nature genetics

LETTERS

https://doi.org/10.1038/s41588-018-0307-5

Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use

Tobacco and alcohol use are leading causes of mortality that influence risk for many complex diseases and disorders[1]. They are heritable[2,3] and etiologically related[4,5] behaviors that have been resistant to gene discovery efforts[6–11]. In sample sizes up to 1.2 million individuals, we discovered 566 genetic variants in 406 loci associated with multiple stages of tobacco use (initiation, cessation, and heaviness) as well as alcohol use, with 150 loci evidencing pleiotropic association. Smoking phenotypes were positively genetically correlated with many health conditions, whereas alcohol use was negatively correlated with these conditions, such that increased genetic risk for alcohol use is associated with lower disease risk. We report evidence for the involvement of many systems in tobacco and alcohol use, including genes involved in nicotinic, dopaminergic, and glutamatergic neurotransmission. The results provide a solid starting point to evaluate the effects of these loci in model organisms and more precise substance use measures.

An analysis overview is provided in Supplementary Fig. 1; all independent associated variants are in Supplementary Tables 1–5; and quantile-quantile, Manhattan, and LocusZoom plots are shown in Supplementary Figs. 2–12. Smoking initiation phenotypes included age of initiation of regular smoking (AgeSmk; $n = 341{,}427$; 10 associated variants) and a binary phenotype indicating whether an individual had ever smoked regularly (SmkInit; $n = 1{,}232{,}091$; 378 associated variants). Heaviness of smoking was measured with cigarettes per day (CigDay; $n = 337{,}334$; 55 associated variants). Smoking cessation (SmkCes; $n = 547{,}219$; 24 associated variants) was a binary variable contrasting current versus former smokers. Available measures of alcohol use were simpler, with drinks per week (DrnkWk; $n = 941{,}280$; 99 associated variants) widely available and similarly measured across studies. See the Supplementary Note and Supplementary Tables 6 and 7 for phenotype definition details.

The four smoking phenotypes were genetically correlated with one another (Fig. 1 and Supplementary Table 8). DrnkWk was not highly genetically correlated with the smoking phenotypes ($r_g \sim 0.10$) except for SmkInit ($r_g \sim 0.34$, $p = 6.7 \times 10^{-63}$), suggesting that sequence variations affecting alcohol use and those affecting initiation of smoking overlap substantially. The phenotypes were highly genetically correlated across constituent studies (Supplementary Table 9), suggesting a minor effect of phenotypic heterogeneity in the present results, even across Western Europe and the United States. Smoking phenotypes were genetically correlated in expected directions with many behavioral, psychiatric, and medical phenotypes (Fig. 1 and Supplementary Table 10). Genetic variation associated with increased alcohol use was associated with greater levels of risky behavior ($r_g = 0.20$, $p = 1.8 \times 10^{-7}$) and cannabis use ($r_g = 0.36$, $p = 6.2 \times 10^{-10}$), but with less risk of disease for almost all diseases (Fig. 1 and Supplementary Table 10).

Using a novel method to evaluate multivariate genetic correlation at the locus (versus global) level, we observed 150 loci that affected multiple substance use phenotypes (Fig. 2 and Supplementary Table 11). Patterns of pleiotropy across phenotypes were highly diverse, with only three loci significantly associated with all five phenotypes. These three loci included associations implicating phosphodiesterase 4B (*PDE4B*) and cullin 3 (*CUL3*). PDE4B regulates cyclic AMP second messenger availability and thereby affects signal transduction, and it is downregulated by chronic nicotine administration in rats[12]. CUL3 has wide-ranging effects, including on ubiquination and protein degradation, and de novo mutations in *CUL3* are associated with rare diseases affecting response to the mineralocorticoid aldosterone[13], which itself is affected by smoking[14] and is associated with alcohol use[15]. In addition to testing for pleiotropy, we also used MTAG[16] to leverage the observed genetic correlations to increase power for locus discovery. Using this method, we discovered 1,193 independent, genome-wide significantly associated common variants (minor allele frequency (MAF), >1%; AgeSmk, 173; CigDay, 89; SmkCes, 83; SmkInit, 692; DrnkWk, 156) listed in Supplementary Table 12 and described further in the supplementary information.

Phenotypic variation accounted for by our initial 566 conditionally independent genome-wide significant variants from the initial genome-wide association study (GWAS) ranged from 0.1% (SmkCes) to 2.3% (SmkInit; see Fig. 3). SNP heritability calculated using linkage disequilibrium (LD) score regression[17] ranged from 4.2% for DrnkWk to 8.0% for CigDay (Fig. 3 and Supplementary Table 13), consistent with estimates made using individual-level data[18], SNP heritabilities calculated from the largest individual contributing studies (Supplementary Table 13), and prior work[19]. The results suggest that these phenotypes are highly polygenic and that the majority of the heritability is accounted for by variants below standard GWAS thresholds.

To further investigate the polygenicity, polygenic risk scores (PRS; Supplementary Table 14) were computed on the National Longitudinal Study of Adolescent to Adult Health (Add Health)[20] and the Health and Retirement Study (HRS)[21] datasets, which are representative of their birth cohorts in the United States and represent exposures to different tobacco policy environments. Add Health participants were born, on average, in 1979; average birth year in the HRS was 1938. Despite these generational differences, the polygenic score performed similarly in both samples. It accounted for approximately 1%, 4%, 1%, 4%, and 2.5% of variance in AgeSmk, CigDay, SmkCes, SmkInit, and DrnkWk, respectively, about half of the estimated SNP heritability of these traits (Fig. 3). More concretely, in Add Health and the HRS, respectively, a 1 s.d. increase in the CigDay risk score resulted in two and three additional daily cigarettes; a 1 s.d. increase on the SmkInit risk score resulted in a 12% and 10% increased risk of regularly smoking; and a 1 s.d. increase on the DrnkWk risk score reflected one additional drink per week in both datasets.

A full list of authors and affiliations appears at the end of the paper.

AgeSmk	CigDay	SmkInit	SmkCes	DrnkWk	
$h^2 = 0.05$	−0.38**	−0.71**	−0.31**	−0.10*	Age of smoking initiation (AgeSmk)
−0.38**	$h^2 = 0.08$	0.33**	0.42**	0.07*	Cigarettes per day (CigDay)
−0.71**	0.33**	$h^2 = 0.08$	0.40**	0.34**	Smoking initiation (SmkInit)
−0.31**	0.42**	0.40**	$h^2 = 0.05$	0.11**	Smoking cessation (SmkCes)
−0.10*	0.07*	0.34**	0.11**	$h^2 = 0.04$	Drinks per week (DrnkWk)
0.04	−0.01	−0.03	−0.10**	−0.02	Height
0.22**	−0.09*	−0.04	−0.02	0.08*	Age at menarche
0.67**	−0.40**	−0.48**	−0.46**	0.01	Age of first birth
0.55**	−0.26**	−0.40**	−0.51**	0.01	Years of education
−0.21	0.63*	0.16	0.43*	−0.02	Cotinine
−0.32**	0.15**	0.28**	0.12*	0.20**	General risk tolerance
−0.43**	0.09	0.60**	0.06	0.36**	Lifetime cannabis use
−0.31*	0.20*	0.41**	0.17*	0.17	ADHD
0.06	−0.04	0.01	−0.08	−0.03	Autism spectrum disorder
0.02	0.06	0.06	−0.10*	0.04	Bipolar disorder
−0.17*	0.12*	0.19**	0.26**	−0.06	Major depressive disorder
−0.17**	0.13**	0.20**	0.20**	0.02	Neuroticism
−0.05	0.10**	0.14**	0.06*	0.01	Schizophrenia
−0.05	−0.02	−0.06	0.08	0.13	Alzheimer's
−0.03	−0.01	−0.04	0.04	0.03	Multiple sclerosis
−0.02	0.02	0.02	0.01	−0.10*	Parkinson's
−0.16**	0.19**	0.12**	0.12**	−0.08*	Body mass index
−0.20**	0.24**	0.13**	0.17**	−0.11**	Obesity class I
0.03	−0.04	−0.08*	0.02	0.03	Bone density: femoral neck
0.03	0.01	−0.06	0.04	0.02	Lumbar spine
0.16**	−0.17**	−0.09**	−0.15**	0.17**	Cholesterol: HDL
−0.06	0.07*	0.02	0.10*	−0.03	LDL
−0.03	0.07*	0.03	0.08*	−0.01	Total
0.01	0.05	−0.08	0.16*	−0.11	Chronic kidney disease
−0.27**	0.25**	0.19**	0.21**	−0.01	Coronary artery disease
−0.16*	0.15*	0.07*	0.06	−0.08*	Diabetes: type 2
−0.13*	0.17*	0.07*	0.11*	−0.01	Fasting main effect: glucose
−0.24*	0.16*	0.09*	0.10	−0.24**	Insulin
−0.17	0.22*	0.11	0.20	−0.01	Proinsulin
−0.03	0.14*	0.04	0.11*	−0.04	Heart rate
0.04	0.03	0.01	−0.05	−0.06*	Inflammatory bowel disease
0.08	−0.01	−0.03	−0.13*	−0.06	Ulcerative colitis
−0.04	0.08	0.08	0.02	−0.07	Primary biliary cirrhosis
0.06	0.07	0.01	0.06	−0.06	Systemic lupus erythematosus

Fig. 1 | Genetic correlations between substance use phenotypes and phenotypes from other large GWAS. Genetic correlations between each of the phenotypes are shown in the first five rows, with heritability estimates displayed down the diagonal. All genetic correlations and heritability estimates were calculated using LD score regression. Purple shading represents negative genetic correlations, and red shading represents positive correlations, with increasing color intensity reflecting increasing correlation strength. A single asterisk reflects a significant genetic correlation at the $P < 0.05$ level. Double asterisks reflect a significant genetic correlation at the Bonferroni-correction $P < 0.000278$ level (corrected for 180 independent tests). Note that SmkCes was oriented such that higher scores reflected current smoking, and for AgeSmk, lower scores reflect earlier ages of initiation, both of which are typically associated with negative outcomes.

