MultiPopPred: A Trans-Ethnic Disease Risk Prediction Method, and its Application to the South Asian Population

Ritwiz Kamal[1,2,3] and Manikandan Narayanan[1,2,3,*]

[1]Department of Computer Science & Engineering, Indian Institute of Technology (IIT) Madras, Chennai, India.
[2]Centre for Integrative Biology and Systems Medicine, IIT Madras, Chennai, India.
[3]Wadhwani School of Data Science and AI, IIT Madras, Chennai, India.

*Corresponding author. E-mail: nmanik@cse.iitm.ac.in
Contributing author: ritwiz@cse.iitm.ac.in

Abstract

Genome-wide association studies (GWAS) aimed at estimating the disease risk of genetic factors have long been focusing on homogeneous Caucasian populations, at the expense of other understudied non-Caucasian populations. Therefore, active efforts are underway to understand the differences and commonalities in exhibited disease risk across different populations or ethnicities. There is, consequently, a pressing need for computational methods that efficiently exploit these population specific vs. shared aspects of the genotype-phenotype relation. We propose MultiPopPred, a novel trans-ethnic polygenic risk score (PRS) estimation method, that taps into the shared genetic risk across populations and transfers information learned from multiple well-studied auxiliary populations to a less-studied target population. MultiPopPred employs a specially designed Nesterov-smoothed penalized shrinkage model and a L-BFGS optimization routine. We present three variants of MultiPopPred based on the availability of individual-level vs. summary-level data and the weightage of each auxiliary population. Extensive comparative analyses performed on simulated genotype-phenotype data reveal that MultiPopPred improves PRS prediction in the South Asian population by 65% on settings with low target sample sizes and by 21% overall across all simulation settings, when compared to state-of-the-art trans-ethnic PRS estimation methods. This performance trend is promising and encourages application and further assessment of MultiPopPred under other simulation and real-world settings.

Keywords: Target Population, Auxiliary Population, Bayesian Linear Regression, Shrinkage, Optimization, Polygenic Risk Score (PRS), South Asians, Complex Traits, Genome Wide Association Studies (GWAS)

1 Introduction

Humans across the world are known to have differences in their genetic makeup alongside a number of various environmental factors; these differences in genetic and environmental factors can translate to different populations or ethnicities having different degrees of predispositions to complex diseases [1–3]. Complex (polygenic) diseases usually manifest as a consequence of a large number of common genetic factors (single nucleotide polymorphisms or SNPs) contributing in miniscule to moderate amounts simultaneously to the disease trait [4, 5]. Typical examples of such diseases include type 2 diabetes, cardiovascular diseases, and neurological disorders. For these diseases, studies have reported that people of South Asian ancestry exhibit different risk profiles than Caucasian individuals - for instance, South Asians exhibit a higher risk for cardiovascular diseases [6] and a higher prevalence of diabetes [7] as compared to Caucasians.

Traditionally, genome-wide association studies (GWAS), employing linear regression models, have served as a popular and established means to study associations between genetic factors or SNPs and traits of interest. Studies with sufficiently large sample sizes may be used to construct polygenic risk scores (PRS) that are single-valued estimates aggregating effects of SNPs across the human genome to give us insights into an individual's predisposition to develop a given trait [8]. While large GWAS (with $\geq 10^6$ individuals in the study) conducted on Caucasian (mainly European) populations have been significantly insightful [9, 10], there is a clear lack of representation of non-Caucasian populations (like South Asian) in the field. GWAS conducted on South Asian populations usually have around a few hundred to few thousand individuals only in the study [11–16], thereby rendering the study statistically underpowered to get estimates that are reliable enough to compute PRS worthy of downstream utility. Moreover, owing to differences in linkage disequilibrium (LD) patterns, allele frequencies and varying trait heritability across populations, estimates from European GWAS cannot simply be plugged in to a South Asian population [17, 18], without risking exacerbation in health disparities [19]. There are active efforts underway to address this under-representation of low-resource non-Caucasian populations like South Asian.

A seemingly straightforward solution to the above problem would be to collect more data on South Asian individuals before conducting GWAS, albeit at an unnecessarily higher cost of time, resources, and efforts. Recent efforts (such as PROSPER [20], PRS-CSx [21], and TL-Multi [22]) instead aim to understand and harness the differences and commonalities in per-SNP effect sizes (or risk towards a disease) between populations or ethnicities. Subsequently, they aim to utilize this information on population-specific vs. shared genetic disease risk to perform a transfer learning from well studied informative auxiliary populations (such as European) to a less studied target population (such

as South Asian). Kachuri et al. [23] present a comprehensive review of other similar published works in this domain. While each pre-existing method (including PROSPER [20], PRS-CSx [21], TL-Multi [22], Lassosum [24, 25]) adds an important new perspective to the problem at hand, they leave ample scope to integrate the information from multiple auxiliary populations in a manner that allows for accurate estimation of per-SNP effect sizes and consequently disease risk in the target population. As also observed before [20, 23], there continues to be a strong need for complementary trans-ethnic methods to estimate disease risk, because no single existing method unanimously outperforms all other methods in estimating the risk of diseases across diverse populations with different genetic architectures and sample sizes.

We propose MultiPopPred, a novel trans-ethnic polygenic risk score (PRS) estimation method, that simultaneously integrates information on per SNP effect sizes learned from multiple well-studied auxiliary populations (like Caucasian populations) and transfers the same to a less-studied target population. MultiPopPred employs a specially designed penalized L1 shrinkage model and a fast and efficient L-BFGS optimization routine for this purpose. We present three variants of MultiPopPred based on the availability of individual-level vs. summary-level data and the weightage of each auxiliary population. Extensive comparative analyses performed on simulated genotype-phenotype data reveal that MultiPopPred improves PRS prediction in the South Asian population by 65% on settings with low target sample sizes and by 21% overall across all simulation settings, when compared to state-of-the-art trans-ethnic PRS estimation methods. This performance trend is promising and encourages application and further assessment of MultiPopPred under other simulation and real-world settings.

2 Results

2.1 Overview of MultiPopPred

MultiPopPred builds upon the advancements made by previous trans-ethnic PRS estimation methods, while overcoming their shortcomings, to provide improved per SNP effect size estimates for low-resource populations. MultiPopPred (Figure 1) employs a L1-penalized lasso regression approach to close the gap between a target population's effect sizes and a set of aggregated effect sizes over multiple ($K \geq 1$) auxiliary populations. By doing so, our method capitalizes on the shared genetic makeup across populations and aims to perform transfer learning rather than learning weights from scratch when there is a dearth of trainable data for the target population. Depending upon the availability of the quantity and type of auxiliary and target data (i.e. individual level vs. summary level data) and the manner of quantification of similarity between a given pair of populations, we propose three variants of MultiPopPred, (i) MultiPopPred-Vanilla, (ii) MultiPopPred-Admixture, and (iii) MultiPopPred-ExtLD. MultiPopPred-Vanilla and MultiPopPred-Admixture require individual level target data for training along with summary level auxiliary data. MultiPopPred-ExtLD, on the other hand, mimics the more commonly occurring real-world scenario and operates only with the target and auxiliary summary statistics along with an external LD reference panel for the target population. MultiPopPred-Vanilla and MultiPopPred-ExtLD

both integrate information from multiple auxiliary populations by equally weighting each source of information while MultiPopPred-Admixture weighs each auxiliary population in proportion to their admixture component in target individuals (see Methods, Figure 1B). All three variants of MultiPopPred employ the L-BFGS optimizer for accurate and effective convergence to produce precise per-SNP β estimates.

