Rare genetic variants explain missing heritability in smoking

Seon-Kyeong Jang1, Luke Evans2,3, Allison Fialkowski1, Donna K. Arnett4, Allison E. Ashley-Koch5, Kathleen C. Barnes6, Diane M. Becker7, Joshua C. Bis8, John Blangero9, Eugene R. Bleecker10, Meher Preethi Boorgula9, Donald W. Bowden11, Jennifer A. Brody9, Brian E. Cade12, Brenda W. Campbell Jenkins13, April P. Carson14, Sameer Chavan6, L. Adrienne Cupples15, Brian Custer16, Scott M. Damrauer17,18, Sean P. David19,20, Mariza de Andrade21, Carla L. Dinardo22, Tasha E. Fingerlin23,24, Myriam Fornage25, Barry I. Freedman26, Melanie E. Garrett15, Sina A. Gharib6,27, David C. Glahn28, Jeffrey Haessler29, Susan R. Heckbert30,31, John E. Hokanson32, Lifang Hou33, Shih-Jen Hwang34, Matthew C. Hyman35, Renae Judy17, Anne E. Justice36, Robert C. Kaplan29,37, Sharon L. R. Kardia38, Shannon Kelly39, Wonji Kim40, Charles Kooperberg29, Daniel Levy34,41, Donald M. Lloyd-Jones33, Ruth J. F. Loos42,43, Ani W. Manichaikul44, Mark T. Gladwin45, Lisa Warsinger Martin46, Mehdi Nouraie45, Olle Melander47,48, Deborah A. Meyers10, Courtney G. Montgomery49, Kari E. North50, Elizabeth C. Oelsner51, Nicholette D. Palmer51, Marinnelle Payton52, Anna L. Pelito53, Patricia A. Peyser38, Michael Preuss42,43, Bruce M. Psaty54, Dandi Qiao40, Daniel J. Rader35,55, Nicholas Rafaels6, Susan Redline12, Robert M. Reed56, Alexander P. Reiner29, Stephen S. Rich44, Jerome I. Rotter57, David A. Schwartz58,59, Aladdin H. Shadyab60, Edwin K. Silverman40, Nicholas L. Smith30,31, J. Gustav Smith61,62, Albert V. Smith63, Jennifer A. Smith38, Weihong Tang64, Kent D. Taylor57, Marilyn J. Telen5, Ramachandran S. Vasan65,66, Victor R. Gordeuk67, Zhe Wang42,43, Kerri L. Wiggins8, Lisa R. Yanek7, Ivana V. Yang53, Kendra A. Young32, Kristin L. Young50, Yingze Zhang45, Dajiang J. Liu68, Matthew C. Keller2 and Scott Vrieze1

Common genetic variants explain less variation in complex phenotypes than inferred from family-based studies, and there is a debate on the source of this ‘missing heritability’. We investigated the contribution of rare genetic variants to tobacco use with whole-genome sequences from up to 26,257 unrelated individuals of European ancestries and 11,743 individuals of African ancestries. Across four smoking traits, single-nucleotide-polymorphism-based heritability \( h^2_{SNP} \) was estimated from 0.13 to 0.28 (s.e., 0.10–0.13) in European ancestries, with 35–74% of it attributable to rare variants with minor allele frequencies between 0.01% and 1%. These heritability estimates are 1.5–4 times higher than past estimates based on common variants alone and accounted for 60% to 100% of our pedigree-based estimates of narrow-sense heritability \( h^2_{ped} \) (0.18–0.34). In the African ancestry samples, \( h^2_{SNP} \) was estimated from 0.03 to 0.33 (s.e., 0.09–0.14) across the four smoking traits. These results suggest that rare variants are important contributors to the heritability of smoking.

characterizing the genetic architecture of complex phenotypes is fundamental to understanding genetic influences on human individual differences and has implications in diverse fields, including medicine and psychology. The joint distribution of genetic effect sizes and minor allele frequencies (MAFs) is shaped by natural selection and population history1,2 and informs our understanding of human evolution and gene-mapping strategies for complex traits. The past decade of genome-wide association studies (GWAS) of common genetic variants has expanded our knowledge of this genetic architecture by discovering tens of thousands of genomic loci associated with a wide range of complex phenotypes such as height1,2, lipids3, diabetes4 and schizophrenia5, to name a few. However, the aggregate effect of common genetic variation typically explains less phenotypic variation than that inferred from traditional family-based studies (for example, twins, siblings or larger pedigrees)2, a difference often referred to as ‘missing heritability’. Tobacco use is a complex behavioural trait of high public health concern7–10 and demonstrated genetic and environmental (for example, policy) influences11,12. Not only is tobacco use a leading cause of global morbidity and mortality, but measures of tobacco
use are strong indicators of addiction to nicotine. For example, the number of cigarettes smoked per day is genetically highly correlated \((r = 0.95)\) with dependence on nicotine\(^1\) and other commonly used substances\(^2,3\). Tobacco use is also considered a part of externalizing dimensions of personality and psychopathology along with other substance use and hyperactive behaviours\(^4,5\). The heritability of smoking behaviours has been estimated at approximately 50\% (s.e., 5\%) in twin studies, comparable to many other complex behavioural traits.

At the same time, estimates of tobacco use heritability from GWAS of single nucleotide polymorphisms (SNPs) have routinely found much lower SNP-based heritability \((h^2_{SNP})\) estimates\(^1,6,7\). Such analyses to date have been based on common variants with MAF > 1\% from GWAS of imputed microarrays. In a recent GWAS of tobacco use in up to 1.1 million individuals, Liu et al.\(^8\) reported \(h^2_{SNP}\) estimates ranging between 5\% and 11\%, with smoking initiation and age of smoking initiation showing the highest and the lowest estimates, respectively. Even more recently, Evans et al.\(^9\) reported \(h^2_{SNP}\) estimates of 5–18\% for smoking traits in UK Biobank imputed genotypes of up to 323,068 individuals. In this latter study, the estimated contribution of rare variants to the heritability was minimal, probably due to poor imputation of rare variants with MAF < 1\%. Similar to results for other complex traits\(^10,11\), some but far from all of the twin-based heritability of smoking behaviours can be attributed to common variants obtained through the imputation of microarray genotypes.

