Beyond SNP Heritability: Polygenicity and Discoverability Estimated for Multiple Phenotypes with a Univariate Gaussian Mixture Model

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

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

Abstract

Of signal interest in the genetics of human traits is estimating their polygenicity, the proportion of causally associated single nucleotide polymorphisms (SNPs), and the effect size variance, or discoverability, of the causal SNPs. Narrow-sense heritability is proportional to the product of these quantities. Here we present the first model, using detailed linkage disequilibrium structure, to estimate polygenicity, discoverability, and narrow-sense heritability from genome-wide association studies (GWAS) summary statistics. We apply the model to multiple diverse phenotypes and validate it with extensive simulations. We find polygenicities ranging from $\simeq 2\times10^{-5}$ to $\simeq 4\times10^{-3}$, 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 at current sample sizes, and map out sample sizes required to fully explain additive SNP heritability. The model also allows for estimating – and controlling for – residual inflation in GWAS summary statistics. Thus the model is able to describe the broad genetic architecture of phenotypes where both polygenicity and discoverability range over several orders of magnitude, shows why only small proportions of heritability have been explained for discovered SNPs, and provides a roadmap for future GWAS discoveries.

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

INTRODUCTION

The genetic components of complex human traits and diseases arise from hundreds to likely many thousands of single nucleotide polymorphisms (SNPs) (Visscher et al., 2012), 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 SNPs (polygenicity) involved in any particular phenotype; their effective strength of association (discoverability); 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 (Stahl et al., 2012; Yang et al., 2015; So et al., 2011; Speed et al., 2012; Lee et al., 2011; Yang et al., 2011a; Kumar et al., 2016; Palla and Dudbridge, 2015). However, the effects of population structure (Price et al., 2010), combined with high polygenicity and linkage disequilibrium (LD), leading to spurious degrees of SNP association, or inflation, considerably complicate matters, and are also areas of much focus (Yang et al., 2011c; Bulik-Sullivan et al., 2015; Kang et al., 2010). Yet, despite recent significant advances, it has been difficult to develop a mathematical model of polygenic architecture based on GWAS that can be used for power estimated across human phenotypes.

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 (Pasaniuc and Price, 2016)) to estimate polygenicity (π_1) and discoverability (σ^2_β), as well as elevation of z-scores due to any residual inflation of the z-scores arising from variance distortion induced by cryptic relatedness (σ^2_0), which remains a concern in large-scale studies (Price et al., 2010). We estimate π_1, σ^2_β, and σ^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 (Witte et al., 2014); to pre-

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

dict 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 (Holland et al., 2016) 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 phenotypes, in each case estimating the three model parameters and auxiliary quantities, including the overall inflation factor λ, (traditionally referred to as genomic control (Devlin and Roeder, 1999)) for SNP sets, 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.

METHODS

The Model: Probability Distribution for Z-Scores

To establish notation, we note briefly that in GWAS, ignoring covariates, one traditionally performs simple linear regression in relating genotype to phenotype (Holland et al., 2016; Thompson et al., 2015). That is, assume a linear vector equation

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

for phenotype vector y over N samples (mean-centered and normalized to unit variance), mean-centered genotype vector g_i for the ith of n SNPs, true effect (regression coefficient) β_i, and residual vector e_i containing the effects of all the other causal SNPs, the independent environmental component, and random error. 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 (Yang et al., 2011c), 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 (Yang et al., 2011b)), and that their β coefficients in the GWAS framework are distributed normally with variance σ_β^2:

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

(We use the symbol β to refer to a scalar or vector, with context indicating which.) Taking into account all SNPs (the remaining ones are all null by definition), this is equivalent to the two-component Gaussian mixture model

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

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, the association z-scores for SNPs can be decomposed into an effect δ and a residual term $\epsilon \sim \mathcal{N}(0, \sigma_0^2)$, assumed to be independent (Holland et al., 2016):

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

with

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

so that

$$\begin{aligned} \text{var}(z) &= \text{var}(\delta) + \text{var}(\epsilon) \\ &\equiv \sigma^2 + \sigma_0^2 \end{aligned} \tag{7}$$

where

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

By construction, $\sigma_0^2 = 1$ unless the z-scores are elevated by variance distortion due to cryptic relatedness in the sample. When there is no genetic effect, $\delta = 0$ Implicit in Eq. 8 is approximating the denominator, $1 - q^2$, of the χ^2 statistic noncentrality 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 δ is

$$\delta = \sqrt{N}q/\sqrt{1-q^2}. \tag{9}$$

Taylor expanding in q and then taking the variance gives

$$\text{var}(\delta) = \sigma_\beta^2 NH[1 + (15/4)\sigma_\beta^4 H^2 + O(\sigma_\beta^8 H^4)]. \tag{10}$$

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 tag SNP i, $\hat{\beta}_i$, and hence the GWAS z-score, implicitly incorporates contributions due to LD with neighboring causal SNPs. (A tag 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 tag 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, β_j is its true regression coefficient, and ε is the independent true environmental and error residual vector. Thus, explicitly including all causal true β's, Eq. 2 becomes

$$\begin{aligned} \hat{\beta}_i &= \frac{\sum_j g_i^T g_j\beta_j}{NH_i} + \frac{g_i^T\varepsilon}{g_i^T g_i} \\ &\equiv \beta_i' + \varepsilon_i' \end{aligned} \tag{11}$$

(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 latent (true) causal effects and the latent environmental (plus error) component; β_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, β_i' in Eq. 11 can be written as

$$\beta_i' = \sum_j \sqrt{\frac{H_j}{H_i}} r_{ij}\beta_j. \tag{12}$$

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

$$\begin{aligned} z_i &= \sqrt{NH_i}\beta_i' + \epsilon \\ &= \sqrt{N}\sum_j \sqrt{H_j} r_{ij}\beta_j + \epsilon. \end{aligned} \tag{13}$$

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 σ^2 will be given by

$$\sigma^2 = kr^2\sigma_\beta^2 NH. \tag{14}$$

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

$$f_1(z; N, \mathcal{H}, \sigma_\beta, \sigma_0) = \phi(z; 0, kr^2\sigma_\beta^2 NH + \sigma_0^2) \tag{15}$$

where $\phi(\cdot, \mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2, and $\mathcal{H}$ is shorthand for the LD and heterozygosity structure of such SNPs (in this case, denoting exactly k causals with LD given by r^2 and heterozygosity given by H). If a proportion α of all tag 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

$$f(z) = (1-\alpha)\phi(z; 0, \sigma_0^2) + \alpha f_1(z), \tag{16}$$

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 tag 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 tag 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 should be accurate for sufficiently fine binning. 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.

Model PDF: Multinomial Expansion

As in our previous work, we incorporate the model parameter π_1 for the fraction of all SNPs that are causal (Holland et al., 2016). 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. 15 for different combinations of numbers of causals among the w LD windows, each Gaussian scaled by the probability of the corresponding combination of causals among the LD windows, i.e., by the appropriate multinomial distribution term.

For w r^2-windows, we must consider the possibilities where the tag 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 π_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 causals ($\sum_{i=1}^w k_i = K$ 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

$$M(k_0, ..., k_w; n_0, ..., n_w; \pi_1) = \frac{n!}{k_0!...k_w!} p_0^{k_0}...p_w^{k_w} \tag{17}$$

and the corresponding Gaussian is

$$\phi(z; 0, (k_1 H_1 r_1^2 + ... + k_w H_w r_w^2)\sigma_\beta^2 N + \sigma_0^2). \tag{18}$$

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 causals in LD with the SNP, and all possible distributions of those causals among the w r^2-windows:

$$\text{pdf}(z; N, \mathcal{H}, \pi_1, \sigma_\beta, \sigma_0) = \sum_{K=0}^{K_{max}} \sum_{k_1,...,k_w} \frac{n!}{k_0!...k_w!} p_0^{k_0} ... p_w^{k_w} \times \phi(z; 0, (k_1 H_1 r_1^2 + ... + k_w H_w r_w^2)\sigma_\beta^2 N + \sigma_0^2), \quad (19)$$

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 π_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 π_1 is small (often $\leq 10^{-3}$), the upper bound on the first summation over total number of potential causals 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,

$$\sum_{K=0}^{K_{max}} \sum_{k_1,...,k_w} M(k_0, ..., k_w; n_0, ..., n_w; \pi_1) \simeq 1. \quad (20)$$

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

Model PDF: Convolution

From Eq. 13, there is an alternative and 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. 13 z is a sum of correlation- and heterozygosity-weighted random variables $\{\beta_j\}$ and the random variable ϵ, where $\{\beta_j\}$ denotes the set of true causal parameters for each of the SNPs in LD with the tag 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 (2), for each SNP j in LD with the tag SNP ($1 \leq j \leq b$, where b is the tag SNP's block size),

$$\sqrt{NH_j} r_j \beta_j \sim \pi_1 \mathcal{N}(0, NH_j r_j^2 \sigma_\beta^2) + (1 - \pi_1)\mathcal{N}(0, 0). \quad (21)$$

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

$$\pi_1 \exp(-2\pi^2 k^2 N H_j r_j^2 \sigma_\beta^2),$$

while that of the second term is simply $(1 - \pi_1)$. Additionally, the environmental term is $\epsilon \sim \mathcal{N}(0, \sigma_0^2)$, and its Fourier transform is $\exp(-2\pi^2 \sigma_0^2 k^2)$. For each tag 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 tag 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 griding instead of with respect to every individual tag SNP z-score; in the section "Parameter Estimation" below we describe using a 10×10 grid. Additionally, for each H-L bin the LD r^2 and heterozygosity histogram structure for each tag SNP is built, using w_{max} equally-spaced r^2 bins for $r_{min}^2 \leq r^2 \leq 1$; $w_{max} = 20$ is large enough to allow for converged results; $r_{min}^2 = 0.05$ 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 σ_β^2 (high heritability with low polygenicity) that we do not encounter for the 12 phenotypes we analyze here. In any case, this can be calibrated for using simulations.

