Investigating heterogeneous genetic effects contributing to smoking behaviors

Scott A Funkhouser[1*], Jason D Boardman[2,3], John K Hewitt[1,4], Michael C Stallings[1,4], Christian J Hopfer[1,5], Sandra A Brown[6,7], Tamara L Wall[6], Chandra A Reynolds[8], Matthew C Keller[1,4], Luke M Evans[1,9]

[1] Institute for Behavioral Genetics, University of Colorado, Boulder, Colorado, United States

[2] Department of Sociology, University of Colorado, Boulder, Colorado, United States

[3] Institute for Behavioral Science, University of Colorado, Boulder, Colorado, United States

[4] Department of Psychology and Neuroscience, University of Colorado, Boulder, Colorado, United States

[5] Department of Psychiatry, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States

[6] Department of Psychiatry, University of California San Diego, La Jolla, California, United States

[7] Department of Psychology, University of California San Diego, La Jolla, CA, United States

[8] Department of Psychology, University of California Riverside, Riverside, CA, United States

[9] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, United States

[*] Corresponding Authors:

E-mail: Scott.Funkhouser@colorado.edu, luke.m.evans@colorado.edu

Abstract

The genetic architecture of numerous smoking behaviors is highly polygenic, but these genetic effects are heterogeneous and depend on moderating factors. Here, we used common SNPs from the UK Biobank to investigate heterogeneous genetic effects for smoking heaviness using cigarettes per day (CPD) records, smoking initiation (SI), and smoking cessation (SC). We assessed heterogeneous effects across levels of sex, age of smoking initiation, major depressive disorder (MDD) DSM-V-like diagnosis, generalized anxiety disorder (GAD) DSM-V-like diagnosis, and whether an individual has seen a psychiatrist for nerves, anxiety, tension, or depression. We observed suggestive evidence of heterogeneous genetic effects for CPD and SC, moderated by MDD and GAD, respectively [CPD $\hat{r}_g$ = 0.69 (SE = 0.15) between MDD cases and controls, and SC $\hat{r}_g$ = 0.38 (SE = 0.28) between GAD cases and controls, $p < 0.05$]. We detected 5 SNPs with genome-wide significant evidence of heterogeneous effects moderated by either MDD or GAD (p-value < $5\text{x}10^{-8}$) for CPD and SC. We also observed strong evidence for heterogeneous genetic effects for SI between sexes (between-sex $\hat{r}_g$ = 0.82, SE = 0.02). While we detected no individual SNPs that were moderated by sex at genome-wide significance (all p-value > $5\text{x}10^{-8}$), we observed evidence of novel genome-wide significant SI-SNP associations using sex-stratified GWAS; six loci were discovered in either men or women separately that were not identified in a previous smoking meta-analysis that had a 6-fold larger, sex-combined sample. Furthermore, using several independent testing samples, there was suggestive evidence that the prediction ability of a polygenic risk score (PRS) for smoking initiation improved through the utilization of sex-specific SNP effects. This work suggests that a more nuanced approach to GWAS analyses is warranted, as potential heterogeneous effects can complicate variant discovery and polygenic risk score accuracy.

Author Summary

Smoking imposes a heavy health burden and is highly polygenic in architecture. Consistent with most complex traits, many causal loci have yet to be identified, even when using the largest available samples. One possible reason for the difficulty in inferring genetic variants associated with such complex traits is that common genetic variants possess context-dependent (or heterogeneous) effects. Utilizing the UK Biobank, we find evidence for heterogeneous SNP effects on smoking initiation, heaviness, and cessation among psychiatric disorder cases and controls and between sexes. Failure to model such heterogeneity (when accounting for sample size) resulted in lower independent sample predictive ability. This work encourages a more nuanced approach to GWAS and polygenic risk prediction. The assumption that all genetic effects are homogeneous limits our understanding of complex traits when heterogeneous effects are present.

Introduction

Tobacco smoking has contributed to more than 20 million preventable deaths since 1964[1] and disproportionally affects certain groups within populations. Individuals with depression and anxiety are at an increased risk of becoming nicotine dependent[2–4] and early-onset smokers are more likely to smoke heavily compared to late-onset smokers[5]. Furthermore, sex differences in smoking behaviors are well-documented, with a consensus showing men exhibit greater rates of smoking initiation, cessation, and tobacco usage than women[6–8]. While numerous environmental factors likely contribute to tobacco's heterogeneous usage, smoking behaviors are known to be heritable[9,10] with a highly polygenic architecture comprised of

many loci with very small effects[11–15]. Family studies have suggested that the genetic component to smoking is itself heterogeneous[16–19], *i.e.* dependent on certain factors that moderate one's genetic risk to smoke, smoke heavily, or persist in smoking.

Heterogeneous genetic effects could complicate efforts to identify genetic loci that influence smoking behaviors. While many genome-wide significant (GWS; p-value < $5x10^{-8}$) SNPs have been identified, their individual effects are small and collectively explain only about one-third of the SNP-heritability[20]. If many of the genetic effects for smoking are heterogeneous, this could partially explain i) differing tobacco usage across groups and ii) the exceedingly small size of average SNP effects when modeled as having a single effect across all groups or conditions. This second point may be especially true when the effect of an allele takes place only in a rare group (*e.g.,* psychiatric disorder cases), where the average effect of an allele substitution (*i.e.*, averaged across a random sample of psychiatric disorder cases and controls) would be weighted toward zero. Prior evidence from a Japanese population suggests sex-dependent SNP effects for smoking behaviors[13], but few studies have utilized genome-wide SNP data to infer heterogeneous genetic effects for smoking behaviors across potential moderating factors. Other candidate gene studies have implicated individual loci with heterogeneous smoking effects[21], but these findings have not replicated in biobank-scale data[22].

In this study, we estimated heterogeneous genetic effects for smoking heaviness [*i.e.*, cigarettes per day (CPD)], smoking cessation (SC), and smoking initiation (SI) using the UK Biobank[23]. Heterogeneity of effects was assessed across levels of sex, age of smoking initiation (ASI), major depressive disorder (MDD) DSM-V-like diagnosis, generalized anxiety disorder (GAD) DSM-V-like diagnosis, and whether an individual had seen a psychiatrist for

nerves, anxiety, tension, or depression (PSYCH). We evaluated evidence for heterogeneous genetic effects at multiple scales: genome-wide, within functional annotations, and at single SNPs. Specifically, for each smoking trait we i) estimated the genetic correlation between groups (*e.g.*, between MDD cases and controls) and the proportion of phenotypic variance explained by heterogeneous SNP effects genome-wide, ii) estimated differences in heritability enrichment among cell-type-specific annotations, and iii) conducted a GWAS to test for heterogeneous effects at individual SNPs. Given clear genome-wide evidence for differential genetic architectures for SI between sexes (as evident from a between-sex $\hat{r}_g$ = 0.82 SE = 0.02), we then evaluated polygenic risk score (PRS) accuracy for SI when allowing for SNP effects to be heterogeneous, as opposed to homogenous, between sexes.

Results

Clear genome-wide evidence of heterogenous SNP effects for SI between sexes

Using common (MAF > 0.01) genome-wide SNPs, we modeled each smoking trait (CPD, SI, and SC) across levels of potential moderators (*e.g.*, across males and females or MDD-like cases and controls) using a bivariate GREML model (see Methods). CPD is often transformed to different scales prior to analysis[16,21]; because SNP effect heterogeneity may depend on the chosen scale, we considered four different transformations of CPD for all analyses: the raw scale, binned, log transformed, and dichotomized (see Methods for more details). For all trait-by-moderator combinations, we restricted our analyses to unrelated individuals (estimated relatedness <0.05, sample sizes in Table 1).

