B. *et al.* Role of Alström syndrome 1 in the regulation of blood pressure and renal function. *JCI Insight* **3**, (2018).

33. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

34. Li, Y. R. *et al.* Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat. Med.* **21**, 1018–1027 (2015).

35. Sanchez-Roige, S. *et al.* Genome-wide association studies of impulsive personality traits (BIS-11 and UPPS-P) and drug experimentation in up to 22,861 adult research participants identify loci in the CACNA1I and CADM2 genes. *J. Neurosci.* **39**, 2562–2572 (2019).

36. Zhao, B. *et al.* Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat. Genet.* **51**, 1637–1644 (2019).

37. Edenberg, H. J. *et al.* Variations in GABRA2, Encoding the α2 Subunit of the GABAA Receptor, Are Associated with Alcohol Dependence and with Brain Oscillations. *Am. J. Hum. Genet.* **74**, 705–714 (2004).

38. Dick, D. M. *et al.* The Role of GABRA2 in Risk for Conduct Disorder and Alcohol and Drug Dependence across Developmental Stages. *Behav. Genet.* **36**, 577–590 (2006).

39. Duman, R. S., Sanacora, G. & Krystal, J. H. Altered connectivity in depression: GABA and glutamate neurotransmitter deficits and reversal by novel treatments. *Neuron* **102**, 75–90 (2019).

40. Brambilla, P., Perez, J., Barale, F., Schettini, G. & Soares, J. C. GABAergic dysfunction in mood disorders. *Mol. Psychiatry* **8**, 721–737 (2003).

41. Hill, W. D. *et al.* A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol. Psychiatry* **24**, 169–181 (2019).

42. Lee, P. H. *et al.* Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469–1482 (2019).

43. Schrode, N. *et al.* Synergistic effects of common schizophrenia risk variants. *Nat. Genet.* **51**, 1475–1485 (2019).

44. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

45. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via

Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).

46. Derzon, J. H. The correspondence of family features with problem, aggressive, criminal, and violent behavior: A meta-analysis. *J. Exp. Criminol.* (2010). doi:10.1007/s11292-010-9098-0

47. O'Brien, D. T., Farrell, C. & Welsh, B. C. Broken (windows) theory: A meta-analysis of the evidence for the pathways from neighborhood disorder to resident health outcomes and behaviors. *Social Science and Medicine* (2019). doi:10.1016/j.socscimed.2018.11.015

48. Chang, L. Y., Wang, M. Y. & Tsai, P. S. Neighborhood disadvantage and physical aggression in children and adolescents: A systematic review and meta-analysis of multilevel studies. *Aggress. Behav.* (2016). doi:10.1002/ab.21641

49. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).

50. Ruderfer, D. M. *et al.* Significant shared heritability underlies suicide attempt and clinically predicted probability of attempting suicide. *Mol. Psychiatry* (2019). doi:10.1038/s41380-018-0326-8

51. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).

52. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018).

53. Selzam, S. *et al.* Comparing Within- and Between-Family Polygenic Score Prediction. *Am. J. Hum. Genet.* **105**, 351–363 (2019).

54. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

55. Stahl, E. A. *et al.* Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).

56. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

57. Tielbeek, J. J. *et al.* Genome-wide association studies of a broad spectrum of antisocial behavior. *JAMA Psychiatry* **74**, 1242 (2017).

58. Kranzler, H. R. *et al.* Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat. Commun.* **10**, 1499 (2019).

59. Bulik-Sullivan, B. K. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

60. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain.

Science **360**, eaap8757 (2018).

61. Gage, S. H., Smith, G. D., Ware, J. J., Flint, J. & Munafò, M. R. G = E: What GWAS Can Tell Us about the Environment. *PLOS Genet.* **12**, e1005765 (2016).

62. Fox, D. Subversive science. *Penn State Law Rev.* **124**, 153–191 (2019).

63. Goddard, H. H. *The Kallikak family: A study in the heredity of feeble mindedness*. (MacMillan, 1912).

64. Kevles, D. J. *In the Name of Eugenics: Genetics and the Uses of Human Heredity*. (Harvard University Press, 1998).

65. Ladd-Taylor, M. *Fixing the Poor: Eugenic Sterilization and Child Welfare in the Twentieth Century*. (Johns Hopkins University Press, 2017).

