HHS Public Access
Author manuscript
Nat Genet. Author manuscript; available in PMC 2015 September 01.

Published in final edited form as:
Nat Genet. 2015 March ; 47(3): 291–295. doi:10.1038/ng.3211.

LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies

Brendan K. Bulik-Sullivan[1,2,3], **Po-Ru Loh**[4,5], **Hilary Finucane**[6], **Stephan Ripke**[2,3], **Jian Yang**[7,8], **Schizophrenia Working Group of the Psychiatric Genomics Consortium**[9], **Nick Patterson**[1], **Mark J. Daly**[1,2,3], **Alkes L. Price**[1,4,5], and **Benjamin M. Neale**[1,2,3,*]

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA

[2]Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA

[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA

[4]Department of Epidemiology, Harvard School of Public Health, Boston, MA

[5]Department of Biostatistics, Harvard School of Public Health, Boston, MA

[6]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA

[7]Brain Institute, University of Queensland, Brisbane, Queensland, Australia

[8]University of Queensland Diamantina Institute, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland, Australia

Abstract

Both polygenicity (*i.e.*, many small genetic effects) and confounding biases, such as cryptic relatedness and population stratification, can yield an inflated distribution of test statistics in genome-wide association studies (GWAS). However, current methods cannot distinguish between inflation from true polygenic signal and bias. We have developed an approach, LD Score regression, that quantifies the contribution of each by examining the relationship between test statistics and linkage disequilibrium (LD). The LD Score regression intercept can be used to estimate a more powerful and accurate correction factor than genomic control. We find strong evidence that polygenicity accounts for the majority of test statistic inflation in many GWAS of large sample size.

---

[*]To whom correspondence ought to be addressed: bneale@broadinstitute.org.
[9]A list of members and affiliations appears in the Supplementary Note

AUTHOR CONTRIBUTIONS

BBS conceived of the idea, analyzed the data, performed the analyses and drafted the manuscript. BMN conceived of the idea and drafted the manuscript. MJD conceived of the idea and supplied reagents. NP conceived of the idea and supplied reagents. ALP conceived of the idea and supplied reagents. PRL analyzed the data and performed the analyses. HF analyzed the data and performed the analyses. JY provided software. All authors provided input and revisions for the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

INTRODUCTION

Variants in LD with a causal variant show elevated test statistics in association analysis proportional to the LD (measured by r^2) with the causal variant[1–3]. The more genetic variation an index variant tags, the higher the probability that this index variant will tag a causal variant. In contrast, inflation from cryptic relatedness within or between cohorts[4,5,6] or population stratification purely from genetic drift will not correlate with LD Score.

Under a polygenic model, such that effect sizes are drawn independently from distributions with variance proportional to $p(1\text{-}p)^{-1/2}$ where p is minor allele frequency (MAF), then the expected χ^2-statistic of variant j is

$$E[\chi^2|\ell_j]=Nh^2\ell_j/M+Na+1, \quad (1)$$

where N is sample size; M is the number of SNPs, such that h^2/M is the average heritability explained per SNP; a measures the contribution of confounding biases, such as cryptic relatedness and population stratification; and $\ell_j := \Sigma_k r_{jk}^2$ is the *LD Score* of variant j, which measures the amount of genetic variation tagged by j and (a full derivation of this equation is provided in the Supplementary Note). This relationship holds for meta-analyses, and also for ascertained studies of binary phenotypes, in which case h^2 is on the observed scale. Consequently, if we regress χ^2statistics from GWAS against LD Score (LD Score regression), the intercept minus one is an estimator of the mean contribution of confounding bias to the inflation in the test statistics.

RESULTS

Overview of Methods

We estimated LD Scores from the European ancestry samples in the 1000 Genomes Project[7] (EUR) using an unbiased estimator[8] of r^2 with 1 centiMorgan (cM) windows, singletons excluded (MAF > 0.13%) and no r^2 cutoff. Standard errors were estimated by jackknifing over blocks of individuals, and we used these standard errors to correct for attenuation bias in LD Score regression (*i.e.*, the downward bias in the magnitude of the regression slope that results when the regressor is measured noisily, see Online Methods).

For LD Score regression, we excluded variants with EUR MAF < 1% because the LD Score standard errors for these variants were very high (note: variants included in LD Score regression are a subset of variants included in LD Score estimation). In addition, we excluded loci with extremely large effect sizes or extensive long-range LD from all regressions, because these can be considered outliers in such an analysis and would have disproportionate influence on the regression (Online Methods).

An important consideration in the estimation of LD Score is the extent to which the sample from which we estimate LD Score matches the sample for the association study. If there is mismatch between LD Scores from the reference population and the target population used for GWAS, then LD Score regression can be biased in two ways. First, if LD Scores in the reference population are equal to LD Scores in the target population plus mean-zero noise, then the intercept will be biased upwards and the slope downwards. This is conceptually equivalent to increasing the measurement error of LD Score. Secondly and perhaps more importantly, consider the scenario where there is a directional bias in average LD Score such that the LD Scores in the reference population are systematically higher or lower than in the target population. Under such a scenario, then the LD Score regression intercept will be biased downwards or upwards, respectively (Online Methods).

To explore the stability of LD Score across European populations, we estimated LD Scores using each of the 1000 Genomes EUR subpopulations separately (Utah Residents with Northern and Western European Ancestry (CEU), British in England and Scotland (GBR), Toscani in Italia (TSI) and Finnish in Finland (FIN)). The LD Scores from all four subpopulations were highly correlated, but mean LD Score increased with latitude (Supplementary Table 8), consistent with the observation that Southern European populations have gone through less severe bottlenecks than Northern European populations[9]. For example, in comparison to the combined EUR LD Score, the mean LD Score for FIN was 7% larger, and the mean LD Score for TSI was 8% smaller. We evaluated the impact of these differences on the behavior of the LD Score regression analysis and find that the EUR reference panel is adequate for studies in outbred populations of predominantly northern European ancestry, such as European American or UK populations (see Online methods). For other populations, a different reference panel should be used.

Under strong assumptions about the effect sizes of rare variants, the slope of the LD Score regression can be re-scaled to be an estimate of the heritability explained by all SNPs used in the estimation of the LD Scores (Supplementary Table 1). Relaxing these assumptions in order to obtain a robust estimate of the heritability explained by all 1000 Genomes SNPs is a direction for further research; however, we note that the LD Score regression intercept is robust to these assumptions.

Simulations with Polygenic Genetic Architectures

To verify the relationship between linkage disequilibrium and χ^2statistics, we performed a variety of simulations to model scenarios with population stratification, cryptic relatedness and polygenic architecture.

To model a polygenic quantitative trait, we assigned per-allele effect sizes drawn from $N(0, h^2/(2p(1-p))^{-1/2}/M)$ to varying numbers of causal variants and for varying heritabilities in an approximately unstructured cohort of 1000 Swedes. In all simulation settings, the average LD Score regression intercept was close to one. We note that if there are few causal variants, the LD Score regression estimates are still unbiased, but the standard errors become very large, meaning that this approach is best suited to polygenic traits (Supplementary Figures 3–5).

