BMC Bioinformatics

RESEARCH Open Access

Incorporating functional annotation with bilevel continuous shrinkage for polygenic risk prediction

Yongwen Zhuang[1], Na Yeon Kim[2], Lars G. Fritsche[1], Bhramar Mukherjee[1*] and Seunggeun Lee[2*]

*Correspondence:
bhramar@umich.edu;
lee7801@snu.ac.kr

[1] University of Michigan, Ann Arbor, USA
[2] Seoul National University, Seoul, Republic of Korea

Abstract

Background: Genetic variants can contribute differently to trait heritability by their functional categories, and recent studies have shown that incorporating functional annotation can improve the predictive performance of polygenic risk scores (PRSs). In addition, when only a small proportion of variants are causal variants, PRS methods that employ a Bayesian framework with shrinkage can account for such sparsity. It is possible that the annotation group level effect is also sparse. However, the number of PRS methods that incorporate both annotation information and shrinkage on effect sizes is limited. We propose a PRS method, PRSbils, which utilizes the functional annotation information with a bilevel continuous shrinkage prior to accommodate the varying genetic architectures both on the variant-specific level and on the functional annotation level.

Results: We conducted simulation studies and investigated the predictive performance in settings with different genetic architectures. Results indicated that when there was a relatively large variability of group-wise heritability contribution, the gain in prediction performance from the proposed method was on average 8.0% higher AUC compared to the benchmark method PRS-CS. The proposed method also yielded higher predictive performance compared to PRS-CS in settings with different overlapping patterns of annotation groups and obtained on average 6.4% higher AUC. We applied PRSbils to binary and quantitative traits in three real world data sources (the UK Biobank, the Michigan Genomics Initiative (MGI), and the Korean Genome and Epidemiology Study (KoGES)), and two sources of annotations: ANNOVAR, and pathway information from the Kyoto Encyclopedia of Genes and Genomes (KEGG), and demonstrated that the proposed method holds the potential for improving predictive performance by incorporating functional annotations.

Conclusions: By utilizing a bilevel shrinkage framework, PRSbils enables the incorporation of both overlapping and non-overlapping annotations into PRS construction to improve the performance of genetic risk prediction. The software is available at https://github.com/styvon/PRSbils.

Keywords: PRSbils, Polygenic risk score, Functional annotation, KEGG pathway, Biobank data, Continuous shrinkage

Background

Genetic data are important resources to improve the risk prediction for complex diseases [1]. The genetic effects of variants across the genome can be summarized in the form of polygenic risk scores (PRS) that estimate individuals' genetic liability. The wide availability of summary statistics from large-scale genome-wide association studies (GWAS), which test the associations between the genetic variants and phenotypes of interest across the genome, has facilitated the application of PRS [2]. Early PRS approaches such as pruning and thresholding (P + T) included only a selection of SNPs that reach genome-wide significance [3, 4], while it was suggested by later studies that including all SNPs and applying shrinkage on their weights would increase the heritability estimates [5]. Although PRS has shown great promise in the early identification and prediction of disease risks, its predictive performance for explaining the full genetic contribution to diseases remains limited. Studies have indicated that polygenic risk scores explain only a small amount of total phenotypic variability of complex traits [6, 7], and improvements in the construction of PRS, especially selecting and shrinking SNP effects to better capture genetic heritability, remain an area of rigorous development.

Genetic variants in different functional categories can have different shares of contribution to the heritability of complex traits, and recent studies have shown that the incorporation of functional annotation can improve the predictive performance of PRS [8, 9]. In addition, for certain phenotypes where only a small proportion of variants are causal, PRS methods with Bayesian continuous shrinkage framework have been proposed to account for such sparsity while yielding higher performance [10]. It is possible that the annotation group level effect is also sparse and accounting for this factor holds potential for additional performance improvement. Although existing PRS methods improve prediction accuracy through utilizing GWAS summary statistics and accounting for the potential sparsity of the genetic architectures, the number of PRS methods that incorporate annotation information and apply continuous shrinkage on effect sizes is limited. For example, the sparsity assumption for the underlying genetic architectures is generally made either on a global level [10] or by partitioning variants into bins with similar sum of squared posterior mean effect sizes [11, 12], and there is a lack of study on addressing sparsity across the annotation groups.

In this paper, we propose a PRS method with bilevel continuous shrinkage prior to leverage the functional annotation information to bridge the above-mentioned gap. This prior accommodates the varying genetic architectures both on the variant-specific level and on the functional annotation level, and the posterior update is conducted using a Gibbs sampler. PRSbils uses GWAS summary statistics instead of individual-level data and accounts for local LD patterns through an external LD reference panel. We conducted simulation studies and investigated the predictive performance in settings with different genetic architectures. Results indicated that PRSbils outperformed the benchmark method in all settings. We applied PRSbils to binary and quantitative traits in three real world data sources, and demonstrated that the proposed method generally improved predictive performance. In summary, our study showed that PRSbils holds the potential for improving predictive performance by incorporating functional annotations using a novel bilevel shrinkage approach.

Methods

Overview of Bayesian continuous shrinkage regression model

Denote y as an N-vector of standardized phenotype, G as an $N \times M$ matrix of standardized genotypes, ϵ as an N-vector of random noise, and β as an M-vector of genetic effect sizes. Then, a regression model of genotypes and phenotypes can be expressed as

$$y = G\beta + \epsilon$$

where $\epsilon \sim MVN(0, \sigma^2 I_N), p(\sigma^2) \propto \sigma^{-2}$, i.e., $y|G, \beta, \sigma^2$ follows a multivariate Normal distribution with mean $G\beta$ and covariance matrix $\sigma^2 I_N$.

For high dimensional genetic data, the number of genetic variants M is much larger than the number of individuals N, and it is often assumed that the genetic effect vector β is sparse, meaning that only a small amount of the variants are associated with the outcome phenotype. Under this sparsity assumption, the prior distribution of β can be chosen to be either a discrete or a continuous mixture of Normal distributions. The discrete mixture type of prior is also known as the spike-and-slab prior [13], and is a combination of a point mass at 0 and a density for the non-zero part.

The continuous mixture type of prior assigns β with a continuous distribution centered at 0. One commonly used set of priors is the global–local shrinkage priors [14], which utilizes both a global shrinkage parameter τ^2 and local marker-specific parameters $\lambda_1^2, ..., \lambda_M^2$ to model the prior distribution of β. Specifically:

$$\beta|\sigma^2, \tau^2, \lambda_1^2, ..., \lambda_M^2 \sim MVN(0, \frac{\sigma^2}{N}\tau^2 V_\lambda)$$

$$\lambda_j^2 \sim \pi_1(\lambda_j^2) d\lambda_j^2,\ j \in \{1, ..., M\}$$

$$\tau^2 \sim \pi_2(\tau^2) d\tau^2$$

where $V_\lambda \equiv diag\{\lambda_1^2, ..., \lambda_M^2\}$ is an M × M diagonal matrix, and π_1 and π_2 are absolutely continuous functions and have a wide range of choices. For example, the model becomes Lasso when λ_j^2 follows the standard exponential distribution; a horseshoe prior is constructed when both τ and λ_j follow a standard half-Cauchy distribution.

