ARTICLE

Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores

Bjarni J. Vilhjálmsson,[1,2,3,4,*] Jian Yang,[5,6] Hilary K. Finucane,[1,2,3,7] Alexander Gusev,[1,2,3] Sara Lindström,[1,2] Stephan Ripke,[8,9,10] Giulio Genovese,[3,8,11] Po-Ru Loh,[1,2,3] Gaurav Bhatia,[1,2,3] Ron Do,[12,13] Tristan Hayeck,[1,2,3] Hong-Hee Won,[3,14] Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Sekar Kathiresan,[3,14] Michele Pato,[15] Carlos Pato,[15] Rulla Tamimi,[1,2,16] Eli Stahl,[3,13,17,18] Noah Zaitlen,[19] Bogdan Pasaniuc,[20] Gillian Belbin,[12,13] Eimear E. Kenny,[12,13,18,21] Mikkel H. Schierup,[4] Philip De Jager,[3,22,23] Nikolaos A. Patsopoulos,[3,22,23] Steve McCarroll,[3,8,11] Mark Daly,[3,8] Shaun Purcell,[3,13,17,18] Daniel Chasman,[22,24] Benjamin Neale,[3,8] Michael Goddard,[25,26] Peter M. Visscher,[5,6] Peter Kraft,[1,2,3,27] Nick Patterson,[3] and Alkes L. Price[1,2,3,27,*]

Polygenic risk scores have shown great promise in predicting complex disease risk and will become more accurate as training sample sizes increase. The standard approach for calculating risk scores involves linkage disequilibrium (LD)-based marker pruning and applying a p value threshold to association statistics, but this discards information and can reduce predictive accuracy. We introduce LDpred, a method that infers the posterior mean effect size of each marker by using a prior on effect sizes and LD information from an external reference panel. Theory and simulations show that LDpred outperforms the approach of pruning followed by thresholding, particularly at large sample sizes. Accordingly, predicted R^2 increased from 20.1% to 25.3% in a large schizophrenia dataset and from 9.8% to 12.0% in a large multiple sclerosis dataset. A similar relative improvement in accuracy was observed for three additional large disease datasets and for non-European schizophrenia samples. The advantage of LDpred over existing methods will grow as sample sizes increase.

Introduction

Polygenic risk scores (PRSs) computed from genome-wide association study (GWAS) summary statistics have proven valuable for predicting disease risk and understanding the genetic architecture of complex traits. PRSs were used for predicting genetic risk in a schizophrenia (SCZ) GWAS for which there was only one genome-wide-significant locus[1] and have been widely used for predicting genetic risk for many traits.[1–14] PRSs can also be used for drawing inferences about genetic architectures within and across traits.[11,12,15–17] As GWAS sample sizes grow, the prediction accuracy of PRSs will increase and might eventually yield clinically actionable predictions.[15,18–20] However, as noted in recent work,[18] current PRS methods do not account for the effects of linkage disequilibrium (LD), which limits their predictive value, especially for large samples. Indeed, our simulations show that, in the presence of LD, the prediction accuracy of the widely used approach of LD pruning followed by p value thresholding (P+T)[1,6,8,9,11,12,14,15,18,19] falls short of the heritability explained by the SNPs (Figure 1 and Figure S1; see Material and Methods).

One possible solution to this problem is to use one of the many available prediction methods that require genotype data as input. These include genomic BLUP—which assumes an infinitesimal distribution of effect sizes—and its extensions to non-infinitesimal mixture priors.[21–28] However, these methods are not applicable to GWAS summary statistics when genotype data are unavailable

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [2]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [3]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [4]Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark; [5]Queensland Brain Institute, University of Queensland, Brisbane, 4072 QLD, Australia; [6]Diamantina Institute, Translational Research Institute, University of Queensland, Brisbane, 4101 QLD, Australia; [7]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [8]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [9]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; [10]Deptartment of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Campus Mitte, 10117 Berlin, Germany; [11]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [12]Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [13]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [14]Cardiovascular Research Center, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; [15]Department of Psychiatry and Behavioral Sciences, Keck School of Medicine at University of Southern California, Los Angeles, CA 90089, USA; [16]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA; [17]Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [18]Center of Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [19]Lung Biology Center, Department of Medicine, University of California, San Francisco, San Francisco, CA 94143, USA; [20]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA; [21]Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [22]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA; [23]Program in Translational NeuroPsychiatric Genomics, Ann Romney Center for Neurologic Diseases, Department of Neurology, Brigham and Women's Hospital, Boston, MA 02115, USA; [24]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA 02215, USA; [25]Department of Food and Agricultural Systems, University of Melbourne, Parkville, 3010 VIC, Australia; [26]Biosciences Research Division, Department of Primary Industries, Bundoora, 3083 VIC, Australia; [27]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
*Correspondence: bjarni.vilhjalmsson@gmail.com (B.J.V.), aprice@hsph.harvard.edu (A.L.P.)
http://dx.doi.org/10.1016/j.ajhg.2015.09.001.

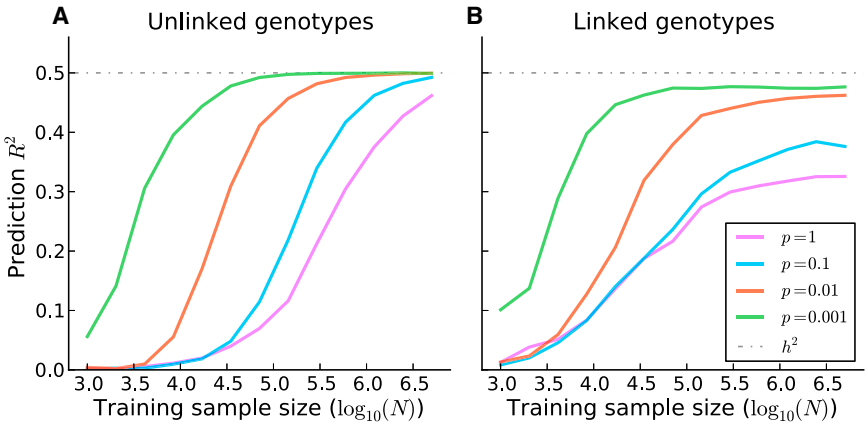

Figure 1. Prediction Accuracy of P+T Applied to Simulated Genotypes with and without LD

The performance of P+T, PRSs based on LD-pruned SNPs ($r^2 < 0.2$) followed by p value thresholding with an optimized threshold, when applied to simulated genotypes with and without LD. The prediction accuracy, as measured by squared correlation between the true phenotypes and the PRSs (prediction R^2), is plotted as a function of the training sample size. The results are averaged over 1,000 simulated traits with 200,000 simulated genotypes, where the fraction of causal variants p was allowed to vary. In (A), the simulated genotypes are unlinked. In (B), the simulated genotypes are linked; we simulated independent batches of 100 markers while fixing the squared correlation between adjacent variants in a batch at 0.9.

because of privacy concerns or logistical constraints, as is often the case. In addition, many of these methods become computationally intractable at the very large sample sizes (>100,000 individuals) that would be required for achieving clinically relevant predictions for most common diseases.[15,18,19]

In this study, we propose LDpred, a Bayesian PRS that estimates posterior mean causal effect sizes from GWAS summary statistics by assuming a prior for the genetic architecture and LD information from a reference panel. By using a point-normal mixture prior[25,29] for the marker effects, LDpred can be applied to traits and diseases with a wide range of genetic architectures. Unlike P+T, LDpred has the desirable property that its prediction accuracy converges to the heritability explained by the SNPs as sample size grows (see below). Using simulations based on real genotypes, we compare the prediction accuracy of LDpred to that of the widely used approach of P+T,[1,6,8,9,11,12,14,15,18,19,30] as well as other approaches that train on GWAS summary statistics. We apply LDpred to seven common diseases for which raw genotypes are available in small sample size and to five common diseases for which only summary statistics are available in large sample size.

Material and Methods

Overview of Methods

LDpred calculates the posterior mean effects from GWAS summary statistics by conditioning on a genetic architecture prior and LD information from a reference panel. The inner product of these re-weighted and the test-sample genotypes is the posterior mean phenotype and thus, under the model assumptions and available data, an optimal (minimum variance and unbiased) predictor.[31] The prior for the effect sizes is a point-normal mixture distribution, which allows for non-infinitesimal genetic architectures. The prior has two parameters: the heritability explained by the genotypes and the fraction of causal markers (i.e., the fraction of markers with non-zero effects). The heritability parameter is estimated from GWAS summary statistics and accounts for sampling noise and LD[32–34] (see details below). The fraction of causal markers is allowed to vary and can be optimized with respect to prediction accuracy in a validation dataset, analogous to how P+T is applied in practice. Hence, similar to P+T (where p value thresholds are varied and multiple PRSs are calculated), multiple LDpred risk scores are calculated with the use of priors with varying fractions of markers with non-zero effects. The value that optimizes prediction accuracy can then be determined in an independent validation dataset. We approximate LD by using data from a reference panel (e.g., independent validation data). We estimate the posterior mean effect sizes via the Markov chain Monte Carlo (MCMC) method and apply them to validation data to obtain PRSs. In the special case of no LD, posterior mean effect sizes with a point-normal prior can be viewed as a soft threshold and can be computed analytically (Figure S2; see details below). We have released open-source software implementing the method (see Web Resources).

A key feature of LDpred is that it relies on GWAS summary statistics, which are often available even when raw genotypes are not. In our comparison of methods, we therefore focus primarily on PRSs that rely on GWAS summary statistics. The main approaches that we compare with LDpred are listed in Table S1. These include PRS based on all markers (unadjusted PRS), P+T, and LDpred specialized to an infinitesimal prior (LDpred-inf) (see details below). We note that LDpred-inf is an analytic method, given that posterior mean effects are closely approximated by

$$E(\beta \mid \tilde{\beta}, D) \approx \left(\frac{M}{Nh_g^2} I + D\right)^{-1} \tilde{\beta}, \qquad \text{(Equation 1)}$$

where D denotes the LD matrix between the markers in the training data, and $\tilde{\beta}$ denotes the marginally estimated marker effects (see details below). LDpred-inf (using GWAS summary statistics) is analogous to genomic BLUP (using raw genotypes) because it assumes the same prior.

Phenotype Model

Let Y be a $N \times 1$ phenotype vector and X be a $N \times M$ genotype matrix, where N is the number of individuals, and M is the number of genetic variants. For simplicity, we will assume throughout that the phenotype Y and individual genetic variants X_i have been

mean centered and standardized to have variance 1. We model the phenotype as a linear combination of M genetic effects and an independent environmental effect ε, i.e., $Y = \sum_{i=1}^{M} X_i\beta_i + \varepsilon$, where X_i denotes the i^{th} genetic variant, β_i is its true effect, and ε is the environmental and noise contribution. In this setting, the (marginal) least-squares estimate of an individual marker effect is $\widehat{\beta}_i = X_i'Y/N$. For clarity, we implicitly assume that we have the standardized effect estimates available to us as summary statistics. In practice, we usually have other summary statistics, including the p value and direction of the effect estimates, from which we infer the standardized effect estimates. First, we exclude all markers with ambiguous effect directions, i.e., A/T and G/C SNPs. Second, from the p values we obtain Z scores and multiply them by the sign of the effects (obtained from the effect estimates or effect direction). Finally, we approximate the least-squares estimate for the effect by $\widehat{\beta}_i = s_i(z_i/\sqrt{N})$, where s_i is the sign, and z_i is the Z score obtained from the p value. If the trait is a case-control trait, this transformation from the p value to the effect size can be thought of as being an effect estimate for an underlying quantitative liability or risk trait.[35]

Unadjusted PRS

The unadjusted PRS is simply the sum of all the estimated marker effects for each allele, i.e., the standard unadjusted polygenic score for the i^{th} individual is $S_i = \sum_{j=1}^{M} X_{ji}\widehat{\beta}_j$, where X_{ji} denotes the genotype for the i^{th} individual and the j^{th} genetic variant.

P+T

In practice, the prediction accuracy is improved if the markers are LD pruned and p value pruned a priori. Informed LD pruning (also known as LD clumping), which preferentially prunes the less significant marker, often yields much more accurate predictions than pruning random markers. Applying a p value threshold, i.e., using only markers that achieve a given significance threshold, also improves prediction accuracies for many traits and diseases. In this paper, P+T refers to the strategy of first applying informed LD pruning with r^2 threshold 0.2 and subsequently applying p value thresholding, where the p value threshold is optimized over a grid with respect to prediction accuracy in the validation data.

Bpred: Bayesian Approach in the Special Case of No LD

Under a model, the optimal linear prediction given some statistic is the posterior mean prediction. This prediction is optimal in the sense that it minimizes the prediction error variance.[36] Under the linear model described above, the posterior mean phenotype given GWAS summary statistics and LD is

$$E\left(Y \mid \tilde{\beta}, \widehat{D}\right) = \sum_{i=1}^{M} X_i' E\left(\beta_i \mid \tilde{\beta}, \widehat{D}\right).$$

Here, $\tilde{\beta}$ denotes a vector of marginally estimated least-squares estimates obtained from the GWAS summary statistics, and $\widehat{D}$ refers to the observed genome-wide LD matrix in the training data, i.e., the samples for which the effect estimates are calculated. Hence, the quantity of interest is the posterior mean marker effect given LD information from the GWAS sample and the GWAS summary statistics. In practice, we might not have this information available to us and are forced to estimate the LD from a reference panel. In most of our analyses, we estimated the local LD structure in the training data from the independent validation data. Although this choice of LD reference panel can lead to small bias when one estimates individual prediction accuracy, this choice is valid whenever the aim is to calculate accurate PRSs for a cohort without knowing the case-control status a priori. In other words, it is an unbiased estimate for the PRS accuracy when the validation data are used as an LD reference, which we recommend in practice.

The variance of the trait can be partitioned into a heritable part and the noise, i.e., $\text{Var}(Y) = h_g^2\Theta + (1 - h_g^2)\text{I}$, where h_g^2 denotes the heritability explained by the genotyped variants, and $\Theta = XX'/M$ is the SNP-based genetic relationship matrix. We can obtain a trait with the desired covariance structure if we sample the betas independently with mean 0 and variance h_g^2/M. Note that if the effects are independently sampled, then this also holds true for correlated genotypes, i.e., when there is LD. However, LD will increase the variance of heritability explained by the genotypes as estimated from the data (as a result of fewer effective independent markers).

If all samples are independent and all markers are unlinked and have effects drawn from a Gaussian distribution, i.e., $\beta_i \sim_{iid} N(0, (h_g^2/M))$, then this is an infinitesimal model,[37] where all markers are causal. Under this model, the posterior mean can be derived analytically, as shown by Dudbridge[15]:

$$E(\beta_i \mid \tilde{\beta}) = E(\beta_i \mid \tilde{\beta}_i) = \left(\frac{h_g^2}{h_g^2 + \frac{M}{N}}\right)\tilde{\beta}_i.$$

Interestingly, with unlinked markers, the infinitesimal shrink factor times the heritability, i.e.,

$$\left(\frac{h_g^2}{h_g^2 + \frac{M}{N}}\right)h_g^2,$$

is the expected squared correlation between the unadjusted PRS (with unlinked markers) and the phenotype, regardless of the underlying genetic architecture.[38,39]

An arguably more reasonable prior for the effect sizes is a non-infinitesimal model, where only a fraction of the markers are causal. For this, consider the following Gaussian mixture prior:

$$\beta_i \sim_{iid} \begin{cases} N\left(0, \dfrac{h_g^2}{Mp}\right) \text{with probability } p \\ 0 \text{ with probability } (1-p), \end{cases}$$

where p is the probability that a marker is drawn from a Gaussian distribution, i.e., the fraction of causal markers. Under this model, the posterior mean can be derived as (see Appendix A)

$$\text{E}(\beta_i \mid \tilde{\beta}_i) = \left(\frac{h_g^2}{h_g^2 + \frac{Mp}{N}}\right)\overline{p}_i\tilde{\beta}_i,$$

where $\overline{p}_i$ is the posterior probability that the i^{th} marker is causal and can be calculated analytically (see Equation A1 in Appendix A). In our simulations, we refer to this Bayesian shrink without LD as Bpred.

LDpred: Bayesian Approach in the Presence of LD

If we allow for loci to be linked, then we can derive posterior mean effects analytically under a Gaussian infinitesimal prior (described above). We call the resulting method LDpred-inf, and it represents a computationally efficient special case of LDpred. If we assume that distant markers are unlinked, the posterior mean for the effect sizes within a small region l under an infinitesimal model is well approximated by

$$E(\beta^l \mid \tilde{\beta}^l, D) \approx \left(\frac{M}{Nh_g^2} I + D_l\right)^{-1} \tilde{\beta}^l.$$

Here, D_l denotes the regional LD matrix within the region of LD, and $\tilde{\beta}^l$ denotes the least-squares-estimated effects within that region. The approximation assumes that the heritability explained by the region is small and that LD with SNPs outside of the region is negligible. Interestingly, under these assumptions the resulting effects approximate the standard mixed-model genomic BLUP effects. LDpred-inf is therefore a natural extension of the genomic BLUP to summary statistics. A more detailed derivation is given in Appendix A. In practice, we do not know the LD pattern in the training data, and we need to estimate it by using LD in a reference panel.

