Supplementary Note for:

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

Contents

1. Supplementary Figures

Figure S1. Overview of Analyses.

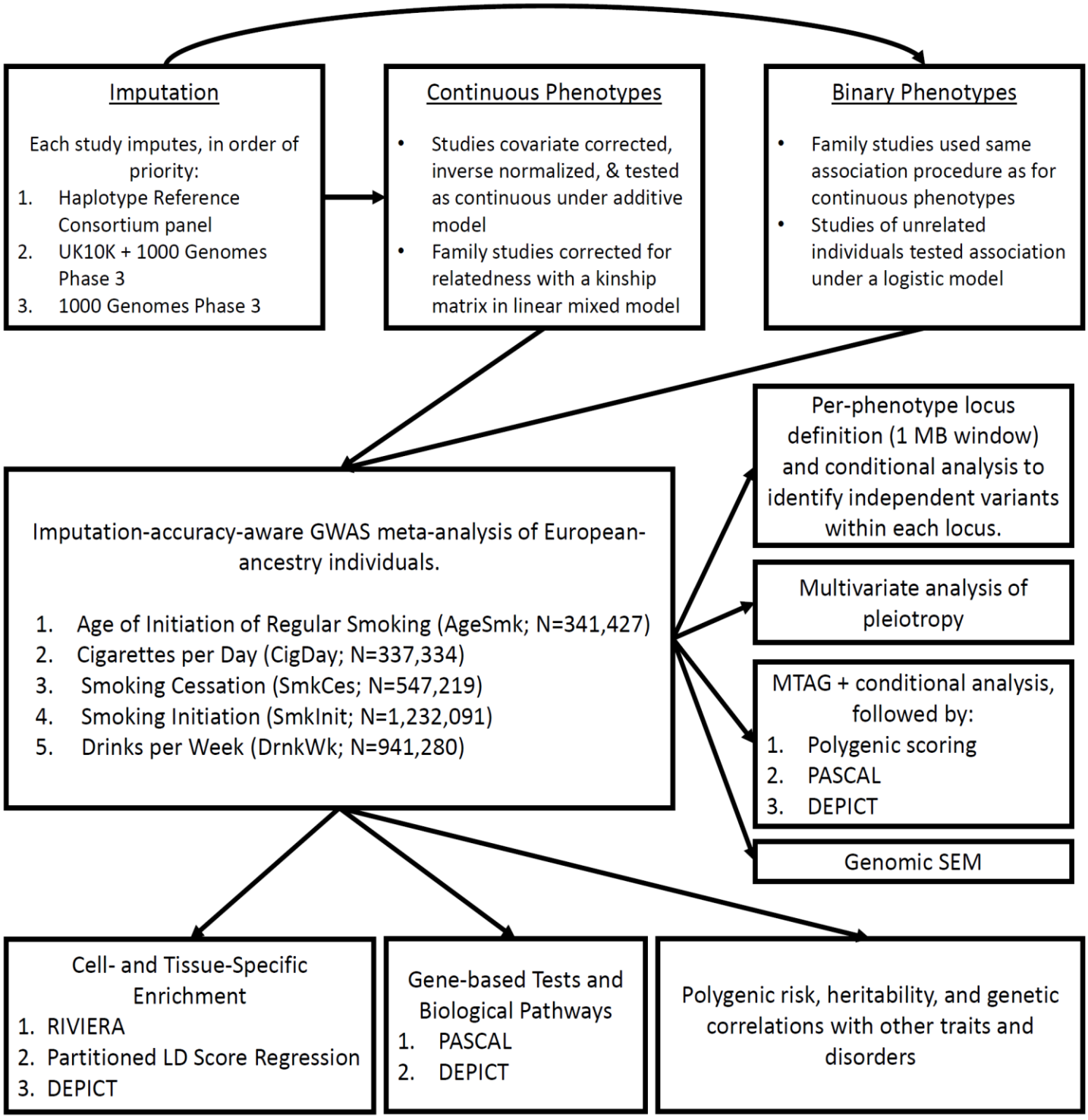
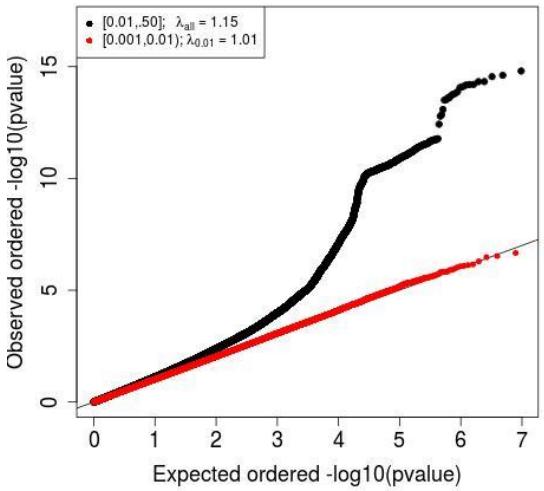
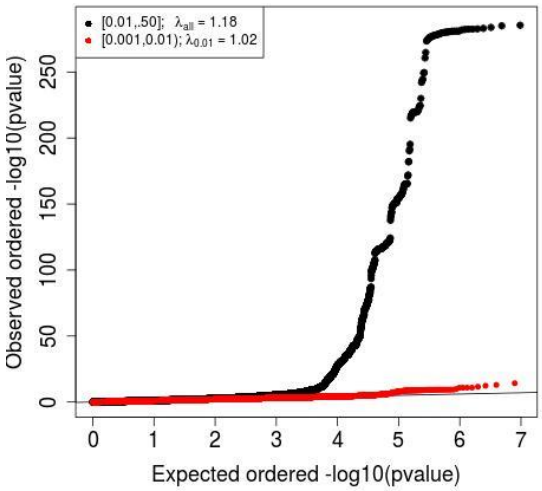


Figure S2. Meta-Analysis QQ Plots of Common (MAF ≥1%) and Low Frequency (.1% ≤ MAF < 1%) Variants. All LDSC intercepts were less than 1 (see **Table S25**), indicating that any observed inflation is unlikely due to population stratification, but to polygenic inheritance. Note, the most significant variant in DrnkWk had p-value $<1e^{-324}$ and is not plotted here. λ = genomic control.

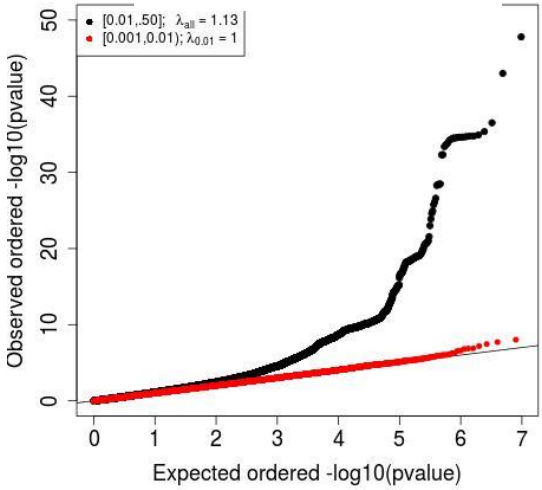
Age of Initiation of Smoking (AgeSmk)



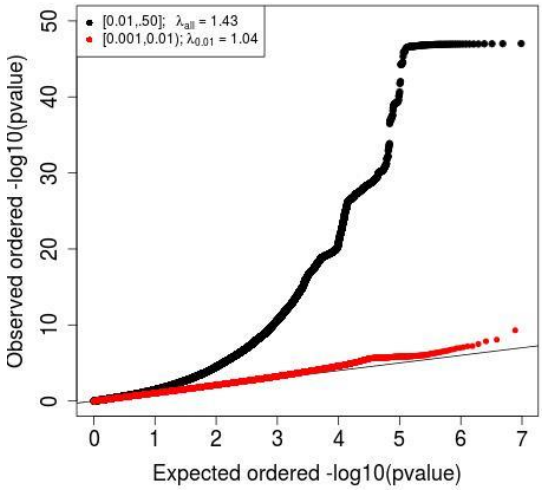
Cigarettes per Day (CigDay)



Smoking Cessation (SmkCes)



Smoking Initiation (SmkInit)



Drinks per Week (DrnkWk)

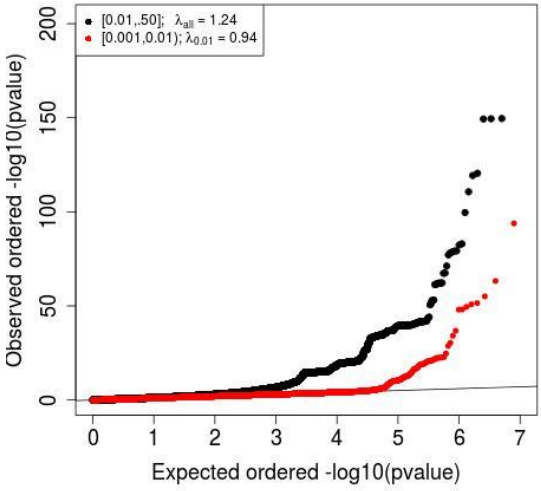
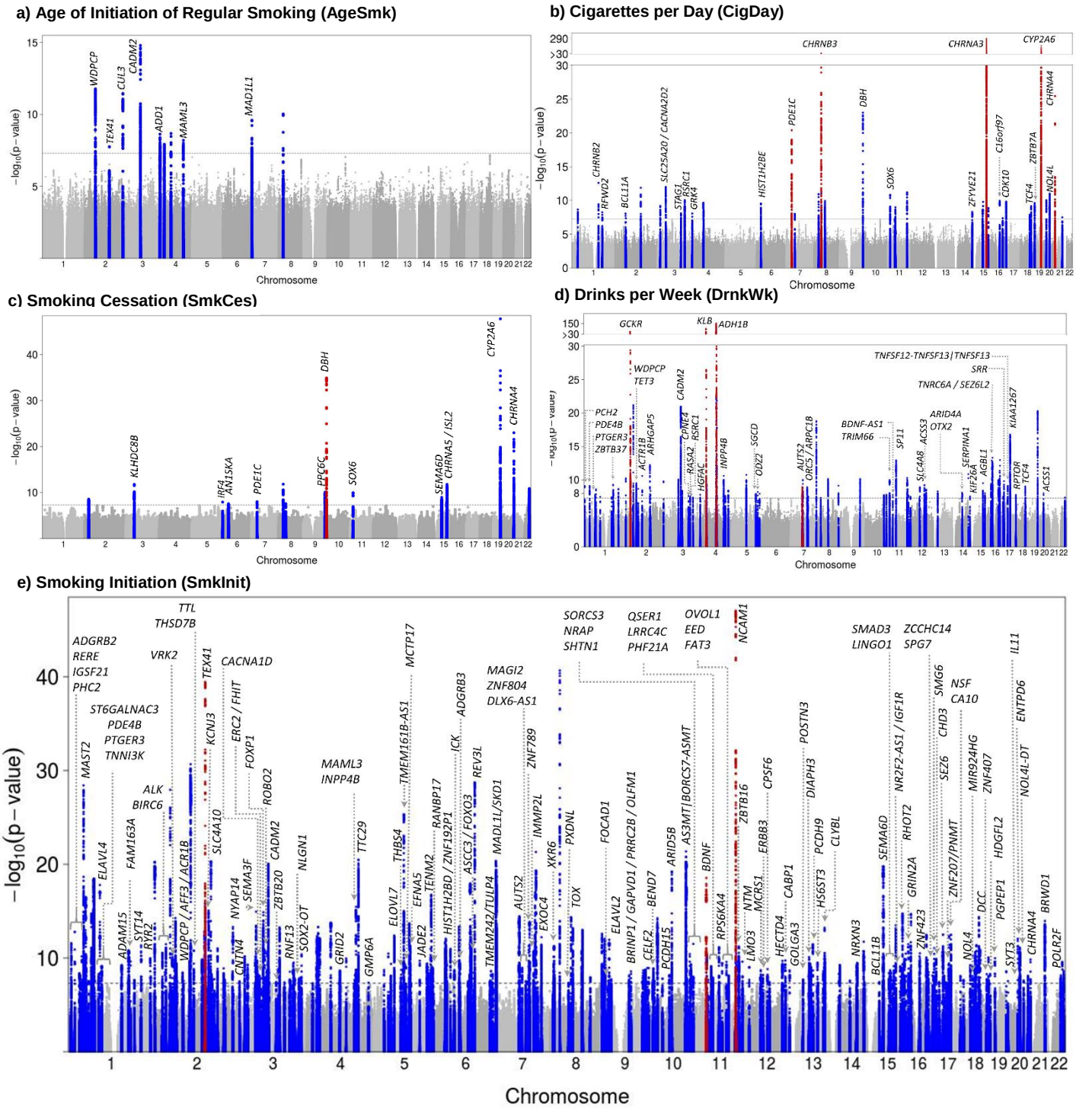


Figure S3. Manhattan plots of all five phenotypes. Gene names are listed only if the sentinel variant – the most significant variant – within the locus lies within the gene (5' to 3' UTR). Variants are dots, and are colored blue or red if they lie within the locus, defined as ± 500KB around the sentinel variant. Loci in red are known. Loci in blue are novel. The grey line corresponds to $p=5\times10^{-8}$, the standard GWAS significance threshold in individuals of European ancestry.



Figures S4-S12. Marginal and conditional LocusZoom plots of GWAS results.

Figure S13. Cell group enrichment from LDSC and RiVIERA-ridge. In the left panel are enrichment results across all cell types within each of the 10 cell groups (enrichment for specific cell types are available in **Figure S14**). On the right are results from a complementary method, RiVIERA-ridge, which imposes a different, sparser, model on the cell-type-specific enrichments, accounting for correlations among cell types. CNS cell group enrichment appears especially robust for AgeSmk, CigDay, DrnkWk, and SmkInit.

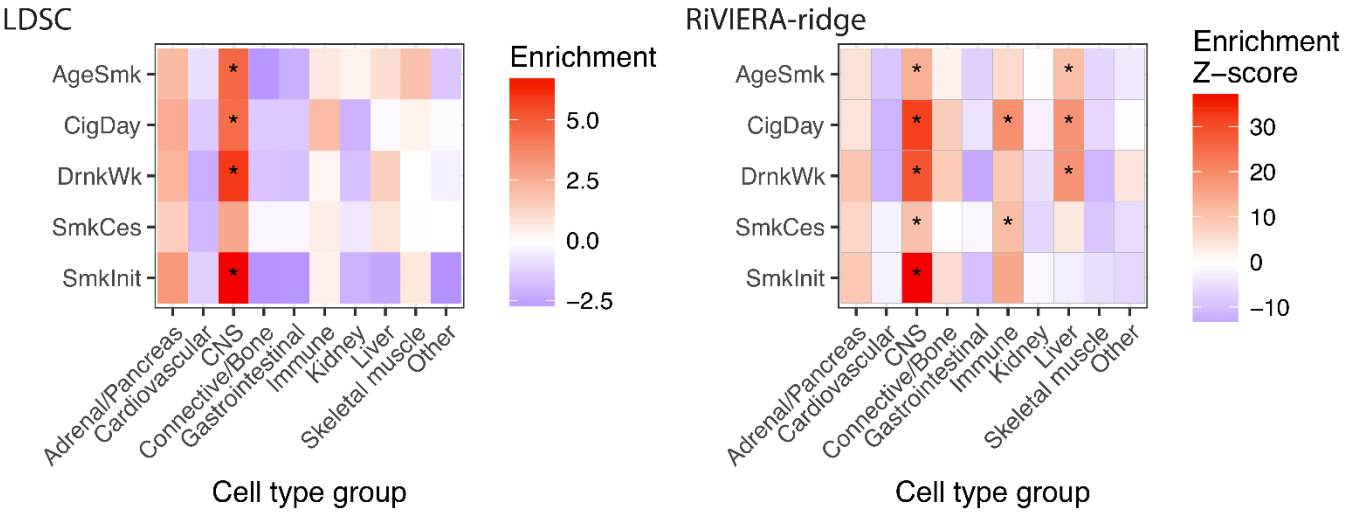


Figure S14. Cell-type-specific enrichment from LDSC. The figure shows cell-type specific enrichment from LD Score Regression for each of the four histone marks across 100 cell types and all five phenotypes. Enrichment with FDR < .05 is asterisked. Cells that are blank are those where no data is available for that cell x histone mark combination.

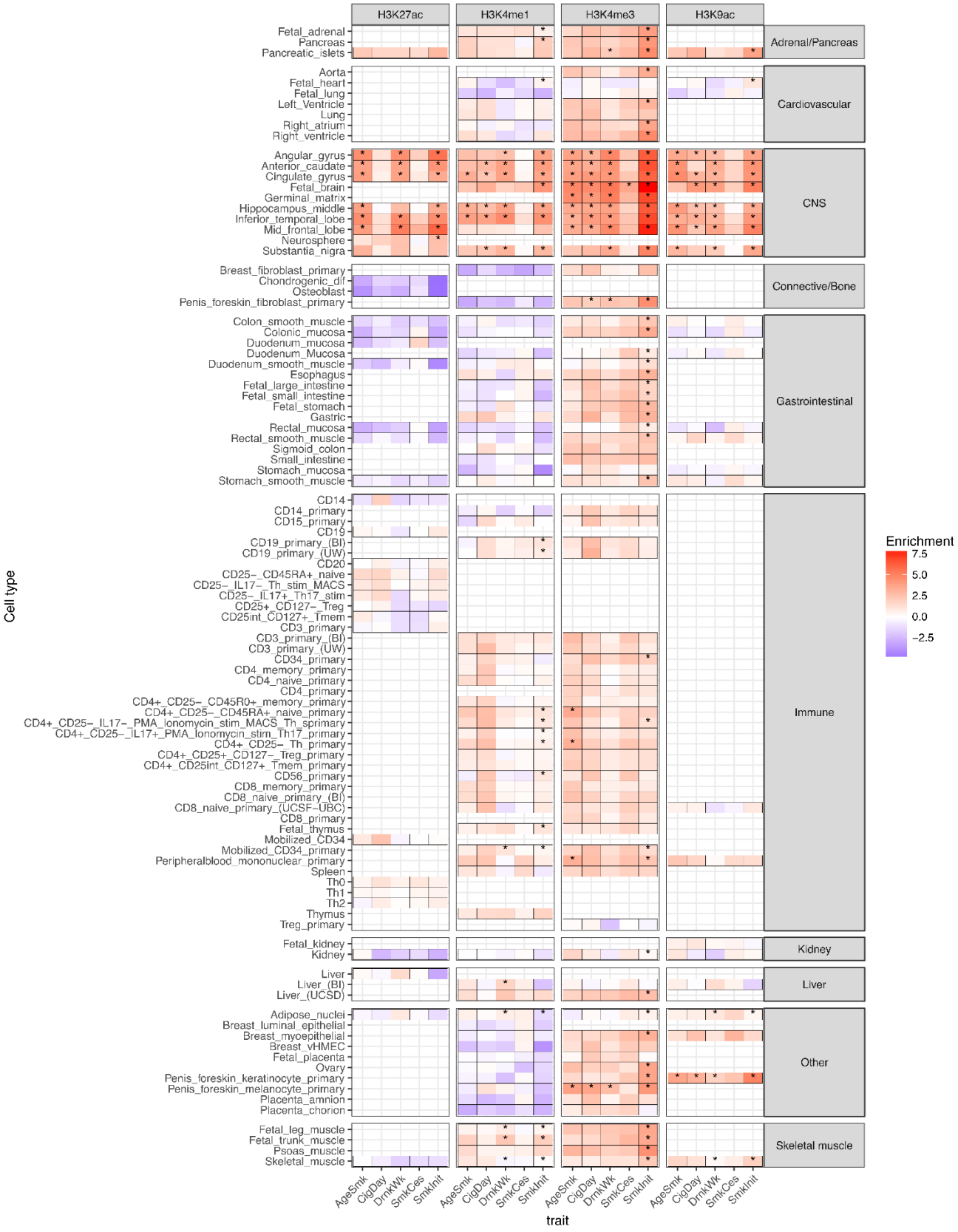


Figure S15. Cell/tissue expression in DEPICT. The figure shows the degree to which genes in significant loci from each phenotype (in columns) are overexpressed, relative to genes in random sets of loci, for each of the tissue/cell types from the Gene Expression Omnibus available in DEPICT. Tissue/cell types are grouped by color-coded rows for Medical Subject Headings (MeSH) first-level terms. The degree of red shading indicates the one-sided P value from DEPICT on a –log10 scale. Significant overexpression is indicated with single asterisks (False Discovery Rate < 0.05) or double asterisks (False Discovery Rate < 0.01).

