nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-021-00908-3

Check for updates

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

Richard Karlsson Linnér[1,41], Travis T. Mallard[2,41], Peter B. Barr[3,41], Sandra Sanchez-Roige[4,5,41], James W. Madole[2], Morgan N. Driver[6], Holly E. Poore[7], Ronald de Vlaming[1], Andrew D. Grotzinger[2], Jorim J. Tielbeek[8], Emma C. Johnson[9], Mengzhen Liu[10], Sara Brin Rosenthal[11], Trey Ideker[12], Hang Zhou[13,14], Rachel L. Kember[15,16], Joëlle A. Pasman[17], Karin J. H. Verweij[18], Dajiang J. Liu[19,20], Scott Vrieze[10], COGA Collaborators*, Henry R. Kranzler[15,16], Joel Gelernter[13,14,21,22], Kathleen Mullan Harris[23,24], Elliot M. Tucker-Drob[2,25], Irwin D. Waldman[7,26], Abraham A. Palmer[4,27,42], K. Paige Harden[2,25,42], Philipp D. Koellinger[1,28,42] and Danielle M. Dick[3,6,42]

Behaviors and disorders related to self-regulation, such as substance use, antisocial behavior and attention-deficit/hyperactivity disorder, are collectively referred to as externalizing and have 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 loci were enriched for genes expressed in the brain and related to nervous system development. A polygenic score constructed from our results predicts a range of behavioral and medical outcomes that were not part of 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 trait with complex and far-reaching social and health correlates.

Behaviors 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 studies have demonstrated that externalizing liability is highly heritable (~80%)[4,5]. To date, however, no large-scale molecular genetic studies have utilized the extensive degree of genetic overlap among externalizing traits to aid gene discovery, as most studies have focused on individual disorders[6]. For many high-cost, high-risk behaviors with an externalizing component—opioid use disorder and suicide attempts[7] being salient examples—there are limited genotyped cases available 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

[1]Department of Economics, Vrije Universiteit Amsterdam, Amsterdam, the 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, the 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]Center for Computational Biology and Bioinformatics, Department of Medicine, University of California San Diego, La Jolla, CA, USA. [12]Department of Medicine, University of California San Diego, La Jolla, CA, USA. [13]Department of Psychiatry, Yale University School of Medicine, West Haven, CT, USA. [14]Department of Psychiatry, VA CT Healthcare System, West Haven, CT, USA. [15]Center for Studies of Addiction, University of Pennsylvania School of Medicine, Philadelphia, PA, USA. [16]Mental Illness Research Education and Clinical Center, Crescenz VA Medical Center, Philadelphia, PA, USA. [17]Behavioural Science Institute, Radboud University Nijmegen, Nijmegen, the Netherlands. [18]Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands. [19]Department of Public Health Sciences, Penn State University, Hershey, PA, USA. [20]Institute of Personalized Medicine, Penn State University, Hershey, PA, USA. [21]Department of Genetics, Yale University School of Medicine, West Haven, CT, USA. [22]Department of Neuroscience, Yale University School of Medicine, West Haven, CT, USA. [23]Department of Sociology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [24]Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [25]Population Research Center, University of Texas at Austin, Austin, TX, USA. [26]Center for Computational and Quantitative Genetics, Emory University, Atlanta, GA, USA. [27]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA. [28]La Follette School of Public Affairs, University of Wisconsin–Madison, Madison, WI, USA. [41]These authors contributed equally: Richard Karlsson Linnér, Travis T. Mallard, Peter B. Barr, Sandra Sanchez-Roige. [42]These authors jointly supervised this work: Abraham A. Palmer, K. Paige Harden, Philipp D. Koellinger, Danielle M. Dick. *A list of authors and their affiliations appears at the end of the paper. e-mail: koellinger@wisc.edu; ddick@vcu.eduddick@vcu.edu

Table 1 | Summary of seven externalizing-related disorders and behaviors with GWAS summary statistics

Phenotype	*N*	h^2 (s.e.)	λ_{GC}	Mean χ^2	Intercept	Ratio	Ref.
ADHD	53,293	0.235 (0.015)	1.253	1.297	1.034	0.113	13
ALCP	164,684	0.055 (0.004)	1.149	1.174	1.013	0.073	14,15
CANN	186,875	0.066 (0.004)	1.230	1.267	1.026	0.098	16
FSEX[a]	357,187	0.115 (0.004)	1.623	1.869	1.036	0.041	17
NSEX	336,121	0.097 (0.004)	1.492	1.682	1.027	0.041	17
RISK	426,379	0.053 (0.002)	1.372	1.461	1.019	0.041	17
SMOK	1,251,809	0.078 (0.002)	2.328	3.152	1.126	0.058	18

The statistics reported in this table were all estimated with LD Score regression[12]. Heritability (h^2) is on the observed scale[12]. The genomic inflation factor, λ_{GC}, is the median χ^2 statistic divided by the expected median of the χ^2 distribution with 1 d.f.[62]. Mean χ^2 is the average χ^2 statistic. The intercept is the estimated LD Score regression intercept. The ratio measures stratification bias, defined as: (intercept − 1)/(mean χ^2 − 1)[12]. [a]Reverse-coded (see Methods).

genetically correlated traits[10]. Multivariate approaches can use summary statistics from genome-wide association studies (GWAS) to discover connections between phenotypes not typically studied together because they span different domains, fields of study or life stages. New statistical methods can increase the effective sample size by adjusting for sample overlap. Elucidating the shared genetic basis of externalizing liability can advance our understanding of the developmental etiology of self-regulation and enables mapping the pathways by which genetic risk and socio-environmental factors contribute to the development of externalizing outcomes.

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

Results

Genomic SEM of externalizing liability. Following our preregistered analysis plan (https://doi.org/10.17605/OSF.IO/XKV36), we collated summary statistics from GWAS on externalizing-related traits (Supplementary Methods). For an exhaustive description of the phenotype selection procedure and GWAS protocol, see the Supplementary Methods. All phenotypes considered for inclusion are listed in Supplementary Table 1. We first applied quality control (Supplementary Table 2) and excluded summary statistics based on power considerations (that is, linkage disequilibrium (LD) Score regression $h^2 < 0.05$, or mean $\chi^2 < 1.05$)[12]. After applying these filters, 11 externalizing phenotypes remained, with sample sizes >50,000 (N = 53,293–1,251,809; Supplementary Table 3). All samples were of European ancestry. The following seven phenotypes made it to the final multivariate model specification (Table 1 and Supplementary Table 4): (1) attention-deficit/hyperactivity disorder (ADHD)[13], (2) problematic alcohol use (ALCP; a meta-analysis of alcohol dependence and 'alcohol use disorder identification test problem items' (AUDIT-P))[14,15], (3) lifetime cannabis use (CANN)[16], (4) reverse-coded age at first sexual intercourse (FSEX; Methods)[17], (5) number of sexual partners (NSEX)[17], (6) general risk tolerance (RISK)[17] and (7) lifetime smoking initiation (SMOK)[18].

For a complete description of the model selection procedure, see the Supplementary Methods. In summary, before genomic SEM, we first applied hierarchical clustering to a matrix of LD) score genetic correlations, which identified three (k) clusters (Supplementary Table 5). An exploratory factor analysis benchmarked four factor models, specifying one to four (k+1) latent factors, with the aim to best explain the genetic correlations among the 11 phenotypes (Supplementary Table 6). The three-factor solution was determined to be the best-fitting exploratory model, which aligned with the hierarchical clustering.

We proceeded with confirmatory factor analysis to formally model genetic covariances with genomic SEM, which is unbiased by sample overlap and sample-size imbalances[10,19]. As indicated by its model fit indices ($\chi^2(44) = 8007.35$; Akaike information criterion (AIC) = 8051.35; comparative fit index (CFI) = 0.662; standardized root mean square residual (SRMR) = 0.161), we found that a common factor model with 11 phenotypes did not satisfy our preregistered criteria (that is, CFI and SRMR were >0.9 and <0.08, respectively). Two more complicated specifications were tested, a correlated three-factor model (that is, akin to the best-fitting exploratory model) and a bifactor model (Supplementary Table 7), but neither of these two models met the criteria or provided a parsimonious interpretation. Finally, we estimated a revised and less complex common factor model with the 7 phenotypes (Table 1 and Fig. 1a) that displayed moderate-to-large (that is, ≥0.5) loadings on the single factor estimated in the first common factor model with 11 phenotypes. The revised common factor model with 7 externalizing phenotypes provided the best fit across all specifications, and it closely approximated the observed genetic covariance matrix (that is, $\chi^2(12) = 390.234$, AIC = 422.234, CFI = 0.957 and SRMR = 0.079). This model was selected as our final factor model because it identified a genetic factor of externalizing that was suitable for genome-wide association analysis, offered an easily interpretable factor solution and satisfied the model fit criteria. We hereafter refer to it as 'the externalizing factor' (EXT).

The common factor captures a shared genetic liability to the final seven externalizing traits (Fig. 1b), and genetic variants associated with EXT predict central externalizing disorders and a range of behavioral and medical outcomes that were not in the model (see below). We performed a leave-one-phenotype-out genomic SEM analysis to ensure that no single phenotype, for example, the phenotype with the largest N, was unduly influencing the genetic architecture estimated for EXT (Supplementary Methods). We found that the genetic correlations between EXT and each of seven leave-one-phenotype-out models were not distinguishable from unity ($r_g \sim 0.984–0.999$, s.e. ~ 0.028–0.035), which suggests that none of the phenotypes are driving the genetic architecture of EXT.

We extended genomic SEM to estimate genetic correlations between EXT and 91 preregistered phenotypes with GWAS summary statistics that were not 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 (Fig. 1c): As anticipated, EXT showed strong positive genetic correlations with drug exposure ($r_g = 0.91$, s.e. = 0.09), antisocial behavior ($r_g = 0.65$, s.e. = 0.17) and impulsivity

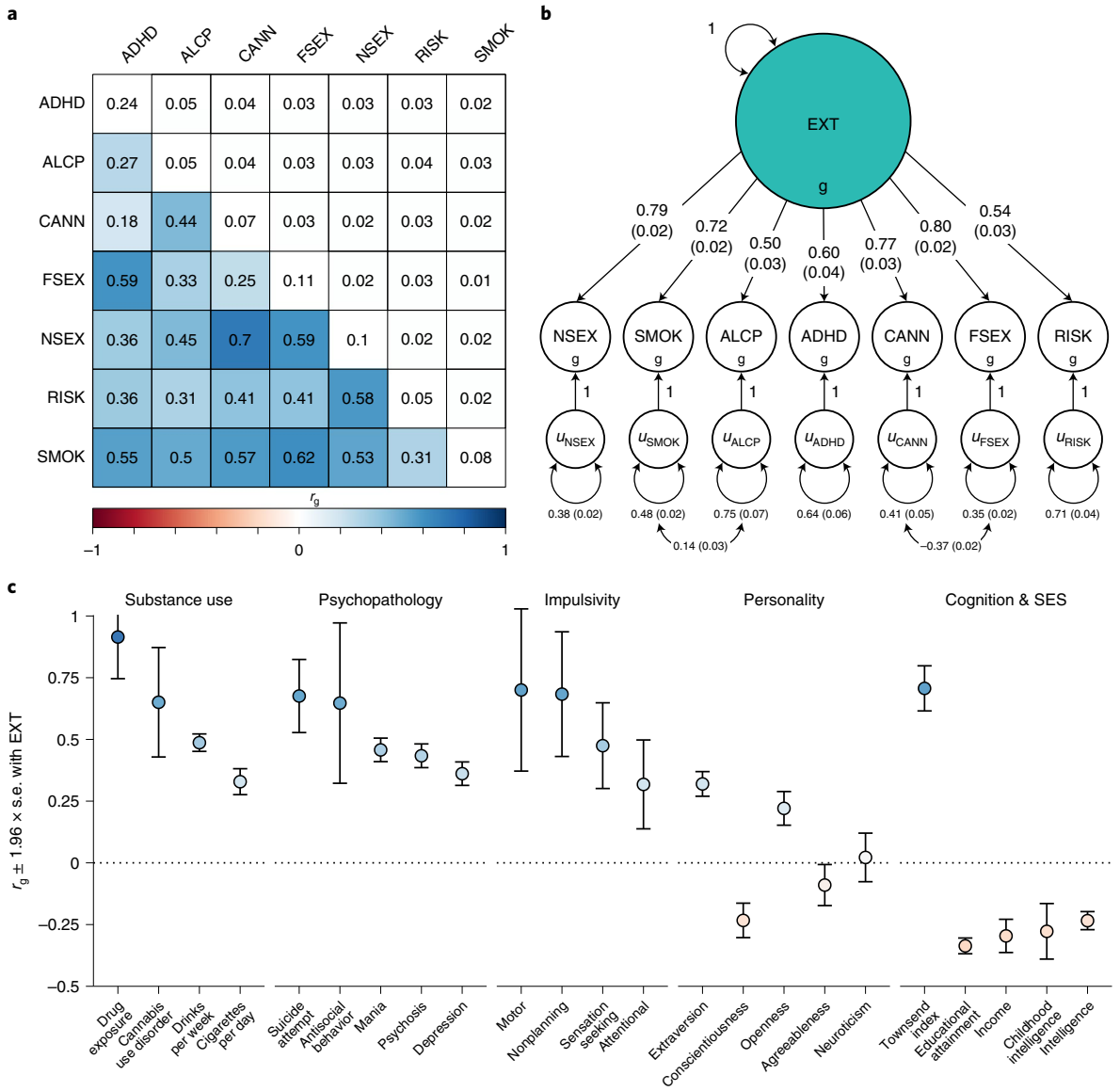


Fig. 1 | Genetic correlations and structural equation modeling with genomic SEM. a, The lower and upper triangles display pairwise LD Score genetic correlation (r_g) estimates and their s.e., respectively, among the final seven discovery phenotypes (Table 1), and the diagonal displays observed-scale SNP heritabilities (h^2; see Table 1 for s.e.). **b**, Path diagram of the final revised common factor model estimated with genomic SEM. The factor loadings were standardized, and s.e. are presented in parentheses. **c**, r_g estimates between the genetic EXT (N = 1,492,085) and a subset of phenotypes selected to establish convergent and discriminant validity (Supplementary Table 8 reports all 91 estimated genetic correlations together with the exact number of independent samples used to derive each estimate), where blue and red bars represent positive and negative r_g estimates, respectively, using the same color scale as in **a**. Error bars represent 95% confidence intervals (CIs) centered on the r_g estimate, computed as 1.96 times the s.e. ADHD, N = 53,293; ALCP, N = 164,864; CANN, N = 186,875; FSEX, N = 357,187; NSEX, N = 336,121; RISK, N = 426,379; SMOK, N = 1,251,809.

measures, including motor impulsivity (r_g = 0.70, s.e. = 0.17) and failures to plan (r_g = 0.68, s.e. = 0.13). We estimated similar genetic correlations with personality domains (based on 23andMe[20]) as to those reported in twin studies, that is, positive correlation with extraversion (r_g = 0.32, s.e. = 0.03), and negative with conscientiousness (r_g = −0.23, s.e. = 0.04) and agreeableness (r_g = −0.09, s.e. = 0.04)[11,21]. However, prior work has found neuroticism but not openness to be correlated with externalizing[21], while we found a positive correlation with openness (r_g = 0.22, s.e. = 0.04) but not with neuroticism (r_g = 0.02, s.e. = 0.05). Notably, EXT was also correlated with suicide attempts (r_g = 0.68, s.e. = 0.08) and post-traumatic stress disorder (r_g = 0.53, s.e. = 0.06). EXT showed more modest inverse correlations with educational attainment (r_g = −0.32, s.e. = 0.02) and intelligence (r_g = −0.23, s.e. = 0.02), indicating that EXT is not simply reflecting genetic influences on cognitive ability. Finally, there was a significant correlation with the Townsend index (r_g = 0.71, s.e. = 0.05), a measure of neighborhood deprivation that reflects high concentrations of unemployment, household overcrowding and lower home ownership and car ownership[22]. Genetic correlations can reflect correlated social processes or variables that are nonrandomly distributed with respect to genotypes, such as genetic nurture or neighborhood conditions, and we return to this topic in within-family analyses below.