Deriving an analytical expression for the posterior mean under a non-infinitesimal Gaussian mixture prior is difficult, and thus LDpred approximates it numerically by using an approximate MCMC Gibbs sampler. This is similar to the Gauss-Seidel approach, except that instead of using the posterior mean to update the effect size, we sample the update from the posterior distribution. Compared to the Gauss-Seidel method, this seems to lead to less serious convergence issues. The approximate Gibbs sampler is described in detail in Appendix A. To ensure convergence, we shrink the posterior probability of being causal by a fixed factor at each big iteration step i, where the shrinkage factor is defined as $c = \min(1, (\hat{h}_g^2/(\tilde{h}_g^2)_i))$, where $\hat{h}_g^2$ is the estimated heritability based on an aggregate approach (see below), and $(\tilde{h}_g^2)_i$ is the estimated genome-wide heritability at each big iteration. To speed up convergence in the Gibbs sampler, we used Rao-Blackwellization and observed that good convergence was usually attained with fewer than 100 iterations in practice (see Appendix A).

Estimation of the Heritability Parameter

In the absence of population structure and assuming independent and identically distributed mean-zero SNP effects, the following equation has been shown to hold:

$$E(\chi_j^2) = 1 + \frac{Nh_g^2 l_j}{M},$$

where χ_j^2 is the χ^2-distributed test statistic at the j^{th} SNP, and $l_j = \sum_k [r^2(j,k) - (1 - r^2(j,k)/N - 2)]$, summing over k neighboring SNPs in LD, is the LD score for the j^{th} SNP. Taking the average of both sides over SNPs and rearranging, we obtain a heritability estimate:

$$\tilde{h}_g^2 = \frac{(\overline{\chi^2} - 1)M}{\bar{l}N},$$

where $\overline{\chi^2} = \sum_j (\chi_j^2/M)$ and $\bar{l} = \sum_j (l_j/M)$. We call this the aggregate estimator, and it is equivalent to LD-score regression[32–34] with intercept constrained to 1 and SNP j weighted by $1/l_j$. Prediction accuracy is not predicated on the robustness of this estimator, which will be evaluated elsewhere. Following the conversion proposed by Lee et al.,[40] we also report the heritability on the liability scale.

Practical Considerations

When LDpred is applied to real data, two parameters need to be specified beforehand. The first parameter is the LD radius, i.e., the number of SNPs that we adjust for on each side of a given SNP. There is a trade-off when we decide on the LD radius. If the LD radius is too large, then errors in LD estimates can lead to apparent LD between unlinked loci, which can lead to worse effect estimates and poor convergence. If the LD radius is too small, then we risk not accounting for LD between linked loci. We found that a LD radius of approximately $M/3{,}000$ (the default value in LDpred), where M is the total number of SNPs used in the analysis, works well in practice; this corresponds to a 2 Mb LD window on average in the genome. We also note that LDpred is implemented with a sliding window along the genome, whereas LDpred-inf is implemented with tiling LD windows, because this is computationally more efficient and does not affect accuracy. Regarding choice of the LD panel, its LD structure should ideally be similar to the training data for which the summary statistics are calculated. In simulations, we found that the LD reference panel should contain at least 1,000 individuals.

The second parameter is the fraction p of non-zero effects in the prior. This parameter is analogous to the p value threshold used in P+T. Our recommendation is to try a range of values for p (e.g., 1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 3E−4, 1E−4, 3E−5, 1E−5; these are default values in LDpred). This will generate 11 sets of SNP weights, which can be used for calculating polygenic scores. One can then use independent validation data to optimize the parameter, analogous to how the p value threshold is optimized in the P+T method.

When using LDpred, we recommend that SNP weights (posterior mean effect sizes) are calculated for exactly the SNPs used in the validation data. This ensures that all SNPs with non-zero weights are in the validation dataset. In practice, we use the intersection of SNPs present in the summary-statistics dataset, the LD reference genotypes, and the validation genotypes. If the validation cohort contains more than 1,000 individuals, with the same ancestry as the individuals used for the GWAS summary statistics, then we suggest using the validation cohort as the LD reference as well. These steps are implemented in the LDpred software package.

Simulations

We performed three types of simulations: (1) simulated traits and simulated genotypes; (2) simulated traits, simulated summary statistics, and simulated validation genotypes; and (3) simulated traits based on real genotypes. For most of the simulations, we used the point-normal model for effect sizes as described above:

$$\beta_i \sim_{iid} \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) \text{with probability } p \\ 0 \text{ with probability } (1-p). \end{cases}$$

For some of our simulations (Figure S5), we sampled the non-zero effects from a Laplace distribution instead of a Gaussian distribution. For all of our simulations, we used four different values for p (the fraction of causal loci). For some of our simulations (Figure S1), we sampled the fraction of causal markers within a region from a Beta(p, $1-p$) distribution. This simulates a genetic architecture where causal variants cluster in certain regions of the genome. We then obtained the simulated trait by summing up the allelic effects for each individual and adding a Gaussian-distributed noise term to fix the heritability. The simulated genotypes were sampled from a standard Gaussian distribution. To emulate LD, we simulated one genotype or SNP at a time to generate batches of 100 correlated SNPs. Each SNP was defined as the sum of the preceding adjacent SNP and some noise, where they were scaled to correspond to a fixed squared correlation between two adjacent SNPs within a batch. We simulated genotypes with the adjacent squared correlation between SNPs set to 0 (unlinked SNPs) and 0.9 (SNPs in LD).

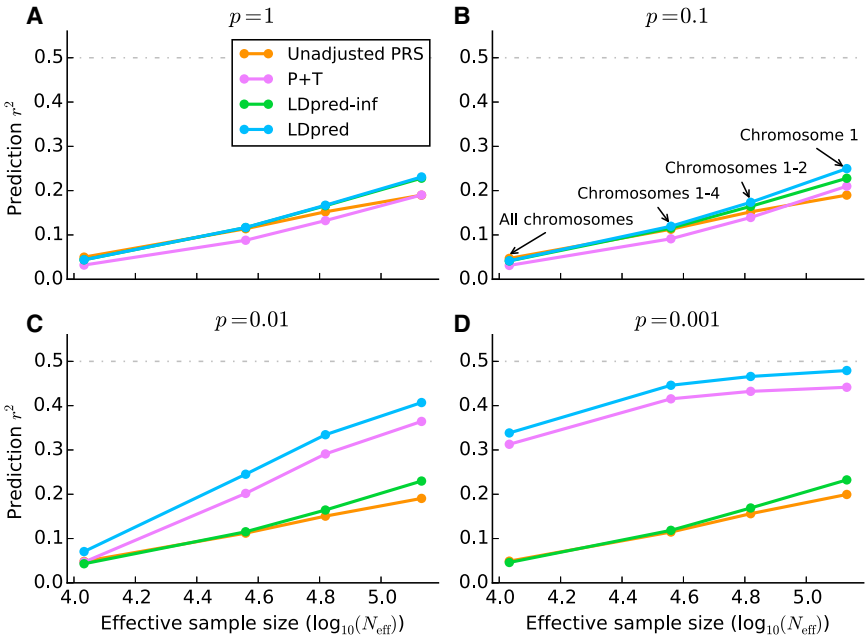

Figure 2. Comparison of Four Prediction Methods Applied to Simulated Traits

Prediction accuracy of the four different methods listed in Table S1 when applied to simulated traits with WTCCC genotypes. The four subfigures correspond to $p = 1$ (A), $p = 0.1$ (B), $p = 0.01$ (C), and $p = 0.001$ (D) for the fraction of simulated causal markers with (non-zero) effect sizes sampled from a Gaussian distribution. To aid interpretation of the results, we plot the accuracy against the effective sample size, defined as $N_{\text{eff}} = (N/M_{\text{sim}})M$, where $N = 10{,}786$ is the training sample size, $M = 376{,}901$ is the total number of SNPs, and M_{sim} is the actual number of SNPs used in each simulation: 376,901 (all chromosomes), 112,185 (chromosomes 1–4), 61,689 (chromosomes 1 and 2), and 30,004 (chromosome 1). The effective sample size is the sample size that maintains the same N/M ratio if all SNPs are used.

In order to compare the performance of our method at large sample sizes, we simulated summary statistics that we used as training data for the PRSs. We also simulated two smaller samples (2,000 individuals) representing independent validation data and a LD reference panel. When there is no LD, the least-squares effect estimates (summary statistics) are sampled from a Gaussian distribution, $\widehat{\beta}_i \mid \beta_i \sim_{iid} N(\beta_i, (1/N))$, where β_i are the true effects. To simulate marginal effect estimates without genotypes in the presence of LD, we first estimate the LD pattern empirically by simulating 100 linked SNPs for 1,000 individuals for a given value (as described above) and average over 1,000 simulations. This matrix captures the LD pattern in the validation data given that we simulate it by using the same procedure. Using this LD matrix D, we then sample the marginal least-squares estimates within a region of LD (SNP chunk) as $\widehat{\beta} \mid \beta \sim_{iid} N(D\beta, (D/N))$, where D is the LD matrix.

For the simulations in Figure 1B and Figures S1, S3, and S4, we simulated least-squares effect estimates for 200,000 variants in batches of LD regions with 100 variants each (as described above). We then simulated genotypes for 2,000 validation individuals and averaged over 100–3,000 simulated phenotypes to ensure smooth curves. Depending on the simulation parameters, the actual number of repeats required for achieving a smooth curve varied. For the simulations in Figure 1A and Figure S2, we simulated the least-squares estimates independently by adding an appropriately scaled Gaussian noise term to the true effects.

When simulating traits by using the Wellcome Trust Case Control Consortium (WTCCC) genotypes (Figure 2), we performed simulations under four different scenarios representing different number of chromosomes: (1) all chromosomes, (2) chromosomes 1–4, (3) chromosomes 1 and 2, and (4) chromosome 1. We used 16,179 individuals in the WTCCC data and 376,901 SNPs that passed quality control (QC). In our simulations, we used 3-fold cross-validation, whereby 1/3 of the data were validation data and 2/3 were training data.

WTCCC Genotype Data

We used the WTCCC genotypes[41] for both simulations and analysis. After performing QC, pruning variants with missing rates above 1%, and removing individuals with genetic relatedness coefficients above 0.05, we were left with 15,835 individuals genotyped for 376,901 SNPs, including 1,819 individuals with bipolar disease (BD), 1,862 individuals with coronary artery disease (CAD), 1,687 individuals with Crohn disease (CD), 1,907 individuals with hypertension (HT), 1,831 individuals with rheumatoid arthritis (RA), 1,953 individuals with type 1 diabetes (T1D), and 1,909 individuals with type 2 diabetes (T2D). For each of the seven diseases, we performed 5-fold cross-validation on affected individuals and 2,867 control individuals. For each of these analyses, we used the validation data as the LD reference data when using LDpred and when performing LD pruning.

Summary Statistics and Independent Validation Datasets

Six large summary-statistics datasets were analyzed in this study. The Psychiatric Genomics Consortium 2 (PGC2) SCZ summary statistics[14] consisted of 34,241 affected and 45,604 control individuals. For our purposes, we calculated GWAS summary statistics while excluding the ISC (International Schizophrenia Consortium) cohorts and the MGS (Molecular Genetics of Schizophrenia) cohorts. All subjects in these cohorts provided informed consent for this research, and procedures followed were in accordance with ethical standards. The summary statistics were calculated on a set of 1000 Genomes imputed SNPs, resulting in 16.9 million statistics. The two independent validation datasets, the ISC and MGS datasets, both consist of multiple cohorts with individuals of European descent. For both of the validation datasets, we used the chip genotypes and filtered individuals with more than 10% of genotype calls missing and filtered SNPs that had a missing rate more than 1% and a minor allele frequency (MAF) greater than 1%. In addition, we removed SNPs that had ambiguous nucleotides, i.e., A/T and G/C. We matched the SNPs between the validation and GWAS summary-statistics datasets on the basis of the SNP rsID and excluded triplets, SNPs for which one nucleotide was unknown, and SNPs that had different nucleotides in different datasets. This was our QC procedure for all large summary-statistics datasets that we analyzed. After QC, the ISC

cohort consisted of 1,562 affected and 1,994 control individuals genotyped on ~518,000 SNPs that overlapped with the GWAS summary statistics. The MGS dataset consisted of 2,681 affected and 2,653 control individuals after QC and had ~549,000 SNPs that overlapped with the GWAS summary statistics.

For multiple sclerosis (MS), we used the International Multiple Sclerosis Genetics Consortium summary statistics.[42] These were calculated with 9,772 affected and 17,376 control individuals (27,148 individuals in total) for ~465,000 SNPs. As an independent validation dataset, we used the BWH/MIGEN chip genotypes with 821 affected and 2,705 control individuals.[43] All subjects provided informed consent for this research, and procedures followed were in accordance with ethical standards. After QC, the overlap between the validation genotypes and the summary statistics only consisted of ~114,000 SNPs, which we used for our analysis.

For breast cancer (BC), we used the Genetic Associations and Mechanisms in Oncology (GAME-ON) BC GWAS summary statistics, consisting of 16,003 affected and 41,335 control individuals (both estrogen-receptor-negative [ER^-] and -positive [ER^+] individuals were included in this analysis).[44–47] These summary statistics were calculated for 2.6 million HapMap2 imputed SNPs. As validation genotypes, we combined genotypes from five different datasets: (1) ER^- and control individuals from the Breast and Prostate Cancer Cohort Consortium (BPC3),[44] (2) individuals from the Nurses' Health Study 2 (NHS2) breast cancer study (BrCa), (3) affected and control individuals from the Nurses' Health Study 1 (NHS1) mammographic density study, (4) NHS1 individuals from the Cancer Genetic Markers of Susceptibility (CGEMS) study,[48] and (5) control individuals from the NHS2 kidney stone study. All subjects in each cohort provided informed consent for this research, and procedures followed were in accordance with ethical standards. None of these 307 affected or 560 control individuals were included in the GWAS summary-statistics analysis, and they thus represent an independent validation dataset. We used the chip genotypes that overlapped the GWAS summary statistics, which resulted in ~444,000 genotypes after QC.

For CAD, we used the transatlantic Coronary Artery Disease Genome-wide Replication and Meta-analysis (CARDIoGRAM) consortium GWAS summary statistics. These were calculated with 22,233 affected and 64,762 control individuals (86,995 individuals in total) for 2.4 million SNPs.[10] For T2D, we used the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium GWAS summary statistics. These were calculated with 12,171 affected and 56,862 control individuals (69,033 individuals in total) for 2.5 million SNPs.[49] For both CAD and T2D, we used the Women's Genomes Health Study (WGHS) dataset as validation data,[50] where we randomly down-sampled the control individuals. For CAD, we validated the predictions in 923 individuals with cardiovascular disease and 1,428 control individuals, and for T2D we used 1,673 affected and 1,434 control individuals. We used the genotyped SNPs that overlapped the GWAS summary statistics, which amounted to about ~290,000 SNPs for both CAD and T2D after QC. All WGHS subjects provided informed consent for this research, and procedures followed were in accordance with ethical standards.

For height, we used the Genetic Investigation of Anthropometric Traits (GIANT) GWAS summary statistics as published in Lango Allen et al.[6] These were calculated with 133,653 individuals and imputed to 2.8 million HapMap2 SNPs. As a validation cohort, we used the Mount Sinai Medical Center BioMe cohort, which consists of 2,013 individuals and was genotyped at ~646,000 SNPs. All subjects provided informed consent for this research, and procedures followed were in accordance with ethical standards. After QC, the remaining ~539,000 SNPs that overlapped the GWAS summary statistics were used for the analysis.

For all six of these traits, we used the validation dataset as the LD reference data when using LDpred and when performing LD pruning. By using the validation dataset as LD reference data, we were only required to coordinate two different datasets, i.e., the GWAS summary statistics and the validation dataset. We calculated P+T risk scores for different p value thresholds by using grid values (1E−8, 1E−6, 1E−5, 3E−5, 1E−4, 3E−4, 1E−3, 3E−3, 0.01, 0.03, 0.1, 0.3, 1), and for LDpred we used the mixture probability (fraction of causal markers) values (1E−4, 3E−4, 1E−3, 3E−3, 0.01, 0.03, 0.1, 0.3, 1). We then reported the optimal prediction value from a validation dataset for LDpred and P+T.

SCZ Validation Datasets with Non-European Ancestry

For the non-European validation datasets, we used the MGS dataset as an LD reference, given that the summary statistics were obtained with individuals of European ancestry. This required us to coordinate across three different datasets: the GWAS summary statistics, the LD reference genotypes, and the validation genotypes. To ensure sufficient overlap of genetic variants across all three datasets, we used 1000 Genomes imputed MGS genotypes and the 1000 Genomes imputed validation genotypes for the three Asian validation datasets (JPN1, TCR1, and HOK2). To limit the number of markers for these datasets, we only considered markers that had a MAF > 0.1. After performing QC and removing variants with a MAF < 0.1, we were left with 1.38 million SNPs and 492 affected and 427 control individuals in the JPN1 dataset, 1.88 million SNPs and 898 affected and 973 control individuals in the TCR1 dataset, and 1.71 million SNPs and 476 affected and 2,018 control individuals in the HOK2 dataset.

