# Beyond SNP Heritability: Polygenicity and Discoverability of Phenotypes Estimated with a Univariate Gaussian Mixture Model

Dominic Holland[a,b,*], Oleksandr Frei[d], Rahul Desikan[c], Chun-Chieh Fan[a,e,f], Alexey A. Shadrin[d], Olav B. Smeland[a,d,g], V. S. Sundar[a,f], Paul Thompson[h], Ole A. Andreassen[d,g], Anders M. Dale[a,b,f,i]

[a]*Center for Multimodal Imaging and Genetics, University of California at San Diego, La Jolla, CA 92037, USA,*
[b]*Department of Neurosciences, University of California, San Diego, La Jolla, CA 92093, USA,*
[c]*Department of Radiology, University of California, San Francisco, San Francisco, CA 94158, USA,*
[d]*NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo 0424 Oslo, Norway,*
[e]*Department of Cognitive Sciences, University of California at San Diego, La Jolla, CA 92093, USA,*
[f]*Department of Radiology, University of California, San Diego, La Jolla, CA 92093, USA,*
[g]*Division of Mental Health and Addiction, Oslo University Hospital, 0407 Oslo, Norway,*
[h]*Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA,*
[i]*Department of Psychiatry, University of California, San Diego, La Jolla, CA 92093, USA,*

## Abstract

Estimating the polygenicity (proportion of causally associated single nucleotide polymorphisms (SNPs)) and discoverability (effect size variance) of causal SNPs for human traits is currently of considerable interest. SNP-heritability is proportional to the product of these quantities. We present a basic model, using detailed linkage disequilibrium structure from an extensive reference panel, to estimate these quantities from genome-wide association studies (GWAS) summary statistics. We apply the model to diverse phenotypes and validate the implementation with simulations. We find model polygenicities ranging from $\simeq 2 \times 10^{-5}$ to $\simeq 4 \times 10^{-3}$, with discoverabilities similarly ranging over two orders of magnitude. A power analysis allows us to estimate the proportions of phenotypic variance explained additively by causal SNPs reaching genome-wide significance at current sample sizes, and map out sample sizes required to explain larger portions of additive SNP heritability. The model also allows for estimating residual inflation (or deflation from over-correcting of z-scores), and assessing compatibility of replication and discovery GWAS summary statistics.

*Keywords:* GWAS, Polygenicity, Discoverability, Heritability, Causal SNPs, Effect size, Linkage Disequlilbrium

## Author Summary

There are ∼10 million common variants in the genome of humans with European ancestry. For any particular phenotype a number of these variants will have some causal effect. It is of great interest to be able to quantify the number of these causal variants and the strength of their effect on the phenotype.

Genome wide association studies (GWAS) produce very noisy summary statistics for the association between subsets of common variants and phenotypes. For any phenotype, these statistics collectively are difficult to interpret, but buried within them is the true landscape of causal effects. In this work, we posit a probability distribution for the causal effects, and assess its validity using simulations. Using a detailed reference panel of ∼11 million common variants – among which only a small fraction are likely to be causal, but allowing for non-causal variants to show an association with the phenotype due to correlation with causal variants – we implement an exact procedure for estimating the number of causal variants and their mean strength of association with the phenotype. We find that, across different phenotypes, both these quantities – whose product allows for lower bound estimates of heritability – vary by orders of magnitude.

## INTRODUCTION

The genetic components of complex human traits and diseases arise from hundreds to likely many thousands of single nucleotide polymorphisms (SNPs) [1], most of which have weak effects. As sample sizes increase, more of the associated SNPs are identifiable (they reach genome-wide significance), though power for discovery varies widely across phenotypes. Of particular interest are estimating the proportion of common SNPs from a reference panel (polygenicity) involved in any particular phenotype; their effective strength of association (discoverability, or causal effect size variance); the proportion of variation in susceptibility, or phenotypic variation, captured additively by all common causal SNPs (approximately, the narrow sense heritability), and the fraction of that captured by genome-wide significant SNPs – all of which are active areas of research [2, 3, 4, 5, 6, 7, 8, 9]. The effects of population structure [10], combined with high polygenicity and linkage disequilibrium (LD), leading to spurious degrees

*Corresponding author:
email: dominic.holland@gmail.com
Phone: 858-822-1776
Fax: 858-534-1078

of SNP association, or inflation, considerably complicate matters, and are also areas of much focus [11, 12, 13]. Despite these challenges, there have been recent significant advances in the development of mathematical models of polygenic architecture based on GWAS [14, 15]. One of the advantages of these models is that they can be used for power estimation in human phenotypes, enabling prediction of the capabilities of future GWAS.

Here, in a unified approach explicitly taking into account LD, we present a model relying on genome-wide association studies (GWAS) summary statistics (z-scores for SNP associations with a phenotype [16]) to estimate polygenicity ($\pi_1$, the proportion of causal variants in the underlying reference panel of approximately 11 million SNPs from a sample size of 503) and discoverability ($\sigma^2_\beta$, the causal effect size variance), as well as elevation of z-scores due to any residual inflation of the z-scores arising from variance distortion ($\sigma^2_0$, which for example can be induced by cryptic relatedness), which remains a concern in large-scale studies [10]. We estimate $\pi_1$, $\sigma^2_\beta$, and $\sigma^2_0$, by postulating a z-score probability distribution function (pdf) that explicitly depends on them, and fitting it to the actual distribution of GWAS z-scores.

Estimates of polygenicity and discoverability allow one to estimate compound quantities, like narrow-sense heritability captured by the SNPs [17]; to predict the power of larger-scale GWAS to discover genome-wide significant loci; and to understand why some phenotypes have higher power for SNP discovery and proportion of heritability explained than other phenotypes.

In previous work [18] we presented a related model that treated the overall effects of LD on z-scores in an approximate way. Here we take the details of LD explicitly into consideration, resulting in a conceptually more basic model to predict the distribution of z-scores. We apply the model to multiple phenotype datasets, in each case estimating the three model parameters and auxiliary quantities, including the overall inflation factor $\lambda$, (traditionally referred to as genomic control [19]) and narrow sense heritability, $h^2$. We also perform extensive simulations on genotypes with realistic LD structure in order to validate the interpretation of the model parameters. A discussion of the relation of the present paper to other work is provided in the Supplementary Material.

## METHODS

### Overview

Our basic model is a simple postulate for the distribution of causal effects (denoted $\beta$ below) [20]. Our model assumes that only a fraction of all SNPs are in some sense causally related to any given phenotype. We work with a reference panel of approximately 11 million SNPs with 503 samples, and assume that all common causal SNPs (minor allele frequency (MAF) $> 0.002$) are contained in it. Any given GWAS will have z-scores for a subset of these reference SNPs (we use the term "typed" below to refer to GWAS SNPs with z-scores, whether they were directly genotyped or their genotype was imputed). When a z-score partially involves a latent causal component (i.e., not pure noise), we assume that it arises through LD with neighboring causal SNPs, or that it itself is causal.

We construct a pdf for z-scores that directly follows from the underlying distribution of effects. For any given typed SNP's z-score, it is dependent on the other SNPs the focal SNP is in LD with (SNPs that are "tagged" by the focal SNP), taking into account their LD with the focal SNP and their heterozygosity (i.e., it depends not just on the focal typed SNP's total LD and heterozygosity, but also on the distribution of neighboring reference SNPs in LD with it and their heterozygosities). We present two ways of constructing the model pdf for z-scores, using multinomial expansion, and using convolution. The former is perhaps more intuitive, but the latter is more numerically tractable, yielding an exact solution, and is used here to obtain all reported results. The problem then is finding the three model parameters that give a maximum likelihood best fit for the model's prediction of the distribution of z-scores to the actual distribution of z-scores. Because we are fitting three parameters typically using $\gtrsim 10^6$ data points, it is appropriate to incorporate some data reduction to facilitate the computations. To that end, we bin the data (z-scores) into a $10\times10$ grid of heterozygosity-by-total LD (having tested different grid sizes to ensure convergence of results). Also, when building the LD and heterozygosity structures of reference SNPs, we fine-grained the LD range ($0 \leq r^2 \leq 1$), again ensuring that bins were small enough that results were well converged. To fit the model to the data we bin the z-scores (within each heterozygosity/total LD window) and calculate the multinomial probability for having the actual distribution of z-scores (numbers of z-scores in the z-score bins) given the model pdf for the distribution of z-scores, and adjusting the model parameters using a multidimensional unconstrained nonlinear minimization (Nelder-Mead), so as to maximize the likelihood of the data, given the parameters.

A visual summary of the predicted and actual distribution of z-scores is obtained by making quantile-quantile plots showing, for a wide range of significance thresholds going well beyond genome-wide significance, the proportion (x-axis) of typed SNPs exceeding any given threshold (y-axis) in the range. It is important also to assess the quantile-quantile sub-plots for SNPs in the heterozygosity-by-total LD grid elements (see Supplementary Material).

With the pdf in hand, various quantities can be calculated: the number of causal SNPs; the expected genetic effect (denoted $\delta$ below, where $\delta^2$ is the non-centrality parameter of a Chi-squared distribution) at the current sample size for a typed SNP given the SNP's z-score and its full LD and heterozygosity structure; the estimated SNP heritability, $h^2_{SNP}$ (excluding contributions from rare reference SNPs, i.e., with MAF<0.2%); and the sample size required to explain any percentage of that with genome-

wide significant SNPs. The model can easily be extended using a more complex distribution for the underlying $\beta$'s, with multiple-component mixtures for small and large effects, and incorporating selection pressure through both heterozygosity dependence on effect sizes and linkage disequilibrium dependence on the prior probability of a SNP's being causal – issues we will address in future work.

**The Model: Probability Distribution for z-Scores**

To establish notation, we consider a bi-allelic genetic variant, $i$, and let $\beta_i$ denote the effect size of allele substitution of that variant on a given quantitative trait. We assume a simple additive generative model (simple linear regression, ignoring covariates) relating genotype to phenotype [18, 21]. That is, assume a linear vector equation (no summation over repeated indices)

$$y = g_i\beta_i + e_i \tag{1}$$

for phenotype vector $y$ over $N$ samples (mean-centered and normalized to unit variance), mean-centered genotype vector $g_i$ for the $i^{\text{th}}$ of $n$ SNPs (vector over samples of the additively coded number of reference alleles for the $i^{\text{th}}$ variant), true fixed effect $\beta_i$ (regression coefficient) for the SNP, and residual vector $e_i$ containing the effects of all the other causal SNPs, the independent random environmental component, and random error. Variants with non-zero fixed effect $\beta_i$ are said to be "causal". For SNP $i$, the estimated simple linear regression coefficient is

$$\hat{\beta}_i = g_i^T y/(g_i^T g_i) = \text{cov}(g_i, y)/\text{var}(g_i), \tag{2}$$

where $T$ denotes transpose and $g_i^T g_i/N = \text{var}(g_i) = H_i$ is the SNP's heterozygosity (frequency of the heterozygous genotype): $H_i = 2p_i(1-p_i)$ where $p_i$ is the frequency of either of the SNP's alleles.

Consistent with the work of others [11, 15], we assume the causal SNPs are distributed randomly throughout the genome (an assumption that can be relaxed when explicitly considering different SNP categories, but that in the main is consistent with the additive variation explained by a given part of the genome being proportional to the length of DNA [22]). In a Bayesian approach, we assume that the parameter $\beta$ for a SNP has a distribution (in that specific sense, this is similar to a random effects model), representing subjective information on $\beta$, not a distribution across tangible populations [23]. Specifically, we posit a normal distribution for $\beta$ with variance given by a constant, $\sigma_\beta^2$:

$$\beta \sim \mathcal{N}(0, \sigma_\beta^2). \tag{3}$$

This is also how the $\beta$ are distributed across the set of causal SNPs. Therefore, taking into account all SNPs (the remaining ones are all null by definition), this is equivalent to the two-component Gaussian mixture model we originally proposed [20]

$$\beta \sim \pi_1\mathcal{N}(0, \sigma_\beta^2) + (1-\pi_1)\mathcal{N}(0, 0) \tag{4}$$

where $\mathcal{N}(0, 0)$ is the Dirac delta function, so that considering all SNPs, the net variance is $\text{var}(\beta) = \pi_1\sigma_\beta^2$. If there is no LD (and assuming no source of spurious inflation), the association z-score for a SNP with heterozygosity $H$ can be decomposed into a fixed effect $\delta$ and a residual random environment and error term, $\epsilon \sim \mathcal{N}(0, 1)$, which is assumed to be independent of $\delta$ [18]:

$$z = \delta + \epsilon \tag{5}$$

with

$$\delta = \sqrt{NH}\beta \tag{6}$$

so that

$$\begin{aligned} \text{var}(z) &= \text{var}(\delta) + \text{var}(\epsilon) \\ &\equiv \sigma^2 + 1 \end{aligned} \tag{7}$$

where

$$\sigma^2 = \sigma_\beta^2 NH. \tag{8}$$

By construction, under null, i.e., when there is no genetic effect, $\delta = 0$, so that $\text{var}(\epsilon) = 1$.

If there is no source of variance distortion in the sample, but there is a source of bias in the summary statistics for a subset of markers (e.g., the sample is composed of two or more subpopulations with different allele frequencies for a subset of markers – pure population stratification in the sample [24]), the marginal distribution of an individual's genotype at any of those markers will be inflated. The squared z-score for such a marker will then follow a non-central Chi-square distribution (with one degree of freedom); the non-centrality parameter will contain the causal genetic effect, if any, but biased up or down (confounding or loss of power, depending on the relative sign of the genetic effect and the SNP-specific bias term). The effect of bias shifts, arising for example due to stratification, is nontrivial, and currently not explicitly in our model; it is usually accounted for using standard methods [25].

