ARTICLE

Non-parametric Polygenic Risk Prediction via Partitioned GWAS Summary Statistics

Sung Chun,[1,2,3,4,12,13] Maxim Imakaev,[1,2,3,4,12] Daniel Hui,[1,3,5] Nikolaos A. Patsopoulos,[1,3,5] Benjamin M. Neale,[3,6,7] Sekar Kathiresan,[3,7,8,14] Nathan O. Stitziel,[9,10,11,*] and Shamil R. Sunyaev[1,2,3,4,*]

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 the 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 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 allow for linkage disequilibrium in summary statistics of 5 million dense genome-wide markers and consistently improves prediction accuracy. We show that NPS 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.

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 Goddard and Hayes[2] and Falke et al.[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 (GWASs) 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.[9–15] These methods include extensions of BLUP,[9,10,16] 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.

To leverage summary-level data from large-scale GWAS projects, an alternative approach to construct polygenic risk scores based on summary statistics has been introduced.[3,12,14,17–21] The originally proposed version is

[1] Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA; [2] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA; [3] Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [4] Altius Institute for Biomedical Sciences, Seattle, WA 98121, USA; [5] Systems Biology and Computer Science Program, Ann Romney Center for Neurological Diseases, Department of Neurology, Brigham & Women's Hospital, Boston, MA 02115, USA; [6] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; [7] Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA; [8] Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA 02114, USA; [9] Cardiovascular Division, Department of Medicine, Washington University School of Medicine, Saint Louis, MO 63110, USA; [10] Department of Genetics, Washington University School of Medicine, Saint Louis, MO 63110, USA; [11] McDonnell Genome Institute, Washington University School of Medicine, Saint Louis, MO 63110, USA

[12] These authors contributed equally to this work

[13] Present address: Division of Pulmonary Medicine, Boston Children's Hospital, Boston, MA 02115, USA

[14] Present address: Verve Therapeutics, Cambridge, MA 02142, USA

*Correspondence: nstitziel@wustl.edu (N.O.S.), ssunyaev@rics.bwh.harvard.edu (S.R.S.)

https://doi.org/10.1016/j.ajhg.2020.05.004.

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.[22] In practice, however, implementing the conditional mean predictor poses a dilemma. The GWAS-estimated effect sizes capture genetic effects of all SNPs in linkage disequilibrium (LD), so these marginal estimates have to be first deconvoluted into genetic contribution of individual causal SNPs. Furthermore, 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 extensively sampling likely combinations of causal genetic effects under a simplified model of genetic architecture. However, these methods often ignore the correlation of sampling errors of estimated effects between SNPs in LD for the sake of computational efficiency.[18,20] Such approximation can lead to a suboptimal prediction model due to double-counting of correlated sampling errors. In case of dense high-resolution GWAS data, this effect can be severe due to extensive and rank-deficient LD structures. Recent approaches account for correlated sampling errors by applying a Metropolis-Hastings technique to reject proposed states based on a full multivariate likelihood or by assuming a continuous shrinkage prior for the allelic architecture, but their prediction accuracy still depend on the convergence of high-dimensional combinatorial sampling processes, and it remains challenging to extend these models to incorporate additional complexity of true architecture.[19,21]

In spite of this methodological complexity, polygenic scores trained on large-scale datasets show some promise for practical applications in medical genetics. Polygenic scores have been used to analyze the UK Biobank, the largest epidemiological cohort that includes genetic data.[23] 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. Our method accounts for both types of correlations induced by LD in GWAS summary statistics, namely the correlations of true genetic effects as well as sampling errors, by using eigenvalue decomposition of LD matrix instead of relying on a high-dimensional sampling technique. Despite growing interest in non-parametric prediction models, thus far there has been no non-parametric polygenic score that can fully allow for LD under the conditional mean effect framework.[24–29] 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.

Material and Methods

Overview

Our approach is to partition SNPs into groups and determine the relative weights based on predictive value of each partition estimated in the training data (Figure 1A). Intuitively, when there is no LD between SNPs, a partition dominated by non-causal variants will have low power to distinguish case subjects from control subjects, whereas the partition enriched with strong signals will be more informative for predicting the phenotype. This is equivalent to approximating the conditional mean effect curve by piecewise linear interpolation. Because of LD, however, we cannot apply the partitioning method directly to GWAS effect sizes. True genetic effects as well as sampling noise are correlated between adjacent SNPs. To prevent estimated genetic signals smearing across partitions, we first transform GWAS data into an orthogonal domain, which we call "eigenlocus" (Figure 1B). Specifically, we use a decorrelating linear transformation obtained by eigenvalue decomposition of the local LD matrix. Both genotypes and sampling errors are uncorrelated in the eigenlocus representation. In this representation, however, true genetic effects do not follow analytically tractable distributions except under infinitesimal and extremely polygenic architectures. Therefore, 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.

Decorrelating Projection

We split the genome into L non-overlapping windows of m SNPs each. By default, m was set to 4,000 SNPs (~2.5 Mb on average). The window size was chosen to be large enough to capture the majority of 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 each genomic window $l \in \{1, \ldots, L\}$, let $\mathbf{X}_l$ be an $N \times m$ genotype matrix of N individuals and m SNPs in the window. We assume that the genotypes are standardized to the mean of 0 and variance of 1. Let $\widehat{\boldsymbol{\beta}}_l$ be an m-dimensional vector of observed effect sizes from a GWAS and $\boldsymbol{\beta}_l$ be an m-dimensional vector of true underlying genetic effects in window l. The scales of $\widehat{\boldsymbol{\beta}}_l$ and $\boldsymbol{\beta}_l$ are defined with respect to the standardized genotypes. Then, the LD matrix $\mathbf{D}_l$ is given by $\mathbf{D}_l = (1/N)\mathbf{X}_l^T \mathbf{X}_l$ and can be factorized by eigenvalue decomposition into $\mathbf{D}_l = \mathbf{Q}_l \, \boldsymbol{\Lambda}_l \, \mathbf{Q}_l^T$, where $\mathbf{Q}_l$ is an orthonormal matrix of eigenvectors and $\boldsymbol{\Lambda}_l$ is a diagonal matrix of eigenvalues.

Now we introduce a linear decorrelating transformation $\mathcal{P}_l$, which projects summary statistics $\widehat{\boldsymbol{\beta}}_l$ and genotypes $\mathbf{X}_l$ into a decorrelated space which we call "**eigenlocus space**." We call the projection $\mathcal{P}_l$ an "**eigenlocus projection.**" $\mathcal{P}_l$ is defined as the following:

$$\mathcal{P}_l := \boldsymbol{\Lambda}_l^{-\frac{1}{2}} \mathbf{Q}_l^T$$

By applying the eigenlocus projection on $\widehat{\boldsymbol{\beta}}_l$ and $\mathbf{X}_l$, we obtain the estimated effect sizes $\widehat{\boldsymbol{\eta}}_l$ and projected genotypes $\mathbf{X}_l^P$ in this eigenlocus space as follows:

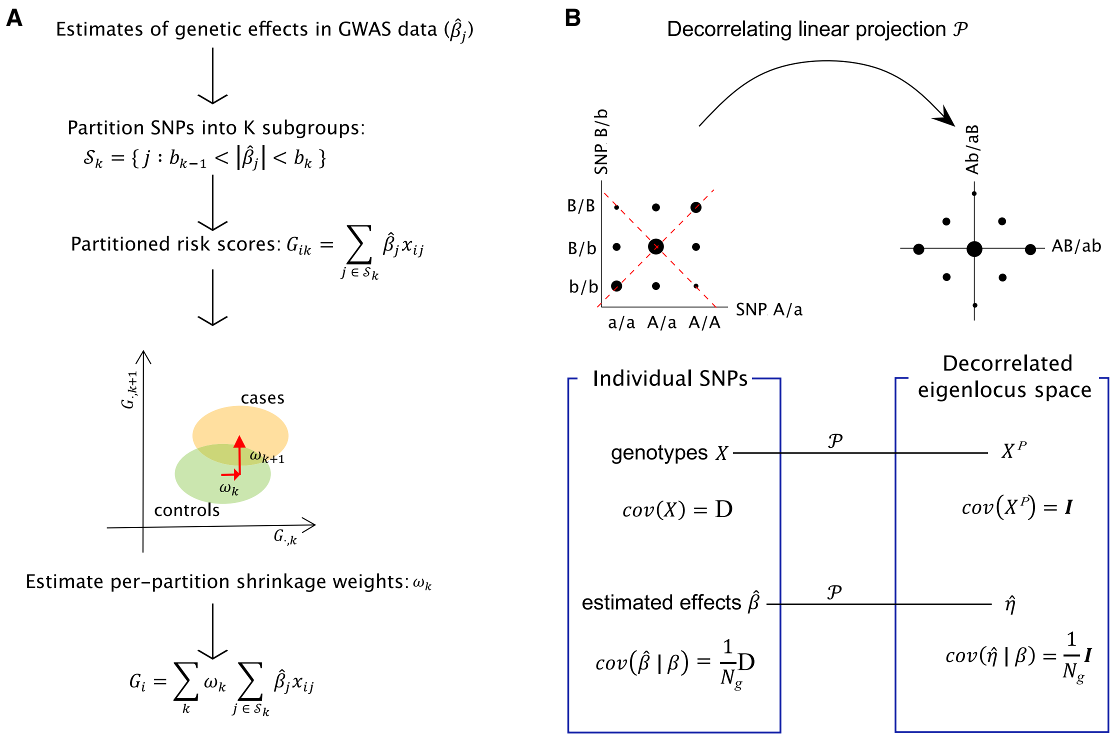
A
Estimates of genetic effects in GWAS data ($\hat{\beta}_j$)
Partition SNPs into K subgroups:
$\mathcal{S}_k = \{ j : b_{k-1} < |\hat{\beta}_j| < b_k \}$
Partitioned risk scores: $G_{ik} = \sum_{j \in \mathcal{S}_k} \hat{\beta}_j x_{ij}$
$G_{\cdot,k+1}$
cases
ω_{k+1}
ω_k
controls
$G_{\cdot,k}$
Estimate per-partition shrinkage weights: ω_k
$G_i = \sum_k \omega_k \sum_{j \in \mathcal{S}_k} \hat{\beta}_j x_{ij}$
B
Decorrelating linear projection $\mathcal{P}$
SNP B/b
B/B
B/b
b/b
a/a A/a A/A
SNP A/a
Ab/aB
AB/ab
Individual SNPs
Decorrelated eigenlocus space
genotypes X
$\mathcal{P}$
X^P
$cov(X) = \mathrm{D}$
$cov(X^P) = \boldsymbol{I}$
estimated effects $\hat{\beta}$
$\mathcal{P}$
$\hat{\eta}$
$cov(\hat{\beta} \mid \beta) = \frac{1}{N_g}\mathrm{D}$
$cov(\hat{\eta} \mid \beta) = \frac{1}{N_g}\boldsymbol{I}$

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

(A) For unlinked markers, NPS partitions SNPs into K subgroups splitting the GWAS effect sizes ($\widehat{\beta}_j$) 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 case subjects and control subjects. Estimating the per-partition shrinkage 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.

(B) For markers in LD, genotypes and estimated effects are decorrelated first by a linear projection $\mathcal{P}$ in non-overlapping windows of ~2.5 Mb in length, and then NPS is applied to the data. The size of black dots indicates genotype frequencies in population. Before projection, genotypes at SNP 1 and 2 are correlated due to LD (**D**), and thus sampling errors of estimated effects ($\widehat{\beta}_j \mid \beta_j$) 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_g is the sample size of GWAS cohort.

$$\begin{aligned} \widehat{\boldsymbol{\eta}}_l &:= \mathcal{P}_l \widehat{\boldsymbol{\beta}}_l \\ \mathbf{X}_l^P &:= \mathbf{X}_l \mathcal{P}_l^T \end{aligned} \qquad \text{(Equation 1)}$$

This projection will remove the correlation structure induced by LD in the genotypes $\mathbf{X}_l^P$ and in the sampling error of estimated effects $\widehat{\boldsymbol{\eta}}_l$. Specifically, in the eigenlocus space, $\widehat{\boldsymbol{\eta}}_l$ and $\mathbf{X}_l^P$ follow the following multivariate normal distributions (see Appendices A and B for the derivation):

$$\mathbf{X}_l^P \sim N(0, \ \mathbf{I})$$

$$\widehat{\boldsymbol{\eta}}_l \mid \boldsymbol{\beta}_l \sim N\left(\boldsymbol{\eta}_l, \ \frac{1}{N_g}\mathbf{I}\right)$$

where N_g is the sample size of GWAS from which summary statistics $\widehat{\boldsymbol{\beta}}_l$ was obtained and $\boldsymbol{\eta}_l$ is the true underlying genetic effect defined by $\boldsymbol{\eta}_l = \boldsymbol{\Lambda}_l^{1/2}\mathbf{Q}_l^T\boldsymbol{\beta}_l$.

