Estimating degree of polygenicity, causal effect size variance, and confounding bias in GWAS summary statistics

Dominic Holland[a,b,*], Chun-Chieh Fan[a,c,d], Oleksandr Frei[e], Alexey A. Shadrin[e], Olav B. Smeland[e,f,a], V. S. Sundar[a,d], Enhancing Neuro Imaging Genetics through Meta Analysis Consortium, Ole A. Andreassen[e,f], Anders M. Dale[a,b,d,g]

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

Abstract

Of signal interest in the genetics of traits are estimating the proportion, π_1, of causally associated single nucleotide polymorphisms (SNPs), and their effect size variance, σ^2_β, which are components of the mean heritabilities captured by the causal SNP. Here we present the first model, using detailed linkage disequilibrium structure, to estimate these quantities from genome-wide association studies (GWAS) summary statistics, assuming a Gaussian distribution of SNP effect sizes, β. We apply the model to three diverse phenotypes – schizophrenia, putamen volume, and educational attainment – and validate it with extensive simulations. We find that schizophrenia is highly polygenic, with $\simeq 5 \times 10^4$ causal SNPs distributed with small effect size variance, $\sigma^2_\beta = 3.5 \times 10^{-5}$ (in units where the phenotype variance is normalized to 1), requiring a GWAS study with more than 1/2-million samples in each arm for full discovery. In contrast, putamen volume involves only $\simeq 3 \times 10^2$ causal SNPs, but with $\sigma^2_\beta = 1.2 \times 10^{-3}$, indicating a much larger proportion of the causal SNPs that are strongly associated. Educational attainment has similar polygenicity to schizophrenia, but with effects that are substantially weaker, $\sigma^2_\beta = 5 \times 10^{-6}$, leading to much lower heritability. Thus the model is able to describe the broad genetic architecture of phenotypes where both polygenicity and effect size variance 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, Causal SNPs, Effect size, Linkage Disequlilbrium

INTRODUCTION

The genetic components of complex traits or 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 any residual inflation of the z-scores from uncorrected population structure or 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

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

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 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 (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 pruned 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

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:

$$\beta \sim \mathcal{N}(0, \sigma_\beta^2). \tag{1}$$

(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

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

where $\mathcal{N}(0, 0)$ is the Dirac delta function, so that considering all SNPs, the net variance is $\mathrm{var}(\beta) = \pi_1 \sigma_\beta^2$. Ignoring LD, the association z-scores for causal 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):

$$z = \delta + \epsilon \tag{3}$$

with

$$\delta = \sqrt{NH}\beta \tag{4}$$

where N is the sample size and H is the SNP's heterozygosity (frequency of the heterozygous genotype, $H = 2p(1-p)$ where p is the frequency of either of the SNP's alleles), so that

$$\begin{aligned} \mathrm{var}(z) &= \mathrm{var}(\delta) + \mathrm{var}(\epsilon) \\ &\equiv \sigma^2 + \sigma_0^2 \end{aligned} \tag{5}$$

where

$$\sigma^2 \equiv \sigma_\beta^2 NH. \tag{6}$$

Now consider the effects of LD on z-scores. Let β_{eff} be the true effective β-coefficient for a tag SNP arising due to LD with neighboring causal SNPs. It is given by the sum of neighboring causal SNP β-coefficients, each weighted by its correlation with the tag SNP:

$$\beta_{eff} = \sum_j r_j \beta_j. \tag{7}$$

Then, from Eq. 3, the z-score for the tag SNP's association with the phenotype is given by:

$$z = \sqrt{NH}\beta_{eff} + \epsilon. \tag{8}$$

Thus, for example, if the SNP itself were not causal but were in LD with k known causal SNPs, where its LD with each of these was the same, given by some value r^2 ($0 < r^2 \leq 1$), then σ^2 will be given by

$$\sigma^2 = kr^2 \sigma_\beta^2 NH. \tag{9}$$

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

$$f_a(z|N, H, L; \sigma_\beta, \sigma_0) = \phi(z; 0, kr^2\sigma_\beta^2 NH + \sigma_0^2) \tag{10}$$

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

$$f(z) = (1 - a)\phi(z; 0, \sigma_0^2) + af_a(z), \tag{11}$$

dropping the parameters for convenience.

For real genotypes, however, the LD structure is far more complicated, and of course the causal SNPs are generally numerous and unknown. 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 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 0.05 and 1; we use L again as a shorthand to represent all this. The value $r^2_{min} = 0.05$ was chosen as a lower-bound for LD above the noise threshold; we find that $w \simeq 10$ 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, H, L 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. 10 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 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

$$M(k_0, ..., k_w; n_0, ..., n_w; \pi_1) = \frac{n!}{k_0!...k_w!} p_0^{k_0} ... p_w^{k_w} \qquad (12)$$

and the corresponding gaussian is

$$\phi(z; 0, (k_1 r_1^2 + ... + k_w r_w^2)\sigma_\beta^2 NH + \sigma_0^2). \qquad (13)$$

For a SNP with heterozygosity H and LD structure L, 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:

$$\begin{aligned} &\mathrm{pdf}(z|N, H, L; \pi_1, \sigma_\beta, \sigma_0) = \\ &\sum_{K=0}^{K_{max}} \sum_{k_1,...,k_w} \frac{n!}{k_0!...k_w!} p_0^{k_0} ... p_w^{k_w} \times \\ &\phi(z; 0, (k_1 r_1^2 + ... + k_w r_w^2)\sigma_\beta^2 NH + \sigma_0^2), \end{aligned} \qquad (14)$$

where K_{max} in bounded above by n. Note again that L is shorthand for the linkage-disequilibrium structure of the SNP, giving the set $\{n_i\}$, 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 – see below – $K_{max} \ll n_i$. The number of ways of dividing K causal SNPs amongst w LD windows is given by the binomial coefficient $\binom{m}{a}$, where $m = K + w - 1$ and $a = w - 1$, so the number of terms in the second summation grows rapidly with K and w. However, because π_1 is small (often $\leq 10^{-3}$), we find that 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 $K_{max} < \min(10, n)$, even for large blocks with $n \simeq 10^3$. That is,

$$\sum_{K=0}^{K_{max}} \sum_{k_1,...,k_w} M(k_0, ..., k_w; n_0, ..., n_w; \pi_1) \simeq 1. \qquad (15)$$

