Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores

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

1. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
2. Program for Genetics and Statistical Genomics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
3. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.
4. Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark.
5. Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia.
6. The Diamantina Institute, The Translational Research Institute, University of Queensland, Brisbane, Queensland, Australia.
7. Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA.
8. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
9. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA.
10. Department of Genetics, Harvard Medical School, Boston, MA, USA.
11. Department of Medicine, Harvard Medical School, Boston, MA, USA.
12. Cardiovascular Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
13. Department of Psychiatry and Behavioral Sciences, Keck School of Medicine at University of Southern California, Los Angeles, CA, USA.
14. Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA.

15. The Department of Psychiatry at Mount Sinai School of Medicine, New York, NY, USA
16. Department of Medicine, Lung Biology Center, University of California San Francisco, San Francisco, CA, USA.
17. Department of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, CA, USA.
18. Program in Translational NeuroPsychiatric Genomics, Ann Romney Center for Neurologic Diseases, Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA
19. Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA.
20. Department of Food and Agricultural Systems, University of Melbourne, Parkville, Victoria, Australia.
21. Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria, Australia.
22. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, 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 causal 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 Online 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[22-28]. 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[26,29] 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[1,6,8,9,12,13,15,16,19,20,30], 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.

Results

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, the best unbiased prediction (see Online Methods). 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[31-33] (see Online Methods). 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 *P*-value thresholds are varied in standard PRS. 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 Online Methods). 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 **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 Online Methods). We note that LDpred-inf is an analytic method, since posterior mean effects are closely approximated by:

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

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

Simulations

We first considered simulations with simulated genotypes (see Online 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 Online 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**). 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[34] (15,835 samples and 376,901 markers, after QC), we simulated infinitesimal traits with heritability set to 0.5 (see Online 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 Online 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 6**; in particular, LDpred outperforms other recently proposed approaches that use LD from a reference panel[14,35].

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 7**. 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[34], using 5-fold cross validation (see Online 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. [36]] and AUC[37] in **Supplementary Table 1**; the relationship between these metrics is discussed in Online 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]. 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 and BSLMM both explicitly model non-infinitesimal architectures; however, unlike LDpred, BSLMM requires full genotype data and cannot be applied to large summary statistic data sets (see below). 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 2**. The calibration of the predictions for the different approaches is shown in **Supplementary Table 3** Consistent with our simulations, LDpred provides much better calibration than other approaches.

Application to five large summary statistic data sets

We applied LDpred to five diseases—schizophrenia (SCZ), multiple sclerosis (MS), breast cancer (BC), type 2 diabetes (T2D) and 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 Online Methods). Prediction accuracies for LDpred and other methods are reported in **Figure 4** (Nagelkerke R^2) and **Supplementary Table 4** (other metrics).

For all 5 diseases, 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, and 0.017 for CAD). The relative increase in Nagelkerke R^2 over other approaches ranged from 11% for T2D to >25% for SCZ. 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]. 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]).

Parameters inferred by LDpred and other methods are provided in **Supplementary Table 5**, and calibration results are provided in **Supplementary Table 6**, 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 Online Methods). Although prediction accuracies were lower in absolute terms, we observed similar relative improvements for LDpred vs. other methods (**Supplementary Tables 7 and 8**).

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[23,25-28,38-40] as training data. Some of those methods employ a Variational Bayes (Iterative Conditional Expectation) approach to reduce their running time[25,26,38,40] (and report that results are similar to MCMC[40]), 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]. One recent study proposed an elegant approach for handling case-control ascertainment while including genome-wide significant associations as fixed effects[39]; 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[41]. 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, joint analysis of multiple traits—which can potentially increase prediction accuracy—is not considered here, and remains as a future direction[42]. Sixth, 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[43]. Seventh, 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. Finally, we have not considered the advantages of different prior distributions across genomic regions[28] or functional annotation classes[44], whose incorporation into methods for polygenic prediction remains as a future direction. Despite these limitations, LDpred is likely to be broadly useful in leveraging summary statistic data sets for polygenic prediction.

Online Methods

Phenotype model

Let Y be a $N\times1$ 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[45].

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 ith individual is $S_i = \sum_{j=1}^{M} X_{ji}\,\hat{\beta}_j$.

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 and is unbiased in the Bayesian sense[46]. Under the linear model described above, the posterior mean phenotype given GWAS summary statistics and LD is

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

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 our analysis we used the independent validation data set to estimate the local LD structure in the training data.

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 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[47] 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[48,49].

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

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

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 **Supplementary Note**):

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

where

$$\bar{p}_i = \frac{\frac{p}{\sqrt{h_g^2/Mp + 1/N}} e^{\frac{-\hat{\beta}_i^2}{2(h_g^2/Mp+1/N)}}}{\frac{p}{\sqrt{h_g^2/Mp+1/N}} e^{\frac{-\hat{\beta}_i^2}{2(h_g^2/Mp+1/N)}} + \frac{1-p}{\sqrt{1/N}} e^{\frac{-\hat{\beta}_i^2}{2(1/N)}}}$$

is the posterior probability of an individual marker being causal. 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

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

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 **Supplementary Note**. 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 we therefore approximate it numerically in

LDpred. The Bayesian shrink under the infinitesimal model implies that we can solve it either using a Gauss-Seidel method[50,51] , 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. The approximate Gibbs sampler is described in detail in the **Supplementary Note.** 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 **Supplementary Note**).

