Non-parametric polygenic risk prediction using partitioned GWAS summary statistics

Sung Chun[1,2,3,4,¶], Maxim Imakaev[1,2,3,4,¶], Daniel Hui[1,3,5], Nikolaos A. Patsopoulos[1,3,5], Benjamin M. Neale[3,6,7], Sekar Kathiresan[3,7,8], Nathan O. Stitziel[9,10,11,*], Shamil R. Sunyaev[1,2,3,4,*]

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

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

[3] Broad Institute of Harvard and MIT, Cambridge, Massachusetts, 02142, USA

[4] Altius Institute for Biomedical Sciences, Seattle, Washington, 98121, USA

[5] Systems Biology and Computer Science Program, Ann Romney Center for Neurological Diseases, Department of Neurology, Brigham & Women's Hospital, Boston, 02115 MA, USA

[6] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA

[7] Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA

[8] Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA

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

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

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

¶ These authors contributed equally to this work.

* Corresponding authors

nstitziel@wustl.edu (NOS)

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

Abstract

In complex trait genetics, the ability to predict phenotype from genotype is the ultimate measure of our understanding of genetic architecture underlying the heritability of a trait. A complete understanding of the genetic basis of a trait should allow for predictive methods with accuracies approaching the trait's heritability. The highly polygenic nature of quantitative traits and most common phenotypes has motivated the development of statistical strategies focused on combining myriad individually non-significant genetic effects. Now that predictive accuracies are improving, there is a growing interest in practical utility of such methods for predicting risk of common diseases responsive to early therapeutic intervention. However, existing methods require individual level genotypes or depend on accurately specifying the genetic architecture underlying each disease to be predicted. Here, we propose a polygenic risk prediction method that does not require explicitly modeling any underlying genetic architecture. We start with a set of summary statistics in the form of SNP effect sizes from a large GWAS cohort. We then remove the correlation structure across summary statistics arising due to linkage disequilibrium and apply a piecewise linear interpolation on conditional mean effects. In both simulated and real datasets, this new non-parametric shrinkage (NPS) method can reliably correct for linkage disequilibrium in summary statistics of 5 million dense genome-wide markers and consistently improves prediction accuracy. We show that NPS significantly improves the identification of groups at high risk for Breast Cancer, Type 2 Diabetes, Inflammatory Bowel Disease and Coronary Heart Disease, all of which have available early intervention or prevention treatments. The NPS software is available at http://github.com/sgchun/nps/.

Introduction

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 same logic applies to many other disease areas with available prophylactic interventions including cancers and diabetes.

The field of phenotypic prediction was conceived in plant and animal genetics (reviewed in refs. [2,3]). The first approaches relied on "major genes" – allelic variants of large effect sizes readily detectable by genetic linkage or association. These efforts were quickly followed by strategies adopting polygenic models, most notably the genomic version of the Best Linear Unbiased Predictor (BLUP) [4].

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 [5–7]. Almost immediately, after realization that a multitude of small effect alleles play an important role in complex trait genetics [2,3,8], these methods were extended to accommodate very large (or even all) genetic markers [9–14]. These methods include extensions of BLUP [9,10], or Bayesian approaches that extend both shrinkage techniques and random effect models [11]. 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" [14–18] represent an alternative approach based on summary statistics. The originally proposed version is additive over genotypes weighted by apparent effect sizes exceeding a given *p*-value threshold. In theory, the risk predictor based on expected true genetic effects given the genetic effects observed in GWAS (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 [19]. 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 simplified model of genetic architecture. This methodology has been successfully used to analyze the UK Biobank, the largest epidemiological cohort that includes genetic data [20]. Individuals with extreme values of polygenic score were shown to have a substantially elevated risk for corresponding diseases, generating enthusiasm for clinical applications of the method.

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. We evaluate the performance of this new approach under a simulated genetic architecture of 5 million dense SNPs across the genome. We also test the method using real data in four disease areas: breast cancer, type 2 diabetes, inflammatory bowel disease and coronary heart disease.

Results

Method Overview

If true genetic effects of all variants on the trait were known, adding these effects for all alleles in an individual would provide the ideal linear predictor of the phenotype. The accuracy of such a predictor would equal narrow sense heritability. However, true genetic effects are unknown and their statistical estimates deviate from the true values even in expectation. Estimates of genetic effects in GWAS are strongly affected by sampling noise, and the variants with smallest effect sizes are difficult to distinguish from the background noise of non-causal SNPs. Another complication arises from extensive linkage disequilibrium (LD). Estimated genetic effects are strongly influenced by effects of neighboring variants. Since true genetic effects are unknown, they have to be approximated based on available data. Formally, the best possible linear predictor would rely on expected genetic effects conditional on summary statistics[19]. Sampling noise increases absolute values of estimated genetic effects compared to the true effects. The expected true effects can be expressed as “shrinking” the estimated effects towards zero via differential weighting of the estimated effects.

Our approach outlined in Figure 1 is to partition SNPs into groups of similar observed effect sizes in GWAS data ($\hat{\beta}$) and determine the relative weight based on predictive value of each partition estimated in the training data. Intuitively, a partition dominated by non-causal variants will have low power to distinguish cases from controls whereas the partition enriched with strong signals will be better able to predict a phenotype. In the absence of LD, this is equivalent to approximating the conditional mean effect curve by piecewise linear interpolation (Methods). Note that estimating the per-partition weights is a far easier problem than estimating per-SNP effects. The training sample size is small but still larger than the number of partitions, whereas for per-SNP effects, the GWAS sample size is considerably smaller than the number of markers in the genome. This procedure "shrinks" the estimated effect sizes not relying on any specific assumption about the distribution of true effect sizes. Thus, we call it "Non-Parametric Shrinkage" (NPS).

In the presence of LD, we cannot apply the partitioning method directly to GWAS effect sizes since true genetic effects as well as sampling noise are correlated between adjacent SNPs. To prevent estimated genetic signals smearing across partitions, we transform GWAS data into an orthogonal domain, which we call "eigenlocus" (Fig. 1b and Methods). Specifically, we use a decorrelating linear transformation obtained by eigenvalue decomposition of local LD matrix. Both genotypes and sampling errors are uncorrelated in the eigenlocus representation. We apply our partitioning-based non-parametric shrinkage to the estimated effect sizes in the eigenlocus, and then restore them back to the original per-SNP effects.

In general, NPS requires double partitioning on both eigenvalues of the decorrelating projection and GWAS effect sizes in the eigenlocus space (Methods and Supplementary Note). Since the full combinatorial optimization of partitioning cut-offs is neither necessary nor practical, we place the cut-offs for 10 by 10 double-partitioning based on heuristics without optimizing them on individual datasets. In simulations, we can show that shrinkage weights estimated by the NPS approach closely track the conditional mean effects in the eigenlocus space (Supplementary Figs. 1-4).

Simulated benchmark

To benchmark the accuracy of NPS, we simulated the genetic architecture using the real LD structure of 5 million dense common SNPs from the 1000 Genomes Project (Methods). We considered the causal fraction of SNPs from 1% to 0.01%, dependency of heritability on minor allele frequency (MAF) and enrichment of heritability in DNase I hypersensitive sites (DHS) based on the previous literature[21–23]. The prediction accuracy of NPS remained robust across the simulated genetic architectures (Table 1 and Supplementary Tables 1). We evaluated the performance of NPS vis-a-vis a comparable successful parametric technique (Supplementary Tables 1-4). 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. 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 some of the simulated allelic architectures are the spike-and-slab allelic architecture for which LDPred is expected to be optimal as a Bayesian method. However, we found that in most simulations, LDPred adopted the infinitesimal or extremely polygenic model irrespective of the true simulated regime, pointing to the challenge of computational optimization in the parametric case (Supplementary Table 3). The simulations suggest that the well-optimized parametric model is capable of generating good predictions, but NPS is much more robust and does not suffer from optimization issues. Overall, our method significantly outperformed LDPred as well as the commonly-used Pruning and Thresholding (P+T) approach (Table 1).

Application to real data

We benchmarked the accuracy of NPS and other methods using publicly available GWAS summary statistics and training and validation cohorts assembled with UK Biobank samples (Methods) [24–29]. For all three phenotypes we examined, NPS showed significantly higher accuracy than LDPred or P+T. (Table 2, Supplementary Tables 5-7 and Supplementary Figs. 5-9). In particular, our method outperformed the other methods by greater magnitudes with more recent GWAS summary statistics with finer resolution. For example, the latest breast cancer

GWAS study has twice as large sample size as the previous study and used a custom genotyping array to densely genotype known cancer susceptibility loci. The R^2 of our method increased by 1.42-fold with the latest breast cancer data whereas the accuracy of P+T and LDPred improved only 1.14 and 1.12-fold, respectively.

Since our method estimates a large number of parameters from the training data, it might be particularly vulnerable to overfitting cryptic genetic features common to both training and testing data which may result in inflated prediction accuracy. To eliminate this possibility, we benchmarked the prediction models in Partners Biobank, as an independent validation cohort (Methods) [30]. For all phenotypes except early-onset coronary artery disease (for which there were very few cases in the validation cohort), NPS outperformed both P+T and LDPred in terms of the prediction R^2 (Table 3 and Supplementary Tables 8-10).

A recent study reported that the extreme tails of the polygenic score distribution are associated with risk that is similar to monogenic mutations [20]. At the highest 5% tail in polygenic risk score distribution, NPS yielded odds ratios that were higher than the other methods across all phenotypes (Tables 2 and 3). Overall, the odds ratios of disease for the upper 5% tail (compared to the remainder of the distribution) produced by NPS were significantly higher than those of LDPred and P+T (Fisher's method, P=0.002 and 0.0002, respectively), indicating an ability to identify an even higher risk subset of the population than previously appreciated.

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 [20,31]. 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 [18,32]. Also, our method assumes – similar to other methods – that all datasets come from a homogeneous population. It has been shown that polygenic risk models are not transferrable between populations due to differences in allele frequencies and patterns of linkage disequilibrium [33], which is a problem that should be addressed by future work in this field.

Human phenotypes vary in the degree of polygenicity [34], in the fraction of heritability attributable to low-frequency variants [21] and in other aspects of allelic architecture [22,35]. The optimality of a Bayesian 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 [21–23,34,35]. Several studies have proposed to extend polygenic scores by incorporating additional complexity into the parametric Bayesian models, yet these methods were not applied to genome-wide sets of markers due to computational challenges [36,37]. Recently, there has been a growing interest in non-parametric or semi-parametric approaches, such as those based on modeling of latent variables or kernel-based estimation of prior or marginal distributions, however, thus far they cannot leverage summary statistics or directly account for the linkage disequilibrium (LD) structure in the data [38–41]. To address these issues, we developed NPS, a non-parametric method which is agnostic to allelic architecture. In simulations, we show that this approach should be advantageous across a wide range of phenotypes and traits with differing underlying architectures, and find that it outperforms existing prediction methods in UK Biobank for four different traits of medical interest. Finally, as demonstrated in the prediction accuracy using two different breast cancer GWAS summary statistics, with increasing size and marker density in case-control association studies across a range of diseases, our NPS method should continue to outperform traditional parametric approaches for identifying individuals at increased risk.

Methods

Overview of Non-Parametric Shrinkage (NPS). In the absence of LD, conditional mean effects, namely, the expected true genetic effects given observed GWAS data, can be approximated by piecewise linear interpolation. We partition SNPs into *K* disjoint intervals based on observed GWAS effect sizes ($\hat{\beta}_j$) and fit a linear function $f(\hat{\beta}_j) = \omega_k \hat{\beta}_j$ on each interval of $k = 1, \dots, K$. Specifically, when x_{ij} is the genotype of individual i at SNP $j = 1, \dots, M$ 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 follows:

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

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:

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

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 (1), we can use 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.

In the presence of LD, we transform genotypes and GWAS effect sizes into the eigenlocus representation defined by a decorrelating linear projection $\mathcal{P}$. Specifically, the decorrelating projection $\mathcal{P}$ is defined as follows:

$$\mathcal{P} = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{Q}^{\boldsymbol{T}} \quad (2)$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{Q}$ are matrices of eigenvalues and eigenvectors, respectively, of a local reference LD matrix $\boldsymbol{D}$. For the axis of projection j defined by eigenvalue λ_j and eigenvector q_j, $\mathcal{P}$ yields the following projected genotype x_{ij}^P and estimated effect $\hat{\eta}_j$:

$$x_{ij}^P = \frac{1}{\sqrt{\lambda_j}} q_j^T x_i \qquad \text{and} \qquad \hat{\eta}_j = \frac{1}{\sqrt{\lambda_j}} q_j^T \hat{\beta} \tag{3}$$

where x_i and $\hat{\beta}$ are genotypes and estimated GWAS effects, respectively, in the original SNP space. In the eigenlocus representation, both x_{ij}^P and sampling errors of $\hat{\eta}_j$ are uncorrelated across axes of projection j, therefore we can apply the partitioning-based non-parametric shrinkage on x_{ij}^P and $\hat{\eta}_j$ similarly for the case without LD (Supplementary Note).