Still, the number of terms is large; e.g., for $K = 8$ and $w = 5$ there are 495 terms. We approximate the sums in Eq. 14 with the simpler expression involving only sums over terms where the causal SNPs all reside in the same r^2-window, plus a null term. The probability that any K of the n SNPs in the block are causal while the remainder $n - K$ are null is given by the binomial distribution, $B(K, n; \pi_1)$:

$$B(K; n; \pi_1) = (1-\pi_1)^{n-K} \pi_1^K \frac{n!}{(n-K)!K!}. \qquad (16)$$

Multiplying this by n_i/n approximates the probability of their being in the i-th r^2-window. Multiplying these into the gaussian corresponding to K causals in window i, summing over both indices, and incorporating the null term, leads to the following approximation that is in good numerically agreement with Eq. 14:

$$\begin{aligned} &\mathrm{pdf}(z|N, H, L; \pi_1, \sigma_\beta, \sigma_0) = \\ &\sum_{K=1}^{K_{max}} \sum_{i=1}^{w} \frac{n_i}{n} B(K; n; \pi_1) \phi(z; 0, K r_i^2 \sigma_\beta^2 NH + \sigma_0^2) \\ &+ (1-\pi_1)^n \phi(z; 0, \sigma_0^2). \end{aligned} \qquad (17)$$

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 $r^2 < 0.05$ we set that r^2 to zero (zeroing out the noise). For each SNP we also built a histogram giving the numbers of SNPs in $w = 8$ equally-spaced r^2-windows covering the range $0.05 \leq r^2 \leq 1$. These steps were carried out independently for both 1000 Genomes phase 3 and for HapGen2.

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 restricted to SNPs for which TLD ≤ 600, MAF ≥ 0.005, and LD block size (defined by $r^2 \geq 0.05$) ≤ 2000.

We analyzed summary statistics for participants with European ancestry for: (1) schizophrenia from the Psychiatric Genomics Consortium (PGC) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), with 35,476 cases and 46,839 controls ($N_{eff} \equiv 4/(1/N_{cases} + 1/N_{controls}) = 76,326$) across 52 separate substudies, with imputation of SNPs using the 1000 Genomes Project reference panel (1000 Genomes Project Consortium, 2010) for a total of approximately 5,369,285 genotyped and imputed SNPs passing the above restrictions (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014); (2) putamen volume using data from the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) consortium (Hibar et al., 2015), with 12,596 samples and a total of 4,196,831 SNPs; and (3) educational attainment, measured as the number of years of schooling completed, with 328,917 samples and a total of 5,361,110 SNPs, available at `https://www.thessgac.org` (Okbay et al., 2016). Examples of SNP histograms for schizophrenia are in Supporting Material Fig. 5.

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 constructed as $Y = Y_g + \epsilon$, 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 heterozygosity, H, and detailed LD structure, i.e., its LD histogram, L. Given the data – the set of z-scores for all SNPs, as well as their heterozygosities and LD-structures – and the H- and L-dependent pdf for z-scores, the objective is to find the model parameters that best predict the distribution of all z-scores. H ranges between 0.05 and 0.5, and the amplitudes of L will vary over a wide range. A useful one-dimensional proxy for L is TLD, which ranges from 1 to 600. Since the model pdf explicitly predicts z-score distributions for particular values of H and L, instead of taking all the SNPs at once, we bin the SNPs with respect to a grid of these quantities; for any given (H,TLD) bin there will be a range of z-scores whose distribution the model it intended to predict. We find that a 5×5-grid of equally spaced bins is adequate for converged results. In lieu of or in addition to TLD binning, one can bin SNPs with respect to their total LD block size (total number of SNPs in LD, ranging from 1 to 2,000).

To find the model parameters that best fit the data, for a given (H,TLD) bin we binned the selected SNPs z-scores into equally-spaced bins of width dz=0.4 (between z_{min}=−38 and z_{max}=38, allowing for p-values near the numerical limit of 10^{-315}), and from Eq. 17 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 prameter values will be those that maximize the accrual of this probability over all (H,TLD) bins. We constructed a cost function by calculating, for a givem (H,TLD) bin, $-\ln(p_m)$ and averaging over prunings, and then accumulating this over all (H,TLD) bins. Model parameters minimizing the cost were obtained from Nelder-Mead multidimensional unconstrained nonlinear minimization of the cost function, using the Matlab function `fminsearch()`.