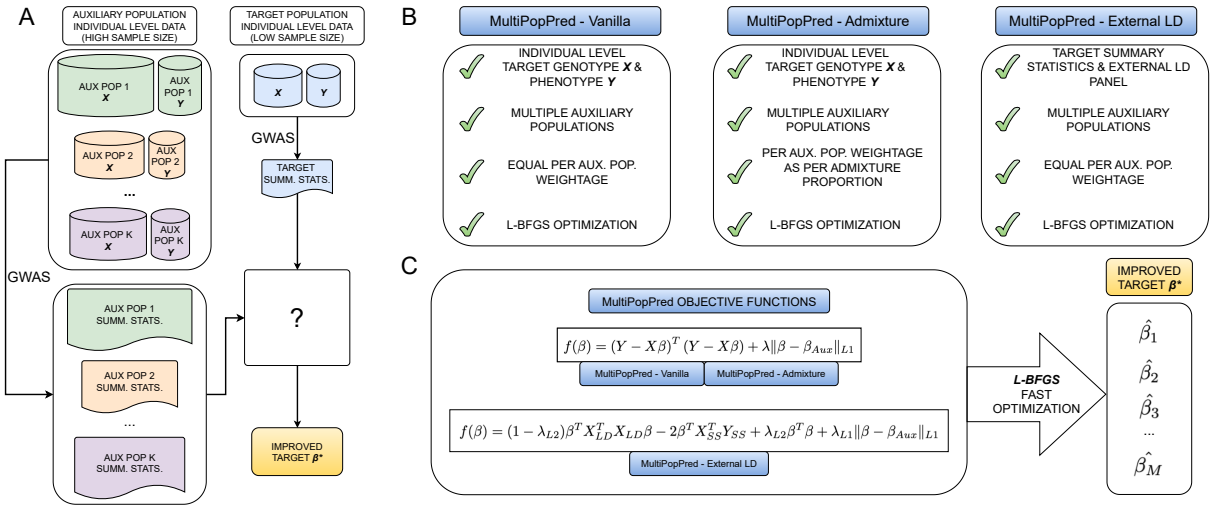

Fig. 1 Overview of MultiPopPred Methodology: (A) GWAS conducted on auxiliary and target populations of varying sample sizes are shown. Summary statistics of a population with larger sample sizes are statistically more reliable. MultiPopPred aims to transfer effect sizes learned from auxiliary populations to the target population to improve prediction accuracy. (B) Three variants of MultiPopPred based on the availability of individual-level vs. summary-level data and the weightage of each auxiliary population. (C) MultiPopPred objective functions employed by the different variants. L-BFGS optimization routine produces robust and reliable target effect size estimates. Abbreviations: Aux: Auxiliary Population, Tar: Target Population, Summ. Stats.: Summary Statitics.

2.2 Overview of Comparative Analyses

We have performed a comprehensive comparative analysis of MultiPopPred with several existing and established state-of-the-art (SOTA) PRS estimation methods that can be broadly categorized into three types depending on their input requirements. First, we have the single population based methods such as Baseline GWAS [26] and Lassosum2 [20, 24]. Next we have (one auxiliary, one target) population based methods such as TL-Multi [22]. Finally, we have (multiple auxiliary, one target) population based methods such as PRS-CSx [21] and PROSPER [20]. Brief descriptions of these methods and the parameters used in each case can be found in Methods. We assess the accuracy of our method's PRS estimates under several simulation settings, involving a range of sample sizes and heritability values (Section 2.3). Wherever applicable, like

in any traditional machine learning setup, the available data has been split into training (to get predictions), validation (to tune hyper-parameters) and testing (to report performance on) sets. Hyperparameter tuning results for the L1 and L2 penalizations in case of MultiPopPred can be found in Supplementary Table 1, Supplementary Fig. 1-2. We used 1000 Genomes Project (1KG) Phase 3 [27] defined super-populations (EUR, EAS, AMR, AFR, SAS) as reference panels wherever applicable (Methods).

2.3 Simulated Data Analysis

We first evaluated the performance of MultiPopPred and other SOTA methods on fully simulated genotype and phenotype data. Individual level genotypes for each population were simulated using HAPGEN2 [28] and the corresponding per population 1KG samples as reference (Methods). The ground truth per SNP effect sizes were generated using a K dimensional multivariate Gaussian distribution (K corresponding to the number of populations) with a fixed covariance matrix (Σ) (Methods). Subsequently, the phenotypes were generated in accordance with a linear additive model: $Y_{(n\times 1)} = X_{(n\times m)}\beta_{(m\times 1)} + \epsilon_{(n\times 1)}$ (Methods). We assumed each SNP to be having some contribution towards the phenotype, no matter how miniscule, so as to resemble the real world setting.

We designed three major analyses to understand the importance and optimality, w.r.t real-world resemblance, of three major parameters that controlled the different simulation settings viz., (i) heritability (h^2), (ii) number of target population samples, and (iii) the number of auxiliary population samples. Here, heritability (more precisely, narrow-sense heritability) h^2 refers to the proportion of the variation in Y explained by the genetic component alone i.e. $h^2 = \frac{\text{var}(X\beta)}{\text{var}(Y)}$. For each parameter configuration, the simulation was repeated 5 times and for each simulated setting MultiPopPred was run under 15 different configurations of auxiliary-target population tuples. Figure 2 depicts the comparative performances of the different PRS estimation methods for a subset of these configurations. Performances were evaluated on test datasets using the metric Correlation Ratio (see Methods). (We define Correlation Ratio as the correlation between the actual and predicted phenotypes, divided by the maximum achievable value of that correlation. Thus, a higher value of Correlation Ratio indicates better performance.) MultiPopPred was observed to report an overall 21% improvement in correlation ratio as compared to Baseline GWAS, PRS-CSx and PROSPER (averaged across the three MultiPopPred variants, the four simulated settings and the four auxiliary-target tuple configurations depicted in 2).

