Little Evidence of Modified Genetic Effect of rs16969968 on Heavy Smoking Based on Age of Onset of Smoking

Short title: No rs16969968 x age of smoking initiation interaction effect on heaviness of smoking.

Christine Adjangba[1,‡], Richard Border, Ph.D.[1,2,†], Pamela N. Romero Villela[1,3], Marissa A. Ehringer, Ph.D.[1,4], and Luke M. Evans, Ph.D.[1,5*]

[1] Institute for Behavioral Genetics, University of Colorado Boulder
[2] Department of Applied Mathematics, University of Colorado Boulder
[3] Department of Psychology & Neuroscience, University of Colorado Boulder
[4] Department of Integrative Physiology, University of Colorado Boulder
[5] Department of Ecology & Evolutionary Biology, University of Colorado Boulder
[‡] Current address: Department of Biology, Duke University
[†] Current address: Department of Neurology, David Geffen School of Medicine, University of California Los Angeles
* Corresponding Author: Luke M. Evans, Ph.D., Institute for Behavioral Genetics, Boulder, CO 80303; luke.m.evans@colorado.edu

Word Count: 3,558

ABSTRACT

Tobacco smoking is the leading cause of preventable death globally. Smoking quantity, measured in cigarettes per day (CPD), is influenced both by the age of onset of regular smoking (AOS) and by genetic factors, including a strong effect of the non-synonymous single nucleotide polymorphism rs16969968. A previous study by Hartz et al. reported an interaction between these two factors, whereby rs16969968 risk allele carriers who started smoking earlier showed increased risk for heavy smoking compared to those who started later. This finding has yet to be replicated in a large, independent sample. We performed a preregistered, direct replication attempt of the rs16969968×AOS interaction on smoking quantity in 128,383 unrelated individuals from the UK Biobank, meta-analyzed across ancestry groups. We fit statistical association models mirroring the original publication as well as formal interaction tests on multiple phenotypic and analytical scales. We replicated the main effects of rs16969968 and AOS on CPD but failed to replicate the interaction using previous methods. Nominal significance of the rs16969968×AOS interaction term depended strongly on the scale of analysis and the particular phenotype, as did associations stratified by early/late AOS. No interaction tests passed genome-wide correction (α=5e-8), and all estimated interaction effect sizes were much smaller in magnitude than previous estimates. We failed to replicate the strong rs16969968×AOS interaction effect previously reported. If such gene-moderator interactions influence complex traits, they likely depend on scale of measurement, and current biobanks lack the power to detect significant genome-wide associations given the minute effect sizes expected.

IMPLICATIONS:

We failed to replicate the strong rs16969968×AOS interaction effect on smoking quantity previously reported. If such gene-moderator interactions influence complex traits, current biobanks lack the power to detect significant genome-wide associations given the minute effect sizes expected. Furthermore, many potential interaction effects are likely to depend on the scale of measurement employed.

INTRODUCTION

Approximately 20% of deaths every year in the United States can be attributed to cigarette smoking, and smokers have life expectancies at least 10 years shorter than nonsmokers[1]. Furthermore, the rise in use among adolescents of various electronic cigarettes has emerged as a potentially dangerous trend about which little is known regarding long-term health and addiction consequences[2]. There is strong evidence from adoption, family, and twin studies that both genetic and environmental factors contribute to risk for smoking behaviors, with heritability estimates for nicotine dependence, ever becoming a regular smoker, and smoking quantity ranging between 33% and 71%[3-7]. Recently, genome-wide association studies (GWAS) have identified common variants associated with smoking[8-13]. In particular, the nicotinic acetylcholine receptor subunit genes *CHRNA5-CHRNA3-CHRNB4* on chromosome 15 have been implicated by well-powered GWAS of smoking behaviors[12,14,15]. Within *CHRNA5*, which codes for the α5 receptor subunit, the nonsynonymous G/A single nucleotide polymorphism (SNP) rs16969968 has been replicated through both large-scale GWAS[8-13,16] and functional assays[17-20] to influence smoking quantity, as measured by the number of cigarettes smoked per day, and nicotine dependence. The rs16969968-A risk allele has the largest estimated allelic effect on smoking quantity known to date[12]. While GWAS have identified many additional smoking-associated variants, rs16969968 remains a focus of individual functional studies and genetic epidemiological studies, with 292 publications reporting analyses of rs16969968 indexed by dbSNP (https://www.ncbi.nlm.nih.gov/snp/rs16969968#publications) and 454 publications (198 within the last five years) listed on LitVar (www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar), at the time of this writing.