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 j'th SNP summing over k neighboring SNPs in LD. Taking the average of both sides over SNPs and rearranging, we obtain a heritability estimate

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

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[31-33] 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.*[52], we also reported the heritability on the liability scale.

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:

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

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 p parameter from a Beta(p,1- p) distribution. 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 a smaller sample (2000 individuals) representing an independent validation data. 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 100-500 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[34] 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.

Summary statistics and independent validation data sets

Five 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. 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[53]. 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[54]. 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)[55-58]. 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[55], BRCA NHS2 cases, NHS1 cases and controls from a mammographic density study, CGEMS NHS1 cases[59], and Kidney Stone NHS2 controls. 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[60]. For both CAD and T2D we used the Womens Genomes Health Study (WGHS) data set as validation data[61], 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.

For all of these data sets we used the validation data set as an LD-reference for LDpred and when LD-pruning. 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. 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).

Prediction accuracy metrics

For simulated 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 1**, **4**, and **7:** 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.*[36]. 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.

Web Resources

- LDpred software: http://www.hsph.harvard.edu/alkes-price/software/ and 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[60]: 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. 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

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. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. 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. 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. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **advance online publication**(2015).
30. CARDIoGRAMplusC4D Consortium. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics* **45**, 25–33 (2013).
31. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* **19**, 807-812 (2011).
32. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **advance online publication**(2015).
33. Finucane, H.K. *et al.* Partitioning heritability by functional category using GWAS summary statistics. *bioRxiv* (2015).
34. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

35. 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).
36. 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).
37. 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).
38. 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).
39. Golan, D. & Rosset, S. Effective genetic-risk prediction using mixed models. *Am J Hum Genet* **95**, 383-93 (2014).
40. Loh, P.R. *et al.* Efficient Bayesian mixed model analysis increases association power in large cohorts. *Submitted* (2014).
41. Liu, D.J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat Genet* **46**, 200-4 (2014).
42. 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.
43. 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).
44. 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).
45. 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).
46. Goddard, M.E., Wray, N.R., Verbyla, K. & Visscher, P.M. Estimating Effects and Making Predictions from Genome-Wide Marker Data. 517-529 (2009).
47. Fisher, R. The correlation between relatives: on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh* (1918).
48. Daetwyler, H., Villanueva, B. & Woolliams, J. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* **3**, e3395 (2008).
49. Visscher, P. & Hill, W. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genetics* **5**, e1000628 (2009).
50. Hageman, L.A. & Young, D.M. *Applied Iterative Methods*, (Dover Publications, 2004).
51. Legarra, A. & Misztal, I. Technical Note: Computing Strategies in Genome-Wide Selection. *Journal of Dairy Science* **91**, 360-366.

52. 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).
53. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214-219 (2011).
54. Patsopoulos, N.A., *et al*. Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Annals of Neurology* **70**, 897-912 (2011).
55. 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).
56. Ghoussaini, M. *et al.* Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat Genet* **44**, 312-318 (2012).
57. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative–specific breast cancer risk loci. *Nature Genetics* **45**, 392–398 (2013).
58. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353-361 (2013).
59. 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).
60. 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).
61. 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).

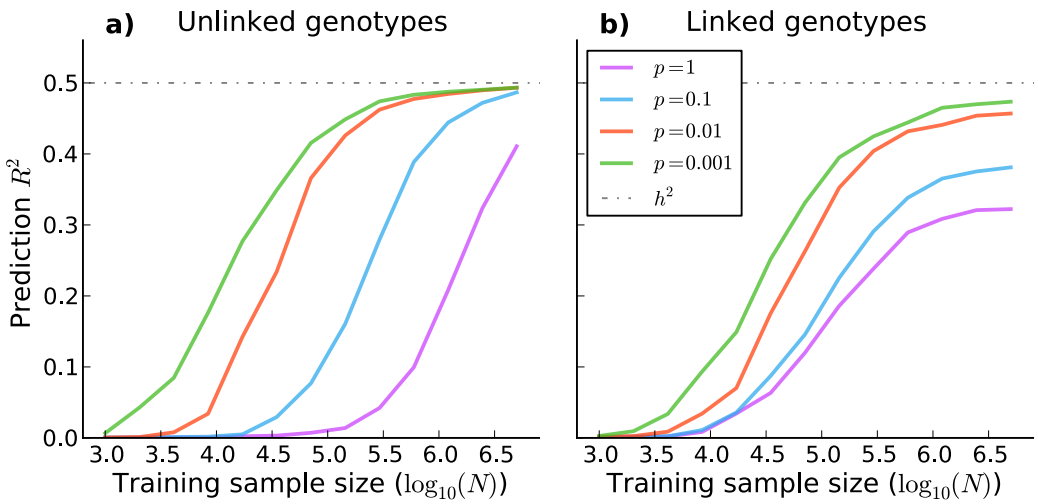


Figure 1: The performance of polygenic risk scores using LD-pruning (r^2<0.2) followed by thresholding (P+T) with optimized threshold when applied to simulated genotypes with and without LD. The prediction accuracy, as measured by squared correlation between the true phenotypes and the polygenic risk scores (prediction R^2), is plotted as a function of the training sample size. The results are averaged over 2000 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.

