# Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology

Ying Wang[1,2,3,*], Masahiro Kanai[1,2,3,4,5], Taotao Tan[6], Mireille Kamariza[7], Kristin Tsuo[1,2,3], Kai Yuan[1,2], Wei Zhou[1,2,3], Yukinori Okada[5,8,9,10,11], the BioBank Japan Project, Hailiang Huang[1,2], Patrick Turley[12, 13], Elizabeth G. Atkinson[6], Alicia R. Martin[1,2,3,*]

1. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
2. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
3. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
4. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
5. Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan
6. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
7. Society of Fellows, Harvard University, Cambridge, MA, 02138 USA
8. Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
9. Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan
10. Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan
11. Center for Infectious Disease Education and Research (CiDER), Osaka University, Suita 565-0871, Japan
12. Department of Economics, University of Southern California, Los Angeles, CA, USA
13. Center for Economic and Social Research, University of Southern California, Los Angeles, CA, USA

*Correspondence: yiwang@broadinstitute.org, armartin@broadinstitute.org

## Abstract

Polygenic risk scores built from multi-ancestry genome-wide association studies (GWAS, $PRS_{multi}$) have the potential to improve PRS accuracy and generalizability across populations. To provide the best practice to leverage the increasing diversity of genomic studies, we used large-scale simulated and empirical data to investigate how ancestry composition, trait-specific genetic architecture, and PRS methodology affect the performance of $PRS_{multi}$ as compared to PRS constructed from single-ancestry GWAS ($PRS_{single}$). In both simulations on 6 various scenarios and empirical analyses on 17 anthropometric and blood panel traits, we showed that the accuracy of $PRS_{multi}$ overall outperformed $PRS_{single}$ in the understudied target populations, except for a few comparisons where the understudied population only accounted for a very small proportion of the multi-ancestry GWAS. Further, using substantially fewer samples for traits such as height and mean corpuscular volume from Biobank Japan (BBJ) may achieve comparable accuracies to using 320,000 European (EUR) individuals from UK Biobank (UKBB). Finally, we find that incorporating PRS based on local ancestry-informed GWAS and large-scale EUR-based PRS improved predictive performance than using EUR-based PRS alone in understudied African (AFR) population, especially for less polygenic traits when there are variants with large ancestry-specific effects. Overall, our study provides insights into how ancestry composition and genetic architecture impact polygenic prediction across populations, particularly across imbalanced sample sizes. Our work also highlights the need for increasing diversity in genetic studies to achieve equitable PRS performance across ancestral populations and provides practical guidance on developing PRS from multiple resources.

# Introduction

Polygenic risk scores (PRS) are useful tools for approximating the cumulative genetic susceptibility to complex traits and diseases. PRS are typically calculated as the weighted sum of the number of risk variants, with weights based on their association in genome-wide association studies (GWAS). Using well-powered GWAS and advanced statistical methodology, PRS have shown early promise in predicting traits and disease risks, with accuracies comparable to monogenic variants and traditional clinical risk factors[1–5]. However, current GWAS have vast Eurocentric study biases, resulting in attenuated PRS accuracies in other populations, with performance declining with increasing genetic distance between the discovery and target populations[6–9]. Such accuracy differences could be attributable to various factors, such as demographic history, environment, phenotypic heterogeneity, and between-ancestry linkage disequilibrium (LD) and/or minor allele frequency (MAF) differences[6,7]. The current reduced performance of PRS across populations impedes their equitable applications and may even exacerbate health disparities especially for minority populations that tend to experience the greatest burden of disease[10,11].

To achieve the most accurate and generalizable PRS, we would require access to large-scale and diverse GWAS, especially with representation that matches the specific target population. However, GWAS in European (EUR) populations are currently much larger than in other populations, and although efforts are underway to rectify these gaps, it will be many years before the global population is fully represented. Helpfully, studies have shown that using GWAS data with even a small proportion of non-European individuals has the potential to improve the predictive accuracy of PRS in underrepresented populations[12–14]. This could largely be due to the fact that common variants explain a large proportion of heritable variation and that causal variants underlying complex traits and diseases are expected to be largely shared across ancestries[7,15–17]. With the increasing availability and scale of genomic data from underrepresented and ancestrally diverse populations, we are especially interested in how this greater diversity could improve the generalizability of PRS.

In particular, recently admixed populations consisting of chromosomal segments of mosaic ancestries, often systematically excluded from current genomic studies due to their complicated population structure[18,19], could provide unique opportunities to develop more generalizable PRS as their genetic effects are estimated in more consistent environments, reducing confounding relative to estimates across ancestry groups. Further, deep phenotyping is generally lacking or inconsistently measured in diverse populations across continents, but phenotypes can be measured more comparably in recently admixed populations. Recent methodological advancements in local ancestry inference and association testing have enabled us to conduct ancestry-specific GWAS in admixed populations[20–22]. It remains unclear how PRS based on such local ancestry-informed summary statistics perform in underrepresented populations and how to integrate them with available large-scale EUR-based PRS.

Recently developed statistical methodologies leverage the increasing diversity of GWAS data to improve PRS portability, including PolyPred[23], PRS-CSx[24] and CT-SLEB[25]. However, the effect of genetic architecture, ancestry composition of GWAS discovery cohorts, and PRS construction methodologies on cross-ancestry predictive accuracy remains largely unclear. For example, a recent study found no increase in accuracy when meta-analyzing GWAS from a relatively small Ugandan cohort with the large EUR-based data from UK Biobank (UKBB)[12]. Furthermore, theoretical frameworks for approximating expected PRS accuracy from multi-ancestry GWAS are lacking. Current theoretical calculations for PRS accuracy implicitly assume homogeneous-ancestry discovery samples[26,27], leaving out factors that are expected to play a role with multi-ancestry cohorts. Such factors may include between-ancestry LD and MAF differences, between-ancestry genetic correlation, and heritability and sample sizes from different ancestries.

To provide insights into those issues, we first explored the impact of ancestry compositions in discovery GWAS on predictive accuracy of PRS using large-scale population genetic simulations and real genomic data from the BioBank Japan (BBJ)[28] and UKBB. The overall study design is shown in **Figure S1**. In what follows, we use the expression **single-ancestry GWAS** to refer to a GWAS including only one ancestry; we use **multi-ancestry GWAS** to refer to those including two or more ancestries. We meta-analyzed **EUR GWAS** and GWAS in other minority populations (**Minor GWAS**) with different ratios of sample sizes to mimic multi-ancestry GWAS with varying ancestry composition. Specifically, we focused on Asians (EAS) and Africans (AFR) minority populations in this study. We compared PRS performance constructed from single-ancestry GWAS (**PRS$_{\text{single}}$**) and multi-ancestry GWAS (**PRS$_{\text{multi}}$**), respectively. We find that PRS$_{\text{multi}}$ generally outperforms PRS$_{\text{single}}$ (mostly large-scale EUR GWAS-derived PRS), but that performance depends on trait-specific genetic architecture and ancestry composition of discovery GWAS. As admixed populations are understudied yet disproportionately yield novel genetic findings[29], we further conducted local ancestry inference to explore whether, how, and to what extent PRS generalizability can be improved using GWAS discovery data from AFR-EUR admixed individuals. We find that PRS constructed from local ancestry-informed GWAS can improve PRS performance in the underrepresented AFR population for those less polygenic traits with large-effect ancestry-enriched variants. Overall, we show the PRS predictive performance is usually but not always improved using multi-ancestry GWAS as compared to using single-ancestry GWAS, which is highly dependent on ancestry-composition, trait specific genetic architecture, and PRS construction methods.

# Results

## Evaluating the effects of imbalanced sample sizes across ancestries on PRS accuracy through simulations

We simulated genotypes using HapGen2 and phenotypes according to six different scenarios with varying trait heritability ($h^2$ = 0.03, 0.05) and number of causal variants ($M_c$ = 100, 500, 1000), such that the polygenicity ranged from ~0.1% to ~1%. We assumed that the causal variants and their effect sizes are shared across ancestries (i.e., cross-ancestry genetic correlation is 1). To mimic the imperfect tagging of causal variants by genotyped or imputed variants, we excluded the causal variants when performing GWAS. As for single-ancestry discovery GWAS, we ran

GWAS or meta-analyzed GWAS in different numbers of bins, varying from 1 to 52 in each ancestry. As for multi-ancestry discovery GWAS, we meta-analyzed EUR GWAS and Minor GWAS (EAS or AFR GWAS) to vary the ancestry composition. We used different numbers of bins from EUR GWAS (from 4 to 52 with 4 increments, each bin with 10,000). We also varied the contribution from minority populations, ranging from 1 to 52 bins from EAS or AFR GWAS. We constructed PRS by P+T using varying *p*-value thresholds and reported the accuracy based on the optimal threshold fine-tuned in the validation cohort. The simulation setup is shown in detail in **Figure S1** and **Methods**.

## PRS predictive accuracy improved with more individuals from target populations included in the multi-ancestry GWAS but varying with genetic architecture

We first explored how different LD reference panels impact PRS predictive accuracy of P+T when the ancestry composition of the multi-ancestry GWAS varied. Specifically, we used three sets of LD reference panels, including two single-ancestry datasets (N=10,000) that matched the ancestry composition of each population contributing to the discovery GWAS, and one combined dataset (N=10,000) with individuals proportional to the ancestry composition of the discovery GWAS. Overall, we observed that the impact of the LD reference panel was subtle for more polygenic traits compared to less polygenic ones (**Figure S2 and Table S1**). When using single-ancestry LD reference panels, we found that using the one matching the majority ancestry in the discovery GWAS provided better predictive performance. Furthermore, we found that such single-ancestry LD reference panels generally provided comparable predictive accuracy to the proportional combined ancestry panel, especially when the ancestry composition was increasingly disproportionate. In particular, the proportional combined ancestry LD panel did not yield significantly better PRS accuracy compared to the optimal single-ancestry LD panel, and minor differences were smallest for the most polygenic scenarios. In our simulation setup, the proportion of understudied populations could go above 50% although this was always not the case in current multi-ancestry GWAS. We hereafter will report the results based on the estimates using the combined LD reference panel to avoid arbitrariness when ancestry proportions for multi-ancestry GWAS are similar.

We observed consistent upward trends of predictive accuracy in the understudied target populations with increasing target-ancestry matched samples included in discovery GWAS (**Figure S2**). Such improvement varied between different genetic architectures. Specifically, we found the accuracy reached a plateau sooner in smaller numbers of bins from minority populations for less polygenic traits with larger per-variant explained variance when compared to more polygenic traits with lower variance explained for each variant.

## PRS predictive accuracy is higher with multi-ancestry GWAS than with single-ancestry GWAS

When constructed PRS using single-ancestry GWAS, we found that using more ancestry-matched GWAS outperformed other discovery populations (**Figure S3**). Compared to using EUR GWAS, the benefit of using ancestry-matched GWAS was generally more obvious for more polygenic traits and larger GWAS. Relative to PRS accuracy attained using EUR GWAS only, we observed substantial accuracy improvements in the target population by including more individuals from the target ancestry in multi-ancestry GWAS; this trend was clearer for more polygenic traits (**Figure 1 and Figure S4**). However, we did not consistently observe such accuracy gains for the majority EUR population, or the other understudied ancestry not included in the multi-ancestry discovery GWAS. In our simulations but unlike in most GWAS, populations typically understudied in current genomic studies can be the majority in the discovery GWAS. Nevertheless, we still observed substantial PRS accuracy improvements when the proportion of understudied populations in the discovery GWAS was less than 50%. We expected to observe similar relative improvements in the target populations using $\text{PRS}_{\text{multi}}$ compared to using EUR GWAS-derived PRS ($\text{PRS}_{\text{EUR\_GWAS}}$) with the same number of bins from EUR populations. Specifically, the relative accuracy here was calculated as the difference in PRS $R^2$ between the PRS derived from multi-ancestry GWAS and EUR GWAS divided by the PRS $R^2$ in EUR ancestry from the EUR GWAS, i.e. $RA = \frac{R^2_{target}\ using\ PRS_{multi} - R^2_{target}\ using\ PRS_{EUR\_GWAS}}{R^2_{EUR}\ using\ PRS_{EUR\_GWAS}}$. Compared with using large-scale EUR only GWAS, we found that multi-ancestry GWAS with much smaller sample sizes could achieve comparable or better predictive accuracy (**Table S1**). Overall, adding fewer individuals from the target populations improved accuracy for less polygenic traits versus more polygenic traits. Similarly, larger sample sizes from AFR populations were required compared to EAS populations especially for more polygenic traits, likely due to the larger effective population size in AFR populations and larger genetic divergence between EUR and AFR populations.

