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) have guided significant contributions towards identifying disease associated Single Nucleotide Polymorphisms (SNPs) in Caucasian populations, albeit with limited focus on other understudied low-resource non-Caucasian populations. There have been active efforts over the years to understand and exploit the population specific vs. shared aspects of the genotype-phenotype relation across different populations or ethnicities to bridge this gap. However no single approach that unanimously outperforms all other methods exists. Furthermore, the efficacy of transfer learning models that are simpler than existing approaches remains an open question. We propose MultiPopPred, a suite of novel and simple trans-ethnic polygenic risk score (PRS) estimation methods, that tap into the shared genetic risk across populations and transfer 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. Extensive comparative analyses performed on simulated genotype-phenotype data reveal that MultiPopPred improves PRS prediction in the South Asian population by 69% on settings with low target sample sizes, by 19% overall across all simulation settings, and by 73% overall across all semi-simulated settings when compared to state-of-the-art trans-ethnic PRS estimation methods. We further observe a 44% overall improvement in PRS

prediction across 8 quantitative real-world traits from UKBiobank. This performance trend is promising and encourages application of MultiPopPred for reliable PRS estimation under resource constrained real-world settings.

Keywords: Target Population, Auxiliary Population, 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 effect sizes of SNPs (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 effect sizes of SNPs 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 suite of novel and simple trans-ethnic polygenic risk score (PRS) estimation methods, that simultaneously integrate information on effect sizes of SNPs learned from multiple well-studied auxiliary populations (like Caucasian populations) and transfer the same to a less-studied target population. MultiPopPred employs a specially designed Nesterov-smoothed penalized shrinkage model and a L-BFGS optimization routine. We present five versions 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 69% on settings with low target sample sizes, by 19% overall across all simulation settings, and by 73% overall across all semi-simulated settings when compared to state-of-the-art trans-ethnic PRS estimation methods. We further observe a 44% overall improvement in PRS prediction across 8 quantitative real-world traits from UKBiobank. This performance trend is promising and encourages application of MultiPopPred for reliable PRS estimation under resource constrained real-world settings.

2 Results

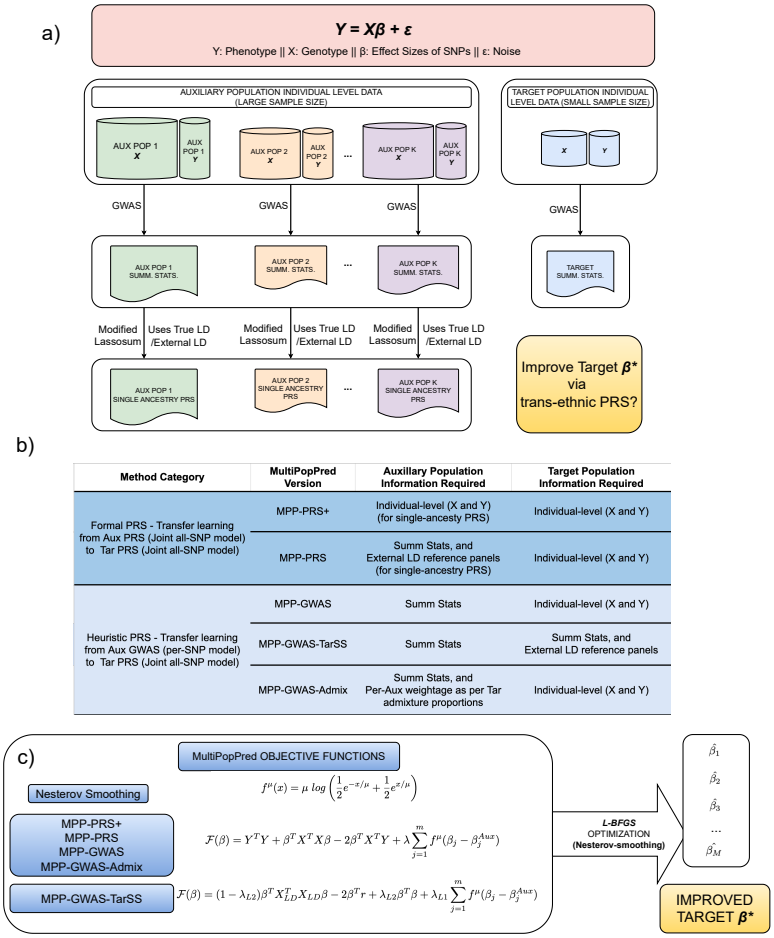
2.1 Overview of MultiPopPred

MultiPopPred builds upon the advancements made by previous trans-ethnic PRS estimation methods, while overcoming their shortcomings, to provide improved effect size estimates of SNPs 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 five versions of MultiPopPred. First,

we propose two formal PRS versions that aim to perform transfer learning from joint all-SNP PRS models to a joint all-SNP PRS model. These include (i) MultiPopPred-PRS+ (MPP-PRS+) and, (ii) MultiPopPred-PRS (MPP-PRS). MPP-PRS+ assumes availability of complete information for training, requiring individual level data for the target as well as auxiliary populations. MPP-PRS, on the other hand requires only the target population's individual level data while using external LD reference panels for the auxiliary populations' PRS computation. Next we propose three heuristic PRS versions that perform transfer learning directly from per-SNP models (GWAS summary statistics) to a joint all-SNP PRS model. These include (i) MultiPopPred-GWAS (MPP-GWAS), (ii) MultiPopPred-GWAS-Admixture (MPP-GWAS-Admix), and (iii) MultiPopPred-GWAS-TarSS (MPP-GWAS-TarSS). MPP-GWAS and MPP-GWAS-Admix require individual level target data for training along with summary level auxiliary data. MPP-GWAS-TarSS 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. Except MPP-GWAS-Admix, all four versions of MPP integrate information from multiple auxiliary populations by equally weighting each source of information. MPP-GWAS-Admix weighs each auxiliary population in proportion to their admixture component in target individuals. Further details can be found in Methods and Figure 1. All five versions of MultiPopPred employ a Nesterov-smoothed penalized shrinkage model and a L-BFGS optimizer for accurate and effective convergence to produce precise β estimates of SNPs. A comprehensive overview of all simulation and real-world benchmarks that we have used to evaluate our method against and performed comparative analysis have been detailed in section 2.2 and Supplementary Data 1-2.

2.2 Overview of Comparative Benchmarks

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] besides MultiPopPred. 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, semi-simulation and real-world settings, involving a range of sample sizes and heritability values (Sections 2.3, 2.4 and 2.5). Wherever applicable, like in any traditional machine learning setup, the available data has been split into training (to learn a predictive model), validation (to tune hyper-parameters) and testing (to report performance on) sets. Hyper-parameter tuning results for the L1 and L2 penalizations in case of MultiPopPred can be found in Supplementary Table 1, Supplementary Fig. 1-4. We tuned MultiPopPred's hyper-parameters for a subset of simulated settings and froze the optimal hyper-parameter values for the rest of the simulated settings (Supplementary Data 3). In the real-world setting, we tuned MultiPopPred's hyper-parameters for each trait separately (Supplementary Data 3). We used 1000 Genomes



Method Category	MultiPopPred Version	Auxillary Population Information Required	Target Population Information Required
Formal PRS - Transfer learning from Aux PRS (Joint all-SNP model) to Tar PRS (Joint all-SNP model)	MPP-PRS+	Individual-level (X and Y) (for single-ancesty PRS)	Individual-level (X and Y)
	MPP-PRS	Summ Stats, and External LD reference panels (for single-ancestry PRS)	Individual-level (X and Y)
Heuristic PRS - Transfer learning from Aux GWAS (per-SNP model) to Tar PRS (Joint all-SNP model)	MPP-GWAS	Summ Stats	Individual-level (X and Y)
	MPP-GWAS-TarSS	Summ Stats	Summ Stats, and External LD reference panels
	MPP-GWAS-Admix	Summ Stats, and Per-Aux weightage as per Tar admixture proportions	Individual-level (X and Y)

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) Five versions 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 versions. L-BFGS optimization of the Nesterov-smoothed objective function produces robust and reliable target effect size estimates. Here X, Y, and β refer to the genotype, phenotype and effect sizes of SNPs respectively. For more details on the formulae, see Methods. Abbreviations: Aux: Auxiliary Population, Tar: Target Population, Summ. Stats.: Summary Statitics.