Simulations with Confounding

The model assumes that that there is no systematic correlation between FST and LD Score (see Supplementary Note). This assumption may be violated in practice as a result of linked selection (*i.e.*, positive selection[10] and background selection[11]). If there were a positive correlation between LD Score and F_{ST} , the LD Score regression intercept would underestimate the contribution of population stratification to the inflation in χ^2statistics. To quantify the bias that this might introduce into the LD Score regression intercept, we performed a series of simulations with real population stratification.

We obtained un-imputed genotypes from Psychiatric Genomics Consortium (PGC) controls from seven European cohorts genotyped on the same array (Supplementary Table 2). To simulate population stratification on a continental scale, we assigned case/control status based on cohort membership, then computed association statistics for each pair of cohorts (note that in this simulation setup the expected mean χ^2-statistic is $1+bNF_{ST}$, where b is the correlation between phenotype and ancestry and N is sample size, ref [12]). To simulate population stratification on a national scale, we computed the top three principal components within each cohort, then computed association statistics using each of these principal components as phenotypes. Quantile-quantile (QQ) plots from simulations with population stratification and polygenicity show indistinguishable patterns of inflation (Fig. 1a,b), but the average LD Score regression intercept was approximately equal to λ_{GC} in simulations with population stratification (see Supplementary Table 3a for simulations with continental-scale stratification and Supplementary Table 4a for simulations with national-scale stratification), and near 1 in simulations with polygenicity (Supplementary Figures 1–5). Furthermore the qualitative appearance of the pattern of inflation as a function of LD Score was completely different in each set of simulations (Fig 1c,d). The observed correlations between F_{ST} and LD Score in all simulations were negligible (generally 10^{-5} to 10^{-4}, see Supplementary Tables 3b and 4b). We note that in simulations with population stratification, the LD Score regression slope was slightly greater than zero on average (Supplementary Tables 3c, 4c), likely a result of linked selection. Nevertheless, the performance of the LD Score regression intercept was comparable to λ_{GC}, and so would be suitably conservative if used as a correction factor, despite the small bias in the slope.

Simulations with Confounding and Polygenicity

To simulate a more realistic scenario where both polygenicity and bias contribute simultaneously to test statistic inflation, we obtained genotypes from approximately 22,000 individuals from throughout Europe from the Wellcome Trust Case-Control Consortium 2[13]. We simulated polygenic phenotypes with causal SNPs drawn from the first halves of chromosomes, leaving all SNPs on the second halves of chromosomes null. In addition, we included an environmental stratification component aligned with the first principal component of the genotype data, representing Northern vs. Southern European ancestry. In this setup, the mean χ^2 among SNPs on the second halves of chromosomes measures the average contribution of stratification. We performed similar simulations with cryptic relatedness using data from the Framingham Heart Study[14], which includes close relatives. In all simulation replicates, the LD Score regression intercept was approximately equal to the mean χ^2 among null SNPs (Supplementary Table 5), which demonstrates that LD Score regression can partition the inflation in test statistics even in the presence of both bias and polygenicity.

Finally, we modeled studies of a polygenic binary phenotype with case-control ascertainment using a simulated genotypes and a liability threshold model, and verified that LD Score regression is not noticeably biased by case-control ascertainment (Supplementary Table 6).

Frequency-Dependent Genetic Architectures

LD Score regression works optimally when variance explained per SNP is uncorrelated with LD Score (this means that rare variants have larger effect sizes than common variants, which may be appropriate for a disease phenotype under moderate negative selection). A potential limitation of LD Score regression is that variance explained per SNP may be correlated with LD Score for some phenotypes. For an example where this might occur, consider a phenotype that is selectively neutral, so that per-allele effect size is uncorrelated with MAF (which means that variance explained is positively correlated with MAF, as additive genetic variance is defined as $2pqa^2$ where p and q are the major and minor allele frequency and a is the additive genetic effect). Since LD Score is also positively correlated with MAF, in this case we would expect variance explained to be positively correlated with LD Score, which will introduce downward bias in the LD Score regression intercept and upward bias in the LD Score regression slope, leading to an underestimate of potential bias.

To quantify the magnitude of the bias that MAF-dependent genetic architectures could introduce, we simulated a frequency-dependent genetic architecture where effect size was uncorrelated with MAF (Online Methods). For most phenotypes, this model should represent a reasonable bound of the genetic architecture. We observed minimal bias: in these simulations, the mean LD Score regression intercept was 0.994 (Supplementary Figure 6, Supplementary Table 7). Nevertheless, there exist extreme genetic architectures where LD Score regression is not effective: for instance if all causal variants are rare (MAF < 1%, which may be an appropriate model for a phenotype under extreme negative selection), then LD Score regression will often generate a negative slope, and the intercept will be exceed the mean χ^2 (Supplementary Figure 7).

Real Data

Finally, we applied LD Score regression to summary statistics from GWAS representing more than 20 different phenotypes[15–32] (see Table 1 and Supplementary Figures 8a–w. Metadata about the studies in the analysis are presented in Supplementary Tables 10a,b). For all studies, the slope of the LD Score regression was significantly greater than 0, and the LD Score regression intercept was substantially less than λ_{GC} (mean difference 0.11), suggesting that polygenicity significantly contributes to the increase in mean χ^2 and confirming that correcting test statistics by dividing by λ_{GC} is unnecessarily conservative. As an example, Figure 2 displays the LD Score regression for the most recent schizophrenia GWAS, restricted to ~70,000 European individuals[33]. The low intercept of 1.07 and indicates at most a small contribution of bias, and that the mean χ^2 of 1.613 results mostly from polygenicity. LD Score plots for all other GWAS included in table 1 can be found in Supplementary Figures 8a–w. As with any inference procedure that relies on a model of genetic architecture, it is possible that our results may be biased by model misspecifications other than those that we have simulated directly (e.g., if independent effect sizes are a poor model, perhaps because coupled alleles have a tendency to have effects in the same direction). This may explain the moderate inflation in the LD Score regression intercept that we observe in some large GWAS that are likely well-calibrated. Note that upward bias in the LD Score regression intercept means only that the intercept may be conservative as a correction factor.

DISCUSSION

Whenever possible, it is preferable to obtain all relevant genotype data and correct for confounding biases directly[34–38]; post-hoc correction of test statistics is no substitute for diligent quality control. However, in the event that only summary data are available, or if a conservative correction is desired, we propose that the LD Score regression intercept provides a more robust quantification of the extent of inflation from confounding bias than λ_{GC} (or intergenic λ_{GC}, Supplementary Table 8). Since λ_{GC}increases with sample size in the presence of polygenicity (even without confounding bias)[3], the gain in power obtained by correcting test statistics with the LD Score regression intercept instead of λ_{GC} will become even more substantial for larger GWAS. Extending this method to non-European populations such as East Asians or West Africans is straightforward given appropriate reference panels, but extension to admixed populations is the subject of future research.