Due to the rank-deficiency of LD matrix $\mathbf{D}_l$ and application of regularization on $\mathbf{D}_l$ (described below), the dimension of eigenlocus space m_l can be lower than the total number of SNPs m in a given window l. Specifically, we set the LD between SNPs to 0 unless the absolute value of estimated LD was greater than $5/\sqrt{N}$. This is to suppress sampling noises in off-diagonal entries of LD matrix. 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 in each window. In addition, projections corresponding to eigenvalues less than 0.5 were truncated for the computational efficiency since they were dominated by noises. 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 between projected genotypes belonging to neighboring windows. For the pairs with the absolute Pearson correlation > 0.3, we kept the one yielding a larger absolute effect size and eliminated the other.

By applying the above processing steps in each genomic window l, we obtained m_l-dimensional vector of estimated effect sizes $\widehat{\boldsymbol{\eta}}_l = \left\{\widehat{\eta}_{lj}\right\}$ and $N \times m_l$ matrix of genotypes $\mathbf{X}_l^P = \left\{x_{lij}^P\right\}$ in the eigenlocus space. Here, the index $j \in \{1, \ldots, m_l\}$ indicates an individual genetic variation yielded by applying an eigenlocus projection (Equation 1) with eigenvalues $\boldsymbol{\Lambda}_l = \{\lambda_{lj}\}$. In this representation, we can operate on each genetic variation independently from each other since they are decorrelated.

Partitioning Strategy

Since the SNPs with largest effect sizes span a wide range of values but are sampled only sparsely, we cannot reliably estimate the conditional mean effect for this large-effect tail without assuming *a priori* parametric assumption on its distribution. This issue is particularly the case for genome-wide significant SNPs. To solve this problem, we handled the genome-wide significant SNPs as a separate partition from the rest of SNPs and treat them as fixed effect estimates. Specifically, 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 $\mathbf{I}$ with eigenvalues of 1. To avoid LD across SNPs in $\mathcal{S}_0$, genome-wide significant SNPs were selected into $\mathcal{S}_0$ only if the LD between them is low ($r^2 < 0.3$). Then, we residualized the effects of SNPs in $\mathcal{S}_0$ from estimated effects of the rest of SNPs in order to avoid double-counting their genetic effects.

The genetic variants which were not selected to $\mathcal{S}_0$ were projected into the eigenlocus space and then grouped into 10 × 10 double-partitions on intervals of eigenvalues λ_{lj} and absolute estimated effect sizes $\left|\widehat{\eta}_{lj}\right|$. This is because in the eigenlocus space, conditional mean effect $E[\eta_{lj}|\widehat{\eta}_{lj}]$ depends not only on the absolute value of estimated genetic effect $\left|\widehat{\eta}_{lj}\right|$ but also on eigenvalue of projection λ_{lj}. The eigenvalue of projection tracks the scale of true genetic effect in the eigenlocus space (Appendix C). In total, we used 101 partitions in this study including the partition of genome-wide significant SNPs $\mathcal{S}_0$.

While fully optimizing the partitioning cut-offs can potentially improve the accuracy of prediction model, this becomes rapidly impractical as the number of partitions increases. NPS requires a large enough number of partitions to closely approximate conditional mean effects, thus the combinatorial search for optimal cut-offs is computationally intractable. Therefore, we applied the following general heuristic, which worked well across our simulation datasets. First, the partitioning cut-offs were selected on the intervals of eigenvalues, equally distributing $\sum\limits_{l}\sum\limits_{j=1}^{m_l}\lambda_{lj}$ across partitions. This partition scheme evenly distributes the tagged heritability across partitions. The partitions on eigenvalues are denoted here by $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ from the lowest to the highest. Then, each partition of eigenvalues $\mathcal{S}_k$ was further partitioned on intervals of $\left|\widehat{\eta}_{lj}\right|$, equally distributing $\sum\limits_{l}\sum\limits_{j=1}^{m_l}\widehat{\eta}_{lj}^2$ across partitions. This second partitioning scheme is intended to evenly distribute the overall variance in polygenic scores, namely, $var\left(\sum\limits_{l}\sum\limits_{j=1}^{m_l}\widehat{\eta}_{lj}x_{lij}^P\right)$, across the partitions. This second partitions of $\mathcal{S}_k$ are denoted by $\mathcal{S}_{k,1}, \ldots, \mathcal{S}_{k,10}$ from the lowest to the highest $\left|\widehat{\eta}_{lj}\right|$.

Estimation of Conditional Mean Effect

The predicted genetic risk scores of individual $i \in \{1, \ldots, N\}$ can be represented by the sum of conditional mean effects $E[\eta_{lj}|\widehat{\eta}_{lj}]$ multiplied by genetic dosages x_{lij}^P across all genomic windows $l \in \{1, \ldots, L\}$ and genetic variations $j \in \{1, \ldots, m_l\}$ in each window. Instead of deriving conditional mean effects under a genetic architecture prior, we interpolate the conditional mean effects by fitting a linear function $f(\widehat{\eta}_{lj}) = \omega_k\widehat{\eta}_{lj}$ for each partition $k = 0, \ldots, K-1$ as follows:

$$\widehat{y}_i = \sum_{l=l}^{L}\sum_{j=1}^{m_l}E\left[\eta_{lj}\,|\,\widehat{\eta}_{lj}\right]x_{lij}^P \approx \sum_{l=l}^{L}\sum_{j=1}^{m_l}\left(\sum_{k=0}^{K-1}\omega_k\widehat{\eta}_{lj}I\left(\left(\lambda_{lj}, \left|\widehat{\eta}_{lj}\right|\right)\in\mathcal{S}_k\right)\right)x_{lij}^P \qquad \text{(Equation 2)}$$

where $I(\cdot)$ is an indicator function for the membership of genetic variations to partition k, $\mathcal{S}_k$ is the set of all genetic variations assigned to partition k, and K is the total number of partitions, set to 101 by default. The equation (Equation 2) can be further simplified by changing the order of summation as below:

$$\widehat{y}_i \approx \sum_{k=0}^{K-1}\omega_k\left(\sum_{l=l}^{L}\sum_{j=1}^{m_l}\widehat{\eta}_{lj}I\left(\left(\lambda_{lj}, \left|\widehat{\eta}_{lj}\right|\right)\in\mathcal{S}_k\right)x_{lij}^P\right)$$

$$= \sum_{k=0}^{K-1}\omega_k\left(\sum_{\left(\lambda_{lj}, \left|\widehat{\eta}_{lj}\right|\right)\in\mathcal{S}_k}\widehat{\eta}_{lj}x_{lij}^P\right) = \sum_{k=0}^{K-1}\omega_k G_{ik} \qquad \text{(Equation 3)}$$

where G_{ik} is a partitioned polygenic score of individual i calculated using only genetic variations belonging to the partition k. Then, ω_k becomes equivalent to the per-partition shrinkage weight. Based on Equation 3, we can estimate ω_k by fitting known phenotypes y_i with partitioned scores G_{ik} across individuals i in a small genotype-level training cohort.

For dichotomous phenotypes without covariates, we used a linear discriminant analysis (LDA) to estimate ω_k. The partitioned scores G_{ik} calculated in a training cohort form K-dimensional feature space, and LDA guarantees the optimal accuracy of the classifier when case and control subgroups follow multivariate normal distributions in the feature space. Since each partition consists of a sufficient number of projected genetic variations, partitioned scores of case and control subjects, namely $G_{ik}|\,y_i$, follow approximately normal distributions.[30] The variance of partitioned scores is approximately equal between case and control subjects since G_{ik} of an individual partition explains only a small fraction of phenotypic variation on the observed scale in typical GWAS data.[31] Furthermore, due to the decorrelating property of eigenlocus projection, the covariance of G_{ik} and $G_{ik'}$ can be assumed to be approximately 0 between different partitions k and k'. Although in theory, the liability thresholding effect induces slight non-zero covariance between partitions, this effect is typical small and negligible. Thus, LDA-derived shrinkage weights can be independently estimated for each partition and simplify to:

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

Similarly, for continuous phenotypes or the case of dichotomous traits with covariates, we can estimate per-partition shrinkage weights ω_k by applying the following linear regression model to the training data:

$$y_i = \omega_k G_{ik} + covariates$$

independently for each partition.

In the special case of infinitesimal genetic architecture, in which all SNPs are causal with normally distributed effect sizes, the conditional mean effects have been analytically derived and are predicted to depend only on eigenvalues λ_{lj};[18] therefore, we can cross-check the accuracy of our shrinkage weights ω_k estimated by NPS in simulations (Appendix D). To apply NPS, we first partitioned genetic variations in the eigenlocus space into ten subgroups on intervals of their eigenvalues λ_{lj} as described above but without separating out the genome-wide significant SNPs (Figure 2A). The per-partition shrinkage weights ω_k trained by NPS closely tracked the theoretical optimum in most of the bins. Interestingly, in the lowest and highest partitions of eigenvalues,

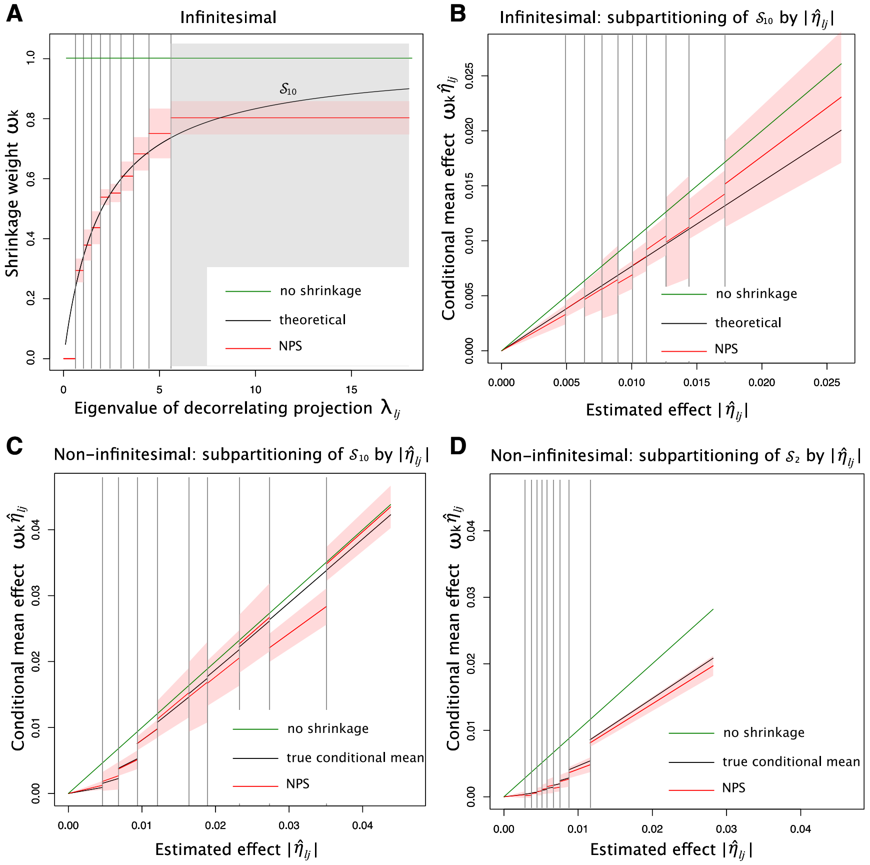

Figure 2. Per-Partition Shrinkage Weights Estimated by Non-Parametric Shrinkage (NPS) Approximate the Conditional Mean Effects in the Decorrelated Space

(A) NPS shrinkage weights ω_k (red line) compared to the theoretical optimum (black line), $\lambda_{lj}/\left(\lambda_{lj}+\frac{M}{N_g h^2}\right)$, under infinitesimal architecture. The partition of largest eigenvalues $\mathcal{S}_{10}$ is marked by gray box.