66. Carlson, J. & Harris, K. Quantifying and contextualizing the impact of bioRxiv preprints through automated social media audience segmentation. *bioRxiv* (2020). doi:10.1101/2020.03.06.981589

67. Murray, C. *Human Diversity: The Biology of Gender, Race, and Class*. (Twelve, 2020).

68. NPR.org. What Happened When Dylann Roof Asked Google For Information About Race? (2017).

69. Kvaale, E. P., Gottdiener, W. H. & Haslam, N. Biogenetic explanations and stigma: A meta-analytic review of associations among laypeople. *Soc. Sci. Med.* **96**, 95–103 (2013).

70. Lebowitz, M. S., Tabb, K. & Appelbaum, P. S. Asymmetrical genetic attributions for prosocial versus antisocial behaviour. *Nat. Hum. Behav.* **3**, 940–949 (2019).

71. McSwiggan, S., Elger, B. & Appelbaum, P. S. The forensic use of behavioral genetics in criminal proceedings: Case of the MAOA-L genotype. *Int. J. Law Psychiatry* **50**, 17–23 (2017).

72. Morris, T. T., Davies, N. M. & Davey Smith, G. Can education be personalised using pupils' genetic data? *Elife* **9**, e49962 (2020).

73. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

74. Tucker-Drob, E. M., Briley, D. A. & Harden, K. P. Genetic and environmental influences on cognition across development and context. *Curr. Dir. Psychol. Sci.* **22**, 349–355 (2013).

75. Fletcher, J. M. Why have tobacco control policies stalled? Using genetic moderation to examine policy impacts. *PLoS One* **7**, e50576 (2012).

76. Papageorge, N. W., Ronda, V. & Zheng, Y. *The Economic Value of Breaking Bad: Misbehavior, Schooling and the Labor Market*. (2019).

77. Lundberg, S. The College Type: Personality and Educational Inequality. *J. Labor Econ.* **31**, 421–441 (2013).

78. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

79. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).

80. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

81. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

82. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

83. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).

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

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

86. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

87. Rietveld, C. A. *et al.* Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13790–13794 (2014).

88. Harris, K. M., Halpern, C. T., Haberstick, B. C. & Smolen, A. The National Longitudinal Study of Adolescent Health (Add Health) sibling pairs data. *Twin Res. Hum. Genet.* **16**, 391–8 (2013).

89. McQueen, M. B. *et al.* The National Longitudinal Study of Adolescent to Adult Health (Add Health) sibling pairs genome-wide data. *Behav. Genet.* **45**, 12–23 (2015).

90. Begleiter, H. The Collaborative Study on the Genetics of Alcoholism. *Alcohol Health Res. World* **19**, 228–236 (1995).

91. Edenberg, H. J. The collaborative study on the genetics of alcoholism: An update. *Alcohol Res. Heal.* (2002).

92. Bucholz, K. K. *et al.* Comparison of Parent, Peer, Psychiatric, and Cannabis Use Influences Across Stages of Offspring Alcohol Involvement: Evidence from the COGA Prospective Study. *Alcohol. Clin. Exp. Res.* (2017). doi:10.1111/acer.13293

93. Calkins, M. E. *et al.* The Philadelphia Neurodevelopmental Cohort: constructing a

deep phenotyping collaborative. *J Child Psychol Psychiatry* **56**, 1356–1369 (2016).

94. Satterthwaite, T. D. *et al.* The Philadelphia Neurodevelopmental Cohort: a publicly available resource for the study of normal and abnormal brain development in youth. *Neuroimage* **124**, 1115–1119 (2016).

95. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

96. Davis, L. Psychiatric Genomics, Phenomics, and Ethics Research In A 270,000-Person Biobank (BioVU). *Eur. Neuropsychopharmacol.* **29**, S739–S740 (2019).

97. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, (2013).

98. Altshuler, D. M., Gibbs, R. A. & Peltonen, L. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

99. Lee, J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

100. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* **12**, e0175508 (2017).

101. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).