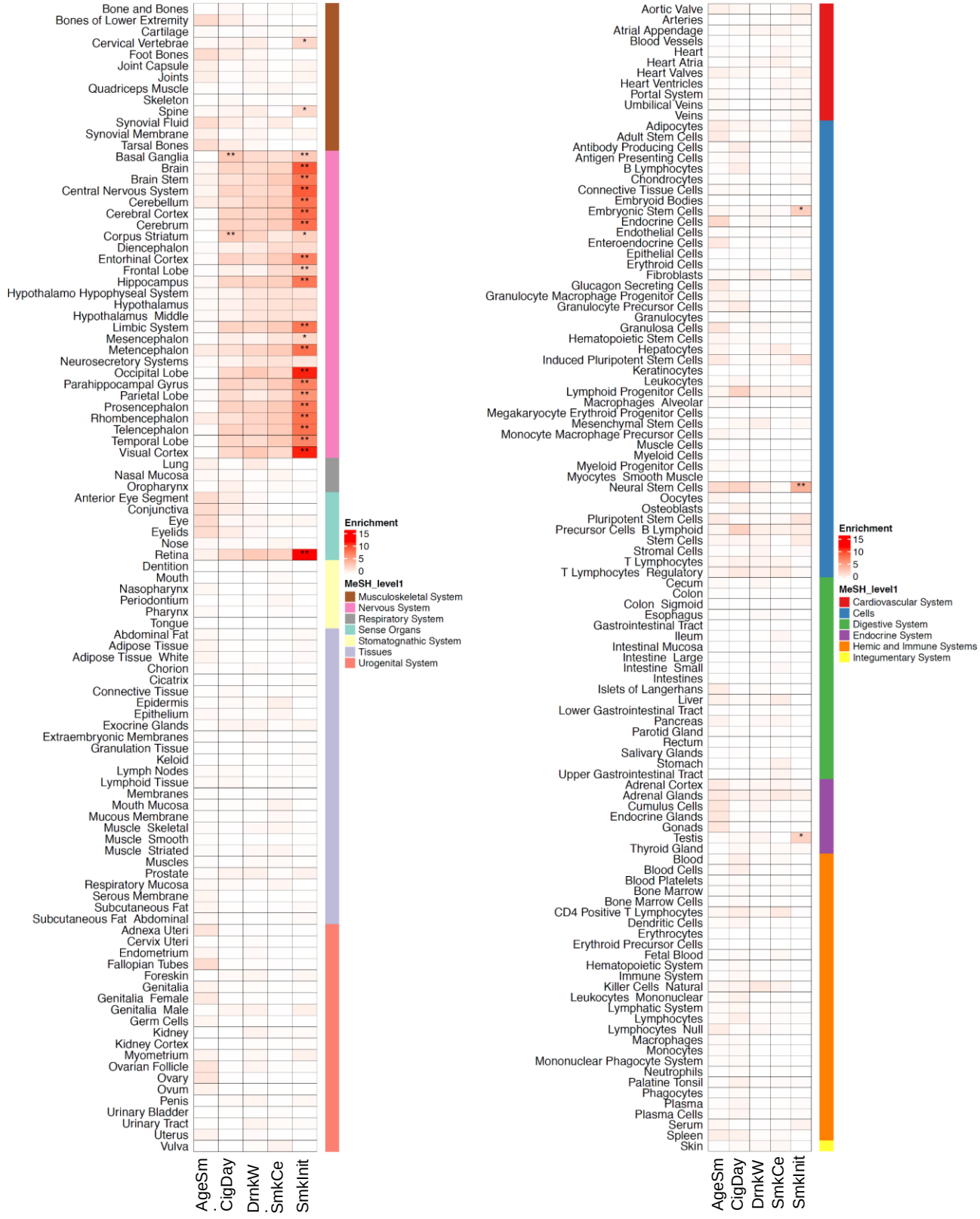


Figure S16. Manhattan plots of MTAG results.

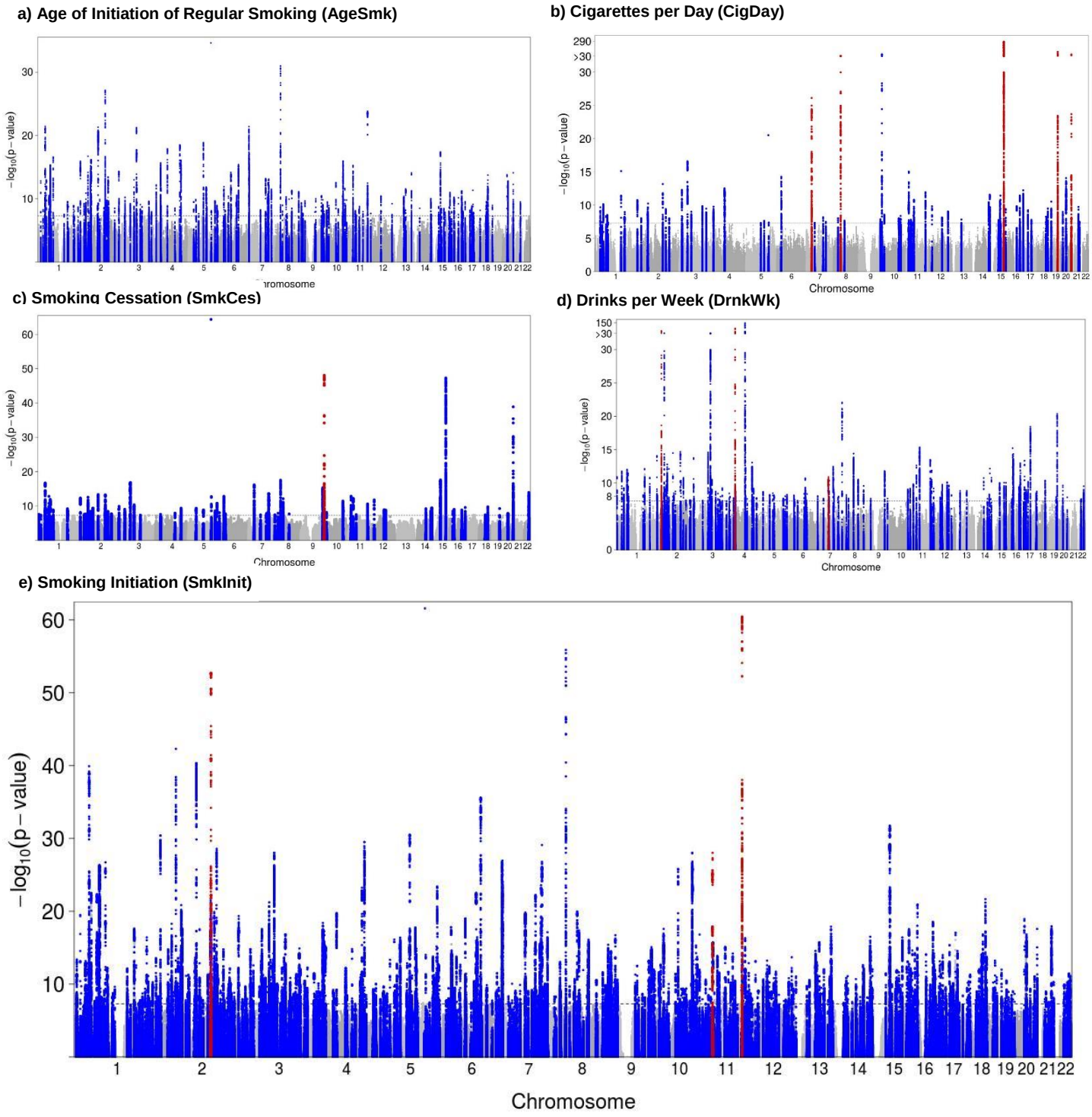
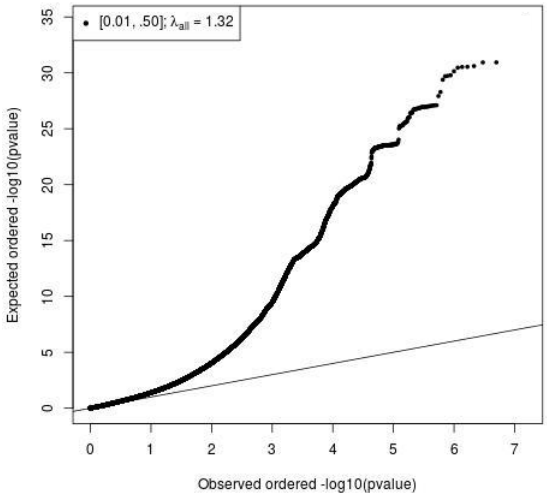
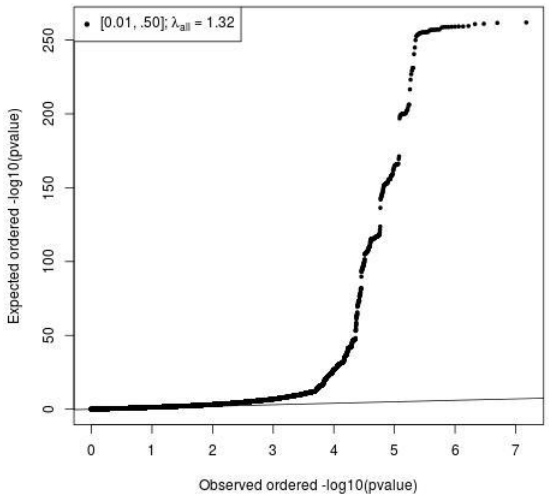


Figure S17. QQ plots of MTAG results

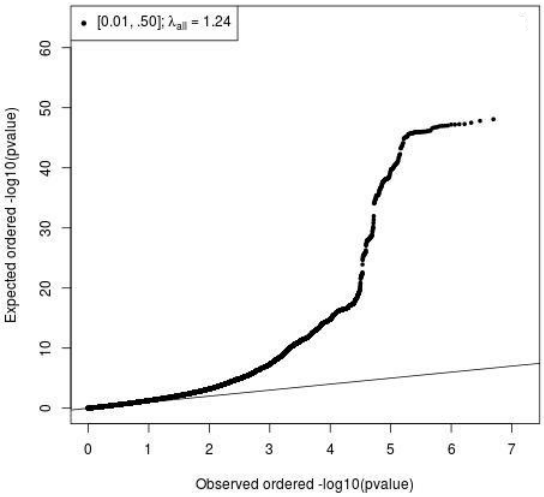
a) Age of Initiation of Smoking (AgeSmk)



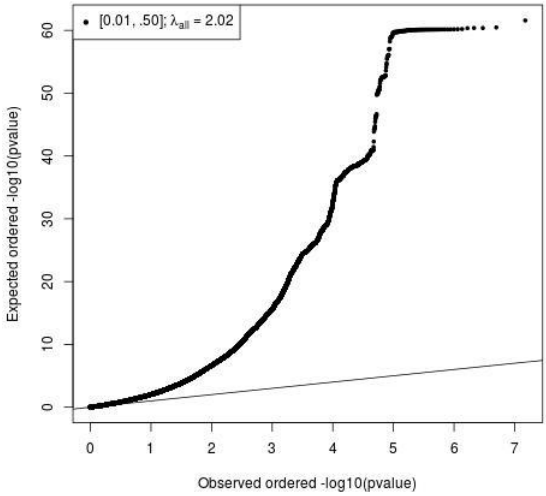
b) Cigarettes per Day (CigDay)



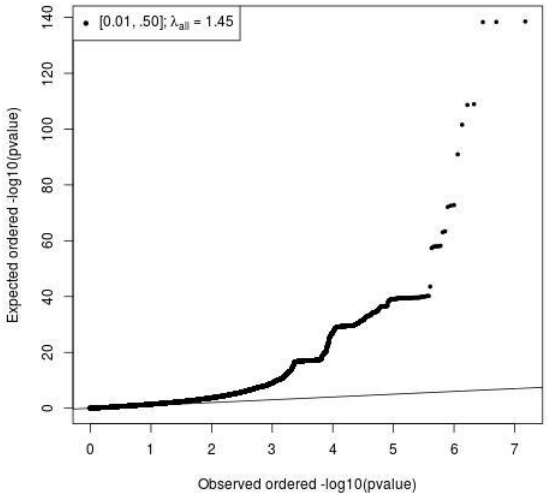
c) Smoking Cessation (SmkCes)



d) Smoking Initiation (SmkInit)

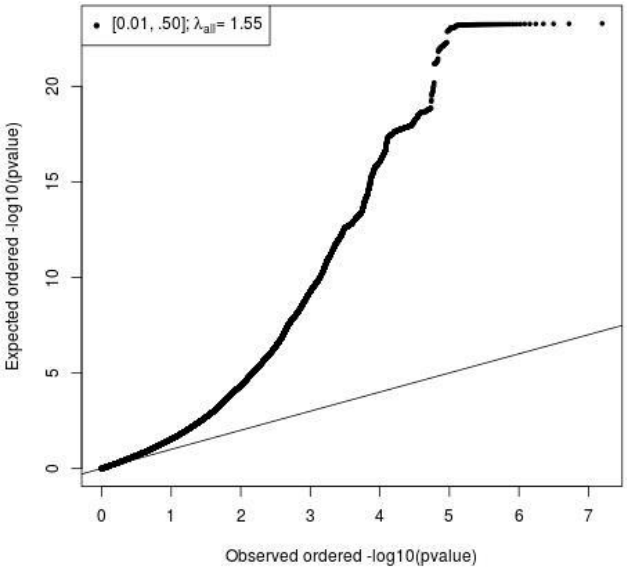


e) Drinks per Week (DrnkWk)



Figures S18-S22. LocusZoom plots of MTAG results

Figure S23 QQplot of Genomic SEM



2. Phenotypes

2.1. Overview

Phenotypes were defined by a workgroup consisting of Drs. Laura J. Bierut, Marilyn C. Cornelis, David A. Hinds, Jaakko Kaprio, Eric Jorgenson, Dajiang J. Liu, Matt McGue, Marcus R. Munafo, Scott Vrieze, and Luisa Zuccolo. Phenotypes were selected based on the availability of a phenotype in a large number of participating studies, previous genetic association research on the phenotype, and clinical relevance. After gathering relevant information and multiple teleconferences, phenotypes were selected and defined.

All studies were asked to use age, age squared, sex, and genetic principal components as covariates in their genetic association analysis. Studies were also asked that cigarettes per day be corrected for current v. former smoker status (if available). Finally, all studies were requested to consider the use of other study-specific covariates in their analyses (e.g., case-control status, site in multi-site studies).

2.2. Phenotype definitions

Age of Initiation of Regular Smoking (AgeSmk)

1. Age at which an individual started smoking cigarettes regularly
2. Does not include information about pipes/cigars/chew, or other non-cigarette forms of tobacco use.
3. Measured in a variety of ways:
    a. *At what age did you begin smoking regularly?*
    b. *How long have you smoked?* combined with *What is your current age?*

Cigarettes per Day (CigDay)

1. Defined as the average number of cigarettes smoked per day, either as a current smoker or former smoker, and whether self-rolled or manufactured are smoked (most studies did not distinguish). Individuals who either never smoked, or for whom there is no available data (e.g., someone was a former smoker, but for whom former smoking was never assessed) were set to missing.
2. For studies that collected a quantitative measure of cigarettes per day, where the respondent is free to provide any integer (e.g., 13 cigarettes per day) responses were binned as follows.
    a. 1 = 1-5
    b. 2 = 6-15
    c. 3 = 16-25
    d. 4 = 26-35
    e. 5 = 36+
3. For studies with pre-defined bins, the pre-defined bins were used.
4. Does not include information about pipes/cigars/chew, or other non-cigarette forms of tobacco use.
5. Cigarettes per day was measured with a single question for most contributing studies using, for example:
    a. *How many cigarettes do you smoke per day?*
    b. *How many cigarettes did you smoke per day?*

Smoking Cessation (SmkCes)

1. Binary phenotype with current smokers coded as “2” and former smokers coded as “1”, and never smokers are coded as missing.
2. Does not include information about pipes/cigars/chew, or other non-cigarette forms of tobacco use.
3. Usually measured through a combination of questions, including:
    a. *Do you currently smoke?* and *Have you ever smoked regularly?*
    b. *Do you smoke?* and *Have you smoked over 100 cigarettes in your entire life?*

Smoking Initiation (SmkInit)

1. This is a binary phenotype. Any participant reporting ever being a regular smoker in their life (current or former) were coded “2”, while any participant who reported never being a regular smoker in their life were coded “1”.
2. Does not include information about pipes/cigar/chew, or other non-cigarette forms of tobacco use.

3. This phenotype was measured in a variety of ways.
    a. *Have you smoked over 100 cigarettes over the course of your life?*
    b. *Have you ever smoked every day for at least a month?*
    c. *Have you ever smoked regularly?*

Drinks per Week (DrnkWk)

1. Defined as the average number of drinks a participant reported drinking each week, aggregated across all types of alcohol. If a study recorded binned response ranges (e.g., 1-4 drinks per week, 5-10 drinks per week) we used the midpoint of the range. For example, if an individual reported 1-5 drinks per week, we assume they drank 2.5 drinks per week on average.
2. This was measured in a variety of ways.
    a. *In the past week, how many alcoholic beverages did you have?*
    b. *Thinking about the past year, on the average how many drinks did you have each week?*
3. This phenotype was left-anchored at 1 and log-transformed prior to analysis, in order to prevent outliers from having undue leverage on analyses.

2.3. Measurement validity and reliability

There are several indications that our measurements, while imperfect, are reliable and valid. We selected our phenotypes in part based on reports from the Tobacco and Genetics (TAG) consortium[1], which was the last large-scale GWAS meta-analysis of tobacco use phenotypes. TAG found that age of smoking initiation, cigarettes per day, smoking cessation, ever/never regular smoker were heritable, had sufficient variation in population samples, and mild differences in questionnaire language had small effects. Indeed, the heritabilities estimated from our results are quite similar to heritabilities estimated from the 2010 TAG results, although with tighter standard errors.