2.3.1 Increase in the number of auxiliary populations leads to better target estimates

Before elaborating on the specific observations related to each of the three analyses mentioned above, we would like to point out a general trend noticed across all the different simulation settings and experiments viz., an increase in the performance as the number of auxiliary populations used for transfer learning was increased from 1

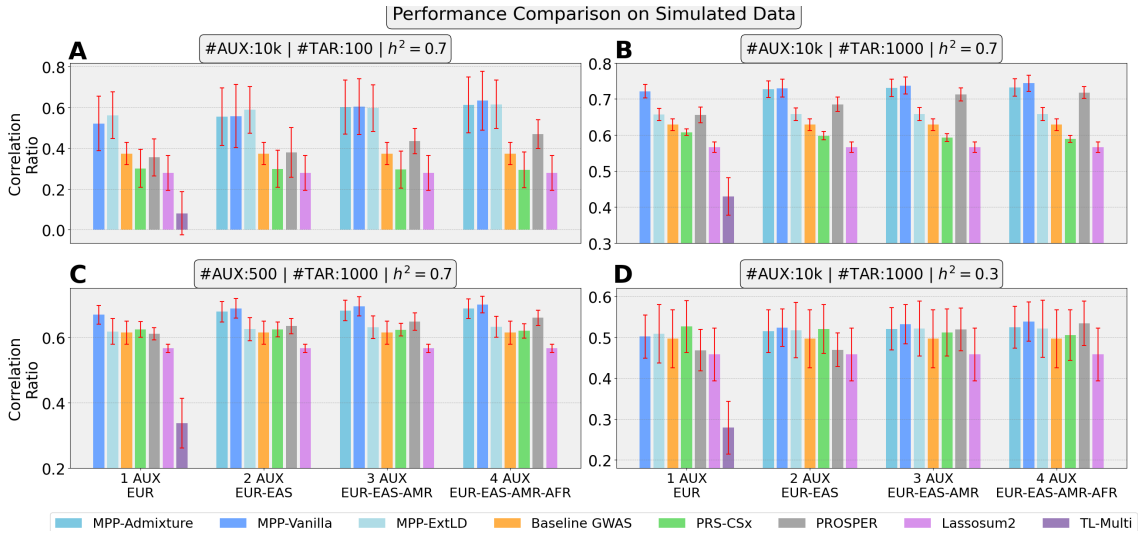

Fig. 2 Performance Comparison on Simulated Data (Phenotype Prediction): We depict the comparative performance of MultiPopPred with state-of-the-art (SOTA) methods under four different simulation settings and four auxiliary-target population tuples. (A) The number of samples per auxiliary population was set to 10,000 while the number of target population samples was kept very low i.e. 100 to represent an extreme condition. h^2 was set to 70% akin to heritability values exhibited by most complex traits. Subsequently, each of these simulation parameters were tweaked to observe trends in the comparative performance. (B) The number of auxiliary samples and h^2 remained the same as (A) but the number of target samples was increased to 1000. (C) The number of target samples and h^2 remained the same as (B) but the number of samples per auxiliary population was reduced to 500. (D) The number of auxiliary and target samples remained the same as (B) but h^2 was reduced to 30%. It must be noted that MultiPopPred-Admixture is not applicable under the single auxiliary population setting since it does not make sense to calculate admixture proportions for a single refernce population. On the other hand, TL-Multi is not applicable for > 1 auxiliary population settings. Abbreviations: Aux: Auxiliary Population, Tar: Target Population, h^2: Heritability.

to 4 in steps of 1. The increase was observed to be subtle but consistent. The subtlety of the increase can be attributed to the way simulations were designed such that the correlation between the ground truth per-SNP effect sizes for each pair of populations used in the study was fixed at a relatively high value of 0.8. This observation also validates our hypothesis that information from multiple auxiliary populations, if aggregated appropriately lends to efficient transfer learning that in turn translates to improved PRS estimates for a target population.

2.3.2 MultiPopPred performs on-par or better than SOTA methods at varying heritabilities

In our first major analysis, we fixed the number of auxiliary samples per population at 10,000 (mimicking the high resources available for well-studied populations), the number of target samples at 1000 (mimicking the typically available low resources for an understudied population) and varied the heritability (h^2) from 10% to 100% in steps of 20 ($h^2 = 10\%$ representing a phenotype having a much higher contribution

from the noise component and $h^2 = 100\%$ representing a phenotype having no contribution from the noise component). We observed higher variability in performance at very low h^2 values (2D), expectedly so, due to the true phenotype having very little contribution from the genetic component. As the h^2 value and consequently the contribution of the genetic component to the phenotype increased, each method was observed to report an improvement in performance both in terms of the observed Correlation Ratio and the stability of the same. Moreover, at low h^2 values (10%-30%) almost all methods were noted to have more or less comparable performances (2D) except for TL-Multi. At moderate h^2 (50%-70%) and higher h^2 (90%-100%) values, MultiPopPred (specifically the Vanilla and Admixture variants) was observed to have a clear edge over other methods (2A, 2B, 2C). This can be attributed to (i) better aggregation of effect sizes from the auxiliary populations before performing transfer of the learned information to the target population and (ii) use of the L-BFGS optimizer which seems to be more efficient at locating the optima than an optimization routine like coordinate descent which was employed by most of the other methods. It must be noted that MultiPopPred-ExtLD and PRS-CSx base their predictions on GWAS summary statistics alone as opposed to directly using genotypes and phenotypes. This explains the slight dip in performance noted for these two methods as compared to MultiPopPred-Vanilla, MultiPopPred-Admixture and PROSPER. Between MultiPopPred-ExtLD and PRS-CSx, the former was seen to consistently outperform the latter.

2.3.3 MultiPopPred outperforms SOTA methods even at low target sample sizes

In the second analysis, h^2 was fixed at 70% (mimicking the typically estimated heritability for known complex traits such as height, schizophrenia, etc. [29, 30]) along with 1000 target samples while the number of auxiliary samples per population was varied from 500 to 20,000. As expected, with an increase in the number of auxiliary samples the corresponding GWAS estimates became more statistically powerful, thereby resulting in an improved transfer learning to the target population. It was also observed that around 5000-10,000 auxiliary samples per population should suffice to do an efficient transfer learning operation, following which the performance saturates. 2B shows MultiPopPred demonstrating a 12%, 18% and 1.5% average improvement in correlation ratio over Baseline GWAS, PRS-CSx and PROSPER respectively. In the third analysis, h^2 was fixed at 70% along with 10,000 auxiliary samples per population while the number of target samples was varied from 100 to 2000. All three variants of MultiPopPred were observed to clearly outperform other methods at low target samples depicting its robustness (Fig. 2A). In the low target sample setting (2A), MultiPopPred demonstrated a 57%, 97% and 41% average improvement in correlation ratio in comparison to Baseline GWAS, PRS-CSx and PROSPER respectively. While in general the higher the number of target samples, the better it is for prediction, through this experiment we observed that MultiPopPred can be used reliably with a number of target samples even 10x to 100x times lesser than the auxiliary populations.