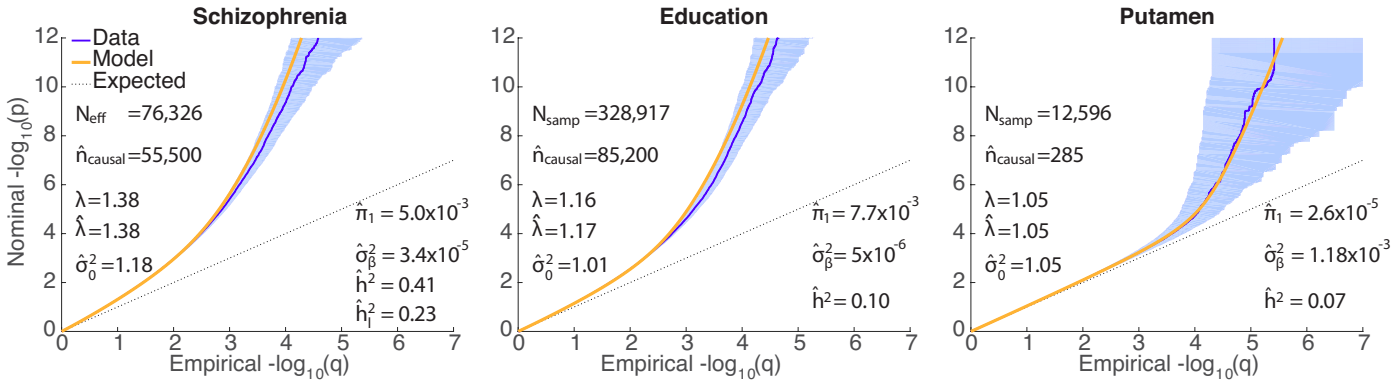

Figure 1: QQ plots of z-scores for (A) schizophrenia, (B) educational attainment, and (C) putamen volume, (dark blue, 95% confidence interval in light blue) with model prediction (yellow). The dashed line is the expected QQ plot under null (no SNPs associated with the phenotype). λ and $\hat{\lambda}$ are the overall nominal genomic control factors calculated from the data and model plots, respectively. 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 (reexpressed as h^2_l on the liability scale for schizophrenia assuming a prevalence of 1% in adult populations), and $\hat{n}_{causal}$ is the estimated number of causal SNPs. $n_{snp} = 11,015,833$ is the total number of SNPs, whose LD and MAF underlie the model; the GWAS z-scores are for subsets of these SNPs. Though the phenotypes are diverse (examples of a categorical mental disorder, a behavioural phenotype, and a cerebral subregional tissue volume), the model nevertheless provides good fits, even though estimated polygenicities differ by two orders of magnitude and discoverabilities differ by almost three orders of magnitude. N_{samp} is the sample size, expressed as N_{eff} for schizophrenia – 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).

Posterior Effect Sizes

Model posterior effect sizes were calculated using numerical integration over the random variable δ:

$$\delta_{expected} \equiv E(\delta|z) = \int P(\delta|z)\delta d\delta = \frac{1}{P(z)} \int P(z|\delta)P(\delta)\delta d\delta. \quad (18)$$

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

$$P(z|\delta) = \phi(z; \delta, \sigma_0^2). \quad (19)$$

$P(z)$ is shorthand for $\text{pdf}(z|N, H, L; \pi_1, \sigma_\beta, \sigma_0)$, given by Eq. 17, and, also from Eq. 17, $P(\delta)$ is

$$P(\delta) \equiv \text{pdf}(\delta|N, H, L; \pi_1, \sigma_\beta) = \sum_{K=1}^{K_{max}} \sum_{i=1}^{w} \frac{n_i}{n} B(K; n; \pi_1)\phi(\delta; 0, Kr_i^2\sigma_\beta^2 NH). \quad (20)$$

Similarly,

$$\delta^2_{expected} \equiv E(\delta^2|z) = \int P(\delta|z)\delta^2 d\delta = \frac{1}{P(z)} \int P(z|\delta)P(\delta)\delta^2 d\delta, \quad (21)$$

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 that reach genome-wide significance, $p \leq 5\times10^{-8}$ (i.e., for which $|z|>z_t=5.33$) 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 $\text{var}(Y|g)=\text{var}(\beta g)=2\beta^2 H$ and $\delta=\sqrt{NH}\beta$. Using Eq. 21, let $C \equiv E(\delta^2|z, N)P(z, N)$, emphasizing dependence on sample size, N. Then the proportion of chip heritability captured additively by genome-wide significant SNPs is

$$S(N; z_t) = \frac{\sum_{z:|z|>z_t} C(z, N)}{\sum_{all\ z} C(z, N)}. \quad (22)$$

The ratio in Eq. 22 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 $S(N; z_t)$ 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 (also known as PP 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 population structure – 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. 17, 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 5×5 grid of H×TLD bins, and the cdf vector (with elements corresponding to the z-values in $\vec{z}_{nom}$) accumulated for each such bin (using mean values

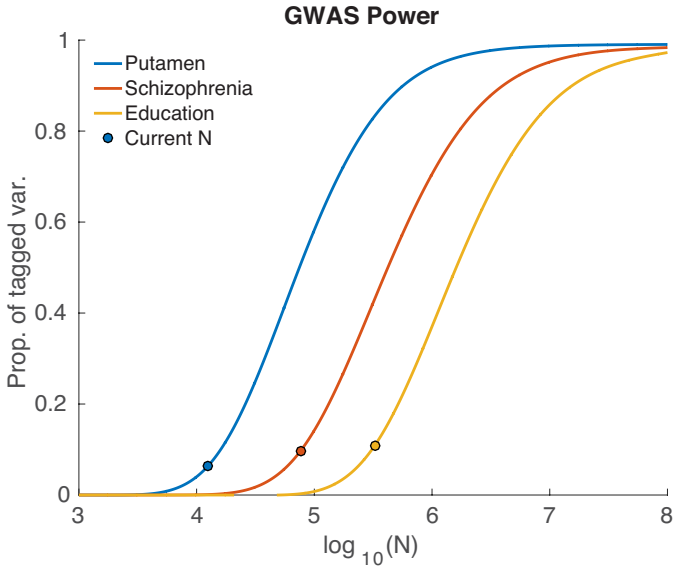

Figure 2: Proportion of narrow-sense chip heritability captured by genome-wide significant SNPs as a function of sample size, N. Left-to-right plot order is determined by decreasing σ^2_β. For current sample sizes, the proportions are: putamen volume, 0.064; schizophrenia, 0.096; educational attainment, 0.109.

of H and TLD 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 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, $q \pm 1.96\sqrt{q(1-q)/n_{tot}/f}$.