There is additional external evidence that the chosen measures are reliable and valid, based on decades of research into the effects of smoking and alcohol use on morbidity and mortality[2-6]. Using these measures, smoking and alcohol are considered as major contributors to premature death, excess morbidity, and substantial public costs. In smaller studies with biomarkers, the same conclusions on risk effects are reached but with somewhat enhanced risk ratios[7]. In addition, measures like cigarettes per day correlates with aspects of smoking topography and blood metabolites (e.g., our binned measure of CigDay correlates with cotinine levels at ~.30[8]). Self-reported smoking status (of current smoking) is generally quite valid when confirmed by exhaled carbon monoxide and cotinine levels (as biomarkers of smoking)[9]. There are also moderate to high correlations between cigarettes per day and nicotine dependence (r=.8)[10], as well as between drinking frequency and alcohol (r=.5) dependence[10], which themselves have extensive validation literatures.

While evidence for reliability and validity exists—indeed, the present results support the reliability and validity of these measures—our substance use measures likely suffer from greater unreliability than some classical complex phenotypes such as height, years of education, or schizophrenia. Unlike smoking and alcohol use, such phenotypes are either easy to reliably measure (education, height) or are highly salient and chronic (schizophrenia), and therefore quite noticeable. Some of our smoking and drinking measures (SmkCes, CigDay, and DrnkWk) are mutable and likely change from year to year. The genetic etiology of these phenotypes may differ in important ways depending on the nature of the environment, and the age/birth year of the participant[11,12]. Stages of smoking development (from experimentation to weekly to daily smoking) may have specific genetic components as suggested by a relatively small GWAS of family data[13].

Future work with more precise measures (e.g., improved self report and/or biomarker data) in developmentally informative samples will be poised to evaluate such hypotheses. All of these potential measurement issues are expected to increase measurement error and decrease statistical and interpretive power.

2.4. Individual study phenotypic descriptive statistics

Each participating study followed the same analysis plan. 23andMe, deCODE, HUNT, and GERA followed the phenotype definitions but used their own analysis pipelines for analysis, given the unique complexities of data analysis with those large samples. Consistency in phenotype definition is high, given the strong

between-study genetic correlations observed between the largest available studies, including those that used their own analysis pipelines (**Table S9**).

3. Individual study generation of summary statistics

3.1. Genotyping and imputation

All studies were genotyped on genome-wide arrays. We left quality control procedures to the discretion of the participating studies, almost all of which have extensive experience in conducting genetic association analyses with genome-wide array data.

The majority of studies conducted imputation according to instructions in the analysis plan (available on https://genome.psych.umn.edu). The plan requested that each participating site conduct imputation on the University of Michigan imputation server, which applies quality control checks (strand flips, allele flips, missingness, etc.) and, if the data are sufficiently clean, the server phases, imputes, and makes the results available for download[14]. Most studies imputed to the Haplotype Reference Consortium panel[15]. The Haplotype Reference Consortium panel includes >39 million SNPs based on 32,488 low- to moderate-pass whole genome sequences, including the 1000 Genomes phase 3, providing the best imputation accuracy currently available[15].

Several studies did not impute using the imputation server, primarily due to data use restrictions, computational, and/or resource limitations. 23andMe, due to contractual obligations to their research participants, could not use the online imputation server, which, at the time of this writing, is the only way to access the Haplotype Reference Consortium panel. 23andMe imputed using a combination of 1000 Genomes phase 3 and the UK10K. The UK Biobank genotypes, as well as study data obtained from dbGaP could not be transferred for imputation on the imputation server. For the UK Biobank we simply used the imputed genotypes made available by the Biobank (custom reference panel including 1000 Genomes phase 3 and the UK10K for the initial release, and Haplotype Reference Consortium for the second release). For studies obtained through dbGaP, we imputed to 1000 Genomes phase 3.

Regardless of reference panel, all studies used either Minimac3[14] or IMPUTE2[16]. Both programs estimate the squared correlation between an imputed estimated number of alternative alleles and the true number of alternative alleles, known as the INFO score (IMPUTE2) or RSQ (Minimac3).

3.2. Association analysis

Summary statistics were generated in each study sample using RVTEST[17] according to the analysis plan. Studies composed primarily of classically related individuals (e.g., family studies) first regressed out covariates under a linear model, inverse-normalized the residuals, and tested for an additive effect of each variant. Family studies followed this analysis for all phenotypes, even binary phenotypes such as smoking initiation and cessation. Studies of classically unrelated individuals followed the same analysis for quasi-continuous phenotypes (i.e., cigarettes per day, drinks per week, age of initiation), but estimated additive genetic effects under a logistic model for binary phenotypes.

Accounting for population stratification and cryptic relatedness was addressed during the generation of summary statistics by each local study through the use of kinship-based linear mixed models[18] and genetic principal components[19].

3.3. Quality control of per-study summary statistics

For each study, we performed a wide range of checks to ensure data quality of both the phenotypes and genotypes. For each phenotype and covariate, we obtained distribution statistics including the minimum, maximum, quartiles, median, mean, and standard deviation. We ensured that each of these statistics was within expected limits given the phenotype definitions and any scale transformations. We also evaluated simple relationships among phenotypes (e.g., per the phenotype definitions, the smoking cessation phenotype cannot have a larger sample size than the proportion of ever smokers from the smoking initiation phenotype). When discrepancies were noted we contacted the original study for clarification or re-analysis. In some cases, the discrepancy was not noted until a later stage, in which case the association summary

statistics for that phenotype were excluded from analysis. Phenotypic statistics are presented in **Table S6 and Table S7.**

To help ensure genetic data quality we conducted extensive quality control and filtering on the contributed summary statistics from each cohort. We removed imputed variants with imputation quality less than .3, in order to ensure that genetic variants with low imputation quality would not confound meta-analysis results. Moreover, we examined the allele labels from each cohort and compared them with reference panels. We also compared the reported allele frequencies with that of the reference panels, and we removed the variant site from the participating study if the reported allele frequency differs drastically from the reference panel, in order to avoid potential allele flips or strand flips. For quantitative traits, we plotted the variance of the score statistics against the sample size, and made sure that the trait residuals in each study are properly normalized and the trait analyzed between studies were measured and analyzed using the same unit. For each variant in the meta-analysis, we calculated the effective sample size $N_{eff} = \sum_k N_k r_k^2$, where N_k is the sample size in study k and r_k^2 is the imputation quality. We removed variants with effective sample sizes < 10% of the total sample size.

4. Meta-analysis methods

4.1. Meta-analysis

Meta-analysis was performed centrally using the software package rareGWAMA (https://github.com/dajiangliu/rareGWAMA). Given that rare variants may have higher degree of between-study heterogeneity in allele frequencies and imputation qualities, we extended existing methods and developed a novel approach that is aware of these between-study heterogeneities. Specifically, the methods aggregated weighted Z-score statistics, i.e. $Z_{META} = \frac{\sum_k w_k Z_k}{(\sum_k w_k^2)^{1/2}}$, where Z_k is the Z-score statistic in study k. The weight w_k is defined by $w_k = N_k p_k (1-p_k) R_k^2$, where p_k is the variant allele frequency, R_k^2 is the imputation quality, and N_k is the sample size for study k. The weights are proportional to the sample genotype variance. When the trait is uniformly measured and the allele frequency is similar, the method is approximately equivalent to the meta-analysis method that weights score statistics by the sample sizes. Yet, our method is imputation aware and more robust against between-study heterogeneity in the allele frequencies and imputation qualities.

At the meta-analysis step we accounted for any residual population stratification that may have existed even after genetic PCs and linear mixed models were used by local studies. To do this, we applied a stratified genomic control[20] correction to each study's results prior to meta-analysis. The correction was stratified because common variants (MAF ≥ 1% per the pooled allele frequency calculated across all studies) were corrected by genomic controls based on only common variants; whereas low frequency variants, here defined as variants with MAF between 0.1% and 1% based on the pooled allele frequency, were corrected by genomic controls based on only low frequency variants. One cohort, Genetic Epidemiology Research on Adult Health and Aging, only provided results for variants with MAF ≥ 1%, so the genomic control based on common variants was applied to GERA summary statistics even for the rare cases where the pooled minor allele frequency for a variant in GERA happened to be <1%.

First, we conducted a full joint meta-analysis of all available data, using the methods described above. These are presented as our primary findings, and reported in the main text. The statistical significance threshold applied to the meta-analysis of variants with MAF > 1% was 5×10^{-8}, consistent with widely held statistical significance thresholds. We also tested association for variants with 0.1%>MAF>1% to which we applied a statistical significance threshold of $p < 5\times10^{-9}$. The latter threshold applies a correction for ~10 million tests, which is approximately the number of conditionally independent variants tested once the MAF lower bound was extended from 1% to 0.1%. Traditionally, when analyzing common variants (MAF > 1%), there are approximately 1 million independent tests, and p-value thresholds of 5×10^{-8} is conventionally adopted. In sequence-based genetic studies, substantially more low frequency/rare variants are analyzed. Here, after filtering, around 15 million genetic variants with MAF > 0.1% were retained and analyzed. We expect that the number of independent association tests can be larger than a GWAS of common variants. As a result, we calculated the number of independent tests using a combination of existing methods. Specifically, Li and Ji[21],

Gao[22], Chen and Liu[23] have developed methods to calculate the effective number of independent tests for a set of correlated SNPs in GWAS studies. These methods make use of the eigenvalues of the matrix of linkage disequilibrium (measured in R^2) between SNPs, which can be calculated directly based upon spectral decomposition. We estimated the number of independent tests using the genotype data from a subset of the Haplotype Reference Consortium panel[15]. To proceed, we first calculated LD blocks across the genome using the algorithm implemented in PLINK version 1.9 with default settings. We lowered the MAF threshold to 0.01% to accommodate all low frequency variants that we tested in GSCAN. Next, we calculated the effective number of independent tests within each LD block and between LD blocks using the three methods, which we aggregated to get the total number of independent tests.

Typically, the methods used to estimate the number of independent tests uses the correlation matrix of the entire set of SNPs to be analyzed to calculate the significance threshold. The matrix is difficult to calculate efficiently for a large number of variants, and may generate spurious correlation between distant SNPs not actually in LD. We only analyzed the correlation matrix of SNPs in short segments (LD blocks as well as between the LD blocks), making our approach computationally feasible, and biologically sensible. This method is likely more conservative than approaches that use the correlation matrix for an arbitrarily large number of variants at a time, because we do not consider correlations between variants in different LD blocks when estimating the number of independent tests. Yet, based upon our assessment, the correlations between SNPs inside and outside the LD block are very weak. Our estimates should closely approximate the true number of independent tests. Gao's as well as Li and Ji's method gave extremely similar results overall, estimating the effective independent number of variants to be 10.1 million independent tests. The estimate from Chen and Liu was slightly lower at 9.8 million.

4.2. Locus definition and conditional analysis

A locus was defined as a 1MB region surrounding the top *p*-value. In the few instances where such loci overlapped or abutted one another, they were collapsed into a single locus. For each locus thus defined, we conducted conditional analysis using marginal association statistics, as well as approximated covariance matrices between them based upon the linkage disequilibrium information estimated from HRC panel.

For each locus, we performed sequential forward selection, in order to identify independently associated variants. Specifically, we initialized the set of independently associated variants (denoted by Φ), starting with the top association signal in the locus. For each iteration, conditioning on variants in Φ, we performed conditional association analyses for all remaining variants. If the top association signal after the conditional analysis remained significant, we added the top variant to the set Φ, and then repeated the conditional association analysis. If the top variant after the conditional analysis was no longer significant, we stopped and reported variants in the set Φ as the final set of independent variants for that locus. We used the same single variant significance threshold ($P < 5\times10^{-8}$ for variants with MAF≥.01, 5×10^{-9} for low-frequency variants) to determine statistical significance with the sequential forward selection results.

We developed and applied an improved conditional analysis method to identify independently associated variants[24]. Unlike existing conditional meta-analysis methods that make use of final meta-analysis results[25], we made use of cohort level summary statistics. Therefore, our method is able to better estimate correlations between score statistics than existing methods when contributed summary association statistics contain missing values. For example, consider a simple scenario of meta-analysis of two studies. The first variant is only measured in study 1 and the second variant is only measured study 2. The meta-analysis statistics from the two variants will be independent. Existing methods that approximate the correlation between the two variants with linkage disequilibrium coefficients will incorrectly estimate the correlation between test statistics, lead to inflated type I error and reduce power.

Conditional analysis using summary statistics requires external estimates of non-independence among effects of distinct variants. Non-independence between genetic variants was estimated based on linkage disequilibrium patterns in a subset of the Haplotype Reference Consortium (HRC) dataset[15]. This is a large subset (N=21,500) of the same dataset used for the vast majority of imputation into individual study samples. The full HRC comprises roughly 32,500 individual whole genome sequences from multiple whole-genome sequencing studies, with phased genotype calls available at all sites with a minor allele count of at least 5.

The HRC contains worldwide populations, but the majority are of European (EUR) origin. We used a subset of the full HRC, including: AMD (European ancestry and worldwide; N=3,189), BIPOLAR (European ancestry; N=2,487), GECCO (European ancestry; N=1,112), GOT2D (Europe, N=2,709), HUNT (Norway; N=1,023), SardiNIA (Sardinia; N=3,445), MCTFR (Minnesota, US; N=1,325), 1000 Genomes (worldwide; N=2,495), UK10K (UK; N=3,715). The subset of the HRC data we accessed totaled 21,500 whole genome sequences comprising 38,913,048 biallelic SNPs.

Our conditional analysis approach is expected to have well controlled type I errors and improved power over existing methods. Despite this, we sought to compare our conditional analysis results to those generated with a more widely used tool as a reference. To do this, we used Genome-wide Complex Trait Analysis software (GCTA) to conduct a conditional and joint multiple-SNP analysis (COJO)[25]. Using the same summary statistics and LD reference panel, we performed a stepwise selection process as suggested by the authors[25]. We chose to use the locus definition (1MB, adjacent such windows are collapsed) as in our primary conditional analysis approach outlined above. A basic comparison between our approach and COJO, as applied to these loci, is shown in **Table S27**, where it indeed appears that our approach has improved power. For all loci, rareGWAMA identified as many or more conditionally independent variants than COJO.

4.3. Identifying loci with pleiotropic effects

We adapted the polygenic model to estimate the extent of pleiotropy. The method can be viewed as an extension of the variance component model to incorporate the uncertainty of whether a gene region is associated with the trait or not.

Model Specification

We consider region-based association analysis multiple traits. The genotypes for locus *j* is denoted by $\mathbf{G_j}$. The genetic effects for locus j are denoted by $\boldsymbol{\beta_j} = (\boldsymbol{\beta_{j1}}, \boldsymbol{\beta_2}, \dots, \boldsymbol{\beta_{jK}})$, where $\boldsymbol{\beta_{jk}}$ is the effects of region j for trait k.

$$\mathbf{Y} = \mathbf{G_j}\boldsymbol{\beta_j} + \mathbf{e_j} \quad (1)$$

Different regions are assumed to be independent. We use an indicator γ_{jk} to denote the states of association, with $\gamma_{jk} = 1$ if the gene j is associated with trait k. The genetic effect satisfies:

$$\beta_{jk}|\gamma_{jk} \sim \begin{cases} N(0, \tau_k^2) & \text{if } \gamma_{jk} = 1 \\ 0 & \text{if } \gamma_{jk} = 0 \end{cases} \quad (2)$$

The genetic effects for region j between different traits are assumed to be correlated, and satisfies

$$cov(\beta_{jk_1}, \beta_{jk_2}|\gamma_{jk_1}, \gamma_{jk_2}) = r_{k_1k_2}\tau_{k_1}\tau_{k_2}\gamma_{jk_1}\gamma_{jk_2} \quad (3)$$

$r_{k_1k_2}$ is the correlation between genetic effects if the locus is causal for both traits k_1 and k_2. The genetic effects for different loci are assumed to be independent.

Finally, the indicator γ_{jk} is assumed to be independently and identically distributed, and follows a Bernoulli distribution, i.e. $\gamma_{jk} \sim Bernoulli(p_k)$.

Together, the full model is specified as follows:

$$\mathbf{Y} = \mathbf{G_j}\boldsymbol{\beta_j} + \mathbf{e_j} \quad (4)$$

$$\boldsymbol{e_j} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\rho}) \quad (5)$$

$$\boldsymbol{\beta_j}|\boldsymbol{\gamma} \sim MVN(\mathbf{0}, \mathbf{V_\gamma}), \text{ where } \mathbf{V_\gamma}(k_1, k_2) = cov(\beta_{jk_1}, \beta_{jk_2}|\boldsymbol{\gamma}) = r_{k_1k_2}\tau_{k_1}\tau_{k_2}\gamma_{jk_1}\gamma_{jk_2} \quad (6)$$

$$\gamma_{jk} \sim Bernoulli(p_k) \quad (7)$$

The above model can be fitted with only summary level data, i.e. the z-scores. Without loss of generality, we assume that the genotypes and phenotypes are standardized to have mean value of 0 and variance of 1. So the Z-score statistic is equal to $Z_{jk} = \frac{1}{\sqrt{N}}\mathbf{G_j'Y_k}$.

We noted that the parameters $r_{k_1k_2}\tau_{k_1},\tau_{k_2},\gamma_{jk}$ can be estimated using method of moment from the joint distribution for the score statistics based. The following estimating equation for the variance and covariances will be used:

$$\boldsymbol{var}\left(\mathbf{G_j'Y_k}\right)|\gamma_{\mathrm{jk}} = \mathbf{G_j'G_jG_j'G_j}\boldsymbol{\tau_k^2}\gamma_{jk} + \mathbf{G_j'G_j} \quad (8)$$

$$cov\left(\mathbf{G_j'Y_{k_1}}, \mathbf{G_j'Y_{k_2}}\right)|\gamma_{\mathrm{jk}_1},\gamma_{\mathrm{jk}_2} = \mathbf{G_j'G_jG_j'G_j}r_{k_1k_2}\tau_{k_1}\tau_{k_2}\gamma_{jk} + \mathbf{G_j'G_j}\rho_{k_1k_2} \quad (9)$$

Model Estimation

The variance of trait residuals was estimated using the z-score statistics from variant sites that do not show evidence of association (p-value >.5).

We fitted model using an expectation moment of moment algorithm (EMOM) The algorithm iteratively alternate between the E and MOM step until the estimated parameters converge.

1) E-step: We estimate the hyper-parameters p_{jk} using Bayes formula for the marginal distribution for the Z-score statistic, i.e.

$$\hat{p}_{jk} = \frac{\prod_m \phi\left([\mathbf{G_j'Y_k}](m);0,[\mathbf{G_j'G_jG_j'G_j}\boldsymbol{\tau_k^2}+\mathbf{G_j'G_j}](m,m)\right)}{\prod_m \phi\left(\left[\mathbf{G_j'Y_k}\right](m);0,\mathbf{G_j'G_jG_j'G_j}\boldsymbol{\tau_k^2}+\mathbf{G_j'G_j}(m,m)\right)+\prod_m \phi\left(\mathbf{G_j'Y_k}(m);0,\left[\mathbf{G_j'G_j}\right](m,m)\right)} \quad (10)$$

Where $[\cdot](m)$ denotes the mth element of a vector and $[\cdot](m,m)$ denotes the (m,m)th element of a matrix. The assignment of the association status is given by $\gamma_{jk} = I\left(\hat{p}_{jk} > .5\right)$.

2) MoM-step:

To estimate τ_k, we use trait with $\gamma_{jk} = 1$, and apply method-of-moment estimation based upon (8). To estimate parameters $r_{k_1k_2}$, we use trait pairs with $\gamma_{jk_1} = \gamma_{jk_2} = 1$, and apply method-of-moment estimation based upon (9).