Project (1KG) Phase 3 [27] defined super-populations (EUR, EAS, AMR, AFR, SAS) as reference panels wherever applicable (Methods).

2.3 Benchmarking using Simulated Data

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 effect sizes of SNPs 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 (Supplementary Data 4-6). 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.)

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 to 4. 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 effect sizes of SNPs 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.

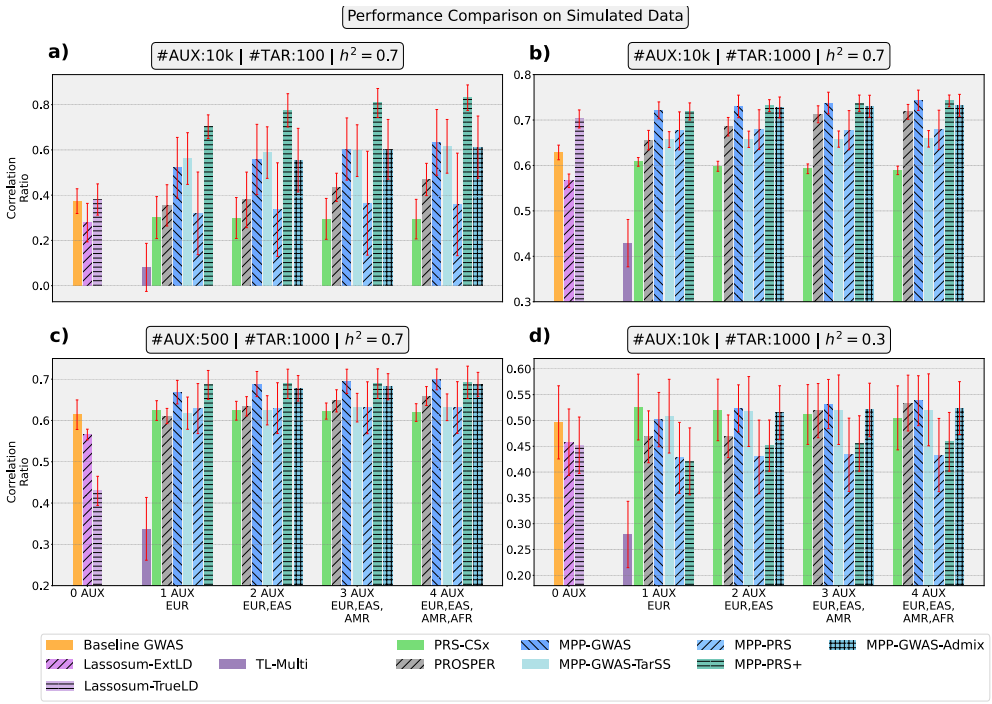


Fig. 2 Performance Comparison on Simulated Data (Phenotype Prediction): We depict the comparative performance of MultiPopPred with state-of-the-art (SOTA) methods under four different simulation settings and four auxiliary-target population tuples. (A) The number of samples per auxiliary population was 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.

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

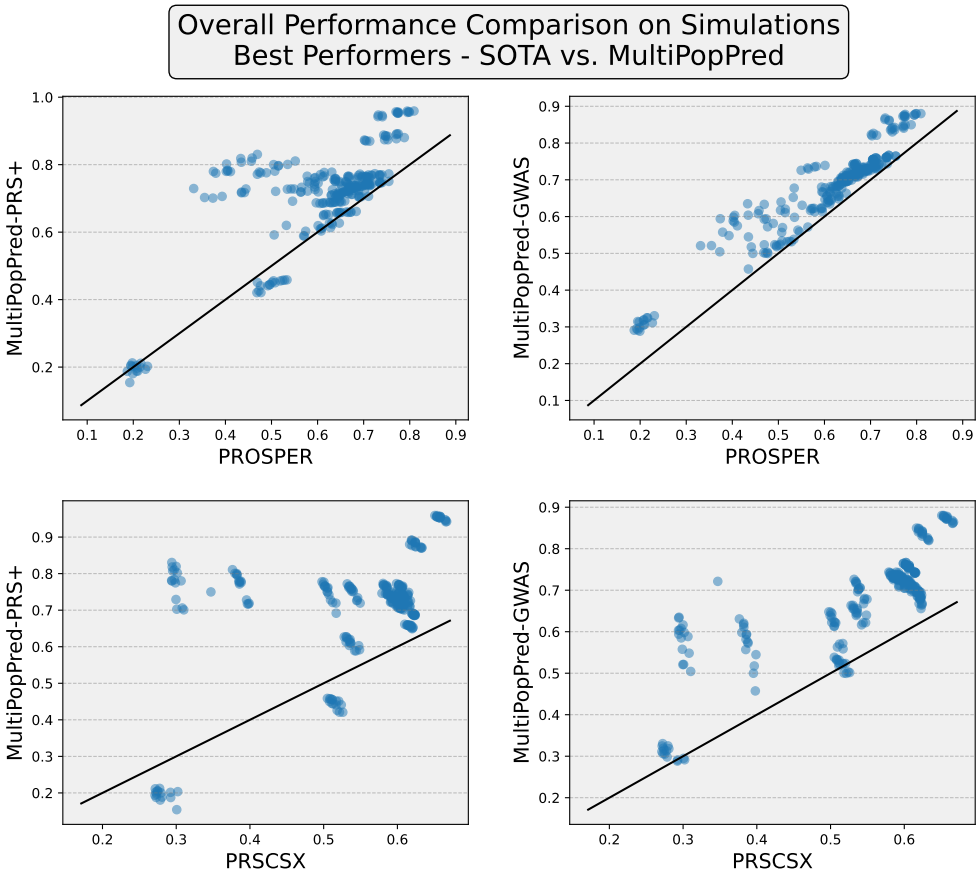


Fig. 3 Overall Performance Comparison on all simulation scenarios tested (wrt Phenotype Prediction using Correlation Ratio): We depict the comparative performance of the two best performing versions of MultiPopPred (MultiPopPred-PRS+ and MultiPopPred-GWAS) against the two best performing and leading SOTA methods (PROSPER and PRS-CSx). This figure has been plotted using a superset of the data shown in Figure 2 and a subset of data available in Supplementary Data 4-6. Here, each point in each scatter plot corresponds to the average performance measure over 5 replicate datasets per simulation configuration. There are 3 simulation parameters, 6 different values per simulation parameter, 15 different auxiliary-target tuples per simulation configuration making a total of 270 points per scatter plot.

low h^2 values (Figure 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 (Figure 2d) except for TL-Multi. At moderate h^2 (50%-70%) and higher h^2 (90%-100%) values, MultiPopPred (specifically the MPP-PRS+, MPP-GWAS and MPP-GWAS-Admix versions) was observed to have a clear edge over other methods (Figures 2a, 2b, 2c). It

must be noted that MPP-GWAS-TarSS 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 other MultiPopPred versions. Between MPP-GWAS-TarSS and PRS-CSx, the former was seen to consistently outperform the latter.

