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

1. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, 02115 MA, USA.
2. Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, 02115 MA, USA.
3. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, 02142 MA, USA.
4. Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark.
5. Queensland Brain Institute, The University of Queensland, Brisbane, 4072 Queensland, Australia.
6. The Diamantina Institute, The Translational Research Institute, University of Queensland, Brisbane, 4101 Queensland, Australia.
7. Department of Mathematics, Massachusetts Institute of Technology, Cambridge, 02139 MA, USA.
8. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, 02142 MA, USA.
9. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, 02114 MA, USA.
10. Department of Genetics, Harvard Medical School, Boston, 02115 MA, USA.
11. The Charles Bronfman Institute of Personalized Medicine, The Icahn School of Medicine at Mount Sinai, New York, NY, USA.
12. Department of Genetics and Genomic Sciences, The Icahn School of Medicine at Mount Sinai, New York, NY, USA.

13. Cardiovascular Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, 02114 MA, USA.
14. Department of Psychiatry and Behavioral Sciences, Keck School of Medicine at University of Southern California, Los Angeles, 90089 CA, USA.
15. Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, 02115 USA.
16. The Department of Psychiatry at Mount Sinai School of Medicine, New York, 10029 NY, USA
17. Department of Medicine, Lung Biology Center, University of California San Francisco, San Francisco, 94143 CA, USA.
18. Department of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, 90095 CA, USA.
19. The Icahn Institute of Genomics and Multiscale Biology, The Icahn School of Medicine at Mount Sinai, New York, NY, USA.
20. The Center of Statistical Genetics, The Icahn School of Medicine at Mount Sinai, New York, NY, USA.
21. Department of Medicine, Harvard Medical School, Boston, 02115 MA, USA.
22. Program in Translational NeuroPsychiatric Genomics, Ann Romney Center for Neurologic Diseases, Department of Neurology, Brigham and Women's Hospital, Boston, 02115 MA, USA
23. Division of Preventive Medicine, Brigham and Women's Hospital, Boston, 02215 MA, USA.
24. Department of Food and Agricultural Systems, University of Melbourne, Parkville, 3010 Victoria, Australia.
25. Biosciences Research Division, Department of Primary Industries, Bundoora, 3083 Victoria, Australia.
26. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, 02115 MA, USA.

∗ Correspondence should be addressed to B.J.V. (bjarni.vilhjalmsson@gmail.com) or A.L.P. (aprice@hsph.harvard.edu).

Abstract:

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 LD-pruning markers and applying a P-value threshold to association statistics, but this discards information and may reduce predictive accuracy. We introduce a new method, LDpred, which infers the posterior mean effect size of each marker using a prior on effect sizes and LD information from an external reference panel. Theory and simulations show that LDpred outperforms the pruning/thresholding approach, particularly at large sample sizes. Accordingly, prediction R^2 increased from 20.1% to 25.3% in a large schizophrenia data set and from 9.8% to 12.0% in a large multiple sclerosis data set. A similar relative improvement in accuracy was observed for three additional large disease data sets and when predicting in non-European schizophrenia samples. The advantage of LDpred over existing methods will grow as sample sizes increase.

Introduction

Polygenic risk scores (PRS) computed from genome-wide association study (GWAS) summary statistics have proven valuable for predicting disease risk and understanding the genetic architecture of complex traits. PRS were used to predict genetic risk in a schizophrenia GWAS for which there was only one genome-wide significant locus[1] and have been widely used to predict genetic risk for many traits[1-15]. PRS can also be used to draw inferences about genetic architectures within and across traits[12,13,16-18]. As GWAS sample sizes grow the prediction accuracy of PRS will increase and may eventually yield clinically actionable predictions[16,19-21]. However, as noted in recent work[19], current PRS methods do not account for 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[1,6,8,9,12,13,15,16,19,20] falls short of the heritability explained by the SNPs (**Figure 1** and **Supplementary Figure 1**; see Materials and Methods).

One possible solution to this problem is to use one of the many available prediction methods that require genotype data as input, including genomic BLUP—which assumes an infinitesimal distribution of effect sizes—and its extensions to non-infinitesimal mixture priors[22-29]. However, these methods are not applicable to GWAS summary statistics when genotype data are unavailable due to 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 (>100K individuals) that would be required to achieve clinically relevant predictions for most common diseases[16,19,20].

In this study we propose a Bayesian polygenic risk score, LDpred, which estimates posterior mean causal effect sizes from GWAS summary statistics assuming a prior for the genetic architecture and LD information from a reference panel. By using a point-normal mixture prior[26,30] for the marker effects, LDpred can be applied to traits and diseases with a wide range of genetic architectures. Unlike LD-pruning followed by *P*-value thresholding, 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 the widely used approach of LD-pruning followed by *P*-value thresholding[1,6,8,9,12,13,15,16,19,20,31], 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.

Materials and Methods

Overview of Methods

LDpred calculates the posterior mean effects from GWAS summary statistics conditional on a genetic architecture prior and LD information from a reference panel. The inner product of these re-weighted effect sizes with test sample genotypes is the posterior mean phenotype and thus, under the model assumptions and available data, an optimal (minimum variance and unbiased) predictor[32]. 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, accounting for sampling noise and LD[33-35] (see details below). The fraction of causal markers is allowed to vary and can be optimized with respect to prediction accuracy in a validation data set, analogous to how LD-pruning followed by *P*-value thresholding (P+T) is applied in practice. Hence, similar to P+T, where *P*-value thresholds are varied and multiple PRS are calculated, multiple LDpred risk scores are calculated using priors with varying fractions of markers with non-zero effects. The value optimizing prediction accuracy can then be determined in an independent validation data set. We approximate LD using data from a reference panel (e.g. independent validation data). The posterior mean effect sizes are estimated via Markov Chain Monte Carlo (MCMC), and applied to validation data to obtain polygenic risk scores. 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 (**Supplementary Figure 2**; 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 polygenic risk scores that rely on GWAS summary statistics. The main approaches that we compare LDpred with are listed in **Supplementary Table 1**. These include Polygenic Risk Score using all markers (PRS-all), LD-pruning followed by *P*-value thresholding (P+T) and LDpred specialized to an infinitesimal prior (LDpred-inf) (see details below). We note that LDpred-inf is an analytic method, since posterior mean effects are closely approximated by:

$$E(\beta|\tilde{\beta}, D) \approx \left(\frac{M}{Nh_g^2} I + D\right)^{-1} \tilde{\beta}, \quad (1)$$

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), as it assumes the same prior.

Phenotype model

Let Y be a $N \times 1$ phenotype vector and X a $N \times M$ genotype matrix, where the 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 its true effect, and ε the environmental and noise contribution. In this setting the (marginal) least square estimate of an individual marker effect is $\hat{\beta}_i = X_i' Y / N$. 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 square estimate for the effect by $\hat{\beta}_i = s_i \frac{z_i}{\sqrt{N}}$, where s_i is the sign, and z_i is the Z-score as 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[36].

Polygenic risk score using all markers (PRS-all)

The polygenic risk score using all genotyped markers 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 $\mathrm{S}_i = \sum_{j=1}^{M} X_{ji}\, \hat{\beta}_j$.

LD-pruning followed by thresholding (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. only markers that achieve a given significance thresholds are used, also improves prediction accuracies for many traits and diseases. In this paper the LD-pruning followed by thresholding approach refers to the strategy of first applying informed LD-pruning with r^2 threshold of 0.2, and subsequently *P*-value thresholding where the *P*-value threshold is optimized over a grid with respect to prediction accuracy in the validation data.