Application of NPS to genome-wide datasets. The estimated effect sizes $\hat{\beta}_j$ at SNPs j = 1, …, M are available as summary statistics from a large discovery GWAS study. When $\hat{\beta}_j$ was a per-allele effect, we converted them relative to standardized genotypes by multiplying by $\sqrt{2f_j(1-f_j)}$, where f_j is the allele frequency of SNP j in the discovery GWAS cohort. For case/control GWAS, logistic log odds ratios were used for $\hat{\beta}_j$.

Because of the difficulty to finely partition the largest-effect tail, we handled the genome-wide significant SNPs separately from the rest of SNPs. The genome-wide significant SNPs were set aside to a special partition $\mathcal{S}_0$, for which the decorrelating projection was set to the identity matrix I. To avoid LD between SNPs in $\mathcal{S}_0$, genome-wide significant SNPs were selected into $\mathcal{S}_0$ keeping the minimum distance of 500 kb from each other. Secondary GWAS peaks within 500 kb from a SNP included in $\mathcal{S}_0$ were handled together with the rest of polygenic signals regardless of their conditional significance. In order to avoid double-counting the effects of SNPs set aside to $\mathcal{S}_0$, GWAS effect sizes were residualized on the estimated effect of each SNP in $\mathcal{S}_0$ up to 500 kb in both directions.

Then, we processed the residualized effect size estimates $\hat{\beta}_j$ and genotypes of individuals in a 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 eigenvalue 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 necessarily non-negative semi-definite, $\boldsymbol{\Lambda}$ can include negative eigenvalues. Negative eigenvalues were truncated along with those that are positive but smaller than 0.5 since they were dominated by noise. Applying the eigenlocus projection $\mathcal{P}$ (equation 2), 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 3), 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 cross-window correlations. To eliminate such correlations, we applied LD pruning in the eigenlocus space between adjacent windows. Specifically, we calculated Pearson correlations $\rho_{jj'}$ between decorrelated genotypes x_{ij}^P and $x_{ij'}^P$, where j and j' are the indices of projection, belonging to neighboring windows. For the pairs with $|\rho_{jj'}| > 0.3$, we kept the one yielding a larger absolute effect size 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 $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ from the lowest to the highest. Each partition of eigenvalues $\mathcal{S}_k$ was sub-partitioned on intervals of $|\hat{\eta}_j|$, equally distributing $\sum_j \hat{\eta}_j^2$ across partitions, and split to partitions $\mathcal{S}_{k,1}, \ldots, \mathcal{S}_{k,10}$. 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 total, we used 101 partitions including the partition of genome-wide significant SNPs $\mathcal{S}_0$.

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

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

where $k = 0$ or $k \in \{1 \dots 10\} \times \{1 \dots 10\}$. Given the phenotype y_i of each individual in the training cohort, we estimated per-partition shrinkage weights ω_k by linear discriminant analysis (LDA) using the equation (1). Each $\hat{\eta}_j$ was reweighted by the shrinkage weight of corresponding partition to obtain conditional mean decorrelated effects as follows:

$$E[\eta_j \mid \hat{\eta}_j] = \omega_k \hat{\eta}_j \quad \text{for } j \in \mathcal{S}_k$$

Then, we back-transformed the effect sizes from the eigenlocus representation to the original per-SNP effect space in each window (Supplementary Note).

Because the accuracy of eigenlocus 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 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 of genetic architecture with dense genome-wide markers. For simulated benchmarks, we generated genetic architecture with 5 million dense 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 (n=404) to populate LD structures in simulated genetic data. Due to the limited sample size of the LD panel, we regularized the LD matrix by applying Schur product with a tapered banding matrix so that the LD smoothly tapered off to 0 starting from 150 kb up to 300 kb [42].

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. Since we assume that LD travels no farther than 300 kb, as long as we simulate genotypes in blocks of length greater than 300 kb, 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 1,250 SNPs (average 750 kb in length) were sampled directly out of multivariate normal distribution $N(\mu = 0, \Sigma = \boldsymbol{D}_{(1)})$. From the next block, we sampled the genotypes of 1,250 SNPs 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:

$$\begin{pmatrix} \boldsymbol{D}_l & \boldsymbol{D}_{l,l+1} \\ \boldsymbol{D}_{l+1,l} & \boldsymbol{D}_{l+1} \end{pmatrix}$$

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

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

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 Hardy-Weinberg equilibrium. n is the number of simulated samples, and f_j is the allele frequency of SNP j. This MVN-based simulator can efficiently generate a very large cohort with realistic LD structure across the genome and guarantees to produce homogenous population without stratification.

We simulated three different sets of genetic architecture: point-normal mixture, MAF dependency and DNase I hypersensitive sites (DHS). The point-normal mixture 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 being 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 allowed the scale of causal effect sizes to vary across SNPs in proportion to $\left(f_j(1-f_j)\right)^{\alpha}$ with $\alpha = -0.25$ [21] 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 point-normal 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 [23]. 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 with Plink by testing for the association with the total liability instead of case/control status; this is computationally easier than to generate a large case/control GWAS cohort directly, and the estimated effect sizes are equivalent. 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 [43]. 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.

GWAS summary statistics. GWAS summary statistics are publicly available for phenotypes of breast cancer [24,25], inflammatory bowel disease (IBD) [26], type 2 diabetes (T2D) [27] and coronary artery disease (CAD) [29]. These GWAS summary statistics were based only on Caucasian samples with an exception of CAD, for which 13% of discovery cohort comprised of non-European ancestry.

UK Biobank. UK Biobank samples were used for training and validation purposes. Case and control samples were defined as follows. Breast cancer cases were identified by ICD10 codes of diagnosis. Controls were selected from females who were not diagnosed with or did not self-report history of breast cancer. We excluded individuals with history of any other cancers, *in situ* neoplasm or neoplasm of unknown nature or behavior from both cases and controls. For IBD, we identified case individuals by ICD10 or self-reported disease codes of Crohn's disease, ulcerative colitis or IBD. Controls were randomly selected excluding participants with history of any auto-immune disorders. For T2D, cases were identified by ICD10 diagnosis codes or by questionnaire on history of diabetes combined with the age of diagnosis over 30. For early-onset CAD, case individuals were identified by ICD10 codes of diagnosis or cause of death. The early-onset was determined by the age of heart attack on the questionnaire ($\leq$ 55 for men and $\leq$ 65 for women). Individuals with history of CAD were excluded from controls regardless of the age of onset. The latest CAD summary statistics include UK Biobank samples in the interim release; thus, to avoid sample overlap, we used only post-interim samples, which were identified by genotyping batch IDs.

For genotype QC, we filtered out SNPs with MAF below 5% or INFO score less than 0.4. We also excluded tri-allelic SNPs and InDels. We discarded SNPs if MAFs deviate by more than 0.1 between UK Biobank and GWAS discovery cohorts.

For all phenotypes, we filtered out participants who were retracted, not from white British ancestry, or had indication of any QC issue in UK Biobank. We included only samples which were genotyped with Axiom array. Related samples were excluded to avoid potential confounding. Controls were down-sampled to meet the case to control ratio of 1:1. The selected samples were randomly split to training and validation cohorts. Because of case/control ascertainment, we determined the 5% cut-off in polygenic risk score distribution indirectly by over-sampling control samples while accounting for the known prevalence of disease in UK Biobank (1,000 iterations).

Partners Biobank. We used Partners Biobank [30] to evaluate the accuracy of prediction models in an independent validation cohort. These genotyping data were previously generated using the

MEGA-Ex array. Markers with monomorphic allele frequency, complementary alleles, less than 99.5% genotyping rate, or deviation from Hardy-Weinberg equilibrium ($P < 0.05$) were removed. Then, statistical imputation was conducted to infer genotypes at missing markers using Eagle v2.4 and IMPUTE v4 on the reference panel (1000 Genomes Phase 3). Excluding samples of non-European ancestry, a total of 16,839 samples from US white population were available for use. Participants with breast cancer, IBD, T2D and CAD were identified using a phenotype query algorithm with the PPV parameter of 0.90 [44]. To obtain early-onset CAD, both cases and controls were restricted to men with age ≤ 55 and women with age ≤ 65. Since the definition of early-onset CAD is sex-dependent, we included the sex covariate in the genetic risk model for CAD. The coefficient of sex covariate was estimated in the training cohort.

LDPred. The accuracy of LDPred was evaluated in simulated and real datasets using the default parameter setting. The underlying causal fraction parameter was optimized using the training cohort, which is available as individual-level genotype data. Specifically, the causal SNP fractions of 1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 0.0003 and 0.0001 were tested in the training data, and the prediction model yielding the highest prediction R^2 was selected for validation. The training genotypes were also used as a reference LD panel.

LDPred accepts only hard genotype calls as inputs. Thus, for real data we converted imputed allelic dosages to most likely genotypes after filtering out SNPs with genotype probability < 0.9. SNPs with the missing rate > 1% or deviation from Hardy-Weinberg equilibrium ($P < 10^{-5}$) were also excluded. Prediction models were trained using only SNPs which passed all QC filters in both training and validation datasets, as recommended by the authors. SNPs with complementary alleles were excluded automatically by LDPred. In simulations, all genotypes were generated as hard calls, and complementary alleles were avoided; thus, the exactly same set of SNPs were used for both LDPred and NPS. In a subset of datasets, we further examined the accuracy of LDPred when it was run only with directly genotyped SNPs. In simulated datasets, we assumed that both training and validation cohorts were genotyped with Illumina HumanHap550v3 array, restricting the genotype data to 490,504 common SNPs. For UK Biobank

datasets, prediction models were constrained to up to 354,110 common SNPs in UK Biobank Axiom array. In the case of validation in Partners Biobank, we did not consider running LDPred only with genotyped SNPs since too few SNPs were directly genotyped in both UK Biobank and Partners Biobank.

LD Pruning and Thresholding. LD Pruning and Thresholding (P+T) algorithm was evaluated using PRSice software in the default setting [45]. In real data, imputed allelic dosages were converted to hard-called genotypes similarly as for LDPred. A training cohort was used as a reference LD panel and to optimize pruning and thresholding parameters. The best prediction model suggested by PRSice was evaluated in validation cohorts.

References

1. Grundy, S. M. *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* (2018). doi:10.1016/j.jacc.2018.11.003
2. Goddard, M. E. & Hayes, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* **10,** 381–391 (2009).
3. Falke, K. C. *et al.* The spectrum of mutations controlling complex traits and the genetics of fitness in plants. *Curr Opin Genet Dev* **23,** 665–671 (2013).
4. Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. *Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps*. (2001).
5. Ripatti, S. *et al.* A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* **376,** 1393–1400 (2010).
6. Wacholder, S. *et al.* Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* (2010). doi:10.1056/NEJMoa0907727
7. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17,** 1520–8 (2007).
8. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–9 (2010).
9. Golan, D. & Rosset, S. Effective Genetic-Risk Prediction Using Mixed Models. *Am. J. Hum. Genet.* **95,** 383–393 (2014).
10. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24,** 1550–1557 (2014).
11. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9,** e1003264 (2013).
12. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45,** 400–5, 405e1–3 (2013).
13. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
14. Stahl, E. a *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44,** 483–9 (2012).
15. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet* **45,** 400–5, 405e1–3 (2013).

16. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–52 (2009).
17. Shi, J. *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* **12,** e1006493 (2016).
18. Vilhjalmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97,** 576–592 (2015).
19. Goddard, M. E., Wray, N. R., Verbyla, K. & Visscher, P. M. Estimating Effects and Making Predictions from Genome-Wide Marker Data. *Stat. Sci.* **24,** 517–529 (2009).
20. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50,** 1219–1224 (2018).
21. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49,** 986–992 (2017).
22. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50,** 746–753 (2018).
23. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–52 (2014).
24. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47,** 373–380 (2015).
25. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551,** 92–94 (2017).
26. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47,** 979–986 (2015).
27. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66,** 2888–2902 (2017).
28. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018).
29. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49,** 1385–1391 (2017).
30. Karlson, E. *et al.* Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. *J. Pers. Med.* **6,** 2 (2016).
31. Riglin, L. *et al.* Schizophrenia risk alleles and neurodevelopmental outcomes in childhood: a population-based cohort study. *The Lancet Psychiatry* **4,** 57–62 (2017).
32. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14,** 507–15 (2013).
33. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100,** 635–649 (2017).
34. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
35. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49,** 1421–1427 (2017).
36. Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.* **13,** 1–16 (2017).
37. Hu, Y. *et al.* Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genet.* **13,** 1–22 (2017).
38. Zeng, P. & Zhou, X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat. Commun.* **8,** 1–11 (2017).
39. Efron, B. Empirical bayes estimates for large-scale prediction problems. *J. Am. Stat. Assoc.* **104,** 1015–1028 (2009).
40. So, H. C. & Sham, P. C. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. *Sci. Rep.* **7,** 1–11 (2017).
41. Gianola, D., Fernando, R. L. & Stella, A. Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. *Genetics* **173,** 1761–1776 (2006).