Figure 3 depicts the overall improvements, across all simulation settings, in performance w.r.t. Correlation Ratio when the two best performing versions of MultiPopPred (MPP-PRS+ and MPP-GWAS) were compared against the two best performing and leading SOTA methods (PROSPER and PRS-CSx). Averaging across the simulated scenarios, MPP-PRS+ demonstrated 17% and 32% improvements over PROSPER and PRS-CSx respectively while MPP-GWAS was observed to obtain a 13% and 26% improvement over PROSPER and PRS-CSx respectively. Further comparisons of the MultiPopPred versions with SOTA methods can be found in Supplementary Figure 5.

2.3.3 MultiPopPred outperforms SOTA methods especially 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 in our simulations, around 5000-10,000 auxiliary samples per population should suffice to do an efficient transfer learning operation, following which the performance saturates. Figure 2c shows MultiPopPred demonstrating a 8%, 6% and 5% average improvement in correlation ratio over Baseline GWAS, PRS-CSx and PROSPER respectively. Figure 2b shows MultiPopPred demonstrating a 12%, 18% and 3% 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 five versions of MultiPopPred were observed to clearly outperform other methods at low target samples depicting its robustness (Figure 2a). In the low target sample setting (Figure 2a), MultiPopPred demonstrated a 54%, 93% and 42% 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.

2.3.4 Evaluating MultiPopPred on PROSPER's Benchmarks

The results presented thus far focused on our simulation strategies for genotypes and phenotypes where the ground-truth effect sizes of SNPs for each population (target as well as auxiliary) were designed to be spread around 0 as per a normal distribution. This design choice was made to maintain close resemblance to the infinitesimal model

or omnigenic model [31, 32] of phenotypes such as height, that are known to have a high degree of polygenicity. Another model of SNP effect size distribution popular in literature is the sparse model which assumes that all non-causal SNPs have zero contribution towards a phenotype. Such sparse models have been observed to be exhibited by gene-expression traits [33]. While both models of SNP effect size distribution have their own merits and shortcomings, we incline towards the infinitesimal/omnigenic model in our simulations so as to cater to our focus on highly polygenic traits rather than sparse gene-expression based traits. PROSPER's simulations follow the sparse model of SNP effect sizes and hence methods that learn tighter estimates of βs which are concentrated around 0 with a high peak at the mean are expected to perform better. On the other hand, in our simulations, methods that learn β estimates spread around 0 are expected to perform better.

To test the applicability of MultiPopPred under a different simulation framework, we performed an additional set of analyses wherein MultiPopPred was subject to three different simulation benchmarks [34] proposed by PROSPER [20] and its predecessor CT-SLEB [34] (see Supplementary Table 2). Similar to the trends observed in our simulation analyses, MPP-PRS+ was observed to be among the better performers than the other MPP versions, while MPP-GWAS-Admix was observed to emerge as the best out of the five MPP versions. MultiPopPred-GWAS-TarSS was also observed to closely follow MPP-PRS+ in terms of performance. All versions of MPP were observed to perform worse than PROSPER. While the gap in MPP and PROSPER's performance was relatively smaller for the first benchmark (0.01 fraction of causal SNPs), it was seen to increase significantly as the fraction of causal SNPs was reduced progressively to 0.001 and 0.0005 in benchmarks two and three respectively. This can be attributed to the fact that PROSPER produces β estimates more concentrated around 0 than MPP, thereby aligning more with their simulation framework (Supplementary Figure 6). To verify our understanding on the likely cause for the gap in MPP and PROSPER's performance, we examined the distribution of SNP effect sizes produced by the two methods in our simulations. Supplementary Figure 7 confirms that MPP performs better than PROSPER in our simulations because of its ability to learn more spread out βs, which is in line with our simulation framework. The insights from this analysis indicate that depending on the underlying sparse or infinitesimal distribution of SNP effect sizes for a given trait, PROSPER or MPP could be used appropriately to compute PRS.

2.3.5 Secondary Simulations

We performed a set of secondary simulation analyses to further validate the results and observations mentioned above and to refine our understanding of the working of MultiPopPred under various conditions. First, we benchmarked the performance of MultiPopPred against HAPNEST [35], a publicly available simulated genotype dataset, in low and high target sample size settings similar to Fig. 2a and 2b. This simulation was performed with two different sets of SNPs with and without LD structures. In accordance with previously observed results, MultiPopPred was seen to clearly outperform SOTA methods (Supplementary Fig. 8) with PROSPER being the second best performer. Next, we changed the covariance matrix used during simulations to

account for poorly correlated pairs of populations. The ground truth per-SNP effect sizes for the target population (SAS) were simulated to have a high correlation (0.9) with a single auxiliary population (EUR) and a poor correlation with other auxiliary populations. It was observed that the presence of a single strongly correlated pair of auxiliary-target population pair had a more pronounced effect on performance in low sample size settings (Supplementary Fig. 9). The final few secondary simulation analyses were conducted to support our choice of specific elements, such as the optimization routine and penalization strategy, in the design of MultiPopPred (Supplementary Fig. 10-11).

To summarize, MultiPopPred was observed to report an overall 19% improvement in correlation ratio in our simulations as compared to all SOTA methods (averaged across the five MultiPopPred versions, 18 simulated settings and 15 auxiliary-target tuple configurations Figure 2, Supplementary Data 4-6). Individually, each MultiPopPred version reported an overall improvement of 29% (MPP-PRS+), 10% (MPP-PRS), 24% (MPP-GWAS), 10% (MPP-GWAS-TarSS), and 23% (MPP-GWAS-Admix), over all SOTA methods and all simulation settings.

2.4 Benchmarking using Semi-Simulated Data

Semi-simulated data analysis was performed using real-world genotype data on EUR, EAS, AFR and SAS populations from UK Biobank [36] while phenotypes were simulated as per the procedure described in section 2.3 (also, see Methods). We performed a comparative analysis of MultiPopPred and other SOTA methods with respect to phenotype predictions under two different semi-simulated settings corresponding to low and high target sample sizes. Similar to the fully simulated analysis, each semi-simulation was repeated 5 times and for each setting MultiPopPred and other methods were run under 3 different configurations of auxiliary-target population tuples (see Fig. 4, Supplementary Data 7). On average, MultiPopPred was observed to outperform PRS-CSx and PROSPER by a significant margin of 126% in the low target sample size setting, by 19% in the high target sample size setting and by 73% overall. These observations have been reported as percentage improvements in the average correlation ratio over the 5 different semi-simulated settings and the 3 different auxiliary-target population tuples. It must still be maintained that the drop in PROSPER's performance, especially in setting with low target sample sizes, can be attributed to the non-suitability of the extreme setting for the method.