(B) Conditional mean effects estimated by NPS (red line) in sub-partitions of $\mathcal{S}_{10}$ by $|\hat{\eta}_{lj}|$ under infinitesimal architecture. The theoretical line (black) is the average over all λ_{lj} in $\mathcal{S}_{10}$.

(C and D) Conditional mean effects estimated by NPS (red line) in sub-partitions of $\mathcal{S}_{10}$ (C) and $\mathcal{S}_2$ (D) on intervals of $|\hat{\eta}_{lj}|$ under non-infinitesimal architecture with the causal SNP fraction of 1%. The true conditional means (black) were estimated over 40 simulation runs.

The mean NPS shrinkage weights (red line) and their 95% CIs (red shade) were estimated from five replicates. Grey vertical lines indicate partitioning cut-offs. No shrinkage line (green) indicates $\omega_k = 1$. The number of markers M is 101,296. The discovery GWAS size N_g equals to M. The heritability h^2 is 0.5.

$\mathcal{S}_1$ and $\mathcal{S}_{10}$, the estimated shrinkage was significantly biased away from the optimal curve. The smallest eigenvalues are too noisy to estimate with the reference LD panel. Therefore, it is correct to down-weight ω_1 almost to 0. In case of partition $\mathcal{S}_{10}$, it spans the widest interval of eigenvalues but consists of the fewest number of SNPs. While it is ideal to apply a finer partitioning in this interval so as to better interpolate the theoretical curve, the total numbers of SNPs and independent projection vectors in the genome are the fundamental limiting factor.

In the case of infinitesimal architecture, theory predicts that per-partition shrinkage weights are independent of estimated effect sizes $\hat{\eta}_{lj}$. To examine the robustness of NPS, we applied the general 10-by-10 double partitioning on λ_{lj} and $|\hat{\eta}_{lj}|$ collected under infinitesimal simulations. In overall, the shrinkage weights estimated by double partitioning agree with the theoretical expectation. The estimated conditional mean effects, interpolated with $\omega_k \hat{\eta}_{lj}$, follow the linear trajectory (Figures 2B and S1).

For non-infinitesimal genetic architecture, we do not have an analytic derivation of conditional mean effects; therefore, we empirically estimated the conditional means using the true underlying effects η_{lj} and true LD structure of the population. Here, 1% of SNPs were simulated to be causal with normally distributed effect sizes. As expected, the true conditional mean dips for the lowest values of $|\hat{\eta}_{lj}|$ but approaches no shrinkage ($\omega_k = 1$) with increasing values of $|\hat{\eta}_{lj}|$ (Figures 2C and 2D). A notable difference between the partitions of largest eigenvalues and second smallest eigenvalues is that the true conditional mean is very close to no shrinkage for large $|\hat{\eta}_{lj}|$ in the former. This is because eigenvalues are proportional to the scale of true effects η_{lj}; therefore, with large enough eigenvalues, the sampling error becomes relatively small and the estimated effect sizes more accurate. In all partitions, conditional mean effects estimated by NPS stayed very close to the true conditional means (Figure S2).

Back-Conversion from the Eigenlocus Space to Per-SNP Effects

Rewriting Equation 2 using matrix operations, we can reformulate the N-dimensional vector of predicted genetic risk scores $\hat{\boldsymbol{y}}$ using the original SNP genotypes $\mathbf{X}_l$ instead of eigenlocus genotypes $\mathbf{X}_l^P$ as follows:

$$\hat{\boldsymbol{y}} = \sum_{l=1}^{L} \mathbf{X}_l^P E[\boldsymbol{\eta}_l \mid \hat{\boldsymbol{\eta}}_l] = \sum_{l=1}^{L} \mathbf{X}_l \left(\boldsymbol{\Lambda}_l^{-\frac{1}{2}} \mathbf{Q}_l^T\right)^T E[\boldsymbol{\eta}_l \mid \hat{\boldsymbol{\eta}}_l]$$

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

from the definition of $\mathbf{X}_l^P$ (Equation 1). We obtain the conditional mean effects by non-parametric shrinkage in the following form:

$$E[\boldsymbol{\eta}_l \mid \hat{\boldsymbol{\eta}}_l] \approx \mathbf{W}_l \, \hat{\boldsymbol{\eta}}_l$$

where $\mathbf{W}_l$ is an $m_l \times m_l$ diagonal matrix with diagonal entries $\{w_{jj}\}$ defined as:

$w_{jj} = \omega_k$ *with* with the k such that $(\lambda_{lj}, |\hat{\eta}_{lj}|) \in \mathcal{S}_k$

where k is the partition to which the j^{th} projected genetic variation belong in the eigenlocus space. Therefore, the reweighted effects in the original per-SNP scale can be retrieved back by computing $\mathbf{Q}_l \boldsymbol{\Lambda}_l^{-\frac{1}{2}} \mathbf{W}_l \, \hat{\boldsymbol{\eta}}_l$.

Application of NPS to Genome-wide Datasets

The estimated effect size at each SNP is available as summary statistics from a large discovery GWAS. As these estimated effects were represented as per-allele effects, we converted them relative to standardized genotypes by multiplying by $\sqrt{2f(1-f)}$, where f is the allele frequency of each SNP in the discovery GWAS cohort.

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. When NPS was run in parallel on up to 88 processors (22 chromosomes × 4 window shifts), it took total computation time of 3 to 6 h for each dataset.

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 non-Finnish 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.[32]

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 $N(\mu = 0,\ \Sigma = \mathbf{D}_1)$. 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:

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

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

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

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 and lowest $n(1-f)^2$ genotypes as homozygotes and the rest as heterozygotes under Hardy-Weinberg equilibrium. n is the number of simulated samples and f is the allele frequency of each SNP. This MVN-based simulator can efficiently generate a very large cohort with realistic LD structure across the genome and is guaranteed to produce homogeneous 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 for SNP 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 $(f_j(1-f_j))^\alpha$ with $\alpha = -0.25$[33] as follows:

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

Finally, for the DHS model, we further extended the MAF-dependent point-normal architecture to exhibit clumping of causal SNPs within DHS peaks. Fifteen percent 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 the genome in order to simulate the enrichment of per-SNP heritability in DHS reported in the previous study.[34] Specifically, β_j was sampled from the following distribution:

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

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 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 approximately equivalent by a common scaling factor. 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 GWASs of same sample sizes.[35] For the training dataset, we assembled a cohort of 2,500 case subjects and 2,500 control subjects by down-sampling control subjects 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,[36,37] inflammatory bowel disease (IBD),[38] type 2 diabetes (T2D),[39] and coronary artery disease (CAD).[40] These GWAS summary statistics were based only on white (European) 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. Control subjects 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 case and control subjects. For IBD, we identified case individuals by ICD10 or self-reported disease codes of Crohn disease, ulcerative colitis, or IBD. Control subjects were randomly selected excluding participants with history of any auto-immune disorders. For T2D, case subjects were identified by ICD10 diagnosis codes or by questionnaire on history of diabetes combined with the age of diagnosis over 30. However, our T2D case subjects may include a small fraction of type 1 diabetic case subjects misdiagnosed as T2D (3.7%) as previously reported.[41] 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 all phenotypes, our case definition includes both prevalent and incident cases.

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. For all phenotypes, we filtered out participants who were retracted, were not from white British ancestry, or had indication of any QC issue in UK Biobank. We included only samples that were genotyped with Axiom array. Related samples were excluded to avoid potential confounding. The samples were randomly split to training and validation cohorts. Controls were down-sampled to the case to control ratio of 1:1 to assemble training cohorts, but no down-sampling was applied to validation cohorts to keep the original case prevalence.

Partners Biobank

We used Partners Biobank[42] 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 v.2.4 and IMPUTE v.4 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.[43] To obtain early-onset CAD, both case and control subjects were restricted to men with age $\leq$ 55 and women with age $\leq$ 65. Since the prevalence of early-onset CAD and T2D are sex dependent, we included the sex covariate in the genetic risk model for CAD and T2D. For all methods, the coefficient of sex covariate was estimated in the training cohort of UK Biobank.

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 at the training step. 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 that 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; thus, we validated LDPred only using overlapping markers in imputed data of two cohorts.

LD Pruning and Thresholding

LD Pruning and Thresholding (P+T) algorithm was evaluated using PRSice software in the default setting.[44] 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.

PRS-CS

PRS-CS algorithm was benchmarked using the default parameter setting.[21] The optimal ϕ parameter values were optimized in training cohorts, and the highest performing model was evaluated in validation cohorts. For the reference LD panel, we used a set of simulated genotypes produced by our MVN simulator in order to accurately capture the underlying LD structure of our simulated datasets; in real data, we used the "EUR" reference LD panel provided in the software. Imputed allelic dosages were converted to hard-called genotypes similarly as recommended by the authors.

Results

Application to Simulated Data

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 (Material and 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.[33,34,45] The prediction accuracy of NPS remained robust across the simulated genetic architectures (Tables 1 and S1). We measured prediction accuracy using Nagelkerke R^2 and odds ratio at the highest 5% tail of the polygenic score distribution. The latter measure has been

Table 1. Comparison of Prediction Accuracy in Simulated Genetic Architecture

				5%	**NPS R^2_{Nag} Compared to**		
% Causal SNPs	**Method**	**$R^2_{Nagelkerke}$**	**% h^2 Explained**	**Tail OR**	**P+T**	**LDPred**	**PRS-CS**
1%	P+T	0.050	14.8	3.18			
	LDPred	0.068	20.6	3.66			
	PRS-CS	0.075	22.0	4.02			
	NPS	0.085	24.6	4.27	1.68*	1.25*	1.13*
0.1%	P+T	0.136	40.8	6.32			
	LDPred	0.080	23.0	4.08			
	PRS-CS	0.156	44.8	7.03			
	NPS	0.179	51.2	8.09	1.31*	2.22*	1.14*
0.01%	P+T	0.213	61.4	9.92			
	LDPred	0.153 (0.268)[a]	43.8 (74.6)[a]	7.66 (13.37)[a]			
	PRS-CS	0.228	65.3	10.35			
	NPS	0.328	92.6	17.19	1.54*	2.14*	1.44*

Non-parametric shrinkage (NPS) is more robust and accurate compared to other methods in simulated datasets. 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 case subjects and 2,500 control subjects. 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 asterisk (*) indicates a significant improvement in Nagelkerke's R^2 (paired t test; $p < 0.05$).

[a] The accuracy of LDPred varies widely depending on the convergence of prediction model; thus, we report the maximum R^2 in parentheses as well as the average performance.

popularized by a recent study that reported that the tails of the polygenic score distribution are associated with risk that is similar to monogenic mutations.[23]

We evaluated the performance of NPS vis-à-vis two popular methods, LDPred and P+T, as well as the newest method PRS-CS with the superior reported accuracy[13,18,21] (Tables S1–S5). LDPred is the state-of-the-art Bayesian parametric method, which is similarly based on summary statistics estimated in large GWAS datasets and an independent training set with individual-level data. PRS-CS is a new sophisticated extension of the Bayesian strategy. We found that our method resulted in more accurate predictions than all three methods across a range of genome-wide simulations. PRS-CS was shown to be more accurate than P+T and LDPred on simulated data, although less accurate than NPS. The improvement over LDPred 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 (Table S3). The simulations suggest that the well-optimized parametric models are capable of generating good predictions, but NPS is much more robust and does not suffer from optimization issues. Overall, NPS improves accuracy consistently for all simulated allelic architectures for both Negelkerke R^2 and odds ratios at 5% tail (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 (Material and Methods).[36–40,46] For all three phenotypes except coronary artery disease, NPS showed significantly higher accuracy than LDPred or P+T (Tables 2 and S6–S9 and Figures S3–S7) and highly similar (statistically indistinguishable) accuracy compared to PRS-CS. In particular, our method and PRS-CS outperformed the other two methods by greater magnitudes with more recent GWAS summary statistics with finer resolution. For example, the latest breast cancer GWAS 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 1.5-fold with the latest breast cancer data whereas the accuracy of LDPred did not improve at all. The R^2 of P+T increased by 1.25-fold, but the gain is mainly due to the inferior accuracy with older GWAS data.