The above-mentioned Bayesian model can be generalized to the bilevel global–local shrinkage models to account for additional group information [15]. We consider the situation where each variant belongs to one of K mutually-exclusive annotation groups, i.e., $A_j = k$ if variant j belongs to annotation group k. Then the Bayesian bilevel global–local shrinkage regression model can be expressed as:

$$\beta|\sigma^2, \delta_{A_1}^2, ..., \delta_{A_M}^2, \lambda_1^2, ..., \lambda_M^2 \sim MVN(0, \frac{\sigma^2}{N} V_\delta V_\lambda)$$

$$\lambda_j^2 \sim \pi_1\left(\lambda_j^2\right) d\lambda_j^2,\ j \in \{1, \ldots, M\}$$

$$\delta_k^2 \sim \pi_2(\delta_k^2) d\delta_k^2,\ k \in \{1, ..., K\}$$

where $V_\delta \equiv diag\{\delta^2_{A_1}, ..., \delta^2_{A_M}\}$, with $\delta^2_{A_j}$ being the group-level shrinkage parameter when the corresponding group for the jth variant is A_j.

PRSbils

Based on the Bayesian bilevel global–local shrinkage regression model, we incorporate functional annotation as the group-level information and assume a standard half-Cauchy prior $C^+(0,1)$ for each group-level shrinkage parameter δ_k and each local shrinkage parameter λ_j. The Bayesian regression model can then be specified as:

$$y|G,\beta,\sigma^2 \sim MVN(G\beta, \sigma^2 I_N)$$

$$\beta|\sigma^2, \delta^2_1, ..., \delta^2_K, \lambda^2_1, ..., \lambda^2_M \sim MVN(0, \frac{\sigma^2}{N} V_\delta V_\lambda)$$

$$\sigma^2 \sim \sigma^{-2} d\sigma^2$$

$$\delta_k \sim C^+(0,1)$$

$$\lambda_j \sim C^+(0,1)$$

The posterior distribution of β and σ^2 can then be derived:

$$\beta|\cdot \sim MVN(B^{-1}G^T y, \sigma^2 B^{-1})$$

$$\sigma^2|\cdot \sim IG(\frac{n+p-1}{2}, \frac{1}{2}[(y-G\beta)^T(y-G\beta) + \beta^T(V_\delta V_\lambda)^{-1}\beta)]$$

where $B = G^T G + \mathrm{N}(V_\delta V_\lambda)^{-1}$.

When the LD matrix D and the summary-level estimation of genetic effect size $\widehat{\beta}$ are available, we can obtain an approximation for the posterior distribution with $\widehat{\beta} = G^T y/N$ and $D = G^T G/N$:

$$\beta|\cdot \sim MVN(\frac{N}{\sigma^2}\Sigma\widehat{\beta}, \Sigma)$$

$$\sigma^2|\cdot \sim IG(\frac{1}{2}(N+M), \frac{N}{2}\{1 - 2\beta^T\widehat{\beta} + \beta^T[D + (V_\delta V_\lambda)^{-1}]\beta\})$$

where $\Sigma = \frac{\sigma^2}{N}[D + (V_\delta V_\lambda)^{-1}]^{-1}$.

To derive the posterior distributions for shrinkage parameters δ and λ, we note that the standard half-Cauchy distribution can be decomposed into a scale mixture of inverse Gamma distributions [16]. Let x, a be random variables satisfying

$$x^2|a \sim IG(\frac{1}{2}, \frac{1}{a}), a \sim IG(\frac{1}{2}, 1)$$

where $IG(\alpha, \beta)$ is the inverse Gamma distribution with shape parameter α and scale parameter β, then $x \sim C^{+}(0,1)$.

PRSbils with non-overlapping annotation assignment

We first consider the situation where each variant has only one annotation. Using the scale mixture representation of the standard half-Cauchy distribution, we obtain an alternative representation of the priors for δ and λ:

$$\delta_k | t_k \sim IG(\frac{1}{2}, \frac{1}{t_k}), t_k \sim IG(\frac{1}{2}, 1),\ k \in \{1, ..., K\}$$

$$\lambda_j | c_j \sim IG(\frac{1}{2}, \frac{1}{c_j}), c_j \sim IG(\frac{1}{2}, 1),\ j \in \{1, ..., M\}$$

The posteriors for δ and hyper-parameter t can be expressed as

$$\delta_k^2 | \cdot \sim IG(\frac{M_k + 1}{2}, \frac{\sum_{j \in \{j: A_j = k\}} \frac{N\beta_j^2}{\lambda_j^2}}{2\sigma^2} + \frac{1}{t_k})$$

$$t_k | \delta_k^2 \sim IG(1, \delta_k^{-2} + 1)$$

where M_k denotes the number of variants within group k, $k = 1, ..., K$.

We obtain the posteriors for λ and c in a similar fashion:

$$\lambda_j^2 | \beta_j, \sigma^2, \delta_{A_j}^2, c_j \sim IG(1, \frac{N\beta_j^2}{2\sigma^2} + \frac{1}{c_j})$$

$$c_j | \lambda_j^2, \delta_{A_j}^2 \sim IG(1, \lambda_j^{-2} + 1)$$

Since all the conditional distributions for the parameters are known, posterior samples of β can be obtained by a Gibbs sampler. After the posterior β values (denoted as $\widetilde{\beta}$) are achieved, we construct the final PRS score by combing the group-wise scores across all annotations

$$PRS = \sum_{k=1}^{K} \alpha_k PRS_k$$

where $PRS_k = \sum_{\{j: A_j = k\}} G_j \widetilde{\beta}_j$ is the group-wise score, and α_k is the group-specific weight for group k. We obtain the estimates for $\alpha_1, ..., \alpha_k$ through tenfold cross-validation using a separate validation data, which includes individual-level genotype and phenotype data. We use this group-wise combination approach based on the observation that signal-enriched annotations are more informative for prediction, and weighting each partition differently can further improve PRS performance [11].

We also investigated a hybrid method that is a combination of PRSbils and conventional PRS methods. In this case, we construct the PRS score by

$$PRS = \sum_{k=1}^{K} \alpha_k PRS_k + \gamma \sum_{j=1}^{M} G_j \widetilde{\beta}_j{}'$$

where $\widetilde{\beta}_j{}'$ denotes the genetic effect size generated from conventional PRS methods.

PRSbils with overlapping annotation assignment

Under the situation where a variant belongs to multiple annotation groups, for example, one variant can be involved in multiple pathways, we account for such a variant separately in each of the annotation groups it belongs to. The assumption underlying this is that variants with more annotations have potentially larger contribution to the heritability and therefore will be shrunk less and have larger posterior effect sizes in general [17]. The total number of variants in the overlapping setting will become $M' = \sum_{k=1}^{K} \sum_{j=1}^{M} I\left(k \in A_j\right)$ (Additional file 1: Figure S10), and the local shrinkage parameters will be assigned to each of the M' variants:

$$\lambda^2_{j,A_j} | \beta_{j,A_j}, \sigma^2, \delta^2_{A_j}, c_{j,A_j} \sim IG(1, \frac{N\beta^2_{j,A_j}}{2\sigma^2} + c_{j,A_j})$$

$$c_{j,A_j} | \lambda^2_{j,A_j}, \delta^2_{A_j} \sim IG(1, \lambda^{-2}_{j,A_j} + 1)$$

The group-wise score can be calculated by $PRS_k = \sum_{\{j:A_j=k\}} G_j \widetilde{\beta}_{j,A_j}$.