2.5 Benchmarking using Real World Data

Real world analysis was performed on 8 quantitative traits (four having relatively higher heritability values and four having relatively lower heritability values) from UKBiobank viz., (i) Height, (ii) Body Mass Index (BMI), (iii) Systolic Blood Pressure (SBP), (iv) Diastolic Blood Pressure (DBP), (v) High-Density Lipoprotein (HDL), (vi) Low-Density Lipoprotein (LDL), (vii) Total Cholesterol (TC) and (viii) Triglycerides (TG) (see Supplementary Data 8-10). These traits were chosen to cover different heritability (h^2) values. The h^2 values for these traits were observed from [37] and have been consolidated in Supplementary Table 3. SAS was set as the target population

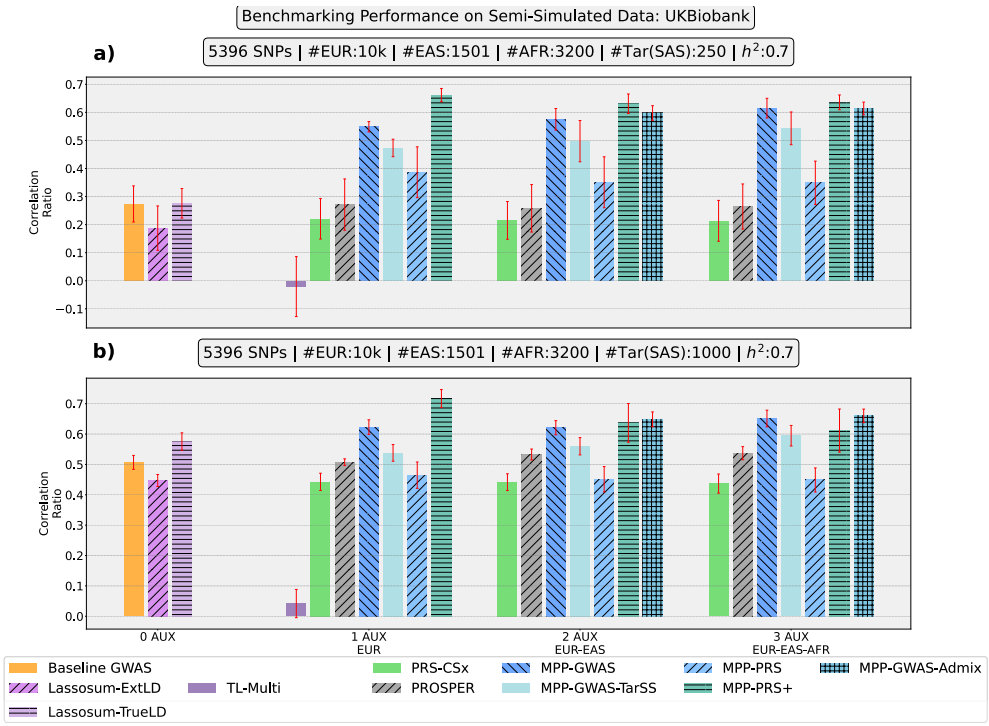


Fig. 4 Performance Comparison on Semi-Simulated Data (Phenotype Prediction): We depict the comparative performance of MultiPopPred with state-of-the-art (SOTA) methods under two different semi-simulation settings and three auxiliary-target population tuples. First, the number of samples per auxiliary population was set to 10,000 for EUR, 1501 for EAS, 3200 for AFR and the number of target population samples was set to 250 with h^2 being 70%. Next, the number of auxiliary samples and h^2 remained the same as before but the number of target samples was increased to 1000. The analysis was restricted to a set of 5396 SNPs (from chromosome 22) that were common across UK Biobank and 1000 Genomes. Abbreviations: Aux: Auxiliary Population, Tar: Target Population, h^2: Heritability.

while EUR, EAS and AFR were used as auxiliary populations wherever applicable. All five versions of MultiPopPred and the different SOTA methods were evaluated across 3 different datasets per trait and the performance was reported in terms of R^2 instead of Correlation Ratio due to the absence of ground-truth effect sizes of SNPs.

2.5.1 Heuristic PRS versions show promise alongside formal PRS estimation approaches

Traditionally whenever a PRS estimate is talked about, one refers to a joint all-SNP or all-SNPs model for training and subsequent predictions, as opposed to a per-SNP model. We observe in our real world analysis that the Baseline GWAS approach, despite being a per-SNP model, performs decently enough across all 8 UKBiobank traits (see Figure 6). This is further reflected in the performance of MultiPopPred

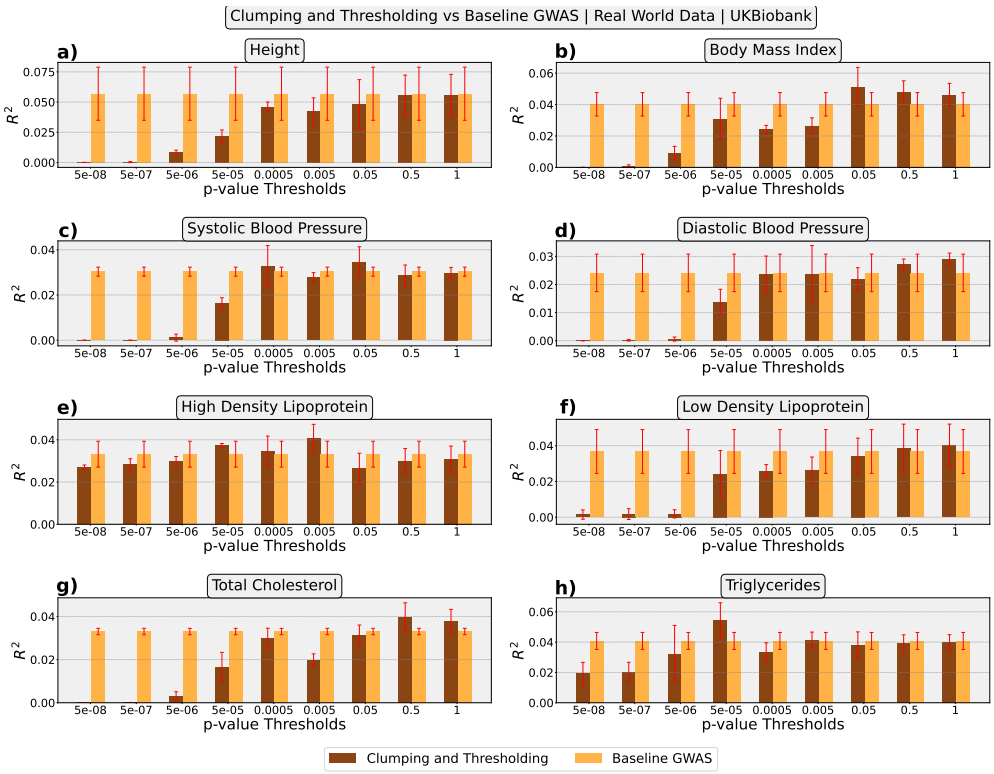


Fig. 5 Performance Comparison on Real-World Data (Phenotype Prediction): We depict the comparative performance of Baseline GWAS against a traditionally used Clumping and Thresholding PRS computation method, when applied to 8 different quantitative traits from UKBiobank.

versions that build directly upon Baseline GWAS. The heuristic PRS versions of MultiPopPred namely, MPP-GWAS, MPP-GWAS-Admix and MPP-GWAS-TarSS reported an overall improvement in R^2 over SOTA methods by a margin of 28%, 12%, and 28% respectively. This is still lesser than a 44% overall improvement reported by MPP-PRS+, the leading formal PRS version of MultiPopPred, but the promise shown demands further exploration of heuristic based PRS approaches. It must be noted that within the heuristic PRS versions of MultiPopPred, very minimal variation in performance was observed potentially indicating a weaker dependence on the exact manner of aggregation of information from the auxiliary populations. MPP-GWAS-TarSS was observed to perform on-par or slightly worse than the other two heuristic PRS MultiPopPred versions, likely due to the use of external LD as opposed to true LD of the target population.