From the bivariate model we estimated the genetic correlation (r_g) between strata to infer the presence of heterogeneous genetic effects that show disproportionality between strata (Table

115 1). Using a likelihood ratio test under the null hypothesis of $r_g = 1$, we found limited evidence of
116 different genetic effects across most moderators. However, we found strong evidence that
117 genetic effects for SI differed between sexes ($\hat{r}_g = 0.82$, SE = 0.02; p-value = 4.22×10^{-9}), and
118 nominally significant evidence that genetic effects for log-transformed CPD differed between
119 MDD-like cases and controls ($\hat{r}_g = 0.69$, SE = 0.15; p-value = 0.047). We also found suggestive
120 evidence that genetic effects for SC differed between GAD-like cases and controls, but due to the
121 relatively small number of GAD cases, the standard error was quite large ($\hat{r}_g = 0.38$, SE = 0.28;
122 p-value = 0.078). Estimates of genetic correlations using an alternative method, cross-trait
123 LDSC[24], were consistent with GREML-based estimates (Supplementary Table S1).
124 We next decomposed the total variance across strata to estimate the proportion of
125 phenotypic variance explained by heterogeneous effects [denoted PVE(σ^2_{het})] using a GREML
126 interaction model[25] (Table 1). For SI, we estimated sex-dependent effects to account for 3.7%
127 of the liability scale phenotypic variance (SE = 0.4%; p-value = 3.44×10^{-13}). Likewise, for
128 binned CPD, we estimated MDD-dependent genetic effects to account for 5.2% of the total
129 variance but with larger uncertainty (SE = 2.3%; p-value = 0.007).
130
131
132
133
134
135
136
137

Table 1 Genome-wide estimates of heterogeneous genetic effects

Trait[a]	Moderator[b]	N_1[c]	N_2	r_g[d]		PVE(σ^2_{het})[e]		LRT p-value[f]	
				Estimate	SE	Estimate	SE	$H_0: r_g = 1$	$H_0: \sigma^2_{het} = 0$
CPD (binned)	ASI	58,767	57,893	1	0.06	0.0024	0.005	0.5	0.282
CPD (binned)	GAD	3,760	32,390	0.98	0.34	$1x10^{-6}$	0.026	0.477	0.5
CPD (binned)	MDD	9,679	23,400	0.69	0.16	0.0518	0.023	0.05	0.007
CPD (binned)	PSYCH	17,219	101,007	0.85	0.08	0.0106	0.008	0.056	0.088
CPD (binned)	SEX	62,062	56,585	1	0.05	$1x10^{-6}$	0.006	0.5	0.5
CPD (dichotomized)	ASI	28,237	26,794	1	0.07	0.0213	0.017	0.5	0.06
CPD (dichotomized)	GAD	1,800	15,414	1	2.9	0.146	0.089	0.5	0.038
CPD (dichotomized)	MDD	4,525	11,127	0.85	0.28	0.107	0.077	0.315	0.073
CPD (dichotomized)	PSYCH	8,402	47,666	0.97	0.11	0.0005	0.023	0.386	0.492
CPD (dichotomized)	SEX	29,744	26,536	0.99	0.07	0.0178	0.019	0.425	0.178
CPD (log)	ASI	58,767	57,893	1	0.05	0.006	0.005	0.5	0.082
CPD (log)	GAD	3,760	32,390	1	0.42	$1x10^{-6}$	0.026	0.5	0.5
CPD (log)	MDD	9,679	23,400	0.69	0.15	0.0402	0.023	0.047	0.032
CPD (log)	PSYCH	17,219	101,007	0.94	0.09	0.0024	0.008	0.255	0.376
CPD (log)	SEX	62,062	56,585	0.98	0.05	$1x10^{-6}$	0.006	0.376	0.5
CPD (raw)	ASI	58,767	57,893	1	0.06	0.0055	0.005	0.5	0.102
CPD (raw)	GAD	3,760	32,390	1	0.4	$1x10^{-6}$	0.026	0.5	0.5
CPD (raw)	MDD	9,679	23,400	0.75	0.15	0.0337	0.022	0.069	0.057
CPD (raw)	PSYCH	17,219	101,007	0.87	0.08	0.0119	0.008	0.082	0.059
CPD (raw)	SEX	62,062	56,585	1	0.05	0.0037	0.006	0.5	0.279
SC	ASI	58,552	57,606	1	0.11	0.0061	0.01	0.5	0.254
SC	GAD	5,080	46,659	0.38	0.28	0.0403	0.051	0.078	0.208
SC	MDD	13,385	34,204	0.64	0.35	0.0462	0.047	0.215	0.176
SC	PSYCH	21,798	137,902	1	0.15	$2.3x10^{-6}$	0.011	0.5	0.5
SC	SEX	82,356	77,866	0.96	0.09	0.004	0.009	0.304	0.339
SI	GAD	10,694	109,741	0.91	0.12	$1.6x10^{-6}$	0.015	0.245	0.5
SI	MDD	29,395	83,501	1	0.08	0.0022	0.01	0.5	0.409
SI	PSYCH	36,843	284,352	0.93	0.05	0.0024	0.005	0.287	0.44
SI	SEX	146,663	175,448	0.82	0.02	0.0372	0.004	$4.22x10^{-9}$	$3.44x10^{-13}$

[a] CPD: cigarettes per day (transformation); SC: smoking cessation; SI: smoking initiation

[b] ASI: age of smoking initiation; GAD: generalized anxiety disorder; MDD: major depressive disorder; PSYCH: whether an individual has seen a psychiatrist for nerves, anxiety, tension, or depression; SEX: males vs females.

[c] Sample sizes, consisting of unrelated individuals. For MDD and GAD, N_1 is the number of cases; for SEX, N_1 is the number of males; for ASI, N_1 is the number of late (> 17 y.o.) onset smokers; for PSYCH, N_1 is the number of individuals who have seen a psychiatrist for nerves, anxiety, tension, or depression.

[d] r_g= the between-strata genetic correlation, as estimated from a bivariate GREML model.

[e] PVE(σ^2_{het}) = the proportion of variance explained by heterogeneous effects, as estimated from a separate interaction GREML model (see Methods). For binary traits (SI, SC, and dichotomized CPD), the estimate is on the liability scale using sample trait prevalences.

[f] P-values obtained from a likelihood ratio test (LRT).

To determine if genetic variances themselves differed between strata, we estimated fold differences in variance components as estimated from the bivariate model (Fig. 1). For CPD

(binned, dichotomized, and raw scales), we observed nominal evidence of differing genetic variances between males and females, and between late-onset smokers and early-onset smokers (95% confidence interval of $\hat{\sigma}^2_{g_1}/\hat{\sigma}^2_{g_2}$ did not include 1). No other smoking phenotype showed evidence of differing genetic variances between moderator groups. Given that allele frequencies between strata are highly correlated (the genome-wide correlations of allele frequencies between strata were all >0.999), we hypothesized that differing genetic variances may reflect SNP effects that depend on the trait variance itself, for instance, the greater total variance of CPD in males than females (Fig 1). We repeated the bivariate analysis after standardizing the trait within strata (centering to zero mean and scaling to unit variance) and observed no clear differences in genetic or residual variances between strata. This indicates that while SNP effects for CPD may differ between sexes or between late-onset and early-onset smokers, such differences in SNP effects for CPD can be accounted for by differences in trait variance rather than differing genetic mechanisms.