Bayesian approach in the special case of no LD (Bpred)

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[37]. Under the linear model described above, the posterior mean phenotype given GWAS summary statistics and LD is

$$\mathrm{E}\big(Y\big|\tilde{\beta},\widehat{D}\big) = \sum_{i=1}^{M} X_i{}'\mathrm{E}(\beta_i|\tilde{\beta},\widehat{D}).$$

Here $\tilde{\beta}$ denotes a vector of marginally estimated least square estimates as 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 may not have this information available to us and are forced to estimate the LD from a reference panel. In most of our analysis 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 estimating individual prediction accuracy, this choice is valid whenever the aim is to calculate accurate polygenic risk scores for a cohort without knowing the case-control status a priori. In other words, it is an unbiased estimate for the polygenic risk score accuracy when using the validation data 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. $\mathrm{Var}(Y) = h_g^2\Theta + (1-h_g^2)\mathrm{I}$, where h_g^2 denotes the heritability explained by the genotyped variants, and $\Theta = \frac{XX\prime}{M}$ 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 $\frac{h_g^2}{M}$. 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 (due to fewer effective independent markers).

If we assume that all samples are independent, and that all markers are unlinked and have effects drawn from a Gaussian distribution, i.e. $\beta_i \sim_{iid} N\left(0,\frac{h_g^2}{M}\right)$. This is an infinitesimal model[38] where all markers are causal and under it the posterior mean can be derived analytically, as shown by Dudbridge[16]:

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

Interestingly, with unlinked markers this infinitesimal shrink factor times the heritability, i.e. $\left(\frac{h_g^2}{h_g^2+{}^{M}/_{N}}\right)h_g^2$, is the expected squared correlation between the polygenic risk score using all (unlinked) markers and the phenotype, regardless of the underlying genetic architecture[39,40].

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

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

where p is the fraction of markers that is causal, is an unknown parameter. Under this model the posterior mean can be derived as (see **Appendix A**):

$$\mathrm{E}(\beta_i|\tilde{\beta}_i) = \left(\frac{h_g^2}{h_g^2+{}^{M\bar{p}_i}/_{N}}\right)\tilde{\beta}_i\,,$$

Where $\bar{p}_i$ is the posterior probability of an individual marker being causal and can be calculated analytically (see equation (A.1) in **Appendix A**). In our simulations we refer to this Bayesian shrink without LD as Bpred.

Bayesian approach in the presence of LD (LDpred)

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

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

Here D_l denotes the regional LD matrix within the region of LD and $\tilde{\beta}^l$ denotes the least square 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. The detailed derivation is given in the **Appendix A.** In practice we do not know the LD pattern in the training data, and we need to estimate it 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 using an approximate MCMC Gibbs sampler. This is similar 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 the **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 $c = \min(1, \frac{\hat{h}_g^2}{(\tilde{h}_g^2)_i})$, where $\hat{h}_g^2$ is the estimated heritability using 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 less than 100 iterations in practice (see **Appendix A**).

Estimation of heritability parameter

In the absence of population structure and assuming i.i.d. mean-zero SNP effects, the following equation has been shown to hold

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

where $l_j = \sum_k \left[r^2(j,k) - \frac{1-r^2(j,k)}{N-2} \right]$, is the LD score for the jth SNP summing over k neighboring SNPs in LD. Taking the average of both sides over SNPs and rearranging, we obtain a heritability estimate

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

where $\overline{\chi^2} = \sum_j \frac{\chi_j^2}{M}$, and $\bar{l} = \sum_j \frac{l_j}{M}$. We call this the aggregate estimator, and it is equivalent to LD score regression[33-35] with intercept constrained to 1 and SNP j weighted by $\frac{1}{l_j}$. Prediction accuracy is not predicated on the robustness of this estimator, which will be evaluated elsewhere. Following the conversion proposed by Lee *et al.*[41], we also reported the heritability on the liability scale.

Practical considerations

When applying LDpred to real data there are two parameters that need to be specified beforehand. The first parameter is the LD-radius, i.e. the number of SNPs on each side of a given SNP that we adjust for. There is a trade-off when deciding 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 $M/3000$ to work well in practice (this is the default value in LDpred), where M is the total number of SNPs; this corresponds to 2Mb LD-window on average in the genome. We also note that LDpred is implemented using a sliding window along the genome, whereas LDpred-inf is implemented using tiling LD windows, as this was computationally more efficient and does not affect accuracy.

The second parameter is the fraction p of non-zero effects in the prior. This parameter is analogous to the P-value threshold when conducting LD-pruning followed by *P*-value thresholding (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 to calculate 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 1000 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; (3) simulated traits using real genotypes. For most of the simulations we used the point-normal model for effect sizes as described above:

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

For some of our simulations (**Supplementary Figure 5**) 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 (**Supplementary Figure 1**) 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. The simulated trait was then obtained 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 linkage disequilibrium (LD) we simulated one genotype or SNP at a time generating 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 when simulating LD.

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 polygenic risk scores. We also simulated two smaller samples (2000 individuals) representing an independent validation data and a LD reference panel. When there is no LD, the least square effect estimates (summary statistics) are sampled from a Gaussian distribution $\hat{\beta}_i|\beta_i \sim_{iid} N\left(\beta_i, \frac{1}{N}\right)$, 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 SNPs for 1000 individuals for a given value (as described above) and average over 1000 simulations. This matrix captures the LD pattern in the validation data since we simulate it using the same procedure (described earlier). Using this LD matrix D we then sample the marginal least square estimates within a region of LD (SNP chunk) as $\hat{\beta}|\beta \sim_{iid} N\left(D\beta, \frac{D}{N}\right)$, where D is the LD matrix.

For the simulations in **Figure 1 b)** and **Supplementary Figures 1**, **3**, and **4**, we simulated least square effect estimates for 200K variants in batches of LD regions with 100 variants each (as described above). We then simulated genotypes for 2000 validation individuals and averaged over 100-3000 simulated phenotypes to ensure smooth curves. Depending on the simulation parameters, the actual number of repeats required to achieve a smooth curve varied. For the simulations in **Figure 1 a)** and **Supplementary Figure 2**, we simulated the least square estimates independently by adding an appropriately scaled Gaussian noise term to the true effects.

When simulating traits using the 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 1-2; (4) chromosome 1. We used 16,179 individuals in the WTCCC data, and 376,901 SNPs that passed quality control. In our simulations we used 3-fold cross validation, using 1/3 of the data as validation data and 2/3 as training data.

WTCCC Genotype data

We used the Wellcome Trust Case Control Consortium (WTCCC) genotypes[42] for both simulations and analysis. After quality control, pruning variants with missing rates above 1%, and removing individuals that had genetic relatedness coefficients above 0.05, we were left with 15,835 individuals genotyped for 376,901 SNPs, including 1,819 cases for bipolar disease (BD), 1,862 cases for coronary artery disease (CAD), 1,687 cases for Chron's disease (CD, 1,907 cases for hypertension (HT), 1,831 cases for rheumatoid arthritis (RA), 1,953 cases for type-1 diabetes (T1D), and 1,909 cases for type-2 diabetes (T2D). For each of the 7 diseases, we performed 5-fold cross-validation on disease cases and 2,867 controls. 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 data sets

Six large summary statistics data sets were analyzed in this paper. The Psychiatric Genomics Consortium (PGC) 2 schizophrenia summary statistics[15] consists of

34,241 cases and 45,604 controls. For our purposes we calculated GWAS summary statistics while excluding the ISC (International Schizophrenia Consortium) cohorts and the MGS (Molecular Genetics of Schizophrenia) cohorts respectively. 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 16.9M statistics. The two independent validation data sets, the ISC and the MGS data sets, both consist of multiple cohorts with individuals of European descent. For both of the validation data sets we used the chip genotypes and filtered individuals with more than 10% of genotype calls missing and filtered SNPs that had more than 1% missing rate and a minor allele frequency greater than 1%. In addition we removed SNPs that had ambiguous nucleotides, i.e. A/T and G/C SNPs. We matched the SNPs between the validation and the GWAS summary statistics data sets based on the SNP *rs*-ID and excluded triplets, SNPs where one nucleotide was unknown, and SNPs that had different nucleotides in different data sets. This was our quality control (QC) procedure for all large summary statistics data sets that we analyzed. After QC, the ISC consisted of 1562 cases and 1994 controls genotyped on 518K SNPs that overlapped with the GWAS summary statistics. The MGS data set consisted of 2681 cases and 2653 controls after QC and had 549K SNPs that overlapped with the GWAS summary statistics.