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^2_β 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

$$h^2 = \pi_1 n_{snp} \overline{H} \sigma^2_\beta. \tag{23}$$

For all-or-none traits like disease status, the estimated h^2 from Eq. 23 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^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 (Dempster and Lerner, 1950; Lee et al., 2011):

$$h^2_l = h^2 \frac{K(1-K)}{P(1-P)} \times \frac{K(1-K)}{a^2}, \tag{24}$$

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

Figure 1 shows QQ plots for the z-scores for schizophrenia, educational attainment, and putamen volume, along with model estimates. In all cases, the model fit (yellow) closely tracks the data (dark blue). Figure 6 in Supporting Material shows QQ subplots for a 5×5 grid of $H \times TLD$ ranges for schizophrenia.

The estimated number of causal SNPs is given by the polygenicity, π_1, times the total number of SNPs, n_{snp}; the latter is given by the total number of SNPs that went into building the LD structure, L in Eq. 17, i.e., the approximately 11 million SNPs selected from the 1000 Genomes Phase 3 reference panel, not the number of SNPs in the particular GWAS. Thus, for schizophrenia, $\pi_1 = 5.0 \times 10^{-3}$, so that $\hat{n}_{causal} = 5.5 \times 10^4$, not all of which are in linkage equilibrium. Educational attainment has slightly greater polygenicity than schizophrenia, $\pi_1 = 7.7 \times 10^{-3}$. In contrast, for putamen volume $\pi_1 = 2.6 \times 10^{-5}$, so that $\hat{n}_{causal} = 285$.

The effective strength of SNP association with the phenotype (mean β^2 for causals, the effective SNP "discoverability") is $\hat{\sigma}^2_\beta = 3.5 \times 10^{-5}$ for schizophrenia (in units where the variance of the phenotype is normalized to 1). It is an order of magnitude smaller for educational attainment, $\hat{\sigma}^2_\beta = 5 \times 10^{-6}$, but two orders of magnitude bigger for putamen volume: $\hat{\sigma}^2_\beta = 1.2 \times 10^{-3}$.

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	0.11 (0.01)	1E-5	1.2E-5 (2E-6)	4.3E-3 (7E-4)	3.8E-3 (8E-4)	1.01 (0.002)	110	136	(18)
0.1	0.08 (0.01)	1E-4	9.9E-5 (1E-5)	4.2E-4 (2E-5)	3.5E-4 (4E-5)	1.01 (0.002)	1101	1089	(160)
0.1	0.08 (0.01)	1E-3	1.2E-3 (2E-4)	4.2E-5 (5E-7)	3.1E-5 (5E-6)	1.02 (0.002)	11015	12846	(2543)
0.1	0.08 (0.01)	1E-2	1.3E-2 (2E-3)	4.2E-6 (4E-8)	2.8E-6 (5E-7)	1.02 (0.002)	110158	141095	(24296)
0.4	0.57 (0.05)	1E-5	1.9E-5 (2E-6)	1.7E-2 (3E-3)	1.3E-2 (2E-3)	1.01 (0.001)	110	209	(21)
0.4	0.38 (0.01)	1E-4	1.1E-4 (8E-6)	1.7E-3 (7E-5)	1.4E-3 (1E-4)	1.03 (0.003)	1101	1226	(92)
0.4	0.32 (0.01)	1E-3	9.8E-4 (6E-5)	1.7E-4 (2E-6)	1.4E-4 (9E-6)	1.05 (0.003)	11015	10753	(705)
0.4	0.31 (0.01)	1E-2	1.3E-2 (5E-4)	1.7E-5 (2E-7)	9.7E-6 (4E-7)	1.06 (0.003)	110158	145964	(5959)
0.7	1.07 (0.09)	1E-5	2.3E-5 (2E-6)	3.0E-2 (5E-3)	2.0E-2 (3E-3)	1.02 (0.002)	110	257	(23)
0.7	0.71 (0.03)	1E-4	1.2E-4 (8E-6)	2.9E-3 (1E-4)	2.4E-3 (2E-4)	1.05 (0.003)	1101	1349	(93)
0.7	0.58 (0.02)	1E-3	9.6E-4 (5E-5)	2.9E-4 (4E-6)	2.6E-4 (1E-5)	1.08 (0.006)	11015	10524	(526)
0.7	0.52 (0.01)	1E-2	1.3E-2 (4E-4)	2.9E-5 (3E-7)	1.6E-5 (4E-7)	1.09 (0.005)	110158	145692	(4605)

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 $Y{=}Y_g{+}\epsilon$, 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.

Note that for logistic linear regression coefficient β, the odds ratio for disease is $\mathrm{OR} = e^\beta$; for a rare disease, this is approximately equal to the genotypic relative risk: $\mathrm{GRR} \simeq \mathrm{OR}$. 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, the mean relative risk is $\simeq 1.0000175$.

The narrow sense heritability from the ascertained case-control schizophrenia GWAS is estimated as h^2=0.41 (with mean heterozygosity from the ∼11 million SNPs, $\overline{H} = 0.2165$). Taking adult population prevalence of schizophrenia to be K=0.01 (Purcell et al., 2009; Whiteford et al., 2013), and given that there are 35,476 cases and 46,839 controls in the study, so that P=0.43, the heritability on the liability scale for schizophrenia from Eq. 24 is h^2_l=0.23; for K=0.005 (Kinney et al., 2009), h^2_l=0.20. For the quantitative endophenotype putamen volume, the heritability is estimated to be 7%, while for educational attainment the heritability is estimated to be 10%.