Relative to accuracy using Minor GWAS only, we found that in the ancestry-matched minority population, the accuracy improvement of using multi-ancestry GWAS gradually diminished and remained similar to using Minor GWAS even when the sample size of multi-ancestry GWAS was much larger (**Figure S5 and Table S1**). We showed that in general no obvious improvement was achieved by $\text{PRS}_{\text{multi}}$ when the understudied target populations accounted for more than half sample sizes of the multi-ancestry GWAS except for the least polygenic traits where a much smaller Minor GWAS outperformed multi-ancestry GWAS. Interestingly, we observed consistent accuracy improvements for target populations of EUR and the ancestry not included in the multi-ancestry GWAS when compared to using PRS derived from Minor GWAS ($\text{PRS}_{\text{Minor\_GWAS}}$), although such improvement decreased with larger numbers of bins from minority populations. This could be due to multi-ancestry GWAS being more genetically similar to those populations as compared to Minor GWAS.

## Empirical analysis of PRS accuracy within and across ancestries using 17 quantitative phenotypes

### Genetic architecture of 17 studied phenotypes

To understand how trait genetic architecture influences predictive accuracy of PRS across ancestries, we first estimated several parameters influencing different aspects of genetic architecture for 17 phenotypes in the UKBB and BBJ (**Table S2 and Table S3**). Specifically, we estimated SNP-based heritability, polygenicity ($\pi$, the proportion of SNPs with nonzero effects) and a coefficient of negative selection (*S*, measuring the relationship between MAF and estimated effect sizes) using SBayesS.

The phenotypes included in this study varied widely in genetic architecture across these estimated parameters, with polygenicity estimates ranging from low (e.g., mean corpuscular hemoglobin concentration [MCHC], basophil count [basophil], mean corpuscular hemoglobin [MCH], mean corpuscular volume [MCV]) to high (e.g., height and body mass index [BMI]) (**Figure 2 and Table S3**). SNP-based heritability estimates similarly ranged from <0.1 for basophil and MCHC to 0.54 and 0.33 for height using UKBB and BBJ, respectively, regardless of polygenicity. These polygenicity estimates are relative and cannot be directly interpreted as the number of causal variants. Rather, we used them here to quantify the relative degree of polygenicity between phenotypes with estimates based on the same set of SNPs as well as using marginal effects from GWAS conducted in a consistent manner. The median *S* parameters were -0.63 and -0.47 using UKBB and BBJ, respectively. While the negative *S* values indicate negative selection (i.e., rarer variants have larger effects), it remains unclear to what degree population stratification could confound its estimates[30,31]. We found that the polygenicity estimates using UKBB were mostly higher than those using BBJ, which could be due to the higher statistical power with larger sample sizes in the UKBB resulting in more variants with small effects being detected. Similarly, we observed significantly higher SNP-based heritability in the UKBB compared to BBJ except for MCHC and basophil, indicating possible phenotype heterogeneity between the two cohorts. Specifically, BBJ is a hospital-based cohort with participants recruited with certain diseases, whereas UKBB is a population-based cohort with overall healthier participants. This is also consistent with the previous study using estimates from LD score regression (LDSC) and stratified-LDSC[6]. Moreover, as described previously[6], the estimated cross-ancestry genetic correlations between UKBB and BBJ for those traits were not statistically different from 1 (p-value > 0.05/17) except for a few including basophil (0.5945, SE = 0.1221), height (0.6932, SE = 0.0172), BMI, (0.7474, SE = 0.0230), diastolic blood pressure (DBP, 0.8354, SE = 0.0509), and systolic blood pressure (SBP, 0.8469, SE = 0.0430).

## PRS accuracy using smaller target ancestry-matched GWAS versus larger-scale EUR GWAS may be comparable depending on methodology and trait-specific genetic architecture

We first constructed PRS using P+T and PRS-CS for different phenotypes in the target populations using single-ancestry discovery GWAS from UKBB and BBJ, respectively.

Overall, there was a clear increasing trend in the target populations between PRS accuracy and a larger discovery GWAS (**Figure 3** and **Table S4**). However, such patterns differed by ancestry and PRS methods in a trait-specific manner. For example, the upward trend in the UKBB-EAS was not obviously witnessed for basophil, a rare cell type, when using BBJ. This might be attributable to smaller GWAS sample sizes, ascertainment bias and lower heritability in the BBJ. Moreover, we observed that the more sophisticated method PRS-CS overall significantly outperformed the classic P+T method across traits especially for more polygenic traits and larger sample sizes (one-side Wilcoxon test, *p*-value < 0.05). Specifically, the median accuracy of PRS derived from BBJ in the UKBB-EAS was 0.013 and 0.010 using PRS-CS and P+T, respectively. The corresponding values were 0.046 and 0.032 when the discovery GWAS was UKBB. However, we observed that accuracy of PRS using P+T outperformed PRS-CS for MCH and MCV when BBJ was the discovery GWAS, which could be due to ancestry-enriched variants with large effects for such traits. Further, we showed that for most traits when using full UKBB GWAS with much larger sample sizes provided better predictive accuracy in the UKBB-EAS than using full BBJ. However, for traits such as height, MCV and MCH, using target-ancestry matched GWAS presented consistently better predictive performance but dependent on PRS methods. Specifically, the pattern was witnessed using both P+T and PRS-CS for height but only P+T for MCV and MCH. Moreover, PRS derived from BBJ for those traits with a much smaller sample size achieved similar or even better performance than full UKBB-derived PRS.

Consistent with previous work[6–9], $\text{PRS}_{\text{single}}$ was generally more transferable (as measured by relative accuracy, the ratio of predictive accuracy between target populations) when the target population was more genetically related to the discovery GWAS (**Figure S6**). Interestingly, we observed that in comparison with predictive accuracy, there was no obvious increasing trend between PRS relative accuracy and larger UKBB-based GWAS sample sizes while there was more variation using BBJ-based GWAS due to its smaller sample size and lower SNP-based heritability. These results suggest that the PRS transferability issue is unlikely to be improved by just using larger EUR GWAS.

### Multi-ancestry GWAS-derived PRS usually improves predictive performance relative to single-ancestry GWAS-derived PRS

To explore PRS predictive performance using multi-ancestry GWAS, we meta-analyzed single-ancestry GWAS from UKBB and BBJ. Similar to the simulation setup, we mimicked proportional ancestry composition in the multi-ancestry GWAS by meta-analyzing EUR GWAS from various bins in the UKBB, ranging from 8 to 64 with an increment of 8 (each bin of 5,000), and GWAS in the BBJ (see **Methods**, **Figure S1** and **Table S2**). The ratio of EUR/EAS samples was between 64:1 to 8/$\text{Bin}_{\text{Total}}$ (total number of bins for the specific trait as shown in **Table S2**), thus ~85% multi-ancestry GWAS having a EUR proportion larger than 50%. We performed P+T and PRS-CS using different LD reference panels and evaluated the performance in the target populations.

Similar to the phenomenon we observed in our simulations of predictive accuracy being less affected by the choice of LD reference panel for more polygenic traits, we found that there was only a slight difference between using the combined LD reference panel proportional to the ancestries included in the multi-ancestry GWAS and using the panel matched with the majority population of discovery GWAS for P+T (**Figure S7 and Table S5**). Moreover, the majority of PRS was constructed from GWAS with more EUR individuals; we hereafter reported the results using 1KG-EUR as the LD reference for both P+T and PRS-CS.

Compared to PRS using single-ancestry GWAS from UKBB ($PRS_{EUR\_GWAS}$), we found it was heartening that 99.7% and 92.4% of $PRS_{multi}$ improved predictive accuracy in the UKBB-EAS when using P+T and PRS-CS, respectively (**Table S6 and Figure S8**). With more EAS samples added into the discovery GWAS, we found that the PRS accuracy in the UKBB-EAS also increased (**Figure 4**). For example, the largest absolute accuracy improvements of $PRS_{multi}$ compared to $PRS_{EUR\_GWAS}$ using P+T were 0.038 (0.085 VS 0.047), 0.035 (0.058 VS 0.023) and 0.034 (0.071 VS 0.037) for platelet count (PLT), BMI and height, respectively, when the number of bins from BBJ was or was close to the total number of bins and the number of bins from UKBB was 64. Whilst PRS-CS witnessed corresponding improvements of 0.020 (0.0126 VS 0.101), 0.025 (0.075 VS 0.050) and 0.013 (0.097 VS 0.084) for the three traits. Moreover, P+T showed overall more improvement as compared to PRS-CS regardless of the number of bins from EUR GWAS, with the median $R^2$ improvement being 0.014 and 0.008, respectively. The upward trend was not consistently shown between PRS accuracy in the UKBB-EUR, especially using PRS-CS (**Figure S9 and Table S6**). This pattern was consistent with our simulation results and previous reports that PRS accuracy for the minority populations included in the multi-ancestry GWAS benefited more from adding more ancestry-matched individuals compared to other populations including EUR populations[32]. We noted that the accuracy of $PRS_{multi}$ could remain largely unchanged or slightly decrease when the number of bins from BBJ was small, which was consistent with previous studies[12,32].

## PRS derived from meta-analyzed multi-ancestry GWAS often outperform weighted PRS in understudied populations

Linearly combining PRS constructed from GWAS with different ancestries has also previously been proposed to improve prediction in diverse populations[33]. Here, we constructed the weighted PRS (**$PRS_{weighted}$**) by linearly combined PRS derived from single-ancestry GWAS from UKBB and BBJ (see **Methods**). We then compared the accuracy of $PRS_{multi}$ and $PRS_{weighted}$ using both P+T and PRS-CS.

Among the comparisons in the UKBB-EAS, 91.4% and 78.0% showed accuracy improvement of $PRS_{multi}$ compared to $PRS_{weighted}$ when using P+T and PRS-CS, respectively. We found that $PRS_{multi}$ achieved better performance than $PRS_{weighted}$, especially in the UKBB-EAS (**Figure S10 and Table S7**, *p*-value < 0.05, one-side Wilcoxon test). The median improvement of $PRS_{multi}$ was 0.011 and 0.003 using P+T and PRS-CS, respectively. We observed the largest improvement of $PRS_{multi}$ in the UKBB-EAS using P+T were 0.045 (0.065 VS 0.020) and 0.036 (0.048 VS 0.012) for monocyte count (monocyte) with a ratio of bins from UKBB and BBJ being 56:15 and DBP with bin ratio being 40:25, respectively. While using PRS-CS, we found that the accuracy of $PRS_{multi}$ greatly improved for PLT (0.091 VS 0.073) with bin ratio being 24:1 and lymphocyte (0.044 VS 0.028) with bin ratio being 16:1. We did not observe a consistent pattern between accuracy differences and GWAS sample sizes. Moreover, although overall better performance was shown for $PRS_{multi}$, we found that $PRS_{weighted}$ instead significantly outperformed $PRS_{multi}$ for PLT using P+T (0.086 VS 0.081) and for height using PRS-CS (0.091 VS 0.082). For the accuracy differences between the two PRS strategies in the UKBB-EUR, we observed slight improvement of $PRS_{multi}$ (0.003) using P+T, but higher accuracy of $PRS_{weighted}$ (0.002) using PRS-CS. The different pattern in the UKBB-EAS and UKBB-EUR might be due to the overall higher SNP-based heritability in the UKBB than the BBJ, resulting in more information being borrowed for EAS samples when meta-analyzing with EUR samples. This is also consistent with the multi-trait analyses that those traits with smaller sample sizes and SNP-based heritability benefited more from shared genetic components[34].