42. Cai, T. T., Zhang, C. H. & Zhou, H. H. Optimal rates of convergence for covariance matrix estimation. *Ann. Stat.* **38,** 2118–2144 (2010).
43. Yang, J., Wray, N. R. & Visscher, P. M. Comparing apples and oranges: Equating the power of case-control and quantitative trait association studies. *Genet. Epidemiol.* **34,** 254–257 (2010).
44. Gainer, V. S. *et al.* The Biobank Portal for Partners Personalized Medicine: A Query Tool for Working with Consented Biobank Samples, Genotypes, and Phenotypes Using i2b2. *J. Pers. Med.* **6,** (2016).
45. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31,** 1466–1468 (2015).

Table 1. Comparison of prediction accuracy in simulated genetic architecture.

% causal SNPs	Method	$R^2_{Nagelkerke}$	h^2 explained (%)	R^2 gain over P+T	R^2 gain over LDPred
1%	P+T	0.050	14.8		
	LDPred	0.068	20.6		
	NPS	0.078	23.4	**1.55 ***	**1.15 ***
0.1%	P+T	0.136	40.8		
	LDPred	0.080	23.0		
	NPS	0.167	47.8	**1.23 ***	**1.42 ***
0.01%	P+T	0.213	61.4		
	LDPred	0.153 (0.268)[1]	43.8 (74.6)[1]		
	NPS	0.315	88.8	**1.48 ***	**1.94 ***

Non-parametric shrinkage (NPS) is more robust and accurate than Pruning and Thresholding (P+T) and Bayesian parametric method (LDPred). The simulations incorporate the dependency of heritability on minor allele frequency and clumping of causal SNPs in known DHS elements. The heritability was 0.5, and the prevalence was 5%. The number of markers was 5,012,500. 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 measured in the validation cohort of 50,000 samples. The h^2 explained stands for the proportion of heritability on the liability scale explained by polygenic scores. The star (*) indicates a significant improvement in Nagelkerke's R^2 (paired t-test). [1]The accuracy of LDPred varies widely depending on the convergence of prediction model; thus, we report the maximum R^2 in parenthesis as well as the average performance.

Table 2. Accuracy of polygenic prediction in real data.

Discovery GWAS	Training (UK Biobank)	Validation (UK Biobank)	Method	R^2_{Nag}	Tail OR	R^2 gain over P+T	R^2 gain over LDPred
Breast Cancer 2015 (N=~120,000)	N=3,956/3,956	N=3,957/3,957	P+T	0.051	2.28		
			LDPred	0.061	2.33		
			NPS	0.060	2.50	**1.19 ***	0.98
Breast Cancer 2017 (N=~230,000)			P+T	0.064	2.25		
			LDPred	0.059	2.54		
			NPS	0.085	2.86	**1.33 ***	**1.44 ***
Inflammatory Bowel Disease (N=~35,000)	N=2,483/2,483	N=2,482/2,482	P+T	0.085	2.85		
			LDPred	0.076	2.71		
			NPS	0.094	3.19	1.11	**1.24 ***
Type 2 Diabetes (N=~160,000)	N=7,298/7,298	N=7,298/7,298	P+T	0.057	2.29		
			LDPred	0.081	2.78		
			NPS	0.094	2.93	**1.65 ***	**1.16 ***

Non-Parametric Shrinkage (NPS) outperforms both Pruning and Thresholding (P+T) and LDPred in real data. Both training and validation cohorts were sampled from UK Biobank. The tail Odds Ratio (OR) stands for the odds ratios of cases over controls at the 5% tail in polygenic score distribution compared to the rest. The star (*) indicates a significant improvement in Nagelkerke's R^2 (R^2_{Nag}) by bootstrapping.

Table 3. Accuracy of polygenic prediction in independent validation cohorts.

Discovery GWAS	Training (UK Biobank)	Validation (Partners)	Method	R^2_{Nag}	Tail OR	R^2 gain over P+T	R^2 gain over LDPred
Breast Cancer 2017 (N=~230,000)	N=3,956/3,956	N=754/8,324	P+T	0.016	1.56		
			LDPred	0.015	1.78		
			NPS	0.024	2.08	**1.52 ***	**1.64 ***
Inflammatory Bowel Disease (N=~35,000)	N=2,483/2,483	N=839/16,000	P+T	0.050	3.57		
			LDPred	0.038	3.07		
			NPS	0.057	3.81	1.15	**1.53 ***
Type 2 Diabetes (N=~160,000)	N=7,298/7,298	N=2,026/14,813	P+T	0.016	1.78		
			LDPred	0.024	1.81		
			NPS	0.029	2.04	**1.84 ***	**1.20 ***
Coronary Artery Disease (N=~330,000)	N=2,773/2,773	N=268/7,107	P+T	0.020	3.05		
			LDPred	0.016	2.22		
			NPS	0.019	3.27	0.94	1.19

Non-Parametric Shrinkage (NPS) outperforms both Pruning and Thresholding (P+T) and LDPred in completely independent validation cohorts from US white population (Partners Biobank). The same cohorts from UK Biobank was used for training prediction models (Table 2). The tail Odds Ratios (OR) stand for the odds ratios of cases over controls at the 5% tail in polygenic score distribution compared to the rest. The star (*) indicates a significant improvement in Nagelkerke's R^2 by bootstrapping.

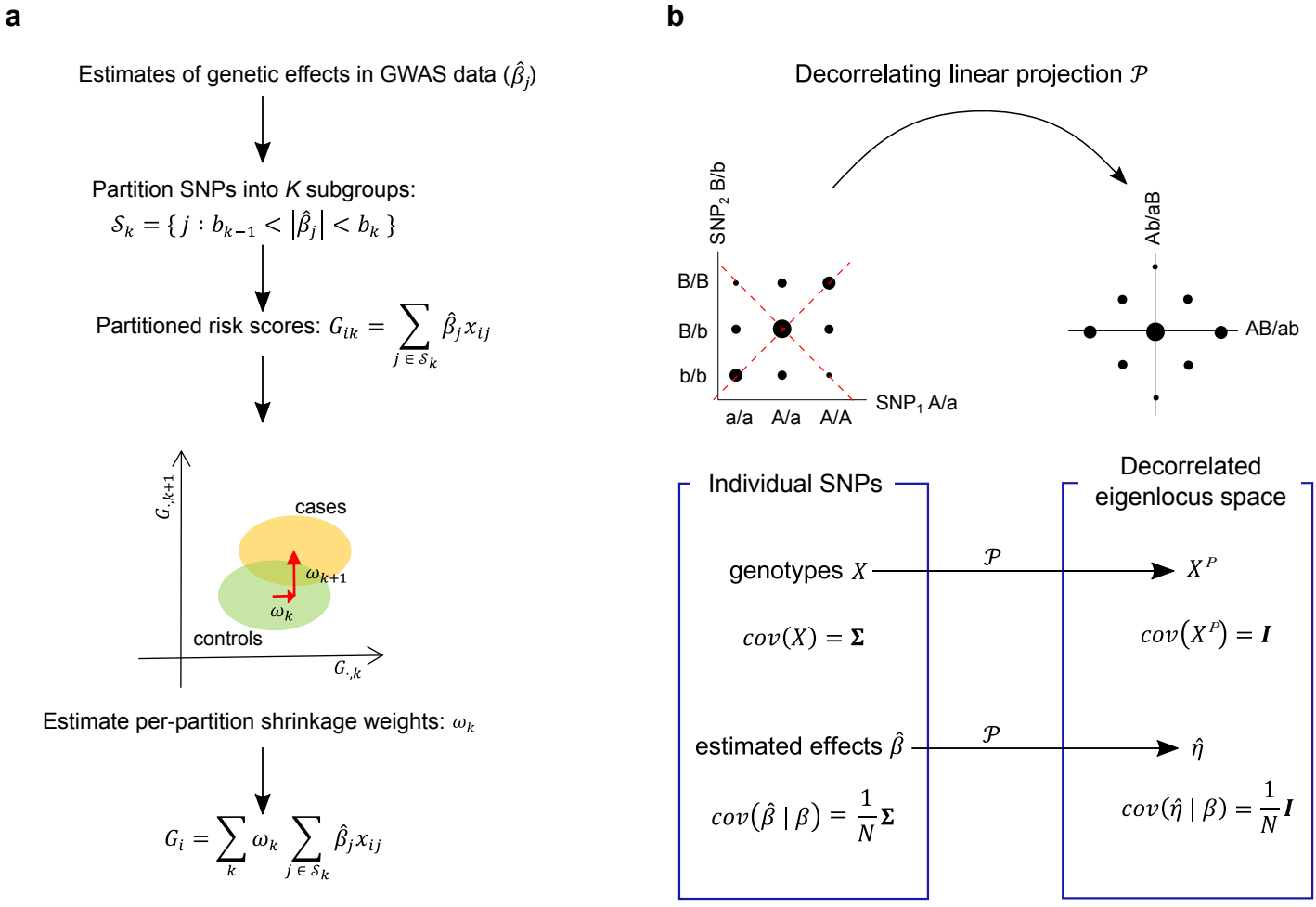


Figure 1. Overview of Non-Parametric Shrinkage (NPS).

(a) For unlinked markers, NPS partitions SNPs into *K* subgroups splitting the GWAS effect sizes ($\hat{\beta}_j$) 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 Mb in length, and then NPS is applied to the data. The size of black dots indicates genotype frequencies in population. Before projection, genotypes between SNP 1 and 2 are correlated due to LD ($\boldsymbol{\Sigma}$), and thus sampling errors of estimated effects ($\hat{\beta}_j \mid \beta_j$) are also correlated between adjacent SNPs. The projection $\mathcal{P}$ neutralizes both correlation structures. The axes of projection are marked by red dashed lines. β_j denotes the true genetic effect at SNP j. N is the sample size of GWAS cohort.

Supplementary Information for:

Non-parametric polygenic risk prediction using partitioned GWAS summary statistics

Sung Chun[1,2,3,4,¶], Maxim Imakaev[1,2,3,4,¶], Daniel Hui[1,3,5], Nikolaos A. Patsopoulos[1,3,5], Benjamin M. Neale[3,6,7], Sekar Kathiresan[3,7,8], Nathan O. Stitziel[9,10,11,*], Shamil R. Sunyaev[1,2,3,4,*]

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

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

[3] Broad Institute of Harvard and MIT, Cambridge, Massachusetts, 02142, USA

[4] Altius Institute for Biomedical Sciences, Seattle, Washington, 98121, USA

[5] Systems Biology and Computer Science Program, Ann Romney Center for Neurological Diseases, Department of Neurology, Brigham & Women's Hospital, Boston, 02115 MA, USA

[6] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA

[7] Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA

[8] Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA

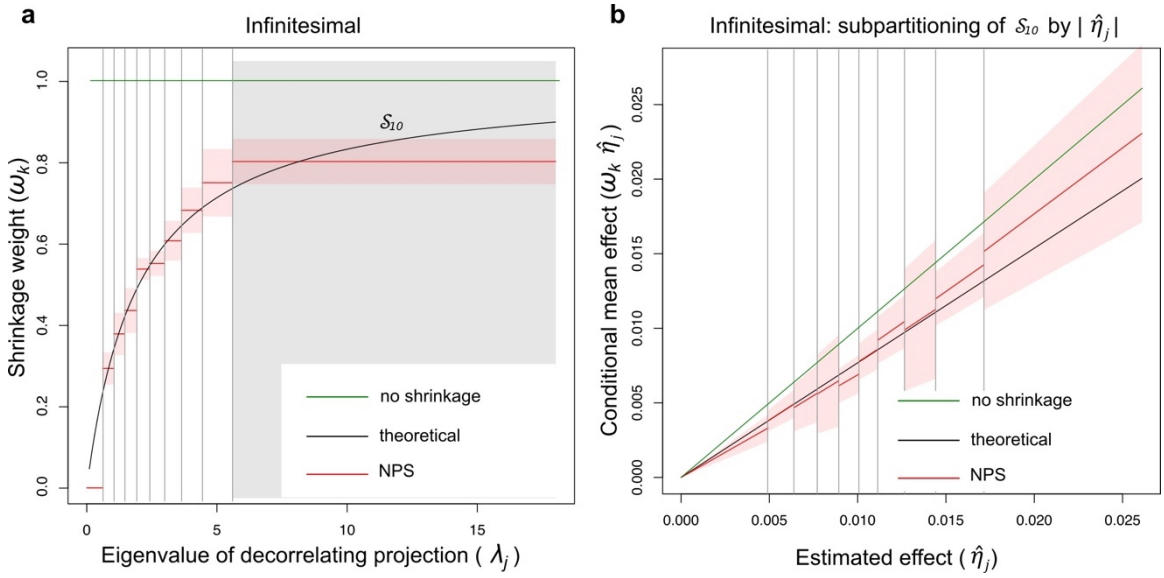
[9] Cardiovascular Division, Department of Medicine, Washington University School of Medicine, Saint Louis, Missouri, 63110, USA

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

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

[¶] These authors contributed equally to this work.