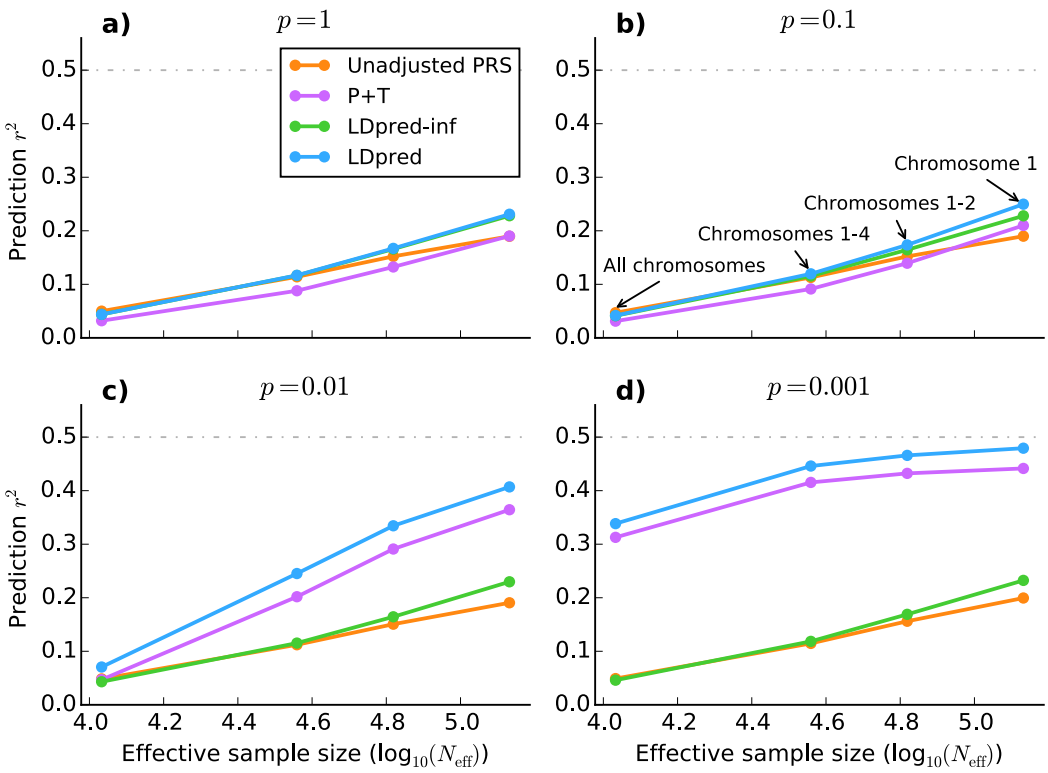


Figure 2: Comparison between the four different methods listed in Table 1 when applied to simulated traits with WTCCC genotypes. The four subfigures **a-d**, correspond to different values of the fraction of simulated causal markers (p) with (non-zero) effect sizes sampled from a Gaussian distribution. 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.