Figure 2 shows the sample size required to reach a given proportion of chip heritability for the phenotypes (assuming equal numbers of cases and controls for schizophrenia: $N_{eff} = 4/(1/N_{cases} + 1/N_{controls})$, so that when $N_{cases} = N_{controls}$, $N_{eff} = N_{cases} + N_{controls} = N$, the total sample size). At current sample sizes, only 10%, 10%, and 7% of narrow-sense chip heritability is captured for schizophrenia, educational attainment, and putamen volume, respectively. And to capture the preponderance of chip heritability for schizophrenia, for example, a sample with approximately half a million each of cases and controls would be needed.

The estimated total inflation factor for the pruned data, λ, is almost exactly predicted by the model. E.g., for schizophrenia, $\lambda = \hat{\lambda} = 1.38$, whereas for educational attainment the values are $\lambda = 1.16$ and $\hat{\lambda} = 1.17$. Higher polygenicity, π_1, mean strength of association, σ^2_β, and sample size, N, will all contribute to higher λ. Residual population structure will also contribute to genomic inflation. For schizophrenia, inflation from population structure is estimated to be $\hat{\sigma}^2_0 = 1.179$. In contrast, for educational attainment $\hat{\sigma}^2_0 = 1.01$, indicating essentially no residual inflation due to population structure.

Simulations

Table 1 shows the simulation results, comparing true and estimated values for the model parameters, heritability, and the number of causal SNPs. In supporting material, Figure 3 shows QQ plots for a randomly chosen β-vector and phenotype instantiation for each of the twelve (π_1, h^2) scenarios. Most of the $\hat{\pi}_1$ estimated are in reasonable agreement with the true values, though for $\pi_1 = 10^{-5}$ they are larger by about a factor of two for h^2 equal to 0.4 and 0.7. The number of estimated causals are in correspondingly good agreement with the true values, ranging in increasing powers of 10 from 110 through 110,158. While the estimated polygenicities tend to be slight over-estimates, the estimated discoverabilities, $\hat{\sigma}^2_\beta$'s, tend to be under-estimates. From Supporting Material Figure 3, the tails of the QQ plots for the true parameters (dashed dark blue curves), particularly for the larger π_1's, deviate from the simulated data plots (solid dark blue curves), consistently over-estimating the proportion of SNPs with more extreme z-scores. The model fit, however, bends these curves down toward the data curves. Note that steeper tails have larger σ^2_β's, and larger π_1's lead to earlier departure from the null line. In all cases, σ^2_0 is close to 1, indicating no population structure. Estimates of heritability, $\hat{h}^2$, show a tendency to decrease with increasing π_1. In all cases, however, the value for genomic control, λ, es-

timated from the model is in very good agreement with the value estimated from the simulated data; these values increase both as π_1 and σ_β^2 (or h^2, for fixed π_1) increase. E.g., for $\pi_1 = 10^{-5}$ and $h^2 = 0.1$, $\lambda = 1.01$ and $\hat{\lambda} = 1.01$, while for $\pi_1 = 10^{-2}$ and $h^2 = 0.7$, $\lambda = 1.26$ and $\hat{\lambda} = 1.25$.

DISCUSSION

Building on our previous work and the work of others, 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 residual inflation of the association statistics due to sources of bias like uncorrected population structure, σ_0^2.

We apply the model to three diverse phenotypes, one qualitative and two quantitative: schizophrenia, educational attainment, putamen volume. In each case, we estimate the polygenicity, discoverability, and residual inflation due to bias; we also estimate the number of causal SNPs, n_{causal}, and the SNP heritability, h^2 (for schizophrenia, we reexpress 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 $\pi_1 = 5 \times 10^{-3}$. This leads to an estimate of $n_{causal} \simeq 55,000$, which is in scale-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, $\sigma_\beta^2 = 3.5 \times 10^{-5}$, indicating that most associations are of week effect, i.e., have low discoverability.

For educational attainment (Rietveld et al., 2013; Okbay et al., 2016; Cesarini and Visscher, 2017), the polygenicity is somewhat greater, $\pi_1 = 7.7 \times 10^{-3}$, leading to an estimate of $n_{causal} \simeq 85,000$, which also is in scale-agreement with a recent estimate of the number of loci contributing to heritability of $\simeq 70,000$ (Rietveld et al., 2013). The variance of the distribution for causal effect sizes is an order of magnitude smaller than for schizophrenia, $\sigma_\beta^2 = 5 \times 10^{-6}$, indicating lower discoverability.

In marked contrast is putamen volume, which has very low polygenicity: $\pi_1 = 2.6 \times 10^{-5}$, so that only 285 SNPs (out of ∼11 million) are estimated to be causal. However, these SNPs are characterized by high discoverability, two-orders of magnitude larger than for schizophrenia: $\sigma_\beta^2 = 1.2 \times 10^{-3}$.

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.

Despite the much stronger effects in putamen volume, the very high polygenicity for schizophrenia leads to its being more than three times as heritable. Our point estimate for liability-scale heritability of schizophrenia is $h_l^2 = 0.23$ (assuming a population risk of 0.01), and that 10% of this (i.e., 2.3% 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 good agreement with a recent result, $h_l^2 = 0.27$ (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 $h_l^2 = 0.23$ 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 2.3% of overall variation on the liability scale for schizophrenia explainable by genome-wide significant loci is a little lower than the corresponding estimate of 3.4% based on risk profile scores (RPS) (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). 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 thst 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). Noteworthy in this respect is 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_β^2 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 2) will shift to the left.

