Non-parametric polygenic risk prediction using partitioned GWAS summary statistics

Sung Chun[1,2,3,¶], Maxim Imakaev[1,2,3,¶], Nathan O. Stitziel[4,5,6,*], Shamil R. Sunyaev[1,2,3,7,*]

[1] Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts, 02115, USA

[2] Department of Medicine, Harvard Medical School, Boston, Massachusetts, 02115, USA

[3] Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, 02142, USA

[4] Cardiovascular Division, Department of Medicine, Washington University School of Medicine, Saint Louis, Missouri, 63110, USA

[5] Department of Genetics, Washington University School of Medicine, Saint Louis, Missouri, 63110, USA

[6] McDonnell Genome Institute, Washington University School of Medicine, Saint Louis, Missouri, 63110, USA

[7] Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, 02115, USA

¶ These authors contributed equally to this work.

* Corresponding authors

nstitziel@wustl.edu (NOS)

ssunyaev@rics.bwh.harvard.edu (SRS)

Abstract

Traditional methods for polygenic risk prediction of complex disease have historically focused on the set of genetic markers associated with disease at a genome-wide level of significance. Newer approaches have been developed to incorporate information from all markers across the genome, however these methods typically require individual level genotypes or depend on accurately specifying the underlying genetic architecture in order to optimally apply shrinkage methods on estimated effect sizes. Here, we propose a novel partitioning-based risk prediction model to achieve non-parametric shrinkage that does not require explicitly modeling the genetic architecture. We start with a set of summary statistics in the form of SNP effect sizes from a large GWAS cohort. We remove the correlation structure across summary statistics arising due to linkage disequilibrium. Then, we partition the decorrelated summary statistics and determine the appropriate shrinkage factors for each partition using an independent small cohort with individual level data. We show that the resulting non-parametric shrinkage is equivalent to applying piecewise linear interpolation on conditional mean effects. Based on simulated datasets, we show that our new non-parametric shrinkage (NPS) can reliably correct for linkage disequilibrium in summary statistics of dense 5 million genome-wide markers and considerably outperforms a state-of-the-art method across wide range of genetic architecture.

Introduction

In complex trait genetics, the ability to predict phenotype from genotype is the ultimate measure of our understanding of the allelic architecture underlying the heritability of a trait. Complete understanding of the genetic basis of a trait should allow for predictive methods having accuracies approaching the trait's broad sense heritability. In addition to improving our fundamental understanding of basic genetics, phenotypic prediction has obvious practical utility, ranging from crop and livestock applications in agriculture to estimating the genetic component of risk for common human diseases in medicine. For example, a portion of the current guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk focuses on estimating a patient's risk of developing disease [1]; in theory, genetic predictors have the potential to reveal a substantial proportion of this risk early in life (even before clinical risk factors are evident) enabling prophylactic intervention for high-risk individuals.

The field of phenotypic prediction originated in plant and animal genetics (reviewed in refs. [2] and [3]). The first approaches relied on "major genes" – allelic variants of large effect sizes readily detectable by genetic linkage or association. Similarly, after the early results of human genome-wide association studies (GWAS) became available, the first risk predictors in humans were based on combining the effects of markers significantly and reproducibly associated with the trait, typically those with association statistics exceeding a genome-wide level of significance [4,5].

More recent studies have demonstrated that a multitude of small effect alleles, not detectable at a genome-wide level of statistical significance using currently existing sample sizes, play an important role in complex trait genetics [2,3,6]. This has led to increasing popularity of highly polygenic and even infinitesimal models of the allelic architecture underlying complex traits [7,8]. Under the assumption that numerous alleles of small effect are involved in the trait, predictors based only on GWAS-variants are expected to have low accuracy because they do not include most of the DNA variants underlying the trait. New methods of phenotype prediction leverage this insight by fitting all variants simultaneously [8–13]. These methods include extensions of BLUP [10,11], or Bayesian approaches that extend both shrinkage techniques and

random effect models [8]. Newer methods benefited from allowing for classes of alleles with vastly different effect size distributions. However, these methods require individual level genotype data that do not exist for large meta-analyses and are computationally expensive.

"Polygenic scores" [7,14–17] represent an alternative approach based on summary statistics. These scores, that are generally additive over genotypes weighted by apparent effect sizes exceeding a given *p*-value threshold, have the potential to incorporate many relevant variants undetected by GWAS, albeit at the expense of including numerous variants unrelated to the trait. In theory, the risk predictor based on conditional mean effects can achieve the optimal accuracy of linear risk models regardless of underlying genetic architecture by properly down-weighting noise introduced by non-causal variants [18]. In practice, however, implementing the conditional mean predictor poses a dilemma. In order to estimate the conditional mean effects, we need to know the underlying genetic architecture first, but the true architecture is unknown and difficult to model accurately. The current methods circumvent this issue by deriving conditional means under a highly simplified model of genetic architecture. However, the optimality of resulting risk predictor is not guaranteed when the true underlying genetic architecture deviates from the assumed prior. In particular, recent studies have revealed complex dependencies of heritability on minor allele frequency (MAF) and local genomic features such as regulatory landscape and intensity of background selections [19–23]. Several studies proposed to extend polygenic scores by incorporating additional complexity into the parametric Bayesian models, however, these methods have not been demonstrated with the genome-wise set of markers due to computational challenges [17,24,25]. Recently, there have been growing interests in non-parametric and semi-parametric approaches, such as those based on modeling of latent variables and kernel-based estimation of prior or marginal distributions, however, thus far these models cannot leverage summary statistics or account for the linkage disequilibrium (LD) structure in the data [26–29].

Here, we propose a novel risk prediction approach called partitioning-based non-parametric shrinkage (NPS). Without specifying a parametric model of underlying genetic architecture, we aim to estimate the conditional mean effects directly from the data. Large GWAS

studies often provide estimated effect sizes at each SNP in a discovery cohort as summary statistics. We will show that using a second small independent training cohort with available genotype-level data, we can learn the shrinkage weights for apparent effect sizes from the discovery GWAS. In this way, we can leverage the large sample size of the discovery cohort, which is typically available only as summary statistics rather than individual genotypes. We evaluate the performance of this new approach under a simulated genetic architecture of dense 5 million SNPs across the genome.

Results

Method Overview

We start from a simple observation that conditional mean effects can be approximated by piece-wise linear interpolation in the absence of LD (Fig 1A). We can partition SNPs into *K* disjoint intervals based on observed effect sizes ($\hat{\beta}_j$) and fit a linear function $f(\hat{\beta}_j) = \omega_k \hat{\beta}_j$ on each interval of $k = 1, \ldots, K$. Specifically, when x_{ij} is the genotype of individual i at marker j and β_j is the true effect size at marker j, the predicted phenotype $\hat{y}_i$ based on conditional mean effects $E[\beta_j \mid \hat{\beta}_j]$ can be interpolated as the following:

$$\hat{y}_i = \sum_{j=1}^{M} E[\beta_j \mid \hat{\beta}_j]\, x_{ij} \approx \sum_{j=1}^{M} \left(\sum_{k=1}^{K} \omega_k \hat{\beta}_j I\left(b_{k-1} < \hat{\beta}_j \leq b_k\right) \right) x_{ij} \tag{1}$$

where b_{k-1} and b_k are partition boundaries and $I(\cdot)$ is an indicator function for partition k. This equation can be further simplified by changing the order of summation as below:

$$= \sum_{k=1}^{K} \omega_k \left(\sum_{j \in \mathcal{S}_k}^{M} \hat{\beta}_j x_{ij} \right) = \sum_{k=1}^{K} \omega_k G_{ik} \tag{2}$$

where $\mathcal{S}_k$ is the set of all markers assigned to partition k. If we define a partitioned risk score G_{ik} to be a risk score of individual i calculated using only SNPs in partition k, ω_k becomes equivalent

to the per-partition shrinkage weight. Based on equation (2), we can utilize a small genotype-level training cohort to estimate ω_k by fitting phenotypes y_i with partitioned risk scores G_{ik} of training individual i. We do not need to specify the underlying distribution of true genetic effects β_j since ω_k can be directly learned from the data.

We rely on a small independent cohort with full genotype information to estimate the shrinkage weights. For quantitative traits, the per-partition shrinkage weights can be estimated by linear regression. For binary phenotypes such as case/control status, ω_k can be derived from a linear discriminant analysis (LDA)-based classifier in the *K*-dimensional feature space formed by partitioned risk scores G_{ik}. LDA guarantees the optimal accuracy of classifier when case and control subgroups follow multivariate normal distributions in the feature space. The partitioned risk scores of cases and controls indeed follow approximately normal distributions as far as each partition consists of an enough number of SNPs [30]. Furthermore, we can assume that the covariance of partitioned risk scores approximately equal between cases and controls and is a diagonal matrix because 1) G_{ik} of a single partition explains only a small fraction of phenotypic variation on the observed scale with typical GWAS sample sizes [31] and 2) simulations suggest that covariance across partitions is negligibly small despite the liability thresholding effect. Under these conditions, LDA-derived shrinkage weights can be independently estimated for each partition and simplify to:

$$\omega_k \approx 2\frac{E[\,G_{ik}|\,y_i = 1] - E[\,G_{ik}|\,y_i = 0]}{var[\,G_{ik}|\,y_i = 1] + var[\,G_{ik}|\,y_i = 0]} \tag{3}$$

where the phenotype of an individual y_i is encoded in 0 or 1 depending on being a control or a case. Note that estimating ω_k is a far easier problem than estimating per-SNP effects. The training sample size is small but still larger than the number of parameters to estimate, namely *K*, whereas for $\hat{\beta}_j$, the GWAS sample size is considerably smaller than the number of markers.