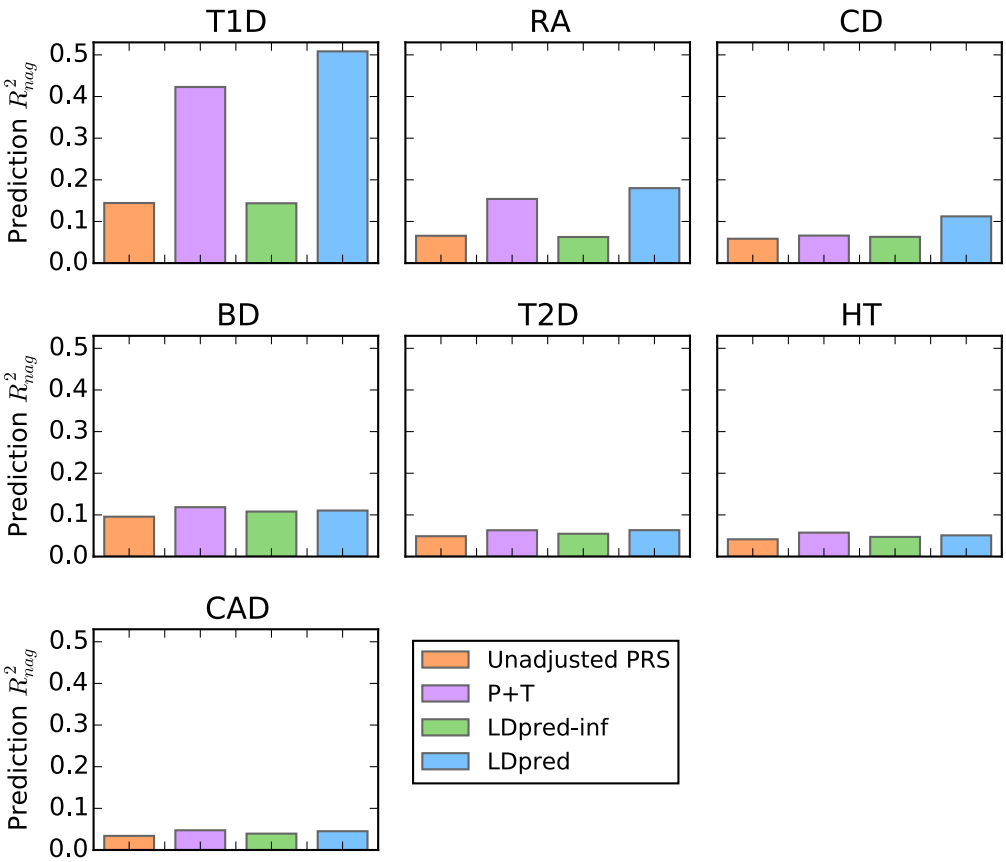


Figure 3: Comparison of methods when applied to 7 WTCCC disease data sets, type-1 diabetes (T1D), rheumatoid arthritis (RA), Chron's disease (CD), bipolar disease (BD), type-2 diabetes (T2D), hypertension (HT), coronary artery disease (CAD). The Nagelkerke prediction R^2 is shown on the y-axis, see **Supplementary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

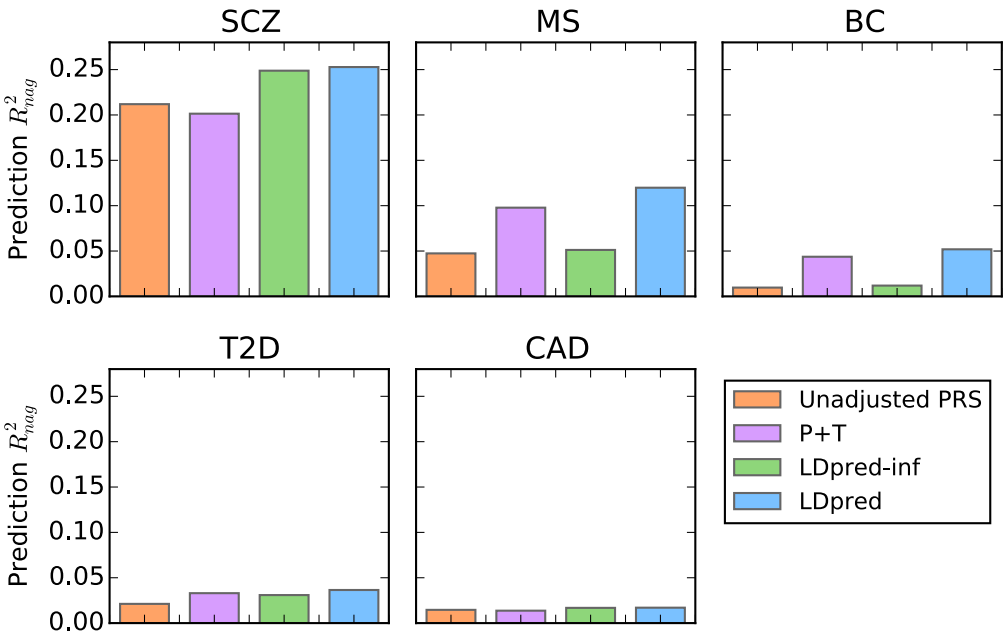
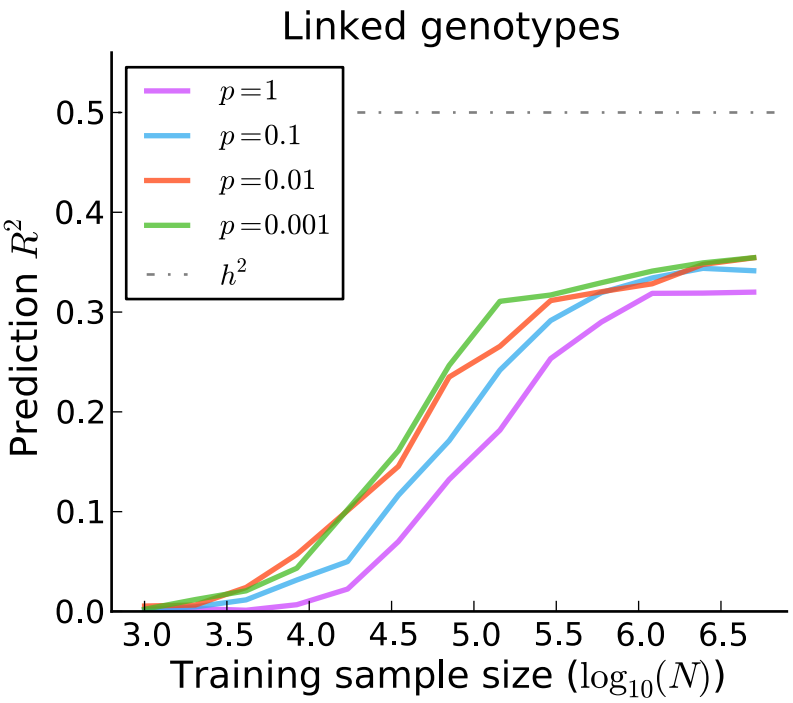