There is an extensive literature regarding the possible contributors to missing heritability, including inflated family-based heritability estimates\(^12,13\), increased phenotypic heterogeneity in GWAS compared with family-based studies\(^14\), epistasis\(^14\) and genetic variants not in linkage disequilibrium (LD) with common variants, including structural\(^15\) and rare variants\(^16\). Rare genetic variants are one compelling explanation for the missing heritability of fitness-related traits, as one expects negative selection to force deleterious alleles to low frequencies\(^16,17\). Common variants, in contrast, are expected to explain most trait variance under a neutral model where most mutations have little selective effect\(^17,18\). The contribution of rare variants can inform competing explanations for the missing heritability as well as competing population genetics models, but this requires large samples of extremely well-imputed microarrays or whole-genome sequencing (WGS).

To date, SNP-based heritability estimates of complex behaviours have been based on a few million common variants. With imputed variants, the quality of imputation depends on the reference panel used\(^19,20\), and even the best imputation strategies perform poorly for variants with MAF < 1\% in population-based samples\(^11,12\) of unrelated individuals. With the advent of relatively affordable deep WGS, it is now possible to directly genotype variants of lower frequencies in larger samples. While genetic association studies may be underpowered to detect an association between a given single rare variant and a complex trait\(^21\), recent extensions of mixed-effects models allow one to estimate a random effect representing the aggregate contribution of rare variants to phenotypic variance over and above common variation\(^11,12,13,22\). To date, a small number of recent WGS studies have reported evidence that rare variants account for a substantial part of the heritability of anthropometric, transcriptomic and medical phenotypes\(^23–25\) (but see also refs. \(^26,27\) for counterexamples). Notably, rare variants, especially those in regions of low LD, capture a large part of the missing heritability for height and body mass index, albeit with large standard errors\(^28\).

Here we used deep WGS (mean depth 30x) from the Trans-Omics for Precision Medicine (TOPMed) programme to estimate the heritability explained by variants as rare as MAF of 1\% in 10,000 for individuals of European ancestry (up to 26,257 individuals) and 1 in 1,000 for African admixed individuals (up to 11,743 individuals). We performed extensive sensitivity analyses to test the influence of rare-variant population structure along with various analytic parameters on our SNP-based heritability estimates for tobacco use. The study protocol was approved by the Institutional Review Board at the University of Minnesota and the TOPMed consortium. Informed consent was obtained originally by the participating studies.

### Results

#### Heritability estimates

We estimated the heritability of four smoking phenotypes in samples of European ancestry: (1) age of smoking initiation \((AgeSmk, N = 14,709)\), which measures the age at which an individual started regularly smoking; (2) cigarettes smoked per day \((CigDay, N = 15,384)\), an index of heaviness of smoking and the average number of cigarettes smoked per day as a current or former smoker, grouped into five bins, with higher numbers indicating greater use; (3) smoking cessation \((SmkCes, N = 17,827)\), a binary variable indicating whether a person is a current smoker; and (4) smoking initiation \((SmkInit, N = 26,257)\), a binary variable indicating people who ever smoked regularly (that is, over 100 cigarettes) in their life. We stratified SNPs and short indels by their MAF and LD so that we had a total of six bins (MAF 5–50\% and high LD, MAF 1–5\% and high LD, MAF 1–5\% and low LD, MAF 5–50\% and low LD, MAF 1–5\% and low LD, MAF 0.1–1\%, and MAF 0.01–0.1\%). Variants were assigned to the high-LD group when their LD scores were higher than the median of the variants in the same MAF bin (Table 1). After we adjusted for the fixed effects of ten common variant principal components (PCs), ten rare variant PCs, age, sex and a random effect of study, the total SNP-based heritability \((h^2_{SNP})\) from variants of all MAFs was 0.226 (s.e., 0.116) for AgeSmk, 0.134 (0.095) for CigDay, 0.283 (0.127) for SmkCes and 0.225 (0.096) for SmkInit (Fig. 1). Heritability estimates from the four common-variant bins (that is, MAF 1–5\%)—which include two bins for variants with high LD scores and two bins for variants with low LD scores—were summed to compute the total heritability attributable to common variants \((h^2_{common})\). Rare-variant heritability \((h^2_{rare})\) was computed likewise from the two MAF 0.01–1\% bins, which were not grouped by LD score because most rare variants have low LD (Methods). Across the four smoking phenotypes, rare variants accounted for 35–74\% of the

| Table 1 | Sample size and number of variants per MAF/LD bin in the European ancestry sample |
|---------|-----------------------------------------------|
|         | Sample size | 5–50\%, HI | 5–50\%, LO | 1–5\%, HI | 1–5\%, LO | 0.1–1\% | 0.01–0.1\% |
| AgeSmk  | 14,709      | 3,092,517   | 3,092,534  | 1,342,734 | 1,342,736 | 5,392,813 | 28,280,118 |
| CigDay  | 15,384      | 3,092,240   | 3,092,269  | 1,341,881 | 1,341,882 | 5,435,505 | 20,415,037 |
| SmkCes  | 17,827      | 3,092,593   | 3,092,594  | 1,340,764 | 1,340,771 | 5,413,019 | 23,483,166 |
| SmkInit | 26,257      | 3,092,454   | 3,092,475  | 1,341,068 | 1,341,071 | 5,395,579 | 21,108,704 |

* This shows the number of variants per bin in one-interquartile-range (IQR) unrelated samples \((k < 0.025)\). † Sample size of unrelated individuals. HI, high LD; LO, low LD.
total heritability. The majority of the heritability attributable to rare variants was in the rarest frequency bin (MAF 0.01–0.1%), except for CigDay, where most $\hat{h}^2_{SNP}$ was attributed to variants with MAF 0.1–1% (Supplementary Table 1). When we further partitioned rare variants according to their functional impact, most of the heritability was localized to lower-functional-impact (non-protein-altering) variants (Supplementary Table 2).