## PRS derived from local ancestry-informed GWAS can improve accuracy for some less polygenic traits

We utilized local ancestry-informed summary statistics generated by Tractor[21] from the admixed AFR-EUR individuals to construct PRS in the understudied AFR population across 17 traits. We referred to PRS derived from such local ancestry-informed ancestry specific GWAS summary statistics in AFR ancestry as **$\mathbf{AFR_{Tractor}}$**. Two different PRS methods, P+T and PRS-CS, were used to benchmark performance of ancestry-specific PRS as compared to PRS build off of large-scale traditional summary statistics. Here, we denoted such traditional large-scale EUR GWAS performed with standard linear regression as **$\mathbf{EUR_{Standard}}$**. To compare with PRS performance derived from different GWAS, we further constructed weighted PRS ($PRS_{weighted}$) by leveraging existing large-scale EUR GWAS as well as $AFR_{Tractor}$ and compared with PRS derived from multi-ancestry meta-analyzed GWAS (**$\mathbf{Meta_{Standard}}$**, see **Methods**).

Local ancestry-informed ancestry-specific GWAS had a much smaller sample size relative to the EUR-inclusive GWAS, as is typical for GWAS of underrepresented populations. As expected, we did not observe significant predictive accuracy of PRS derived from such AFR-specific GWAS ($AFR_{Tractor}$) for most traits such as height and BMI (**Figure 5 and Table S8**). Notably, $AFR_{Tractor}$ provided better performing PRS for 5 traits including white blood cell count (WBC), neutrophil count (neutrophil), MCV, MCH and MCHC; their accuracies using P+T were significantly higher than those from using $EUR_{Standard}$ (one-side Wilcoxon test, *p*-value = 0.004) despite $EUR_{standard}$ having much larger-scale discovery data (**Figure 5 and Table S8**). This might be attributable to those traits containing large-effect AFR-enriched variants, especially for MCV, MCH and MCHC, which are captured by Tractor GWAS[12,21]. Consistent with previous findings, P+T overall outperformed PRS-CS for these traits with much sparser genetic architectures. Given that heritability bounds predictive accuracy, which can vary among populations and contexts, we also compared heritability estimates in the Pan-UK Biobank Project (https://pan.ukbb.broadinstitute.org/docs/heritability/index.html) among AFR and EUR populations. Consistent with our PRS accuracy results, we observed higher but not statistically different SNP-based heritability estimated using LDSC in AFR than in EUR for WBC (0.41, SE = 0.19 VS 0.17, SE = 0.01), neutrophil (0.44, SE = 0.26 VS 0.15, SE = 0.01), and MCHC (0.15, SE = 0.11 VS 0.06, SE = 0.01). The lack of statistical difference stems from the large standard error likely due to the small sample size of AFR, although sparser genetic architectures also lead to less stable heritability estimates with LDSC.

We also showed that using weighted linear regression to combine $AFR_{Tractor}$ and $EUR_{Standard}$, improved predictive accuracy for those above-mentioned 5 traits with ancestry-enriched variants.

This result is similar to the findings in previous sections that for some traits with large effect ancestry-enriched variants, weighting PRS by linearly combining discovery GWAS from multiple populations performed better compared to the meta-analysis strategy; for traits without these ancestry-enriched variants, the meta-analysis strategy showed overall higher performance. Specifically, the mean accuracy of $\text{PRS}_{\text{weighted}}$ for those 5 traits was 0.044, 0.031, and 0.028 using P+T, PRS-CS and PRS-CSx, respectively; and the differences between the three PRS construction methods were not significant. The mean accuracy of $\text{Meta}_{\text{Standard}}$ was 0.016 and 0.008 using PRS-CS and P+T, respectively. Lastly, we did not observe significant differences between running standard linear regression with covariates in admixed populations and $\text{AFR}_{\text{Tractor}}$, although it is worth noting that the effective sample size of local ancestry-informed GWAS is ~20% smaller due to the reduction from deconvolving ancestral tracts when generating ancestry-specific GWAS summary statistics. We also note that in-sample LD was usually required for PRS derived from traditional GWAS performed with linear regression in admixed populations with complicated LD structure, whereas we can utilize external LD reference panels for PRS derived from local ancestry-informed GWAS as shown here, eliminating the need for direct access to the individual-level genotypes of admixed populations (**Figure 5 and Table S8**).

## Discussion

In this study, we performed extensive evaluations of PRS performance through both simulation and empirical analyses to explore the impact of ancestry composition, trait-specific genetic architecture and PRS methodology on PRS predictive accuracy and generalizability across populations.

Our simulations demonstrated that predictive accuracy in the understudied target population benefited from increasing genetic diversity of discovery GWAS, and that this pattern varied across trait genetic architectures and ancestry composition. Compared to using EUR GWAS, we showed that there were considerable improvements from adding a smaller proportion of understudied populations for less polygenic traits, whereas for more polygenic traits, accuracy continued to improve more as a function of sample size. Moreover, the generalizability of PRS was also improved by using multi-ancestry GWAS. On the other hand, we found that a much smaller underrepresented target-ancestry matched GWAS could achieve comparable predictive accuracy to a large multi-ancestry GWAS.

We recapitulated the main findings from our simulations in empirical analyses for phenotypes across a range of genetic architectures. Specifically, we showed that the addition of samples from an underrepresented target ancestry - even with small proportions - may improve the predictive accuracy in the target ancestry. However, the extent of the improvement was affected by various factors such as the sample size ratios between EUR GWAS and Minor GWAS, trait genetic architecture, and LD reference panels. Among those factors, between-ancestry genetic architecture differences, in particular, ancestry-enriched variants with large effects, affected accuracy improvement more than sample sizes and LD reference panels. We note that the advantage of PRS constructed from multi-ancestry GWAS is likely to dwindle when the sample size of understudied populations continues to increase. It is still recommended to leverage large-scale EUR GWAS for current scale of understudied populations, although we may not expect accuracy improvement when meta-analyzing extremely small Minor GWAS.

We also found that leveraging information from multiple ancestries by directly meta-analyzing the datasets could improve predictive performance more than linearly combining PRS through an optimized weighting strategy in understudied populations, especially for P+T. This has also been shown using a more sophisticated genome-wide PRS method, PRS-CSx, which jointly analyzes multiple GWAS while accounting for LD from different ancestries[35]. We think improvements from meta-analyzed GWAS could be due to the fact that $\text{PRS}_{\text{multi}}$ implicitly assumed that the causal variants are shared between ancestries, and thus, the underrepresented target ancestry, especially when its SNP-based heritability is lower, borrows more genetic information from the other ancestry with larger sample sizes. Although the predictive performance of $\text{PRS}_{\text{multi}}$ in the UKBB-EAS is better overall with this approach, we note that its accuracy could be affected by the choice of LD reference panel, while $\text{PRS}_{\text{weighted}}$ was not limited by this factor.

We also showed that these findings from simulations and empirical analyses on 17 traits using BBJ and UKBB were largely generalizable when incorporating PRS derived from local ancestry-informed GWAS and large-scale EUR GWAS. Specifically, we found that $\text{PRS}_{\text{weighted}}$ provided overall better performance for traits with ancestry-enriched variants, such as MCHC and MCV, compared to $\text{PRS}_{\text{multi}}$. We have shown the advantage of leveraging GWAS in admixed populations by accounting for local-ancestry, and without direct access to individual genotypes of admixed populations to improve PRS predictive performance in understudied populations. However, the sample size of admixed individuals here was relatively small, and we expect that further guidance on optimal PRS strategies for improved generalizability using PRS derived from local ancestry-informed GWAS will follow from future analyses of datasets with larger sample size such as *All of Us*.

While some previous studies have shown the benefits of leveraging increasing genetic diversity to improve PRS accuracy in global populations[14,36], most have used GWAS with primarily European ancestry. In this study, we have provided additional best practices for developing PRS for understudied populations using different discovery cohorts, particularly when GWAS have different ancestry compositions across various trait genetic architectures (**Figure 6**). Our suggestions focus on general guidelines when constructing $\text{PRS}_{\text{single}}$ and $\text{PRS}_{\text{multi}}$ (or $\text{PRS}_{\text{weighted}}$) depending on genetic architecture, ancestry composition, sample sizes and statistical power, PRS methodology, and LD reference panels.

First, when developing $\text{PRS}_{\text{single}}$, the choice of input GWAS, i.e., whether using large-scale EUR GWAS or using underrepresented target-ancestry matched GWAS, is dependent on cross-ancestry genetic correlation ($r_g$), SNP-based heritability in discovery ($h_d^2$) and target populations ($h_t^2$), discovery GWAS sample size ($N_d$) and the number of genome-wide independent segments in the discovery population ($M_d$). We further illustrate the relationship between PRS accuracy and single-ancestry discovery GWAS sample size for traits studied here in **Figure S11**. For traits with relatively low $r_g$ and a sizable ancestry-matched GWAS (e.g., > 20-40% of EUR GWAS), such as

BMI and height, PRS accuracy in the target population benefits from using ancestry-matched GWAS; for traits with high $r_g$ and SNP-based heritability, larger-scale EUR GWAS will likely perform better than smaller-scale ancestry-matched GWAS. We note that these results could be affected by characteristics of the target cohort and phenotype precision. We provide a theoretical equation to estimate the expected accuracy using discovery GWAS with ancestry different from target population, thus enabling comparisons with accuracy using EUR GWAS based on prior information of different parameters. We expect that Bayesian methods adaptive to trait genetic architecture are expected to show better performance compared to classic P+T methods unless there are target ancestry-enriched variants or traits with very sparse genetic architecture, as shown in previous studies[36–39].

Second, relative to $\text{PRS}_{\text{single}}$ using EUR GWAS, we recommend using $\text{PRS}_{\text{multi}}$ except when the target ancestry-matched GWAS is extremely small. We showed that there was little to no improvement comparing $\text{PRS}_{\text{multi}}$ to $\text{PRS}_{\text{single}}$ when the sample size from the target population was only a few thousands (e.g., < 10,000). The theoretical equation derived for cross-ancestry prediction mentioned above is also applicable for prediction using multi-ancestry GWAS. Therefore, $\text{PRS}_{\text{multi}}$ is also generally preferred for traits with high $r_g$ and SNP-based heritability and large sample size. There is increasing evidence showing that most common variants are shared between-ancestries, thus supporting high cross-ancestry $r_g$ for most traits[7,16]. However, estimates of $r_g$ can be affected by phenotypic and environmental heterogeneity between different populations[15,40]. A consideration when constructing PRS based on multi-ancestry GWAS using summary-level based methods, such as P+T and PRS-CS, is which LD reference panel best approximates the LD structure between SNPs while being most readily available to researchers. We have shown that when EUR is still the majority population in the discovery GWAS, using the EUR-based reference panel can approximate the LD of discovery GWAS well compared to a combined panel with multiple ancestries proportional to the discovery GWAS, which are consistent with our previous findings[14].

Third, although it is common practice to develop weighted linear combinations of PRS from ancestry-specific GWAS due to the easy access to external ancestry-matched LD reference panels, we suggest constructing PRS using multi-ancestry GWAS rather than through linear combinations based on our results. The difference between these two strategies was subtler using PRS-CS with some notable exceptions, including higher accuracy with $\text{PRS}_{\text{weighted}}$ for traits with low $r_g$ such as height. We also showed that $\text{PRS}_{\text{weighted}}$ outperformed $\text{PRS}_{\text{multi}}$ in the UKBB-AFR for traits with AFR-enriched variants, such as WBC and MCHC, when incorporating local ancestry-informed GWAS and large-scale EUR GWAS. More practically, $\text{PRS}_{\text{weighted}}$ is more efficient which can directly use PRS weights from resources such as PGS Catalog[41].

In summary, there is no one-size-fits all method or approach for constructing PRS, as the optimal approach depends on genetic architecture, ancestry composition, statistical power, and other factors. These factors can be complex, particularly as a deluge of methods are being developed to address the PRS generalizability problem. To inform optimal approaches across a wide variety of scenarios, we have distilled the results of a wide range of simulations and empirical analyses across trait genetic architectures, ancestries, and methods into a set of guidelines from parameters that are typically evaluated at the outset of a genetic study.