To further investigate this trend observed in heuristic PRS versions, we performed a comparative analysis of Baseline GWAS alone with Clumping and Thresholding (CT) [34] - a traditional and very simplistic PRS estimation method that has served as the baseline for most previous SOTA methods existing in literature. Figure 5 shows

that at stringent p-value cutoffs ($5 \times 10^{-8}, 5 \times 10^{-7}, 5 \times 10^{-6}$), CT performs poorly compared to Baseline GWAS (see Supplementary Data 11). As the p-value threshold increases, the performance of CT improves reaching an on-par or in some cases a better level than Baseline GWAS. This indicates that while filtering the set of available SNPs prior to PRS computation is essential, it must not come at the cost of truly causal or truly informative SNPs. Retaining additional SNPs during PRS computation does not seem to harm the predictive performance by a huge margin. A logical solution to this situation could be to use the complete available set of SNPs for PRS estimation while employing sufficiently regularized models (like MultiPopPred or PROSPER) to ensure that the contribution of non-informative SNPs towards prediction is controlled. Furthermore, the use of simple GWAS as the baseline should be promoted, as opposed to CT, while proposing novel PRS estimation methods.

2.5.2 MultiPopPred-PRS+ outperforms SOTA methods for all real world traits

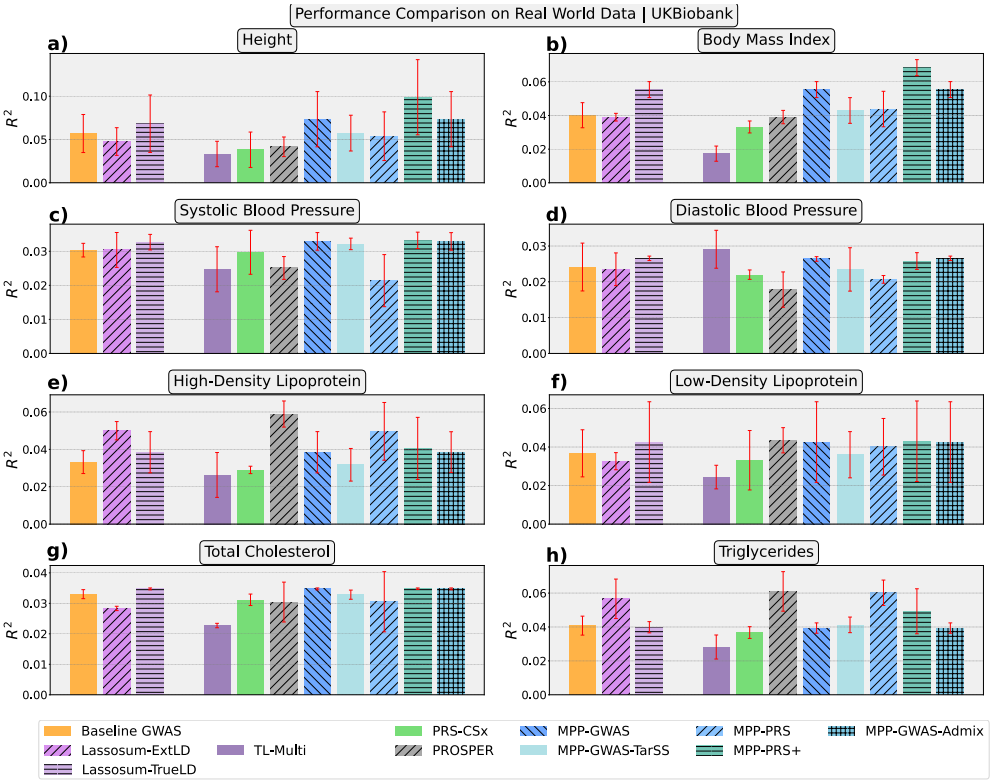


Fig. 6 Performance Comparison on Real-World Data (Phenotype Prediction): We depict the comparative performance of MultiPopPred with state-of-the-art (SOTA) methods when applied to 8 different quantitative traits from UKBiobank.

MPP-PRS+ was observed to be the best performer among the five MultiPopPred versions, which is consistent with the trends observed in our simulation and semi-simulation analyses. This improved performance is also in line with the full information, i.e., individual level target an auxiliary information available to MPP-PRS+. PROSPER uses individual level target information only during hyperparameter tuning in the validation dataset while PRS-CSx uses only summary-level information from the target and auxiliary populations. These results suggest that individual-level information should be used as much as possible, whenever available, so that the LD structure used by the method matches the true LD structure of the population being studied.

MPP-PRS+ was further observed to outperform SOTA methods by an overall margin of 123%, 112%, 19%, 13%, 17%, 26%, 19%, and 23% for Height, BMI, SBP, DBP, HDL, LDL, TC, and TG respectively (see Fig. 6). It must be noted, however, that for HDL and TG respectively, MPP-PRS+ reported a drop in R^2 by 32% and 20% when compared to PROSPER alone (Fig. 6e, 6h). This may be attributed to the baseline method on top of which MPP-PRS+ or PROSPER build, i.e., Lassosum-TrueLD for MPP-PRS+ and Lassosum2 (or Lassosum-ExtLD) for PROSPER. In both HDL and TG, Lassosum2 is seen to outperform Lassosum-TrueLD by 39% and 42% respectively. MPP-PRS, on the other hand, builds on Lassosum2 and hence was observed to perform better than MultiPopPred-PRS+ for these two specific traits.

Summarizing all the results in the real-world analyses, the five MultiPopPred versions were observed to report a 44% (MPP-PRS+), 18% (MPP-PRS), 28% (MPP-GWAS), 12% (MPP-GWAS-TarSS), and 28% (MPP-GWAS-Admix) overall improvement in R^2 averaged across the 24 (8×3) different trait-dataset configurations and the SOTA methods. (Supplementary Data 8, Supplementary Fig. 12). Compared to the leading SOTA method - PROSPER alone, MPP versions reported an overall improvement of 33% (MPP-PRS+), 3% (MPP-PRS), 19% (MPP-GWAS), 3% (MPP-GWAS-TarSS), and 19% (MPP-GWAS-Admix) across all traits. Similarly, compared to PRS-CSx alone, the MPP versions were observed to report an overall improvement of 55% (MPP-PRS+), 28% (MPP-PRS), 38% (MPP-GWAS), 20% (MPP-GWAS-TarSS), and 38% (MPP-GWAS-Admix) across all traits. Figure 7 depicts the improvement attained by best two MultiPopPred versions over the leading SOTA methods PROSPER and PRS-CSx respectively.

3 Discussion

In this work, we have proposed MultiPopPred, a suite of novel trans-ethnic polygenic risk score (PRS) estimation methods that aim to harness the transferability of learned information from multiple well-studied auxiliary populations to an understudied target population. This work presents five versions of MultiPopPred that cater to varying real-world scenarios determined by the availability of (i) individual level vs. summary level data, (ii) high resource vs. low resource data, and (iii) genetically similar vs. different population tuples. MultiPopPred utilizes a L1 penalized shrinkage model that strives to close the gap between the target and aggregated auxiliary effect sizes of SNPs while minimizing the least squared loss. MultiPopPred also employs a

Fig. 7 Overall Performance Comparison on all 8 traits from UKBiobank (wrt Phenotype Prediction using R^2): We depict the comparative performance of the two best performing versions of MultiPopPred (MultiPopPred-PRS+ and MultiPopPred-GWAS) against the two best performing and leading SOTA methods (PROSPER and PRS-CSx). This figure has been plotted using a subset of the data shown in Figure 6. Here, each point in each scatter plot corresponds to the performance measure for an individual replicate dataset for a trait. There are 8 traits and 3 replicate datasets per traits, making a total of 24 points per scatter plot.