Multivariate GWAS of externalizing liability. We next used genomic SEM[10] to conduct a GWAS on the shared genetic liability

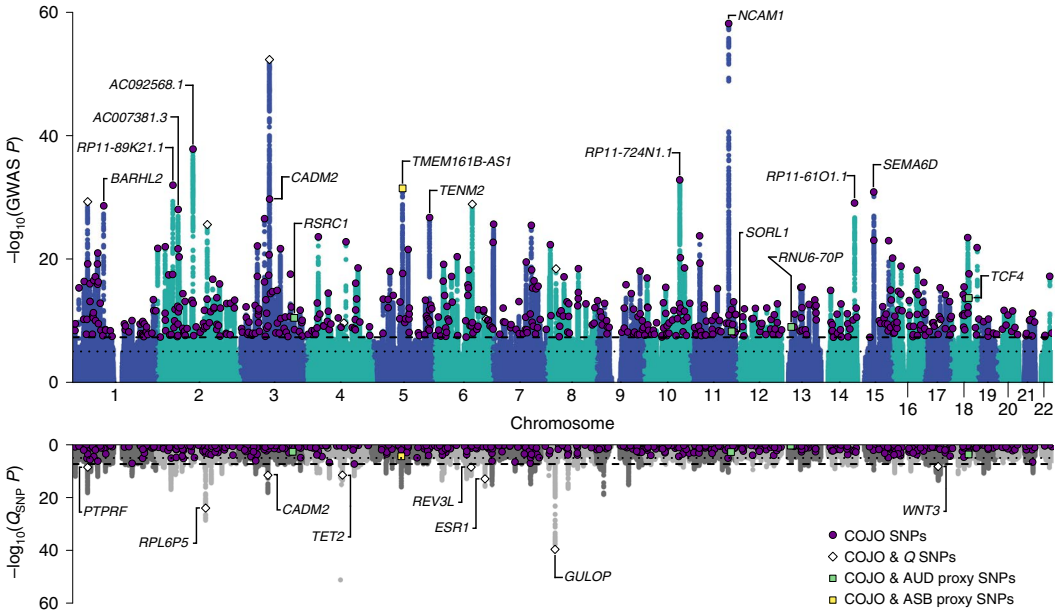


Fig. 2 | Multivariate genome-wide association analysis of EXT with genomic SEM. Scatterplot of $-\log_{10}$(*P* value for two-sided *z*-test) for weighted least-squares regression to estimate GWAS associations (top) and $-\log_{10}$(*P* value for one-sided χ^2 test with 7–1 d.f.) for Q_{SNP} tests of heterogeneity (bottom) for EXT. Purple dots represent the 579 EXT SNPs that were COJO at genome-wide significance (two-sided $P<5\times10^{-8}$; Supplementary Table 9). White diamonds represent 8 of the 579 SNPs that also showed significant Q_{SNP} heterogeneity. Four green squares and one yellow square represent 5 of the 579 SNPs that also were Bonferroni-significant proxy-phenotype associations with AUD and antisocial behavior (ASB), respectively (Supplementary Tables 11 and 12). Gene names refer to the closest gene based on genomic location, displayed for a selection of the findings (Supplementary Table 9 reports the nearest gene for all 579 EXT SNPs).

EXT (Fig. 2 and Extended Data Fig. 2). This analysis estimated single-nucleotide polymorphism (SNP) associations directly with EXT, with an effective sample size of $N=1{,}492{,}085$ individuals (Supplementary Methods). These analyses are different in their approach and substantially increase sample size, statistical power and the range of findings compared to previous work (Supplementary Methods). After applying conditional and joint multiple-SNP analysis on a set of near-independent, genome-wide significant (two-sided $P<5\times10^{-8}$) lead SNPs[23], we identified 579 conditionally and jointly associated (COJO) 'EXT SNPs' (Supplementary Tables 9 and 9B), 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 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 new, as they have not been reported previously for any trait in the GWAS literature (that is, neither of these 41 SNPs nor any SNPs in LD ($r^2>0.1$) were reported for any traits at two-sided $P<1\times10^{-5}$ in the NHGRI-EBI GWAS Catalog[6] (version e96 2019-05-03; Supplementary Table 10).

Genomic SEM was used to perform SNP-level tests of heterogeneity (Q_{SNP}; Supplementary Methods and Supplementary Data 1 and 2) to investigate whether each SNP had consistent, pleiotropic effects on the seven input phenotypes that effectively only operate via EXT. If the EXT loci really index a shared genetic externalizing liability, we would expect to identify heterogeneity mostly in regions of the genome not associated with EXT. In the absence of heterogeneity, it is expected that a given SNP's GWAS effects on the input phenotypes will scale proportionally to the factor loadings[24] (Supplementary Methods). The genome-wide Q_{SNP} analysis was adequately powered (mean $\chi^2_{(1)}=1.864$; Extended Data Fig. 2), and at one-sided Q_{SNP} $P<5\times10^{-8}$, we identified 160 Q_{SNP} loci (Supplementary Methods). Importantly, only 8 of these 160 loci overlapped with EXT loci (~1% = 8/579; Fig. 2 and Supplementary Table 9). Reassuringly, we identified 3.6 times more EXT loci than Q_{SNP} loci (579/160). Using a less stringent significance threshold by focusing specifically on the 579 EXT loci, only 7% (41/579) were significant for Q_{SNP} (one-sided Q_{SNP} $P<0.05/579$). The observation that a small minority of the EXT loci were heterogeneous at either significance threshold, and that the vast majority of the 160 Q_{SNP} loci were found outside EXT loci, provide evidence that the EXT loci primarily index a unitary dimension of genetic liability rather than representing an amalgamation of variants with divergent associations across the discovery phenotypes. Notably, the strongest Q_{SNP} and most salient example of a heterogeneous, trait-specific association is SNP rs1229984 (one-sided Q_{SNP} $P=1.67\times10^{-51}$; Supplementary Data 1). This particular SNP, located in the gene *ADH1B*, is a missense variant with a well-established role in alcohol metabolism[25], and it was not associated with EXT (two-sided $P=0.022$) but only with problematic alcohol use (two-sided $P=6.43\times10^{-57}$). Additionally, for each of the 579 EXT SNPs, we investigated the concordance in direction of SNP effects (that is, the sign) on the seven phenotypes (Supplementary Methods). For 317 of the 579 EXT SNPs (54.7%), the concordance was perfect (that is, the same direction of effect on all seven phenotypes), and for 203 (35.1%), 47 (8.1%) and 12 (2.1%) EXT SNPs, we observed six, five and four concordant effects, respectively. Thus, the analysis of sign concordance lends further support to our interpretation that the EXT loci primarily index a shared genetic liability to externalizing.

Quasi-replication analyses. Because the discovery stage effectively exhausted large study cohorts available for replication, we performed a series of preregistered quasi-replication analyses (Supplementary Tables 11 and 12). As quasi-replication analyses of the 579 SNPs (Supplementary Methods), a three-step method tested their association with two independent, GWAS meta-analyses on externalizing

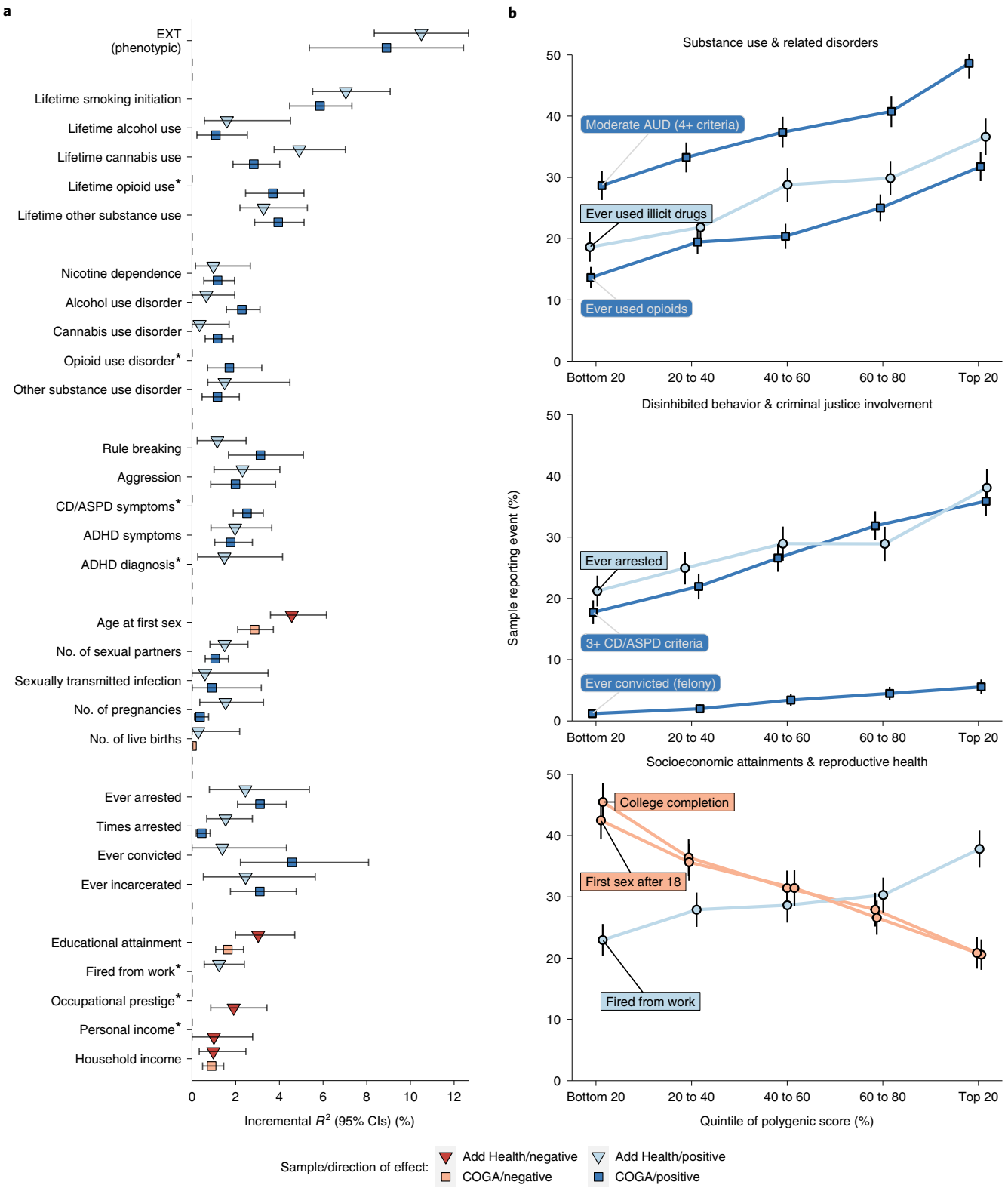


Fig. 3 | Genome-wide EXT polygenic score associations with behavioral, psychiatric and social outcomes in the independent Add Health and COGA datasets. **a**, Scatterplots illustrating the incremental proportion of variance (incremental R^2 or ΔR^2) explained by the genome-wide PRS-CS polygenic score. Light and dark hue indicates the Add Health (N=5,107) and COGA (N=7,594) cohorts, respectively. Blue and red bars indicate positive and negative associations, respectively. The error bars represent 95% CIs centered on ΔR^2, computed as 1.96 times the s.e. (estimated using percentile method bootstrapping over 1,000 bootstrap samples). Asterisks indicate phenotypes that were available only in one of the holdout samples. **b**, Line charts illustrating the relative risks across quintiles of the polygenic score for eight (binary or dichotomized) illustrative outcomes: (1) meeting four or more criteria for AUD, (2) lifetime use of an illicit substance other than cannabis, (3) lifetime opioid use, (4) ever being arrested, (5) meeting three 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. The error bars represent 95% CIs centered on the per-quintile prevalence, computed as 1.96 times the analytical s.e.

phenotypes: (1) AUD (r_g with EXT = 0.52; N = 202,004) and (2) antisocial behavior (r_g with EXT = 0.69; N = 32,574). We had preregistered to hold out antisocial behavior from the externalizing GWAS to enable quasi-replication with a central externalizing trait that was not included in the model. First, we tested whether the 579 SNPs (or an LD proxy for missing SNPs, $r^2 > 0.8$) showed sign concordance, that is, the same direction of effect between EXT and AUD or antisocial behavior: 75.4% of SNPs showed sign concordance with AUD (two-sided test $P = 6.84 \times 10^{-36}$) and 66.9% with antisocial behavior (two-sided test $P = 1.39 \times 10^{-15}$; 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 for 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% versus 6.6%) with AUD (two-sided $P = 1.87 \times 10^{-31}$) and 58 (10.5% versus 4.7%) with antisocial behavior ($P = 1.64 \times 10^{-8}$). In the third test, the 579 SNPs were jointly more strongly enriched for association with AUD (one-sided Mann–Whitney test $P = 5.89 \times 10^{-26}$) and antisocial behavior ($P = 1.10 \times 10^{-5}$) compared to their empirical null distributions. Overall, the three exercises consistently suggested that the GWAS of EXT is not spurious overall, and that it is enriched for genetic signal with two phenotypes of central importance to the literature on externalizing. Below, we perform further quasi-replication of the 579 EXT SNPs in an auxiliary polygenic score analyses (also in within-family models).