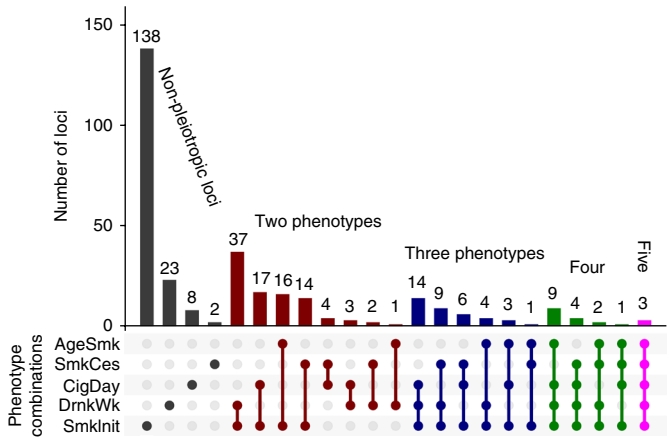

Fig. 2 | Pleiotropy. Depicted here are results from the multivariate analysis of pleiotropy. For each locus, the method returns the best-fitting solution of which phenotypes were associated with that locus. All loci with one or more associated phenotypes are shown here. For example, every locus associated with AgeSmk was found to be pleiotropic for other phenotypes (green, blue, red, purple, and fuchsia bars), and no locus showed association with only AgeSmk (no dark gray bar for AgeSmk). When sample sizes are unequal across phenotypes, the method also improves power for those phenotypes with smaller samples. The total numbers of loci associated with each trait (whether pleiotropic or not) from these analyses were 40 (AgeSmk), 48 (SmkCes), 72 (CigDay), 111 (DrnkWk), and 278 (SmkInit). Full information is in Supplementary Table 11.

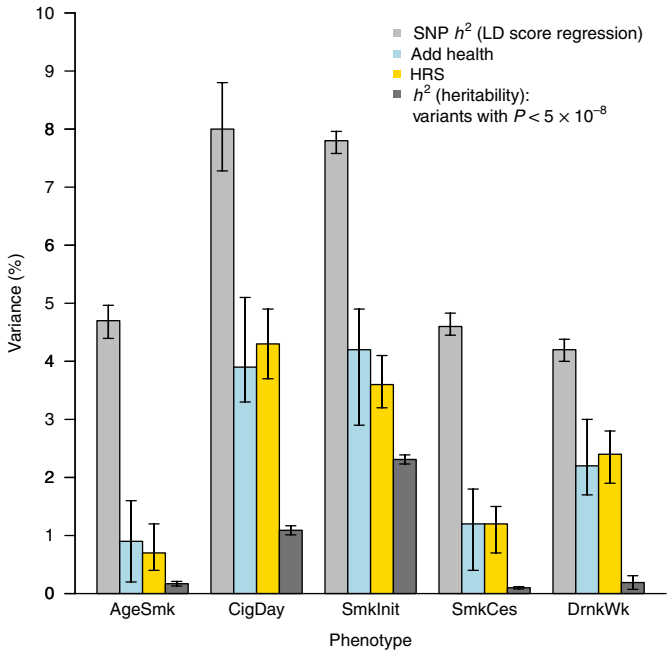

Fig. 3 | Heritability and polygenic prediction. The light gray bars reflect SNP heritability, estimated with LD score regression. The light blue and gold bars reflect the predictive power of a PRS in Add Health and the HRS, respectively. Despite the 41 year generational gap between participants from these two studies, and major tobacco-related policy changes during that time, the polygenic scores are similarly predictive in both samples. Error bars are 95% confidence intervals estimated with 1,000 bootstrapped repetitions. Dark gray bars represent the total phenotypic variance explained by only genome-wide significant SNPs.

Cell and tissue enrichment[22] was observed across all five phenotypes within core histone marks from multiple central nervous system tissues (Supplementary Figs. 13–15 and Supplementary Tables 15 and 16). Enrichment was observed in tissues from cortical and sub-cortical regions in the central nervous system. Structure and function of these regions have been robustly associated with individual differences in frequencies, magnitudes, and clinical characteristics of alcohol use, and substance use/misuse generally, in human imaging research. Our results include significant enrichment across phenotypes and histone marks in the hippocampus[23], inferior temporal pathways[24], dorsolateral and medial prefrontal cortex[25], caudate, and striatum[26]. Consistent with gene and pathway findings described below, alcohol and nicotine use affect dopaminergic and glutamatergic neurotransmission among these brain regions, compromising reward-based learning and facilitating drug-seeking behavior[26]. Enrichment within other cell or tissue groups and specific cell or tissue types included immune and liver cells, but was less consistent across analytical approaches.

We manually reviewed all of the genes implicated by the GWAS or gene-based tests (see Supplementary Tables 1–5 for the full catalog of implicated genes and Supplementary Tables 17–21 for gene and gene set test results). We replicated known associations between multiple variants in the nicotine metabolism gene *CYP2A6*

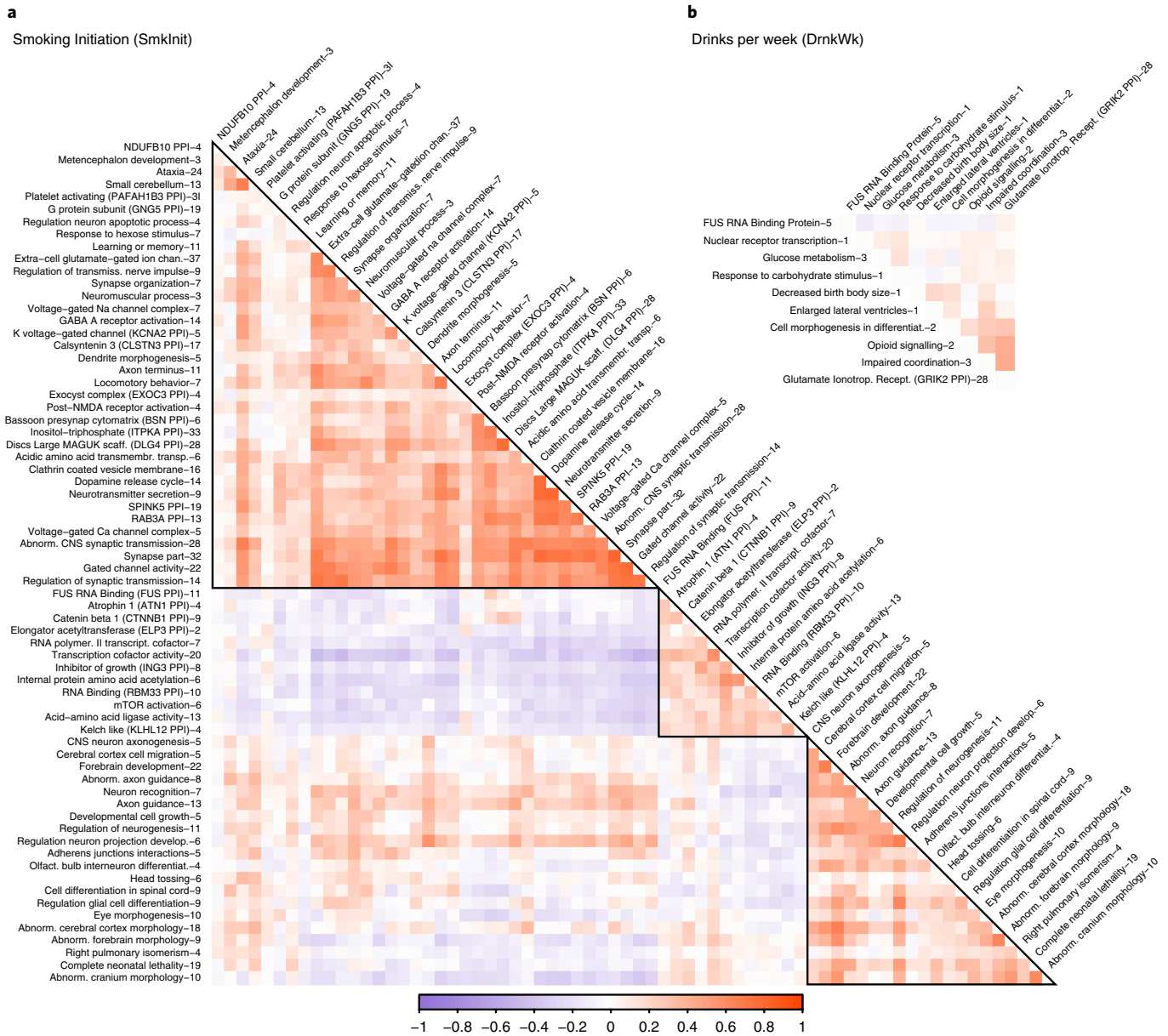

Fig. 4 | Correlations among exemplary DEPICT gene sets. (**a**,**b**) There were 68 clusters available for SmkInit (**a**) and 10 for DrnkWk (**b**) (CigDay, AgeSmk, and SmkCes did not have >1 exemplary set). Purple shading represents negative correlations, and red shading represents positive correlations, with increasing color intensity reflecting increasing correlation strength. Cluster names are truncated for space, with a full list of all names in Supplementary Table 18. The number after each name is the number of gene sets in each cluster. The matrix naturally falls into three red superclusters along the diagonal. The largest supercluster contains primarily gene sets related to neurotransmitter receptors, ion channels (sodium, potassium, calcium), learning/memory, and other aspects of central nervous system function. The middle supercluster includes gene sets defined by regulation of transcription and translation, including RNA binding and transcription factor activity. The final supercluster is composed primarily of gene sets related to development of the nervous system.

with CigDay ($P=4.0\times10^{-99}$) and SmkCes ($P=1.6\times10^{-48}$). We replicated an association signal in the alcohol metabolism gene *ADH1B* associated with DrnkWk, identifying in that locus 11 conditionally independently associated variants (lowest $P<2.2\times10^{-303}$).