3 Discussion

We have proposed MultiPopPred, a novel trans-ethnic polygenic risk score estimation method that aims to harness the transferability of learned information from multiple well-studied auxiliary populations to an under-studied target population. This work presents three variants of MultiPopPred, viz. MultiPopPred-Vanilla, MultiPopPred-Admixture and MultiPopPred-ExtLD that cater to varying real-world scenarios determined by the availability of (i) individual level vs. summary statistics data, (ii) high resource vs. low resource data, and (iii) genetically similar vs different population tuples. All three variants of MultiPopPred utilize a L1 penalized shrinkage model that strives to close the gap between the target and carefully aggregated auxiliary effect sizes while minimizing a least squares based loss. MultiPopPred also employs a L-BFGS optimization routine for fast and efficient prediction of robust per SNP effect sizes, for a given target population, that has the potential to in turn translate to informed disease risk prediction.

We performed extensive comparative analysis to evaluate the performance of MultiPopPred against existing and established state-of-the-art (SOTA) methods under various fully simulated settings. Key findings from simulation studies show that MultiPopPred robustly outperforms other methods in extreme conditions, i.e. when there is a dearth of target samples that otherwise would have allowed enough statistical power to estimate effect sizes reliably. MultiPopPred's efficient performance can be quantified by a 21% overall improvement in the correlation ratio metric across all simulated settings and a 65% improvement in a low target sample setting. As a future work, we intend to subject MultiPopPred to evaluation under a series of secondary simulation settings such as (i) changing the covariance matrix used during simulations to account for poorly correlated pairs of populations, (ii) benchmarking against other existing simulated datasets such as HAPNEST [31] and Harvard Dataverse [20, 32], (iii) runtime comparisons and (iv) performance comparison w.r.t. per-SNP effect size predictions.

While the observed results are promising, further assessment of MultiPopPred on semi-simulated and real-world settings needs to be conducted. Furthermore, MultiPopPred has certain limitations as well. First, while the manner of aggregation of auxiliary effect sizes by MultiPopPred has been observed to be more efficient than the existing SOTA methods, it might not be ideal. For example, the Vanilla and Ext-LD variants which assume equal weightage for each auxiliary population. Furthermore, weighing auxiliary populations by admixture proportions also presents it's own challenges. The admixture proportions themselves are determined solely based on the genotypic information available. However, the same may not hold true for all phenotypes. Such phenotypic dependencies in admixture proportions is something that needs further investigation in the future. Some more challenging aspects of MultiPopPred that demand future investigation are (i) inclusion of rare or population specific variants, (ii) inclusion of non-linear effects (epistasis) of SNPs towards the phenotype, and (iii) inclusion of non-genetic, environmental (often unobserved confounders) while determining a target individual's disease risk. Careful consideration and redressal of such concerns could be valuable to refine MultiPopPred thereby enabling us to address disease risk prediction in understudied populations to the best extent possible.

4 Methods

This section provides a detailed account of the methodology proposed and employed in MultiPopPred and data simulations. The section concludes with a brief account of all state-of-the-art (SOTA) methods used for comparative analysis.

4.1 MultiPopPred

We propose MultiPopPred as a trans-ethnic polygenic risk score estimation method that aims to tackle the under representation of low resource and consequently, under studied populations (such as South Asian) for a better informed and inclusive approach to genomics based disease risk prediction. MultiPopPred aims to do so by carefully utilizing knowledge learned from multiple high resource and well studied populations (such as European, East Asian, etc.) simultaneously to improve accuracy of estimates. We call such high resource populations Auxiliary populations, while the low resource population is called the Target population. Before elaborating further on the exact mechanism employed by MultiPopPred to integrate information coming from the auxiliary populations, let us revisit the shortcomings of traditional GWAS based approaches for the concerned task.

4.1.1 Background, Preliminaries and Notations

Traditional GWAS rely solely on a simple linear regression approach (as depicted in Equation 1 below) to arrive at per-SNP effect sizes or βs. Here, X represents the $n \times m$ genotype matrix (where n is the number of individuals in the study and m is the number of SNPs), Y is the $n \times 1$ phenotype (or disease trait) vector, β is a $m \times 1$ vector containing per SNP effect sizes and ϵ is the $n \times 1$ noise vector. Traditionally, the genotype matrix X is normalized so that each column of the matrix has a zero mean and unit variance. Similarly, the phenotype vector Y is also normalized so that it has a zero mean and unit variance. In addition to the standard assumptions associated with a simple linear regression model [33], this approach assumes that (i) the SNPs have a linear and additive relationship with the phenotype, (ii) there are no interactions among the SNPs themselves, (iii) the noise term includes contributions from other non-additive and environmental factors and follows a normal distribution with zero mean and unit variance i.e. $\mathcal{N}(0, 1)$. Such an approach works well if we have a large number of samples (n) to conduct the study leading to reliable β estimates, as is the case with high-resource (large n) auxiliary populations. However, in case of low resource populations (small n, usually in the range of a few hundreds) such an approach fails to produce reliable estimates of β.

$$\begin{aligned} Y &= X\beta + \epsilon \\ \mathcal{F}(\beta) &= (Y - X\beta)^T (Y - X\beta) \end{aligned} \tag{1}$$

4.1.2 MultiPopPred Objective Function

An obvious solution to the above stated problem with low resource populations is to increase the number of samples used for the study by collecting more data. This, however, requires more time, cost and efforts. MultiPopPred, instead, taps into the potential hosted in the shared genetic makeup among individuals across multiple populations. The idea is to leverage the statistically reliable per-SNP effect sizes (βs) from high resource populations and perform a transfer of this learned information, to improve β estimates for the target population despite limited resources. Let us say, we have K high resource, well studied auxiliary populations with statistically significant β estimates ($\beta^{Aux1}, \beta^{Aux2}, ..., \beta^{AuxK}$), each having dimension $m \times 1$, from their respective GWAS. Additionally, we have the target population's GWAS and/or genotype ($X_{n \times m}$) and phenotype ($Y_{n \times 1}$) information. We propose to solve Equation 1 above through a L1 penalized regression approach (Equation 2) that strives to make the target β estimates as close as possible to the auxiliary population βs, while minimizing a least-squares based loss.

$$\mathcal{F}(\beta^{Tar}) = \left(Y^{Tar} - X^{Tar}\beta^{Tar}\right)^T \left(Y^{Tar} - X^{Tar}\beta^{Tar}\right) + \lambda \|\beta^{Tar} - \beta^{Aux}\|_{L1} \tag{2}$$