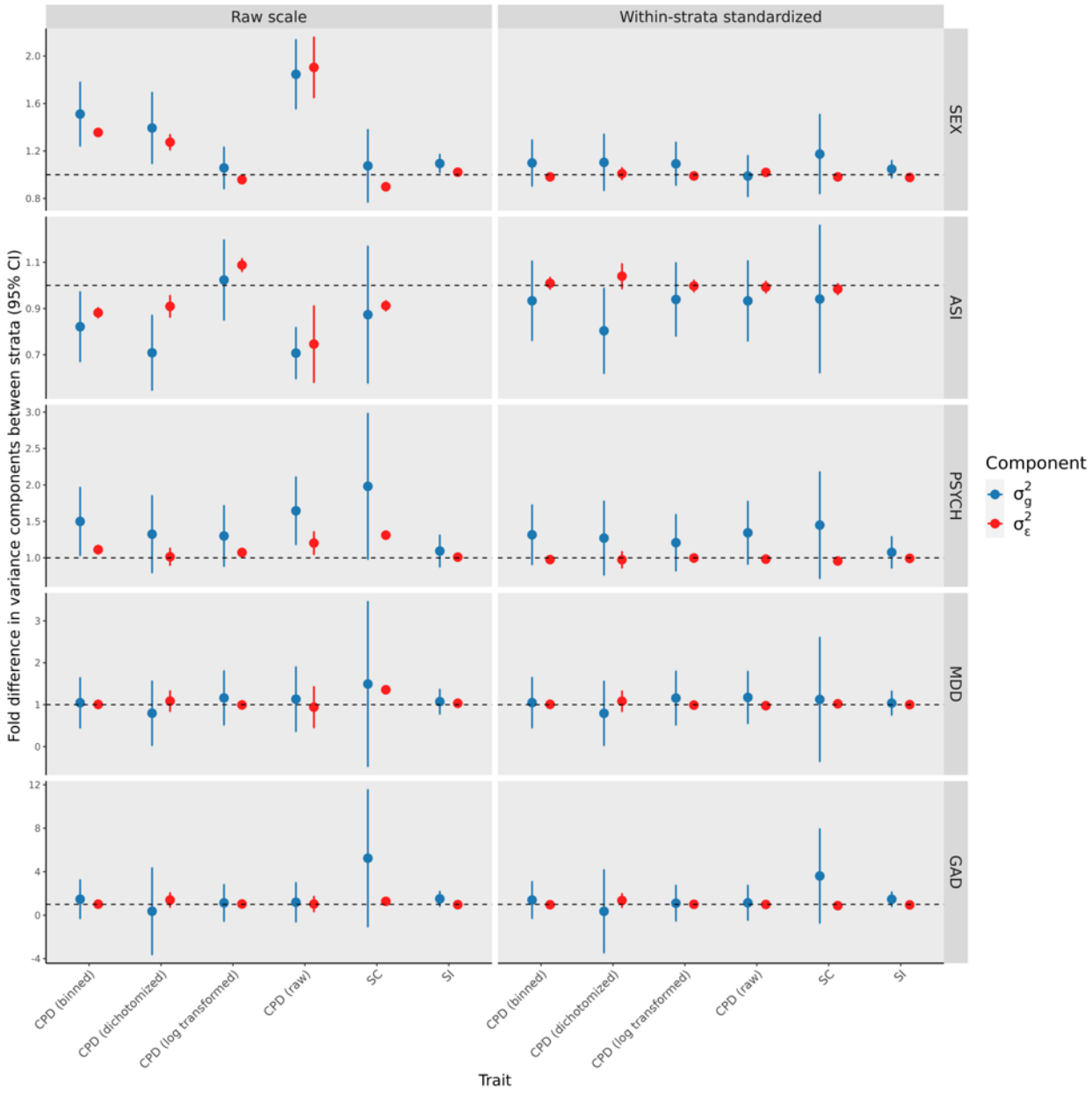


Fig 1. Fold differences in variance components between strata. Shown are point estimates and 95% confidence intervals. The dashed horizontal line at 1 indicates equal variance components. Each horizontal facet indicates a different moderator. For ASI, larger values than 1 indicate larger variances in late onset (Age 17-35) smokers than early onset smokers (Age 10-16). For both GAD and MDD, larger values than 1 indicate larger variances in cases than controls. For PSYCH, larger values than 1 indicate larger variances in those who have seen a psychiatrist for nerves than those who have not. For SEX, larger values than 1 indicate larger variances among males than females. For "Raw scale" the bivariate model is fit without standardizing the trait. For "Within-strata standardized", the bivariate model is fit after centering and scaling the trait within strata.

No evidence of differing partitioned heritabilities within cell-type-specific annotations.

Given evidence for disproportional genome-wide genetic effects, particularly for SI between sexes, we next asked whether evidence for differing SNP effects can be localized to functional genomic annotations. We performed stratified (*e.g.*, by sex) GWAS using BOLT-LMM[26], then used LD score regression (LDSC)[27,28] to partition the SNP-based heritability (h^2_{SNP}) within each strata, estimating strata-specific LDSC regression coefficients and h^2_{SNP} enrichment scores. When dichotomizing CPD, we were unable to use BOLT-LMM with GAD cases due to a limited sample size (N = 1854; full BOLT-LMM sample sizes shown in Supplementary Table S2) and therefore were unable to compare partitioned heritabilities for dichotomized CPD between GAD cases and controls. We tested a total of 221 annotations derived from four gene expression datasets[28], with each annotation corresponding to a set of SNPs within 100-kb of genes uniquely expressed in a particular tissue. We used a z-score to test for differences in LDSC coefficients and h^2_{SNP} enrichment scores between strata (see Methods). Across all 6188 tests (221 annotations by 28 trait-by-moderator combinations), we found no significant differences in LDSC regression coefficients or h^2_{SNP} enrichment scores after controlling for the false-discovery rate (Supplementary Figures S1-S2), indicating no evidence that groups differ in the heritable contribution of cell-type specific or other functional annotations.

Individual SNPs show evidence of heterogeneous effects

We next tested for differences in marginal SNP effect estimates between strata using a z-score of the difference in effect sizes and a two-sided *p*-value (*p*-diff; see methods;

Supplementary Figures S3-S58). Across all trait-by-moderator combinations tested, we observed 5 loci with genome-wide significant (GWS) evidence of heterogenous effects (p-diff < $5x10^{-8}$; Table 2). Notably, all 5 loci showed differing directions of effects of lead SNPs between strata, and none of these loci reached genome-wide significance within strata. Miami-plots that compare within-strata GWAS results can be found in Supplementary Figures S59-S86.

Table 2 Genome-wide significant differences in SNP effects (p-diff < $5x10^{-8}$)

Trait[a]	Moderator	Lead SNP	CHR	BP	$FREQ_1$[b]	$FREQ_2$[b]	Nearest Gene[c]	Control / early onset smoker		Case / late onset smoker		p-value		
								$\hat{\beta}$ (OR)	SE[d]	$\hat{\beta}$ (OR)	SE[d]	p-control/EOS[e]	p-case/LOS[e]	p-diff
CPD (binned)	MDD	rs13144992	4	145176474	0.14	0.14	*GYPA*	-0.033	0.011	0.0822	0.0177	0.003	$3.31x10^{-6}$	$3.23x10^{-8}$
CPD (log)	MDD	rs13144992	4	145176474	0.14	0.14	*GYPA*	-0.0218	0.00692	0.0506	0.0111	0.002	$5.58x10^{-6}$	$3.43x10^{-8}$
CPD (dichotomized)	ASI	rs78459872	6	124367555	0.96	0.96	*NKAIN2*	-0.0411 (0.85)	0.00924	0.0265 (1.13)	0.00823	$8.80x10^{-6}$	0.001	$4.71x10^{-8}$
SC[c]	GAD	rs112615043	3	150225437	0.97	0.97	*SERP1*	0.0131 (1.16)	0.00575	-0.11 (0.43)	0.0214	0.023	$3.07x10^{-7}$	$3.12x10^{-8}$
SC	GAD	rs77271903	4	67315863	0.91	0.91	*EPHA5*	0.00675 (1.08)	0.00329	-0.06 (0.63)	0.0118	0.04	$3.55x10^{-7}$	$4.87x10^{-8}$
SC	PSYCH	rs139501724	18	12831814	0.97	0.97	*PTPN2*	0.00564 (1.04)	0.0039	-0.0594 (0.72)	0.0112	0.148	$1.26x10^{-7}$	$4.59x10^{-8}$

[a] Coding of binary traits: CPD (dichotomized) compares heavy smokers (> 20 CPD; coded 1) to light smokers (<= 10; coded 0). SC compares current smokers (coded 1) to former smokers (coded 0).
[b] $FREQ_1$ = Frequency of effect allele in MDD/GAD/PSYCH controls, or early onset smokers. Vise-versa for $FREQ_2$.
[c] Nearest gene based on UCSC annotations. Gene is in bold if lead SNP is within the body of the gene.
[d] SE is on the observed scale (SE of $\hat{\beta}$)
[e] EOS = early onset smokers, LOS = late onset smokers

For CPD, SNPs with significant differences in effects between strata were largely dependent on the CPD transformation; both binned CPD and log transformed CPD showed GWS evidence of differing SNP effects at rs13144992 (p-diff = 3.23×10^{-8} and 3.43×10^{-8}, respectively) between MDD cases and controls, however this signal decayed with raw CPD (p-diff = 1.4×10^{-7}) and dichotomized CPD (p-diff = 1.89×10^{-4}). Likewise, when dichotomizing CPD, the only heterogenous signal reaching GWS was between late-onset and early-onset smokers, located at rs78459872, a SNP that exhibited no GWS evidence of ASI-dependent effects under different CPD transformations (p-diff > 4.3×10^{-5}). We further observed two loci with differing effects for SC between GAD cases and controls, and one locus with differing SC effects between PSYCH cases and controls. Given evidence that several traits possess differences in variance between strata (Fig. 1), we re-tested these five SNPs after standardizing the corresponding trait within strata to see if SNP effect differences may reflect differences in trait scale (Supplementary Table S3). After normalizing and re-testing, we found very little differences in p-diff values, however both the ASI-dependent SNP associated with dichotomized CPD (rs78459872) and the PSYCH-dependent SNP associated with SC (rs139501724) were no longer GWS (both p-diff = 5.3×10^{-8} after re-testing). Additional loci reached GWS in one stratum but not the other, however most of these instances (*e.g.* a GWS signal in MDD-controls but not in MDD-cases) are likely explained by differences in power (see Supplementary Figures S59-S86).