Sensitivity analyses. We used a variety of sensitivity analyses to evaluate the influence of residual population structure and cryptic relatedness. First, we evaluated the influence of increasing the number of PCs from 20 to 40 and 100 (half of which derived from common and rare variants) on $\hat{h}^2_{SNP}$ to test whether greater numbers of PCs account for potential residual population structure. The results for AgeSmk, SmkCes and SmkInit were unaffected (Fig. 2 and Supplementary Table 3). We observed model convergence issues when correcting CigDay for a large number of PCs, resulting in negative $\hat{h}^2_{SNP}$ (for example, $-0.12$) attributable to the rarest MAF bin, which we deemed well outside the plausible range.

Second, we tested whether $\hat{h}^2_{SNP}$, especially that of rare variants, was driven by individuals sharing long identity-by-descent (IBD) segments. Using FastSMC42, we detected about 7.5 million IBD segments longer than 2 cM and shared among at least two individuals (mean length, 2.50 cM; s.d., 0.50 cM). After we removed up to 178 individuals whose shared IBD segments stretched longer than 2.5% of the total genome length, $\hat{h}^2_{SNP}$ did not show meaningful differences, with changes in $\hat{h}^2_{SNP}$ ranging from $-0.025$ to 0.031 (Fig. 2 and Supplementary Table 3). Third, to assess the influence of distant

---

**Fig. 1 | SNP-based heritability estimates in the European ancestry sample for each of the six MAF/LD bins, and sums across bins.** The error bars represent standard errors. ‘Rare’ is the sum of the MAF 0.1–1% and MAF 0.01–0.1% bins. ‘Common’ is the sum of the other MAF bins. ‘Total’ is the sum of ‘Rare’ and ‘Common’. All estimates were adjusted for demographic variables and 20 PCs (half of them from rare variants) as fixed effects along with a random effect of cohort, except for CigDay, which was adjusted for five common PCs to allow model convergence.
Fifth, we obtained the approximate null distribution of $\hat{h}^2_{SNP}$ under the assumption that population structure explains the observed SNP-based heritability (that is, the genotypes of ancestrally close individuals are exchangeable). This was done by estimating $\hat{h}^2_{SNP}$ in samples randomly permuted within individuals close in ancestry as determined by a weighted sum of the top ten PCs with weights proportional to their corresponding eigenvalues. We computed $\hat{h}^2_{null}$ from 100 permuted trials and found limited evidence of inflation, as mean heritability estimates from the permuted trials were near zero across all bins of the four smoking phenotypes (Supplementary Table 4).

Finally, we estimated $\hat{h}^2_{SNP}$ in samples with decreasing control of population stratification by ascertaining samples with less family structure on heritability estimates; we adjusted for the top 20 PCs from an IBD-based genetic relatedness in addition to the 20 PCs from rare and common variants and found that $\hat{h}^2_{SNP}$ was not affected, with the largest change in $\hat{h}^2_{SNP}$ being 0.013 (Fig. 2 and Supplementary Table 3).

Fourth, previous studies indicated that rare variants are more likely to show geographical specificity. While such geographic specificity was at least partially captured by our inclusion of study as a random effect, we went further by including a random effect of recruitment site (which is nested within a given study) in the model. After adjustment of site, $\hat{h}^2_{rare}$ of SmkCes was about 50% of the original estimate, while $\hat{h}^2_{SNP}$ of the other smoking traits showed little change (Fig. 2 and Supplementary Table 3).

**Fig. 2 | SNP-based heritability estimates in the European ancestry sample from sensitivity analyses.** The error bars represent standard errors. The figure shows SNP-based heritability estimates from different sensitivity conditions. Heritability was estimated after adjusting for 20 common and 20 rare variant PCs (‘40 PCs’), or 50 common and 50 rare variant PCs (‘100 PCs’), after removing individuals who share IBD segments more than 2.5% of the total genome length (‘Long IBD’), after adjusting for the top 20 PCs from the IBD-based GRM (‘IBD PCs’), and after adjusting for recruitment site as a random effect (‘Site’).
The relatedness of closely related individuals was estimated from included all individuals with at least one close relative in TOPMed. These heritability estimates were derived using a single GRM that error bars denote standard errors. ‘Pedigree’ and ‘Pedigree-FHS’ refer the figure shows SNP heritability estimates across different studies.

The same selective pressures may apply to the heritability patterns found for common variants in this study. We found that most of the heritability from common variants was attributed to low-LD variants in the European ancestry samples. Consistent with this, Gazal et al. reported that common variants with low LD in low-recombination-rate regions had larger per-SNP heritability than those with high LD. One possible explanation for this finding is that low-LD common variants are more likely to have arisen recently than high-LD variants. The low-LD variants have therefore had less time to be pushed to low frequency or fixation by selection pressure, comprising a major source of the heritability attributed to common variants.

Our suggestion that smoking traits have been under negative selection might seem surprising. While nicotine itself has been pervasive in the environment for millions of years, tobacco in its

stringent PC-based ancestry filtering thresholds (Methods). The largest change in was observed for SmkCes, where the estimates were up to 0.10 higher in samples with less stringent PC-based filtering than in the primary analysis, possibly reflecting sampling variations and residual population structure (Supplementary Fig. 1 and Supplementary Table 3).

In summary, in 36 comparisons (4 phenotypes×9 conditions), showed relatively large changes only for SmkCes when we adjusted for geographical site or relaxed ancestry filtering. Our primary results presented in Fig. 1 are from the most stringent PC-filtering condition.