The final association states are determined by $I\left(\hat{p}_{jk} > .5\right)$, where $\hat{p}_{jk}$ is the estimated probability of association when the algorithm converges.

This method was applied to each of the 405 loci discovered in the GWAS (it is inapplicable to the MTAG results, as it depends on the genetic correlation between traits). The posterior probability of locus association was estimated for all possible combinations of the five phenotypes, and the combination with the highest posterior was selected and reported in **Table S12** along with information about which genes within that locus had been implicated. The patterns of association across all loci are plotted in **Figure 2**.

5. Meta-analysis results

5.1. Filters applied after meta-analysis

We applied multiple post-meta-analysis variant filters to ensure robustness of reported findings. To reduce artifacts arising from a small number of studies, we excluded any variant that was present in only two or fewer studies. We excluded all variants with a minor allele frequency less than 0.001, the lower bound of what we expect will be imputable with passable accuracy with the currently best available imputation reference panel (the HRC)[15]. Association power is a function of multiple things, including minor allele frequency, imputation quality, and sample size. There will be some variants that are extremely poorly imputed. To account for imputation quality, we considered the product of imputation quality and sample size, summed across studies. This we termed the “effective” sample size, which is directly proportional to statistical power. We removed any variant with an effective sample size of 10% or less of the total sample size. Results from the application of these filters are displayed in **Table S24**.

After applying these three variant filters, we calculated genomic controls and maximum/median per-variant sample sizes. Sample sizes ranged from 337,334 for cigarettes per day to 1,232,091 for smoking initiation. Genomic controls values indicate that Type I error rates are well controlled, for both common and low-frequency variants.

5.2. Assessing population stratification using the LD intercept test

Table S25 shows the results of the LD Score Regression intercept test, which provides an estimate of the extent to which association statistic *p*-values in GWAS may be affected by population stratification[26]. Here we assess the extent to which population stratification is driving the inflation we observe in our summary statistics. In the first set of intercept test results where we have applied no genomic control to our results (Panel A), in row 4, we observed intercept values from 1.03 for AgeSmk to 1.14 for SmkInit. The ratio (row 6) indicates the proportion of the inflation in the mean χ^2 (row 5) that the LD score regression intercept ascribes to causes other than polygenicity, such as population stratification. The ratios range from 0.06 (*SE* = 0.01) for DrnkWk to 0.13 (*SE* = 0.02) for SmkCes. Based on these results, we find evidence for population stratification, and thus we applied a cohort-level genomic control to the summary statistics from each of the participating studies before conducting meta-analysis. The results for the intercept test using these summary statistics are shown in Panel B. This time, the ratio is below 0 for each of our phenotypes, indicating that a cohort-level genomic control is likely sufficient to account for any population stratification biases in our meta-analytic results.

5.3. Quality control and manual review of all significant loci

We applied our locus definition (1MB windows and collapsing adjacent windows) and evaluated a meta-analytic Cochrane's *Q* and I^2 statistic to evaluate effect heterogeneity across contributing studies for the most significant variant within each locus. Q was considered significant after application of a Bonferroni correction for all tests within a phenotype. No variants were significantly heterogeneous for AgeSmk, SmkCes, or DrnkWk. One variant was heterogeneous for CigDay (rs28813180, chr 3, pos 158,191,316; *Q* $p=5.6\times10^{-6}$, I^2=15.4%) and two variants were significantly heterogeneous for SmkInit (rs2279829, chr 3, pos 147,716,635, *Q* $p=2.7\times10^{-5}$, I^2=70.3%; and rs2526390, chr 3, pos 50,181,136, *Q* $p=1.4\times10^{-4}$, I^2=80.3%). These variants are reported and remained in downstream analyses, but caution is advised in the interpretation of their effect. *Q* *p*-values, df, and I^2 are reported for all top variants in genome-wide significant loci in **Tables S3-S7**.

We conducted a detailed review of all conditionally independent variants. All genome-wide significant loci were plotted with LocusZoom[27] and included in **Figures S4-S8** for AgeSmk, CigDay, SmkCes, SmkInit, and DrnkWk, respectively. All LocusZoom plots were manually reviewed and suspicious loci evaluated in detail. Seven total loci were identified as highly suspicious and likely spurious given the pattern of association between the lead associated variant and neighboring variants in high linkage disequilibrium. Three loci associated with SmkInit were identified as suspicious and removed from further evaluation (lead variants: rs1247037, rs1318004, rs2622165); four loci associated with SmkInit were removed (lead variants: rs828867, rs1169243, rs1140501, rs1318004); one locus associated with DrnkWk was removed (lead variant: rs8178967). Despite removing these variants from further consideration, we retain their LocusZoom plots in **Figures S4-S8** and tabulate these loci in **Table S28**, for the interested reader.

LocusZoom images were made using the LocusZoom Standalone software (https://genome.sph.umich.edu/wiki/LocusZoom_Standalone) using a subset of the HRC as the reference panel (described elsewhere) for LD information and the UCSC genome browser for dbSNP and gene positions. We set the most significant variant in the region as the reference variant upon which LD in the window is based. Some reference variants were not present in the HRC, but were present in studies that imputed using UK10K + 1000 Genomes phase 3, or 1000 Genomes phase 3 alone. For these variants we rendered the LocusZoom plots using the most significant HRC variant in the region. In some instances these were in high LD with the sentinel variant, in other cases they were not. These variants are listed in **Table S29**, along with the LD between them.

For all loci that harbored 2 or more conditionally independent variants, we plotted them using the conditional result utility in LocusZoom to visually inspect the pattern of conditional independence. These plots are included in **Figures S9-S12** for CigDay, SmkCes, SmkInit, and DrnkWk, respectively. (AgeSmk had no loci with more than a single conditionally independent variant.)

The remaining full set of genome-wide significant loci, and all conditionally independent variants within those loci, can be found in **Tables S1-S5**.

To evaluate whether a locus was novel or known, we conducted a literature search of previous GWAS and highly powered candidate gene studies. The list of known loci compiled through this process is provided in **Table S22**.

5.4. LocusZoom website

A website to share interactive LocusZoom plots was created to facilitate exploration of the meta-analysis results. The site is located at http://gscan.psych.umn.edu/. The interactive site displays considerably more information than static LocusZoom plots, including allele frequencies, p-values, LD, etc., for the top variant and all other plotted variants. The website uses the LocusZoom.js application, a javascript/d3 embeddable plugin for interactively visualizing statistical genetic data[27] (https://github.com/statgen/locuszoom), to create the plots. A wrapper script was created to take as input a list of phenotypes, a list of significant variants per phenotype, and a list of variants in loci of interest for the phenotype. The input is converted into a comma-separated (csv) file, and a customized html file is created for each phenotype for the LocusZoom.js app. A web server is also included in the wrapper. It uses Python's flask and pandas packages, parses json web requests, gets results from the csv file, and formats the results into a json response for LocusZoom. The wrapper is available at https://github.com/dattagargi/LZ_Wrapper.

The LocusZoom.js app contains a web page with six tabs. The first tab depicts significant loci for each phenotype. All significant variants across phenotypes are listed on the left. Clicking a significant variant generates a plot depicting the significant locus for that variant (500kb upstream and downstream), with each phenotype differentiated by its colors. When hovering over a variant in the plot, information about the variant's location, p-value, minor allele frequency, and which phenotype it is associated with are shown. Genes in the locus are depicted under the plot along with a summary of the gene, including expected and observed variants, and the predicted variant function.

Each of the five phenotypes are captured in five additional tabs in the LocusZoom website. Each phenotype tab has a list of clickable entries for a phenotype's independently significant variants on the left. When clicking on a significant variant, a plot with all variants for the phenotype in that locus (500kb upstream, 500kb downstream) is displayed. The variants in this plot are stratified and colored by their pairwise LD with the top significant variant in that region. As with the first tab, genes in the locus and their summaries are depicted.

Given restrictions from broad sharing of our full meta-analysis results (per contract with 23andMe), we have included on the LocusZoom website the meta-analytic results excluding 23andMe. LocusZoom plots of all significant signals from the full meta-analysis, including 23andMe, are included as static images in the supplementary materials.

6. MTAG

MTAG[28] is a method to increase power for GWAS through analysis of multiple correlated phenotypes. MTAG does not deliver a multivariate test statistic in the usual sense, and does not provide direct evidence of whether a given variant is associated with multiple phenotypes. Rather, MTAG delivers separate results for each phenotype. In our case, MTAG uses the genetic correlations among our substance use phenotypes to increase power to deliver a separate set of summary statistics for each of our five phenotypes. As MTAG relies on LDSC to estimate the genetic correlation, it by default does not test variants with MAF < 1%, and these were therefore excluded from the MTAG analysis. We applied MTAG to each of our five substance use phenotypes in turn, using the remaining four phenotypes to increase power. MTAG was run with default settings, except that we set the minimum N to zero and used the effective sample size (observed N*imputation RSQ) as the per-marker sample size and took the square root of the chi-squared statistics as the Z score (using beta to assign the direction). After coordinating summary statistics across phenotypes and filtering variants with MAF < 1%, 9,732,723 SNPs remained.

The MTAG-estimated effective sample size for AgeSmk (original GWAS N = 341,425) was 931,815. This increase is large due to the high genetic correlation between AgeSmk and SmkInit, the latter phenotype having the largest observed sample size in the original GWAS. Effective sample size for CigSmk increased from 337,333 to 403,928; DrnkWk increased from 941,279 to 1,039,210; SmkCes increased from 547,219 to 820,192; and SmkInit increased from 1,232,093 to 1,359,002. Genomic controls increased from 1.15 to 1.32 (AgeSmk), 1.18 to 1.32 (CigDay), 1.14 to 1.24 (SmkCes), 1.45 to 2.02 (SmkInit), and 1.26 to 1.45 (DrnkWk). Despite these increases in genomic controls, the LD intercept test for population stratification revealed no appreciable effect of stratification for any phenotype other than SmkInit (**Table S26**).

We used Genome-wide Complex Trait Analysis software (GCTA) to conduct conditional analysis per chromosome (the 'joint' option in GCTA-COJO)[25]. Using the same subset of the Haplotype Reference Consortium described elsewhere, we identified a total of 1,193 independently associated significant variants across the five phenotypes: 173 for AgeSmk (1730% increase over the original GWAS results), 89 for CigDay (162% increase), 83 for SmkCes (289% increase), 692 for SmkInit (184% increase) and 156 for DrnkWk (156% increase). The results are detailed in **Table S12**, with QQ plots, Manhattan plots, and LocusZoom plots in **Figures S18-S22**. LocusZoom plots were manually reviewed to identify loci with problematic LD support. Loci without apparent support were flagged, removed from **Table S12**, listed in **Table S30**, but retained in the LocusZoom supplementary figures for interested readers.

One strong assumption on which MTAG relies is that the estimated variance-covariance matrix (homogenous **Ω**, using notation from the original paper[28]) of a given variant's effects is homogeneous. It is recommended to evaluate this assumption through calculation of a maximum false discovery rate (FDR) for each phenotype. Higher max FDRs suggest violation of the assumption, although explicit decision thresholds are not described or evaluated in the original publication describing the method. Max FDRs were estimated as 0.06 for AgeSmk, 0.007 for CigDay, .07 for SmkCes, .0002 for SmkInit, and .0003 for DrnkWk, suggesting that MTAG assumptions were violated to some extent for AgeSmk and SmkCes. This is consistent with the observation that apparent power increase from MTAG was quite remarkable for these two phenotypes, with the number of associated variants increasing by 1730% for AgeSmk and 289% for SmkCes. MTAG results should therefore be interpreted with caution, perhaps especially for these two phenotypes.

7. Genomic SEM and the correlational structure of substance use

Measures of psychopathology and related behaviors are correlated in systematic ways, such that analyses of the covariance structure of psychiatric disorders delivering three groups of disorder, often termed psychosis, externalizing, and internalizing[29,30]. Paradigmatic examples of externalizing psychopathology include antisocial behavior, substance dependence, and measures of disinhibitory behavior, such as ADHD and personality measures of behavioral constraint[31,32]. Such results are based largely on factor analysis of observed covariance matrices, including genetically informed covariance matrices (e.g., in twins[33]). Recent advances in estimating genetic correlations based on GWAS summary statistics, such as LD Score Regression[34], have facilitated factor analysis of the resulting genetic correlation matrices. A new method, Genomic SEM[35], takes advantage of the LDSC technique to estimate genetic correlation and sampling matrices, and applies confirmatory factor analysis to them.

We fit confirmatory factor models to the genetic correlation matrix of our five substance use phenotypes. Since our phenotypes are all related to substance use, we expected the analysis results may support an externalizing-type model, where correlations among our five phenotypes could be modeled as resulting from the influence of a single factor. We compared model fit using standard criteria including absolute (e.g., chi square tests, CFI[36], and SRMR[36]) and relative fit indices (likelihood ratio tests, Akaike Information Criterion[37]). Fit of a single factor model to our five phenotypes did not satisfy standard cutoffs for absolute measures of fit (χ^2=259.7, df=5, p=4.5×10^{-54}; AIC=279.7, SRMR=.081; CFI=.895). Common cutoffs are CFI≥.90 and SRMR≤.08. There are limited modifications one can make to a confirmatory model of five phenotypes, but inspection of the genetic correlation matrix suggested a strong residual correlation between age of smoking initiation and ever/never smoker. When the model was re-fit allowing this correlation to be freely estimated, the fit was statistically significantly improved but global indices improved only slightly (χ^2=252.6, df=4, p=1.8×10^{-53}; AIC=274.6; SRMR=.079; CFI=.897; likelihood ratio test relative to initial model, p=.008), indicating that fit was borderline acceptable by convention, although barely. Standardized loadings (and standard errors) of the phenotypes onto

the common factor were -.69 (.06), .41 (.03), .44 (.04), .85 (.05), and .29 (.03) for AgeSmk, CigDay, SmkCes, SmkInit, and DrnkWk. The estimated residual correlation between AgeSmk and SmkInit was -.10 (.07). Despite marginal fit, we conducted a GWAS on this model, regressing the common factor onto each genetic variant included in the univariate GWASs. Like MTAG, because Genomic SEM relies on LDSC to define the genetic correlation matrix on which the confirmatory factor model was fit, we restricted the GWAS only to common variants (MAF>1%).

In all, 7,863,978 variants were included in the GWAS. **Figure S23** shows the QQ plot from Genomic SEM. The genomic control for all variants with MAF > 1% was 1.55. Genomic SEM also provides a Q-statistic test of association heterogeneity, which contrasts the effect of a variant on a common factor, versus the direct effect of the variant on the individual phenotypes over and above variant effects mediated by the common factor. Applying a p-value threshold of 5×10^{-8} to this Q statistic p-value, 30% of these 7,863,978 variants were significantly heterogeneous. We applied GCTA COJO to identify all variants conditionally genome-wide significantly associated with the common factor. This process identified 213 variants, 155 (73%) of which also had heterogeneity Q statistics with $p<5\times10^{-8}$. Application of a Bonferroni correction for 213 tests (p<.0002) resulted in 84% of the variants having statistically significant Q values. These results indicate that three-fourths to four-fifths of associated SNPs showed highly heterogeneous effects across our five phenotypes. All conditionally independent genome-wide significant variants identified with Genomic SEM are listed in **Table S31** for the interested reader, along with their Q statistic.

Genomic SEM appears to be effective in recovering the common factors underlying genetically related phenotypes[35] with simple correlation structures. However, for our substance use phenotypes, the Genomic SEM results do not appear to strongly support the hypothesis that a general externalizing vulnerability is a suitable representation of the shared genetic etiology among our five substance use phenotypes. This conclusion is perhaps expected given that we selected our substance use phenotypes not to represent an externalizing factor, but rather to reflect various stages of substance use (initiation, heaviness, and cessation) and multiple substances (nicotine and alcohol). The Genomic SEM results are also consistent with our multivariate association test (described in **section 4.3**), where extensive pleiotropy was observed among our phenotypes, but only three loci were associated simultaneously with all five phenotypes. In particular, there are especially strong associations (for complex behavioral phenotypes, at least) that appear to be largely phenotype-specific. Associations in the alcohol-metabolizing *ADH1B* locus appear to be specific to alcohol use. The lowest p-value between the Genomic SEM common factor and any variant within one megabase of ADH1B was .001. The most significant association between rs56113850 in *CYP2A6*, a gene involved in nicotine metabolism and known to be associated with cigarettes per day, and the Genomic SEM common factor was .004 (the univariate p-value for CigDay was 4×10^{-99}, with SmkCes it was 1.6×10^{-48}).

8. Heritability and genetic correlations with related diseases and traits

8.1. Overview

We used univariate and bivariate LD Score Regression[26] to assess the heritability of each phenotype and to estimate a variety of genetic correlations. Analyses included (1) LD Score Regression intercept tests to evaluate the extent to which population stratification or cryptic relatedness may artificially inflate our summary statistics; (2) estimation of genetic correlations across our five phenotypes; (3) estimation of genetic correlations computed within a phenotype, between the larger contributing studies, as an estimate of the extent to which phenotypes were measuring the same genetic risk in different studies; and (4) estimation of genetic correlation between the five phenotypes and a wide variety of other phenotypes related to smoking and alcohol behaviors, and for which GWAS have already been made publicly available.

8.2. Heritability

Table S13 provides several heritability estimates for each of the five phenotypes. In row 3, we find that the heritability is 0.047 for AgeSmk (*SE* = 0.003), 0.080 for CigDay (*SE* = 0.008), 0.078 for SmkInit (*SE* = 0.002), 0.046 for SmkCes (*SE* = 0.002), and 0.042 for DrnkWk (*SE* = 0.002). These estimates are similar to previous narrow-sense heritability estimates reported from genome-side studies using LD Score Regression[3]. In rows 4-8 of **Table S13** our estimates of heritability for the overall meta-analysis results are compared to similar

estimates of heritability for individual studies with large sample sizes, including 23andMe, HUNT, Kaiser (DrnkWk only), UKB Axiom and UKB 350. “UKB Axiom” includes individuals from the first release of the UK Biobank, but excludes the UK BiLEVE cohort. “UKB 350” includes the 350,000 additional individuals contained in the second release by the UK Biobank. For biologically related individuals across these cohorts, one member of each pair of individuals related at the level of second cousin or closer were removed. Though there is some variation in the heritability across studies—perhaps related to differences in smoking prevalence and alcohol use across studies and countries/cultures—the estimates are comparable to those for the overall meta-analytic summary statistics. Finally, in rows 9-10, we used GCTA, a GREML-based approach[4], to compare estimates of heritability calculated using summary statistics (LD Score Regression) to estimates calculated using individual-level data (GCTA). GCTA relies on a different set of assumptions than LD Score Regression, and provides a complementary method to calculate heritability in large study samples in which individual level data was accessible. For both UKB Axiom and UKB 350, we find similar estimates between the GCTA estimates and the LD Score Regression estimates, suggesting that there are no gross biases in our heritability estimates.

8.3. Genetic correlation

Under standard assumptions, bivariate LD Score Regression produces unbiased estimates of genetic correlation, even in the presence of sample overlap[34]. Accordingly, to estimate the extent of genetic correlation between each of our phenotypes, and between our phenotypes and other phenotypes related to nicotine and alcohol use, we used LD Score Regression and followed standard procedures[38]. We used data from European-ancestry samples in the 1000 Genomes Project, restricted to only HapMap3 variants with minor allele frequency > 0.01. Standard errors of LD Score Regression estimates were calculated using a block jackknife over all variants.