For multiple sclerosis we used the International Multiple Sclerosis (MS) Genetics Consortium summary statistics[43]. These were calculated with 9,772 cases and 17,376 controls (27,148 individuals in total) for 465K SNPs. As an independent validation data set we used the BWH/MIGEN chip genotypes with 821 cases and 2705 controls[44]. 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 114K SNPs, which we used for our analysis.

For breast cancer we used the Genetic Associations and Mechanisms in Oncology (GAME-ON) breast cancer GWAS summary statistics, consisting of 16,003 cases and 41,335 controls (both ER- and ER+ were included in this analysis)[45-48]. These summary statistics were calculated for 2.6M HapMap2 imputed SNPs. As validation genotypes we combined genotypes from five different data sets, BPC3 ER- cases and controls[45], BRCA NHS2 cases, NHS1 cases and controls from a mammographic density study, CGEMS NHS1 cases[49], and Kidney Stone NHS2 controls. All subjects in each cohort provided informed consent for this research, and procedures followed were in accordance with ethical standards. None of these 307 cases and 560 controls were included in the GWAS summary statistics analysis and thus represent an independent validation data set. We used the chip genotypes that overlapped with the GWAS summary statistics, which resulted in 444K genotypes after QC.

For coronary artery disease we used the transatlantic Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) consortium GWAS summary statistics. These were calculated using 22,233 cases and 64,762 controls

(86,995 inviduals in total) for 2.4M SNPs[10]. For the type-2 diabetes we used the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) consortium GWAS summary statistics. These were calculated using 12,171 cases and 56,862 controls (69,033 individuals in total) for 2,5M SNPs[50]. For both CAD and T2D we used the Womens Genomes Health Study (WGHS) data set as validation data[51], where we randomly down-sampled the controls. For CAD we validated in 923 cases CVD and 1428 controls, and for T2D we used 1673 cases and 1434 controls. We used the genotyped SNPs that overlapped with the GWAS summary statistics, which amounted to about 290K SNPs for both CAD and T2D after quality control. All WGHS subjects provided informed consent for this research, and procedures followed were in accordance with ethical standards.

For height we used the GIANT (Genetic Investigation of ANthropometric Traits) GWAS summary statistics as published in the Lango Allen *et al.*[6], which are calculated using 133,653 individuals and imputed to 2.8M HapMap2 SNPs. As validation cohort we used the BioMe cohort from Mount Sinai Medical Center, consisting of 2013 individuals and genotyped at 646K SNPs. All subjects provided informed consent for this research, and procedures followed were in accordance with ethical standards. After QC, the remaining 539K SNPs that overlapped with the GWAS summary statistics were used for the analysis.

For all six of these traits, we used the validation data set as the LD reference data when using LDpred and when performing LD-pruning. By using the validation as LD-reference data, we were only required to coordinate two different data sets, i.e. the GWAS summary statistics and the validation dataset. We calculated risk scores for different *P*-value thresholds 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 data for LDpred and P+T respectively.

Schizophrenia validation data sets with non-European ancestry

For the non-European validation data sets we used the MGS data set as an LD-reference, as the summary statistics were obtained using individuals of European ancestry. This required us to coordinate across three different data sets, the GWAS summary statistics, the LD reference genotypes and the validation genotypes. To ensure sufficient overlap of genetic variants across all three data sets we used 1000 genomes imputed MGS genotypes and the 1000 genomes imputed validation genotypes for the three Asian validation data sets (JPN1, TCR1, and HOK2). To limit the number of markers for these data sets we only considered markers that had MAF>0.1. After QC, and removing variants with MAF<0.1, we were left with 1.38 million SNPs and 492 cases and 427 controls in the JPN1 data set, 1.88 million SNPs and 898 cases and 973 controls in the TCR1 data set, and 1.71 million SNPs and 476 cases and 2018 controls in the HOK2 data set.

For the African American validation data set (AFAM) we used the reported GWAS summary statistics data set[15] to train on. The AFAM data set consisted of 3361 schizophrenia cases and 5076 controls. Since the AFAM data set 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 between the 1000 genomes imputed MGS genotypes, the HapMap 3 imputed AFAM genotypes and the PGC2 reported summary statistics had 482K SNPs after QC (with a MAF>0.01). All subjects in the JPN1, TCR1, HOK2, and AFAM data sets 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 data sets 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[1,9,13,15], but also report three other commonly used metrics in **Supplementary Tables 2**, **5**, **7**, and **10:** 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 5 principal components (PCs) in the validation sample (top 3 PCs for breast cancer). The relationship between observed scale R^2, liability scale R^2, and AUC is described in Lee *et al.*[52]. We note that Nagelkerke R^2 is similar to observed scale R^2 (i.e. is also affected by case-control ascertainment), but generally has slightly larger values.

Results

Simulations

We first considered simulations with simulated genotypes (see Materials and Methods). Accuracy was assessed using squared correlation (prediction R^2) between observed and predicted phenotype. The Bayesian shrink imposed by LDpred generally performed well in simulations without LD (**Supplementary Figure 3**); in this case, posterior mean effect sizes can be obtained analytically (see Materials and Methods). However, LDpred performed particularly well in simulations with LD (**Supplementary Figure 4**); the larger improvement (e.g. vs. 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 **Supplementary Figure 5**). We also evaluated the prediction accuracy as a function of the LD reference panel sample size (**Supplementary Figure 6**). LDpred performs best with an LD reference panel of at least 1000 individuals. These results also highlight the importance of using an LD reference population with LD patterns similar to the training sample, as an inaccurate reference sample will have effects similar to a reference sample of small size. Below we focus on simulations with real Wellcome Trust Case Control Consortium genotypes, which have more realistic LD properties.

Using real Wellcome Trust Case Control Consortium (WTCCC) genotypes[42] (15,835 samples and 376,901 markers, after QC), we simulated infinitesimal traits with heritability set to 0.5 (see Materials 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 $\frac{Nh_g^2}{M}$ is large. On the other hand, P+T performs less well, consistent with the intuition that pruning markers loses information.

We next simulated non-infinitesimal traits using real WTCCC genotypes, varying the proportion p of causal markers (see Materials and Methods). Results are displayed in **Figure 2b-d**. 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 PRS and LDpred-inf) perform poorly, and P+T is second best among these methods. Comparisons to additional methods are provided in **Supplementary Figure 7**; in particular, LDpred outperforms other recently proposed approaches that use LD from a reference panel[14,53] (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 vs. the predictor yields a slope of 1, and mis-calibrated otherwise; calibration is particularly important for risk prediction in clinical settings. In general, unadjusted PRS 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 **Supplementary Figure 8**. As expected, LDpred provides much better calibration than other approaches.

Application to WTCCC disease data sets

We compared LDpred to other summary statistic based methods across the 7 WTCCC disease data sets[42], using 5-fold cross validation (see Materials and Methods). Results are displayed in **Figure 3**. (We used Nagelkerke R^2 as our primary figure of merit in order to be consistent with previous work[1,9,13,15], but we also provide results for observed-scale R^2, liability-scale R^2 [ref. [52]] and AUC[54] in **Supplementary Table 2**; the relationship between these metrics is discussed in Materials 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 on the same data using BSLMM[27], BayesR[29] and MultiBLUP[28]. For these three immune-related disorders the MHC region explains a large amount of the overall variance, representing 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 statistic data sets (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 P+T, potentially due to insufficient training sample size for modeling LD to have a large impact. The inferred heritability parameters and optimal *p* parameters for LDpred, as well as the optimal thresholding parameters for P+T, are provided in **Supplementary Table 3**. The calibration of the predictions for the different approaches is shown in **Supplementary Table 4** Consistent with our simulations, LDpred provides much better calibration than other approaches.