PRS-CS

PRS-CS is a PRS method which infers the posterior genetic effect size β using summary-level $\widehat{\beta}$ from existing GWAS studies as well as LD information from a reference panel. It is based on the Bayesian continuous shrinkage regression model without group information and obtains posterior samples using a Gibbs sampler. A general gamma-gamma distribution is assigned to the local shrinkage parameters λ^2_j:

$$\lambda_j | c_j \sim G(a_0, c_j), c_j \sim G(b_0, 1)$$

with $G(\cdot, \cdot)$ representing a Gamma distribution with shape and scale parameters, and a_0 and b_0 being pre-specified constants. When $a_0 = 0.5$ and $b_0 = 1$, it is equivalent to the scale mixture representation of the standard half-Cauchy distribution.

When a prior guess of the global shrinkage parameter τ is not available, PRS-CS either uses a grid search for the best performing value in an additional validation set (PRS-CS), or assigns a standard half-Cauchy prior on τ in the fully Bayesian model (PRS-CS-auto).

For the posterior sampling part, PRSbils is an extension of the PRS-CS-auto approach to differentiate the shrinkage across different groups. When the number of groups $K = 1$, our approach is equivalent to PRS-CS-auto with hyper-parameters $a_0 = 0.5$ and $b_0 = 1$.

LDpred-funct

LDpred-funct incorporates functional annotation as priors for the genetic effects using the baseline-LD model which includes non-overlapping annotations [12]. It assumes a prior distribution $\beta_j \sim N\left(0, c\sigma_j^2\right)$ for the normalized genetic effects, where σ_j^2 represents per-SNP heritability obtained from stratified LD score regression [18] and c is a normalizing constant. The posterior mean of β is

$$E[\beta|\cdot] = W^{-1}N\widehat{\beta}$$

where $W^{-1} = \left[ND + \frac{1}{c}diag\left(\frac{1}{\sigma_1^2}, ..., \frac{1}{\sigma_M^2}\right)\right]^{-1}$. The SNPs are then ranked by the absolute posterior mean effect sizes and partitioned into L bins with approximately the same sum of squared posterior mean effect sizes. The PRS is generated by

$$PRS = \sum_{l=1}^{L} \alpha_l PRS(l)$$

where the weights are determined via tenfold cross-validation.

To make LDpred-funct applicable to our study which uses different annotations from the baseline LD model, we used stratified LD score regression to obtain the per-SNP heritability under the functional annotations being used, obtained the posterior mean of β, and get the PRS with the number of bins L fixed at 40.

Biobank data overview

UK Biobank is a large-scale database with biomedical information from UK participants recruited from 2006 to 2010 [18]. The genetic data from UK Biobank consists of over 90 million genetic variants imputed from the Haplotype Reference Consortium (HRC) [19] among 488,377 individuals. Data from the Michigan Genomics Initiative (MGI) [20] and the Korean Genome and Epidemiology Study (KoGES) data [21] were also analyzed in our study. We used Data Freeze 3 of the MGI data, which includes 56,984 genotyped participants at the University of Michigan with over 32 million genome-wide variants imputed from the HRC [22]. The KoGES data includes a total of 72,298 Korean individuals, with over 8 million genetic variants imputed from 1000 Genome project phase 3 + Korean reference genome (397 samples) and with minor allele frequency (MAF) > 0.01, HWE p-value > 1×10^{-6}, variant call rate > 95% [23].

For all genetic data in the UK Biobank, MGI and KoGES, NCBI Build 37/UCSC hg19 was used for genomic coordinates. We further restricted our analysis to HapMap3 SNPs with minimum MAF > 0.01, HWE p-value > 1×10^{-6}, variant call rate > 95%, individual missing rate < 1%, and LD-pruning R2 < 0.99. LD information from 503 European samples in the 1000 Genomes Project (1 KG) [24] was used as an external reference panel for the UK Biobank and MGI data, and 1 KG East Asian reference panel was used for the KoGES data.

Simulation studies

We conducted simulation studies to compare the performance of PRSbils to PRS-CS. We also evaluated a hybrid method of PRSbils with PRS-CS, in which the scores from PRS-CS was combined with the one from the proposed method. We also compared

the predictive performance with LDpred-funct for non-overlapping annotation groups. A total of $M = 125,000$ SNPs were sampled from the UK Biobank data with above-mentioned quality control filters, with 1 KG as LD reference panel. The sampled variants were then assigned to K different annotation groups, which explain $q_1, ..., q_K\%$ of the total heritability h^2 respectively. For each annotation group k, the proportion of causal variants is denoted as p_k. Genetic effect sizes were generated from a mixture of point-Normal models specified as:

$$\beta_j \sim \begin{cases} N\left(0, \frac{q_{A_j} h^2}{p_{A_j} M}\right), & \text{with probability } p_{A_j} \\ 0, & \text{with probability } 1 - p_{A_j} \end{cases}$$

We investigated five simulation settings with different K, p_k and q_k with non-overlapping annotation groups, i.e., each variant is mapped to one and only one annotation (Table 1). For settings 1–4, we fixed the number of annotation groups to 4 ($K = 4$), the proportion of causal variants in each group to 0.5%, 1%, 1.5%, 2% respectively, and vary the proportion of the total heritability explained by each group from a relatively sparse scenario ($q = (0, 0, 10\%, 90\%)$) to a more balanced scenario ($q = (25\%, 25\%, 25\%, 25\%)$). For setting 5, we changed the number of annotation groups to $K = 10$, and considered a situation with more group-wise sparsity where only two of the groups contribute to the total heritability.

In addition, we simulated two settings (settings 6 and 7 in Table 1) with different overlapping patterns (Additional file 1: Figure S1). For both settings, we used four annotation groups contributing 0, 0, 10%, 90% to the total heritability. Overlapping pattern I was used in setting 6, where the intersection over union metric (IOU) was higher among the annotation groups with low heritability contribution. For setting 7, overlapping pattern II was used, where IOU was higher among the annotation groups with high heritability contribution.

We then simulated the phenotypes using the sum of all SNPs weighted by their corresponding genetic effect sizes, together with a Normal random error term to fix the heritability at $h^2 = 0.7$.