All drugs of abuse activate the mesolimbic dopamine system reward pathway[27], and dopamine-related genes have long been popular candidate genes. We found that variants near the widely studied dopamine receptor D2 (*DRD2*)[28] were associated across phenotypes, including CigDay, SmkCes, and DrnkWk ($P=6.5\times10^{-12}$, 1.1×10^{-10}, and 4.9×10^{-11}, respectively), but not with AgeSmk or SmkInit, suggesting that these variants are less relevant in early stages of nicotine use. Other specific dopamine-related genes only showed associations with smoking phenotypes, including multiple associations between CigDay and SmkCes with dopamine β-hydroxylase (*DBH*; $P=9.8\times10^{-24}$ and 1.2×10^{-35}, respectively)[9], an enzyme necessary to convert dopamine to norepinephrine. SmkInit was associated with variation near protein phosphatase 1 regulatory subunit 1B (*PPP1R1B*; $P=3.9\times10^{-8}$), a signal transduction gene that affects synaptic plasticity and reward-based learning in the striatum[29,30] and contributes to the behavioral effects of nicotine in mice[31]. In pathway analyses, dopamine gene sets were enriched only in SmkInit, where the exemplar 'reactome dopamine neurotransmitter release cycle' pathway was enriched ($P=9.2\times10^{-5}$; Fig. 4 and Supplementary Table 18).

Table 1 | Non-synonymous sentinel variants

Phenotype	Gene	rsID	Chr	Position	REF	ALT	AF	Beta	*P*	*N*	*Q*
CigDay (SmkCes)	*CHRNA5*	rs16969968[a]	15	78,882,925	G	A	0.34	0.075	1.2×10^{-278}	330,721	0.34
CigDay	*HIST1H2BE*	rs7766641	6	26,184,102	G	A	0.27	−0.014	2.9×10^{-10}	335,553	0.78
CigDay (AgeSmk)	*GRK4*	rs1024323	4	3,006,043	C	T	0.38	−0.012	8.7×10^{-9}	337,334	0.17
SmkInit	*REV3L*	rs462779[a]	6	111,695,887	G	A	0.81	−0.019	4.5×10^{-29}	1,232,091	0.67
SmkInit (DrnkWk)	*BDNF*	rs6265	11	27,679,916	C	T	0.20	−0.016	2.8×10^{-19}	1,232,091	0.13
SmkInit	*RHOT2*	rs1139897	16	720,986	G	A	0.23	−0.012	1.8×10^{-15}	1,232,091	0.61
SmkInit (DrnkWk)	*ZNF789*	rs6962772[a]	7	99,081,730	A	G	0.15	−0.015	2.1×10^{-14}	1,232,091	0.92
SmkInit	*BRWD1*	rs4818005[a]	21	40,574,305	A	G	0.58	−0.010	3.9×10^{-14}	1,232,091	0.75
SmkInit	*ENTPD6*	rs6050446	20	25,195,509	A	G	0.97	0.035	8.8×10^{-13}	1,225,969	0.33
SmkInit	*RPS6KA4*	rs17857342[a]	11	64,138,905	T	G	0.38	−0.010	9.8×10^{-12}	1,232,091	0.16
SmkInit	*FAM163A*	rs147052174	1	179,783,167	G	T	0.02	0.037	2.3×10^{-10}	1,232,091	0.59
SmkInit	*PRRC2B*	rs34553878	9	134,907,263	A	G	0.11	0.016	1.2×10^{-9}	1,232,091	0.28
SmkInit	*ADAM15*	rs45444697[a]	1	155033918	C	T	0.21	0.010	5.3×10^{-9}	1,232,091	0.46
SmkInit	*MMS22L*	rs9481410[a]	6	97,677,118	G	A	0.76	0.010	1.1×10^{-8}	1,232,091	0.04
SmkInit	*QSER1*	rs62618693	11	32,956,492	C	T	0.04	−0.020	2.1×10^{-8}	1,232,091	1.00
DrnkWk	*ADH1B*	rs1229984	4	100,239,319	T	C	0.96	0.060	2.2×10^{-308}	941,280	0.05
DrnkWk	*GCKR*	rs1260326	2	27,730,940	T	C	0.60	0.008	8.1×10^{-45}	941,280	0.10
DrnkWk	*SLC39A8*	rs13107325	4	103,188,709	C	T	0.07	−0.009	1.5×10^{-22}	941,280	0.33
DrnkWk	*SERPINA1*	rs28929474	14	94,844,947	C	T	0.02	−0.012	1.3×10^{-11}	941,280	0.50
DrnkWk (SmkInit)	*ACTR1B*	rs11692465	2	98,275,354	G	A	0.09	0.008	2.5×10^{-11}	937,516	0.40
DrnkWk	*TNFSF12-13*	rs3803800	17	7,462,969	A	G	0.79	0.004	1.5×10^{-10}	941,280	0.67
DrnkWk	*HGFAC*	rs3748034	4	3,446,091	G	T	0.14	−0.005	1.7×10^{-8}	941,280	0.65

The sentinel variant in approximately 4% of loci was non-synonymous. Shown here are all non-synonymous sentinel variants, and all non-synonymous variants in near-perfect LD with a sentinel variant. If the listed gene was also associated (through single variant or gene-based test) with another phenotype, that phenotype is listed in parentheses. Several genes have been implicated in previous studies of substance use/addiction, including *CHRNA5*, *BDNF*, *GCKR*, and *ADH1B*. Phenotype abbreviations are defined in Fig. 1. Chr, chromosome; REF, reference allele; ALT, alternate allele; AF, allele frequency of ALT; *Q*, Cochrane's *Q* statistic *P* value. [a]These variants were not themselves sentinel, but were in near-perfect LD with a sentinel variant ($r^2 > 0.99$, from the 1000 Genomes European population). The scale of Beta is on the unit of the standard deviation of the phenotype. For binary phenotypes the standard deviation was calculated from the weighted average prevalence across all studies included in the meta-analysis (available in Supplementary Table 7).

Neuronal acetylcholine nicotinic receptors are the initial site of nicotine action in the brain and have long been implicated in nicotine use and dependence[32]. With the exception of *CHRNA7*, all central-nervous-system-expressed nicotinic receptor genes were significantly associated with one or more smoking phenotypes, many reported here for the first time. Enrichment was also noted for nicotinic-receptor-related pathways and genes in smoking phenotypes (Supplementary Tables 17–21). There was no evidence of association between nicotinic receptor genes or pathways with DrnkWk, despite the use of nicotinic receptor partial agonists (for example, varenicline) in the treatment of alcohol dependence[33].

Associations with SmkInit highlighted structures and functions related to long-term potentiation and reward-related learning and memory, systems that affect reward processing and addiction[28,34,35]. Glutamate is an important neurotransmitter mediating these processes, and exemplar pathways related to glutamate were significantly enriched in SmkInit (for example, 'extracellular-glutamate-gated ion channel', $P=9.9\times10^{-7}$; 'post-NMDA receptor activation events', $P=5.5\times10^{-5}$; and 'DLG4 PPI subnetwork', $P=4.5\times10^{-12}$; Supplementary Table 18). DLG4 affects NMDA receptors and potassium channel clusters and has a central role in glutamatergic models of reward-related learning[35]. Individual associated genes related to these pathways included glutamate ionotropic receptor NMDA type subunit 2 (*GRIN2A*; $P=3.4\times10^{-11}$) and homer scaffolding protein 2 (*HOMER2*; $P=3.1\times10^{-14}$), which affects addictive behavior in mice[35,36] and regulates glutamate metabotropic receptor 1 (*GRM1*). Pathways enriched in SmkInit also included sodium-, potassium-, and calcium voltage-gated channels (Fig. 4 and Supplementary Table 18), essential to neuronal excitability and signaling.