Application to six large summary statistic data sets

We applied LDpred to five diseases—schizophrenia (SCZ), multiple sclerosis (MS), breast cancer (BC), type 2 diabetes (T2D), coronary artery disease (CAD)—for which we had GWAS summary statistics for large sample sizes (ranging from 27K to 86K individuals) and raw genotypes for an independent validation data set (see Materials and Methods). Prediction accuracies for LDpred and other methods are reported in **Figure 4** (Nagelkerke R^2) and **Supplementary Table 5** (other metrics). We also applied LDpred to height, a quantitative trait, for which we had GWAS summary statistics calculated using 134K individuals[6], and an independent validation dataset. The height prediction accuracy for LDpred and other methods are reported in Supplementary Table 6.

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 our simulations showing larger improvements when the trait is highly polygenic, as is known to be the case for SCZ[15] and height[55]. We note that for both CAD and T2D, the accuracy attained using >60K training samples from large meta-analyses (**Figure 4**) is actually lower than the accuracy attained using <5K training samples from 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[20]). To examine this further, we trained CAD and T2D PRS on the WTCCC data and validated in the WGHS data, and determined that the prediction accuracy in external WGHS validation data is substantially smaller than within the WTCCC data set (**Supplementary Table 7**). Possible explanations for this discrepancy include differences in sample ascertainment in the WGHS and WTCCC data sets, or unadjusted data artifacts in the WTCCC training/validation data.

Parameters inferred by LDpred and other methods are provided in **Supplementary Table 8**, and calibration results are provided in **Supplementary Table 9**, with LDpred again attaining the best calibration. Finally, we applied LDpred to predict SCZ risk in non-European validation samples of both African and Asian descent (see Materials and Methods). Although prediction accuracies were lower in absolute terms, we observed similar relative improvements for LDpred vs. other methods (**Supplementary Tables 10 and 11**).

Discussion

Polygenic risk scores are likely to become clinically useful as GWAS sample sizes continue to grow[16,19]. 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 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 polygenic risk scores over LD-pruning followed by *P*-value thresholding (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[23,25-30,56,57] as training data. Some of those methods employ a Variational Bayes (Iterative Conditional Expectation) approach to reduce their running time[25,26,30,56] (and report that results are similar to MCMC[30]), but we found that MCMC generally obtains more robust results than Variational Bayes when analyzing 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[26], although different priors were used by other studies, e.g. a mixture of normals[24,27,29]. One recent study proposed an elegant approach for handling case-control ascertainment while including genome-wide significant associations as fixed effects[57]; however, the correlations between distal causal SNPs induced by case-control ascertainment do not impact summary statistics from marginal analyses, and explicit modeling of non-infinitesimal effect size distributions will appropriately avoid shrinking genome-wide significant associations (**Supplementary Figure 2**).

While LDpred is a substantial improvement on existing methods for conducting polygenic prediction using summary statistics, 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 may be compromised. One potential solution is the broad sharing of summary LD statistics, which has previously been advocated in other settings[58]. 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 may not accurately model the true genetic architecture, and it is possible that other prior distributions may perform better in some settings. Third, in those instances where raw genotypes are available, fitting all markers simultaneously (if computationally tractable) may achieve higher accuracy than methods based on marginal summary statistics. Fourth, as with other prediction methods, heterogeneity across cohorts may hinder prediction accuracy; our results suggest that this could be a major concern in some data sets. Fifth, we assume that summary statistics have been appropriately corrected for genetic ancestry, but if this is not the case then the prediction accuracy may be misinterpreted[20], or may even decrease[59]. 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 polygenic risk scores. Despite these limitations, LDpred is likely to be broadly useful in leveraging summary statistic data sets 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.*[60], joint analysis of multiple traits can increase prediction accuracy. In addition, using different prior distributions across genomic regions[28] or functional annotation classes[61], may further improve the prediction. Finally, although LDpred attains a similar relative improvement when predicting into non-European samples, the lower absolute accuracy than in European samples motivates further efforts to improve prediction in diverse populations.

Web Resources

- LDpred software: http://www.hsph.harvard.edu/alkes-price/software/
- LDpred code repository: https://bitbucket.org/bjarni_vilhjalmsson/ldpred
- Genetic Associations and Mechanisms in Oncology (GAME-ON) breast cancer GWAS summary statistics: http://gameon.dfci.harvard.edu
- Type-2 diabetes summary statistics[50]: www.diagram-consortium.org
- Coronary artery disease summary statistics[10]: http://www.cardiogramplusc4d.org
- Schizophrenia summary statistics[15]: http://www.med.unc.edu/pgc/downloads

Acknowledgments

We thank Shamil Sunayev, Brendan Bulik-Sullivan, Liming Liang, Naomi Wray, Daniel Sørensen, and Esben Agerbo for useful discussions. This research was supported by NIH grants R01 GM105857, R03 CA173785, and U19 CA148065-01. BJV was supported by a Danish Council for Independent Research grant DFF-1325-0014. HKF was supported by the Fannie and John Hertz Foundation. Members of the Schizophrenia Working Group of the Psychiatric Genomics Consortium and the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study are listed in the **Supplementary Note.** 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. Authors have no conflict of interest to declare.

Appendix A: Posterior mean phenotype estimation

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 up 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 classical example is Fisher's infinitesimal model[38], which assumes genotypes are unlinked effect sizes have a Gaussian distribution (after normalizing by allele frequency).

Gaussian prior (infinitesimal model): Assume that β_i are independently drawn from a Gaussian distribution $\beta_i \sim N\left(0, \frac{h_2}{M}\right)$, where M denotes the total number of causal effects (β_i). Then we can derive a posterior mean given the ordinary least square estimate $\beta_i = \frac{X_i Y}{N}$. The least square estimate is approximately distributed as

$$\hat{\beta}_i \sim N\left(\beta_i, \frac{1-\frac{h_2}{M}}{N}\right),$$

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

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

This suggest that a uniform Bayesian shrink by a factor of $\frac{1}{1+\frac{M}{h^2 N}}$ is appropriate under Fisher's infinitesimal model.

Laplace prior (infinitesimal model): Under the Fisher/Orr model, causal effects are approximately exponentially distributed[62]. Empirical evidence largely supports this for human diseases, but also points to a genetic architecture in which there are fewer large effects[63]. Regardless, a double Exponential or a Laplace distribution is arguably a reasonable prior distribution for the effect sizes, where the variance is $\frac{h^2}{M}$ (so that they sum up to the total heritability). Under this model, the probability density function for β_i becomes

$$f(\beta_i) = \sqrt{\frac{M}{2h^2}}\exp\left(-|\beta_i|\sqrt{\frac{2M}{h^2}}\right).$$

Using the Bayes theorem we can write out the posterior density given the ordinary least square estimate as follows

$$f(\beta_i|\hat{\beta}_i) = \frac{f(\hat{\beta}_i|\beta_i)f(\beta_i)}{\int_{-\infty}^{\infty} f(\hat{\beta}_i|\beta_i)f(\beta_i)d\beta_i}.$$

Using the fact that the ordinary least square estimate is Gaussian distributed, we can write out the term in the integral as follows

$$\int_{-\infty}^{\infty} f(\hat{\beta}_i|\beta_i)f(\beta_i)d\beta_i = \frac{1}{2}\sqrt{\frac{M}{2h^2}}\int_{-\infty}^{\infty}\exp\left(-\frac{N(\hat{\beta}_i-\beta_i)^2}{2} - |\beta_i|\frac{2M}{h^2}\right)d\beta_i.$$

This integral is non-trivial, however we can solve it numerically[64]. Similarly, the posterior mean, $E(\beta_i|\widehat{\beta_\iota})$, also yields a non-trivial integral that can be evaluated numerically.

LASSO shrink: When the effects have a Gaussian prior distribution the posterior prior is symmetric, causing mean and mode to be equal. This is not the case when we use a Laplace prior for the effects. Although the posterior mean requires numerical integration, it turns out that the posterior mode has a simple analytical form[65]. The posterior mode under a Laplace prior is in fact the LASSO estimate[66]. 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

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

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[38]. 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 using mixture distributions with a mixture parameter p that denotes the fraction of causal markers. More specifically, we will consider a spike and slab prior with a 0-spike and Gaussian slab (see **Supplementary Figure 9**).