Novel SI-associated loci detected using sex-stratified GWAS

We found strong genome-wide evidence of heterogeneous genetic effects for smoking initiation between sexes, consistent with evidence from an independent Japanese sample [13]. However, we observed no sex differences in individual SNP effects for SI at GWS (see

Supplementary Figures S57-S58). Due to this surprising lack of genome-wide significant effect differences at individual SNPs, we performed additional analyses to characterize the genetic architecture of SI between sexes. Genome-wide, we observed an inverse relationship between MAF and estimated sex-differences in SNP effects, and simulations indicated a lack of power to detect GWS SNP effect differences across the MAF spectrum (Supplementary Figure S87).

Despite the lack of power to detect significant differences in effect estimates, our sample was well-powered to detect effects in either sex alone. In the sex-stratified GWAS for SI, we observed 51 GWS risk loci, consisting of 24 male risk loci not overlapping with a female risk locus, 23 female risk loci not overlapping with a male risk locus, and 4 risk loci that overlapped between males and females (Fig 2; Supplementary Tables S4 and S5). To determine whether novel SI-associated loci may be identified through sex-stratified GWAS, in which genetic effects are allowed to be heterogeneous between sexes, we compared our sex-stratified GWAS results to sex-combined GWAS results, in which a single effect is assumed to be shared by sexes (N = 418,329). We observed 14 independent ($r^2 < 0.1$) lead SNPs using sex-stratified GWAS that were not within a risk locus identified from the sex-combined GWAS, despite the roughly 2-fold larger sample size of the sex-combined analysis. Furthermore, six of these 14 loci were also not detected in a prior European ancestry meta-analysis of SI (the trait definition was identical to this study, see Methods for more details) involving roughly 6-fold more individuals than either sex-stratified analysis (N ~ 1.2 million[11], Fig 2 and Table 3). To quantify the degree that sex-stratified GWAS can lead to increased statistical power in detecting loci bearing heterogenous effects, we performed power analyses, varying sample size, MAF, and degree of SNP effect heterogeneity (Supplementary Figure S88). Sex-stratified GWAS consistently showed equal or greater power to detect any effect (whether it affects males, females, or both) at a sex-

heterogeneous effect locus than sex-combined GWAS. For example, we observed 3-fold increase in power at somewhat rare SNPs (MAF = 0.05) where the fold-difference in sex-specific odds ratios was 1.04; when observing real data, we found such differences in effects at MAF = 0.05 to be within a plausible range, indicating that our present sample sizes are simply underpowered to detect true heterogeneous effects between sexes (see Supplementary Figure S87).

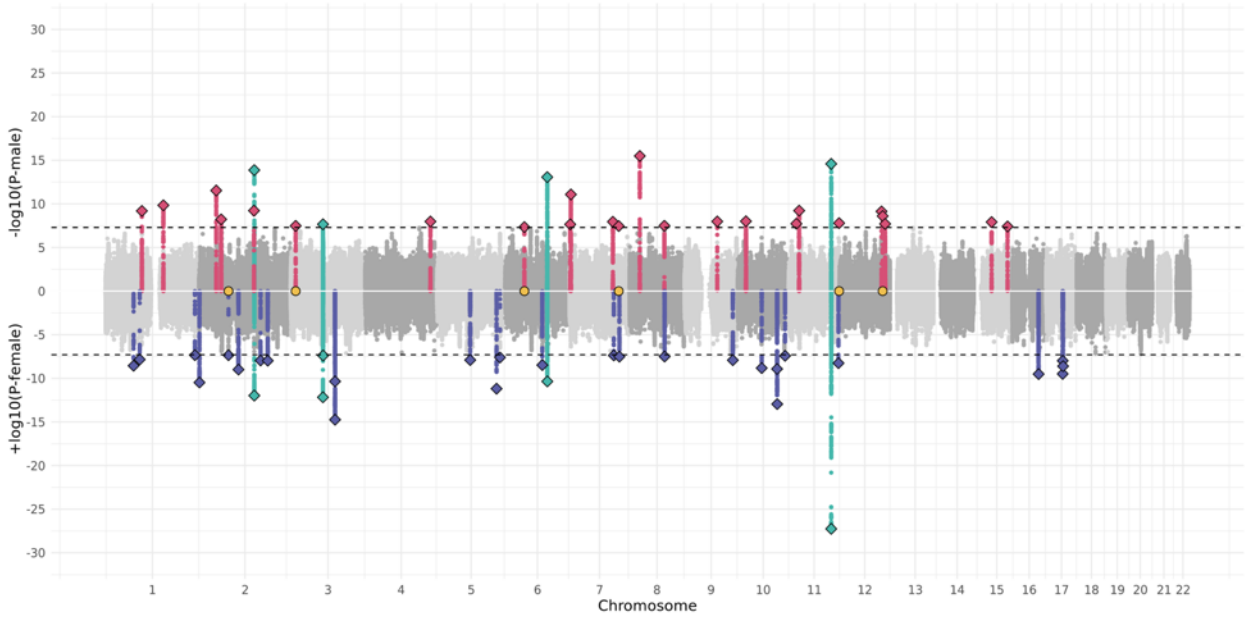


Fig 2. Miami-plot showing sex-specific GWAS signals for smoking initiation. The male manhattan plot for SI is shown above 0 on the x-axis, while the female manhattan plot for SI is shown below 0 on the x-axis. SNPs are highlighted red if they were within a male-specific genomic risk locus (a locus that did not overlap with any female-specific GWS risk locus), and vice versa for blue SNPs. In teal are SNPs within risk loci that overlapped between males and females. Diamonds indicate independent ($r^2 < 0.1$), sex-specific lead SNPs. Yellow circles mark the position of novel signals (Table 3)—lead SNPs that reached GWS in either males or females but were not within a detectable risk locus when performing a sex-combined GWAS in the UK Biobank nor within a risk locus in a prior meta-analysis of SI (N ~ 1.2M)[11].

Table 3 Novel[a] SI-associated loci discovered using sex-stratified GWAS

Lead SNP[b]	CHR	BP	FREQ[c]	Nearest Gene(s)[d]	Female		Male		*p*-value			Power to detect difference at GWS[e]
					$\hat{\beta}$ ($\widehat{OR}$)	SE	$\hat{\beta}$ ($\widehat{OR}$)	SE	*p*-female	*p*-male	*p*-diff	
rs6547148	2	77747513	0.46	***LRRTM4***	-0.00776 (0.968)	0.00142	-0.000971 (0.996)	0.00158	4.50×10^{-8}	0.538	1.38×10^{-3}	0.014
rs360892	3	13094397	0.39	***IQSEC1***	0.00267 (1.01)	0.00144	0.00888 (1.04)	0.00161	0.064	3.33×10^{-8}	4.05×10^{-3}	0.002
rs7770532	6	50940775	0.62	*TFAP2B*	0.00136 (1.01)	0.00146	0.00891 (1.04)	0.00163	0.354	4.57×10^{-8}	5.59×10^{-4}	0.017
rs76759272	7	130612803	0.96	*FLJ43663*	-0.00122 (0.995)	0.00379	-0.0233 (0.911)	0.00423	0.748	3.48×10^{-8}	9.90×10^{-5}	0.058
rs669257	11	133836301	0.80	*IGSF9B*	0.00335 (1.01)	0.00177	0.0111 (1.05)	0.00197	0.059	1.61×10^{-8}	3.27×10^{-3}	0.003
rs11066972	12	114603439	0.86	*RBM19/TBX5*	0.00239 (1.01)	0.00202	0.0135 (1.06)	0.00226	0.236	2.50×10^{-9}	2.56×10^{-4}	0.025

[a] "Novel" indicates that the lead SNP was not within any genomic risk loci identified from a sex-combined GWAS using the UK Biobank (N = 418,329), nor within any genomic risk loci from a prior sex-combined meta-analysis of SI (N = 1,232,091) [11].
[b] Lead SNP, either within a risk locus identified only in males, or only in females.
[c] Allele frequency of effect allele. Shown is the allele frequency obtained in females, with the corresponding male allele frequency not differing by more than 0.2%
[d] Nearest gene based on UCSC annotations. Gene is bold if SNP is within the body of the gene.
[e] Estimated power to detect a difference in sex-specific effects at *p*-diff < 5×10^{-8}, assuming point estimates of male and female effects are the true values, and assuming effect allele frequencies in [c].