4.1.3 MultiPopPred Optimization

Going forward we will drop the superscript Tar from X^{Tar}, Y^{Tar} and β^{Tar} for the sake of brevity. Here, $\mathcal{F}$ denotes the objective function to be optimized, λ denotes the L1 shrinkage parameter, and β^{Aux} refers to an aggregate of ($\beta^{Aux1}, \beta^{Aux2}, ..., \beta^{AuxK}$). The way the aggregation of these betas is performed renders different variants of MultiPopPred, which have been elaborated on in the subsection 4.1.4. The use of $L1$ penalization here can be attributed to two major factors: (i) theoretically, efficient shrinkage that drives non-important features (SNPs) towards near-zero effect sizes, (ii) empirically, L1 penalization has been observed to demonstrate better performance on the concerned prediction task than its popular counterpart L2 penalization, which has been employed by SOTA such as PRS-CSx (Results). For optimization of our objective function (Equation 2), we experimented with several different methods and decided to go ahead with the Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method [34]. Although L-BFGS is suited for optimizing smooth, differentiable functions, studies have highlighted its applicability and efficiency even in case of some non-smooth functions and some smooth approximations of non-smooth functions [35]. In our case, we employ Nesterov smoothing (smoothing parameter μ fixed at 0.1) [36] with an entropy prox function $f^{\mu}(x)$ (Equation 3) [37] to produce a smoothed approximation of the objective function as depicted in Equation 4 below. We supply the first order derivative or Jacobian of Equation 4 to the L-BFGS method for faster convergence, whenever smoothing is used.

$$f^{\mu}(x) = \mu \; log\left(\frac{1}{2}e^{-x/\mu} + \frac{1}{2}e^{x/\mu}\right) \tag{3}$$

$$\mathcal{F}(\beta) = (Y - X\beta)^T (Y - X\beta) + \lambda \sum_{j=1}^{m} f^{\mu}(\beta_j - \beta_j^{Aux}) \tag{4}$$

4.1.4 MultiPopPred Variants

MultiPopPred-Vanilla

This variant of MultiPopPred requires the target population's genotype (X) and phenotype (Y) along with each auxiliary population's GWAS summary statistics as inputs. In this variant we give equal weightage to each auxiliary population. In other words, if there are K auxiliary populations each having their own set of βs (each of size $m \times 1$) viz., $\beta^{Aux1}, \beta^{Aux2}, ..., \beta^{AuxK}$ from the summary staistics, then the aggregated β^{Aux} is defined as per Equation 5 below.

$$\beta^{Aux} = \frac{1}{K}\ \beta^{Aux1} + \frac{1}{K}\ \beta^{Aux2} + ... + \frac{1}{K}\ \beta^{AuxK} \tag{5}$$

If Nesterov smoothing is used along with this variant, the Jacobian or gradient vector $g^{\mu}(x)$ (again of size $m \times 1$) supplied to the L-BFGS optimizer (ignoring the constant terms) can be derived to get Equations 6 and 7 below.

$$\frac{\partial \mathcal{F}(\beta)}{\partial \beta_j} = 2X^T X\beta - 2X^T Y + \lambda\ g^{\mu}(\beta_j - \beta_j^{Aux}) \tag{6}$$

$$g^{\mu}(x) = \frac{\partial f^{\mu}(x)}{\partial x} = \frac{-e^{-x/\mu} + e^{x/\mu}}{e^{-x/\mu} + e^{x/\mu}} \tag{7}$$

MultiPopPred-Admixture

Like the vanilla variant, this variant of MultiPopPred also requires the target X and Y along with auxiliary GWAS summary statistics as inputs. In this variant, however, the auxiliary populations are given weightage in accordance to their admixture proportions in the target population individuals. Suppose there are K auxiliary populations and n individuals in the target population. Then for each individual, we compute the admixture proportions of the K auxiliary populations using an appropriate software such as ADMIXTURE [38]. (We employ the ADMIXTURE tool in a supervised manner with quasi-Newton convergence acceleration [39].)

$$\beta^{(i)}_{AuxAdmx} = w^{(i)}_{Aux1}\ \beta^{Aux1} + w^{(i)}_{Aux2}\ \beta^{Aux2} + ... + w^{(i)}_{AuxK}\ \beta^{AuxK} \tag{8}$$

Here, $w^{(i)}_{Auxk}$ corresponds to the admixture proportion of auxiliary population k in the ith target individual such that $\sum_{k=1}^{K} w^{(i)}_{Auxk} = 1$. Furthermore, $\beta^{Aux1}, \beta^{Aux2}, ..., \beta^{AuxK}$ (each $m \times 1$) refer to each of the K auxiliary populations' βs from their respective summary statistics. $\beta^{(i)}_{AuxAdmx}$ $(m \times 1)$ refers to the admixed β for i^{th} target individual. The aggregated β^{Aux} is computed by taking the median across all n individuals in the target population as depicted in Equations 9 and 10 below.

$$\beta^{(i)}_{AuxAdmx} = \left[\beta^{(i)}_{SNP1}\ \beta^{(i)}_{SNP2}\ \beta^{(i)}_{SNP3}\ ...\ \beta^{(i)}_{SNPj}\ ...\ \beta^{(i)}_{SNPm}\ \right]^T \tag{9}$$

$$\beta_{Aux} = \left[\beta'_{SNP1}\ \beta'_{SNP2}\ \beta'_{SNP3}\ ...\ \beta'_{SNPj}\ ...\ \beta'_{SNPm}\right]^T \quad where, \\ \beta'_{SNPj} = median(\beta^{(1)}_{SNPj}, \beta^{(2)}_{SNPj}, ..., \beta^{(i)}_{SNPj}, ..., \beta^{(n)}_{SNPj}) \tag{10}$$

MultiPopPred-ExtLD

This variant of MultiPopPred is designed to take only the target and auxiliary populations' summary statistics as input along with an external target LD reference panel that may be derived from 1000 Genomes [27] (or any other appropriate resources). The intuition behind this variant is to cater to the commonly encountered situation of non-availability of individual level genotype and phenotype data for the target population due to scarcity, privacy, or ethical constraints. In such scenarios, MultiPopPred-ExtLD offers an efficient and fast (owing to Nesterov smoothing) alternative without a considerable loss in performance. Akin to the vanilla variant, the aggregated β^{Aux} is computed as per Equation 5 giving equal weightage to each auxiliary population.

$$\mathcal{F}(\beta) = Y^TY + \beta^TX^TX\beta - 2\beta^TX^TY + \lambda\sum_{j=1}^{m} f^{\mu}(\beta_j - \beta_j^{Aux}) \tag{11}$$