Variance distortion in the distribution of z-scores can arise from cryptic relatedness in the sample (drawn from a population mixture with at least one subpopulation with identical-by-descent marker alleles, but no population stratification) [19]. If $z_u$ denotes the uninflated z-scores, then the inflated z-scores are

$$z = \sigma_0 z_u, \tag{9}$$

where $\sigma_0 \geq 1$ characterizes the inflation. Thus, from Eq. 7, in the presence of inflation in the form of variance distortion

$$\begin{aligned} \text{var}(z) &= \sigma_0^2(\sigma^2 + 1) \\ &\equiv \tilde{\sigma}^2 + \sigma_0^2 \\ &\equiv \tilde{\sigma}_\beta^2 NH + \sigma_0^2 \end{aligned} \tag{10}$$

where $\tilde{\sigma}_\beta^2 \equiv \sigma_0^2\sigma_\beta^2$, so that $\text{var}(\delta) = \tilde{\sigma}^2 \equiv \tilde{\sigma}_\beta^2 NH$ and $\epsilon \sim \mathcal{N}(0, \sigma_0^2)$. In the presence of variance distortion one is

dealing with inflated random variables $\tilde{\beta} \sim \mathcal{N}(0, \tilde{\sigma}^2_\beta)$, but we will drop the tilde on the $\beta$'s in what follows.

Since variance distortion leads to scaled z-scores [19], then, allowing for this effect in some of the extremely large data sets, we can assess the ability of the model to detect this inflation by artificially inflating the z-scores (Eq. 9), and checking that the inflated $\hat{\sigma}^2_0$ is estimated correctly while the other parameter estimates remain unchanged.

Implicit in Eq. 8 is approximating the denominator, $1 - q^2$, of the $\chi^2$ statistic non-centrality parameter to be 1, where $q^2$ is the proportion of phenotypic variance explained by the causal variant, i.e., $q \equiv \sqrt{H}\beta$. So a more correct $\delta$ is

$$\delta = \sqrt{N}q/\sqrt{1-q^2}. \tag{11}$$

Taylor expanding in $q$ and then taking the variance gives

$$\mathrm{var}(\delta) = \sigma^2_\beta NH[1 + (15/4)\sigma^4_\beta H^2 + O(\sigma^8_\beta H^4)]. \tag{12}$$

The additional terms will be vanishingly small and so do not contribute in a distributional sense; (quasi-) Mendelian or outlier genetic effects represent an extreme scenario where the model is not expected to be accurate, but SNPs for such traits are by definition easily detectable. So Eq. 8 remains valid for the polygenicity of complex traits.

Now consider the effects of LD on z-scores. The simple linear regression coefficient estimate for typed SNP $i$, $\hat{\beta}_i$, and hence the GWAS z-score, implicitly incorporates contributions due to LD with neighboring causal SNPs. (A typed SNP is a SNP with a z-score, imputed or otherwise; generally these will compose a smaller set than that available in reference panels like 1000 Genomes used here for calculating the LD structure of typed SNPs.) In Eq. 1, $e_i = \sum_{j \neq i} g_j \beta_j + \varepsilon$, where $g_j$ is the genotype vector for SNP $j$, $\beta_j$ is its true regression coefficient, and $\varepsilon$ is the independent true environmental and error residual vector (over the $N$ samples). Thus, explicitly including all causal true $\beta$'s, Eq. 2 becomes

$$\begin{aligned}\hat{\beta}_i &= \frac{\sum_j g_i^T g_j \beta_j}{NH_i} + \frac{g_i^T \varepsilon}{g_i^T g_i} \\ &\equiv \beta'_i + \varepsilon'_i \end{aligned} \tag{13}$$

(the sum over $j$ now includes SNP $i$ itself). This is the simple linear regression expansion of the estimated regression coefficient for SNP $i$ in terms of the independent latent (true) causal effects and the latent environmental (plus error) component; $\beta'_i$ is the effective simple linear regression expression for the true genetic effect of SNP $i$, with contributions from neighboring causal SNPs mediated by LD. Note that $g_i^T g_j/N$ is simply $\mathrm{cov}(g_i, g_j)$, the covariance between genotypes for SNPs $i$ and $j$. Since correlation is covariance normalized by the variances, $\beta'_i$ in Eq. 13 can be written as

$$\beta'_i = \sum_j \sqrt{\frac{H_j}{H_i}} r_{ij} \beta_j. \tag{14}$$

where $r_{ij}$ is the correlation between genotypes at reference SNP $j$ and typed SNP $i$. Then, from Eq. 5, the z-score for the typed SNP's association with the phenotype is given by:

$$\begin{aligned} z_i &= \sqrt{NH_i}\beta'_i + \epsilon_i \\ &= \sqrt{N}\sum_j \sqrt{H_j} r_{ij}\beta_j + \epsilon_i. \end{aligned} \tag{15}$$

We noted that in the absence of LD, the distribution of the residual in Eq. 5 is assumed to be univariate normal. But in the presence of LD (Eq. 15) there are induced correlations. Letting $\boldsymbol{\epsilon}$ denote the vector of residuals (with element $\epsilon_i$ for SNP $i$, $i = 1 \ldots n$), and $\boldsymbol{M}$ denote the (sparse) $n \times n$ LD-$r^2$ matrix, then, ignoring inflation, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{M})$ [26]. Since the genotypes of two unrelated individuals are marginally independent, this multivariate normal distribution for $\boldsymbol{\epsilon}$ is contingent on the summary statistics for all SNPs being determined from the *same* set of individuals, which generally is overwhelmingly, if not in fact entirely, the case (in the extreme, with an independent set of individuals for each SNP, $\boldsymbol{M}$ would be reduced to the identity matrix). A limitation of the present work is that we do not consider this complexity. This may account for the relatively minor misfit in the simulation results for cases of high polygenicity – see below.

Thus, for example, if the SNP itself is not causal but is in LD with $k$ causal SNPs that all have heterozygosity $H$, and where its LD with each of these is the same, given by some value $r^2$ ($0 < r^2 \leq 1$), then $\tilde{\sigma}^2$ in Eq. 10 will be given by

$$\tilde{\sigma}^2 = kr^2\tilde{\sigma}^2_\beta NH. \tag{16}$$

For this idealized case, the marginal distribution, or pdf, of z-scores for a set of such associated SNPs is

$$f_1(z; N, \mathcal{H}, \sigma_\beta, \sigma_0) = \phi(z; 0, kr^2\tilde{\sigma}^2_\beta NH + \sigma^2_0) \tag{17}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the normal distribution with mean $\mu$ and variance $\sigma^2$, and $\mathcal{H}$ is shorthand for the LD and heterozygosity structure of such SNPs (in this case, denoting exactly $k$ causal SNPs with LD given by $r^2$ and heterozygosity given by $H$). If a proportion $\alpha$ of all typed SNPs are similarly associated with the phenotype while the remaining proportion are all null (not causal and not in LD with causal SNPs), then the marginal distribution for all SNP z-scores is the Gaussian mixture

$$f(z) = (1-\alpha)\phi(z; 0, \sigma^2_0) + \alpha f_1(z), \tag{18}$$

dropping the parameters for convenience.

For real genotypes, however, the LD and heterozygosity structure is far more complicated, and of course the causal SNPs are generally numerous and unknown. Thus, more generally, for each typed SNP $\mathcal{H}$ will be a two-dimensional histogram over LD ($r^2$) and heterozygosity ($H$), each grid element giving the number of SNPs falling within the edges of that $(r^2, H)$ bin. Alternatively, for each typed SNP it can be built as two one-dimensional histograms, one giving the LD structure (counts of neighboring SNPs in each LD

$r^2$ bin), and the other giving, for each $r^2$ bin, the mean heterozygosity for those neighboring SNPs, which will be accurate for sufficiently fine binning – within a bin, the heterozygosities of the tagged referene SNPs wll be in a vary narrow range. We use the latter in what follows. We present two consistent ways of expressing the *a posteriori* pdf for z-scores, based on multinomial expansion and on convolution, that provide complementary views. The multinomial approach perhaps gives a more intuitive feel for the problem, but the convolution approach is considerably more tractable numerically and is used here to obtain all reporter results. All code used in the analyses, including simulations, is publicly available on GitHub [27].

**Model PDF: Multinomial Expansion**

As in our previous work, we incorporate the model parameter $\pi_1$ for the fraction of all SNPs that are causal [18]. Additionally, we calculate the actual LD and heterozygosity structure for each SNP. That is, for each SNP we build a histogram of the numbers of other SNPs in LD with it for $w$ equally-spaced $r^2$-windows between $r^2_{min}$ and 1 where $r^2_{min} = 0.05$ (approximately the noise floor for correlation when LD is calculated from the 503 samples in 1000 Genomes), and record the mean heterozygosity for each bin; as noted above, we use $\mathcal{H}$ as shorthand to represent all this. We find that $w \simeq 20$ is sufficient for converged results. For any given SNP, the set of SNPs thus determined to be in LD with it constitute its LD block, with their number given by $n$ (LD with self is always 1, so $n$ is at least 1). The pdf for z-scores, given $N, \mathcal{H}$, and the three model parameters $\pi_1, \sigma_\beta, \sigma_0$, will then be given by the sum of Gaussians that are generalizations of Eq. 17 for different combinations of numbers of causal SNPs among the $w$ LD windows, each Gaussian scaled by the probability of the corresponding combination of causal SNPs among the LD windows, i.e., by the appropriate multinomial distribution term.

For $w$ $r^2$-windows, we must consider the possibilities where the typed SNP is in LD with all possible numbers of causal SNPs in each of these windows, or any combination thereof. There are thus $w+1$ categories of SNPs: null SNPs (which $r^2$-windows they are in is irrelevant), and causal SNPs, where it does matter which $r^2$-windows they reside in. If window $i$ has $n_i$ SNPs ($\sum_{i=1}^{w} n_i = n$) and mean heterozygosity $H_i$, and the overall fraction of SNPs that are causal is $\pi_1$, then the probability of having simultaneously $k_0$ null SNPs, $k_1$ causal SNPs in window 1, and so on through $k_w$ causal SNPs in window $w$, for a nominal total of $K$ causal SNPs ($\sum_{i=1}^{w} k_i = K$ and $k_0 = n - K$), is given by the multinomial distribution, which we denote $M(k_0, ..., k_w; n_0, ..., n_w; \pi_1)$. For an LD block of $n$ SNPs, the prior probability, $p_i$, for a SNP to be causal and in window $i$ is the product of the independent prior probabilities of a SNP being causal and being in window $i$: $p_i = \pi_1 n_i/n$. The prior probability of being null (regardless of $r^2$-window) is simply $p_0 = (1-\pi_1)$. The probability of a given breakdown $k_0, ..., k_w$ of the neighboring SNPs into the $w+1$ categories is then given by

$$M(k_0, ..., k_w; n_0, ..., n_w; \pi_1) = \frac{n!}{k_0!...k_w!} p_0^{k_0} ... p_w^{k_w} \qquad (19)$$

and the corresponding Gaussian is

$$\phi(z; 0, (k_1 H_1 r_1^2 + ... + k_w H_w r_w^2)\tilde{\sigma}_\beta^2 N + \sigma_0^2). \qquad (20)$$

For a SNP with LD and heterozygosity structure $\mathcal{H}$, the pdf for its z-score, given $N$ and the model parameters, is then given by summing over all possible numbers of total causal SNPs in LD with the SNP, and all possible distributions of those causal SNPs among the $w$ $r^2$-windows:

$$\begin{aligned} &\mathrm{pdf}(z; N, \mathcal{H}, \pi_1, \sigma_\beta, \sigma_0) = \\ &\sum_{K=0}^{K_{max}} \sum_{k_1,...,k_w} \frac{n!}{k_0!...k_w!} p_0^{k_0} ... p_w^{k_w} \times \\ &\phi(z; 0, (k_1 H_1 r_1^2 + ... + k_w H_w r_w^2)\tilde{\sigma}_\beta^2 N + \sigma_0^2), \end{aligned} \qquad (21)$$

where $K_{max}$ is bounded above by $n$. Note again that $\mathcal{H}$ is shorthand for the heterozygosity and linkage-disequilibrium structure of the SNP, giving the set $\{n_i\}$ (as well as $\{H_i\}$), and hence, for a given $\pi_1$, $p_i$. Also there is the constraint $\sum_{i=1}^{w} k_i = K$ on the second summation, and, for all $i$, $\max(k_i) = \max(K, n_i)$, though generally $K_{max} \ll n_i$. The number of ways of dividing $K$ causal SNPs amongst $w$ LD windows is given by the binomial coefficient $\binom{a}{b}$, where $a \equiv K+w-1$ and $b \equiv w-1$, so the number of terms in the second summation grows rapidly with $K$ and $w$. However, because $\pi_1$ is small (often $\leq 10^{-3}$), the upper bound on the first summation over total number of potential causal SNPs $K$ in the LD block for the SNP can be limited to $K_{max} < \min(20, n)$, even for large blocks with $n \simeq 10^3$. That is,

$$\sum_{K=0}^{K_{max}} \sum_{k_1,...,k_w} M(k_0, ..., k_w; n_0, ..., n_w; \pi_1) \simeq 1. \qquad (22)$$

Still, the number of terms is large; e.g., for $K = 10$ and $w = 10$ there are 92,378 terms.

For any given typed SNP (whose z-score we are trying to predict), it is important to emphasize that the specific LD $r^2$ and the heterozygosity of each underlying causal (reference) SNP tagged by it need to be taken into account, at least in an approximate sense that can be controlled to allow for arbitrary finessing giving converged results. This is the purpose of our $w = 20$ LD-$r^2$ windows, which inevitably leads to the multinomial expansion. Which window the causal SNP is in matters, leading to $w+1$ SNP categories, as noted above. Setting $w = 1$ would result in only a very rough approximation for the model pdf, reducing our multinomial to a binomial involving just two categories of SNPs: null and causal, with all causal SNPs treated the same, regardless of their LD with the tag SNP and their heterozygosity, as is done for the "M2" and "M3"

models in [14]. The effect of this is shown in two examples in the Supplementary Material ("Relation to Other Work").

**Model PDF: Convolution**

From Eq. 15, there exists an efficient procedure that allows for accurate calculation of a z-score's *a posteriori* pdf (given the SNP's heterozygosity and LD structure, and the phenotype's model parameters). Any GWAS z-score is a sum of unobserved random variables (LD-mediated contributions from neighboring causal SNPs, and the additive environmental component), and the pdf for such a composite random variable is given by the convolution of the pdfs for the component random variables. Since convolution is associative, and the Fourier transform of the convolution of two functions is just the product of the individual Fourier transforms of the two functions, one can obtain the *a posteriori* pdf for z-scores as the inverse Fourier transform of the product of the Fourier transforms of the individual random variable components.

