Title

Estimating Effect Sizes and Expected Replication Probabilities from GWAS Summary Statistics.

Journal

Frontiers in genetics, 7(FEB)

ISSN

1664-8021

Authors

Holland, Dominic
Wang, Yunpeng
Thompson, Wesley K
et al.

Publication Date

2016

DOI

10.3389/fgene.2016.00015

Peer reviewed

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 February 2016
doi: 10.3389/fgene.2016.00015

Estimating Effect Sizes and Expected Replication Probabilities from GWAS Summary Statistics

Dominic Holland[1,2], Yunpeng Wang[1,2,3,4], Wesley K. Thompson[5], Andrew Schork[1,6], Chi-Hua Chen[1,7], Min-Tzu Lo[1,7], Aree Witoelar[3,4], Schizophrenia Working Group of the Psychiatric Genomics Consortium, Enhancing Neuro Imaging Genetics through Meta Analysis Consortium, Thomas Werge[8], Michael O'Donovan[9], Ole A. Andreassen[3,4] and Anders M. Dale[1,2,5,7]*

[1] Multimodal Imaging Laboratory, University of California, San Diego, La Jolla, CA, USA, [2] Department of Neurosciences, University of California, San Diego, La Jolla, CA, USA, [3] NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway, [4] Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway, [5] Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA, [6] Department of Cognitive Sciences, University of California, San Diego, La Jolla, CA, USA, [7] Department of Radiology, University of California, San Diego, La Jolla, CA, USA, [8] Institute of Biological Psychiatry, MHC, Sct. Hans Hospital and University of Copenhagen, Copenhagen, Denmark, [9] MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Cardiff, UK

OPEN ACCESS

Edited by:
Steven J. Schrodi, Marshfield Clinic Research Foundation, USA

Reviewed by:
Ina Hoeschele, Virginia Tech, USA
Duncan C. Thomas, University of Southern California, USA

***Correspondence:**
Dominic Holland
dominic.holland@gmail.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *16 December 2015*
Accepted: *28 2016*
Published: *16 February 2016*

Citation:
Holland D, Wang Y, Thompson WK, Schork A, Chen C-H, Lo M-T, Witoelar A, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Enhancing Neuro Imaging Genetics through Meta Analysis Consortium, Werge T, O'Donovan M, Andreassen OA and Dale AM (2016) Estimating Effect Sizes and Expected Replication Probabilities from GWAS Summary Statistics. Front. Genet. 7:15. doi: 10.3389/fgene.2016.00015

Genome-wide Association Studies (GWAS) result in millions of summary statistics ("z-scores") for single nucleotide polymorphism (SNP) associations with phenotypes. These rich datasets afford deep insights into the nature and extent of genetic contributions to complex phenotypes such as psychiatric disorders, which are understood to have substantial genetic components that arise from very large numbers of SNPs. The complexity of the datasets, however, poses a significant challenge to maximizing their utility. This is reflected in a need for better understanding the landscape of z-scores, as such knowledge would enhance causal SNP and gene discovery, help elucidate mechanistic pathways, and inform future study design. Here we present a parsimonious methodology for modeling effect sizes and replication probabilities, relying only on summary statistics from GWAS substudies, and a scheme allowing for direct empirical validation. We show that modeling z-scores as a mixture of Gaussians is conceptually appropriate, in particular taking into account ubiquitous non-null effects that are likely in the datasets due to weak linkage disequilibrium with causal SNPs. The four-parameter model allows for estimating the degree of polygenicity of the phenotype and predicting the proportion of chip heritability explainable by genome-wide significant SNPs in future studies with larger sample sizes. We apply the model to recent GWAS of schizophrenia ($N = 82{,}315$) and putamen volume ($N = 12{,}596$), with approximately 9.3 million SNP z-scores in both cases. We show that, over a broad range of z-scores and sample sizes, the model accurately predicts expectation estimates of true effect sizes and replication probabilities in multistage GWAS designs. We assess the degree to which effect sizes are over-estimated when based on linear-regression association coefficients. We estimate the polygenicity of schizophrenia to be 0.037 and the putamen to be 0.001, while the respective sample sizes required to approach fully explaining the chip heritability

are 10^6 and 10^5. The model can be extended to incorporate prior knowledge such as pleiotropy and SNP annotation. The current findings suggest that the model is applicable to a broad array of complex phenotypes and will enhance understanding of their genetic architectures.

Keywords: GWAS, Gaussian mixture model, effect size, schizophrenia, SNP discovery, putamen, heritability

INTRODUCTION

Many complex traits and common phenotypes have a genetic component that arises from large numbers of genetic loci (Visscher et al., 2012). The total effect of the genetic component on phenotypic expression is often substantial, as indicated by measures of heritability (Tenesa and Haley, 2013; Witte et al., 2014) obtained from twin and family studies and genome-wide association studies (GWAS) for multiple phenotypes. For example, heritability of schizophrenia is estimated to be ∼80% from twin studies (Sullivan et al., 2003), ∼64% from family studies (Lichtenstein et al., 2009; Wray and Gottesman, 2012), with a lower bound of ∼33% from recent GWAS (Ripke et al., 2013a). For any phenotype, GWAS provide a platform for uncovering the underlying genetic architecture, but this poses a substantial challenge, compounded by the complexity of the datasets: $\sim 10^4$–10^5 individuals with $\sim 10^7$ genetic markers (single nucleotide polymorphisms, or SNPs) in various levels of correlation (linkage disequilibrium, or LD), $\sim 10^6$ of which are estimated to be independent (Dudbridge and Gusnanto, 2008; Pe'er et al., 2008), with multiple possible roles for SNPs in mechanistic pathways.

Mathematical modeling is important for statistical genetics to capture, both broadly and in detail, the complexity of the datasets (Schork, 2002; So et al., 2010; Stahl et al., 2012). Indeed, with the number of markers much larger than number of individuals in GWAS, modeling assumptions are required so as to estimate parameters of interest and thereby obtain realistic descriptions of the numbers, distributions, and effect sizes of causal SNPs—and the considerably larger number of SNPs in strong LD with causal SNPs—which in turn can assist in causal SNP discovery and individual risk prediction, and inform mechanistic understanding of genetic effects in phenotypic expression.

Better understanding of the genetic architecture of complex traits will be facilitated both by many more individuals being genotyped, by fine-mapping, and by developing more advanced and realistic modeling techniques. Standard GWAS approaches, however, are designed for discovering a small number of common variants with relatively large effects (i.e., low polygenicity), and so are not optimized for analyzing the large numbers of small effects in highly polygenic traits. Thus, there is a need for the development of analytical methods appropriate for the many phenotypes that have, or are expected to have, high polygenicity (Andreassen et al., 2013a,b; Schork et al., 2013).

Recently, methods have been developed to explore the combined contributions of many low-penetrance effects that do not reach genome-wide significance at current sample sizes. These include polygenic risk score profiling (Purcell et al., 2009; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014); mixed linear modeling to estimate the genetic variance in unrelated individuals, where the distribution of effect sizes is modeled as a single normal (Yang et al., 2010; Lee et al., 2011, 2012a); a related Bayesian hierarchical model where the z-scores (or summary statistics for SNP association with phenotype), given the effect sizes, are assumed to follow a single normal distribution (So et al., 2011); modeling the distribution of the estimated genetic variance of *known* discoveries for a trait as a mixture of exponentials distribution, analogous to a scale mixture of normals distribution for the regression coefficients (Park et al., 2011); and an analysis that combines this later work with polygenic risk score profiling and heritability estimates from GWAS (Chatterjee et al., 2013). Additionally, multivariate linear mixed models have been developed (Yang et al., 2011a; Speed and Balding, 2014; Zhou and Stephens, 2014). The focus here, however, is on standard univariate analysis, but the empirical method for estimating regression coefficients in replication samples is also applicable to those arising from multivariate analysis.

Mixture densities (Efron, 2013), particularly mixtures of normals, have previously been used in various forms to estimate effect sizes from individual-level data (Meuwissen et al., 2001; Goddard et al., 2009; Erbe et al., 2012; Zhou et al., 2013). Here we expand on a version of this relatively simple model for the distribution of z-scores (Thompson et al., 2015), and apply it to genome-wide summary statistics for schizophrenia and also to putamen volume, which provides an illustrative contrast.