In conclusion, we have developed LD Score regression, a method to distinguish between inflated test statistics from confounding bias and polygenicity. Application of LD Score regression to over 20 complex traits confirms that polygenicity accounts for the majority of test statistic inflation in GWAS results and this approach can be used to generate a correction factor for GWAS that retains more power than λ_{GC}, especially at large sample sizes. We have made available for download a Python command line tool for estimating LD Score and performing LD Score regression, and a database of LD Scores suitable for European-ancestry samples ([URLs]). Research in progress aims to apply this method to estimation of components of heritability, genetic correlation and the calibration of mixed model association statistics.

ONLINE METHODS

Estimation of LD Score

We estimated European LD Scores from 378 phased European individuals (excluding one individual from a pair of cousins) from the 1000 Genomes Project reference panel using the –ld-mean-rsq option implemented in the GCTA[39] software package (with flags --ld-mean-rsq –ld-rsq-cutoff 0 –maf 0.00001; we implemented a 1centiMorgan (cM) window using the –ld-wind flag and modified .bim files with physical coordinates replaced with genetic coordinates as described in the next paragraph – note that a 1cM window be achieved more conveniently using the flags –l2 and –ld-wind-cm in the LDSC software package by the authors). The primary rationale for using a sequenced reference panel containing several hundred individuals for LD Score estimation rather than a genotyped GWAS control panel with several thousand individuals was that even after imputing off-chip genotypes, the variants available from a genotyping array only account for a subset of all variants. Using only a subset of all variants for estimating LD Score produces estimates that are biased downwards.

We used a window of radius 1cM around the index variant for the sum of r^2's (using the genetic map and phased genotypes from the IMPUTE2 website, see URLs), no r^2 cutoff, and excluded singletons (MAF < 0.13%). The standard estimator of the Pearson correlation coefficient has upward bias of approximately 1 / N, where N is sample size, so we employed an approximately unbiased estimator of LD Score given by $r_{adj}^{\hat{2}} := r^{\hat{2}} - \frac{1 - r^{\hat{2}}}{N - 2}$ where $r^{\hat{2}}$ denotes the standard, biased estimator of the squared Pearson correlation. Note that it is possible to have $r_{adj}^{\hat{2}} < 1$, which is a mathematically necessary feature of any unbiased estimator of r^2. Thus, some estimated LD Scores will be less than 1. In practice, almost all variants with estimated LD Score less than 1 were rare: only 0.01% of variants with MAF > 5% had estimated LD Scores below 1.

We examined the effect of varying the window size on our estimates of LD Score, and found that our estimates of LD Score were robust to choice of window size. The mean difference in LD Scores estimated with a 1 cM window and a 2 cM window was less than 1% of the mean LD Score (Supplementary Figure 9), and all LD Scores estimated with window sizes larger than 1 cM had squared correlations > 0.99 (Supplementary Table 7). This observation also addresses concerns about inflation in the LD Score from the intra-European population structure in the 1000 Genomes reference panel. The mean inflation in the 1 cM LD Score from population structure can be approximately bounded by the mean difference between a 1 cM LD Score and a 2 cM LD Score. Since this difference is < 1% of the mean LD Score,

---

URLs

1. 1000 Genomes genetic map and haplotypes: http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html
2. LD Score database: ftp://atguftp.mgh.harvard.edu/brendan/1k_eur_r2_hm3snps_se_weights.RDS
3. Simulation and regression code for this paper: https://github.com/bulik/ld_score
4. Software tool for LD Score estimation and estimation of variance components from summary statistics: http://github.com/bulik/ldsc/
5. GIANT Consortium summary statistics: http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
6. PGC and TAG Consortium summary statistics: https://pgc.unc.edu/Sharing.php#SharingOpp
7. IIBDGC summary statistics (NB these summary statistics are meta-analyzed with immunochip data, which is not appropriate for LD Score regression): http://www.ibdgenetics.org/downloads.html
8. CARDIoGRAM summary statistics: http://www.cardiogramplusc4d.org/downloads/
9. DIAGRAM summary statistics: http://diagram-consortium.org/downloads.html
10. Rheumatoid Arthritis summary statistics: http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/
11. Blood Pressure summary statistics: http://www.georgehretlab.org/icbp_088023401234-9812599.html
12. MAGIC consortium summary statistics: http://www.magicinvestigators.org/downloads/
13. GEFOS consortium summary statistics: http://www.gefos.org/?q=content/data-release
14. SSGAC summary statistics: http://ssgac.org/Data.php

we conclude that bias from population structure is not significantly inflating our estimates of LD Score.

We estimated LD Score standard error via a delete-one jackknife over the 378 phased individuals in the 1000 Genomes European reference panel. We found that the LD Score standard error was positively correlated with MAF and with LD Score itself. Jackknife estimates of LD Score standard error became extremely large for variants with MAF < 1%, so we excluded variants with 1000 Genomes European sample MAF < 1% from all LD Score regressions.

Intra-European LD Score Differences

In order to quantify the magnitude of intra-European differences in LD Score, we estimated LD Scores using each of the 1000 Genomes European subpopulations: Utah Residents with Northern and Western European Ancestry (CEU), British in England and Scotland (GBR), Toscani in Italia (TSI) and Finnish in Finland (FIN). The LD Scores from the four subpopulations were all highly correlated but the mean LD Score was not constant across populations. The mean LD Scores (MAF > 1%) were EUR, 110; CEU, 109; GBR, 104; FIN, 117; TSI, 102. The observation that the mean LD Score in the Finnish (FIN) population was elevated is consistent with a recent bottleneck in the genetic history of Finland[40], and the observation that the mean LD Score in the Southern European TSI population is lower is consistent with reports that Southern European populations have gone through less severe bottlenecks than Northern European populations.

Intra-European differences in LD Score can be a source of bias in the LD Score regression intercept. For instance, if one attempts to perform LD Score regression using the 1000 Genomes European LD Score on a GWAS with all samples from Finland, then the LD Score regression intercept may be biased upwards. Similarly, if one attempts to perform LD Score regression using the 1000 Genomes European LD Score on a GWAS with all samples from Italy, the LD Score regression intercept may be biased downwards. If we make the approximation that the intra-European differences in LD Score can be described by an additive term plus 5% noise (*i.e.*, if we assume that the FIN LD Score equals the pan-European LD Score plus seven, which is a worst-case scenario among linear relationships between the two LD Scores in terms of bias in the intercept), then the bias introduced into the LD Score regression intercept by using the pan-European LD Score to perform LD Score regression on a Finnish GWAS will be 7 multiplied by the slope of the LD Score regression plus 5% of mean(χ^2)-1, where 7 is the difference between the reference population LD Score and the GWAS population LD Score. Since all of the mean European subpopulation LD Scores that we have estimated are within ± 8 of the mean pan-European LD Score, we estimate that the bias in the LD Score regression intercept from intra-European LD Score differences is at most ±10 times the LD Score regression slope. For the real GWAS analyzed in Table 1, this corresponds to a worst-case difference of approximately ±10% in the estimate of the proportion of the inflation in the mean χ^2 that results from confounding bias, with a higher probability of upward bias (because the noise term in the relationship between target and reference LD Score always causes upward bias in the LD Score regression intercept, while systematic directional differences in target and reference LD Scores can bias the LD Score regression intercept in either direction).