While this variant also aims to optimize Equation 4 rewritten as Equation 11, an issue encountered is the source of X in X^TX and X^TY being different (as pointed out by Mak et al. [24]). Since we do not have access to the actual target X and Y, X^TX is substistuted by $X_{LD}^TX_{LD}$ computed using an external LD panel (1000 Genomes) while X^TY is substituted by the SNP-wise correlation computed from the target population's summary statistics as $r = (n_{Tar}\beta^{Tar})_{m\times 1}$ (n_{Tar} referring to the number of samples used for target GWAS). We address this issue in a fashion similar to Mak et al. [24], by replacing $X_{LD}^TX_{LD}$ with $(1-\lambda_{L2})X_{LD}^TX_{LD} + \lambda_{L2}I$ where I is an identity matrix and $0 < \lambda_{L2} < 1$. Ignoring the constant terms, Equation 11 can thus be rewritten as Equation 12 transforming our lasso optimization problem into an elastic-net one. The Jacobian supplied to L-BFGS also gets transformed as per Equation 13.

$$\mathcal{F}(\beta) = (1-\lambda_{L2})\beta^TX_{LD}^TX_{LD}\beta - 2\beta^Tr + \lambda_{L2}\beta^T\beta + \lambda_{L1}\sum_{j=1}^{m} f^{\mu}(\beta_j - \beta_j^{Aux}) \tag{12}$$

$$\frac{\partial\mathcal{F}(\beta)}{\partial\beta_j} = 2(1-\lambda_{L2})X_{LD}^TX_{LD}\beta - 2r + 2\lambda_{L2}\beta + \lambda_{L1}\ g^{\mu}(\beta_j - \beta_j^{Aux}) \tag{13}$$

4.2 Data Simulation

Genotypes: We used HAPGEN2 [28] to simulate individual level genotypes for five different populations viz., South Asian (SAS), European (EUR), East Asian (EAS), American (AMR), and African (AFR). HAPGEN2 requires at minimum the following resources for simulations: (i) known haplotypes encoded as 0/1 (with one row per SNP and one column per haplotype) for a given reference population for resampling from, (ii) a legend file containing per SNP rs IDs, base pair positions, alternate and reference alleles, and (iii) a map file detailing fine-scale recombination rates across the given region. The known haplotypes for each of the five super-populations were obtained from 1000 Genomes [27, 40, 41] with SAS containing GIH samples, EUR containing CEU, FIN, GBR, IBS, TSI samples, EAS containing CDX, CHB, CHS, JPT, KHV samples, AMR containing CLM, MXL, PEL, PUR samples, and AFR containing ACB, ASW, LWK, YRI samples grouped together respectively. The selection of these specific sub-populations can be attributed to the availability of their respective genetic maps and recombination rates (Data Availability). For a given super-population, the genetic maps and recombination rates were linearly interpolated at each position (Code Availability) followed by averaging across the sub-populations and then proceeding with the final simulations. On consultation with the authors of PRS-CSx [21], who followed a similar workflow for data simulation, we noted that simulation of sub-population level genotype data followed by merging them into a super-population cohort would be an equally valid approach. Although we simulated genotype data for multiple chromosomes (1-22 for SAS and 19-22 for EUR, EAS, AMR, AFR, we have restricted all analysis on simulations (Results) to only chromosome 22 in favour of consistency and computational efficiency. In total we have simulated 40K samples for SAS and 50K samples each for EUR, EAS, AMR and AFR for chromosome 22 alone. The set of SNPs used for simulation were restricted to the 1,297,432 HapMap3 variants [42] which were further subjected to a series of filters viz., (i) removal of triallelic and strand ambiguous variants and (ii) retention of SNPs having a MAF $> 1\%$. Finally, a custom filter was applied (for computational efficiency) to select SNPs known to have association with diseases such as Type 2 Diabetes and Coronary Artery Disease [9, 10, 15, 16]. This resulted in a final SNPs tally of 8215 in SAS, 9887 in EUR, 9353 in EAS, 9873 in AMR and 9731 in AFR for chromosome 22.

Phenotypes: For simulating phenotypes we followed the linear model depicted in Equation 1. We sampled per SNP effect sizes or βs for each population from a 5-dimensional (corresponding to the 5 super-populations) multivariate Gaussian distribution having mean zero and where the covariance matrix Σ was designed to model the correlation or covariance (ρ) between each population's effect sizes. For all major experiments ρ was set to 0.8 for each pair of populations, unless otherwise specified. In other words, $\Sigma = 0.8$ if $i \neq j$ and $\Sigma = 1$ if $i = j$, where i, j refers to auxiliary populations i and j respectively. The noise component of the phenotype was sampled from a zero mean and unit variance standard univariate gaussian distribution. Finally, the contribution of the genetic components (SNPs) and the noise component (ϵ, including non-additive and environmental factors) towards the phenotype was controlled using

a heritability parameter $h^2 \in [0, 1]$, as depicted in Equation 14. Here X corresponds to a $(n \times m)$ normalized genotype matrix ($\bar{X}_j = 0$, $\sigma^2(X_j) = 1 \ \forall j \in (1, m)$), β is a $(m \times 1)$ effect size vector, $\epsilon \sim \mathcal{N}(0, 1)$ is the $(n \times 1)$ noise vector and $\sigma(.)$ refers to standard deviation. A higher value of heritability indicates greater contribution of the genetic components towards the phenotype and vice-versa. GWAS was performed on the simulated data using `big_univLinReg` (`bigstatsr`) in R [43]. All scripts to perform data simulation as per the above description are available on our Github Code Availability).

$$Y \quad = \quad \frac{X\beta}{\sigma(X\beta)} * \sqrt{h^2} \quad + \quad \epsilon * \sqrt{1 - h^2} \tag{14}$$

Evaluation Metric (Phenotype Prediction): For comparing performances w.r.t phenotype prediction we define Correlation Ratio as $\frac{R(Y_{Truth}, \hat{Y})}{R(Y_{Truth}, Y_{Truth} - \epsilon)}$, the ratio of the correlation between the true (ground-truth) and predicted phenotypes with the correlation between the true phenotype and the predicted phenotype if we somehow had access to the ground truth effect sizes (β_{Truth}). In other words, $Y_{Truth} - \epsilon = X\beta_{Truth}$. Correlation Ratio gives us the ratio of how well a particular method is doing versus the best that could be done w.r.t phenotype prediction. Ideally, we would want the Correlation Ratio to be as close to 1 as possible.

4.3 Related Works: Other State-of-the-art Methods

The various state-of-the-art (SOTA) methods that were used for performing comparative analysis with MultiPopPred can be categorised broadly into three sets based on whether and how many auxiliary population they make use of to estimate target effect sizes: (i) multiple auxiliary population based methods viz., PROSPER, PRS-CSx, (ii) single auxiliary population based methods viz., TL-Multi and (iii) solely target population based methods viz., Lassosum2, Baseline GWAS. All the above mentioned methods were used either at their default hyperparameter values or were allowed to learn their own hyperparameter values from the input data, if applicable.