Table S8 provides genetic correlations for the full meta-analyzed summary statistics between each of our five phenotypes. All correlations were statistically significant at the $p<.05$ level, and all genetic correlations are in the expected direction of effect. For example, age of initiation was negatively correlated with the other four phenotypes, providing evidence that there is overlap between variants associated with an earlier onset of smoking and those associated with ever engaging in regular smoking behavior (SmkInit), smoking persistence (SmkCes), and amount of substance use (CigDay and DrnkWk). Among the remaining correlations, we observed the strongest genetic correlation between Age of Initiation of regular smoking and Smoking Initiation ($\hat{r}_g$= -0.713, *SE* = 0.026). Our remaining smoking and drinking phenotypes were all positively correlated, again as expected. We observed moderate correlations among nearly all of our smoking phenotypes, which on average were genetically correlated at ~0.4. We observed the weakest correlations with DrnkWk, which was significantly correlated in absolute magnitude at about 0.1 with our four smoking phenotypes. In general, the analysis indicates substantial overlap between the variants that influence smoking phenotypes. Further, while there is evidence of overlap between the variants that influence smoking and drinking phenotypes, these results also demonstrate greater differences between smoking and drinking, than between different smoking behaviors.

Smoking and drinking have been implicated as risk phenotypes in a number of other diseases and disorders[39-43]. Accordingly, we computed genetic correlations between our five substance use phenotypes and other phenotypes related to nicotine and alcohol use, and for which large GWAS have been performed and summary statistics were publicly available. **Table S10** presents the results for these correlations across a large range of phenotypes, including height; a variety of sociodemographic and developmental phenotypes that we expect to be correlated with nicotine and alcohol use behavior (age of menarche, age of first birth, years of education); risk behavior and other substance use phenotypes (cotinine, general risk tolerance, lifetime cannabis use); psychiatric phenotypes (ADHD, autism spectrum disorder, bipolar disorder, major depressive disorder, neuroticism, schizophrenia); non-psychiatric neurological diseases (Alzheimer’s Disease, Multiple Sclerosis, and Parkinson’s Disease) and medical traits and disease for which smoking and alcohol use are prominent risk factors (BMI; Class I obesity, femoral neck bone mineral density, lumbar spine bone mineral density, HDL cholesterol, LDL cholesterol, total cholesterol, chronic kidney disease, coronary artery disease, Type 2 diabetes, fasting glucose, fasting insulin, fasting proinsulin, heart rate, inflammatory bowel disease, ulcerative colitis, primary biliary cirrhosis, and systematic lupus erythematosus). **Figure 1** shows a heat map of the results. Note that the first five rows display correlations among the five tobacco and

alcohol use phenotypes from **Table S10**, with the LD Score Regression heritabilities of each phenotype shown down the diagonal of these first five rows. In **Figure 1**, we display correlations significant at the .05 alpha level with a single asterisk, and correlations that remain significant after a Bonferroni correction for 180 tests (p<.000278; the number of cells in the matrix after subtracting the redundant genetic correlations of the five substance use phenotypes as well as subtracting the heritability estimates) are displayed with two asterisks.

For height we found low and generally non-significant correlations with our five phenotypes, as expected. For the sociodemographic and developmental phenotypes, we observed moderate correlations, especially for age of first birth and years of education, such that the variants associated with earlier years of smoking, for engaging or persisting in smoking behavior, or for increased cigarette or alcohol use are significantly and moderately associated with lower age of first birth or lower years of education, as one would expect[44,45]. The correlations for sociodemographic and developmental phenotypes are among the strongest observed for our five phenotypes.

Across the psychiatric phenotypes, risk and other substance use phenotypes, non-psychiatric neurological disease phenotypes, and medical comorbidity phenotypes, we continue to see similar emergent patterns of substance use phenotypes being correlated in the expected direction with less healthy outcomes (**Table S10**). **Figure 1** best conveys this message, with its broad but clear patterns of correlations. For our smoking phenotypes, the strongest genetic correlations emerge for phenotypes like cotinine levels, general risk tolerance, lifetime cannabis use, ADHD, major depressive disorder, neuroticism, BMI and obesity, HDL cholesterol, coronary artery disease, Type 2 Diabetes, and insulin and proinsulin levels. Our drinking phenotype, drinks per week, is generally less strongly genetically correlated across these other phenotypes, with correlations that are often around zero, especially for non-psychiatric neurological diseases and medical comorbidities. This is perhaps expected, as smoking is probably the major behavioral risk factor for these medical outcomes, and hence enriched among the cases incorporated into the respective disease-GWAS studies. We also observe distinct patterns of genetic correlation between smoking and drinking phenotypes for psychiatric traits, where smoking phenotypes generally have stronger genetic correlations than does drinking. The strongest correlations for drinking emerge for general risk tolerance, lifetime cannabis use, BMI and obesity, HDL cholesterol, and insulin levels. Overall, the genetic correlation results from this analysis provide evidence that, at the genetic level, our substance use phenotypes are indeed partially genetic, as distinct from purely environmental, risk factors in a number of other psychiatric, substance use, and medical phenotypes. However, it is difficult to draw strong inferences about causal pathways or the role of mediators from simple bivariate genetic correlations (e.g., variant causes disease which causes smoking cessation). When considering only the most significant genetic correlations after Bonferroni correction (shown with two asterisks in **Figure 1**), age of first birth, years of education, general risk tolerance, lifetime cannabis use, major depressive disorder, neuroticism, schizophrenia, BMI, obesity, HDL cholesterol, and coronary artery disease are generally the most statistically significantly correlated with our substance use phenotypes.

8.4. Possibility of phenotypic heterogeneity

As described above, the five phenotypes showed significant but low heritability attributable to genotyped and imputed variants, as computed by LD Score Regression[26]. Heritabilities ranged from 4.2% for DrnkWk to 8.0% for CigDay, which are consistent with those found previously in much smaller samples[46], and were confirmed with univariate LD Score Regression and GREML-based approaches for individual large studies included in the meta-analysis (**Table S13**). These meta-analytic heritability estimates may be affected by phenotype heterogeneity across constituent studies. To evaluate this possibility further, we compared our meta-analytic heritabilities to those calculated using individual level data[47] in the UK Biobank. The meta-analytic heritabilities were consistently ~4% lower than results from individual-level data in the UK Biobank, suggesting the possibility of phenotypic or other heterogeneity across constituent studies in the meta-analysis, or bias in either the LD Score Regression or GREML estimates. We further evaluated heterogeneity by computing genetic correlations for each phenotype across all of the largest contributing studies (23andMe, UK Biobank, HUNT, GERA) in **Table S13**. For example, we estimated the genetic correlation for cigarettes per day in 23andMe and UKB 350. The result ($\hat{r}_g$= 0.866, *SE* = 0.051), indicates that this phenotype has a similar genetic etiology in the United-States-oriented 23andMe and the United-Kingdom-oriented UKB samples. Lower genetic correlations we observe in this analysis might be explained by country-level

population differences in nicotine metabolism, diversity in public policy (e.g., cigarette prices), acceptability of alcohol consumption, and so on.

More generally, this analysis allows an estimate of the extent to which the phenotypes in each of the larger studies are genetically similar and, by extension, whether these independent studies are measuring similar constructs. In **Table S9,** each of the Panels (A-E) show results for one phenotype across each of the larger studies. For each of our five phenotypes, we nearly always observe high genetic correlations between each of our datasets, often well over 0.8. The binary phenotypes, SmkInit and SmkCes, show particularly high correlations across the datasets, although there is nearly the same level of consistency across AgeSmk, CigDay, and DrnkWk. Lower genetic correlations were largely restricted to comparisons with the Norwegian biobank study, HUNT. For instance, the genetic correlation for age of initiation of regular smoking in HUNT and UKB Axiom was 0.674 (*SE* = 0.438). Other lower correlations included the correlation for drinks per week between HUNT and Kaiser ($\hat{r}_g$= 0.538, *SE* = 0.150) and for drinks per week between Kaiser and UKB Axiom ($\hat{r}_g$= 0.694, *SE* = 0.108), indicating the possibility of only a partial overlap in genetic etiology between alcohol use in Norway versus the US or the UK. Overall, this analysis demonstrated consistency in phenotype measurement and genetic etiology across our studies and, provides additional confidence in the generalizability of genetic effects observed in the larger meta-analysis.

Data from twin, family, and adoption literature strongly supports a substantial genetic influence on smoking initiation and maintenance of smoking. A review summarized that liability to initiation of smoking results from about 60% genetic influences, about 20% shared environmental influences, and about 20% influences specific to individuals[48]. Further, for those who then progress to nicotine dependence, genetic factors played a prominent role in this transition, with liability to nicotine dependence resulting from about 70% genetic influences (with negligible environmental influences). The overlap between smoking initiation and dependence was substantial, but not complete. Further, a twin study of tobacco initiation, regular tobacco use, and nicotine dependence showed heritability estimates of 75%, 80%, and 60% for these phenotypes, respectively[49]. Though studied less frequently than smoking, drinking behavior follows similar patterns in the twin and family literature, with a heritability of about 60% for alcohol use[50,51]. These values are in contrast to those discovered here using genome-wide common variants. The present results suggest that a small to moderate fraction of the total heritability of these phenotypes is detectable with the present study design.

9. Polygenic risk scoring

9.1. Overview

In what follows, we assess how well polygenic risk scores (PRS) for each of our five phenotypes predicted those phenotypes, in holdout prediction samples composed European-descent individuals. First, we describe the methodology used to create the scores; next, we assess the predictive power of our polygenic scores, which come from our summary statistics of AgeSmk, CigDay, SmkInit, SmkCes, and DrnkWk.

9.2. Score construction methods

9.2.1. LDpred

A polygenic score for an individual is a weighted sum of a person's genotypes at *J* loci,

$$\hat{g}_i = \sum_{j=1}^{J} \hat{\beta}_j g_{ij}.$$

where $\hat{g}_i$ denotes the polygenic score of individual i, $\hat{\beta}_j$ is the estimated additive effect size of the effect-coded allele at variant j, and g_{ij} is the genotype of individual i at variant j (coded as having 0, 1, or 2 instances of the effect-coded allele).

The scores in this analysis, rather than using a pruning and thresholding method, are constructed using a Bayesian score generation method called LDpred[52], a model to generate polygenic scores while accounting for linkage disequilibrium between variants. Since we do not know the variance-covariance matrix of the effects in the training sample, we replace this matrix with a block diagonal matrix estimated

using LD patterns from a reference sample of unrelated individuals with European ancestry. Our LD reference samples are cohort-specific; to estimate the block diagonal LD matrix, we use the genotyped data from each of our hold-out prediction cohorts, after dropping cryptically-related individuals or ancestry outliers, to create our LDpred reference samples. All variants that passed our meta-analytic filters (described above) were used when constructing our scores.

Smoking and alcohol use rates are significantly influenced by cultural norms. With this in mind, we selected two independent prediction samples, the Health and Retirement Study (HRS)[53] and the National Longitudinal Study of Adolescent to Adult Health (Add Health)[54]. The HRS is a panel study of U.S. households that began in 1992 with the intention of monitoring physical, emotional, and economic wellbeing during the transition into retirement and older age; it is now representative of the U.S. population over age 50, and its participants are surveyed approximately every two years. In the HRS, the mean birth year of respondents is 1938.4 (SD = 9.3), and the mean age at the time of assessment is 57.6 (SD = 8.9). Add Health includes a much later range of birth years than the HRS. Add Health originated as an in-school survey of a nationally representative sample of U.S. adolescents enrolled in grades 7 through 12 during the 1994-1995 school year. Respondents were born between 1974 and 1983. A subset of the original Add Health respondents has been followed up with in-home interviews, which allows researchers to assess correlates of outcomes in the transition to early adulthood. In Add Health, the mean birth year of respondents is 1979.0 (SD = 1.8), and the mean age at the time of assessment (Wave 4) is 29.0 (SD = 1.8). Given the differences in birth years between the two datasets, the respondents grew up with different cultural norms and policies related to substance use, especially smoking. In the HRS, ~57% of respondents reported ever smoking regularly, and these respondents smoked about ~13 cigarettes per day (**Table S14,** Panel B). In the fourth wave of Add Health, when the mean age of respondents was 29, about 53% of respondents reported ever smoking regularly, and these respondents smoked ~11 cigarettes a day on average (**Table S14**, Panel A).

For the purpose of these PRS analyses, we used meta-analytic results generated on all studies except for HRS and Add Health. After imputing the genetic data from both datasets to the Haplotype Reference Consortium (HRC)[15], we used only HapMap3 variants, which were well imputed and provide good coverage of common variation across the genome, so that we can make more equivalent comparisons across HRS and Add Health. We limited all analyses in the prediction cohorts to European-ancestry individuals. Finally, we used only HapMap3 variants with a call rate above 98% and a minor allele frequency > 1%. In our final score generating step, we used PLINK[55] to multiply the genotype probability of each variant by the corresponding LDpred posterior mean over all variants. For each of our five phenotype scores in the HRS, we used 1,102,465 HapMap3 variants to construct our scores. For Add Health, 1,496,011 HapMap3 variants were used.

We also generated polygenic scores for our five sets of summary statistics from MTAG using the same method (LDpred) as that used in the creation of our original GWAS polygenic scores, since one of the primary motivations for the development of MTAG was increased polygenic prediction[28].

9.3. Measuring prediction accuracy

Prediction accuracy in these analyses was based on an ordinary least squares regression of a given phenotype (AgeSmk, CigDay, SmkInit, SmkCes, or DrnkWk) on the polygenic score along with a set of standard controls, which include age, sex, interaction between age and sex, and the first ten principle components of the variance-covariance matrix of the genetic data.

Prediction accuracy comes from a two-step process where we first regress the phenotype on a standard set of controls without including the PRS. Then, the PRS predictor is added and the difference in R^2 is calculated. For our quantitative phenotypes, AgeSmk, CigDay, and DrnkWk, the predictive power of the PRS is the change in the R^2 in going from the regression without the PRS to the regression with the PRS. For our two binary phenotypes, SmkInit and SmkCes, we measure the incremental pseudo-R^2 from probit regressions. 95% confidence intervals around all R^2 values are bootstrapped with 1000 repetitions each.

9.4. Polygenic scoring results

Table S14 reports the results of our polygenic scoring analyses. All scores significantly predicted the relevant phenotypes for which they were constructed. For AgeSmk, we estimate that a one-standard-deviation increase in the polygenic risk score is associated with an additional 0.31 years in the age of smoking initiation in Add Health and an additional 0.50 years in the HRS. The associated incremental R^2 is low for the AgeSmk score at 0.9% in Add Health and 0.7% in the HRS. For cigarettes per day, an equivalent one-standard-deviation increase in the PRS is associated with about two additional daily cigarettes in Add Health and about three additional cigarettes in the HRS. The R^2 values are 3.9% in Add Health and 4.3% in the HRS. For our final continuous phenotype, DrnkWk, a one-SD increase in the score reflects about an additional drink per week in both datasets, with an associated R^2 of 2.2% in Add Health and 2.4% in the HRS.

For our binary phenotypes, the rows labeled "*Score*" in **Table S14** represent incremental effects, in other words the effect of a one-standard-deviation-increase in the risk score on the probability of the outcome over and above all other covariates. Thus for SmkInit, a one-standard deviation increase in the score was associated with a 12% increase in the probability of being a regular smoker; the comparable increase is 10% in the HRS. The pseudo-R^2 of the SmkInit score was 4.2% in Add Health and 3.6% in the HRS. Finally, for SmkCes, a one-SD increase in the score was associated with a 5% increase in the probability of being a current (versus former) smoker in Add Health and a 3% increase in the HRS. The SmkCes score had a lower incremental pseudo-R^2 at 1.2% in both datasets. We note that the comparability of the R^2 or pseudo-R^2 values was surprising, given that the ages in these datasets were very different, and secular trends in smoking have changed drastically over time.

In **Figure 3**, we show how the R^2 of the PRSs compared to the upper bound on the score's predictive power, which is the SNP heritability of a given phenotype (see **Table S13** for the heritabilities). Comparing the values for Add Health and HRS to the SNP heritability values, we find that our scores for AgeSmk and SmkCes performed most poorly of the five scores, with low R^2 values and smaller percentages of the variance accounted for relative to the SNP heritability of these phenotypes. This is perhaps best explained by lower samples sizes for the AgeSmk and SmkCes phenotypes relative to, for instance, SmkInit or DrnkWk. Further, phenotypes like AgeSmk and SmkCes are more prone to measurement error than a phenotype like SmkInit, driving down the overall explainable (non-error) variance in AgeSmk and SmkCes.

However, our scores for CigDay, SmkInit, and DrnkWk performed well relative to their SNP heritability estimates and account for, on average, about half of the SNP heritability in each of these phenotypes. This was particularly exciting for CigDay, since this phenotype has a smaller sample size that is more comparable to AgeSmk. For all smoking phenotypes, the scores are well-powered and represent significant gains in prediction compared to scores made from the 2010 TAG scores[1], with the highest levels of prediction from the CigDay and SmkInit scores.

Results from polygenic scores derived from the MTAG results are reported in **Table S32**. In comparing the polygenic scores from our original GWAS results (**Table S14**) to those derived from MTAG, there is generally negligible increase in predictive power (measured by incremental R^2 or incremental pseudo-R^2), and these slight gains are similar between the two hold-out prediction datasets (the HRS and Add Health). The increase in R^2 ranges from ~0.1% for Age of Initiation of Regular Smoking (AgeSmk) to a little over a full percentage point for Cigarettes Per Day (CigDay). The increase for Cigarettes Per Day is notably by far the largest increase in predictive power from use of MTAG.

9.5. Variance accounted for by genome-wide significant loci

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

The proportion of variance accounted for by the set of all conditionally independent variants is 0.00173 for AgeSmk (*SE* = 0.00017), 0.01087 for CigDay (*SE* = 0.00039), 0.02315 for SmkInit (*SE* = 0.00044), 0.00096 for SmkCes (*SE* = 0.00011), and 0.00187 for DrnkWk (*SE* = 0.00057). These results are also depicted as percentages in the dark gray bars in **Figure 3**. While the variance explained by conditionally independent genome-wide significant variants is small for AgeSmk, SmkCes, and DrnkWk compared to CigDay and SmkInit, for a complex trait it should be roughly proportional to the LDSC-based heritability of the phenotypes. CigDay and SmkInit have nearly double the LDSC heritability of the other three phenotypes (see **Table S15**).

10. GWAS catalogue lookups

To investigate other GWAS phenotypes for which our loci have been implicated, we performed lookups of our conditionally independent loci for each of our five phenotypes in the NHGRI-EBI GWAS Catalog database[56], which catalogues genome-wide significant associations from previous GWAS.

The results of this exercise are reported in **Table S33**. In this table, we include information, by phenotype, about each conditionally independent locus that has been implicated in another GWAS. For ease of comparison, where relevant we included information about both the association reported in the present meta-analysis and the association reported in a given catalogue GWAS. We also manually checked and reported each GWAS catalogue effect allele to determine whether the beta or odds ratio reported in the catalogue had a concordant sign with the beta from our phenotypes.