One of our main objectives is to model, in a descriptive and accurate yet parsimonious way, the distribution of genetic summary statistics of traits for which a significant portion of the genome is involved, and thus help illuminate the genetic architecture of polygenic traits. To test for accuracy, we present a methodology for non-parametric estimation of quantities of interest which can then be compared with model predictions. This includes obtaining realistic estimates of the true effect size given the observed z-score and minimal other information: sample size and the model parameters that capture the statistics of the distribution—essentially, estimates of the conditional expectancy of regression coefficients β given z. Of particular importance, we want accurately to predict replication probabilities in multistage GWAS, i.e., the probability that a given z-score in a discovery sample will pass a nominal p-value significance threshold in a replication sample (that might include the discovery sample as a subset), a quantity that has not hitherto been a focus of much research. This quantity requires knowing the full distribution of test-statistics. In line with the parsimony of the model, the parameters will be directly interpretable—for example, one gives an index of the polygenicity—and being able

accurately to estimate them and their uncertainties is a central component in this study. These will be used, for example, in power calculations to predict the proportion of additive chip heritability (which in turn is the proportion of phenotypic variance explainable by additive genetic effects of common SNPs assayed by GWAS arrays) that is discoverable as a function of sample size. (Additive chip heritability arises from additive contributions to phenotypic variance from tagged SNPs; below we will interchangeably refer to proportion of chip heritability and proportion of tagged variance explained by genome-wide significant SNPs.) Other recently developed methods that enable estimating chip heritability and proportion of variance explained are LD Scoring (Bulik-Sullivan et al., 2015) and Additive Variance Explained and Number of Genetic Effects Method of Estimation (AVENGEME) (Palla and Dudbridge, 2015).

It is the relatively large effects that GWAS have discovered in recent years, yet for many phenotypes it appears that very large numbers of much smaller effects remain unidentified (Yang et al., 2010; Sklar et al., 2011; Ripke et al., 2013a,b). Thus, it will be necessary to include large (or sparse) and small (or ubiquitous) effects, a breakdown that naturally can be captured in a mixture model for the distribution of SNP z-scores: one Gaussian for each. (Since only a very small fraction of all SNPs are expected to be causal or mechanistically associated with a given phenotype, we use the word "sparse" to characterize these. Since many other SNPs can be expected to exhibit *attenuated* apparent effects through LD with the sparse SNPs, we also use the word "large" as a descriptor for the sparse effects. In contrast, we use "ubiquitous" and "small" to characterize the SNPs in LD with the sparse SNPs and their attenuated apparent effects.) In these Gaussians, it is important to incorporate the allele SNP heterozygosity (variance of the allele count). Null and non-null effects distributed throughout the genome could be captured by using only a single Gaussian; a two-groups mixture of Gaussians distribution has been used for ubiquitous null and sparse non-null effects; modifying this slightly will additionally allow for ubiquitous non-null effects: dedicating one Gaussian to ubiquitous null and non-null (small) effects, and the other to sparse (large) non-null effects. Intuitively, sparse effects represent SNPs that are in strong LD with causal SNPs (or more generally, with SNPs that are mechanistically associated with the phenotype), while the ubiquitous non-null effects largely arise from weak LD with causal SNPs, and the null effects—effects that do not replicate—arise from environmental and error contributions.

Here, we develop a unified framework for power calculations, relying on only four parameters and their uncertainties, that enables prediction of effect sizes, replication probabilities, and fraction of chip heritability explained by genome-wide significant SNPs as a function of sample size. We show that model variations that do not take into account ubiquitous non-null effects do not provide a good match to actual data, and propose a minor modification with important implications for the discovery of SNPs affecting phenotypes. We apply the model to the Psychiatric Genomics Consortium (PGC; Sullivan, 2010) schizophrenia sample: 35,476 cases and 46,839 controls 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 9.3 million genotyped and imputed SNPs (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). We also apply the model to putamen volume using data from the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) consortium (Hibar et al., 2015), with 12,596 subjects and the same set of SNPs as for schizophrenia. For these two phenotypes, using nonparametric methods described herein, we directly compare empirical with model results for expected replication effect sizes, variances, and replication probabilities, as a function of sample size, and map out the estimated proportion of chip heritability explained by genome-wide significant SNPs as a function of sample size.

METHODS

Proposed Gaussian Mixture Model

Assuming a linear relationship between a quantitative phenotype and genotype (logistic relationship for case-control designs), a massively univariate, or marginal regression, approach with effective sample size N (see Supplementary Material for definition) shows that the z-score for a given SNP can be written as a sum of genetic effect, δ, and a remainder term, ϵ, encompassing environmental and error contributions, assumed to be independent of δ: $z = \delta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_0^2)$, σ_0^2 possibly being slightly different from 1 due to population substructure (Devlin and Roeder, 1999). In the context of the model, δ is the "true" effect size. Assuming Hardy-Weinberg equilibrium, for any particular SNP the effect size $\delta \propto \sqrt{N \cdot H}$, where H is the heterozygosity (allele count variance in the population), $H = 2p(1-p)$, p being the allele frequency for either of the two SNP alleles (see Supplementary Material for further details). Thus, the variance of z is $\mathrm{var}(z) = \mathrm{var}(\delta) + \sigma_0^2$, with $\mathrm{var}(\delta) \propto N \cdot H$. We introduce a four-parameter two-component Gaussian mixture model for the marginal distribution of z-scores assuming SNPs belong to a class of ubiquitous effects or to a class of sparse effects:

$$f(z) = \pi_0 \phi(z, 0, \sigma_0^2 + \sigma_1^2) + \pi_1 \phi(z, 0, \sigma_0^2 + \sigma_1^2 + \sigma_2^2). \quad (1)$$

Here, π_0 is the prior probability (after uniform pruning so that large LD blocks are not over-represented with respect to small LD blocks—see below) that a SNP is in the ubiquitous class ($\pi_0 \approx 1$) of "small" replicating effects (described by σ_1^2); $\pi_1 = 1 - \pi_0$ is the prior probability that a SNP is in the sparse class of "large" replicating effects (described by $\sigma_1^2 + \sigma_2^2$), i.e., π_1 is the fraction of independent SNPs characterized by the broader (sparse) normal probability distribution function (PDF), which we denote the index of polygenicity ($\pi_1 \ll 1$); $\phi(\cdot, \mu, \sigma^2)$ is the normal PDF with mean μ and variance σ^2; all SNPs have a component that is a null effect, i.e., a non-replicating error/environmental contribution (associated with σ_0^2); $\sigma_1^2 = \sigma_a^2 N \cdot H$ is the additional variance associated with non-null ubiquitous effects, and $\sigma_2^2 = \sigma_b^2 N \cdot H$ is the additional variance associated with sparse effects, with σ_a^2 and σ_b^2 being the corresponding per-allele variances (assumed to be independent of allele frequency). The four parameters of the

model then are $\pi_1, \sigma_0^2, \sigma_a^2$, and σ_b^2. In the two-groups formalism, the non-null effect size is given by

$$\delta = \begin{cases} \delta_a, & \text{with prior prob. } \pi_0, \\ \delta_a + \delta_b, & \text{with prior prob. } \pi_1, \end{cases} \quad (2)$$

with δ_a and δ_b independent. The components of δ can be written as

$$\begin{aligned} \delta_a &= a\sqrt{N \cdot H}, \quad a \sim \mathcal{N}(0, \sigma_a^2), \\ \delta_b &= b\sqrt{N \cdot H}, \quad b \sim \mathcal{N}(0, \sigma_b^2). \end{aligned} \quad (3)$$

The more usual two-groups mixture model would have $\sigma_a \equiv 0$, so that SNPs would be categorized as either truly null and ubiquitous, or non-null and sparse (Efron, 2013; see also Greenland and Poole, 2013 for arguments against "spike ad smear" priors).

In the usual terminology applied to two-groups mixture models (Efron, 2013), the Bayesian local true discovery rate, $tdr(z)$, is the probability that z corresponds to the π_1 arm of Equation 1, i.e., the posterior probability that a SNP with this z-score has a sparse effect:

$$\begin{aligned} tdr(z) &= Pr(b \neq 0|z) \\ &= \pi_1 \phi(z, 0, \sigma_0^2 + \sigma_1^2 + \sigma_2^2)/f(z). \end{aligned} \quad (4)$$

Then, from Equation 1, it is easy to show that the posterior distribution of effect sizes δ, given z, is a weighted sum of sparse and ubiquitous contributions

$$\begin{aligned} Pr(\delta|z) = (1 - tdr(z))\phi(\delta, \mu_u(z), \sigma_u^2) + \\ tdr(z)\phi(\delta, \mu_s(z), \sigma_s^2) \end{aligned} \quad (5)$$

where,

$$\mu_u(z) = z\sigma_1^2/(\sigma_0^2 + \sigma_1^2) \quad (6)$$

$$\mu_s(z) = z(\sigma_1^2 + \sigma_2^2)/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2) \quad (7)$$

$$\sigma_u^2 = \sigma_0^2\sigma_1^2/(\sigma_0^2 + \sigma_1^2) \quad (8)$$

$$\sigma_s^2 = \sigma_0^2(\sigma_1^2 + \sigma_2^2)/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2) \quad (9)$$

(see Supplementary Material for details). Note that from Equations 2, 3, and 5,

$$\begin{aligned} E(\delta_a|z) = z \cdot [tdr(z) \cdot \sigma_1^2/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2) + \\ (1 - tdr(z)) \cdot \sigma_1^2/(\sigma_0^2 + \sigma_1^2)], \end{aligned} \quad (10)$$

$$E(\delta_b|z) = z \cdot tdr(z) \cdot \sigma_2^2/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2). \quad (11)$$

where E denotes expectation. Alternatively, based on Equation 5, we can write $\delta = \delta_u + \delta_s$ for non-sparse and sparse effects, respectively, where

$$E(\delta_u|z) = (1 - tdr(z))\mu_u(z) \quad (12)$$

$$E(\delta_s|z) = tdr(z)\mu_s(z). \quad (13)$$

The objective, then, is to determine the empirical distribution of z-scores and the four model parameters that best characterize the data, assess the quality of the resulting model fit, and address its implications. Below we describe how the replication probability can be measured empirically. Of particular interest will be the accuracy of the model prediction of this quantity.