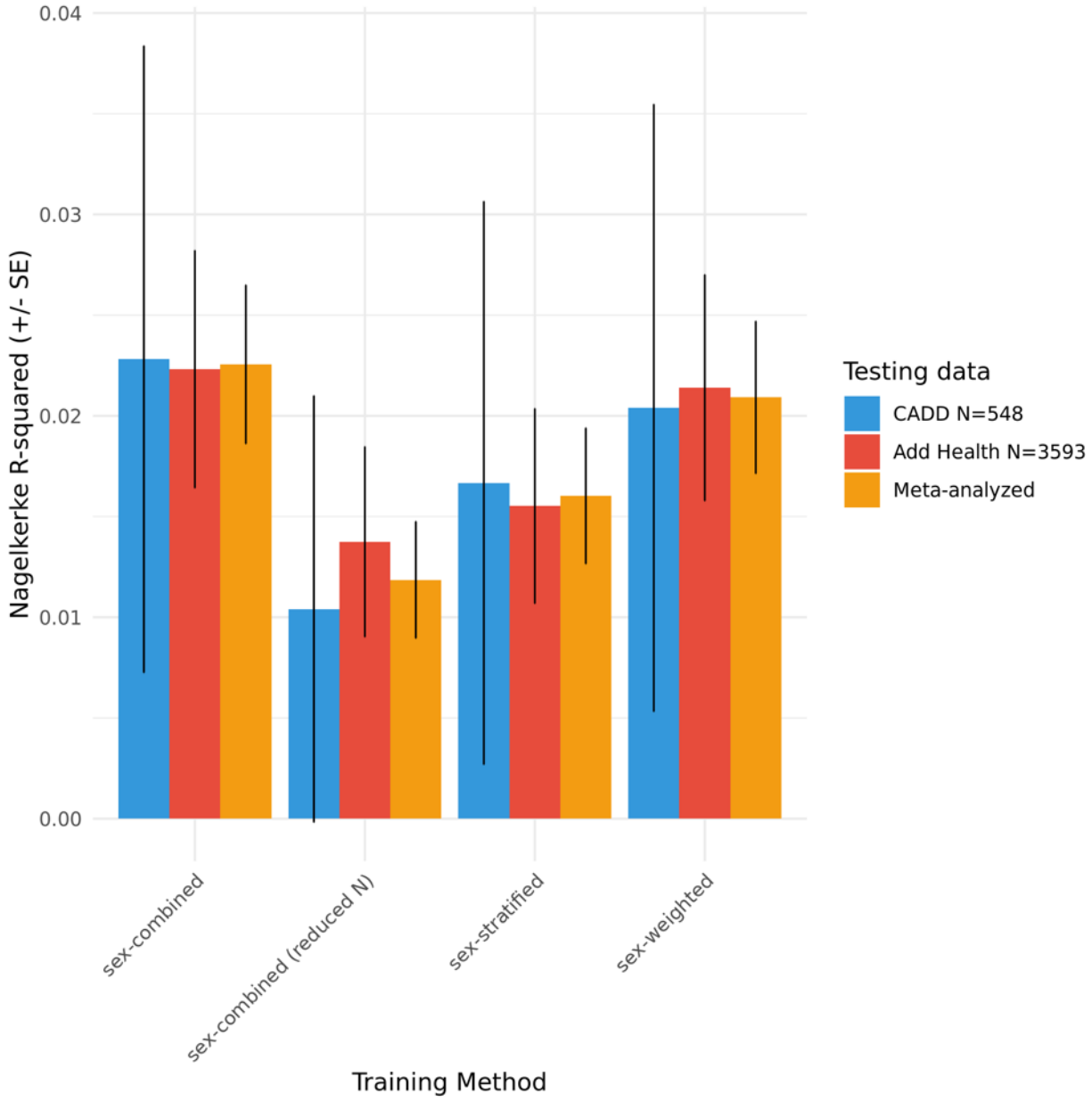
For males and females separately, we then estimated the genetic correlation between SI and 757 sex-combined traits on LDhub [29]. Despite observing very different GWS signals between males and females for SI, we observed that for all 757 traits, the male-specific SI genetic correlation estimate (95% CI) overlapped with the female-specific SI genetic correlation estimate (95% CI) (Supplementary Tables S6-S7).

Suggestive evidence of enhancing SI PRS accuracy using sex-specific SNP effects

Using independent target data from the National Longitudinal Study of Adolescent to Adult Health (Add Health) [30] and from the Center on Antisocial Drug Dependence (CADD) [31], we tested whether allowing for sex-specific SNP effects can enhance polygenic risk score (PRS) accuracy for SI, when compared to assuming SNP effects are shared between sexes. We computed male-specific and female-specific PRSs from the corresponding sex-stratified GWAS summary stats (derived from the UK Biobank) using SBLUP in GCTA [25]. For comparison, we computed PRS from the sex-combined GWAS summary statistics and a sex-combined GWAS in which sample size was halved to resemble the sex-stratified sample sizes. Additionally, we computed another PRS derived from weighted sex-specific GWAS statistics, as implemented in SMTpred[32]. We computed prediction accuracy of PRS when compared to a covariate-only model using logistic regression and Nagelkerke's R^2 (see Methods).

Prediction accuracy was similar between the two independent target datasets across all training methods. Prediction accuracy was highest using the full, sex-combined training sample (Fig. 3; Supplementary Table S8). Sex-stratification of the GWAS slightly improved prediction compared to the sex-combined analysis when the training sample sizes were approximately equal, but the prediction standard errors were large and the improvement was not statistically

397 significant. Using a weighted index to borrow information across sex-specific SNP effects
398 improved prediction accuracy to a level comparable to that obtained with the full, sex-combined
399 training sample.
400



401

402 Fig 3. Accuracy of polygenic risk scores for smoking initiation using two testing datasets. Shown
403 is the Nagelkerke R^2 and intervals show one standard error obtained from 1000 bootstrap
404 samples. Meta-analysis across datasets was done using inverse-variance weighting. Sex-
405 combined (reduced N) used a reduced training sample size to roughly match the sample size
406 obtained through sex-stratification. Sex-combined training N = 418,329; sex-combined (reduced
407 N) training N = 209,157; sex-weighted and sex-stratified training Ns = 189693 males and
408 228636 females.

409

Discussion

We observed modest evidence of genome-wide genetic moderation of smoking behaviors due to psychiatric disorders and age of smoking initiation, and several individual loci with differing strata-specific SNP effects. However, we observed strong genome-wide evidence that sex moderates the genetic contribution to smoking initiation, and identified numerous novel loci associated with SI when we stratified the GWAS by sex. These novel loci were not detected in a prior sex-combined GWAS of the same phenotype [11], despite utilizing a 6-fold increase in sample size than the sex-stratified GWAS presented in this study. This suggests that more nuanced analyses can aid in uncovering the genetic factors that contribute to the initiation of regular smoking and possibly in complex traits more generally.

Psychiatric disorder moderation

Twin-based studies have found depression moderates the genetic variance of smoking heaviness[16]. We observed suggestive evidence that the genetic correlations of CPD between MDD cases and controls and SC between GAD cases and controls are less than one. A lack of stronger evidence is likely due to the low power to detect genetic variance differences between cases and controls (*e.g.*, the number of unrelated GAD cases with an SC record was ~5K, Table 1).

Consistent with these findings, we observed relatively few genome-wide significant SNP effects that differed among groups, with at least one depending on the trait scale. Within 100kb of these SNP-by-moderator associations are loci previously associated with COPD and lung cancer[33], forced vital capacity[34], waist circumference [35], forced vital capacity in COPD patients[36], sleep quality[37], and dinner intake in a Hispanic population[38]. Collectively,

433 these results are consistent with a modest genetic effect moderation on smoking by depression
434 and generalized anxiety, which is consistent with reports using twin samples, though of smaller
435 magnitude[16–19]. Larger, independent samples with well-phenotyped psychiatric data—which
436 are currently limited—will be necessary to identify heterogeneous effect loci, but we expect the
437 difference in magnitude of individual effects will be small.
438
439 **Moderation by sex and age of smoking initiation**
440 As seen in Fig. 1, depending on how one measures smoking heaviness (through different
441 transformations of CPD), we observed different variances but no evidence of differential SNP
442 heritabilities of CPD between males and females, and between late and early onset smokers.
443 Non-genetic factors may largely be the direct cause of differing variances in CPD, with genetic
444 effects differing proportionally due to these differences in scale. We did observe a single SNP
445 with differing CPD effects between late-onset and early-onset smokers, however again this effect
446 depended on the chosen CPD scale and transformation.
447 Alternatively, we observed clear evidence of disproportional genetic effects for SI
448 between males and females, indicative of sex differences in genetic mechanisms that contribute
449 to smoking initiation risk. This is consistent with a previous, independent study that estimated
450 the between-sex genetic correlation to be < 1 for smoking initiation (measured as ever vs. never
451 smokers) using common SNPs in a Japanese population[13]. Intriguingly, we identified no
452 individual SNP effects at that differed between sexes at genome-wide significance (p-diff <
453 $5x10^{-8}$), indicating that such effects are exceedingly small. Furthermore, we observed no
454 evidence of differing genetic correlations with SI between males and females when testing over
455 700 traits on LD Hub. This could indicate that while males and females possess partially distinct