Bioinformatic analyses highlight relevant neurobiology. We performed bioinformatic analyses to explore biological processes underlying EXT (Supplementary Methods, Supplementary Tables 9, 10, 13–26 and Extended Data Figs. 4–8). Multi-marker analysis of genomic annotation (MAGMA) gene-property analyses and gene-network analysis with a parsimonious composite network (PCNet) suggested an abundance of enrichment in genes expressed in brain tissues, particularly during prenatal developmental stages (Extended Data Figs. 6 and 8), with the strongest enrichment seen in the cerebellum, followed by the frontal cortex, limbic system tissues and pituitary gland tissues (Extended Data Fig. 5). Furthermore, MAGMA gene-set analysis and PCNet network 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[26], which suggested 587 genes; (2) MAGMA gene-based association analysis[27], which identified 928 Bonferroni-significant genes (one-sided $P < 2.74 \times 10^{-6}$); (3) H-MAGMA[28], a method that assigns noncoding SNPs to cognate genes based on chromatin interactions in adult brain tissue, identifying 2,033 Bonferroni-significant genes (one-sided $P < 9.84 \times 10^{-7}$); and (4) S-PrediXcan[29], which uses transcriptome-based analyses of predicted gene expression in 13 brain tissues and which identified 348 Bonferroni-significant gene–tissue pairs (two-sided $P < 2.73 \times 10^{-7}$).

We found 34 genes that were consistently identified by all four methods, while 741 overlapped across two or more methods (Supplementary Table 22 and Extended Data Fig. 7). Several of the 34 implicated genes are new discoveries for the psychiatric/behavioral literature and have previously been identified only in relation to nonpsychiatric biomedical diseases. Such discoveries include *ALMS1* (previously associated with kidney function and urinary metabolites[30]) and *ERAP2* (blood protein levels and autoimmune disease[31,32]). 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[17,33]), Zic family member 4 (*ZIC4*; associated with brain volume[34]), 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[35,36] and candidate gene for psychiatric disorders[37,38]), neuronal growth regulator 1 (*NEGR1*; associated with intelligence and educational attainment[39,40]) and paired basic amino acid cleaving enzyme (*FURIN*; associated with schizophrenia, risk tolerance and vulnerability to psychiatric disorders[19,41]).

Polygenic score analyses. We created genome-wide polygenic scores for EXT with ~1 million SNPs, adjusted for LD with PRS-CS[42] (Supplementary Methods), among individuals from two hold-out samples selected for their detailed phenotypes related to externalizing and substance use (Supplementary Methods): (1) the National Longitudinal Study of Adolescent to Adult Health (Add Health; N = 5,107), a US-based study of adolescents recruited from secondary schools in the mid-1990s; and (2) the Collaborative Study on the Genetics of Alcoholism (COGA; N = 7,594), a US-based study on genetic contributions to AUDs.

To investigate the validity of EXT, in each of these two samples, we generated a phenotypic EXT by fitting a factor model to phenotypic data corresponding to the seven discovery phenotypes (Extended Data Fig. 9 and Supplementary Table 27). Controlling for age, sex, and ten genetic principal components (PCs), the genome-wide polygenic score was associated with the phenotypic factor in both datasets ($\beta_{\text{Add Health}} = 0.33$, 95% CI: 0.30–0.36, $\Delta R^2 = 10.5\%$; $\beta_{\text{COGA}} = 0.30$, 95% CI: 0.27–0.34, $\Delta R^2 = 8.9\%$; Fig. 3a and Supplementary Table 28). The variance explained by the EXT polygenic score ($\Delta R^2 \sim 8.9$–10.5%) is commensurate with many conventional variables used in social science research, including parental socioeconomic status (SES), family income or structure and neighborhood disadvantage/disorder[43–45]. Next, as further quasi-replication, we created a polygenic score using only the 579 EXT SNPs (this score was only used for this quasi-replication exercise), and also this polygenic score was found to be associated with the phenotypic EXT, explaining ~3–4% of the variance ($\beta_{\text{Add Health}} = 0.20$, 95% CI: 0.17–0.23, $\Delta R^2 = 4.1\%$; $\beta_{\text{COGA}} = 0.17$, 95% CI: 0.13–0.20, $\Delta R^2 = 3.0\%$).

In Add Health, COGA and the Philadelphia Neurodevelopmental Cohort (PNC), we next explored to what extent genome-wide polygenic scores for EXT were associated with childhood externalizing disorders and a variety of phenotypes that reflect difficulty with self-regulation or its consequences (Fig. 3b and Supplementary Tables 29–31; see the tables for standard errors (s.e.) per hold-out sample). Polygenic scores for EXT explained significant variance (ΔR^2) in criteria counts of ADHD (mean $\Delta R^2 = 1.65\%$), conduct disorder (mean $\Delta R^2 = 3.1\%$) and oppositional defiant disorder ($\Delta R^2 = 1.96\%$), as well as in the categories substance use initiation (mean $\Delta R^2 = 1.3$–6.5%), substance use disorders (mean $\Delta R^2 = 0.8$–1.7%), disinhibited behaviors (mean $\Delta R^2 = 1.5$–2.5%), criminal justice system involvement (mean $\Delta R^2 = 1.0$–3.0%), reproductive health (mean $\Delta R^2 = 0.3$–3.7%) and socioeconomic attainment (mean $\Delta R^2 = 0.1$–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. The associations between the EXT polygenic score and this broad range of phenotypes represent an affirmative test of the hypothesis that genetic variants associated with externalizing liability generalize to a variety of behavioral and social outcomes related to self-regulation.

Phenome-wide association study with externalizing polygenic score. To evaluate medical outcomes associated with EXT, we conducted a phenome-wide association study (PheWAS) in 66,915

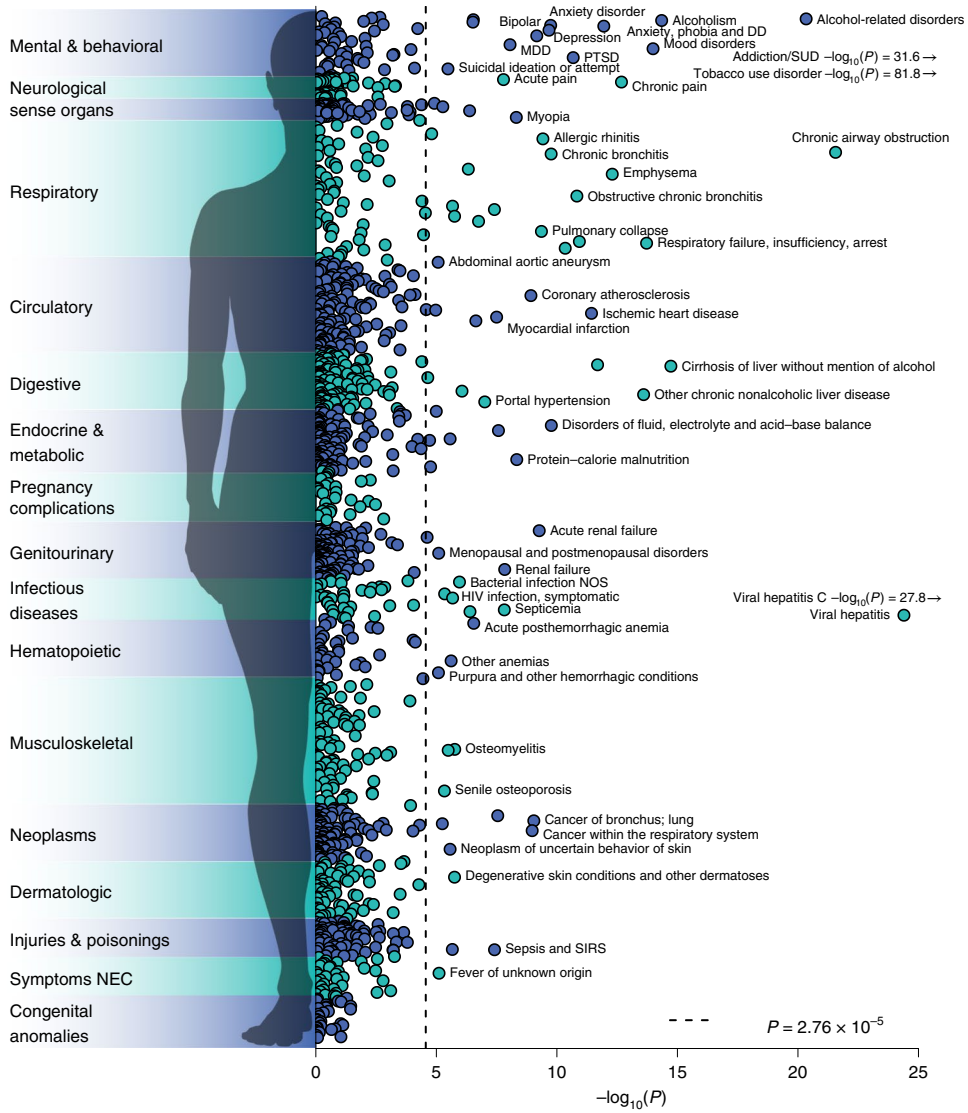


Fig. 4 | Phenome-wide association study in the BioVU biorepository. $-\log_{10} P$ values for two-sided *z*-test of the log of the odds ratio for the genome-wide PRS-CS polygenic score for EXT with 1,335 medical outcomes, estimated with logistic regression in up to 66,915 participants, adjusted for sex, median age in the EHR data and the first ten 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 complete results, including case–control counts, effect sizes and s.e., are reported in Supplementary Table 20. DD, dissociative disorders; MDD, major depressive disorder; NEC, not elsewhere classified; NOS, not otherwise specified; PTSD, post-traumatic stress disorder; SIRS, systemic inflammatory response syndrome; SUD, substance use disorder.

genotyped individuals of European ancestry in the BioVU biorepository, a US-based biobank of electronic health records (EHRs) from the Vanderbilt University Medical Center[46]. A logistic regression was fit to 1,335 case–control disease phenotypes. In total, 255 disease phenotypes were associated with the EXT polygenic score at false discovery rate < 0.05, with odds ratios ranging from 0.8 to 1.4 per standard deviation increase in the score (Fig. 4 and Supplementary Table 32). 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 outcomes 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 the liver, sepsis and lung cancer. Behaviors related to self-regulation, for example, smoking, drinking, drug use, condomless sex and overeating, contribute to many of these medical outcomes.

Within-family analyses demonstrate robustness to confounding. Genetic associations detected in GWAS can be due to direct genetic effects, but can also be confounded by population stratification, indirect genetic effects from, for example, parental environment and assortative mating[47,48]. While reducing statistical power, sibling comparisons overcome these methodological challenges, because meiosis randomizes genotypes to siblings[47,49]. 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), and a sibling sample from the UK Biobank (UKB; $N = 39{,}640$), which were held out from the discovery stage (Supplementary Methods).

In Add Health and COGA, the phenotypic EXT corresponding to the seven discovery phenotypes (see above) was regressed on the genome-wide EXT polygenic scores in a within-family model (Supplementary Table 33). Parameter estimates from the within-family model ($\beta_{\text{WF Add Health}} = 0.12$, 95% CI: 0.04–0.20;

$\beta_{WFCOGA} = 0.14$, 95% CI: 0.08–0.20) were smaller compared to ordinary least-squares models without family-specific intercepts ($\beta_{Add\ Health} = 0.20$, 95% CI, 0.16–0.24; $\beta_{COGA} = 0.16$, 95% CI, 0.12–0.20), but remained statistically significant (Add Health two-sided $P = 4.89 \times 10^{-3}$; COGA two-sided test $P = 1.87 \times 10^{-6}$). As a formal test of attenuation, we evaluated the standardized difference between β_{WF} and β (that is, a z-statistic assumed to be normally distributed; Supplementary Methods) and found that it was −1.988 (two-sided $P = 0.047$) and −0.704 (two-sided $P = 0.481$) for the PRS-CS polygenic score in Add Health and COGA, respectively. Thus, we conclude that there was some, but not extreme, attenuation when predicting the phenotypic EXT within families. Also, the association of the quasi-replication polygenic score constructed with the 579 EXT SNPs remained significant and basically did not attenuate in within-family models (for this score, the standardized difference between β_{WF} and β was −0.338 (two-sided $P = 0.735$) and 0.07 (two-sided $P = 0.944$) in Add Health and COGA, respectively).

In the UKB sibling hold-out sample, we conducted analyses of the genome-wide EXT polygenic scores with 37 phenotypes from the domains of (a) risky behavior, (b) overall and reproductive health, (c) cognitive ability, (d) personality and (e) SES (Supplementary Methods and Supplementary Table 34). We evaluated the per-category mean of the standardized difference between β_{WF} and β, and found that within-family estimates were, on average, the same for the risky behavior category (mean attenuation = 0.08; 95% CI: −1.67 to 1.83), and only attenuated modestly for personality (mean attenuation = −0.35; 95% CI: −1.06 to 0.36). However, the within-family estimates attenuated more for cognitive ability (mean attenuation −6.55; 95% CI: −9.93 to −3.17), SES (mean attenuation −2.43; 95% CI: −4.39 to −0.48), and overall and reproductive health (mean attenuation −2.20; 95% CI: −4.18 to −0.21). Nonetheless, the EXT polygenic score remained nominally significant (two-sided $P < 0.05$) with 24 outcomes across the five categories, showing that the externalizing GWAS captures genetic effects that are 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 an understanding of the molecular genetic underpinnings of externalizing has lagged behind progress made in other areas of medical and psychiatric genetics. For example, dozens of genetic loci have been discovered for schizophrenia (>100 loci)[50], bipolar disorder (30 loci)[51] and major depressive disorders (44 loci)[52], whereas for antisocial behavior[53], AUDs[54] and opioid use disorders[8], only a very small number of loci have been discovered. We used multivariate genomic analyses to accelerate genetic discovery, identifying 579 genome-wide significant loci associated with a liability toward externalizing outcomes, 121 of which are entirely new discoveries for any of the seven phenotypes analyzed. Follow-up bioinformatic analyses suggest the implicated genes have early neurodevelopmental effects, which are then associated with behavioral patterns that have repercussions across the lifespan.

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[55,56]; new analytic methods now allow us to capitalize on these genetic correlations. Pragmatically, non-disease phenotypes such as the ones we use here (for example, 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 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 a phenotypic EXT. These effect sizes rival the associations observed with 'traditional' covariates used in social science research.

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. These 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[57]. Linking biology with socially valued behavioral outcomes can be politically sensitive[58]. 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[59]. At the same time, failing to understand how individual genetic differences contribute to vulnerability to externalizing can increase stigma and blame for these behaviors[60]. 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 individuals of European ancestry) must be interpreted with great care. Our 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[61]. Please see our frequently asked questions and supporting materials at https://externalizing.org/.

In conclusion, our analyses demonstrate the far-reaching toll of human suffering borne by people with high genetic liabilities to externalizing. Future work will be needed to tease apart the pathways by which biological and social risks unfold within and across generations, and our findings can contribute to that effort.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-021-00908-3.

Received: 16 October 2020; Accepted: 13 July 2021;
Published online: 26 August 2021

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.* **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. 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).
7. Swann, A. C., Lijffijt, M., O'Brien, B. & Mathew, S. J. Impulsivity and suicidal behavior. *Curr. Top. Behav. Neurosci.* **47**, 179–195 (2020).