For the African-American (AFAM) validation dataset, we used the reported GWAS summary-statistics dataset[14] to train on. The AFAM dataset consisted of 3,361 SCZ-affected and 5,076 control individuals. Because the AFAM dataset was not included in that analysis, this allowed us to leverage a larger sample size, but at the cost of having fewer SNPs. The overlap among the 1000 Genomes imputed MGS genotypes, the HapMap 3 imputed AFAM genotypes, and the PGC2 reported summary statistics included ~482,000 SNPs (with a MAF > 0.01) after QC. All subjects in the JPN1, TCR1, HOK2, and AFAM datasets provided informed consent for this research, and procedures followed were in accordance with ethical standards.

Prediction-Accuracy Metrics

For quantitative traits, we used squared correlation (R^2). For case-control traits, which include all of the disease datasets analyzed, we used four different metrics. We used Nagelkerke R^2 as our primary figure of merit in order to be consistent with previous work,[1,9,12,14] but we also report three other commonly used metrics in Tables S2, S5, S7, and S10: observed-scale R^2, liability-scale R^2, and the area under the curve (AUC). All of the reported prediction R^2 values were adjusted for the top five principal components (PCs) in the validation sample (top three PCs for BC). The relationship among the observed-scale R^2, liability-scale R^2, and AUC is described in Lee et al.[51] We note that Nagelkerke R^2 is similar to the observed-scale R^2 (i.e., is also affected by case-control ascertainment) but generally has slightly larger values.

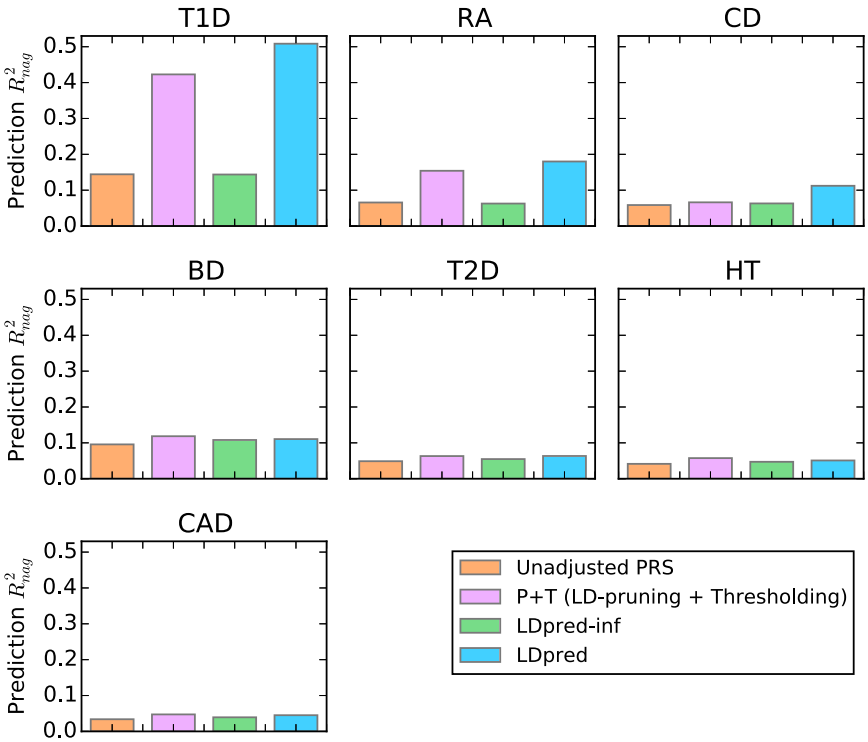

Figure 3. Comparison of Methods Applied to Seven WTCCC Disease Datasets
The prediction accuracy of different methods as estimated from 5-fold cross-validation in seven WTCCC disease datasets: type 1 diabetes (T1D), rheumatoid arthritis (RA), Crohn disease (CD), bipolar disease (BD), type 2 diabetes (T2D), hypertension (HT), and coronary artery disease (CAD). The Nagelkerke prediction R^2 is shown on the y axis (see Table S2 for other metrics). LDpred significantly improved the prediction accuracy for the immune-related diseases T1D, RA, and CD (see main text).

Results

Simulations

We first considered simulations with simulated genotypes (see Material and Methods). We assessed accuracy by using squared correlation (prediction R^2) between observed and predicted phenotypes. The Bayesian shrink imposed by LDpred generally performed well in simulations without LD (Figure S3); in this case, posterior mean effect sizes can be obtained analytically (see Material and Methods). However, LDpred performed particularly well in simulations with LD (Figure S4); the larger improvement (e.g., versus P+T) in this case indicates that the main advantage of LDpred is in its explicit modeling of LD. Simulations under a Laplace mixture distribution prior gave similar results (see Figure S5). We also evaluated the prediction accuracy as a function of the sample size of the LD reference panel (Figure S6). LDpred performs best with an LD reference panel of at least 1,000 individuals. These results also highlight the importance of using an LD reference population with LD patterns similar to the training sample, given that an inaccurate reference sample will have effects similar to those of a small reference sample. Below, we focus on simulations with real WTCCC genotypes, which have more realistic LD properties.

Using real WTCCC genotypes[41] (15,835 samples and 376,901 markers after QC), we simulated infinitesimal traits with the heritability set to 0.5 (see Material and Methods). We extrapolated results for larger sample sizes (N_{eff}) by restricting the simulations to a subset of the genome (smaller M), leading to larger N/M. Results are displayed in Figure 2A. LDpred-inf and LDpred (which are expected to be equivalent in the infinitesimal case) performed well in these simulations—particularly at large values of N_{eff}, consistent with the intuition from Equation 1 that the LD adjustment arising from the reference-panel LD matrix (D) is more important when Nh^2_g/M is large. On the other hand, P+T performed less well, consistent with the intuition that pruning markers loses information.

We next simulated non-infinitesimal traits by using real WTCCC genotypes and varying the proportion p of causal markers (see Material and Methods). Results are displayed in Figures 2B–2D. LDpred outperformed all other approaches, including P+T, particularly at large values of N/M. For $p = 0.01$ and $p = 0.001$, the methods that do not account for non-infinitesimal architectures (unadjusted PRSs and LDpred-inf) performed poorly, and P+T was second best among these methods. Comparisons to additional methods are provided in Figure S7; in particular, LDpred outperformed other recently proposed approaches that use LD from a reference panel[13,52] (see Appendix B).

Besides accuracy (prediction R^2), another measure of interest is calibration. A predictor is correctly calibrated if a regression of the true phenotype versus the predictor yields a slope of 1 and is miscalibrated otherwise; calibration is particularly important for risk prediction in clinical settings. In general, unadjusted PRSs and P+T yield poorly calibrated risk scores. On the other hand, the Bayesian approach provides correctly calibrated predictions (if the prior accurately models the true genetic architecture and the LD is appropriately accounted for), avoiding the need for re-calibration at the validation stage. The calibration slopes for the simulations using WTCCC genotypes are given in Figure S8. As expected, LDpred provides much better calibration than other approaches.

Application to WTCCC Disease Datasets

We compared LDpred to other summary-statistics-based methods across the seven WTCCC disease datasets[41] by using 5-fold cross-validation (see Material and Methods). Results are displayed in Figure 3. (We used Nagelkerke R^2

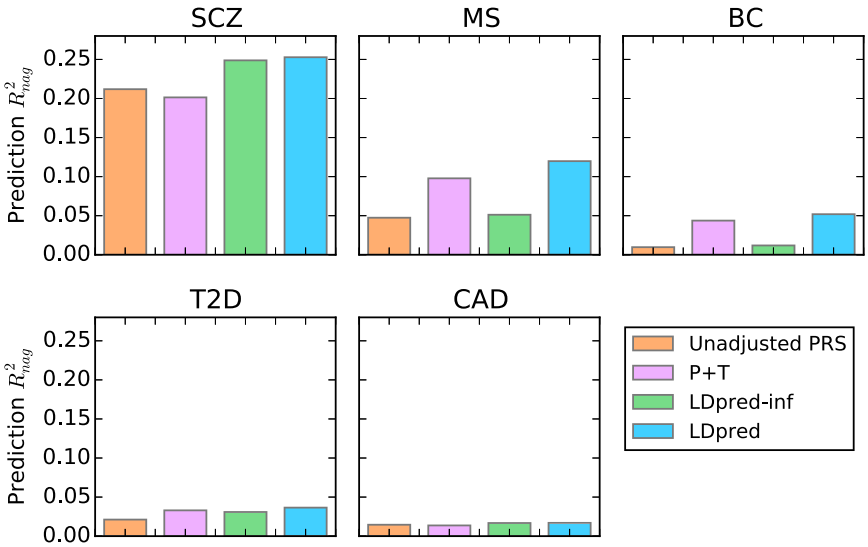

Figure 4. Comparison of Methods Training on Large GWAS Summary Statistics for Five Different Diseases

The prediction accuracy is shown for five different diseases: schizophrenia (SCZ), multiple sclerosis (MS), breast cancer (BC), type 2 diabetes (T2D), and coronary artery disease (CAD). The risk scores were trained with large GWAS summary-statistics datasets and used for predicting disease risk in independent validation datasets. The Nagelkerke prediction R^2 is shown on the y axis (see Table S5 for other metrics). Compared to LD pruning + thresholding (P+T), LDpred improved the prediction R^2 by 11%–25%. SCZ results are shown for the SCZ-MGS validation cohort used in recent studies,[9,12,14] but LDpred also produced a large improvement for the independent SCZ-ISC validation cohort (Table S5).

as our primary figure of merit in order to be consistent with previous work,[1,9,12,14] but we also provide results for observed-scale R^2, liability-scale R^2,[51] and AUC[53] in Table S2; the relationship between these metrics is discussed in the Material and Methods.)

LDpred attained significant improvement in prediction accuracy over P+T for T1D (p value = 4.4E−15), RA (p value = 1.2E−5), and CD (p value = 2.7E−8), similar to previous results from BSLMM[26], BayesR,[28] and MultiBLUP[27] on the same data. For these three immune-related disorders, the major histocompatibility complex (MHC) region explains a large amount of the overall variance, such that it represents an extreme special case of a non-infinitesimal genetic architecture. We note that LDpred, BSLMM, and BayesR all explicitly model non-infinitesimal architectures; however, unlike LDpred, BSLMM and BayesR require full genotype data and cannot be applied to large summary-statistics datasets (see below). MultiBLUP, which also requires full genotype data, assumes an infinitesimal prior that varies across regions and thus benefits from a different modeling extension; the possibility of extending multiBLUP to work with summary statistics is a direction for future research. For the other diseases with more-complex genetic architectures, the prediction accuracy of LDpred was similar to that of P+T, potentially because the training sample size was not sufficiently large enough for modeling LD to have a sizeable impact. The inferred heritability parameters and optimal p parameters for LDpred, as well as the optimal thresholding parameters for P+T, are provided in Table S3. The calibration of the predictions for the different approaches is shown in Table S4. Consistent with our simulations, LDpred provides much better calibration than other approaches.

Application to Six Large Summary-Statistics Datasets

We applied LDpred to five diseases—SCZ, MS, BC, T2D, and CAD—for which we had GWAS summary statistics for large sample sizes (ranging from ~27,000 to ~86,000 individuals) and raw genotypes for an independent validation dataset (see Material and Methods). Prediction accuracies for LDpred and other methods are reported in Figure 4 (Nagelkerke R^2) and Table S5 (other metrics). We also applied LDpred to height (a quantitative trait), for which we had GWAS summary statistics calculated with ~134,000 individuals[6] and an independent validation dataset. The height-prediction accuracy for LDpred and other methods is reported in Table S6.

For all six traits, LDpred provided significantly better predictions than other approaches (for the improvement over P+T, the p values were 6.3E−47 for SCZ, 2.0E−14 for MS, 0.020 for BC, 0.004 for T2D, 0.017 for CAD, and 1.5E−10 for height). The relative increase in Nagelkerke R^2 over other approaches ranged from 11% for T2D to 25% for SCZ, and we observed a 30% increase in prediction R^2 for height. This is consistent with the fact that our simulations showed larger improvements for highly polygenic traits, such as SCZ[14] and height.[54] We note that for both CAD and T2D, the accuracy attained with >60,000 training samples from large meta-analyses (Figure 4) is actually lower than the accuracy attained with <5,000 training samples from the WTCCC (Figure 3). This result is independent of the prediction method applied and demonstrates the challenges of potential heterogeneity in large meta-analyses (although prediction results based on cross-validation in a single cohort should be viewed with caution[19]). To examine this further, we trained CAD and T2D PRSs on the WTCCC data, validated them in the WGHS data, and determined that the prediction accuracy in external WGHS validation data is substantially smaller than within the WTCCC dataset (Table S7). Possible explanations for this discrepancy include differences in sample ascertainment in the WGHS and WTCCC datasets or unadjusted data artifacts in the WTCCC training and validation data.

Parameters inferred by LDpred and other methods are provided in Table S8, and calibration results are provided

in Table S9; again, LDpred attained the best calibration. Finally, we applied LDpred to predict SCZ risk in non-European validation samples of both African and Asian descent (see Material and Methods). Although prediction accuracies were lower in absolute terms, we observed similar relative improvements for LDpred over other methods (Tables S10 and S11).

Discussion

PRSs are likely to become clinically useful as GWAS sample sizes continue to grow.[15,18] However, unless LD is appropriately modeled, their predictive accuracy will fall short of their maximal potential. Our results show that LDpred is able to address this problem—even when only summary statistics are available—by estimating posterior mean effect sizes by using a point-normal prior and LD information from a reference panel. Intuitively, there are two reasons for the relative gain in prediction accuracy of LDpred PRSs over P+T. First, LD pruning discards informative markers and thereby limits the overall heritability explained by the markers. Second, LDpred accounts for the effects of linked markers, which can otherwise lead to biased estimates. These limitations hinder P+T regardless of the LD pruning and p value thresholds used.

Although we focus here on methods that only require summary statistics, we note the parallel advances that have been made in methods that require raw genotypes[22,24–29,55,56] as training data. Some of those methods employ a variational Bayes (iterative conditional expectation) approach to reduce their running time[24,25,29,55] (and report that results are similar to those of MCMC[29]), but we found that MCMC generally obtains more robust results than variational Bayes in the analysis of summary statistics, perhaps because the LD information is only approximate. Our use of a point-normal mixture prior is consistent with some of those studies,[25] although different priors, e.g., a mixture of normal, were used by other studies.[23,26,28] One recent study proposed an elegant approach for handling case-control ascertainment while including genome-wide-significant associations as fixed effects;[56] however, the correlations between distal causal SNPs induced by case-control ascertainment do not affect summary statistics from marginal analyses, and explicit modeling of non-infinitesimal effect-size distributions will appropriately avoid shrinking genome-wide-significant associations (Figure S2).

Although LDpred is a substantial improvement over existing methods for using summary statistics to conduct polygenic prediction, it still has limitations. First, the method's reliance on LD information from a reference panel requires that the reference panel be a good match for the population from which summary statistics were obtained; in the case of a mismatch, prediction accuracy might be compromised. One potential solution is the broad sharing of summary LD statistics, which has previously been advocated in other settings.[57] If LDpred uses the true LD pattern from the training sample, and there is no unaccounted long-range LD, then we expect little or no gain in prediction accuracy with individual-level genotype information. Second, the point-normal mixture prior distribution used by LDpred might not accurately model the true genetic architecture, and it is possible that other prior distributions might perform better in some settings. Third, in those instances where raw genotypes are available, fitting all markers simultaneously (if computationally tractable) might achieve higher accuracy than methods based on marginal summary statistics. Fourth, as with other prediction methods, heterogeneity across cohorts might hinder prediction accuracy; our results suggest that this could be a major concern in some datasets. Fifth, we assume that summary statistics have been appropriately corrected for genetic ancestry, but if this is not the case, then the prediction accuracy might be misinterpreted[19] or might even decrease.[58] Sixth, our analyses have focused on common variants; LD reference panels are likely to be inadequate for rare variants, motivating future work on how to treat rare variants in PRSs. Despite these limitations, LDpred is likely to be broadly useful in leveraging summary-statistics datasets for polygenic prediction of both quantitative and case-control traits.

As sample sizes increase and polygenic predictions become more accurate, their value increases, both in clinical settings and for understanding genetics. LDpred represents substantial progress, but more work remains to be done. One future direction would be to develop methods that combine different sources of information. For example, as demonstrated by Maier et al.,[59] joint analysis of multiple traits can increase prediction accuracy. In addition, using different prior distributions across genomic regions[27] or functional annotation classes[60] could further improve the prediction. Finally, although LDpred attains a similar relative improvement when using non-European samples as validation samples, the lower absolute accuracy than in European samples motivates further efforts to improve prediction in diverse populations.

Appendix A: Estimating the Posterior Mean Phenotype

Under the assumption that the phenotype has an additive genetic architecture and is linear, then estimating the posterior mean phenotype boils down to estimating the posterior mean effects of each SNP and then summing their contribution in a risk score.

Posterior Mean Effects Assuming Unlinked Markers and an Infinitesimal Model

We will first consider the infinitesimal model, which represents a genetic architecture where all genetic variants are causal. The classic example is Fisher's infinitesimal

model,[37] which assumes that genotypes are unlinked and that effect sizes have a Gaussian distribution (after normalizing by allele frequency).