456 genetic loci for SI, the functional consequences of each may not differ dramatically, genome-
457 wide.
458
459 **Novel SI-associated loci identified through sex-stratified GWAS**
460 While increasing sample size is one strategy to uncover novel risk loci for SI, the presence
461 of heterogenous SNP effects could enable a more nuanced, sex-stratified analysis to uncover
462 certain SI-associated loci more efficiently (see Fig 2; Supplementary Fig S88). We identified 6
463 GWS loci using sex-stratified GWAS for SI that were not detected in a sex-combined GWAS of
464 the UK Biobank, nor detected in the largest known GWAS meta-analysis of SI[11]. These loci
465 highlight the fact that, in addition to the improved power gained from increasing sample size such
466 as in Liu et al.[11], incorporating nuanced analyses to investigate possible heterogeneous effects
467 among groups can identify novel associations and provide a better understanding of the underlying
468 trait architecture. Please see the Supplementary Note for discussion about these six novel signals.
469 A more nuanced association analysis may also improve genomic prediction. Leveraging
470 sex-specific SNP effects for sex-stratified SI prediction appeared to increase accuracy when
471 compared to sex-combined prediction using a comparable training sample size. When borrowing
472 information between sex-specific SNP BLUPs using a weighted index as implemented in
473 SMTpred[32], prediction accuracy increased to a similar level obtained when training with the
474 full, sex-combined training sample. Future work may seek to develop additional means to borrow
475 information between males and females to optimize prediction accuracy while allowing for
476 heterogeneous genetic effects. Although the large standard errors complicate distinguishing the
477 optimal approach among sex-combined, sex-weighted, and sex-stratified training methods, our

478 results are consistent with improved prediction when allowing for heterogeneous effects, when
479 accounting for training sample size.
480
481 **Limitations**
482 Requiring individuals to possess both smoking and moderator phenotypes reduced our
483 sample sizes, in some cases, severely (*e.g.*, the aforementioned 5K individuals possessing both a
484 SC and GAD record), which led to reduced power and greater uncertainty in effect estimates. For
485 example, we detected GWS evidence of heterogeneous SNP effects associated with CPD and
486 SC, however none of these SNPs reached GWS within strata. Furthermore, we cautiously note
487 that sex-specific risk loci identified from two independent GWAS do not necessarily imply
488 heterogeneous effects between sexes. In particular, if two non-overlapping risk loci are in close
489 proximity, identification in one but not the other sex-specific GWAS may result from random
490 sampling of genotypes rather than heterogeneous genetic effects, or from subtle differences in
491 power (N_{males} = 189,693 and $N_{females}$ = 228,636). Differential calling of genotypes or genotype
492 sampling could partially explain why some sex-specific SNP effects reaching GWS in this study
493 did not reach GWS in a prior meta-analyzed sample of the same phenotype. For studying the
494 effects of MDD or GAD moderation, we emphasize the pressing need for an independent,
495 replication dataset, however, currently there are very few samples containing large numbers of
496 genotyped individuals with both smoking records and MDD-DSMV-like/GAD-DSMV-like
497 records. Further work will be crucial in investigating MDD- and GAD-dependent genetic effects
498 that contribute to smoking behaviors.
499
500

Summary

For highly complex traits such as smoking behaviors, incorporating more nuanced analyses, such as a careful consideration of possible context-dependent effects, may provide a more complete picture of their genetic architecture. Such heterogenous genetic effects may contribute to estimated allelic effects that are infinitesimally small (and difficult to detect) within the population as a whole, even in very large samples. Given smoking's heavy burden on human health, there is a strong incentive to continue to pursue evidence of heterogeneous effects that can disproportionally burden certain groups.

Materials and Methods

Genotypes, phenotypes and moderators

All genotypes, phenotypes, and moderators were obtained from the UK Biobank[23]. Phenotype definitions for smoking behaviors matched exactly GSCAN[11] definitions: Smoking initiation (SI) was a binary phenotype that compared individuals who had smoked at least 100 cigarettes to individuals who had never smoked. Smoking cessation (SC) was a binary phenotype that compared current smokers to former smokers (fields 1239 and 1249). Cigarettes per day (CPD) was based on "Number of cigarettes currently smoked daily (current cigarette smokers)", "Number of cigarettes previously smoked daily", or "Number of cigarettes previously smoked daily (current cigar/pipe smokers)" (UK Biobank data fields 2887, 3456, and 6183). Different transformations of CPD were considered, raw CPD, natural log transformed CPD, binned CPD (matching GSCAN defined CPD) included five bins [1 – individuals who smoke(d) 1 to 5; 2 – individuals who smoke(d) 6 to 15; 3 – individuals who smoke(d) 16 to 25; 4 – individuals who smoke(d) 26 to 35; 5 – individuals who smoke(d) 36 to 140], and dichotomized CPD (individuals

who smoke more than 20 cigarettes per day vs individuals who smoke 10 or less, excluding remaining individuals). Age of smoking initiation (ASI) was defined as the age at which an individual began smoking regularly (fields 3426 and 2867). Data from the UK Biobank mental health questionnaire was used to construct MDD DSMV-like, and GAD DSMV-like records. GAD DSMV-like cases required endorsement of either Field IDs 20425 or 20542, and endorsement of 20421 with 20420 reported as ≥56 months, and endorsement of 20540 or 20543≥2, and endorsement of 20541 or 20537 or 20539, as well as three or more "Yes" responses to the following symptom Field IDs: 20426 or 20423, 20429, 20419, 20422, 20417, 20427, and endorsement of 'a little' or more of field 20418 (impairment or impact). Individuals with complete data but who did not meet the above criteria were treated as controls. Supplementary Figure S89 visually describes the assignment of GAD DSMV-like cases and controls. Similarly, MDD DSMV-like cases required "Yes" responses to 5 or more of the following symptom Field IDs: 20446, 20441, 20533, 20534, 20535, 20449, 20536, 20450, 20435, and 20437, as well as "somewhat" or more response to field 20440, a "almost every day" or more response to field 20439, and a "about half of the day" or more response to 20436.

Estimating within-strata variance components and between-strata genetic correlations

Using GCTA's[25] bivariate model implementation (https://cnsgenomics.com/software/gcta/#BivariateGREMLanalysis), we fit a bivariate model treating the same trait—measured in different strata—as two different phenotypes. For example, we modeled the genetic (co)variance of CPD measured in MDD-like cases and CPD measured in MDD-like controls. For a particular trait-by-moderator combination, we built a genetic relationship matrix (GRM) using filtered, genotyped SNPs. Filtering of SNPs was performed

using 436,065 European-ancestry individuals after removing individuals with mismatched self-reported and genetic sex, $|F_{het}| \geq 0.2$, and/or no phenotypic information, where SNPs were removed if they had a genotyping rate or MAF less than 0.05 or had a *p*-value from a Hardy-Weinberg test smaller than $1x10^{-8}$. GRM entries were then pruned using a relatedness cutoff of 0.05. Fixed effect covariates consisted of sex, batch, assessment center, education level, Townsend deprivation index, age, age squared, and the first 10 genomic principle components, derived from both the European subset of the UK Biobank and the whole UK biobank. When analyzing CPD, current vs former smoker status was included as an additional covariate. Point estimates and standard errors of within-strata variance components and heritabilities were obtained from GCTA, as were between-strata genetic covariances and correlations. Standard errors of estimated variance component fold differences (e.g. $\frac{\widehat{\sigma^2_{g_1}}}{\sigma^2_{g_2}}$, where 1 and 2 index opposing strata) were approximated using the delta method[39], utilizing the sampling variances/covariances of model parameters obtained from GCTA. Specifically, the variance of a ratio of genetic variance estimates was approximated by $\left(\frac{\hat{\sigma}^2_{g_1}}{\hat{\sigma}^2_{g_2}}\right)^2 \left(\frac{\mathrm{var}(\hat{\sigma}^2_{g_1})}{(\hat{\sigma}^2_{g_1})^2} + \frac{\mathrm{var}(\hat{\sigma}^2_{g_2})}{(\hat{\sigma}^2_{g_2})^2} - 2\frac{\mathrm{cov}(\hat{\sigma}^2_{g_1},\hat{\sigma}^2_{g_2})}{(\hat{\sigma}^2_{g_1})(\hat{\sigma}^2_{g_2})}\right)$, with the ratio of residual variances between strata approximated similarly. The genetic correlation is defined as $r_g = \frac{\sigma_{g_1 g_2}}{\sqrt{\sigma^2_{g_1}\sigma^2_{g_2}}}$ and testing $H_0: r_g = 1$ was done using a likelihood ratio test, comparing the full model (above) to one in which r_g is constrained to one.