Table 1 Summary of parameter settings in the simulation study

Setting	K	M_k	p_k(%)	q_k(%)	Overlap pattern
1	4	49,750, 37,500, 25,125, 12,625	0.5, 1, 1.5, 2	0, 0, 10, 90	Non-overlap
2	4	49,750, 37,500, 25,125, 12,625	0.5, 1, 1.5, 2	0, 0, 50, 50	Non-overlap
3	4	49,750, 37,500, 25,125, 12,625	0.5, 1, 1.5, 2	10, 20, 30, 40	Non-overlap
4	4	49,750, 37,500, 25,125, 12,625	0.5, 1, 1.5, 2	25,25,25,25	Non-overlap
5	10	12,375 × 6, 12,625, 12,625, 12,750, 12,750	0 × 6, 2, 2, 3, 3	0 × 8,10, 90	Non-overlap
6	4	49,750, 37,500, 25,125, 12,625	0.5, 1, 1.5, 2	0, 0, 10, 90	Overlap I
7	4	49,750, 37,500, 25,125, 12,625	0.5, 1, 1.5, 2	0, 0, 10, 90	Overlap II

A total of $M = 125,000$ SNPs were sampled from the UK Biobank data with quality control filters, with 1 KG as LD reference panel. The sampled variants were then assigned to K different annotation groups, which explain $q_1, ..., q_K\%$ of the total heritability h^2 respectively. For each annotation group k, the proportion of causal variants is denoted as p_k. Overlapping pattern I was used in setting 6, where the intersection over union metric (IOU) was higher among the annotation groups with low heritability contribution. For setting 7, overlapping pattern II was used, where IOU was higher among the annotation groups with high heritability contribution

To obtain the summary statistics, we performed GWAS to calculate the marginal genetic effect size estimates $\widehat{\beta}$ using SAIGE[20] version 0.44.3, which is a computationally efficient method that controls for case–control imbalance as well as potential sample relatedness, among $N_{sumstat} = 50,000$ simulated individuals. The summary statistics were used as input for PRSbils and PRS-CS.

The prediction performance was evaluated for both methods on a separate test set consisting of $N_{test} = 24,000$ simulated individuals. AUC and R^2 were used to measure the prediction accuracy. To obtain AUC, we binarize the phenotypes assuming those with top 10% highest phenotype values as the true at-risk population.

Biobank data analysis

We analyzed both binary and quantitative traits for three data sources, i.e., the UK Biobank data, the MGI data, and the KoGES data. For the UK Biobank and the MGI data, we studied type II diabetes as a binary trait, and BMI and LDL as quantitative traits. For the KoGES data, we assessed the results for type II diabetes. The binary type II diabetes trait for genotyped individuals were defined by the PheWAS codes [25] aggregated from ICD codes in the electronic health records for the UK Biobank data and the MGI data, while for the KoGES data it is identified from questionnaire-based interviews. The quantitative traits were obtained as a physical measure in the target data at the initial assessment visit of the participants. For each trait in the analysis, we used a common set of SNPs from the summary statistics, the 1000G reference panel, and the test set. The total number of SNPs used was 1,093,109 for type II diabetes, 986,885 for BMI, and 926,775 for LDL.

The summary statistics used in the analysis were from existing study results. For type II diabetes, we used the result from a GWAS analysis of 407,701 white British UK Biobank participants using SAIGE [26] when analyzing UK Biobank and MGI data, while summary statistics from Biobank Japan [27] was used for the KoGES data. For BMI, we used the GWAS results from GIANT Consortium with 332,153 participants with European ancestry [28]. For LDL, GWAS results from the GLGC Consortium with 188,578 participants with European ancestry [29] were used. A summary of the data souce used in the biobank data analysis is presented in Additional file 1: Table S1.

To avoid the overlapping of samples between the test samples and the samples in the UK Biobank summary statistics for type II diabetes, we applied the PRS methods to a sample of 7,528 white individuals with non-British origin in the UK Biobank. For quantitative traits, since the summary statistics does not overlap with the UK Biobank population, we applied the methods to a test set consisting of around 80,000 white British individuals in the UK Biobank. Two types of group information were used for PRSbils. The first was 186 pathway annotations from the Kyoto Encyclopedia of Genes and Genomes (KEGG), which includes the networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development [30]. Variants were first mapped to Ensembl genes by position, and then from genes to KEGG pathways. The mapping of genetic variants to KEGG annotations is not unique, which means each variant can have multiple KEGG annotations. The second one was six non-overlapping Refseq gene-based functional annotations (i.e. exonic/splicing; ncRNA; UTR5/UTR3; intronic; upstream/

downstream; intergenic) obtained using ANNOVAR [31]. For variants without any available annotations, PRSbils assigned them to a separate group for no annotations.

For each trait, we adopted a tenfold cross-validation approach where the α parameters were estimated using a random sample of 9/10 of the test set, and the performance was validated on the rest 1/10 of the samples in terms of AUC and R^2. For the binary trait, we used Efron's pseudo R^2 [32] instead of the ordinary least square R^2. For the continuous traits BMI and LDL, we calculated the AUC by binarizing them with thresholds of 25 (threshold for overweight) and 4.1 mmol/L (or 160 mg/dL, threshold for high LDL) respectively.

Results

Simulation study results

We evaluated prediction performance of PRSbils, PRS-CS and the hybrid method in seven simulation settings (Table 1). For the settings with non-overlapping annotation groups (Settings 1–5), we also evaluated the performance of LDpred-funct. Since the estimated per-SNP heritability might not be stable due to the relatively small sample size, we used the true stratified heritability values for LDpred-funct.

Simulation Settings 1–4 consider a fixed number of total annotation groups $K = 4$. Figure 1 shows that when there is a relatively large variability of group-wise heritability contribution, the gain in prediction performance from PRSbils is the largest compared to PRS-CS. For example, in setting 1, where each annotation group contribute 0%, 0%, 10% and 90% of the total heritability, PRSbils yielded an average AUC of 0.559 (95% CI [0.546, 0.572]), the hybrid PRSbils + PRS-CS method yielded an average AUC of 0.561 (95% CI [0.548, 0.575]), compared to an average AUC of 0.527 (95% CI [0.500, 0.555]) from PRS-CS and 0.540 (95% CI [0.523, 0.558]) from LDpred-funct. Similar patterns of AUC were shown in other settings. The improved average performance over the benchmark methods can be explained by the group-wise shrinkage parameter allowing for shrinking the group with high heritability contribution differently than other annotation

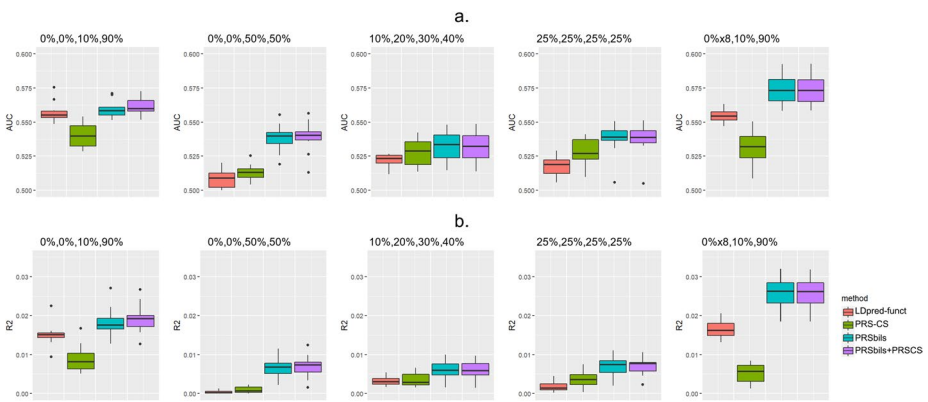

Fig. 1 Comparison of prediction performance in simulation studies with non-overlapping annotation groups (Settings 1–5), measured by AUC (**a**) and R^2 (**b**). A total of $M = 125,000$ SNPs were sampled from the UK Biobank data with 1 KG as LD reference panel. Genetic effects were generated using a mixture of point-Normal models with total heritability fixed at 0.7. GWAS results from $N_{sumstat} = 50,000$ simulated individuals were used as summary statistics. Prediction accuracy was evaluated in a test sample of $N_{test} = 24,000$ simulated individuals with tenfold cross-validation

groups, instead of making a uniform shrinkage at the global level. When the difference in group-wise heritability contribution was small, such as in Setting 4 where all groups contribute equally to the total heritability, the prediction performance of PRSbils was similar to PRS-CS. This is expected as the group-wise shrinkage parameters from PRSbils would behave similarly as the global shrinkage parameter in PRS-CS.