102. Consortium, T. G. O. The Gene Ontology project in 2008. *Nucleic Acids Res.* **36**, D440–D444 (2007).

103. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–40 (2011).

104. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).

105. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–5 (2013).

106. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).

107. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).

108. Rajarajan, P. *et al.* Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* **362**, (2018).

109. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1–20

(2018).

Acknowledgements This research was carried out under the auspices of the Externalizing Consortium. The study was classified as secondary research of de-identified subjects, and the study was awarded ethical approval by the internal review board (IRB) of Virginia Commonwealth University (VCU), with reference number HM20019386. These analyses were made possible by the generous public sharing of summary statistics from published GWAS from the Psychiatric Genomics Consortium (PGC), the Million Veterans Program (MVP), the International Cannabis Consortium (ICC), the GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN), the Social Science Genetics Association Consortium (SSGAC), the Genetics of Personality Consortium (GPC), and the Broad Antisocial Behavior Consortium (BroadABC). We would like to thank the many studies that made these consortia possible, the researchers involved, and the participants in those studies, without whom this effort would not be possible. We would also like to thank the research participants and employees of 23andMe for making this work possible. This research was conducted in part using the UK Biobank Resource under applications 40830 and 11425. We thank all UKB cohort participants for making this study possible. We thank Lea K. Davis for providing access to BioVU. Finally, we thank the Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, H. Edenberg, L. Bierut, includes eleven different centers: University of Connecticut (V. Hesselbrock); Indiana University (H.J. Edenberg, J. Nurnberger Jr., T. Foroud; Y. Liu); University of Iowa (S. Kuperman, J. Kramer); SUNY Downstate (B. Porjesz); Washington University in St. Louis (L. Bierut, J. Rice, K. Bucholz, A. Agrawal); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield, A. Brooks); Department of Biomedical and Health Informatics, The Children's

Hospital of Philadelphia; Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA (L. Almasy), Virginia Commonwealth University (D. Dick), Icahn School of Medicine at Mount Sinai (A. Goate), and Howard University (R. Taylor). Other COGA collaborators include: L. Bauer (University of Connecticut); J. McClintick, L. Wetherill, X. Xuei, D. Lai, S. O'Connor, M. Plawecki, S. Lourens (Indiana University); G. Chan (University of Iowa; University of Connecticut); J. Meyers, D. Chorlian, C. Kamarajan, A. Pandey, J. Zhang (SUNY Downstate); J.C. Wang, M. Kapoor, S. Bertelsen (Icahn School of Medicine at Mount Sinai); A. Anokhin, V. McCutcheon, S. Saccone (Washington University); J. Salvatore, F. Aliev, B. Cho (Virginia Commonwealth University); and Mark Kos (University of Texas Rio Grande Valley). A. Parsian and H. Chen are the NIAAA Staff Collaborators. All studies included in the externalizing GWAS are listed in the Supplementary Information.

Funding Initial analyses by the Externalizing Consortium were funded by the National Institute of Alcohol Abuse and Alcoholism through an administrative supplement to R01AA015146. D.M.D. was supported through funding from the National Institute of Alcohol Abuse and Alcoholism (K02AA018755, U10AA008401, and P50AA0022527). P.D.K. was supported through a European Research Council Consolidator Grant (647648 EdGe). K.P.H. was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development: (R01HD092548 and R01HD083613) and the Jacobs Foundation. A.A.P. was supported by the National Institute of Alcohol Abuse and Alcoholism (R01AA026281) and the National Institute of Drug Abuse (P50DA037844). S.S-R. was supported through a NARSAD Young Investigator Award from the Brain and Behavior Foundation (grant number 27676). Both A.A.P. and S.S-R. were supported by funds from the California Tobacco-Related Disease Research Program (TRDRP, grant numbers 28IR-0070 and T29KT0526). The content of this article is solely the responsibility