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 tag SNPs having n_w SNPs (not all of which necessarily are tag SNPs) with LD r^2 in the wth r^2 bin whose average heterozygosity is H_w. Thus, one can re-express z-scores for an H-L bin as

$$z = \sqrt{N} \sum_{w=1}^{w_{max}} \left(\sqrt{H_w} r_w \sum_{j=0}^{n_w} \beta_j \right) + \epsilon \quad (22)$$

where β_j and ϵ 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 Δz ($z_n = -z_1 - \Delta z$, and $z_{n/2+1} = 0$). Taking $z_1 = -38$ allows for the minimum p-values of 5.8×10^{-316} (near the numerical limit); with $n = 2^{10}$, $\Delta z = 0.0742$. Given Δz, 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 Δk, where $k_1 = -f_c$ ($k_n = -k_1 - \Delta k$, and $k_{n/2+1} = 0$; the DFT pair Δz and Δk are related by $\Delta z \Delta k = 1/n$). Define

$$A_w \equiv -2\pi^2 N H_w r_w^2 \sigma_\beta^2. \tag{23}$$

(see Eq. 21). 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

$$G(k_j) = \prod_{w=1}^{w_{max}} \left(\pi_1 \exp(A_w k_j^2) + (1 - \pi_1)\right)^{n_w}, \tag{24}$$

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

$$E(k_j) = \exp(-2\pi^2 \sigma_0^2 k_j^2). \tag{25}$$

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

$$\mathbf{pdf_z} = \mathcal{F}^{-1}\left[\mathbf{F_z}\right]. \tag{26}$$

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 (Consortium et al., 2015, 2012; Sveinbjornsson et al., 2016). For simulations, we used HAPGEN2 (Li and Stephens, 2003; Spencer et al., 2009; Su et al., 2011) 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 for which more than 99% of genotypes were identical; 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.

Sequentially moving through each chromosome in contiguous blocks of 5,000 SNPs, for each SNP in the block we calculated its Pearson r^2 correlation coefficients with all SNPs in the central bock itself and with all SNPs in the pair of flanking blocks of size up to 50,000 each. For each SNP we calculated its total LD (TLD), given by the sum of LD r^2's thresholded such that if $r^2 < r^2_{min}$ we set that r^2 to zero. For each SNP we also built a histogram giving the numbers of SNPs in w_{max} equally-spaced r^2-windows covering the range $r^2_{min} \leq r^2 \leq 1$. These steps were carried out independently for 1000 Genomes phase 3 and for HAPGEN2 (for the latter, we used 1000 simulated samples).

Employing a similar procedure, we also built binary (logical) LD matrices identifying all pairs of SNPs for which LD $r^2 > 0.8$, a liberal threshold for SNPs being "synonymous".

In applying the model to summary statistics, we calculated histograms of TLD and LD block size (using 100 bins in both cases) and ignoring SNPs whose TLD or block size was so large that their frequency was less than a hundredth of the respective histogram peak; typically this amounted to restricting to SNPs for which TLD ≤ 600 and LD block size $\leq 1,500$. We also ignored SNPs for which MAF ≤ 0.01,

We analyzed summary statistics for twelve phenotypes (in what follows, where sample sizes varied by SNP, we quote the median value): (1) bipolar disorder ($N_{cases} = 20,352$, $N_{controls} = 31,358$) (Stahl et al., 2018); (2) schizophrenia ($N_{cases} = 35,476$, $N_{controls} = 46,839$) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014); (3) major depressive disorder ($N_{cases} = 59,851$, $N_{controls} = 113,154$) (Wray and Sullivan, 2017); (4) late onset Alzheimer's disease (LOAD; $N_{cases} = 17,008$, $N_{controls} = 37,154$) (Lambert et al., 2013) (in the Supplementary Material we present results for a more recent GWAS with $N_{cases} = 71,880$ and $N_{controls} = 383,378$ (Jansen et al., 2018)); (5) Crohn's disease (N_{cases} = 12,194, $N_{controls} = 34,915$); (6) ulcertive colitis ($N_{cases} = 12,366$, $N_{controls} = 34,915$) (de Lange et al., 2017); (7) number of years of formal education ($N = 293,723$) (Okbay et al., 2016); (8) intelligence ($N = 262,529$) (Sniekers et al., 2017; Savage et al., 2018); (9) low- ($N = 89,873$) and (10) high-density lipoprotein ($N = 94,295$) (Willer et al., 2013); (11) height ($N = 707,868$) (Yengo et al., 2018; Wood et al., 2014); and (12) putamen volume (normalized by intracranial volume, $N = 11,598$) (Hibar et al., 2015). Most participants were of European ancestry.

For schizophrenia, for example, there were 6,610,991 SNPs with finite z-scores out of the 11,015,833 SNPs from the 1000 Genomes reference panel that underlie the model; the genomic control factor for these SNPs was $\lambda_{GC} = 1.466$. Of these, 314,857 were filtered out due to low maf or very large LD block size. The genomic control factor for the remaining SNPs was $\lambda_{GC} = 1.468$; for the pruned subsets, with $\simeq 1.49 \times 10^6$ SNPs each, it was $\lambda = 1.30$. (Note that genomic control values for pruned data are always lower than for unpruned data.)

Simulations

We generated genotypes for 10^5 unrelated simulated samples using HAPGEN2 (Su et al., 2011). For narrow-sense heritability h^2 equal to 0.1, 0.4, and 0.7, we considered polygenicity π_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 β-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 β's. Defining $Y_g = G\beta$, where G is the genotype matrix and β here is the vector of coefficients over all SNPs, the total phenotype vector is con-

structed as $Y{=}Y_g{+}\varepsilon$, where the residual random vector ε for each instantiation is drawn from a normal distribution such that $h^2 = \mathrm{var}(Y_g)/\mathrm{var}(Y)$. For each of the instantiations this implicitly defines the "true" value σ^2_β.

The regression slope, β, and the Pearson correlation coefficient, r, are assumed to be t-distributed. These quantities have the same t-value: $t = \beta/\mathrm{se}(\beta) = r/\mathrm{se}(r) = r\sqrt{N-2}/\sqrt{1-r^2}$, with corresponding p-value from Student's t cumulative distribution function (cdf) with $N-2$ degrees of freedom: $p = 2{\times}\mathrm{tcdf}(-|t|, N-2)$. 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: $z = -\Phi^{-1}(p/2) \times \mathrm{sign}(r)$, where Φ is the normal cdf (z and t have 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 10×10-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. 26 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 N, $\mathcal{H}$, and the model parameters) were calculated using numerical integration over the random variable δ:

$$\begin{aligned}\delta_{expected} \equiv E(\delta|z) &= \int P(\delta|z)\delta d\delta \\ &= \frac{1}{P(z)}\int P(z|\delta)P(\delta)\delta d\delta. \qquad (27)\end{aligned}$$

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

$$P(z|\delta) = \phi(z; \delta, \sigma_0^2). \qquad (28)$$

$P(z)$ is shorthand for $\mathrm{pdf}(z|N, \mathcal{H}, \pi_1, \sigma_\beta, \sigma_0)$, given by Eq. 26. $P(\delta)$ is calculated by a similar procedure that lead to Eq. 26 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 $f_i \equiv \mathrm{pdf}(\delta_i|\mathcal{H})$, is

$$\mathbf{pdf}_\delta = \mathcal{F}^{-1}\left[\mathbf{F}_\delta\right]. \qquad (29)$$

Similarly,

$$\begin{aligned}\delta^2_{expected} \equiv E(\delta^2|z) &= \int P(\delta|z)\delta^2 d\delta \\ &= \frac{1}{P(z)}\int P(z|\delta)P(\delta)\delta^2 d\delta, \qquad (30)\end{aligned}$$

which is used in power calculations.