From Eq. 15 $z$ is a sum of correlation- and heterozygosity-weighted random variables $\{\beta_j\}$ and the random variable $\epsilon$, where $\{\beta_j\}$ denotes the set of true causal parameters for each of the SNPs in LD with the typed SNP whose z-score is under consideration. The Fourier transform $F(k)$ of a Gaussian $f(x) = c \times \exp(-ax^2)$ is $F(k) = c\sqrt{\pi/a} \times \exp(-\pi^2 k^2/a)$. From Eq. 4, for each SNP j in LD with the typed SNP ($1 \leq j \leq b$, where $b$ is the typed SNP's block size),

$$\sqrt{NH_j} r_j \beta_j \sim \pi_1 \mathcal{N}(0, NH_j r_j^2 \tilde{\sigma}_\beta^2) + (1-\pi_1)\mathcal{N}(0,0). \quad (23)$$

The Fourier transform (with variable $k$ – see below) of the first term on the right hand side is

$$F(k) = \pi_1 \exp(-2\pi^2 k^2 NH_j r_j^2 \tilde{\sigma}_\beta^2), \quad (24)$$

while that of the second term is simply $(1-\pi_1)$. Additionally, the environmental term is $\epsilon \sim \mathcal{N}(0, \sigma_0^2)$ (ignoring LD-induced correlation, as noted earlier), and its Fourier transform is $\exp(-2\pi^2\sigma_0^2 k^2)$. For each typed SNP, one could construct the *a posteriori* pdf based on these Fourier transforms. However, it is more practical to use a coarse-grained representation of the data. Thus, in order to fit the model to a data set, we bin the typed SNPs whose z-scores comprise the data set into a two-dimensional heterozygosity/total LD grid (whose elements we denote "H-L" bins), and fit the model with respect to this coarse grid instead of with respect to every individual typed SNP z-score; in the section "Parameter Estimation" below we describe using a $10 \times 10$ grid. Additionally, for each H-L bin the LD $r^2$ and heterozygosity histogram structure for each typed SNP is built, using $w_{max}$ equally-spaced $r^2$ bins for $r^2_{min} \leq r^2 \leq 1$ (this is a change in notation from the previous section: $w_{max}$ here plays the role of $w$ there; in what follows, $w$ will be used as a running index, $1 \leq w \leq w_{max}$); $w_{max} = 20$ is large enough to allow for converged results; $r^2_{min} = 0.05$ is generally small enough to capture true causal associations in weak LD while large enough to exclude spurious contributions to the pdf arising from estimates of $r^2$ that are non-zero due to noise. This points up a minor limitation of the model stemming from the small reference sample size ($N_R = 503$ for 1000 Genomes) from which $\mathcal{H}$ is built. Larger $N_R$ would allow for more precision in handling very low LD ($r^2 < 0.05$), but this is an issue only for situations with extremely large $\sigma_\beta^2$ (high heritability with low polygenicity) that we do not encounter for the 16 phenotypes we analyze here. In any case, this can be calibrated for using simulations.

We emphasize again that setting $w_{max} = 1$ would result in only an approximation for the model pdf (see "Relation to Other Work" in the Supplementary Material).

For any H-L bin with mean heterozygosity $H$ and mean total LD $L$ there will be an average LD and heterozygosity structure with a mean breakdown for the typed SNPs having $n_w$ reference SNPs (not all of which necessarily are typed SNPs, i.e., have a z-score) with LD $r^2$ in the $w^{\text{th}}$ $r^2$ bin whose average heterozygosity is $H_w$. Thus, one can re-express z-scores for an H-L bin as

$$z = \sqrt{N} \sum_{w=1}^{w_{max}} \left( \sqrt{H_w} r_w \sum_{j=0}^{n_w} \beta_j \right) + \epsilon \quad (25)$$

where $\beta_j$ and $\epsilon$ are unobserved random variables.

In the spirit of the discrete Fourier transform (DFT), discretize the set of possible z-scores into the ordered set of $n$ (equal to a power of 2) values $z_1, ..., z_n$ with equal spacing between neighbors given by $\Delta z$ ($z_n = -z_1 - \Delta z$, and $z_{n/2+1} = 0$). Taking $z_1 = -38$ allows for the minimum p-values of $5.8 \times 10^{-316}$ (near the numerical limit); with $n = 2^{10}$, $\Delta z = 0.0742$. Given $\Delta z$, the Nyquist critical frequency is $f_c = \frac{1}{2\Delta z}$, so we consider the Fourier transform function for the z-score pdf at $n$ discrete values $k_1, ..., k_n$, with equal spacing between neighbors given by $\Delta k$, where $k_1 = -f_c$ ($k_n = -k_1 - \Delta k$, and $k_{n/2+1} = 0$; the DFT pair $\Delta z$ and $\Delta k$ are related by $\Delta z \Delta k = 1/n$). Define

$$A_w \equiv -2\pi^2 NH_w r_w^2 \tilde{\sigma}_\beta^2. \quad (26)$$

(see Eq. 24). Then the product (over $r^2$ bins) of Fourier transforms for the genetic contribution to z-scores, denoted $G_j \equiv G(k_j)$, is

$$G(k_j) = \prod_{w=1}^{w_{max}} \left( \pi_1 \exp(A_w k_j^2) + (1-\pi_1) \right)^{n_w}, \quad (27)$$

and the Fourier transform of the environmental contribution, denoted $E_j \equiv E(k_j)$, is

$$E(k_j) = \exp(-2\pi^2 \sigma_0^2 k_j^2). \quad (28)$$

Let $\mathbf{F_z} = (G_1 E_1, ..., G_n E_n)$ denote the vector of products of Fourier transform values, and let $\mathcal{F}^{-1}$ denote the inverse Fourier transform operator. Then, the vector of pdf values for z-score bins (indexed by $i$) in the H-L bin with mean

LD and heterozygosity structure $\mathcal{H}$, $\mathbf{pdf_z} = (f_1, ..., f_n)$ where $f_i \equiv \text{pdf}(z_i|\mathcal{H})$, is

$$\mathbf{pdf_z} = \mathcal{F}^{-1}\left[\mathbf{F_z}\right]. \tag{29}$$

**Data Preparation**

For real phenotypes, we calculated SNP minor allele frequency (MAF) and LD between SNPs using the 1000 Genomes phase 3 data set for 503 subjects/samples of European ancestry [28, 29, 30]. In order to carry out realistic simulations (i.e., with realistic heterozygosity and LD structures for SNPs), we used HAPGEN2 [31, 32, 33] to generate genotypes; we calculated SNP MAF and LD structure from 1000 simulated samples. We elected to use the same intersecting set of SNPs for real data and simulation. For HAPGEN2, we eliminated SNPs with MAF<0.002; for 1000 Genomes, we eliminated SNPs for which the call rate (percentage of samples with useful data) was less than 90%. This left $n_{snp}$=11,015,833 SNPs. See Supplementary Material for further details.

We analyzed summary statistics for sixteen phenotypes (in what follows, where sample sizes varied by SNP, we quote the median value): (1) major depressive disorder ($N_{cases}$ = 59,851, $N_{controls}$ = 113,154) [34]; (2) bipolar disorder ($N_{cases}$ = 20,352, $N_{controls}$ = 31,358) [35]; (3) schizophrenia ($N_{cases}$ = 35,476, $N_{controls}$ = 46,839) [36]; (4) coronary artery disease ($N_{cases}$ = 60,801, $N_{controls}$ = 123,504) [37]; (5) ulcerative colitis ($N_{cases}$ = 12,366, $N_{controls}$ = 34,915) and (6) Crohn's disease ($N_{cases}$ = 12,194, $N_{controls}$ = 34,915) [38]; (7) late onset Alzheimer's disease (LOAD; $N_{cases}$ = 17,008, $N_{controls}$ = 37,154) [39] (in the Supplementary Material we present results for a more recent GWAS with $N_{cases}$ = 71,880 and $N_{controls}$ = 383,378 [40]); (8) amyotrophic lateral sclerosis (ALS) ($N_{cases}$ = 12,577, $N_{controls}$ = 23,475) [41]; (9) number of years of formal education ($N$ = 293,723) [42]; (10) intelligence ($N$ = 262,529) [43, 44]; (11) body mass index ($N$ = 233,554) [45]; (12) height ($N$ = 251,747) [46]; (13) putamen volume (normalized by intracranial volume, $N$ = 11,598) [47]; (14) low- ($N$ = 89,873) and (15) high-density lipoprotein ($N$ = 94,295) [48]; and (16) total cholesterol ($N$ = 94,579) [48]. Most participants were of European ancestry.

For height, we focused on the 2014 GWAS [46], not the more recent 2018 GWAS [49], although we also report below model results for the latter. There are issues pertaining to population structure in the various height GWAS [50, 51], and the 2018 GWAS is a combination of GIANT and UKB GWAS, so some caution is warranted in interpreting results for these data.

For the ALS GWAS data, there is very little signal outside chromosome 9: the data QQ plot essentially tracks the null distribution straight line. The QQ plot for chromosome 9, however, shows a significant departure from the null distribution. Of 471,607 SNPs on chromosome 9 a subset of 273,715 have z-scores, of which 107 are genome-wide significant, compared with 114 across the full genome. Therefore, we restrict ALS analysis to chromosome 9.

A limitation in the current work is that we have not taken account of imputation inaccuracy, where lower MAF SNPs are, through lower LD, less certain. Thus, the effects from lower MAF causal variants will be noisier than for higher MAF variants.

**Simulations**

We generated genotypes for $10^5$ unrelated simulated samples using HAPGEN2 [33]. For narrow-sense heritability $h^2$ equal to 0.1, 0.4, and 0.7, we considered polygenicity $\pi_1$ equal to $10^{-5}$, $10^{-4}$, $10^{-3}$, and $10^{-2}$. For each of these 12 combinations, we randomly selected $n_{causal} = \pi_1 \times n_{snp}$ "causal" SNPs and assigned them $\beta$-values drawn from the standard normal distribution (i.e., independent of $H$), with all other SNPs having $\beta = 0$. We repeated this ten times, giving ten independent instantiations of random vectors of $\beta$'s. Defining $Y_G = G\beta$, where $G$ is the genotype matrix and $\beta$ here is the vector of true coefficients over all SNPs, the total phenotype vector is constructed as $Y$=$Y_G$+$\varepsilon$, where the residual random vector $\varepsilon$ for each instantiation is drawn from a normal distribution such that $h^2 = \text{var}(Y_G)/\text{var}(Y)$. For each of the instantiations this implicitly defines the "true" value $\sigma^2_\beta$.

The sample simple linear regression slope, $\hat{\beta}$, and the Pearson correlation coefficient, $\hat{r}$, are assumed to be t-distributed. These quantities have the same t-value: $t = \hat{\beta}/\text{se}(\hat{\beta}) = \hat{r}/\text{se}(\hat{r}) = \hat{r}\sqrt{N-2}/\sqrt{1-\hat{r}^2}$, with corresponding p-value from Student's $t$ cumulative distribution function (cdf) with $N-2$ degrees of freedom: $p = 2\times\text{tcdf}(-|t|, N-2)$ (see Supplementary Material). Since we are not here dealing with covariates, we calculated $p$ from correlation, which is slightly faster than from estimating the regression coefficient. The t-value can be transformed to a z-value, giving the z-score for this $p$: $z = -\Phi^{-1}(p/2) \times \text{sign}(\hat{r})$, where $\Phi$ is the normal cdf ($z$ and $t$ have the same p-value).

**Parameter Estimation**

We randomly pruned SNPs using the threshold $r^2 > 0.8$ to identify "synonymous" SNPs, performing ten such iterations. That is, for each of ten iterations, we randomly selected a SNP (not necessarily the one with largest z-score) to represent each subset of synonymous SNPs. For schizophrenia, for example, pruning resulted in approximately 1.3 million SNPs in each iteration.

The postulated pdf for a SNP's z-score depends on the SNP's LD and heterozygosity structure (histogram), $\mathcal{H}$. Given the data – the set of z-scores for available SNPs, as well as their LD and heterozygosity structure – and the $\mathcal{H}$-dependent pdf for z-scores, the objective is to find the model parameters that best predict the distribution of z-scores. We bin the SNPs with respect to a grid of heterozygosity and total LD; for any given H-L bin there will be a range of z-scores whose distribution the model it intended to predict. We find that a $10 \times 10$-grid of equally spaced

bins is adequate for converged results. (Using equally-spaced bins might seem inefficient because of the resulting very uneven distribution of z-scores among grid elements – for example, orders of magnitude more SNPs in grid elements with low total LD compared with high total LD. However, the objective is to model the effects of H and L: using variable grid element sizes so as to maximize balance of SNP counts among grid elements means that the true H- and L-mediated effects of the SNPs in a narrow range of H and L get subsumed with the effects of many more SNPs in a much wider range of H and L – a misspecification of the pdf leading to some inaccuracy.) In lieu of or in addition to total LD (L) binning, one can bin SNPs with respect to their total LD block size (total number of SNPs in LD, ranging from 1 to ∼1,500).

To find the model parameters that best fit the data, for a given H-L bin we binned the selected SNPs z-scores into equally-spaced bins of width $dz$=0.0742 (between $z_{min}$= −38 and $z_{max}$=38, allowing for p-values near the numerical limit of $10^{-316}$), and from Eq. 29 calculated the probability for z-scores to be in each of those z-score bins (the prior probability for "success" in each z-score bin). Then, knowing the actual numbers of z-scores (numbers of "successes") in each z-score bin, we calculated the multinomial probability, $p_m$, for this outcome. The optimal model parameter values will be those that maximize the accrual of this probability over all H-L bins. We constructed a cost function by calculating, for a given H-L bin, $-\ln(p_m)$ and averaging over prunings, and then accumulating this over all H-L bins. Model parameters minimizing the cost were obtained from Nelder-Mead multidimensional unconstrained nonlinear minimization of the cost function, using the Matlab function `fminsearch()`.