Alcohol is known to affect glutamatergic signaling pathways[37], and more than half of the enriched pathways for DrnkWk clustered within the exemplar 'glutamate ionotropic receptor kainate type subunit 2 (GRIK2) PPI subnetwork' (Fig. 4 and Supplementary Table 18). However, not all DrnkWk-enriched pathways involved the brain as glucose and carbohydrate processing pathways were associated with DrnkWk but no smoking phenotype, perhaps suggesting that alcohol consumption is influenced by individual differences in one's ability to process calorie-rich alcoholic beverages. Finally, we discovered variation in and around gene-rich regions, including corticotropin-releasing hormone receptor 1 (*CRHR1*; $P=1.6\times10^{-17}$) and urocortin (*UCN*; $P=8.1\times10^{-45}$), associated with DrnkWk, but not smoking. *UCN* encodes an endogenous ligand for *CRHR1* and *CRHR2* (ref. [38]). CRH affects hormones involved in the stress response, including cortisol, and has been associated with the stress response and relapse to drug taking in animals[39,40].

Specific mechanisms by which implicated genes influence substance use in humans are largely unknown, even for those genes reported above involving systems, such as neurotransmission, reward-related learning and memory, and the stress response. To prioritize genes for functional experimentation, we tabulated conditionally independent genome-wide significant non-synonymous variants (Table 1). In the 406 GWAS loci, 4% of sentinel variants were non-synonymous, representing a significant enrichment ($P=2.5\times10^{-10}$; 0.4% of variants with MAF>0.1% in the imputation panel[41] were non-synonymous). Several genes in Table 1 have been previously associated with substance use/addiction (see Supplementary Table 22 for a list of previous associations), and two variants have been functionally validated (rs1229984 and rs16969968)[42,43]. The others have

not, but in some cases their genes interact with established molecular targets of addiction and may themselves be suitable targets for further investigation. For example, rs1024323 in G-protein-coupled receptor kinase 4 (*GRK4*) was associated with CigDay ($P = 8.7 \times 10^{-9}$) and lies within a locus associated with AgeSmk. *GRK4* is involved in the regulation of G-protein-coupled receptors, including metabotropic glutamate receptor 1 (*GRM1*)[44], $GABA_B$ receptors[45], and dopamine receptors D1 (*DRD1*) and D3 (*DRD3*) in the kidneys and cerebellum, and is involved in essential hypertension[46]. *GRK4* is also expressed in the midbrain and forebrain[46,47], but no research has evaluated its impact on substance use behavior. To take one more example, the non-synonymous variant in *SLC39A8* affects zinc and manganese transport, is highly pleiotropic for complex phenotypes, and may impair inflammation, glutamatergic neurotransmission, and regulation of various metals in the body[48].

Ultimately, substance use is embedded in a complex web of causal relations[49] (for example, see Fig. 1), and caution must be exercised in drawing strong causal conclusions. However, our findings represent a major step forward in understanding the etiology of these complex, disease-relevant behaviors. In particular, statistical and interpretive power were both enabled by simultaneously studying multiple related substance use behaviors representing different stages of use and different substances. More precise measurements, including evaluating age and environment as moderators for these dynamic phenotypes[50], functional research, and complementary gene mapping approaches (for example, sequencing) will aid in the discovery of mechanisms by which implicated genes may affect substance use and related disease risk.

URLs. GSCAN website (with summary statistics and LocusZoom plots for MTAG loci), https://genome.psych.umn.edu/index.php/GSCAN; ANNO, https://github.com/zhanxw/anno/; APIGenome, https://github.com/hyunminkang/apigenome/; BCFtools, http://samtools.github.io/bcftools/; BOLT-LMM, https://data.broadinstitute.org/alkesgroup/BOLT-LMM/; DEPICT, https://data.broadinstitute.org/mpg/depict/; GCTA, http://cnsgenomics.com/software/gcta/; GenomicSEM, https://github.com/MichelNivard/GenomicSEM/; LDpred, https://github.com/bvilhjal/ldpred/; LDSC, https://github.com/bulik/ldsc/; LocusZoom, https://github.com/statgen/locuszoom-standalone/; Michigan Imputation Server, http://imputationserver.sph.umich.edu/; Minimac3, https://genome.sph.umich.edu/wiki/Minimac3; MTAG: https://github.com/omeed-maghzian/mtag/; PASCAL, https://www2.unil.ch/cbg/index.php?title=Pascal; PLINK, https://www.cog-genomics.org/plink/1.9/; PriorityPruner, http://prioritypruner.sourceforge.net/; R, https://www.r-project.org/; rareGWAMA, https://github.com/dajiangliu/rareGWAMA/; RiVIERA, https://github.com/yueli-compbio/RiVIERA/; RVTESTS, https://github.com/zhanxw/rvtests/; SEQMINER, https://github.com/zhanxw/seqminer/; SHAPEIT, http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at http://doi.org/https://doi.org/10.1038/s41588-018-0307-5.

Received: 1 April 2018; Accepted: 6 November 2018;
Published online: 14 January 2019

References

1. Ezzati, M. et al. Selected major risk factors and global and regional burden of disease. *Lancet* **360**, 1347–1360 (2002).
2. Hicks, B. M., Schalet, B. D., Malone, S. M., Iacono, W. G. & McGue, M. Psychometric and genetic architecture of substance use disorder and behavioral disinhibition measures for gene association studies. *Behav. Genet.* **41**, 459–475 (2011).
3. Polderman, T. J. et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
4. Kendler, K. S., Schmitt, E., Aggen, S. H. & Prescott, C. A. Genetic and environmental influences on alcohol, caffeine, cannabis, and nicotine use from early adolescence to middle adulthood. *Arch. Gen. Psychiatry* **65**, 674–682 (2008).
5. Kendler, K. S., Prescott, C. A., Myers, J. & Neale, M. C. The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. *Arch. Gen. Psychiatry* **60**, 929–937 (2003).
6. Bierut, L. J. et al. *ADH1B* is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry. *Mol. Psychiatry* **17**, 445–450 (2012).
7. Thorgeirsson, T. E. et al. Sequence variants at *CHRNB3–CHRNA6* and *CYP2A6* affect smoking behavior. *Nat. Genet.* **42**, 448–453 (2010).
8. Thorgeirsson, T. E. et al. A rare missense mutation in *CHRNA4* associates with smoking behavior and its consequences. *Mol. Psychiatry* **21**, 594–600 (2016).
9. Furberg, H. et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).
10. Schumann, G. et al. *KLB* is associated with alcohol drinking, and its gene product β-Klotho is necessary for FGF21 regulation of alcohol preference. *Proc. Natl Acad. Sci. USA* **113**, 14372–14377 (2016).
11. Jorgenson, E. et al. Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Mol. Psychiatry* **22**, 1359–1367 (2017).
12. Polesskaya, O. O., Smith, R. F. & Fryxell, K. J. Chronic nicotine doses down-regulate PDE4 isoforms that are targets of antidepressants in adolescent female rats. *Biol. Psychiatry* **61**, 56–64 (2007).
13. Boyden, L. M. et al. Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. *Nature* **482**, 98–102 (2012).
14. Wang, W. et al. Forced expiratory volume in the first second and aldosterone as mediators of smoking effect on stroke in African Americans: the Jackson Heart Study. *J. Am. Heart Assoc.* **5**, e002689 (2016).
15. Aoun, E. G. et al. A relationship between the aldosterone-mineralocorticoid receptor pathway and alcohol drinking: preliminary translational findings across rats, monkeys and humans. *Mol. Psychiatry* **23**, 1466–1473 (2018).
16. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
17. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
18. Yang, J. A., 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).
19. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
20. 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).
21. Sonnega, A. et al. Cohort profile: the Health and Retirement Study (HRS). *Int. J. Epidemiol.* **43**, 576–585 (2014).
22. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
23. Wilson, S., Bair, J. L., Thomas, K. M. & Iacono, W. G. Problematic alcohol use and reduced hippocampal volume: a meta-analytic review. *Psychol. Med.* **47**, 2288–2301 (2017).
24. Ewing, S. W. F., Sakhardande, A. & Blakemore, S. J. The effect of alcohol consumption on the adolescent brain: a systematic review of MRI and fMRI studies of alcohol-using youth. *Neuroimage Clin.* **5**, 420–437 (2014).
25. Goldstein, R. Z. & Volkow, N. D. Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nat. Rev. Neurosci.* **12**, 652–669 (2011).
26. Volkow, N. D. & Morales, M. The brain on drugs: from reward to addiction. *Cell* **162**, 712–725 (2015).
27. Koob, G. F. & Volkow, N. D. Neurocircuitry of addiction. *Neuropsychopharmacology* **35**, 217–238 (2010).
28. Koob, G. F. & Volkow, N. D. Neurobiology of addiction: a neurocircuitry analysis. *Lancet Psychiatry* **3**, 760–773 (2016).
29. Fernandez, E., Schiappa, R., Girault, J. A. & Le Novere, N. DARPP-32 is a robust integrator of dopamine and glutamate signals. *PLoS Comput. Biol.* **2**, 1619–1633 (2006).
30. Yagishita, S. et al. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**, 1616–1620 (2014).
31. Zhu, H. W. et al. DARPP-32 phosphorylation opposes the behavioral effects of nicotine. *Biol. Psychiatry* **58**, 981–989 (2005).
32. Stoker, A. K. & Markou, A. Unraveling the neurobiology of nicotine dependence using genetically engineered mice. *Curr. Opin. Neurobiol.* **23**, 493–499 (2013).