## Limitations of the study

Last but not least, we note a few limitations and future directions in our study. First, we are focused on common variants present in different populations, while population-enriched variants by definition have lower frequencies and larger effect sizes in some populations. The role of such variants on polygenic prediction are worth exploring across phenotypes when there are sufficient sample sizes for different ancestral populations. We have shown that for traits with target ancestry-enriched variants where their effect sizes are larger in minority populations, substantially smaller target-ancestry matched GWAS can yield comparable or better predictive performance than using larger-scale EUR GWAS. This highlights again the importance of diversifying genomic studies. Second, as we used external LD reference panels for PRS construction, PRS performance decreases with LD mismatch between the discovery population and LD reference panel, especially using multi-ancestry GWAS. While we show that LD reference panel differences have a relatively modest effect on PRS accuracy, they have a much larger effect on fine-mapping[42], so future efforts are warranted to share in-sample LD without direct access to individual-level genotypes, especially for large consortia with numerous and diverse cohorts. Alternatively, developing more sophisticated individual-level PRS methods that preserve privacy and are scalable to current biobank-scale genomics data is also promising. Third, we are focused on quantitative phenotypes with a range of genetic architectures, but we expect the findings are generally applicable to binary traits, as we have investigated previously[14]. However, there are some caveats for studying binary phenotypes which may be more susceptible to different factors, such as variable case/control ratios, phenotype definitions, environmental differences, and smaller effective sample sizes or lower statistical power. Fourth, we have provided theoretical expectations of cross-ancestry prediction, but they are to some extent limited by reliable estimates for different parameters such as cross-ancestry genetic correlation and the effective number of independent genome-wide segments, which can prove especially challenging to estimate for multi-ancestry discovery GWAS with highly imbalanced sample sizes. We also observed a discrepancy between expected and observed accuracies. The most straightforward explanation might be that the assumptions of trait genetic architecture are different between PRS construction methods and theory. Thus, expected accuracy models should be adaptive to trait-specific genetic architecture. Finally, as there is no one-size-fits-all method, we focus on P+T and PRS-CS in this study. Although we show that trends are generally consistent between the two methods and we expect they are mostly generalizable to other methods, there are still some slight differences especially regarding the choice of using meta-analysis and weighted strategies. Despite the limitations, our findings have shown the benefits of leveraging increasing diversity of current genomics studies to improve polygenic prediction across populations. We also highlight the necessity of diversifying the ancestry as well as phenotype spectrum when collecting genomics data from global populations to achieve more equitable use of PRS for traits with varying genetic architectures.

## Acknowledgements

A.R.M. and Y.W. were supported by funding from the National Institutes of Health (K99/R00MH117229 to A.R.M.) as well as funding from European Union's Horizon 2020 research and innovation program under grant agreement 101016775. A.R.M. was also supported by funding from U01HG011719.

## Declaration of interests

All authors declare no competing interests.

## Figure Legends

### Figure 1. Predictive accuracy improvement of PRS using meta-analyzed multi-ancestry (EUR and EAS) GWAS compared to using EUR GWAS in 6 simulated genetic architectures.

We illustrated the results using 32 EUR bins as an example. PRS was evaluated in AFR, EAS and EUR, respectively. Full results are shown in **Table S1**. The red vertical dashed line in each panel indicates the point where the number of bins from EUR and EAS populations is the same. The black horizontal dashed line indicates y=0. The error bars represent the standard errors of predictive accuracy differences using PRS derived from multi-ancestry GWAS ($PRS_{multi}$) and EUR GWAS ($PRS_{EUR\_GWAS}$), respectively.

### Figure 2: Genetic architecture of 17 studied traits between Biobank Japan (BBJ) and UK Biobank (UKBB).

The phenotypes were ranked according to their polygenicity estimates using GWAS from UKBB. The error bar was the standard deviation of the corresponding estimate. Trait abbreviations are shown in **Table S2**. The vertical dashed line was the median estimate. Full results are shown in **Table S3**.

### Figure 3. Predictive performance of 17 traits in the UKBB-EAS using P+T and PRS-CS.

We used GWAS from both Biobank Japan (BBJ) and UK Biobank (UKBB) to construct PRS. We reported the predictive accuracy in the UKBB-EAS using the auto model for PRS-CS and optimal *p*-value for P+T (see **Methods**). We showed the results for 7 traits with SNP-based heritability > 0.1 in both BBJ and UKBB, while they were ranked by polygenicity estimated using UKBB (**Figure 2**). Trait abbreviations are all described further in **Table S2**. Full results for all traits are shown in **Table S4**.

## Figure 4. Accuracy improvement of PRS in the UKBB-EAS using multi-ancestry GWAS relative to using EUR GWAS for P+T and PRS-CS.

We constructed PRS using P+T and PRS-CS and evaluated them in the UKBB-EAS. The y-axis is the accuracy difference of PRS between using multi-ancestry GWAS ($\text{PRS}_{\text{multi}}$) and using EUR GWAS ($\text{PRS}_{\text{EUR\_GWAS}}$) when the number of bins from EUR GWAS is 64. The x-axis is the number of bins from BBJ included in the multi-ancestry GWAS. The error bars indicate the standard error of mean accuracy improvement. The red dashed line is y=0. The red dashed line is y=0. We showed the results for 7 traits with SNP-based heritability > 0.1 in both Biobank Japan (BBJ) and UK Biobank (UKBB), while they were ranked by polygenicity estimates using UKBB (**Figure 2**). Trait abbreviations are all described further in **Table S2**. Full results are shown in **Table S6**.

## Figure 5. Predictive accuracy for P+T and PRS-CS/PRS-CSx in the UK Biobank African population using various discovery GWAS.

$\text{AFR}_{\text{Tractor}}$ denotes the AFR-specific GWAS performed using Tractor. $\text{EUR}_{\text{Standard}}$ refers to standard GWAS performed in the European population in the UKBB. $\text{Meta}_{\text{Standard}}$ is the meta-analysis performed on $\text{AFR}_{\text{Tractor}}$ and $\text{EUR}_{\text{Standard}}$. The weighted PRS was constructed through a linear combination of PRS generated from $\text{AFR}_{\text{Tractor}}$ and $\text{EUR}_{\text{Standard}}$, respectively, using various methods including P+T and PRS-CS. Further, we also constructed weighted PRS using PRS-CSx where the input GWAS were $\text{AFR}_{\text{Tractor}}$ and $\text{EUR}_{\text{Standard}}$. This figure shows the results for traits with SNP-based heritability > 0.1 in the UK Biobank African population (UKBB-AFR); full results are shown in **Table S8**.

## Figure 6. General practices for developing PRS using different discovery GWAS.

We summarized the general practice for developing PRS A) using single-ancestry GWAS ($\text{PRS}_{\text{single}}$); and B) using GWAS from multiple ancestries ($\text{PRS}_{\text{multi}}$ or $\text{PRS}_{\text{weighted}}$). For $\text{PRS}_{\text{single}}$, we can compare the expected accuracies either using underrepresented target-ancestry matched GWAS (Minor GWAS) or large-scale European-based GWAS (EUR GWAS) and choose the input GWAS for PRS method based on prior information including cross-ancestry genetic correlation ($r_g$), SNP-based heritability in discovery ($h_d^2$) and target populations ($h_t^2$), discovery GWAS sample size ($N_d$) and the number of genome-wide independent segments in the discovery population ($M_d$). For $\text{PRS}_{\text{multi}}$, meta-analysis is generally recommended whilst the linear weighted combination shows its superiority for traits with ancestry-enriched variants.

# Methods

## Simulations

### Simulated genotypes in three populations

To explore whether the predictive accuracy in the underrepresented target ancestry could be improved with additional samples included in the multi-ancestry discovery GWAS, we simulated genotypes of chromosome 22 for 560,000 individuals in each population including European ancestry (EUR), East Asian ancestry (EAS) and African ancestry (AFR) using the software HapGen2[43]. We used the haplotypes from 1000 Genome Project (1KG, Phase 3)[44] as the sample pool. We excluded Americans of African Ancestry in SW USA and African Caribbeans in Barbados from the AFR samples due to their high degree of recent admixture. We used default parameters in HapGen2 with effective sample sizes of 11,375, 12,239 and 17,380 for EUR, EAS and AFR, respectively[43]. After simulating the genotypes on chromosome 22, we ran analyses with a total of 87,938 overlapping SNPs across the three ancestries which passed quality control filters: minor allele frequency (MAF) > 0.01, Hardy-Weinberg Equilibrium (HWE) $p$-value > $10^{-6}$ and genotype missingness rates across individuals < 0.05. We then removed 2nd-degree related individuals using the software KING[45], resulting in 534,352, 533,996 and 537,498 unrelated individuals from EUR, EAS and AFR, separately. We randomly sampled 10K and 520K individuals from each ancestry as the withheld target population and discovery population, respectively.

### Simulated phenotypes with varying trait genetic architecture

For the sake of simplicity, we assumed that causal variants are shared across populations and their effect sizes are perfectly correlated ($r_g = 1$). We simulated phenotypes based on the simple additive model: $y = g + e$, where $g = \sum_{j=1}^{M_c} x_{ij}\beta_j$. $M_c$ is the number of causal variants, $x_{ij}$ is the genotype coded as 0, 1, or 2 for the $j$th SNP in the $i$th population. The effect size of $j$th SNP is drawn from a multivariate normal distribution, $\beta_j \sim MVN(0, \Sigma)$,where the diagonal and off-diagonal elements of $\Sigma$ were $\frac{h^2}{2f_{ij}(1-f_{ij})M_C}$ and $r_g \times \frac{h^2}{2f_{ij}(1-f_{ij})M_C}$, respectively. We denoted $f_{ij}$ as the MAF of $j$th SNP in the $i$th population and $h^2$ as the trait heritability. We simulated the environmental effects to follow a normal distribution with 0 mean and $1 - h^2$ variance, $e \sim N(0, 1 - h^2)$. We simulated different levels of heritability for chromosome 22 ($h^2 =$ 0.03, 0.05) and various numbers of causal variants ($M_c$ = 100, 500, 1000) randomly sampled from all the 87,938 SNPs, resulting in a total of 6 simulation scenarios that span a range of realistic polygenicity from ~0.1% to ~1% causal variants.

### Downsampling GWAS

We split the 520,000 unrelated individuals included in the discovery population into 52 evenly distributed bins (each with N =10,000). We labeled each bin from 1 to the total number of bins ($\text{Bin}_{\text{total}}$ = 52), i.e., $\text{Bin}_1$, $\text{Bin}_2$, …, $\text{Bin}_{\text{total}}$. We ran GWAS using simple linear regression implemented in PLINK v2.0[46] in each of those 52 bins in the three populations, respectively. We excluded the causal variants when running GWAS to mimic the phenomenon of imperfect tagging. We then iteratively meta-analyzed a different number of bins using inverse-variance weighted meta-analysis in METAL[47]. Specifically, we first ran meta-analyses on $\text{Bin}_1$+$\text{Bin}_2$, $\text{Bin}_1$+$\text{Bin}_2$+$\text{Bin}_3$, …, and $\text{Bin}_1$+$\text{Bin}_2$+$\text{Bin}_3$+...+$\text{Bin}_{\text{total}}$ in each population.

To mimic a multi-ancestry meta-analysis scenario with different proportions of ancestries, we arbitrarily selected a subset of bins from EUR GWAS, ranging from 4 to 52 bins with increments of 4. We iteratively added different numbers of bins, ranging from 1 to 52 in EAS and AFR, respectively, into EUR GWAS through meta-analysis using the inverse-variance weighted fixed effects model implemented in METAL. By doing this, the ratio of sample sizes of EUR/EAS and EUR/AFR included in the meta-analyzed multi-ancestry GWAS (**Meta**) ranged from 52:1 to 4:52. This simulation setup is illustrated in **Figure S1**.