In addition to genetic risk factors for heavy smoking, earlier age at onset of regular smoking (AOS) is well-known to predict risk for later heavy use and nicotine dependence[21,22]. In light of previous findings, Hartz et al.[13] conducted a meta-analysis of 33,348 individuals across 43 European and American data sets to test whether genetic vulnerability to heavy smoking and nicotine dependence at rs16969968 depends on AOS, and found a strong, significant interaction between early AOS and the rs16969968-A allele on heavy smoking (OR=1.16). Additional studies[23,24] have focused on rs16969968 interactions with other variables, highlighting the continued interest in rs16969968 interactions on behavior. Notably, these include those evaluating rs16969968×age of nicotine exposure[20,25], and include a report that early intervention to prevent adolescent smoking reduces the genetic risk of rs16969968 for heavy smoking later in life, a gene-by-intervention

interaction[26]. The original finding of rs16969968×AOS has been referenced in reviews citing the need for evaluation of gene-by-environment (G×E) interaction effects on nicotine dependence [27], and suggesting that direct replication of methods is needed to rigorously evaluate G×E interactions on smoking[28]. However, despite the large rs16969968×AOS interaction effect size originally reported, animal model evidence to support the plausibility of such an interaction[20], and continued interest in rs16969968, we are aware of no large-scale replication attempt in an independent sample. Here, we assessed whether there is an rs16969968×age of onset of smoking interaction in a well-powered (Fig. S1), independent sample, in an attempt to directly replicate the original findings.

METHODS

We preregistered our analyses through the Open Science Framework (osf.io/ynh2j) after we had obtained the UK Biobank data, but before we analyzed CPD or AOS.

Study Population

We used the UK Biobank, a large sample with rich phenotype and genome-wide genotype data[29]. We included all participants with available genomic data who had reported CPD and AOS data. The participants were either current or former smokers aged 40 years or older. To avoid confounding influences of population stratification[30], we initially, and following our preregistration, performed analyses using only individuals of European (EUR) ancestry, the largest subsample within the UK Biobank, identified by those whose first scores on the first four principal components (PCs; UK Biobank data field ID 22009) fell within the range of the UK Biobank identified individuals of European ancestry (field 22006). Following this, we expanded our analysis to include all available individuals within the UK Biobank. We identified relatively genetically homogeneous groups of individuals within the UK Biobank after excluding the EUR-ancestry individuals noted above using K-means clustering, from K=2-10, applied to the first 10 PC axes (data field ID 22009). The percent variance explained plateaued at K=10 clusters (Fig. S2). All analyses were subsequently performed within these 11 genetic clusters (EUR-ancestry + K=10 clusters). We note that the purpose of this clustering was solely to identify relatively genetically homogeneous groups of individuals within which to perform association analyses, and not to make population genetic inferences.

We only included unrelated individuals in our primary analyses to avoid possible confounding due to

shared environmental factors. Relatedness was estimated within each genetic cluster using MAF- and LD-pruned array markers (plink2[31] command: --maf 0.01 --hwe 1e-8 --indep-pairwise 50 5 0.2) after excluding those individuals with self-report and genetic sex mismatch (fields 31 and 22001), those with unusually high inbreeding coefficients ($|F_{het}| > 0.2$), and those identified by the UK Biobank and Affymetrix as having poor-quality genomic data (fields 220010 and 22051). Unrelated individuals (estimated relatedness<0.05) were identified with GCTA[32] v1.91.3 within each cluster. After removing individuals with missing phenotype and covariate data (see below), a total of 128,383 unrelated individuals across all genetic clusters were included.

Variables

Smoking quantity, as measured by CPD, was the primary dependent variable in analyses. Data on CPD (fields 2887, 3456, and 6183) were obtained from current or former smokers by asking the question "About how many cigarettes do/did you smoke on average each day?" These data were highly skewed; therefore, we also analyzed $\log_{10}$-transformed CPD (Fig. S3). Because of observed evidence of scale dependence[33] (see results below), we also analyzed heavy/light CPD on an additive scale. These two additional procedures were the only deviations from our preregistered analyses. Final analyses considered untransformed CPD, $\log_{10}$(CPD), heavy/light (analyzed on both multiplicative, i.e., logistic, and additive scales), and binned encodings. The dichotomous encoding defined smoking quantity as light smoking (CPD ≤ 10) versus heavy smoking (CPD > 20), mirroring the definition used by Hartz et al. The binned encoding defined smoking quantity as a linear variable consisting of 0 (CPD ≤ 10), 1 (11-20 CPD), 2 (21-30 CPD), or 3 (CPD > 30), also matching their secondary analysis.

Age of onset of regular smoking (AOS) was determined from fields 3426 and 2867, where participants were asked "How old were you when you first started smoking on most days?" AOS was analyzed based on a dichotomous encoding, a binned encoding, and the raw AOS data, again replicating the methods of Hartz et al.[13]. The dichotomous encoding defined early as AOS ≤ 16 years and late as AOS > 16 years. The binned encodings were 0 (AOS ≤ 15 years), 1 (AOS = 16 years), 2 (17-18 years AOS), or 3 (AOS > 18 years). We note that, matching Hartz et al., the median AOS was 16 in the UK Biobank (Fig. S3), making a reasonable and comparable age at which to separate early vs. late initiating smokers.