Figure 4: Comparison of prediction accuracy for 5 different diseases, schizophrenia (SCZ), multiple sclerosis (MS), breast cancer (BC), type-2 diabetes (T2D), and coronary artery disease (CAD). The risk scores were trained using large GWAS summary statistics data sets and used to predict in independent 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)**.

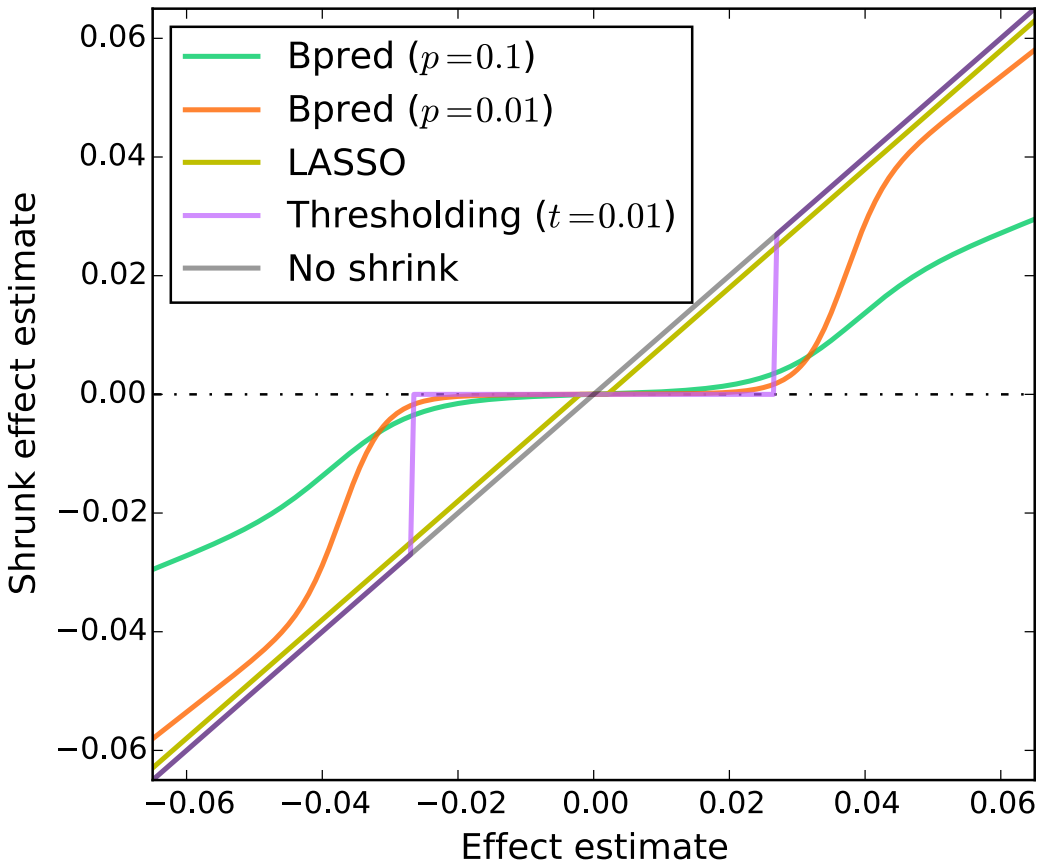
Prediction Method	Accounts for LD?	Accounts for non-infinitesimal genetic architecture?	Comments
Unadjusted polygenic risk score	No.	No.	
LD-pruning followed by *P*-value thresholding (P+T)	Yes*.	Yes.	A heuristic that discards information from pruned and thresholded markers.
LDpred-inf	Yes.	No.	An analytical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

Table 1: 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.)