Pedigree-based heritability estimates. We removed closely related individuals when estimating the heritabilities described above. However, we were able to leverage thousands of related family members in TOPMed to independently estimate the total narrow-sense heritability of our four smoking phenotypes in pedigrees. These heritability estimates were derived using a single GRM that included all individuals with at least one close relative in TOPMed. The relatedness of closely related individuals was estimated from common variants, which, when estimated for close family members, serves as a proxy for rare and common variant sharing. The relatedness of classically unrelated individuals (relatedness estimate, serves as a proxy for rare and common variant sharing. The low-LD variants in the European ancestry samples. Consistent with this, Gazal et al. reported that common variants with low LD in low-recombination-rate regions had larger per-SNP heritability than those with high LD. One possible explanation for this finding is that low-LD common variants are more likely to have arisen recently than high-LD variants. The low-LD variants have therefore had less time to be pushed to low frequency or fixation by selection pressure, comprising a major source of the heritability attributed to common variants.

Using related individuals in TOPMed across cohorts (N up to 21,546), narrow-sense heritabilities for AgeSmk, CigDay, SmkCes and SmkInit were estimated at 0.343 (0.075), 0.175 (0.066), 0.179 (0.067) and 0.288 (0.057), respectively. To check whether was downwardly estimated due to a relatively large proportion of unrelated pairs, we also estimated pedigree-based heritability in the Framingham Heart Study (FHS) (N up to 3,024), which is composed of families as part of the study design. Similar estimates (AgeSmk, 0.225; CigDay, 0.244; SmkCes, 0.195; SmkInit, 0.299) were obtained in the FHS as in the larger analysis of all available families, albeit with larger standard errors. The estimates from the pedigree analysis are presented in Fig. 3 and Supplementary Table 5.

Heritability estimates in African ancestries. We explored attributable to variants with MAF down to 0.1–1% for SmkInit (N=11,743) and common variants for the other smoking phenotypes (N=6,796–7,549) in samples of African ancestries. After we adjusted for 100 PCs, admixture proportions of five ancestries, local ancestry and the same demographic covariates as in European ancestry, common-variant-based SNP heritability was estimated as follows: 0.098 (0.091) for AgeSmk, 0.028 (0.113) for CigDay, 0.075 (0.120) for SmkCes and 0.085 (0.10) for SmkInit. Variants from MAF 0.1–1% additionally accounted for about 24% of the phenotypic variance of SmkInit, leading to total SNP-based heritability (combining both common and low-frequency variants) of SmkInit of 0.329 (0.144). Heritability estimates for individual bins are reported in Table 2 and Supplementary Table 6.

Discussion

Using up to 26,257 whole-genome sequences, we found that about 13–28% of the phenotypic variance of four smoking phenotypes in individuals of European ancestry could be attributed to genetic variants with minimum MAFs as low as 1 in 10,000. This is about 1.5–4.5 times larger than the previous SNP-based heritability estimates based on common variants alone.

This increase was largely driven by the inclusion of rare variants with MAF 0.1–0.01%, which accounted for 35–74% of the estimated SNP-based heritability of smoking phenotypes in samples of European ancestries. Smoking cessation showed the highest (0.209; s.e., 0.123), followed by age of smoking initiation (0.153; s.e., 0.114), smoking initiation (0.079; s.e., 0.091) and cigarettes per day (0.054; s.e., 0.092). Different smoking phenotypes would in principle vary in their genetic architectures if they differentially relate to fitness or fitness-related traits. However, the standard errors associated with these estimates prevent us from drawing strong conclusions about their relative magnitudes.

The contributions of rare variants observed here seem consistent with the action of negative selection. Under negative selection, harmful mutations are maintained at low frequency in the population, whereas under the neutral model, common variants are expected to explain the majority (>90%) of the heritability.

The same selective pressures may apply to the heritability patterns found for common variants in this study. We found that most of the heritability from common variants was attributed to low-LD variants in the European ancestry samples. Consistent with this, Gazal et al. reported that common variants with low LD in low-recombination-rate regions had larger per-SNP heritability than those with high LD. One possible explanation for this finding is that low-LD common variants are more likely to have arisen recently than high-LD variants. The low-LD variants have therefore had less time to be pushed to low frequency or fixation by selection pressure, comprising a major source of the heritability attributed to common variants.

Our suggestion that smoking traits have been under negative selection might seem surprising. While nicotine itself has been pervasive in the environment for millions of years, tobacco in its
cigarette form represents an evolutionarily novel environment for humans, and the selection against such traits is likely to be weak due to its recency. Nevertheless, genetic variants influencing smoking can affect multiple different traits more directly related to evolutionary fitness. For example, smoking is highly genetically correlated with age at onset of reproductive behaviours (especially age of smoking initiation) and with many social and health outcomes such as education and metabolic diseases. Selection pressure on these correlated traits can therefore influence the genetic architecture of smoking phenotypes. More studies are needed to understand the mechanisms that maintain the contribution of rare genetic variants to the heritability of smoking phenotypes in the population.

We found that almost all the contribution from rare variants to $h^2_{SNP}$ was attributed to genetic variants classified as having relatively low expected protein-altering consequences. This may suggest an importance of rare regulatory genetic variation in the etiology of smoking behaviours. Despite limited contributions to heritability from protein-altering variants, it is still possible that they explain higher heritability individually given the smaller number of such variants (about 0.2 million protein-altering versus 20 million non-protein-altering variants in MA \{1–50\%\}) and the presumably higher selective pressure against them. Unfortunately, high standard errors and overall small heritabilities from the protein-altering variants bin prevented us from drawing strong conclusions from the present data.