GWAS Power

It is of interest to estimate the proportion of additive phenotypic variance arising from the n_{snp} SNPs under study (the chip heritability (Witte et al., 2014)) that can be explained by SNPs reaching genome-wide significance, $p{\leq}5{\times}10^{-8}$ (i.e., for which $|z|{>}z_t{=}5.45$) at a given sample size (Pe'er et al., 2008; McCarthy et al., 2008). For a SNP with genotype vector g (over N samples) and heterozygosity H, one has $\mathrm{var}(Y|g){=}\mathrm{var}(\beta g){=}2\beta^2 H$ and $\delta{=}\sqrt{NH}\beta$. 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. For each z from

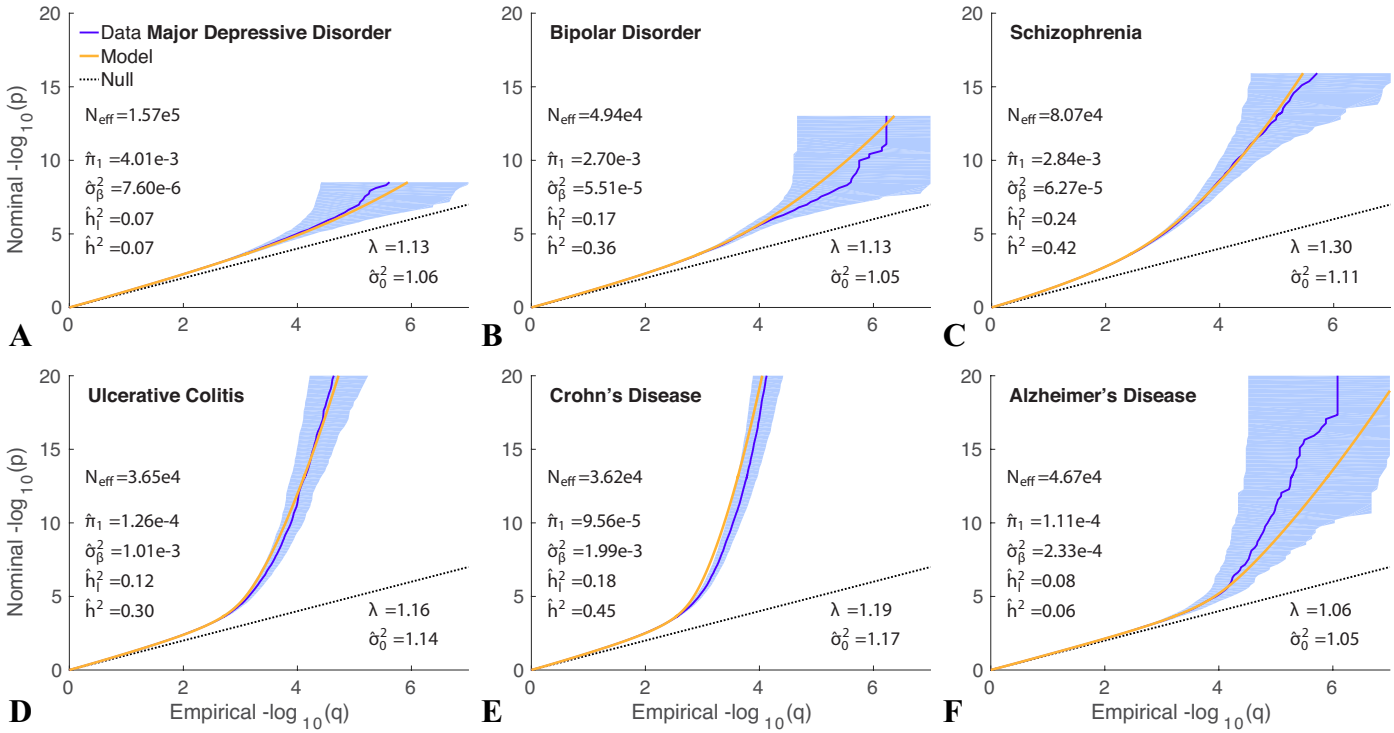

Figure 1: QQ plots of (pruned) z-scores for qualitative phenotypes (dark blue, 95% confidence interval in light blue) with model prediction (yellow): (A) major depressive disorder, (B) bipolar disorder, (C) schizophrenia, (D) ulcerative colitis, (E) Crohn's disease, and (F) late onset Alzheimer's disease (excluding APOE; see also Supplementary Material Fig. 7). The dashed line is the expected QQ plot under null (no SNPs associated with the phenotype). λ 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, $\hat{\sigma}^2_\beta$; and SNP association χ^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%, BIP 0.5%, SCZ 1%, UC 0.1%, CD 0.1%, AD 14% (for people aged 71 and older in the USA (Plassman et al., 2007)). 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).

the fixed set (and, as before, employing data reduction by averaging so that H and L denote values for the 10×10 grid), use Eq. 30 to define

$$C(z|N,H,L) \equiv E(\delta^2|z,N,H,L)P(z|N,H,L) \qquad (31)$$

(emphasizing dependence on N, H, and L). Then the proportion of chip heritability captured additively by genome-wide significant SNPs is estimated by

$$S(N) = \frac{\sum_{H,L}\sum_{z:|z|>z_t} C(z,N,H,L)}{\sum_{H,L}\sum_{all\ z} C(z,N,H,L)} \qquad (32)$$

where $\sum_{H,L}$ denotes sum over the H-L grid elements. The ratio in Eq. 32 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 $S(N)$ 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 cryptic relatedness – 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. 26, 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 10×10 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, λ, for the z-scores is defined as the median z^2 divided by the median for the null distribution, 0.455 (Devlin and Roeder, 1999). 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

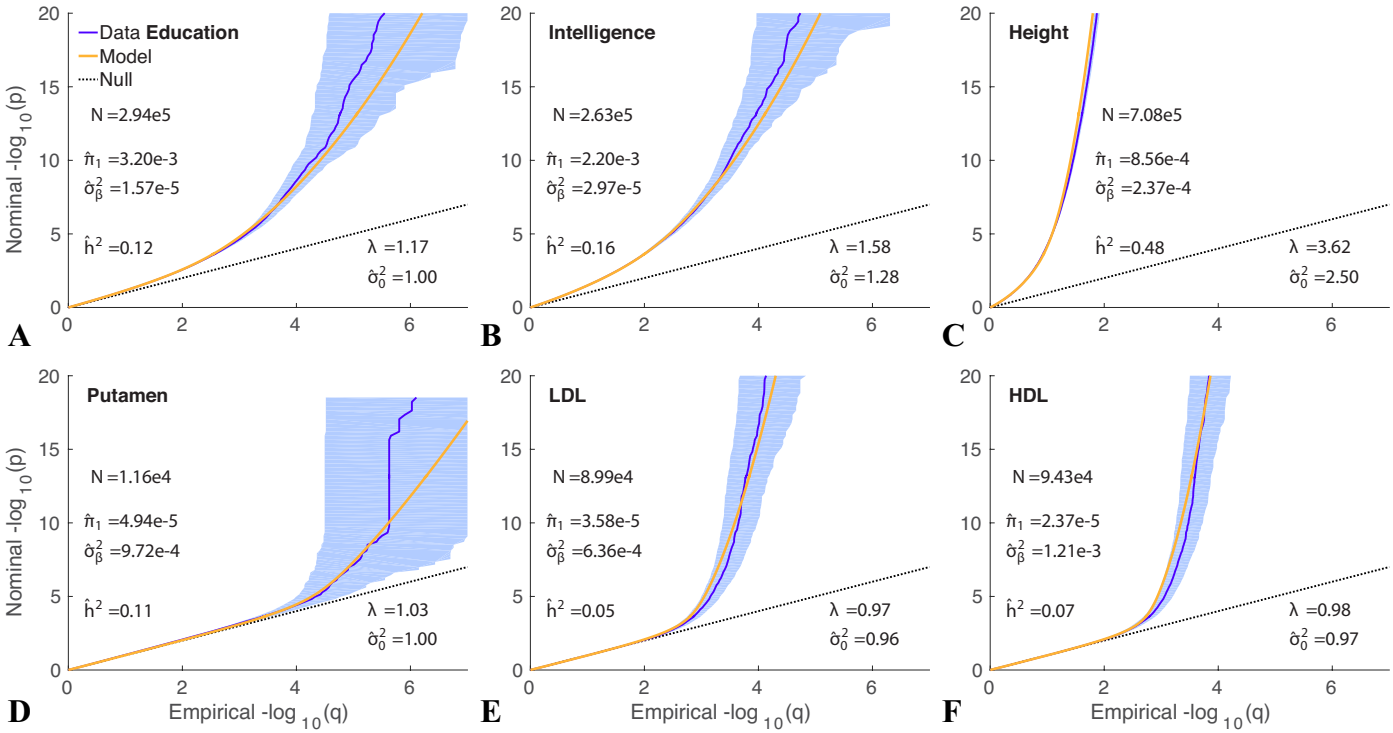

Figure 2: QQ plots of (pruned) z-scores for quantitative phenotypes: (A) educational attainment, (B) intelligence, (C) height, (D) putamen volume, (E) low-density lipoprotein (LDL), and (F) high-density lipoprotein (HDL). N is the sample size. See Fig. 1 for further description.

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 λ reported here are for pruned SNP 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, $n_p = qn_{tot}$, 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 (Clopper and Pearson, 1934), 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, π_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. 26, i.e., the approximately 11 million SNPs selected from the 1000 Genomes Phase 3 reference panel, not the number of tag SNPs in the particular GWAS. We emphasize that the more meaningful quantity estimated is the polygenicity; a reference panel with a substantially larger number of samples would allow for inclusion of more SNPs (non-zero maf), and thus the actual number of causal SNPs estimated would change slightly. Still, n_{causal} provides a valid metric for comparing phenotypes.