As there might be confusion about the terms bias and effect size, a note of clarification is in order. A true regression coefficient β for association between genotype and phenotype (see Supplementary Material) can be thought of as a true effect size. An estimate $\hat{\beta}$ of this from data will then be an estimated effect size. Simple linear regression, the standard approach in GWAS, provides an unbiased estimate of effect sizes. That is, the marginal expectation of $\hat{\beta}$ is β: $E(\hat{\beta}) = \beta$. This is usually what is understood when the term "unbiased" is used for estimated quantities. However, the quantity of practical interest in association studies, because it is directly related to predicted z-scores in replication samples, is the conditional expectancy of the true effect size, given the estimate from GWAS: $E(\beta|\hat{\beta})$. This quantity we call the adjusted effect size; our results below show that $|E(\beta|\hat{\beta})| < |\hat{\beta}|$. In other words, $\hat{\beta}$ provides an inflated estimate of the true effect size. We show below that the adjusted effect size can be expressed as a function of $\hat{\beta}$, sample size N, heterozygosity H, and the four model parameters: $E(\beta|\hat{\beta}) = f(\hat{\beta}, N, H; \sigma_0, \sigma_a, \sigma_b, \pi_1)$. Since the Wald statistic (or z-score) corresponding to $\hat{\beta}$ is simply a scaling of $\hat{\beta}$, $z = \hat{\beta}/\text{se}(\hat{\beta})$, we will also refer to the component δ of z as the adjusted effect size, with context making it clear what is meant. (Note that "effect size" is also often used to denote the portion of phenotypic variance due to genotype, i.e., $\hat{\beta}^2 H$ Park et al., 2011). We will show below that δ provides an unbiased estimate for the true effect size: given a discovery sample z-score z_d for a particular SNP, the expected value for the z-score in a replication sample is $\delta = E(z_r|z_d)$.

Finally, we note that natural variation in genotypes vis-a-vis phenotype does not lead to bias but will effect power: larger variation (larger se(β)) will result in smaller Wald statistics, all else equal. Genotype measurement error, however, may induce correlation among the genotypes and have a biased effect, operating in a manner similarly to population structure, resulting in a positive contribution to the null variance component σ_0. Independent of correlated errors, random genotyping errors will lead to underestimation of the regression coefficient—the effect of regression dilution, or regression attenuation (Fuller, 2009).

Empirical Estimation

We analyzed summary statistics (z-scores) from the PGC schizophrenia sample of 35,476 cases and 46,839 controls across 52 separate substudies, with 9,279,485 genotyped and imputed SNPs; restricting allele frequency to be greater than 0.005 reduced the number of SNPs by 2% (to 9,083,435). We randomly and repeatedly divide the data into complementary discovery and replication sets, and calculate the empirical expected z-score, and the expected square of the z-score, in the replication set, given a z-score in the discovery set. We also calculate empirical posterior estimates of the variance of the effect size in the replication set given a z-score in the discovery set, and the replication probability (defined below) for z-scores in the discovery set.

Note that while π_1 and σ_0^2 do not depend on sample size N, σ_1^2 and σ_2^2 are proportional to N. Thus, to separate out these contributions and to examine the effects of sample size on posterior expectation values, the division of the data into complementary discovery and replication sets is not just a single division, e.g., split-half. Rather, we repeatedly and randomly divide the data with 10% of the effective sample in the discovery set (90% replication), then do the same in increments of 10%, through 90% of the effective sample in the discovery set (10% in replication). In the current work we do not use raw genotype data, but summary statistics from 52 studies with uneven distribution of sample sizes. So, for example, random draws are made from the 52 studies so that the studies selected for the discovery set will comprise approximately 10% of the total effective sample size, and so on for the other percentage breakdowns. The number of random draws per each percentage breakdown was 100, sufficient to provide smooth empirical a posteriori estimates of the quantities of interest. For each random draw, the SNPs were randomly pruned: for SNPs in LD, corresponding to correlation coefficient $r^2 \geq 0.8$, the SNP selected was randomly chosen—not necessarily the one with largest z-score. Using a fixed correlation coefficient allows for uniform pruning such that each LD block is treated equivalently (large block not over-represented); randomly selecting the representative SNP explicitly avoids the "winner's curse" (Zöllner and Pritchard, 2007; Ghosh et al., 2008) in estimating discovery sample z-scores. Pruning at $r^2 \geq 0.8$ reduced the number of SNPs analyzed from ~ 9.3 to ~ 2.8 million, i.e., $\sim 30\%$ of the total.

Note, each iteration of the procedure produces an unbiased estimate of the posterior effect size means and variances, conditional on the discovery z-scores. The purpose of averaging across 100 random iterations is to smooth out the random differences present in each arbitrary partition of the sample into discovery and replication samples. Since each iteration is unbiased, the average across all iterations is again unbiased for the conditional posterior means and variances.

Empirical Posterior Effect Sizes and Variances

For a given discovery-replication division of the data, z-scores from the discovery set were binned in 200 equally-spaced bins between $z_{min} = -6$ and $z_{max} = 6$. For each bin, the mean z-score for the corresponding SNPs was estimated from the replication set. For a given bin, denote the mean z-score in the discovery set as z_d, and the corresponding mean z-score in the replication set as z_r. Averaging these over the 100 repetitions (for a given percentage breakdown) provides an empirical estimate for the posterior expectation value of z_r given z_d, $E(z_r|z_d)$. Note that the empirical z_r, being a mean in the replication set, is a direct estimate of the effect size δ_r in the replication set corresponding to z_d in the discovery set: $E(z_r|z_d) = E(\delta_r|z_d)$. Similarly, empirical estimates for $E(z_r^2|z_d) = E(\delta_r^2|z_d) + \sigma_0^2$ and $\mathrm{var}(z_r|z_d) = \mathrm{var}(\delta_r|z_d) + \sigma_0^2$ were calculated.

Empirical Replication Probabilities

Given a discovery sample z-score z_d for some SNP, the effect can be deemed to replicate if the corresponding z-score z_r in the replication sample has the same sign as z_d, and its p-value from a one-tailed test, based on the standard normal cumulative distribution, is less than a chosen threshold, say $p_t = 0.05$, corresponding to $-|z_r| < z_t = -1.645$. For binned discovery sample z-scores, the empirical replication probability for a given bin is defined as the fraction of z-scores in the bin that replicate, i.e., have replication-sample p-value $p_r < p_t$. As before, averaging these over the 100 repetitions (for a given percentage breakdown) provides an empirical estimate of the replication probabilities, $R(z_d; z_t)$, which we also denote $Pr(p_r < p_t)$.

