Genetic architecture of smoking: Evaluating rare variant contribution from deep whole-genome sequencing of up to 26,000 individuals

Seon-Kyeong Jang
University of Minnesota

Luke Evans

Allison Fialkowski
University of Minnesota

Donna Arnett
University of Kentucky College of Public Health

Diane Becker
Johns Hopkins University School of Medicine

Joshua Bis
University of Washington

John Blangero
Department of Human Genetics, University of Texas Rio Grande Valley School of Medicine

Eugene Bleecker
Department of Medicine, University of Arizona

Jennifer Brody
University of Washington

L. Adrienne Cupples
Boston University School of Public Health

Scott Damrauer
Department of Surgery, Perelman School of Medicine, University of Pennsylvania

Sean David
University of Chicago

Mariza de Andrade
Department of Health Sciences Research, Mayo Clinic

Tasha Fingerlin
Colorado School of Public Health, University of Colorado Denver - Anschutz Medical Campus

Sina Gharib
University of Washington

David Glahn
Boston Children's Hospital

Jeffrey Haessler
Fred Hutchinson Cancer Research Center

Susan Heckbert
University of Washington

John Hokanson
University of Colorado Anschutz Medical Campus

Shih-Jen Hwang
Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health

Matthew Hyman
University of Pennsylvania

Renae Judy
University of Pennsylvania

Anne Justice
Department of Population Health Sciences, Geisinger Health System

Robert Kaplan
Department of Epidemiology and Population Health, Albert Einstein College of Medicine

Wonji Kim
Channing Division of Network Medicine, Brigham and Women&#39;s Hospital and Harvard Medical School

Charles Kooperberg
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center

Dan Levy
Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health

Ruth Loos
Ichan School of Medicine at Mount Sinai https://orcid.org/0000-0002-8532-5087

Ani Manichaikul
Center for Public Health Genomics, School of Medicine, University of Virginia

Mark Gladwin
University of Pittsburgh School of Medicine

Lisa Martin
Division of Cardiology, School of Medicine and Health Sciences

Mehdi Nouraie
University of Pittsburgh School of Medicine

Olle Melander
Lund University

Deborah Meyers
Department of Medicine, University of Arizona

Kari North
University of North Carolina at Chapel Hill, Chapel Hill

Elizabeth Oelsner
Division of General Medicine, Columbia University Irving Medical Center, Columbia University

Anna Peljto
Department of Medicine, University of Colorado School of Medicine

Michael Preuss
Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai

Bruce Psaty
Cardiovascular Health Research Unit, Departments of Epidemiology, Medicine and Health Services, University of Washington, Seattle, WA

Dandi Qiao
Channing Division of Network Medicine, Brigham and Women&#39;s Hospital and Harvard Medical School

Daniel Rader
Department of Genetics, Perelman School of Medicine, University of Pennsylvania

Robert Reed
University of Maryland School of Medicine

Alexander Reiner
Fred Hutchinson Cancer Research Center

Stephen Rich
Center for Public Health Genomics, School of Medicine, University of Virginia

Jerome Rotter
The Institute for Translational Genomics and Population Sciences

David Schwartz
University of Colorado Denver

Aladdin Shadyab
University of California

Edwin Silverman
Channing Division of Network Medicine, Brigham and Women&#39;s Hospital and Harvard Medical School

Nicholas Smith
University of Washington

Gustav Smith
Gothenburg University

Albert Smith
University of Michigan, Ann Arbor

Weihong Tang
University of Minnesota

Kent Taylor

The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center

Ramachandran Vasan
Boston University School of Medicine

Victor Gordeuk
University of Illinois at Chicago

Zhe Wang
Icahn School of Medicine at Mount Sinai

Kerri Wiggins
University of Washington https://orcid.org/0000-0003-2749-1279

Lisa Yanek
Johns Hopkins University School of Medicine

Ivana Yang
University of Colorado School of Medicine

Kendra Young
University of Colorado Anschutz Medical Campus

Kristin Young
University of North Carolina at Chapel Hill

Yingze Zhang
Department of Medicine, University of Pittsburgh School of Medicine

Dajiang Liu
Penn State College of Medicine

Matthew Keller
University of Colorado

Scott Vrieze (vrieze@umn.edu)
University of Minnesota https://orcid.org/0000-0003-3861-7930

Article

Keywords: Genetic architecture, smoking, cigarettes, GCTA, missing heritability

Posted Date: May 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-475149/v1

Version of Record: A version of this preprint was published at Nature Human Behaviour on August 4th, 2022. See the published version at https://doi.org/10.1038/s41562-022-01408-5.

Genetic architecture of smoking: Evaluating rare variant contribution from deep whole-genome sequencing of up to 26,000 individuals

Authors:

Seon-Kyeong Jang[1], Luke Evans[2], Allison Fialkowski[1], Donna K. Arnett[3], Diane M. Becker[4], Joshua C. Bis[5], John Blangero[6], Eugene R. Bleecker[7], Jennifer A Brody[5], L. Adrienne Cupples[8], Scott M. Damrauer[9,10], Sean P. David[11,12], Mariza de Andrade[13], Tasha E. Fingerlin[14,15], Sina A. Gharib[5,16], David C Glahn[17], Jeffrey Haessler[18], Susan R. Heckbert[19,20], John E. Hokanson[21], Shih-Jen Hwang[22], Matthew C. Hyman[23], Renae Judy[9], Anne E. Justice[24], Robert C Kaplan[18,25], Wonji Kim[26], Charles Kooperberg[18], Dan Levy[22], Ruth J.F. Loos[27,28], Ani W. Manichaikul[29], Mark T. Gladwin[30], Lisa Warsinger Martin[31], Mehdi Nouraie[30], Olle Melander[32,33], Deborah A. Meyers[7], Kari E. North[34], Elizabeth C. Oelsner[35], Anna L. Peljto[36], Michael Preuss[27,28], Bruce M Psaty[37], Dandi Qiao[26], Daniel J. Rader[23,38], Robert M. Reed[39], Alexander P. Reiner[18], Stephen S. Rich[29], Jerome I. Rotter[40], David A. Schwartz[41,42], Aladdin H. Shadyab[43], Edwin K. Silverman[26], Nicholas L. Smith[19,20], J. Gustav Smith[44,45], Albert V. Smith[46], Weihong Tang[47], Kent D. Taylor[40], Ramachandran S. Vasan[48,49], Victor R. Gordeuk[50], Zhe Wang[27,28], Kerri L. Wiggins[5], Lisa R. Yanek[4], Ivana V. Yang[36], Kendra A. Young[21], Kristin L. Young[34], Yingze Zhang[30], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Dajiang J. Liu[51], Matthew Keller[2], Scott Vrieze[1]

Affiliations:

1 Department of Psychology, University of Minnesota, Minneapolis, MN, USA

2 Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, CO, USA

3 Dean's Office, University of Kentucky College of Public Health, Lexington, KY, USA

4 Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

5 Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA

6 Department of Human Genetics, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA

7 Department of Medicine, University of Arizona, Tucson, AZ, USA

8 Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA

9 Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

10 Department of Surgery, Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, USA

11 Department of Family Medicine, Prtizker School of Medicine, University of Chicago, Chicago, IL, USA

12 NorthShore University HealthSystem, Evanston, IL, USA

13 Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.

14 Colorado School of Public Health, University of Colorado Denver - Anschutz Medical Campus, Aurora, CO, USA

15 Center for Genes Environment and Health, National Jewish Health, Denver, CO, USA.

16 Center for Lung Biology, Division of Pulmonary, Critical Care and Sleep Medicine, University of Washington, Seattle, WA, USA

17 Department of Psychaitry, Boston Children's Hosptial and Harvard Medical School, Boston, MA, USA

18 Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

19 Department of Epidemiology, University of Washington, Seattle WA, USA

20 Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle WA, USA

21 Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

22 Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA

23 Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

24 Department of Population Health Sciences, Geisinger Health System, Danville, PA, USA

25 Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA

26 Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

27 Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

28 The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

29 Center for Public Health Genomics, School of Medicine, University of Virginia, Charlottesville, VA, USA

30 Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

31 Division of Cardiology, School of Medicine and Health Sciences, Washington, DC, USA

32 Department of Clinical Sciences, Lund University, Malmö, Sweden

33 Department of Internal Medicine, Skåne University Hospital, Malmö, Sweden

34 Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

35 Division of General Medicine, Columbia University Irving Medical Center, Columbia University, New York, NY, USA

36 Department of Medicine, University of Colorado School of Medicine, Aurora, CO, USA

37 Cardiovascular Health Research Unit, Department of Medicine, Epidemiology and Health Services, University of Washington, Seattle, WA, USA

38 Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

39 University of Maryland School of Medicine, Baltimore, MD, USA

40 The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA

41 Department of Medicine, School of Medicine, University of Colorado Denver, Aurora, CO, USA

42 Department of Immunology, School of Medicine, University of Colorado Denver, Aurora, CO, USA

43 Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, La Jolla, CA, USA

44 Wallenberg Laboratory/Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University, Sweden

45 Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden

46 Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA

47 Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA

48 Sections of Preventive medicine and Epidemiology and cardiovascular medicine, Department of medicine, Boston University School of Medicine, Boston, MA, USA

49 Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA

50 Department of Medicine, University of Illinois at Chicago, Chicago, IL, USA

51 Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA, USA

Funding:

R01DA044283, R01DA037904, R01HG008983 to SIV

R01MH100141 to MCK

Cohort-wise acknowledgement in Supplementary Note

Keywords:

Genetic architecture, smoking, cigarettes, GCTA, missing heritability

Acknowledgement

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). See the TOPMed Omics Support Table (Supplementary Note) for study specific omics support information. Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Abstract

Background: Across complex traits, common variants explain only a modest amount of variance, with SNP-heritability consistently below heritability estimates from close relatives. Here, we examined the contribution of rare variant to tobacco use risk in up to 26,000 individuals of European ancestry in the Trans-Omics for Precision Medicine (TOPMed) program with whole genome sequence (WGS;~30X coverage).