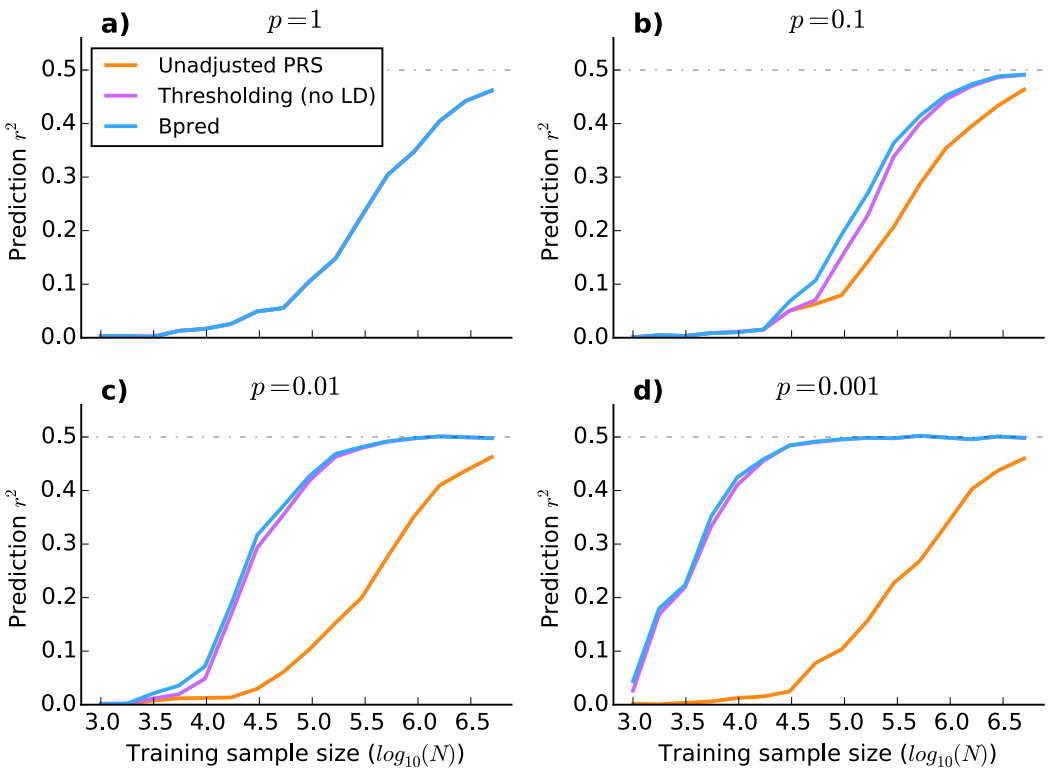
Supplementary Figures and Tables



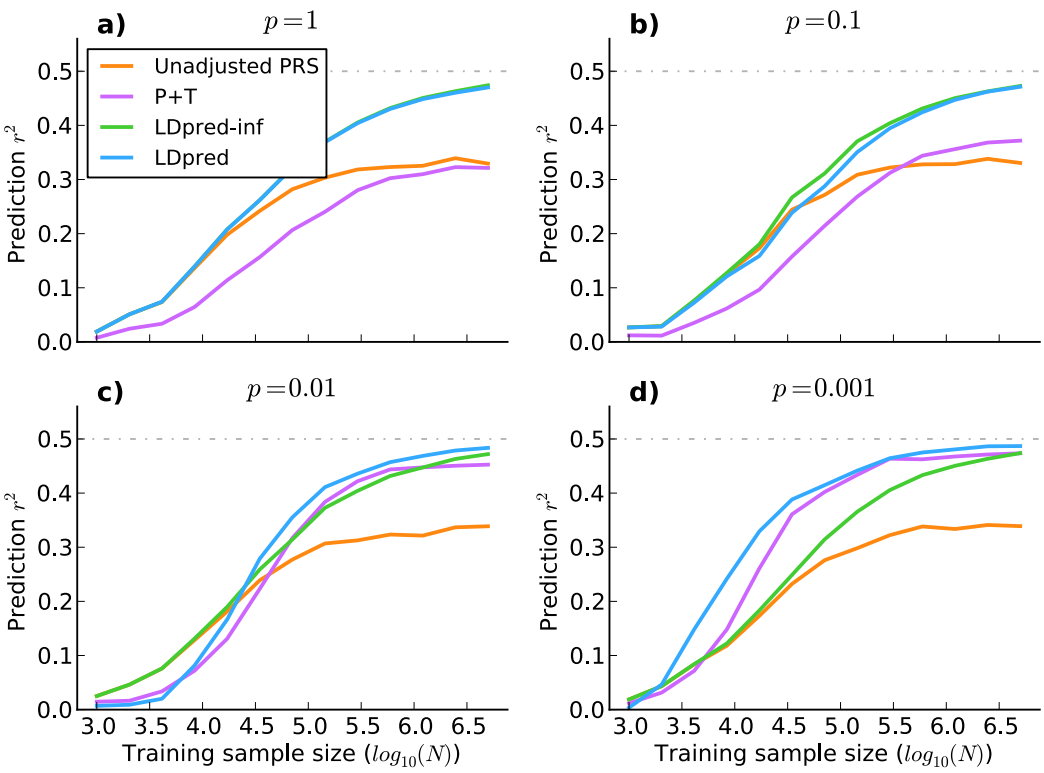
Supplementary Figure 1. Performance of P+T (LD-pruning followed by thresholding) for an alternative genetic architecture where causal markers cluster. The results are averaged over 3000 simulated traits with 200K simulated genotypes where the average fraction of causal variants *p* was let vary. The 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 *p* parameter from a Beta(*p*,*1-p*) distribution. 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*.