Model Posterior Effect Sizes and Variances

For a discovery sample of effective sample size N_d and a new replication sample of effective sample size N_r, and noting that effect sizes are proportional to the square root of effective sample sizes, the posterior distribution for z_r given z_d is given by a modification of Equation 5:

$$Pr(z_r|z_d) = (1 - tdr)\phi(z_r, m_u, s_u^2) + tdr \cdot \phi(z_r, m_s, s_s^2), \quad (14)$$

where

$$m_u = \sqrt{N_r/N_d} \cdot \mu_u \quad (15)$$

$$m_s = \sqrt{N_r/N_d} \cdot \mu_s \quad (16)$$

$$s_u^2 = \sigma_0^2 + (N_r/N_d) \cdot \sigma_u^2 \quad (17)$$

$$s_s^2 = \sigma_0^2 + (N_r/N_d) \cdot \sigma_s^2 \quad (18)$$

(the explicit dependence of tdr, μ_u, and μ_s on z_d and N_d, and of σ_u and σ_s on N_d, has been dropped to simplify the notation). Since $z = \delta + \epsilon$, with δ and ϵ assumed to be independent and $E(\epsilon) = 0$, the expected effect size δ_r in the replication sample, given a z-score z_d in the discovery sample, can be read off from Equation 14:

$$E(\delta_r|z_d) = \sqrt{N_r/N_d}[(1 - tdr)\mu_u + tdr \cdot \mu_s]. \quad (19)$$

Additionally, from Equation 14 and using standard properties of mixture distributions (see Supplementary Material), it also follows that

$$\mathrm{var}(\delta_r|z_d) = (N_r/N_d)[(1 - tdr)\sigma_u^2 + tdr \cdot \sigma_s^2 + tdr(1 - tdr)(\mu_s - \mu_u)^2]. \quad (20)$$

Model Replication Probabilities

The replication rate $R(z_d; z_t)$ for a given z_d is the probability, in the complementary replication sample, of getting a z-score z_r (with corresponding p-value p_r) more significant than a chosen threshold, $z_t \leq 0$ (corresponding to p-value p_t), which is simply the proportion of z-scores more significant than the threshold, which in turn is given by the cumulative distribution function (CDF) version of Equation 14:

$$R(z_d; z_t) = (1 - tdr)\Phi(z_t, -|m_u|, s_u^2) + tdr \cdot \Phi(z_t, -|m_s|, s_s^2)$$

$$\equiv Pr(p_r < p_t) \tag{21}$$

corresponding to the normal PDF with mean m and variance s^2, evaluated at z_t.

The likelihood $Pr(z|\delta)$ is simply

$$Pr(z|\delta) = \phi(z; \delta, \sigma_0^2).$$

Thus, the probability of having a z-score (with p-value p) that will pass significance threshold p_t, given the expected effect size δ, is just the CDF:

$$Pr(p < p_t, \delta) = \Phi(z_t, -|\delta|, \sigma_0^2). \tag{22}$$

Parameter Estimation

The four model parameters were estimated by minimizing a convex cost function that was composed of a sum of two terms: the weighted sum of the squares of the differences between empirical estimates and model estimates of effect sizes (expected z-scores) and of the expected z-scores-squared (minimizing only with respect to the former does not allow for sufficient precision in determining the parameters). More specifically, denoting empirical replication z-scores as z, and model (predicted) posterior expectation values as δ and η, the cost function $c(\pi_1, \sigma_0^2, \sigma_a^2, \sigma_b^2)$ is

$$c(\pi_1, \sigma_0^2, \sigma_a^2, \sigma_b^2) = \sum_{i,j,k} [w_{ijk} \cdot (\bar{z}_{ijk} - \delta_{ijk})^2 + w_{ijk} \cdot (\overline{z^2}_{ijk} - \eta_{ijk})^2]. \tag{23}$$

Here, the sum over i is over the 9 different percentage breakdowns of the dataset into complementary discovery-replication fractions; the sum over j is over the 100 repetitions for each breakdown; and the sum over k is over the 200 discovery sample z-score bins. For a given percentage breakdown i and repetition j, $\bar{z}_{ijk}$ is the mean empirical replication z-score for discovery sample SNPs in bin k (note again that the mean replication z-score for a given bin provides a direct estimate of the effect size), and $\delta_{ijk} \equiv \delta_{ijk}(N_r/N_d, z_d, H; \pi_1, \sigma_0^2, \sigma_a^2, \sigma_b^2)$ is the corresponding model prediction (only weakly dependent on repetition j through variation in heterozygosity H from repetition to repetition); $\overline{z^2}_{ijk}$ is the mean of the empirical replication z-score-squared for discovery sample SNPs in bin k, and η_{ijk} is the corresponding model prediction. For given i and j, the weighting w_{ijk} is the number of SNPs (z-scores) in the k-th bin. With i indexing the ratio N_r/N_d and k indexing z_d (and dropping the explicit dependence of δ_{ijk} and η_{ijk} on N_r/N_d, H, and the model parameters, so as to simplify the notation), $\delta_{ijk} \equiv E(z_r|z_d)$ is given by Equation 19 ($E(\epsilon) = 0$), and $\eta_{ijk} \equiv E(z_r^2|z_d) = E(\delta_r^2|z_d) + \sigma_0^2$ can be calculated from Equations 19 and 20, noting that

$$\begin{aligned} E(z_r^2|z_d) - [E(z_r|z_d)]^2 &= \mathrm{var}(z_r|z_d) \\ &= \mathrm{var}(\delta_r|z_d) + \sigma_0^2. \end{aligned} \tag{24}$$

In particular, for a given sample of effective size N, the posterior expectation of the square of the "true" effect size for that sample is

$$\begin{aligned} E(\delta^2|z) = {} & (1 - tdr)\sigma_u^2 + tdr \cdot \sigma_s^2 + \\ & tdr(1 - tdr)(\mu_s - \mu_u)^2 + \\ & [(1 - tdr)\mu_u + tdr \cdot \mu_s]^2 \end{aligned} \tag{25}$$

(note again that the explicit dependence of tdr, μ_u, and μ_s on z and N, and of σ_u and σ_s on N, has been dropped to simplify the notation). Best-fit model parameters were determined by Nelder-Mead minimization of $c(\pi_1, \sigma_0^2, \sigma_a^2, \sigma_b^2)$, using the Matlab function `fminsearch()`.

Proportion of Genetically-Determined Phenotypic Variance Explained, as a Function of *N*

For a given sample size N, the expected proportion of the total additive genetic variance (approximately the proportion of chip heritability) explained by sparse effects, $S(N; z_t)$, for SNPs with $-|z| < z_t$ for some threshold $z_t < 0$, or equivalently with p-value less than the corresponding threshold p_t, can be estimated by simulating z-scores for all SNPs, whereby the unobservable null, ubiquitous, and sparse effects can explicitly be assigned to individual SNPs. From Equation 1 and the implicit decomposition $z = \delta + \epsilon$, all of the simulation SNPs i ($i = 1, \ldots, \simeq 2.8 \times 10^6$) are assigned an environmental/error component ϵ_i drawn from a normal distribution with mean zero and variance given by the estimated σ_0^2. Then, a proportion π_1 of the SNPs (indexed by k)—conceptually, SNPs that are in strong LD with causal SNPs—are assigned an additional component $\delta_{c,k}$ drawn from a normal distribution with mean zero and variance $(\sigma_a^2 + \sigma_b^2)N\overline{H}$, where $\overline{H}$ is the mean heterozygosity, so that the corresponding z-score for such a SNP k is $z_k = \delta_{c,k} + \epsilon_k$. (Though not necessary for the calculation of S, the remaining proportion, $1 - \pi_1$, of the SNPs can be assigned an additional (ubiquitous) component $\delta_{u,j}$ drawn from a normal distribution with mean zero and variance $\sigma_a^2 N\overline{H}$, giving z-scores $z_j = \delta_{u,j} + \epsilon_j$. Note that the total effect size for a SNP in the π_1 category can be decomposed into "ubiquitous" and "sparse" components: $\delta_{c,k} = \delta_{u,k} + \delta_{s,k}$.) The proportion of chip heritability explained by SNPs with sparse effects is given by the ratio

$$S(N; z_t) = \frac{\sum_{k:-|z_k|<z_t} \delta_{c,k}^2(N)}{\sum_{all\ k} \delta_{c,k}^2(N)} \tag{26}$$

where $\delta_{c,k}^2(N)$ denotes the square of the "true" effect size component of the z-score for the kth SNP, emphasizing its dependence on N (see Supplementary Material). The numerator and denominator can be averaged over several repetitions for a smooth estimate of $S(N; z_t)$. Alternatively, given a set of z-scores $\{z_k\}$, replace $\delta_{c,k}^2$ with the expectation $E(\delta_k^2|z_k)$ (see Equation 25). The corresponding ratio in Equation 26 should be accurate if the average effect of LD cancels between the numerator and denominator, which will always occur for N large enough so that S approaches 1. In any case, the effects of LD can increasingly be mitigated by higher levels of pruning.