### LD clumping (P+T)

We used PLINK v1.90 to clump quasi-independent SNPs with LD $r^2 < 0.1$ in 500Kb windows. We tested a total of four different LD reference panels (one for single-ancestry and three for multi-ancestry GWAS) with consideration to the ancestry composition of the discovery GWAS and target population to explore the impact of various LD reference panels on predictive accuracy. For the single-ancestry GWAS, we used the 10,000 withheld ancestry-matched target populations as the LD reference panel. For the multi-ancestry GWAS, we used three LD reference panels. Specifically, we used two LD reference panels composed of a single ancestry that did not mirror the makeup of the discovery GWAS, including one panel of 10,000 withheld EUR individuals and the other from understudied populations (either 10,000 EAS or 10,000 AFR in this study). The third panel consisted of individuals from different ancestries that were proportional to discovery GWAS with a total of 10,000 samples. We calculated PRS in the withheld target population using 8 different *p*-value thresholds: $5 \times 10^{-8}$, $1 \times 10^{-6}$, $1\times 10^{-4}$, $1 \times 10^{-3}$, 0.01, 0.05, 0.1, and 1. We denoted PRS constructed from single-ancestry GWAS as single-ancestry PRS ($\text{PRS}_{\text{single}}$) and those from meta-analyzed multi-ancestry GWAS as multi-ancestry PRS ($\text{PRS}_{\text{multi}}$). We calculated the predictive accuracy as the variance explained by the PRS ($R^2$) through linear regression: $y \sim$ PRS and computed the corresponding 95% confidence intervals (CIs) through bootstrap. When selecting the optimal *p*-value threshold with the highest predictive accuracy, we evenly split the target population into a test cohort and a validation cohort. We hyper-tuned the *p*-value threshold in the validation cohort and evaluated the accuracy in the test cohort.

## Empirical analysis of 17 quantitative traits in the UKBB and BBJ

We further explored how the findings from simulations generalized in real data using 17 quantitative traits shared between UKBB and BBJ, including anthropometric traits (BMI and height) and blood panel traits studied previously (**Table S2**)[6]. We investigated these traits due to their widespread availability in biobanks as well as their high statistical power given their quantitative nature.

## Datasets and Quality Control (QC)

### UK Biobank (UKBB)

The details of assigning ancestry for each individual in the UKBB are described in the Pan-UK Biobank Project (Pan UKBB: https://pan.ukbb.broadinstitute.org/). Briefly, a random forest classifier trained on reference data from 1KG and Human Genome Diversity Project (HGDP)[48] was used to classify cohort individuals under continental population labels based on the top 6 principal components (PCs). In this study, we used a total of 361,144 and 2,684 unrelated EUR and EAS participants, respectively. We obtained unrelated individuals through running hl.maximal_independent_set using Hail (https://hail.is/). Specifically, within each population, we ran PC-Relate[49] with k=10 and min_individual_maf=0.05. We used the individuals assigned EAS ancestry as the target dataset. For EUR samples, we first randomly withheld 5,000 individuals with complete phenotype information for all 17 studied phenotypes as the target population. We split the remaining individuals into evenly distributed bins (each of N = 5,000) for each phenotype. The number of total bins for each studied phenotype ranged from 68 to 71 according to phenotype missingness (**Table S2**). We labeled each bin from 1 to the total number of bins in the same way as described in simulations.

### BioBank Japan (BBJ)

BBJ is a multi-institutional hospital-based biobank which has recruited approximately 200,000 participants from 12 medical institutions in Japan between fiscal years 2003 and 2007[28]. Written informed consents were obtained from all the participants, as approved by the ethics committees of the RIKEN Center for Integrative Medical Sciences, and the Institute of Medical Sciences, the University of Tokyo. The participants were genotyped using either (i) the Illumina HumanOmniExpressExome BeadChip or (ii) a combination of the Illumina HumanOmniExpress and HumanExome BeadChips. The genotypes were then prephased using Eagle[50] and imputed using Minimac3[51] with a reference panel that consists of 1KG samples (N = 2,504) and whole-genome sequencing (WGS) data of Japanese individuals (N = 1,037)[52]. Standard quality controls of participants and genotypes were applied as described elsewhere[52]. Briefly, we excluded samples with low call rates (< 98%), closely related individuals (PLINK PI_HAT > 0.175), or non-Japanese outliers based on the PCA. We then excluded genotyped variants with call rate < 98%, HWE $P$-value $< 1.0 \times 10^{-6}$, number of heterozygotes < 5, or low concordance rate (< 99.5%) with WGS for a subset of individuals (N = 939). Phenotypes were retrieved from medical records and prepared as described previously[53].

### 1000 Genomes Project Phase 3 (1KG)

We used 1KG phase 3 data as LD reference panels in this study. Specifically, we kept 495 unrelated EUR, 498 unrelated EAS, and 484 unrelated AFR individuals from 1KG. The AFR individuals were used in the recently admixed population analysis only.

### Quality Control

The imputation strategies for UKBB and BBJ have been described in detail elsewhere[54,55]. After imputation, we first excluded ambiguous variants (e.g., A/T and C/G) and further filtered to keep those variants with imputation INFO score > 0.3, MAF > 0.01, HWE $p$-value > $10^{-6}$, and genotyping missing rates across individuals < 0.05. A total of ~8.6M and ~6.6M SNPs were retained for UKBB and BBJ, respectively. We used SNPs passing these quality controls in our analyses, resulting in ~3.6M SNPs that overlapped between two biobanks and 1KG.

## PRS construction

### Discovery GWAS

All phenotypes were curated and transformed to be normally distributed as described previously[6]. We then performed GWAS on the rank normalized phenotypes using simple linear regression implemented in PLINK v2.0. We included age, sex, $age^2$, age × sex, $age^2$ × sex, and the first 20 PCs as the covariates. Similar to the GWAS strategy described in *Simulations*, we first ran GWAS in each bin and then iteratively meta-analyzed different numbers of bins using inverse-variance weighted meta-analysis in METAL in the UKBB and BBJ, respectively. When meta-analyzing the single-ancestry GWAS from UKBB and BBJ (denoted as **Meta**), the number of bins from EUR GWAS we used for 17 traits ranged from 8 to 64 with an increment of 8 and we iteratively added bins from GWAS in the BBJ.

### PRS construction methods

We used two methods to construct PRS in the target populations (UKBB-EAS and UKBB-EUR) including P+T, as described in *Simulations*, and PRS-CS[39], which infers posterior mean effects of SNPs by placing a continuous shrinkage prior through a Bayesian regression framework. To reduce the overall computational burden, we first ran PRS-CS using GWAS summary statistics from UKBB with varying numbers of bins (from 8 to 64, with an increment of 8) for 17 traits. We explored how the hyper-parameter (*phi*, the proportion of SNPs with nonzero effects) affects PRS performance with different GWAS sample sizes as well as trait genetic architectures. Specifically, we ran both the grid model with various *phi* parameters ($1\times 10^{-6}$, $1\times 10^{-4}$, 0.01 and 1) and the auto model which automatically estimates the *phi* parameter based on the input GWAS. We used default settings for all other parameters. We found that in the UKBB, PRS-CS-auto provided comparable predictive accuracy across all traits compared to using the optimal phi parameter in the grid model (**Figure S12**). Therefore, we used the PRS-CS-auto model for BBJ and Meta to construct PRS when using PRS-CS. We used LD reference panels in ancestry-matched populations from 1KG for $PRS_{single}$. For $PRS_{multi}$, we used 1KG-EUR as the LD reference panel.

To further explore the performance of PRS leveraging discovery GWAS from multiple ancestries, we used a previously developed method by linearly combining PRS based on optimized weights[33]. Specifically, the weighted PRS is calculated as $\mathbf{PRS_{weighted}} = w_1 * PRS_{UKBB} + w_2 * PRS_{BBJ}$. The weights $w_1$ and $w_2$ were optimal incremental $R^2$ in the validation cohort where we split the target population into two even parts.

### PRS performance evaluation

We used the incremental $R^2$ from the linear regression after regressing out the impact of covariates to evaluate the predictive accuracy. We computed the corresponding 95% confidence intervals (CIs) through bootstrap.

### Measures of genetic architecture using summary-data-based BayesS (SBayesS)[56]

To better understand the impact of trait genetic architecture on PRS predictive performance, we evaluated three parameters including the polygenicity ($\pi$, proportion of SNPs with nonzero effects), SNP-based heritability and $S$ (the relationship between MAF and effect sizes) for 17 studied phenotypes using SBayesS implemented in the software GCTB (https://cnsgenomics.com/software/gctb/) (**Table S2).** We used meta-analyzed GWAS across the full UKBB and BBJ datasets. We used the LD reference panel provided by GCTB for UKBB GWAS. We constructed a shrunk LD matrix using 50,000 unrelated individuals from BBJ as the LD reference panel for BBJ GWAS. We used 4 chains for the MCMC process which calculated the Gelman-Rubin convergence diagnostic (also known as potential scale reduction factor) for these three parameters. We performed the analyses using other default settings for SBayesS. As Bayesian models might suffer from convergence issues, we considered a threshold < 1.2 of the Gelman-Rubin convergence diagnostic as good convergence for the estimated parameters.

## UK Biobank recent admixture ancestry analysis

To investigate one explanation for poor transferability of PRS across populations – genetic divergence between the discovery and target cohorts – we further explored whether PRS constructed from ancestry-specific summary statistics generated with local ancestry-informed GWAS in admixed populations improves predictive performance in underrepresented populations. Specifically, we used the Tractor method[21], accounting for both local ancestry and risk allele information, to run GWAS in two-way admixed AFR-EUR individuals from the UKBB (N = 4,576). The average AFR proportion was 62.9%. We used 4,022 unrelated relatively homogeneous AFR individuals, which are independent from the admixed individuals, as the target cohort.

We followed the same criteria for QC and individual selection as described in Atkinson et al.[21]. For sample QC, we excluded individuals that had <95% call rate, withdrew from the study, had closer than 2nd degree relatives present in the sample, or that had sex chromosome aneuploidies. For variant QC we restricted to biallelic SNPs with >90% call rate, Hardy-Weinberg Equilibrium $p$ value > $10^{-6}$, and MAF of at least 0.5%. We selected two-way admixed AFR-EUR individuals from the UKBB by first using the PC loadings from the reference dataset described previously for ancestry inference (1KG + HGDP) to project UKBB individuals into the same PC space. We applied the same random forest ancestry classifier described previously to the projected UK Biobank PCA data and assigned AFR ancestry if the probability was >50%. We restricted to only two-way admixed AFR-EUR ancestry individuals by selecting those individuals assigned the 'AFR' population label, then filtering to those with at least 12.5% European ancestry, at least 10%

African ancestry, and who did not deviate more than 1 standard deviation from the AFR-EUR cline based on their PC loadings. This resulted in 4,576 individuals.

We ran local ancestry deconvolution on this set of admixed individuals using RFmix v2[20] with 1 EM iteration and a window size of 0.2 cM with the HapMap combined recombination map[57] to inform switch locations. The -n 5 flag (terminal node size for random forest trees) was included to account for an unequal number of reference individuals per reference population. We used the --reanalyze-reference flag, which recalculates admixture in the reference samples for improved ability to distinguish ancestries. As a reference panel, we used continental AFR and EUR individuals from the 1KG.

We then ran Tractor GWAS for those 17 quantitative traits on these UKBB admixed AFR-EUR individuals, which generates ancestry-specific summary statistics for the AFR ($\mathrm{AFR}_{\mathrm{Tractor}}$) and EUR ($\mathrm{EUR}_{\mathrm{Tractor}}$) ancestry components. We compared the PRS performance when calculating using these ancestry-specific effect size estimates versus standard GWAS methods in an admixed discovery cohort by performing GWAS in the same set of admixed individuals using the simple linear regression model as described previously ($\mathrm{ADM}_{\mathrm{Standard}}$). To compare to common practices in statistical genetics, we also used GWAS summary statistics using the UKBB EUR GWAS ($\mathrm{EUR}_{\mathrm{standard}}$, N = 320,000) from previous section and meta-analyzed $\mathrm{AFR}_{\mathrm{Tractor}}$ with $\mathrm{EUR}_{\mathrm{standard}}$ ($\mathrm{Meta}_{\mathrm{standard}}$, N = 324,576).