**Posterior Effect Sizes**

Model posterior effect sizes, given $z$ (along with N, $\mathcal{H}$, and the model parameters), were calculated using numerical integration over the random variable $\delta$:

$$\delta_{expected} \equiv E(\delta|z) = \int P(\delta|z)\delta d\delta = \frac{1}{P(z)} \int P(z|\delta)P(\delta)\delta d\delta. \quad (30)$$

Here, since $z|\delta \sim \mathcal{N}(\delta, \sigma_0^2)$, the posterior probability of $z$ given $\delta$ is simply

$$P(z|\delta) = \phi(z; \delta, \sigma_0^2). \quad (31)$$

$P(z)$ is shorthand for $\text{pdf}(z|N, \mathcal{H}, \pi_1, \sigma_\beta, \sigma_0)$, given by Eq. 29. $P(\delta)$ is calculated by a similar procedure that lead to Eq. 29 but ignoring the environmental contributions $\{E_j\}$. Specifically, let $\mathbf{F}_\delta = (G_1, ..., G_n)$ denote the vector of products of Fourier transform values. Then, the vector of pdf values for genetic effect bins (indexed by $i$; numerically, these will be the same as the z-score bins) in the H-L bin, $\mathbf{pdf}_\delta = (f_1, ..., f_n)$ where $f_i \equiv \text{pdf}(\delta_i|\mathcal{H})$, is

$$\mathbf{pdf}_\delta = \mathcal{F}^{-1}\left[\mathbf{F}_\delta\right]. \quad (32)$$

Similarly,

$$\delta^2_{expected} \equiv E(\delta^2|z) = \int P(\delta|z)\delta^2 d\delta = \frac{1}{P(z)} \int P(z|\delta)P(\delta)\delta^2 d\delta, \quad (33)$$

which is used in power calculations.

**GWAS Replication**

A related matter has to do with whether z-scores for SNPs reaching genome-wide significance in a discovery-sample are compatible with the SNPs' z-scores in a replication-sample, particularly if any of those replication-sample z-scores are far from reaching genome-wide significance, or whether any apparent mismatch signifies some overlooked inconsistency. The model pdf allows one to make a principled statistical assessment in such cases. We present the details for this application, and results applied to studies of bipolar disorder, in the Supplementary Material.

**GWAS Power**

Chip heritability, $h^2_{SNP}$, is the proportion of phenotypic variance that in principle can be captured additively by the $n_{snp}$ SNPs under study [17]. It is of interest to estimate the proportion of $h^2_{SNP}$ that can be explained by SNPs reaching genome-wide significance, $p \leq 5\times10^{-8}$ (i.e., for which $|z|>z_t$=5.45), at a given sample size [60, 61]. In Eq 1, for SNP $i$ with genotype vector $g_i$ over $N$ samples, let $y_{g_i} \equiv g_i\beta_i$. If the SNP's heterozygosity is $H_i$, then $\text{var}(y_{g_i})=\beta_i^2 H_i$. If we knew the full set $\{\beta_i\}$ of true $\beta$-values, then, for z-scores from a particular sample size $N$, the proportion of SNP heritability captured by genome-wide significant SNPs, $A(N)$, would be given by

$$A(N) = \frac{\sum_{i:|z_i|>z_t} \beta_i^2 H_i}{\sum_{all\ i} \beta_i^2 H_i}. \quad (34)$$

Now, from Eq. 15, $\delta_i = \sqrt{N}\sum_j \sqrt{H_j} r_{ij}\beta_j$. If SNP $i$ is causal and sufficiently isolated so that it is not in LD with other causal SNPs, then $\delta_i = \sqrt{N}\sqrt{H_i}\beta_i$, and $\text{var}(y_{g_i}) = \delta_i^2/N$. When all causal SNPs are similarly isolated, Eq. 34 becomes

$$A(N) = \frac{\sum_{i:|z_i|>z_t} \delta_i^2}{\sum_{all\ i} \delta_i^2}. \quad (35)$$

Of course, the true $\beta_i$ are not known and some causal SNPs will likely be in LD with others. Furthermore, due to LD with causal SNPs, many SNPs will have a nonzero (latent or unobserved) effect size, $\delta$. Nevertheless, we can formulate an approximation to $A(N)$ which, assuming the pdf for z-scores (Eq. 29) is reasonable, will be inaccurate to the degree that the average LD structure of genome-wide significant SNPs differs from the overall average LD structure. As before (see the subsection "Model PDF: Convolution"), consider a fixed set of $n$ equally-spaced nominal z-scores covering a wide range of possible values (changing from the summations in Eq. 35 to the uniform summation spacing $\Delta z$ now requires bringing the probability

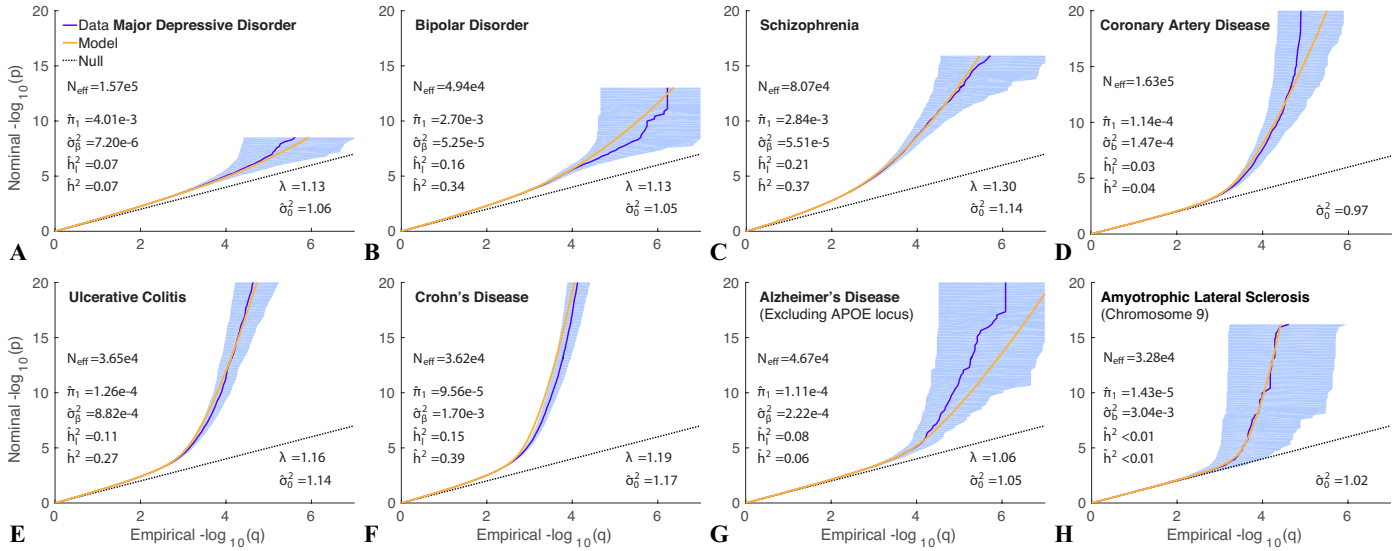


Figure 1: QQ plots of (pruned) z-scores for qualitative phenotypes (dark blue, 95% confidence interval in light blue) with model prediction (yellow): (A) major depressive disorder; (B) bipolar disorder; (C) schizophrenia; (D) coronary artery disease (CAD); (E) ulcerative colitis (UC); (F) Crohn's disease (CD); (G) late onset Alzheimer's disease (AD), excluding APOE (see also Supplementary Material Fig. S7); and (H) amyotrophic lateral sclerosis (ALS), restricted to chromosome 9 (see also Supporting Material Figure S9). The dashed line is the expected QQ plot under null (no SNPs associated with the phenotype). $p$ is a nominal p-value for z-scores, and $q$ is the proportion of z-scores with p-values exceeding that threshold. $\lambda$ is the overall nominal genomic control factor for the pruned data (which is accurately predicted by the model in all cases). The three estimated model parameters are: polygenicity, $\hat{\pi}_1$; discoverability, $\hat{\sigma}^2_\beta$ (corrected for inflation); and SNP association $\chi^2$-statistic inflation factor, $\hat{\sigma}^2_0$. $\hat{h}^2$ is the estimated narrow-sense chip heritability, re-expressed as $h^2_l$ on the liability scale for these case-control conditions assuming a prevalence of: MDD 6.7% [52], BIP 0.5% [53], SCZ 1% [54], CAD 3% [55], UC 0.1% [56], CD 0.1% [56], AD 14% (for people aged 71 and older in the USA [57, 58]), and ALS $5 \times 10^{-5}$ [59]. The estimated number of causal SNPs is given by $\hat{n}_{causal} = \hat{\pi}_1 n_{snp}$ where $n_{snp} = 11,015,833$ is the total number of SNPs, whose LD structure and MAF underlie the model; the GWAS z-scores are for subsets of these SNPs. $N_{eff}$ is the effective case-control sample size – see text. Reading the plots: on the vertical axis, choose a p-value threshold (more extreme values are further from the origin), then the horizontal axis gives the proportion of SNPs exceeding that threshold (higher proportions are closer to the origin). Numerical values for the model parameters are also given in Table 1. See also Supplementary Material Figs. S12-S18.

density into the summations). For each $z$ from the fixed set (and, as before, employing data reduction by averaging so that H and L denote values for the $10 \times 10$ grid), use $E(\delta^2|z, N, H, L)$ given in Eq. 33 to define

$$C(z|N, H, L) \equiv E(\delta^2|z, N, H, L)P(z|N, H, L) \qquad (36)$$

(emphasizing dependence on N, H, and L). Then, for any $N$, $A(N)$ can be estimated by

$$A(N) = \frac{\sum_{H,L} \sum_{z:|z|>z_t} C(z, N, H, L)}{\sum_{H,L} \sum_{all\ z} C(z, N, H, L)} \qquad (37)$$

where $\sum_{H,L}$ denotes sum over the H-L grid elements. The ratio in Eq. 37 should be accurate if the average effects of LD in the numerator and denominator cancel – which will always be true as the ratio approaches 1 for large $N$. Plotting $A(N)$ gives an indication of the power of future GWAS to capture chip heritability.

### Quantile-Quantile Plots and Genomic Control

One of the advantages of quantile-quantile (QQ) plots is that on a logarithmic scale they emphasize behavior in the tails of a distribution, and provide a valuable visual aid in assessing the independent effects of polygenicity, strength of association, and variance distortion – the roles played by the three model parameters – as well as showing how well a model fits data. QQ plots for the model were constructed using Eq. 29, replacing the normal pdf with the normal cdf, and replacing $z$ with an equally-spaced vector $\vec{z}_{nom}$ of length 10,000 covering a wide range of nominal $|z|$ values (0 through 38). SNPs were divided into a $10\times10$ grid of H$\times$L bins, and the cdf vector (with elements corresponding to the z-values in $\vec{z}_{nom}$) accumulated for each such bin (using mean values of H and L for SNPs in a given bin).

For a given set of samples and SNPs, the genomic control factor, $\lambda$, for the z-scores is defined as the median $z^2$ divided by the median for the null distribution, 0.455 [19]. This can also be calculated from the QQ plot. In the plots we present here, the abscissa gives the -$\log_{10}$ of the proportion, $q$, of SNPs whose z-scores exceed the two-tailed significance threshold $p$, transformed in the ordinate as -$\log_{10}(p)$. The median is at $q_{med} = 0.5$, or $-\log_{10}(q_{med}) \simeq 0.3$; the corresponding empirical and model p-value thresholds ($p_{med}$) for the z-scores – and equivalently for the z-scores-squared – can be read off from the plots. The genomic inflation factor is then given by

$$\lambda = [\Phi^{-1}(p_{med}/2)]^2/0.455.$$

Note that the values of $\lambda$ reported here are for pruned SNP

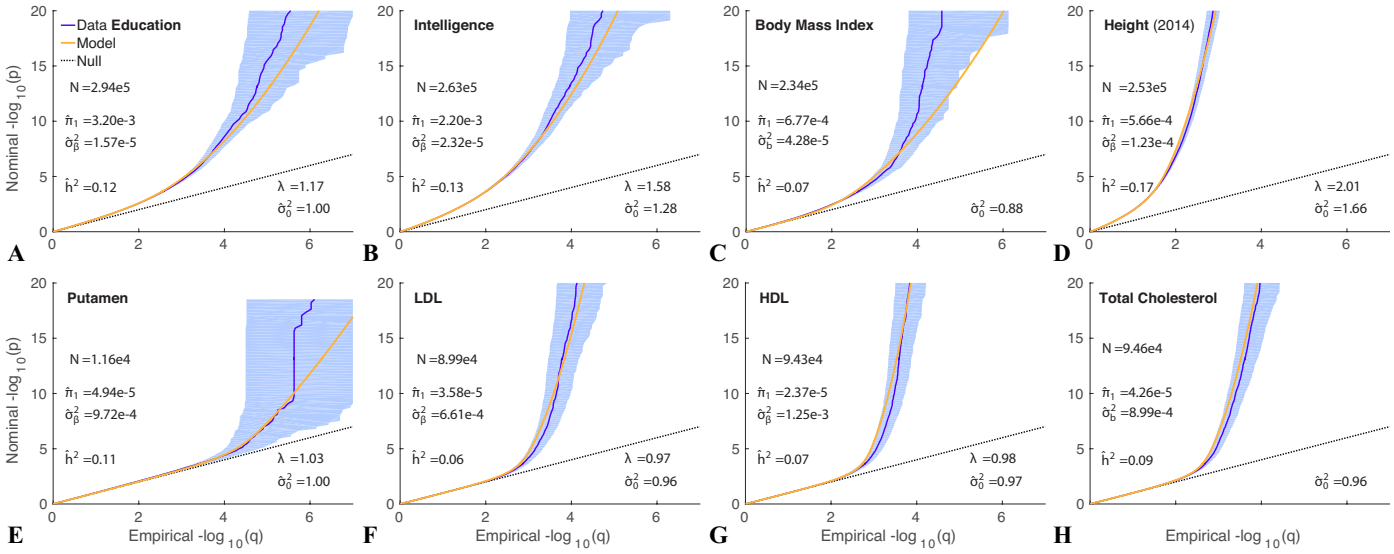


Figure 2: QQ plots of (pruned) z-scores and model fits for quantitative phenotypes: (A) educational attainment; (B) intelligence; (C) body mass index (BMI); (D) height; (E) putamen volume; (F) low-density lipoprotein (LDL); (G) high-density lipoprotein (HDL); and (H) total cholesterol (TC). $N$ is the sample size. See Fig. 1 for further description. Numerical values for the model parameters are also given in Table 1. See also Supplementary Material Figs. S19-S24.

sets; these values will be lower than for the total GWAS SNP sets.