In the presence of LD, we cannot apply piece-wise linear interpolation directly on observed effect sizes because the conditional mean effect at each SNP depends on genetic

effects of all adjacent SNPs in LD. To address the problem, we split the genome into ~2 Mbps windows, and in each window, we project SNP genotypes into a decorrelated space by a linear transformation obtained by the spectral decomposition of local LD matrix (Fig 1B, See Methods). Specifically, when $\boldsymbol{D}$ is a local reference LD matrix obtained from training genotypes and decomposes into $\boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of non-negative eigenvalues λ_j and $\boldsymbol{Q}$ is a matrix of orthonormal eigenvalues q_j, the decorrelating projection is defined as:

$$\mathcal{P} = \boldsymbol{\Lambda}^{-1/2}\,\boldsymbol{Q}^{\boldsymbol{T}} \tag{4}$$

and for the axis of projection defined by q_j, the transformation $\mathcal{P}$ yields the projected genotype x_{ij}^P and decorrelated effect estimate $\hat{\eta}_j$ as the following:

$$x_{ij}^P = \frac{1}{\sqrt{\lambda_j}} q_j^T x_i$$

and (5)

$$\hat{\eta}_j = \frac{1}{\sqrt{\lambda_j}} q_j^T \hat{\beta}$$

where x_i and $\hat{\beta}$ are *m*-dimensional vectors of genotypes and per-SNP estimated effects, respectively, at all SNPs, and *m* is the number of SNPs in the window. Under the projection, both projected genotypes and sampling noises of decorrelated effect estimates become uncorrelated across different axes of projection (Supplementary Note). Although strictly speaking, the lack of correlation does not imply independence, we apply the non-parametric shrinkage technique as if x_{ij}^P and $\hat{\eta}_j$ are from unlinked markers.

Partitioning strategy

We use a higher number of partitions to better interpolate the mean conditional effect curve. Since the full combinatorial optimization of partitioning cut-offs b_k is neither necessary or practical, we place the cut-offs based on heuristics without optimizing them on individual dataset

(See Methods, Supplementary Note and Fig S1). Note that widely-used *p*-value thresholding approach is a special case of non-parametric shrinkage with the number of partition *K* being 2. When SNP effect sizes are measured relative to standardized genotypes, thresholding on *p*-values of association is equivalent to thresholding on estimated SNP effect sizes.

In the presence of LD, non-parametric shrinkage requires double partitioning on both λ_j and $\hat{\eta}_j$ because the conditional mean effects in the decorrelated space, i.e. $E[\eta_j \mid \hat{\eta}_j]$, depend not only on $\hat{\eta}_j$ but also on eigenvalue λ_j of the projection vector. To provide intuition behind this, an eigenvalue captures how many correlated SNPs are tracked by a projection vector, and we can show that the scale of true underlying decorrelated effect sizes, namely $var[\eta_j]$, is proportional to the eigenvalue λ_j (Supplementary Note). Thus, with increasing eigenvalues, the scale of decorrelated effects becomes large relative to the sampling error $\frac{1}{N}$. Therefore, estimated effects become more reliable. On the other hand, with decreasing eigenvalues, $\hat{\eta}_j$ becomes dominated by sampling error. In addition, conditional mean effects depend on $\hat{\eta}_j$ as a function of polygenicity. If we consider genetic architecture with a low proportion of causal SNPs, some of projection vectors may not involve even a single causal SNP, and in this case, their true decorrelated effect size η_j will be 0. Such non-causal projection vectors will be enriched among small estimates of $\hat{\eta}_j$, thus of which conditional mean effects become attracted toward 0.

In the special case of infinitesimal genetic architecture in which all SNPs are causal with normally distributed effect sizes, we already know analytically derived conditional mean effects [17], therefore can cross-check the accuracy of our shrinkage weights ω_k estimated by non-parametric shrinkage (Fig 2A-B, Supplementary Note). For this, we simulated a genome consisted of 1,236 loci of 2 Mb each, of which LD structures were sampled from real human haplotypes but each locus was treated unlinked from each other for simplicity (See Methods). The ratio of discovery cohort size to number of markers were set to 1, and an independent cohort of 2,500 cases and 2,500 controls was set aside for training. To apply the non-parametric shrinkage, we first partitioned decorrelated effect sizes into 10 subgroups on intervals of eigenvalues λ_j while allocating equal amount of $\sum_j \lambda_j$ to each partition (Fig 2A). The per-partition shrinkage

weights ω_k trained by non-parametric shrinkage closely tracked the theoretical optimum in most of the bins. However, interestingly, in the lowest and highest partitions of eigenvalues, $\mathcal{S}_{\lambda 1}$ and $\mathcal{S}_{\lambda 10}$, the estimated shrinkage was significantly biased away from the optimal curve. The smallest eigenvalues are too noisy to estimate with the reference LD panel. Therefore, it is correct to down-weight ω_1 almost to 0. In case of partition $\mathcal{S}_{\lambda 10}$, it spans the widest interval of eigenvalues but consists of the fewest number of SNPs. While it is ideal to apply a finer partitioning in this interval so to better interpolate the theoretical curve, the total number of SNPs and independent projection vectors in the genome are the fundamental limiting factor.

In case of infinitesimal architecture, theory predicts that per-partition shrinkage weights depend only on eigenvalues and not on estimated effect sizes. To examine the robustness of non-parametric shrinkage, we applied general 10-by-10 double partitioning on λ_j and $\hat{\eta}_j$ collected under infinitesimal simulations. Similarly to eigenvalues, we placed partition boundaries on the interval of $\hat{\eta}_j$ keeping each partition to have equal $\sum_j \hat{\eta}_j{}^2$. This is a heuristic approach to distribute partition boundaries toward the intermediate range of $\hat{\eta}_j$ and not to over-partition the largest values of $\hat{\eta}_j$. This approach works reliably well across a wide range of simulation conditions we tested. In overall, the shrinkage weights estimated by double partitioning agree with the theoretical expectation. The estimated conditional mean effects, interpolated with $\omega_k \hat{\eta}_j$, follow the linear trajectory (Fig 2B for $\mathcal{S}_{\lambda 10}$, Fig S2 for other partitions of eigenvalues).

For non-infinitesimal genetic architecture, we do not have an analytic derivation of conditional mean effects, therefore empirically estimated the conditional means using the true underlying effects η_j and true LD structure of the population (See Methods). One percent of SNPs were simulated to be causal with normally distributed effect sizes. The ratio between number of discovery samples to number of markers was set to 1:1. As expected, the true conditional mean dips for the lowest values of $\hat{\eta}_j$ but catches up toward no shrinkage ($\omega_k = 1$) with increasing values of $\hat{\eta}_j$ (Fig 2C-D). A notable difference between the partitions of largest eigenvalues $\mathcal{S}_{\lambda 10}$ (Fig 2C) and second smallest eigenvalues $\mathcal{S}_{\lambda 2}$ (Fig 2D) is that the true conditional mean is very close to almost no shrinkage for large $\hat{\eta}_j$ in the former. This is because

the scale of true effects η_j is proportional to eigenvalues, therefore, with large enough λ_j, the sampling error becomes relatively small and the estimated effect sizes more accurate. In all partitions, conditional mean effects estimated by non-parametric shrinkage stayed very close to the true conditional means (Fig S3 for eigenvalue partitions other than $\mathcal{S}_{\lambda 2}$ and $\mathcal{S}_{\lambda 10}$). The observed consistency between estimated and true conditional means is not because of shared underlying data. At each simulation run, we re-generated the entire dataset starting from fresh sampling of true underlying genetic effects β_j under the same genetic architecture parameters.

Simulated benchmark

To benchmark the accuracy of non-parametric shrinkage, we simulated the genetic architecture using the real LD structure of dense 5 million common SNPs from the 1000 Genomes Project (See Methods). For baseline simulations, we considered a simple mixture-normal genetic architecture with the causal fraction of SNPs being 1% to 0.01% (50,000 to 500 SNPs). This is the model most commonly used as a Bayesian prior. With decreasing causal fractions, the accuracy of non-parametric shrinkage improved in terms of both liability-scale R^2 as well as Nagelkerke's R^2 on the observed scale (Table 1). With fewer causal SNPs, per-SNP effect sizes increase making it easier to predict the phenotype.

We incrementally added more complexity to simulated genetic architecture starting from the baseline model. First, we incorporated the dependency of heritability on minor allele frequency (MAF) into the simulation. Because of this dependency, recent studies have reported that low frequency SNPs contribute less heritability than previously expected under no dependency [20]. Low-frequency SNPs tend to be captured by eigenvectors of small eigenvalues and are challenging to handle with spectral decomposition. Lowering the overall heritability contribution of low-frequency SNPs made non-parametric shrinkage predict more accurately (Table 1).

Further, we introduced clumping of causal SNPs in open chromatins by making the fraction of causal SNPs five-fold higher in DNase I hypersensitive sites (DHS), comprising 15% of the genome, compared to the rest of genome. The prediction accuracy of non-parametric

shrinkage went down slightly compared to simulations of uniformly distributed causal SNPs but still remained robust even if we did not explicitly account for DHS overlap in the current version of non-parametric shrinkage (Table 1).