Gaussian mixture prior (spike and a slab): Assume that the effects are drawn from a mixture distribution as follows:

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

Another way of writing this is to use Dirac's delta function, i.e. write $\beta_i = pu + (1-p)v$, where $u \sim \left(0, \frac{h^2}{Mp}\right)$ 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 $\hat{\beta}_i$ as follows:

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

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

$$E(\beta_i|\hat{\beta}_i) = \int_{-\infty}^{\infty} \frac{\beta_i f(\hat{\beta}_i|\beta_i) f(\beta_i)}{\int_{-\infty}^{\infty} f(\hat{\beta}_i|\beta_i) f(\beta_i) d\beta_i} d\beta_i\,,$$

hence we only need to calculate the following definite integral

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

Thus the posterior mean is

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

where

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

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

$$P\left(\beta_i \sim N(\cdot,\cdot)|\hat{\beta}_i\right) = \frac{f\left(\hat{\beta}_i|\beta_i \sim N(\cdot,\cdot)\right)f\left(\beta_i \sim N(\cdot,\cdot)\right)}{f\left(\hat{\beta}_i\right)}$$

$$= \frac{\frac{p}{\sqrt{\frac{h^2}{Mp}+\frac{1}{N}}}\exp\left\{-\frac{1}{2}\left(\frac{\hat{\beta_i}^2}{\frac{h^2}{Mp}+\frac{1}{N}}\right)\right\}}{\frac{p}{\sqrt{\frac{h^2}{Mp}+\frac{1}{N}}}\exp\left\{-\frac{1}{2}\left(\frac{\hat{\beta_i}^2}{\frac{h^2}{Mp}+\frac{1}{N}}\right)\right\}+\frac{1-p}{\frac{1}{\sqrt{N}}}\exp\left\{-\frac{1}{2}N\hat{\beta_i}^2\right\}} \quad (A.1)$$

we can rewrite the posterior mean in a simpler fashion. If we let $\bar{p_i} = P\left(\beta_i \sim N(\cdot\right.$ $N(\,\hat{\beta}_i)$, denote the posterior probability that β_i is non-zero or Gaussian distributed, then it becomes

$$E\left(\beta_i|\hat{\beta}_i\right) = \left(\frac{1}{1+\frac{Mp}{h^2N}}\right)\bar{p}_i\hat{\beta}_i\,.$$

Posterior mean effects assuming linked markers and an infinitesimal model (LDpred-inf)

Following Yang *et al.*[53], we can obtain the joint least square effect estimates as

$$\hat{\beta}_{\text{joint}} = D^{-1}\hat{\beta}_{\text{marg}}\,, \quad (15)$$

where $D = \frac{XX\prime}{N}$ is the LD correlation matrix. 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, there the joint least square 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 adjust for LD locally instead. To formalize these ideas, let us introduce some notation. Let l_i denote the i^{th} locus or region with M_{l_i} markers, and let $\hat{\beta}$ denote the marginal least square estimate vector. In addition, let $\beta^{(i)}$ denote the vector of true effects that are in the i^{th} region, and similarly let $\hat{\beta}^{(i)}$ denote the corresponding vector of marginal effect estimates in the region. Under this model we can derive the sampling distribution for effect estimates at the i^{th} region, i.e.

$\hat{\beta}^{(i)}|\beta^{(i)}$. The mean is $E(\hat{\beta}^{(i)}|\beta^{(i)}) = D^{(i)}\beta^{(i)}$, where $D^{(i)} = \frac{X^{(i)}X^{(i)\prime}}{N}$ is the LD matrix obtained from the markers in the ith region, i.e. $X^{(i)}$. Furthermore, the conditional covariance matrix is

$$Var(\hat{\beta}^{(i)}|\beta^{(i)}) = E(\hat{\beta}^{(i)\prime}\hat{\beta}^{(i)}|\beta^{(i)}) - E(\hat{\beta}^{(i)}|\beta^{(i)})E(\hat{\beta}^{(i)}|\beta^{(i)})'$$

$$\begin{aligned}
&= \frac{1}{N^2}E\left(X^{(i)}\left(X^{(i)\prime}\,\beta^{(i)} + \epsilon\right)\left(X^{(i)}\left(X^{(i)\prime}\,\beta^{(i)} + \epsilon\right)\right)'\middle|\,\beta^{(i)}\right) \\
&- \left(D^{(i)}\beta^{(i)}\right)\left(D^{(i)}\beta^{(i)}\right)' \\
&= \left(D^{(i)}\beta^{(i)}\right)\left(D^{(i)}\beta^{(i)}\right)'\frac{1}{N}E\left(X^{(i)}\epsilon\left(X^{(i)}\epsilon\right)'|\beta^{(i)}\right) - \left(D^{(i)}\beta^{(i)}\right)\left(D^{(i)}\beta^{(i)}\right)' \\
&= X^{(i)}\frac{1}{N^2}E\left(\epsilon\epsilon'|\beta^{(i)}\right)\left(X^{(i)}\right)' = \frac{1 - {h^2}_{l_i}}{N^2}X^{(i)}\left(X^{(i)}\right)' \\
&= \frac{1 - {h^2}_{l_i}}{N}D^{(i)}\,,
\end{aligned}$$

where ${h^2}_{l_i}$ 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 $Var(\hat{\beta}^{(i)}|\beta^{(i)}) = \frac{1}{N}D^{(i)}$. 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 square estimates. Similarly, for the joint estimate we have

$$E(\hat{\beta}^{(i)}_{\text{joint}}|\beta^{(i)}) = \beta^{(i)},$$

and

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

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

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

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

$$E(\beta|\hat{\beta}) = \left(\frac{1}{1 - h_l^2}D + \frac{M}{Nh^2}I\right)^{-1}\hat{\beta}\,,$$

$$\text{Var}(\beta|\hat{\beta}) = \frac{1}{N}\left(\frac{1}{1 - h_l^2}D + \frac{M}{Nh^2}I\right)^{-1},$$

where h^2 denotes the heritability explained by the M causal variants and $h_l^2 \approx \frac{kh^2}{M}$ is the heritability of the k effects, or variants in the region of LD. If $M \gg k$, then $1 - h_l^2$ becomes approximately one, 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 (LDpred)

The Bayesian shrink under the infinitesimal model implies that we can solve it either using a Gauss- Seidel method[67,68], 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 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.

Gaussian distributed effects: Define q as follows

$$q \sim \begin{cases} 1 & \text{w. prob. } p \\ 0 & \text{w. prob. } (1-p) \end{cases},$$

then we can write $\beta = qu$ where $u \sim N\left(0, \frac{h^2}{Mp} I\right)$. 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 $\hat{\beta}$ given β is

$$f(\hat{\beta}|\beta) = \frac{1}{\sqrt{|D|}} \left(\frac{N}{2\pi(1-h_l^2)} \right)^{\frac{M}{2}} \exp\left\{ \frac{N(\hat{\beta} - D\beta)' D^{-1} (\hat{\beta} - D\beta)}{2(1-h_l^2)} \right\}. \quad (A.2)$$

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

$$E(\beta|\hat{\beta}) = \int \frac{\beta_i f(\hat{\beta}|\beta) f(\beta)}{\int f(\hat{\beta}|\beta) f(\beta) d\beta} d\beta \ ,$$

which now consists of two M-dimensional integrands. Any multiplicative term that does not involve β in the two integrands factors out. Since 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.

Metropolis Hastings Markov Chain Monte Carlo: 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.