8. Zhou, H. et al. Association of OPRM1 functional coding variant with opioid use disorder: a genome-wide association study. *JAMA Psychiatry* https://doi.org/10.1001/jamapsychiatry.2020.1206 (2020).
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.* **3**, 513–525 (2019).
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. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
13. Demontis, D. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
14. Walters, R. K. et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656–1669 (2018).
15. Sanchez-Roige, S. et al. Genome-wide association study meta-analysis of the alcohol use disorders identification test in two population-based cohorts. *Am. J. Psychiatry* **176**, 107–118 (2018).
16. 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).
17. 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).
18. 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).
19. Lee, P. H. et al. Genomic relationships, novel loci and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469–1482 (2019).
20. Lo, M.-T. et al. Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nat. Genet.* **49**, 152–156 (2016).
21. Rosenström, T. et al. Joint factorial structure of psychopathology and personality. *Psychol. Med.* **49**, 2158–2167 (2019).
22. Townsend, P. *Health and Deprivation: Inequality and the North* (Croom Helm, 1988).
23. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
24. de la Fuente, J., Davies, G., Grotzinger, A. D., Tucker-Drob, E. M. & Deary, I. J. A general dimension of genetic sharing across diverse cognitive traits inferred from molecular data. *Nat. Hum. Behav.* **5**, 49–58 (2021).
25. Hart, A. B. & Kranzler, H. R. Alcohol dependence genetics: lessons learned from genome-wide association studies (GWAS) and post-GWAS analyses. *Alcohol. Clin. Exp. Res.* **39**, 1312–1327 (2015).
26. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
27. 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).
28. 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).
29. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
30. Jaykumar, A. B. et al. Role of Alström syndrome 1 in the regulation of blood pressure and renal function. *JCI Insight* **3**, e95076 (2018).
31. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
32. Li, Y. R. et al. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat. Med.* **21**, 1018–1027 (2015).
33. 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).
34. 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).
35. Edenberg, H. J. et al. Variations in *GABRA2*, encoding the α2 subunit of the $GABA_A$ receptor, are associated with alcohol dependence and with brain oscillations. *Am. J. Hum. Genet.* **74**, 705–714 (2004).
36. 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).
37. 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).
38. Brambilla, P., Perez, J., Barale, F., Schettini, G. & Soares, J. C. GABAergic dysfunction in mood disorders. *Mol. Psychiatry* **8**, 721–737 (2003).
39. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
40. 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).
41. Schrode, N. et al. Synergistic effects of common schizophrenia risk variants. *Nat. Genet.* **51**, 1475–1485 (2019).
42. 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).
43. Derzon, J. H. The correspondence of family features with problem, aggressive, criminal and violent behavior: a meta-analysis. *J. Exp. Criminol.* https://doi.org/10.1007/s11292-010-9098-0 (2010).
44. 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. *Soc. Sci. Med.* https://doi.org/10.1016/j.socscimed.2018.11.015 (2019).
45. 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.* https://doi.org/10.1002/ab.21641 (2016).
46. Davis, L. Psychiatric genomics, phenomics and ethics research in a 270,000-person Biobank (BioVU). *Eur. Neuropsychopharmacol.* **29**, S739–S740 (2019).
47. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype–phenotype associations in humans. *Science* **365**, 1396–1400 (2019).
48. Kong, A. et al. The nature of nurture: effects of parental genotypes. *Science* **359**, 424–428 (2018).
49. Selzam, S. et al. Comparing within- and between-family polygenic score prediction. *Am. J. Hum. Genet.* **105**, 351–363 (2019).
50. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
51. Stahl, E. A. et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).
52. 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).
53. Tielbeek, J. J. et al. Genome-wide association studies of a broad spectrum of antisocial behavior. *JAMA Psychiatry* **74**, 1242–1250 (2017).
54. 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).
55. Bulik-Sullivan, B. K. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
56. Anttila, V. et al. Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
57. 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).
58. Fox, D. Subversive science. *Penn State Law Rev.* **124**, 153–191 (2019).
59. American Society of Human Genetics. ASHG denounces attempts to link genetics and racial supremacy. *Am. J. Hum. Genet.* **103**, 636 (2018).
60. 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).
61. 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).
62. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

COGA Collaborators

Bernice Porjesz[29], Victor Hesselbrock[30], Tatiana M. Foroud[31], Arpana Agrawal[32], Howard J. Edenberg[31], John I. Nurnberger Jr[31], Yunlong Liu[31], Samuel Kuperman[33], John Kramer[33], Jacquelyn L. Meyer[29], Chella Kamarajan[29], Ashwini K. Pandey[29], Laura Bierut[32], John Rice[32], Kathleen K. Bucholz[32], Marc A. Schuckit[34], Jay Tischfield[35], Andrew Brooks[35], Ronald P. Hart[35], Laura Almasy[36], Danielle M. Dick[3,6,42], Jessica E. Salvatore[35], Allison Goate[37], Manav Kapoor[37], Paul Slesinger[37], Denise M. Scott[38], Lance Bauer[30], Leah Wetherill[31], Xiaoling Xuei[31], Dongbing Lai[31], Sean J. O'Connor[31], Martin H. Plawecki[31], Spencer Lourens[31], Laura Acion[33], Grace Chan[30,33], David B. Chorlian[29], Jian Zhang[29], Sivan Kinreich[29], Gayathri Pandey[29], Michael J. Chao[37], Andrey P. Anokhin[32], Vivia V. McCutcheon[32], Scott Saccone[32], Fazil Aliev[39], Peter B. Barr[3,41], Hemin Chin[40] and Abbas Parsian[40]

[29]SUNY Downstate, Brooklyn, NY, USA. [30]University of Connecticut, Farmington, CT, USA. [31]Indiana University, Indianapolis, IN, USA. [32]Washington University in St. Louis, St. Louis, MO, USA. [33]University of Iowa, Iowa City, IA, USA. [34]University of California San Diego, San Diego, CA, USA. [35]Rutgers University, Newark, NJ, USA. [36]Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA. [37]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [38]Howard University, Washington, DC, USA. [39]Virginia Commonwealth University, Richmond, VA, USA. [40]National Institute on Alcohol Abuse and Alcoholism, Bethesda, MD, USA. A full list of investigators and their affiliations appears in the Supplementary Information.

Methods

The article is accompanied by Supplementary Information. The study followed a preregistered analysis plan (https://doi.org/10.17605/OSF.IO/XKV36), which specified that we would generate new, or collect existing, single-phenotype GWAS summary statistics on externalizing phenotypes (Supplementary Methods). Summary statistics were to be analyzed with genomic SEM to (a) estimate a genetic factor structure underlying externalizing liability, (b) identify SNPs and genes involved in a shared genetic liability to externalizing rather than individual traits, and (c) increase the accuracy of polygenic scores for specific externalizing phenotypes that are difficult or intractable to study in large samples. To ensure satisfying statistical power, we preregistered a minimum sample size of $N > 15{,}000$, and that additional exclusions would be based on negligible SNP-based heritability or GWAS signal. All considered traits are discussed in the preregistration and listed in Supplementary Table 1, while the following sections focus on 11 phenotypes that were not excluded due to negligible SNP-based heritability or GWAS signal. The study did not manipulate an experimental condition or collect any new individual-level data, and thus, was neither randomized nor blinded.

Collecting single-phenotype GWAS on externalizing phenotypes. A detailed definition of 'externalizing phenotypes' was preregistered to delimit the collection of single-phenotype summary statistics (Supplementary Methods). Summary statistics from existing studies were provided by, or downloaded from, the public repositories of 23andMe, the Psychiatric Genomics Consortium (PGC), the Million Veterans Program, the International Cannabis Consortium, the GWAS & Sequencing Consortium of Alcohol and Nicotine Use, the Social Science Genetics Association Consortium, the Genetics of Personality Consortium and the Broad Antisocial Behavior Consortium (Supplementary Methods). All considered GWAS are listed in Supplementary Table 1, and Supplementary Table 4 reports the 67 underlying cohorts of the summary statistics in the final Genomic SEM specification (see below).

GWAS in the UK Biobank. For the Genomic SEM analyses, we conducted a total of ten GWAS in the UKB (Supplementary Table 1). These GWAS were conducted for two reasons: (1) to generate summary statistics for phenotypes that had not yet been studied in the full genetic data release, or (2) to generate hold-out summary statistics that excluded participants for follow-up analyses. The hold-out summary statistics were used to replace, in our genomic SEM analyses, summary statistics from existing studies that had included UKB data. With respect to (1), summary statistics for 'age at first sexual intercourse' and 'AUDIT-P' were later included in the final genomic SEM specification (the latter as a meta-analysis with a GWAS on alcohol dependence by the PGC). With respect to (2), the final specification included replacement summary statistics on 'lifetime cannabis use', 'general risk tolerance' and 'lifetime smoking initiation' and 'number of sexual partners'. The seventh phenotype in the final specification—ADHD by the PGC—did not include analyses from UKB. For a detailed description, see Supplementary Methods.

The GWAS in the UKB were conducted with linear mixed models (BOLT-LMM version 2.3.2) and were adjusted for sex, birth year, sex-specific birth-year dummies, genotyping array and batch and 40 genetic PCs estimated with FlashPCA2 (version 2.0). Two partly overlapping hold-out sub-samples of UKB participants were excluded from all single-phenotype GWAS summary statistics that included UKB data, and the participants were instead retained as a hold-out sample for polygenic score analyses (Supplementary Methods). Genetic relatives (pairwise KING coefficient ≥ 0.0442, version 2.1.5) of the held-out individuals were excluded from the study altogether to ensure independence between the discovery and follow-up analyses. In summary, whenever an existing GWAS (or GWAS meta-analysis) was based on UKB, we re-conducted it using the same phenotype definition to generate summary statistics that excluded the hold-out sample and their genetic relatives.

GWAS inclusion criteria, quality control and meta-analysis. All GWAS were conducted among individuals that (a) were of European ancestry, (b) were not missing any relevant covariates, (c) were successfully genotyped and passed standardized sample-level quality control (according to study-specific protocols[13–16,18]) and (d) were unrelated (unless a particular GWAS was conducted with linear mixed models). Genotypes were imputed with reference data from either the 1000 Genomes Consortium[63], the Haplotype Reference Consortium[64], the UK10K Consortium[65] or a combination thereof. We performed quality control with EasyQC (version 9.1)[66]. For that purpose, we used a whole-genome sequenced reference panel, assembled from 1000 Genomes Consortium[63] and UK10K Consortium[65] data by using BCFtools (version 1.8; Supplementary Methods). Our quality-control procedure applied recommended[66] SNP filtering to remove (1) rare SNPs (minor allele frequency < 0.005), (2) SNPs with an IMPUTE imputation quality (INFO) score < 0.9, (3) SNPs that could not be mapped to or had discrepant alleles with the reference panel and (4) otherwise low-quality variants (Supplementary Table 2). For a complete description of the quality-control procedures, see Supplementary Methods.

We performed sample-size weighted meta-analysis with METAL (versions 2011-03-25 and 2020-05-05)[67], while ensuring absence of sample overlap (Supplementary Methods). We excluded any summary statistics with negligible SNP-based heritability ($h^2 < 0.05$) or GWAS signal ($\chi^2 < 1.05$), estimated with LD Score regression (version 1.0.0)[12,55]. At this stage, we had collected or generated well-powered summary statistics for 11 phenotype-specific GWAS (or meta-analysis) that satisfied our inclusion criteria (Supplementary Table 3): (1) ADHD ($N = 53{,}293$), (2) FSEX ($N = 357{,}187$), (3) ALCP ($N = 164{,}684$), (4) automobile speeding propensity (DRIV, $N = 367{,}151$), (5) alcoholic drinks per week (DRIN, $N = 375{,}768$), (6) reverse-coded educational attainment (EDUC, $N = 725{,}186$), (7) CANN ($N = 186{,}875$), (9) SMOK ($N = 1{,}251{,}809$), (9) RISK ($N = 426{,}379$), (10) irritability (IRRT, $N = 388{,}248$) and (11) NSEX ($N = 336{,}121$; Supplementary Table 4). The GWAS effect sizes of age at first sexual intercourse and educational attainment were reversed to anticipate positive correlations with externalizing liability.

Exploratory factor analysis. As an initial analysis to guide the multivariate analyses, we performed hierarchical clustering of a matrix with pairwise LD Score (version 1.0.0) genetic correlations (Supplementary Methods). The 11 phenotypes displayed appreciable genetic overlap with at least one other phenotype ($\max|r_g| = 0.245$–0.773; Supplementary Table 5). Three (k) clusters were identified: (1) ADHD, EDUC, FSEX, IRRT and SMOK; (2) ALCP and DRIN; and (3) CANN, DRIV, NSEX and RISK.

Exploratory factor analysis tested four factor solutions, specifying 1 to $k+1$ factors with the 'factanal' function of R ('stats' package version 3.5.1; Supplementary Methods), where k is the number of clusters identified in the genetic correlation matrix, while retaining factors that explained $\geq$15% variance (preregistered threshold). The fourth factor explained only 12.5% variance, and thus, the three-factor solution was considered the best-fitting exploratory model (Supplementary Table 6). The factor loadings were consistent with the hierarchical clustering. However, as detailed in Supplementary Methods, the second and third factor 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 exploratory analysis that some of the 11 indicators may not be optimal for identifying a common genetic liability to externalizing, and that a less complex model with fewer indicators may perform better in subsequent confirmatory analyses.

Confirmatory factor analyses with Genomic SEM. We formally modeled genetic covariances (rather than r_g) in confirmatory factor analyses using genomic SEM, versions 0.0.2a-c[10] (Supplementary Methods). Genomic SEM is unbiased by sample overlap and imbalanced sample size, and by applying to summary statistics allows for genetic analyses of latent factors with more observations than is typically possible with individual-level data[10]. We estimated and benchmarked four models: (1) a common factor model with the 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 finally, (4) a revised common factor model that only included seven of the phenotypes that satisfied moderate-to-large (that is, $\geq$0.50) loadings on the single latent factor in model 1 (Supplementary Table 7). We found that model 4 was the only model that closely approximated the observed genetic covariance matrix ($\chi^2(12 = 390.234$, AIC = 422.234, CFI = 0.957, SRMR = 0.079), that fulfilled our preregistered model fit criteria, and that coalesced with theoretical expectations of a common genetic liability to externalizing. This model was selected as final specification, and is referred to as EXT. To explore the convergent and discriminant validity of EXT, we estimated its genetic correlation with 91 traits from various domains (Supplementary Table 8).