[*] Correspondence to – nstitziel@wustl.edu (NOS) and ssunyaev@rics.bwh.harvard.edu (SRS)



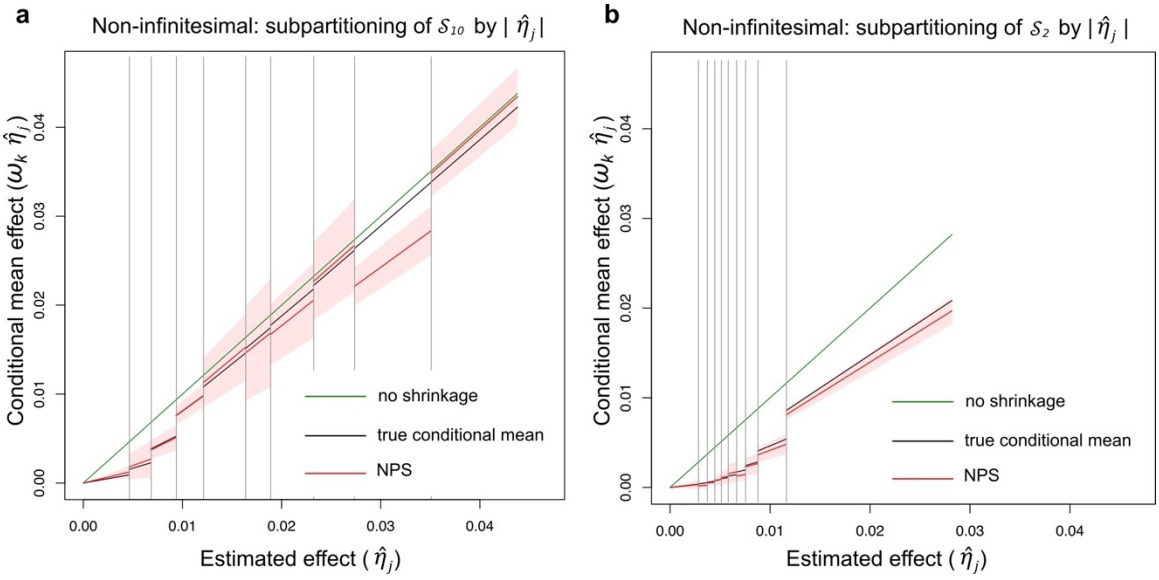
Supplementary Figure 1. Non-parametric shrinkage (NPS) approximates the conditional mean effects: infinitesimal genetic architecture. For the infinitesimal architecture, the analytic solution for conditional mean effect is known and can be reformulated as follows (See ref. [1] and Supplementary Note):

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

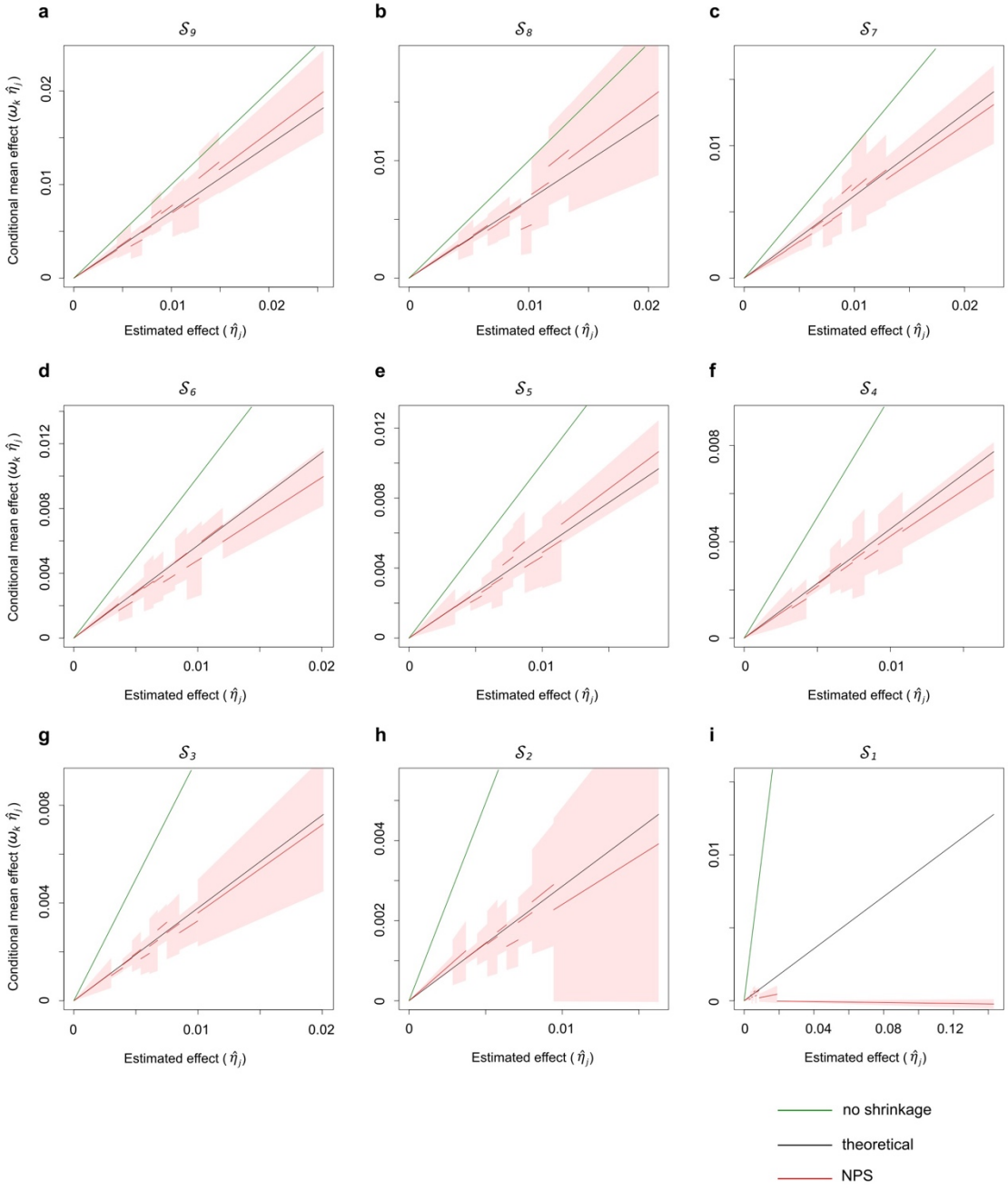
where λ_j is the eigenvalue of eigenlocus projection j, M is the number of markers, N is the sample size of discovery GWAS, h^2 is the heritability, and η_j and $\hat{\eta}_j$ are the true and estimated genetic effects, respectively, in the eigenlocus space. Using this theoretical derivation, we examined the accuracy of NPS in simulated datasets. **(a)** We partitioned the eigenlocus space into 10 subgroups, $\mathcal{S}_k = \{j \mid b_{k-1} < \lambda_j \leq b_k\}$, on intervals of eigenvalues λ_j and then estimated per-partition shrinkage weight ω_k in each partition $k = 1, \ldots, 10$ by NPS. In effect, NPS is equivalent to applying the following linear interpolation in each partition:

$$E_{j \in \mathcal{S}_k}\left[\eta_j \mid \hat{\eta}_j\right] \approx \omega_k \hat{\eta}_j$$

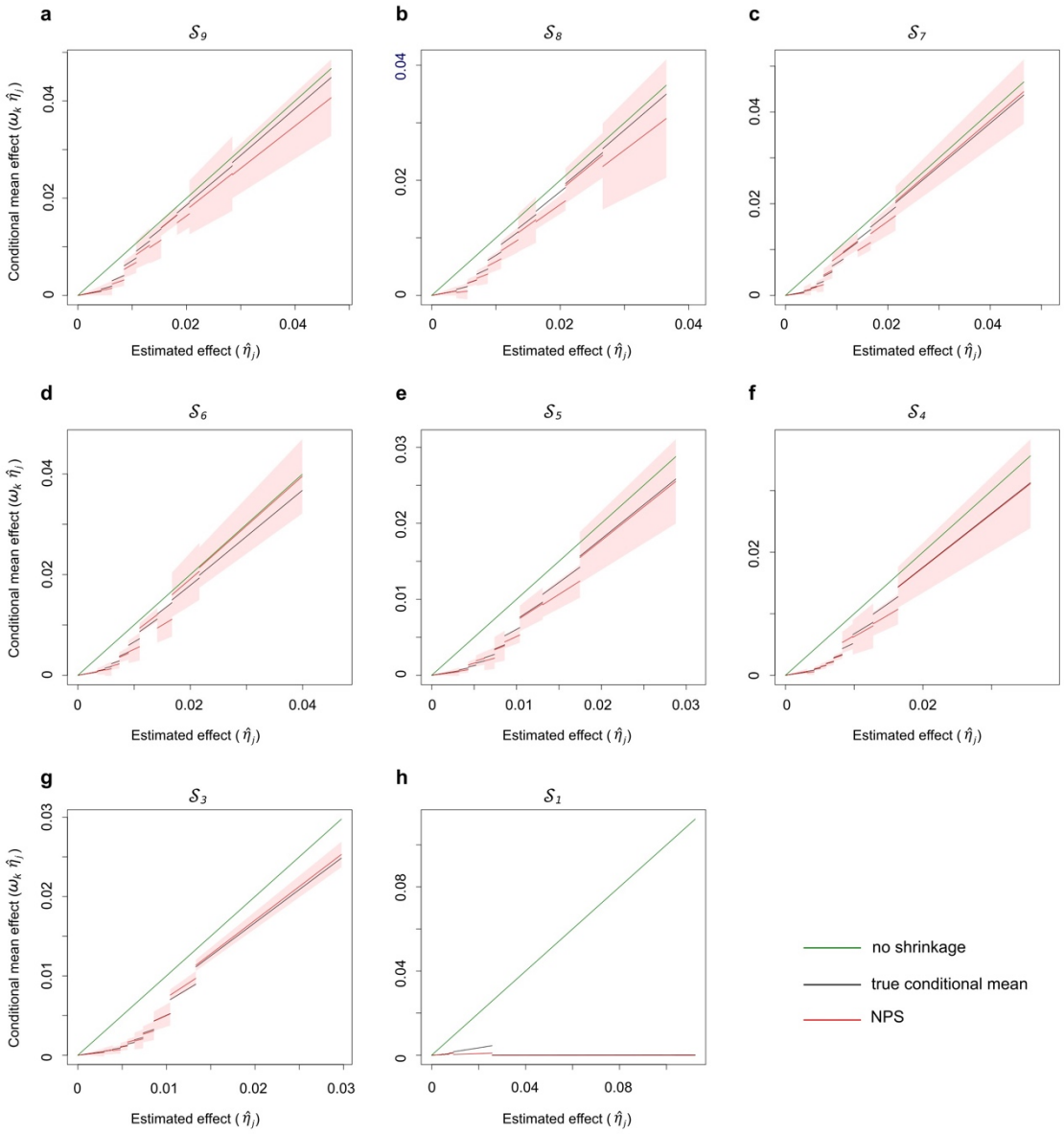
As expected, the estimated ω_k (red line) closely tracked the theoretical optimum, $\lambda_j/(\lambda_j + \frac{M}{Nh^2})$ (black line). However, in the partitions of $\mathcal{S}_1$ and $\mathcal{S}_{10}$, ω_k deviated significantly from theoretical expectation. In $\mathcal{S}_1$, $\omega_1 \approx 0$ since the eigenvectors of smallest eigenvalues are too noisy to estimate using the reference LD panel. $\mathcal{S}_{10}$ (grey box) 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 to better interpolate the theoretical curve, the total numbers of SNPs and independent projection vectors in the genome are the fundamental limiting factor. **(b)** To examine the robustness of NPS, we applied general 10-by-10 double partitioning on λ_j and $\hat{\eta}_j$. The NPS approximated the theoretical conditional mean effect, $E\left[\eta_j \mid \hat{\eta}_j\right]$, across all intervals of $\left|\hat{\eta}_j\right|$ sub-partitioning $\mathcal{S}_{10}$ (See Supplementary Fig. 3 for sub-partitions of $\mathcal{S}_1, \ldots, \mathcal{S}_9$). For both **(a)** and **(b)**, the estimated ω_k and their 95% CIs (red shade) were estimated from 5 replicates. Grey vertical lines indicate partitioning boundaries $\{b_k\}$. No shrinkage line (green) indicates $\omega_k = 1$. *M*=101,296. *N*=101,296. h^2=0.5.