Knowing the total number, $n_{tot}$, of p-values involved in a QQ plot (number of GWAS z-scores from pruned SNPs), any point $(q, p)$ (log-transformed) on the plot gives the number, $n_p = q\, n_{tot}$, of p-values that are as extreme as or more extreme than the chosen p-value. This can be thought of as $n_p$ "successes" out of $n_{tot}$ independent trials (thus ignoring LD) from a binomial distribution with prior probability $q$. To approximate the effects of LD, we estimate the number of independent SNPs as $n_{tot}/f$ where $f \simeq 10$. The 95% binomial confidence interval for $q$ is calculated as the exact Clopper-Pearson 95% interval [62], which is similar to the normal approximation interval, $q \pm 1.96\sqrt{q(1-q)/n_{tot}/f}$.

## Number of Causal SNPs

The estimated number of causal SNPs is given by the polygenicity, $\pi_1$, times the total number of SNPs, $n_{snp}$: $n_{causal} = \pi_1 n_{snp}$. $n_{snp}$ is given by the total number of SNPs that went into building the heterozygosity/LD structure, $\mathcal{H}$ in Eq. 29, i.e., the approximately 11 million SNPs selected from the 1000 Genomes Phase 3 reference panel, not the number of typed SNPs in the particular GWAS. The parameters estimated are to be seen in the context of the reference panel, which we assume contains all common causal variants. Stable quantities (i.e., fairly independent of the reference panel size. e.g., using the full panel or ignoring every second SNP), are the estimated effect size variance and number of causal variants – which we demonstrate below – and hence the heritability. Thus, the polygenicity will scale inversely with the reference panel size. A reference panel with a substantially larger number of samples would allow for inclusion of more SNPs (non-zero MAF), and thus the actual polygenicity estimated would change slightly.

## Narrow-sense Chip Heritability

Since we are treating the $\beta$ coefficients as fixed effects in the simple linear regression GWAS formalism, with the phenotype vector standardized with mean zero and unit variance, from Eq. 1 the proportion of phenotypic variance explained by a particular causal SNP whose reference panel genotype vector is $g$, $q^2$=var$(y; g)$, is given by $q^2 = \beta^2 H$. The proportion of phenotypic variance explained additively by all causal SNPs is, by definition, the narrow sense chip heritability, $h^2$. Since E$(\beta^2)$=$\sigma^2_\beta$ and $n_{causal} = \pi_1 n_{snp}$, and taking the mean heterozygosity over causal SNPs to be approximately equal to the mean over all SNPs, $\overline{H}$, the chip heritability can be estimated as

$$h^2 = \pi_1 n_{snp} \overline{H} \sigma^2_\beta. \tag{38}$$

Mean heterozygosity from the ∼11 million SNPs is $\overline{H} = 0.2165$.

For all-or-none traits like disease status, the estimated $h^2$ from Eq. 38 for an ascertained case-control study is on the observed scale and is a function of the prevalence in the adult population, $K$, and the proportion of cases in the study, $P$. The heritability on the underlying continuous liability scale [63], $h^2_l$, is obtained by adjusting for ascertainment (multiplying by $K(1-K)/(P(1-P))$, the ratio of phenotypic variances in the population and in the study) and rescaling based on prevalence [64, 6]:

$$h^2_l = h^2 \frac{K(1-K)}{P(1-P)} \times \frac{K(1-K)}{a^2}, \tag{39}$$

where $a$ is the height of the standard normal pdf at the

| Phenotype | $\pi_1$ | $\sigma^2_\beta$ | $\sigma^2_0$ | $n_{causal}$ | $h^2_{(l)}$ |
|---|---|---|---|---|---|
| MDD | 4.01E-3 | 7.20E-6 | 1.06 | 4.4E4 | 0.07 |
| Bipolar Disorder | 2.70E-3 | 5.25E-5 | 1.05 | 3.0E4 | 0.16 |
| Schizophrenia | 2.84E-3 | 5.51E-5 | 1.14 | 3.1E4 | 0.21 |
| CAD | 1.14E-4 | 1.47E-4 | 0.97 | 1.3E3 | 0.03 |
| Ulcerative Colitis | 1.26E-4 | 8.82E-4 | 1.14 | 1.4E3 | 0.11 |
| Crohn's Disease | 9.56E-5 | 1.70E-3 | 1.17 | 1.1E3 | 0.15 |
| AD (no APOE)* | 1.11E-4 | 2.22E-4 | 1.05 | 1.2E3 | 0.08 |
| ALS† | 1.43E-5 | 3.04E-3 | 1.02 | 7 | 0.00 |
| Education | 3.20E-3 | 1.57E-5 | 1.00 | 3.5E4 | 0.12 |
| Intelligence | 2.20E-3 | 2.32E-5 | 1.28 | 2.4E4 | 0.13 |
| BMI | 6.44E-4 | 4.28E-5 | 0.88 | 709 | 0.07 |
| Height (2010)$ | 4.32E-4 | 1.66E-4 | 0.94 | 4.8E3 | 0.17 |
| Height (2014) | 5.66E-4 | 1.23E-4 | 1.66 | 6.2E3 | 0.17 |
| Height (2018)$ | 8.56E-4 | 9.46E-5 | 2.50 | 9.4E3 | 0.19 |
| Putamen Volume | 4.94E-5 | 9.72E-4 | 1.00 | 540 | 0.11 |
| LDL | 3.58E-5 | 6.61E-4 | 0.96 | 390 | 0.06 |
| HDL | 2.37E-5 | 1.25E-3 | 0.97 | 260 | 0.07 |
| TC | 4.26E-5 | 8.99E-4 | 0.96 | 469 | 0.09 |

Table 1: Summary of model results for phenotypes shown in Figures 1 and 2. The subscript in $h^2_{(l)}$ indicates that for the qualitative phenotypes (the first eight) the reported SNP heritability is on the liability scale. MDD: Major Depressive Disorder; CAD: coronary artery disease; AD: Alzheimer's Disease (excluding APOE locus; *for the full autosomal reference panel, i.e., including APOE, $h^2_l = 0.15$ for AD – see Supplementary Material Figure S7 (A) and (B)); BMI: body mass index; †ALS: amyotrophic lateral sclerosis, restricted to chromosome 9; LDL: low-density lipoproteins; HDL: high-density lipoproteins. $In addition to the 2014 height GWAS (N=251,747 [46]), we include here model results for the 2010 (N=133,735 [65]) and 2018 (N=707,868 [49]) height GWAS; there is remarkable consistency for the 2010 and 2014 GWAS despite very large differences in the sample sizes – see Supporting Material Figure S8.

truncation point $z_K$ defined such that the area under the curve in the region to the right of $z_K$ is $K$.

## RESULTS

### Simulations

Table 2 shows the simulation results, comparing true and estimated values for the model parameters, heritability, and the number of causal SNPs, for twelve scenarios where $\pi_1$ and $\sigma^2_\beta$ both range over three orders of magnitude, encompassing the range of values for the phenotypes; in Supplementary Material, Figure S3 shows QQ plots for a randomly chosen (out of 10) $\beta$-vector and phenotype instantiation for each of the twelve ($\pi_1$, $h^2$) scenarios. Most of the $\hat{\pi}_1$ estimates are in very good agreement with the true values, though for the extreme scenario of high heritability and low polygenicity it is overestimated by factors of two-to-three. The numbers of estimated causal SNPs (out of ∼11 million) are in correspondingly good agreement with the true values, ranging in increasing powers of 10 from 110 through 110,158. The estimated discoverabilities ($\hat{\sigma}^2_\beta$) are also in good agreement with the true values. In most cases, $\hat{\sigma}^2_0$ is close to 1, indicating little or no global inflation, though it is elevated for high heritability with high polygenicity, suggesting it is capturing some ubiquitous effects.

In Supplementary Material, we examine the issue of model misspecification. Specifically, we assign causal effects $\beta$ drawn from a Gaussian whose variance is not simply a constant but depends on heterozygosity, such that rarer causal SNPs will tend to have larger effects [15]. The results – see Supplementary Material Table S1 – show that the model still makes reasonable estimates of the underlying genetic architecture. Additionally, we tested the scenario where true causal effects are distributed with respect to two Gaussians [14], a situation that allows for a small number of the causal SNPs to have quite large effects – see Supplementary Material Table S2. We find that heritabilities are still reasonably estimated using our model. In all these scenarios the overall data QQ plots were accurately reproduced by the model. As a counter example, we simulated summary statistics where the prior probability of a reference SNP being causal decreased linearly with total LD (see Supplementary Material Table S3). In this case, our single Gaussian fit (which assumes no LD dependence on the prior probability of a reference SNP being causal) did not produce model QQ plots that accurately tracked the data QQ plots (see Supplementary Material Figure S11). The model parameters and heritabilities were also poor. But this scenario is highly artificial; in contrast, in situations where the data QQ plots were accurately reproduced by the model, the estimated model parameters and heritability were plausible.

### Phenotypes

Figures 1 and 2 show QQ plots for the pruned z-scores for eight qualitative and eight quantitative phenotypes, along with model estimates (Supplementary Material Figs. S12-S28 each show a 4×4 grid breakdown by heterozygosity × total-LD of QQ plots for all phenotypes studied here; the 4×4 grid is a subset of the 10×10 grid used in the calculations). In all cases, the model fit (yellow) closely tracks the data (dark blue). For the sixteen phenotypes, estimates for the model polygenicity parameter (fraction of reference panel, with ≃11 million SNPs, estimated to have non-null effects) range over two orders of magnitude, from $\pi_1 \simeq 2 \times 10^{-5}$ to $\pi_1 \simeq 4 \times 10^{-3}$. The estimated SNP discoverability parameter (variance of $\beta$, or expected $\beta^2$, for causal variants) also ranges over two orders of magnitude from $\sigma^2_\beta \simeq 7 \times 10^{-6}$ to $\sigma^2_\beta \simeq 2 \times 10^{-3}$ (in units where the variance of the phenotype is normalized to 1).

We find that schizophrenia and bipolar disorder appear to be similarly highly polygenic, with model polygenicities $\simeq 2.84 \times 10^{-3}$ and $\simeq 2.70 \times 10^{-3}$, respectively. The model polygenicity of major depressive disorder, however, is 40% higher, $\pi_1 \simeq 4 \times 10^{-3}$ – the highest value among the sixteen phenotypes. In contrast, the model polygenicities of late

| $h^2$ | $\hat{h}^2$ | $\pi_1$ | $\hat{\pi}_1$ | $\sigma^2_\beta$ | $\hat{\sigma}^2_\beta$ | $\hat{\sigma}^2_0$ | $n_{causal}$ | $\hat{n}_{causal}$ |
|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.12 (0.01) | 1E-5 | 1.4E-5 (2E-6) | 4.3E-3 (7E-4) | 3.6E-3 (5E-4) | 1.01 (0.002) | 110 | 151 (20) |
| 0.1 | 0.10 (0.01) | 1E-4 | 1.0E-4 (2E-5) | 4.2E-4 (2E-5) | 4.1E-4 (5E-5) | 1.01 (0.003) | 1101 | 1130 (206) |
| 0.1 | 0.09 (0.01) | 1E-3 | 0.9E-3 (1E-4) | 4.2E-5 (5E-7) | 4.1E-5 (4E-6) | 1.02 (0.003) | 11015 | 10340 (1484) |
| 0.1 | 0.09 (0.01) | 1E-2 | 0.8E-2 (2E-3) | 4.2E-6 (4E-8) | 5.6E-6 (2E-6) | 1.02 (0.002) | 110158 | 83411 (25448) |
| 0.4 | 0.52 (0.05) | 1E-5 | 2.3E-5 (2E-6) | 1.7E-2 (3E-3) | 9.1E-3 (1E-3) | 1.02 (0.002) | 110 | 259 (20) |
| 0.4 | 0.45 (0.02) | 1E-4 | 1.2E-4 (8E-6) | 1.7E-3 (7E-5) | 1.5E-3 (9E-5) | 1.04 (0.002) | 1101 | 1310 (92) |
| 0.4 | 0.39 (0.01) | 1E-3 | 1.0E-3 (5E-5) | 1.7E-4 (2E-6) | 1.6E-4 (8E-6) | 1.05 (0.003) | 11015 | 10607 (578) |
| 0.4 | 0.37 (0.01) | 1E-2 | 0.9E-2 (1E-3) | 1.7E-5 (2E-7) | 1.7E-5 (2E-6) | 1.06 (0.003) | 110158 | 95135 (10851) |
| 0.7 | 0.91 (0.09) | 1E-5 | 2.9E-5 (2E-6) | 3.0E-2 (5E-3) | 1.3E-2 (2E-3) | 1.02 (0.003) | 110 | 324 (24) |
| 0.7 | 0.82 (0.02) | 1E-4 | 1.4E-4 (7E-6) | 2.9E-3 (1E-4) | 2.4E-3 (1E-4) | 1.05 (0.002) | 1101 | 1493 (79) |
| 0.7 | 0.70 (0.01) | 1E-3 | 1.0E-3 (4E-5) | 2.9E-4 (4E-6) | 2.8E-4 (1E-5) | 1.08 (0.003) | 11015 | 10866 (406) |
| 0.7 | 0.66 (0.01) | 1E-2 | 0.9E-2 (7E-4) | 2.9E-5 (3E-7) | 2.9E-5 (2E-6) | 1.09 (0.003) | 110158 | 95067 (8191) |

Table 2: Simulation results: comparison of mean (std) true and estimated (ˆ) model parameters and derived quantities. Results for each line, for specified heritability $h^2$ and fraction $\pi_1$ of causal SNPs, are from 10 independent instantiations with random selection of the $n_{causal}$ causal SNPs that are assigned a $\beta$-value from the standard normal distribution. Defining $Y_g = G\beta$, where $G$ is the genotype matrix, the total phenotype vector is constructed as $Y{=}Y_g{+}\varepsilon$, where the residual random vector $\varepsilon$ for each instantiation is drawn from a normal distribution such that $\text{var}(Y) = \text{var}(Y_g)/h^2$ for predefined $h^2$. For each of the instantiations, $i$, this implicitly defines the true value $\sigma^2_{\beta i}$, and $\sigma^2_\beta$ is their mean. An example QQ plot for each line entry is shown in Supplementary Material, Figure S3.

onset Alzheimer's disease and Crohn's disease are almost thirty times smaller than that of schizophrenia.