Narrow-sense Chip Heritability

Since we are treating the β coefficients as fixed effects in the simple linear regression GWAS formalism, with the phenotype vector standardized with mean zero and unit variance, the proportion of phenotypic variance explained by a particular causal SNP, 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(β^2)=σ_β^2 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

$$h^2 = \pi_1 n_{snp} \overline{H} \sigma_\beta^2. \tag{33}$$

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. 33 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 (Falconer, 1965), h_l^2, is obtained by adjusting for ascertainment (multiplying by $K(1-K)/(P(1-P))$,

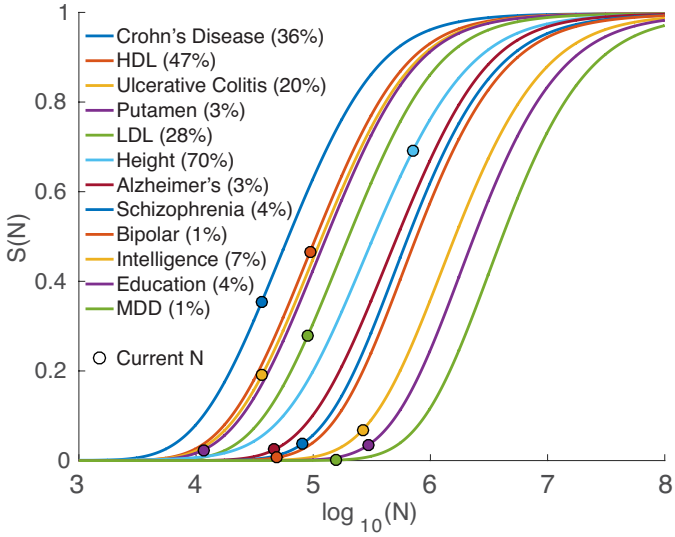

Figure 3: Proportion of narrow-sense chip heritability, $S(N)$ (Eq. 32), captured by genome-wide significant SNPs as a function of sample size, N, for phenotypes shown in Figures 1 and Figure 2. Values for current sample sizes are shown in parentheses. Left-to-right curve order is determined by decreasing σ_β^2. (The curve for AD excludes the APOE locus.)

the ratio of phenotypic variances in the population and in the study) and rescaling based on prevalence (Dempster and Lerner, 1950; Lee et al., 2011):

$$h_l^2 = h^2 \frac{K(1-K)}{P(1-P)} \times \frac{K(1-K)}{a^2}, \tag{34}$$

where a is the height of the standard normal pdf at the truncation point z_K defined such that the area under the curve in the region to the right of z_K is K.

RESULTS

Phenotypes

Figures 1 and 2 show QQ plots for the pruned z-scores for six qualitative and six quantitative phenotypes, along with model estimates. In all cases, the model fit (yellow) closely tracks the data (dark blue). For the twelve phenotypes polygenicities range over two orders of magnitude, from $\pi_1 \simeq 2 \times 10^{-5}$ to $\pi_1 \simeq 4 \times 10^{-3}$. The SNP discoverability (mean strength of SNP association with the phenotype, mean β^2 for causals) also ranges over two orders of magnitude from $\sigma_\beta^2 \simeq 7.6 \times 10^{-6}$ to $\sigma_\beta^2 \simeq 2 \times 10^{-3}$ (in units where the variance of the phenotype is normalized to 1).

We find that schizophrenia and bipolar disorder are similarly highly polygenic, with polygenicities $\simeq 2.84 \times 10^{-3}$ and $\simeq 2.70 \times 10^{-3}$, respectively. The polygenicity of major depressive disorder, however, is 40% higher, $\pi_1 \simeq 4 \times 10^{-3}$ – the highest value among the twelve phenotypes. In contrast, the polygenicities of late onset Alzheimer's disease and Crohn's disease are almost thirty times smaller than that of schizophrenia; ulcerative colitis is slightly more polygenic than Crohn's disease, $\pi_1 \simeq 1.26 \times 10^{-4}$ versus $\pi_1 = 9.56 \times 10^{-5}$.

In Supplementary Material Figure 7 we show results for Alzheimer's disease exclusively for chromosome 19 (which contains APOE), and for all autosomal chomosomes 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). These GWAS give consistent estimates of polygenicity: $\pi_1 \sim 1.1 \times 10^{-4}$ to 1.2×10^{-4}.

Of the quantitative traits, educational attainment has the highest polygenicity, $\pi_1 = 3.2 \times 10^{-3}$, slightly larger than for schizophrenia. Intelligence is less polygenic, $\pi_1 = 2.2 \times 10^{-3}$. Approximately two orders of magnitude lower in polygenicity are the endophenotypes putamen volume and low- and high-density lipoproteins; HDL had the lowest polygenicity of the twelve phenotypes, $\pi_1 = 2.37 \times 10^{-5}$. Height exhibited a medium polygenicity less than half that of intelligence, $\pi_1 = 8.56 \times 10^{-4}$.

The estimated number of causal SNPs for schizophrenia is $\hat{n}_{causal} \sim 3.1 \times 10^4$, but major depressive disorder involves significantly more causal SNPs: $\hat{n}_{causal} \simeq 4.4 \times 10^4$. Alzheimer's disease, however, involves only a relatively small number of causal SNPs, $\hat{n}_{causal} \simeq 1,200$ (with ~ 10 on chromosome 19 – see Supporting Material Figure 7), while for height $\hat{n}_{causal} \simeq 9,400$. There is no underlying assumption in the model that the causal SNPs are in linkage equilibrium, and indeed for high-polygenicity phenotypes they can not all be in linkage equilibrium.

The effective SNP discoverability for schizophrenia is $\hat{\sigma}_\beta^2 = 6.27 \times 10^{-5}$, slightly larger than 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 highest value of the twelve phenotypes: $\hat{\sigma}_\beta^2 = 1.99 \times 10^{-3}$.

Additionally, for Alzheimer's disease we show in Supplementary Material Figure 7 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 effective SNP discoverability for education is $\hat{\sigma}_\beta^2 = 1.57 \times 10^{-5}$; for intelligence it is almost two times larger,

h^2	$\hat{h}^2$	π_1	$\hat{\pi}_1$	σ^2_β	$\hat{\sigma}^2_\beta$	$\hat{\sigma}^2_0$	n_{causal}	$\hat{n}_{causal}$
0.1	0.12 (0.01)	1E-5	1.37E-5 (2E-6)	4.3E-3 (7E-4)	3.67E-3 (5E-4)	1.01 (0.002)	110	151 (20)
0.1	0.10 (0.01)	1E-4	1.03E-4 (2E-5)	4.2E-4 (2E-5)	4.13E-4 (5E-5)	1.01 (0.003)	1101	1130 (206)
0.1	0.09 (0.01)	1E-3	0.94E-3 (1E-4)	4.2E-5 (5E-7)	4.22E-5 (4E-6)	1.02 (0.003)	11015	10340 (1484)
0.1	0.09 (0.01)	1E-2	0.76E-2 (2E-3)	4.2E-6 (4E-8)	5.75E-6 (2E-6)	1.02 (0.002)	110158	83411 (25448)
0.4	0.52 (0.05)	1E-5	2.35E-5 (2E-6)	1.7E-2 (3E-3)	0.93E-2 (1E-3)	1.02 (0.002)	110	259 (20)
0.4	0.45 (0.02)	1E-4	1.19E-4 (8E-6)	1.7E-3 (7E-5)	1.59E-3 (9E-5)	1.04 (0.002)	1101	1310 (92)
0.4	0.39 (0.01)	1E-3	0.96E-3 (5E-5)	1.7E-4 (2E-6)	1.69E-4 (8E-6)	1.05 (0.003)	11015	10607 (578)
0.4	0.37 (0.01)	1E-2	0.86E-2 (1E-3)	1.7E-5 (2E-7)	1.83E-5 (2E-6)	1.06 (0.003)	110158	95135 (10851)
0.7	0.91 (0.09)	1E-5	2.95E-5 (2E-6)	3.0E-2 (5E-3)	1.30E-2 (2E-3)	1.02 (0.003)	110	324 (24)
0.7	0.82 (0.02)	1E-4	1.36E-4 (7E-6)	2.9E-3 (1E-4)	2.54E-3 (1E-4)	1.05 (0.002)	1101	1493 (79)
0.7	0.70 (0.01)	1E-3	0.99E-3 (4E-5)	2.9E-4 (4E-6)	2.97E-4 (1E-5)	1.08 (0.003)	11015	10866 (406)
0.7	0.66 (0.01)	1E-2	0.86E-2 (7E-4)	2.9E-5 (3E-7)	3.21E-5 (2E-6)	1.09 (0.003)	110158	95067 (8191)

Table 1: 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 π_1 of causal SNPs, are from 10 independent instantiations with random selection of the n_{causal} causal SNPs that are assigned a β-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 ε for each instantiation is drawn from a normal distribution such that $\mathrm{var}(Y) = \mathrm{var}(Y_g)/h^2$ for predefined h^2. For each of the instantiations, i, this implicitly defines the true value $\sigma^2_{\beta i}$, and σ^2_β is their mean. An example QQ plot for each line entry is shown in Supplementary Material, Figure 4.

and for height it is an order of magnitude larger still. However, high-density lipoprotein has a much higher SNP discoverability, over five times that of height.

Note that for logistic linear regression coefficient β, the odds ratio for disease is OR $= e^\beta$; for a rare disease, this is approximately equal to the genotypic relative risk: GRR $\simeq$ OR. Since $\mathbb{E}\left[\beta^2\right] = \sigma^2_\beta$, the mean relative risk $\mathbb{E}\left[\mathrm{GRR}\right] \simeq 1 + \sigma^2_\beta/2$. Thus, for schizophrenia for example, the mean relative risk is $\simeq 1.00003$.