For educational attainment, we estimate SNP heritability $h^2 = 0.10$, in good agreement with the estimate of 11.5% given in (Okbay et al., 2016). As with schziophrenia, 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 $h^2 = 0.07$, in reasonable agreement with an earlier estimate of 0.1 for the same overall data set (Hibar et al., 2015; So et al., 2011).

To assess the validity of the model, we conduct extensive simulations over a wide range of polygenicities and heritabilities for simulated quantitative traits, using the full set of SNPs used in the phenotype analyses with realistic LD structure. The simulations in general validate the model: with the true number of causals ranging over three orders of magnitude, 10^2-10^5 (while heritabilities range from 0.1 to 0.7), the estimated number of causals in each case is in reasonable agreement with the corresponding true value. Similarly, the true σ_β^2's range over four orders of magnitued, and the estimated values are generally well within a factor of two of the corresponding true value. It should be noted that for all simulations, σ_0^2 is close to 1.0 (indicating no confounding inflation, as expected in HapGen), though there is a trend toward larger values for higer heritability and polygenicity. Thus, the higher inflation found for schizophrenia is unlikely to be an artifact of the model. The simulation QQ model plots in general agree with the simulation QQ data plots, though there is an overestimation of the proportion of more extreme z-scores, particularly at very high polygenicities. This might be an artifact of using the computationally simpler but less accurate Eq. 17 instead of Eq. 14, which is currently a limitation in the implementation of the model. A Monte Carlo approach to calculating the pdf in Eq. 14 might lead to more accurate QQ model plots.

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. The model was not designed to handle situations where the reversal of short sections of DNA underlies SNP association, as appears to be the case for some phenotypes, e.g, in chromosome 8p for neuroticism (Lo et al., 2016). 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 Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC) for making available their GWAS summary statistics for schizophrenia; the Enhancing Neuro Imaging Genetics through Meta Analysis Consortium (ENIGMA) for making available their GWAS summary statistics for putamen volume; and the Social Science Genetic Association Consortium (SSGAC) for GWAS summary statistics on educational attainment.

Funding

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

References

1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature 467 (7319), 1061–1073.

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.

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.

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

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.

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., 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.

Kang, H. M., Sul, J. H., Service, S. K., 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.

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.

Lo, M.-T., Hinds, D. A., Tung, J. Y., Franz, C., Fan, C.-C., Wang, Y., Smeland, O. B., Schork, A., Holland, D., Kauppi, K., et al., 2016. Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. Nature genetics.

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.

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.

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.

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.

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.

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. 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.

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.

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.

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.