While some reported associations are in general difficult to draw logical connections between, or are implicated in underpowered GWAS (e.g., rs12203592, from Smoking Cessation being implicated in an underpowered GWAS of tanning), many of the results are confirmatory of the phenotypes we study. For instance, we find association with lung cancer for Age of Initiation; fibrinogen levels, squamous cell lung cancer, and nicotine metabolite and dependence for Cigarettes Per Day; lung cancer, breast cancer, and smoking behavior for Smoking Initiation; cutaneous squamous cell carcinoma, basal cell carcinoma, and local histogram emphysema pattern for Smoking Cessation; and alcohol consumption and liver enzyme levels in Drinks Per Week. Further, we find many shared associations between our results and negative health outcomes like worsening pulmonary function and schizophrenia for smoking phenotypes and higher adiposity measures for Drinks Per Week.

11. Functional enrichment

We used multiple methods to evaluate the genes, cell types, tissues, and biological pathways in which our meta-analytic findings show enrichment. Many tools have been created and employed for these purposes, with known and unknown strengths and weaknesses. To help avoid the possibility of potentially spurious results based on the employ of a single tool, we implemented multiple tools for all analyses. A result that was consistently detected as statistically significant by all methods was assigned the highest confidence; one detected in only one method was assigned less confidence; and a result with no significant support from any method was not considered further.

11.1. Cell Group enrichment

11.1.1. Processing summary statistics

Due to data sharing restrictions, these analyses are based on a subset of the data that do not include 23andMe. For each of the 5 phenotypes, we processed summary statistics uniformly as follows. Consistent with the approach used to filter variants for all other analyses, we first filtered out variants with effective sample size lower than 10% of the maximum effective sample size over all SNPs. The remaining variants were then selected for those also found in 1000 Genome Project Phase 3 (version 5). In total, we retained about 14 million out of the 17 million SNPs in the original summary statistics for each trait for subsequent functional enrichment analyses.

11.1.2. LD pruning

We ran PriorityPruner (version 0.1.4) (http://prioritypruner.sourceforge.net) to recursively remove variants that are within 100-kb distance and in high LD ($r^2 > 0.6$) with the most significant variants. The 100-bp window was chosen to remove fewer SNPs compared to using a larger window size although in general

the global enrichments are robust to different window sizes. As a result, all of the SNPs from each trait were either labeled as selective SNPs or as tagged (i.e., pruned) and removed from the subsequent analyses or selected SNPs. We took only the selected SNPs for the genome-wide enrichment analyses as described next.

11.1.3. Functional genomic and cell-type-specific annotations

Each variant was annotated by the 272 annotations obtained from LD score regression database (https://data.broadinstitute.org/alkesgroup/LDSCORE/)[38]. There are 52 baseline annotations[57] and 220 cell-type specific annotations over 10 cell groups. The 220 cell-type-specific annotations were originally generated by the ENCODE and Roadmap Epigenomic Consortium[58] and represent 100 well defined cell types over four different histone marks namely H3K4me1, H3K4me3, H3K27ac, and H3K9ac. The 100 cell types can be grouped into 10 cell groups depending upon their primary tissues of origin. Note that not all of the cell types have four histone marks measured, as indicated by the presence of 220, rather than 400, annotations.

11.1.4. Partitioned LD score regression

To detect genome-wide functional and tissue-specific epigenomic enrichments, we performed enrichment analyses by heritability stratification using LD score regression (LDSC) implemented in the LDSC software (https://github.com/bulik/ldsc/). LDSC differs from some other existing methods such as the restricted maximum (REML) implemented in GCTA in estimating heritability as it only requires summary statistics using the reference LD panel as surrogate to the sample-specific genotype covariance matrix. LDSC also differs from functional enrichment methods[59-61] in that it operates on genome-wide SNPs instead of only the GWAS loci based on the fact that the Z-score association statistic for a given variant is a function of the effects of all of the variants it tags via linkage disequilibrium.

Specifically, annotation-stratified LD scores were estimated using dichotomized/binary annotations, 1000 Genomes Project samples with European ancestry, and one million base-pairs LD window by default. LDSC then determines functional enrichments of the GWAS traits by partitioning heritability according to the variance explained by the LD-linked SNPs belonging to each functional category[38]. Statistical enrichment was defined as the ratio between the percentage of heritability explained by variants in each annotated category and the percentage of variants covered by that category. A resampling-based approach was then used to assess standard error estimates[38]. In practice, LDSC only uses HapMap3 SNPs to fit the LD score regression model and assumes that HapMap3 SNPs are sufficient for tagging all 1KG SNPs through LD.

As suggested in the online LDSC user manual, the 53 baseline annotations of LD score regression were always included in the model across all analyses to account for a non-tissue-specific effect. We then followed the suggested protocol of LDSC and kept baseline annotations in the model while adding each annotation track one at a time to test for enrichment. In particular, We obtained the LD score regression software LDSC (v1.0.0) (https://github.com/bulik/ldsc) to determine functional enrichments of the GWAS traits by partitioning heritability according to the variance explained by the LD-linked SNPs belonging to each functional category[38]. Following the online LDSC manual, we first trained a baseline LDSC model using the 52 non-cell-type specific functional categories (plus one category that includes all SNPs) using the observed Z-scores of HapMap3 SNPs for each trait. We then trained 220 models on cell-type-specific annotations including 4 histone marks (H3K4me1, H3K4me3, H3K9ac, H3K27ac) and 100 well-defined cell types. In particular, we tested cell-group enrichments over 10 pre-defined cell-group annotations, namely: adrenal/pancreas, cardiovascular, central nervous system (CNS), connective/bone, gastrointestinal, immune, kidney, liver, skeletal muscle, and others. The cell-group annotations are the results of aggregating 220 cell-type-specific annotations over 4 histone marks (H3K4me1, H3K4me3, H3K9ac, H3K27ac) and 100 well-defined cell types. To detect which specific epigenomes contribute to the group-level enrichment, we also performed 220 tests over each individual annotation. We then reported the p-values Bonferroni-corrected for multiple testings over 50 tests for the cell-group annotation enrichment analyses (10 annotations times 5 addiction traits) and 1100 tests for the cell-specific enrichment analyses (220 annotations times 5 traits).

11.1.5. RIVIERA

As a complementary method to the LDSC analyses described above, we previously developed an efficient mixture model learning approach[62]. Briefly, we take into account three lines of evidence: (1) genome-wide genetic signals in terms of marginal summary statistics of each SNP; (2) functional genomic and epigenomic reference annotations overlapping the SNPs; and (3) linkage disequilibrium, by estimating SNP-by-SNP Pearson correlations from the 1000 Genome Project European reference panel (phase 1 version 3). The general scheme of our learning algorithm follows expectation-maximization (EM) updates: at E-step, we infer the posterior probabilities of causal SNPs by fixing their functional enrichments; at M-step we learn the functional enrichments over all of the annotations in concordance to the posterior probabilities of the SNP associations. We then alternate between the E and M-step until some convergence criterion is satisfied. Here the SNP posteriors are defined by the likelihood dictated by the genetic signals and the prior dictated by a logistic function, which is a weighted linear combination of the annotations. Because of the linearity of the model, we interpret the highly positive coefficient weights as functional enrichments of the corresponding annotations. To account for non-cell-type specific enrichments, we also included 52 baseline annotations as the linear intercepts.

11.2. Cell group enrichment results

11.2.1. Epigenomic enrichments at the cell-group level

We first investigated what cell-group epigenomes are enriched for genetics signals of the five traits using LDSC. To this end, we tested 10 pre-defined cell-group annotations, namely: adrenal/pancreas, cardiovascular, central nervous system (CNS), connective/bone, gastrointestinal, immune, kidney, liver, skeletal muscle, and others -- according to a previous study[38]. CNS cell-group epigenome exhibits the strongest enrichments for all five traits significantly enriched for four of the five traits (except for SmkCes) at Bonferroni-corrected p-value <0.01 (**Figure S12**). Interestingly, in RiVIERA (a probabilistic model using regularization), CigDay, DrnkWk, and AgeSmk are also strongly enriched for liver cell types, potentially related to well-known metabolic pathways for nicotine and ethanol possibly including aldehyde (e.g., the region around *ALDH2* was significantly associated in the GWAS). The primary point of agreement, and the results in which perhaps the most confidence is warranted, is that both methods agree that CNS is strongly enriched in all of the five addiction traits. For the LD Score Regression 220 cell-type-specific enrichment results, we observed that most significantly enriched annotations are in the CNS and/or Liver categories (**Figure S13**).

11.2.2. Genes implicated by genome-wide significant loci

For each phenotype, we used SEQMINER[63] and UCSC genome browser annotations (refFlat.txt.gz; obtained December 15 2017) to annotate all conditionally independent genome-wide significant variants. These annotations are listed in **Tables S1-S5**. Using these same annotations, we also identified all genes harboring at least one variant within LD r^2>0.3 with any conditionally independent variant. Gene boundaries were defined as spanning the 5' to 3' untranslated regions. The genes corresponding to each conditionally independent variant are also listed in **Tables S1-S5** for convenience.

11.2.3. PASCAL

We ran PASCAL[64] on our meta-analytic results for each of the 5 summary statistics to estimate the gene scores and subsequently the enrichments of canonical pathways from MSigDb. Default settings were used, except we set –maxsnp=-1 to ensure that results were calculated for all genes regardless of the number of variants in each gene.

Depending on the phenotype, between 21,873 and 21,907 genes were tested for association. After Bonferroni correction for multiple testing of 21,900 genes (i.e., $p<2.3\times10^{-6}$), we identified 19 genes significantly associated with AgeSmk, 153 genes with CigDay, 127 with SmkCes, 736 with SmkInit, and 310 with DrnkWk. Significant genes within 500kb of a conditionally independent variant from the full meta-analysis are also listed along with the variant in **Tables S1-S5**, for ease of reference. All significant genes for all phenotypes are included in **Table S20**.

PASCAL also tests for gene-based enrichment within 1,077 predefined canonical pathways from MSigDb. After Bonferroni correction of the χ^2 p-value for 1,077 tests (i.e., $p < 4.6\times10^{-5}$), we discovered three pathways enriched for CigDay, all involving acetylcholine (Reactome presynaptic nicotinic acetylcholine receptors, Reactome highly calcium permeable postsynaptic nicotinic acetylcholine receptors, and Reactome acetylcholine binding and downstream events). We identified two pathways in SmkCes, again both related to acetylcholine (Reactome presynaptic nicotinic acetylcholine receptors, and Reactome highly calcium permeable postsynaptic nicotinic acetylcholine receptors). We discovered one pathway enriched in SmkInit, Reactome Developmental Biology, a gene set of 395 genes involved in organism development. No pathways were enriched at this Bonferroni level of significance for AgeSmk or DrnkWk. Results for all tested pathways for all phenotypes are listed in **Table S21**. We also applied a false discovery rate to the PASCAL pathway results, which resulted in no change to the identified pathways.

We also ran PASCAL for the MTAG summary statistics, and report these findings in **Table S34**. Many more genes were significantly associated than for the original GWAS results, mirroring the increase in the number of genome-wide significant variants. The count included 539 genes associated with AgeSmk, 334 for CigDay, 338 for SmkCes, 1,614 for SmkInit, and 515 for DrnkWk.

PASCAL pathway analysis of MTAG summary statistics resulted in some small apparent gain over the original GWAS pathway results. The only notable differences were that the pathway 'Reactome Developmental Biology' was associated with AgeSmk and SmkInit. New to the MTAG results, 'Reactome immune system' and 'Reactome hemostasis' were associated with SmkInit. These results are provided in **Table S35**.

11.2.4. DEPICT

DEPICT, or Data-driven Expression Prioritized Integration for Complex Traits[65], is a bioinformatics toolkit that is used to prioritize reconstituted gene sets and identify the tissues and cell types in which causal genes are likely to be active. DEPICT was developed using gene expression data derived from 77,840 samples (gene expression microarrays) and public pathway annotation datasets (GO, KEGG, REACTOME, MGI, and InWeb). DEPICT uses 10,968 reconstituted gene sets to identify pathways enriched by association findings.

Here we ran DEPICT using the default settings and the GWAS summary statistics for our five phenotypes: age of initiation (AgeSmk), cigarettes per day (CigDay), smoking initiation (SmkInit), smoking cessation (SmkCes), and drinks per week (DrnkWk). For DEPICT input, summary statistics were clumped for each trait using a 500 kb flanking region and an LD cutoff of r^2 >0.1. Here, we use DEPICT as a supplement to understand genetic signals beyond the genome-wide significant loci. We excluded variants with a *p*-value greater than 5×10^{-5}. Finally, and again as a default setting in DEPICT, samples of European ancestry from the 1000 Genomes project (phase 1 release V3) were used to compute LD between variants.

We also applied DEPICT to the MTAG summary statistics. As expected, the number of loci discovered by applying DEPICT to MTAG increased the number of pathways, genes, and tissues associated with each phenotype. Given that MTAG results are simply a weighted linear combination of the effect sizes from the original GWAS it is even less straightforward than usual to interpret the biological meaning of any given associated gene/pathway/tissue. This, combined with the elevated max FDR rates for AgeSmk and SmkCes, should engender some caution when interpreting these results.

11.2.4.1. Tissue Enrichment

Tissue enrichment results are reported for each phenotype in **Table S16** and displayed in **Figure S14**. Using gene expression data, enrichment is shown for 209 Medical Subject Headings (MeSH) annotations. In the table and figure, results are classified more broadly into 13 MeSH first-level terms, which help classify findings into physiologically-relevant tissue and cell systems. In **Figure S14**, the results are displayed the most clearly. Here each column represents one of our phenotypes, and MeSH first-level terms are color coded. Relative to genes in random sets of loci, enrichment is

indicated by red shading, and significant enrichment is noted with single asterisks (False Discovery Rate (*FDR*) < 0.05) or double asterisks (*FDR* < 0.01).

At the *FDR* < 0.05 and *FDR* < 0.01 levels, these analyses implicated zero (AgeSmk), two (CigDay), 28 (SmkInit), zero (SmkCes), and zero (DrnkWk) tissues/cells, depending on the phenotype. Nearly all implicated tissues or cell types were related to the central nervous system (CNS) including, for CigDay, the corpus striatum and basal ganglia, two regions known to be involved in addiction-related networks. Findings for SmkInit implicated many central nervous system tissues including brain, central nervous system, cerebral cortex, cerebrum, metencephalon, rhombencephalon, telencephalon, cerebellum, limbic system, hippocampus, prosencephalon, brain stem, parahippocampal gyrus, entorhinal cortex, temporal lobe, parietal lobe, basal ganglia, frontal lobe, corpus striatum, and the mesencephalon. Embryonic stem cells and neural stem cells also showed enrichment.

Perhaps interestingly, the three most strongly enriched tissues in our DEPICT analyses were in the visual system, including the retina, occipital lobe, and visual cortex. Cigarette smoking is known to increase risk for a host of eye diseases including macular degeneration[66], glaucoma[67], and cataracts[68]. Broadly, the present results may therefore suggest complex pathways between smoking and eye-related disorders or visual processing, though additional research will of course be needed to understand these potential relationships.

DEPICT results based on MTAG summary statistics expanded the number of associated tissues at the FDR < .05 level. There were 26 tissues showing enrichment for AgeSmk, 24 for CigDay, 26 for SmkCes, 30 for SmkInit, and 28 for DrnkWk. The results showed highly consistent enrichment across essentially all neural tissues enriched with SmkInit in the DEPICT results described above. The only non-neural tissues were retina (enriched in all five phenotype MTAG results), embryonic stem cells (SmkInit only), male genitalia (DrnkWk), and prostate (DrnkWk). All enrichments at FDR < .05 are listed in **Table S36**.

11.2.4.2. Prioritized genes and gene sets

DEPICT-prioritized genes are reported in **Table S17**. There were zero (AgeSmk), 40 (CigDay), 513 (SmkInit), zero (SmkCes), and one (DrnkWk) prioritized genes across our five phenotypes with *FDR* < 0.05. While our GWAS strongly implicated nicotinic receptor genes associated with CigDay, SmkCes, and to a lesser extent SmkInit, nicotinic receptor genes identified by DEPICT were only prioritized for SmkInit at *FDR* < 5%. These genes included *CHRNB2*, *CHRNA5*, *CHRNA3*, *CHRNA4*, *CHRNB4*, and *CHRNA2*. Again, the MTAG results were far more numerous, including 330 genes prioritized for AgeSmk at FDR < .05, 109 for CigDay, 152 for SmkCes, 1,121 for SmkInit, and 81 for DrnkWk. Nicotinic receptor genes were prioritized across all smoking phenotypes in the MTAG DEPICT results. The additional genes are extensive and listed in **Table S37**.

DEPICT gene sets are taken from public databases GO, KEGG, REACTOME, MGI, and InWeb. They have been “reconstituted” by using co-expression data to estimate evidence of a gene belonging to a gene set. The 25 genes in the DEPICT-defined loci with the highest membership scores with respect to each significantly enriched gene set are found in columns F-AD of **Table S18**. (MTAG DEPICT results are in the same columns in **Table S38**.)

DEPICT results included zero (AgeSmk), 12 (CigDay), 766 (SmkInit), zero (SmkCes), and 51 (DrnkWk) reconstituted gene sets with enrichment at FDR < 0.05. Pathways implicated in CigDay include 11 protein-protein interaction subnetworks (GNAO1, SIN3A, HDAC3, EYA1, GATA1, RUNX1T1, ACTL6A, RIC8A, MTA2, ATXN1, and BCL7A), which appear to be involved in DNA machinery including transcription factors and chromatin remodeling (**Table S18**).

Significant gene set enrichment results for SmkInit were numerous, implicating about 7% of all tested pathways and significantly enriched for pathways that affect CNS activity and development. Among the top 20 pathways at FDR < 0.05, all of them affected the brain, including GO presynaptic membrane, GO dendrite development, GO dendrite, GO neuron recognition, MP abnormal neurite

morphology, MP impaired coordination, DLG4 PPI subnetwork, GO brain development, MP abnormal CNS synaptic transmission, GO growth cone, GO axon, GO site of polarized growth, GO forebrain development, GO synapse, GO synaptic membrane, MP abnormal locomotor activation, GO axonogenesis, MP abnormal olfactory bulb morphology, and GO dendrite morphogenesis. This general trend of enrichment exists for hundreds of additional pathways enriched in SmkInit, and generally implicates many basic biological functions related to the propensity to smoke cigarettes regularly.

Gene set enrichments in DrnkWk were varied. Of the top 10 implicated pathways, half were directly related to non-AMPA glutamate signaling. Each of these reconstituted gene sets were protein-protein interaction subnetworks including GRIN2A, GRIN1, GRIN2B, DLG3, and GRM1. Other reconstituted gene sets related to glutamatergic neurotransmission significant at FDR < 0.05 included protein-protein interaction subnetworks defined by GRIK2, DLG4, and DLG1. Also among the top 10 were protein-protein interaction subnetworks defined by proteins involved in stress, hormones, or immune function (RPS6KA3, DUSP4, and PPP5C). Other gene sets showing significant enrichment were related to dopamine (Reactome DARPP:32 events); opioid signaling (Reactome opioid signaling); cell adhesion and differentiation (CDH2 PPI subnetwork, Reactome nuclear receptor transcription pathway, GO regulation of cell morphogenesis involved in differentiation, and GO regulation of cell morphogenesis); gross brain structure and function (MP abnormal hippocampus morphology, GO axon, and MP enlarged lateral ventricles); processing of sugars (Reactome gluconeogenesis, GO response to carbohydrate stimulus, Reactome metabolism of carbohydrates, and Reactome glucose metabolism); decreased birth body size; and abnormal behavior (MP impaired coordination, MP abnormal contextual conditioning behavior, and MP abnormal locomotor activation).