Method: We grouped about 35million genetic variants by their minor allele frequencies (MAF) and linkage disequilibrium (LD) and estimated SNP-heritability for age of smoking initiation (N=14,747), cigarettes smoked per day (N=15,425), smoking cessation (N=17,871) and initiation (N=26,340) using linear mixed model. Rare variant population structure is detected and adjusted for by permutation procedure. We estimated an upper bound for narrow-sense heritability for tobacco use using available pedigrees consisting of close relatives in TOPMed.

Results: Rare variants with MAF 0.1% to 0.01%, mostly from non-protein altering region, accounted for 26% of variation in age of initiation and 15% for cessation. Follow-up analysis indicated that about one-third of these rare variants contribtion is potentially confounded with rare variants structure even after adjusting for principal components. After further conservative adjustment of population structure, we estimated SNP-based heritability to be 0.21 (SE=0.08) for age of initiation, 0.15 (0.06) for cigarettes per day, 0.21 (0.09) for cessation, and 0.24 (0.07) for initiation, 1.8-4.5 times higher than previous SNP-based estimates. Our pedigree-based upper-bound for SNP-based heritability ranged from 0.18-0.35.

Conclusion: The substantial contribution of rare variants for several smoking phenotypes sheds light on the missing heritability and genetic etiology of tobacco use. It also informs fine-mapping strategies since the majority of the rare variant contribution was located in non-coding regulatory regions.

Introduction

Characterizing the genetic architecture of complex phenotypes has long been an important goal of genetic epidemiology, with implications for diverse fields including biology, medicine, and psychology. One aspect of this work involves characterizing the joint distribution of effect sizes and minor allele frequency (MAF), which is shaped by natural selection and population history[1,2]. Genome-wide association studies (GWAS) have discovered tens of thousands of genomic loci associated with complex phenotypes, providing new and basic insights into the genetic architecture of complex phenotypes, including rampant polygenicity[3]. Aggregating across loci typically explains only small fractions of phenotypic variance, fractions much lower than have been obtained in family-based studies (e.g., twins, or siblings, or larger pedigrees)[4]. This difference has been coined the "missing heritability".

Tobacco use is a complex behavioral trait of high public health concern[5], with demonstrated genetic and environmental (e.g., policy, cultural) influences. Not only does tobacco use influence risk of many diseases and represents a leading causes of global morbidity and mortality, but measures of tobacco use are strong indicators of addiction to nicotine (e.g., number of cigarettes smoked per day is genetically highly correlated (r=.95) with nicotine dependence[6]) and other commonly used substances. Heritability of smoking behaviors has been estimated at approximately 50% (SE 5%)[7] in twin studies, comparable to many other complex behavioral traits. On the other hand, estimates of tobacco use heritability from GWAS of single nucleotide polymorphisms (SNPs) have routinely found much lower SNP-based heritability (h^2_{SNP}) estimates[8,9]. Such analyses to date have been based on common variants (e.g., MAF > 1%) from GWAS of imputed microarrays. In a recent GWAS of tobacco use in up to 1.1 million individuals, Liu et al. reported h^2_{SNP} estimates ranging between 5% and 11%[8] with smoking initiation and age of smoking initiation showing the highest and the lowest common variants-based heritability, respectively[10]. Even more recently, Evans et al. reported h^2_{SNP} estimates of 5%-18% for smoking traits, using individual-level UK Biobank imputed genotypes of up to 323,068 individuals, finding that common variants contributed significantly to the overall heritability. The contribution of rare variant was minimal, although estimates were limited to imputed variants, which is not highly accurate for variants <1% MAF. Similar to results for other complex traits[7,11], some but far from all of the twin-based heritability of smoking can be attributed to common variants obtained through imputation of microarray genotypes.

There are many possible contributors to missing heritability, including inflated family-based heritability estimates[12], epistasis[13], structural variation[11], and rare variants[14]. Rare variants are one compelling explanation, as one expects negative selection to force strongly deleterious alleles to low frequencies[15]. Current SNP-based heritability estimates are based on a few million common variants. With imputed variants, the quality of imputation depends on the reference panel used[16,17], and even the best imputation strategies perform poorly for rare variants (e.g., MAF < 1%) in population samples[18,19]. With the advent of relatively affordable deep whole genome sequencing (WGS), it is now possible to directly genotype variants of any frequency. While genetic association studies may be underpowered to detect an association between a given single rare variant and a complex trait[20], we can test the contribution of rare variants in aggregate to phenotypic heritability over and above common variation[18,19] [20,21]. A small number of recent WGS studies reported evidence that low-frequency and rare variants contribute to heritability of anthropometric, transcriptomic, and medical phenotypes[22–25] (but see also [26,27]). Notably, Wainschtein et al[22]. reported that rare variants, especially those in regions of low LD, entirely captures the missing heritability for height and nearly so for BMI, albeit with large standard errors. To date, no previous study has used WGS in large population samples (e.g., >20,000) to estimate rare variant heritability for complex behavioral phenotypes, such as tobacco use.

Genetic principal components and kinship-based mixed models tend to reduce confounding due to stratification for tests of common variants[28,29]. However, it remains unclear the extent to which these techniques will satisfactorily work for rare variant analyses[30–32]. For example, rare variants may be more likely to be shared among individuals living in close proximity, possibly confounding heritability when rare genotype sharing coincides with geographically clustered environmental risk factors[33]. Thus, novel approaches are needed to control for population stratification in rare variant analyses[34–36].

Here, we used deep WGS from the Trans-Omics for Precision Medicine (TOPMed) initiative to evaluate the genetic architecture of four smoking phenotypes down to minor allele frequencies of 1 in 10,000 (MAF≥0.0001). We also evaluated issues of rare variant-based population structure using a new permutation method developed for rare variant associations[37,38].

Methods

Sample

We considered individuals of European ancestry in TOPMed (freeze 8, mean depth >30)[21] measured for at least one of four smoking phenotypes for inclusion. We determined European ancestry in two steps. First, we identified an initial ancestry-inclusive set by projecting TOPMed genotypes (N=137,977) onto genetic principal (PC) axes from the 1000 Genomes project[8] (1000G) then used a k-nearest neighbor method to assign ancestry of TOPMed individuals with 1000G as a reference set. More specifically, we used online augmentation-decomposition-transformation (OADP) to calculate PC scores of TOPMed individuals, which implements Procrustes transformation with an augmented data set (i.e., combining TOPMed and 1000G reference genomes together)[39]. Then, for a given TOPMed sample, we chose the top 20 reference individuals in 1000G who were closest in terms of the Euclidean distance of 20 PC scores and assigned European ancestry when at least 87.5% of the reference individuals had European ancestry (Supplementary Note). This resulted in 38,915 individuals classified as European ancestry who also had at least one smoking phenotype. Second, after visually inspecting PCs 1-4 of the selected individuals, we suspected residual population heterogeneity in PC 4 (Figure S5). We then further restricted samples to those whose summed Euclidean distance of PCs 1-4 fell within the 1 interquartile range (IQR) of the European sample (N=38,915) identified in the first step. We additionally created samples using 1.5, 2, 3 IQR and reserved them for sensitivity analysis, described further in that section (Figure S1-2). After IQR filtering, we only retained unrelated individuals, resulting in following final sample size per phenotype in Table 1 (N ranging from 14,749 to 26,347). Relatedness was estimated with HapMap3 variants (HWE p-value > 10^{-6}, MAF > 0.01) using `GCTAv1.92` to obtain a list of nominally unrelated individuals with pairwise $\widehat{\pi} < .025$.

Phenotypes

TOPMed is a consortium of independent studies, where DNA samples were sequenced and called in a unified way. Smoking phenotypes had previously been collected independently in each of the constituent TOPMed studies. Four smoking phenotypes, each representing self-report survey questions assessing different stages of tobacco use, were available across most TOPMed studies. We used the same definition and coding scheme as Liu et al., 2019[8]. Age of smoking initiation (AgeSmk) was the age at which an individual started regularly smoking. Cigarettes per day (CigDay) was the average number of cigarettes smoked per day as a current or former smoker, and grouped into five bins, with higher numbers indicating greater use. For both AgeSmk and CigDay, lifelong non-smokers are excluded (set to missing). Smoking cessation (SmkCes) and initiation (SmkInit) are binary variables indicating former versus current smokers and never versus ever smoker, with case defined as current and ever smoker, respectively. A linear mixed-effects regression analysis indicated that these four variables are correlated but not redundant, each measuring distinct aspects of smoking behavior (Supplementary Note, Table S2). Descriptive statistics for each phenotype across cohorts are presented in Table S3.

Genotypes, LD scores, GRM, and GREML-LDMS-I

Genome-based restricted maximum likelihood (GREML) estimates heritability by comparing phenotypic similarity to observed genetic similarity among distantly related individuals using a linear mixed model[40]. It can yield biased estimates when causal variants are unevenly distributed as a function of LD and MAF[18]. To mitigate this bias, GREML-LDMS-I partitions genomes into different LD × MAF bins[19]. We initially considered ~710 million phased genotypes that have passed strict quality filters[41]. We additionally removed 95,750 variants with Hardy-Weinberg equilibrium *p*-values less than 10^{-6} in the European sample (N=38,915). Then, we calculated allele frequency separately for each phenotype using `plink1.9` in a final sample that went through PC, IQR, and relatedness filtering. We then stratified variants by MAF, and additionally by median linkage disequilibrium (LD) scores within the two most common MAF bins[19]. This resulted in the following six bins: MAF (0.05, 0.5] high LD, MAF (0.05, 0.5] low LD, MAF (0.01, 0.05] high LD, MAF (0.01, 0.05] low LD, MAF (0.001, 0.01], and finally MAF (0.0001, 0.001]. We stratified only the common variant bins by LD because most low-frequency and rare variants have low LD scores, and to limit the number of bins to retain power for heritability estimation. LD scores of individual variants were calculated using `GCTA1.92` with default 10Mb window in the final sample combined across four smoking phenotypes (Table S4). This process resulted in approximately 35 million SNPs and indels (Table 1; Supplementary Note).