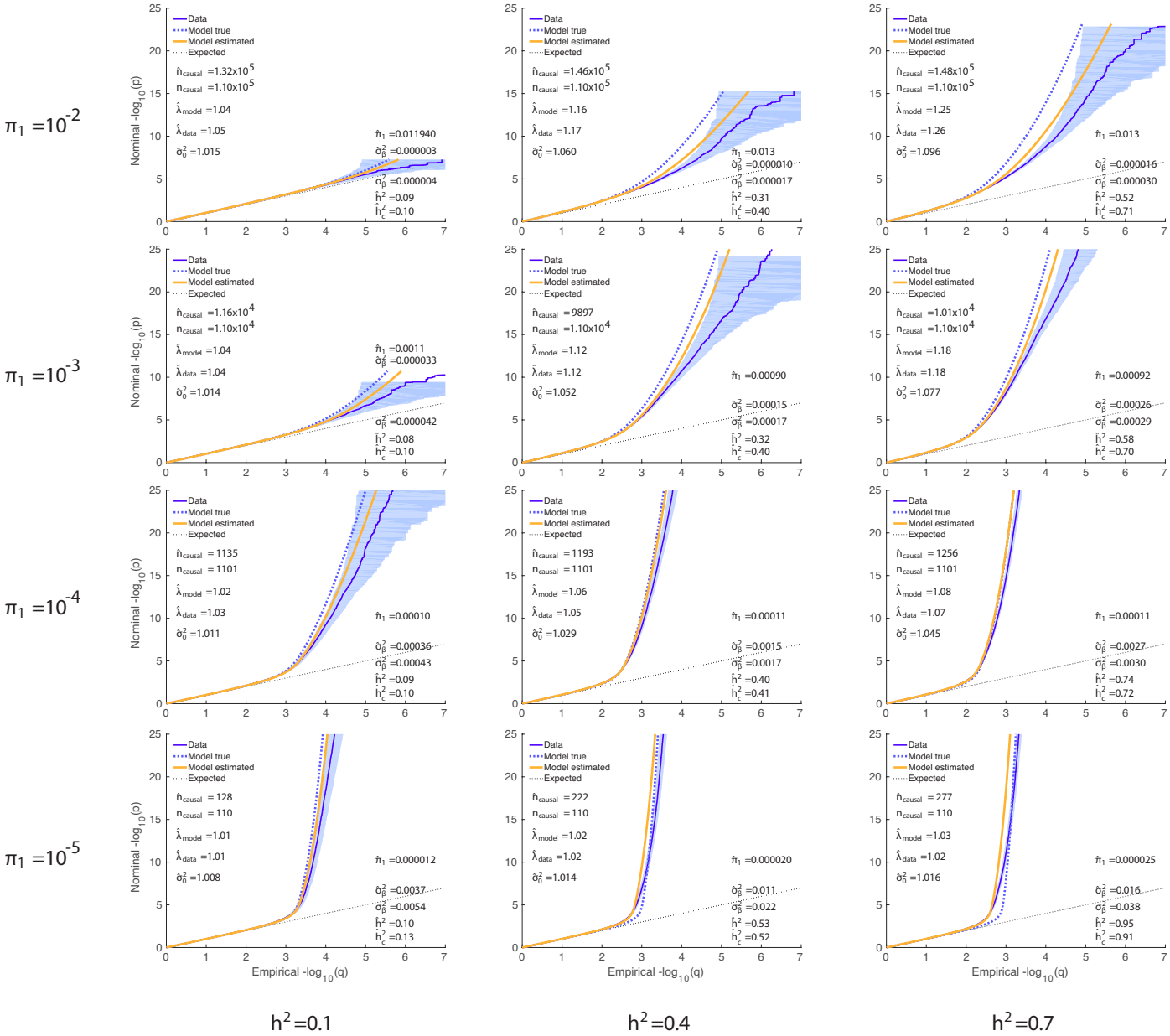

Figure 3: Quantile-quantile plots for simulations. True polygenicity is specified for each row, and true heritability is specified for each column. QQ-plots for simulated data in dark blue, with 95% confidence interval in light blue; model prediction in yellow. The dashed blue curve is the QQ plot corresponding to the true parameters. $\lambda \equiv \hat{\lambda}_{data}$ and $\hat{\lambda}_{model}$ are the overall nominal genomic control factors calculated from the plots. 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, and $\hat{n}_{causal}$ is the estimated number of causal SNPs. The dotted black line is the expected plot under null. $\hat{h}^2_c$ is the same as $\hat{h}^2$ but with $\overline{H}$ calculated from the known causal SNPs (instead of from all SNPs). 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).

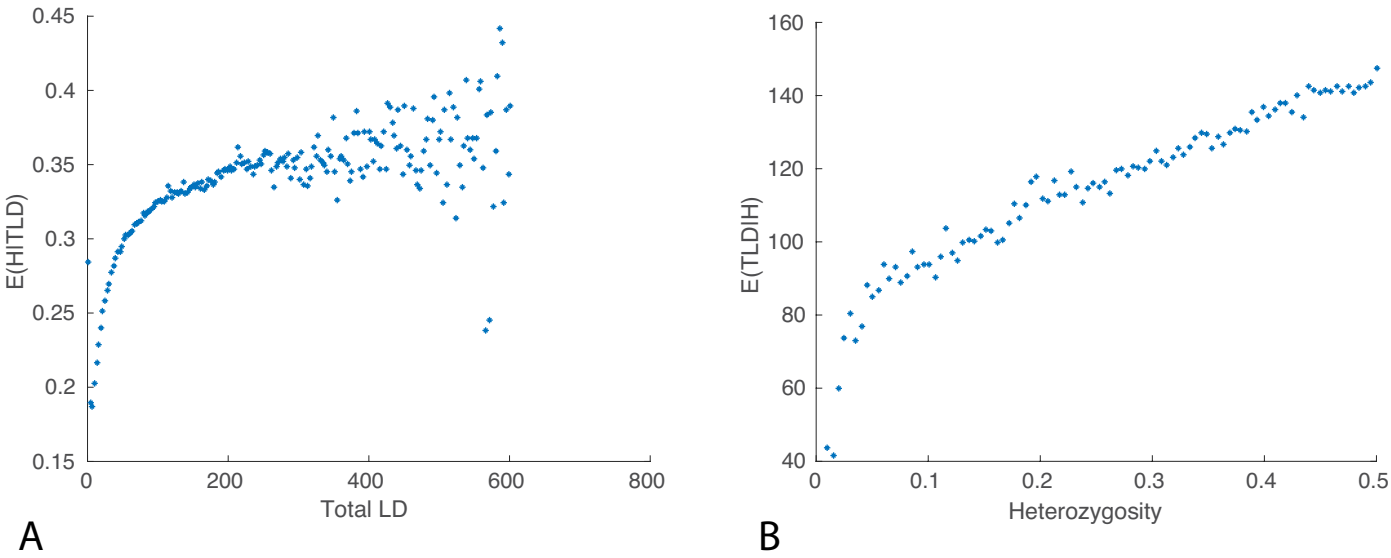

Figure 4: (A) Mean value of heterozygosity for given total LD (SNPs were binned based on TLD and the mean TLD for each bin plotted on the x-axis; the corresponding mean heterozygosity for SNPs in each bin was then plotted on the y-axis). (B) Mean value of total LD for given heterozygosity. Plots made for SNPs in the PGC2 schizophrenia GWAS; TLD and H calculated from 1000 Genomes phase 3 reference panel.