Covariates included were sex (field 31), age at time of assessment (field 21003), age^2, Townsend

Deprivation Index (field 21003), educational attainment ("qualification", categorical, field 6138), genotyping batch (field 22000), assessment center (field 54), and the first 10 genetic principal components as estimated with flashpca[34] applied to the MAF- and LD-pruned SNPs as described above. High collinearity of covariates within this sample resulted in a rank-deficient design matrix, which we addressed by performing a principal components analysis of the c=141 fixed effects using the prcomp function in R v3.2.2[35]. We then estimated the rank of the resulting eigenvector matrix (rank $r < c$) using the *matrix* R package[36] and included the first r=140 principal components as covariates in all analyses.

Statistical Analyses

All statistical analyses were performed within each genetic ancestry cluster separately. For dichotomized light/heavy CPD, we performed logistic regression using *glm* (family='binomial') in R[35] to assess the multiplicative scale interaction. The model included the rs16969968 genotype (coded as 0, 1, or 2), AOS, and rs16969968×AOS. All genotype×covariate and AOS×covariate interactions were included within the models to appropriately control for confounding[37]. For continuous variables (raw, binned, and log-transformed CPD) and the additive scale interaction model of the dichotomous heavy/light phenotype, we tested the same model using linear regression with the R *lm* function. Because many of the non-EUR-ancestry clusters had relatively few unrelated individuals within them, including all 140 covariates and their interactions resulted in a model that could not be fitted. We therefore reduced the number of covariates to be the scores from the first five PC scores of the covariate design matrix for the K=10 non-EUR-ancestry clusters.

The above model varied from that tested by Hartz et al., who tested rs16969968 effects on smoking phenotypes stratified by AOS (early versus late), using logistic regression (i.e., multiplicative scale). To recapitulate their methods, we performed secondary association tests of rs16969968 stratified by early versus late AOS using BOLT-LMM v2.3.2[38], with 339,444 genome-wide SNPs (quality control as described above, but without LD-pruning) to control for cryptic relatedness. All covariates were included in the BOLT-LMM models, excluding interaction terms. Because BOLT-LMM is not recommended for samples of less than 5,000 (see documentation from ref. [38]), we used GCTA leave-one-chromosome-out (--mlma-loco) approach[39] for the non-EUR-ancestry genetic clusters. Finally, to directly replicate previous methods, we performed AOS-stratified

logistic regression of heavy/light CPD using only rs16969968 and sex as independent variables.

We meta-analyzed the results using the inverse variance weighting approach in METAL[40]. We report meta-analyzed results below, and all cluster-specific results in the Supplementary Material.

We also performed several power analyses, to determine the power to detect the previously reported effect size[13], as well as to determine the sample size needed to achieve 80% power at specified effect sizes and α. To estimate the power to detect the previously reported effect size in the UK Biobank sample under a multiplicative scale interaction model, we simulated 61,077 diploid genotypes and early/late AOS in R, with linear predictors simulated using the previously reported main effect sizes as,

$$\mathrm{lp} = 0.33 + \log(1.28)g + \log(2.63)a + \log(\mathrm{OR}_{\mathrm{AOS*rs16969968}})ag \quad (1)$$

where genotypes, *g*, were simulated from a binomial distribution with MAF=0.34, the observed frequency of the A allele in the UK Biobank, early versus late AOS status, *a*, was randomly assigned to individuals. We varied the interaction effect size, $\log(\mathrm{OR}_{\mathrm{AOS*rs16969968}})$ between 0.005 and 0.4, reflecting a range of plausible effect sizes and encompassing the previously reported interaction effect (OR=1.16). Binary phenotypes, y, were then simulated in R as,

y = rbinom(61077,1,exp(lp)/(1 + exp(lp))). (2)

For each simulated interaction effect size, we performed 1,000 replicate simulations, estimating the interaction effect using logistic regression as above, and recorded the number of observations with an interaction *p*-value below either nominal significance, α=0.05, or genome-wide significance, α=5e-8. We performed similar simulations with the main AOS and rs16969968 effect sizes estimated within the UK Biobank (see below). We varied the sample size from 1e3 to 2e6, varying interaction effect size (previously reported $\mathrm{OR}_{\mathrm{AOS*rs16969968}}$=1.16 versus our meta-analyzed estimate $\mathrm{OR}_{\mathrm{AOS*rs16969968}}$=1.004), and nominal versus genome-wide significance thresholds (α=0.05 versus 5e-8, respectively).