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 (Material and Methods).[42] For all phenotypes, NPS outperformed both P+T and LDPred and showed similar accuracy as PRS-CS (Tables 3 and S10–S13). NPS also has a higher odds ratio at 5%

Table 2. Accuracy of Polygenic Prediction in Real Data

Discovery GWAS	Training (UK Biobank)	Validation (UK Biobank)	Method	R^2_{Nag}	5% Tail OR
Breast cancer 2015 (n = ∼120,000)	n = 3,956/3,956	n = 3,957/73,652	P+T	0.021	2.28
			LDPred	0.026	2.42
			PRS-CS	0.030	2.60
			NPS	0.030	2.53
Breast cancer 2017 (n = ∼230,000)			P+T	0.027	2.37
			LDPred	0.026	2.33
			PRS-CS	0.043	2.96
			NPS	0.045	3.01
Inflammatory bowel disease (n = ∼35,000)	n = 2,483/2,483	n = 2,482/157,272	P+T	0.028	3.00
			LDPred	0.027	2.77
			PRS-CS	0.040	3.67
			NPS	0.035	3.60
Type 2 diabetes (n = ∼160,000)	n = 7,298/7,298	n = 7,298/144,020	P+T	0.046	3.04
			LDPred	0.059	3.51
			PRS-CS	0.066	3.99
			NPS	0.065	3.81
Coronary artery disease (n = ∼330,000)	n = 2,000/2,000	n = 773/62,512	P+T	0.063	5.17
			LDPred	0.078	5.65
			PRS-CS	0.075	4.92
			NPS	0.073	5.21

Non-parametric shrinkage (NPS) and PRS-CS outperform 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 case subjects over control subjects at the 5% tail in polygenic score distribution compared to the rest. For CAD and T2D, all prediction models were trained and validated with the sex covariate to account for the difference of disease prevalence by sex.

distribution tail than PRS-CS consistently for all phenotypes, although this improvement is not statistically significant (Table 3).

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 polygenic risk prediction.[23,47] 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.[18,48] 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,[49] which is a problem that should be addressed by future work in this field.

Human phenotypes vary in the degree of polygenicity,[50] in the fraction of heritability attributable to low-frequency variants[33] and in other aspects of allelic architecture.[45,51] 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.[33,34,45,50,51] 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.[52,53] Recently, there has been a

Table 3. Accuracy of Polygenic Prediction in Independent Validation Cohorts

Discovery GWAS	Training (UK Biobank)	Validation (Partners)	Method	R^2_{Nag}	5% Tail OR
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
			PRS-CS	0.034	2.23
			NPS	0.034	2.32
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
			PRS-CS	0.065	4.11
			NPS	0.069	4.32
Type 2 diabetes (n = ~160,000)	n = 7,298/7,298	n = 2,026/14,813	P+T	0.038	2.10
			LDPred	0.046	2.51
			PRS-CS	0.058	2.80
			NPS	0.054	2.97
Coronary artery disease (n = ~330,000)	n = 2,000/2,000	n = 268/7,107	P+T	0.018	2.72
			LDPred	0.016	2.31
			PRS-CS	0.027	3.16
			NPS	0.025	4.10

Non-parametric shrinkage (NPS) and PRS-CS outperform 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. For CAD and T2D, all prediction models were trained and validated with the sex covariate to account for the difference of disease prevalence by sex.

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 structure in the data.[24–27] To address these issues, we developed NPS, a non-parametric method that 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. NPS is flexible to incorporate additional complexity of true genetic architecture. Our non-parametric approach has been recently adopted by LDPred-funct, an extension of LDPred to incorporate functional annotations.[54] 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 outperform traditional parametric approaches for identifying individuals at increased risk.

Appendix A. Distribution of Projected Genotypes in the Eigenlocus Space

Let X_i be an m-dimensional genotype vector of all SNPs in genomic window l and individual i. We drop the subscript for genomic window for the sake of simplicity when it is clear from the context. The standardized genotype X_i is approximated by the following multivariate normal distribution:

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

where $\mathbf{D}$ is a LD matrix of the window. Since the projected genotype X_i^P is derived by applying eigenlocus projection $\mathcal{P}$ on X_i by definition (Equation 1), X_i^P also follows a multivariate normal distribution. Specifically, the distribution of X_i^P is:

$$X_i^P \sim N\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\,\mathbf{Q}^T 0,\ \left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\,\mathbf{Q}^T\right)\mathbf{D}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\,\mathbf{Q}^T\right)^T\right)$$

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

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

Appendix B. Distribution of Effect Size Estimates in the Eigenlocus Space

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

$$\widehat{\beta} = \frac{1}{N_g}\mathbf{X}^T\boldsymbol{y}$$

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

$$\widehat{\boldsymbol{\beta}} = \frac{1}{N_g}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \varepsilon) = \mathbf{D}\boldsymbol{\beta} + \frac{1}{N_g}\mathbf{X}^T\varepsilon$$

where the residual ε follows an N_g-dimensional multivariate normal distribution $N(0,\ \sigma_e^2\mathbf{I})$. In an individual window, the genetic effects explain only a small fraction of phenotypic variation, so we can assume that $\sigma_e^2 \approx \mathrm{var}(\mathbf{y}) = 1$. The distribution of sampling noise in $\widehat{\boldsymbol{\beta}}$, namely the distribution of $\widehat{\boldsymbol{\beta}}$ given $\boldsymbol{\beta}$, follows:

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

$$\approx N\left(\mathbf{D}\boldsymbol{\beta},\ \frac{1}{N_g}\mathbf{D}\right)$$

since $\mathbf{D} = (1/N_g)\mathbf{X}^T\mathbf{X}$. Since the estimated effect size $\widehat{\boldsymbol{\eta}}$ in the eigenlocus space is obtained by applying $\mathcal{P}$ on $\widehat{\boldsymbol{\beta}}$ by definition (Equation 1), the distribution of $\widehat{\boldsymbol{\eta}}$ given $\boldsymbol{\beta}$ also follows a multivariate normal distribution:

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

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

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

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

Appendix C. Interpretation of Eigenvalues

Let $\boldsymbol{\beta}$ be the m-dimensional vector of true genetic effect at m SNPs in a genomic window. We assume that $\boldsymbol{\beta}$ is symmetric at 0 and independent at each SNP. Then, the distribution of true genetic effects $\boldsymbol{\eta} = \{\eta_j\}$ in the eigenlocus space will follow:

$$E[\eta_j] = E\left[\sqrt{\lambda_j}\ \boldsymbol{q}_j^T\boldsymbol{\beta}\right] = \sqrt{\lambda_j}\ \boldsymbol{q}_j^T E[\boldsymbol{\beta}] = 0$$

where λ_j and $\boldsymbol{q}_j$ are the eigenvalue and eigenvector, respectively, projecting $\boldsymbol{\beta}$ to η_j by Equation 1. If we put that eigenvector $\boldsymbol{q}_j$ is $(q_{1j}\ldots q_{mj})^T$ and $\boldsymbol{\beta}$ is $(\beta_1\ldots\beta_m)^T$, the variance of true genetic effects for an eigenlocus is:

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

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

Therefore, in general, $var[\eta_j]$, is directly proportional to eigenvalue λ_j. 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 $\sum_{s=1}^{m} q_{sj}^2 = 1$.

Appendix D. 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.:[18]

$$E\left[\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}\right] = \left(\frac{M}{N_g h^2}\mathbf{I} + \mathbf{D}\right)^{-1}\widehat{\boldsymbol{\beta}} \qquad \text{(Equation S1)}$$

where N_g is the sample size of GWAS cohort, h^2 is the heritability of trait, M is the total number of SNPs, and $\mathbf{D}$ is the LD matrix of full rank. Then, $\mathbf{D}$ can be factorized into $\mathbf{D} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$ with eigenvalues $\boldsymbol{\Lambda}$ and eigenvectors $\mathbf{Q}$. Since

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

and

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

we can reformulate Equation S1 as follows:

$$E\left[\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}\right] = \mathbf{Q}\left(\frac{M}{N_g h^2}\mathbf{I} + \boldsymbol{\Lambda}\right)^{-1}\mathbf{Q}^T\widehat{\boldsymbol{\beta}}$$

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

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

by the definition of $\widehat{\boldsymbol{\eta}}$ (Equation 1). Hence,

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

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

$$= \left(\frac{M}{N_g h^2}\mathbf{I} + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{\Lambda}\, \widehat{\boldsymbol{\eta}}$$

by the definition of $\boldsymbol{\eta}$. Therefore, for the j^{th} eigenlocus projection defined by eigenvalue λ_j and eigenvector $\boldsymbol{q_j}$, the conditional mean effect is given as the following:

$$E[\eta_j \mid \widehat{\eta}_j] = \frac{\lambda_j}{\lambda_j + \frac{M}{N_g h^2}} \widehat{\eta}_j$$

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

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

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.05.004.

Acknowledgments

N.O.S. was supported in part by NIH grants K08HL114642, R01HL131961, and UM1HG008853 and by The Foundation for Barnes-Jewish Hospital. S.K. was supported by a Research Scholar award from the Massachusetts General Hospital, the Donovan Family Foundation, NIH R01HL107816, a grant from Fondation Leducq, and an investigator-initiated grant from Merck. S.R.S. was supported by NIH R35GM127131, R01MH101244, and U01HG006500. S.C., M.I., and S.R.S. were supported by a grant from the Altius Institute for Biomedical Sciences. This research has been conducted using the UK Biobank Resource under Application Number 31063.

Declaration of Interests

S.K. is a co-founder, chief executive officer, and a board member of Verve Therapeutics.

Received: June 15, 2019
Accepted: May 1, 2020
Published: May 28, 2020

Web Resources

NPS software, https://github.com/sgchun/nps/

References

1. Grundy, S.M., Stone, N.J., Bailey, A.L., Beam, C., Birtcher, K.K., Blumenthal, R.S., Braun, L.T., de Ferranti, S., Faiella-Tommasino, J., Forman, D.E., et al. (2018). 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. Circulation *139*, e1082–e1143.
2. Goddard, M.E., and Hayes, B.J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. *10*, 381–391.
3. Falke, K.C., Glander, S., He, F., Hu, J., de Meaux, J., and Schmitz, G. (2013). The spectrum of mutations controlling complex traits and the genetics of fitness in plants. Curr. Opin. Genet. Dev. *23*, 665–671.
4. Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819–1829.
5. Ripatti, S., Tikkanen, E., Orho-Melander, M., Havulinna, A.S., Silander, K., Sharma, A., Guiducci, C., Perola, M., Jula, A., Sinisalo, J., et al. (2010). A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet *376*, 1393–1400.
6. Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H.S., Diver, W.R., Thun, M.J., Cox, D.G., Hankinson, S.E., Kraft, P., et al. (2010). Performance of common genetic variants in breast-cancer risk models. N. Engl. J. Med. *362*, 986–993.
7. Wray, N.R., Goddard, M.E., and Visscher, P.M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. *17*, 1520–1528.
8. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
9. Golan, D., and Rosset, S. (2014). Effective genetic-risk prediction using mixed models. Am. J. Hum. Genet. *95*, 383–393.
10. Speed, D., and Balding, D.J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. *24*, 1550–1557.
11. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.
12. Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S.J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. *45*, 400–405, e1–e3.
13. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P.; and International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
14. Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A., et al.; Diabetes Genetics Replication and Meta-analysis Consortium; and Myocardial Infarction Genetics Consortium (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. *44*, 483–489.
15. Abraham, G., Tye-Din, J.A., Bhalala, O.G., Kowalczyk, A., Zobel, J., and Inouye, M. (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. PLoS Genet. *10*, e1004137.
16. Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet. *11*, e1004969.