In Setting 1 and Setting 5, two groups contribute 10% and 90% to the total heritability, but the total number of annotation groups differ ($K = 4$ in Setting 1 and $K = 10$ in Setting 5). PRSbils yields similar performance in these two settings, but has a slightly larger variability in Setting 5, which yielded an average AUC of 0.574 (95% CI [0.552, 0.595]) (Fig. 1). This is likely because the proposed method only shrinks groups with no heritability contribution to a small value but not exactly to zero, and therefore the large proportion of no heritability annotation groups, the noisier the PRS value will be, making the performance to fluctuate more. In comparison, the PRS-CS method does not incorporate the annotation group information and yielded stable AUCs in Setting 1 and Setting 5 that are consistently lower than PRSbils. LDpred-funct also yielded similar AUCs in Setting 1 and Setting 5, with the performance improvement of PRSbils over LDpred-funct larger in Setting 5, confirming that when the group-wise heritability distribution is sparser, the continuous shrinkage approach adopted by PRSbils performs better in capturing the sparsity pattern.

In Settings 6 and 7, we investigated the influence of using different overlapping patterns of annotation groups on the performance (Fig. 2). PRSbils yielded higher predictive performance compared to the benchmark method in both settings, with an average 8.4% gain in AUC for Setting 6 and 0.8% for Setting 7. An explanation for the difference in the performance gain is that the high IOU between annotation group 3 and group 4 in Setting 7 resulted in a higher correlation in the group-wise shrinkage parameters, which reduced PRSbils's ability to differentiate between these groups with different heritability contribution. These results suggest that under the overlapping annotation scenario, choosing an annotation mapping which better separates the potential heritability-contributing sets may improve the prediction accuracy.

Biobank data analysis

We used the summary statistics as the training data to obtain the posterior genetic effect estimates, and evaluated the performance of both the proposed and existing PRS methods using the UK Biobank data (Fig. 3, Additional file 1: Figure S2), the MGI data (Fig. 4, Additional file 1: Figure S3) and the KoGES data (Additional file 1: Figure S4) as the test data. Each box plot contains the 10 results from the tenfold cross validation using the test data.

For the UK Biobank data, PRSbils with KEGG annotation outperformed PRS-CS for both type II diabetes and BMI. PRSbils yielded a 9.9% improvement in AUC over PRS-CS for type II diabetes on average, and a 1.5% average improvement for BMI. When gene-based annotations derived from ANNOVAR were used, *AUC* from PRSbils was on average 3.8% higher than PRS-CS for BMI, yet the predictive performance for type II diabetes was similar among the methods. For only one scenario with LDL trait using gene-based annotation, PRSbils yielded lower prediction performance than PRS-CS. We also note that the prediction performance varies across different

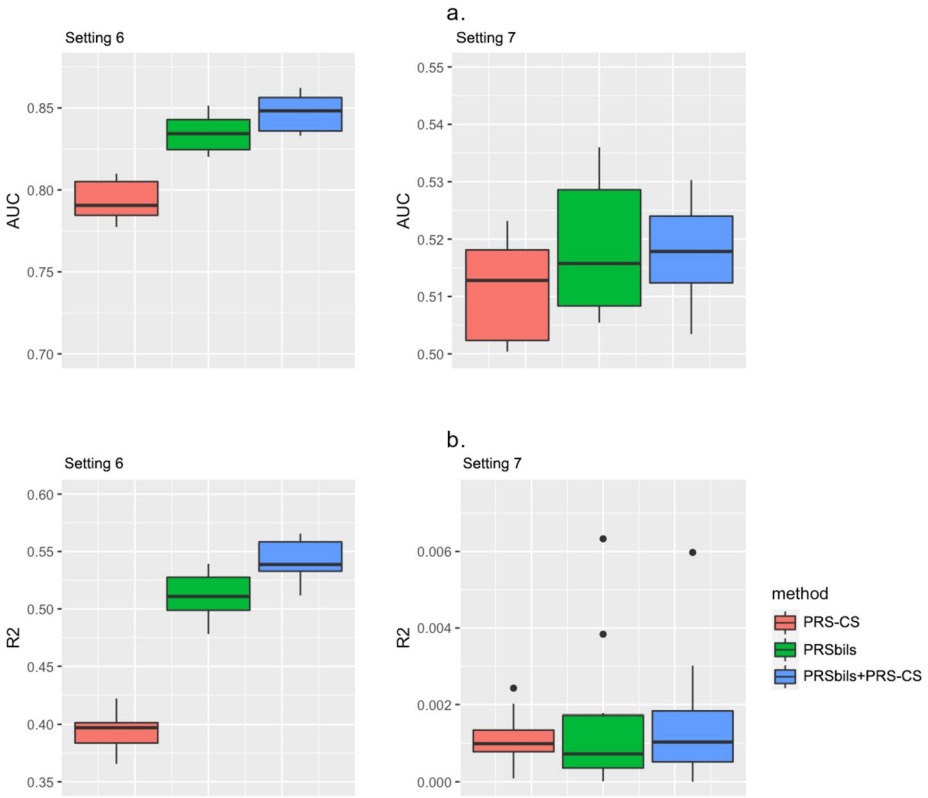

Fig. 2 Comparison of prediction performance in simulation studies with overlapping annotation groups (Settings 6–7), measured by AUC (**a**) and R^2 (**b**). Left: Setting 6 with overlapping pattern I, in which IOU is higher among the annotation groups with low heritability contribution. Right: Setting 7 with overlapping pattern II where IOU is higher among the annotation groups with high heritability contribution

populations for different annotation information. For the MGI and KoGES data, PRSbils did not yield better prediction performance than the benchmark method for the traits analyzed. One potential factor that can lead to this difference in performance is the cohort difference: The UKB is population-based and consists of UK participants mainly of European ancestry, while the MGI is patient-based, and the KoGES is an East Asian cohort. Different individual characteristics in these cohorts can have varying influence on the predictive performance. In contrast, the hybrid approach of PRS-CS and PRSbils performed robustly, showed high performance in most of the analysis.

It is likely that different annotation types vary in their contribution to the total heritability for diseases of interest, and as indicated by the simulation studies, such a difference affects the performance of the group-wise shrinkage parameter from the proposed method compared to the global shrinkage parameter in PRS-CS. We investigated the overall shrinkage of the two methods across the three biobank dataset for type II diabetes and presented the results in Additional file 1: Figure S5, which showed an overall consistency in the shrinkage between PRSbils and PRS-CS especially when the shrinkage value is relatively large, and PRSbils has a more variable pattern than PRS-CS when the shrinkage value is relatively small, indicating the differentiating effect of the group-wise shrinkage parameter.