Multivariate GWAS analyses with genomic SEM. Using genomic SEM, we performed multivariate genome-wide association analysis by estimating SNP associations with EXT, which is our main discovery analysis (Supplementary Methods). The estimated effective sample size of the 'externalizing GWAS' is $N_{eff} = 1{,}492{,}085$, and the mean χ^2 and λ_{GC} are 3.114 and 2.337, respectively. LD Score regression suggested polygenicity rather than bias from population stratification[10,12], as the (default settings) LD Score intercept and attenuation ratio were estimated to be 1.115 (s.e. = 0.019) and 0.054 (s.e. = 0.009), respectively.

Conventional 'clumping' was applied with PLINK (v1.90b6.13)[68] to identify near-independent genome-wide significant lead SNPs, with the primary (two-sided) P-value threshold of 5×10^{-8}, the secondary P-value threshold (for computational efficiency) of 1×10^{-4} and an r^2 threshold of 0.1 together with a wide SNP window (1,000,000 kb). Before counting the number of hits to report, we first subjected the 855 lead SNPs to 'multi-SNP-based conditional & joint association analysis using GWAS summary data' (GCTA-COJO, version 1.93.1beta[23,69]; Supplementary Methods). This procedure identified 579 lead SNPs that were COJO with EXT (Supplementary Table 9). We performed lookups of these '579 EXT SNPs', and any correlated SNPs ($r^2 > 0.1$) in the NHGRI-EBI GWAS Catalog[6] (e96 2019-05-03) to investigate whether the loci have previously been reported with other traits (at two-sided $P < 1 \times 10^{-5}$; Supplementary Table 10). To evaluate whether each SNP acted through the EXT, we estimated Q_{SNP} heterogeneity statistics genome wide with genomic SEM (Supplementary Methods). The null hypothesis of the Q_{SNP} test is that SNP effects on the constituent phenotypes operate (that is, are statistically mediated) via the EXT factor, so a significant Q_{SNP} test indicates that the SNP effects are better explained by pathways independent of

EXT. The Q_{SNP} analysis identified substantial heterogeneity (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 10). An analysis of sign concordance further supported homogeneity among the EXT SNPs (Supplementary Methods).

Proxy-phenotype and quasi-replication analysis. We conducted proxy-phenotype[70] and quasi-replication[71] analyses by investigating the 579 EXT SNPs for association in two independent, second-stage GWAS on (1) AUD ($N=202{,}004$, $r_g=0.52$) and (2) antisocial behavior ($N=32{,}574$, $r_g=0.69$; Supplementary Methods and Supplementary Tables 11 and 12). For SNPs missing from the second-stage GWAS, we analyzed proxy SNPs ($r^2>0.8$). Significant proxy-phenotype associations were evaluated for Bonferroni-corrected significance (two-sided test $P<0.05/579$). For the quasi-replication, we generated empirical null distributions for the second-stage GWAS by randomly selecting 250 near-independent ($r^2<0.1$) SNPs matched on minor allele frequency (± 1 percentage point) for each of the 579 SNPs. The quasi-replication included three steps: (1) a binomial test of sign concordance to test whether the direction of effect of the 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 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, using a nonparametric (one-sided) Mann–Whitney test of the null hypothesis that the P values of the SNPs are derived from the empirical null distribution. We strongly rejected the null hypotheses in 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.

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 hold-out cohorts: (1) Add Health[72,73], (2) COGA[74–76], (3) PNC[77,78], (4) the UKB siblings hold-out cohort[79] and (5) BioVU[46] (Supplementary Methods). In each dataset, we generated three scores, of which two were adjusted for LD: (1) PRS-CS (version 20 October 2019; default Bayesian gamma-gamma prior of 1 and 0.5, and 1,000 Monte Carlo iterations with 500 burn-in iterations)[80], (2) LDpred (version 0.9.09; infinitesimal Bayesian prior)[81] and (3) unadjusted scores[82], while using SNPs that overlapped the HapMap 3 Consortium consensus set[83] (for comparability across methods and with previous work and because PRS-CS imposes that restriction). We evaluated the incremental R^2/pseudo-R^2 (ΔR^2) attained by adding the polygenic score to a regression model with baseline covariates, as in previous efforts[17]. The baseline model included covariates for sex, age and genetic PCs, and genotyping array and batch. The choice of statistical model (for example, least squares versus logit) and adjustment of s.e. depended on (1) the phenotype distribution and (2) the cohort data structure (independent versus clustered observations; Supplementary Methods. We estimated 95% CIs for ΔR^2 using percentile method bootstrapping (1,000 iterations).

In Add Health and COGA, we performed out-of-sample validation of EXT by modeling a latent phenotypic EXT corresponding to the seven Genomic SEM phenotypes (Supplementary Methods and Supplementary Tables 27 and 28). In Add Health, COGA, PNC and the UKB siblings hold-out cohort, we performed exploratory analyses with a range of preregistered phenotypes from various domains (Supplementary Tables 29–31 and 34). In BioVU, we performed a PheWAS of medical outcomes by fitting a logistic regression to 1,335 case–control disease 'phecodes'[84] ($N=66{,}915$; Supplementary Table 32), adjusted for sex, median age in the EHR data and the first ten genetic PCs.

We performed within-family analyses among full siblings in Add Health, COGA and the UKB siblings hold-out cohort (Supplementary Methods). 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 least-squares regression on a single outcome: the factor scores of the phenotypic EXT (a continuous variable) while adjusting for family-specific dummy variables (Supplementary Table 33), and calculated the standardized difference (that is, a z-statistic) between the within-family coefficient ($\hat{\beta}_{WF}$) to the coefficient from a model without family dummies ($\hat{\beta}$ Supplementary Methods). In the UKB siblings hold-out cohort, we performed an analogous analysis of exploratory phenotypes (Supplementary Table 34). We analyzed heteroskedasticity-consistent and cluster-robust s.e., clustered at the family level.

Bioannotation. We conducted bioannotation and bioinformatic analyses (Supplementary Methods). The method FUMA (version 1.3.5e)[26] was applied to explore the functional consequences of the 579 SNPs (Supplementary Table 9), which included ANNOVAR categories (that is, the functional consequence of SNPs on genes), combined annotation dependent depletion scores, RegulomeDB scores, expression quantitative trait loci and chromatin states. The default external reference data for FUMA is described elsewhere[26].

Gene-based analyses was performed with MAGMA (version 1.08)[27] (Supplementary Methods). Genome-wide SNPs were first mapped to 18,235 protein-coding genes from Ensembl (build 85)[85], and SNPs within each gene were jointly tested for association with EXT. We evaluated Bonferroni-corrected significance, adjusted for the number of genes (one-sided $P<2.74\times10^{-6}$; Supplementary Table 13). Next, MAGMA gene-set analysis was performed using 15,481 curated gene sets and Gene Ontology[86] terms obtained from the Molecular Signatures Database (version 7.0)[87]. We evaluated Bonferroni-corrected significance, adjusted for the number of gene sets (one-sided $P<3.23\times10^{-6}$; Supplementary Table 14). 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 15), and between brain gene expression profiles and gene associations across 11 brain tissues from BrainSpan[88] (Supplementary Table 16). Gene expression values were $\log_2$-transformed average reads per kilobase million (RPKM) per tissue type (after replacing RPKM > 50 with 50) based on Genotype-Tissue Expression (GTEx) RNA-sequencing data (version 8.0)[89]. 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' (based on MAGMA version 1.08)[28], to assign noncoding (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[90–92]. We evaluated Bonferroni-corrected significance, adjusted for the number of tests within each of the four Hi-C datasets (one-sided $P<9.83$–9.86×10^{-7}; Supplementary Tables 17–20).

The method S-PrediXcan (version 0.6.2)[93] tested the association of EXT with gene expression in brain tissues. We used precomputed tissue weights from the GTEx database (version 8.0) as the reference dataset[89]. As inputs, 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 $P<2.77\times10^{-7}$, which was Bonferroni corrected for 13 tissues times 13,876 tested genes (180,388 gene–tissue pairs; Supplementary Table 21). In Supplementary Table 22, we summarize the gene findings. Finally, we followed up on the subset of gene findings that were consistently implicated in all gene-based methods, by generating an 'externalizing gene network' as a PCNet and an 'externalizing systems map' with Cytoscape (version 3.8.2)[94,95], and applied tissue-specific expression analysis (version 1.0) and specific expression analysis (version 1.1) to explore tissue and brain region specificity (Supplementary Tables 23–26).

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

Data availability

All data sources are described in the Supplementary Information and are listed in the Reporting Summary. No new data were collected. Only data from existing studies or study cohorts were analyzed, some of which have restricted access to protect the privacy of the study participants (see Reporting Summary for accession codes or URLs). The minimum dataset necessary to interpret, verify and extend the research, that is, the GWAS summary statistics for the EXT GWAS (our main discovery analysis), can be obtained by following the procedures detailed at https://externalizing.org/request-data/. In brief, summary statistics are derived from analyses based in part on 23andMe data, for which we are restricted to only publicly available 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 GWAS summary statistics.

References

63. Yang, J. et al. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
64. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
65. Walter, K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
66. Winkler, T. W. et al. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
67. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
68. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
69. 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).

70. Rietveld, C. A. et al. Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc. Natl Acad. Sci. USA* **111**, 13790–13794 (2014).
71. 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).
72. 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**, 391–398 (2013).
73. 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).
74. Begleiter, H. The Collaborative Study on the Genetics of Alcoholism. *Alcohol Health Res. World* **19**, 228–236 (1995).
75. Edenberg, H. J. The collaborative study on the genetics of alcoholism: an update. *Alcohol Res. Health* **26**, 214–218 (2002).
76. 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.* https://doi.org/10.1111/acer.13293 (2017).
77. Calkins, M. E. et al. The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. *J. Child Psychol. Psychiatry* **56**, 1356–1369 (2016).
78. 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).
79. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
80. 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).
81. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
82. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
83. Altshuler, D. M., Gibbs, R. A. & Peltonen, L. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
84. 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).
85. Hubbard, T. et al. The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
86. Consortium, T. G. O. The Gene Ontology project in 2008. *Nucleic Acids Res.* **36**, D440–D444 (2007).
87. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
88. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
89. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
90. Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
91. Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).
92. Rajarajan, P. et al. Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* **362**, eaat4311 (2018).
93. 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).
94. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
95. Singhal, A. et al. Multiscale community detection in Cytoscape. *PLoS Comput. Biol.* **16**, e1008239 (2020).

Acknowledgements

This research was carried out under the auspices of the Externalizing Consortium. The study was classified as secondary research of de-identified participants, and the study was awarded ethical approval by the internal review board 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 PGC, the Million Veterans Program, the International Cannabis Consortium, the GWAS & Sequencing Consortium of Alcohol and Nicotine use, the Social Science Genetics Association Consortium, the Genetics of Personality Consortium and the Broad Antisocial Behavior Consortium. We 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 also thank the research participants and employees of 23andMe for making this work possible. This research was conducted in part using the UKB resource under applications 40830 and 11425. We thank all UKB cohort participants for making this study possible. We thank L. K. Davis for providing access to BioVU. Finally, we thank COGA; principal investigators B. Porjesz, V. Hesselbrock, H. Edenberg, L. Bierut; and collaborators at eleven different centers: University of Connecticut (V. Hesselbrock); Indiana University (H. J. Edenberg, J. Nurnberger Jr., T. Foroud and Y. Liu); University of Iowa (S. Kuperman and J. Kramer); SUNY Downstate (B. Porjesz); Washington University in St. Louis (L. Bierut, J. Rice, K. Bucholz and A. Agrawal); University of California, San Diego (M. Schuckit); Rutgers University (J. Tischfield and A. Brooks); Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia; Department of Genetics, Perelman School of Medicine, University of Pennsylvania (L. Almasy); Virginia Commonwealth University (D.M.D); 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 and S. Lourens (Indiana University); G. Chan (University of Iowa and University of Connecticut); J. Meyers, D. Chorlian, C. Kamarajan, A. Pandey and J. Zhang (SUNY Downstate); J. C. Wang, M. Kapoor and S. Bertelsen (Icahn School of Medicine at Mount Sinai); A. Anokhin, V. McCutcheon and S. Saccone (Washington University); J. Salvatore, F. Aliev and B. Cho (Virginia Commonwealth University); and M. Kos (University of Texas Rio Grande Valley). A. Parsian and H. Chen are the National Institute on Alcohol Abuse and Alcoholism (NIAAA) staff collaborators. All studies included in the externalizing GWAS are listed in the Supplementary Information. Funding: The Externalizing Consortium has been supported by the NIAAA through an administrative supplement (R01AA015416) and by the National Institute of Drug Abuse (R01DA050721). D.M.D. was supported through funding from the NIAAA (K02AA018755, U10AA008401 and P50AA022527). 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 (NICHD; R01HD092548 and R01HD083613) and the Jacobs Foundation. A.A.P. was supported by the NIAAA (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 no. 27676). Both A.A.P. and S.S.-R. were supported by funds from the California Tobacco-Related Disease Research Program (grant nos. 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 NICHD, 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 (https://addhealth.cpc.unc.edu/). This research used Add Health GWAS data funded by Eunice Kennedy Shriver 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 the National Institutes of Health (NIH) grant U10AA008401 from the NIAAA and the National Institute on Drug Abuse. Data were obtained from Vanderbilt University Medical Center's BioVU, which is supported by numerous sources, including 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, including U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962 and R01HD074711; and additional funding sources listed at https://victr.vumc.org/biovu-funding/. Support for data collection for the PNC, acquired through dbGaP (accession no. phs000607, v3.p2), was provided by grant RC2MH089983 awarded to R. Gur and RC2MH089924 was awarded to H. Hakonarson. Participants 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 available in the Supplementary Information.

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 GWAS, quality control, meta-analysis, genetic correlations and multivariate analyses with genomic SEM, with assistance from A.D.G. R.K.L. led the proxy-phenotype and quasi-replication 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. R.d.V. derived analytical s.e. for the within-family analyses. S.S.-R. led the bioinformatics analyses, and R.K.L, S.B.R. and T.I. 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 the 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 Add Health study cohort. A.D.G., E.M.T.-D. and I.D.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.

Competing interests

H.R.K. is a member of the American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative, which was supported in the last 3 years by AbbVie, Alkermes,

Ethypharm, Indivior, Lilly, Lundbeck, Otsuka, Pfizer, Arbor and Amygdala Neurosciences. H.R.K. and J.G. are named as inventors on PCT patent application no. 15/878,640 entitled 'genotype-guided dosing of opioid agonists,' filed on 24 January 2018. J.G. did paid editorial work for the journal Complex Psychiatry. The authors declare no other competing interests.