$f(\beta|\hat{\beta}) \propto f(\beta,\hat{\beta}) = f(\hat{\beta}|\beta)f(\beta_i|\beta_{-i})f(\beta_{-i})$,

where β_{-i} denotes all the other effects except for the effect of the ith marker. Note that $(\beta_i|\beta_{-i})f(\beta_{-i}) = f(\beta)$. We can use this fact to sample efficiently in a Markov chain Monte Carlo setting where we sample one marker effect at a time in an iterative fashion (the conditional proposal distribution is therefore univariate). This ensures that the Metropolis-Hastings acceptance ratio $\alpha(\beta \to \beta^*) = \alpha(\beta^* \to \beta)$ only depends on local LD, and not the distributions of other effects, i.e.

$$\alpha(\beta_i \to \beta_i^*) = \min\left\{1, \frac{f(\beta^*,\hat{\beta})g(\beta_i^* \to \beta_i)}{f(\beta,\hat{\beta})g(\beta_i \to \beta_i^*)}\right\} = \min\left\{1, \frac{f(\hat{\beta}|\beta^*)f(\beta_i^*)g(\beta_i^* \to \beta_i)}{f(\hat{\beta}|\beta)f(\beta_i)g(\beta_i \to \beta_i^*)}\right\},$$

where the asterisk denotes the proposed effect as sampled from the conditional proposal distribution g. Since Dirac's delta density is infinite for a zero value, this ratio is undefined under the previously proposed infinitesimal model. Therefore, we consider an alternative mixture distribution with two Gaussians, one with variance $\frac{(1-\tau)h^2}{Mp}$ and the other with variance $\frac{\tau h^2}{M(1-p)}$ where τ is a small number, say $\tau = 10^{-3}$. Hence the prior distribution becomes

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

The conditional distribution $f(\hat{\beta}|\beta)$ is still the same and is given in equation (A.2). Together this gives us all the quantities needed to implement the Metropolis Hastings MCMC.

Approximate Gibbs sampler (LDpred): The general MH MCMC described above is tedious to implement and can also be computationally inefficient if proposal distributions are not carefully chosen. As a more efficient MCMC approach, we also considered a Gibbs sampler. This requires us to derive the marginal conditional posterior distributions for effects, i.e. $f(\beta|\hat{\beta},\beta_{-i})$, where β_{-i} refers to the vector of betas excluding the ith beta. We can write the posterior distribution as follows

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

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

$$P\left(\beta_i = 0 \middle| \hat{\beta}, \beta_{-i}\right) = \frac{P(\beta_i = 0, \hat{\beta}, \beta_{-i})}{P(\hat{\beta}, \beta_{-i})} = \frac{P(\beta_i = 0, \hat{\beta} | \beta_{-i})}{P\left(\beta_i = 0, \hat{\beta} \middle| \beta_{-i}\right) + \int_{\beta_i \neq 0} f\left(\hat{\beta} \middle| \beta\right) f(\beta_i) d\beta_i}.$$

Obtaining an analytical solution to this is non-trivial, however, if we assume that $P\left(\beta_i = 0 \middle| \hat{\beta}, \beta_{-i}\right) \approx P\left(\beta_i = 0 \middle| \hat{\beta}_i, \beta_{-i}\right)$, then we can simply extract out the effects of LD from other effects on the effect estimate $\hat{\beta}_i$ and then use the marginal posterior probability of the marker being causal from equation (A.1) instead, i.e. $P\left(\beta_i = 0 \middle| \hat{\beta}_i, \beta_{-i}\right) \approx \bar{p}_i$. If we sample the effect to be non-zero, and again make the simplifying assumption that $f\left(\beta_i \middle| \hat{\beta}, \beta_{-i}\right) \approx f\left(\beta_i \middle| \hat{\beta}_i, \beta_{-i}\right)$ 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

$$f\left(\beta_i \middle| \hat{\beta}, \beta_{-i}\right) \approx f\left(\beta_i \middle| \hat{\beta}_i, \beta_{-i}\right) = (1 - \bar{p}_i)\delta_{\beta_i} + \bar{p}_i h(\beta_i),$$

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

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

Practical considerations for LDpred: Throughout the derivation of LDpred above we assumed that the LD information in the training data was known. However, in practice that information may not be available and instead we need to estimate the LD pattern from a reference panel. In simulations we found that the accuracy of this estimation does affect the performance of LDpred, and we recommend that the LD be estimated from reference panels with at least 1000 individuals. In the current implementation of LDpred we fixed an LD window around the genetic variant when calculating the posterior mean effect. This is a parameter in the model that the user can set, and the optimal value may depend on the number of markers and other factors. For our analysis we accounted for LD between the SNP and a fixed window of SNPs of each side. The actual number of SNPs that were used to account for LD depends on the total number of SNPs used in each analysis, with larger windows for larger datasets.

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. As 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 the 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 converge issues arise when applying LDpred to data, then it may be worthwhile to explore higher values for the 0-jump probability.

Finally, an important parameter that LDpred assumes to known is p, the fraction of "causal markers". This parameter may of course not actually reflect the true fraction of causal markers as the model assumptions are, as always, flawed and the causal markers may not necessarily be genotyped. However, it is likely related to the true number of causal sites and may give valuable insight into the genetic architecture. Analogous to P-value thresholding we recommend that users calculate generate multiple LDpred polygenic risk scores for different values of p and then inferring and/or optimize on it in an independent validation data.

Appendix B: Conditional joint analysis

To understand the conditional joint (COJO) analysis as proposed by Yang *et al.*[53], we implemented a stepwise conditional joint analysis method in LDpred. The COJO analysis estimates the joint least square estimate from the marginal least square estimate (obtained from GWAS summary statistics). If we define $D = \frac{XX\prime}{N}$, then we have the following relationship

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

This matrix D has dimensions $M \times M$ and may 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) using N samples. Hence, if $N < M$, we do not expect the method to perform particularly well. We verified this in simulations (see **Supplementary Figure 7a)**). By restricting to "top" SNPs and accounting for LD using a stepwise approach (as proposed by Yang *et al.*[53]) 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.

References

1. Purcell, S. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).
2. Pharoah, P., Antoniou, A., Easton, D. & Ponder, B. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med* **358**, 2796-2803 (2008).
3. Evans, D.M., Visscher, P.M. & Wray, N.R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet* **18**, 3525-31 (2009).

4. Wei, Z. *et al.* From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. *PLoS Genet* **5**, e1000678 (2009).

5. Speliotes, E. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**, 937-948 (2010).

6. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 832-838 (2010).

7. Bush, W. *et al.* Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. *Am J Hum Genet* **86**, 621-625 (2010).

8. Machiela, M. *et al.* Initial impact of the sequencing of the human genome. *Genet Epidemiol* **35**, 506-514 (2011).

9. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43**, 969-976 (2011).

10. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43**, 333-338 (2011).

11. The International Multiple Sclerosis Genetics Consortium *et al.* Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. *Am J Hum Genet* **86**, 621-5 (2010).

12. Stahl, E. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet* **44**, 483-489 (2012).

13. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**, 1150-1159 (2013).

14. Rietveld, C.A. *et al.* GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science* **340**, 1467-1471 (2013).

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

16. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genetics* **9**, e1003348 (2013).

17. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483-95 (2013).

18. Ruderfer, D.M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol Psychiatry* **19**, 1017-24 (2014).

19. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics* **45**, 400-405 (2013).

20. Wray, N.R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* **14**, 507-15 (2013).

21. Plenge, R.M., Scolnick, E.M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov* **12**, 581-94 (2013).

22. de los Campos, G., Gianola, D. & Allison, D. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat Rev Genet* **11**, 880-886 (2010).

23. Abraham, G., Kowalczyk, A., Zobel, J. & Inouyes, M. SparSNP: Fast and memory-efficient analysis of all SNPs for phenotype prediction. *BMC Bioinformatics* **13**, 88 (2012).

24. Erbe, M. *et al.* Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of dairy science* **95**, 4114–4129 (2012).

25. Logsdon, B.A., Carty, C.L., Reiner, A.P., Dai, J.Y. & Kooperberg, C. A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging. *Bioinformatics* **28**, 1738-44 (2012).

26. Carbonetto, P. & Stephens, M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and its Accuracy in Genetic Association Studies. *Bayesian Analysis* **7**, 73-108 (2012).

27. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics* **9**, e1003264 (2013).

28. Speed, D. & Balding, D.J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res* **24**, 1550-7 (2014).

29. Moser, G. *et al.* Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLoS Genet* **11**, e1004969 (2015).

30. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-290 (2015).