For each phenotype, we performed GREML-LDMS-I with the GRMs for above-mentioned six bins and cohort indicator matrix as random effects. The cohort matrix was a N x N matrix indicating whether a given pair of individuals belongs to the same study (1, otherwise 0). For AgeSmk and CigDay, we inverse-rank normalized residuals of these phenotypes after regressing out age, age^2, sex and their two-way interaction terms[42–44], and entered 11 PCs and sequencing center as fixed effects (Figure S3). We used PCs released by TOPMed consortium, which were calculated by `pcair` function in `GENESIS` package in R using 638,486 SNPs with $|LD| < 0.32$ and $MAF > 0.01$. SmkCes and SmkInit are binary phenotypes, thus no transformation was applied. Age, age^2, sex and their interaction terms, 11 PCs and sequencing center were entered as fixed effects for binary phenotypes. We used `GCTAv1.92` for both construction of GRM and GREML-LDMS-I analysis. We allowed the estimates to be negative to obtain unbiased estimates of heritability and standard error. Heritabilities for binary phenotypes were analyzed under a liability threshold model[45]. Population prevalence was set at 0.15 and 0.42 for SmkCes and SmkInit, respectively, based on smoking prevalence in the UK Biobank dataset, to allow for ready comparison with this publicly available and widely- used dataset. For all traits, total heritability was calculated by adding heritability estimates of the six LDMS bins with SEs approximated by the delta method[46].

Partitioning rare variants heritability

To further interrogate sources of rare variant heritability, we divided variants in the low-frequency and rare variants bins into protein altering versus non-protein altering variants bins[22]. Functional impact of variants was assessed by snpEff 4.3 annotation with “HIGH” and “MODERATE” categorized as protein altering while “LOW” and “MODIFIER” categorized as non-protein altering[47]. Variance components were then estimated for a total of eight bins including four low-frequency/rare variants and the four common variant bins.

Sensitivity analyses

To evaluate the robustness of our results, we tested the effect of different decisions with respect to filtering on ancestry, relatedness, and phenotype transformation. First, we created three additional samples with different levels of ancestral variation by gradually relaxing the sample inclusion thresholds based on ancestry PCs. Specifically, we created three samples whose PC-based Euclidean distance lay within 1.5 IQR, 2 IQR, and 3 IQR of the European sample (N=38,915). The greater the IQR threshold, the more ancestral variation is present in the resulting sample, the larger the sample, and the greater the chance of observing effects of population stratification. Finally, we evaluated yet another alternative ancestry-based filter for comparison with our main result. We selected European samples whose PCs were within 6 standard deviations (SD) of the mean of PC1 and PC2 of 1000G CEU sample (Utah residents with Northern and Western European ancestry), an approach used previously in a past study of height[22]. This resulted in similar sample size with that from main analysis (Table S1). More details for the sensitivity analysis procedure is in Supplementary Note.

We also evaluated two relatedness thresholds: $\hat{\pi} < .05$ and $< .025$, which correspond approximately to being related less than first and second cousin, respectively. In addition, we evaluated sensitivity to phenotype transformation methods for quantitative phenotypes (AgeSmk and CigDay). For AgeSmk, we compared rank-based inverse normal transformation (IVRT), log transformation, and log transformation after removing outliers defined as observations lying beyond 3SD from the mean. For CigDay, we compared IVRT and log transformation only, given that CigDay is binned to five groups and thus had no outliers.

Permutation to further control effects of population stratification

Past research has indicated that rare variants can show different patterns of confounding than common variants[34]. To deal with this possibility, we applied a recently developed permutation method designed to control for type I error in genetic association tests of rare variants[23,37,38]. Specifically, we created an N × N distance matrix populated by scaled Euclidian distance of PC1-11 between each individual. Then, we randomly exchanged genotypes of a given individual with one of their 100 nearest neighbors[23,38,48]. We created a total of 200 replicates of permuted genotypes and applied GREML with the same set of fixed and random covariates used in the main analysis. Mean ($\hat{h}^2{}_{\text{null}}$) and SD of heritability estimates from 200 permutation replicates were calculated for each bin and were tested against zero using a one-sided Z-test. Mean heritability greater than zero across replicates would indicate significant bias induced by population stratification. We also tested the estimates from main analysis against the permuted null distribution using two-sided Z-test (Supplementary Note).

Pedigree-based heritability

Missing heritability is often quantified as the difference between GREML results and biometric variance decompositions based on families (e.g., twins). Indeed, GREML as described thus far was applied only to distantly related individuals, so as to avoid confounding with non-additive and shared environmental effects. Here, we take advantage of thousands of closely related pairs of participants in TOPMed to estimate the heritability of our four smoking phenotypes in close pedigrees ($\hat{h}^2{}_{\text{ped}}$). $\hat{h}^2{}_{\text{ped}}$ is same as $K_{IBS>t}$ in Zaitlen et al 2013[12], which is estimated from a single GRM including non-zero values for closely related individuals and zero values for distantly related individuals. This provides an upper bound on the narrow-sense heritability, to which we can compare our GREML estimates to quantify any of the remaining missing heritability. This quantity is analogous to twin heritability, but unlike traditional pedigree or twin estimates, $\hat{h}^2{}_{\text{ped}}$ uses measured genotypes to estimate relatedness rather than expected relatedness of relatives based solely on their pedigree.

For this analysis, we created a GRM with all available samples after excluding pairs related greater than .80 to exclude identical twins and duplicates. To aid in model identification, we included cohorts that had at least 10 first-degree relatives ($\hat{\pi} > .375$). A list of cohorts and details of the procedure are presented in Table S5 and Supplementary Note. This GRM was fitted together with a cohort matrix and the same set of fixed effect covariates used in the main analysis. To test whether resulting $\hat{h}^2{}_{\text{ped}}$ is underestimated due to relatively low level of relatedness structure in the sample, we conducted the same analysis again using only Framingham Heart Study (FHS) which consists of family samples (i.e., high level of relatedness structure) without a random effect of cohort.

Results

Heritability estimates

SNP-based heritability ($\hat{h}^2{}_{\text{WGS}}$) for the four smoking phenotypes was initially estimated as following: 0.31 (.075), 0.146 (.062), 0.252 (.087), 0.242 (.069) for AgeSmk, CigDay, SmkCes, and

SmkInit, respectively. These are unadjusted values calculated by summing up estimates of all six LD x MAF bins; values adjusted for residual stratification using our permutation procedure are described in the next section. Heritability estimates from the common variant bins (i.e., MAF (.01, 0.5] including both high and low LD bins) were summed to compute heritability attributable to common variants ($\hat{h}^2{}_{\text{common}}$). Likewise, heritability estimates from the rare variant bins (i.e., MAF (0.0001, 0.01]) were summed to obtain heritability attributable to rare variants ($\hat{h}^2{}_{\text{rare}}$). Common variants accounted for more than half of the heritability of CigDay (.087; SE .038) and SmkInit (.169; SE .038), with low LD common variants contributing to the majority of $\hat{h}^2{}_{\text{common}}$ for both phenotypes. Rare variants accounted for more than half of the heritability of AgeSmk (.217; SE .067) and SmkCes (.177; SE .079), with the majority of the variance attributable to the rarest bin (MAF (0.0001, 0.001]). Estimates for the six MAF × LD bins are presented in Figure 1 and Table S6.

We further partitioned heritability of two rare variant bins into protein-altering and non-altering bins. For MAF (.001-.01], estimates were close to zero (-.047 ~ .039) regardless of the protein altering property for all phenotypes other than CigDay which showed 9.1% (SE 4.7%) of phenotypic variance accounted for by non-protein altering variant. For MAF (.0001-.001]) bin, non-protein altering variants accounted for 25.1% (SE 6.4%) and 13.6% (SE 7%) of the phenotypic variance of AgeSmk and SmkCes (see Table S7 for full results). Overall, there was very limited evidence for a prominent role of rare protein coding variants in the genetic etiology of these smoking phenotypes.

Permuted mean heritabilities were mostly close to zero across different bins and phenotypes (Figure 2, Table S8). Only the rarest bin of AgeSmk had permuted mean heritability significantly different from zero with weaker evidence for SmkCes (AgeSmk: Mean=.102, SE=.050, SmkCes: Mean=.043, SE=.041). We also tested heritability estimates from main analysis against the permuted null distribution (Figure 1, Table S6). The rarest bins of AgeSmk and SmkCes were unlikely to be drawn from the null distribution (*p*=.002 and *p*=.012, respectively), indicating that the initial estimates are not entirely accounted for by residual population structure. We adjusted partial impact of residual stratification by subtracting the permuted mean from $\hat{h}^2{}_{\mathrm{WGS}}$. Adjusted $\hat{h}^2{}_{\mathrm{WGS}}$ for AgeSmk and SmkCes was 0.212 (SE 0.075) and 0.209 (SE 0.087). Note that the adjusted $\hat{h}^2{}_{\mathrm{WGS}}$ is conservative, as permuted heritability may partly capture true rare variant effects among individuals sharing recent ancestors.

Estimates from pedigree analysis ($\hat{h}^2{}_{\mathrm{ped}}$) are presented in Figure 3 and Table S9. All $\hat{h}^2{}_{\mathrm{ped}}$ were greater than our (both adjusted and unadjusted) $\hat{h}^2{}_{\mathrm{WGS}}$, except for SmkCes. When we compare $\hat{h}^2{}_{\mathrm{WGS}}$ and $\hat{h}^2{}_{\mathrm{ped}}$, it would appear that the inclusion of rare variants from WGS accounts for much of the missing heritability for smoking phenotypes (60%-100%). Using the Framingham Heart Study (FHS) cohort, which has a large proportion of related individuals, we obtained generally comparable $\hat{h}^2{}_{\mathrm{ped}}$, albeit with lower estimate for AgeSmk and overall greater SEs.

Sensitivity analyses

We explored how sensitive the results were to ancestral filtering thresholds, relatedness cut-off thresholds, and phenotype transformation methods (Figure 4; Figures S1-4; Table S10). The point estimates tended to slightly increase as we applied stricter ancestry-based filtering for AgeSmk and SmkInit. Notably, $\hat{h}^2{}_{\mathrm{WGS}}$ was estimated higher when we included extreme AgeSmk observations compared to when excluding outliers or applying rank-based inverse normalization, both of which effectively controlled the influence of outliers. Across all phenotypes, $\hat{h}^2{}_{\mathrm{WGS}}$ tended to be higher when using a relatedness cut-off of .05 compared to using .025 with maximum difference in $\hat{h}^2{}_{\mathrm{WGS}}$ being .055 for SmkInit in 2IQR sample. Finally, heritability estimates using alternative sample selection strategy based on the 1000 Genomes CEU group were generally comparable to the main $\hat{h}^2{}_{\mathrm{WGS}}$ estimates across six bins (Table S10).