Additional information

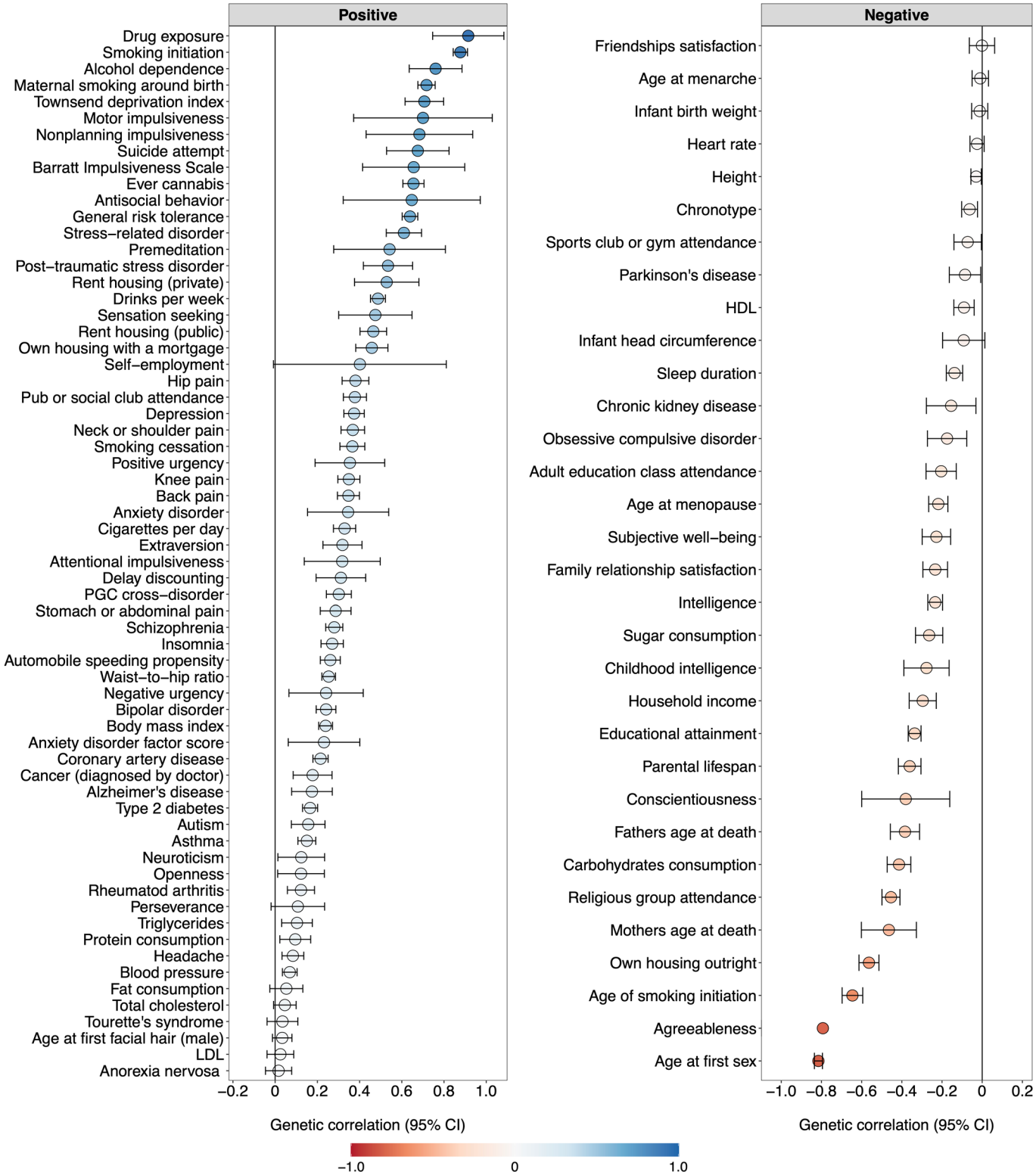
Extended data is available for this paper at https://doi.org/10.1038/s41593-021-00908-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-021-00908-3.

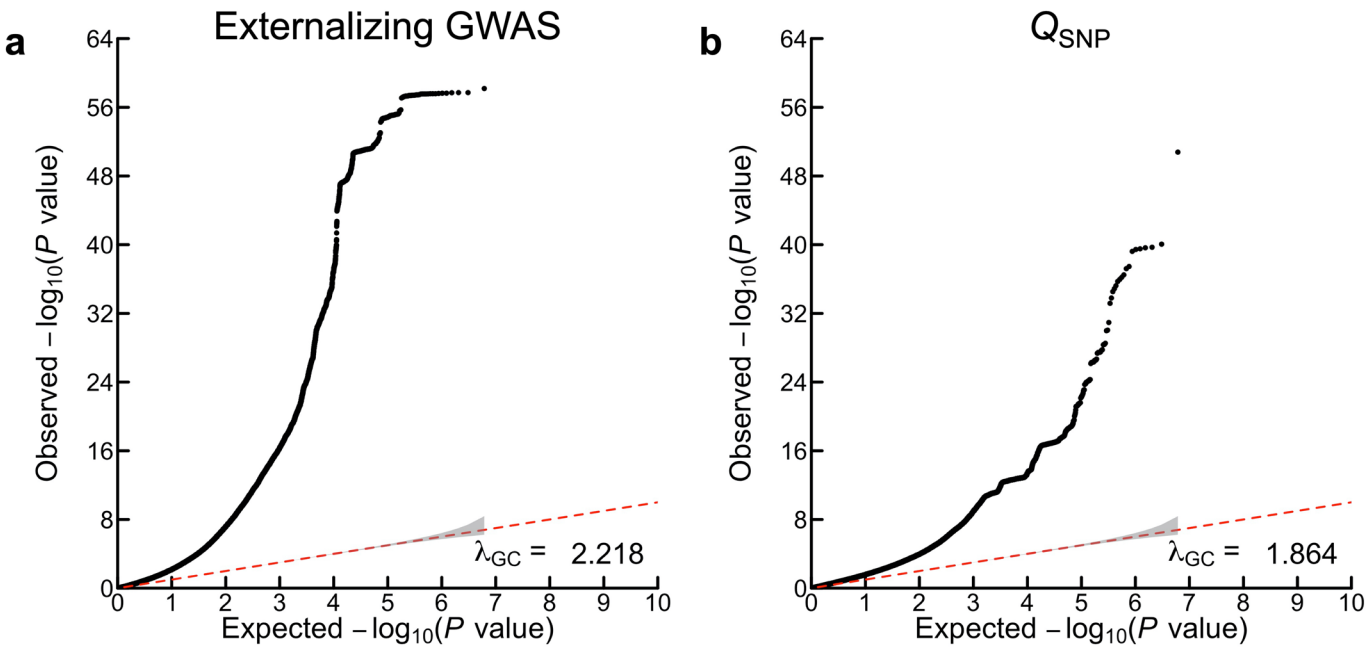
Correspondence and requests for materials should be addressed to P.D.K. or D.M.D.

Peer review information *Nature Neuroscience* thanks Eske Derks and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

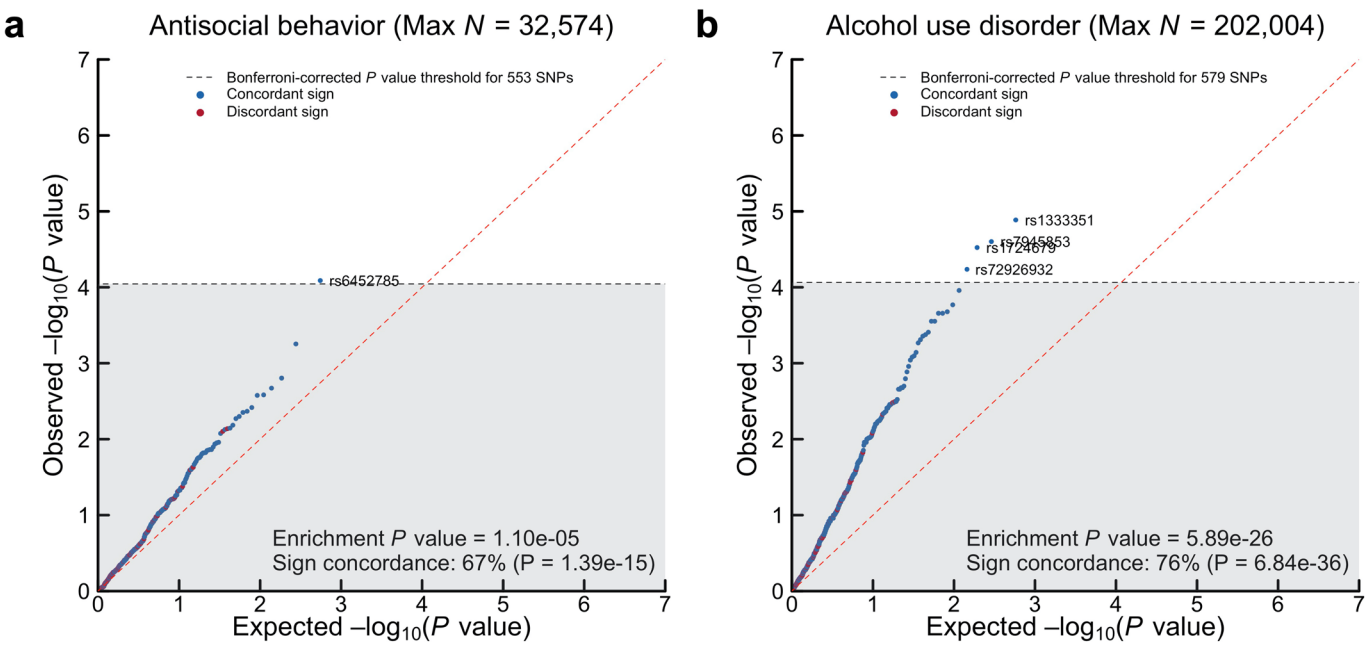
Reprints and permissions information is available at www.nature.com/reprints.



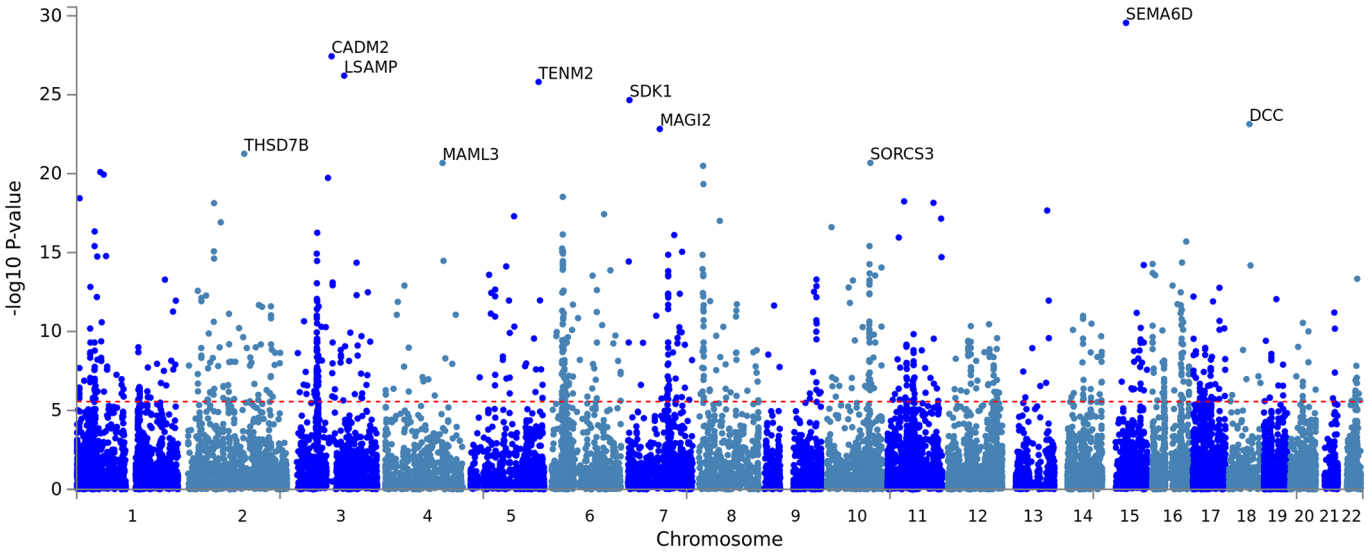
Extended Data Fig. 1 | Genetic correlations with the genetic externalizing factor (EXT). Dot plot of genetic correlations (r_g) estimated with Genomic SEM between the genetic externalizing factor (*EXT*) with 91 other complex traits (Supplementary Methods). Error bars are 95% confidence intervals, calculated as 1.96 × *SE*, centered on the r_g estimate (omitted for Agreeableness). The estimates are also reported in Supplementary Table 8, together with the exact number of independent samples used to derive each estimate. This figure displays genetic correlations with personality measures based on GWAS summary statistics from the Genomics of Personality Consortium, while Fig. 1 instead reports genetic correlations with personality measures based on more recent and substantially larger GWAS provided by 23andMe.



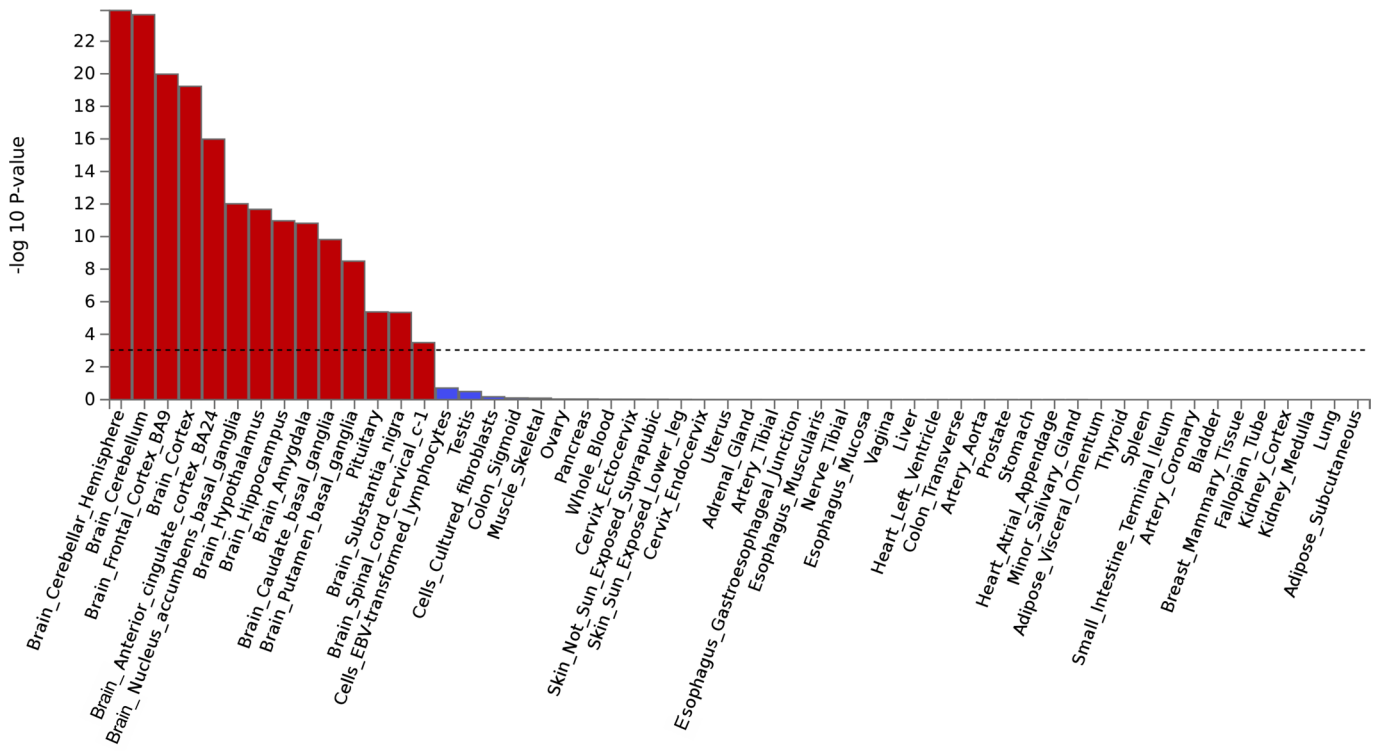
Extended Data Fig. 2 | Quantile-quantile (Q-Q) plots of the externalizing GWAS and QSNP results. The panels display Q-Q plots for (**a**) the externalizing GWAS (N_{eff} = 1,492,085), and (**b**) SNP-level tests of heterogeneity (Q_{SNP}) with respect to the SNP-effects estimated in the externalizing GWAS (for more details see Supplementary Information section 3). The y-axis is the observed association *P* value on the $-\log_{10}$ scale (based on a two-sided Z-test in **a**, and based on a one-sided χ^2 test scaled to 1 degree of freedom in **b**). The gray shaded areas represent 95% confidence intervals centered on the expected $-\log_{10}(P)$ of the null distribution. The genomic inflation factors displayed here, λ_{GC}, is defined as the median χ^2 association test statistic divided by the expected median of the χ^2 distribution with 1 degree of freedom, and were calculated with 6,132,068 and 6,107,583 SNPs for (**a**) and (**b**), respectively. Although there is a noticeable early 'lift-off', the estimated LD Score regression intercepts of (**a**) 1.115 (*SE* = 0.019) and (**b**) 0.9556 (*SE* = 0.013) suggest that most of the inflation of the test statistics is attributable to polygenicity rather than bias from population stratification.