Assume that β_i are independently drawn from a Gaussian distribution $\beta_i \sim N(0,(h_2/M))$, where M denotes the total number of causal effects (β_i). Then, we can derive a posterior mean given the marginal ordinary least-squares estimate $\tilde{\beta}_i = (X_i Y)/N$. The least-squares estimate is approximately distributed as

$$\tilde{\beta}_i \sim N\left(\beta_i, \frac{1-\frac{h^2}{M}}{N}\right),$$

where N is the number of individuals. The variance can be approximated further, $\mathrm{Var}(\beta_i) \approx 1$, when M is large. With this variance, the posterior distribution for β_i is

$$\beta_i \mid \tilde{\beta}_i \sim N\left(\left(\frac{1}{1+\frac{M}{h^2N}}\right)\tilde{\beta}_i,\ \frac{1}{N}\left(\frac{1}{1+\frac{M}{h^2N}}\right)\right).$$

This suggests that a uniform Bayesian shrink by a factor of

$$\frac{1}{1+\frac{M}{h^2N}}$$

is appropriate under Fisher's infinitesimal model.

Other possible choices of prior distributions for the effects include Laplace distributions. However, calculating the posterior mean under this model is non-trivial but can be solved numerically.[61] Alternatively, the posterior mode has a simple analytical form.[62] The posterior mode under a Laplace prior is in fact the LASSO estimate.[63] If we assume that the sum of the effects has variance h^2 and that the genetic markers are uncorrelated, then the posterior mode estimate is

$$\dot{\beta}_i = \mathrm{sign}(\beta_i)\max\left(0, |\beta_i| - \sqrt{\frac{h^2}{2M}}\right).$$

Interestingly, the posterior mode effects for estimated effects below a given threshold are set to 0, even though all betas are causal in the model.

Posterior Mean Effects Assuming Unlinked Markers and a Non-infinitesimal Model

Most diseases and traits are not likely to be strictly infinitesimal, i.e., follow Fisher's infinitesimal model.[37] Instead, a non-infinitesimal model, where only a fraction of the genetic variants are truly causal and affect the trait, is more likely to describe the underlying genetic architecture. We can model non-infinitesimal genetic architectures by using mixture distributions with a mixture parameter p, which denotes the fraction of causal markers. More specifically, we will consider a spike-and-slab prior with a 0 spike and a Gaussian slab (see Figure S9).

Assume that the effects are drawn from a mixture distribution as follows:

$$\beta_i \sim \begin{cases} N\left(0, \frac{h^2}{Mp}\right) \text{with probability } p \\ 0 \text{ with probability } (1-p) \end{cases}.$$

Another way of writing this is to use Dirac's delta function, i.e., write $\beta_i = pu + (1-p)v$, where $u \sim (0,(h^2/Mp))$ and $v \sim \delta_{\beta_i}$. Here, δ_{β_i} denotes the point density at $\beta_i = 0$, which integrates to 1. We can then write out the density for $\tilde{\beta}_i$ as follows:

$$\begin{aligned} f(\tilde{\beta}_i) &= \int_{-\infty}^{\infty} f(\tilde{\beta}_i \mid \beta_i) f(\beta_i) d\beta_i \\ &= \frac{p}{2\pi}\left(\sqrt{\frac{NMp}{h^2}}\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(N(\tilde{\beta}_i-\beta_i)^2 + \frac{Mp}{h^2}\beta_i^2\right)\right\} d\beta_i\right) + (1-p)\left(\sqrt{\frac{N}{2\pi}}\exp\left\{-\frac{1}{2}N\tilde{\beta}_i^2\right\}\right) \\ &= \frac{1}{\sqrt{2\pi}}\left(\frac{p}{\sqrt{\frac{h^2}{Mp}+\frac{1}{N}}}\exp\left\{-\frac{1}{2}\left(\frac{\tilde{\beta}_i^2}{\frac{h^2}{Mp}+\frac{1}{N}}\right)\right\}\right) + \frac{1-p}{\frac{1}{\sqrt{N}}}\exp\left\{-\frac{1}{2}N\tilde{\beta}_i^2\right\}. \end{aligned}$$

We are interested in the posterior mean, which can be expressed as

$$E(\beta_i \mid \tilde{\beta}_i) = \int_{-\infty}^{\infty} \frac{\beta_i f(\tilde{\beta}_i \mid \beta_i) f(\beta_i)}{\int_{-\infty}^{\infty} f(\tilde{\beta}_i \mid \beta_i) f(\beta_i) d\beta_i} d\beta_i .$$

Hence, we only need to calculate the following definite integral:

$$\int_{-\infty}^{\infty} \beta_i f(\tilde{\beta}_i \mid \beta_i) f(\beta_i) d\beta_i = \frac{p}{2\pi}\sqrt{\frac{NMp}{h^2}}\int_{-\infty}^{\infty} \beta_i \times \exp\left\{-\frac{1}{2}\left(N(\tilde{\beta}_i-\beta_i)^2 + \frac{Mp}{h^2}\beta_i^2\right)\right\} d\beta_i.$$

Thus, the posterior mean is

$$\begin{aligned} E(\beta_i \mid \tilde{\beta}_i) &= C\int_{-\infty}^{\infty} \beta_i \exp\left\{-\frac{1}{2}\left(N(\beta_i^2 - 2\beta_i\tilde{\beta}_i) + \frac{Mp}{h^2}\beta_i^2\right)\right\} d\beta_i \\ &= C\sqrt{\frac{2\pi}{N}}\left(\frac{1}{1+\frac{Mp}{Nh^2}}\right)^{3/2} \exp\left\{\frac{N}{2}\left(\frac{1}{1+\frac{Mp}{Nh^2}}\right)\tilde{\beta}_i^2\right\}\tilde{\beta}_i, \end{aligned}$$

where

$$C = \frac{\frac{p}{\sqrt{2\pi}}\sqrt{\frac{NMp}{h^2}}\exp\left\{-\frac{1}{2}N\tilde{\beta}_i^2\right\}}{\frac{p}{\sqrt{\frac{h^2}{Mp}+\frac{1}{N}}}\exp\left\{-\frac{1}{2}\left(\frac{\tilde{\beta}_i^2}{\frac{h^2}{Mp}+\frac{1}{N}}\right)\right\}+\frac{1-p}{\frac{1}{\sqrt{N}}}\exp\left\{-\frac{1}{2}N\tilde{\beta}_i^2\right\}}.$$

Alternatively, by realizing that the posterior probability that β_i is sampled from the Gaussian distribution given $\tilde{\beta}_i$ is exactly

$$P(\beta_i \sim N(\bullet,\bullet)\,|\,\tilde{\beta}_i) = \frac{f(\tilde{\beta}_i \mid \beta_i \sim N(\bullet,\bullet))f(\beta_i \sim N(\bullet,\bullet))}{f(\tilde{\beta}_i)} = \frac{\frac{p}{\sqrt{\frac{h^2}{Mp}+\frac{1}{N}}}\exp\left\{-\frac{1}{2}\left(\frac{\tilde{\beta}_i^2}{\frac{h^2}{Mp}+\frac{1}{N}}\right)\right\}}{\frac{p}{\sqrt{\frac{h^2}{Mp}+\frac{1}{N}}}\exp\left\{-\frac{1}{2}\left(\frac{\tilde{\beta}_i^2}{\frac{h^2}{Mp}+\frac{1}{N}}\right)\right\}+\frac{1-p}{\frac{1}{\sqrt{N}}}\exp\left\{-\frac{1}{2}N\tilde{\beta}_i^2\right\}}, \quad \text{(Equation A1)}$$

we can rewrite the posterior mean in a simpler fashion. If we let $\overline{p}_i = P(\beta_i \sim N(\bullet,\bullet)|\tilde{\beta}_i)$ denote the posterior probability that β_i is non-zero or Gaussian distributed (Equation A1), then it becomes

$$E(\beta_i \mid \tilde{\beta}_i) = \left(\frac{1}{1+\frac{Mp}{h^2N}}\right)\overline{p}_i\tilde{\beta}_i .$$

Posterior Mean Effects Assuming Linked Markers and an Infinitesimal Model

Following Yang et al.,[52] we can obtain the joint least-squares effect estimates as

$$\widehat{\beta}_{\text{joint}} = D^{-1}\tilde{\beta}_{\text{marg}},$$

where $D = XX'/N$ is the LD correlation matrix, and $\tilde{\beta}$ denotes the vector of marginal least-squares effects (which is approximately equal to the joint least-squares estimate if SNPs are unlinked). In practice, the LD matrix is $M \times M$ and possibly singular, e.g., if two (or more) markers are in perfect linkage. If the LD matrix D is singular, the joint least-squares estimate does not have a unique solution. However, if the individuals in the training data do not display family or population structure, the genome-wide LD matrix is approximately a banded matrix, which allows local adjustment for LD instead. To formalize these ideas, we introduce some notation. Let l_i denote the i^{th} locus or region with M_{l_i} markers. In addition, let $\beta^{(i)}$ denote the vector of true effects that are in the i^{th} region, and similarly let $\tilde{\beta}^{(i)}$ denote the corresponding marginal least-squares estimates in the region. Under this model, we can derive the sampling distribution for effect estimates at the i^{th} region, i.e., $\tilde{\beta}^{(i)} \,|\, \beta^{(i)}$. The mean is $E(\tilde{\beta}^{(i)} \,|\, \beta^{(i)}) = D^{(i)}\beta^{(i)}$, where $D^{(i)} = X^{(i)}X^{(i)\prime}/N$ is the LD matrix obtained from the markers in the i^{th} region, i.e., $X^{(i)}$. Furthermore, the conditional covariance matrix is

$$\begin{aligned}
\text{Var}(\tilde{\beta}^{(i)} \mid \beta^{(i)}) &= E(\tilde{\beta}^{(i)\prime}\tilde{\beta}^{(i)} \mid \beta^{(i)}) - E(\tilde{\beta}^{(i)} \mid \beta^{(i)})E(\tilde{\beta}^{(i)} \mid \beta^{(i)})' \\
&= \frac{1}{N^2}E\left(X^{(i)}\left(X^{(i)\prime}\,\beta^{(i)} + \varepsilon\right)\left(X^{(i)}\left(X^{(i)\prime}\,\beta^{(i)} + \varepsilon\right)\right)' \mid \beta^{(i)}\right) \\
&\quad -\left(D^{(i)}\beta^{(i)}\right)\left(D^{(i)}\beta^{(i)}\right)' \\
&= \left(D^{(i)}\beta^{(i)}\right)\left(D^{(i)}\beta^{(i)}\right)'\frac{1}{N}E\left(X^{(i)}\varepsilon\left(X^{(i)}\varepsilon\right)' \mid \beta^{(i)}\right) \\
&\quad -\left(D^{(i)}\beta^{(i)}\right)\left(D^{(i)}\beta^{(i)}\right)' \\
&= X^{(i)}\frac{1}{N^2}E\left(\varepsilon\varepsilon' \mid \beta^{(i)}\right)\left(X^{(i)}\right)' = \frac{1-h_{l_i}^2}{N^2}X^{(i)}\left(X^{(i)}\right)' \\
&= \frac{1-h_{l_i}^2}{N}D^{(i)} ,
\end{aligned}$$

where $h_{l_i}^2$ denotes the heritability explained by the markers in the region, i.e., $X^{(i)}$. If we assume that the heritability explained by an individual region is small, then this simplifies to $\text{Var}(\tilde{\beta}^{(i)} \,|\, \beta^{(i)}) = D^{(i)}/N$. This equation is

particularly useful for performing efficient simulations of effect sizes without simulating the genotypes. Given an LD matrix, D, we can simulate effect sizes and corresponding least-squares estimates. Similarly, for the joint estimate, we have

$$E\left(\widehat{\beta}^{(i)}_{\text{joint}} \mid \beta^{(i)}\right) = \beta^{(i)}$$

and

$$\text{Var}\left(\widehat{\beta}^{(i)}_{\text{joint}} \mid \beta^{(i)}\right) = \frac{1-h^2_{l_i}}{N}\left(D^{(i)}\right)^{-1}.$$

In the following, we let β (and $\tilde{\beta}$) denote the effects within a region of LD. We furthermore assume that these markers only explain a fraction, h^2_l, of the total phenotypic variance, and $h^2_l \leq h^2$. Given a Gaussian prior distribution $\beta \sim N(0,\ (h^2/M))$ for the effects and the conditional distribution $\tilde{\beta}\,|\,\beta$, we can derive the posterior mean by considering the joint density:

$$f(\tilde{\beta},\beta) = \frac{1}{\sqrt{|D|}}\left(\frac{N}{2\pi\left(1-h^2_l\right)}\right)^{\frac{M}{2}} \times \exp\left\{\frac{N(\tilde{\beta}-D\beta)'D^{-1}(\tilde{\beta}-D\beta)}{2\left(1-h^2_l\right)}\right\}\left(\frac{Mp}{2\pi h^2}\right)^{-\frac{M}{2}} \times \exp\left\{\frac{M}{2h^2}\beta'\beta\right\}.$$

We can now obtain the posterior density for $\tilde{\beta}\,|\,\beta$ by completing the square in the exponential. This yields a multivariate Gaussian with mean and variance as follows:

$$E(\beta \mid \tilde{\beta}) = \left(\frac{1}{1-h^2_l}D + \frac{M}{Nh^2}I\right)^{-1}\tilde{\beta}\ ,$$

$$\text{Var}(\beta \mid \tilde{\beta}) = \frac{1}{N}\left(\frac{1}{1-h^2_l}D + \frac{M}{Nh^2}I\right)^{-1},$$

where h^2 denotes the heritability explained by the M causal variants, and $h^2_l \approx kh^2/M$ is the heritability of the k effects or variants in the region of LD. If $M \gg k$, then $1-h^2_l$ becomes approximately 1, and the equations above can be simplified accordingly. As expected, the posterior mean approaches the maximum-likelihood estimator as the training sample size grows.

Posterior Mean Effects Assuming Linked Markers and a Non-infinitesimal Model

The Bayesian shrink under the infinitesimal model implies that we can solve it either by using a Gauss-Seidel method[64,65] or via MCMC Gibbs sampling. The Gauss-Seidel method iterates over the markers and obtains a residual effect estimate after subtracting the effect of neighboring markers in LD. It then applies a univariate Bayesian shrink, i.e., the Bayesian shrink for unlinked markers (described above). It then iterates over the genome multiple times until convergence is achieved. However, we found the Gauss-Seidel approach to be sensitive to model assumptions, i.e., if the LD matrix used differed from the true LD matrix in the training data, we observed convergence issues. We therefore decided to use an approximate MCMC Gibbs sampler instead to infer the posterior mean. The approximate Gibbs sampler used by LDpred is similar to the Gauss-Seidel approach, except that instead of using the posterior mean to update the effect size, we sample the update from the posterior distribution. Compared to the Gauss-Seidel method, this seems to lead to less serious convergence issues. Below, we describe the Gibbs sampler used by LDpred.

Define q as follows:

$$q \sim \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } (1-p) \end{cases}.$$

Then, we can write $\beta = qu$, where $u \sim N(0,(h^2/Mp)I)$. Hence, we can write the multivariate density for β as

$$f(\beta) = \prod_{i=1}^{M}\left(p\sqrt{\frac{Mp}{2\pi h^2}}\exp\left\{-\frac{Mp}{2h^2}\beta_i^2\right\} + (1-p)\delta_{\beta_i}\right).$$

The sampling distribution for $\tilde{\beta}$ given β is

$$f(\tilde{\beta} \mid \beta) = \frac{1}{\sqrt{|D|}}\left(\frac{N}{2\pi\left(1-h^2_l\right)}\right)^{\frac{M}{2}} \times \exp\left\{\frac{N(\tilde{\beta}-D\beta)'D^{-1}(\tilde{\beta}-D\beta)}{2\left(1-h^2_l\right)}\right\}. \quad \text{(Equation A2)}$$

As usual, we want to calculate the posterior mean, i.e.,

$$E(\beta \mid \tilde{\beta}) = \int \frac{\beta_i f(\tilde{\beta} \mid \beta) f(\beta)}{\int f(\tilde{\beta} \mid \beta) f(\beta) d\beta} d\beta,$$

which now consists of two M-dimensional integrands. Any multiplicative term that does not involve β in the two integrands factors out. Because the integrand consists of 2^M nontrivial additive terms, we result to numerical approximations to sample from the posterior and estimate the posterior mean effects.

An alternative approach to obtaining the posterior mean is to sample from the posterior distribution and then average over the samples to obtain the posterior mean. In our case, we know the posterior up to a constant, i.e.,

$$f(\beta \mid \tilde{\beta}) \propto f(\beta,\tilde{\beta}) = f(\tilde{\beta} \mid \beta) f(\beta_i \mid \beta_{-i}) f(\beta_{-i}),$$

where β_{-i} denotes all the other effects except for the effect of the i^{th} marker. Note that $(\beta_i\,|\,\beta_{-i})f(\beta_{-i}) = f(\beta)$. We can use this fact to sample efficiently in a MCMC setting, where we sample one marker effect at a time in an iterative

fashion (the conditional proposal distribution is therefore univariate).