The narrow sense SNP heritability from the ascertained case-control schizophrenia GWAS is estimated as h^2=0.42. Taking adult population prevalence of schizophrenia to be K=0.01 (Purcell et al., 2009; Whiteford et al., 2013) (but see also Kinney et al. (2009), 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. 34 is $\hat{h}^2_l$=0.24. For bipolar disorder, with K=0.005 (Merikangas et al., 2011), 20,352 cases and 31,358 controls, $\hat{h}^2_l$=0.17. Major depressive disorder appears to be much less heritable than schizophrenia: $\hat{h}^2_l$=0.07. Crohn's disease is slightly more heritable than ulcerative colitis: $\hat{h}^2_l$=0.18 versus $\hat{h}^2_l$=0.12. The SNP heritability of intelligence is greater than that of educational attainment, 16% versus 12%. For the quantitative endophenotype putamen volume, the heritability is estimated to be 11%. Of the twelve phenotypes considered, height had the highest SNP heritability, 48%.

Excluding APOE, our estimate of the liability-scale SNP heritability for Alzheimer's disease is 8%, with a similar contributions from APOE.

Figure 3 shows the sample size required to reach a given proportion of chip heritability 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 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. The proportion of chip heritability captured for height at the current sample size is considerably higher, 70%, and for HDL it is 47%. Major depressive disorder GWAS currently is greatly underpowered, as shown in Figure 3(A). For other phenotypes, the proportions of total SNP heritability captured at the available sample sizes are given in Figure 3.

Simulations

Table 1 shows the simulation results, comparing true and estimated values for the model parameters, heritability, and the number of causal SNPs, for twelve scenarios where π_1 and σ^2_β both range over three orders of magnitude, encompassing the range of values for the phenotypes; in Supporting Material, Figure 4 shows QQ plots for a randomly chosen (out of 10) β-vector and phenotype instantiation for each of the twelve (π_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 $\sim$11 million) are in correspondingly good agreement with the true values, ranging in increasing powers of 10 from 110 through 110,158. The estimated discoverabilities ($\hat{\sigma}^2_\beta$) are also in good agreement with the true values. In most

cases, $\hat{\sigma}_0^2$ 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.

DISCUSSION

Here we present the first unified method based on GWAS summary statistics, incorporating detailed LD structure from a reference panel, for directly estimating phenotypic polygenicity, π_1, SNP discoverability or strength of association (specifically, the variance of the underlying causal effects), σ_β^2, and narrow-sense SNP heritability. In addition the model can be used to estimate residual inflation of the association statistics due to variance distortion induced by cryptic relatedness, σ_0^2.

We apply the model to twelve diverse phenotypes, six qualitative and seven quantitative. In each case we estimate the polygenicity, discoverability, and residual inflation due to variance distortion; we also estimate the number of causal SNPs, n_{causal}, and the SNP heritability, h^2 (for qualitative phenotypes, we re-express this as the proportion of population variance in disease liability, h_l^2, under a liability threshold model, adjusted for ascertainment). 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 causals is >20,000 (Loh et al., 2015). The SNP associations, however, are characterized by a narrow distribution, $\sigma_\beta^2 = 6.27 \times 10^{-5}$, indicating that most associations are of weak effect, i.e., have low discoverability. Bipolar disorder has similar parameters, though slightly smaller discoverability. The smaller sample size for bipolar disorder, combined with its slightly smaller σ_β^2, 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 (Rietveld et al., 2013; Okbay et al., 2016; Cesarini and Visscher, 2017), 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 (Rietveld et al., 2013). The variance of the distribution for causal effect sizes is a quarter that of schizophrenia, indicating lower discoverability. Intelligence, a related phenotype (Sniekers et al., 2017; Plomin and von Stumm, 2018), has almost twice the dscoverability of 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 are characterized by relatively high discoverability, respectively 17-times and 31-times larger than for schizophrenia.

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.

Twin studies estimate the heritability of Alzheimer's disease to be in the range 60-80% (Gatz et al., 2006). A recent raw genotype-based analysis (GCTA), including genes that contain rare variants that affect risk for AD, reported SNP heritability of 53% (Ridge et al., 2016; Yang et al., 2011a); an earlier related study that did not include rare variants and had only a quarter of the common variants estimated SNP heritability of 33% (Ridge et al., 2013). It should be noted that 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 5×10^{-3}; for AD the polygenicity is $\simeq 10^{-4}$. Thus, though the GCTA approach yields a heritability estimate closer to the twin-based (broad sense) value, it is not clear how to interpret GCTA estimates of heritability when the underlying polygenicity in fact is $\ll 1$. For the 2013 data analyzed here (Lambert et al., 2013), a summary-statistics-based method applied to a subset of 54,162 of the 74,046 samples gave SNP heritability of almost 7% (Zheng et al., 2017; Bulik-Sullivan et al., 2015). Our estimate for the full 2013 dataset is 16%, half from APOE.

Onset and clinical progression of sporadic Alzheimer's disease is strongly age-related (Holland et al., 2012; Desikan et al., 2017), with prevalence in differential age groups increasing at least up through the early 90s (Plassman et al., 2007). 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. choice of prevalence parameter in calculating liability-scale heritability, as well as 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. The consistent deviations between the model fits and the data shown in Figure 1 (F) and Supporting Material Figure 7 (A) and (D) suggest that the model does not fully capture the genetic architecture of AD (ignoring APOE/chromosome 19). Future work will

examine annotation-specific SNP categories, and the possibility of higher locality among causal SNPs – where one is, another might more likely be in the neighborhood than indicated by the random placement underlying the current version of the model. Regarding chromosome 19, the very low polygenicity coupled with the very high discoverability is a region where the model is less reliable in the precision of the parameter estimates; apart from proposed improvements just described, accuracy in this region of parameter space will be enhanced by more precise estimates of LD and heterozygosity structure of SNPs (enhanced signal to noise in the low LD ($r^2 < 0.05$) and/or low heritability ($H < 0.05$) range), so that r^2_{min} used in the model pdf (Eq. 26) can reliably be taken lower than 0.05. This can be achieved by using reference panels with larger sample sizes.

Despite the much stronger effects in putamen volume, the very high polygenicity for schizophrenia leads to its being almost four times as heritable. Our point estimate for liability-scale heritability of schizophrenia is $h^2_l = 0.28$ (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^2_l estimate is in reasonable agreement with a recent result, $h^2_l = 0.27$ (Loh et al., 2015; Golan et al., 2014), 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 $h^2_l = 0.23$ from PGC1 raw genotype data for 915,354 markers for 9,087 schizophrenia cases and 12,171 controls (Lee et al., 2012; Yang et al., 2011a). Our estimate of 1% of overall variation on the liability scale for schizophrenia explainable by genome-wide significant loci is less than the related estimate based on risk profile scores (RPS) of 3.4% obtained at the smaller sample size (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Two possible explanations for this are: (A) our result is based on the distribution of many ($\sim 3.1 \times 10^4$) causal SNPs, not specific measures of the most extreme 108 SNPs in the tails of the distribution; (B) inflation may play a role. Nevertheless, 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). The identified SNPs then need to be mapped to genes and their modality (e.g., regulatory or functional effects) determined, so that targeted therapeutics can be developed (Schubert et al., 2015). Greater power for discovery is achievable by using prior information involving SNP functional categories (Schork et al., 2013; Andreassen et al., 2013; Sveinbjornsson et al., 2016). However, it is not yet clear how significant a role genomics can play in psychiatric precision medicine (Breen et al., 2016). It is noteworthy that estimates of broad-sense heritability of schizophrenia from twin and family studies are in the range 0.6-0.8 (Sullivan et al., 2003; Lichtenstein et al., 2009), considerably higher than the narrow-sense chip heritability estimates from GWAS. Additionally, schizophrenia is considered a spectrum disorder with multiple phenotypic dimensions and diverse clinical presentation (MacDonald and Schulz, 2009; Peralta and Cuesta, 2001); GWAS might therefore benefit from considering continuous phenotypes rather than dichotomous variables in such situations (Edwards et al., 2016). More specifically in the context of the present model, if a nominally categorical phenotype can be decomposed into more than one subcategory, there is potential for enhanced power for discovery. The heritability estimated in a binary case-control design would be an average over heritabilities for the case subcategories. If those heritabilities are similar, then, since the union of the subcategory polygenicities gives the total polygenicity over all cases, the σ^2_β for any subcategory will be larger by a factor equal to the ratio of overall polygenicity to the subcategory polygenicity, and the corresponding power curve (as in Figure 3) will shift to the left.

For educational attainment, we estimate SNP heritability $h^2 = 0.12$, in good agreement with the estimate of 11.5% given in Okbay et al. (2016). 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 (Branigan et al., 2013; Rietveld et al., 2013).

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 (Hibar et al., 2015; So et al., 2011).

For height we find that its polygenicity is $\pi_1 = 8.56 \times 10^{-4}$, two-fifths that of intelligence, while its discoverability is over an order of magnitude greater than that of intelligence, leading to a SNP heritability of 48%. This is close to the GCTA estimate of 50% obtained from a subset of the raw genotype data (Wood et al., 2014). Power analysis shows that for height 70% of the narrow-sense heritability arising from common SNPs is explained by genome-wide significant SNPs at the current sample size, i.e., almost 33% of total phenotypic variance. Note that a polygenicity of $\sim 8 \times 10^{-4}$ is very far from "omnigenic" (Boyle et al., 2017).