Discussion

Using the largest sample of whole genome sequences (N=14,747 - 26,340) to date for complex behavioral traits, we found that rare variants with MAF 0.01% to 0.1% accounted for approximately 15% and 10% of phenotypic variation of AgeSmk and SmkCes after correcting for potential influence of population structure at rare variants. These estimates can be seen conservative as our permutated mean might have partially included true rare variant effects. Compared to AgeSmk and SmkCes, contribution of rare variants to CigDay and SmkInit was lower (6% and 7%, respectively), with common variants explaining a greater proportion of the total phenotypic variation (9% and 17%, respectively). After adjustment, the total heritability estimate ($\hat{h}^2{}_{\mathrm{WGS}}$) was 0.21, 0.15, 0.21, and 0.24 for AgeSmk, CigDay, SmkCes, and SmkInit, respectively which is 1.8 to 4.5 times higher than past heritability estimates based on common variants alone[8].

A handful of studies have now reported evidence for rare variants contributing to phenotypic variance of anthropometric[22], medical[17], and transcriptomic phenotypes[16]. For example, rare variation may explain all of the missing heritability for height and BMI[22]. Our $\hat{h}^2{}_{\mathrm{WGS}}$ generally falls below past twin estimates for tobacco use phenotypes ($\hat{h}^2{}_{\mathrm{twin}}$ = .48; SE .05)[7]. However, after possibly overcorrecting for population stratification, current $\hat{h}^2{}_{\mathrm{WGS}}$ estimates accounted for 61% to 100% of our pedigree-based heritability estimate ($\hat{h}^2{}_{\mathrm{ped}}$) across the four phenotypes, closing the gap on the missing heritability for these phenotypes. In the present study, we consider our pedigree estimate of heritability as the most relevant benchmark by which to judge the GREMLresults. While twin studies – especially same-sex twins – control perfectly for standard covariates (e.g., age, sex, birth year), family studies, especially of multiple generations, typically incorporate only linear combinations of such predictors. This results in a statistically adjusted phenotype that contains more noise, and proportionally less heritable variance than a twin study. Future studies will benefit from considering how different analytic approaches, such as using twin-only samples versus diverse classes of relatives (e.g., grandparents-grandchild, siblings, half-sibling etc) and using SNP-based versus expected relatedness, influence heritability estimation and potentially capture different aspects of heritable variation[12].

Published studies suggested that population structure induced by rare variants may not be sufficiently accounted for by PC correction[32,34,49,50]. Similarly, we found that mean permuted heritability substantially departed from zero (M=.10, SE=.05) for the MAF [0.01-0.1%] bin of AgeSmk and to lesser degree, for that of SmkCes (M=.04, SE=.04), indicating the possibility of partial confounding due to residual population structure. We did not find evidence for this confounding for other phenotypes or MAFs. Existence of geographically localized non-genetic risk or systematic measurement bias in different cohorts could lead to rare variant stratification[34]. Without substantial knowledge of causal risk factors that align with rare variant sharing and the availability of such data, it is difficult to directly identify the source of the confounding. Given increasing interest in the role of rare variation in complex disease, research into the nature and method to detect and adjust for rare variants will be crucial. For example, future studies should consider extending the current permutation approach and improve computational efficiency.

The majority of the rare variant heritability (90%-100%) was attributable to non-protein altering regions for AgeSmk and SmkCes. This suggests that most genetic variation is likely to be located outside protein-coding regions, which themselves comprise only ~1% of base pairs in the human genome. For common variants, a majority of the heritability appeared attributable to low-LD variants for all smoking traits. This seems consistent with the action of negative selection, indicating that these variants are relatively young in the genealogical history, and are still being pushed to lower frequencies. Consistent with this idea, Gazal et al 2017 reported that common variants with low LD in low-recombination rate regions had larger per-SNP heritability than those with high LD as they are more likely to be recent and thus have less time to be removed by negative selection[51].

In sensitivity analyses, most heritability estimates from varying combinations of ancestral variation, relatedness cut-offs, and phenotype transformation typically showed differences of 1%-5% in each sensitivity condition. AgeSmk tended to show higher rare variant heritability when extreme observations were included than they were not. Extreme observations may induce model misspecification or alternatively, are enriched in causal rare variants, possibly leading to higher heritability estimates[23,52,53]. The influence phenotypic scaling was reported in recent study of Evans et al. 2021 where CigDay showed higher heritability when it was dichotomized versus the ordinal version analyzed here[52]. Estimation of heritability for binned variable (e.g., CigDay in this study) may be improved by different modeling choice including liability threshold model[54]. Finally, relaxing relatedness thresholds tended to be associated with higher estimates, an effect one would expect if shared environment in distantly related individuals were influencing smoking traits (e.g., state-level policy or regional culture)[55].

Our findings should be interpreted in light of several limitations. First, while our sample size is the largest to date for heritability analysis using WGS, even larger datasets or more precise phenotypic measures are required. Larger studies would provide greater precision in estimation and a more comprehensive assessment of genetic architecture of these complex traits by finer partitioning by MAF and functional annotations. Second, even with the use of deep sequences, we did not fully "recover" trait heritability, either as estimated using available pedigrees, or twin heritability reported in the literature. There remain many explanations, including ultra-rare variants (MAF <= .0001) and other types of genetic variations (e.g., copy number variations) that may constitute additional sources of heritability. Next, our pedigree-based heritability estimates may be inflated by shared environment, as we were unable to model genetic similarity and environmental similarity separately. Therefore, the pedigree estimates should be interpreted as likely upward biased upper-bounds for SNP-based narrow-sense heritability. Finally, smoking phenotypes were measured by one or two questions and were limited to those commonly collected in biomedical studies like those in TOPMed. This allows accumulations of large sample sizes across multiple independently- collected samples, but also may reflect individual study characteristics, whose variance ultimately can be captured by our cohort random effect.

In conclusion, our results indicate that rare variants in regulatory region contribute substantially to the heritability of smoking phenotypes. Common variant influences seem to be overrepresented by more recently arisen alleles that on average have lower LD[51]. Nicotine has been pervasive in the environment for millions of years acting as a pesticide, such that one expects many organisms, including humans, to have evolved systems to handle nicotine exposures. The genetic variants that influence addiction to nicotine appear to be highly polygenic and under negative selection, despite the fact that tobacco in its cigarette form represents an evolutionarily novel environment for humans. Smoking phenotypes can also be considered a manifestation of psychologically disinhibited and externalizing tendencies (e.g., age at onset of reproductive behaviors), such that selection pressure on them can also influence genetic architecture of smoking by pleiotropic effects[56]. The current study informs the genetic etiology of nicotine addiction and provides a benchmark for the future study of other complex behavioral traits.

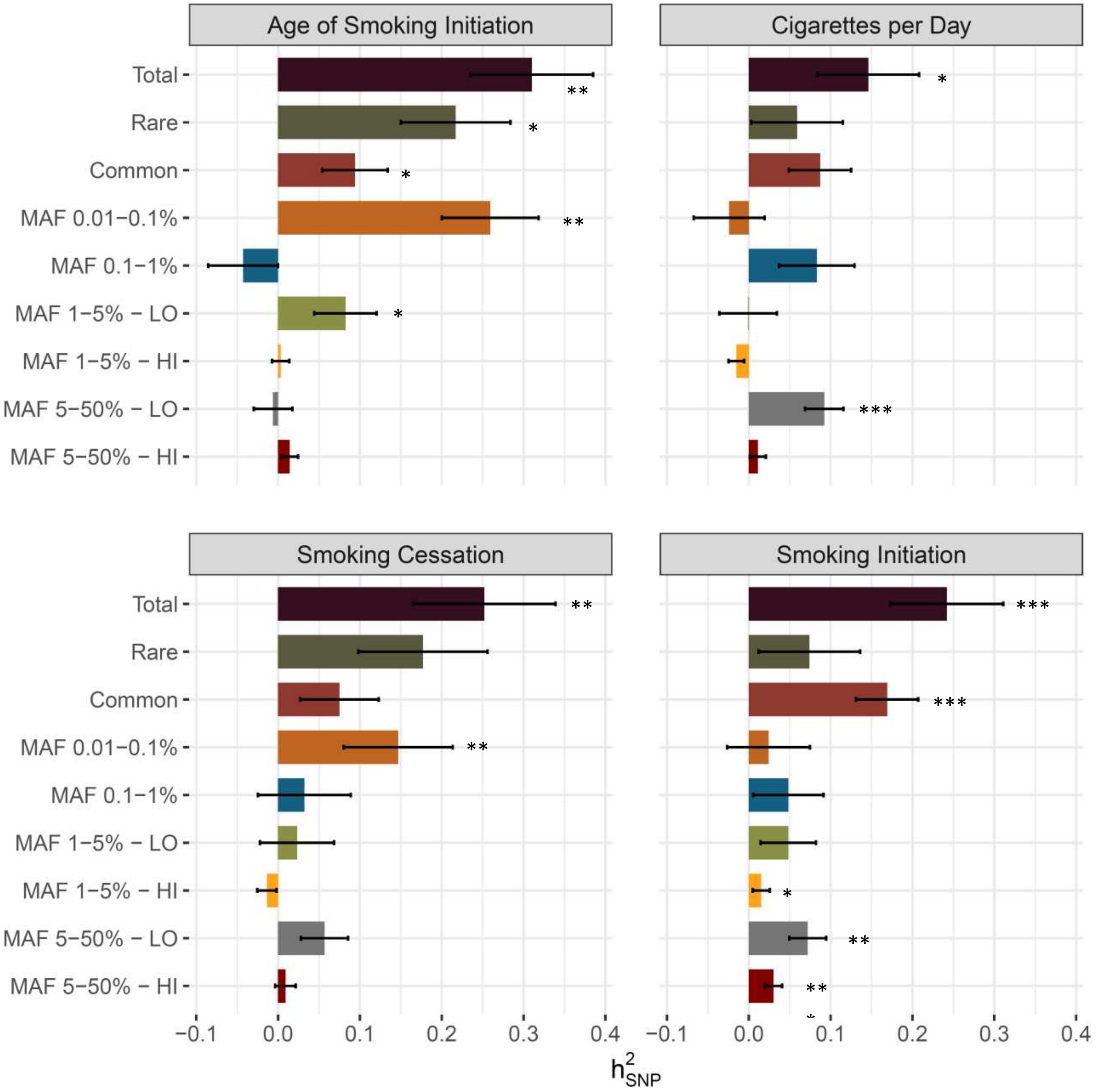
Data availability

Phenotype data are available through authorized access portal in dbgap (https://dbgap.ncbi.nlm.nih.gov/) or direct request to Principal Investigators (PIs). Accession numbers and email addresses of PIs are presented in Supplementary Note. Genetic data are available through dbgap.