33. Litten, R. Z. et al. A double-blind, placebo-controlled trial assessing the efficacy of varenicline tartrate for alcohol dependence. *J. Addiction Med.* **7**, 277–286 (2013).
34. Hyman, S. E., Malenka, R. C. & Nestler, E. J. Neural mechanisms of addiction: the role of reward-related learning and memory. *Annu. Rev. Neurosci.* **29**, 565–598 (2006).
35. Kalivas, P. W. The glutamate homeostasis hypothesis of addiction. *Nat. Rev. Neurosci.* **10**, 561–572 (2009).
36. Szumlinski, K. K. et al. Methamphetamine addiction vulnerability: the glutamate, the bad, and the ugly. *Biol. Psychiatry* **81**, 959–970 (2017).
37. Gass, J. T. & Olive, M. F. Glutamatergic substrates of drug addiction and alcoholism. *Biochem. Pharmacol.* **75**, 218–265 (2008).
38. Vaughan, J. et al. Urocortin, a mammalian neuropeptide related to fish urotensin I and to corticotropin-releasing factor. *Nature* **378**, 287–292 (1995).
39. Logrip, M. L., Koob, G. F. & Zorrilla, E. P. Role of corticotropin-releasing factor in drug addiction: potential for pharmacological intervention. *CNS Drugs* **25**, 271–287 (2011).
40. Volkow, N. D., Koob, G. F. & McLellan, A. T. Neurobiologic advances from the brain disease model of addiction. *N. Engl. J. Med.* **374**, 363–371 (2016).
41. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
42. Lassi, G. et al. The *CHRNA5–A3–B4* gene cluster and smoking: from discovery to therapeutics. *Trends Neurosci.* **39**, 851–861 (2016).
43. Edenberg, H. J. The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Res. Health* **30**, 5–13 (2007).
44. Sallese, M. et al. The G-protein-coupled receptor kinase GRK4 mediates homologous desensitization of metabotropic glutamate receptor 1. *FASEB J.* **14**, 2569–2580 (2000).
45. Perroy, J., Adam, L., Qanbar, R., Chenier, S. & Bouvier, M. Phosphorylation-independent desensitization of $GABA_B$ receptor by GRK4. *EMBO J.* **22**, 3816–3824 (2003).
46. Yang, J., Villar, V. M., Armando, I., Jose, P. A. & Zeng, C. Y. G. G protein–coupled receptor kinases: crucial regulators of blood pressure. *J. Am. Heart Assoc.* **5**, e003519 (2016).
47. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017). erratum **553**, 530 (2018).
48. Costas, J. The highly pleiotropic gene SLC39A8 as an opportunity to gain insight into the molecular pathogenesis of schizophrenia. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **177**, 274–283 (2018).
49. Kong, A. et al. The nature of nurture: effects of parental genotypes. *Science* **359**, 424–428 (2018).
50. Vrieze, S. I., Hicks, B. M., Iacono, W. G. & McGue, M. Decline in genetic influence on the co-occurrence of alcohol, marijuana, and nicotine dependence symptoms from age 14 to 29. *Am. J. Psychiatry* **169**, 1073–1081 (2012).

Acknowledgements

This study was designed and carried out by the GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN). It was conducted by using the UK Biobank Resource under application number 16651. This study was supported by funding from US National Institutes of Health awards R01DA037904 to S.V., R01HG008983 to D. J. Liu., and R21DA040177 to D. J. Liu. Ethical review and approval was provided by the University of Minnesota institutional review board; all human subjects provided informed consent. A full list of acknowledgements is provided in the Supplementary Note.

Author contributions

G.A., D.J.L., and S.V. designed the study. D.J.L. and S.V. led and oversaw the study. M. Liu was the study's lead analyst. She was assisted by Y.J., D.J.L., S.V., R.W., D.M.B., and G.D. Bonferroni thresholds were calculated by D.M. Phenotype definitions were developed by L.J.B., M.C.C., D.A.H., J.K., E.J., D.J.L., M.M., M.R.M., S.V., and L.Z. Software development was carried out by Y.J., D.J.L., and X.Z. Conditional analyses were performed by Y.J. and M. Liu. Heritability, genetic correlation, and polygenic scoring analyses were performed by R.W. Multivariate analyses were performed by Y.J., M. Liu, and D.J.L. Bioinformatics analyses were performed and interpreted by F. Chen, J.D., J.J.L., Y. Li, M. Liu, J. A. Stitzel, S.V., and R.W. The LocusZoom website was designed by G.D. Figures were created by M. Liu, R.W., Y. Li, and S.V. M.A.E. and M.C.K. helped with data access. R.W. coordinated authorship and acknowledgement details. M.C.C., S.P.D., E.J., J.K., and J. A. Stitzel provided helpful advice and feedback on study design and the manuscript. All authors contributed to and critically reviewed the manuscript. Y. Li, D.J.L., M. Liu, S.V., and R.W. made major contributions to the writing and editing.

Competing interests

L.J.B. and the spouse of N.L.S. are listed as inventors on issued US patent number 8,080,371, 'Markers for Addiction', covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction. S.P.D. is a scientific advisor to BaseHealth, Inc. G.B., D.F.G., G.W.R., H.S., K.S., and T.E.T. are employees of deCODE Genetics/Amgen, Inc. C.T. and D.H. are employees of 23andMe, Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0307-5.

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

Correspondence and requests for materials should be addressed to D.J.L. or S.V.

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

Mengzhen Liu[1,76], Yu Jiang[2,3,76], Robbee Wedow[4,5,6,76], Yue Li[7,8,76], David M. Brazel[4,9,10], Fang Chen[2,3], Gargi Datta[1], Jose Davila-Velderrain[7,8], Daniel McGuire[2,3], Chao Tian[11], Xiaowei Zhan[12,13], 23andMe Research Team[14], HUNT All-In Psychiatry[14], Hélène Choquet[15], Anna R. Docherty[16,17], Jessica D. Faul[18], Johanna R. Foerster[19], Lars G. Fritsche[19], Maiken Elvestad Gabrielsen[20], Scott D. Gordon[21], Jeffrey Haessler[22], Jouke-Jan Hottenga[23], Hongyan Huang[24,25], Seon-Kyeong Jang[1], Philip R. Jansen[26,27], Yueh Ling[2,9], Reedik Mägi[28], Nana Matoba[29], George McMahon[30], Antonella Mulas[31], Valeria Orrù[31], Teemu Palviainen[32], Anita Pandit[19], Gunnar W. Reginsson[33], Anne Heidi Skogholt[20], Jennifer A. Smith[18,34], Amy E. Taylor[30], Constance Turman[24,25], Gonneke Willemsen[23], Hannah Young[1], Kendra A. Young[35], Gregory J. M. Zajac[19], Wei Zhao[34], Wei Zhou[36], Gyda Bjornsdottir[33], Jason D. Boardman[4,5,6], Michael Boehnke[19], Dorret I. Boomsma[23], Chu Chen[22], Francesco Cucca[31], Gareth E. Davies[37], Charles B. Eaton[38], Marissa A. Ehringer[4,39], Tõnu Esko[8,28], Edoardo Fiorillo[31], Nathan A. Gillespie[16,21], Daniel F. Gudbjartsson[33,40], Toomas Haller[28], Kathleen Mullan Harris[41,42], Andrew C. Heath[43], John K. Hewitt[4,44], Ian B. Hickie[45], John E. Hokanson[35], Christian J. Hopfer[4,46], David J. Hunter[24,25,47], William G. Iacono[1], Eric O. Johnson[48], Yoichiro Kamatani[29], Sharon L. R. Kardia[34], Matthew C. Keller[4,44], Manolis Kellis[7,8], Charles Kooperberg[22], Peter Kraft[24,25,49], Kenneth S. Krauter[4,9], Markku Laakso[50,51], Penelope A. Lind[52], Anu Loukola[32], Sharon M. Lutz[53], Pamela A. F. Madden[43], Nicholas G. Martin[21], Matt McGue[1], Matthew B. McQueen[4,39], Sarah E. Medland[52], Andres Metspalu[28], Karen L. Mohlke[54], Jonas B. Nielsen[55], Yukinori Okada[29,56], Ulrike Peters[22,57], Tinca J. C. Polderman[26], Danielle Posthuma[26,58], Alexander P. Reiner[22,57], John P. Rice[59], Eric Rimm[25,60], Richard J. Rose[61], Valgerdur Runarsdottir[62], Michael C. Stallings[4,44], Alena Stančáková[50], Hreinn Stefansson[33], Khanh K. Thai[15], Hilary A. Tindle[63], Thorarinn Tyrfingsson[62], Tamara L. Wall[64], David R. Weir[18], Constance Weisner[15], John B. Whitfield[21], Bendik Slagsvold Winsvold[65], Jie Yin[15], Luisa Zuccolo[30,66], Laura J. Bierut[59], Kristian Hveem[20,67,68], James J. Lee[1], Marcus R. Munafò[66,69], Nancy L. Saccone[70], Cristen J. Willer[36,55,71], Marilyn C. Cornelis[72], Sean P. David[73], David A. Hinds[11], Eric Jorgenson[15], Jaakko Kaprio[32,74], Jerry A. Stitzel[4,39], Kari Stefansson[33,75], Thorgeir E. Thorgeirsson[33], Gonçalo Abecasis[19], Dajiang J. Liu[2,3,77*] and Scott Vrieze[1,77*]