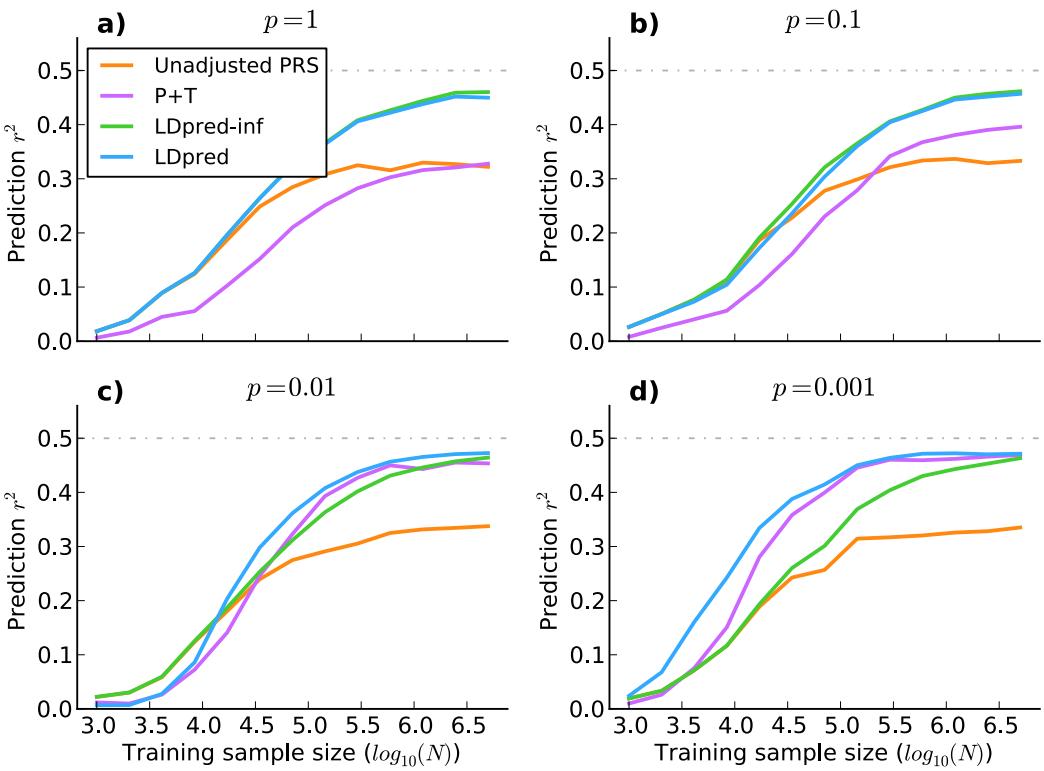
Supplementary Figure 2. Comparison of five different shrinks in the absence of LD. Bpred corresponds to LDpred without LD and can be derived analytically (see Online Methods and Supplementary Note for details). The marginal (least square) effect estimate is plotted against the shrunk estimate for the five different shrinks. Bpred denotes the analytical solution to LDpred, which can be derived in the absence of LD (see Supplementary Note 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.



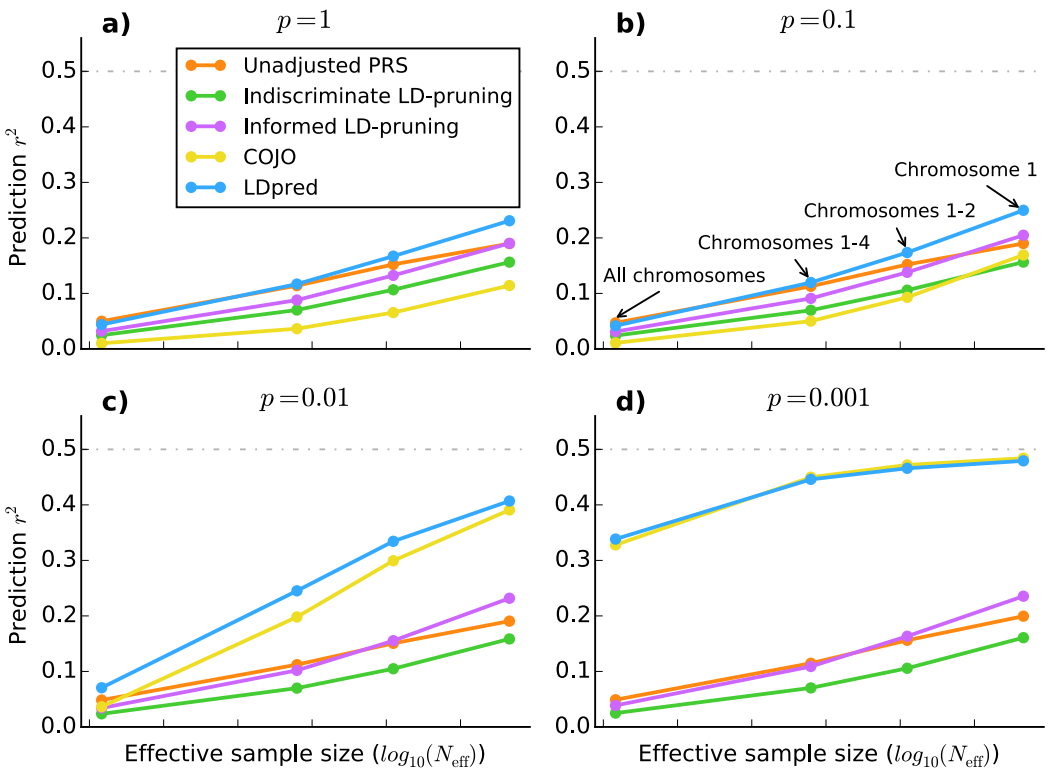
Supplementary Figure 3. Comparison of methods using simulated genotypes without LD. The four subfigures **a-d** correspond to different genetic architectures where we vary *p*, the fraction of variants with (non-zero) effects drawn from a Gaussian distribution. Bpred denotes the analytical solution to LDpred, which can be derived in the absence of LD (see Supplementary Note for details). As expected, Bpred outperforms *P*-value thresholding in the absence of LD, although not by much.