Code availability

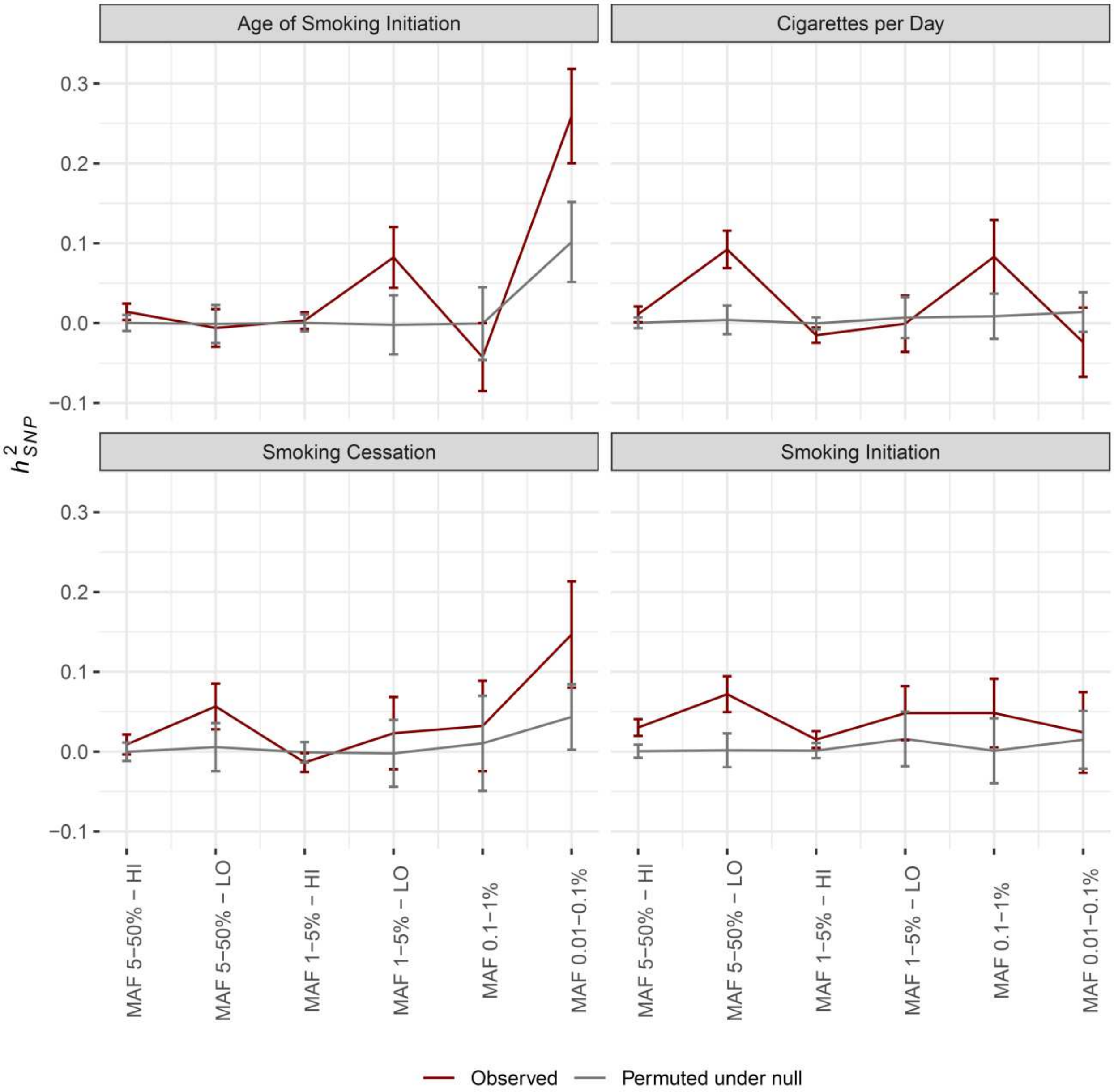
GCTA software is available at https://cnsgenomics.com/software/gcta/. We obtained a code used to generate permutation sequences through email correspondence with Aurélie Cobat (aurelie.cobat@inserm.fr).

Figure 1. SNP-based heritability estimates for AgeSmk, CigDay, SmkCes, and SmkInit, for each of the six MAF/LD bins, as well as sums across bins.



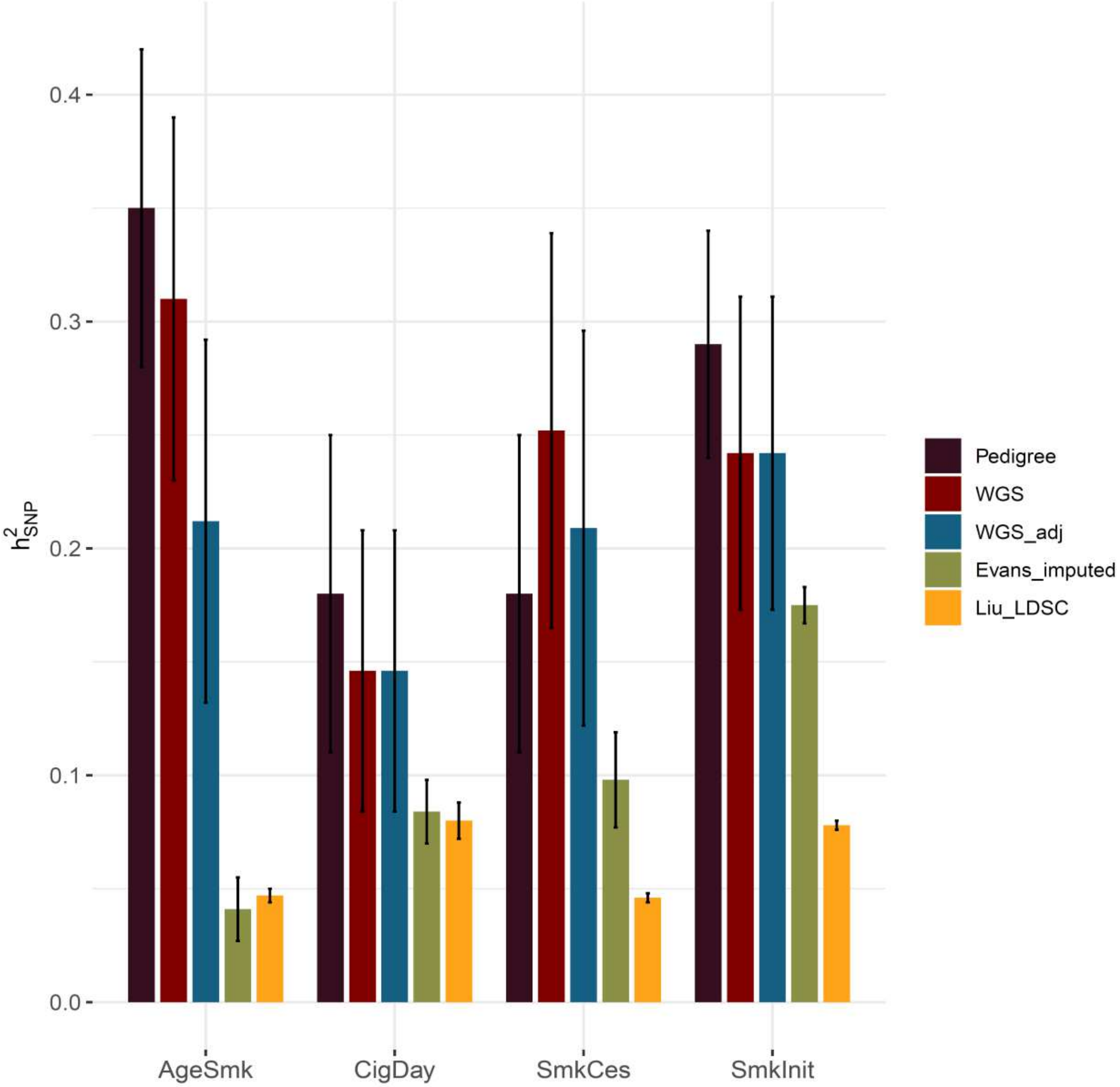
Note. Bars are standard errors. The “Rare” bin is the sum of the MAF .1=1% and MAF .01-.1%. “Common” is the sum of the other MAF bins. Total is the sum of Rare and Common. No estimates shown here were adjusted by results from permutation procedure. For adjusted results, please see Figure 2. Asterisks were added to the components that are significantly different from permuted mean under null distribution (see Table S6). **p* < .05, ***p* < .01, ****p* < .001.