[1]Department of Psychology, University of Minnesota Twin Cities, Minneapolis, MN, USA. [2]Department of Public Health Sciences, College of Medicine, Pennsylvania State University, Hershey, PA, USA. [3]Institute of Personalized Medicine, College of Medicine, Pennsylvania State University, Hershey, PA, USA. [4]Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, CO, USA. [5]Department of Sociology, University of Colorado Boulder, Boulder, CO, USA. [6]Institute of Behavioral Science, University of Colorado Boulder, Boulder, CO, USA. [7]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [9]Department of Molecular, Cellular, and Developmental Biology, University of Colorado Boulder, Boulder, CO, USA. [10]Interdisciplinary Quantitative Biology Graduate Group, University of Colorado Boulder, Boulder, CO, USA. [11]23andMe, Inc., Mountain View, CA, USA. [12]Quantitative Biomedical Research Center, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA. [13]Center for the Genetics of Host Defense, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA. [14]A full list of members and affiliations appears at the end of the paper. [15]Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA. [16]Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA. [17]Department of Psychiatry and Human Genetics, University of Utah, Salt Lake City, UT, USA. [18]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA. [19]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [20]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway. [21]Genetic Epidemiology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [22]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [23]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands. [24]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [25]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [26]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands. [27]Department of Child and Adolescent Psychiatry, Erasmus MC Rotterdam, Rotterdam, the Netherlands. [28]Estonian Genome Center, University of Tartu, Tartu, Estonia. [29]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama City, Japan. [30]Department of Population Health Science, Bristol Medical School, Oakfield Grove, Bristol, UK. [31]Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Monserrato, Italy. [32]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [33]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland. [34]Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA. [35]Department of Epidemiology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [36]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [37]Avera Institute for Human Genetics, Sioux Falls, SD, USA. [38]Department of Family Medicine and Community Health, Alpert Medical School, Brown University, Providence, RI,

USA. [39]Department of Integrative Physiology, University of Colorado Boulder, Boulder, CO, USA. [40]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [41]Department of Sociology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [42]Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [43]Department of Psychiatry, Washington University in St. Louis, St. Louis, MO, USA. [44]Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, USA. [45]Brain and Mind Centre, University of Sydney, Sydney, New South Wales, Australia. [46]Department of Psychiatry, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [47]Nuffield Department of Population Health, University of Oxford, Oxford, UK. [48]Fellows Program, RTI International, Research Triangle Park, NC, USA. [49]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [50]Department of Internal Medicine, Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland. [51]Department of Medicine, Kuopio University Hospital, Kuopio, Finland. [52]Psychiatric Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [53]Department of Biostatistics and Bioinformatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [54]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [55]Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI, USA. [56]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan. [57]Department of Epidemiology, University of Washington, Seattle, WA, USA. [58]Department of Clinical Genetics, VU Medical Centre Amsterdam, Amsterdam, the Netherlands. [59]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA. [60]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [61]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA. [62]SAA—National Center of Addiction Medicine, Vogur Hospital, Reykjavik, Iceland. [63]Department of Medicine, Vanderbilt University, Nashville, TN, USA. [64]Department of Psychiatry, University of California, San Diego, San Diego, CA, USA. [65]FORMI and Department of Neurology, Oslo University Hospital, Oslo, Norway. [66]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [67]HUNT Research Centre, Department of Public Health and Nursing, Norwegian University of Science and Technology, Levanger, Norway. [68]Department of Medicine, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway. [69]UK Centre for Tobacco and Alcohol Studies, School of Psychological Science, University of Bristol, Bristol, UK. [70]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. [71]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [72]Department of Preventative Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [73]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [74]Department of Public Health, University of Helsinki, Helsinki, Finland. [75]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [76]These authors contributed equally: Mengzhen Liu, Yu Jiang, Robbee Wedow, Yue Li. [77]These authors jointly supervised this work: Dajiang Liu, Scott Vrieze. *e-mail: dajiang.liu@psu.edu; vrieze@umn.edu

23andMe Research Team

Michelle Agee[11], Babak Alipanahi[11], Adam Auton[11], Robert K. Bell[11], Katarzyna Bryc[11], Sarah L. Elson[11], Pierre Fontanillas[11], Nicholas A. Furlotte[11], David A. Hinds[11], Bethann S. Hromatka[11], Karen E. Huber[11], Aaron Kleinman[11], Nadia K. Litterman[11], Matthew H. McIntyre[11], Joanna L. Mountain[11], Carrie A. M. Northover[11], J. Fah Sathirapongsasuti[11], Olga V. Sazonova[11], Janie F. Shelton[11], Suyash Shringarpure[11], Chao Tian[11], Joyce Y. Tung[11], Vladimir Vacic[11], Catherine H. Wilson[11] and Steven J. Pitts[11]

HUNT All-In Psychiatry

Amy Mitchell[65], Anne Heidi Skogholt[20], Bendik S. Winsvold[65,78], Børge Sivertsen[79,80,81], Eystein Stordal[80,82], Gunnar Morken[80,83], Håvard Kallestad[80,83], Ingrid Heuch[81], John-Anker Zwart[65,78,84], Katrine Kveli Fjukstad[85,86], Linda M. Pedersen[65], Maiken Elvestad Gabrielsen[20], Marianne Bakke Johnsen[65,84], Marit Skrove[87], Marit Sæbø Indredavik[80,87], Ole Kristian Drange[80,83], Ottar Bjerkeset[80,88], Sigrid Børte[65,84] and Synne Øien Stensland[65,89]

[78]Department of Neurology, Oslo University Hospital, Oslo, Norway. [79]Department of Health Promotion, Norwegian Institute of Public Health, Bergen, Norway. [80]Department of Mental Health, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway. [81]Department of Research and Innovation, Helse-Fonna HF, Haugesund, Norway. [82]Department of Psychiatry, Hospital Namsos, Nord-Trøndelag Health Trust, Namsos, Norway. [83]Division of Mental Health Care, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [84]Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [85]Department of Psychiatry, Nord-Trøndelag Hospital Trust, Levanger Hospital, Levanger, Norway. [86]Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology, Trondheim, Norway. [87]Regional Centre for Child and Youth Mental Health and Child Welfare, Department of Mental Health, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway. [88]Faculty of Nursing and Health Sciences, Nord University, Levanger, Norway. [89]Norwegian Centre for Violence and Traumatic Stress Studies, Oslo, Norway.

Methods

Generation of summary statistics. Participants in all studies were genotyped on genome-wide arrays. The majority of studies imputed their genotypes to the Haplotype Reference Consortium[41] using the University of Michigan Imputation Server (see URLs)[51]. Several studies did not impute using the imputation server, due to data sharing restrictions, computational limitations, and/or resource limitations (described in the Supplementary Note). All studies used either Minimac3[51] or IMPUTE2[52] for imputation.

GWAS summary statistics were generated in each study sample using RVTESTS[53] according to a standard analysis plan. Studies composed primarily of classically related individuals (for example, family studies) first regressed out covariates including genetic principal components under a linear model, inverse-normalized the residuals (except for 23andMe), and tested for an additive effect of each variant under a linear mixed model with a genetic kinship matrix. Family studies followed this analysis for all phenotypes, even binary phenotypes such as smoking initiation and cessation. Studies of entirely classically unrelated individuals followed the same analysis for quasi-continuous phenotypes (AgeSmk, CigDay, DrnkWk), but estimated additive genetic effects under a logistic model for binary phenotypes (SmkInit and SmkCes).

Quality control checks were applied to ensure quality of both the phenotypes and the genotypes. For each phenotype and covariate, distribution statistics including the minimum, maximum, quartiles, median, mean, and standard deviation were examined. We ensured that these statistics were within expected limits given the phenotype definitions and any scale transformations per the analysis plan. We also evaluated simple relationships among phenotypes. When discrepancies were noted, we contacted the original study for clarification or re-analysis, or the data were removed from further analysis. Phenotypic statistics are presented in Supplementary Tables 6 and 7.

Extensive genetic quality control and filtering were performed on the contributed summary statistics from each cohort. We removed imputed variants with imputation quality less than 0.3 (the estimated squared correlation between the imputed dosage and true dosage). We compared the per-study allele labels and allele frequencies with those of the imputation reference panels and removed or reconciled mismatches. For quantitative traits, we plotted the variance of the score statistics against the sample size and tested whether the trait residuals in each study were properly normalized and whether the trait analyzed between studies was measured and analyzed using the same unit.

Meta-analysis. Meta-analysis was performed centrally using the software package rareGWAMA (see URLs). All statistical tests in the meta-analysis or secondary analyses of the meta-analytic results (for example, PRS, functional enrichment, MTAG, GenomicSEM) were two-sided. Given that rarer variants and/or behavioral phenotypes may show between-study heterogeneity in allele frequencies, imputation qualities, or genetic architecture, we extended existing methods and developed a novel fixed effects approach that accounts for between-study heterogeneity. Specifically, the methods aggregated weighted Z-score statistics, that is, $Z_{\mathrm{META}} = \frac{\sum_k w_k Z_k}{\left(\sum_k w_k^2\right)^{1/2}}$, where Z_k is the Z-score statistic in study k. The weight w_k is defined by $w_k = N_k p_k (1-p_k) R_k^2$, where p_k is the variant allele frequency, R_k^2 is the imputation quality, and N_k is the sample size for study k. Under the null and with the present sample sizes, Z_{META} is normally distributed. The weights are proportional to the sample genotype variance. When the trait is uniformly measured and the allele frequencies are similar, the method is approximately equivalent to meta-analysis of sample-size-weighted Z scores. Yet the method accounts for between-study heterogeneity in imputation accuracy and allele frequencies. The use of a fixed effects model, the most common approach in GWAS meta-analysis of single-ancestry groups, appeared acceptable given the apparent lack of substantial meta-analytic effect heterogeneity (see Cochrane's Q and I^2 statistics in Supplementary Tables 1–5).