PROSPER: Polygenic Risk scOres based on enSemble of PEnalized Regression models was proposed by Zhang et al. in 2024 [20] as a PRS computation method that makes use of multiple populations to improve predictions on an understudied target population. The workflow can be broken down into three major steps. First, independent per population GWAS analysis is performed via Lassosum2 [20, 24] using training and validation datasets. The next step involves modelling a joint cross-population penalized regression that makes use of the GWAS and optimal parameters computed in the previous step. Finally, a super learning model is trained to compute an ensemble of PRS computed in the second step. PROSPER has been known to be on par with or outperform all previously established PRS estimation methods in the domain, to the best of our knowledge. Hence, any novel approach being on par with or outperforming PROSPER can be considered to have similar trends for such previously existing methods.

PRS-CSx: Polygenic Risk Scores - Continuous Shrinkage was proposed by Ruan et al. in 2022 [21] as an extension to their previously published PRS-CS [44] single-ancestry PRS computation method. PRS-CSx makes use of global and local shrinkage parameters with only GWAS summary statistics from multiple populations and external LD refernce panels per population to estimate per population effect sizes via a Markov Chain Monte Carlo (MCMC) approximation strategy. PRS-CSx can be said to be following an approach, in theory, similar to a L2 penalized regression model. It has served as a gold-standard benchmark for comparing and contrasting newly developed PRS estimation methods over the years.

TL-Multi: Transfer Learning - Multi was proposed by Tian et al. in 2022 [22] as PRS estimation method that makes use of L1 penalization to minimize the difference between a single auxiliary population and a target population's effect sizes. Similar to PROSPER, TL-Multi also makes use of Lassosum (Mak et al. [24]) as a subroutine to compute per population GWAS followed by another instance of Lassosum to transfer learned effect sizes from the auxiliary to the target population. TL-Multi has also served as the basis for developing our MultiPopPred method.

Lassosum2: Lassosum was proposed originally by Mak et al. in 2017 [24] as a single-ancestry PRS estimation method that utilizes L1 penalization, as suggested by the name. We make use of Lassosum2 which was implemented by Zhang et al. alongside PROSPER to facilitate our comparative analyses.

Baseline GWAS: This method refers to a simple GWAS performed on the target population training data without any additional information from other auxiliary populations. Like in any traditional GWAS, this method employs a simple linear regression based approach without any penalization. We include this method in our comparative analysis to observe the relative improvements in performance that come with the inclusion of informative auxiliary populations.

4.4 Data Availability

The data used in this study have been referred to appropriately. Data generated in this study will be made available upon reasonable request to the authors.

4.5 Code Availability

The code used in this study will be made available via the following Github link: https://github.com/BIRDSgroup/MultiPopPred .

4.6 Acknowledgments

We thank members of our BIRDS research group for their valuable inputs during the course of this work. This work was supported by the Prime Minister's Research Fellowship (PMRF) Grant SB21221945CSPMRF008892 awarded to RK and Wellcome

Trust/DBT India Alliance Intermediate Fellowship Grant IA/I/17/2/503323 awarded to MN.

References

[1] Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., *et al.*: Insights into human genetic variation and population history from 929 diverse genomes. Science **367**(6484), 5012 (2020)

[2] Benton, M.L., Abraham, A., LaBella, A.L., Abbot, P., Rokas, A., Capra, J.A.: The influence of evolutionary history on human health and disease. Nature Reviews Genetics **22**(5), 269–283 (2021)

[3] Hou, K., Ding, Y., Xu, Z., Wu, Y., Bhattacharya, A., Mester, R., Belbin, G.M., Buyske, S., Conti, D.V., Darst, B.F., *et al.*: Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nature Genetics **55**(4), 549–558 (2023)

[4] Zwir, I., Arnedo, J., Del-Val, C., Pulkki-Råback, L., Konte, B., Yang, S.S., Romero-Zaliz, R., Hintsanen, M., Cloninger, K.M., Garcia, D., *et al.*: Uncovering the complex genetics of human character. Molecular Psychiatry **25**(10), 2295–2312 (2020)

[5] Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Psychiatric Genomics Consortium, S.W.G., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.: Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics **47**(3), 291–295 (2015)

[6] Patel, A.P., Wang, M., Kartoun, U., Ng, K., Khera, A.V.: Quantifying and understanding the higher risk of atherosclerotic cardiovascular disease among south asian individuals: results from the uk biobank prospective cohort study. Circulation **144**(6), 410–422 (2021)

[7] Narayan, K.V., Kondal, D., Daya, N., Gujral, U.P., Mohan, D., Patel, S.A., Shivashankar, R., Anjana, R.M., Staimez, L.R., Ali, M.K., *et al.*: Incidence and pathophysiology of diabetes in south asian adults living in india and pakistan compared with us blacks and whites. BMJ Open Diabetes Research and Care **9**(1), 001927 (2021)

[8] Lambert, S.A., Abraham, G., Inouye, M.: Towards clinical utility of polygenic risk scores. Human Molecular Genetics **28**(R2), 133–142 (2019)

[9] Mahajan, A., Spracklen, C.N., Zhang, W., Ng, M.C., Petty, L.E., Kitajima, H., Yu, G.Z., Rüeger, S., Speidel, L., Kim, Y.J., *et al.*: Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. Nature Genetics **54**(5), 560–572 (2022)

[10] Aragam, K.G., Jiang, T., Goel, A., Kanoni, S., Wolford, B.N., Atri, D.S., Weeks, E.M., Wang, M., Hindy, G., Zhou, W., *et al.*: Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. Nature Genetics **54**(12), 1803–1815 (2022)

[11] Zhao, W., Smith, J.A., Wang, Y.Z., Chintalapati, M., Ammous, F., Yu, M., Moorjani, P., Ganna, A., Gross, A., Dey, S., *et al.*: Polygenic risk scores for alzheimer's disease and general cognitive function are associated with measures of cognition in older south asians. The Journals of Gerontology: Series A **78**(5), 743–752 (2023)

[12] Harshfield, E.L., Fauman, E.B., Stacey, D., Paul, D.S., Ziemek, D., Ong, R.M., Danesh, J., Butterworth, A.S., Rasheed, A., Sattar, T., *et al.*: Genome-wide analysis of blood lipid metabolites in over 5000 south asians reveals biological insights at cardiometabolic disease loci. BMC Medicine **19**, 1–17 (2021)

[13] Yajnik, C.S., Wagh, R., Kunte, P., Asplund, O., Ahlqvist, E., Bhat, D., Shukla, S.R., Prasad, R.B.: Polygenic scores of diabetes-related traits in subgroups of type 2 diabetes in india: a cohort study. The Lancet Regional Health-Southeast Asia **14** (2023)