Regression Weights

In order to produce an efficient regression estimator, we must deal with two problems. First, χ^2-statistics at SNPs in LD are correlated. Second, the χ^2-statistics of variants with high LD Score have higher variance than the χ^2-statistics of variants with low LD Score (heteroskedasticity).

The statistically optimal solution to the correlation problem is to perform generalized lease squares (GLS) with the variance-covariance matrix of χ^2-statistics. However, this matrix is intractable under our model. As an approximation, we correct for correlation by weighting variant *j* by the reciprocal of the LD Score of variant *j* counting LD only with other SNPs included in the regression. Precisely, if we let *S* denote the set of variants included in the LD Score regression then the LD Score of variant *j* counting LD only with other SNPs included in the regression is $\ell_j(S) := 1+\sum_{k\in S} r_{jk}^2$ Weighting by $1/l_j(S)$ would be equivalent to GLS with the full variance-covariance matrix of χ^2-statistics if the genome consisted of LD blocks and r^2 (in the population) was either zero or one. We estimate $l_j(S)$ for the set of variants S described in the section *Application to Real Data* using the same procedure we used to estimate the full 1000 Genomes LD Score. Since our estimates of $\hat{l_j}$ can be negative and regression weights must be positive, we weight by $1/\max(\hat{l_j},1)$.

To account for heteroskedasticity, we weight by $1(1+Nh_g^2\ell_j/M)^2$, which is the reciprocal of the conditional variance function $\mathrm{Var}[\chi_j^2|\ell_j]$ under our model if we make the additional assumption that per-normalized genotype effect sizes are normally distributed (note that violation of this assumption does not bias the regression, it only increases the standard error. A derivation is provided in the Supplementary Note).

Attenuation Bias

Standard least-squares and weighted least-squares regression theory assumes that the explanatory variable (also referred to as the independent variable, or *X*) is measured without error. If the explanatory variable is measured with error, then the magnitude of the regression slope will be biased toward zero. This form of bias is known as attenuation bias. If the explanatory variable is measured with error, but the variance of this error is known, then it is possible to produce an unbiased regression slope by multiplying the slope by a disattenuation factor, which is equal to the squared weighted Pearson correlation between the noisy estimates of the explanatory variable and the true value of the explanatory variable. We provide an R script that can estimate this disattenuation factor given LD Scores and jackknife estimates of LD Score standard errors (see URLs).

Simulations

When performing simulations with polygenic genetic architectures using genotyped or imputed data, variants in the 1000 Genomes reference panel not included in the set of genotypes used for simulation cannot contribute to the simulated phenotypes, and so should not contribute to the LD Score used for simulations. Precisely, for the simulations with polygenicity and the simulations with polygenicity and bias, we used LD Scores where estimates of r^2 were derived from the 1000 Genomes European reference panel, but the sum of r^2's was taken over only those SNPs included in the simulations. For the simulations with frequency-dependent genetic architecture, we estimated LD Scores from the same genotypes used for simulations, because we wanted to quantify the bias introduced by frequency-dependent genetic architecture even when LD Scores are estimated with little noise. For the simulations with pure population stratification, we used an LD Score estimated from all 1000 Genomes variants, since there was no simulated polygenic architecture in these simulations. For simulations with pure population stratification, the details of the cohorts used are given in supplementary table 1.

It is difficult to use real genotypes to simulate ascertained studies of a binary phenotype with low population prevalence: to obtain 1000 cases with a simulated 1% phenotype, one would need to sample on expectation 100,000 genotypes, which is not feasible. We therefore generated simulated genotypes at 1.1 million SNPs with mean LD Score 110 and a simplified LD structure where r^2 is either 0 or 1, and all variants had 50% minor allele frequency. We generated phenotypes under the liability threshold model with all per-normalized genotype effect sizes (*i.e.*, effects on liability) drawn *i.i.d.* from a normal distribution, then sampled individuals at random from the simulated population until the desired number of cases and controls for the study had been reached. The R script that performs these simulations is available online ([URLs]).

Application to Real Data

The majority of the sets of summary statistics that we analyzed did not contain information about sample minor allele frequency or imputation quality. In order to restrict to a set of common, well-imputed variants, we retained only those SNPs in the HapMap 3 reference panel[41] for the LD Score regression. To guard against underestimation of LD Score from summing only LD with variants within a 1cM window, we removed variants in regions with exceptionally long-range LD[42] from the LD Score regression (NB LD with these variants were included in the estimation of LD Score). Lastly, we excluded pericentromeric regions (defined as ± 3 cM from a centromere) from the LD Score regression, because these regions are enriched for sequence gaps, which may lead to underestimation of LD Score, and depleted for genes, which may reduce the probability of association to phenotype[43,44]. The final set of variants retained for LD Score regression on real data consisted of approximately 1.1 million variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