A Gibbs sampler is an efficient MCMC that can be used whenever the marginal conditional posterior distributions can be derived. For our purposes, these are the conditional posterior distributions of the effects, i.e., $f(\beta \mid \tilde{\beta}, \beta_{-i})$, where β_{-i} refers to the vector of betas excluding the i^{th} beta. We can write the posterior distribution as follows:

$$f(\beta \mid \tilde{\beta}, \beta_{-i}) = \frac{f(\tilde{\beta}, \beta)}{f(\tilde{\beta}, \beta_{-i})} = \frac{f(\tilde{\beta} \mid \beta) f(\beta)}{f(\tilde{\beta} \mid \beta_{-i}) f(\beta_{-i})} = \frac{f(\tilde{\beta} \mid \beta) f(\beta_i)}{f(\tilde{\beta} \mid \beta_{-i})}$$
$$= \frac{f(\tilde{\beta} \mid \beta) f(\beta_i)}{\int f(\tilde{\beta} \mid \beta) f(\beta_i) d\beta_i} .$$

Sampling from this distribution is not trivial. However, we can partition the sampling procedure into two parts, such that we first sample whether the effect is different from 0 and then if it is different from 0, we can assume it has a Gaussian prior. To achieve this, we first need to calculate the posterior probability that a marker is causal, i.e.,

$$P(\beta_i = 0 \mid \tilde{\beta}, \beta_{-i}) = \frac{P(\beta_i = 0, \tilde{\beta}, \beta_{-i})}{P(\tilde{\beta}, \beta_{-i})}$$
$$= \frac{P(\beta_i = 0, \tilde{\beta} \mid \beta_{-i})}{P(\beta_i = 0, \tilde{\beta} \mid \beta_{-i}) + \int_{\beta_i \neq 0} f(\tilde{\beta} \mid \beta) f(\beta_i) d\beta_i} .$$

Obtaining an analytical solution to this is non-trivial; however, if we assume that $P(\beta_i = 0 \mid \tilde{\beta}, \beta_{-i}) \approx P(\beta_i = 0 \mid \tilde{\beta}_i, \beta_{-i})$, then we can simply extract the effects of LD from other effects on the effect estimate $\tilde{\beta}_i$ and then use the marginal posterior probability that the marker is causal from Equation A1 instead, i.e., $P(\beta_i = 0 \mid \tilde{\beta}_i, \beta_{-i}) \approx \overline{p}_i$. If we sample the effect to be non-zero and again make the simplifying assumption that $f(\beta_i \mid \tilde{\beta}, \beta_{-i}) \approx f(\beta_i \mid \tilde{\beta}_i, \beta_{-i})$, then we can write out its posterior distribution, extract the effects of LD on the effect estimate, and sample from the marginal (without LD) posterior distribution derived above. More specifically, the marginal posterior distribution for β_i becomes

$$f(\beta_i \mid \tilde{\beta}, \beta_{-i}) \approx f(\beta_i \mid \tilde{\beta}_i, \beta_{-i}) = (1 - \overline{p}_i) \delta_{\beta_i} + \overline{p}_i h(\beta_i),$$

where $h(\beta_i)$ is the Gaussian density for the posterior distribution conditional on $\beta_i \neq 0$, i.e.,

$$\beta_i \mid \tilde{\beta}_i, \beta_{-i}, \beta_i \neq 0 \sim N\left(\left(\frac{1}{1 + \frac{M}{h^2 N}} \right) \tilde{\beta}_i, \ \frac{1}{N} \left(\frac{1}{1 + \frac{M}{h^2 N}} \right) \right).$$

Other Considerations for LDpred

Although LDpred aims to estimate the posterior mean phenotype (the best unbiased prediction), it is only guaranteed to do so if all the assumptions hold. Because LDpred relies on a few assumptions (both regarding LD and mathematical approximations), it is an approximate Gibbs sampler, which can lead to robustness issues. Indeed, we found LDpred to be sensitive to inaccurate LD estimates, especially for very large sample sizes. To address this, we set the probability of setting the effect size to 0 in the Markov chain to be at least 5%. This improved the robustness of LDpred as observed in both simulated and real data. If convergence issues arise when LDpred is applied to data, then it might be worthwhile to explore higher values for the 0-jump probability.

Finally, throughout the above derivation of LDpred, we assumed that the LD information in the training data was known. However, in practice that information might not be available, and instead we need to estimate the LD pattern from a reference panel. As long as the LD reference panel is representative and contains at least 1,000 individuals, this assumption does not seem to affect performance in simulations.

Appendix B: Conditional Joint Analysis

To understand the conditional joint (COJO) analysis as proposed by Yang et al.,[52] we implemented a stepwise COJO analysis in LDpred. The COJO analysis estimates the joint least-squares estimate from the marginal least-squares estimate (obtained from GWAS summary statistics). If we define $D = XX'/N$, then we have the following relationship:

$$\widehat{\beta}_{\text{joint}} = (D)^{-1} \tilde{\beta} .$$

This matrix D has dimensions $M \times M$ and might be singular. However, as for LDpred, we can adjust for LD locally if the individuals in the training data do not display family or population structure, in which case the genome-wide LD matrix is approximately a banded matrix. In practice, COJO analysis with all SNPs suffers a fundamental problem of statistical inference, i.e., it infers a large number of parameters (M) by using N samples. Hence, if $N < M$, we do not expect the method to perform particularly well. We verified this in simulations (see Figure S7A). By restricting to "top" SNPs and accounting for LD by using a stepwise approach (as proposed by Yang et al.[52]) we alleviate this concern. However, although this reduces overfitting when $N < M$, this approach also risks discarding potentially informative markers from the analysis. Nevertheless, by optimizing the stopping threshold via cross-validation in an independent dataset, the method performs reasonably well in practice, especially when the number of causal markers in the genome is small. In contrast, LDpred conditions on the sample size and accounts for the noise term appropriately (under the model), leading to improved prediction accuracies regardless of training sample size.

Supplemental Data

Supplemental Data include 9 figures and 11 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.09.001.

Consortia

Members of the Schizophrenia Working Group of the Psychiatric Genetics Consortium are Stephan Ripke, Benjamin M. Neale, Aiden Corvin, James T.R. Walters, Kai-How Farh, Peter A. Holmans, Phil Lee, Brendan Bulik-Sullivan, David A. Collier, Hailiang Huang, Tune H. Pers, Ingrid Agartz, Esben Agerbo, Margot Albus, Madeline Alexander, Farooq Amin, Silviu A. Bacanu, Martin Begemann, Richard A. Belliveau, Jr., Judit Bene, Sarah E. Bergen, Elizabeth Bevilacqua, Tim B. Bigdeli, Donald W. Black, Richard Bruggeman, Nancy G. Buccola, Randy L. Buckner, William Byerley, Wiepke Cahn, Guiqing Cai, Dominique Campion, Rita M. Cantor, Vaughan J. Carr, Noa Carrera, Stanley V. Catts, Kimberly D. Chambert, Raymond C.K. Chan, Ronald Y.L. Chen, Eric Y.H. Chen, Wei Cheng, Eric F.C. Cheung, Siow Ann Chong, C. Robert Cloninger, David Cohen, Nadine Cohen, Paul Cormican, Nick Craddock, James J. Crowley, David Curtis, Michael Davidson, Kenneth L. Davis, Franziska Degenhardt, Jurgen Del Favero, Lynn E. DeLisi, Ditte Demontis, Dimitris Dikeos, Timothy Dinan, Srdjan Djurovic, Gary Donohoe, Elodie Drapeau, Jubao Duan, Frank Dudbridge, Naser Durmishi, Peter Eichhammer, Johan Eriksson, Valentina Escott-Price, Laurent Essioux, Ayman H. Fanous, Martilias S. Farrell, Josef Frank, Lude Franke, Robert Freedman, Nelson B. Freimer, Marion Friedl, Joseph I. Friedman, Menachem Fromer, Giulio Genovese, Lyudmila Georgieva, Elliot S. Gershon, Ina Giegling, Paola Giusti-Rodrguez, Stephanie Godard, Jacqueline I. Goldstein, Vera Golimbet, Srihari Gopal, Jacob Gratten, Jakob Grove, Lieuwe de Haan, Christian Hammer, Marian L. Hamshere, Mark Hansen, Thomas Hansen, Vahram Haroutunian, Annette M. Hartmann, Frans A. Henskens, Stefan Herms, Joel N. Hirschhorn, Per Hoffmann, Andrea Hofman, Mads V. Hollegaard, David M. Hougaard, Masashi Ikeda, Inge Joa, Antonio Julia, Rene S. Kahn, Luba Kalaydjieva, Sena Karachanak-Yankova, Juha Karjalainen, David Kavanagh, Matthew C. Keller, Brian J. Kelly, James L. Kennedy, Andrey Khrunin, Yunjung Kim, Janis Klovins, James A. Knowles, Bettina Konte, Vaidutis Kucinskas, Zita Ausrele Kucinskiene, Hana Kuzelova-Ptackova, Anna K. Kahler, Claudine Laurent, Jimmy Lee Chee Keong, S. Hong Lee, Sophie E. Legge, Bernard Lerer, Miaoxin Li, Tao Li, Kung-Yee Liang, Jeffrey Lieberman, Svetlana Limborska, Carmel M. Loughland, Jan Lubinski, Jouko Lnnqvist, Milan Macek, Jr., Patrik K.E. Magnusson, Brion S. Maher, Wolfgang Maier, Jacques Mallet, Sara Marsal, Manuel Mattheisen, Morten Mattingsdal, Robert W. McCarley, Colm McDonald, Andrew M. McIntosh, Sandra Meier, Carin J. Meijer, Bela Melegh, Ingrid Melle, Raquelle I. Mesholam-Gately, Andres Metspalu, Patricia T. Michie, Lili Milani, Vihra Milanova, Younes Mokrab, Derek W. Morris, Ole Mors, Preben B. Mortensen, Kieran C. Murphy, Robin M. Murray, Inez Myin-Germeys, Bertram Mller-Myhsok, Mari Nelis, Igor Nenadic, Deborah A. Nertney, Gerald Nestadt, Kristin K. Nicodemus, Liene Nikitina-Zake, Laura Nisenbaum, Annelie Nordin, Eadbhard O'Callaghan, Colm O'Dushlaine, F. Anthony O'Neill, Sang-Yun Oh, Ann Olincy, Line Olsen, Jim Van Os, Psychosis Endophenotypes International Consortium, Christos Pantelis, George N. Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T. Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O. Perkins, Olli Pietilinen, Jonathan Pimm, Andrew J. Pocklington, John Powell, Alkes Price, Ann E. Pulver, Shaun M. Purcell, Digby Quested, Henrik B. Rasmussen, Abraham Reichenberg, Mark A. Reimers, Alexander L. Richards, Joshua L. Roffman, Panos Roussos, Douglas M. Ruderfer, Veikko Salomaa, Alan R. Sanders, Ulrich Schall, Christian R. Schubert, Thomas G. Schulze, Sibylle G. Schwab, Edward M. Scolnick, Rodney J. Scott, Larry J. Seidman, Jianxin Shi, Engilbert Sigurdsson, Teimuraz Silagadze, Jeremy M. Silverman, Kang Sim, Petr Slominsky, Jordan W. Smoller, Hon-Cheong So, Chris C.A. Spencer, Eli A. Stahl, Hreinn Stefansson, Stacy Steinberg, Elisabeth Stogmann, Richard E. Straub, Eric Strengman, Jana Strohmaier, T. Scott Stroup, Mythily Subramaniam, Jaana Suvisaari, Dragan M. Svrakic, Jin P. Szatkiewicz, Erik Sderman, Srinivas Thirumalai, Draga Toncheva, Paul A. Tooney, Sarah Tosato, Juha Veijola, John Waddington, Dermot Walsh, Dai Wang, Qiang Wang, Bradley T. Webb, Mark Weiser, Dieter B. Wildenauer, Nigel M. Williams, Stephanie Williams, Stephanie H. Witt, Aaron R. Wolen, Emily H.M. Wong, Brandon K. Wormley, Jing Qin Wu, Hualin Simon Xi, Clement C. Zai, Xuebin Zheng, Fritz Zimprich, Naomi R. Wray, Kari Stefansson, Peter M. Visscher, Wellcome Trust Case Control Consortium, Rolf Adolfsson, Ole A. Andreassen, Douglas H.R. Blackwood, Elvira Bramon, Joseph D. Buxbaum, Anders D. Børglum, Sven Cichon, Ariel Darvasi, Enrico Domenici, Hannelore Ehrenreich, Tonu Esko, Pablo V. Gejman, Michael Gill, Hugh Gurling, Christina M. Hultman, Nakao Iwata, Assen V. Jablensky, Erik G. Jonsson, Kenneth S. Kendler, George Kirov, Jo Knight, Todd Lencz, Douglas F. Levinson, Qingqin S. Li, Jianjun Liu, Anil K. Malhotra, Steven A. McCarroll, Andrew McQuillin, Jennifer L. Moran, Preben B. Mortensen, Bryan J. Mowry, Markus M. Nthen, Roel A. Ophoff, Michael J. Owen, Aarno Palotie, Carlos N. Pato, Tracey L. Petryshen, Danielle Posthuma, Marcella Rietschel, Brien P. Riley, Dan Rujescu, Pak C. Sham, Pamela Sklar, David St. Clair, Daniel R. Weinberger, Jens R. Wendland, Thomas Werge, Mark J. Daly, Patrick F. Sullivan, and Michael C. O'Donovan.

Members of the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study are Peter Kraft, David J. Hunter, Muriel Adank, Habibul Ahsan, Kristiina Aittomäki, Laura Baglietto, Sonja Berndt, Carl Blomquist, Federico Canzian, Jenny Chang-Claude, Stephen J. Chanock, Laura Crisponi, Kamila Czene, Norbert Dahmen, Isabel dos Santos Silva, Douglas Easton, A. Heather Eliassen, Jonine Figueroa, Olivia Fletcher, Montserrat Garcia-Closas, Mia M. Gaudet, Lorna Gibson, Christopher A. Haiman, Per Hall, Aditi Hazra, Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON), Rebecca Hein, Brian E. Henderson, Albert Hofman, John L. Hopper, Astrid Irwanto, Mattias Johansson, Rudolf Kaaks, Muhammad G. Kibriya, Peter Lichtner, Sara Lindström, Jianjun Liu, Eiliv Lund, Enes Makalic, Alfons Meindl, Hanne Meijers-Heijboer, Bertram Müller-Myhsok, Taru A. Muranen, Heli Nevanlinna, Petra H. Peeters, Julian Peto, Ross L. Prentice, Nazneen Rahman, María José Sánchez, Daniel F. Schmidt, Rita K. Schmutzler, Melissa C. Southey, Rulla Tamimi, Ruth Travis, Clare Turnbull, Andre G. Uitterlinden, Rob B. van der Luijt, Quinten Waisfisz, Zhaoming Wang, Alice S. Whittemore, Rose Yang, and Wei Zheng.

Acknowledgments

We thank Shamil Sunayev, Brendan Bulik-Sullivan, Liming Liang, Naomi Wray, Daniel Sørensen, and Esben Agerbo for useful discussions. We would also like to thank Toni Clarke for useful comments on the software. This research was supported by NIH grants R01 GM105857, R03 CA173785, and U19 CA148065-01. B.J.V. was supported by Danish Council for Independent Research grant

DFF-1325-0014. H.K.F. was supported by the Fannie and John Hertz Foundation. This study made use of data generated by the Wellcome Trust Case Control Consortium (WTCCC) and the Wellcome Trust Sanger Institute. A full list of the investigators who contributed to the generation of the WTCCC data is available at www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113.