Our SNP-based heritability estimates were largely robust to our sensitivity checks. Heritability changed little after we increased the number of PCs and tested finer-scale population structure via permuting individuals that shared several long IBD segments or adjusting for geographical site as a binomial variable. Cigarettes per day showed the lowest SNP-based heritability reported in the literature. There remain many explanations, including ultra-rare variants (MA \{1–50\%\}) across the four phenotypes, closing the missing heritability gap for smoking phenotypes. Note that our $h^2_{SNP}$ values are estimated from pedigrees from multiple cohorts, who also may span different generations (for example, parents–offspring) and share markedly different smoking environments. We consider our $h^2_{SNP}$ as the most relevant benchmark by which to compare $h^2_{SNP}$ from our genome-based restricted maximum likelihood (GREML) results and past GWAS studies, which also have varying sources of sample heterogeneity, including gender and environment interaction and varying heritability. Twin studies, in contrast, usually consist of large, homogeneous cohorts with well-defined pedigrees sharing most of the demography (that is, monozygotic and dizygotic twins).

For individuals of African ancestries, we found that common SNPs and indels (MA > 1%) accounted for 3–15% of the phenotypic variance of smoking phenotypes, similar in magnitude to existing studies. We observed an additional contribution of rare variants (MA \{0.1–1\%\}) for smoking initiation, accounting for about 25% of smoking initiation liability. Note that estimating heritability in samples with complex population structure is in its infancy, and how recent admixture influences causal allele frequency spectra is not well understood. More research on this topic will be welcome.

Our findings should be interpreted in light of several limitations. First, although we used the largest WGS sample reported to date for heritability analysis, even larger datasets will be important. Larger sample sizes would allow greater precision in estimation and a more comprehensive assessment of the genetic architecture of these complex traits by finer partitioning by MA and functional annotations. Second, even with the use of deep sequences, we did not fully reach the trait heritability as estimated using available pedigrees or twin heritability reported in the literature. There remain many explanations, including ultra-rare variants (MA \{0.01–0.1\%\}), other types of genetic variations (for example, copy number variations) that are not in LD with the variants used here and still other sources of heritable variation. Cigarettes per day showed the lowest SNP-based heritability and often encountered model convergence issues, especially with a large number of covariates. Since it is a binned variable, its heritability may be more robustly estimated on a liability scale.

Third, although we carefully assessed the potential influence of population structure in extensive sensitivity analyses, existing methods may not be entirely capturing the population stratification of rare variants. More studies are needed to understand the extent and the best ways to handle it, although we are also sensitive.
to the possibility that such corrections may be counterproductive, removing meaningful heritable variance. Fourth, our pedigree-based heritability estimates may be inflated by shared environment, as we were unable to model genetic similarity and environmental similarity separately. The pedigree estimates therefore may best be interpreted as an upper bound on the SNP-based narrow-sense heritability. However, we expect the magnitude of inflation due to shared environment to be smaller than in typical twin studies given that our pedigree sample includes multiple classes of family relations who share much less of their environments than twins would.

Fifth, there is ample documentation of spousal similarity in smoking behaviours14–16, and one of the potential sources of this similarity, assortative mating, can artefactually inflate marker-based heritability estimates21. While recent studies have not found evidence for primary assortative mating on smoking traits in the genome22, more research is needed on this topic.

Sixth, genetic association signals of human behavioural traits can arise from multiple sources, not only from variants’ downstream biological influences on a phenotype of interest but also from its genetically correlated traits (for example, education and income) along with indirect effects of parents’ and siblings’ genetics23–25. Our SNP-based heritability estimates include influences from multiple sources and could overlap with polygenic signals of socio-economic factors given their substantial genetic correlations with smoking (|r| = 0.26–0.55)26. With bigger sample sizes and studies of close families, future research can localize the origins of SNP-based heritability and examine the genetic architecture of shared and distinct genetic components of smoking and its correlated phenotypes (such as education and income)27.

Seventh, smoking phenotypes were measured by one or two questions and were limited to those commonly collected in biomedical studies such as those in TOPMed. This allows accumulations of large sample sizes across multiple independently collected samples, but each measure is noisy and may reflect idiosyncrasies of the participating studies. We attempted to account for the latter through including a random effect of study. In conclusion, the present study expanded our understanding of the genetic architecture of tobacco use by showing that much of the missing heritability of smoking phenotypes can be explained by rare genetic variants. The use of WGS allows less biased estimation of SNP-based heritability by including virtually all potential causal SNPs down to a MAF of 0.01% in individuals of European ancestries and 0.1% in those of African ancestors, rather than relying on the degree to which imputed variants tag causal variants. The current study informs the genetic etiology of nicotine addiction and provides a benchmark for the future study of other complex behavioural traits.

**Methods**

**Sample.** We considered individuals of European ancestries in TOPMed (freeze 8; mean depth, >30×) measured for at least one of four smoking phenotypes for inclusion. Our data access application was approved by relevant cohorts in TOPMed. We determined European ancestries in two steps. First, we identified an initial ancestry-inclusive set by projecting TOPMed genotypes (N = 137,977) onto genetic PC axes from the 1000 Genomes project (1000 G) and then used a k-nearest-neighbour method to assign the ancestry of TOPMed individuals with 1000 G as a reference set. More specifically, we used online augmentation–decomposition–transformation to calculate the PC scores of TOPMed individuals, which implements Procrustes transformation with an augmented dataset (that is, combining TOPMed and 1000 G reference genomes; https://github.com/daviddaiwa/zheng/frappe)28. Then, for a given TOPMed sample, we chose the top 20 reference individuals in 1000 G who were the 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 initially classified as having European ancestry who also had at least one smoking phenotype. Second, we partitioned the data such that PCs 1–4 of the selected individuals, we suspected residual population heterogeneity, especially in PC 4 (Supplementary Fig. 2). We then further restricted the samples to those whose summed Euclidean distance of PCs 1–4 fell in the upper 1 IQR of the European sample (N = 38,915) identified in the first step. We additionally created samples using 1.5, 2 and 3 IQR and reserved them for sensitivity analysis (Supplementary Figs. 1 and 2). After IQR filtering, we retained only unrelated individuals. Relatedness was estimated with HapMap3 variants (Hardy–Weinberg equilibrium P > 10^-4, MAF > 0.01) using GCTA v1.92 to obtain a list of nominally unrelated individuals with pairwise relatedness (r < 0.025). We additionally removed 76–190 individuals who showed very high rare-variant sharing (r > 0.25 in the MAF 0.01–0.1% bin) who seemed to have inherited small IBDS from distant non-European ancestors. This resulted in the final sample size per phenotype shown in Table 1 (N ranging from 14,709 to 26,257).