Modeling the proportion of variance explained by heterogeneous effects

GCTA's univariate model[25], using the –gxe argument (https://cnsgenomics.com/software/gcta/#GREMLanalysis) was used to decompose the total

variance across strata into a shared component (σ_g^2), a deviation from the shared component (σ_{het}^2), and a residual component (σ_ε^2) using the same GRM and fixed effects covariates as before. Using GCTAs model parameter estimates, we estimated the proportion of phenotypic variance explained by heterogeneous effects, $PVE(\sigma_{het}^2) = \frac{\sigma_{het}^2}{(\sigma_g^2+\sigma_{het}^2+\sigma_\varepsilon^2)}$, with the standard error provided by GCTA, and tested $H_0: \sigma_{het}^2 = 0$ using a likelihood ratio test, comparing the likelihood of the full model to one in which σ_{het}^2 is constrained to 0. For binary traits (dichotomized CPD, SI, and SC), we transformed estimates of $PVE(\sigma_{het}^2)$ to the liability scale[40], using sample prevalences.

GWAS, heterogeneous effect inference, and novel SI-loci detection

Stratified GWAS was performed using BOLT-LMM[26]. For all GWAS we used all individuals of European descent, including related individuals. For each BOLT-LMM model fit, fixed effect covariates were identical to those used in GREML based models. GWAS was performed using imputed SNPs with a 0.9 INFO score and 0.01 MAF cutoff, resulting in 7,749,105 SNPs with which to obtain within-strata estimated SNP effects and their standard errors. A z-score was used to infer differences in SNP effects between strata 1 and 2: $z_{snp} = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{\sqrt{SE(\widehat{\beta}_1)^2 + SE(\widehat{\beta}_2)^2}}$, with a two-sided p-value: p-diff = $2\Phi(-|z_{snp}|)$, where Φ is the normal cumulative distribution function. For all BOLT-LMM model fits, the random polygenic component was estimated from the infinitesimal model as opposed to BOLT's mixture model. When performing sex-combined GWAS for SI, we combined males and females and ran BOLT similarly while including sex as a covariate. Genomic risk loci and lead SNP identification was performed using FUMA[41] with default parameters. To identify potentially novel SI-associations, we performed sex-combined

GWAS using the UK Biobank, then used FUMA to identify genomic risk loci and lead SNPs. We then determined whether a lead SNPs identified from sex-stratified GWAS (associating with males, females, or both) was within an identified sex-combined genomic risk locus. Similarly, we determined whether lead SNPs from the sex-stratified GWAS were within risk loci reported in Liu et al. 2019[11].

Partitioning stratified heritability estimates to functional categories

Using stratified GWAS results obtained from BOLT-LMM (see above), we performed cell-type specific LDSC analysis[28] to partition strata-specific SNP heritability along functional annotations, *i.e.*, SNPs within 100kb of genes uniquely expressed in a particular tissue. This was done using within-annotation LD scores, with annotations derived from the baseline model[27], the Cahoy et al. gene-expression dataset[42], GTEx (both multi-tissue assessment of cell-type specific genes and brain-specific assessment of cell-type specific genes)[43], and the Franke Lab gene expression dataset[44,45]. LD scores were downloaded and analysis was carried out according to the steps outlined at https://github.com/bulik/ldsc/wiki/Cell-type-specific-analyses. Only hapmap3 SNPs were used in all LDSC analyses. To infer differences in partitioned heritabilities between strata, firstly we fit the partitioned LDSC model separately for each strata, thus obtaining strata-specific partitioned LDSC estimated coefficients (representing the per-SNP contribution to heritability from a particular annotation) and estimated heritability enrichments (heritability of an annotation divided by the number of SNPs in the annotation). We then inferred differences in LDSC estimated coefficients and heritability enrichments between strata, using the same z-score and two-sided testing approach outlined in the previous section.

614

615 **Simulations**

616 All simulations utilized a single causal variant model, whereby genotypes at the causal variant x
617 were sampled from the binomial distribution: $x \sim B(2, p)$, where p is the minor allele frequency.
618 Simulating a sex-specific binary trait y was done by sampling: $y \sim B\left(1, e^{x\beta}/\left(1 + e^{x\beta}\right)\right)$, with
619 β being the sex-specific log odds ratio of the causal variant. To transform linear coefficients
620 b (like those obtained from BOLT-LMM) to log odds ratios, we used the approximation $\beta \approx$
621 $\frac{b}{\mu(1-\mu)}$, with μ being the sex-specific case fraction. For each simulation, sex-specific causal
622 variant effects were fixed and genotypes at the causal variant were randomly sampled for 5000
623 replicates.

624

625 **Polygenic risk scoring and prediction accuracy**

626 To compute SNP effects used in polygenic risk scoring, we used marginal SNP effects from
627 BOLT-LMM model fits, then obtained best linear unbiased predictions (BLUPs) of SNP effects
628 (for all 7,749,105 SNPs) that account for LD using SBLUP [25,46]. In both CADD and Add
629 Health datasets, we used randomly sampled, unrelated individuals of European descent. In each
630 dataset, PRS were computed from SNPs imputed from haplotype reference consortium data
631 (MAF > 0.01, INFO R^2 > 0.95). In CADD, we predicted the response to "Have you smoked at
632 least 20 cigarettes in your lifetime?" and in Add Health we predicted the response to "Have you
633 ever smoked cigarettes regularly, that is, at least 1 cigarette every day for 30 days?". To assess
634 prediction accuracy, we compared a full model including PRS and covariates to a reduced model
635 including covariates only. For both datasets, covariates were age, age-squared, sex, educational
636 attainment (coded categorically), and the first 10 genomic principal components.

637

638 **Acknowledgements**

639 This work was supported by NIMH Training Grant 5T32MH016880-38, R01 AG046938-06
640 (MPIs: Reynolds, Wadsworth), R01 DA044283-01 (Vrieze), R01 MH100141-06 (Keller), and
641 awards DA042755 and DA032555 (Hopfer). CADD and GADD data were supported by
642 DA011015, DA035804, DA012845, DA035804, and DA021692. This work utilized the Summit
643 supercomputer, which is supported by the National Science Foundation (awards ACI-1532235
644 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The
645 Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State
646 University. Data storage supported by the University of Colorado Boulder "PetaLibrary". This
647 research has been conducted using the UK Biobank Resource (application number 1665). We
648 thank the UK Biobank and the participants of the UK Biobank, and the participants of the
649 AddHealth (PI: Harris) and CADD (PI: Hewitt) studies. Lastly, we thank Robin Corley for
650 valuable contributions to the writing of this manuscript.

651

652

653 **References**

654 1. US Department of Health and Human Services. The health consequences of smoking – 50
655 years of progress: a report of the Surgeon General. Rep Surg Gen. 2014.
656 doi:10.1016/S0025-7125(16)30355-8
657 2. Swendsen J, Conway KP, Degenhardt L, Glantz M, Jin R, Merikangas KR, et al. Mental
658 disorders as risk factors for substance use, abuse and dependence: results from the 10-year
659 follow-up of the National Comorbidity Survey. Addiction. 2010;105: 1117–1128.

doi:10.1111/j.1360-0443.2010.02902.x

3. Brook J, Brook D, Zhang C. Psychosocial predictors of nicotine dependence in Black and Puerto Rican adults: A longitudinal study. Nicotine Tob Res. 2008;10: 959–967. doi:10.1080/14622200802092515

4. Karp I, O'Loughlin J, Hanley J, Tyndale RF, Paradis G. Risk factors for tobacco dependence in adolescent smokers. Tob Control. 2006;15: 199–204. doi:10.1136/tc.2005.014118

5. Kendler KS, Myers J, Damaj MI, Chen X. Early smoking onset and risk for subsequent nicotine dependence: A monozygotic co-twin control study. Am J Psychiatry. 2013;170: 408–413. doi:10.1176/appi.ajp.2012.12030321

6. Higgins ST, Kurti AN, Redner R, White TJ, Gaalema DE, Roberts ME, et al. A literature review on prevalence of gender differences and intersections with other vulnerabilities to tobacco use in the United States, 2004–2014. Prev Med (Baltim). 2015;80: 89–100. doi:10.1016/j.ypmed.2015.06.009