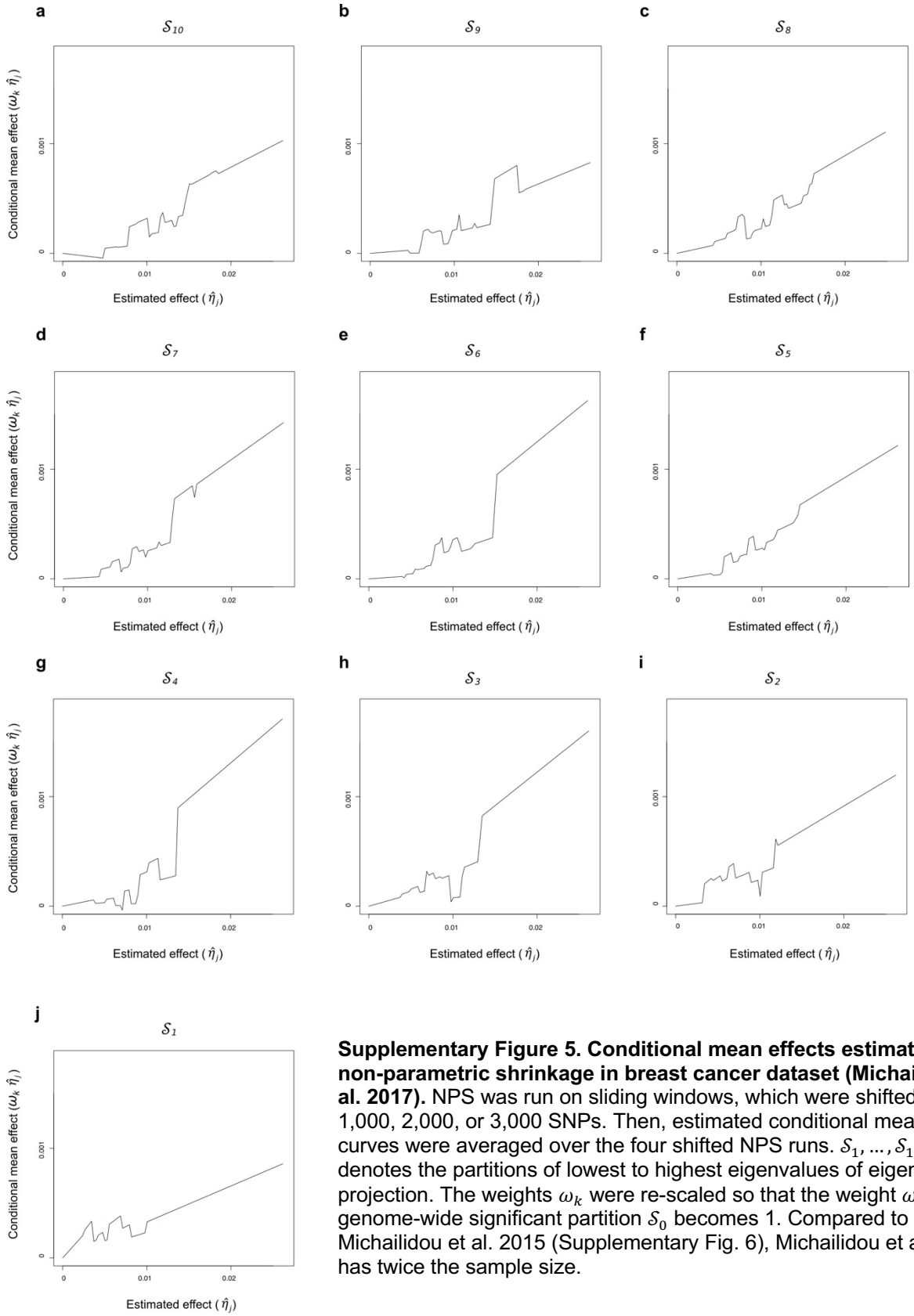
Supplementary Figure 2. Non-parametric shrinkage (NPS) approximates the conditional mean effects: non-infinitesimal genetic architecture. For non-infinitesimal genetic architectures, we do not have an analytic derivation of conditional mean effect $E[\eta_j \mid \hat{\eta}_j]$; therefore we empirically estimated the conditional means in simulations using the true underlying effects η_j and true LD structure of the population (Supplementary Note). We applied general 10-by-10 double partitioning on λ_j and $\hat{\eta}_j$. Shown here are sub-partitions for **(a)** partition of largest eigenvalues $\mathcal{S}_{10}$ and **(b)** partition of second smallest eigenvalues $\mathcal{S}_2$ (See Supplementary Fig. 4 for the rest of partitions). As expected, the true conditional mean (black line) dips for the lowest values of $\hat{\eta}_j$ but approaches no shrinkage ($\omega_k = 1$, green line) with increasing values of $\hat{\eta}_j$. A notable difference between **(a)** $\mathcal{S}_{10}$ and **(b)** $\mathcal{S}_2$ is that the true conditional mean is very close to no shrinkage for large $\hat{\eta}_j$ in the former. This is because eigenvalues are proportional to the scale of true effects η_j; therefore, with large enough eigenvalues, the sampling error becomes relatively small and the estimated effect sizes more accurate (Supplementary Note). In all partitions, conditional mean effects estimated by NPS (red line) stayed very close to the true conditional means. For both **(a)** and **(b)**, the estimated ω_k and their 95% CIs (red shade) were estimated from 5 replicates. The true conditional means were estimated over 40 simulation runs. Simulations to obtain true conditional means were completely independent from simulations to run NPS; only the genetic architecture parameters and underlying LD structure were shared between two sets of simulations. One percent of SNPs were simulated to be causal with normally distributed effect sizes. Grey vertical lines indicate partitioning boundaries $\{b_k\}$. No shrinkage line (green) indicates $\omega_k = 1$. *M*=101,296. *N*=101,296. h^2=0.5.



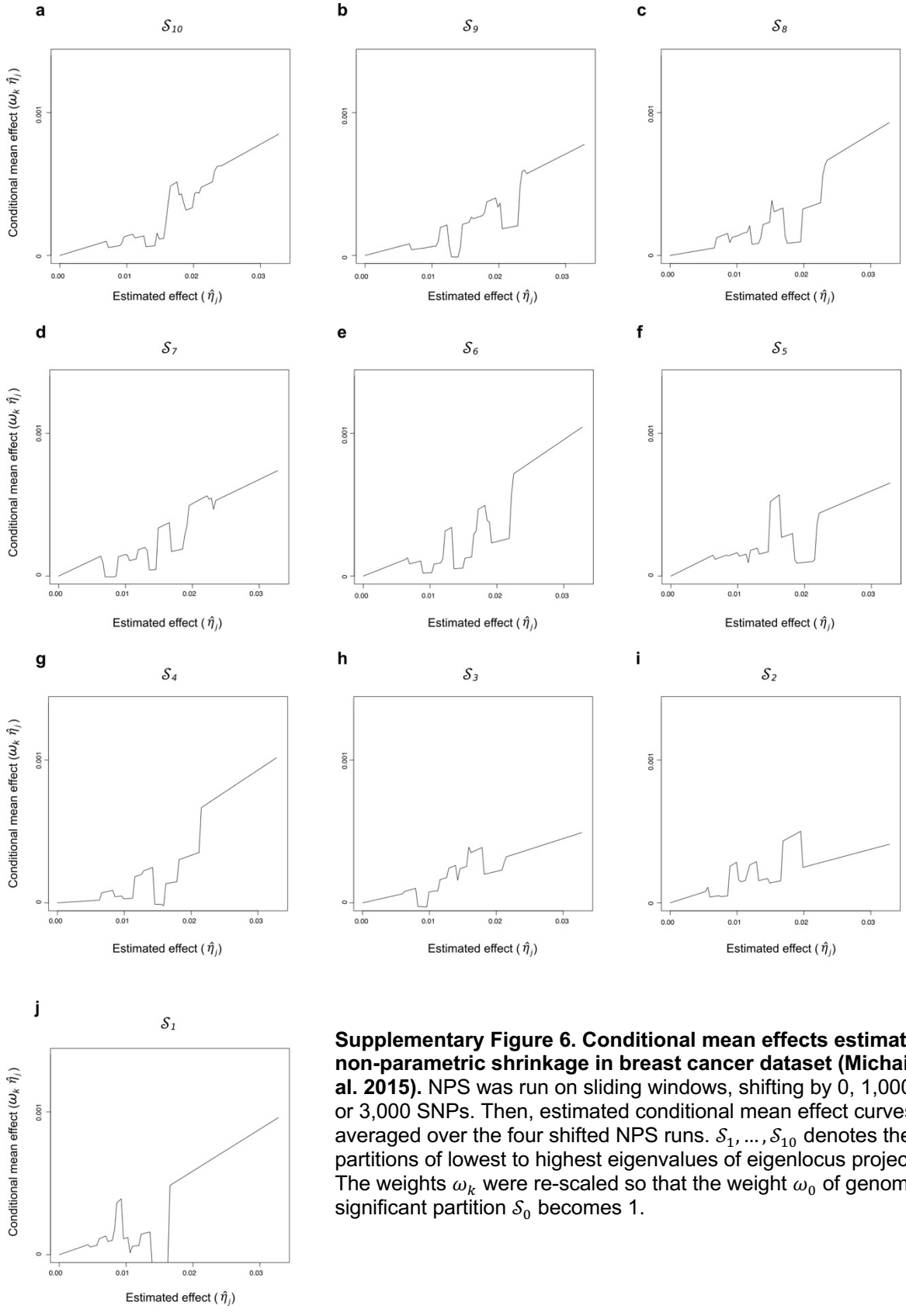
Supplementary Figure 3. Non-parametric shrinkage (NPS) approximates the conditional mean effects: infinitesimal genetic architecture ($\mathcal{S}_1, \ldots, \mathcal{S}_9$). NPS shrinkage weights ω_k (red line) were compared to the theoretical optimum (black line), $\lambda_j/(\lambda_j + \frac{M}{Nh^2})$, under the infinitesimal architecture. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ indicates the partitions of lowest to highest eigenvalues of projection (See Supplementary Fig. 1 for $\mathcal{S}_{10}$). The mean NPS shrinkage weights (red line) and their 95% CIs (red shade) were estimated from 5 replicates. 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.



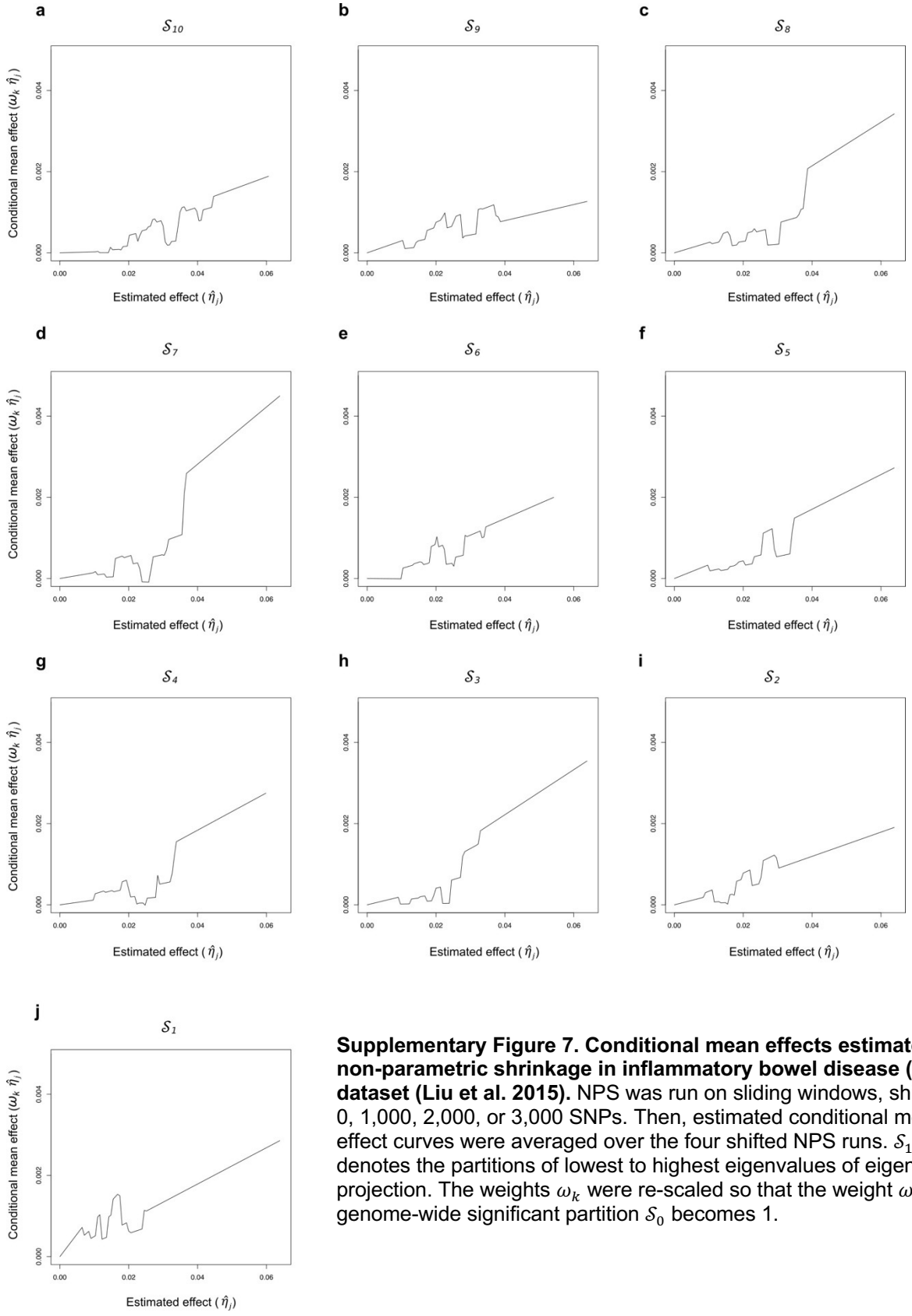
Supplementary Figure 4. Non-parametric shrinkage (NPS) approximates the conditional mean effects: non-infinitesimal genetic architecture ($\mathcal{S}_1, \mathcal{S}_3, \ldots, \mathcal{S}_9$). NPS shrinkage weights ω_k (red line) were compared to the true conditional means (black line), which were estimated empirically in 40 simulation runs. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ indicates the partitions of lowest to highest eigenvalues of projection (See Supplementary Fig. 2 for $\mathcal{S}_2$ and $\mathcal{S}_{10}$). The mean NPS shrinkage weights (red line) and their 95% CIs (red shade) were estimated from 5 replicates. 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. The fraction of causal SNPs is 1%.