**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 questions at different stages of tobacco use, were available across most TOPMed studies. We used the same definition and coding scheme as Liu et al.29. AgeSmk measures the age at which an individual started regularly smoking. CigDay is the average number of cigarettes smoked per day as a current or former smoker and is grouped into five bins, with higher numbers indicating greater use. For both AgeSmk and CigDay, lifelong non-smokers are excluded (set to missing). SmkCes and SmkMin are binary variables indicating former versus current smokers and never versus ever regular smokers, with case defined as current and ever smoker, respectively. These four variables were correlated but not redundant, each measuring distinct aspects of smoking (Supplementary Note and Supplementary Table 7). Descriptive statistics for each phenotype across cohorts can be found in Supplementary Table 8.

**Genotypes, LD scores, GRMs and GREML-LDMS-I.** GREML estimates heritability by comparing phenotypic similarity to observed genetic similarity among distantly related individuals using a linear mixed model30. GREML yields biased estimates when causal variants are unevenly distributed as a function of LD and MAF31. To mitigate this bias, the GREML-LDMS-I method partitions SNPs into different LD×MAF bins. We initially considered ~710 million genotypes that passed strict quality filters described elsewhere32. We additionally removed 95,750 variants with Hardy–Weinberg equilibrium P values less than 10^-16 in the European ancestry 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 calculated LD scores then further stratified the above-described LD scores into median LD scores within the two most common MAF bins. This resulted in six bins: MAF (0.05, 0.5) and high LD, MAF (0.05, 0.5) and low LD, MAF (0.01, 0.05) and high LD, MAF (0.01, 0.05) and 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 because more bins decrease estimate precision. LD scores, defined as the sum of squared correlations (r^2) between a variant and all the variants in a region, were calculated using GCTA v1.92 with the default 10 Mb window in the final sample combined across the four smoking phenotypes (a summary of LD scores is presented in Supplementary Table 9). This process resulted in approximately 35 million SNPs and indels (Table 1). For each phenotype, we performed GREML-LDMS-I analysis using the three methods for the above-mentioned six bins and a cohort indicator matrix as random effects. The cohort matrix was an NxN matrix indicating whether a given pair of individuals belonged to the same study (1, otherwise 0). The variance explained by cohort was included in the total phenotypic variance, which could yield somewhat smaller heritability estimates than when the heritability is conditioned on cohort. We regressed out age, squared age, sex and their two-way interaction terms33–37 from the quantitative phenotypes (that is, AgeSmk and CigDay) and used residuals. The residuals of CigDay were further inverse-rank normalized to facilitate model convergence by smoothing its distribution. GREML-LDMS-I results from other types of transformations for AgeSmk and CigDay are presented in Supplementary Table 3. We calculated 10 PCs from common variants (MAF ≥ 1%) and another 10 PCs from rare variants (0.01% < MAF < 1%) after LD pruning (plink –indep-pairwise 100 5 0.2 for common and 2,000 10 0.02 for rare variants). These 20 PCs and the sequencing centre were entered as fixed effects in the linear mixed model. The same demographic variables and 20 PCs were entered as fixed effects for the binary variables. We used GCTA v1.92 for both the construction of GRMs and the GREML-LDMS-I analysis. We replaced the original GRM diagonals with 1+ inbreeding coefficients (Fhat3 computed from the GCTA software) because the original formula tended to yield far higher variance of diagonals for rare variants than expected in theory (Supplementary Note and Supplementary Table 10). We allowed the estimates to be negative to obtain unbiased estimates. Heritability estimates were analyzed under various using models30. Population prevalence was set at 0.15 and 0.42 for SmkCes and SmkMin, respectively, on the basis of 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 the heritability estimates of the six bins with standard errors approximated by the delta method38.

**Partitioning rare-variant heritability.** To further interrogate sources of rare-variant heritability, we divided rare-variant bins into protein-altering versus non-protein-altering variant bins39. Functional impact of variants was assessed by
We initially had 19,788 individuals with SNP-based heritability in African ancestries. For population structure, we additionally excluded 58 individuals who showed by the ancestry assignment procedure described earlier. We selected unrelated individuals at each genomic segment of two homologous chromosomes, using the Human Genome Diversity Panel as a reference panel (Supplementary Fig. 3). Given the smaller variance of African ancestry individuals in variants to be analysed by RFMix75, which estimates the ancestry of admixed individuals expected to influence the effectiveness of the protein, such as missense and splice region variants. ‘MODIFIER’ includes variants that are more than two-way admixture as determined by global ancestry proportions of individuals of African ancestries, as SNP-based relatedness is no longer estimated due to the relatively low level of relatedness structure in the population.

We then randomly exchanged genotypes of a given individual with one of their 100 nearest neighbours using sampling algorithms from LocPerm, which was done to control for population structure in the rare-variant association test. We estimated heritability in 100 replicates of permuted genotypes with the same demographic covariates and cohort used in the main analysis. The greater the IQR threshold, the more ancestral variation can be attributed to rare variants in a sample resulting in the sample sizes in Table 2. We then derived SNPs from the dbgap TOPMed exchange area.