Received: March 26, 2015
Accepted: September 1, 2015
Published: October 1, 2015

Web Resources

The URLs for data presented herein are as follows:

CARDIoGRAMplusC4D Consortium: coronary artery disease summary statistics, http://www.cardiogramplusc4d.org
DIAGRAM Consortium: type 2 diabetes summary statistics, http://www.diagram-consortium.org/
Genetic Associations and Mechanisms in Oncology (GAME-ON) breast cancer GWAS summary statistics, http://gameon.dfci.harvard.edu
LDpred code repository, https://bitbucket.org/bjarni_vilhjalmsson/ldpred
LDpred software, http://www.hsph.harvard.edu/alkes-price/software/
Psychiatric Genomics Consortium: schizophrenia summary statistics, http://www.med.unc.edu/pgc/downloads

References

1. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
2. Pharoah, P.D., Antoniou, A.C., Easton, D.F., and Ponder, B.A. (2008). Polygenes, risk prediction, and targeted prevention of breast cancer. N. Engl. J. Med. *358*, 2796–2803.
3. Evans, D.M., Visscher, P.M., and Wray, N.R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum. Mol. Genet. *18*, 3525–3531.
4. Wei, Z., Wang, K., Qu, H.Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J.T., Chiavacci, R., et al. (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. *5*, e1000678.
5. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Mägi, R., et al.; MAGIC; Procardis Consortium (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. *42*, 937–948.
6. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.
7. Bush, W.S., Sawcer, S.J., de Jager, P.L., Oksenberg, J.R., McCauley, J.L., Pericak-Vance, M.A., and Haines, J.L.; International Multiple Sclerosis Genetics Consortium (IMSGC) (2010). Evidence for polygenic susceptibility to multiple sclerosis–the shape of things to come. Am. J. Hum. Genet. *86*, 621–625.
8. Machiela, M.J., Chen, C.Y., Chen, C., Chanock, S.J., Hunter, D.J., and Kraft, P. (2011). Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet. Epidemiol. *35*, 506–514.
9. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.
10. Schunkert, H., König, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., Preuss, M., Stewart, A.F., Barbalic, M., Gieger, C., et al.; Cardiogenics; CARDIoGRAM Consortium (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat. Genet. *43*, 333–338.
11. Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A., et al.; Diabetes Genetics Replication and Meta-analysis Consortium; Myocardial Infarction Genetics Consortium (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. *44*, 483–489.
12. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al.; Multicenter Genetic Studies of Schizophrenia Consortium; Psychosis Endophenotypes International Consortium; Wellcome Trust Case Control Consortium 2 (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. *45*, 1150–1159.
13. Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al.; LifeLines Cohort Study (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science *340*, 1467–1471.
14. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
15. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.
16. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. *14*, 483–495.
17. Ruderfer, D.M., Fanous, A.H., Ripke, S., McQuillin, A., Amdur, R.L., Gejman, P.V., O'Donovan, M.C., Andreassen, O.A., Djurovic, S., Hultman, C.M., et al.; Schizophrenia Working Group of Psychiatric Genomics Consortium; Bipolar Disorder Working Group of Psychiatric Genomics Consortium; Cross-Disorder Working Group of Psychiatric Genomics Consortium (2014). Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. Mol. Psychiatry *19*, 1017–1024.
18. 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.e1–e3.
19. Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. *14*, 507–515.
20. Plenge, R.M., Scolnick, E.M., and Altshuler, D. (2013). Validating therapeutic targets through human genetics. Nat. Rev. Drug Discov. *12*, 581–594.

21. de los Campos, G., Gianola, D., and Allison, D.B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. *11*, 880–886.
22. Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2012). SparSNP: fast and memory-efficient analysis of all SNPs for phenotype prediction. BMC Bioinformatics *13*, 88.
23. Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A., and Goddard, M.E. (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.
24. Logsdon, B.A., Carty, C.L., Reiner, A.P., Dai, J.Y., and Kooperberg, C. (2012). A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging. Bioinformatics *28*, 1738–1744.
25. Carbonetto, P., and Stephens, M. (2012). Scalable Variational Inference for Bayesian Variable Selection in Regression, and its Accuracy in Genetic Association Studies. Bayesian Anal. *7*, 73–108.
26. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.
27. Speed, D., and Balding, D.J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. *24*, 1550–1557.
28. Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet. *11*, e1004969.
29. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
30. Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., Thompson, J.R., Ingelsson, E., Saleheen, D., Erdmann, J., Goldstein, B.A., et al.; CARDIoGRAMplusC4D Consortium; DIAGRAM Consortium; CARDIOGENICS Consortium; MuTHER Consortium; Wellcome Trust Case Control Consortium (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. Nat. Genet. *45*, 25–33.
31. Grimmett, G.R., and Stirzaker, D.R. (2001). Probability and Random Processes (Oxford University Press).
32. Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V., Ingelsson, E., O'Connell, J.R., Mangino, M., et al.; GIANT Consortium (2011). Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. *19*, 807–812.
33. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
34. Finucane, H.K., et al. (2015). Partitioning heritability by functional category using GWAS summary statistics. bioRxiv, http://dx.doi.org/10.1101/014241.
35. Pirinen, M., Donnelly, P., and Spencer, C. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Ann. Appl. Stat. *7*, 369–390.
36. Goddard, M.E., Wray, N.R., Verbyla, K., and Visscher, P.M. (2009). Estimating Effects and Making Predictions from Genome-Wide Marker Data. Statist. Sci. *24*, 517–529.
37. Fisher, R.A. (1918). The correlation between relatives: on the supposition of Mendelian inheritance. Philos. Trans. R. Soc. Edinb. *52*, 399–433. https://digital.library.adelaide.edu.au/dspace/bitstream/2440/15097/1/9.pdf.
38. Daetwyler, H.D., Villanueva, B., and Woolliams, J.A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE *3*, e3395.
39. Visscher, P.M., and Hill, W.G. (2009). The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet. *5*, e1000628.
40. 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.
41. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
42. Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S.E., et al.; International Multiple Sclerosis Genetics Consortium; Wellcome Trust Case Control Consortium 2 (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature *476*, 214–219.
43. Patsopoulos, N.A., Esposito, F., Reischl, J., Lehr, S., Bauer, D., Heubach, J., Sandbrink, R., Pohl, C., Edan, G., Kappos, L., et al.; Bayer Pharma MS Genetics Working Group; Steering Committees of Studies Evaluating IFNβ-1b and a CCR1-Antagonist; ANZgene Consortium; GeneMSA; International Multiple Sclerosis Genetics Consortium (2011). Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann. Neurol. *70*, 897–912.
44. Siddiq, A., Couch, F.J., Chen, G.K., Lindström, S., Eccles, D., Millikan, R.C., Michailidou, K., Stram, D.O., Beckmann, L., Rhie, S.K., et al.; Australian Breast Cancer Tissue Bank Investigators; Familial Breast Cancer Study; GENICA Consortium (2012). A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. Hum. Mol. Genet. *21*, 5373–5384.
45. Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M.K., Dicks, E., Dennis, J., Wang, Q., Humphreys, M.K., Luccarini, C., et al.; Netherlands Collaborative Group on Hereditary Breast and Ovarian Cancer (HEBON); Familial Breast Cancer Study (FBCS); Gene Environment Interaction of Breast Cancer in Germany (GENICA) Network; kConFab Investigators; Australian Ovarian Cancer Study Group (2012). Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat. Genet. *44*, 312–318.
46. Garcia-Closas, M., Couch, F.J., Lindstrom, S., Michailidou, K., Schmidt, M.K., Brook, M.N., Orr, N., Rhie, S.K., Riboli, E., Feigelson, H.S., et al.; Gene ENvironmental Interaction and breast CAncer (GENICA) Network; kConFab Investigators; Familial Breast Cancer Study (FBCS); Australian Breast Cancer Tissue Bank (ABCTB) Investigators (2013). Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat. Genet. *45*, 392–398, e1–e2.
47. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al.; Breast and Ovarian Cancer Susceptibility Collaboration; Hereditary Breast and Ovarian Cancer

Research Group Netherlands (HEBON); kConFab Investigators; Australian Ovarian Cancer Study Group; GENICA (Gene Environment Interaction and Breast Cancer in Germany) Network (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. *45*, 353–361.e1–e2.

48. Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat. Genet. *39*, 870–874.
49. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al.; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990.
50. Ridker, P.M., Chasman, D.I., Zee, R.Y., Parker, A., Rose, L., Cook, N.R., and Buring, J.E.; Women's Genome Health Study Working Group (2008). Rationale, design, and methodology of the Women's Genome Health Study: a genome-wide association study of more than 25,000 initially healthy american women. Clin. Chem. *54*, 249–255.
51. Lee, S.H., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2012). A better coefficient of determination for genetic profile analysis. Genet. Epidemiol. *36*, 214–224.
52. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375, S1–S3.
53. Wray, N.R., Yang, J., Goddard, M.E., and Visscher, P.M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. *6*, e1000864.
54. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
55. Meuwissen, T.H., Solberg, T.R., Shepherd, R., and Woolliams, J.A. (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet. Sel. Evol. *41*, 2.
56. Golan, D., and Rosset, S. (2014). Effective genetic-risk prediction using mixed models. Am. J. Hum. Genet. *95*, 383–393.
57. Liu, D.J., Peloso, G.M., Zhan, X., Holmen, O.L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P.L., Goel, A., Zhang, H., et al. (2014). Meta-analysis of gene-level tests for rare variant association. Nat. Genet. *46*, 200–204.
58. Chen, C.-Y., Han, J., Hunter, D.J., Kraft, P., and Price, A.L. (2015). Explicit modeling of ancestry improves polygenic risk scores and BLUP prediction. Genet. Epidemiol. *39*, 427–438.
59. Maier, R., Moser, G., Chen, G.B., Ripke, S., Coryell, W., Potash, J.B., Scheftner, W.A., Shi, J., Weissman, M.M., Hultman, C.M., et al.; Cross-Disorder Working Group of the Psychiatric Genomics Consortium (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am. J. Hum. Genet. *96*, 283–294.
60. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
61. Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica *136*, 245–257.
62. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning (Springer).
63. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B Stat. Methodol. *58*, 267–288.
64. Hageman, L.A., and Young, D.M. (2004). Applied Iterative Methods (Dover Publications).
65. Legarra, A., and Misztal, I. (2008). Technical note: Computing strategies in genome-wide selection. J. Dairy Sci. *91*, 360–366.

The American Journal of Human Genetics

Supplemental Data

Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores

Bjarni J. Vilhjálmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E. Kenny, Mikkel H. Schierup, Philip De Jager, Nikolaos A. Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M. Visscher, Peter Kraft, Nick Patterson, and Alkes L. Price

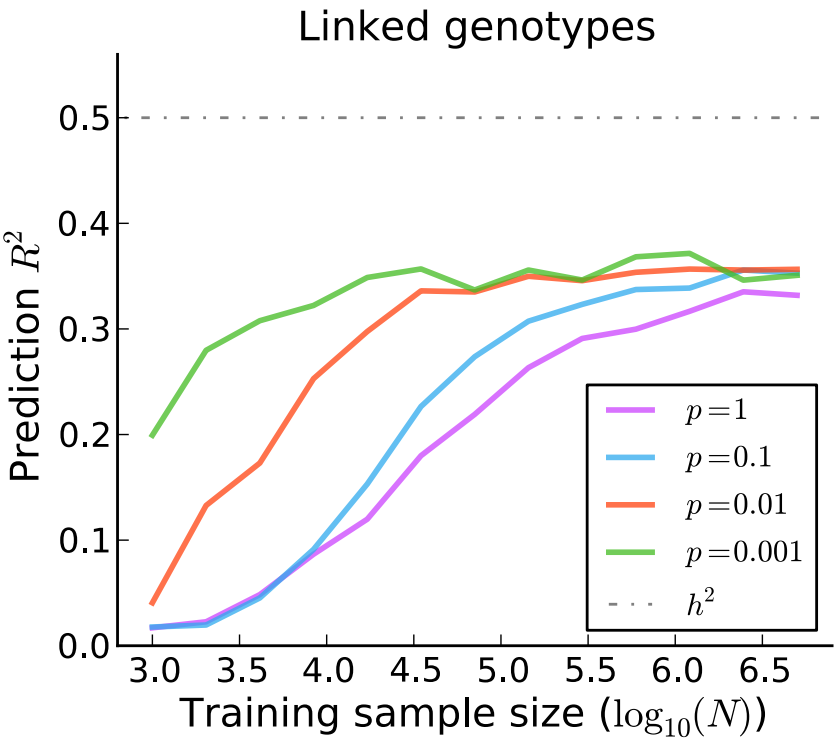

Figure S1. Performance of P+T when causal markers cluster. Performance of P+T (LD-pruning followed by thresholding) for an alternative genetic architecture where causal markers cluster. The results are averaged over 3000 simulated traits with 200K simulated genotypes where the average fraction of causal variants *p* was let vary. The simulated genotypes are linked, where we simulated independent batches of 100 markers where the squared correlation between adjacent variants in a batch was fixed to 0.9. For each simulated 100 SNP region of LD, we sampled the fraction of causal markers within a region from a Beta(*p*,*1-p*) distribution, ensuring that the expected fraction of causal markers across the genome is still *p*. This will cause causal variants to cluster in some regions of the genome. As expected, the impact of LD on the prediction accuracy of P+T is greater when causal variants cluster, and still substantial for small values of *p*.

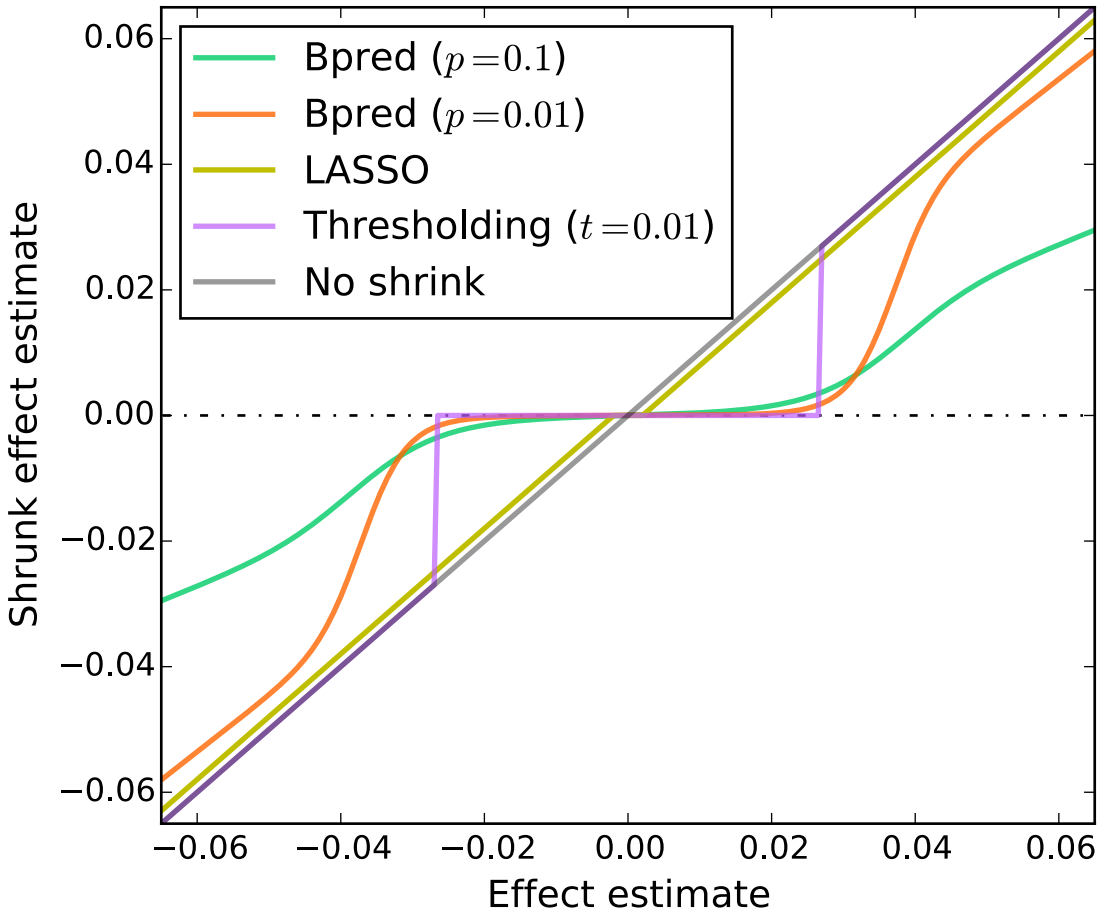

Figure S2. Comparison of five different shrinks in the absence of LD. Bpred corresponds to LDpred without LD and can be derived analytically (see Materials and Methods for details). The marginal (least square) effect estimate is plotted against the shrunk estimate for the five different shrinks. Bpred denotes the analytical solution to LDpred, which can be derived in the absence of LD (see Appendix A for details). The Bpred shrink shown here assumes that the heritability is 0.5 and the training sample size is 10,000 and the number of markers is 60,000. Similarly, the LASSO shrink shown here corresponds to the (marginal) posterior mode effect under a Laplace prior for the causal effects. Compared to *P*-value thresholding, and LASSO, Bpred can be viewed as a smoother shrink.