Nesterov-smoothed objective function and a L-BFGS optimization routine for effective PRS prediction in a given target population. Through extensive simulation and semi-simulation analyses, covering a range of trait heritabilities and available auxiliary and target population sample sizes, we have observed that MultiPopPred robustly outperforms other methods in extreme conditions where there is a dearth of target samples. MPP-PRS+ was seen to emerge as the best performing version of MultiPopPred, likely owing to its use of full information on target and auxiliary individual-level data. Subsequent evaluations of MultiPopPred on application to quantitative real-world traits from UKBiobank also reported consistent trends.

MultiPopPred is expected to report larger improvements in performance for a given trait when the underlying distribution of SNP effect sizes follows a non-sparse, i.e. infinitesimal/omnigenic model. For traits with a sparse underlying distribution of SNP effect sizes, MultiPopPred reported comparable performance on one benchmark while PROSPER was observed to perform better on the other two benchmarks. This indicates that MultiPopPred or PROSPER could be chosen for PRS computation depending on the type of trait under consideration. Other than that, as in any transfer learning approach, MultiPopPred is expected to perform better when the auxiliary populations, from which transfer of knowledge occurs, are sufficiently closely related to the target population in terms of LD structure and hence the underlying distribution of SNP effect sizes. In the absence of such genetically similar auxiliary populations, MultiPopPred's predictive performance is expected to be guided largely by the number of target and auxiliary samples available. MultiPopPred's effective performance could potentially 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 Nesterov-smoothed objective function with 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. Both these potential reasons could be verified via ablation studies as a future work. A surprising takeaway from the simulation and semi-simulation analyses was the impressive performance of the Heuristic PRS versions of MultiPopPred despite performing a direct transfer of information from per-SNP GWAS models to a joint all-SNP PRS model. This observation demonstrates potential and scope for further exploration of direct GWAS to PRS trans-ethnic methods in the future.

A major bottleneck observed while estimating PRS in low resource populations such as South Asian was the limited availability of samples for training, validation and testing. While methods like MultiPopPred provide a reasonable alternative in such extreme situations, we must set realistic expectations on the maximum achievable performance in such cases. To set a benchmark on the variation of predictive performance as the sample size increases, we conducted a simple analysis on the EUR samples from UK Biobank for the trait Height. While all our previous analyses have treated SAS as the target population, this time EUR was chosen as the target owing to the flexibility offered by EUR in switching from very low sample sizes to very high sample sizes through simple subsetting of samples at random. The performance of Baseline GWAS (baseline for heuristic PRS versions) and Lassosum2 (baseline for formal PRS versions) was noted as the number of training samples were increased. Supplementary Data 12 shows that the best performance (test $R^2 = 0.139$ for Baseline GWAS and 0.097 for Lassosum2) was observed with training samples as high as $4,32,526$. As the sample size was decreased, the performance was also observed to dip proportionately (test $R^2 = 0.025$ for Baseline GWAS and 0.011 for Lassosum2 with $10,000$ training samples). This shows that although at the outset, the observed absolute R^2 values in our real world analyses might seem pretty low from a standard machine learning point of view (and may even indicate overfitting when training and test performances are compared), on a relative scale to the best achievable performance, they demonstrate good predictive power.

While the observed results are promising, 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 default and Tar-SS versions all assume equal weightage for each auxiliary population. Furthermore, weighing auxiliary populations by admixture proportions also presents its 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. Besides that, as a future work, we intend to subject MultiPopPred to evaluation under a series of real-world binary traits and some additional secondary simulation settings to explore potential improvements offered by the method in terms of computational space and time. Some more challenging aspects of MultiPopPred could benefit from future investigation are inclusion of rare or population specific SNPs, inclusion of non-linear effects (epistasis) of SNPs towards the phenotype, and 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. Nevertheless, the existing versions of MultiPopPred still perform comparably or better than existing SOTA methods in most scenarios, and can pave the way for its application to understand the genetic architecture of multiple phenotypes in understudied populations.

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 (PRS) estimation method that aims to tackle the under-representation of low-resource and consequently, understudied 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, and the low-resource population 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 [38], 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. 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 to a couple of thousands) such an approach fails to produce reliable estimates of β.

$$\begin{aligned} Y &= X\beta + \epsilon \\ \hat{\beta} &= \text{argmin}_{\beta}(Y - X\beta)^T(Y - X\beta) \end{aligned} \tag{1}$$

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 genotype-phenotype relationship 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 reliable β estimates ($\beta^{Aux_1}, \beta^{Aux_2}, ..., \beta^{Aux_K}$), each having dimension $m \times 1$, from their respective GWAS. Additionally, we have the target population's GWAS summary statistics 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 the standard least-squares based loss. Here, $\mathcal{F}$ denotes the objective function to be optimized, λ denotes the L1 shrinkage parameter, and β^{Aux} refers to an aggregate of ($\beta^{Aux_1}, \beta^{Aux_2}, ..., \beta^{Aux_K}$). The way the aggregation of these βs is performed renders different versions of MultiPopPred, which have been elaborated on in section 4.1.4.

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

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. 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 (Supplementary Fig. 8), which has been employed by SOTA such as PRS-CSx (see 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 [39] (Supplementary Fig. 9). 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 [40]. In our case, we employ Nesterov smoothing (smoothing parameter μ fixed at 0.1) [41] with an entropy prox function $f^{\mu}(x)$ (Equation 3) [42] 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 Versions

We have proposed five versions of MultiPopPred based on 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. First, we propose two formal PRS versions that perform a joint model PRS to PRS transfer learning. These include (i) MultiPopPred-PRS+, and (ii) MultiPopPred-PRS. Next we propose three heuristic PRS versions that aim to perform transfer learning directly from per-SNP model GWAS summary statistics to a joint PRS model. These include (i) MultiPopPred-GWAS, (ii) MultiPopPred-GWAS-TarSS, and (iii) MultiPopPred-GWAS-Admixture. Each of these versions is described in detail in the following subsections. It must be noted that we did not develop admixture and summary-level-data-only based versions of MPP-PRS because the advantages offered by MPP-GWAS-TarSS and MPP-GWAS-Admix over MPP-GWAS were observed to be quite subtle in our simulation studies.