RESULTS

We observed significant main effects of the rs16969968 A allele and AOS on CPD (Figures 1A, 1B, S4; Tables 1, S1-2). When estimated as predictors of heavy vs. light smoker status, the meta-analyzed estimated genetic effect, $\mathrm{OR}_{\mathrm{rs16969968}}$=1.12 (*p*=4.8e-28), was similar to but lower than the previous estimate[13] of 1.28. The effect of early AOS, $\mathrm{OR}_{\mathrm{AOS}}$=1.19 (*p*=3.6e-45), was less than previously reported[13] ($\mathrm{OR}_{\mathrm{AOS}}$=2.63). However,

both main effects were significantly associated with CPD in the expected direction, regardless of the CPD or AOS encoding, and represent strong evidence that both the rs16969968 A allele and early AOS are positively associated with heavier smoking, replicating previous findings. We note that, like previous findings[12,13], rs16969968 is not associated with AOS (Fig. S3D)

Conversely, the interaction between rs16969968 genotype and AOS was only nominally significant (α=0.05) and only in some combinations of CPD and AOS encodings (Figures 1C, S4; Tables 1, S1-2). Specifically, when treating both CPD and AOS as binary phenotypes the logistic model interaction was not significant ($OR_{rs16969968\times AOS}$=1.004, *p*=0.82) and the effect was notably lower than the previously reported estimate of 1.16. Interestingly, the interaction effect was nominally significant (p<0.05) for the binned CPD phenotype and dichotomized AOS, and when heavy/light CPD was analyzed on the additive scale, but not when the CPD phenotype was either heavy vs. light analyzed on the multiplicative scale model or when CPD was log-transformed. Across all tests and all CPD and AOS encodings, no interaction effects reached genome-wide significance (*p*>0.028).

Associations of rs16969968 stratified by AOS also produced mixed results. 95% confidence intervals (α=0.05) of the meta-analyzed effect sizes were non-overlapping only for binned and binary CPD encodings (Fig. 2, Table S3). When examining meta-analyzed genetic effects of rs16969968 on heavy versus light CPD, OR_{Early}/OR_{Late} was much lower than previously reported (OR_{Early}/OR_{Late}=1.016, Table S3). Within the largest ancestry cluster (EUR-ancestry), 95% CIs of the rs16969968 effects were non-overlapping in early vs. late AOS individuals for all CPD encodings except $\log_{10}$(CPD) (Fig S5; Table S3-S5). For all other ancestry clusters, we found no evidence of different rs16969968 effects using either a genome-wide (α=5e-8) or nominal (α=0.05) significance threshold (Table S4).

The direct replication test using Hartz et al.[13] methods with only rs16969968 and sex as independent variables found no evidence of different allelic effects between early and late smokers (p=0.41; Table S6-S8).

Our power analyses yielded two main results. First, our sample was well powered (>99%) to detect an interaction effect of the size previously reported at nominal significance (Figs. 3, S1), though not at genome-wide significance (power ~5%), even with over 61,000 subjects. Second, a sample drastically larger than that analyzed here would be required to detect an interaction effect of $OR_{rs16969968\times AOS}$=1.004, as estimated within our sample, with 80% power at α=0.05 (Fig. 3). Even applying the upper 95% CI limit of our estimate

($OR_{rs16969968\times AOS}$=1.035), or the estimate within the largest genetic cluster (1.03) would require a sample of approximately 1.3-2 million participants to achieve 80% power at a genome-wide threshold (α=5e-8) (Figs. S6, S7).

DISCUSSION

We replicated the substantial main effects of rs16969968 and early age of onset of smoking on CPD, across all phenotypic and analytical scales (Table 1). Estimates were in the same direction and of roughly similar magnitude as those previously reported[13].

Conversely, we found limited evidence of an rs16969968×AOS interaction effect. Formal interaction model results were mixed and depended heavily on measurement scale and phenotype encoding. Notably, our attempt to directly replicate the methods of Hartz et al. failed to identify a significant difference in the rs16969968-A allele effect on heavy smoking between early and late AOS (p=0.41; Table S6). This is similar to the results from stratified linear mixed model analyses, where the genetic effect in early AOS individuals was 1.016-fold higher than in late AOS individuals, despite greater statistical power of linear mixed models[39], as well as more control of potential confounding variables, such as genetic ancestry and geographic variation throughout the UK. Patterns within the largest genetic ancestry cluster, individuals with primarily EUR ancestry, were similar to the trans-ethnic meta-analyzed results. While some tests of differences in the stratified associations did reach nominal significance, the results suggest only minute differences in rs16969968 effects between early and late initiating smokers (Table S3). Across multiple analytic frameworks and phenotype encodings, the majority of our results were incongruent with an interaction between rs16969968 and AOS.

Magnitude of Effects and Power

No interaction test, and no comparison of stratified estimates, reached genome-wide significance (α=5e-8) despite the comparatively large sample size of our study. With genome-wide genotyping arrays and imputation commonly applied[41], and as genome-wide interaction associations and heritability studies have become more frequent[42-48], focusing on genome-wide significance thresholds is paramount to avoid false positives, even in situations where there are *a priori* hypotheses of interaction, as in rs16969968×AOS. Applying sufficiently stringent significance thresholds in initial studies, whether genome-wide 5e-8 or another

specified threshold, is a best practice for replication of association studies[49], and we believe that as GWAS interactions (including G×E and G×G) studies become more frequent, the question of applicable significance thresholds should be revisited.