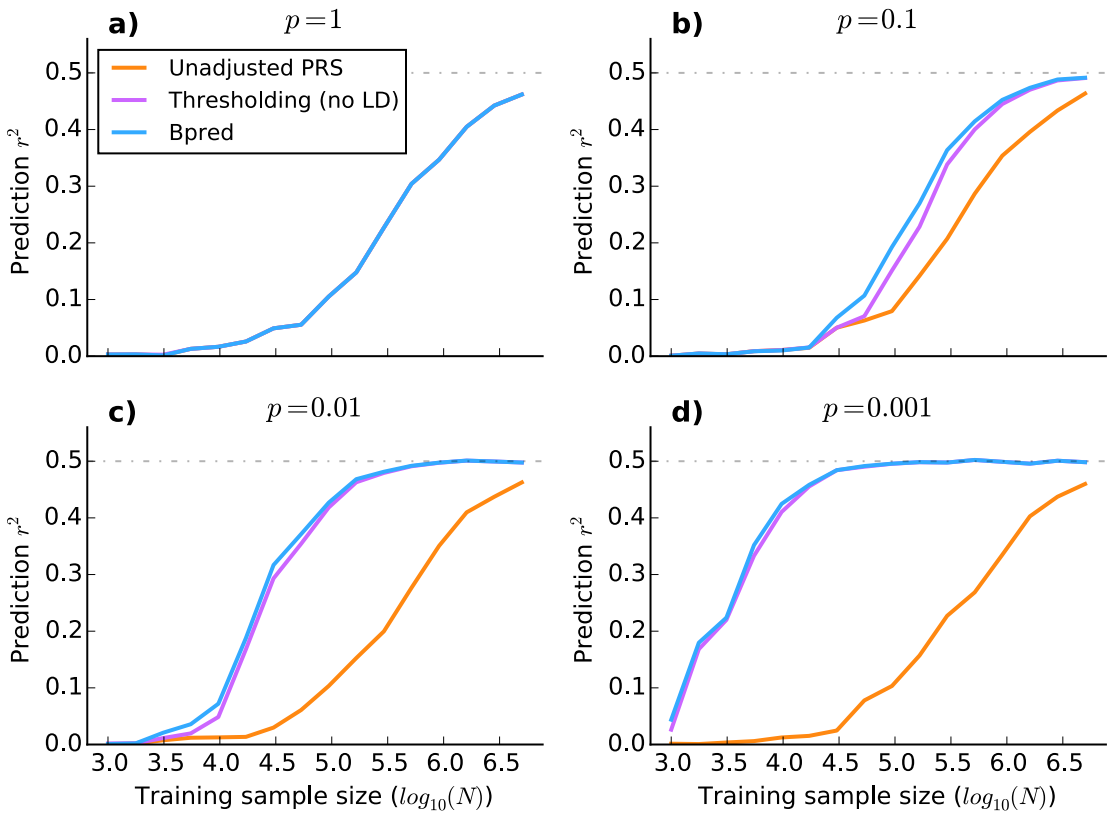

Figure S3. Comparison of methods using simulated genotypes without LD. The four subfigures **a-d** correspond to different genetic architectures where we vary *p*, the fraction of variants with (non-zero) effects drawn from a Gaussian distribution. Bpred denotes the analytical solution to LDpred, which can be derived in the absence of LD (see Appendix A for details). As expected, Bpred outperforms *P*-value thresholding in the absence of LD, although not by much.

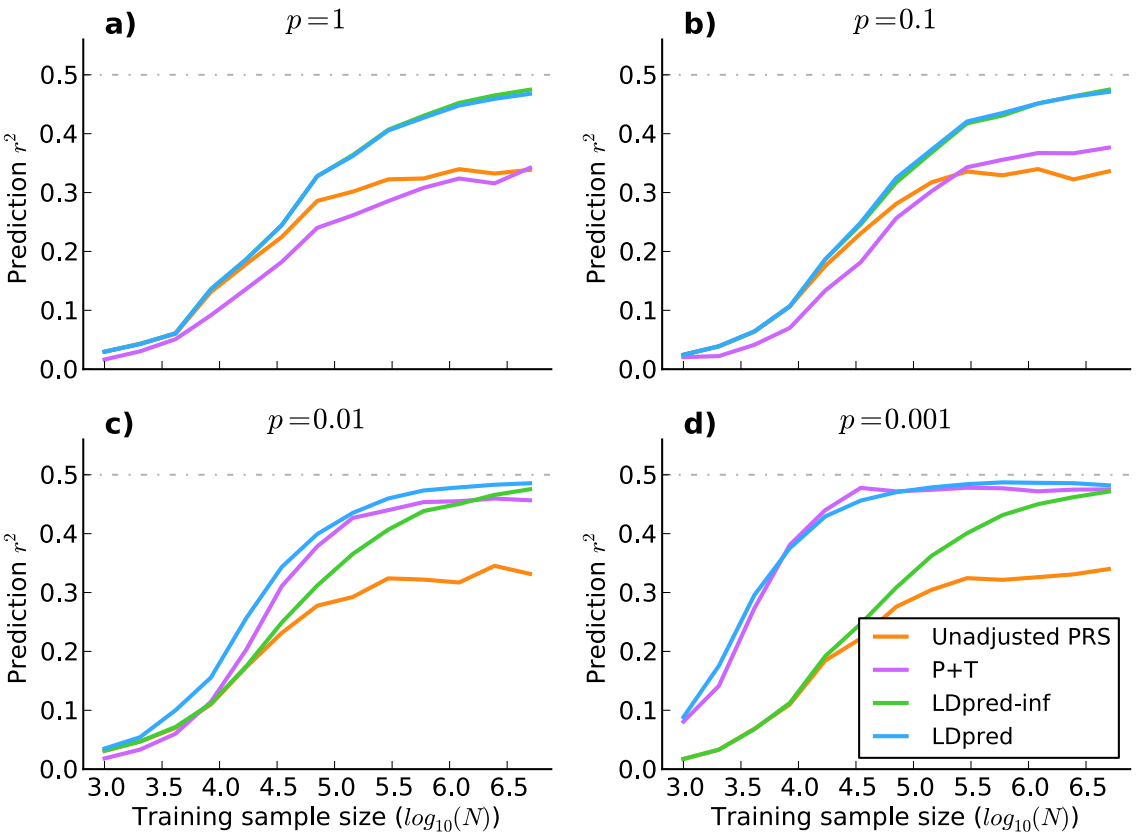

Figure S4. Comparison of methods using simulated genotypes with LD. The four subfigures **a-d** correspond to different genetic architectures where we vary *p*, the fraction of variants with (non-zero) effects drawn from a Gaussian distribution. We simulated marginal least square effect estimates with LD (see Materials and Methods for details). This enabled us to evaluate the behavior of the methods at large sample sizes. The LD structure consisted of 100 SNP regions where adjacent markers had $r^2=0.9$. For validation we simulated 200000 SNPs in 2000 individuals. For each point in the plot we averaged the results over 100 independent phenotype simulations keeping the simulated genotypes fixed (see Materials and Methods for details). Note that when p=0.001, the chance of two causal variants being in LD is very small (~1%), and thus the improvement from accounting for LD in LDpred is negligible compared to P+T.

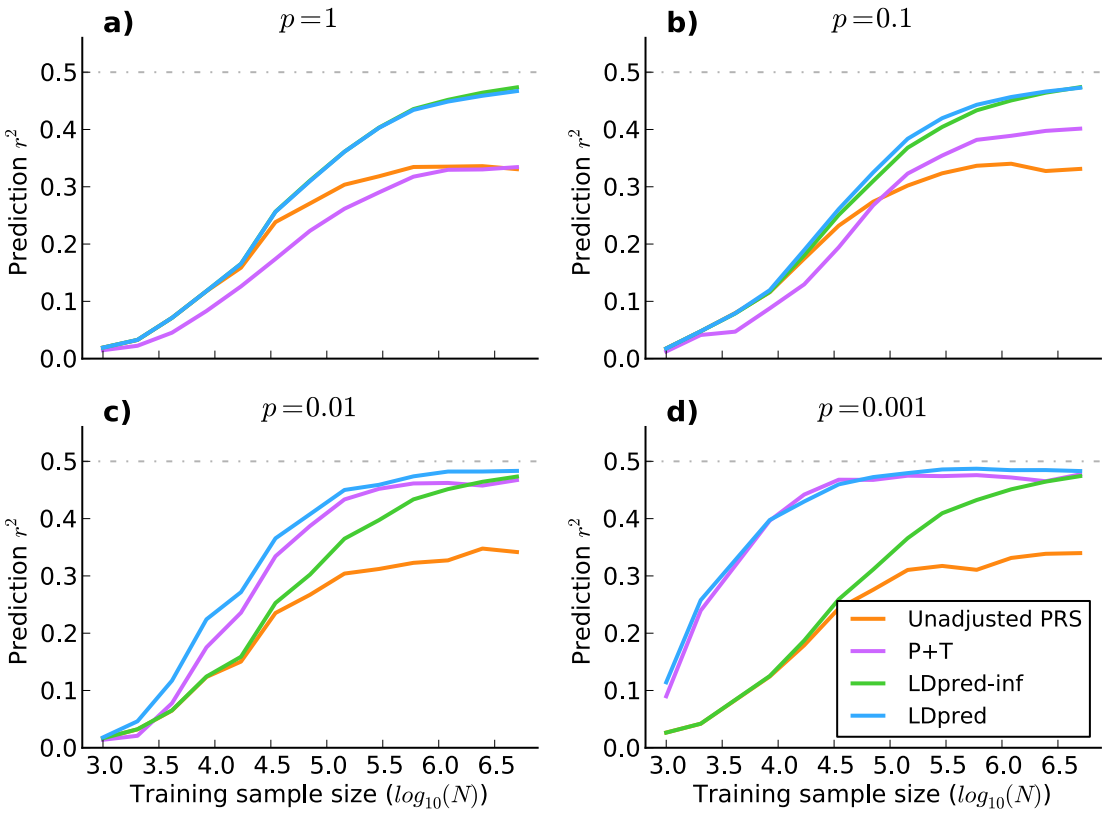

Figure S5. Comparison of methods when effects follow a Laplace distribution. Here the genotypes were simulated with LD using same simulation setup as in **Figure S4**, except the effect estimates were drawn from a Laplace mixture distribution instead of Gaussian mixture distribution. The change in prior appears to have minimal effect on the shape of the curve and the relative performance.

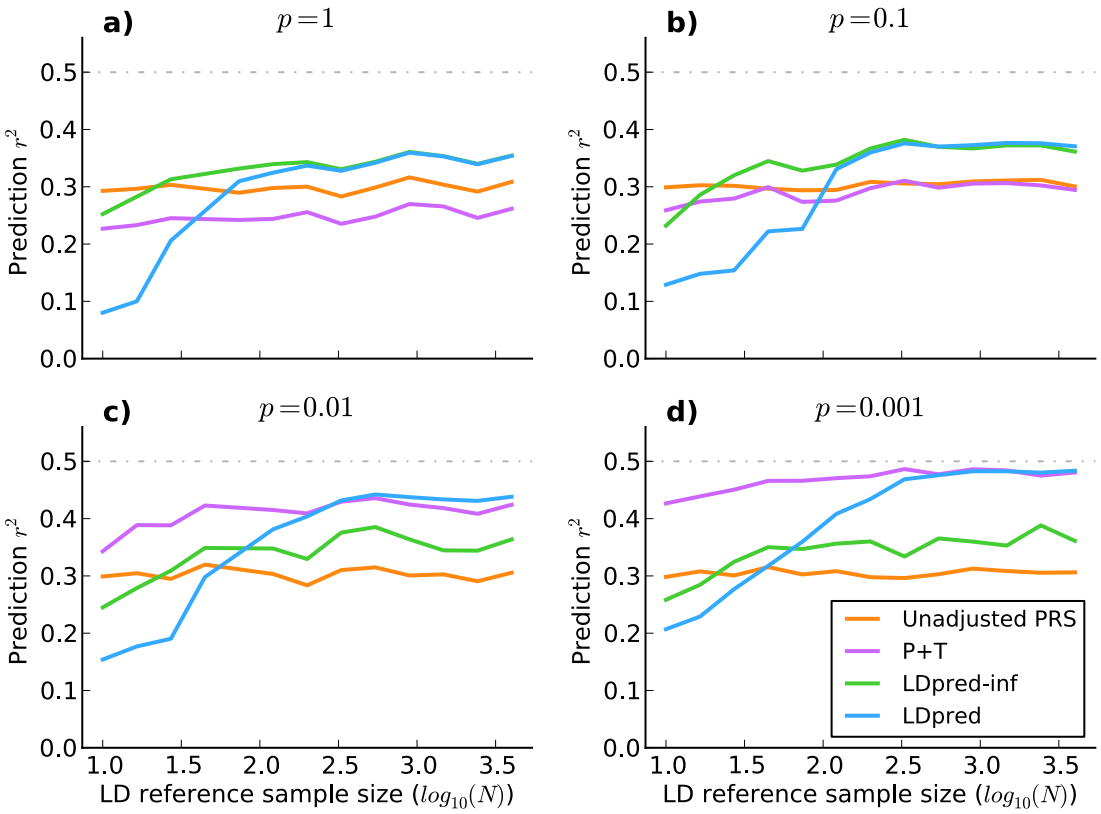

Figure S6. Prediction accuracy for methods as a function of LD reference sample size. Following the simulation setup from before (see **Figure S4.**) we simulated marginal least square effect estimates with LD (see Materials and Methods for details). We simulated segments of 100 SNPs where adjacent markers had $r^2=0.9$. We simulated in total 200000 SNPs and 2000 validation individuals. For each point in the plot we averaged the results over 100 independent phenotype simulations keeping the simulated validation genotypes fixed (see Materials and Methods for details). In addition, we simulated an LD reference panel with varying sample size along the x-axis. From these plots we see that a LD reference panel with more than 1000 individuals is necessary to ensure accurate LDpred scores. The accuracy of LDpred appears to be more sensitive to poor LD estimates than both P+T and LDpred-inf. Note that the Unadjusted PRS does not depend on LD information and is therefore expected to be a straight line, and thus providing a baseline.

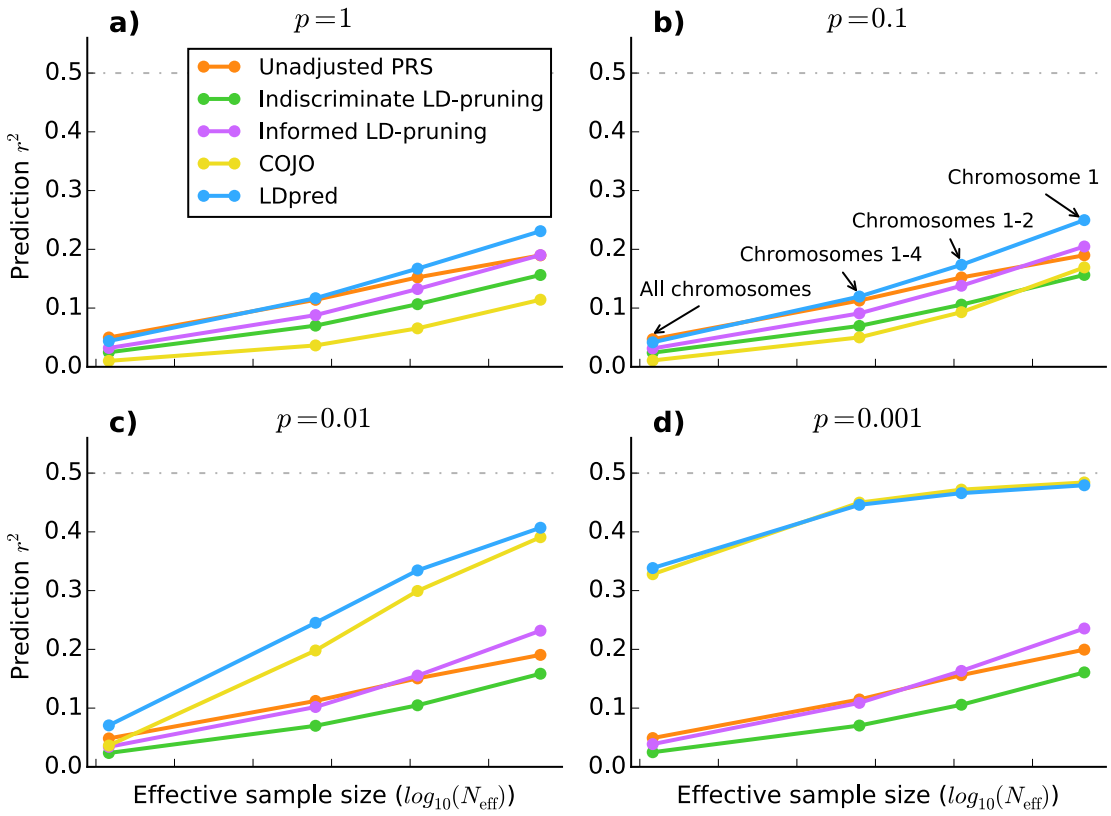

Figure S7. Comparisons to other methods using simulated traits and real WTCCC genotypes. As expected COJO[1,2] performs close to optimal with sufficient training data, or more precisely, when the ratio *(Nh²)/(Mp)* is approximately larger than 10. The comparison between the two types of LD-pruning clearly demonstrates the advantage of informed LD-pruning over indiscriminate LD-pruning, which randomly prunes either marker of a pair of markers in LD. For both LD-pruning strategies a pair of markers was considered in LD if $r^2>0.2$. When LDpred is compared to conditional joint analysis (COJO), LDpred outperforms COJO as long as the data does not overwhelm the prior, i.e. when *(Nh²)/(Mp)* is not sufficiently large (<10). For most of the diseases considered in this paper, current sample sizes are still not large enough for joint estimates to yield accurate risk scores.