Figure 2. Mean h^2_{SNP} estimates from permutation trials



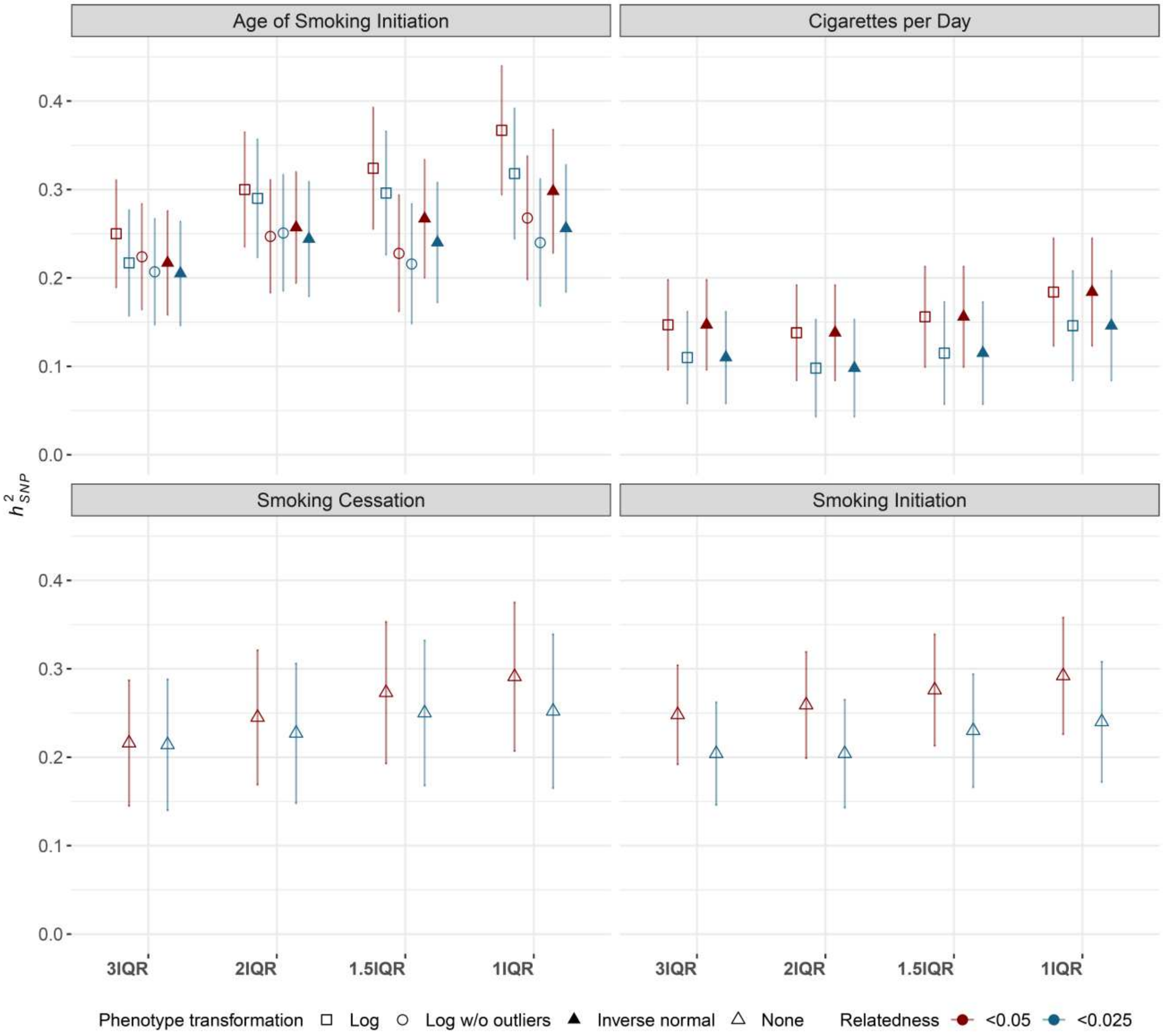
Note. Red lines are mean h^2_{SNP} estimates and standard errors for each variants bin from GREML-LDMS analysis. Gray lines are mean h^2_{SNP} estimates and standard errors from permutation trails.

Figure 3. Comparison of heritability estimates between current and published studies



Note. Heritability estimates and their standard errors. It shows SNP heritability estimates across different studies. Pedigree, WGS, and WGS_adj refer to pedigree-based, WGS-based, and permutation adjusted SNP heritability estimates in current study. Evans_imputed and Liu_LDSC each refer to SNP heritability estimates from Evans et al. (MAF:0.5-0.01, relatedness threshold=.02[52]) and LDSC analysis from a recent meta-analysis of tobacco use[8].

Figure 4. Sensitivity analysis of h^2_{WGS} estimates



Note. X-axis indicates different ancestry filtering thresholds. The shape of the points indicates phenotype transformation methods with “Log”, “Log w/o outliers”, “Inverse normal”, “None” indicating log transformation including outliers, log transformation without outliers, rank-based inverse normal transformation, and no transformation, respectively. Red and blue color each indicate relatedness thresholds .025 and .05. Dots and whiskers each represent heritability estimates and their SEs. Note that here we presented $\hat{h}^2_{\text{WGS}}$ for AgeSmk (1IQR and $\hat{\pi} < .025$) estimated excluding variants with MAC 3 as these variants fall below the MAF threshold for all other comparison $\hat{h}^2_{\text{WGS}}$ for AgeSmk.

Table 1. Sample size and number of variants per MAF/LD bin.

	Sample size (unrelated individuals only)					
	Method to Select Samples Based on Ancestry					
	1IQR[a]	1.5IQR	2IQR	3IQR	CEU 6SD	
AgeSmk	14,749	15,133	15,432	15,706	15,052	
CigDay	15,434	15,832	16,138	16,434	15,748	
SmkCes	17,872	18,319	18,662	18,988	18,223	
SmkInit	26,347	26,958	27,402	27,884	26,812	
	Number of variants per bin (MAF and Linkage Disequilibrium)[b]					
	5-50% - HI	5-50% - LO	1-5% - HI	1-5% - LO	0.1-1%	0.01-0.1%
AgeSmk	3,092,517	3,092,534	1,342,734	1,342,736	5,392,813	28,280,118
CigDay	3,092,240	3,092,269	1,341,881	1,341,882	5,435,505	20,415,037
SmkCes	3,092,593	3,092,594	1,340,764	1,340,771	5,413,019	23,483,166
SmkInit	3,092,454	3,092,475	1,341,068	1,341,071	5,395,579	21,108,704

[a] 1IQR is used for main analysis and the rest of samples was used for sensitivity analysis. 1IQR = 1 * the interquartile range of PCs 1-4, and is the most restrictive choice, 1.5IQR is 1.5 * the interquartile range of PCs 1-4, and so on. CEU 6SD is an alternative way to select samples based on ancestry of the CEU group of 1000 Genomes.

[b] This shows the number of variants per bin in 1IQR unrelated samples ($\hat{\pi} < .025$).

References

1. Johnson, T. & Barton, N. Theoretical models of selection and mutation on quantitative traits. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1411–1425 (2005).
2. Keinan, A. & Clark, A. G. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336**, 740–743 (2012).
3. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**, 5–22 (2017).
4. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
5. Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S. & Murray, C. J. Selected major risk factors and global and regional burden of disease. *The Lancet* **360**, 1347–1360 (2002).
6. Quach, B. C. *et al.* Expanding the Genetic Architecture of Nicotine Dependence and its Shared Genetics with Multiple Traits: Findings from the Nicotine Dependence GenOmics (iNDiGO) Consortium. *bioRxiv* 2020.01.15.898858 (2020) doi:10.1101/2020.01.15.898858.
7. Polderman, T. J. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics* **47**, 702 (2015).
8. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics* **51**, 237–244 (2019).
9. Erzurumluoglu, A. M. *et al.* Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Molecular Psychiatry* **25**, 2392–2409 (2020).
10. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47**, 1236 (2015).
11. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11**, 446–450 (2010).

12. Zaitlen, N. *et al.* Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLOS Genetics* **9**, e1003520 (2013).

13. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS* **109**, 1193–1198 (2012).

14. Young, A. I. Solving the missing heritability problem. *PLOS Genetics* **15**, e1008222 (2019).

15. Gibson, G. Rare and Common Variants: Twenty arguments. *Nature Review Genetics* **13**, 135–145 (2012).

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

17. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics* **48**, 1443–1448 (2016).

18. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics* **47**, 1114–1120 (2015).

19. Evans, L. M. *et al.* Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics* **50**, 737 (2018).

20. Derkach, A., Zhang, H. & Chatterjee, N. Power Analysis for Genetic Association Test (PAGEANT) provides insights to challenges for rare variant association studies. *Bioinformatics* **34**, 1506–1513 (2018).

21. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

22. Recovery of trait heritability from whole genome sequence data | bioRxiv. https://www.biorxiv.org/content/10.1101/588020v1.

23. Hernandez, R. D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression. *Nature Genetics* **51**, 1349–1355 (2019).

24. Sul, J. H. *et al.* Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates. *Translational Psychiatry* **10**, 1–10 (2020).

25. Halvorsen, M. *et al.* Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nature Communications* **11**, 1–13 (2020).

26. Luo, Y. *et al.* Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nature Genetics* **49**, 186–192 (2017).

27. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).

28. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006).

29. Haworth, S. *et al.* Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nature Communications* **10**, 333 (2019).

30. Browning, S. R. & Browning, B. L. Population Structure Can Inflate SNP-Based Heritability Estimates. *Am J Hum Genet* **89**, 191–193 (2011).

31. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702 (2019).

32. Liu, Q., Nicolae, D. L. & Chen, L. S. Marbled Inflation From Population Structure in Gene-Based Association Studies With Rare Variants. *Genet Epidemiol* **37**, (2013).

33. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* **108**, 11983–11988 (2011).

34. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **44**, 243–246 (2012).

35. O'Connor, T. D. *et al.* Fine-Scale Patterns of Population Stratification Confound Rare Variant Association Tests. *PLOS ONE* **8**, e65834 (2013).

36. Persyn, E., Redon, R., Bellanger, L. & Dina, C. The impact of a fine-scale population stratification on rare variant association test results. *PLOS ONE* **13**, e0207677 (2018).

37. Mullaert, J. *et al. Taking population stratification into account by local permutations in rare-variant association studies on small samples*. http://biorxiv.org/lookup/doi/10.1101/2020.01.29.924977 (2020) doi:10.1101/2020.01.29.924977.

38. Bouaziz, M. *et al.* Controlling for Human Population Stratification in Rare Variant Association Studies. *bioRxiv* 2020.02.28.969477 (2020) doi:10.1101/2020.02.28.969477.

39. Zhang, D., Dey, R. & Lee, S. Fast and robust ancestry prediction using principal component analysis. *bioRxiv* 713172 (2019) doi:10.1101/713172.

40. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42**, 565 (2010).

41. Taliun, D. *et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program*. http://biorxiv.org/lookup/doi/10.1101/563866 (2019) doi:10.1101/563866.

42. Fidler, J., Ferguson, S. G., Brown, J., Stapleton, J. & West, R. How does rate of smoking cessation vary by age, gender and social grade? Findings from a population survey in England. *Addiction* **108**, 1680–1685 (2013).

43. Karp, I., O'loughlin, J., Paradis, G., Hanley, J. & Difranza, J. Smoking Trajectories of Adolescent Novice Smokers in a Longitudinal Study of Tobacco Use. *Annals of Epidemiology* **15**, 445–452 (2005).

44. Mathew, A. R. *et al.* Life-Course Smoking Trajectories and Risk for Emphysema in Middle Age: The CARDIA Lung Study. *Am J Respir Crit Care Med* **199**, 237–240 (2018).

45. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics* **88**, 294–305 (2011).

46. Powell, L. A. Approximating Variance of Demographic Parameters Using the Delta Method: A Reference for Avian BiologistsAproximación De La Varianza Para Parámetros Demográficos Utilizando El Método Delta: Una Referencia Para Biólogos De AvesShort Communications. *Condor* **109**, 949–954 (2007).

47. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

48. Taking population stratification into account by local permutations in rare-variant association studies on small samples | bioRxiv. https://www.biorxiv.org/content/10.1101/2020.01.29.924977v1.

49. Zawistowski, M. *et al.* Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. *Eur J Hum Genet* **22**, 1137–1144 (2014).

50. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat Genet* **44**, 623–630 (2012).

51. Gazal, S. *et al.* Linkage disequilibrium dependent architecture of human complex traits shows action of negative selection. *Nat Genet* **49**, 1421–1427 (2017).

52. Evans, L. M. *et al. Genetic architecture of four smoking behaviors using partitioned* h^2_{SNP} . http://medrxiv.org/lookup/doi/10.1101/2020.06.17.20134080 (2020) doi:10.1101/2020.06.17.20134080.

53. Peloso, G. M. *et al.* Phenotypic extremes in rare variant study designs. *European Journal of Human Genetics* **24**, 924–930 (2016).

54. Neale, B. Liability Threshold Models. in *Wiley StatsRef: Statistics Reference Online* (American Cancer Society, 2014). doi:10.1002/9781118445112.stat06439.

55. Conley, D. *et al.* Testing the Key Assumption of Heritability Estimates Based on Genome-wide Genetic Relatedness. *J Hum Genet* **59**, 342–345 (2014).

56. Identification of 370 loci for age at onset of sexual and reproductive behaviour, highlighting common aetiology with reproductive biology, externalizing behaviour and longevity | bioRxiv. https://www.biorxiv.org/content/10.1101/2020.05.06.081273v1.full.

Figures

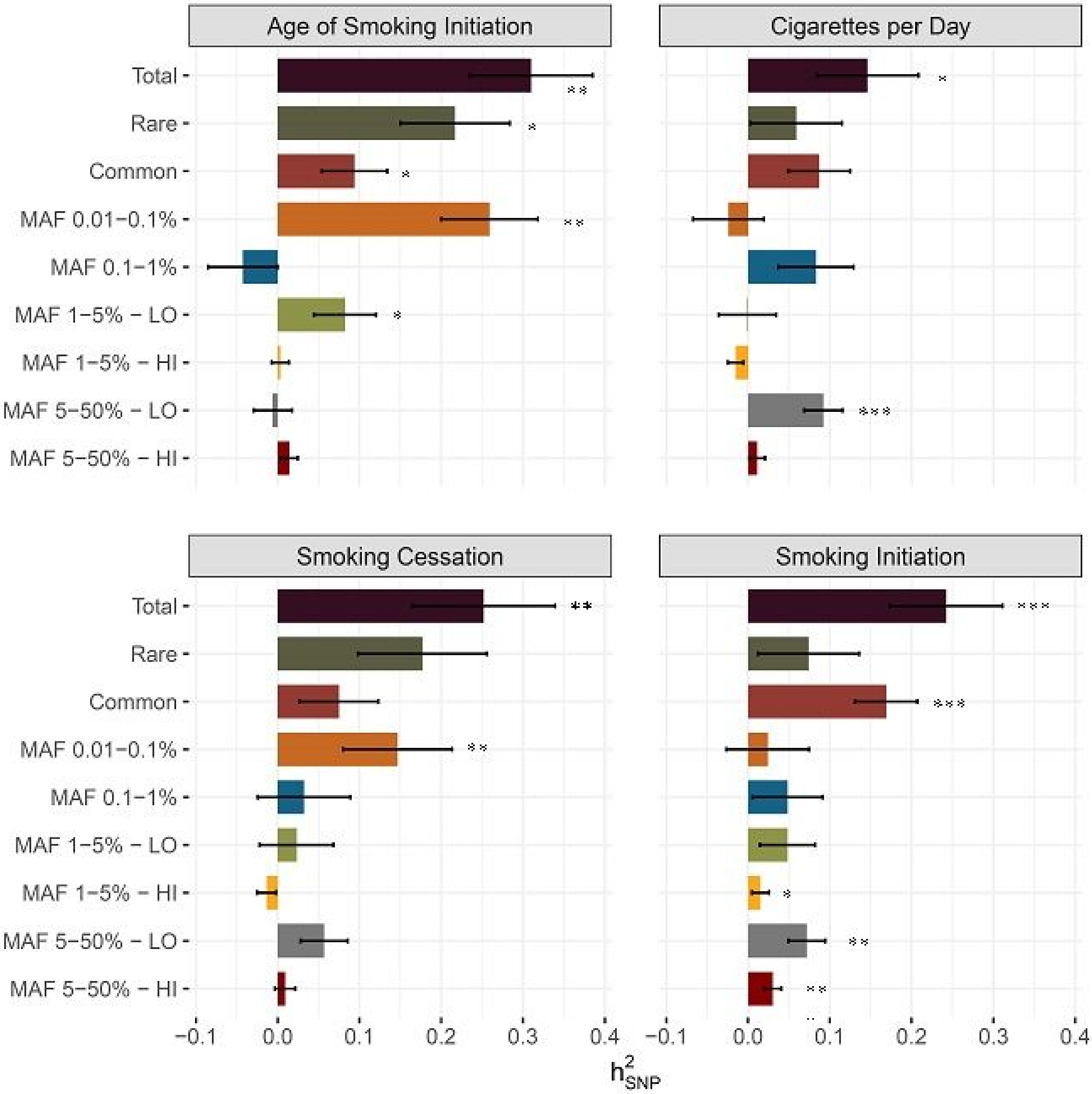


Figure 1

SNP-based heritability estimates for AgeSmk, CigDay, SmkCes, and SmkInit, for each of the six MAF/LD bins, as well as sums across bins. Note. Bars are standard errors. The “Rare” bin is the sum of the MAF .1=1% and MAF .01-.1%. “Common” is the sum of the other MAF bins. Total is the sum of Rare and Common. No estimates shown here were adjusted by results from permutation procedure. For adjusted results, please see Figure 2. Asterisks were added to the components that are significantly different from permuted mean under null distribution (see Table S6). *p < .05, **p < .01, ***p < .001.

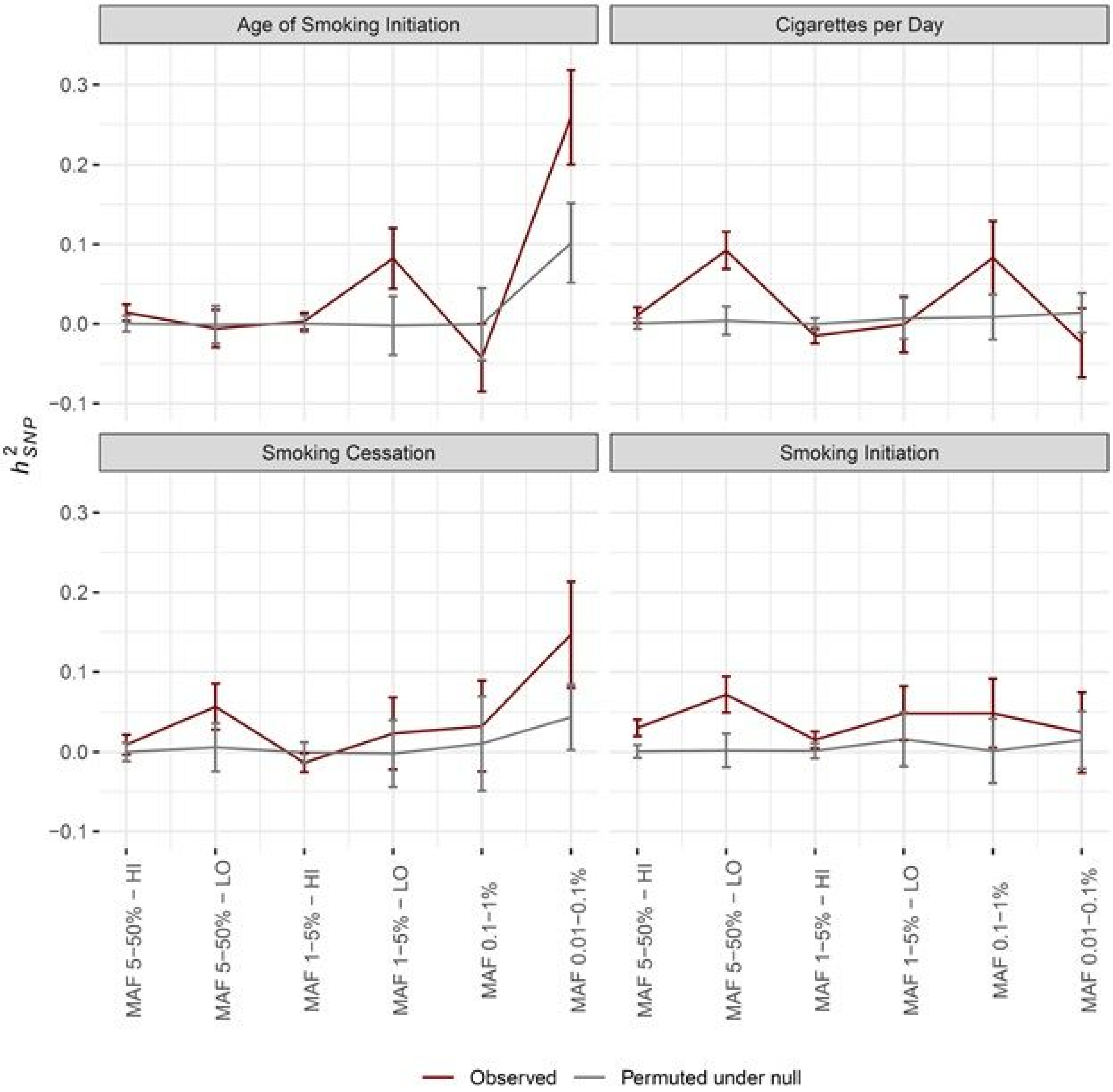


Figure 2

Mean h2SNP estimates from permutation trials. Note. Red lines are mean h2SNP estimates and standard errors for each variants bin from GREML-LDMS analysis. Gray lines are mean h2SNP estimates and standard errors from permutation trails.