Population stratification and cryptic relatedness were addressed during the generation of summary statistics by each local study through the use of kinship-based linear mixed models[54] and genetic principal components[55]. Residual stratification was further corrected at the meta-analytic level with study-specific genomic controls[56] (calculated separately for variants with MAF ≥ 1% and 0.1% ≤ MAF < 1%; Supplementary Table 23) applied to each study's results prior to meta-analysis.

A locus was defined as a 1-Mb region surrounding the 'sentinel' variant (the variant in the locus with the lowest P value). When any two such loci overlapped or abutted, they were collapsed into a single locus. Variants within each locus were subjected to conditional analysis using a novel partial correlation-based score statistic using cohort-level summary statistics[57] implemented in a sequential forward selection framework. The method requires marginal association statistics and approximated covariance matrices among them and performs favorably compared with existing methods[57] (Supplementary Table 24). Covariances among effects were based upon the linkage disequilibrium information estimated from a subset of the Haplotype Reference Consortium[41].

We applied multiple post-meta-analysis variant filters to ensure robustness of reported findings. To reduce artifacts arising from a small number of studies, we excluded any variant that was present in only two or fewer studies. For each variant in the meta-analysis, we calculated the effective sample size $N_{\mathrm{eff}} = \sum_k N_k r_k^2$, where N_k is the sample size in study k and r_k^2 is the imputation quality. We removed variants with effective sample sizes <10% of the total sample size to ensure only well-imputed variants with a modicum of power were included. We also excluded all variants with MAF > 0.001, the lower bound of moderate imputation accuracy with the current best available imputation reference panel[41]. Variants with MAF > 1% are expected to be imputed with high accuracy. Results from the application of post-meta-analysis filters are displayed in Supplementary Table 25.

After applying variant filters and obtaining our final meta-analytic results, we calculated genomic controls and maximum/median per-variant sample sizes. Sample sizes ranged from 337,334 for cigarettes per day to 1,232,091 for smoking initiation. Quantile–quantile plots, LD intercept tests, and genomic control values indicate that Type I error rates were well controlled for common and low-frequency variants (Supplementary Fig. 2 and Supplementary Table 26). All conditionally independent variants were plotted in LocusZoom and included in Supplementary Figs. 1–12. All plots were visually inspected, and suspicious loci were identified (see Supplementary Table 27) and removed from further consideration. To ensure LD information was available between sentinel variants and others in the locus, we used surrogate variants for eight loci (Supplementary Table 28).

We estimated the extent of pleiotropy for each genome-wide associated locus from our GWAS using an empirical Bayes approach (that is, whether a given locus is simultaneously associated with multiple phenotypes). Using summary association statistics from a given locus as input, the method estimated the 5 × 5 genetic correlation of the locus and the posterior probability of association for all possible phenotype configurations, while accounting for genome-wide genetic correlations and trait residual correlations. In cases in which loci associated with different phenotypes overlapped, the locus was expanded in size. Statistical details are available in Section 3.3 of the Supplementary Note.

We applied MTAG[16] to variants with MAF > 1% from the final meta-analysis results for each phenotype, using the other four phenotypes to increase power for locus discovery. Genomic controls and LD intercept tests of the MTAG results were well controlled (Supplementary Table 29), and Manhattan and quantile–quantile plots were well behaved (Supplementary Figs. 16 and 17). GCTA-COJO[58] was used to identify conditionally independent variants (listed in Supplementary Table 12). All loci were plotted with LocusZoom and visually inspected, with suspicious loci identified (for example, those without LD support; see Supplementary Table 30) and removed from further consideration. Additional details, including testing of MTAG model assumptions, are provided in the Supplementary Note. Finally, we also applied GenomicSEM[59] to our five phenotypes to formally model and factor their correlation structure. See Supplementary Fig. 18, Supplementary Table 31, and the Supplementary Note for further details.

Genome-wide significance threshold. The primary focus was to test variants with MAF ≥ 1%, as these will be imputed with high confidence. The statistical significance threshold applied to meta-analysis of all variants with MAF ≥ 1% was 5×10^{-8}, consistent with widespread convention in GWAS of European individuals. Since our imputation procedure is expected to provide some marginal level of accuracy down to MAF of 0.1%, we also conducted an exploratory association test for low-frequency variants with 0.1% < MAF < 1%, to which we applied a statistical significance threshold of $P < 5 \times 10^{-9}$. Only two such low-frequency variants surpassed the conventional common variant threshold of $P < 5 \times 10^{-8}$. Of these two, one low-frequency variant, associated with SmkInit, survived the more stringent multiple testing correction (rs181508347, intergenic, MAF = 0.0096, $P = 5 \times 10^{-10}$), and it is included in our count of discovered loci and listed in Supplementary Table 4. The more stringent threshold applies a correction for ~10 million tests, which is approximately the number of conditionally independent variants tested once the MAF lower bound was extended from 1% to 0.1%. We calculated this threshold using three existing methods[60–62]. These methods make use of the eigenvalues of the matrix of LD (measured in R^2) between SNPs, calculated with a spectral decomposition. We estimated the number of independent tests using the genotype data from a subset of the Haplotype Reference Consortium panel[41]. We first calculated LD blocks across the genome using the algorithm implemented in PLINK v.1.9[63] with default settings, and then we lowered the MAF threshold to 0.1% to accommodate all low-frequency variants. Next, we calculated the effective number of independent tests within each LD block and between LD blocks using the aforementioned three methods, which we aggregated to get the total number of independent tests. The three techniques estimated the number of independent variants at 9.8 million to 10.1 million independent tests, similar to other independent estimates[64]. A total of 278 sentinel variants (including the one genome-wide significant low-frequency variant) had $P < 5 \times 10^{-9}$, out of the original 406 with $P < 5 \times 10^{-8}$.

Heritability. We used univariate and bivariate LD score regression[17] to assess the heritability of each phenotype and to estimate a variety of genetic correlations. Analyses included (1) LD score regression intercept tests to evaluate the extent to which population stratification or cryptic relatedness may artificially inflate our summary statistics; (2) estimation of genetic correlations across our five phenotypes; (3) estimation of genetic correlations computed within a phenotype

but between the larger contributing studies, as an estimate of the extent to which phenotypes were measuring the same genetic risk in different studies; and (4) estimation of genetic correlation between the five phenotypes and a wide variety of other phenotypes related to smoking and alcohol behaviors, and for which GWAS have already been made publicly available.

Under standard assumptions, bivariate score regression produces unbiased estimates of genetic correlation, even in the presence of sample overlap[65]. Accordingly, to estimate the extent of genetic correlation between each of our phenotypes, and between our phenotypes and other phenotypes related to nicotine and alcohol use, we used standard procedures in LD score regression[22]. To be included in these analyses, variants were restricted to those present in HapMap3 with MAF>0.01. Standard errors were estimated with a block jackknife over all variants.

We estimated the proportion of variance explained by the set of all conditionally independently associated variants. The joint effects of variants in a locus were approximated by $\hat{\vec{\beta}}_{\mathrm{JOINT}} = \mathbf{V}_{\mathbf{META}}^{-1}\vec{U}_{\mathrm{META}}$, where $\vec{U}_{\mathrm{META}}$ is the single variant score statistics and $\mathbf{V}_{\mathbf{META}}$ is the covariance matrix between them. The phenotypic variance explained by the independently associated variants in a locus is given by $\hat{\vec{\beta}}^{T}_{\mathrm{joint}}\mathrm{cov(G)}\,\hat{\vec{\beta}}_{\mathrm{JOINT}}$, where cov(G) is the genotype covariance estimated from the Haplotype Reference Consortium panel.

Polygenic scoring. PRS were computed using LDpred[66], which accounts for linkage disequilibrium between variants. Since we do not know the variance–covariance matrix of the effects in the training sample (here, the GWAS results), we replace this matrix with a block diagonal matrix estimated using LD patterns from the prediction cohorts, after dropping cryptically related individuals and ancestry outliers.

Smoking and alcohol use rates are influenced by secular trends and policy changes over the past half-century. We therefore selected two independent prediction cohorts, the HRS[21] and Add Health[20]. The HRS is a nationally representative study of US households that began in 1992; the mean birth year of respondents is 1938 (s.d. = 9.3), and the mean age at the time of assessment is 57.6 (s.d. = 8.9). Add Health is a nationally representative sample of US adolescents enrolled in grades 7 through 12 during the 1994–1995 school year. The mean birth year of respondents was 1979 (s.d. = 1.8), and the mean age at assessment (here, wave 4) was 29.0 (s.d. = 1.8). In the HRS, ~57% of respondents reported ever smoking regularly, and these respondents smoked ~13 cigarettes per day. In Add Health, slightly fewer (~53%) of respondents reported ever smoking regularly, and these respondents smoked ~11 cigarettes per day on average (Supplementary Table 14). For each of our five phenotype scores, we used variants that overlapped with HapMap3 (~1.1 million) to construct the scores. Prediction accuracy was estimated using ordinary least squares regression of a given phenotype (AgeSmk, CigDay, SmkInit, SmkCes, or DrnkWk) on the polygenic score and covariates including age, sex, age × sex interaction, and the first ten genetic principal components.