Our liability-scale heritability estimate for schizophrenia is substantially lower than the estimate from LD Score regression, $h^2_l = 0.555$ (Bulik-Sullivan et al., 2015). When the per-allele effect size, β, is independent of the MAF, LD Score regression will lead to a biased rotation of the regression line, increasing the slope, which is proportional to the estimated heritability, and decreasing the intercept, which gives the residual inflation. Thus, the discrepancy with LD Score regression in principle might arise due to the assumption in LD Score regression that effect sizes for variants are drawn independently from normal distributions with variance inversely proportional to heterozygos-

ity (so that the variance explained per SNP is uncorrelated with TLD, since TLD is positively correlated with MAF – Supporting Material Figure 5(B)). However, we show that the expected z^2 increases linearly with H (as well as TLD, Supporting Material Figure 6), indicating that $E(\beta^2) = \sigma_\beta^2$ is, to a first approximation, independent of MAF. Thus, the assumption of larger effects (β) for less common variants appears unwarranted, and would lead to both a lower estimate of residual inflation from population structure/cryptic relatedness and a higher estimate of heritability.

CONCLUSION

The SNP-level causal effects model we have presented is based on GWAS summary statistics and detailed LD structure, and assumes a Gaussian distribution of effect sizes at a fraction of SNPs randomly distributed across the autosomal genome. 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. In addition, the model provides a roadmap for discovery in future GWAS. Future extensions and refinements include modeling specific polygenicities and effect size variances for different SNP functional annotation categories (Schork et al., 2013; Andreassen et al., 2013; Sveinbjornsson et al., 2016), possible modified pdf for non-Gauassian distribution of effects at the tails of the z-score distributions, examining individual chromosomes and possible allele frequency dependencies in different phenotypes, and extension to pleiotropic analyses. Higher accuracy in characterizing causal alleles in turn will enable greater power for SNP discovery.

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

Andreassen, O. A., Djurovic, S., Thompson, W. K., Schork, A. J., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., van de Bunt, M., Morris, A. P., et al., 2013. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. The American Journal of Human Genetics 92 (2), 197–209.

Boyle, E. A., Li, Y. I., Pritchard, J. K., 2017. An expanded view of complex traits: From polygenic to omnigenic. Cell 169 (7), 1177–1186.

Branigan, A. R., McCallum, K. J., Freese, J., 2013. Variation in the heritability of educational attainment: An international meta-analysis. Social Forces, 109–140.

Breen, G., Li, Q., Roth, B. L., O'Donnell, P., Didriksen, M., Dolmetsch, R., O'Reilly, P. F., Gaspar, H. A., Manji, H., Huebel, C., et al., 2016. Translating genome-wide association findings into new therapeutics for psychiatry. Nature Neuroscience 19 (11), 1392–1396.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., Neale, B. M., of the Psychiatric Genomics Consortium, S. W. G., et al., 2015. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics 47 (3), 291–295.

Cesarini, D., Visscher, P. M., 2017. Genetics and educational attainment. npj Science of Learning 2 (1), 4.

Clopper, C. J., Pearson, E. S., 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26 (4), 404–413.

Consortium, . G. P., et al., 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491 (7422), 56–65.

Consortium, . G. P., et al., 2015. A global reference for human genetic variation. Nature 526 (7571), 68–74.

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., et al., 2017. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nature genetics 49 (2), 256.

Dempster, E. R., Lerner, I. M., 1950. Heritability of threshold characters. Genetics 35 (2), 212.

Desikan, R. S., Fan, C. C., Wang, Y., Schork, A. J., Cabral, H. J., Cupples, L. A., Thompson, W. K., Besser, L., Kukull, W. A., Holland, D., et al., 2017. Genetic assessment of age-associated alzheimer disease risk: Development and validation of a polygenic hazard score. PLoS medicine 14 (3), e1002258.

Devlin, B., Roeder, K., Dec 1999. Genomic control for association studies. Biometrics 55 (4), 997–1004.

Edwards, A. C., Bigdeli, T. B., Docherty, A. R., Bacanu, S., Lee, D., De Candia, T. R., Moscati, A., Thiselton, D. L., Maher, B. S., Wormley, B. K., et al., 2016. Meta-analysis of positive and negative symptoms reveals schizophrenia modifier genes. Schizophrenia bulletin 42 (2), 279–287.

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

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., Fiske, A., Pedersen, N. L., 2006. Role of genes and environments for explaining alzheimer disease. Archives of general psychiatry 63 (2), 168–174.

Golan, D., Lander, E. S., Rosset, S., 2014. Measuring missing heritability: inferring the contribution of common variants. Proceedings of the National Academy of Sciences 111 (49), E5272–E5281.

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivières, S., Jahanshad, N., Toro, R., Wittfeld, K., Abramovic, L., Andersson, M., et al., 2015. Common genetic variants influence human subcortical brain structures. Nature.

Holland, D., Desikan, R. S., Dale, A. M., McEvoy, L. K., Initiative, A. D. N., et al., 2012. Rates of decline in alzheimer disease decrease with age. PloS one 7 (8), e42325.

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

Jansen, I., Savage, J., Watanabe, K., Bryois, J., Williams, D., Steinberg, S., Sealock, J., Karlsson, I., Hagg, S., Athanasiu, L., et al., 2018. Genetic meta-analysis identifies 10 novel loci and functional pathways for alzheimer's disease risk. bioRxiv, 258533.

Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E., et al., 2010. Variance component model to account for sample structure in genome-wide association studies. Nature genetics 42 (4), 348–354.

Kinney, D. K., Teixeira, P., Hsu, D., Napoleon, S. C., Crowley, D. J., Miller, A., Hyman, W., Huang, E., 2009. 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, sbp023.

Kumar, S. K., Feldman, M. W., Rehkopf, D. H., Tuljapurkar, S., 2016. Limitations of gcta as a solution to the missing heritability problem. Proceedings of the National Academy of Sciences 113 (1), E61–E70.

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., et al., 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. Nature genetics 45 (12), 1452–1458.

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., Keller, M. C., Visscher, P. M., Wray, N. R., Consortium, S. P. G.-W. A. S., et al., 2012. Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. Nature genetics 44 (3), 247–250.

Lee, S. H., Wray, N. R., Goddard, M. E., Visscher, P. M., 2011. Estimating missing heritability for disease from genome-wide association studies. The American Journal of Human Genetics 88 (3), 294–305.

Li, N., Stephens, M., 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165 (4), 2213–2233.

Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., Hultman, C. M., 2009. Common genetic determinants of schizophrenia and bipolar disorder in swedish families: a population-based study. The Lancet 373 (9659), 234–239.

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., et al., 2015. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nature genetics.

MacDonald, A. W., Schulz, S. C., 2009. What we know: findings that every theory of schizophrenia should explain. Schizophrenia Bulletin 35 (3), 493–508.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., Hirschhorn, J. N., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews genetics 9 (5), 356–369.

Merikangas, K. R., Jin, R., He, J.-P., Kessler, R. C., Lee, S., Sampson, N. A., Viana, M. C., Andrade, L. H., Hu, C., Karam, E. G., et al., 2011. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Archives of general psychiatry 68 (3), 241–251.

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S. F. W., et al., 2016. Genome-wide association study identifies 74 loci associated with educational attainment. Nature 533 (7604), 539–542.

Palla, L., Dudbridge, F., 2015. 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.

Pasaniuc, B., Price, A. L., 2016. Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics.

Pe'er, I., Yelensky, R., Altshuler, D., Daly, M. J., 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic epidemiology 32 (4), 381–385.

Peralta, V., Cuesta, M. J., 2001. How many and which are the psychopathological dimensions in schizophrenia? issues influencing their ascertainment. Schizophrenia research 49 (3), 269–285.

Plassman, B. L., Langa, K. M., Fisher, G. G., Heeringa, S. G., Weir, D. R., Ofstedal, M. B., Burke, J. R., Hurd, M. D., Potter, G. G., Rodgers, W. L., et al., 2007. Prevalence of dementia in the united states: the aging, demographics, and memory study. Neuroepidemiology 29 (1-2), 125–132.

Plomin, R., von Stumm, S., 2018. The new genetics of intelligence. Nature Reviews Genetics.

Price, A. L., Zaitlen, N. A., Reich, D., Patterson, N., 2010. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11 (7), 459–463.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., Sklar, P., Purcell, S. M., Stone, J. L., Sullivan, P. F., et al., 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460 (7256), 748–752.

Ridge, P. G., Hoyt, K. B., Boehme, K., Mukherjee, S., Crane, P. K., Haines, J. L., Mayeux, R., Farrer, L. A., Pericak-Vance, M. A., Schellenberg, G. D., et al., 2016. Assessment of the genetic variance of late-onset alzheimer's disease. Neurobiology of aging 41, 200–e13.

Ridge, P. G., Mukherjee, S., Crane, P. K., Kauwe, J. S., et al., 2013. Alzheimer2019s disease: analyzing the missing heritability. PloS one 8 (11), e79771.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al., 2013. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. science 340 (6139), 1467–1471.