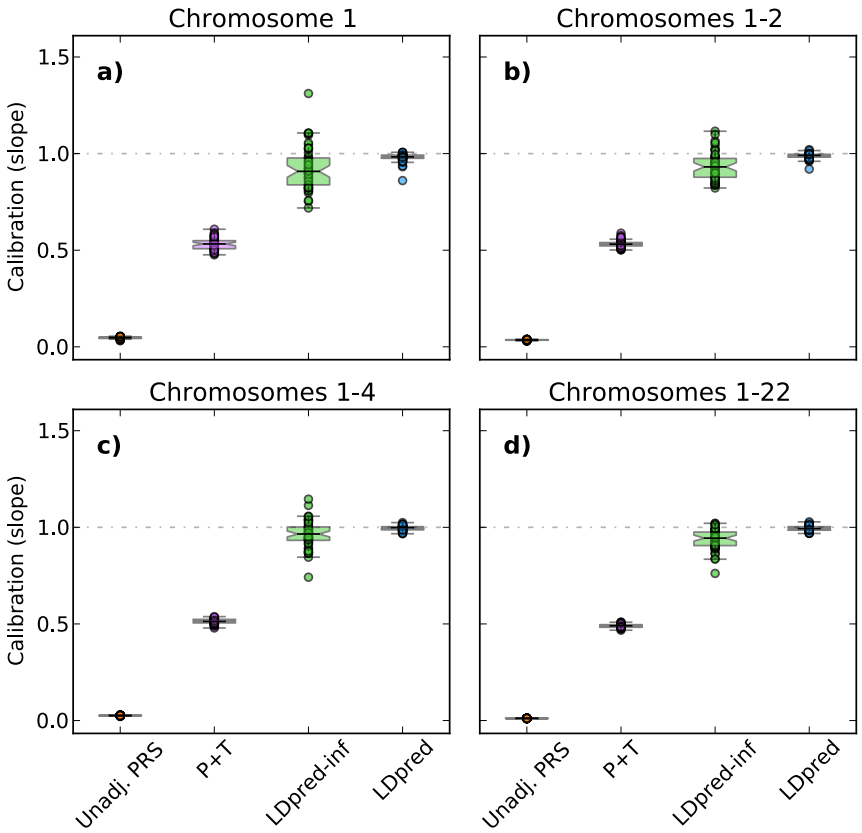
Supplementary Figure 4. Comparison of methods using simulated genotypes with LD. The four subfigures **a-d** correspond to different genetic architectures where we vary *p*, the fraction of variants with (non-zero) effects drawn from a Gaussian distribution. We simulated marginal least square effect estimates with LD (see Supplementary Note 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 2000 independent phenotype simulations keeping the simulated genotypes fixed (see Supplementary Note for details). The behavior of LDpred in the subfigure **c)** for small sample sizes is due to a LD window-size mismatch between the simulated data and the LDpred and P+T methods.



Supplementary Figure 5. Comparison of methods using simulated genotypes (see **Supplementary Figure 4.**) with LD with Laplace mixture distributed effects instead of Gaussian mixture distributed effects. The change in prior appears to have minimal effect on the shape of the curve and the relative performance.



Supplementary Figure 6. Comparisons to other methods using simulated traits and real WTCCC genotypes. As expected COJO[14,35] performs close to optimal with sufficient training data, or more precisely, when the ratio $(Nh^2)/(Mp)$ 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 $(Nh^2)/(Mp)$ 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.



Supplementary Figure 7. 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).

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

Supplementary Table 1. Numerical values of results displayed in Figure 3, 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[36] 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

Supplementary Table 2. 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[52] 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

Supplementary Table 3. 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

Supplementary Table 4. Numerical values of results displayed in Figure 4, on four different R^2 or AUC scales.

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

Supplementary Table 5. 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

Supplementary Table 6. 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

Supplementary Table 7. Numerical values of results displayed in Figure 4, 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

Supplementary Table 8. Parameters inferred or assumed by P+T and LDpred for analysis of the non-European validation samples in **Supplementary Table 7**.