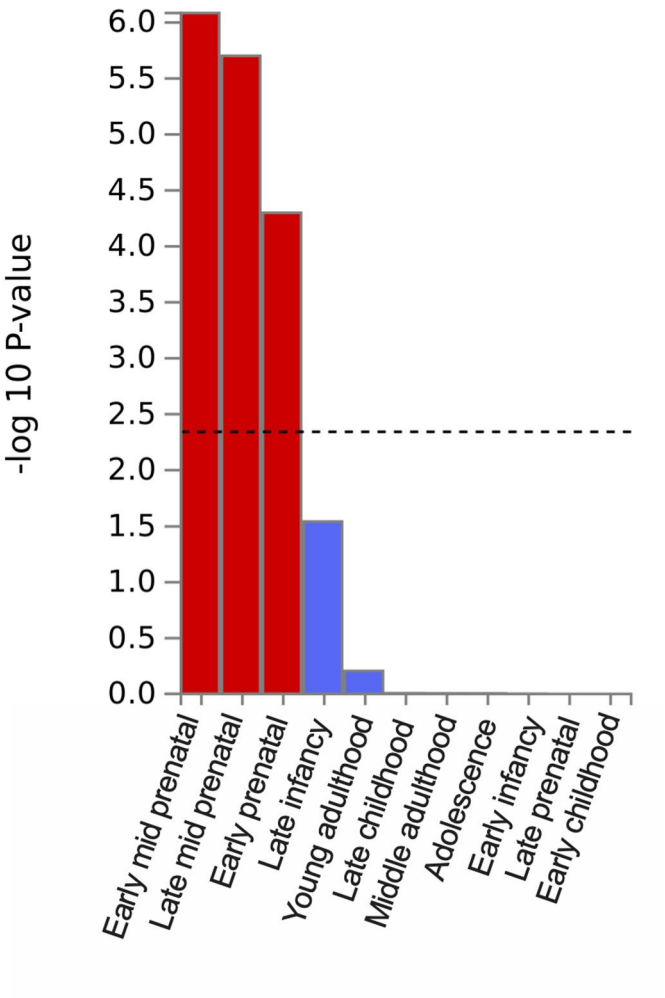
Extended Data Fig. 3 | Quantile-quantile (Q-Q) plots of the proxy-phenotypes analyses. Panels (**a–b**) show -log10(*P* values from a two-sided *Z*-test) for linear regression of the 553 and 579 *EXT* SNPs (or such SNPs that could be proxied in case of missingness, $r^2 > 0.8$) that were looked up in independent, second-stage GWAS samples on (1) antisocial behavior ($N = 32,574$) and (2) alcohol use disorder ($N = 202,400$), respectively (Supplementary Information section 4). Dashed line denotes experiment-wide significance at $P < 0.05/553$ and $0.05/579$ for (1) and (2), respectively. Enrichment *P* value is the result of a one-sided test of joint enrichment with the non-parametric Mann-Whitney test against an empirical null distribution of 138,250 and 144,750 near-independent ($r^2 < 0.1$) SNPs, matched on MAF, that were randomly selected from the GWAS on (1) and (2), respectively. Sign concordance is the proportion of looked-up SNPs with concordant direction of effect sizes across the externalizing GWAS and the second-stage GWAS, and the sign concordance *P* value is from a one-sided binomial tests of the sign concordance for the 579 SNPs (against the null hypothesis of 50% concordance that is expected by chance).



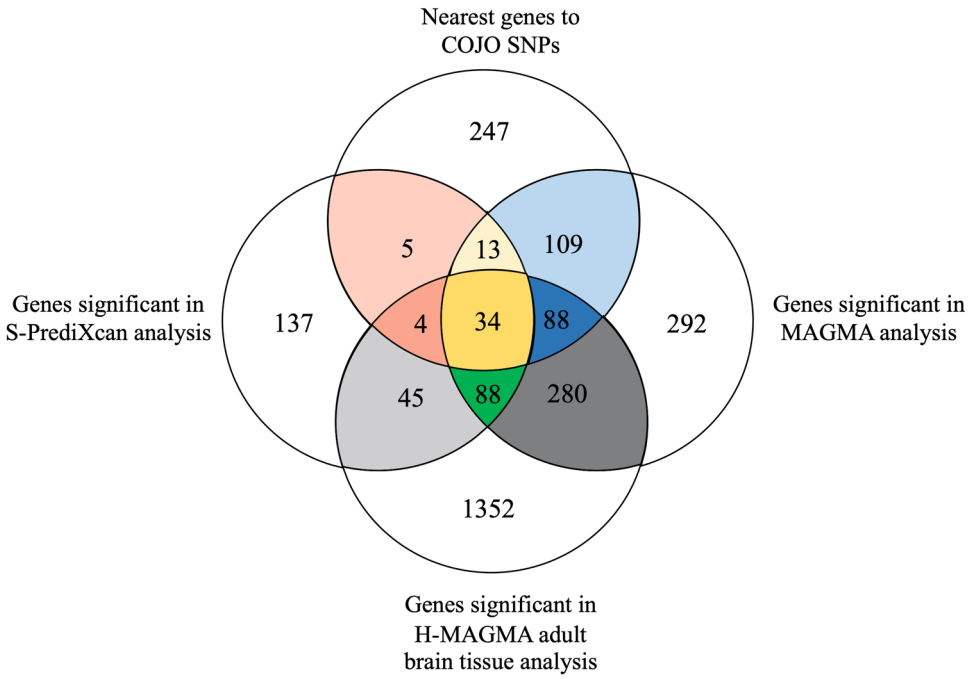
Extended Data Fig. 4 | MAGMA gene-based association analysis. Manhattan plot of the -log10(*P* from a one-sided *Z*-test) of 18,093 genes that were tested for association in the MAGMA (v.1.08) gene-based association analysis (Supplementary Information section 6). The 10 most significant genes are labeled with gene names. Red dashed line represents Bonferroni-significance, adjusted for the number of tested genes (one-sided $P = 2.74 \times 10^{-6}$). 928 genes were found to be significant, of which 244 have one or more genome-wide significant SNPs from the externalizing GWAS within their gene breakpoints. The results are also report in Supplementary Table 13.



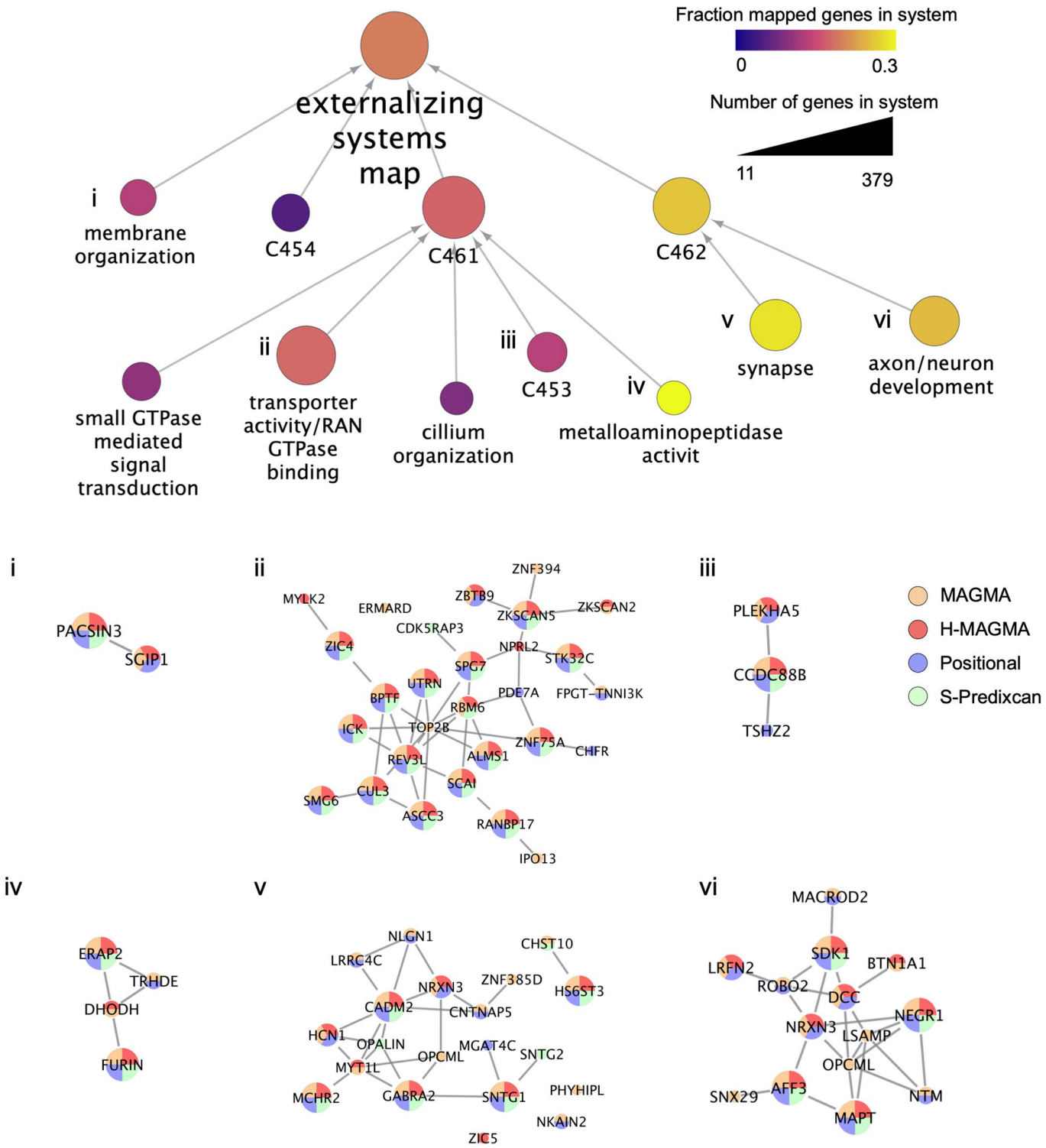
Extended Data Fig. 5 | MAGMA gene-property analysis. Bar plot of the $-\log_{10}$(*P* from one-sided *Z*-tests) of the point estimate from a generalized least squares regression. The analysis identified that the externalizing GWAS is significantly enriched in brain and pituitary gland tissues (Supplementary Information section 6). Dashed line denotes Bonferroni-corrected significance, adjusted for testing 54 tissues (one-sided $P < 9.26 \times 10^{-4}$). 14 tissues were significantly associated with the externalizing GWAS, including 13 brain related tissues and the pituitary tissue. The results are also report in Supplementary Table 15.



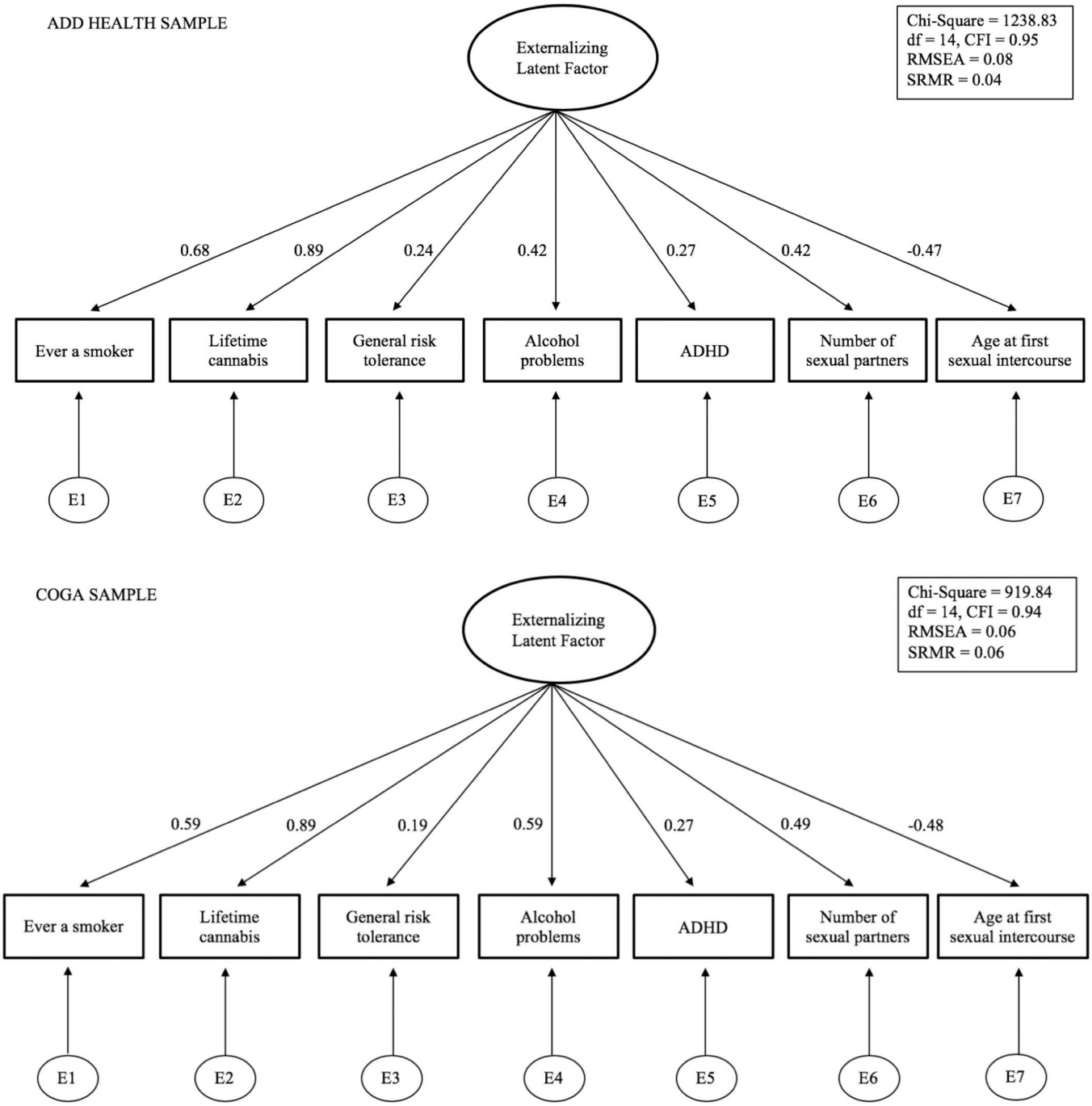
Extended Data Fig. 6 | MAGMA gene-property analysis of enrichment in brain tissues across 11 developmental stages (BrainSpan). Bar plot of the $-\log_{10}(P$ from one-sided Z-tests) of the point estimate from a generalized least squares regression. The analysis identified that the externalizing GWAS is significantly enriched during prenatal developmental stages (Supplementary Information section 6). Dashed line denotes Bonferroni-corrected significance, adjusted for testing 54 tissues (one-sided $P < 9.26 \times 10^{-4}$). The results are also report in Supplementary Table 16.