We evaluated the performance of non-parametric shrinkage vis-a-vis a comparable successful parametric technique (Table 1). LDPred is the state-of-the-art parametric method, which is similarly based on summary statistics estimated in large GWAS datasets and an independent training set with individual-level data. As a Bayesian method, it is optimal for the spike and slab allelic architecture. We found that our method resulted in more accurate predictions than LDPred across a range of genome-wide simulations. This is seemingly surprising given that the simulated allelic architectures are close to the LDPred assumptions. However, we found that in most simulations, LDPred adopted the infinitesimal model irrespectively to the true simulated regime pointing to the challenge of the computational optimization in the parametric case. Notably, in a single simulation, where LDPred correctly selected a non-infinitesimal model with a low causal fraction (0.01%), it achieved high prediction accuracy. This is reflected in Table 1 as a high confidence interval in one of the LDPred simulation sets. The simulations suggest that the well optimized parametric model is capable to generate good predictions, but non-parametric shrinkage is much more robust and does not suffer from optimization issues. This results in a higher performance of NPS in practice. Overall, in our most realistic simulations incorporating DHS clumping and MAF dependency, our method showed 1.20 and 1.38-fold higher accuracy in Nagelkerke's R^2 compared to LDPred for causal fractions of 1% and 0.1% (P = 0.0007 and 0.001, respectively).

Discussion

Understanding how phenotype maps to genotype has always been a central question of basic genetics. With the explosive growth in the amount of training data, there is also a clear prospect and enthusiasm for clinical applications of the polygenic risk prediction [32,33]. The current reality is, however, that most large-scale GWAS datasets are available in the form of summary statistics only. Nonetheless, data on a limited number of cases are frequently available

from epidemiological cohorts such as UK Biobank or from public repositories with a secured access such as dbGaP. This motivated us to develop a method that is primarily based on summary statistics but also benefits from smaller training data at the raw genotype resolution. Although we heavily rely on the training data to construct a prediction model, the requirement for out-of-sample training data is not unique for our method. Widely-used thresholding-based polygenic scores and Bayesian parametric methods also need genotype-level data to optimize their model parameters [17,34].

Human phenotypes vary in the degree of polygenicity [22], in the fraction of heritability attributable to low-frequency variants [20] and in other aspects of allelic architecture [21,23]. Our method is agnostic with respect to the underlying allelic architecture. This makes it potentially equally applicable to a wide range of phenotypes.

A stream of recent publications point to a disproportionate contribution to heritability of protein coding genes and regulatory regions highlighted by chromatin accessibility and epigenetic modifications [19,35–38]. This heritability enrichment is often observed to be cell-type specific. While in practice, we do not have the comprehensive annotations of functional genomics data relevant to the phenotype of interest, incorporating available external annotations into the model could increase the accuracy of polygenic risk prediction [24,25]. Although we do not explore this direction in the present work, our method opens a clear perspective for the incorporation of the functional data. The current scheme partitions summary statistics only based on the apparent effect size, however, the same partitioning strategy can be easily applied to any functional genomics variable.

Similarly to other methods, our method assumes that all datasets come from a homogeneous population. It is well appreciated that polygenic risk models are not transferrable between populations due to differences in allele frequencies and patterns of linkage disequilibrium [39]. This problem should be addressed by the future work in the field.

Materials and Methods

Non-parametric shrinkage

The estimated effect sizes $\hat{\beta}_j$ at SNPs j = 1, …, *M* are available as summary statistics from a large discovery GWAS study of *N* samples. When the estimated effects were prepared as per-allele effects, we converted them relative to standardized genotypes by multiplying $\sqrt{2f_j(1-f_j)}$, where f_j is the allele frequency of SNP j in the discovery GWAS cohort. The logistic per-allele effect sizes obtained from case/control GWAS were treated in the same way. In addition, non-parametric shrinkage (NPS) requires a small independent genotype-level training cohort in order to build the prediction model.

We processed the effect size estimates of discovery GWAS and training cohort of sample size N' in non-overlapping windows of 4,000 SNPs each (~2.4 Mb in length). In each window, given an N' × 4,000 standardized genotype matrix $\boldsymbol{X}$, the raw reference LD matrix $\boldsymbol{D} = \frac{1}{N'}\boldsymbol{X}^T\boldsymbol{X}$ was regularized in order to suppress sampling noise particularly in off-diagonal entries. Specifically, pairwise LD was set to 0 if the SNPs were separated by > 500 kb or the absolute value of estimated LD was smaller than $5/\sqrt{N'}$. Since the standard error of pairwise LD is approximately $1/\sqrt{N'}$ under no correlation, we expect that on average, only 1.7 uncorrelated SNP pairs escape the above regularization threshold per window. The regularized LD matrix $\boldsymbol{D}^*$ was factorized into the following by spectral decomposition:

$$\boldsymbol{D}^* = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^T$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues and $\boldsymbol{Q}$ is an orthonormal matrix of eigenvectors. Since $\boldsymbol{D}^*$ is not non-negative semi-definite, $\boldsymbol{\Lambda}$ contained negative eigenvalues. Negative eigenvalues were truncated along with those positive but smaller than 0.5 since they were dominated by noise. Applying the decorrelating linear transformation $\mathcal{P}$ (equation 4), we obtained decorrelated genotypes x_{ij}^{P} and decorrelated effect size estimates $\hat{\eta}_j$ for each projection defined by eigenvalue λ_j and corresponding eigenvector q_j (equation 5), where i is an individual and j is the index for decorrelating projection.

Although we chose the window size to be large enough to capture the majority of local LD patterns, some LD structures, particularly near the edge, span across windows, which in turn yield residual cross-window correlations. To eliminate such correlations, we applied LD pruning in decorrelated space between adjacent windows. Specifically, we calculated Pearson correlations $\rho_{jj'}$ between decorrelated genotypes x_j^P and $x_{j'}^P$, where j and j' are the axes of projection, belonging to distinct windows. For the pairs with $|\rho_{jj'}| > 0.3$, we kept the one yielding larger decorrelated effect sizes and eliminated the other.

Next, we merged decorrelated effect sizes $\hat{\eta}_j$ across all windows and defined the 10 × 10 double-partitioning boundaries on intervals of λ_j and $|\hat{\eta}_j|$. The eigenvalues were split to 10 intervals of λ_j, equally distributing $\sum_j \lambda_j$ across partitions. The partitions on eigenvalues are denoted here by $\mathcal{S}_{\lambda 1}$, …, $\mathcal{S}_{\lambda 10}$ from the lowest to the highest. Each partition of eigenvalues was sub-partitioned on intervals of $|\hat{\eta}_j|$, equally distributing $\sum_j \hat{\eta}_j^2$ across partitions. The partition boundaries of $|\hat{\eta}_j|$ were defined separately for each partition of eigenvalues because the distribution of $|\hat{\eta}_j|$ is dependent on λ_j.

In each sub-partition k, we calculated a partitioned risk score G_{ik} of training individual i as the following:

$$G_{ik} = \sum_{j \in \mathcal{S}_k} \hat{\eta}_j \, x_{ij}^P$$

Given the phenotype y_i of each individual in training cohort, we estimated per-partition shrinkage weights ω_k by linear discriminant analysis (LDA) using the equation (3). After re-weighting each $\hat{\eta}_j$ by shrinkage weight of corresponding partition, we converted the decorrelated effect sizes back to the original SNP effect space. This back-conversion has to be done window-by-window using all $\hat{\eta}_j$ belonging to each window. Specifically, when a given window has m decorrelated effects, $\hat{\eta}_{j_o}$, $\hat{\eta}_{j_o+1}$, …, $\hat{\eta}_{j_o+m-1}$, we applied the following back-transformation:

$$\boldsymbol{Q}\,\boldsymbol{\Lambda}^{-1/2} \begin{pmatrix} \omega_{k(j_0)}\,\hat{\eta}_{j_o} \\ \vdots \\ \omega_{k(j_0+m-1)}\,\hat{\eta}_{j_o+m-1} \end{pmatrix}$$

where $k(j)$ stands for the partition such that $j \in \mathcal{S}_k$ (Supplementary Note). The resulting re-weighted SNP effect sizes is relative to standardized genotypes and can be converted to per-allele effects by correcting for allele frequencies.

Because the accuracy of decorrelating projection declines near the edge of windows, the overall performance of NPS is affected by the placement of window boundaries relative to locations of strong genome-wide association peaks. To alleviate such dependency, we repeated the same NPS procedure shifting by 1,000, 2,000, and 3,000 SNPs and took the average reweighted effect sizes across four NPS runs.

Simulation with LD confined to each locus

To show that shrinkage weights estimated by NPS approximate conditional mean effects in the decorrelated space, we simulated genetic architecture with SNPs in LD. The LD matrix was calculated using the genotypes of the 1000 Genomes Project CEU panel (2n=198) for a total of 101,296 SNPs, which were obtained by 10-fold down-sampling of Illumina HumanHap550 genotyping SNPs. SNPs with minor allele frequency (MAF) < 5% were filtered out. The genome was broken down to 2 Mb loci, and SNP-poor loci with less than 40 SNPs were excluded. Overall, a total of 1,236 loci with 82 SNPs on average were used for this simulation. Since the raw LD matrix was calculated with the reference LD panel of a small sample size, we suppressed spurious long-range LD by setting the LD between SNPs separated by > 500 kb to 0. For simplicity, we confined LD structure to each locus and disallowed LD spanning across loci.