In Supplementary Material Figure S7 we show results for Alzheimer's disease exclusively for chromosome 19 (which contains APOE), and for all autosomal chromosomes excluding chromosome 19. We also show results with the same chromosomal breakdown for a recent GWAS involving 455,258 samples that included 24,087 clinically diagnosed LOAD cases and 47,793 AD-by-proxy cases (individuals who were not clinically diagnosed with LOAD but for whom at least one parent had LOAD) [66]. These GWAS give consistent estimates of polygenicity: $\pi_1 \sim 1 \times 10^{-4}$ excluding chromosome 19, and $\pi_1 \sim 6 \times 10^{-5}$ for chromosome 19 exclusively.

Of the quantitative traits, educational attainment has the highest model polygenicity, $\pi_1 = 3.2 \times 10^{-3}$, similar to intelligence, $\pi_1 = 2.2 \times 10^{-3}$. Approximately two orders of magnitude lower in polygenicity are the endophenotypes putamen volume and low- and high-density lipoproteins.

The model effective SNP discoverability for schizophrenia is $\hat{\sigma}^2_\beta = 5.51 \times 10^{-5}$, similar to that for bipolar disorder. Major depressive disorder, which has the highest polygenicity, has the lowest SNP discoverability, approximately one-eighth that of schizophrenia; it is this low value, combined with high polygenicity that leads to the weak signal in Figure 1 (A) even though the sample size is relatively large. In contrast, SNP discoverability for Alzheimer's disease is almost four times that of schizophrenia. The inflammatory bowel diseases, however, have much higher SNP discoverabilities, 16 and 31 times that of schizophrenia respectively for ulcerative colitis and Crohn's disease – the latter having the second highest value of the sixteen phenotypes: $\hat{\sigma}^2_\beta = 1.7 \times 10^{-3}$.

Additionally, for Alzheimer's disease we show in Supplementary Material Figure S7 that the discoverability is two orders of magnitude greater for chromosome 19 than for the remainder of the autosome. Note that since two-thirds of the 2018 "cases" are AD-by-proxy, the discoverabilities for the 2018 data are, as expected, reduced relative to the values for the 2013 data (approximately 3.5 times smaller).

The narrow sense SNP heritability from the ascertained case-control schizophrenia GWAS is estimated as $h^2{=}0.37$. Taking adult population prevalence of schizophrenia to be $K{=}0.01$ [67, 68] (but see also [69], for $K{=}0.005$), and given that there are 51,900 cases and 71,675 controls in the study, so that the proportion of cases in the study is $P{=}0.42$, the heritability on the liability scale for schizophrenia from Eq. 39 is $\hat{h}^2_l{=}0.21$. For bipolar disorder, with $K{=}0.005$ [53], 20,352 cases and 31,358 controls, $\hat{h}^2_l{=}0.16$. Major depressive disorder appears to have a much lower model-estimated SNP heritability than schizophrenia: $\hat{h}^2_l = 0.07$. The model estimate of SNP heritability for height is 17%, lower than the oft-reported value ∼50% (see Discussion). However, despite the huge differences in sample size, we find the same value, 17%, for the 2010 GWAS ($N{=}133{,}735$ [65]), and 19% for the 2018 GWAS ($N = 707{,}868$ [49, 46]) – see Table 1.

Figure 3 shows the sample size required so that a given proportion of chip heritability is captured by genome-wide significant SNPs for the phenotypes (assuming equal numbers of cases and controls for the qualitative phenotypes: $N_{eff} = 4/(1/N_{cases} + 1/N_{controls})$, so that when $N_{cases} = N_{controls}$, $N_{eff} = N_{cases} + N_{controls} = N$, the total sample size, allowing for a straightforward comparison with quantitative traits). At current sample sizes, only 4% of narrow-sense chip heritability is captured for schizophrenia

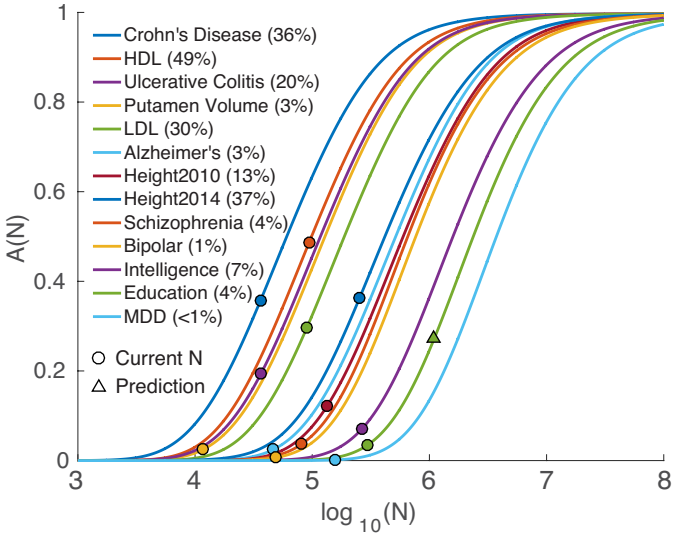


Figure 3: Proportion of narrow-sense chip heritability, $A(N)$ (Eq. 37), captured by genome-wide significant SNPs as a function of sample size, $N$, for phenotypes shown in Figures 1 and Figure 2. Values for current sample sizes are shown in parentheses. Left-to-right curve order is determined by decreasing $\sigma_\beta^2$. The prediction for education at sample size N=1.1 million is $A(N) = 0.27$, so that the proportion of phenotypic variance explained is predicted to be 3.5%, in good agreement with 3.2% reported in [70]. (The curve for AD excludes the APOE locus. For HDL, see Supplementary Material for additional notes.)

and only 1% for bipolar disorder; using current methodologies, a sample size of $N_{eff} \sim 1$ million would be required to capture the preponderance of SNP heritability for these phenotypes. Major depressive disorder GWAS currently is greatly under-powered, as shown in Figure 3(A). For education, we predict that 3.5% of phenotypic variance would be explained at $N = 1.1$ million, in good agreement with the value found from direct computation of 3.2% [70]. For other phenotypes, the proportions of total SNP heritability captured at the available sample sizes are given in Figure 3.

The sample size for ALS was quite low, and we restricted the analysis to chromosome 9, which had most of the genome-wide significant typed SNPs; we estimate that there are ∼7 causal SNPs with high discoverability on chromosome 9 [71, 72], with very high discoverability, $\sigma_\beta^2 \simeq 0.003$. In contrast, for AD restricted to chromosome 19, there were an estimated 14 causal SNPs with discoverability $\sigma_\beta^2 \simeq 0.02$ (see Supplementary Material Figure S7 (B)).

In this study, we assume that population stratification in the raw data has been corrected for in the publicly-available summary statistics. However, given that some of the sample sizes are extremely large, we allow for the possibility of residual cryptic relatedness. This would result in a scaling of the z-scores, Eq. 9 [19]. Thus, to test the modeling of inflation due to cryptic relatedness, we scaled the simulation z-scores as described earlier ($z = \sigma_0 z_u$ with $\sigma_0 > 1$, where $z_u$ are the original z-scores, i.e., not artificially inflated) and reran the model. E.g., for education and schizophrenia we inflated the z-scores by a factor of 1.2. For schizophrenia we found $\sigma_0^2 = 1.366$, which is almost exactly as predicted ($1.14 \times 1.2 = 1.368$), while the polygenicity and discoverability parameters are essentially unchanged: $\pi_1 = 2.81 \times 10^{-3}$, and $\sigma_\beta^2 = 5.56 \times 10^{-5}$. For education we found $\sigma_0^2 = 1.206$, which again is almost exactly as predicted ($1.0 \times 1.2 = 1.2$), while the polygenicity and discoverability parameters are again essentially unchanged: $\pi_1 = 3.19 \times 10^{-3}$, and $\sigma_\beta^2 = 1.57 \times 10^{-5}$.

A comparison of our results with those of [14] and [15] is in Supplementary Material Table S4. Critical methodological differences with model M2 in [14] are that we use a full reference panel of 11 million SNPs from 1000 Genomes Phase 3, we allow for the possibility of inflation in the data, and we provide an exact solution, based on Fourier Transforms, for the z-score pdf arising from the posited distribution of causal effects, resulting in better fits of the model and the data QQ plots – as can be seen by comparing our QQ plots with those reported in S4. Although our estimated number of causal are often within a factor of two of those from the nominally equivalent model M2 of Zhang et al, there is no clear pattern to the mismatch.

### GWAS Replication

In the Supplementary Material, we provide an extensive example of testing the compatibility of summary statistics from two large bipolar disorder GWASs. Because z-scores are so noisy, it is possible for a typed SNP with a highly significant p-values in one GWAS to completely fail to reach significance in a subsequent GWAS, and for these outcomes to be statistically consistent. SNP heterozygosity and total LD, as well as sample sizes, are relevant in making such assessments.

### Dependence on Reference Panel

Given a liberal MAF threshold of 0.002, our reference panel should contain the vast majority of common SNPs for European ancestry. However, it does not include other structural variants (such as small insertions/deletions, or haplotype blocks) which may also be causal for phenotypes. To validate our parameter estimates for an incomplete reference, we reran our model on real phenotypes using the culled reference where we exclude every other SNP. The result is that all estimated parameters are as before except that $\hat{\pi}_1$ doubles, leaving the estimatde number of causal SNPs and heritability as before. For example, for schizophrenia we get $\pi_1 = 5.3 \times 10^{-3}$ and $\sigma_\beta^2 = 5.8 \times 10^{-5}$ for the reduced reference panel, versus $\pi_1 = 2.8 \times 10^{-3}$ and $\sigma_\beta^2 = 5.5 \times 10^{-5}$ for the full panel, with heritability remaining essentially the same (37% on the observed scale).

## DISCUSSION

Here we present a unified method based on GWAS sum-

mary statistics, incorporating detailed LD structure from an underlying reference panel of SNPs with MAF>0.002, for estimating: phenotypic polygenicity, $\pi_1$, expressed as the fraction of the reference panel SNPs that have a non-null true $\beta$ value, i.e., are "causal"; and SNP discoverability or mean strength of association (the variance of the underlying causal effects), $\sigma^2_\beta$. In addition the model can be used to estimate residual inflation of the association statistics due to variance distortion induced by cryptic relatedness, $\sigma^2_0$. The model assumes that there is very little, if any, inflation in the GWAS summary statistics due to population stratification (bias shift in z-scores due to ethnic variation).

We apply the model to sixteen diverse phenotypes, eight qualitative and eight quantitative. From the estimated model parameters we also estimate the number of causal common-SNPs in the underlying reference panel, $n_{causal}$, and the narrow-sense common-SNP heritability, $h^2$ (for qualitative phenotypes, we re-express this as the proportion of population variance in disease liability, $h^2_l$, under a liability threshold model, adjusted for ascertainment); in the event rare SNPs (i.e., not in the reference panel) are causal, $h^2$ will be an underestimate of the true SNP heritability. In addition, we estimate the proportion of SNP heritability captured by genome-wide significant SNPs at current sample sizes, and predict future sample sizes needed to explain the preponderance of SNP heritability.

We find that schizophrenia is highly polygenic, with $\pi_1 = 2.8 \times 10^{-3}$. This leads to an estimate of $n_{causal} \simeq 31,000$, which is in reasonable agreement with a recent estimate that the number of causal SNPs is >20,000 [73]. The SNP associations, however, are characterized by a narrow distribution, $\sigma^2_\beta = 6.27 \times 10^{-5}$, indicating that most associations are of weak effect, i.e., have low discoverability. Bipolar disorder has similar parameters. The smaller sample size for bipolar disorder has led to fewer SNP discoveries compared with schizophrenia. However, from Figure 3, sample sizes for bipolar disorder are approaching a range where rapid increase in discoveries becomes possible. For educational attainment [74, 42, 75], the polygenicity is somewhat greater, $\pi_1 = 3.2 \times 10^{-3}$, leading to an estimate of $n_{causal} \simeq 35,000$, half a recent estimate, $\simeq 70,000$, for the number of loci contributing to heritability [74]. The variance of the distribution for causal effect sizes is a quarter that of schizophrenia, indicating lower discoverability. Intelligence, a related phenotype [43, 76], has a larger discoverability than education while having lower polygenicity ($\sim 10,000$ fewer causal SNPs).

In marked contrast are the lipoproteins and putamen volume which have very low polygenicity: $\pi_1 < 5 \times 10^{-5}$, so that only 250 to 550 SNPs (out of ~11 million) are estimated to be causal. However, causal SNPs for putamen volume and HDL appear to be characterized by relatively high discoverability, respectively 17-times and 23-times larger than for schizophrenia (see Supplementary Material for additional notes on HDL).

The QQ plots (which are sample size dependent) reflect these differences in genetic architecture. For example, the early departure of the schizophrenia QQ plot from the null line indicates its high polygenicity, while the steep rise for putamen volume after its departure corresponds to its high SNP discoverability.

For Alzheimer's disease, our estimate of the liability-scale SNP heritability for the full 2013 dataset [39] is 15% for prevalence of 14% for those aged 71 older, half from APOE, while the recent "M2" and "M3" models of Zhang et al [14] gave values of 7% and 10% respectively – see Supplementary Materials Table S4. A recent report from two methods, LD Score Regression (LDSC) and SumHer [77], estimated SNP heritability of 3% for LDSC and 12% for SumHer (assuming prevalence of 7.5%). A raw genotype-based analysis (GCTA), including genes that contain rare variants that affect risk for AD, reported SNP heritability of 53% [78, 7]; an earlier related study that did not include rare variants and had only a quarter of the common variants estimated SNP heritability of 33% for prevalence of 13% [79]. GCTA calculations of heritability are within the domain of the so-called infinitesimal model where all markers are assumed to be causal. Our model suggests, however, that phenotypes are characterized by polygenicities less than $5 \times 10^{-3}$; for AD the polygenicity is $\simeq 10^{-4}$. Nevertheless, the GCTA approach yields a heritability estimate closer to the twin-based (broad sense) value, estimated to be in the range 60-80% [80]. The methodology appears to be robust to many assumptions about the distribution of effect sizes [81, 82]; the SNP heritability estimate is unbiased, though it has larger standard error than methods that allow for only sparse causal effects [65, 83]. For the 2013 data analyzed here [39], a summary-statistics-based method applied to a subset of 54,162 of the 74,046 samples gave SNP heritability of almost 7% on the observed scale [84, 12]; our estimate is 12% on the observed scale – see Supplementary Material Figure S7 A and B.