We constructed PRS based on HapMap3 SNPs for P+T and PRS-CS, as previous work showed similar performance with P+T using reliable HapMap3 SNPs only to using genome-wide SNPs[14]. Given the ancestry composition of discovery GWAS, we used different sets of reference panels for various discovery GWAS. Specifically, we used 1KG-EUR as the LD reference panel for $\mathrm{EUR}_{\mathrm{Tractor,}}$ $\mathrm{EUR}_{\mathrm{standard}}$ and $\mathrm{Meta}_{\mathrm{standard}}$, and 1KG-AFR for $\mathrm{AFR}_{\mathrm{Tractor}}$. We used an in-sample LD panel for $\mathrm{ADM}_{\mathrm{Standard.}}$ We optimized *p*-value thresholds for P+T and *phi* parameters for PRS-CS, respectively, in the validation cohort. To leverage information from multi-ancestry GWAS, we also constructed weighted PRS using GWAS of $\mathrm{AFR}_{\mathrm{Tractor}}$ and $\mathrm{EUR}_{\mathrm{Standard}}$, for P+T and PRS-CS, respectively. We further compared the weighted PRS to that using PRS-CSx which accounts for between-ancestry LD. We evenly split the target AFR cohort into two random sets to serve as independent validation and test datasets. We calculated the predictive accuracy using incremental $R^2$ as previously described. We repeated the process 100 times and reported the standard error of predictive accuracy across 100 estimates.

# Data and code availability

1000 Genome Phase 3 data can be accessed at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data. We used UK Biobank data via application 31063. The software used in this study can be found at: Plink (https://www.cog-genomics.org/plink/), PRS-CS (https://github.com/getian107/PRScs), PRS-CSx (https://github.com/getian107/PRScsx), Tractor (https://github.com/Atkinson-Lab/Tractor), and SBayesS/GCTB (https://cnsgenomics.com/software/gctb/). The Pan UK Biobank Project can be accessed at: Pan-UK Biobank Project https://pan.ukbb.broadinstitute.org. The codes used in this study have been deposited to https://github.com/ywangleo/multi-ancestry-PRS.

# References

1. Inouye, M., Abraham, G., Nelson, C.P., Wood, A.M., Sweeting, M.J., Dudbridge, F., Lai, F.Y., Kaptoge, S., Brozynska, M., Wang, T., et al. (2018). Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J. Am. Coll. Cardiol. *72*, 1883–1893. 10.1016/j.jacc.2018.07.079.

2. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224. 10.1038/s41588-018-0183-z.

3. Mars, N., Widén, E., Kerminen, S., Meretoja, T., Pirinen, M., Della Briotta Parolo, P., Palta, P., FinnGen, Palotie, A., Kaprio, J., et al. (2020). The role of polygenic risk and susceptibility genes in breast cancer over the course of life. Nat. Commun. *11*, 6383. 10.1038/s41467-020-19966-5.

4. Maas, P., Barrdahl, M., Joshi, A.D., Auer, P.L., Gaudet, M.M., Milne, R.L., Schumacher, F.R., Anderson, W.F., Check, D., Chattopadhyay, S., et al. (2016). Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. JAMA Oncol *2*, 1295–1302. 10.1001/jamaoncol.2016.1025.

5. Craig, J.E., Han, X., Qassim, A., Hassall, M., Cooke Bailey, J.N., Kinzy, T.G., Khawaja, A.P., An, J., Marshall, H., Gharahkhani, P., et al. (2020). Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. Nat. Genet. *52*, 160–166. 10.1038/s41588-019-0556-y.

6. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591. 10.1038/s41588-019-0379-x.

7. Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P.M., and Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat. Commun. *11*, 3865. 10.1038/s41467-020-17719-y.

8. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse

human populations. Nat. Commun. *10*, 3328. 10.1038/s41467-019-11112-0.

9. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649. 10.1016/j.ajhg.2017.03.004.

10. Cunningham, T.J., Croft, J.B., Liu, Y., Lu, H., Eke, P.I., and Giles, W.H. (2017). Vital Signs: Racial Disparities in Age-Specific Mortality Among Blacks or African Americans - United States, 1999-2015. MMWR Morb. Mortal. Wkly. Rep. *66*, 444–456. 10.15585/mmwr.mm6617e1.

11. Hales, C.M. (2020). Prevalence of Obesity and Severe Obesity Among Adults: United States, 2017-2018 (U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics).

12. Majara, L., Kalungi, A., Koen, N., Zar, H., Stein, D.J., Kinyanda, E., Atkinson, E.G., and Martin, A.R. (2021). Low generalizability of polygenic scores in African populations due to genetic and environmental diversity. Cold Spring Harbor Laboratory, 2021.01.12.426453. 10.1101/2021.01.12.426453.

13. Ruan, Y., Anne Feng, Y.-C., Chen, C.-Y., Lam, M., Sawa, A., Martin, A.R., Qin, S., Huang, H., Ge, T., and Initiatives, S.G.A. (2021). Improving polygenic prediction in ancestrally diverse populations. bioRxiv. 10.1101/2020.12.27.20248738.

14. Wang, Y., Namba, S., Lopera, E., Kerminen, S., Tsuo, K., Läll, K., Kanai, M., Zhou, W., Wu, K.-H., Favé, M.-J., et al. (2021). Global biobank analyses provide lessons for computing polygenic risk scores across diverse cohorts. bioRxiv. 10.1101/2021.11.18.21266545.

15. Guo, J., Bakshi, A., Wang, Y., Jiang, L., Yengo, L., Goddard, M.E., Visscher, P.M., and Yang, J. (2021). Quantifying genetic heterogeneity between continental populations for human height and body mass index. Sci. Rep. *11*, 5240. 10.1038/s41598-021-84739-z.

16. Shi, H., Burch, K.S., Johnson, R., Freund, M.K., Kichaev, G., Mancuso, N., Manuel, A.M., Dong, N., and Pasaniuc, B. (2020). Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. Am. J. Hum. Genet. *106*, 805–817. 10.1016/j.ajhg.2020.04.012.

17. Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B.J., Loohuis, L.O., and Pasaniuc, B. (2022). Polygenic scoring accuracy varies across the genetic ancestry continuum in all human populations. bioRxiv, 2022.09.28.509988. 10.1101/2022.09.28.509988.

18. Pfaff, C.L., Parra, E.J., Bonilla, C., Hiester, K., McKeigue, P.M., Kamboh, M.I., Hutchinson, R.G., Ferrell, R.E., Boerwinkle, E., and Shriver, M.D. (2001). Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am. J. Hum. Genet. *68*, 198–207. 10.1086/316935.

19. Pritchard, J.K., and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. *69*, 1–14. 10.1086/321275.

20. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a

discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288. 10.1016/j.ajhg.2013.06.020.

21. Atkinson, E.G., Maihofer, A.X., Kanai, M., Martin, A.R., Karczewski, K.J., Santoro, M.L., Ulirsch, J.C., Kamatani, Y., Okada, Y., Finucane, H.K., et al. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. Nat. Genet. *53*, 195–204. 10.1038/s41588-020-00766-y.

22. Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G.K., Tandon, A., Kao, W.H.L., Ruczinski, I., Fornage, M., Siscovick, D.S., Zhu, X., et al. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS Genet. *7*, e1001371. 10.1371/journal.pgen.1001371.

23. Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W.J., Khera, A.V., Okada, Y., Biobank Japan Project, Martin, A.R., Finucane, H.K., et al. (2022). Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat. Genet. *54*, 450–458. 10.1038/s41588-022-01036-9.

24. Ruan, Y., Lin, Y.-F., Feng, Y.-C.A., Chen, C.-Y., Lam, M., Guo, Z., Stanley Global Asia Initiatives, He, L., Sawa, A., Martin, A.R., et al. (2022). Improving polygenic prediction in ancestrally diverse populations. Nat. Genet. *54*, 573–580. 10.1038/s41588-022-01054-7.

25. Zhang, H., Zhan, J., Jin, J., Ahearn, T.U., Yu, Z., O'Connell, J., Jiang, Y., Chen, T., Garcia-Closas, M., Lin, X., et al. Novel Methods for Multi-ancestry Polygenic Prediction and their Evaluations in 3.7 Million Individuals of Diverse Ancestry. 10.1101/2022.03.24.485519.

26. Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. *14*, 507–515. 10.1038/nrg3457.

27. Daetwyler, H.D., Villanueva, B., and Woolliams, J.A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One *3*, e3395. 10.1371/journal.pone.0003395.

28. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al. (2017). Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. *27*, S2–S8. 10.1016/j.je.2016.12.005.

29. Morales, J., Welter, D., Bowler, E.H., Cerezo, M., Harris, L.W., McMahon, A.C., Hall, P., Junkins, H.A., Milano, A., Hastings, E., et al. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol. *19*, 21. 10.1186/s13059-018-1396-2.

30. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K., et al. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. Elife *8*. 10.7554/eLife.39725.

31. Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W., Hirschhorn, J., Daly, M.J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife *8*. 10.7554/eLife.39702.

32. Lehmann, B.C.L., Mackintosh, M., McVean, G., and Holmes, C.C. Optimal strategies for learning multi-ancestry polygenic scores vary across traits. 10.1101/2021.01.15.426781.

33. Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, and Price, A.L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. *41*, 811–823. 10.1002/gepi.22083.

34. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., Furlotte, N.A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet. *50*, 229–237. 10.1038/s41588-017-0009-4.

35. Tsuo, K., Zhou, W., Wang, Y., Kanai, M., Namba, S., Gupta, R., Majara, L., Nkambule, L.L., Morisaki, T., Okada, Y., et al. (2022). Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. Cell Genomics *2*, 100212. 10.1016/j.xgen.2022.100212.

36. Graham, S.E., Clarke, S.L., Wu, K.-H.H., Kanoni, S., Zajac, G.J.M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. Nature *600*, 675–679. 10.1038/s41586-021-04064-3.

37. Ni, G., Zeng, J., Revez, J.A., Wang, Y., Zheng, Z., Ge, T., Restuadi, R., Kiewa, J., Nyholt, D.R., Coleman, J.R.I., et al. (2021). A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. Biol. Psychiatry *90*, 611–620. 10.1016/j.biopsych.2021.04.018.

38. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 5086. 10.1038/s41467-019-12653-0.

39. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776. 10.1038/s41467-019-09718-5.

40. Hou, K., Ding, Y., Xu, Z., Wu, Y., Bhattacharya, A., Mester, R., Belbin, G., Conti, D., Darst, B.F., Fornage, M., et al. (2022). Causal effects on complex traits are similar across segments of different continental ancestries within admixed individuals. bioRxiv. 10.1101/2022.08.16.22278868.

41. Lambert, S.A., Gil, L., Jupp, S., Ritchie, S.C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., et al. (2021). The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat. Genet. *53*, 420–425. 10.1038/s41588-021-00783-5.

42. Kanai, M., Elzur, R., Zhou, W., Zhou, W., Kanai, M., Wu, K.-H.H., Rasheed, H., Tsuo, K., Hirbo, J.B., Wang, Y., et al. (2022). Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. Cell Genomics, 100210. 10.1016/j.xgen.2022.100210.

43. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease

SNPs. Bioinformatics *27*, 2304–2305. 10.1093/bioinformatics/btr341.

44. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. 10.1038/nature15393.

45. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. 10.1093/bioinformatics/btq559.

46. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7. 10.1186/s13742-015-0047-8.

47. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191. 10.1093/bioinformatics/btq340.

48. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104. 10.1126/science.1153717.

49. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free Estimation of Recent Genetic Relatedness. Am. J. Hum. Genet. *98*, 127–148. 10.1016/j.ajhg.2015.11.022.

50. Loh, P.-R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816. 10.1038/ng.3571.

51. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287. 10.1038/ng.3656.

52. Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat. Commun. *10*, 4393. 10.1038/s41467-019-12276-5.

53. Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. Nat. Genet. *53*, 1415–1424. 10.1038/s41588-021-00931-x.

54. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O’Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. 10.1038/s41586-018-0579-z.

55. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400. 10.1038/s41588-018-0047-6.

56. Zeng, J., Xue, A., Jiang, L., Lloyd-Jones, L.R., Wu, Y., Wang, H., Zheng, Z., Yengo, L., Kemper, K.E., Goddard, M.E., et al. (2021). Widespread signatures of natural selection across human complex traits and functional genomic categories. Nat. Commun. *12*, 1164. 10.1038/s41467-021-21446-3.