We considered two genetic architectures: infinitesimal model with normally distributed effect sizes and non-infinitesimal model for which 1% SNPs are casual with normally distributed effect sizes. The discovery GWAS summary statistics were directly sampled from the following *m*-dimensional multivariate normal distribution one locus each time:

$$\hat{\beta} \sim N(\mu = \beta \boldsymbol{D}, \Sigma = \frac{1}{N}\boldsymbol{D})$$

where $\hat{\beta}$ and β are *m*-dimensional vectors of true and estimated effect sizes of the locus, $\boldsymbol{D}$ is a local LD matrix, and *N* is the discovery GWAS sample size. *N* was set to equal to the genome-

wide number of markers *M*. The standardized genotypes of training cohort were also generated using a multivariate normal distribution:

$$X_i \sim N(\mu = 0, \Sigma = \boldsymbol{D})$$

where X_i is an *m*-dimensional genotype vector of individual i. We generated genotypes for 50,000 individuals and simulated phenotypes under a liability threshold model with the heritability h^2 of 0.5 and prevalence of 5%. By down-sampling controls, we assembled a training case/control cohort of 2,500 cases and 2,500 controls.

For NPS, each locus was treated as a single analysis window without averaging over sliding windows. The true underlying LD matrix ***D*** was hidden, and NPS had to estimate the reference LD from the training cohort. For infinitesimal model, the theoretically optimal shrinkage weight ω_k^o is known (Supplementary Note) [17]:

$$\omega_k^o = \lambda_j / (\lambda_j + \frac{M}{Nh^2})$$

For non-infinitesimal model, we do not have analytically known optimal shrinkage, therefore instead empirically estimated it by regressing conditional mean effects against $\hat{\eta}_j$ with the fixed intercept of 0 as follows:

$$E\left[\, \eta_j \mid \hat{\eta}_j \in \mathcal{S}_k \right] \sim \omega_k^o \, \hat{\eta}_j + 0$$

Then, we took the average of ω_k^o over 40 runs of simulations under the same genetic architecture parameters. $E\left[\, \eta_j \mid \hat{\eta}_j \in \mathcal{S}_k \right]$ was estimated by taking the sample mean of true decorrelated effects η_j in each partition $\mathcal{S}_k$. We calculated η_j by:

$$\eta_j = \sqrt{\lambda_j}\, \boldsymbol{q}_j^T \beta$$

using the true genetic effects β and spectral decomposition of true population LD matrix. The shrinkage weights estimated by NPS are not in the same absolute scale as theoretical values. Thus, we rescaled NPS estimates to the same scale as theoretical values by multiplying by a global scaling factor. The accuracy of prediction model is not affected by global rescaling. For infinitesimal model, the global scaling factor was $\sqrt{\sum_j (\omega_k^o \, \hat{\eta}_j)^2 \,/\, \sum_j (\omega_k \, \hat{\eta}_j)^2}$. For non-infinitesimal

model, it was obtained by regressing ω_k^o against estimated ω_k with fixed intercept of 0 by fitting $\omega_k^o \sim \omega_k + 0$ across 100 overall partitions.

Simulation of genetic architecture with dense genome-wide markers

For simulated benchmarks, we generated genetic architecture with dense ~5 million genome-wide markers from the 1000 Genomes Project. We kept only SNPs with MAF > 5% and Hardy-Weinberg equilibrium test *p*-value > 0.001. We used EUR panel (2n=808) to populate LD structures in simulated genetic data. Due to the limited sample size of LD panel, we regularized the LD matrix by Schur product with a tapered banding matrix [40]. Briefly, we preserved the raw pairwise LD for SNP pairs less than 150 kb apart, and then proportionally tapered them to 0 for SNP pairs up to 300 kb apart, and finally set to 0 for distances over 300 kb. The resulting regularized LD matrix $\boldsymbol{D}^*$ is not positive semi-definite. Thus, we approximated it to a positive semi-definite LD matrix $\boldsymbol{D}^o$ constructed by:

$$\boldsymbol{D}^o = \boldsymbol{Q}\,\boldsymbol{\Lambda}'\boldsymbol{Q}^T + 0.01\,\boldsymbol{I}$$

where $\boldsymbol{D}^* = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^T$ by spectral decomposition and $\boldsymbol{\Lambda}'$ is the matrix of eigenvalues $\boldsymbol{\Lambda}$ with negative values masked out to 0.

Next, we generated genotypes across the entire genome simulating the genome-wide patterns of LD. We assume that the standardized genotypes follow a multivariate normal distribution (MVN). Since we assume that LD travels no farther than 300 kb, as far as we simulate genotypes in blocks of length greater than 300kb, we can simulate the entire chromosome without losing any LD patterns by utilizing a conditional multivariate normal distribution as the following. The genotypes for the first block of 2,500 SNPs were sampled directly out of $N(\mu = 0, \Sigma = \boldsymbol{D}_{(1)}^o)$. From the next block, we sampled the genotypes of 1,250 SNPs (average 0.75 Mb in length) each, conditional on the genotypes of previous 1,250 SNPs. When the genotype of block l is x_l and the LD matrix spanning block l and $l+1$ is split into submatrices as the following:

$$\begin{pmatrix} \boldsymbol{D}_{(l)}^o & \boldsymbol{D}_{l,l+1}^o \\ \boldsymbol{D}_{l+1,l}^o & \boldsymbol{D}_{(l+1)}^o \end{pmatrix}$$

then, the genotype of next block $l+1$ follows a conditional multivariate normal distribution as:

$$\mathrm{X}_{l+1}|X_l = \mathrm{x}_l \sim N\left(\mu = \boldsymbol{D}^o_{l+1,l}(\boldsymbol{D}^o_l)^{-1}\mathrm{x}_l, \Sigma = \boldsymbol{D}^o_{(l+1)} - \boldsymbol{D}^o_{l+1,l}(\boldsymbol{D}^o_l)^{-1}\boldsymbol{D}^o_{l,l+1}\right)$$

After the genotype of entire chromosome was generated in this way, the standardized genotype values were converted to allelic genotypes by taking the highest Nf_j^2 and lowest $N(1-f_j)^2$ genotypes as homozygotes and the rest as heterozygotes under the assumption of Hardy-Weinberg equilibrium. N is the number of simulated samples, and f_j is the allele frequency of SNP j.

We simulated three different sets of genetic architecture: mixture normal, MAF dependency and DNase I sensitive sites (DHS). The mixture normal is a spike-and-slab architecture in which a fraction of SNPs have normally distributed causal effects β_j as below:

$$\beta_j \sim pN(0,1) + (1-p)\delta_0$$

where p is the fraction of causal SNPs (1, 0.1 or 0.01%) and δ_0 is a point mass at the effect size of 0. For the MAF-dependent model, we extended the mixture normal so that the scale of causal effect sizes varied across SNPs in proportion to $\left(f_j(1-f_j)\right)^\alpha$ with $\alpha = -0.25$ [20] as follows:

$$\beta_j \sim p\,N\left(0, \left(f_j(1-f_j)\right)^\alpha\right) + (1-p)\delta_0$$

Finally, for the DHS model, we further extended the MAF-dependent architecture to exhibit clumping of causal SNPs within DHS peaks. Fifteen per cents of simulated SNPs were located in the master DHS sites that we downloaded from the ENCODE project. We assumed a five-fold higher causal fraction in DHS (p_{DHS}) compared to the rest of genome in order to simulate the enrichment of per-SNP heritability in DHS reported in the previous study [19]. Specifically, β_j was sampled from the following distribution:

$$\beta_j \sim \begin{cases} p_{DHS}\,N\left(0, \left(f_j(1-f_j)\right)^\alpha\right) + (1-p_{DHS})\delta_0 & \text{if SNP } j \text{ is in DHS} \\ \frac{1}{5}\,p_{DHS}N\left(0, \left(f_j(1-f_j)\right)^\alpha\right) + \left(1-\frac{1}{5}p_{DHS}\right)\delta_0 & \text{otherwise} \end{cases}$$

In each genetic architecture, we simulated phenotypes for discovery, training and validation populations of 100,000, 50,000 and 50,000 samples, respectively, using a liability threshold model of the heritability of 0.5 and prevalence of 0.05. In the discovery population, we obtained GWAS summary statistics by testing for the association with the total liability instead of

case/control status. This is computationally easier than to generate a case/control GWAS cohort of a realistic sample size because of the low prevalence. With the prevalence of 0.05, statistical power of quantitative trait association studies using the total liability is roughly similar to those of dichotomized case/control GWAS studies of same sample sizes [41]. For the training dataset, we assembled a cohort of 2,500 cases and 2,500 controls by down-sampling controls out of the simulated population of 50,000 samples. The validation population was used to evaluate the accuracy of prediction model in terms of R^2 of the liability explained and Nagelkerke's R^2 to explain case/control outcomes.

Benchmark evaluation of LDPred

We benchmarked LDPred on the exactly same datasets tested for NPS. The window size was set to 1,670 SNPs in order to split the genome into 3,000 windows as recommended [17]. LDPred requires to optimize the causal fraction parameter using the training cohort available as individual-level genotype data. We tested the casual SNP fractions of 1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 0.0003 and 0.0001 and selected the model parameter yielding the highest prediction accuracy in the training cohort. This model was benchmarked in the validation cohort. For fair comparison, we did not generate complementary alleles in simulations since they were filtered out by LDPred automatically.