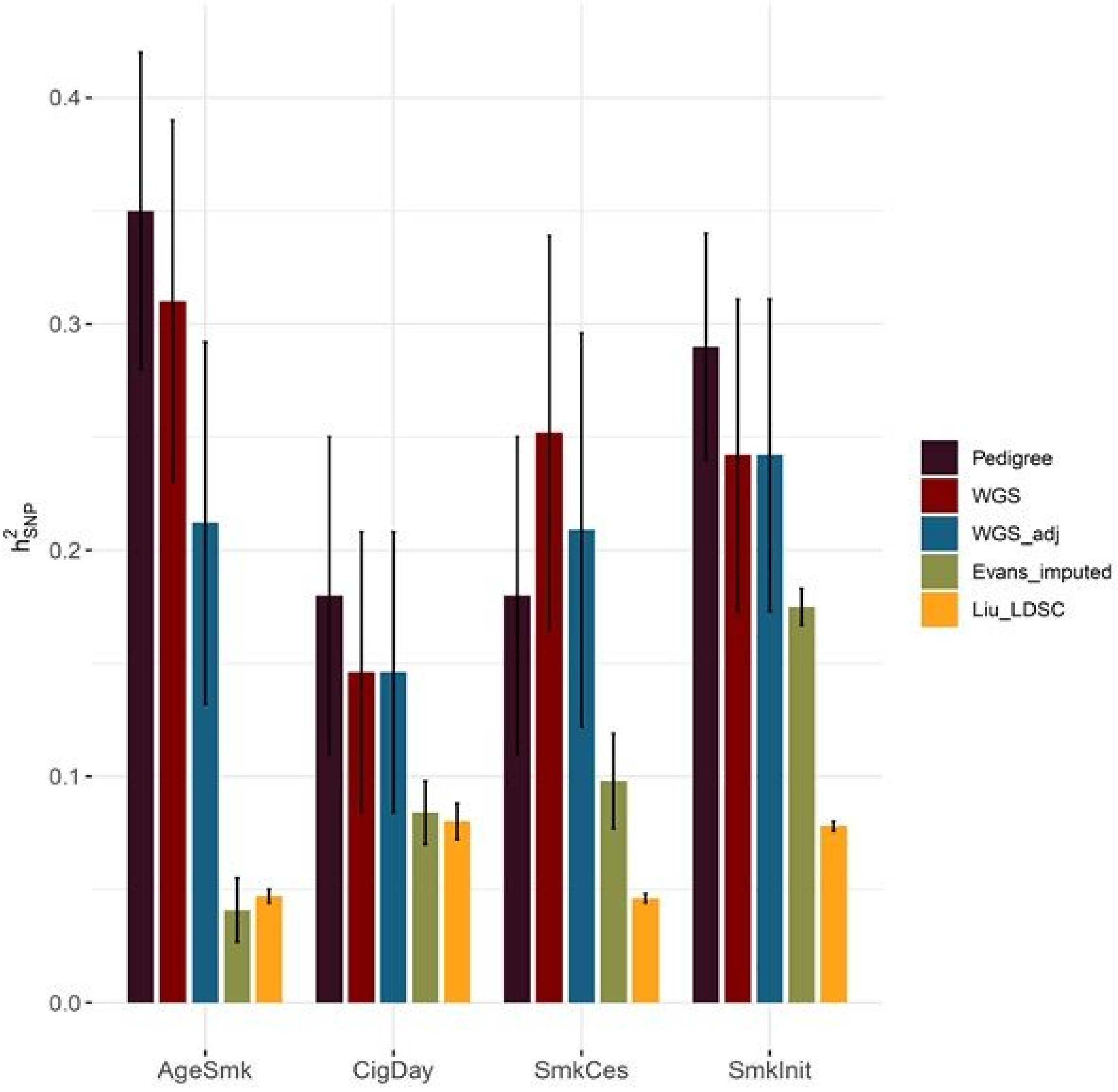


Figure 3

Comparison of heritability estimates between current and published studies. Note. Heritability estimates and their standard errors. It shows SNP heritability estimates across different studies. Pedigree, WGS, and WGS_adj refer to pedigree-based, WGS-based, and permutation adjusted SNP heritability estimates in current study. Evans_imputed and Liu_LDSC each refer to SNP heritability estimates from Evans et al. (MAF:0.5-0.01, relatedness threshold=.0252) and LDSC analysis from a recent meta-analysis of tobacco use8.

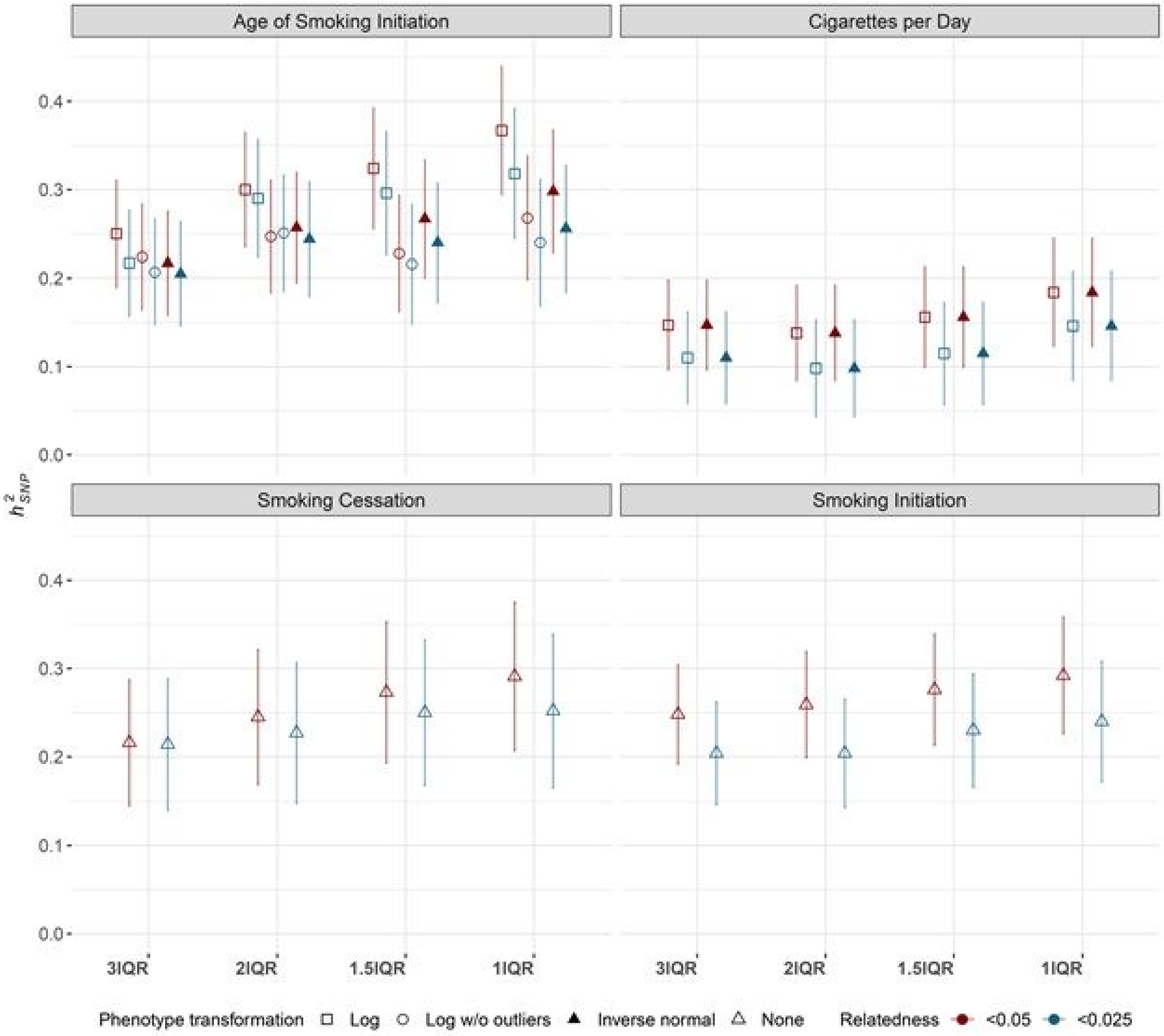


Figure 4

Sensitivity analysis of h2WGS estimates. Note. X-axis indicates different ancestry filtering thresholds. The shape of the points indicates phenotype transformation methods with “Log”, “Log w/o outliers”, “Inverse normal”, “None” indicating log transformation including outliers, log transformation without outliers, rank-based inverse normal transformation, and no transformation, respectively. Red and blue color each indicate relatedness thresholds .025 and .05. Dots and whiskers each represent heritability estimates and their SEs. Note that here we presented h ^2_WGS for AgeSmk (1IQR and π < .025) estimated excluding variants with MAC 3 as these variants fall below the MAF threshold for all other comparison h ^2_WGS for AgeSmk.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryNote032321.docx

- SupplementaryTables042521.xlsx
- SupplementaryTables05042116201838354.xlsx
- SupplementaryFiguresuncompressed042821.docx