Multistage Design Combining Independent Discovery and Replication Datasets

It is common practice to estimate effect sizes in a multistage design where, at stage-1, candidate SNPs are selected in a discovery dataset using a liberal p-value (e.g., $p = 10^{-6}$), and then reassessed at stage-2 in an independent replication data set, with the final assessment for significance being from the combined datasets (Satagopan et al., 2004; Skol et al., 2007; Thomas et al., 2009; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011; Lambert et al., 2013; Ripke et al., 2013a). This strategy is usually employed when the independent replication dataset is not directly available to researchers, but where z-scores for candidate SNPs can be requested. The model presented here allows for predictions in this scenario. Specifically, let N_d and N_r be as before (sample sizes for independent discovery and replication datasets, respectively), and let $N_{dr} = N_d + N_r$ denote the sample size for the combined dataset. Define the inverse-variance weights $w_d = \sqrt{N_d/N_{dr}}$ and $w_r = \sqrt{N_r/N_{dr}}$. Then, given a discovery sample z-score z_d, the posterior expectation for the z-score z_{dr} in the combined dataset is

$$\begin{aligned} E(z_{dr}|z_d) &= E(\delta_{dr}|z_d) \\ &= w_d z_d + w_r E(\delta_r|z_d), \end{aligned} \tag{27}$$

where δ_{dr} is the effect size and $E(\delta_r|z_d)$ is given by Equation 19. The variance is simply

$$\mathrm{var}(z_{dr}|z_d) = w_r^2 \mathrm{var}(z_r|z_d), \tag{28}$$

where $\mathrm{var}(z_r|z_d)$ is given by Equations 20 and 24, thus allowing $E(z_{dr}^2|z_d)$ to be calculated. The replication probability $R_{dr}(z_d; z_t)$ in the combined dataset for z_d in the discovery dataset, which we can also write as $Pr(p_{dr} < p_t)$ where p_{dr} is the p-value for the SNP in the combined data set, is given by Equation 21, only replacing z_t on the right-hand side with $z_t' = (z_t + w_d|z_d|)/w_r$:

$$\begin{aligned} R_{dr}(z_d; z_t) &= (1 - tdr)\Phi(z_t', -|m_u|, s_u^2) + \\ &\quad tdr \cdot \Phi(z_t', -|m_s|, s_s^2) \\ &\equiv Pr(p_{dr} < p_t) \end{aligned} \tag{29}$$

(see Supplementary Material for further details). Thus, for a SNP whose discovery sample z-score is z_d, the probability of its z-score in the combined dataset reaching geome-wide significance is given by $R_{dr}(z_d; z_t)$, with $z_t = -5.33$ (i.e., $Pr(p_{dr} < p_t)$ with $p_t = 5 \times 10^{-8}$).

RESULTS

For a range of z-scores in the discovery sample between −6 and 6, **Figure 1** shows the empirical estimates (solid black curves) for schizophrenia of (A) expected effect sizes and (B) variances in the replication sample, and (C) the replication rate at $z_t = -1.64$ (i.e., $p_t = 0.05$), for split-half discovery and replication samples. It is significant that in (A), in a neighborhood of approximately ±3 around $z_d = 0$ the expected effect size is non-zero: the black line has a positive slope. This implies that there exists ubiquitous non-null "small" effects (small z-scores have high probability densities, i.e., the corresponding SNPs are highly abundant). As the neighborhood extends further, "large" effects found in the discovery sample correspond to "large" effects in the replication sample.

The parameter estimates (with 95% confidence intervals in square brackets) for schizophrenia were:

$$\begin{aligned} \pi_1 &= 0.037 \quad [0.017; \quad 0.079] \\ \sigma_0 &= 1.014 \quad [1.011; \quad 1.017] \\ \sigma_a &= 0.0057 \quad [0.0051; \quad 0.0063] \\ \sigma_b &= 0.020 \quad [0.015; \quad 0.025] \end{aligned} \tag{30}$$

(the procedure for calculating the standard errors is described in the Supplementary Material). The model fit with these parameters is shown as the solid red curve in **Figure 1**: there is an excellent fit to posterior effect size and variance, and to the replication rate. In contrast, if the model assumes no non-null ubiquitous effects ($\sigma_a = 0$), then without changing the other parameters the fit corresponds to the green dashed line: in the ±3 neighborhood around $z_d = 0$, small discovery effects do not replicate (i.e., they are null), while sparse "large" effects do replicate, in only approximate agreement with the empirical estimates. The alternative scenario assumes no sparse effects, $\sigma_b = 0$: without altering the other parameters, the

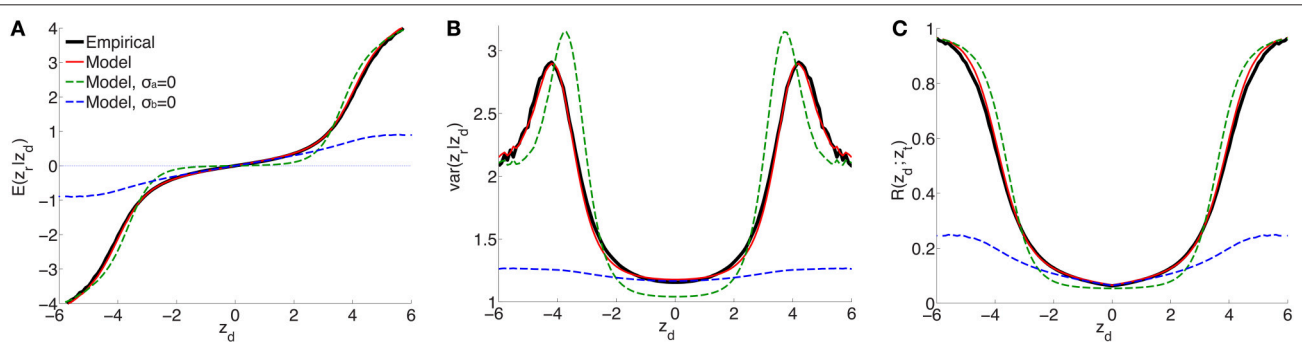

FIGURE 1 | For schizophrenia, posterior estimate of (A) effect size and (B) variance; (C) estimate of replication probability for $z_t = -1.64$ (i.e., $p_t = 0.05$): empirical (black solid lines), current model (red solid lines), model with no ubiquitous effects (green dashed lines), and model with no sparse effects (blue dashed lines), for split-half discovery and replication data.

model fit is shown by the blue dashed curves in **Figure 1**. In this case, the model provides a reasonable match with the empirical small effects. However, it completely fails to capture the posterior empirical effects for large z_d. Corresponding curves for explicit model fits with parameters $(\pi_1, \sigma_0^2, \sigma_b^2)$ and (σ_0^2, σ_a^2), respectively, are shown in Supplementary Material Figure S1 (note that in the case of $\sigma_b = 0$ the model is degenerate with respect to π_1).

For putamen volume, the parameter estimates (with 95% confidence intervals) were:

$$\begin{aligned} \pi_1 &= 0.0010 \quad [0.0001; \quad 0.0082] \\ \sigma_0 &= 1.002 \quad [1.001; \quad 1.003] \\ \sigma_a &= 0.0033 \quad [0.0029; \quad 0.0039] \\ \sigma_b &= 0.034 \quad [0.019; \quad 0.062]. \end{aligned} \tag{31}$$

Thus, putamen volume is approximately 40-times less polygenic than schizophrenia.

Discovery and replication sample sizes have a pronounced influence on the empirical estimates of effect size, variance, and replication rate, but these dramatic changes across a wide range of z-scores are remarkably well matched by the model estimates (Supplementary Material Figure S3). Additionally, the empirical PDF for the z-scores is accurately reproduced by the model fit, Equation 1, for z-scores divided up into five heterozygosity windows (Supplementary Material Figure S2), validating the basic model definition, Equation 1.

Figure 2 shows the posterior effect size components δ_u (green curve) and δ_s (red curve), given by Equations 12 and 13, for z-scores between -6 and 6, for an effective sample size $N \simeq 34,000$. Also shown is the total effect size (black) and the posterior variance of the effect size (blue curve). The individual components can be seen to behave as expected: the green curve shows the ubiquitous non-null effects increasing away from the origin, reaching a peak, and falling to zero for large values of z; the red curve shows the sparse effects component essentially flat near the origin, indicating a lack of sparse effects in this neighborhood, and then monotonically increasing, beginning near where the ubiquitous effects start falling to zero. The peaks in the variance can be seen to arise from the regions where sparse effects are already prominent but the small effects have not yet died off. (Note that the variance is drawn on the same scale as the effect sizes.)

To test the extent to which the small ubiquitous effects arose from LD with sparse large effects, in addition to light random pruning as before at $r^2 \geq 0.8$, we further restricted to SNPs whose total linkage disequilibrium (TLD), given by the sum of their LD r^2's with neighboring SNPs, was less than 15 (for reference, the median TLD was 54.5), reducing the number of SNPs to $\simeq$ 1 million (10% of the total). The empirical effect sizes, for a sample breakdown of 50% discovery and 50% replication, are shown in **Figure 3** (the red plot). For comparison, also shown (in black) is the empirical effect size plot for random pruning at $r^2 \geq 0.8$ without restricting by TLD (same as in **Figure 1A**). There is a pronounced diminution in the extent of ubiquitous (small) effects: the decreased slope near the origin suggests that a substantial portion of the ubiquitous effects arises due to LD with large effects, particularly large LD blocks (Yang et al., 2011b): if the distance between causal SNPs is comparable or smaller than typical LD block sizes, then "ubiquitous" effects are expected. Note that the shift from the black to the red curve in **Figure 3** is consonant with the shift from the black (ubiquitous and sparse effects) to the red (sparse effects only) curve in **Figure 2**.