References

1. Stone NJ, Robinson JG, Lichtenstein AH, Bairey Merz CN, Blum CB, Eckel RH, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129: S1-45. doi:10.1161/01.cir.0000437738.63853.7a
2. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10: 381–391. doi:10.1038/nrg2575
3. Falke KC, Glander S, He F, Hu J, de Meaux J, Schmitz G. The spectrum of mutations controlling complex traits and the genetics of fitness in plants. Curr Opin Genet Dev. 2013;23: 665–671. doi:10.1016/j.gde.2013.10.006
4. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet. 2010/10/26. 2010;376: 1393–1400. doi:S0140-6736(10)61267-6 [pii]10.1016/S0140-6736(10)61267-6
5. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, et al. Performance of common genetic variants in breast-cancer risk models. N Engl J Med. 2010; doi:10.1056/NEJMoa0907727
6. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. Nature Publishing Group; 2010;42: 565–9. doi:10.1038/ng.608
7. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 2013;45: 400–5, 405e1–3. doi:10.1038/ng.2579
8. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. Visscher PM, editor. PLoS Genet. Public Library of Science; 2013;9: e1003264. doi:10.1371/journal.pgen.1003264
9. Abraham G, Tye-Din JA, Bhalala OG, Kowalczyk A, Zobel J, Inouye M. Accurate and

robust genomic prediction of celiac disease using statistical learning. PLoS Genet. 2014;10: e1004137. doi:10.1371/journal.pgen.1004137

10. Golan D, Rosset S. Effective Genetic-Risk Prediction Using Mixed Models. Am J Hum Genet. Elsevier; 2014;95: 383–393. doi:10.1016/j.ajhg.2014.09.007

11. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24: 1550–1557. doi:10.1101/gr.169375.113

12. Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, et al. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS One. 2014;9: e93017. doi:10.1371/journal.pone.0093017

13. Weissbrod O, Geiger D, Rosset S. Multikernel linear mixed models for complex phenotype prediction. Genome Res. 2016;26: 969–979. doi:10.1101/gr.201996.115

14. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460: 748–52. doi:10.1038/nature08185

15. Stahl E a, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet. Nature Publishing Group; 2012;44: 483–9. doi:10.1038/ng.2232

16. Shi J, Park JH, Duan J, Berndt ST, Moy W, Yu K, et al. Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data. PLoS Genet. 2016;12: e1006493. doi:10.1371/journal.pgen.1006493

17. Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97: 576–592. doi:10.1016/j.ajhg.2015.09.001

18. Goddard ME, Wray NR, Verbyla K, Visscher PM. Estimating Effects and Making Predictions from Genome-Wide Marker Data. Stat Sci. 2009;24: 517–529. doi:10.1214/09-STS306

19. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning

heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. Elsevier; 2014;95: 535–52. doi:10.1016/j.ajhg.2014.10.004

20. Speed D, Cai N, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. Nat Genet. Nature Publishing Group; 2017;49: 986–992. doi:10.1038/ng.3865

21. Gazal S, Finucane HK, Furlotte NA, Loh PR, Palamara PF, Liu X, et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat Genet. Nature Publishing Group; 2017;49: 1421–1427. doi:10.1038/ng.3954

22. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. Elsevier; 2017;169: 1177–1186. doi:10.1016/j.cell.2017.05.038

23. Zeng J, de Vlaming R, Wu Y, Robinson MR, Lloyd-Jones LR, Yengo L, et al. Signatures of negative selection in the genetic architecture of human complex traits. Nat Genet. Nature Publishing Group; 2018;50: 746–753. doi:10.1038/s41588-018-0101-4

24. Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput Biol. 2017;13: 1–16. doi:10.1371/journal.pcbi.1005589

25. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS Genet. 2017;13: 1–22. doi:10.1371/journal.pgen.1006836

26. Zeng P, Zhou X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat Commun. Springer US; 2017;8: 1–11. doi:10.1038/s41467-017-00470-2

27. Efron B. Empirical bayes estimates for large-scale prediction problems. J Am Stat Assoc. 2009;104: 1015–1028. doi:10.1198/jasa.2009.tm08523

28. So HC, Sham PC. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Sci Rep. 2017;7: 1–11. doi:10.1038/srep41262

29. Gianola D, Fernando RL, Stella A. Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. Genetics. 2006;173: 1761–1776.

doi:10.1534/genetics.105.049510

30. Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. Schork NJ, editor. PLoS Genet. Public Library of Science; 2010;6: e1000864. doi:10.1371/journal.pgen.1000864

31. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9: e1003348. doi:10.1371/journal.pgen.1003348

32. Khera A V, Chaffin M, Aragam KG, Emdin CA, Klarin D, Haas M, et al. Genome - wide polygenic score to identify a monogenic risk - equivalent for coronary disease. bioRxiv. 2017; doi:10.1101/218388

33. Riglin L, Collishaw S, Richards A, Thapar AK, Maughan B, O'Donovan MC, et al. Schizophrenia risk alleles and neurodevelopmental outcomes in childhood: a population-based cohort study. The Lancet Psychiatry. Elsevier Ltd; 2017;4: 57–62. doi:10.1016/S2215-0366(16)30406-0

34. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. Nature Publishing Group; 2013;14: 507–15. doi:10.1038/nrg3457

35. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–5. doi:10.1126/science.1222794

36. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. Nature Publishing Group; 2014; doi:10.1038/nature13835

37. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. Nature Publishing Group; 2013;45: 124–30. doi:10.1038/ng.2504

38. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47: 1228–1235. doi:10.1038/ng.3404

39. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet. Elsevier; 2017;100: 635–649. doi:10.1016/j.ajhg.2017.03.004

40. Cai TT, Zhang CH, Zhou HH. Optimal rates of convergence for covariance matrix estimation. Ann Stat. 2010;38: 2118–2144. doi:10.1214/09-AOS752

41. Yang J, Wray NR, Visscher PM. Comparing apples and oranges: Equating the power of case-control and quantitative trait association studies. Genet Epidemiol. 2010;34: 254–257. doi:10.1002/gepi.20456

1 **Table 1. Simulated benchmark comparison of non-parametric shrinkage (NPS) and LDPred.**

Simulated genetic architecture	Fraction of causal SNPs	Method	R^2 Liability	R^2 Nagelkerke	Fold improvement over LDpred	
					R^2_L	R^2_{Nag}
Mixture normal	1 %	NPS	0.122	0.081	1.12	1.10
	0.1 %		0.139	0.101	1.35	1.34
	0.01 %		0.234	0.162	2.49	2.54
	1 %	LDPred	0.109	0.074		
	0.1 %		0.103	0.075		
	0.01 %		0.094	0.064		
+ MAF dependency α = -0.25	1 %	NPS	0.131	0.095	1.28	1.25
	0.1 %		0.161	0.113	1.31	1.30
	0.01 %		0.247	0.170	1.93	1.89
	1 %	LDPred	0.102	0.077		
	0.1 %		0.123	0.087		
	0.01 %		0.128	0.090		
+ 5-fold higher causal fraction in DHS	1% (3 in DHS, 0.6 outside)	NPS	0.122 +/- 0.003	0.082 +/- 0.006	**1.19 ***	**1.20 ***
	0.1% (0.3 in DHS, 0.06 outside)		0.153 +/- 0.013	0.109 +/- 0.009	**1.38 ***	**1.38 ***
	0.01% (0.03 in DHS, 0.006 outside)		0.241 +/- 0.004	0.167 +/- 0.003	1.10	1.09
	1% (3 in DHS, 0.6 outside)	LDPred	0.103 +/- 0.006	0.068 +/- 0.008		
	0.1% (0.3 in DHS, 0.06 outside)		0.111 +/- 0.012	0.079 +/- 0.011		
	0.01% (0.03 in DHS, 0.006 outside)		0.219 +/- 0.332	0.153 +/- 0.247		

2
3 Non-parametric shrinkage (NPS) is more robust and accurate than the state-of-the-art parametric method (LDPred). The number of markers was 5,012,500.
4 The GWAS sample size was 100,000. Prediction models were optimized in the training cohort of 2,500 cases and 2,500 controls. R^2 of prediction was
5 measured in the validation cohort of 50,000 samples. The 95% confidence intervals were estimated from three replicate simulation runs. Simulations of mixture
6 normal and MAF dependency were not replicated. The star(*) indicates a significant improvement in R^2 (paired t-test).

Figure Legends

Figure 1. Overview of non-parametric shrinkage (NPS).

(A) For unlinked markers, NPS partitions the estimated effect sizes ($\hat{\beta}_j$) into *K* subgroups at cut-offs of $b_0, b_1, \ldots, b_K$. Partitioned risk scores G_{ik} are calculated for each partition k and individual i using an independent genotype-level training cohort. The per-partition shrinkage weights ω_k are determined by the separation of G_{ik} between training cases and controls. (B) For markers in LD, genotypes and estimated effects are decorrelated first by a linear projection $\mathcal{P}$ in non-overlapping windows of ~ 2.4 Mb in length, and then NPS is applied to the data. The size of black dots indicates genotype frequencies in the training cohort. Before projection, genotypes are correlated between SNP 1 and 2 by LD, and thus sampling noise of estimated effects ($\hat{\beta}_j \mid \beta_j$) is also correlated between SNPs. The projection $\mathcal{P}$ neutralizes both correlation structures. The axes of projection are marked by red dashed lines.

Figure 2. Per-partition shrinkage weights estimated by non-parametric shrinkage (NPS) approximate the conditional mean effects in the decorrelated space.