To help condense and visualize the results, for Smoking Initiation and Drinks Per Week we used affinity propagation clustering[69] to cluster related DEPICT reconstituted gene sets. First, we removed duplicate gene sets that could potentially affect our false discovery rates. We then calculated a correlation matrix of all significant (FDR < 0.05) DEPICT reconstituted gene sets, using only DEPICT prioritized genes with FDR < 0.2. We finally clustered the DEPICT reconstituted gene sets for SI and DrnkWk using the apcluster function in R. In total, after dropping duplicates, we found 68 gene set clusters for SmkInit, which include 740 reconstituted gene sets found by DEPICT to be significantly enriched; and 10 gene set clusters for DrnkWk, representing 50 significantly enriched reconstituted gene sets.

Figure 4 summarizes these results. For each of the 68 clusters in Panel A and each of the 10 clusters in Panel B, clusters are named by the exemplar gene set for that cluster as determined by the affinity propagation algorithm. In the figure, correlations between exemplary gene sets are shown for each cluster. Full names of clusters, as well as the gene sets that were assigned to each cluster, are listed in **Table S19**.

DEPICT results based on MTAG summary statistics included 588 (AgeSmk), 215 (CigDay), 766 (SmkInit), 1322 (SmkCes), and 309 (DrnkWk) reconstituted gene sets with enrichment at FDR < 0.05. These pathways are listed in **Table S38**. We applied the same affinity propogation clustering algorithm to the results for each phenotype, and provide these results in **Table S39**. The general thrust of these results was highly similar to that provided in Figure 4 for SmkInit based on the initial GWAS results. The vast majority of enriched pathways, especially the most significantly enriched ones, were involved in neurotransmitter receptors, ion channels, learning/memory, brain structure and other aspects of CNS function. A large number of pathways were also related to regulation of transcription and translation, and development of the nervous system. One perhaps notable change was for DrnkWk, where pathways related to glucose and carbohydrate processing were no longer associated.

12. Contributions and acknowledgements

12.1. Detailed author contributions

Dajiang J. Liu and Scott Vrieze led and oversaw the study.

Goncalo Abecasis, Dajiang J. Liu, and Scott Vrieze designed the study.

Mengzhen Liu was the study's lead analyst, responsible for quality control and meta-analyses; she was primarily assisted by Yu Jiang, and she was also assisted by Dajiang J. Liu, Scott Vrieze, and Robbee Wedow. Bonferroni thresholds were calculated by Daniel McGuire.

Phenotype definitions were developed by Laura J. Bierut, Marilyn C. Cornelis, David A. Hinds, Jaakko Kaprio, Eric Jorgenson, Dajiang J. Liu, Matt McGue, Marcus R. Munafo, Scott Vrieze, and Luisa Zuccolo.

Software development and implementation for the study were carried out by Yu Jiang, Dajiang J. Liu, and Xiaowei Zhan.

Conditional analyses were performed by Yu Jiang and Mengzhen Liu.

Analyses in the UK Biobank were performed by David M. Brazel and Gargi Datta.

Heritability, genetic correlation, and polygenic scoring analyses were performed by Robbee Wedow.

Pleiotropy analyses were designed and executed by Dajiang J. Liu.

Bioinformatics and biological insights analyses were performed by Fang Chen (DEPICT), Yue Li (Partitioned LD Score Regression; PASCAL; RiVIERA); Mengzhen Liu (DEPICT; GWAS catalogue lookup; manual review of implicated genes; PASCAL), Jose Davila-Velderrain (PASCAL; RiVIERA), Scott Vrieze (DEPICT; PASCAL; manual review of implicated genes), and Robbee Wedow (DEPICT; GWAS catalogue lookup; manual review of implicated genes). In interpreting results, they were assisted in major ways by James J. Lee and Jerry A. Stitzel.

The LocusZoom website was designed and implemented by Gargi Datta.

The majority of figures were created by Mengzhen Liu and Robbee Wedow.

Yueh Ling, Mengzhen Liu, Scott Vrieze, Robbee Wedow, and Hannah Young helped with coordinating among the participating cohorts. Marissa A. Ehringer and Matthew C. Keller helped with data access. Robbee Wedow managed and coordinated all authorship and acknowledgement details for the study.

Marilyn C. Cornelis, Sean P. David, Eric Jorgenson, Jaakko Kaprio, and Jerry A. Stitzel provided particularly helpful advice and feedback on various aspects of the study design and the manuscript.

All authors contributed to and critically reviewed the manuscript. Yue Li, Dajiang J. Liu, Mengzhen Liu, Scott Vrieze, and Robbee Wedow made major contributions to the writing and editing.

12.2. Cohort-level contributions

Cohort	Author	Design & Management	Data Collection	Genotyping	Genotype Prep.	Phenotype Prep.	Data Analysis
23andMe	Chao Tian					X	X
23andMe	David Hinds	X	X	X	X		
Add Health	Jason D. Boardman	X					
Add Health	Kathleen Mullan Harris	X	X				
Add Health	Matthew B. McQueen	X					
ALSPAC	George McMahon					X	X
ALSPAC	Amy E. Taylor					X	X
ALSPAC	Marcus R. Munafò	X					
ALSPAC	Luisa Zuccolo	X					
BBJ	Nana Matoba					X	X
BBJ	Yoichiro Kamatani	X					
BBJ	Yukinori Okada	X					
BLTS	Anna R. Docherty					X	X
BLTS	Nathan A. Gillespie	X	X			X	X
BLTS	Ian B. Hickie	X	X				
CADD	John K. Hewitt	X	X			X	X
CADD	Christian J. Hopfer	X	X				
CADD	Kenneth S. Krauter	X	X	X	X		
CADD	Michael C. Stallings	X	X		X	X	X
CADD	Tamara L. Wall	X	X				X
COGEND	Laura J. Bierut	X	X	X	X	X	X
COGEND	Eric O. Johnson	X	X	X	X	X	X
COGEND	John P. Rice	X	X	X	X	X	X
COGEND	Nancy L. Saccone	X	X	X	X	X	X
COPDGene	Kendra A. Young				X	X	X

COPDGene	John E. Hokanson	X	X	X		X	X
COPDGene	Sharon M. Lutz	X			X	X	X
deCODE	Gunnar W. Reginsson						X
deCODE	Gyda Bjornsdottir	X	X			X	
deCODE	Daniel F. Gudbjartsson	X		X			X
deCODE	Valgerdur Runarsdottir		X			X	
deCODE	Hreinn Stefansson	X		X	X		
deCODE	Kari Stefansson	X					
deCODE	Thorgeir E. Thorgeirsson	X					X
deCODE	Thorarinn Tyrfingsson		X			X	
EGCUT	Reedik Mägi				X		X
EGCUT	Tõnu Esko	X					
EGCUT	Toomas Haller						X
EGCUT	Andres Metspalu		X				
Finntwin & NAG-FIN	Teemu Palviainen					X	X
Finntwin & NAG-FIN	Jaakko Kaprio	X	X	X		X	
Finntwin & NAG-FIN	Anu Loukola			X	X		
Finntwin	Richard J. Rose	X	X				
GERA	Helene Choquet	X				X	X
GERA	Eric Jorgenson	X			X	X	X
GERA	Khan K. Thai				X	X	X
GERA	Constance Weisner	X					
GERA	Jie Yin					X	X
GfG	Johanna R. Foerster	X	X				
GfG	Anita Pandit	X	X		X		
GfG	Gregory J.M. Zajac		X		X	X	X
HRS	Jessica D. Faul	X	X	X		X	
HRS	Jennifer A. Smith			X	X		X

HRS	Wei Zhao			X	X		X
HRS	Sharon L. R. Kardia			X	X		
HRS	David Weir	X	X	X			
HUNT	Maiken Elvestad Gabrielsen						X
HUNT	Anne Heidi Skogholt						X
HUNT	Wei Zhou						X
HUNT	Kristian Hveem				X	X	X
HUNT	Jonas B. Nielsen	X					
HUNT	Cristen J. Willer	X					
HUNT	Bendik Slagsvold Winsvold	X					
MCTFR	William G. Iacono	X	X	X		X	
MCTFR	Matt McGue	X	X	X	X	X	X
METSIM	Michael Boehnke		X	X			
METSIM	Markku Laakso		X			X	
METSIM	Karen L. Mohlke		X	X			
METSIM	Alena Stančáková		X				
NESCOG	Philip R. Jansen				X		X
NESCOG	Danielle Posthuma	X	X	X			
NESCOG	Tinca J. C. Polderman	X	X	X		X	
NHS, NHS2, & HPFS	Hongyan Huang				X	X	X
NHS, NHS2, & HPFS	Constance Turman				X		X
NHS, NHS2, & HPFS	Marilyn C. Cornelis	X	X	X	X	X	X
NHS, NHS2, & HPFS	David J. Hunter	X	X	X			
NHS, NHS2, & HPFS	Peter Kraft	X	X	X	X	X	X
NHS, NHS2, & HPFS	Eric Rimm	X	X	X			

NTR	Jouke-Jan Hottenga				X		X
NTR	Gonneke Willemsen		X			X	
NTR	Dorret I. Boomsma	X	X				
NTR	Gareth E. Davies			X			
OZALC	Scott Gordon						X
OZALC	Andrew C. Health	X	X		X	X	
OZALC	Penelope A. Lind		X		X	X	
OZALC & NAG-FIN	Pamela A. F. Madden	X	X		X	X	
OZALC	Nicholas G. Martin	X	X		X	X	
OZALC	Sarah E. Medland	X	X		X	X	
OZALC	John B. Whitfield	X	X		X	X	
SardiNIA	Antonella Mulas				X		
SardiNIA	Valeria Orrù					X	
SardiNIA	Francesco Cucca	X	X	X			
SardiNIA	Edoardo Fiorillo		X	X			
WHI	Jeffrey Haessler				X		X
WHI	Chu Chen	X					
WHI	Sean P. David	X					
WHI	Charles B. Eaton	X					
WHI	Charles Kooperberg	X	X	X		X	
WHI	Ulrike Peters	X		X		X	
WHI	Alexander P. Reiner	X		X		X	
WHI	Hilary A. Tindle	X					

12.3. Additional acknowledgements

23andMe, Inc. — 23andMe research participants provided informed consent to take part in this research under a protocol approved by the AAHRPP-accredited institutional review board, Ethical and Independent Review Services. We would like to thank the research participants and employees of 23andMe for making this work possible. The full GWAS summary statistics for the 23andMe datasets will be made available to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please contact apply.research@23andme.com for more information and to apply to access the data.

Add Health (National Longitudinal Study of Adolescent to Adult Health) — The National Longitudinal Study of Adolescent to Adult Health (*Add Health*) is supported by grant P01 HD031921 to Kathleen Mullan Harris from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. *Add Health* GWAS data were funded by NICHD grants to Harris (R01 HD073342) and to Harris, Boardman, and McQueen (R01 HD060726). For information about access to the data from this study, contact addhealth@unc.edu.

ALSPAC (Avon Longitudinal Study of Parents and Children) — We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. This research was specifically funded by the UK Medical Research Council and the University of Bristol (MC_UU_12013/1, MC_UU_12013/6). LZ is supported by the UK Medical Research Council (G0902144). AET and MRM are members of the UK Centre for Tobacco Control Studies, a UKCRC Public Health Research: Centre of Excellence. Funding from British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. This publication is the work of the GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN), which serves as guarantors for the contents of this paper.

Descriptions of the ALSPAC cohort can be found in the two following articles: (1) Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, Molloy L, Ness A, Ring S, Davey Smith G. Cohort Profile: The 'Children of the 90s'; the index offspring of The Avon Longitudinal Study of Parents and Children (ALSPAC). *International Journal of Epidemiology* 2013; 42:111-127; (2) Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, Ring S, Nelson SM, Lawlor DA. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology* 2013; 42:97-110. Study data for individuals 22 years and older were collected and managed using REDCap

electronic data capture tools hosted at the University of Bristol. REDCap (Research Electronic Data Capture) is a secure, web-based application designed to support data capture for research studies, providing 1) an intuitive interface for validated data entry; 2) audit trails for tracking data manipulation and export procedures; 3) automated export procedures for seamless data downloads to common statistical packages; and 4) procedures for importing data from external sources. The tool is described in detail in the following article: Paul A. Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, Jose G. Conde, Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support, *Journal of Biomedical Informatics* 2009; 42(2):377-381.

Please note that the ALSPAC study website contains details of all the data that is available through a fully searchable data dictionary available here: http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Details on the ethics committee/institutional review board that approved aspects of the study can be found here: http://www.bristol.ac.uk/alspac/researchers/research-ethics/. For more information about this dataset, see http://www.bristol.ac.uk/alspac/.

ARIC (Atherosclerosis Risk in Communities) — The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions. Funding for GENEVA was provided by National Human Genome Research Institute grant U01HG004402 (E. Boerwinkle). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000280.v3.p1.

BBJ (BioBank Japan Project) — The Biobank Japan (BBJ) Project was established in 2003 with the aim of the implementation of personalized medicine as a leading project of Ministry of Education, Culture, Sports, Science and Technology (MEXT). In collaboration with twelve cooperating institutes, the BBJ has recruited a total of 200,000 people, suffering from at least one of the 47 target common diseases, in the first phase (5-year period). BBJ has collected biospecimens including DNA and serum as well as various clinical and lifestyle information through interview or medical records by using standardized questionnaire. All participants gave written informed consent to this project and this study was approved by ethical committees of RIKEN and participating institutes. For more information about this study, please see https://biobankjp.org/english/plan/summary.html.

BEAGESS (The Barrett's and Esophageal Adenocarcinoma. Genetic Susceptibility Study) — This study made use of data generated by investigators in the BEACON consortium through a grant funded by the US National Institutes of Health (NIH) (RO1CA136725) to Thomas L. Vaughan and David C. Whiteman (multiple PIs). In support of this work, T.L.V. was also supported

by NIH grant KO5CA124911 and D.C.W. by a Future Fellowship grant FT0990987 from the Australia Research Council. Additional collaborators, sources of support and origin of the data and biospecimens are listed in the following publication: Levine DM, Ek WE, Zhang R, Liu X, Onstad L, Sather C, et al. A genome-wide association study identifies new susceptibility loci for esophageal adenocarcinoma and Barrett's esophagus. Nat Genet. 2013 Dec;45(12):1487-93. The dataset used for the analyses described in this manuscript was obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000869.v1.p1.

BLTS (Brisbane Longitudinal Twin Study) — The Brisbane Longitudinal Study acknowledges funding from the Australian National Health and Medical Research Council grants 496682, 1009064, 1009064, and 49662. Nathan A. Gillespie is supported by NIH R00DA023549, and Anna R. Docherty is supported by MH109765. For more information about this study, contact Nathan A. Gillespie (nathan.gillespie@vcuhealth.org).

CADD (Center on Antisocial Drug Dependence) — The Center on Antisocial Drug Dependence (CADD) data were funded by grants from the National Institute on Drug Abuse (P60 DA011015, R01 DA012845, R01 DA021913, R01 DA021905, R01 DA035804). For more information about this study, contact John K. Hewitt (john.hewitt@colorado.edu).

COGEND (Collaborative Genetic Study of Nicotine Dependence) — This research was supported by P01 CA089392, U01 HG004422, and R01 DA036583. Funding support for genotyping which was performed at the Johns Hopkins University Center for Inherited Disease Research was provided by the NIH "Genome-wide Association Studies in the Genes and Environment Initiative" (U01HG004438) and the NIH contract "High throughput genotyping for studying the genetic contributions to human disease" (HHSN268200782096C). We also acknowledge funding from R01 DA026911, and we thank Weimin Duan for analytic assistance. For more information about this study, contact Laura J. Bierut (laura@wustl.edu).

COPDGene (Genetics of Chronic Obstructive Pulmonary Disease) — The COPDGene® project was supported by award numbers R01 HL089897, R01 HL089856 and U01 HL089897-06A from the NHLBI. Research reported in this publication was also supported by the NHLBI award number K01 HL125858. The COPDGene® project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, Siemens, Sunovion, and GlaxoSmithKline. The authors acknowledge investigators of the COPDGene® project core units. For more information about this study, see http://www.copdgene.org/.

deCODE (deCODE Genetics/AMGEN, Inc.) — The authors are thankful to the Icelandic participants and staff at the Patient Recruitment Center. The work at deCODE genetics / Amgen was supported in part by the National Institute of Drug Abuse (NIDA grants, R01-DA017932 and R01-DA034076). For more information about this study, email info@decode.is.

EGCUT (Estonian Genome Center) — The EGCUT studies were financed by Estonian Government (grants IUT20-60 and IUT24-6) and by European Commission through the European Regional Development Fund in the frame of grant Estonian Center of Genomics/Roadmap II (project

No. 2014-2020.4.01.16-0125) and grant GENTRANSMED (Project No. 2014-2020.4.01.15-0012) and through H2020 grant no 692145 (ePerMed). For more information, please contact Tõnu Esko (tonu.esko@ut.ee).

eMERGE (Electronic Medical Records and Genomics) — Samples and associated genotype and phenotype data used in this study were provided by the Mayo Clinic. Funding support for the Mayo Clinic was provided through a cooperative agreement with the National Human Genome Research Institute (NHGRI), Grant #: UOIHG004599; and by grant HL75794 from the National Heart Lung and Blood Institute (NHLBI). Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000203.v1.p1. Funding support for the Personalized Medicine Research Project (PMRP) was provided through a cooperative agreement (U01HG004608) with the National Human Genome Research Institute (NHGRI), with additional funding from the National Institute for General Medical Sciences (NIGMS) The samples used for PMRP analyses were obtained with funding from Marshfield Clinic, Health Resources Service Administration Office of Rural Health Policy grant number D1A RH00025, and Wisconsin Department of Commerce Technology Development Fund contract number TDF FYO10718. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000360.v3.p1.

Finntwin & NAG-FIN (Finnish Twin Cohort) — The Finnish Twin Cohort/Nicotine Addiction Genetics-Finland study was supported by Academy of Finland (grants # 213506, 129680 to JK), NIH DA12854 (Pamela A F Madden), Global Research Award for Nicotine Dependence / Pfizer Inc. (JK), Wellcome Trust Sanger Institute, UK and the European Community's Seventh Framework Programme ENGAGE Consortium (HEALTH-F4-2007- 201413). In Finntwin12, support for data collection and genotyping has come from National Institute of Alcohol Abuse and Alcoholism (grants AA-12502, AA-00145, and AA-09203 to RJR and AA15416 and K02AA018755 to Danielle M Dick), the Academy of Finland (grants 100499, 205585, 118555, 141054 and 264146 to JK) & Wellcome Trust Sanger Institute, UK. JK has been supported by the Academy of Finland (grants # 265240 & 263278) and the Sigrid Juselius Foundation. For more information about this study, contact Jaakko Kaprio (jaakko.kaprio@helsinki.fi).

FHS (Framingham Heart Study) — The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195 and HHSN268201500001I). This manuscript was not prepared in

collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000007.v28.p10. Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL- 64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. Funding for Affymetrix genotyping of the FHS Omni cohorts was provided by Intramural NHLBI funds from Andrew D. Johnson and Christopher J. O'Donnell.