Pedigree-based heritability. We created a GRM with all available samples, including related ones but excluding pairs related greater than 0.80 to exclude identical twins and duplicates. To aid in model identification, we created a panel of SNPs that had at least 5% of SNPs that are related greater than 0.375. A list of cohorts included in this analysis is presented in Supplementary Table 11. A set of demographic and cohort covariates as in the European ancestry analyses. We performed GREML-LDMS-I with the same set of demographic and cohort covariates as in the European ancestry analyses. Sensitivity analyses. To evaluate the degree of sensitivity of our results to residual population structure (especially that of rare variants) and various analytic decisions, we conducted extensive sensitivity analyses. First, we performed GREML-LDMS-I with 10 PCs (20 common and 20 rare) and 100 PCs (50 common and 50 rare) as fixed effects. Second, we estimated heritability after removing individuals whose total length of shared IBD segments covered more than 11,744) and common variants (MAF > 0.025) by applying the pcairPartition function in the GENESIS package. Accession numbers and email addresses of the principal investigators are presented in the Supplementary Note. Genetic data are available through the dbgap TOPMed exchange area.

Received: 28 April 2021; Accepted: 10 June 2022; Published online: 4 August 2022

References
1. Johnson, T. & Barton, N. Theoretical models of selection and mutation on quantitative traits. *Phil. Trans. R. Soc. B* 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. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in 100,000 individuals of European ancestry. *Hum. Mol. Genet.* 27, 3641–3649 (2018).
4. Graham, S. E. et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature* 600, 675–679 (2021).
5. Vujkovic, M. et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* 52, 680–691 (2020).
6. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427 (2014).
7. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* 467, 786–792 (2009).
8. Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S. & Murray, C. J. Comparative risk assessment of disease and injury attributable to 7000 death and disability conditions in 21 world regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 380, 225–234 (2012).
9. Carter, B. D. et al. Smoking and mortality—beyond established causes. *N. Engl. J. Med.* 372, 631–640 (2015).
10. Maes, H. H. et al. A genetic epidemiological mega analysis of smoking initiation in adolescents. *Nicotine Tob. Res.* 19, 401–409 (2017).
11. Vink, J. M. & Boomsma, D. I. Interplay between heritability of smoking and environmental conditions? A comparison of two birth cohorts. *BMC Public Health* 11, 316 (2011).
12. Quach, B. C. et al. Expanding the genetic architecture of nicotine dependence and its shared genetics with multiple traits. *Nature communications* 11, 1–13 (2020).
13. Degenhardt, L. & Hall, W. The relationship between tobacco use–substance-use disorders and mental health: results from the National Survey of Mental Health and Well-Being. *Nicotine Tob. Res.* 3, 225–234 (2001).
14. McCabe, S. E., West, B. T. & McCabe, V. V. Associations between early onset of e-cigarette use and cigarette smoking and other substance use among US adolescents: a national sample. *Addiction* 103, 751–759 (2014).
15. King, S. M., Iacono, W. G. & McGue, M. Childhood externalizing and internalizing psychopathology in the prediction of early substance use. *Addiction* 99, 1548–1559 (2004).
16. Polderman, T. J. C. et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* 47, 702–709 (2015).
17. Liu, M. et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* 51, 237–244 (2019).
19. Errazuriz-Aguuia, A. M. et al. Meta-analysis of up to 622,409 individuals identifies 40 loci influencing behaviour associated genetic loci. Mol. Psychiatry 25, 2392–2409 (2020).

20. Evans, L. M. et al. Genetic architecture of four smoking behaviors using partitioned SNP heritability. Addiction 116, 2498–2508 (2021).

21. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11, 446–450 (2010).

22. Zaitlen, N. et al. Using extended genotype to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet. 9, e1003520 (2013).

23. Wray, N. R. & Maier, R. Genetic basis of complex trait and its implications for fitness and genome-wide association studies. Proc. Natl Acad. Sci. USA 107, 1752–1756 (2010).

24. Visscher, P. M., Goddard, M. E., Derks, E. M. & Wray, N. R. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. Mol. Psychiatry 17, 474–485 (2012).

25. Young, A. I. Solving the missing heritability problem. Nat. Commun. 5, 446–450 (2014).

26. Gibson, G. Rare and common variants: twenty arguments. Nat. Rev. Genet. 13, 135–145 (2012).

27. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl Acad. Sci. USA 107, 1752–1756 (2010).

28. Visscher, P. M., Goddard, M. E., Derks, E. M. & Wray, N. R. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. Mol. Psychiatry 17, 474–485 (2012).

29. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 47, 1114–1120 (2015).

30. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl Acad. Sci. USA 107, 1752–1756 (2010).

31. YANG, J. et al. Recovery of trait heritability from whole genome imputation. Nat. Genet. 49, 186–192 (2017).

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

33. Wainwright, P. et al. Recovery of trait heritability from whole genome sequence data. Preprint at bioRxiv https://doi.org/10.1101/588020 (2021).

34. Nait Saada, J. et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nat. Commun. 11, 6130 (2020).

35. Nelson, M. R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 350, 290–299 (2021).

36. Wainschtein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat. Genet. 54, 263–273 (2022).

37. Hernandez, R. D. et al. Ultrarare variants drive substantial cis heritability of human gene expression. Nat. Genet. 51, 1349–1355 (2019).

38. Sul, J. H. et al. Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates. Transl. Psychiatry 10, 74 (2020).

39. Halvorsen, M. et al. Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. Nat. Commun. 11, 1842 (2020).

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

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

42. Wainschtein, P. et al. Recovery of trait heritability from whole genome sequence data. Preprint at bioRxiv https://doi.org/10.1101/588020 (2021).

43. Nait Saada, J. et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nat. Commun. 11, 6130 (2020).

44. Nelson, M. R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 350, 290–299 (2021).

45. Muller, J. et al. Taking population stratification into account by local permutations in rare-variant association studies on small samples. Genet. Epidemiol. 45, 821–829 (2021).