(A) NPS shrinkage weights ω_k (red line) compared to the theoretical optimum (black line), $\lambda_j/(\lambda_j + \frac{M}{Nh^2})$, under infinitesimal architecture. The partition of largest eigenvalues $\mathcal{S}_{\lambda 10}$ is marked by grey box. (B) Conditional mean effects estimated by NPS (red line) in sub-partitions of $\mathcal{S}_{\lambda 10}$ by $\hat{\eta}_j$ under infinitesimal architecture. The theoretical line (black) is the average over all λ_j in $\mathcal{S}_{\lambda 10}$. (C-D) Conditional mean effects estimated by NPS (red line) in sub-partitions of $\mathcal{S}_{\lambda 10}$ (C) and $\mathcal{S}_{\lambda 2}$ (D) on intervals of $\hat{\eta}_j$ under non-infinitesimal architecture with the causal SNP fraction of 1%. The true conditional means (black) were estimated over 40 simulation runs. (All) The mean NPS shrinkage weights (red line) and their 95% CIs (red shade) were estimated from 5 replicates. Grey vertical lines indicate partitioning cut-offs. No shrinkage line (green) indicates $\omega_k = 1$. The number of markers *M* is 101,296. The discovery GWAS size *N* equals to *M*. The heritability h^2 is 0.5.

Fig 1

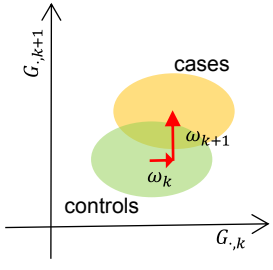
A

Estimated per-SNP effects from discovery GWAS ($\hat{\beta}_j$)

Partition SNPs into *K* subgroups:

$$\mathcal{S}_k = \{ j : b_{k-1} < |\hat{\beta}_j| < b_k \}$$

Partitioned risk scores: $G_{ik} = \sum_{j \in \mathcal{S}_k} \hat{\beta}_j x_{ij}$

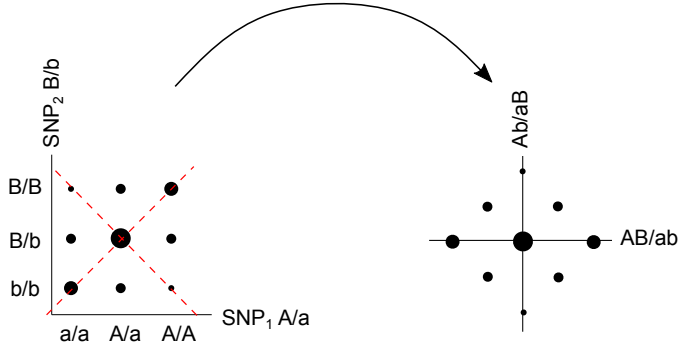


Estimate per-partition mean weights: ω_k

$$G_i = \sum_k \omega_k \sum_{j \in \mathcal{S}_k} \hat{\beta}_j x_{ij}$$

B

Decorrelating linear projection $\mathcal{P}$



Individual SNPs

genotypes X

$cov(X) = \mathbf{\Sigma}$

estimated effects $\hat{\beta}$

$cov(\hat{\beta} \mid \beta) = \frac{1}{N}\mathbf{\Sigma}$

$X^{P^T} := \mathcal{P}X^T$

$\hat{\eta} := \mathcal{P}\hat{\beta}$

Decorrelated space

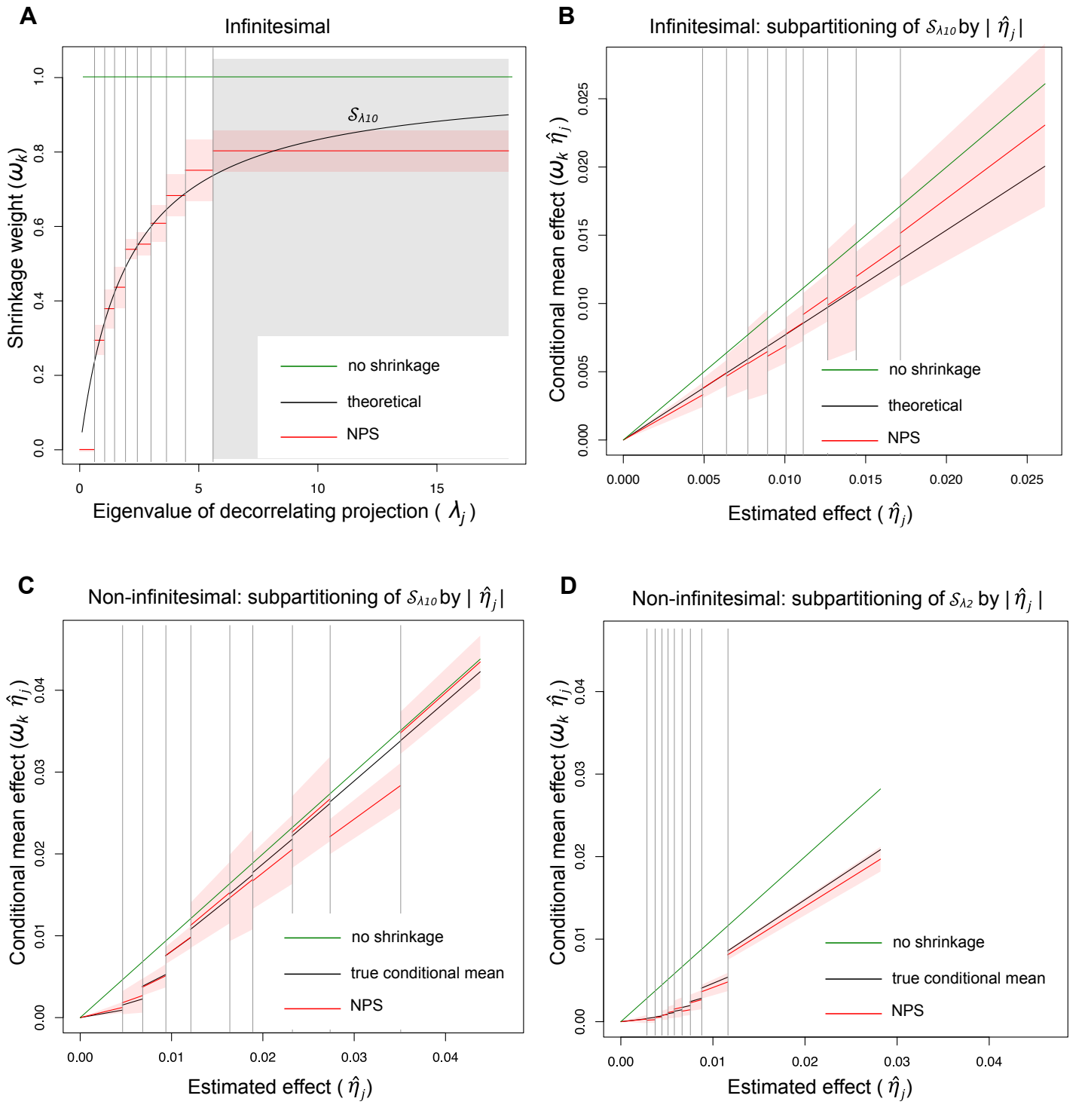
X^P

$cov(X^P) = \boldsymbol{I}$

$\hat{\eta}$

$cov(\hat{\eta} \mid \beta) = \frac{1}{N}\boldsymbol{I}$

Fig 2


A
Infinitesimal
$\mathcal{S}_{\lambda 10}$
Shrinkage weight (ω_k)
Eigenvalue of decorrelating projection (λ_j)
no shrinkage
theoretical
NPS
B
Infinitesimal: subpartitioning of $\mathcal{S}_{\lambda 10}$ by | $\hat{\eta}_j$ |
Conditional mean effect (ω_k $\hat{\eta}_j$)
Estimated effect ($\hat{\eta}_j$)
no shrinkage
theoretical
NPS
C
Non-infinitesimal: subpartitioning of $\mathcal{S}_{\lambda 10}$ by | $\hat{\eta}_j$ |
Conditional mean effect (ω_k $\hat{\eta}_j$)
Estimated effect ($\hat{\eta}_j$)
no shrinkage
true conditional mean
NPS
D
Non-infinitesimal: subpartitioning of $\mathcal{S}_{\lambda 2}$ by | $\hat{\eta}_j$ |
Conditional mean effect (ω_k $\hat{\eta}_j$)
Estimated effect ($\hat{\eta}_j$)
no shrinkage
true conditional mean
NPS

Supplementary Note for

Non-parametric polygenic risk prediction using partitioned GWAS summary statistics

Decorrelating projection

We split the genome into L non-overlapping windows of m SNPs each. An individual window is large enough to capture the majority of linkage disequilibrium (LD) patterns except near the edge. For the sake of simplicity, we assume that LD is confined to each window and there exists no LD across windows.

In genomic window l, let $\boldsymbol{X}_l$ be an $N \times m$ genotype matrix of a discovery cohort and $\boldsymbol{X}'_l$ be an $N' \times m$ genotype matrix of a training cohort. The sample sizes of discovery and training cohorts are N and N', respectively. The genotypes are standardized to the mean of 0 and variance of 1. Let $\hat{\beta}_l$ be an m-dimensional vector of apparent effect sizes at all SNPs from the discovery GWAS. $\hat{\beta}_l$ is also defined with respect to the standardized genotypes. Let β_l be an m-dimensional vector of true underlying genetic effects at all SNPs in window l. For convenience, we omit the subscript l when it is clear from the context.