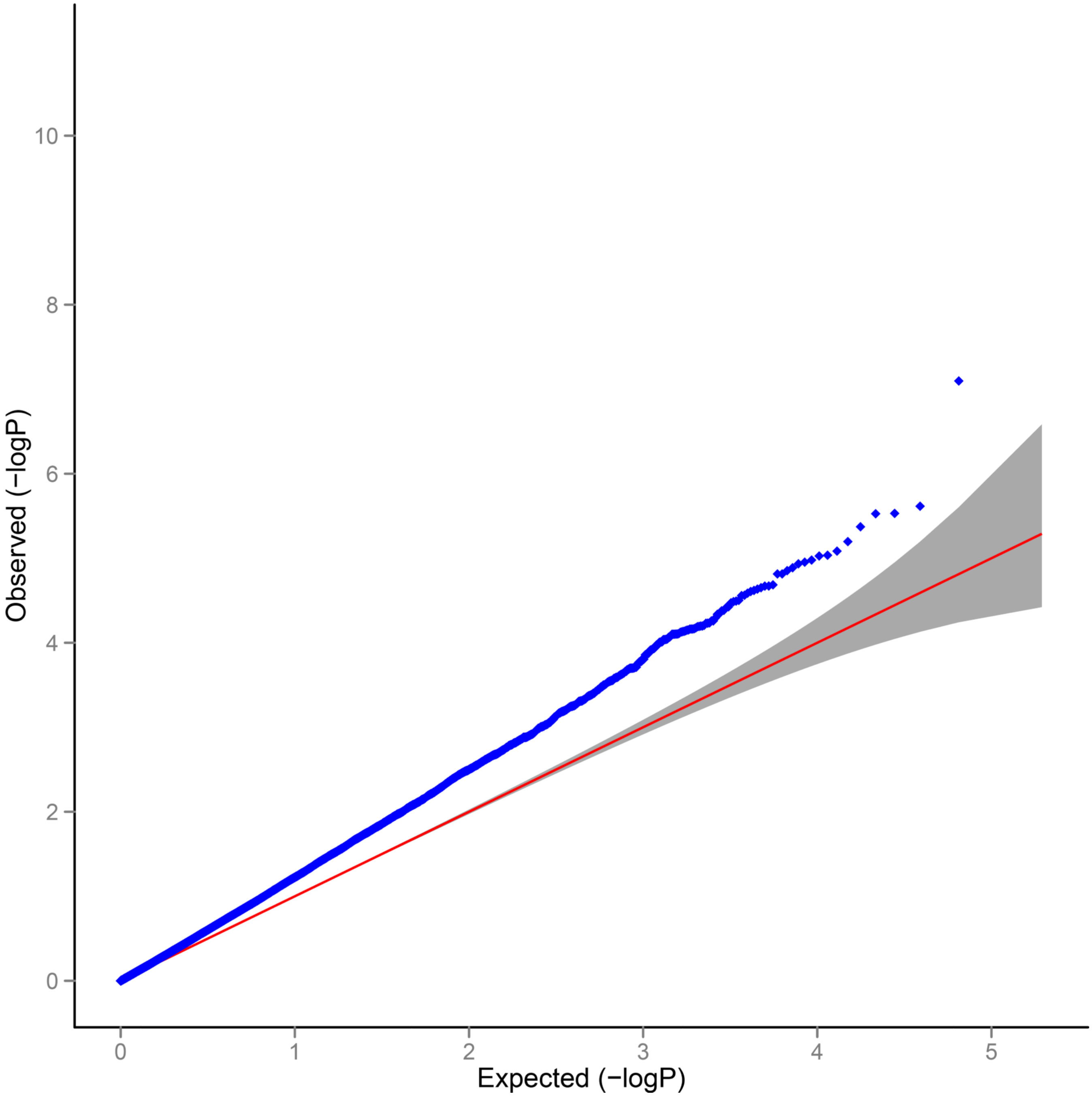
ACKNOWLEDGEMENTS

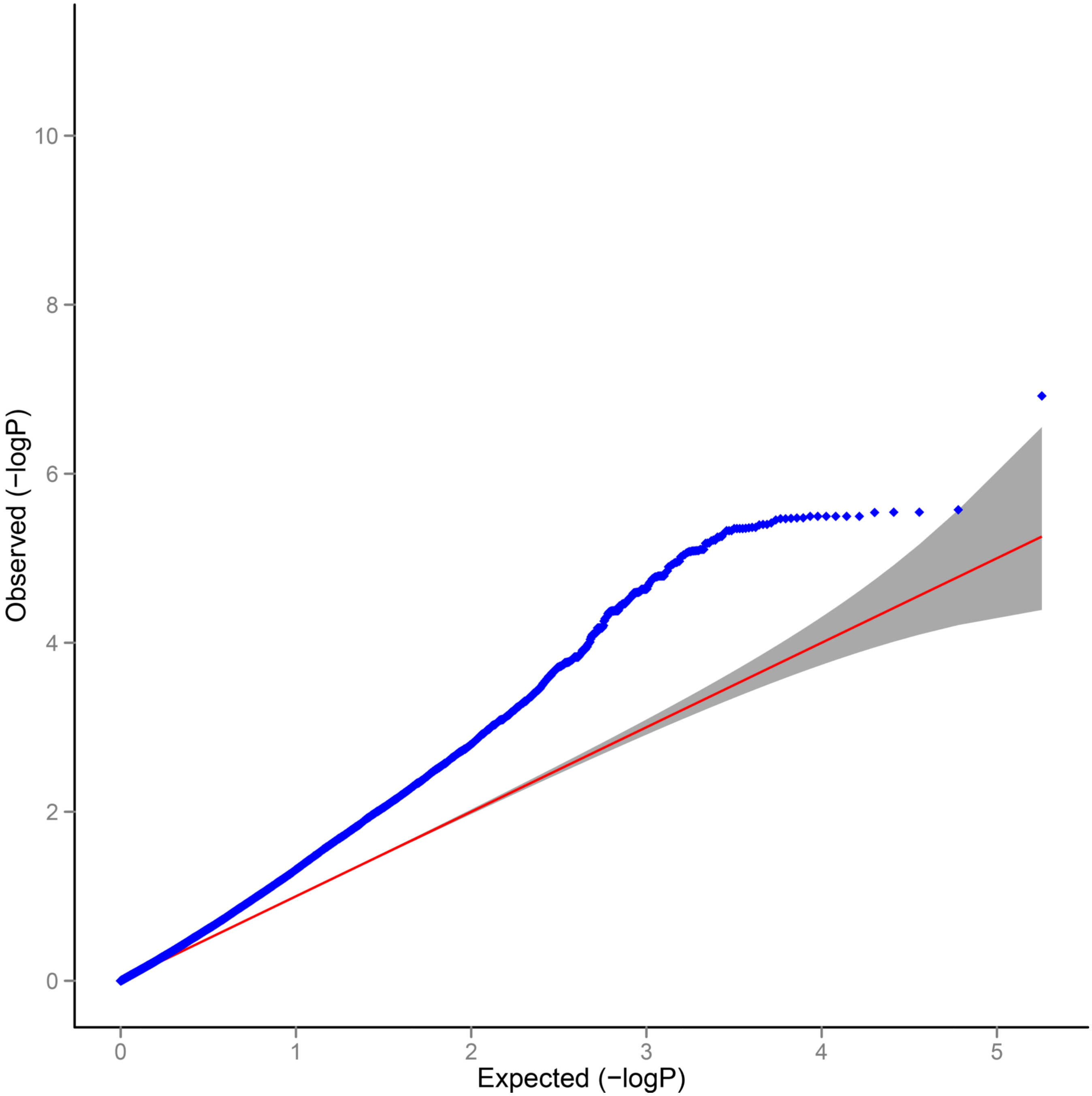
We would like to thank P. Sullivan for helpful discussion. This work was supported by NIH grants R01 HG006399 (ALP), R03 CA173785 (HF) and R01 MH094421 (PGC) and by the Fannie and John Hertz Foundation (HF). Data on coronary artery disease / myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.CARDIOGRAMPLUSC4D.ORG. Finally, we thank the coffee machine in the ATGU common area for inspiration.