In all tests related to interaction effects and stratified associations, the estimated interaction effect sizes were much smaller than previously reported[13]. Despite substantially greater power than the original study, which had a sample size of ~30,000, (Fig. 4, S6-S7), we estimated the effect to be only 1.004 (or 1.016 in the stratified associations). The lack of replication when using the exact same methods suggests that there is no true interaction at this locus. It is important to recognize that both replicated main effects were strong, significant, and in the expected direction, reflecting the strongest single-locus genetic effect on CPD[12] and a strong, consistent risk factor of heavy smoking (early age of initiation). This suggests that if the interaction were to exist, its effect would be much less than previously expected. Importantly, with an OR=1.004, it would be insignificant for possible clinical interventions, such as targeted smoking awareness based on rs16969968 genotype[26].

The discrepancy between our results and those reported by Hartz et al.[13] could additionally reflect differences between the study populations and models used for analyses. The study by Hartz et al.[13] exemplified a tremendous effort to collect the largest available sample size at the time. They were able to do so by meta-analyzing multiple individual studies together, a highly coordinated endeavor that must be recognized and applauded. One possible outcome of this approach is heterogeneity of effect estimates, which they found and noted. Our analysis focused on a single, relatively homogeneous dataset instead of many studies, removing potential heterogeneity that could have influenced the previous results. Our meta-analysis of relatively homogenous ancestry clusters also attempted to minimize any confounding of stratification. Second, although 33% of the Hartz et al. data were European datasets, consistent cultural differences may exist between American and UK samples, such as general attitudes towards smoking, and any potential impact would be difficult to assess. Such differences between samples could lead to true heterogeneity in the effects[50] and the different estimated effects we observed, though Hartz et al. reported no significant difference between OR estimates from American versus EU studies. A possible source of bias, in both the initial and the current study, is that of collider bias[51,52]. The UK Biobank is healthier and wealthier[51] than the general UK Population, leading to ascertainment and the potential for colliders. Genetic effects on education could lead to false

negative genetic associations in the UK Biobank with smoking traits when controlling for education[51], but we view this as an unlikely explanation for our failure to replicate results, as both main effects were replicated, and because whether we controlled for education or not, we found little evidence of rs16969968×AOS interaction. Selection bias in general could lead to false positive or negative associations. Additional methodological differences include testing a full statistical interaction model with complete covariate×AOS and covariate×genotype terms and using a linear mixed model in our stratified analyses, neither of which were previously employed. Mixed model approaches generally improve power[39], and including the covariate interaction terms should lead to unbiased estimates of the rs16969968×AOS interaction[37]. On the other hand, comparing estimates across different subsamples, as in stratified linear mixed model analysis, introduces an additional potential source of confounding. However, the respective strengths and weaknesses of these methods cannot account for our failure to directly replicate the original finding; our stratified association tests with only sex as a covariate (mirroring the approach of Hartz et al.) failed to identify significant differences in allelic effect sizes between early and late AOS individuals (p=0.41; Table S6), despite being well-powered to do so.

Regardless, with respect to particular phenotype encodings and analyses (e.g., stratified analyses of heavy vs. light smoker status, with linear mixed models), we did find nominally significant, very small differences in allelic effect size estimates between early- and late-onset smokers. These findings are thus potentially congruent with a small interaction between rs1696968 and AOS. If there is a true rs16969968×AOS interaction of roughly the magnitude we estimated (OR=1.004), it would a drastically larger sample size to detect it (Fig. 4). We must therefore conclude that any such interactions specific to an individual locus are likely of very small effect, will be very difficult to identify even with the largest available biobanks, and likely contribute minimally to phenotypic variance.

Conclusions

We found limited support for the rs16969968×AOS interaction. To the extent that AOS might moderate the effect of rs16969968, we estimate this effect to be far smaller than previously reported. We suggest that even larger sample sizes will be required to identify, with genome-wide significance, interactions at individual loci given the expected magnitude of the interaction effects. On the other hand, our unambiguous replications of

the main effects of both rs16969968 and AOS on smoking quantity support epidemiological evidence that individuals who begin regularly smoking at a young age are at a higher risk for nicotine dependence later in life[21,22]. This provides further evidence in support of public health interventions for adolescent smoking that could help reduce tobacco use, which would in turn lower the number of tobacco-related deaths and illnesses.

ACKNOWLEDGEMENTS

Ms. Adjangba was supported by the Summer Multicultural Access to Research Training Program at the University of Colorado. Drs. Evans and Border are supported by National Institute of Mental Health R01 MH100141-06 (PI: Matthew C. Keller) and Dr. Evans is supported by National Institute on Drug Abuse R01 DA044283-01A1 (PI: Scott I. Vrieze) and National Institute on Aging R01 AG046938 (PI: C.A. Reynolds/S.M. Wadsworth).

Conflict of Interest Disclosures: The authors declare no conflict of interest.