The LD matrix $\boldsymbol{D}$ is given by $\boldsymbol{D} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$. Let us assume for a moment that $\boldsymbol{D}$ has full rank. In this case, $\boldsymbol{D}$ is symmetric and positive semi-definite, thus can be factorized by spectral decomposition into the following form:

$$\boldsymbol{D} = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^{\boldsymbol{T}}$$

where $\boldsymbol{Q}$ is an orthonormal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix of positive eigenvalues. The extension to rank-deficient LD matrix is straight-forward and will be discussed in later section.

When the discovery cohort has the infinite sample size, the true genetic effects β can be directly recovered from the summary statistics $\hat{\beta}$ by correcting for LD:

$$\beta = \boldsymbol{D}^{-\mathbf{1}}\,\hat{\beta} = \boldsymbol{Q}\boldsymbol{\Lambda}^{-\mathbf{1}}\boldsymbol{Q}^{\boldsymbol{T}}\hat{\beta} = \boldsymbol{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}\hat{\beta}\right) \qquad (Eq\ S1)$$

We call the linear transformation $\mathcal{P} \coloneqq \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}$ a *decorrelating projection* because of the reasons presented in the next section. The inverse transformation $\boldsymbol{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}}$ will retrieve SNP effects back from the decorrelated space defined by $\mathcal{P}$. Note that the sample size does not have to be infinite to order for $\mathcal{P}$ to be well-defined.

By applying $\mathcal{P}$ on summary statistics, we obtain *decorrelated effect size estimates* $\hat{\eta}$, and by applying $\mathcal{P}$ on genotypes, we obtain *projected genotypes* $\boldsymbol{X}'^{P}$ as follows:

$$\hat{\eta} := \mathcal{P}\hat{\beta}$$
$$(\boldsymbol{X}'^{P})^{T} := \mathcal{P}\boldsymbol{X}'^{T} \qquad (Eq\ S2)$$

Distribution of projected genotypes

Let X_i' be an m-dimensional genotype vector of individual i in window l. Then, X_i' follows the following multivariate normal distribution:

$$X_i' \sim N(\mathbf{0}, \boldsymbol{D})$$

Since the projected genotype $\mathrm{X}_i'^{P}$ is derived by multiplying $\mathcal{P}$ on X_i' by definition (Eq S2), it also follows a multivariate normal distribution. Specifically, the distribution of $\mathrm{X}_i'^{P}$ is:

$$X_i'^{P} \sim N\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}\mathbf{0}, \left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}\right)\boldsymbol{D}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}\right)^{\boldsymbol{T}}\right)$$

$$= N\left(\mathbf{0}, \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}}\right) = N(\mathbf{0}, \boldsymbol{I})$$

since $\boldsymbol{D} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\boldsymbol{T}}$ and $\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{Q} = \boldsymbol{I}$. The projected genotypes are now decorrelated with the covariance of $\boldsymbol{I}$.

Distribution of decorrelated effect size estimates

In the discovery GWAS, the estimated effect sizes $\hat{\beta}$ are calculated by linear regression as below:

$$\hat{\beta} = \frac{1}{N}\boldsymbol{X}^{T}y$$

where y is an N-dimensional phenotype vector. For convenience, we assume that y is standardized to the mean of 0 and variance of 1. At this time, we treat genotypes as fixed variables and model the genetic effects β and residuals ϵ as random. Since $y = \boldsymbol{X}\beta + \epsilon$,

$$\hat{\beta} = \frac{1}{N}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{X}\beta + \epsilon) = \boldsymbol{D}\beta + \frac{1}{N}\boldsymbol{X}^{\boldsymbol{T}}\epsilon$$

where the residual ϵ follows an N-dimensional multivariate normal distribution $N(\mathbf{0}, \sigma_e^2\boldsymbol{I})$. In an individual window, the genetic effects explain only a small fraction of phenotypic variation,

therefore $\sigma_e^2 \approx \mathrm{var(y)} = 1$. The distribution of sampling noise in $\hat{\beta}$, namely the distribution of $\hat{\beta}$ given β, follows:

$$\hat{\beta} \mid \beta \sim N(\boldsymbol{D}\beta + \frac{1}{N}\boldsymbol{X^T}\boldsymbol{0}, \frac{\sigma_e^2}{N^2}\boldsymbol{X^T}\boldsymbol{I}\boldsymbol{X})$$
$$\approx N(\boldsymbol{D}\beta, \frac{1}{N}\boldsymbol{D})$$

since $\boldsymbol{D} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$. Since the decorrelated effect size estimate $\hat{\eta}$ is a product of $\mathcal{P}$ and $\hat{\beta}$ by definition (Eq S2), the distribution of $\hat{\eta}$ given β also follows a multivariate normal distribution:

$$\hat{\eta} \mid \beta \sim N\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q^T}\boldsymbol{D}\beta, \frac{1}{N}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q^T}\boldsymbol{D}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q^T}\right)^{\boldsymbol{T}}\right)$$

$$= N\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q^T}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q^T}\beta, \frac{1}{N}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q^T}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q^T}\boldsymbol{Q}\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)$$

$$= N\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Q^T}\beta, \frac{1}{N}\boldsymbol{I}\right) \qquad (Eq\ S3)$$

since $\boldsymbol{D} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q^T}$ and $\boldsymbol{Q^T}\boldsymbol{Q} = \boldsymbol{I}$. The sampling noise of $\hat{\eta}$ is now decorrelated with the covariance of $\frac{1}{N}\boldsymbol{I}$. Hence, the projection $\mathcal{P}$ removes correlations in both genotypes and summary statistics.

Interpretation of eigenvalues

Based on Eq S3, $\hat{\eta} \mid \beta$ approaches to $\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Q^T}\beta$ as the sample size goes to the infinity. Thus, we define true decorrelated effect η as:

$$\eta \coloneqq \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Q^T}\beta \qquad (Eq\ S4)$$

Let us assume that the distribution of β is symmetric at 0 and independent at each SNP. Then,

$$E[\eta_j] = E\left[\sqrt{\lambda_j}\, q_j^T\beta\right] = \sqrt{\lambda_j}\, q_j^T E[\beta] = 0$$

and

$$var[\eta_j] = E\left[\left(\sqrt{\lambda_j}\, q_j^T\beta\right)^2\right] - E[\eta_j]^2$$

$$= \lambda_j \sum_{s=1}^{m} q_{sj}^2\, E[\beta_s^2]$$

where the j-th eigenvector q_j is $(q_{1j} \quad \cdots \quad q_{mj})^T$. Therefore, the scale of η_j, namely, $var[\eta_j]$, is proportional to eigenvalue λ_j. Furthermore, in particular cases that all SNPs have the same variance of per-SNP effect sizes σ_g^2,

$$var[\eta_j] = \lambda_j \sigma_g^2$$

since $\sum_{s=1}^{m} q_{sj}^2 = 1$. In both non-infinitesimal and infinitesimal architecture, if the genetic effects are sampled i.i.d. for each SNP uniformly across the genome, $\sigma_g^2 = var(\beta_j) = h^2/M$. M is the total number of markers, and h^2 is the heritability of the phenotype.

Conditional mean effects under infinitesimal genetic architecture

Under infinitesimal genetic architecture, the conditional mean effect has been analytically derived in [1]:

$$E[\beta \mid \hat{\beta}] = \left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{D}\right)^{-\mathbf{1}} \hat{\beta} \qquad (Eq\ S5)$$

under the assumption that $\boldsymbol{D}$ is the LD matrix of full rank. Since

$$\left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{D}\right) = \boldsymbol{Q}\left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{\Lambda}\right)\boldsymbol{Q}^{\boldsymbol{T}}$$

and

$$\left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{D}\right)^{-\mathbf{1}} = \boldsymbol{Q}\left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{\Lambda}\right)^{-\mathbf{1}}\boldsymbol{Q}^{\boldsymbol{T}}$$

we can reformulate Eq S5 as follows:

$$E[\beta \mid \hat{\beta}] = \boldsymbol{Q}\left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{\Lambda}\right)^{-\mathbf{1}}\boldsymbol{Q}^{\boldsymbol{T}}\hat{\beta} \qquad (Eq\ S6)$$

$$= \boldsymbol{Q}\left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{\Lambda}\right)^{-\mathbf{1}}\boldsymbol{\Lambda}^{\frac{1}{2}}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}\hat{\beta}\right)$$

$$= \boldsymbol{Q}\left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{\Lambda}\right)^{-\mathbf{1}}\boldsymbol{\Lambda}^{\frac{1}{2}}\hat{\eta}$$

by the definition of $\hat{\eta}$ (Eq S2). Hence,

$$E[\eta \mid \hat{\eta}] = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}E[\beta \mid \hat{\eta}] = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}E[\beta \mid \hat{\beta}]$$

$$= \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{Q}\left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{\Lambda}\right)^{-\mathbf{1}}\boldsymbol{\Lambda}^{\frac{1}{2}}\hat{\eta}$$

$$= \left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{\Lambda}\right)^{-\mathbf{1}} \boldsymbol{\Lambda}\,\hat{\eta} \qquad (Eq\ S7)$$

using the definition of η (Eq S4). For the j-th projection defined by λ_j and q_j, the conditional mean effect is given as the following:

$$E\left[\eta_j \mid \hat{\eta}_j\right] = \frac{\lambda_j}{\lambda_j + \frac{M}{Nh^2}}\hat{\eta}_j$$