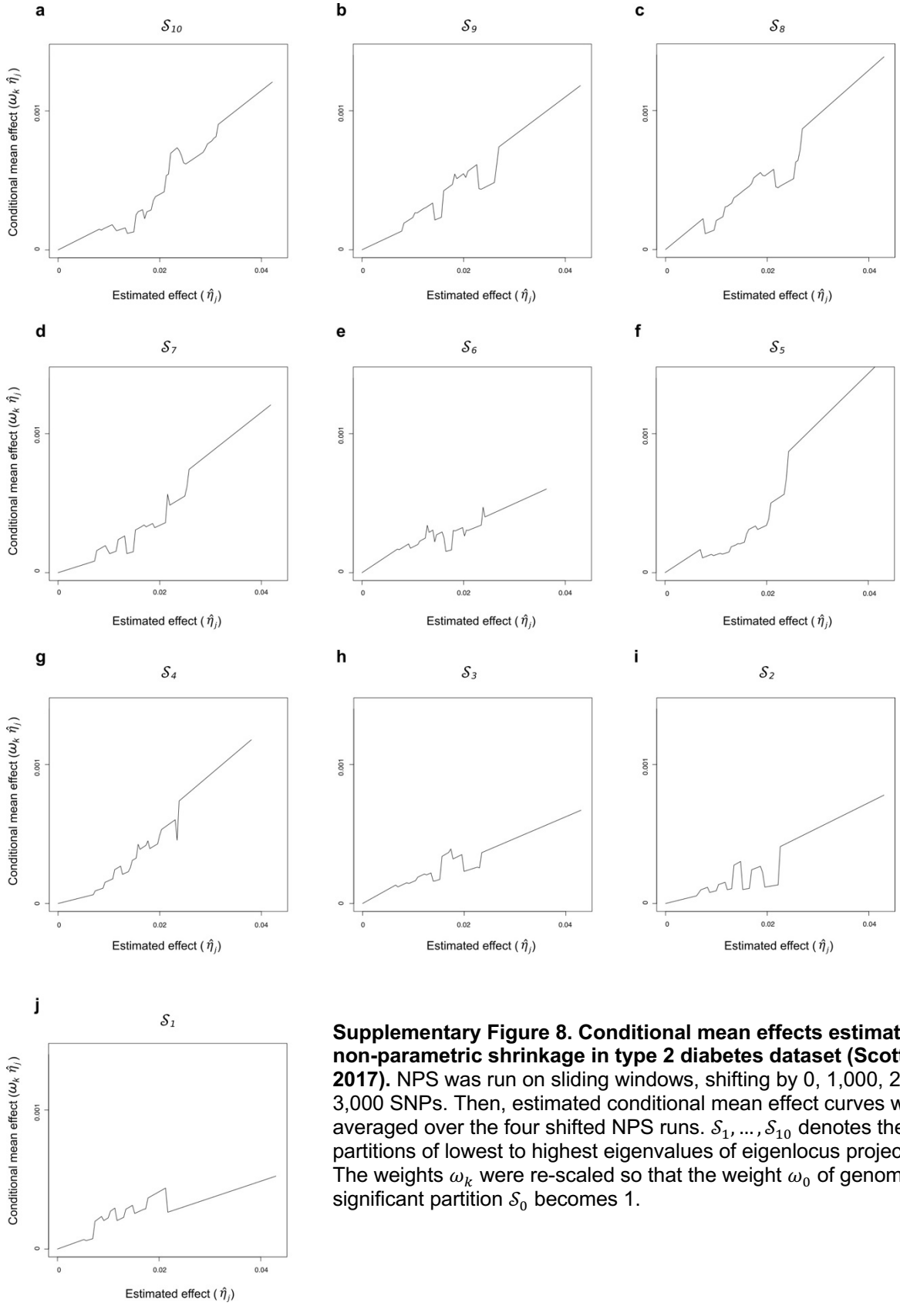
Supplementary Figure 5. Conditional mean effects estimated by non-parametric shrinkage in breast cancer dataset (Michailidou et al. 2017). NPS was run on sliding windows, which were shifted by 0, 1,000, 2,000, or 3,000 SNPs. Then, estimated conditional mean effect curves were averaged over the four shifted NPS runs. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denotes the partitions of lowest to highest eigenvalues of eigenlocus projection. The weights ω_k were re-scaled so that the weight ω_0 of genome-wide significant partition $\mathcal{S}_0$ becomes 1. Compared to Michailidou et al. 2015 (Supplementary Fig. 6), Michailidou et al. 2017 has twice the sample size.



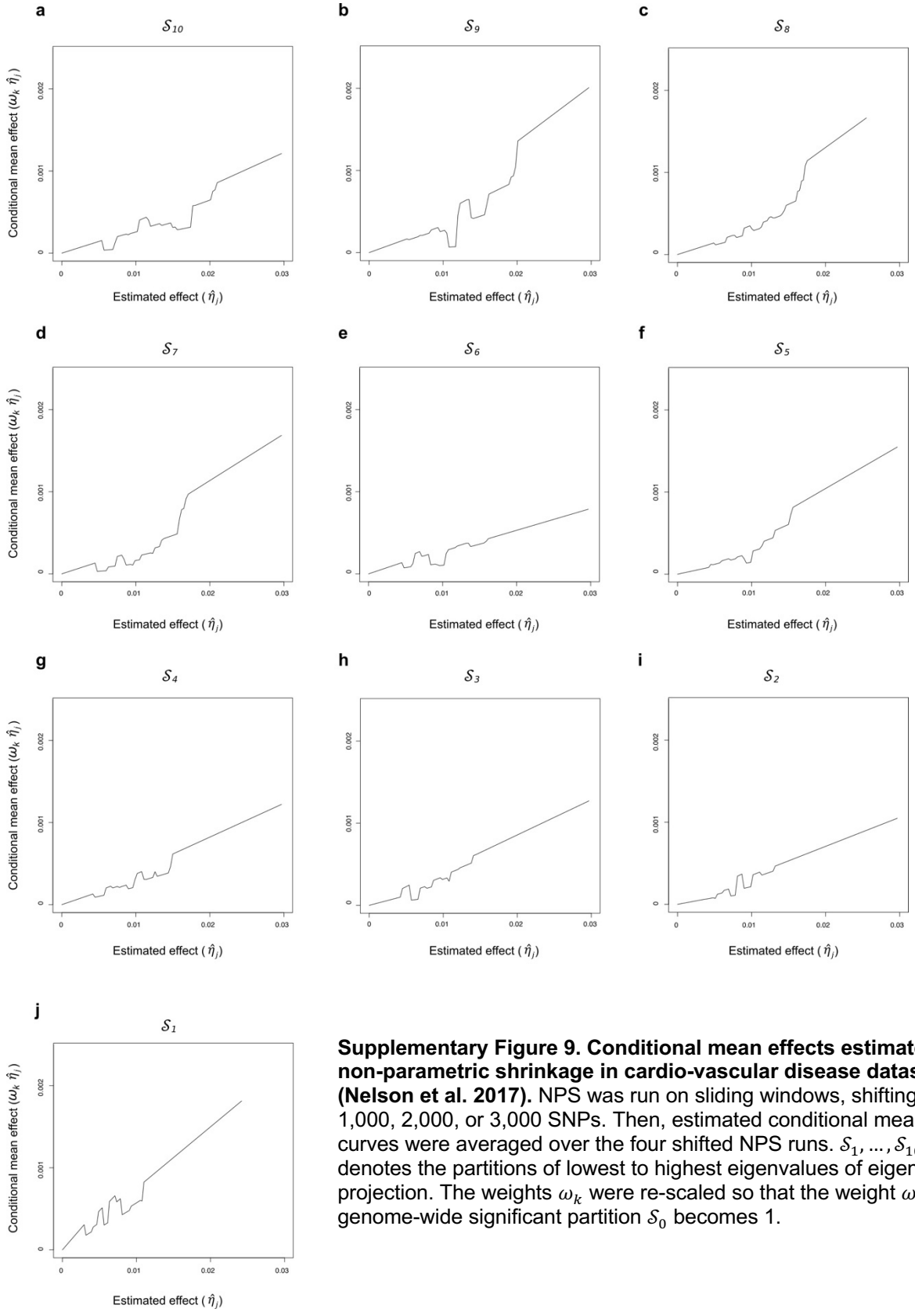
Supplementary Figure 6. Conditional mean effects estimated by non-parametric shrinkage in breast cancer dataset (Michailidou et al. 2015). NPS was run on sliding windows, shifting by 0, 1,000, 2,000, or 3,000 SNPs. Then, estimated conditional mean effect curves were averaged over the four shifted NPS runs. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denotes the partitions of lowest to highest eigenvalues of eigenlocus projection. The weights ω_k were re-scaled so that the weight ω_0 of genome-wide significant partition $\mathcal{S}_0$ becomes 1.



Supplementary Figure 7. Conditional mean effects estimated by non-parametric shrinkage in inflammatory bowel disease (IBD) dataset (Liu et al. 2015). NPS was run on sliding windows, shifting by 0, 1,000, 2,000, or 3,000 SNPs. Then, estimated conditional mean effect curves were averaged over the four shifted NPS runs. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denotes the partitions of lowest to highest eigenvalues of eigenlocus projection. The weights ω_k were re-scaled so that the weight ω_0 of genome-wide significant partition $\mathcal{S}_0$ becomes 1.



Supplementary Figure 8. Conditional mean effects estimated by non-parametric shrinkage in type 2 diabetes dataset (Scott et al. 2017). NPS was run on sliding windows, shifting by 0, 1,000, 2,000, or 3,000 SNPs. Then, estimated conditional mean effect curves were averaged over the four shifted NPS runs. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denotes the partitions of lowest to highest eigenvalues of eigenlocus projection. The weights ω_k were re-scaled so that the weight ω_0 of genome-wide significant partition $\mathcal{S}_0$ becomes 1.



Supplementary Figure 9. Conditional mean effects estimated by non-parametric shrinkage in cardio-vascular disease dataset (Nelson et al. 2017). NPS was run on sliding windows, shifting by 0, 1,000, 2,000, or 3,000 SNPs. Then, estimated conditional mean effect curves were averaged over the four shifted NPS runs. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denotes the partitions of lowest to highest eigenvalues of eigenlocus projection. The weights ω_k were re-scaled so that the weight ω_0 of genome-wide significant partition $\mathcal{S}_0$ becomes 1.

Supplementary Table 1. Comparison of prediction accuracy in genetic architectures simulating uniformly distributed causal SNPs.

Genetic Architecture	% causal SNPs	Method	Validation		R^2_{Nag} gain over	
			$R^2_{Nagelkerke}$	$R^2_{Liability}$	P+T	LDPred
(**a**) Point-Normal (GCTA)	1%	P+T	0.049	0.072		
		LDPred	0.071	0.103		
		NPS	0.080	0.116	1.63	1.12
	0.1%	P+T	0.141	0.205		
		LDPred	0.071	0.102		
		NPS	0.156	0.224	1.11	2.19
	0.01%	P+T	0.189	0.273		
		LDPred	0.076	0.110		
		NPS	0.313	0.444	1.66	4.14
(**b**) Point-Normal with MAF dependency ($\alpha = -0.25$)	1%	P+T	0.050	0.071		
		LDPred	0.073	0.101		
		NPS	0.090	0.125	1.81	1.23
	0.1%	P+T	0.142	0.206		
		LDPred	0.076	0.112		
		NPS	0.160	0.232	1.13	2.11
	0.01%	P+T	0.199	0.293		
		LDPred	0.087	0.126		
		NPS	0.310	0.444	1.56	3.55

Non-parametric shrinkage (NPS) is more accurate than Pruning and Thresholding (P+T) and Bayesian parametric method (LDPred). Here, two sets of Point-Normal architectures were simulated: (**a**) a spike-and-slab GCTA model which assumes the independence of heritability on minor allele frequency (MAF) and (**b**) an architecture incorporating the dependency of heritability on MAF ($\alpha = -0.25$). Under each model and for each causal fraction, three instances of genetic architecture were generated. Recent studies have found that low frequency SNPs contribute less heritability than previously expected under no dependency [2,3]. Low-frequency SNPs tend to be captured by eigenvectors of small eigenvalues and are challenging to handle with spectral decomposition. More realistic simulations (**b**) lowering the overall heritability contribution of low-frequency SNPs made non-parametric shrinkage prediction slightly more accurate than (**a**) GCTA models. The heritability was set to 0.5 on the liability scale, and the prevalence of was 5%. The number of markers was 5,012,500. The GWAS sample size was 100,000. Prediction models were optimized in the training cohort of 2,500 cases and 2,500 controls. The prediction R^2 was measured in the validation cohort of 50,000 samples and averaged over three simulations.