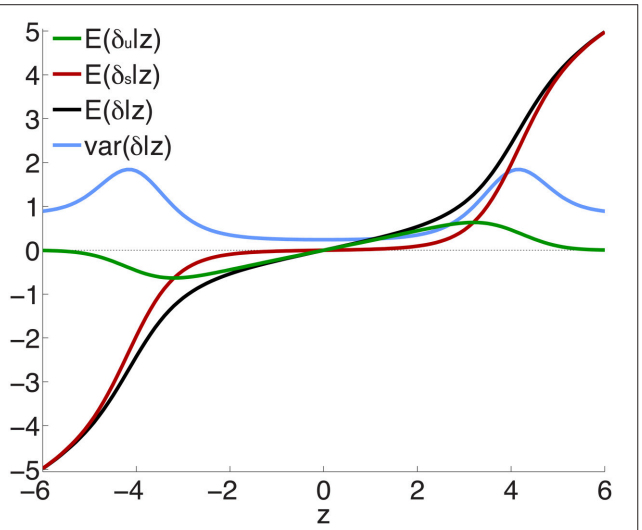

FIGURE 2 | Posterior effect size and variance, calculated for effective sample size $N_d = N_r = N \simeq 34,000$—see Equations 12, 13, 20 and Supplementary Material. Note that sparse effects have a component that arises from ubiquitous effects. $z = \delta + \epsilon$, where $\delta = \delta_U + \delta_S$ and $E(\epsilon) = 0$; δ_U are ubiquitous effects, while δ_S are additional contributions to total sparse effects.

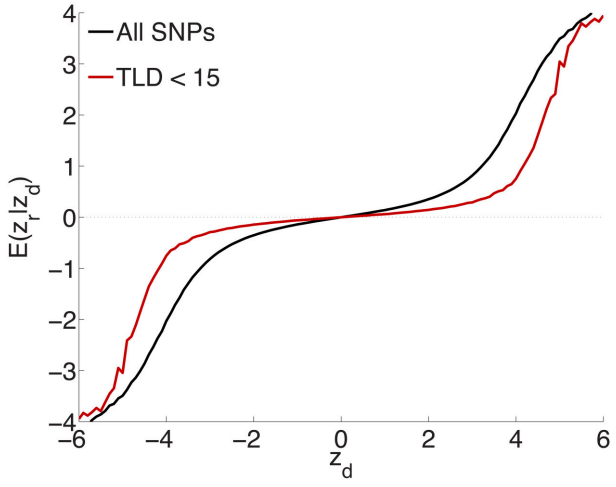

FIGURE 3 | Randomly culling SNPs with LD $r^2 \geq 0.8$ and further restricting to SNPs with total LD (TLD) less than 15 (approximately 1 million SNPs remaining) shows a diminution in the extent of ubiquitous effects (decreased slope near the origin for the red curve), consistent with an interpretation that the ubiquitous effects arise due to LD with causal SNPs. The black plot is for light random pruning at $r^2 \geq 0.8$, shown in **Figure 1A**.

Figure 4A shows quantile-quantile (QQ) plots for SNP p-values, comparing empirical and model fits for the full schizophrenia and putamen data sets, averaged over 100 repetitions with random pruning. For any threshold p-value, given as the ordinate in $\log_{10}$ units (to emphasize tail, or small p, behavior), the abscissa gives the proportion of SNPs whose p-values are at least as significant as the ordinate value; the dashed line at 45° corresponds to the null hypothesis where the distribution of SNP z-scores is assumed to follow a standard normal distribution (the threshold p-value then being synonymous with the proportion of SNPs exceeding that value). The model plots are given by Equation 1, replacing the PDF with the CDF, and provide a remarkably good fit to the empirical plots. The earlier deviation of the schizophrenia plot, compared with the putamen plot, from the null line is due to the higher polygenicity of schizophrenia.

Figure 4B shows the projected proportion of tagged variance explained by sparse SNPs reaching genome-wide significance ($p \leq 5 \times 10^{-8}$) for schizophrenia and putamen volume, as a function of the total number N of subjects in the samples (assuming equal numbers of cases and controls for schizophrenia), as given by Equation 26. The fraction of tagged variance explained by GWAS is expected to approximately equal the fraction of additive genetically-determined phenotypic variance, or narrow-sense heritability, explained. The blue asterisk indicates the sample size from current ENIGMA data ($N = 12,596$); the green asterisk, for schizophrenia, gives $N = 76,326$, assuming the effective sample size from the current PGC2, $N_{eff} = 38,163$, arose from an equal number of cases and controls: $N/2 = N_{cases} = N_{controls} = N_{eff}$. Thus, we estimate that 15% of chip heritability for schizophrenia is currently explainable by genome-wide significant SNPs in PGC2; for these SNPs, the replication rate at $p_t = 0.05$ for our split-half sample is 97% or higher. GWAS on approximately half a million each of cases and controls would need to be performed to explain all the chip heritability for schizophrenia. For putamen volume, 14% of chip heritability appears to be explainable by genome-wide significant SNPs given the current sample size in ENIGMA. In contrast to schizophrenia, however, approximately only 100,000 people would need to be assessed to fully explain chip heritability for putamen volume. The higher sample size requirements for schizophrenia are due partly to its higher polygenicity.

The per-allele contribution of a locus, with z-score z, to the phenotypic variance v_a of the trait is usually estimated to be $v_a = \hat{\beta}^2 H$, where $\hat{\beta}$ is the corresponding regression coefficient in the univariate setting (Park et al., 2010). This is proportional to z^2/N, since $z = \hat{\beta}/se(\hat{\beta})$. However, the "true" effect more correctly is related to the non-centrality parameter δ ($z = \delta + \epsilon$), so that the per-allele contribution to phenotypic variance is proportional to $E(\delta^2|z)/N$, not z^2/N. **Figure 5A** plots $E(\delta^2|z)$ vs. z^2 (or scaled $\hat{\beta}^2$) for three illustrative total sample sizes. For an independent sample, the degree by which v_a is an overestimate when based on z^2 instead of $E(\delta^2|z)$ is given by the ratio of the height on the black dotted line to the height on the appropriate curve, at z^2. For example, for schizophrenia with a total sample size of N=50,000 and a z-score on the threshold of genome-wide significance ($z \simeq \pm 5.33$), v_a will be over-estimated by a factor of 2.2 (the height of the black dot relative to the blue dot).

Figure 5B shows the probability, given by Equation 29, of reaching genome-wide significance in a combined discovery and replication dataset of total sample size corresponding to PGC2 ($N = 76,326$), where the discovery sample size N_d is 20, 50, or 90% of the total, as a function of p-value in the discovery dataset. For example, in discovery samples equal to 50 or 90% of the total, to have a probability of reaching genome-wide significance in the combined dataset approximately equal to 0.8 would require having reached genome-wide significance in the discovery dataset; for a discovery sample only 20% of the total, the same probability of replicating requires the discovery sample p-value to be at least $p = 10^{-9}$, i.e., *more* significant than the genome-wide threshold.

DISCUSSION

Here we present a simple modeling framework, a scale-mixture of Gaussians that is a modification of previously published

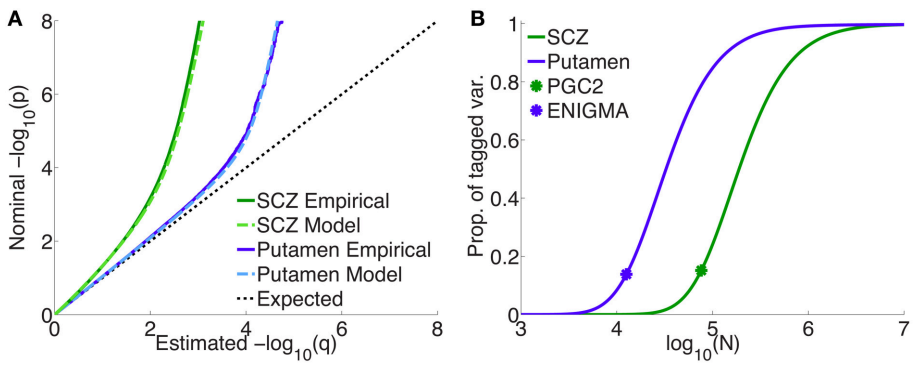

FIGURE 4 | (A) Empirical and model QQ plots for putamen volume and schizophrenia. **(B)** Proportion of total additive genetic variance or chip heritability explained by sparse effects for all "tagged" SNPs with p-value less than the GWAS p-value threshold ($p_t = 5 \times 10^{-8}$), as a function of effective sample size, for putamen volume and schizophrenia (the asterisks correspond to the current effective sample sizes for ENIGMA and PGS2). Of the total variance that is explained by sparse effects for all SNPs, the proportion explained by SNPs currently reaching the usual GWAS significance level is approximately 15% for both phenotypes.