31. CARDIoGRAMplusC4D Consortium. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics* **45**, 25–33 (2013).

32. Grimmett, G.R. & Stirzaker, D.R. *Probability and Random Processes*, (Oxford University Press, 2001).

33. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* **19**, 807-812 (2011).

34. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015).

35. Finucane, H.K. *et al.* Partitioning heritability by functional category using GWAS summary statistics. *Nat Genet* **In press**(2015).

36. Pirinen, M., Donnelly, P. & Spencer, C. Efficient computation with a linear mixed model on large-scale data sets with applications to
genetic studies. *Ann Appl Stat* **7**, 369-390 (2013).

37. Goddard, M.E., Wray, N.R., Verbyla, K. & Visscher, P.M. Estimating Effects and Making Predictions from Genome-Wide Marker Data. 517-529 (2009).

38. Fisher, R. The correlation between relatives: on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh* (1918).

39. Daetwyler, H., Villanueva, B. & Woolliams, J. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* **3**, e3395 (2008).

40. Visscher, P. & Hill, W. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genetics* **5**, e1000628 (2009).

41. 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).

42. Consortium, W.T.C.C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

43. The International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Control Consortium 2. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214-219 (2011).

44. Patsopoulos, N.A. *et al.* Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Annals of Neurology* **70**, 897-912 (2011).

45. Siddiq, A. *et al.* A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Human Molecular Genetics* **21**, 5373-5384 (2012).

46. Ghoussaini, M. *et al.* Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat Genet* **44**, 312-318 (2012).

47. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative–specific breast cancer risk loci. *Nature Genetics* **45**, 392–398 (2013).

48. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353-361 (2013).

49. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**, 870-874 (2007).

50. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**, 981-990 (2012).

51. Ridker, P.M. *et al.* 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. *Clinical Chemistry* **54**, 249-55 (2008).

52. 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).

53. 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).

54. Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling. *PLoS Genet* **6**, e1000864 (2010).

55. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-1186 (2014).

56. Meuwissen, T.H., Solberg, T.R., Shepherd, R. & Woolliams, J.A. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet Sel Evol* **41**, 2 (2009).

57. Golan, D. & Rosset, S. Effective genetic-risk prediction using mixed models. *Am J Hum Genet* **95**, 383-93 (2014).

58. Liu, D.J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat Genet* **46**, 200-4 (2014).

59. Chen, C.-Y., Han, J., Hunter, D.J., Kraft, P. & Price, A.L. Explicit modeling of ancestry improves polygenic risk scores and BLUP prediction. *bioRxiv* (2014).

60. Maier, R. *et al.* Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder. *The American Journal of Human Genetics* **96**, 283-294 (2015).

61. Gusev, A. *et al.* Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *The American Journal of Human Genetics* **95**, 535-552 (2014).

62. Barton, N.H. & Keightley, P.D. Understanding quantitative genetic variation. *Nat Rev Genet* **3**, 11-21 (2002).

63. Park, J.-H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* **42**, 570-575 (2010).

64. Goddard, M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245-257 (2009).

65. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning*, (Springer, 2009).

66. Tibshirani, R. Regression shrinkage and selection via the lasso. *J Royal Statist Soc B* **58**, 267-288 (1996).

67. Hageman, L.A. & Young, D.M. *Applied Iterative Methods*, (Dover Publications, 2004).

68. Legarra, A. & Misztal, I. Technical Note: Computing Strategies in Genome-Wide Selection. *Journal of Dairy Science* **91**, 360-366.

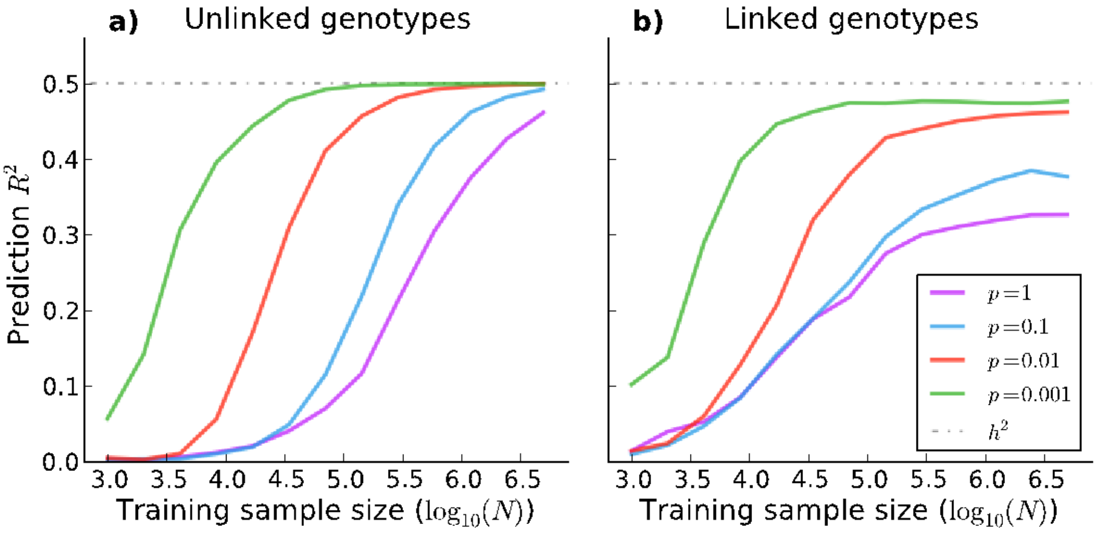

Figure 1: The performance of polygenic risk scores using LD-pruning (r^2<0.2) followed by thresholding (P+T) with optimized threshold when applied to simulated genotypes with and without LD. The prediction accuracy, as measured by squared correlation between the true phenotypes and the polygenic risk scores (prediction R^2), is plotted as a function of the training sample size. The results are averaged over 1000 simulated traits with 200K simulated genotypes where the fraction of causal variants *p* was let vary. In **a)** the simulated genotypes are unlinked. In **b)** 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.