Prediction accuracy comes from a two-step process in which we first regress the phenotype on a standard set of covariates without including the PRS. Then, the PRS predictor is added, and the difference in the coefficient of determination (R^2) is calculated. For our quantitative phenotypes, AgeSmk, CigDay, and DrnkWk, the predictive power of the PRS is the change in the R^2 in going from the regression without the PRS to the regression with the PRS. For our two binary phenotypes, SmkInit and SmkCes, we measure the incremental pseudo-R^2 from probit regressions. 95% confidence intervals around all R^2 values are bootstrapped with 1,000 repetitions each. The same polygenic scoring procedure was applied to the MTAG results (Supplementary Table 32).

Epigenomic enrichment. To detect genome-wide functional and tissue-specific epigenomic enrichments, we performed enrichment analyses by heritability stratification using LD score regression, implemented in the LDSC v.1.0.0 software. Annotation-stratified LD scores were estimated using dichotomized/binary annotations, 1000 Genomes Project samples with European ancestry, and 1 million–base pair LD windows by default. LDSC then determines functional enrichment of the GWAS traits by partitioning heritability according to the variance explained by the LD-linked SNPs belonging to each functional category[22]. Statistical enrichment was defined as the ratio between the percentage of heritability explained by variants in each annotated category and the percentage of variants covered by that category. A resampling approach was used to estimate standard errors[22].

Following standard procedure, we trained a baseline LDSC model using the 52 non-cell-type-specific functional categories (plus one category that includes all SNPs) and used the observed Z-scores of HapMap3 SNPs for each trait. We tested cell-group enrichments over 10 predefined cell-group annotations[22]. The cell-group annotations are the result of aggregating 220 cell-type-specific annotations over 4 histone marks (H3K4me1, H3K4me3, H3K9ac, H3K27ac) and 100 well-defined cell types. To detect which specific epigenomes contribute to the group-level enrichment, we performed 220 tests over each individual annotation. Multiple testing was accounted for through Bonferroni correction within phenotype with 10 tests for the cell-group annotation enrichment analyses and 220 tests for the cell-specific enrichment analyses. As a complementary method to LDSC, we also applied a recently developed mixture model learning approach[67], and we report these results in Supplementary Fig. 13.

Gene and gene-set tests. For each phenotype, we used SEQMINER[68] and the University of California, Santa Cruz genome browser annotations (refGene; retrieved 15 December 2017) to annotate all conditionally independent genome-wide significant variants. We identified all genes (all variants 5′ to 3′ UTR) harboring at least one variant within LD $r^2>0.3$ with any conditionally independent variant. See Supplementary Tables 1–5.

We conducted a manual review of all genes implicated within each locus, overlap with the GWAS catalog (Supplementary Table 33), and all pathways identified by PASCAL and DEPICT (described below). We considered a gene to be implicated if it harbored variation in LD with a conditionally independent genome-wide significant variant, or if a gene was located within the locus and was significant by the PASCAL gene-based test. PASCAL[69] was used for gene-based and pathway analysis to test genes and canonical pathways from MSigDb (Supplementary Tables 20 and 21). Default settings were used to test all variants within all genes. DEPICT[70] was used to identify enrichment within tissues/cell types and reconstituted gene sets (also known as pathways). For each phenotype, variants from the GWAS were clumped using 500-kilobase flanking regions with the LD cutoff $r^2>0.1$ (based on 1000 Genomes phase 1 release v.3, the default in DEPICT). We used DEPICT to understand genetic signals beyond the genome-wide significant loci that surpass the conventional 5×10^{-8}, and so included all variants with $P<5\times10^{-5}$. DEPICT tissue enrichment results are displayed in Supplementary Fig. 15, where enrichment relative to genes in random sets of loci is indicated by red shading. To cluster DEPICT reconstituted gene sets, we used affinity propagation clustering[71] and calculated the correlation between each resulting 'exemplary gene set' in Fig. 4. Genes, gene sets, and tissue/cell enrichments were considered significant when their false discovery rate was below 0.05. All such significant DEPICT results are reported in Supplementary Tables 17–19. PASCAL and DEPICT were also applied in the same fashion to the MTAG summary statistics (Supplementary Tables 34–39).

Statistics. The GWAS meta-analysis was conducted using χ^2 statistics based upon an imputation-quality-aware fixed-effect meta-analysis approach. Two-sided P values were calculated. The MTAG and GenomicSEM analysis test statistics were determined using the GWAS meta-analysis results, and two-sided P values were similarly calculated from the χ^2 distribution. The pleiotropic analysis was conducted based upon an empirical Bayes approach. The prior distributions for the effect sizes were assumed to follow a mixture distribution, with a point mass at zero (representing the possibility that the locus is not associated with the trait) and a normal distribution (representing the possibility that the locus is associated). The hyperparameters were estimated by maximizing the marginal likelihood. The method properly accounts for the local genetic correlation and residual correlation between phenotypes. The posterior probability of association (PPA) for each locus was estimated for each possible combination of five phenotypes, and the combination with the highest PPA was reported for each locus.

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

Code availability

All software used to perform these analyses is available online.

Data availability

GWAS summary statistics can be downloaded online (https://genome.psych.umn.edu/index.php/GSCAN). We provide association results for all SNPs that passed quality-control filters in a GWAS meta-analysis of each of our five substance use phenotypes that excludes the research participants from 23andMe.

References

51. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
52. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
53. Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423–1426 (2016).
54. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
55. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
56. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
57. Jiang, Y. et al. Proper conditional analysis in the presence of missing data identified novel independently associated low frequency variants in nicotine dependence genes. *PLoS Genet.* **14**, e1007452 (2018).

58. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, S1–S3 (2012).
59. Grotzinger, A. D. et al. Genomic sem provides insights into the multivariate genetic architecture of complex traits. Preprint at https://doi.org/10.1101/305029 (2018).
60. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221–227 (2005).
61. Gao, X. Y., Becker, L. C., Becker, D. M., Starmer, J. D. & Province, M. A. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet. Epidemiol.* **34**, 100–105 (2010).
62. Chen, Z. X. & Liu, Q. Z. a new approach to account for the correlations among single nucleotide polymorphisms in genome-wide association studies. *Hum. Hered.* **72**, 1–9 (2011).
63. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
64. Wu, Y., Zheng, Z. L., Visscher, P. M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* **18**, 86 (2017).
65. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
66. Vilhjalmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
67. Li, Y., Davila-Velderrain, J. & Kellis, M. A probabilistic framework to dissect functional cell-type-specific regulatory elements and risk loci underlying the genetics of complex traits. Preprint at https://doi.org/10.1101/059345 (2017).
68. Zhan, X. & Liu, D. J. SEQMINER: an R-package to facilitate the functional interpretation of sequence-based associations. *Genet. Epidemiol.* **39**, 619–623 (2015).
69. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol.* **12**, e1004714 (2016).
70. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
71. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).

Corresponding author(s): Dajiang Liu
Scott Vrieze

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 Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	All studies used either ShapeIt2, EAGLE or Finch to phase genotypes and used either Minimac3 or IMPUTE2V3 for imputation. Summary statistics were generated using RVTESTS release v1.9.7 or v1.9.9 or BOLT-LMM v2.2. Meta-analysis and conditional analysis was performed using rareGWAMA_0.4 in R. LD Score Regression v1.0.0 was used to measure heritability, test for population stratification and cryptic relatedness, estimate genetic correlations and enrichment analyses. RiVIERA-ridge was also used for enrichment analyses. LDpred v0.9.09 was used to construct the polygenic scores. PASCAL was used for gene based and pathway analysis and DEPICT was used to identify enrichment within tissues/cell types and reconstituted gene sets. Locuszoom plots were made using LocusZoom standalone software v1.3. GenomicSEM was used for the Genomic SEM analyses. MTAG software was used for the MTAG analysis.

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/reviewers upon request. 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

Upon acceptance, results excluding the 23andMe substudy will be available from the GSCAN Wiki page (https://genome.psych.umn.edu/), and posted on dbGaP. The 23andMe substudy itself is available upon request to 23andMe.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	No sample size calculation was done but we tried to increase our sample size as much as possible. We contacted as many studies (with our phenotypes of interest) as possible and applied for relevant studies available in public repositories. Our meta-analysis includes the largest sample size of similar phenotypes to date and therefore, we believe our results are sufficiently powered.
Data exclusions	We excluded any non-European sample as population differences may lead to spurious results. We also excluded results for some phenotypes from smaller studies when those results were severely inflated or deflated per the genomic control, and there was no alternative explanation (e.g., inflation was due to polygenic signal). We applied filters to the genomic data post meta-analysis (minor allele frequency > .1%, effective sample size of at least 10% per phenotype and at least 3 studies must be included for each variant) in order to only report variants on which we had robust results.
Replication	Our results have replicated 26/27 previous known loci as detailed in the manuscript. In order to maximize power to detect the variants, we did not separate our sample into a separate discovery and replication set.
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	European ancestry with 52.2% female.
Recruitment	We did not do any recruitment. Analysis was of existing de-identified data.