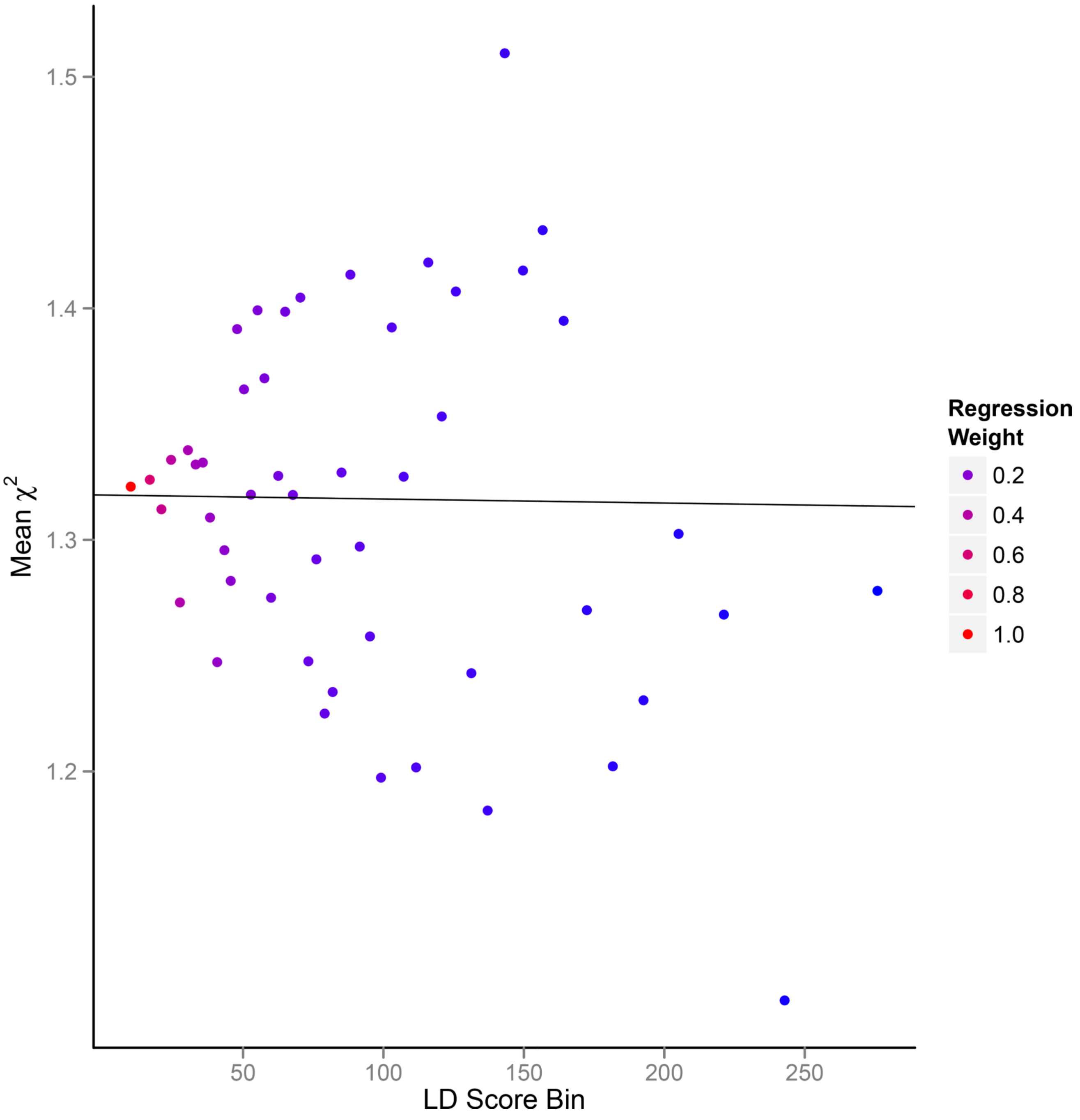
References

1. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. Am J Hum Genet. 2001; 69:1–14. [PubMed: 11410837]
2. Sham PC, Cherny SS, Purcell S, Hewitt JK. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. Am J Hum Genet. 2000; 66:1616–1630. [PubMed: 10762547]
3. Yang J, et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 2011; 19:807–812. [PubMed: 21407268]
4. Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 2005; 1:e32. [PubMed: 16151517]
5. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]
6. Lin DY, Sullivan PF. Meta-analysis of genome-wide association studies with overlapping subjects. Am J Hum Genet. 2009; 85:862–872. [PubMed: 20004761]
7. Consortium TGP, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]
8. Yin PFX. Estimating R2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods. The Journal of Experimental Education. 2001; 69:203–224.
9. Ralph P, Coop G. The geography of recent genetic ancestry across Europe. PLoS Biol. 2013; 11:e1001555. [PubMed: 23667324]
10. Bersaglieri T, et al. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004; 74:1111–1120. [PubMed: 15114531]
11. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009; 5:e1000471. [PubMed: 19424416]
12. Price AL, et al. The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet. 2009; 5:e1000505. [PubMed: 19503599]
13. International Multiple Sclerosis Genetics, C. et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011; 476:214–219. [PubMed: 21833088]
14. Splansky GL, et al. The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am J Epidemiol. 2007; 165:1328–1335. [PubMed: 17372189]
15. Sullivan PF, et al. Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. Mol Psychiatry. 2009; 14:359–375. [PubMed: 19065144]
16. Heid IM, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet. 2010; 42:949–960. [PubMed: 20935629]
17. Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–838. [PubMed: 20881960]
18. Neale BM, et al. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry. 2010; 49:884–897. [PubMed: 20732625]
19. Speliotes EK, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010; 42:937–948. [PubMed: 20935630]
20. Stahl EA, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010; 42:508–514. [PubMed: 20453842]
21. Tobacco & Genetics, C. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010; 42:441–447. [PubMed: 20418890]