57. International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58. 10.1038/nature09298.

# Main Figures

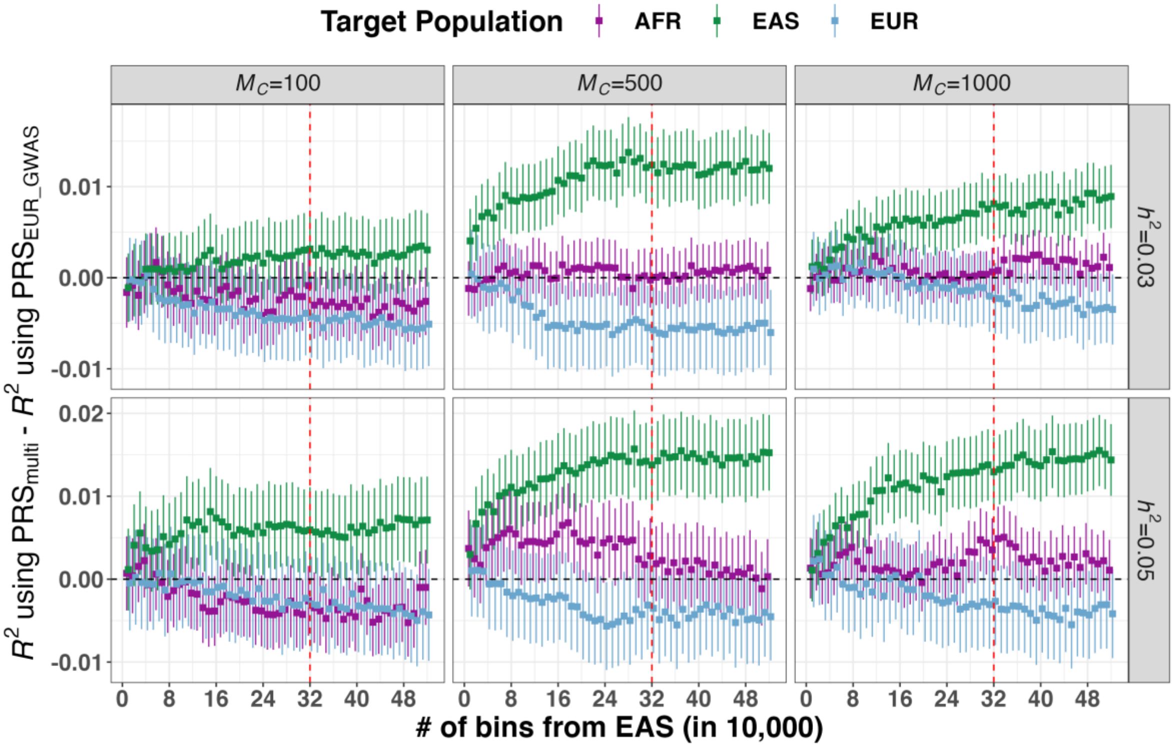


Figure 1. Predictive accuracy improvement of PRS using meta-analyzed multi-ancestry (EUR and EAS) GWAS compared to using EUR GWAS in 6 simulated genetic architectures.

We illustrated the results using 32 EUR bins as an example. PRS was evaluated in AFR, EAS and EUR, respectively. Full results are shown in **Table S1**. The red vertical dashed line in each panel indicates the point where the number of bins from EUR and EAS populations is the same. The black horizontal dashed line indicates y=0. The error bars represent the standard errors of predictive accuracy differences using PRS derived from multi-ancestry GWAS ($PRS_{multi}$) and EUR GWAS ($PRS_{EUR\_GWAS}$), respectively.

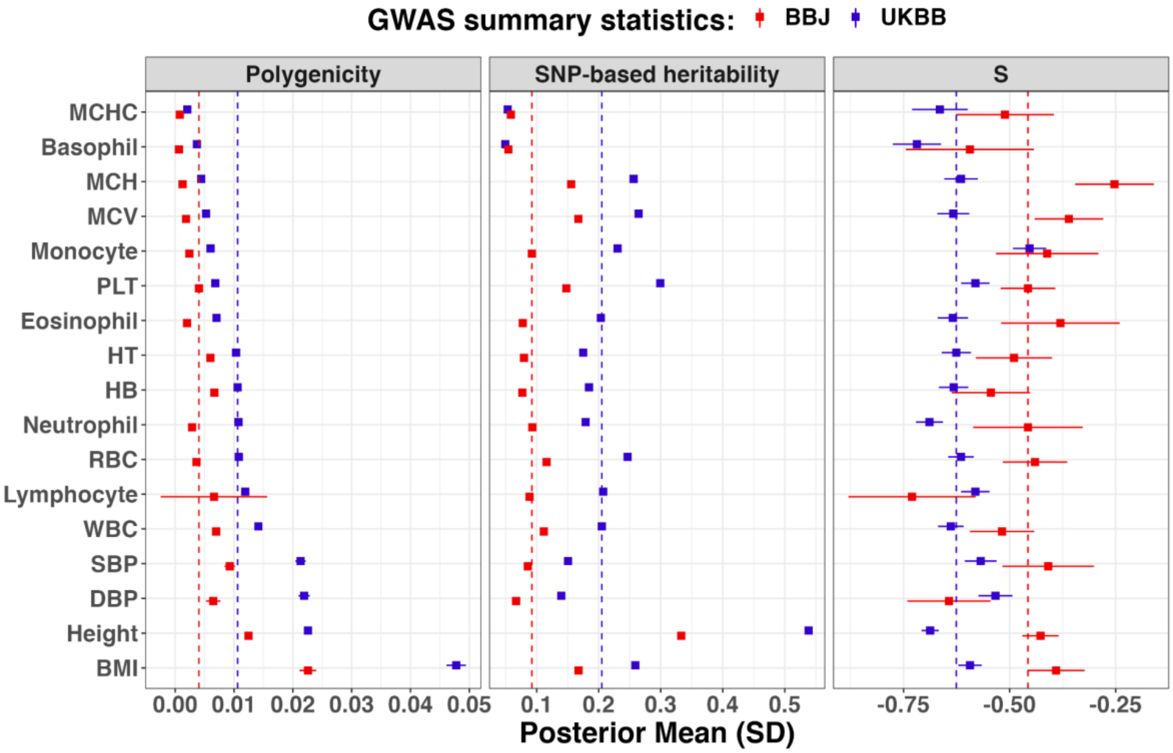
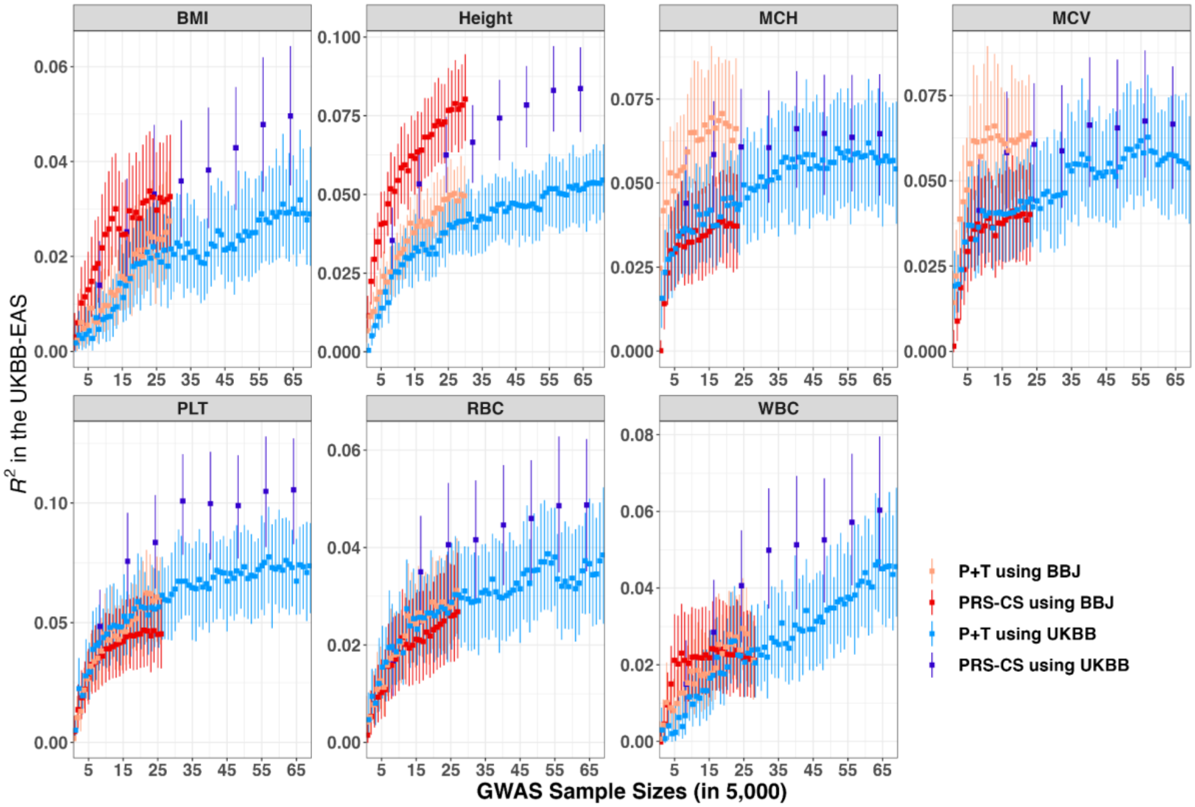


Figure 2: Genetic architecture of 17 studied traits between Biobank Japan (BBJ) and UK Biobank (UKBB).

The phenotypes were ranked according to their polygenicity estimates using GWAS from UKBB. The error bar was the standard deviation of the corresponding estimate. Trait abbreviations are shown in **Table S2**. The vertical dashed line was the median estimate. Full results are shown in **Table S3**.



## Figure 3. Predictive performance of 17 traits in the UK Biobank East-Asian population using P+T and PRS-CS.

We used GWAS from both Biobank Japan (BBJ) and UK Biobank (UKBB) to construct PRS. We reported the predictive accuracy in the UK Biobank East-Asian population (UKBB-EAS) using the auto model for PRS-CS and optimal *p*-value for P+T (see **Methods**). We showed the results for 7 traits with SNP-based heritability > 0.1 in both BBJ and UKBB, while they were ranked by polygenicity estimated using UKBB (**Figure 2**). Trait abbreviations are all described further in **Table S2**. Full results for all traits are shown in **Table S4**.

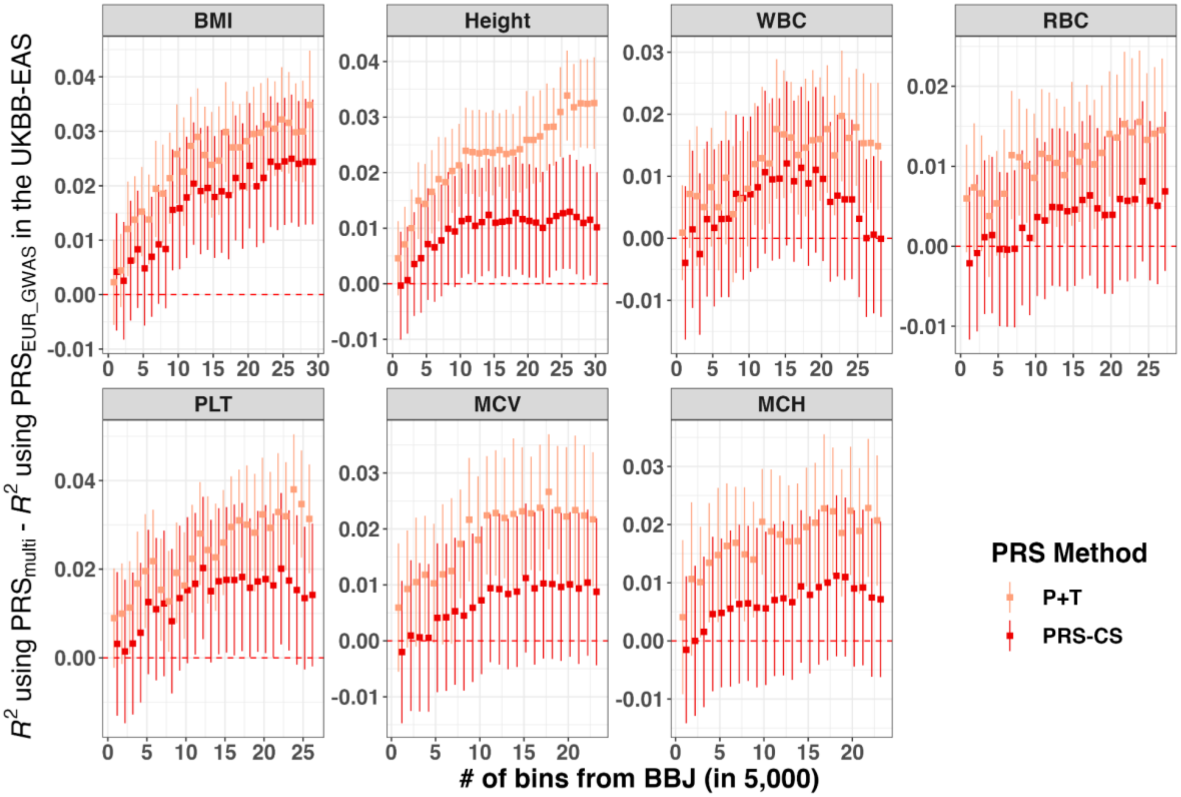


Figure 4. Accuracy improvement of PRS in the UK Biobank East-Asian population using multi-ancestry GWAS relative to using EUR GWAS for P+T and PRS-CS.