Thus, in infinitesimal architecture, the conditional mean effect $E\left[\eta_j \mid \hat{\eta}_j\right]$ simplifies to $\omega\,\hat{\eta}_j$, where ω is the optimal shrinkage weight and depends only on eigenvalues as follow:

$$\omega = \frac{\lambda_j}{\lambda_j + \frac{M}{Nh^2}}$$

Back-conversion from the decorrelated space to the SNP space

Let $\hat{y}$ be an N'-dimensional vector of predicted phenotypes. We construct $\hat{y}$ by summing over all projected genotypes multiplied by conditional mean effects in the decorrelated space as follows:

$$\hat{y} = \sum_{l=1}^{L} X_l'^P E[\eta_l \mid \hat{\eta}_l]$$

where the conditional mean effect $E[\eta_l \mid \hat{\eta}_l]$ is obtained by non-parametric shrinkage. By the definition of $X_l'^P$ (Eq S2),

$$\hat{y} = \sum_{l=1}^{L} X_l' \left(\boldsymbol{\Lambda}_l^{-\frac{1}{2}}\boldsymbol{Q}_l^T\right)^T E[\eta_l \mid \hat{\eta}_l]$$

$$= \sum_{l=1}^{L} X_l' \left(\boldsymbol{Q}_l\boldsymbol{\Lambda}_l^{-\frac{1}{2}} E[\eta_l \mid \hat{\eta}_l]\right)$$

Note that X_l' is the genotype matrix in the original SNP space. Thus, $E[\eta_l \mid \hat{\eta}_l]$ can be converted back to the SNP space by the following transformation:

$$\boldsymbol{Q}_l\boldsymbol{\Lambda}_l^{-\frac{1}{2}} E[\eta_l \mid \hat{\eta}_l]$$

In the infinite sample size, $E[\eta_l \mid \hat{\eta}_l] = \hat{\eta}_l$. Hense, using Eq S1, we can recover

$$\hat{y} = \sum_{l=1}^{L} X_l' \left(\boldsymbol{Q}_l \boldsymbol{\Lambda}_l^{-\frac{1}{2}} \hat{\eta}_l \right) = \sum_{l=1}^{L} X_l' \left(\boldsymbol{Q}_l \boldsymbol{\Lambda}_l^{-\frac{1}{2}} \boldsymbol{\Lambda}_l^{-\frac{1}{2}} \boldsymbol{Q}_l^T \hat{\beta} \right) = \sum_{l=1}^{L} X_l' \left(\boldsymbol{D}^{-1} \hat{\beta} \right) = \sum_{l=1}^{L} X_l' \beta$$

as expected.

Rank deficiency of LD matrix

Even when the LD matrix $\boldsymbol{D}$ is not full rank, it is symmetric and non-negative semi-definite. In this case, spectral decomposition on $\boldsymbol{D}$ yields only r positive eigenvalues, where r is the rank of the matrix and $r < m$, and the rest of eigenvalues are 0. Without the loss of generality, we can reorder the eigenvalues and corresponding eigenvectors in such a way that only the first r eigenvalues are positive. We truncate the components corresponding to eigenvalues $r + 1, \dots, m$ and reduce the dimension to r. Specifically, the truncated matrices are defined as the following:

$$\boldsymbol{Q}' = (q_1 \quad \cdots \quad q_r)$$

$$\boldsymbol{\Lambda}' = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_r \end{pmatrix}$$

where λ_j and q_j are the j-th eigenvalues and eigenvectors, respectively. Since $\boldsymbol{Q}'$ and $\boldsymbol{\Lambda}'$ satisfy the following:

$$\boldsymbol{D} = \boldsymbol{Q}' \boldsymbol{\Lambda}' \boldsymbol{Q}'^{\boldsymbol{T}}$$

and

$$\boldsymbol{Q}'^{\boldsymbol{T}} \boldsymbol{Q}' = \boldsymbol{I}_r$$

all results we derived in the previous sections hold with $\boldsymbol{Q}'$ and $\boldsymbol{\Lambda}'$ in place of $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$, respectively.

However, the analysis of infinitesimal model (Eqs S5-S7) requires further discussion since it is non-trivial to generalize to rank deficient $\boldsymbol{D}$. Note that Eq S5 was derived under the assumption that $\boldsymbol{D}$ has full rank [1], thus cannot be used directly for rank deficient $\boldsymbol{D}$ even though the matrix $\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{D}$ is always invertible.

We re-derived the posterior mean effects from joint probability density function of $\hat{\eta}$ and η in reduced r-dimensional space. Now $E[\eta \mid \hat{\eta}]$ of Eq S7 becomes the following equation:

$$E[\eta \mid \hat{\eta}] = \left(\frac{M}{Nh^2} \boldsymbol{I}_r + \boldsymbol{\Lambda}' \right)^{-\mathbf{1}} \boldsymbol{\Lambda}' \hat{\eta}$$

Therefore,

$$E\left[\beta \mid \hat{\beta}\right] = \boldsymbol{Q}' \boldsymbol{\Lambda}'^{-\frac{1}{2}} E[\eta \mid \hat{\eta}]$$

$$= \boldsymbol{Q}'\boldsymbol{\Lambda}'^{-\frac{1}{2}}\left(\frac{M}{Nh^2}\boldsymbol{I}_r + \boldsymbol{\Lambda}'\right)^{-\mathbf{1}}\boldsymbol{\Lambda}'\left(\boldsymbol{\Lambda}'^{-\frac{1}{2}}\boldsymbol{Q}'^{\boldsymbol{T}}\hat{\beta}\right)$$
$$= \boldsymbol{Q}'\left(\frac{M}{Nh^2}\boldsymbol{I}_r + \boldsymbol{\Lambda}'\right)^{-\mathbf{1}}\boldsymbol{Q}'^{\boldsymbol{T}}\hat{\beta} \qquad (Eq\ S8)$$

Note that this result is not identical to the previous Eq S6, which was derived for full-rank $\boldsymbol{D}$:

$$E[\beta \mid \hat{\beta}] = \boldsymbol{Q}\left(\frac{M}{Nh^2}\boldsymbol{I} + \boldsymbol{\Lambda}\right)^{-\mathbf{1}}\boldsymbol{Q}^{\boldsymbol{T}}\hat{\beta} \qquad (Eq\ S6)$$

In Eq S6, M/Nh^2 term remains for $q_{r+1}, \ldots, q_m$ whereas it was truncated in Eq S8.

Empirical analysis of partitioning schemes based on simulations of unlinked markers

To explore an effective partitioning strategy for non-parametric shrinkage, we compared the prediction accuracy of different partitioning approaches under simulated genetic architecture of unlinked markers. While a higher number of partitions is desirable to better interpolate the mean conditional effect curve, fully optimizing the partitioning cut-offs rapidly becomes impractical due to the combinatorial search space and limited training sample size. To explore this issue, we benchmarked fully optimized three-way partitioning, of which two cut-off values were optimized in the training cohort, against seven-way partitioning, of which cut-off points were evenly placed without further optimization. In our simulation, the seven-way partitioning performed equivalently as or significantly more accurately than the three-way partitioning (Fig S1, $P < 10^{-4}$ for the causal SNP fractions simulated at 1% and 3%).

Note that the widely-used *p*-value thresholding approach is a special case of non-parametric shrinkage with the number of partition *K* being 2. When the effect sizes are measured relative to the standardized genotypes, thresholding on $|\hat{\beta}_j|$ is equivalent to thresholding on *p*-values of association. This is because $|\hat{\beta}_j|$ has the following monotonic relationship with *p*-values:

$$|\hat{\beta}_j| = \frac{1}{N}\Phi^{-1}\left(1 - \frac{1}{2}p_j\right)$$

where Φ is the standard normal cumulative distribution function and p_j is the *p*-value of association at SNP j. As expected, thresholding was always significantly inferior to finer partitioning (Fig S1, $P < 0.01$).

For this analysis, we simulated a spike-and-slab genetic architecture under a liability threshold model. The total number of markers was 100,000, and 0.3, 1, or 3% of all markers were simulated to be causal SNPs with effect sizes randomly sampled from a normal distribution. SNPs were independent with each other without LD in this simulation. We assumed the heritability of 0.5 and disease prevalence of 5% and over-sampled cases relative to

controls to generate cohorts with case:control ratio of 1:1. For each simulated causal fraction, we generated 500 sets of three cohorts: discovery, training and validation with sample sizes of 20,000, 4,000 and 4,000, respectively. Discovery cohort data were accessed only as summary statistics. Prediction models were optimized in training cohort and evaluated in validation cohort. For the *p*-value thresholding, the optimal *p*-value cut-off for polygenic risk score was scanned from 10^{-7} to 1 at increments of 0.25 on the $\log_{10}$ scale. The cut-off yielding the highest Area Under the Curve (AUC) in the training cohort was selected for the final model tested in the validation cohort. For seven-way partitioning, we used fixed partition boundaries on the interval of $|\hat{\beta}|$ corresponding to every order of magnitudes in *p*-values of association. Specifically, the intervalof $|\hat{\beta}|$ was split at six *p*-values of association: $5x10^{-7}$, $5x10^{-6}$, …, and $5x10^{-2}$. For three-way partitioning, all 15 combinations of the above six *p*-value cut-off values were examined to find the most optimal partition. In all three partitioning schemes, the shrinkage weight for the partition of smallest $|\hat{\beta}|$ was set to 0.

References

[1] Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015 Oct 01;97(4):576-92. PubMed PMID: 26430803. Pubmed Central PMCID: 4596916.