GERA (Genetic Epidemiology Research in Adult Health and Aging) — GERA acknowledges the following funding support: RC2 AG036607 from the National Institutes of Health (NIH); R01 EY027004 from the National Eye Institute (NEI); R21 AA021223 from the National Institute of Alcohol Abuse and Alcoholism (NIAAA); Kaiser Permanente Research Program on Genes, Environment, and Health, funded by the Wayne and Gladys Valley Foundation, The Ellison Medical Foundation, the Robert Wood Johnson Foundation, and Kaiser Permanente National and Northern California Community Benefit Programs. All study procedures were approved by the Institutional Review Board of the Kaiser Foundation Research Institute. Researchers interested in this data should contact rpgeh-collab@kp.org.

GfG (Genes for Good) — The Genes for Good study is funded through discretionary funds, provided to Dr. Gonçalo Abecasis by the University of Michigan. The authors sincerely thank all study participants for their time and dedication, as well as the hard-working Genes for Good administrative staff and colleagues at the UM DNA Sequencing Core. For more information about this study, see https://genesforgood.sph.umich.edu/ or contact the study directly at genesforgood@umich.edu.

HUNT (The Nord-Trøndelag Health Study) — HUNT is a collaboration between HUNT Research Centre (Faculty of Medicine, NTNU, Norwegian University of Science and Technology), the Nord-Trøndelag County Council, the Central Norway Health Authority, and the Norwegian Institute of Public Health. Bendik Slagsvold Winsvold is funded by the University of Oslo and the South Eastern Norway Health Trust. Anne Heidi Skogholt, Maiken Elvestad Gabrielsen, and Kristian Hveem are funded by NTNU, the Central Norway Health Trust, and the K.J. Jebsen Foundation. For more information about this study, email hunt@medisin.ntnu.no.

HRS (Health and Retirement Study) — HRS is supported by the National Institute on Aging (NIA U01AG009740). The genotyping was funded separately by the National Institute on Aging (RC2 AG036495, RC4 AG039029). Our genotyping was conducted by the NIH Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotyping quality control and final preparation of the data were performed by the University of Michigan School of Public Health. See the HRS website (http://hrsonline.isr.umich.edu/gwas) for details.

MCTFR (Minnesota Center for Twin and Family Research) — MCTFR was supported in part by USPHS Grants from the National Institute on Alcohol Abuse and Alcoholism (R01 AA09367 and R01 AA11886) and from the National Institute on Drug Abuse (R01 DA05147, R01 DA13240, and

U01 DA024417. GWAS and phenotypic data for MCTFR subjects who provided consent to place their data in a public repository are deposited into the database of Genotypes and Phenotypes (dbGaP, www.ncbi.nlm.nih.gov/gap) under phs000620. For further information, please contact Matt McGue (mcgue001@umn.edu).

MESA (Multi-Ethnic Study of Atherosclerosis) — MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-RR-025005, and UL1-TR-000040. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. The MESA CARe data used for the analyses described in this manuscript were obtained through dbGaP phs000209.v13.p1. Funding for CARe genotyping was provided by NHLBI Contract N01-HC-65226. This study is part of the NHLBI Grand Opportunity Exome Sequencing Project (GO-ESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). HeartGO gratefully acknowledges the following groups and individuals who provided biological samples or data for this study. DNA samples and phenotypic data were obtained from the following studies supported by the NHLBI: the Atherosclerosis Risk in Communities (ARIC) study, the Coronary Artery Risk Development in Young Adults (CARDIA) study, Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Jackson Heart Study (JHS) and the Multi-Ethnic Study of Atherosclerosis (MESA). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000209.v13.p1.

METSIM (Metabolic Syndrome in Men) — METSIM was supported by in part by NIH grant number U01 DK062370 to Michael Boehnke. For information about the METSIM study, contact Markku Laakso at markku.laakso@kuh.fi.

NESCOG (Netherlands Study on Cognition, Environment and Genes) — This research was part of Science Live, the innovative research program of science center NEMO that enables scientists to carry out real, publishable, peer-reviewed research using NEMO visitors as volunteers. For more information about this study, email info@nescog.nl.

NHS, NHS2, and HPFS (Nurses' Health Study, Nurses' Health Study II, and Health Professionals' Follow-up Study) — The contributions from the Nurses' Health Study, Nurses Health Study II, and Health Professionals' Follow-up Study were supported by the National Institute of Health grants P01CA87969, P01CA055075, P01DK070756, U01HG004728, UM1CA186107, UM1CA176726, UM1CA167552, R01CA49449, R01CA50385, R01CA67262, R01HL034594, R01HL088521, R01HL35464, R01EY015473, R01EY022305, P30EY014104, R03DC013373, and

R03CA165131. For information about these studies, contact Peter Kraft pkraft@hsph.harvard.edu) or Marilyn C. Cornelis (marilyn.cornelis@northwestern.edu).

NINDS SiGN (The National Institute of Neurological Disorders and Stroke Genetics Network) — The NINDS International Stroke Genetics Consortium Study dataset was funded by the National Institute of Neurological Disorders and Stroke Cooperative Agreement Award 1U01NS069208 (Steven Kittner). The dataset used for the analyses described in this manuscript was obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000615.v1.p1.

NTR (Netherlands Twin Register) — NTR would like to thank all the twins and family members for their participation. This work was supported by the Netherlands Organization for Scientific Research (NWO: MagW/ZonMW grants 904-61-090, 985-10-002,904-61-193,480-04-004, 400-05-717, Addiction-31160008 Middelgroot-911-09-032, Spinozapremie 56-464-14192), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI –NL, 184.021.007), the VU University's Institute for Health and Care Research (EMGO+) and Neuroscience Campus Amsterdam (NCA), the European Science Council (ERC Advanced, 230374), the Avera Institute for Human Genetics, Sioux Falls, South Dakota (USA) and the National Institutes of Health (NIH, R01D0042157-01A). Part of the genotyping was funded by the Genetic Association Information Network (GAIN) of the Foundation for the US National Institutes of Health (NIMH, MH081802) and by the Grand Opportunity grants 1RC2MH089951-01 and 1RC2 MH089995-01 from the NIMH. Part of the analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org), which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003), the Dutch Brain Foundation, and the department of Psychology and Education of the VU University Amsterdam. For access to NTR results, please contact Dorret I. Boomsma at di.boomsma@vu.nl.

OZALC (Australian Twin-Family Studies on Nicotine and Alcohol Genetics) — OZALC acknowledges the work over many years of staff of the Genetic Epidemiology group at QIMR Berghofer Medical Research Institute (formerly the Queensland Institute of Medical Research) in managing the studies which generated the data used in this analysis. We also acknowledge and appreciate the willingness of study participants to complete multiple, and sometimes lengthy, questionnaires and interviews. Many of the participants were contacted originally through the Australian Twin Registry. Funding for the original studies in which information on alcohol use and smoking status was obtained came from the US National Institutes of Health (AA07535, AA07728, AA11998, AA13320, AA13321, AA14041, AA17688, DA012854 and DA019951); the Australian National Health and Medical Research Council (241944, 339462, 389927, 389875, 389891, 389892, 389938, 442915, 442981, 496739, 552485 and 552498); and the Australian Research Council (A7960034, A79906588, A79801419, DP0770096, DP0212016 and DP0343921). Researchers interested in using this dataset may apply for access using dbGaP Study Accession phs000181.v1.p1.

SardiNIA (SardiNIA project) — We thank all the volunteers who generously participated in this study and made this research possible. This research was supported by National Human

Genome Research Institute grants HG005581, HG005552, HG006513, HG007022 and HG007089; by National Heart, Lung, and Blood Institute grant HL117626; by the Intramural Research Program of the US National Institutes of Health, National Institute on Aging, contracts N01-AG-1-2109 and HHSN271201100005C; by Sardinian Autonomous Region (L.R. 7/2009) grant cRP3-154; by the PB05 InterOmics MIUR Flagship Project; by grant FaReBio2011 'Farmaci e Reti Biotecnologiche di Qualità'. For information about this study, contact David Schlessinger (SchlessingerD@mail.nih.gov) or Francesco Cucca (fcucca@irgb.cnr.it).

UK Biobank — This research has been conducted using the UK Biobank Resource under Application Number 16651. Informed consent was obtained from UK Biobank subjects.

WHI (Women's Health Initiative) — The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. Personal funding for Sean P. David from National Institute on Minority Health and Health Disparities grant U54-MD010724. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whi.org/researchers/Documents_Write_a_Paper/WHI_Investigator_Short_List.pdf. For more information about this study, please contact nm9o@nih.gov.

Individual acknowledgements

Scott Vrieze, Mengzhen Liu, Gargi Datta, Hannah Young, and the Vrieze Lab were partially supported by the National Institutes of Health grant numbers R01DA042755, R01AA023974, R01DA037904, R21DA040177, U01DA041120, and R01HG008983. Dajiang J. Liu was partially support by grant R01HG008983 from the National Genome Research Institute of the National Institutes of Health. Both Dajiang J. Liu and Yu Jiang were partially supported by grants R21DA040177 and R01DA037904 from the National Institute of Drug Abuse of the National Institutes of Health. Manolis Kellis and Yue Li were partially supported by the National Institutes of Health grant numbers R01HG008155, R01MH109978, U01HG009088, and R01DA037904. David M. Brazel was supported by grant number 5T3DA017637-13 from the National Institute on Drug Abuse (NIDA) at the National Institutes of Health. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of NIDA. Jerry A. Stitzel was supported, in part, by NIH grant UH2DA040142. Marissa A. Ehringer was supported by the National Institute of Alcohol Abuse and Alcoholism (NIAAA) grant number R01AA017889. Matthew C. Keller was supported by the National Institute of Mental Health (NIMH) grant number R01MH100141. Robbee Wedow was generously supported by the National Science Foundation's Graduate Research Fellowship Program (DGE 1144083). Any opinion, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

13. References

1. Furberg H, Kim Y, Dackor J, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genet.* 2010;42(5):441-U134.
2. Jacobs DR, Adachi H, Mulder I, et al. Cigarette smoking and mortality risk - Twenty-five-year follow-up of the seven countries study. *Arch Intern Med.* 1999;159(7):733-740.
3. Jacobs DR, Adachi H, Mulder I, et al. Cigarette smoking and mortality risk 25-year follow-up of the seven countries study. *Circulation.* 1998;98(17):582-582.
4. Mclaughlin JK, Dietz MS, Mehl ES, Blot WJ. Reliability of Surrogate Information on Cigarette-Smoking by Type of Informant. *Am J Epidemiol.* 1987;126(1):144-146.
5. Williams GD, Aitken SS, Malin H. Reliability of self-reported alcohol consumption in a general population survey. *J Stud Alcohol.* 1985;46(3):223-227.
6. Brigham J, Lessov-Schlaggar CN, Javitz HS, McElroy M, Krasnow R, Swan GE. Reliability of adult retrospective recall of lifetime tobacco use. *Nicotine Tob Res.* 2008;10(2):287-299.
7. Timofeeva MN, McKay JD, Smith GD, et al. Genetic Polymorphisms in 15q25 and 19q13 Loci, Cotinine Levels, and Risk of Lung Cancer in EPIC. *Cancer Epidem Biomar.* 2011;20(10):2250-2261.
8. Krebs NM, Chen A, Zhu JJ, et al. Comparison of Puff Volume With Cigarettes per Day in Predicting Nicotine Uptake Among Daily Smokers. *Am J Epidemiol.* 2016;184(1):48-57.
9. Vartiainen E, Seppala T, Lillsunde P, Puska P. Validation of self reported smoking by serum cotinine measurement in a community-based study. *J Epidemiol Commun H.* 2002;56(3):167-170.
10. Hicks BM, Schalet BD, Malone SM, Iacono WG, McGue M. Psychometric and genetic architecture of substance use disorder and behavioral disinhibition measures for gene association studies. *Behav Genet.* 2011;41(4):459-475.
11. Vrieze SI, Iacono WG, McGue M. Confluence of genes, environment, development, and behavior in a post Genome-Wide Association Study world. *Dev Psychopathol.* 2012;24(4):1195-1214.
12. Vrieze SI, Hicks BM, Iacono WG, McGue M. Decline in genetic influence on the co-occurrence of alcohol, marijuana, and nicotine dependence symptoms from age 14 to 29. *Am J Psychiatry.* 2012;169(10):1073-1081.
13. He L, Pitkaniemi J, Heikkila K, et al. Genome-wide time-to-event analysis on smoking progression stages in a family-based study. *Brain Behav.* 2016;6(5).
14. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016.
15. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* 2016.
16. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genet.* 2012;44(8):955-+.
17. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics.* 2016;32(9):1423-1426.
18. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet.* 2010;42(4):348-354.
19. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* 2006;38(8):904-909.
20. Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999;55:997-1004.
21. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity.* 2005;95(3):221-227.
22. Gao XY, Becker LC, Becker DM, Starmer JD, Province MA. Avoiding the High Bonferroni Penalty in Genome-Wide Association Studies. *Genet Epidemiol.* 2010;34(1):100-105.
23. Chen ZX, Liu QZ. A New Approach to Account for the Correlations among Single Nucleotide Polymorphisms in Genome-Wide Association Studies. *Hum Hered.* 2011;72(1):1-9.
24. Jiang Y, Chen S, McGuire D, et al. Proper Conditional Analysis in the Presence of Missing Data Identified Novel Independently Associated Low Frequency Variants in Nicotine Dependence Genes. *PLoS Genetics.* 2018.

25. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet.* 2012;44(4):369-375, S361-363.
26. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genet.* 2015;47(3):291-+.
27. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 2010;26(18):2336-2337.
28. Turley P, Walters RK, Maghzian O, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genet.* 2018;50(2):229-+.
29. Kotov R, Krueger RF, Watson D, et al. The Hierarchical Taxonomy of Psychopathology (HiTOP): A Dimensional Alternative to Traditional Nosologies. *J Abnorm Psychol.* 2017;126(4):454-477.
30. Krueger RF, Caspi A, Moffitt TE, Silva PA. The structure and stability of common mental disorders (DSM-III-R): a longitudinal-epidemiological study. *J Abnorm Psychol.* 1998;107(2):216-227.
31. Krueger RF, Hicks BM, Patrick CJ, Carlson SR, Iacono WG, McGue M. Etiologic connections among substance dependence, antisocial behavior, and personality: Modeling the externalizing spectrum. *J Abnorm Psychol.* 2002;111(3):411-424.
32. Young SE, Stallings MC, Corley RP, Krauter KS, Hewitt JK. Genetic and environmental influences on behavioral disinhibition. *Amercan Journal of Medical Genetics.* 2000;96(5):684-695.
33. Martin NG, Eaves LJ. Genetical Analysis of Covariance Structure. *Heredity.* 1977;38(Feb):79-95.
34. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genet.* 2015;47(11):1236-+.
35. Grotzinger AD, Rhemtulla M, de Vlaming R, et al. Genomic SEM Provides Insights into the Multivariate Genetic Architecture of Complex Traits. *bioRxiv.* 2018.
36. Hu LT, Bentler PM. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Struct Equ Modeling.* 1999;6(1):1-55.
37. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol Methods.* 2012;17(2):228-243.
38. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genet.* 2015;47(11):1228-+.
39. Wolf PA, Dagostino RB, Kannel WB, Bonita R, Belanger AJ. Cigarette-Smoking as a Risk Factor for Stroke - the Framingham-Study. *Jama-J Am Med Assoc.* 1988;259(7):1025-1029.
40. Anstey KJ, von Sanden C, Salim A, O'Kearney R. Smoking as a risk factor for dementia and cognitive decline: a meta-analysis of prospective studies. *Am J Epidemiol.* 2007;166(4):367-378.
41. Bagnardi V, Blangiardo M, La Vecchia C, Corrao G. A meta-analysis of alcohol drinking and cancer risk. *Brit J Cancer.* 2001;85(11):1700-1705.
42. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *Lancet.* 2007;370(9589):765-773.
43. Modig K, Silventoinen K, Tynelius P, Kaprio J, Rasmussen F. Genetics of the association between intelligence and nicotine dependence: a study of male Swedish twins. *Addiction.* 2011;106(5):995-1002.
44. Broms U, Silventoinen K, Lahelma E, Koskenvuo M, Kaprio J. Smoking cessation by socioeconomic status and marital status: The contribution of smoking behavior and family background. *Nicotine Tob Res.* 2004;6(3):447-455.
45. Wennerstad KM, Silventoinen K, Tynelius P, Bergman L, Kaprio J, Rasmussen F. Associations between IQ and cigarette smoking among Swedish male twins. *Soc Sci Med.* 2010;70(4):575-581.
46. Zheng J, Erzurumluoglu AM, Elsworth BL, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics.* 2017;33(2):272-279.
47. Yang JA, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88(1):76-82.
48. Sullivan PF, Kendler KS. The genetic epidemiology of smoking. *Nicotine Tob Res.* 1999;1 Suppl 2:S51-57; discussion S69-70.

49. Maes HH, Sullivan PF, Bulik CM, et al. A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence. *Psychological Medicine.* 2004;34(7):1251-1261.
50. Han C, McGue M, Iacono WG. Lifetime tobacco, alcohol and other substance use in adolescent Minnesota twins: Univariate and multivariate behavioural genetic analyses. *Addiction.* 1999;94(7):981-993.
51. Heath AC, Jardine R, Martin NG. Interactive effects of genotype and social environment on alcohol consumption in female twins. *J Stud Alcohol.* 1989;50(1):38-48.
52. Vilhjalmsson BJ, Yang J, Finucane HK, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet.* 2015;97(4):576-592.
53. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JWR, Weir DR. Cohort Profile: the Health and Retirement Study (HRS). *Int J Epidemiol.* 2014;43(2):576-585.
54. Harris KM, Halpern CT, Haberstick BC, Smolen A. The National Longitudinal Study of Adolescent Health (Add Health) Sibling Pairs Data. *Twin Research and Human Genetics.* 2013;16(1):391-398.
55. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559-575.
56. Hindorff LA, MacArthur J, Morales J, et al. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies.
57. Gusev A, Lee SH, Trynka G, et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *American Journal of Human Genetics.* 2014;95(5):535-552.
58. Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518(7539):317-330.
59. Trynka G, Sandor C, Han B, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet.* 2013;45(2):124-130.
60. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet.* 2014;94(4):559-573.
61. Kichaev G, Pasaniuc B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am J Hum Genet.* 2015;97(2):260-271.
62. Li Y, Davila-Velderrain J, Kellis M. A probabilistic framework to dissect functional cell-type-specific regulatory elements and risk loci underlying the genetics of complex traits. *BioRxiv.* 2017;059345.
63. Zhan X, Liu DJ. SEQMINER: An R-Package to Facilitate the Functional Interpretation of Sequence-Based Associations. *Genet Epidemiol.* 2015;39(8):619-623.
64. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *Plos Comput Biol.* 2016;12(1).
65. Pers TH, Karjalainen JM, Chan Y, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications.* 2015;6.
66. Milton RC, Clemons TE, Kurinij N, Sperduto RD, Grp AREDSR. Risk factors for the incidence of advanced age-related macular degeneration in the age-elated eye disease study (AREDS) - AREDS report no.19. *Ophthalmology.* 2005;112(4):533-539.
67. Kang JH, Pasquale LR, Rosner BA, et al. Prospective study of cigarette smoking and the risk of primary open-angle glaucoma. *Arch Ophthalmol-Chic.* 2003;121(12):1762-1768.
68. Cumming RG, Mitchell P. Alcohol, smoking, and cataracts - The Blue Mountains Eye Study. *Arch Ophthalmol-Chic.* 1997;115(10):1296-1303.
69. Frey BJ, Dueck D. Clustering by passing messages between data points. *Science.* 2007;315(5814):972-976.