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

47. Mills, M. C. et al. Identification of 371 genetic variants for age at first sex and birth linked to externalising behaviour. Nat. Hum. Behav. 5, 1717–1730 (2021).

48. Domingue, B. W., Rehkopf, D. H., Conley, D. & Boardman, J. D. Geographic clustering of polygenic scores at different stages of the life course. RSF 4, 137–149 (2018).

49. Abdellouaou, A. et al. Genetic correlates of social stratification in Great Britain. Nat. Hum. Behav. 3, 1332–1342 (2019).

50. Tropf, F. C. et al. Hidden heritability due to heterogeneity across seven populations. Nat. Hum. Behav. 1, 757–765 (2017).

51. Boardman, J. D., Blalock, C. L. & Pampel, F. C. Trends in the genetic influence of smoking. J. Health Soc. Behav. 51, 108–123 (2010).

52. Bi, W. et al. Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes. Am. J. Hum. Genet. 108, 825–839 (2021).

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

54. Zaidi, A. A. & Mathieson, I. Demographic history mediates the effect of stratification on polygenic scores. elife 9, e61548 (2020).

55. Treur, J. L., Vink, J. M., Boomsma, D. I. & Middeldorp, C. M. Spousal resemblance for smoking: underlying mechanisms and effects of cohort and age. Drug Alcohol Depend. 184, 923–931 (2018).

56. Agrawal, A. et al. Assortative mating for cigarette smoking and for alcohol consumption in female Australian twins and their spouses. Behav. Genet. 36, 553–566 (2006).

57. Vink, J. M., Willemsen, G. & Boomsma, D. I. The association of current smoking behavior with the smoking behavior of parents, siblings, friends and spouses. Addiction 98, 923–931 (2003).

58. Border, R. et al. Assortative mating biases marker-based heritability estimators. Nat. Commun. (in the press).

59. Visscher, P. M., Goddard, M. E., Derks, E. M. & Wray, N. R. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. Mol. Psychiatry 17, 474–485 (2012).

60. Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait locus mapping. Genetics 138, 985–993 (1994).

61. Wray, N. R. & Maier, R. Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability. Curr. Epidemiol. Rep. 1, 220–227 (2015).

62. Hauser, R. F. & Steinberg, S. R. A new look at human population genetics in the genome era. Annu. Rev. Genet. 48, 249–276 (2014).
Acknowledgements

The molecular data for the TOPMed programme was supported by the National Heart, Lung and Blood Institute. 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, was provided by the TOPMed Informatics Research Center (R01HL-117626-02S1; contract no. HHSN268201800002C). Core support including phenotype harmonization, data management, sample-identity quality control and general programme coordination was provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract no. HHSN268201800001I).

We acknowledge the studies and participants who provided biological samples and data for TOPMed. Funding for this project included grant nos R01DA044283, R01DA037904 and R01HG008983 to S.V. and grant no. R01HL100141 to M.C.K. Cohort-wise acknowledgement is provided in the Supplementary Note. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.-K.J., S.V., M.C.K. and L.E. designed the study. A.F. contributed to the data analysis. All authors contributed to the data collection and curation and critically reviewed the manuscript.

Competing interests

B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. E.K.S. has received grant support from GSK and Bayer Research support to the University of Pennsylvania from Renalyx. AI and personal fees from Calico Labs, both outside the current work. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01408-5.

Correspondence and requests for materials should be addressed to Scott Vrieze.

Peer review information Nature Human Behaviour thanks Roseann Peterson, Abdel Abdellaoui and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

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

© The Author(s), under exclusive licence to Springer Nature Limited 2022
Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

- Data collection: Softwares and codes used in the collection and quality control procedure of sequence data are described here: https://www.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-8.

- Data analysis: GCTA1.92 was used to calculate LD scores and perform GCTA-REML analysis. We generated permutation sequences using LocPerm software. Plink1.9 was used to perform miscellaneous quality controls including genotype and sample missingness filtering and hardy-weinberg equilibrium statistics filtering. We used FRAPSOA to infer ancestry of individuals in TOPMed and GENESIS software package in R to identify unrelated individuals in African ancestry.

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

Data

Policy information about availability of data

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

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

Phenotype and genetic data are available through authorized access portal in dbgap [https://dbgap.ncbi.nlm.nih.gov/] or direct request to TOPMed Principal Investigators. Accession numbers and email addresses of Principal Investigators are presented in Supplementary Note.
Field-specific reporting

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

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

**Sample size**
- We simulated standard errors under a range of sample sizes and allele frequencies of genetic variants using GREML power calculator (https://cnsgenomics.com/software/gcta/GREMLpowercalculator). We found that current sample size would yield up-to 6-9% of SEs for SNP-based heritability when we include genetics variants with minor allele frequency (MAF) down to 0.01%. We decided that the results are still informative of the genetic architecture of behavioral traits and future research on this topic given current study includes the largest sample of WGS to date for heritability analysis of complex behavioral traits.

**Data exclusions**
- We excluded individuals with ancestry other than European or African ancestries. We excluded genetic variants with MAF lower than 0.01% based on the power analysis.

**Replication**
- Given lack of large independent samples of whole genome sequencing of smoking at the time, our results could not be reproduced in another dataset. We tested robustness of our results through various sensitivity analysis.

**Randomization**
- This study is observational study and no randomization was performed.

**Blinding**
- This study is observational study and no blinding procedure was applied.

Reporting for specific materials, systems and methods

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

### Materials & experimental systems

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

### Methods

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

Human research participants

Policy information about studies involving human research participants

**Population characteristics**

Community and clinical samples with European ancestry (56~64% females; mean age around 60 (SD: 11)) and those with African ancestry (50~60% females; mean age around 54 (SD: 12~13)).

**Recruitment**

We did not do any recruitment. We used existing de-identified data.

*This study was approved of Institutional Review Board at the University of Minnesota. Access to genetic and phenotype data was approved by individual cohorts in TOPMed consortium.*