Extended Data Fig. 7 | Gene overlap across multiple gene-association methods. Venn diagram illustrating the overlap between (1) the nearest genes to the 579 jointly associated lead SNPs (denoted as the COJO *EXT* SNPs, see Supplementary Table 9), (2) the genes significant in the MAGMA gene-based analysis (Supplementary Table 13), (3) the genes significant in the H-MAGMA adult brain tissue analysis (Supplementary Table 17), and (4) the genes significant in the S-PrediXcan analysis (Supplementary Table 21). Across these four approaches, 34 genes were consistently implicated; these genes include *CADM2*, *PACSIN3*, *ZIC4*, *MAPT*, and *GABRA2*. Colored regions of this diagram correspond to the coloring shown in Supplementary Table 22, which lists all identified genes. No new statistical test was performed to generate this figure, and the statistical test used in each gene-based approach is reported in the notes of Supplementary Tables 9, 13, 17, and 21.



Extended Data Fig. 8 | Externalizing systems map estimated with the Order Statistics Local Optimization Method (OSLOM) algorithm. Representation of the externalizing network neighborhood estimated with PCNet as modular gene systems. In the top panel, circles represent distinct systems, with size indicating the number of genes belonging to each system (min 11 for 'cilium organization', and max 379 for the 'externalizing systems map'). System color indicates the fraction of genes in each system that have been mapped to the externalizing phenotype by at least one of the four gene mapping methods (positional, MAGMA, H-MAGMA, and S-PrediXcan). Systems have been annotated with significantly enriched gene ontology terms. Systems without significant enrichment of biological pathways are labeled with a unique system ID (C454, C461, C453, C462), and may represent novel pathways. **(i-vi)** Visualization of genes within selected systems that have been mapped to the externalizing phenotype by one or more gene mapping methods, and their molecular interactions. In the bottom panel, the gene size is mapped to the number of methods in which the gene was found associated with externalizing (with the largest genes indicating the gene was identified by all 4 methods), and gene color(s) indicates which method(s) have mapped the gene.



Extended Data Fig. 9 | Confirmatory factor analysis of phenotypic externalizing factor in Add Health and COGA. Path diagram of confirmatory factor analysis (CFA) models in (top panel) Add Health (*N* = 15,107) and (bottom panel) COGA (N = 16,857) (Supplementary Information section 5). The reported model fit statistics and fit indices are degrees of freedom (*df*), comparative fit index (*CFI*), root mean square error (*RMSEA*), standardized root mean squared residual (*SRMR*). Standardized factor loadings presented as numbers on the paths.

nature research

Corresponding author(s): Danielle M. Dick
Philipp D. Koellinger

Last updated by author(s): Jun 25, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection (because the study only analyzed existing data resources).
Data analysis	No custom algorithms or software was developed in this study. Software and code for the genetic data analysis we report can be found at: BOLT-LMM (version 2.3.2) https://alkesgroup.broadinstitute.org/BOLT-LMM/BOLT-LMM_manual.html KING (version 2.1.5): https://www.kingrelatedness.com/ FlashPCA2 (version 2.0): https://github.com/gabraham/flashpca (FlashPCA2), EasyQC (version 9.1): https://www.uni-regensburg.de/medizin/epidemiologie-praeventivmedizin/genetische-epidemiologie/software/ BCFtools (version 1.8): http://samtools.github.io/bcftools/bcftools.html LD Score regression (version 1.0.0): https://github.com/bulik/ldsc Genomic SEM (versions 0.0.2a-c): https://github.com/GenomicSEM/GenomicSEM GCTA-COJO (version 1.93.1beta): https://cnsgenomics.com/software/gcta/ METAL (versions 2011-03-25 & 2020-05-05) https://genome.sph.umich.edu/wiki/METAL_Documentation PLINK1.9 (version v1.90b6.13): https://www.cog-genomics.org/plink/ PRS-CS (version October 20, 2019): https://github.com/getian107/PRScs LDpred (version 0.9.09): https://github.com/bvilhjal/ldpred MAGMA (version 1.08): https://ctg.cncr.nl/software/magma R "base" and "stats" packages (version 3.5.1): https://cran.r-project.org/ Software, code, or webtools for the bioinformatic analyses we report can be found at: FUMA (version 1.3.5e): https://fuma.ctglab.nl/ H-MAGMA (version version June 14, 2019): https://github.com/thewonlab/H-MAGMA PrediXcan (version v0.6.2): https://github.com/hakyimlab/MetaXcan Cytoscape (version 3.8.2): https://cytoscape.org/

Tissue Specific Expression Analysis (TSEA, version 1.0): http://genetics.wustl.edu/jdlab/tsea/
Specific Expression Analysis (SEA, version 1.1): http://genetics.wustl.edu/jdlab/csea-tool-2/

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No new data was collected. Only data from existing studies or study cohorts were analyzed, some of which are restricted access to protect the privacy of the study participants. The minimum data set necessary to interpret, verify, and extend the research, i.e., the 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/. In brief, summary statistics are derived from analyses based in part on 23andMe data, for which we are restricted to 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 GWAS summary statistics. No source data is published alongside the paper.

Restricted access individual-level phenotype and genetic data:
Add Health, dbGaP Study Accession: phs001367.v1.p1
Collaborative Study on the Genetics of Alcoholism (COGA), dbGaP Study Accession: phs000763.v1.p1
Philadelphia Neurodevelopmental Cohort, dbGaP Study Accession: phs000607.v3.p2
UK Biobank: https://www.ukbiobank.ac.uk/
Vanderbilt University Medical Center biobank (BioVU): https://victr.vumc.org/biovu-description/

Restricted access reference data:
UK10K, accession code(s) EGAD00001000740 (https://ega-archive.org/datasets/EGAD00001000740); EGAD00001000741 (https://ega-archive.org/datasets/EGAD00001000741)

Publicly available reference data:
1000 Genomes phase 3 (version 5): https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
HapMap 3 (revision 2): https://mathgen.stats.ox.ac.uk/impute/data_download_hapmap3_r2.html
Haplotype Reference Consortium variant site list (version 1.1): http://www.haplotype-reference-consortium.org/site
H-MAGMA reference data: https://github.com/thewonlab/H-MAGMA
Molecular Signatures Database (MsigDB version 7.0): https://www.gsea-msigdb.org/gsea/msigdb/
Genotype-Tissue Expression database (GTEx, version 8.0): https://gtexportal.org/home/datasets
PredictDB Data Repository: http://predictdb.org

Publicly available GWAS summary statistics:
Broad Antisocial Behavior Consortium (Broad ABC): http://broadabc.ctglab.nl/summary_statistics
Genomics of Personality Consortium (GPC): https://tweelingenregister.vu.nl/gpc
GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN): https://conservancy.umn.edu/handle/11299/201564
International Cannabis Consortium (ICC): https://www.ru.nl/bsi/research/group-pages/substance-use-addiction-food-saf/vm-saf/genetics/international-cannabis-consortium-icc/
Psychiatric Genomics Consortium: https://www.med.unc.edu/pgc/download-results/
Social Science Genetic Association Consortium: https://www.thessgac.org/data

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

Field-specific reporting

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

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

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

Behavioural & social sciences study design

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

Study description	A study introduction and motivation is given in Supplementary Information section 1. The study procedure can broadly be categorized into three major stages: 1. We amassed a set of phenotype-specific GWAS summary statistics for different externalizing phenotypes, either by collecting existing results or by performing GWAS in UK Biobank (UKB) (Supplementary Information section 2). The multivariate method

"genomic structural equation modelling" (Genomic SEM) was applied on a subset of the summary statistics (N = 53,293–1,251,809) deemed adequately heritable and statistically powered, in order to estimate a series of model specifications representing different genetic factor structures (Supplementary Information section 3). The best-fitting and most parsimonious solution ("the preferred model specification") specified a single common genetic factor with seven indicator phenotypes (which we hereafter refer to as "the latent genetic externalizing factor", or simply, "the externalizing factor"). We estimated genetic correlations between the externalizing factor and 92 other traits from various research domains. Our main discovery analysis is a GWAS on the latent genetic externalizing factor, which we henceforth refer to as "the externalizing GWAS" (Neff = 1,492,085). The externalizing GWAS results were first clumped and then subjected to "conditional and joint multiple-SNP analysis" (GCTA-COJO) to identify a set of "579 jointly associated lead SNPs", which we consider to be our main GWAS findings.

2. The results of the externalizing GWAS were utilized to perform proxy-phenotype analyses of antisocial behavior and alcohol use disorder (Supplementary Information section 4). Similarly, the results were used for polygenic score analyses of a variety of behavioral, health, criminal justice, and substance use measures, including a phenome-wide association study (PheWAS) of electronic-health records in the biorepository of the Vanderbilt University Medical Center (BioVU) (Supplementary Information section 5).

3. Bioannotation of the externalizing GWAS was performed with the methods "functional mapping and annotation of genetic associations" (FUMA), "multi-marker analysis of genomic annotation" (MAGMA), "Hi-C coupled MAGMA" (H-MAGMA), and "S-PrediXcan" (Supplementary Information section 6).

Research sample

Restricted access individual-level phenotype and genetic data:
Add Health, dbGaP Study Accession: phs001367.v1.p1
Collaborative Study on the Genetics of Alcoholism (COGA), dbGaP Study Accession: phs000763.v1.p1
Philadelphia Neurodevelopmental Cohort, dbGaP Study Accession: phs000607.v3.p2
UK Biobank: https://www.ukbiobank.ac.uk/
Vanderbilt University Medical Center biobank (BioVU): https://victr.vumc.org/biovu-description/

Restricted access reference data:
UK10K, accession code(s) EGAD00001000740 (https://ega-archive.org/datasets/EGAD00001000740); EGAD00001000741 (https://ega-archive.org/datasets/EGAD00001000741)

Publicaly available reference data:
1000 Genomes phase 3 (version 5): https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
HapMap 3 : https://mathgen.stats.ox.ac.uk/impute/data_download_hapmap3_r2.html
Haplotype Reference Consortium variant site list (version 1.1): http://www.haplotype-reference-consortium.org/site
H-MAGMA reference data: https://github.com/thewonlab/H-MAGMA
Molecular Signatures Database (MsigDB version 7.0): https://www.gsea-msigdb.org/gsea/msigdb/
Genotype-Tissue Expression database (GTEx, version 8.0): https://gtexportal.org/home/datasets
PredictDB Data Repository: http://predictdb.org

Publicly available GWAS summary statistics:
Broad Antisocial Behavior Consortium (Broad ABC): http://broadabc.ctglab.nl/summary_statistics
Genomics of Personality Consortium (GPC): https://tweelingenregister.vu.nl/gpc
GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN): https://conservancy.umn.edu/handle/11299/201564
International Cannabis Consortium (ICC): https://www.ru.nl/bsi/research/group-pages/substance-use-addiction-food-saf/vm-saf/genetics/international-cannabis-consortium-icc/
Psychiatric Genomics Consortium: https://www.med.unc.edu/pgc/download-results/
Social Science Genetic Association Consortium: https://www.thessgac.org/data

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

Sampling strategy

The sampling strategies of the existing studies or study cohorts that were analyzed in this study are described in their respective references (see Supplementary Information).

We aimed to attain the largest molecular genetic study on externalizing traits. The preregistered analysis plan, the first version of which was time-stamped on November 8, 2018 (https://doi.org/10.17605/OSF.IO/XKV36), specified a minimum GWAS sample size of N>15,000, which was determined by multiplying the recommended minimum N for LD score regression (i.e., 5,000) by three. After additional exclusions, time-stamped on March 29, 2019 (https://doi.org/10.17605/OSF.IO/XKV36), based on negligible SNP-heritability or GWAS signal, all remaining GWAS summary statistics included more than 50,000 people. In the final Genomic SEM model, we estimated the lower bound of independent observations to be 1,373,240, which makes it among the largest genome-wide association studies ever conducted, and thus, suggests that the study was adequately powered to find replicable SNP associations.

Data collection

No new data was collected in this study, and thus, the study was neither randomized nor blinded.

Timing

Public and restricted access data sources were accessed between June, 2018, and June, 2020.

Data exclusions

Genotype quality-control exclusion criteria were specified in the preregistered analysis plan on Open Science Framework (https://doi.org/10.17605/OSF.IO/XKV36). The study was restricted to analyses in European-ancestry individuals that passed all genotype quality control procedures. Pre-registered SNP quality-control exclusion criteria was applied to exclude low-quality or rare genetic variants (described in detail in Supplementary Information section 2).

Non-participation

No participants dropped out/declined participation.

Randomization	Randomization to an experimental condition was not applied because the study is not experimental. Analyses were statistically adjusted for age, sex, genetic principal components, genotyping array and batch. Within-family analysis were performed to confirm the robustness of the results to potential bias from non-random allotment of genotypes, because meiosis randomizes genotypes to siblings.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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