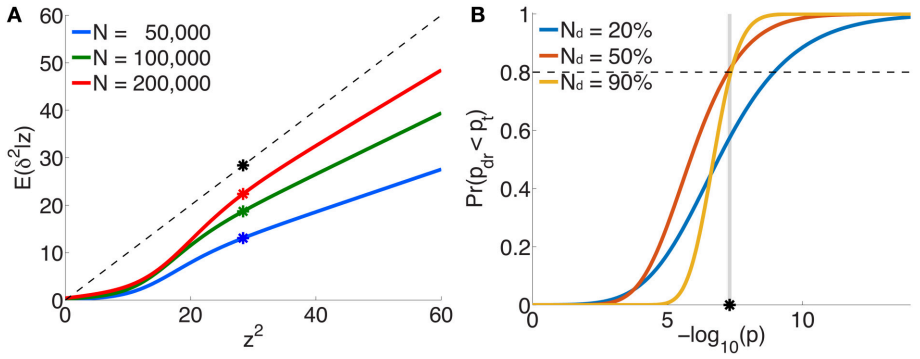

FIGURE 5 | For schizophrenia, (A) posterior estimates of effect-size-squared, as given by Equation 25, vs. z^2 for three total sample sizes. When assuming that the phenotypic variance explained by a SNP is given by $\hat{\beta}^2 H \propto z^2/N$, the degree to which this is an over-estimate is indicated by the ratio of the height of the black dashed line (the assumption $\delta^2 = z^2$) to the height of the corresponding point on the curve for a given sample size. The asterisks correspond to the threshold significant z-score. **(B)** For a multistage GWAS, where discovery is from a subset (20%, 50%, 90%) of the total PGC2 sample, the curves give the probability of a SNP with p-value p in the discovery sample passing genome-wide significance ($p_{dr} < p_t = 5 \times 10^{-8}$) in the combined (total) data set, Equation 29. The vertical gray line is at $p = p_t$.

methodologies (Meuwissen et al., 2001; Goddard et al., 2009; Erbe et al., 2012; Efron, 2013; Zhou et al., 2013), for assessing the contributions of ubiquitous and sparse effects to quantities of interest in GWAS. Additionally, we present a procedure for testing the model empirically. The model has great utility in that it allows for the prediction of the replication probability and expected effect size for each SNP, given the discovery and replication sample sizes, the four model parameters, and the discovery sample z-score of the SNP. Using nonparametric methods applied to large schizophrenia and subcortical brain structure volumes, we show how the empirical replication probabilities and effect sizes can be calculated directly from the data, and demonstrate that the model results are in excellent agreement with these—across a wide range of z-scores and sample sizes, and for multistage GWAS designs, whether splitting the available sub-studies into randomized disjoint sets for discovery and replication, or combining the discovery dataset as a subset in replication.

The parameter π_1 in the model gives the prior probability that SNPs belong in the category of sparse effects, i.e., those likely to have a large and significant association with the phenotype. π_1 thus provides a way of estimation the proportion of "causal" SNPs. An alternative method for doing this is Approximate Bayesian Polygenic Analysis (ABPA; Stahl et al., 2012). With 1000 Genomes SNP imputation for a total of approximately 9.3 million SNPs in the studies, all possible causal common SNPs are likely to be represented, either directly or through strong LD. These SNPs will have the largest expected association summary statistics; the remaining SNPs will either show weaker effects through attenuated LD with the causal SNPs, or have null effects. To a first approximation, a plausible interpretation is that SNPs in the π_1 category are likely to be dominated by SNPs in strong LD with causal SNPs. Our estimate of $\pi_1 \simeq 0.037$ for the PGC2 schizophrenia GWAS suggests that this condition is highly polygenic: about 3.7% of SNPs are potentially significantly associated with the phenotype. It should be noted, however, that variations at different loci exhibiting LD may have independent effects on a phenotype (Malo et al., 2008).

Recently, a schizophrenia GWAS study, using approximately half of the subjects and all of the SNPs employed here, reported an estimate of "SNP heritability" (a lower bound of narrow sense heritability since only variation due to SNP association can be determined; copy-number variants and rare variants, for example, are not included) $h^2 = 33\%$ on the liability scale, adjusted for case-control ascertainment (Lee et al., 2012b; Ripke et al., 2013a). It should be noted that an implicit assumption in the method used for estimating h^2 is that the distribution of effect sizes is given by a single Gaussian (no explicit sparse-effects component—see **Figure 1A**: the blue dashed curve for $\sigma_b = 0$), whereas we have shown here that it is more appropriate to consider the effect sizes being described by a Gaussian mixture distribution, with non-null ubiquitous and sparse components. It is not clear how this affects the result for h^2. A follow-up study, using all of the subjects and SNPs employed here, reported that 3.4% of variation on the liability scale to schizophrenia, adjusted for case-control ascertainment, is explained by genome-wide significant loci (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Therefore, approximately 10% (3.4/33) of SNP heritability (or chip heritability) for schizophrenia was estimated to be captured by genome-wide significant SNPs, a result that comports with our estimate of 15%. For the putamen, a recent report (Hibar et al., 2015) using all of the subjects and SNPs employed here, and a different method for estimating SNP heritability than was used for schizophrenia (So et al., 2011), found that approximately 10% of variance in putamen volume was due to all common variants, while 1.09% was attributable to genome-wide significant SNPs. Thus, again, approximately 10% of SNP heritability for putamen volume was estimated to be captured by genome-wide significant SNPs, a result in broad agreement with our estimate of 14%.

We presented a four-parameter two-groups mixture of normals parametric model for the distribution of GWAS

summary statistics. Additionally, we presented an empirical scheme for estimating replication z-scores, their variances, and replication rates. The model parameters are then estimated by minimizing a cost function that depends on the differences in model and empirical estimates of effect sizes and expected z-scores-squared. Applying the model and empirical scheme to recent large GWAS of schizophrenia and putamen volume, we showed that effect sizes, along with variances and replication rates, are accurately described by the simple model over a wide range of z-scores, and over a wide range of discovery and replication sample sizes when the relationships can change dramatically. We further showed how the model can be used to estimate the fraction of additive SNP heritability (an estimate of the fraction of chip heritability) explainable by genome-wide significant SNPs in a massive univariate setting, as a function of sample size. The model enables estimation of an index for the degree of polygenicity of a phenotype, because structurally it allows for separate distributions of the large number of ubiquitous (small) effects and the relatively small number of sparse (large) effects, postulating that these different classes of effects will be distributed differently. We showed that ignoring the contribution of ubiquitous non-null effects severely degrades the accuracy of the model. A variant of the model, particularly for phenotypes likely to be of low polygenicity, e.g., individual gene expression or methylation, where interpretation of the ubiquitous component as currently presented might become problematic, could incorporate a third gaussian for ubiquitous null effects, with variance parameter σ_0^2. In the current study we note that the four-parameter model is more parsimonious and fits the data well. In future work, we will explore more deeply the relationship between true effects and total LD that will enable us more precisely to pin down the interpretation of the parameters.

Given the model parameters, for any SNP with a *p*-value less than some threshold—indicating a potential candidate for true association with the phenotype—an accurate estimate of its probability of replicating can be calculated for replication samples of different effective numbers of subjects. It should be noted, however, that to have reasonable probability of replicating in a multistage GWAS, it might be necessary that candidate SNPs in the discovery sample substantially *exceed* genome-wide significance. The model can be improved and extended by letting the model coefficients depend on SNP heterozygosity, and by incorporating additional information like SNP functional annotation category (Schork et al., 2013). We showed that schizophrenia is highly polygenic (∼3.7% of SNPs are significantly associated). We estimate that at approximately half a million each of cases and controls genotyped, all SNPs contributing to narrow-sense heritability would have reached genome-wide significance. Putamen volume is approximately 40-times less polygenic than schizophrenia, and only of order one hundred thousand people need to be genotyped to capture all chip heritability with genome-wide significant SNPs.

Deep sequencing of large samples can be expected to advance our understanding of the genetics of complex phenotypes, but it seems more cost-efficient to start with uncovering more risk genes from existing GWAS data by improving analytical tools, as much remains to be explored in the complex landscape where large numbers of SNPs in GWAS effect phenotype. In particular, there is a need for better understanding the distribution of effect sizes in GWAS (Andreassen et al., 2013a,b; Schork et al., 2013). The univariate mixture model we have presented and validated empirically here captures the distribution of SNP effect sizes, and thus much of the genetic architecture of complex phenotypes. It can be used for estimating polygenicity and informing power calculations, in particular for providing estimates of probabilities of reaching genome-wide significance in multistage GWAS, and the proportion of SNP heritability explainable in future larger studies. The model can be extended by incorporating additional prior information, such as SNP annotation and pleiotropy, to enable more accurate estimation of replication probabilities (Lewinger et al., 2007; Wang et al., 2016). Combined with larger sample sizes, this statistical methodology may facilitate improved detection of smaller effect sizes and enhance the ability of GWAS in accurate risk prediction, and to inform human physiology and disease etiology.