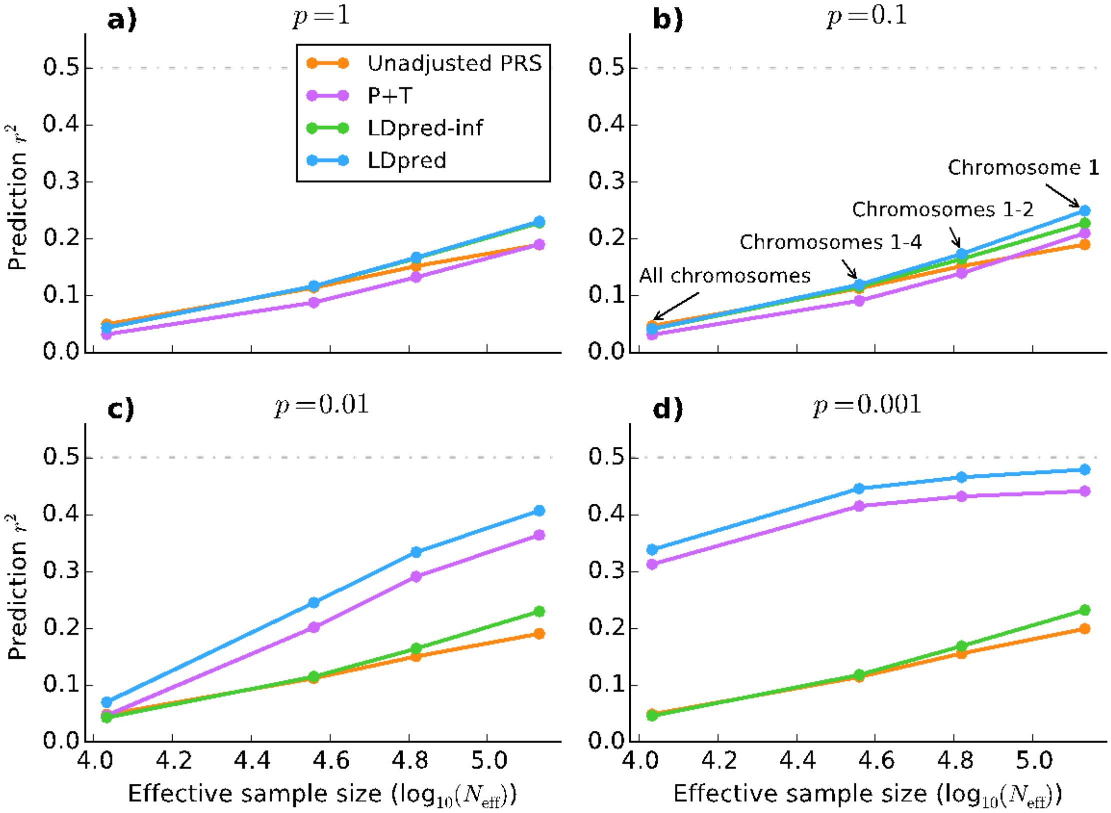

Figure 2: Comparison between the four different methods listed in Table 1 when applied to simulated traits with WTCCC genotypes. The four subfigures **a-d**, correspond to different values of the fraction of simulated causal markers (p) with (non-zero) effect sizes sampled from a Gaussian distribution. To aid interpretation of the results, we plot the accuracy against the effective sample size defined as $N_e = \frac{N}{M_{sim}} M$, 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 1-2) and 30,004 (chromosome 1), respectively. The effective sample size is the sample size that maintains the same N/M ratio if using all SNPs.

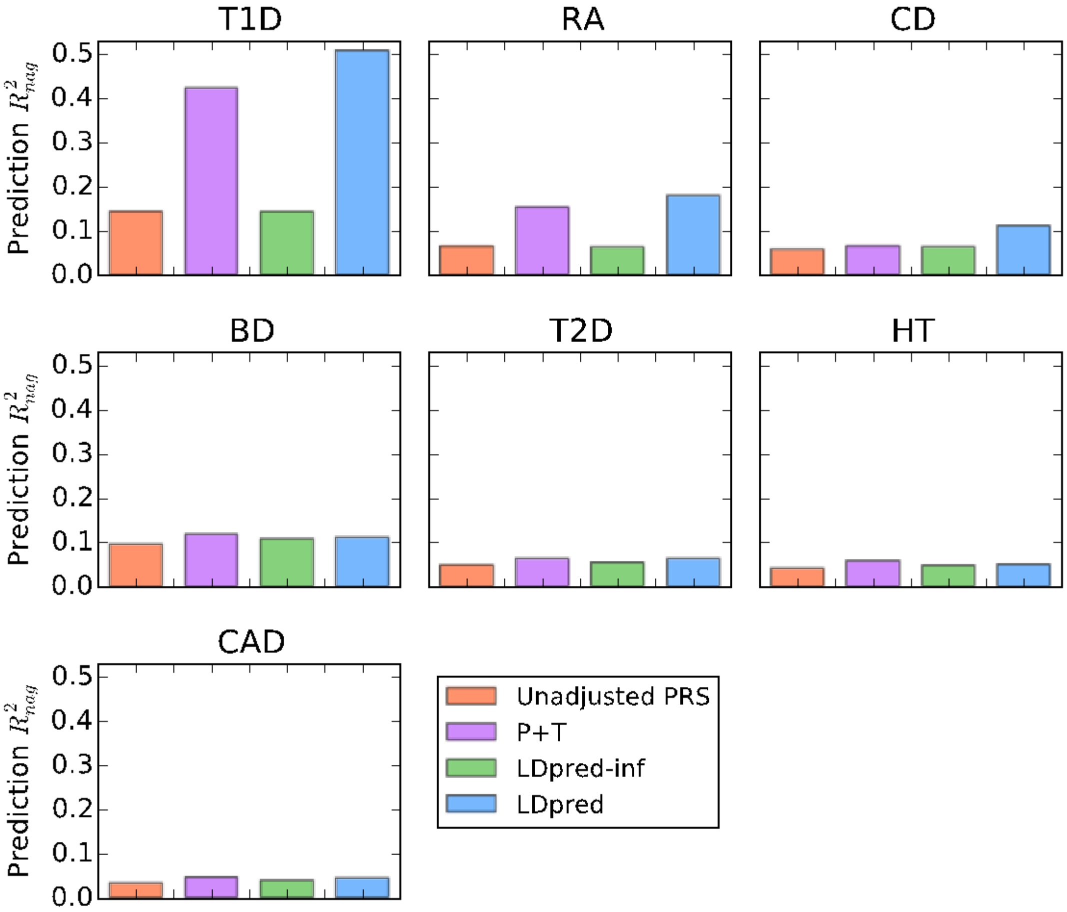

Figure 3: Comparison of methods when applied to 7 WTCCC disease data sets, type-1 diabetes (T1D), rheumatoid arthritis (RA), Chron's disease (CD), bipolar disease (BD), type-2 diabetes (T2D), hypertension (HT), coronary artery disease (CAD). The Nagelkerke prediction R^2 is shown on the y-axis, see **Supplementary Table 1** for other metrics. LDpred significantly improved the prediction accuracy for the immune-related diseases T1D, RA, and CD (see main text).

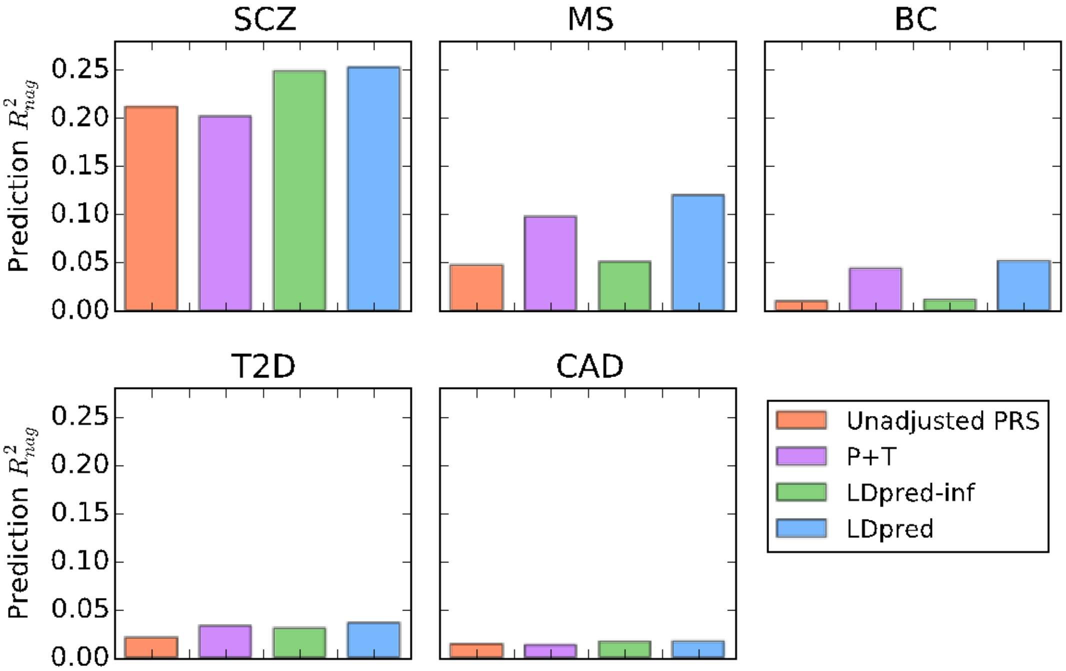

Figure 4: Comparison of prediction accuracy for 5 different diseases, schizophrenia (SCZ), multiple sclerosis (MS), breast cancer (BC), type-2 diabetes (T2D), and coronary artery disease (CAD). The risk scores were trained using large GWAS summary statistics data sets and used to predict in independent validation data sets. The Nagelkerke prediction R^2 is shown on the y-axis (see **Supplementary Table 1** for other metrics). LDpred improved the prediction R^2 by 11-25% compared to LD-pruning + Thresholding (P+T). SCZ results are shown for the SCZ-MGS validation cohort used in recent studies[9,13,15], but LDpred also produced a large improvement for the independent SCZ-ISC validation cohort (**Supplementary Table 4)**.