MultiPopPred - Formal PRS Versions:

MultiPopPred-PRS+

This version of MultiPopPred (MPP-PRS+) assumes availability of complete information on target as well as auxiliary populations, for the purpose of training. In other

words, MPP-PRS+ requires the target as well as the auxiliary populations' genotype (X) and phenotype (Y), along with their respective GWAS summary statistics as inputs. Auxiliary populations' single-ancestry PRS is computed using a modified version of Lassosum (Lassosum-TrueLD) that uses the true LD for each auxiliary population computed from its respective genotype. In this version of MultiPopPred, 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., $\beta^{Aux_1}, \beta^{Aux_2}, ..., \beta^{Aux_K}$ from the single-ancestry PRS, then the aggregated β^{Aux} is defined as per Equation 5 below.

$$\beta^{Aux} = \frac{1}{K}\ \beta^{Aux_1} + \frac{1}{K}\ \beta^{Aux_2} + ... + \frac{1}{K}\ \beta^{Aux_K} \tag{5}$$

If Nesterov smoothing is used along with this MPP version, 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-PRS

This version of MultiPopPred (MPP-PRS) relaxes one of constraints on the input requirements imposed by MPP-PRS+. MPP-PRS assumes that we have the target population's individual level genotype and phenotype along with its GWAS summary statistics as well as the auxiliary populations' GWAS summary statistics available with us. However, the auxiliary populations' genotype and phenotype are not available. For each auxiliary population, MPP-PRS uses its GWAS summary statistics and its external LD reference panel as inputs to Lassosum2 (Lassosum-ExtLD) to compute its single-ancestry PRS. Hereon, MPP-PRS resembles MPP-PRS+ in terms of the weightage given to each auxiliary population and the objective function to be optimized.

MultiPopPred - Heuristic PRS Versions:

MultiPopPred-GWAS This version of MultiPopPred (MPP-GWAS) requires the target population's genotype (X) and phenotype (Y), along with its GWAS summary statistics as well as each auxiliary population's GWAS summary statistics as inputs. Like the preceding MPP versions, MPP-GWAS also assigns equal weightage to each auxiliary population (Equation 5) and uses Equations 6-7 for optimization. The difference between MPP-GWAS and the preceding two versions, as the name suggests, lies in the basis for transfer learning. In case of MPP-GWAS, $\beta^{Aux_1}, \beta^{Aux_2}, ..., \beta^{Aux_K}$ correspond to each auxiliary population's GWAS summary statistics (each of size $m \times 1$), and not the single-ancetry PRS. MPP-GWAS thus aggregates the effect sizes of SNPs from each auxiliary population's summary statistics before proceeding with the optimization as per Equations 6-7.

MultiPopPred-GWAS-Admixture

Like MPP-GWAS, this version of MultiPopPred also requires the target X and Y along with auxiliary GWAS summary statistics as inputs. In this version, 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 [43]. (We employ the ADMIXTURE tool in a supervised manner with quasi-Newton convergence acceleration [44].)

$$\beta^{(i)}_{AuxAdmx} = w^{(i)}_{Aux_1}\ \beta^{Aux_1} + w^{(i)}_{Aux_2}\ \beta^{Aux_2} + ... + w^{(i)}_{Aux_K}\ \beta^{Aux_K} \tag{8}$$

Here, $w^{(i)}_{Aux_k}$ corresponds to the admixture proportion of auxiliary population k in the ith target individual such that $\sum_{k=1}^{K} w^{(i)}_{Aux_k} = 1$. Furthermore, $\beta^{Aux_1}, \beta^{Aux_2}, ..., \beta^{Aux_K}$ (each $m \times 1$) refer to each of the K auxiliary populations' βs from their respective GWAS 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.

$$\beta^{(i)}_{AuxAdmx} = \left[\beta^{(i)}_{SNP_1}\ \beta^{(i)}_{SNP_2}\ \beta^{(i)}_{SNP_3}\ ...\ \beta^{(i)}_{SNP_j}\ ...\ \beta^{(i)}_{SNP_m}\ \right]^T \tag{9}$$

$$\begin{gathered}\beta_{Aux} = \left[\beta^{'}_{SNP_1}\ \beta^{'}_{SNP_2}\ \beta^{'}_{SNP_3}\ ...\ \beta^{'}_{SNP_j}\ ...\ \beta^{'}_{SNP_m}\right]^T \quad where, \\ \beta^{'}_{SNP_j} = median(\beta^{(1)}_{SNP_j}, \beta^{(2)}_{SNP_j}, ..., \beta^{(i)}_{SNP_j}, ..., \beta^{(n)}_{SNP_j})\end{gathered} \tag{10}$$

MultiPopPred-GWAS-TarSS

This version of MultiPopPred (MPP-GWAS-TarSS) 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 version 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, MPP-GWAS-TarSS offers an efficient and fast (owing to Nesterov smoothing) alternative without a considerable loss in performance. Akin to the MPP-GWAS version, 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 version 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 $X_{LD}^T X_{LD}$ computed using an external LD panel (1000 Genomes) while $X^T Y$ is substituted by the SNP-wise correlation (r) computed from the target population's summary statistics using the formula $r = (n_{Tar}\beta^{Tar})_{m\times 1}$ (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 $X_{LD}^T X_{LD}$ with $(1-\lambda_{L2})X_{LD}^T X_{LD} + \lambda_{L2} I$ 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.

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

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

4.1.5 MultiPopPred additional implementation details

MultiPopPred adopts a dynamic approach for estimation of effect sizes of SNPs depending on the number of SNPs available as inputs. When the number of SNPs is small enough ($< 30,000$), MultiPopPred processes the whole set of SNPs together as a single block. On the other hand, when the number of SNPs is large enough ($> 30,000$), MultiPopPred processes the set of SNPs in blocks of a pre-set size. In our real-world analyses, the SNPs are processed in per-chromosome blocks (i.e., all SNPs in a chromosome constitutes a single block). In our comparative analyses on PROSPER's benchmark (Harvard Dataverse [34]), SNPs are processed in blocks of size 100 (i.e., each block containing 100 contiguous SNPs) per chromosome. These implementation decisions on block-wise processing have been made considering computational efficiency in mind, in line with the standard implementation of pipelines employed by PROSPER [20] and PRS-CSx [21]. Please note that in the standard implementation of all these methods, one regression model is fitted per block, and a final regression model is used to regress the phenotype against the predicted phenotypes from all the blocks.

4.2 Datasets used for Benchmarking

4.2.1 Generating Simulated Data

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, 45, 46] 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. These SNPs were selected based on 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 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 [47] which were further subjected to a series of filters viz., (i) removal of triallelic and strand ambiguous variants and (ii) retention of SNPs whose 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 ($\bar{X}_j = 0$, $Var(X_j) = 1\ \forall j \in (1, m)$), β is a $(m \times 1)$ effect size vector, $\epsilon \sim \mathcal{N}(0, 1)$ is the $(n \times 1)$ noise vector. A higher value of heritability indicates greater contribution of the genetic components towards the phenotype and vice-versa. Standard/Baseline (per-SNP) GWAS was performed on the simulated data using `big_univLinReg` (`bigstatsr`) in R [48]. All scripts to perform data simulation as per the above description are available on our Github (Code Availability).

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

4.2.2 Generating Semi-Simulated Data

Semi-simulated Data: For semi-simulated analysis we obtained real-world genotypes from UK Biobank (Project Number 221732). The 502,140 samples in UK Biobank were first grouped into subgroups based on self-identified ancestry (Datafield 21000). We further filtered out samples with any potential ambiguity in self-identified ancestry to retain 455,497 EUR (British and Irish) samples, 1571 EAS (Chinese) samples, 3390 AFR (African) samples and 8015 SAS (Indian, Pakistani and Bangladeshi) samples. The set of SNPs was restricted to the set of common SNPs across the four different ancestry groups (EUR, EAS, AFR and SAS) and 1000 Genomes after application of approproate filters such as MAF $> 1\%$ and removal of non-biallelic and strand-ambiguous variants. This resulted in a set of 5396 SNPs in chromosome 22 that was used for further analysis. Subsequently phenotypes were generated using the multivariate Gaussian approach as described above.