REFERENCES

1. US Department of Health and Human Services. Health Consequences of Smoking—50 Years of Progress A Report of the Surgeon General. *Report of the Surgeon general.* 2014;1081.
2. Fadus MC, Smith TT, Squeglia LM. The rise of e-cigarettes, pod mod devices, and JUUL among youth: Factors influencing use, health implications, and downstream effects. *Drug Alcohol Depend.* 2019;201:85-93.
3. Haberstick BC, Ehringer MA, Lessem JM, Hopfer CJ, Hewitt JK. Dizziness and the genetic influences on subjective experiences to initial cigarette use. *Addiction.* 2011;106(2):391-399.
4. Haberstick BC, Zeiger JS, Corley RP, et al. Common and drug-specific genetic influences on subjective effects to alcohol, tobacco and marijuana use. *Addiction.* 2011;106(1):215-224.
5. Kaprio J. Genetic epidemiology of smoking behavior and nicotine dependence. *COPD.* 2009;6(4):304-306.
6. Rose R.J., Broms U., Korhonen T., Dick D.M., J. K. Genetics of Smoking Behavior. In: YK K, ed. *Handbook of Behavior Genetics*. New York, NY: Springer; 2009.
7. Kendler KS, Schmitt E, Aggen SH, Prescott CA, Virginia V. Genetic and Environmental Influences on Alcohol, Caffeine, Cannabis, and Nicotine Use From Early Adolscence to Middle Adulthood. *Arch Gen Psychiatry.* 2008;65:674-682.
8. Hancock DB, Guo Y, Reginsson GW, et al. Genome-wide association study across European and African American ancestries identifies a SNP in DNMT3B contributing to nicotine dependence. *Mol Psychiatry.* 2018;23(9):1911-1919.
9. Hancock DB, Reginsson GW, Gaddis NC, et al. Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence. *Transl Psychiatry.* 2015;5:e651.

10. Saccone NL, Emery LS, Sofer T, et al. Genome-Wide Association Study of Heavy Smoking and Daily/Nondaily Smoking in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Nicotine Tob Res.* 2018;20(4):448-457.
11. Wen L, Yang Z, Cui W, Li MD. Crucial roles of the CHRNB3-CHRNA6 gene cluster on chromosome 8 in nicotine dependence: update and subjects for future research. *Transl Psychiatry.* 2016;6(6):e843.
12. Liu M, Jiang Y, Wedow R, 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(2):237-244.
13. Hartz SM, Short SE, Saccone NL, et al. Increased genetic vulnerability to smoking at CHRNA5 in early-onset smokers. *Arch Gen Psychiatry.* 2012;69(8):854-860.
14. Bierut LJ, Stitzel JA, Wang JC, et al. Variants in nicotinic receptors and risk for nicotine dependence. *Am J Psychiatry.* 2008;165(9):1163-1171.
15. Berrettini W, Yuan X, Tozzi F, et al. Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol Psychiatry.* 2008;13(4):368-373.
16. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet.* 2010;42(5):441-447.
17. Bailey CD, Tian MK, Kang L, O'Reilly R, Lambe EK. Chrna5 genotype determines the long-lasting effects of developmental in vivo nicotine exposure on prefrontal attention circuitry. *Neuropharmacology.* 2014;77:145-155.
18. Kuryatov A, Berrettini W, Lindstrom J. Acetylcholine receptor (AChR) alpha5 subunit variant associated with risk for nicotine dependence and lung cancer reduces (alpha4beta2)(2)alpha5 AChR function. *Mol Pharmacol.* 2011;79(1):119-125.
19. George AA, Lucero LM, Damaj MI, Lukas RJ, Chen X, Whiteaker P. Function of human alpha3beta4alpha5 nicotinic acetylcholine receptors is reduced by the alpha5(D398N) variant. *J Biol Chem.* 2012;287(30):25151-25162.
20. O'Neill HC, Wageman CR, Sherman SE, Grady SR, Marks MJ, Stitzel JA. The interaction of the Chrna5 D398N variant with developmental nicotine exposure. *Genes Brain Behav.* 2018;17(7):e12474.
21. Lydon DM, Wilson SJ, Child A, Geier CF. Adolescent brain maturation and smoking: what we know and where we're headed. *Neurosci Biobehav Rev.* 2014;45:323-342.
22. 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(4):408-413.
23. Adrian M, Kiff C, Glazner C, et al. Examining gene-environment interactions in comorbid depressive and disruptive behavior disorders using a Bayesian approach. *J Psychiatr Res.* 2015;68:125-133.
24. Schneider KK, Hule L, Schote AB, Meyer J, Frings C. Sex matters! Interactions of sex and polymorphisms of a cholinergic receptor gene (CHRNA5) modulate response speed. *Neuroreport.* 2015;26(4):186-191.
25. Grucza RA, Johnson EO, Krueger RF, et al. Incorporating age at onset of smoking into genetic models for nicotine dependence: evidence for interaction with multiple genes. *Addict Biol.* 2010;15(3):346-357.
26. Vandenbergh DJ, Schlomer GL, Cleveland HH, et al. An Adolescent Substance Prevention Model Blocks the Effect of CHRNA5 Genotype on Smoking During High School. *Nicotine Tob Res.* 2016;18(2):212-220.
27. Dick DM, Barr PB, Cho SB, et al. Post-GWAS in Psychiatric Genetics: A Developmental Perspective on the "Other" Next Steps. *Genes Brain Behav.* 2018;17(3):e12447.
28. Do EK, Maes HH. Genotype x Environment Interaction in Smoking Behaviors: A Systematic Review. *Nicotine Tob Res.* 2017;19(4):387-400.
29. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726):203-209.
30. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38(8):904-909.
31. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.
32. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88(1):76-82.
33. VanderWeele TJ, Knol MJ. A Tutorial on Interaction. *Epidemiologic Methods.* 2014;3(1):33-72.
34. Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. *PLoS One.* 2014;9(4):e93766.