17. Shi, J., Park, J.H., Duan, J., Berndt, S.T., Moy, W., Yu, K., Song, L., Wheeler, W., Hua, X., Silverman, D., et al.; MGS (Molecular Genetics of Schizophrenia) GWAS Consortium; GECCO (The Genetics and Epidemiology of Colorectal Cancer Consortium); GAME-ON/TRICL (Transdisciplinary Research in Cancer of the Lung) GWAS Consortium; PRACTICAL (PRostate cancer AssoCiation group To Investigate Cancer Associated aLterations) Consortium; PanScan Consortium; and GAME-ON/ELLIPSE Consortium (2016). 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.
18. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am. J. Hum. Genet. *97*, 576–592.
19. Zhu, X., and Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Ann. Appl. Stat. *11*, 1561–1592.
20. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 5086.
21. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776.
22. Goddard, M.E., Wray, N.R., Verbyla, K., and Visscher, P.M. (2009). Estimating Effects and Making Predictions from Genome-Wide Marker Data. Stat. Sci. *24*, 517–529.
23. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.
24. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat. Commun. *8*, 456.
25. Efron, B. (2009). Empirical bayes estimates for large-scale prediction problems. J. Am. Stat. Assoc. *104*, 1015–1028.
26. So, H.C., and Sham, P.C. (2017). Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Sci. Rep. *7*, 41262.
27. Gianola, D., Fernando, R.L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics *173*, 1761–1776.
28. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol. *41*, 469–480.
29. Inouye, M., Abraham, G., Nelson, C.P., Wood, A.M., Sweeting, M.J., Dudbridge, F., Lai, F.Y., Kaptoge, S., Brozynska, M., Wang, T., et al.; UK Biobank CardioMetabolic Consortium CHD Working Group (2018). Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J. Am. Coll. Cardiol. *72*, 1883–1893.
30. Wray, N.R., Yang, J., Goddard, M.E., and Visscher, P.M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. *6*, e1000864.
31. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.
32. Cai, T.T., Zhang, C.H., and Zhou, H.H. (2010). Optimal rates of convergence for covariance matrix estimation. Ann. Stat. *38*, 2118–2144.
33. Speed, D., Cai, N., Johnson, M.R., Nejentsev, S., Balding, D.J.; and UCLEB Consortium (2017). Reevaluation of SNP heritability in complex human traits. Nat. Genet. *49*, 986–992.
34. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and SWE-SCZ Consortium (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
35. Yang, J., Wray, N.R., and Visscher, P.M. (2010). Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. Genet. Epidemiol. *34*, 254–257.
36. Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M.J., Maranian, M.J., Bolla, M.K., Wang, Q., Shah, M., et al.; BOCS; kConFab Investigators; AOCS Group; NBCS; and GENICA Network (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat. Genet. *47*, 373–380.
37. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al.; NBCS Collaborators; ABCTB Investigators; and ConFab/AOCS Investigators (2017). Association analysis identifies 65 new breast cancer risk loci. Nature *551*, 92–94.
38. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.; International Multiple Sclerosis Genetics Consortium; and International IBD Genetics Consortium (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.
39. Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D., et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2017). An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes *66*, 2888–2902.
40. Nelson, C.P., Goel, A., Butterworth, A.S., Kanoni, S., Webb, T.R., Marouli, E., Zeng, L., Ntalla, I., Lai, F.Y., Hopewell, J.C., et al.; EPIC-CVD Consortium; CARDIoGRAMplusC4D; and UK Biobank CardioMetabolic Consortium CHD working group (2017). Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat. Genet. *49*, 1385–1391.
41. Thomas, N.J., Jones, S.E., Weedon, M.N., Shields, B.M., Oram, R.A., and Hattersley, A.T. (2018). Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from UK Biobank. Lancet Diabetes Endocrinol. *6*, 122–129.
42. Karlson, E.W., Boutin, N.T., Hoffnagle, A.G., and Allen, N.L. (2016). 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.
43. Gainer, V.S., Cagan, A., Castro, V.M., Duey, S., Ghosh, B., Goodson, A.P., Goryachev, S., Metta, R., Wang, T.D., Wattanasin, N., and Murphy, S.N. (2016). 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*, 6.

44. Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: Polygenic Risk Score software. Bioinformatics *31*, 1466–1468.
45. Zeng, J., de Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. *50*, 746–753.
46. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
47. Riglin, L., Collishaw, S., Richards, A., Thapar, A.K., Maughan, B., O'Donovan, M.C., and Thapar, A. (2017). Schizophrenia risk alleles and neurodevelopmental outcomes in childhood: a population-based cohort study. Lancet Psychiatry *4*, 57–62.
48. Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. *14*, 507–515.
49. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.
50. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186.
51. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., and Price, A.L. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat. Genet. *49*, 1421–1427.
52. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput. Biol. *13*, e1005589.
53. Hu, Y., Lu, Q., Liu, W., Zhang, Y., Li, M., and Zhao, H. (2017). Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS Genet. *13*, e1006836.
54. Marquez-Luna, C., Gazal, S., Loh, P.-R., Furlotte, N., Auton, A., Price, A.L., Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S.S., et al. (2019). Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. bioRxiv. https://doi.org/10.1101/375337.

The American Journal of Human Genetics, Volume *107*

Supplemental Data

Non-parametric Polygenic Risk Prediction

via Partitioned GWAS Summary Statistics

Sung Chun, Maxim Imakaev, Daniel Hui, Nikolaos A. Patsopoulos, Benjamin M. Neale, Sekar Kathiresan, Nathan O. Stitziel, and Shamil R. Sunyaev

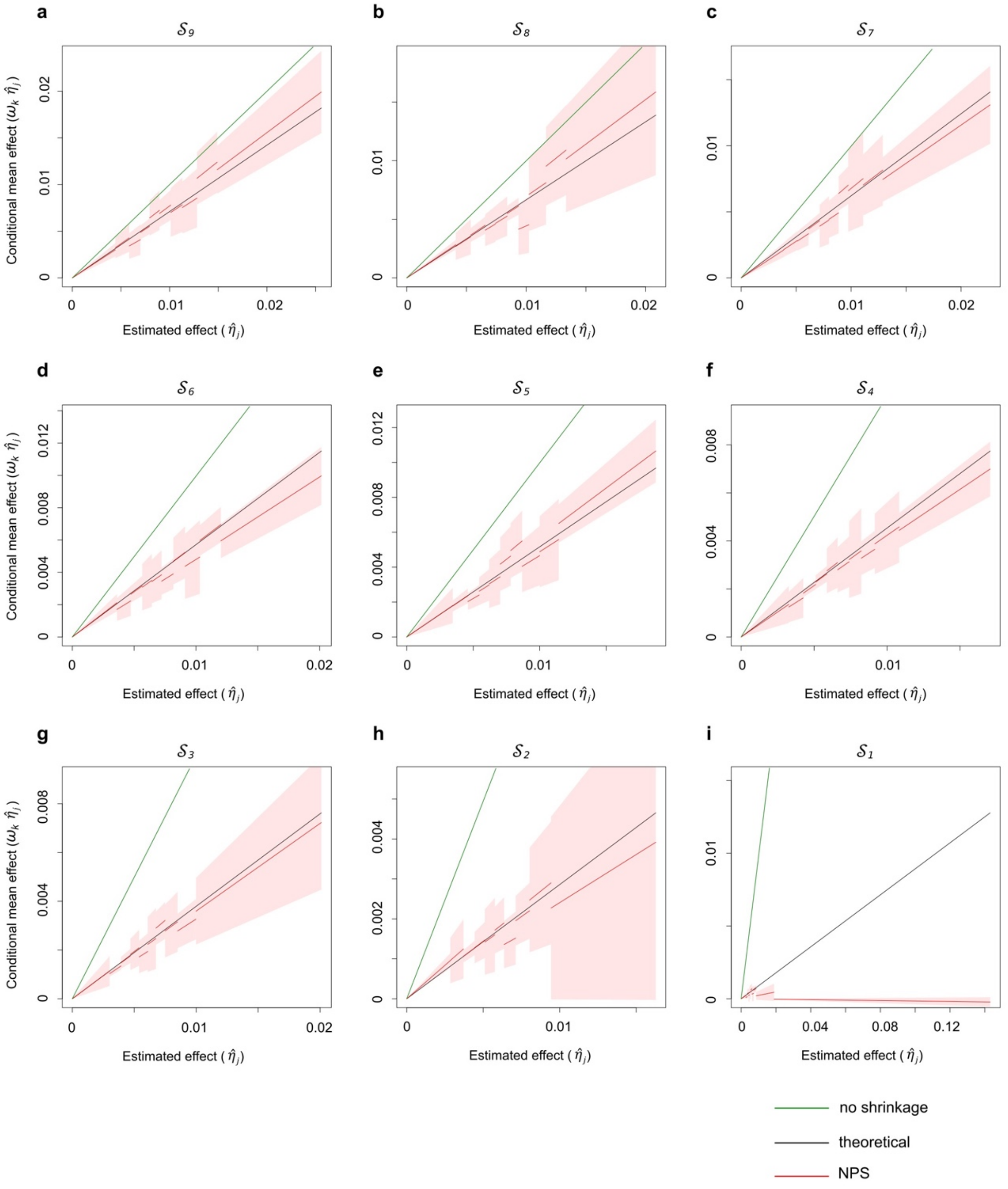

Figure S1. NPS approximates the conditional mean effects: infinitesimal genetic architecture ($\mathcal{S}_1, \dots, \mathcal{S}_9$). NPS shrinkage weights ω_k (red line) are compared to the theoretical optimum (black line), $\lambda_j/(\lambda_j + \frac{M}{N_g h^2})$, under the infinitesimal architecture. $\mathcal{S}_1, \dots, \mathcal{S}_{10}$ indicate the partitions of lowest to highest eigenvalues of projection. 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. See Figure 2B for $\mathcal{S}_{10}$.

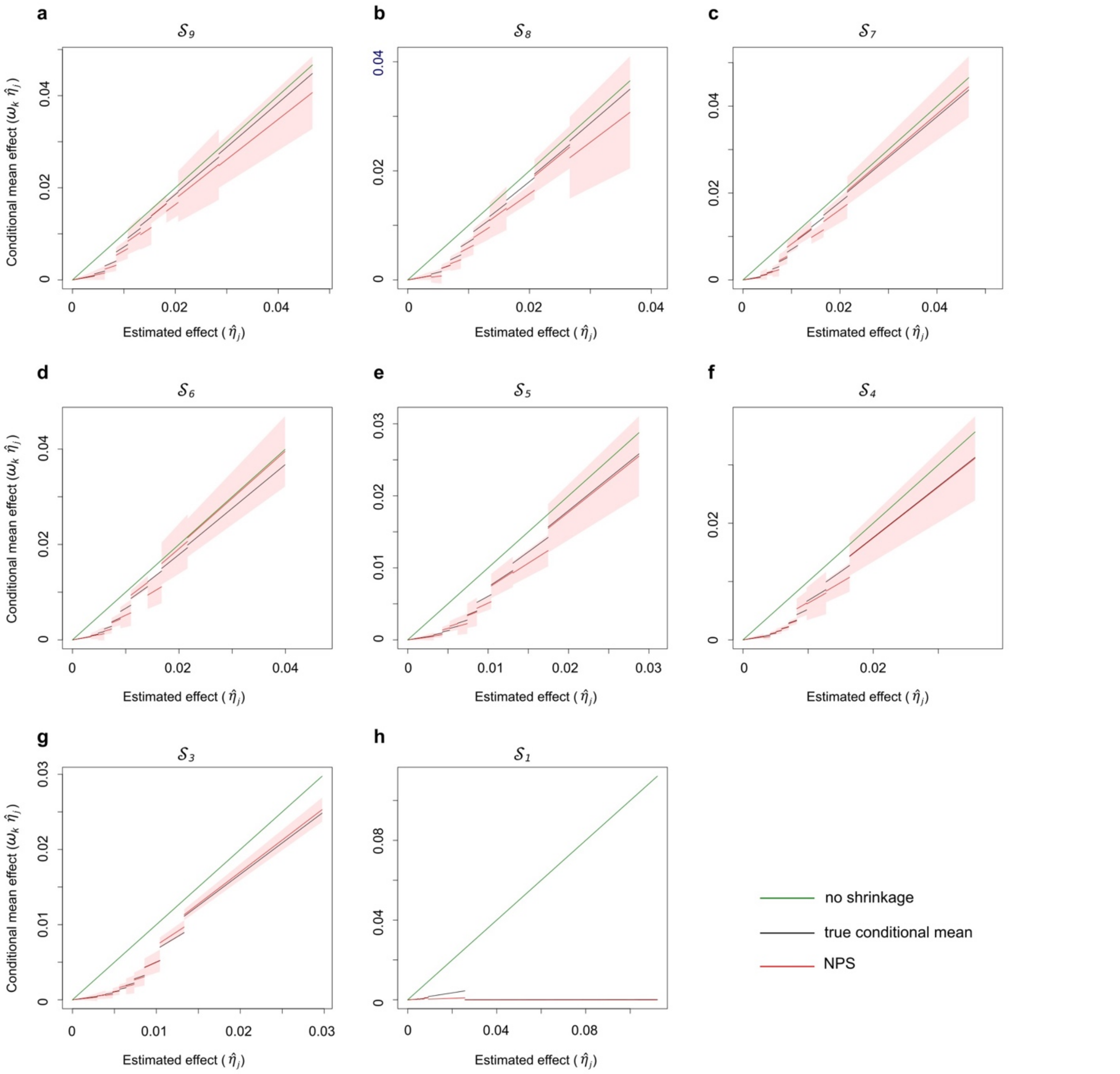

Figure S2. NPS approximates the conditional mean effects: non-infinitesimal genetic architecture ($\mathcal{S}_1, \mathcal{S}_3, \dots, \mathcal{S}_9$). NPS shrinkage weights ω_k (red line) are compared to the true conditional means (black line), which were estimated empirically from 40 simulation runs. $\mathcal{S}_1, \dots, \mathcal{S}_{10}$ indicate the partitions of lowest to highest eigenvalues of projection. 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%. See Figure 2C-D for $\mathcal{S}_2$ and $\mathcal{S}_{10}$, respectively.