AUTHOR CONTRIBUTIONS

DH: wrote manuscript; performed analyses; contributed to study design, interpretation of results, and critical revision of manuscript. YW, CC, ML, AS: Contributed to data preparation, interpretation of results, and critical revision of manuscript. WT: Contributed to study design and interpretation of results. AW: Contributed to interpretation of results. MD, TW: Contributed to study development. OA: Contributed to study design, interpretation of results, and critical revision of manuscript. AD: Contributed to study design, data preparation, analysis, interpretation of results, and critical revision of manuscript.

FUNDING

The study was financially supported by NIH (R01MH100351), NIH GM104400-01A, the Research Council of Norway (#213837, #223273), the South-East Norway Regional Health Authority (#2013-123) and KG Jebsen Foundation.

ACKNOWLEDGMENTS

The authors wish to thank the participants of the studies for their contribution, as well as all researchers in the Schizophrenia Working Group of the Psychiatric Genomics Consortium and in the Enhancing Neuro Imaging Genetics through Meta Analysis Consortium who contributed with GWAS data. A full list of these researchers and their affiliations is provided in the Supplementary Material.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00015

REFERENCES

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

Andreassen, O. A., Djurovic, S., Thompson, W. K., Schork, A. J., Kendler, K. S., O'Donovan, M. C., et al. (2013a). Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am. J. Hum. Genet.* 92, 197–209. doi: 10.1016/j.ajhg.2013.01.001

Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., et al. (2013b). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet.* 9:e1003455. doi: 10.1371/journal.pgen.1003455

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., et al. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi: 10.1038/ng.3211

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* 45, 400–405. doi: 10.1038/ng.2579

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55, 997–1004. doi: 10.1111/j.0006-341X.1999.00997.x

Dudbridge, F., and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* 32, 227–234. doi: 10.1002/gepi.20297

Efron, B. (2013). *Large-scale Inference : Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge; New York: Cambridge University Press.

Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., et al. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95, 4114–4129. doi: 10.3168/jds.2011-5019

Fuller, W. A. (2009). *Measurement Error Models*, Vol. 305. New York, NY: John Wiley & Sons.

Ghosh, A., Zou, F., and Wright, F. A. (2008). Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am. J. Hum. Genet.* 82, 1064–1074. doi: 10.1016/j.ajhg.2008.03.002

Goddard, M. E., Wray, N. R., Verbyla, K., and Visscher, P. M. (2009). Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* 24, 517–529. doi: 10.1214/09-STS306

Greenland, S., and Poole, C. (2013). Living with p values: resurrecting a bayesian perspective on frequentist statistics. *Epidemiology* 24, 62–68. doi: 10.1097/EDE.0b013e3182785741

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivières, S., Jahanshad, N., et al. (2015). Common genetic variants influence human subcortical brain structures. *Nature*. 520, 224–229. doi: 10.1038/nature14101

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nat. Genet.* 45, 1452–1458. doi: 10.1038/ng.2802

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., et al. (2012a). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* 44, 247–250. doi: 10.1038/ng.1108

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2012b). A better coefficient of determination for genetic profile analysis. *Genet. Epidemiol.* 36, 214–224. doi: 10.1002/gepi.21614

Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* 88, 294–305. doi: 10.1016/j.ajhg.2011.02.002

Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J., and Thomas, D. C. (2007). Hierarchical bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidemiol.* 31, 871–882. doi: 10.1002/gepi.20248

Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., et al. (2009). Common genetic influences for schizophrenia and bipolar disorder: a population-based study of 2 million nuclear families. *Lancet* 373, 234–239. doi: 10.1016/S0140-6736(09)60072-6

Malo, N., Libiger, O., and Schork, N. J. (2008). Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am. J. Hum. Genet.* 82, 375–385. doi: 10.1016/j.ajhg.2007.10.012

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Palla, L., and 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. *Am. J. Hum. Genet.* 97, 250–259. doi: 10.1016/j.ajhg.2015.06.005

Park, J. H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., et al. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci. U.S.A.* 108, 18026–18031. doi: 10.1073/pnas.1114759108

Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., et al. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* 42, 570–575. doi: 10.1038/ng.610

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* 32, 381–385. doi: 10.1002/gepi.20303

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752. doi: 10.1038/nature08185

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., et al. (2013a). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* 45, 1150–1159. doi: 10.1038/ng.2742

Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., Breen, G., et al. (2013b). A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* 18, 497–511. doi: 10.1038/mp.2012.21

Satagopan, J. M., Venkatraman, E., and Begg, C. B. (2004). Two-stage designs for gene–disease association studies with sample size constraints. *Biometrics* 60, 589–597. doi: 10.1111/j.0006-341X.2004.00207.x

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* 43, 969–976. doi: 10.1038/ng.940

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427. doi: 10.1038/nature13595

Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., et al. (2013). All snps are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps. *PLoS Genet.* 9:e1003449. doi: 10.1371/journal.pgen.1003449

Schork, N. J. (2002). Power calculations for genetic association studies using estimated probability distributions. *Am. J. Hum. Genet.* 70, 1480–1489. doi: 10.1086/340788

Sklar, P., Ripke, S., Scott, L. J., Andreassen, O. A., Cichon, S., Craddock, N., et al. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near odz4. *Nat. Genet.* 43, 977. doi: 10.1038/ng.943

Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2007). Optimal designs for two-stage genome-wide association studies. *Genet. Epidemiol.* 31, 776–788. doi: 10.1002/gepi.20240

So, H.-C., Li, M., and Sham, P. C. (2011). Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet. Epidemiol.* 35, 447–456. doi: 10.1002/gepi.20593

So, H.-C., Yip, B., and Sham, P. C. (2010). Estimating the total number of susceptibility variants underlying complex diseases from genome-wide association studies. *PLoS ONE* 5:e13898. doi: 10.1371/journal.pone.0013898

Speed, D., and Balding, D. J. (2014). Multiblup: improved snp-based prediction for complex traits. *Genome Res.* 24, 1550–1557. doi: 10.1101/gr.169375.113

Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., et al. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* 44, 483–489. doi: 10.1038/ng.2232

Sullivan, P. F. (2010). The psychiatric GWAS consortium: big science comes to psychiatry. *Neuron* 68, 182–186. doi: 10.1016/j.neuron.2010.10.003

Sullivan, P. F., Kendler, K. S., and Neale, M. C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* 60, 1187–1192. doi: 10.1001/archpsyc.60.12.1187

Tenesa, A., and Haley, C. S. (2013). The heritability of human disease: estimation, uses and abuses. *Nat. Rev. Genet.* 14, 139–149. doi: 10.1038/nrg3377

Thomas, D. C., Casey, G., Conti, D. V., Haile, R. W., Lewinger, J. P., and Stram, D. O. (2009). Methodological issues in multistage genome-wide association studies. *Stat. Sci.* 24, 414. doi: 10.1214/09-STS288

Thompson, W. K., Wang, Y., Schork, A. J., Witoelar, A., Zuber, V., Xu, S., et al. (2015). An empirical bayes mixture model for effect size distributions in genome-wide association studies. *PLoS Genet.* 11:e1005717. doi: 10.1371/journal.pgen.1005717

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of gwas discovery. *Am. J. Hum. Genet.* 90, 7–24. doi: 10.1016/j.ajhg.2011.11.029

Wang, Y., Thompson, W. K., Schork, A. J., Holland, D., Chen, C.-H., Bettella, F., et al. (2016). Leveraging genomic annotations and pleiotropic enrichment for improved replication rates in schizophrenia GWAS. *PLoS Genet.* 12:e1005803. doi: 10.1371/journal.pgen.1005803

Witte, J. S., Visscher, P. M., and Wray, N. R. (2014). The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* 15, 765–776. doi: 10.1038/nrg3786

Wray, N. R., and Gottesman, I. I. (2012). Using summary data from the danish national registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder. *Front. Genet.* 3:118. doi: 10.3389/fgene.2012.00118

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011a). Gcta: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., et al. (2011b). Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* 19, 807–812. doi: 10.1038/ejhg.2011.39

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* 9:e1003264. doi: 10.1371/journal.pgen.1003264

Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* 11, 407–409. doi: 10.1038/nmeth.2848

Zöllner, S., and Pritchard, J. K. (2007). Overcoming the winnerž019s curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* 80, 605–615. doi: 10.1086/512821

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.