We constructed PRS using P+T and PRS-CS and evaluated them in the UK Biobank East-Asian population (UKBB-EAS). The y-axis is the accuracy difference of PRS between using multi-ancestry GWAS ($PRS_{multi}$) and using EUR GWAS ($PRS_{EUR\_GWAS}$) when the number of bins from EUR GWAS is 64. The x-axis is the number of bins from BBJ included in the multi-ancestry GWAS. The error bars indicate the standard error of mean accuracy improvement. The red dashed line is y=0. The red dashed line is y=0. We showed the results for 7 traits with SNP-based heritability > 0.1 in both Biobank Japan (BBJ) and UK Biobank (UKBB), while they were ranked by polygenicity estimates using UKBB (**Figure 2**). Trait abbreviations are all described further in **Table S2**. Full results are shown in **Table S6**.

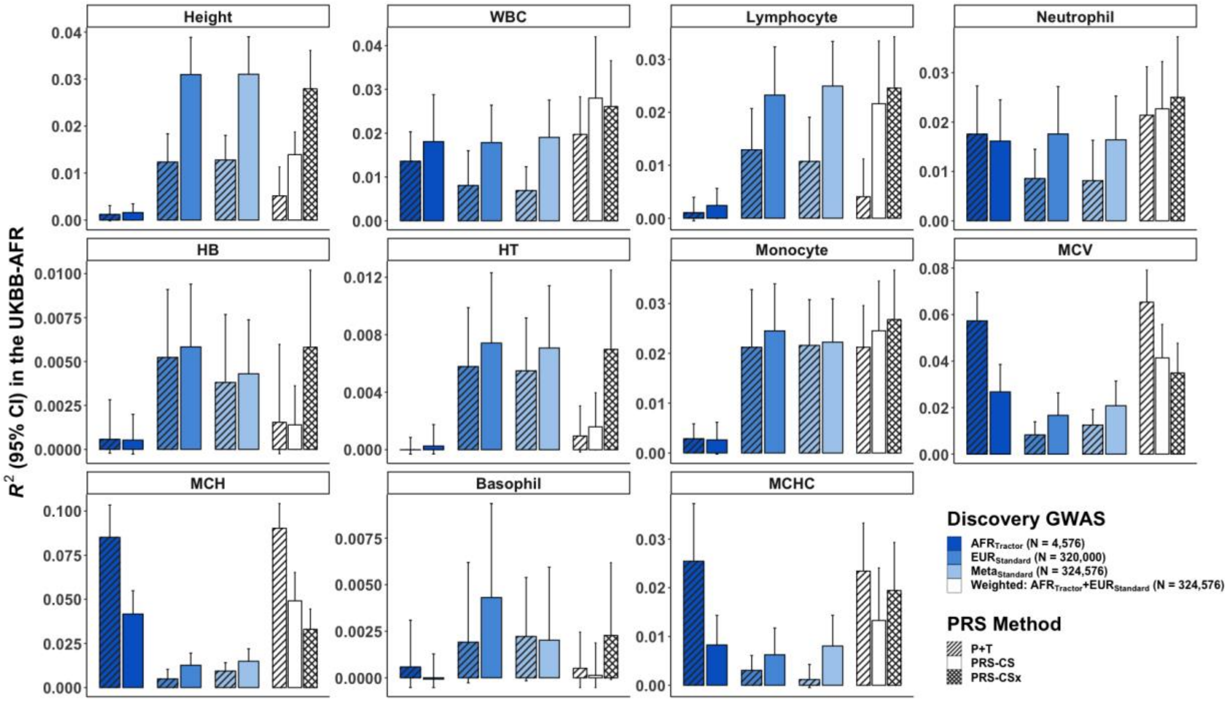


Figure 5. Predictive accuracy for P+T and PRS-CS/PRS-CSx in the UK Biobank African population using various discovery GWAS.

$AFR_{Tractor}$ denotes the AFR-specific GWAS performed using Tractor. $EUR_{Standard}$ refers to standard GWAS performed in the European population in the UKBB. $Meta_{Standard}$ is the meta-analysis performed on $AFR_{Tractor}$ and $EUR_{Standard}$. The weighted PRS was constructed through a linear combination of PRS generated from $AFR_{Tractor}$ and $EUR_{Standard}$, respectively, using various methods including P+T and PRS-CS. Further, we also constructed weighted PRS using PRS-CSx where the input GWAS were $AFR_{Tractor}$ and $EUR_{Standard}$. This figure shows the results for traits with SNP-based heritability > 0.1 in the UK Biobank African population (UKBB-AFR); full results are shown in **Table S8**.

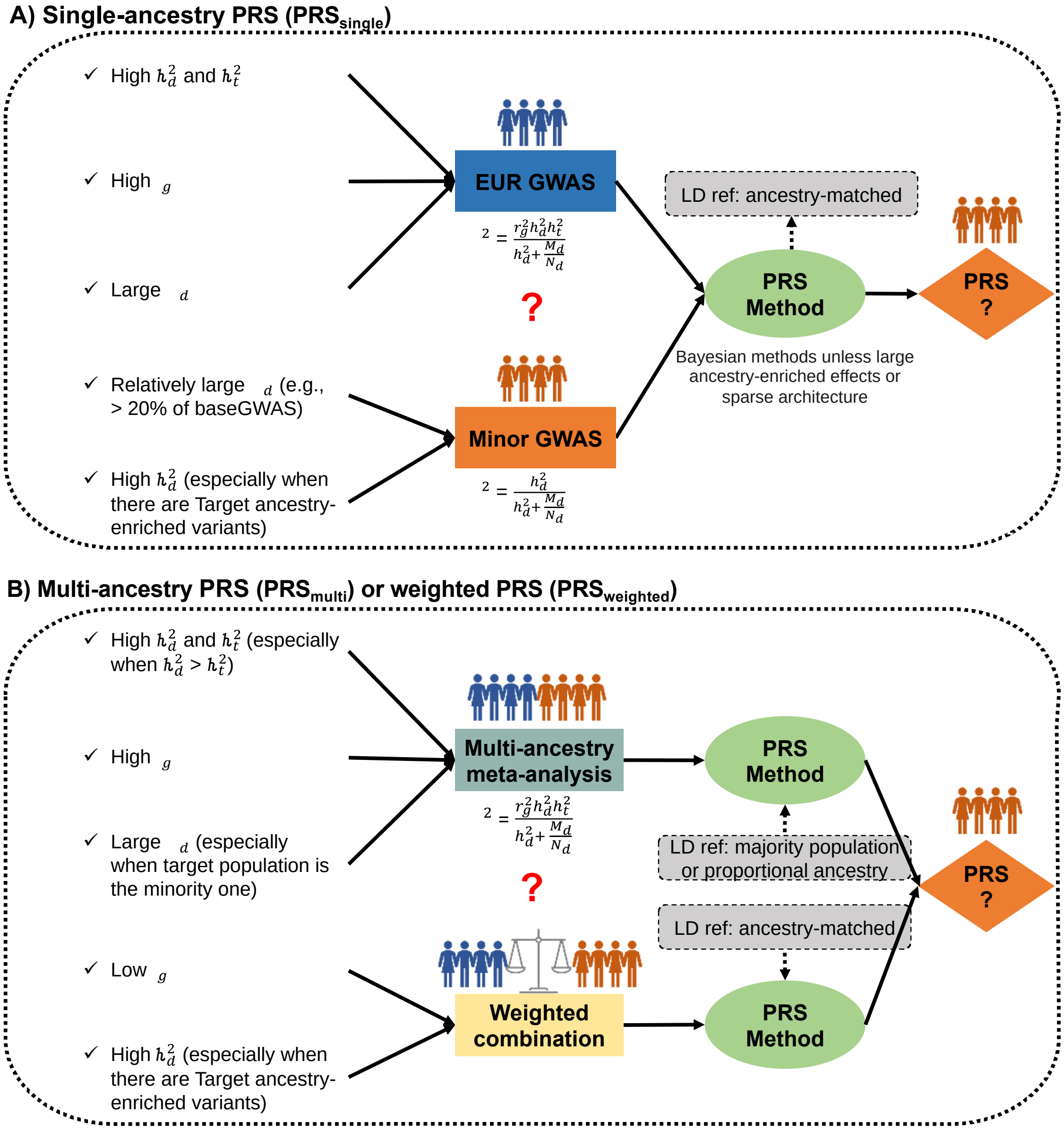

A) Single-ancestry PRS ($PRS_{single}$)
✓ High $h_d^2$ and $h_t^2$
✓ High $_g$
✓ Large $_d$
✓ Relatively large $_d$ (e.g., > 20% of baseGWAS)
✓ High $h_d^2$ (especially when there are Target ancestry-enriched variants)
EUR GWAS
$^2 = \frac{r_g^2 h_d^2 h_t^2}{h_d^2 + \frac{M_d}{N_d}}$
?
Minor GWAS
$^2 = \frac{h_d^2}{h_d^2 + \frac{M_d}{N_d}}$
LD ref: ancestry-matched
PRS Method
Bayesian methods unless large ancestry-enriched effects or sparse architecture
PRS ?
B) Multi-ancestry PRS ($PRS_{multi}$) or weighted PRS ($PRS_{weighted}$)
✓ High $h_d^2$ and $h_t^2$ (especially when $h_d^2 > h_t^2$)
✓ High $_g$
✓ Large $_d$ (especially when target population is the minority one)
✓ Low $_g$
✓ High $h_d^2$ (especially when there are Target ancestry-enriched variants)
Multi-ancestry meta-analysis
$^2 = \frac{r_g^2 h_d^2 h_t^2}{h_d^2 + \frac{M_d}{N_d}}$
?
Weighted combination
PRS Method
LD ref: majority population or proportional ancestry
LD ref: ancestry-matched
PRS Method
PRS ?

Figure 6. General practices for developing PRS using different discovery GWAS.

We summarized the general practice for developing PRS A) using single-ancestry GWAS ($PRS_{single}$); and B) using GWAS from multiple ancestries ($PRS_{multi}$ or $PRS_{weighted}$). For $PRS_{single}$, we can compare the expected accuracies either using underrepresented target-ancestry matched GWAS (Minor GWAS) or large-scale European-based GWAS (EUR GWAS) and choose the input GWAS for PRS method based on prior information including cross-ancestry genetic correlation ($r_g$), SNP-based heritability in discovery ($h_d^2$) and target populations ($h_t^2$), discovery GWAS sample size ($N_d$) and the number of genome-wide independent segments in the discovery population

($M_d$). For PRS$_{\text{multi}}$, meta-analysis is generally recommended whilst the linear weighted combination shows its superiority for traits with ancestry-enriched variants.