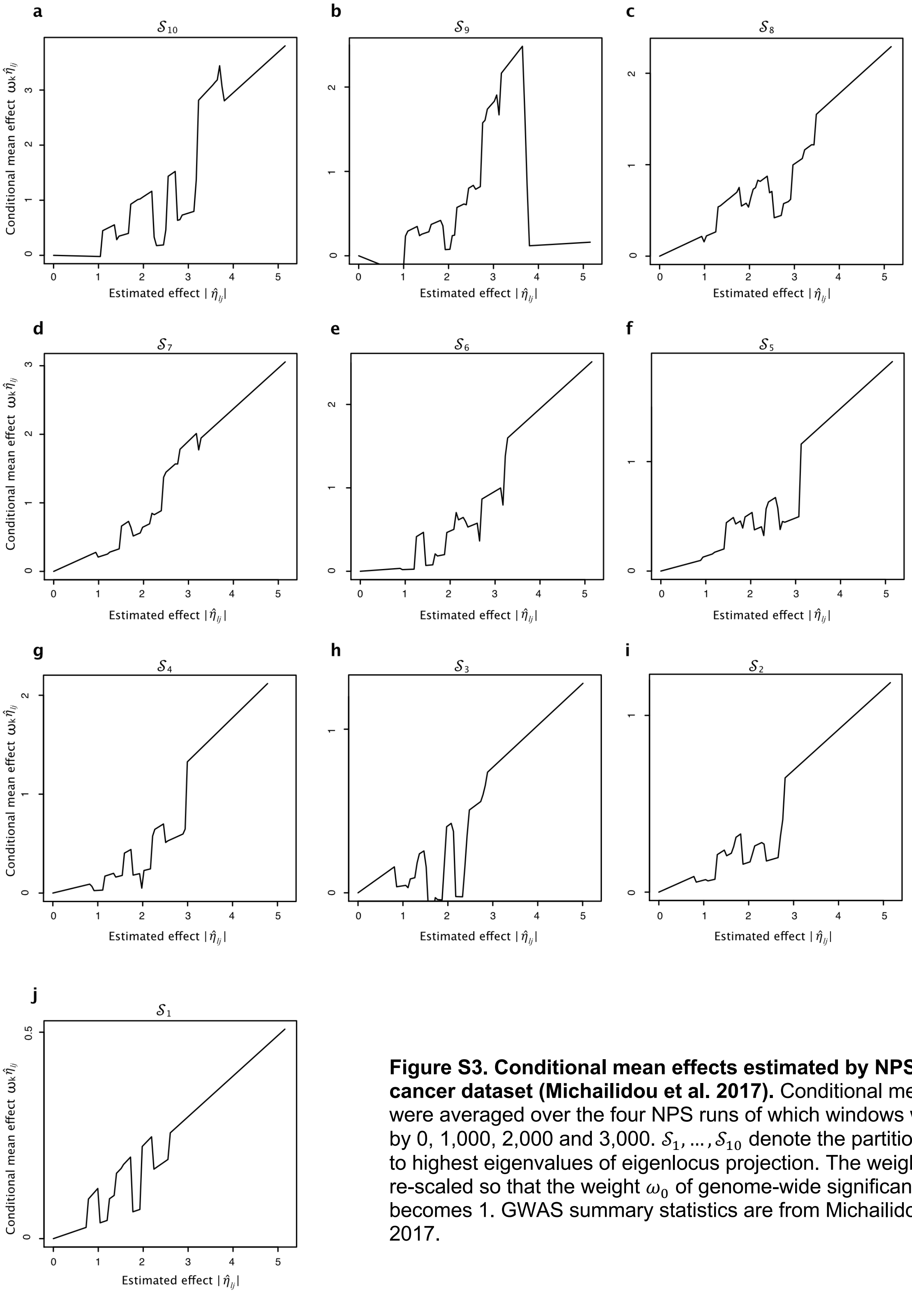

Figure S3. Conditional mean effects estimated by NPS in breast cancer dataset (Michailidou et al. 2017). Conditional mean effects were averaged over the four NPS runs of which windows were shifted by 0, 1,000, 2,000 and 3,000. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denote the partitions of lowest 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. GWAS summary statistics are from Michailidou et al. 2017.

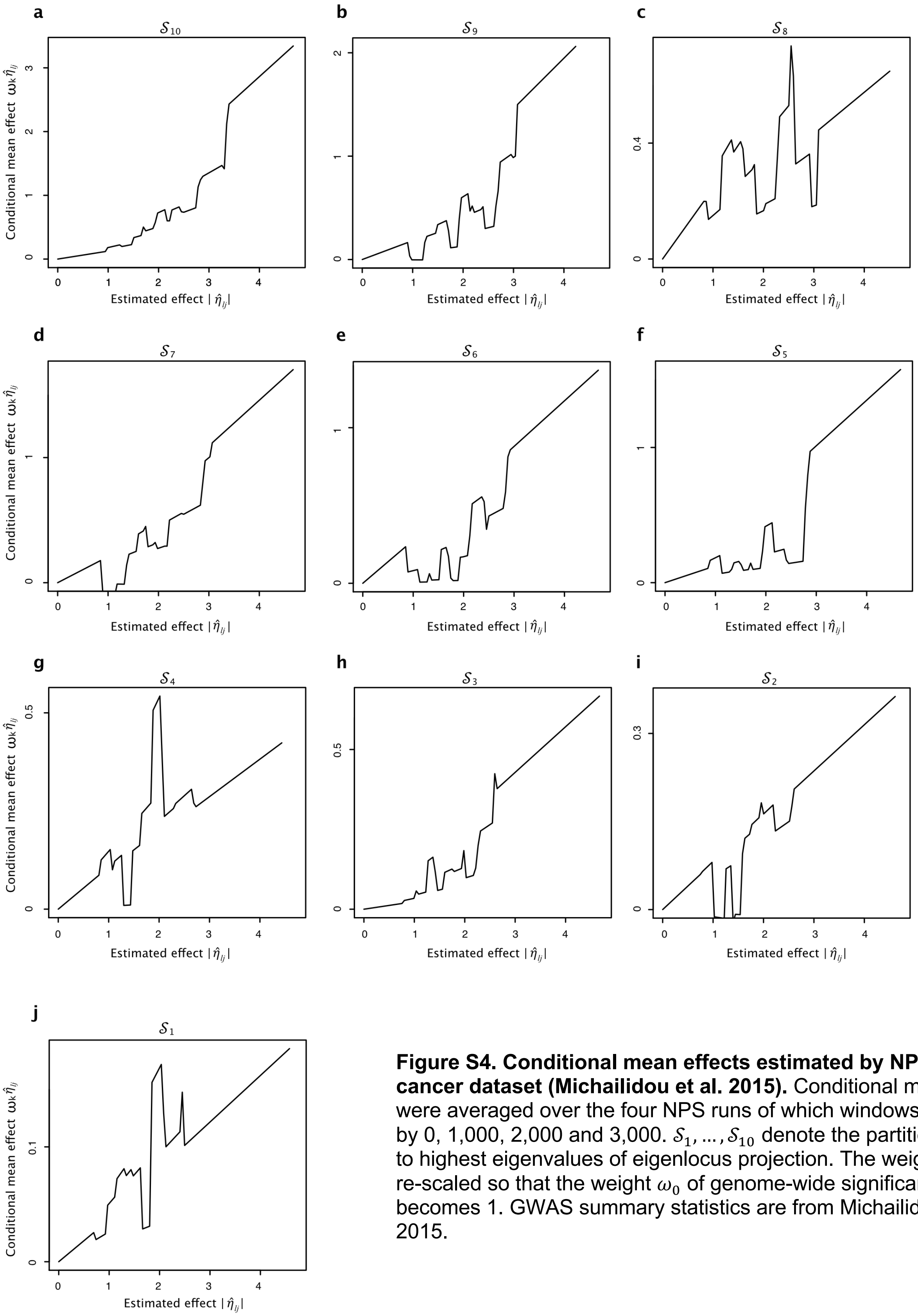

Figure S4. Conditional mean effects estimated by NPS in breast cancer dataset (Michailidou et al. 2015). Conditional mean effects were averaged over the four NPS runs of which windows were shifted by 0, 1,000, 2,000 and 3,000. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denote the partitions of lowest 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. GWAS summary statistics are from Michailidou et al. 2015.

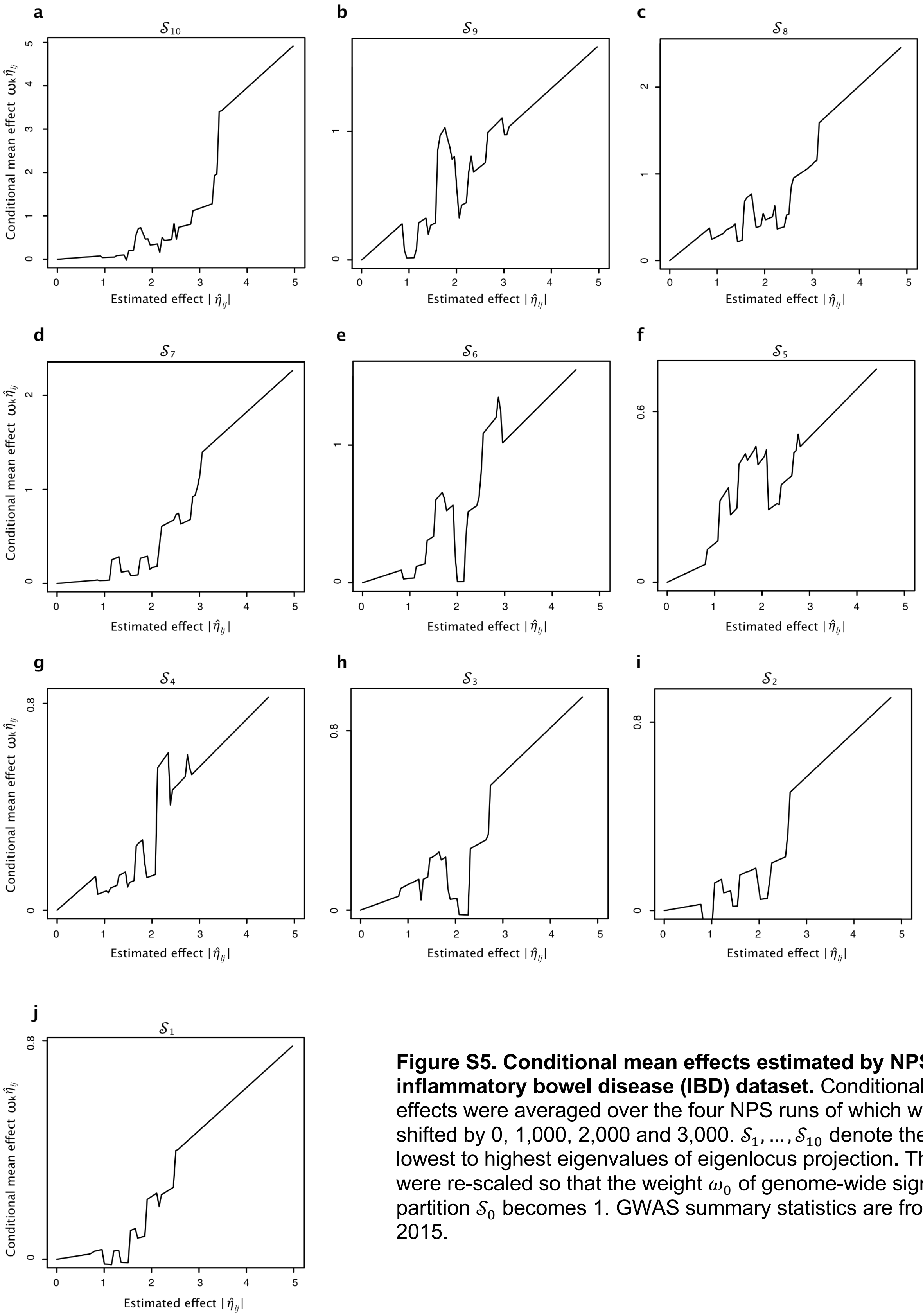

Figure S5. Conditional mean effects estimated by NPS in inflammatory bowel disease (IBD) dataset. Conditional mean effects were averaged over the four NPS runs of which windows were shifted by 0, 1,000, 2,000 and 3,000. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denote the partitions of lowest to highest 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. GWAS summary statistics are from Liu et al. 2015.