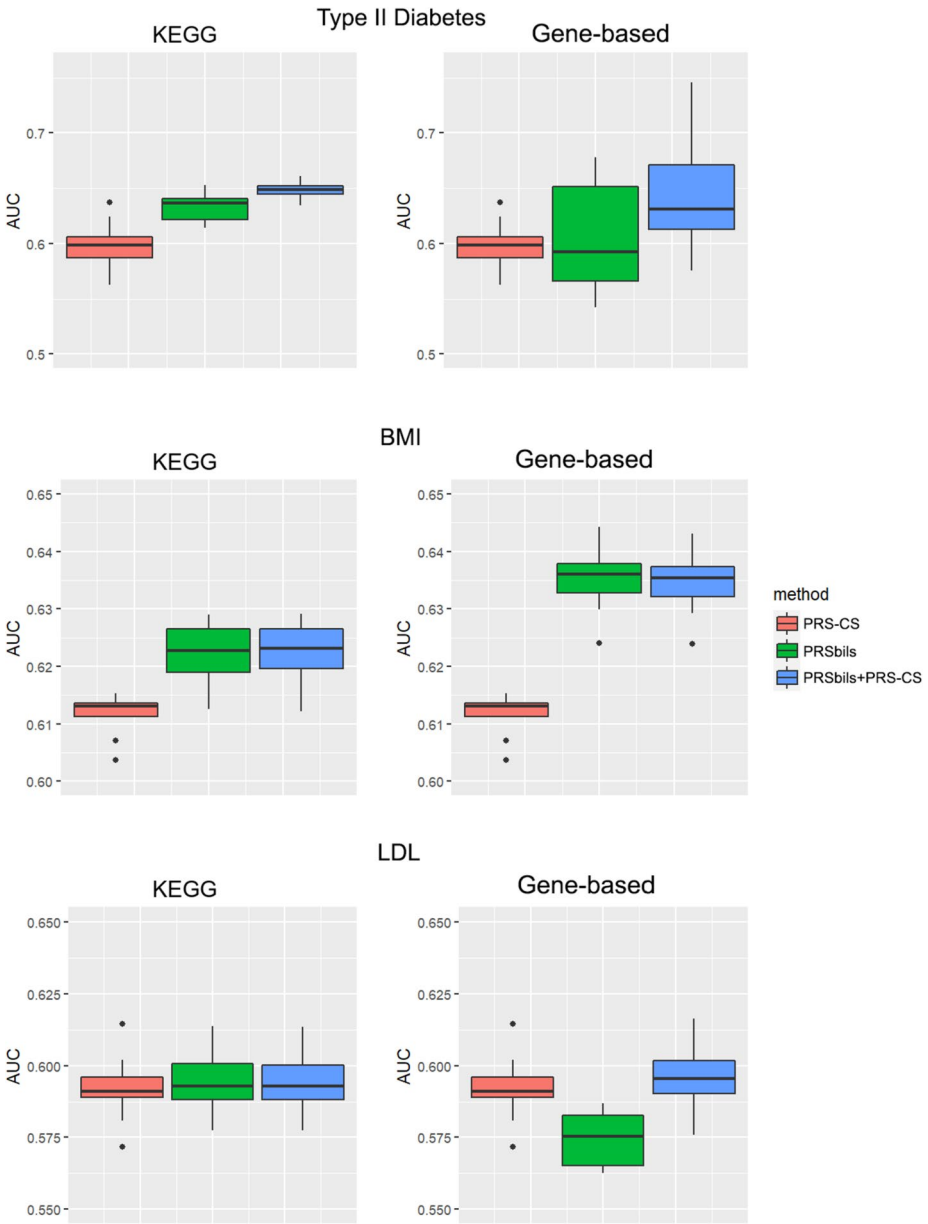

Fig. 3 Evaluation of AUC for UK Biobank analysis results. Left panel: KEGG functional annotations were used for the analysis of the proposed; Right panel: Refseq gene-based functional annotations from ANNOVAR were used for the analysis of PRSbils. From top to bottom: type II diabetes, BMI, LDL. For type II diabetes, summary statistics were obtained result from a GWAS analysis of 407,701 white British UK Biobank participants. For BMI, we used the GWAS results from GIANT Consortium with 332,153 participants with European ancestry. For LDL, GWAS results from the GLGC Consortium with 188,578 participants with European ancestry were used

We also performed an additional analysis to compare the performance with LDpred-funct [12]. Due to the lack of availability of the functional enrichment files for annotations other than those in the baselineLD model as required by LDpred-funct, we were only able to evaluate LDpred-funct's performance using the baselineLD model annotations, with type II diabetes as the phenotype. The predictive performance measured in AUC were similar across all the methods compared (Additional file 1: Figure S6)