[14] Meng, X., Navoly, G., Giannakopoulou, O., Levey, D.F., Koller, D., Pathak, G.A., Koen, N., Lin, K., Adams, M.J., Rentería, M.E., *et al.*: Multi-ancestry genome-wide association study of major depression aids locus discovery, fine mapping, gene prioritization and causal inference. Nature Genetics **56**(2), 222–233 (2024)

[15] Loh, M., Zhang, W., Ng, H.K., Schmid, K., Lamri, A., Tong, L., Ahmad, M., Lee, J.-J., Ng, M.C., Petty, L.E., *et al.*: Identification of genetic effects underlying type 2 diabetes in south asian and european populations. Communications Biology **5**(1), 329 (2022)

[16] Huang, Q.Q., Sallah, N., Dunca, D., Trivedi, B., Hunt, K.A., Hodgson, S., Lambert, S.A., Arciero, E., Wright, J., Griffiths, C., *et al.*: Transferability of genetic loci and polygenic scores for cardiometabolic traits in british pakistani and bangladeshi individuals. Nature Communications **13**(1), 4664 (2022)

[17] Li, Y.R., Keating, B.J.: Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Medicine **6**, 1–14 (2014)

[18] Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., Kenny, E.E.: Human demographic history impacts genetic risk prediction across diverse populations. The American Journal of Human Genetics **100**(4), 635–649 (2017)

[19] Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., Daly, M.J.: Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics **51**(4), 584–591 (2019)

[20] Zhang, J., Zhan, J., Jin, J., Ma, C., Zhao, R., O'Connell, J., Jiang, Y., Research Team, Koelsch, B.L., Zhang, H., *et al.*: An ensemble penalized regression method for multi-ancestry polygenic risk prediction. Nature Communications **15**(1), 3238 (2024)

[21] Ruan, Y., Lin, Y.-F., Feng, Y.-C.A., Chen, C.-Y., Lam, M., Guo, Z., He, L., Sawa, A., Martin, A.R., *et al.*: Improving polygenic prediction in ancestrally diverse populations. Nature Genetics **54**(5), 573–580 (2022)

[22] Tian, P., Chan, T.H., Wang, Y.-F., Yang, W., Yin, G., Zhang, Y.D.: Multiethnic polygenic risk prediction in diverse populations through transfer learning. Frontiers in Genetics **13**, 906965 (2022)

[23] Kachuri, L., Chatterjee, N., Hirbo, J., Schaid, D.J., Martin, I., Kullo, I.J., Kenny, E.E., Pasaniuc, B., Diverse Populations (PRIMED) Consortium Methods Working Group Auer Paul L. 20 Conomos Matthew P. 21 Conti David V. 22 23 Ding Yi 24 Wang Ying 19 25 26 Zhang Haoyu 27 28 Zhang Yuji 29, P.R.M., Witte, J.S., *et al.*: Principles and methods for transferring polygenic risk scores across global populations. Nature Reviews Genetics **25**(1), 8–25 (2024)

[24] Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., Sham, P.C.: Polygenic scores via penalized regression on summary statistics. Genetic Epidemiology **41**(6), 469–480 (2017)

[25] Hahn, G., Prokopenko, D., Lutz, S.M., Mullin, K., Tanzi, R.E., Cho, M.H., Silverman, E.K., Lange, C., Precision Medicine (TOPMed) Consortium: A smoothed version of the lassosum penalty for fitting integrated risk models using summary statistics or individual-level data. Genes **13**(1), 112 (2022)

[26] Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M.P., Foulkes, A.S.: A guide to genome-wide association analysis and post-analytic interrogation. Statistics in Medicine **34**(28), 3769–3792 (2015)

[27] Consortium, .G.P., *et al.*: A global reference for human genetic variation. Nature **526**(7571), 68 (2015)

[28] Su, Z., Marchini, J., Donnelly, P.: Hapgen2: simulation of multiple disease snps. Bioinformatics **27**(16), 2304–2305 (2011)

[29] Rotwein, P.: Revisiting the population genetics of human height. Journal of the Endocrine Society **4**(4), 025 (2020)

[30] Owen, M.J., Legge, S.E., Rees, E., Walters, J.T., O'Donovan, M.C.: Genomic findings in schizophrenia and their implications. Molecular Psychiatry **28**(9), 3638–3647 (2023)

[31] Wharrie, S., Yang, Z., Raj, V., Monti, R., Gupta, R., Wang, Y., Martin, A.,

O'Connor, L.J., Kaski, S., Marttinen, P., *et al.*: Hapnest: efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes. Bioinformatics **39**(9), 535 (2023)

[32] Zhang, H., Zhan, J., Jin, J., Zhang, J., Lu, W., Zhao, R., Ahearn, T.U., Yu, Z., O'Connell, J., Jiang, Y., *et al.*: A new method for multiancestry polygenic prediction improves performance across diverse populations. Nature Genetics **55**(10), 1757–1768 (2023)

[33] James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J.: Linear regression. In: An Introduction to Statistical Learning: With Applications in Python, pp. 69–134. Springer, ??? (2023)

[34] Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical Programming **45**(1), 503–528 (1989)

[35] Asl, A., Overton, M.L.: Behavior of limited memory BFGS when applied to nonsmooth functions and their nesterov smoothings. In: Numerical Analysis and Optimization: NAO-V, Muscat, Oman, January 2020 V, pp. 25–55 (2021). Springer

[36] Nesterov, Y.: Smoothing technique and its applications in semidefinite optimization. Mathematical Programming **110**(2), 245–259 (2007)

[37] Hahn, G., Lutz, S.M., Laha, N., Cho, M.H., Silverman, E.K., Lange, C.: A fast and efficient smoothing approach to lasso regression and an application in statistical genetics: polygenic risk scores for chronic obstructive pulmonary disease (copd). Statistics and Computing **31**, 1–11 (2021)

[38] Alexander, D.H., Novembre, J., Lange, K.: Fast model-based estimation of ancestry in unrelated individuals. Genome Research **19**(9), 1655–1664 (2009)

[39] Zhou, H., Alexander, D., Lange, K.: A quasi-newton acceleration for high-dimensional optimization algorithms. Statistics and Computing **21**, 261–273 (2011)

[40] Howie, B., Marchini, J., Stephens, M.: Genotype imputation with thousands of genomes. G3: Genes— Genomes— Genetics **1**(6), 457–470 (2011)

[41] Howie, B.N., Donnelly, P., Marchini, J.: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics **5**(6), 1000529 (2009)

[42] Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F.L., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al.: The international hapmap project (2003)

[43] Privé, F., Aschard, H., Ziyatdinov, A., Blum, M.G.: Efficient analysis of large-scale genome-wide data with two r packages: bigstatsr and bigsnpr. Bioinformatics **34**(16), 2781–2787 (2018)

[44] Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., Smoller, J.W.: Polygenic prediction via bayesian regression and continuous shrinkage priors. Nature Communications **10**(1), 1776 (2019)