Supplementary Table 2. Accuracy of non-parametric shrinkage in genetic architectures simulating the enrichment of causal SNPs within DNase I Hypersensitive Sites (DHS).

Fraction of causal SNPs	Training	Validation		
	AUC	R^2_{Nag}	$R^2_{Liability}$	AUC
1%	0.737	0.074	0.113	0.698
	0.731	0.079	0.118	0.704
	0.724	0.082	0.119	0.708
0.1%	0.784	0.171	0.238	0.791
	0.793	0.179	0.247	0.799
	0.777	0.166	0.237	0.787
	0.799	0.165	0.240	0.790
	0.786	0.157	0.234	0.784
	0.799	0.163	0.238	0.787
0.01%	0.876	0.311	0.444	0.880
	0.886	0.307	0.443	0.876
	0.879	0.326	0.451	0.884

Each row represents the prediction accuracy of non-parametric shrinkage (NPS) in an individual simulation run. The prediction accuracy of NPS went down slightly compared to simulations of uniformly distributed causal SNPs (Supplementary Table 1) but still remained robust even if we did not explicitly account for DHS overlap in the current version of NPS. The causal fractions of 1% and 0.01% were replicated three times each, and the causal fraction of 0.1% were replicated six times. The simulation incorporates the dependency of heritability on minor allele frequency ($\alpha = -0.25$) and five-fold enrichment of causal SNPs in DHS elements. The heritability was set to 0.5 on the liability scale with the case prevalence of 5%. The number of markers was 5,012,500. The GWAS sample size was 100,000. Prediction models were optimized in the training cohort of 2,500 cases and 2,500 controls. The prediction R^2 was measured in validation cohorts of 50,000 samples. AUC – Area Under the Curve.

Supplementary Table 3. Accuracy of LDPred in genetic architectures simulating the enrichment of causal SNPs within DNase I Hypersensitive Sites (DHS).

Fraction of causal SNPs (p)	Input SNPs	Training		Validation		
		Estimated p	AUC	R^2_{Nag}	$R^2_{Liability}$	AUC
1%	All SNPs (M=5,012,500)	1.0	0.706	0.065	0.100	0.684
		1.0	0.695	0.068	0.102	0.689
		1.0	0.686	0.071	0.105	0.693
0.1%		0.3	0.695	0.080	0.108	0.705
		1.0	0.690	0.083	0.116	0.711
		1.0	0.686	0.075	0.107	0.699
		0.3	0.698	0.078	0.118	0.704
		1.0	0.693	0.069	0.103	0.694
		0.1	0.644	0.098	0.140	0.727
0.01%		0.3	0.726	0.093	0.141	0.721
		0.3	0.723	0.098	0.143	0.729
		0.01	0.840	0.268	0.373	0.854
1%	Genotyped SNPs Only (M=490,504)	1.0	0.699	0.062	0.094	0.680
		1.0	0.683	0.062	0.095	0.680
		1.0	0.674	0.066	0.095	0.687
0.1%		0.003	0.756	0.149	0.210	0.773
		1.0	0.679	0.079	0.106	0.707
		0.0001	0.729	0.116	0.165	0.715
		0.001	0.765	0.138	0.197	0.764
		0.3	0.718	0.100	0.144	0.730
		0.0003	0.753	0.123	0.183	0.753
0.01%		0.0003	0.786	0.150	0.222	0.780
		0.001	0.749	0.115	0.166	0.743
		0.001	0.816	0.222	0.317	0.827

Each row represents the prediction accuracy of LDPred in an individual simulation run. The causal fractions of 1% and 0.01% were replicated three times each, and 0.1% was replicated six times. The simulation incorporates the dependency of heritability on MAF ($\alpha = -0.25$) and five-fold enrichment of causal SNPs in DHS. The h^2 was 0.5 with the prevalence of 5%. LDPred was run using all 5,012,500 SNPs (top) as well as a sparse set of 490,504 SNPs taken from HumanHap550v3 genotyping array (bottom). With sparse SNPs, LDPred converged to closer-to-truth simulated causal fractions and resulted a higher average but lower maximum accuracy than using all markers. The prediction model reaching the highest accuracy in a training cohort was selected for validation. The estimated causal fraction (p) represents the causal fraction of best performing prediction model in training. p=1.0 denotes the infinitesimal model in which all SNPs are causal. The GWAS sample size was 100,000. Prediction models were optimized in the training cohort of 2,500 cases and 2,500 controls. The prediction R^2 was measured in validation cohorts of 50,000 samples. AUC – Area Under the Curve.

Supplementary Table 4. Accuracy of pruning and thresholding in genetic architectures simulating the enrichment of causal SNPs within DNase I Hypersensitive Sites (DHS).

True	Training			Validation		
causal SNPs	P cutoff	# SNPs	AUC	R^2_{Nag}	$R^2_{Liability}$	AUC
1%	0.046	57,816	0.680	0.047	0.072	0.662
	0.097	92,163	0.661	0.050	0.076	0.664
	0.153	121,820	0.664	0.054	0.075	0.670
0.1%	0.0001	2,082	0.783	0.174	0.244	0.793
	0.00015	2,562	0.751	0.133	0.186	0.761
	0.0002	2,765	0.735	0.119	0.164	0.747
	0.0001	2,147	0.795	0.160	0.247	0.787
	0.0001	2,296	0.736	0.105	0.163	0.738
	0.00015	2,529	0.759	0.128	0.190	0.757
0.01%	0.0001	1,662	0.827	0.209	0.305	0.823
	0.0001	1,631	0.807	0.176	0.263	0.797
	0.0001	1,553	0.833	0.252	0.352	0.848

Each row represents the prediction accuracy of pruning and thresholding (P+T) algorithm in an individual simulation run. The causal fractions of 1% and 0.01% were replicated three times each, and the causal fraction of 0.1% were replicated six times. The simulation incorporates the dependency of heritability on minor allele frequency ($\alpha = -0.25$) and five-fold enrichment of causal SNPs in DHS elements. The heritability was set to 0.5 on the liability scale with the case prevalence of 5%. The prediction model reaching the highest accuracy in a training cohort was selected for validation. The P-value cutoff of best-performing model is reported here along with the number of SNPs after pruning and thresholding. The GWAS sample size was 100,000. Prediction models were optimized in the training cohort of 2,500 cases and 2,500 controls. The prediction R^2 was measured in validation cohorts of 50,000 samples. AUC – Area Under the Curve.

Supplementary Table 5. Accuracy of non-parametric shrinkage applied to real GWAS summary statistics and UK Biobank datasets.

	Training		Validation (UK Biobank)	
GWAS	**# SNPs**	**AUC**	**AUC**	**Tail OR (5%)**
Breast Cancer 2015	5,755,927	0.654	0.620 [0.61-0.63]	2.50 [2.1-3.0]
Breast Cancer 2017	6,063,180	0.668	0.643 [0.63-0.66]	2.86 [2.4-3.4]
IBD	5,784,396	0.676	0.649 [0.63-0.66]	3.19 [2.6-4.0]
Type 2 Diabetes	5,827,280	0.661	0.651 [0.64-0.66]	2.93 [2.6-3.3]

GWAS summary statistics for breast cancer, inflammatory bowel disease (IBD) and type 2 diabetes were obtained from Michailidou et al. 2015, Michailidou et al. 2017, Liu et al. 2015 and Scott et al. 2017, respectively. The training and validation cohorts were both assembled using UK Biobank samples (see Table 2 for case/control sample sizes). The tail OR denotes the odds ratio at the 5% highest risk tail compared to the rest of cohort. The 5% cutoff of polygenic score distribution in the unascertained population was determined by resampling at known disease prevalence in UK biobank. The numbers in brackets are the 95% confidence intervals for AUC (Area Under the Curve) and tail OR, which were estimated by bootstrapping.

Supplementary Table 6. Accuracy of LDPred applied to real GWAS summary statistics and UK Biobank datasets.

GWAS	Training			Validation (UK Biobank)	
	# SNPs	Estimated causal fraction	AUC	AUC	Tail OR (5%)
Breast Cancer 2015	3,417,759	0.01	0.630	0.621 [0.61-0.63]	2.33 [1.9-2.8]
Breast Cancer 2017	3,478,993	0.1	0.621	0.618 [0.61-0.63]	2.54 [2.1-3.0]
IBD	3,396,783	0.03	0.640	0.635 [0.62-0.65]	2.71 [2.2-3.4]
Type 2 Diabetes	3,451,818	0.01	0.642	0.639 [0.63-0.65]	2.78 [2.5-3.2]
Breast Cancer 2015	351,917	0.3	0.605	0.600 [0.59-0.61]	2.37 [2.0-2.9]
Breast Cancer 2017	353,627	1.0	0.606	0.608 [0.60-0.62]	2.09 [1.7-2.5]
IBD	353,325	1.0	0.618	0.620 [0.60-0.64]	2.71 [2.2-3.4]
Type 2 Diabetes	354,110	0.1	0.640	0.643 [0.63-0.65]	2.88 [2.5-3.3]

GWAS summary statistics for breast cancer, inflammatory bowel disease (IBD) and type 2 diabetes were obtained from Michailidou et al. 2015, Michailidou et al. 2017, Liu et al. 2015 and Scott et al. 2017, respectively. The training and validation cohorts were both assembled using UK Biobank samples (see Table 2 for case/control sample sizes). LDPred was ran using all hard-called common SNPs (top) as well as directly genotyped SNPs (bottom). LDPred runs only with genotypes and automatically excludes complementary alleles; therefore, the number of input SNPs are fewer than the number of all available imputed SNPs. The estimated causal fraction represents the causal fraction parameter of best performing prediction model in training cohort. The estimated causal fraction of 1.0 denotes the infinitesimal model in which all SNPs are causal. The tail OR denotes the odds ratio at the 5% highest risk tail compared to the rest of cohort. The 5% cutoff of polygenic score distribution in the unascertained population was determined by resampling at known disease prevalence in UK biobank. The numbers in brackets are the 95% confidence intervals for AUC (Area Under the Curve) and tail OR, which were estimated by bootstrapping.

Supplementary Table 7. Accuracy of pruning and thresholding applied to real GWAS summary statistics and UK Biobank datasets.

GWAS	Training			Validation (UK Biobank)	
	P cutoff	# SNPs	AUC	AUC	Tail OR (5%)
Breast Cancer 2015	0.0001	427	0.615	0.611 [0.60-0.62]	2.28 [1.9-2.7]
Breast Cancer 2017	0.0003	1,516	0.627	0.625 [0.61-0.64]	2.25 [1.9-2.7]
IBD	0.0002	621	0.648	0.643 [0.63-0.66]	2.85 [2.3-3.6]
Type 2 Diabetes	0.0004	691	0.613	0.616 [0.61-0.63]	2.29 [2.0-2.6]

GWAS summary statistics for breast cancer, inflammatory bowel disease (IBD) and type 2 diabetes were obtained from Michailidou et al. 2015, Michailidou et al. 2017, Liu et al. 2015 and Scott et al. 2017, respectively. The training and validation cohorts were both assembled using UK Biobank samples (see Table 2 for case/control sample sizes). The prediction model reaching the highest accuracy in a training cohort was selected for validation. The P-value cutoff of best-performing model is reported here along with the number of SNPs after pruning and thresholding. The tail OR denotes the odds ratio at the 5% highest risk tail compared to the rest of cohort. The 5% cutoff of polygenic score distribution in the unascertained population was determined by resampling at known disease prevalence in UK biobank. The numbers in brackets are the 95% confidence intervals for AUC (Area Under the Curve) and tail OR, which were estimated by bootstrapping.

Supplementary Table 8. Accuracy of non-parametric shrinkage in independent validation cohorts.

	Training		Validation (Partners Biobank)	
GWAS	**# SNPs**	**AUC**	**AUC**	**Tail OR (5%)**
Breast Cancer 2017	5,755,927	0.654	0.611 [0.59-0.63]	2.08 [1.6-2.7]
IBD	6,063,180	0.668	0.669 [0.65-0.69]	3.81 [3.1-4.7]
Type 2 Diabetes	5,784,396	0.676	0.606 [0.59-0.62]	2.04 [1.7-2.4]
CAD	5,741,641	0.698	0.603 [0.57-0.64]	3.27 [2.2-4.7]

The polygenic risk models trained in UK Biobank (Table 2 and Supplementary Table 5) were validated in US white population (Partners Biobank). GWAS summary statistics for coronary artery disease (CAD) were obtained from Nelson et al. 2017. See Table 3 for case/control sample sizes of validation cohorts. The tail OR denotes the odds ratio at the 5% highest risk tail compared to the rest of cohort. The numbers in brackets are the 95% confidence intervals for AUC (Area Under the Curve) and tail OR, which were estimated by DeLong's method and bootstrapping, respectively.

Supplementary Table 9. Accuracy of LDPred in independent validation cohorts.