of the authors and does not necessarily represent the official views of the above funding bodies. This research used data from Add Health, a program project directed by K.M.H. (principal investigator) and designed by J. R. Udry, P. S. Bearman, and K.M.H. at the University of North Carolina at Chapel Hill, and funded by grant P01HD031921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Information on how to obtain the Add Health data files is available on the Add Health website (www.cpc.unc.edu/addhealth). This research used Add Health GWAS data funded by Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) grants R01HD073342 to K.M.H. (principal investigator) and R01HD060726 to K.M.H., J. D. Boardman, and M. B. McQueen (multiple principal investigators). COGA is a national collaborative study supported by NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism and the National Institute on Drug Abuse. Data were obtained from Vanderbilt University Medical Center's BioVU which is supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH funded Shared Instrumentation Grant S10RR025141; and CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, R01HD074711; and additional funding sources listed at https://victr.vumc.org/biovu-funding/. Support for data collection for the Philadelphia Neurodevelopment Cohort (PNC), acquired through dbGaP (accession number phs000607.v3.p2), was provided by grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson. Subjects were recruited and genotyped through the Center for Applied Genomics (CAG) at The Children's Hospital in Philadelphia (CHOP). Phenotypic data collection occurred at the CAG/CHOP and at the Brain Behavior

Laboratory, University of Pennsylvania. A full list of funding for investigator effort is listed in the supplementary material.

Author contributions D.M.D. and P.D.K. conceived the study. The study protocol was developed by D.M.D., K.P.H., R.K.L., P.D.K, T.T.M., and A.A.P. D.M.D., K.P.H., P.D.K., and A.A.P. jointly oversaw the study. D.M.D. and R.K.L. led the writing of the manuscript, with substantive contributions to the writing from K.P.H., P.D.K. and A.A.P. R.K.L. and T.T.M. were the lead analysts, responsible for conducting genome-wide association studies, quality control, meta-analysis, genetic correlations, and multivariate analyses with Genomic SEM, with assistance from A.D.G. R.K.L. performed the proxy-phenotype analyses. P.B.B. led the polygenic score analyses, and R.K.L. and T.T.M. contributed to those analyses. S.S-R performed the PheWAS in BioVU. S.S-R. led the bioinformatics analyses, and R.K.L contributed to those analyses. P.B.B., R.K.L, T.T.M., and S.S-R. prepared the tables and figures, with assistance from M.N.D, J.W.M., and H.E.P. J.J.T, E.C.J., M.L., H.Z., R.K., and J.A.P. prepared cohort-level GWAS meta-analyses under supervision of K.J.H.V., D.J.L., S.V., H.R.K., and J.G. K.M.H. assisted with analyses performed in the AddHealth study cohort. A.D.G., E.T-D., and I.W. provided helpful advice and feedback on various aspects of the study design. All authors contributed to and critically reviewed the manuscript. R.K.L., T.T.M, P.B.B., and S.S-R. made especially major contributions to the writing and editing; these authors contributed equally.

Competing interests Dr. Kranzler is a member of the American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative, which was supported in the last three years by AbbVie, Alkermes, Ethypharm, Indivior, Lilly, Lundbeck, Otsuka, Pfizer, Arbor, and Amygdala Neurosciences. Drs. Kranzler and Gelernter are named as inventors on PCT patent application #15/878,640 entitled: "Genotype-guided dosing of opioid agonists,"

filed January 24, 2018. Dr. Gelernter did paid editorial work for the journal Complex Psychiatry. Authors declare no other competing interests.

Data and code availability All data sources are described in the Supplementary Information. No new data was collected as part of this study. Only data from existing studies or study cohorts were analyzed, some of which are restricted access to protect the privacy of the study participants. GWAS summary statistics for the externalizing (*EXT*) GWAS (our main discovery analysis) can be obtained by following the procedures detailed at https://externalizing.org/request-data/. The summary statistics are derived from analyses based in part on 23andMe data, for which we can only publicly report results for up to 10,000 SNPs. The full set of externalizing GWAS summary statistics can be made available to qualified investigators who enter into an agreement with 23andMe that protects participant confidentiality. Once the request has been approved by 23andMe, a representative of the Externalizing Consortium can share the full set of summary statistics. All code necessary to replicate this study is available upon request.

Additional information Supplementary Information is available for this paper. Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper. Correspondence and requests for materials should be addressed to Richard Karlsson Linnér at r.karlssonlinner@vu.nl.