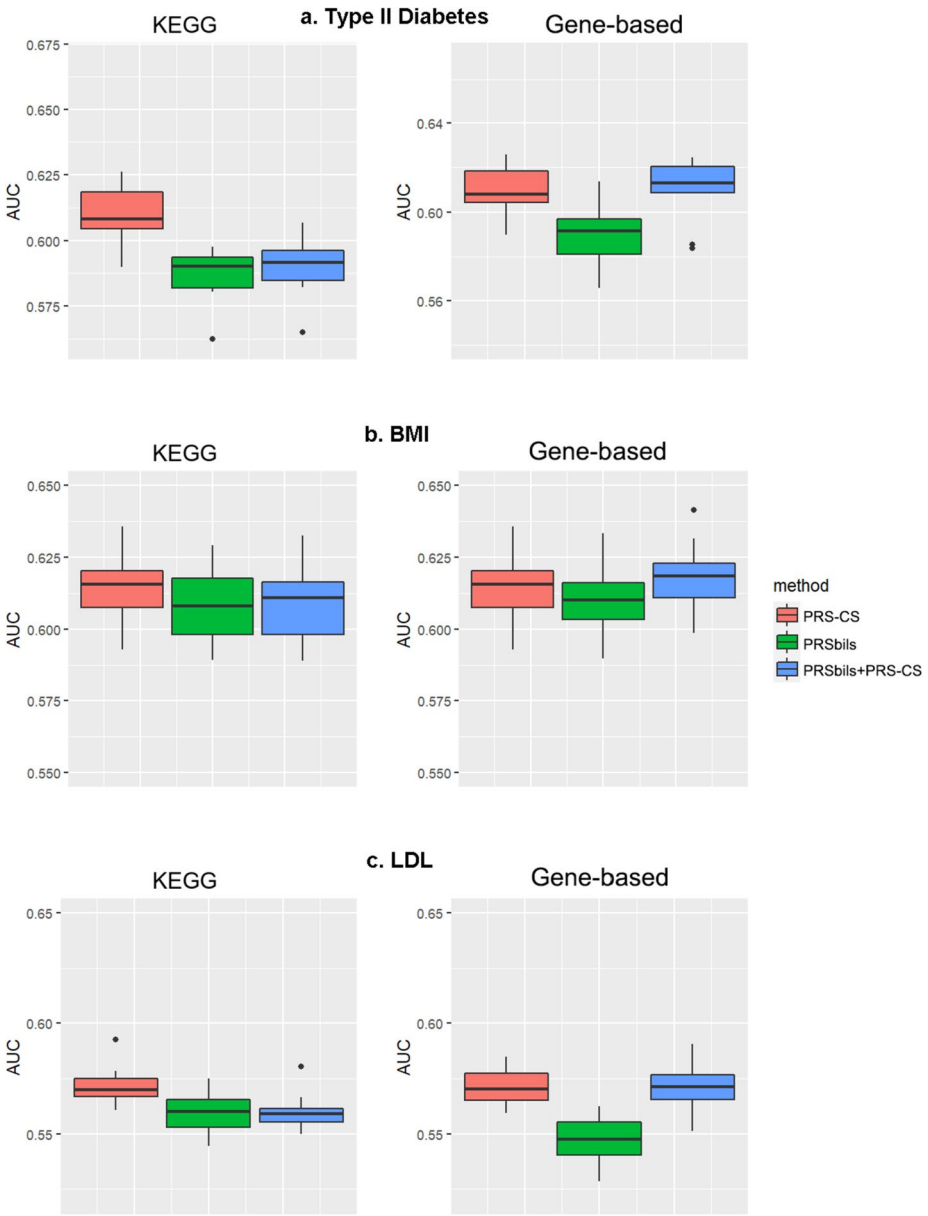

Fig. 4 Evaluation of AUC for the MGI data. Left panel: KEGG functional annotations were used for the analysis of the proposed; Right panel: Refseq gene-based functional annotations from ANNOVAR were used for the analysis of PRSbils. **a** type II diabetes; **b** BMI; **c** LDL. For type II diabetes, summary statistics were obtained result from a GWAS analysis of 407,701 white British UK Biobank participants. For BMI, we used the GWAS results from GIANT Consortium with 332,153 participants with European ancestry. For LDL, GWAS results from the GLGC Consortium with 188,578 participants with European ancestry were used

for the annotations in the baselineLD model, with LDpred-funct yielding a slightly higher performance. However, when compared to the performance using the KEGG annotation which PRSbils is able to incorporate, the performance of LDpred-funct using baselineLD model annotations had a lower performance than PRSbils and the hybrid method PRSbils + PRS-CS.

Computation time

Computation time was evaluated for the UK Biobank data analysis for Type II diabetes with KEGG annotation. PRSbils yielded a computation time of 14.0 CPU hours, compared to 12.9 CPU hours for PRS-CS. The slight increment of computation time for PRSbils is largely due to the additional computation on variants with overlapping annotation groups. All evaluations were computed on an Intel(R) Xeon(R) Gold 6242R CPU (Additional file 1: Supplementary Note 1).

Discussion

PRSbils can be applied to both non-overlapping and overlapping annotations. When the annotation categories overlap with each other (i.e., one SNP can belong to multiple annotation categories), the posterior effect size is calculated and incorporated into the PRS separately for each category a variant belongs to. The underlying assumption for this framework is that SNPs belonging to more annotation categories are prioritized for genetic risk calculation as they are more likely to be causal. It is similar to the idea of penalizing the SNPs with multiple annotations less than those with only one annotation category in a penalized regression framework [17]. As has been illustrated in the simulation studies, sparsity of the underlying heritability enrichment from each annotation group is a key factor for the predictive performance of PRSbils. When multiple groups are included in an annotation categorization, PRSbils is expected to yield a larger performance improvement than the methods not utilizing annotation group information if only a few groups contribute a relatively large proportion of the heritability. This is because the group shrinkage parameter from PRSbils is able to differentiate the degree of shrinkage across the annotation groups, instead of putting a uniform global shrinkage for all variants. Recent studies have shown that such group sparsity patterns are present in human traits. For example, pathway analysis of GWAS suggested that genetically associated variants are enriched in specific genes or pathways for traits such as diabetes, schizophrenia, and Alzheimer's disease [33–35]. We expect that identifying and applying these group-sparse annotations on a disease by disease basis would help further improve predictive performance.

In addition, the application of the proposed method using the KEGG annotations explores the pathway-level knowledge of polygenic risk for complex diseases, and provides an alternative way to stratify genetic liability in addition to the commonly used functional annotation, which adds to existing literature's ongoing investigation into the use of pathway PRS as a more informative way for patient stratification and treatment response prediction [36]. The average shrinkage for each KEGG annotation group can provide biological or clinical interpretations such as how different pathways weigh in terms of their relative importance for disease risk prediction. We illustrate this with the group-wise average shrinkage results from the UKB analysis (Additional file 1: Figure S7).

PRSbils uses a summary-statistics-based approach to obtain the posterior parameters via bilevel continuous shrinkage, and can be applied to both quantitative and binary traits. To construct the final PRS score, a separate validation dataset including individual-level genotypes and phenotypes is needed to regularize the group-wise weights. We currently adopt this validation data approach since it is challenging to model how each

annotation group contributes to the overall phenotype-specific risk. If the group-wise contribution can be captured accurately with prior knowledge, it is possible to apply PRSbils without a validation dataset, which remains a future step of the study.

The shrinkage parameters δ_k in PRSbils control the degree of shrinkage across annotation categories, and are automatically learned from the summary statistics in this study. An alternative way to specify the values for δ_k is to fix them using prior knowledge about the annotation-level sparseness of the genetic architecture. If the sample size of the training set to train α is small, we expect the latter approach to yield higher predictive performance, because the current fully Bayesian approach would generate less stable estimates for the shrinkage parameters under this situation. Indeed, with additional simulations we confirmed that the performance of the current PRSbils approach would have a higher variance when the sample size for the training set was small (Additional file 1: Figure S8).

We also explored the influence of annotation misclassification on prediction performance through additional simulation studies. Genotype and phenotype data were generated using the same settings as in Settings 1–5 (Table 1) with “true” corresponding annotation group assignment. Then, the variants were assigned a random "observed" annotation group with equal probability to train and test for prediction performance. The results showed that misclassification of annotation groups have negative impact on the predictive performance of PRSbils, especially when there is larger group sparsity in each annotation group’s contribution to heritability (Additional file 1: Figure S9). Thus, to achieve good performance, it is important to ensure that the group annotations well depicts the underlying heritability structure.

We note several points that can be further improved in future studies for PRSbils. Firstly, from the predictive performance of PRSbils evaluated in UK Biobank, MGI, and KoGES data, we noted that the results were not consistent across different data sources, which indicates the influence of cohort difference over predictive performance when annotation information is incorporated. It is thus critical to investigate the difference in the underlying architecture of the annotation groups across populations and make the method more robust for transethnic risk prediction. Secondly, although the posterior genetic effect estimates are shrunk towards zero by PRSbils, they are not exactly zero, which can have negative effects on the predictive accuracy. It remains for future work to make the posterior effect estimates sparser by selecting the groups of annotations to be in the final model. Thirdly, when the goal is to estimate the effect of polygenic risk score on quantitative phenotypes, additional information such as treatment effects can be included in the model to make the estimate more accurate. Fourthly, our simulation study assessed the method’s performance using a relatively high heritability value in order to differentiate between different group-wise heritability settings, and further exploration is needed to evaluate the method’s robustness under different overall heritability.