4.2.3 Pre-processing Real-World Data

Real-World Data: For real-world data analysis, we obtained real world genotypes from UK Biobank (Project Number 221732). Eight quantitative phenotypes were also obtained from UK Biobank - (i) Standing Height (Datafield 50), (ii) Body Mass Index (Datafield 21001), (iii) Systolic Blood Pressure (Datafield 4080), (iv) Diastolic Blood Pressure (Datafield 4079), (v) HDL Cholesterol (Datafield 23406), (vi) LDL Cholesterol (Datafield 23405), (vii) Total Cholesterol (Datafield 23400) and (viii) Total Triglycerides (Datafield 23407). All samples with missing values in the phenotypes were discarded. The missing values in the genotypes were imputed with simplistic approaches (mean across individuals) owing to the fraction of missing values being pretty low ($\leq 1\%$). The set of SNPs was restricted to ones remaining post application of standard filters such as intersection with 1000 Genomes SNPs, MAF $> 1\%$ and removal of non-biallelic and strand ambiguous variants. This resulted in a total number of samples and SNPs as reported in Supplementary Data 10-11. All genotype and phenotype data was standardized (0 mean and unit variance) and the phenotypes were adjusted for covariates such as Genetic Principal Components (PCs) 1-10 (Datafields 22009-a1 to 22009-a10), Age at recruitment (Datafield 21022) and Sex (Datafield 31). For all comparative analyses, the number of EUR samples was fixed at 50,000 due to computational resources' limitations. For the target population 1000 randomly selected samples were reserved for validation and another 1000 randomly selected samples for testing while all other available samples were used for training. It was made sure that the set of training, validation and testing samples did not overlap. For each trait, 3 sets of randomly shuffled datasets were used and the performance was reported as the average over these datasets. Baseline GWAS was conducted on each real-world dataset using standard PLINK [49] procedures. Supplementary Figures 13-22 depict the Manhattan and QQ-plots corresponding to the GWAS conducted on the 8 real-world traits. The QQ-plots report genomic inflation factor (scaled for 1000 samples) λ_{1000} values between 0.9 to 1.05 for all cases (except for TG GWAS conducted on AFR) indicating acceptable genomic inflation levels. The Manhattan plots,

with thresholds set at p-value 5×10^{-8} further indicate that our GWAS pass standard quality checks.

4.2.4 Evaluation Metrics used in benchmarking

Evaluation Metric (Simulation and Semi-Simulation): For comparing performances with respect to phenotype prediction in simulations as well as semi-simulations, 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}). Note that, $X\beta_{Truth} = Y_{Truth} - \epsilon$. 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.

Evaluation Metric (Real-World): For comparing performances with respect to phenotype prediction in real-world analyses, we have used R^2, which is the square of the Pearson's correlation between the true (ground-truth) and the predicted phenotypes. Ideally, we would want the R^2 to be as close to the heritability of the trait as possible. Note that we could not use Correlation Ratio since β_{Truth} is not known in real-world data.

4.3 State-of-the-art methods used for comparative analyses and their parameter settings

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 single-ancestry PRS computation is performed via Lassosum2 [20, 24] using training and validation datasets. The next step involves modeling multiple joint cross-population penalized regression tasks (corresponding to multiple populations) that make use of the single-ancestry PRS and optimal parameters computed using Lassosum2 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 [50] 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 reference 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 an established benchmark for comparing and contrasting newly developed PRS estimation methods over the years [23].

TL-Multi: Transfer Learning - Multi was proposed by Tian et al. in 2022 [22] as a 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. The publicly available code for TL-Multi was observed to have consistency issues (difference in their implemented and paper's objective function), and hence a corrected version of the same was used for all analyses in this work.

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. Both Lassosum2 and the original Lassosum are the same in theory and the only difference between them is the implementation.

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. Baseline GWAS follows a per-SNP model as described in Equation 1.

Clumping and Thresholding: This method refers to a clumping and thresholding applied to a simple GWAS (performed on the target population training data without any additional information from other auxiliary populations) to select subsets of SNPs for PRS computation. A total of 540 different configurations were used for this method corresponding to the following parameters - 9 p-value thresholds ($5 \times 10^{-8}, 5 \times 10^{-7}, 5 \times 10^{-6}, 5 \times 10^{-5}, 5 \times 10^{-4}, 5 \times 10^{-3}, 5 \times 10^{-2}, 5 \times 10^{-1}, 1$), 6 r^2 thresholds ($0.001, 0.01, 0.05, 0.1, 0.2, 0.5$), 5 LD window thresholds (50kb, 100kb, 250kb, 500kb, 1000kb) and 2 reference genotypes (UKBiobank and 1000 Genomes). The best configuration was chosen based on a held out validation dataset.

4.4 Data Availability

Data generated in this study will be made available upon reasonable request to the authors. MultiPopPred's Supplementary Data files are available for access here: https://bit.ly/4nemWud. Data used in this study have been referred to appropriately: Access to the UK Biobank data was obtained under the project number 221732 from https://www.ukbiobank.ac.uk/; 1000 Genomes Phase 3 reference panel and genotypes: https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html, https://cncr.nl/research/magma/; Genetic maps for each sub-population as per 1000 Genomes: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates/; HAPNEST: https://www.ebi.ac.uk/biostudies/studies/S-BSST936; PROSPER's Harvard Dataverse Simulations: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/COXHAP.

4.5 Code Availability

The code developed in this study will be made available via the following Github link: https://github.com/BIRDSgroup/MultiPopPred. Codes of other relevant tools developed in other studies are accessible as follows: HAPGEN2: https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html; Interpolation of genetic maps: https://github.com/joepickrell/1000-genomes-genetic-maps; PRS-CSx: https://github.com/getian107/PRScsx; PROSPER and Lassosum2: https://github.com/Jingning-Zhang/PROSPER; TL-Multi: https://github.com/mxxptian/TLMulti; ADMIXTURE: https://dalexander.github.io/admixture/; PLINK: https://zzz.bwh.harvard.edu/plink/plink2.shtml.

5 Acknowledgments

We thank Prof. Tom Michoel and, 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., 23andMe, R.T., 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] Cui, R., Elzur, R.A., Kanai, M., Ulirsch, J.C., Weissbrod, O., Daly, M.J., Neale, B.M., Fan, Z., Finucane, H.K.: Improving fine-mapping by modeling infinitesimal effects. Nature Genetics **56**(1), 162–169 (2024)

[32] Barton, N.H., Etheridge, A.M., Véber, A.: The infinitesimal model: Definition, derivation, and implications. Theoretical Population Biology **118**, 50–73 (2017)

[33] Wheeler, H.E., Shah, K.P., Brenner, J., Garcia, T., Aquino-Michaels, K., Consortium, G., Cox, N.J., Nicolae, D.L., Im, H.K.: Survey of the heritability and sparse architecture of gene expression traits across human tissues. PLoS Genetics **12**(11), 1006423 (2016)

[34] 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)

[35] 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)

[36] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.*: The UK Biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)

[37] Karczewski, K.J., Gupta, R., Kanai, M., Lu, W., Tsuo, K., Wang, Y., Walters, R.K., Turley, P., Callier, S., Shah, N.N., et al.: Pan-UK Biobank gwas improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. MedRxiv, 2024–03 (2024)

[38] James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J.: An introduction to statistical learning: With applications in python. Springer (2023)

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

[40] 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

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

[42] 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)

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

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

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

[46] 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)

[47] 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)

[48] 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)

[49] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., *et al.*: PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics **81**(3), 559–575 (2007)

[50] 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)