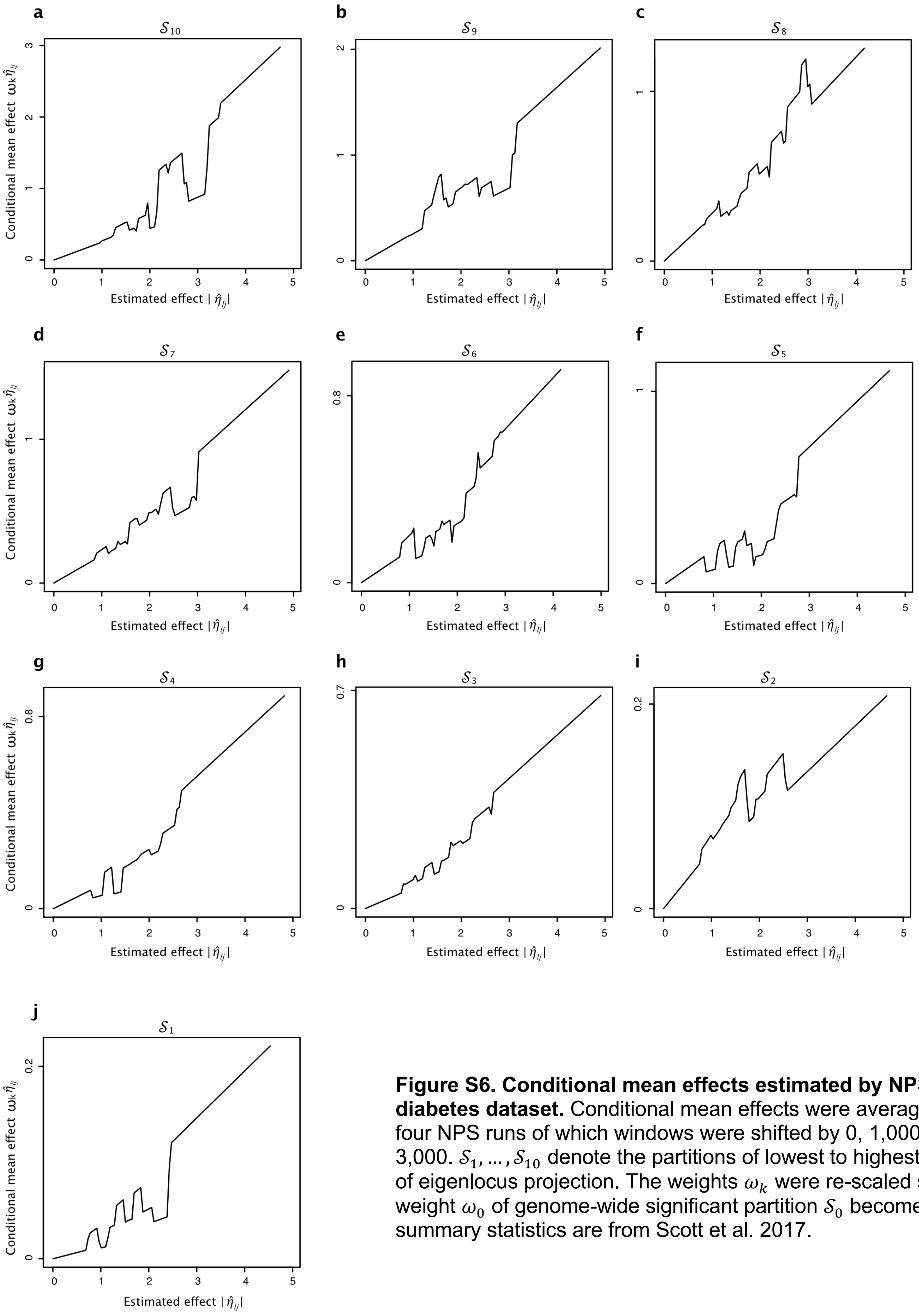

Figure S6. Conditional mean effects estimated by NPS in type 2 diabetes dataset. Conditional mean effects were averaged over the four NPS runs of which windows were shifted by 0, 1,000, 2,000 and 3,000. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denote the partitions of lowest to highest eigenvalues of 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. GWAS summary statistics are from Scott et al. 2017.

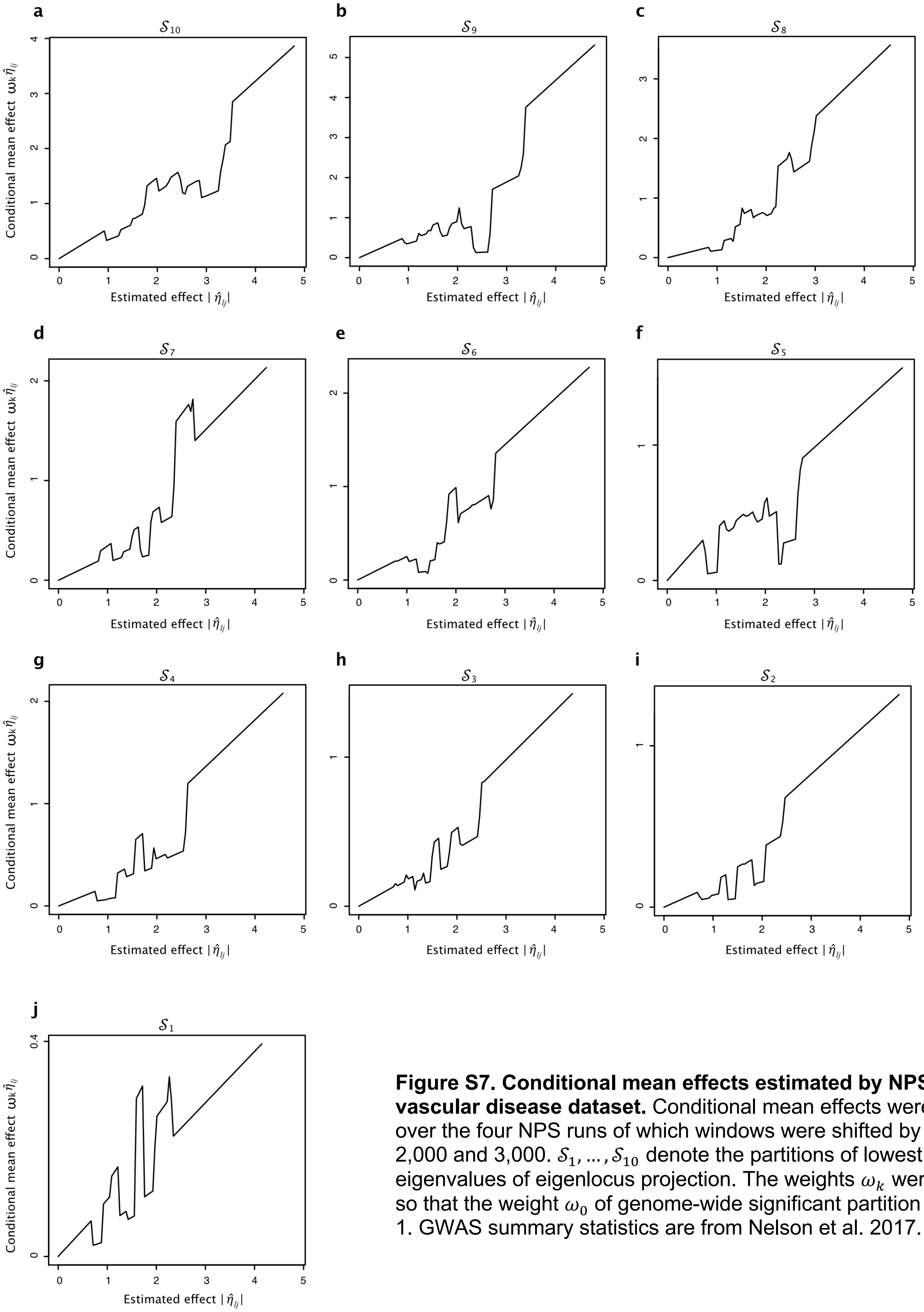

Figure S7. Conditional mean effects estimated by NPS in cardio-vascular disease dataset. Conditional mean effects were averaged over the four NPS runs of which windows were shifted by 0, 1,000, 2,000 and 3,000. $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$ denote the partitions of lowest to highest eigenvalues 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. GWAS summary statistics are from Nelson et al. 2017.

Table S1. Comparison of prediction accuracy in genetic architectures simulating uniformly distributed causal SNPs.

Genetic Architecture	% causal SNPs	Method	Validation		NPS R^2_{Nag} compared to		
			$R^2_{Nagelkerke}$	$R^2_{Liability}$	P+T	LDPred	PRS-CS
(**a**) Point-Normal (GCTA)	1%	P+T	0.049	0.072			
		LDPred	0.071	0.103			
		PRS-CS	0.072	0.105			
		NPS	0.082	0.120	1.66 *	1.15 *	1.14 *
	0.1%	P+T	0.141	0.205			
		LDPred	0.071	0.102			
		PRS-CS	0.140	0.199			
		NPS	0.169	0.241	1.20 *	2.37 *	1.21 *
	0.01%	P+T	0.189	0.273			
		LDPred	0.076	0.110			
		PRS-CS	0.224	0.325			
		NPS	0.329	0.465	1.74 *	4.36 *	1.47 *
(**b**) Point-Normal with MAF dependency ($\alpha = -0.25$)	1%	P+T	0.050	0.071			
		LDPred	0.073	0.101			
		PRS-CS	0.081	0.115			
		NPS	0.093	0.131	1.87 *	1.27 *	1.16 *
	0.1%	P+T	0.142	0.206			
		LDPred	0.076	0.112			
		PRS-CS	0.152	0.220			
		NPS	0.175	0.253	1.24 *	2.31 *	1.15 *
	0.01%	P+T	0.199	0.293			
		LDPred	0.087	0.126			
		PRS-CS	0.230	0.330			
		NPS	0.329	0.471	1.66 *	3.78 *	1.43 *

NPS is more accurate than Pruning and Thresholding (P+T), LDPred and PRS-CS in simulated datasets. 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 (Speed et al. 2017, Zeng et al. 2018). 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 NPS slightly more accurate than under (**a**) GCTA models. Binary phenotypes were simulated with the heritability of 0.5 on the liability scale and 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 accuracies were measured in the validation cohort of 50,000 samples and averaged over three simulations. The star (*) indicates that Nagelkerke's R^2 is significantly different (paired t-test; P < 0.05).

Table S2. Accuracy of NPS in genetic architectures simulating the enrichment of causal SNPs within DNase I Hypersensitive Sites (DHS).