Savage, J., Jansen, P., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C., Nagel, M., Awasthi, S., Barr, P., Coleman, J., Grasby, K., Hammerschlag, A., Kaminski, J., Karlsson, R., et al., 2018. Genome-wiide association meta-analysis (n=269,867) identifies new genetic and functional links to intelligence. Nature genetics forthcoming.
URL `https://ctg.cncr.nl/software/summary_statistics`

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

Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., Kelsoe, J. R., O'Donovan, M. C., Furberg, H., Schork, N. J., et al., 2013. All snps are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps. PLoS genetics 9 (4), e1003449.

Schubert, C. R., O'Donnell, P., Quan, J., Wendland, J. R., Xi, H. S., Winslow, A. R., Domenici, E., Essioux, L., Kam-Thong, T., Airey, D. C., et al., 2015. Brainseq: neurogenomics to drive novel target discovery for neuropsychiatric disorders. Neuron 88 (6), 1078.

Sniekers, S., Stringer, S., Watanabe, K., Jansen, P. R., Coleman, J. R., Krapohl, E., Taskesen, E., Hammerschlag, A. R., Okbay, A., Zabaneh, D., et al., 2017. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. Nature genetics 49 (7), 1107.

So, H.-C., Li, M., Sham, P. C., 2011. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. Genetic epidemiology 35 (6), 447–456.

Speed, D., Hemani, G., Johnson, M. R., Balding, D. J., 2012. Improved heritability estimation from genome-wide snps. The American Journal of Human Genetics 91 (6), 1011–1021.

Spencer, C. C., Su, Z., Donnelly, P., Marchini, J., 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5 (5), e1000477.

Stahl, E., Breen, G., Forstner, A., McQuillin, A., Ripke, S., Cichon, S., Scott, L., Ophoff, R., Andreassen, O. A., Kelsoe, J., Sklar, P., 2018. Genomewide association study identifies 30 loci associated with bipolar disorder. bioRxiv.
URL `https://www.biorxiv.org/content/early/2018/01/24/173062`

Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., Kraft, P., Chen, R., Kallberg, H. J., Kurreeman, F. A., et al., 2012. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nature genetics 44 (5), 483–489.

Su, Z., Marchini, J., Donnelly, P., 2011. Hapgen2: simulation of multiple disease snps. Bioinformatics 27 (16), 2304–2305.

Sullivan, P. F., Kendler, K. S., Neale, M. C., 2003. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies.

Archives of general psychiatry 60 (12), 1187–1192.

Sveinbjornsson, G., Albrechtsen, A., Zink, F., Gudjonsson, S. A., Oddson, A., Másson, G., Holm, H., Kong, A., Thorsteinsdottir, U., Sulem, P., et al., 2016. Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nature genetics.

Thompson, W. K., Wang, Y., Schork, A., Zuber, V., Andreassen, O. A., Dale, A. M., Holland, D., Shujing, X., 2015. An empirical bayes method for estimating the distribution of effects in genome-wide association studies. PLoS Genetics [in press].

Visscher, P. M., Brown, M. A., McCarthy, M. I., Yang, J., 2012. Five years of gwas discovery. The American Journal of Human Genetics 90 (1), 7–24.

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., Charlson, F. J., Norman, R. E., Flaxman, A. D., Johns, N., et al., 2013. 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.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., et al., 2013. Discovery and refinement of loci associated with lipid levels. Nature genetics 45 (11), 1274.

Witte, J. S., Visscher, P. M., Wray, N. R., 2014. The contribution of genetic variants to disease depends on the ruler. Nature Reviews Genetics 15 (11), 765–776.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., et al., 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature genetics 46 (11), 1173–1186.

Wray, N. R., Sullivan, P. F., 2017. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. bioRxiv.
URL `https://www.biorxiv.org/content/early/2017/07/24/167577`

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., Robinson, M. R., Perry, J. R., Nolte, I. M., van Vliet-Ostaptchouk, J. V., et al., 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature genetics.

Yang, J., Lee, S. H., Goddard, M. E., Visscher, P. M., 2011a. Gcta: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics 88 (1), 76–82.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., De Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., et al., 2011b. Genome partitioning of genetic variation for complex traits using common snps. Nature genetics 43 (6), 519–525.

Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'connell, J. R., Mangino, M., et al., 2011c. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics 19 (7), 807–812.

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., Frayling, T. M., Hirschhorn, J., Yang, J., Visscher, P. M., et al., 2018. Meta-analysis of genome-wide association studies for height and body mass index in˜ 700,000 individuals of european ancestry. bioRxiv, 274654.

Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., Haycock, P. C., Hemani, G., Tansey, K., Laurin, C., Pourcain, B. S., et al., 2017. 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.

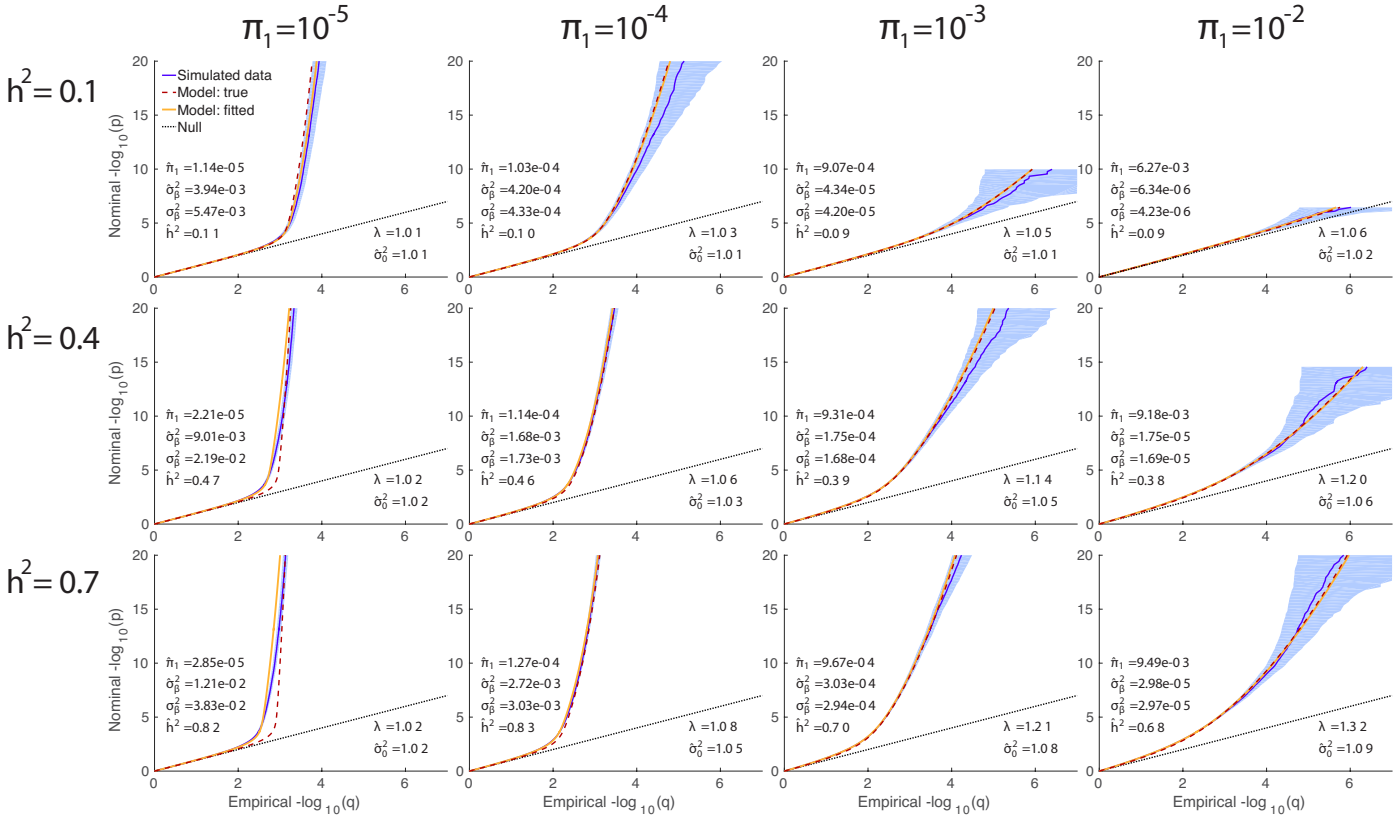

Figure 4: Quantile-quantile plots for simulations. True polygenicity is specified for each column, and true heritability is specified for each row. QQ-plots for simulated data in dark blue, with 95% confidence interval in light blue; model prediction in yellow. The dashed red curve is the QQ plot corresponding to the true parameters. λ is the overall nominal genomic control factor for the pruned simulated data. The three estimated model parameters are: polygenicity, $\hat{\pi}_1$; discoverability, $\hat{\sigma}^2_\beta$; and SNP association χ^2-statistic inflation factor, $\hat{\sigma}^2_0$. $\hat{h}^2$ is the estimated narrow-sense chip heritability. The dotted black line is the expected plot under null. 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).