Onset and clinical progression of sporadic Alzheimer's disease is strongly age-related [85, 86], with prevalence in differential age groups increasing at least up through the early 90s [57]. Thus, it would be more accurate to assess heritability (and its components, polygenicity and discoverability) with respect to, say, five-year age groups beginning with age 65 years, and using a consistent control group of nonagenarians and centenarians. By the same token, comparisons among current and past AD GWAS are complicated because of potential differences in the age distributions of the respective case and the control cohorts. Additionally, the degree to which rare variants are included will affect heritability estimates. The summary-statistic-based estimates of polygenicity that we report here are, however, likely to be robust for common SNPs: $\pi_1 \simeq 1.1 \times 10^{-4}$, with only a few causal SNPs on chromosome 19.

Our point estimate for the liability-scale SNP heritability of schizophrenia is $h^2_l = 0.21$ (assuming a population risk of 0.01), and that 4% of this (i.e., 1% of overall disease

liability) is explainable based on common SNPs reaching genome-wide significance at the current sample size. This $h_l^2$ estimate is in reasonable agreement with a recent result, $h_l^2 = 0.27$ [73, 87], also calculated from the PGC2 data set but using raw genotype data for 472,178 markers for a subset of 22,177 schizophrenia cases and 27,629 controls of European ancestry; and with an earlier result of $h_l^2 = 0.23$ from PGC1 raw genotype data for 915,354 markers for 9,087 schizophrenia cases and 12,171 controls [88, 7]. The recent "M2" (single non-null Gaussian) model estimate is $h_l^2 = 0.29$ [14] (see Supplementary Materials Table S4). No QQ plot was available for the M2 model fit to schizophrenia data, but such plots (truncated on the y-axis at $-\log_{10}(p) = 10$) for many other phenotypes were reported [14]. We note that for multiple phenotypes (height, LDL cholesterol, total cholesterol, years of schooling, Crohn's disease, coronary artery disease, and ulcerative colitis) our single causal Gaussian model appears to provide a better fit to the data than M2: many of the M2 plots show a very early and often dramatic deviation between prediction and data, as compared with our model QQ plots which are also built from a single causal Gaussian, suggesting an upward bias in polygenicity and/or variance of effect sizes, and hence heritability as measured by the M2 model for these phenotypes. The LDSC liability-scale (1% prevalence) SNP heritability for schizophrenia has been reported as $h_l^2 = 0.555$ [12] and more recently as 0.19 [77], in very good agreement with our estimate; on the observed scale it has been reported as 45% [84, 12], in contrast to our corresponding value of 37%. Our estimate of 1% of overall variation on the liability scale for schizophrenia explainable by genome-wide significant loci compares reasonably with the proportion of variance on the liability scale explained by Risk Profile Scores (RPS) reported as 1.1% using the "MGS" sample as target (the median for all 40 leave-one-out target samples analyzed is 1.19% – see Extended Data Figure 5 and Supplementary Tables 5 and 6 in [36]; this was incorrectly reported as 3.4% in the main paper). These results show that current sample sizes need to increase substantially in order for RPSs to have predictive utility, as the vast majority of associated SNPs remain undiscovered. Our power estimates indicate that ∼500,000 cases and an equal number of controls would be needed to identify these SNPs (note that there is a total of approximately 3 million cases in the US alone).

A subtle but important issue is downward bias of large-sample maximum-likelihood estimates of SNP heritability, due to over-ascertainment of cases in case-control studies [87]; it has been examined in the context of restricted maximum likelihood (REML) in GCTA, which assumes a polygenicity of 1, i.e., every SNP is causal. For schizophrenia, this has been assessed in the context of BOLT-REML, which assumes a mixture distribution of small ('spike') and large ('slab') effects [73]: from 22,177 cases and 27,629 controls, the observed-scale heritability is reported as $h^2 = 0.415$, equivalent to $h_l^2 = 0.23$ on the liability scale, assuming 1% disease prevalence. However, using "phenotype correlation-genetic correlation" (PCGC) regression, a moments-based approach requiring raw-genotype data which produces unbiased estimates for case-control studies of disease traits [87], the unbiased liability-scale heritability is reported as $h_g^2 = 0.27$, indicating that the likelihood-maximization estimate is biased down by 15% of the unbiased value (the degree of underestimation decreases for smaller sample sizes). Our estimate for the liability-scale heritability of schizophrenia, from a larger sample than in [73], is $h_l^2 = 0.21$. This at least would be consistent with downward bias operating in point-normal causal distributions, in a manner similar to that in GCTA and BOLT-REML. This would then translate into either an underestimate of the number of causal SNPs, or more likely an underestimate of the variance of the distribution of causal effects.

For educational attainment, we estimate SNP heritability $h^2 = 0.12$, in good agreement with the estimate of 11.5% given in [42]. As with schizophrenia, this is substantially less than the estimate of heritability from twin and family studies of $\simeq 40\%$ of the variance in educational attainment explained by genetic factors [89, 74].

For putamen volume, we estimate the SNP heritability $h^2 = 0.11$, in reasonable agreement with an earlier estimate of 0.1 for the same overall data set [47, 4]. For LDL and HDL, we estimate $h^2 = 0.06$ and $h^2 = 0.07$ respectively, in good agreement with the LDSC estimates $h^2 = 0.08$ and $h^2 = 0.07$ [77], and the M2 model of [14] – see Supporting Material Table S4.

For height (N=251,747 [46]) we find that its model polygenicity is $\pi_1 = 5.66 \times 10^{-4}$, a quarter that of intelligence, while its discoverability is five times that of intelligence, leading to a SNP heritability of 17%. The number of causal SNPs (out of a total of approximately 11 million) is approximately 6k; although this is about one twentieth the estimate reported in [90], it remains large and allows for height to be interpreted as "omnigenic". For the 2010 GWAS (N=133,735 [65]) and 2018 GWAS (N=707,868 [49]), we estimate SNP heritability of 17% and 19% respectively (see Table 1 and Supplementary Material Fig. S8). These heritabilities are in considerable disagreement with the SNP heritability estimate of $\simeq$50% [46] (average of estimates from five cohorts ranging in size from N=1,145 to N=5,668, with ∼1 million SNPs). For the 2010 GWAS, the M2 model [14] gives $h^2 = 0.30$ (see Supporting Material Table S4); the upward deviation of the model QQ plot in [14] suggests that this value might be inflated. For the 2014 GWAS, the M3 model estimate is $h^2 = 33\%$ [14]; the Regression with Summary Statistics (RSS) model estimate is $h^2 = 52\%$ (with $\simeq 11,000$ causal SNPs) [91], which, not taking any inflation into account is definitely a model overestimate; and in [77] the LDSC estimate is reported as $h^2 = 20\%$ while the SumHer estimate is $h^2 = 46\%$ (in general across traits, the SumHer heritability estimates tend to be two-to-five times larger than the LDSC estimates). The M2, M3, and RSS models use

a reference panel of ∼1 million common SNPs, in contrast with the ∼11 million SNPs used in our analysis. Also, it should be noted that the M2, M3, and RSS model estimates did not take the possibility of inflation into account. For the 2014 height GWAS, that inflation is reported as the LDSC intercept is 2.09 in [77], indicating considerable inflation; for the 2018 dataset we find $\sigma_0^2$=2.5, while the LD score regression intercept is 2.1116 (se 0.0458). Given the various estimates of inflation and the controversy over population structure in the height data [50, 51], it is not clear what results are definitely incorrect.

Our power analysis for height (2014) shows that 37% of the narrow-sense heritability arising from common SNPs is explained by genome-wide significant SNPs ($p \leq 5\times10^{-8}$), i.e., 6.3% of total phenotypic variance, which is substantially less than the 16% direct estimate from significant SNPs [46]. It is not clear why these large discrepancies exist. One relevant factor, however, is that we estimate a considerable confounding ($\sigma_0^2$=1.66) in the height 2014 dataset. Our $h^2$ estimates are adjusted for the potential confounding measured by $\sigma_0^2$, and thus they represent what is likely a lower bound of the actual SNP-heritability, leading to a more conservative estimate than what has previously been reported. We note that after adjustment, our $h^2$ estimates are consistent across all three datasets (height 2010, 2014 and 2018), which otherwise would range by more than 2.5-fold. Another factor might be the relative dearth of typed SNPs with low heterozygosity and low total LD (see top left segment in Supporting Material Figure S21, $n = 780$): there might be many causal variants with weak effect that are only weakly tagged. Nevertheless, given the discrepancies noted above, caution is warranted in interpreting our model results for height.

## CONCLUSION

The common-SNP causal effects model we have presented is based on GWAS summary statistics and detailed LD structure of an underlying reference panel, and assumes a Gaussian distribution of effect sizes at a fraction of SNPs randomly distributed across the autosomal genome. While not incorporating the effects of rare SNPs, we have shown that it captures the broad genetic architecture of diverse complex traits, where polygenicities and the variance of the effect sizes range over orders of magnitude.

The current model (essentially Eq. 4) and its implementation (essentially Eq. 29) are basic elements for building a more refined model of SNP effects using summary statistics. Higher accuracy in characterizing causal alleles in turn will enable greater power for SNP discovery and phenotypic prediction.


### Acknowledgments

We thank the consortia for making available their GWAS summary statistics, and the many people who provided DNA samples.

### Funding

Research Council of Norway (262656, 248984, 248778, 223273) and KG Jebsen Stiftelsen; ABCD-USA Consortium (5U24DA041123).


### References


[1] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[2] Eli A Stahl, Daniel Wegmann, Gosia Trynka, Javier Gutierrez-Achury, Ron Do, Benjamin F Voight, Peter Kraft, Robert Chen, Henrik J Kallberg, Fina AS Kurreeman, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature genetics*, 44(5):483–489, 2012.

[3] Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna AE Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John RB Perry, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*, 2015.

[4] Hon-Cheong So, Miaoxin Li, and Pak C Sham. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genetic epidemiology*, 35(6):447–456, 2011.

[5] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[6] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[7] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[8] Siddharth Krishna Kumar, Marcus W Feldman, David H Rehkopf, and Shripad Tuljapurkar. Limitations of gcta as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences*, 113(1):E61–E70, 2016.

[9] Luigi Palla and Frank Dudbridge. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *The American Journal of Human Genetics*, 97(2):250–259, 2015.

[10] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[11] Jian Yang, Michael N Weedon, Shaun Purcell, Guillaume Lettre, Karol Estrada, Cristen J Willer, Albert V Smith, Erik Ingelsson, Jeffrey R O'connell, Massimo Mangino, et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812, 2011.

[12] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[13] Hyun Min Kang, Jae Hoon Sul, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[14] Yan Zhang, Guanghao Qi, Ju-Hyun Park, and Nilanjan Chatterjee. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature genetics*, 50(9):1318, 2018.

[15] Jian Zeng, Ronald Vlaming, Yang Wu, Matthew R Robinson, Luke R Lloyd-Jones, Loic Yengo, Chloe X Yap, Angli Xue, Julia Sidorenko, Allan F McRae, et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nature genetics*, 50(5):746, 2018.

[16] Bogdan Pasaniuc and Alkes L Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 2016.

[17] John S Witte, Peter M Visscher, and Naomi R Wray. The contribution of genetic variants to disease depends on the ruler. *Nature Reviews Genetics*, 15(11):765–776, 2014.

[18] D. Holland, Y. Wang, W. K. Thompson, A. Schork, C. H. Chen, M. T. Lo, A. Witoelar, T. Werge, M. O'Donovan, O. A. Andreassen, and A. M. Dale. Estimating Effect Sizes and Expected Replication Probabilities from GWAS Summary Statistics. *Front Genet*, 7:15, 2016.

[19] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, Dec 1999.

[20] Dominic Holland, Chun-Chieh Fan, Oleksandr Frei, Alexey A. Shadrin, Olav B. Smeland, V. S. Sundar, Ole A. Andreassen, and Anders M. Dale. Estimating degree of polygenicity, causal effect size variance, and confounding bias in gwas summary statistics. *bioRxiv*, 2017.

[21] Wesley K Thompson, Yunpeng Wang, Andrew Schork, Verena Zuber, Ole A Andreassen, Anders M Dale, Dominic Holland, and Xu Shujing. An empirical bayes method for estimating the distribution of effects in genome-wide association studies. *PLoS Genetics*, [in press], 2015.

[22] Jian Yang, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza De Andrade, Bjarke Feenstra, Eleanor Feingold, M Geoffrey Hayes, et al. Genome partitioning of genetic variation for complex traits using common snps. *Nature genetics*, 43(6):519–525, 2011.

[23] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[24] Nan M Laird and Christoph Lange. *The fundamentals of modern statistical genetics*. Springer Science & Business Media, 2010.

[25] Chengqing Wu, Andrew DeWan, Josephine Hoh, and Zuoheng Wang. A comparison of association methods correcting for population stratification in case–control studies. *Annals of human genetics*, 75(3):418–427, 2011.

[26] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[27] Dominic Holland. *GWAS-Causal-Effects-Model*, 2019. `https://github.com/dominicholland/GWAS-Causal-Effects-Model`.

[28] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[29] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[30] Gardar Sveinbjornsson, Anders Albrechtsen, Florian Zink, Sigurjón A Gudjonsson, Asmundur Oddson, Gísli Másson, Hilma Holm, Augustine Kong, Unnur Thorsteinsdottir, Patrick Sulem, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature genetics*, 2016.

[31] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[32] Chris CA Spencer, Zhan Su, Peter Donnelly, and Jonathan Marchini. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5(5):e1000477, 2009.

[33] Zhan Su, Jonathan Marchini, and Peter Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305, 2011.

[34] Naomi R Wray, Stephan Ripke, Manuel Mattheisen, Maciej Trzaskowski, Enda M Byrne, Abdel Abdellaoui, Mark J Adams, Esben Agerbo, Tracy M Air, Till MF Andlauer, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics*, 50(5):668, 2018.

[35] Eli Stahl, Gerome Breen, Andreas Forstner, Andrew McQuillin, Stephan Ripke, Sven Cichon, Laura Scott, Roel Ophoff, Ole A Andreassen, John Kelsoe, and Pamela Sklar. Genomewide association study identifies 30 loci associated with bipolar disorder. *bioRxiv*, 2018.

[36] Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, Jul 2014.

[37] Majid Nikpay, Anuj Goel, Hong-Hee Won, Leanne M Hall, Christina Willenborg, Stavroula Kanoni, Danish Saleheen, Theodosios Kyriakou, Christopher P Nelson, Jemma C Hopewell, et al. A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, 47(10):1121, 2015.