22. International Consortium for Blood Pressure Genome-Wide Association, S. et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011; 478:103–109. [PubMed: 21909115]
23. Psychiatric, G.C.B.D.W.G. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet. 2011; 43:977–983. [PubMed: 21926972]
24. Schunkert H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011; 43:333–338. [PubMed: 21378990]
25. Estrada K, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nat Genet. 2012; 44:491–501. [PubMed: 22504420]
26. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–124. [PubMed: 23128233]
27. Manning AK, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet. 2012; 44:659–669. [PubMed: 22581228]
28. Morris AP, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. 2012; 44:981–990. [PubMed: 22885922]
29. Cross-Disorder Group of the Psychiatric Genomics, C. & Genetic Risk Outcome of Psychosis, C. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013; 381:1371–1379. [PubMed: 23453885]
30. Major Depressive Disorder Working Group of the Psychiatric, G.C.et al. A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry. 2013; 18:497–511. [PubMed: 22472876]
31. Rietveld CA, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science. 2013; 340:1467–1471. [PubMed: 23722424]
32. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–427. [PubMed: 25056061]
33. Consortium., S.W.G.o.t.P.G. Biological Insights from 108 Schizophrenia-Associated Genetic Loci. Nature. 2014; 511:421–427. [PubMed: 25056061]
34. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]
35. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]
36. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–354. [PubMed: 20208533]
37. Lippert C, et al. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011; 8:833–835. [PubMed: 21892150]
38. Korte A, et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012; 44:1066–1071. [PubMed: 22902788]
39. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88:76–82. [PubMed: 21167468]
40. Jakkula E, et al. The genome-wide patterns of variation expose significant substructure in a founder population. Am J Hum Genet. 2008; 83:787–794. [PubMed: 19061986]
41. International HapMap, C. et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]
42. Price AL, et al. Long-range LD can confound genome scans in admixed populations. Am J Hum Genet. 2008; 83:132–135. author reply 135-9. [PubMed: 18606306]
43. Smith AV, Thomas DJ, Munro HM, Abecasis GR. Sequence features in regions of weak and strong linkage disequilibrium. Genome Res. 2005; 15:1519–1534. [PubMed: 16251462]
44. She X, et al. The structure and evolution of centromeric transition regions within the human genome. Nature. 2004; 430:857–864. [PubMed: 15318213]


Observed (−logP)
0
2
4
6
8
10
Expected (−logP)
0
1
2
3
4
5


Observed (−logP)
Expected (−logP)
0
2
4
6
8
10
0
1
2
3
4
5


1.5
1.4
1.3
1.2
Mean χ^2
50
100
150
200
250
LD Score Bin
Regression Weight
0.2
0.4
0.6
0.8
1.0

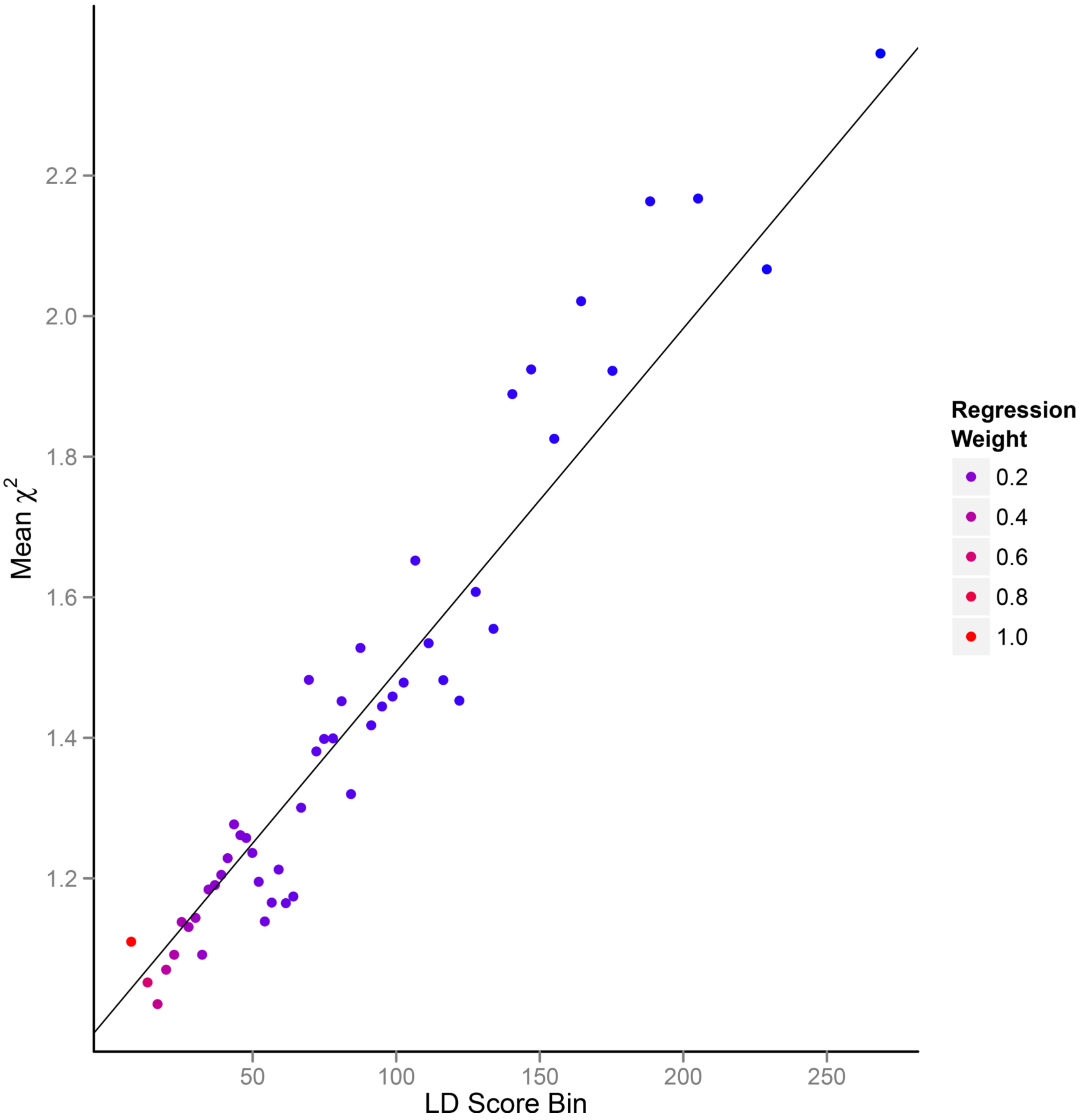


Figure 1.

Results from selected simulations. **(a)** QQ plot with population stratification ($\lambda_{GC} = 1.32$, LD Score regression intercept = 1.30). **(b)** QQ plot with polygenic genetic architecture with 0.1% of SNPs causal ($\lambda_{GC} = 1.32$, LD Score regression intercept = 1.006) **(c)** LD Score plot with population stratification. Each point represents an LD Score quantile, where the *x*-coordinate of the point is the mean LD Score of variants in that quantile and the *y*-coordinate is the mean χ^2 of variants in that quantile. Colors correspond to regression weights, with red indicating large weight. The black line is the LD Score regression line. **(d)** As in panel c but LD Score plot with polygenic genetic architecture.