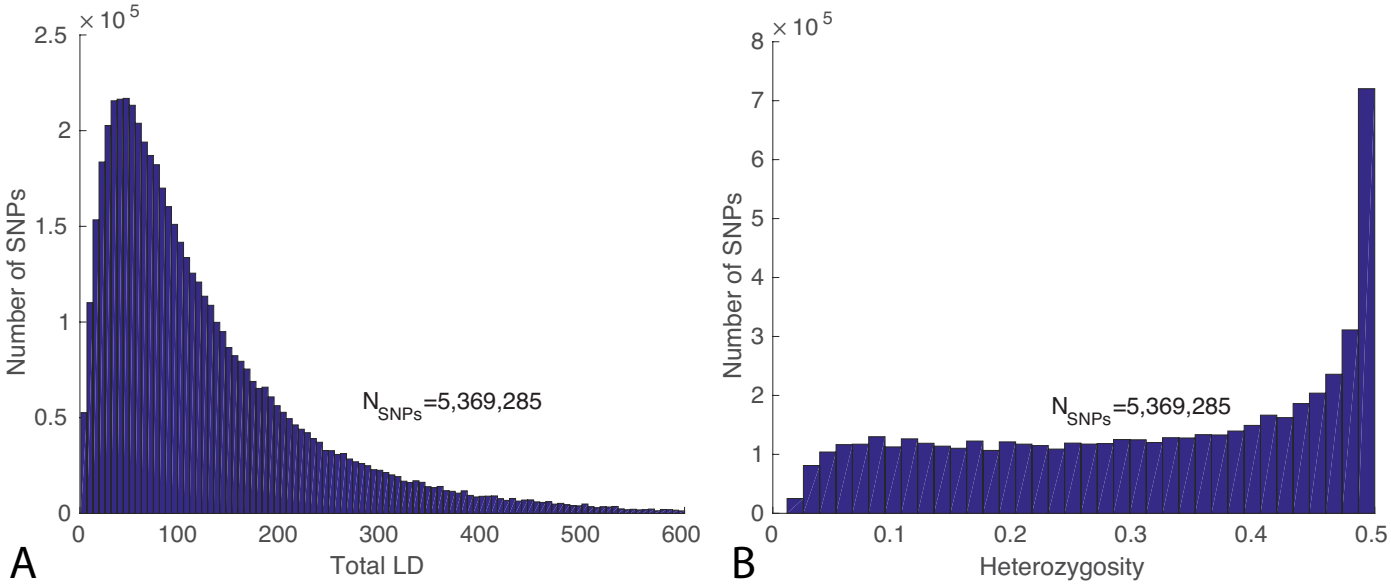

Figure 5: Histograms of SNPs in schizophrenia GWAS, by (A) total LD, and (B) heterozygosity.

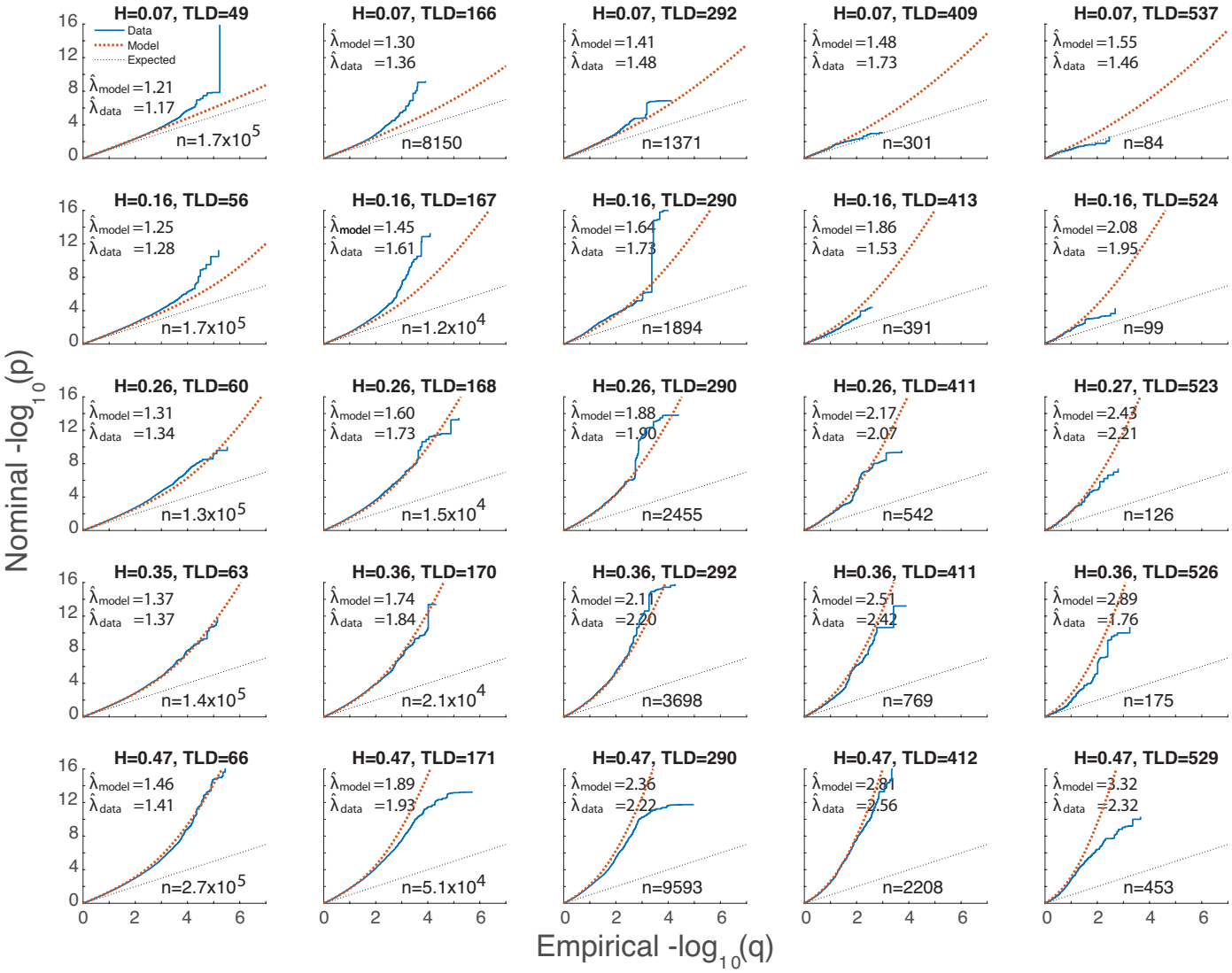

Figure 6: QQ plots for schizophrenia, for a 5X5 grid of total LD X heterozygosity. n is the number of SNPs in each plot. H and TLD are the mean values in each plot. $\hat{\lambda}_{data}$ and $\hat{\lambda}_{model}$ are the genomic control values calculated from the QQ plots for the data and the model, respectively.