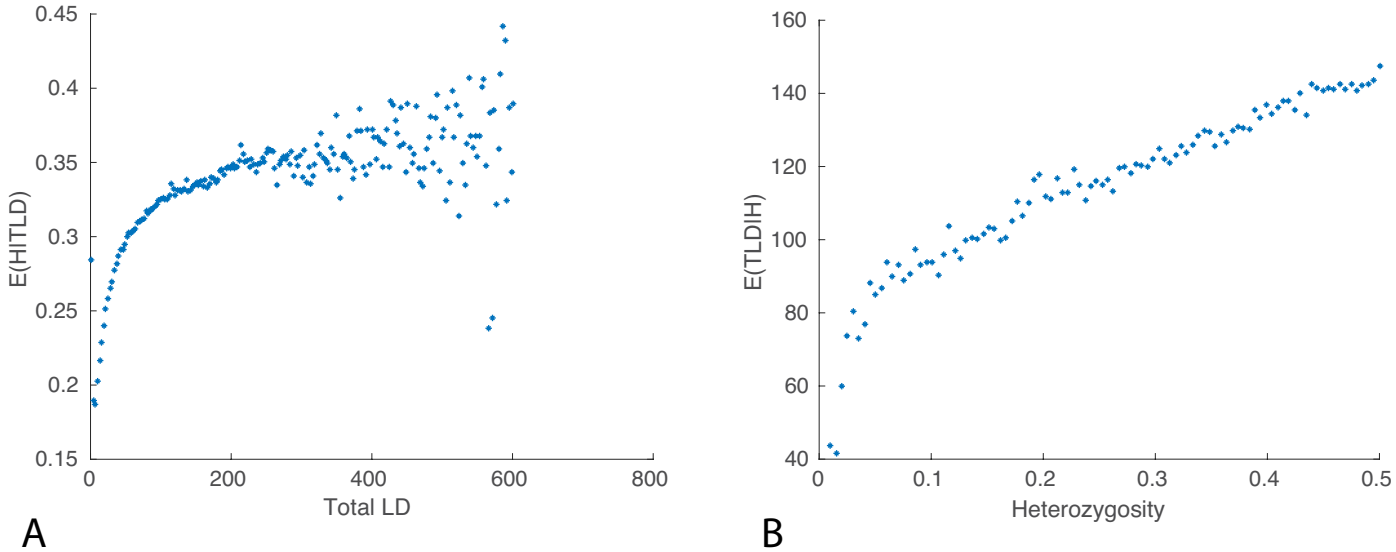

Figure 5: (A) Mean value of heterozygosity for given total LD (SNPs were binned based on TLD and the mean TLD for each bin plotted on the x-axis; the corresponding mean heterozygosity for SNPs in each bin was then plotted on the y-axis). (B) Mean value of total LD for given heterozygosity. Plots made for SNPs in the PGC2 schizophrenia GWAS; TLD and H calculated from 1000 Genomes phase 3 reference panel.

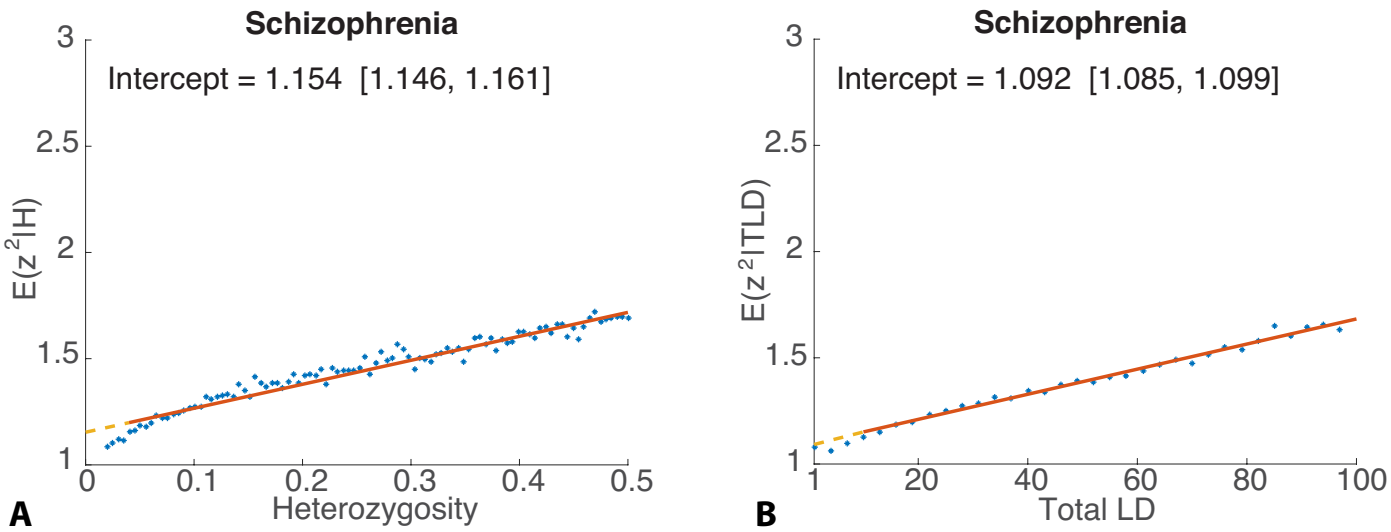

Figure 6: Regressing z^2 on (A) heterozygosity, and (B) total LD for schizophrenia.

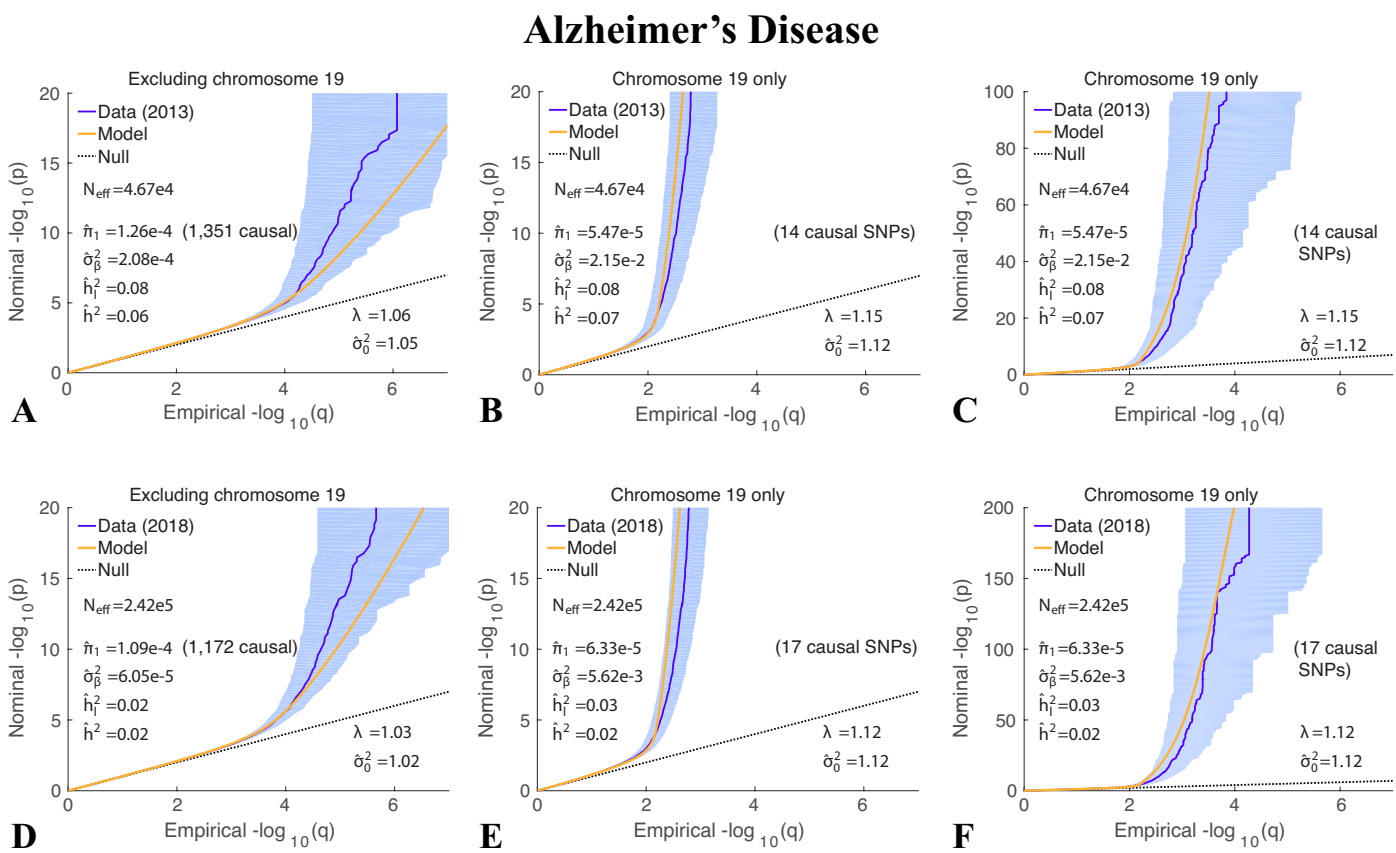

Figure 7: QQ plots for AD, comparing the 2013 data (Lambert et al., 2013) and model results (top row) with those of 2018 (Jansen et al., 2018) (bottom row). For both data sets, we exclude (in (A) and (D)) chromosome 19 from the analysis, and in (B) and (E) we analyze chromosome 19 exclusively. (C) is (B) with the y-axis extended 5×; (F) is (E) with the y-axis extended 10×. The heritabilities reported are for the relevant sections of DNA. See Figure 1 for a further description.