7. Kaleta D, Wojtysiak P, Usidame B, Elzbieta DZ, Fronczak A, Korytkowski P, et al. Heaviness of smoking among employed men and women in Poland. Int J Occup Med Environ Health. 2016;29: 191–208. doi:10.13075/ijomeh.1896.00433

8. Wetter DW, Kenford SL, Smith SS, Fiore MC, Jorenby DE, Baker TB. Gender differences in smoking cessation. J Consult Clin Psychol. 1999;67: 555–562. doi:10.1037/0022-006X.67.4.555

9. Carmelli D, Swan GE, Robinette D, Fabsitz R. Genetic Influence on Smoking — A Study of Male Twins. N Engl J Med. 1992;327: 829–833. doi:10.1056/NEJM199209173271201

10. Heath AC, Martin NG. Genetic models for the natural history of smoking: evidence for a

genetic influence on smoking persistence. Addict Behav. 1993;18: 19–34. doi:10.1016/0306-4603(93)90005-T

11. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, 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. 2019;51: 237–244. doi:10.1038/s41588-018-0307-5

12. Karlsson Linnér R, Biroli P, Kong E, Meddens SFW, Wedow R, Fontana MA, 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. 2019;51: 245–257. doi:10.1038/s41588-018-0309-3

13. Matoba N, Akiyama M, Ishigaki K, Kanai M, Takahashi A, Momozawa Y, et al. GWAS of smoking behaviour in 165,436 Japanese people reveals seven new loci and shared genetic architecture. Nat Hum Behav. 2019;3: 471–477. doi:10.1038/s41562-019-0557-y

14. Xu K, Li B, McGinnis KA, Vickers-Smith R, Dao C, Sun N, et al. Genome-wide association study of smoking trajectory and meta-analysis of smoking status in 842,000 individuals. Nat Commun. 2020;11. doi:10.1038/s41467-020-18489-3

15. Brazel DM, Jiang Y, Hughey JM, Turcot V, Zhan X, Gong J, et al. Exome Chip Meta-analysis Fine Maps Causal Variants and Elucidates the Genetic Architecture of Rare Coding Variants in Smoking and Alcohol Use. Biol Psychiatry. 2019;85: 946–955. doi:10.1016/j.biopsych.2018.11.024

16. Keskitalo-Vuokko K, Korhonen T, Kaprio J. Gene-Environment Interactions between Depressive Symptoms and Smoking Quantity. Twin Res Hum Genet. 2016;19: 322–329. doi:10.1017/thg.2016.36

17. Harden KP, Hill JE, Turkheimer E, Emery RE. Gene-environment correlation and

706 interaction in peer effects on adolescent alcohol and tobacco use. Behav Genet. 2008;38:
707 339–347. doi:10.1007/s10519-008-9202-7
708 18. Lessov CN, Martin NG, Statham DJ, Todorov AA, Slutske WS, Bucholz KK, et al.
709 Defining nicotine dependence for genetic research: Evidence from Australian twins.
710 Psychol Med. 2004;34: 865–879. doi:10.1017/S0033291703001582
711 19. Boardman JD, Saint Onge JM, Haberstick BC, Timberlake DS, Hewitt JK. Do schools
712 moderate the genetic determinants of smoking? Behav Genet. 2008;38: 234–246.
713 doi:10.1007/s10519-008-9197-0
714 20. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to
715 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol
716 use. Nat Genet. 2019;51: 237–244. doi:10.1038/s41588-018-0307-5
717 21. Hartz SM, Short SE, Saccone NL, Culverhouse R, Chen L, Schwantes-An T-H, et al.
718 Increased Genetic Vulnerability to Smoking at CHRNA5 in Early-Onset Smokers. Arch
719 Gen Psychiatry. 2012;69: 854. doi:10.1001/archgenpsychiatry.2012.124
720 22. Adjangba C, Border R, Villela PNR, Ehringer MA, Evans LM. Little Evidence of
721 Modified Genetic Effect of rs16969968 on Heavy Smoking Based on Age of Onset of
722 Smoking. MedRxiv. 2020. doi:10.1101/2020.04.22.20071407
723 23. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An
724 Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases
725 of Middle and Old Age. PLOS Med. 2015;12: e1001779.
726 doi:10.1371/journal.pmed.1001779
727 24. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of
728 genetic correlations across human diseases and traits. Nat Genet. 2015;47: 1236–1241.

729 doi:10.1038/ng.3406

730 25. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex
731 trait analysis. Am J Hum Genet. 2011;88: 76–82. doi:10.1016/j.ajhg.2010.11.011

732 26. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al.
733 Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat
734 Genet. 2015;47: 284–290. doi:10.1038/ng.3190

735 27. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning
736 heritability by functional annotation using genome-wide association summary statistics.
737 Nat Genet. 2015;47: 1228–1235. doi:10.1038/ng.3404

738 28. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al.
739 Heritability enrichment of specifically expressed genes identifies disease-relevant tissues
740 and cell types. Nat Genet. 2018;50: 621–629. doi:10.1038/s41588-018-0081-4

741 29. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, et al. LD
742 Hub: A centralized database and web interface to perform LD score regression that
743 maximizes the potential of summary level GWAS data for SNP heritability and genetic
744 correlation analysis. Bioinformatics. 2017;33: 272–279.
745 doi:10.1093/bioinformatics/btw613

746 30. Harris KM, Halpern CT, Whitsel E, Hussey J, Tabor J, Entzel P, et al. The National
747 Longitudinal Study of Adolescent to Adult Health: Research Design. 2009.

748 31. Rhea SA, Bricker JB, Wadsworth SJ, Corley RP. The colorado adoption project. Twin Res
749 Hum Genet. 2013;16: 358–365. doi:10.1017/thg.2012.109

750 32. Maier RM, Zhu Z, Lee SH, Trzaskowski M, Ruderfer DM, Stahl EA, et al. Improving
751 genetic prediction by leveraging genetic correlations among human diseases and traits.

752 Nat Commun. 2018;9: 1–17. doi:10.1038/s41467-017-02769-6

753 33. Young RP, Whittington CF, Hopkins RJ, Hay BA, Epton MJ, Black PN, et al.
754 Chromosome 4q31 locus in COPD is also associated with lung cancer. Eur Respir J.
755 2010;36: 1375–1382. doi:10.1183/09031936.00033310

756 34. Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, et al. A genome-
757 wide association study of pulmonary function measures in the framingham heart study.
758 PLoS Genet. 2009;5: 1–8. doi:10.1371/journal.pgen.1000429

759 35. Wen W, Kato N, Hwang JY, Guo X, Tabara Y, Li H, et al. Genome-wide association
760 studies in East Asians identify new loci for waist-hip ratio and waist circumference. Sci
761 Rep. 2016;6: 2–10. doi:10.1038/srep17958

762 36. Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, McDonald ML, et al. A genome-
763 wide association study identifies risk loci for spirometric measures among smokers of
764 European and African ancestry. BMC Genet. 2015;16: 1–11. doi:10.1186/s12863-015-
765 0299-4

766 37. Byrne EM, Gehrman PR, Medland SE, Nyholt DR, Heath AC, Madden PAF, et al. A
767 genome-wide association study of sleep habits and insomnia. Am J Med Genet Part B
768 Neuropsychiatr Genet. 2013;162: 439–451. doi:10.1002/ajmg.b.32168

769 38. Comuzzie AG, Cole SA, Laston SL, Voruganti VS, Haack K, Gibbs RA, et al. Novel
770 Genetic Loci Identified for the Pathophysiology of Childhood Obesity in the Hispanic
771 Population. PLoS One. 2012;7. doi:10.1371/journal.pone.0051954

772 39. Fischer TM, Gilmour AR, Werf J. Computing approximate standard errors for genetic
773 parameters derived from random regression models fitted by average information REML.
774 Genet Sel Evol. 2004;36: 363. doi:10.1186/1297-9686-36-3-363

40. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011;88: 294–305. doi:10.1016/j.ajhg.2011.02.002

41. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8: 1–10. doi:10.1038/s41467-017-01261-5

42. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: A new resource for understanding brain development and function. J Neurosci. 2008;28: 264–278. doi:10.1523/JNEUROSCI.4178-07.2008

43. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550: 204–213. doi:10.1038/nature24277

44. Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun. 2015;6. doi:10.1038/ncomms6890

45. Fehrmann RSN, Karjalainen JM, Krajewska M, Westra HJ, Maloney D, Simeonov A, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet. 2015;47: 115–125. doi:10.1038/ng.3173

46. Robinson MR, Kleinman A, Graff M, Vinkhuyzen AAE, Couper D, Miller MB, et al. Genetic evidence of assortative mating in humans. Nat Hum Behav. 2017;1: 1–13. doi:10.1038/s41562-016-0016

798

799