35. *R: A language and environment for statistical computing.* [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2015.
36. *Matrix: Sparse and dense matrix classes and methods. R package version 1.2-2.* [computer program]. 2015.
37. Keller MC. Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol Psychiatry.* 2014;75(1):18-24.
38. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. *Nat Genet.* 2018;50(7):906-908.
39. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet.* 2014;46(2):100-106.
40. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics.* 2010;26(17):2190-2191.
41. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* 2016;48(10):1279-1283.
42. Rawlik K, Canela-Xandri O, Tenesa A. Evidence for sex-specific genetic architectures across a spectrum of human complex traits. *Genome Biol.* 2016;17(1):166.
43. Young AI, Wauthier FL, Donnelly P. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nat Genet.* 2018;50(11):1608-1614.
44. Dahl A, Nguyen K, Cai N, Gandal MJ, Flint J, Zaitlen N. A Robust Method Uncovers Significant Context-Specific Heritability in Diverse Complex Traits. *Am J Hum Genet.* 2020;106(1):71-91.
45. Peterson RE, Cai N, Dahl AW, et al. Molecular Genetic Analysis Subdivided by Adversity Exposure Suggests Etiologic Heterogeneity in Major Depression. *Am J Psychiatry.* 2018;175(6):545-554.
46. Arnau-Soler A, Adams MJ, Generation S, Major Depressive Disorder Working Group of the Psychiatric Genomics C, Hayward C, Thomson PA. Genome-wide interaction study of a proxy for stress-sensitivity and its prediction of major depressive disorder. *PLoS One.* 2018;13(12):e0209160.
47. Nivard MG, Middeldorp CM, Lubke G, et al. Detection of gene–environment interaction in pedigree data using genome-wide genotypes. *European Journal of Human Genetics.* 2016;24(12):1803-1809.
48. Robinson MR, English G, Moser G, et al. Genotype-covariate interaction effects and the heritability of adult body mass index. *Nat Genet.* 2017;49(8):1174-1181.
49. NCI-NHGRI Working Group on Replication in Association Studies, Chanock SJ, Manolio T, et al. Replicating genotype-phenotype associations. *Nature.* 2007;447(7145):655-660.
50. Konig IR. Validation in genetic association studies. *Brief Bioinform.* 2011;12(3):253-258.
51. Munafo MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol.* 2018;47(1):226-235.
52. Pingault JB, O'Reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. *Nat Rev Genet.* 2018;19(9):566-580.

Table 1. Trans-ethnic meta-analysis estimated main and interaction effects (β) and standard errors (SE) for rs16969968, age of smoking initiation (AOS), and their interaction. Shown are estimates for each encoding of CPD and AOS.