Fraction of causal SNPs	Training	Validation		
	AUC	$R^2_{Nagelkerke}$	$R^2_{Liability}$	AUC
1%	0.746	0.082	0.126	0.708
	0.737	0.083	0.125	0.708
	0.725	0.089	0.118	0.716
0.1%	0.800	0.174	0.249	0.793
	0.811	0.188	0.262	0.808
	0.802	0.179	0.261	0.798
	0.810	0.176	0.254	0.798
	0.810	0.176	0.250	0.798
	0.813	0.178	0.259	0.799
0.01%	0.891	0.325	0.463	0.887
	0.894	0.323	0.462	0.885
	0.887	0.336	0.463	0.889

Each row represents the prediction accuracy of NPS in an individual simulation run. The prediction accuracy of NPS decreased slightly compared to simulations of uniformly distributed causal SNPs (Table S1) but still remained robust. We did not train NPS prediction models using functional annotations. The causal fractions of 1% and 0.01% were replicated three times each, and the causal fraction of 0.1% was 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. Binary phenotypes were simulated with the heritability of 0.5 on the liability scale and 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 accuracies were measured in validation cohorts of 50,000 samples. AUC – Area Under the Curve.

Table S3. 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	$R^2_{Nagelkerke}$	$R^2_{Liability}$	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. Binary phenotypes were simulated with the heritability of 0.5 on the liability scale and 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 accuracies were measured in validation cohorts of 50,000 samples. AUC – Area Under the Curve.

Table S4. Accuracy of pruning and thresholding in genetic architectures simulating the enrichment of causal SNPs within DNase I Hypersensitive Sites (DHS).

Fraction of	Training			Validation		
causal SNPs	P cutoff	# SNPs	AUC	$R^2_{Nagelkerke}$	$R^2_{Liability}$	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. Binary phenotypes were simulated with the heritability of 0.5 on the liability scale and 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.

Table S5. Accuracy of PRS-CS in genetic architectures simulating the enrichment of causal SNPs within DNase I Hypersensitive Sites (DHS).

Fraction of causal SNPs	Training		Validation		
	$\hat{\phi}$	AUC	R^2_{Nag}	$R^2_{Liability}$	AUC
1%	0.01	0.720	0.072	0.110	0.693
	0.0001	0.696	0.074	0.107	0.697
	0.01	0.700	0.079	0.113	0.705
0.1%	0.0001	0.771	0.157	0.221	0.780
	0.0001	0.773	0.164	0.227	0.789
	0.0001	0.769	0.155	0.224	0.781
	0.0001	0.782	0.156	0.226	0.781
	0.0001	0.768	0.148	0.217	0.777
	0.0001	0.777	0.157	0.229	0.781
0.01%	0.000001	0.835	0.230	0.332	0.835
	0.000001	0.835	0.222	0.322	0.830
	0.000001	0.819	0.232	0.326	0.833

Each row represents the prediction accuracy of PRS-CS 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. Binary phenotypes were simulated with the heritability of 0.5 on the liability scale and prevalence of 5%. The prediction model reaching the highest accuracy in a training cohort was selected for validation. $\hat{\phi}$ denotes the model parameter ϕ of best-performing model in training. The reference LD panel was derived from a cohort sampled under the same LD structure. 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.

Table S6. Accuracy of NPS applied to real GWAS summary statistics and UK Biobank datasets.

	Training		Validation (UK Biobank)	
GWAS	**# Projections**	**AUC**	**AUC**	**Tail OR (5%)**
Breast Cancer 2015	120,886	0.656	0.627 [0.62-0.64]	2.53 [2.3-2.8]
Breast Cancer 2017	124,061	0.678	0.654 [0.65-0.66]	3.01 [2.7-3.3]
IBD	110,157	0.686	0.659 [0.65-0.67]	3.60 [3.2-4.0]
Type 2 Diabetes	139,106	0.697	0.686 [0.68-0.69]	3.81 [3.6-4.1]
CAD	105,162	0.778	0.738 [0.72-0.76]	5.21 [4.3-6.2]

GWAS summary statistics for breast cancer, inflammatory bowel disease (IBD), type 2 diabetes, coronary artery disease (CAD) were obtained from Michailidou et al. 2015, Michailidou et al. 2017, Liu et al. 2015, Scott et al. 2017, and Nelson et al. 2017, respectively. The training and validation cohorts were both assembled using UK Biobank samples (see Table 2). The number of projections represents the total number of independent projection eigenvectors used for NPS training across the genome. The 5% 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 AUCs (Area Under the Curve) and tail ORs, which were estimated by DeLong’s method and bootstrapping, respectively. T2D and CAD models were trained and validated with the sex covariate.

Table S7. Accuracy of LDPred applied to real GWAS summary statistics and UK Biobank datasets.

	Training			**Validation (UK Biobank)**	
GWAS	**# SNPs**	**Estimated causal fraction**	**AUC**	**AUC**	**Tail OR (5%)**
Breast Cancer 2015	3,417,759	0.01	**0.630**	0.618 [0.61-0.63]	2.42 [2.2-2.7]
Breast Cancer 2017	3,478,993	0.1	**0.621**	0.615 [0.61-0.62]	2.33 [2.1-2.6]
IBD	3,396,783	0.03	**0.640**	0.641 [0.63-0.65]	2.77 [2.4-3.1]
Type 2 Diabetes	3,451,818	0.01	**0.680**	0.679 [0.67-0.68]	3.51 [3.3-3.8]
CAD	3,405,299	0.003	0.753	0.738 [0.72-0.76]	5.17 [4.3-6.1]
Breast Cancer 2015	351,917	0.3	0.605	0.597 [0.59-0.61]	2.25 [2.0-2.5]
Breast Cancer 2017	353,627	1.0	0.606	0.604 [0.60-0.61]	2.03 [1.8-2.3]
IBD	353,325	1.0	0.618	0.622 [0.61-0.63]	2.76 [2.4-3.1]
Type 2 Diabetes	354,110	0.1	0.679	0.680 [0.67-0.69]	3.63 [3.4-3.9]
CAD	329,644	0.03	**0.757**	0.742 [0.72-0.76]	5.65 [4.7-6.7]

GWAS summary statistics for breast cancer, inflammatory bowel disease (IBD), type 2 diabetes, coronary artery disease (CAD) were obtained from Michailidou et al. 2015, Michailidou et al. 2017, Liu et al. 2015, Scott et al. 2017, and Nelson et al. 2017, respectively. The training and validation cohorts were both assembled using UK Biobank samples (see Table 2). LDPred was ran using all hard-called common SNPs (top) as well as directly genotyped SNPs (bottom). The prediction models producing a higher AUCs in training cohorts, indicated in bold, were chosen for Table 2. LDPred runs only with hard-called genotypes and automatically excludes complementary alleles; therefore, the number of input SNPs are fewer than the number of all available imputed SNPs across the genome. 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 numbers in brackets are the 95% confidence intervals for AUCs (Area Under the Curve) and tail ORs, which were estimated by DeLong’s method and bootstrapping, respectively. T2D and CAD models were trained and validated with the sex covariate.

Table S8. 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	0.607 [0.60-0.62]	2.07 [1.9-2.3]
Breast Cancer 2017	0.0003	1,521	0.627	0.621 [0.61-0.63]	2.37 [2.1-2.6]
IBD	0.0002	621	0.648	0.644 [0.63-0.65]	3.00 [2.7-3.4]
Type 2 Diabetes	0.0004	691	0.661	0.659 [0.65-0.67]	3.04 [2.8-3.3]
CAD	0.025	8,915	0.739	0.719 [0.70-0.74]	5.17 [4.3-6.1]

GWAS summary statistics for breast cancer, inflammatory bowel disease (IBD), type 2 diabetes, coronary artery disease (CAD) were obtained from Michailidou et al. 2015, Michailidou et al. 2017, Liu et al. 2015, Scott et al. 2017, and Nelson et al. 2017, respectively. The training and validation cohorts were both assembled using UK Biobank samples (see Table 2). 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 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. T2D and CAD models were trained and validated with the sex covariate.

Table S9. Accuracy of PRS-CS applied to real GWAS summary statistics and UK Biobank datasets.

GWAS	Training			Validation (UK Biobank)	
	# SNPs	$\hat{\phi}$	AUC	AUC	Tail OR (5%)
Breast Cancer 2015	712,303	0.0001	0.635	0.626 [0.62-0.63]	2.60 [2.3-2.9]
Breast Cancer 2017	711,549	0.0001	0.657	0.651 [0.64-0.66]	2.96 [2.7-3.3]
IBD	707,371	0.0001	0.665	0.668 [0.66-0.68]	3.67 [3.3-4.1]
Type 2 Diabetes	715,952	0.0001	0.686	0.688 [0.68-0.69]	3.99 [3.7-4.3]
CAD	708,976	0.0001	0.763	0.739 [0.72-0.76]	4.92 [4.1-5.8]

GWAS summary statistics for breast cancer, inflammatory bowel disease (IBD), type 2 diabetes, coronary artery disease (CAD) were obtained from Michailidou et al. 2015, Michailidou et al. 2017, Liu et al. 2015, Scott et al. 2017, and Nelson et al. 2017, respectively. The training and validation cohorts were both assembled using UK Biobank samples (see Table 2). The prediction model reaching the highest accuracy in a training cohort was selected for validation. $\hat{\phi}$ denotes the model parameter ϕ of best-performing model in training. The European reference LD panel provided with the software was used for training. PRS-CS uses only HapMap 3 SNPs. 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. T2D and CAD models were trained and validated with the sex covariate.

.

Table S10. Accuracy of NPS in independent validation cohorts.

	Training		Validation (Partners Biobank)	
GWAS	**# Projections**	**AUC**	**AUC**	**Tail OR (5%)**
Breast Cancer 2017	124,061	0.678	0.624 [0.60-0.64]	2.32 [1.7-3.0]
IBD	110,157	0.686	0.686 [0.67-0.70]	4.32 [3.5-5.2]
Type 2 Diabetes	139,106	0.697	0.647 [0.63-0.66]	2.97 [2.6-3.5]
CAD	105,162	0.778	0.615 [0.58-0.65]	4.10 [2.8-5.8]

The polygenic risk models trained in UK Biobank (Table 2) were validated in US white population (Table 3; Partners Biobank). The identical polygenic risk prediction models reported in Tables 2 and S6 were validated in Partners Biobank without re-training or model adjustment. 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. T2D and CAD models were trained and validated with the sex covariate.

Table S11. Accuracy of LDPred in independent validation cohorts.

	Training (UK Biobank)			Validation (Partners Biobank)	
GWAS	# SNPs	Estimated causal fraction	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	0.665	0.635 [0.62-0.65]	2.51 [2.1-2.9]
CAD	1,224,034	0.003	0.724	0.595 [0.56-0.63]	2.31 [1.4-3.5]

The polygenic risk models were trained with LDPred in UK Biobank cohorts and validated in US white population (Table 3; Partners Biobank). The training cohorts are identical to those in Tables 2 and S7, 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 hard 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. T2D and CAD models were trained and validated with the sex covariate.

Table S12. 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	0.656	0.623 [0.61-0.64]	2.10 [1.8-2.5]
CAD	0.15	15,908	0.739	0.611 [0.58-0.65]	2.72 [1.8-3.9]

The polygenic risk models were trained with pruning and thresholding algorithm in UK Biobank cohorts and validated in US white population (Table 3; Partners Biobank). The training cohorts are identical to those in Tables 2 and S8, however, the prediction models were reconstructed by re-running P+T on the SNPs found in both training and validation cohorts. 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. T2D and CAD models were trained and validated with the sex covariate.

Table S13. Accuracy of PRS-CS in independent validation cohorts.

GWAS	Training			Validation (Partners Biobank)	
	# SNPs	**Estimated ϕ**	**AUC**	**AUC**	**Tail OR (5%)**
Breast Cancer 2017	512,117	0.0001	0.647	0.624 [0.60-0.64]	2.23 [1.7-2.9]
IBD	509,143	0.0001	0.663	0.682 [0.66-0.70]	4.11 [3.3-5.0]
Type 2 Diabetes	515,164	0.0001	0.685	0.649 [0.64-0.66]	2.80 [2.4-3.3]
CAD	510,103	0.0001	0.751	0.621 [0.58-0.66]	3.16 [2.1-4.4]

The polygenic risk models were trained with PRS-CS in UK Biobank cohorts and validated in US white population (Table 3; Partners Biobank). The training cohorts are identical to those in Tables 2 and S9, however, the prediction models were reconstructed by re-running PRS-CS on the SNPs found in both training and validation cohorts as recommended by the authors. PRS-CS uses only HapMap 3 SNPs. The prediction model reaching the highest accuracy in a training cohort was selected for validation. $\hat{\phi}$ denotes the model parameter ϕ of best-performing model in training. The European reference LD panel provided with the software was used for training. 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. T2D and CAD models were trained and validated with the sex covariate.