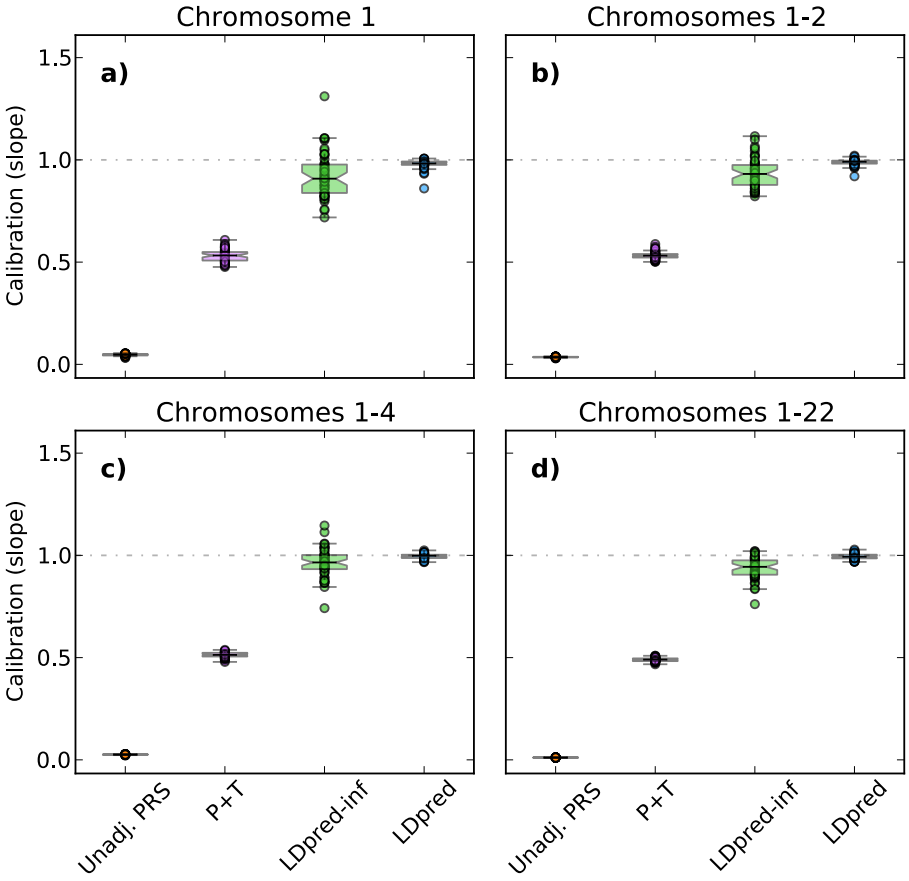

Figure S8. Boxplots of calibration slopes for simulations in Figure 2. Boxplots of calibration slopes for the four prediction methods evaluated in Figure 2 for *p=0.001* (the fraction of variants with non-zero effects). The subfigures **a-d** correspond to different number of SNPs used, ranging from 30,004 SNPs on chromosome 1 in **a)** to 376,901 SNPs or the full genome in **d)**. If the prediction conditional on the true value is unbiased then we expect a slope of one. A slope of less than one implies that the predicted value is mis-calibrated by a factor of 1/slope. Results for other values of *p (p=1; p=0.1; p=0.01)* gave similar results, and even stronger bias for P+T (LD-pruning followed by *P*-value thresholding).

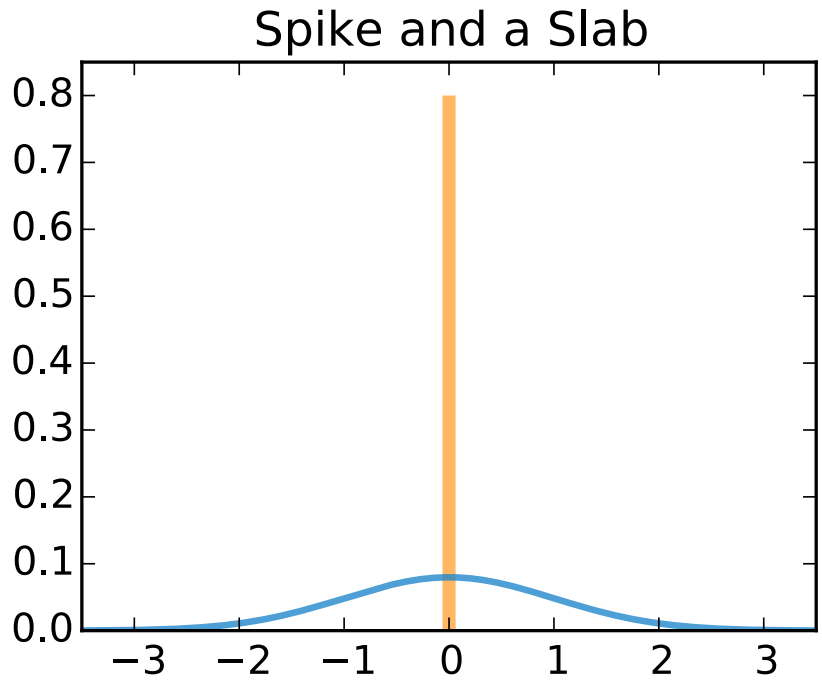

Figure S9. A spike and slab prior. An illustration of a spike and slab prior with a Gaussian slab.

Prediction Method	Accounts for LD?	Accounts for non-infinitesimal genetic architecture?	Comments
Unadjusted polygenic risk score	No.	No.	
LD-pruning followed by *P*-value thresholding (P+T)	Yes*.	Yes.	A heuristic that discards information from pruned and thresholded markers.
LDpred-inf	Yes.	No.	An analytical solution that assumes an infinitesimal prior for effects.
LDpred	Yes.	Yes.	A Gibbs sampler that assumes a point-normal mixture prior for effects.

Table S1. Overview of methods. A list of the main polygenic risk score methods (using summary association statistics as input) considered in this study. (*Although P+T prunes SNPs in high LD, it ignores bias induced by linked causal markers.)

Disease	Prediction accuracy measurement	Unadjusted PRS using all SNPs	Pruning + Thresholding	LDpred-inf	LDpred
T1D	Observed scale R^2	0.1064	0.3195	0.1062	0.3832
	Nagelkerke R^2	0.1442	0.4228	0.1438	0.5084
	Liability scale R^2	0.0426	0.0934	0.0426	0.1037
	AUC	0.6915	0.8410	0.6912	0.8738
T2D	Observed scale R^2	0.0360	0.0465	0.0404	0.0467
	Nagelkerke R^2	0.0488	0.0631	0.0547	0.0633
	Liability scale R^2	0.0257	0.0327	0.0287	0.0329
	AUC	0.6094	0.6243	0.6180	0.6275
CAD	Observed scale R^2	0.0250	0.0349	0.0290	0.0333
	Nagelkerke R^2	0.0338	0.0473	0.0393	0.0451
	Liability scale R^2	0.0191	0.0263	0.0221	0.0253
	AUC	0.5880	0.6087	0.5963	0.6043
CD	Observed scale R^2	0.0428	0.0485	0.0461	0.0824
	Nagelkerke R^2	0.0585	0.0661	0.0630	0.1122
	Liability scale R^2	0.0148	0.0167	0.0159	0.0267
	AUC	0.6212	0.6313	0.6279	0.6693
RA	Observed scale R^2	0.0483	0.1151	0.0462	0.1354
	Nagelkerke R^2	0.0656	0.1540	0.0627	0.1801
	Liability scale R^2	0.0239	0.0508	0.0229	0.0579
	AUC	0.6277	0.6994	0.6267	0.7162
BD	Observed scale R^2	0.0707	0.0876	0.0798	0.0816
	Nagelkerke R^2	0.09578	0.1185	0.1080	0.1105
	Liability scale R^2	0.0308	0.0371	0.0342	0.0349
	AUC	0.6552	0.6744	0.6662	0.6682
HT	Observed scale R^2	0.0306	0.0424	0.0348	0.0376
	Nagelkerke R^2	0.0414	0.0574	0.0471	0.0509
	Liability scale R^2	0.0258	0.0351	0.0292	0.0314
	AUC	0.6005	0.6180	0.6072	0.6109

Table S2. Numerical values of results displayed in Figure 3. The values are displayed on four different R^2 or AUC scales. To transform the prediction R^2 to liability scale we used the Lee *et al.* R^2 transformation[3] using values of disease prevalence specified in Supplementary Table 2.

Disease	Optimal fraction of causal markers used in LDpred	Optimal *P*-value threshold for Pruning + Thresholding	LDpred estimated heritability	LDpred estimated heritability on liability scale	Assumed disease prevalence
T1D	0.001	10^{-6}	1.3250	0.7258	0.005
T2D	0.03	1	0.6206	0.5125	0.03
CAD	0.03	1	0.6160	0.5181	0.035
CD	0.01	0.0001	0.7974	0.2904	0.001
RA	0.0001	10^{-6}	0.9097	0.5145	0.0075
BD	0.1	1	0.9695	0.4959	0.005
HT	0.03	1	0.6216	0.5939	0.05

Table S3. P+T and LDpred parameters for methods evaluated in Figure 3. The heritabilities are calculated as averages over 5 cross validations. The Lee *et al*. heritability transformation[4] was used to obtain the heritability on the liability scale. The LD window size used in the simulations was 400 SNPs.

Disease	Unadjusted PRS using all SNPs	Pruning + Thresholding	LDpred-inf	LDpred
T1D	0.0082	0.4301	3.2282	0.6365
T2D	0.0056	0.0278	1.2678	1.0198
CAD	0.0058	0.0231	2.1214	1.6566
CD	0.0059	0.0231	1.4159	0.8570
RA	0.0069	0.3163	2.3133	0.7755
BD	0.0076	0.0249	1.2348	1.1472
HT	0.0055	0.0301	1.7345	1.7039

Table S4. Calibration comparison for methods evaluated in Figure 3. We report the slope, where a value close to 1 represents a well-calibrated prediction. LDpred yields the most appropriately calibrated predictions.

Disease	Prediction accuracy measurement	Unadjusted PRS using all SNPs	Pruning + Thresholding	LDpred-inf	LDpred
SCZ-MGS	Observed scale R^2	0.1591	0.1510	0.1870	0.1898
	Nagelkerke R^2	0.2119	0.2014	0.2488	0.2528
	Liability scale R^2	0.0616	0.0594	0.0688	0.0694
	AUC	0.7294	0.7248	0.7499	0.7519
SCZ-ISC	Observed scale R^2	0.1169	0.0970	0.1334	0.1367
	Nagelkerke R^2	0.1574	0.1304	0.1803	0.1836
	Liability scale R^2	0.0518	0.0446	0.0578	0.0585
	AUC	0.6988	0.6784	0.7127	0.7165
MS	Observed scale R^2	0.0316	0.0674	0.0363	0.0840
	Nagelkerke R^2	0.0474	0.0978	0.0512	0.1198
	Liability scale R^2	0.0149	0.0302	0.0170	0.0368
	AUC	0.6169	0.6714	0.6187	0.6918
BC	Observed scale R^2	0.0071	0.0324	0.0092	0.0386
	Nagelkerke R^2	0.0097	0.0437	0.0119	0.0519
	Liability scale R^2	0.0040	0.0184	0.0052	0.0220
	AUC	0.5489	0.6052	0.5549	0.6156
T2D	Observed scale R^2	0.0159	0.0247	0.0214	0.0273
	Nagelkerke R^2	0.0212	0.0330	0.0309	0.0365
	Liability scale R^2	0.0112	0.0170	0.0149	0.0187
	AUC	0.5747	0.5854	0.5825	0.5953
CAD	Observed scale R^2	0.0109	0.0101	0.0124	0.0125
	Nagelkerke R^2	0.0146	0.0137	0.0168	0.0170
	Liability scale R^2	0.0085	0.0080	0.0097	0.0098
	AUC	0.5612	0.5557	0.5645	0.5647

Table S5. Numerical values of results displayed in Figure 4. The numerical results are shown on four different R^2 or AUC scales.

Trait	Prediction accuracy measurement	Unadjusted PRS using all SNPs	Pruning + Thresholding	LDpred-inf	LDpred
Height	R^2	0.0927	0.0841	0.0906	0.1014
	PC-adjusted R^2	0.0697	0.0634	0.0656	0.0853
	Risk Score + PC R^2	0.1205	0.1146	0.1166	0.1353

Table S6. Prediction accuracy for height. Height and the polygenic risk score for height is stratified by population structure. The prediction accuracy is therefore substantially reduced when we account for the first 5 principal components. Interestingly, LDpred improves the PC-adjusted prediction accuracy by 30% compared to P+T.

Diesease	Prediction accuracy measurement	Unadjusted PRS using all SNPs	Pruning + Thresholding	LDpred-inf	LDpred
T2D	Observed scale R^2	0.0022	0.0109	0.0023	0.0095
	Nagelkerke R^2	0.0029	0.0175	0.0030	0.0126
	Liability scale R^2	0.0011	0.0055	0.0012	0.0048
	AUC	0.5247	0.5633	0.5249	0.5573
CAD	Observed scale R^2	0.0029	0.0058	0.0030	0.0048
	Nagelkerke R^2	0.0039	0.0078	0.0040	0.0065
	Liability scale R^2	0.0023	0.0046	0.0024	0.0038
	AUC	0.5284	0.5418	0.5289	0.5374

Table S7. Additional validation for T2D and CAD when training on WTCCC data. Prediction accuracy for type-2 diabetes and coronary artery disease when training on WTCCC cases and controls and predicting into the WGHS data.

Disease	Optimal *P*-value threshold for Pruning + Thresholding	Optimal Gaussian mixture weight (fraction of causal markers) for LDpred	LDpred/ LD-pruning window size (# of SNPs)	GWAS sample size used in LDpred	LDpred estimated heritability	LDpred estimated heritability on liability scale	Assumed prevalence
SCZ-MGS	0.1	0.3	500	65K	0.5738	0.4231	0.01
SCZ-ISC	0.1	0.3	500	65K	0.4718	0.3479	0.01
MS	0.001	0.01	400	27K	0.3694	0.1321	0.001
BC	0.00003	0.003	400	50K	0.1934	0.1124	0.01
T2D	0.00003	0.1	300	69K	0.2061	0.1582	0.0075
CAD	0.1	1	300	86K	0.2943	0.2494	0.035

Table S8. Model parameters for results in Figure 4. Parameters inferred or assumed by P+T and LDpred for results displayed in **Figure 4**. The Lee *et al*. heritability transformation[52] was used to obtain the heritability on the liability scale.

Disease	Unadjusted PRS using all SNPs	Pruning + Thresholding	LDpred-inf	LDpred
SCZ-MGS	0.0063	0.0467	0.3845	0.3918
SCZ-ISC	0.0130	0.0407	0.4683	0.4413
MS	0.0089	0.0717	0.9092	0.2011
BC	0.0017	0.1327	1.2323	0.5650
T2D	0.0032	0.1002	0.6421	0.4057
CAD	0.0035	0.0137	0.2244	0.1868

Table S9. Calibration slopes for methods evaluated in Figure 4. We report the slope, where a value close to 1 represents a well-calibrated prediction.

Schizophrenia Cohort	Prediction accuracy measurement	Unadjusted PRS using all SNPs	Pruning + Thresholding	LDpred-inf	LDpred
MGS (European ancestry)	Observed scale R^2	0.1591	0.1510	0.1870	0.1898
	Nagelkerke R^2	0.2119	0.2014	0.2488	0.2528
	Liability scale R^2	0.0616	0.0594	0.0688	0.0694
	AUC	0.7294	0.7248	0.7499	0.7519
JPN1 (Japanese ancestry)	Observed scale R^2	0.0477	0.0702	0.0691	0.0695
	Nagelkerke R^2	0.0635	0.0944	0.0923	0.0929
	Liability scale R^2	0.0232	0.0323	0.0319	0.0320
	AUC	0.6276	0.6527	0.6523	0.6531
TCR1 (Chinese ancestry)	Observed scale R^2	0.0570	0.0616	0.0704	0.0717
	Nagelkerke R^2	0.0761	0.0821	0.0939	0.0956
	Liability scale R^2	0.0274	0.0294	0.0329	0.0336
	AUC	0.6331	0.6391	0.6483	0.6488
HOK2 (Chinese ancestry)	Observed scale R^2	0.0253	0.0306	0.0374	0.0373
	Nagelkerke R^2	0.0414	0.0511	0.0609	0.0609
	Liability scale R^2	0.0187	0.0225	0.0271	0.0271
	AUC	0.6176	0.6250	0. 6352	0.6352
AFAM (African American ancestry)	Observed scale R^2	0.0170	0.0151	0.0279	0.0280
	Nagelkerke R^2	0.0233	0.0202	0.0382	0.0383
	Liability scale R^2	0.0095	0.0084	0.0152	0.0152
	AUC	0.5745	0.5682	0.5936	0.5936

Table S10. Prediction accuracy for schizophrenia risk scores when validating in non-European populations. The accuracy is reported on four different R^2 or AUC scales.

SCZ cohort	Genetic ancestry	Optimal *P*-value threshold for Pruning + Thresholding	Optimal Gaussian mixture weight (fraction of causal markers) for LDpred	LDpred/ LD-pruning window size (# of SNPs)	GWAS sample size used in LDpred
JPN1	Japanese (Tokai)	0.1	0.3	1000	65000
TCR1	Chinese (Singapore)	0.1	0.3	1000	65000
HOK2	Chinese (Hong Kong)	1	1	1000	65000
AFAM	African American	0.3	1	400	69000

Table S11. Parameters inferred or assumed by P+T and LDpred for analysis of the non-European validation samples in **Table S10**.

References

1. Rietveld, C.A. *et al.* GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science* **340**, 1467-1471 (2013).
2. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369-375 (2012).
3. Lee, S.H., Goddard, M.E., Wray, N.R. & Visscher, P.M. A Better Coefficient of Determination for Genetic Profile Analysis. *Genetic Epidemiology* **36**, 214-224 (2012).
4. Lee, S., Wray, N., Goddard, M. & Visscher, P. Estimating missing heritability for disease from genome-wide association studies. *American Journal Of Human Genetics* **88**, 294-305 (2011).