			AOS			rs16969968_A			rs16969968 x AOS			
CPD Coding	N	AOS Coding	β[a]	SE	p	β[a]	SE	p	β[a]	SE	p	
CPD (raw)	128383	raw	-0.296	1.15E-02	9.34E-147	1.271	2.17E-01	4.65E-09	-0.019	1.24E-02	1.17E-01	
		binned	-0.830	3.59E-02	3.89E-118	1.042	6.53E-02	2.57E-57	-0.063	3.81E-02	1.01E-01	
		Early/Late	1.589	8.18E-02	4.99E-84	0.861	5.97E-02	3.41E-47	0.190	8.71E-02	2.87E-02	*
log_{10}(CPD)	128383	raw	-0.008	2.88E-04	6.91E-159	0.026	5.36E-03	1.55E-06	0.000	3.07E-04	7.86E-01	
		binned	-0.021	8.83E-04	1.80E-129	0.025	1.60E-03	4.22E-54	0.000	9.34E-04	8.19E-01	
		Early/Late	0.040	2.01E-03	1.95E-86	0.023	1.46E-03	3.40E-56	0.003	2.13E-03	2.12E-01	
binned	128383	raw	-0.023	9.39E-04	1.10E-130	0.102	1.77E-02	9.89E-09	-0.001	1.01E-03	2.03E-01	
		binned	-0.065	2.93E-03	3.07E-109	0.089	5.32E-03	4.94E-63	-0.006	3.11E-03	6.59E-02	
		Early/Late	0.126	6.67E-03	2.55E-79	0.074	4.87E-03	1.26E-52	0.015	7.10E-03	3.79E-02	*
Heavy/Light	61077	raw	-0.018	1.41E-03	1.15E-38	0.116	3.94E-02	3.16E-03	0.000	2.16E-03	9.00E-01	
		binned	-0.077	4.92E-03	1.23E-54	0.146	1.27E-02	2.78E-30	-0.001	6.80E-03	8.30E-01	
		Early/Late	0.171	1.21E-02	3.65E-45	0.116	1.05E-02	4.82E-28	0.004	1.62E-02	8.22E-01	
Heavy/Light (additive scale)	61077	raw	-0.013	7.19E-04	6.59E-76	0.094	1.43E-02	5.02E-11	-0.002	8.15E-04	2.87E-02	*
		binned	-0.044	2.23E-03	2.40E-88	0.070	4.24E-03	2.40E-61	-0.005	2.45E-03	5.70E-02	*
		Early/Late	0.090	5.17E-03	4.34E-67	0.058	3.88E-03	7.84E-50	0.012	5.69E-03	3.41E-02	

[a] β refers to the regression slope. For CPD coded as heavy/light, exp(β) is the Odds Ratio (OR) when analyzed on the multiplicative scale.

Figure 1. Main effects of rs16969968 genotype (A) and AOS (B) on cigarettes per day (CPD or $\log_{10}$(CPD)), and rs16969968 effect on CPD and $\log_{10}$(CPD) as a function of early or late age of initiation (C).

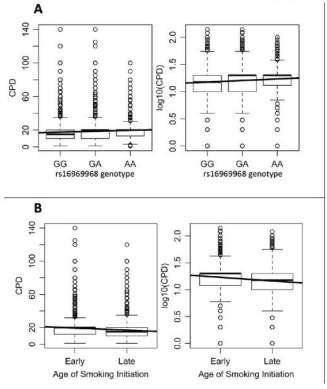

Figure 2. Trans-ethnic meta-analyzed allelic effect size estimates, β, of the rs16969968 risk allele, A, estimated in stratified association analyses by early and late age of initiation, with CIs indicated using either α=0.05 (solid line) or genome-wide Bonferroni-corrected α=5e-8 (dashed line). For the heavy vs. light smoker phenotype, allelic effects (β) were transformed to OR using the BOLT-LMM[38]-suggested transformation, of $e^{\beta / (u(1-u))}$, where u=0.42 is the proportion of cases. Effect size difference Z-test p-values are shown for each comparison. See Table S3 for estimates and statistics.

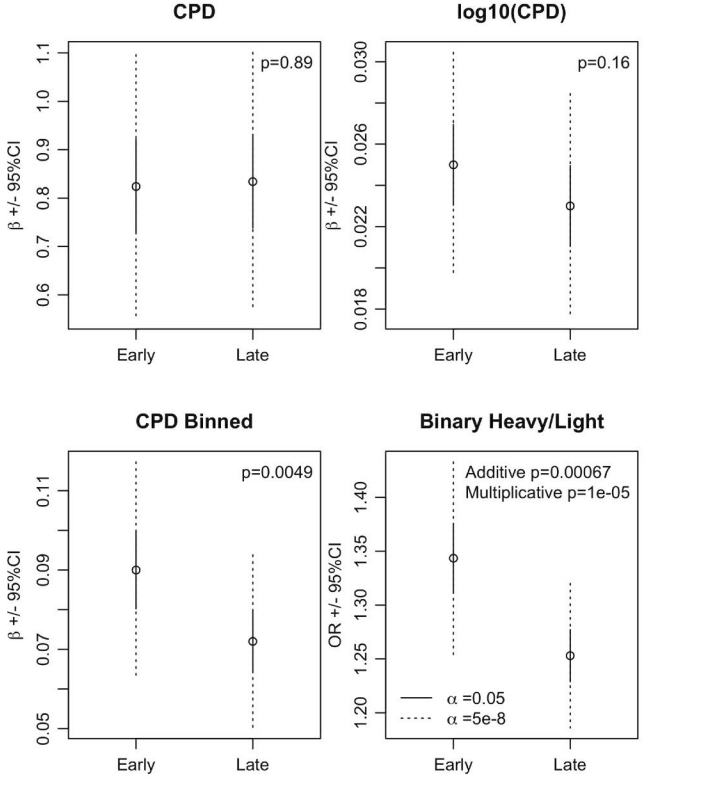

Figure 3. Power to detect the statistical interaction effect across a range of sample sizes for the interaction effect size estimated previously (OR=1.16) and from the current study (OR=1.004), applying either a nominal α=0.05 or genome-wide Bonferroni-corrected α=5e-8.