[38] Katrina M de Lange, Loukas Moutsianas, James C Lee, Christopher A Lamb, Yang Luo, Nicholas A Kennedy, Luke Jostins, Daniel L Rice, Javier Gutierrez-Achury, Sun-Gou Ji, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, 49(2):256, 2017.

[39] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452–1458, 2013.

[40] Iris Jansen, Jeanne Savage, Kyoko Watanabe, Julien Bryois, Dylan Williams, Stacy Steinberg, Julia Sealock, Ida Karlsson, Sara Hagg, Lavinia Athanasiu, et al. Genetic meta-analysis identifies 10 novel loci and functional pathways for alzheimer's disease risk. *bioRxiv*, page 258533, 2018.

[41] Wouter Van Rheenen, Aleksey Shatunov, Annelot M Dekker, Russell L McLaughlin, Frank P Diekstra, Sara L Pulit, Rick AA Van Der Spek, Urmo Võsa, Simone De Jong, Matthew R Robinson, et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature genetics*, 48(9):1043, 2016.

[42] Aysu Okbay, Jonathan P Beauchamp, Mark Alan Fontana, James J Lee, Tune H Pers, Cornelius A Rietveld, Patrick Turley, Guo-Bo Chen, Valur Emilsson, S Fleur W Meddens, et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604):539–542, 2016.

[43] Suzanne Sniekers, Sven Stringer, Kyoko Watanabe, Philip R Jansen, Jonathan RI Coleman, Eva Krapohl, Erdogan Taskesen, Anke R Hammerschlag, Aysu Okbay, Delilah Zabaneh, et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature genetics*, 49(7):1107, 2017.

[44] JE Savage, PR Jansen, S Stringer, K Watanabe, J Bryois, CA de Leeuw, M Nagel, S Awasthi, PB Barr, JRI Coleman, KL Grasby, AR Hammerschlag, J Kaminski, R Karlsson, et al. Genome-wiide association meta-analysis (n=269,867) identifies new genetic and functional links to intelligence. *Nature genetics*, forthcoming, 2018.

[45] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197, 2015.

[46] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common

variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[47] Derrek P Hibar, Jason L Stein, Miguel E Renteria, Alejandro Arias-Vasquez, Sylvane Desrivières, Neda Jahanshad, Roberto Toro, Katharina Wittfeld, Lucija Abramovic, Micael Andersson, et al. Common genetic variants influence human subcortical brain structures. *Nature*, 2015.

[48] Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, Jin Chen, Martin L Buchkovich, Samia Mora, et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274, 2013.

[49] Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, Jian Yang, Peter M Visscher, et al. Meta-analysis of genome-wide association studies for height and body mass index in˜ 700,000 individuals of european ancestry. *bioRxiv*, page 274654, 2018.

[50] Mashaal Sohail, Robert M Maier, Andrea Ganna, Alex Bloemendal, Alicia R Martin, Michael C Turchin, Charleston WK Chiang, Joel Hirschhorn, Mark J Daly, Nick Patterson, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife*, 8:e39702, 2019.

[51] Jeremy J Berg, Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan August Boyle, Xinjun Zhang, Fernando Racimo, Jonathan K Pritchard, et al. Reduced signal for polygenic adaptation of height in uk biobank. *eLife*, 8:e39725, 2019.

[52] NIMH. *Prevalence of Major Depressive Episode Among Adults*, 2016. (accessed December 27, 2018).

[53] Kathleen R Merikangas, Robert Jin, Jian-Ping He, Ronald C Kessler, Sing Lee, Nancy A Sampson, Maria Carmen Viana, Laura Helena Andrade, Chiyi Hu, Elie G Karam, et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Archives of general psychiatry*, 68(3):241–251, 2011.

[54] Doug Speed, Na Cai, Michael R Johnson, Sergey Nejentsev, David J Balding, UCLEB Consortium, et al. Reevaluation of snp heritability in complex human traits. *Nature genetics*, 49(7):986, 2017.

[55] Fabian Sanchis-Gomar, Carme Perez-Quilis, Roman Leischik, and Alejandro Lucia. Epidemiology of coronary heart disease and acute coronary syndrome. *Annals of translational medicine*, 4(13), 2016.

[56] Johan Burisch, Tine Jess, Matteo Martinato, Peter L Lakatos, and ECCO-EpiCom. The burden of inflammatory bowel disease in europe. *Journal of Crohn's and Colitis*, 7(4):322–337, 2013.

[57] Brenda L Plassman, Kenneth M Langa, Gwenith G Fisher, Steven G Heeringa, David R Weir, Mary Beth Ofstedal, James R Burke, Michael D Hurd, Guy G Potter, Willard L Rodgers, et al. Prevalence of dementia in the united states: the aging, demographics, and memory study. *Neuroepidemiology*, 29(1-2):125–132, 2007.

[58] Alzheimer's Association. 2018 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429, 2018.

[59] P Mehta, W Kaye, and J et al. Raymond. Prevalence of amyotrophic lateral sclerosis ˘2014 united states. *MMWR Morb Mortal Wkly Rep*, 67:216–218, 2018.

[60] Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic epidemiology*, 32(4):381–385, 2008.

[61] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369, 2008.

[62] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[63] Douglas S Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, 29(1):51–76, 1965.

[64] Everett R Dempster and I Michael Lerner. Heritability of threshold characters. *Genetics*, 35(2):212, 1950.

[65] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42(7):565–569, Jul 2010.

[66] Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer˘2019s disease risk. *Nature genetics*, page 1, 2019.

[67] Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, Pamela Sklar, Shaun M Purcell, Jennifer L Stone, Patrick F Sullivan, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

[68] Harvey A Whiteford, Louisa Degenhardt, Jürgen Rehm, Amanda J Baxter, Alize J Ferrari, Holly E Erskine, Fiona J Charlson, Rosana E Norman, Abraham D Flaxman, Nicole Johns, et al. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *The Lancet*, 382(9904):1575–1586, 2013.

[69] Dennis K Kinney, Pamela Teixeira, Diane Hsu, Siena C Napoleon, David J Crowley, Andrea Miller, William Hyman, and Emerald Huang. Relation of schizophrenia prevalence to latitude, climate, fish consumption, infant mortality, and skin color: a role for prenatal vitamin d deficiency and infections? *Schizophrenia bulletin*, page sbp023, 2009.

[70] James J Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, 50(8):1112, 2018.

[71] Rubika Balendra and Adrian M Isaacs. C9orf72-mediated als and ftd: multiple pathways to disease. *Nature Reviews Neurology*, page 1, 2018.

[72] Vitalay Fomin, Patricia Richard, Mainul Hoque, Cynthia Li, Zhuoying Gu, Mercedes Fissore-O'Leary, Bin Tian, Carol Prives, and James L Manley. The c9orf72 gene, implicated in amyotrophic lateral sclerosis and frontotemporal dementia, encodes a protein that functions in control of endothelin and glutamate signaling. *Molecular and cellular biology*, 38(22):e00155–18, 2018.

[73] Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, 2015.

[74] Cornelius A Rietveld, Sarah E Medland, Jaime Derringer, Jian Yang, Tõnu Esko, Nicolas W Martin, Harm-Jan Westra, Konstantin Shakhbazov, Abdel Abdellaoui, Arpana Agrawal, et al. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. *science*, 340(6139):1467–1471, 2013.

[75] David Cesarini and Peter M Visscher. Genetics and educational attainment. *npj Science of Learning*, 2(1):4, 2017.

[76] Robert Plomin and Sophie von Stumm. The new genetics of intelligence. *Nature Reviews Genetics*, 2018.

[77] Doug Speed and David J Balding. Sumher better estimates the snp heritability of complex traits from summary statistics. *Nature genetics*, 51(2):277, 2019.

[78] Perry G Ridge, Kaitlyn B Hoyt, Kevin Boehme, Shubhabrata Mukherjee, Paul K Crane, Jonathan L Haines, Richard Mayeux,

Lindsay A Farrer, Margaret A Pericak-Vance, Gerard D Schellenberg, et al. Assessment of the genetic variance of late-onset alzheimer's disease. *Neurobiology of aging*, 41:200–e13, 2016.

[79] Perry G Ridge, Shubhabrata Mukherjee, Paul K Crane, John SK Kauwe, et al. Alzheimer's disease: analyzing the missing heritability. *PloS One*, 8(11):e79771, 2013.

[80] Margaret Gatz, Chandra A Reynolds, Laura Fratiglioni, Boo Johansson, James A Mortimer, Stig Berg, Amy Fiske, and Nancy L Pedersen. Role of genes and environments for explaining alzheimer disease. *Archives of general psychiatry*, 63(2):168–174, 2006.

[81] Luke M Evans, Rasool Tahmasbi, Scott I Vrieze, Gonçalo R Abecasis, Sayantan Das, Steven Gazal, Douglas W Bjelland, Teresa R Candia, Michael E Goddard, Benjamin M Neale, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics*, 50(5):737, 2018.

[82] Jian Yang, Jian Zeng, Michael E Goddard, Naomi R Wray, and Peter M Visscher. Concepts, estimation and interpretation of snp-based heritability. *Nature genetics*, 49(9):1304, 2017.

[83] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 2013.

[84] Jie Zheng, A Mesut Erzurumluoglu, Benjamin L Elsworth, John P Kemp, Laurence Howe, Philip C Haycock, Gibran Hemani, Katherine Tansey, Charles Laurin, Beate St Pourcain, 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*, 33(2):272–279, 2017.

[85] Dominic Holland, Rahul S Desikan, Anders M Dale, Linda K McEvoy, Alzheimers Disease Neuroimaging Initiative, et al. Rates of decline in alzheimer disease decrease with age. *PloS one*, 7(8):e42325, 2012.

[86] Rahul S Desikan, Chun Chieh Fan, Yunpeng Wang, Andrew J Schork, Howard J Cabral, L Adrienne Cupples, Wesley K Thompson, Lilah Besser, Walter A Kukull, Dominic Holland, et al. Genetic assessment of age-associated alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS medicine*, 14(3):e1002258, 2017.

[87] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[88] S Hong Lee, Teresa R DeCandia, Stephan Ripke, Jian Yang, Patrick F Sullivan, Michael E Goddard, Matthew C Keller, Peter M Visscher, Naomi R Wray, Schizophrenia Psychiatric Genome-Wide Association Study Consortium, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. *Nature genetics*, 44(3):247–250, 2012.

[89] Amelia R Branigan, Kenneth J McCallum, and Jeremy Freese. Variation in the heritability of educational attainment: An international meta-analysis. *Social Forces*, pages 109–140, 2013.

[90] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[91] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The annals of applied statistics*, 11(3):1561, 2017.

[92] Eli A Stahl, Gerome Breen, Andreas J Forstner, Andrew McQuillin, Stephan Ripke, Vassily Trubetskoy, Manuel Mattheisen, Yunpeng Wang, Jonathan RI Coleman, Héléna A Gaspar, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics*, page 1, 2019.

[93] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707, 2010.

## Figure Legends

Figure 1: QQ plots of (pruned) z-scores for qualitative phenotypes (dark blue, 95% confidence interval in light blue) with model prediction (yellow): (A) major depressive disorder; (B) bipolar disorder; (C) schizophrenia; (D) coronary artery disease (CAD); (E) ulcerative colitis (UC); (F) Crohn's disease (CD); (G) late onset Alzheimer's disease (AD), excluding APOE (see also Supplementary Material Fig. S7); and (H) amyotrophic lateral sclerosis (ALS), restricted to chromosome 9 (see also Supporting Material Figure S9). The dashed line is the expected QQ plot under null (no SNPs associated with the phenotype). $p$ is a nominal p-value for z-scores, and $q$ is the proportion of z-scores with p-values exceeding that threshold. $\lambda$ is the overall nominal genomic control factor for the pruned data (which is accurately predicted by the model in all cases). The three estimated model parameters are: polygenicity, $\hat{\pi}_1$; discoverability, $\hat{\sigma}_\beta^2$ (corrected for inflation); and SNP association $\chi^2$-statistic inflation factor, $\hat{\sigma}_0^2$. $\hat{h}^2$ is the estimated narrow-sense chip heritability, re-expressed as $h_l^2$ on the liability scale for these case-control conditions assuming a prevalence of: MDD 6.7% [52], BIP 0.5% [53], SCZ 1% [54], CAD 3% [55], UC 0.1% [56], CD 0.1% [56], AD 14% (for people aged 71 and older in the USA [57, 58]), and ALS $5 \times 10^{-5}$ [59]. The estimated number of causal SNPs is given by $\hat{n}_{causal} = \hat{\pi}_1 n_{snp}$ where $n_{snp} = 11,015,833$ is the total number of SNPs, whose LD structure and MAF underlie the model; the GWAS z-scores are for subsets of these SNPs. $N_{eff}$ is the effective case-control sample size – see text. Reading the plots: on the vertical axis, choose a p-value threshold (more extreme values are further from the origin), then the horizontal axis gives the proportion of SNPs exceeding that threshold (higher proportions are closer to the origin). Numerical values for the model parameters are also given in Table 1. See also Supplementary Material Figs. S12-S18.

Figure 2: QQ plots of (pruned) z-scores and model fits for quantitative phenotypes: (A) educational attainment; (B) intelligence; (C) body mass index (BMI); (D) height; (E) putamen volume; (F) low-density lipoprotein (LDL); (G) high-density lipoprotein (HDL); and (H) total cholesterol (TC). $N$ is the sample size. See Fig. 1 for further description. Numerical values for the model parameters are also given in Table 1. See also Supplementary Material Figs. S19-S24.

Figure 3: Proportion of narrow-sense chip heritability, $A(N)$ (Eq. 37), captured by genome-wide significant SNPs as a function of sample size, $N$, for phenotypes shown in Figures 1 and Figure 2. Values for current sample sizes are shown in parentheses. Left-to-right curve order is determined by decreasing $\sigma_\beta^2$. The prediction for education at sample size N=1.1 million is $A(N) = 0.27$, so that the proportion of phenotypic variance explained is predicted to be 3.5%, in good agreement with 3.2% reported in [70]. (The curve for AD excludes the APOE locus. For HDL, see Supplementary Material for additional notes.)

## Supporting Information Legends

In Supporting Information, we provide a detailed description of the methods, as well as providing additional figures showing results for simulations and real phenotypes, and tables, including comparisons of results from other methodologies.