	Training (UK Biobank)			Validation (Partners Biobank)	
GWAS	**# SNPs**	**Est causal**	**AUC**	**AUC**	**Tail OR (5%)**
Breast Cancer 2017	1,261,292	0.1	0.600	0.580 [0.56-0.60]	1.78 [1.3-2.3]
IBD	1,238,654	0.03	0.609	0.639 [0.62-0.66]	3.07 [2.5-3.8]
Type 2 Diabetes	1,243,787	0.01	0.618	0.597 [0.58-0.61]	1.81 [1.5-2.2]
CAD	1,237,683	0.003	0.715	0.595 [0.56-0.63]	2.22 [1.4-3.3]

The polygenic risk models were trained with LDPred in UK Biobank cohorts and validated in US white population (Partners Biobank). The training cohorts for breast cancer, inflammatory bowel disease (IBD) and type 2 diabetes are identical to those in Table 2 and Supplementary Table 6. However, the prediction models were reconstructed by re-running LDPred on the SNPs found in both training and validation cohorts as recommended by the authors. LDPred runs only with genotypes and automatically excludes complementary alleles; therefore, the number of hard-called input SNPs are fewer than the number of all available imputed SNPs. The estimated causal fraction represents the causal fraction parameter of best performing prediction model in training cohort. The estimated causal fraction of 1.0 denotes the infinitesimal model in which all SNPs are causal. See Table 3 for case/control sample sizes of validation cohorts. The tail OR denotes the odds ratio at the 5% highest risk tail compared to the rest of cohort. The numbers in brackets are the 95% confidence intervals for AUC (Area Under the Curve) and tail OR, which were estimated by DeLong's method and bootstrapping, respectively.

Supplementary Table 10. Accuracy of pruning and thresholding in independent validation cohorts.

GWAS	Training (UK Biobank)			Validation (Partners Biobank)	
	P cutoff	# SNPs	AUC	AUC	Tail OR (5%)
Breast Cancer 2017	0.00035	801	0.613	0.589 [0.57-0.61]	1.56 [1.2-2.1]
IBD	0.0002	331	0.629	0.659 [0.64-0.68]	3.57 [2.9-4.4]
Type 2 Diabetes	0.0001	165	0.603	0.577 [0.56-0.59]	1.78 [1.5-2.1]
CAD	0.3878	33,078	0.717	0.612 [0.58-0.65]	3.05 [2.1-4.4]

The polygenic risk models were trained with pruning and thresholding algorithm in UK Biobank cohorts and validated in US white population (Partners Biobank). The training cohorts for breast cancer, inflammatory bowel disease (IBD) and type 2 diabetes are identical to those in Table 2 and Supplementary Table 7. However, the prediction models were reconstructed by re-running pruning and thresholding algorithm on the SNPs found in both training and validation cohorts. The prediction model reaching the highest accuracy in a training cohort was selected for validation. The P-value cutoff of best-performing model is reported here along with the number of SNPs after pruning and thresholding. See Table 3 for case/control sample sizes of validation cohorts. The tail OR denotes the odds ratio at the 5% highest risk tail compared to the rest of cohort. The numbers in brackets are the 95% confidence intervals for AUC (Area Under the Curve) and tail OR, which were estimated by DeLong's method and bootstrapping, respectively.

Supplementary Note

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 $l \in \{1, \dots, L\}$, 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 observed 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 eigenvalue 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 later.

Now we introduce a linear decorrelating transformation $\mathcal{P}$, which projects summary statistics and genotypes into a decorrelated space which we call "**eigenlocus space**." The projection $\mathcal{P}$ is called "**eigenlocus projection**" and defined as the following:

$$\mathcal{P} \coloneqq \boldsymbol{\Lambda}^{-\frac{1}{2}}\,\boldsymbol{Q}^{\boldsymbol{T}} \qquad (Eq\ S1)$$

The effect size estimates $\hat{\eta}$ and projected genotypes $\boldsymbol{X}'^{P}$ in the eigenlocus space are obtained by applying the eigenlocus projection $\mathcal{P}$ on GWAS effect sizes $\hat{\beta}$ and genotypes $\boldsymbol{X}'$ as follows:

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

Distribution of projected genotypes in the eigenlocus space

Let X'_i be an m-dimensional genotype vector of training sample 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 applying $\mathcal{P}$ on X_i' by definition (Eq S2), $\mathrm{X}_i'^P$ 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 in the eigenlocus space are decorrelated with the covariance of $\boldsymbol{I}$.

Distribution of effect size estimates in the eigenlocus space

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:

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

since $\boldsymbol{D} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$. Since the estimated effect size $\hat{\eta}$ in the eigenlocus space is obtained by applying $\mathcal{P}$ on $\hat{\beta}$ by definition (Eq S2), the distribution of $\hat{\eta}$ given β also follows a multivariate normal distribution:

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

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

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

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

Interpretation of eigenvalues

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

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

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 when all SNPs have the same variance of per-SNP effect sizes σ_g^2,

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

since $\sum_{s=1}^{m} q_{sj}^2 = 1$.

Conditional mean effects under infinitesimal genetic architecture in the eigenlocus space

Under infinitesimal genetic architecture, the conditional mean effect has been analytically derived by Vilhjalmsson et al.[1]:

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

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

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

and

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

we can reformulate Eq S5 as follows:

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

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

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

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

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

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

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

by the definition of η (Eq S4). Therefore, for the eigenlocus projection j 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:

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

General partitioning strategy

In general, we approximate the conditional mean effect $E\left[\eta_j \mid \hat{\eta}_j\right]$ by piecewise linear interpolation as follows:

$$E_{j\in\mathcal{S}_k}\left[\eta_j \mid \hat{\eta}_j\right] \approx \omega_k\,\hat{\eta}_j$$

where ω_k is per-partition shrinkage weight in the eigenlocus space.

In the infinitesimal architecture, ω_k will depend only on eigenvalues λ_j, therefore, partitioning needs to be done only on intervals of λ_j. However, in general, ω_k depend on both λ_j and $\hat{\eta}_j$. Therefore, in general we need to apply double-partitioning on λ_j and $\hat{\eta}_j$ in the eigenlocus space. The intuition

behind this double-dependency is the following: 1) The scale of true eigenlocus effects, namely $var[\eta_j]$, is proportional to the eigenvalue λ_j. On the other hand, the sampling error, namely $\hat{\eta}_j \mid \eta_j$, is fixed at $1/N$, where N is the sample size of discovery GWAS cohort. Therefore, with increasing eigenvalues, estimated effects $\hat{\eta}_j$ become more reliable since the scale of η_j becomes larger relative to the sampling error. In contrast, with decreasing eigenvalues, $\hat{\eta}_j$ becomes dominated by sampling error. 2) When underlying genetic architecture has a low polygenicity, i.e. a small proportion of causal SNPs, some of eigenlocus 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 will be enriched among small estimates of $\hat{\eta}_j$. Thus, in such a low-polygenic architecture, $E[\eta_j \mid \hat{\eta}_j] \approx 0$ for smaller values of $\hat{\eta}_j$. In contrast, as the polygenicity approaches the infinitesimal architecture, ω_k will lose the dependency on $\hat{\eta}_j$ and become a constant depending only on λ_j.

Estimation of shrinkage weights using training data

We rely on a small independent cohort with full genotype information (training cohort) to estimate the shrinkage weights ω. Let us assume that the eigenlocus space is partitioned into *K* disjoint subsets, $\mathcal{S}_1, \ldots, \mathcal{S}_k$ on intervals of eigenvalues λ_j and estimated effects $\hat{\eta}_j$. Then, we define a partitioned risk score G_{ik} in the eigenlocus space as follows:

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

Then, the predicted phenotype $\hat{y}_i$ of individual i becomes:

$$\hat{y}_i = \sum_j E[\eta_j \mid \hat{\eta}_j]\, x_{ij}^P = \sum_j \left(\sum_k \omega_k \hat{\eta}_j I(j \in \mathcal{S}_k) \right) x_{ij}^P = \sum_{k=1}^{K} \omega_k \left(\sum_{j \in \mathcal{S}_k} \hat{\eta}_j x_{ij}^P \right) = \sum_{k=1}^{K} \omega_k G_{ik}$$

by applying piece-wide linear interpolation on $E[\eta_j \mid \hat{\eta}_j]$ and changing the order of summation.

For quantitative traits, the per-partition shrinkage weights ω_k can be estimated by applying linear regression to training data. For binary phenotypes, ω_k can be learned 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. We claim that the distributions of *K*-dimensional vector of partitioned risk scores $G_i \mid y_i$ of individual i satisfies the following:

$$G_i \mid y_i = 1 \sim N(\mu_{case}, \Sigma_{case})$$

$$G_i \mid y_i = 0 \sim N(\mu_{control}, \Sigma_{control})$$

and

$$\Sigma_{case} \approx \Sigma_{control}$$

where μ_{case} and $\mu_{control}$ are mean partitioned risk scores among cases and controls, respectively, and Σ_{case} and $\Sigma_{control}$ are $K \times K$ covariance matrices of partitioned risk scores in each subgroup. This is

because the partitioned risk scores of cases and controls, namely $G_{ik} \mid y_i$, follow approximately normal distributions as long as each partition consists of a sufficient number of eigenloci [4]. The variance of partitioned risk scores is approximately equal between cases and controls since G_{ik} of an individual partition explains only a small fraction of phenotypic variation on the observed scale in typical GWAS data [5]. Furthermore, Σ_{case} and $\Sigma_{control}$ are both approximately diagonal; although in theory, the liability thresholding effect induces slight non-zero covariance between partitions, this effect is typical small and negligible.

LDA-derived shrinkage weights can be independently estimated for each partition and simplify to:

$$\omega_k \approx 2\frac{E[\,G_{ik}|\,y_i = 1] - E[\,G_{ik}|\,y_i = 0]}{var[\,G_{ik}|\,y_i = 1] + var[\,G_{ik}|\,y_i = 0]} \qquad (Eq\ S8)$$

The discriminant function is similar when covariates are included [6].

Back-conversion from the eigenlocus space to the original SNP space

Let $\hat{y}$ be an N'-dimensional vector of predicted phenotypes in a training cohort. We construct $\hat{y}$ by summing over all projected genotypes multiplied by conditional mean effects in the eigenlocus 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 of training samples in the original SNP space. Thus, $E[\eta_l \mid \hat{\eta}_l]$ can be converted back to per-SNP effect sizes by the following transformation:

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

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, eigenvalue 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:

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

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

where λ_j and q_j are the j-th positive eigenvalues and corresponding 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 can re-derive 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,

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

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

$$= \boldsymbol{Q}'\left(\frac{M}{Nh^2}\boldsymbol{I}_r + \boldsymbol{\Lambda}'\right)^{-\mathbf{1}} \boldsymbol{Q}'^{\boldsymbol{T}}\hat{\beta} \qquad (Eq\ S9)$$

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

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

In Eq S6, M/Nh^2 term remains for $q_{r+1}, \dots, q_m$ whereas it is truncated in Eq S9.

Simulations to show that NPS approximates the conditional mean effects

To show that shrinkage weights estimated by NPS approximate conditional mean effects in the eigenlocus space (Supplementary Figs. 1-4), we simulated genetic architecture with SNPs in LD. The LD matrix was calculated using the genotypes of the 1000 Genomes Project CEU panel (n=99) for a total of 101,296 SNPs, which were obtained by 10-fold down-sampling of SNPs in Illumina HumanHap550 genotyping array. 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: an infinitesimal model with normally distributed effect sizes and a non-infinitesimal model for which only 1% of SNPs are casual with normally distributed effect sizes. The discovery GWAS summary statistics were directly sampled from the following *m*-dimensional multivariate normal distribution (MVN) one locus each time:

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

where β and $\hat{\beta}$ are *m*-dimensional vectors of true and estimated effect sizes of SNPs in the locus, respectively, $\boldsymbol{D}$ is a local LD matrix, *m* is the number of SNPs in the locus, 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 from an MVN as follows:

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

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 this simulation, we ran NPS treating each locus as a single analysis window without averaging over sliding windows since LD was assumed to be confined in each window. The true underlying LD matrix $\boldsymbol{D}$ was hidden, and NPS estimated the reference LD from the training cohort. For the infinitesimal model, the theoretically optimal shrinkage weight ω_k^o is known [1]:

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

For the non-infinitesimal model, we do not have analytically known optimal shrinkage, therefore instead empirically estimated it by regressing conditional mean effects on $\hat{\eta}_j$ 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 averaged 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 average of true decorrelated effects η_j in each partition $\mathcal{S}_k$. We calculated $\eta_j = \sqrt{\lambda_j}\,\boldsymbol{q}_j^T\beta$ using the true genetic effects β and spectral decomposition of true population LD matrix. To make sure that the consistency between estimated and true mean conditional effects are not due to shared underlying data, at each simulation run, we re-generated the entire dataset starting from fresh sampling of β but under the same genetic architecture parameters.

References

1. Vilhjalmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97,** 576–592 (2015).
2. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49,** 986–992 (2017).
3. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50,** 746–753 (2018).
4. 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).
5. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* **9,** e1003348 (2013).
6. Lachenbruch, P. A. Covariance adjusted discriminant functions. *Ann. Inst. Stat. Math.* **29,** 247–257 (1977).