Despite the limitations, our study has the following strength: Firstly, the proposed PRSbils method utilized a novel bilevel continuous shrinkage framework to integrate functional annotation into the construction of the PRS score, and has shown an overall improvement in performance in simulation studies and biobank data analysis such as the UK Biobank. Secondly, we have showed that the hybrid method

PRSbils + PRS-CS consistently outperformed PRS-CS alone while extending it to incorporate functional annotations. Thirdly, compared to LDpred-funct, the existing method commonly used for integration of functional annotations in the BaselineLD model, PRSbils provides more flexibility in the choice of functional annotations such as KEGG and Refseq gene based annotations to support different research needs, and our study has shown that PRSbils yielded competitive performance to Ldpred-funct when using the annotations in the BaselineLD model while yielding higher performance when using other annotations such as the KEGG annotation. Fourthly, our simulation study shed more light on the less studied area of how the different distribution of group-wise heritability might affect the predictive performance of PRS scores that incorporate functional annotation groups. Finally, our study is one of the first to analyze PRS in a population-based cohort with European ancestry (UK biobank), a population-based cohort with Asian ancestry (KoGES), and a patient-based cohort (MGI) respectively, and sheds light on the potential performance difference due to different cohort characteristics.

Conclusion

We propose PRSbils, a PRS method which incorporates the functional annotation information and accounts for the annotation group-wise sparsity by applying a bilevel continuous shrinkage prior on the genetic effects. PRSbils uses summary statistics to get the posterior genetic effect estimates for each functional annotation group, estimates the combination weights of group-level PRS using a separate set of individual-level data, and generates the final score. We have shown in this study that leveraging the group-wise sparsity architecture of the genetic effects can help improve the performance of polygenic risk prediction. PRSbils is capable of utilizing a wide range of functional annotations, both overlapping and nonoverlapping, into the analysis and remains computationally efficient compared to the benchmark method. In summary, PRSbils enables the efficient and flexible use of different types of annotation information to improve PRS prediction.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05664-2.

Additional file 1. Supplementary Note 1.

Acknowledgements
Not applicable.

Author contributions
YZ, and SL designed the experiments. YZ and SL analyzed the UK Biobank and the MGI data. NK and SL analyzed the KoGES data. LF and BM provided inputs for the MGI data. YZ implemented the program. YZ wrote the manuscript with input and critical feedback from all authors.

Funding
This work was supported by the New Faculty Startup Fund from Seoul National University (N.Y.K. and S.L.) and NIH grants R01-LM012535 and R01-HG008773 (Y.Z.).

Availability of data and materials
The code generated during this study are available at https://github.com/styvon/PRSbils. UK Biobank data were accessed under the accession number UKB: 45227. KoGES data in this study were from the Korean Genome and Epidemiology Study (KoGES; 4851-302), National Research Institute of Health, Centers for Disease Control and Prevention, Ministry for Health and Welfare, Republic of Korea.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 31 March 2023 Accepted: 19 January 2024
Published online: 09 February 2024

References

1. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50:1219–24.
2. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017;18:117–27.
3. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007;17:1520–8.
4. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460:748–52.
5. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42:565–9.
6. Nolte IM, van der Most PJ, Alizadeh BZ, de Bakker PI, Boezen HM, Bruinenberg M, et al. Missing heritability: is the gap closing? An analysis of 32 complex traits in the lifelines cohort study. Eur J Hum Genet. 2017;25:877–85.
7. Young AI, Benonisdottir S, Przeworski M, Kong A. Deconstructing the sources of genotype-phenotype associations in humans. Science. 2019;365:1396–400.
8. Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput Biol. 2017;13:e1005589.
9. Marquez-Luna C, Gazal S, Loh P-R, Kim SS, Furlotte N, Auton A, et al. LDpred-funct: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. 2020. BioRxiv 375337.
10. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10:1–10.
11. Chun S, Imakaev M, Hui D, Patsopoulos NA, Neale BM, Kathiresan S, et al. Non-parametric polygenic risk prediction via partitioned gwas summary statistics. Am J Hum Genet. 2020;107:46–59.
12. Márquez-Luna C, Gazal S, Loh P-R, Kim SS, Furlotte N, Auton A, et al. Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. Nat Commun. 2021;12:1–11.
13. George EI, McCulloch RE. Variable selection via Gibbs sampling. J Am Stat Assoc. 1993;88:881–9.
14. Polson NG, Scott JG. Shrink globally, act locally: sparse Bayesian regularization and prediction. Bayesian Stat. 2010;9:105.
15. Xu Z, Schmidt DF, Makalic E, Qian G, Hopper JL. Bayesian sparse global-local shrinkage regression for selection of grouped variables. 2017. ArXiv Prepr ArXiv170904333.
16. Makalic E, Schmidt DF. A simple sampler for the horseshoe estimator. IEEE Signal Process Lett. 2015;23:179–82.
17. Chen T-H, Chatterjee N, Landi MT, Shi J. A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information. J Am Stat Assoc. 2020;116:1–11.
18. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.
19. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48:1279–83.
20. Fritsche LG, Gruber SB, Wu Z, Schmidt EM, Zawistowski M, Moser SE, et al. Association of polygenic risk scores for multiple cancers in a phenome-wide study: results from the Michigan genomics initiative. Am J Hum Genet. 2018;102:1048–61.
21. Kim Y, Han B-G, Group K. Cohort profile: the Korean genome and epidemiology study (KoGES) consortium. Int J Epidemiol. 2017;46:e20–e20.
22. Zawistowski M, Fritsche LG, Pandit A, Vanderwerff B, Patil S, Schmidt EM, et al. The Michigan genomics initiative: a biobank linking genotypes and electronic clinical records in Michigan medicine patients. Cell Genom. 2023;3:100257.
23. Nam K, Kim J, Lee S. Genome-wide study on 72,298 individuals in Korean biobank data for 76 traits. Cell Genom. 2022;2:10.
24. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
25. Bastarache L. Using phecodes for research with the electronic health record: from PheWAS to PheRS. Annu Rev Biomed Data Sci. 2021;4:1.
26. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50:1335–41.

27. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat Genet. 2018;50:390–400.
28. Turcot V, Lu Y, Highland HM, Schurmann C, Justice AE, Fine RS, et al. Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. Nat Genet. 2018;50:26–41.
29. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45:1274.
30. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
31. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164–e164.
32. Efron B. Regression and ANOVA with zero-one data: measures of residual variation. J Am Stat Assoc. 1978;73:113–21.
33. Mei H, Li L, Griswold M, Mosley T. Gene expression meta-analysis of seven candidate gene sets for diabetes traits following a GWAS pathway study. Front Genet. 2018;9:52.
34. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018;50:381–9.
35. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019;51:404–13.
36. Choi SW, Garcia-Gonzalez J, Ruan Y, Wu HM, Johnson J, Hoggart C, et al. PRSet: Pathway-based polygenic risk score analyses and software. 2023;19:e1010624.

Publisher's Note