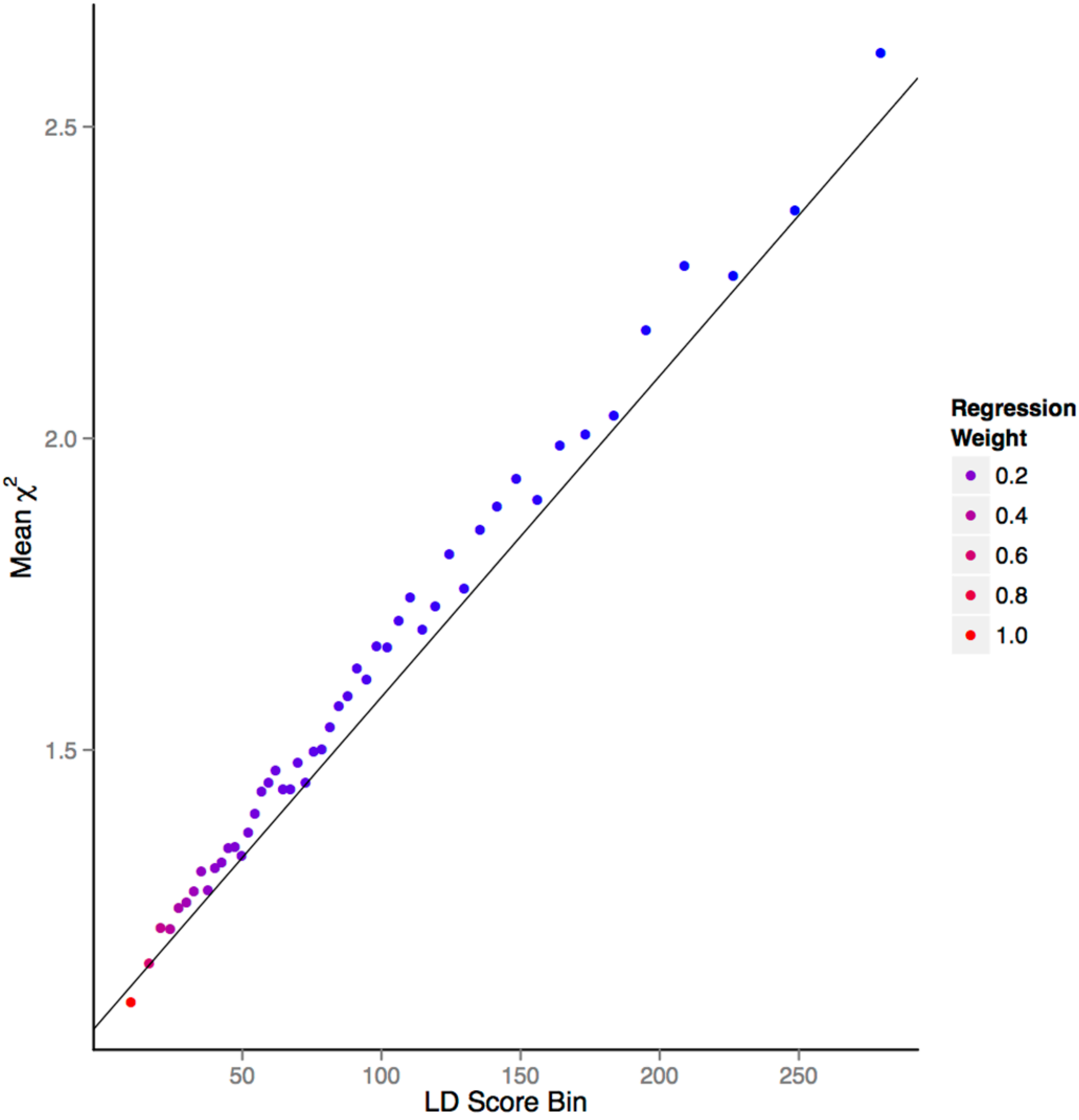


Figure 2.

D Score regression plot for the current schizophrenia meta-analysis[33]. Each point represents an LD Score quantile, where the *x*-coordinate of the point is the mean LD Score of variants in that quantile and the *y*-coordinate is the mean χ^2 of variants in that quantile. Colors correspond to regression weights, with red indicating large weight. The black line is the LD Score regression line. The line appears to fall below the points on the right because this is a weighted regression in which the points on the left receive the largest weights (Online Methods).

Table 1

LD Score regression results for all studies analyzed that either did not apply meta-analysis level GC correction or listed λ_{GC} in the relevant publication. The column labeled “GC” indicates how many rounds of GC correction were performed. For GWAS that applied meta-analysis level GC correction and listed λ_{GC}, we re-inflated all test statistics by the meta-analysis level λ_{GC}. LD Score regression performed on GC-corrected summary statistics will generally yield an intercept less than one. Note that GC correction at the individual study level will also push the expected intercept in the absence of confounding slightly below one (Supplementary Note). Standard errors were obtained via a block jackknife over blocks of ~2000 adjacent SNPs, which provides a robust estimate of standard error in the presence of correlated, heteroskedastic error terms. The column labeled “Type” indicates whether the study was a mega-(raw genotypes shared between studies) or meta-analysis (only summary statistics shared between all contributing studies).

Phenotype	**Mean χ^2**	**λ_{GC}**	**Intercept (SE)**	**Type**	**GC**	**Ref**
Inflammatory bowel disease	1.247	1.164	1.095 (0.010)	mega	0	[26]
Ulcerative Colitis	1.174	1.128	1.079 (0.010)	mega	0	[26]
Crohn’s Disease	1.185	1.122	1.059 (0.008)	mega	0	[26]
Schizophrenia	1.613	1.484	1.070 (0.010)	mega	0	[33]
ADHD	1.033	1.033	1.008 (0.006)	mega	0	[18]
Bipolar Disorder	1.154	1.135	1.030 (0.008)	mega	0	[23]
PGC Cross-Disorder	1.205	1.187	1.018 (0.008)	mega	0	[29]
Major Depression	1.063	1.063	1.009 (0.006)	mega	0	[30]
Rheumatoid Arthritis	1.063	1.033	0.980 (0.007)	mega	2	[20]
Coronary Artery Disease	1.125	1.096	1.033 (0.008)	meta	1	[24]
Type-2 Diabetes	1.116	1.097	1.025 (0.008)	meta	1	[28]
BMI-Adj. Fasting Insulin	1.088	1.072	1.015 (0.007)	meta	1	[27]
Fasting Insulin	1.079	1.067	1.021 (0.007)	meta	1	[27]
College (Yes/No)	1.207	1.180	1.046 (0.009)	meta	1	[31]
Years of Education	1.220	1.188	1.041 (0.009)	meta	1	[31]
Cigarettes Per Day	1.047	1.047	0.998 (0.008)	meta	1	[21]
Ever Smoked	1.097	1.083	1.008 (0.006)	meta	1	[21]
Former Smoker	1.050	1.048	0.999 (0.007)	meta	1	[21]
Age-Onset (Smoking)	1.025	1.030	0.998 (0.006)	meta	1	[21]

Phenotype	**Mean χ^2**	λ_{GC}	**Intercept (SE)**	**Type**	**GC**	**Ref**
FN-BMD	1.163	1.109	1.001 (0.009)	meta	2	[25]
LS-BMD	1.174	1.112	1.032 (0.009)	meta	2	[25]
Waist-Hip Ratio	1.417	1.330	1.040 (0.008)	meta	2	[16]
Height	1.802	1.478	1.149 (0.021)	meta	2	[17]
Body-Mass Index	1.130	1.090	1.033 (0.012)	meta	2	[19]